# Exploiting Experts' Knowledge for Structure Learning of Bayesian Networks

Hossein Amirkhani
http://ceit.aut.ac.ir/~amirkhani

December 2016

# WHAT IS THIS LECTURE ABOUT?

- Part I: Bayesian network fundamentals.

- Part II: A review of the following paper:
  - H. Amirkhani, M. Rahmati, P. Lucas, and A. Hommersom, "Exploiting experts' knowledge for structure learning of Bayesian networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

# Part I

# PROBABILISTIC GRAPHICAL MODELS

- Our goal is to represent a joint distribution $P$ over a set of random variables $X = \{X_1, \ldots, X_n\}$.

- Even in the simplest case where these variables are binary-valued, a joint distribution requires the specification of $2^n - 1$ numbers.

- The **explicit** representation of the joint distribution is **unmanageable** from every perspective:
  - Computationally, Cognitively, and Statistically.

- PGMs exploit **conditional independence** properties of the distribution in order to allow a compact and natural representation:
  - Like what Naïve Bayes does

# PROBABILISTIC GRAPHICAL MODELS

- Two main aspect:
  - Probabilistic: to model our uncertainty.
  - Graphical: a transparent model:

- Nodes are the random variables in our domain.
- Edges correspond, intuitively, to direct influence of one node on another.
- Numeric parameters are probabilities.

- Human experts can have a bidirectional interaction with these models.

# BAYESIAN NETWORKS

- They are a specific type of **probabilistic graphical models**.
  - BNs use directed acyclic graphs (**DAG**).

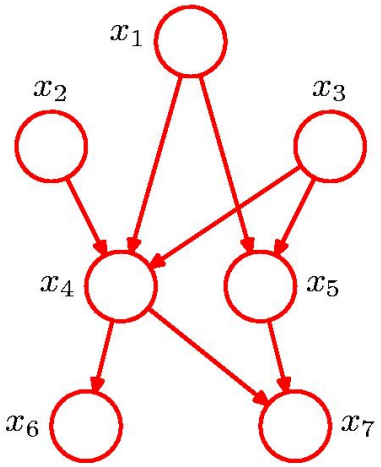
- Two main perspectives:
  - Causal
  - Probabilistic

# BAYESIAN NETWORKS

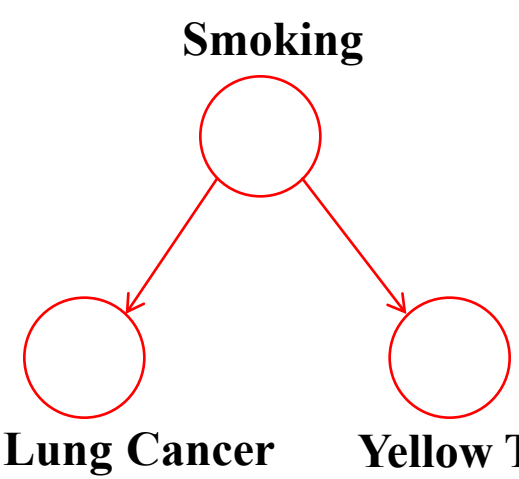
$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)$$

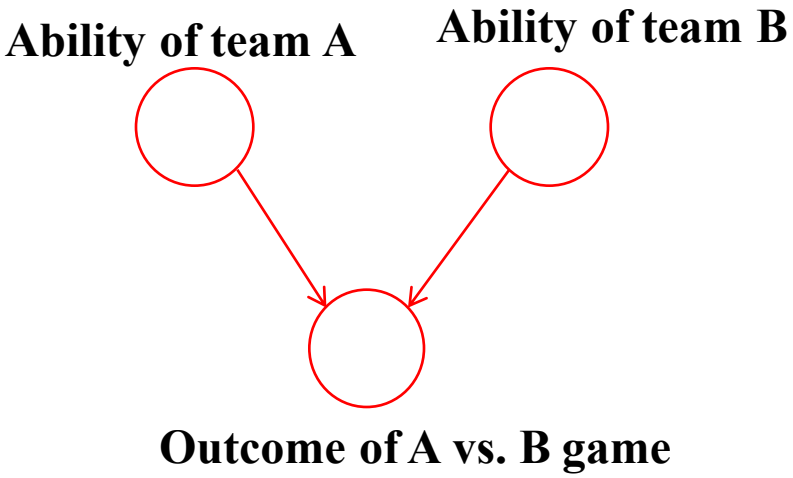
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i))$$

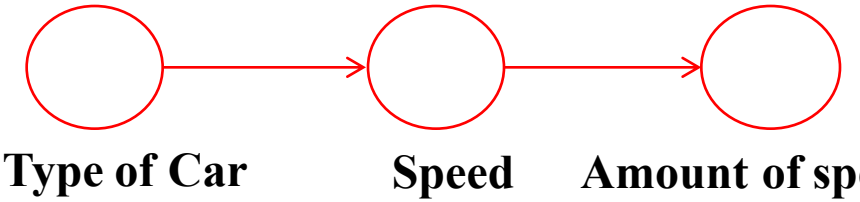
$$\begin{aligned} p(x_1, \ldots, x_7) = {} & p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ & p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5) \end{aligned}$$



# BUILDING BLOCKS



# D-SEPARATION

- One path from A to B is **blocked** by C if it contains a node such that either
  a) the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or
  b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are in the set C.
- If all paths from A to B are blocked, A is said to be d-separated from B by C.
- If A is d-separated from B by C, the joint distribution over all variables in the graph satisfies $A \perp\!\!\!\perp B \mid C$ .

# I-EQUIVALENCE

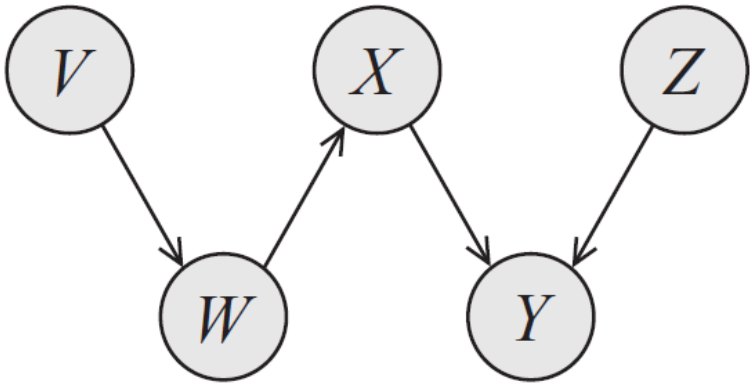
- Let $G$ be a DAG over $X$. We define $I(G)$ to be the set of independence assertions that implied by d-separations.

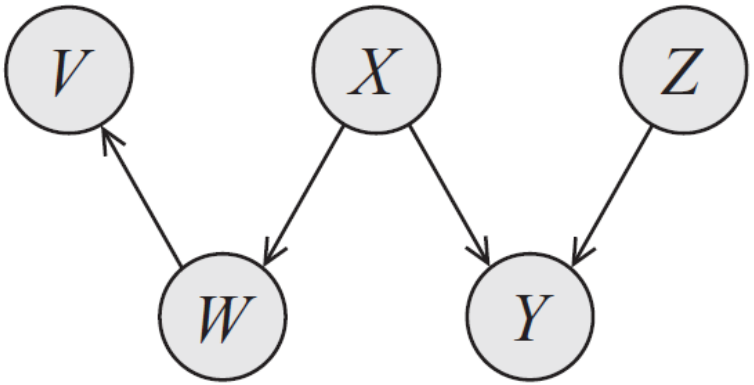
- Two graph structures $G_1$ and $G_2$ over $X$ are **I-equivalent** if $I(G_1) = I(G_2)$.

- The set of all graphs over X is partitioned into a set of mutually exclusive and exhaustive I-equivalence classes.

# THE SKELETON OF A BAYESIAN NETWORK

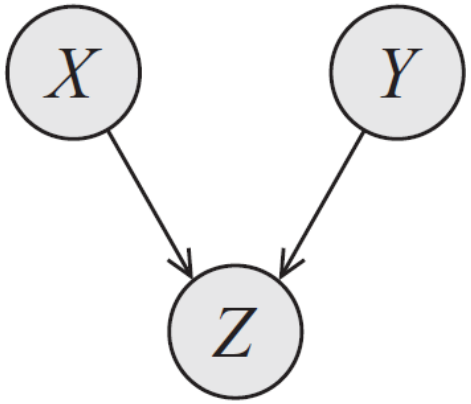
- The **skeleton** of a Bayesian network graph $G$ over $X$ is an undirected graph over $X$ that contains an edge $\{X, Y\}$ for every edge $(X, Y)$ in $G$.



# IMMORALITY

- A **v-structure** $X \rightarrow Z \leftarrow Y$ is an **immorality** if there is no direct edge between X and Y.



# A USEFUL AND INTERESTING THEOREM

- $G_1$ and $G_2$ have the same **skeleton** and the same set of **immoralities** **if and only if** they are **I-equivalent.**

- We can use this theorem to recognize that whether two BNs are I-equivalent or not.

- In addition, this theorem can be used for **learning** the structure of the Bayesian network.

# BAYESIAN NETWORK LEARNING

- Parameter learning:
  - Learning the conditional probability distributions (CPD)

- Structure learning:
  - Score-based methods
  - Constraint-based methods

# Part II

# WHY EXPERTS' KNOWLEDGE?

- Super-exponential structure space:
  - Between $2^{\binom{n}{2}}$ and $3^{\binom{n}{2}}$
    - More than $10^{102}$ for only 24 nodes!
- Lack of enough reliable data.
- I-equivalence structures cannot be distinguished based on data alone.
- Our approach deals with multiple experts with varying levels of expertise rather than an omniscient expert.

# PROBLEM SETTING

- $g_i = \rightarrow$ if $(X \rightarrow Y) \in E$,
- $g_i = \leftarrow$ if $(X \leftarrow Y) \in E$,
- $g_i = \nleftrightarrow$ if neither $(X \rightarrow Y)$ nor $(X \leftarrow Y)$ is in $E$.

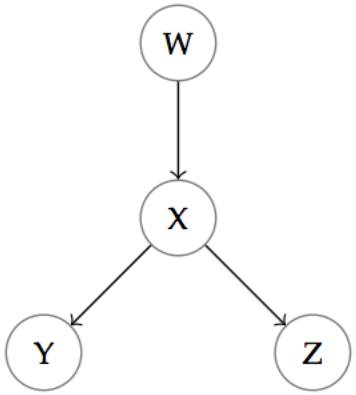



Fig. 1. Simple Bayesian network structure.

Fig. 1. Since there are 4 nodes in this graph, the number of node pairs is $N = 6$. If these pairs are ordered as (X,Y), (X,Z), (X,W), (Y,Z), (Y,W), (Z,W), we have $g_1 = \rightarrow$, $g_2 = \rightarrow$, $g_3 = \leftarrow$, $g_4 = \nleftrightarrow$, $g_5 = \nleftrightarrow$, $g_6 = \nleftrightarrow$.

# PROBLEM SETTING

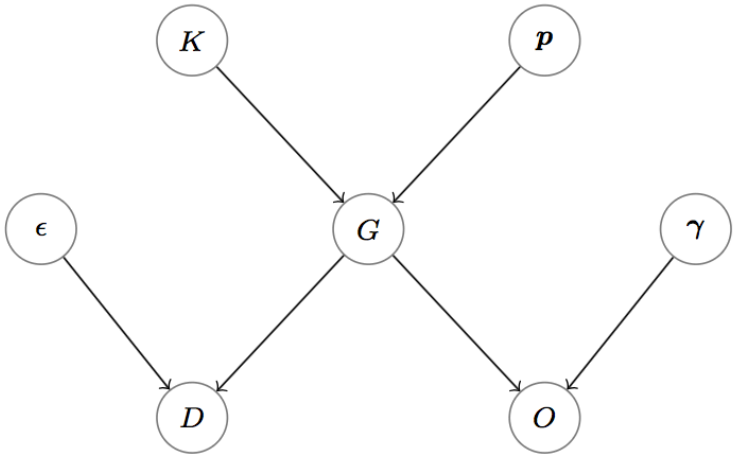
$$O^i_j \in \{\emptyset, \rightarrow, \leftarrow, \nleftrightarrow\}$$

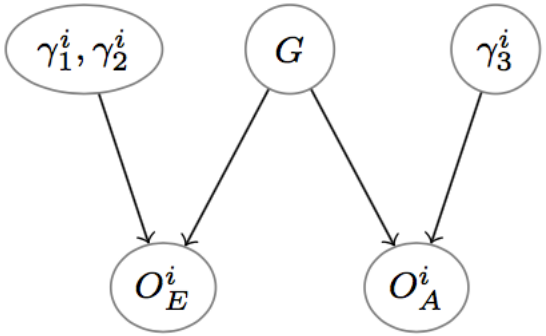
We denote the prior distribution over edge types as $\boldsymbol{p} = \{p_{\rightarrow}, p_{\leftarrow}, p_{\nleftrightarrow}\}$. For example, when $\boldsymbol{p} = \{p_{\rightarrow} = 0.1, p_{\leftarrow} = 0.2, p_{\nleftrightarrow} = 0.7\}$ it means that prior to having any data or experts' knowledge, we believe that 10%, 20%, and 70% of $g_i$s are respectively equal to $\rightarrow$, $\leftarrow$, $\nleftrightarrow$.

# PROBLEM SETTING

- $\gamma_1$: The probability of detecting the existing edges with *correct* directions,
- $\gamma_2$: The probability of detecting the existing edges with *reverse* directions,
- $\gamma_3$: The probability of correctly detecting the *absent* edges.

# PROBLEM MODELLING

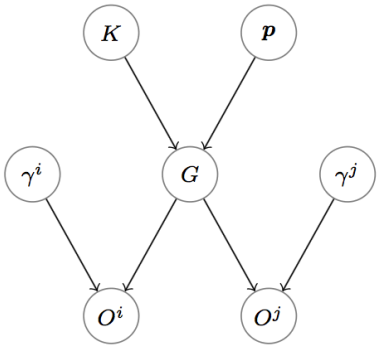


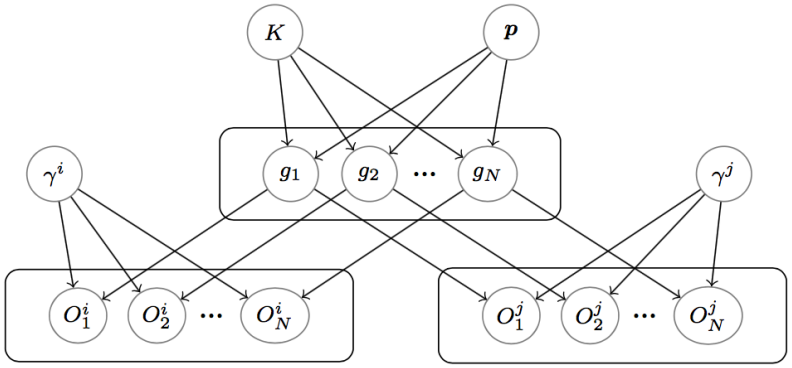

$$G \perp \boldsymbol{\gamma}$$
$$G \perp \boldsymbol{\gamma} \mid \boldsymbol{p}$$
$$D \perp \boldsymbol{\gamma} \mid G$$
$$O \perp D \mid G$$
$$O \perp D \mid G, \boldsymbol{\gamma}$$

$$(\gamma_1^i, \gamma_2^i) \perp \gamma_3^i$$

# MODEL OF EXPERTS' OPINIONS




(a) Abstract model




(b) Detailed model

$$O^i \perp O^j \mid G$$
$$O^i \perp O^j \mid G, \boldsymbol{\gamma}$$
$$O^j \perp \gamma^i \mid G, \gamma^j$$
$$O^j_x \perp O^j_y \mid G, \gamma^j$$
$$O^j_x \perp g_y \mid g_x, \gamma^j$$
$$O^i_x \perp O^j_x \mid g_x, \boldsymbol{p}, \boldsymbol{\gamma}$$
$$O^j_x \perp \gamma^i \mid g_x, \boldsymbol{p}, \gamma^j$$
$$O^j_x \perp \boldsymbol{p} \mid g_x, \gamma^j$$

# SCORES

- Explicit-accuracy-based score:
  - First step: experts' accuracies are estimated,
  - Second step: estimated accuracies are used to score the structures.

- Marginalization-based score:
  - When we are not confident about the estimated accuracies.

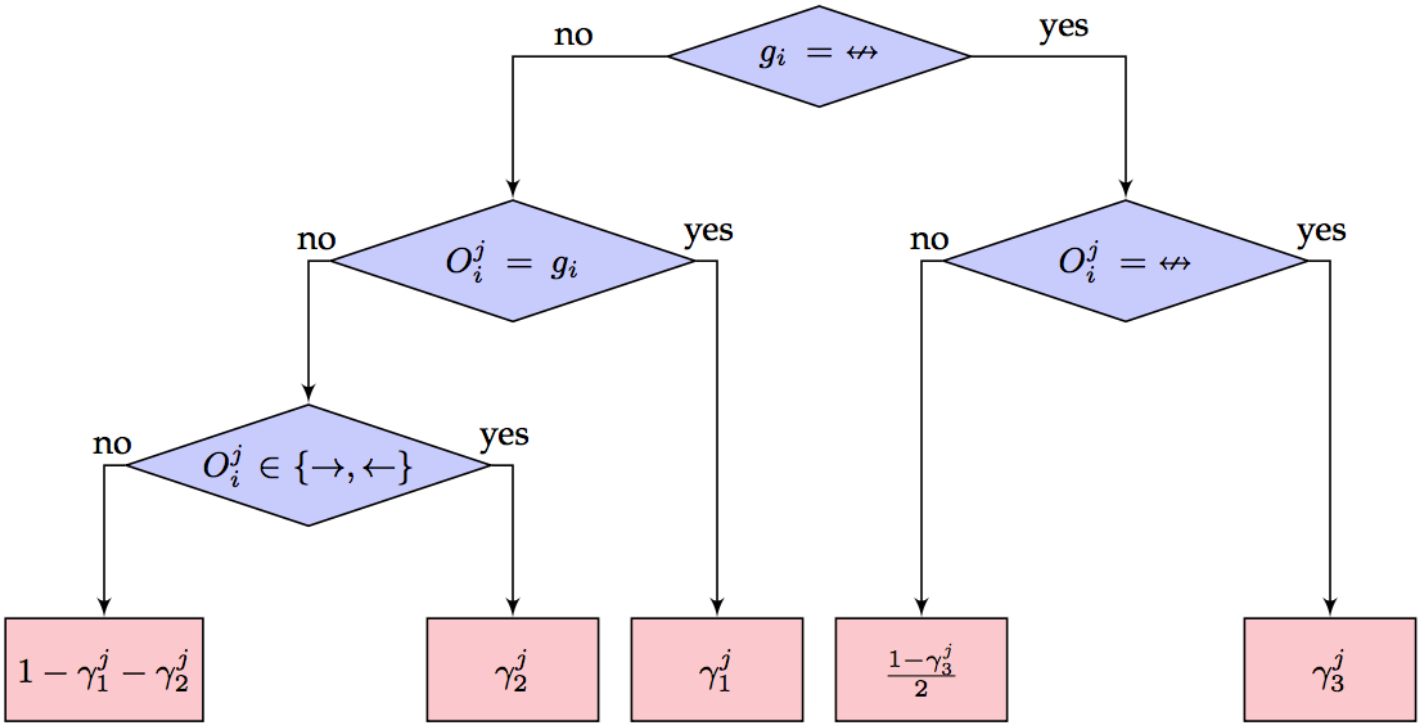
# EXPLICIT-ACCURACY-BASED SCORE

$$P(G \mid D, O, \gamma) \propto P(G, D, O, \gamma) = P(\gamma)\, P(G \mid \gamma)\, P(D \mid G, \gamma)\, P(O \mid G, D, \gamma)$$

$$\text{Score}_{explicit}(G; D, O, \gamma) = \log P(G) + \log P(D \mid G) + \log P(O \mid G, \gamma).$$

$$\log P(O \mid G, \gamma) = \sum_{j=1}^{R} \sum_{i=1}^{N} \log P(O_i^j \mid g_i, \gamma^j)$$

# EXPLICIT-ACCURACY-BASED SCORE



# MARGINALIZATION-BASED SCORE

$$P(G \mid D, O) \propto P(G, D, O) = P(G)\, P(D \mid G)\, P(O \mid G, D)$$

$$\text{Score}_{marg}(G; D, O) = \log P(G) + \log P(D \mid G) + \log P(O \mid G)$$

$$\log P(O \mid G) = \sum_{i=1}^{R} \log P(O^i \mid G).$$

$$P(O^i \mid G) = \int_0^1 \int_0^{1-\gamma_1^i} \int_0^1 \Big( P(\gamma_1^i, \gamma_2^i, \gamma_3^i \mid G) \times P(O^i \mid \gamma_1^i, \gamma_2^i, \gamma_3^i, G) \Big)\, d\gamma_3^i\, d\gamma_2^i\, d\gamma_1^i.$$

# MARGINALIZAITON BASED SCORE

$$
\begin{aligned}
\log P(O^i \mid G) = & \log B(\beta_{i1} + m_{i1}, \beta_{i2} + m_{i2}) \\
& - m_{i2} \log 2 - \log B(\beta_{i1}, \beta_{i2}) \\
& + \log B(\alpha_{i1} + n_{i1}, \alpha_{i2} + n_{i2} + \alpha_{i3} + n_{i3}) \\
& + \log B(\alpha_{i2} + n_{i2}, \alpha_{i3} + n_{i3}) \\
& - \log B(\alpha_{i1}, \alpha_{i2}, \alpha_{i3}).
\end{aligned}
$$

# EVALUATION MEASURE

- There may be three types of errors in the learned DAG (CPDAG):
  - Wrong Connection
  - Missed Edge
  - Wrong Orientation

- The total number of these errors are called the structural hamming distance (SHD).

# SIMULATION EXPERIMENTS

| Name | Description | Nodes | Edges |
|---|---|---|---|
| Asia | Diagnosing some respiratory diseases | 8 | 8 |
| Insurance | Evaluating car insurance risks | 27 | 52 |
| Alarm | Monitoring patients in intensive care | 37 | 46 |
| Hailfinder | Predicting summer hails in northern Colorado | 56 | 66 |

| | Weak | | | Mediocre | | | Good | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ |
| 1 | 0.15 | 0.80 | 0.85 | 0.15 | 0.80 | 0.85 | 0.15 | 0.80 | 0.85 |
| 2 | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 | 0.30 |
| 3 | 0.75 | 0.10 | 0.90 | 0.75 | 0.10 | 0.90 | 0.75 | 0.10 | 0.90 |
| 4 | 0.40 | 0.25 | 0.50 | 0.40 | 0.25 | 0.50 | 0.85 | 0.05 | 0.85 |
| 5 | 0.45 | 0.35 | 0.45 | 0.45 | 0.35 | 0.45 | 0.70 | 0.15 | 0.80 |
| 6 | 0.55 | 0.20 | 0.60 | 0.55 | 0.20 | 0.60 | 0.75 | 0.15 | 0.70 |
| 7 | 0.20 | 0.15 | 0.50 | 0.20 | 0.15 | 0.95 | 0.20 | 0.15 | 0.95 |
| 8 | 0.33 | 0.33 | 0.33 | 0.90 | 0.05 | 0.80 | 0.90 | 0.05 | 0.80 |
| 9 | 0.50 | 0.30 | 0.40 | 0.70 | 0.20 | 0.70 | 0.70 | 0.20 | 0.70 |
| 10 | 0.30 | 0.50 | 0.30 | 0.60 | 0.30 | 0.65 | 0.80 | 0.10 | 0.90 |
| Mean | 0.39 | 0.33 | 0.51 | 0.50 | 0.27 | 0.67 | 0.61 | 0.21 | 0.78 |

# SIMULATION EXPERIMENTS

The Obtained Structural Hamming Distances for the Alarm Network

(a) Weak population

| $\|D\|$ | $\nu$ | DAG Data | DAG Expert | DAG PE | DAG Mean | DAG EM | DAG Marg | CPDAG Data | CPDAG Expert | CPDAG PE | CPDAG Mean | CPDAG EM | CPDAG Marg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1000 | 0.3 | 30.2 | 212.7 | 108.6 | 31.2 | 44.7 | **28.9** | 31.9 | 213.7 | 110.5 | 34.8 | 47.5 | **31.3** |
| | 0.4 | 30.2 | 186.6 | 98.7 | 34.3 | 34.0 | **28.2** | **31.9** | 188.2 | 100.3 | 37.5 | 38.9 | 32.1 |
| | 0.5 | 30.2 | 173.2 | 93.1 | 31.2 | 33.0 | **28.1** | 31.9 | 174.3 | 95.1 | 32.2 | 36.2 | **31.1** |
| | 0.6 | 30.2 | 166.5 | 94.0 | 29.4 | 28.9 | **26.0** | 31.9 | 167.8 | 96.0 | **31.1** | 33.5 | 31.2 |
| | Avg | 30.2 | 184.8 | 98.6 | 31.5 | 35.1 | **27.8** | 31.9 | 186.0 | 100.5 | 33.9 | 39.0 | **31.4** |
| 5000 | 0.3 | 29.5 | 212.7 | 83.8 | 29.9 | 38.4 | **28.4** | **30.5** | 213.7 | 87.0 | 33.2 | 41.3 | **30.5** |
| | 0.4 | 29.5 | 186.6 | 79.9 | 31.1 | 31.5 | **27.8** | **30.5** | 188.2 | 83.0 | 34.0 | 34.9 | **30.5** |
| | 0.5 | 29.5 | 173.2 | 77.0 | 30.2 | 32.6 | **24.8** | 30.5 | 174.3 | 78.9 | 30.9 | 35.0 | **28.3** |
| | 0.6 | 29.5 | 166.5 | 76.8 | 27.0 | 26.4 | **24.8** | 30.5 | 167.8 | 78.8 | 29.6 | 29.0 | **28.0** |
| | Avg | 29.5 | 184.8 | 79.4 | 29.6 | 32.2 | **26.5** | 30.5 | 186.0 | 81.9 | 31.9 | 35.0 | **29.3** |

(c) Good population

| $\|D\|$ | $\nu$ | DAG Data | DAG Expert | DAG PE | DAG Mean | DAG EM | DAG Marg | CPDAG Data | CPDAG Expert | CPDAG PE | CPDAG Mean | CPDAG EM | CPDAG Marg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1000 | 0.3 | 30.2 | 87.9 | 38.3 | 21.7 | **17.6** | **17.6** | 31.9 | 90.6 | 42.4 | 25.8 | **21.3** | 21.5 |
| | 0.4 | 30.2 | 62.4 | 26.5 | 20.8 | **16.1** | 18.0 | 31.9 | 65.4 | 28.7 | 25.1 | **19.6** | 21.1 |
| | 0.5 | 30.2 | 37.7 | 13.3 | 16.3 | **8.2** | 10.6 | 31.9 | 40.8 | 15.0 | 19.6 | **10.0** | 12.1 |
| | 0.6 | 30.2 | 36.1 | 15.1 | 15.7 | **8.8** | 9.0 | 31.9 | 39.8 | 17.3 | 18.8 | **9.6** | 11.2 |
| | Avg | 30.2 | 56.0 | 23.3 | 18.6 | **12.7** | 13.8 | 31.9 | 59.2 | 25.8 | 22.3 | **15.1** | 16.5 |
| 5000 | 0.3 | 29.5 | 87.9 | 33.4 | 20.2 | 15.5 | **14.6** | 30.5 | 90.6 | 37.5 | 24.0 | **18.3** | **18.3** |
| | 0.4 | 29.5 | 62.4 | 26.2 | 19.8 | **16.8** | 17.2 | 30.5 | 65.4 | 29.2 | 23.3 | **20.7** | 21.1 |
| | 0.5 | 29.5 | 37.7 | 13.3 | 11.5 | **8.1** | 9.1 | 30.5 | 40.8 | 16.6 | 14.2 | **9.7** | 10.8 |
| | 0.6 | 29.5 | 36.1 | 15.6 | 15.5 | **7.4** | 10.4 | 30.5 | 39.8 | 18.6 | 18.4 | **8.9** | 12.7 |
| | Avg | 29.5 | 56.0 | 22.1 | 16.8 | **11.9** | 12.8 | 30.5 | 59.2 | 25.5 | 20.0 | **14.4** | 15.7 |

# REAL WORLD EXPERIMENT

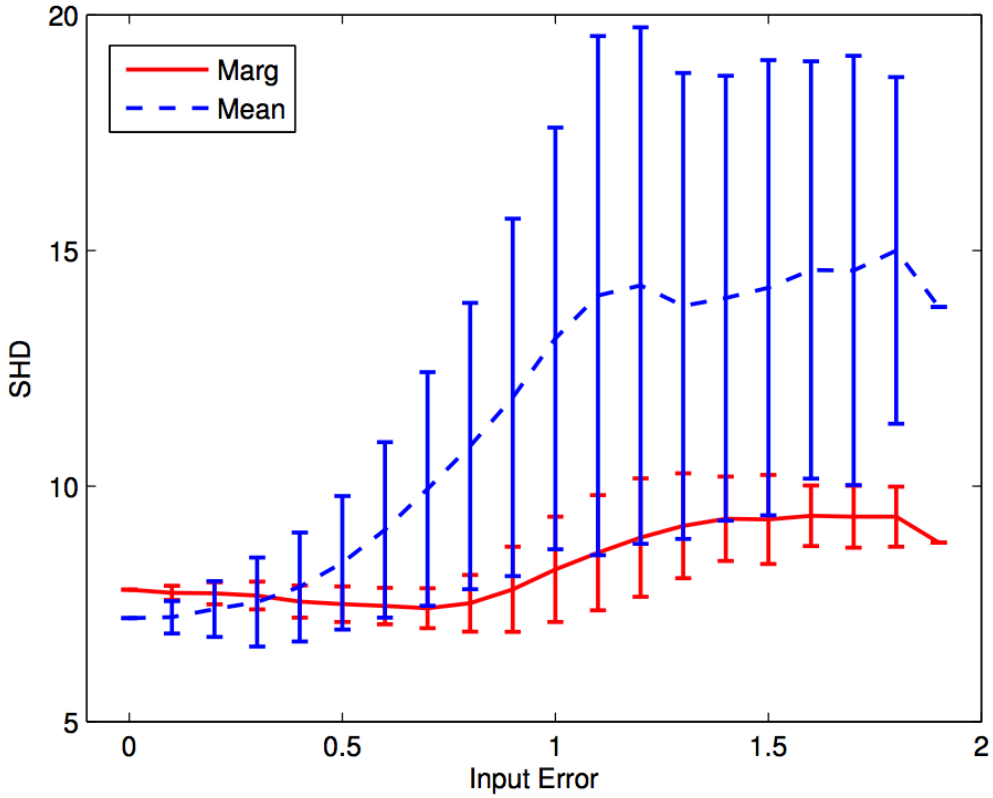
Table that the Radiologists had to Fill in as Part of the Experiment. Entries in the Upper Triangular Part of the Table had to be Filled in by $\rightarrow$, $\leftarrow$, $\leftrightarrow$, or Remain Empty if the Radiologists had no Idea what to Fill in

| | Microcalcifications | Spiculation | Location | Age | Lymph Nodes | Skin Retraction | Shape | Size | Breast cancer | Fibrous Tissue Develop | Breast Density | Margin | Nipple Discharge | Architectural Distort | Metastasis | Mass |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Microcalcifications | | | | | | | | | | | | | | | | |
| Spiculation | | | | | | | | | | | | | | | | |
| Location | | | | | | | | | | | | | | | | |
| Age | | | | | | | | | | | | | | | | |
| Lymph Nodes | | | | | | | | | | | | | | | | |
| Skin Retraction | | | | | | | | | | | | | | | | |
| Shape | | | | | | | | | | | | | | | | |
| Size | | | | | | | | | | | | | | | | |
| Breast cancer | | | | | | | | | | | | | | | | |
| Fibrous Tissue Develop | | | | | | | | | | | | | | | | |
| Breast Density | | | | | | | | | | | | | | | | |
| Margin | | | | | | | | | | | | | | | | |
| Nipple Discharge | | | | | | | | | | | | | | | | |
| Architectural Distort | | | | | | | | | | | | | | | | |
| Metastasis | | | | | | | | | | | | | | | | |
| Mass | | | | | | | | | | | | | | | | |

# REAL WORLD EXPERIMENT

| $\|D\|$ | Data | Expert | PE | Mean | EM | Marg |
|---|---|---|---|---|---|---|
| 100 | 13.2 | 32.0 | 26.4 | **9.7** | 25.3 | 10.2 |
| 400 | 8.5 | 32.0 | 23.4 | **7.2** | 19.6 | **7.2** |
| 700 | 6.8 | 32.0 | 22.0 | **5.8** | 18.2 | **5.8** |
| 1000 | 6.8 | 32.0 | 21.0 | **5.8** | 17.9 | **5.8** |
| Avg | 8.8 | 32.0 | 23.2 | **7.1** | 20.3 | 7.3 |

# REAL WORLD EXPERIMENT



# WHAT'S AHEAD?

- Fuzzy knowledge?
- Correlation knowledge or indirect causal relationships.
- Exploiting knowledge in other learning approaches.

# Thanks