COMPRESSIBLE COMPONENTS REVEAL NETWORK ARCHITECTURES

CABDyN Seminar, University of Oxford
25 February 2014

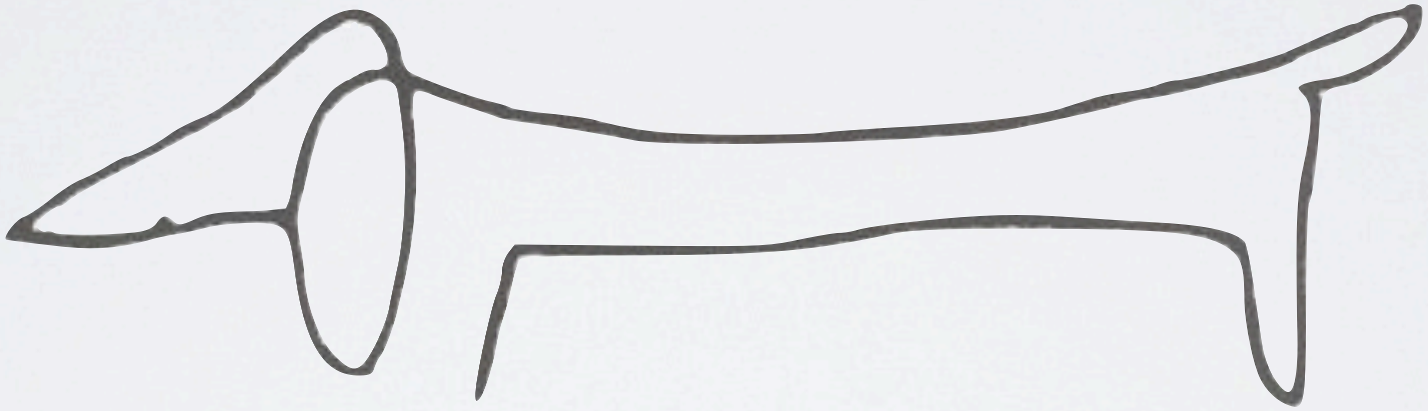
Sebastian Ahnert

Cavendish Laboratory
UNIVERSITY OF CAMBRIDGE

WHAT IS A SKETCH?

WHAT IS A SKETCH?

A sketch concisely captures dominant features.



A NETWORK SKETCH?

A NETWORK SKETCH?

How can we draw a sketch of a network?

A NETWORK SKETCH?

How can we draw a sketch of a network?

This would mean capturing dominant structural features in a concise way.

A NETWORK SKETCH?

How can we draw a sketch of a network?

This would mean capturing dominant structural features in a concise way.

What are dominant structural features?

A NETWORK SKETCH?

How can we draw a sketch of a network?

This would mean capturing dominant structural features in a concise way.

What are dominant structural features?

A NETWORK SKETCH?

How can we draw a sketch of a network?

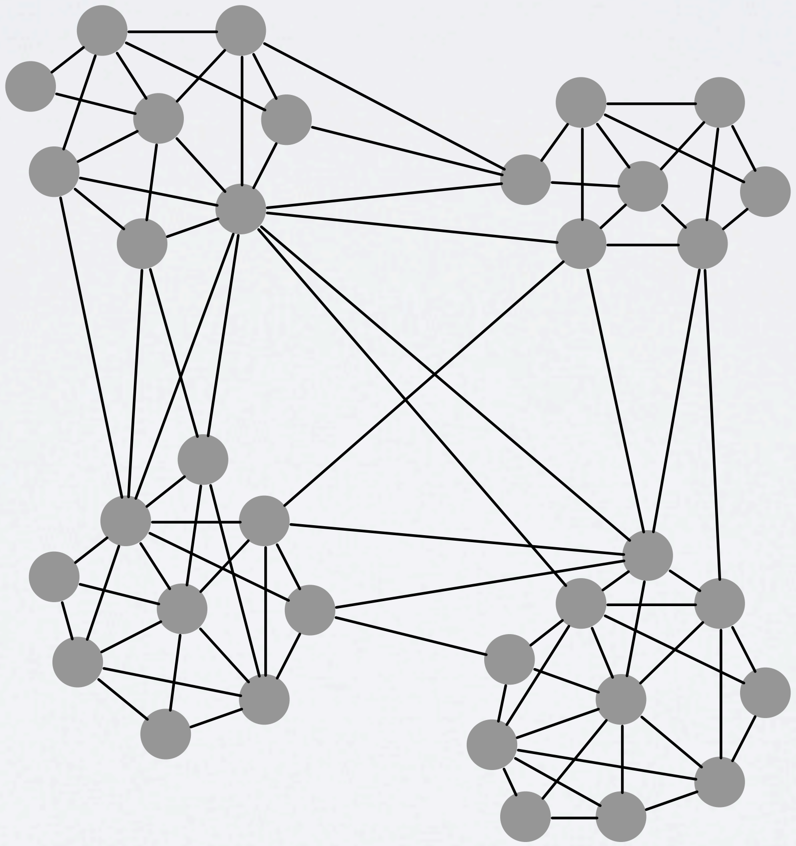
This would mean capturing dominant structural features in a concise way.

What are dominant structural features?

Communities?

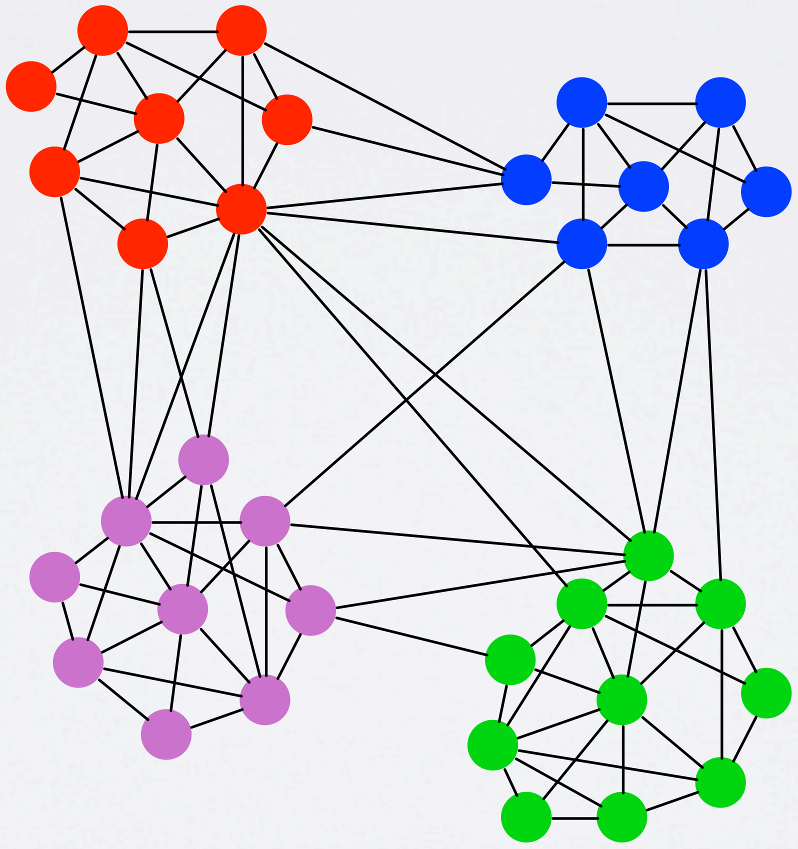
COMMUNITY DETECTION

Communities in networks are typically defined as densely interconnected subsets of nodes.



COMMUNITY DETECTION

Communities in networks are typically defined as densely interconnected subsets of nodes.



COMMUNITY DETECTION

One of the best known methods of this kind is Newman & Girvan's [PRE 69, 026113 (2004)]:

1) calculate betweenness for all edges
2) remove highest betweenness edge
3) recalculate betweenness for all edges and go to 2)

This fragments the network and thus gives rise to a series of partitions.

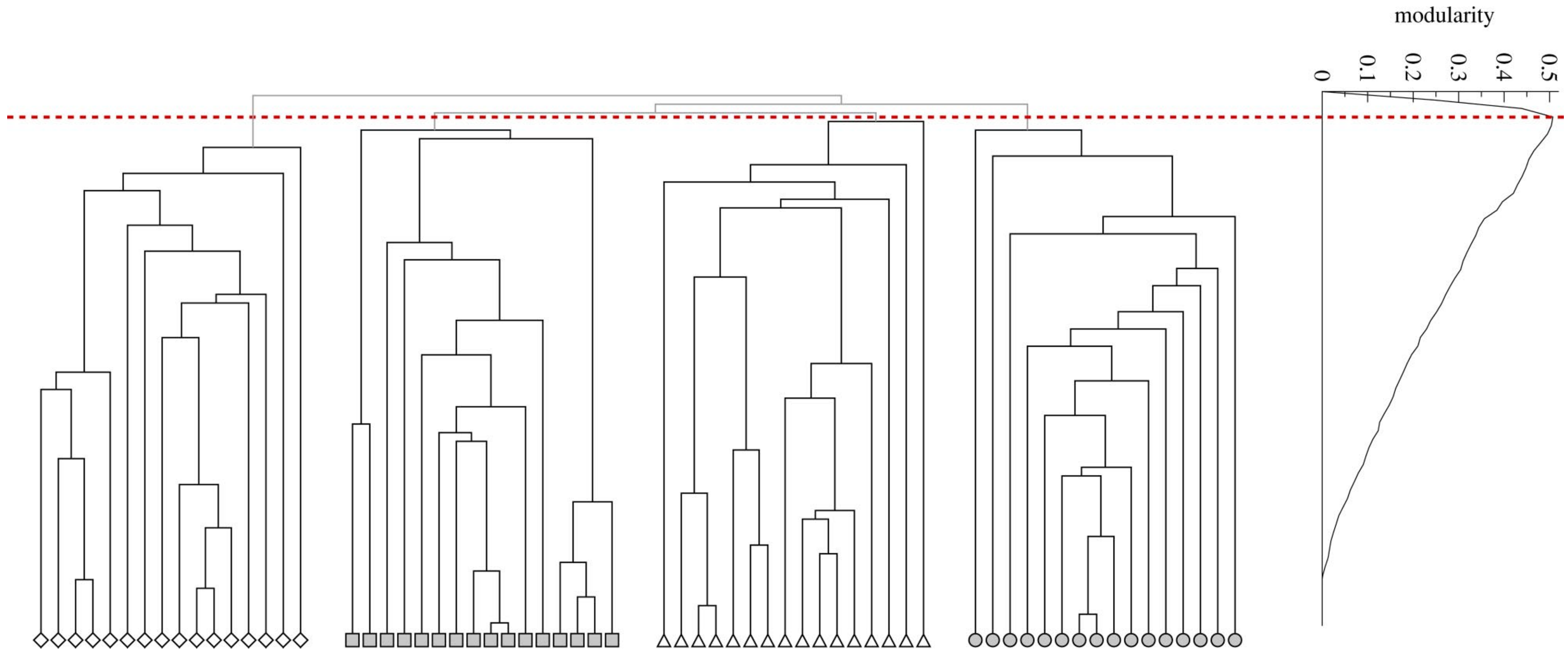
COMMUNITY DETECTION

One then chooses the optimal partition by maximising *modularity*, which is defined as:

$$Q = \sum_i (e_{ii} - a_i^2)$$

where e_{ii} is the fraction of network edges *within* partition i, and a_i is the fraction of edges *connecting to* partition i.

COMMUNITY DETECTION



THE PROBLEMS

Most existing community detection methods:

- are **node-based**
- attempt to **partition** the nodes
- look at connectivity **within** a community.

THE PROBLEMS

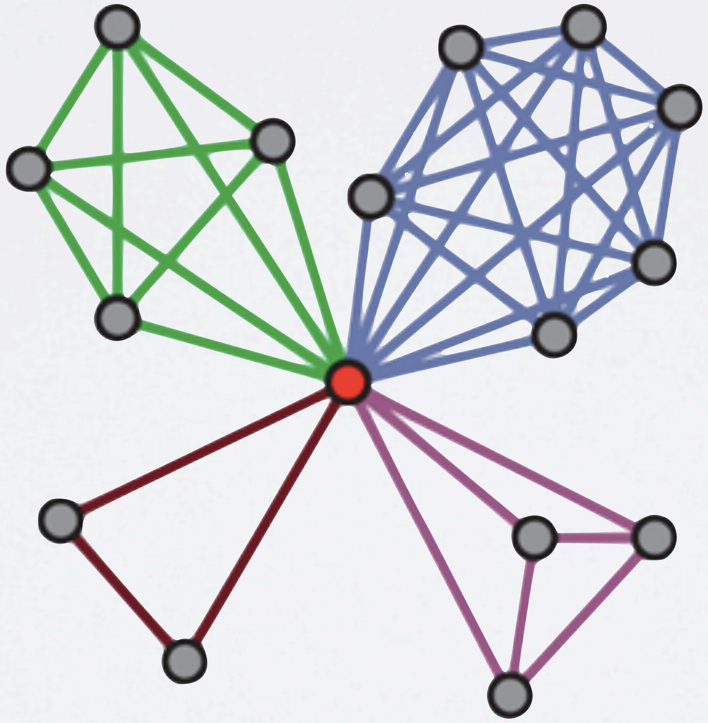
Most existing community detection methods:

- are **node-based**
- attempt to **partition** the nodes
- look at connectivity **within** a community.

This means that they are likely to overlook highly overlapping communities, and communities that are defined by relationships of one node set to another.

EDGE-BASED COMMUNITIES

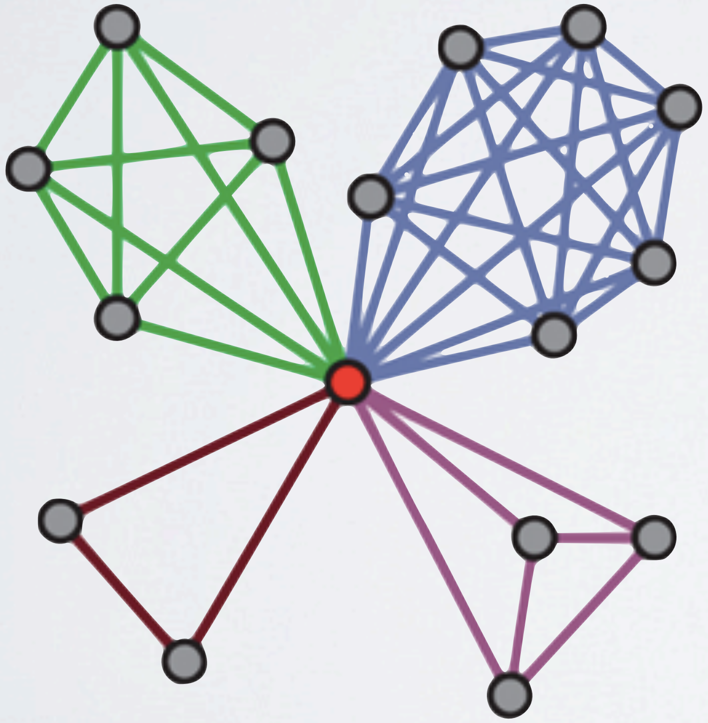
Recent work by Ahn, Bagrow & Lehmann introduced link communities to deal with the problem of overlapping communities.



Nature 466, 761-764 (2010)
(figure from paper)

EDGE-BASED COMMUNITIES

Their approach is to compare pairs of edges that share a node. If the two nodes at the other ends of these edges share many neighbours, the edges are similar.



$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}$$

Nature 466, 761-764 (2010)
(figure from paper)

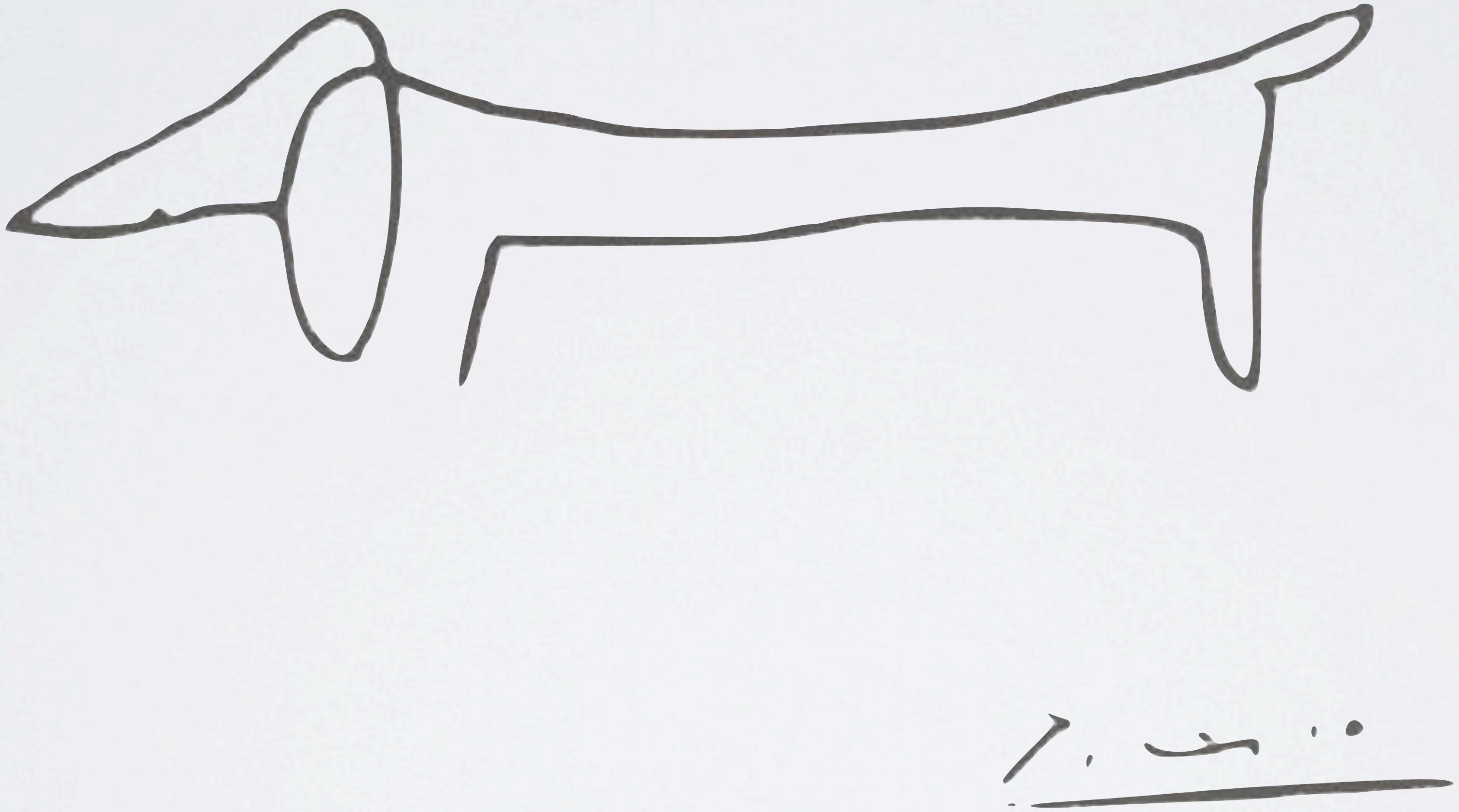
EDGE-BASED COMMUNITIES

Their approach is to compare pairs of edges that share a node. If the two nodes at the other ends of these edges share many neighbours, the edges are similar.

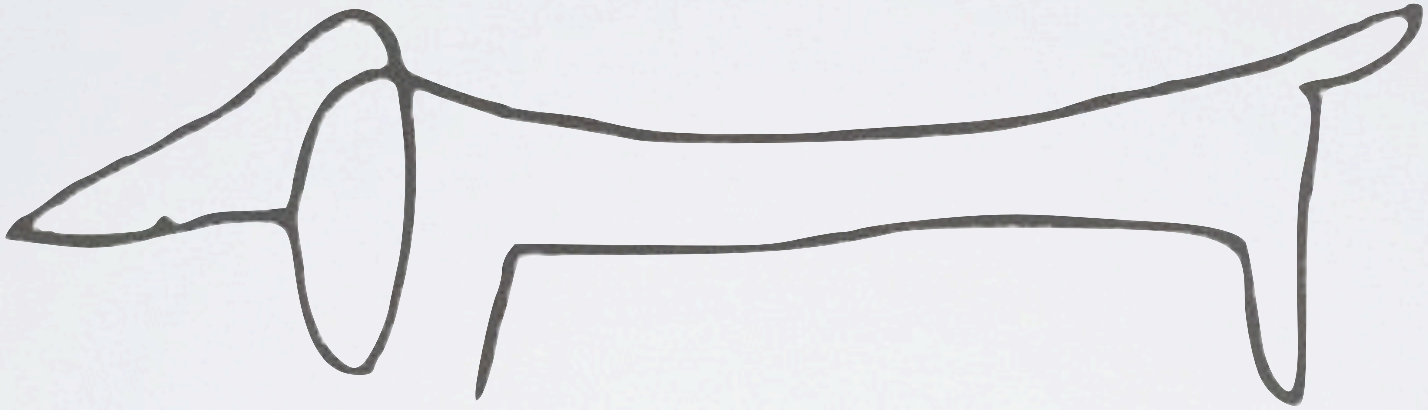
$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}$$

Nature 466, 761-764 (2010)
(figure from paper)

WHAT IS A SKETCH?



WHAT IS A SKETCH?

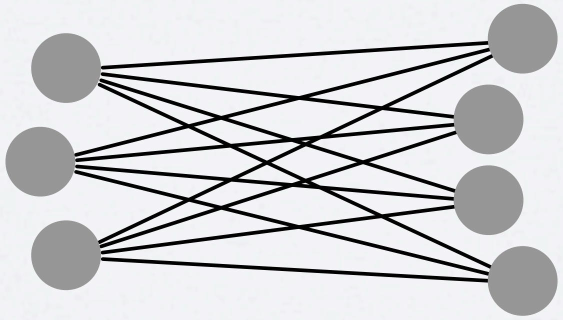


A sketch is a **compression**.

POWER-EDGES

A power-edge is a edge between two sets of nodes, known as power-nodes. It signifies that all the nodes in one set are connected to all the nodes in the other set.

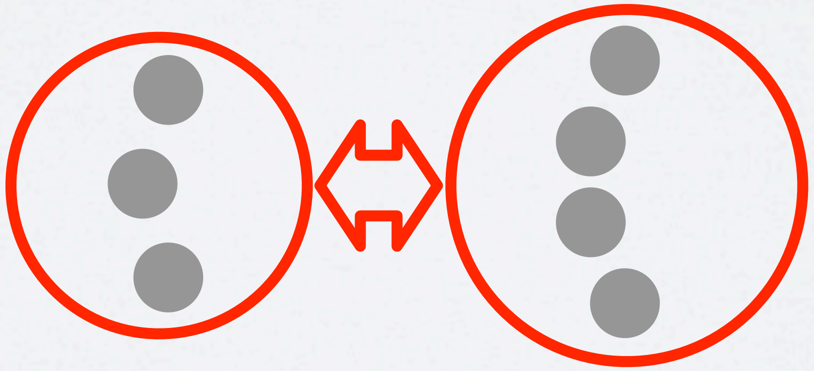
This does not mean that the nodes within a set have to be connected to each other.



POWER-EDGES

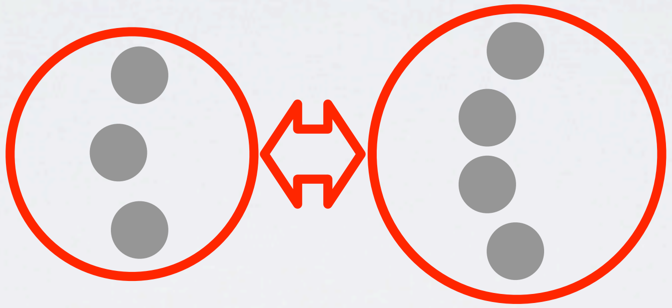
A power-edge is a edge between two sets of nodes, known as power-nodes. It signifies that all the nodes in one set are connected to all the nodes in the other set.

This does not mean that the nodes within a set have to be connected to each other.



COMPRESSING NETWORKS

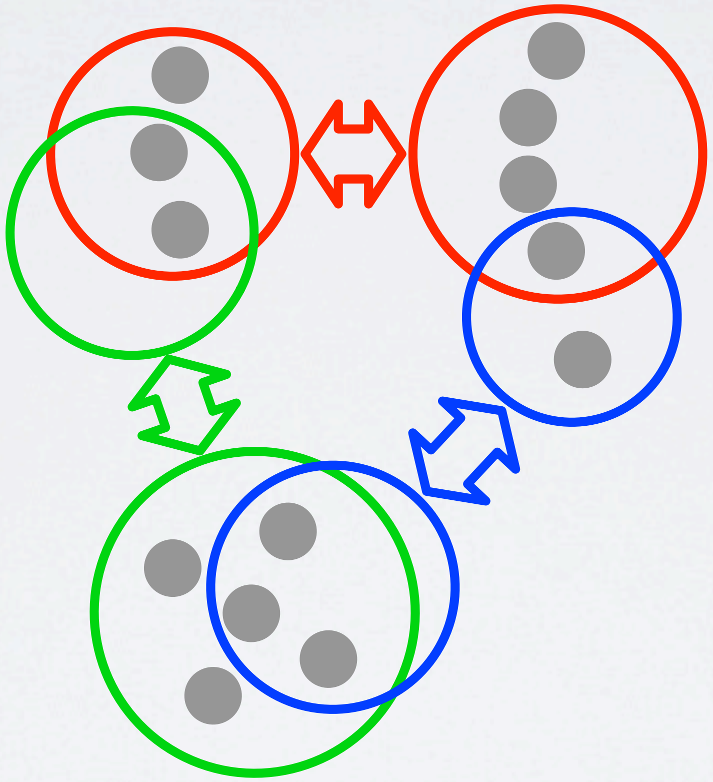
These power-nodes can overlap. Note that self-interactions are also possible.



How do we choose a suitable set of power-edges to represent the graph?

COMPRESSING NETWORKS

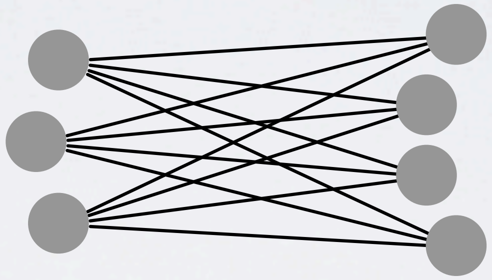
These power-nodes can overlap. Note that self-interactions are also possible.



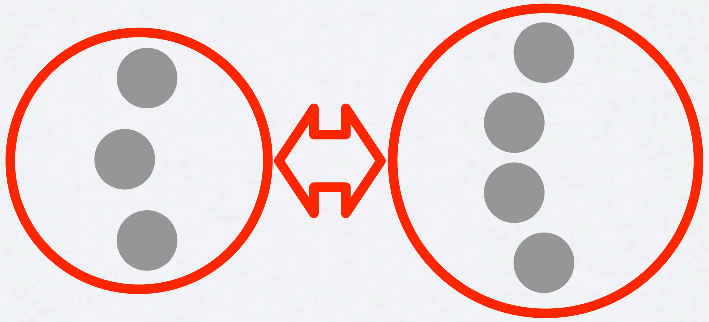
How do we choose a suitable set of power-edges to represent the graph?

COMPRESSIBLE COMPONENTS

Let us compare the original set of edges

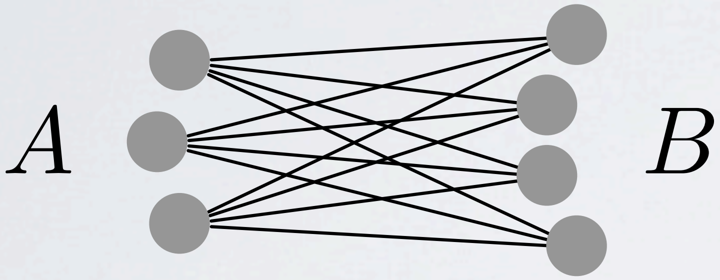


with the power-edge:



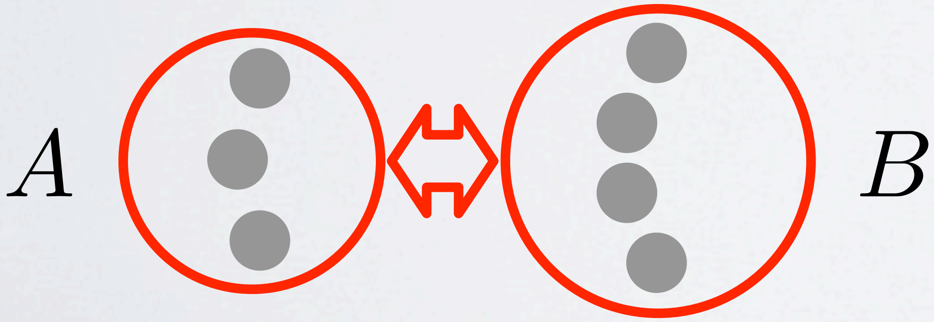
COMPRESSIBLE COMPONENTS

The information required to describe these is:



$$I_{AB} = 2 n_A n_B \log_2 N$$

$$= 24 \log_2 7$$

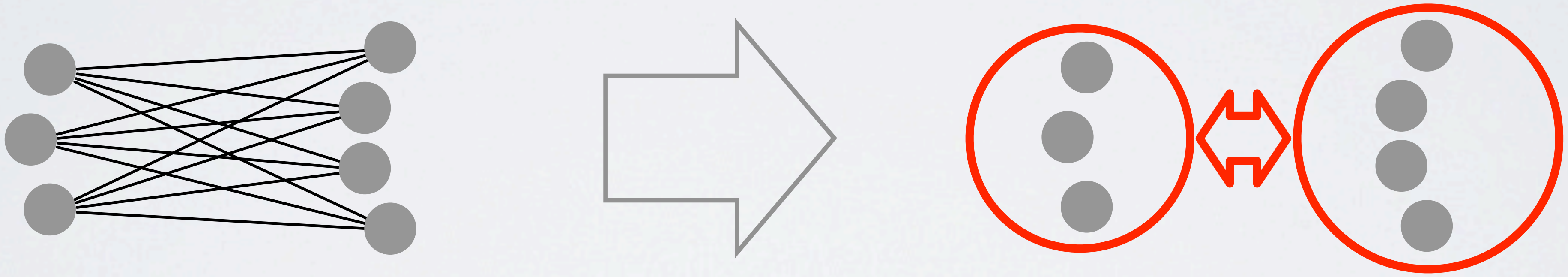


$$I^{(p)}_{AB} = 2(n_A + n_B) \log_2 N$$

$$= 14 \log_2 7$$

COMPRESSIBLE COMPONENTS

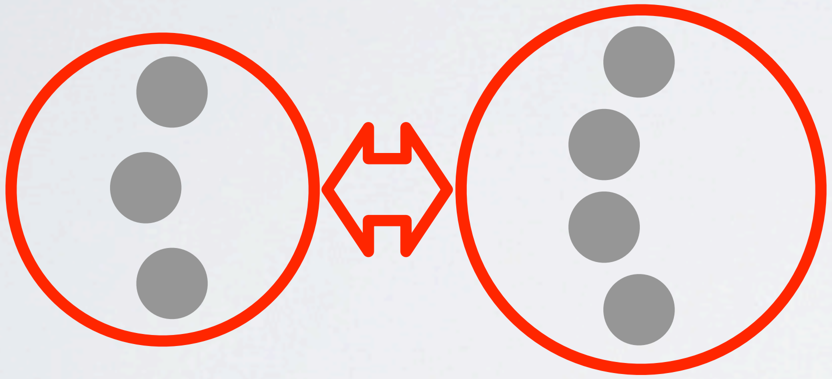
Hence the compression is given by:



$$\Delta I_{AB} = I_{AB}^{(p)} - I_{AB} = 2(n_A n_B - n_A - n_B)\log_2 N$$

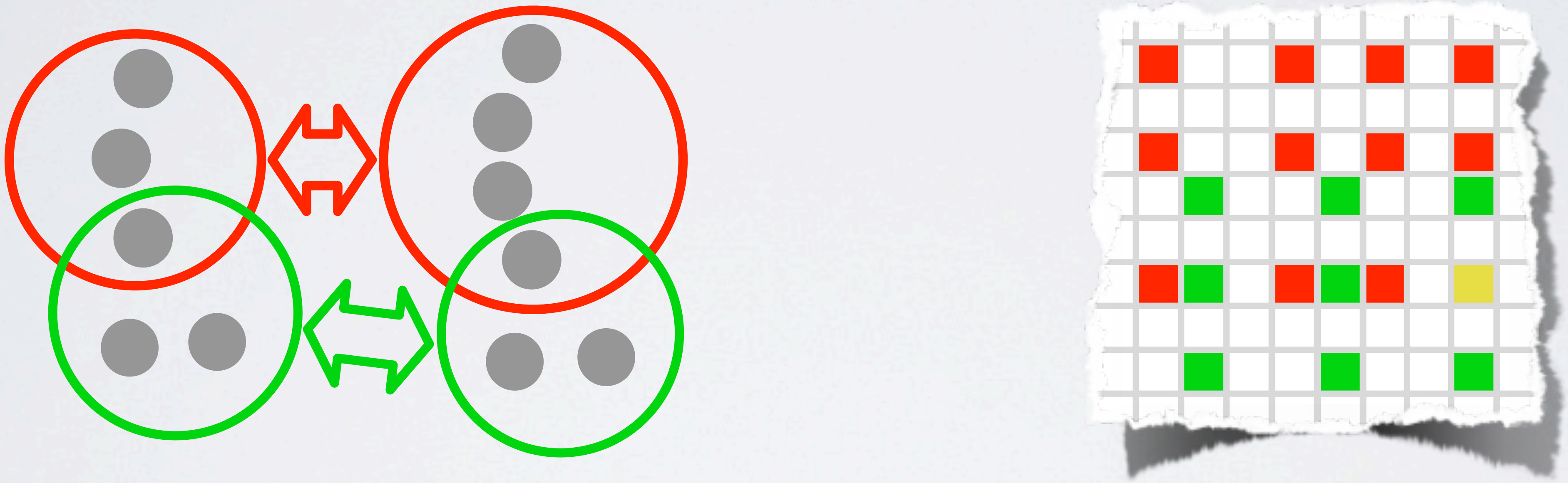
$$= (24 - 14)\log_2 7 = 10\log_2 7$$

COMPRESSIBLE COMPONENTS

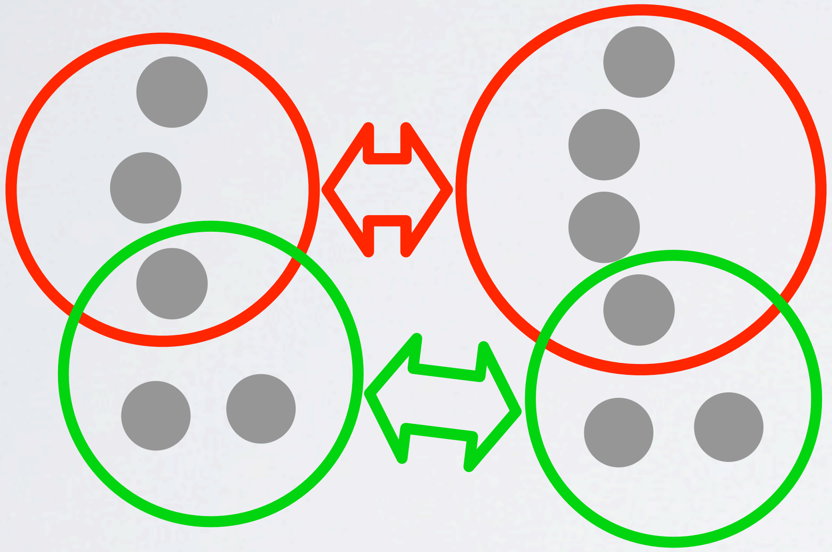


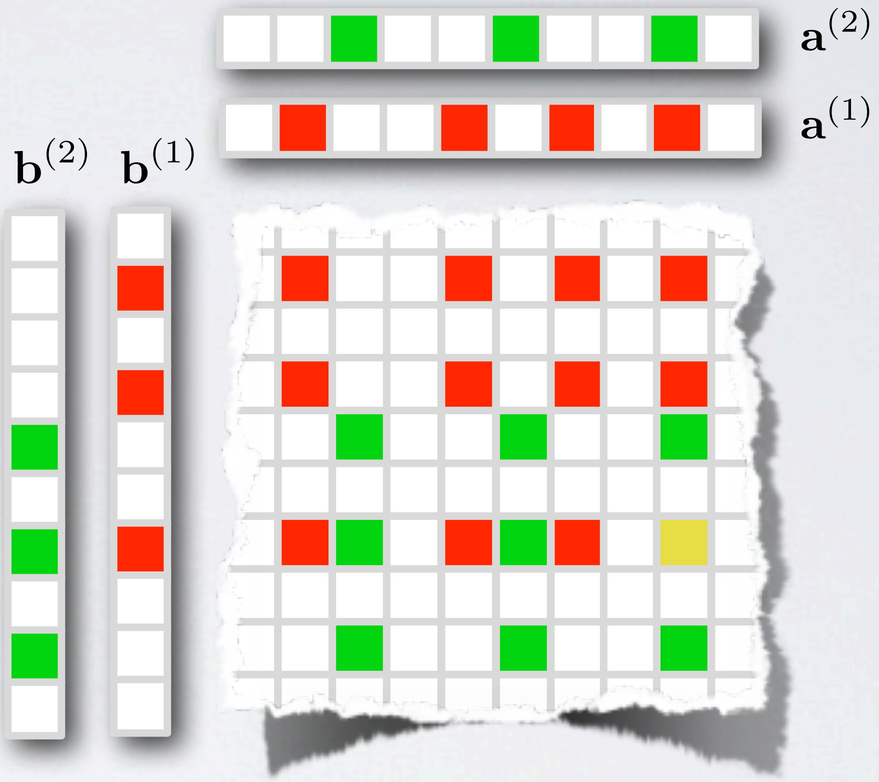
COMPRESSIBLE COMPONENTS

COMPRESSIBLE COMPONENTS



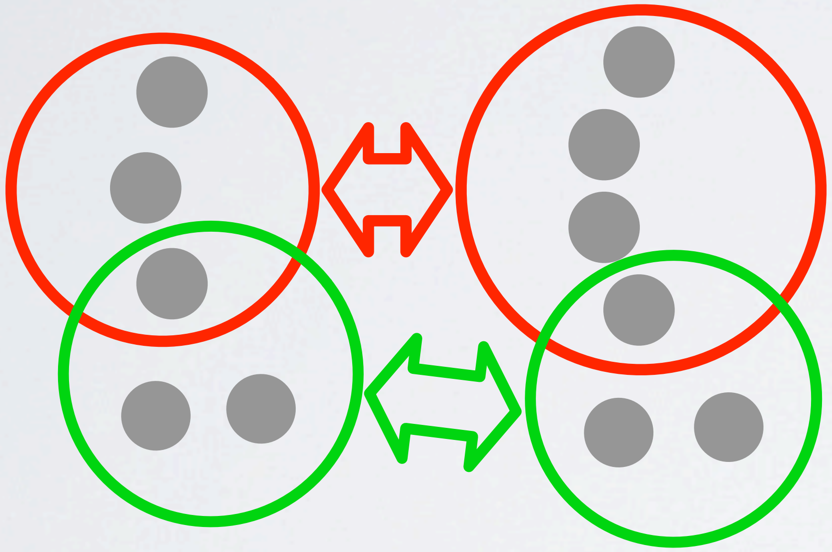
COMPRESSIBLE COMPONENTS

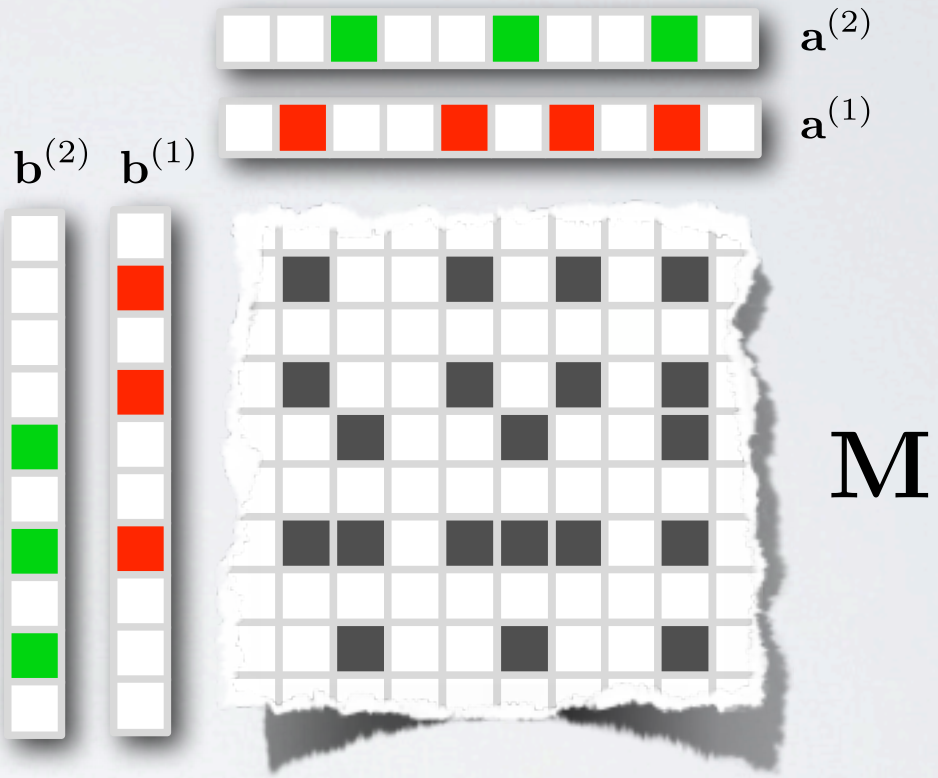




$$v_{ij} = \frac{1}{2}\Big(\sum_{k \in S} a_i^{(k)} b_j^{(k)} - m_{ij} + \big|\sum_{k \in S} a_i^{(k)} b_j^{(k)} - m_{ij}\big|\Big)$$

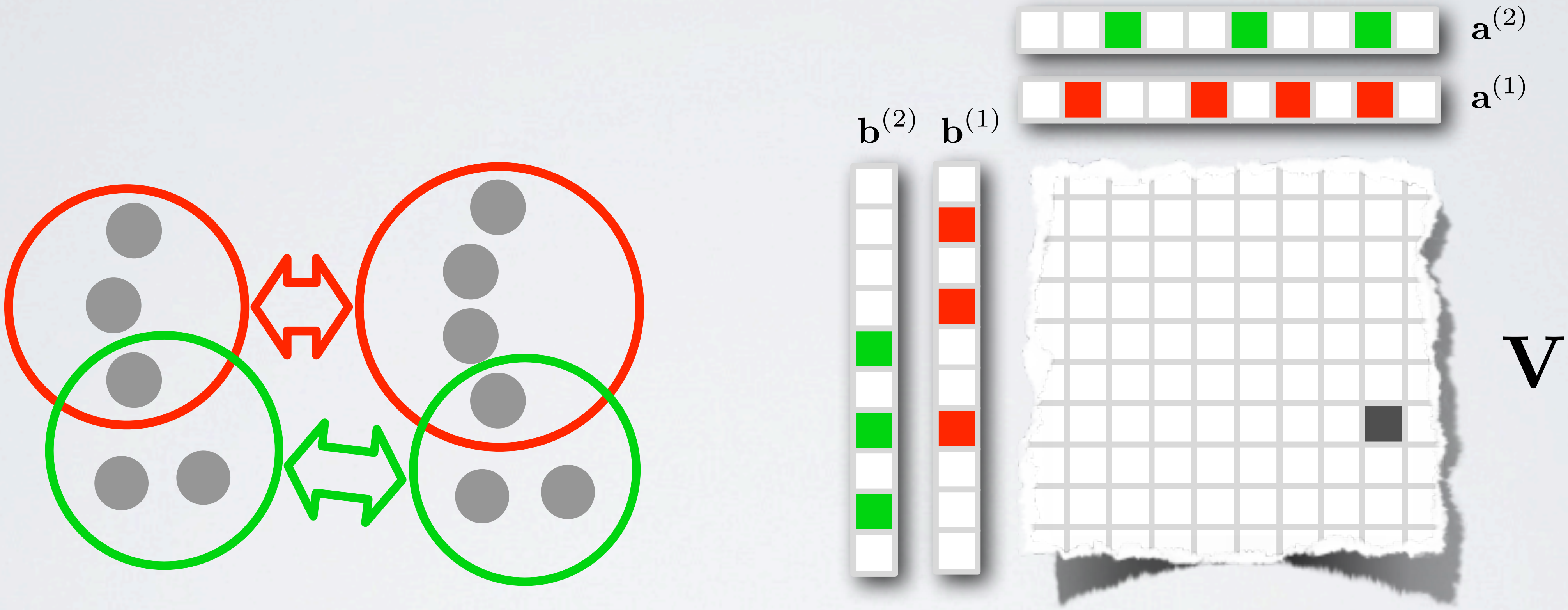
COMPRESSIBLE COMPONENTS





$$v_{ij} = \frac{1}{2}\Big(\sum_{k\in S} a_i^{(k)} b_j^{(k)} - m_{ij} + \big|\sum_{k\in S} a_i^{(k)} b_j^{(k)} - m_{ij}\big|\Big)$$

COMPRESSIBLE COMPONENTS



$$v_{ij} = \frac{1}{2}(\sum_{k\in S} a_i^{(k)} b_j^{(k)} - m_{ij} + |\sum_{k\in S} a_i^{(k)} b_j^{(k)} - m_{ij}|)$$

$$\Delta I_{total} = \sum_{k\in S} \Delta I_k - (\mathbf{1}^T \mathbf{V} \mathbf{1})\, 2 \log_2 N$$

COMPRESSIBLE COMPONENTS

$$\Delta I_{total} = \sum_{k \in S} \Delta I_k - (\mathbf{1}^T \mathbf{V} \mathbf{1})\, 2 \log_2 N$$

We then maximise this quantity using a greedy algorithm:

1) Rank all possible compressible components in order of compressibility.
2) Select most compressible component.
3) Recalculate compressibilities of all others in light of overlaps with selected component.
4) Repeat from step 1) for all compressible components apart from selected one(s).

SOCIAL NETWORKS

A classical social network dataset for community detection is Zachary's karate club network. The shown communities were calculated by Newman & Girvan.

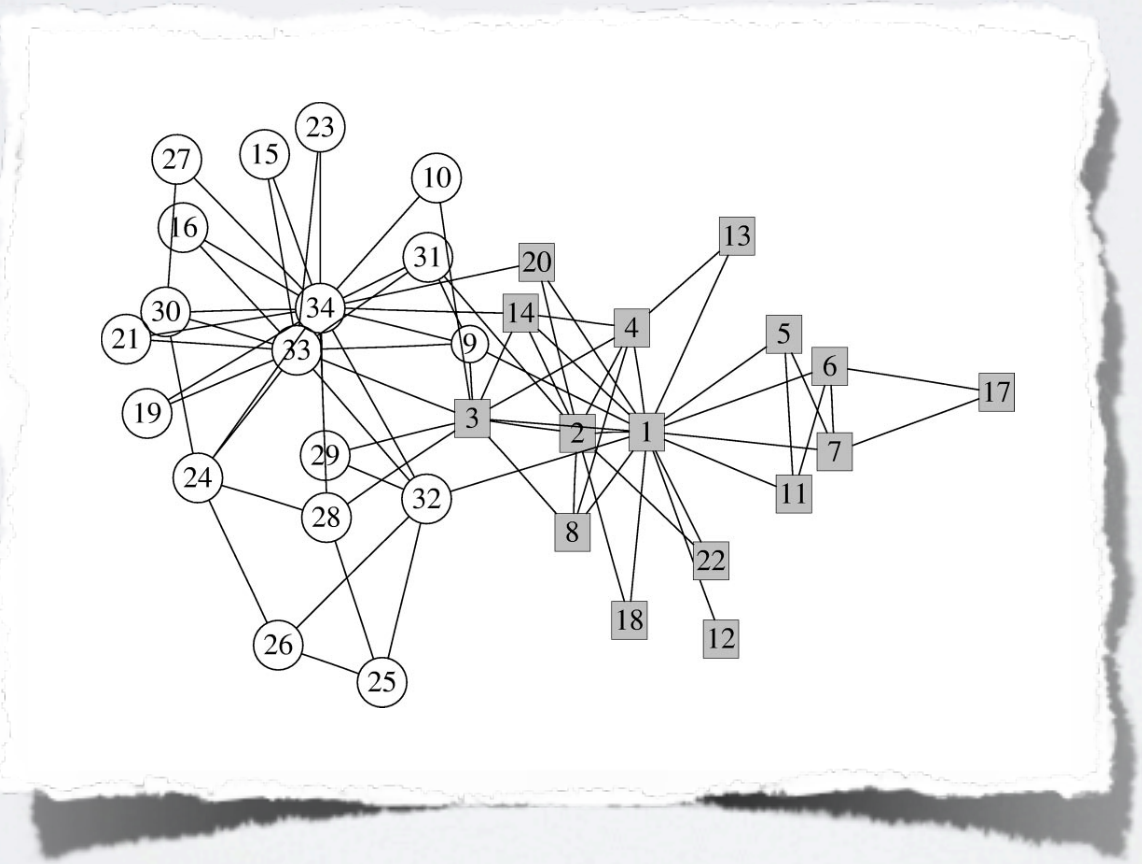


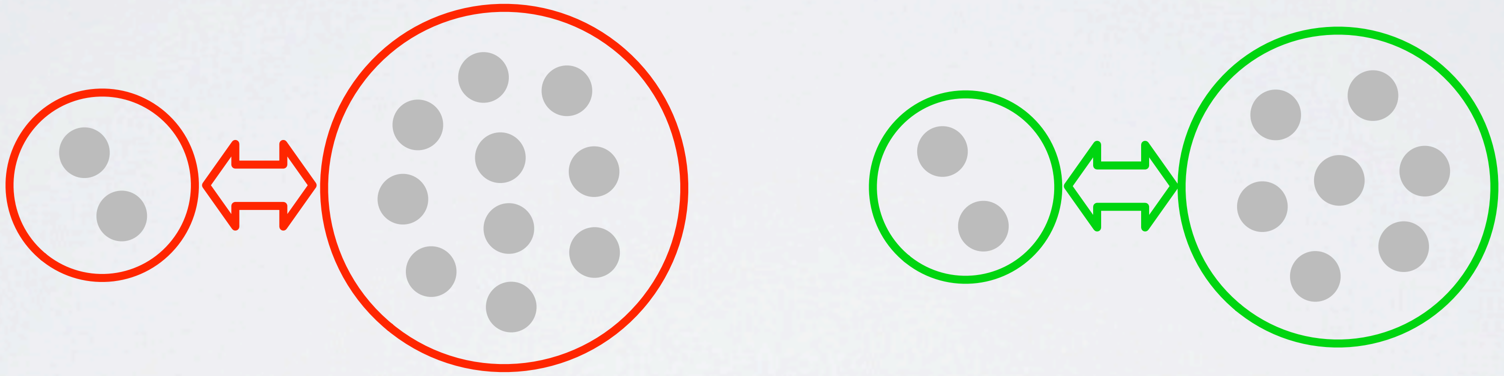
Figure from M. E. J. Newman & M. Girvan, PRE 69, 026113 (2004)

SOCIAL NETWORKS

What are the most compressible components in the karate club network?

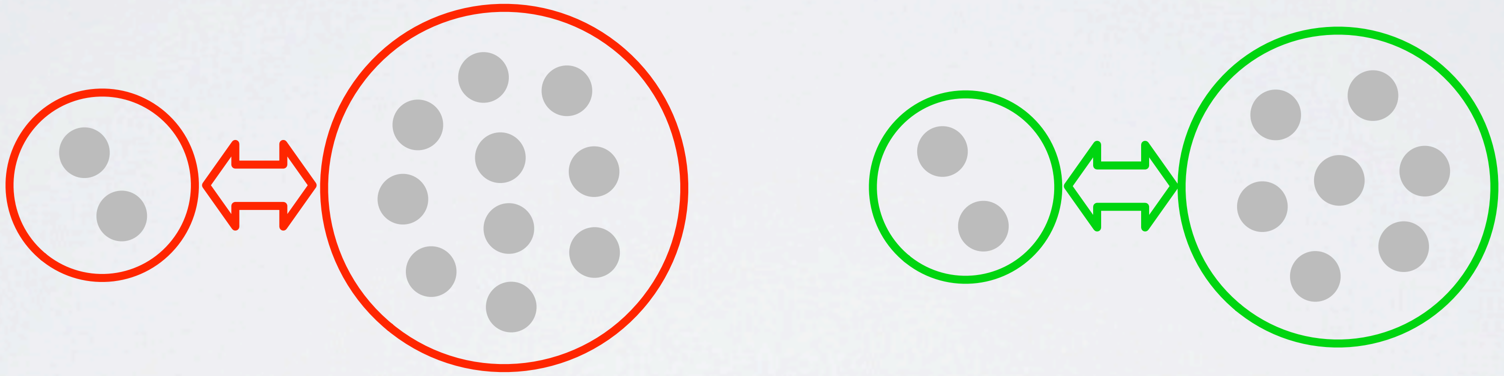
SOCIAL NETWORKS

What are the most compressible components in the karate club network?



SOCIAL NETWORKS

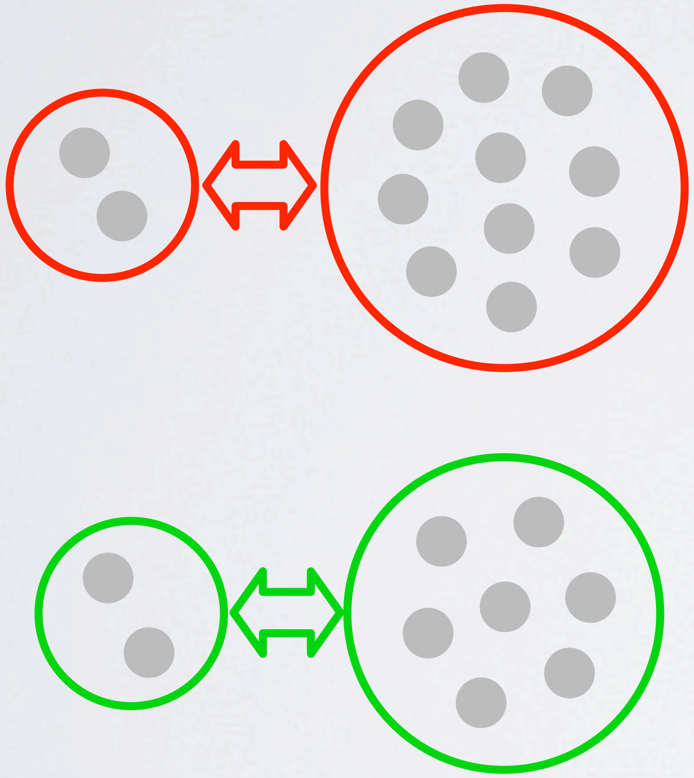
What are the most compressible components in the karate club network?



What is the meaning of these compressible components?

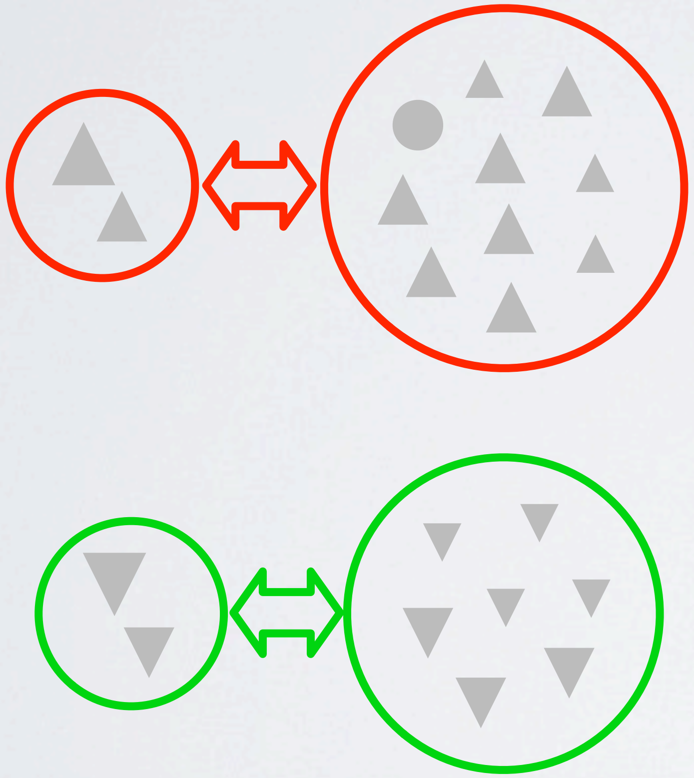
SOCIAL NETWORKS

What is the meaning of these compressible components?

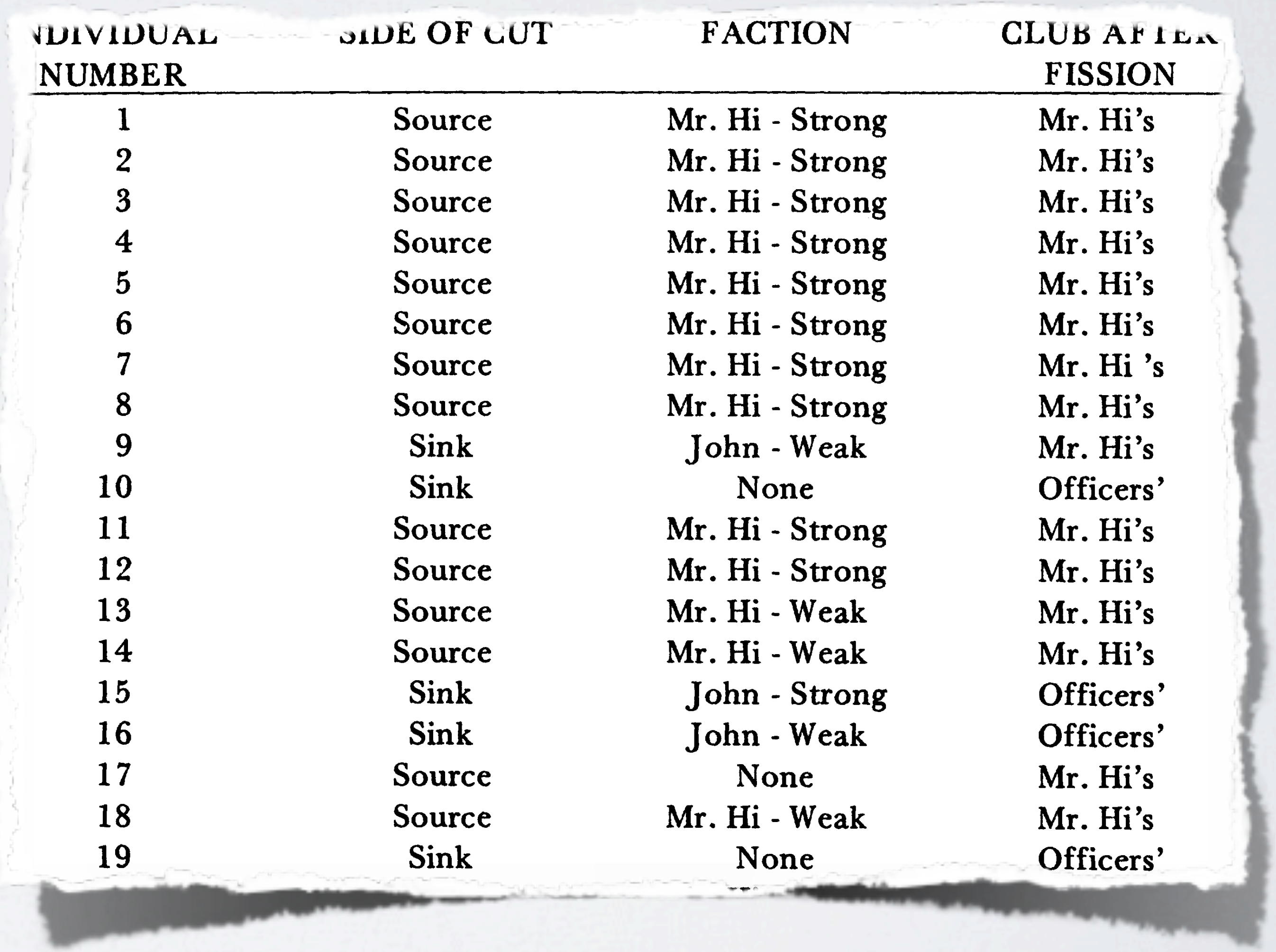


SOCIAL NETWORKS

What is the meaning of these compressible components?



The two factions and their leaders.

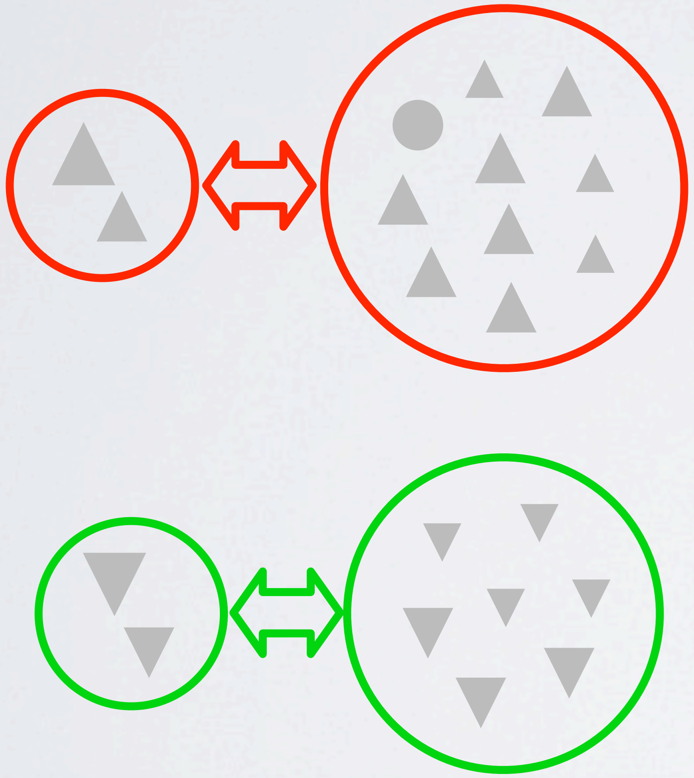


NDIVIDUAL NUMBER	SIDE OF CUT	FACTION	CLUB AFTER FISSION
1	Source	Mr. Hi - Strong	Mr. Hi’s
2	Source	Mr. Hi - Strong	Mr. Hi’s
3	Source	Mr. Hi - Strong	Mr. Hi’s
4	Source	Mr. Hi - Strong	Mr. Hi’s
5	Source	Mr. Hi - Strong	Mr. Hi’s
6	Source	Mr. Hi - Strong	Mr. Hi’s
7	Source	Mr. Hi - Strong	Mr. Hi ’s
8	Source	Mr. Hi - Strong	Mr. Hi’s
9	Sink	John - Weak	Mr. Hi’s
10	Sink	None	Officers’
11	Source	Mr. Hi - Strong	Mr. Hi’s
12	Source	Mr. Hi - Strong	Mr. Hi’s
13	Source	Mr. Hi - Weak	Mr. Hi’s
14	Source	Mr. Hi - Weak	Mr. Hi’s
15	Sink	John - Strong	Officers’
16	Sink	John - Weak	Officers’
17	Source	None	Mr. Hi’s
18	Source	Mr. Hi - Weak	Mr. Hi’s
19	Sink	None	Officers’

from W. Zachary, J. Anthrop. Res. 33, 452-473 (1977)l

SOCIAL NETWORKS

What is the meaning of these compressible components?



The two factions
and their leaders.

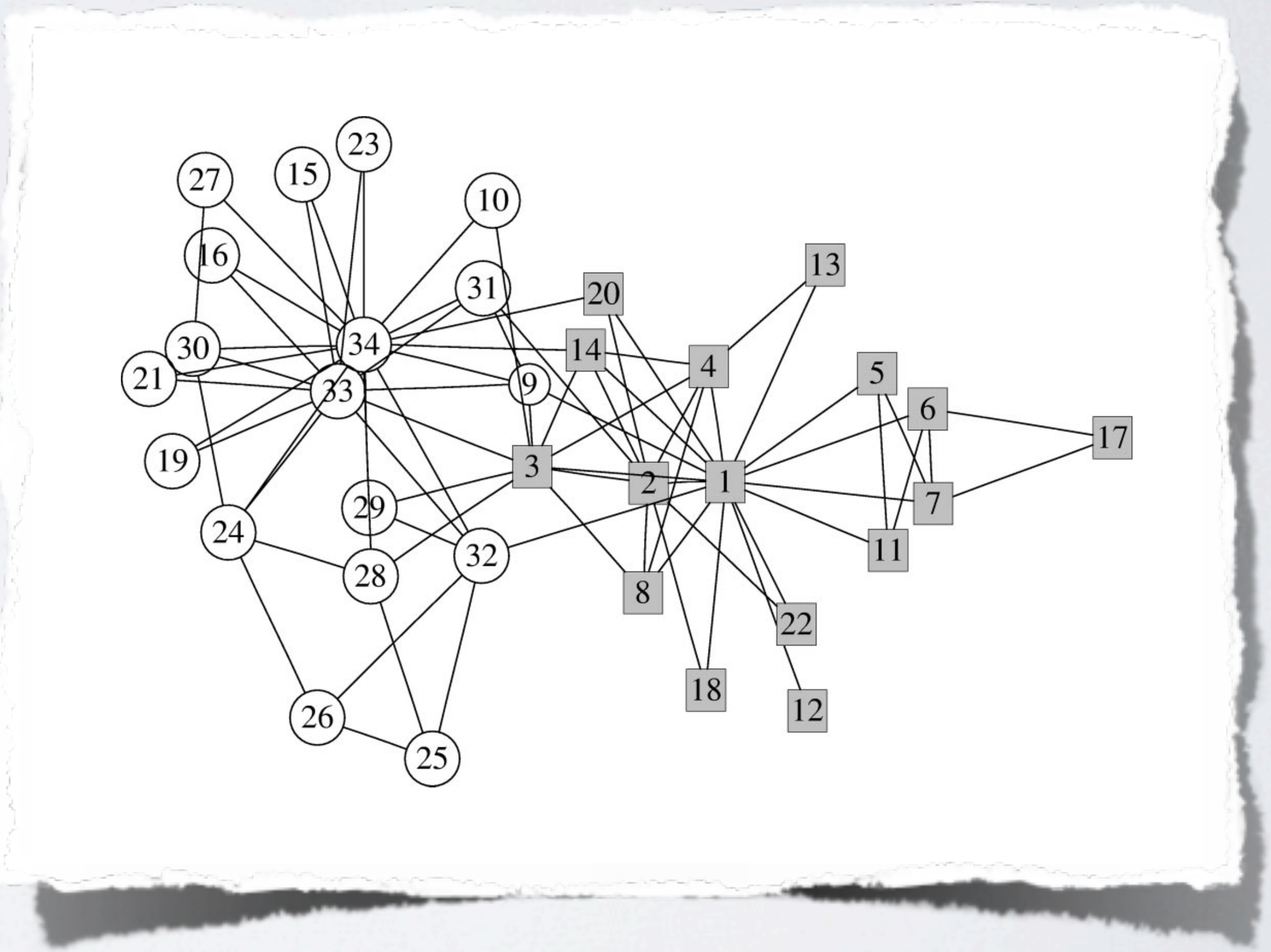


Figure from M. E. J. Newman & M. Girvan, PRE 69, 026113 (2004)

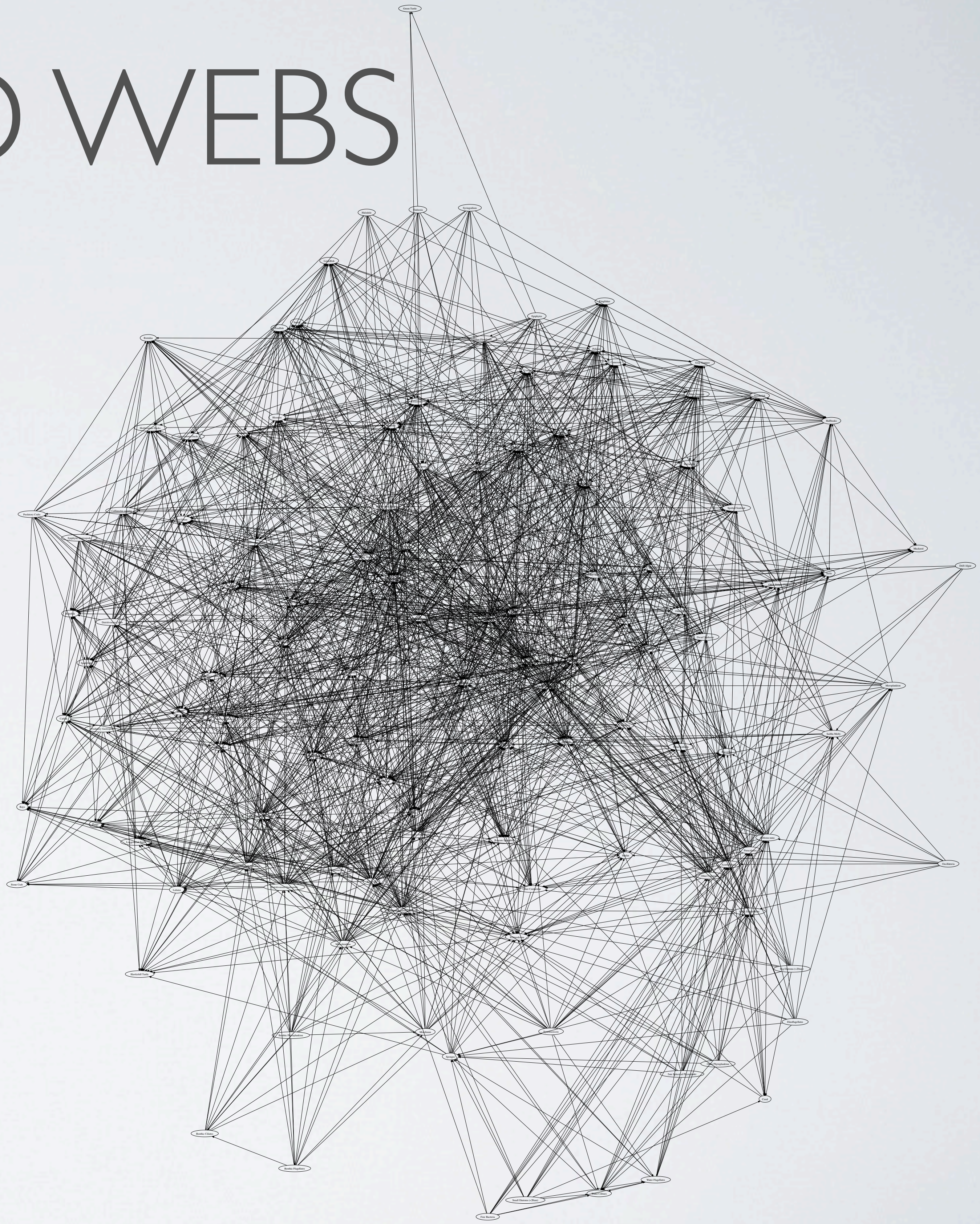
FOOD WEBS

Food webs are ecological networks of predator-prey relationships between organisms.

FOOD WEBS

Food webs are ecological networks of predator-prey relationships between organisms.

Here we will show results for a food web of 122 species in a Florida ecosystem, with 1767 edges.



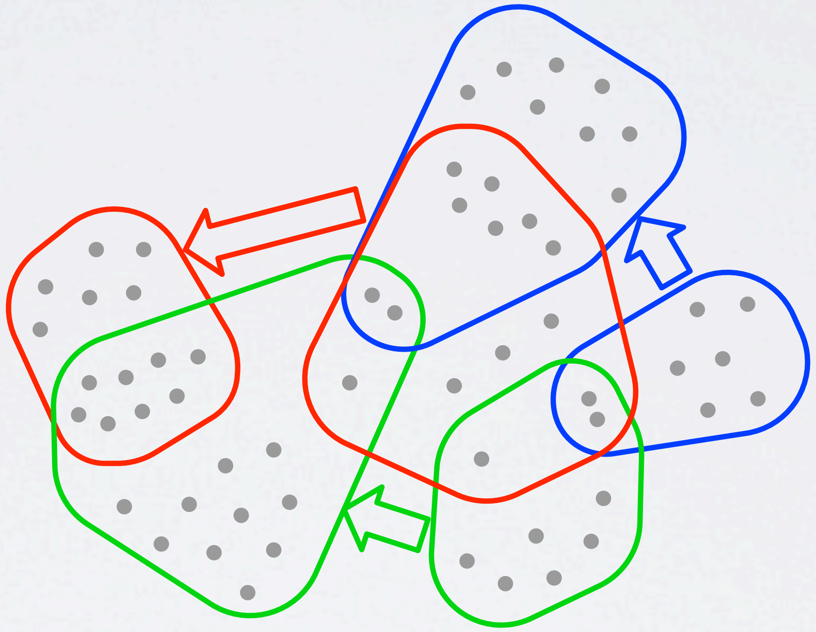
R. E. Ulanowicz, C. Bondavalli, and M. S. Egnotovich, Chesapeake Biological Laboratory, Solomons, MD 20688-0038, USA, Ref. No. UMCESCBL 98-123, 1998

FOOD WEBS

The most compressible components of this food web:

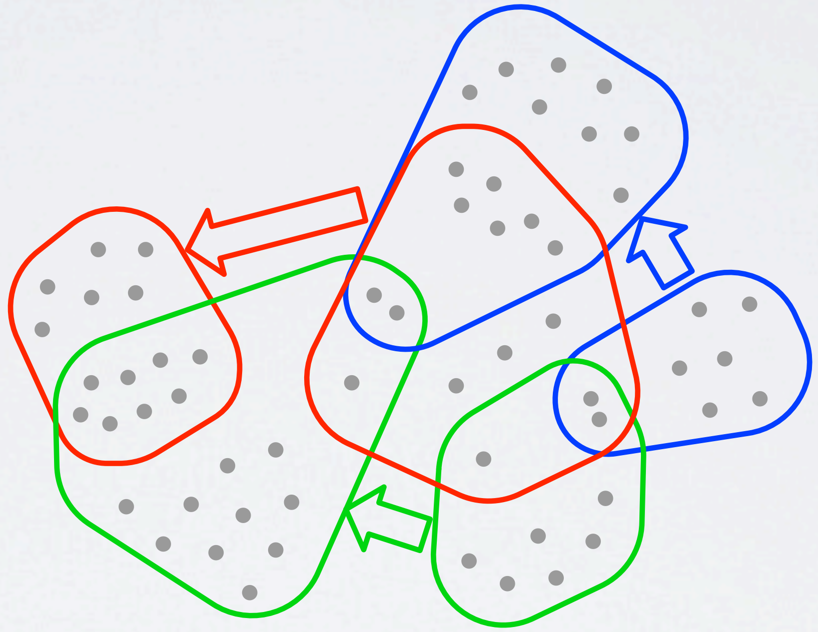
FOOD WEBS

The most compressible components of this food web:



FOOD WEBS

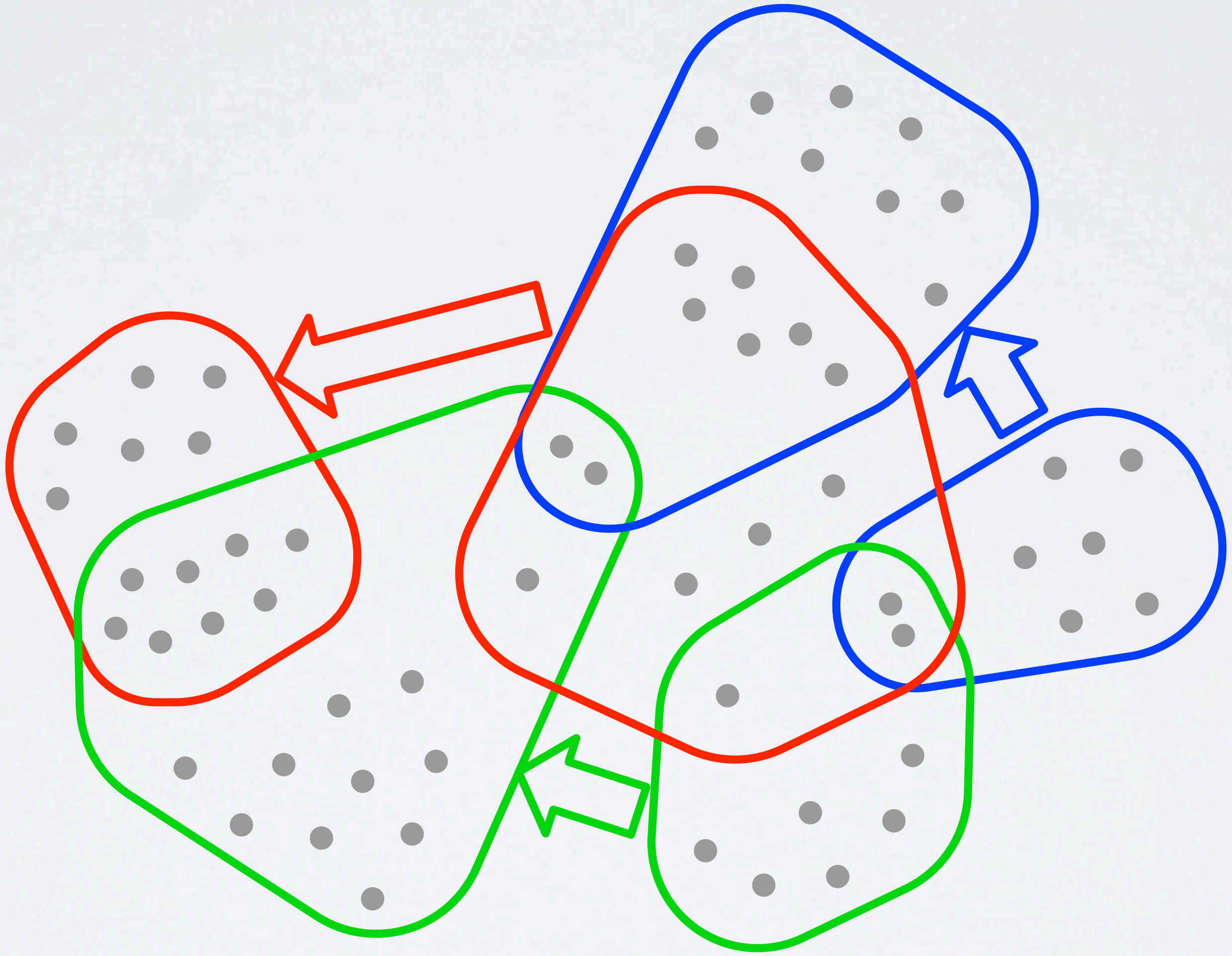
The most compressible components of this food web:



501 edges
28% of network

FOOD WEBS

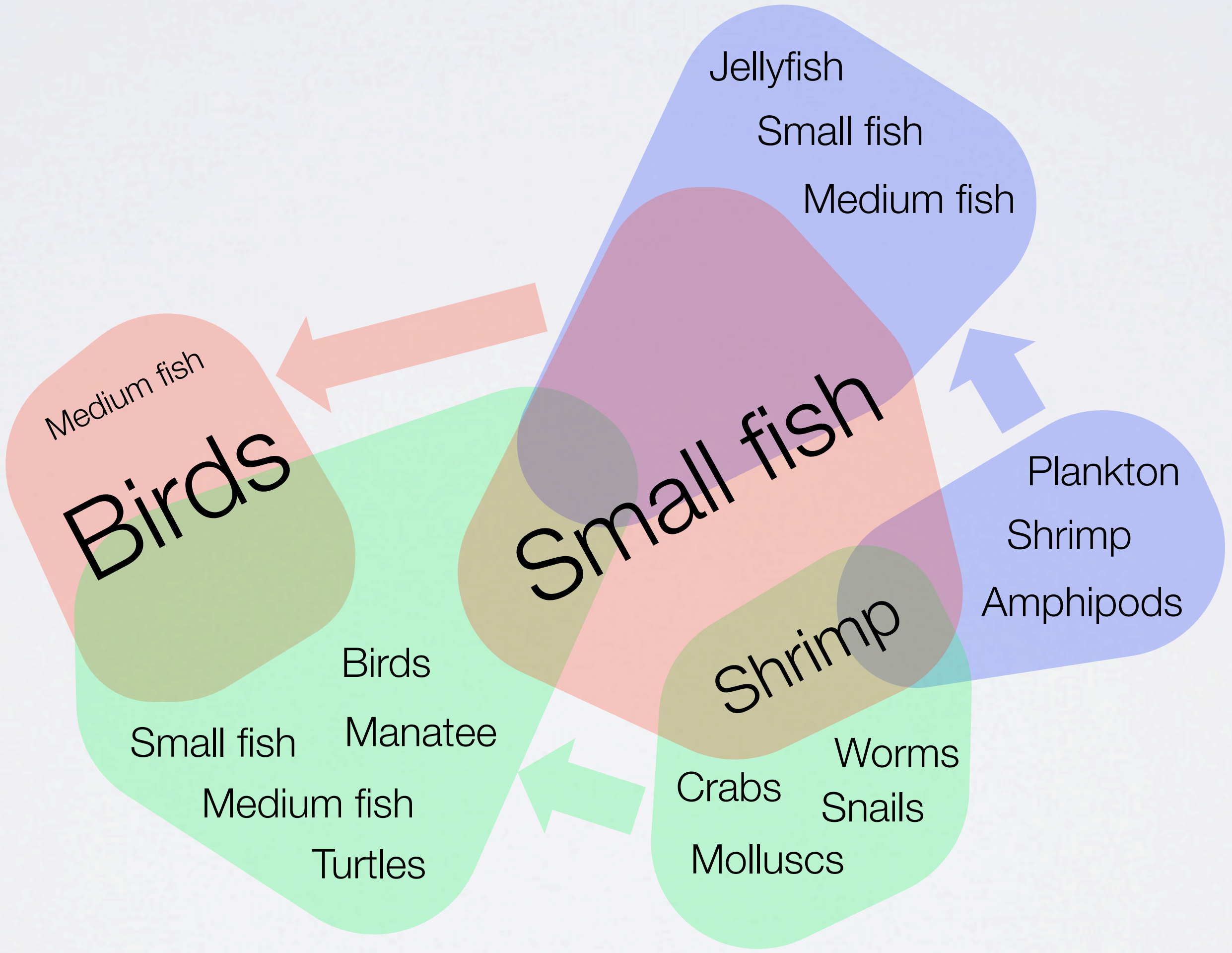
What do these sets mean?



Can we assign meaning in a more rigorous way?

FOOD WEBS

What do these sets mean?



Can we assign meaning in a more rigorous way?

TERM ENRICHMENT

Given a set of terms associated with the nodes we can calculate the enrichment of these relative to a null model.

TERM ENRICHMENT

Given a set of terms associated with the nodes we can calculate the enrichment of these relative to a null model.

The probability of finding term i in set s exactly k times, if the term appears n_i times among all N nodes, is

$$p_i(k,s) = \frac{{}^{n_i}C_k \; {}^{N-n_i}C_{s-k}}{{}^{N}C_s}$$

TERM ENRICHMENT

Given a set of terms associated with the nodes we can calculate the enrichment of these relative to a null model.
The probability of finding term i in set s exactly k times, if the term appears n_i times among all N nodes, is

$$p_i(k,s) = \frac{{}^{n_i}C_k \; {}^{N-n_i}C_{s-k}}{{}^{N}C_s}$$

p-value: $P_i(k^*,s) = 1 - \sum_{k=0}^{k^*-1} p_i(k,s)$

TERM ENRICHMENT

Given a set of terms associated with the nodes we can calculate the enrichment of these relative to a null model.

The probability of finding term i in set s exactly k times, if the term appears n_i times among all N nodes, is

$$p_i(k,s) = \frac{{}^{n_i}C_k \; {}^{N-n_i}C_{s-k}}{{}^{N}C_s}$$

$$P_i(k^*,s) = 1 - \sum_{k=0}^{k^*-1} p_i(k,s) \qquad P_i^{(B)}(k^*,s) = T P_i(k^*,s)$$

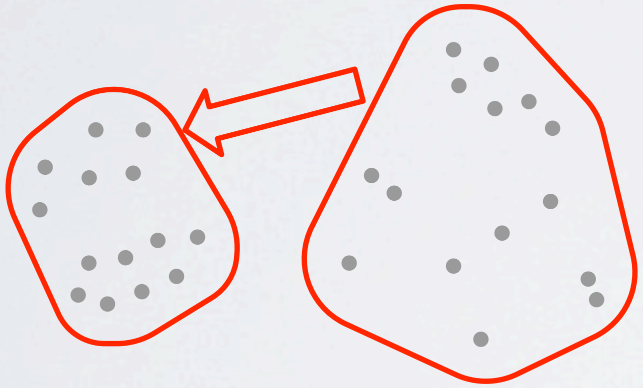
Bonferroni

FOOD WEB TERM ENRICHMENT

Some of the enriched terms in the compressible components of the food web:

FOOD WEB TERM ENRICHMENT

Some of the enriched terms in the compressible components of the food web:



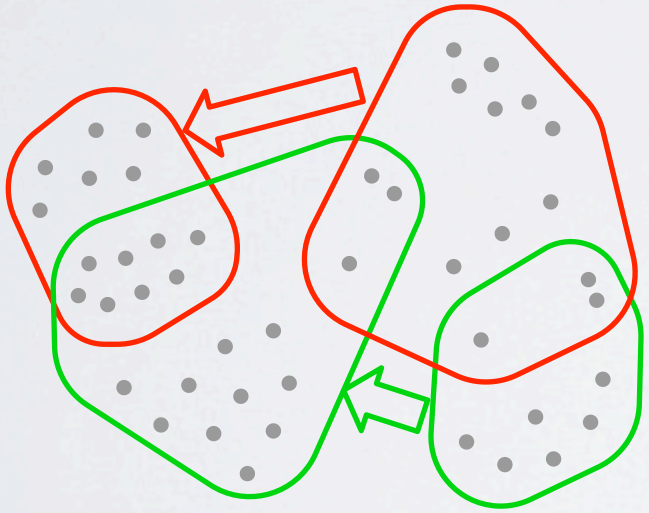
small fish
p-value: 8e-06

⇨

birds
p-value: 2e-12

FOOD WEB TERM ENRICHMENT

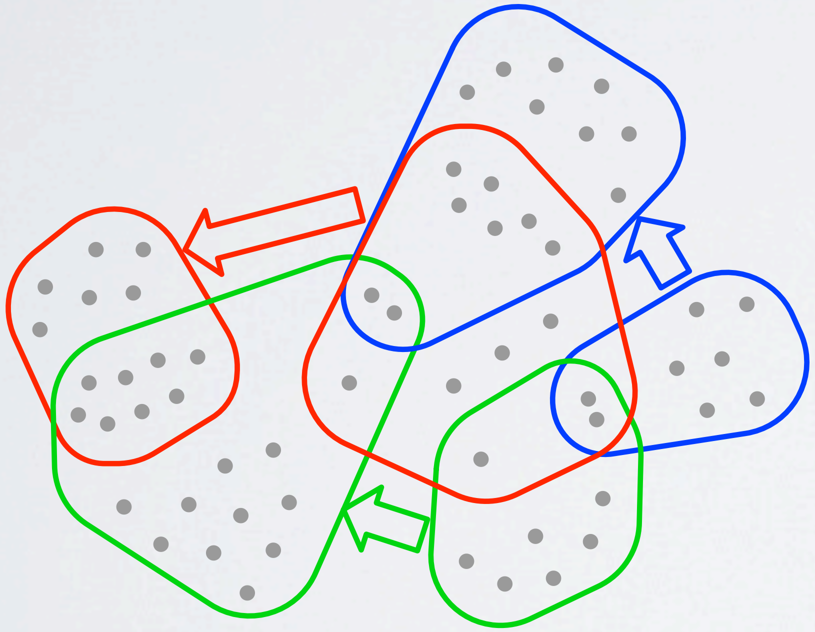
Some of the enriched terms in the compressible components of the food web:



small fish p-value: 8e-06	⇨	birds p-value: 2e-12
worms p-value: 8e-03	⇨	birds p-value: 4e-04

FOOD WEB TERM ENRICHMENT

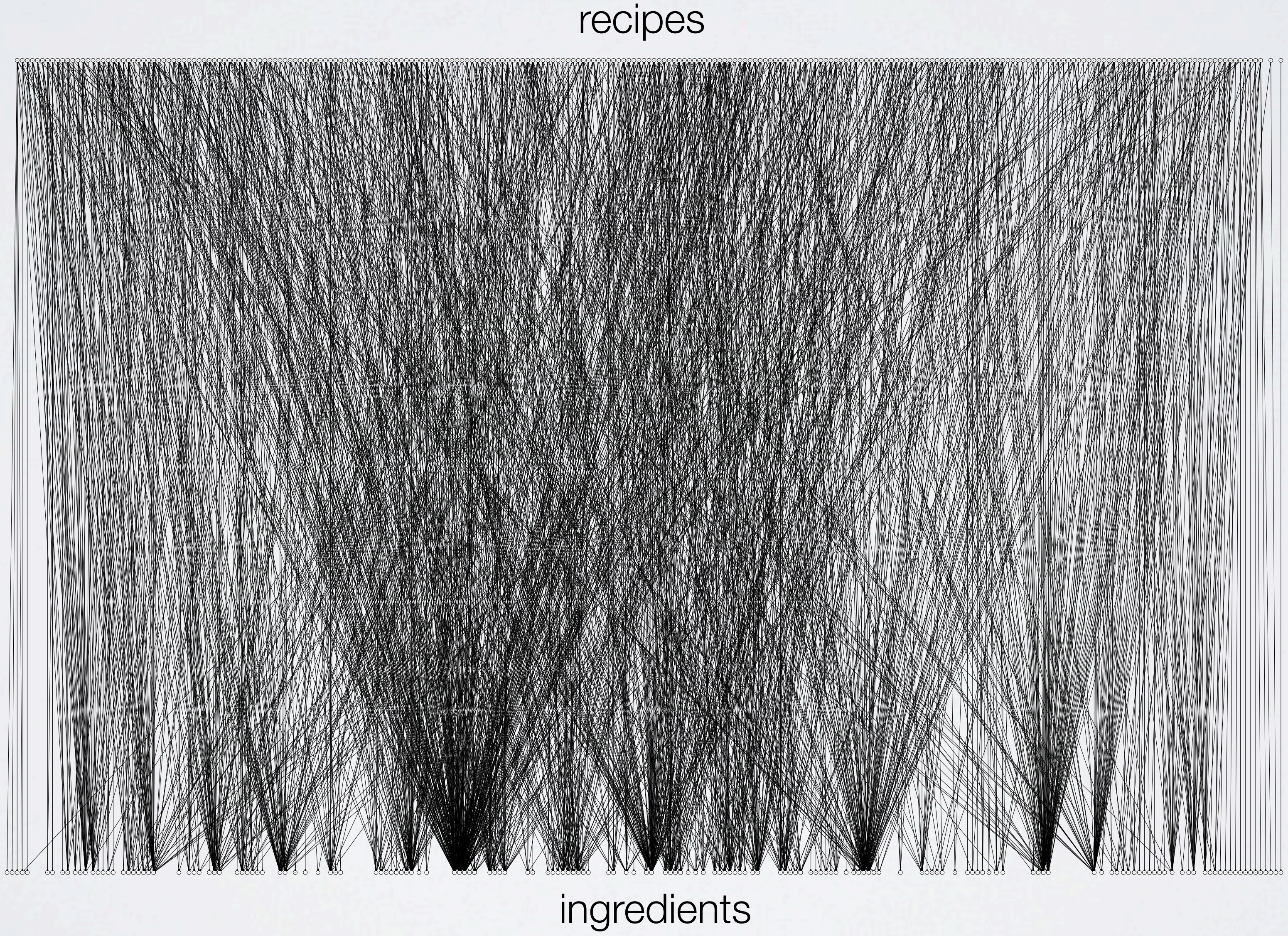
Some of the enriched terms in the compressible components of the food web:



small fish p-value: 8e-06	⇨	birds p-value: 2e-12
worms p-value: 8e-03	⇨	birds p-value: 4e-04
copepods p-value: 4e-04	⇨	small fish p-value: 2e-07

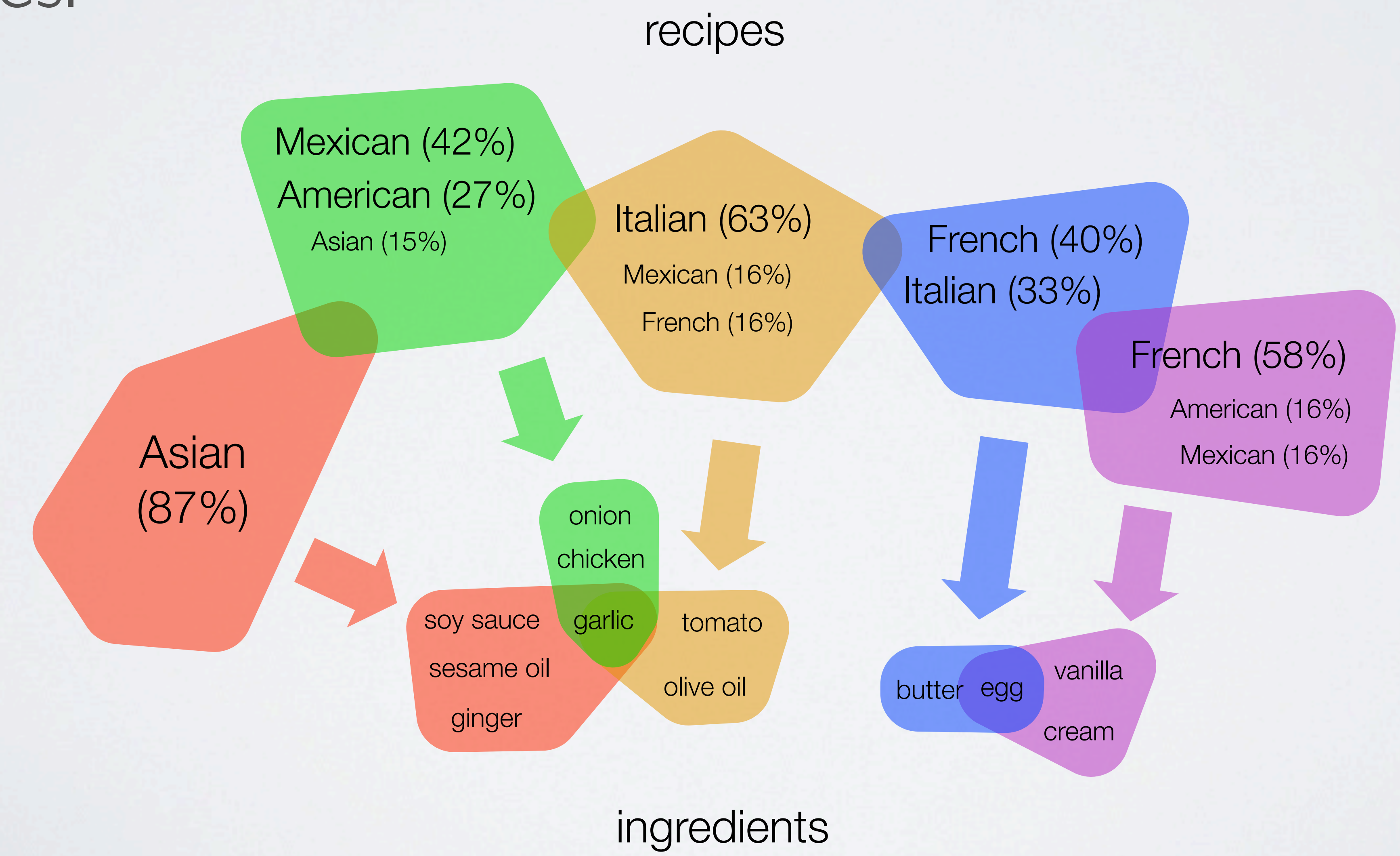
RECIPE NETWORK

Food ingredients and recipes form a bipartite network. Here we study an example with 415 nodes and 1748 edges.



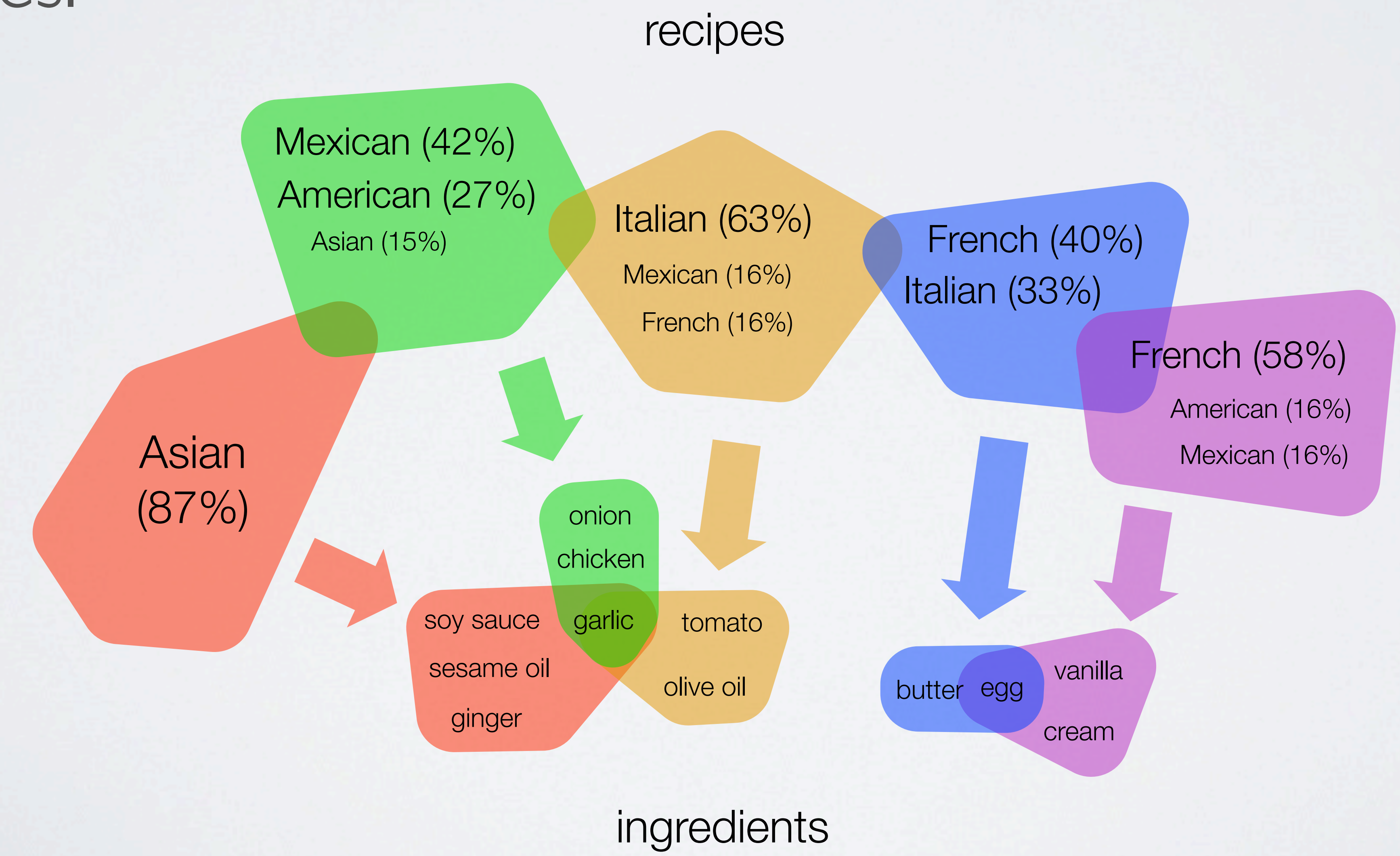
RECIPE NETWORK

The five most compressible components exhibit highly significant enrichment for certain cuisines. The overlap between the power nodes shows the proximity of cuisines.



RECIPE NETWORK

The five most compressible components exhibit highly significant enrichment for certain cuisines. The overlap between the power nodes shows the proximity of cuisines.



YEAST TRANSCRIPTION

YEAST TRANSCRIPTION

In a yeast transcription network the nodes are yeast genes and the edges are regulatory interactions between them.

Our dataset here has 2534 nodes and 6071 edges.

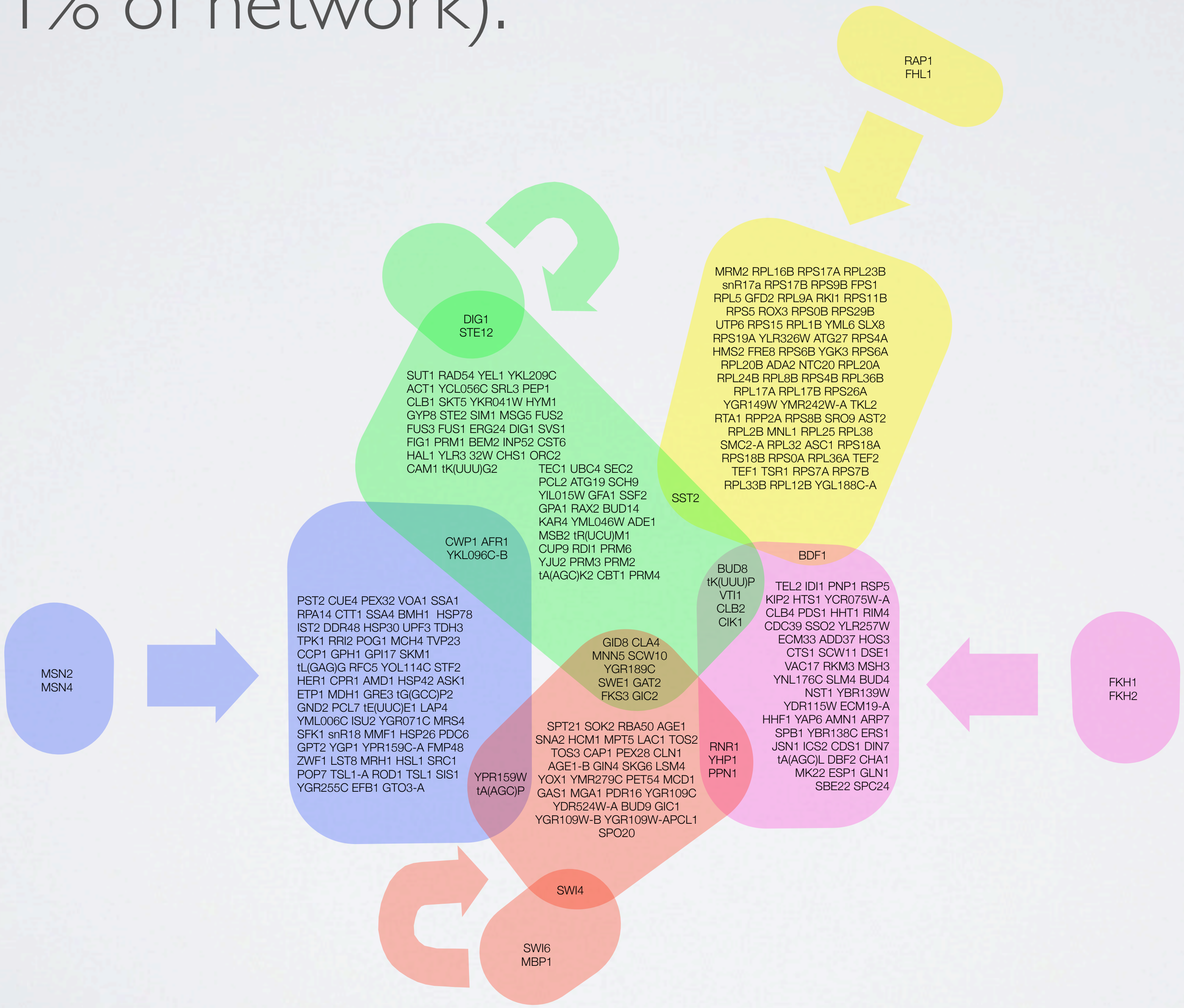
K. D. MacIsaac *et al.*, BMC Bioinformatics 7:113 (2006)

YEAST TRANSCRIPTION

The 5 most compressible components cover 699 edges (11% of network).

YEAST TRANSCRIPTION

The 5 most compressible components cover 699 edges (11% of network).



RAP1
FHL1

MRM2 RPL16B RPS17A RPL23B
snR17a RPS17B RPS9B FPS1
RPL5 GFD2 RPL9A RKI1 RPS11B
RPS5 ROX3 RPS0B RPS29B
UTP6 RPS15 RPL1B YML6 SLX8
RPS19A YLR326W ATG27 RPS4A
HMS2 FRE8 RPS6B YGK3 RPS6A
RPL20B ADA2 NTC20 RPL20A
RPL24B RPL8B RPS4B RPL36B
RPL17A RPL17B RPS26A
YGR149W YMR242W-A TKL2
RTA1 RPP2A RPS8B SRO9 AST2
RPL2B MNL1 RPL25 RPL38
SMC2-A RPL32 ASC1 RPS18A
RPS18B RPS0A RPL36A TEF2
TEF1 TSR1 RPS7A RPS7B
RPL33B RPL12B YGL188C-A

DIG1
STE12

SUT1 RAD54 YEL1 YKL209C
ACT1 YCL056C SRL3 PEP1
CLB1 SKT5 YKR041W HYM1
GYP8 STE2 SIM1 MSG5 FUS2
FUS3 FUS1 ERG24 DIG1 SVS1
FIG1 PRM1 BEM2 INP52 CST6
HAL1 YLR3 32W CHS1 ORC2
CAM1 tK(UUU)G2

TEC1 UBC4 SEC2
PCL2 ATG19 SCH9
YIL015W GFA1 SSF2
GPA1 RAX2 BUD14
KAR4 YML046W ADE1
MSB2 tR(UCU)M1
CUP9 RDI1 PRM6
YJU2 PRM3 PRM2
tA(AGC)K2 CBT1 PRM4

SST2

CWP1 AFR1
YKL096C-B

BDF1

BUD8
tK(UUU)P
VTI1
CLB2
CIK1

MSN2
MSN4

PST2 CUE4 PEX32 VOA1 SSA1
RPA14 CTT1 SSA4 BMH1 HSP78
IST2 DDR48 HSP30 UPF3 TDH3
TPK1 RRI2 POG1 MCH4 TVP23
CCP1 GPH1 GPI17 SKM1
tL(GAG)G RFC5 YOL114C STF2
HER1 CPR1 AMD1 HSP42 ASK1
ETP1 MDH1 GRE3 tG(GCC)P2
GND2 PCL7 tE(UUC)E1 LAP4
YML006C ISU2 YGR071C MRS4
SFK1 snR18 MMF1 HSP26 PDC6
GPT2 YGP1 YPR159C-A FMP48
ZWF1 LST8 MRH1 HSL1 SRC1
POP7 TSL1-A ROD1 TSL1 SIS1
YGR255C EFB1 GTO3-A

TEL2 IDI1 PNP1 RSP5
KIP2 HTS1 YCR075W-A
CLB4 PDS1 HHT1 RIM4
CDC39 SSO2 YLR257W
ECM33 ADD37 HOS3
CTS1 SCW11 DSE1
VAC17 RKM3 MSH3
YNL176C SLM4 BUD4
NST1 YBR139W
YDR115W ECM19-A
HHF1 YAP6 AMN1 ARP7
SPB1 YBR138C ERS1
JSN1 ICS2 CDS1 DIN7
tA(AGC)L DBF2 CHA1
MK22 ESP1 GLN1
SBE22 SPC24

FKH1
FKH2

GID8 CLA4
MNN5 SCW10
YGR189C
SWE1 GAT2
FKS3 GIC2

SPT21 SOK2 RBA50 AGE1
SNA2 HCM1 MPT5 LAC1 TOS2
TOS3 CAP1 PEX28 CLN1
AGE1-B GIN4 SKG6 LSM4
YOX1 YMR279C PET54 MCD1
GAS1 MGA1 PDR16 YGR109C
YDR524W-A BUD9 GIC1
YGR109W-B YGR109W-APCL1
SPO20

YPR159W
tA(AGC)P

RNR1
YHP1
PPN1

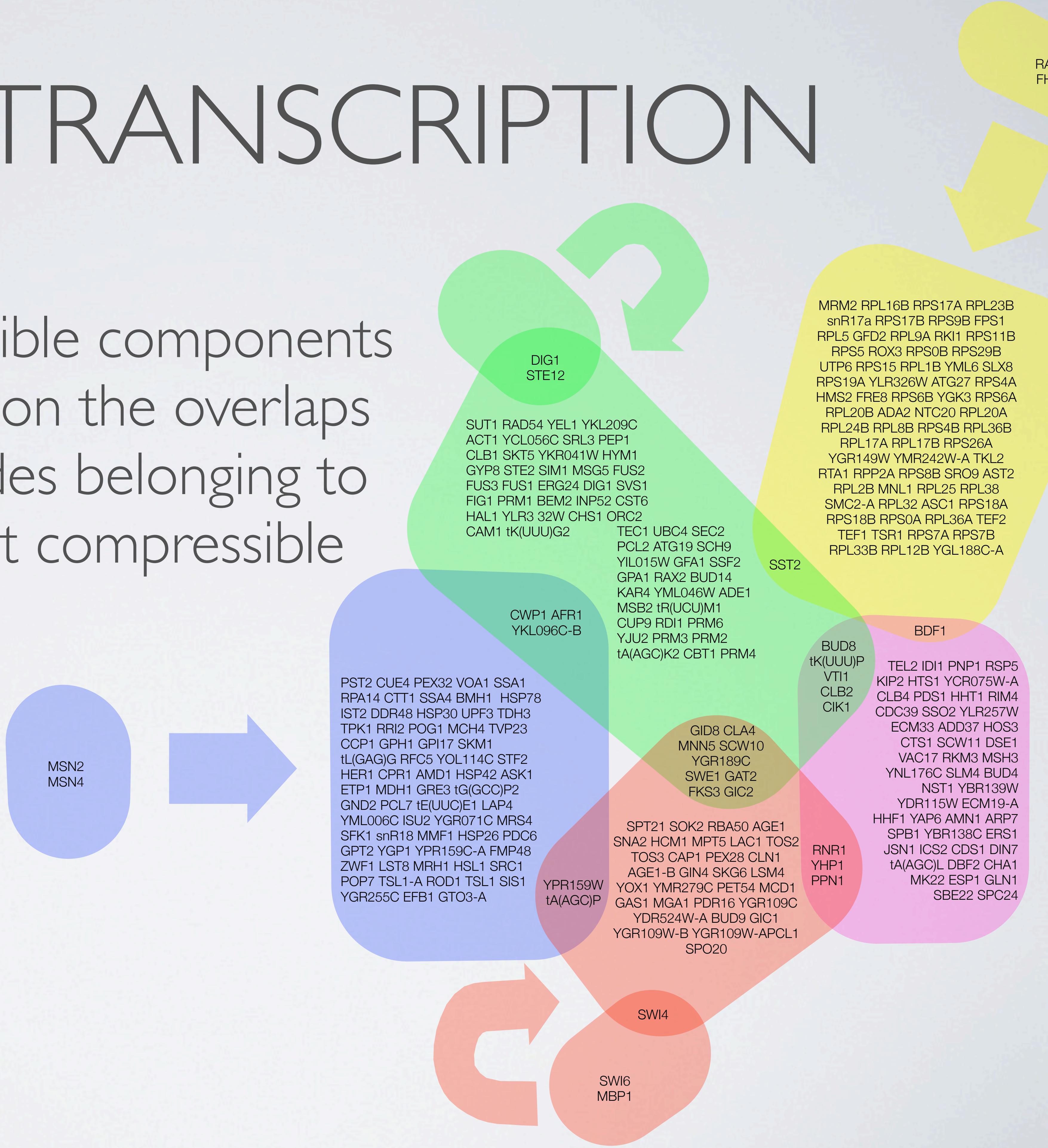
SWI4

SWI6
MBP1

S. E. Ahnert, Mol Biosyst 9, 2681 (2013)

YEAST TRANSCRIPTION

Note that compressible components have no restrictions on the overlaps between power nodes belonging to the same or different compressible components.



RA
FH

DIG1
STE12

SUT1 RAD54 YEL1 YKL209C
ACT1 YCL056C SRL3 PEP1
CLB1 SKT5 YKR041W HYM1
GYP8 STE2 SIM1 MSG5 FUS2
FUS3 FUS1 ERG24 DIG1 SVS1
FIG1 PRM1 BEM2 INP52 CST6
HAL1 YLR3 32W CHS1 ORC2
CAM1 tK(UUU)G2

TEC1 UBC4 SEC2
PCL2 ATG19 SCH9
YIL015W GFA1 SSF2
GPA1 RAX2 BUD14
KAR4 YML046W ADE1
MSB2 tR(UCU)M1
CUP9 RDI1 PRM6
YJU2 PRM3 PRM2
tA(AGC)K2 CBT1 PRM4

SST2

MRM2 RPL16B RPS17A RPL23B
snR17a RPS17B RPS9B FPS1
RPL5 GFD2 RPL9A RKI1 RPS11B
RPS5 ROX3 RPS0B RPS29B
UTP6 RPS15 RPL1B YML6 SLX8
RPS19A YLR326W ATG27 RPS4A
HMS2 FRE8 RPS6B YGK3 RPS6A
RPL20B ADA2 NTC20 RPL20A
RPL24B RPL8B RPS4B RPL36B
RPL17A RPL17B RPS26A
YGR149W YMR242W-A TKL2
RTA1 RPP2A RPS8B SRO9 AST2
RPL2B MNL1 RPL25 RPL38
SMC2-A RPL32 ASC1 RPS18A
RPS18B RPS0A RPL36A TEF2
TEF1 TSR1 RPS7A RPS7B
RPL33B RPL12B YGL188C-A

CWP1 AFR1
YKL096C-B

BDF1

BUD8
tK(UUU)P
VTI1
CLB2
CIK1

TEL2 IDI1 PNP1 RSP5
KIP2 HTS1 YCR075W-A
CLB4 PDS1 HHT1 RIM4
CDC39 SSO2 YLR257W
ECM33 ADD37 HOS3
CTS1 SCW11 DSE1
VAC17 RKM3 MSH3
YNL176C SLM4 BUD4
NST1 YBR139W
YDR115W ECM19-A
HHF1 YAP6 AMN1 ARP7
SPB1 YBR138C ERS1
JSN1 ICS2 CDS1 DIN7
tA(AGC)L DBF2 CHA1
MK22 ESP1 GLN1
SBE22 SPC24

PST2 CUE4 PEX32 VOA1 SSA1
RPA14 CTT1 SSA4 BMH1 HSP78
IST2 DDR48 HSP30 UPF3 TDH3
TPK1 RRI2 POG1 MCH4 TVP23
CCP1 GPH1 GPI17 SKM1
tL(GAG)G RFC5 YOL114C STF2
HER1 CPR1 AMD1 HSP42 ASK1
ETP1 MDH1 GRE3 tG(GCC)P2
GND2 PCL7 tE(UUC)E1 LAP4
YML006C ISU2 YGR071C MRS4
SFK1 snR18 MMF1 HSP26 PDC6
GPT2 YGP1 YPR159C-A FMP48
ZWF1 LST8 MRH1 HSL1 SRC1
POP7 TSL1-A ROD1 TSL1 SIS1
YGR255C EFB1 GTO3-A

MSN2
MSN4

GID8 CLA4
MNN5 SCW10
YGR189C
SWE1 GAT2
FKS3 GIC2

SPT21 SOK2 RBA50 AGE1
SNA2 HCM1 MPT5 LAC1 TOS2
TOS3 CAP1 PEX28 CLN1
AGE1-B GIN4 SKG6 LSM4
YOX1 YMR279C PET54 MCD1
GAS1 MGA1 PDR16 YGR109C
YDR524W-A BUD9 GIC1
YGR109W-B YGR109W-APCL1
SPO20

RNR1
YHP1
PPN1

YPR159W
tA(AGC)P

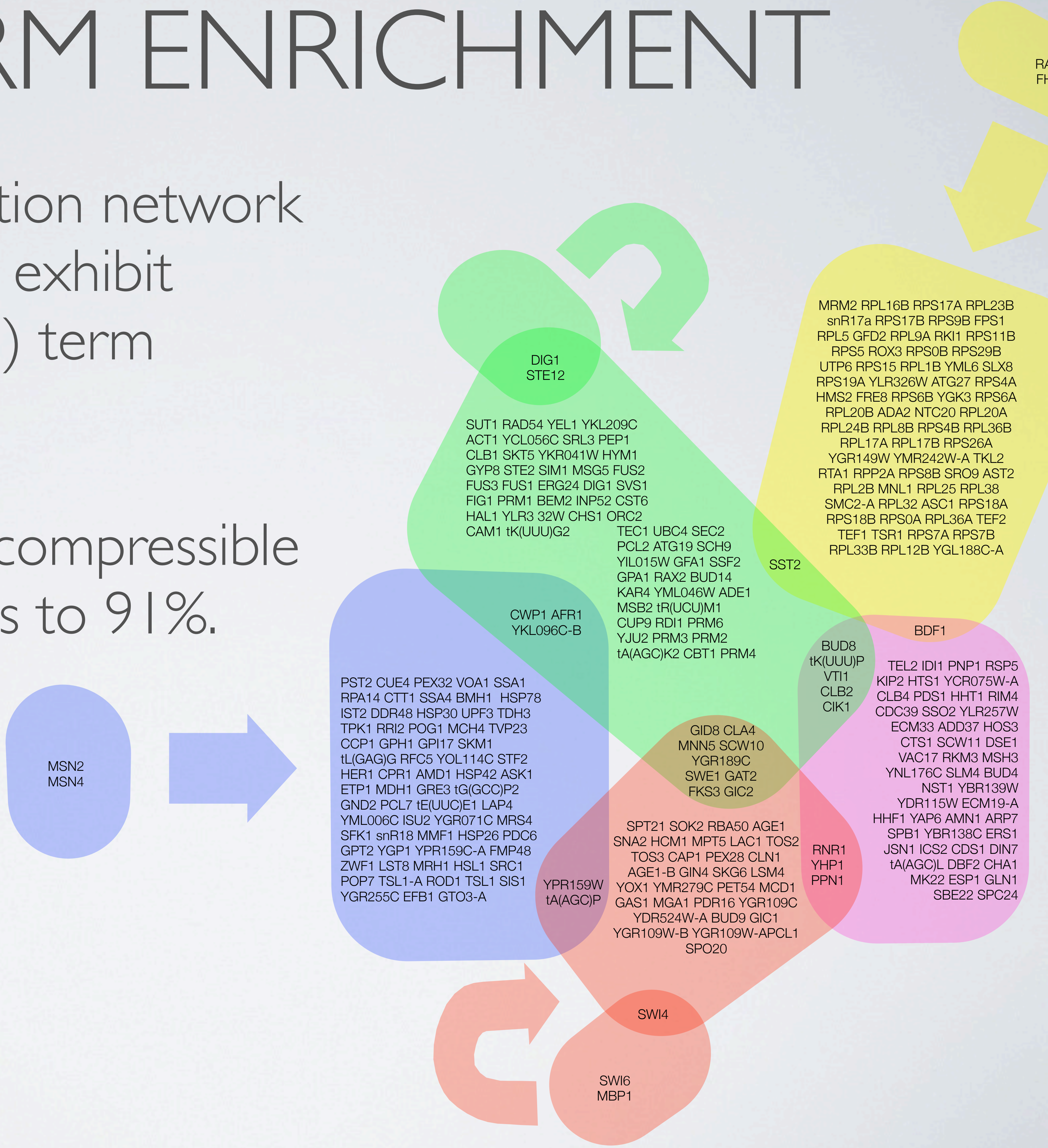
SWI4

SWI6
MBP1

GO TERM ENRICHMENT

In the yeast transcription network 67% of power nodes exhibit Gene Ontology (GO) term enrichment.

Among the top 100 compressible components, this rises to 91%.



DIG1
STE12

SUT1 RAD54 YEL1 YKL209C
ACT1 YCL056C SRL3 PEP1
CLB1 SKT5 YKR041W HYM1
GYP8 STE2 SIM1 MSG5 FUS2
FUS3 FUS1 ERG24 DIG1 SVS1
FIG1 PRM1 BEM2 INP52 CST6
HAL1 YLR3 32W CHS1 ORC2
CAM1 tK(UUU)G2

TEC1 UBC4 SEC2
PCL2 ATG19 SCH9
YIL015W GFA1 SSF2
GPA1 RAX2 BUD14
KAR4 YML046W ADE1
MSB2 tR(UCU)M1
CUP9 RDI1 PRM6
YJU2 PRM3 PRM2
tA(AGC)K2 CBT1 PRM4

SST2

RA
FH

MRM2 RPL16B RPS17A RPL23B
snR17a RPS17B RPS9B FPS1
RPL5 GFD2 RPL9A RKI1 RPS11B
RPS5 ROX3 RPS0B RPS29B
UTP6 RPS15 RPL1B YML6 SLX8
RPS19A YLR326W ATG27 RPS4A
HMS2 FRE8 RPS6B YGK3 RPS6A
RPL20B ADA2 NTC20 RPL20A
RPL24B RPL8B RPS4B RPL36B
RPL17A RPL17B RPS26A
YGR149W YMR242W-A TKL2
RTA1 RPP2A RPS8B SRO9 AST2
RPL2B MNL1 RPL25 RPL38
SMC2-A RPL32 ASC1 RPS18A
RPS18B RPS0A RPL36A TEF2
TEF1 TSR1 RPS7A RPS7B
RPL33B RPL12B YGL188C-A

CWP1 AFR1
YKL096C-B

MSN2
MSN4

PST2 CUE4 PEX32 VOA1 SSA1
RPA14 CTT1 SSA4 BMH1 HSP78
IST2 DDR48 HSP30 UPF3 TDH3
TPK1 RRI2 POG1 MCH4 TVP23
CCP1 GPH1 GPI17 SKM1
tL(GAG)G RFC5 YOL114C STF2
HER1 CPR1 AMD1 HSP42 ASK1
ETP1 MDH1 GRE3 tG(GCC)P2
GND2 PCL7 tE(UUC)E1 LAP4
YML006C ISU2 YGR071C MRS4
SFK1 snR18 MMF1 HSP26 PDC6
GPT2 YGP1 YPR159C-A FMP48
ZWF1 LST8 MRH1 HSL1 SRC1
POP7 TSL1-A ROD1 TSL1 SIS1
YGR255C EFB1 GTO3-A

BDF1

BUD8
tK(UUU)P
VTI1
CLB2
CIK1

TEL2 IDI1 PNP1 RSP5
KIP2 HTS1 YCR075W-A
CLB4 PDS1 HHT1 RIM4
CDC39 SSO2 YLR257W
ECM33 ADD37 HOS3
CTS1 SCW11 DSE1
VAC17 RKM3 MSH3
YNL176C SLM4 BUD4
NST1 YBR139W
YDR115W ECM19-A
HHF1 YAP6 AMN1 ARP7
SPB1 YBR138C ERS1
JSN1 ICS2 CDS1 DIN7
tA(AGC)L DBF2 CHA1
MK22 ESP1 GLN1
SBE22 SPC24

GID8 CLA4
MNN5 SCW10
YGR189C
SWE1 GAT2
FKS3 GIC2

YPR159W
tA(AGC)P

SPT21 SOK2 RBA50 AGE1
SNA2 HCM1 MPT5 LAC1 TOS2
TOS3 CAP1 PEX28 CLN1
AGE1-B GIN4 SKG6 LSM4
YOX1 YMR279C PET54 MCD1
GAS1 MGA1 PDR16 YGR109C
YDR524W-A BUD9 GIC1
YGR109W-B YGR109W-APCL1
SPO20

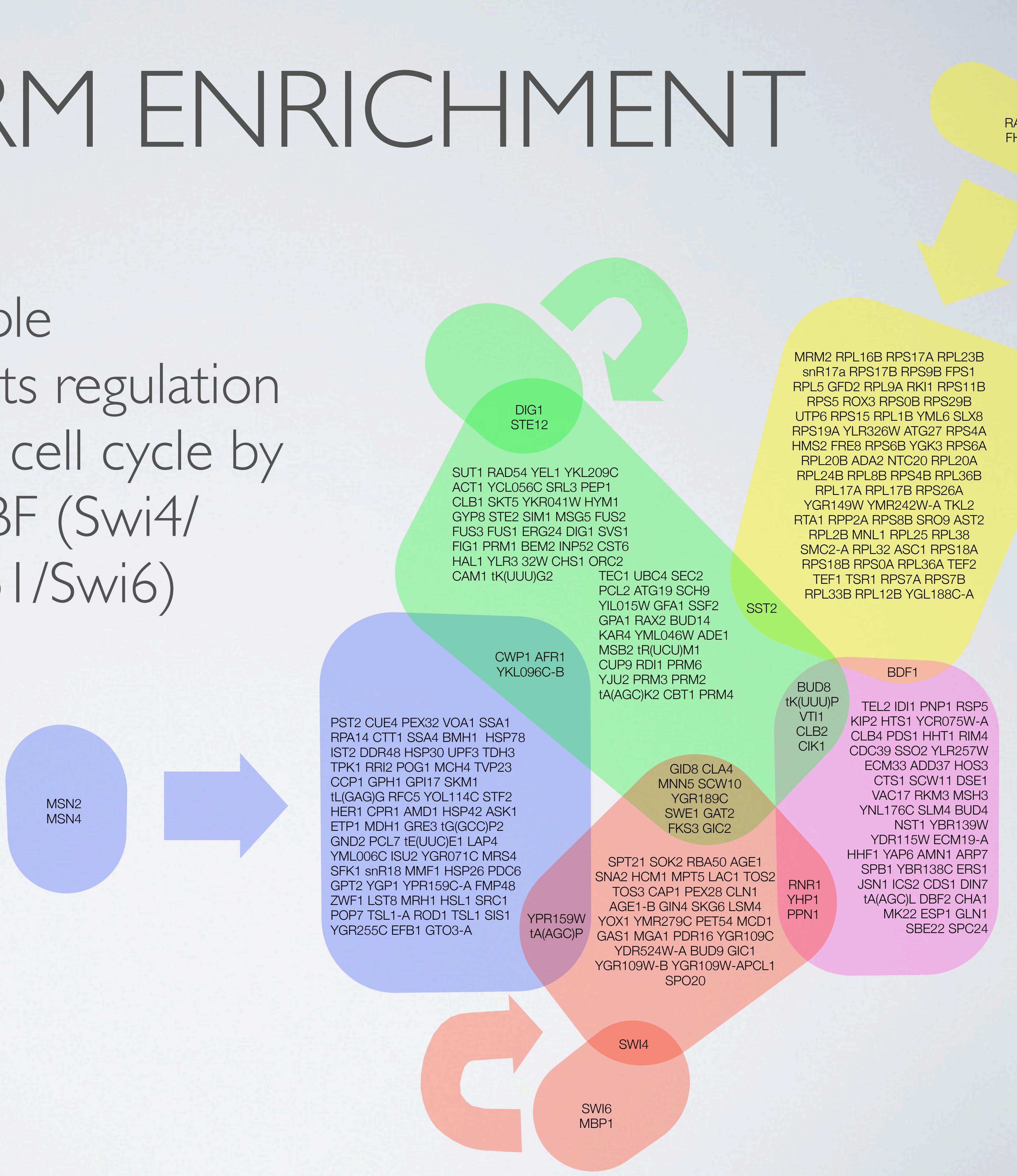
RNR1
YHP1
PPN1

SWI4

SWI6
MBP1

GO TERM ENRICHMENT

The most compressible component represents regulation of the G1/S phase of cell cycle by two heterodimers: SBF (Swi4/Swi6) and MBF (Mbp1/Swi6)



RA
FH

MRM2 RPL16B RPS17A RPL23B
snR17a RPS17B RPS9B FPS1
RPL5 GFD2 RPL9A RKI1 RPS11B
RPS5 ROX3 RPS0B RPS29B
UTP6 RPS15 RPL1B YML6 SLX8
RPS19A YLR326W ATG27 RPS4A
HMS2 FRE8 RPS6B YGK3 RPS6A
RPL20B ADA2 NTC20 RPL20A
RPL24B RPL8B RPS4B RPL36B
RPL17A RPL17B RPS26A
YGR149W YMR242W-A TKL2
RTA1 RPP2A RPS8B SRO9 AST2
RPL2B MNL1 RPL25 RPL38
SMC2-A RPL32 ASC1 RPS18A
RPS18B RPS0A RPL36A TEF2
TEF1 TSR1 RPS7A RPS7B
RPL33B RPL12B YGL188C-A

DIG1
STE12

SUT1 RAD54 YEL1 YKL209C
ACT1 YCL056C SRL3 PEP1
CLB1 SKT5 YKR041W HYM1
GYP8 STE2 SIM1 MSG5 FUS2
FUS3 FUS1 ERG24 DIG1 SVS1
FIG1 PRM1 BEM2 INP52 CST6
HAL1 YLR3 32W CHS1 ORC2
CAM1 tK(UUU)G2

TEC1 UBC4 SEC2
PCL2 ATG19 SCH9
YIL015W GFA1 SSF2
GPA1 RAX2 BUD14
KAR4 YML046W ADE1
MSB2 tR(UCU)M1
CUP9 RDI1 PRM6
YJU2 PRM3 PRM2
tA(AGC)K2 CBT1 PRM4

SST2

CWP1 AFR1
YKL096C-B

BDF1

BUD8
tK(UUU)P
VTI1
CLB2
CIK1

TEL2 IDI1 PNP1 RSP5
KIP2 HTS1 YCR075W-A
CLB4 PDS1 HHT1 RIM4
CDC39 SSO2 YLR257W
ECM33 ADD37 HOS3
CTS1 SCW11 DSE1
VAC17 RKM3 MSH3
YNL176C SLM4 BUD4
NST1 YBR139W
YDR115W ECM19-A
HHF1 YAP6 AMN1 ARP7
SPB1 YBR138C ERS1
JSN1 ICS2 CDS1 DIN7
tA(AGC)L DBF2 CHA1
MK22 ESP1 GLN1
SBE22 SPC24

MSN2
MSN4

PST2 CUE4 PEX32 VOA1 SSA1
RPA14 CTT1 SSA4 BMH1 HSP78
IST2 DDR48 HSP30 UPF3 TDH3
TPK1 RRI2 POG1 MCH4 TVP23
CCP1 GPH1 GPI17 SKM1
tL(GAG)G RFC5 YOL114C STF2
HER1 CPR1 AMD1 HSP42 ASK1
ETP1 MDH1 GRE3 tG(GCC)P2
GND2 PCL7 tE(UUC)E1 LAP4
YML006C ISU2 YGR071C MRS4
SFK1 snR18 MMF1 HSP26 PDC6
GPT2 YGP1 YPR159C-A FMP48
ZWF1 LST8 MRH1 HSL1 SRC1
POP7 TSL1-A ROD1 TSL1 SIS1
YGR255C EFB1 GTO3-A

GID8 CLA4
MNN5 SCW10
YGR189C
SWE1 GAT2
FKS3 GIC2

YPR159W
tA(AGC)P

SPT21 SOK2 RBA50 AGE1
SNA2 HCM1 MPT5 LAC1 TOS2
TOS3 CAP1 PEX28 CLN1
AGE1-B GIN4 SKG6 LSM4
YOX1 YMR279C PET54 MCD1
GAS1 MGA1 PDR16 YGR109C
YDR524W-A BUD9 GIC1
YGR109W-B YGR109W-APCL1
SPO20

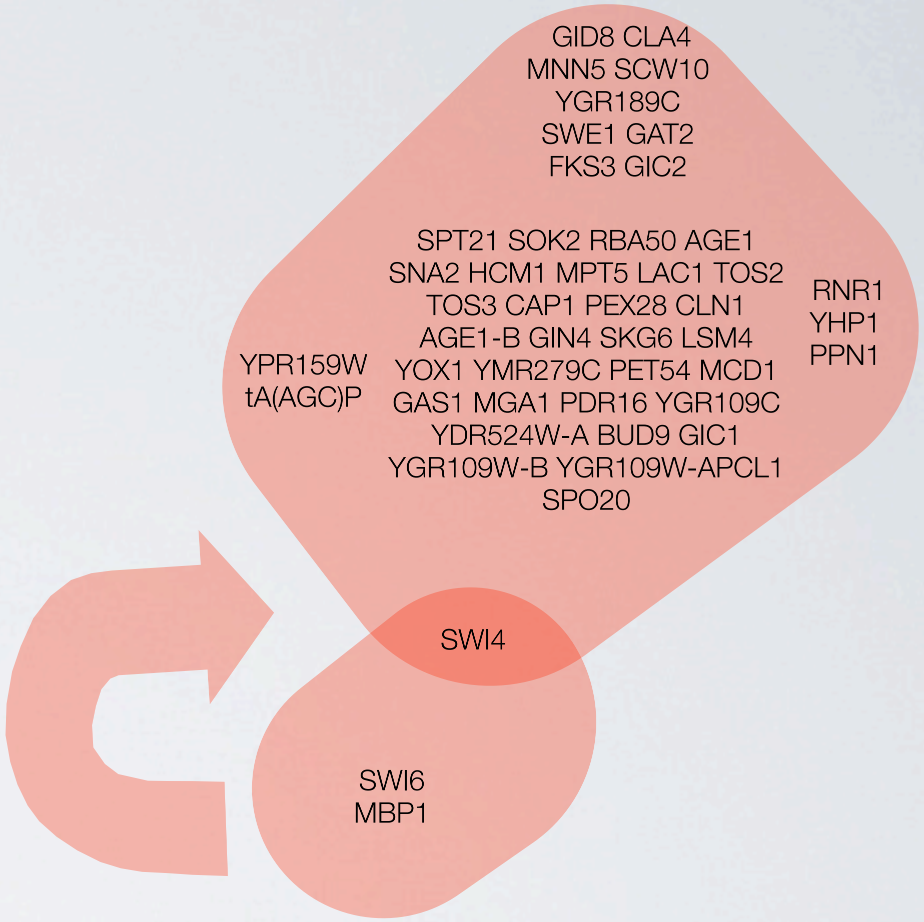
RNR1
YHP1
PPN1

SWI4

SWI6
MBP1

GO TERM ENRICHMENT

The most compressible component represents regulation of the G1/S phase of cell cycle by two heterodimers: SBF (Swi4/Swi6) and MBF (Mbp1/Swi6)



Set A	**GO terms**	**Set B**	**GO terms**
MBP1 SWI4 SWI6	regulation of transcription involved in G1/S phase of mitotic cell cycle (2.05E-04)	47 genes	cell cycle process (1.69E-04) cell division (1.87E-04) incipient cellular bud site (3.03E-04)

GO TERM ENRICHMENT

The second most compressible component represents regulation of two developmental processes, mating and filamentation, regulated by the **Ste12/Dig1**/Dig2 and Tec1/**Ste12/Dig1** complexes.

GO TERM ENRICHMENT

The second most compressible component represents regulation of two developmental processes, mating and filamentation, regulated by the **Ste12/Dig1**/Dig2 and Tec1/**Ste12/Dig1** complexes.

DIG1
STE12

SUT1 RAD54 YEL1 YKL209C
ACT1 YCL056C SRL3 PEP1
CLB1 SKT5 YKR041W HYM1
GYP8 STE2 SIM1 MSG5 FUS2
FUS3 FUS1 ERG24 DIG1 SVS1
FIG1 PRM1 BEM2 INP52 CST6
HAL1 YLR3 32W CHS1 ORC2
CAM1 tK(UUU)G2

TEC1 UBC4 SEC2
PCL2 ATG19 SCH9
YIL015W GFA1 SSF2
GPA1 RAX2 BUD14
KAR4 YML046W ADE1
MSB2 tR(UCU)M1
CUP9 RDI1 PRM6
YJU2 PRM3 PRM2
tA(AGC)K2 CBT1 PRM4

SST2

CWP1 AFR1
YKL096C-B

BUD8
tK(UUU)P
VTI1
CLB2
CIK1

GID8 CLA4
MNN5 SCW10
YGR189C
SWE1 GAT2
FKS3 GIC2

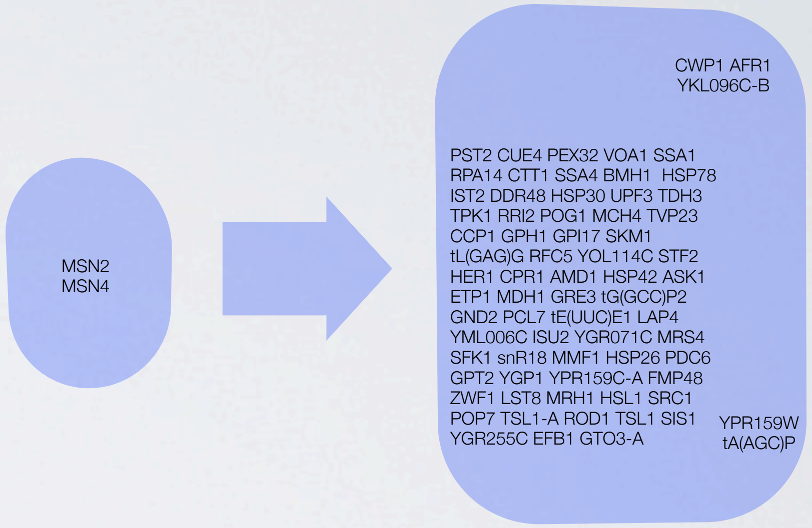
Set A	GO terms	Set B	GO terms
STE12 DIG1	regulation of transcription by pheromones (4.93E-04) cell growth (7.54E-04) regulation of transcription from RNAp II promoter by pheromones (9.45E-04)	78 genes	multi-organism process (6.35E-14) site of polarized growth (7.59E-11) reproduction (2.65E-10)

GO TERM ENRICHMENT

The third most compressible component represents regulation of a broad stress response by **Msn2** and **Msn4**.

GO TERM ENRICHMENT

The third most compressible component represents regulation of a broad stress response by **Msn2** and **Msn4**.



MSN2
MSN4

CWP1 AFR1
YKL096C-B

PST2 CUE4 PEX32 VOA1 SSA1
RPA14 CTT1 SSA4 BMH1 HSP78
IST2 DDR48 HSP30 UPF3 TDH3
TPK1 RRI2 POG1 MCH4 TVP23
CCP1 GPH1 GPI17 SKM1
tL(GAG)G RFC5 YOL114C STF2
HER1 CPR1 AMD1 HSP42 ASK1
ETP1 MDH1 GRE3 tG(GCC)P2
GND2 PCL7 tE(UUC)E1 LAP4
YML006C ISU2 YGR071C MRS4
SFK1 snR18 MMF1 HSP26 PDC6
GPT2 YGP1 YPR159C-A FMP48
ZWF1 LST8 MRH1 HSL1 SRC1
POP7 TSL1-A ROD1 TSL1 SIS1
YGR255C EFB1 GTO3-A

YPR159W
tA(AGC)P

Set A	GO terms	Set B	GO terms
MSN2 MSN4	response to hydrostatic pressure (6.96E-05) response to nitrosative stress (6.96E-05) response to water (8.60E-05)	72 genes	unfolded protein binding (2.06E-03) chaperonin-containing T-complex (2.28E-03) hydrogen peroxide catabolic process (5.38E-03)

GO TERM ENRICHMENT

The fact that the second component describes **highly specific** developmental programmes...

GO TERM ENRICHMENT

The fact that the second component describes **highly specific** developmental programmes...

STE12 DIG1	regulation of transcription by pheromones (4.93E-04) cell growth (7.54E-04) regulation of transcription from RNAp II promoter by pheromones (9.45E-04)	78 genes	multi-organism process (6.35E-14) site of polarized growth (7.59E-11) reproduction (2.65E-10)

GO TERM ENRICHMENT

The fact that the second component describes **highly specific** developmental programmes...

STE12 DIG1	regulation of transcription by pheromones (4.93E-04) cell growth (7.54E-04) regulation of transcription from RNAp II promoter by pheromones (9.45E-04)	78 genes	multi-organism process (6.35E-14) site of polarized growth (7.59E-11) reproduction (2.65E-10)

...and the third component describes a **broad** stress response...

MSN2 MSN4	response to hydrostatic pressure (6.96E-05) response to nitrosative stress (6.96E-05) response to water (8.60E-05)	72 genes	unfolded protein binding (2.06E-03) chaperonin-containing T-complex (2.28E-03) hydrogen peroxide catabolic process (5.38E-03)

GO TERM ENRICHMENT

The fact that the second component describes **highly specific** developmental programmes...

STE12 DIG1	regulation of transcription by pheromones (4.93E-04) cell growth (7.54E-04) regulation of transcription from RNAp II promoter by pheromones (9.45E-04)	78 genes	multi-organism process (6.35E-14) site of polarized growth (7.59E-11) reproduction (2.65E-10) **(6.35E-14)** **(7.59E-11)** **(2.65E-10)**

...and the third component describes a **broad** stress response...

MSN2 MSN4	response to hydrostatic pressure (6.96E-05) response to nitrosative stress (6.96E-05) response to water (8.60E-05)	72 genes	unfolded protein binding (2.06E-03) chaperonin-containing T-complex (2.28E-03) hydrogen peroxide catabolic process (5.38E-03) **(2.06E-03)** **(2.28E-03)** **(5.38E-03)**

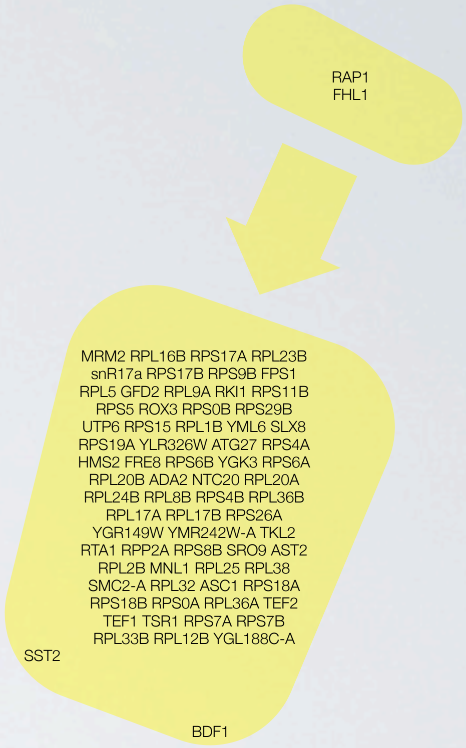
...can also be seen from the p-values of the GO term enrichment.

GO TERM ENRICHMENT

The fourth most compressible component represents highly specific regulation of the ribosome.

GO TERM ENRICHMENT

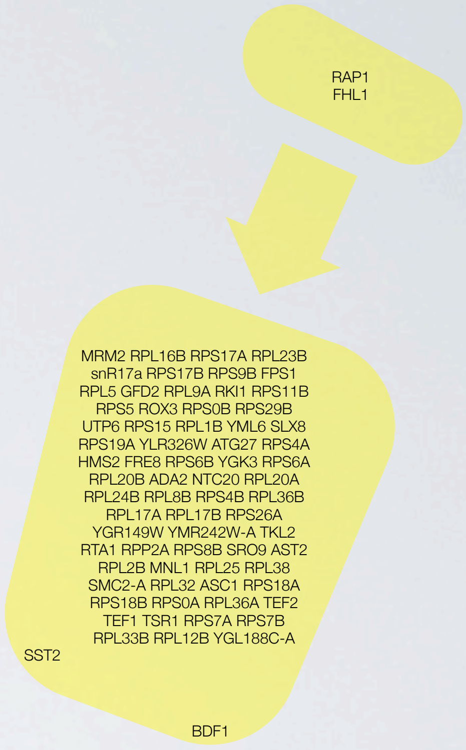
The fourth most compressible component represents highly specific regulation of the ribosome.



Set A	**GO terms**	**Set B**	**GO terms**
RAP1 FHL1	CURI complex (6.08E-03) nucleic acid binding transcription factor activity (7.90E-03)	71 genes	ribonucleoprotein complex (1.40E-37) ribosome (5.46E-37) ribosomal subunit (4.93E-36)

GO TERM ENRICHMENT

The fourth most compressible component represents highly specific regulation of the ribosome.



RAP1
FHL1

MRM2 RPL16B RPS17A RPL23B
snR17a RPS17B RPS9B FPS1
RPL5 GFD2 RPL9A RKI1 RPS11B
RPS5 ROX3 RPS0B RPS29B
UTP6 RPS15 RPL1B YML6 SLX8
RPS19A YLR326W ATG27 RPS4A
HMS2 FRE8 RPS6B YGK3 RPS6A
RPL20B ADA2 NTC20 RPL20A
RPL24B RPL8B RPS4B RPL36B
RPL17A RPL17B RPS26A
YGR149W YMR242W-A TKL2
RTA1 RPP2A RPS8B SRO9 AST2
RPL2B MNL1 RPL25 RPL38
SMC2-A RPL32 ASC1 RPS18A
RPS18B RPS0A RPL36A TEF2
TEF1 TSR1 RPS7A RPS7B
RPL33B RPL12B YGL188C-A

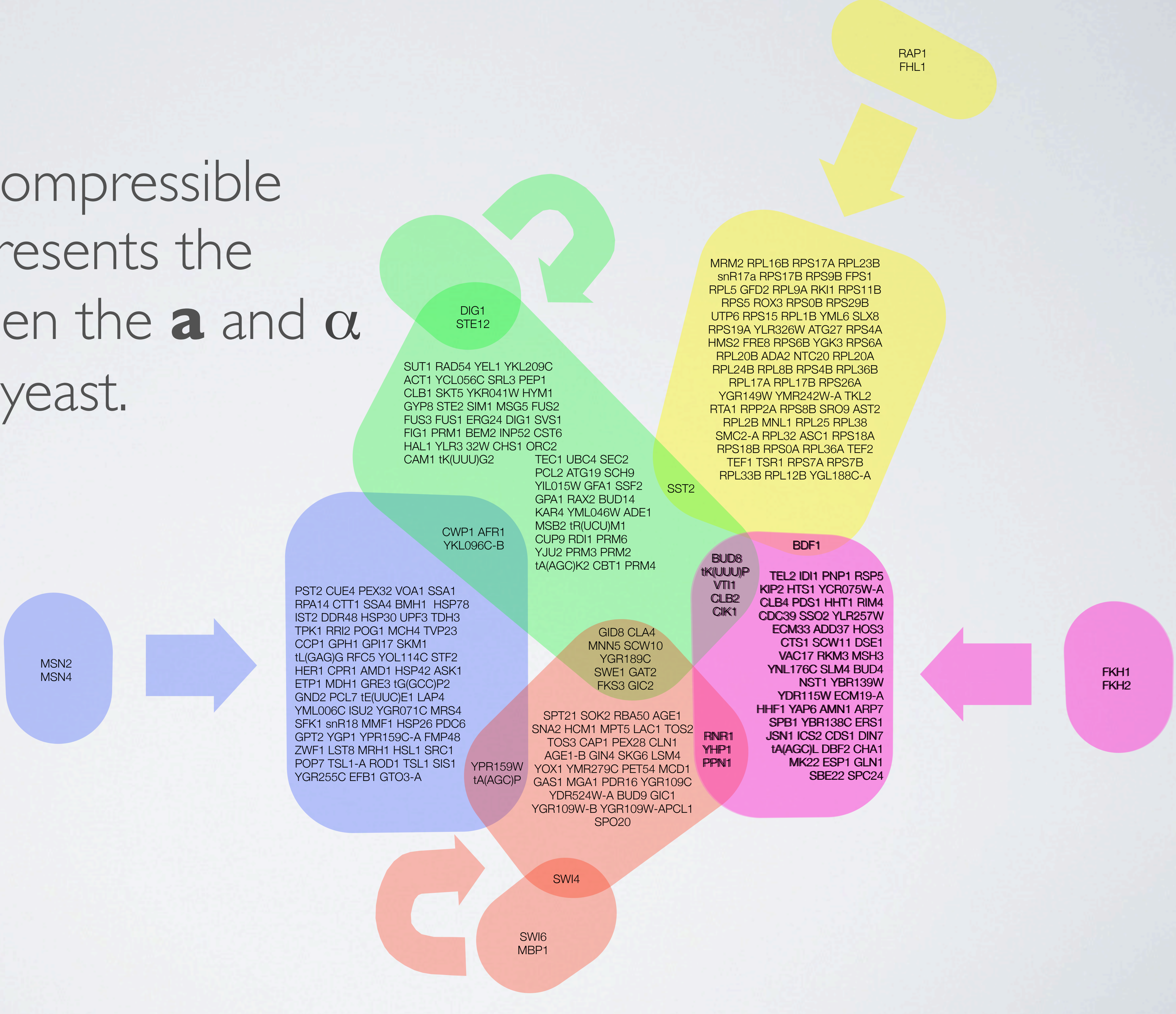
SST2

BDF1

Set A	GO terms	Set B	GO terms
RAP1 FHL1	CURI complex (6.08E-03) nucleic acid binding transcription factor activity (7.90E-03)	71 genes	ribonucleoprotein complex (1.40E-37) **(1.40E-37)** ribosome (5.46E-37) **(5.46E-37)** ribosomal subunit (4.93E-36) **(4.93E-36)**

GO TERM ENRICHMENT

The fifth most compressible component represents the switching between the **a** and α mating types of yeast.



RAP1
FHL1

MRM2 RPL16B RPS17A RPL23B
snR17a RPS17B RPS9B FPS1
RPL5 GFD2 RPL9A RKI1 RPS11B
RPS5 ROX3 RPS0B RPS29B
UTP6 RPS15 RPL1B YML6 SLX8
RPS19A YLR326W ATG27 RPS4A
HMS2 FRE8 RPS6B YGK3 RPS6A
RPL20B ADA2 NTC20 RPL20A
RPL24B RPL8B RPS4B RPL36B
RPL17A RPL17B RPS26A
YGR149W YMR242W-A TKL2
RTA1 RPP2A RPS8B SRO9 AST2
RPL2B MNL1 RPL25 RPL38
SMC2-A RPL32 ASC1 RPS18A
RPS18B RPS0A RPL36A TEF2
TEF1 TSR1 RPS7A RPS7B
RPL33B RPL12B YGL188C-A

DIG1
STE12

SUT1 RAD54 YEL1 YKL209C
ACT1 YCL056C SRL3 PEP1
CLB1 SKT5 YKR041W HYM1
GYP8 STE2 SIM1 MSG5 FUS2
FUS3 FUS1 ERG24 DIG1 SVS1
FIG1 PRM1 BEM2 INP52 CST6
HAL1 YLR3 32W CHS1 ORC2
CAM1 tK(UUU)G2

TEC1 UBC4 SEC2
PCL2 ATG19 SCH9
YIL015W GFA1 SSF2
GPA1 RAX2 BUD14
KAR4 YML046W ADE1
MSB2 tR(UCU)M1
CUP9 RDI1 PRM6
YJU2 PRM3 PRM2
tA(AGC)K2 CBT1 PRM4

SST2

CWP1 AFR1
YKL096C-B

MSN2
MSN4

PST2 CUE4 PEX32 VOA1 SSA1
RPA14 CTT1 SSA4 BMH1 HSP78
IST2 DDR48 HSP30 UPF3 TDH3
TPK1 RRI2 POG1 MCH4 TVP23
CCP1 GPH1 GPI17 SKM1
tL(GAG)G RFC5 YOL114C STF2
HER1 CPR1 AMD1 HSP42 ASK1
ETP1 MDH1 GRE3 tG(GCC)P2
GND2 PCL7 tE(UUC)E1 LAP4
YML006C ISU2 YGR071C MRS4
SFK1 snR18 MMF1 HSP26 PDC6
GPT2 YGP1 YPR159C-A FMP48
ZWF1 LST8 MRH1 HSL1 SRC1
POP7 TSL1-A ROD1 TSL1 SIS1
YGR255C EFB1 GTO3-A

BUD8
tK(UUU)P
VTI1
CLB2
CIK1

BDF1

TEL2 IDI1 PNP1 RSP5
KIP2 HTS1 YCR075W-A
CLB4 PDS1 HHT1 RIM4
CDC39 SSO2 YLR257W
ECM33 ADD37 HOS3
CTS1 SCW11 DSE1
VAC17 RKM3 MSH3
YNL176C SLM4 BUD4
NST1 YBR139W
YDR115W ECM19-A
HHF1 YAP6 AMN1 ARP7
SPB1 YBR138C ERS1
JSN1 ICS2 CDS1 DIN7
tA(AGC)L DBF2 CHA1
MK22 ESP1 GLN1
SBE22 SPC24

FKH1
FKH2

GID8 CLA4
MNN5 SCW10
YGR189C
SWE1 GAT2
FKS3 GIC2

YPR159W
tA(AGC)P

SPT21 SOK2 RBA50 AGE1
SNA2 HCM1 MPT5 LAC1 TOS2
TOS3 CAP1 PEX28 CLN1
AGE1-B GIN4 SKG6 LSM4
YOX1 YMR279C PET54 MCD1
GAS1 MGA1 PDR16 YGR109C
YDR524W-A BUD9 GIC1
YGR109W-B YGR109W-APCL1
SPO20

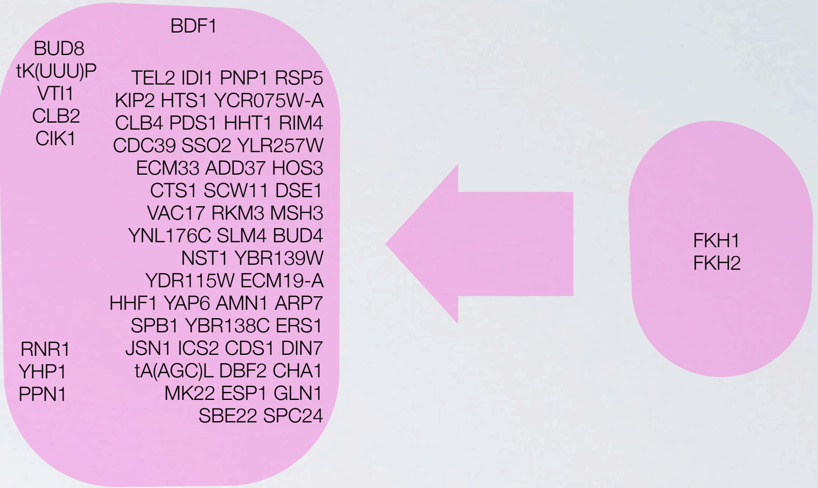
RNR1
YHP1
PPN1

SWI4

SWI6
MBP1

GO TERM ENRICHMENT

The fifth most compressible component represents the switching between the **a** and α mating types of yeast.



BDF1

BUD8
tK(UUU)P
VTI1
CLB2
CIK1

TEL2 IDI1 PNP1 RSP5
KIP2 HTS1 YCR075W-A
CLB4 PDS1 HHT1 RIM4
CDC39 SSO2 YLR257W
ECM33 ADD37 HOS3
CTS1 SCW11 DSE1
VAC17 RKM3 MSH3
YNL176C SLM4 BUD4
NST1 YBR139W
YDR115W ECM19-A
HHF1 YAP6 AMN1 ARP7
SPB1 YBR138C ERS1
JSN1 ICS2 CDS1 DIN7
tA(AGC)L DBF2 CHA1
MK22 ESP1 GLN1
SBE22 SPC24

RNR1
YHP1
PPN1

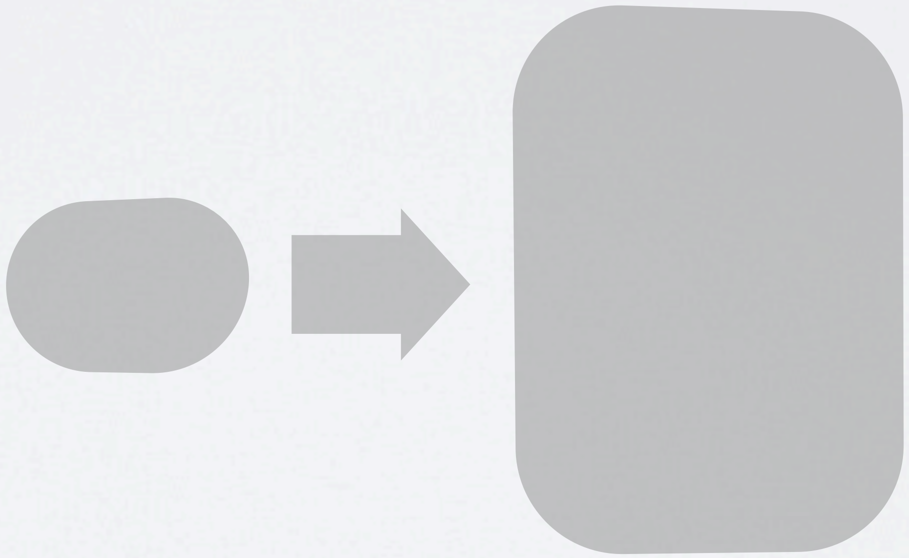
FKH1
FKH2

Set A	GO terms	Set B	GO terms
FKH1 FKH2	regulation of transcription involved in G2/M-phase of mitotic cell cycle (1.03E-04) regulation of transcription elongation from RNA polymerase II promoter (2.80E-04) regulation of transcription elongation, DNA-dependent (2.93E-04)	58 genes	cell cycle (1.64E-05) cell cycle process (9.68E-05) establishment of organelle localization (4.81E-04)

GO TERM ENRICHMENT

GO TERM ENRICHMENT

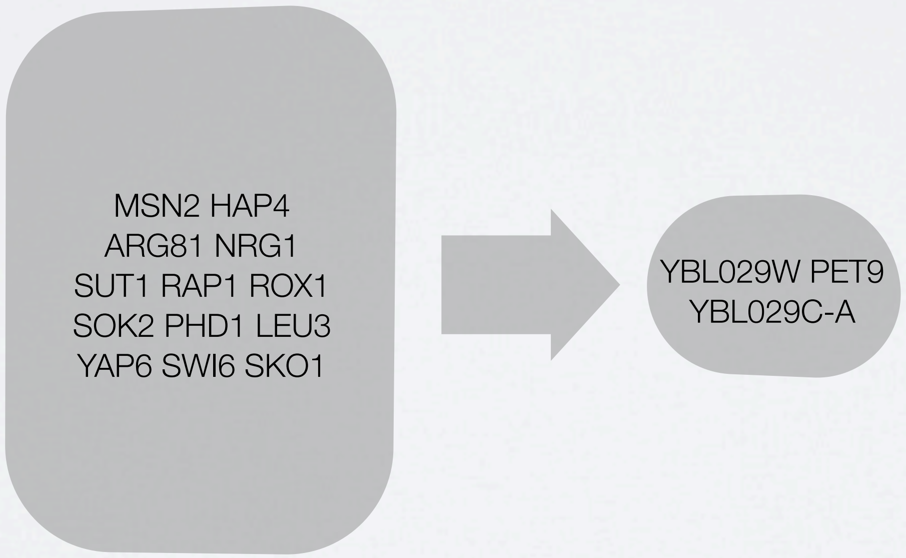
Most compressible components in transcription networks have few genes in Set A and many genes in Set B.



GO TERM ENRICHMENT

Most compressible components in transcription networks have few genes in Set A and many genes in Set B.

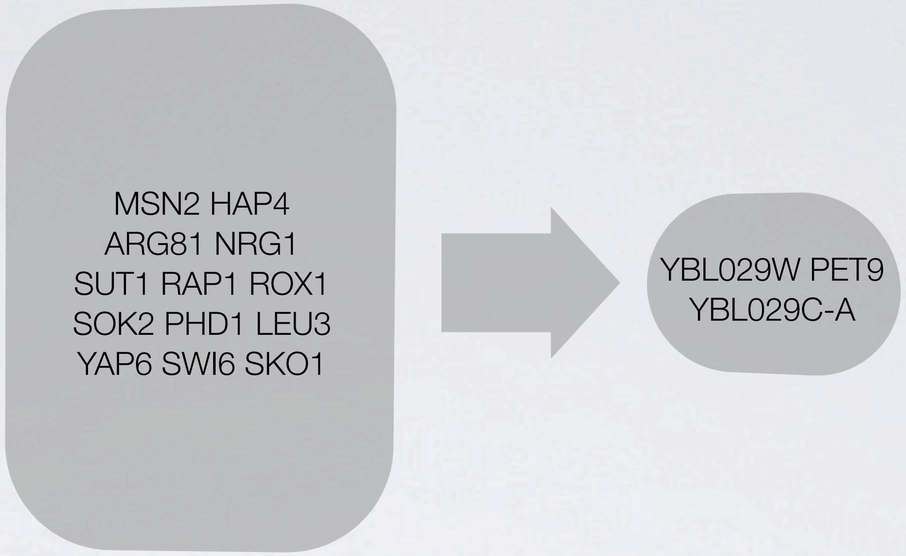
A notable exception is the 18th compressible component.



MSN2 HAP4
ARG81 NRG1
SUT1 RAP1 ROX1
SOK2 PHD1 LEU3
YAP6 SWI6 SKO1

YBL029W PET9
YBL029C-A

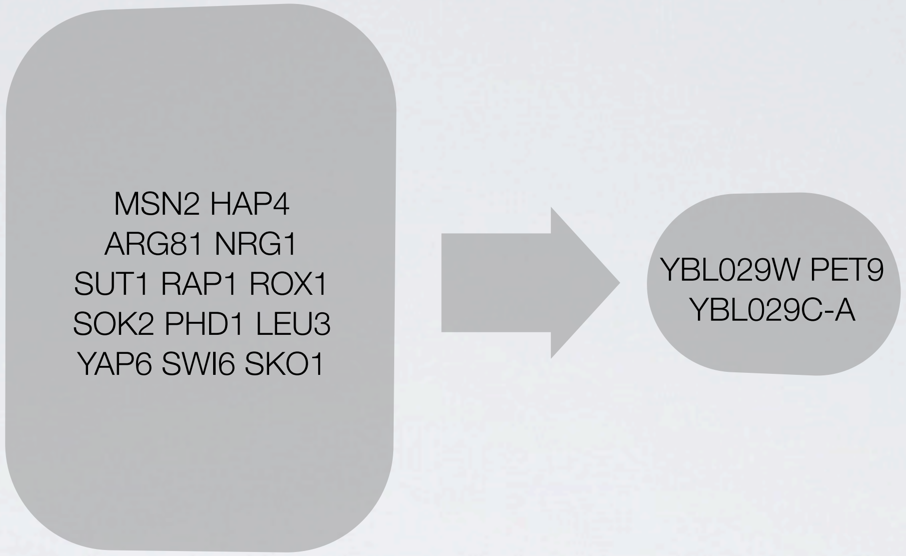
GO TERM ENRICHMENT



MSN2 HAP4
ARG81 NRG1
SUT1 RAP1 ROX1
SOK2 PHD1 LEU3
YAP6 SWI6 SKO1

YBL029W PET9
YBL029C-A

GO TERM ENRICHMENT



MSN2 HAP4
ARG81 NRG1
SUT1 RAP1 ROX1
SOK2 PHD1 LEU3
YAP6 SWI6 SKO1

YBL029W PET9
YBL029C-A

Set A	**GO terms**	**Set B**	**GO terms**
MSN2 HAP4 ARG81 NRG1 SUT1 RAP1 ROX1 SOK2 PHD1 LEU3 YAP6 SWI6 SKO1	nucleic acid binding transcription factor activity (1.6E-12) protein binding transcription factor activity (1.2E-7)	PET9 YBL029W YBL029C-A	cell death (8.9E-03) cell death (8.9E-03)

This compressible component represents processes that lead to cell death, forming a bottleneck of regulatory information as no downstream processing is required - quite literally a regulatory dead end.

GO TERM ENRICHMENT

Set A	**GO terms**	**Set B**	**GO terms**
MSN2 HAP4 ARG81 NRG1 SUT1 RAP1 ROX1 SOK2 PHD1 LEU3 YAP6 SWI6 SKO1	nucleic acid binding transcription factor activity (1.6E-12) protein binding transcription factor activity (1.2E-7)	PET9 YBL029W YBL029C-A	cell death (8.9E-03)

GO TERM ENRICHMENT

Set A	**GO terms**	**Set B**	**GO terms**
MSN2 HAP4 ARG81 NRG1 SUT1 RAP1 ROX1 SOK2 PHD1 LEU3 YAP6 SWI6 SKO1	nucleic acid binding transcription factor activity (1.6E-12) protein binding transcription factor activity (1.2E-7)	PET9 YBL029W YBL029C-A	cell death (8.9E-03)

PET9 is well-known to be involved in cell death (apoptosis).

GO TERM ENRICHMENT

Set A	**GO terms**	**Set B**	**GO terms**
MSN2 HAP4 ARG81 NRG1 SUT1 RAP1 ROX1 SOK2 PHD1 LEU3 YAP6 SWI6 SKO1	nucleic acid binding transcription factor activity (1.6E-12) protein binding transcription factor activity (1.2E-7)	PET9 YBL029W YBL029C-A	cell death (8.9E-03)

PET9 is well-known to be involved in cell death (apoptosis).

YBL029W has been speculatively linked to apoptosis as one of several genes that are over-expressed in apoptotic temperature-sensitive yeast cells.

GO TERM ENRICHMENT

Set A	**GO terms**	**Set B**	**GO terms**
MSN2 HAP4 ARG81 NRG1 SUT1 RAP1 ROX1 SOK2 PHD1 LEU3 YAP6 SWI6 SKO1	nucleic acid binding transcription factor activity (1.6E-12) protein binding transcription factor activity (1.2E-7)	PET9 YBL029W YBL029C-A	cell death (8.9E-03)

PET9 is well-known to be involved in cell death (apoptosis).

YBL029W has been speculatively linked to apoptosis as one of several genes that are over-expressed in apoptotic temperature-sensitive yeast cells.

YBL029C-A has never been linked to apoptosis. Its membership in this set suggests that it might be.

E. COLI TRANSCRIPTION NETWORK

Another widely-studied transcription network is that of the bacterium *E. coli*. The one we use here has 889 nodes and 1465 edges.

In *E. coli* many of the genes are organised into overlapping DNA regions, known as operons. These groups of genes correspond to functional modules.

E. COLI TRANSCRIPTION NETWORK

Another widely-studied transcription network is that of the bacterium *E. coli*. The one we use here has 889 nodes and 1465 edges.

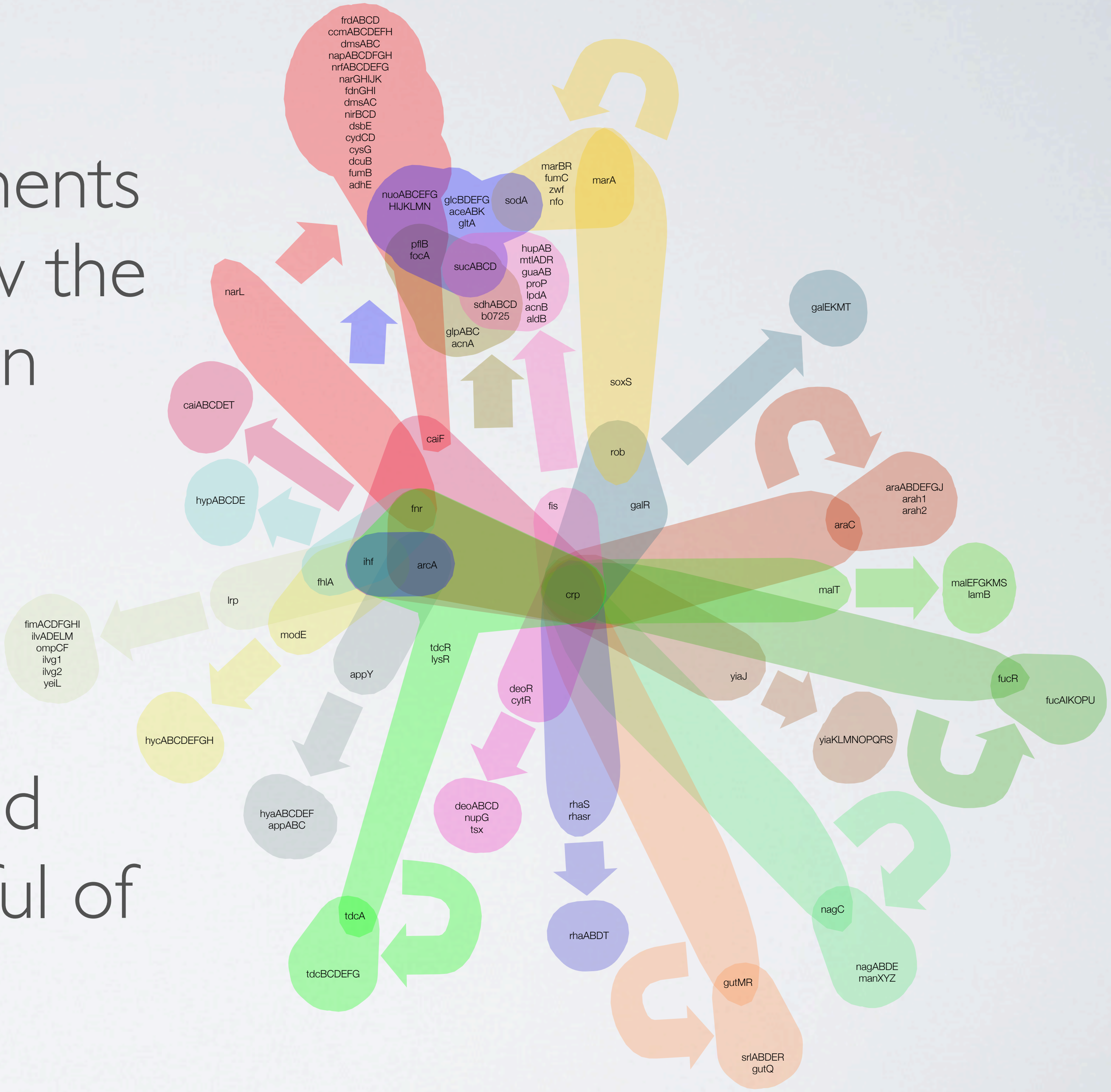
In *E. coli* many of the genes are organised into overlapping DNA regions, known as operons. These groups of genes correspond to functional modules.

M. M. Babu, S. A. Teichmann, Nucl. Acids Res. (2003) 31 (4): 1234-1244.

E. COLI TRANSCRIPTION NETWORK

The compressible components of the *E. coli* network show the organisation of the genes in terms of operons.

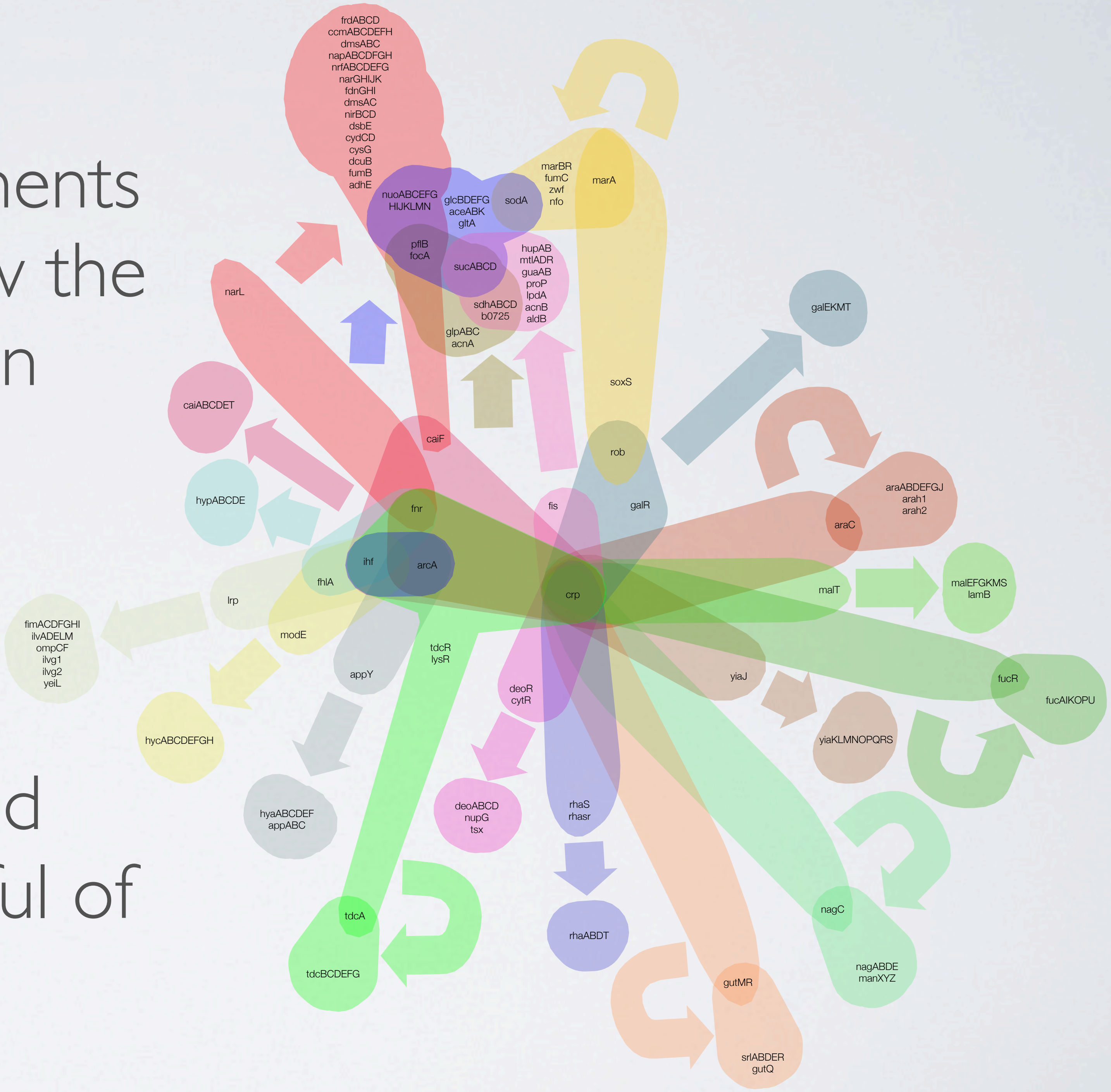
The operons are controlled combinatorially by a handful of key regulatory genes.



E. COLI TRANSCRIPTION NETWORK

The compressible components of the *E. coli* network show the organisation of the genes in terms of operons.

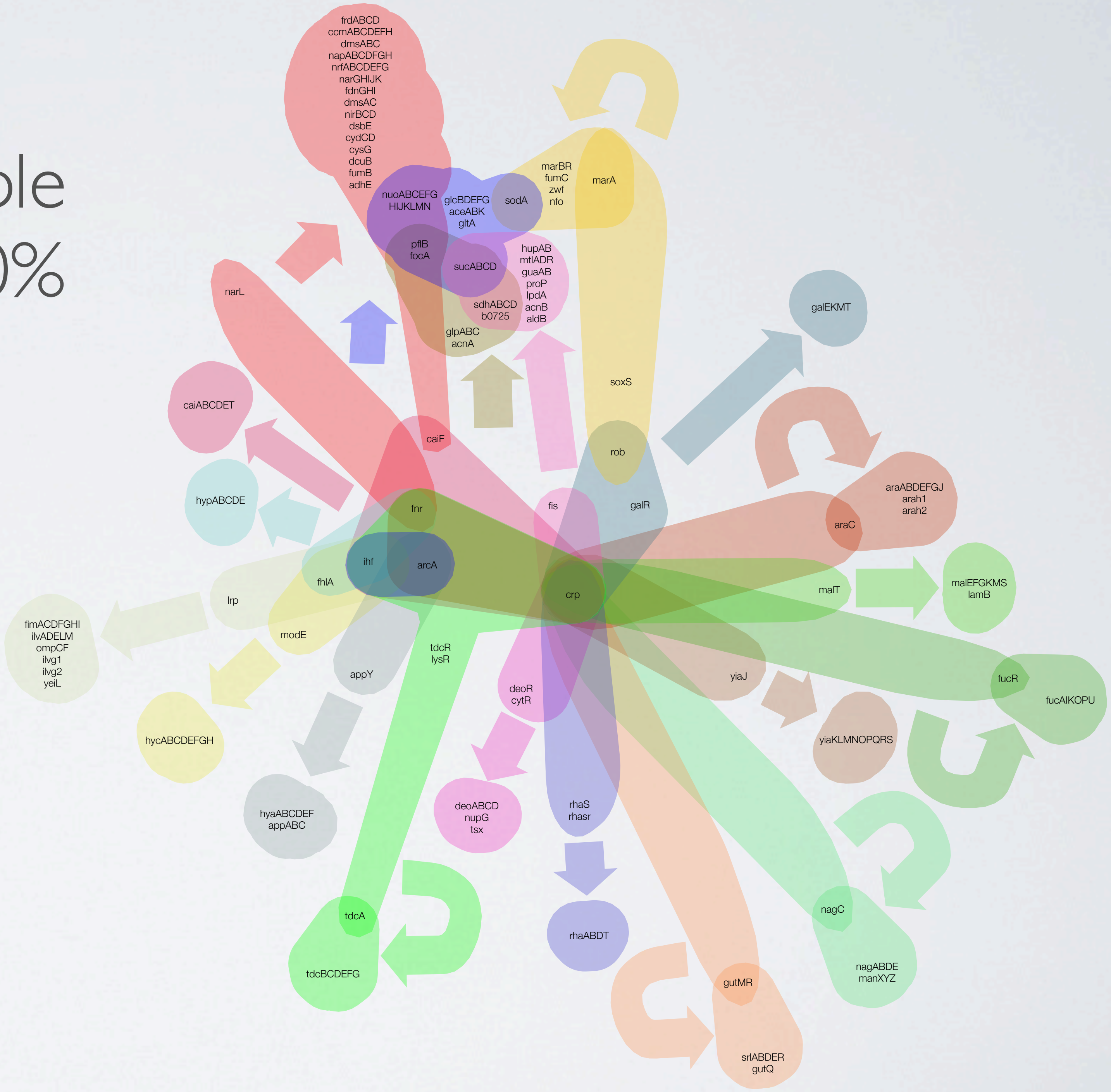
The operons are controlled combinatorially by a handful of key regulatory genes.



S. E. Ahnert, Mol Biosyst 9, 2681 (2013)

E. COLI TRANSCRIPTION NETWORK

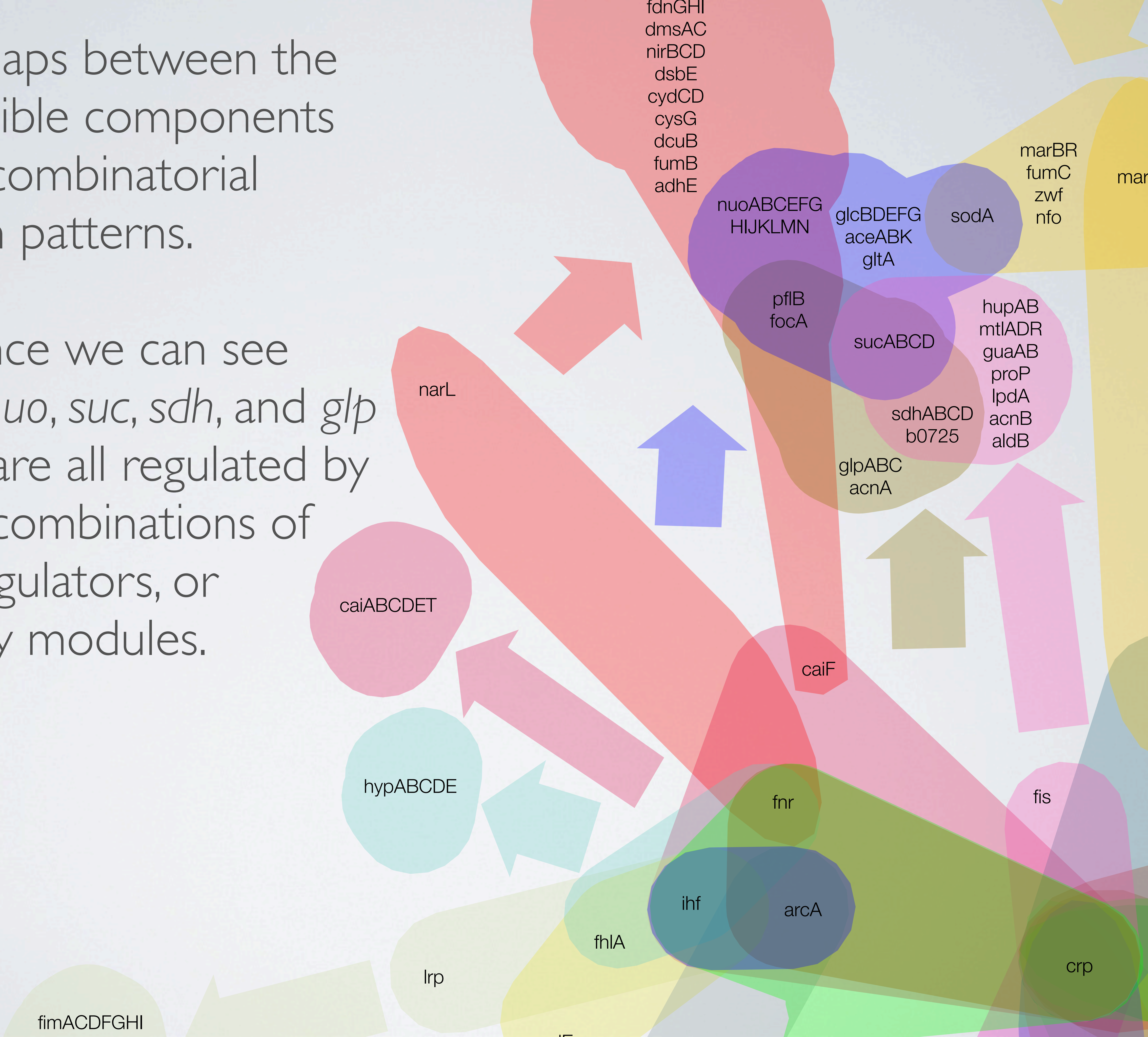
In the 20 most compressible components (pictured), 70% of power nodes exhibited significant GO term enrichment.

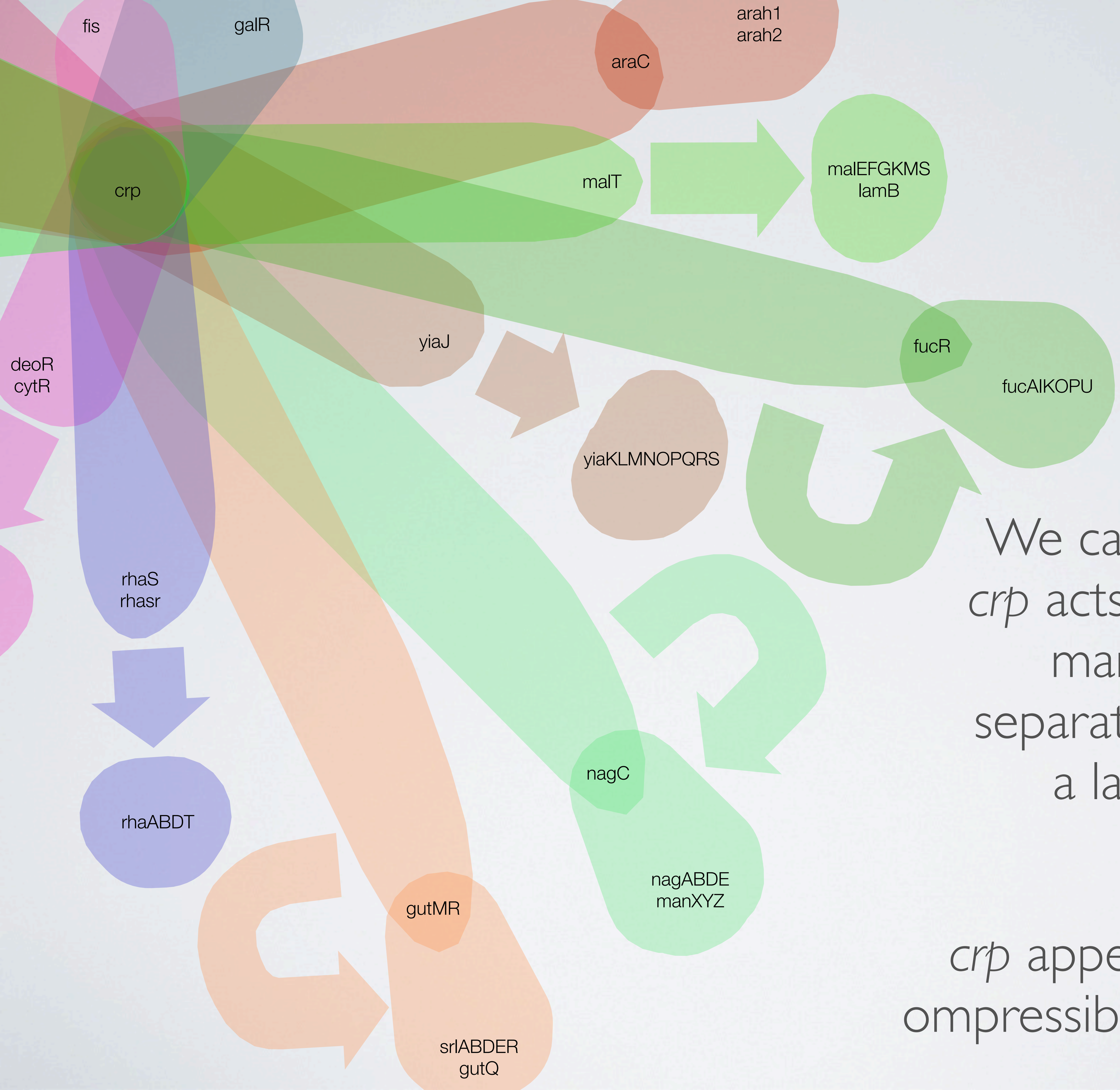


S. E. Ahnert, Mol Biosyst 9, 2681 (2013)

The overlaps between the compressible components highlight combinatorial regulation patterns.

For instance we can see that the *nuo*, *suc*, *sdh*, and *glp* operons are all regulated by different combinations of sets of regulators, or regulatory modules.





We can also see that *crp* acts together with many other genes separately to regulate a large number of operons.

crp appears in 29 of 62 ompressible components.

GLOBAL EFFICIENCY

GLOBAL EFFICIENCY

Global efficiency of a network is defined as the inverse of the harmonic mean of the shortest path length across all possible node pairs in the network.

$$E_{global} = \frac{1}{N(N-1)} \sum_{i,j} \frac{1}{L_{ij}}$$

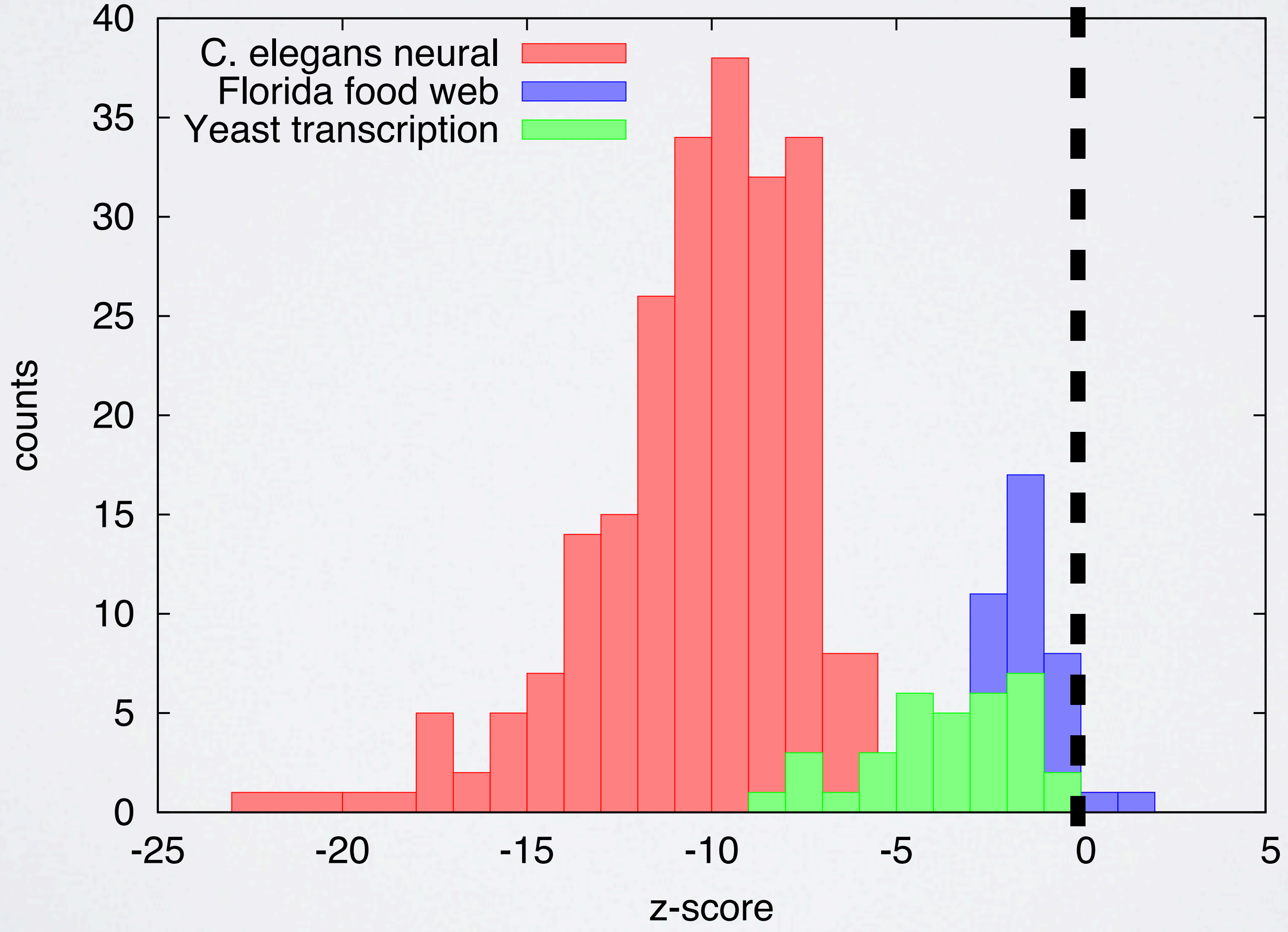
What role do compressible components play in terms of global efficiency?

GLOBAL EFFICIENCY

If we remove compressible components, the global efficiency is almost always reduced more than for an equally sized set of randomly chosen edges.

GLOBAL EFFICIENCY

If we remove compressible components, the global efficiency is almost always reduced more than for an equally sized set of randomly chosen edges.



FUTURE DIRECTIONS

FUTURE DIRECTIONS

One possible future direction would be to introduce the possibility of imperfect power edges, in which the compression information would include small numbers of edges that are missing from an even more compressible component.

Another would be to construct inference approaches which allow the classification of unknown nodes based on term enrichment and their membership in compressible components.

CONCLUSIONS

Compressible components provide a way to identify and visualise dominant connection structures in networks.

This allows the identification of node sets that are defined by their relationships with other node sets, rather than internal connectivity of sets.

The potentially highly overlapping node sets that emerge from this approach can be classified using term enrichment analysis.

CONCLUSIONS

CONCLUSIONS

In transcription networks this approach identifies sets of genes that regulate other sets of genes in combination.

GO term enrichment can be used to assign biological meaning to these overlapping sets.

The partial overlaps of power nodes are a crucial characteristic for the description of transcription networks.

PUBLICATIONS

PUBLICATIONS

The work on transcription networks
has been published in:

S. E. Ahnert, Mol. Biosyst. 9, 2681 (2013)

A more general publication is currently under review.

ACKNOWLEDGMENTS

This work was supported by the Royal Society,
and has been applied in collaboration with:

Emma Towlson, Ed Bullmore, and Petra Vertes
(Cambridge University)

Siobhan Brady and Mallorie Taylor-Teeples
(University of California at Davis)

THANK YOU

for your attention!