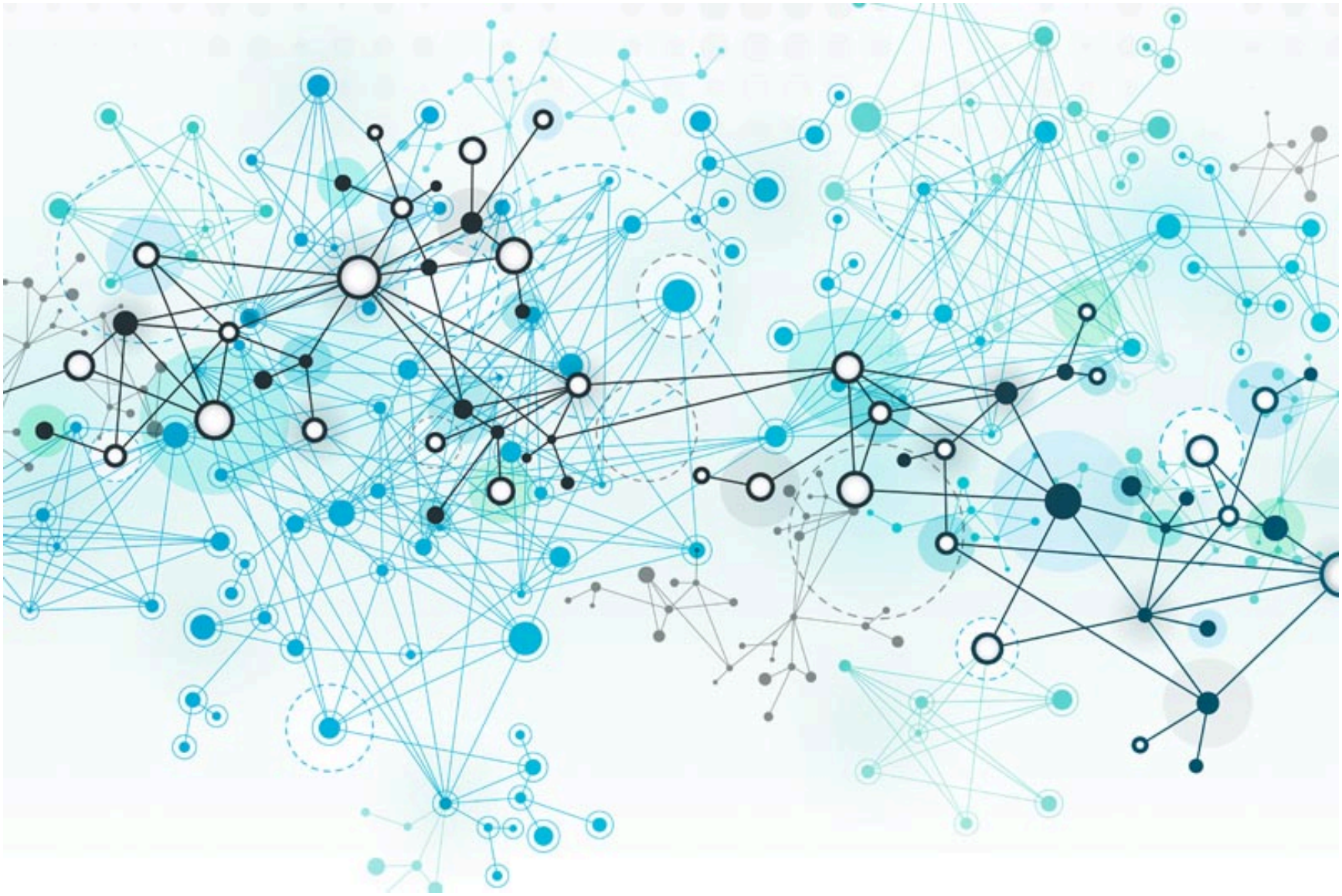
BIG NETWORK DATA
ECE 289 UC San Diego



Where we are

Random graphs

Random geometric graphs

Random connection models

Full and alpha-Connectivity, crossing paths

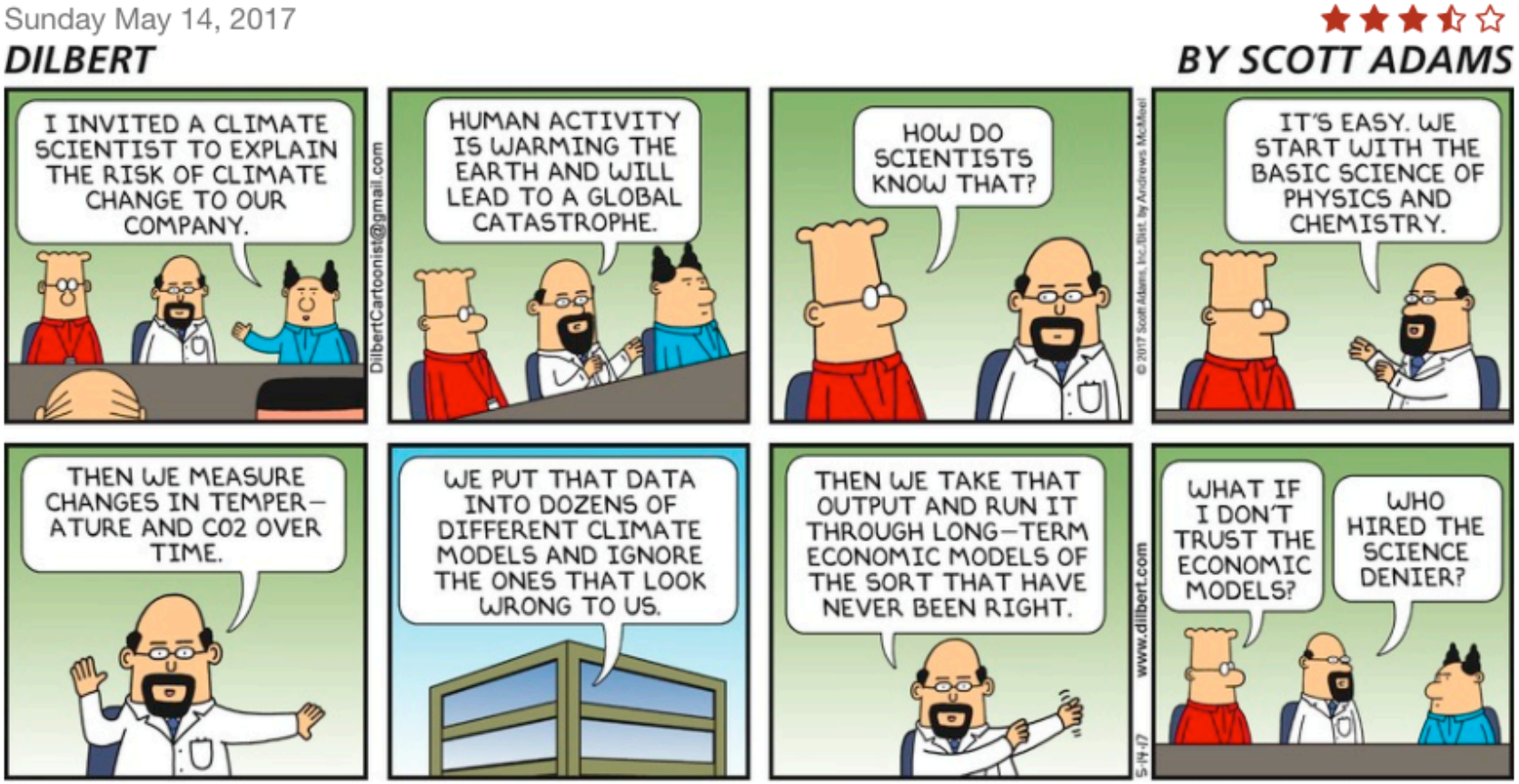
Small-worlds and network navigability

The importance of scale-invariance

We can come up with many models with interesting properties…

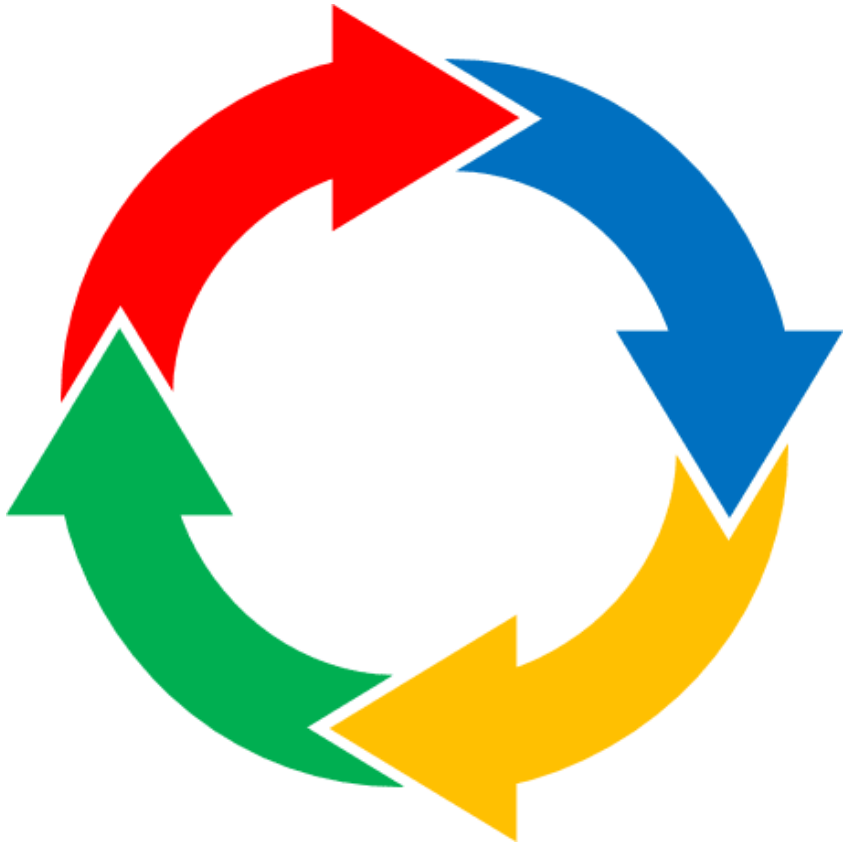
What is a a good model?

Good models


Sunday May 14, 2017
DILBERT
BY SCOTT ADAMS
I INVITED A CLIMATE SCIENTIST TO EXPLAIN THE RISK OF CLIMATE CHANGE TO OUR COMPANY.
HUMAN ACTIVITY IS WARMING THE EARTH AND WILL LEAD TO A GLOBAL CATASTROPHE.
HOW DO SCIENTISTS KNOW THAT?
IT'S EASY. WE START WITH THE BASIC SCIENCE OF PHYSICS AND CHEMISTRY.
THEN WE MEASURE CHANGES IN TEMPER-ATURE AND CO2 OVER TIME.
WE PUT THAT DATA INTO DOZENS OF DIFFERENT CLIMATE MODELS AND IGNORE THE ONES THAT LOOK WRONG TO US.
THEN WE TAKE THAT OUTPUT AND RUN IT THROUGH LONG-TERM ECONOMIC MODELS OF THE SORT THAT HAVE NEVER BEEN RIGHT.
WHAT IF I DON'T TRUST THE ECONOMIC MODELS?
WHO HIRED THE SCIENCE DENIER?
DilbertCartoonist@gmail.com
© 2017 Scott Adams, Inc./Dist. by Andrews McMeel
www.dilbert.com
5-14-17

Network Science Approach

- Observation: from real data
- Analysis: node centrality, degree distribution, clustering, diameter, modularity classes, connectivity, component sizes, dynamic processes
- Generative models
- Validation
- Prediction



Good Models

I argue that a good model should be data-driven

Don’t look for data to support your favorite model

There may be competing models that better explain additional features of the data

Example: Power-law degree distributions

may be just one aspect apparent from the data, but any model producing a power law might not be the right one!

Preferential attachment

- Yule (1925): number of species per genus of flowering plants
- Simon (1955): proposed it as a fundamental mechanism of complexity
- Price (1976): citation networks
- Albert, Barabasi (1999): made it extremely popular in recent times

Biometrika, Vol. 42, No. 3/4 (Dec., 1955), pp. 425-440

ON A CLASS OF SKEW DISTRIBUTION FUNCTIONS

By HERBERT A. SIMON†

Carnegie Institute of Technology

I. Introduction

It is the purpose of this paper to analyse a class of distribution functions that appears in a wide range of empirical data—particularly data describing sociological, biological and economic phenomena. Its appearance is so frequent, and the phenomena in which it appears so diverse, that one is led to the conjecture that if these phenomena have any property in common it can only be a similarity in the structure of the underlying probability mechanisms. The empirical distributions to which we shall refer specifically are: (A) distributions of words in prose samples by their frequency of occurrence, (B) distributions of scientists by number of papers published, (C) distributions of cities by population, (D) distributions of incomes by size, and (E) distributions of biological genera by number of species.

No one supposes that there is any connexion between horse-kicks suffered by soldiers in the German army and blood cells on a microscope slide other than that the same urn scheme provides a satisfactory abstract model of both phenomena. It is in the same direction that we shall look for an explanation of the observed close similarities among the five classes of distributions listed above.

Preferential attachment

$$t = 1 \implies 2 \text{ nodes } m \text{ edges}$$

$$t \geq 2 \implies \text{new node having } m \text{ new edges}$$

$$PA \implies \text{connect one edge at the time}$$

$$p_i = \frac{k_i}{\sum_j k_j}$$

$$\lim_{n \to \infty} P(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \sim k^{-3}$$

The model can also be made more general to obtain a power law degree distribution of any power in (2,infinity)

Preferential attachment

Does preferential attachment really occurs?

It has been proposed to explain power laws in WWW, Internet, collaboration networks, sexual partner networks, protein networks…

"What do the proteins in our bodies, the Internet, a cool collection of atoms and sexual networks have in common? One man thinks he has the answer and it is going to transform the way we view the world."

13 April 2002 issue of *The Scientist*

In reality they have very little in common. If you carefully look at the data, there are fundamental differences that cannot be explained using a single model.


THE RICH GET RICHER, THE POOR GET POORER.
AND THE COMFORTABLY OFF STAY COMFORTABLY OFF!
Search:.07635319

Preferential attachment

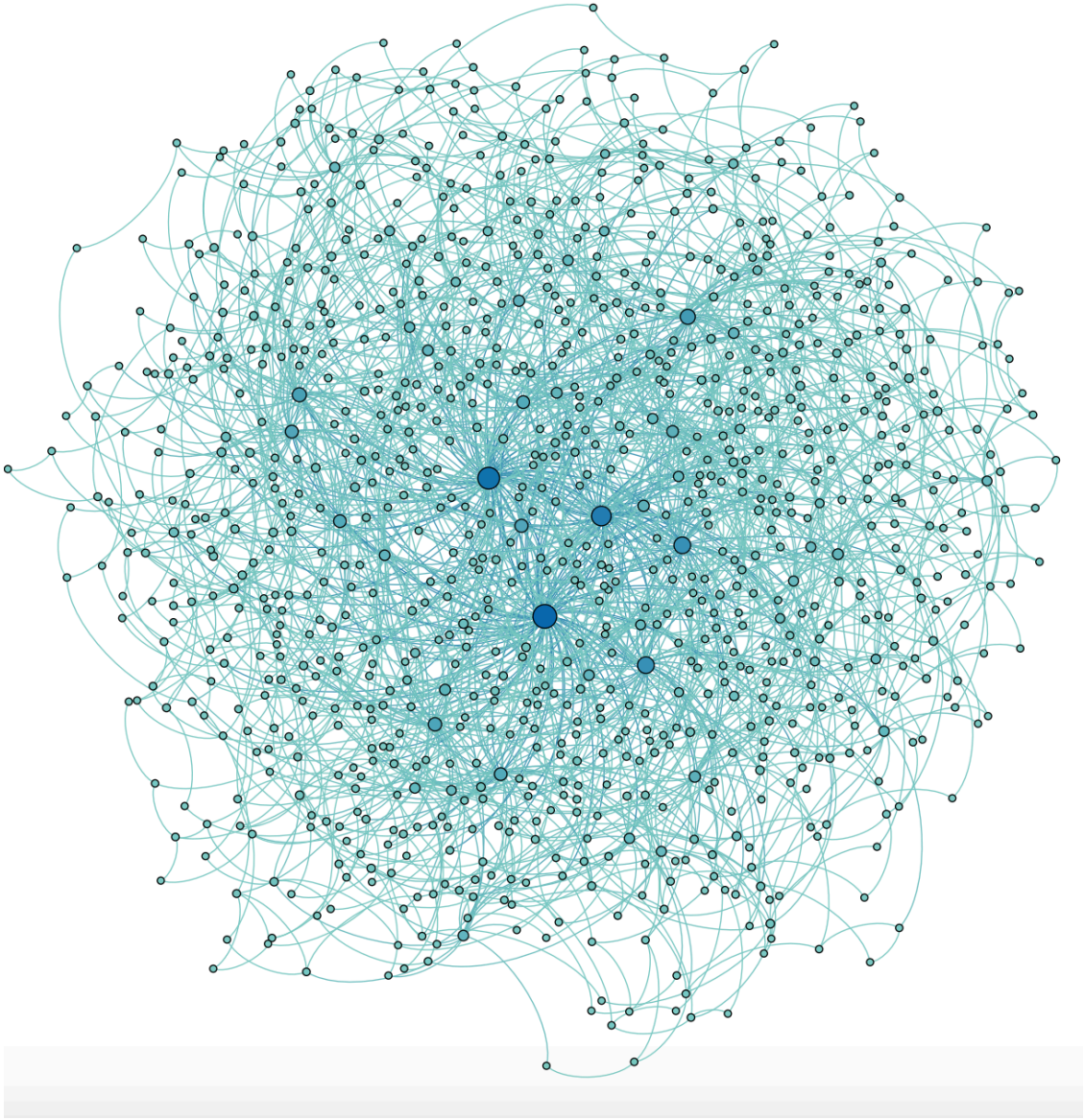
Properties of preferential attachment:

Power law degree distribution

Small diameter $\sim \frac{\log n}{\log \log n}$

Emergence of "hubs": these highly connected nodes appear to be at the core of the network, hence, the network is robust to random attacks but vulnerable to malicious attacks

Preferential attachment “hubs”



Preferential attachment “hubs”

Preferential attachment “hubs”

Removing 95% of the links makes “little damage”

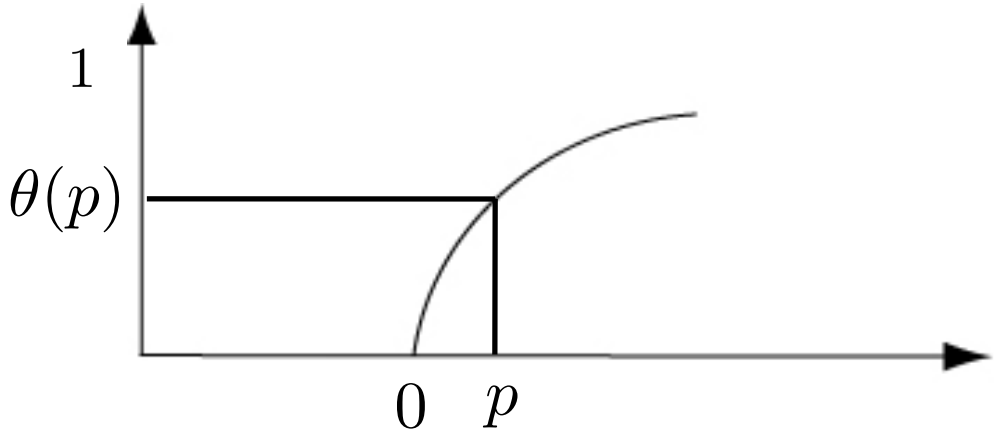
Removing 2% of the hubs breaks the network in a multitude of small components

Robust yet fragile?

A more relevant question regards the size of the giant component

Assume to perform independent thinning of the graph, keeping each edge with probability p and deleting it with probability (1-p)

Bollobas and Riordan (2004)



$$\exp[-c/p^2] < \theta(p) < \exp[-c'/p]$$

$$\forall\, 0 < p \leq 1 \quad |C_n| = \theta(p)n + o(n) \quad \text{w.h.p}$$

Robust yet fragile?

The critical percolation threshold is close to zero

From our previous lectures we know what this implies:

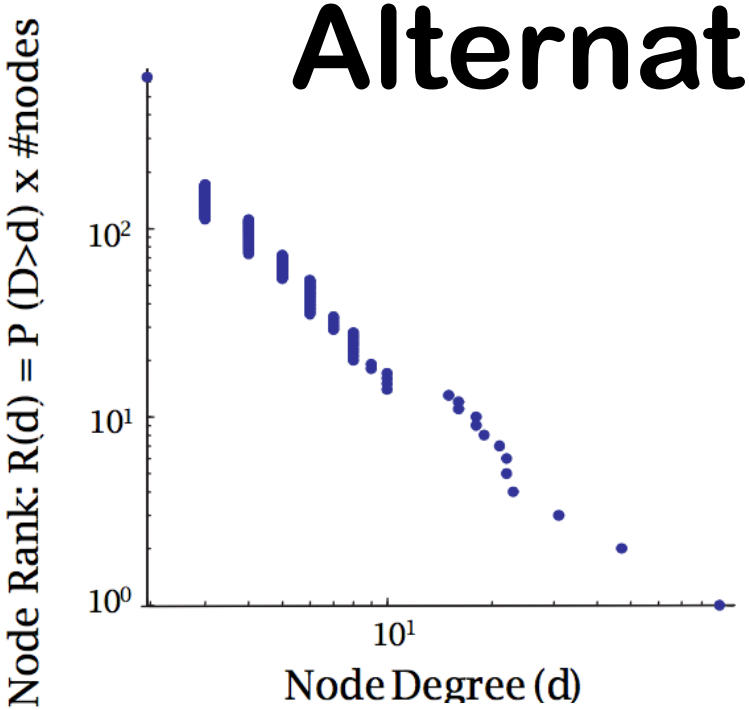
If p is very small the giant component is tiny!

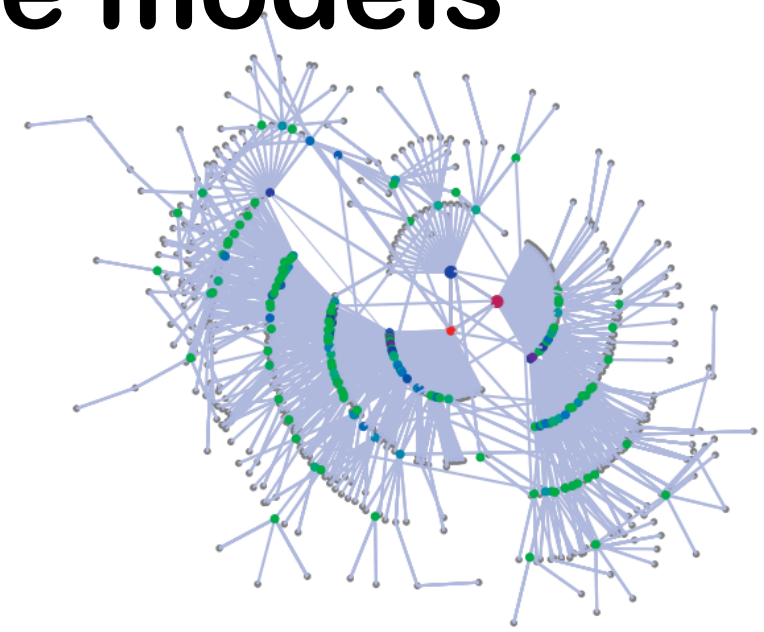
If billions of nodes are connected initially, removing 95% of the links leaves only a handful of nodes connected.

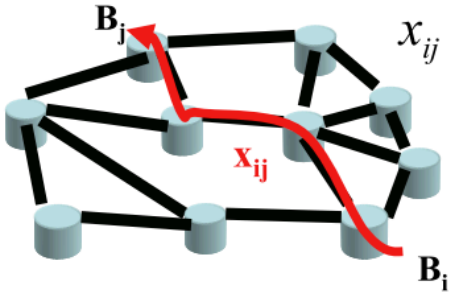
A similar detailed analysis reveals that a considerable fraction of the hubs need to be removed in practice and this result is also very dependent on the details of the model

It is the model that is not robust!

Alternative models





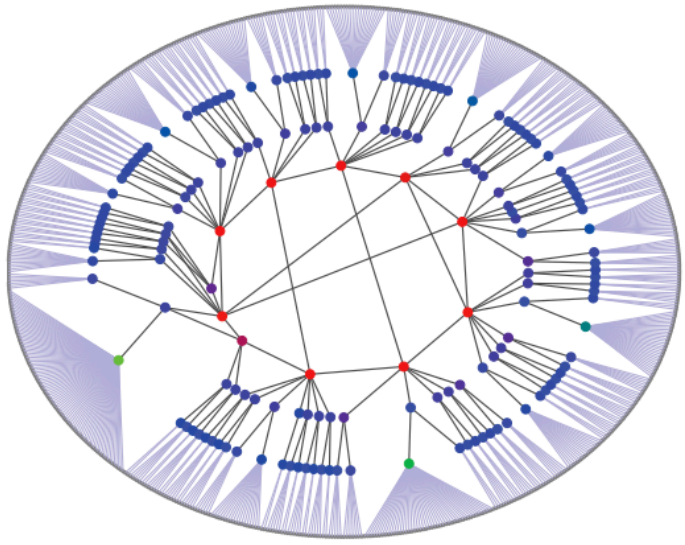


$$x_{ij} \propto B_i B_j$$

$$\max_{\alpha} \sum_{i,j} x_{ij} = \max \sum_{i,j} \alpha B_i B_j$$

$$s.t. \sum_{i,j:k \in r_{ij}} x_{ij} \le B_k, \forall k$$

(a)



(b)

Figure 5. Generating networks using constrained optimization. (a) Engineers view netw

Alternative models

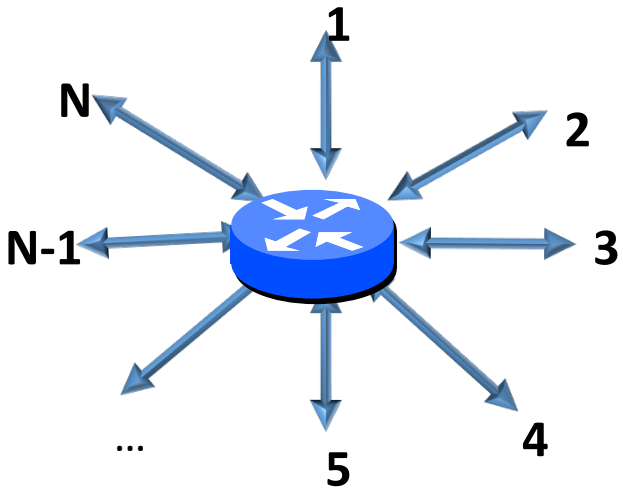
Alternative models must capture more details of the real system

In the case of the internet is clear that the backbone cannot be composed of highly connected hubs

We cannot build fast switches serving a large number of ports

Routers capacity

- N = number of external router “ports”
- R = speed (“line rate”) of a port
- Router capacity = N x R



Routers capacity

Juniper T4000

- R= 10/40 Gbps
- NR = 4 Tbps

Cisco CRS

- R=10/40/100 Gbps
- NR = 322 Tbps

72 racks, 1MW

Routers capacity

Cisco ASR 1006

- R=1/10 Gbps
- NR = 40 Gbps

Juniper M120

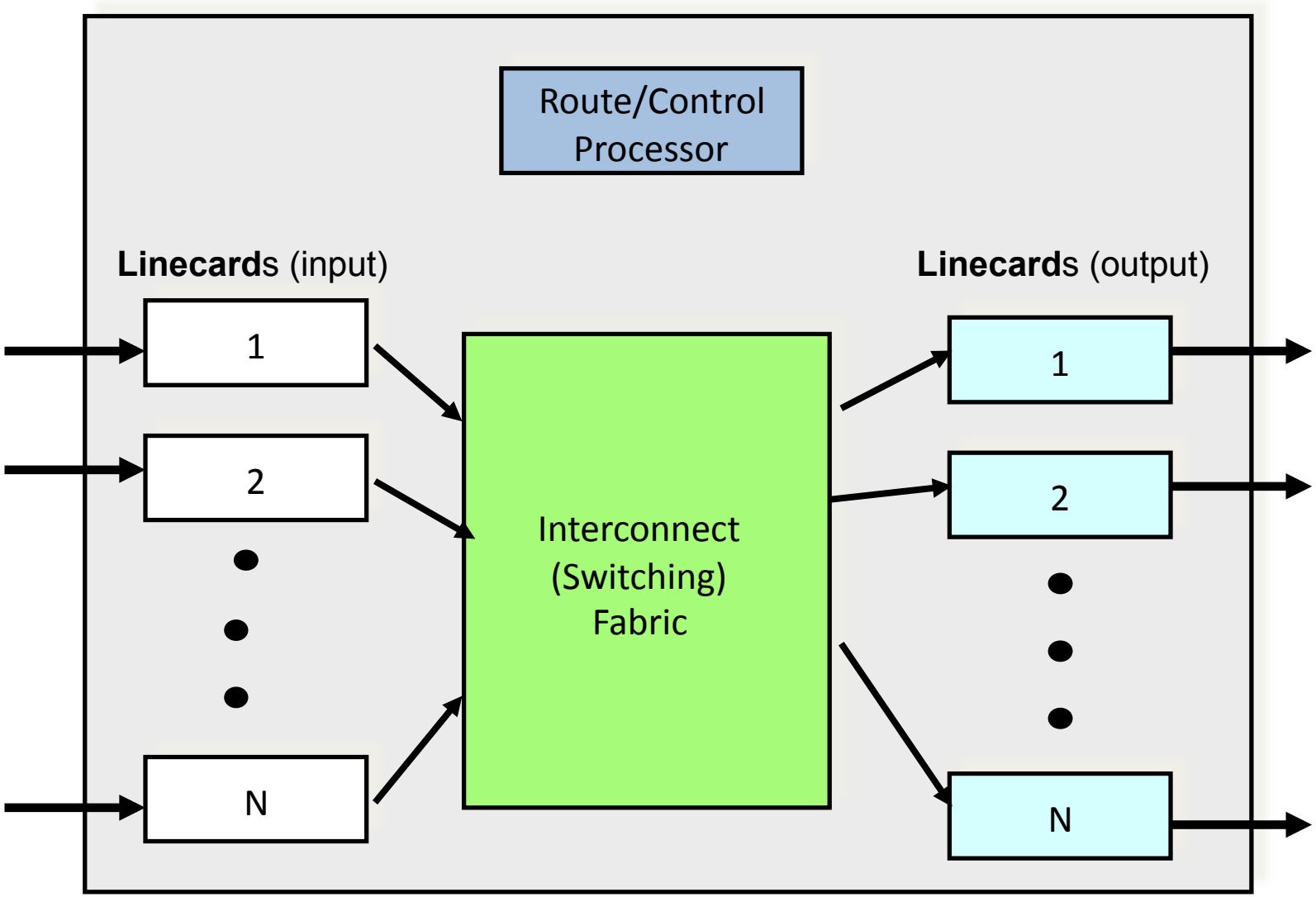
- R= 2.5/10 Gbps
- NR = 120 Gbps

Routers capacity

Cisco 3945E

- R = 10/100/1000 Mbps
- NR < 10 Gbps

Routers capacity



Constrained optimization models

ISP exploit traffic aggregation

Many links with small bandwidth

A few links with large bandwidth

The real architecture arises as the solution to a constrained optimization problem

We will examine generative graph models based on constrained optimization next

The Internet: 2002