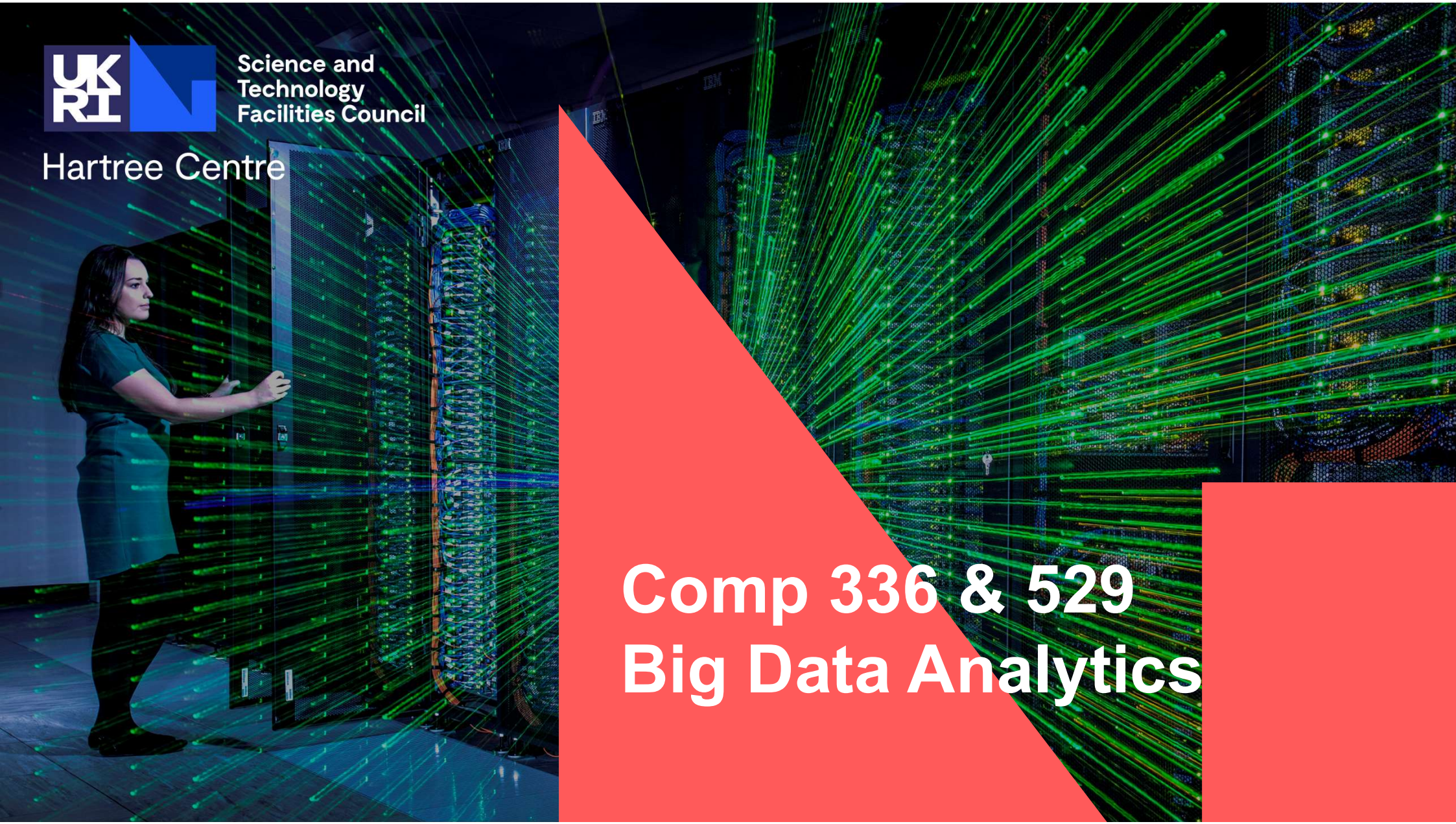


Comp 336 & 529
Big Data Analytics

Hartree Centre

Science and Technology Facilities Council

Hartree Centre

INTELLECTUAL PROPERTY RIGHTS NOTICE:

Many thanks to Prof. Laszlo Barabasi for agreeing to use his lectures and lecture material.

Please note that the corresponding Intellectual Property Rights apply as per Northeastern University, USA and Cambridge University Press for any material from Network Science book.

Week 4: Introduction to Network Science and Network Science algorithms for data analysis

Prof Vassil Alexandrov

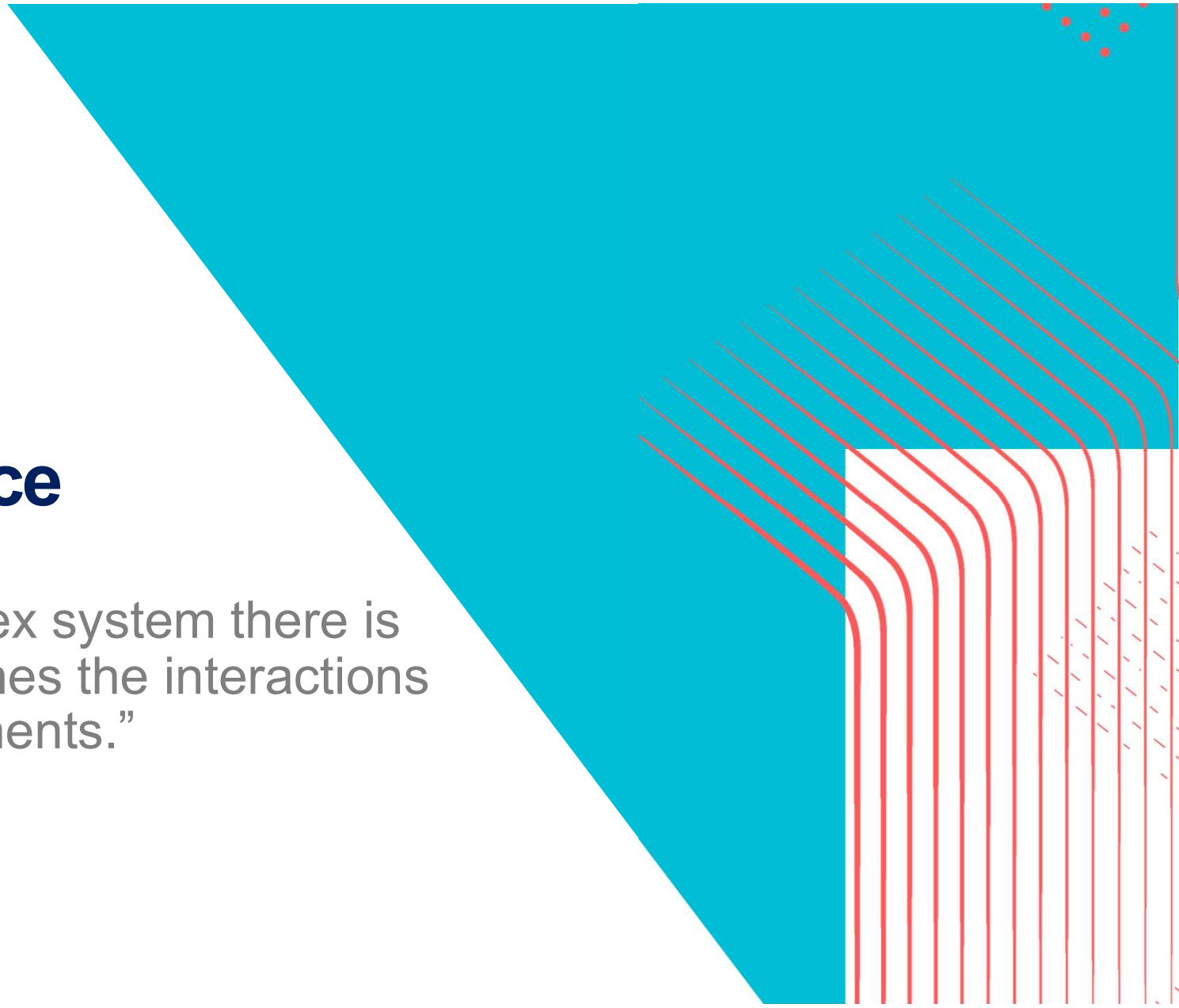
With material from Prof. Laszlo Barabasi's lecture notes in Network Science

Science and Technology Facilities Council

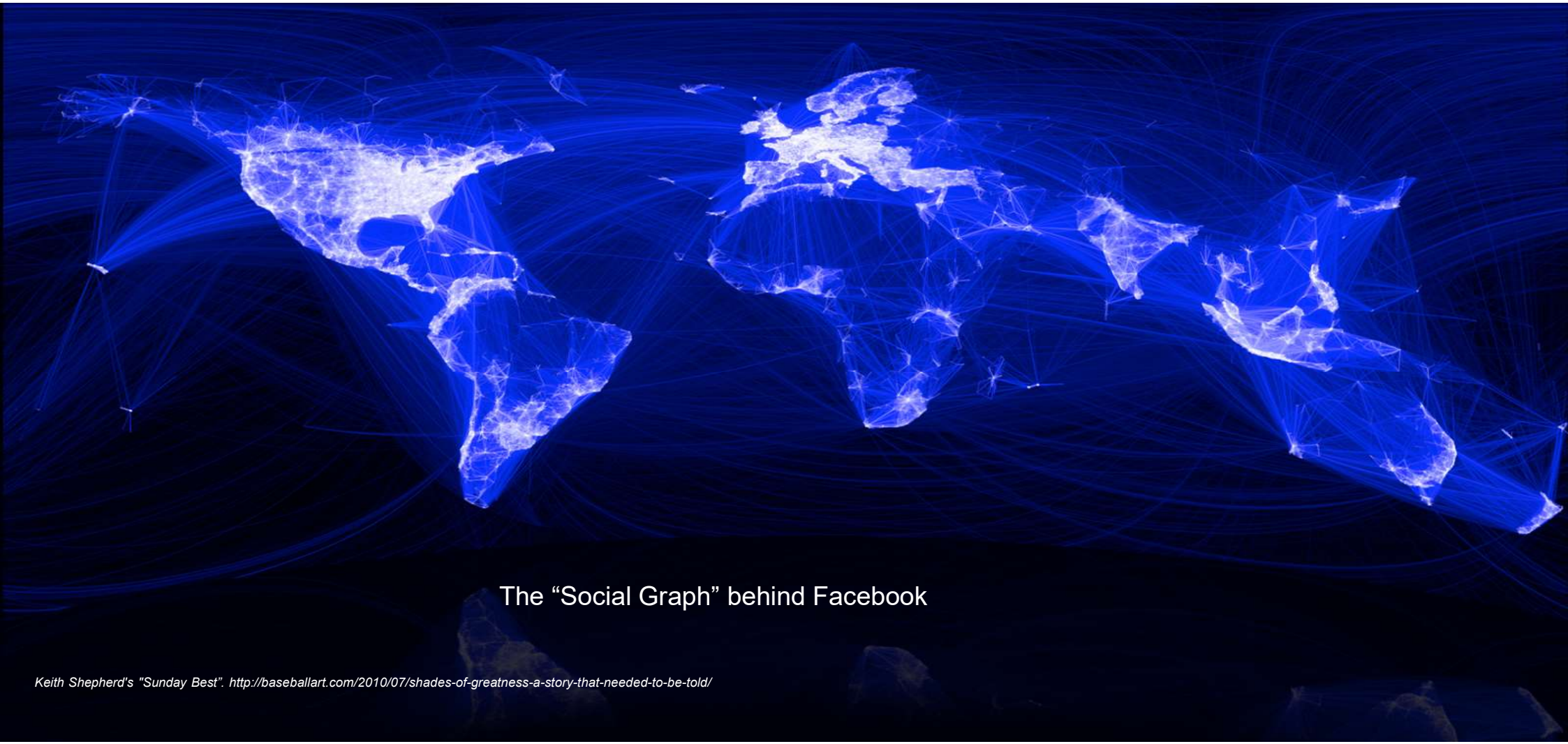
Hartree Centre

Network Science

"Behind each complex system there is a **network**, that defines the interactions between the components."



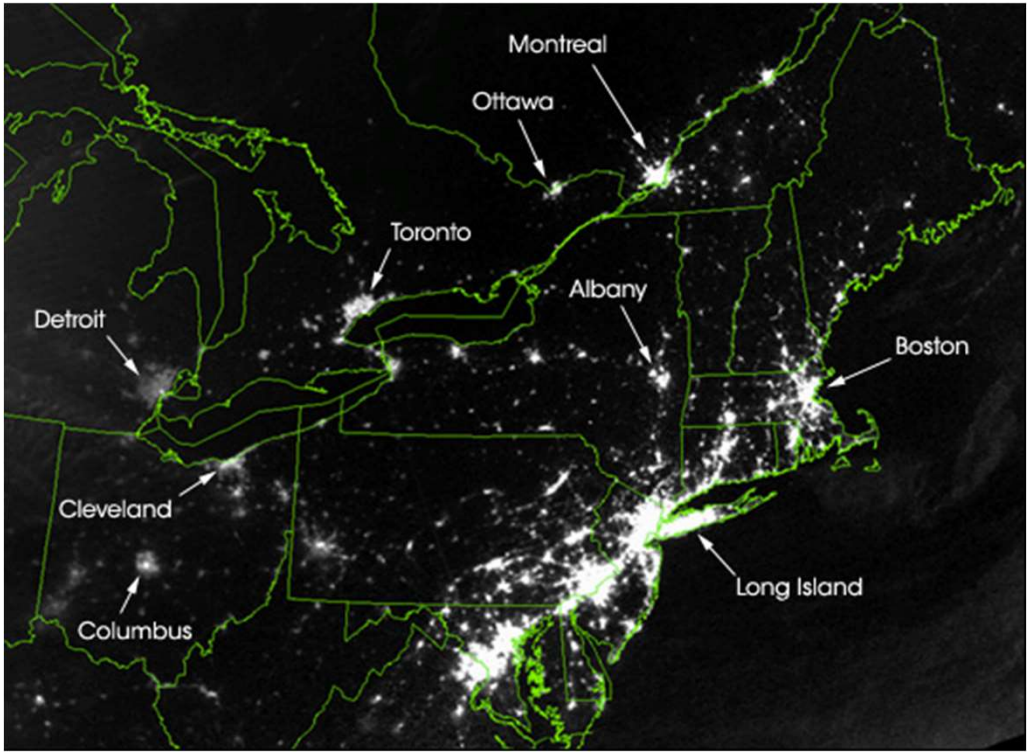
The role of networks



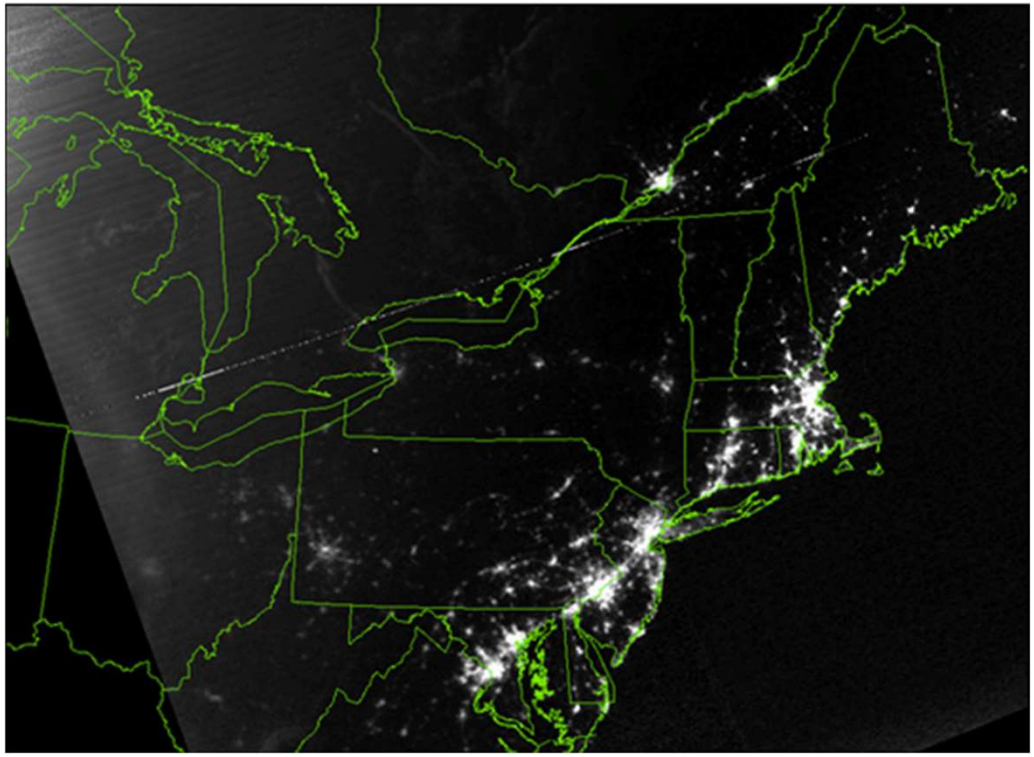
The “Social Graph” behind Facebook

Keith Shepherd's "Sunday Best". http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/

Vulnerability due to interconnectivity



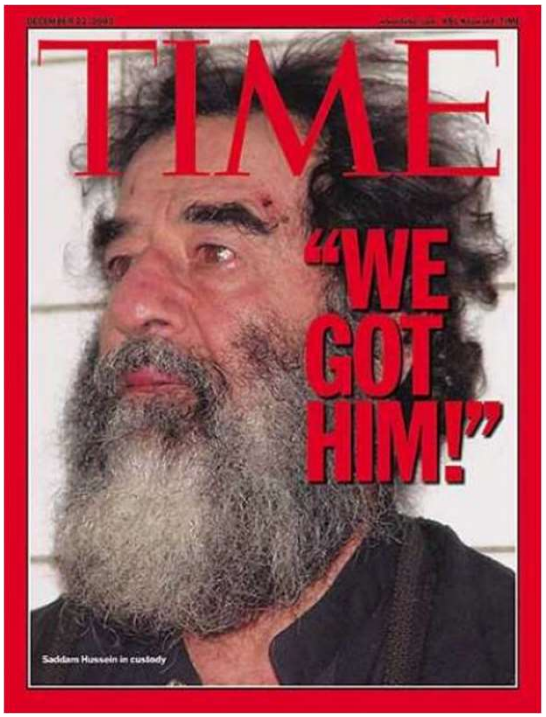
August 14, 2003: 9:29pm EDT
20 hours before

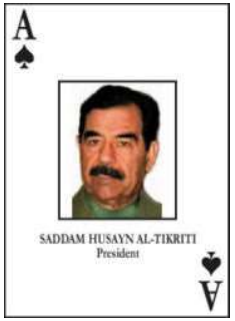


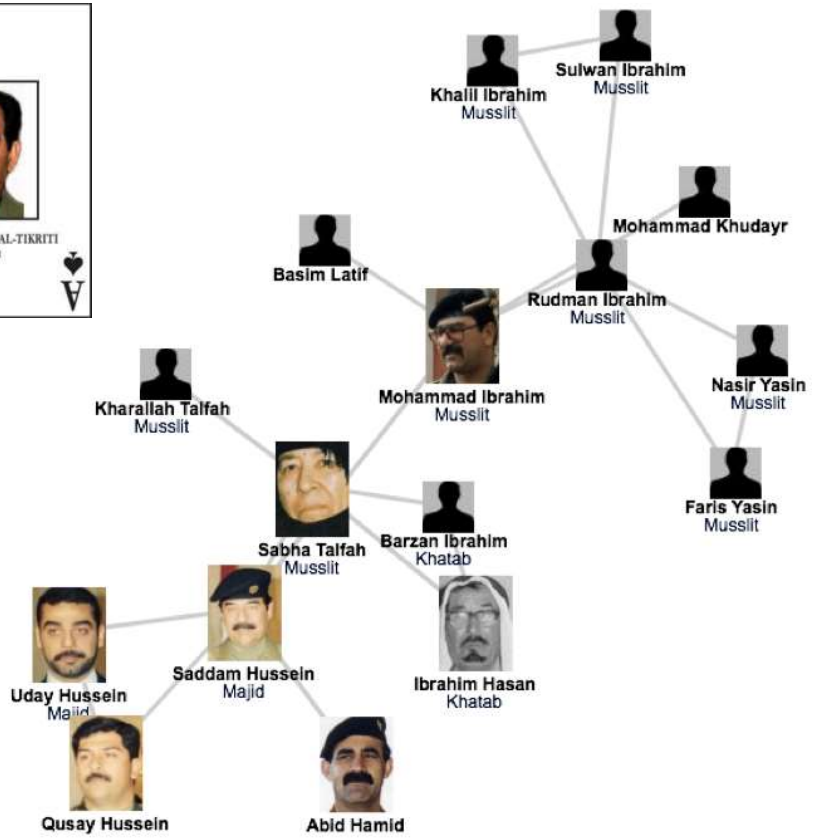
August 15, 2003: 9:14pm EDT
7 hours after

Science and Technology Facilities Council
Hartree Centre

The fate of Saddam and Network Science







Science and Technology Facilities Council

Hartree Centre

Lazslo Barabasi - Network Science, Cambridge University Press, 2016 & L. Barabasi's Lecture Notes

Networks: enable us to understand Complex Systems

Complex

1.

composed of many interconnected parts; compound; composite: a complex highway system.

2.

characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.

3.

so complicated or intricate as to be hard to understand or deal with: a complex problem.

Source: Dictionary.com

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

Source: John L. Casti, Encyclopædia Britannica

Complexity

Network Science characteristics

- Interdisciplinary
- Empirical and Data Driven
- Quantitative and Mathematical
- Computational

Applications and societal impact

- Healthcare: Drug design and development, forecasting and evaluating the spread off viruses

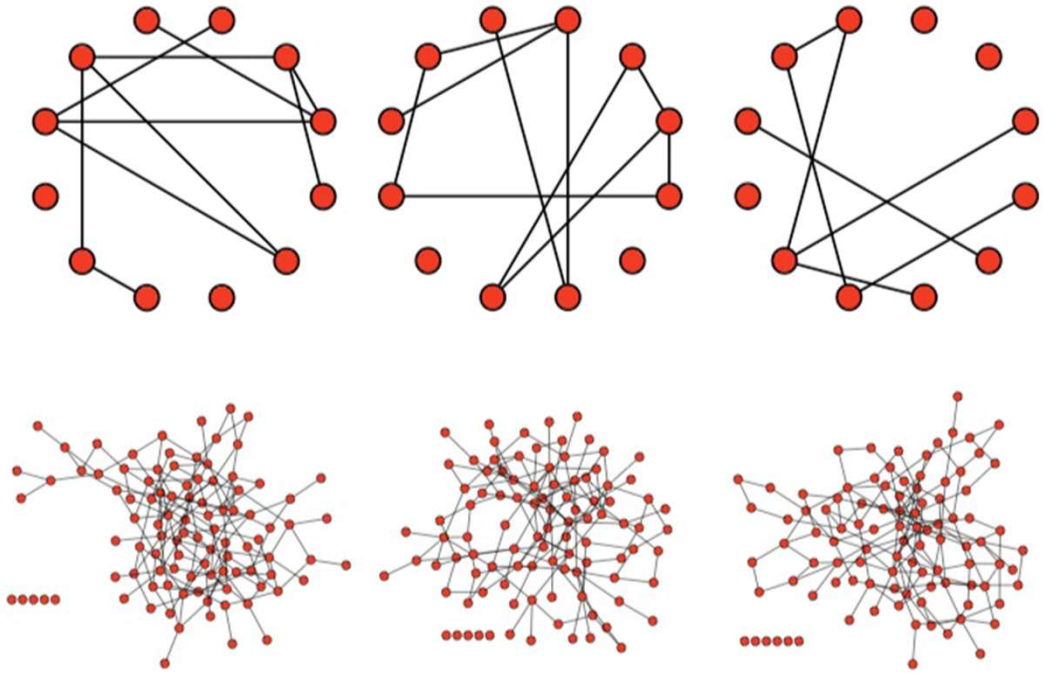
- Security and Cyber security: fighting terrorism

- Economics and management: many companies base their business on network based approach, Amazon,
- Google, etc;

- Social Networks and Analysis: Facebook, Twitter (now X), LinkedIn

Random networks

Random networks a truly random ones:



Laszlo Barabasi, Network Science, 2016

Random networks

"Random network consists of *N* nodes, where each node pair is connected with a probability *p* "

For a random network the degree distribution P_k - the probability that the randomly chosen node has degree *k* follows a binomial distribution:

$$p_k = \binom{N-1}{k} P^k (1-p)^{N-1-k}$$

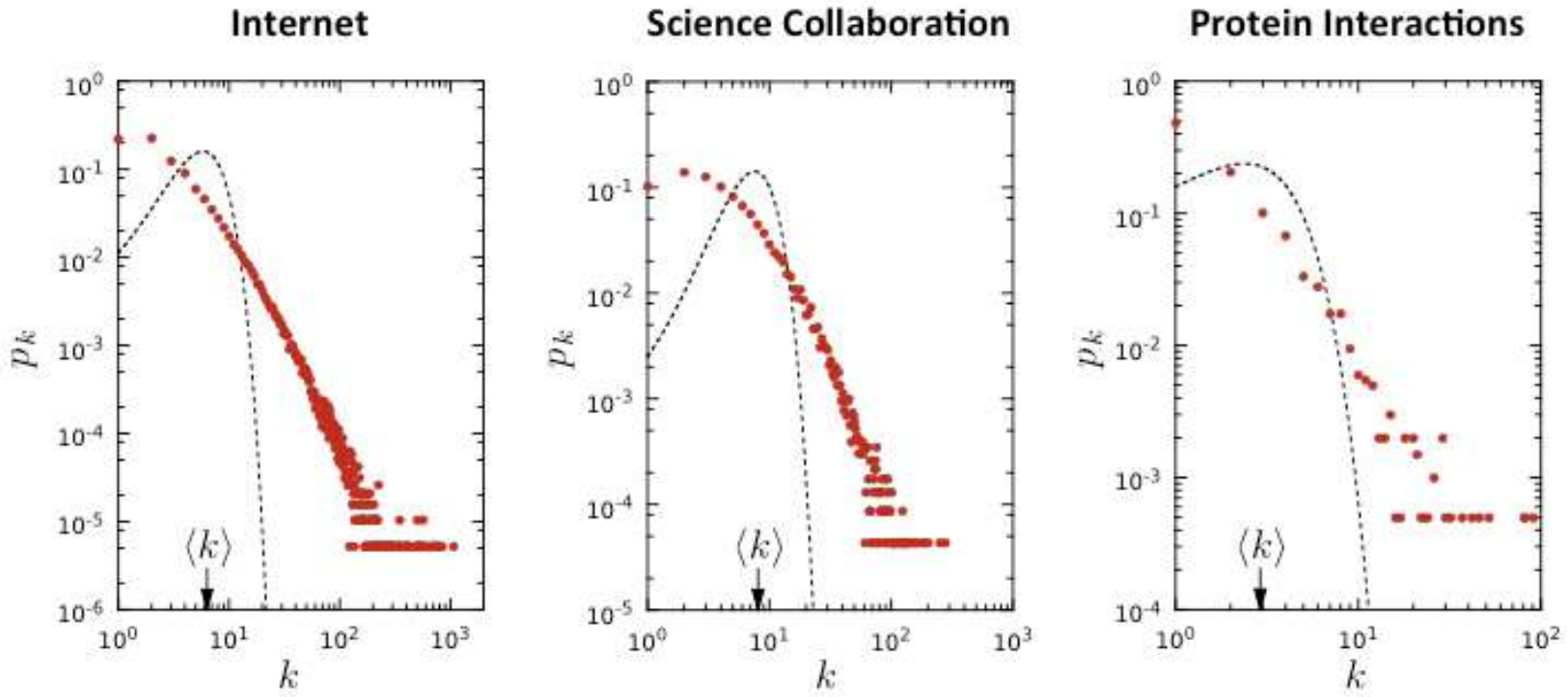
And since most real networks are sparse, e.g. k<<N it is approximated by the Poisson distribution:

$$p_k = e^{-k} \frac{k^k}{k!}$$

Laszlo Barabasi, Network Science, 2016

Scale-Free networks

Real networks are not random!



Laszlo Barabasi, Network Science, 2016

Scale-Free networks

Real networks are not random!

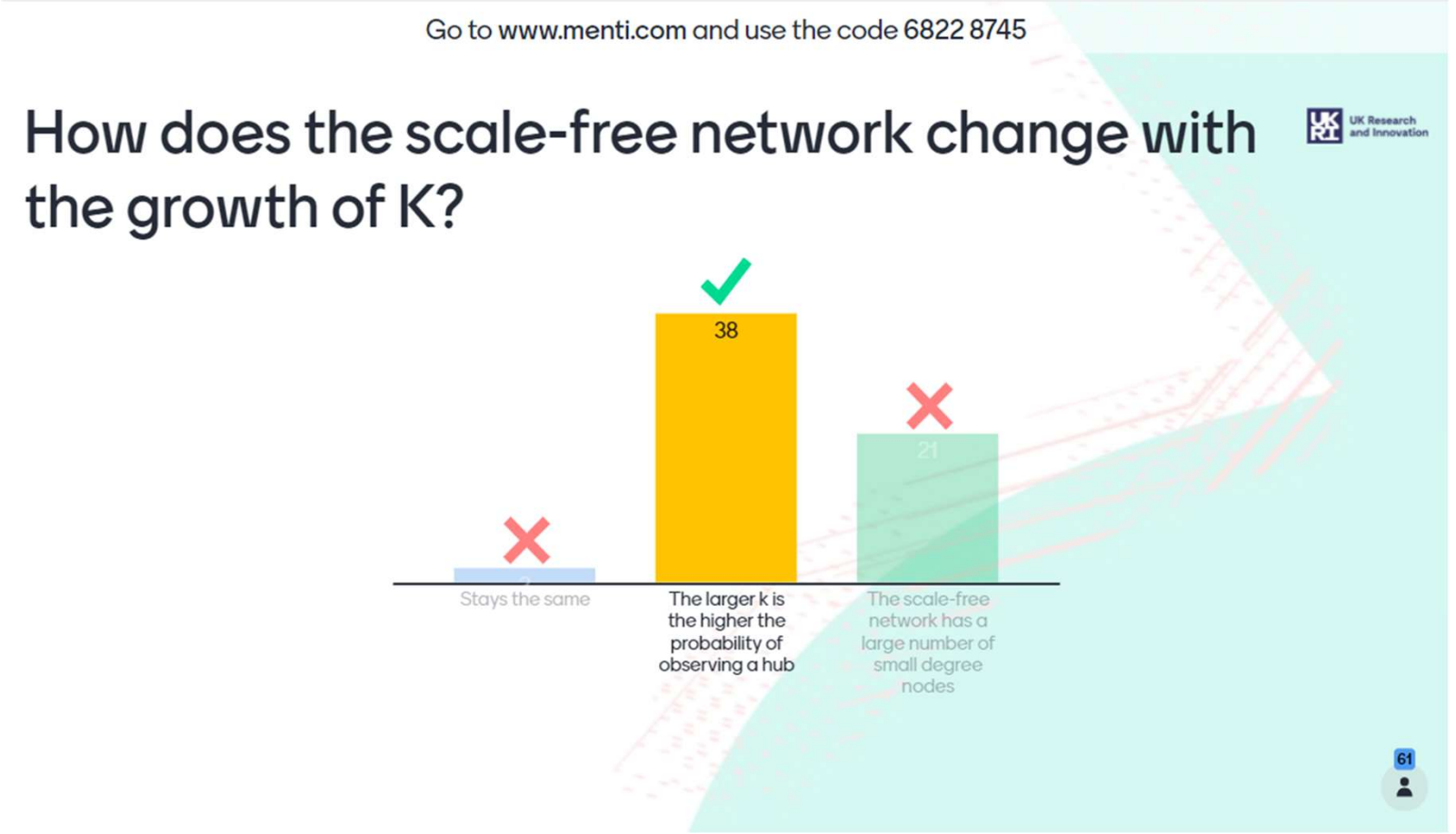
For example the WWW is modeled by a scale-free network where

$$p_k \sim k^{-\gamma}$$

This is the power law distribution and γ is its degree exponent.

The difference between the scale-free and random networks is in the **Hubs** as per power law distribution the larger is *k* the higher is the probability observing a hub. For small *k*, the scale-free network has large number of small degree nodes.

Quiz



Science and Technology Facilities Council

Hartree Centre

Network Science

BA model

Albert-László Barabási

With

Roberta Sinatra

www.BarabasiLab.com

Section 1

Hubs represent the most striking difference between a random and a scale-free network. Their emergence in many real systems raises several fundamental questions:

•Why does the random network model of Erdős and Rényi fail to reproduce the hubs and the power laws observed in many real networks?

• Why do so different systems as the WWW or the cell converge to a similar scale-free architecture?

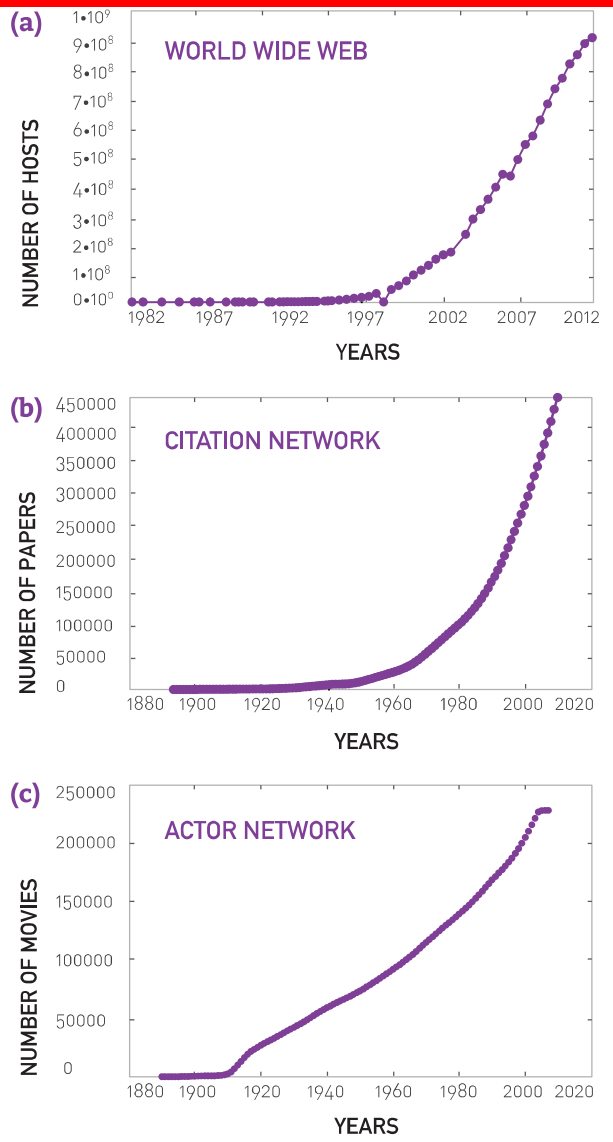
Section 2

Growth and preferential attachment

BA MODEL: Growth

ER model:
the number of nodes, N, is fixed (static models)

networks expand through the addition of new nodes



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

BA MODEL: Preferential attachment

ER model: links are added randomly to the network

New nodes prefer to connect to the more connected nodes

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

Section 2: Growth and Preferential Sttachment

The random network model differs from real networks in two important characteristics:

Growth: While the random network model assumes that the number of nodes is fixed (time invariant), real networks are the result of a growth process that continuously increases.

Preferential Attachment: While nodes in random networks randomly choose their interaction partner, in real networks new nodes prefer to link to the more connected nodes.

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

Section 3

The Barabási-Albert model

Origin of SF networks: Growth and preferential attachment

(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

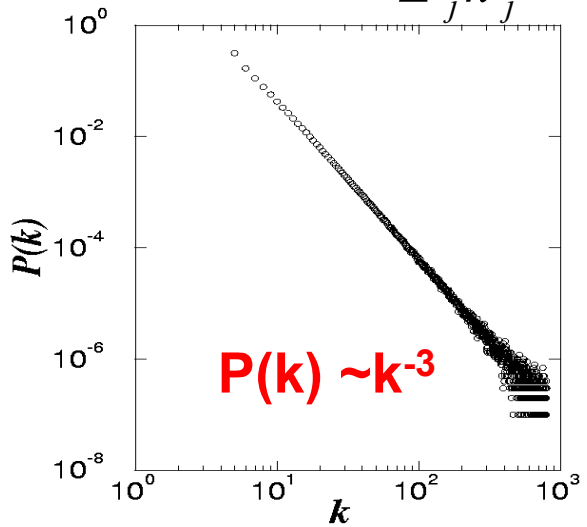
GROWTH:

add a new node with m links

PREFERENTIAL ATTACHMENT:

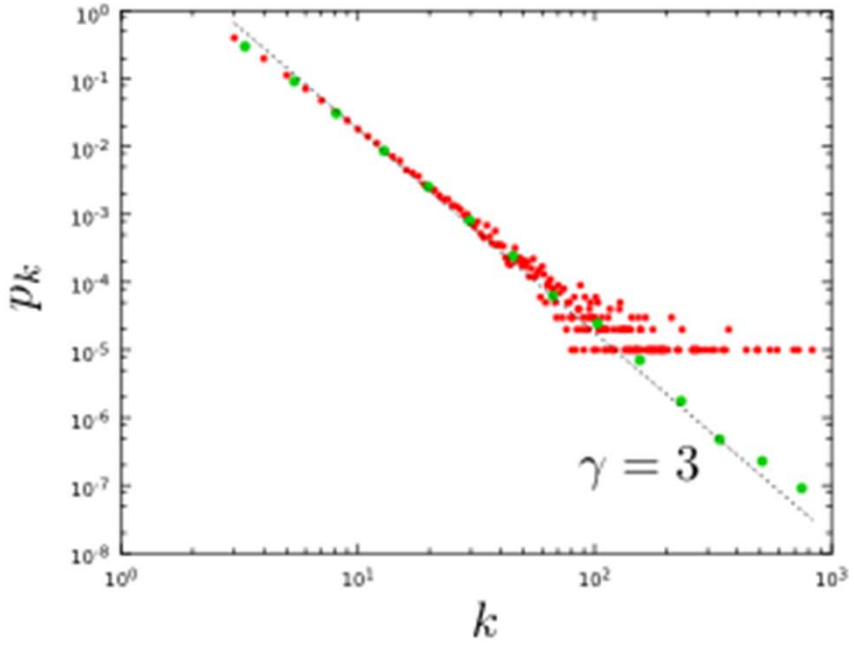
the probability that a node connects to a node with *k* links is proportional to *k*.

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



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

Section 4



György Pólya
PÓLYA PROCESS
MATHEMATICIAN

George Kinsley Zipf
WEALTH DISTRIBUTION
ECONOMIST

Herbert Alexander Simon
MASTER EQUATION
POLITICAL SCIENTIST

Robert Merton
MATTHEW EFFECT
SOCIOLOGIST

Albert-László Barabási & Réka Albert
PREFERENTIAL ATTACHMENT
NETWORK SCIENTISTS

George Udmy Yule
YULE PROCESS
STATISTICIAN

Robert Gibrat
PROPORTIONAL GROWTH
ECONOMIST

Derek de Solla Price
CUMULATIVE ADVANTAGE
PHYSICIST

MILESTONES

PUBLICATION DATE

1923 1925 1931 1935 1941 1945 1950 1955 1960 1968 1970 1976 1980 1985 1990 1995 1999 XXI 2000 2005 2010

György Pólya (1887-1985)
Preferential attachment made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya [2]. Hence, in mathematics preferential attachment is often called a **Pólya process**.

George Udmy Yule (1871-1951)
used preferential attachment to explain the power-law distribution of the number of species per genus of flowering plants [3]. Hence, in statistics preferential attachment is often called a **Yule process**.

Robert Gibrat (1904-1980)
proposed that the size and the growth rate of a firm are independent. Hence, larger firms grow faster [4]. Called **proportional growth**, this is a form of preferential attachment.

George Kinsley Zipf (1902-1950)
used preferential attachment to explain the fat tailed distribution of wealth in the society [5].

Herbert Alexander Simon (1916-2001)
used preferential attachment to explain the fat-tailed nature of the distributions describing city sizes, word frequencies, or the number of papers published by scientists [6].

Derek de Solla Price (1922-1983)
used preferential attachment to explain the citation statistics of scientific publications, referring to it as **cumulative advantage** [7].

Robert Merton (1910-2003)
In sociology preferential attachment is often called the **Matthew effect**, named by Merton [8] after a passage in the Gospel of Matthew.

Barabási (1967) & **Albert** (1972)
introduce the term **preferential attachment** in the context of networks [1] to explain the origin of their power-law degree distribution.

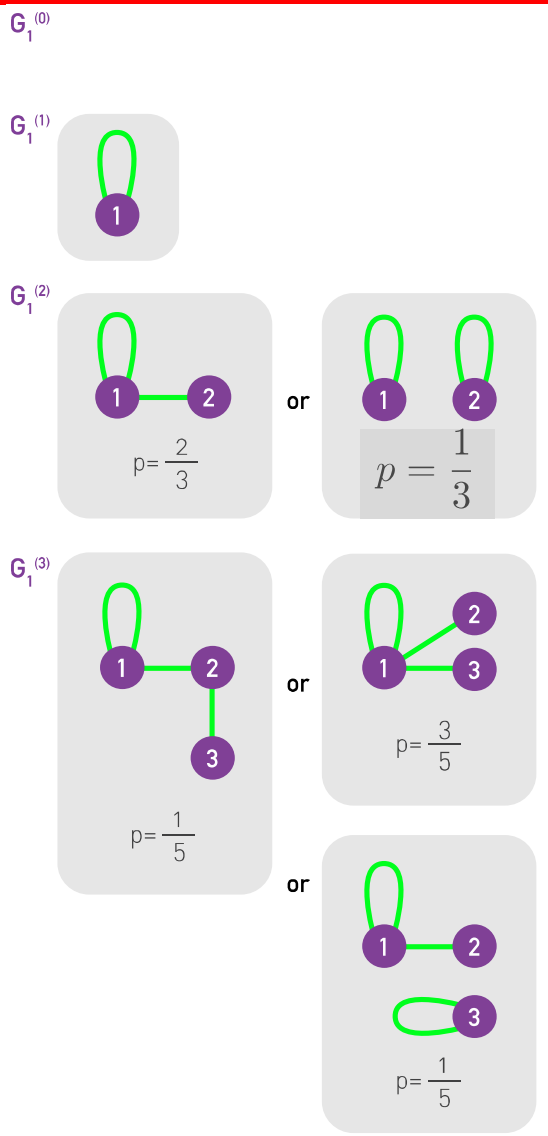
Section 4

Linearized Chord Diagram

The definition of the Barabási-Albert model leaves many mathematical details open:

- It does not specify the precise initial configuration of the first m_0 nodes.
- It does not specify whether the m links assigned to a new node are added one by one, or simultaneously. This leads to potential mathematical conflicts: If the links are truly independent, they could connect to the same node i, leading to multi-links.

$$p(i=s)=\begin{cases} \dfrac{k_i}{2t-1} & if\ 1 \le s \le t-1 \\ \dfrac{1}{2t-1}, & if\ s=t \end{cases}$$



Section 4

Degree dynamics

All nodes follow the same growth law

$$\frac{\partial k_i}{\partial t} \propto \Pi(k_i) = A\frac{k_i}{\sum_j k_j}$$

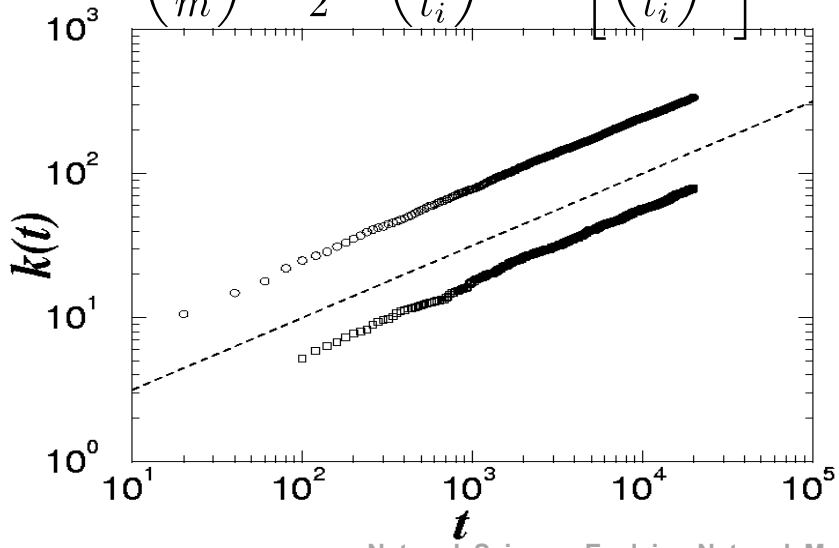
Use: $\sum_j k_j = 2mt$

During a unit time (time step): $\Delta k = m \rightarrow A = m$

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t} \qquad \frac{\partial k_i}{k_i} = \frac{\partial t}{2t} \qquad \int_m^k \frac{\partial k_i}{k_i} = \int_{t_i}^t \frac{\partial t}{2t}$$

$$\ln\left(\frac{k}{m}\right) = \frac{1}{2}\ln\left(\frac{t}{t_i}\right) = \ln\left[\left(\frac{t}{t_i}\right)^{\frac{1}{2}}\right]$$

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \qquad \beta = \frac{1}{2}$$

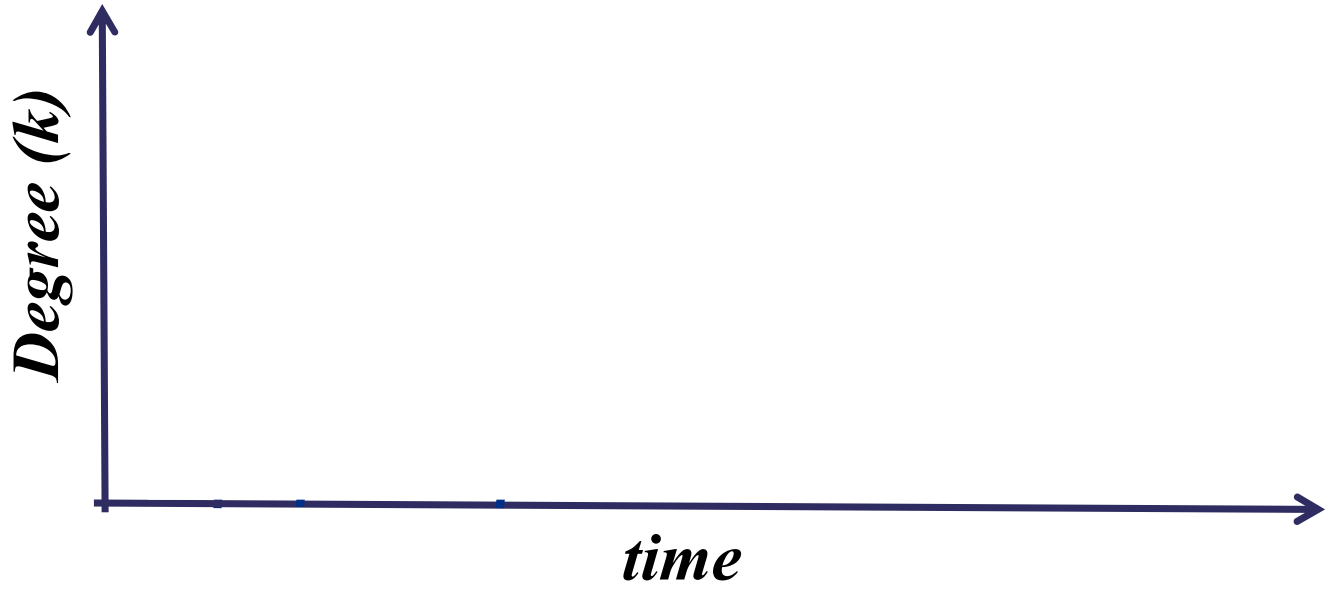


β: dynamical exponent

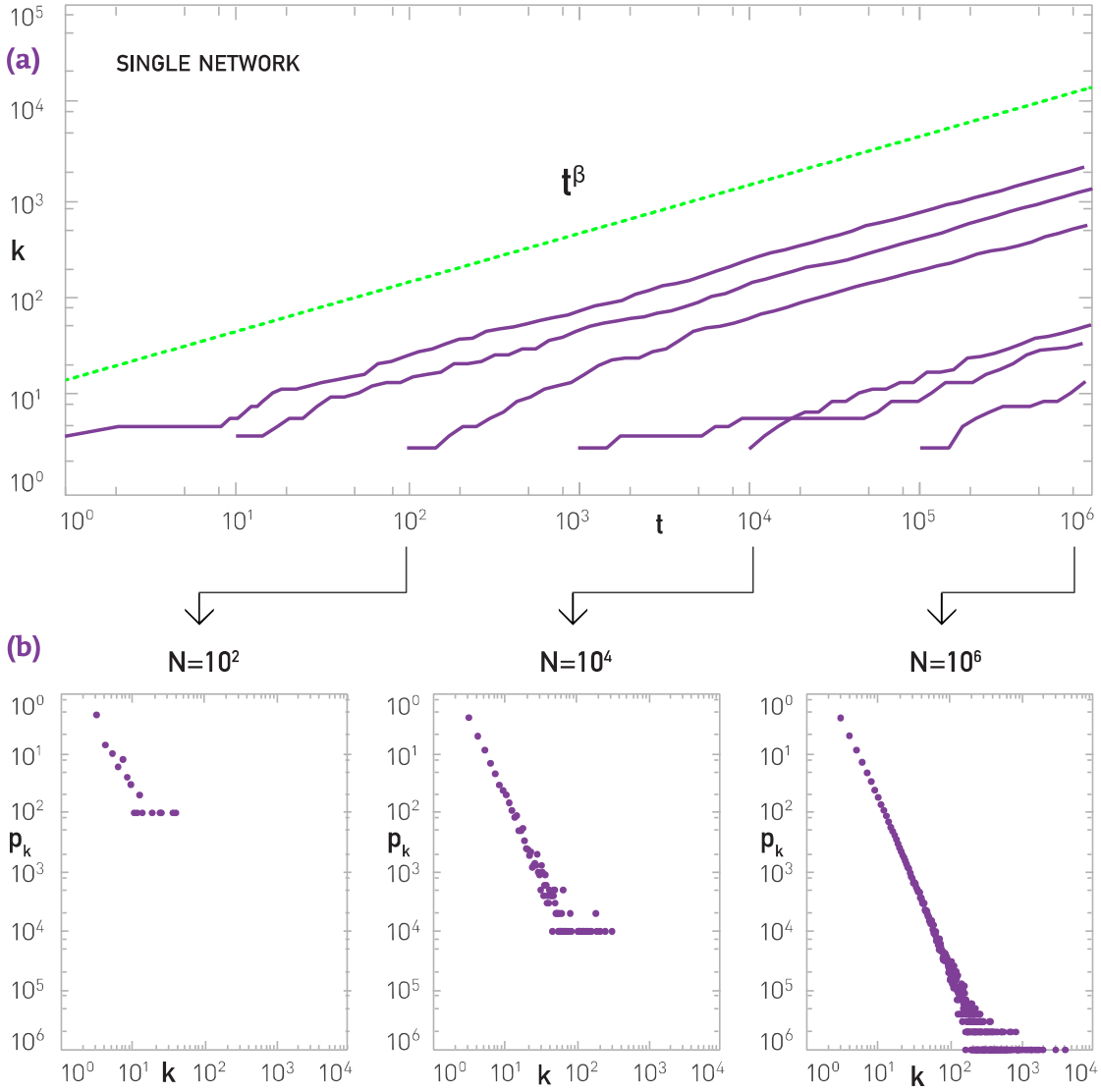
A.-L.Barabási, R. Albert and H. Jeong, *Physica* A **272,** 173 (1999)

All nodes follow the same growth law

SF model: $k(t) \sim t^{1/2}$ **(first mover advantage)**



Section 5.3



- The degree of each node increases following a power-law with the same dynamical exponent β =1/2 (Figure 5.6a). Hence all nodes follow the same dynamical law.

- The growth in the degrees is sublinear (i.e. $\beta < 1$). This is a consequence of the growing nature of the Barabási-Albert model: Each new node has more nodes to link to than the previous node. Hence, with time the existing nodes compete for links with an increasing pool of other nodes.

- The earlier node i was added, the higher is its degree $k_i(t)$. Hence, hubs are large because they arrived earlier, a phenomenon called *first-mover advantage* in marketing and business.

- The rate at which the node i acquires new links is given by the derivative of (5.7)

$$\frac{dk_i(t)}{dt} = \frac{m}{2}\frac{1}{\sqrt{t_i t}}, \qquad (5.8)$$

indicating that in each time frame older nodes acquire more links (as they have smaller t_i). Furthermore the rate at which a node acquires links decreases with time as $t^{-1/2}$. Hence, fewer and fewer links go to a node.

Section 5

Degree distribution

Degree distribution

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \qquad \beta = \frac{1}{2}$$

A node *i* can come with equal probability any time between $t_i=m_0$ and t, hence:

$$P(t_i) = \frac{1}{m_0 + t} \qquad P(t_i < \tau) = \frac{1}{m_0 + t}\int_0^{\tau} dt_i = \frac{\tau}{m_0 + t}$$

$$P(k) = P\left(t_i \le \frac{m^{1/\beta} t}{k^{1/\beta}}\right) = 1 - \frac{m^{1/\beta} t}{k^{1/\beta}(t + m_0)}$$

$$\therefore P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{2m^2 t}{m_o + t}\frac{1}{k^3} \sim k^{-\gamma}$$

$$\gamma = 3$$

A.-L.Barabási, R. Albert and H. Jeong, *Physica* A **272,** 173 (1999)

Degree distribution

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \qquad \beta = \frac{1}{2}$$

$$P(k) = \frac{2m^2 t}{m_o + t}\frac{1}{k^3} \sim k^{-\gamma} \qquad \boxed{\gamma = 3}$$

(i) The degree exponent is independent of m.

(ii) As the power-law describes systems of rather different ages and sizes, it is expected that a correct model should provide a time-independent degree distribution. Indeed, asymptotically the degree distribution of the BA model is independent of time (and of the system size N)
→ the network reaches a stationary scale-free state.

(iii) The coefficient of the power-law distribution is proportional to m^2.

A.-L.Barabási, R. Albert and H. Jeong, *Physica* A **272,** 173 (1999)

The mean field theory offers the correct scaling, BUT it provides the wrong coefficient of the degree distribution.

So assymptotically it is correct ($k \rightarrow \infty$), but not correct in details (particularly for small k).

To fix it, we need to calculate P(k) exactly, which we will do next using a rate equation based approach.

MFT - Degree Distribution: Rate Equation

$< N(k,t) >= tP(K,t)$ Number of nodes with degree *k* at time *t*.

Since at each timestep we add one node, we have N=t (total number of nodes =number of timesteps)

$$\Pi(k) = \frac{k}{\sum_j k_j} = \frac{k}{2mt}$$

2m: each node adds *m* links, but each link contributed to the degree of 2 nodes

Number of links added to degree *k* nodes after the arrival of a new node:

$$\frac{k}{2mt} \times NP(k,t) \times m = \frac{k}{2} P(k,t)$$

Total number of k-nodes

Preferential attachment

New node adds m new links to other nodes

Nr. of degree *k-1* nodes that acquire a new link, becoming degree *k*

$$\frac{k-1}{2} P(k-1,t)$$

Nr. of degree *k* nodes that acquire a new link, becoming degree *k+1*

$$\frac{k}{2} P(k,t)$$

$$(N+1)P(k,t+1) = NP(k,t) + \frac{k-1}{2} P(k-1,t) - \frac{k}{2} P(k,t)$$

k-nodes at time t+1

k-nodes at time t

Gain of k-nodes via k-1→ k

Loss of k-nodes via k→ k+1

MFT - Degree Distribution: Rate Equation

$$(N+1)P(k,t+1) = NP(k,t) + \frac{k-1}{2}P(k-1,t) - \frac{k}{2}P(k,t)$$

- $(N+1)P(k,t+1)$: # k-nodes at time t+1
- $NP(k,t)$: # k-nodes at time t
- $\frac{k-1}{2}P(k-1,t)$: Gain of k-nodes via k-1→ k
- $\frac{k}{2}P(k,t)$: Loss of k-nodes via k→ k+1

We do not have *k=0,1,...,m-1* nodes in the network (each node arrives with degree *m*)
→ We need a separate equation for degree *m* modes

$$(N+1)P(m,t+1) = NP(m,t) + 1 - \frac{m}{2}P(m,t)$$

- $(N+1)P(m,t+1)$: # m-nodes at time t+1
- $NP(m,t)$: # m-nodes at time t
- 1: Add one m-degeree node
- $\frac{m}{2}P(m,t)$: Loss of an m-node via m→ m+1

MFT - Degree Distribution: Rate Equation

$$(N+1)P(k,t+1) = NP(k,t) + \frac{k-1}{2}P(k-1,t) - \frac{k}{2}P(k,t) \qquad \text{k>m}$$

$$(N+1)P(m,t+1) = NP(m,t) + \quad 1 \quad - \frac{m}{2}P(m,t)$$

We assume that there is a stationary state in the N=t→∞ limit, when P(k,∞)=P(k)

$$(N+1)P(k,t+1) - NP(k,t) \to NP(k,\infty) + P(k,\infty) - NP(k,\infty) = P(k,\infty) = P(k)$$

$$(N+1)P(m,t+1) - NP(m,t) \to P(m)$$

$$P(k) = \frac{k-1}{2}P(k-1) - \frac{k}{2}P(k) \qquad P(k) = \frac{k-1}{k+2}P(k-1) \qquad \text{k>m}$$

$$P(m) = 1 - \frac{m}{2}P(m) \qquad P(m) = \frac{2}{2+m}$$

MFT - Degree Distribution: Rate Equation

$$P(k) = \frac{k-1}{k+2} P(k-1) \quad \rightarrow \quad P(k+1) = \frac{k}{k+2} P(k)$$

$$P(m) = \frac{2}{m+2}$$

$$P(m+1) = \frac{m}{m+3} P(m) = \frac{2m}{(m+2)(m+3)}$$

$$P(m+2) = \frac{m+1}{m+4} P(m+1) = \frac{2m(m+1)}{(m+2)(m+3)(m+4)}$$

$$P(m+3) = \frac{m+2}{m+5} P(m+2) = \frac{2m(m+1)}{(m+3)(m+4)(m+5)}$$

m+3 → *k*

...

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}$$

$$P(k) \sim k^{-3}$$ for large *k*

Krapivsky, Redner, Leyvraz, PRL 2000
Dorogovtsev, Mendes, Samukhin, PRL 2000
Bollobas et al, Random Struc. Alg. 2001

MFT - Degree Distribution: A Pretty Caveat

Start from eq. $$P(k) = \frac{k-1}{2} P(k-1) - \frac{k}{2} P(k)$$

$$2P(k) = (k-1)P(k-1) - kP(k) = -P(k-1) - k[P(k) - P(k-1)]$$

$$2P(k) = -P(k-1) - k\frac{P(k) - P(k-1)}{k - (k-1)} = -P(k-1) - k\frac{\partial P(k)}{\partial k}$$

$$P(k) = -\frac{1}{2}\frac{\partial [kP(k)]}{\partial k}$$

Its solution is: $P(k) \sim k^{-3}$

Dorogovtsev and Mendes, 2003

Degree distribution

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \qquad \beta = \frac{1}{2}$$

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \qquad \boxed{\gamma = 3}$$

$$P(k) \sim k^{-3} \qquad \text{for large } k$$

(i) The degree exponent is independent of m.

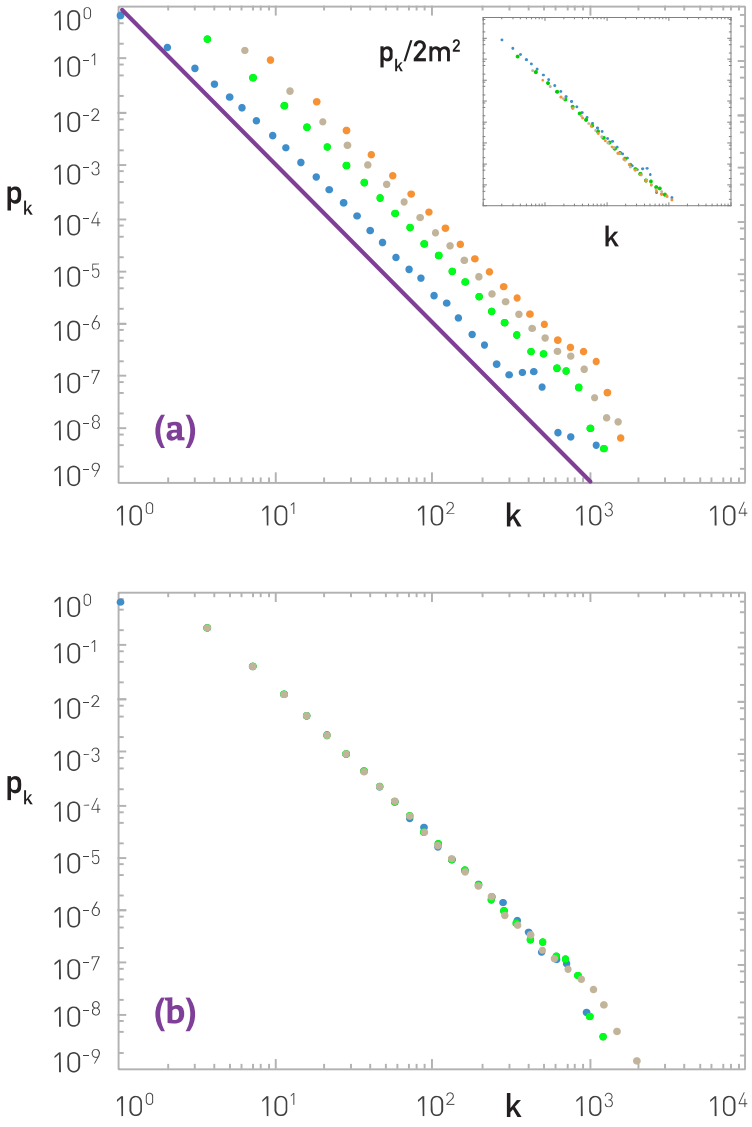
(ii) As the power-law describes systems of rather different ages and sizes, it is expected that a correct model should provide a time-independent degree distribution. Indeed, asymptotically the degree distribution of the BA model is independent of time (and of the system size N)

→ the network reaches a stationary scale-free state.

(iii) The coefficient of the power-law distribution is proportional to m^2.

NUMERICAL SIMULATION OF THE BA MODEL

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}$$



(a) We generated networks with N=100,000 and m_0=m=1 (blue), 3 (green), 5 (grey), and 7 (orange). The fact that the curves are parallel to each other indicates that γ is independent of m and m_0. The slope of the purple line is -3, corresponding to the predicted degree exponent γ=3. Inset: (5.11) predicts $p_k \sim 2m^2$, hence $p_k/2m^2$ should be independent of m. Indeed, by plotting $p_k/2m^2$ vs. k, the data points shown in the main plot collapse into a single curve.

(b) The Barabási-Albert model predicts that p_k is independent of N. To test this we plot p_k for N = 50,000 (blue), 100,000 (green), and 200,000 (grey), with m_0=m=3. The obtained p_k are practically indistinguishable, indicating that the degree distribution is stationary, i.e. independent of time and system size.

Section 6

absence of growth and preferential attachment

MODEL A

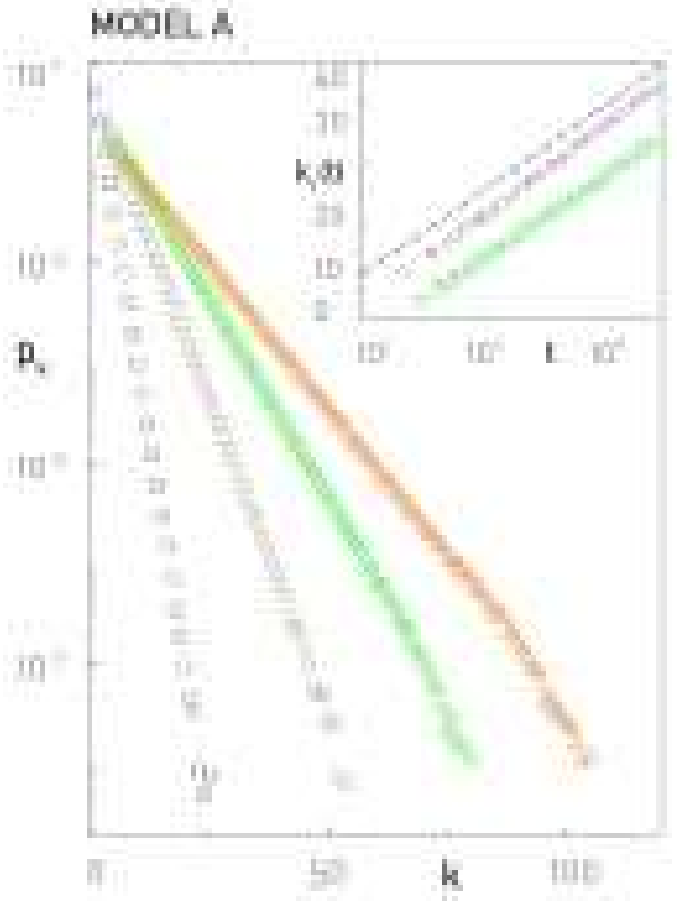
growth ~~preferential attachment~~

$\Pi(k_i)$: uniform

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) = \frac{m}{m_0 + t - 1}$$

$$k_i(t) = m \ln\left(\frac{m_0 + t - 1}{m + t_i - 1}\right) + m$$

$$P(k) = \frac{e}{m}\exp\left(-\frac{k}{m}\right) \sim e^{-k}$$

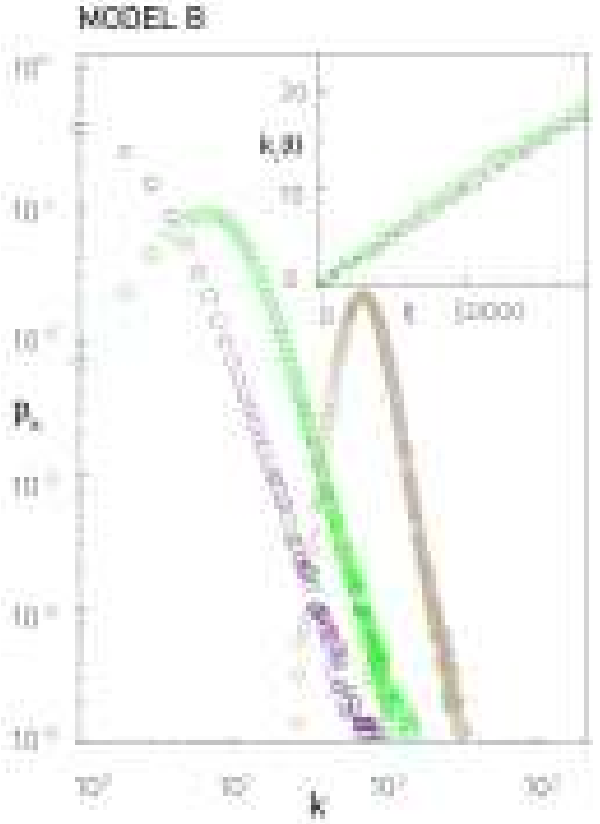


MODEL B

~~growth~~ preferential attachment

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) + \frac{1}{N} = \frac{N}{N-1}\frac{k_i}{2t} + \frac{1}{N}$$

$$k_i(t) = \frac{2(N-1)}{N(N-2)}t + Ct^{\frac{N}{2(N-1)}} \sim \frac{2}{N}t$$



p_k : power law (initially) →

→ Gaussian → Fully Connected

Do we need both growth and preferential attachment?

YEP.

Section 7

Measuring preferential attachment

Measuring preferential attachment

$$\frac{\partial k_i}{\partial t} \propto \Pi(k_i) \sim \frac{\Delta k_i}{\Delta t}$$

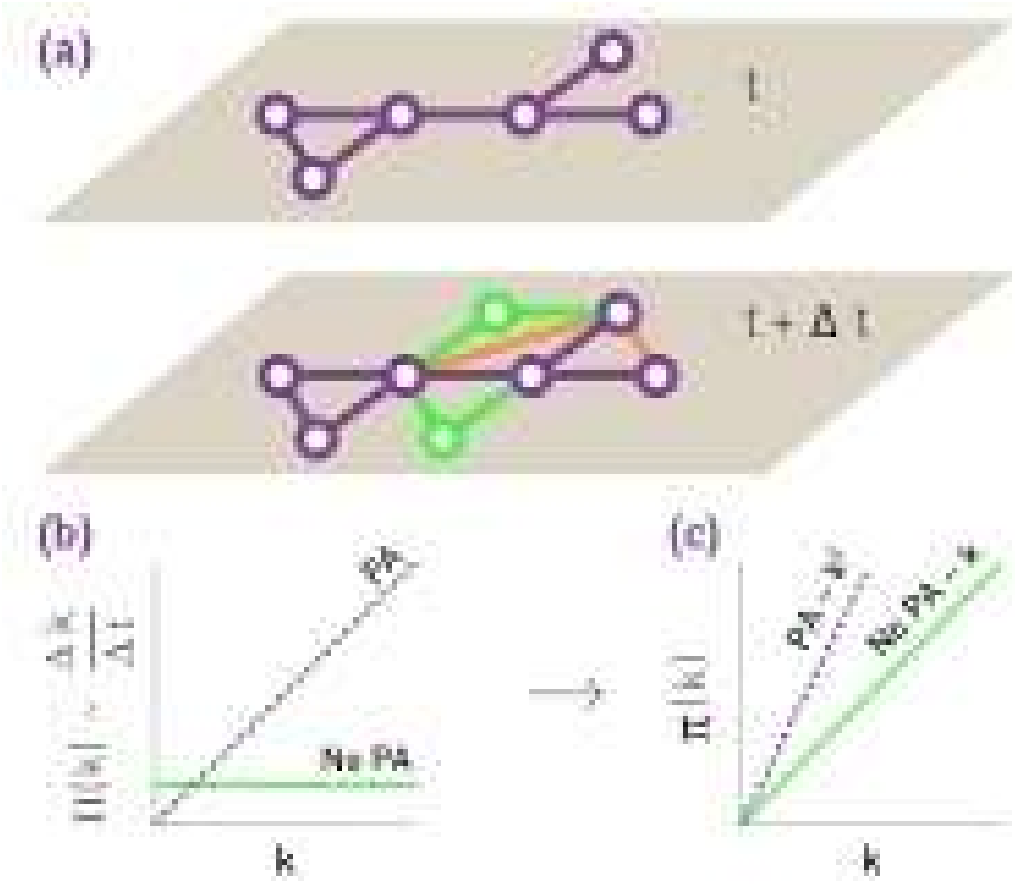
Plot the change in the degree Δk during a fixed time Δt for nodes with degree k.

To reduce noise, plot the integral of $\Pi(k)$ over k:

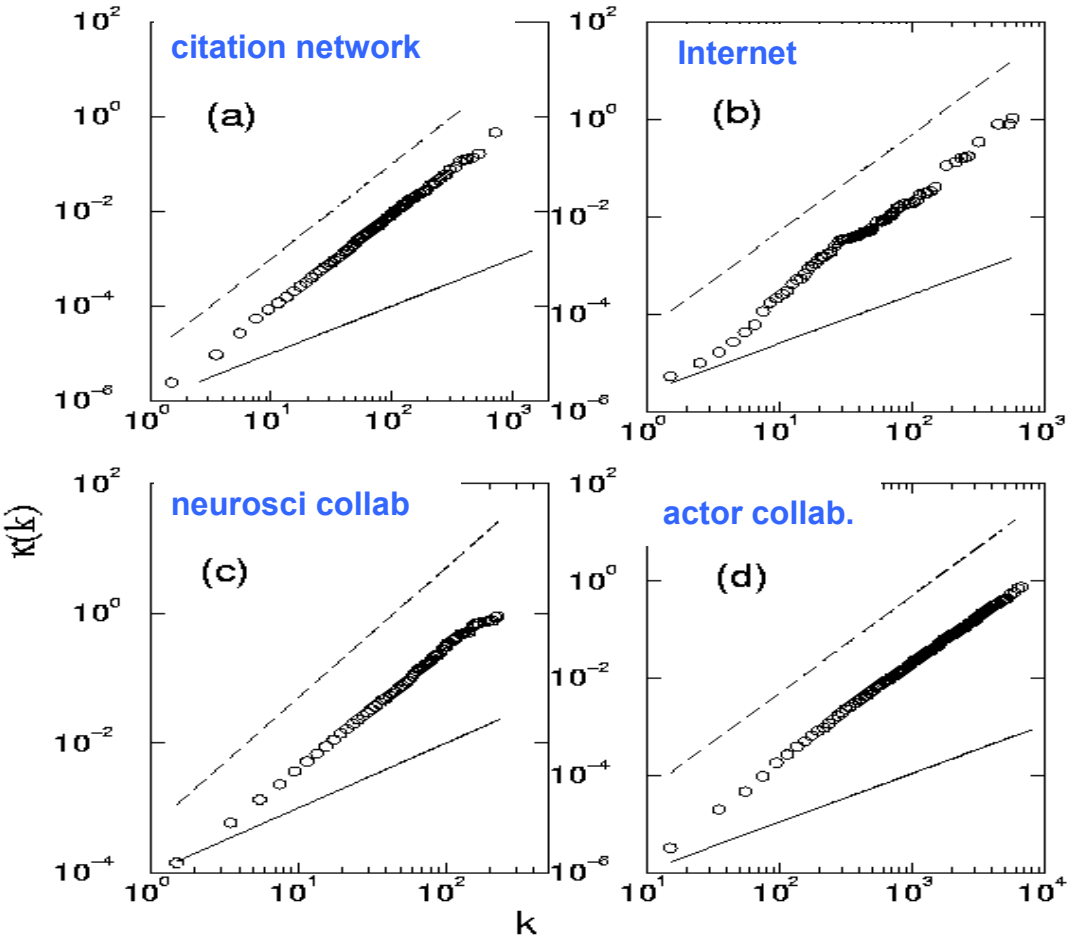
$$\kappa(k) = \sum_{K<k} \Pi(K)$$

No pref. attach:
$\kappa \sim k$ ———

Linear pref. attach:
$\kappa \sim k^2$ - - - - -



(Jeong, Neda, A.-L. B, Europhys Letter 2003; cond-mat/0104131)



Plots shows the integral of Π(k) over k:

$$\kappa(k) = \sum_{K<k} \Pi(K)$$

No pref. attach:
$\kappa \sim k$ ——

Linear pref. attach:
$\kappa \sim k^2$ - - - -

$$\Pi(k) \approx A + k^{\alpha}, \quad \alpha \leq 1$$

Section 8

Nonlinear preferenatial attachment

Section 8 | Nonlinear preferential attachment

$$\Pi(k) \sim k^{\alpha}$$

α=0: Reduces to Model A discussed in Section 5.4. The degree distribution follows the simple exponential function.

α=1: Barabási-Albert model, a scale-free network with degree exponent 3.

0<α<1: Sublinear preferential attachment. New nodes favor the more connected nodes over the less connected nodes. Yet, for the bias is not sufficient to generate a scale-free degree distribution. Instead, in this regime the degrees follow the stretched exponential distribution:

$$p_k \sim k^{-\alpha} \exp\left(\frac{2\mu(\alpha)}{\langle k \rangle (1-\alpha)} k^{1-\alpha} \right)$$

$$k_{\max} \sim (\ln t)^{1/(1-\alpha)}$$

Section 8 Nonlinear preferential attachment

$$\Pi(k) \sim k^{\alpha}$$

α=0: Reduces to Model A discussed in Section 5.4. The degree distribution follows the simple exponential function.

α=1: Barabási-Albert model, a scale-free network with degree exponent 3.

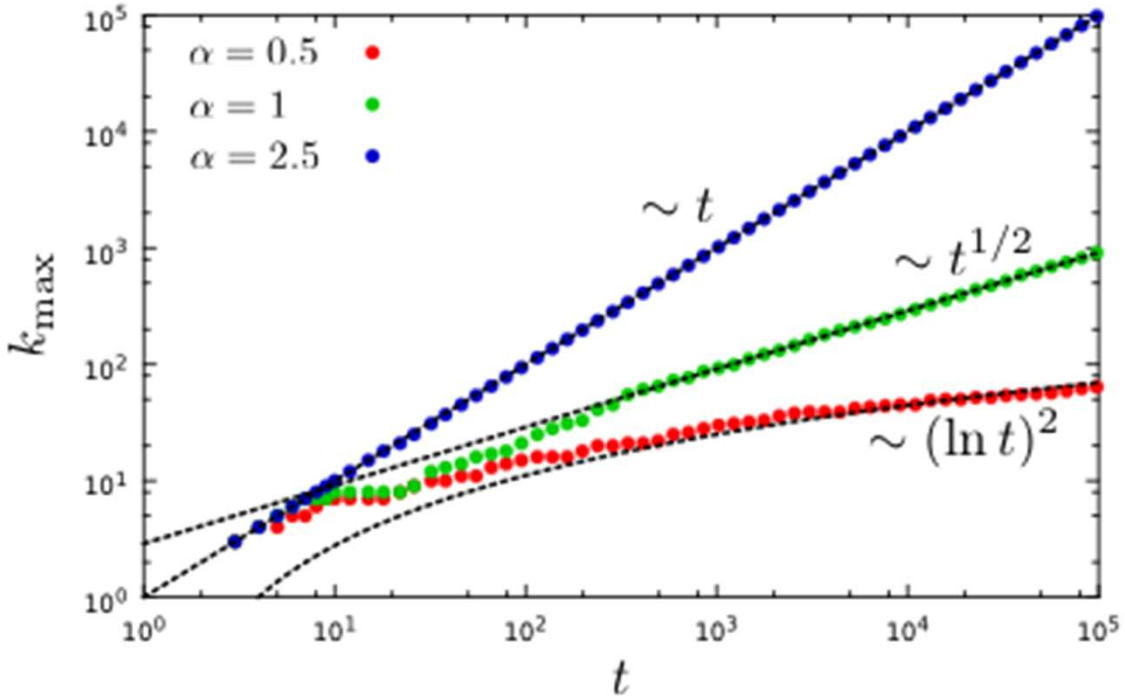
α>1: Superlinear preferential attachment. The tendency to link to highly connected nodes is enhanced, accelerating the "rich-gets-richer" process. The consequence of this is most obvious for , when the model predicts a *winner-takes-all phenomenon: almost all nodes connect to a single or a few super-hubs.*

$$k_{\max} \sim t.$$

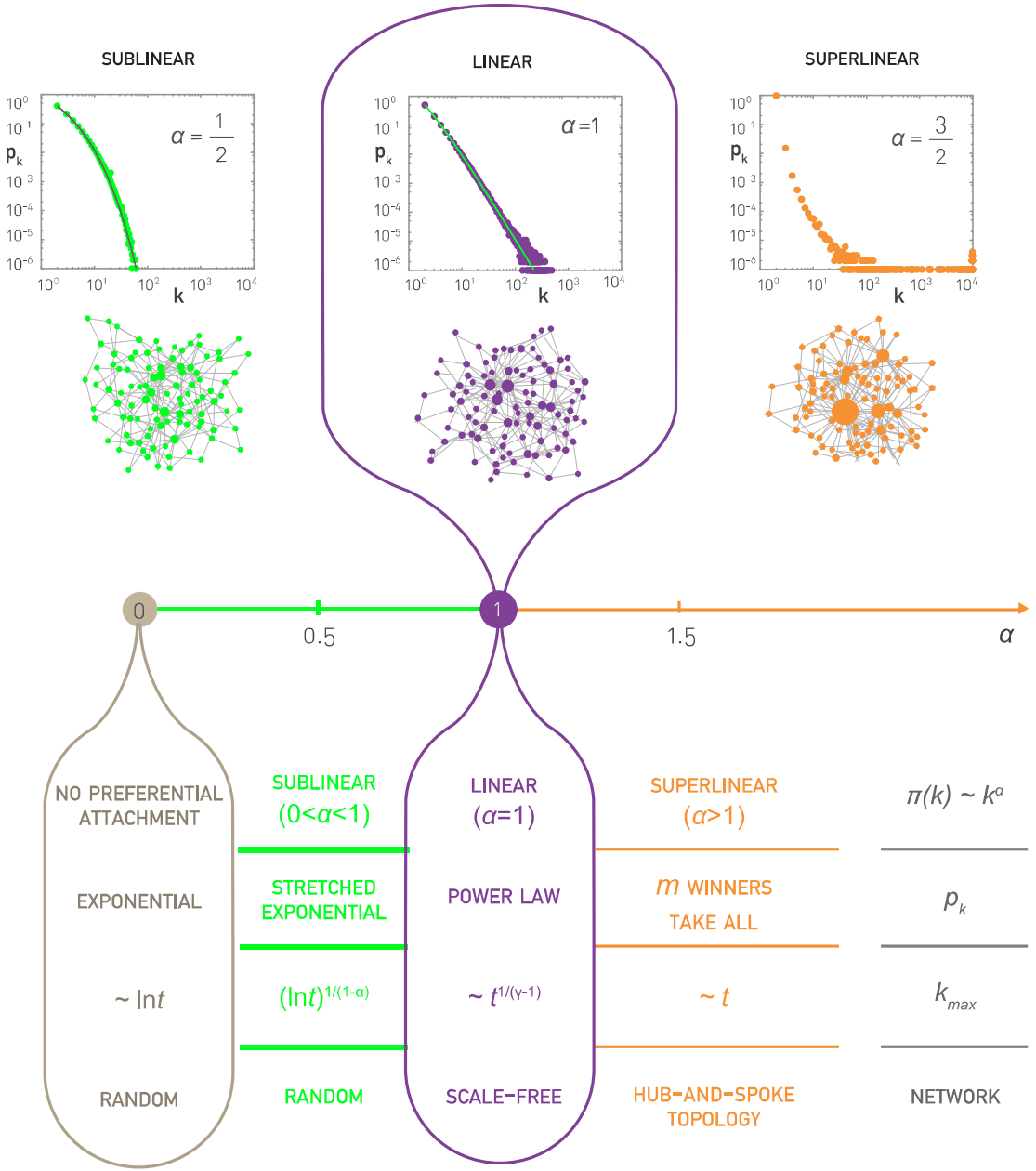
Section 8

Nonlinear preferential attachment

$$\Pi(k) \sim k^{\alpha}$$



The growth of the hubs. The nature of preferential attachment affects the degree of the largest node. While in a scale-free network the biggest hub grows as (green curve), for sublinear preferential attachment this dependence becomes logarithmic (red curve). For superlinear preferential attachment the biggest hub grows linearly with time, always grabbing a finite fraction of all links (blue curve)). The symbols are provided by a numerical simulation; the dotted lines represent the analytical predictions.


SUBLINEAR
LINEAR
SUPERLINEAR
$\alpha = \frac{1}{2}$
$\alpha = 1$
$\alpha = \frac{3}{2}$
p_k
k
0
0.5
1
1.5
α
NO PREFERENTIAL ATTACHMENT
EXPONENTIAL
~ lnt
RANDOM
SUBLINEAR (0<α<1)
STRETCHED EXPONENTIAL
$(\ln t)^{1/(1-\alpha)}$
RANDOM
LINEAR (α=1)
POWER LAW
$\sim t^{1/(\gamma-1)}$
SCALE-FREE
SUPERLINEAR (α>1)
m WINNERS TAKE ALL
~ t
HUB-AND-SPOKE TOPOLOGY
$\pi(k) \sim k^{\alpha}$
p_k
k_{max}
NETWORK

Section 11: Summary

Number of Nodes
$N = t$

Number of Links
$N = mt$

Average Degree
$\langle k \rangle = 2m$

Degree Dynamics
$k_i(t) = m\,(t/t_i)^{\beta}$

Dynamical Exponent
$\beta = 1/2$

Degree Distribution
$p_k \sim k^{\gamma}$

Degree Exponent
$\gamma = 3$

Average Distance
$\langle d \rangle \sim \log N / \log\log N$

Clustering Coefficient
$\langle C \rangle \sim (\ln N)^2 / N$

- The model predicts γ=3 while the degree exponent of real networks varies between 2 and 5 (Table 4.2).

- Many networks, like the WWW or citation networks, are directed, while the model generates undirected networks.

- Many processes observed in networks, from linking to already existing nodes to the disappearance of links and nodes, are absent from the model.

- The model does not allow us to distinguish between nodes based on some intrinsic characteristics, like the novelty of a research paper or the utility of a webpage.

- While the Barabási-Albert model is occasionally used as a model of the Internet or the cell, in reality it is not designed to capture the details of any particular real network. It is a minimal, proof of principle model whose main purpose is to capture the basic mechanisms responsible for the emergence of the scale-free property. Therefore, if we want to understand the evolution of systems like the Internet, the cell or the WWW, we need to incorporate the important details that contribute to the time evolution of these systems, like the directed nature of the WWW, the possibility of internal links and node and link removal.

Science and Technology Facilities Council

Hartree Centre

Examples

Science and Technology Facilities Council

Hartree Centre

Science and Technology Facilities Council

Hartree Centre

Psychological Warfare Analysis

Ilya Blokh and Vassil Alexandrov

Science and Technology Facilities Council

Hartree Centre

Psychological Warfare

The demonstration of multi-directed regular actions aiming at influencing the social opinion in certain areas and consisting of mass media publication's policy, in an Internet setting in particular.

Science and Technology Facilities Council
Hartree Centre

Ilya Blokh and Vassil Alexandrov, 2016

Research Questions

1. What are the laws of information spreading in situation of PW? How it influences people?
2. How propaganda ideas evolve and spread in the Internet?
3. What factors could influence the spreading dynamics of such ideas?
4. How to find and reveal hidden connections between published news and facts from different sources? What kinds of hidden connections could we find?

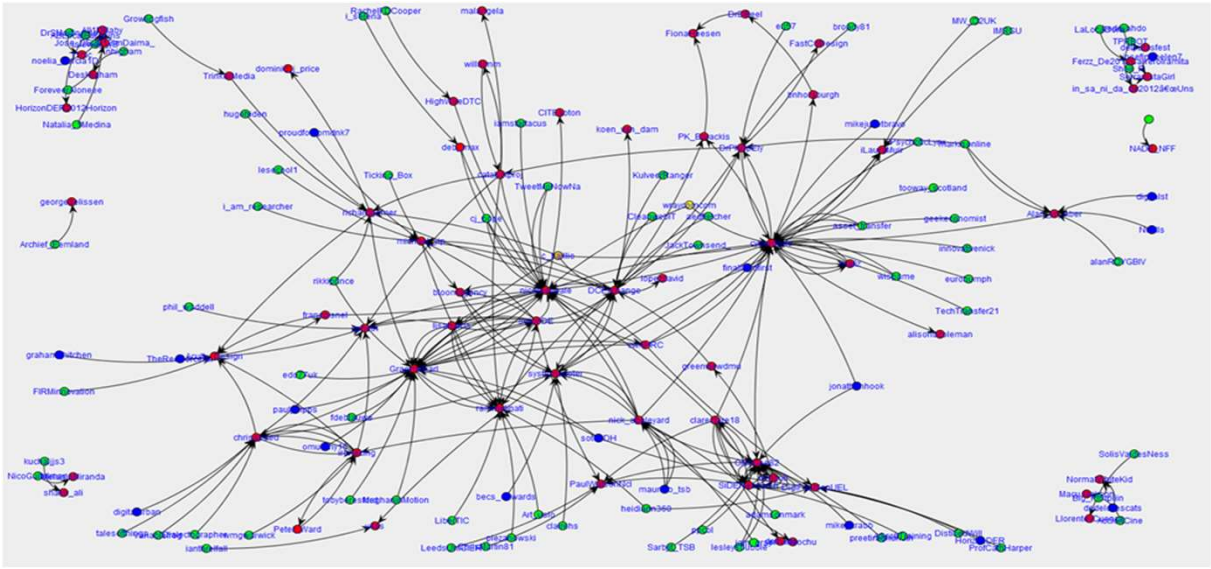
Network Science Approach

Analyze data from social networks -> network science is convenient tool

Twitter network - one of the most popular social networking platforms for news publishing and distribution

Twitter's hashtag mechanism could be a natural way for keyword extraction.

Discovering Most Significant News

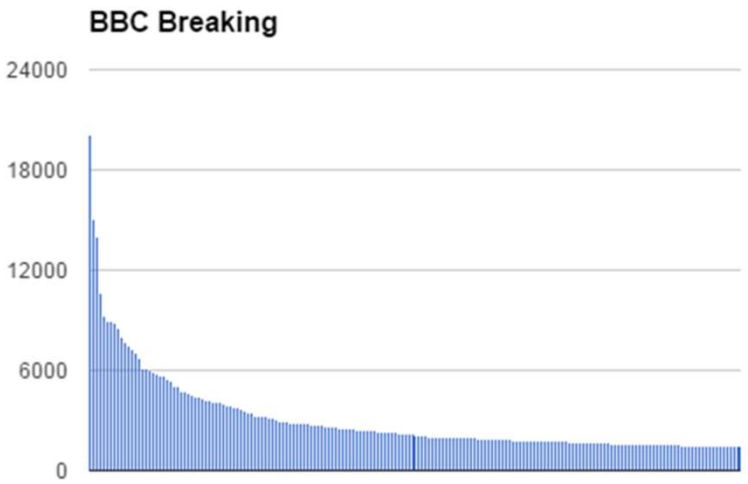


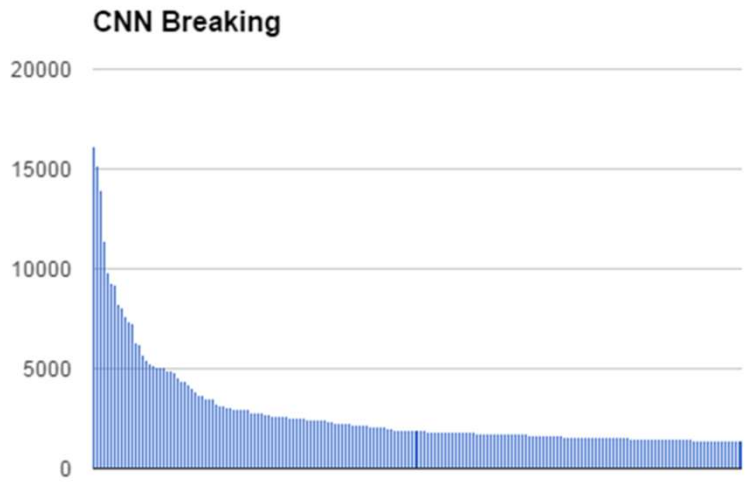
Twitter network

Science and Technology Facilities Council
Hartree Centre

Ilya Blokh and Vassil Alexandrov, 2016

Discovering Most Significant News





Science and Technology Facilities Council
Hartree Centre

Ilya Blokh and Vassil Alexandrov, 2016

What we do?

1. Define observation subject.
2. Define research geography (countries, languages).
3. Define keywords.
4. Define time period.
5. Extract data from Twitter and use Network Science methods construct the network that embraces selected theme.
6. Analyze and compare the data spreading in different parts of chosen geographical regions.
7. Use additional analysis to understand the news ideological affiliation with some of the present sides in PW.

The algorithm

Twitter hashtags = keywords

Keywords -> clusters of keywords

1 cluster = some point of view

⇩

Using hierarchical agglomerative clustering algorithm

$T = \{t_1, t_2, \ldots t_s\}$ - set of tweets

$K = \{k_1, k_2, \ldots k_n\}$ - keywords

α_i - quantity of keyword k_i appearances in tweets

α_{ij}- quantity of simultaneous entries of keywords k_i and k_j

$\omega_{ij} = \frac{\alpha_{ij}}{\alpha_j}$ - relative rate of keyword k_i concerning keyword k_j

metric for clustering: $\overline{\omega_{ij}} = \overline{\omega_{ji}} = (\omega_{ij} + \omega_{ji})/2$

proximity threshold L:

If $\overline{\omega_{ij}} > L$ we put k_i and k_j in one cluster

Ilya Blokh and Vassil Alexandrov, 2016

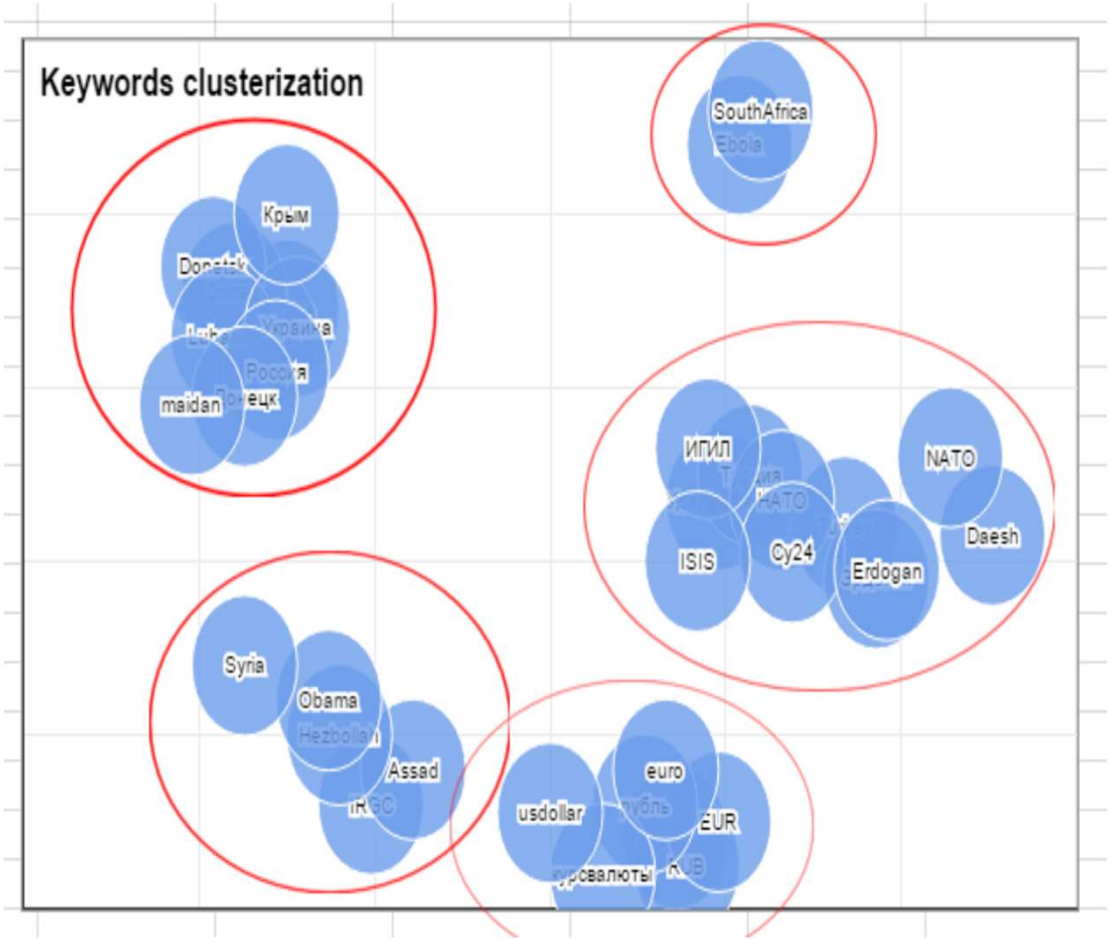
Keywords processing

Covering themes:

- events in Ukraine and Crimea
- discord between Russia and Turkey
- war in Syria
- Russian national currency collapse
- Ebola epidemic.

1.04.15 – 31.12.15

300 000 tweets

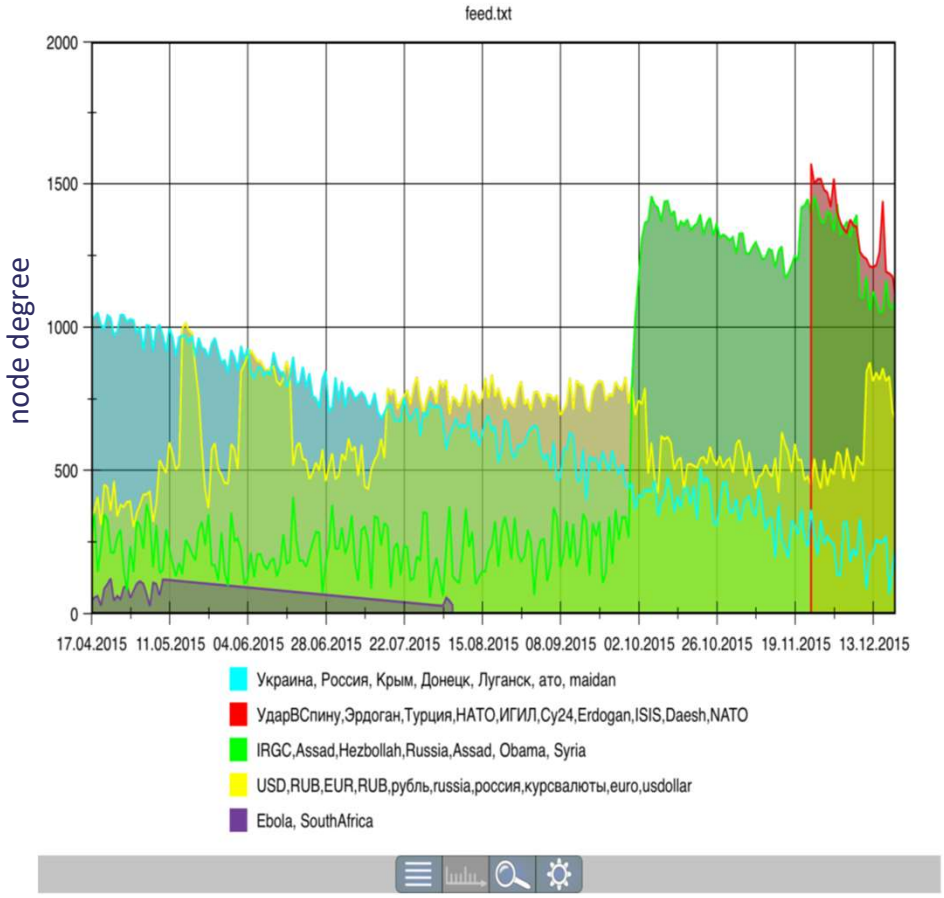


Ilya Blokh and Vassil Alexandrov, 2016

Science and Technology Facilities Council
Hartree Centre

Data analysis using Network Science

Node degree distribution of keyword clusters



Science and Technology Facilities Council

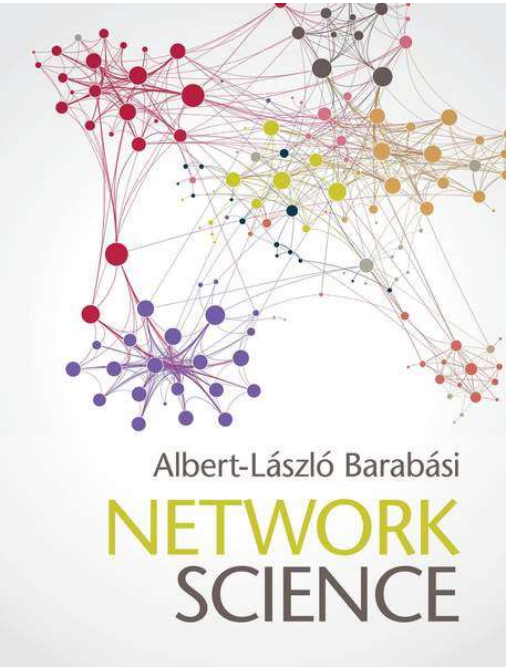
Hartree Centre

Ilya Blokh and Vassil Alexandrov, 2016

Hartree Centre

Resources

Recommended Reading



www.BarabasiLab.com https://slideplayer.com/slide/3218489/

http://snap.stanford.edu/class/cs224w-2015/handouts.html

Guido Caldarelli and Alessandro Chessa, Data Science and Complex Networks Real Case Studies with Python, Oxford University Press, 2016

Science and Technology Facilities Council

Hartree Centre

Questions?

Science and Technology Facilities Council

Hartree Centre

Thank you

vassil.alexandrov@stfc.ac.uk