

# Algorithms for Very Large Scale Causal Discovery & Feature Selection

Constantin F. Aliferis & Ioannis Tsamardinos

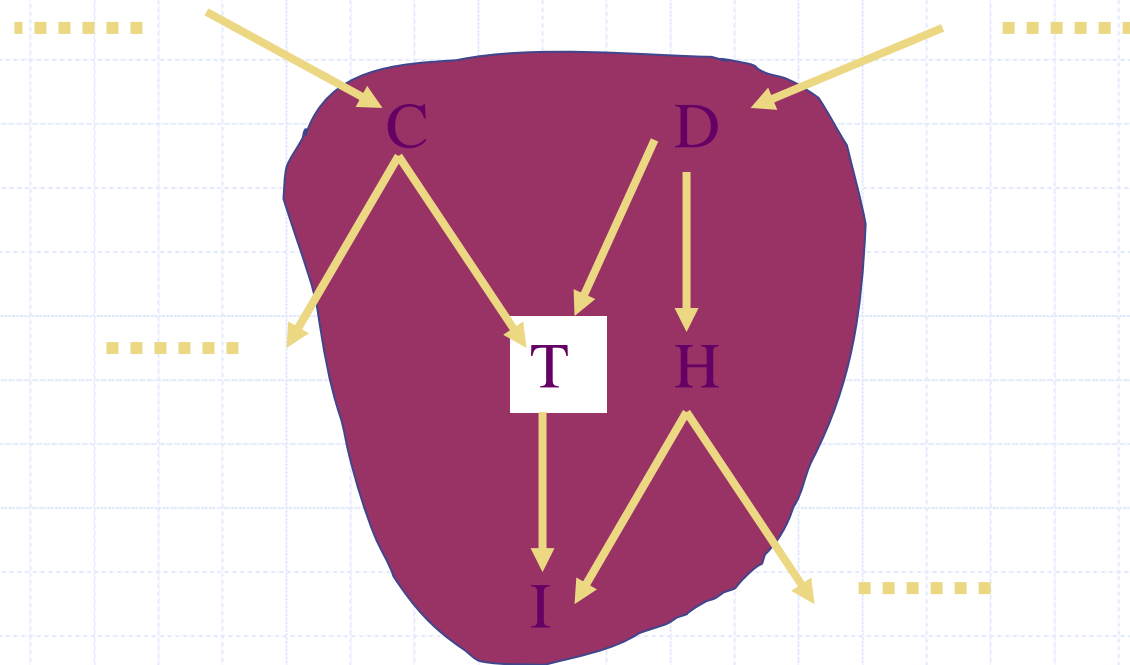
Discovery Systems Laboratory,  
Department of Biomedical Informatics,  
Informatics Center,  
Vanderbilt University

Update of 9-12-2002

# Algorithms for Discovery: Causal Neighborhood

## Definition:

- The Markov Blanket of some variable of interest **T** ("MB(T)") is the set of the immediate causes, immediate effects, and immediate causes of the immediate effects of **T**.

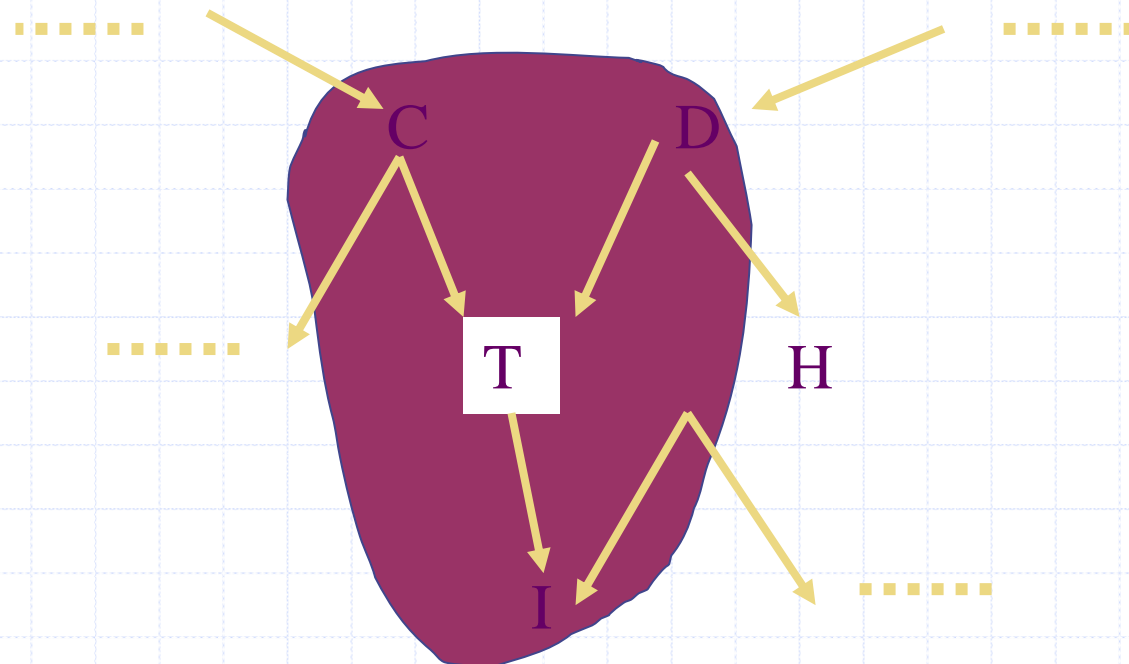


Note: C causes T, T causes I, etc.

# Algorithms for Discovery : Causal Neighborhood

## Definition:

- The Direct Causal Neighborhood can alternatively be defined as the set of Direct Causes and Effects of some variable of interest  $T$  ("DCE( $T$ )").



Note: C causes T, T causes I ,etc.

# Algorithms for Discovery: Problem Statement

## ◆ Goal:

- Given: Data (observations of  $\mathbf{T}$  and a set of variables  $V$ )
- Find:  $MB(\mathbf{T})$  or  $DCE(\mathbf{T})$

## ◆ Why?

# Algorithms for Discovery: Motivation

## ◆ Applications:

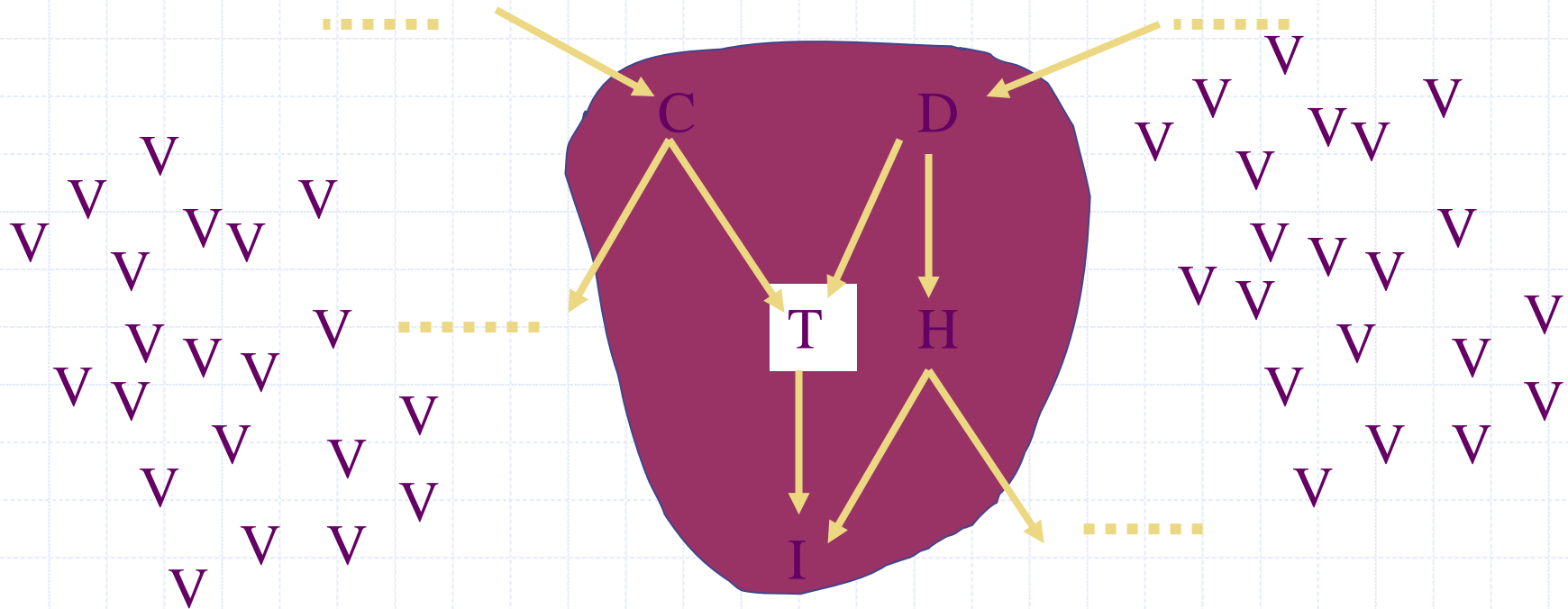
- MB(T) is the minimal set of predictor variables needed for classification (diagnosis, prognosis, etc.) of the target variable **T**
- MB(T) and DCE(T) help us discover immediate causes and effects of **T**
- MB(T) and DCE(T) can be used to discover total causal structure of domain (what variable is causing/is caused by what other variables)
- DCE(T<sub>i</sub>) are specific/fine-grain “causal clusters” of variables T<sub>i</sub>

# DSL Algorithms for Discovery

- ◆ Previously MB(T) and DCE(T) could be discovered using a full-network induction algorithm, or various heuristic procedures
- ◆ Characteristics of newly-developed algorithms :
  - Sound given broad and well-defined assumptions
  - Scale up to hundreds of thousands of variables
  - Quality of output insensitive to errors in learning about the rest of the variables
  - Computational performance insensitive to structure beyond the target **T**
  - Behave well when confounders are not observed

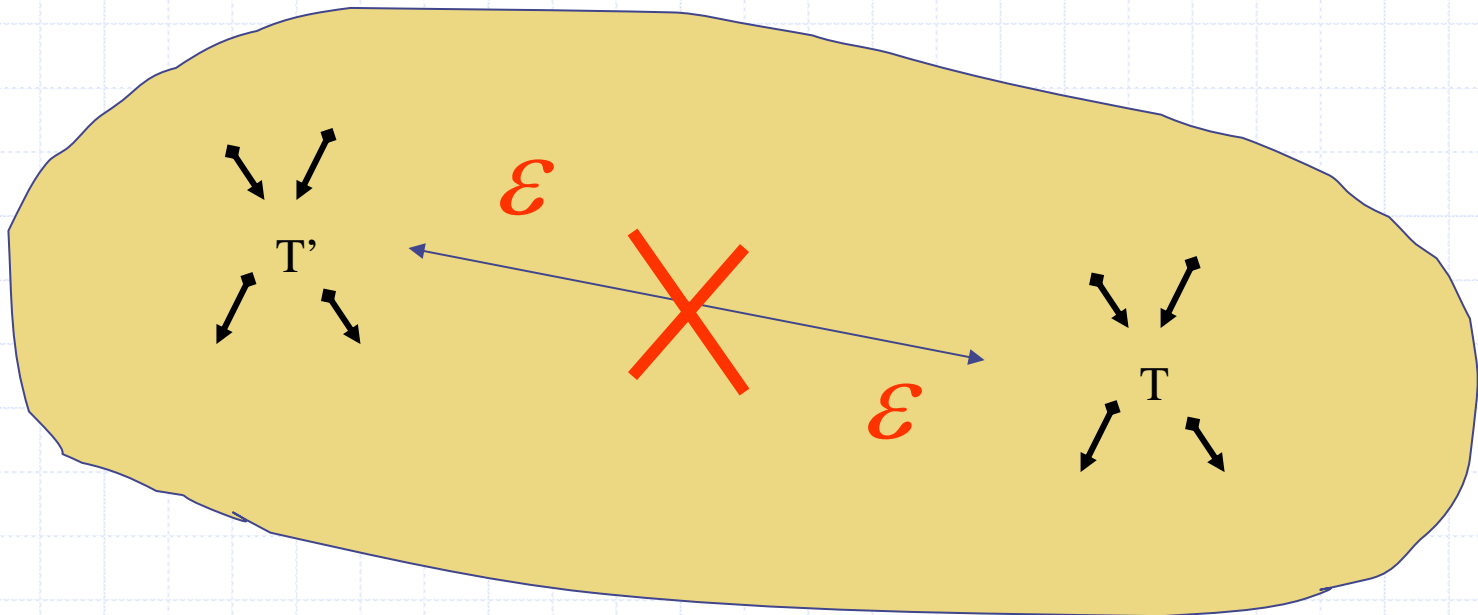
# Scalability

- ◆ As long as  $MB(T)/DCE(T)$  is small relative to the available sample size, we can discover  $MB(T)/DCE(T)$  regardless of how many variables are present in the data; our methods scale up to extremely large numbers of total variables without sacrificing soundness;
- ◆ The state-of-the-art (full-network) algorithms try to learn the whole network and are not tractable for large networks



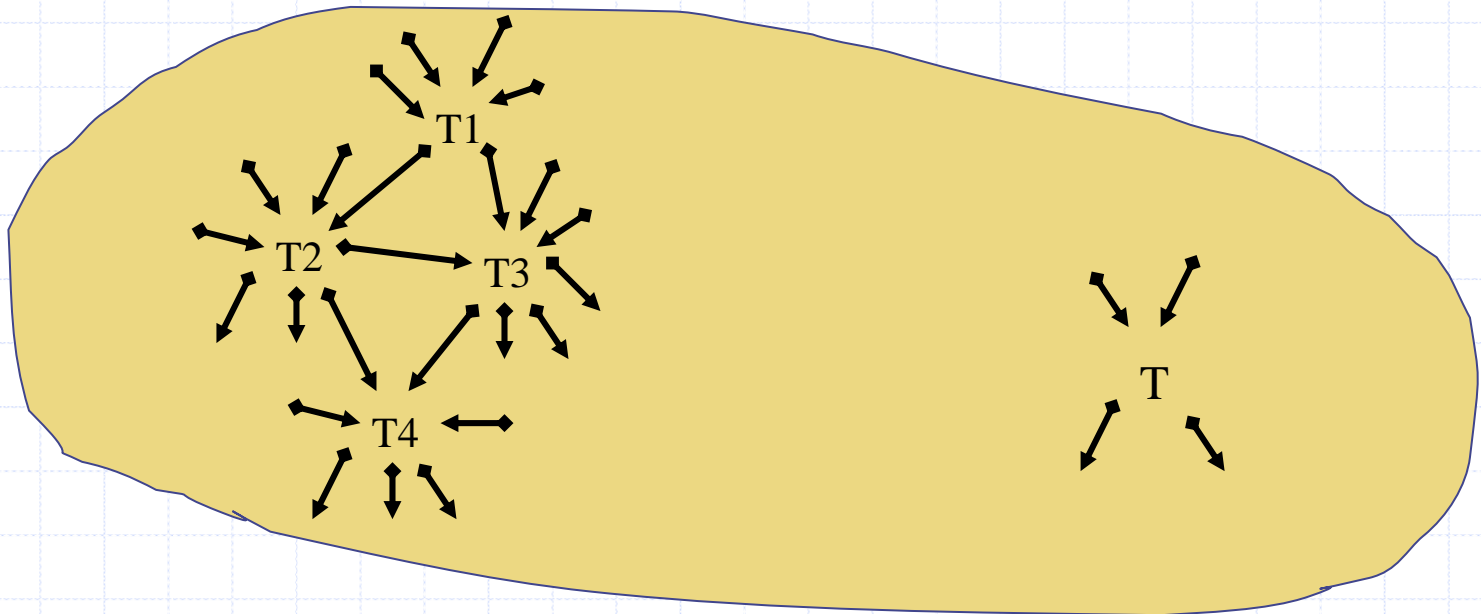
# Insensitivity to errors in other parts of the network

- ◆ As long as  $MB(T)/DCE(T)$  is small relative to the available sample size, we can discover it regardless of how successfully one can infer the MB/DCE of other features (i.e., remote errors do not contaminate local discovery);
- ◆ The state-of-the-art full-network algorithms do not have this property.



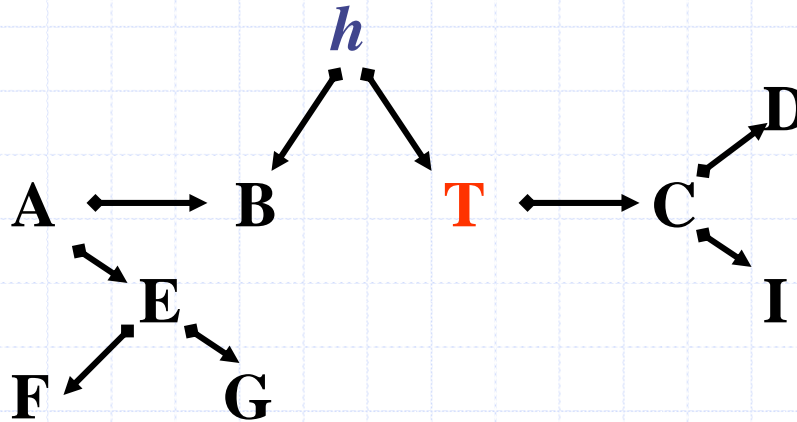
# Insensitivity to variable relationships in other parts of the network

- ◆ As long as  $MB(T)/DCE(T)$  is small relative to the available sample size, we can discover it regardless of how complex is the rest of the process.
- ◆ The state-of-the-art full-network algorithms do not have this property.



# Discovery “Monotonicity”

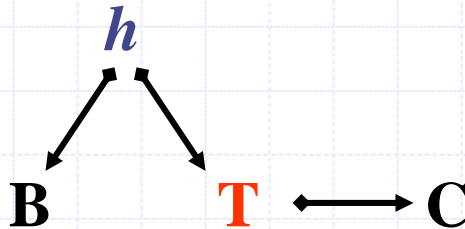
- ◆ Say true structure is:



- ◆ With  $h$  unobserved, any sound algorithm (including DSL algorithms) will return:  $DCE(T) = \{B, C\}$

# Discovery “Monotonicity”

- ◆ Say now that we can observe  $h$ .



- ◆ By analyzing *only* the previous  $DCE(T) = \{B, C\}$ , and  $h$  we get the new  $DCE(T) = \{h, C\}$ .
- ◆ State-of-the-art full-network algorithms either try to learn the hidden structure (an intractable and incomplete task) or have to re-analyze the full network (since local errors propagate as explained earlier)

# Ongoing Work

## ◆ Forthcoming:

- Discover total causal structure of domain
- Discover generalized-XOR (Gx) and quasi-Gx functions (e.g., for multi-gene etiology) as well as non-faithful distributions of any kind
- Filter causal hypotheses generated by application of other methods
- Combine with clustering to identify new molecular disease types
- Discover MB(T)/DCE(T) in dynamic domains/temporal data
- **Problem-driven applications/evaluations**