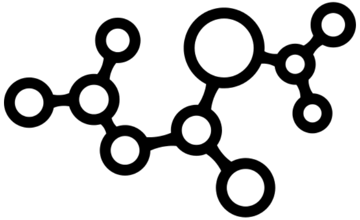
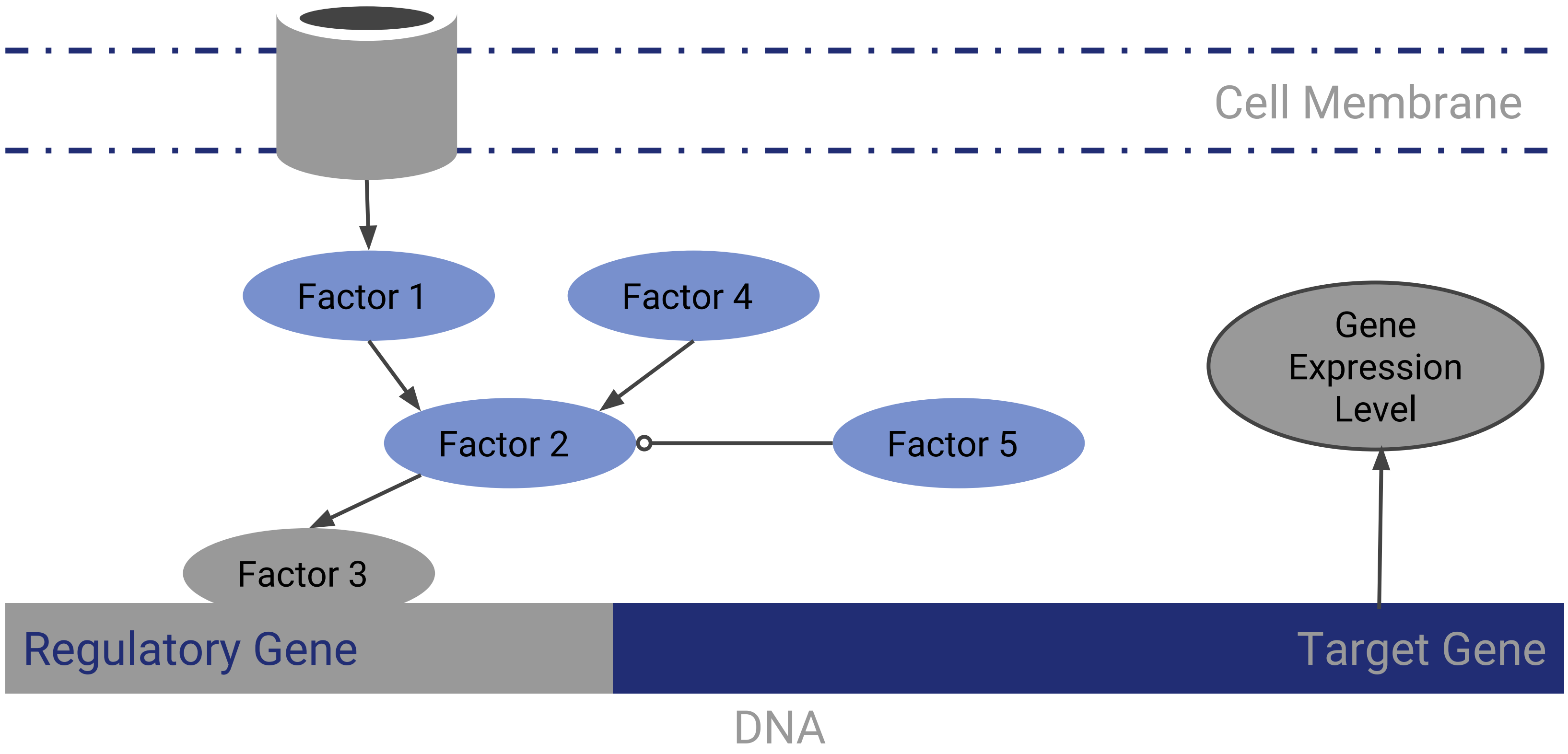
Causal Computational Models for
Gene Regulatory Networks

Sahil Loomba
Parul Jain

Advisors
Dr. Sumeet Agarwal
Dr. Parag Singla



Reintroducing the GRN Problem



Where BTP1 finished…

- Correlation $$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$
- Granger Causality $$x_t = a_0 + \sum_{i=1}^{m} a_i x_{t-i} + \sum_{i=1}^{q} b_i y_{t-i} + \epsilon_t$$
- Mutual Information $$I(X,Y) = H(X) - H(X|Y)$$
- Transfer Entropy $$T(X,Y) = T_{Y \to X} = H(X_t | X_{t-1:t-d}) - H(X_t | X_{t-1:t-d}, Y_{t-1:t-d})$$

Parameters: Size, quantisation, time, lag

Asides: Grid Search, Laplace Smoothing

TE ~ MI > GC > CO

… is where BTP2 picks up

	Linear	Non Linear
Non Predictive	Correlation	Mutual Information
Predictive	Granger Causality	Transfer Entropy

As Random Variable

Convergent Cross Mapping

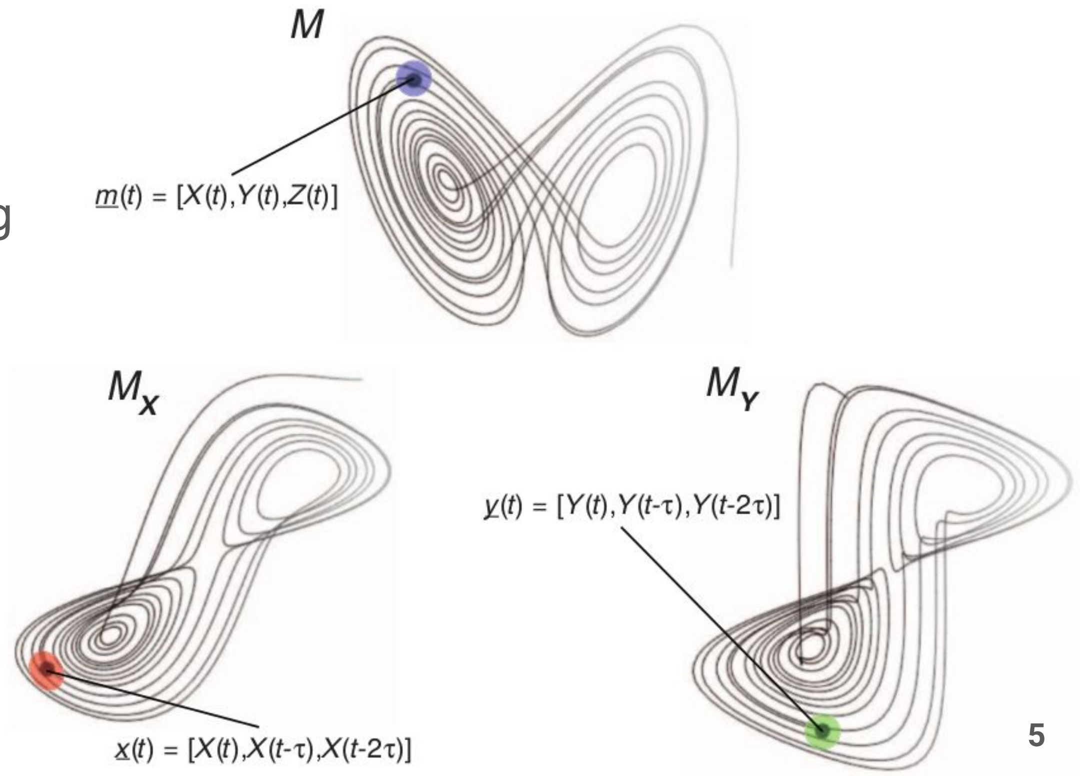
As Dynamical System

Convergent Cross Mapping [Sugihara et al. 2012]

Diffeomorphism across Shadow Manifolds

- For weakly coupled dynamical systems
- New notion of causality: belonging to **same dynamical system**
- Library size ∝ Time series length
- Parameters:
  - Dimensions: M, E where E ≥ M
  - Lag: τ

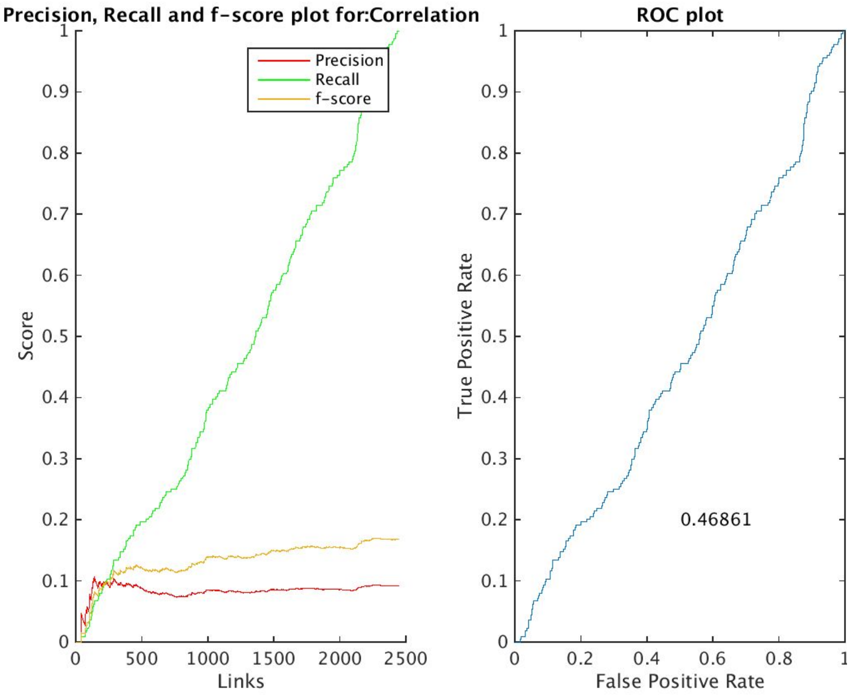
$$C(X,Y) = \rho(X, \hat{X}_Y)$$

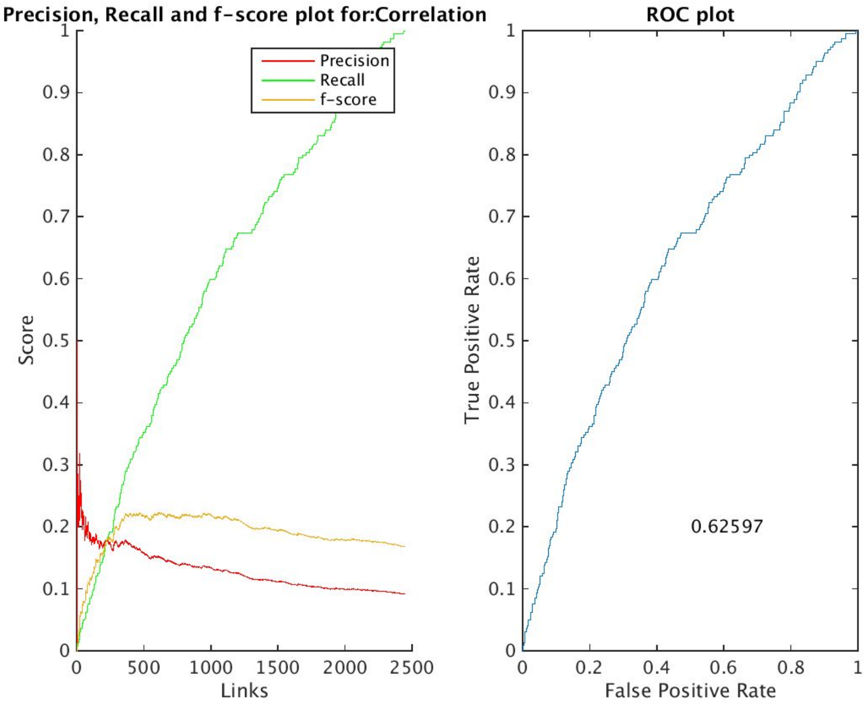


Simulated Data

- SysGenSIM
  - Steady state data
  - Gene expression of different individuals
  - Size - 50, 100   Series - 500, 1000
- DREAM4 dataset [Young et al. 2014]
  - GeneNetWeaver software
  - Time series data
  - Size - 10, 100   Series - 21

Normalisation should be the norm!

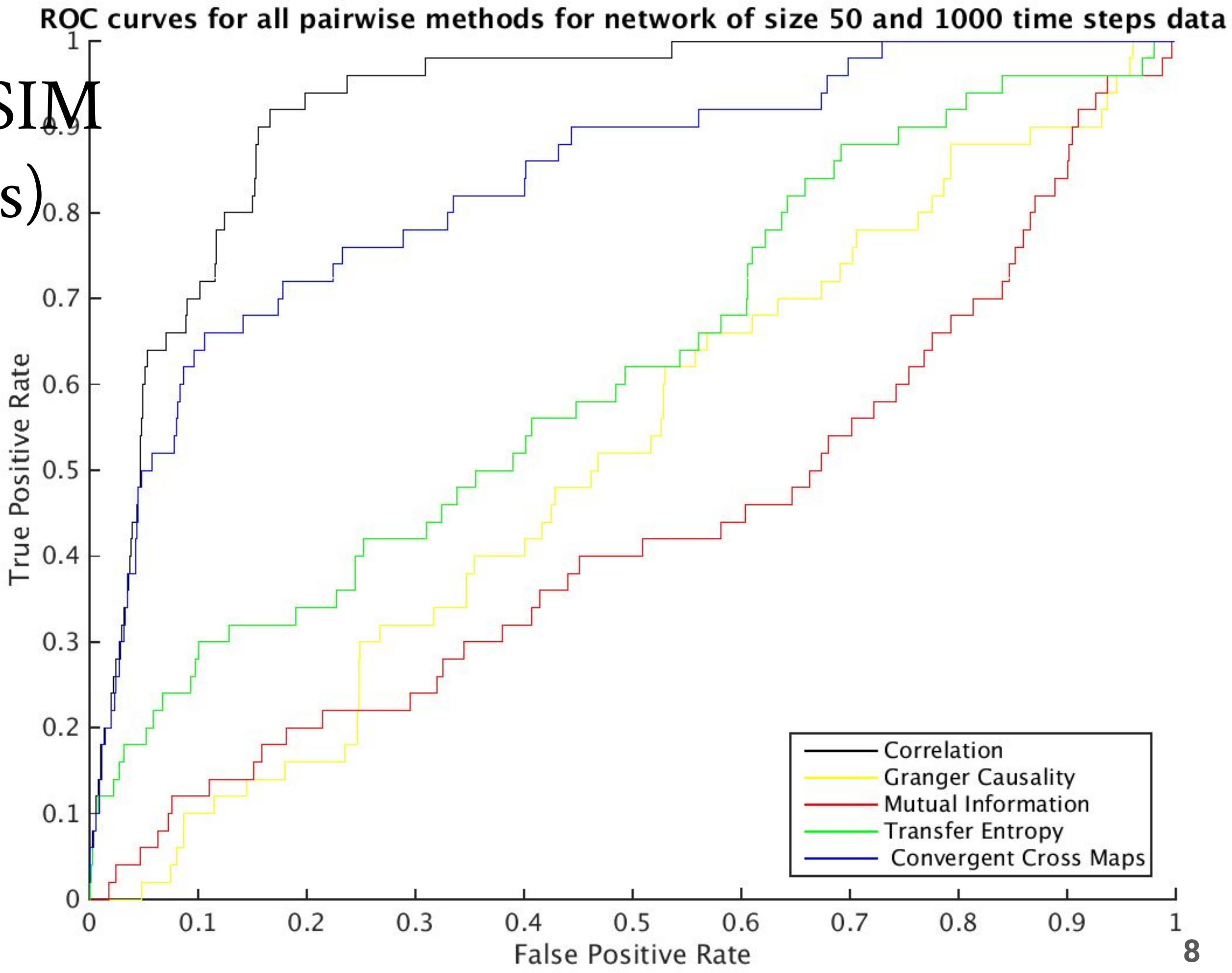




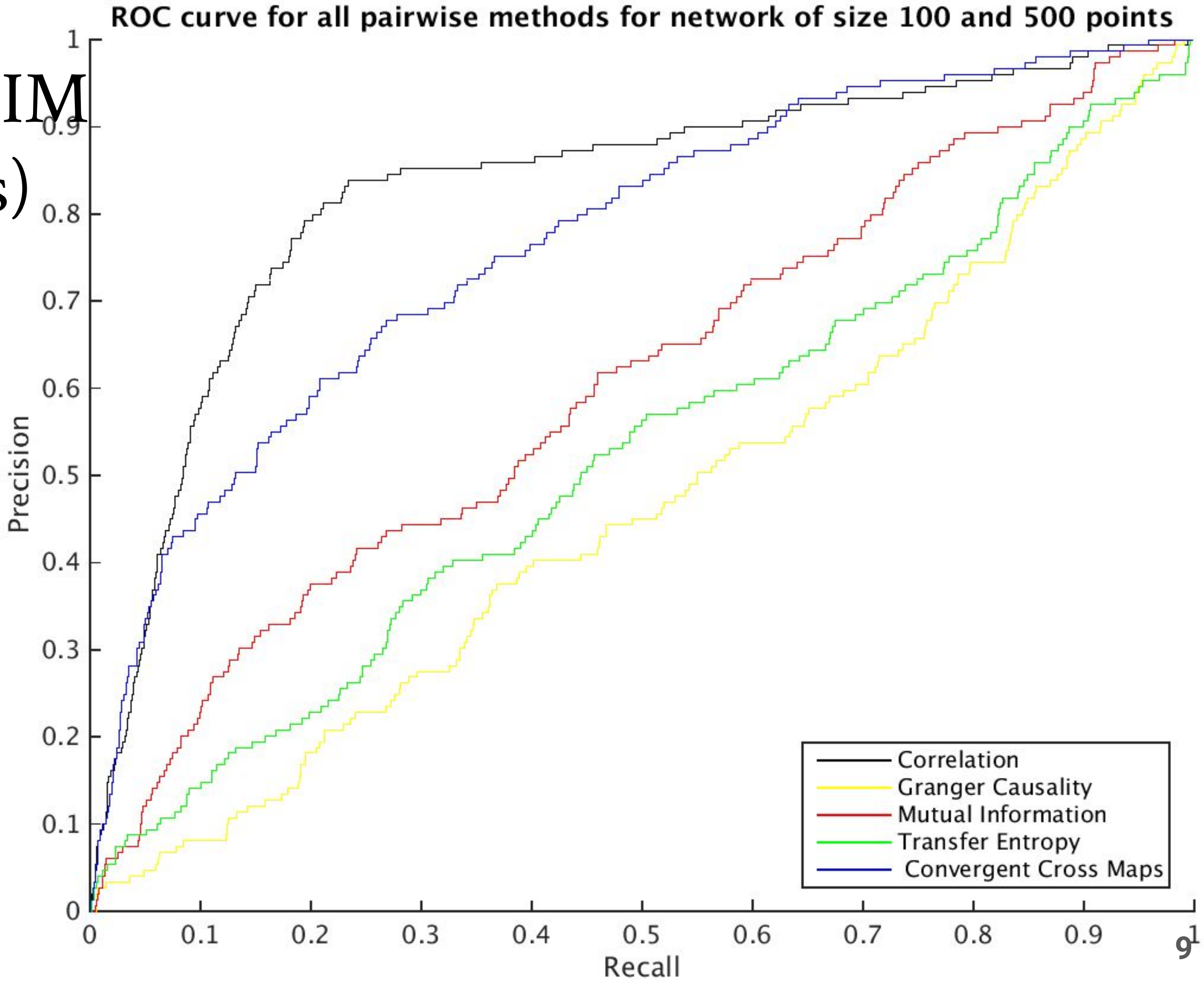
Network Size 50, Time Series Length = 1000

Normalised Time Series data

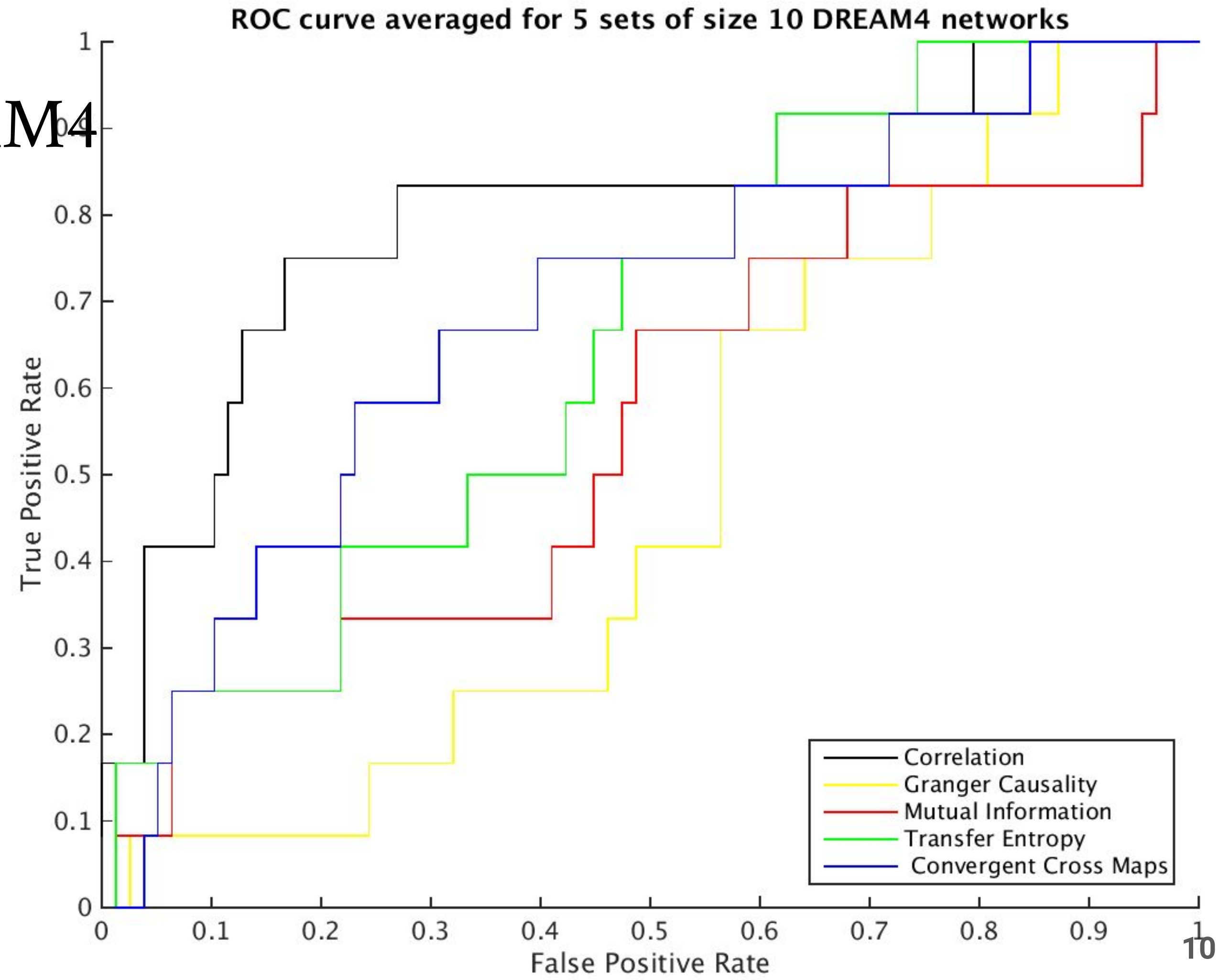
Results for SysGenSIM (size 50, 1000 points)



Results for SysGenSIM (size 100, 500 points)



Results for DREAM4 (size 10)

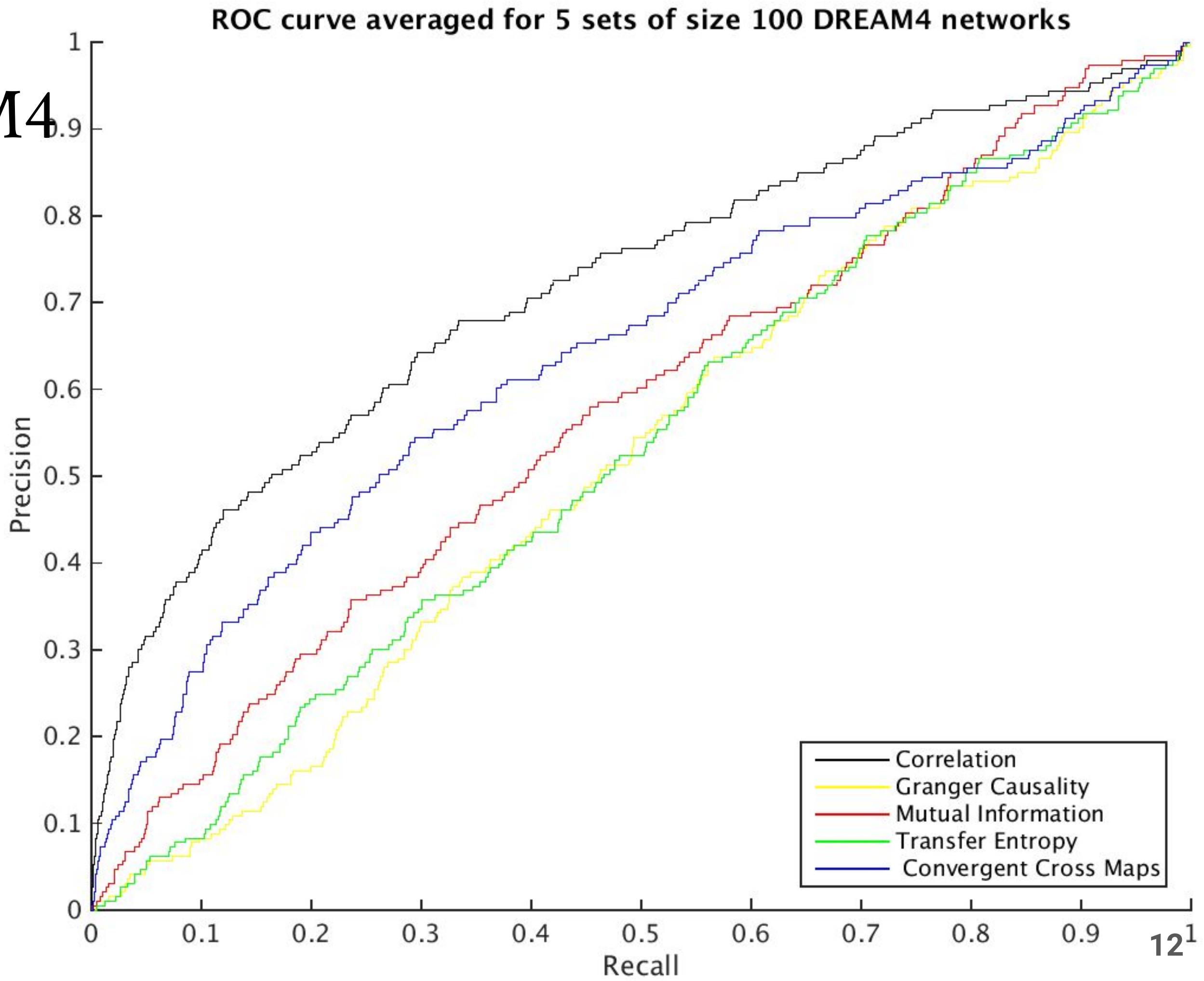


Results for DREAM4 (size 10)

	CCM	Correlation	GC	MI	TE
1	0.55644	**0.70667**	0.55467	0.43733	0.69511
2	**0.75676**	0.60135	0.60557	0.52449	0.61149
3	0.55467	**0.74756**	0.4	0.49333	0.65067
4	0.67333	**0.74625**	0.46853	0.57742	0.43057
5	0.69231	**0.79915**	0.47436	0.55342	0.65491
Average	**0.646702**	**0.720196**	**0.500626**	**0.517198**	**0.60855**

ARACNE : AUROC = 0.668 (Young et al. 2014)

Results for DREAM4 (size 100)



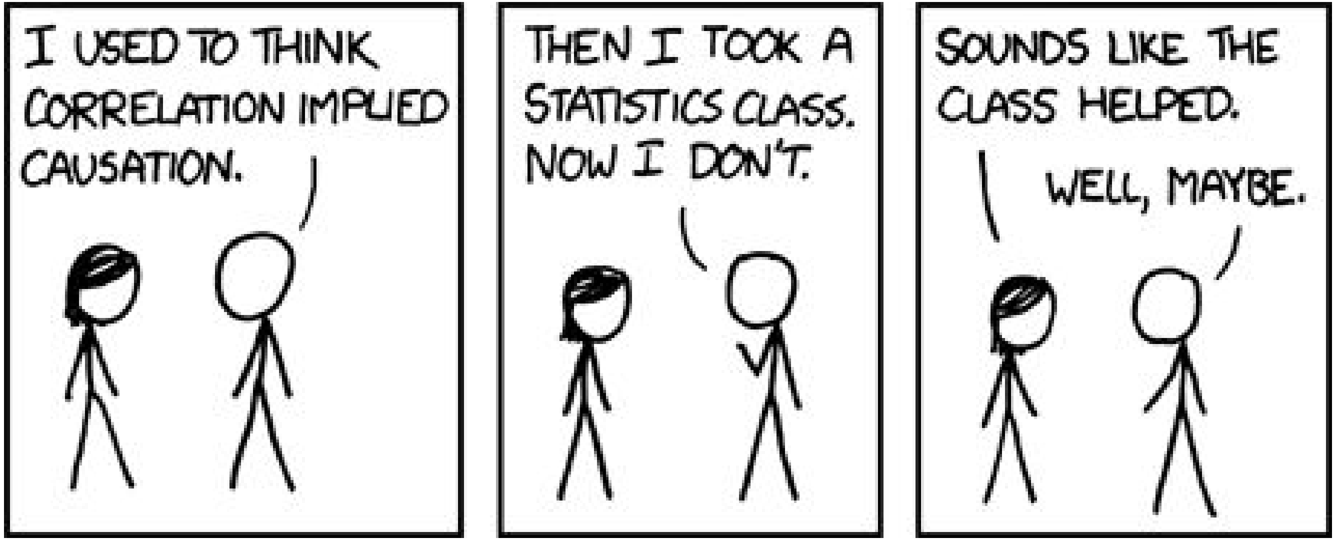
Results for DREAM4 (size 100)

	CCM	Correlation	GC	MI	TE
1	**0.7279**	0.75607	0.45483	0.5949	0.47637
2	**0.64951**	0.64629	0.52156	0.56015	0.53877
3	0.69577	**0.71094**	0.50553	0.54045	0.53193
4	0.61088	**0.68933**	0.53896	0.54633	0.52694
5	0.64136	**0.71993**	0.51951	0.57414	0.52895
Average	**0.665084**	**0.704512**	**0.508078**	**0.563194**	**0.520592**

ARACNE : AUROC = 0.589 (Young et al, 2014)

Pairwise Metrics of Causality

A Summary



CCM and Correlation *seem* to work the best at a pairwise level

Naive Edge Selection

1. (Since all pairwise metrics are directly proportional to the strength of causality,) sort ${}^{n}C_{2}$ edges by metric value in decreasing order.
2. Choose top-k edges and output as graph G.

Smart Edge Selection

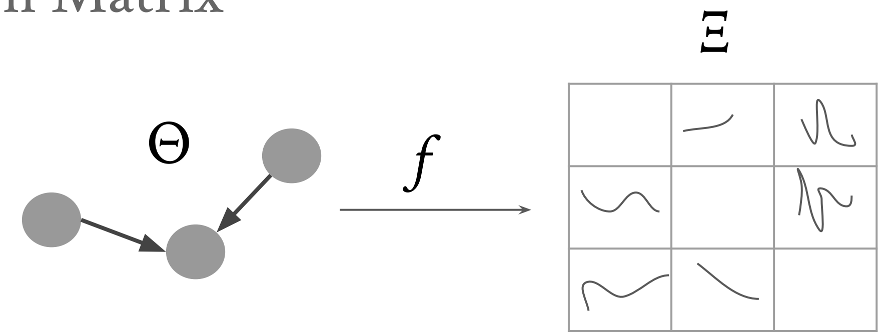
Future Work

Use a “sophisticated” algorithm which selects top-k edges, by making use of graph connectivity and other constraints information.

Intrinsic Graph Structure Estimation [Hino et al. 2015]

Adjacency Matrix to Observation Matrix

$$f : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$$



$$\Theta \mapsto f(\Theta) = \Xi$$

(Pairwise Metrics used here)

$$\xi_{ij} = c_i + c_{ij}\theta_{ij} + \sum_{k \in V} c_{ij}^{k}\theta_{ik}\theta_{kj} + \sum_{k,l \in V} c_{ij}^{kl}\theta_{ik}\theta_{kl}\theta_{lj} + \ldots$$

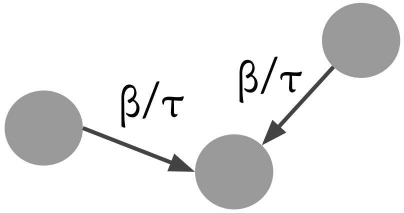
$$t_{ij} = \xi_{ij} + \epsilon$$

Intrinsic Graph Structure Estimation [Hino et al. 2015]

The Random Walk Model

$$L(\Theta) = \begin{bmatrix} \sum \theta_{1k} & -\theta_{12} & \cdots & -\theta_{1n} \\ -\theta_{21} & \sum \theta_{2k} & \cdots & -\theta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\theta_{n1} & -\theta_{n2} & \cdots & \sum \theta_{nk} \end{bmatrix}$$

Digraph Laplacian



$$f(\Theta) = \alpha e^{\beta L(\Theta)}$$

$$\rho = \{\alpha, \beta\}$$

Intrinsic Graph Structure Estimation [Hino et al. 2015]

Parameter Estimation Algorithm

$$J(\rho, \Theta) = \sum_{i,j \in V, i \neq j} \left(t_{ij} - [f(\Theta)]_{ij} \right)^2$$

In an EM style algorithm, iterating over k (number of edges in graph):

$$\Theta^k = f^{-1}(\Xi)$$

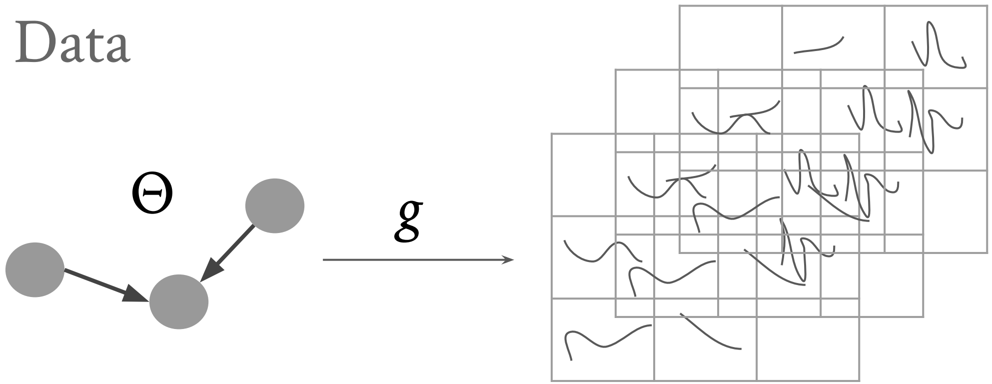
$$\rho_k = \arg\min_{\rho} J(\rho, \Theta^k)$$

Intrinsic Graph Structure Estimation

Moving towards Multi-attribute Data

$$g : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n \times r}$$

$$\Theta \mapsto g(\Theta) = \Xi$$



(Multiple Pairwise Metrics used here)

$$^{q}\xi_{ij} = {}^{q}c_i + {}^{q}c_{ij}\theta_{ij} + \sum_{k \in V} {}^{q}c_{ij}^{k}\theta_{ik}\theta_{kj} + \sum_{k,l \in V} {}^{q}c_{ij}^{kl}\theta_{ik}\theta_{kl}\theta_{lj} + \ldots$$

$$^{q}t_{ij} = {}^{q}\xi_{ij} + \epsilon$$

Intrinsic Graph Structure Estimation

Moving towards Multi-attribute Data

$$g(\Theta) = cat\left({}^1f(\Theta), {}^2f(\Theta), \ldots, {}^rf(\Theta)\right)$$

$$J(\rho, \Theta) = \sum_{1 \le q \le r} \left(\sum_{i,j \in V, i \ne j} \left({}^q t_{ij} - {}^q[g(\Theta)]_{ij} \right)^2 \right)$$

$$\Theta^k = \frac{\sum_{1 \le q \le r} {}^q f^{-1}(\Xi)}{r}$$

$${}^q\rho_k = \arg\min_{\rho} J(\rho, \Theta^k)$$

Ideally, Θ should be exactly mapped by every ${}^qf^{-1}(\Xi)$

Intrinsic Graph Structure Estimation

Problems!

$$
\begin{tikzcd}
\Theta \arrow[rr, bend left, "\text{Exponential Map}"] & & \Xi \arrow[ll, bend left, "\text{Logarithmic Map}"]
\end{tikzcd}
$$

Many-to-one mapping

Q. Does logarithm exist?

Q. If yes, is it principal log?

Intrinsic Graph Structure Estimation

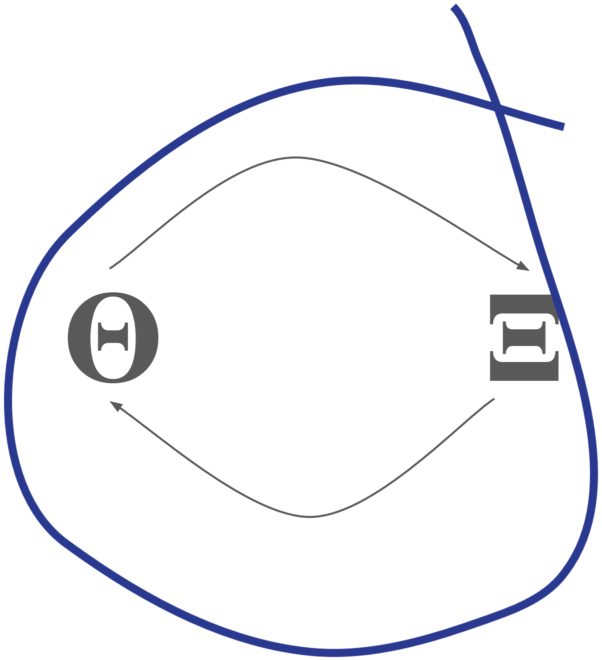
Intervention 1

$$
\begin{tikzcd}
\mathbf{\Theta} \arrow[rr, bend left, "\text{Polynomial Map}"] & & \mathbf{\Xi} \arrow[ll, bend left, "\text{Inverse Polynomial Map}"]
\end{tikzcd}
$$

Q. How to find the (unique) inverse of a polynomial of matrices?

Intrinsic Graph Structure Estimation

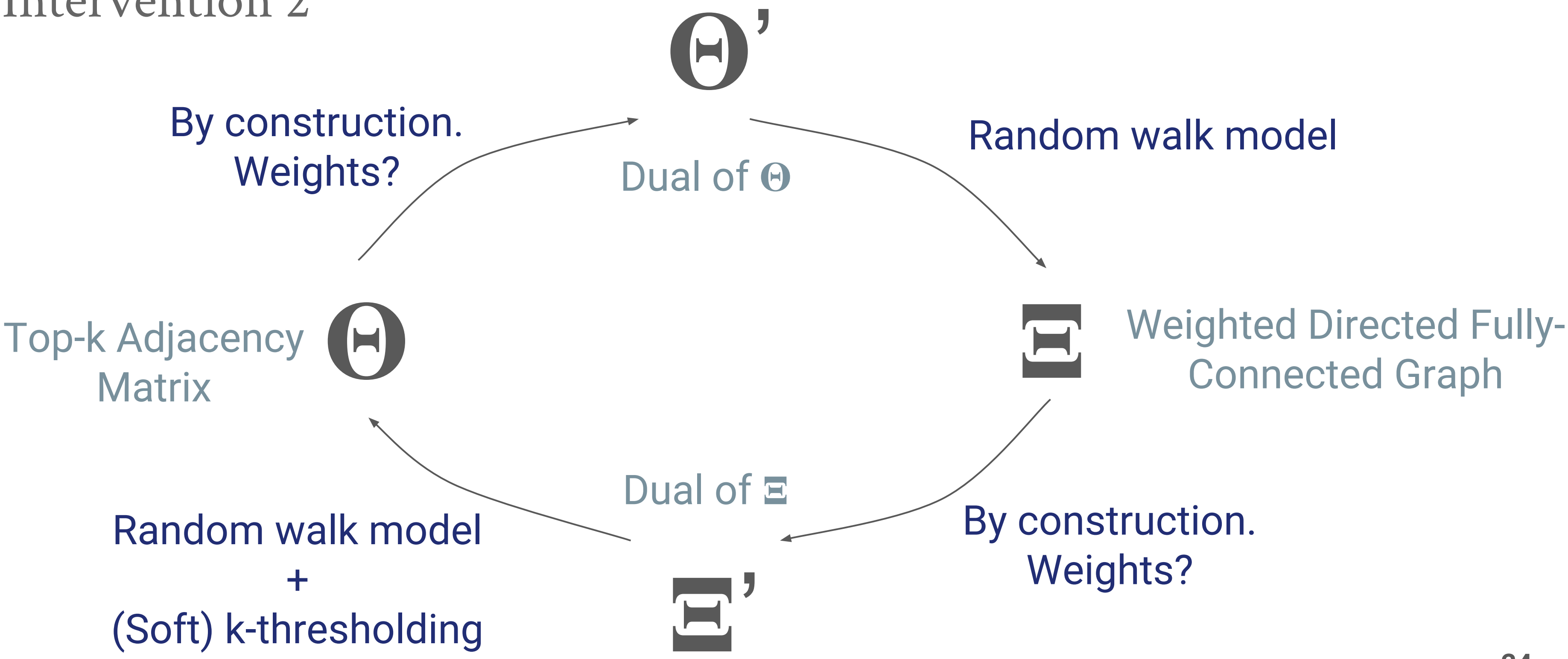
Intervention 2



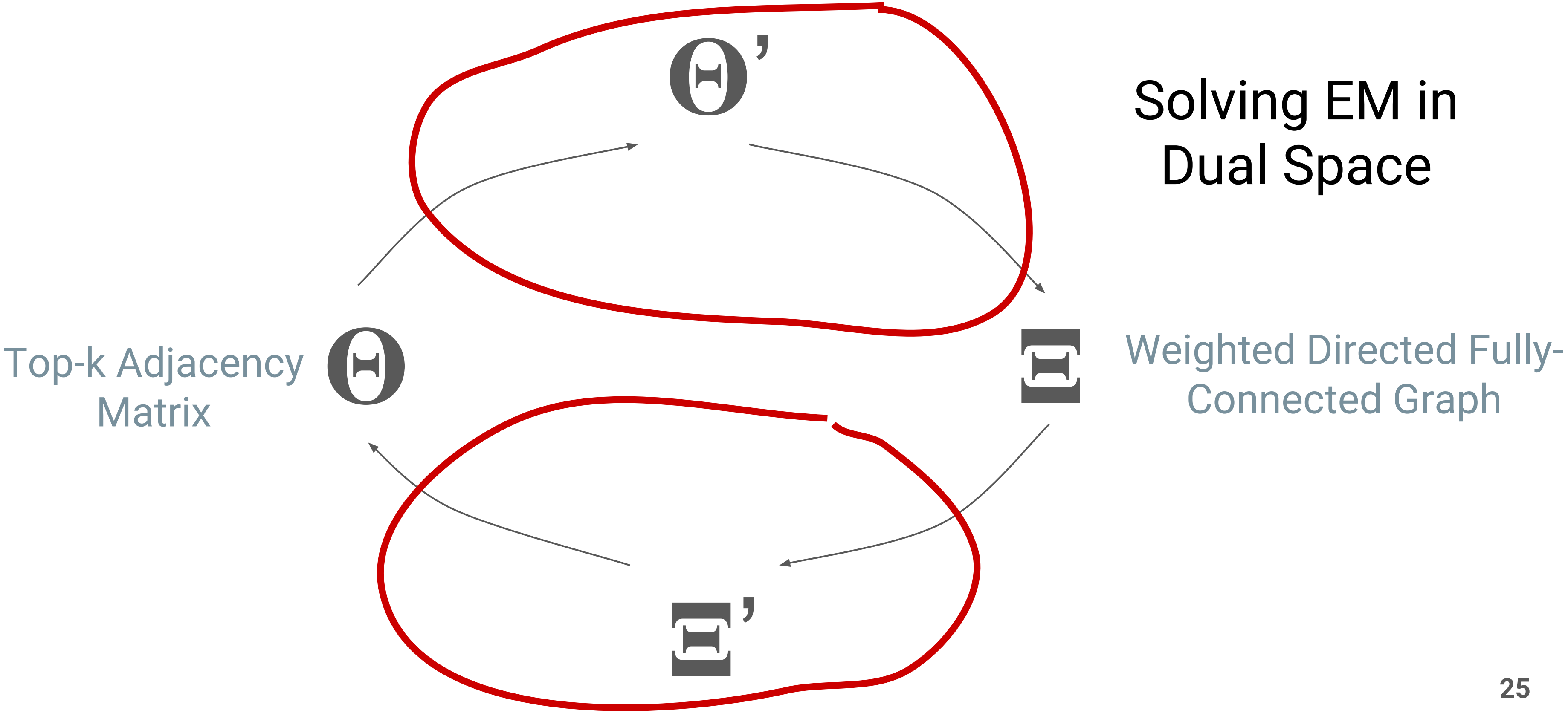
Solving EM in Primal Space

Intrinsic Graph Structure Estimation

Intervention 2



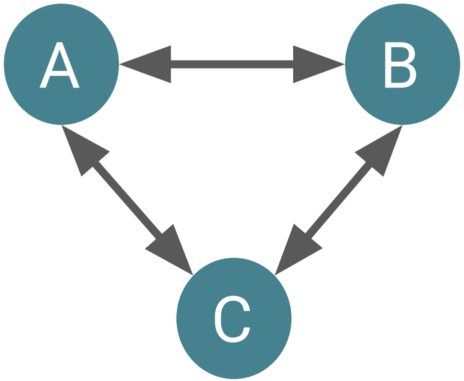
Pagerank on Dual for Graph Estimation

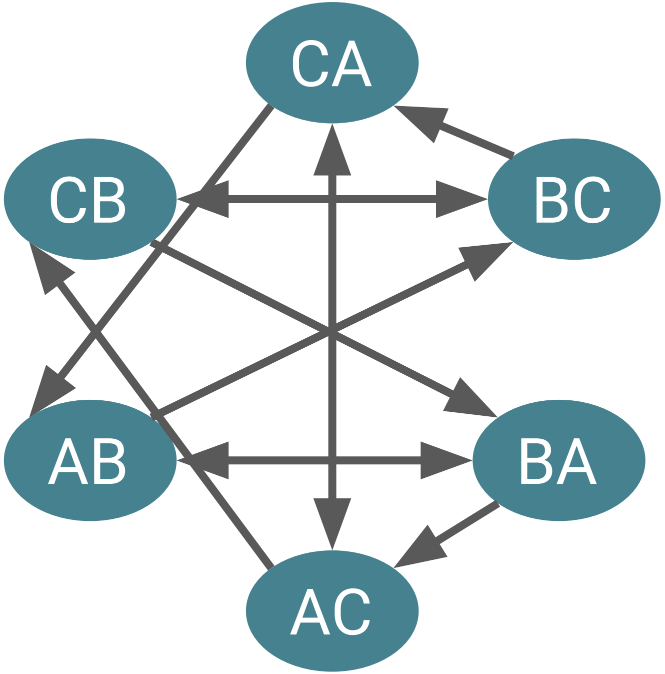


Solving EM in Dual Space

Random Walk on Dual for Graph Estimation

Dual Graph Construction



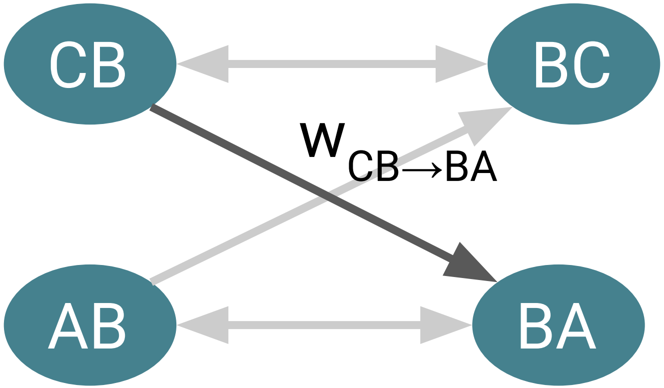


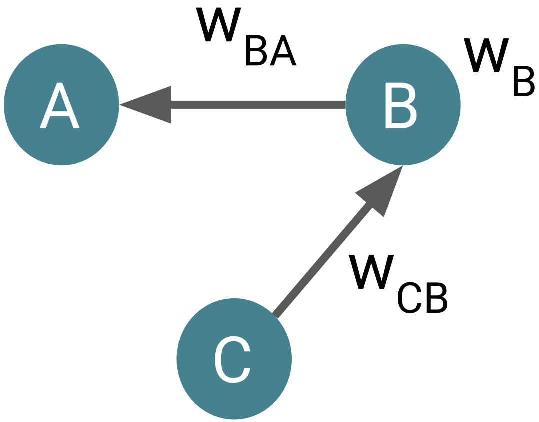
`n(n-1)` links for `n` nodes

`n(n-1)`^{2} metalinks for `n(n-1)` linknodes

Random Walk on Dual for Graph Estimation

Dual Graph Construction





$$w_{CB\to BA} = w_{CB} s_B w_{BA}$$

Random Walk on Dual for Graph Estimation

Use the Pagerank Random Walk Model

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

Pagerank on Dual for Graph Estimation

Empirical Validation on DREAM4

	CCM	Correlation	GC	MI	TE
Pairwise	0.55644	0.63378	0.72711	**0.54489**	**0.38844**
IGE	**0.47289**	**0.58756**	**0.60044**	**0.35022**	0.45244
Pagerank Dual	**0.64089**	**0.63733**	**0.728**	0.51111	**0.48533**

(There's still some) Future Work

- Some mysteries for Pairwise Metrics
- Mathematical validation of graph estimation
    - Why are the "important" links the "actual" links of causality?
- Essentially we're doing a clustering of edges
    - Can we fit this in a regular clustering paradigm?
- Biological information still hasn't been used!

Thank You

Questions and Feedback