Physics of the Web

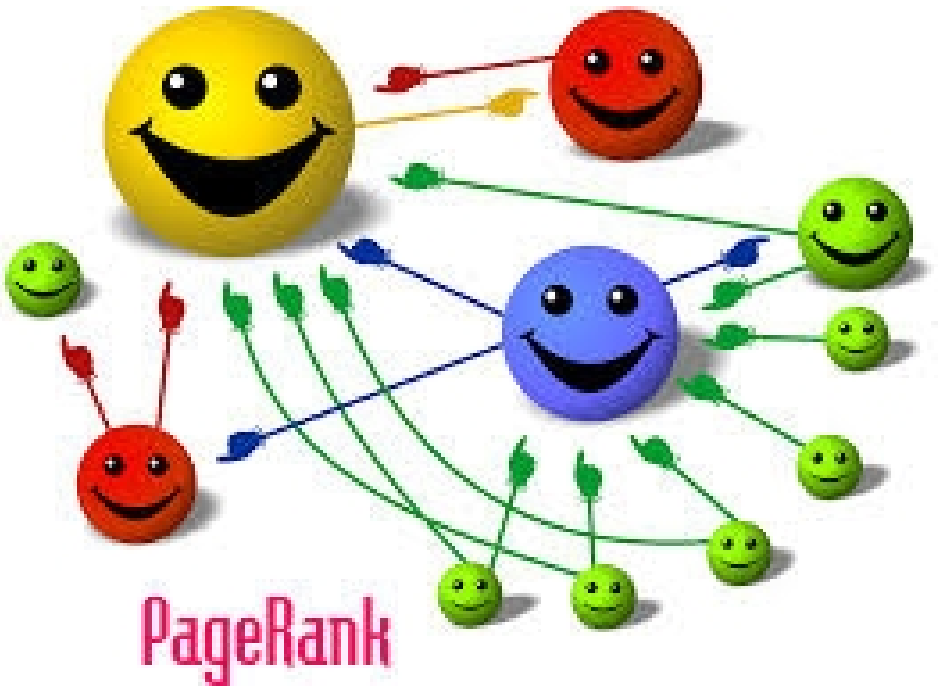
G Santhosh Kumar
Cochin University

Birthday of a Giant

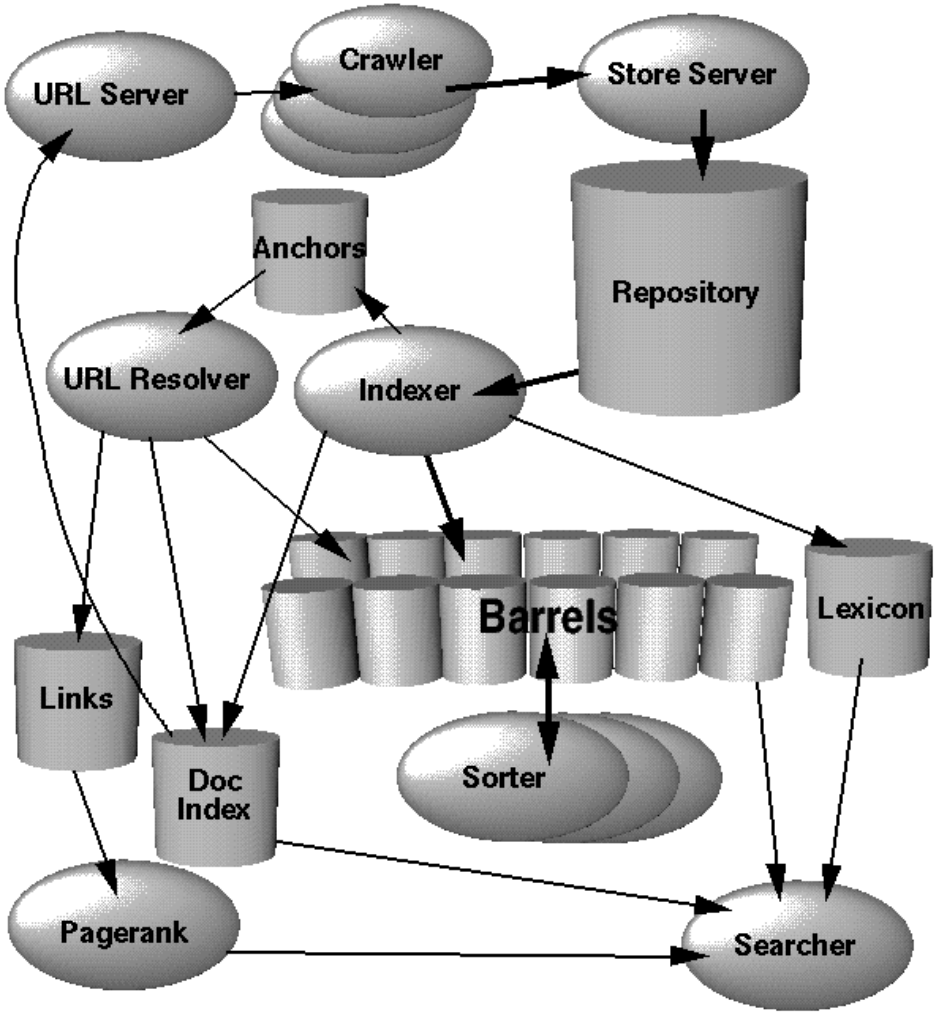
Whose slogan is this?

Stand on the shoulders of giant

Idea of PageRank



Anatomy of a Search Engine



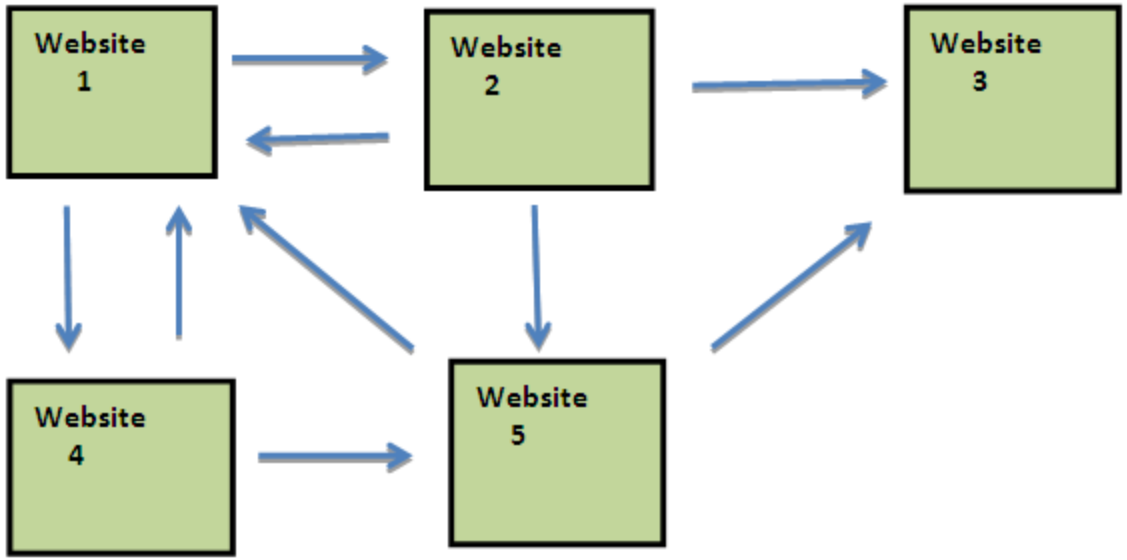
Source:
http://infolab.stanford.edu/~backrub/google.html

Random Surfer model

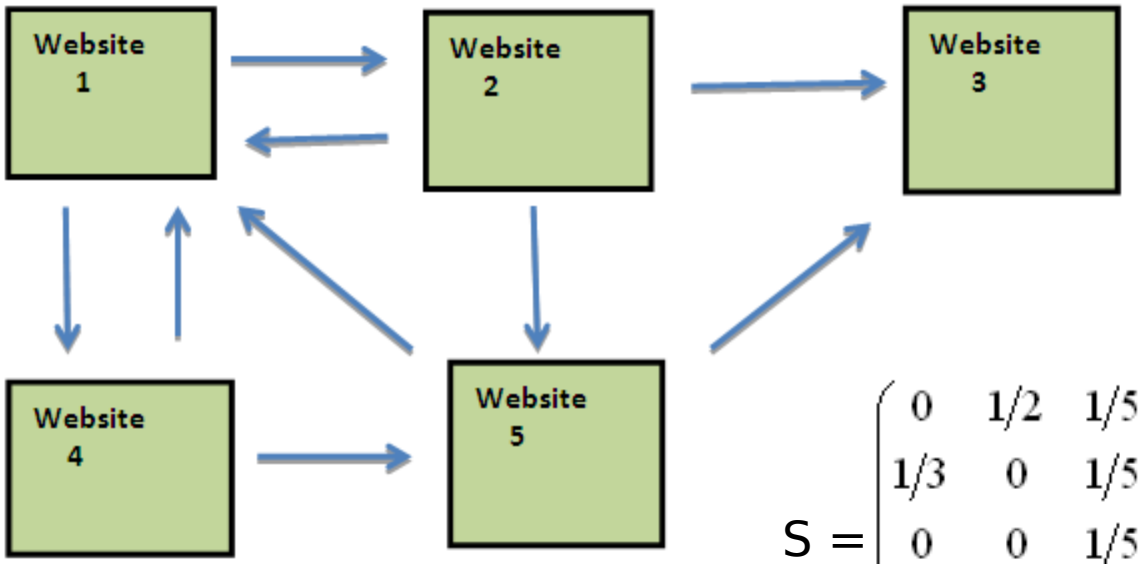
The random surfer visits a web page with a certain probability which derives from the page's PageRank. The probability that the random surfer clicks on one link is solely given by the number of links on that page

the probability for the random surfer reaching one page is the sum of probabilities for the random surfer following links to this page

World's largest Eigen value Problem



Idea is to compute the Principle Eigen vector of the system

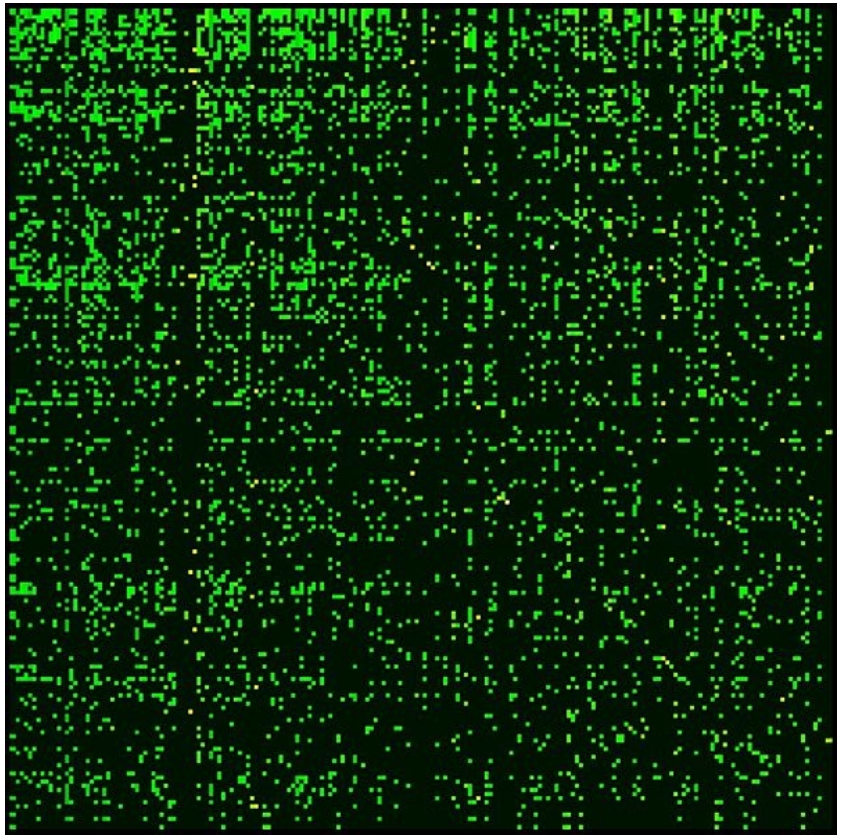


$$S = \begin{pmatrix} 0 & 1/2 & 1/5 & 1/3 & 0 \\ 1/3 & 0 & 1/5 & 1/3 & 0 \\ 0 & 0 & 1/5 & 0 & 1/2 \\ 1/3 & 0 & 1/5 & 0 & 1/2 \\ 1/3 & 1/2 & 1/5 & 1/3 & 0 \end{pmatrix}$$

Google Matrix $G_{ij} = \alpha S_{ij} + (1-\alpha)\frac{1}{N}$

Markov matrix S is irreducible and stochastic

The rank of each page can be generated iteratively
from the Google matrix using the power method



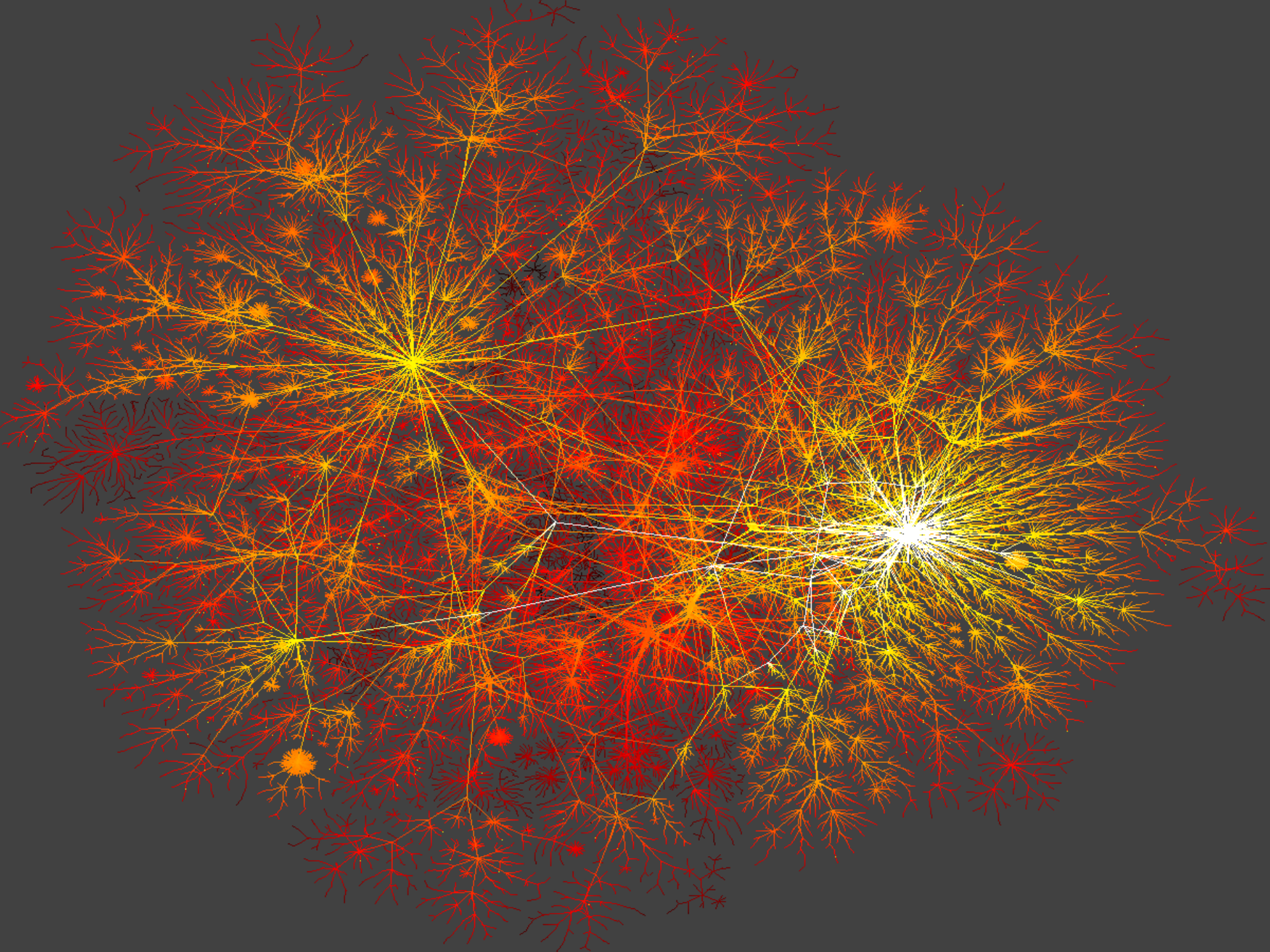
Google matrix of Wikipedia articles network, written in the bases of PageRank index; fragment of top 200 X 200 matrix elements is shown, total size N=3282257

People are interested in Spectrum and eigen states of *G* matrix

Towards Google matrix of ...

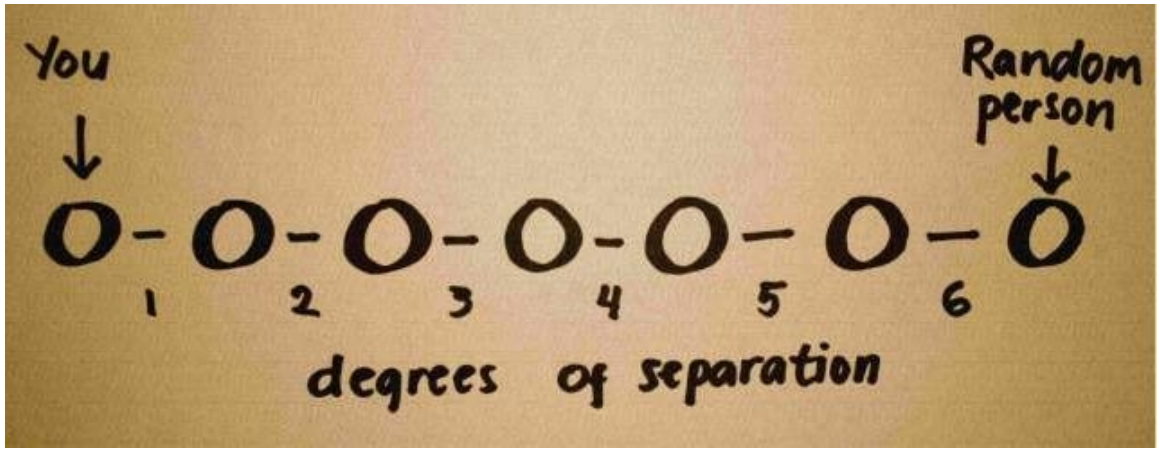
Brain: The Google matrix G is constructed on the basis of neuronal network of a brain model

DNA: Google Matrix Analysis of DNA Sequences



An old experiment

- Milgram in 1967



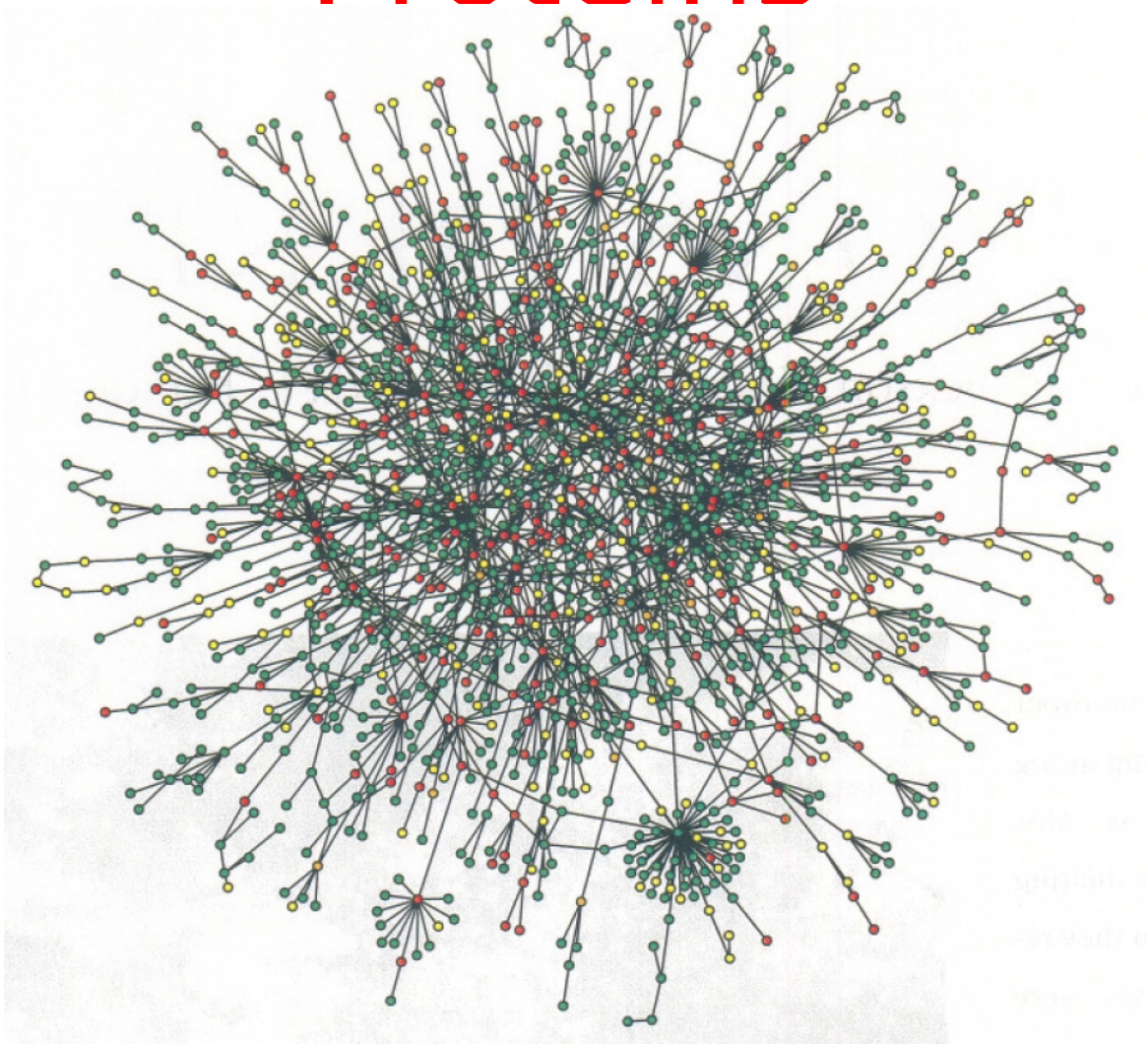
Any two strangers in the world are separated by an average of six

In 2008, a study by Microsoft showed that the average chain of contacts between users of its Messenger Service was 6.6 people

It’s Small World, after all

small diameter of the web means that all that information is just a few clicks away

Map of interacting Proteins

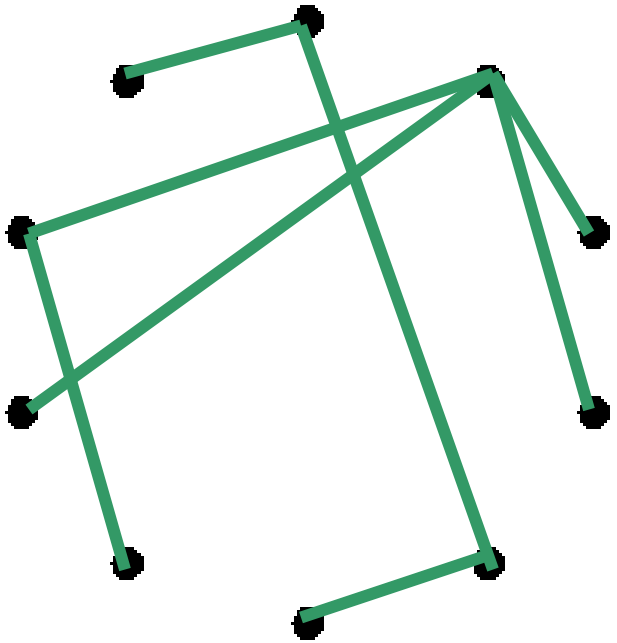


Networks without scale

NETWORK	NODES	LINKS
Cellular metabolism	Molecules involved in burning food for energy	Participation in the same biochemical reaction
Hollywood	Actors	Appearance in the same movie
Internet	Routers	Optical and other physical connections
Protein regulatory network	Proteins that help to regulate a cell's activities	Interactions among proteins
Research collaborations	Scientists	Co-authorship of papers
Sexual relationships	People	Sexual contact
World Wide Web	Web pages	URLs

Random Graphs

Erdös-Rényi model (1960)

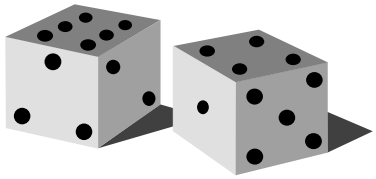


Connect with probability p

p=1/6

N=10

$\langle k \rangle \sim 1.5$



Erdös-Rényi model

Some properties:

- Average number of edges $\langle E \rangle = p\frac{N(N-1)}{2}$

- Average degree $\langle k \rangle = p(N-1)$

Finite average degree $\Rightarrow p \propto \frac{1}{N}$

Erdös-Rényi model

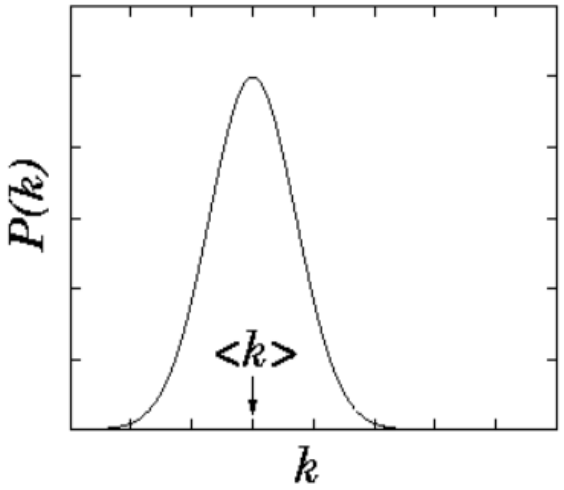
Proba to have a node of degree k=

- connected to k vertices,
- not connected to the other N-k-1

$$P(k) = C_{N-1}^{k} p^{k} (1-p)^{N-1-k}$$

Large N, fixed $pN = \langle k \rangle$: Poisson distribution

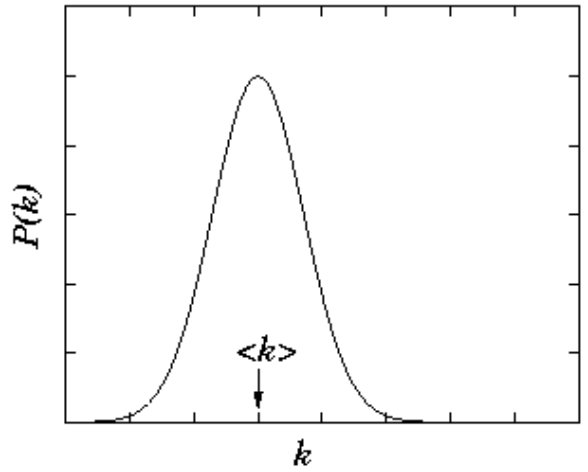
$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^{k}}{k!}$$



Exponential decay at large k

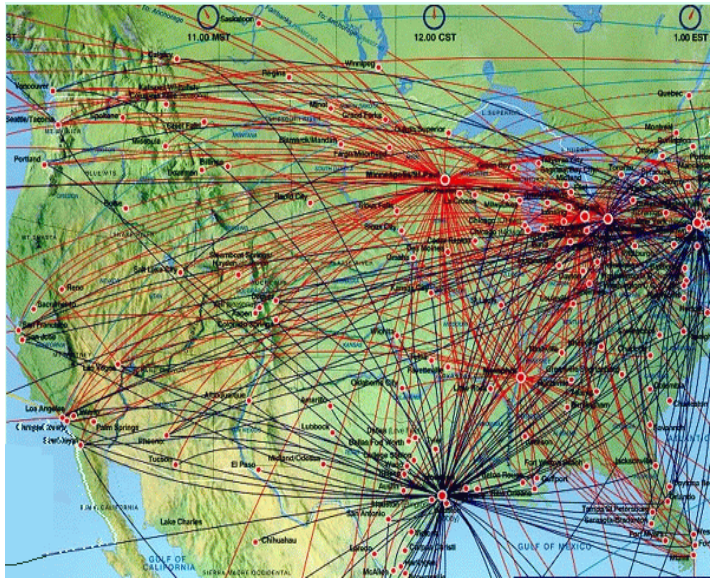
World Wide Web

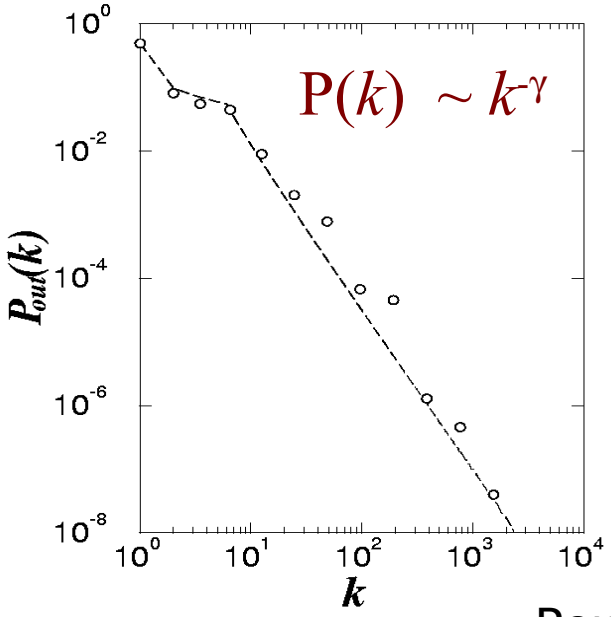
Exponential Network



Expected

Scale-free Network





Found

Power Law

R. Albert, H. Jeong, A-L Barabasi, *Nature*, **401** 130 (1999).

Scale free networks

(1) Networks continuously expand by the addition of new nodes

WWW : addition of new documents

(2) New nodes prefer to link to highly connected nodes.

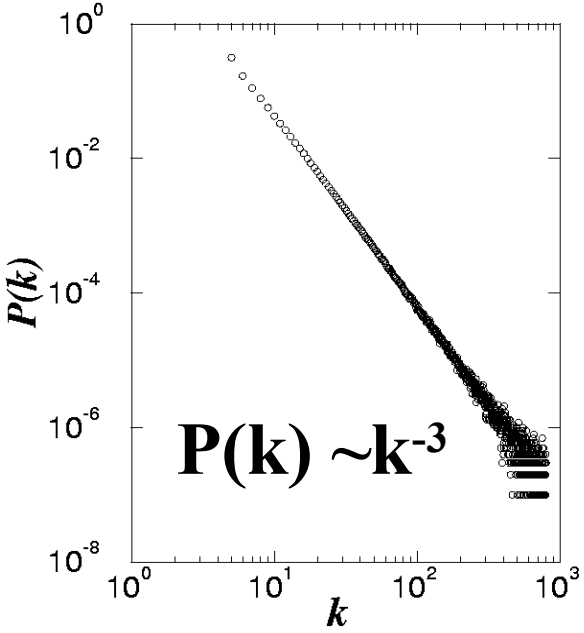
WWW : linking to well known sites

GROWTH:
add a new node with m links

PREFERENTIAL ATTACHMENT: the probability that a node connects to a node with k links is proportional to k.

$$\Pi(k_i) = \frac{k_i}{\Sigma_j k_j}$$

Preferential attachment



Barabási & Albert, *Science* **286,** 509 (1999)

What about late comers?

Fitness model is model of the evolution of a network:

how the links between nodes change over time depends on the **fitness** of nodes. Fitter nodes attract more links at the expense of less fit nodes

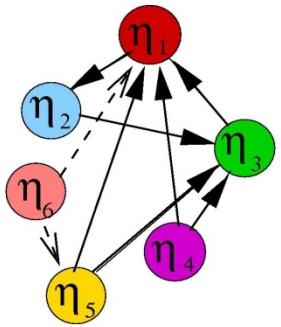
$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j}.$$

Bose-Einstein Condensation in evolving networks

Network

$$\Pi_i = \frac{\eta_i k_i}{\Sigma_j \eta_j k_j}$$

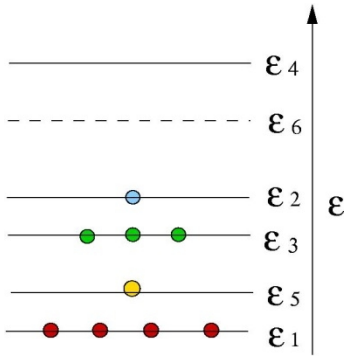
Bose gas



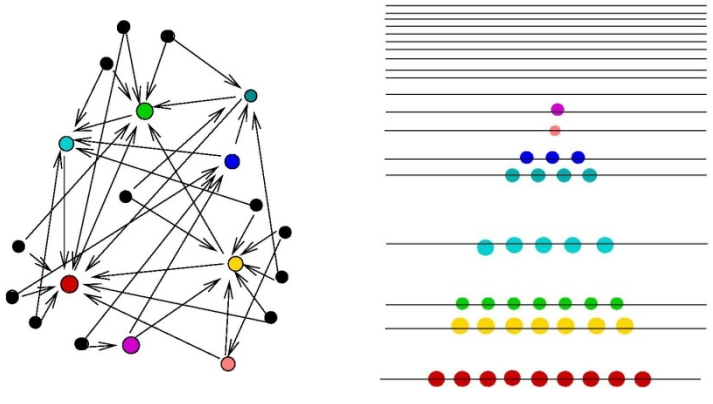
$$\eta \longrightarrow e^{-\beta\varepsilon}$$

$$k_{in}(\eta) \longrightarrow n(\varepsilon) = \frac{1}{e^{-\beta\varepsilon} - 1}$$

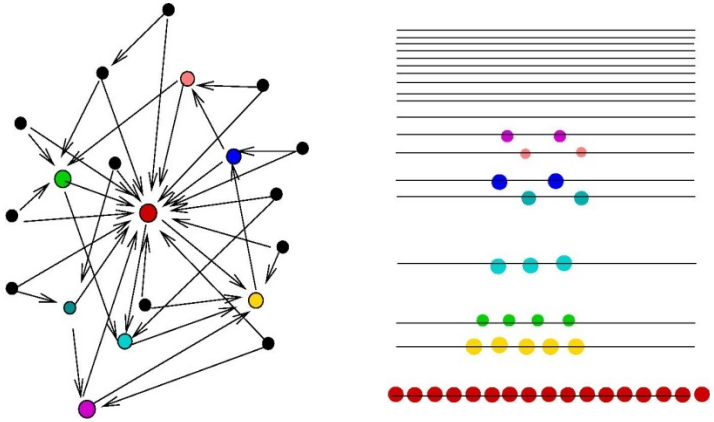
$$\rho(\eta) \longrightarrow g(\varepsilon)$$



Fit-gets-rich



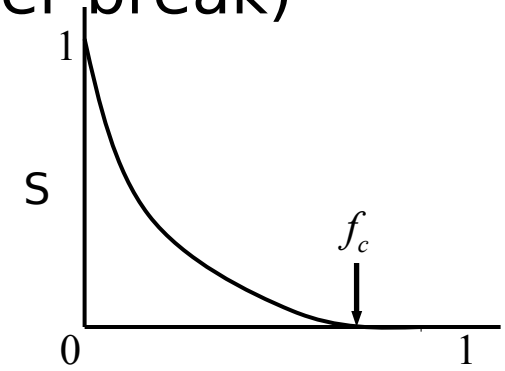
Bose-Einstein condensation

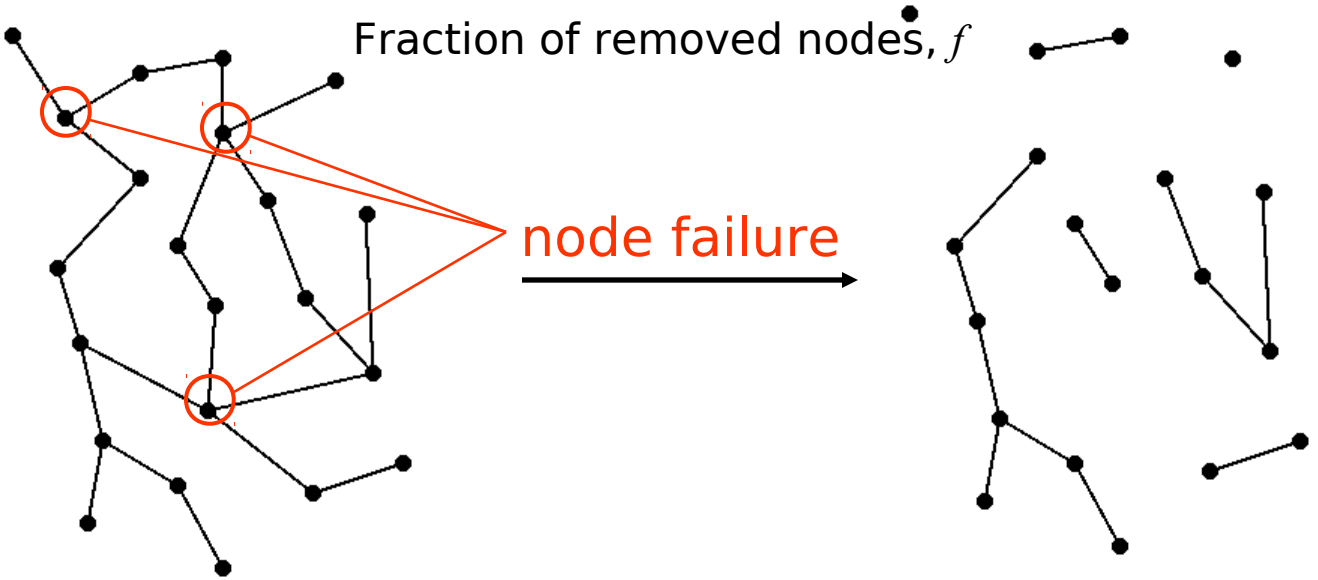


G. Bianconi and A.-L. Barabási, *Physical Review Letters* 2001; cond-mat/0011029

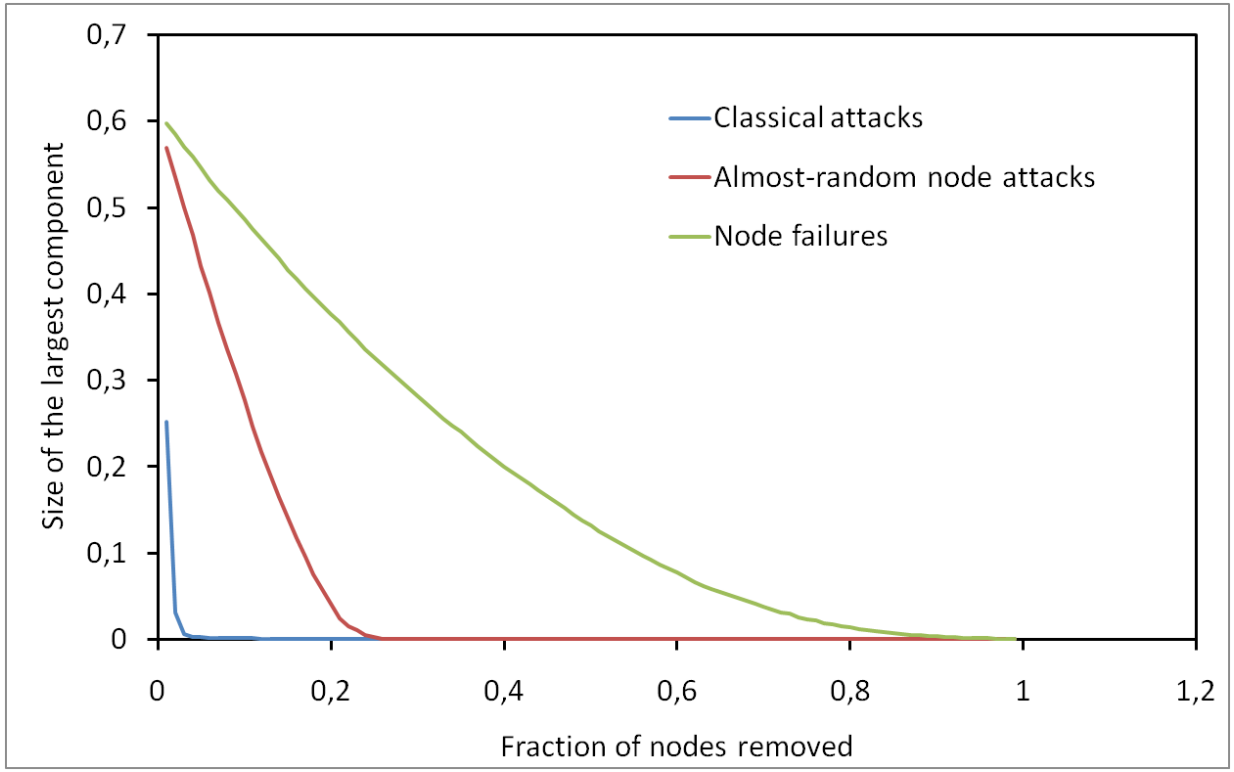
Robustness of Scale free networks

Complex systems maintain their basic functions even under errors and failures (cell → mutations; Internet → router break)



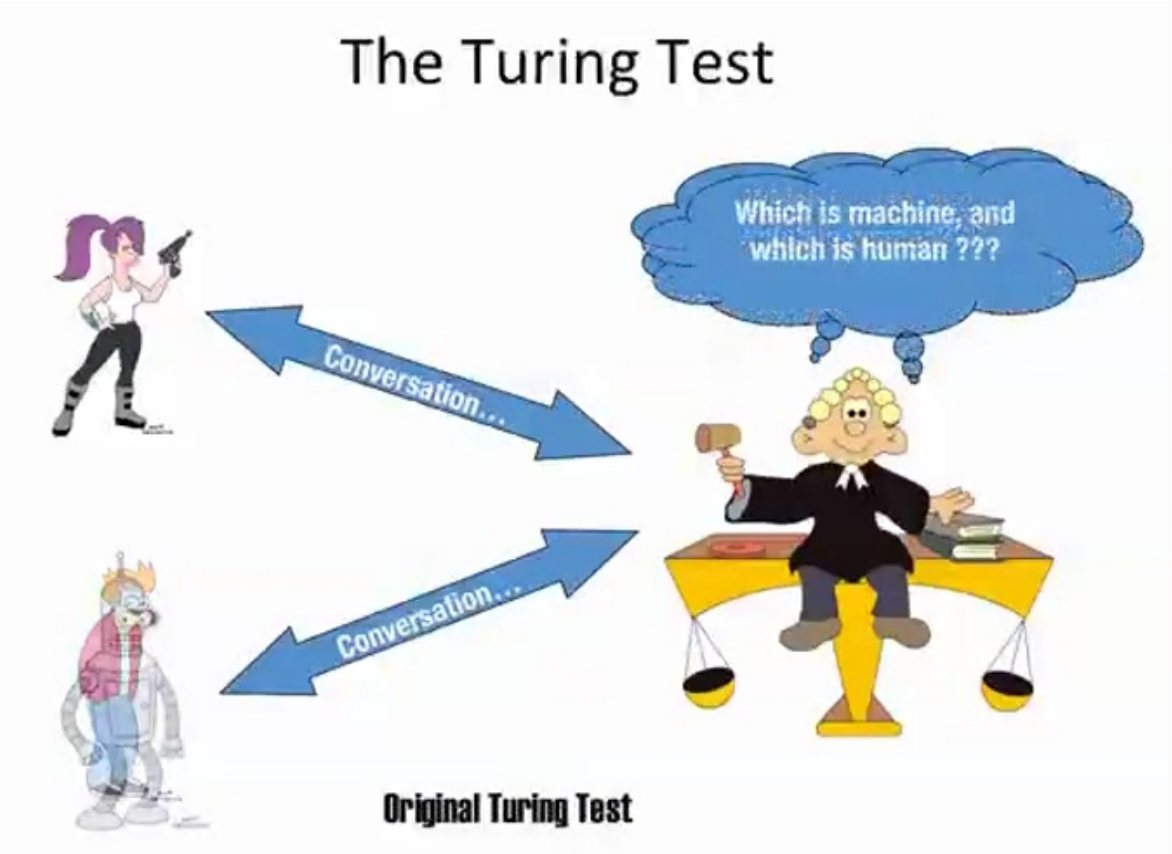


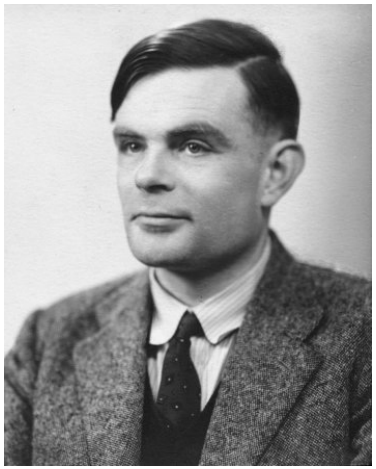
Robustness of Scale free networks



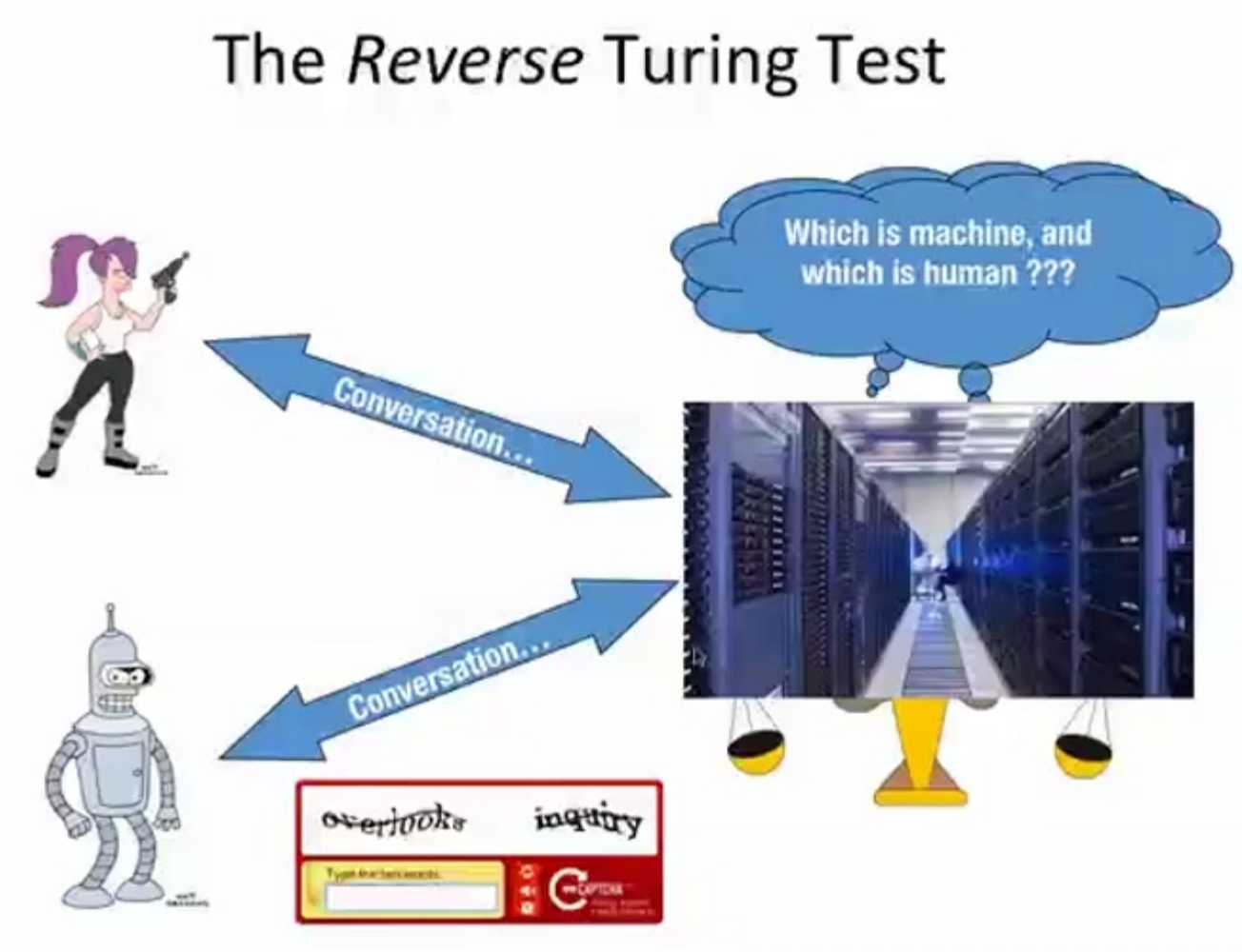
Robustness case
Attack case

Is a computer Intelligent?



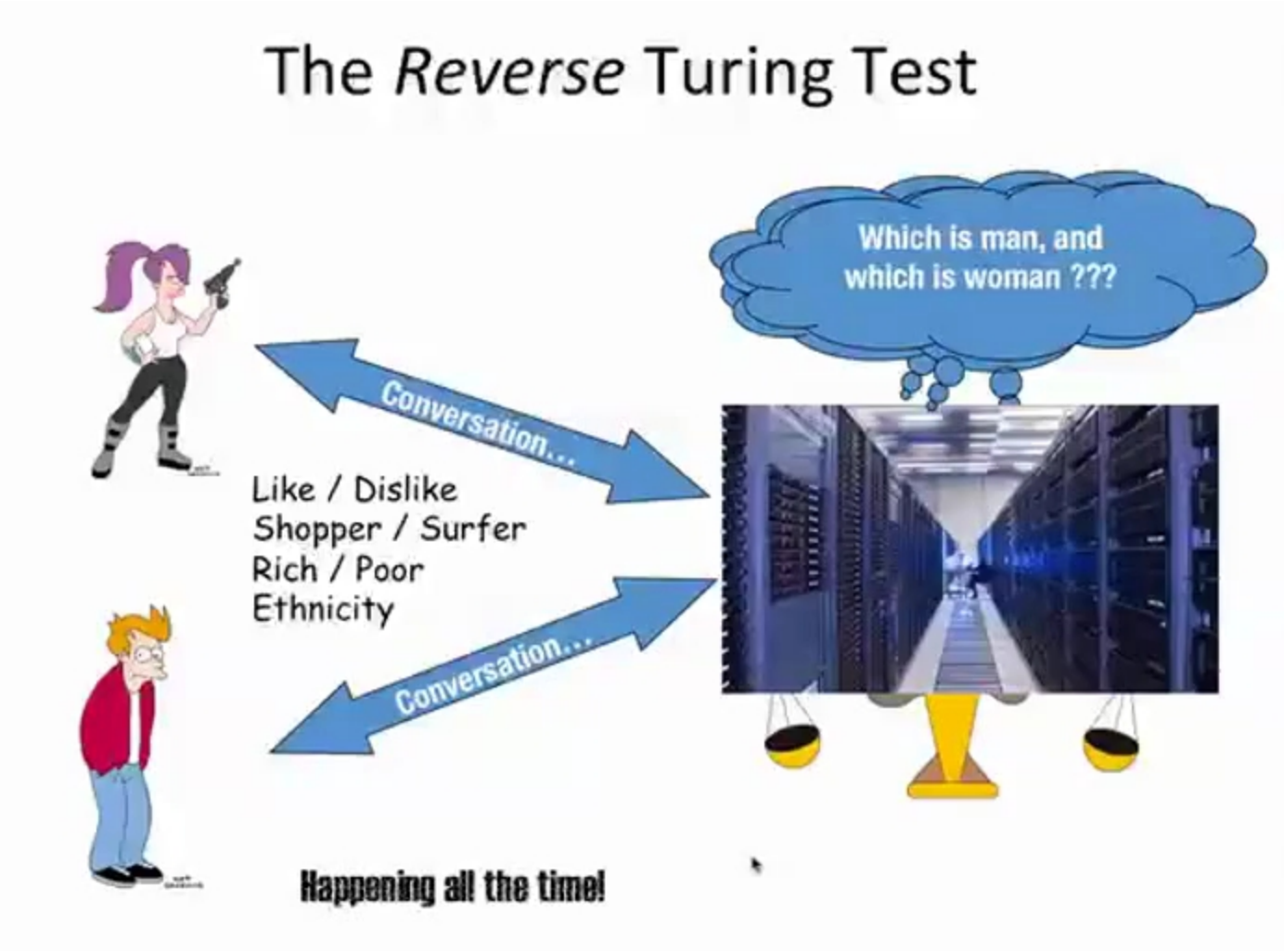


Is a computer Intelligent?



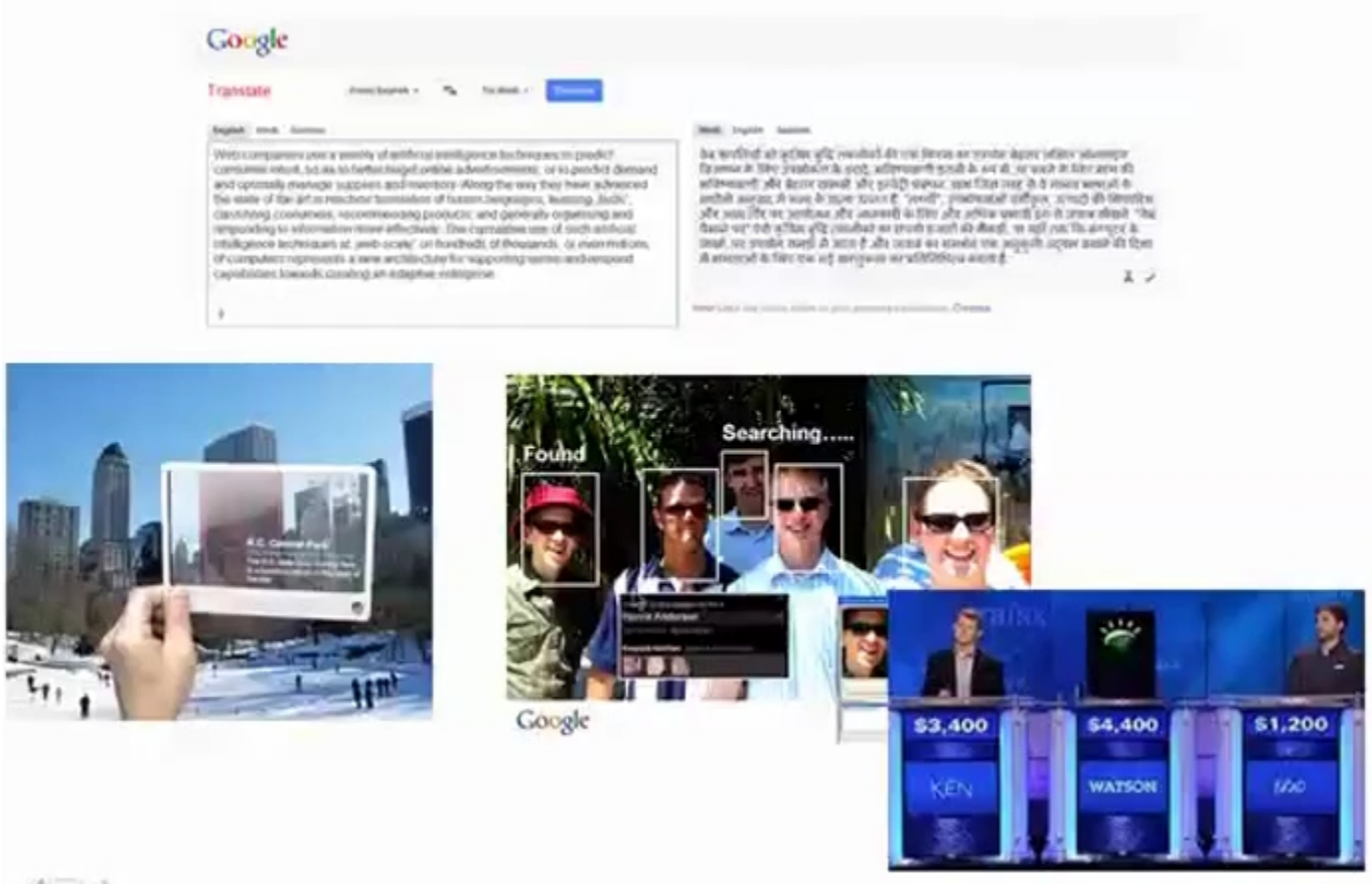
Dr. Gautham Shroff: Course on Web Intelligence on coursera.org

Is a computer Intelligent?



Dr. Gautham Shroff: Course on Web Intelligence on coursera.org

Web Intelligence @ Web Scale AI is here



IBM Watson at *Jeopardy 2011*

IBM

WATSON Goes to Work (For You)

270 BILLION

customer calls are handled annually[1]

Nearly

50 PERCENT

of all incoming service calls require escalation, dispatch, or go unresolved[2]

61 PERCENT

of customer calls could have been resolved with better access to information[2]

FINANCIAL SERVICES

Data Science?

Big-Data *technology*

traditional `business intelligence' using databases:

Google, Facebook, Linkedin, eBay, Amazon ...
did *not* use `traditional' databases for `big data'
why?
what?

- *massive parallelism*
- *Map-Reduce paradigm*

Data Science?

measuring *information* … what is "news"?

why did they do this?
so that *you* read the story!
"dog bites man" – not news
"man bites dog" – interesting!
why?


The New York Times

Europe

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION

AFRICA AMERICAS ASIA PACIFIC EUROPE MIDDLE EAST

At British Inquiry, Murdoch Apologizes Over Scandal

By ALAN COWELL

LONDON — After a day of testimony at a British judicial inquiry over his ties, friendships and disputes with British politicians, Rupert Murdoch returned to the witness stand on Thursday, saying he apologized for failing to take measures to avert the hacking scandal that has convulsed his media outpost here.

FACEBOOK TWITTER GOOGLE+ E-MAIL

Claude Shannon (1948): *information* is related to surprise
a message informing us of an event that has probability p conveys

$-\log_2 p$ bits of *information*

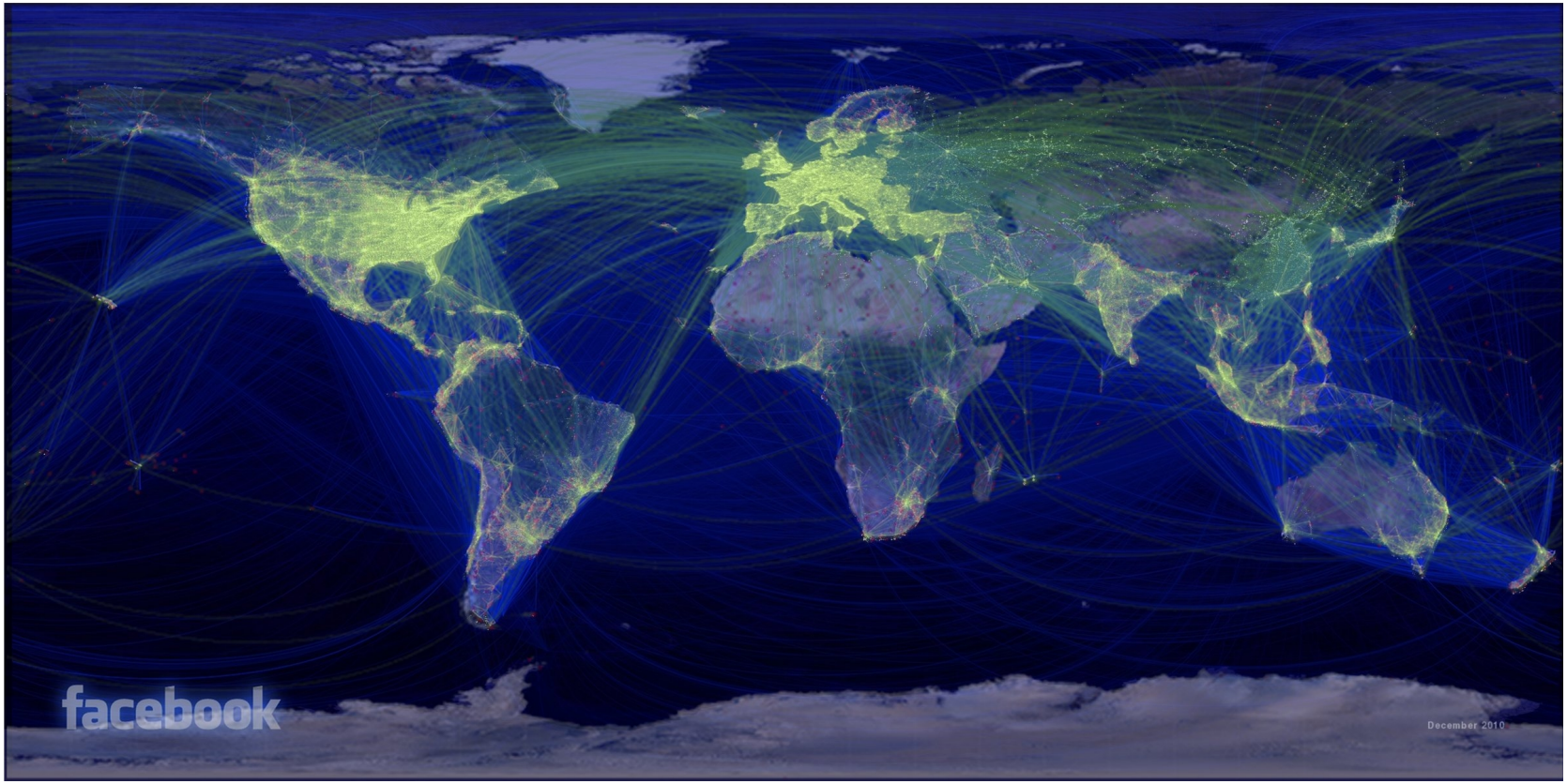
a, in, the, ..
information
miscellaneous

Predicting Scientific Laws?

Eurequa : Already predicted fundamental equations

Patterns in data ...

facebook connection



Network Science?

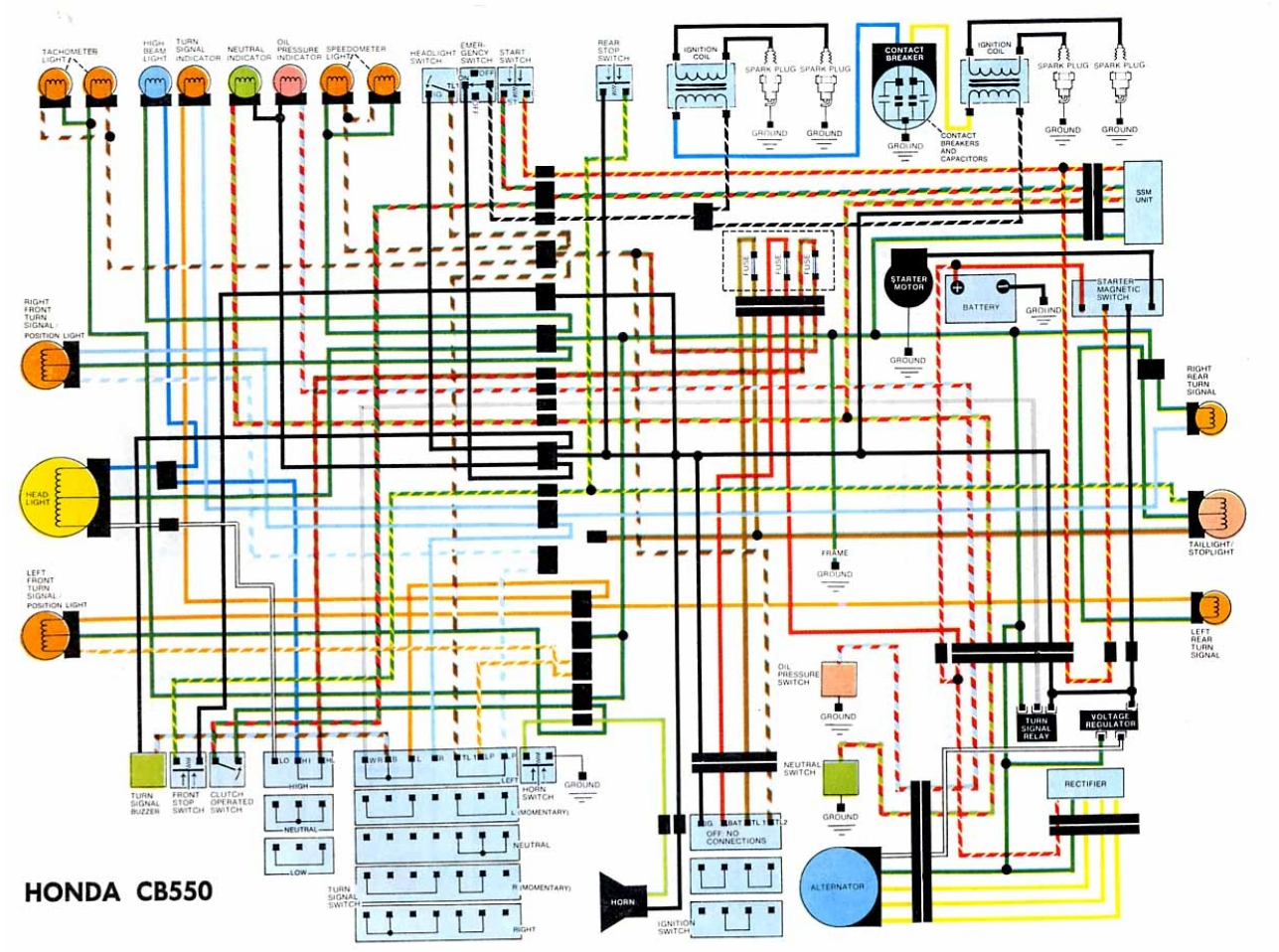
- Watch this

Network Science?

- What is the dynamics of these network?
- How to control the complex network?

Principles shall be drawn from Control Theory

Inverse Problem



Dynamical Systems

- State variables: What is the number (min) of control points required to drive the system?
- Linear systems: Kalaman Filter
- What about Non-linear systems?

Tail End

The 21st century," physicist Stephen Hawking has said, "will be the century of complexity."

Likewise, the physicist Heinz Pagels has said that "the nations and people who master the new sciences of complexity will become the economic, cultural, and political superpowers of the 21st century."

References

- Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life... by Albert-Laszlo Barabasi (Apr 29, 2003)
- The Structure and Dynamics of Networks: (Princeton Studies in Complexity) by Mark Newman, Albert-László Barabási and Duncan J. Watts (Apr 17, 2006)
- Bursts: The Hidden Patterns Behind Everything We Do, from Your E-mail to Bloody Crusades by Albert-Laszlo Barabasi (May 31, 2011)

Thank You