Measuring Connection Strengths and Link Strengths in Discrete Bayesian Networks

Georgia Tech Research Report: GT-IIC-07-01
January 29, 2007

Imme Ebert-Uphoff *(ebert@me.gatech.edu)
Adjunct Associate Professor, Robotics and Intelligent Machines Center
College of Computing, Interactive & Intelligent Computing Division
Georgia Institute of Technology
Atlanta, Georgia 30308, USA

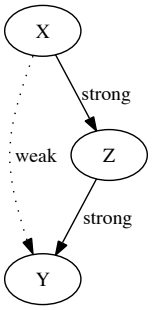
Abstract

This paper discusses measures for connection strength (strength between any two nodes) and link strength (strength along a specific edge) in Discrete Bayesian Networks. The typical application is to visualize the connections in a Bayesian Network learned from data to learn more about the inherent properties of the system (e.g. in earth sciences, biology or medicine).

The paper focuses on measures based on mutual information and conditional mutual information. The goal is to provide an easy-to-read document that gives clear reasoning for existing measures, provides some simple extensions (modified measures for different applications), discusses the limitations of the measures, provides enough interpretation to aid a scientist in selecting the most appropriate one and suggests some new uses for link strength.

1 LINK STRENGTH AND CONNECTION STRENGTH

Boerlage was the first to formally introduce the concepts of link strength versus connection strength for Bayesian Networks with binary nodes (Boerlage 1992). Boerlage defines *connection strength* for any pair of nodes (adjacent or not) to measure the strength between those nodes taking any possible path between them into account. In contrast *link strength* (also known as *arc weight*) is defined for a specific edge and measures the strength of connection only along that single edge.



$\mathbf{X}$:
$P(X = True) = 0.5$
$\mathbf{Z}$:
$P(Z = True|X = True) = 0.9$
$P(Z = True|X = False) = 0.1$
$\mathbf{Y}$:
$P(Y = True|X = True, Z = True) = 0.9$
$P(Y = True|X = False, Z = True) = 0.89$
$P(Y = True|X = True, Z = False) = 0.1$
$P(Y = True|X = False, Z = False) = 0.11$

Figure 1: Sample BN with weak link from X to Y, but strong links from X to Z and Z to Y.

To demonstrate the difference between these concepts in particular for adjacent nodes consider the network in Figure 1. Each of the three nodes only has two states, $True$ and $False$. Let us focus on the connection between nodes X and Y. For this sample network the *direct* link from X to Y is weak[1] , while the indirect link from X to Y *through* Z is very strong. According to the above (vague) concept definitions, the connection strength, CS, between X and Y is strong here, but the link strength, LS, of the edge $X \rightarrow Y$ is weak:

$$\begin{aligned} \mathrm{CS}(X, Y) &= \text{strong}, \\ \mathrm{LS}(X \rightarrow Y) &= \text{weak}. \end{aligned}$$

Any pair of measures for link strength and connection strength should yield this result for the above example.

1.1 APPLICATIONS

Bayesian Networks have become a tool to learn about the inherent structure of systems in disciplines ranging from earth sciences to medicine, biology and the social sciences. For this purpose both the network structure

*Joint appointment with the George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0405.

[1]This becomes obvious by noting that the state of X has little effect on the values of $P(Y = True|X, Z)$.

and probabilities are often learned from data and the resulting structures are visualized as graphs to learn about potential connections between the variables. For this purpose it is helpful to visualize not only the *existence* of arrows, but also the *strength* of the various connections. Both connection strength and link strength are useful for that purpose (Boerlage 1992). They can also be used to generate explanations for reasoning in Bayesian Networks.

In the context of causal discovery, mutual information was used by Cheng et al. (2002) to identify node pairs with strong connections in a proposed constraint-based learning algorithm. Unfortunately, Cheng et al.'s algorithm is based on an unrealistic assumption and thus not reliable, as shown by Chickering and Meek (2003). Nevertheless, the basic idea is a good one and is likely to eventually lead to correct and more efficient structure learning algorithms. Furthermore, we believe that there are several new applications for link strength in causal discovery as outlined in the future work section (Section 7).

Connection strength has also been used for approximate inference. Namely, connection strength can be used to determine the influence of variables, i.e. to determine which ones can be neglected in the approximation (Jitnah and Nicholson 1998). Since connection strength is generally computationally expensive to calculate, link strength measures are sometimes employed to approximate connection strength quickly. For that purpose the link strength measures must be computationally efficient and combinations of them must yield either bounds on or a decent approximation of connection strengths between any two nodes (Jitnah and Nicholson 1998).

The measures in this paper are targeted toward interpretation of Bayesian Networks and potentially for causal discovery, rather than approximate inference. Thus computational complexity takes a back seat.

1.2 OVERVIEW OF MEASURES

The following measures are discussed in this paper:
(1) Entropy (Shannon 1949) is used to measure the uncertainty in a single node.
(2) Mutual information (Shannon 1949, Pearl 1988) is used to measure connection strength.
(3) Two variations of the link strength measure in (Nicholson and Jitnah 1998) are presented: *True Average Link Strength* and *Blind Average Link Strength*.
(4) Mutual Information *Percentage* and Link Strength *Percentage* are proposed to measure the *percentage* of the existing uncertainty that has been removed.
All measures are defined in this document only for *discrete* Bayesian Networks.

1.3 ADDITIONAL LITERATURE

The most important work related to this article, as evident from the previous sections, consists of Shannon's definitions of Entropy and Mutual Information (Shannon 1949), Pearl's use of Mutual Information for Bayesian Networks (Pearl 1988), Boerlage's definition of link strength and connection strength for Bayesian networks with binary nodes (Boerlage 1992) and Jitnah and Nicholson's definition of link strength for fast approximate inference (Jitnah and Nicholson 1998).

Other work of interest – although not used here – is the work by Lacave and Diez (Lacave and Diez 2004) proposing a measure for the "magnitude of influence" of two ordinal variables and displaying it by the thickness of an arc. Other visualization techniques are reviewed in (Lacave and Diez 2002) and (Zapata-Riviera et al. 1999), but no other measures for link or connection strengths are presented in those articles.

2 ENTROPY AS UNCERTAINTY MEASURE

Entropy was already defined by Shannon in the late 1940s (Shannon 1949) and has become the most commonly used measure for the uncertainty of a random variable.

Definition The *entropy* of a discrete random variable, X, is defined as

$$U(X) \quad = \quad \sum_{x_i} P(x_i) \log_2 \frac{1}{P(x_i)}. \qquad (1)$$

Some readers may be more familiar with the expression $U(X) = -\sum_{x_i} P(x_i) \log_2 P(x_i)$, which is identical to (1).

Interpretation How much uncertainty is there in X if no evidence is given for any of the nodes?

Entropy forms the basis for all connection and link strength measures discussed in this paper and thus it is very important to understand its limitations. Those are rarely discussed in textbooks and thus briefly reviewed in Appendix A. While it is important to *understand* these limitations, it is unlikely, based on the nature of the limitations, that they can be overcome by other generally applicable uncertainty measures and entropy thus remains by far the most popular measure for uncertainty.

3 MEASURES FOR CONNECTION STRENGTH

Connection strength between X and Y measures how strongly information on the state of X affects the state of Y (and vice versa). The standard approach is to compare the distribution of Y *without* any evidence to the distribution of Y if there *is* evidence for X. Mutual Information is the most common implementation of this idea: one simply calculates $U(Y)$ and $U(Y|X)$ and compares them (see Section 3.1).

An alternative is to apply a divergence measure between the two probability distributions of Y and $Y|X$. For example, in their earlier work Nicholson and Jitnah apply the Bhattacharyya distance (Nicholson and Jitnah 1997) to the distributions. However, that approach yields less suitable results than Mutual Information (Nicholson and Jitnah 1998).

3.1 MUTUAL INFORMATION

Shannon (Shannon 1949) introduced Mutual Information for the purpose of communication theory. Pearl (Pearl 1988) was the first to propose the use of mutual information to measure connection strength in Bayesian Networks to determine the relevance of some nodes on others.

Definition *Mutual Information* is defined as

$$MI(X,Y) \quad = \quad U(Y) - U(Y|X), \tag{2}$$

where $U(Y|X)$ is calculated by averaging $U(Y|x_i)$ over all possible states x_i of X, taking $P(x_i)$ into account:

$$U(Y|X) \quad = \quad \sum_{x_i} P(x_i) U(Y|x_i). \tag{3}$$

Simple arithmetic transformations yield the formula:

$$MI(X,Y) \quad = \quad \sum_{x,y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)} \right).$$

Mutual Information is symmetric in X and Y, i.e. $MI(X,Y) = MI(Y,X)$.

Interpretation How much is the uncertainty in Y reduced by knowing the state of X? How much is the uncertainty in X reduced by knowing the state of Y?

3.2 MUTUAL INFORMATION PERCENTAGE

In some cases the *absolute amount* of uncertainty reduction in a variable may provide less insight than the *percentage* of the original uncertainty that was removed. Thus we propose a simple extension of Mutual Information, namely Mutual Information Percentage, to be used in conjunction with Mutual Information.

Definition *Mutual Information Percentage* is defined for $U(Y) \neq 0$ as

$$\begin{aligned} MI\%(X,Y) \quad &= \quad \frac{MI(X,Y)}{U(Y)} \cdot 100 \\ &= \quad \frac{U(Y) - U(Y|X)}{U(Y)} \cdot 100. \end{aligned}$$

Mutual Information is *not* symmetric in X and Y, i.e. $MI\%(X,Y) \neq MI\%(Y,X)$. $MI\%(X,Y)$ is undefined for $U(Y) = 0$, which makes perfect sense: if there is zero uncertainty to begin with, then it makes no sense to ask what percentage of it was removed.

Interpretation By how many percentage points is uncertainty in Y reduced by knowing the state of X?

4 MEASURES FOR LINK STRENGTH

There is much less literature on link strength than on connection strength and it appears to be harder to measure. Boerlage (Boerlage 1992) defined measures for both link strength and connection strength, but those only apply to two-state variables. (Nicholson and Jitnah 1998) and (Jitnah 1999) derived a measure based on conditional mutual information that applies for any discrete Bayesian Network. They did not, however, provide a derivation of their measure, which we try to do in the following subsection.

4.1 TRUE AVERAGE LINK STRENGTH

A definition of link strength of an edge $X \rightarrow Y$ can be derived from the definition of connection strength. When considering a link $X \rightarrow Y$, we need to decide how to deal with the *other* parents of Y in order to focus on the connection from parent X to child Y *solely* along edge $X \rightarrow Y$. The approach used here is to instantiate all *other* parents of Y, leaving the direct connection from X to Y as only pathway through which information can travel from X to Y.

Denoting the set of *other* parents of Y as $\mathbf{Z} = \{Z_1, \ldots Z_n\}$, we can adjust Equation (2) of Mutual Information by conditioning both terms on the right on $\mathbf{Z}$, resulting in the following definition. (We use boldface for $\mathbf{Z}$ and $\mathbf{z}$ to indicate that each represents a *set* of zero, one or more variables.)

Definition *True Average Link Strength* of edge $X \rightarrow Y$ is defined as the mutual information of (X,Y) conditioned on all other parents of Y, namely

$$LS^{true}(X \rightarrow Y) \quad = \quad MI(X,Y|\mathbf{Z})$$

$$= \quad U(Y|\mathbf{Z}) - U(Y|X,\mathbf{Z}),$$

where $U(Y|X,\mathbf{Z})$ is the average over the states of all parents and is defined as

$$U(Y|X,\mathbf{Z}) \;=\; \sum_{x,\mathbf{z}} P(x,\mathbf{z}) U(Y|x,\mathbf{z}) \qquad (4)$$

$$= \quad \sum_{x,\mathbf{z}} P(x,\mathbf{z}) \sum_{y} P(y|x,\mathbf{z}) \log_2 \left(\frac{1}{P(y|x,\mathbf{z})} \right),$$

and $U(Y|\mathbf{Z})$ is defined analogously as the average over all *other* parents:

$$U(Y|\mathbf{Z}) \;=\; \sum_{\mathbf{z}} P(\mathbf{z}) U(Y|\mathbf{z}), \qquad (5)$$

where $\mathbf{z}$ represents all possible state *combinations* of the set of other parents, $\mathbf{Z}$.

Instantiating all other parents of Y in $MI(X,Y|\mathbf{Z})$ essentially blocks all information flow through the other parents, $\mathbf{Z}$. We still need to ensure that there remain no other indirect open pathways between Y and X, e.g. through descendants of Y. The following theorem shows that indeed the only pathway that remains open between X and Y once all other parents are instantiated is the direct link from X to Y.

Theorem 4.1 *Consider a BN (G,P) consisting of DAG G and joint probability P. Let $X \rightarrow Y$ be an edge in G and denote the set of all other parents of Y as Z. Let $\widehat{G}$ be the modified DAG generated by deleting edge $X \rightarrow Y$ in G. Then X and Y are conditionally independent given $\mathbf{Z}$ in BN ($\widehat{G}$,$\widehat{P}$) for any joint probability $\widehat{P}$.*

Proof Since edge $X \rightarrow Y$ does not exist in $\widehat{G}$, set $\mathbf{Z}$ represents *all* parents of Y in $\widehat{G}$. Furthermore, X is not a descendent of Y in $\widehat{G}$ - otherwise the original DAG G would contain a directed cycle. Due to the Markov condition any node in a BN is conditionally independent of its non-descendents given its parents. Therefore for any BN with DAG $\widehat{G}$, node Y is conditionally independent of X given $\mathbf{Z}$. ∎

Since X and Y are conditionally independent given $\mathbf{Z}$ if the edge from X to Y is removed, it is clear that edge $X \rightarrow Y$ is indeed the *only* path along which information can flow from X to Y in G if $\mathbf{Z}$ is instantiated. This fact ensures that the True Average Link Strength indeed only measures information flow along the considered edge.

Using (4) and (5) and some transformations yields

$$LS^{true}(X \rightarrow Y) = \sum_{x,\mathbf{z}} P(x,\mathbf{z}) \sum_{y} P(y|x,\mathbf{z}) \log_2 \frac{P(y|x,\mathbf{z})}{P(y|\mathbf{z})}. \qquad (6)$$

Interpretation By how much is the uncertainty in Y reduced by knowing the state of X, if the states of all other parent variables are known (averaged over the parent states using their *actual* joint probability)?

Comparison to Measure by Jitnah and Nicholson: By converting to our notation the measure presented by (Nicholson and Jitnah 1998) and (Jitnah 1999) can be written as

$$LS^{Jitnah-Nicholson}(X \rightarrow Y) = \sum_{x,\mathbf{z}} P_{pr}(\mathbf{z}) P_{pr}(x) \sum_{y} P(y|x,\mathbf{z}) \log_2 \frac{P(y|x,\mathbf{z})}{P_{pr}(y|\mathbf{z})},$$

where the term P_{pr} indicates an approximation of probability that avoids using any inference. Thus Equation (6) presents the exact formula, while the measure by (Nicholson and Jitnah 1998) already employs some approximations to allow for faster evaluation.

4.2 BLIND AVERAGE LINK STRENGTH

This measure is new and is derived from True Average Link Strength by disregarding the actual frequency of occurrence of the parent states. Namely we assume that $X, \mathbf{Z}$ are independent and all uniformly distributed:

$$\hat{P}(x,\mathbf{z}) = P(x)P(\mathbf{z}),\; \hat{P}(x) = \frac{1}{\#(X)},\; \hat{P}(\mathbf{z}) = \frac{1}{\#(\mathbf{Z})}, \quad (7)$$

where $\#(X)$ denotes the number of discrete states of X, etc.

Essentially, this approximation goes one step further in simplifications than the approximations by Jitnah and Nicholson. However, these additional simplifications have a justification of their own. Namely an interesting property of the set of assumptions (7) is that it creates a local measure that depends only on the child node and its conditional probability table, but nothing else in the network. One may argue that for some applications such a local measure is actually more natural, since the connection between parents and child should be independent of any changes in probabilities elsewhere in the network. This discussion is continued in Section 6.3.

Definition *Blind Average Link Strength* is defined as

$$LS^{blind}(X \rightarrow Y) \quad = \quad \hat{U}(Y|\mathbf{Z}) - \hat{U}(Y|X,\mathbf{Z}),$$

where

$$\hat{U}(Y|\mathbf{Z}) = \frac{1}{\#(X)\#(\mathbf{Z})} \sum_{x,y,\mathbf{z}} P(y|x,\mathbf{z}) \log_2 \frac{\#(X)}{\sum_x P(y|x,\mathbf{z})},$$

$$\hat{U}(Y|X,\mathbf{Z}) = \frac{1}{\#(X)\#(\mathbf{Z})} \sum_{x,y,\mathbf{z}} P(y|x,\mathbf{z}) \log_2 P(y|x,\mathbf{z}).$$

Note that $\hat{U}(Y|\mathbf{Z})$ and $\hat{U}(Y|X,\mathbf{Z})$ are obtained from $U(Y|\mathbf{Z})$ and $U(Y|X,\mathbf{Z})$ simply by applying assumptions (7). This definition yields the simple formula

$$LS^{blind}(X \to Y) = \frac{1}{\#(X)\#(\mathbf{Z})} \sum_{x,y,\mathbf{z}} P(y|x,\mathbf{z}) \log_2 \left(\frac{P(y|x,\mathbf{z})}{\frac{1}{\#(X)} \sum_x P(y|x,\mathbf{z})} \right),$$

where $P(y|x,\mathbf{z})$ is given by the conditional probability table of Y and *no* inference is required at all.

Interpretation By how much is the uncertainty in Y reduced by knowing the state of X, if the states of all other parent variables are known (averaged over the parent states assuming all parents are independent of each other and uniformly distributed)?

Comment: This is the simplest and computationally least expensive measure. It is also a local measure, taking only the child and *its* conditional probabilities into account, thus allowing for isolated analysis of child and parents, regardless of the rest of the network.

4.3 LINK STRENGTH PERCENTAGES

Just as percentage of uncertainty reduction can be important in mutual information, the same holds for True Average and Blind Average Link Strength. Therefore we suggest the following two simple extensions of Link Strength:

Definition *True Average Link Strength Percentage* is defined for $U(Y|\mathbf{Z}) \neq 0$ as

$$LS\%^{true}(X \to Y) = \frac{LS^{true}(X \to Y)}{U(Y|\mathbf{Z})} \cdot 100 = \frac{U(Y|\mathbf{Z}) - U(Y|X,\mathbf{Z})}{U(Y|\mathbf{Z})} \cdot 100. \quad (8)$$

Applying independence and uniformity assumptions (7) to the True Average Link Strength Percentage (8) yields the Blind Average Link Strength Percentage.

Definition *Blind Average Link Strength Percentage* is defined for $\hat{U}(Y|\mathbf{Z}) \neq 0$ as

$$LS\%^{blind}(X \to Y) = \frac{LS^{blind}(X \to Y)}{\hat{U}(Y|\mathbf{Z})} \cdot 100 = \frac{\hat{U}(Y|\mathbf{Z}) - \hat{U}(Y|X,\mathbf{Z})}{\hat{U}(Y|\mathbf{Z})} \cdot 100.$$

Analogously to $MI\%$ (and for the same reasons), $LS\%^{true}(X \to Y)$ is undefined if $U(Y|\mathbf{Z}) = 0$ and $LS\%^{blind}(X \to Y)$ is undefined if $\hat{U}(Y|\mathbf{Z}) = 0$.

Interpretation By how many percentage points is the uncertainty in Y reduced by knowing the state of X, if the states of all other parent variables are known (averaged over the parent states using their *actual* joint probability (for True Average) or assuming all parents are independent of each other and uniformly distributed (for Blind Average))?

5 COMPUTATIONAL ISSUES

All the measures discussed in the previous sections plus several visualization routines were implemented by the author as add-ons for two different software packages.
(1) LinkConnectionStrength package (Ebert-Uphoff 2006) is an add-on for Intel's Open-Source Probabilistic Network Library (PNL).
(2) LinkStrength package is an add-on for Kevin Murphy's Bayes Net Toolbox (BNT) for Matlab.
Sources and documentation for both packages are available at www.DataOnStage.com.

5.1 DEGENERATE CASES

Considering the formulas for entropy, mutual information and link strengths turns up a variety of potential degenerate cases that would lead to either (1) division by zero, (2) calculating the logarithm of zero, or (3) calculating an undefined expression such as $P(y|x)$ for $P(x) = 0$. Fortunately, careful analysis shows that in all of those cases the expressions in question converge towards zero when approaching the degenerate case and can thus be handled by simple if-statements in the code.

5.2 COMPUTATIONAL COMPLEXITY

The computation with the highest computational complexity in all of the connection strength and link strength formulas appears to be the inference used to calculate the various required joint probabilities. The inference requirements are as follows:
- $CS(X,Y)$ requires $P(X,Y)$.
- $LS^{true}(X \to Y)$ requires P(all parents of Y).
- $LS^{blind}(X \to Y)$ requires no inference at all.
- Each percentage measure requires the same probabilities as the corresponding absolute measure above.

6 MORE PROPERTIES AND INTERPRETATION

This section provides additional intuition on the measures by presenting some properties and illustrating them by several examples.

Table 1: Results for Sample Network in Figure 1

	LS^{true}	LS^{blind}	MI
$X \to Y$	0.000	0.000	0.311
$X \to Z$	0.531	0.531	0.531
$Z \to Y$	0.204	0.516	0.515

6.1 DO OUR MEASURES BEHAVE AS DESIRED?

Let us revisit the network from Section 1 (Figure 1) used to demonstrate the desired difference in behavior between connection strength and link strength and see whether the measures defined here actually behave in the desired way. Table 1 shows the results for the network from Figure 1 for True Average and Blind Average Link Strength for each edge, as well as Mutual Information for each node pair. The values are consistent with the expectations for link strength and connection strength specified in Section 1, specifically:

Link Strength: No matter which formula is used (True Average or Blind Average), the link strengths of the arcs from X to Z and from Z to Y are significant, while the strength of arc $X \to Y$ nearly vanishes.

Connection Strength: Each pair of nodes, (X, Y), (X, Z) and (Y, Z), is strongly connected. In particular, the pair of nodes (X, Y) receives a strong connectivity value, because X and Y are strongly connected through the chain $X \to Z \to Y$.

6.2 MUTUAL INFORMATION VERSUS TRUE AVERAGE LINK STRENGTH

It is clear from their definitions that for a node with only one parent Mutual Information and True Average Link Strength yield the same value. Mutual Information Percentage and True Average Link Strength Percentage also coincide in this case.

In contrast, let us consider the node pair (X, Y) in the simple 3-node network

$$X \to Y \leftarrow Z.$$

Since the *only* connection between X and Y is the arc $X \to Y$, one may initially expect that Mutual Information and True Average Link Strength would also coincide for that arc. However, mutual information measures how much uncertainty is removed from Y by knowing the state of X *if nothing else is known.* In contrast, True Average Link Strength measures how much uncertainty is removed from Y by knowing the state of X *if the state of* Z *is known.*

In summary, Mutual Information and True Average Link Strength generally only coincide if the child has only one parent.

6.3 TRUE VERSUS BLIND AVERAGE LINK STRENGTH

Let us consider the *Visit to Asia* network introduced in (Lauritzen and Spiegelhalter 1998). Figure 2, 3 and 4 show True Average Link Strength, Blind Average Link Strength and some selected Mutual Information graphs. In the link strength graphs, the value of the link strength is indicated both by the number next to the arrow and by the gray scale of the arrow (if the arrow would otherwise be invisible, a dashed light gray line is used instead). The mutual information graphs indicate the target node by an octagonal shape and the mutual information of all other nodes relative to that one is indicated both by the value underneath each node and by its gray scale.

As indicated by the True Average Percentages on the right of Figure 2 most links are quite strong. Keeping the comments on scale from Section 6.5 in mind all connections except for the one from *Visit to Asia* to *Tuberculosis* can be classified as significant.

The Blind Average Value Percentage for *Visit to Asia* is much higher though, indicating that the reason for the low True Average Percentage is the low probability of state *True* for *Visit to Asia*. In a nutshell, one could say that in this example **True Average Link Strength (and Percentage) only considers the benefit of the information of variable** *Visit to Asia* **for the** *average* **patient. In contrast Blind Average Link Strength (and Percentage) considers all patient categories equally – in this case the small group of patients actually having traveled to Asia is given equal weight to the large group not having traveled there – and thus gives more attention to special cases (small groups) and the value of information of variable** *Visit to Asia* **for that special group.**

This difference is typical of the different viewpoints of True Average and Blind Average. One should be aware of those viewpoints when choosing one measure for a particular application.

6.4 DETECTING DETERMINISTIC RELATIONSHIPS

This section illustrates interesting properties of the Link Strength *Percentages* for deterministic functions. By deterministic function we mean that the state of a child is completely known if the states of all of its parents are known, i.e. there is *no* uncertainty involved.

Definition A node Y is a *deterministic child* of its

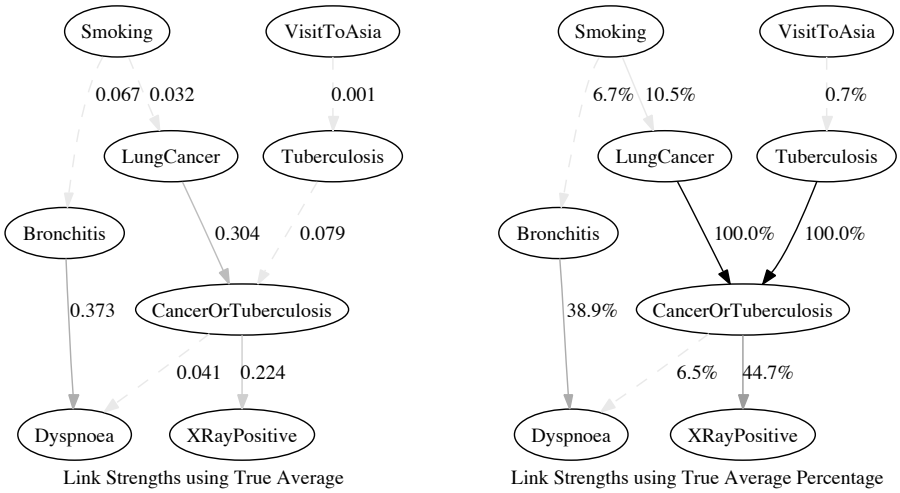


Figure 2: True Average Link Strength (left) and Percentage (right) for Asia Model.

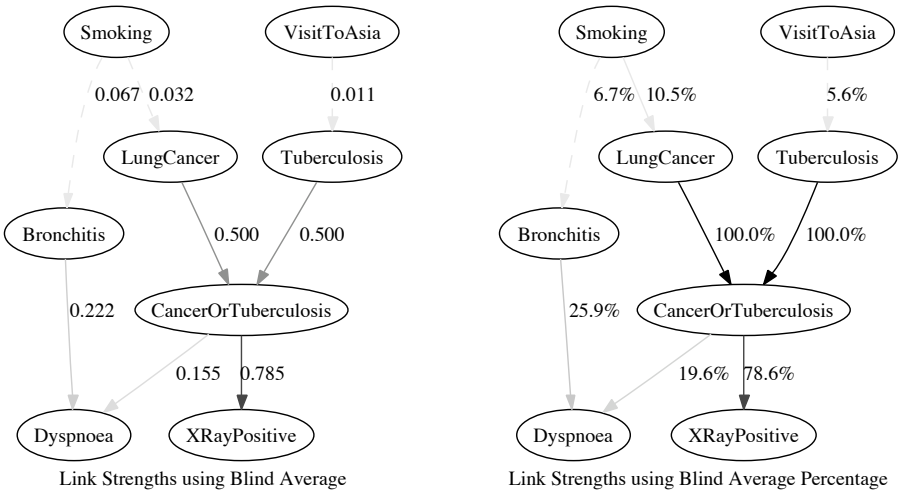


Figure 3: Blind Average Link Strength (left) and Percentage (right) for Asia Model.

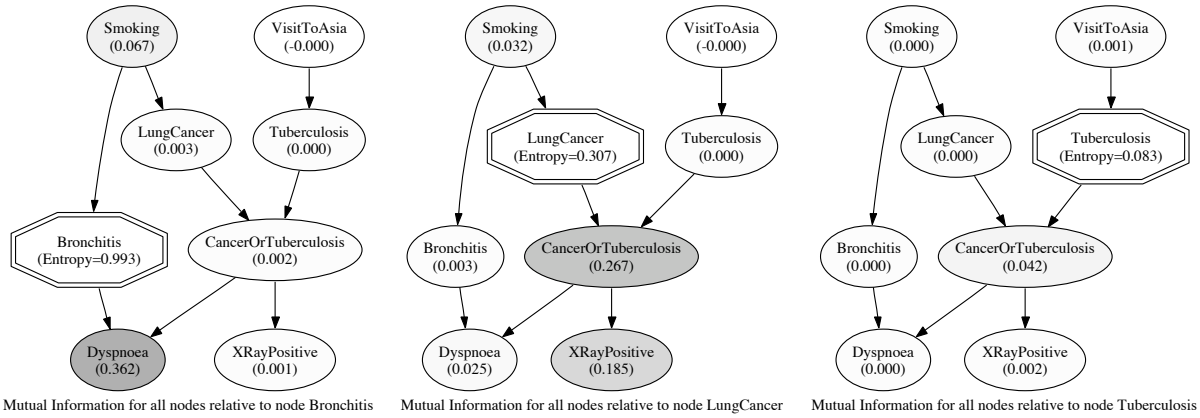


Figure 4: Connection Strength (Mutual Information) relative to node *Bronchitis* (left), *LungCancer* (center) and *Tuberculosis* (right) for Asia Model.

parents, $P_1, \ldots, P_n$, if

$$\forall \text{ states } y, \forall \text{ parent states } p_1, \ldots, p_n : \\ P(y|p_1, \ldots, p_n) \in \{0, 1\}.$$

Proposition 6.1 *If Y is a deterministic child of its parents, then both its True Average and Blind Average Link Strength Percentage from any parent P is* 100%*:*

$$\forall P \in parents(Y): \quad LS^{true}\%(P \to Y) = 100\%$$
$$\forall P \in parents(Y): \quad LS^{blind}\%(P \to Y) = 100\%.$$

Proof See Appendix B.

The question arises whether the reverse is also true, i.e. if the link strength percentages of all parents to a child are 100% does that imply that the child is deterministic? This is indeed the case for Blind Average Percentage, but not for True Average Percentage, as evident from the following two Propositions.

Proposition 6.2 *If $LS^{blind}\%(P \to Y) = 100\%$ for at least one parent P of a node Y, then Y is a deterministic child of its parents.*

Proof See Appendix B.

Remark: it follows that if $LS^{blind}\%(P \to Y) = 100\%$ for *one* of Y's parents, that the same must hold for *all* of Y's parents.

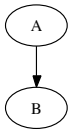
Proposition 6.3 *Even if $LS^{true}\%(P \to Y) = 100\%$ for all parents P of node Y, then Y is not necessarily a deterministic child of its parents.*

Proof See Appendix B.

To see the usefulness in particular of Proposition 6.2 we revisit the Visit to Asia Example. Looking at the plot for the Blind Average Link Strength Percentage (right plot in Figure 3) immediately shows that *CancerOrTuberculosis* is a deterministic child of its parents – which, admittedly, in this case could have been guessed from its name, too. Other cases are less obvious, in particular if a network is learned from data and this property can be helpful to identify deterministic and nearly deterministic child nodes.

6.5 WHICH NUMBERS INDICATE A "STRONG" RELATIONSHIP?

This question cannot be fully answered here, but we try to shed some light on it by considering the trivial example in Figure 5. Nodes A and B are both binary with states $True$ and $False$ and b is a free parameter. Note that for this trivial system it is



$$\mathbf{A}: \; P(A = True) = 0.5$$
$$\mathbf{B}: \; P(B = True|A = True) = b$$
$$P(B = True|A = False) = 1 - b$$

Figure 5: Two-Node Network with parameter b.

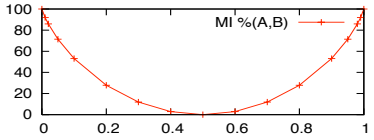


Figure 6: $MI\%(A, B)$ vs. b for Network in Figure 5.

$MI(A, B) = LS^{true}(A \to B) = LS^{blind}(A \to B)$, since B only has a single, uniformly distributed parent (A).

How do mutual information and link strength "scale" for this network, i.e. what values do they result in for varying b? Table 2 (on next page) and Figure 6 show results for $MI(A, B)$ and $MI\%(A, B)$ – and thus also for $LS^{true/blind}(A \to B)$ and $LS\%^{true/blind}(A \to B)$ – for a variety of values of b. Notice *how* quickly $MI(A, B)$ decreases when increasing b from zero. For example, for $b = 0.1$ we know that in 90% of cases B is $True$ if and only if A is $False$. However, the connection/link strength value is only 0.531 with a percentage value of 53.1%. Similarly, even for $b = 0.4$ we know that A still has a *significant* effect on B, but the percentage value of removed uncertainty is only 2.9%.

The lesson from this is that while the values of the measures increase monotonously when uncertainty is reduced, the *scale* of the actual values is *not* linear and *not* intuitive. This needs to be considered when choosing a threshold for when a connection is considered "strong".

7 FUTURE WORK

Much work remains to be done to develop more interpretation and specific guidelines for the use of the measures discussed in this document. Many questions about alternative measures also arise: Are there other measures that have a more intuitive scale? Which other functions $U(X)$ (other than entropy) would be suitable as basis for these measures? Is there another averaging technique – other than True Average or Blind Average – that would yield interesting results?

Furthermore we believe that Link Strength measures may be useful in the context of constraint-based structure learning algorithms **to derive hypotheses of a system's primary causal pathways** from data. One problem when using constraint-based structure

Table 2: Connection and Link Strengths for varying b in Figure 5.

b	0.0	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
$MI(A,B) = LS^{true/blind}(A \to B)$	1.0	0.919	0.859	0.714	0.531	0.278	0.119	0.029	0
$MI\%(A,B) = LS\%^{true/blind}(A \to B)$	100	91.9	85.9	71.4	53.1	27.8	11.9	2.9	0.0

learning algorithms is the generally large number of possible DAGs returned by the algorithm. One way in which link strength may be helpful is that it could help one reduce the number of models to look at. Specifically, we plan to conduct some case studies where we look at the different models delivered by structure learning, and use link strength visualization to see how different they really are if one focuses only on the strong connections. It is likely that some of them only differ in minor connections and this will reduce the number of models in some cases enough so that one can identify only a few major causal hypotheses.

We believe that link strengths measures could also be used **to evaluate the quality of structure learning algorithms**. Currently structure learning algorithms are evaluated by counting the number of incorrect arrows when identifying known systems. We believe that it may be more appropriate to weigh those counts by the link strength of the incorrect arrows. Details of the weighing remain to be determined.

8 CONTRIBUTIONS

This paper reviewed link strength and connection strength measures for discrete Bayesian Networks. The primary contributions are a clean derivation of True Average Link Strength, newly proposed Blind Average Link Strength, newly proposed Percentage Measures for Mutual Information and Link Strength, derivation of several properties of the various measures, and proposed new uses for link strength measures in the context of causal discovery.

References

Boerlage, B., 1992, "Link Strengths in Bayesian Networks ," Master's thesis, Dept. of Computer Science, The University of British Columbia.

Cheng, J., Greiner, R., Kelly, J., Bell, D., and Liu, W., 2002, "Learning Bayesian Networks from Data: An Information-Theory Based Approach ," *Artificial Intelligence Journal*, 1-2(137):43–90.

Chickering, D. and Meek, C., 2003, "Monotone DAG Faithfulness: A Bad Assumption ," Technical Report MST-TR-2003-16, Microsoft Research.

Ebert-Uphoff, I., 2006, "User's Guide for the LinkConnectionStrength Package (Version 1.0, Jan 2006) ," Available at www.DataOnStage.com.

Jitnah, N., 1999, *Using Mutual Information for Approximate Evaluation of Bayesian Networks* PhD thesis, School of Computer Science and Software Engineering, Monash University, Clayton, Victoria, Australia.

Lacave, C. and Diez, F., 2002, "A review of explanation methods for Bayesian networks ," *The Knowledge Engineering Review*, 17(2).

Lacave, C. and Diez, F., 2004, "The Elvira GUI: a tool for generating explanations for Bayesian networks ," *submitted journal paper*.

Lauritzen, S. L. and Spiegelhalter, S., 1998, "Local computations with probabilities on graphical structures and their application to expert systems ," *J. Royal Statistics Society B*, 50(2):157–194.

Nicholson, A. and Jitnah, N., 1997, "Treenets: A framework for anytime evaluation of belief networks ," In *First International Joint Conference on Qualitative and Quantitative Pracical Reasoning (ECSQARU-FAPR'97)*. Springer.

Nicholson, A. and Jitnah, N., 1998, "Using Mutual Information to determine Relevance in Bayesian Networks ," In *5th Pacific Rim International Conference on Artificial Intelligence (PRICAI'98)*, pages 399–410, Singapore. Springer.

Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Interference* Morgan Kaufman Publishers, San Mateo, CA, revised second printing edition.

Shannon, C. and Warren, W., 1949, *The Mathematical Theory of Communication* University of Illinois Press, Urbana and Chicago.

Uffink, J., 1995, "Can the maximum entropy principle be explained as consistency requirement? ," *Studies in History and Philosophy of Modern Physics*, 26(B):223–261.

Zapata-Rivera, J., Neufeld, E., and Greer, J., 1999, "Visualization of Bayesian Belief Networks ," In *IEEE Visualization 1999 Late Breaking Hot Topics Proceeding*, pages 85 – 88.

A Weaknesses of Entropy

Pearl explained a weak point of using entropy – or any other measure that is a function of only the *probabilities* of a random variable's states – to measure uncertainty (Pearl 1988, pp. 322-323):

The main weakness of Shannon's measure is that it does not reflect the ordering or scale information relative to the values that a variable may take. For example, the uncertainty associated with the belief "The temperature is between 37^o *and* 39^o *would have the same entropy measure as the uncertainty associated with "The temperature is either between* 0^o *and* 1^o *or between* 99^o *and* 100^o*" (assuming uniform distribution over the intervals specified). Entropy is invariant to reordering or renaming the values in the domain, so it cannot reflect the fact that we perceive an error between* 37^o *and* 38^o *to be much less critical than an error between* 0^o *and* 100^o*. [...]*

The source of this peculiar behavior is that entropy, contrary to folklore, does not measure the harm caused by uncertainty; it measures the cost of removing the uncertainty (by querying an oracle and paying the same fee for all binary queries). This is why Shannon's mutual information measure endows equal penalty to all errors.

It should also be noted that Pearl's interpretation above of entropy *as the approximate number of required binary queries to determine the state of the variable* has an additional advantage. It seems to be the *simplest* one that fully explains the exact formula of entropy, including the logarithmic scale.

In contrast, most textbooks motivate entropy by stating that the term $\left(\log \frac{1}{P(x_i)}\right)$ measures the *surprise* if event x_i occurs. Then $U(X)$ is the *expected (average) surprise*, if infinitely many trials are performed. This interpretation is helpful, but does not directly explain the logarithmic scale, since one can think of many measures of *surprise* that are not logarithmic. Shannon (1949) stipulated some additional properties for his entropy measure that force the scale to be logarithmic. However, it is still a question of discussion whether those properties are indeed required for an uncertainty measure (Uffink (1995)). Nevertheless, no convincing alternative has yet emerged to measure uncertainty and entropy remains by far the most common choice.

B Proofs for Propositions 6.1 to 6.3

Proof of Proposition 6.1

Proof If node Y is a deterministic child of its parents then it follows $U(Y|X,\mathbf{Z}) = 0$ and $\hat{U}(Y|X,\mathbf{Z}) = 0$ in the definitions of True/Blind Average Link Strengths, which then yields the desired result.

Proof of Proposition 6.2

Proof From $LS^{blind}\%(P \rightarrow Y) = 100\%$ follows $\hat{U}(Y|X,\mathbf{Z}) = 0$, thus

$$\sum_{x,y,\mathbf{z}} P(y|x,\mathbf{z}) \log_2 P(y|x,\mathbf{z}) = 0.$$

Each term $P(y|x,\mathbf{z}) \log_2 P(y|x,\mathbf{z})$ is positive and vanishes if and only if $P(y|x,\mathbf{z}) = 0$ or $P(y|x,\mathbf{z}) = 1$. Thus in order for the whole sum to vanish, we must have $\forall x, y, \mathbf{z}: \ P(y|x,\mathbf{z}) \in \{0,1\}$. Thus Y is a deterministic child of its parents.

Proof of Proposition 6.3

Proof The following degenerate case serves as a counter example. Y has two parents, X, Z, which each can only take states 0 and 1. Let us say that $x = 0$ and $z = 0$ always, thus $P(x = 0, z = 0) = 1$ and $P(x, z) = 0$ otherwise. Define $Y = (x + z) * (\text{random number})$, then $U(Y|x = 0, z = 0) = 0$ and $U(Y|x, z) \neq 0$ otherwise. Thus all products $P(x, z)U(Y|x, z)$ vanish and $U(Y|X, Z) = 0$, although Y is clearly *not* a deterministic child of its parents.

Comment: The inability of the True Average Link Strength Percentage to guarantee that a node is a deterministic child comes from the fact that the definition of whether a child is deterministic is *independent of the joint probability of the node's parents*, while True Average Link Strength Percentage *disregards parent state combinations with zero joint probability.* Thus one may argue that this difference is philosophical in nature and that True Average Link Strength Percentage is also a good indicator for deterministic relationships. Nevertheless, it is more prudent to use Blind Average Link Strength Percentage for that purpose.