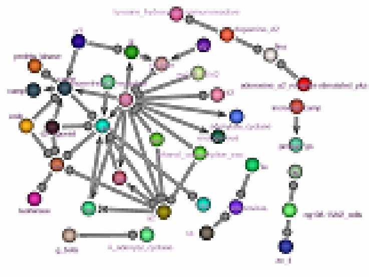
Translating Content – Text Mining

Andrey Rzhetsky

Columbia University



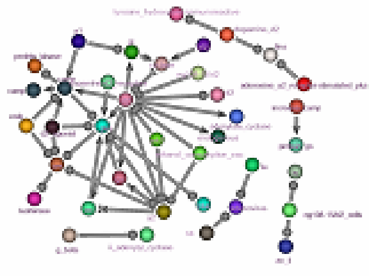
I hope to cover...

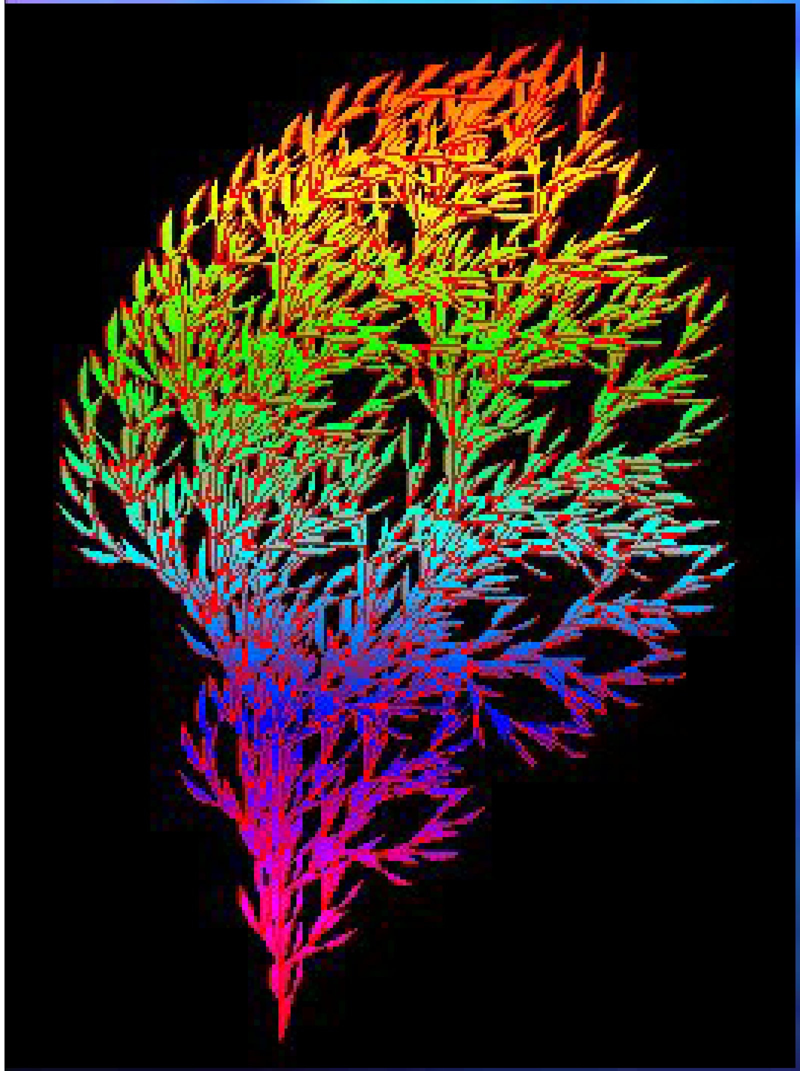
GeneWays 1

Chains of reasoning 2

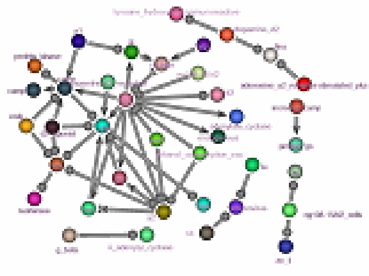
Knowledge as a coral 3

Complex traits 4



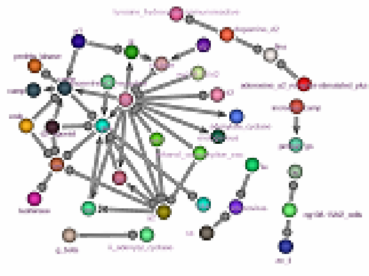


Fractal analogy



Numbers

- Index Medicus®: a monthly subject/author guide to articles in 4,000 medical journals.
- BIOSIS®: Approximately 560,000 new records added each year from 5,000 biological journals
- Chemical Abstracts®: provides references to articles in over 14,000 journals in the field of chemistry
- ...

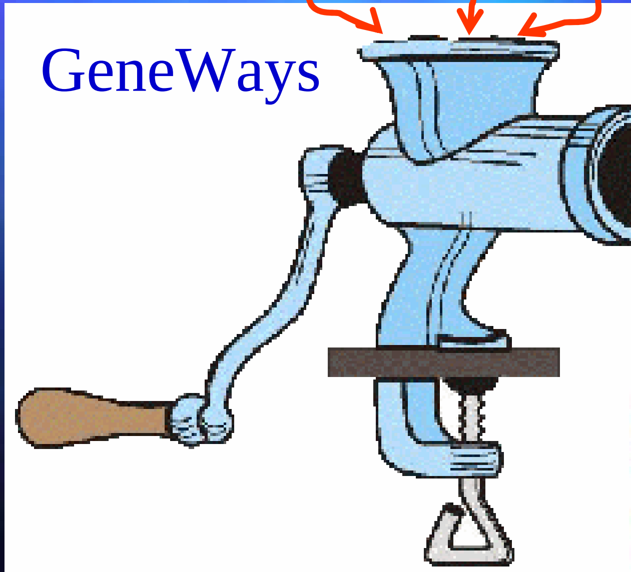


GeneWays as an info-grinder

The Journal of Biological Chemistry
Cell
Blue Light Photoreceptors in Purple Bacteria
nature
Plasmodium genomics

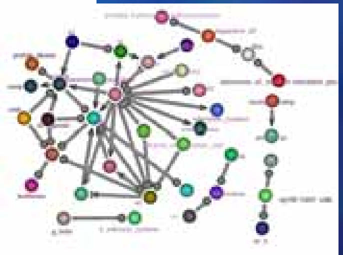
On-line Journals

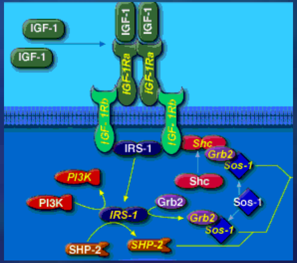
text extraction
↓
automatic curation
↓
interaction database

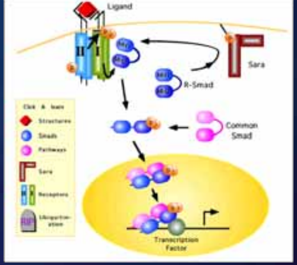


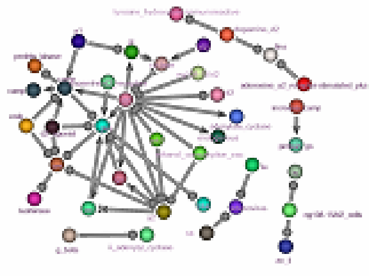
GeneWays

Pathways

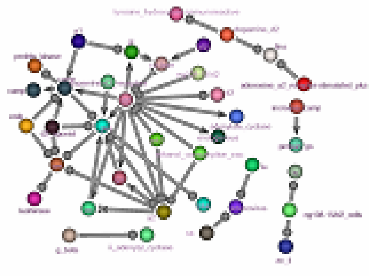








Networks in the core



4 Complex traits

GeneWays 1

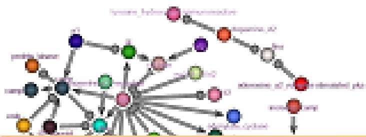
I hope to cover...

Knowledge as a coral

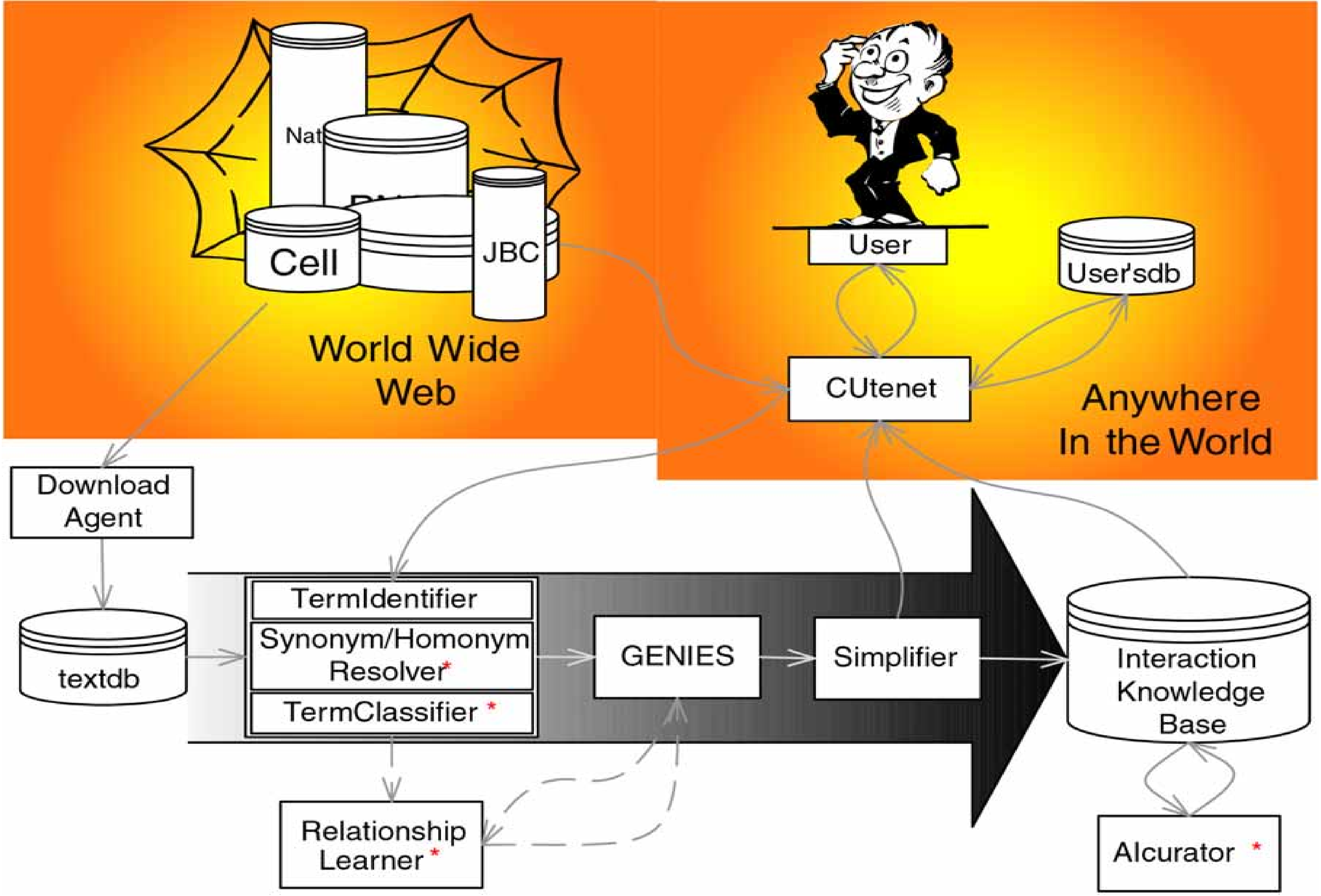
3

Chains of reasoning

2



GeneWays Architecture



World Wide Web

User

User'sdb

CUtenet

Anywhere In the World

Download Agent

textdb

TermIdentifier

Synonym/Homonym Resolver*

TermClassifier *

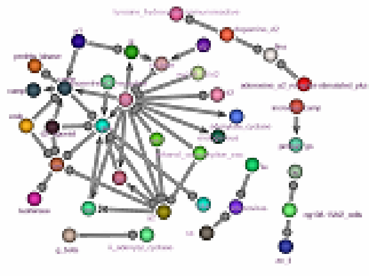
GENIES

Simplifier

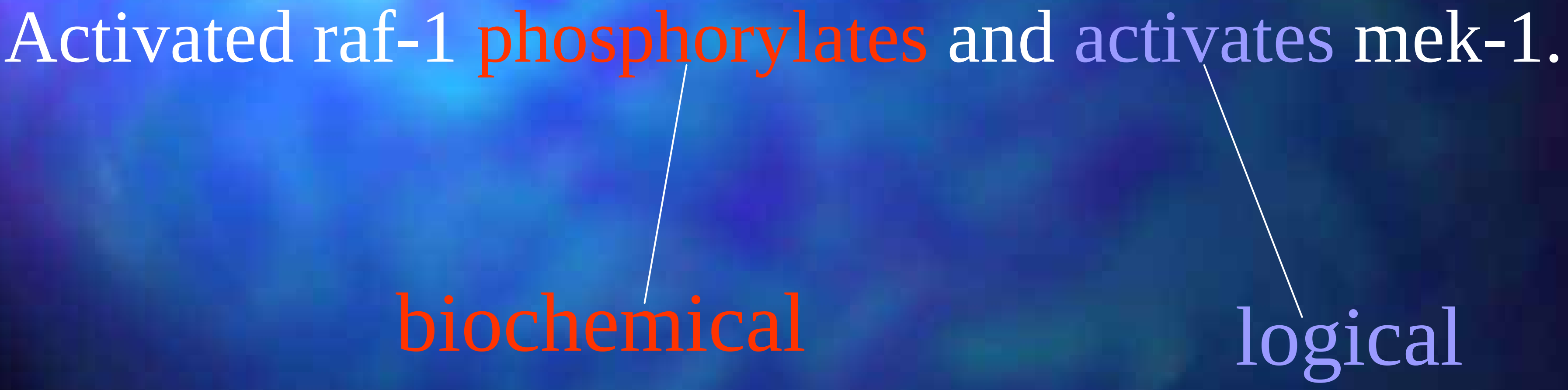
Interaction Knowledge Base

Relationship Learner *

Alcurator *



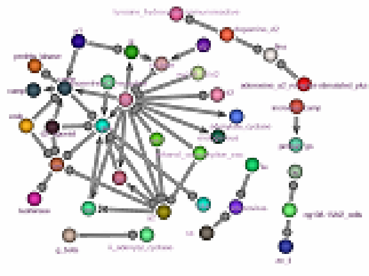
Both logical and biochemical descriptions can be combined in the same sentence:



Activated raf-1 phosphorylates and activates mek-1.

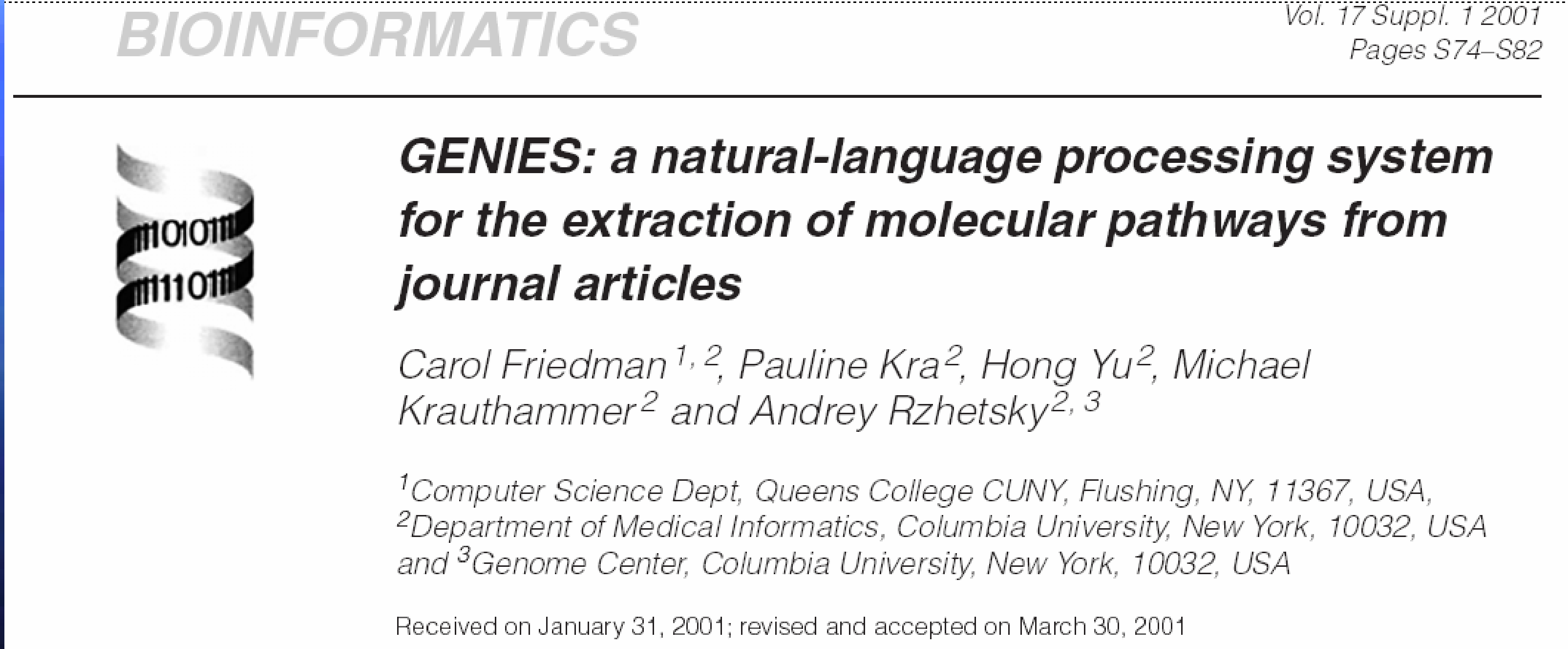
biochemical

logical



GENIES

- Obtains a full parse of the sentence



BIOINFORMATICS

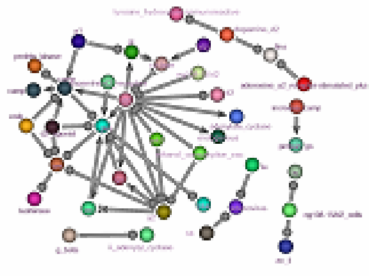
Vol. 17 Suppl. 1 2001
Pages S74–S82

GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles

Carol Friedman[1, 2], Pauline Kra[2], Hong Yu[2], Michael Krauthammer[2] and Andrey Rzhetsky[2, 3]

[1]Computer Science Dept, Queens College CUNY, Flushing, NY, 11367, USA, [2]Department of Medical Informatics, Columbia University, New York, 10032, USA and [3]Genome Center, Columbia University, New York, 10032, USA

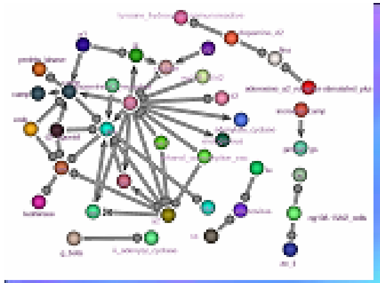
Received on January 31, 2001; revised and accepted on March 30, 2001



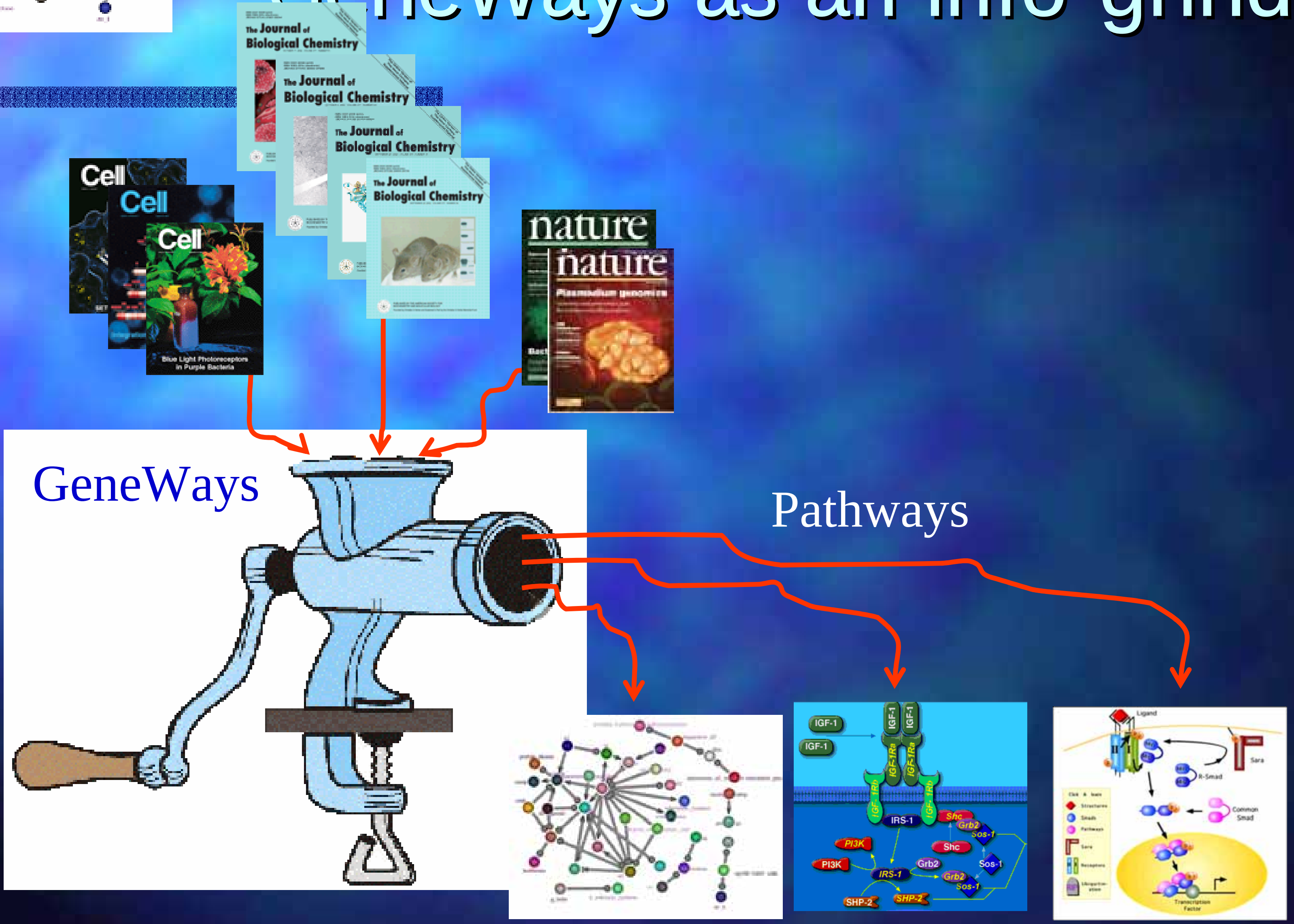
GENIES example

- Mediation of <sonic hedgehog>-induced expression of <Coup-Tfii> by a <protein phosphatase>

- [action, promote, [protein, phosphatase],
  [action, activate, [protein, sonic hedgehog],
  [action, express, [gene, Coup-Tfii]]]]

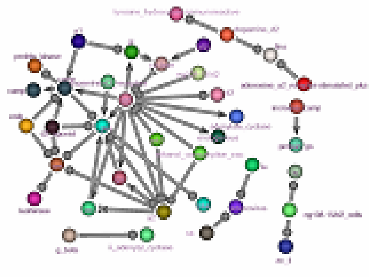


GeneWays as an info-grinder



GeneWays

Pathways

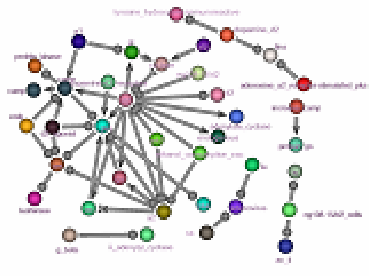


Actions

most relevant to proteins

1001,'bind'
1004,'suppress'
1011,'replace'
1018,'interact'
1020,'activate'
1022,'stimulate'
1023,'phosphorylate'
1027,'increase'
1028,'associate'
1034,'up-regulate'
1036,'inhibit'
1040,'promote'
1041,'down-regulate'
1043,'trigger'

1049,'block'
1054,'modify'
1057,'digest'
1058,'degrade'
1062,'link'
1071,'cleave'
1072,'release'
1074,'catalyze'
1083,'inactivate'
1106,'repress'
1110,'acetylate'
1117,'methylate'



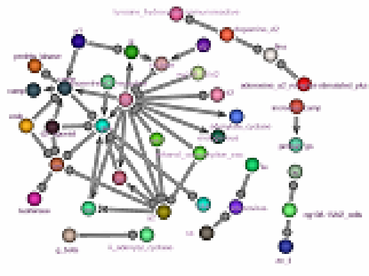
Typical "nodes" of the pathway graph

17767,'calcium channel antagonists'
20324,'hsp70 chaperone'
17467,'activator protein 1'
13194,'tyrosyl-phosphorylated'
4190,'immunodeficiency'
8552,'human fcgammarii'
13151,'ikaros'
7277,'virus-triggered p-dcs'
12290,'anti-alpha4 mabs'

database ID

16044,'alk5'
10393,'mek-1 inhibitor'
13262,'pro-matrilysin'
6584,'gi-type g-protein'
4708,'cell surface: vla-5'
19378,'hla protein'
7145,'tissue proteases'
7653,'smad-7'
9918,'ephb6'
12584,'th2-driven airway inflammation

5104,'daunorubicin'
9689,'paroxonase'
4478,'iga2'
4472,'iga1'
9820,'caveolin 1'
4366,'complexes pr-3'
2258,'gal4-mef2d'
14464,'polyneuropathy'
2253,'gal4-mef2a'
6874,'via l'
19253,'pro-b'



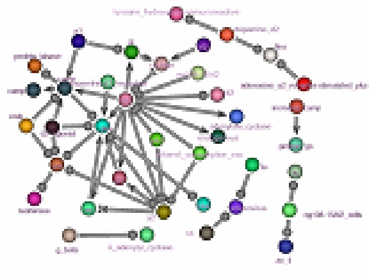
4 Complex traits

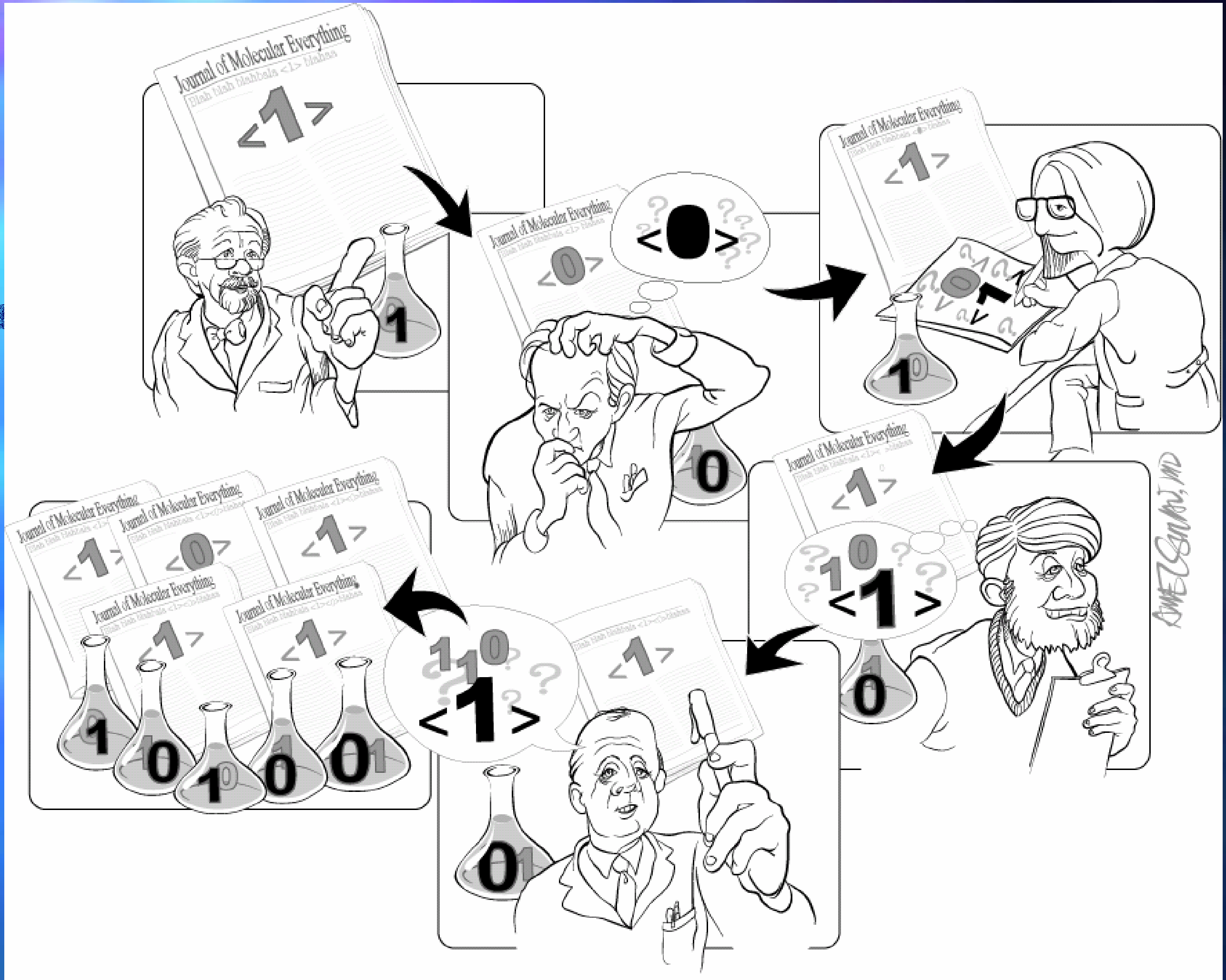
GeneWays 1

I hope to cover...

Knowledge as a coral 3

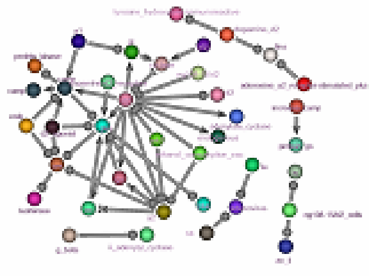
Chains of reasoning 2



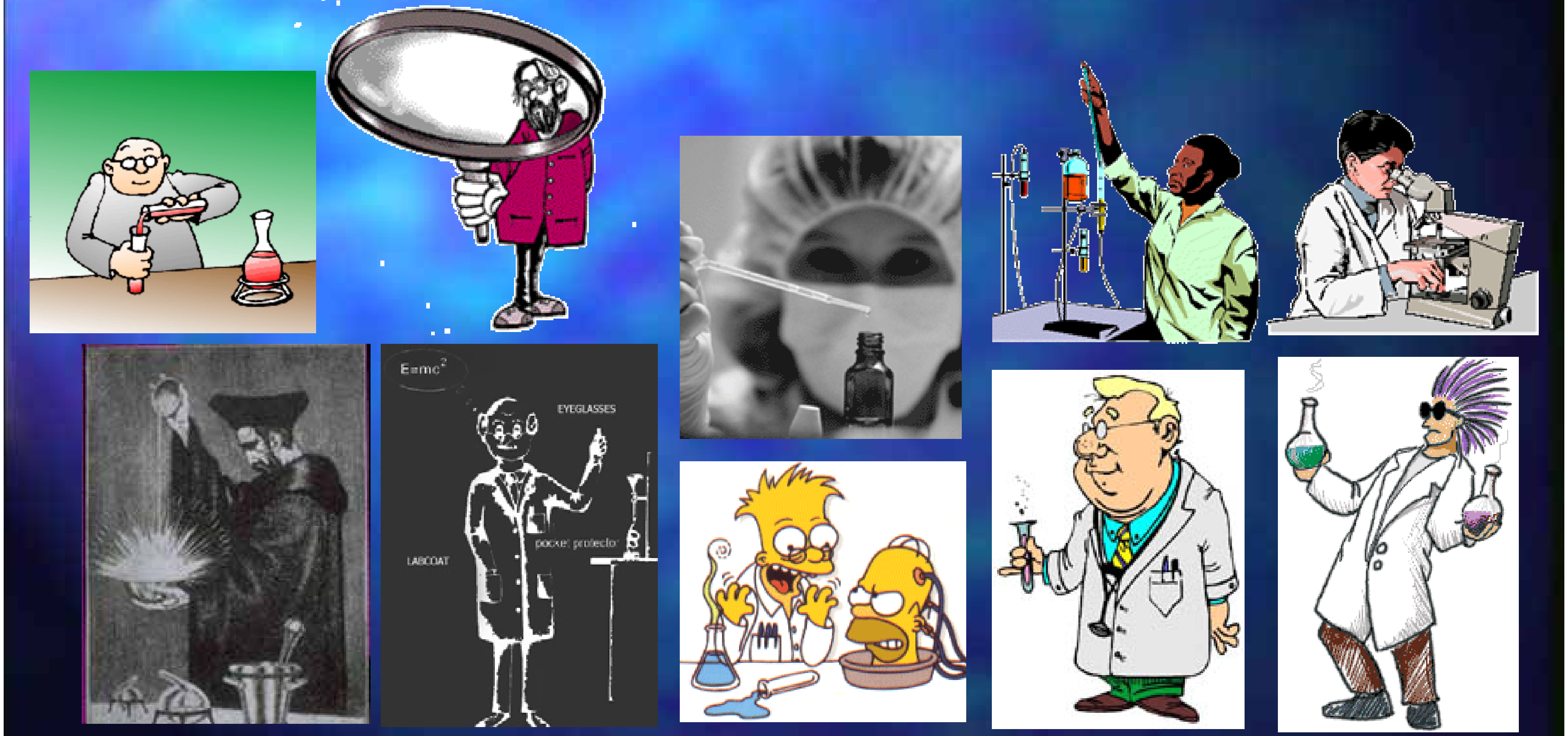

Journal of Molecular Everything
<1>
1
Journal of Molecular Everything
<0>
<0>
0
Journal of Molecular Everything
<1>
10
Journal of Molecular Everything
<1>
<1>
0
Journal of Molecular Everything
<1>
110
<1>
0
Journal of Molecular Everything
<1>
<0>
<1>
1
0
1
0
0

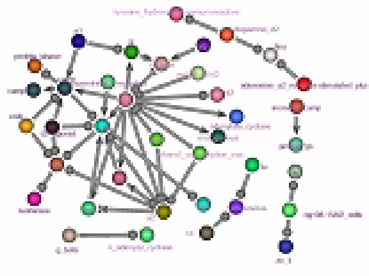
"Chains of collective reasoning" model

Andrey Rzhetsky, Ivan Iossifov, Ji Meng Loh, Kevin P. White

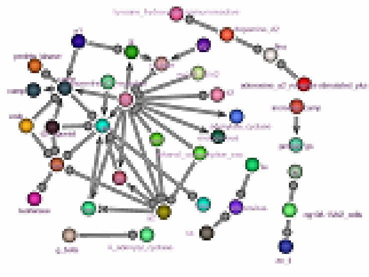


Modeling science...



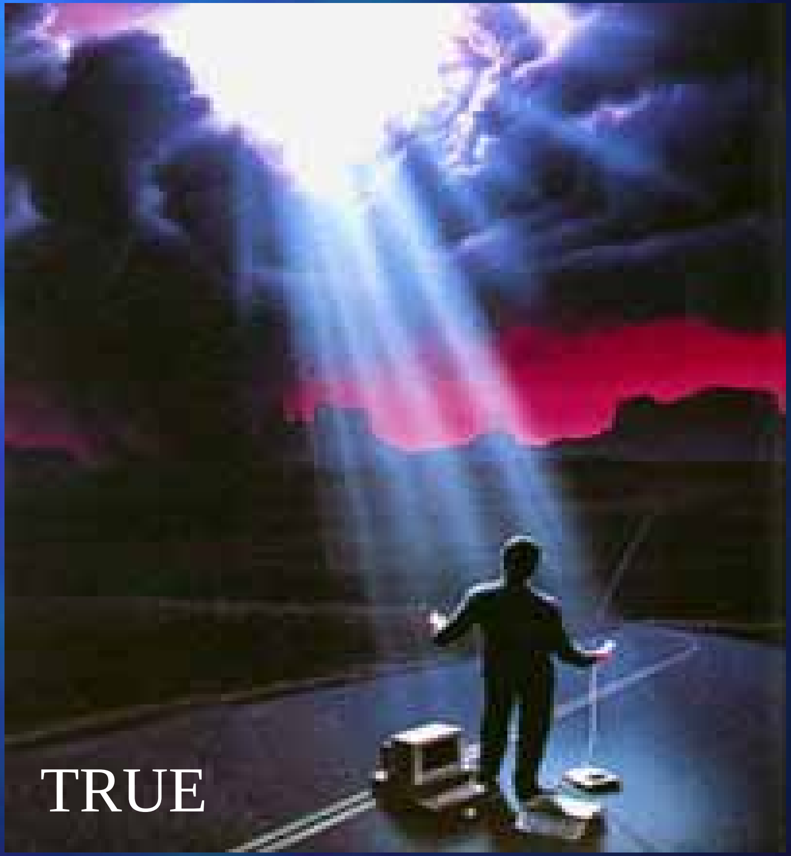


Inferring the truth

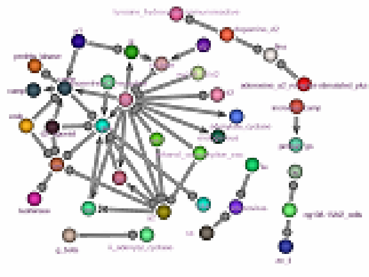


OUR GOAL IS TO DISTINGUISH:

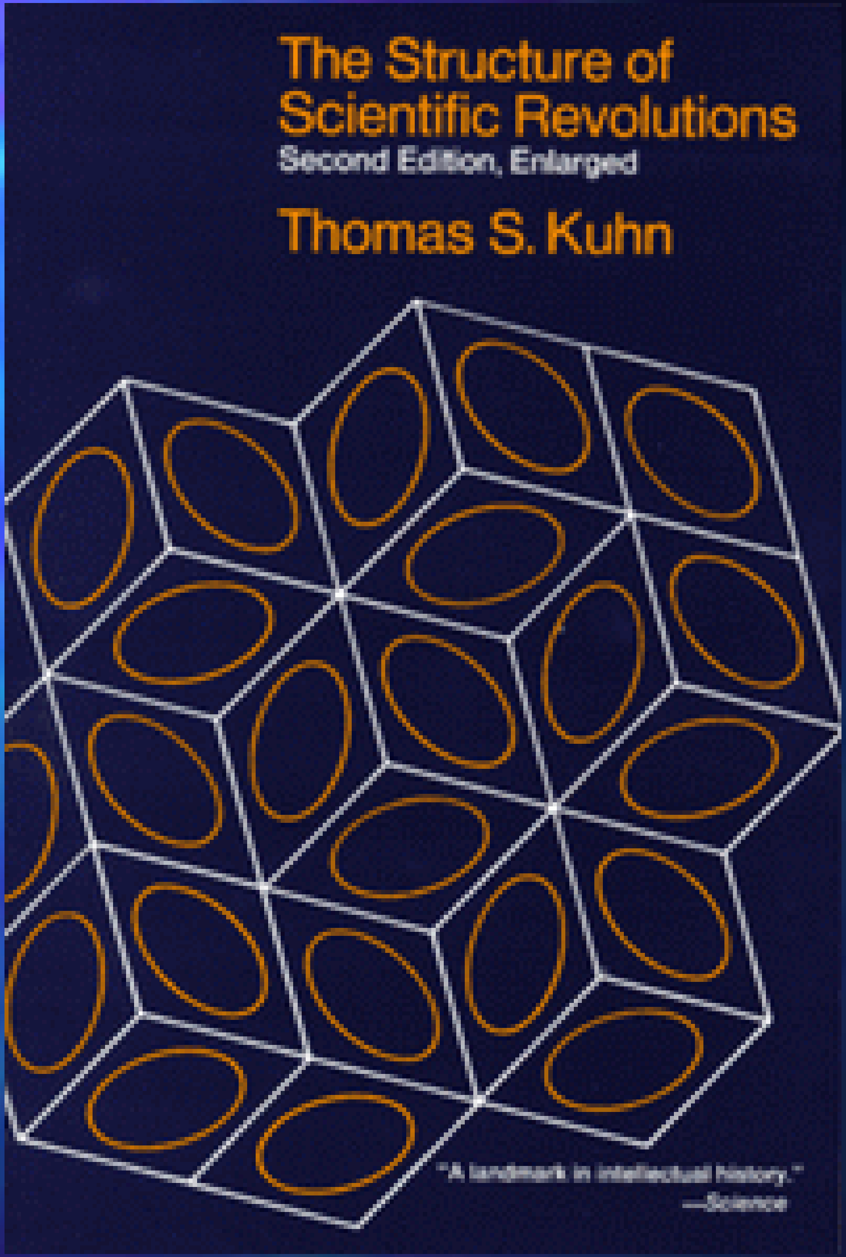
FALSE



TRUE



Scientific revolutions, paradigms

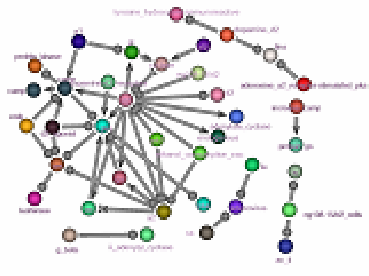


The Structure of Scientific Revolutions

Second Edition, Enlarged

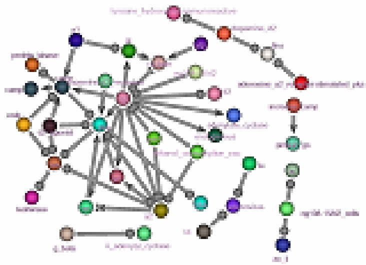
Thomas S. Kuhn

"A landmark in intellectual history."
—Science



In a nutshell ...

Kuhn distinguishes two major states of science: paradigmal or normal science (paradigm is the currently dominant theory that shapes scientist’s perception of the world) and scientific revolution (a process of a rapid change of one paradigm with a new one).



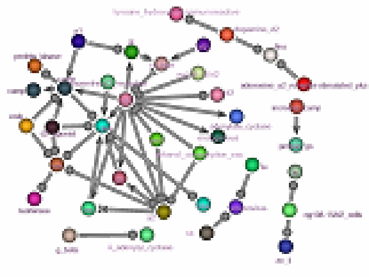
Perception & paradigms

The study of history of science shows that

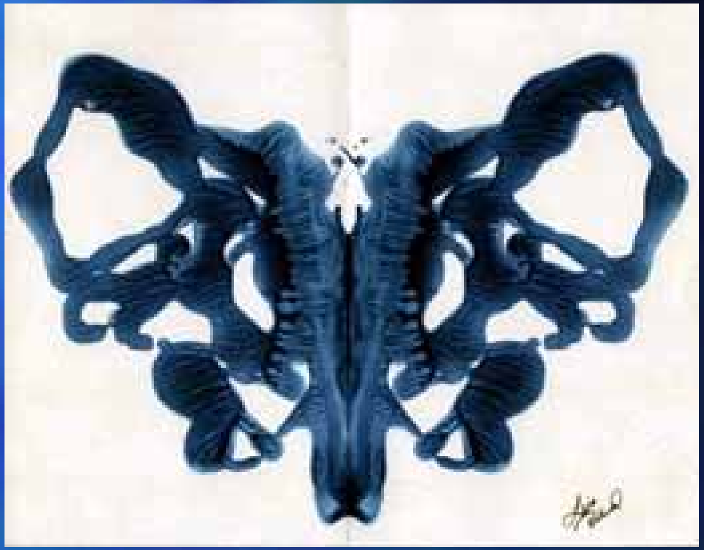
"...paradigm changes do cause scientists to see the world of their research-engagement differently.

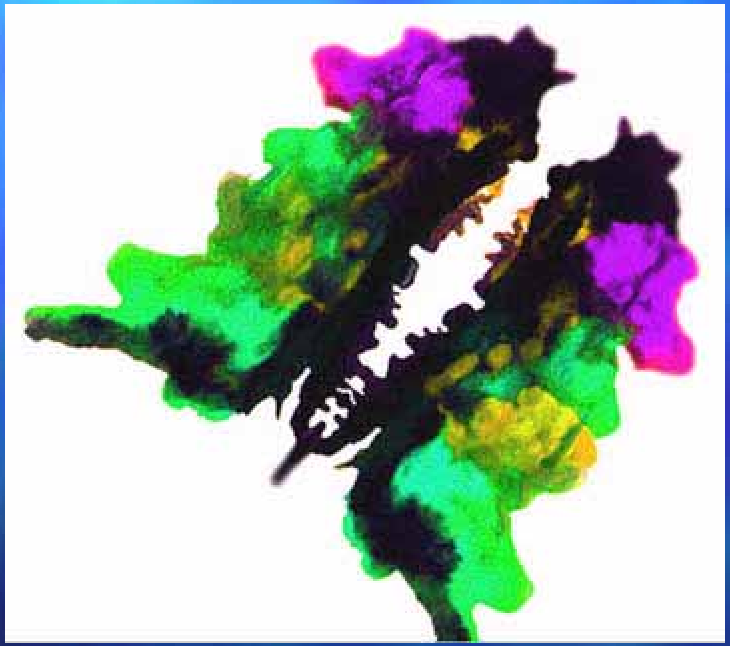
[...] It is as elementary prototypes for these transformations of the scientist's world that the familiar demonstrations of a switch in visual gestalt prove so suggestive. Where were ducks in the scientist's world before the revolution are rabbits afterwards. The man who first saw the exterior of the box from above later sees it from below. Transformations like this, though usually more gradual and almost always irreversible, are common concomitants of scientific training."

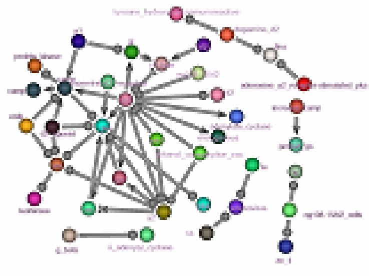
T.S. Kuhn (p.111)



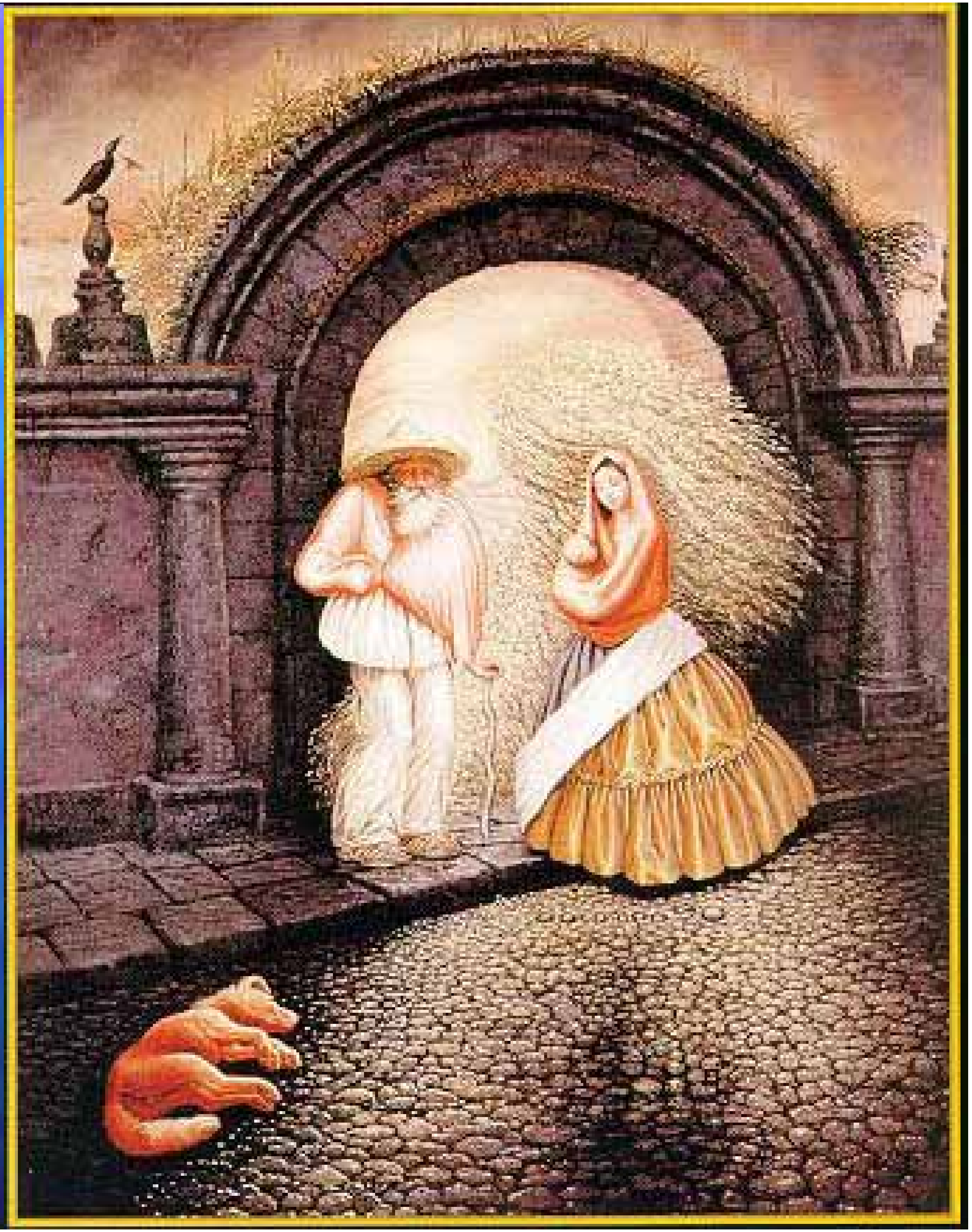
Associations...

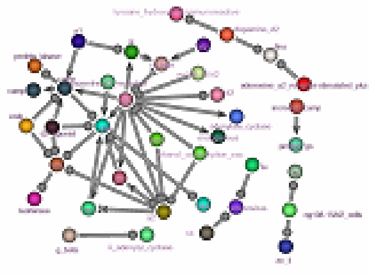






Multiple messages in the same image

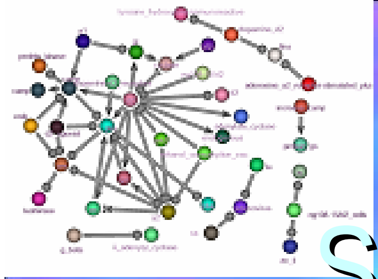




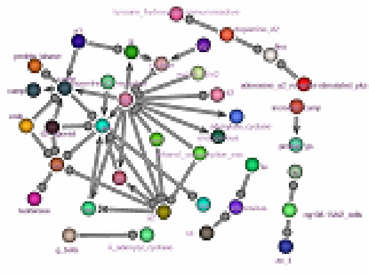
Fuzzy experimental result

0 1

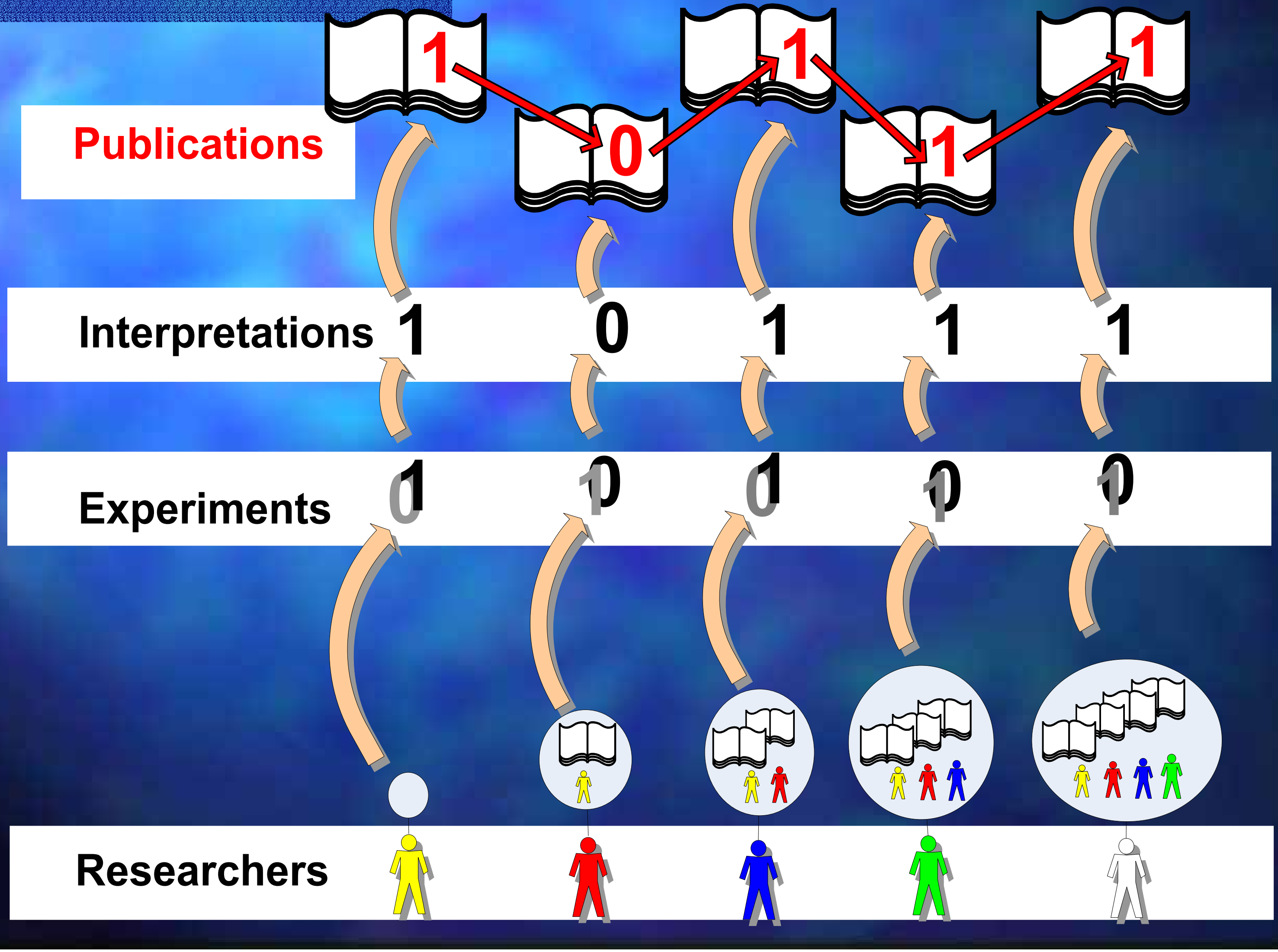
1 0

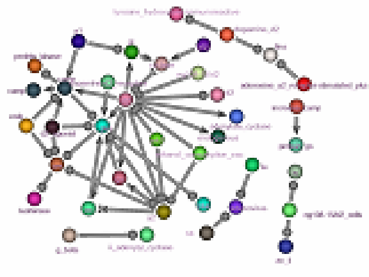


Suggesting a new model...



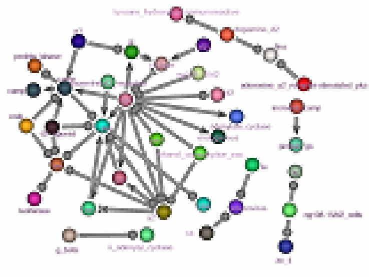
Dependences among publications



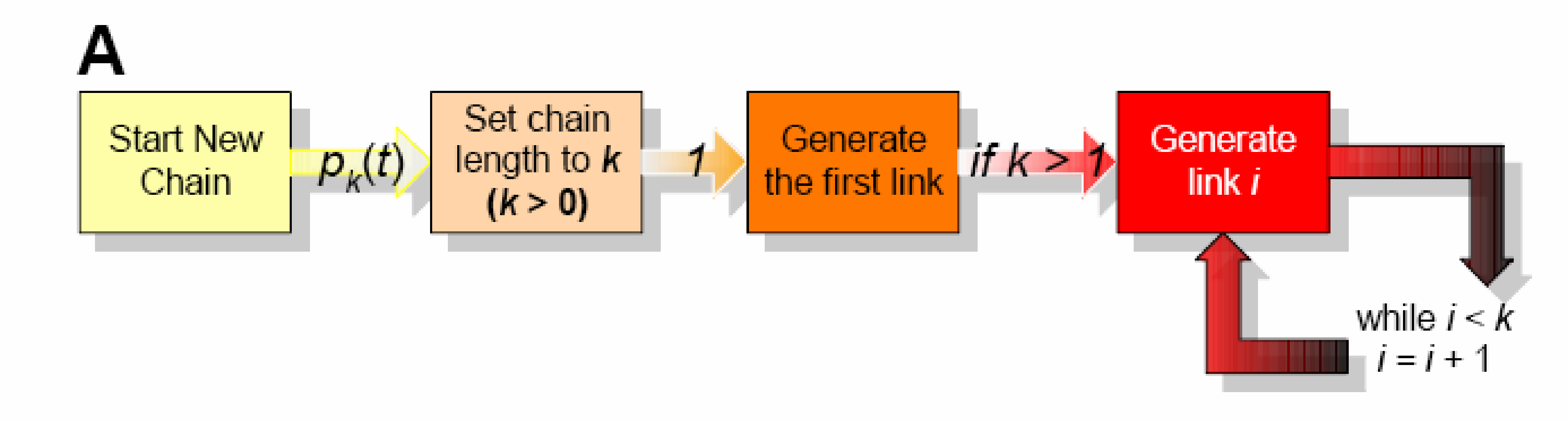


Conform? No!

c1999 Lee Williams/Rick London
London's Times
CONFORM, HELL NO!!
METAL
SAY NO TO CONFORMANCE
METAL
WE WILL NOT CONFORM!!
METAL
CONFORM? HELL NO!!
METAL



More complex/realistic model



A

Start New Chain

$p_k(t)$

Set chain length to ***k*** (***k*** > 0)

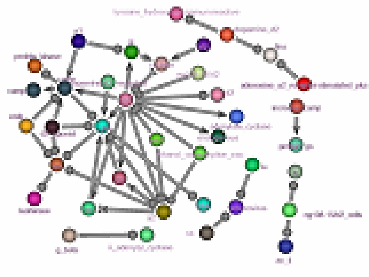
1

Generate the first link

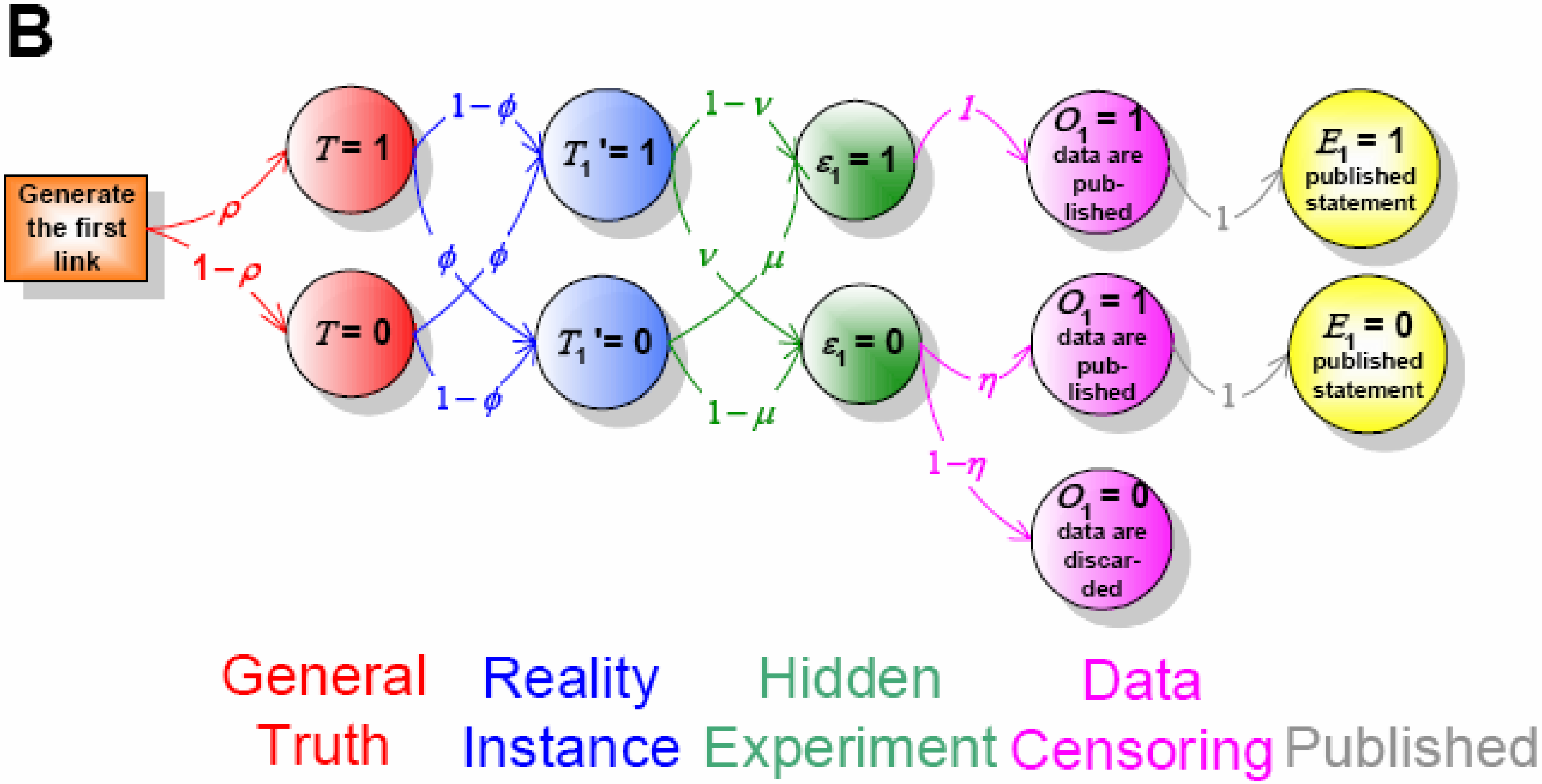
if k > 1

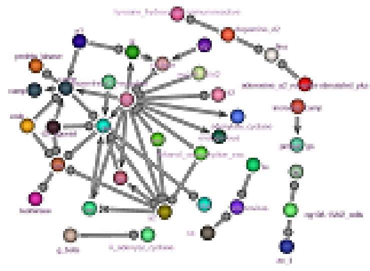
Generate link *i*

while *i* < *k*
i = *i* + 1

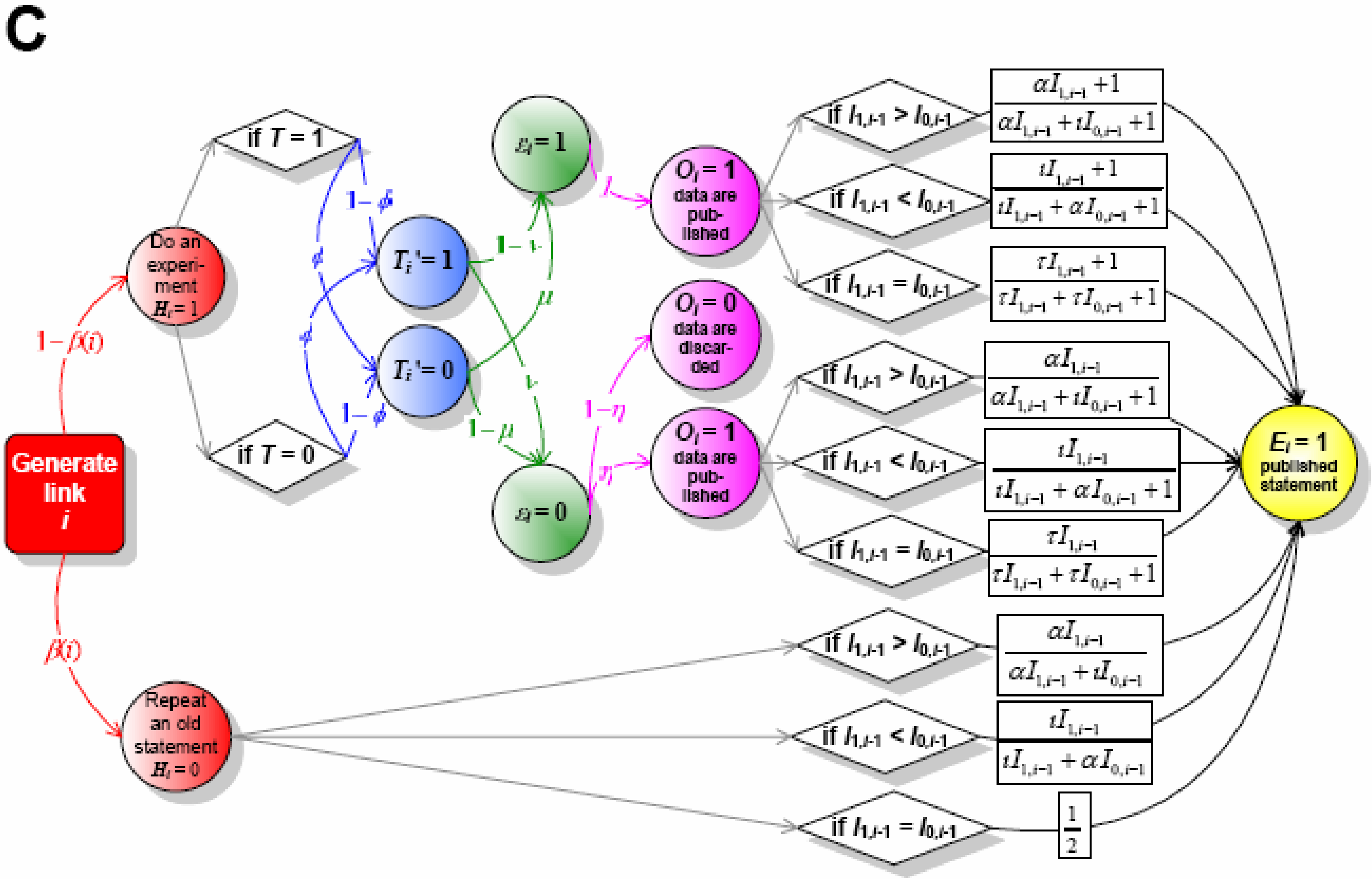


First Link

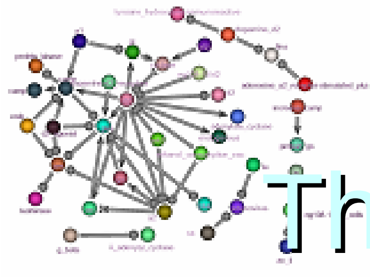

B
Generate the first link
ρ
$1-\rho$
$T = 1$
$T = 0$
$1-\phi$
ϕ
ϕ
$1-\phi$
$T_1' = 1$
$T_1' = 0$
$1-\nu$
ν
μ
$1-\mu$
$\varepsilon_1 = 1$
$\varepsilon_1 = 0$
1
η
$1-\eta$
$O_1 = 1$ data are published
$O_1 = 1$ data are published
$O_1 = 0$ data are discarded
1
1
$E_1 = 1$ published statement
$E_1 = 0$ published statement
General Truth
Reality Instance
Hidden Experiment
Data Censoring
Published



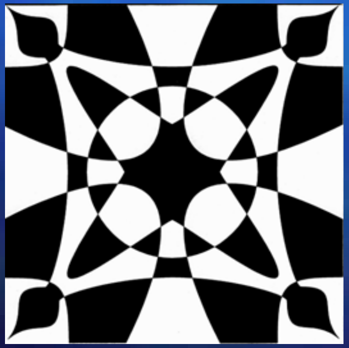
Next link

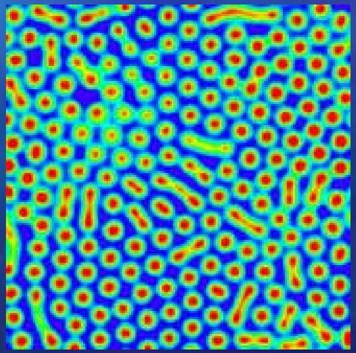


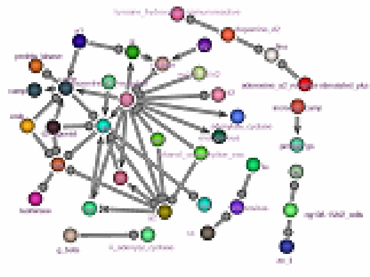
C
Generate link i
$1-\beta(i)$
$\beta(i)$
Do an experiment $H_i = 1$
Repeat an old statement $H_i = 0$
if T = 1
if T = 0
$1-\phi$
ϕ
ϕ
$1-\phi$
$T_i' = 1$
$T_i' = 0$
$1-\iota$
μ
ι
$1-\mu$
$\varepsilon_i = 1$
$\varepsilon_i = 0$
1
$1-\eta$
η
$O_i = 1$ data are published
$O_i = 0$ data are discarded
$O_i = 1$ data are published
if $I_{1,i-1} > I_{0,i-1}$
$\frac{\alpha I_{1,i-1}+1}{\alpha I_{1,i-1}+\iota I_{0,i-1}+1}$
if $I_{1,i-1} < I_{0,i-1}$
$\frac{\iota I_{1,i-1}+1}{\iota I_{1,i-1}+\alpha I_{0,i-1}+1}$
if $I_{1,i-1} = I_{0,i-1}$
$\frac{\tau I_{1,i-1}+1}{\tau I_{1,i-1}+\tau I_{0,i-1}+1}$
if $I_{1,i-1} > I_{0,i-1}$
$\frac{\alpha I_{1,i-1}}{\alpha I_{1,i-1}+\iota I_{0,i-1}+1}$
if $I_{1,i-1} < I_{0,i-1}$
$\frac{\iota I_{1,i-1}}{\iota I_{1,i-1}+\alpha I_{0,i-1}+1}$
if $I_{1,i-1} = I_{0,i-1}$
$\frac{\tau I_{1,i-1}}{\tau I_{1,i-1}+\tau I_{0,i-1}+1}$
if $I_{1,i-1} > I_{0,i-1}$
$\frac{\alpha I_{1,i-1}}{\alpha I_{1,i-1}+\iota I_{0,i-1}}$
if $I_{1,i-1} < I_{0,i-1}$
$\frac{\iota I_{1,i-1}}{\iota I_{1,i-1}+\alpha I_{0,i-1}}$
if $I_{1,i-1} = I_{0,i-1}$
$\frac{1}{2}$
$E_i = 1$ published statement



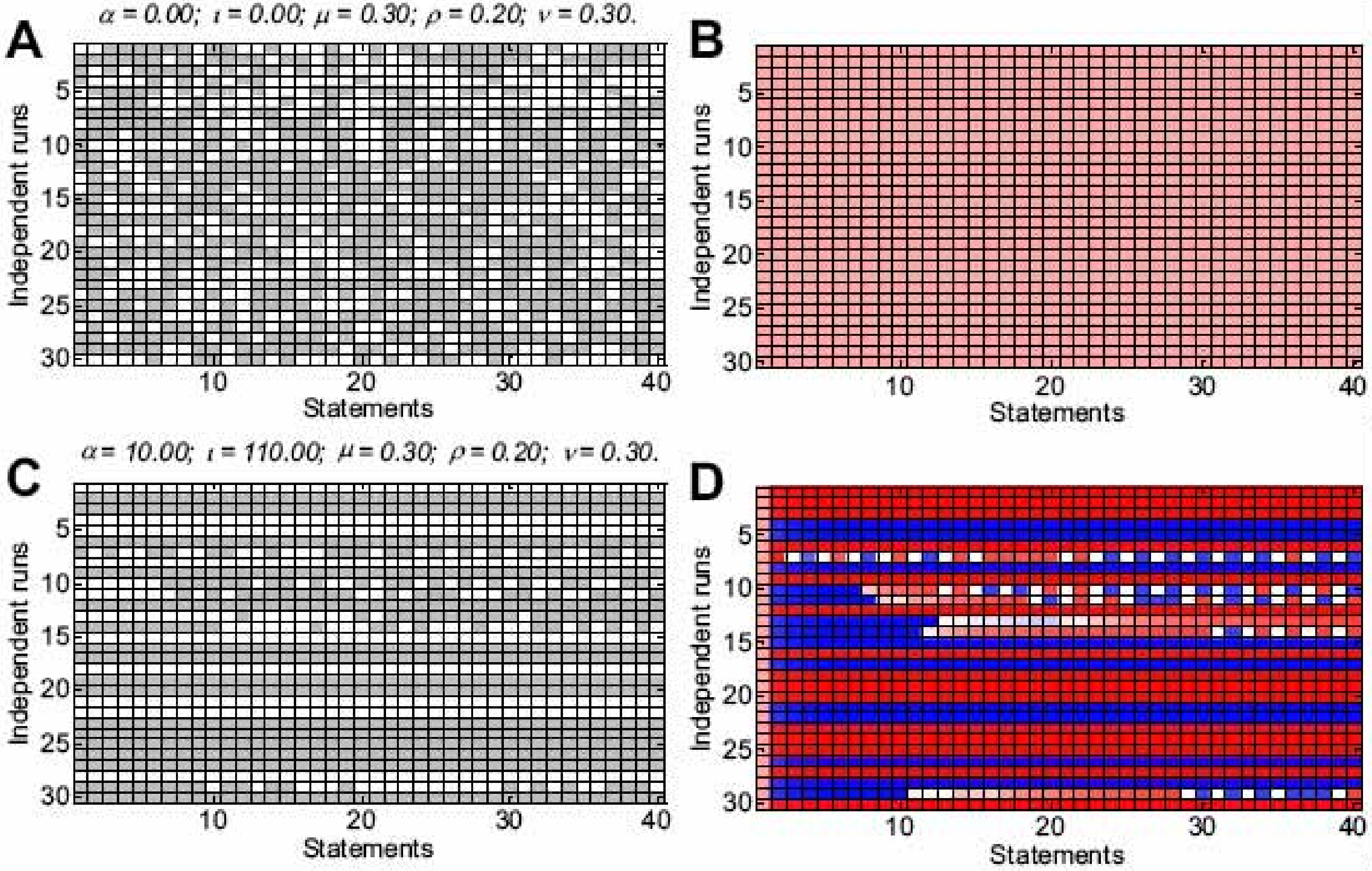
This model is capable of generating diverse patterns (series of zeros and ones) in publications





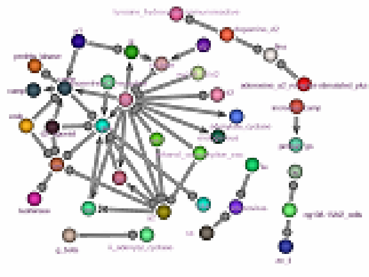


Possible patterns...

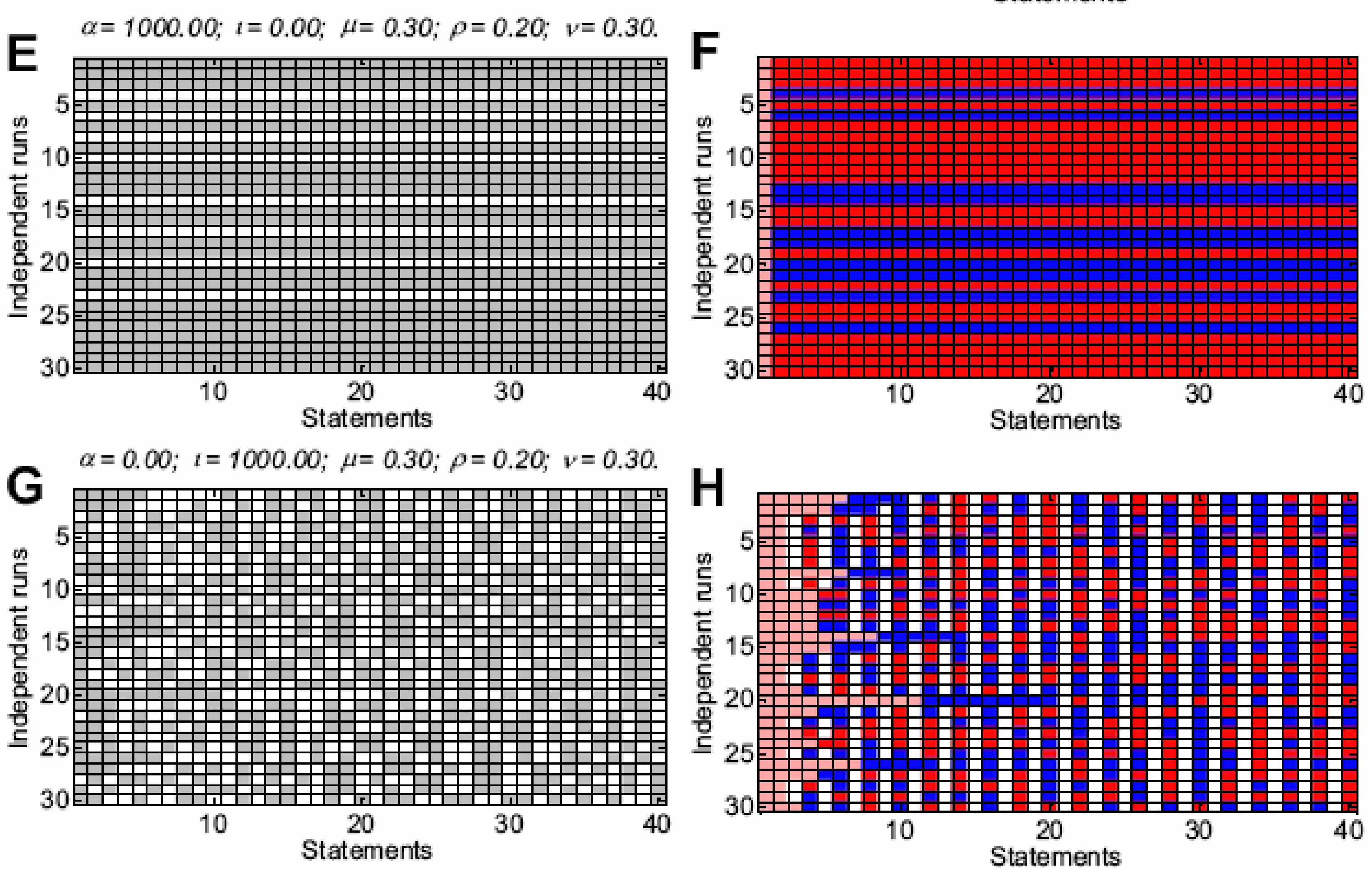


“trust nobody”

“anti-conformism with an inferiority complex”



Patterns (continued)

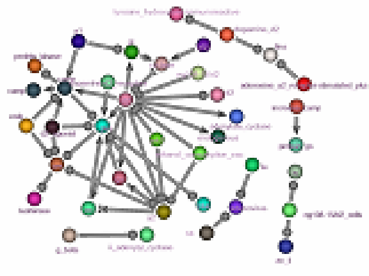


$\alpha = 1000.00;\ \iota = 0.00;\ \mu = 0.30;\ \rho = 0.20;\ \nu = 0.30.$

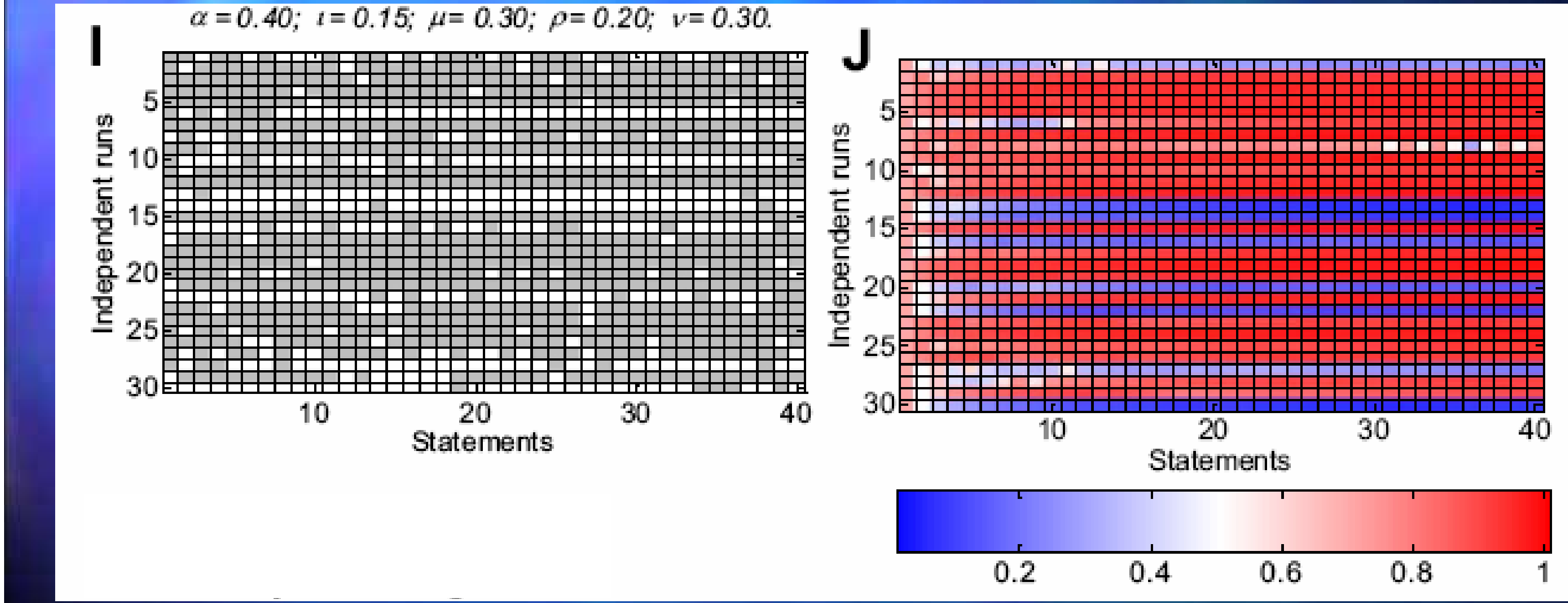
$\alpha = 0.00;\ \iota = 1000.00;\ \mu = 0.30;\ \rho = 0.20;\ \nu = 0.30.$

“super-conformism”

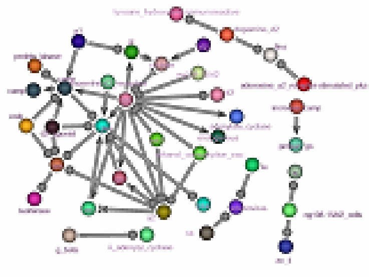
“super-anti-conformism”



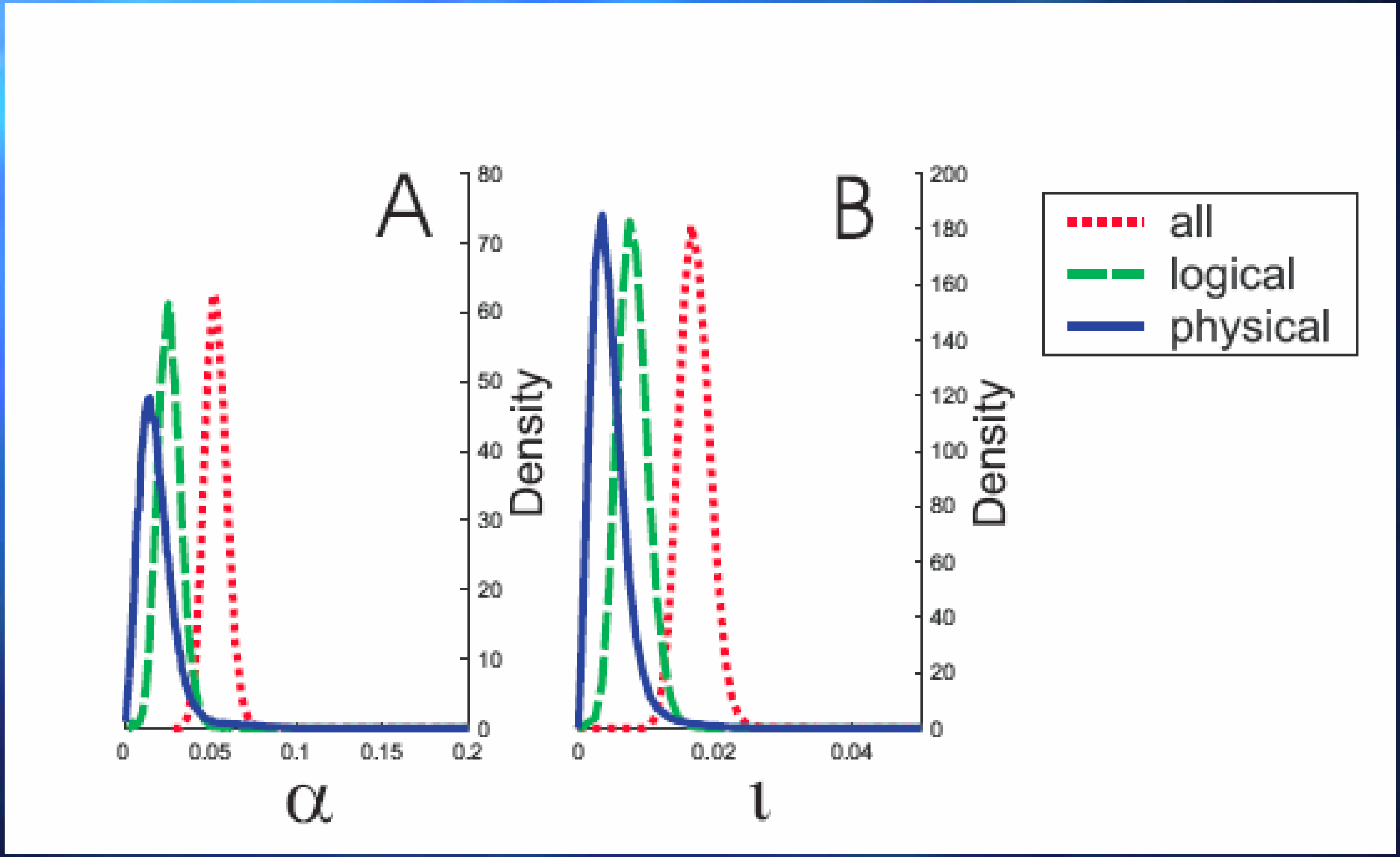
(We have more...)

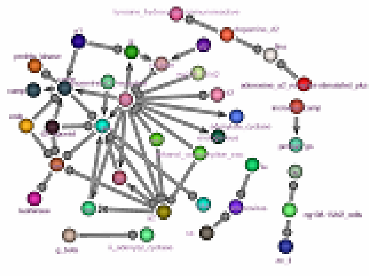


"mild skepticism"

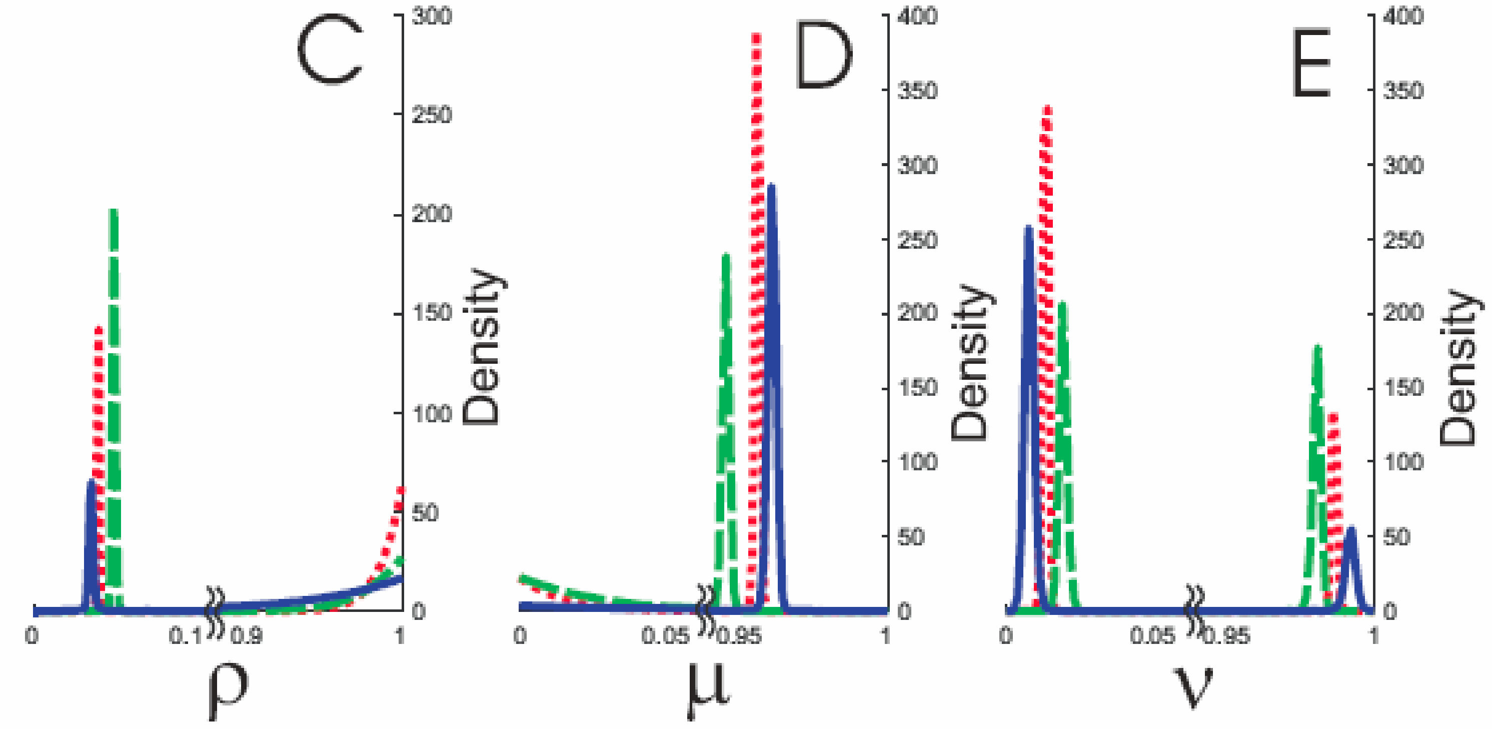


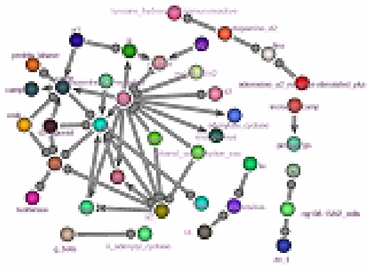
Parameter estimation



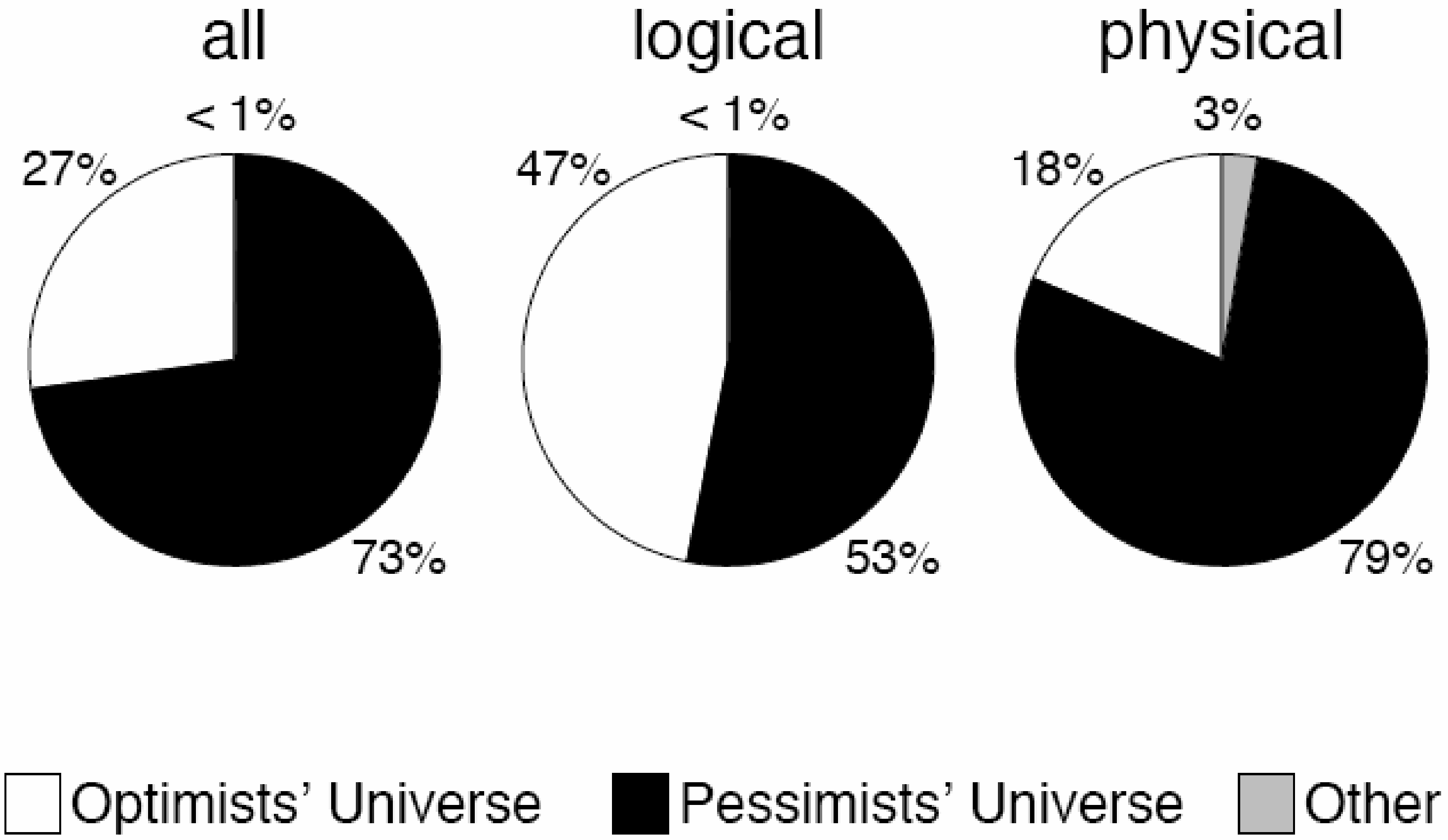


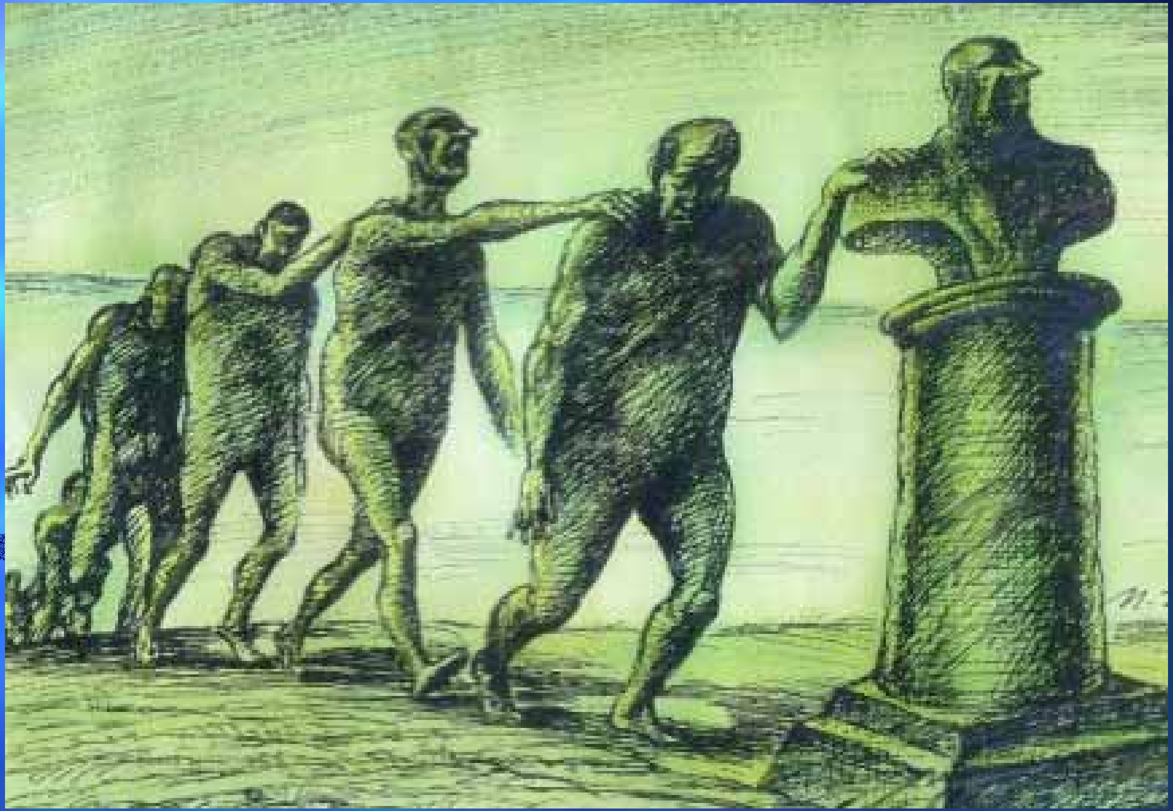
Parameter estimation

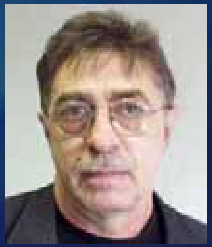




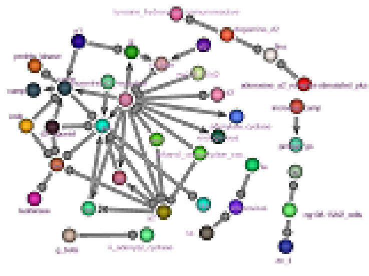
Parameter estimation







(By Michail Zlatkovsky)



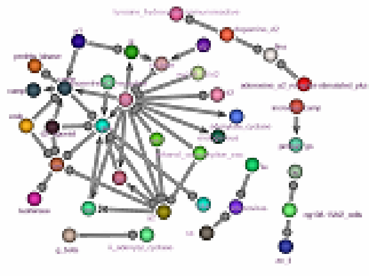
I hope to cover...

GeneWays 1

Chains of reasoning 2

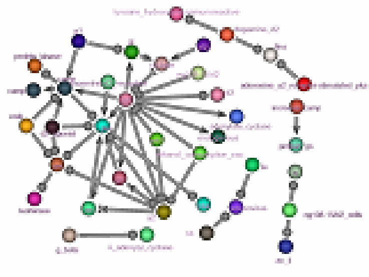
Knowledge as a coral 3

Complex traits 4

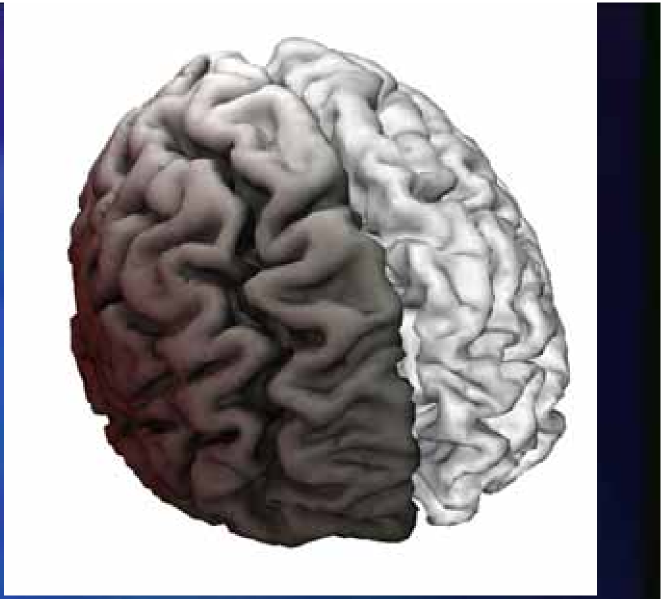


On distributed thinking about molecular networks

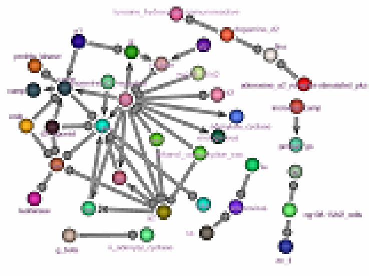
Murat Cokol, Ivan Iossifov, Chani Weinreb, Andrey Rzhetsky



Brain coral analogy

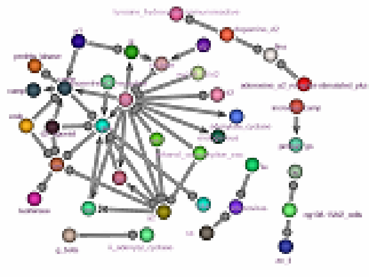


- Mode of knowledge growth
- Surface versus inside
- Knowledge pockets/involutions on the surface
- Coral volume



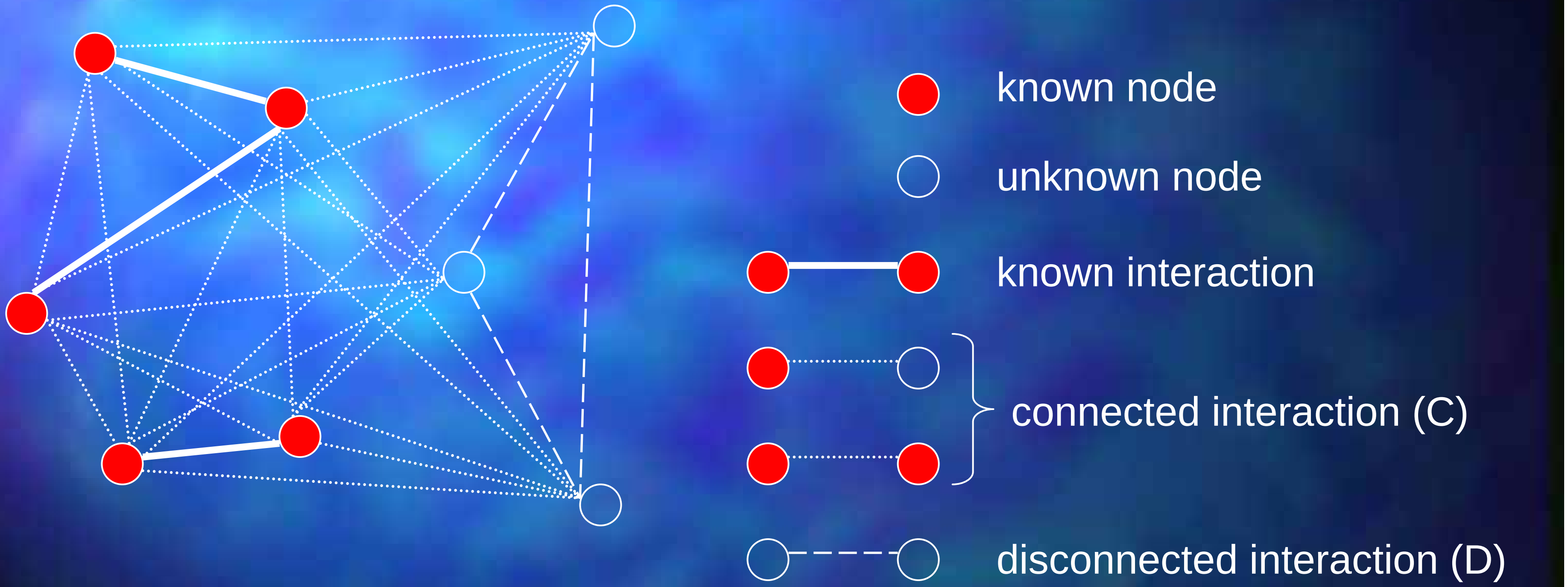
Brain coral analogy

- Mode of knowledge growth
- Surface versus inside
- Knowledge pockets/involutions on the surface
- Coral volume



Jump or crawl?

Connected and disconnected



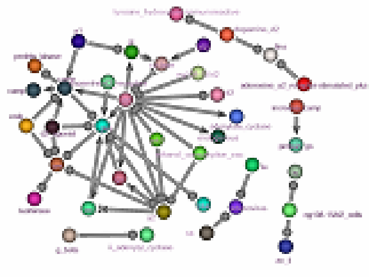
known node

unknown node

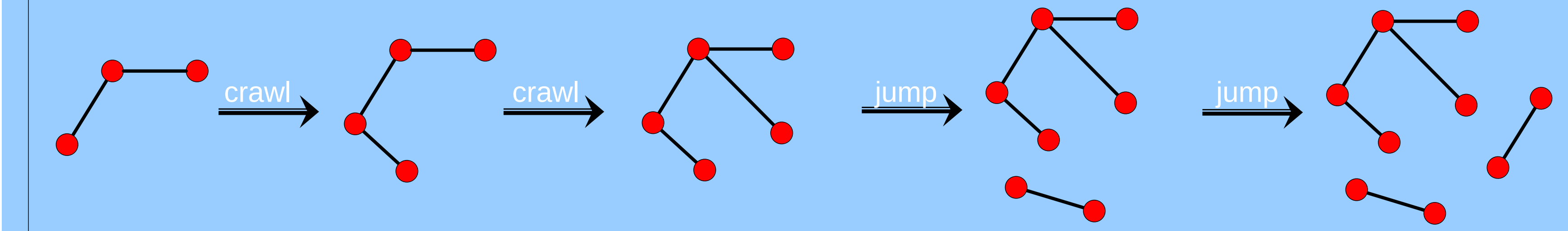
known interaction

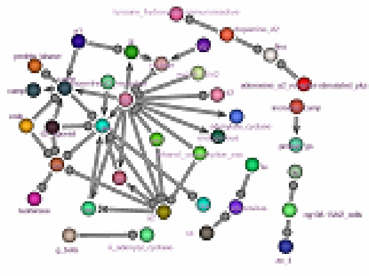
connected interaction (C)

disconnected interaction (D)



Examples: growing knowledge with jump and crawl steps

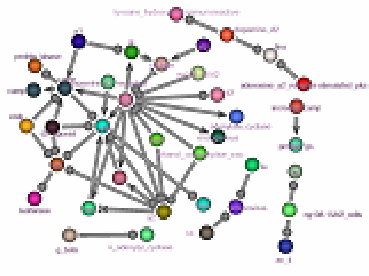




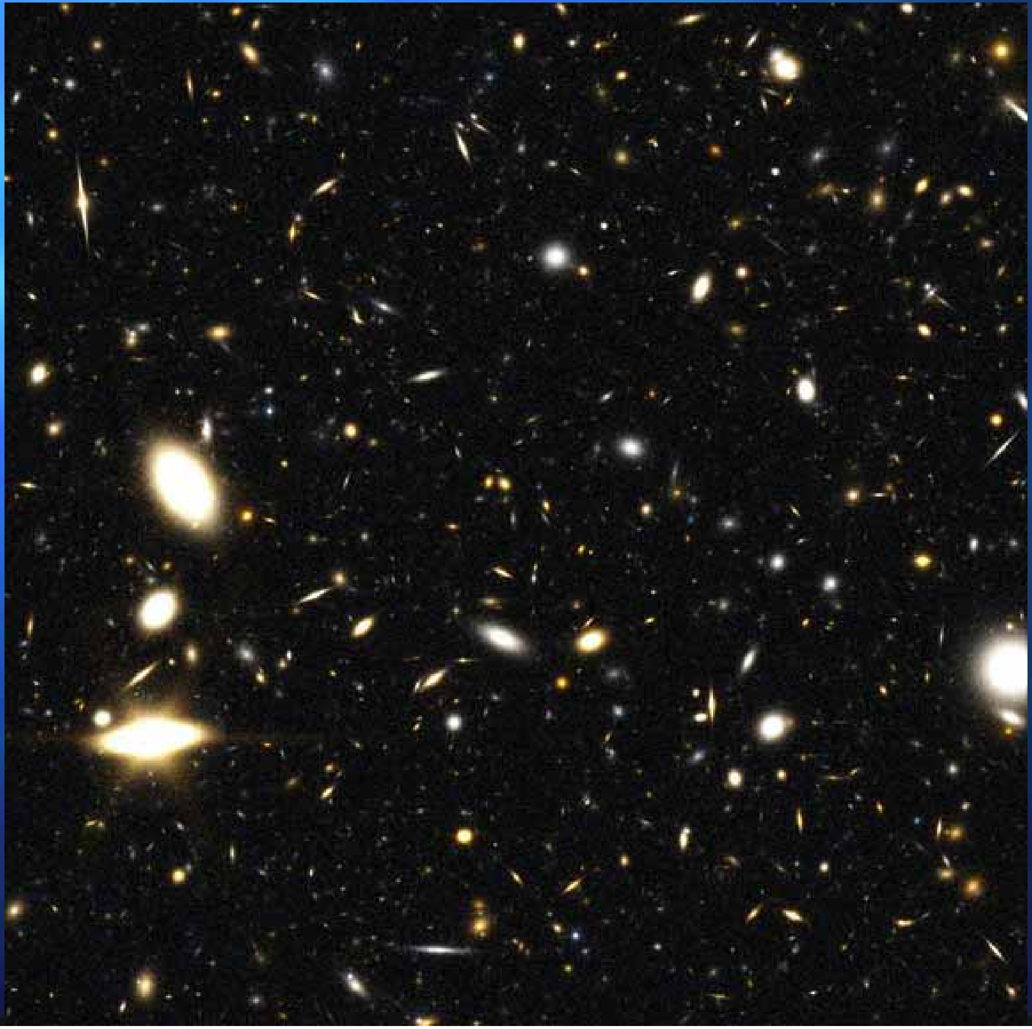
C_t - unknown *connected* fact
D_t - unknown *disconnected*
facts

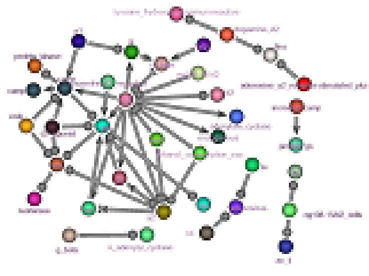
$$p_C(t)=\frac{\theta\cdot C_t}{\theta\cdot C_t+D_t},\text{ and }\ p_D(t)=\frac{D_t}{\theta\cdot C_t+D_t}$$

When θ is set to 1, the choice is random

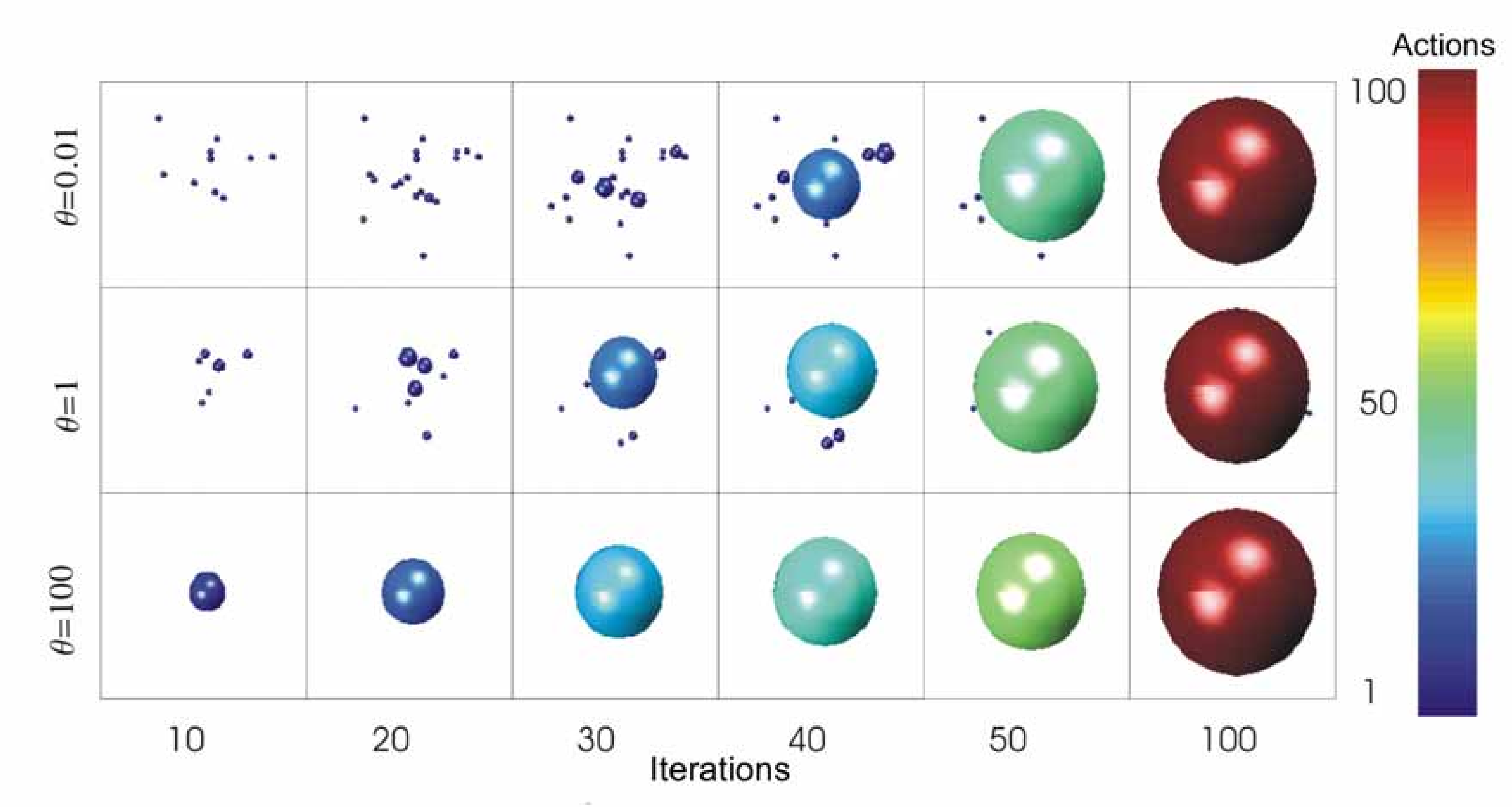


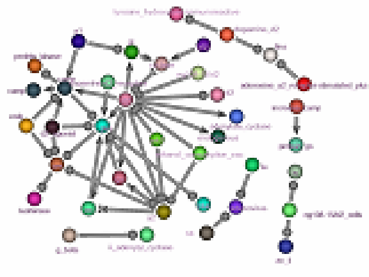
Each of the 3 models allows for multiple possible “universes”





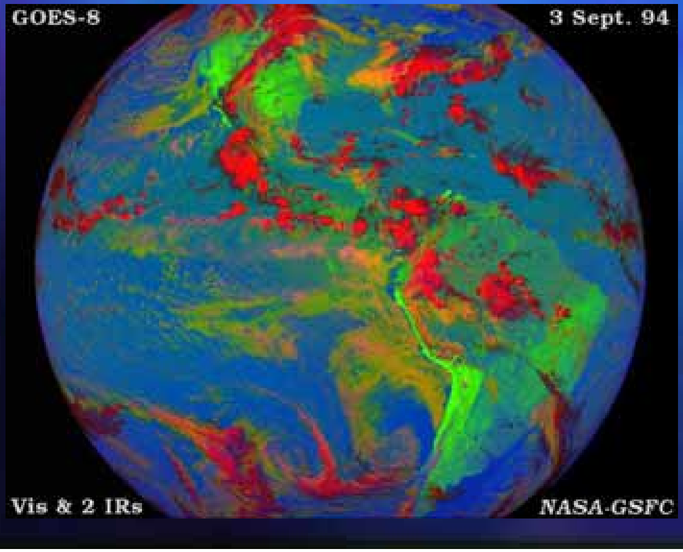
Low to high crawliness (theta) (=High to low jumpiness)

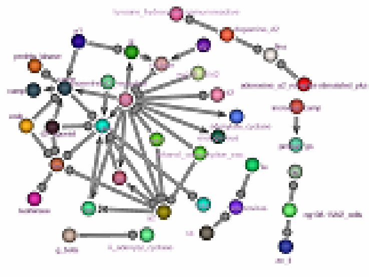




Our universe...

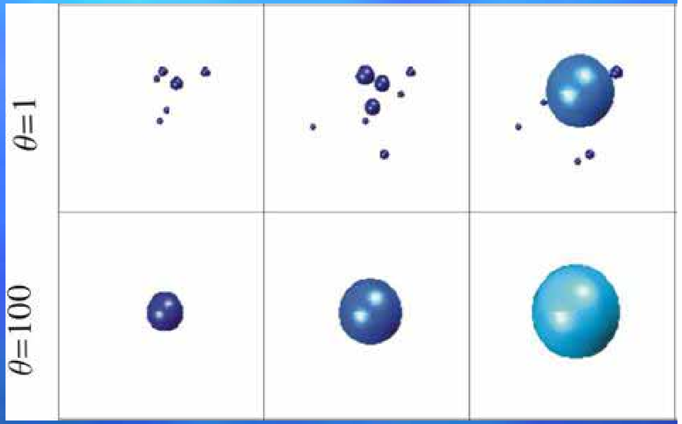
For the real interaction data from biological journals, the value of θ is found to be 5 (95% CI = [4.96, 5.04]). Hence, some jumps occur (but rare!).

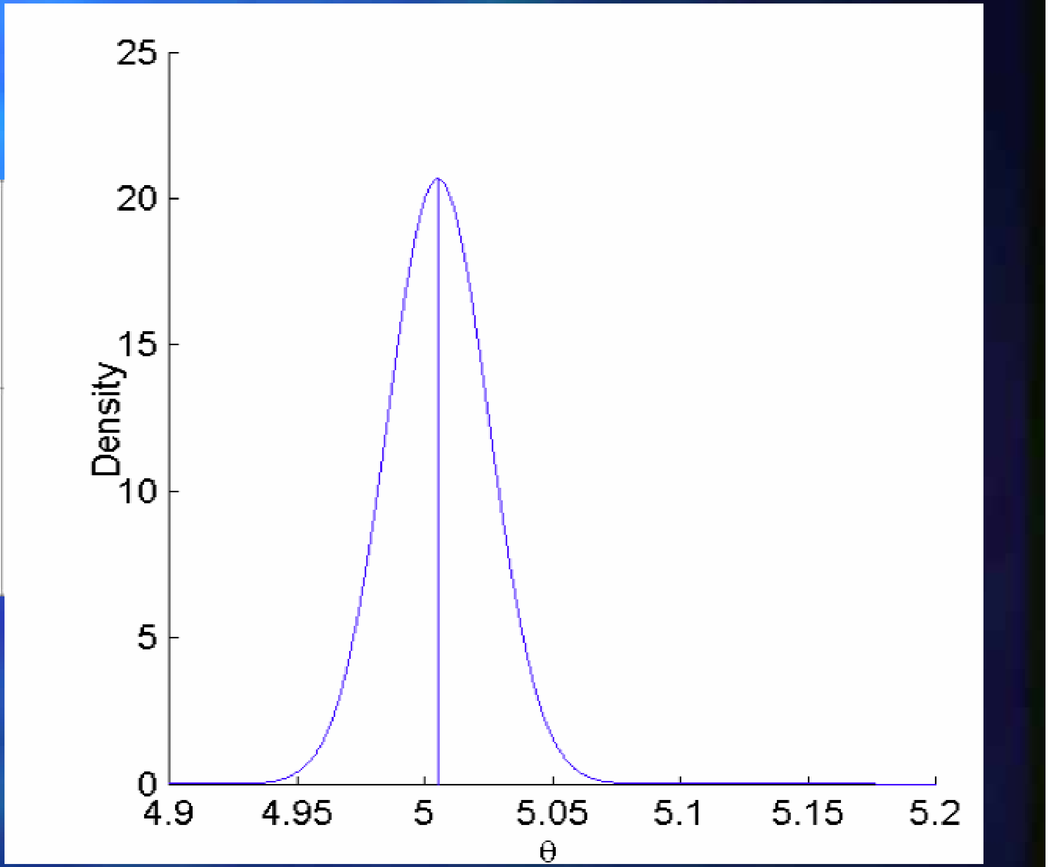




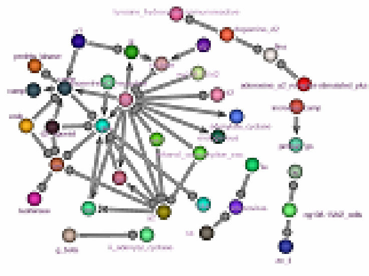
θ Estimated with MCMC:

- Estimated θ ~ 5

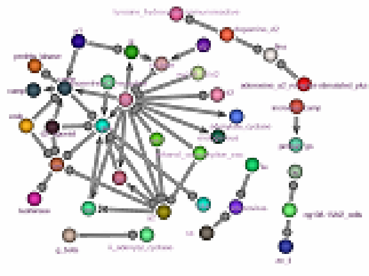




- Mostly crawling!
- But with occasional jumps...

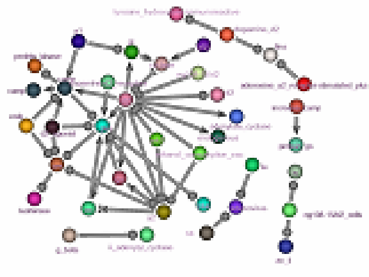


Jump or crawl?



Brain coral analogy

- Mode of knowledge growth
- Surface versus inside
- Knowledge pockets/involutions on the surface
- Coral volume

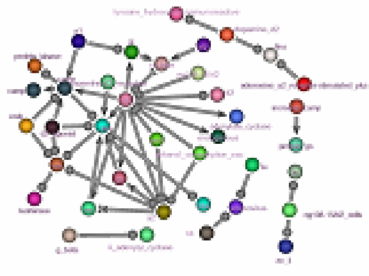


Awareness of Scientists about Potentially Relevant Scientific Results

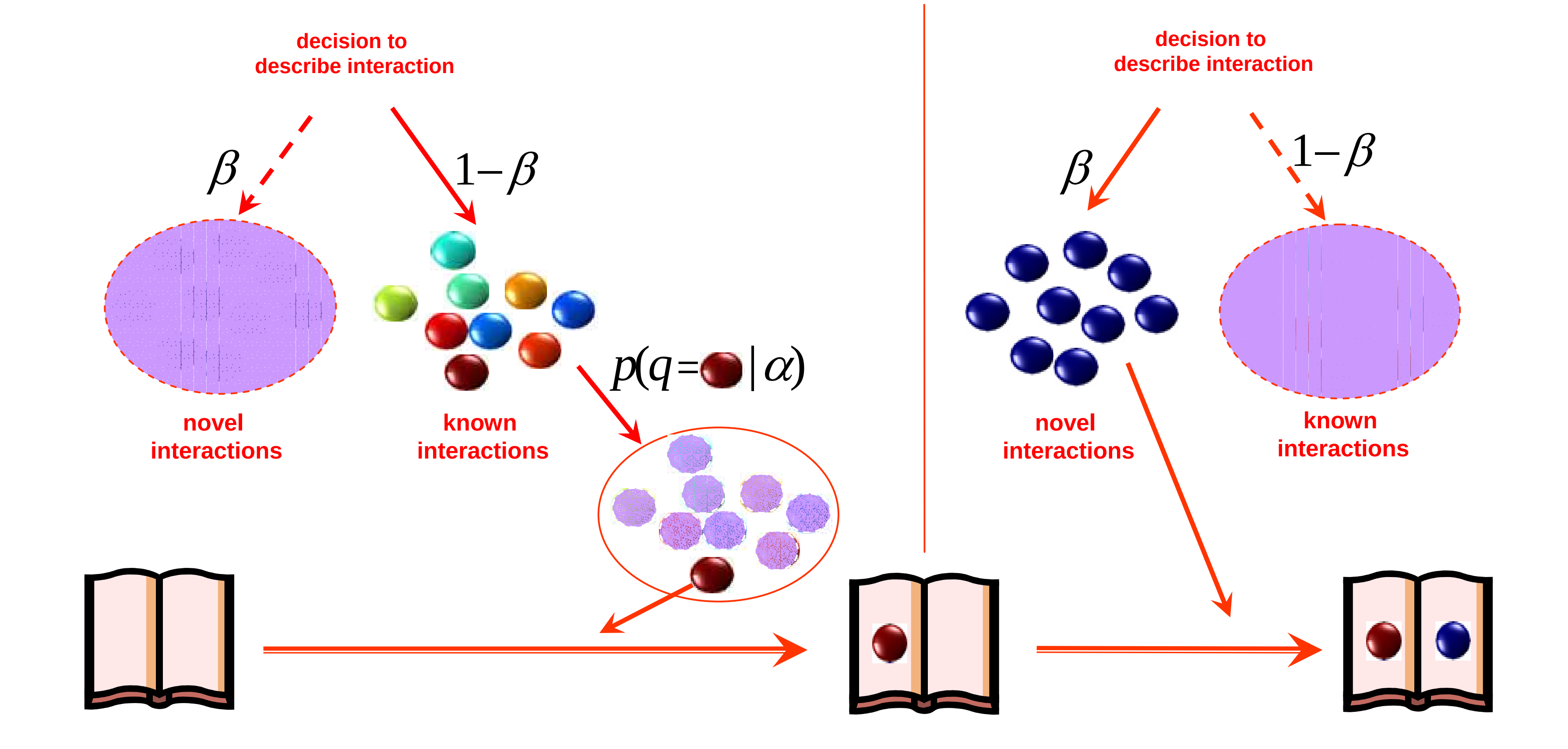
q (popularity) : number of times an interaction was mentioned in the literature

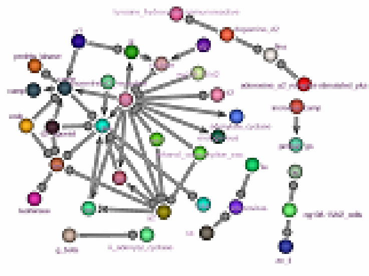
α (temperature) : tendency to include popular interactions in a journal. A value of 1 means no bias towards more or less popular interactions

β (novelty) : tendency to include new interactions in a paper
β = 1 means that all interactions are novel

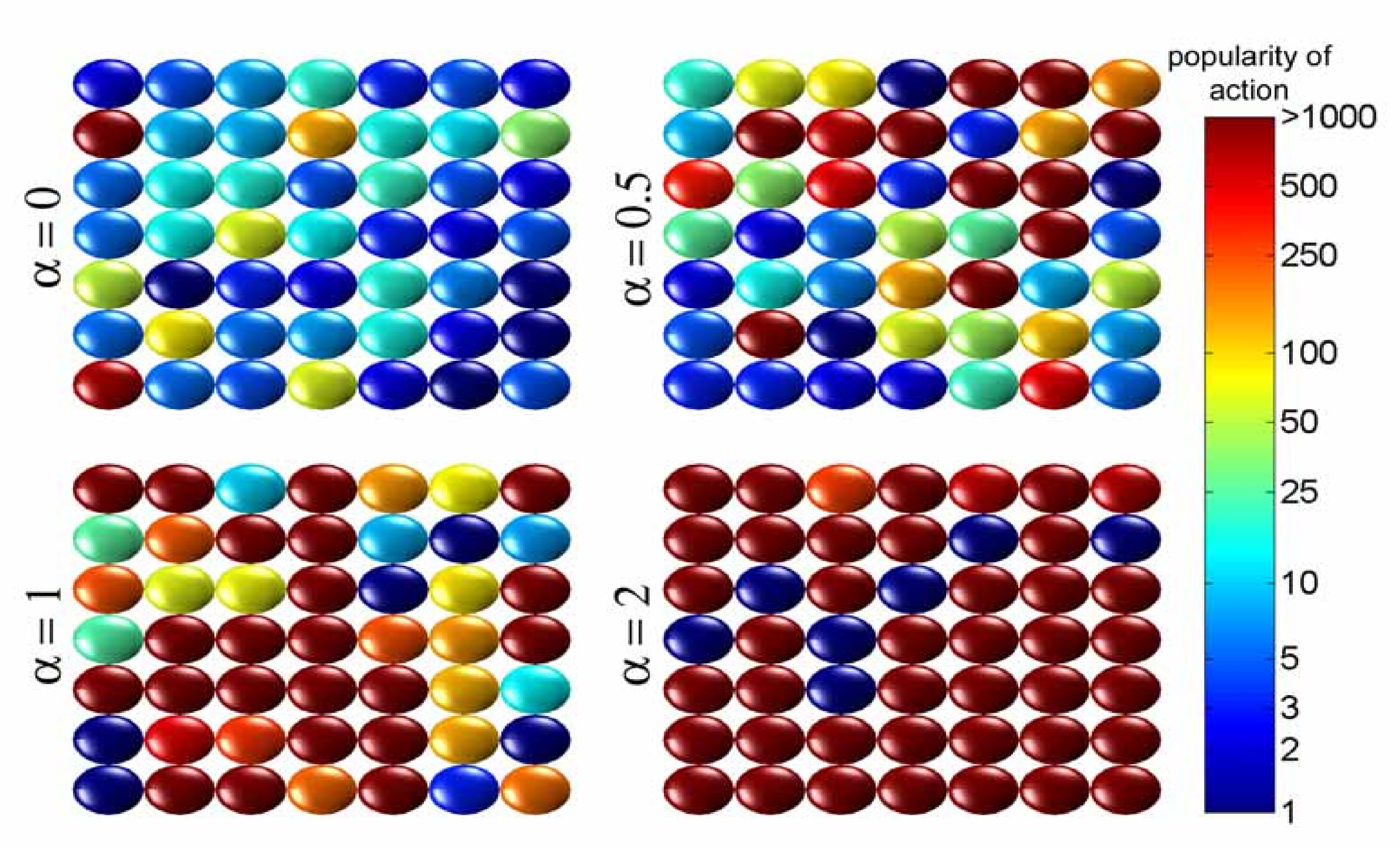


Model for generating a manuscript

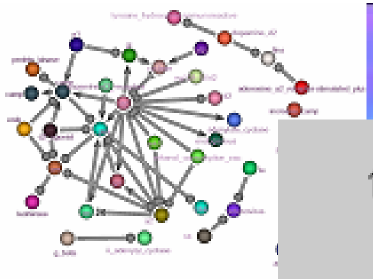




4 imaginary papers from different universes (journals)



Temperature vs. Novelty



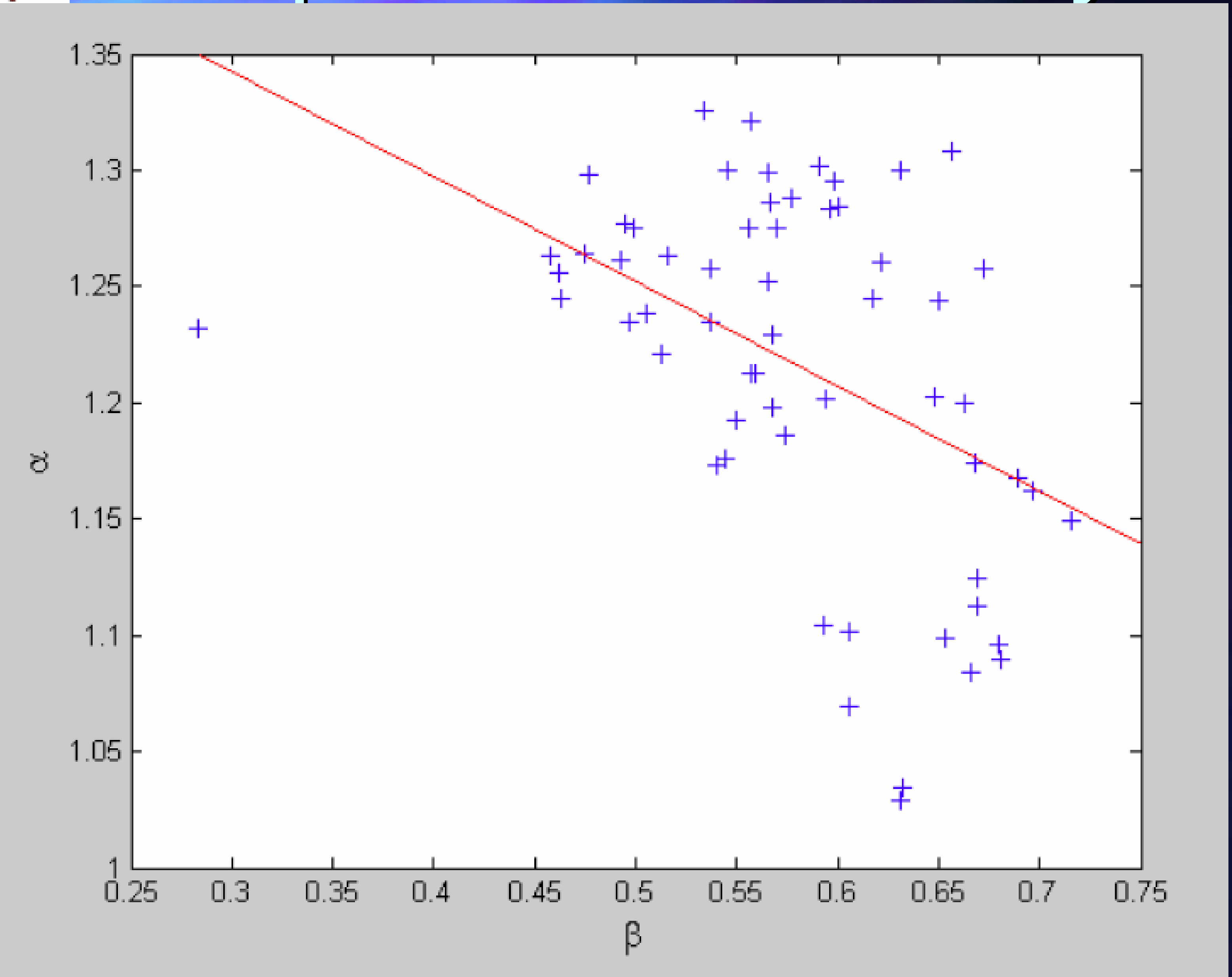
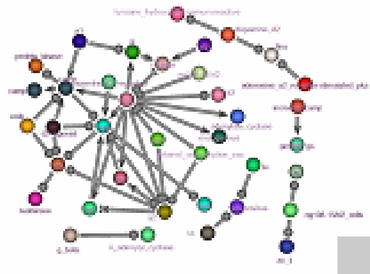


Figure 9: The scatter plot and linear regression function for the journal-specific estimates of α and β. Correlation = -0.43, p-value = 0.00027.



Impact Factor vs. Temperature

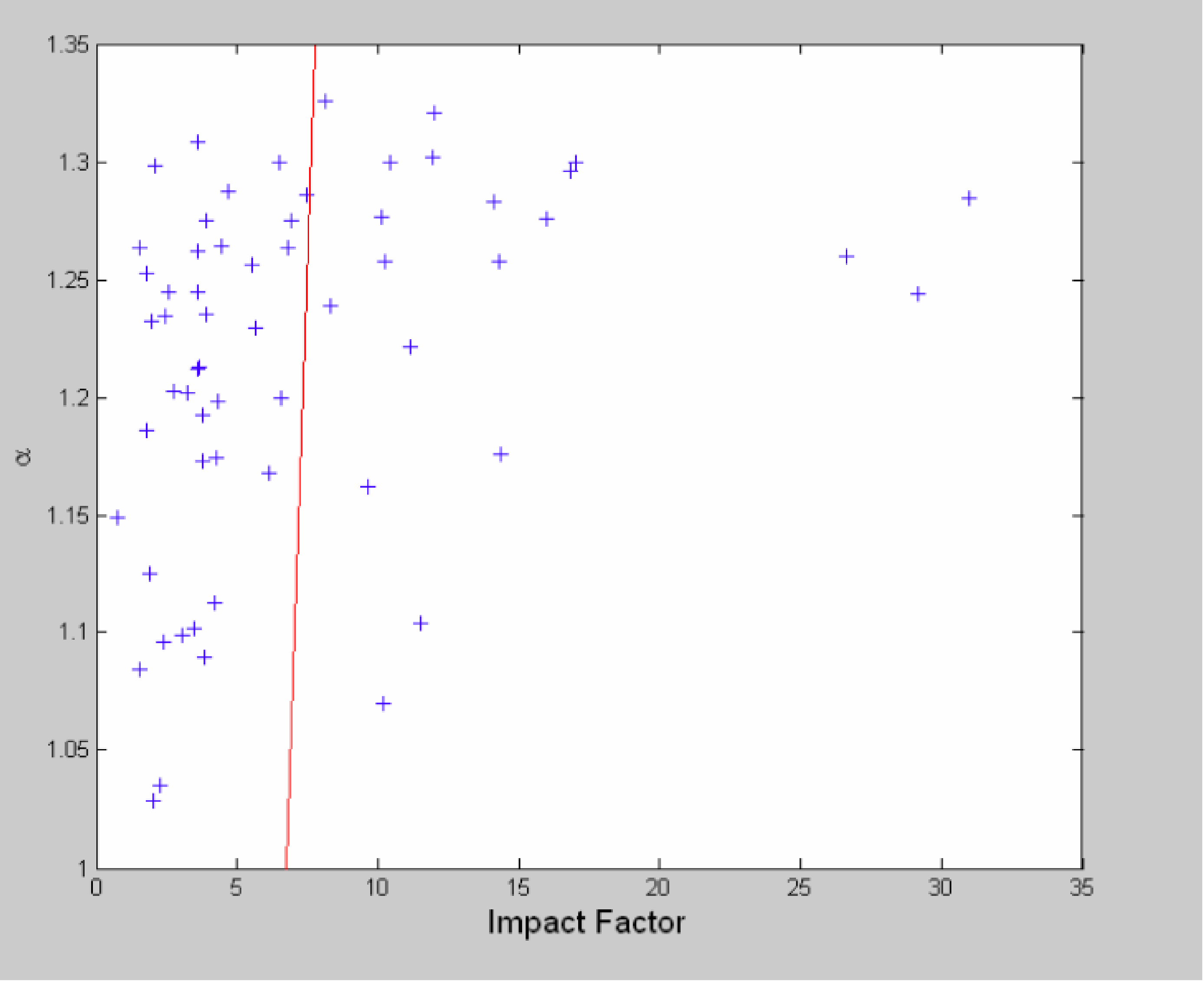
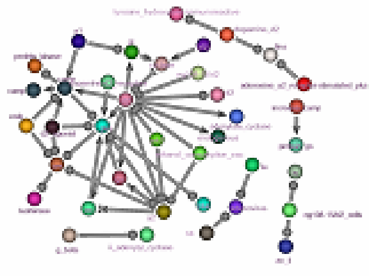


Figure 10: Scatter plot and linear regression for α and IF values for 60 journals. There is a significant linear correlation with correlation coefficient 0.35 and p-value 0.0063.



Impact Factor vs. Novelty

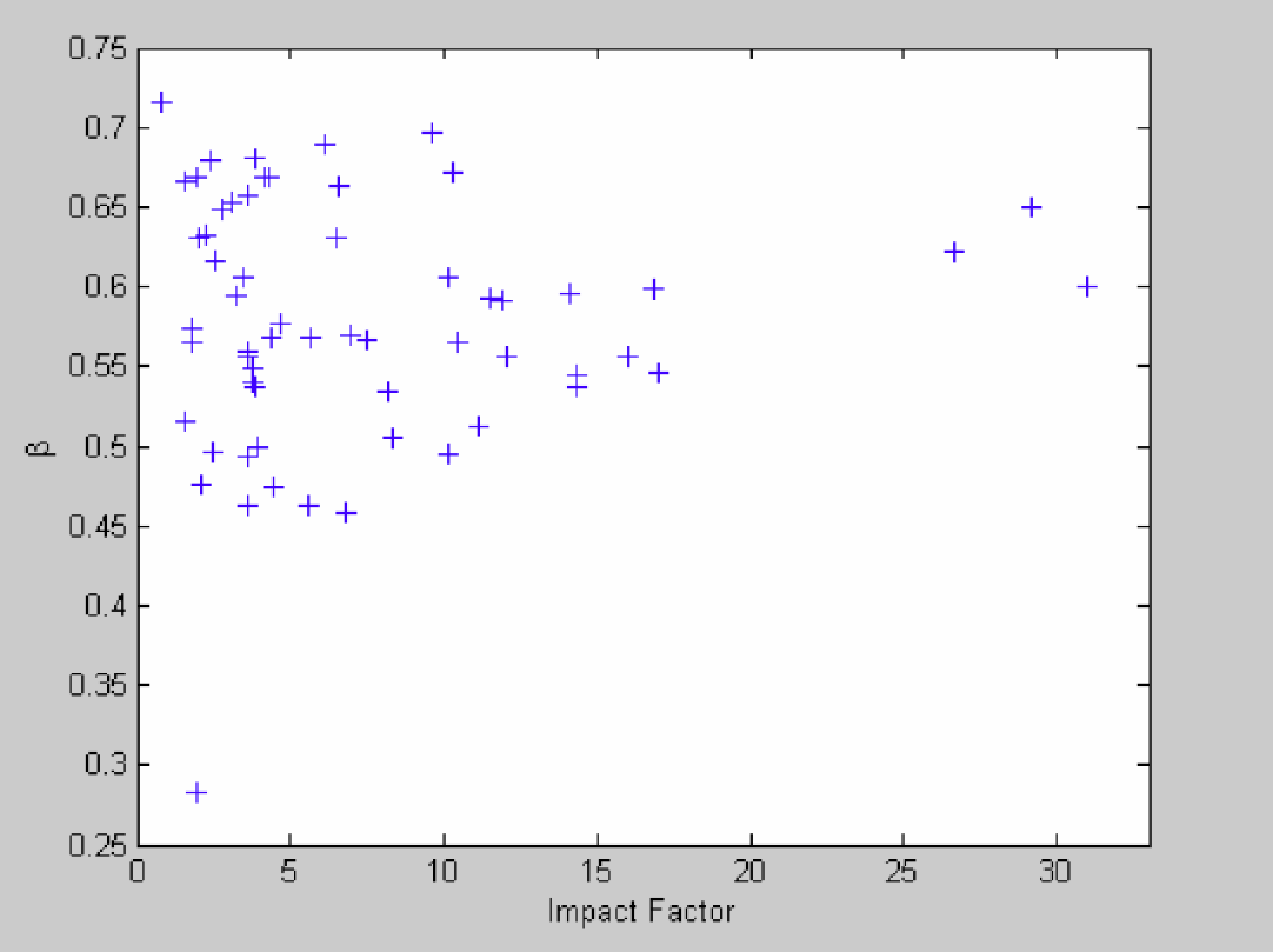
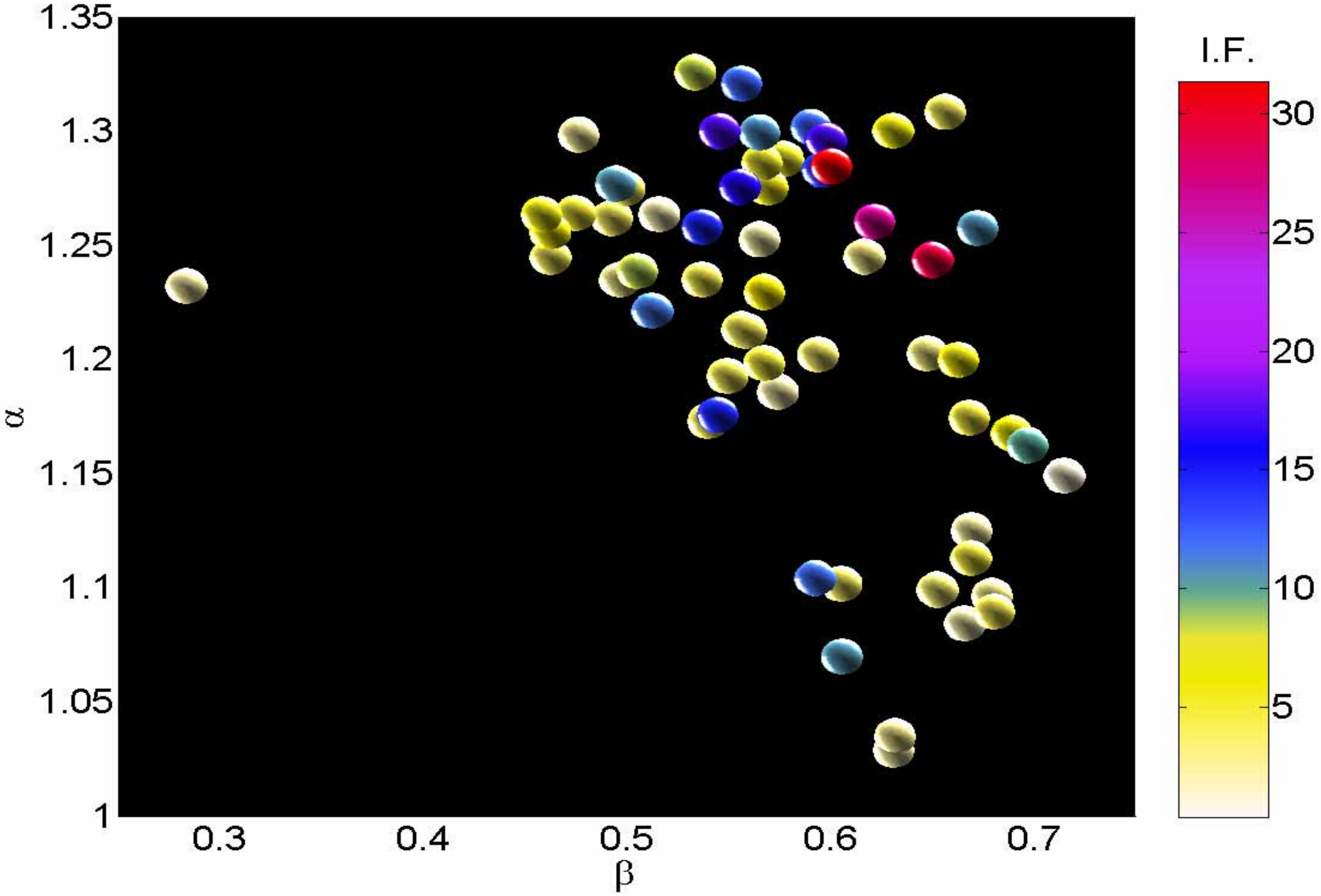
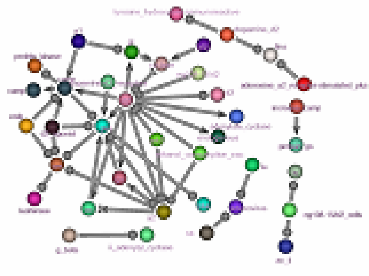


Figure 11: Scatter plot of β and *IF* values for 60 journals. There is no significant linear correlation (correlation coefficient 0.06, *p*-value 0.61).

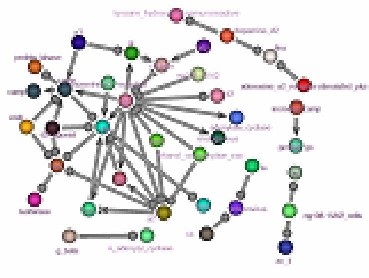
Impact Factor (IF) versus temperature (alpha) and novelty (beta)



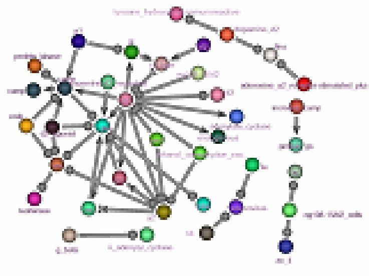


"Winning combination" for a paper in a high-impact journal:

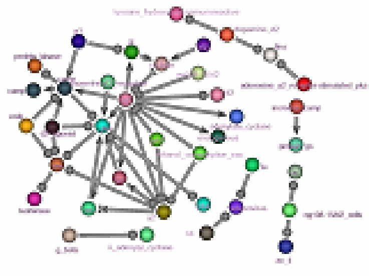
a very high temperature + at least a moderate degree of novelty



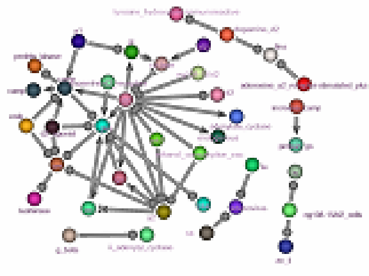
Temperature (popularity) is more important!



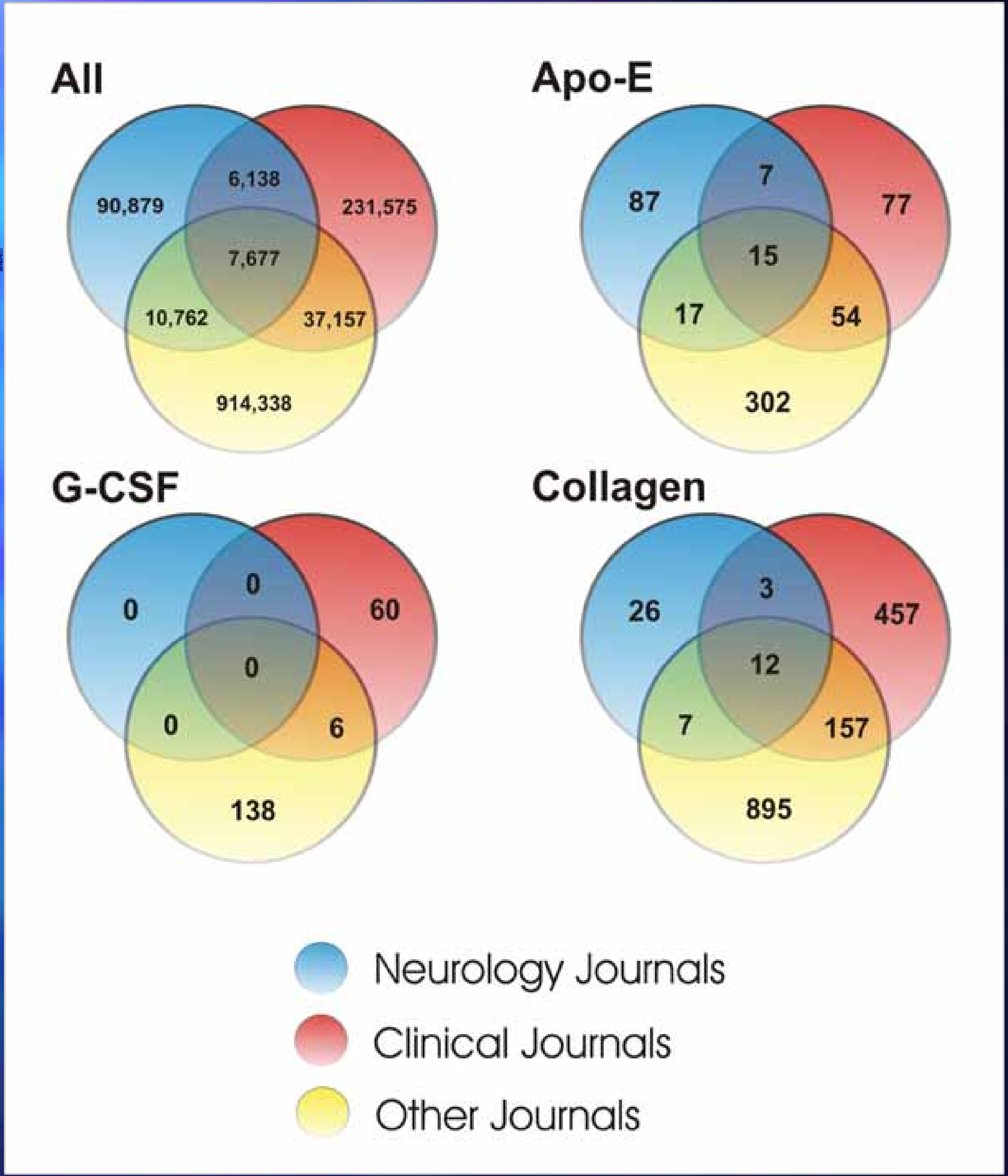
If we are trying to maximize “temperature” + novelty of publications, why journals are only slightly warm on average?

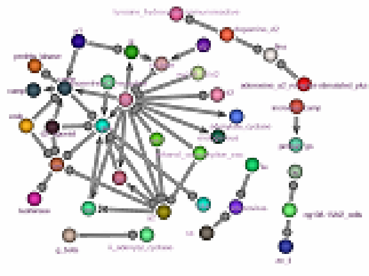


Because of the knowledge pockets! (We think so...)



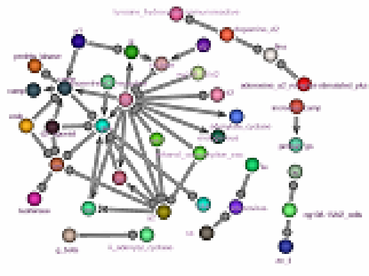
Real pockets



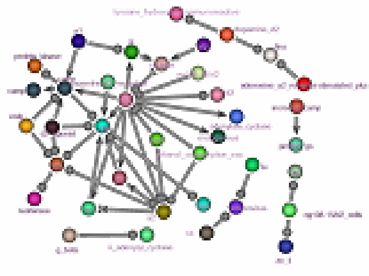


Brain coral analogy

- Mode of knowledge growth
- Surface versus inside
- Knowledge pockets/involutions on the surface
- Coral volume



In a brain coral only surface is growing and alive--inside is dead



Estimating the number of useful interactions that are “out there” (in the center of the coral)

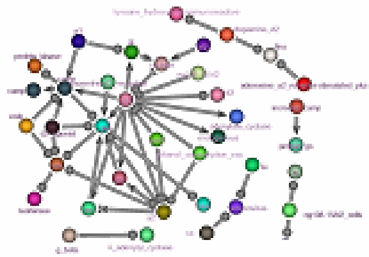
$$y = A \cdot x^{B} + error_{x}$$

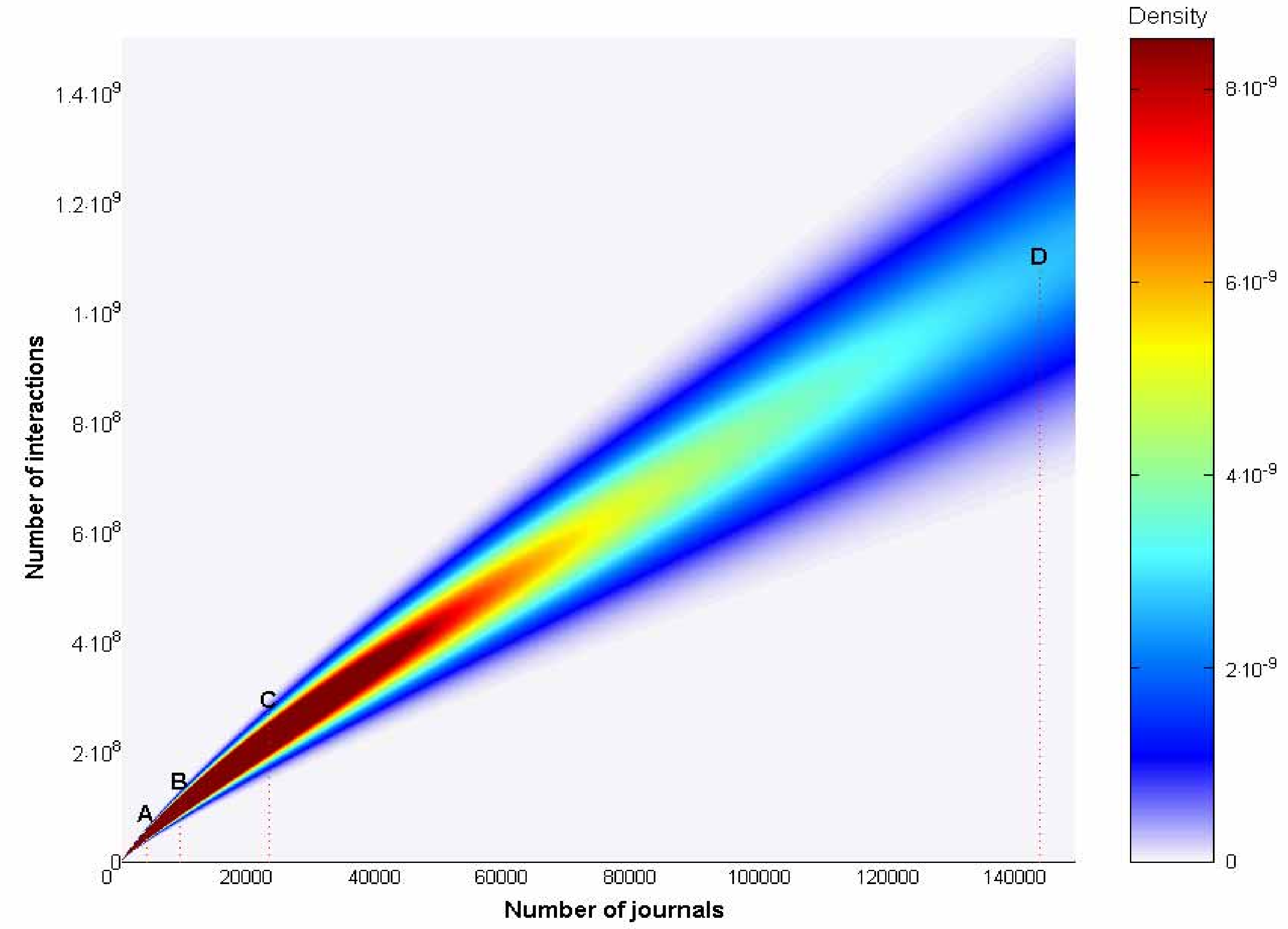
y is the number of facts extracted

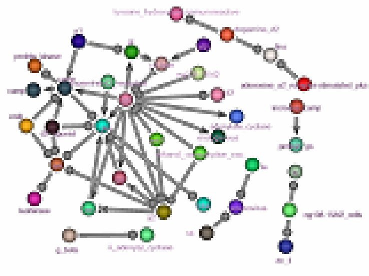
x is the number of average “journals” analyzed

A and *B* are parameters

$error_{x}$ is a normally distributed error term with variance growing proportionally to square of *x*

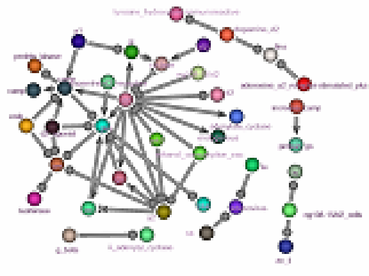



Density
8·10⁻⁹
6·10⁻⁹
4·10⁻⁹
2·10⁻⁹
0
Number of interactions
1.4·10⁹
1.2·10⁹
1·10⁹
8·10⁸
6·10⁸
4·10⁸
2·10⁸
0
A
B
C
D
0
20000
40000
60000
80000
100000
120000
140000
Number of journals



We conclude that ...

- There is mostly crawling
- Mostly “surface” is growing
- There are “knowledge pockets”
- The total volume of information is enormous compared to the living surface



I hope to cover...

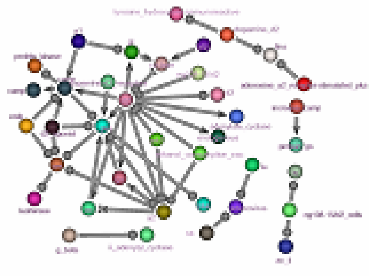
Complex traits 4

GeneWays 1

Chains of reasoning 2

Knowledge as a coral 3

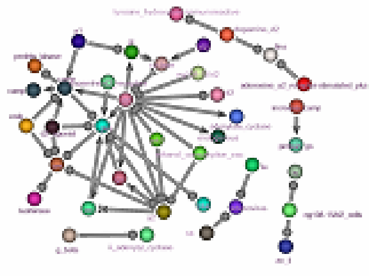
Application to analysis of complex disorders

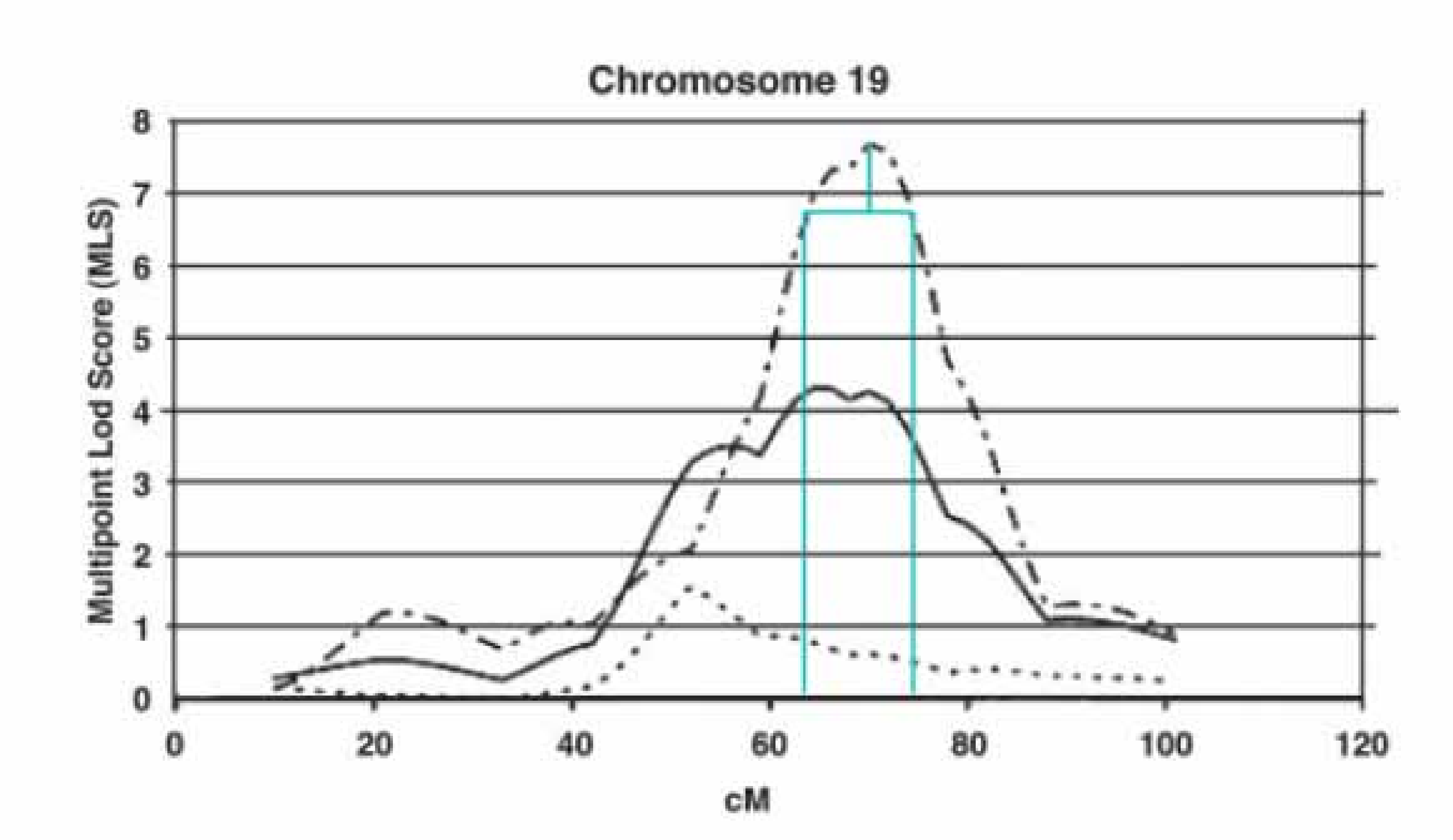


Goal: finding candidate genes for a complex trait

genetic linkage studies

LOD score (logarithmic odds)= log_{10}(likelihood under a linkage model/likelihood under no linkage)

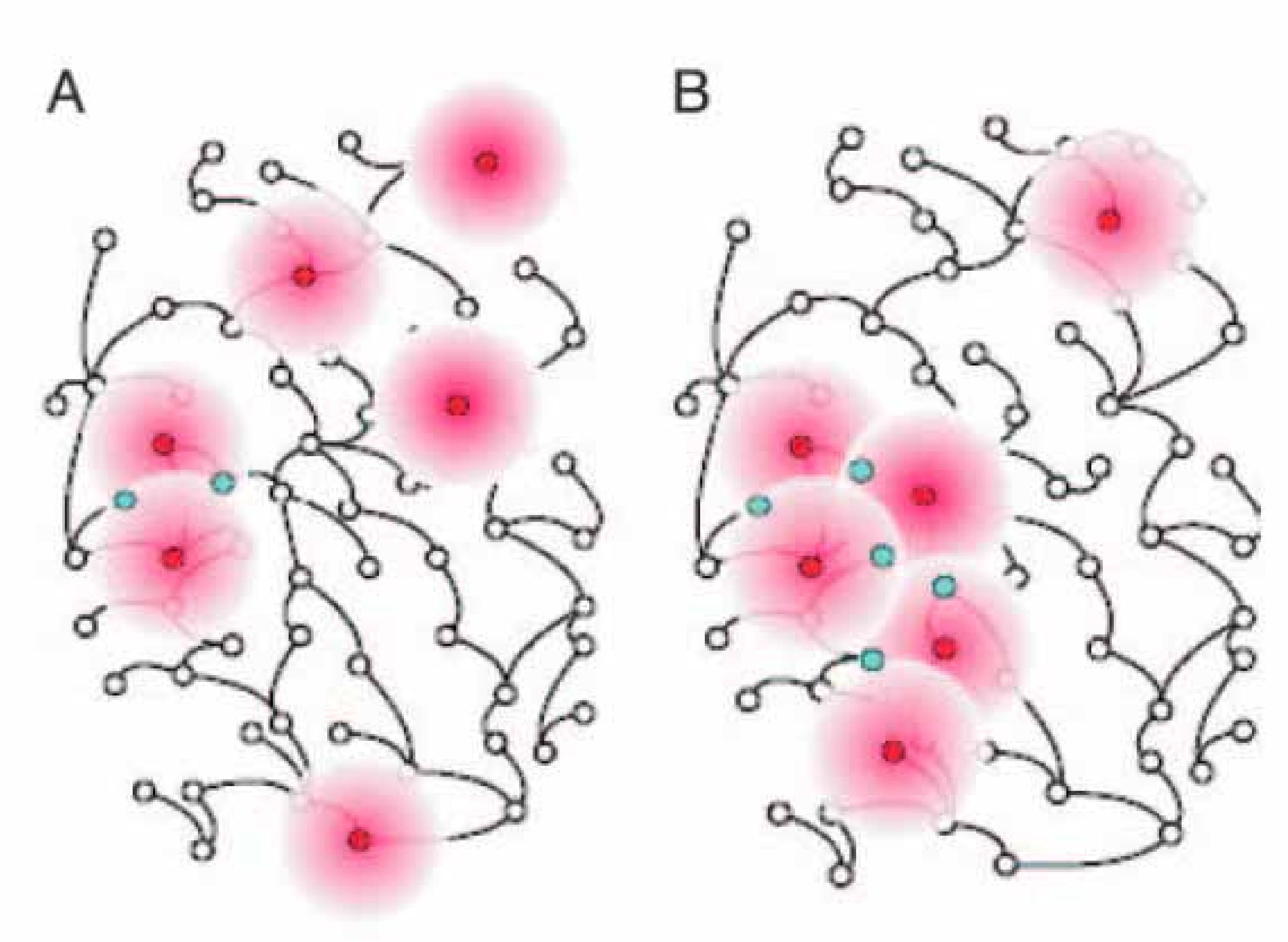


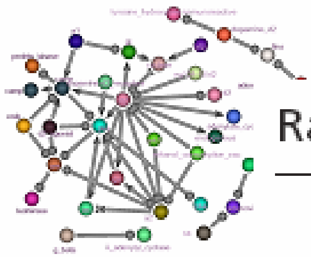

Chromosome 19
Multipoint Lod Score (MLS)
8
7
6
5
4
3
2
1
0
0
20
40
60
80
100
120
cM

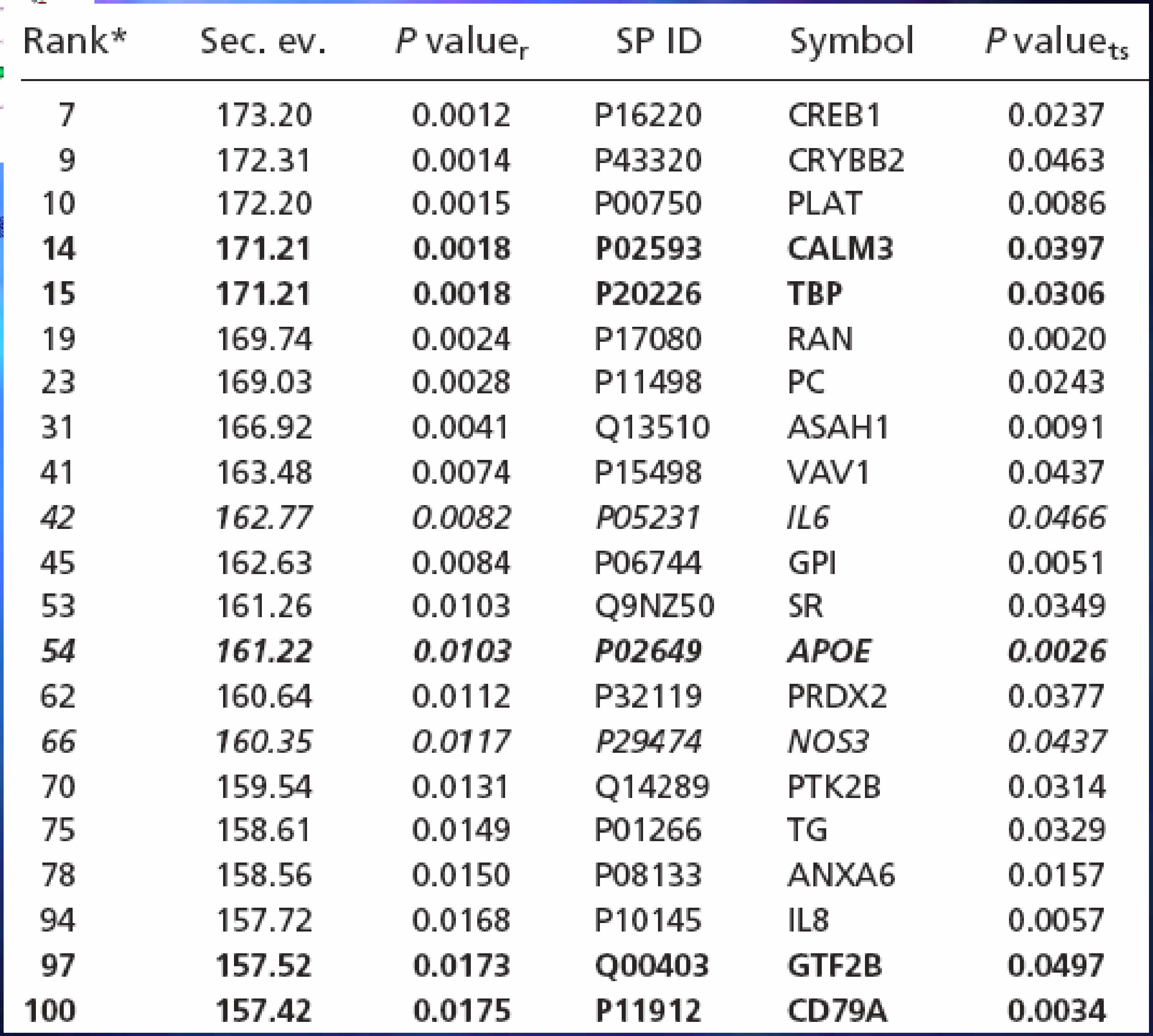
Assumptions

1. The functional molecular module is compact

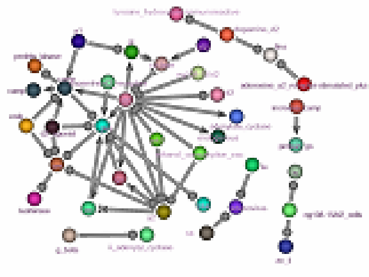
2. The noise is uniformly distributed over the network nodes

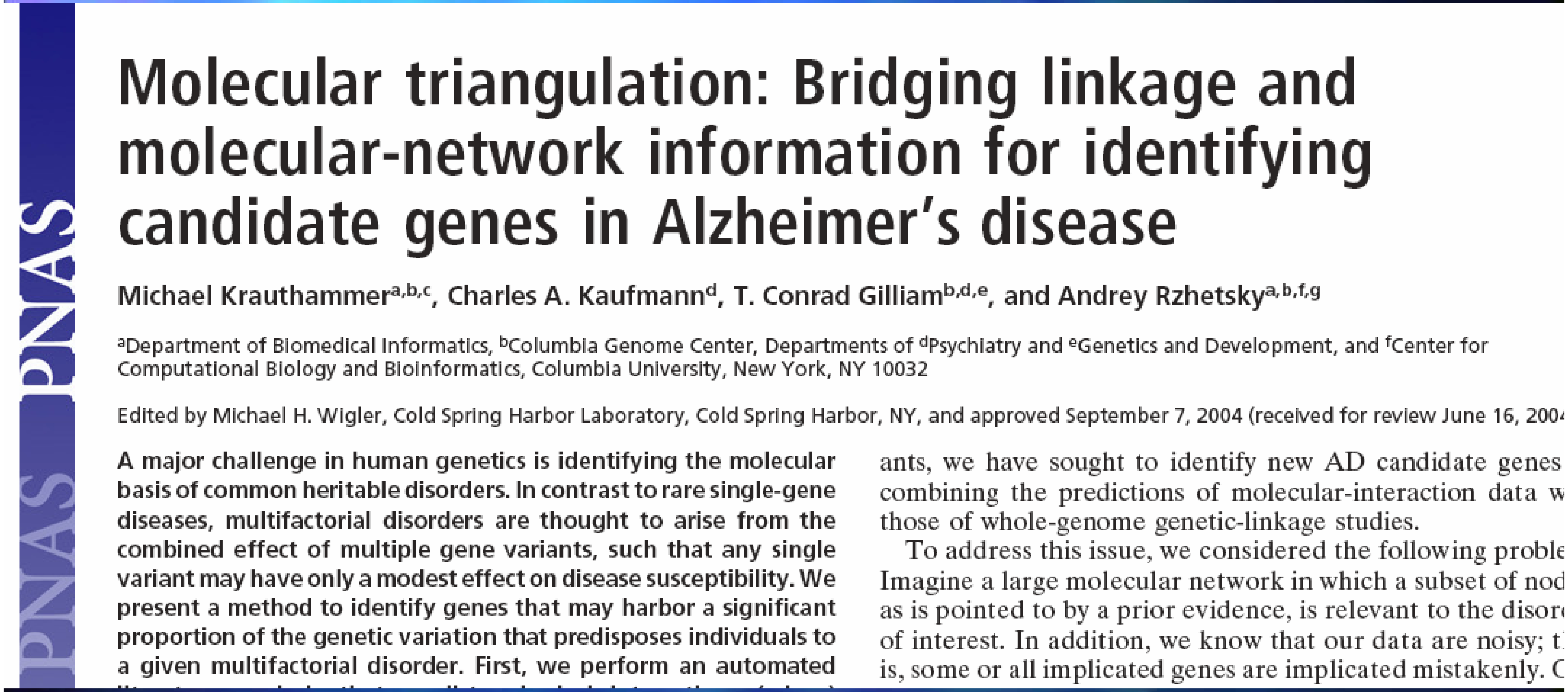

A
B





Rank*	Sec. ev.	P value$_r$	SP ID	Symbol	P value$_{ts}$
7	173.20	0.0012	P16220	CREB1	0.0237
9	172.31	0.0014	P43320	CRYBB2	0.0463
10	172.20	0.0015	P00750	PLAT	0.0086
14	**171.21**	**0.0018**	**P02593**	**CALM3**	**0.0397**
15	**171.21**	**0.0018**	**P20226**	**TBP**	**0.0306**
19	169.74	0.0024	P17080	RAN	0.0020
23	169.03	0.0028	P11498	PC	0.0243
31	166.92	0.0041	Q13510	ASAH1	0.0091
41	163.48	0.0074	P15498	VAV1	0.0437
42	*162.77*	*0.0082*	*P05231*	*IL6*	*0.0466*
45	162.63	0.0084	P06744	GPI	0.0051
53	161.26	0.0103	Q9NZ50	SR	0.0349
54	***161.22***	***0.0103***	***P02649***	***APOE***	***0.0026***
62	160.64	0.0112	P32119	PRDX2	0.0377
66	*160.35*	*0.0117*	*P29474*	*NOS3*	*0.0437*
70	159.54	0.0131	Q14289	PTK2B	0.0314
75	158.61	0.0149	P01266	TG	0.0329
78	158.56	0.0150	P08133	ANXA6	0.0157
94	157.72	0.0168	P10145	IL8	0.0057
97	**157.52**	**0.0173**	**Q00403**	**GTF2B**	**0.0497**
100	**157.42**	**0.0175**	**P11912**	**CD79A**	**0.0034**





PNAS PNAS

Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer’s disease

Michael Krauthammer[a,b,c], Charles A. Kaufmann[d], T. Conrad Gilliam[b,d,e], and Andrey Rzhetsky[a,b,f,g]

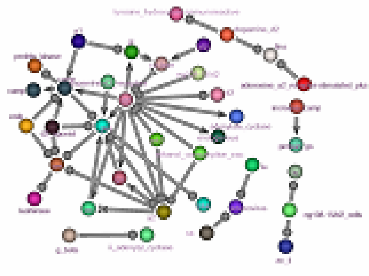
[a]Department of Biomedical Informatics, [b]Columbia Genome Center, Departments of [d]Psychiatry and [e]Genetics and Development, and [f]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032

Edited by Michael H. Wigler, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, and approved September 7, 2004 (received for review June 16, 200

A major challenge in human genetics is identifying the molecular basis of common heritable disorders. In contrast to rare single-gene diseases, multifactorial disorders are thought to arise from the combined effect of multiple gene variants, such that any single variant may have only a modest effect on disease susceptibility. We present a method to identify genes that may harbor a significant proportion of the genetic variation that predisposes individuals to a given multifactorial disorder. First, we perform an automated

ants, we have sought to identify new AD candidate genes combining the predictions of molecular-interaction data w those of whole-genome genetic-linkage studies.

To address this issue, we considered the following proble Imagine a large molecular network in which a subset of nod as is pointed to by a prior evidence, is relevant to the disor of interest. In addition, we know that our data are noisy; t is, some or all implicated genes are implicated mistakenly. C



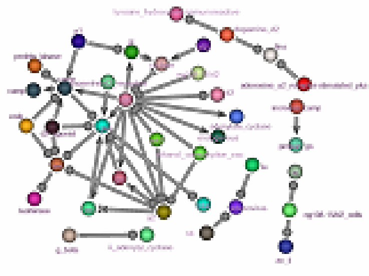
I hope to cover...

Complex traits 4

GeneWays 1

Knowledge as a coral 3

Chains of reasoning 2

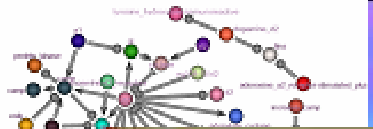


The army that produced these results...

GeneWays team

Professor Carol Friedman,
Co-PI
GENIES

Dr. Vasilis Hatzivassiloglou,
Co-PI
Semantic pattern discovery,
Sense disambiguation

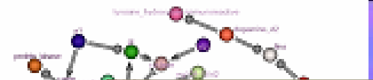


Dr. Pauline Kra
GENIES

Dr. Michael O. Krauthammer
Term recognition,
Data cleansing
(Noisy truth generator)

Mr. Ivan Iossifov
All issues related
to GeneWays

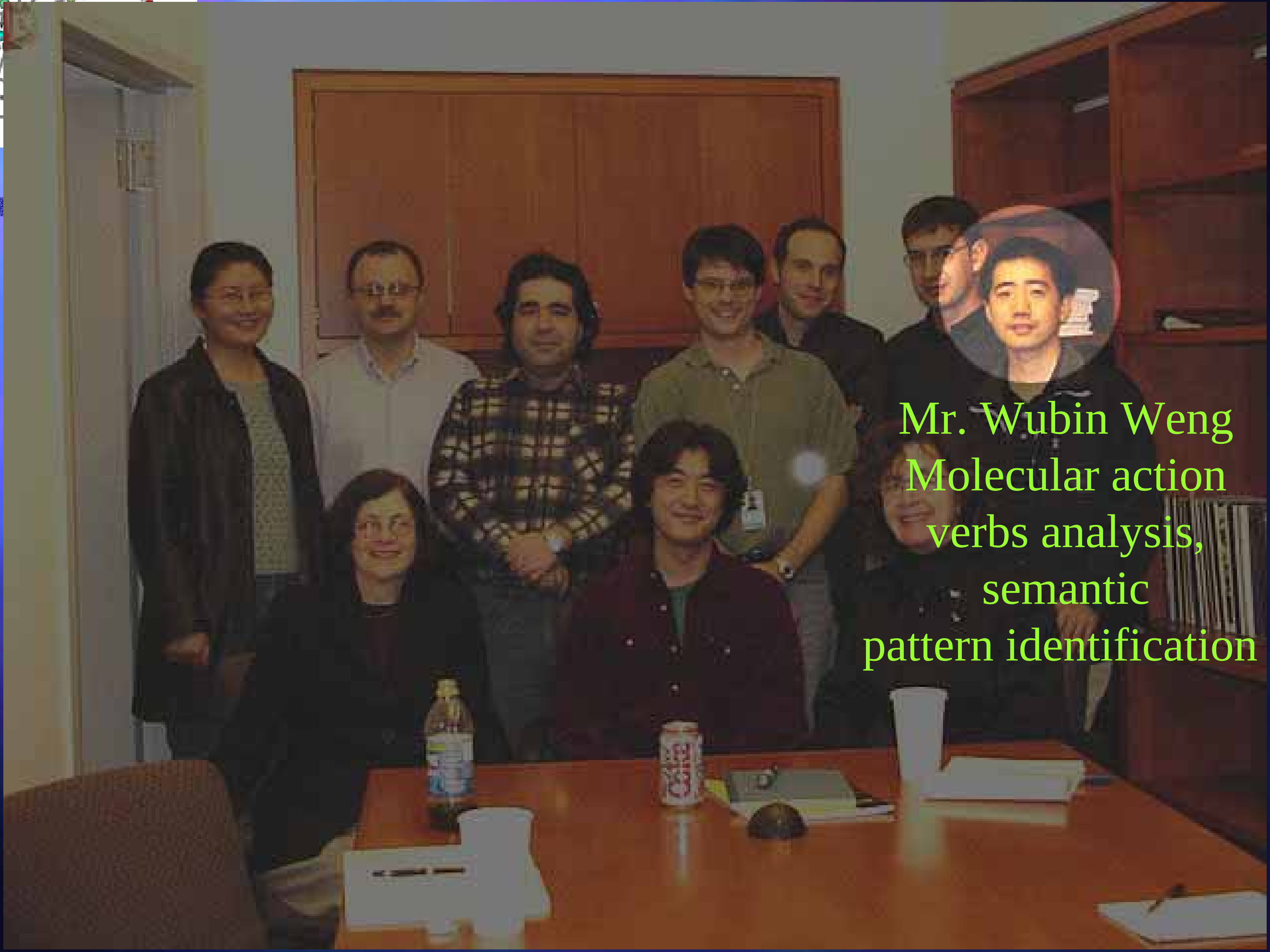
Dr. Hong Yu
Synonym/homonym resolution
Abbreviation disambiguation



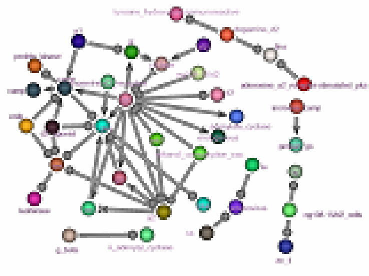
Dr. Shawn M. Gomez
Protein-protein
interaction prediction

Mr. Tomohiro Koike

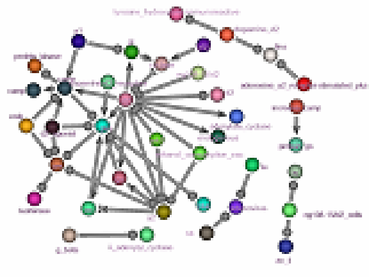
CUtenet



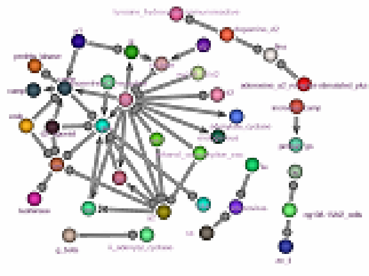
Mr. Wubin Weng
Molecular action
verbs analysis,
semantic
pattern identification



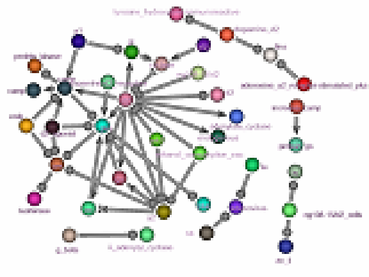
Murat Cokol



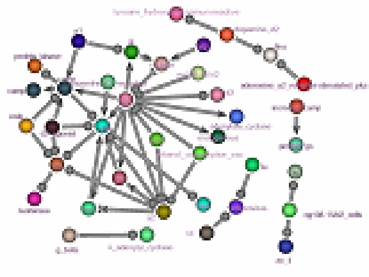
Chani
Weinreb



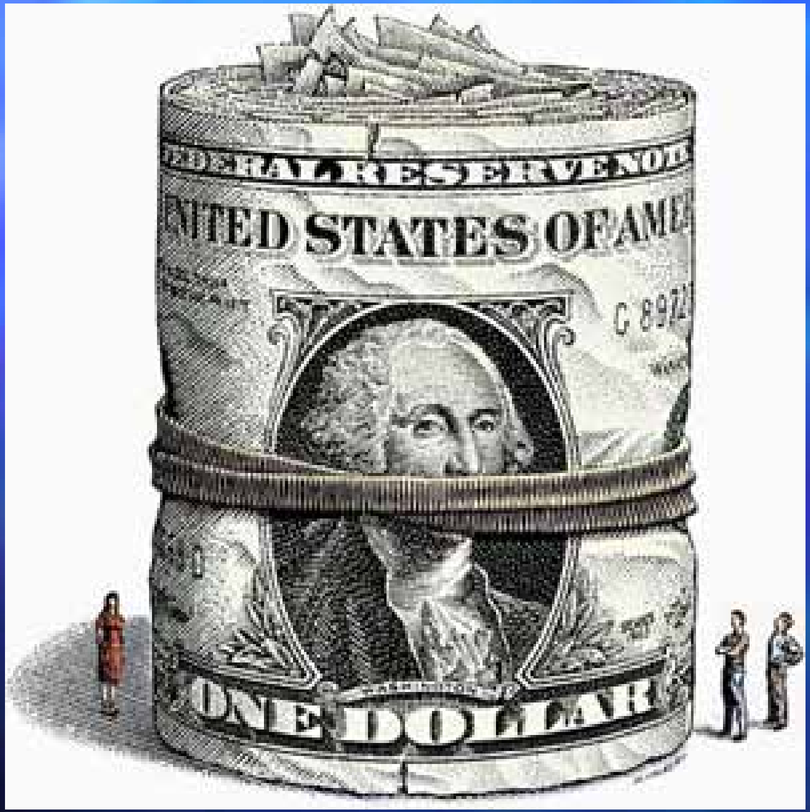
Mitzi Morris



Marc Hadfield



Financial support comes from



DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA

DARPA

NATIONAL INSTITUTES OF HEALTH

NSF

cure Autism now
FOUNDATION