

# ARACNE: An algorithm for the reconstruction of transcriptional regulatory networks in a mammalian cellular context

Ilya Nemenman

JCSB, Columbia

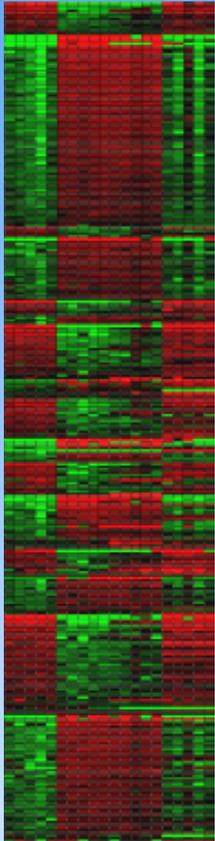
Andrea Califano, Adam Margolin, Kai Wang, Riccardo Dalla Favera,  
Katia Basso, Ulf Klein, Nila Banerjee, Chris Wiggins, Gustavo  
Stolovitzky

[q-bio.MN/0411003](#), [q-bio.MN/0410037](#)

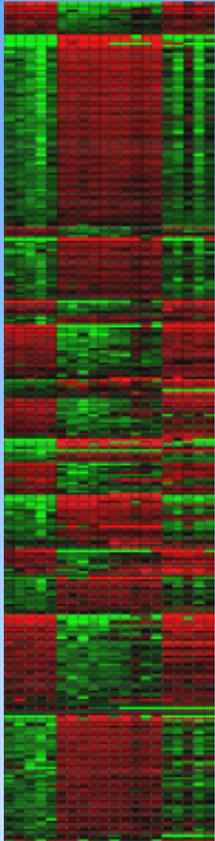
[q-bio.MN/0410036](#), [q-bio.QM/0406015](#)



# Gene expression analysis

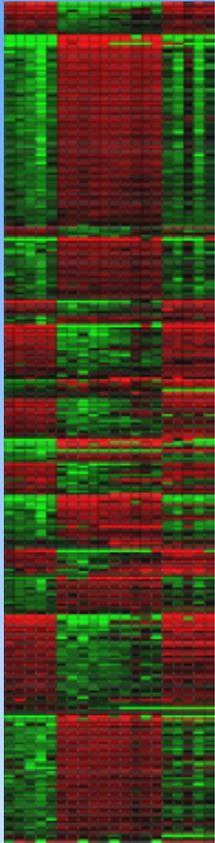


## Gene expression analysis



- clustering – too coarse
- reconstructing networks – Holy Grail!

## Gene expression analysis

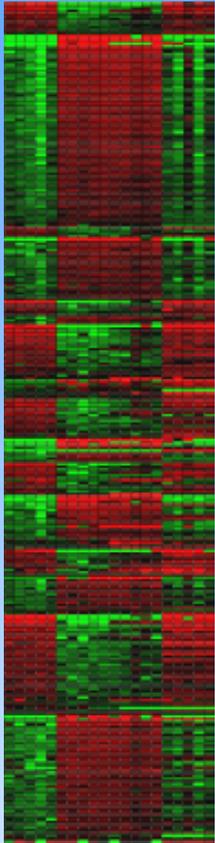


- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:

- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?

## Gene expression analysis



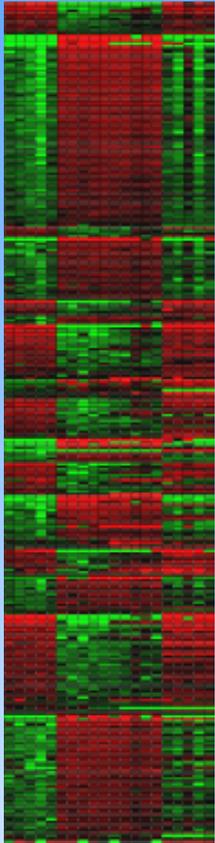
- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:

- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?

Different conditions – (small) fluctuations around a steady state (in behavior, not expression).

## Gene expression analysis



- clustering – too coarse
- reconstructing networks – Holy Grail!

Many methods exist, but:

- loops?
- what dependence (arrows) means?
- what approximations being made? controlling them?
- are approximations biologically sound?
- guarantees?

Different conditions – (small) fluctuations around a steady state (in behavior, not expression).

Akin to having covariance matrix and needing to invert it –  
**extremely many false positives** (e. g. joint coregulation).

## Model of dependence

No time series  $\rightarrow$  no directionality, **steady state statistical dependencies only.**

## Model of dependence

No time series  $\rightarrow$  no directionality, **steady state statistical dependencies only.**

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \dots$$

## Model of dependence

No time series  $\rightarrow$  no directionality, **steady state statistical dependencies only.**

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \dots$$

- use MaxEnt to define  $\phi$
- connections with spin glasses (reverse problem), generalizes MNs

## Model of dependence

No time series  $\rightarrow$  no directionality, **steady state statistical dependencies only.**

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \sum_{ijk} \phi_{ijk}(g_i, g_j, g_k) + \dots$$

- use MaxEnt to define  $\phi$
- connections with spin glasses (reverse problem), generalizes MNs
- enough data to evaluate 2-way marginals only;

## Model of dependence

No time series  $\rightarrow$  no directionality, **steady state statistical dependencies only.**

$$-\log P(g_i) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_i, g_j) + \dots$$

- use MaxEnt to define  $\phi$
- connections with spin glasses (reverse problem), generalizes MNs
- enough data to evaluate 2-way marginals only;
- truncate at 2nd order potential (cannot reconstruct XOR), Bethe approximation
- Mutual information  $I(g_i, g_j) = \langle \log P(g_i, g_j) / P(g_i)P(g_j) \rangle$  is enough to establish dependencies.

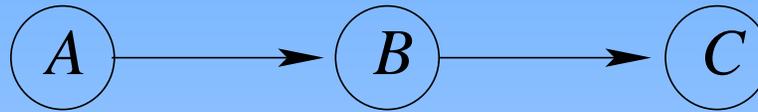
## Notes

- introducing extra  $\psi(g_i)$  describes response to perturbations (but: directionality)

## Notes

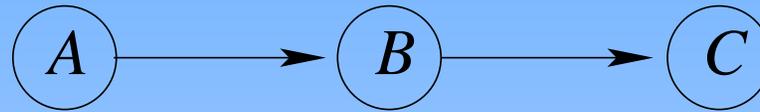
- introducing extra  $\psi(g_i)$  describes response to perturbations (but: directionality)
- biochemical dependencies persist as steady state statistical dependencies, but orders of interactions may change

# Removing false positives – Data Processing inequality



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

# Removing false positives – Data Processing inequality



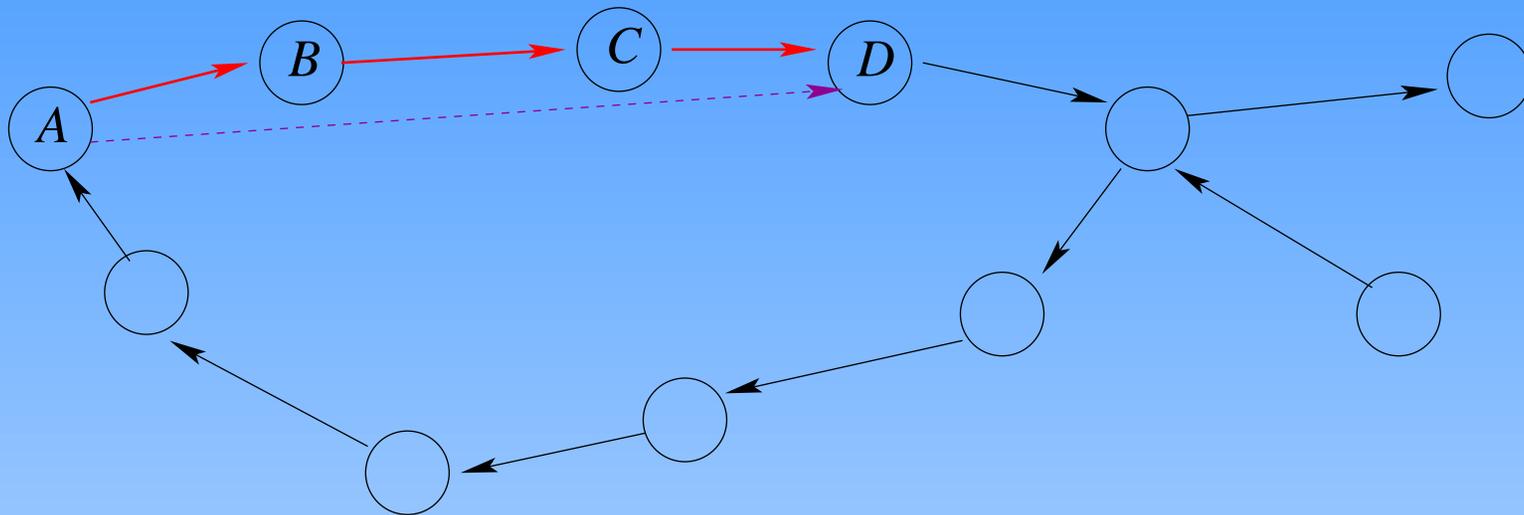
$$I(A, C) \leq \min[I(A, B), I(B, C)]$$

ARACNE: Look at every triplet and remove the weakest link.

# Guarantees

Theorem. 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. The maximum Mutual Information spanning tree is a subnetwork of the network reconstructed by ARACNE.



Theorem. Let  $\pi_{ik}$  be the set of nodes forming the shortest path in the network between nodes  $i$  and  $k$ . Then, if MIs can be estimated without errors, ARACNE reconstructs an interaction network without false positives edges, provided: (a) the network consists only of pairwise interactions, (b) for each  $j \in \pi_{ik}$ ,  $I_{ij} \geq I_{ik}$ .

Further, ARACNE does not produce any false negatives, and the network reconstruction is exact iff (c) for each directly connected pair  $ij$  and for any other node  $k$ , we have  $I_{ij} > \min[I_{ik}, I_{jk}]$ .

## Why should it work?

- higher order interactions project into lower order ones
- large loops are locally trees (biological signals decorrelate very fast:  
 $I(\text{cMYK}, \text{cMYK}) \approx 8$  bits,  $I(\text{cMYK}, \text{second best}) \approx 1$  bit.
- small loops (e. g., feed forward) are often transient

## Gaussian kernels for MI estimation

$$f(x, y) = \frac{1}{2\pi h^2 M} \sum_i \exp \left[ -\frac{(x - x_i)^2 + (y - y_i)^2}{2h^2} \right]$$

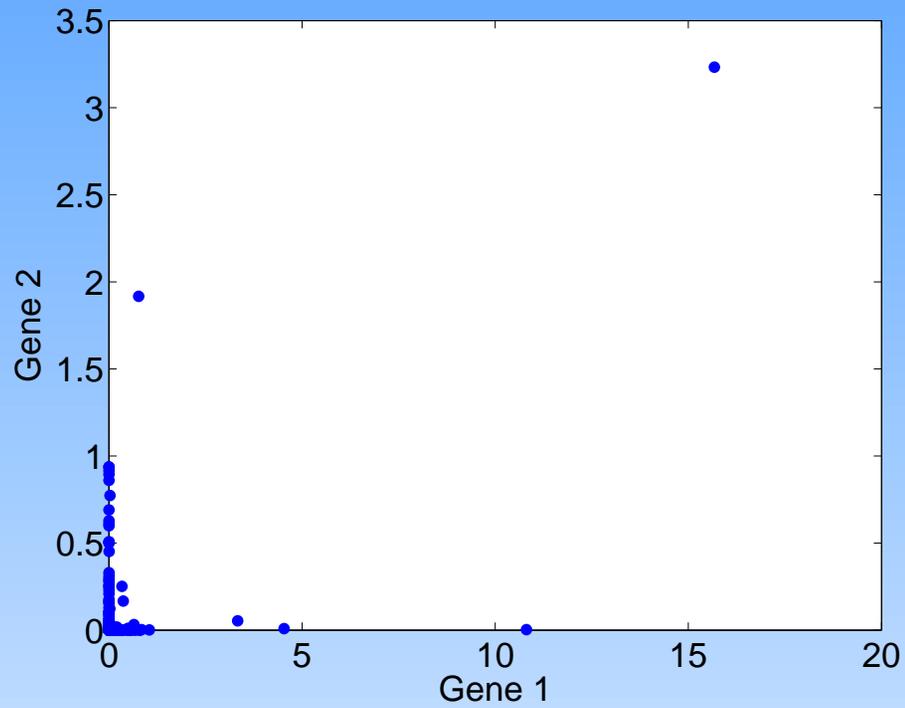
$$f(x) = \frac{1}{\sqrt{2\pi} h M} \sum_i \exp \left[ -\frac{(x - x_i)^2}{2h^2} \right]$$

Consistent for  $M \rightarrow \infty$  for  $h(M) \rightarrow 0$  and  $[h(M)]^2 M \rightarrow \infty$ .

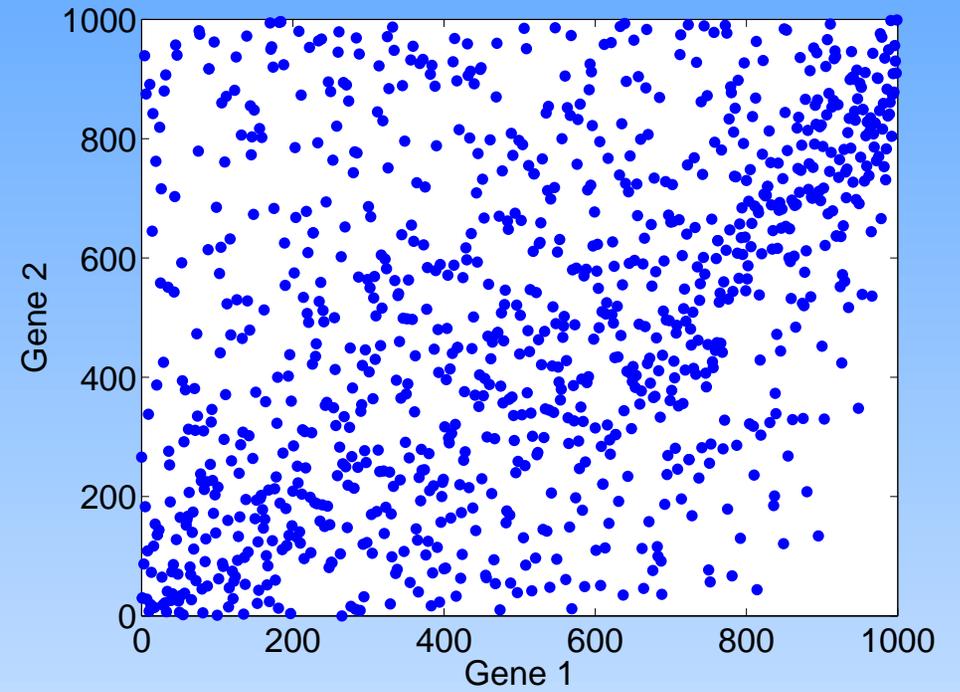
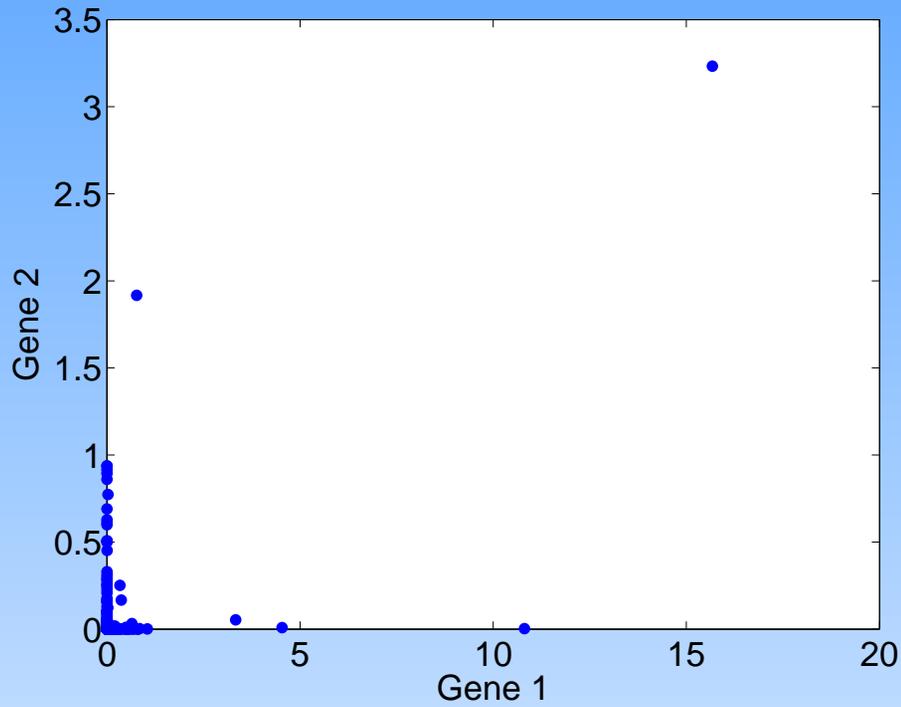
How to select  $h$ ?

Maybe  $h = h(x, y)$ ?

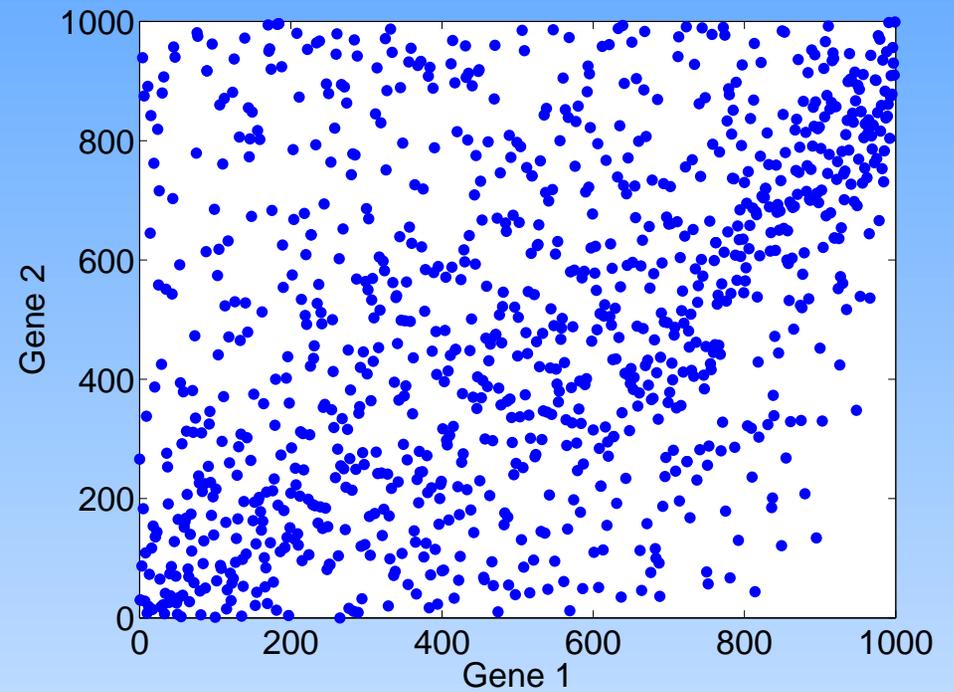
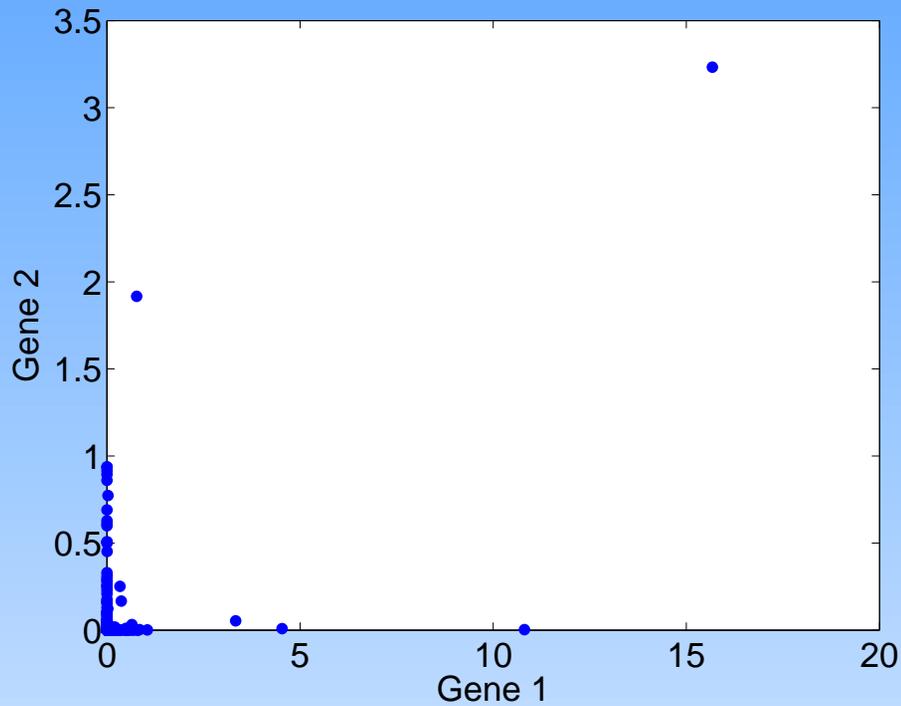
# Copula transform



# Copula transform

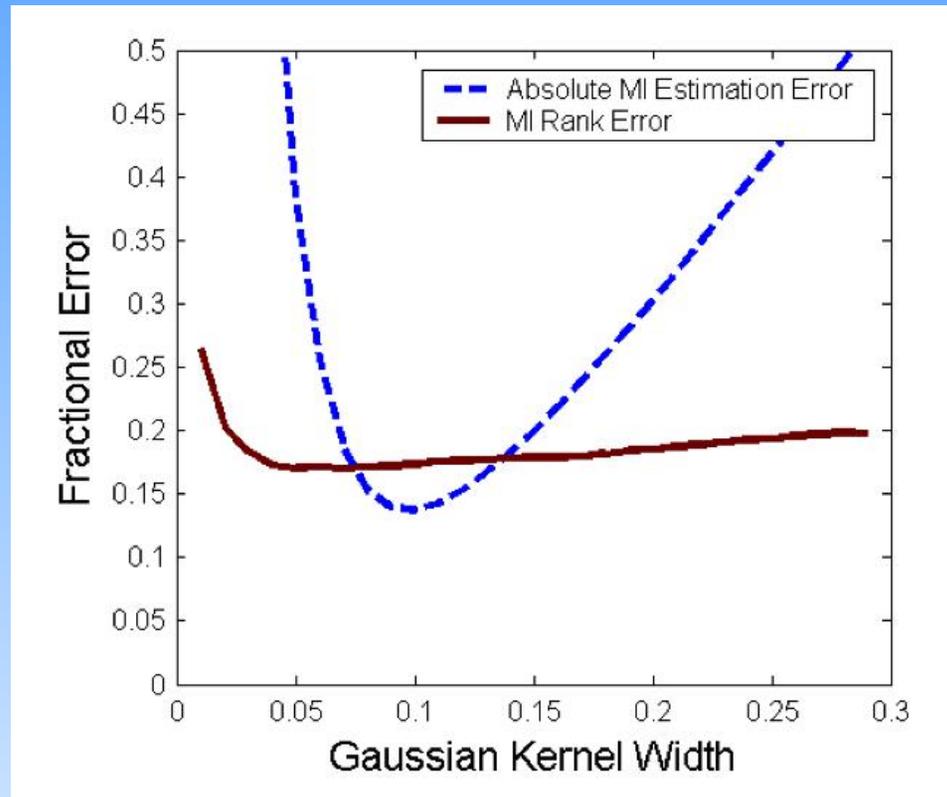


# Copula transform

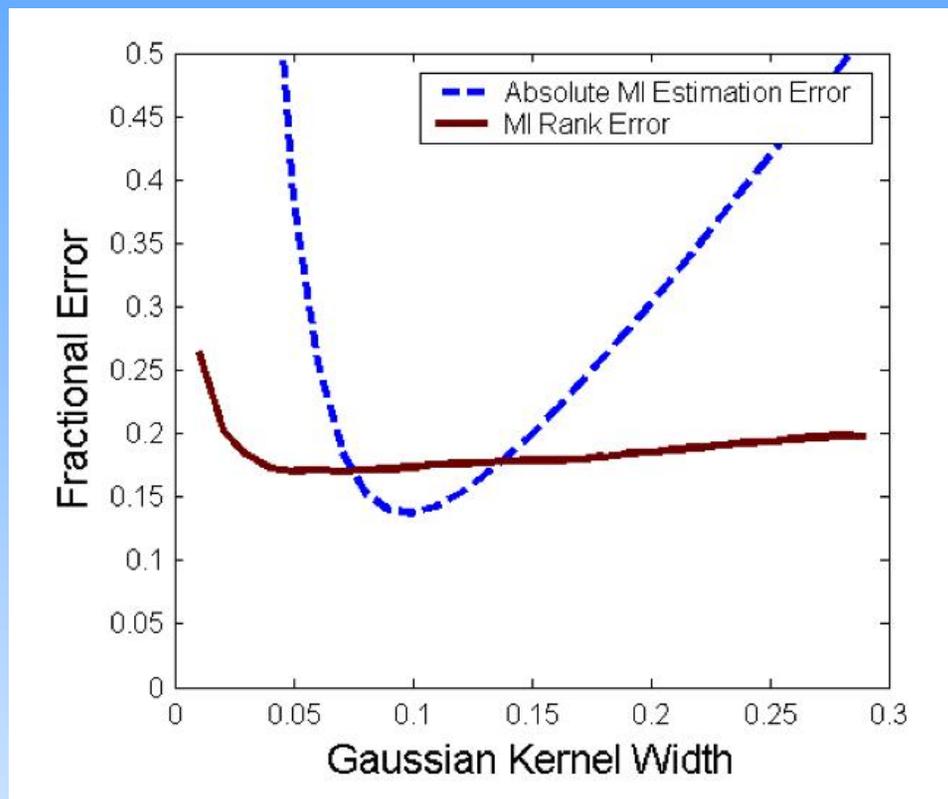


No need to have  $h = h(x, y)$ .

# Mutual information error vs. ranking error

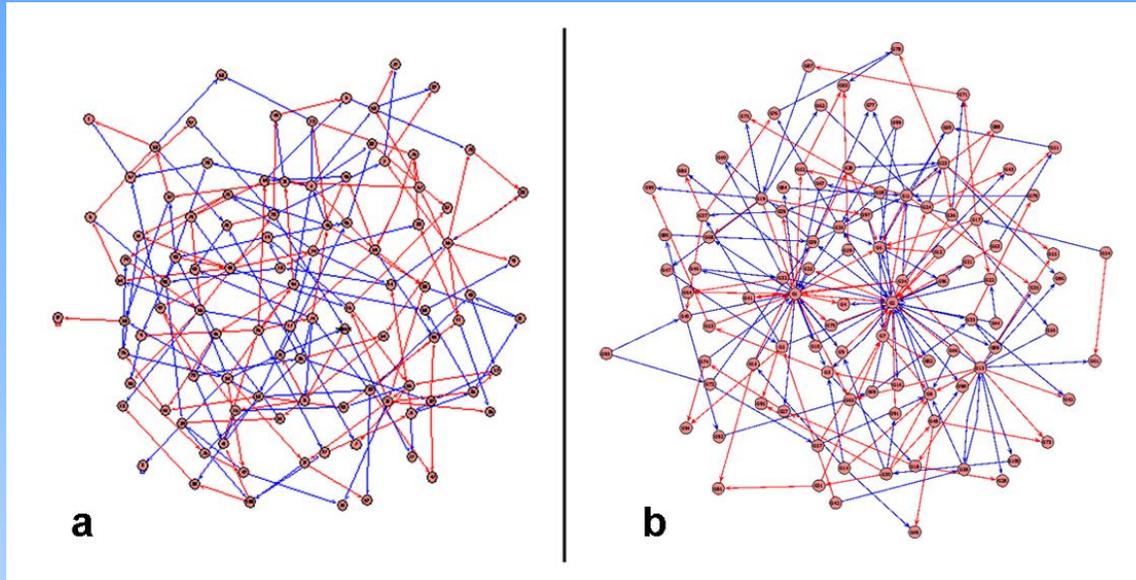


# Mutual information error vs. ranking error

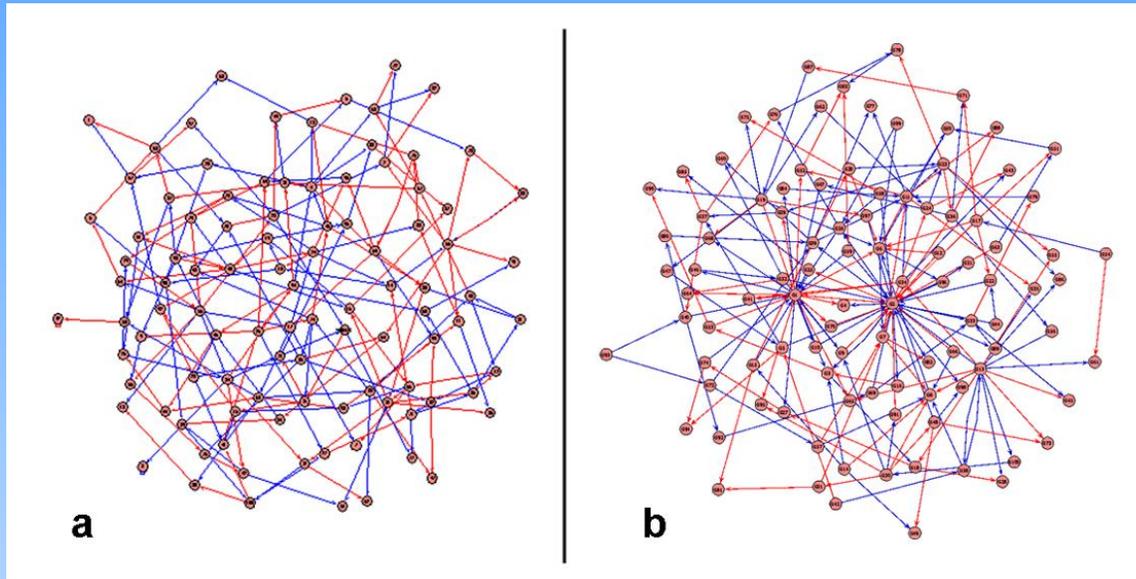


Can use universal best  $h$ .

# Synthetic networks

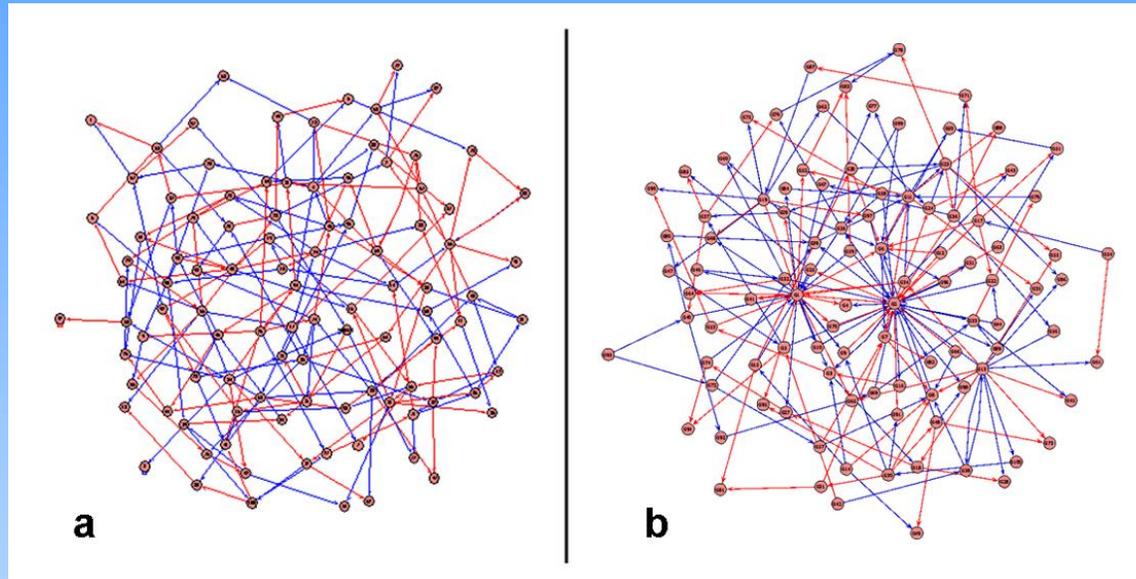


# Synthetic networks



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

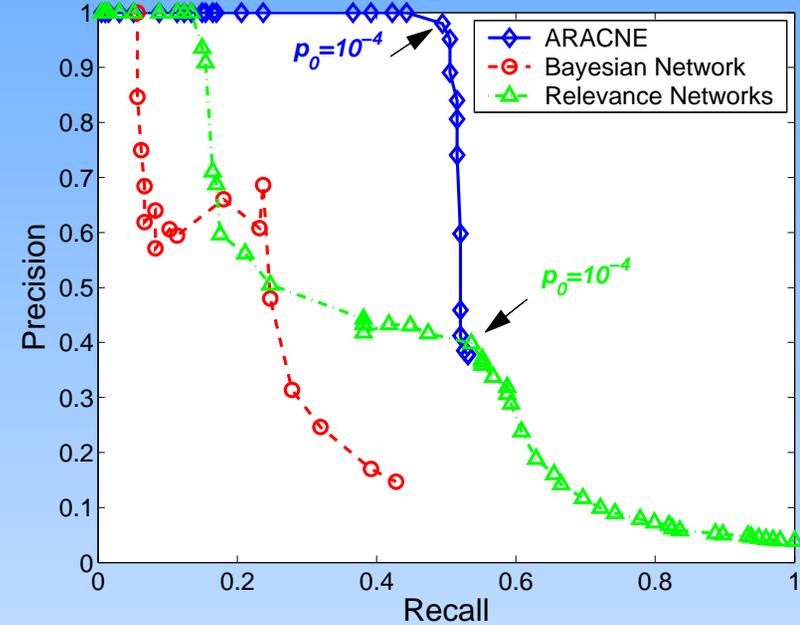
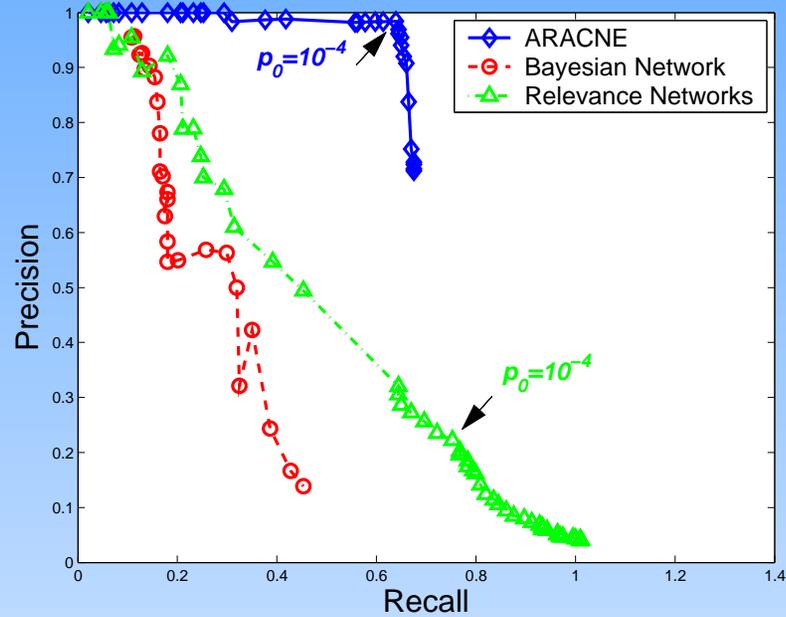
# Synthetic networks



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

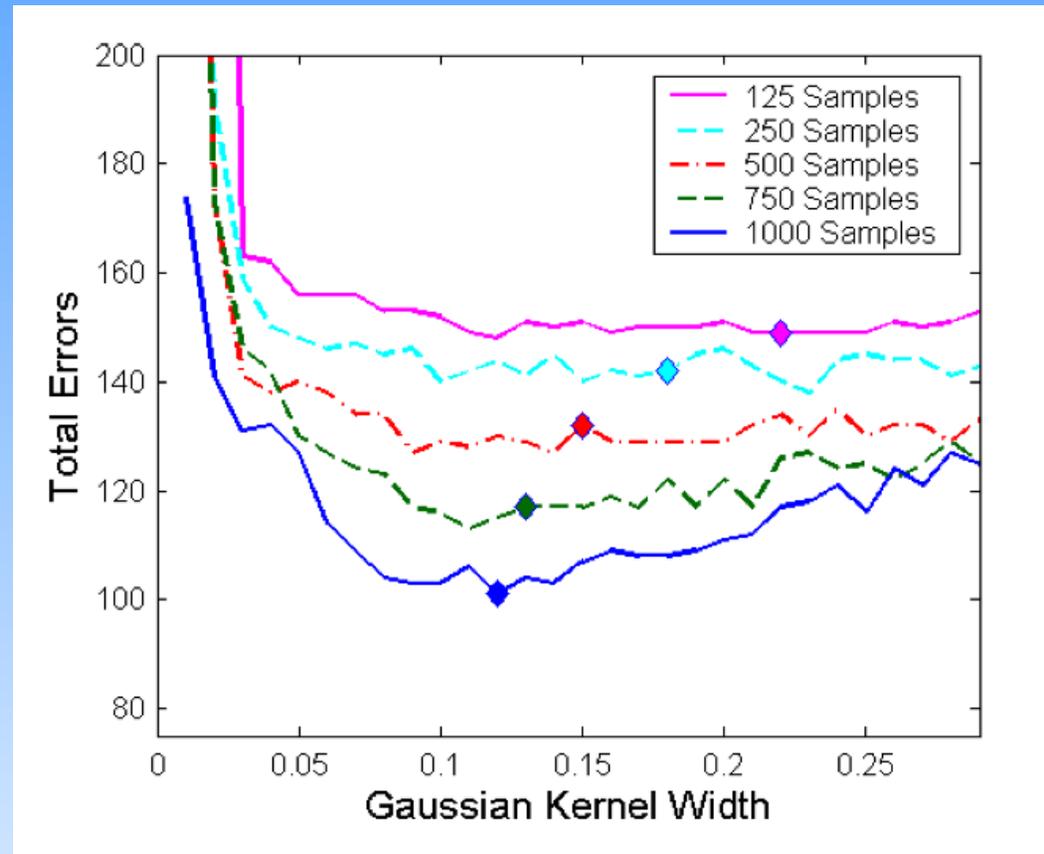
To simulate phenotypes and conditions, randomize  $a_i$  and  $b_i$  (and, possibly,  $I_{j,0}$ ,  $A_{j,0}$ ).

# Benchmarks

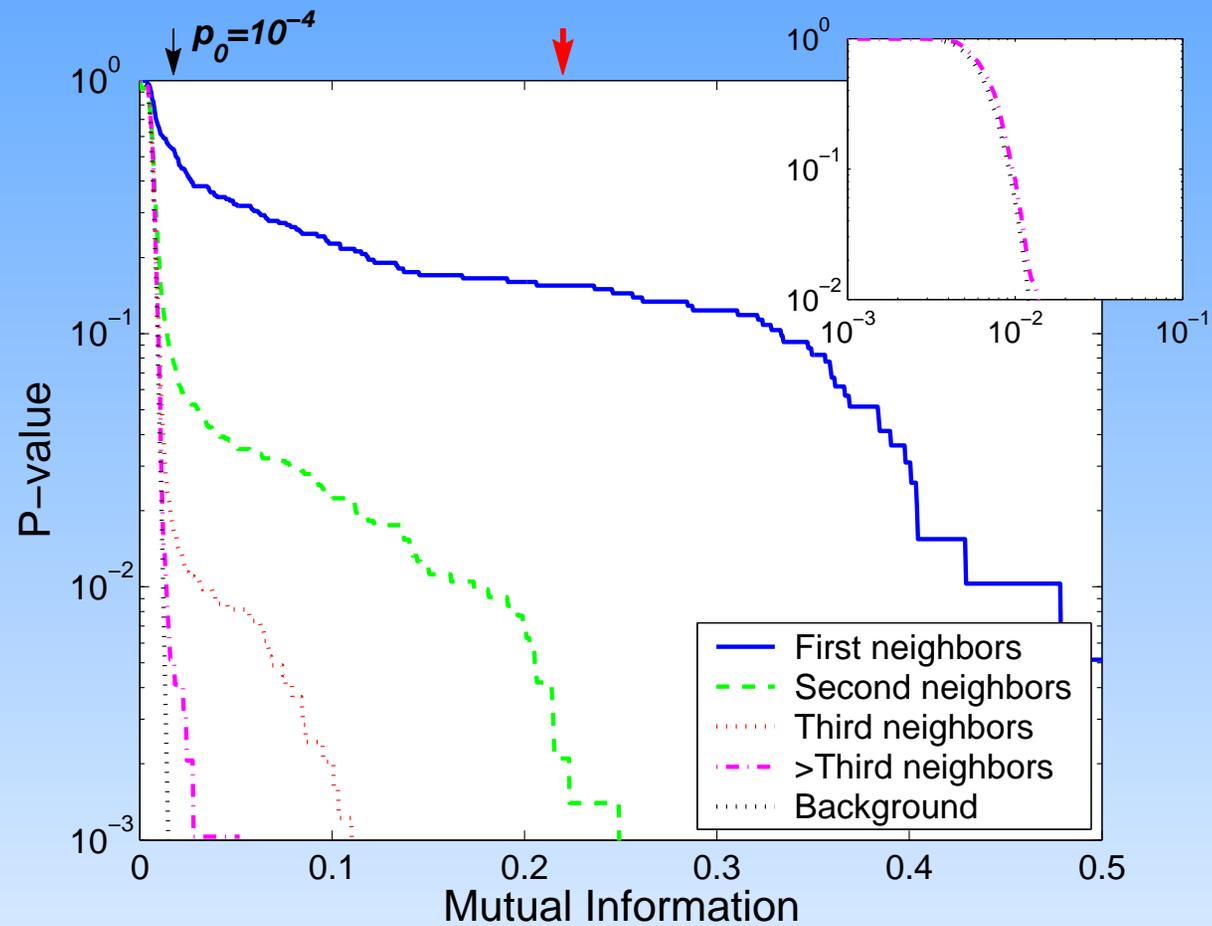


$$N_{TP} - N_{FP} = \max \text{ at } p = 10^{-4}.$$

# No sampling catastrophe!



# Why RN's fail

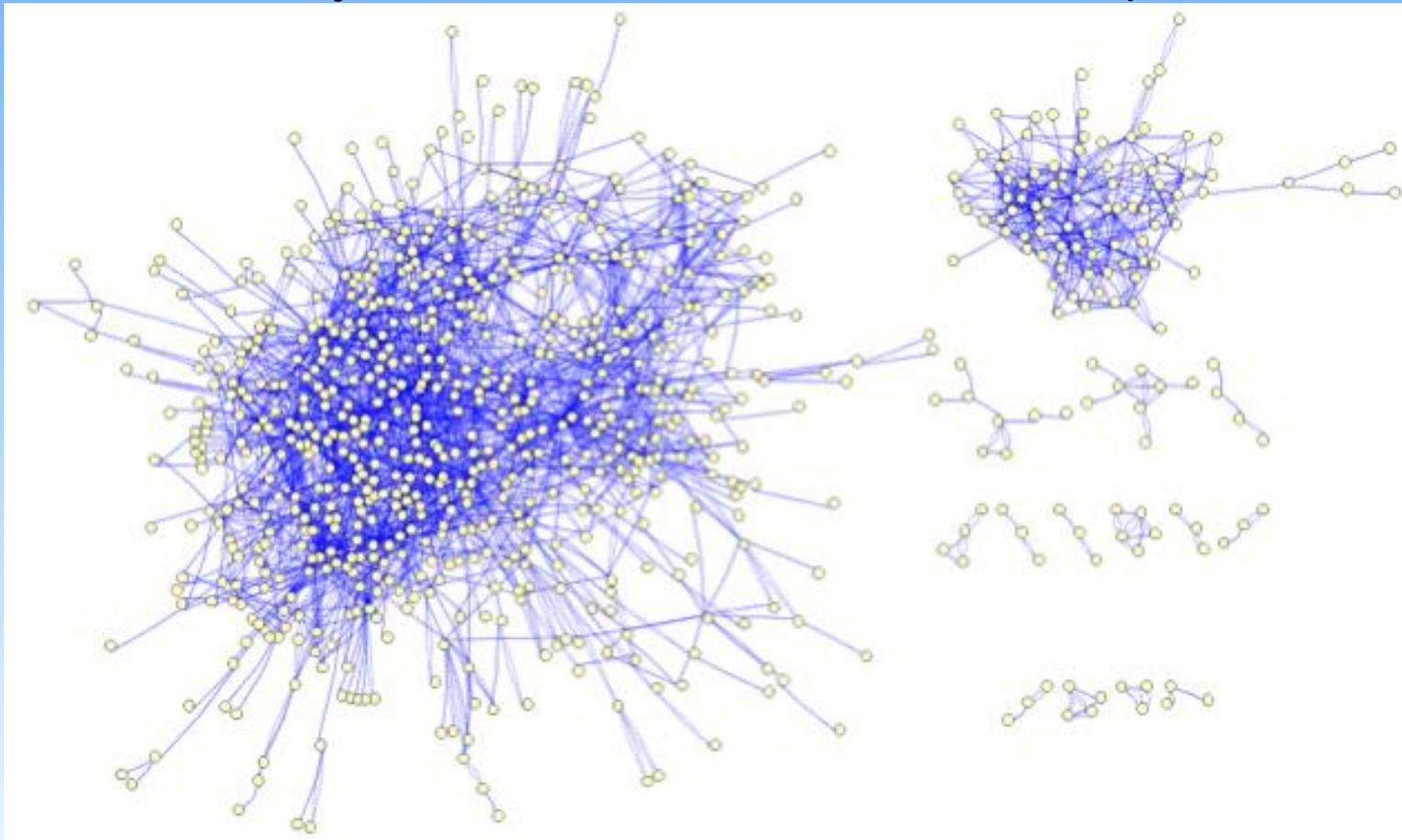


# Complete B-cell network

~ 129000 interactions (possibly scale-free).

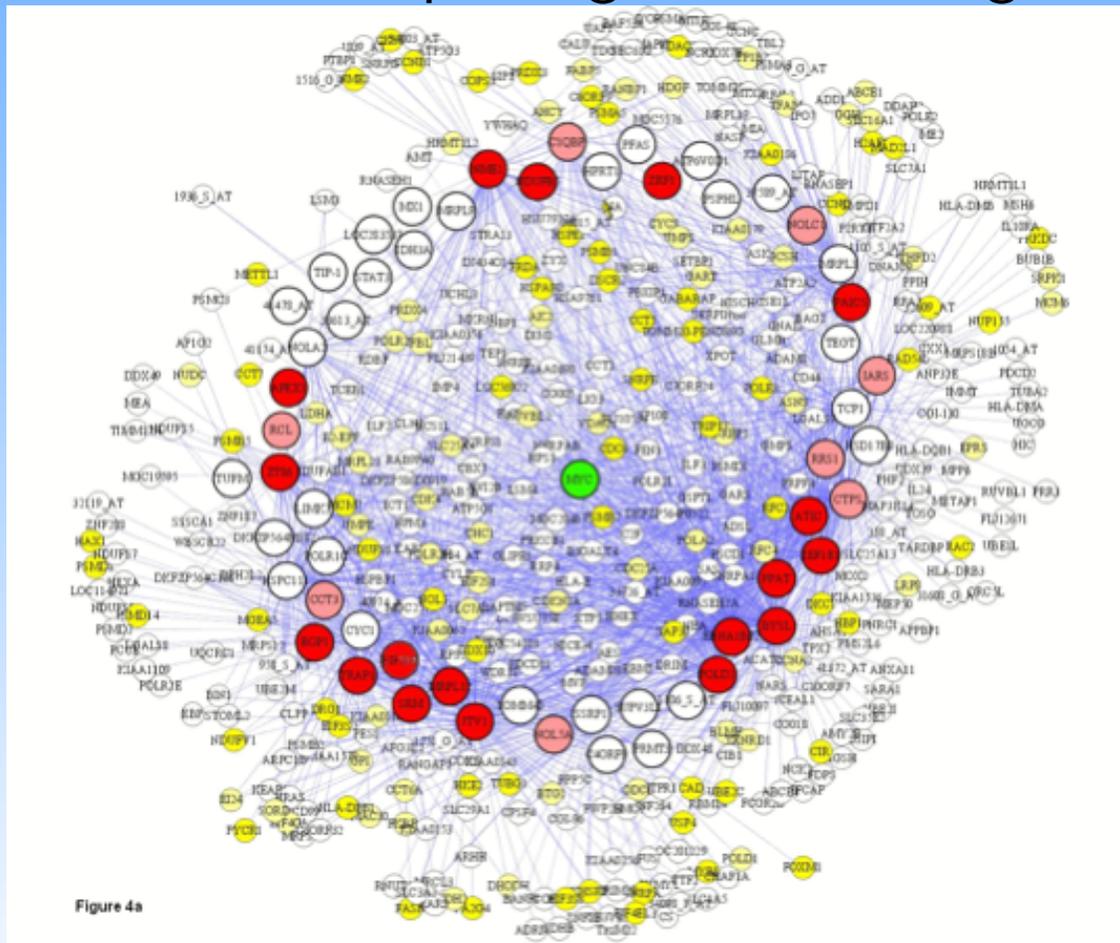
Cell cycle

Ribosomal complex



# c-MYC TF centered network

Protooncogene, involved in many cellular processes, 12% background interactions, top 5% genetic hub, significant MI with  $\sim 2000$  genes.



# c-MYC TF centered network

Protooncogene, involved in many cellular processes, 12% background interactions, top 5% genetic hub, significant MI with  $\sim 2000$  genes.

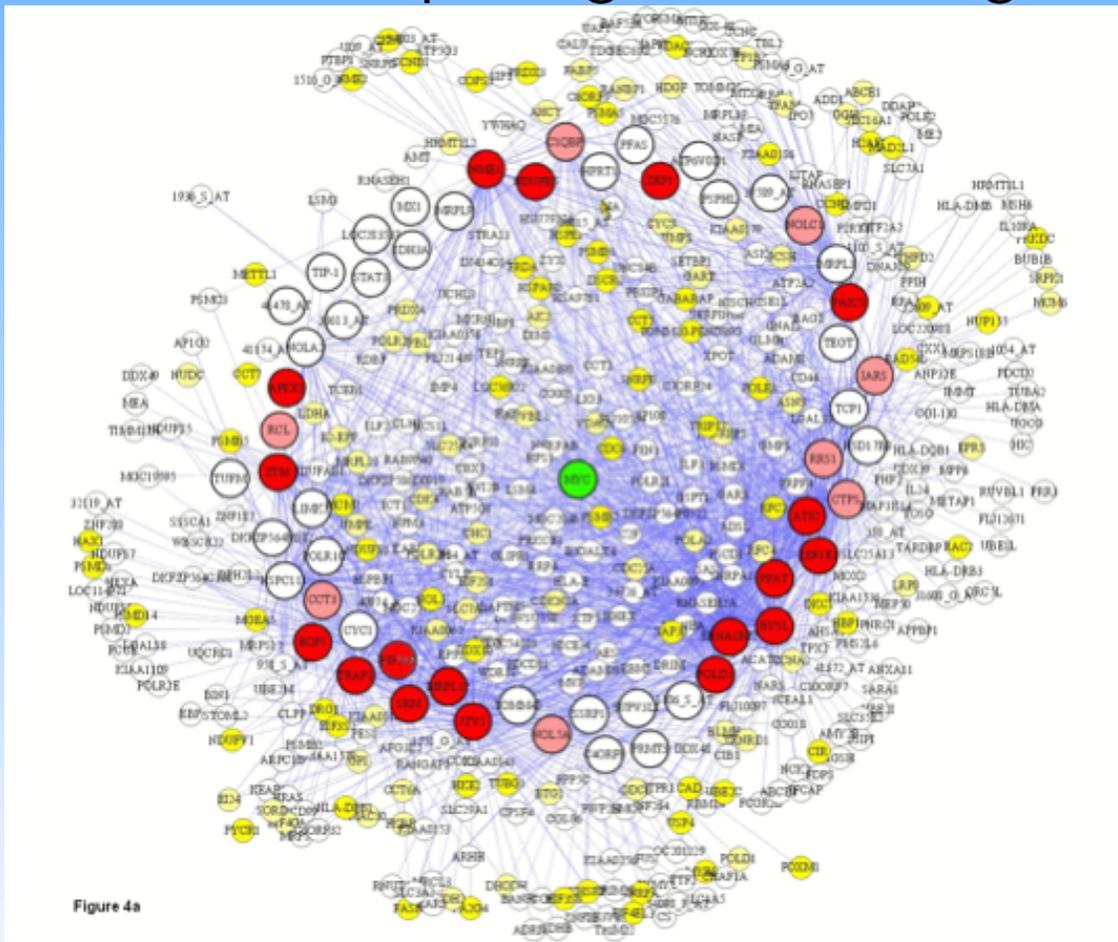
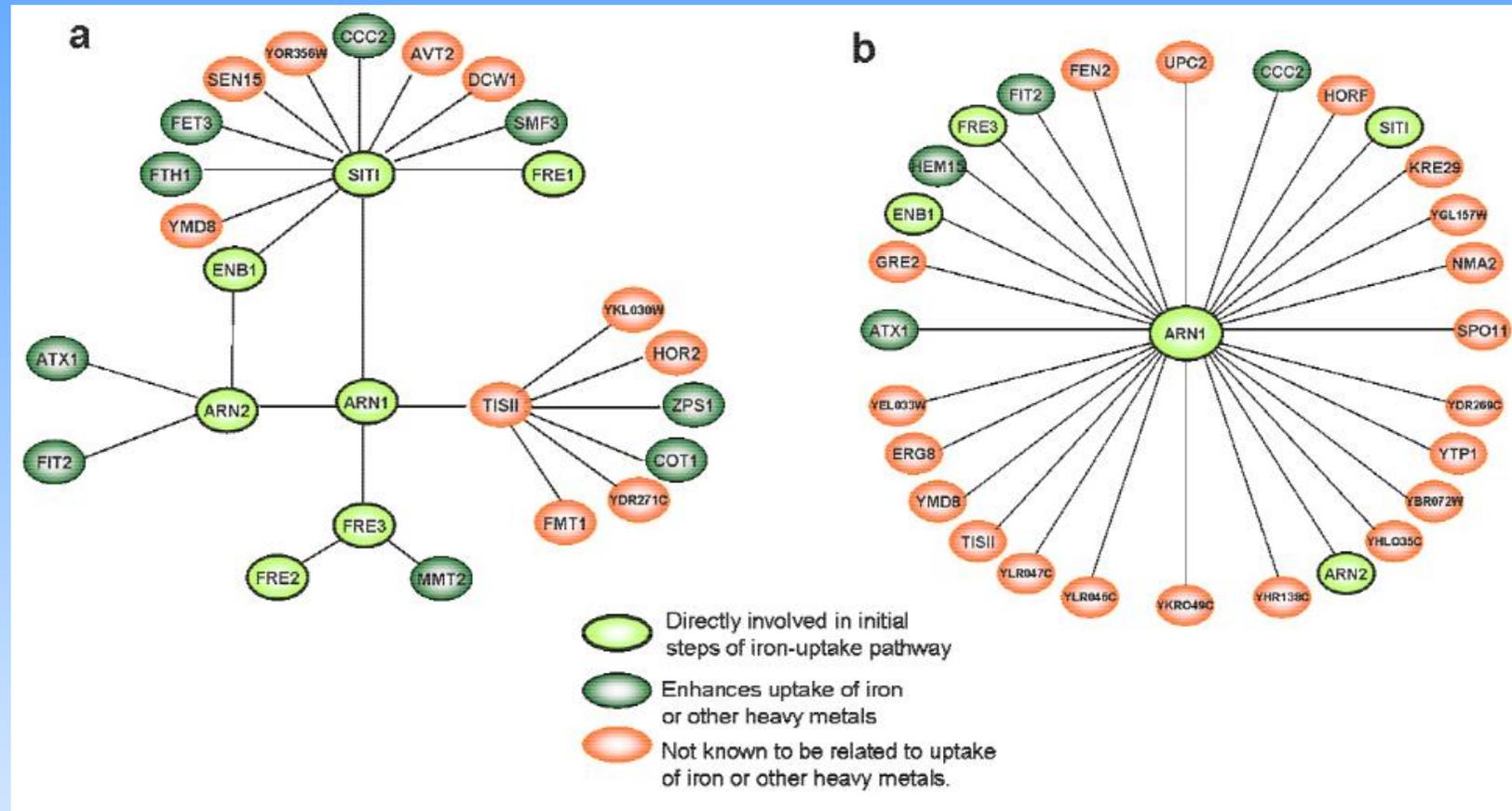


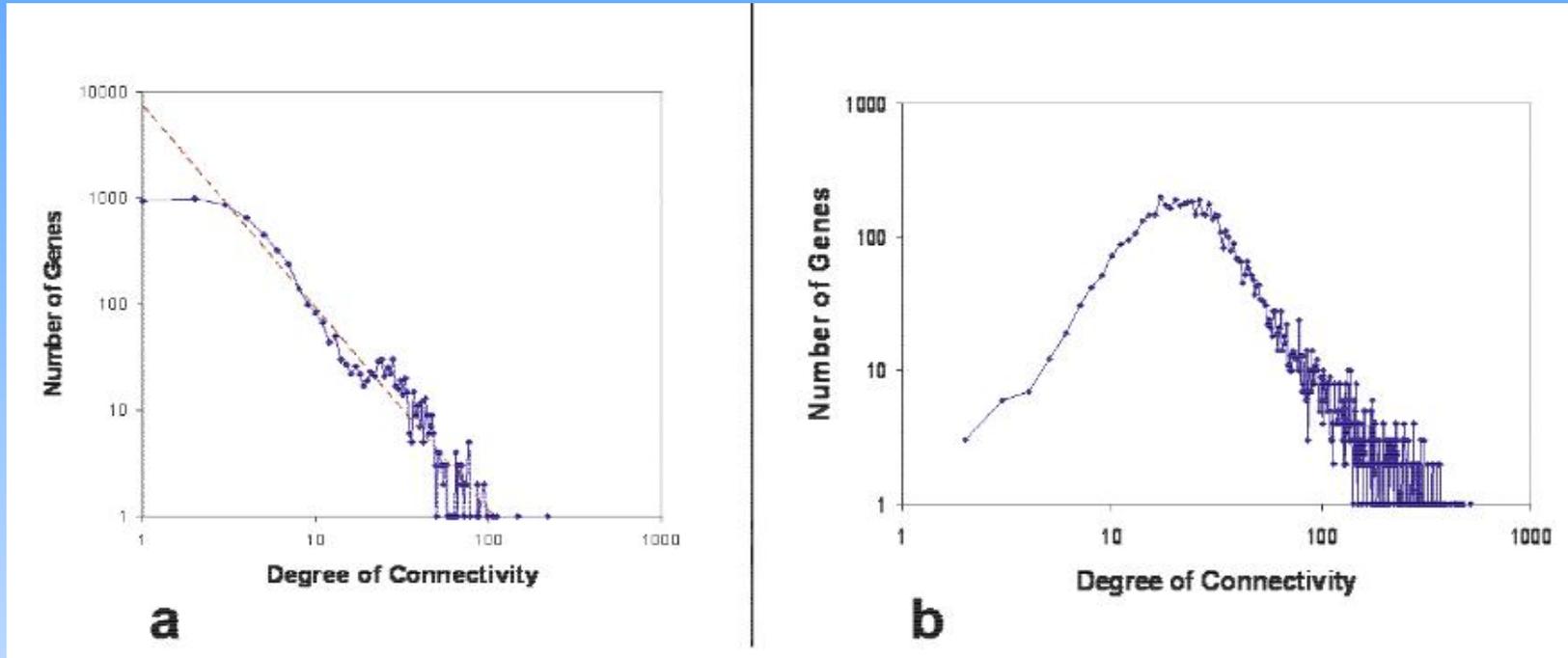
Figure 4a

- 56 1st neighbors
- pre-known targets – 22
- ChIP-proven targets – 11/12
- 2nd neighbors – weaker enrichments
- Most 1st – major hubs

# ARN1 iron uptake in Yeast



# Global network properties



## Hub 3-way interactions (conditional analysis)

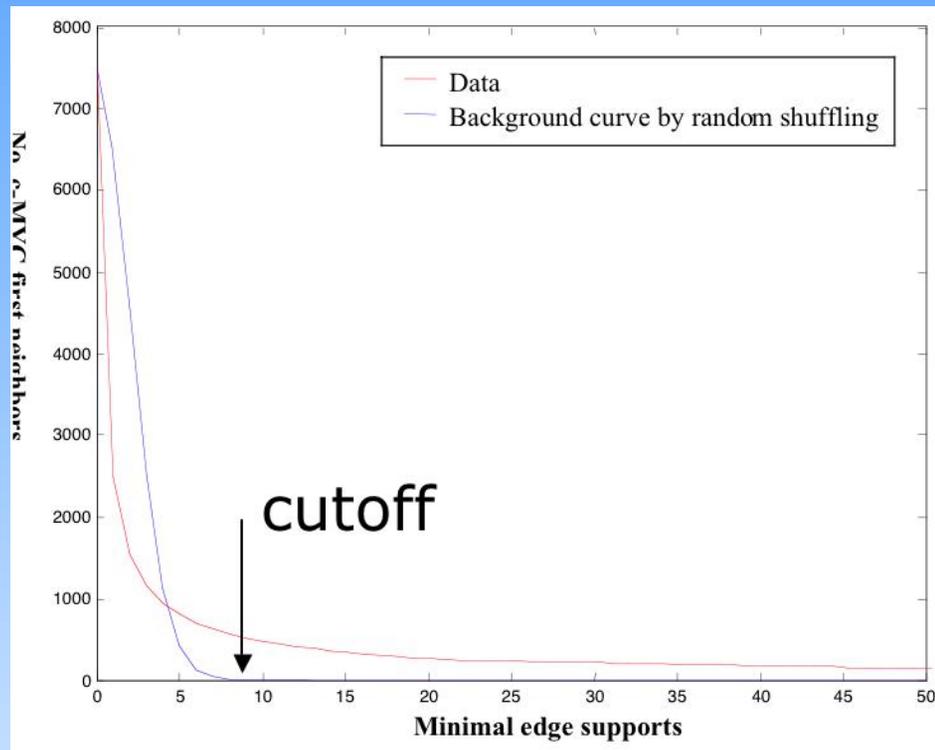
$G_{\mu}^*$  – coarse conditions (+/-) of correlated gene clusters

$$I(g_i, g_j | G_{\mu}^*)$$

- Independent of the hub (true 3-way interactions)
- Large dynamic range

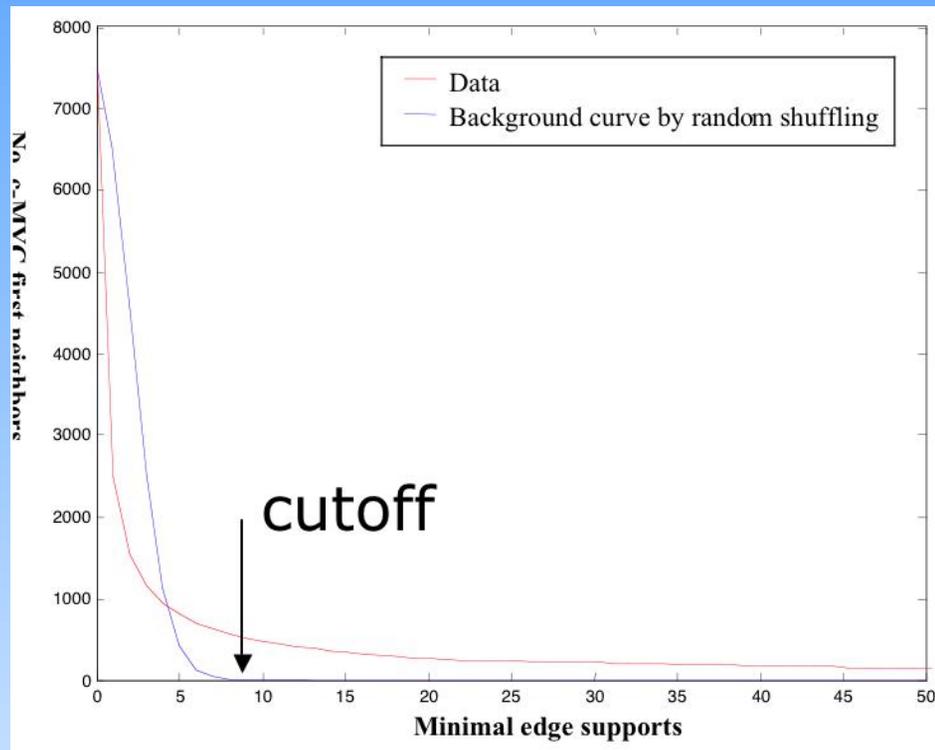
	$G_1^+$	$G_1^-$	$G_2^+$	$G_2^-$	...	...	$G_M^+$	$G_M^-$
<b>Edge 1</b>	1	0	1	0	0	0	1	0
<b>Edge 2</b>	0	1	0	0	0	1	0	0
<b>:</b>	0	0	1	0	1	0	0	0
<b>Edge N</b>	1	0	0	1	0	0	0	1

# Edge support conditions set size



#	$N_P$	$N_V$	$E.$	$N_{FP}$	$P$
1	2422	437	0.18	6520.1	1
2	1458	278	0.19	4541.5	1
3	1066	224	0.21	2514.1	1
4	847	182	0.21	1131.6	1
5	710	157	0.22	423.13	0.60
6	591	132	0.22	136.04	0.23
7	511	119	0.23	37.03	0.072
8	459	110	0.24	9.18	0.02
9	406	104	0.26	1.9	< 0.01
10	367	96	0.26	0.37	< 0.01

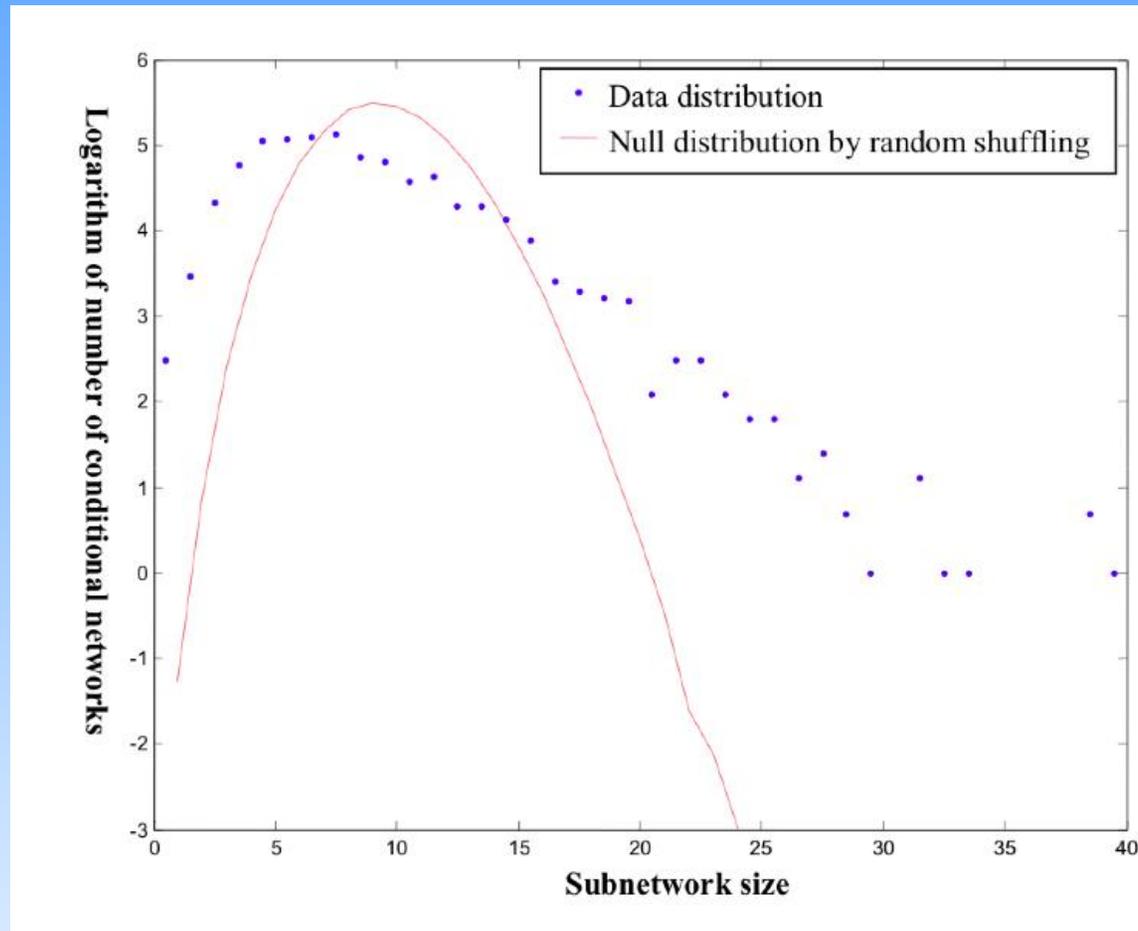
# Edge support conditions set size



#	$N_P$	$N_V$	$E.$	$N_{FP}$	$P$
1	2422	437	0.18	6520.1	1
2	1458	278	0.19	4541.5	1
3	1066	224	0.21	2514.1	1
4	847	182	0.21	1131.6	1
5	710	157	0.22	423.13	0.60
6	591	132	0.22	136.04	0.23
7	511	119	0.23	37.03	0.072
8	459	110	0.24	9.18	0.02
9	406	104	0.26	1.9	< 0.01
10	367	96	0.26	0.37	< 0.01

Probably better than original algorithm.

# Conditional network sizes



# Regulators, indeed

Of 168 c-MYC regulators:

GO Category	$N_C$	$N_{Tot}$	GO $N_C$	GO $N_{Tot}$	$P$
Transcription Regulator Activity (MF)	21	116	2089	21014	0.0049
Protein Kinase (MF)	13	116	1004	21014	0.003
IKBK/NFKB cascade (BP)	5	117	222	24373	0.004
Immune Response (BP)	19	117	1664	24373	0.0003
Humoral IR (BP)	9	117	378	24373	0.0001
Reg. of Transcription, DNA-dep. (BP)	26	117	1697	24373	$1.2 \times 10^{-7}$