Internet Research with "Big (Internet) Data"

PART I (of II)

Walter Willinger
NIKSUN, Inc.

wwillinger@niksun.com

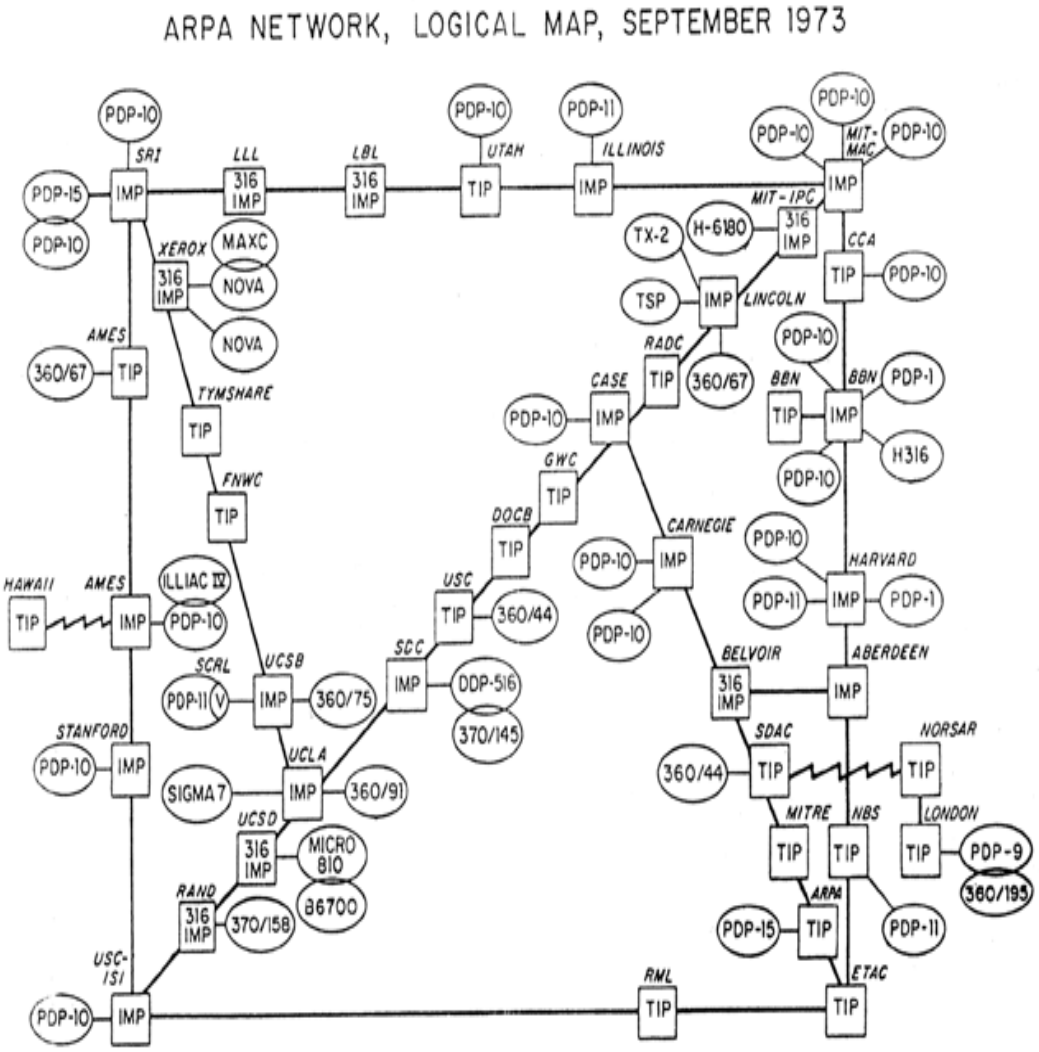
Collaborators

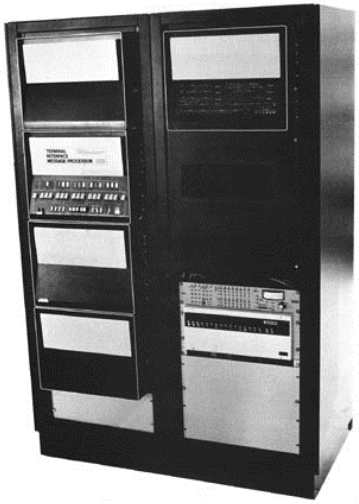
- TU Berlin
  - Anja Feldmann, George Smaragdakis, Philipp Richter
- Akamai/Duke University
  - Nikos Chatzis, Jan Boettger
  - Bruce Maggs, Bala Chandrasekaran
- Northwestern University
  - Fabian Bustamante, Mario Sanchez
- University of Oregon (joint NSF grant, 2013/15)
  - Reza Rejaie, Reza Motamedi
- AT&T Labs-Research
  - Balachander Krishnamurthy, Jeff Erman
- University of Adelaide (joint ARC grant 2011/14)
  - Matt Roughan
- USC/ISI
  - John Heidemann, Xue Cai

Focus on two connectivity structures

- The Internet as a physical construct
  - The Internet as a physical infrastructure
  - Infrastructure = routers/switches and links/cables
  - Router-level topology of the Internet
- The Internet as a logical/virtual construct
  - The Internet as a "network of networks"
  - Network = Autonomous System/Domain (AS)
  - AS-level topology of the Internet

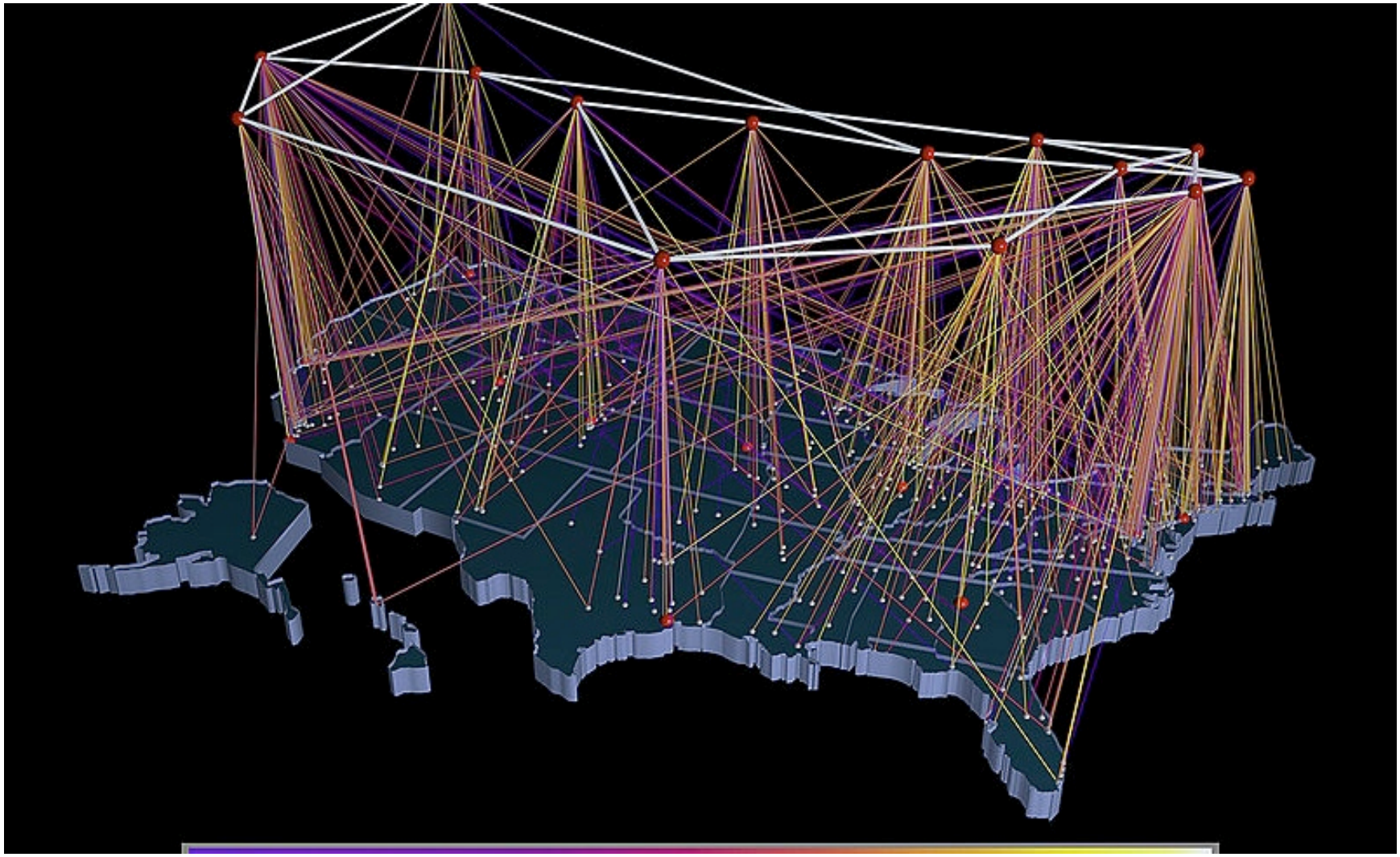
The **physical** Internet (early 1970s)





http://www.computerhistory.org/internet_history/full_size_images/1973_net_map.gif

The **physical** (US) Internet (mid 1990s)



http://commons.wikimedia.org/wiki/File:NSFNET-traffic-visualization-1991.jpg

The **physical** Internet (before 1995)

- Visualization efforts
  - Geography is implicit or explicit
  - The "meaning" of a node is clear
  - Individuals/organizations have complete view of the network

- Insights gained
  - Highly structured connectivity
  - Details matter (e.g., meaning of a node, geography)
  - Rich enough connectivity to "route around failures"

The **physical** Internet (after 1995)

- New challenges (due to decommissioning of NSFNET)
  - No one entity has a complete view of the network
  - The "meaning" of a node has become fuzzy
  - Geography is gone (an after-thought, at best)

- New appealing approach to visualize the physical Internet
  - Step 1: Use traceroute as measurement technique-of-choice
  - Step 2: Perform large-scale traceroute campaigns
  - Step 3: Combine traceroute-derived Internet paths to obtain the Internet's router-level topology

Step 1: traceroute

- Developed by V. Jacobson (1988)
  - Designed to trace out the route to a host
  - Discovers compliant (i.e., IP) routers along path between selected network host computers
- General appeal
  - Everyone can run traceroute
  - traceroute results in lots of useful information

traceroute from NJ to 130.126.0.201

1. wireless_broadband_router (192.168.1.1)
2. 173.63.208.1 (173.63.208.1)
3. g0-3-3-1.nwrknj-lcr-22.verizon-gni.net (130.81.179.194)
4. 130.81.162.84 (130.81.162.84)
5. 0.xe-3-2-0.br2.nyc4.alter.net (152.63.20.213)
6. 204.255.168.114 (204.255.168.114)
7. be2063.mpd22.jfk02.atlas.cogentco.com (154.54.47.57)
8. be2117.mpd22.ord01.atlas.cogentco.com (154.54.7.58)
9. te0-0-2-0.rcr12.ord09.atlas.cogentco.com (154.54.31.230)
10. university-of-illinios-urbana.demarc.cogentco.com (38.104.99.42)
11. t-ch2rtr.ix.ui-iccn.org (72.36.126.77)
12. t-710rtr.ix.ui-iccn.org (72.36.126.81)
13. 72.36.127.86 (72.36.127.86)
14. iccn-ur1rtr-uiuc1.gw.uiuc.edu (72.36.127.2)
15. t-exite1.gw.uiuc.edu (130.126.0.201)

Step 2: traceroute campaigns

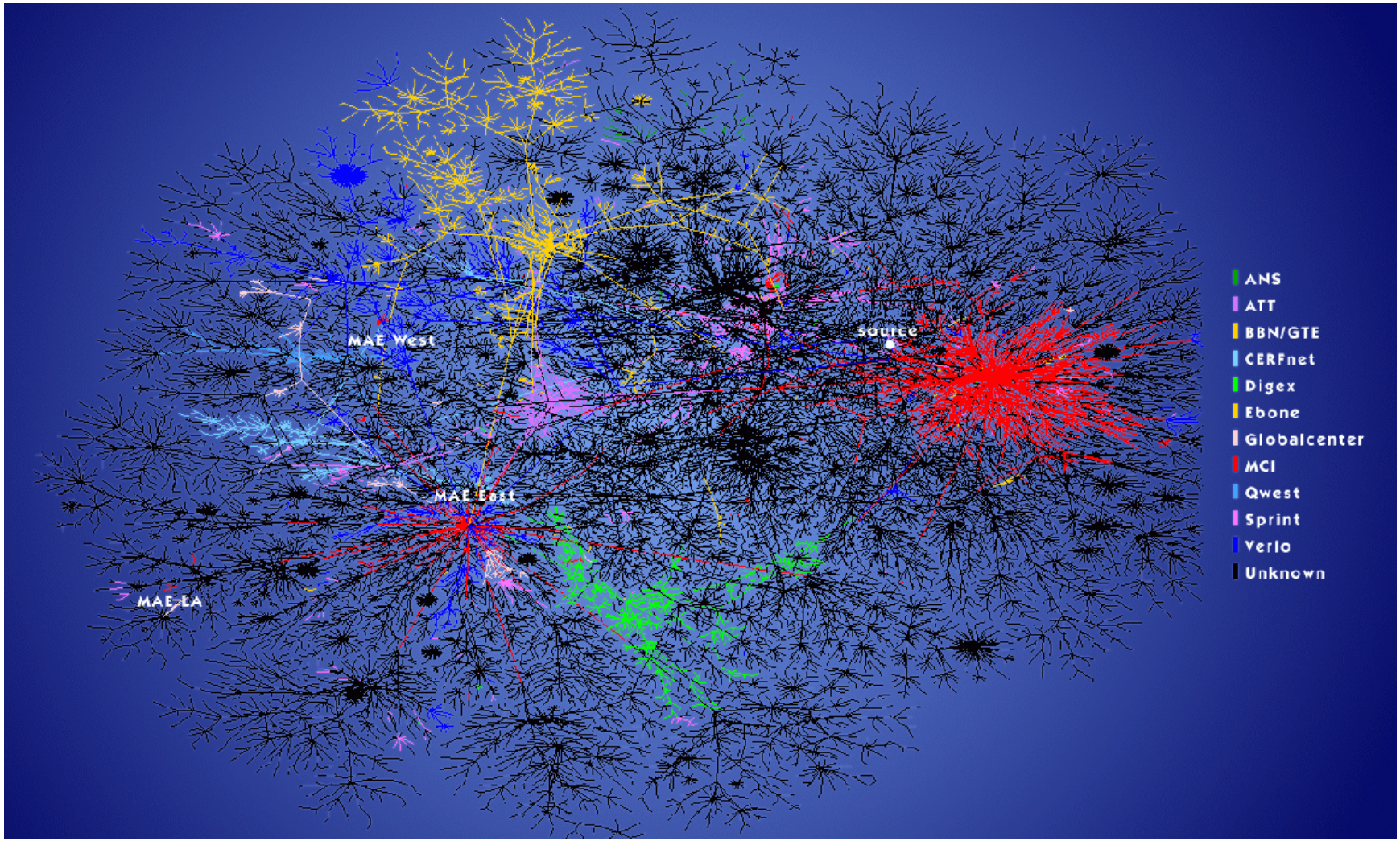
- Perform large-scale traceroute campaign
  - Requires Internet-wide measurement platform/infrastucture
  - Challenge of vantage point selection (sources and targets)
  - First reported large-scale campaign: Pansiot and Grad (1995)

- Example: Archipelago Measurement Infrastructure (Caida)
  - 3 teams (~20 monitors each) independently probe some 20M /24's (full routed IPv4 address space) at 100pps in 2-3days
  - http://www.caida.org/projects/ark/

Step 3: Combine traceroute paths

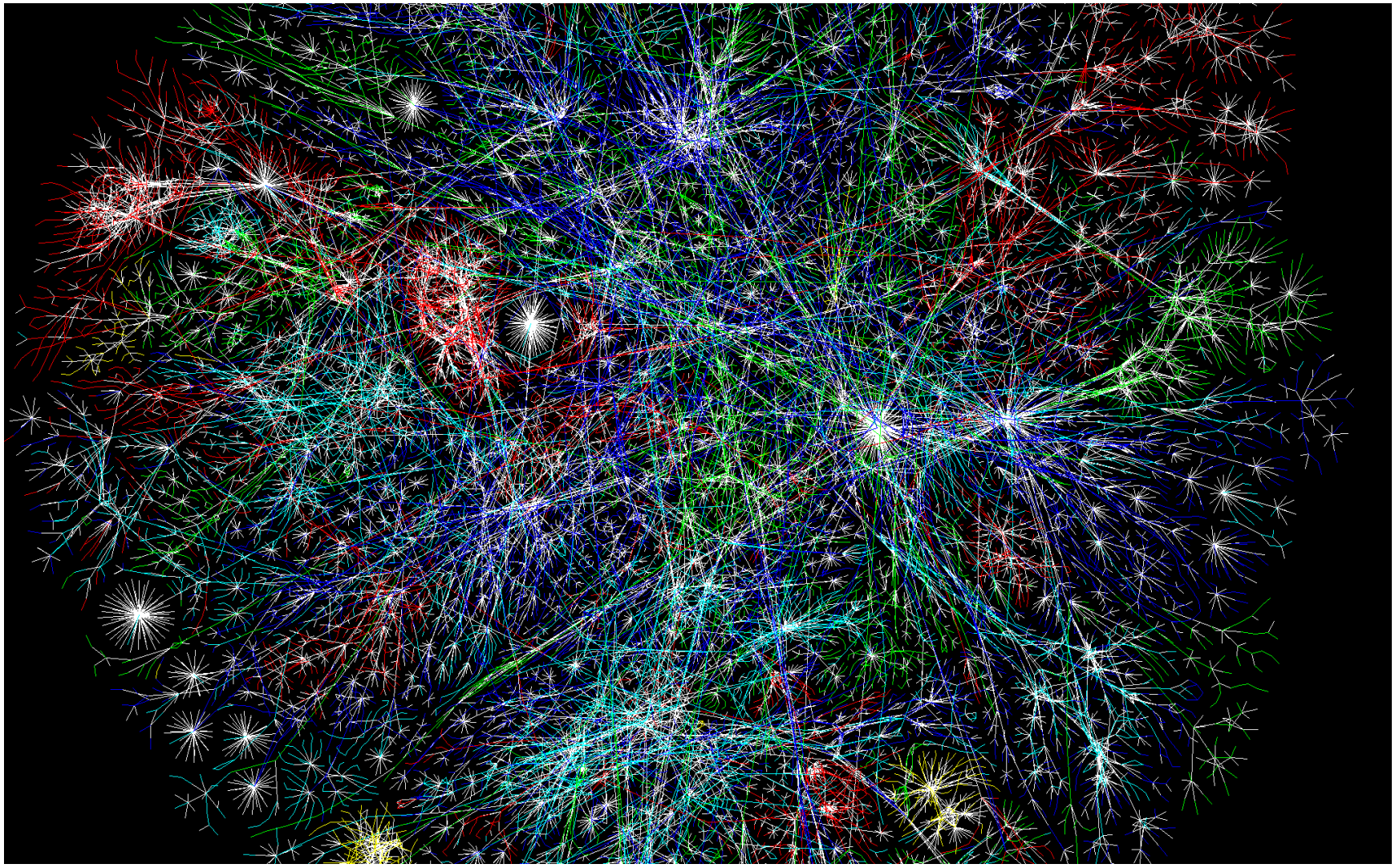
- An early example of "big (Internet) data"
  - Archipelago measurement campaign started in late 2007
  - As of early 2011, the campaign has resulted in some 10 billion traceroute measurements (about 4TB of data) collected from about 60 different vantage points across the Internet
- Working assumption
  - With billions of traceroute-derived Internet paths, it is possible to recover the Internet's router-level topology
  - The produced visualizations provide "insight" into the Internet's router-level topology

The “physical” Internet (late 1990s)



http://www.caida.org/publications/papers/1999/webmatters99/images/caida14_sml.gif

The “physical” Internet (~2010)



http://research.blogs.lincoln.ac.uk/files/2011/02/map-of-internet.png

The "physical" Internet (after 1995)

- "Insights" and "discoveries"
  - Random (e.g., scale-free) graphs appear to be suitable models
  - There are "obvious" high-degree nodes in the Internet
  - Removal of high-degree nodes is an "obvious" vulnerability
  - Discovery of the Internet's "Achilles' heel"

- Questions and issues
  - What is the quality of this "big (traceroute) data" …?
  - How do the new "insights" compare to Internet reality …?
  - What exactly is "physical" about the resulting Internet maps …?

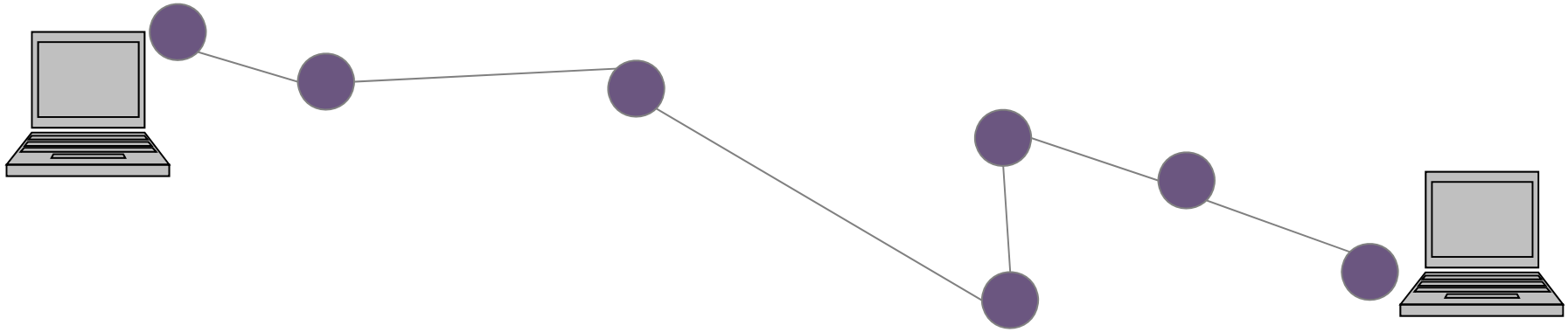
Getting to know your data …

- The "Network Scientist's" perspective
  - Available data is taken at face value ("don't ask …")
  - No or only little domain knowledge is required
  - The outcome often leaves little room for further efforts
- The "Engineer's" perspective
  - Available data tends to be scrutinized (not enough, though)
  - Domain knowledge is "king" – details matter!
  - The results often give rise to new questions/problems

Internet Router-level Connectivity

- Nodes
  - IP routers or switches
- Links
  - Physical connection between two IP routers or switches
- Measurement technique
  - `traceroute` tool
  - traceroute discovers compliant (i.e., IP) routers along path between selected network host computers

The Network Scientist's View

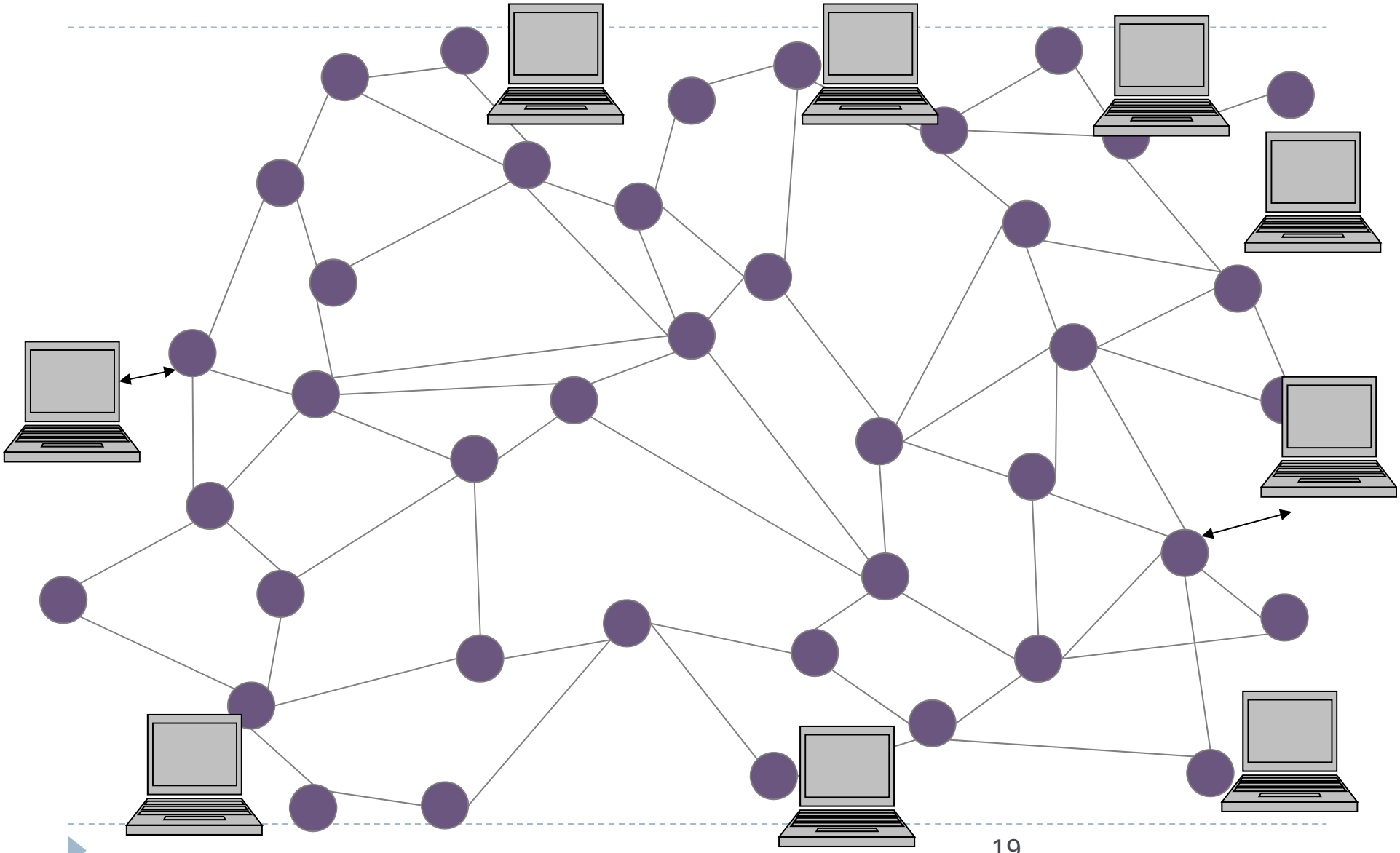


- Basic "experiment"
  - Select a source and destination
  - Run traceroute tool
- Example
  - Run traceroute from my machine in Florham Park, NJ, USA to www.iet.unipi.it

Run traceroute from NJ to www.iet.unipi.it

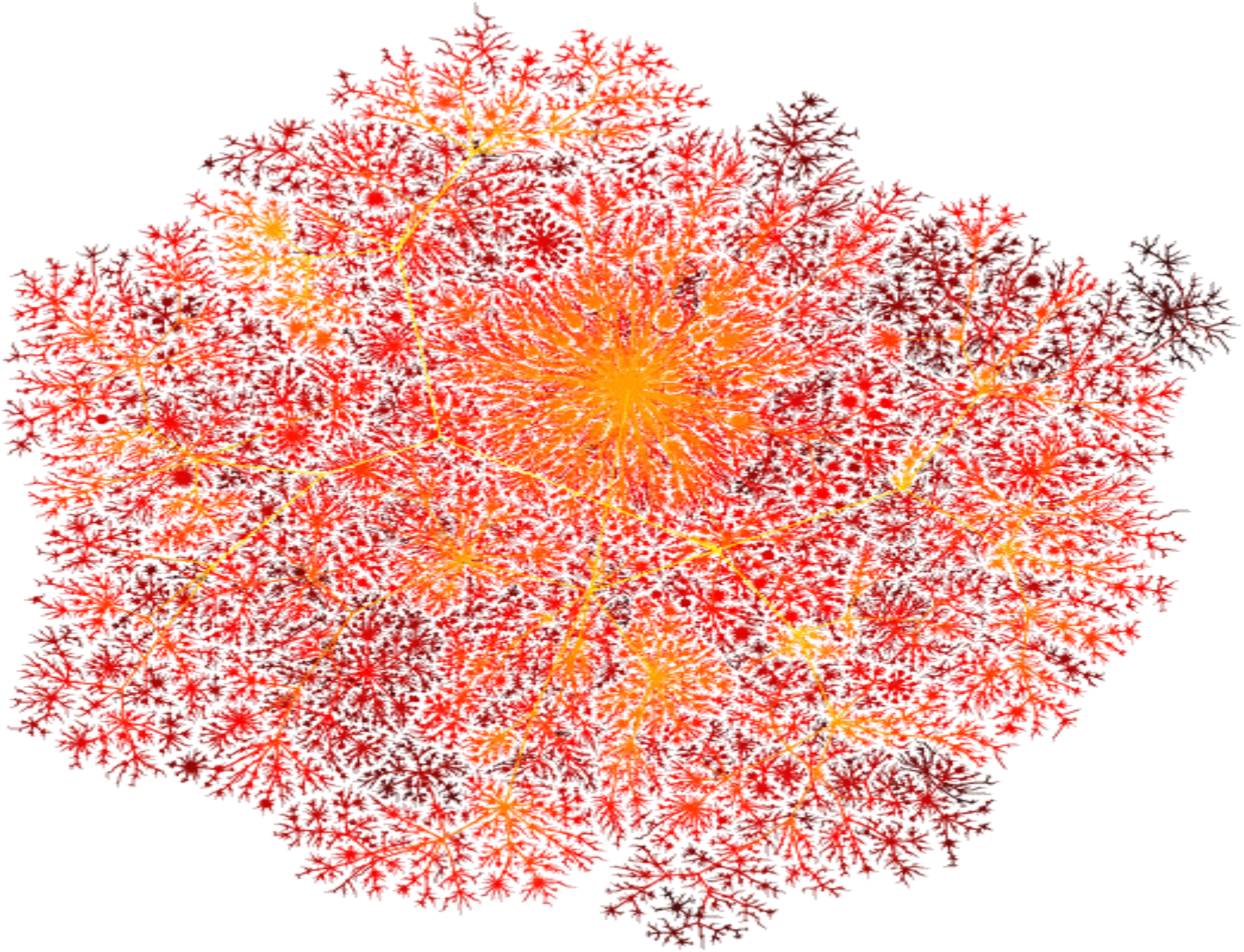
- 1 135.207.176.3 2 ms 1 ms 1 ms
- 2 fp-core.research.att.com (135.207.3.1) 1 ms 1 ms 1 ms
- 3 ngx19.research.att.com (135.207.1.19) 1 ms 0 ms 0 ms
- 4 12.106.32.1 1 ms 1 ms 1 ms
- 5 12.119.12.73 2 ms 2 ms 2 ms
- 6 cr2.n54ny.ip.att.net (12.122.130.94) 4 ms 3 ms 3 ms
- 7 ggr4.n54ny.ip.att.net (12.122.130.33) 3 ms 3 ms 3 ms
- 8 192.205.34.54 3 ms 3 ms 3 ms
- 9 nyk-bb1-link.telia.net (80.91.249.17) 3 ms 3 ms 3 ms
- 10 prs-bb1-link.telia.net (80.91.251.97) 89 ms 89 ms 89 ms
- 11 mno-b1-link.telia.net (80.91.249.39) 101 ms 101 ms 101 ms
- 12 213.248.71.162 96 ms 96 ms 96 ms
- 13 rt-mi2-rt-to1.to1.garr.net (193.206.134.42) 98 ms 98 ms 98 ms
- 14 rt-to1-rt-pi1.pi1.garr.net (193.206.134.74) 132 ms 132 ms 132 ms
- 15 rt-pi1-ru-unipi-1.pi1.garr.net (193.206.136.14) 133 ms 133 ms 133 ms
- 16 ing-ser.unipi.it (131.114.191.130) 143 ms 144 ms 143 ms
- 17 docenti.ing.unipi.it (131.114.28.20) 133 ms 133 ms 133 ms

The Network Scientist's View (cont.)

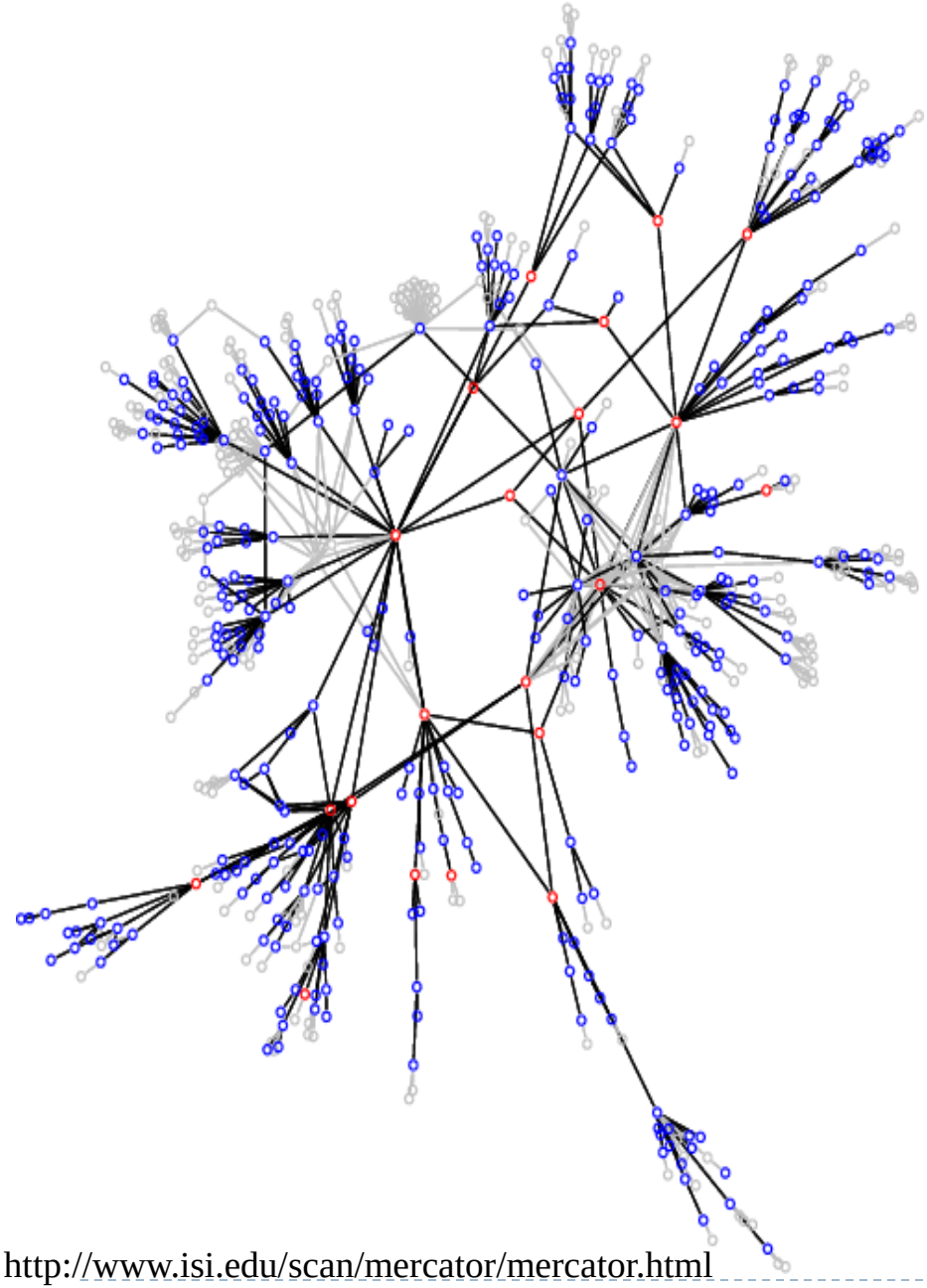


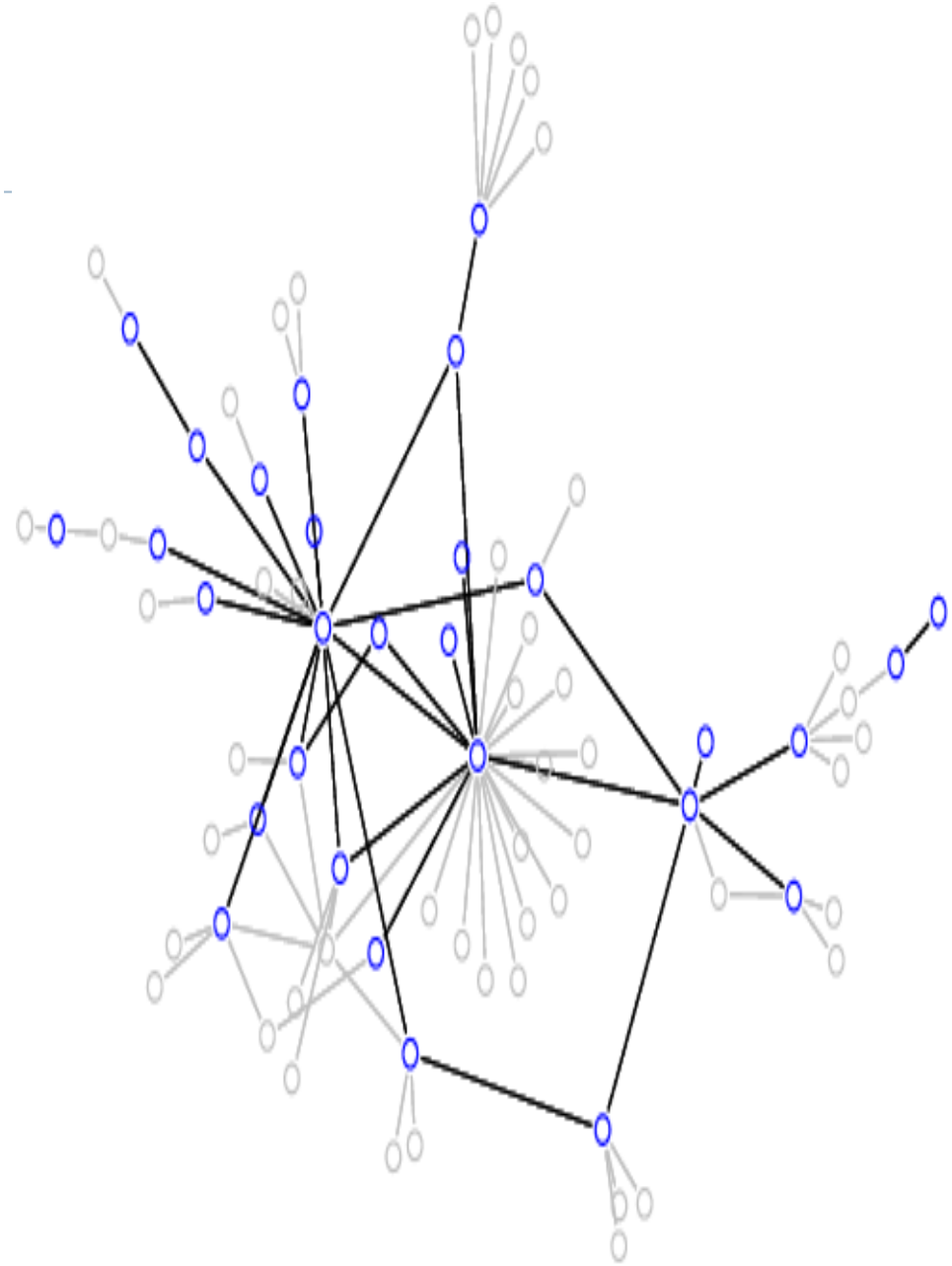
The Network Scientist's View (cont.)

- Measurement technique
  - `traceroute` tool
  - traceroute discovers compliant (i.e., IP) routers along path between selected network host computers
- Available data: from large-scale traceroute experiments
  - Pansiot and Grad (router-level, around 1995, France)
  - Cheswick and Burch (mapping project 1997--, Bell-Labs)
  - Mercator (router-level, around 1999, USC/ISI)
  - Skitter (CAIDA/UCSD), became Ark (in 2008)
  - Rocketfuel (early 2000, router-level maps of ISPs, UW Seattle)
  - Dimes (ongoing EU project)
  - TraceNet, xnet (~2008, Univ. of Texas)
  - Ono (~2008, Northwestern Univ.)
  - Merlin (~2010, Univ. of Strasbourg)
  - ...



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

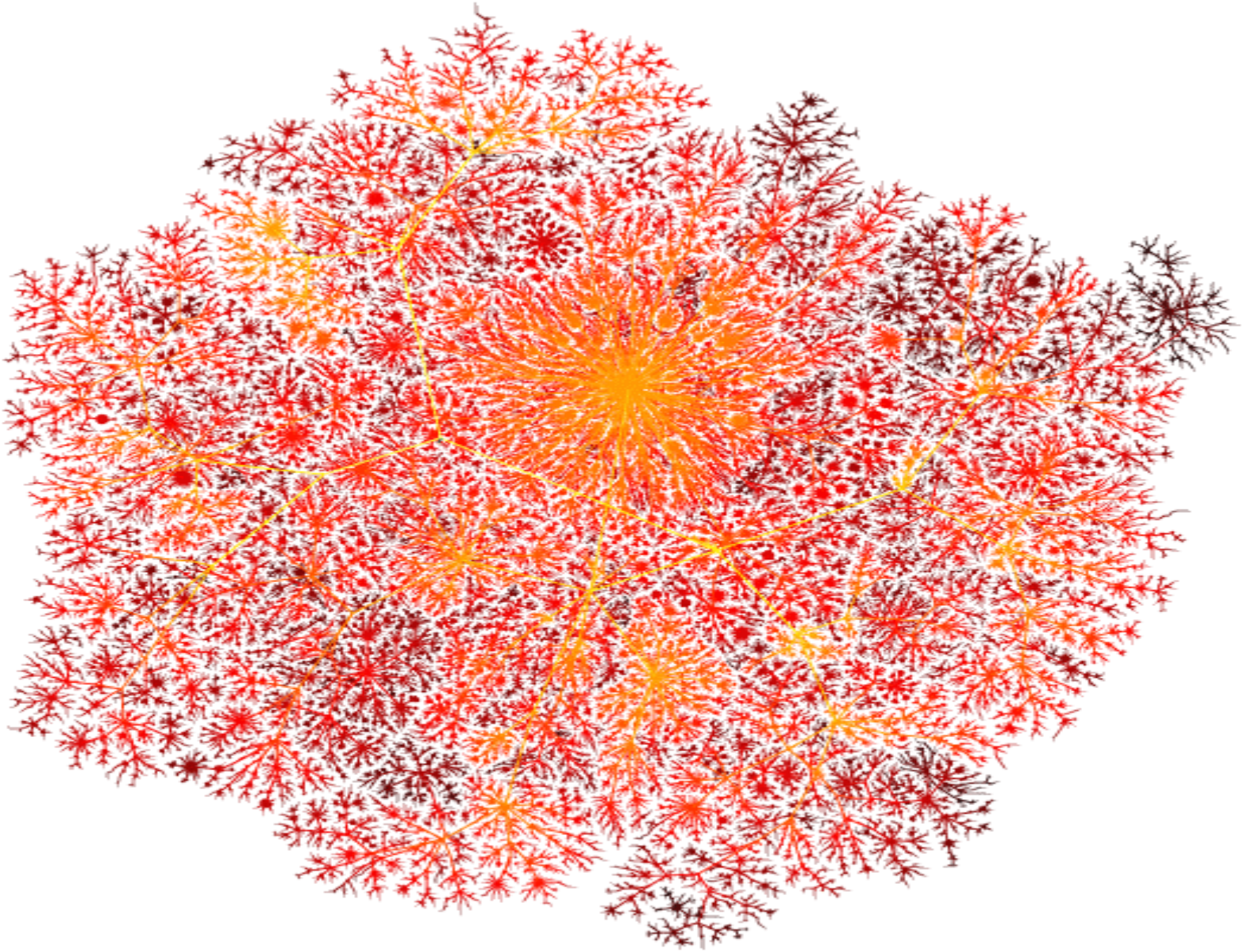




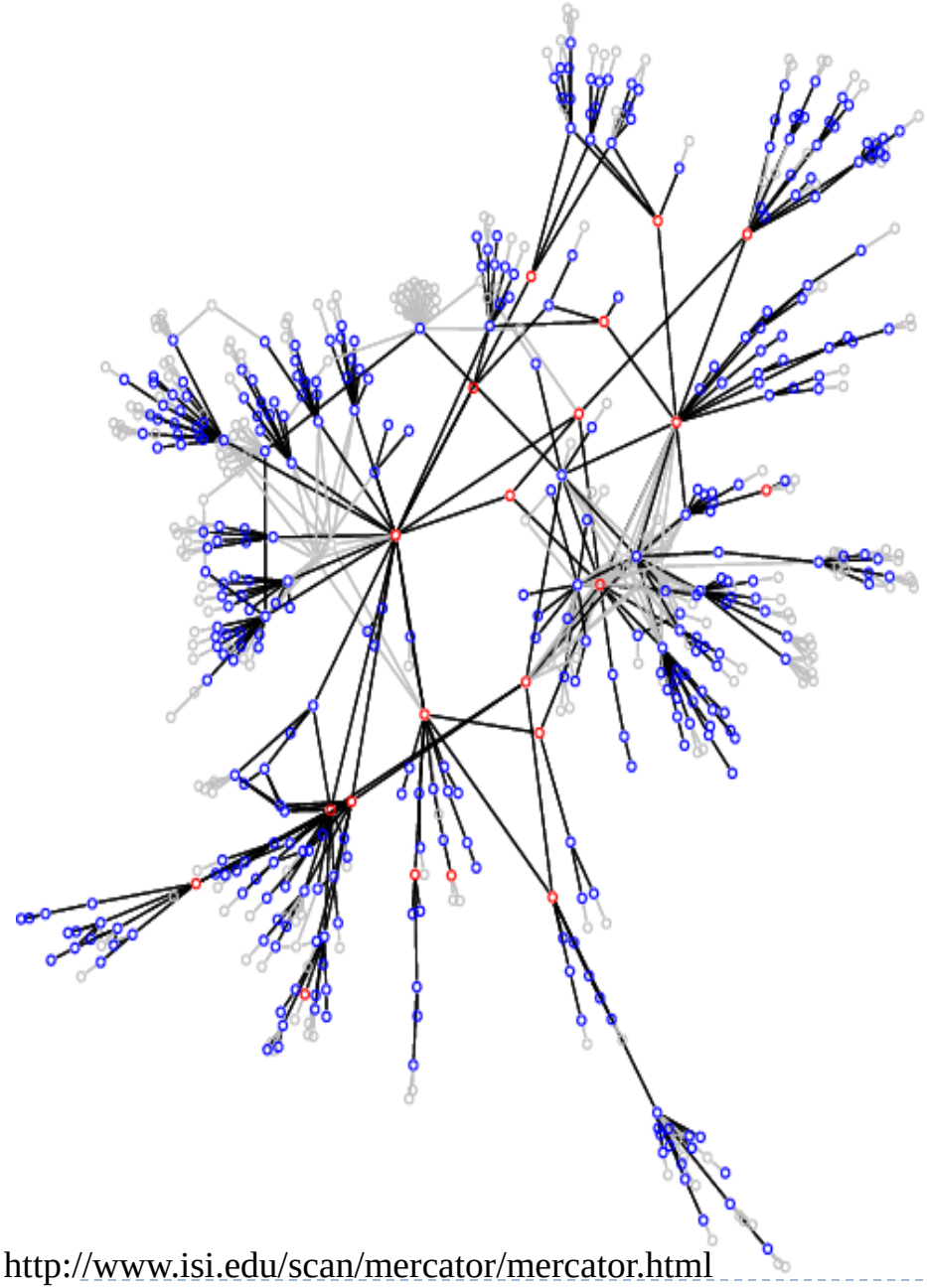
http://www.isi.edu/scan/mercator/mercator.html

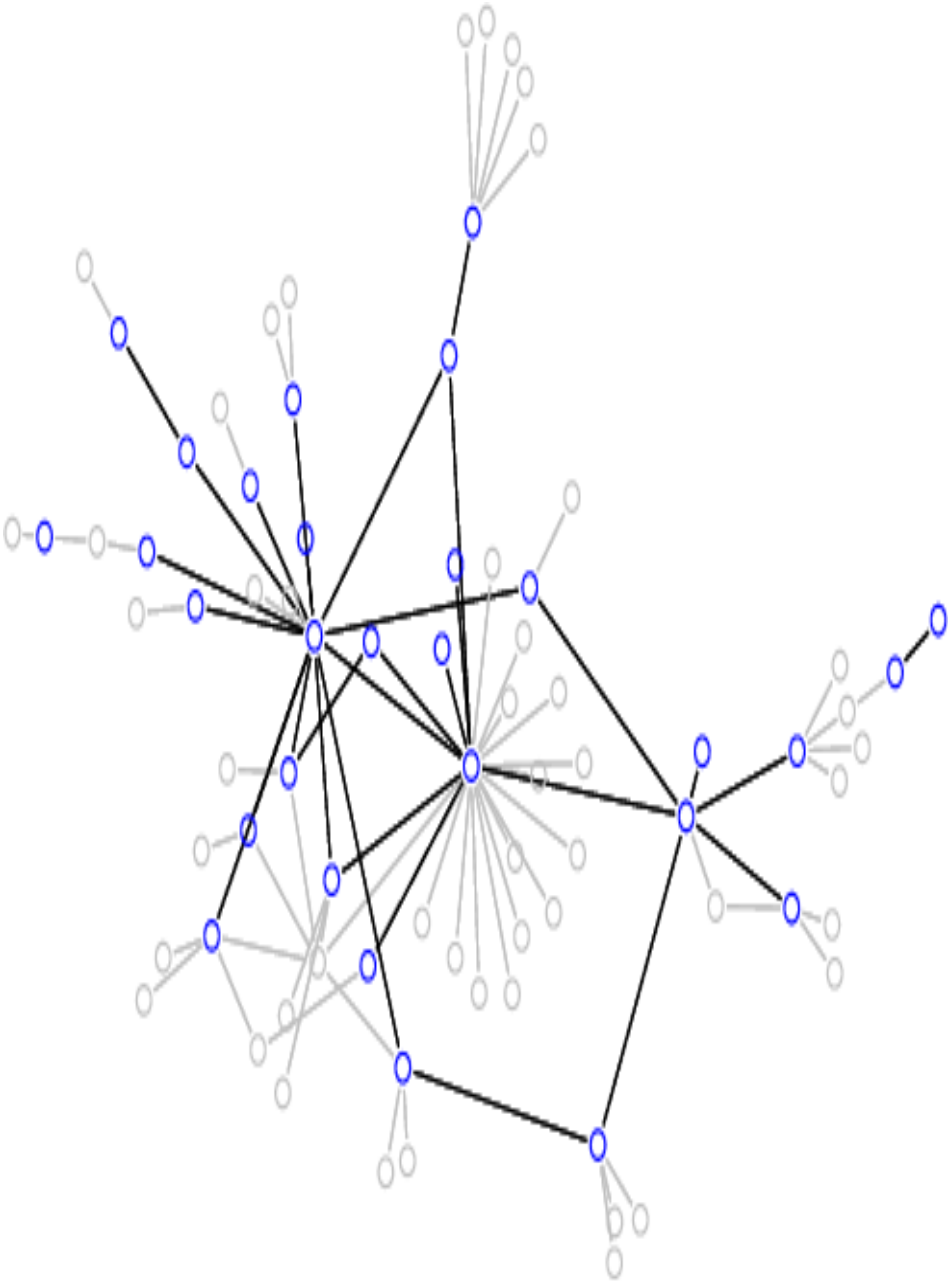
The Network Scientist's View (cont.)

- Inference
  - Given: traceroute-based map (graph) of the router-level Internet (Internet service provider)
  - Wanted: Metric/statistics that characterizes the inferred connectivity maps
  - Main metric-of-choice: Node degree distribution



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





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

The Network Scientist's View (cont.)

- Inference
  - Given: traceroute-based map (graph) of the router-level Internet (Internet service provider)
  - Wanted: Metric/statistics that characterizes the inferred connectivity maps
  - Main metric-of-choice: Node degree distribution

- Modeling
  - Power-law node degree distributions
  - Scale-free networks …
- Predictions …
  - The Achilles' heel of the Internet ….

The Engineer's View

- Measurement technique
  - `traceroute` tool
  - traceroute discovers compliant (i.e., IP) routers along path between selected network host computers
  - The reported IP addresses are not the routers' IP addresses, but the IP addresses of the routers' interfaces (outgoing packet)

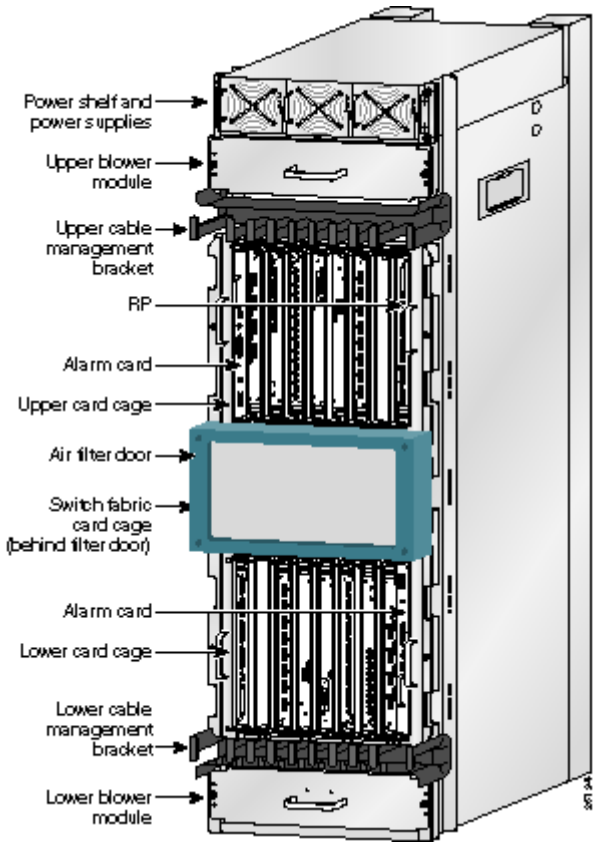
Run trace route from NJ to www.iet.unipi.it

- 1 135.207.176.3 2 ms 1 ms 1 ms
- 2 fp-core.research.att.com (135.207.3.1) 1 ms 1 ms 1 ms
- 3 ngx19.research.att.com (135.207.1.19) 1 ms 0 ms 0 ms
- 4 12.106.32.1 1 ms 1 ms 1 ms
- 5 12.119.12.73 2 ms 2 ms 2 ms
- 6 cr2.n54ny.ip.att.net (12.122.130.94) 4 ms 3 ms 3 ms
- 7 ggr4.n54ny.ip.att.net (12.122.130.33) 3 ms 3 ms 3 ms
- 8 192.205.34.54 3 ms 3 ms 3 ms
- 9 nyk-bb1-link.telia.net (80.91.249.17) 3 ms 3 ms 3 ms
- 10 prs-bb1-link.telia.net (80.91.251.97) 89 ms 89 ms 89 ms
- 11 mno-b1-link.telia.net (80.91.249.39) 101 ms 101 ms 101 ms
- 12 213.248.71.162 96 ms 96 ms 96 ms
- 13 rt-mi2-rt-to1.to1.garr.net (193.206.134.42) 98 ms 98 ms 98 ms
- 14 rt-to1-rt-pi1.pi1.garr.net (193.206.134.74) 132 ms 132 ms 132 ms
- 15 rt-pi1-ru-unipi-1.pi1.garr.net (193.206.136.14) 133 ms 133 ms 133 ms
- 16 ing-ser.unipi.it (131.114.191.130) 143 ms 144 ms 143 ms
- 17 docenti.ing.unipi.it (131.114.28.20) 133 ms 133 ms 133 ms

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.

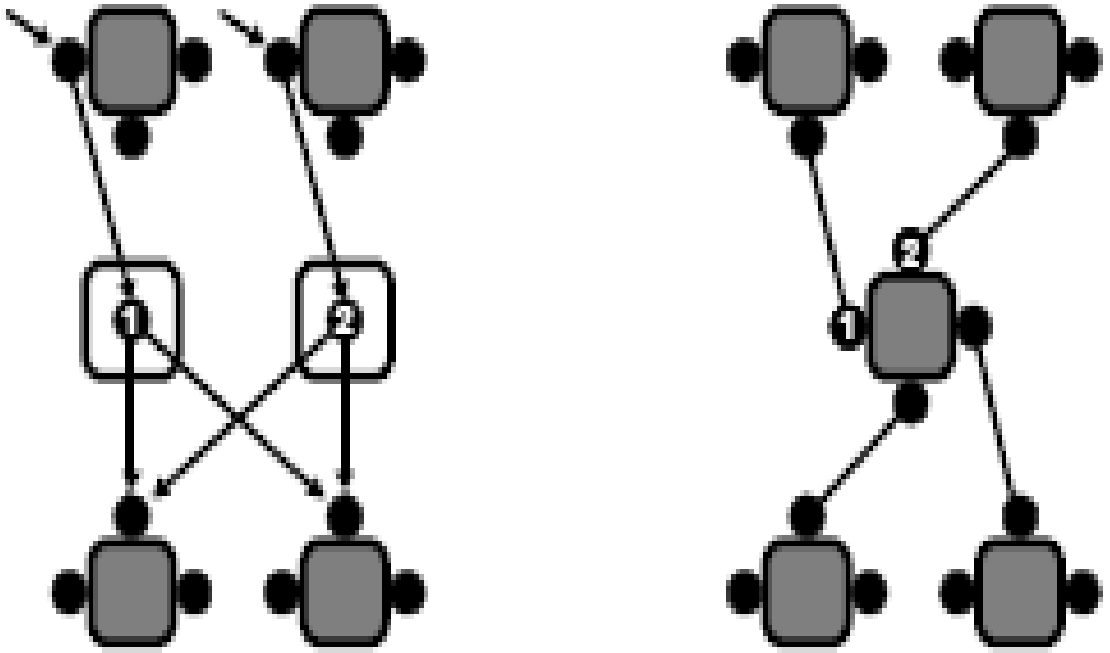
Chassis	Rack size	Slots	Switching Capacity
12416	Full	16	320 Gbps
12410	1/2	10	200 Gbps
12406	1/4	6	120 Gbps
12404	1/8	4	80 Gbps



Source: www.cisco.com

The Engineer's View (cont.)

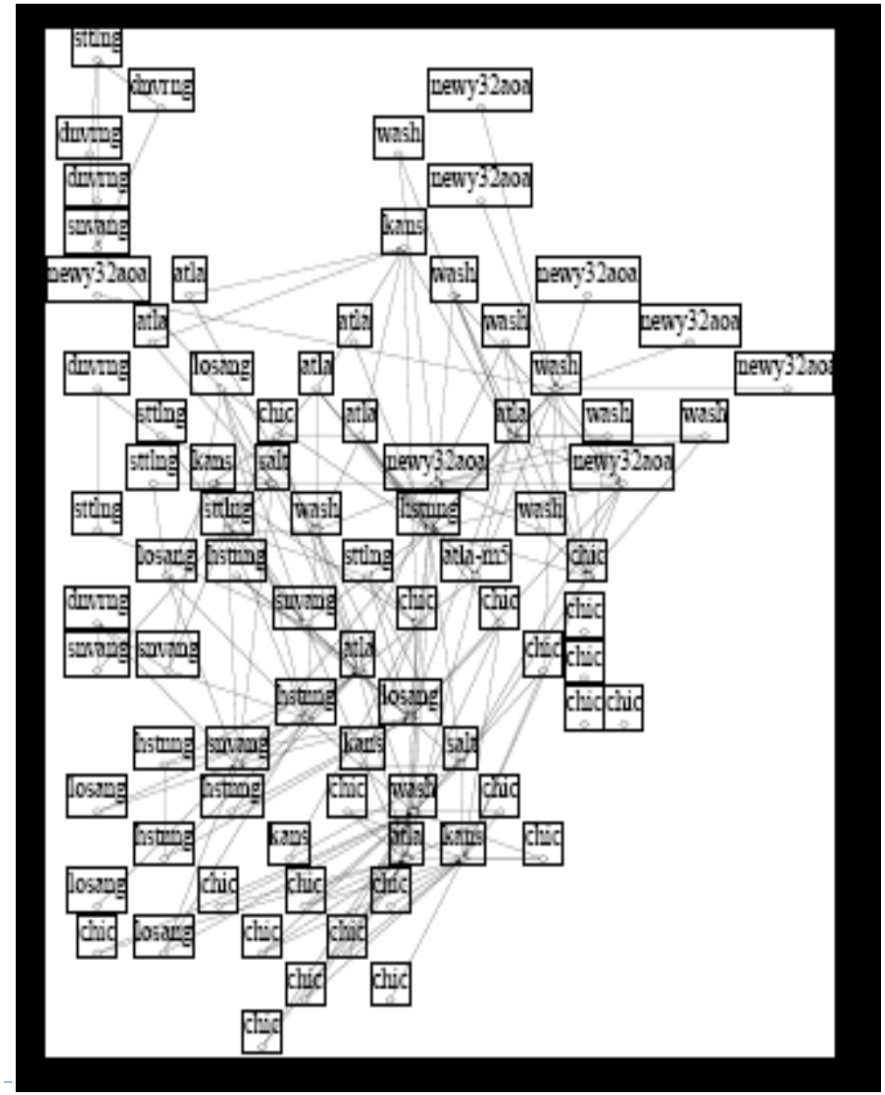
- traceroute is strictly about IP-level connectivity
  - Originally developed by Van Jacobson (1988)
  - Designed to trace out the route to a host
- Using traceroute to map the router-level topology
  - Engineering hack
  - Example of what we can measure, not what we want to measure!
- Basic problem #1: IP alias resolution problem
  - How to map interface IP addresses to IP routers
  - Largely ignored or badly dealt with in the past
  - New efforts in 2008 for better heuristics …



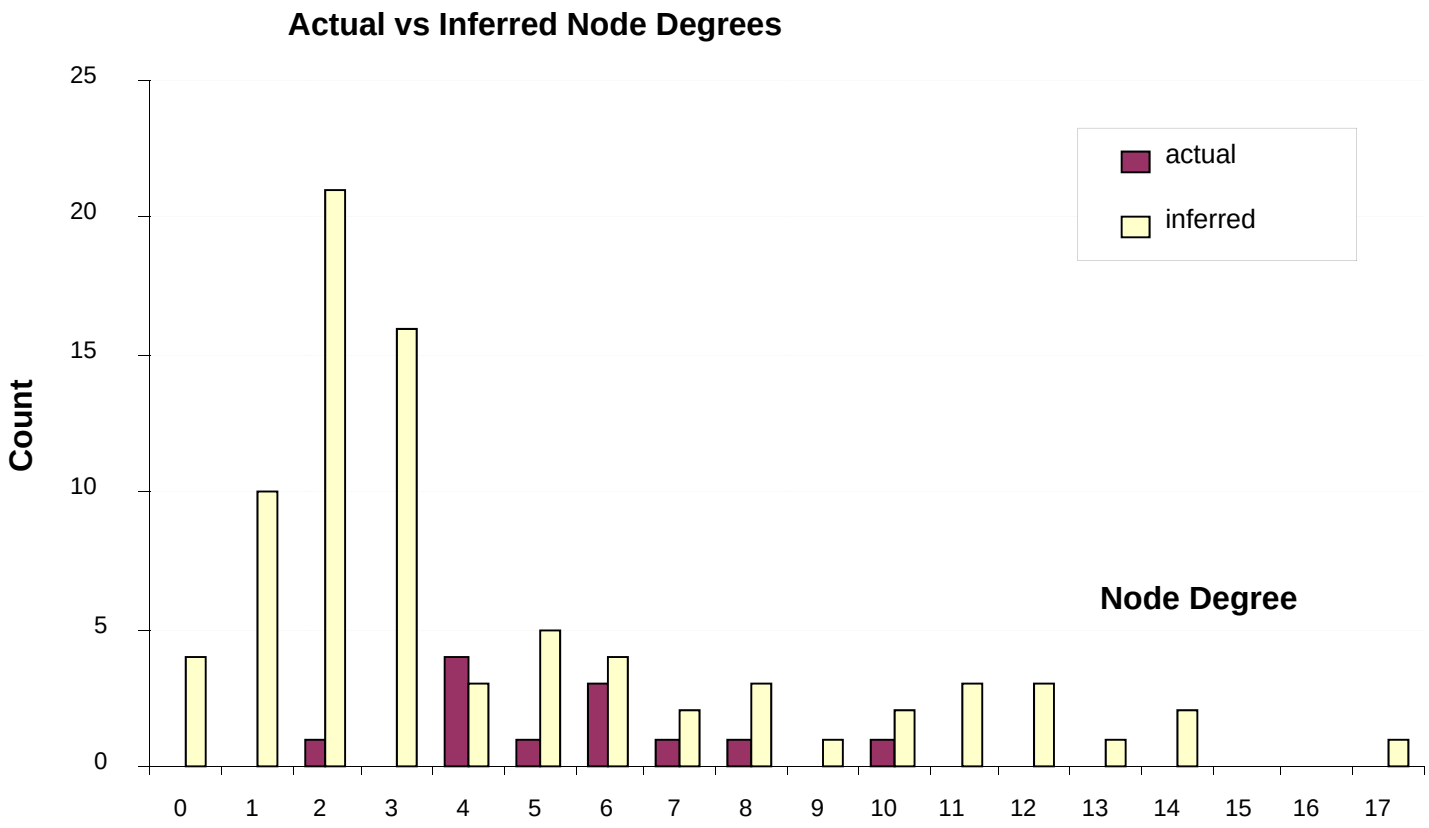
Interfaces 1 and 2 belong to the same router

Example: Abilene Network

Reality check: Abilene/Internet2

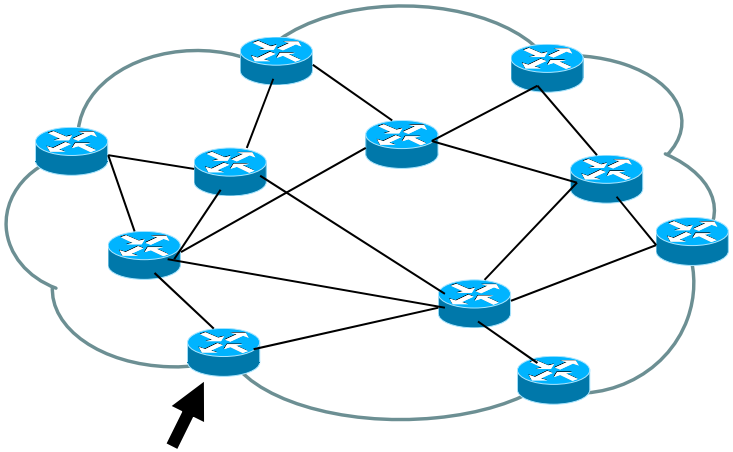


 Courtesy Adam Bender

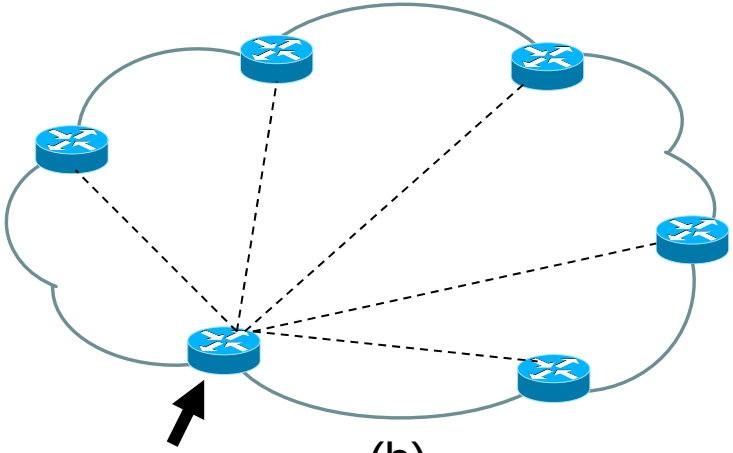

Actual vs Inferred Node Degrees
Count
Node Degree
actual
inferred
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

The Engineer's View (cont.)

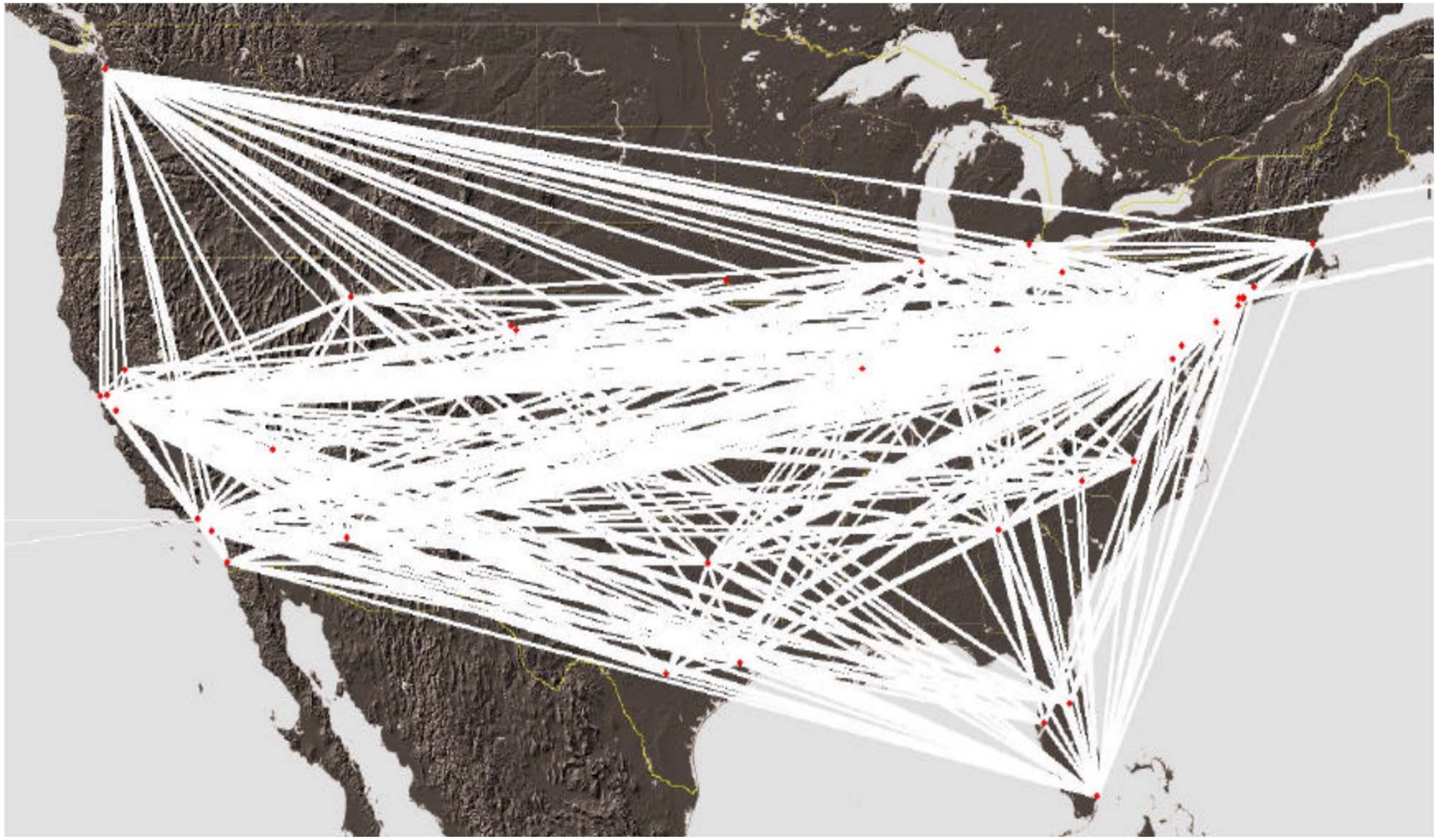
- traceroute is strictly about IP-level connectivity
- Basic problem #2: Layer-2 technologies (e.g., MPLS, ATM)
  - MPLS is an example of a circuit technology that hides the network's physical infrastructure from IP
  - Sending traceroutes through an opaque Layer-2 cloud results in the "discovery" of high-degree nodes, which are simply an artifact of an imperfect measurement technique.
  - This problem has been largely ignored in all large-scale traceroute experiments to date.



(a)

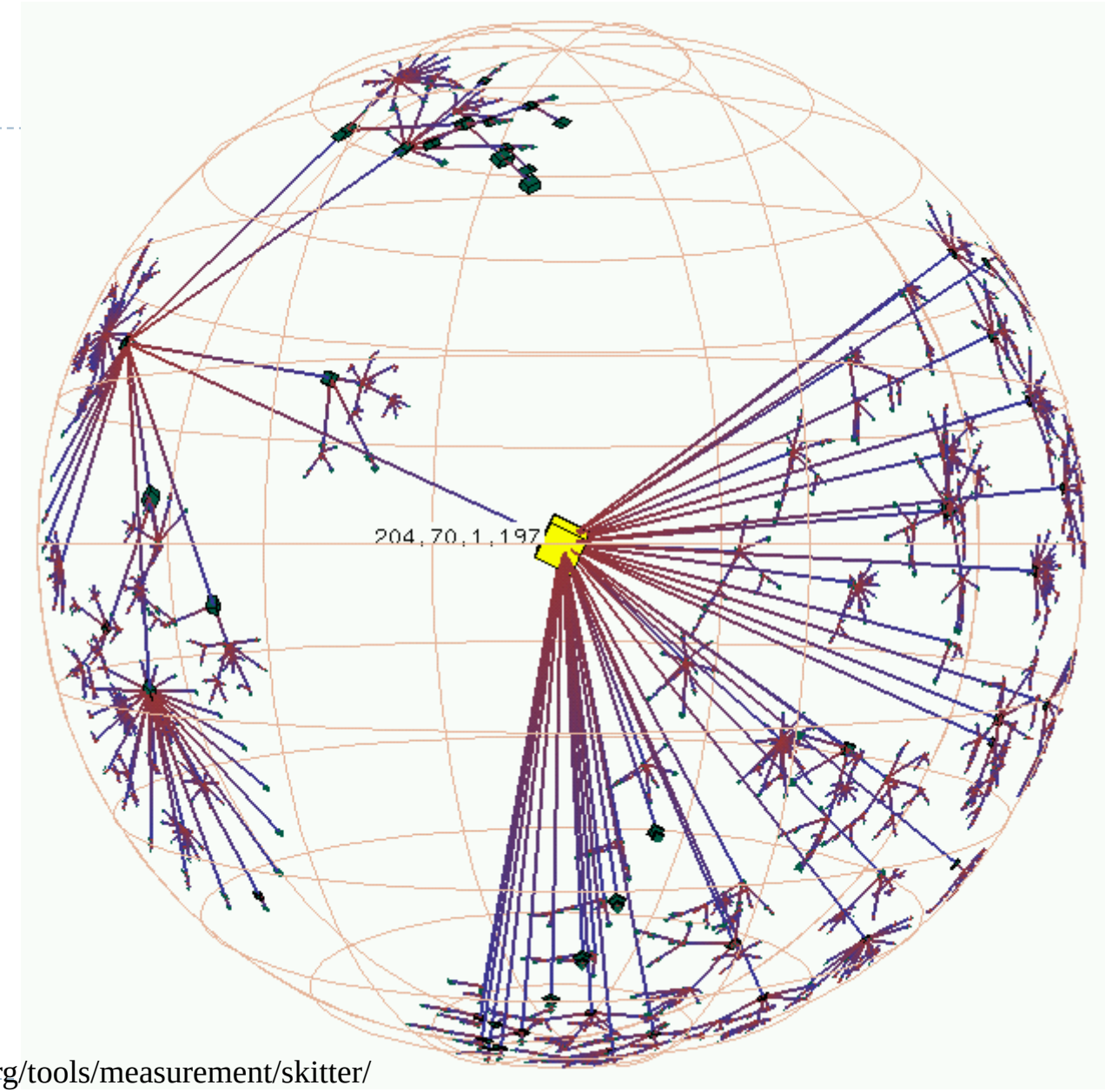


(b)

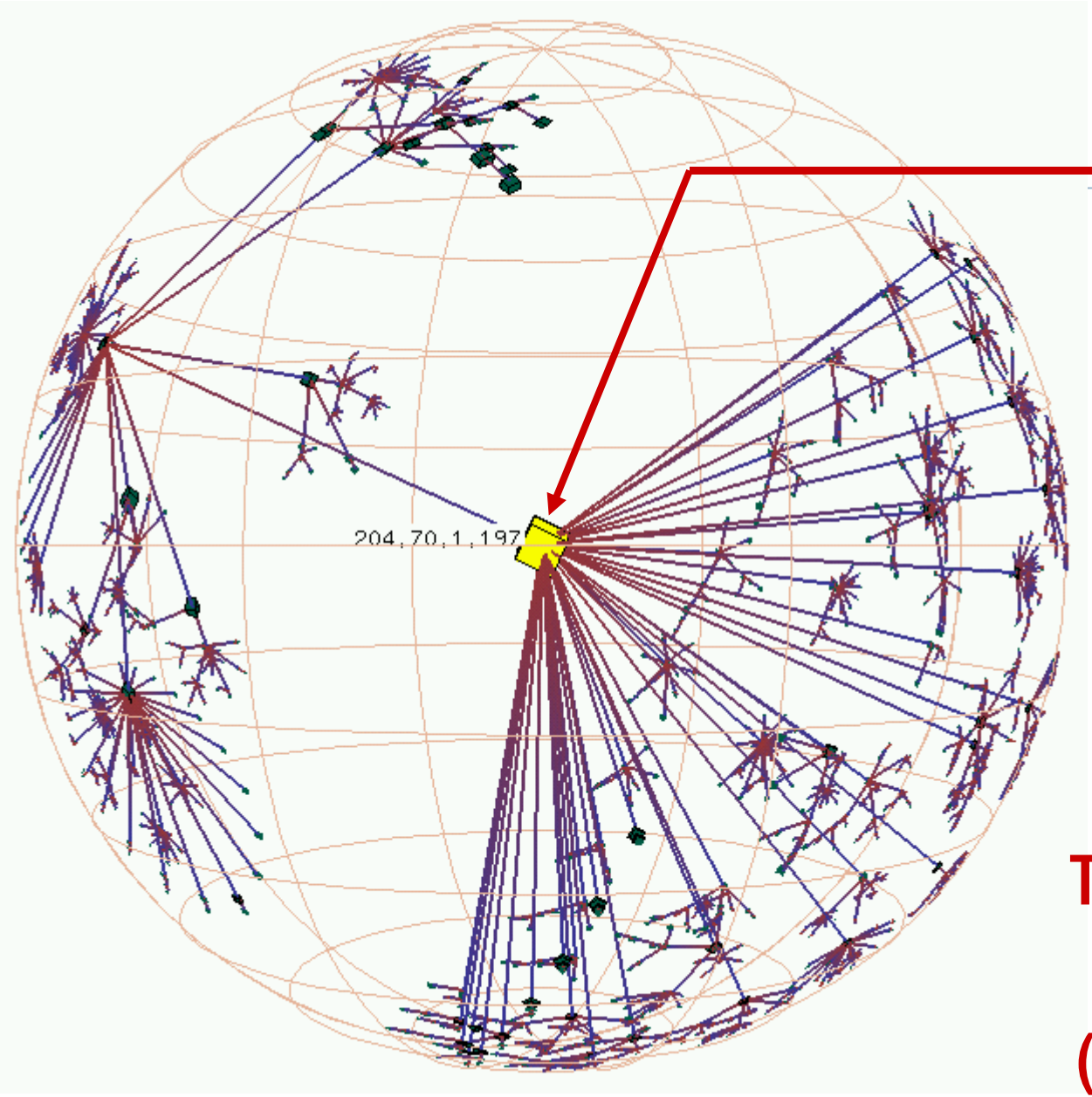


Background image courtesy JHU, applied physics labs

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



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



- 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)

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

The Engineer's View (cont.)

- Additional sources of errors
  - Bias in (mathematical abstraction of) traceroute
  - Has been a major focus within CS/Networking literature
  - Non-issue in the presence of above-mentioned problems

- The irony of traceroute measurements
  - The high-degree nodes in the middle of the network that traceroute reveals are not for real …
  - If there are high-degree nodes in the network, they can only exist at the edge of the network where they will never be revealed by generic traceroute-based experiments …

The Engineer's View on traceroute data

- Bottom line
  - (Current) traceroute measurements are of little use for accurately mapping router-level connectivity
  - Unless significant progress is made, it is unlikely that future traceroute measurements will be more useful for the purpose of router-level mapping

- Lessons learned
  - Key question: Can you trust the available data?
  - Critical role of Data Hygiene in the Petabyte Age
  - Corollary: Petabytes of garbage = garbage
  - Data hygiene is often viewed as "dirty/unglamorous" work

But all this was well-known …!

- J.-J. Pansiot and D. Grad, 1998. On routes and multicast trees in the Internet. Computer Communication Review 28 (1), 41—50.

- From the Pansiot & Grad paper to the "discovery" of the "scale-free Internet"

Recap: Step 1 - Measurements

On Routes and Multicast Trees in the Internet

Jean-Jacques PANSIOT

Dominique GRAD

Université Louis Pasteur - LSIIT URA-CNRS 1871
Computer Science Department
7, rue Descartes 67084 Strasbourg Cedex, France
{pansiot, grad}@dpt-info.u-strasbg.fr
http ://dpt-info.u-strasbg.fr/~{pansiot, grad}

Abstract : **Multicasting has an increasing importance for network applications such as groupware or videoconferencing. Several multicast routing protocols have been defined. However they cannot be used directly in the Internet since most inter-domain routers do no implement multicasting. Thus these protocols are mainly tested either on a small scale inside a domain, or through the Mbone, whose topology is not really the same as Internet topology. The purpose of this paper is to construct a graph using actual routes of the Internet, and then to use this graph to compare some parameters - delays, scaling in term of state or traffic concentration - of multicast routing trees constructed by different algorithms - source shortest path trees and shared trees.**

Key words : Routing, routes, Internet, multicast, shortest path trees, centered trees

Introduction

Multicast routing is an active research area. The problem is to transmit a data packet from one source to K receivers. have therefore no state information to maintain. Newer protocols, usable on a larger scale are now developed. Some are based on a unique centered tree per group, such as CBT [BFC 93], others may also include source rooted trees, such as PIM-SM [EFD 97]. In these two cases, routers not part of a tree do not incur any cost for maintaining trees. On the other hand, intermediate routers with degree 2 in the multicast tree must maintain tree state and signaling, although their role is only to forward multicast packets in much the same way as unicast packets. Solutions [GPZ 96] have been proposed to free these degree 2 nodes from any cost in maintaining multicast trees.

The goal of this paper is twofold. Firstly to get some experimental data on the shape of multicast trees one can actually obtain in Internet: node degree, route length,... These data could be used in particular to calibrate tree and graph generators used to simulate or validate network protocols. Secondly to get more directly usable information for people working on multicast tree construction. For example, are there many nodes of degree 2 ? Are trees rooted in different sources in the same graph very different

Reference: J.-J. Pansiot and D. Grad, 1998. On routes and multicast trees in the Internet. Computer Communication Review 28 (1), 41–50.

Recap: Step 2 - Analysis

On Power-Law Relationships of the Internet Topology

Michalis Faloutsos
U.C. Riverside
Dept. of Comp. Science
michalis@cs.ucr.edu

Petros Faloutsos
U. of Toronto
Dept. of Comp. Science
pfal@cs.toronto.edu

Christos Faloutsos *
Carnegie Mellon Univ.
Dept. of Comp. Science
christos@cs.cmu.edu

Abstract

Despite the apparent randomness of the Internet, we discover some surprisingly simple power-laws of the Internet topology. These power-laws hold for three snapshots of the Internet, between November 1997 and December 1998, despite a 45% growth of its size during that period. We show that our power-laws fit the real data very well resulting in correlation coefficients of 96% or higher.

Our observations provide a novel perspective of the structure of the Internet. The power-laws describe concisely skewed distributions of graph properties such as the node outdegree. In addition, these power-laws can be used to estimate important parameters such as the average neighborhood size, and facilitate the design and the performance analysis of protocols. Furthermore, we can use them to generate and select realistic topologies for simulation purposes.

1 Introduction

"What does the Internet look like?" "Are there any topological properties that don't change in time?" "How will it look like a year from now?" "How can I generate Internet-like graphs for my simulations?" These are some of the questions motivating this work.

In this paper, we study the topology of the Internet and we identify several power-laws. Furthermore, we discuss

hops) that are useful for the analysis of protocols and for speculations of the Internet topology in the future.

Modeling the Internet topology[1] is an important open problem despite the attention it has attracted recently. Paxson and Floyd consider this problem as a major reason "Why We Don't Know How To Simulate The Internet" [16]. Several graph-generator models have been proposed [23] [5] [27], but the problem of creating realistic topologies is not yet solved; the selection of several parameter values are left to the intuition and the experience of each researcher.

As our primary contribution, we identify three power-laws for the topology of the Internet over the duration of a year in 1998. Power-laws are expressions of the form $y \propto x^a$, where a is a constant, x and y are the measures of interest, and $\propto$ stands for "proportional to". Some of those exponents do not change significantly over time, while some exponents change by approximately 10%. However, the important observation is the existence of power-laws, i.e., the fact that there is *some* exponent for each graph instance. During 1998, these power-laws hold in three Internet instances with good linear fits in log-log plots; the correlation coefficient of the fit is at least 96% and usually higher than 98%. In addition, we introduce a graph metric to quantify the density of a graph and propose a rough power-law approximation of that metric. Furthermore, we show how to use our power-laws and our approximation to estimate useful parameters of the Internet, such as the average number of neighbors

Reference: M. Faloutsos, P. Faloutsos, and C. Faloutsos, 1999. On power-law relationships in the Internet topology. Proc. ASM Sigcomm '99, Computer Communication Review 29 (4), 251–262.

Recap: Step 3 - Modeling

The Internet's Achilles' Heel:

Error and attack tolerance of complex networks

Réka Albert, Hawoong Jeong, Albert-László Barabási

Department of Physics, University of Notre Dame, Notre Dame, IN 46556

Many systems that we perceive as truly complex display an amazing degree of tolerance against errors. For example, relatively simple organisms - such as various species of bacteria - grow, persist and reproduce despite large variations in their environment, or drastic pharmaceutical interventions, an error tolerance attributed to the robustness of the underlying cellular (metabolic) network [1]. The increasingly complex communication networks responding to the demand generated by the addition of diverse communication devices to the Internet [2] display a surprising degree of robustness: while key components (routers, lines) regularly malfunction, local failures rarely lead to the loss of the global information-carrying ability of the network. The stability of these and other complex systems against local errors and failures is often attributed to the redundant wiring of the functional web defined by the systems' components, guaranteeing multiple alternative routes between most pairs of nodes. In this paper we demonstrate that such error tolerance is not shared by all redundant systems, but it is displayed only by a class of inhomogeneously wired networks, called scale-free networks. We find that scale-free networks, describing a number of systems, such as the www [3–5], Internet [6], social networks [7] or a cell [8], display an unexpected degree of robustness, the ability of their nodes to communicate being unaffected by even unrealistically high failure rates. However, this error tolerance comes at a high price: these networks are extremely vulnerable to attacks, i.e. to the selection and removal of a few nodes that play the most

Reference: R. Albert, H. Jeong, A.-L. Barabasi, 2000. The Internet's Achilles' heel: Error and attack tolerance of complex networks. Nature 406, 378–382.

Recap: Step 4 – Prediction/Implications

Cover Story: Nature 406, 2000.

Revisiting Pansiot & Grad 1998 paper

- The purpose for performing their traceroute measurements is explicitly stated

The goal of this paper is twofold. Firstly to get some experimental data on the shape of multicast trees one can actually obtain in Internet: node degree, route length,... These data could be used in particular to calibrate tree and graph generators used to simulate or validate network protocols. Secondly to get more directly usable information for people working on multicast tree construction. For example, are there many nodes of degree 2 ? Are trees rooted in different sources in the same graph very different ? In the following, we are interested in sparse groups, that is groups where the average distance between members is high, and with membership ranging up to a few thousands.

In Section 1, we describe how we constructed a graph from actual Internet routes. We mention some problems we found in tracing routes, and we discuss the realism of the graph we obtained. In Section 2, we analyze more precisely the structure of our graph. In Section 3, we compare different types of multicast trees such as source rooted shortest path trees (SPT) or shared trees (ST), in terms of scalability. We compare for example the average delay,

Reference: J.-J. Pansiot and D. Grad, 1998. On routes and multicast trees in the Internet. Computer Communication Review 28 (1), page 41.

Revisiting Pansiot & Grad 1998 paper

- The purpose for performing their traceroute measurements is explicitly stated

- The main problems with the traceroute measurements are explicitly mentioned (IP alias resolution and Layer-2 technology)

Traceroute basically produces the list of IP addresses (and when this is possible, domain names) of routers along the route. For leaves of the graph (that is sources and destinations), we considered only nodes whose domain name was known. However for intermediate nodes, we also kept nodes known only by their IP address. In practice, over more than 10 000 different IP addresses, more than 1000 (10%) remained anonymous (failure of the inverse DNS query). A more serious problem is to determine if two identifiers (name or address) correspond to the same node or not. One may assume that if two different addresses have the same name, they correspond to the same node (via different interfaces). Unfortunately, the converse is not true, two different names (such as border2-hssi1-0.chicago.mci.net and border2-fddi-0.chicago.mci.net) may correspond to two different interfaces of the same host. Worse, for two different addresses, one cannot tell a priori if they correspond to the same host.

In theory, a solution could be to query all addresses using SNMP to discover the address of other interfaces. In practice this is not generally feasible, in particular because routers do not permit SNMP access from everywhere. We have adopted a partial solution, based on the fact that when a router sends an ICMP message [Pos81b], it generally uses as source address the address of the emitting interface, rather than the address of the interface where the original packet arrived. Therefore, we have sent an UDP packet with an unused port number (same principle as *traceroute*) to all IP addresses obtained by *traceroute*.

We then verified if the source address of the ICMP Port Unreachable message (say A) was the same as the destination address of the UDP packet (say B). If this is not the case, A and B are two addresses of the same node. Note that this is likely to occur since we trace routes using source routing. In the above example, A is the interface of the normal route to the router, whereas B is the incoming interface of a source route. With this method around 200 synonyms (different addresses of the same host) were found. Obviously this method is not perfect, and in our resulting graph, some apparently different nodes are actually the same.

Reference: J.-J. Pansiot and D. Grad, 1998. On routes and multicast trees in the Internet. Computer Communication Review 28 (1), page 43.

If we look back at our original data, containing all routes before destination selection (see 1.2), we find nodes with higher degrees : 45 (node connected to the German academic X25 network Win) and 37 (node connected to the English academic SMDS network Janet). These networks use IP over a switched circuit technology. All routers connected to such networks are potential direct neighbors at the IP level. Therefore there is almost no limit on the degree of a node even if the number of physical interfaces is limited. This phenomenon may become even more common with the widespread use of ATM networks in large network backbones. More generally graph edges may correspond to:

- a point to point link between two nodes
- a link within a broadcast network, such an Ethernet or Fddi LAN. Note that these LANs may be found not only on user's sites, but also within backbones for router interconnection.
- a link within a non broadcast multiple access (NBMA) network, such as X25, SMDS, Frame relay or ATM. It could be also a pure switched circuit network such as the phone network.

Reference: J.-J. Pansiot and D. Grad, 1998. On routes and multicast trees in the Internet. Computer Communication Review 28 (1), pages 45/46.

Revisiting Pansiot & Grad 1998 paper

- The purpose for performing their traceroute measurements is explicitly stated
- The main problems with the traceroute measurements are explicitly mentioned (IP alias resolution and Layer-2 technology)
- The Pansiot and Grad paper is an early textbook example for what information a measurement paper should provide.

Revisiting Pansiot & Grad 1998 paper

- The purpose for performing their traceroute measurements is explicitly stated
- The main problems with the traceroute measurements are explicitly mentioned (IP alias resolution and Layer-2 technology)
- The Pansiot and Grad paper is an early textbook example for what information a measurement paper should provide.
- Unfortunately, subsequent papers in this area have completely ignored the essential details provided by Pansiot and Grad and ultimately don't even cite this work anymore!

Although we focus on the Internet topology at the inter-domain level, we also examine an instance at the router level. The graph represents the topology of the routers of the Internet in 1995, and was tediously collected by Pansiot and Grad [14].

- Rout-95: the routers of the Internet in 1995 with 3888 nodes, 5012 edges, and an average outdegree of 2.57.

Clearly, the above graph is considerably different from the first three graphs. First of all, the graphs model the topology at different levels. Second, the Rout-95 graph comes from a different time period, in which Internet was in a fairly early phase.

Reference: M. Faloutsos, P. Faloutsos, and C. Faloutsos, 1999. On power-law relationships in the Internet topology. Proc. ASM Sigcomm ' 99, Computer Communication Review 29 (4), p. 253.

The increasing availability of topological data on large networks, aided by the computerization of data acquisition, had led to great advances in our understanding of the generic aspects of network structure and development[9–16]. The existing empirical and theoretical results indicate that complex networks can be divided into two major classes based on their connectivity distribution $P(k)$, giving the probability that a node in the network is connected to k other nodes. The first class of networks is characterized by a $P(k)$ that peaks at an average $\langle k \rangle$ and decays exponentially for large k. The most investigated examples of such exponential networks are the random graph model of Erdös and Rényi[9,10] and the small-world model of Watts and Strogatz[11], both leading to a fairly homogeneous network, in which each node has approximately the same number of links, $k \simeq \langle k \rangle$. In contrast, results on the World-Wide Web (WWW)[3–5], the Internet[6] and other large networks[17–19] indicate that many systems belong to a class of inhomogeneous networks, called scale-free networks, for which $P(k)$ decays as a power-law, that is $P(k) \sim k^{-\gamma}$, free of a characteristic scale. Whereas the probability that a node has a very large number of connections ($k \gg \langle k \rangle$) is practically prohibited in exponential networks, highly connected nodes are statistically significant in scale-free networks (Fig. 1).

Reference: R. Albert, H. Jeong, A.-L. Barabasi, 2000. The Internet's Achilles' heel: Error and attack tolerance of complex networks. Nature 406, 378—382.

Faloutsos *et al.*[6] investigated the topological properties of the Internet at the router and inter-domain level, finding that the connectivity distribution follows a power-law, $P(k) \sim k^{-2.48}$. Consequently, we expect that it should display the error tolerance and attack vulnerability predicted by our study. To test this, we used the latest survey of the Internet topology, giving the network at the inter-domain (autonomous system) level. Indeed, we find that the diameter of the Internet is unaffected by the random removal of as high as 2.5% of the nodes (an order of magnitude larger than the failure rate (0.33%) of the Internet routers[23]), whereas if the same percentage of the most connected nodes are eliminated (attack), d more than triples (Fig. 2b). Similarly, the large connected cluster persists for high rates of random node removal, but if nodes are removed in the attack mode, the size of the fragments that break off increases rapidly, the critical point appearing at $f_c^1 \simeq 0.03$ (Fig. 3b).

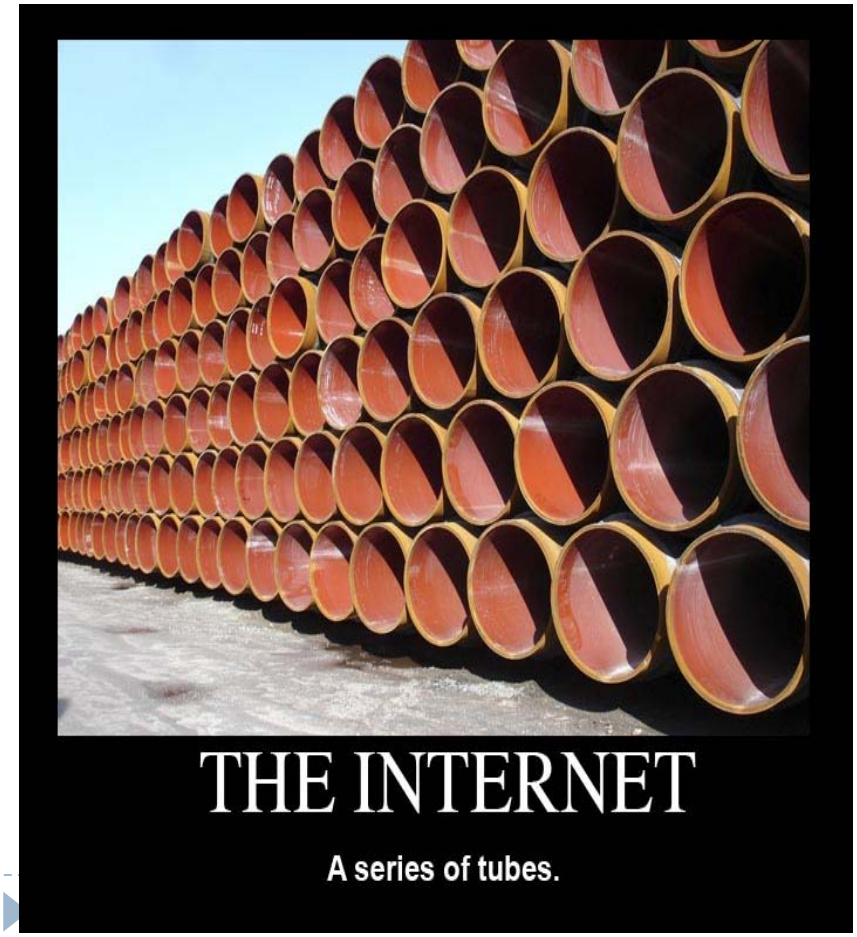
Reference: R. Albert, H. Jeong, A.-L. Barabasi, 2000. The Internet's Achilles' heel: Error and attack tolerance of complex networks. Nature 406, 378–382.

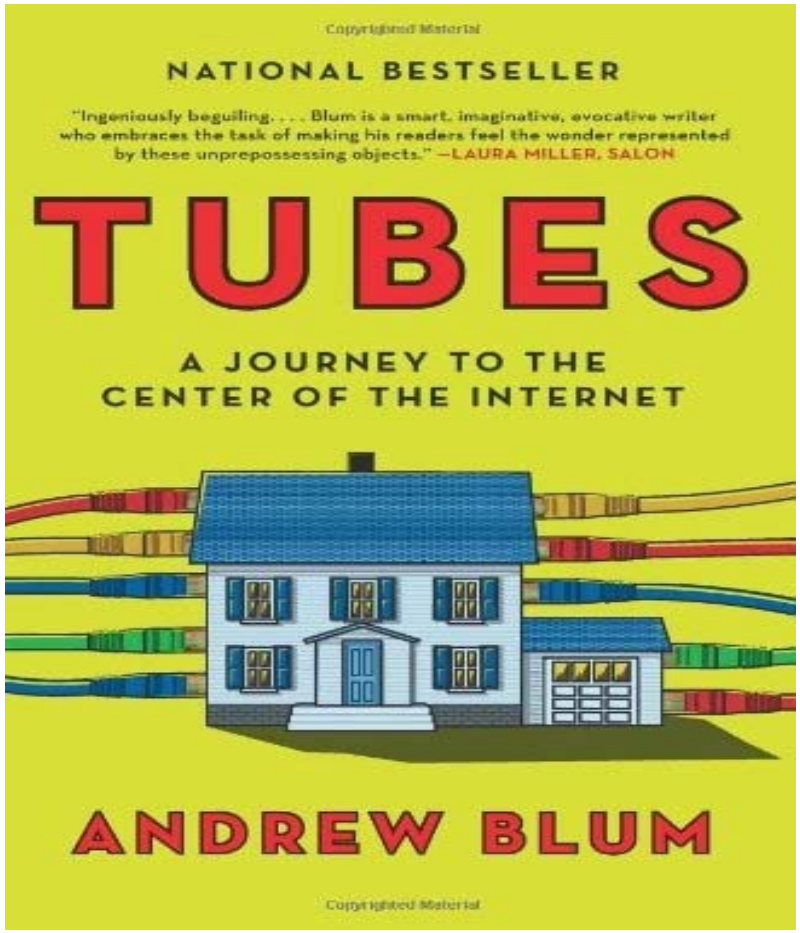
Discussion

- **Details do matter!**
- **Implications for analyzing traceroute data?**
- **Every networking student knows the problems with traceroute, so why is this domain knowledge not used?**
- **Why has the 1998 Pansiot and Grad paper never been referenced in subsequent Internet topology papers?**

Revisiting the **physical** Internet

- Renewed public interest





Revisiting the **physical** Internet

- Renewed public interest

- The physical aspects of the "physical" Internet
  - Renewed focus on the "meaning" of a node
  - Bring back geography
  - Emphasis on structure and not randomness

- Alternative data sources
  - traceroute measurements as one of many potential sources
  - Use other (publicly) available information

Back to basics: From routers/switches …

$$$$$

$$$$$$$

$

$$ to $$$$$

… to racks/cabinets/cages/suites …

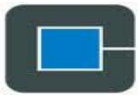
… to colocation (colo) companies …

EQUINIX

DIGITAL REALTY
Data Center Solutions

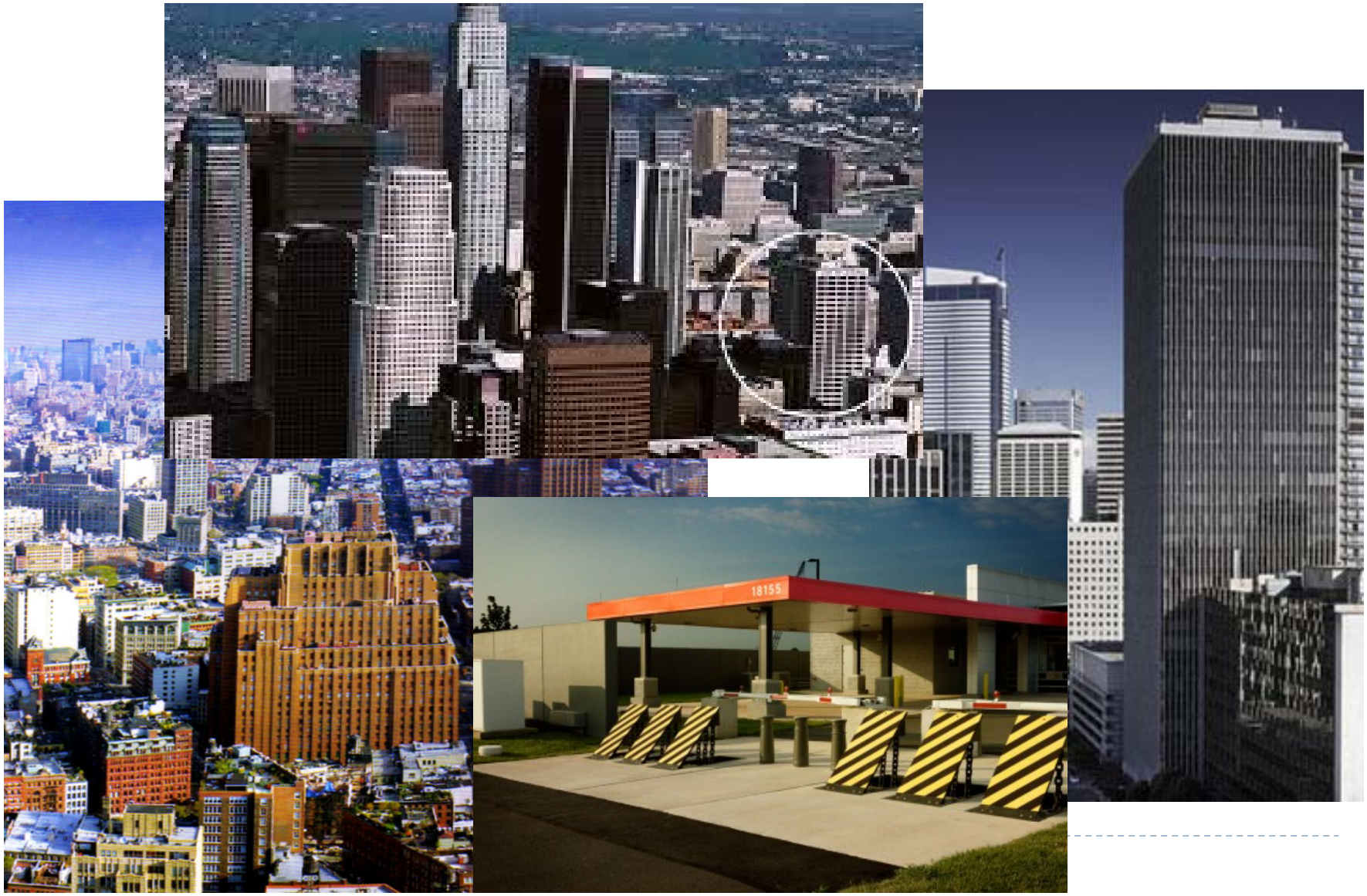
DuPont Fabros Technology

TELEHOUSE
HOME TO YOUR GLOBAL ICT SOLUTIONS



CORESITE

… to carrier hotels



About colos …

- Colos provide space, power, cooling, and physical security for third-party networking equipment and facilitate the interconnection of those third-party networks
- About 1-2K colocation/data center/interconnection facilities in the US
- An informed estimate: Some 10-20% of all US colos house some 80-90% of all networking (routing) equipment

About carrier hotels …

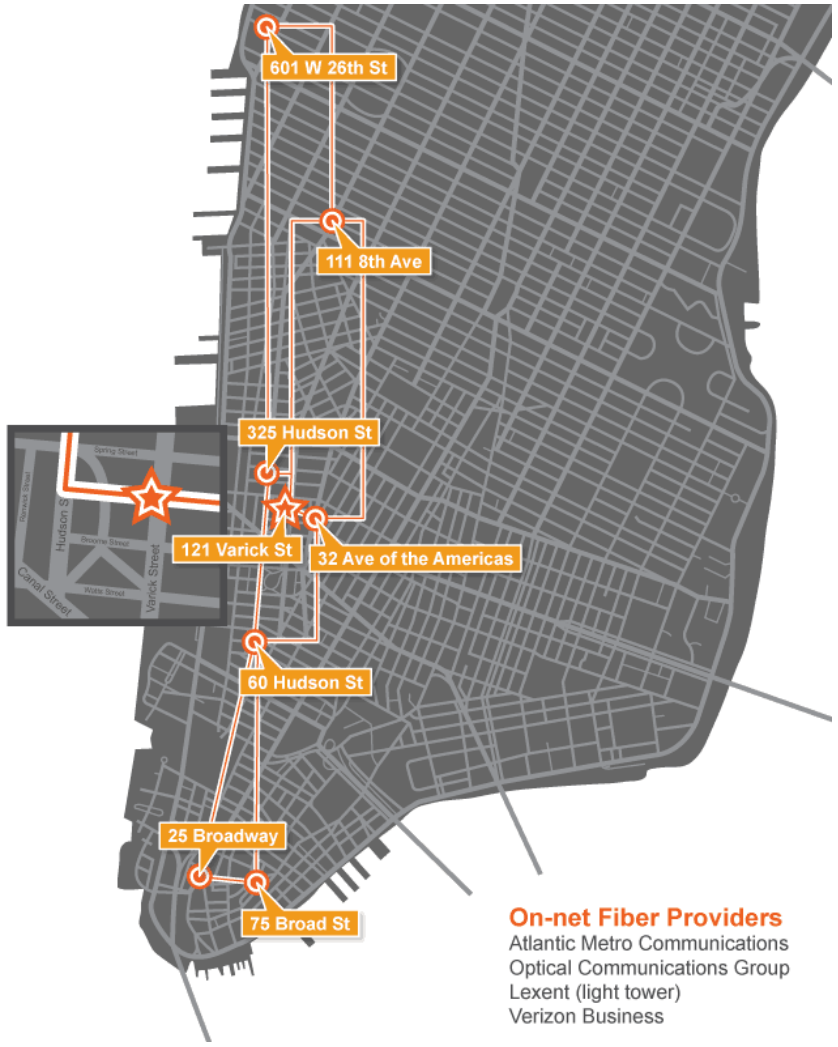
- Many of 1-2K colo facilities in the US are located in one and the same physical building (carrier hotel) in a city
- There are a few hundreds of carrier hotels across the US where most of the routers are located
- These buildings have publicly-known street addresses

Two well-known carrier hotels

- 60 Hudson Street, NYC
  - Built in the late 1920s; Western Union Building
  - Tenants include Telx, Equinix, DataDryd, zColo
- One Wilshire, LA
  - Built in 1966 as an office building (law firms)
  - Became a carrier hotel in the 1990s, mainly due to close proximity to AT&T's main switching center on Grand Avenue and Olive Street
  - Was bought in 2013 for about $500M ($660 per square foot)
  - Tenants include Coresite, zColo, Any2 California (IXP)
  - International cable landing point, 40+ Asia/Pacific carriers/ISPs

A look at the NYC carrier hotels …

- NY has about 100 colos
- NYC has some 50
- They are located in a few carrier hotels:
  - 601 W 26th St
  - 111 8th Ave
  - 325 Hudson St
  - 121 Varick St
  - 32 Ave of the Americas
  - 60 Hudson St
  - 25 Broadway
  - 75 Broad St



http://www.datacenternyc.com/images/mapNY_Fiber.png

The **physical** Internet – a roadmap

- A short-term goal – a coarse-grained view …
  - Map the largest 100-200 carrier hotels/colos/datacenters
  - Put on map of North American Fiber-Optic Long-Haul routes
  - Augment with map of international undersea cables

US Fiber-optic long-haul routes

http://www.metronetzing.org/images/dynamic/image/national.jpg?1288882944959

Global may of undersea cables

http://nicolasrapp.com/wp-content/uploads/2012/07/world_map_05_DARK-520x511.jpg

The “real” physical Internet – a roadmap

- A short-term goal – a coarse-grained view …
  - Map the largest 100-200 carrier hotels/colos/datacenters
  - Put on map of North American Fiber-Optic Long-Haul routes
  - Augment with map of international undersea cables
- A longer-term goal – a finer-grained view …
  - Map tenants in carrier hotels (PoP, router, interface IP address)
  - Map intra- and inter-colo network connections
- An end goal – add “bells and whistles” …
  - Eyeballs (end users), server infrastructure (datacenters), …

Some "fun" activities …

- traceroute experiments (I)
  - Run traceroute from a machine you can access …
  - … to a target in a different continent …
- traceroute experiments (II)
  - Run traceroute from a machine you can access …
  - … to a target in different continent and back (i.e., need access to target machine)
- traceroute experiments (III)
  - In case (I) or (II), how would you go about determining which undersea cable was carrying your probe packets?

Internet Research with "Big (Internet) Data"

Part II (of II)

Walter Willinger
NIKSUN, Inc.

wwillinger@niksun.com

Collaborators

- TU Berlin
  - Anja Feldmann, George Smaragdakis, Philipp Richter
- Akamai/Duke University
  - Nikos Chatzis, Jan Boettger
  - Bruce Maggs, Bala Chandrasekaran
- Northwestern University
  - Fabian Bustamante, Mario Sanchez
- University of Oregon (joint NSF grant, 2013/15)
  - Reza Rejaie, Reza Motamedi
- AT&T Labs-Research
  - Balachander Krishnamurthy, Jeff Erman
- University of Adelaide (joint ARC grant 2011/14)
  - Matt Roughan
- USC/ISI
  - John Heidemann, Xue Cai

Focus on two connectivity structures

- The Internet as a physical construct
  - The Internet as a physical infrastructure
  - Infrastructure = routers/switches and links/cables
  - Router-level topology of the Internet
- The Internet as a logical/virtual construct
  - The Internet as a "network of networks"
  - Network = Autonomous System/Domain (AS)
  - AS-level topology of the Internet

The Internet – a network of networks

- The AS-level Internet
  - Nodes = all 40K publicly routed Autonomous Systems (ASes)
  - Edges = the set of all transit and peering relationships

- A logical/virtual construct
  - AS-link: the two ASes exchange reachability information
  - Reachability: "active" BGP session(s) between border routers
  - AS-link is defined via a protocol: Border Gateway Protocol (BGP)
  - AS-link have attributes (type of AS relationship)
    - Internet transit ("customer-provider" relationship)
    - Internet peering ("public/private" peering relationship)

The AS-level Internet (since ~1995)

- Challenges (due to decommissioning of NSFNET)
  - No one entity has a complete view of the network
  - Networks come in many shapes and forms …
  - What geography for networks?

- Popular approach to visualizing the AS-level Internet
  - Step 1: Use BGP measurements (routing tables, updates)
  - Step 2: Obtain the data from multiple route monitors
  - Step 3: Combine BGP-derived AS-level paths to obtain the Internet's AS-level topology

Step 1-2: BGP measurements

- Commonly-used publicly available large BGP datasets
  - RouteViews project (Univ. of Oregon, since ~1997)
    - www.routeviews.org/
  - RIPE RIS project (RIPE NCC, Netherlands, since ~2000)
    - www.ripe.net/data-tools/stats/ris/routing-information-service
- Use BGP RIBs (routing information base)
  - RIBs contain routing information maintained by the router
  - Typical Routing table size: ~200-300K entries
  - Augment with constantly exchanged announcement/withdrawal messages

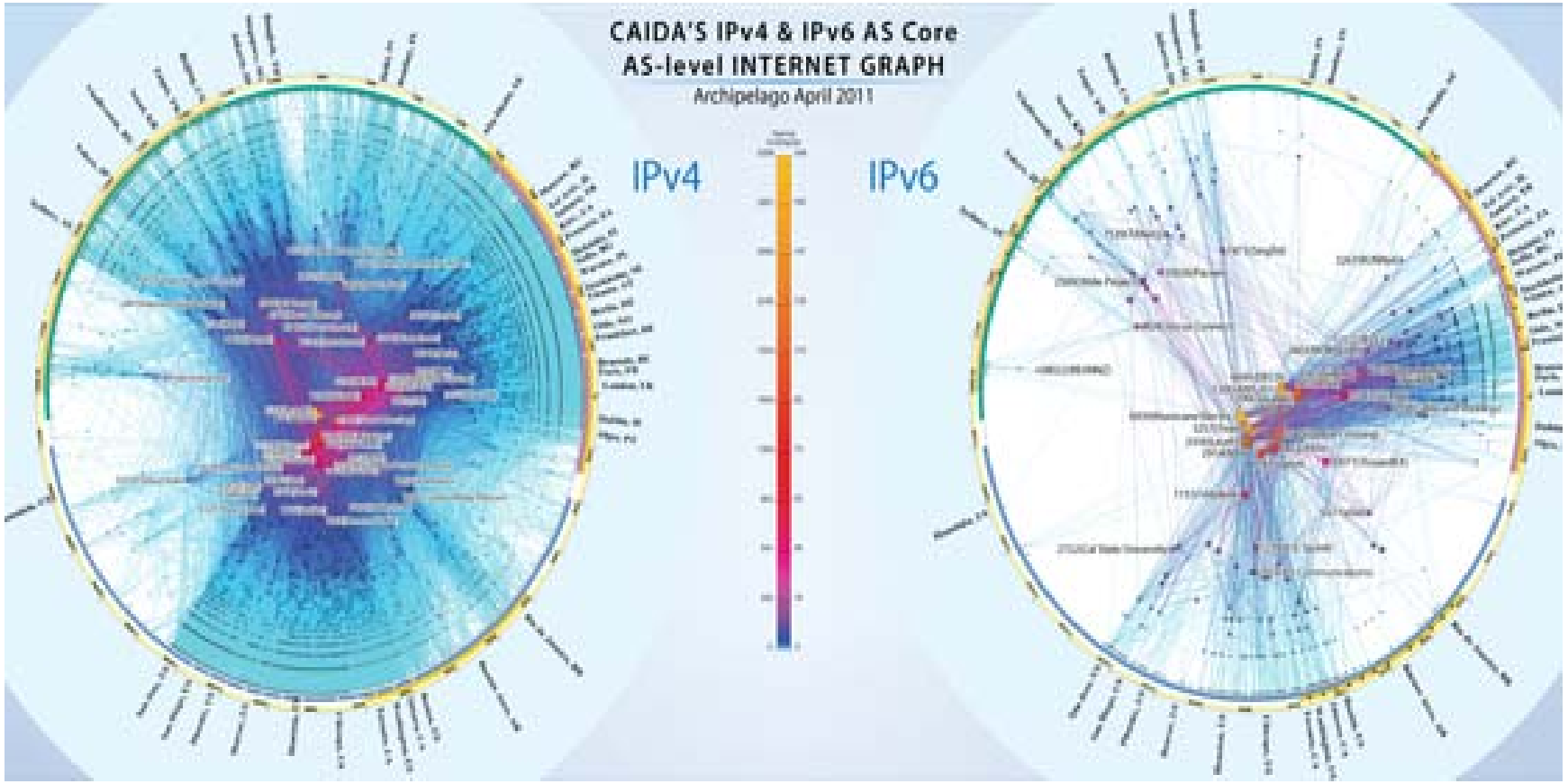
Typical BGP RIB table entry

```
PREFIX:          4.21.252.0/23
FROM:            194.85.4.55  AS3277
TIME:            2004-12-31  20:07:56
TYPE:            MSG_TABLE_DUMP/AFI_IP
VIEW:            0  SEQUENCE:  440
STATUS:          1
ORIGINATED:      Fri Dec 31  06:26:51  2004
AS_PATH:         3277  13062  20764  701
                 6389  8063  19198
NEXT_HOP:        194.85.4.55
COMMUNITIES:     3277:13062  3277:65301
                 3277:65307  20764:3000
                 20764:3011  20764:3020
                 20764:3022
```

Step 3: Combine AS-level paths

- Another example of "big (Internet) data"
  - Currently, there are some 14 RIS route collectors, and each one of them collects an entire BGP routing table every eight hours
  - 1 table (~ 200-400K RIB entries) is about 500MB (uncompressed)
  - Some 4 billion BGP-derived AS-level paths (~ 7 PB of data) per year
- Working assumption
  - With billions of BGP-derived AS paths, it is possible to recover the Internet's AS-level topology
  - The produced visualizations provide "insight" into the Internet's router-level topology

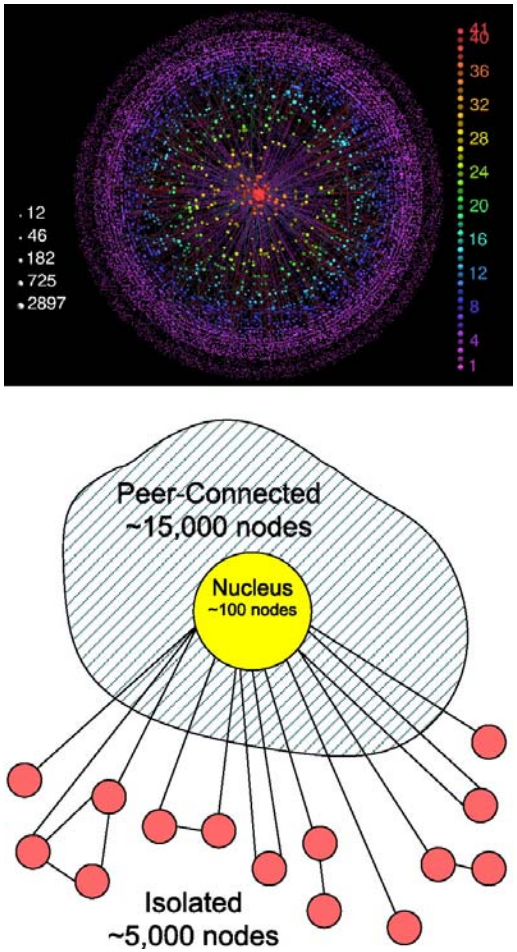
The “AS-level” Internet (caida.org)



The “AS-level” Internet (Peer1.com)

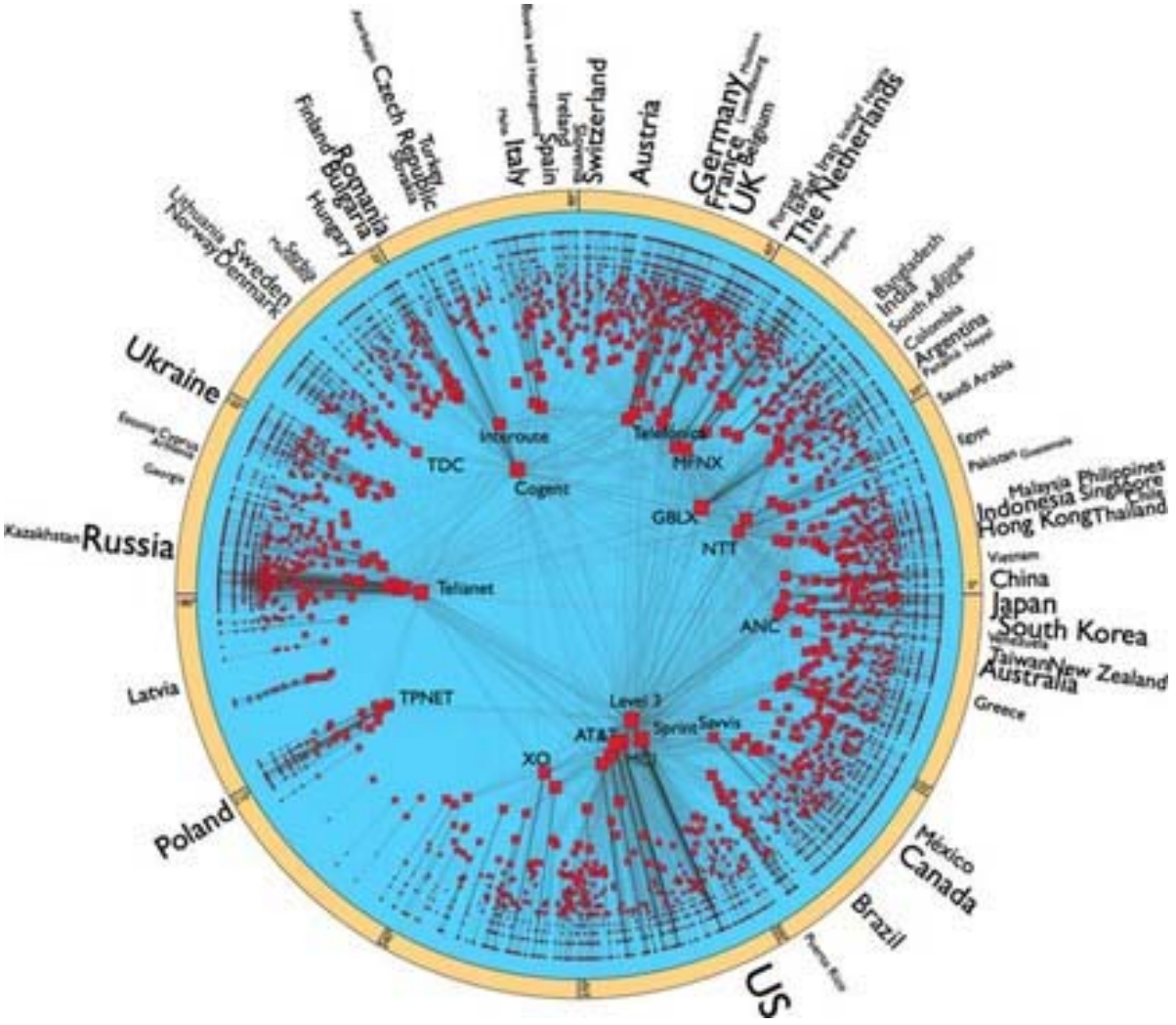
http://www.peer1.com/blog/2011/03/map-of-the-internet-2011

The “AS-level” Internet (PNAS 2007)



S. Carmi, S. Havlin, S. Kirkpatrick, Y.Shavitt, and E. Shir (PNAS 2007)

The “AS-level” Internet (2010)



- M. Boguna, F. Papadopoulos, and D. Krioukov (Nature Communications, 2010)

The AS-level Internet (current trends)

- "Insights" and "discoveries"
  - Random (e.g., scale-free) graphs appear to be suitable models
  - There are "obvious" high-degree nodes in the Internet
  - Removal of high-degree nodes is an "obvious" vulnerability
  - Discovery of the Internet's "Achilles' heel"

- Questions and issues
  - What is the quality of this "big (BGP) data" …?
  - How do the new "insights" compare to Internet reality …?
  - What exactly is "physical" about the resulting Internet maps …?

Getting to know your data …

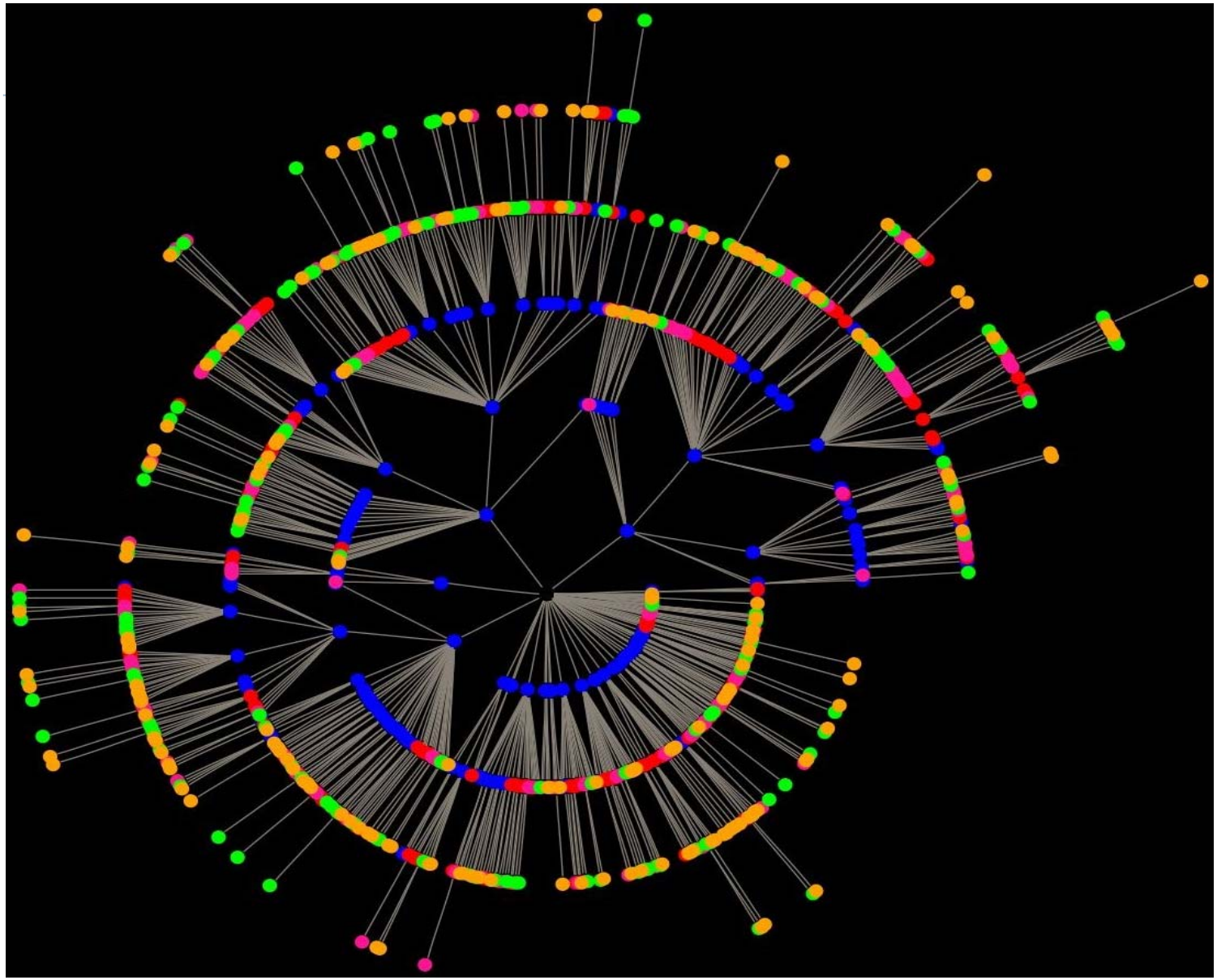
- The "Network Scientist's" perspective
  - Available data is taken at face value ("don't ask …")
  - No or only little domain knowledge is required
  - The outcome often leaves little room for further efforts
- The "Engineer's" perspective
  - Available data tends to be scrutinized (not enough, though)
  - Domain knowledge is "king" – details matter!
  - The results often give rise to new questions/problems

The Network Scientist's View

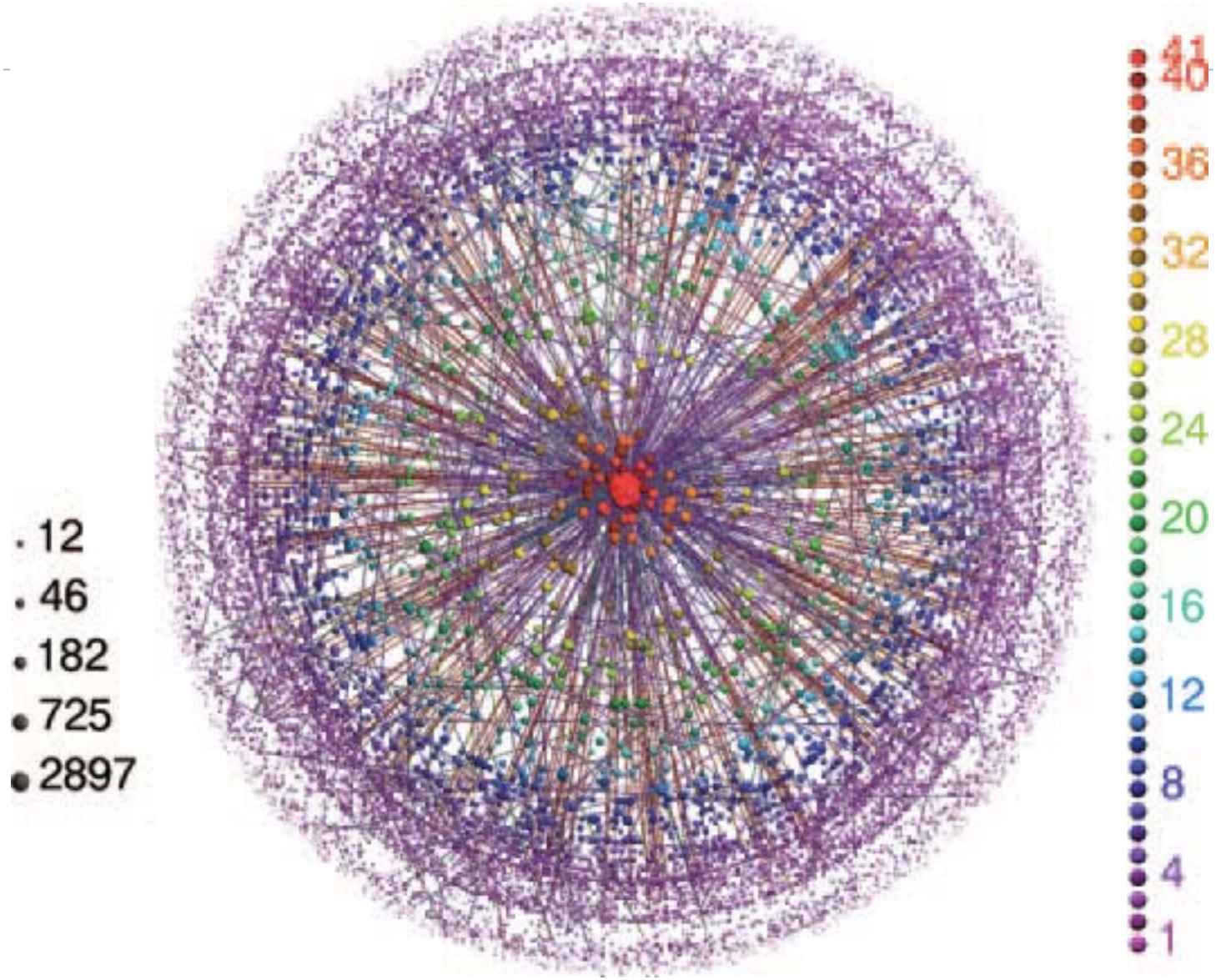
- Easy to download publicly available BGP datasets
- Take the data at "face value"
- Easy to reconstruct a graph (often already provided, courtesy of your friendly networking researchers)
- Resulting graph is taken to represent the Internet's AS-level connectivity ("ground truth")

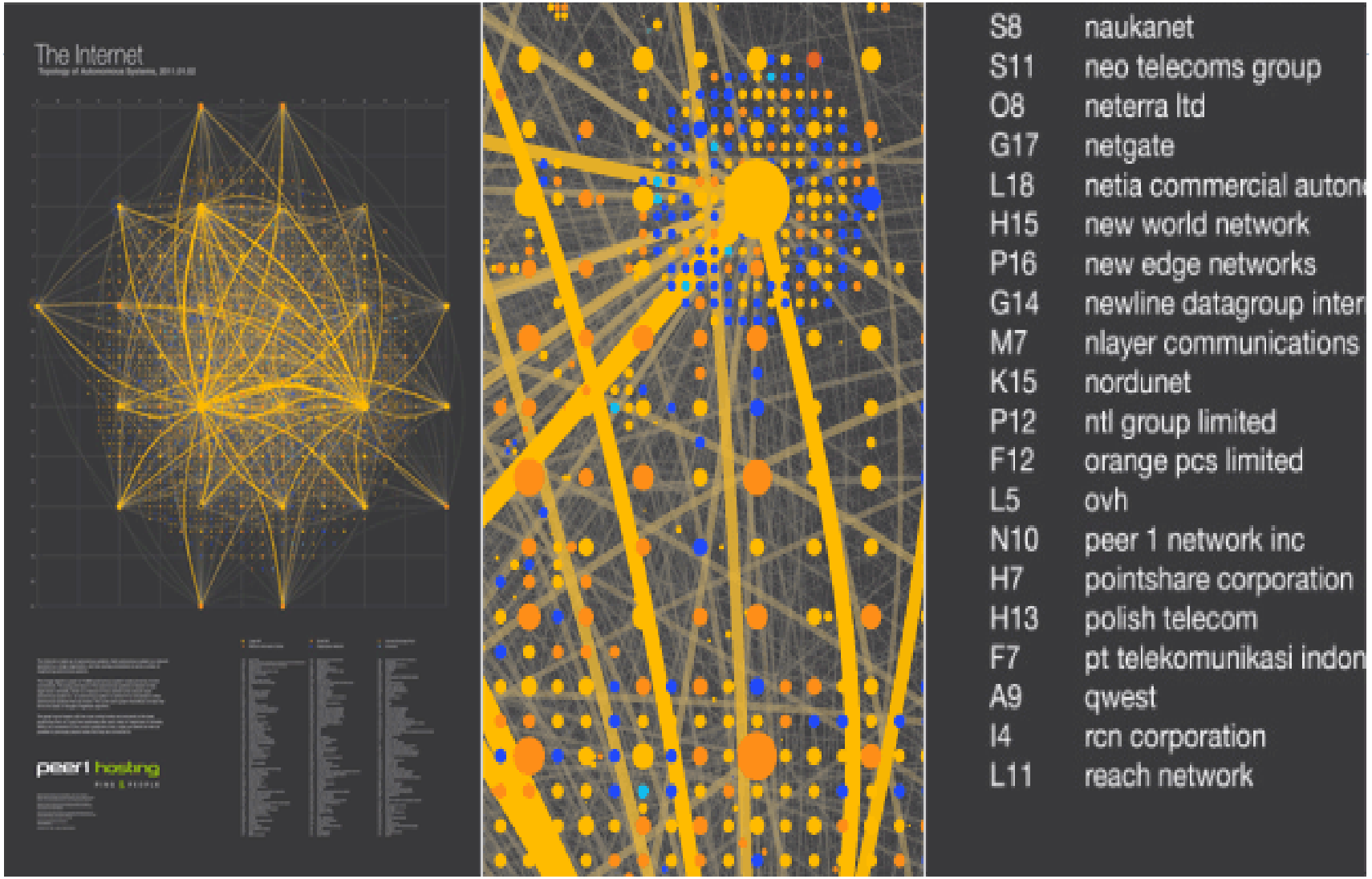
The Network Scientist's view

- Overall appeal for studying AS-level Internet
  - Reduces a "complex" system to a bunch of nodes & links
  - Results in moderate-sized graphs
  - The apparent connection to the Internet makes it an interesting "real-world" graph/network
- Exist "blue prints" for studying graphs
  - Characteristics (e.g., degree distribution, diameter, …)
  - Graph models of the Internet (e.g., scale-free type networks)
  - Model-based predictions
  - AS topology generation
  - Visualization – little else than "eye candy" …



R. D'Souza, C. Borgs, J. Chayes, N. Berger, and R. Kleinberg (PNAS 2007)


12
46
182
725
2897
41
40
36
32
28
24
20
16
12
8
4
1



http://www.peer1.com/blog/2011/03/map-of-the-internet-2011/

The Engineer's view

- The inter-domain routing system
- The inter-domain routing protocol BGP

- BGP-based measurements
- BGP data collection projects for the public good

Re: Inter-Domain Topology

- Inter-domain routing system
  - Foundation for Internet wide-area communication
- Characteristics impacting the performance of this system
  - Inter-domain topology (also called AS-graph)
    - Nodes are ASes
    - Links are AS relationships
    - Links signify route exchange between corresponding ASes, but not necessarily IP traffic exchange!
  - Route stability
    - Transient changes due to router or link failures
    - Router misconfigurations

Re: Measuring AS-level Connectivity

- Basic problem
  - Individual ASes know their (local) AS-level connections
  - AS-specific connectivity data is not publicly available
  - AS-level connectivity cannot be measured directly
- Main Reasons
  - AS-level data are considered proprietary
  - Fear of loosing competitive advantage
  - No central agency exists that collects this data
  - No tool exists to measure AS connectivity directly

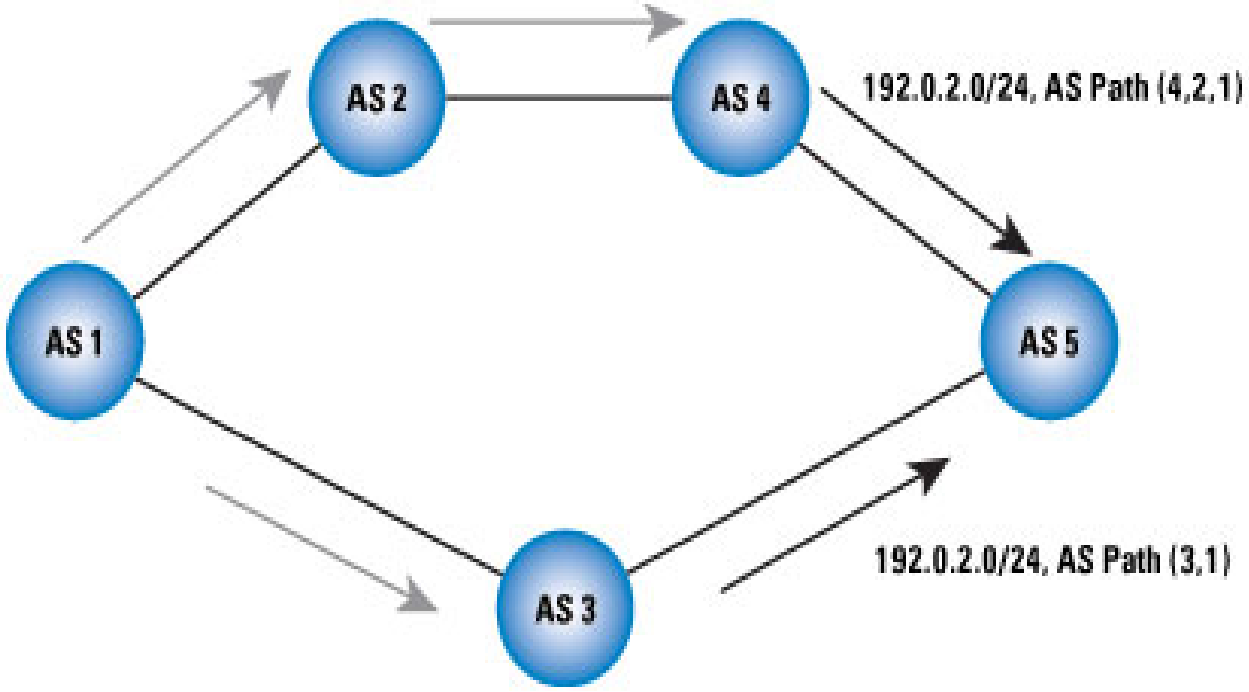
Re: Measuring AS-level Connectivity (cont.)

- Generic approach to overcome basic problem
  - Identify and collect appropriate "surrogate" data
  - Surrogate data should be publicly available/obtainable
  - May require substantial efforts to collect surrogate data
  - What does the surrogate data really say about AS-level connectivity?
- Practical solution
  - Rely on BGP, the de facto inter-domain routing protocol
  - Use BGP RIBs (routing information base)
  - RIBs contain routing information maintained by the router

Re: Inter-Domain Routing Protocol BGP4

- De facto standard inter-domain routing protocol
  - RFC 1771 (1995), RFC 4271 (2006)
- Enables ASes to implement/realize their routing policies
  - An AS may originate one or more routes
  - Routes advertise reachability to IP address prefixes within an AS
  - An AS realises its policies by independently selecting and selectively propagating routes obtained from neighboring ASes
  - Associated with each route is the list of ASes traversed by the route – the route's AS_PATH
- Scalable, expressive, flexible information-hiding protocol
  - Exchange of routing information w/o revealing AS-internals
  - Support for the complex and evolving business policies ASes have with each other

Example of AS Path Generation in BGP



Re: Available BGP measurements

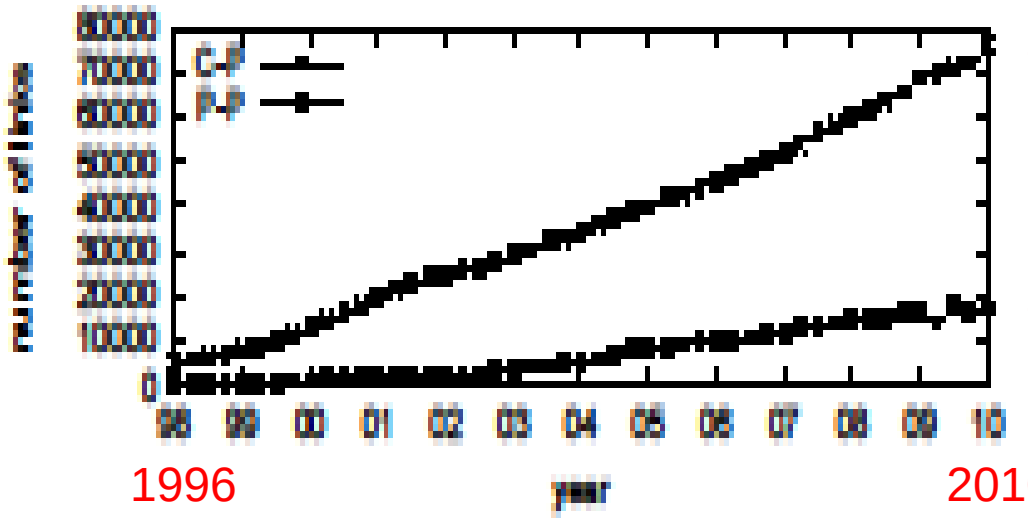
- Use BGP RIBs (routing information base) and updates
  - RIBs contain routing information maintained by the router
  - Typical Routing table size: ~200-300K entries
  - Focus has been on AS_PATH attribute
- Typical BGP RIB table entry

```
PREFIX:         4.21.252.0/23
FROM:           194.85.4.55  AS3277
TIME:           2004-12-31  20:07:56
TYPE:           MSG_TABLE_DUMP/AFI_IP
VIEW:           0   SEQUENCE:  440
STATUS:         1
ORIGINATED:     Fri Dec 31  06:26:51  2004
AS_PATH:        3277  13062  20764  701
                6389  8063  19198
NEXT_HOP:       194.85.4.55
COMMUNITIES:    3277:13062  3277:65301
                3277:65307  20764:3000
                20764:3011  20764:3020
                20764:3022
```

Who is collecting BGP measurements?

- Daily BGP table dumps and updates are collected from multiple monitors that are connected to numerous routers across the Internet
- RouteViews project (University of Oregon)
  - Started ~1997
  - Initially connected to large providers, recently also to IXPs
  - http://www.routeviews.org/
- RIPE RIS project (RIPE NCC, Netherlands)
  - Started data collection around 2000
  - Similar approach as RouteViews
  - http://www.ripe.net/data-tools/stats/ris/routing-information-service

Results from BGP data (1996-2010)



~80,000 C-P links

~20,000 P-P links

- Some 30,000 ASes
- Some 80,000 links of the customer-provider type (80% of all links)
- Some 20,000 links of the peer-peer type (20% of all links)

A. Dhamdhere and C. Dovrolis (2011)

Re: Other data for the AS-level Internet

- Data plane measurements (e.g., traceroute)
  - Archipelago (Ark, previously Skitter), CAIDA
  - Dimes (EU project)
  - Many more …
- Unsolved problem: Mapping traceroutes to AS-routes
  - Problem #1: Mapping interface IP addresses to routers (IP alias resolution problem)
  - Problem #2: Mapping routers to ASes
- Bottom line
  - Without novel solutions to problems #1 and #2, current traceroute-based measurements are of very questionable quality for accurately inferring AS-level connectivity

Re: Other data for the AS-level Internet

- Other available sources
  - Public databases (WHOIS)
  - Internet Routing Registry IRR)
  - Packet Clearing House (PCH), PeeringDB, Euro-IX
- Main problems
  - Voluntary efforts to populate the databases
  - Inaccurate, stale, incomplete information
- Bottom line
  - These databases contain valuable information …
  - These databases are of insufficient quality to even approximately infer AS-level connectivity

Re: Engineer's view - some "details"

- Key observation
  - BGP is not a mechanism by which ASes distribute connectivity information
  - BGP is a protocol by which ASes distribute the reachability of their networks via a set of routing paths that have been chosen by other ASes in accordance with their policies.
- Main challenge
  - BGP measurements are an example of "surrogate" data
  - Using this "surrogate" data to obtain accurate AS-level connectivity information is notoriously hard
  - Examining the hygiene of BGP measurements requires significant commitment and domain knowledge

Re: Engineer's view – some details (cont.)

- Basic problem #1: Incompleteness
  - Many peering links/relationships are not visible from the current set of BGP monitors
  - A well-known problem of vantage point locations
- Basic problem #2: Ambiguity
  - Need heuristics to infer "meaning" of AS links: customer-provider, peer-to-peer, sibling, and a few others
  - Existing heuristics are known to be inaccurate
  - Renewed recent efforts to develop better heuristics …

Re: Engineer's view – some details (cont.)

- The dilemma with current BGP measurements
  - Parts of the available data seem accurate and solid (i.e., customer-provider links, nodes)
  - Parts of the available data are highly problematic and incomplete (i.e., peer-to-peer links)
- Bottom line
  - (Current) BGP-based measurements are of questionable quality for accurately inferring AS-level connectivity
  - It is expected that future BGP-based measurements will be more useful for the purpose of AS-level inference
  - Very difficult to get to the "ground truth"

Re: Engineer's view – some details (cont.)

- RouteViews/RIPE RIS data were never meant to be used to infer the Internet's AS-level connectivty
  - Missing data problem (links)
  - Inaccuracies (AS relationship inference)
  - Ambiguities (due to transients and dynamics)
- BUT: value/benefit of the data for operators is huge!
- Use of BGP-based measurements by the research community for mapping the Internet's AS-level connectivity
  - Engineering hack – BGP is an information-hiding and not an information-revealing protocol
  - An example of "What we can measure is typically not what we want to measure!"

On RouteViews/RIPE-provided datasets

- From the RouteViews/RIPE RIS websites
  - "The RouteViews project was originally conceived as a tool for Internet operators to (i) obtain real-time information about the global routing system from the perspectives of several different backbones and locations around the Internet, and (ii) determine how the global routing system viewed their prefixes and/or AS space."
  - "The goal of the Routing Information Service (RIS) is to collect routing information between ASes and their development over time from a number of vantage points in the Internet. One important application for this data will be debugging. For example, if a user complains that a certain site could not be reached earlier, the RIS will provide the necessary information to discover what caused the problem."
- No mentioning that the obtained data are applicable to inferring the Internet's AS graph, and for good reasons …!!
  - Does provides some info about the AS-level Internet
  - Does not provide the info needed to infer AS-level connectivity

But all this was well-known …!

- R. Govindan and A. Reddy, 1997. An analysis of Internet inter-domain topology and route stability. IEEE INFOCOM.

- The purpose for performing their study is explicitly stated
  - *"To understand the impact of the routing system on wide-area communication, we focus on two characteristics of the routing system: the inter-domain topology and route stability."*
  - *"… we obtain approximate characterizations, called snapshots of the inter-domain topology from three different segments of our [BGP] traces."*

Re: Govindan & Reddy 1997 paper

- The main problems with the BGP measurements are explicitly mentioned
  - *"However, there is still a likelihood of "missing" some of the inter-domain links, and a smaller likelihood of "missing" some domains as well. "*
  - *"In general, we expect that this technique gives a fairly good picture of the topology closer to the trace collection location (i.e., in the North America portion of the Internet). The "fuzziness" of our snapshots is likely to increase with the increasing distance from the trace collection locations."*

- The Govindan & Reddy 1997 paper is an early textbook example for what information a measurement paper should provide.

Re: Govindan & Reddy 1997 paper (cont.)

- Albert et al. (2000) point directly to Faloutsos et al. (1999)
- Faloutsos et al. (1999) cite Govindan&Reddy (1997) but ignore the caveats mentioned in that paper and mis-represent the reported efforts
- Almost all subsequent papers that deal with the AS-level Internet cite Faloutsos et al. (1999)
- An example of the influence that secondary citations can and do have …
- The Govindan & Reddy 1997 paper is now hardly cited and largely forgotten!

Discussion

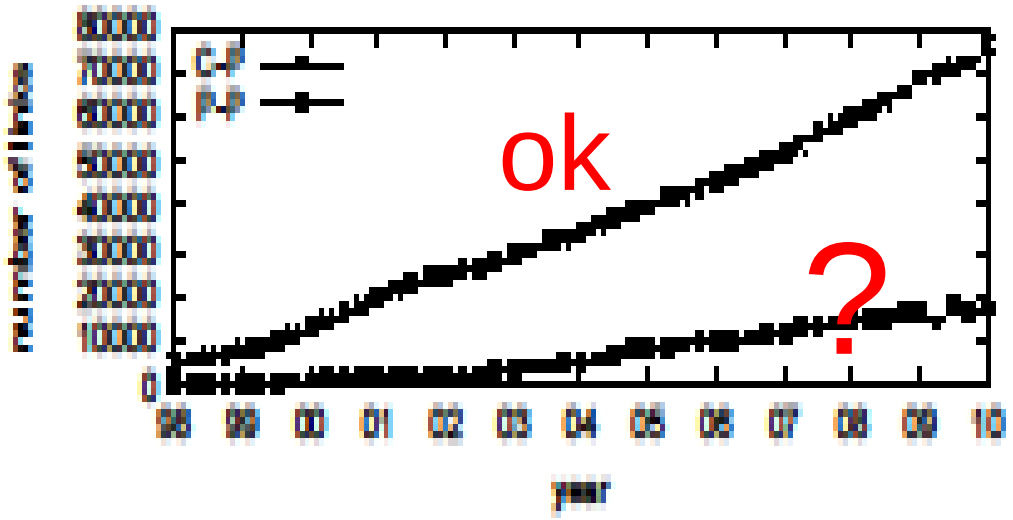
- **Details do matter!**
- **Implications for analyzing BGP data?**
- **Every networking student knows the problems with BGP data, so why is this domain knowledge not used?**
- **Why has the 1997 Govindan & Reddy paper never been referenced in subsequent Internet topology papers?**

Re: Missing link problem in BGP data

- The dilemma with the available BGP measurements
  - Some of the data is accurate and solid
  - Some of the data is highly problematic/incomplete/inaccurate
  - Examining the hygiene of these BGP measurements requires significant commitment and domain knowledge

- 2008 (with R. Oliveira, D. Pei, B. Zhang, and L. Zhang)
  - Good: Using the data over time provides high-quality info about AS links representing customer-provider relationships
  - Bad: Datasets provide low-quality info about AS links representing peer-to-peer relationships (missing link problem)
  - How bad is "bad"?

Re: Missing link problem in BGP data



- Some 30,000 ASes ✓
- Some 80,000 links of the customer-provider type (80% of all links) ✓
- Some 20,000 links of the peer-peer type (20% of all links) ?

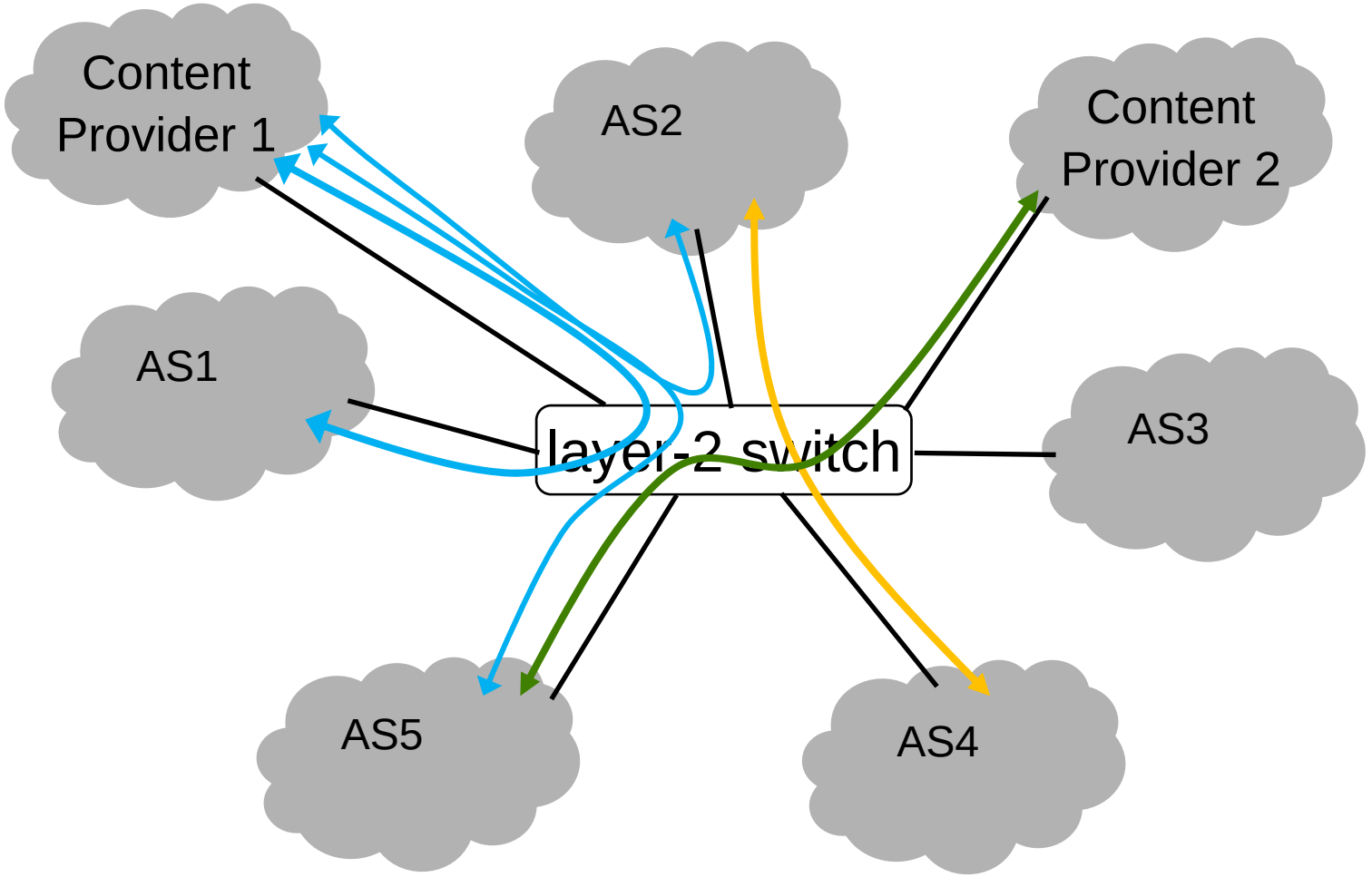
?

A. Dhamdhere and C. Dovrolis (2011)

IXPs and the missing link problem

- An IXP is a physical facility with a switching infrastructure for the primary purpose to enable networks to interconnect and exchange traffic directly (and essentially for free) rather than through one or more 3rd parties (and at a cost).

Internet eXchange Points (IXPs)



IXPs & the missing link problem (~2010)

- An IXP is a physical facility with a switching infrastructure for the primary purpose to enable networks to interconnect and exchange traffic directly (and essentially for free) rather than through one or more 3rd parties (and at a cost).

- Example: European IXP market
  - Operational IXPs: from a few in the mid-1990 to 127 in 2010
  - ~40 participants per IXP (a few 100 for the large ones)
- Simple math to estimate existing number of peering links
  - About 250 * (40*39/2) * .33 ~ 65,000 peer-to-peer links
  - RouteViews/RIPE RIS data only show about 20,000 peerings

- 15 years of AS topology research with almost 50,000 (critical) links missing???

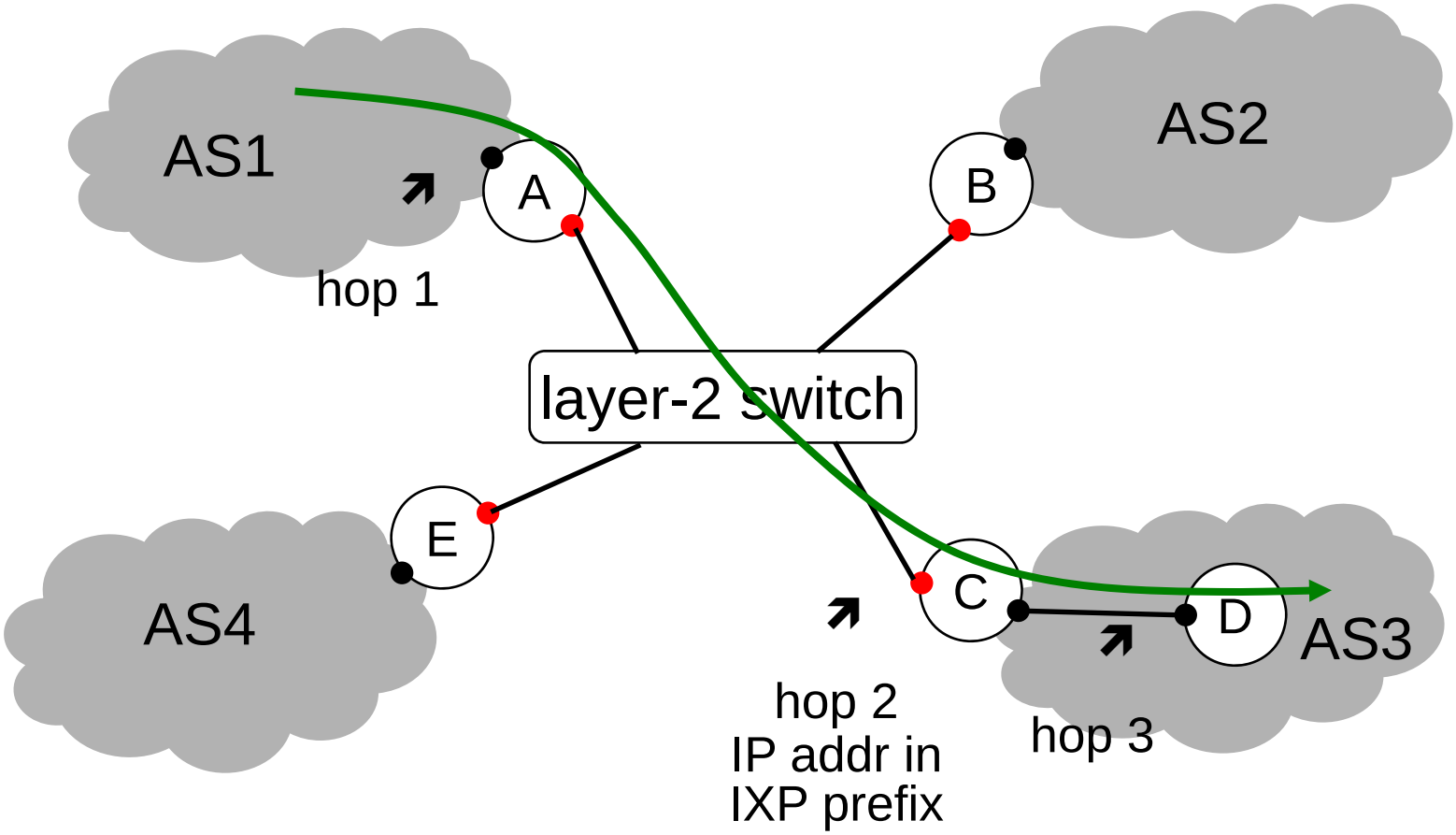
Going after the missing links at IXPs

- IXPs are promising places to look for missing AS links
  - 2002 (with H. Chang, R. Govindan, S. Jamin, and S. Shenker)
- Methodology for identifying IXPs in traceroute probes
  - 2004 (K. Xu, Z. Duan, Z.-L. Zhang, and J. Chandrashekar)
- Initial attempt at discovering new peerings at IXPs
  - 2005 (H. Chang)
- Another attempt at discovering new peering links at IXPs from general-purpose traceroute measurements
  - 2007 (Y. He, G. Siganos, M. Faloutsos, and S.V. Krishnamurthy)
- Explanation for why the Internet's IXP substrate holds the secret concerning the missing links
  - 2008 (with R. Oliveira, D. Pei, B. Zhang, and L. Zhang)

Going after the missing links at IXPs (I)

- Ad-hoc or general-purpose traceroute data
  - Method-of-choice until 2008
  - Detected a few thousands of new links

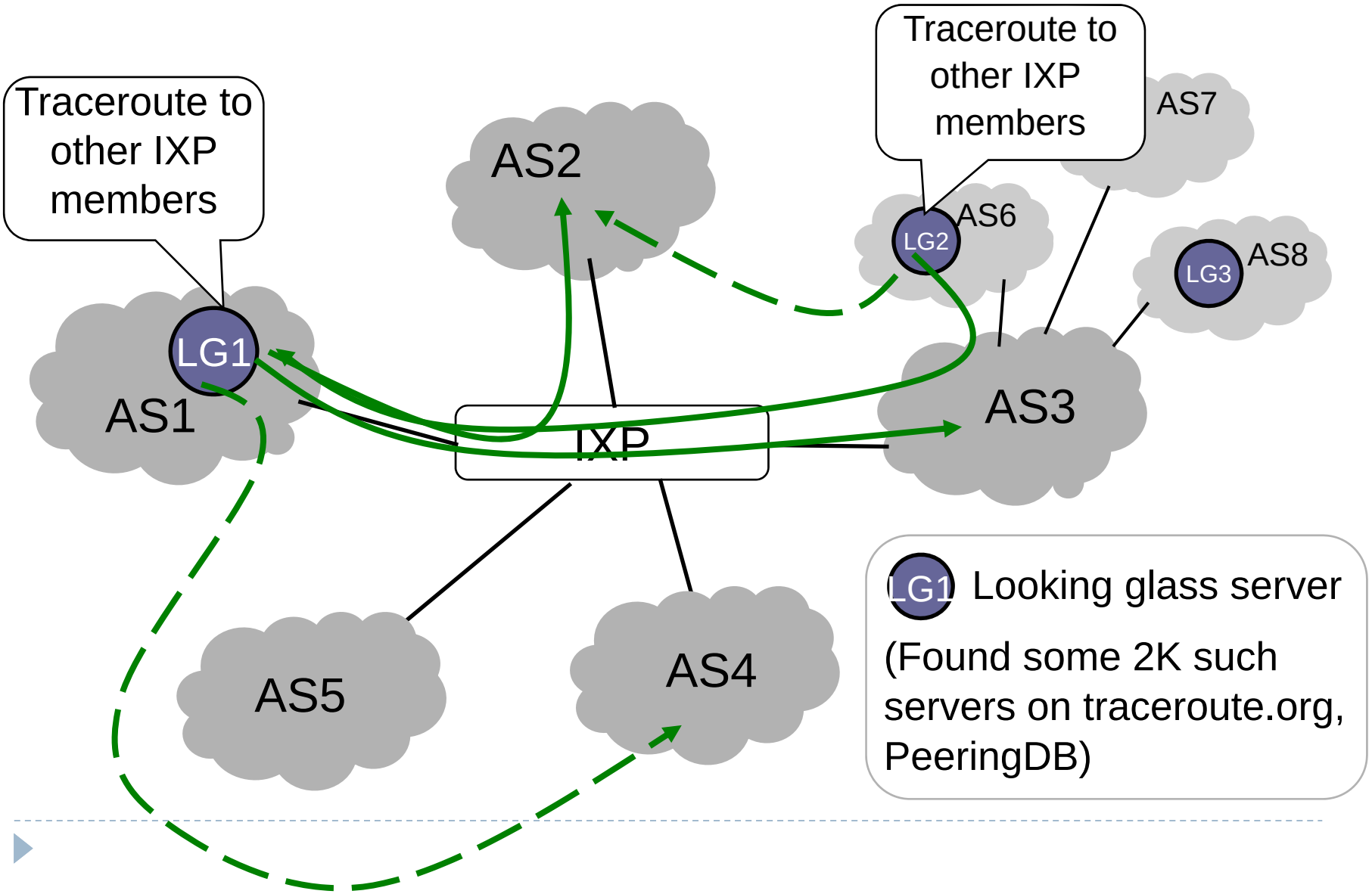
Identifying IXPs in traceroute data



Going after the missing links at IXPs (II)

- Ad-hoc or general-purpose traceroute data
  - Method-of-choice until 2008
  - Detected a few thousands of new links
- Special-purpose traceroute campaigns using LGs
  - 2009 (with B. Augustin and B. Krishnamurthy)
  - Relied on some 1-2K Looking Glasses
  - Detected ~20,000 new P-P links that cannot be seen in the RouteViews/RIPE RIS-provided datasets

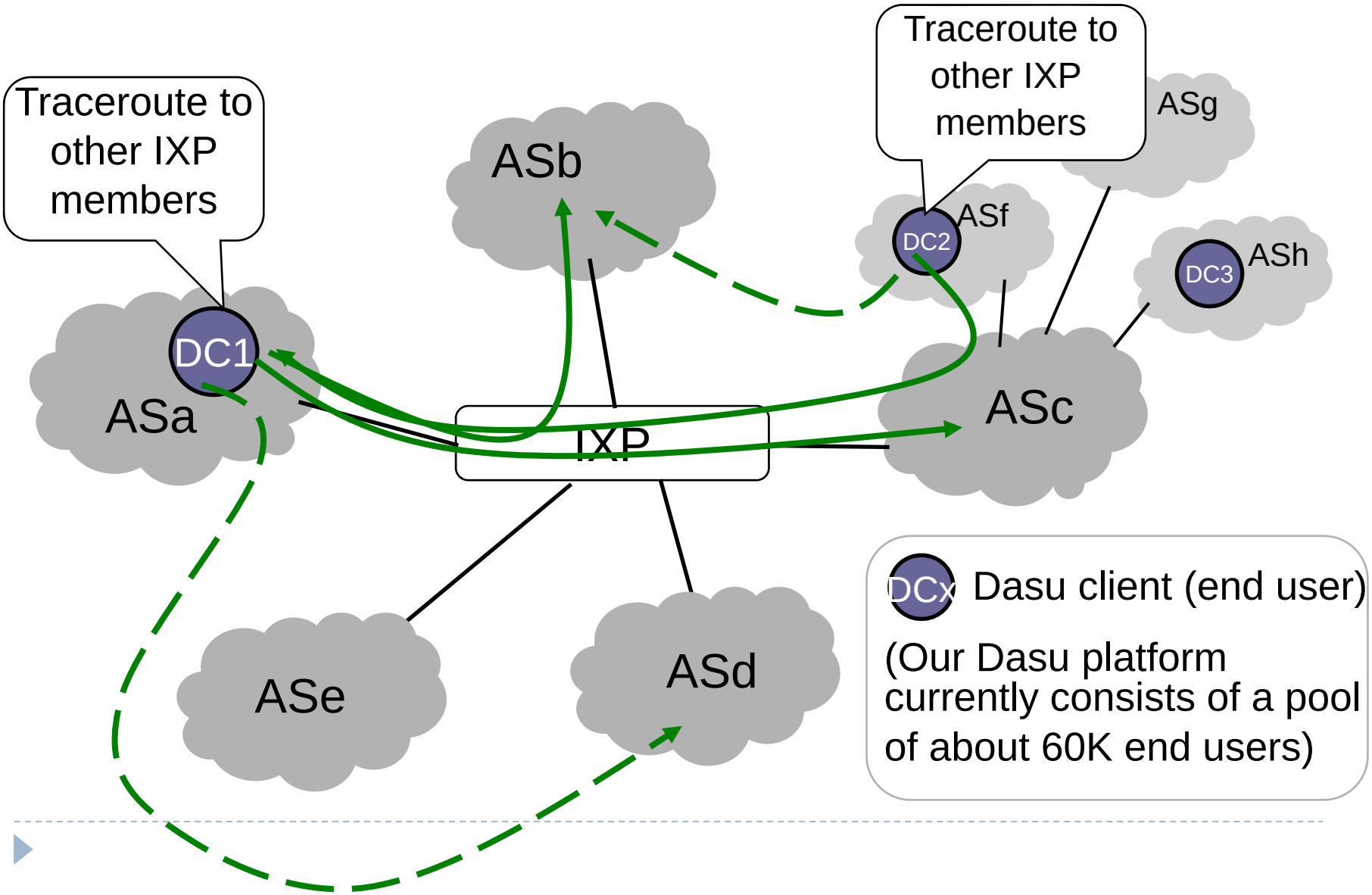
Use of LGs for targeted traceroutes



Going after the missing links at IXPs (III)

- Ad-hoc or general-purpose traceroute data
  - Method-of-choice until 2008
  - Detected a few thousands of new links
- Special-purpose traceroute campaigns using LGs
  - 2009 (with B. Augustin and B. Krishnamurthy)
  - Relied on (mis-used?) some 1-2K Looking Glasses
  - Detected ~20,000 new peering links that cannot be seen in the RouteViews/RIPE RIS-provided datasets
- Special-purpose traceroute campaigns using Dasu
  - 2010 – present (with M. Sanchez, F. Bustamante, and B. Krishnamurthy)
  - Rely on some 30K Dasu clients (i.e., end users)
  - Detected another some 20,000 new P-P links that cannot be seen in the RouteViews/RIPE RIS-provided datasets
  - Importantly: Essentially disjoint from LG-discovered P-P links!

Use of Dasu for targeted traceroutes



Getting closer to “ground truth”? (2012)

Methodology	Number of P-P links in the entire Internet
2010 BGP data (RouteViews/RIPE-RIS)	> 20,000

Getting closer to “ground truth”?

Methodology	Number of P-P links in the entire Internet
2010 BGP data (RouteViews/RIPE-RIS)	> 20,000
2010 Targeted traceroute probes (LGs)	> 40,000
2011 Targeted traceroute probes (Dasu)	> 60,000

Getting closer to "ground truth"?

Methodology	Number of P-P links in the entire Internet
2010 BGP data (RouteViews/RIPE-RIS)	> 20,000
2010 Targeted traceroute probes (LGs)	> 40,000
2011 Targeted traceroute probes (Dasu)	> 60,000
2012 (Proprietary) data from a large IXP	> 200,000

What happened?

- We got lucky …
  - Anja Feldmann's group at T-Labs/TU Berlin obtained high-quality traffic data from on of the largest IXPs in Europe
  - *B. Ager, N. Chatzis, A. Feldmann, N. Sarrar, S. Uhlig, W. Willinger Anatomy of a Large European IXP, ACM Sigcomm 2012*

- A brief summary of our main IXP-specific findings
  - This IXP has some 400 active member ASes (public info)
  - This IXP handles some 10-20 PB traffic on a daily basis (public info) - as much as some of the largest tier-1 ISPs
  - At this IXP alone, there are more than 50,000 links of the peer-peer type, most of which are invisible to the commonly-used BGP and traceroute measurements but are actively used.

Re: Missing link problem (2015)

- Derivation of new lower bound
  - Conservative extrapolation to the European Internet scene
    - 4 such large IXPs in Europe (assume 50% peering): ~160,000 P-P links
    - Remaining 150 or so IXPs in Europe : ~40,000 P-P links
  - Completely ignoring the 150 or so IXPs in the rest of the world

- (Conservative) lower bound on the number of P-P links
  - There are easily more than 200,000 P-P links in today's Internet (as compared to the currently assumed ~ 20,000)

- These numbers require a complete revamping of the mental picture our community has about the AS-level Internet.

What Now?

- 15 years of studies of the AS-level Internet with some 50% of the links missing …
  - Will we learn from this?
- The boring but highly predictable next steps/papers
  - Augment previous AS-graphs with these missing links
  - Repeat the same king of graph-type analysis with this "more complete" AS graph
- The exiting but very difficult next steps/papers
  - Scientific exploration of the AS-level Internet (not a graph!)
  - What network economics to study an economic construct?

Challenge #1: Expect Change!

- Meaning/definition of an AS
  - RFC 1930: A collection of connected IP routing prefixes under the control of one or more network operators that presents a common, clearly defined routing policy to the Internet
  - Reality: ASes are often not homogeneous and/or contiguous entities
  - Examples: multi-AS orgs; one and the same AS can announce different sets of prefixes at different exit points of its network (PoPs)
- Meaning/definition of an AS link
  - Case in point: IXP have no place in a traditional AS graph
  - Requires a single "edge" to connect multiple ASes – need hypergraph structure
- Measurement of AS connectivity
  - Observed new peering arrangements require finer-grained measurement capabilities
- Modeling AS-level connectivity of the Internet
  - NOT a graph!
  - Need models that reflect the importance of economic aspects of this construct

Challenge #2: Measurement is Hard!

- Detecting missing AS links is largely a visibility problem
  - Less about #traceroutes launched, and more about the locations from where they are launched
- Available platforms with VPs
  - PlanetLab infrastructure
    - few, fixed, but powerful nodes
    - limited visibility into the public Internet due to node locations
  - Looking Glass servers
    - a few thousand servers, with limited capabilities (e.g., traceroute, BGP summary)
    - typically found in (and supported by) networks of large NSP or of academic & research/education institutions
    - not intended to be used as "general-purpose" Internet measurement platform – operators are watching them!
  - Dasu platform
    - abundance of nodes/end users in the "interesting" parts of the growing Internet
    - leverages P2P client plug-in to launch active/passive measurement experiments
    - Example of a "good" botnet …

Challenge #3: It's (mostly) about Economics!

- AS-level connectivity of the Internet
  - Much more interesting than what a simple graph can capture
  - The IXP substrate is a very vibrant part of the AS-level Internet
    - IXPs actively vie for (paying) participants
    - IXPs constantly innovate, using latest technologies (e.g., SDN)
    - Economic incentives for IXP participants are often obvious
    - New players enter the picture (e.g., IXP resellers)
- Examples of innovation within the IXP substrate
  - Remote peering service (by IXPs, in combination with NSPs that enable this service)
  - Free use of route server (for multi-lateral peering)
  - Enabler of fine-grained peering relationships
    - Prefix-specific peering
    - Load- or time-of-day-specific peering

Challenge #4: Traffic is Key!

- Cannot understand/model the Internet's AS-level connectivity structure and its evolution without knowing anything about the traffic that is exchanged over this complex structure
- How to perform meaningful measurement experiments and/or inference to provide useful and high-quality traffic-related info?
- Some initial recent attempt
  - 2004 (A. Feldmann et al. : inter-domain Web traffic)
  - 2004 (S. Uhlig et al.: first study of inter-domain traffic traces)
  - 2006 (with H. Chang: on inter-domain connectivity and traffic)
  - 2006 (with H. Chang et al.: inter-AS traffic matrices)
  - 2009 (with Y. Zhang et al.: TMs & compressive sensing/matrix completion)
  - 2010 (V. Bharti et al.: inferring invisible traffic & matrix completion)

Challenge #5: What Internet Hierarchy?

- Our mental picture of "tiered Internet hierarchy" may have been consistent with reality 10-15 years ago, briefly after the decommissioning of the NSFNET
- However, for the last 5-10 years, this mental picture is no longer valid (except maybe for the Tier-1's), nor are the various suggested replacements (commonly referred to as "flattening of the Internet")
  - P. Gill et al., PAM 2008.
  - C. Labovitz et al., SIGCOMM 2010
  - A. Dhamdhere and C. Dovrolis, CoNEXT 2010.
  - A. Ager et al., SIGCOMM 2012

Challenge #6: AS Internet – Not a Graph!

- Reality is more like "everything goes"
  - Wide range of large-to-small content providers, hosting, CDNs
  - Wide range of global-to-local ISPs and NSPs
  - Wide range of IXPs with global/national/local participants
- Hierarchical and flat at the same time
  - Rich upstream (customer-provider) connectivity (e.g., for enterprise/business customers "valuable" traffic)
  - Rich peering (peer-to-peer) connectivity wherever it makes sense and is supported (e.g., for connecting content to eyeballs at IXPs where the demand justifies peering)
- Conventional wisdom vs. reality
  - Well-known "fact": Tier-1's don't show up at IXPs
  - Think again: Tier-1's do show up at IXPs, but in "disguise" (i.e., using different ASNs they own)
  - Need to know: How do ASNs map to organizations/corporations?

Challenge #7: **AS** meets **physical** Internet

- The AS-level Internet
  - IXPs are housed in one or more colo facilities
  - Colos/router hotels house the routers of one or more ASes
  - Inter-AS connectivity can manifest itself in many different physical connections (between distinct border routers)

DE-CIX: Colocation facilities in FRA

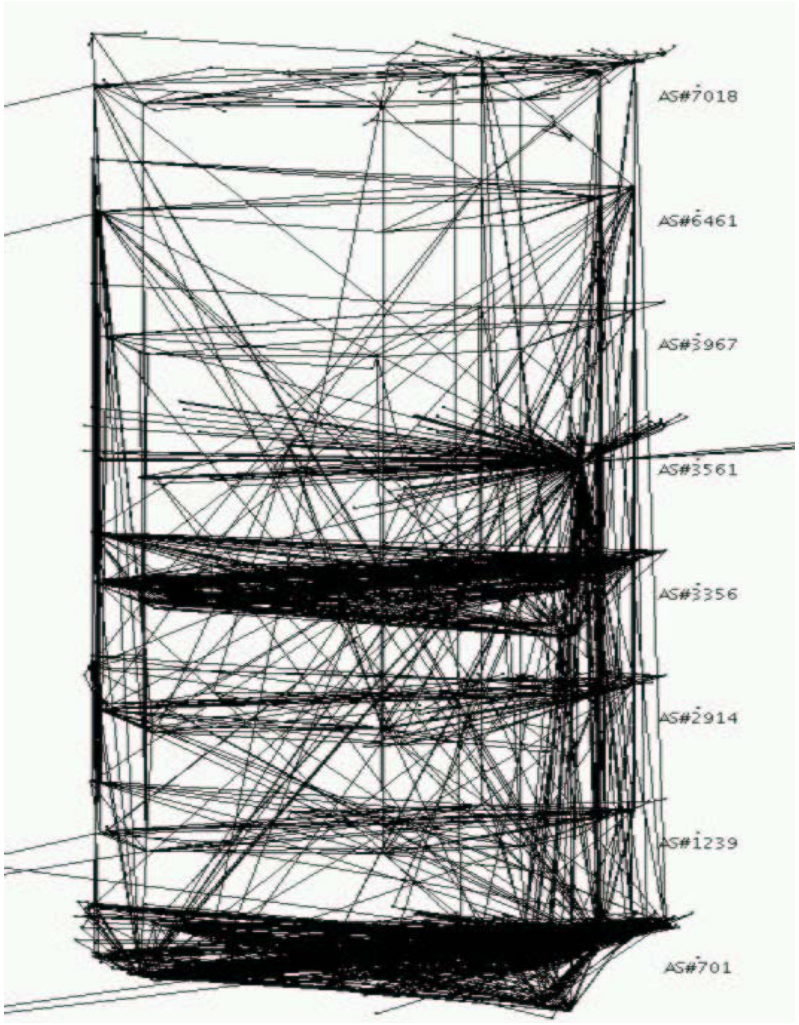
- Equinix, FR4, Lärchenstr. 110
- Equinix, FR5, Kleyerstr. 90
- Equinix, FR2, Kruppstr. 121-127
- e-shelter, Eschborner Landstr. 100
- I.T.E.N.O.S., Rebstöckerstr. 25-31
- Interxion, FRA1, Hanauer Landstr. 302
- Interxion, FRA2, Hanauer Landstr. 304A
- Interxion, FRA3, Weissmüller Str. 21
- Interxion, FRA4, Weissmüller Str. 19
- Interxion, FRA5, Hanauer Landstr. 308a
- Interxion, FRA6, Hanauer Landstr. 300a
- Interxion, FRA7, Hanauer Landstr. 296a
- KPN, Kleyerstr. 90
- Level3, Kleyerstr. 82 (Building A)
- Level3, Kleyerstr. 90
- NewTelco, Rebstöckerstr. 25-31 (Building B, Room B.1.10)
- TelecityGroup, Gutleutstr. 310
- Telehouse, Kleyerstr. 79 (Building K)
- Telehouse, Kleyerstr. 79 (Building I)

Challenge #7: **AS** meets **physical** Internet

- The AS-level Internet
  - IXPs are housed in one or more colo facilities
  - Colos/router hotels house the routers of one or more ASes
  - Inter-AS connectivity can manifest itself in many different physical connections (between distinct border routers)

- The physical Internet
  - Routers of the different ASes are housed in colos/router hotels
  - (Some US Tier-1s have separate facilities/buildings)
  - Intra-AS connectivity is the router-level view of the AS

An early attempt (D. Nicol et al 2003)



http://users.cis.fiu.edu/~liux/research/papers/topo-wsc03.pdf

Grand Challenge – What we have …

- Visualization of the Internet in 1994 (topology & traffic)

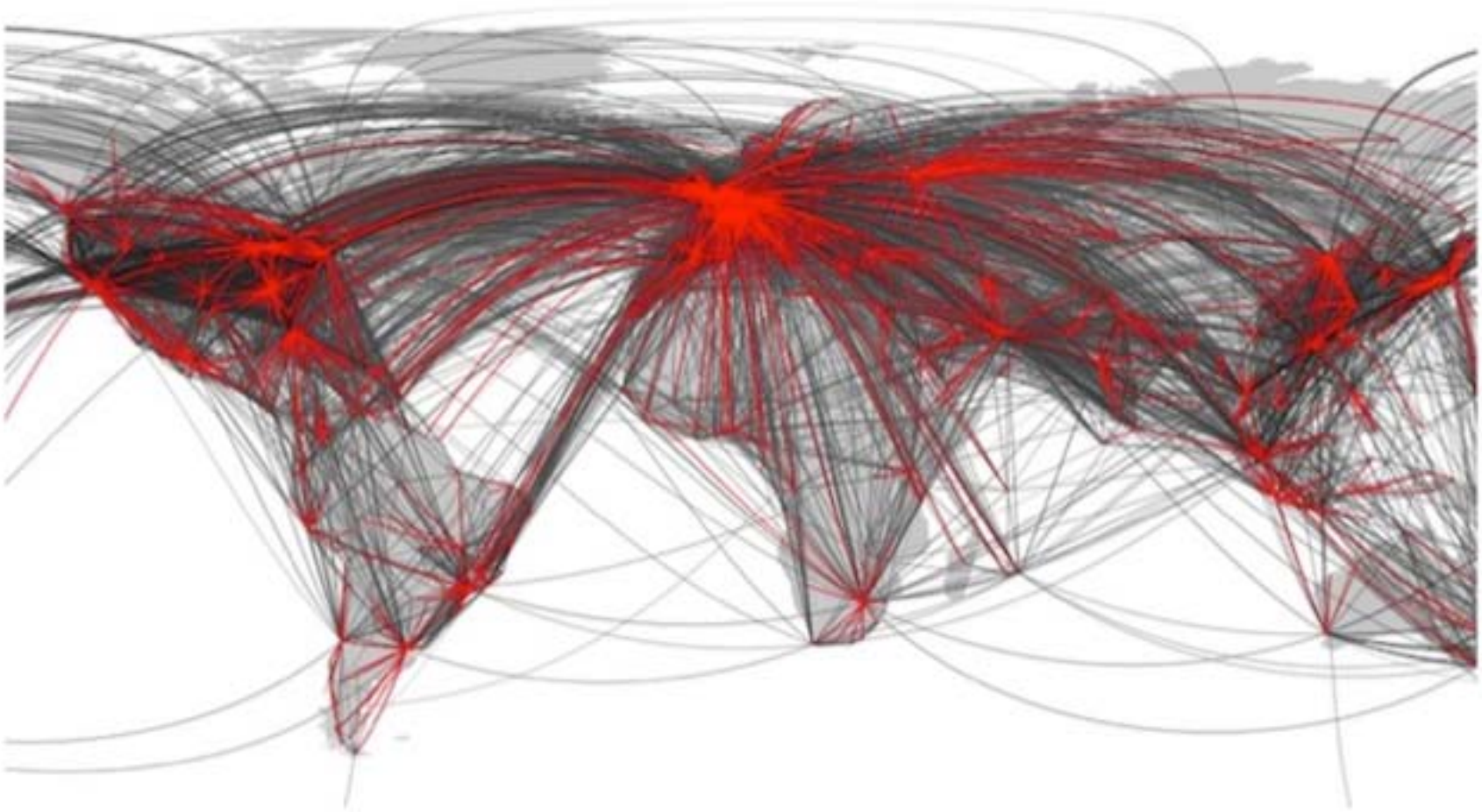
Grand Challenge – What we want …

- Visualization of the Internet in 2012 (topology & traffic)??

Why is this (very) hard?

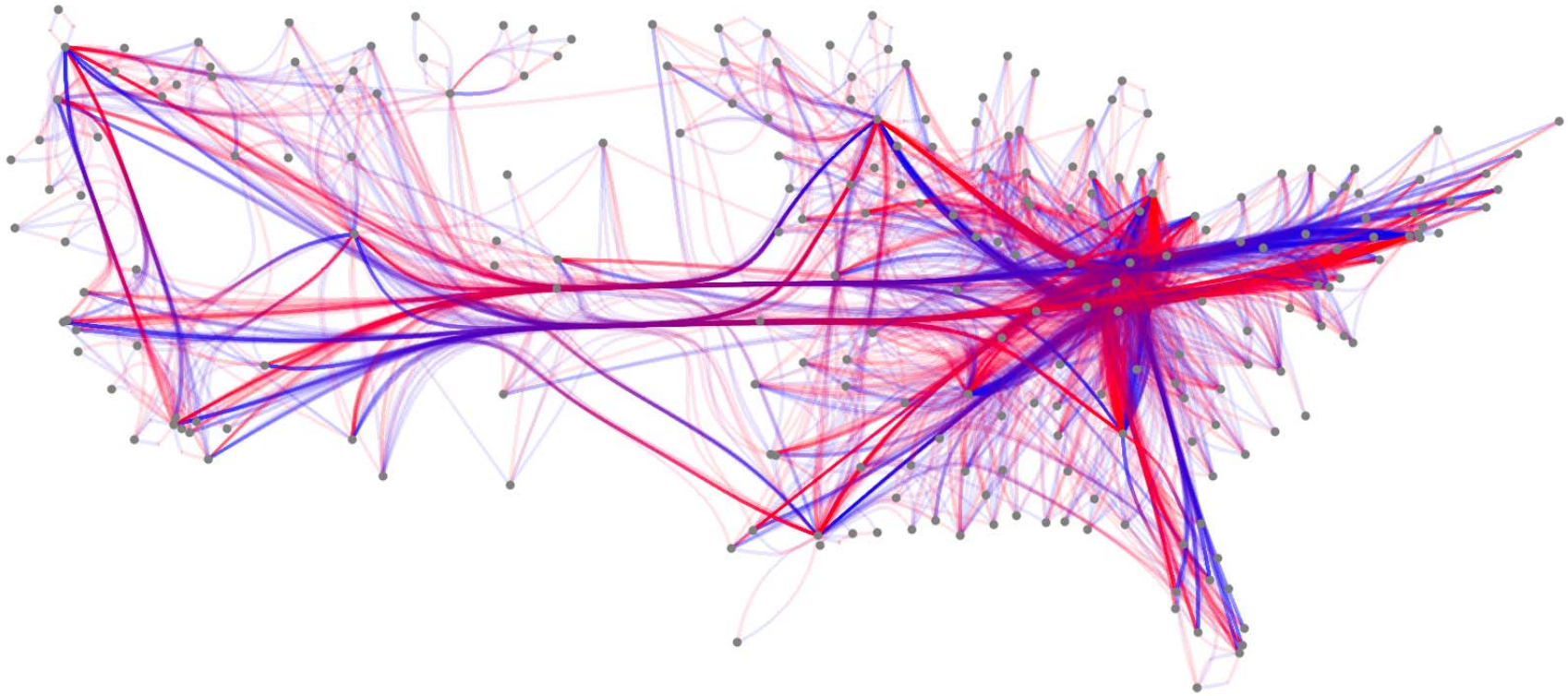
- What topology?
  - AS-level Internet topology (AS graph)
  - Physical Internet topology (router graph)
  - Main focus of some 15 years of Internet topology research
  - We know much less about this than we thought we did …
- What traffic?
  - Inter-domain traffic (AS traffic matrix)
  - How much traffic is exchanged between any pair of ASes?
  - We know next to nothing about this …
- What visualization?
  - ????

An analog: Worldwide airline system …



http://www.northwestern.edu/newscenter/stories/2012/06/complex-networks-skeletons.html

… or US airline traffic



http://vis.stanford.edu/files/2011-DividedEdgeBundling-InfoVis.pdf

Conclusion

- Past 15 years of research on the Internet's AS-level connectivity structure
  - Example of Grossman's (mis)quote of H.L Mencken:
    "Complex problems have simple, easy-to-understand wrong answers."

- Next 15 years of research on the AS-level Internet
  - Emphasis on "network of network" aspect
  - Deal with dynamics of and over this construct
  - What network economics for the AS-level Internet?

  - Major challenge ahead: inter-domain traffic information!

… and finally:

If you start to feel sorry about networking researchers because the reality of Internet measurement makes their lives/jobs difficult,

talk to the biologists or read their papers that describe their measurements, and you will realize what an easy life the networking researches have!

Some “fun” activities …

- traceroute experiments (IV)
  - Run traceroute from a machine you can access …
  - … to a target and make sure it traverses a given AS link …
- traceroute experiments (V)
  - Run traceroute from a machine you can access …
  - … to a target and make sure it traverses a given AS link in a specific city
- traceroute experiments (III)
  - In case (V), how would you go about determining in which colocation facility your probe packets were handed over?

Thanks!

Questions?