International Seminar and Workshop
Causality, Information Transfer and Dynamic Networks
Max Planck Institute for the Physics of Complex Systems

Causal Network Inference by Optimal Causation Entropy

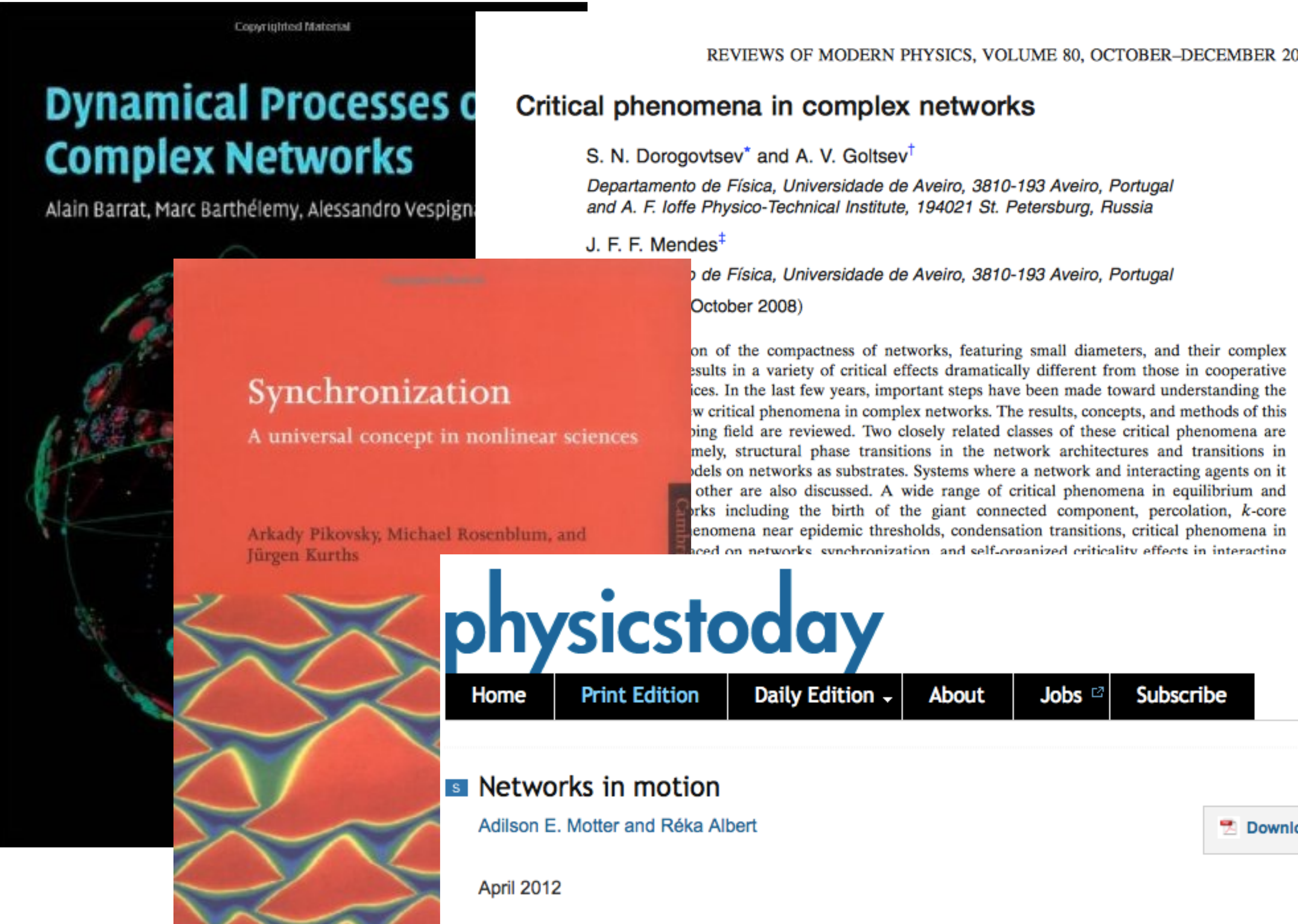
Jie Sun
Department of Mathematics & Computer Science
Clarkson University

Collaborators
Erik Bollt (Clarkson), Carlo Cafaro (Clarkson), and
Dane Taylor (University of North Carolina)

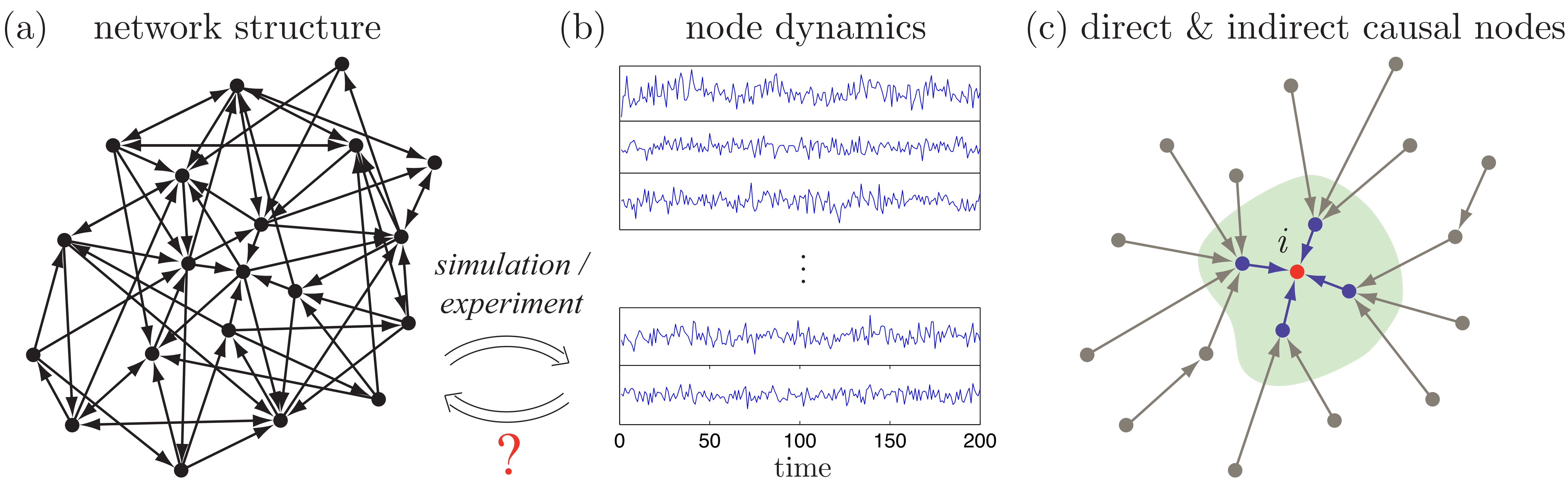
Acknowledgements
Organizers: Jorn Davidsen, Peter Grassberger, Bjorn Schelter, Claudia Ponisch

June 16, 2014 (Dresden, Germany)

Dynamics on Networks



The Problem of Causal Network Inference



Given: (time series) observations at each component of a system

Goal: identify the *direct* causal relationships between the components

…there are several reasons…

medical diagnosis: identify the causes of a disease in order to suggest effective treatments

Procedure of Causal Network Inference

Gather a sufficient amount of relevant data (experimental work)

$\Updownarrow$

Develop an appropriate causal inference measure (theoretical work)

$\Updownarrow$

Accurate and reliable estimate of such a measure (computational/statistical work)

A good causal inference measure should satisfy

…general applicability and neat interpretation...

…immune (in principle) to false positives and false negatives…

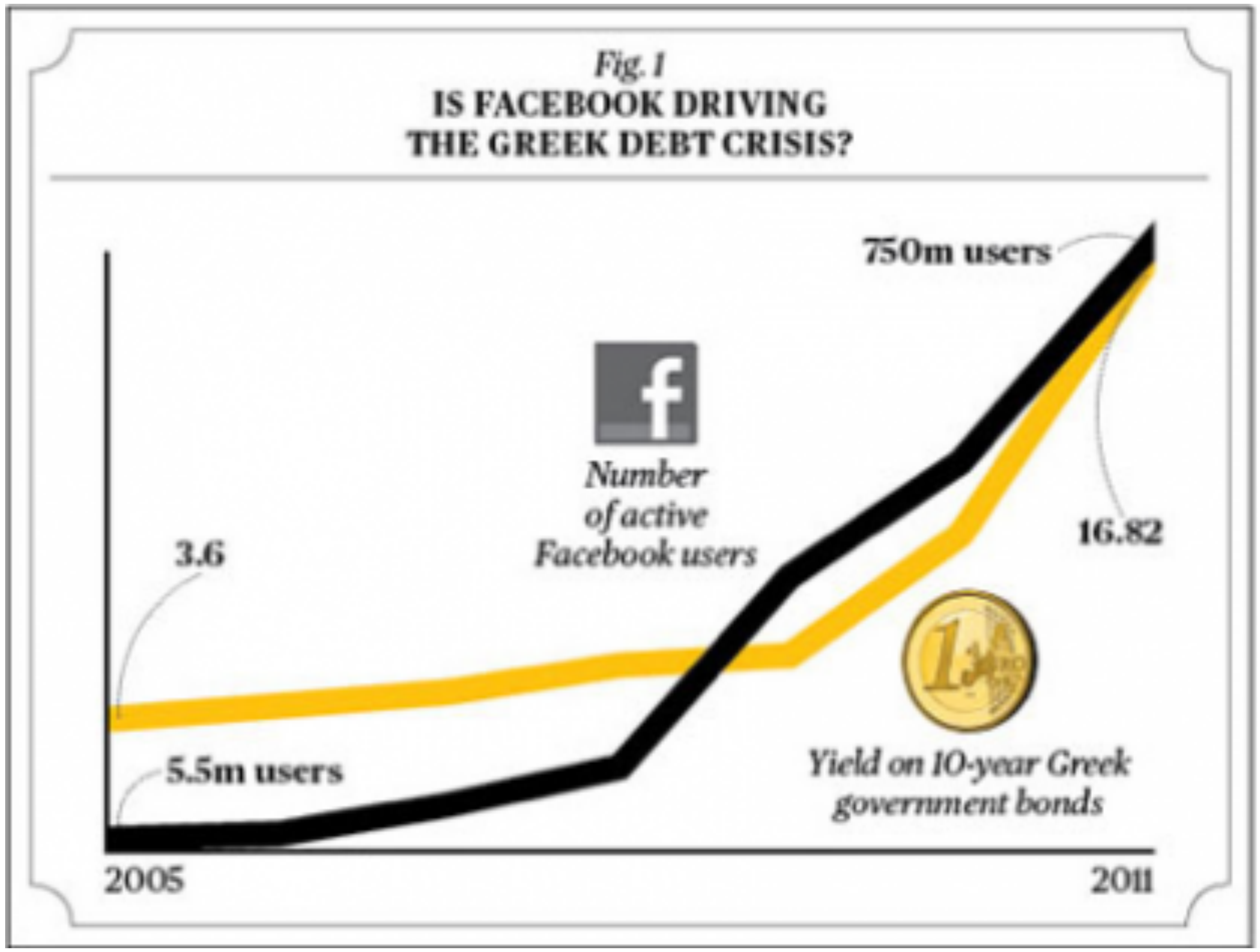
…accurate and fast numerical estimation…

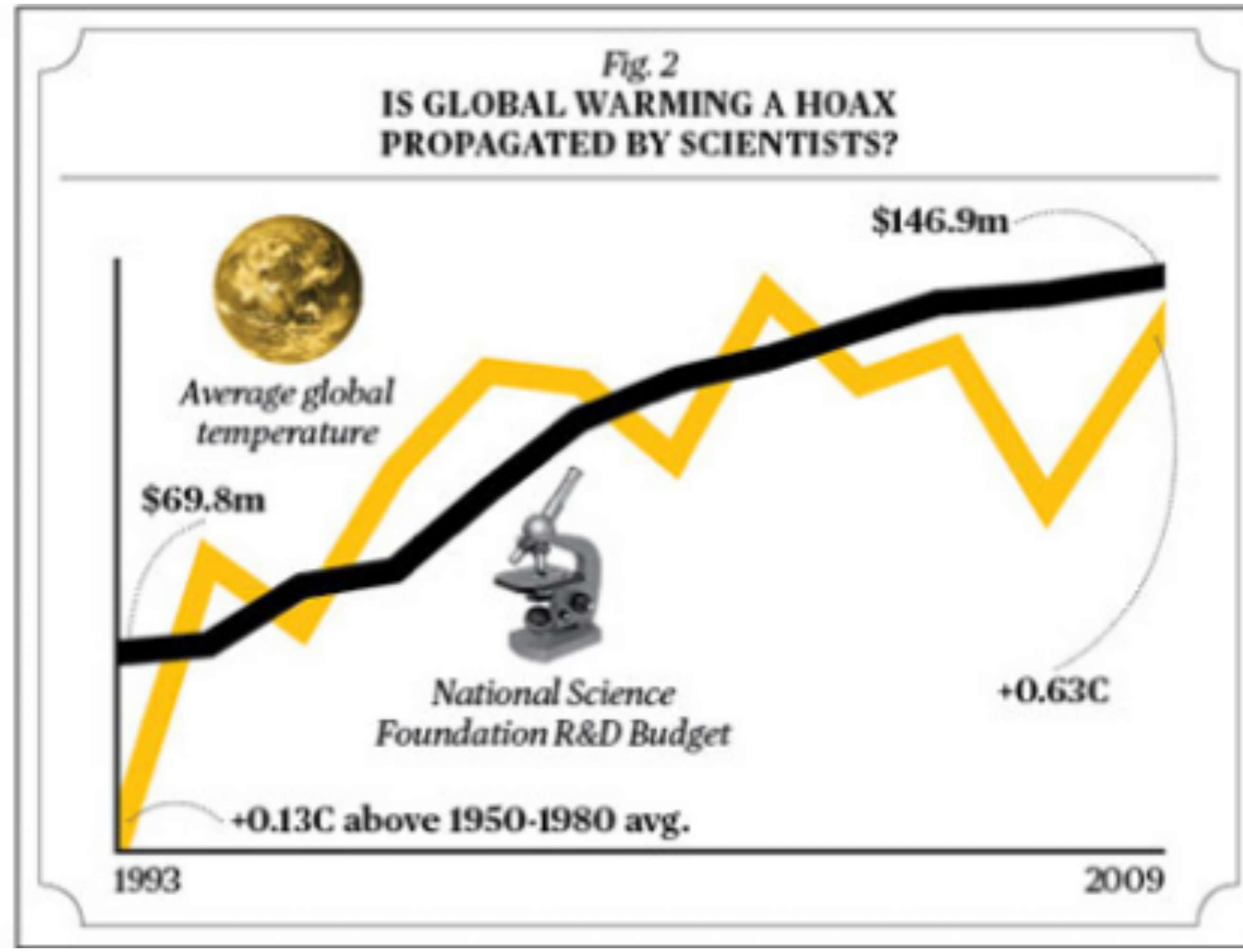
Condition (1) →…linear and nonlinear interactions…

Condition (2) →…correct identification of direct couplings in complex systems with more than two components…

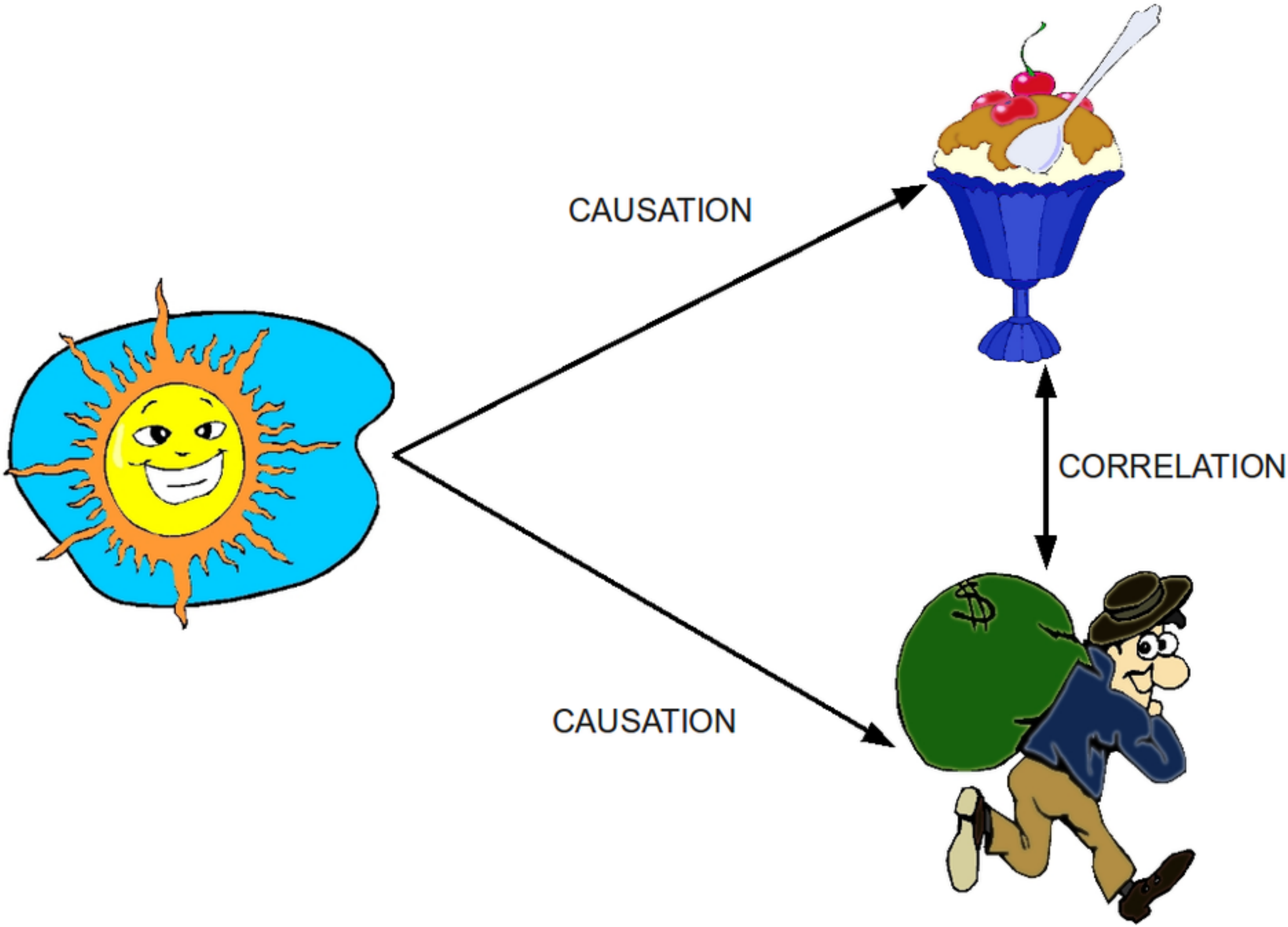
Condition (3) →…appropriate statistical estimation techniques…

Correlation vs. Causality





Correlation does not imply causality.


CAUSATION
CORRELATION
CAUSATION

Information vs. Physical Causality

Paul the Octopus

paul the octopus

Paul in his

Other appellation

Species

Sex

Occupation

Known for

rch tools

nk at a

Results involving Germany [edit]

Opponent	Tournament	Stage	Date	Prediction	Result	Outcome
Poland	Euro 2008	group stage	8 June 2008	Germany	2–0	Correct
Croatia	Euro 2008	group stage	12 June 2008	Germany[2][21]	1–2	Incorrect
Austria	Euro 2008	group stage	16 June 2008	Germany	1–0	Correct
Portugal	Euro 2008	quarter-finals	19 June 2008	Germany	3–2	Correct
Turkey	Euro 2008	semi-finals	25 June 2008	Germany	3–2	Correct
Spain	Euro 2008	final	29 June 2008	Germany[2]	0–1	Incorrect
Australia	World Cup 2010	group stage	13 June 2010	Germany[30]	4–0	Correct
Serbia	World Cup 2010	group stage	18 June 2010	Serbia[30]	0–1	Correct
Ghana	World Cup 2010	group stage	23 June 2010	Germany[30]	1–0	Correct
England	World Cup 2010	round of 16	27 June 2010	Germany[31]	4–1	Correct
Argentina	World Cup 2010	quarter-finals	3 July 2010	Germany[24]	4–0	Correct
Spain	World Cup 2010	semi-finals	7 July 2010	Spain[32]	0–1	Correct
Uruguay	World Cup 2010	3rd place play-off	10 July 2010	Germany	3–2	Correct

Argentine chef Nicolas Bedorrou was so angry after Paul correctly predicted his team would lose its quarter-final clash with **Germany** that he suggested a way to cook the octopus.

2 days ago - Lucknow. The football world, rendered directionless by the demise of soothsayer **Paul the Octopus**, has found its order back again. And in the ...

Mathematical Assumptions

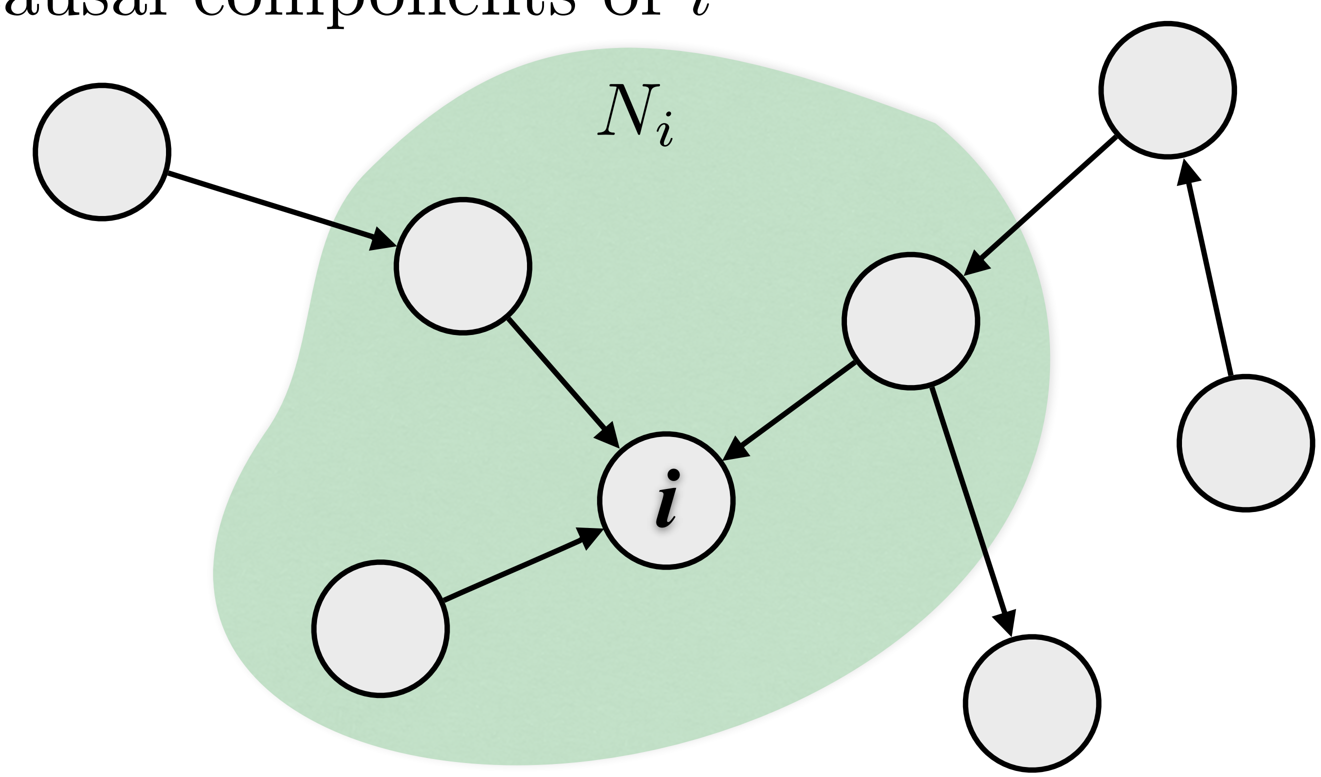
The necessity of making assumptions:

M. Eichler, Graphical modeling of multivariate time series, Probability Theory and Related Fields (2012).

Stochastic process $\{X_t^{(i)}\}_{i=1,2,\ldots,N;t=1,2,\ldots}$

$$\begin{cases} \text{(i) Temporally Markov:} \\ \qquad p(X_t|X_{t-1},X_{t-2},\ldots) = p(X_t|X_{t-1}) = p(X_{t'}|X_{t'-1}) \text{ for any } t \text{ and } t'. \\ \text{(ii) Spatially Markov:} \\ \qquad p(X_t^{(i)}|X_{t-1}) = p(X_t^{(i)}|X_{t-1}^{(N_i)}) \text{ for any } i. \\ \text{(iii) Identifiability:} \\ \qquad p(X_t^{(i)}|X_{t-1}^{(K)}) \neq p(X_t^{(i)}|X_{t-1}^{(L)}) \text{ whenever } (K\cap N_i) \neq (L\cap N_i). \end{cases}$$

N_i : set of direct causal components of i



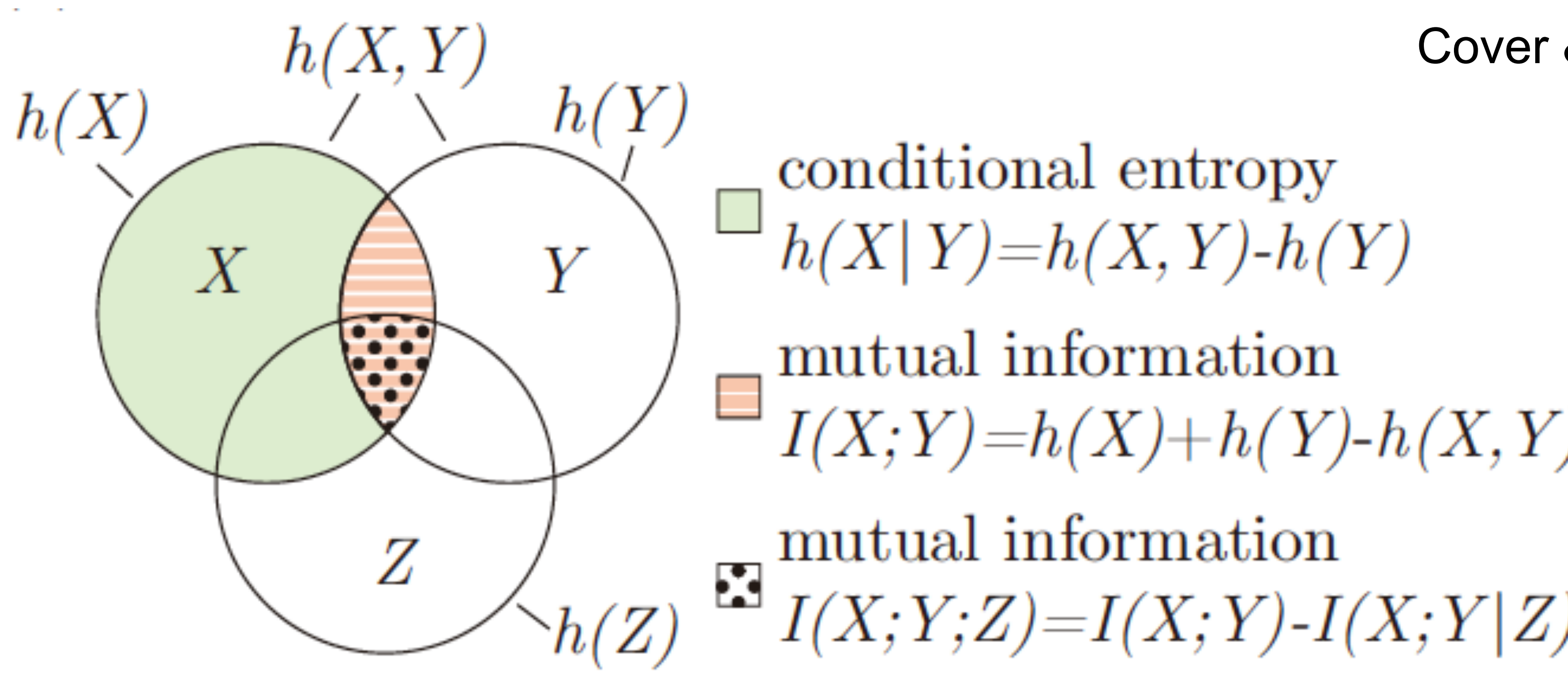
(also assumes full observability of all the components)

Basic Information-theoretic Measures

$$\text{Entropy } h(X) = -\int p(x)\log p(x)dx$$

$$\begin{cases} \text{Joint entropy: } h(X,Y) \equiv h(Y,X) \equiv -\int p(x,y)\log p(x,y)dxdy. \\ \text{Conditional entropies: } \begin{cases} h(X|Y) \equiv -\int p(x,y)\log p(x|y)dxdy; \\ h(Y|X) \equiv -\int p(x,y)\log p(y|x)dxdy. \end{cases} \end{cases}$$

$$\begin{cases} \text{Mutual information: } I(X;Y) \equiv h(X) - h(X|Y) \equiv h(Y) - h(Y|X). \\ \text{Conditional mutual information:} \\ \qquad I(X;Y|Z) \equiv h(X|Z) - h(X|Y,Z) \equiv h(Y|Z) - h(Y|X,Z). \end{cases}$$



Cover & Thomas (2006).

Transfer Entropy, Self-Causality, and Conditioning

Transfer Entropy (TE)

$$T_{Y\to X} \equiv h(X_{t+1}|\mathbf{X}_t) - h(X_{t+1}|\mathbf{X}_t, \mathbf{Y}_t)$$

future of X   past of X   past of Y

$T_{Y\to X}$ measures the reduction of uncertainty about X_{t+1} given knowledge about $\mathbf{Y}_t$ in addition to that of $\mathbf{X}_t$.

T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000)
A. Kaiser & T. Schreiber, Physica D (2002)
M. Palus *et. al.*, Phys. Rev. E (2001)

Remarks:

1. $T_{X\to X} \equiv 0$, TE makes no indication about whether the process $\{X_t\}$ is self-causal or not, and is not designed to address such question.

2. $T_{Y\to X}$ is a bivariate measure (from Y to X), and is therefore not designed to infer direct causality within multiple processes.

Self-causal or not affects system controllability.

Cowan et.al., PLOS ONE (2012)

When there are multiple processes, $\mathbf{Y}_t$ directly causes X_{t+1} only if such "cause" remains after the removal of all other conditions.

Frenzel and Pompe, Phys. Rev. Lett. (2007)
Vejmelka and Palus, Phys. Rev. E (2008)

Benjamin Franklin and Clive Granger

Benjamin Franklin

For want of a nail a shoe was lost,
For want of a shoe a horse was lost,
For want of a horse a battle was lost

and

For want of a battle a kingdom was lost.
And all for the want of a horseshoe nail.

Clive Granger

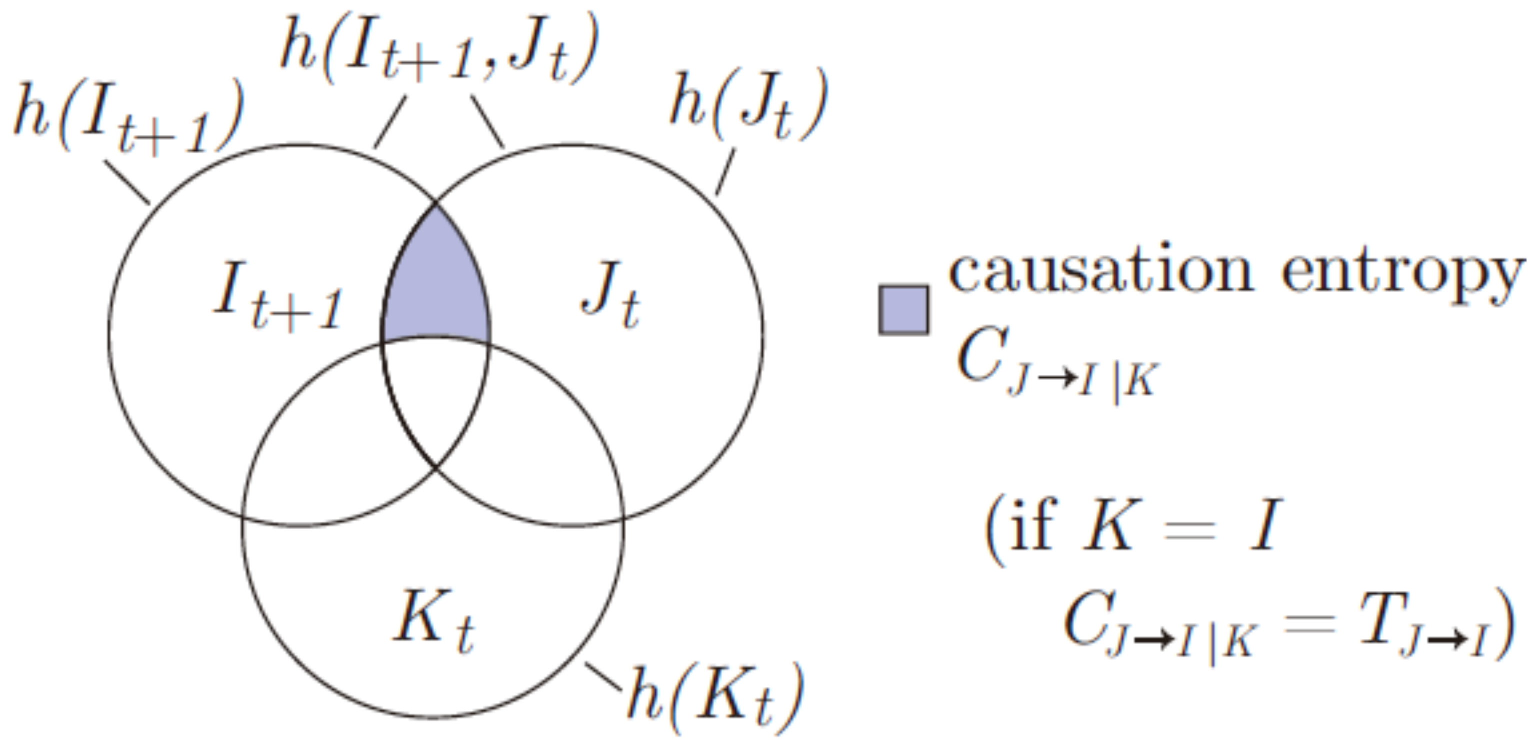
Conclusion - The blacksmith destroyed the kingdom.
– everything is connected to everything
– everything causes (and is caused by) everything
….yes….but not so useful.

Granger's notion of causality: *(Granger, 1969)*
(1) The cause should occur before the effect (caused)
(2) The causal process should carry information—unavailable in other processes—about the effect.

Causation Entropy

The **Causation Entropy (CSE)** from the set of components ***J*** to ***I*** conditioning on ***K*** is defined (explicitly) as:

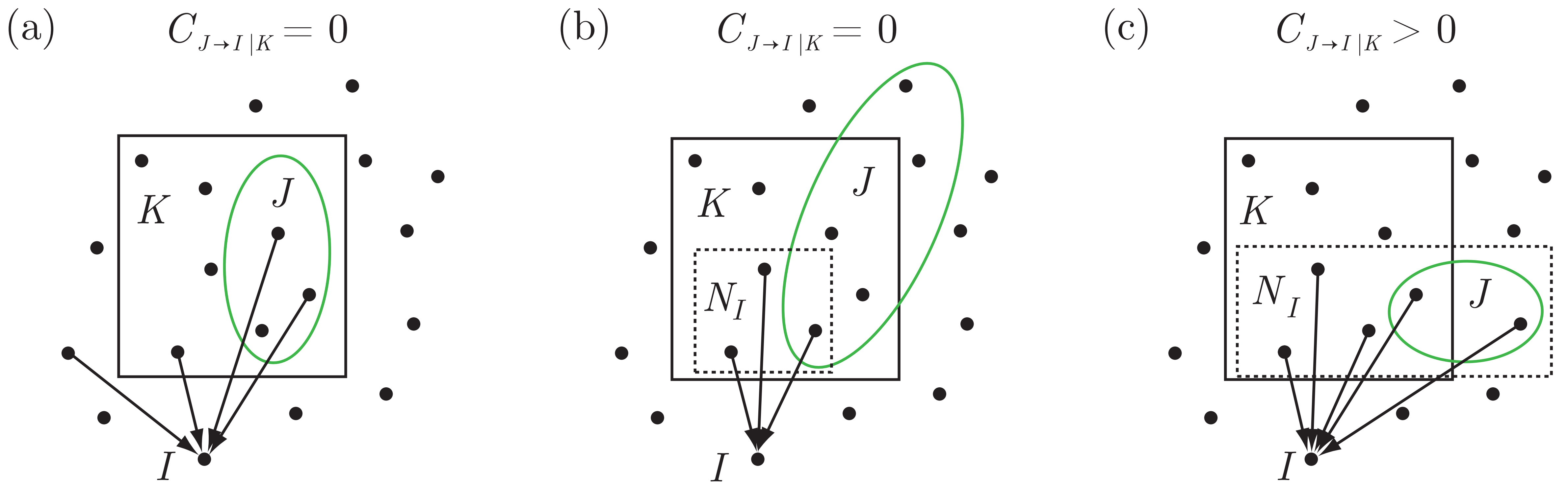
$$C_{J \to I|K} = h(X_{t+1}^{(I)}|X_t^{(K)}) - h(X_{t+1}^{(I)}|X_t^{(K)}, X_t^{(J)})$$



Remarks:

1. CSE can be used to assess whether a process is “self-causal”.
2. CSE does not “solve” the causal inference problem.
3. The definition simply emphasizes the fact that cause-and-effect is not a bivariate question, but rather, involves all three parts (***cause***, ***effect***, and ***conditioning***).

Analytical Properties of CSE



(a) (Redundancy) If $J \subset K$, then $C_{J\to I|K} = 0$.
(b) (No false positive) If $N_I \subset K$, then $C_{J\to I|K} = 0$ for any set of nodes J.
(c) (True positive) If $J \subset N_I$ and $J \not\subset K$, then $C_{J\to I|K} > 0$.
(d) (Decomposition) $C_{J\to I|K} = C_{(K\cup J)\to I} - C_{K\to I}$.

(Optimal Causation Entropy Principle) The set of direct causal neighbors is the minimal set of nodes with maximal Causation Entropy.
Define the family of sets with maximal Causation Entropy as

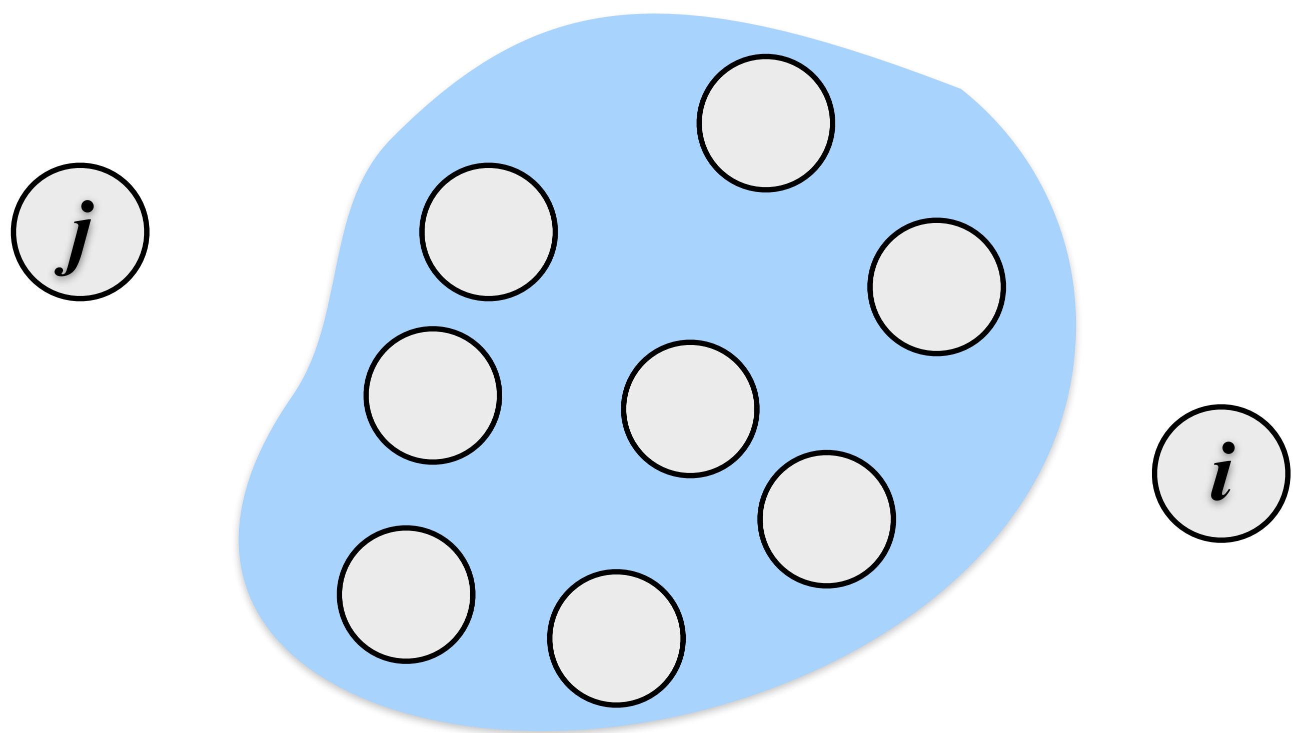
$$\mathcal{K} = \{K | \forall K' \subset \mathcal{V}, C_{K'\to I} \leq C_{K\to I}\}. \tag{2.26}$$

Then the set of direct causal neighbors satisfies

$$N_I = \cap_{K\in\mathcal{K}} K = \mathrm{argmin}_{K\in\mathcal{K}} K. \tag{2.27}$$

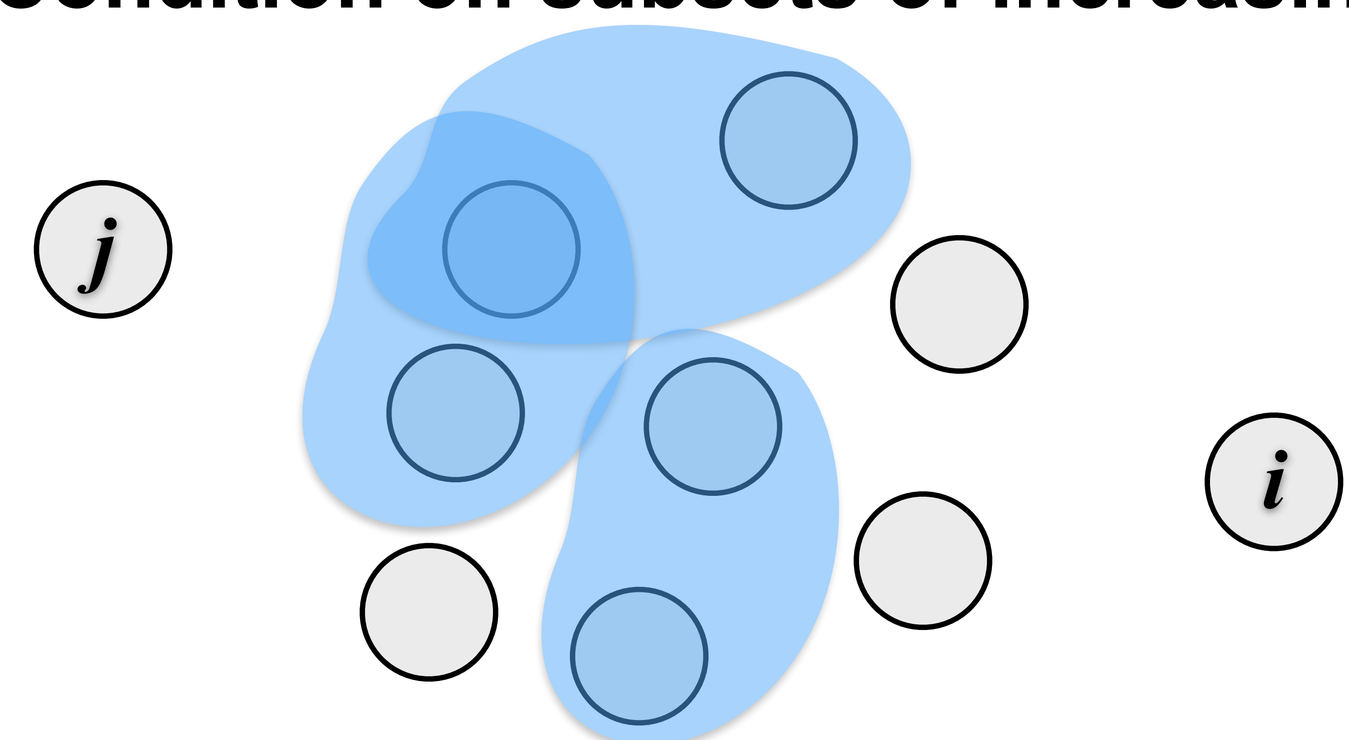
Existing Approaches of Conditioning

1. Condition on everything:



— curse of dimensionality (estimation in the full-dimensional space)

2. Condition on subsets of increasing cardinality (PC-algorithm)

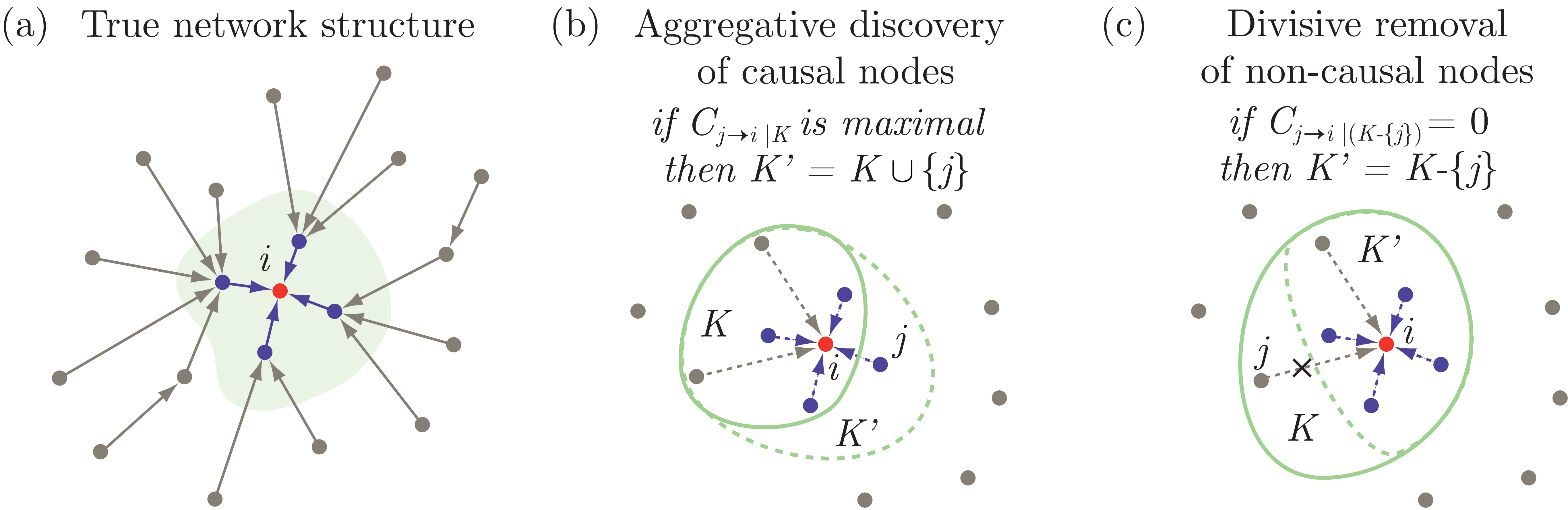


— combinatorial search (high computational burden)

$$\frac{n^2(n-1)^{k-1}}{(k-1)!}$$

Spirtes, Glymour, Scheines (2000). Runge, Heitzig, Petoukhov, Kurths, PRL (2012)

Forward & Backward Conditioning: Discover & Remove



Remarks:

1. The "forward" step (Aggregative Discovery) alone in principle would result in false positives that cannot be mitigated by the increased amount of data. Vlachos, Kugiumtzis, Phys. Rev. E (2010); Kugiumtzis, Phys. Rev. E (2013)

2. The "backward" step (Divisive removal) can be modified to one that is analogous to the PC-algorithm in the statistical inference literature, with the key difference that here enumeration of conditioning subsets needs to be performed only within K, rather than for all nodes.

Spirtes, Glymour, Scheines (2000). Runge, Heitzig, Petoukhov, Kurths, PRL (2012)

Benchmark Systems

$$X_t^{(i)} = \sum_{j \in N_i} A_{ij} X_{t-1}^{(j)} + \xi_t^{(i)}$$

-- *Classical communication channel*
-- *Fluctuations around equilibrium*

stable adjacency matrix

uncorrelated Gaussian noise

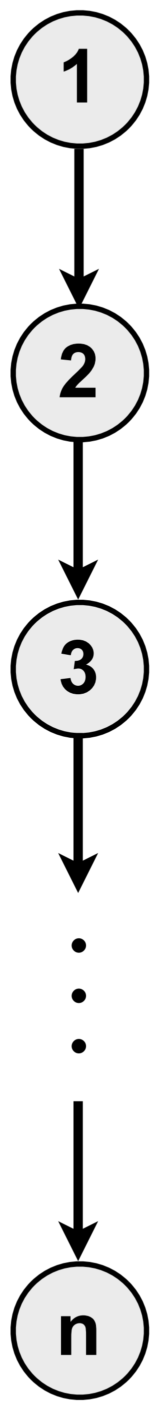
Ahmed and Gokhale, IEEE Trans. Info. Theor. (1989).
Barnett, Phys. Rev. Lett. (2009).

$X_t^{(i)}$ are Gaussian, with covariance $\Phi(\tau, t)_{ij} \equiv \mathrm{cov}[x_{t+\tau}^{(i)}, x_t^{(j)}]$

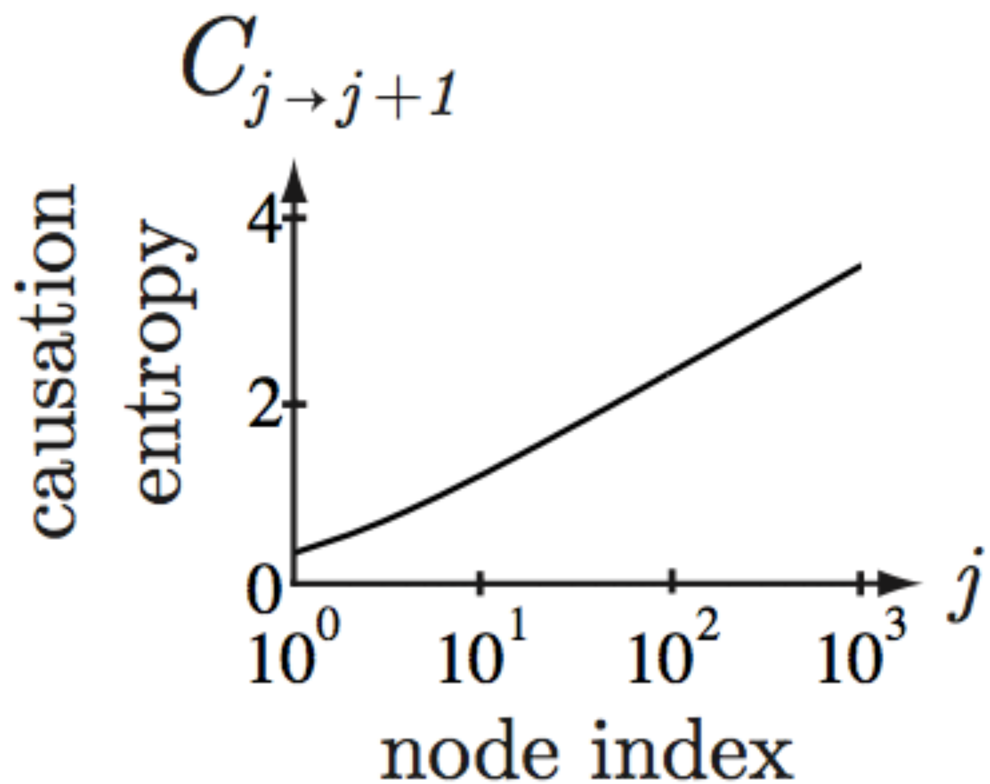
$$\begin{cases} \Phi(0) \equiv \lim_{t\to\infty} \Phi(0,t) = \sum_{k=0}^{\infty} A^k S (A^k)^\top \\ \Phi(\tau) = A\Phi(\tau-1) = A^2\Phi(\tau-2) = \cdots = A^\tau \Phi(0) \end{cases}$$

$$C_{J\to I|K} = \frac{1}{2} \log \left(\frac{\det\left[\Phi(0)_{II} - \Phi(1)_{IK}\Phi(0)_{KK}^{-1}\Phi(1)_{IK}^\top\right]}{\det\left[\Phi(0)_{II} - \Phi(1)_{I,K\cup J}\Phi(0)_{K\cup J,K\cup J}^{-1}\Phi(1)_{I,K\cup J}^\top\right]} \right)$$

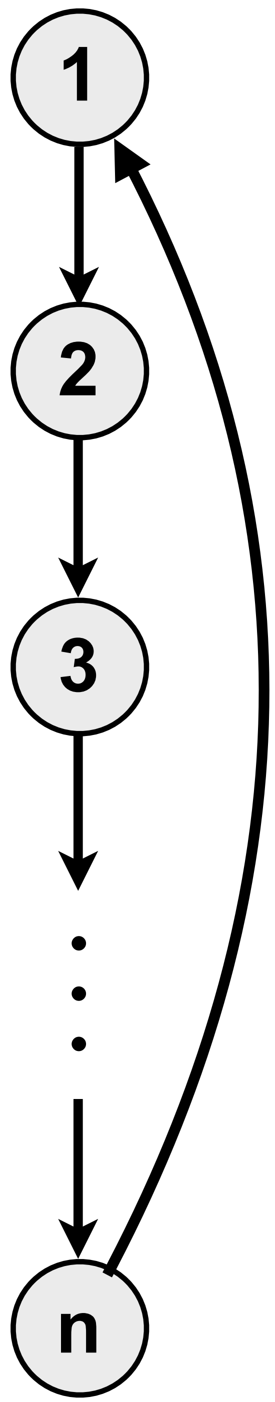
Analytical Calculation: Directed Linear Chain



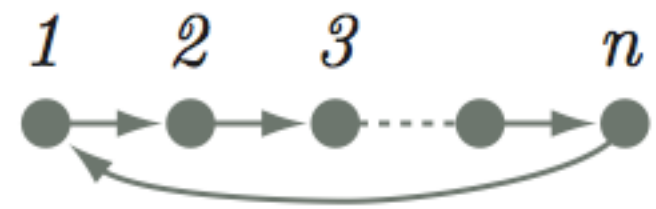
$$C_{j\to i} = \frac{1}{2}\delta_{i,j+1}\log\left(1+\frac{\sum_{k=1}^{j}\sigma_k^2}{\sigma_i^2}\right)$$

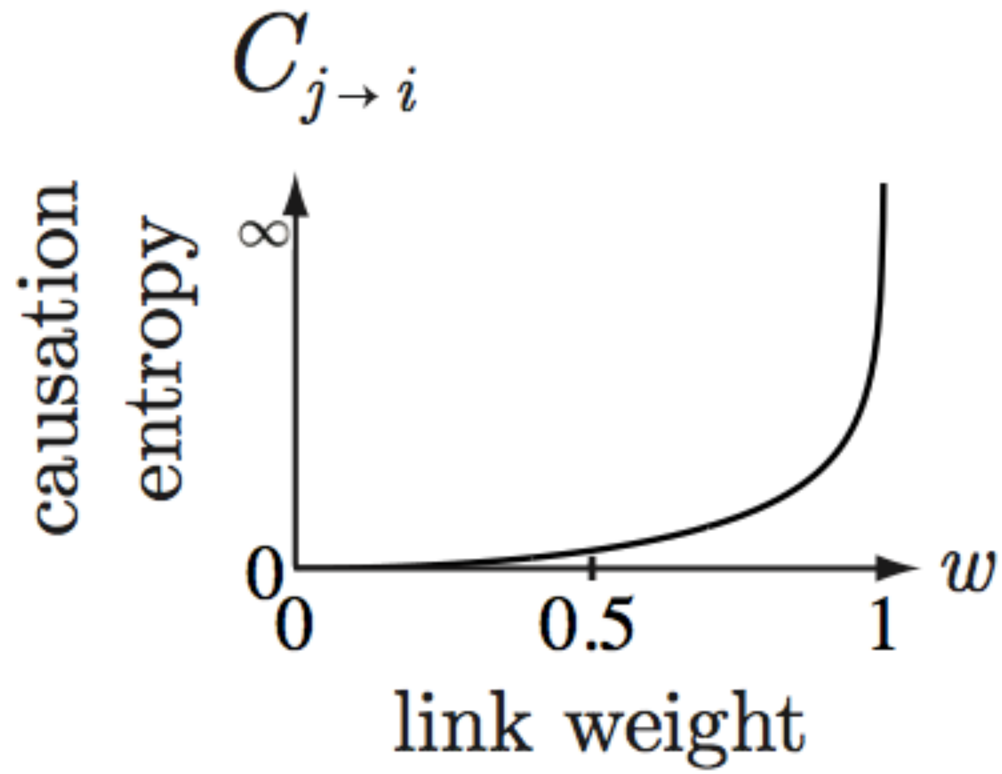


Analytical Calculation: Directed Loop

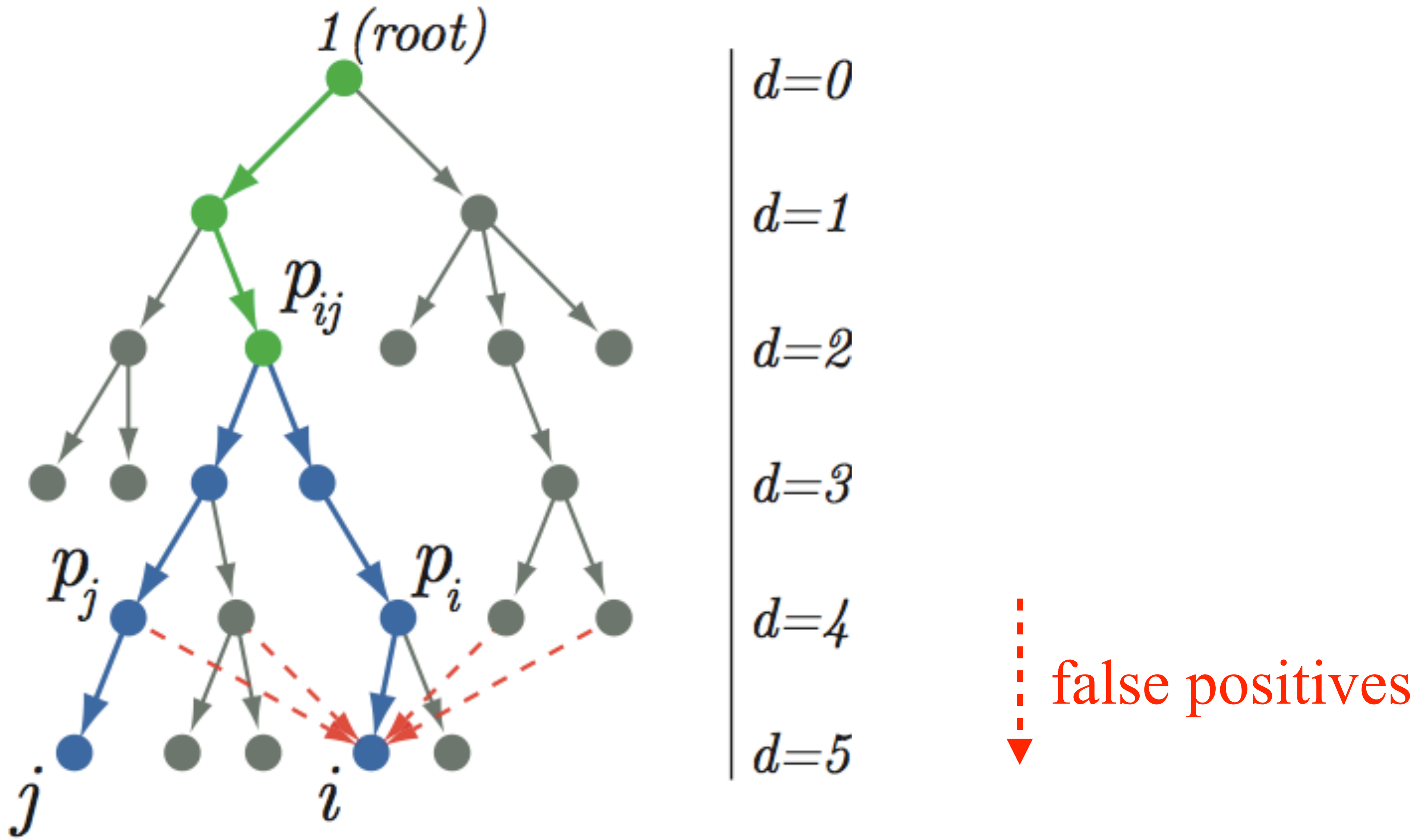


$$C_{j\to i} = \frac{1}{2}\delta_{p_i,j}\log\left(\frac{1}{1-w^2}\right)$$





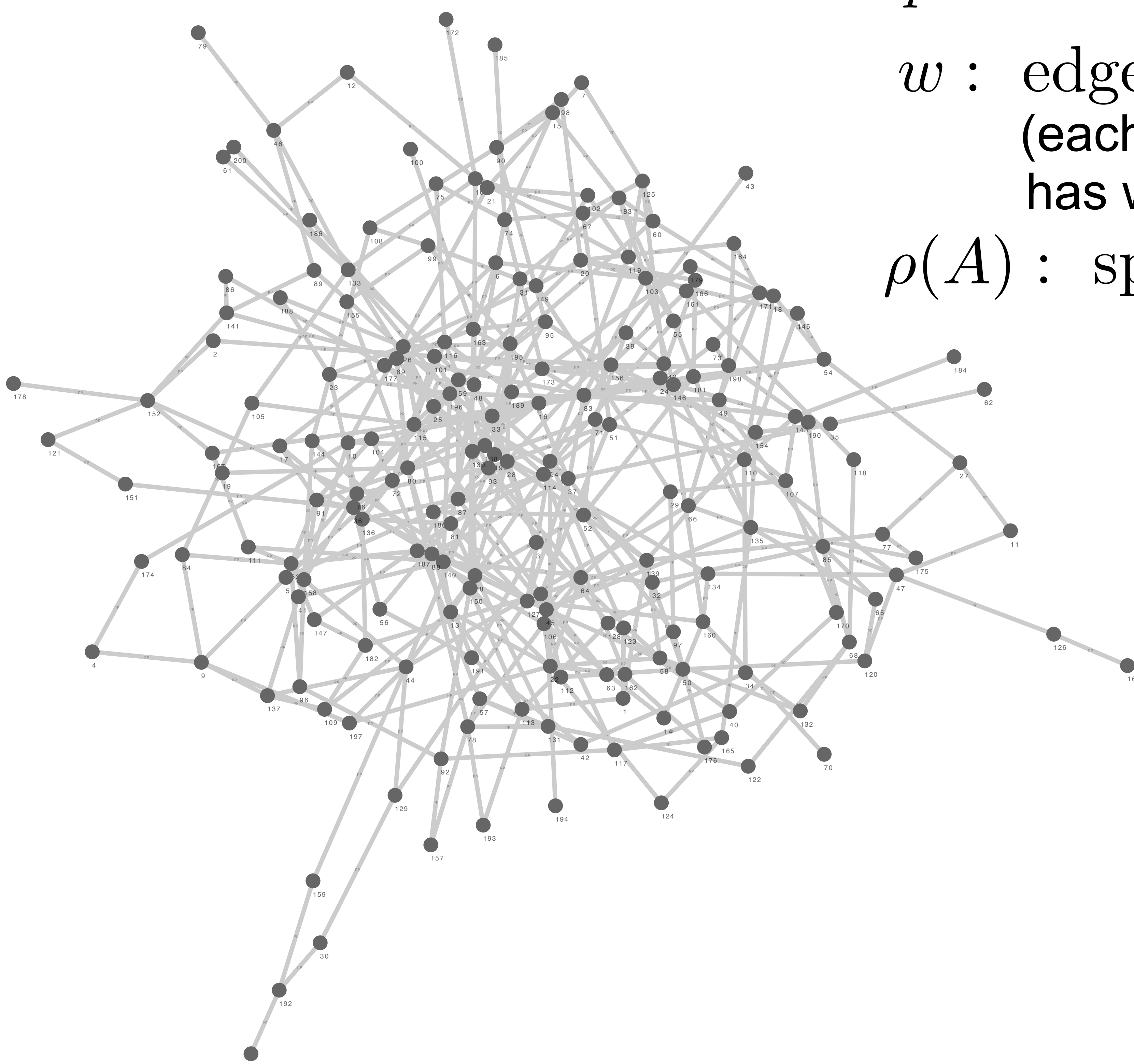
Analytical Calculation: Directed Tree



No conditioning: $$C_{j\to i} = \frac{1}{2}\delta_{d_i,d_j+1}\log\frac{(d_i+1)(d_j+1)}{(d_i+1)(d_j+1)-(d_{p_{ij}}+1)^2}$$

With conditioning: $$\begin{cases} p_i = \arg\max_j C_{j\to i}, \\ C_{j\to i|\{p_i\}} = 0. \end{cases}$$

Numerical Tests for Random Directed Networks



n : number of nodes

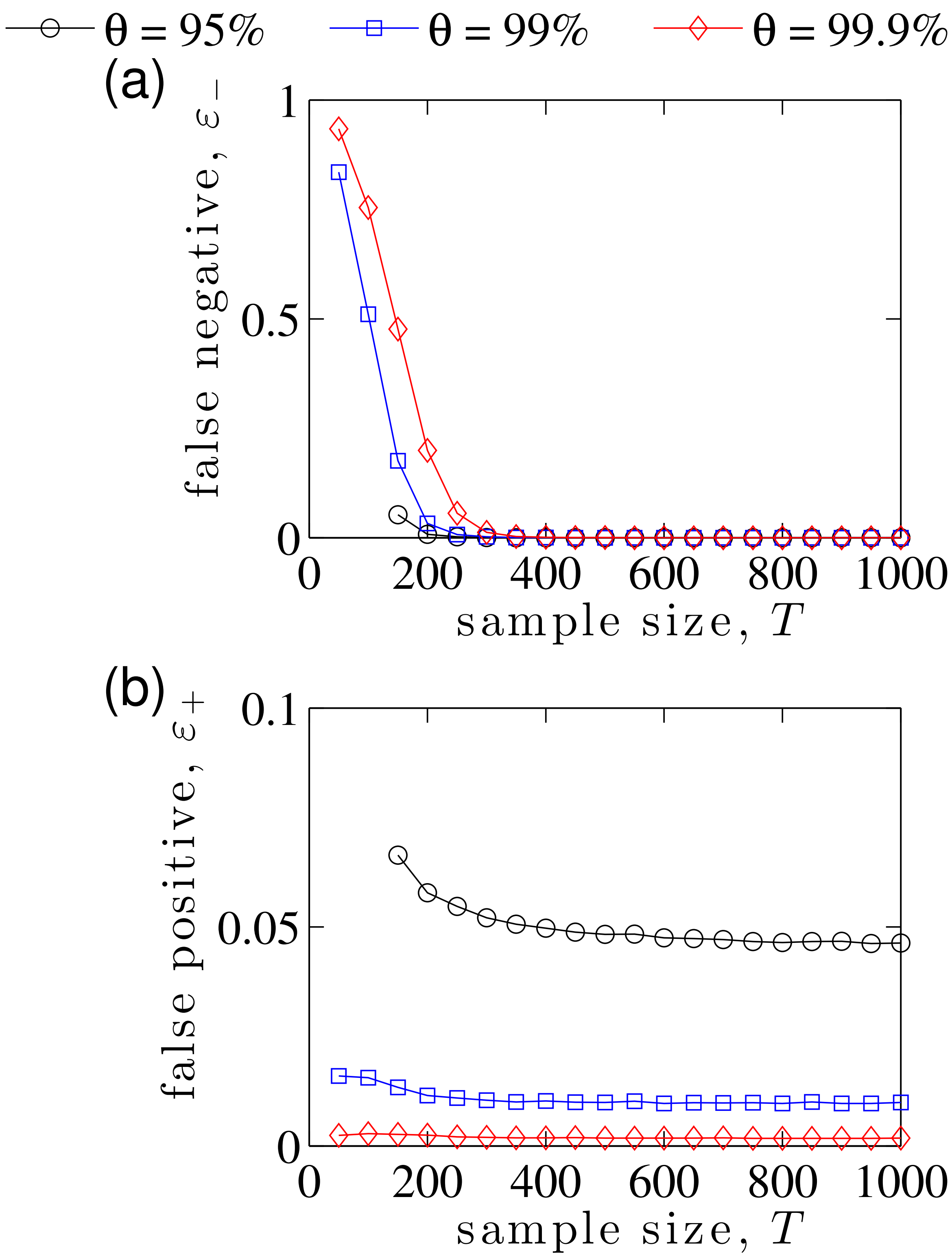
p : connection of probability

w : edge weights
(each directed edge
has weight *w* or *-w*)

$\rho(A)$: spectral radius

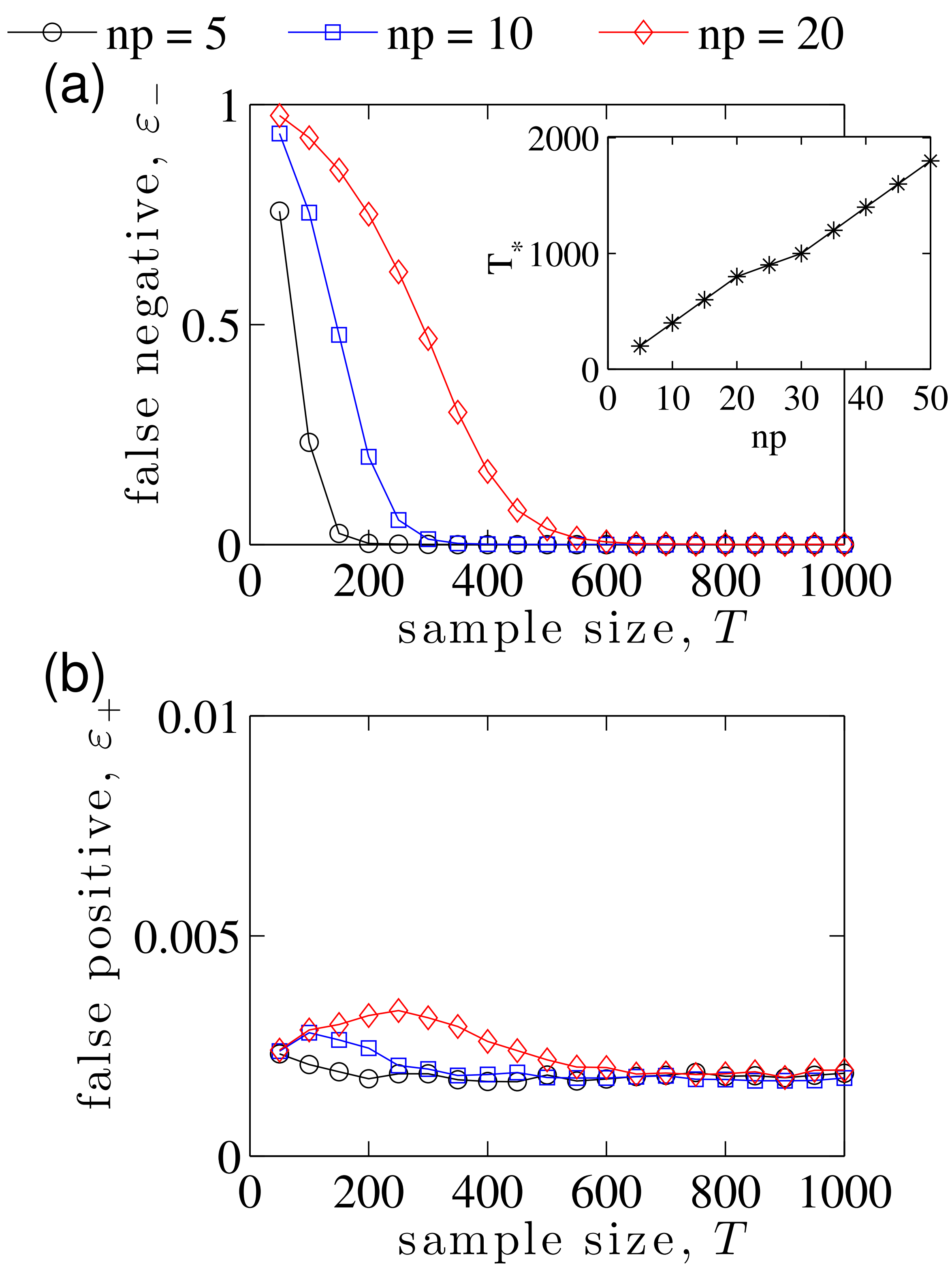
Significance Level and Network Size

$$n = 200, np = 10, \rho(A) = 0.8$$



Effects of Sample Size and Spectral Radius

$$n = 200, \rho(A) = 0.8$$



Conclusions

— oCSE allows for direct causal inference in an efficient way (it can be extended to finite-order Markov process and also process with infinite but fading memory)
— Tradeoff between computation (choosing the conditioning set) and estimation (resulting dimensionality of the random variables)

Challenges

— Latent variables M. Eichler, Journal of Machine Learning Research (2010)
— Non-stationarity Bai and Perron, Journal of Applied Econometrics (2003)
— Reliable estimation Kraskov and Grassberger, PRE (2004); Frenzel and Pompe, PRL (2007)
— Exact statistical test Pethel and Hahs, Entropy (2014, to appear)

Acknowledgements

This work is funded by ARO Grant No. 61386-EG. We thank Dr Samuel Stanton (ARO Complex Dynamics and Systems Program) for his continuous and ongoing support.

References

JS and E.M. Bollt, Physica **D267**, 49 (2014)
JS, D. Taylor, and E.M. Bollt, arXiv:cs.IT/1401.7574 (2014)
JS, C. Cafaro, and E.M. Bollt, Entropy (to appear, 2014)

Thank you!