

“Lies, Damned Lies, and Statistics:”

**Preferential Attachment-type Network Models
of the Internet**

Walter Willinger

AT&T Labs-Research

walter@research.att.com

Acknowledgments

Main Collaborators

- John Doyle (Caltech)
- David Alderson (Caltech)
- Lun Li (Caltech)

Contributions

- Matt Roughan (U. Adelaide, Australia)
- Steven Low (Caltech)
- Ramesh Govindan (USC)
- Reiko Tanaka (RIKEN, Japan)
- Stanislav Shalunov (Abilene)
- Heather Sherman (CENIC)

Recap: What “Network Science” says about the Internet

- Measurements
 - Router-level: large-scale traceroute experiments
 - AS-level: BGP-based, traceroute-based, WHOIS
 - WWW: large-scale web crawling experiments
- Inference
 - (Exclusive) focus on **node degree distribution**
 - Inferred node degree distributions follow a **power law**
- Modeling
 - Preferential attachment-type growth model
 - **Incremental growth**
 - **Preferential attachment:** $p(k) \approx \text{degree of node } k$
 - There exist many variants of this basic PA model

Recap: What “Network Science” says about the Internet (cont)

- Key features of PA-type models
 - Randomness enters via attachment mechanism
 - Exhibit power law node degree distributions with or without exponential cutoffs
- Model validation
 - The model “fits the data ...”
 - Reproduces observed node degree distribution
- Highly publicized claims about Internet topology
 - High-degree nodes form a hub-like core
 - Fragile/vulnerable to targeted node removal
 - Achilles’ heel
 - Zero epidemic threshold

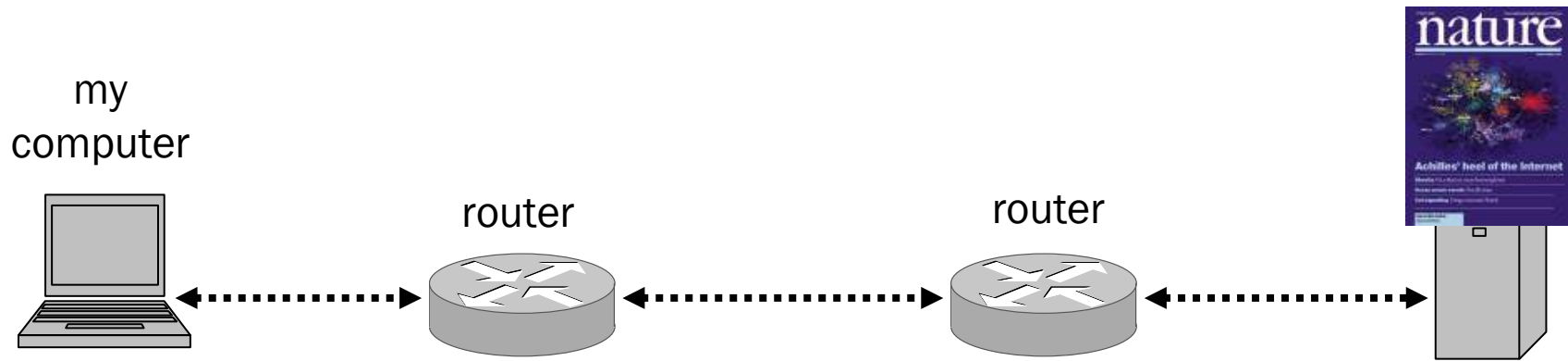
Basic Question

Do the available Internet-related connectivity measurements and their analysis support the sort of claims that can be found in the existing complex networks literature?

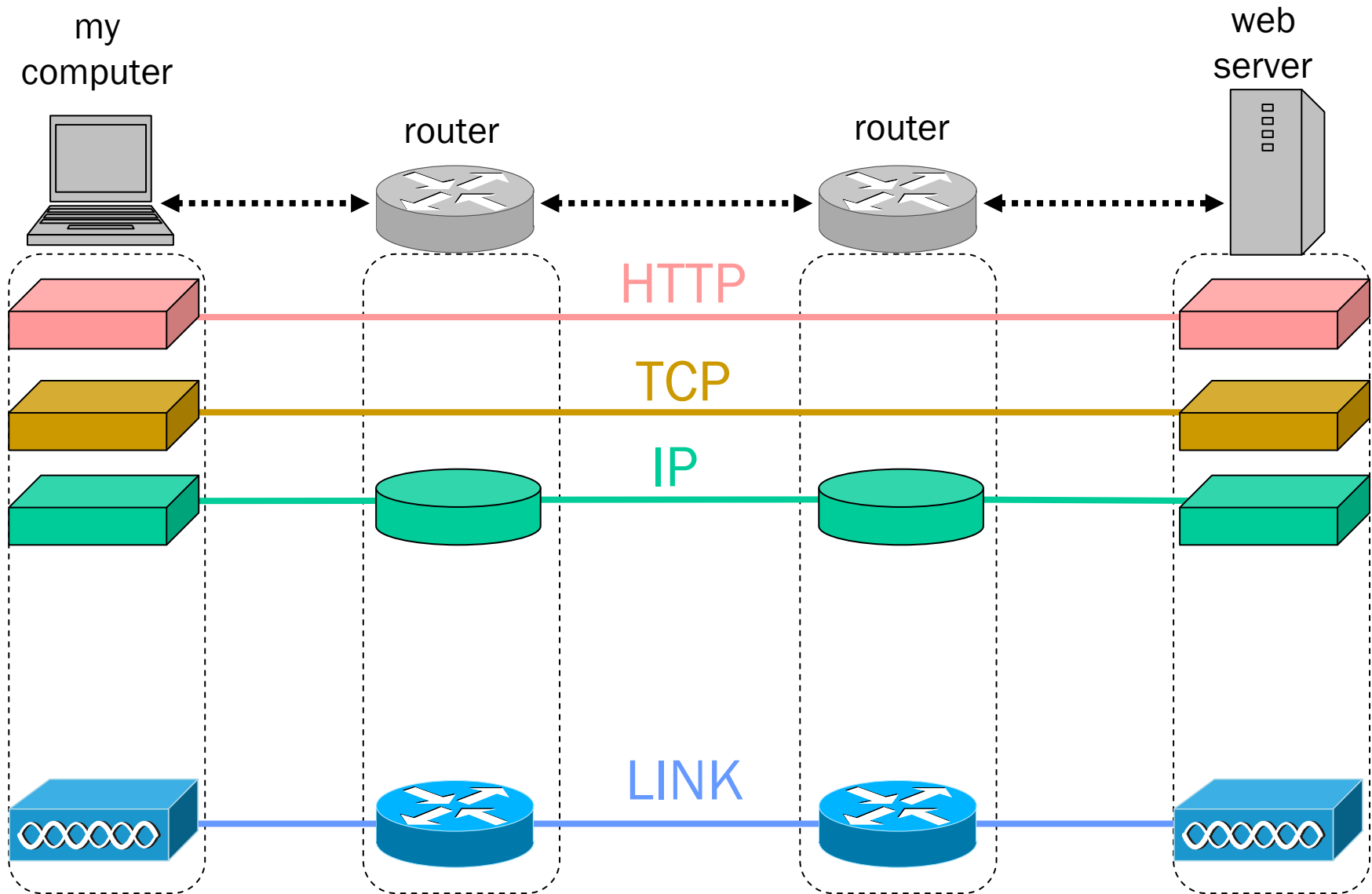
Key Issues

- What about data hygiene?
- What about statistical rigor?
- What about model validation?

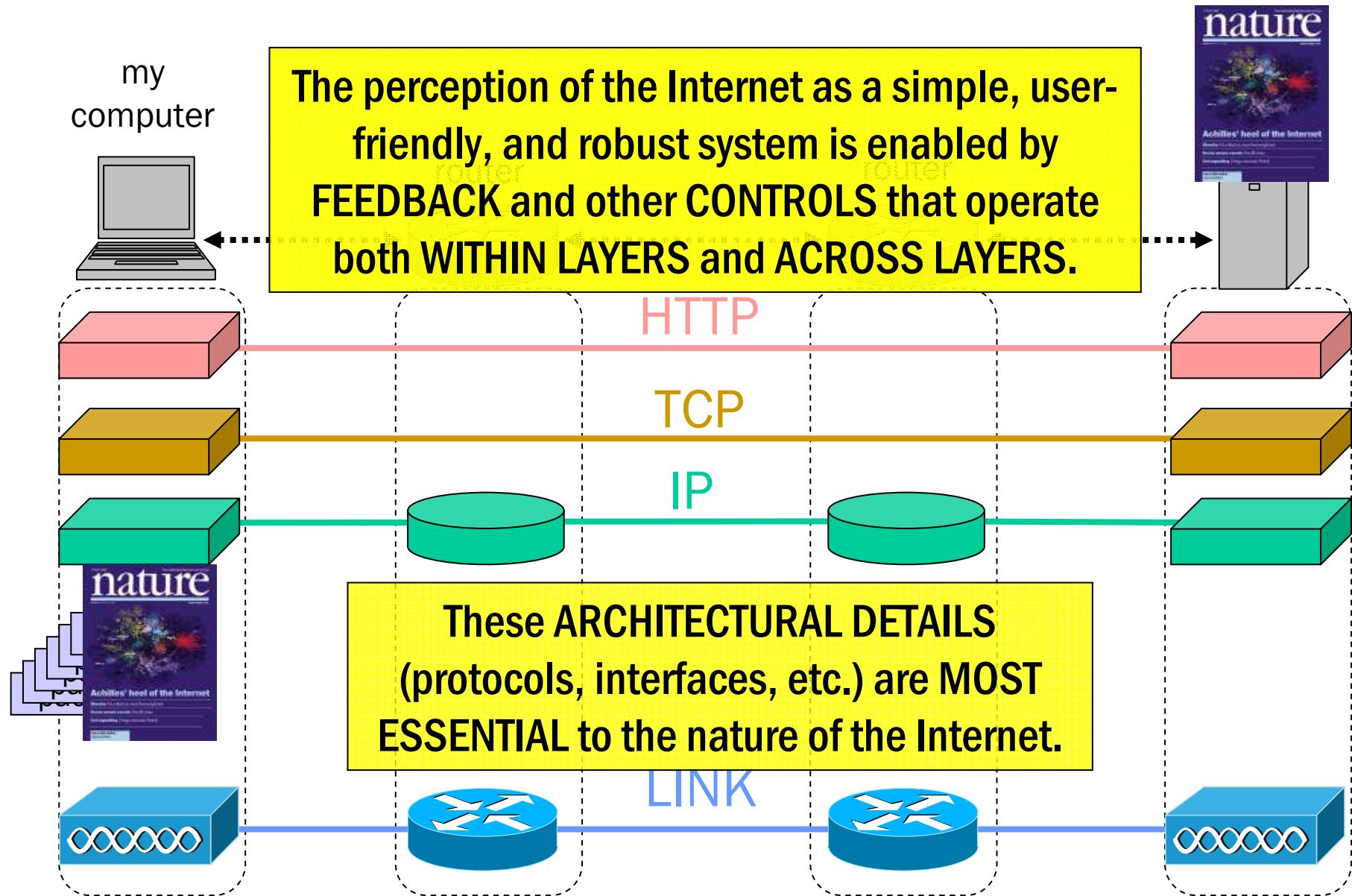
The Internet: The User Perspective



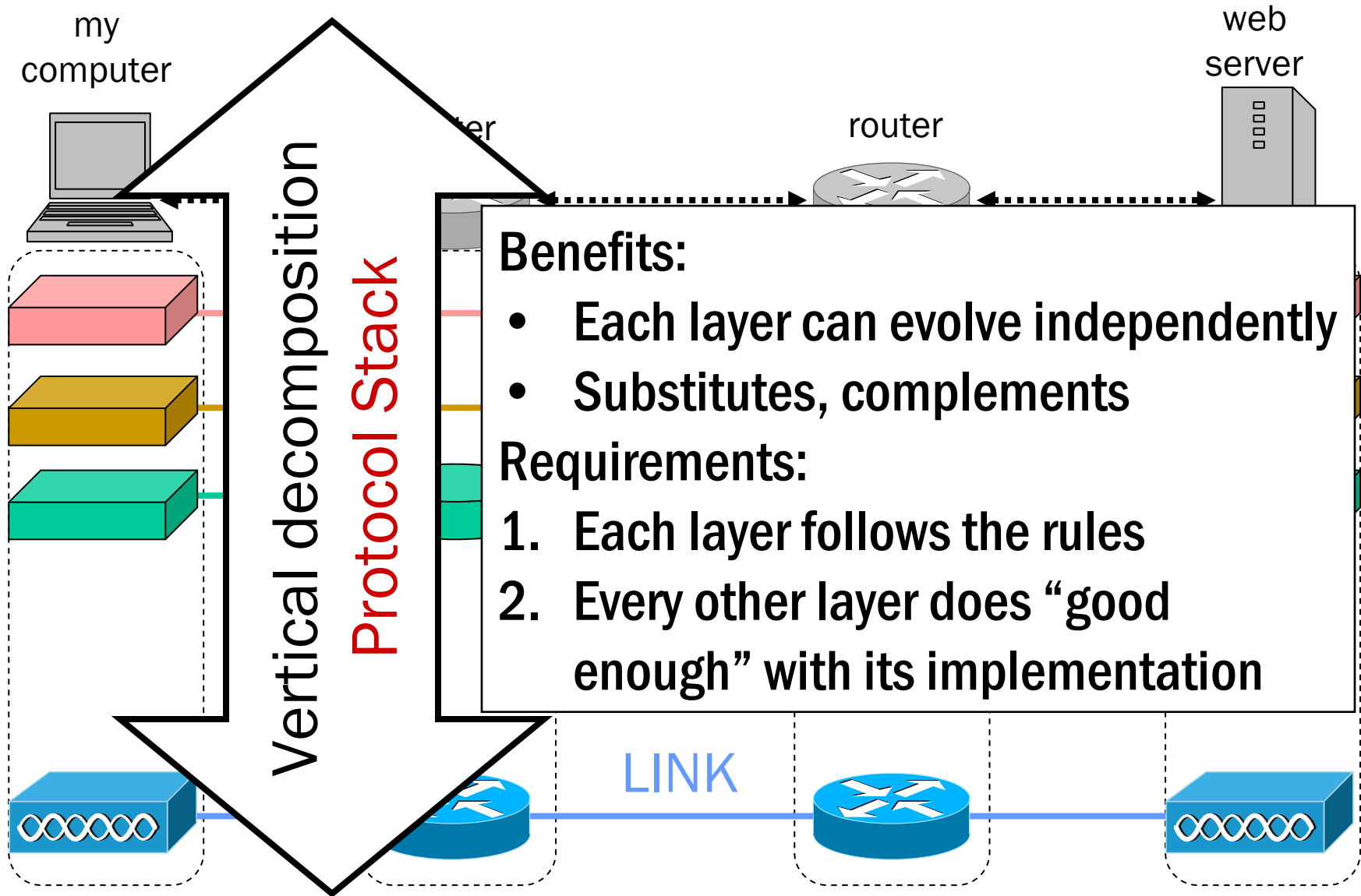
The Internet: The Engineering Perspective



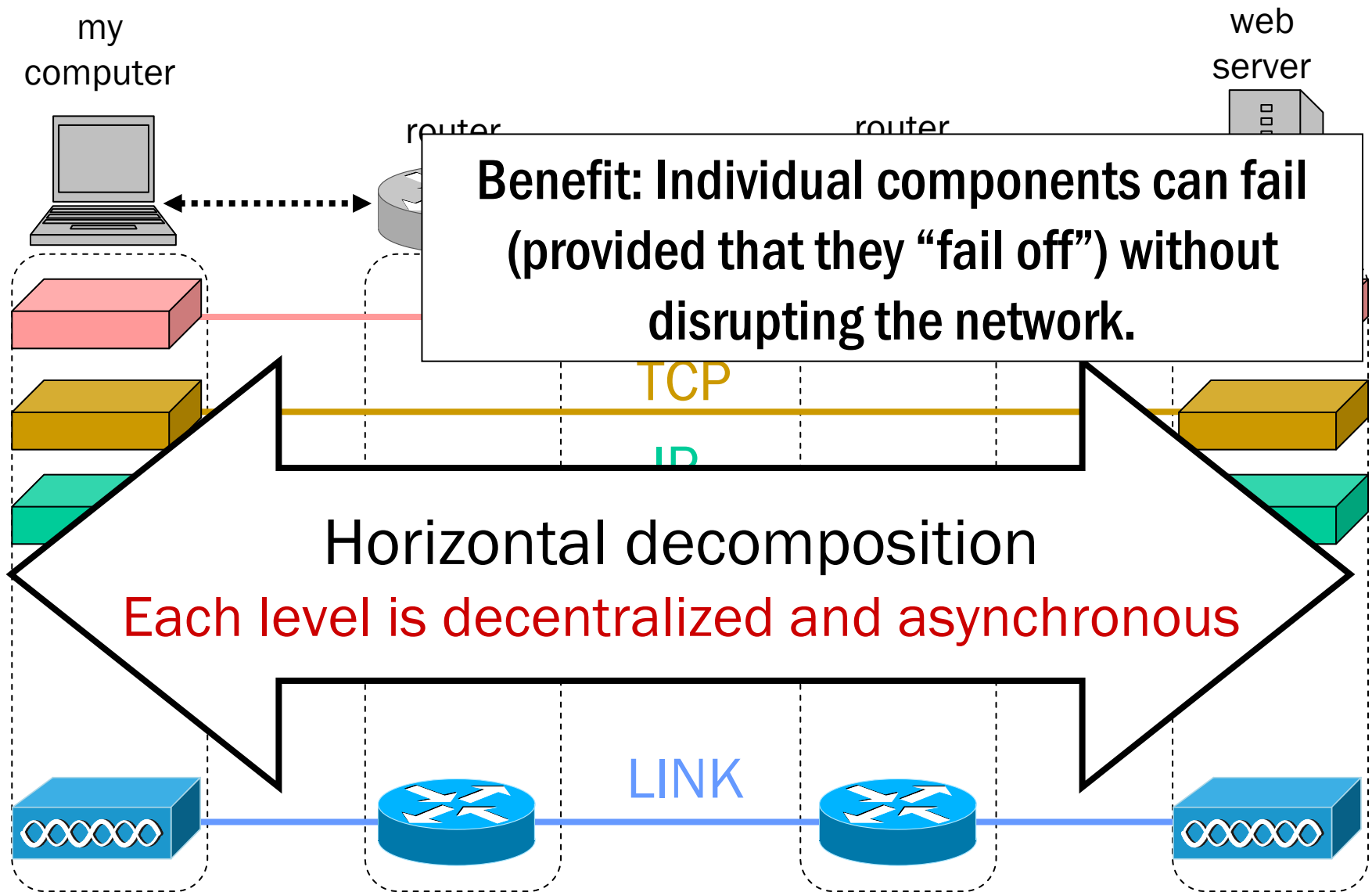
The Internet is a LAYERED Network



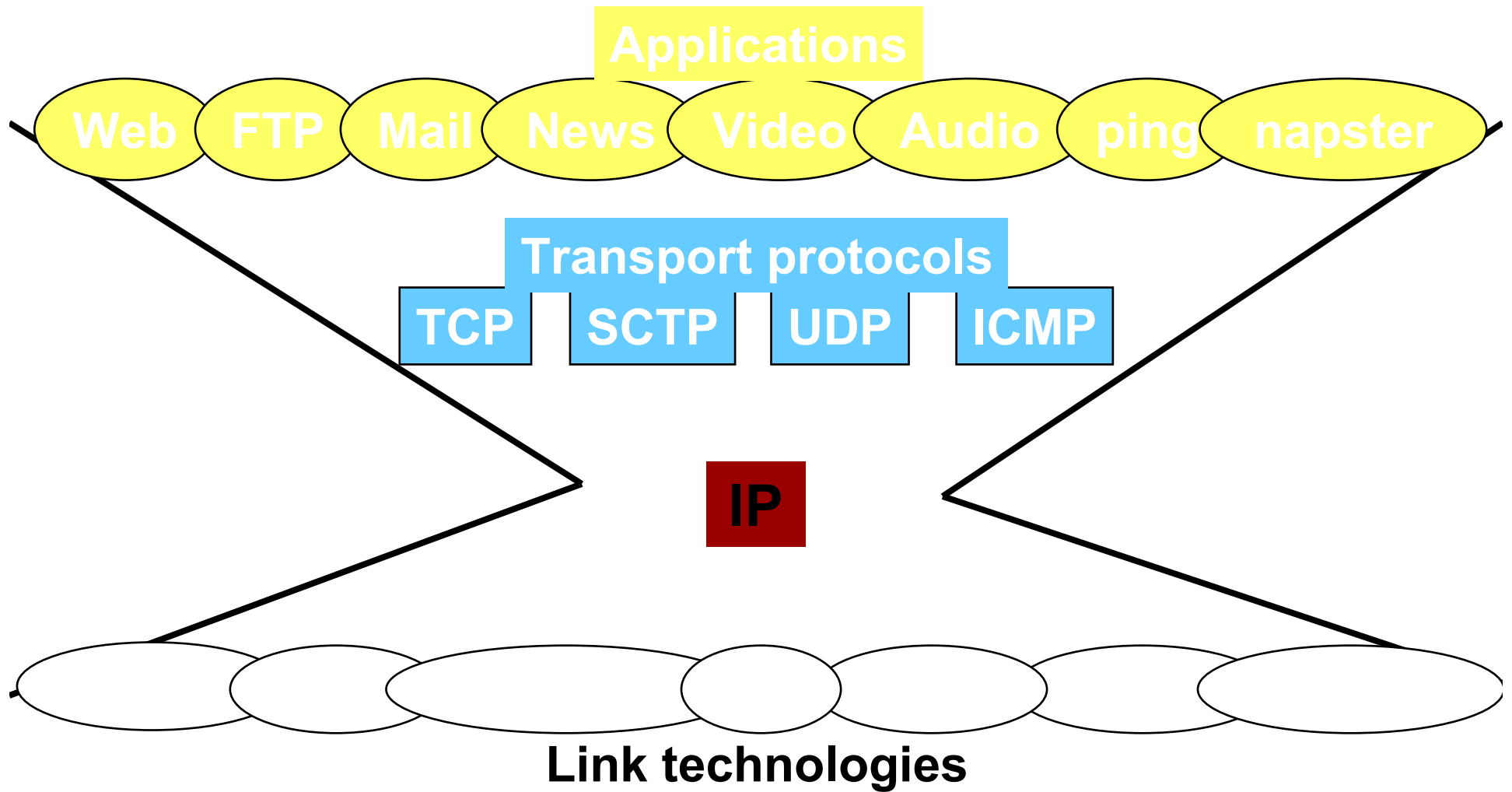
Internet Architecture: Vertical Decomposition



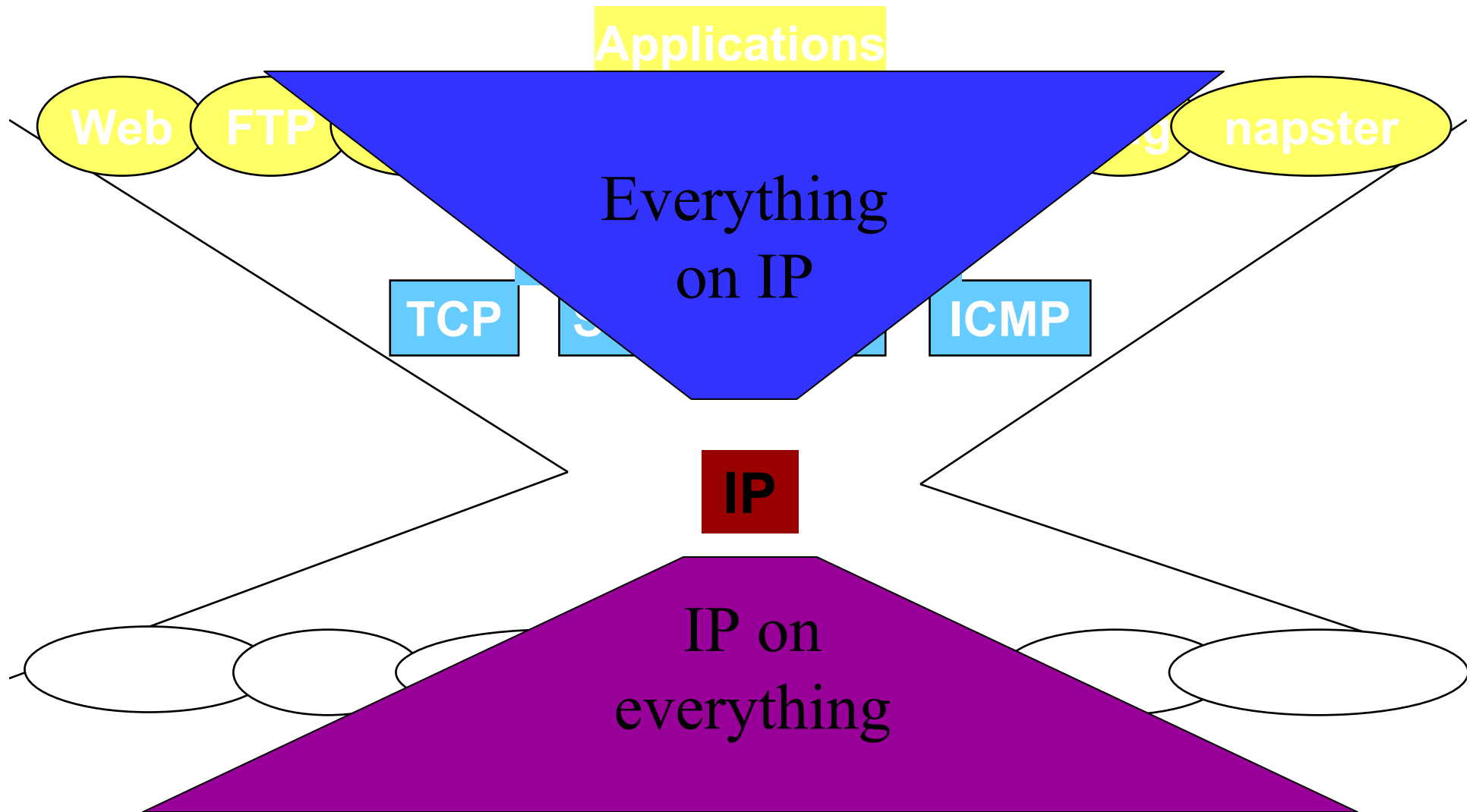
Internet Architecture: Horizontal Decomposition



The Internet hourglass

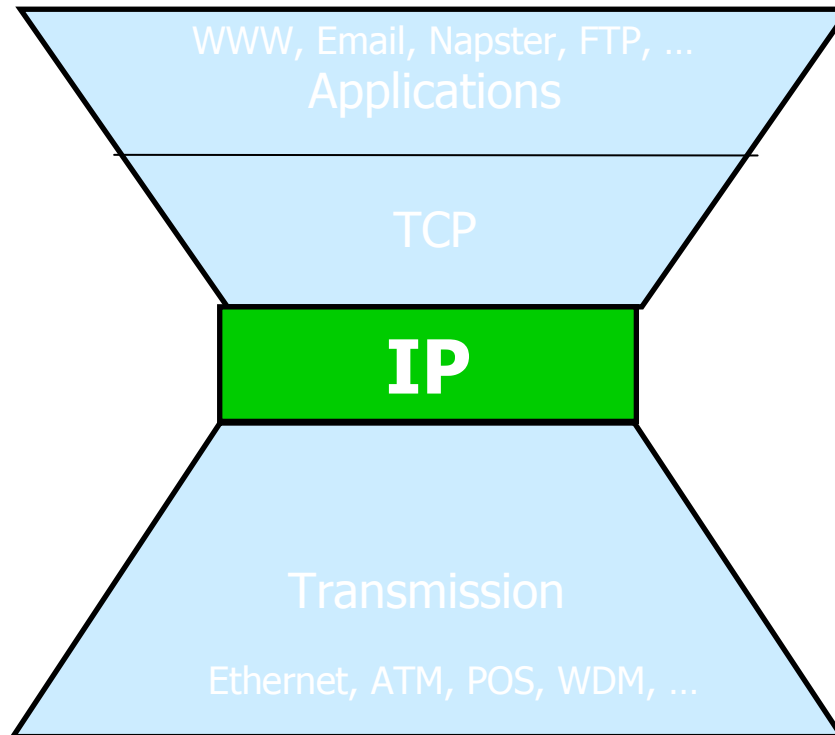


The Internet hourglass

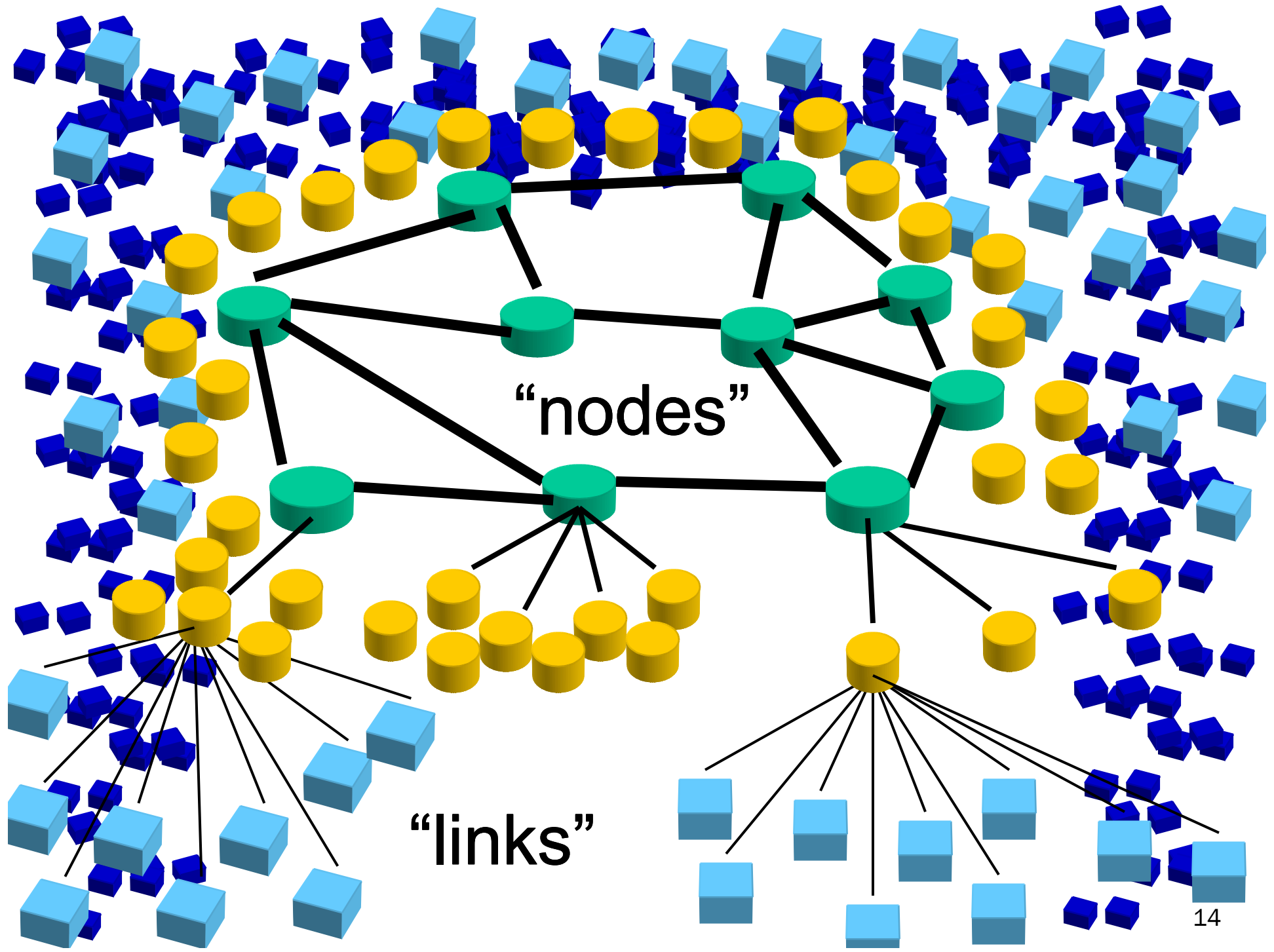


Courtesy Hari Balakrishnan

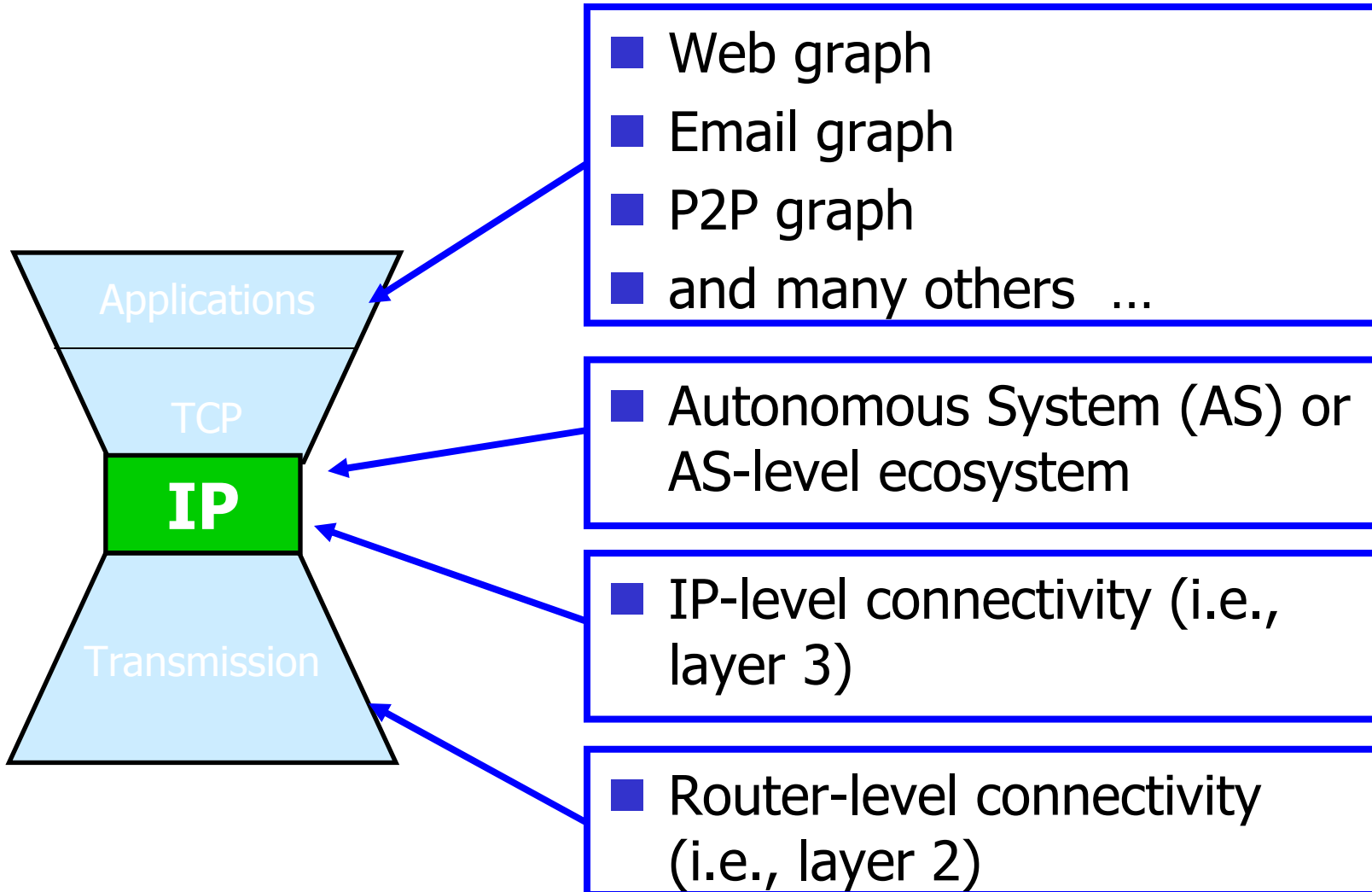
Internet Connectivity/Topology



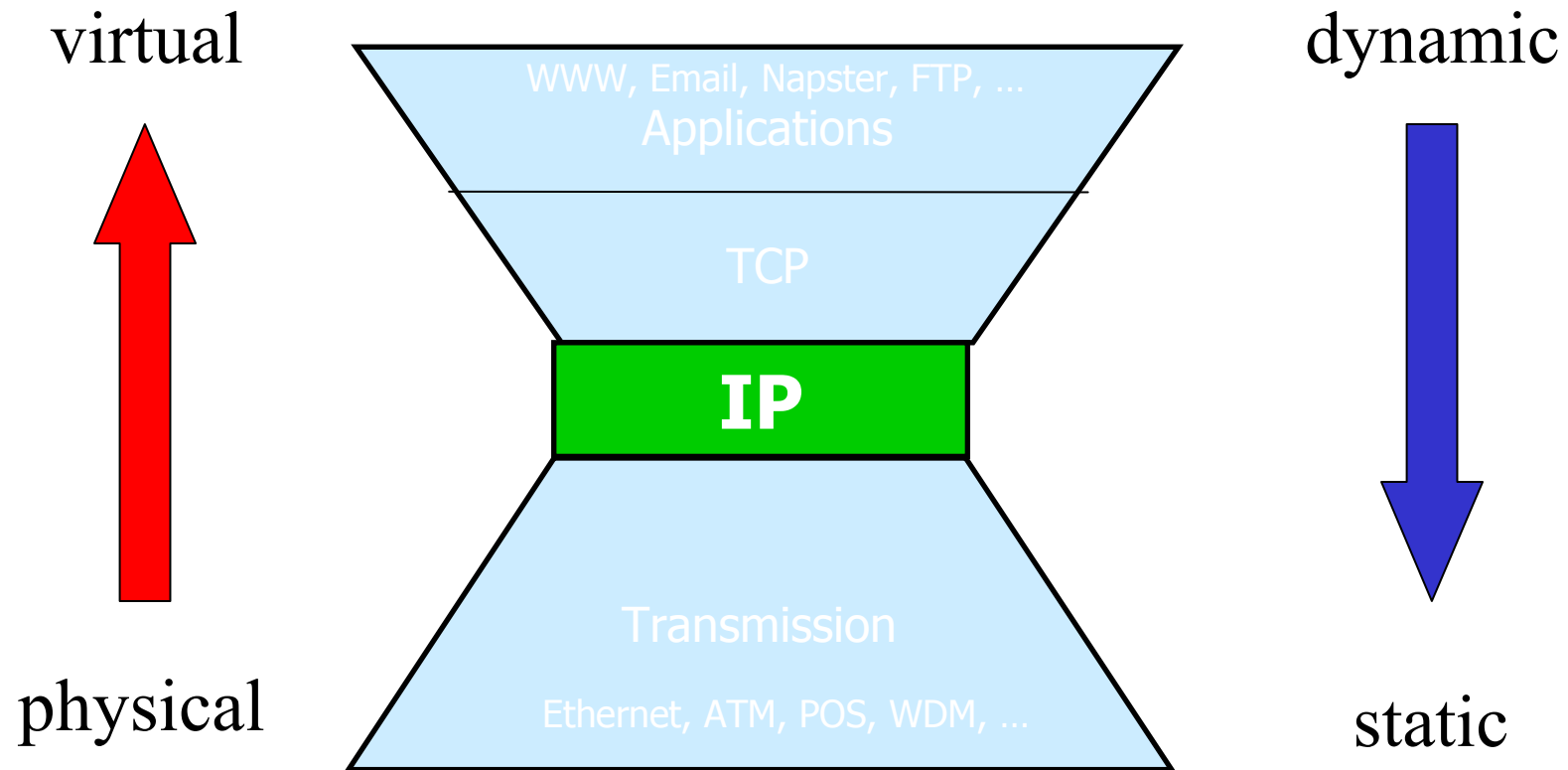
- Consider a (vertical) layer of the Internet hourglass
- Expand it horizontally
- Give layer-specific meaning to “nodes” and “links” 13



The Many Facets of Internet Connectivity/Topology



Internet Connectivity/Topology



On Measuring Internet Connectivity

- No central agency/repository
- Economic incentive for ISPs to obscure network structure
- Direct inspection is typically not possible
- Based on measurement experiments, hacks
- Mismatch between what we want to measure and can measure
- Specific examples covered in this talk
 - Physical connectivity (ISP router-level topology)
 - Logical connectivity (Internet AS-level topology)

Back to our Basic Question

Do the available Internet-related connectivity measurements and their analysis support the sort of claims that can be found in the existing complex networks literature?

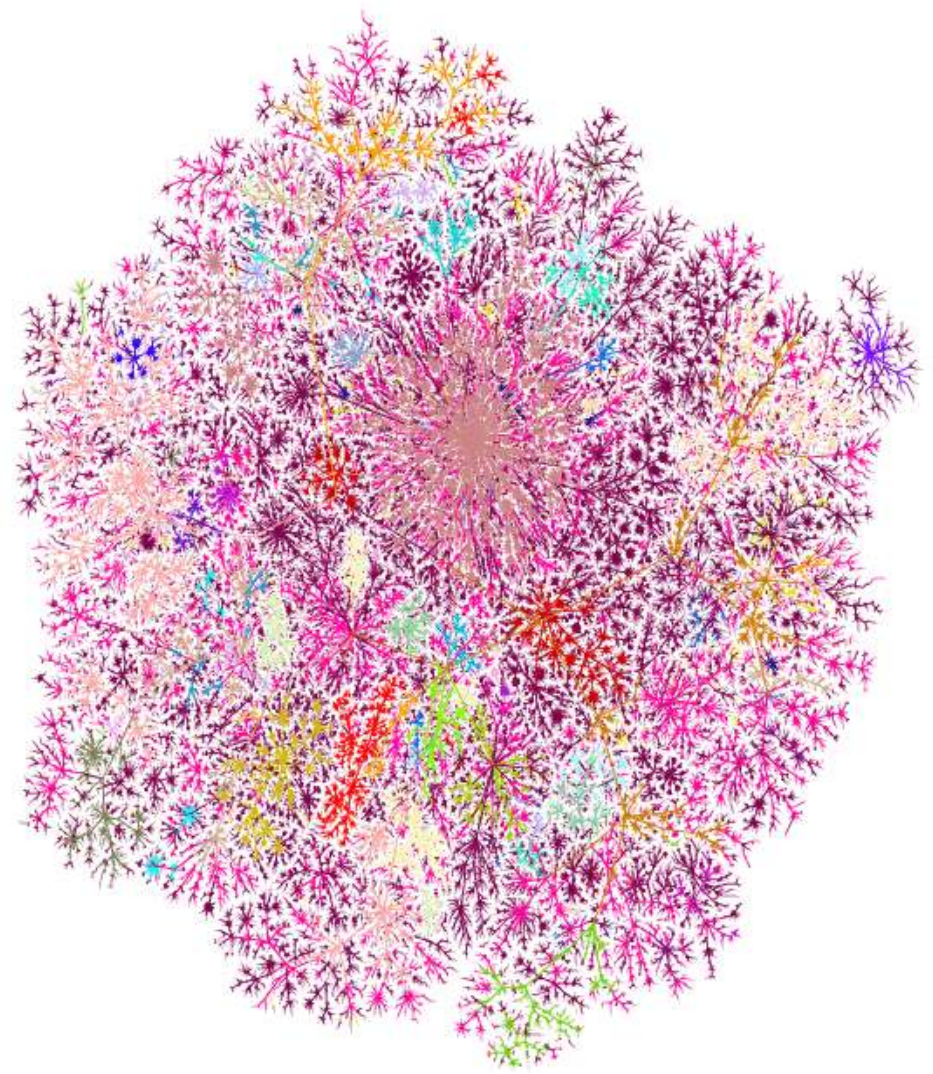
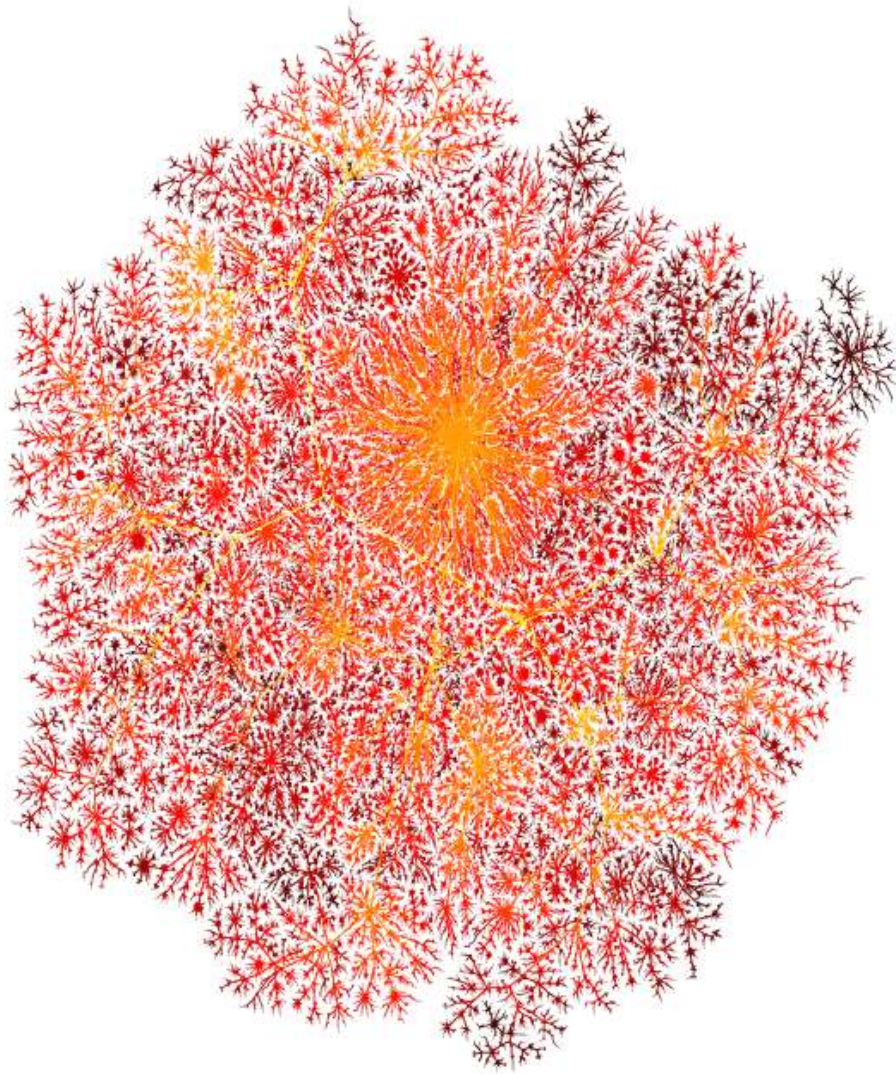
Key Issues

- What about data hygiene?
- What about statistical rigor?
- What about model validation?

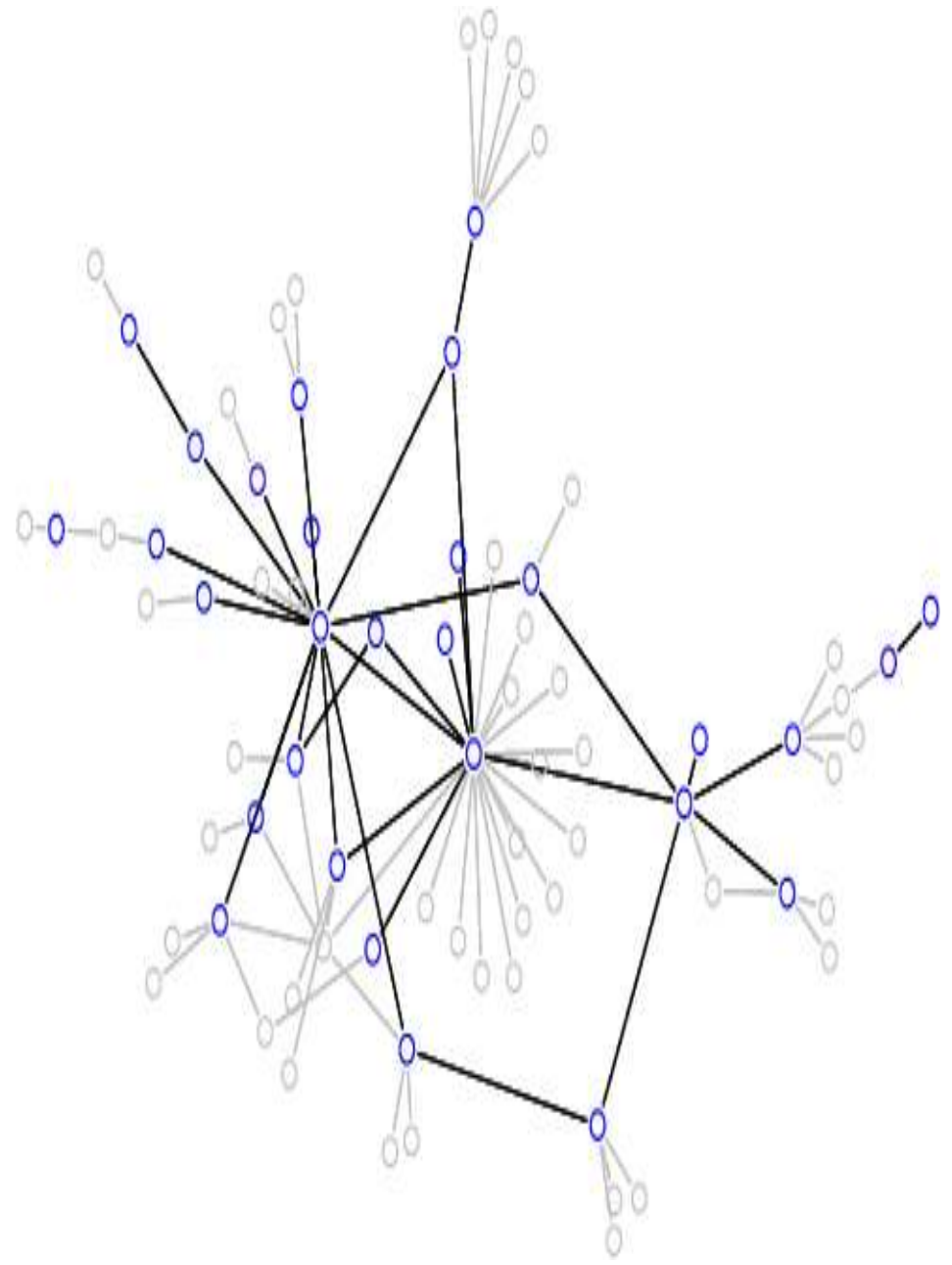
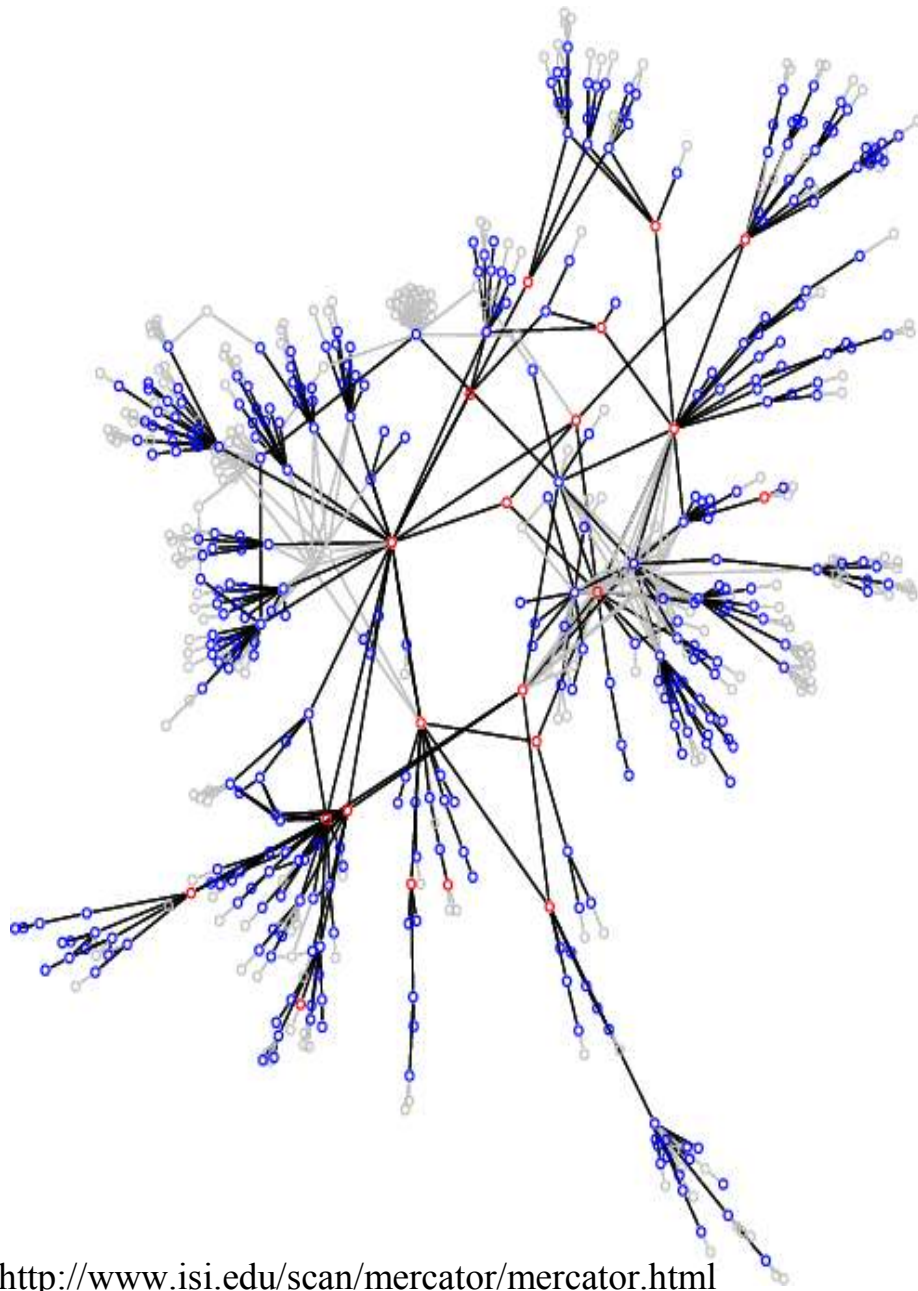
On Data Hygiene

On Measuring the Internet's Router-level Topology

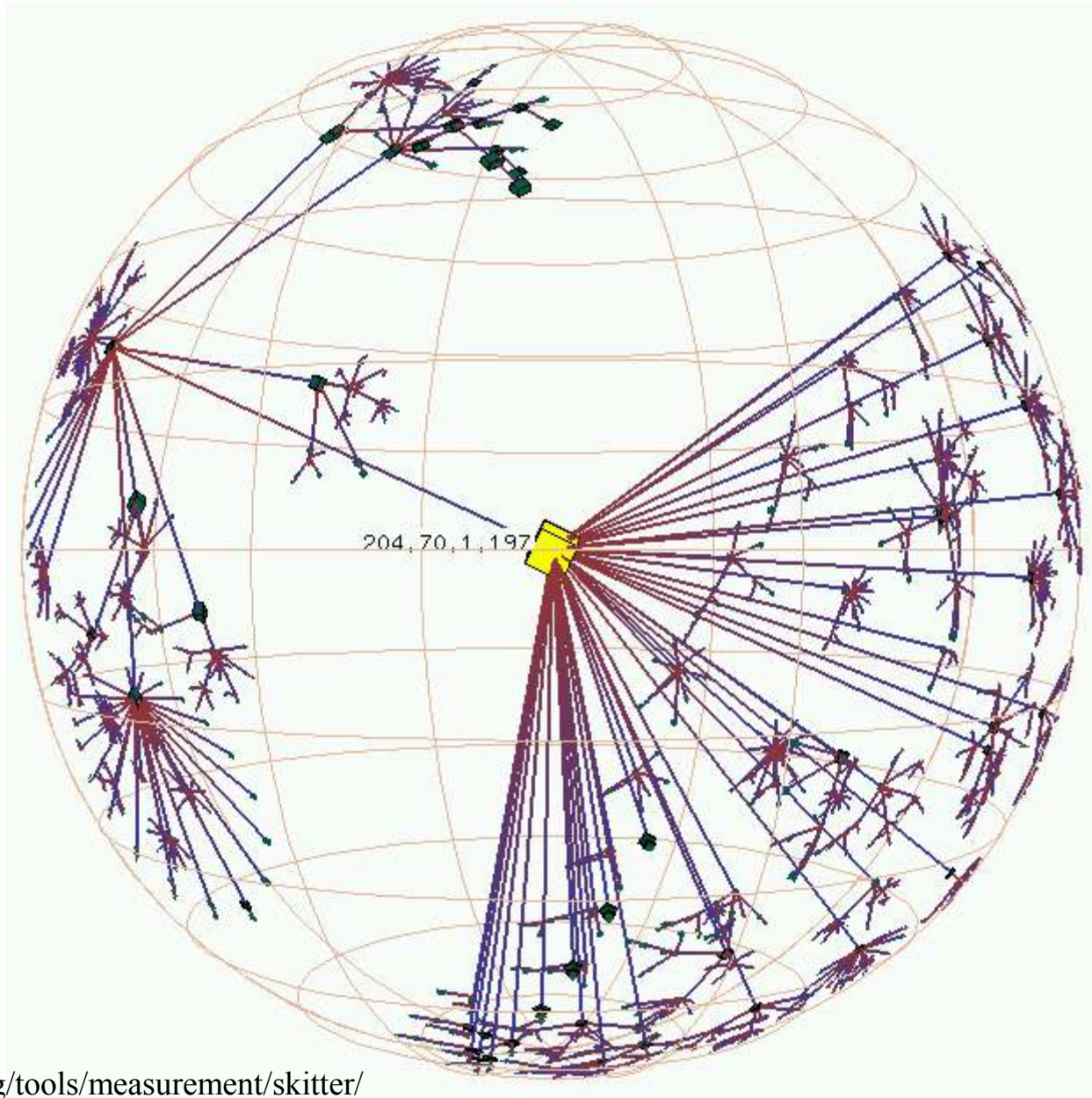
- **traceroute** tool
 - Discovers compliant (i.e., IP) routers along path between selected network host computers
- Large-scale traceroute experiments
 - Pansiot and Grad (router-level map from around 1995)
 - Cheswick and Burch (mapping project 1997--)
 - Mercator (router-level maps from around 1999 by R. Govindan and H. Tangmunarunkit)
 - Skitter (ongoing mapping project by CAIDA folks)
 - Rocketfuel (state-of-the-art router-level maps of individual ISPs by UW folks)
 - Dimes (EU project)



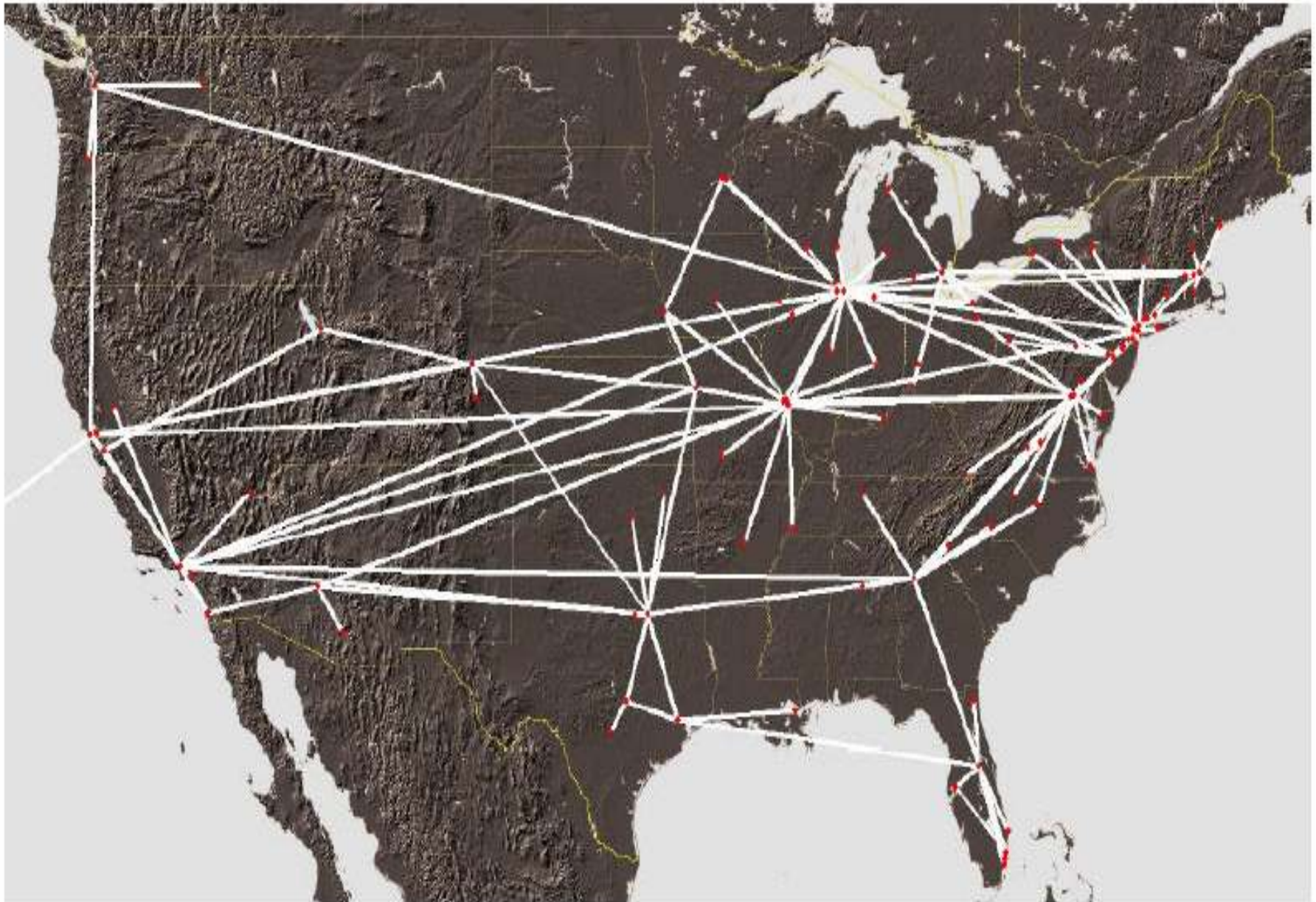
<http://research.lumeta.com/ches/map/>



<http://www.isi.edu/scan/mercator/mercator.html>



<http://www.caida.org/tools/measurement/skitter/>



Background image courtesy JHU, applied physics labs

<http://www.cs.washington.edu/research/networking/rocketfuel/bb>



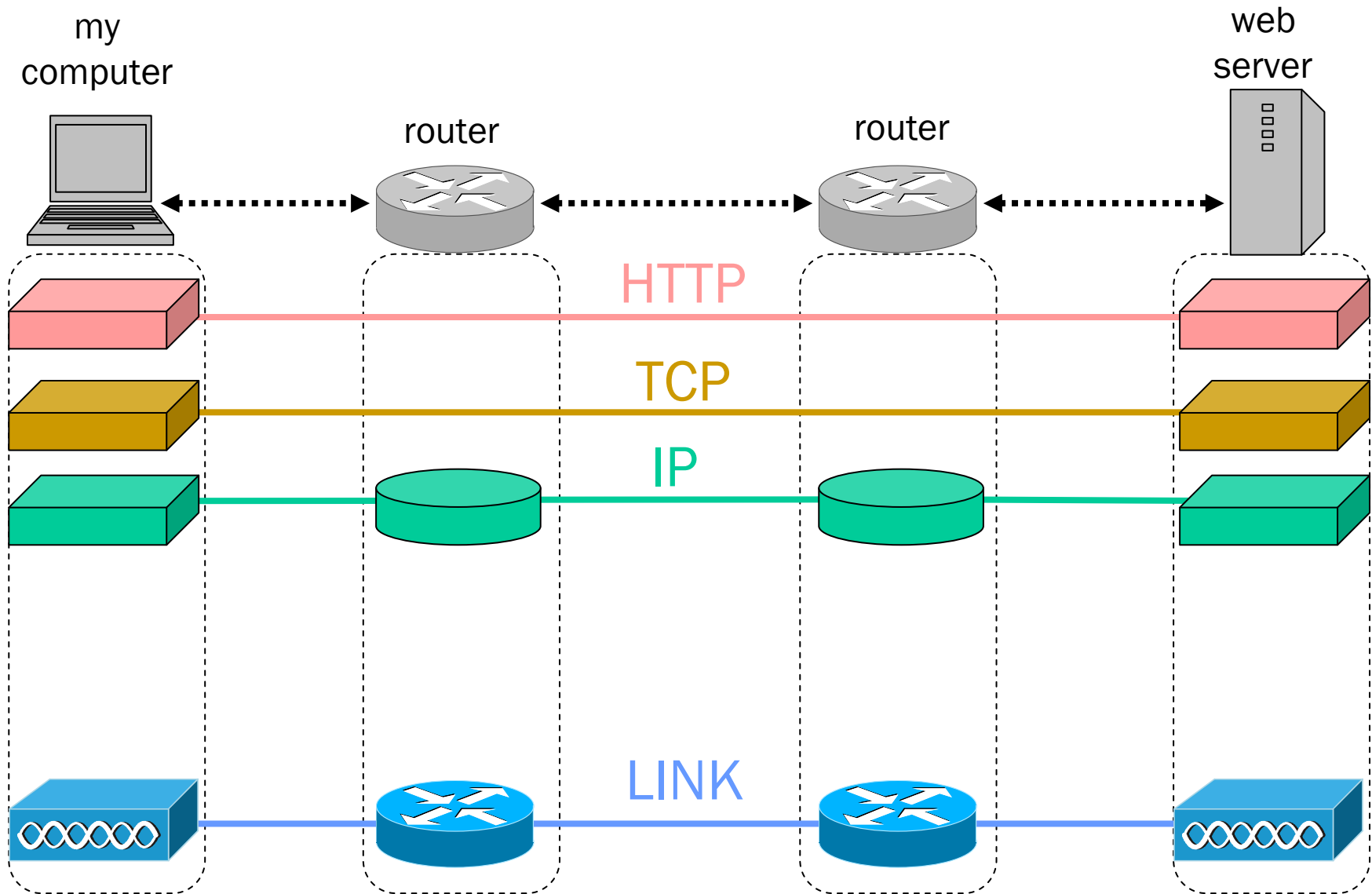
Background image courtesy JHU, applied physics labs

<http://www.cs.washington.edu/research/networking/rocketfuel/>

HOWEVER: Problems with existing measurements

- traceroute-based measurements are **ambiguous**
 - traceroute is strictly about IP-level connectivity
 - traceroute cannot distinguish between high connectivity nodes that are for real and that are fake and due to underlying Layer 2 (e.g., Ethernet, ATM) or Layer 2.5 technologies (e.g., MPLS)

The Internet: The Engineering Perspective





**Illusion of a fully-meshed
Network due to use of MPLS**

Background image courtesy JHU, applied physics labs

<http://www.cs.washington.edu/research/networking/rocketfuel/>

- 
- www.savvis.net
 - managed IP and hosting company
 - founded 1995
 - offering “private IP with ATM at core”

**This “node” is an entire network!
(not just a router)**

HOWEVER: Problems with existing measurements

- traceroute-based measurements are **ambiguous**
 - traceroute is strictly about IP-level connectivity
 - traceroute cannot distinguish between high connectivity nodes that are for real and that are fake and due to underlying Layer 2 (e.g., Ethernet, ATM) or Layer 2.5 technologies (e.g., MPLS)
- traceroute-based measurements are **inaccurate**
 - Requires some guesswork in deciding which IP addresses/interface cards refer to the same router (“alias resolution” problem)
- traceroute-based measurements are **incomplete/biased**
 - IP-level connectivity is more easily/accurately inferred the closer the routers are to the traceroute source(s)
 - Node degree distribution is inferred to be of the power-law type even when the actual distribution is not

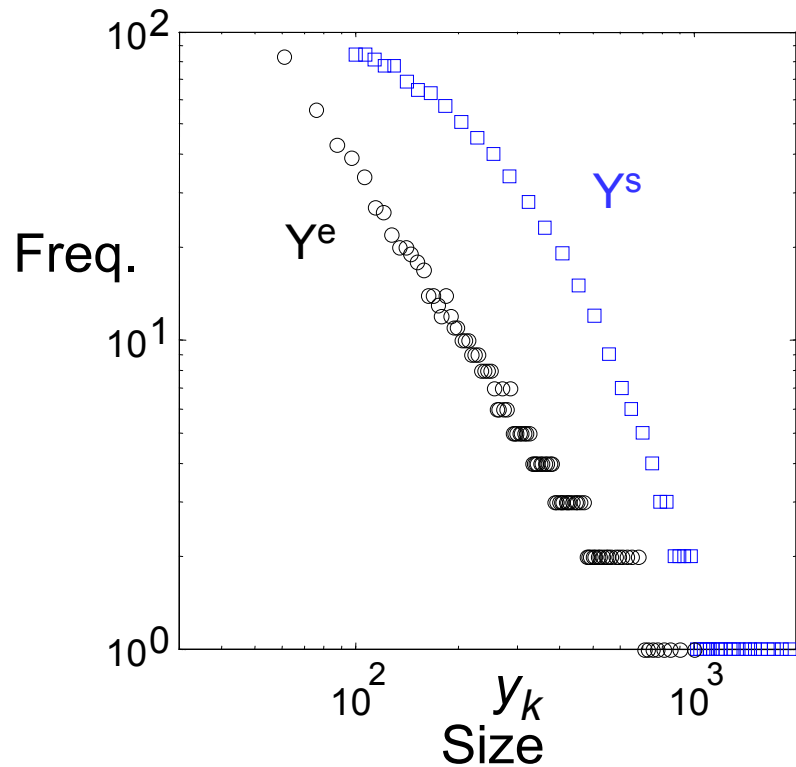
On Statistical Rigor

How to lie with statistics ...

Given: Samples from an exponential distribution

Want: Claim power law behavior

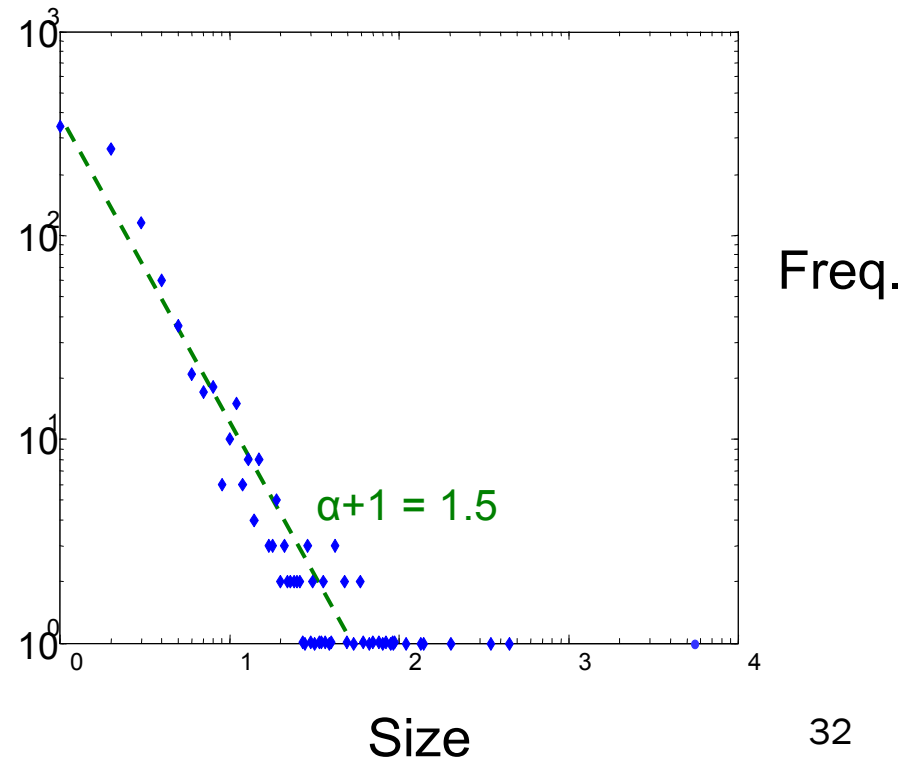
Recipe: Use size-frequency plots!



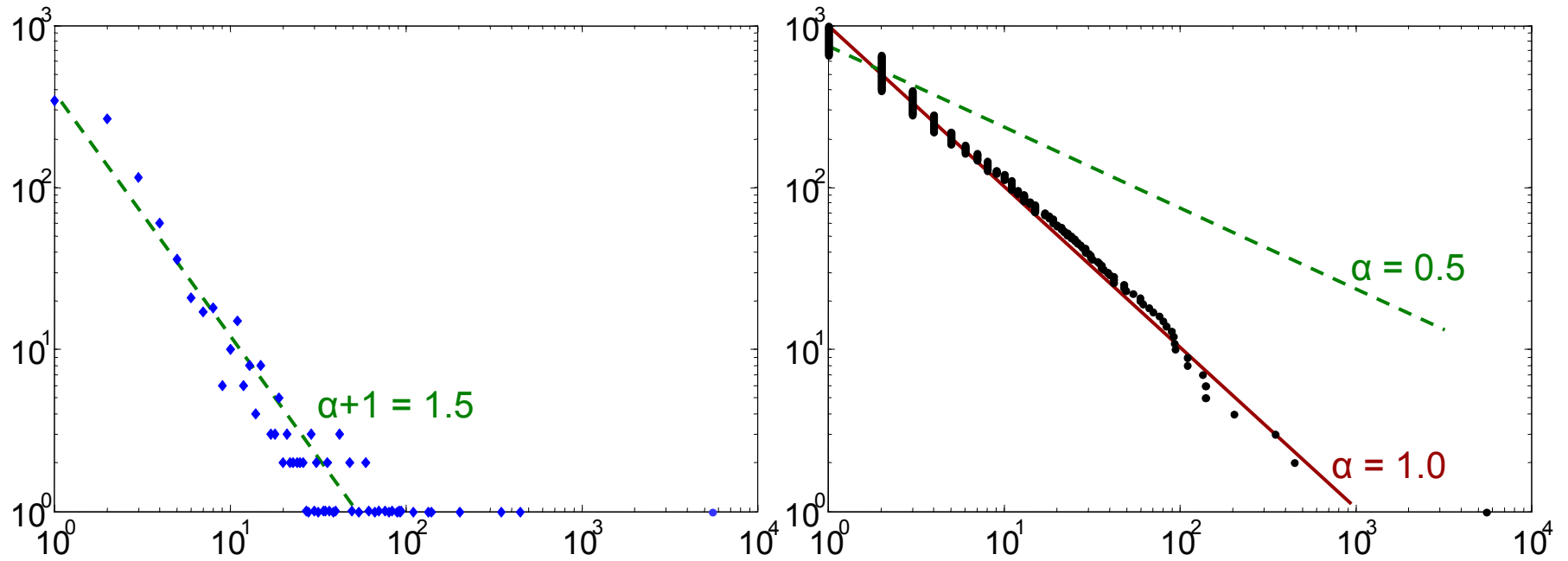
Given: Samples from a Pareto distribution with $\alpha=1.0$

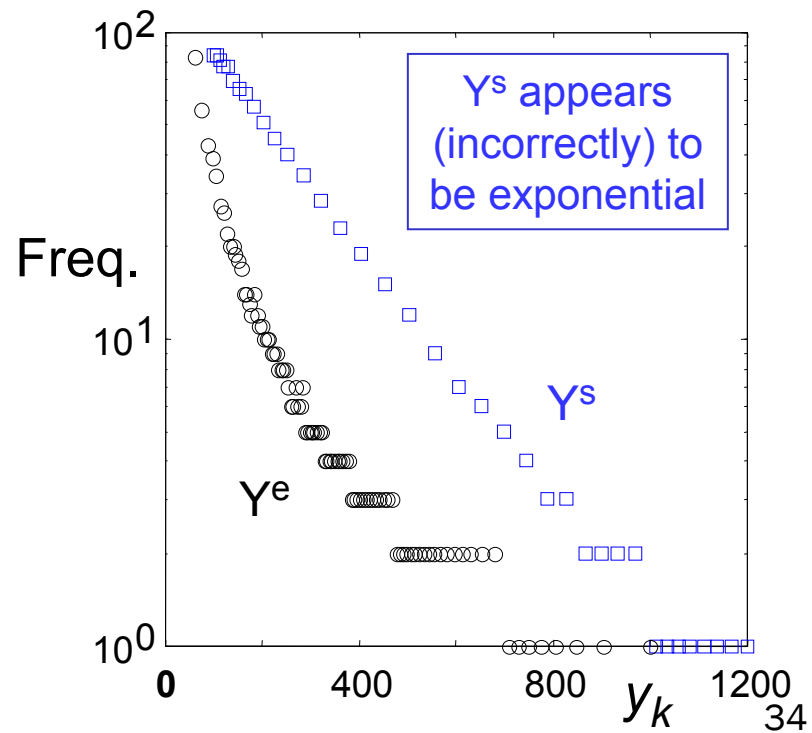
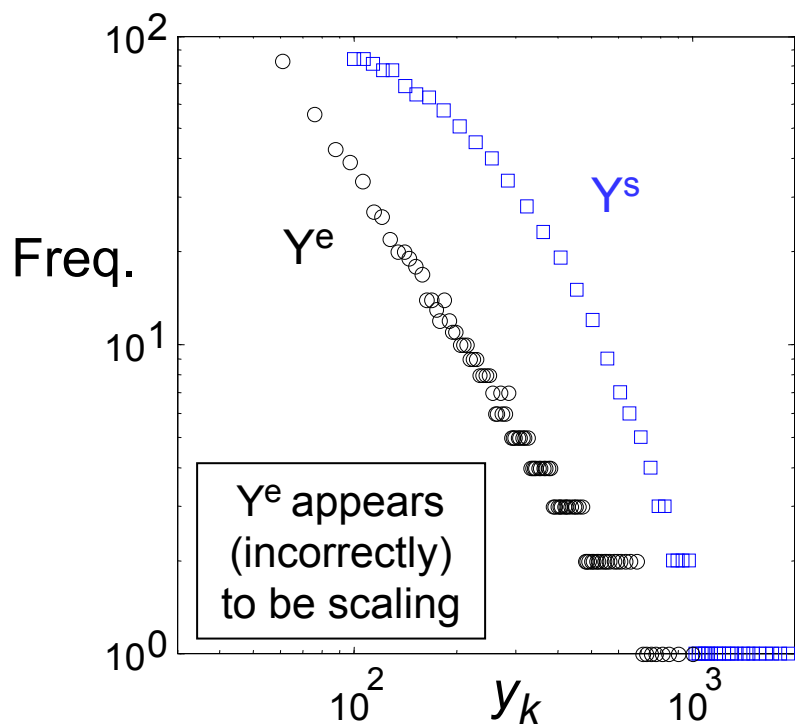
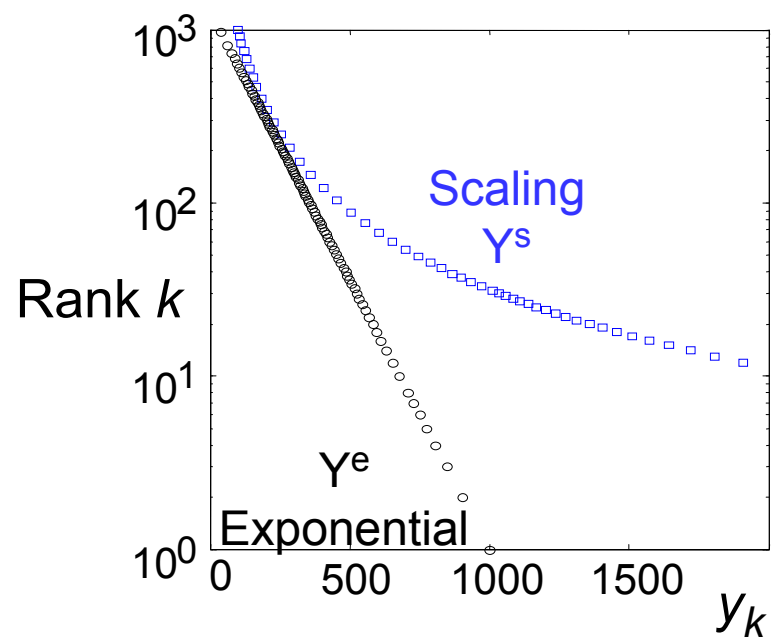
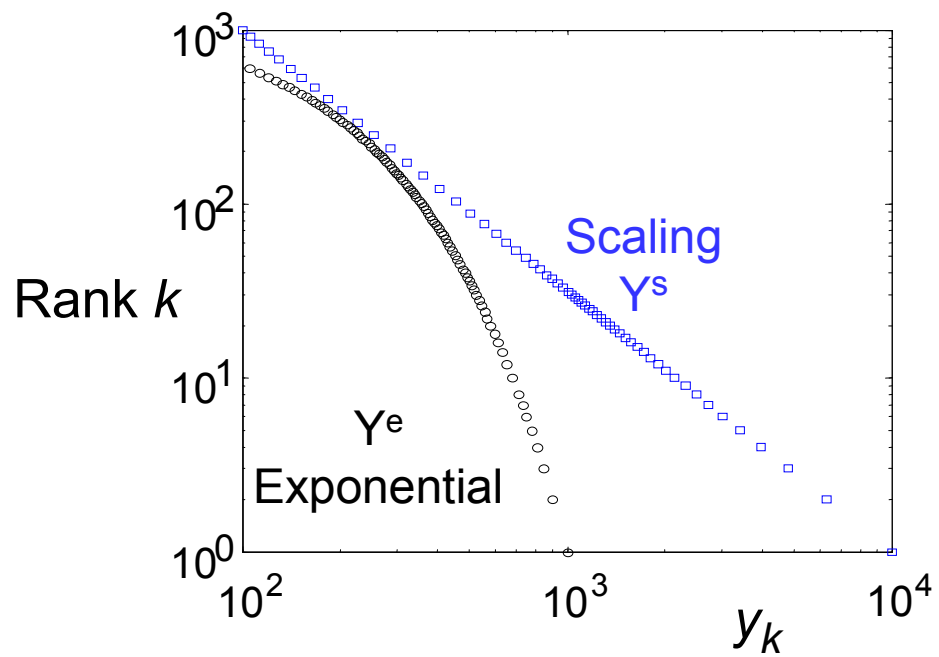
Want: Claim power law with $\alpha=1.5$

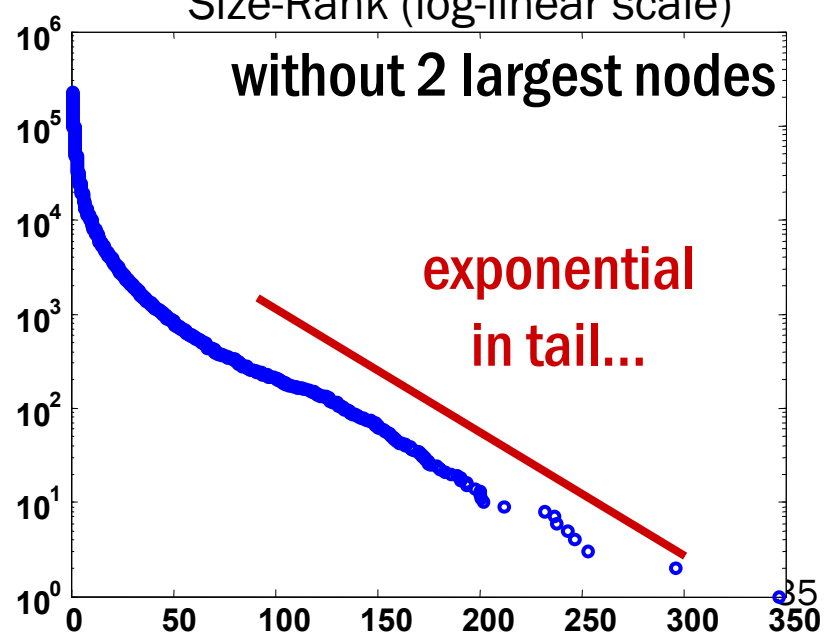
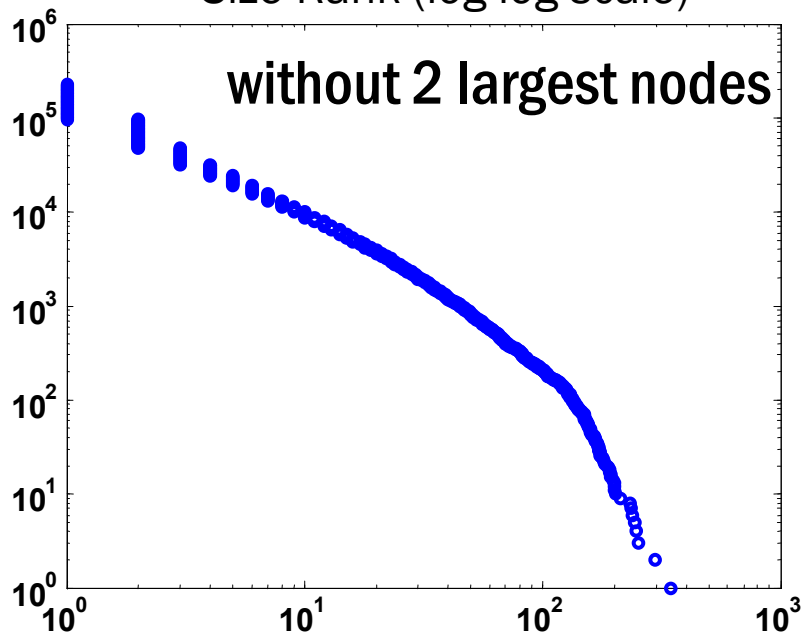
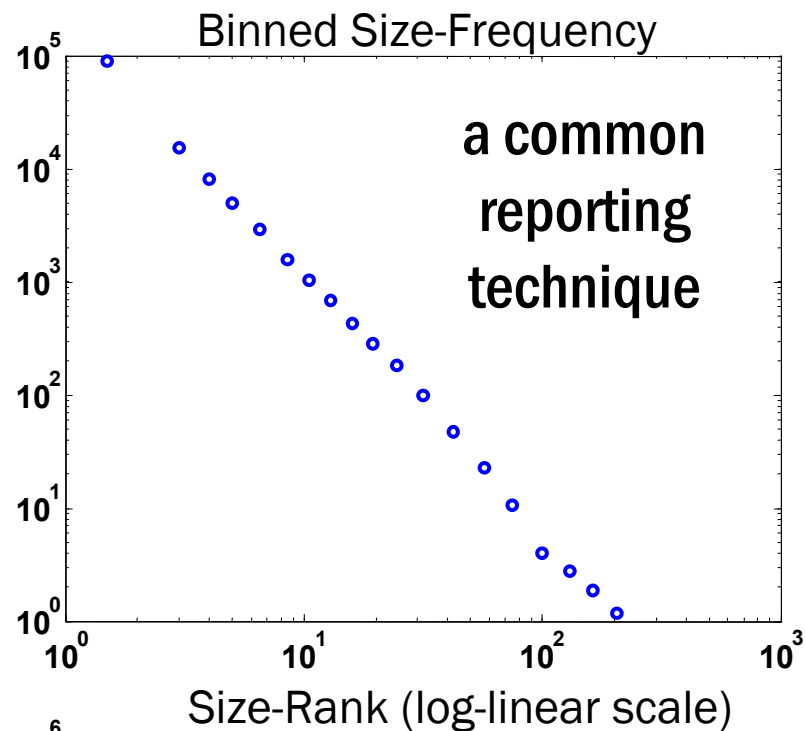
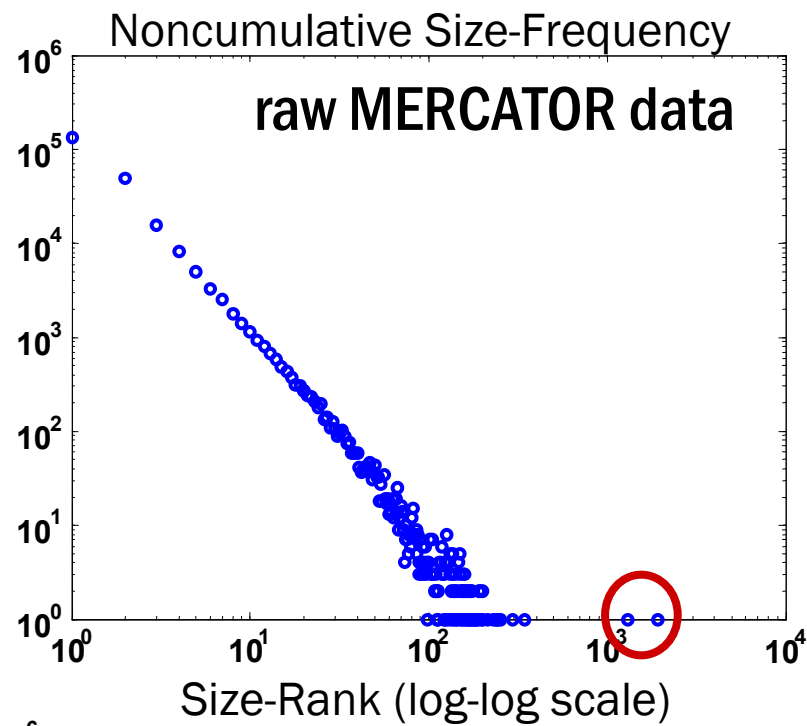
Recipe: Use size-frequency plots!



Size-Frequency vs. Size-Rank Plots or Non-cumulative vs. Cumulative







On Model Validation

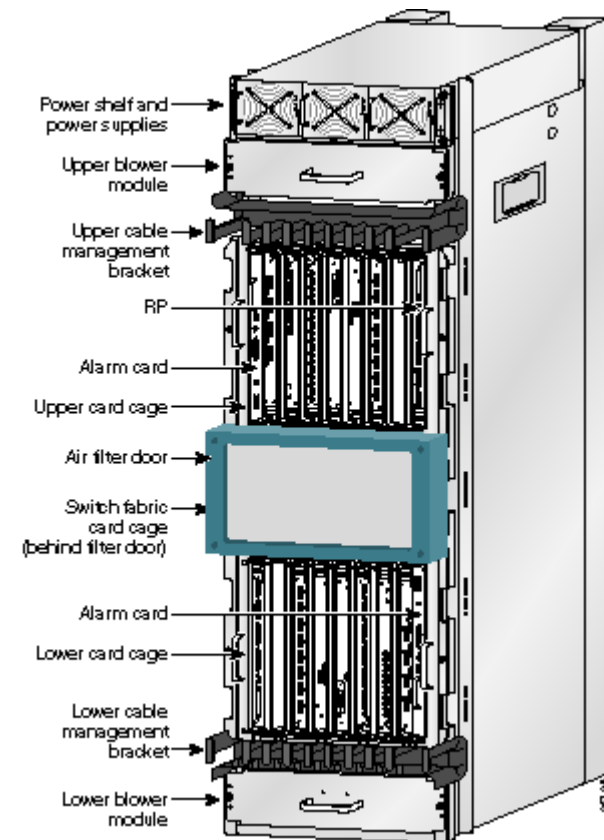
Taking Model validation more serious ...

- Mathematical Modeling 101
 - For one and the same observed phenomenon, there are usually many different explanations/models
 - All models are wrong, but some are “damned lies”
- Model validation \neq data fitting
 - The ability to reproduce a few graph statistics does not constitute “serious” model validation
 - Which of the observed properties does a proposed model have to satisfy before it is deemed “valid”?
- What constitutes “serious” model validation?
 - There is more to networks than connectivity
 - Meaning of “node” and “link”

Cisco 12000 Series Routers

- Modular in design, creating flexibility in configuration.
- Router capacity is constrained by the number and speed of line cards inserted in each slot.

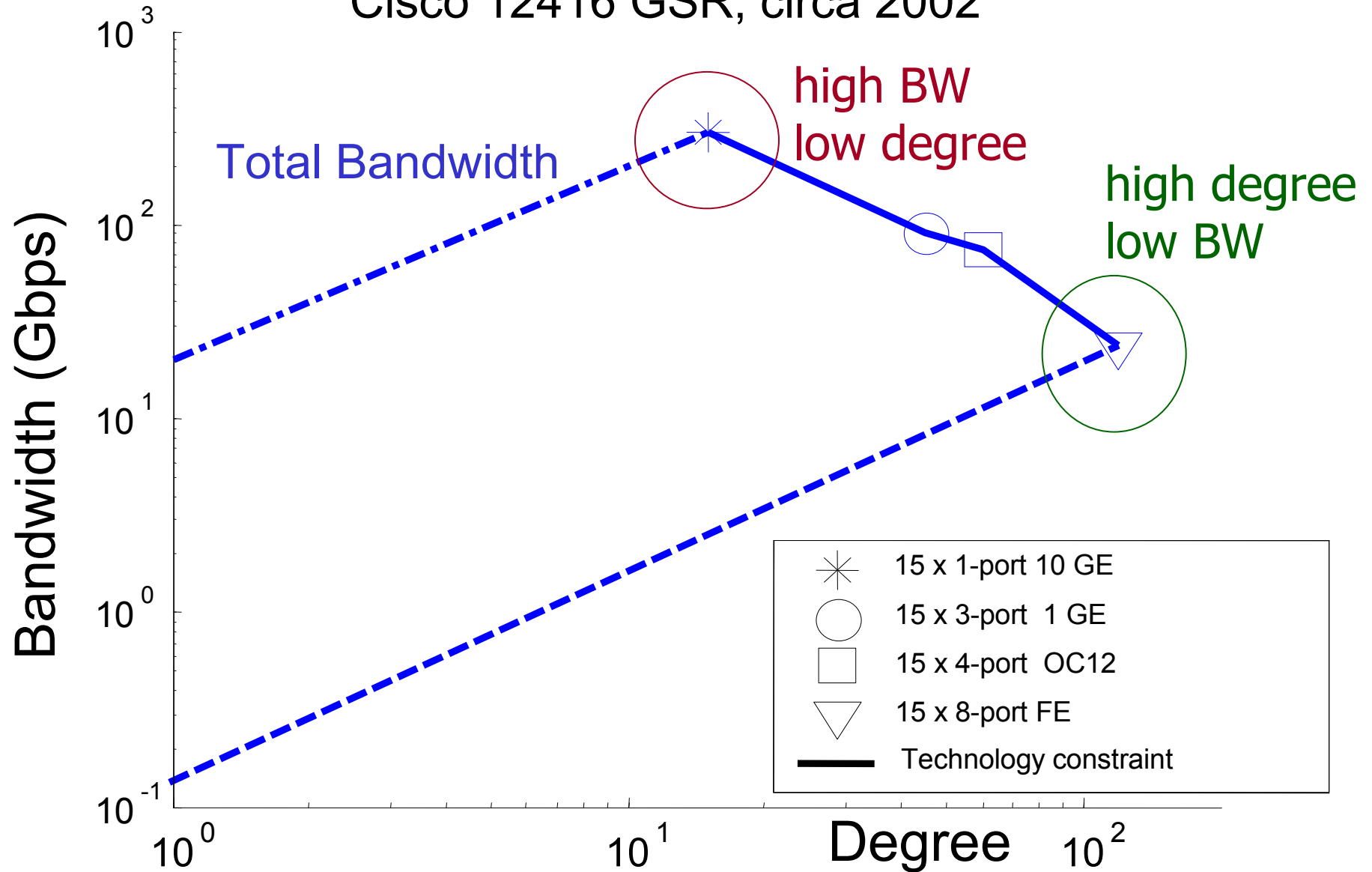
Chassis	Rack size	Slots	Switching Capacity
12416	Full	16	320 Gbps
12410	1/2	10	200 Gbps
12406	1/4	6	120 Gbps
12404	1/8	4	80 Gbps



Source: www.cisco.com

Router Technology Constraint

Cisco 12416 GSR, circa 2002



Back to the Basic Question:

Do the available Internet-related connectivity measurements and their analysis support the sort of claims that can be found in the existing complex networks literature?

Short Answer:

No!

Network Science and the Internet: “Lies, damned lies and statistics”

- Power-law (scale-free) node degree distribution
 - Example of “how to lie with statistics”
- Preferential attachment-type models
 - (White) lies ...
- Highly popularized claims (e.g., Achilles’ heel, fragile/vulnerable to targeted node removal, zero epidemic threshold)
 - Damned lies ...
 - These claims are not “controversial” – they are simply wrong!
- Bad analysis of bad data = bad models (“damned lies”)
“Bad [models] are potentially important: they can be used to stir up public outrage or fear; they can distort our understanding of our world; and they can lead us to make poor policy choices.” (J. Best)

The “Math” Perspective of the Internet

- Assumption
 - Node degree distributions follow a power-law
- Rigorous model definition/formulation
 - Preferential attachment-type models
- Rigorous proofs
 - Achilles’ heel
 - Fragile/vulnerable to targeted node removal
 - Zero epidemic threshold
- End result is the same
 - Rigorous analysis of bad model = “damned lies”

How to avoid such Fallacies?

- Taking model validation more serious
- Applying an engineering perspective to engineered systems

Internet Modeling: An Engineering Perspective

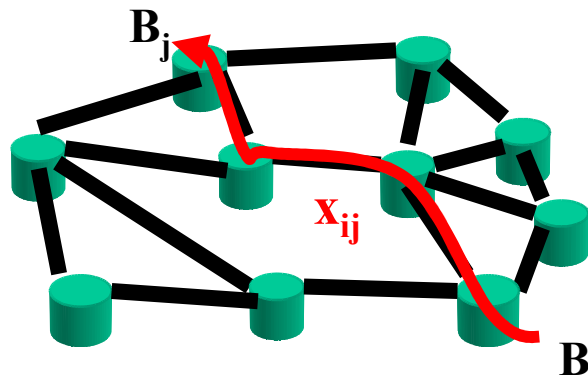
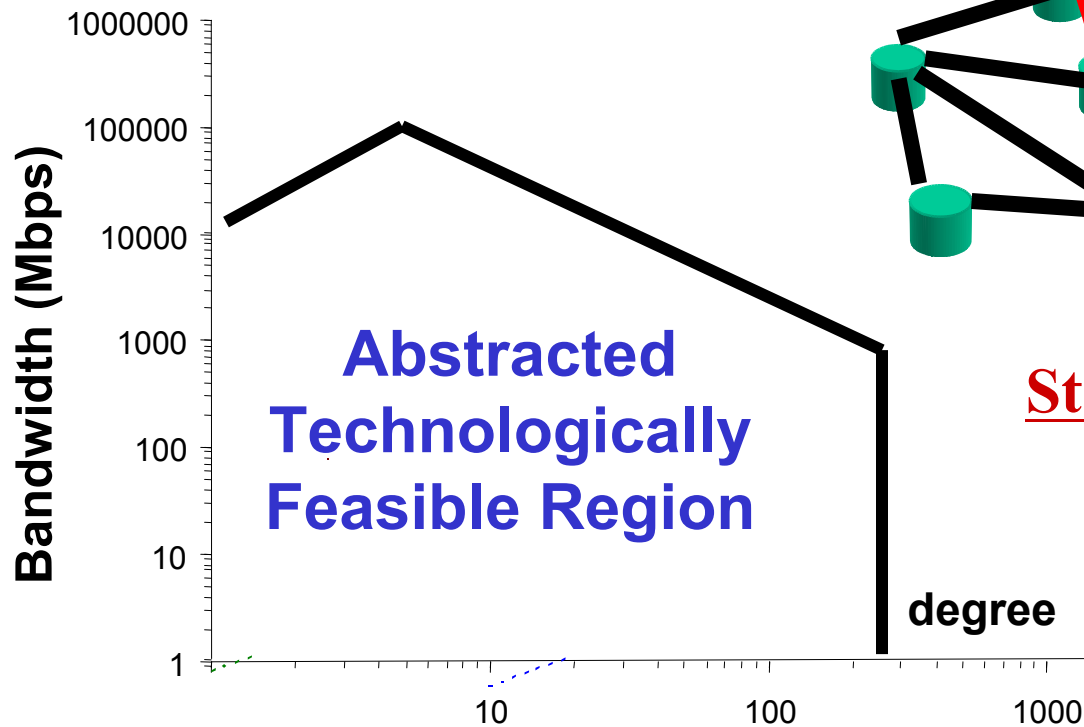
- ISPs design their router-level topology for a purpose, namely to carry an expected traffic demand
- Surely, the way an ISP designs its physical infrastructure is not the result of a series of coin tosses ...
- Randomness enters in terms of uncertainty in traffic demands
- ISPs are constrained in what they can afford to build, operate, and maintain (economics).
- The “nodes” and “links” are physical things that have hard constraints (technology).
- Decisions of ISPs are driven by objectives (performance) and reflect tradeoffs between what is feasible and what is desirable (heuristic optimization)
- Power laws: Full of sound and fury, signifying nothing!

Heuristically Optimized Topologies (HOT)

Given realistic technology constraints on routers, how well is the network able to carry traffic?

Step 1: Constrain to be feasible

Step 2: pick traffic demand model

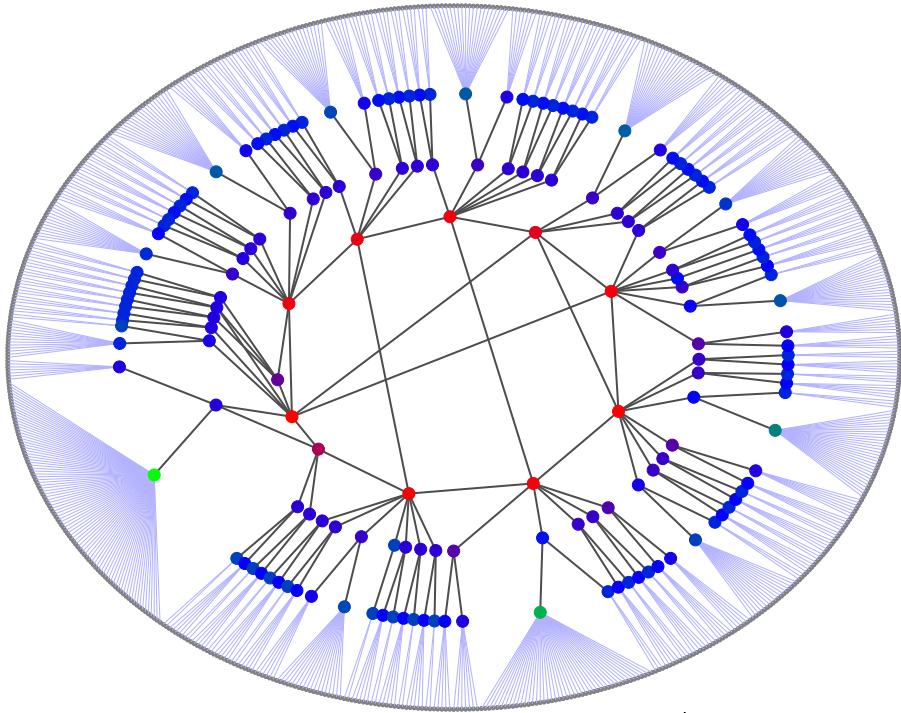


$$x_{ij} \propto B_i B_j$$

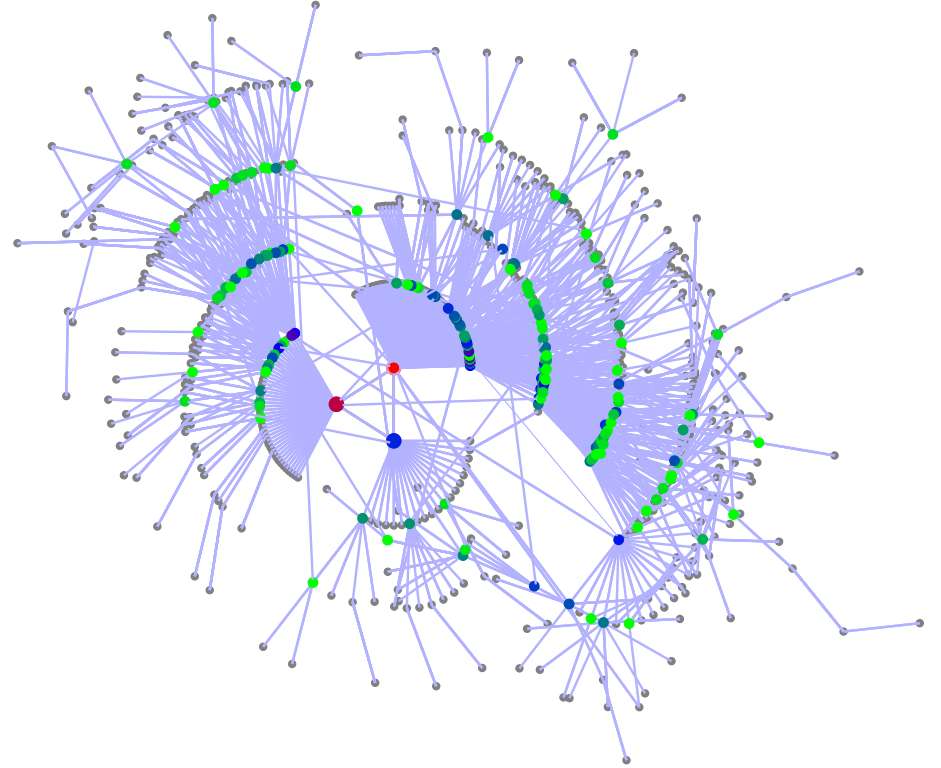
Step 3: Compute max flow

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

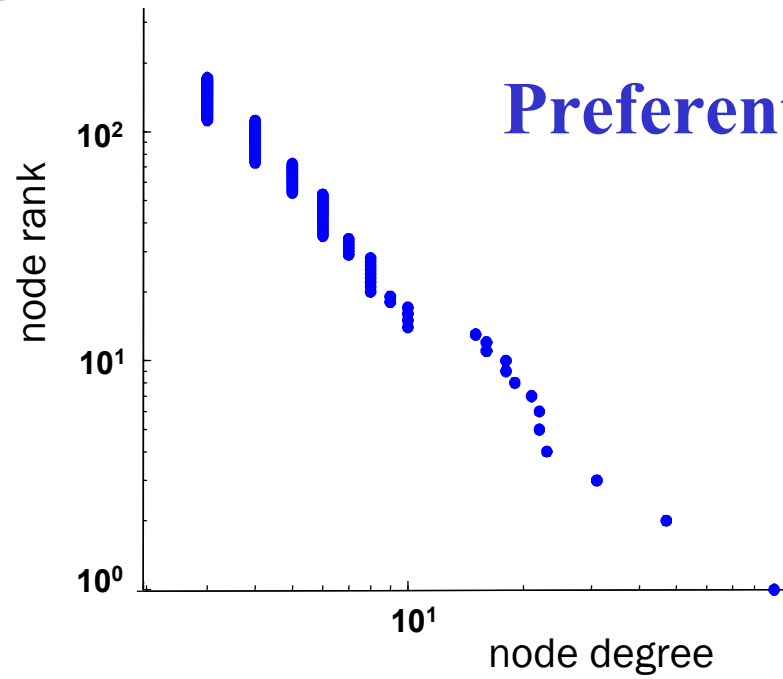
$$s.t. \sum_{i,j:k \in r_{ij}} x_{ij} \leq B_k, \forall k$$



HOT model



Preferential Attachment



HOT-type Network Models

- Very recent alternatives to PA-type models
 - Extremely unlikely to occur at random
- Key features of HOT models
 - Consistent with existing ISP router-level topologies
 - Consistent with existing technologies
 - Consistent with (complementary) measurements
 - Node degree distribution is a non-issue
- Same story for AS-level Internet topology
 - Surely, deciding on whether or not to establish what type of peering relationship and with whom is not the outcome of a series of chance experiments conducted by the different ASs, but is largely based on economic arguments.
 - First HOT-type model by Chang et al. (2006)

Litmus Test for Newly Proposed Network Models

- Make node degree distribution a non-issue
 - Good reasons
 - High-quality data but low variability (e.g., exponential)
 - Low-quality data
 - High-quality data and high variability (e.g., power-laws)
 - PA-type models
 - dead on arrival
 - Only reasonable alternative
 - Bring in and rely on domain knowledge
- What new kinds of measurements does the proposed model suggest for the purpose of model validation
 - PA-type models: none
 - HOT models: get data on existing router technology

Some Implications of this Engineering Perspective

- New paradigm for network modeling
 - Network modeling \neq data fitting
 - Node degree distribution is a non-issue
- Constrained optimization formulation
 - Optimization of tradeoffs between multiple functional objectives of networks
 - Subject to constraints on their components
 - With an explicit source of uncertainty (in the environment) against which solutions must be tolerant or robust

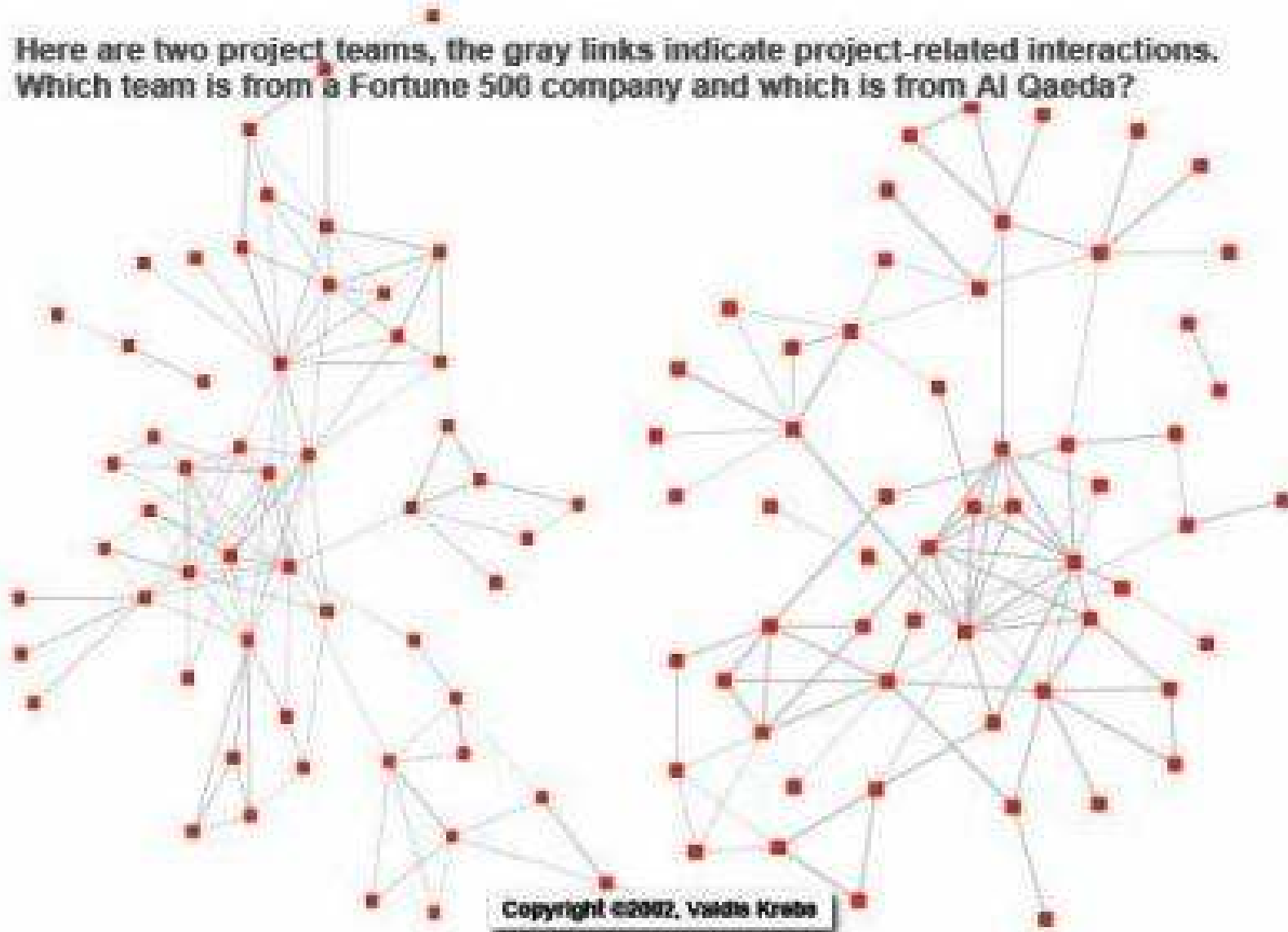
Further Implications of this Engineering Perspective

- Dynamics **of** graphs
 - Evolution of connectivity structures
 - Evolution of (internal) node/link structure
- Dynamics **over** graphs
 - Traffic dynamics (bytes, packets, flows, ...)
- Challenging feedback problem
 - Traffic dynamics/routing impacts network structure
 - Network structure impacts traffic dynamics/routing
- Can't (shouldn't) model connectivity without traffic
- Robustness/fragility considerations only make sense in the context of the broader system, i.e., **protocol stack**
 - Router-level: Inter-AS routing protocol
 - AS-level: Intra-AS routing protocol

Further Implications of this Engineering Perspective (cont.)

- Key question #1: What is the network as whole trying to achieve?
 - Internet router-level: see earlier
 - Internet AS-level: ?
 - WWW, P2P: ??
 - Social Networks: ???

Here are two project teams, the gray links indicate project-related interactions. Which team is from a Fortune 500 company and which is from Al Qaeda?



Further Implications of this Engineering Perspective (cont.)

- Key question #1: What is the network as whole trying to achieve?
 - Internet router-level: see earlier
 - Internet AS-level: ?
 - WWW, P2P: ??
 - Social Networks: ???

- Key question #2: How is the network trying to achieve its objective?
 - Decentralized, distributed
 - Duality gap (“price of anarchy”)

A Reminder

- Past: Modeling in the presence of high-quality data
 - *“All models are wrong ... but some are useful”* (G.E.P. Box)

- Future: Modeling in the presence of highly ambiguous data
 - Take the ambiguities in the data into account
 - *“When exactitude is elusive, it is better to be approximately right than certifiably wrong.”* (B.B. Mandelbrot)

<http://hot.caltech.edu/topology.html>

- L. Li, D. Alderson, J.C. Doyle, W. Willinger. *Toward a Theory of Scale-Free Networks: Definition, Properties, and Implications*. *Internet Mathematics* 2(4), 2006.
- D. Alderson, L. Li, W. Willinger, J.C. Doyle. *Understanding Internet Topology: Principles, Models, and Validation*. *ACM/IEEE Trans. on Networking* 13(6), 2005.
- J.C. Doyle, D. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger. *The "robust yet fragile" nature of the Internet*. *PNAS* 102(41), 2005.
- D. Alderson and W. Willinger. *A contrasting look at self-organization in the Internet and next-generation communication networks*. *IEEE Comm. Magazine*. July 2005.
- W. Willinger, D Alderson, J.C. Doyle, and L. Li, *More "normal" than Normal: scaling distributions in complex systems*. *Proc. Winter Simulation Conf.* 2004.
- W. Willinger, D Alderson, and L. Li, *A pragmatic approach to dealing with high-variability in network measurements*, *Proc. ACM SIGCOMM IMC* 2004.
- L. Li, D. Alderson, W. Willinger, and J. Doyle, *A first-principles approach to understanding the Internet's router-level topology*, *Proc. ACM SIGCOMM* 2004.