Causal Inference Using Composition of Transactions

Abbas Shojaee MD CHDA

2020

Please Do Not Circulate

Outline

- Introduction
- The fundamental question
- Inferring human disease causality using Causal Inference Using Composition of Transactions (CICT)
- Identification of Regulatory Network using CICT
- Research Plan

Fundamental Question

- “The fundamental question behind most research in biology or medical research is a causal question”
- A causal question is difficult, so we often compromise
  - Coexistence, correlation,
  - Differential question
  - Prediction

An even more difficult causal question is when

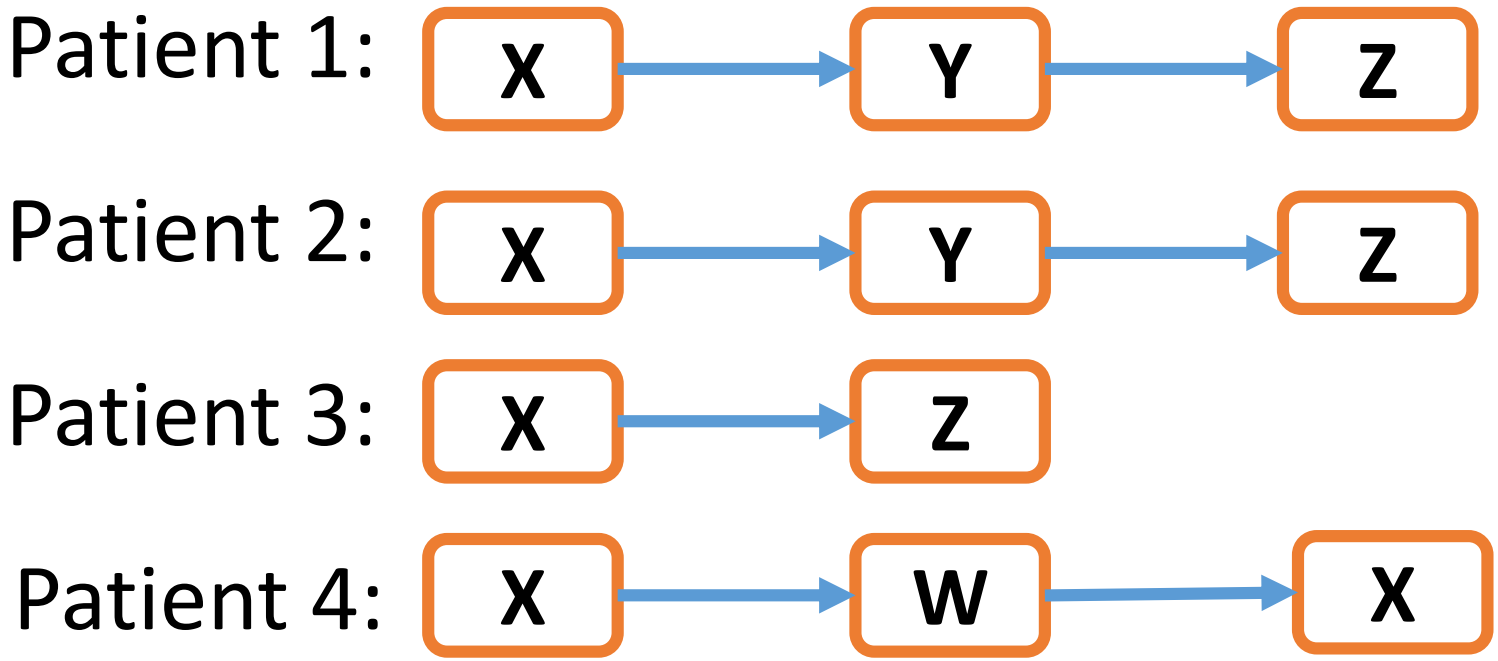
- Many changing parameters
- Many interactions
- Insufficient information on interactions
- The number of observations are low
- Multi layer heterogenous networks

  - Sounds familiar?

Graph Network Inference or System Inference

A Story About Inferring Human Diseases Graph

Each patient has a series of admissions, each admission with a diagnosis

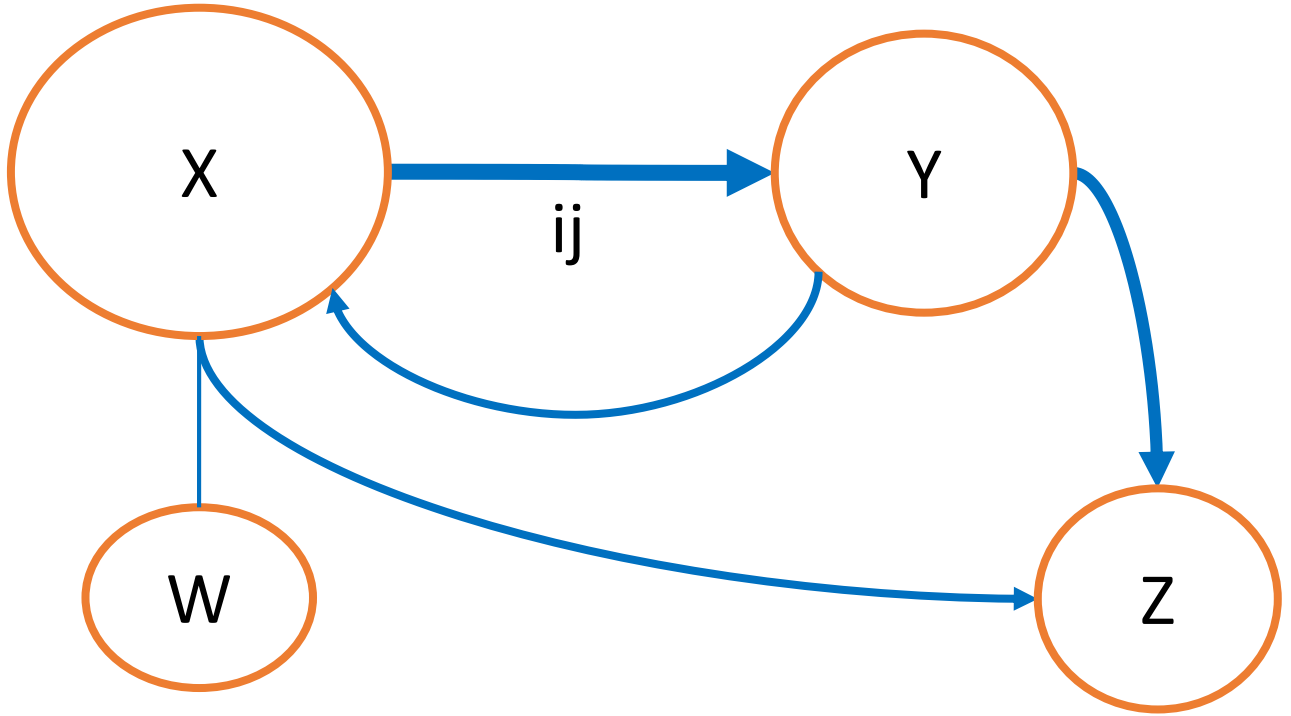


A Story About Inferring Human Diseases Graph

Source is represented by **i**

Target is represented by **j**

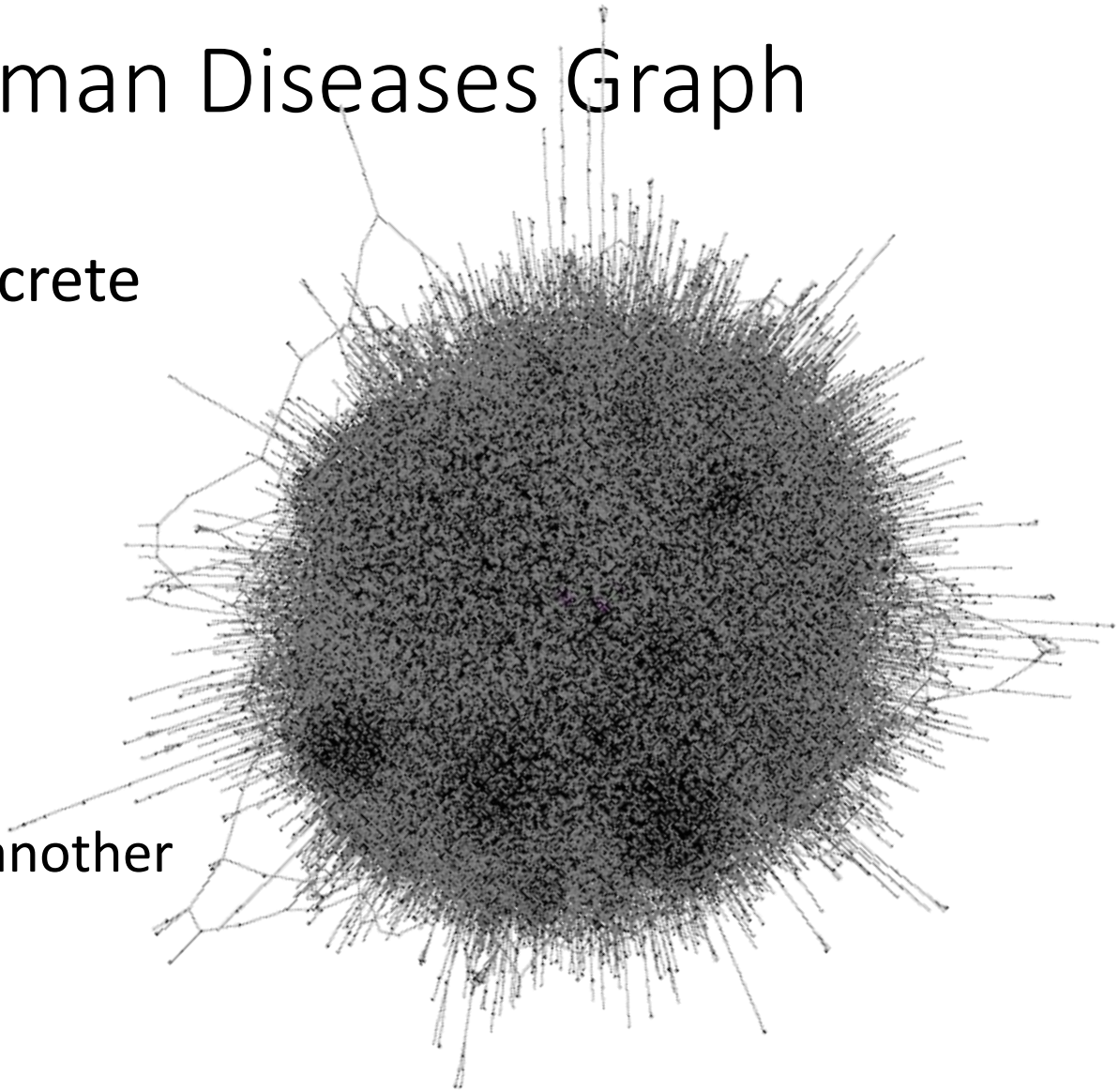
ij is a directed edge from i to j



A Story About Inferring Human Diseases Graph

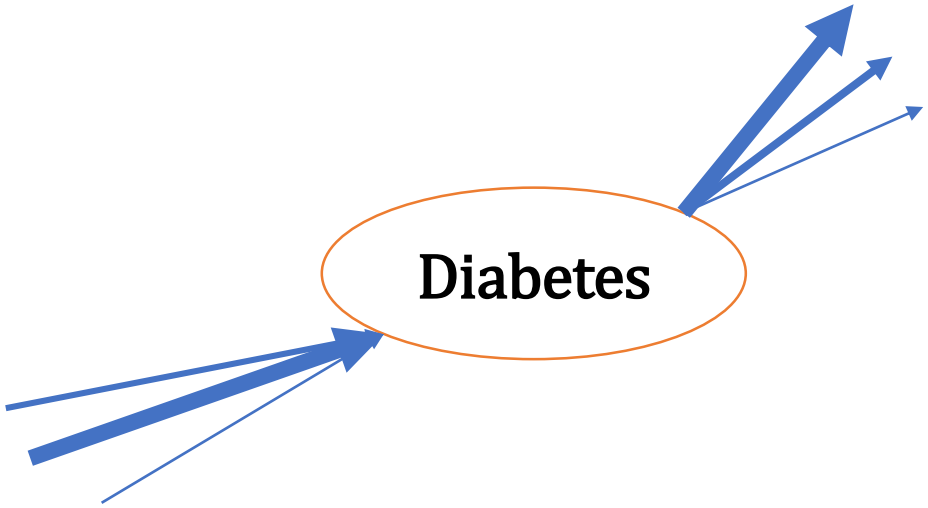
- Lots of very simple, short-length discrete time series data
  - Patients diagnoses and procedures in consecutive hospitalizations
  - 37000 diagnoses and procedures
  - 1.4 billion potential edges
  - 100 million real edges
- Question:
  - which diagnosis or procedure causes another one?

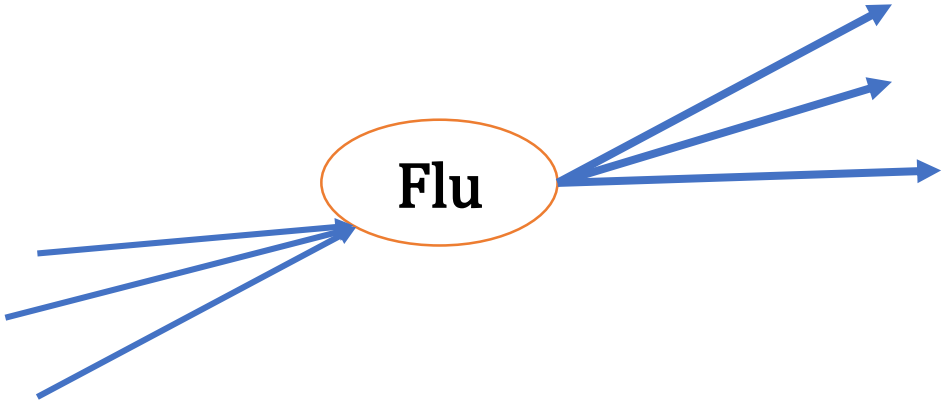
    OR
  - What is the underlying network?



The key idea:

- The set of events before and after a causal phenomenon are different than the set of events before or after a random phenomenon.



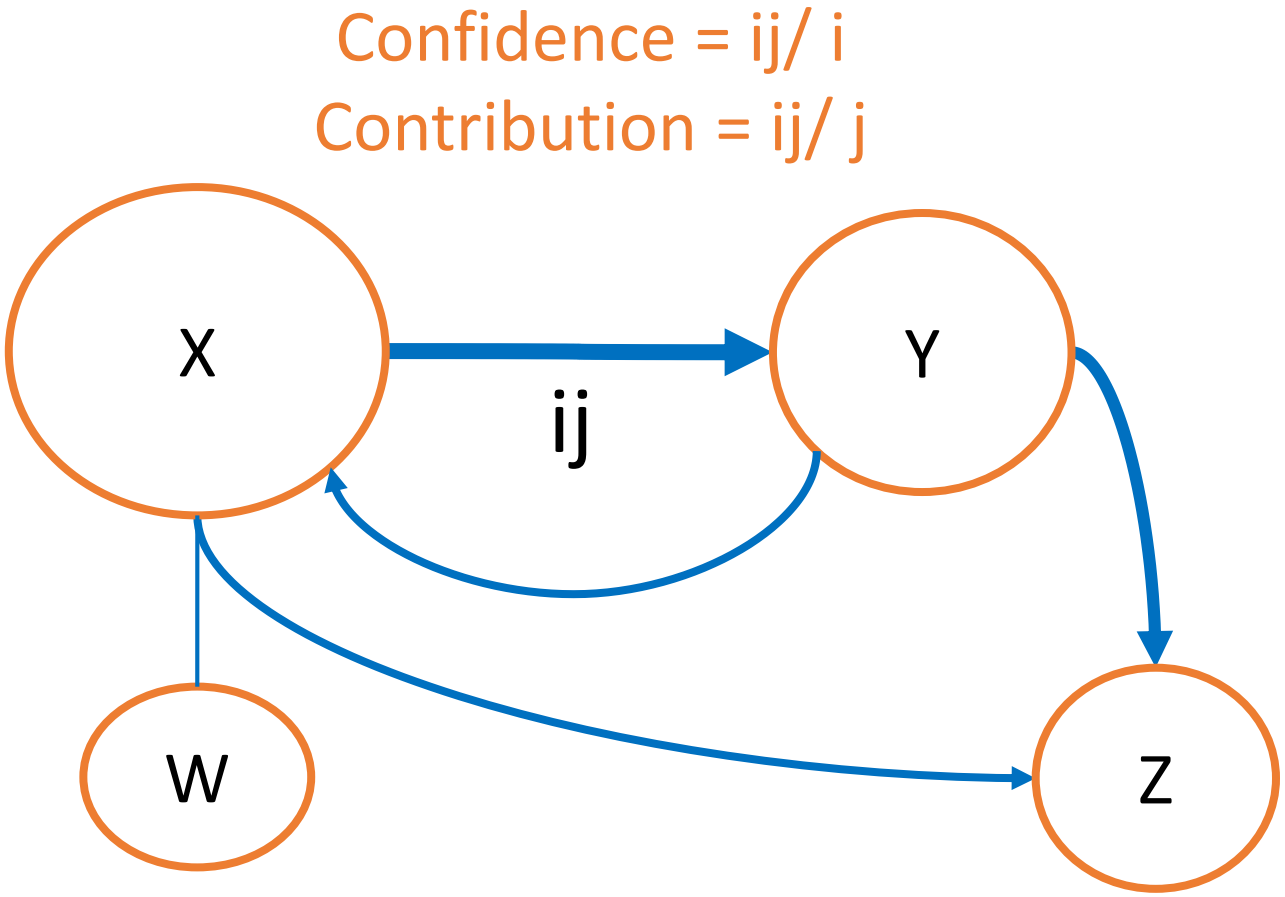


Adding a simple definition

Source is represented by i

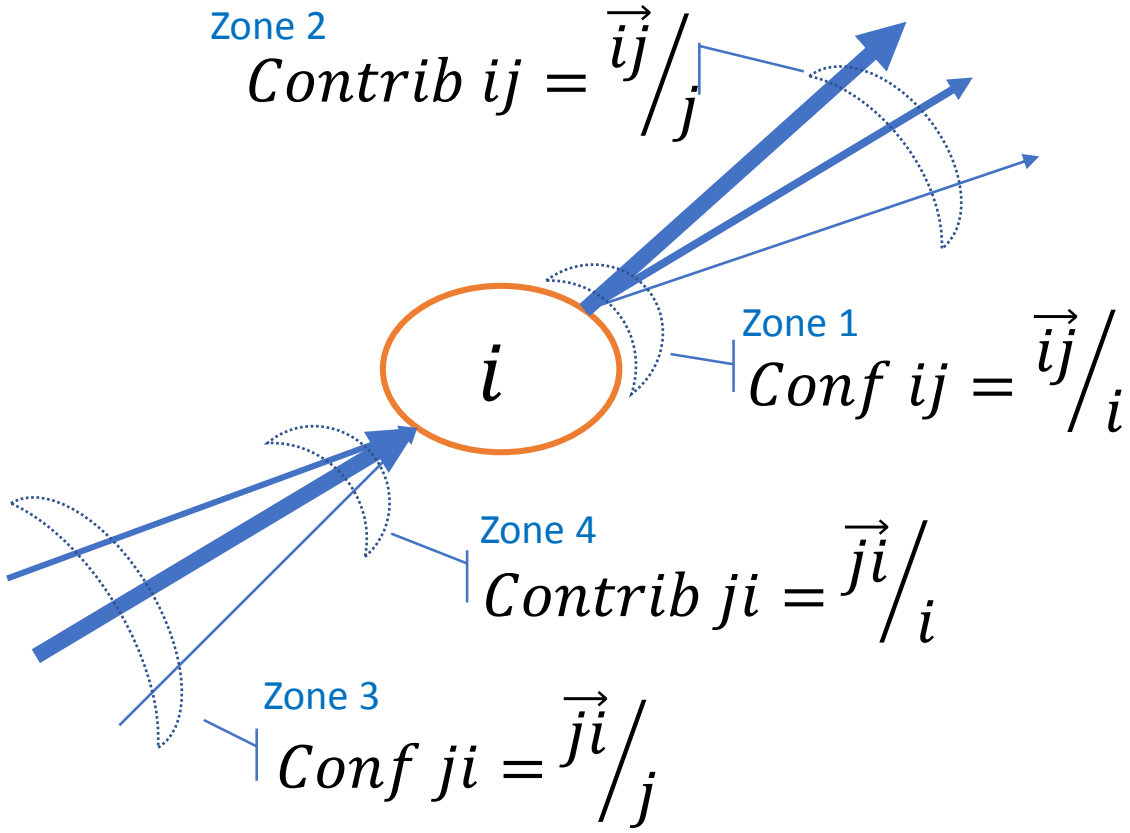
Target is represented by j

ij is a directed edge from i to j

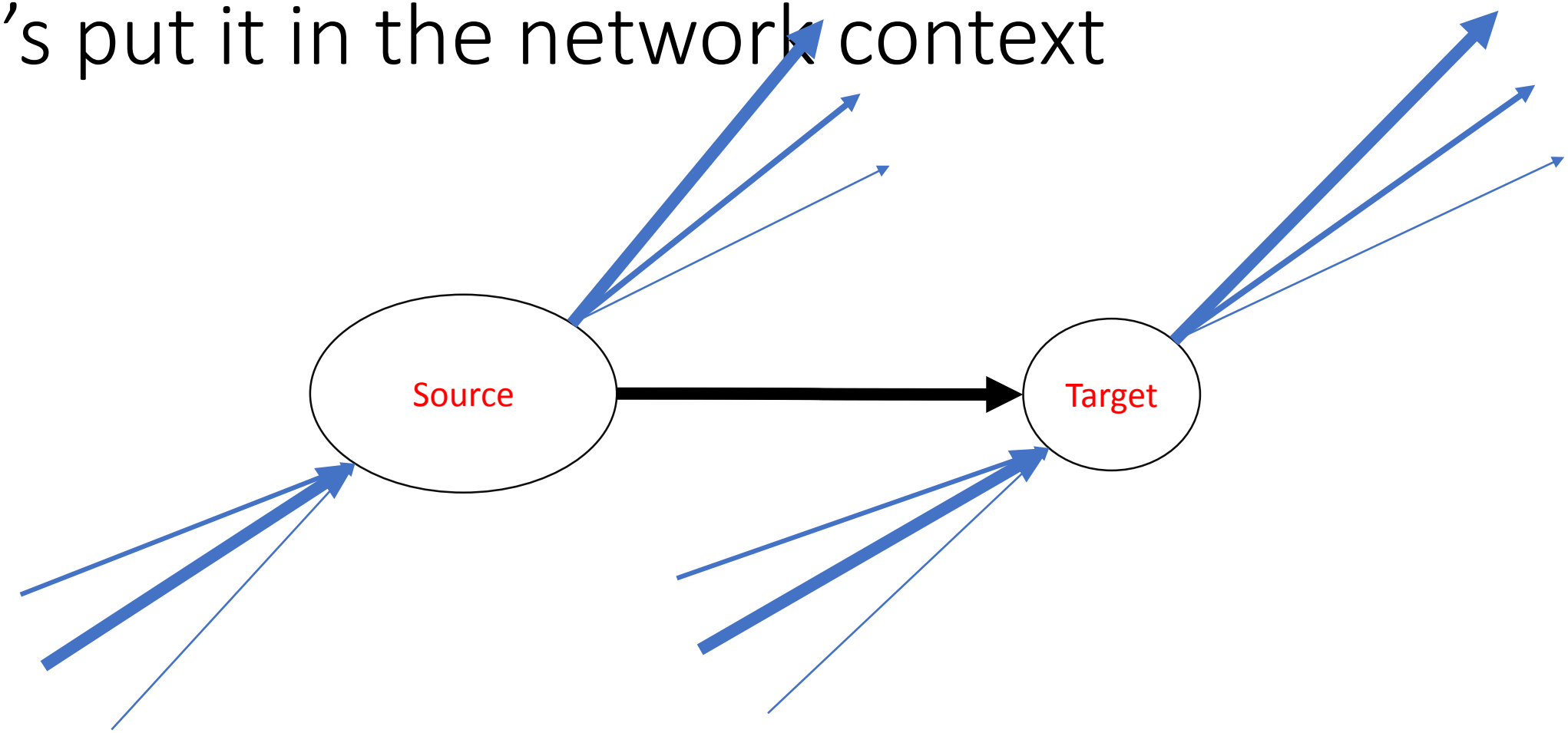


CICT Network Representation

- Directed or undirected
- Unsigned

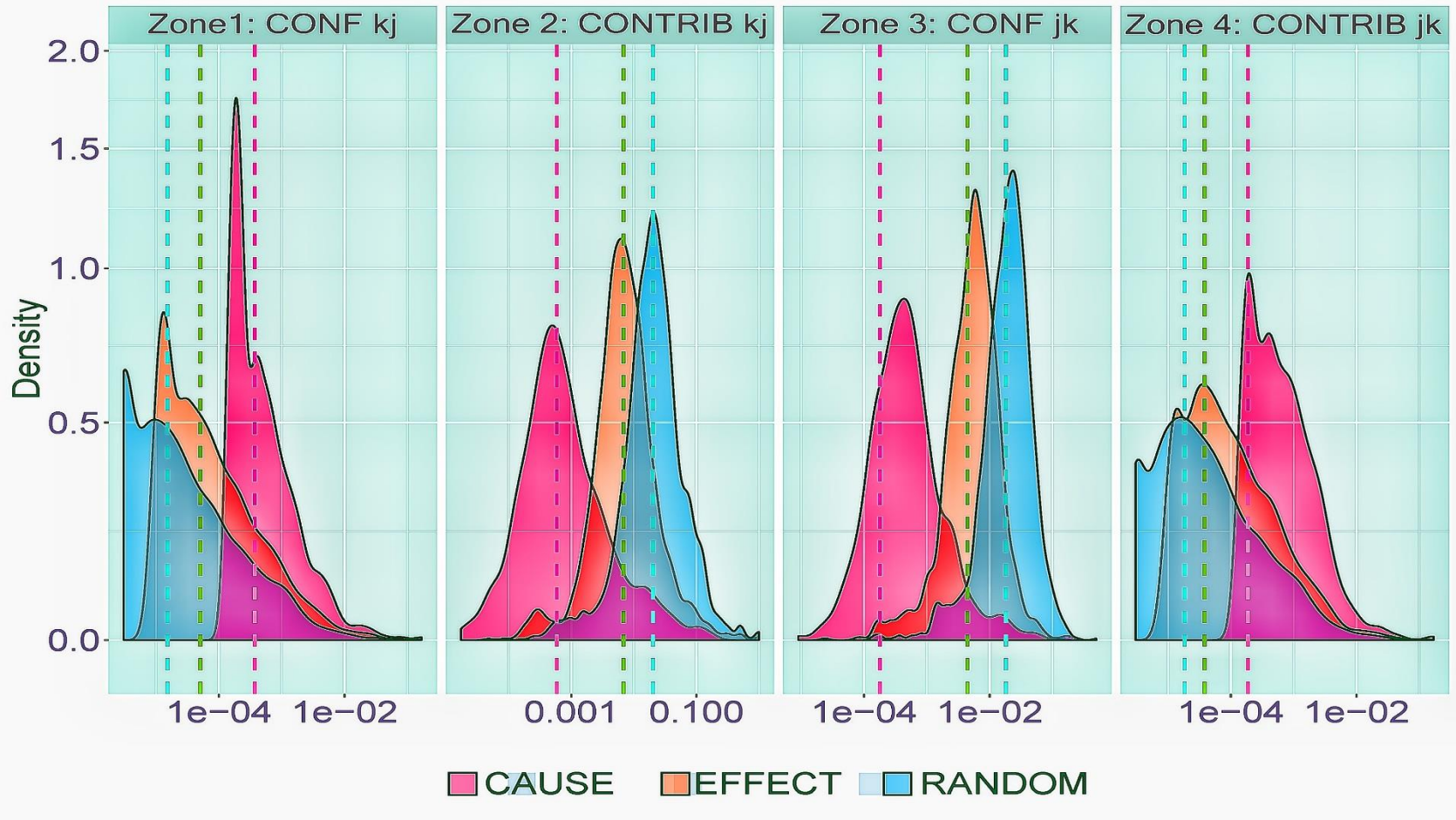


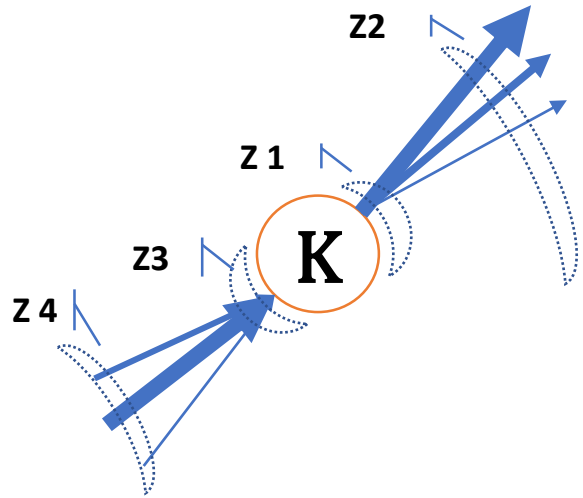
Let’s put it in the network context



Significant in CICT defined zones between causal and random events

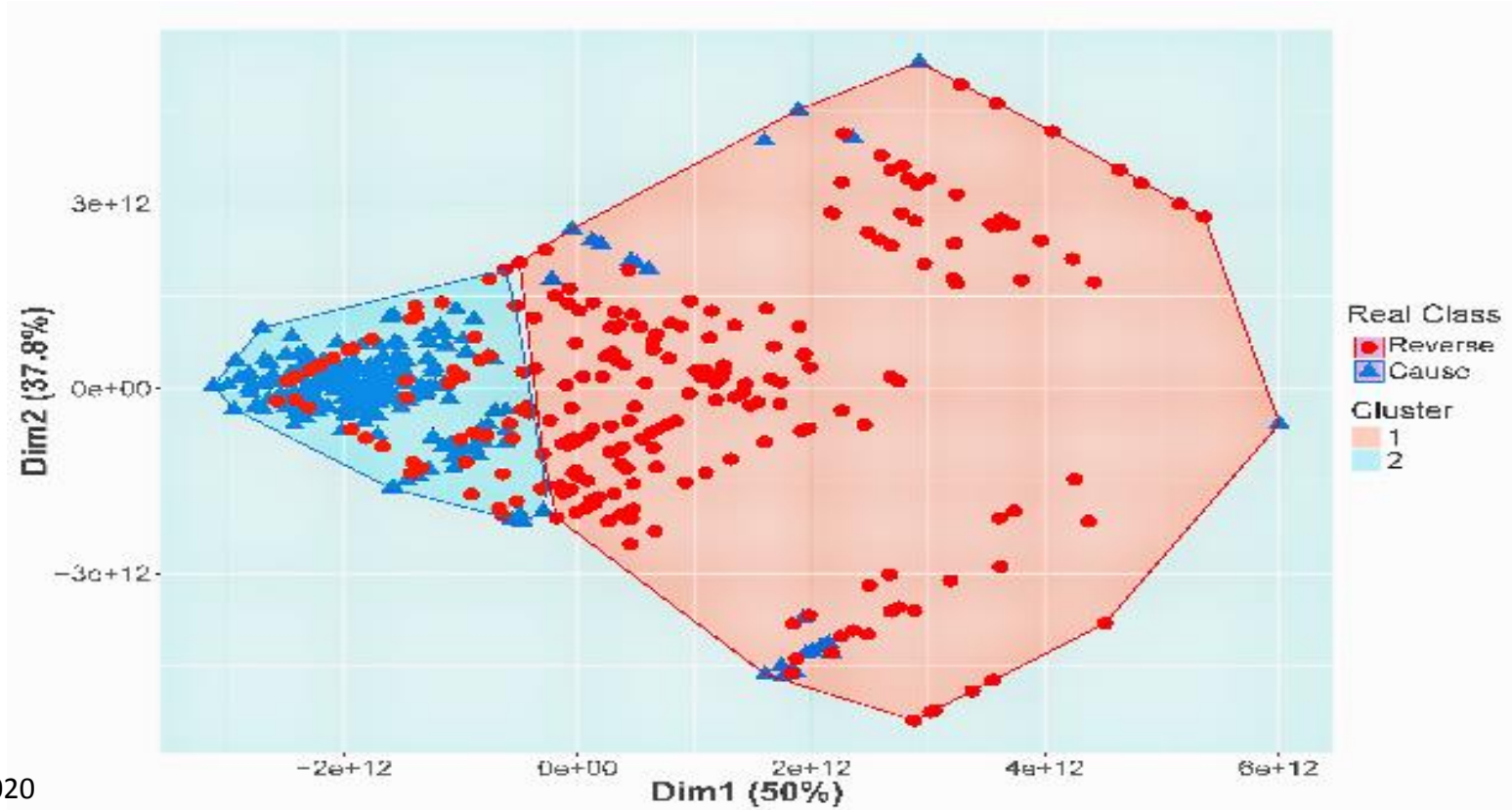
Cause: Rheumatoid Arthritis(red), an effect: Syncope (orange) and a random event: Pneumonia (blue)



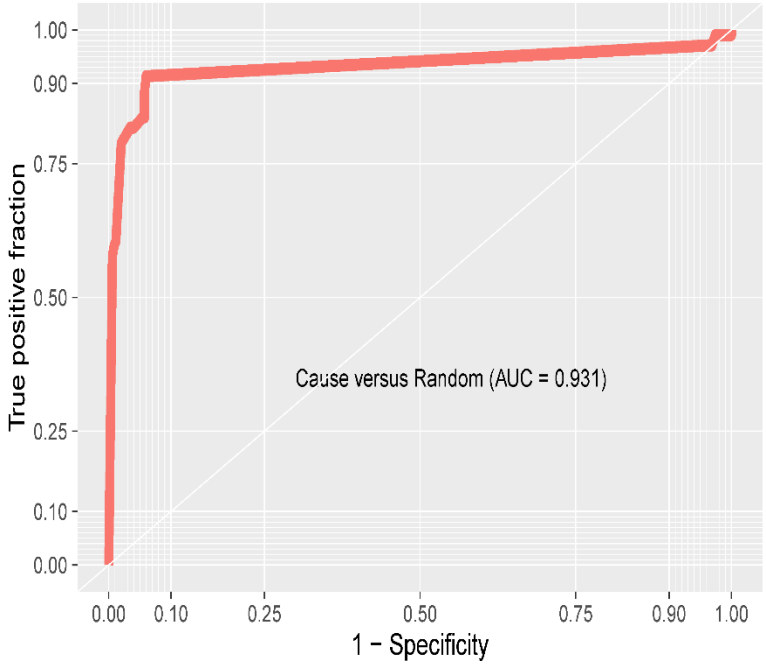


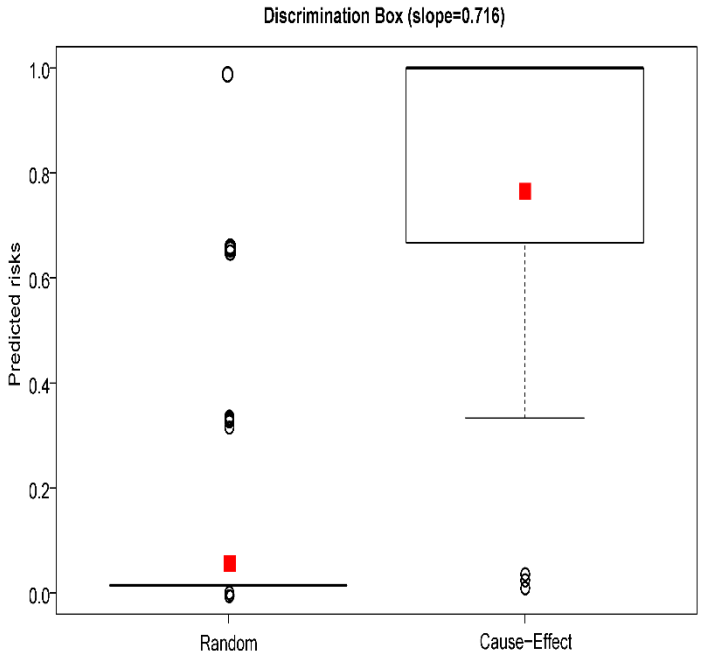
CICT features enables clustering

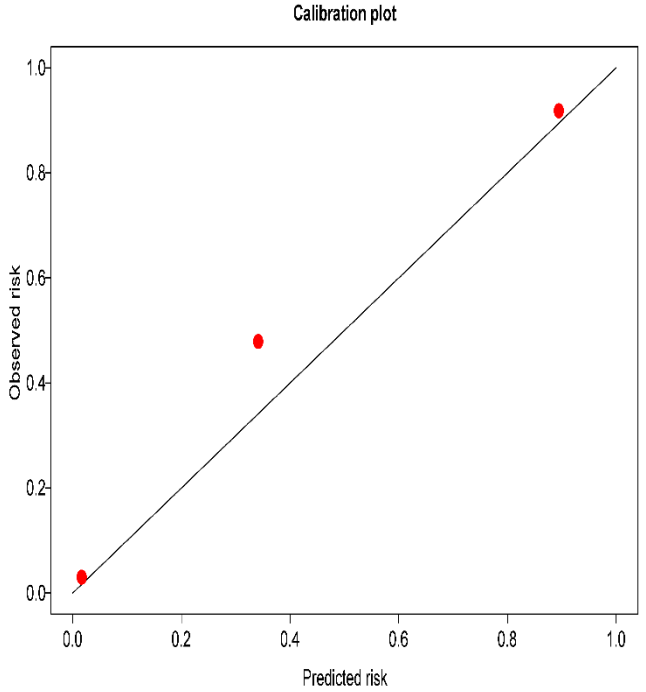
- Two clusters as identified by Partitioning Around Medoids along with the real class of data points displayed on the first and second dimensions of PCA.
- Adjusted Rand Index [9] shows 0.468 agreement between clustering results and real classes



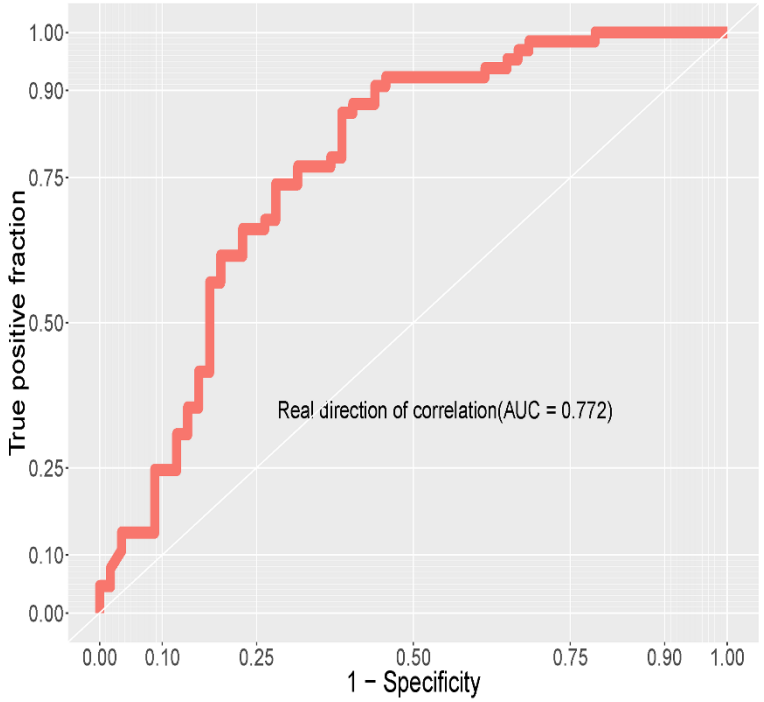
Causal association versus random

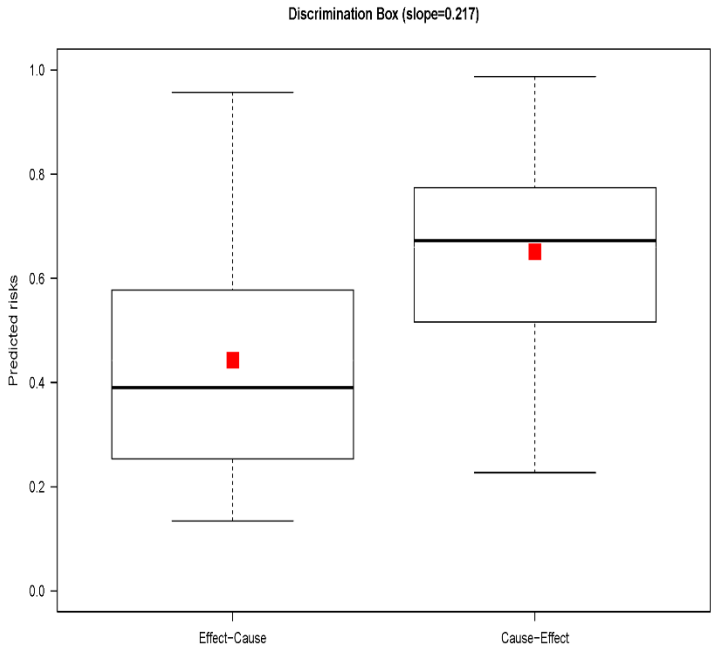


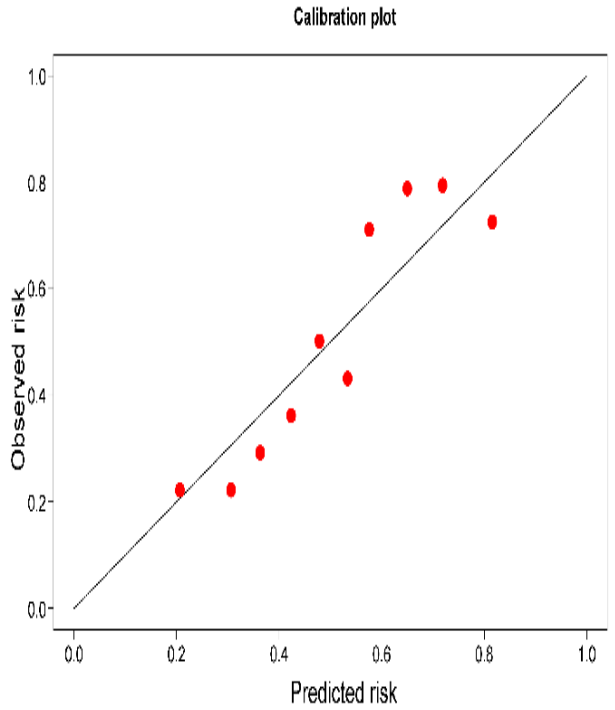




Direction of causality in an association





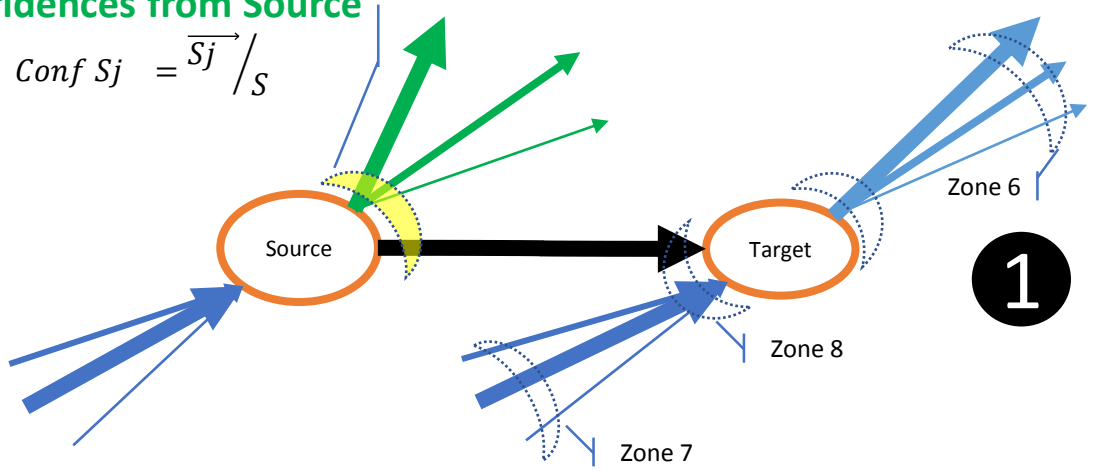


Example of a good and novel discriminator:

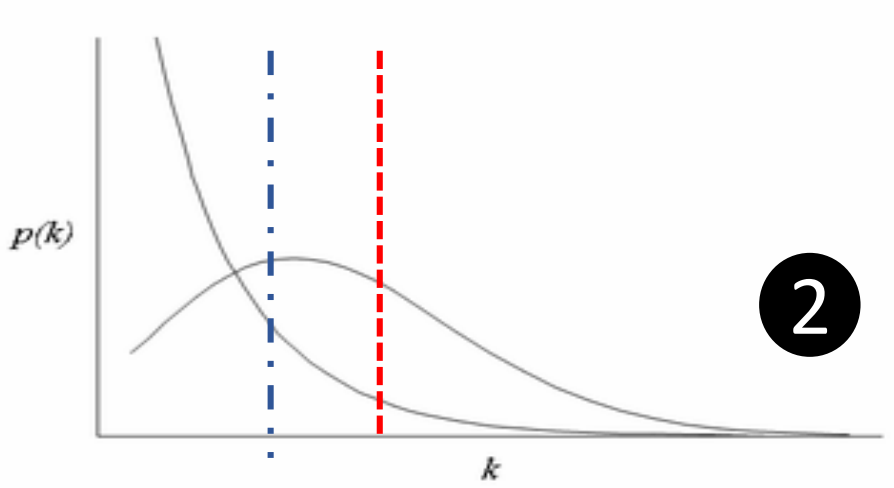
Median absolute deviation of normalized confidences from source

Confidences from Source

$$Conf\ Sj = \overrightarrow{Sj} \Big/ S$$

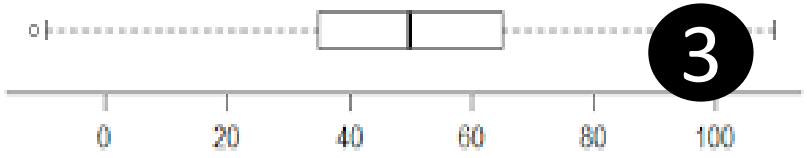


1

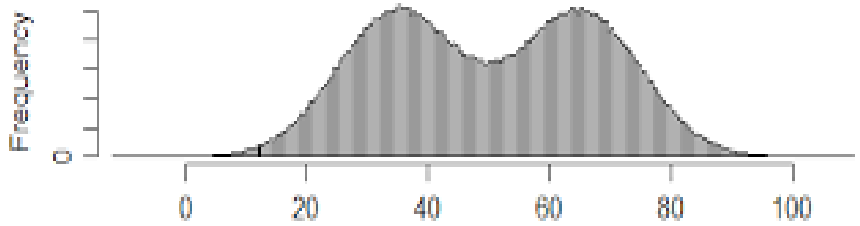


2

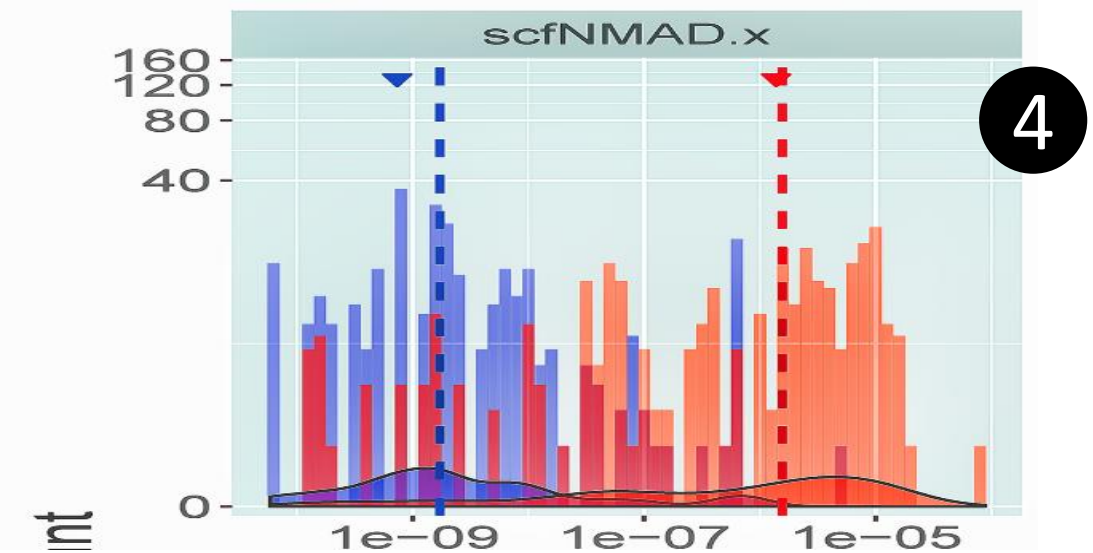
Median of confidence distributions



3

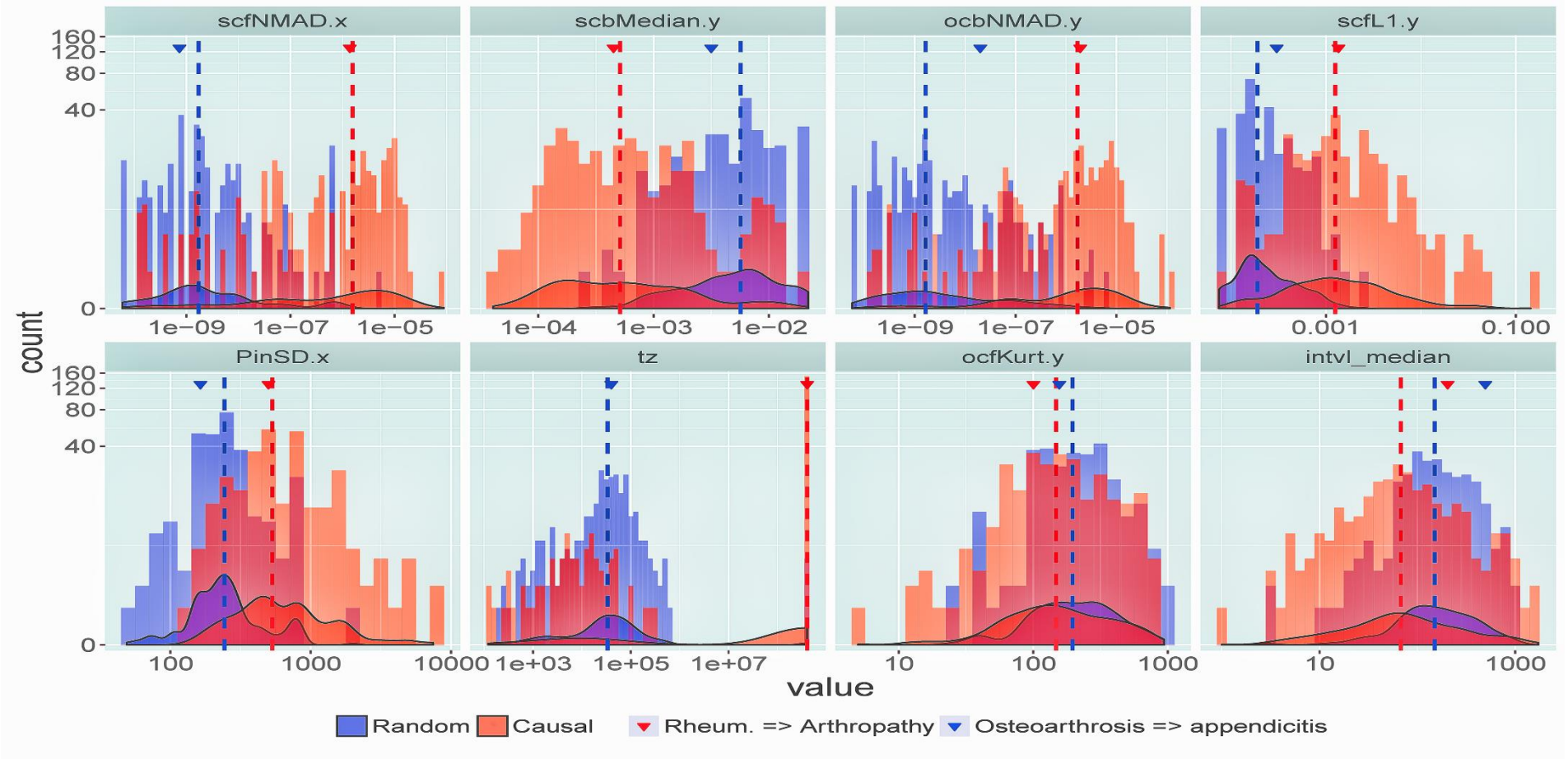


Distribution of medians



4

Decision analysis: top Predictors of causal relations



Results

- Identifies well-known causal relationships
  - Example:
    - Hypertension -> myocardial infarction (hearth attack), smoking -> lung cancer
- Identified several novel findings
  - 11 original findings reported to scientific community as papers or presentations
  - Confirmed by epidemiological time-to-event studies after controlling for all confounders. Examples:
    - Sleep apnea => heart failure
    - Viral pneumonia => pulmonary fibrosis
    - Disorders of coping with stress => heart problems

Significance

- No confounders used
- Highly accurate (AUC ROC > 0.9) discrimination of causal versus random relationships
- Non parametric, no assumptions made on the distribution of input or output
- Applicable to Markov Chain data (directed one-step graph networks) and Markov Networks (undirected one-step graph networks) to infer directional relationships
- Simple method, computationally efficient,
- Scalable at linear time and space complexity in both learning and prediction phases
- Can use numerical and discrete value edges and frequencies
- Not limited to specific constraints on output network structure (e.g. can be cyclic)

CICT – DREAM4

Application of CICT on simulated biological regulatory networks

DREAM4

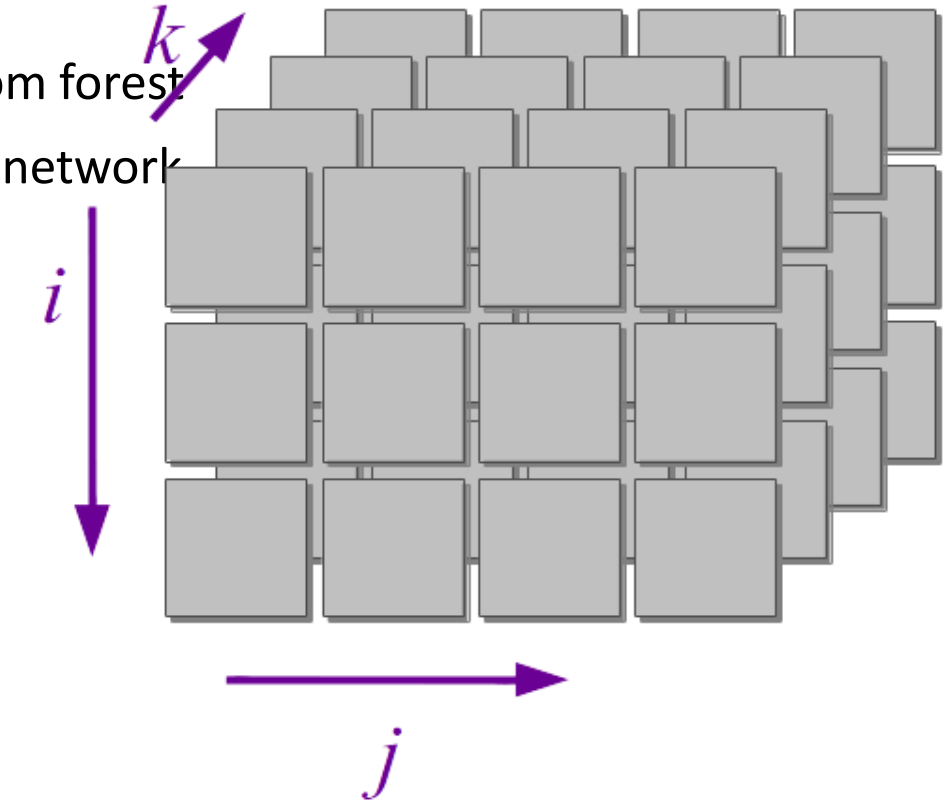
- Systems biology to uncover causal relationships between genotypes and phenotypes
- Identifying Gene Regulatory Networks(GRN) is a main objective
- Dialogue for Reverse Engineering Assessments and Methods (DREAM)
- Annual challenges in systems biology

DREAM4 project

- Inferring gene regulatory networks
- 5 networks each with 100 genes, for each:
  1. Gold Standard (ground truth)
     - Subnetworks from transcriptional regulatory networks of Escherichia coli and Saccharomyces
  2. Simulated wildtype steady state, knockouts, knockdowns, dual knockouts and multifactorial perturbation
  3. Simulated time series
     - 10 time series
       - Each with 21 time points
       - T=0 perturbation happens and continues till time point 10, then perturbation removed and go back to wild type for ten more rounds
       - Perturbation affects one third of all genes
- Objective:
  - predict the underlying network
  - Measures: AUC ROC and AUC PR comparing to Gold Standard data.

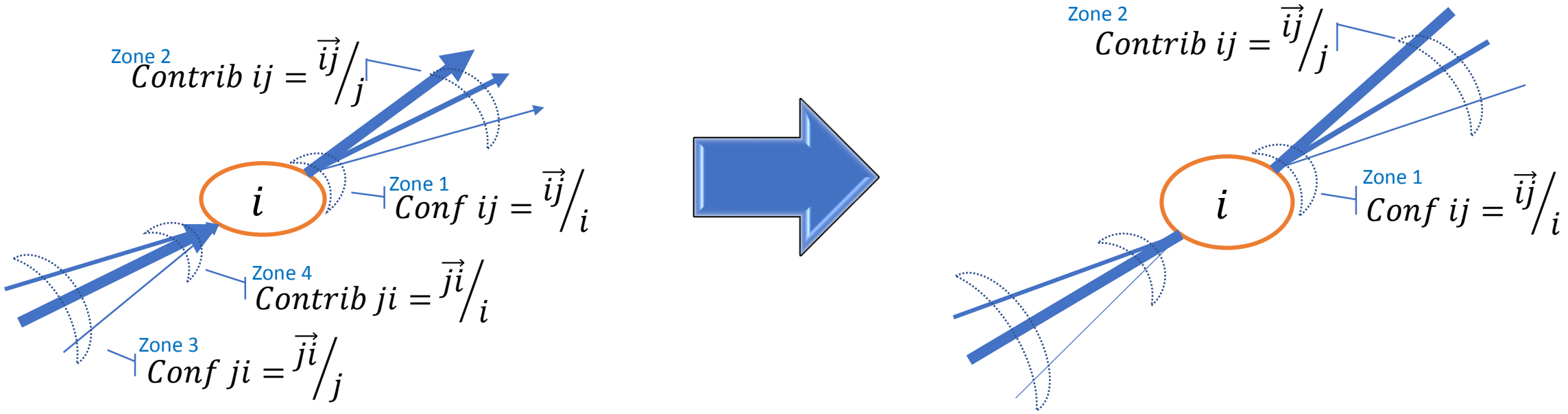
Applying CICT on DREAM 4 Time Series Data

- Calculated mutual information ij between pairs of genes i and j
- Collapsed all K time series data to a CICT network presentation.
- CICT feature production
- Supervised learning with regularized regression and random forest
- Evaluation of the model performance using gold standard network



Applying CICT on DREAM 4 Time Series Data

- CICT network representation has two less distribution zones for a relevance undirected network



Results

- AUC ROC = 0.83

Golden standard Network VS Predicted Network

Research plan

Single Cell RNA seq

- Complex organism
  - Specialized tissues
    - Location: Spatial profiling
    - Timing: reaction profiling
    - Functional profiling of cells
    - Developmental profiling
- Challenges in applying to plant biology
  - Cell walls, vacuoles, chloroplasts and some secondary metabolites
  - Effective ways to identify underlying network

A rich knowledge representation