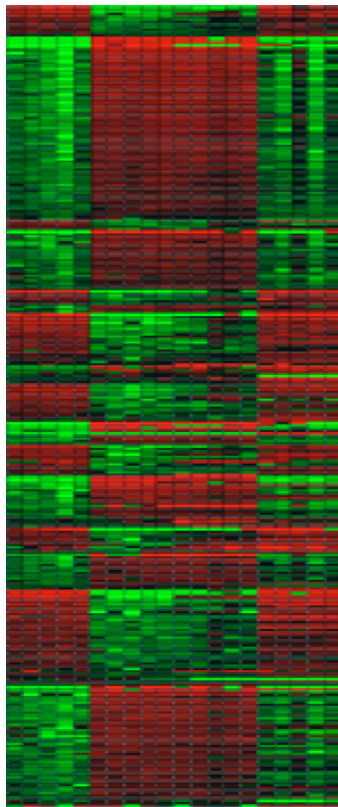


Modeling genetic regulation at different levels: framework, algorithms, applications

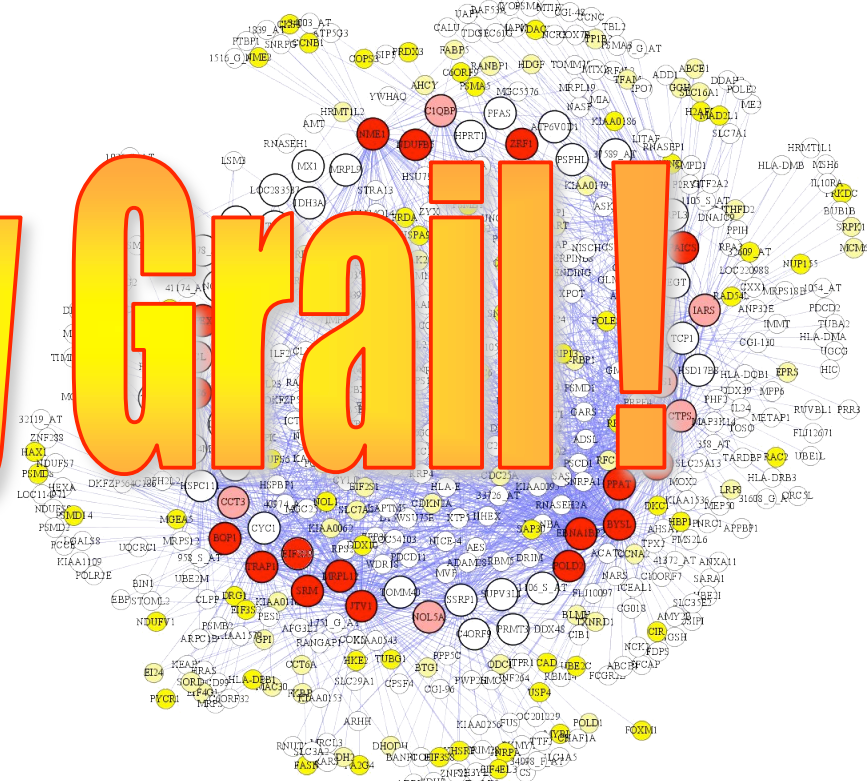


Ilya Nemenman
(JCSB, Columbia)

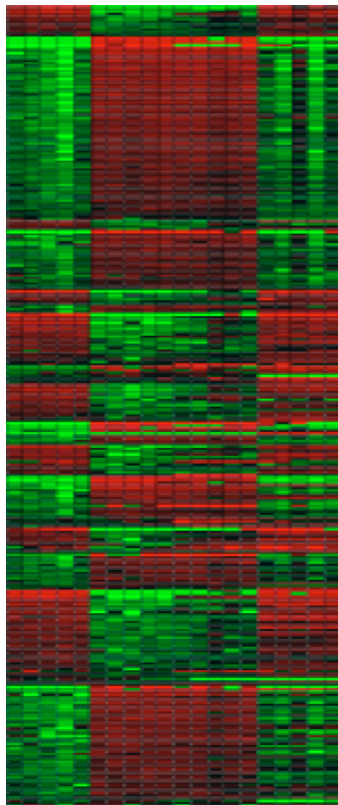
Reconstructing interaction models



Holy Grail!



Reconstruction algorithms: Arms race

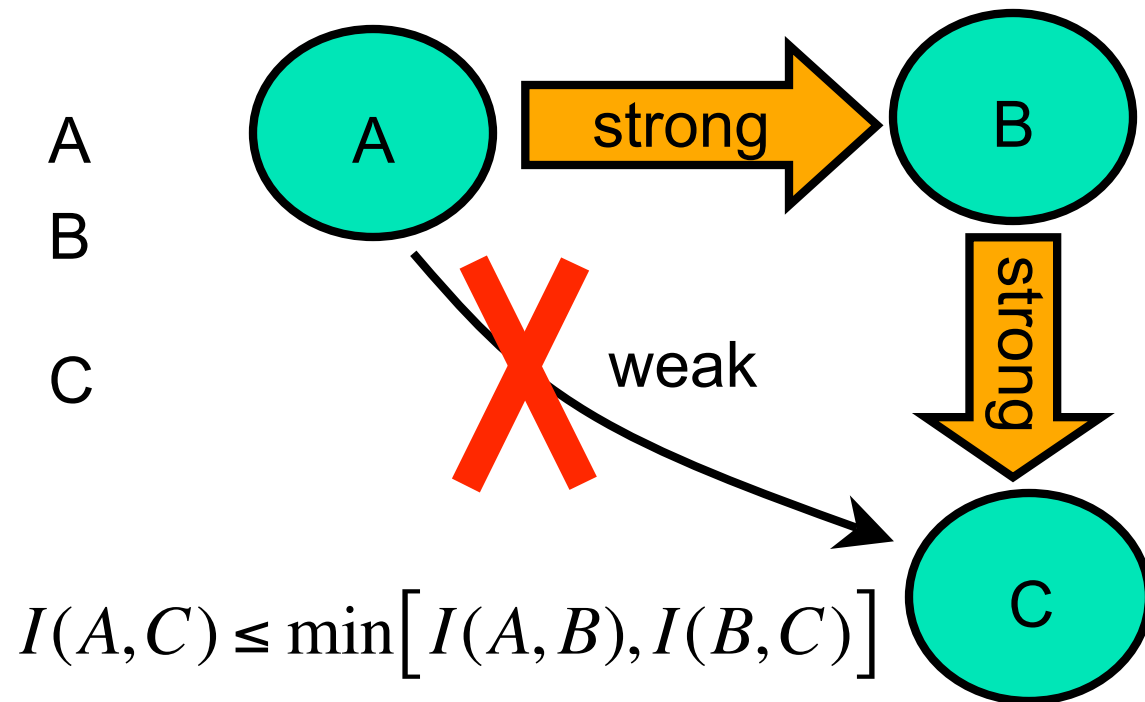
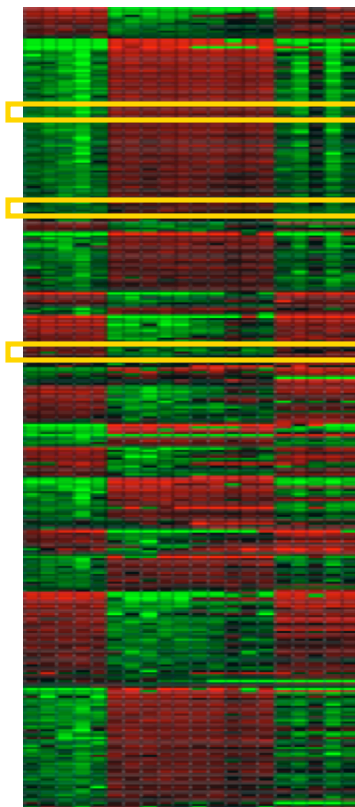


Small data requirements
Robustness to fluct.
Computational complexity
Conditional interactions
Reparam inv., non-param.
Irreducibility

Stat	Co	GM	Biochem.
✗✓	✓	✗✓	✗
✓	✓	✗✓	✗
✗	✓	✗	✗✓
✓	✗✓	✓	✗✓
✗✓	✗✓	✗✓	✓
✓	✗	✓	✗

↑
Influenciomics

Influenciomics (steady state)



What is I (influence)?
Influence vs. interaction?



Two *separate* influenciomics problems

- What is a (statistical, biological) interaction?
 - What does an arrow mean?
 - Higher order dependencies
- Realistic algorithms to uncover them
 - Controlled approximations
 - Biologically sound approximations
 - Performance guarantees
 - Complexity, Robustness, Data requirements...



Defining influence: Variances and Correlations

$$\sigma^2(x)$$

normal

$$\rho(x, x^2) = 0$$

linear

$$\rho(f(x), g(y)) \neq \rho(x, y)$$

not invariant



One-to-one transformations of microarray expression data completely destroys the ranking of correlations. Even sign of correlations may change.



Entropy (unique measure of randomness, in bits)

$$S[X] = - \sum_{x=1}^K p_x \log p_x = - \langle \log p_x \rangle$$

$$0 \leq S[X] \leq \log K \quad (\text{number of "bins"})$$

$$N(x_0, \sigma^2) \Rightarrow S[X] = \frac{1}{2} \log(2\pi e \sigma^2)$$

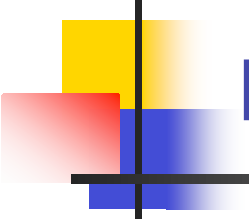


Defining influence: Mutual Information

$$I[X;Y] = \left\langle \log \frac{p_{xy}}{p_x p_y} \right\rangle$$
$$= S[X] + S[Y] - S[X,Y]$$

$$0 \leq I[X;Y] \leq \min(S[X], S[Y])$$

$$N[(x_0, y_0), \Sigma] \Rightarrow I[X;Y] = -\frac{1}{2} \log(1 - \rho_{xy}^2)$$



Why MI as influence measure?

- Captures all dependencies (zero *iff* joint probabilities factorize)
- Reparameterization invariant
- Unique metric-independent measure of “how related”

For 2 variables:

Influence ($I > 0$) **is** interaction.

(Nemenman and Tishby 2005)



Kullback-Leibler divergence

$$D_{KL}[P \parallel Q] = \sum_x p_x \log \frac{p_x}{q_x}$$

$$0 \leq D_{KL}$$

How easy it is to mistake P for Q ?
(KS test, etc.)



MI as MaxEnt

Find least constrained (highest entropy, no interaction) approximation q to p_{xy} , s.t.

$$p_x = q_x$$

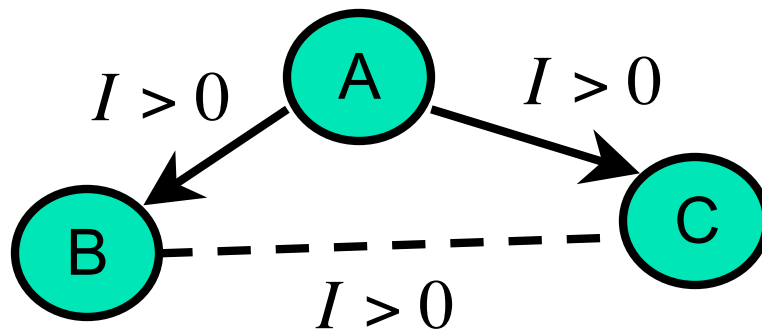
$$p_y = q_y$$



$$q_{xy} = \frac{1}{Z} \exp[-\varphi_x - \varphi_y] = p_x p_y$$

$$I[X; Y] = D_{KL}[P \parallel Q] > 0 \Rightarrow \text{interaction}$$

By analogy: Example of irreducibility



$$P_{ABC} = \frac{P_{AB}P_{AC}}{P_A} = \frac{1}{Z} f_{AB} f_{BC}$$

MaxEnt approximation without BC:

$$Q_{ABC} = \frac{1}{Z} \exp(-\varphi_{AB} - \varphi_{AC}) \Rightarrow D_{KL}[P_{ABC} \parallel Q_{ABC}] = 0$$

No irreducible interaction!

For AB: $Q_{ABC} = \frac{1}{Z} \exp(-\varphi_{AC} - \varphi_{BC}) \quad D_{KL}[P_{ABC} \parallel Q_{ABC}] > 0$

Irreducible interaction.



Higher order influences

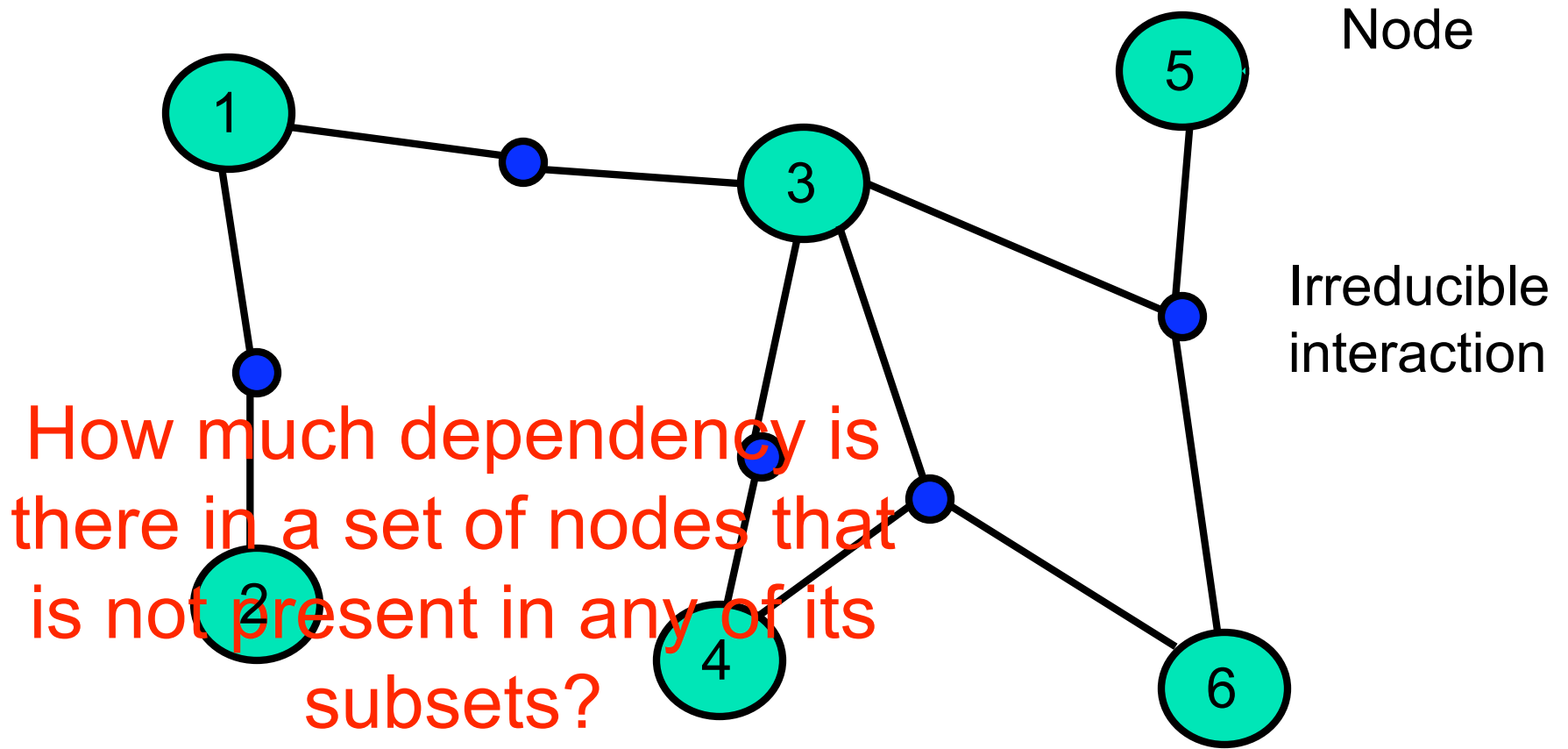
$$I_{XYZ} = \left\langle \log \frac{p_{xyz}}{p_x p_y p_z} \right\rangle$$

(Axiomatically) Amount of *all* influence (in bits) among variables.

But these are not irreducible.

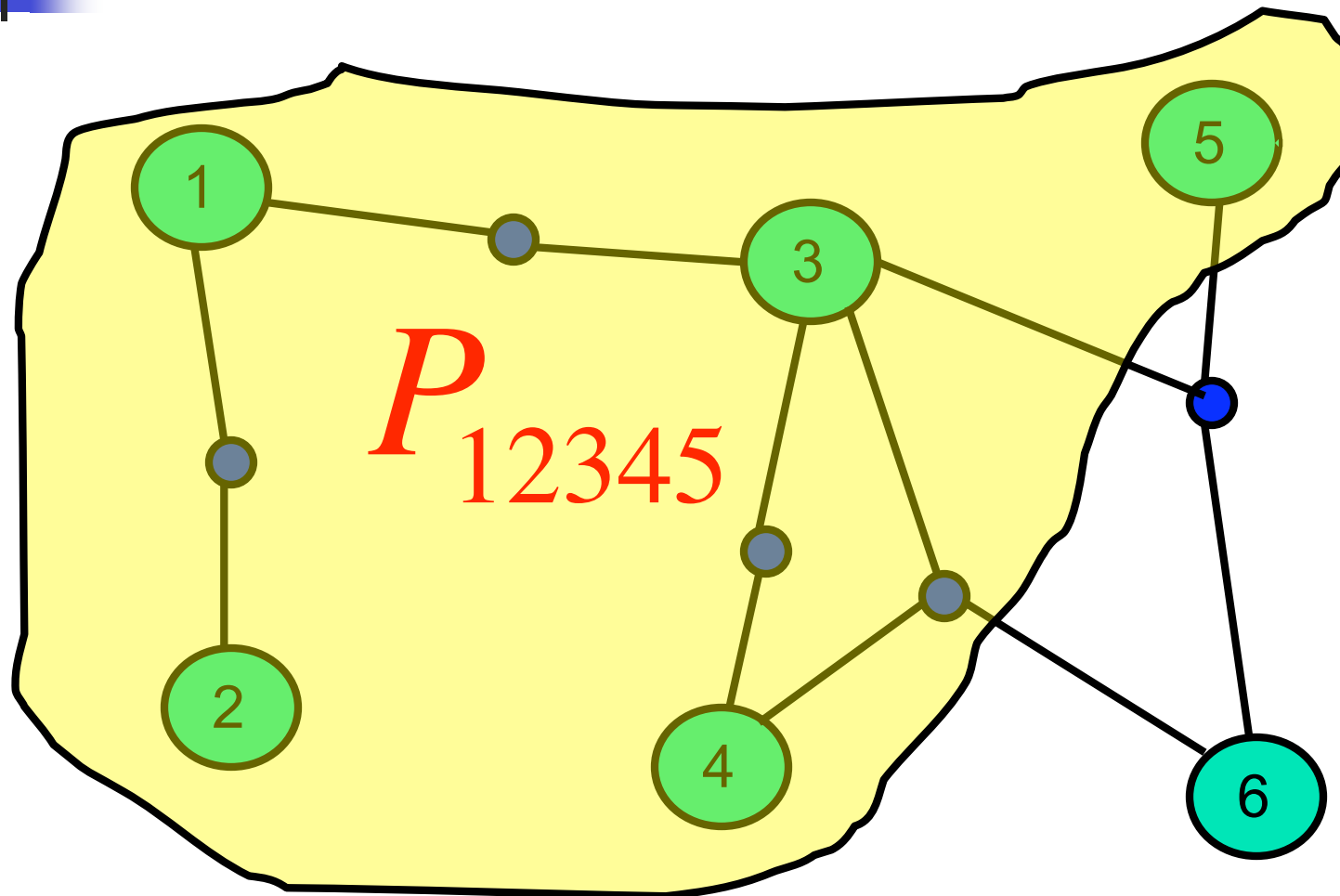
(Nemenman and Tishby 2005)

Higher order irreducible dependencies

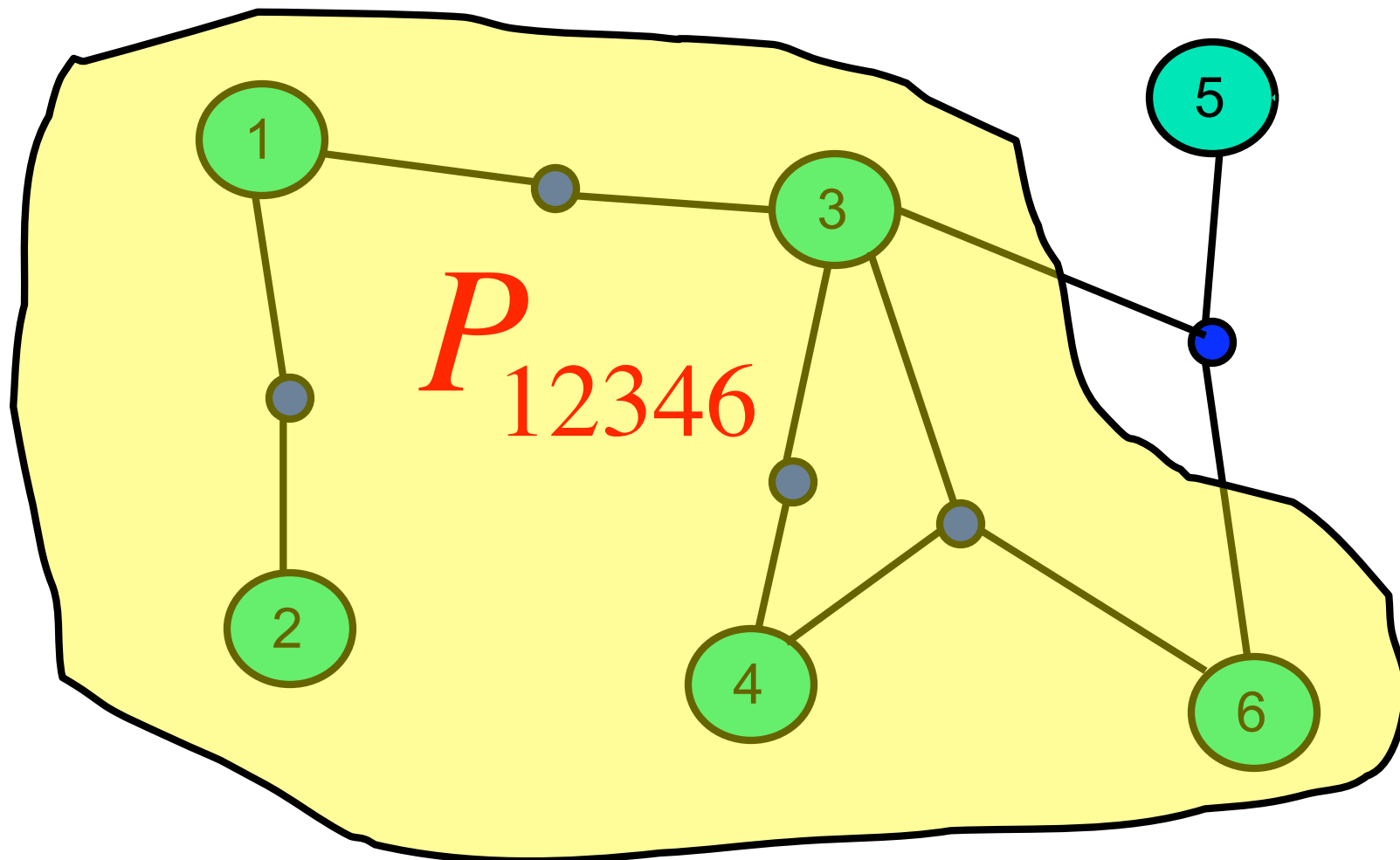


(Schneidman et al. 2003, Nemenman 2004)

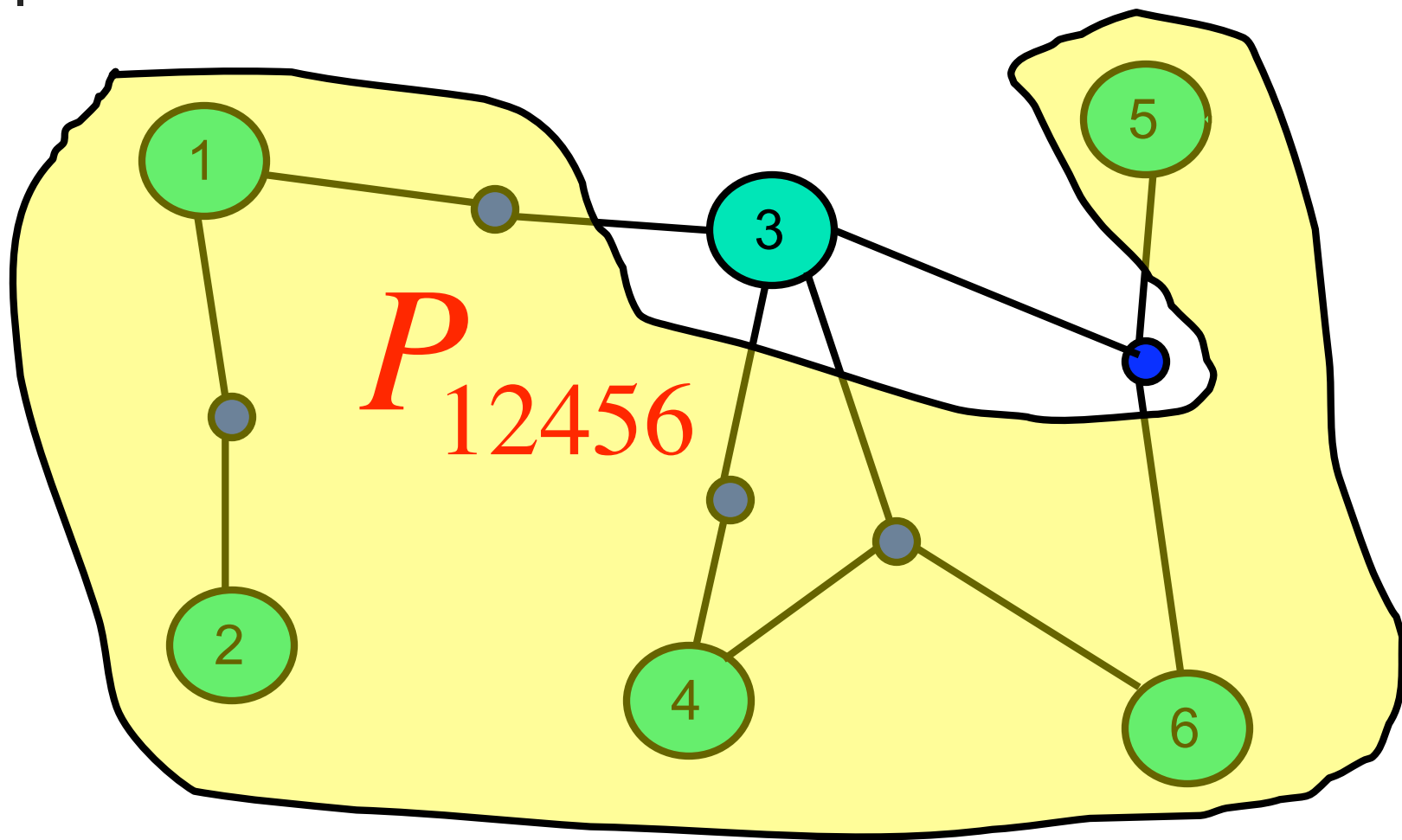
MaxEnt approximations



MaxEnt approximations



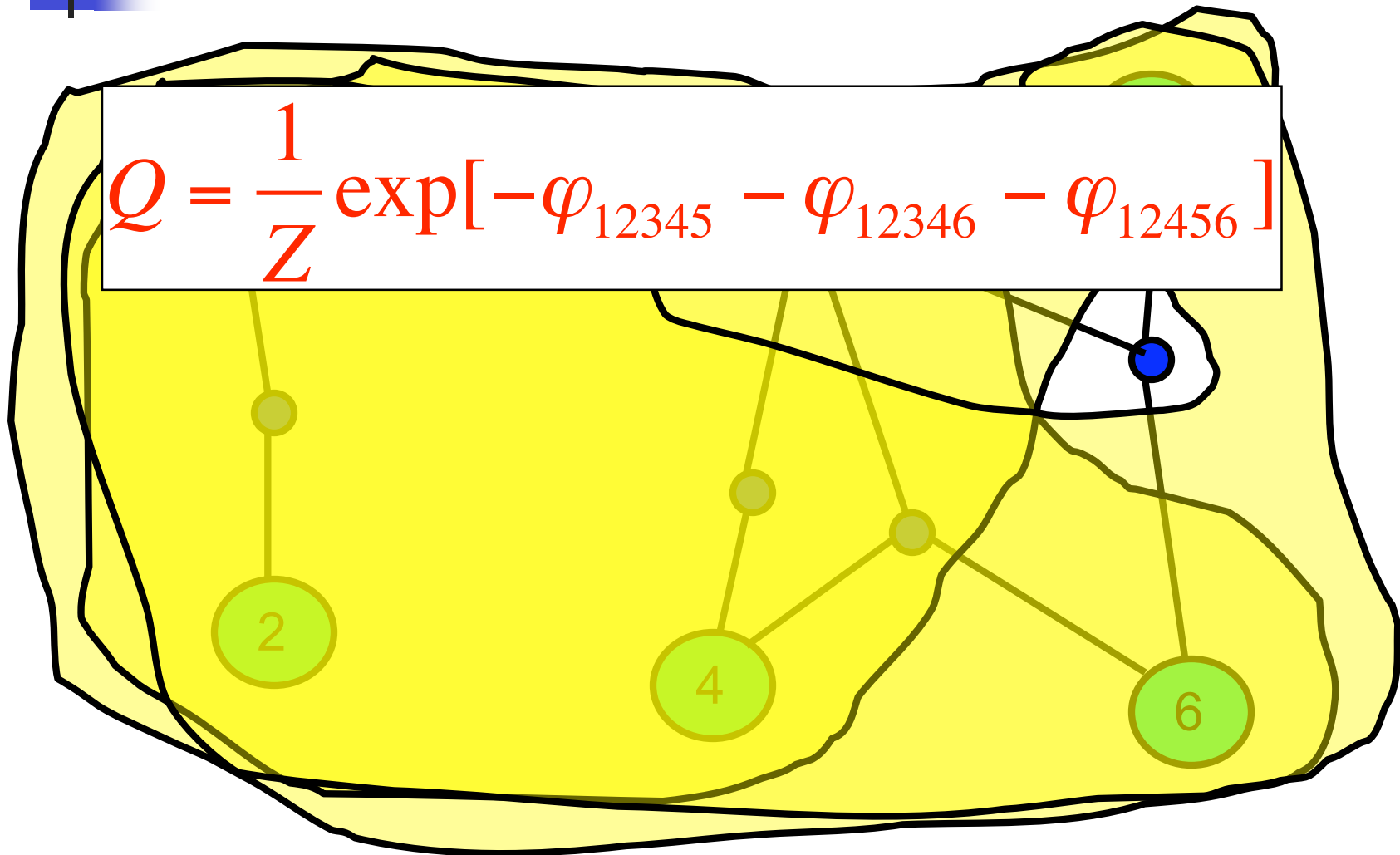
MaxEnt approximations





MaxEnt approximations

$$Q = \frac{1}{Z} \exp[-\varphi_{12345} - \varphi_{12346} - \varphi_{12456}]$$





MaxEnt approximations


$$Q = \frac{1}{Z} \exp[-\varphi_{12345} - \varphi_{12346} - \varphi_{12456}]$$

$$Q' = \frac{1}{Z} \exp[-\varphi_{12345} - \varphi_{12346} - \varphi_{12456} - \varphi_{356}]$$



MaxEnt approximations

$$I'_{356} = D_{KL}[Q' \parallel Q]$$

$I'_{356} > 0 \Rightarrow$ Irreducible interaction present



MaxEnt factorization of PDFs

$$P(x_1, \dots, x_M) =$$
$$= \exp \left[- \sum_i \varphi_i(x_i) - \sum_{ij} \varphi_{ij}(x_i, x_j) - \sum_{ijk} \varphi_{ijk}(x_i, x_j, x_k) - \dots \right]$$

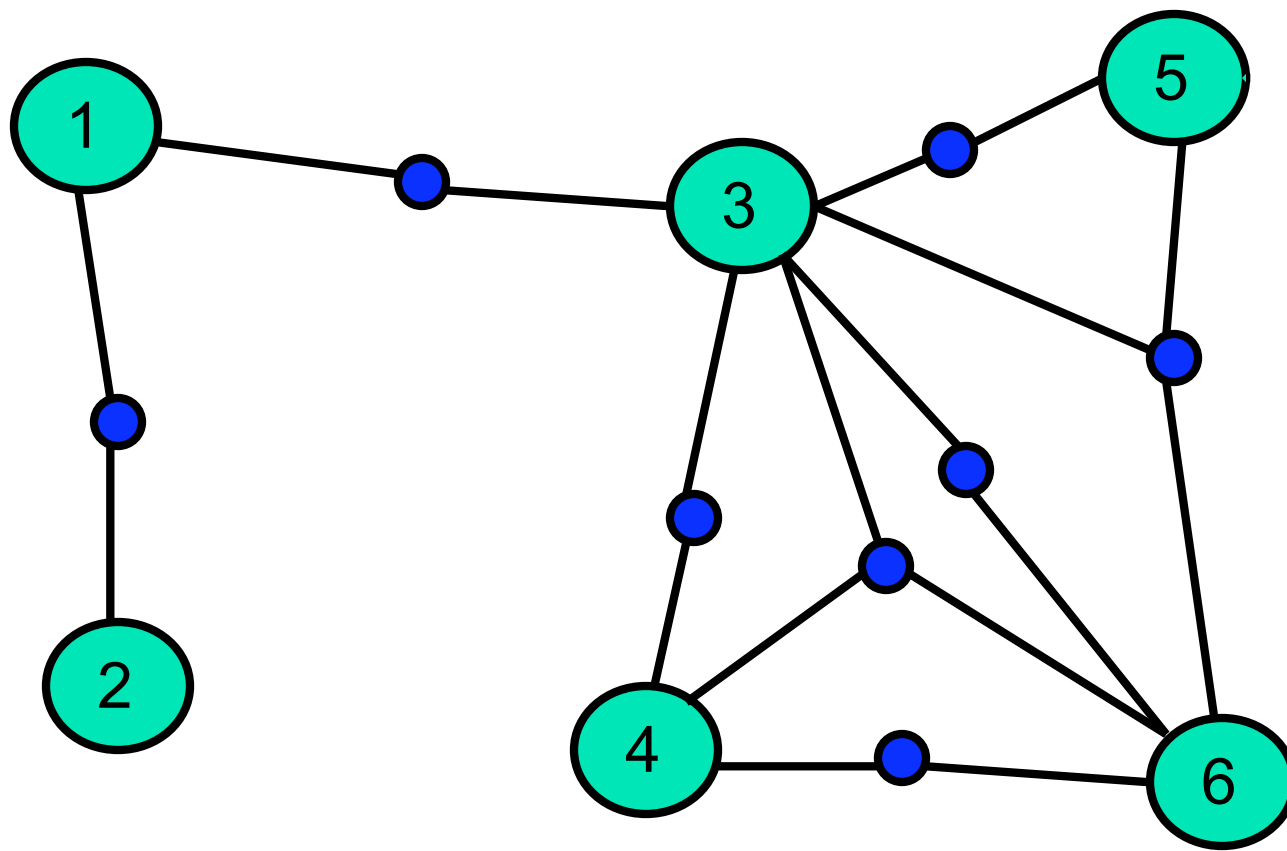
- N -particle potentials
- Spin models -- inverse problem (for discrete variables)
- Random lattices
- Message passing
- Markov Networks



Two *separate* influenziomics problems

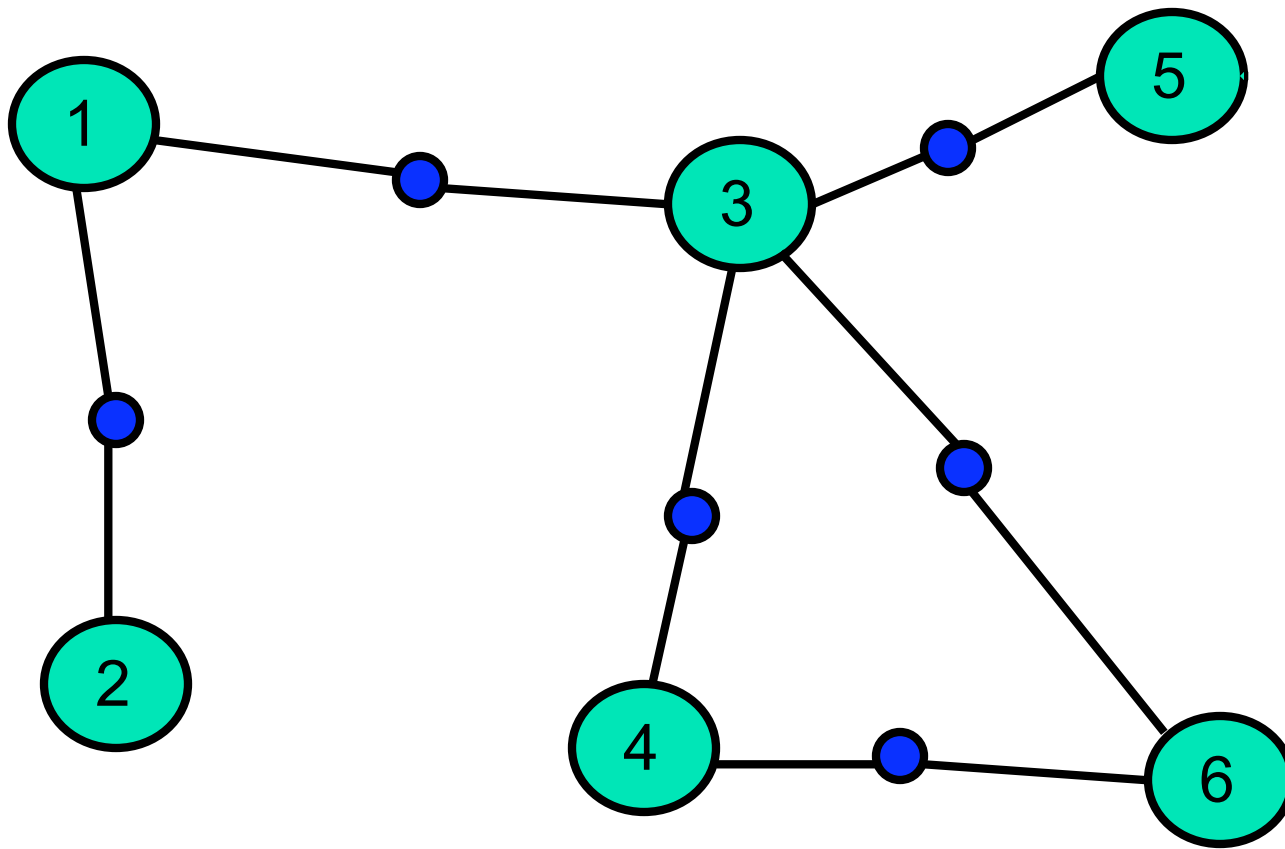
- What is an interaction?
 - What does an arrow mean?
 - Higher order dependencies
- Realistic algorithms to uncover them
 - Controlled approximations (e.g., know the order)
 - Biologically sound assumptions (new knowledge from their verification)
 - Performance guarantees (focus on low false positives for irreducibility)
 - Complexity, Robustness, Data requirements...

Interaction network

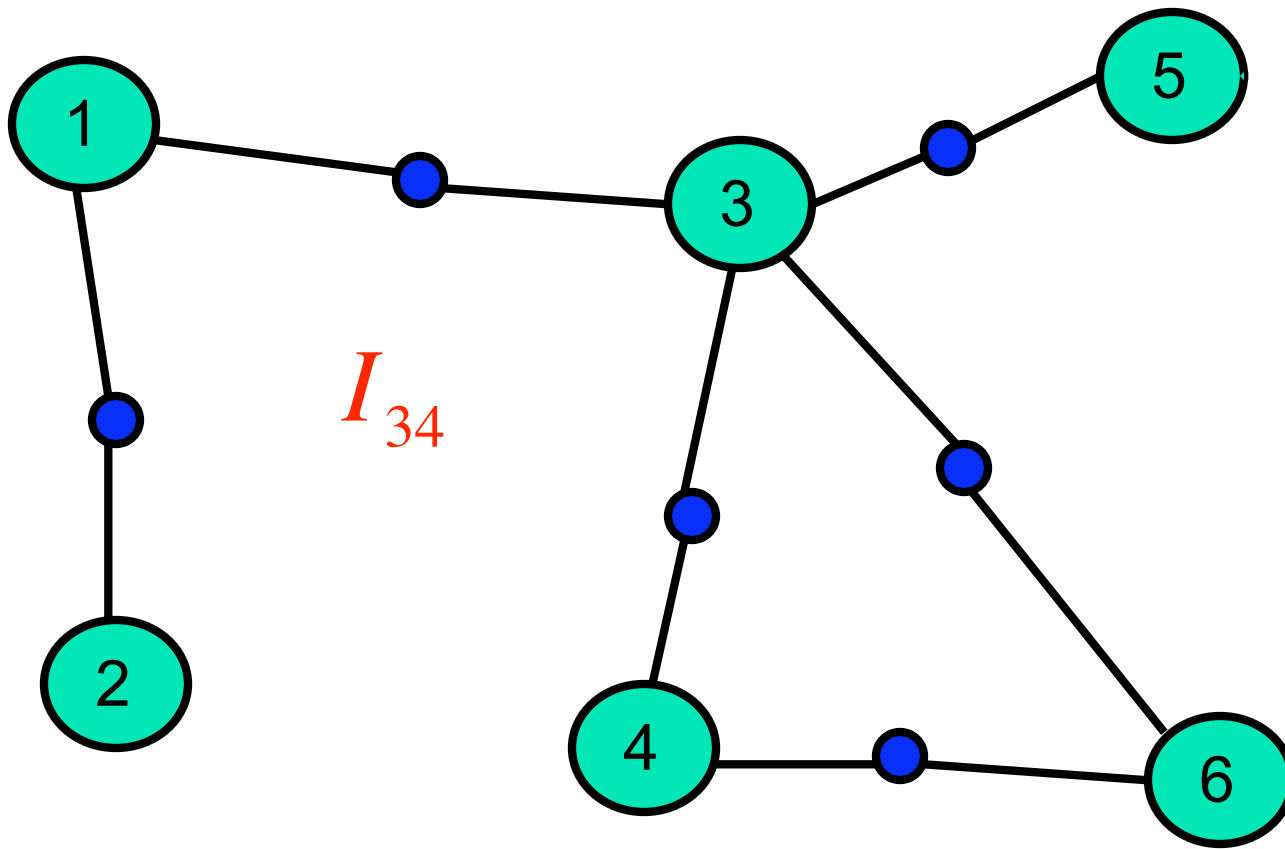


(Basso et al. 2005, Margolin et al. 2005)

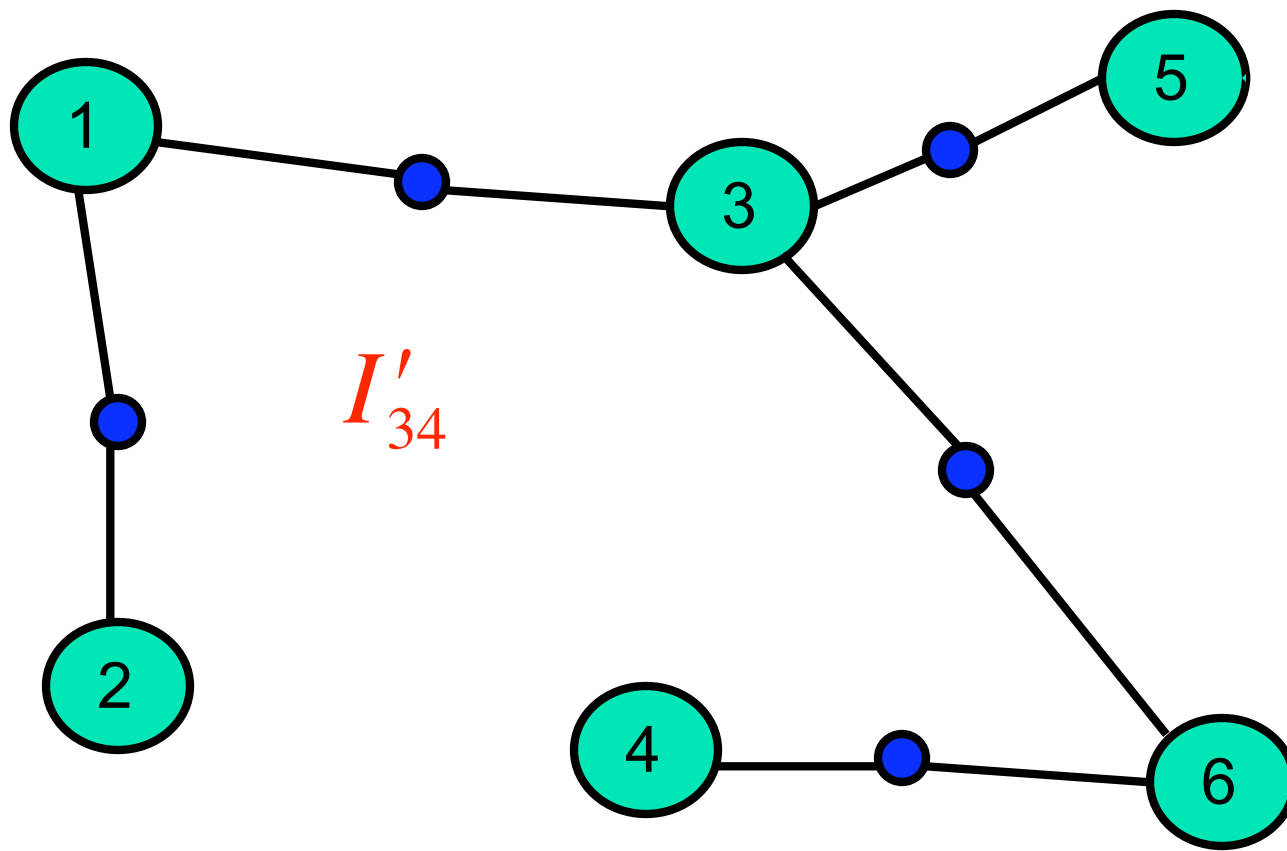
Disregard high orders (undersampling)



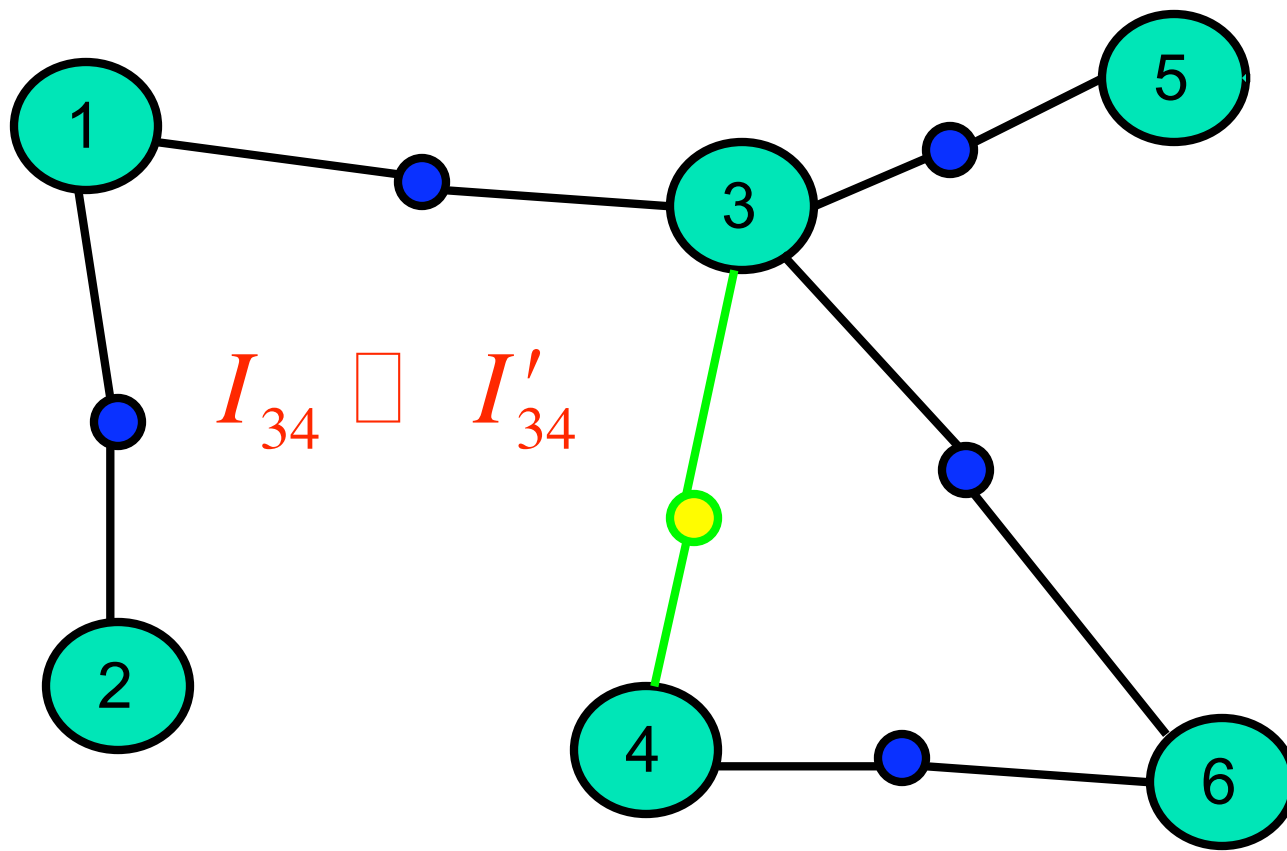
Locally tree-like approximation



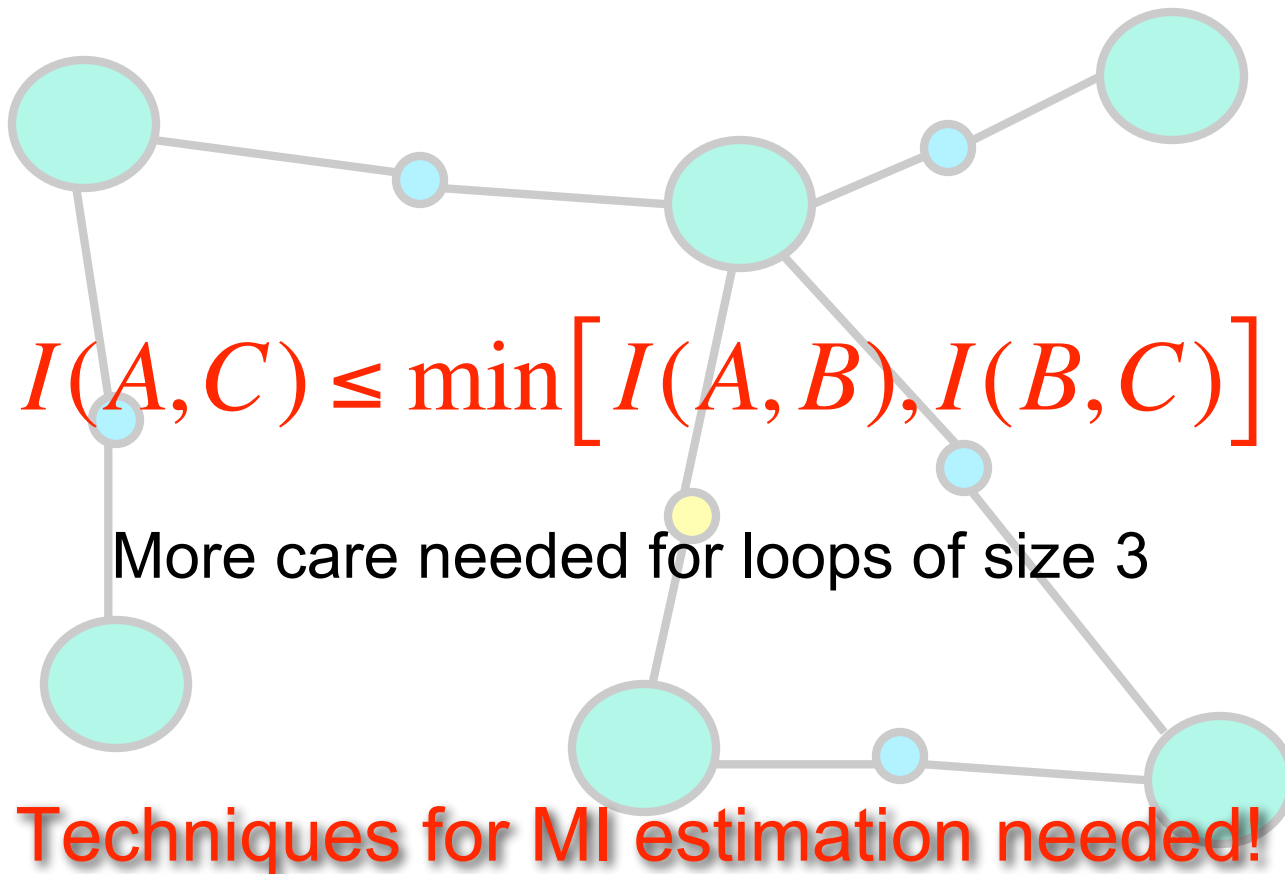
Locally tree-like approximation



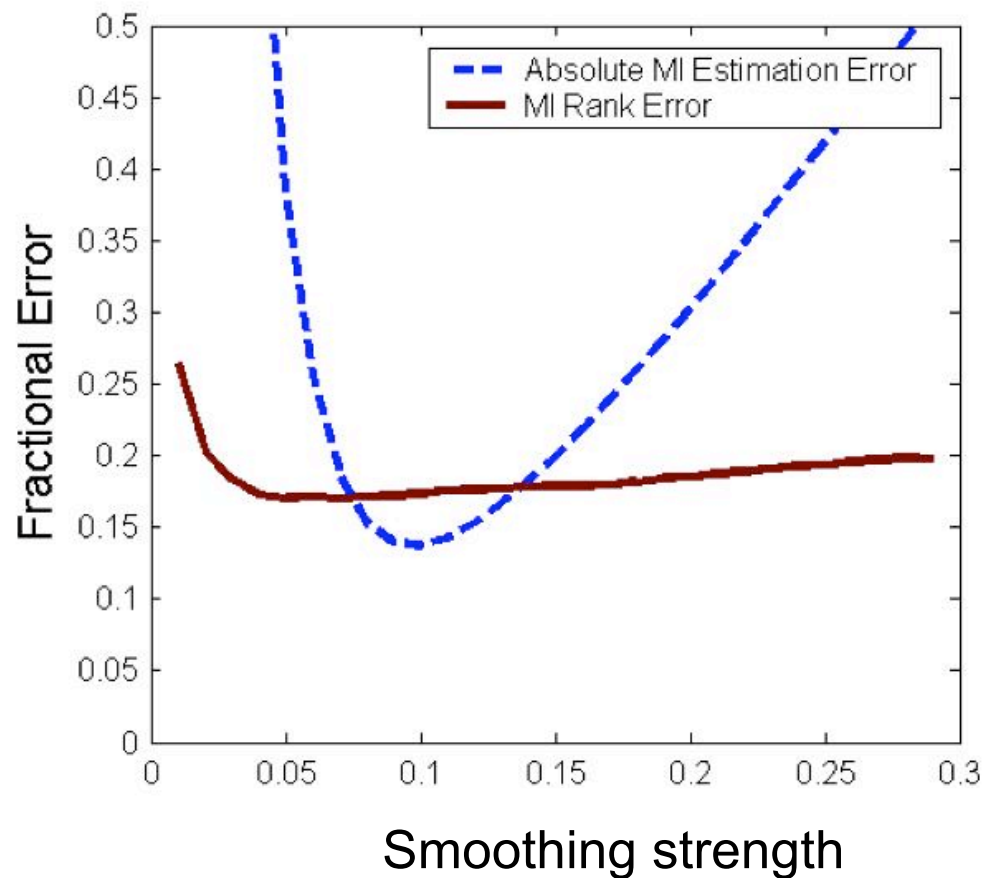
Locally tree-like:
signals decorrelate fast



ARACNE: remove the weakest link in every triplet



Estimating λ : stability of ranks



Also:

- NSB
- copula



No false positives

Where 2-way -- it's 2-way

Theorem 1. If MIs can be estimated with no errors, then ARACNE reconstructs the underlying interaction network exactly, provided this network is a tree and has only pairwise interactions.

Theorem 2. The Chow-Liu maximum mutual information tree is a subnetwork of the network reconstructed by ARACNE.

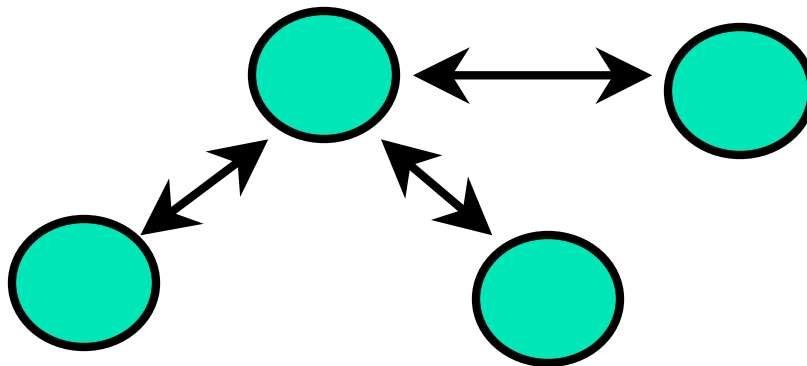
Theorem 3. Locally tree-like -- no false positives (no false negatives under stronger conditions).

Aside: Bethe approximation, Message passing (MP)

$$P(\{x_i\}) = \frac{\prod P(x_i, x_j)}{\prod P(x_i)^{q-1}}$$

Exact for trees

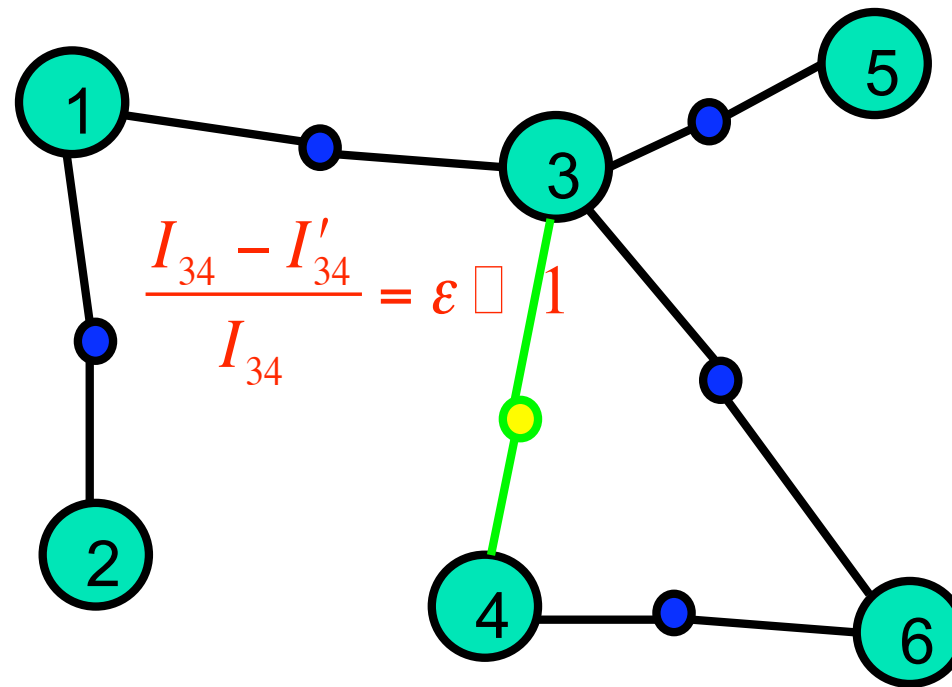
$$P(x_i) = ?$$



MP (belief propagation, transf. matrix) works for trees and *sometimes* for loopy networks. But when exactly?

Conjecture

Locally tree like assumption is what makes MP work!

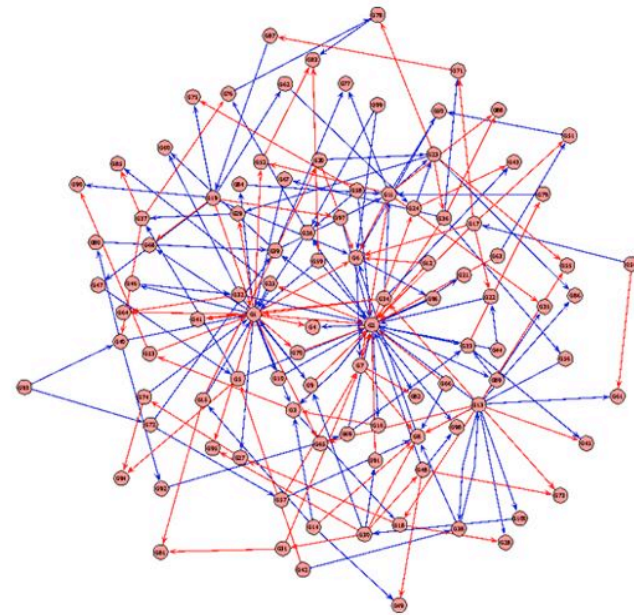
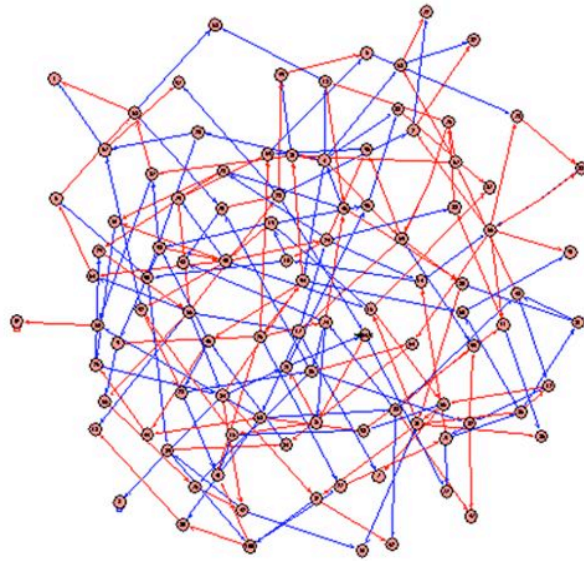




Biological soundness

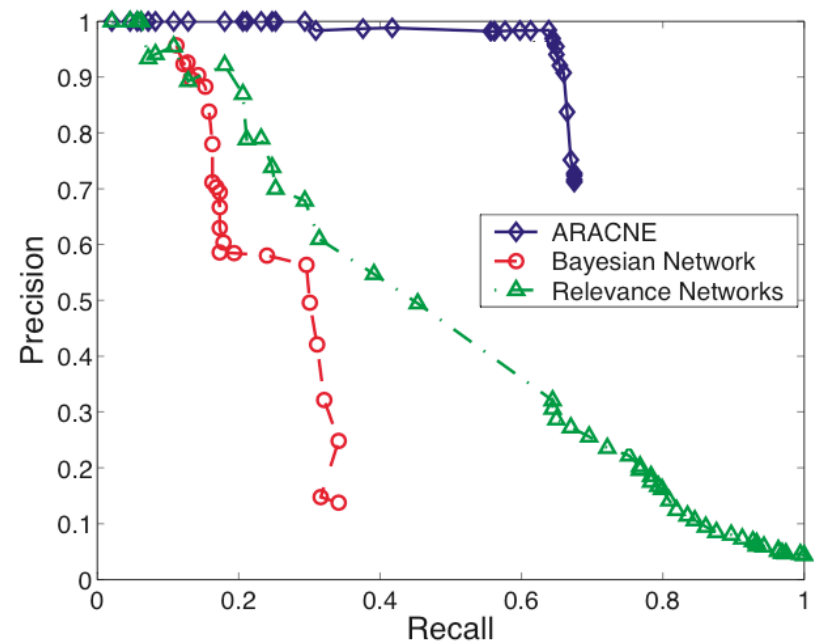
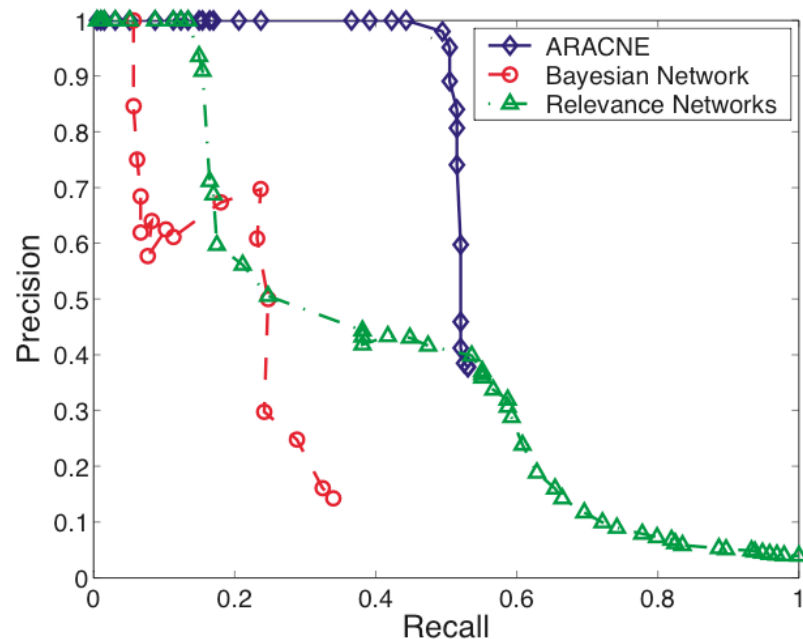
- Higher order interactions project to lower orders
- Fast decorrelation:
 $I(\text{gene}, \text{gene}) \gg I(\text{gene}, \text{second best})$
- Small loops often transient

Synthetic networks



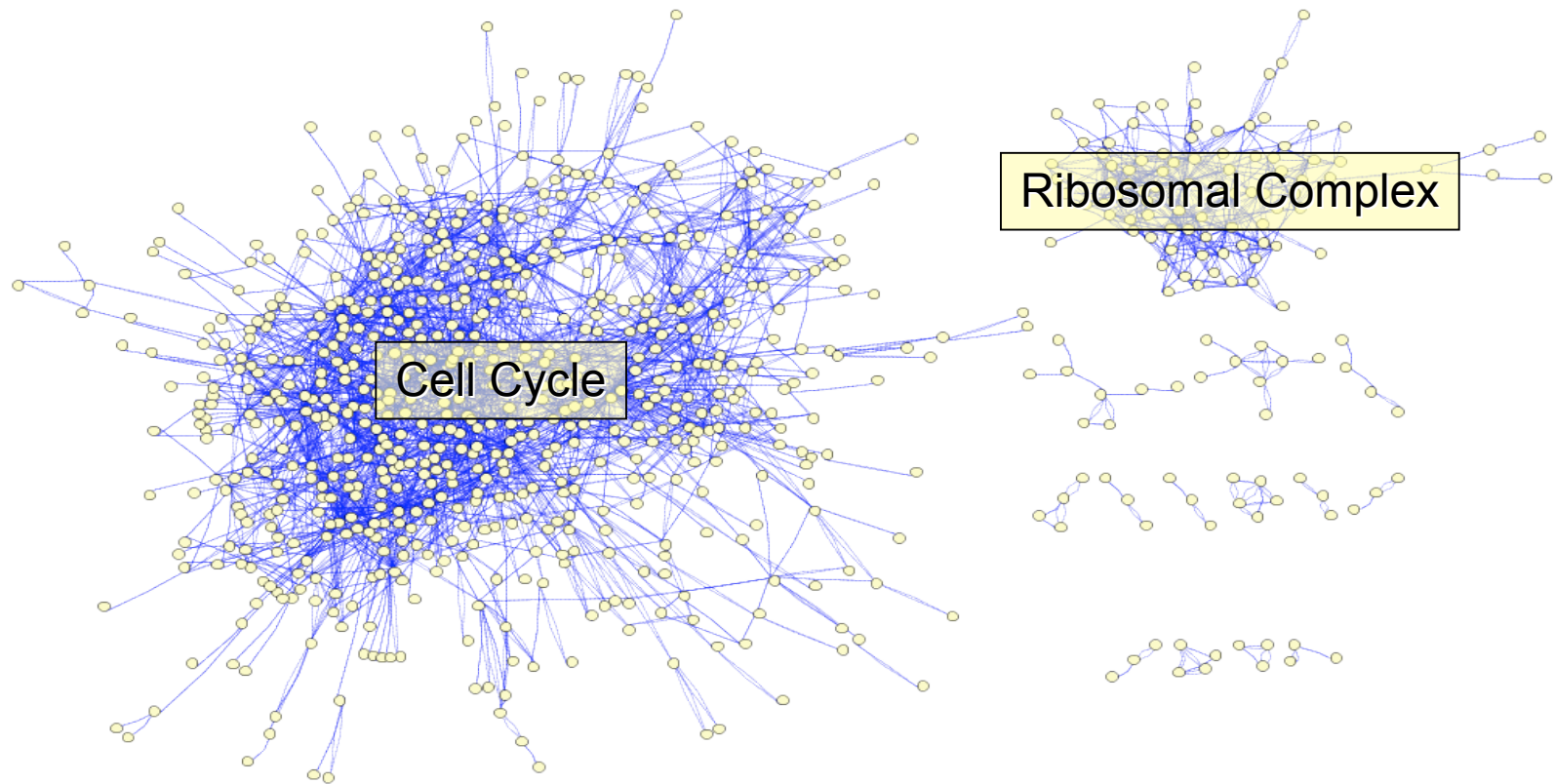
$$\frac{dx_i}{dt} = a_i \prod_j \frac{I_{0,j}^{v_j}}{I_j^{v_j} + I_{0,j}^{v_j}} \prod_j \left(1 + \frac{A_{0,j}^{v_j}}{A_j^{v_j} + A_{0,j}^{v_j}} \right) - b_i x_i$$

Synthetic networks benchmarks ($N=1000$)



Graceful decay for smaller N
Half of all loops kept.

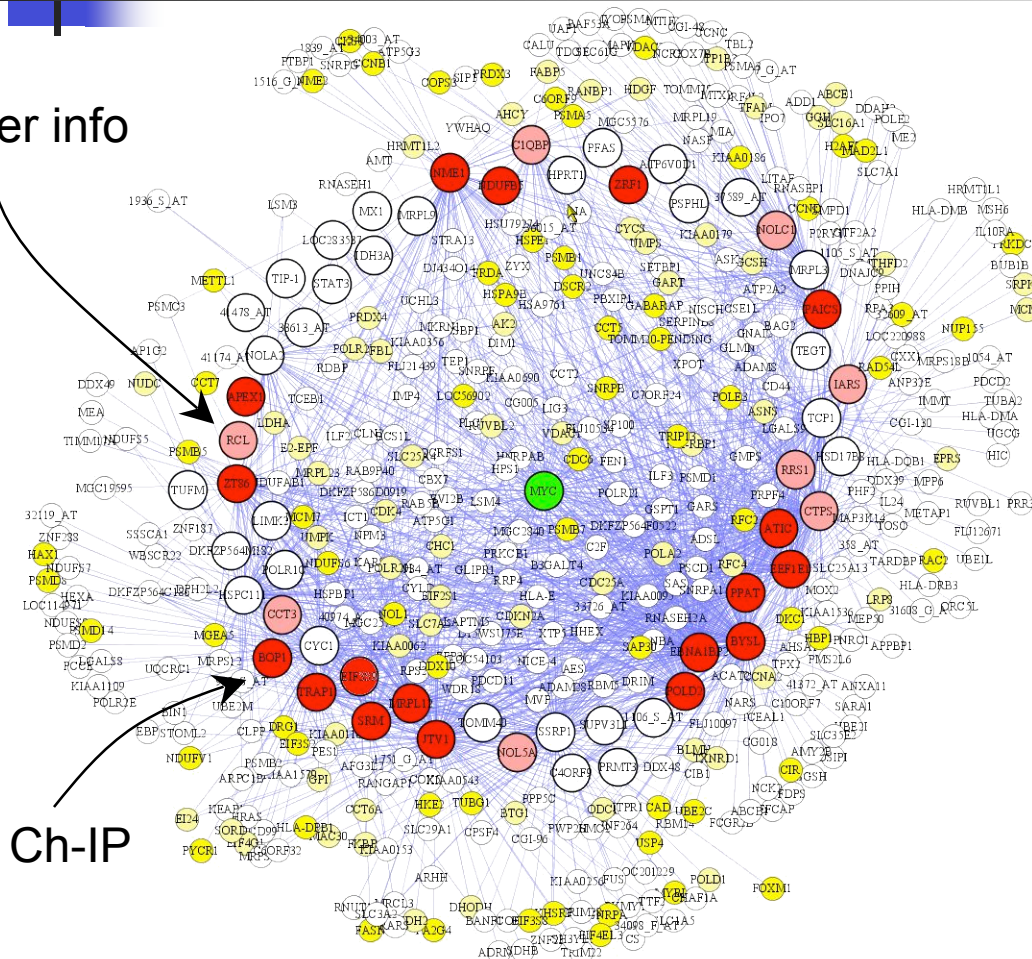
Complete B-cell network (400 arrays)



~129000 interactions

c-MYC subnetwork

other info

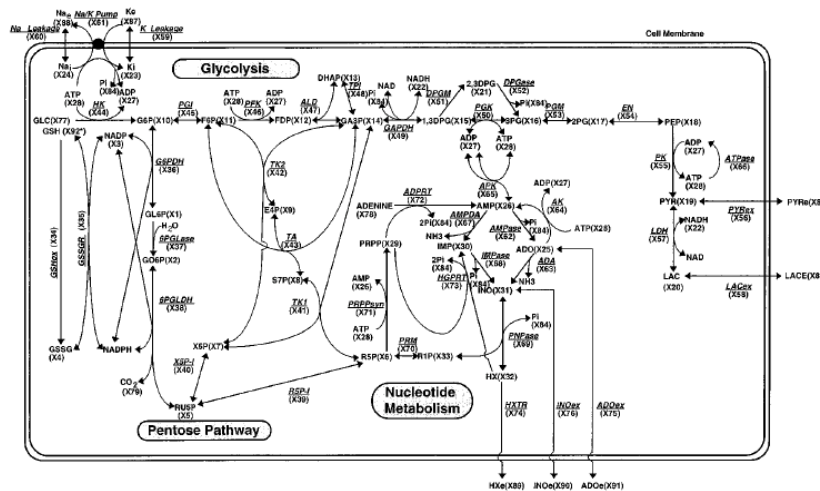


- Protooncogene,
- 12% background binding,
- one of top 5% hubs
- significant MI with 2000 genes

Total interactions: 56
Pre-known: 22
Ch-IP validated: 11/12

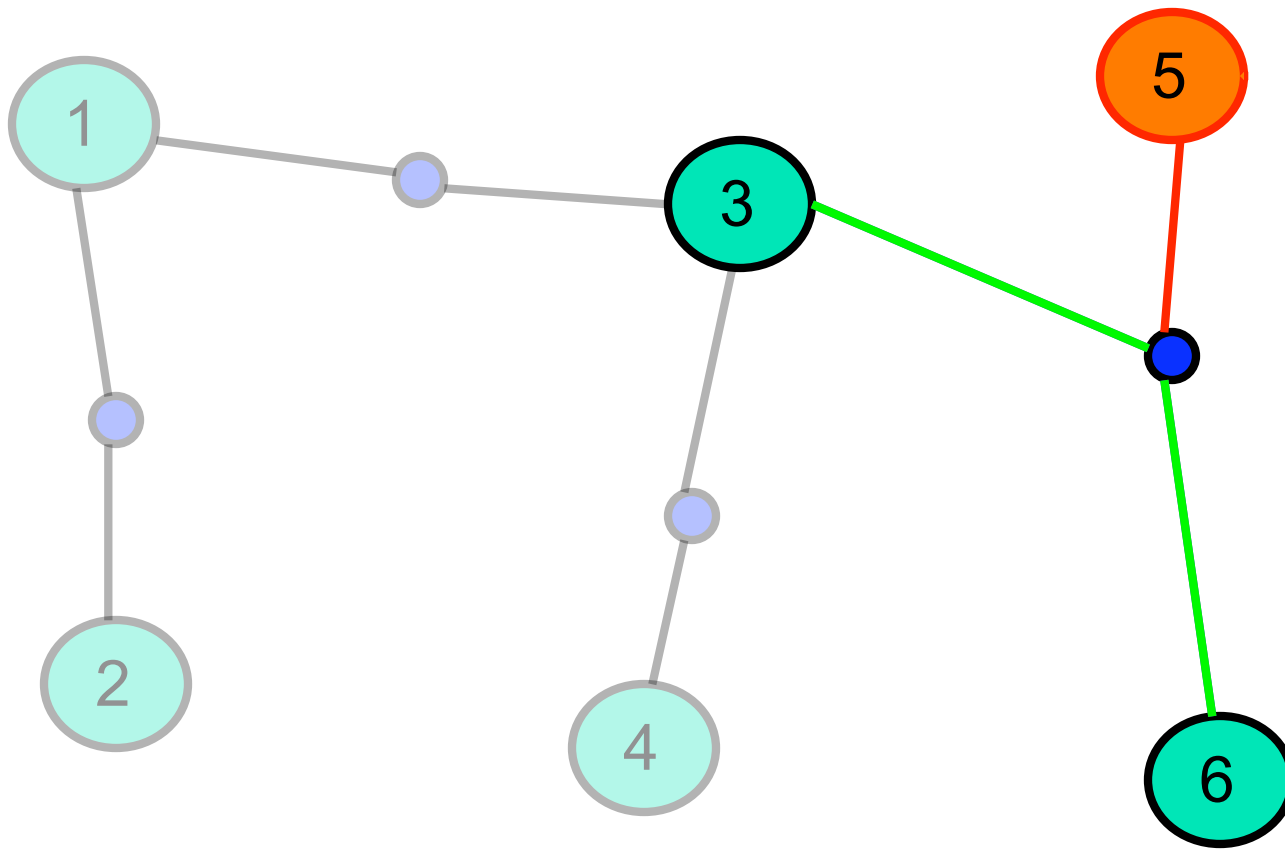
Also validated in...

- Other hubs
- Various yeast data sets
- RBC metabolic network (synthetic)



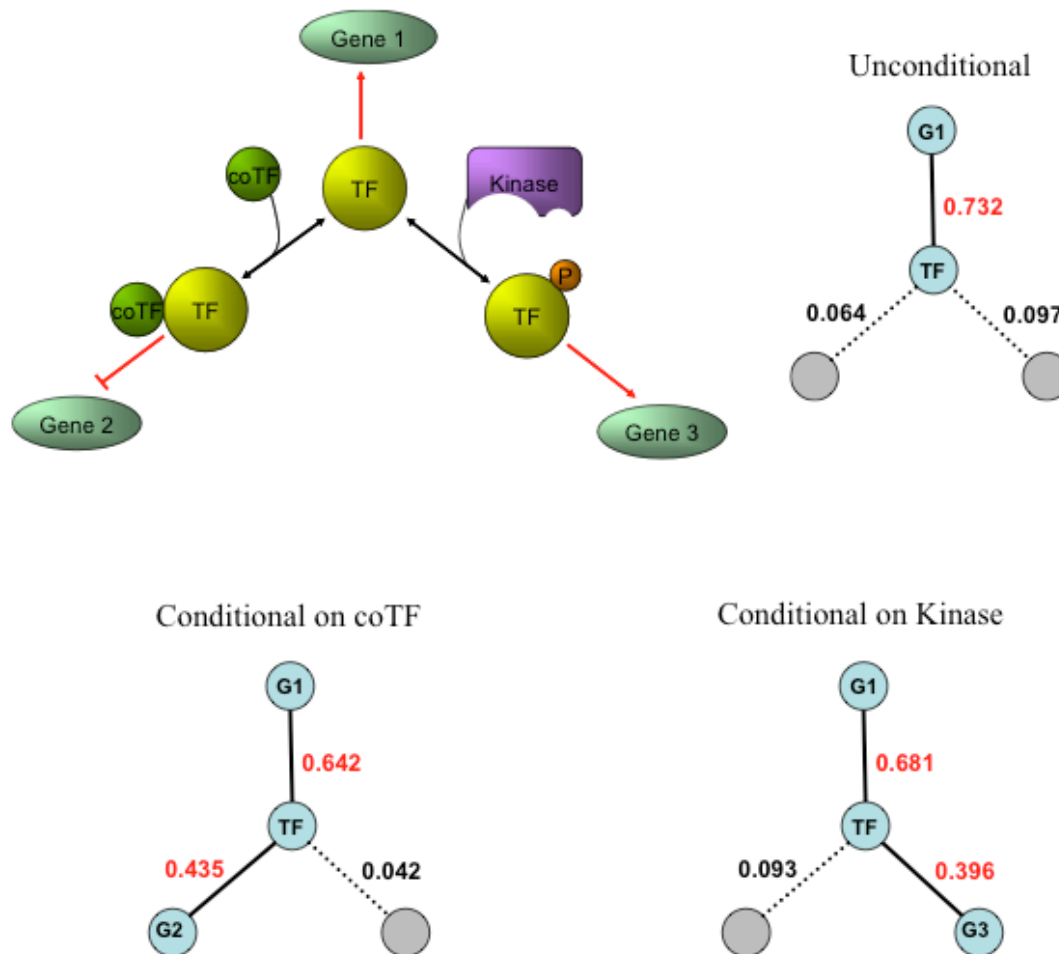
~80% precision
20-80% recall (depending on N)

3rd order interactions (modulated, conditional)

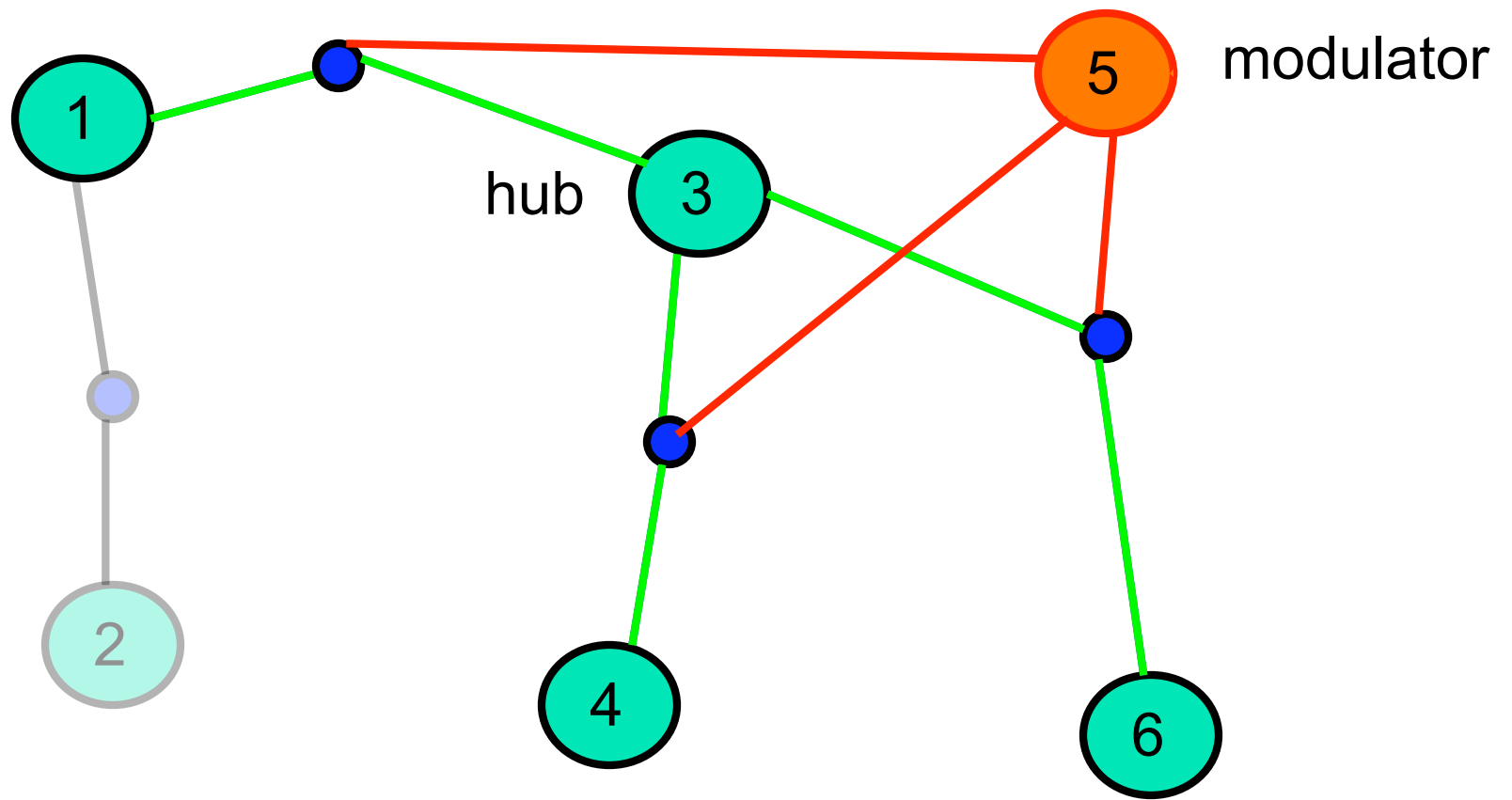


Nontranscriptional modulators from expression data!

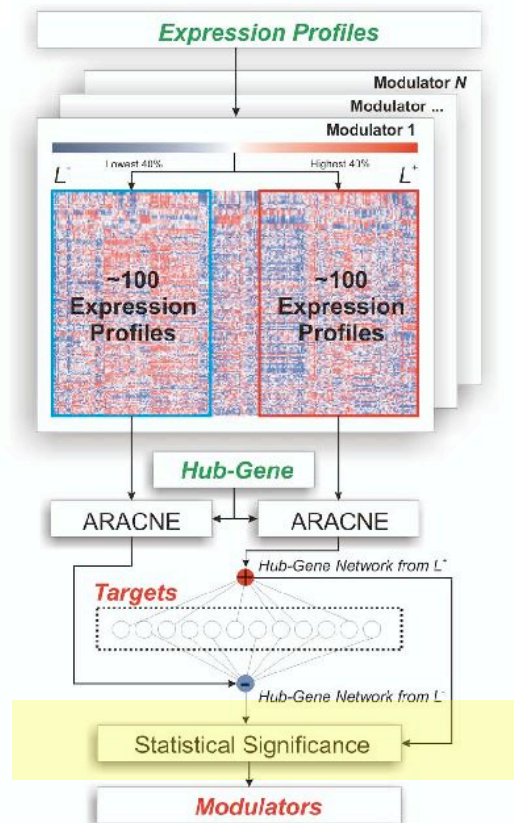
Numerical case study: Non-transcriptional modulation



Large hubs, global (discrete) modulators



Large hubs, global (discrete) modulators



- Focus on important hubs (c-MYC)
- Pre-filter candidate modulators by dynamic range and other conditions.
- Find modulators whose expression inflicts **significant** changes on topology of the ARACNE hubs' interactions
- **No guarantee of irreducibility**
- Validate in GO w.r.t. to transcription factors and kinases among modulators

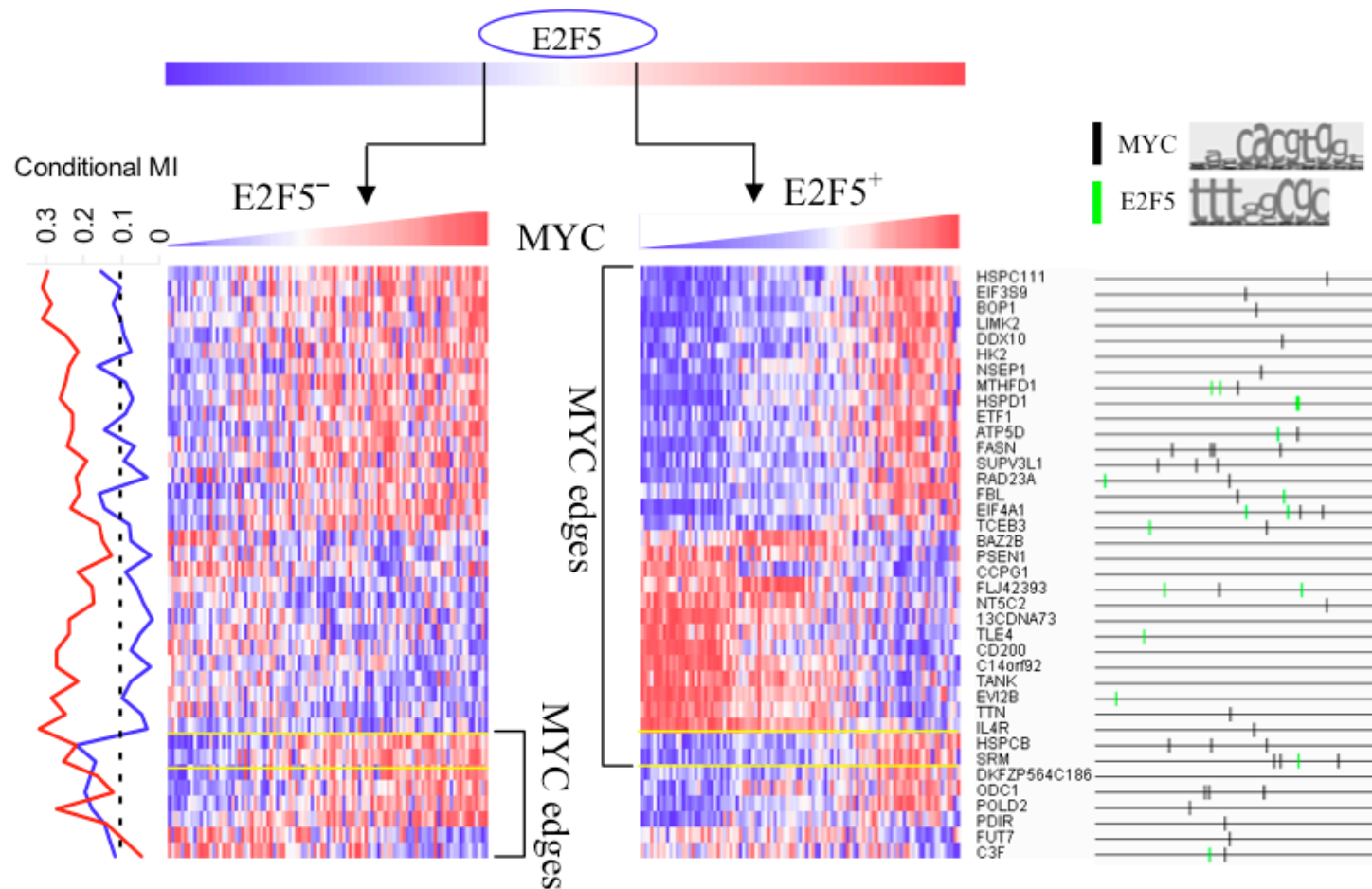
$$|N^+ - N^-| > 0$$



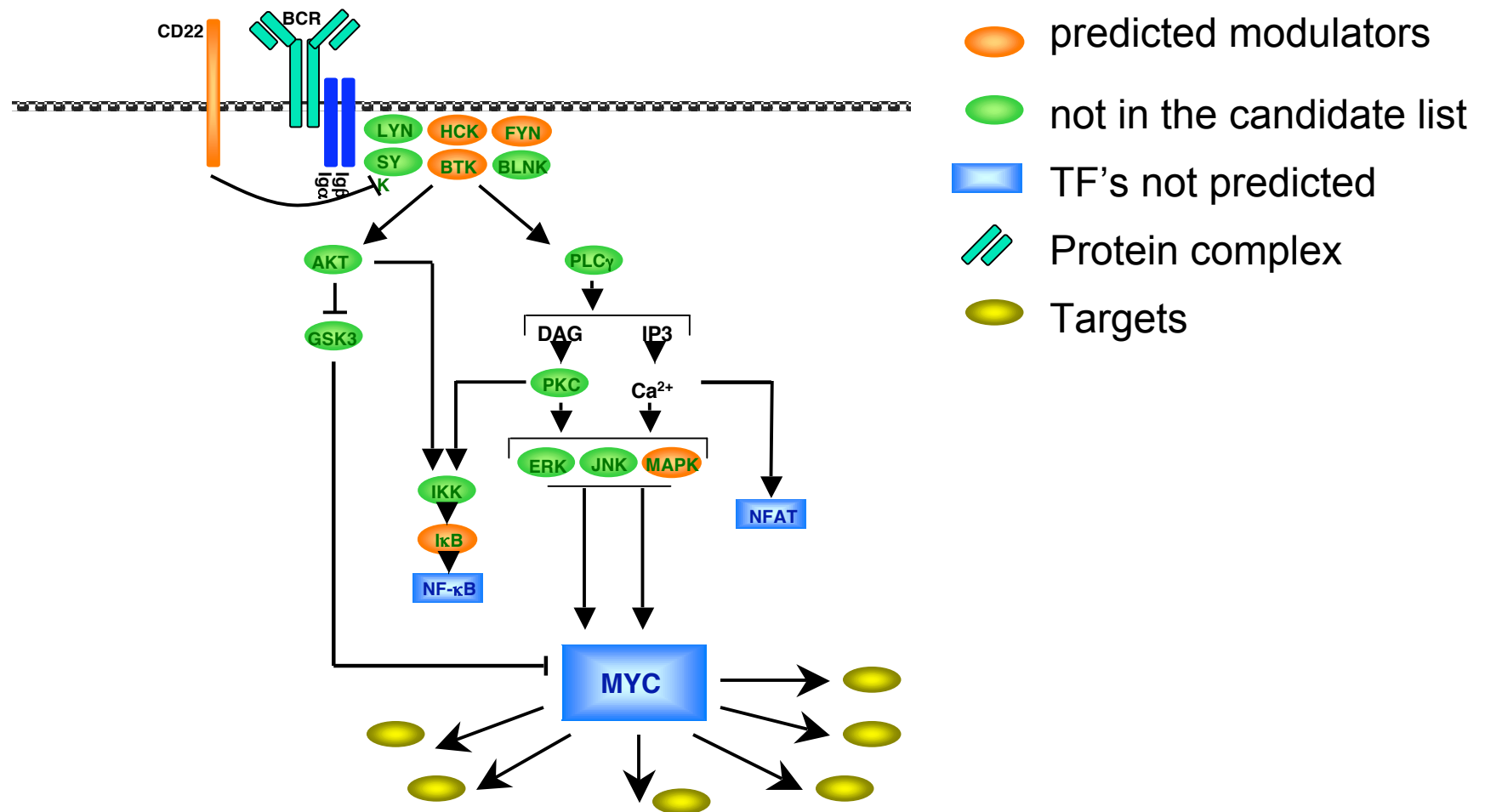
c-MYC modulators

- 1117 candidate modulators (825 with known molecular function in GO)
- 82 (69) candidate modulators identified
- Kinases: 10/69 (backgr. 42/825), $p=1e-3$
- TFs: 15/69 (backgr. 56/825), $p=1e-6$ (validated -- see below).
- Total: 25/69 (backgr. 98/825), $p=3e-8$
- Large scale modulators: ubiquitin conjugating enzyme, mRNA stability, DNA/chromatin modification, etc.

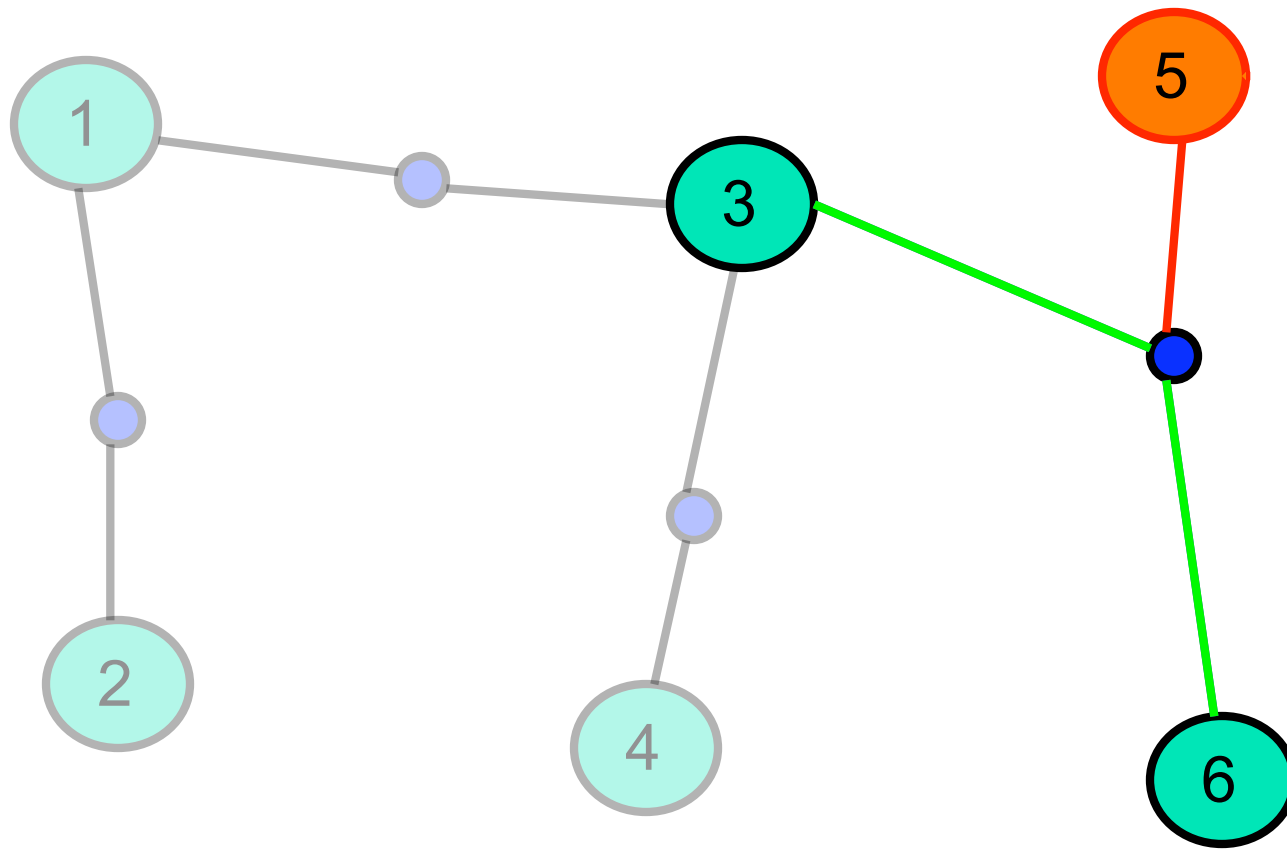
Example: TF co-factor modulator



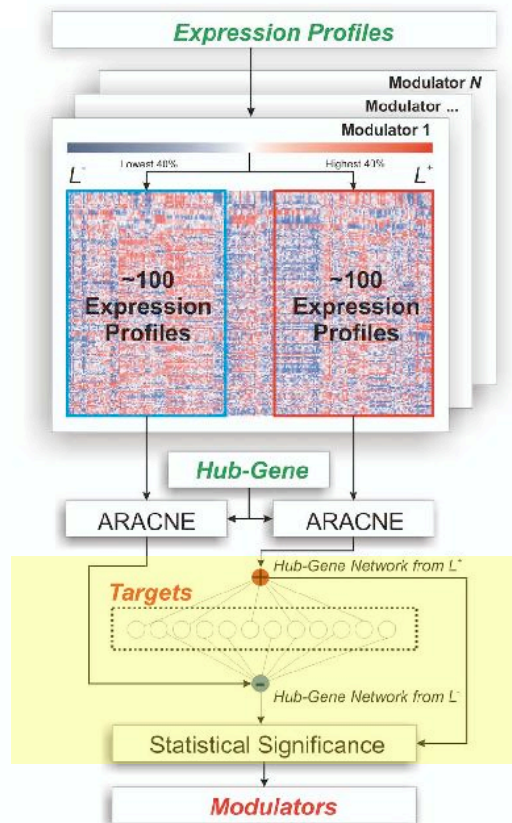
Reducibility: modulating pathways



Large hubs, local modulator (MI change, transistor)



Large hubs, local modulators

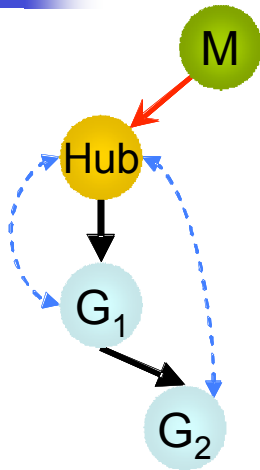


- Focus on important hubs (c-MYC)
- Pre-filter candidate modulators by dynamic range and other conditions.
- Find modulators whose expression inflicts **significant** conditional MI changes for an ARACNE target in at least one conditional topology
- **No guarantee of irreducibility**
- Validate in GO w.r.t. to transcription factors and kinases among modulators

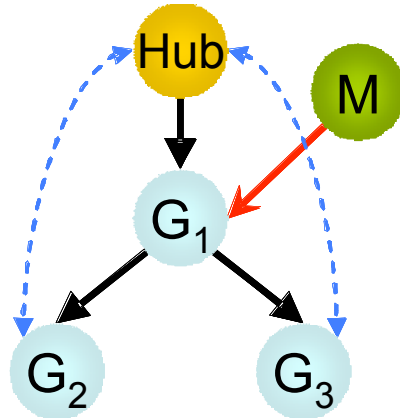
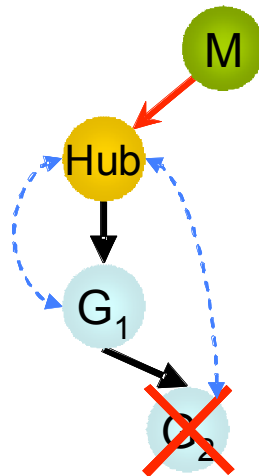
$$\Delta I(g_{TF}, g_t | g_m) =$$

$$= \left| I(g_{TF}, g_t | g_m^+) - I(g_{TF}, g_t | g_m^-) \right| > 0$$

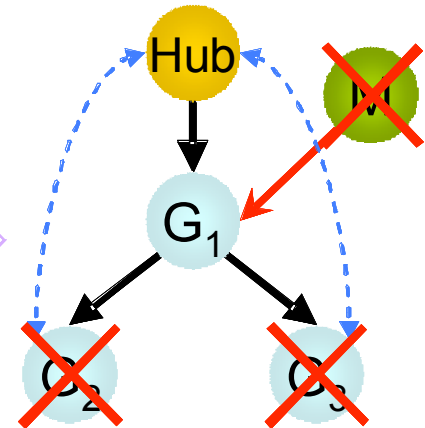
ARACNE helps



DPI



DPI





c-MYC modulators

- 1117 candidate modulators
- 100 (69) candidate modulators identified, modulating 205 interactions with 130 targets
- Modulators enriched in: kinases, acyltransferases, TFs (all at $p < 5\%$); correspond to known MYC modulation pathways.
- TFs: 15/69 (backgr. 56/825), $p = 1e-6$; binding signature for co-TFs (E2F5, MEF2B) found.
- Modulators with largest number of effected targets are not-target-specific (proteolysis, upstream signaling components, receptor signaling molecules); overlap with global modulators.
- Modulators with small number of effected targets are mostly co-TFs, are interaction-specific; no overlap with global modulators.
- About one third of modulators are literature-validated.
- 4 out of 5 TF modulators with TRANSFAC signatures have binding sites in modulated targets promoter regions.



Currently

- Biochemical validation
- Search for irreducible modulators
- Dealing with small loops



Summary

- IT quantities good measures of dependency
- Defined irreducible interactions
- Proposed a set of simplifying assumptions and a corresponding algorithm for second order interactions
- Bootstrapped the algorithm to identify certain third order dependencies
- Validated algorithms in-silico
- Analyzed interaction network of c-MYC, validated in-vivo and through literature



Thanks

- Columbia: Andrea Califano (PI), Adam Margolin (ARACNE, MI estimation), Kai Wang (Modulators 1 and 2, MI estimation), Nila Banerjee (TF signature), Omar Antar (ARACNE on yeast), Riccardo Dalla-Favera (experimental PI), Katia Basso (in-vivo validation), Chris Wiggins (simulations), AMDeC
- IBM: Gustavo Stolovitzky (simulations)
- Jerusalem: Naftali Tishby (framework)
- LANL: Michael Wall (RBC network)