Causal Inference and Its Application in Security

Huang Xiao

Lst. IT Sicherheit (I20)
Technische Universität München

Presented on 29th August 2012

Outline

- What is causality?
  - *Motivation, Example, Intuition.*
- Causal Bayesian Network
  - *Theoretical background, problem statement*
- Approach: Copula based Causal BN
  - *Copula functions*
  - *PICM Structure learning*
- Empirical study on IDS dataset
  - *KDD99 Dataset, experimental results*
- Future work and wrap up.

Approx. 20 min

What is (probabilistic) causality?

From Wikipedia:

Causality (also referred to as **causation**) is the relationship between an event (the *cause*) and a second event (the effect), where the second event is understood as a consequence of the first.

A example question in real life:

Does smoking causes lung cancer?

YES, IT MIGHT DO!

In a probabilistic view

Does smoking causes lung cancer?

Smoking will increase the probability of getting lung cancer.

Why do we need causality?

- Discover the rules of the nature.
- Reasoning
- Decision-making

Fundamental difference with machine learning

Association

- Now we want to find out what causes lung cancer.

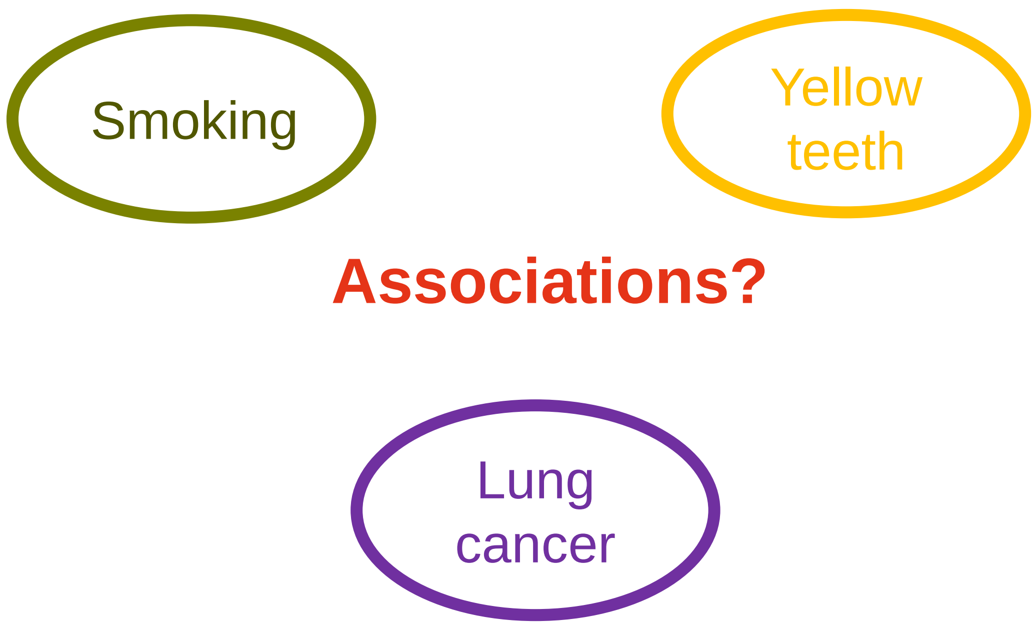
I. Data observations

		Lung cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	100	400
Yes	No	100	400
No	Yes	1	450
No	No	9	8540

Data from 10000 people*

*fictional

Three variables



Measuring Association

Information theory

- *Mutual Information*

Statistics

- *Pearson(linear) correlation*
- *Spearman correlation (continuous variables)*
- *Effect size (between two variables)*
- *Many others..*

From the data…

Obviously…

Yellow teeth and lung cancer are associated.

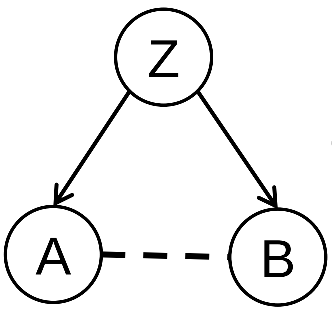
But…

Bleaching the teeth does not help reduce the probability of getting lung cancer.

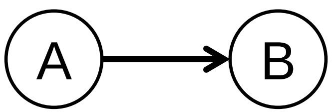
Correlation does not imply Causation!

Common Cause Principle

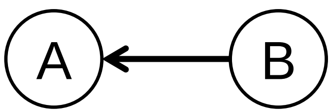
- If **A** and **B** are correlated, then **A** causes **B** or **B** causes **A** or they share a latent common cause.



or



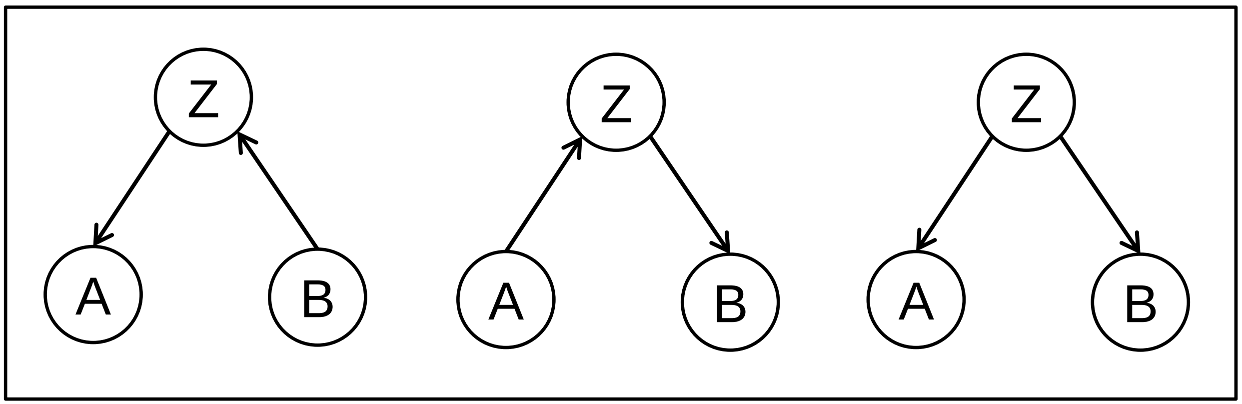
or



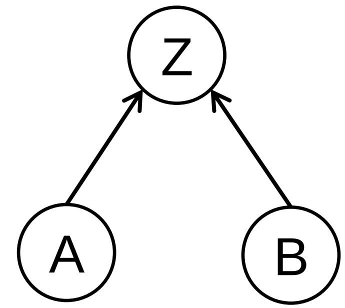
It links causation with probability

HANS REICHENBACH (1891 - 1953)

Conditional Independence



Equivalent class



V-structure

Associations:
- Dep(A, Z | Ø)
- Dep(Z, B | Ø)
- Dep(A, B | Ø)
- Ind(A, B | Z)

Associations:
- Dep(A, Z | Ø)
- Dep(Z, B | Ø)
- Ind(A, B | Ø)
- Dep(A, B | Z)

Possibility of Causal Inference?

Given $Pr(X_1, \dots, X_n)$, can we infer the causal graph $\mathcal{G}$?

Answer:

- Impossible without additional information.

  e.g., expertise knowledge, variable ordering
- Only equivalence class can be recovered!

Bayesian Network

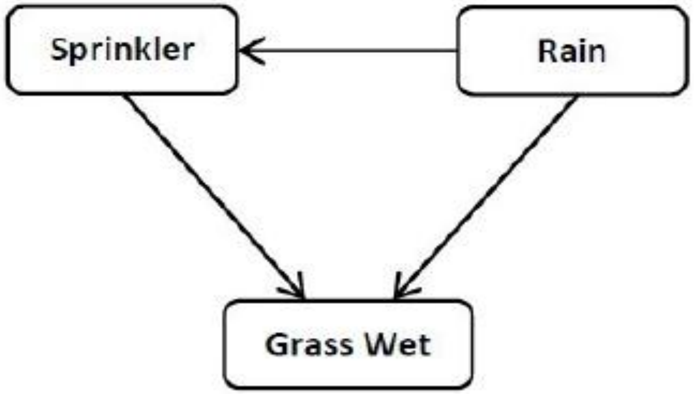
Definition:

Given a set of variables $\{X_1, \dots, X_n\}$, a Bayesian network is a probabilistic graphical model $B = (G, \Theta)$, where G is a directed acyclic graph (DAG) and Θ is the set of the parameters in all conditional probability distributions (CPDs).

Applications: Security engineering, vulnerability detection, intrusion detection, problem diagnosis (trouble shooting)

Example

	Sprinkler	
Rain	T	F
T	0.4	0.6
F	0.05	0.95



Rain	
T	F
0.2	0.8

		Grass Wet	
Sprinkler	Rain	T	F
F	F	0.0	1.0
F	T	0.75	0.25
T	F	0.85	0.15
T	T	0.99	0.01

Assumptions

- **Causal Markov Condition**
  - Every variable is independent of its non-descendants given its parents.
  - Factorization: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_i)$
- **Faithfulness**
  - Causal structure fully determines independences.
- **Acyclic**
  - Needs to be defined in problem setting.
- **Causal sufficiency**
  - Assume no latent common cause.
  - For efficient learning, also for causal interpretation of output.

Learning Bayesian Network

Task:

Given a dataset $\mathcal{D}$, try to learn the structure G and the parameters of all conditional probability distribution Θ.

Traditional method:

1-step: Structure learning

2-step: parameter estimation

Structure learning

I. Constraint based

Conditional independence tests in data and find a DAG faithful to them.

Methods:

- *SGS*
- *PC*
- *TPDA*
- *CPC*

II. Score based

Find a DAG maximizing the posteriori probability given the data.

Methods:

- *K2*
- *Sparse Candidate*
- *GBPS*
- *And many more..*

III. Hybrid

Methods: MMHC, CB, ECOS

Parameter Estimation

Given the structure G learned from last step, factorization will apply according to local terms governed by parameters θ_i

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_i, \theta_i)$$

Any estimator will work here:

e.g., MLE, MAP, and so on.

But…

Only equivalence class can be obtained!

Problems in BN Learning

- Search space is exponentially large in high dimension
- Too many conditional tests
- Local minimum
- Parametric form needed
- Missing values

Copula Treatment – Sklar's theorem

[Sklar 1959] Let $F\left(x_1, \cdots, x_N\right)$ *be any multivariate distribution over real-valued random variables, then there exists a copula function such that*

$$F\left(x_1, \cdots, x_N\right) = C\left(F\left(x_1\right), \cdots, F\left(x_N\right)\right)$$

where $F(X_i)$ *is marginal cumulative density distribution of variable* X_i *and furthermore if each* $F\left(X_i\right)$ *is continuous then* C *is unique.*

A quick sample: Gaussian Copula

Gaussian Copula is a widely explored Copula function:

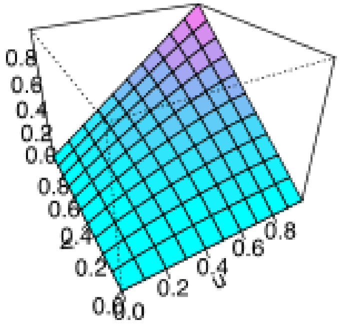
$$C(\{F(x_i)\}) = \Phi_\Sigma\left(\Phi^{-1}(F(x_1)), \cdots, \Phi^{-1}(F(x_N))\right)$$

Φ :standard normal distribution

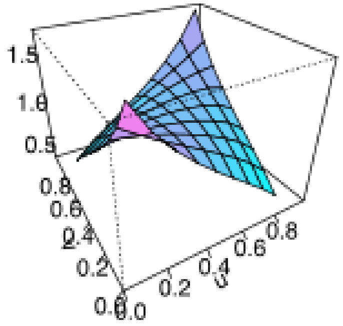
Φ_Σ : zero mean normal distribution

Σ : correlation matrix.

Gaussian copula cumulative



Gaussian copula density



Bivariate cumulative and density distribution of Gaussian Copula with correlation ρ = 0.4

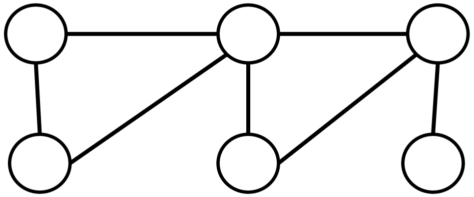
Advantages of Copula Functions

- Totally free choice of marginal distributions of each variable.
- Transform any joint distributions into a Gaussian.
- Non-parametric estimators are allowed, which is an ease for missing values. e.g., kernel density estimator.

Partial Inverse Correlation Matrix

- Instead of many CI-tests, simply inverse the correlation matrix.
- Extremely fast and stable under Gaussian Copula transformation.

$$\Sigma^{-1} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \Rightarrow$$

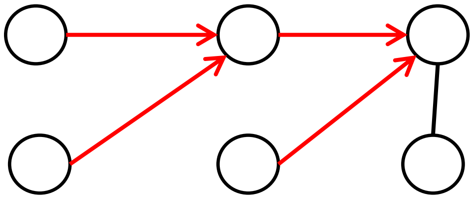


Note that:

$\Sigma^{-1}(i,j) = 0$ *indicates conditional independence, which implies no direct edge between node i and j.*

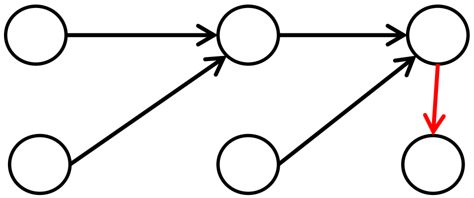
From Skeleton to PDAG

- Find V-structures



Detrianglation

- Constraint propagation



No new V-structure!

Finally, we recovered a causal graph model together with its quantitative factors (probabilistic parameters).

An application on Intrusion Detection System

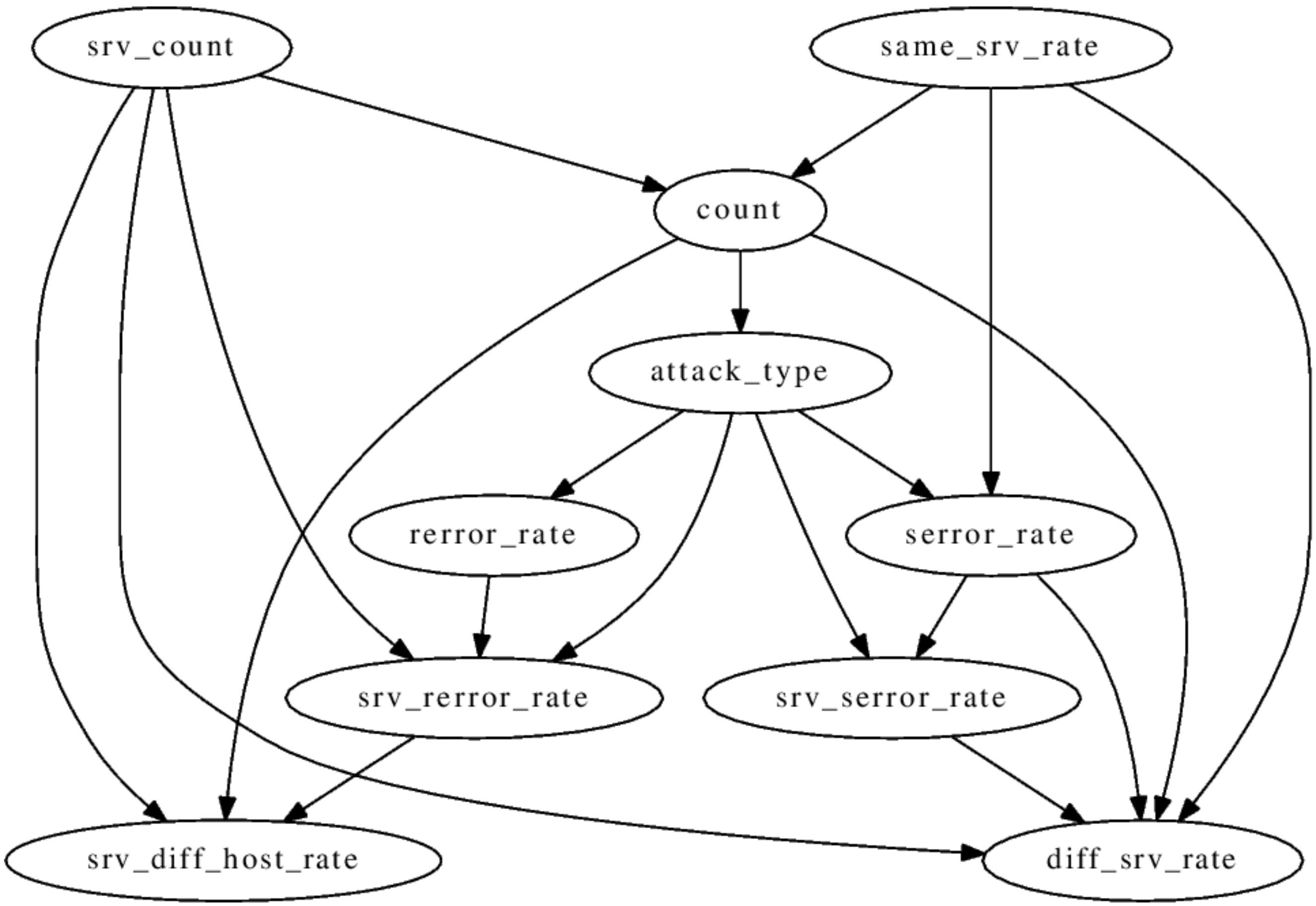
- Dataset: KDD99
- 42 variables in total, e.g.,
  - # of connections to the same host in the past two seconds
  - % of connections that have “REJ” errors
  - # of failed logins
  - Protocol type
- 21 attack types (but 60% are DOS attack)
- Training size: 1000

10 Features

feature name	*description*	*type*
count	number of connections to the same host as the current connection in the past two seconds	continuous
	Note: The following features refer to these same-host connections.	
serror_rate	% of connections that have ``SYN'' errors	continuous
rerror_rate	% of connections that have ``REJ'' errors	continuous
same_srv_rate	% of connections to the same service	continuous
diff_srv_rate	% of connections to different services	continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
	Note: The following features refer to these same-service connections.	
srv_serror_rate	% of connections that have ``SYN'' errors	continuous
srv_rerror_rate	% of connections that have ``REJ'' errors	continuous
srv_diff_host_rate	% of connections to different hosts	continuous

Inferred Causal Graph

10 Nodes only



Other datasets

- DARPA (1998)
  - From MIT Lincoln Labs, simulated in military network environment

Both KDD99 and DARPA are too old..

- ISCX (2012)
  - Gathered data in one week
  - From University of New Brunswick
  - Total 85.33 GB
  - Already got it!

Future work

- Now Copula Functions only work well for the continuous case.
- Most security scenarios are hybrid (both discrete and continuous data, which is still an open problem)
- Real-time causal network updating (DBNs)
- Dynamic feature selection
- Nonlinearity
  - E.g., stochastic process, kernel tricks
- Cyclic Bayesian Network (feedback loop)

Thanks