UNM, 04/05/06

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

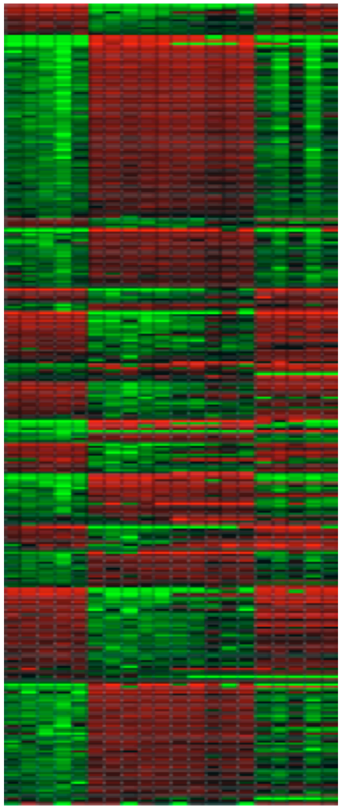
Ilya Nemenman

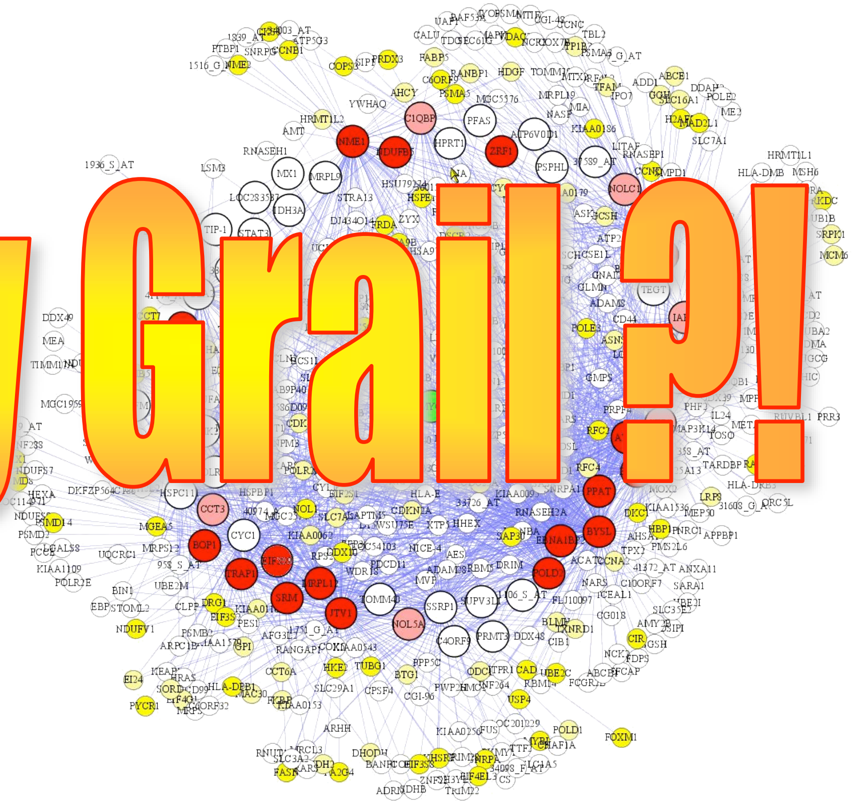
(CCS-3/LANL)

Thanks

- Columbia: *Andrea Califano* (PI), Adam Margolin (ARACNE, MI estimation), Kai Wang (Modulators, 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 (computer support)
- IBM: Gustavo Stolovitzky (simulations)
- Jerusalem: Naftali Tishby (framework)
- LANL: Michael Wall (RBC network)

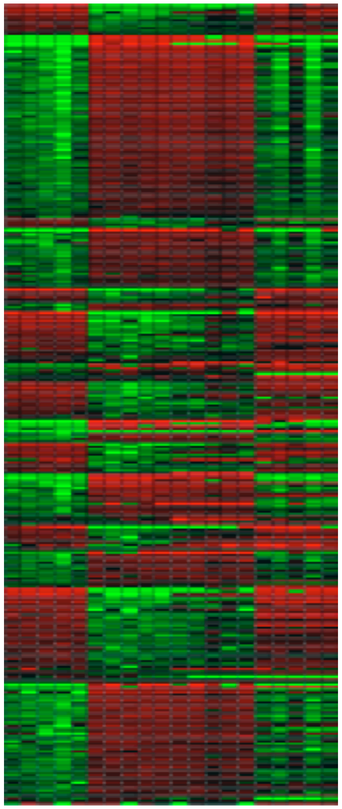
Reconstructing interaction models





Holy Grail?!

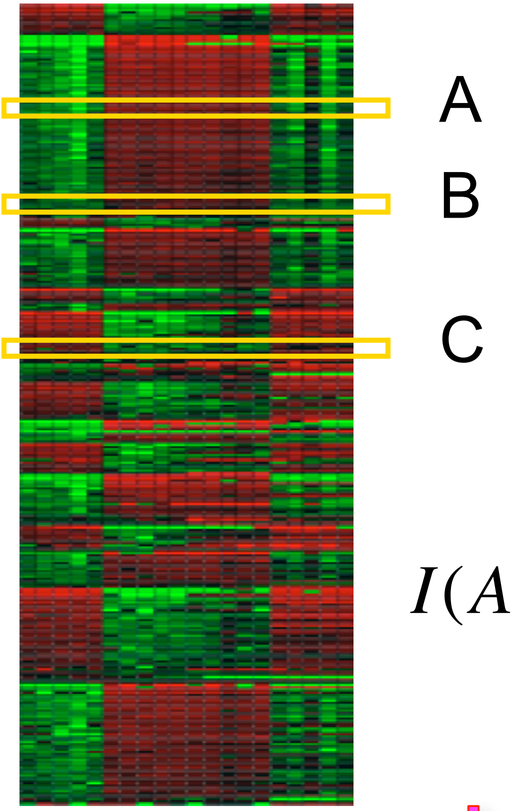
Reconstruction algorithms: The curse of “percent correct”

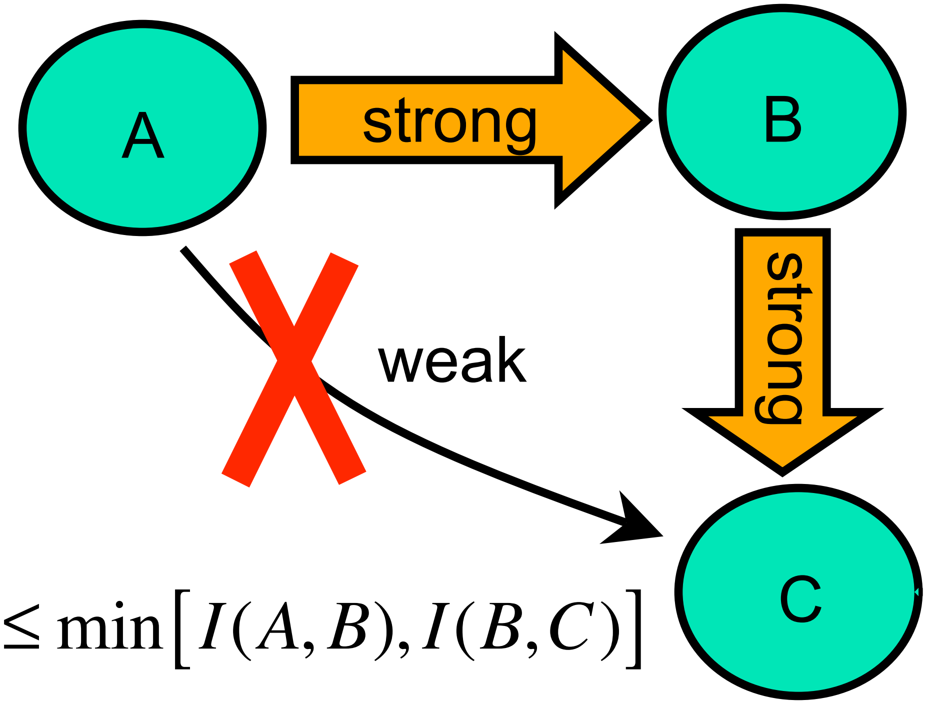


	Stat	Co	GM	Biochem.
Small data requirements	✖✔	✔	✖✔	✖
Robustness to fluct.	✔	✔	✖✔	✖
Computational complexity	✖	✔	✖	✖✔
Conditional interactions	✔	✖✔	✔	✖✔
Reparam inv., non-param.	✖✔	✖✔	✖✔	✔
Irreducibility	✔	✖	✔	✖

Influenciomics

Influenciomics (steady state)





$$I(A,C) \leq \min\left[I(A,B), I(B,C)\right]$$

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
  - Statistical vs. biological?
- 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

$\rho\big(f(x), g(y)\big) \neq \rho(x, y)$ not invariant

! One-to-one transformations of microarray expression data change even signs of the correlations.

Entropy (unique measure of randomness, in bits)

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

$$0 \le S[X] \le \log K$$ (number of "bins")

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

Defining influence: Mutual Information

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

$$= S[X] + S[Y] - S[X,Y]$$

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

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

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, in prep.)

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.

$$p_x = q_x$$

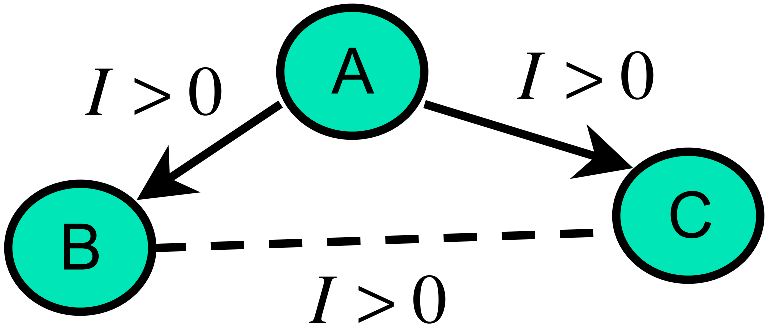
$$p_y = q_y$$

$$\Downarrow$$

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

$I[X;Y] = D_{KL}[P \| Q] > 0 \Rightarrow$ 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:

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

No irreducible interaction!

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

Irreducible interaction.

Higher order influences

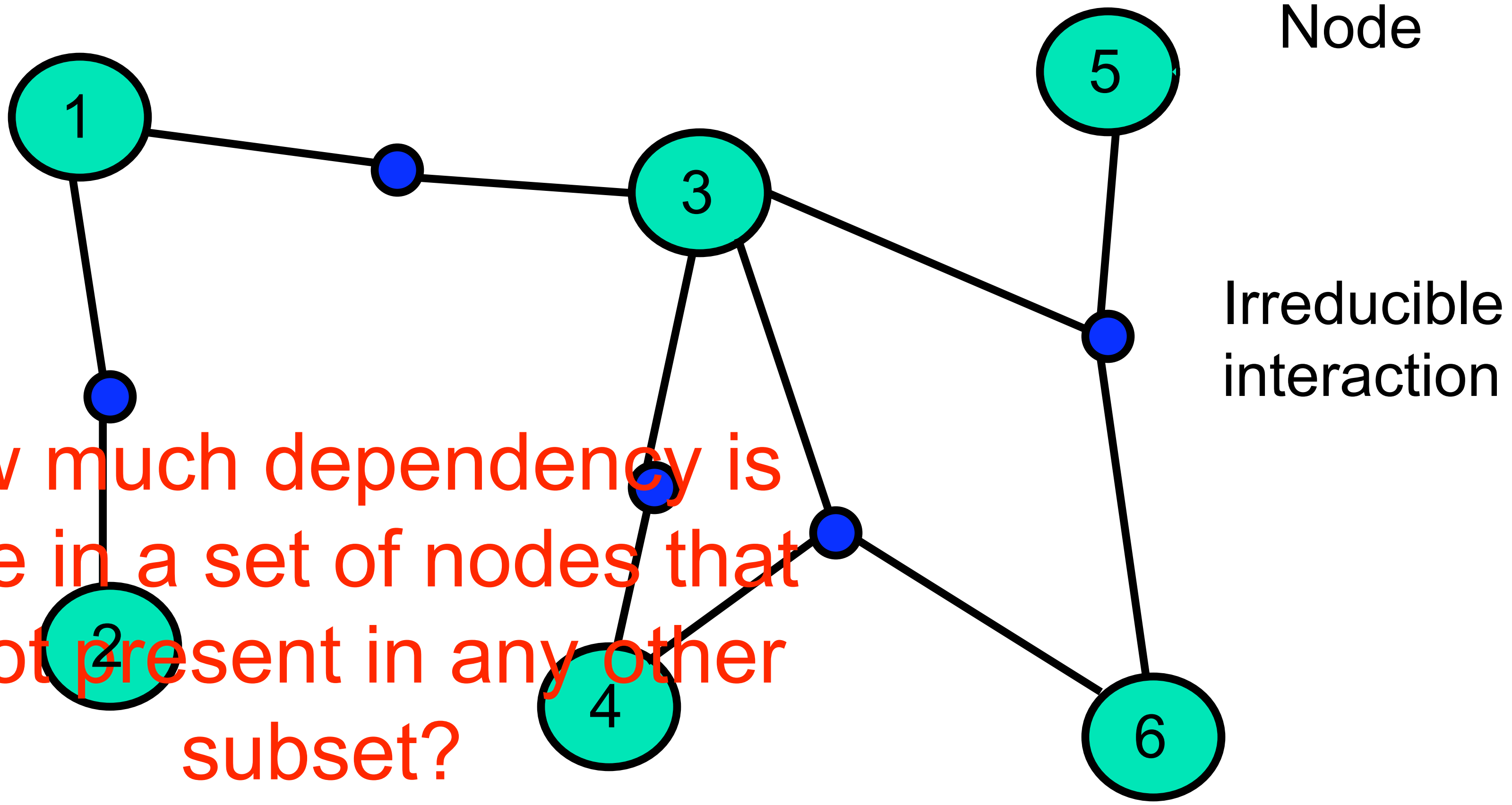
$$I_{XYZ} = \left\langle \log \frac{p_{xyz}}{p_x p_y p_z} \right\rangle$$

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

But these are not irreducible.

(Nemenman and Tishby, in prep.)

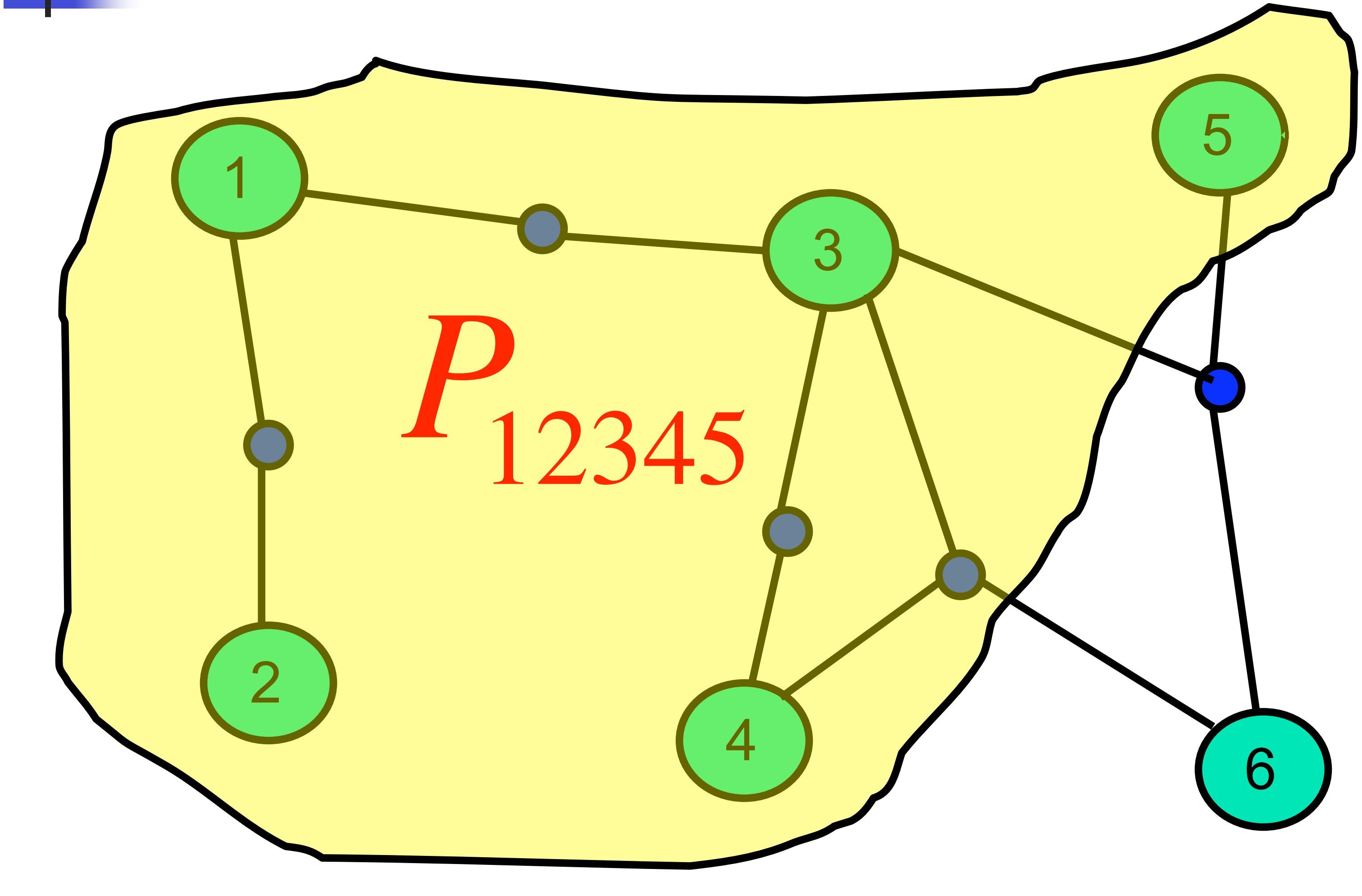
Higher order irreducible dependencies



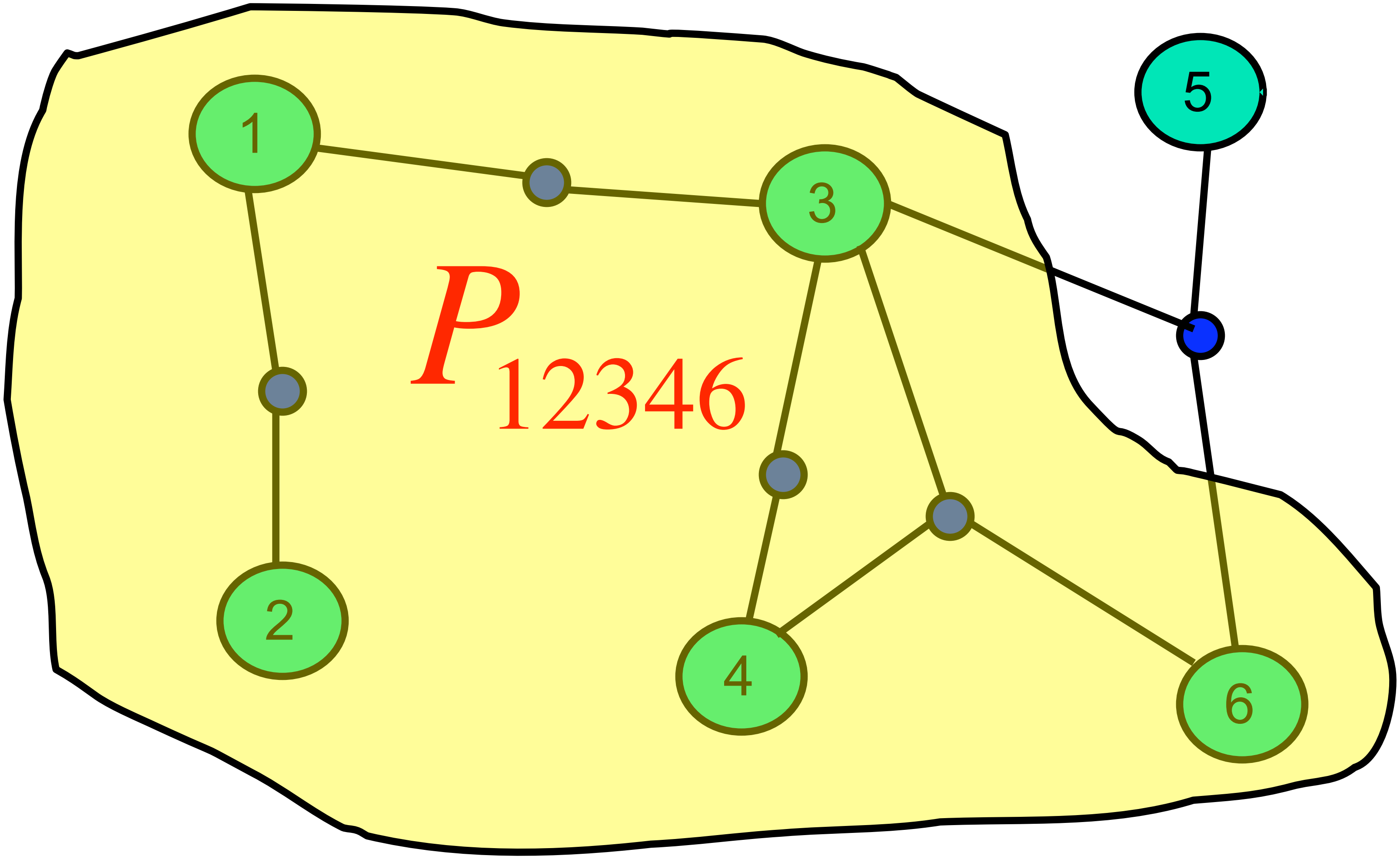
How much dependency is there in a set of nodes that is not present in any other subset?

(Schneidman et al. 2003, Nemenman 2004)

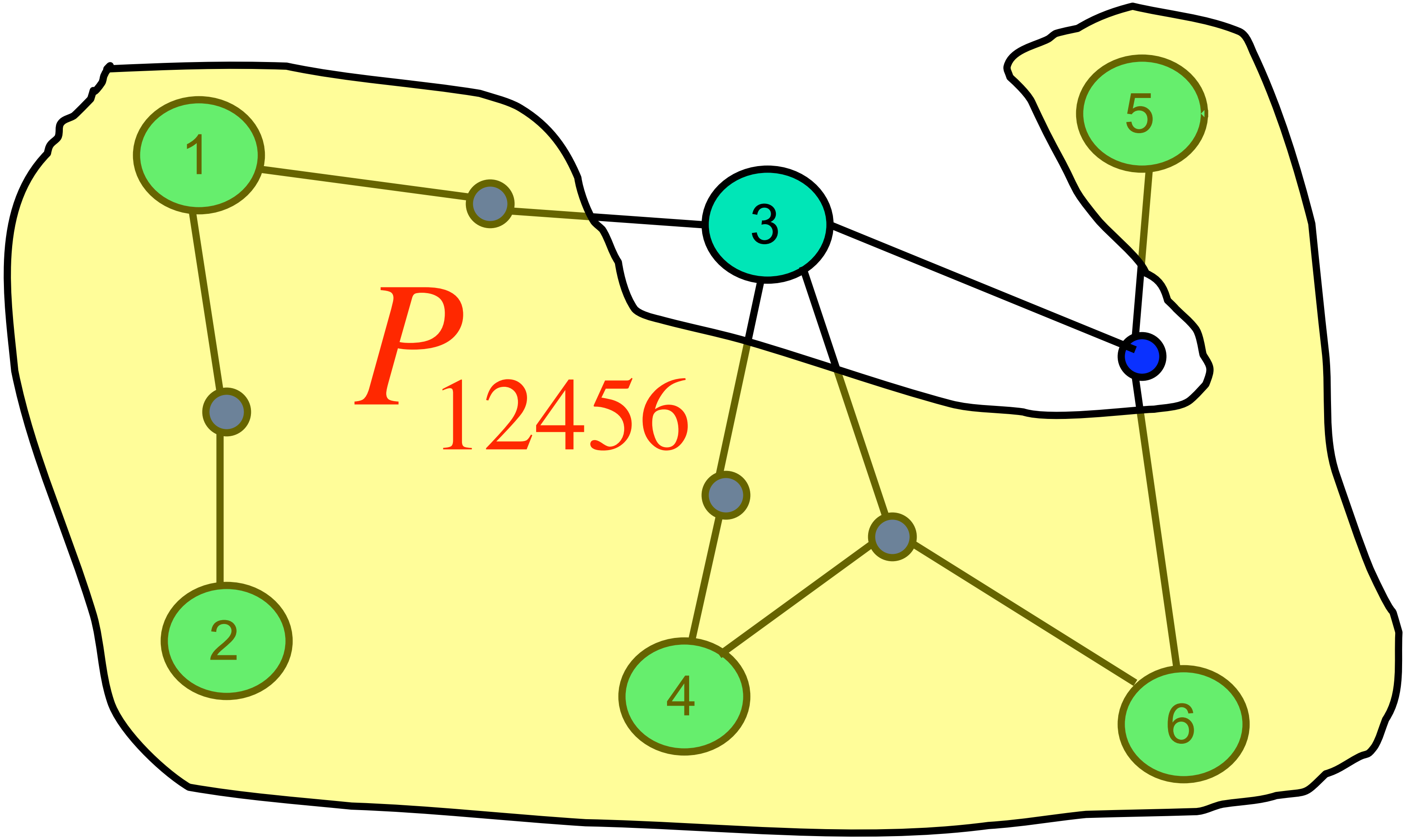
MaxEnt approximations



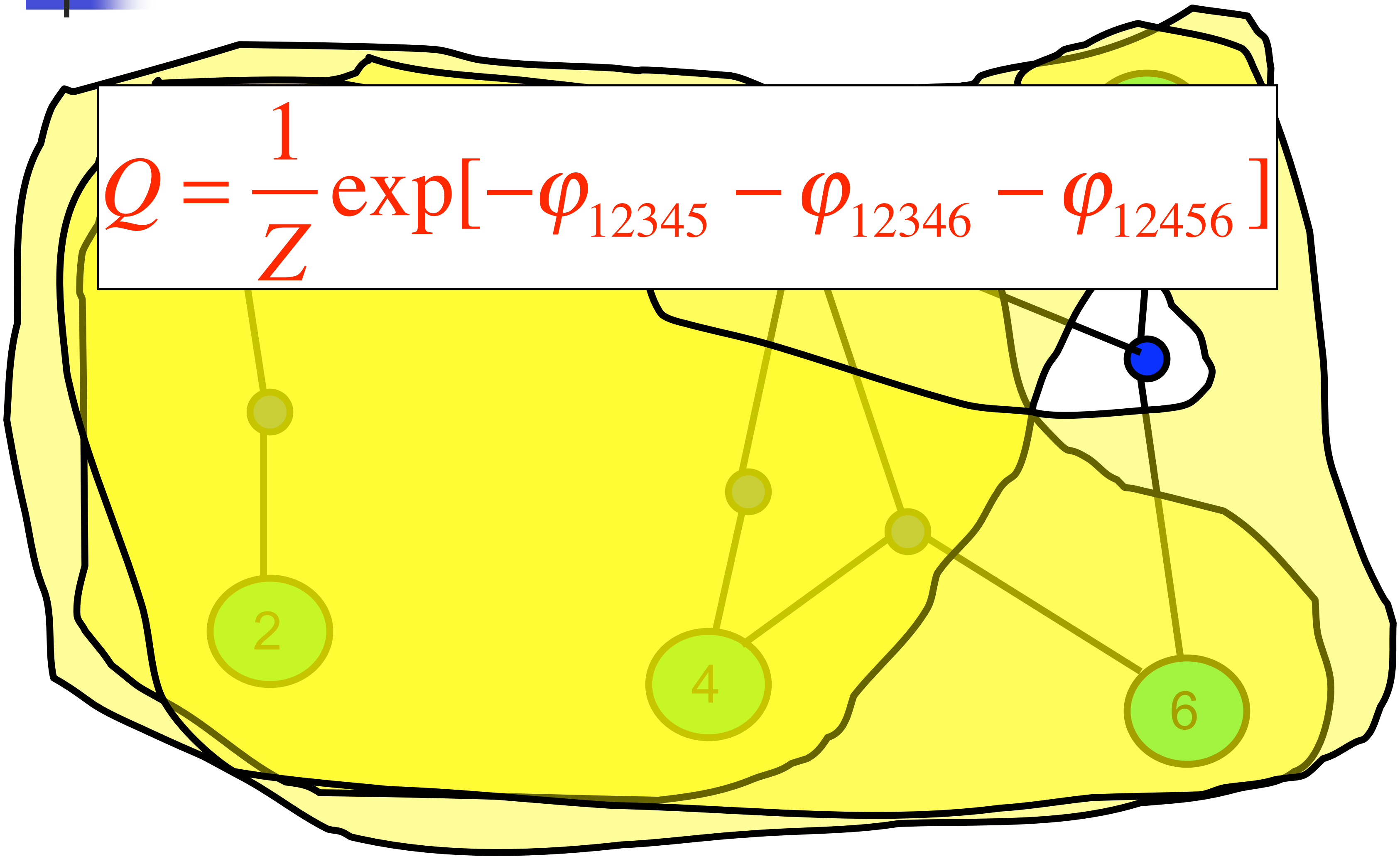
MaxEnt approximations



MaxEnt approximations

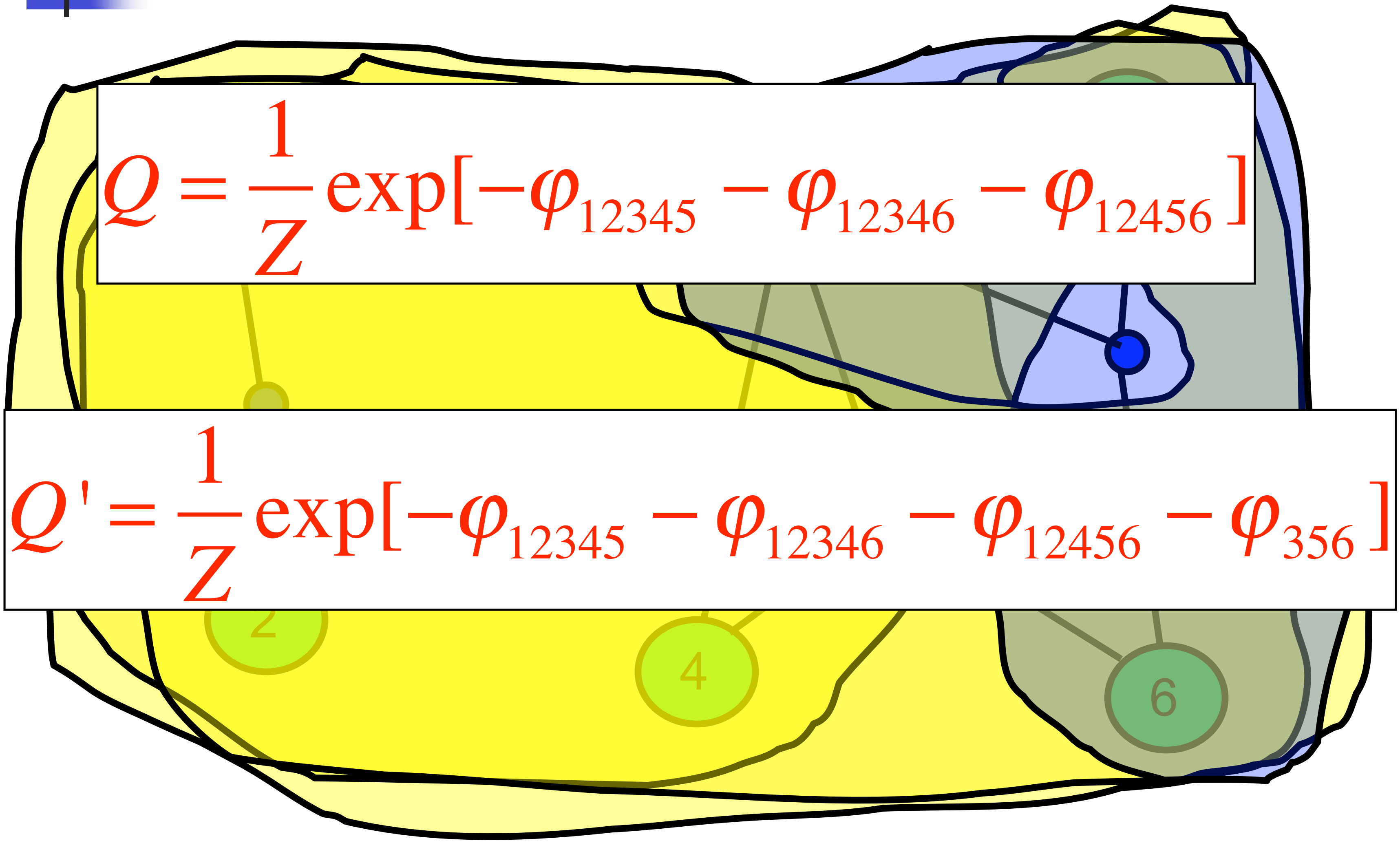


MaxEnt approximations



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

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

$$P(x_1,\ldots x_M)=$$

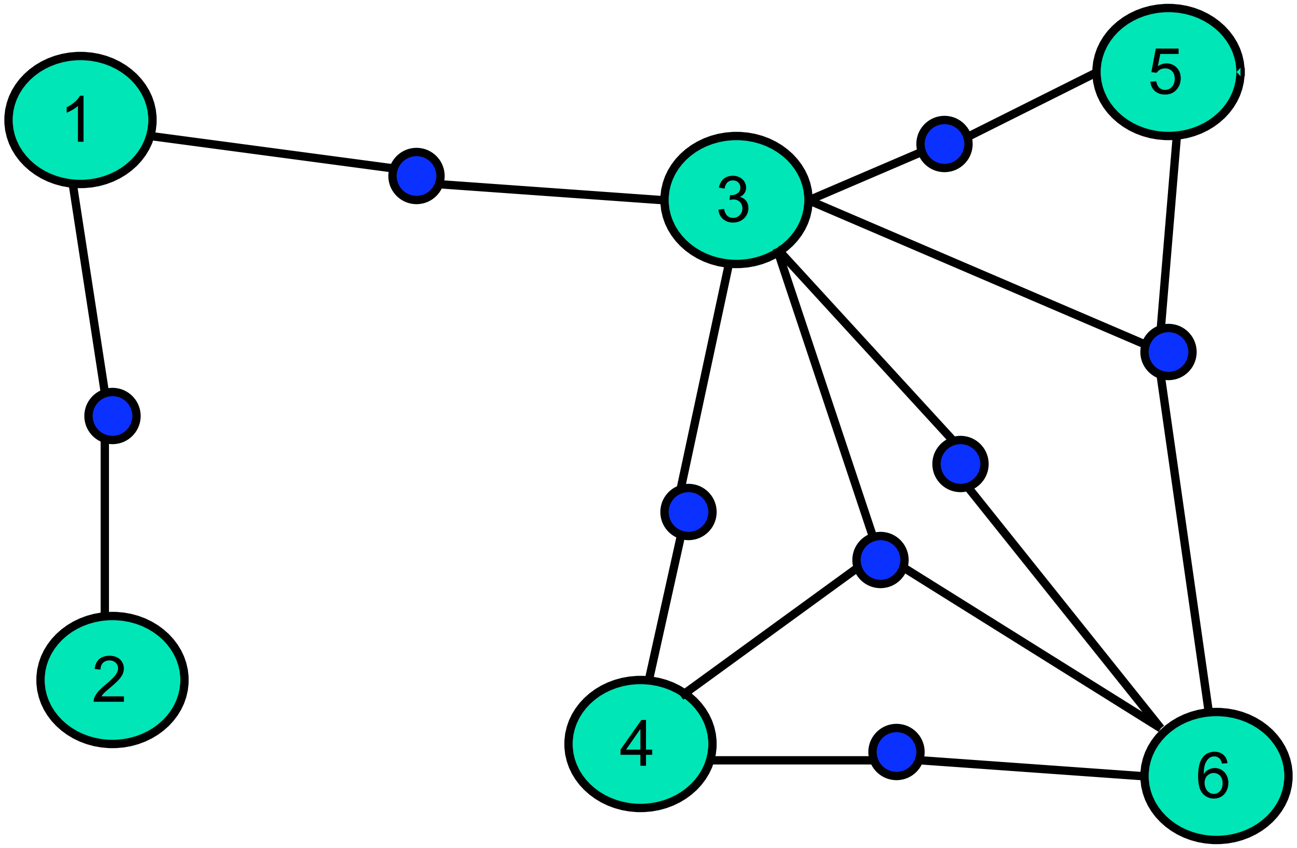
$$=\exp\left[-\sum_i \varphi_i(x_i)-\sum_{ij}\varphi_{ij}(x_i,x_j)-\sum_{ijk}\varphi_{ijk}(x_i,x_j,x_k)-\cdots\right]$$

- *N*-particle potentials
- Spin models -- inverse problem (for discrete variables)
- Random lattices
- Message passing (and if MP works -- ask me later)
- Markov Networks

Two *separate* influenciomics problems

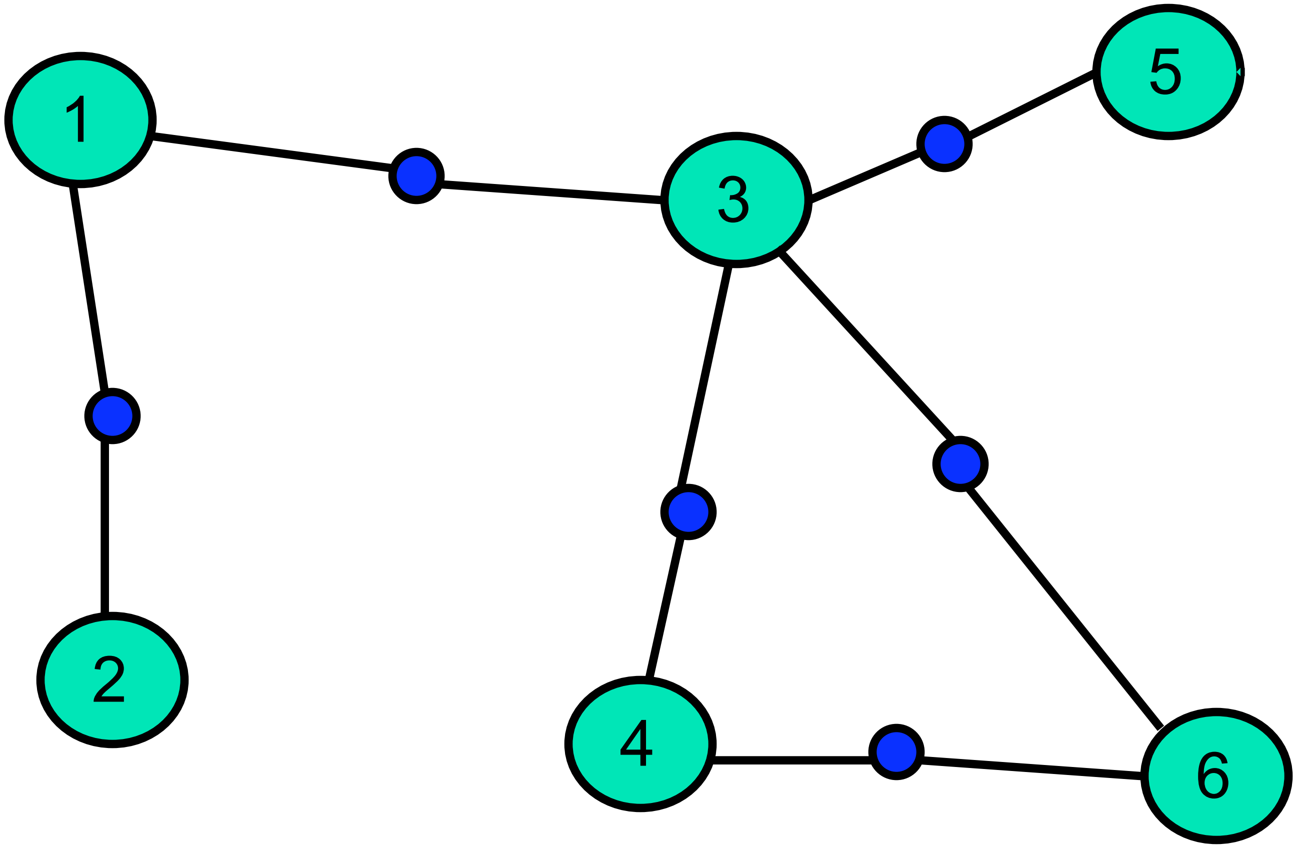
- What is an interaction?
  - What does an arrow mean?
  - Higher order dependencies
- Realistic algorithms to uncover them
  - Controlled approximations (e.g., know the order)
  - Biologically sound assumptions (new knowledge from their verification)
  - Performance guarantees (focus on low false positives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

Interaction network



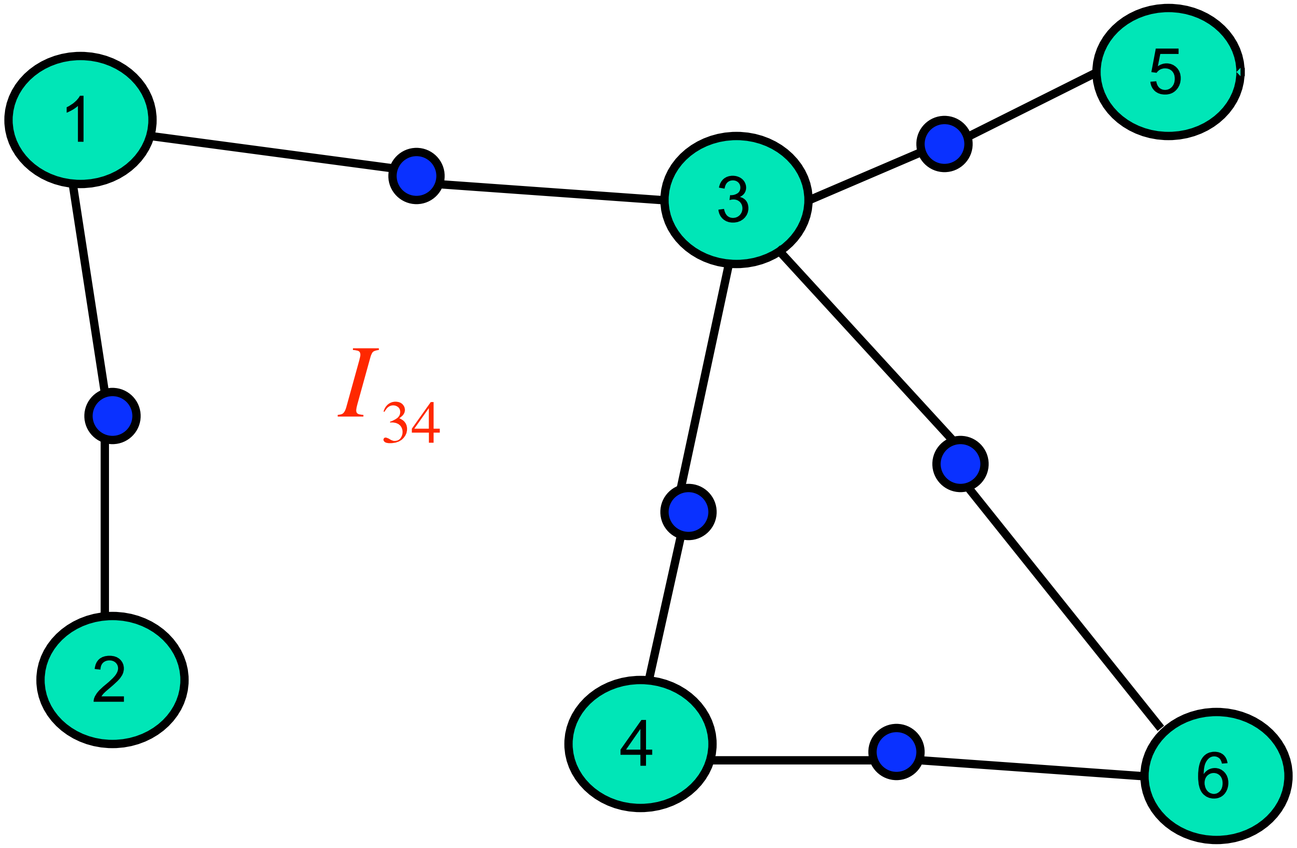
(Basso et al. 2005, Margolin et al. 2005)

Disregard high orders (undersampling)

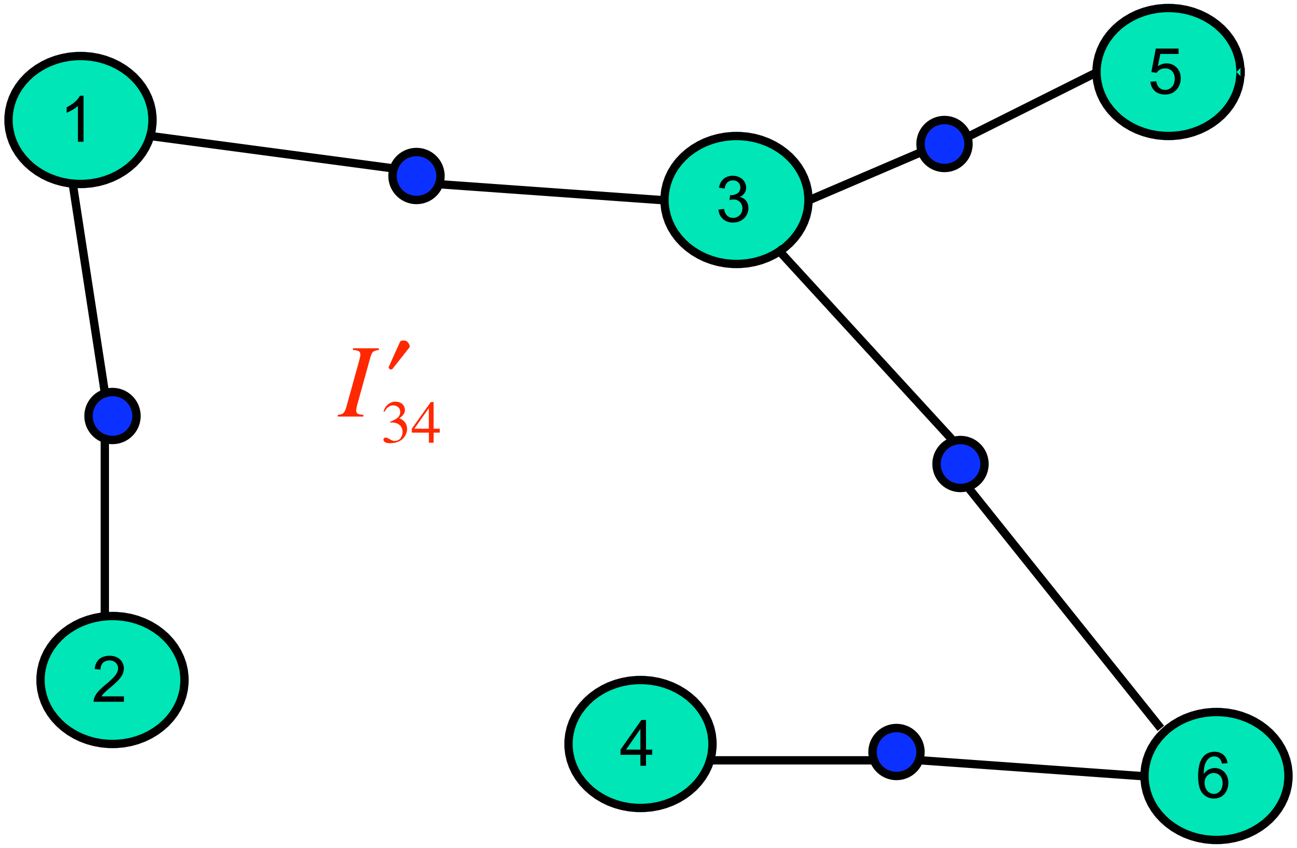


Is second order all we ever need? Cf. Schneidman et al. 2005

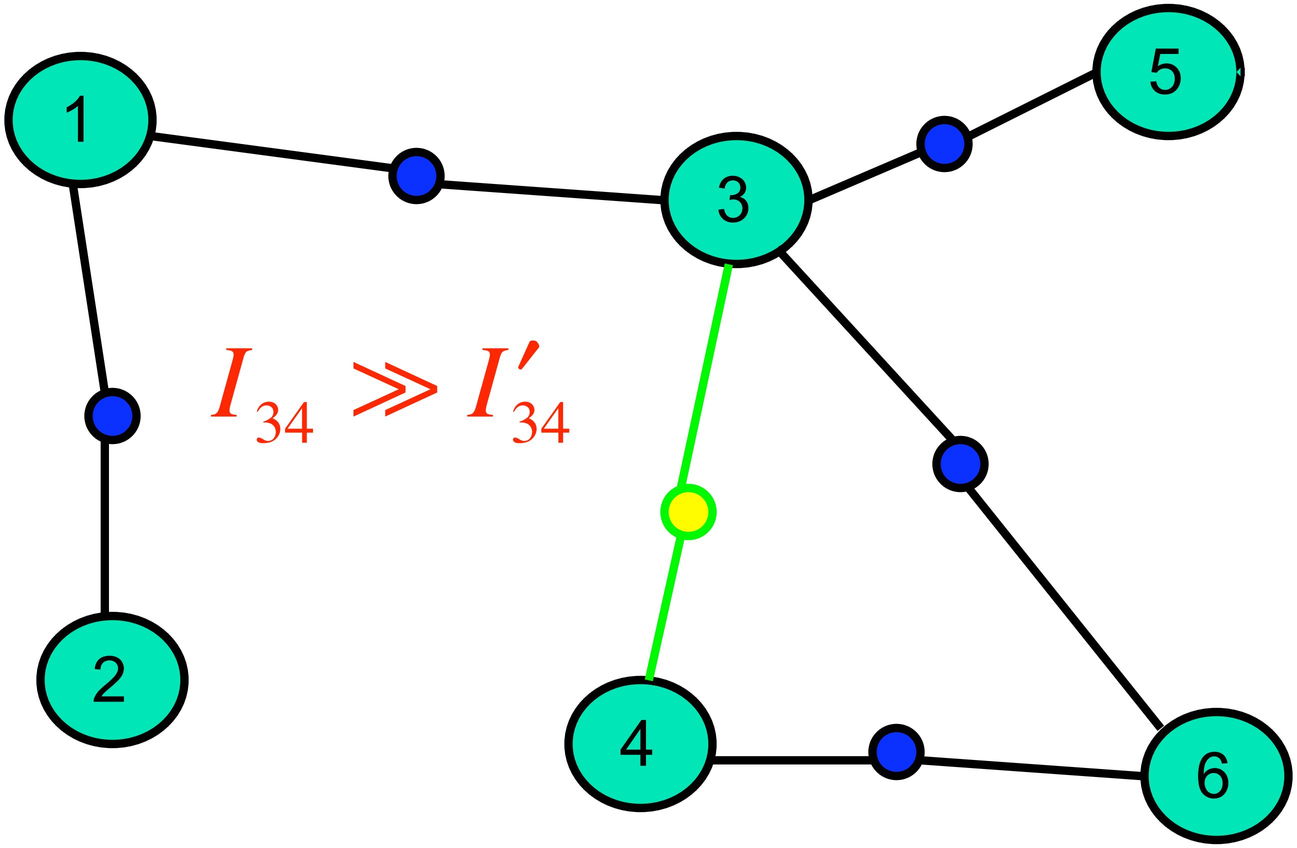
Locally tree-like approximation



Locally tree-like approximation



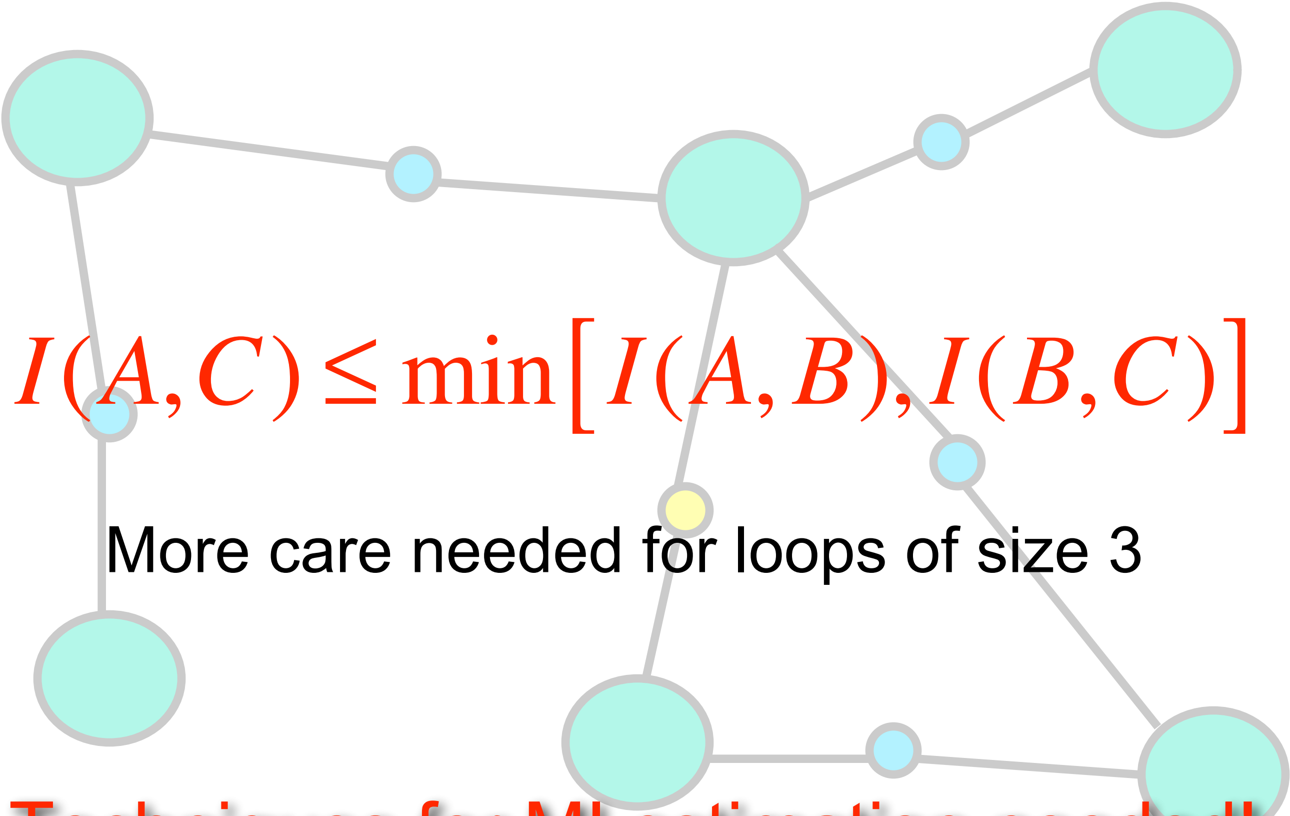
Locally tree-like: signals decorrelate fast



$$I_{34} \gg I'_{34}$$

Conjecture: Message passing works = locally tree-like

ARACNE: remove the weakest link in every triplet



$$I(A,C) \le \min\left[I(A,B), I(B,C)\right]$$

More care needed for loops of size 3

Techniques for MI estimation needed!

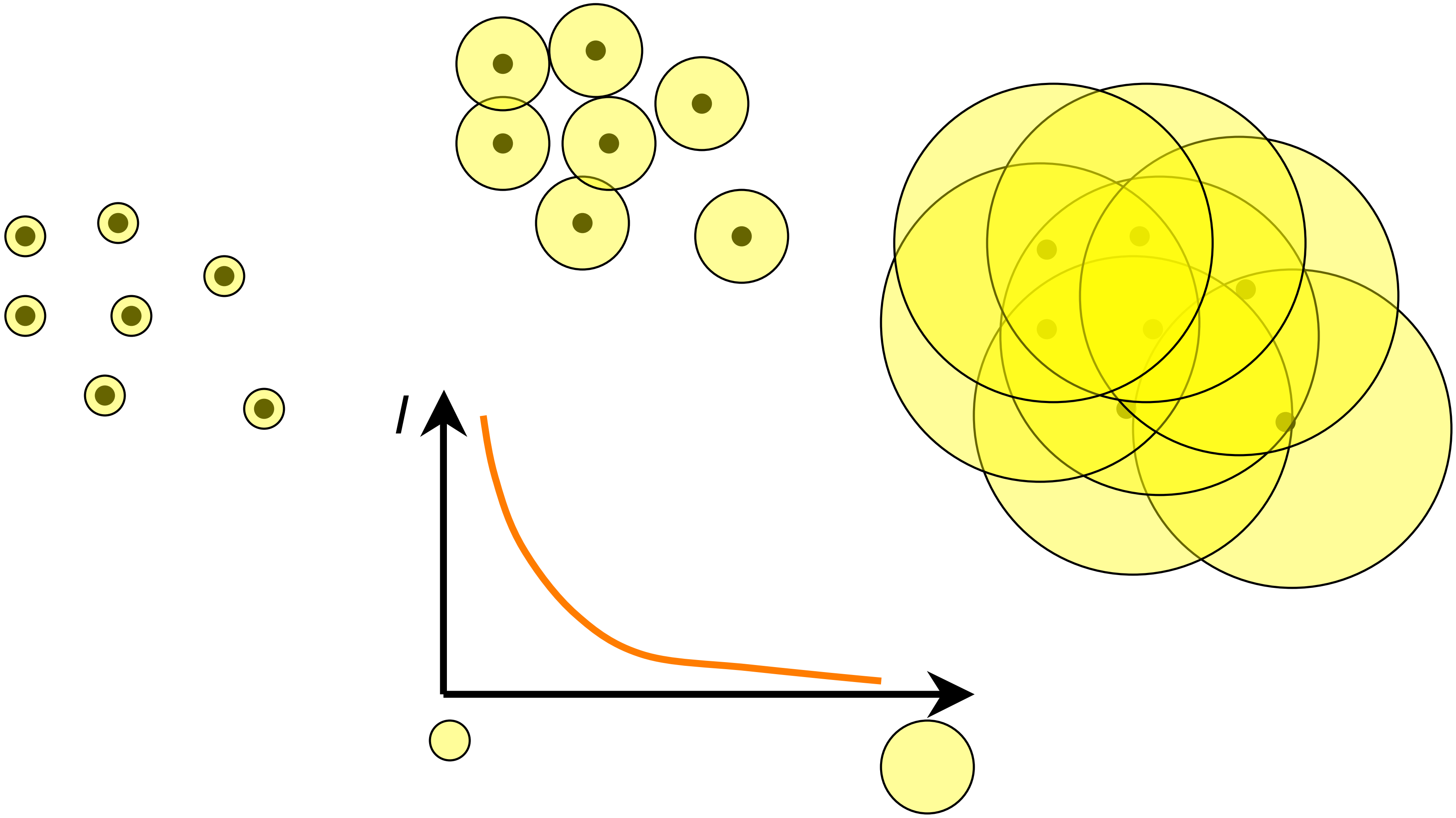
No false positives
Where 2-way -- it's 2-way

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.

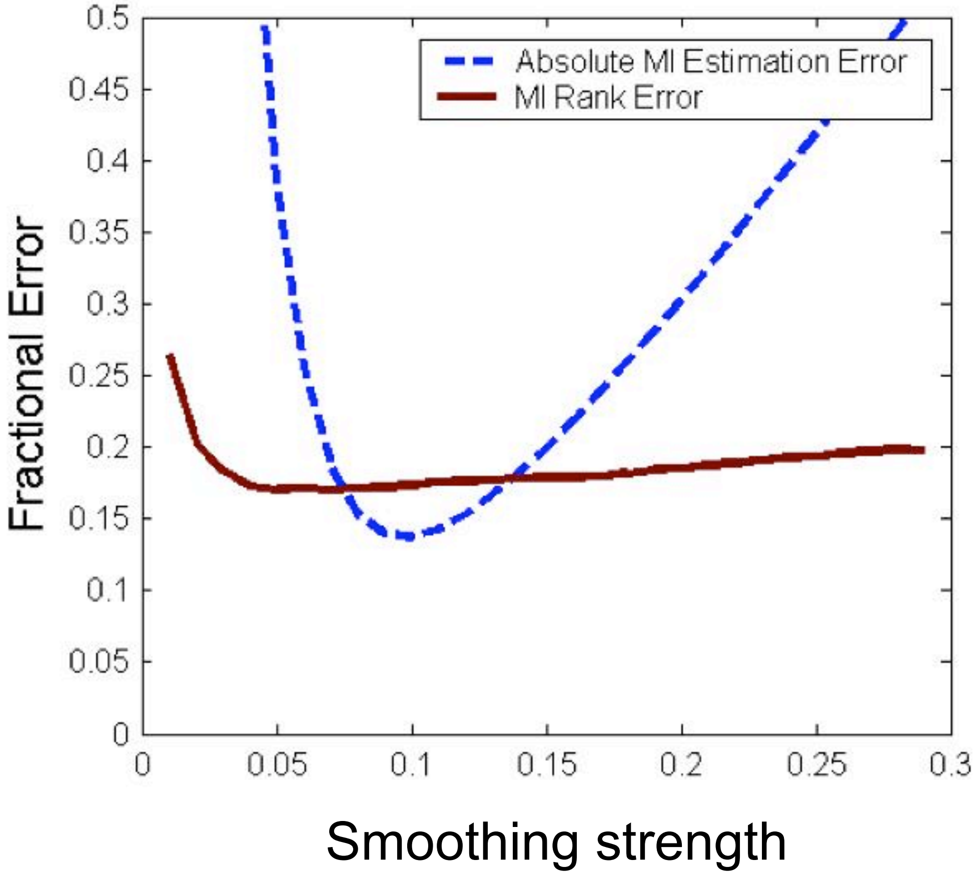
Theorem 2. The Chow-Liu maximum mutual information tree is a subnetwork of the network reconstructed by ARACNE.

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

Estimating *l*: smoothing (e.g., Gaussian Kernels)



Estimating *I*: stability of ranks



Also:

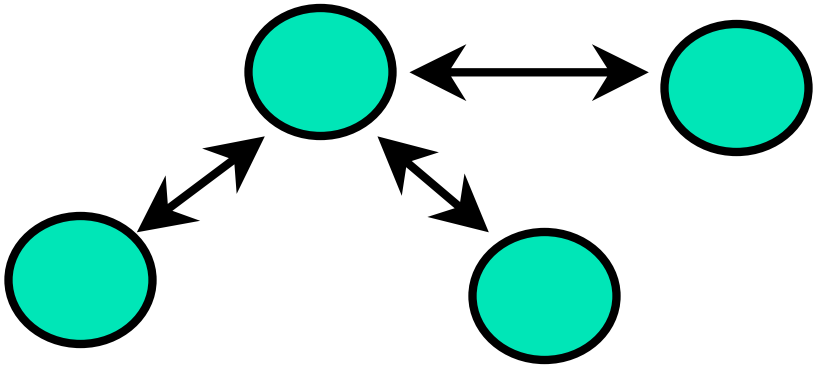
- NSB
- copula

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

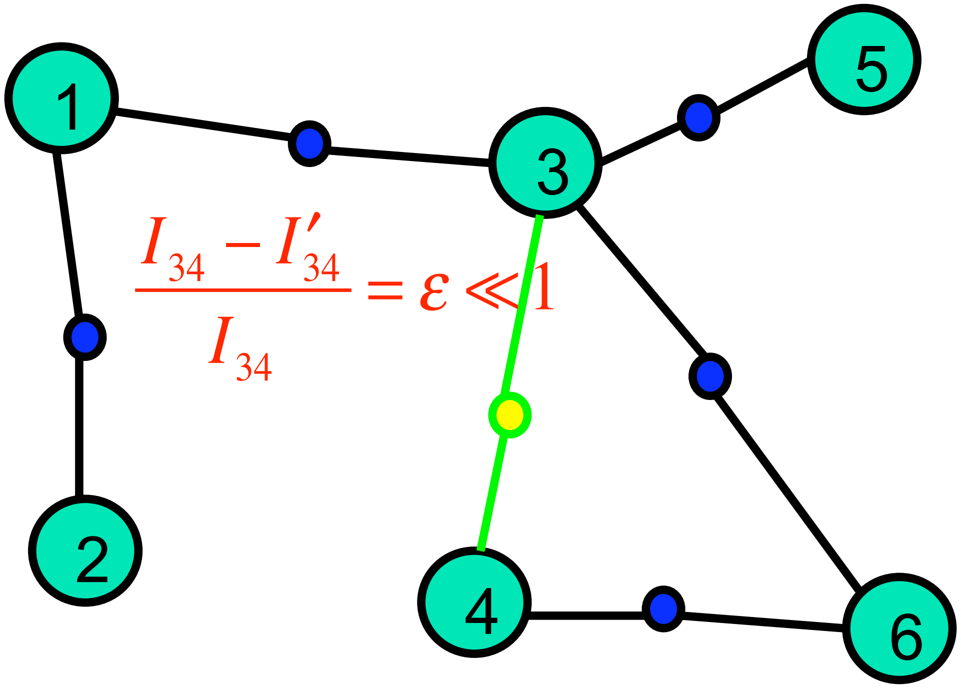
$$P(x_i) = ?$$



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!



$$\frac{I_{34} - I'_{34}}{I_{34}} = \varepsilon \ll 1$$

Biological soundness

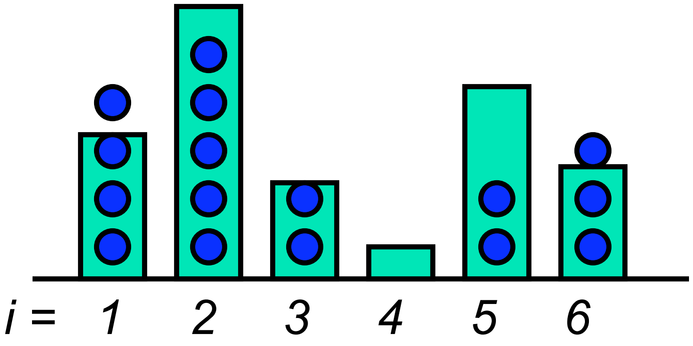
- Higher order interactions project to lower orders
- Fast decorrelation, sparseness: $I(\text{gene,copy}) \gg I(\text{gene,second best})$
- Small loops often transient

Why is IT not common in statistics?

Maximum likelihood estimation:

$p_i,\ \ i = 1\ldots K$

(*K* - # of bins)



$$p_i^{ML} = \frac{n_i}{N}$$

(*N* - sample size)

$$S_{ML} = -\sum_i \frac{n_i}{N}\log\frac{n_i}{N}$$

!

$$\langle S_{ML}\rangle \le -\sum_i \frac{\langle n_i\rangle}{N}\log\frac{\langle n_i\rangle}{N} = S$$

Why is IT not common in statistics?

$$\langle S_{ML} \rangle \le -\sum_i \frac{\langle n_i \rangle}{N} \log \frac{\langle n_i \rangle}{N} = S$$

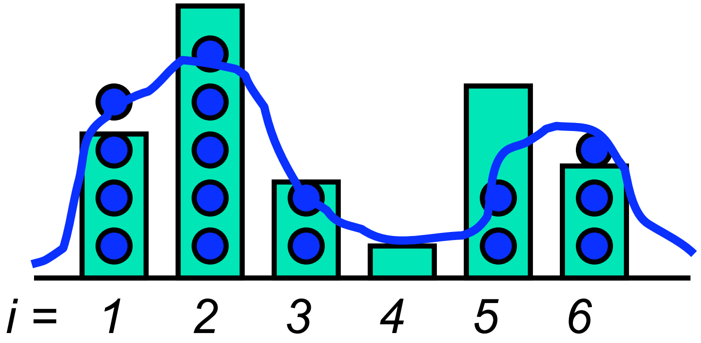
log *K*

$$\text{bias} \propto -\frac{2^S}{N} \gg (\text{variance})^{1/2} \propto \frac{1}{\sqrt{N}}$$

Fluctuations underestimate entropies and overestimate mutual informations.

(Need smoothing.)

Correct smoothing possible



$$S \leq \log N$$

(often not enough)

Incorrect smoothing = over- or underestimation.

Developed for problems ranging from mathematical finance to computational biology.

For estimation of entropy at $K / N \leq 1$ see:
Grassberger 1989, 2003, Antos and Kontoyiannins 2002, Wyner and Foster 2003, Batu et al. 2002, Paninski 2003, Panzeri and Treves 1996, Strong et al. 1998

What if *S*>log*N* ?

But there is hope (Ma, 1981):

For uniform *K*-bin distribution the first coincidence occurs for

$$N_c \sim \sqrt{K} = \sqrt{2^S}$$

$$S \sim 2 \log N_c$$

Time of first coincidence

Can make estimates for square-root-fewer samples!

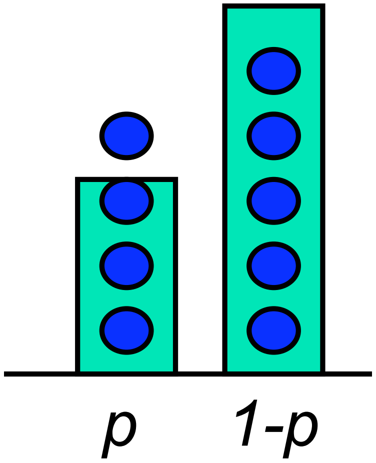
Can this be extended to nonuniform cases?

- Assumptions needed (won’t work always)
- Estimate entropies without estimating distributions.

What is unknown?

Binomial distribution:

$$S = -p \log p - (1-p)\log(1-p)$$



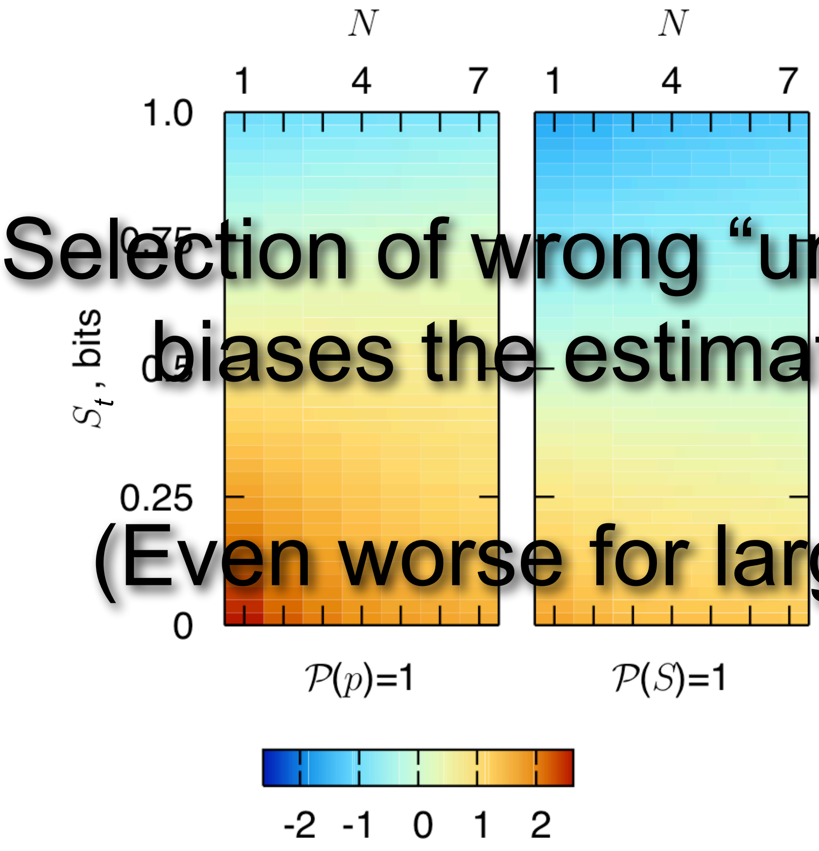
Assume (Bayes)

↓

uniform (no assumptions)

↓ p ↓ S

What is unknown?



$$\varepsilon = \left\langle \frac{S_{est} - S_{true}}{\delta S_{est}} \right\rangle$$

Selection of wrong "unknown" biases the estimation.

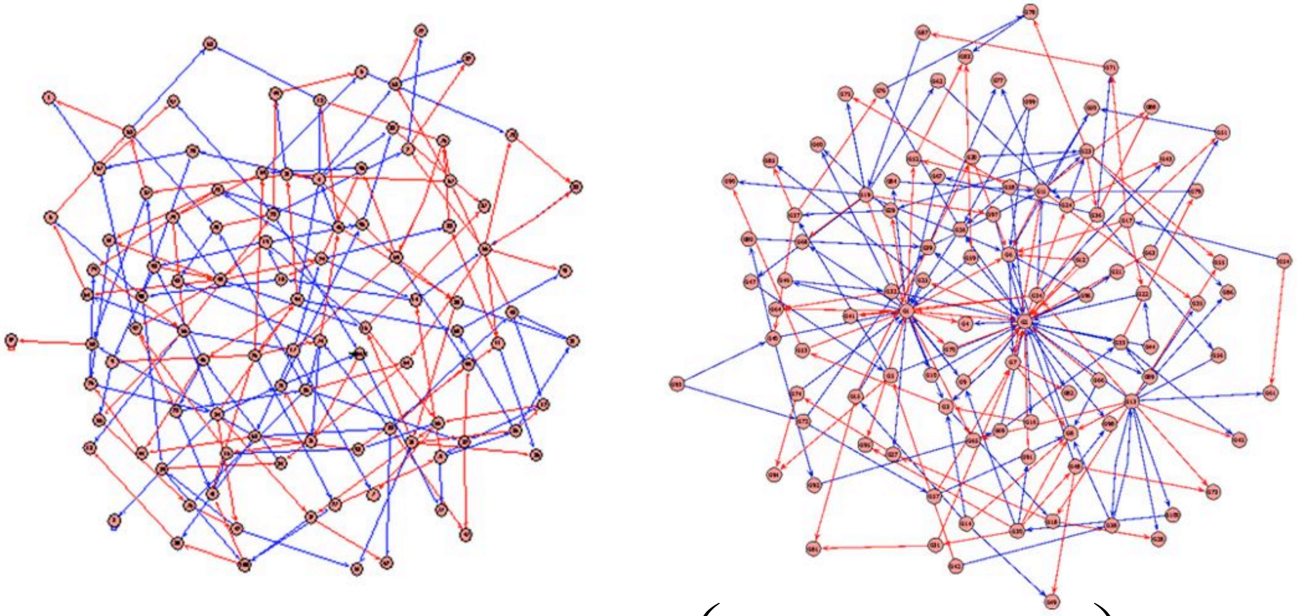
(Even worse for large *K*.)

One possible uniformization strategy for *S* (NSB)

- Posterior variance scales as $1/\sqrt{N}$
- Little bias, except in some known cases.
- Counts coincidences and works in Ma regime (if works).
- Is guaranteed correct for large *N.*
- Allows infinite # of bins.

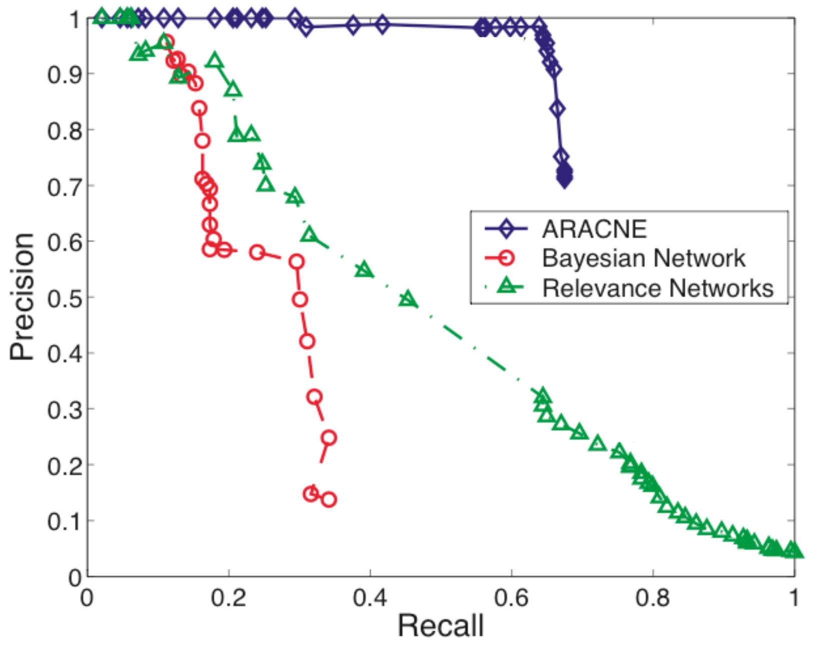
(Nemenman et al. 2002, Nemenman 2003)

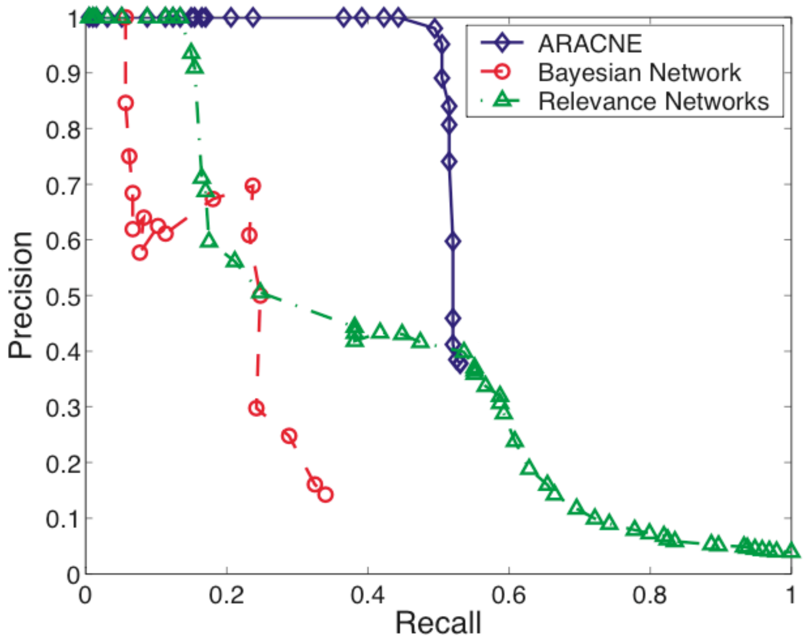
Synthetic networks



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

Synthetic networks (*N*=1000): Biological vs. Statistical Interactions



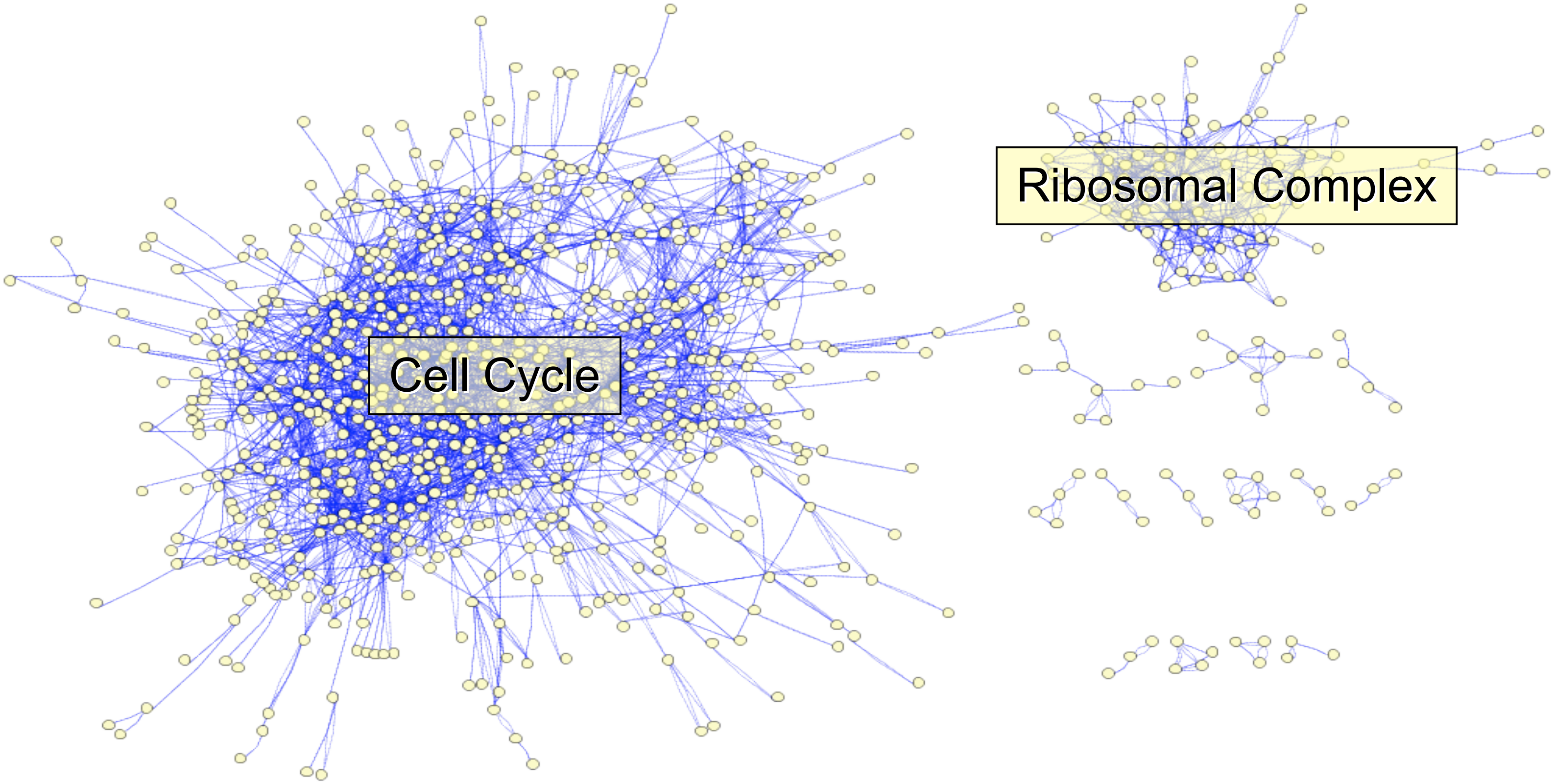


Graceful decay for smaller *N*
Half of all loops kept.

B-cell dataset

- ~400 arrays
- No dynamics
- ~250 naturally occurring, ~150 perturbed
- ~25 phenotypes (normal, tumors, experimental perturbations)
- Expression range due to differential expression in different phenotypes

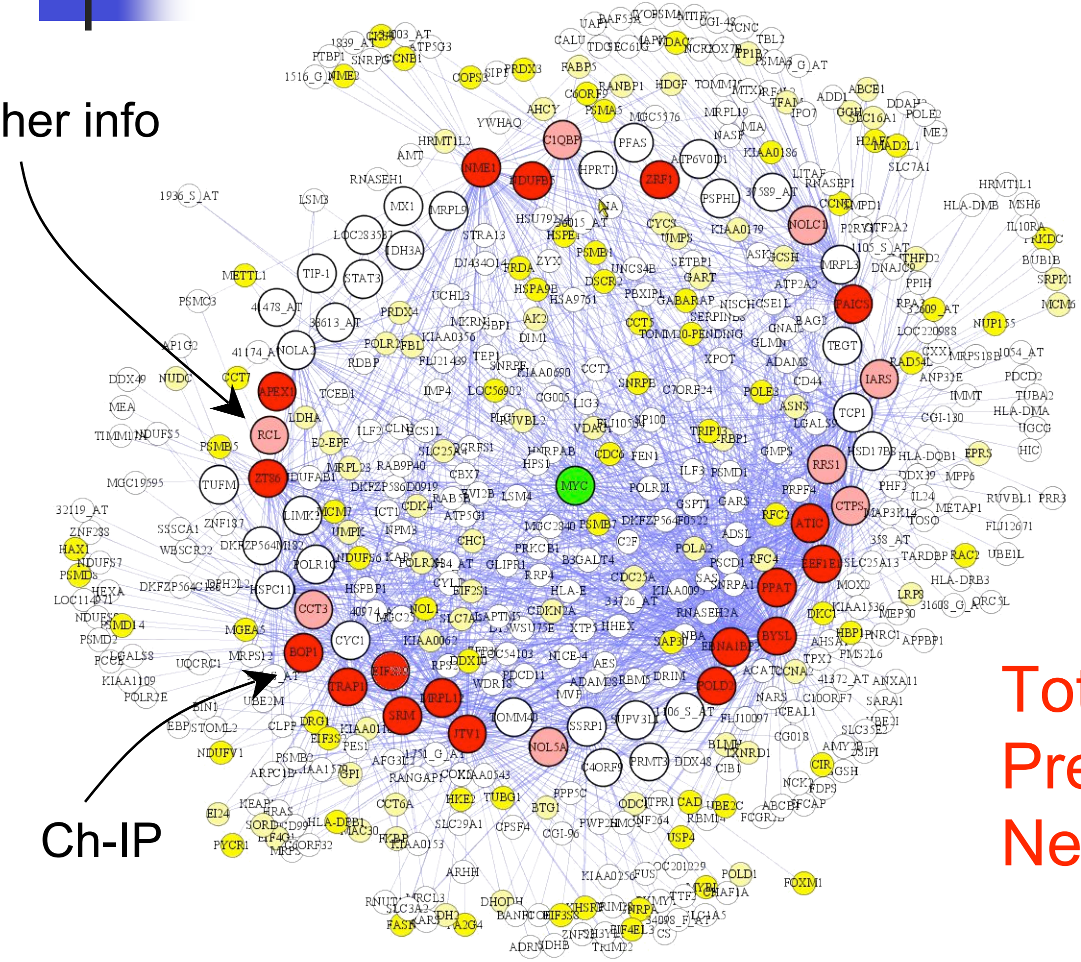
Complete B-cell network



~129000 interactions

c-MYC subnetwork

other info



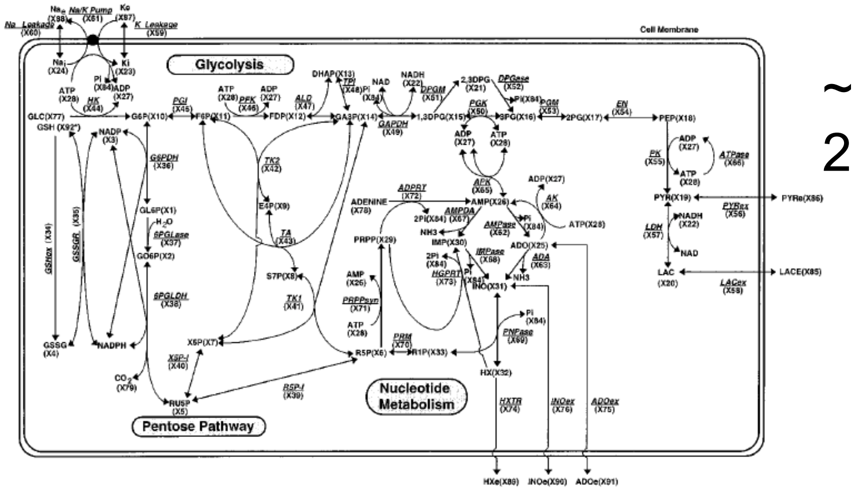
Ch-IP

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

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

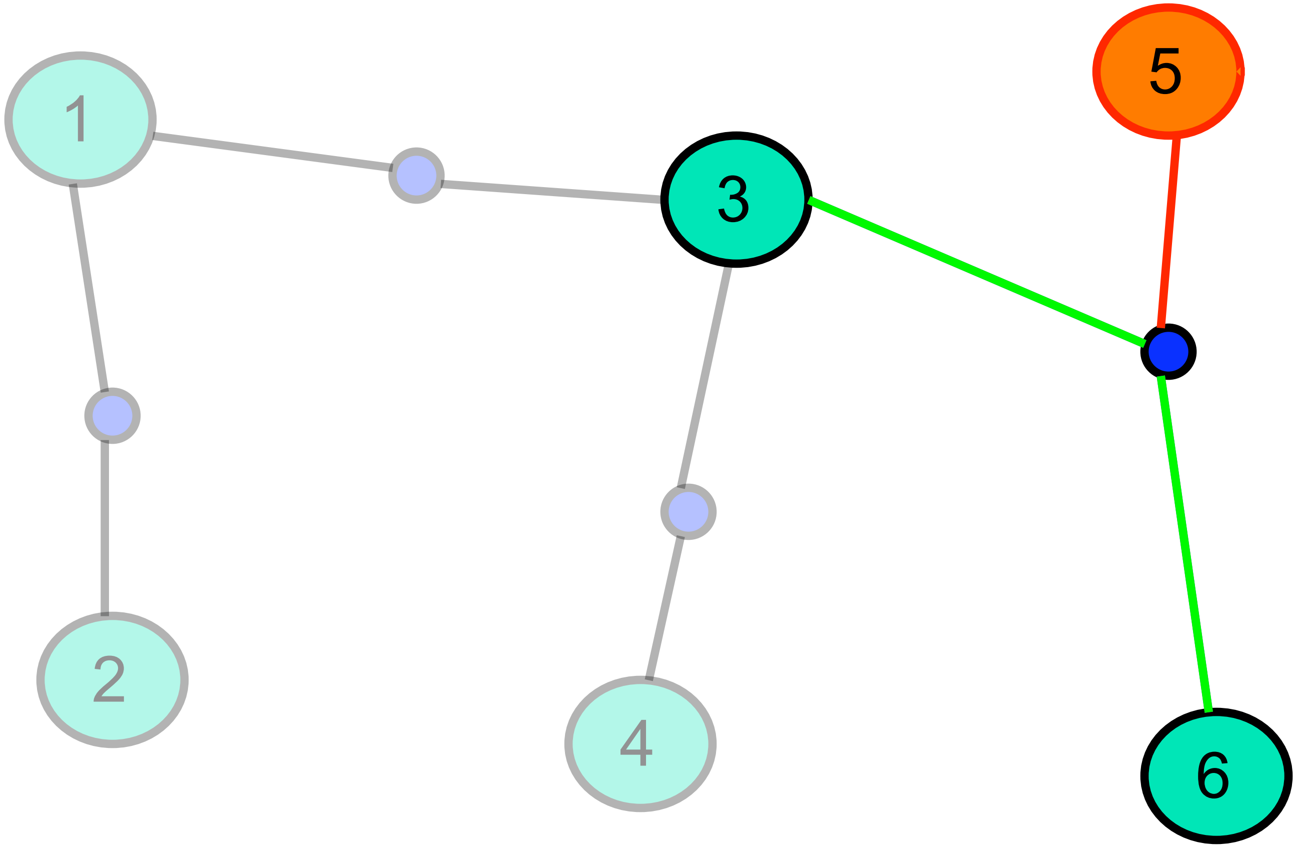
Also validated in…

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



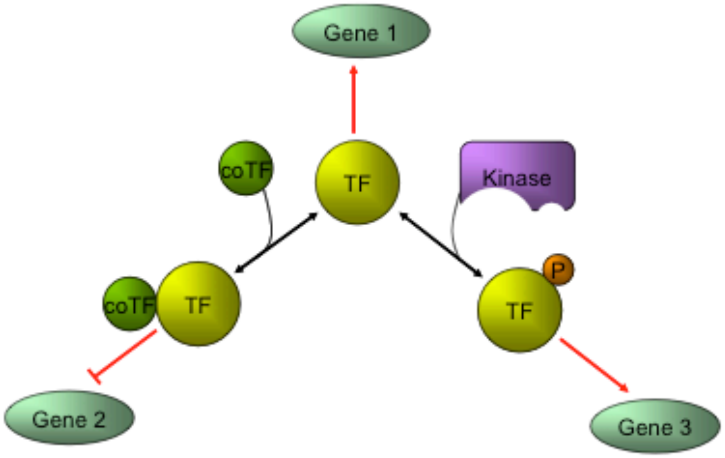
~80% precision
20-80% recall (depending on *N*)

3rd order interactions
(modulated, conditional, transistor)

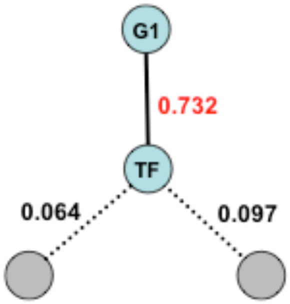


Nontranscriptional modulators from expression data!

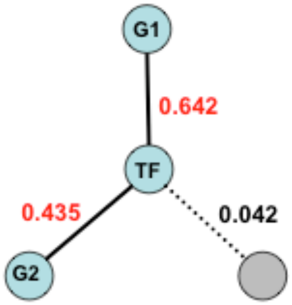
Numerical case study: Non-transcriptional modulation



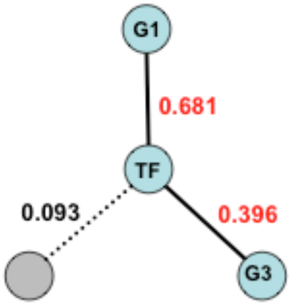
Unconditional



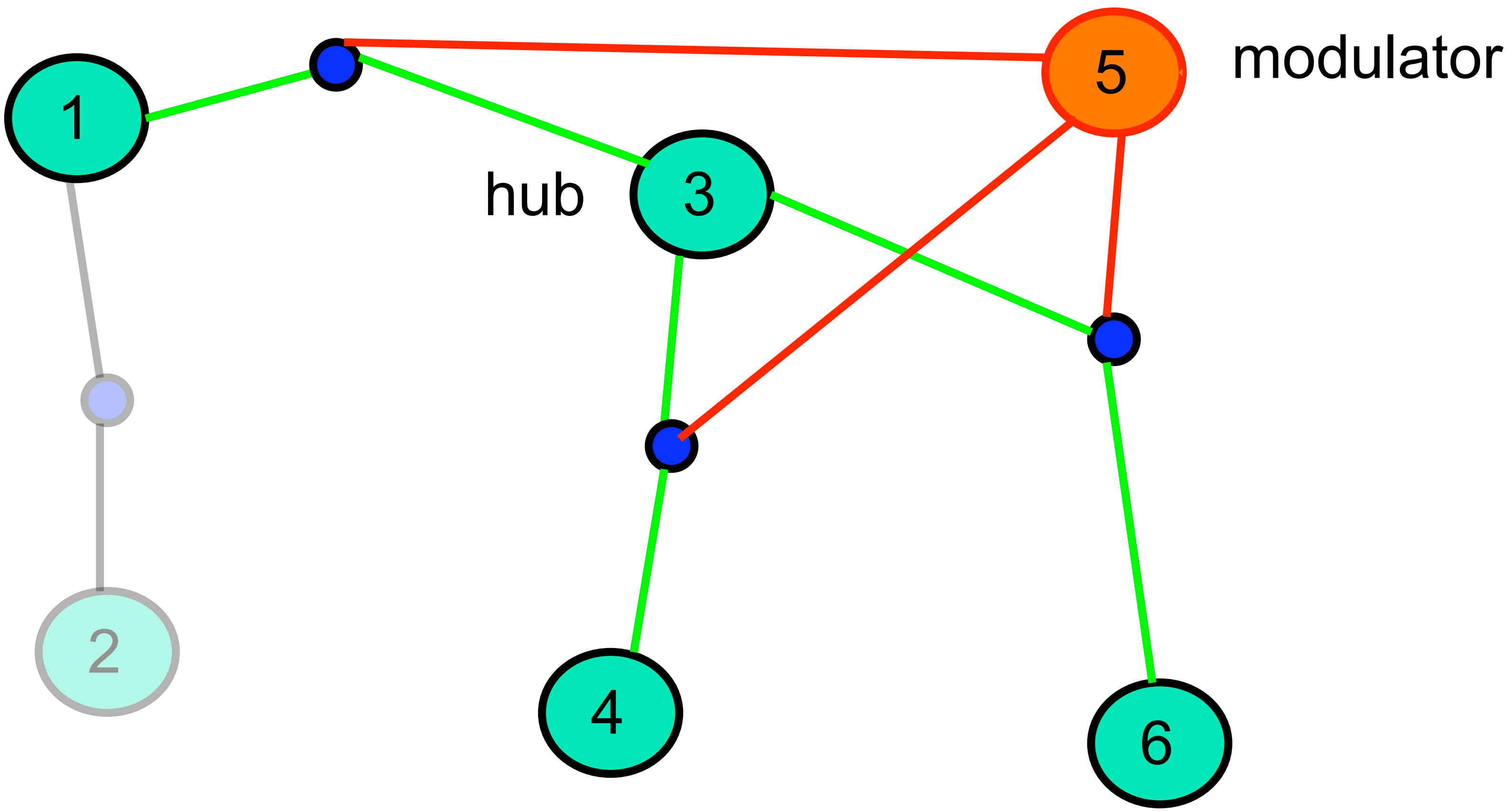
Conditional on coTF



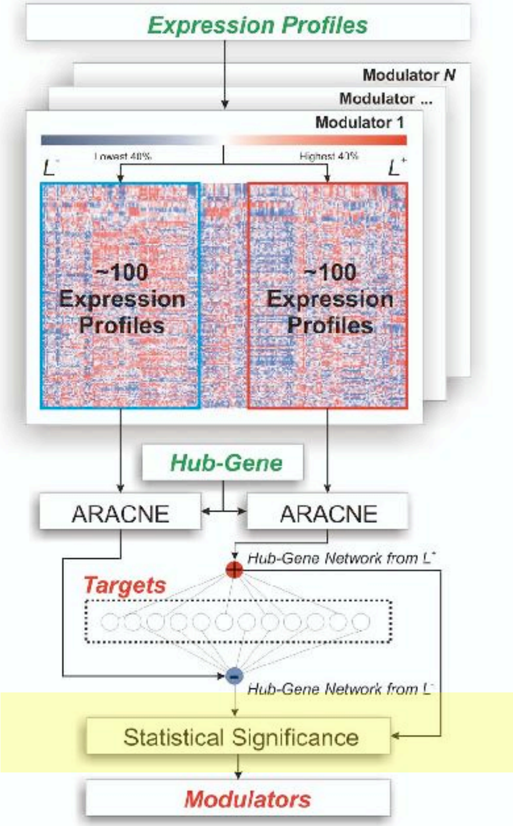
Conditional on Kinase



Large hubs, global (discrete) modulators



Large hubs, global (discrete) modulators



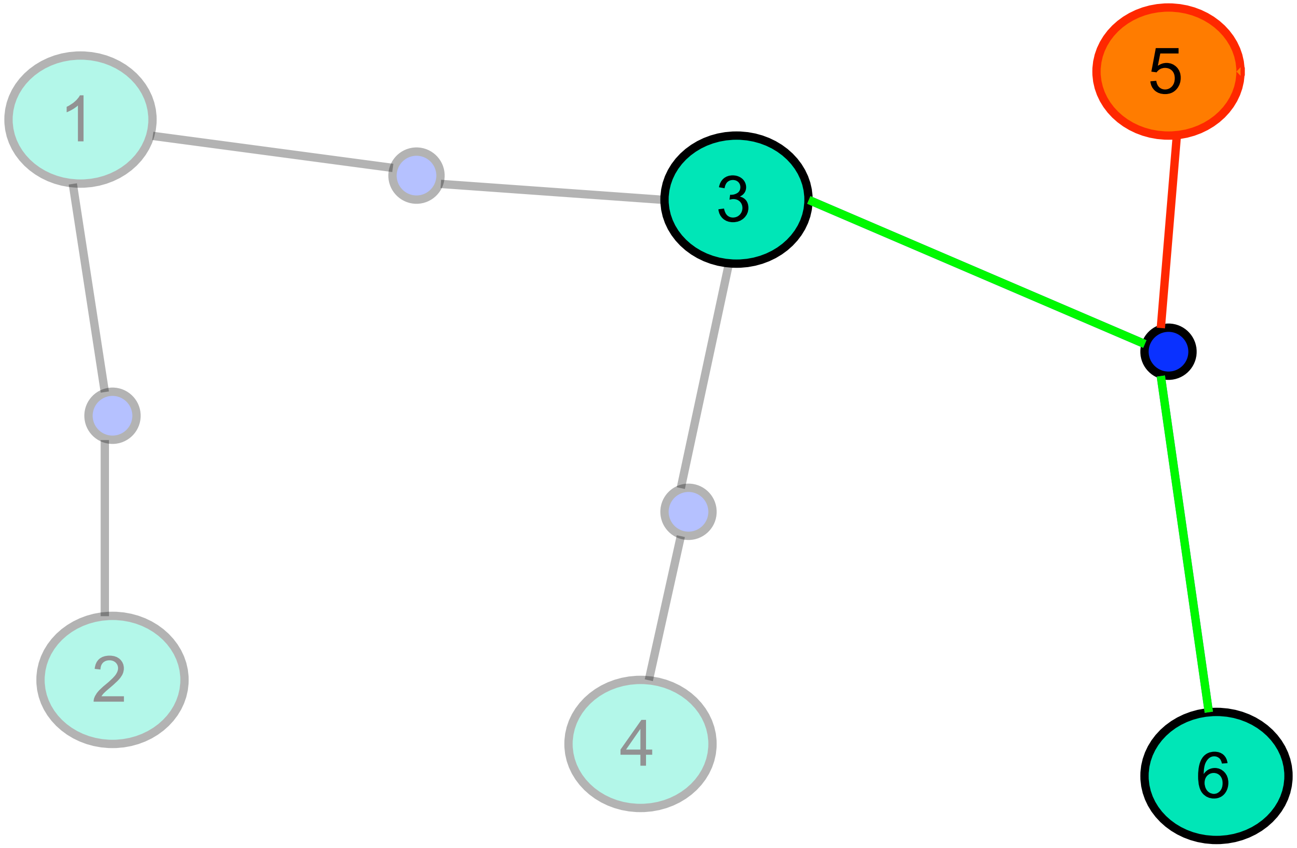
- Focus on important hubs (c-MYC)
- Pre-filter candidate modulators by dynamic range and other conditions.
- Find modulators whose expression inflicts significant changes on topology of the ARACNE hubs' interactions
- No guarantee of irreducibility
- Validate in GO w.r.t. to transcription factors and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

$$\left|N^{+} - N^{-}\right| > 0$$

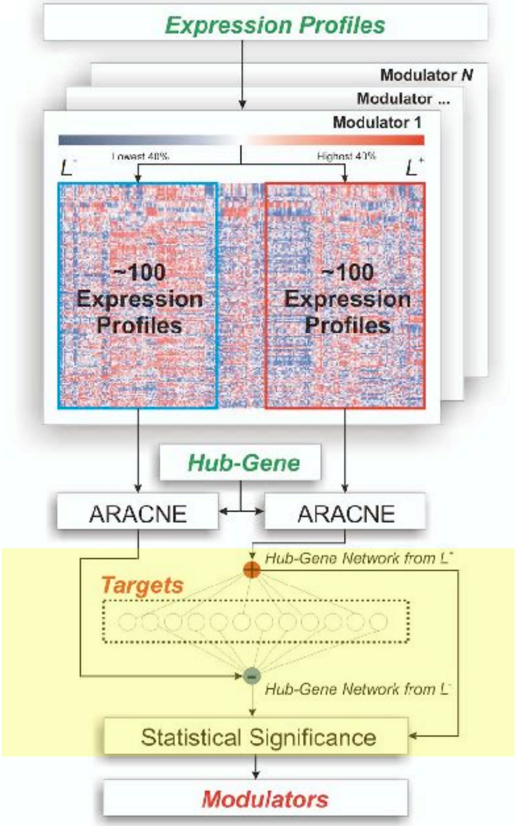
c-MYC modulators

- 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.

Large hubs, local modulator (MI change, transistor)



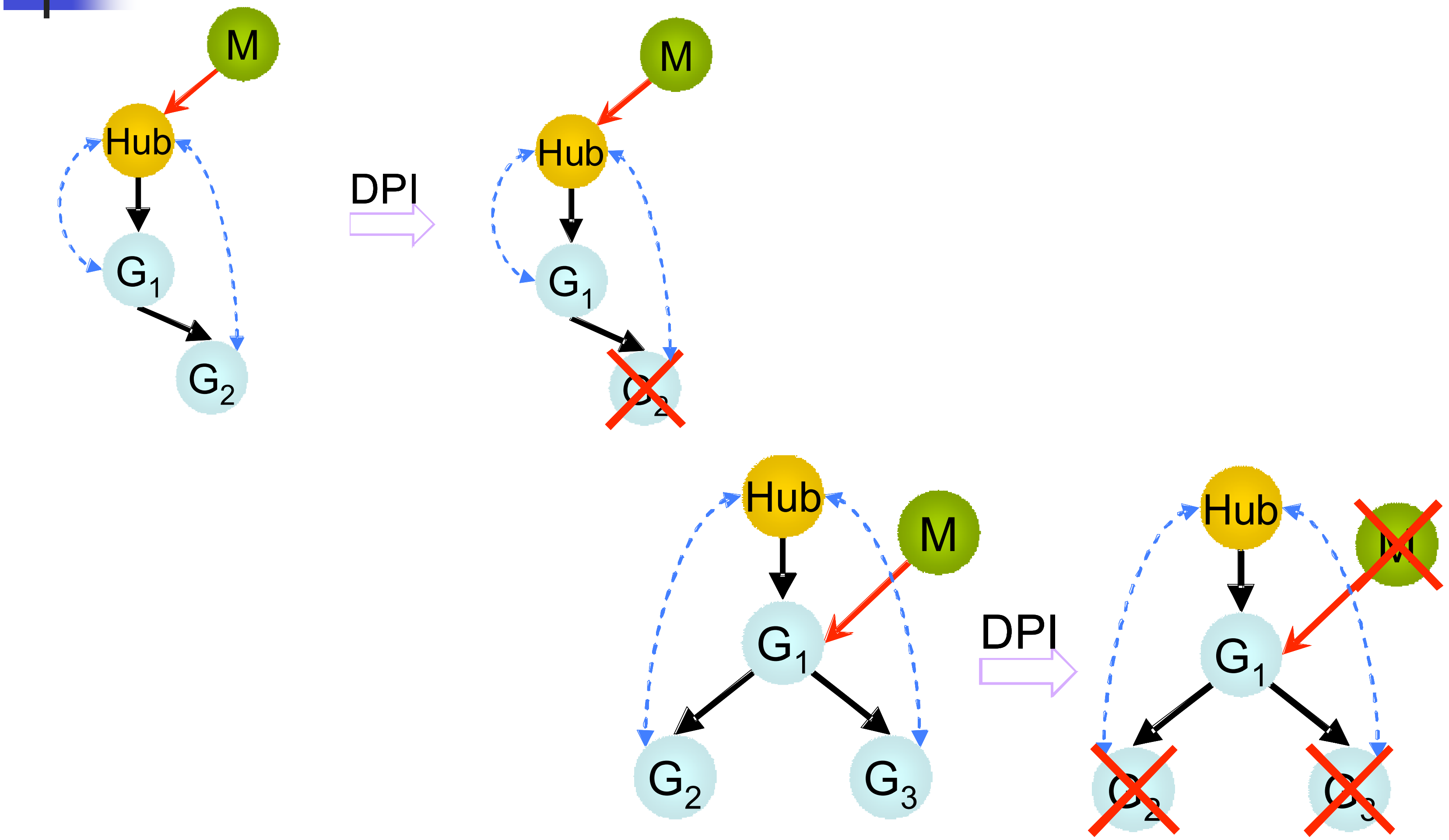
Large hubs, local modulators



- 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

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

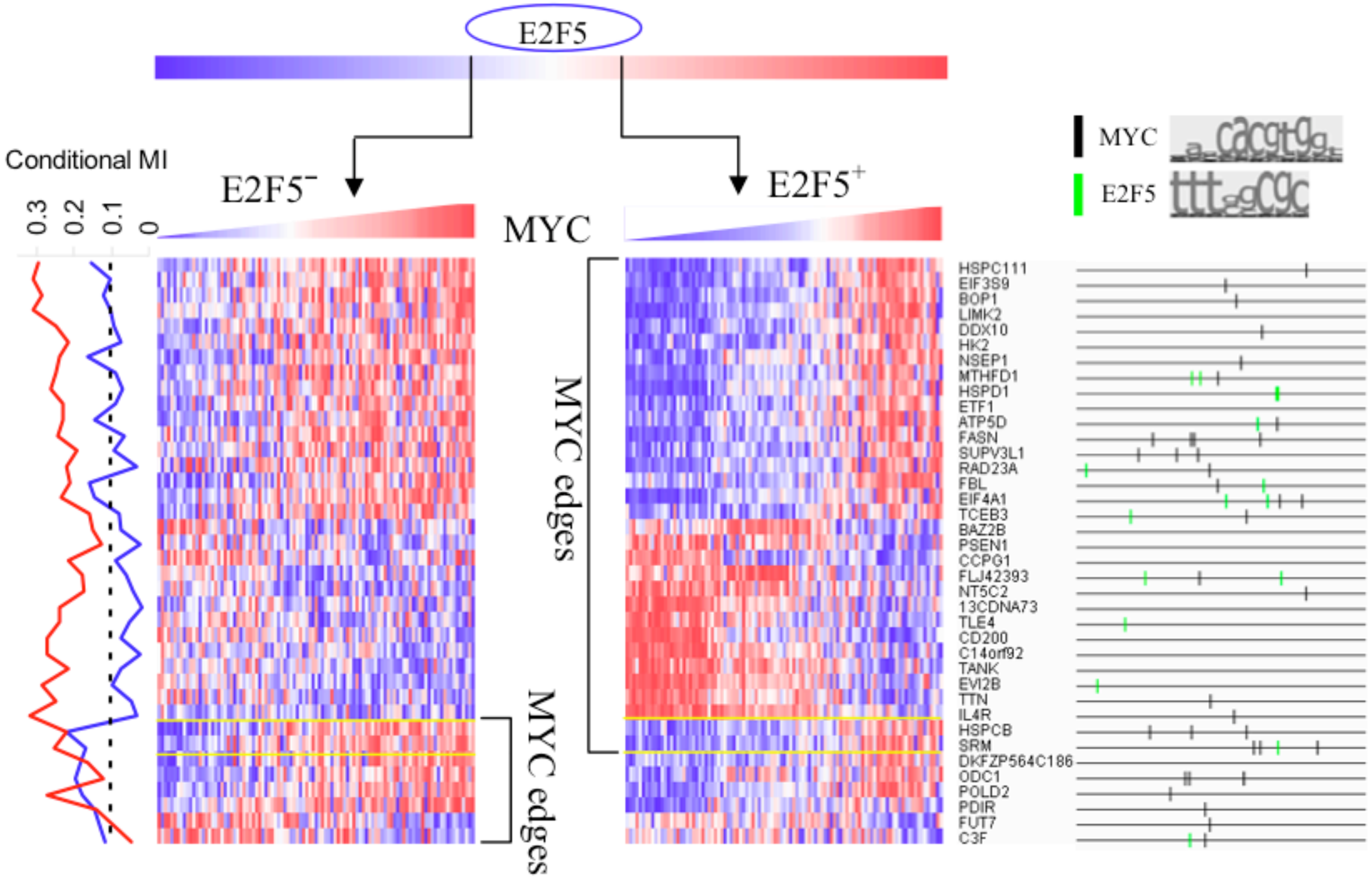
ARACNE helps



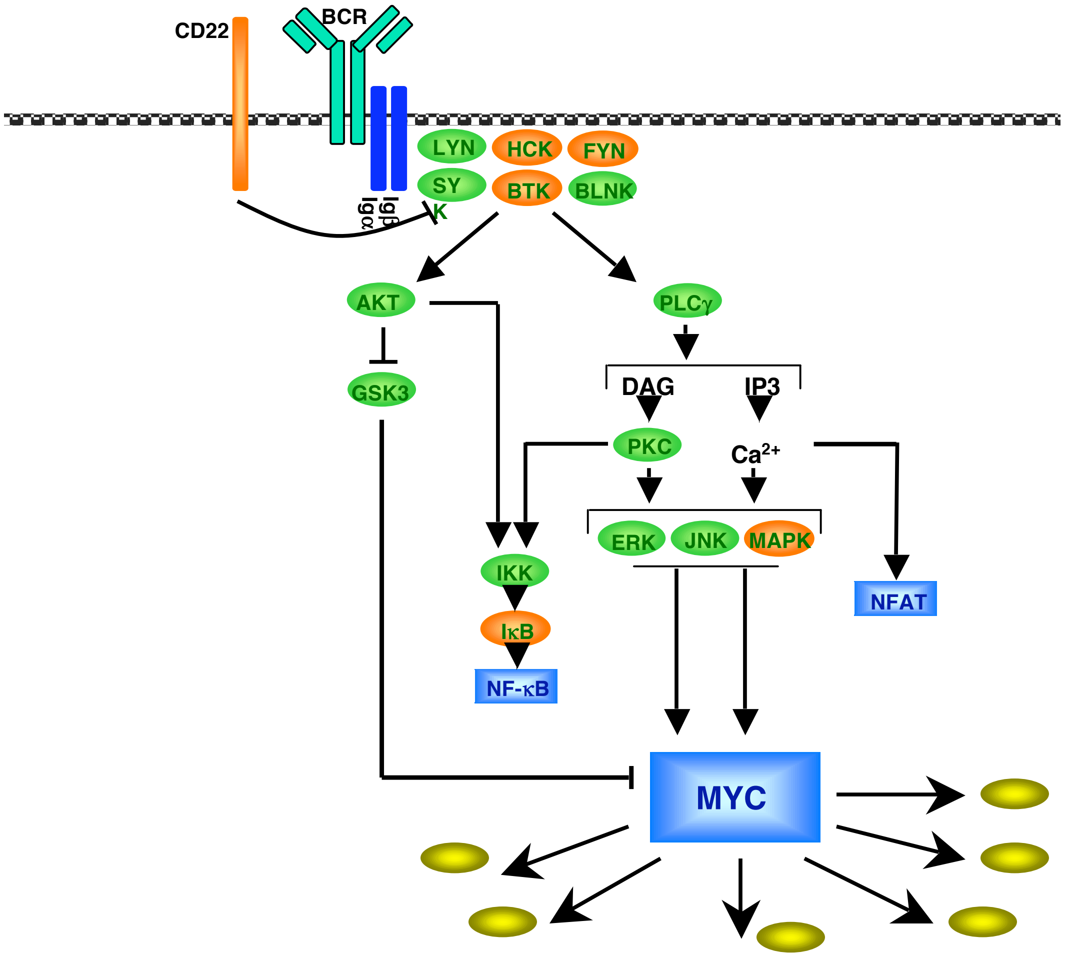
c-MYC modulators

- 1117 candidate modulators
- 100 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, p=1e-6.
- 4 out of 5 TF modulators (e.g., E2F5) with TRANSFAC signatures have binding sites in modulated targets promoter regions.
- Modulators with largest number of effected targets are not-target-specific (proteolisis, upstream signaling components, receptor signaling molecules).
- Modulators with small number of effected targets are mostly co-TFs, are interaction-specific.
- About one third of modulators are literature-validated.

Example:
TF co-factor modulator



Reducibility: modulating pathways



🟠 predicted modulators

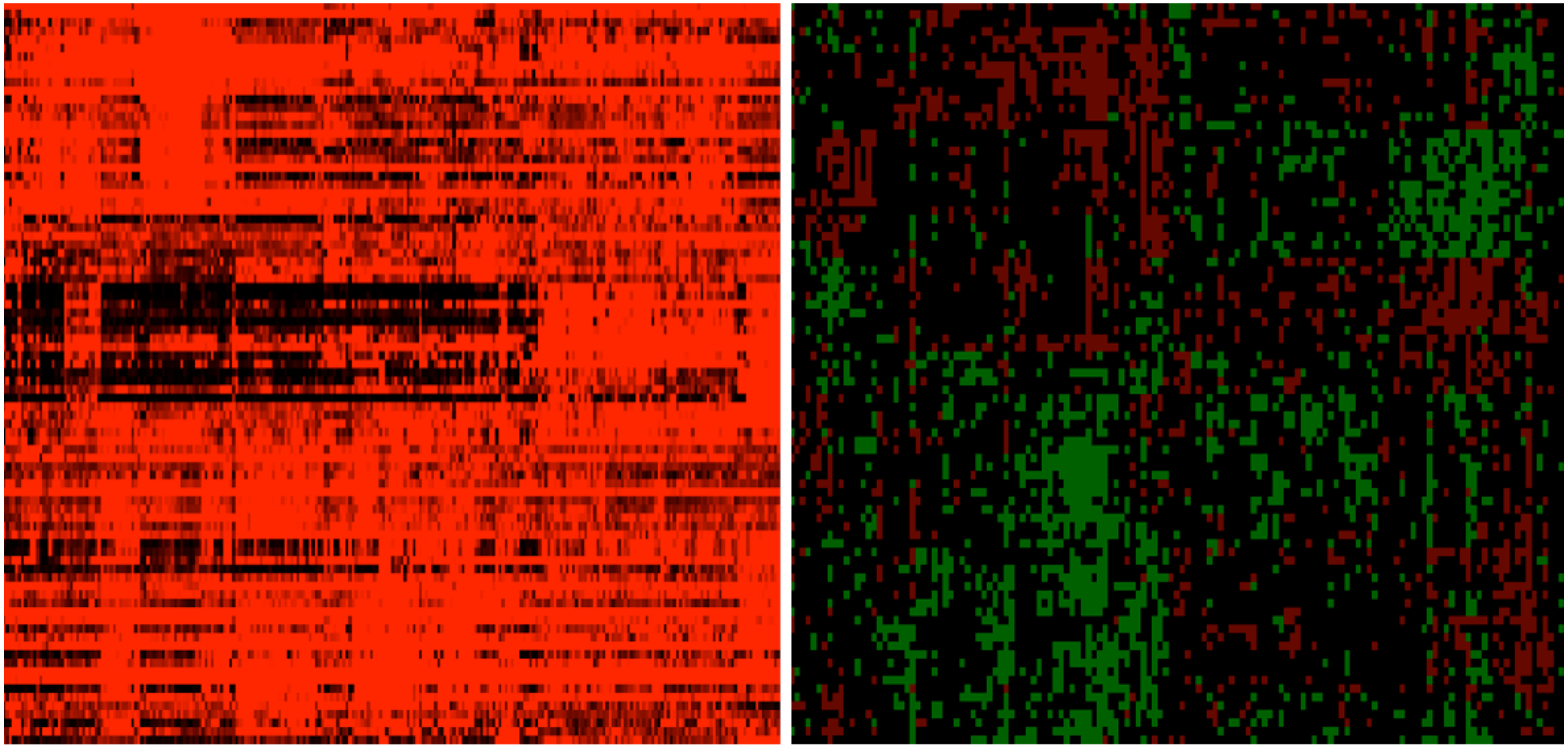
🟢 not in the candidate list

🟦 TF's not predicted

⫽ Protein complex

🟡 Targets

Many correlated modulators



|expression|

change in interactions

Over 70% cluster overlap

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