

Mapping the Internet

hierarchy
topology
ownership

Mads Sas
madalina.sas@pm.me
<https://mis.pm>

1. Hierarchy

1.1 The layered protocol stack

1.2 Modelling the stack's evolution

2. Topology

2.1 A historical detour

2.2 Mapping the Physical layer

2.3 Mapping the Network layer

2.4 Mapping the World Wide Web

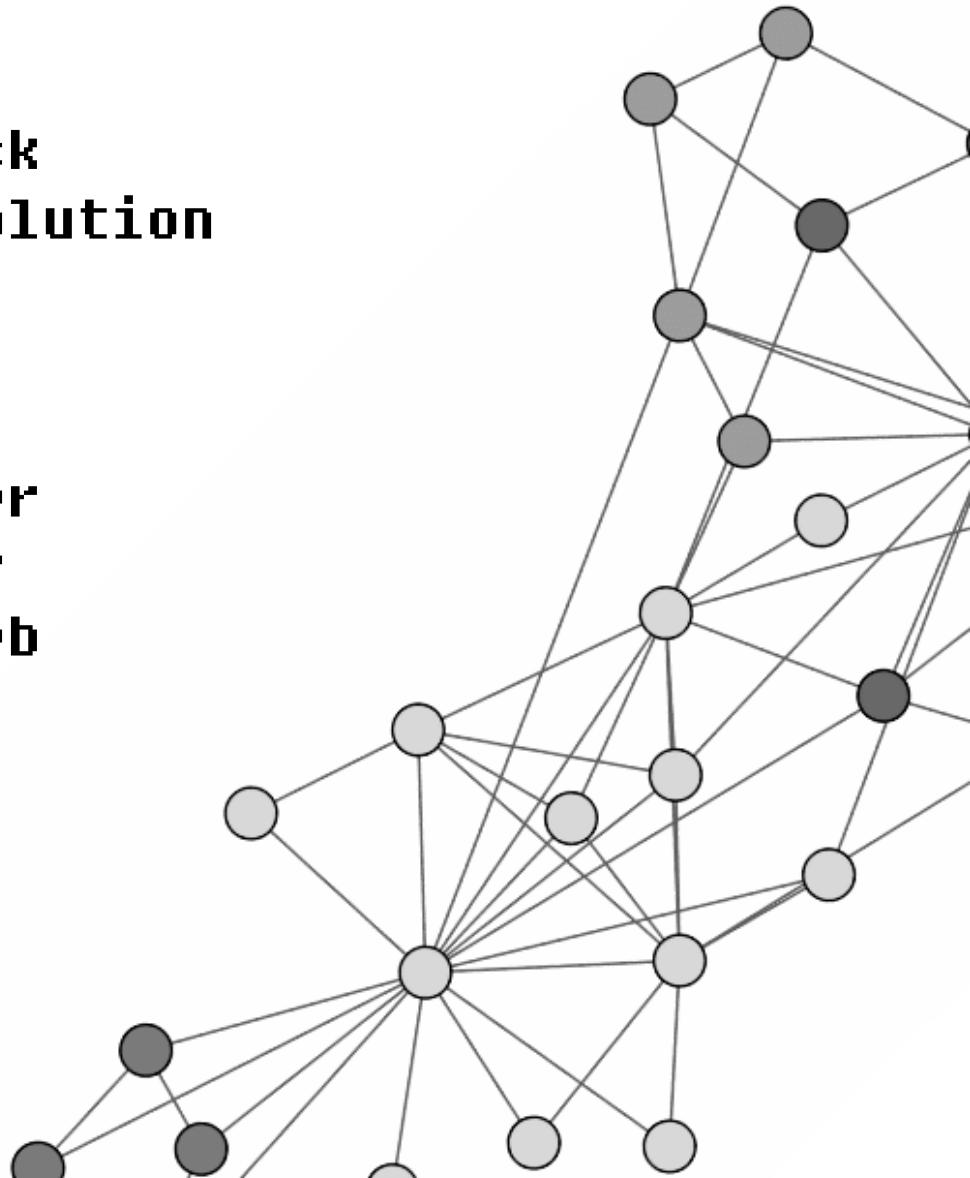
2.5 Networks of Cyber crime

3. Ownership

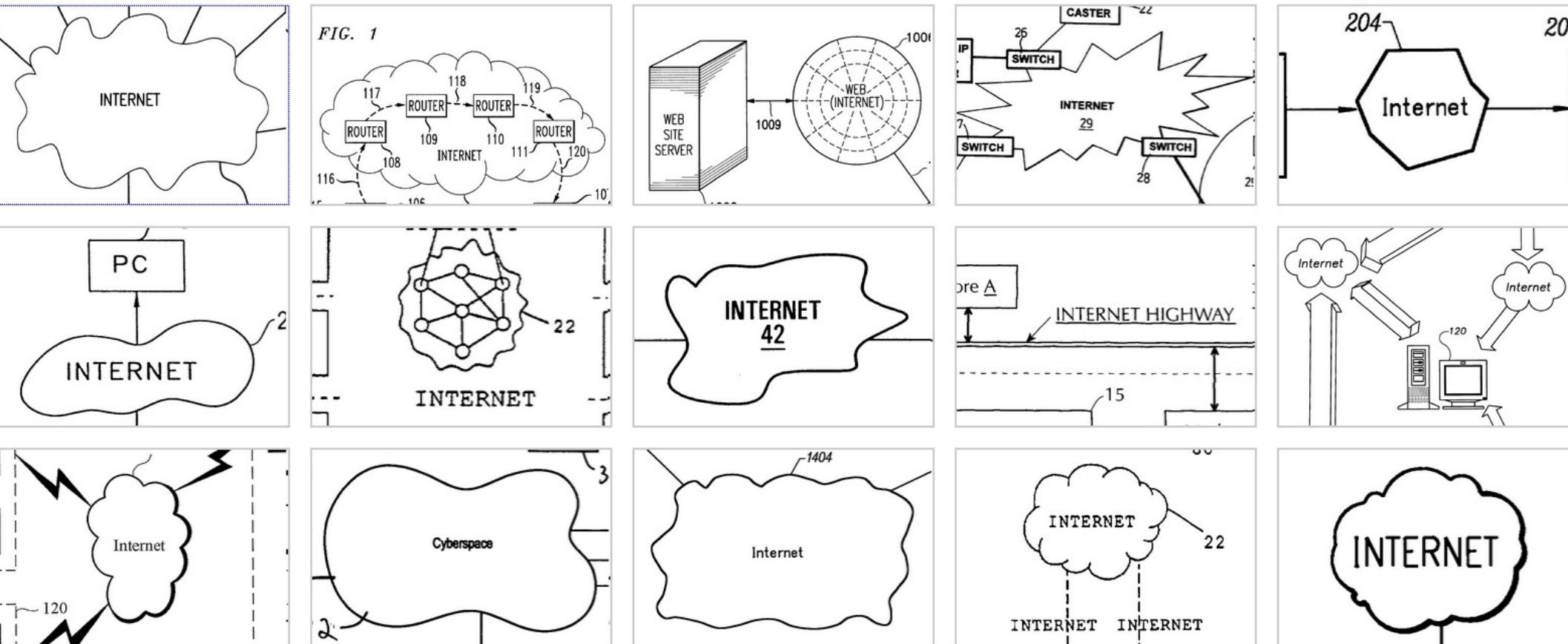
3.1 Management

3.2 Surveillance

3.3 Censorship



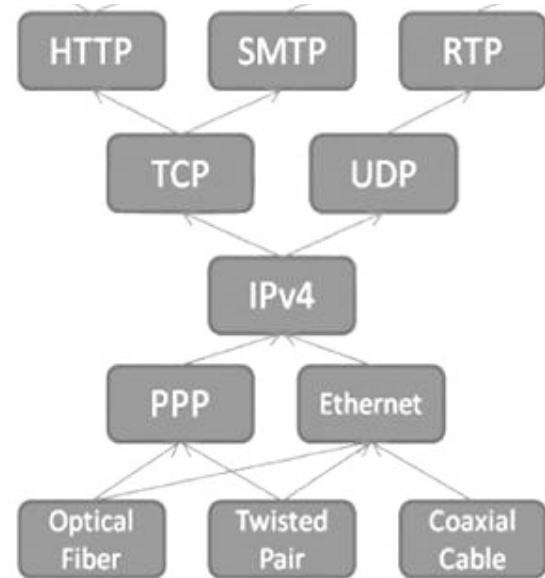
So what's the shape of the Internet?



Hierarchy

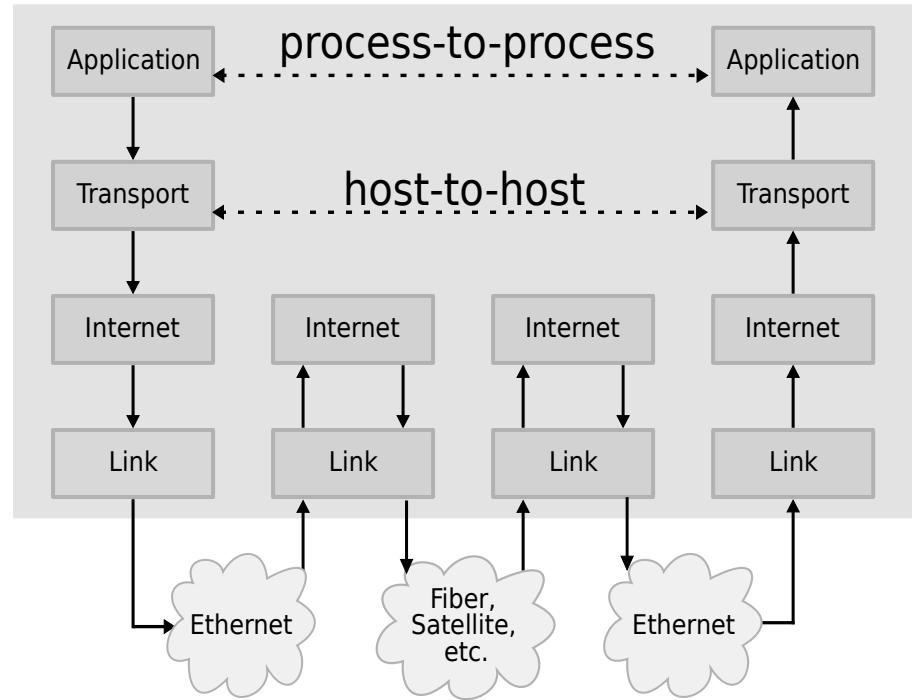
Hierarchical architecture

- TCP/IP stack is split into layers of abstraction. Each layer serves the one above and is facilitated by the one beneath.
- Entities at the same layer but in different hosts communicate with each other via protocols.
- Each layer contains changing and evolving protocols similar to an ecosystem.
- The ecosystem tends to an hourglass shape: innovation tends to survive at the top and bottom layers but not the middle.



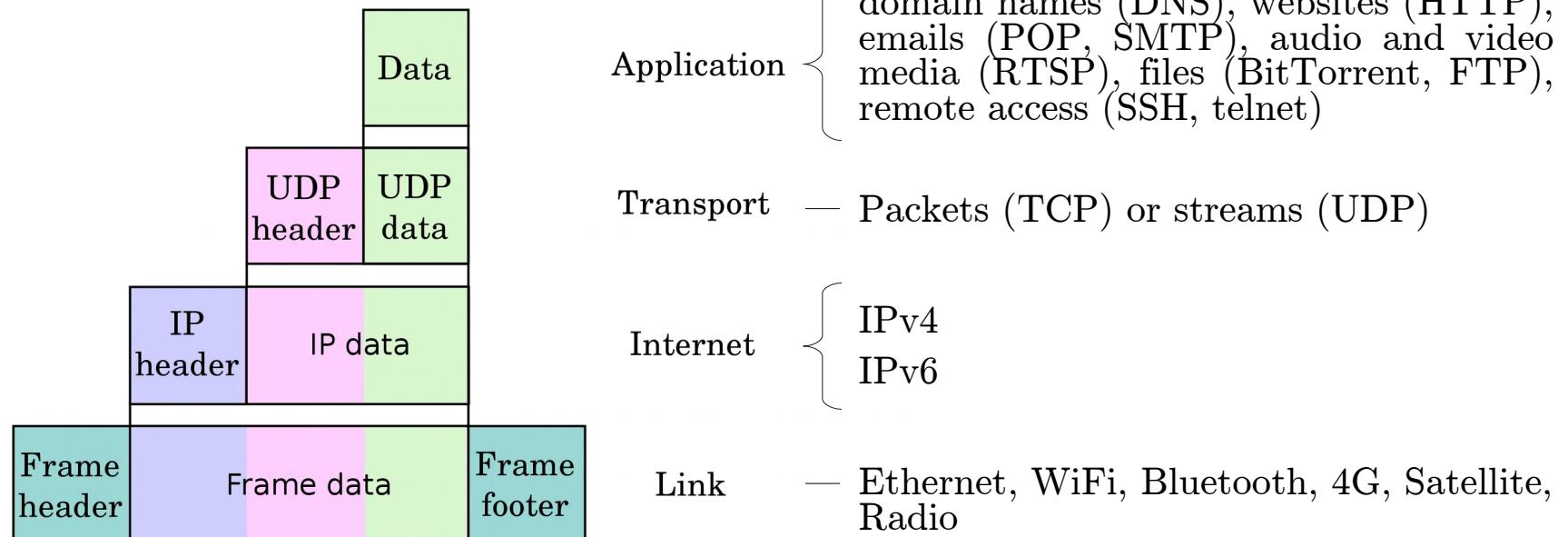
Hierarchical architecture

- Users send and receive data from the application layer
- The transport layer handles splitting it into packets, sequencing them, correcting for errors and detecting dropped packets
- The network layer handles routing the message to the right destination
- The physical layer handles sending the message over a physical channel
- End-to-end principle: network is “dumb”, endpoints communicate



TCP/IP stack

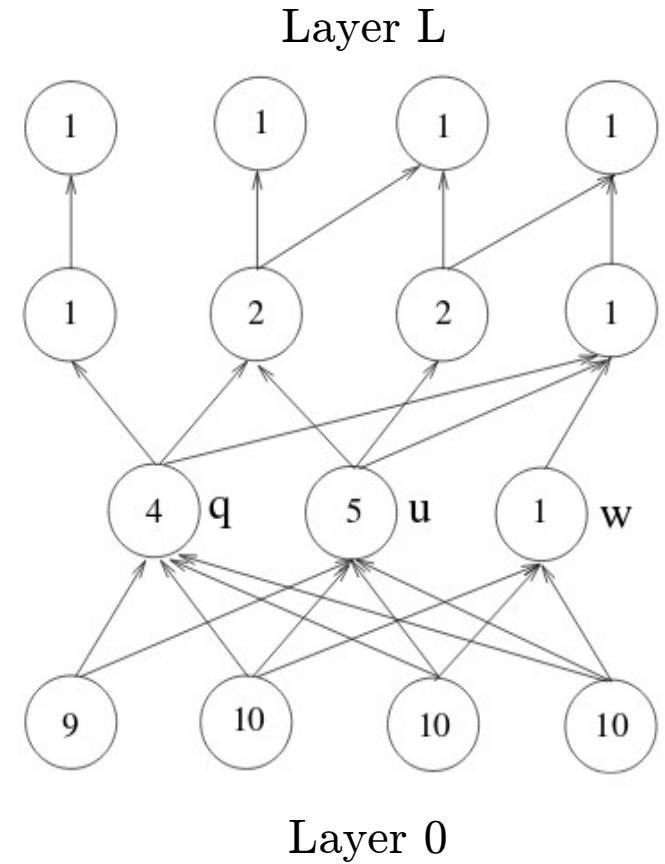
- A lot of protocols at the application and physical (link) layer, only a few in the middle layers – why?



The EvoArch model

- Internet represented as DAG with L layers
- A node u at layer l is connected to some nodes p_i at layer $l+1$ if the protocol p_i is supported by protocol u i.e. p_i is a product of u
- Each node has a value v that captures whether it is more likely to survive its competition
- The value $v(u)$ of a protocol u is driven by the values of the protocols that depend on u

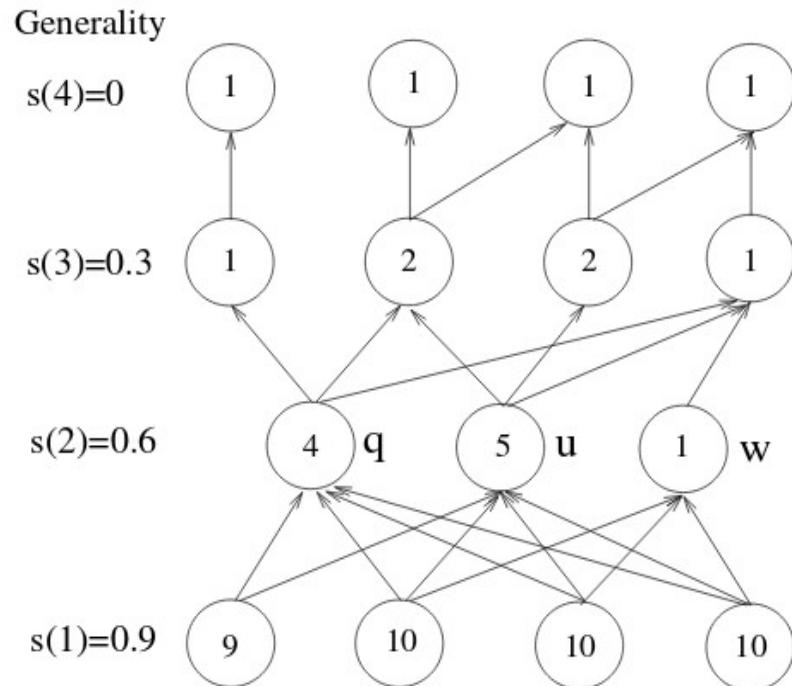
$$v(u) = \begin{cases} \sum_{p \in P(u)} v(p) & l(u) < L \\ 1 & l(u) = L \end{cases}$$



The EvoArch model

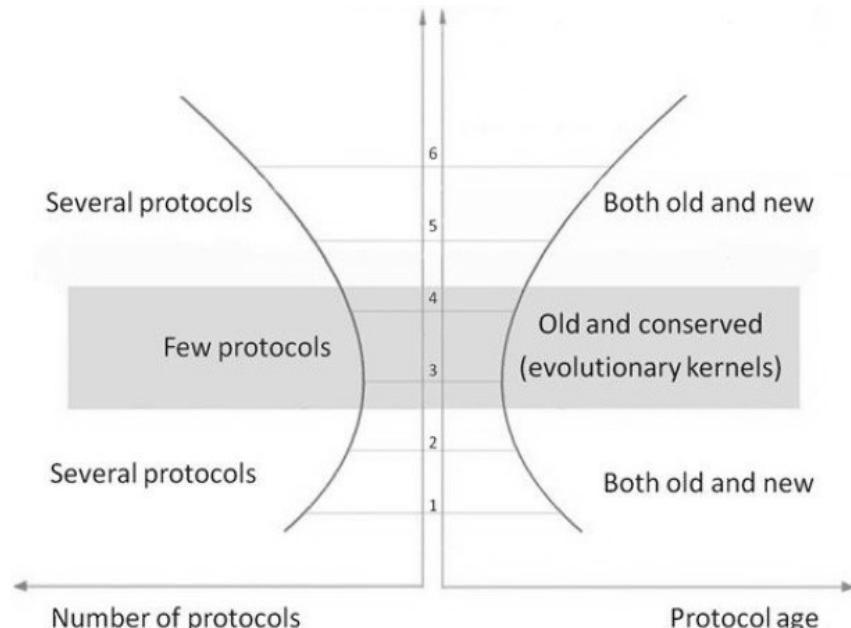
Model parameters

- Generality vector s : captures generalisation or specialisation of each protocol in layer l w.r.t. layer $l-1$. Generality decreases as we move up the stack.
- Competition threshold c : a node u competes with a node w if u shares at least c of w 's products.
- Mortality z : captures intensity of competition.
- When u dies, all products p_i also die if their only substrate is u . This leads to a cascade effect.



The EvoArch model

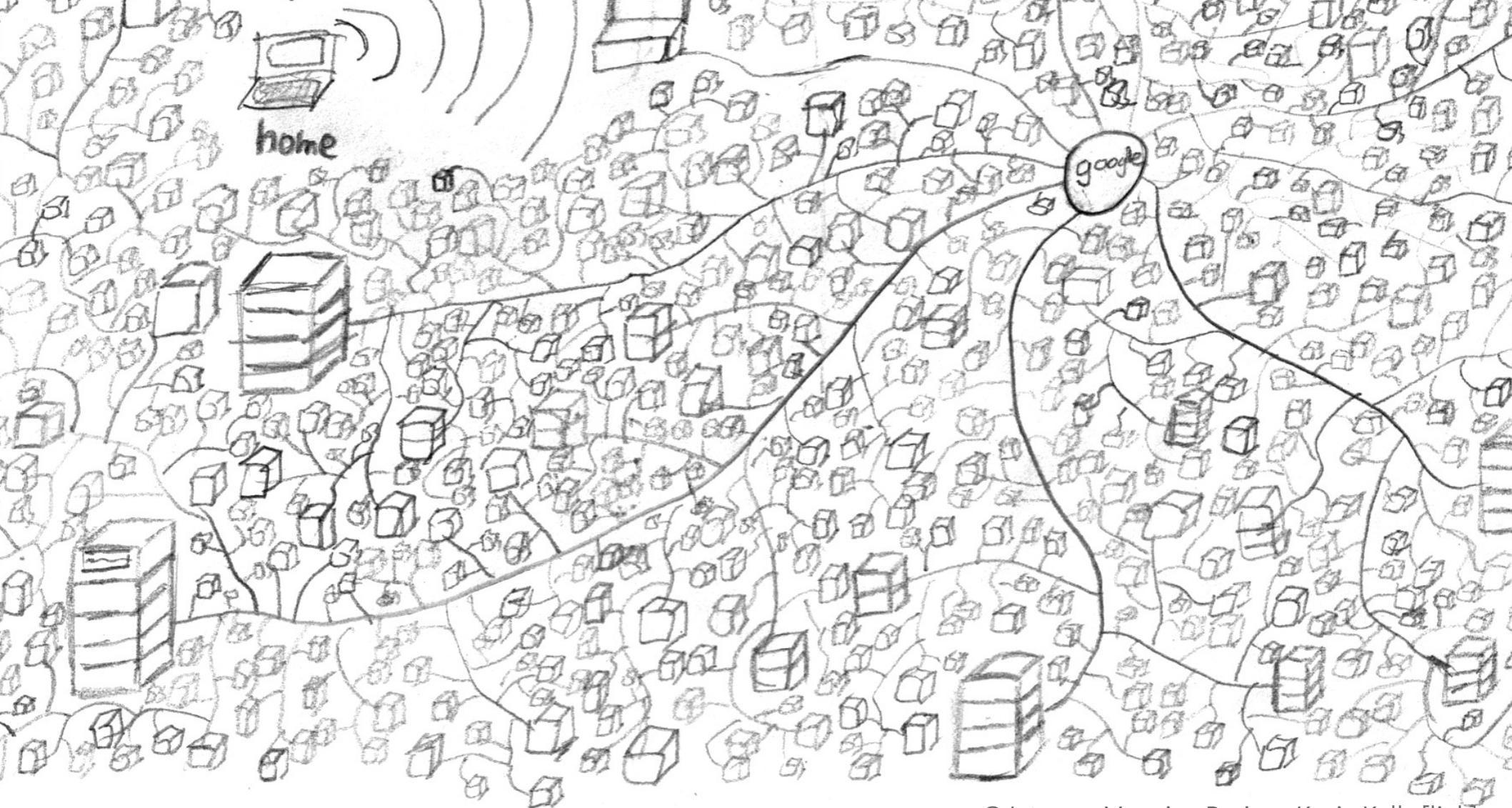
- Randomly generates layered architectures and observes them evolve
- Captures the inherent competition between nodes in the same layer
- The lower the layer, the least generality
- Middle layers appear to be more resilient to competition, with old protocols outliving new attempts (e.g. IP)
- Similar hourglass architectures have been observed in metabolic and gene expression networks as well as the organization of the immune system



Topology

THE INTERNET MAPPING PROJECT

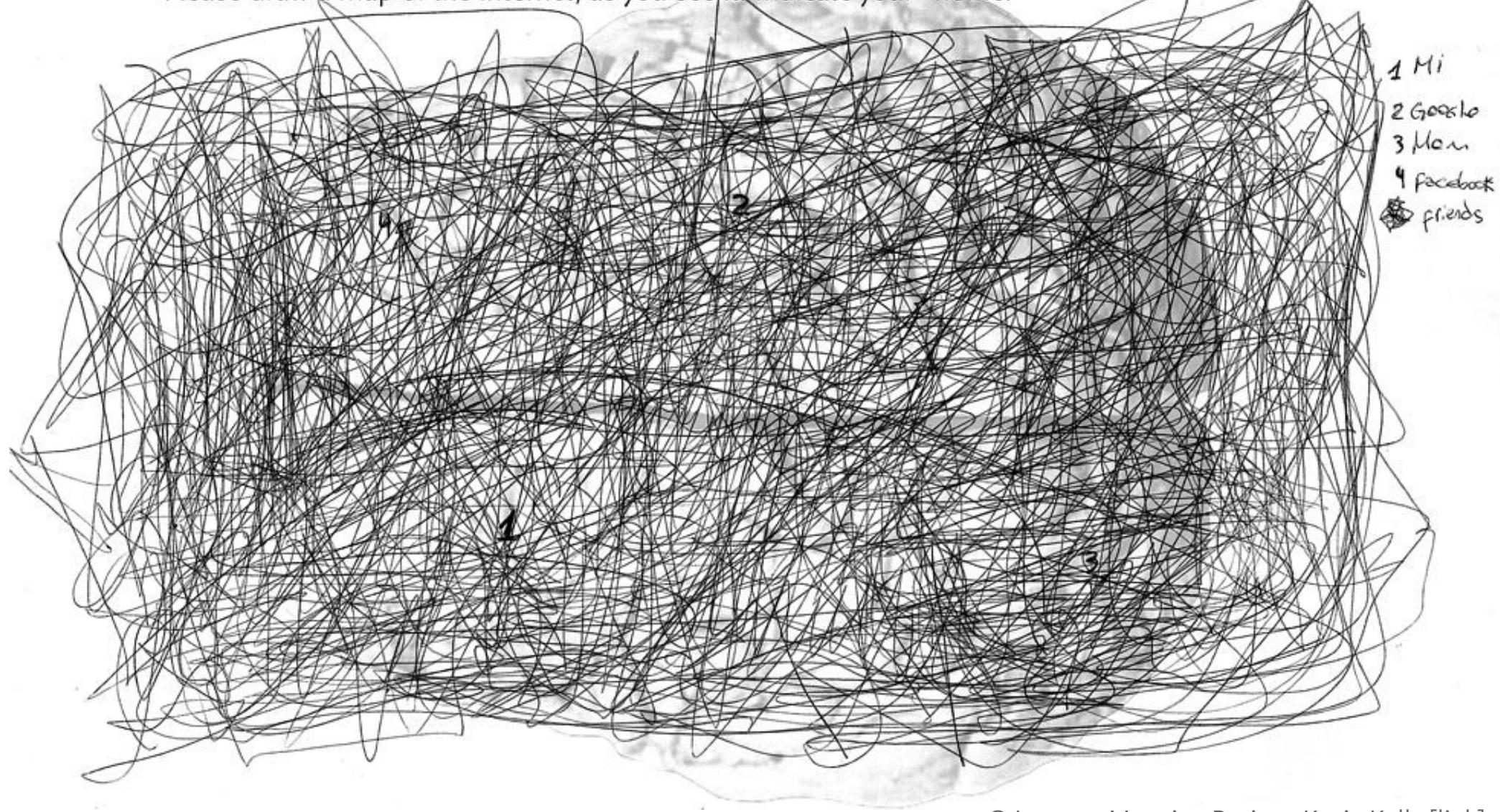
Please draw a map of the internet, as you see it. Indicate your "home."

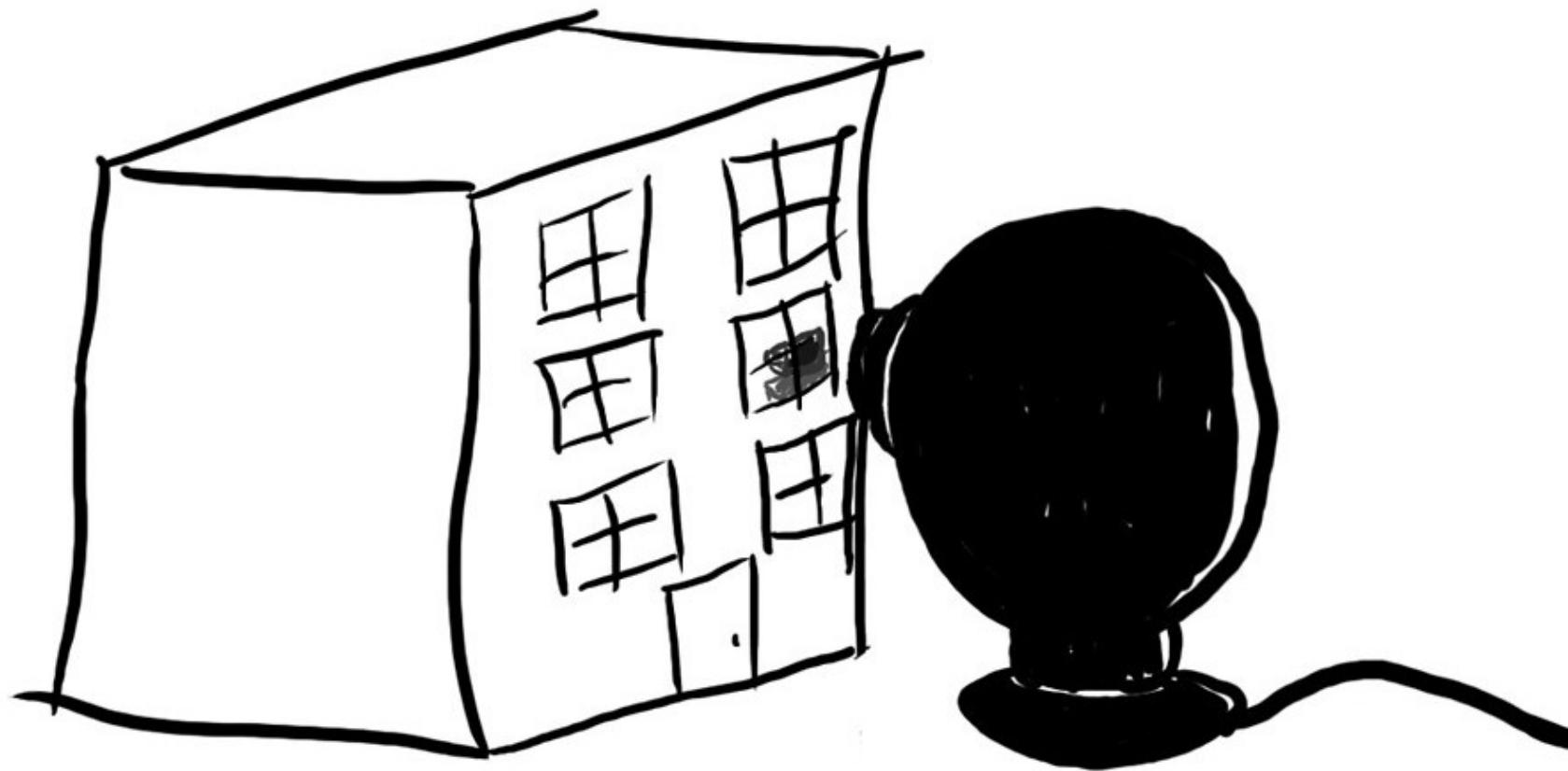


© Internet Mapping Project, Kevin Kelly [[link](#)]



© Internet Mapping Project, Kevin Kelly [link]



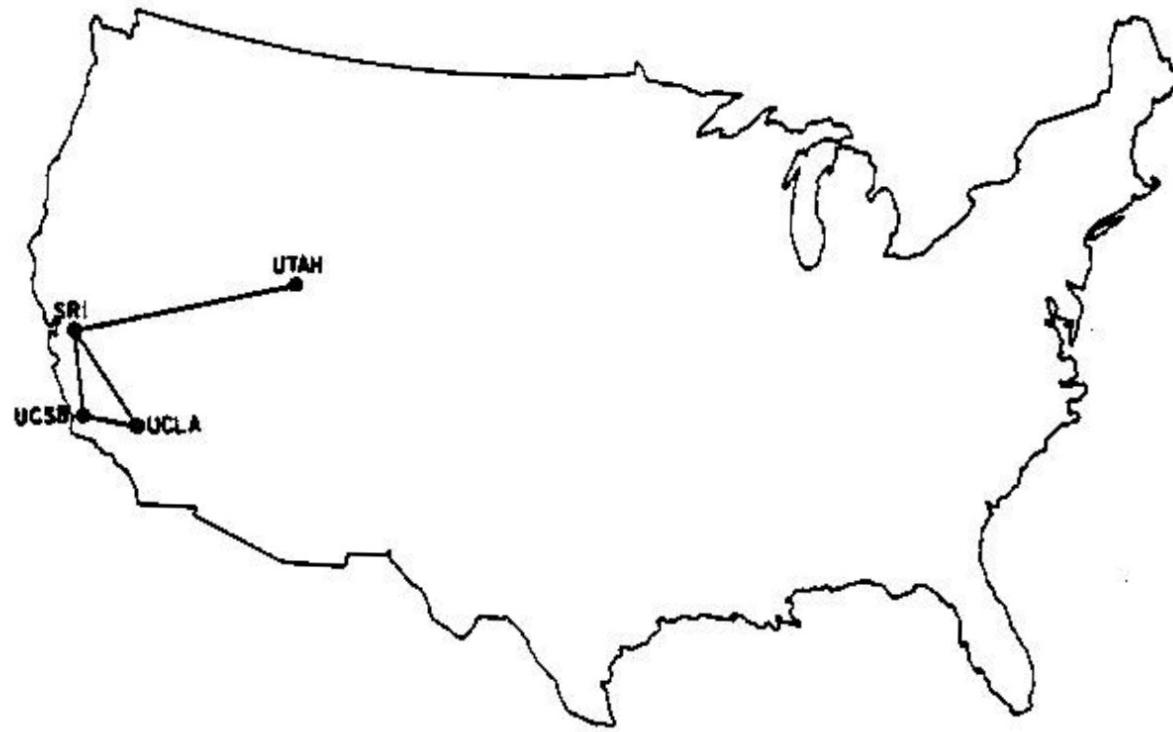


So what's the shape of the Internet again?

A historical detour...

1969

ARPANET

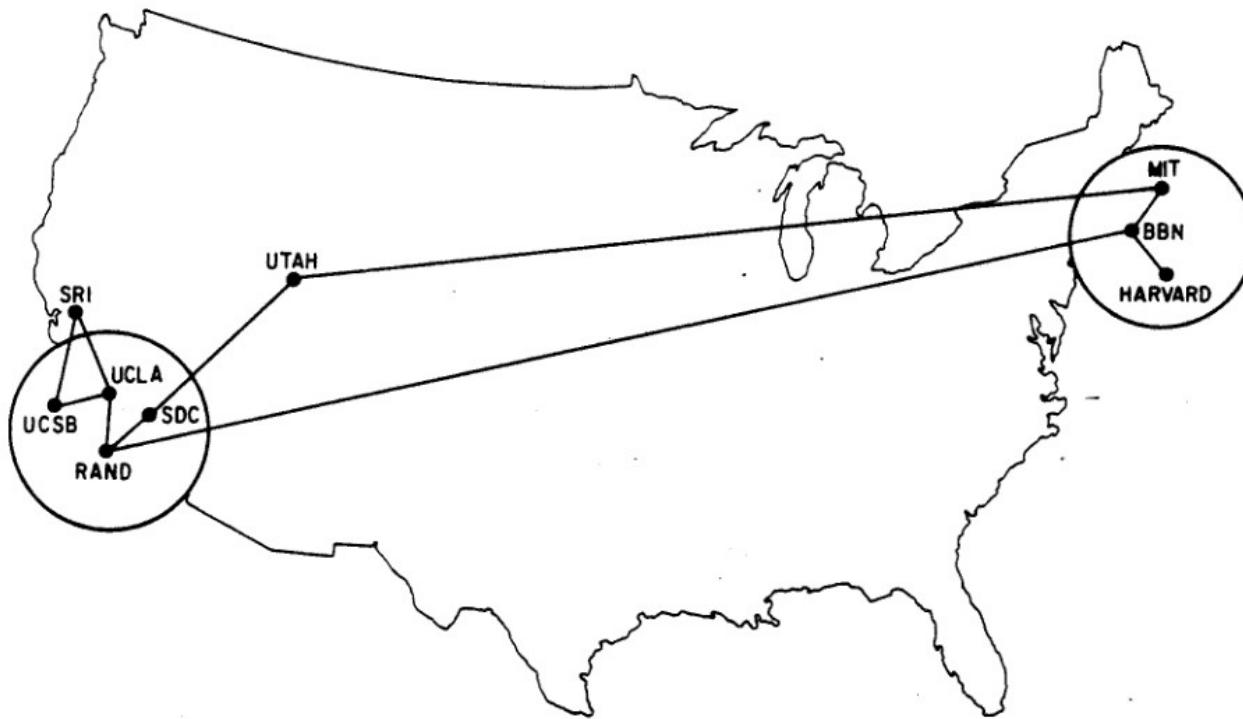


Node – Interface Message Processor (IMP) – gateway – router

Edge – phone line

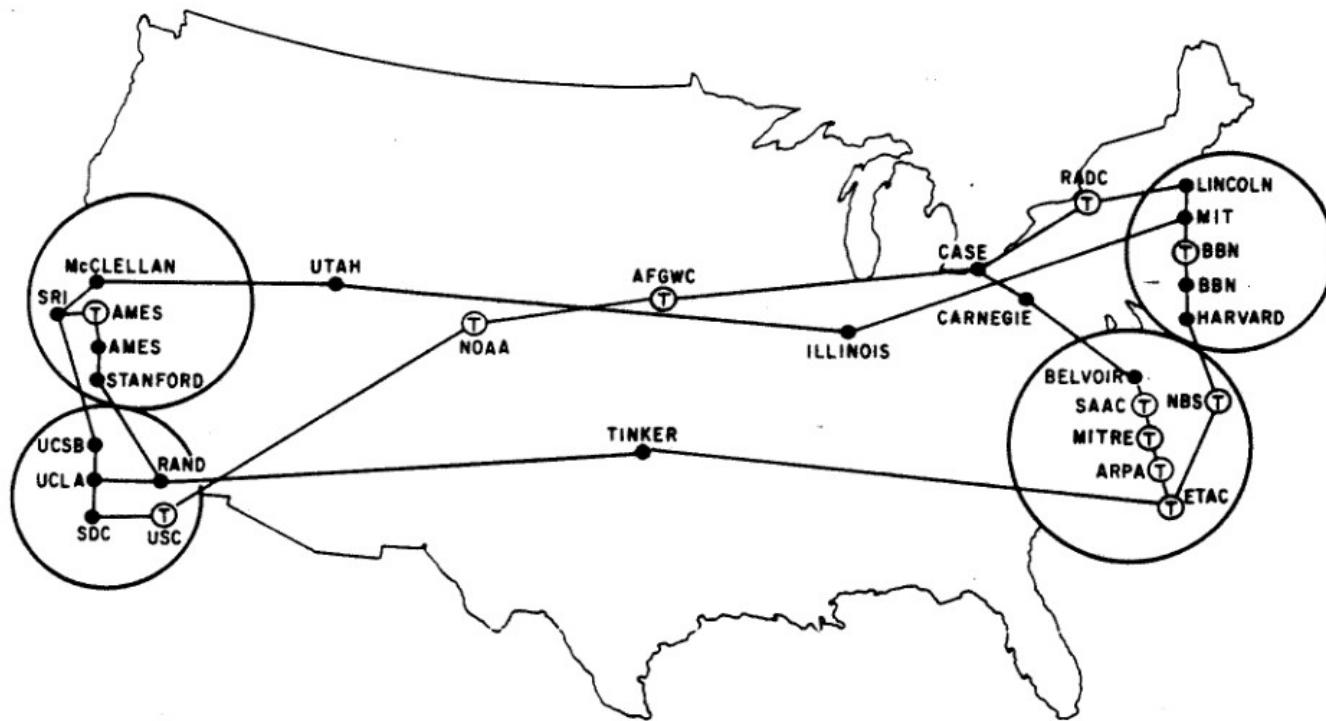
1970

ARPANET



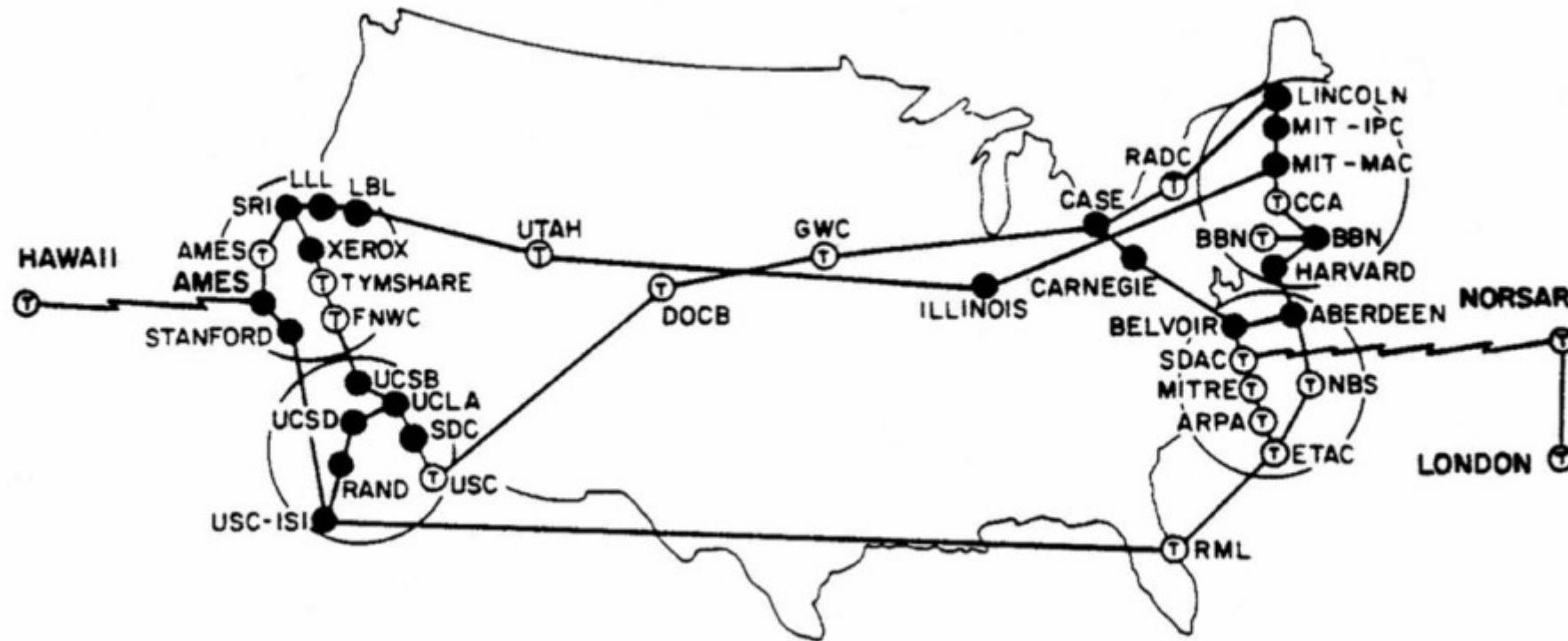
1972

ARPANET



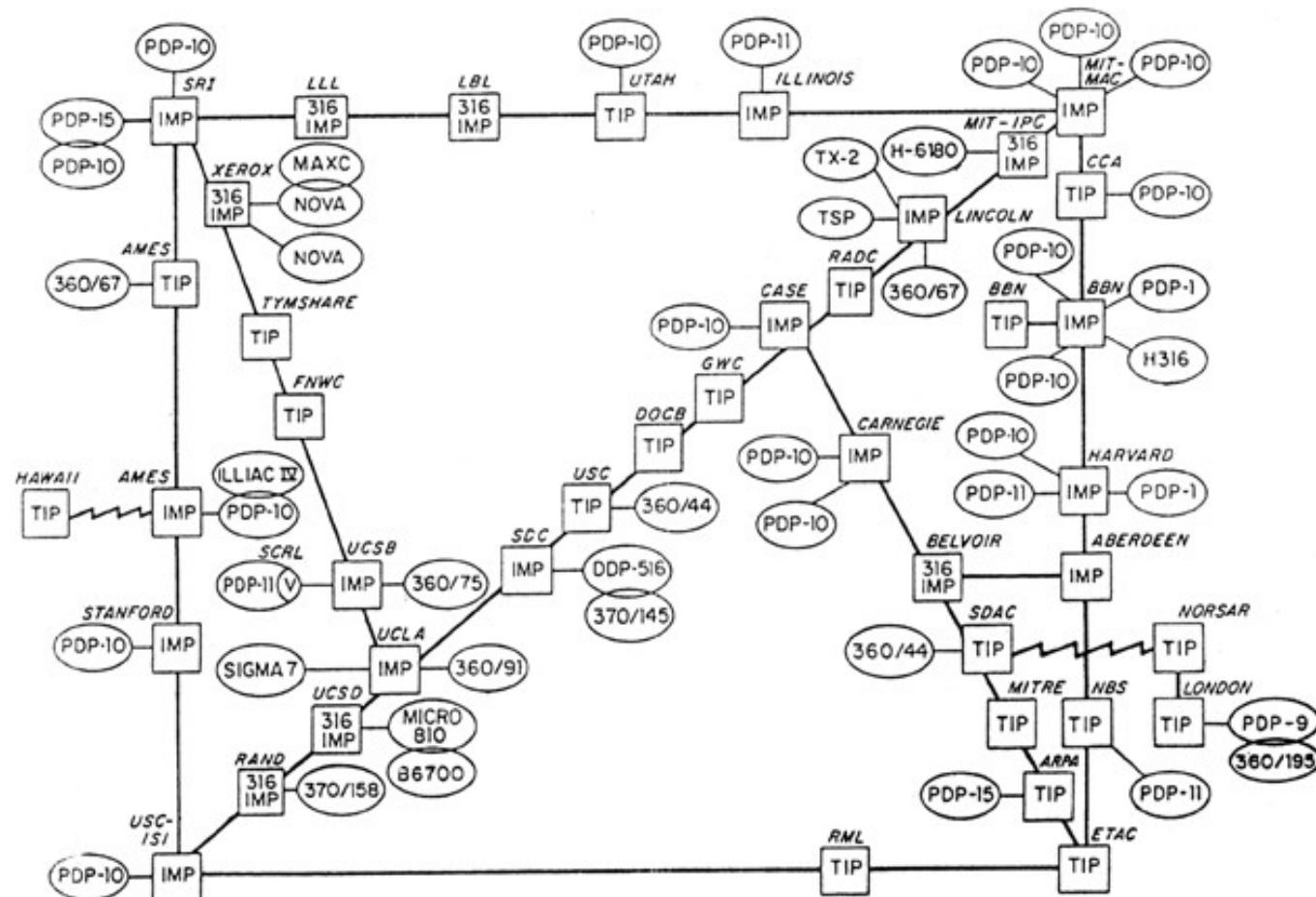
1973

ARPANET



1973

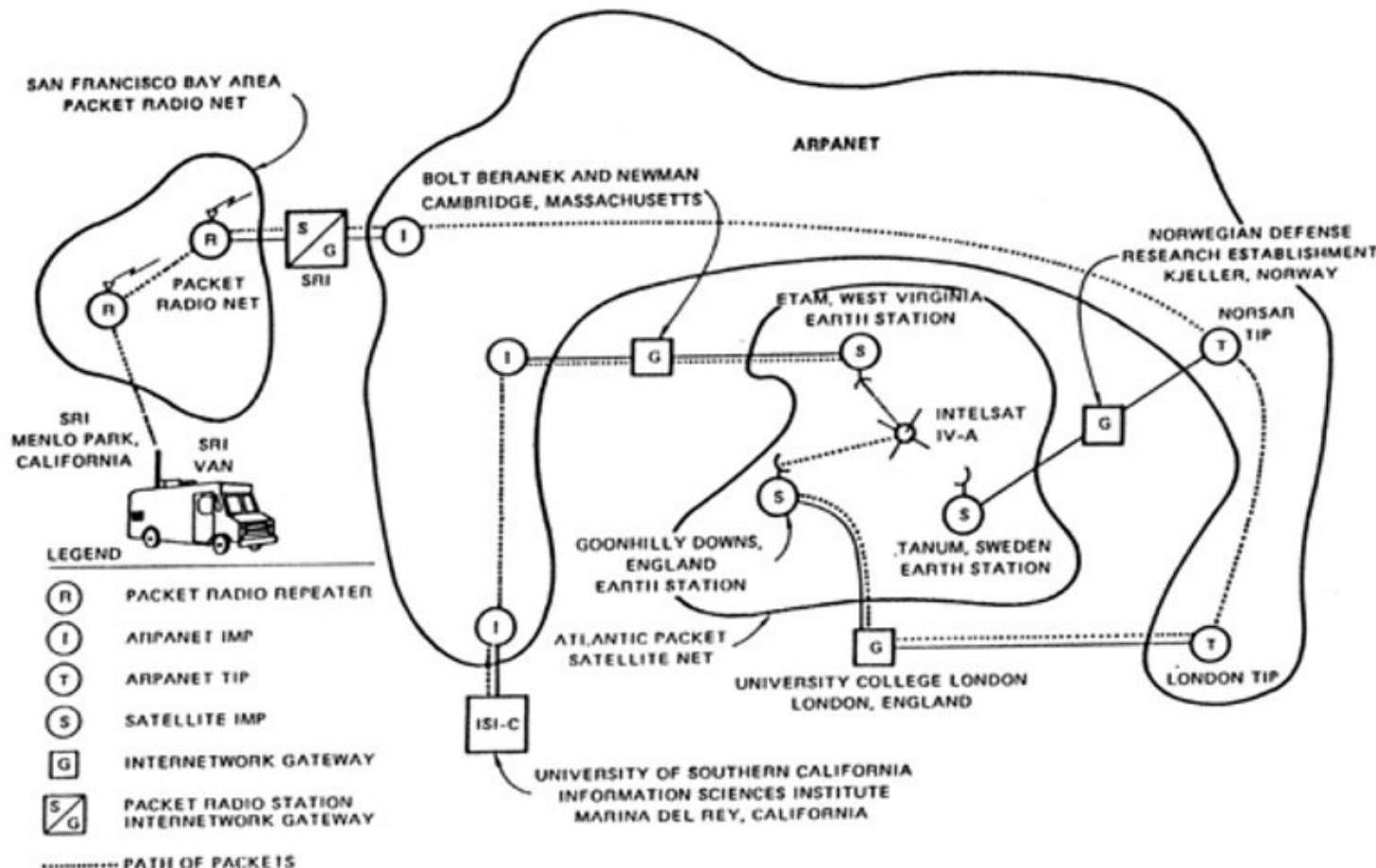
ARPANET + NPL – Logical Map



Node – IMP (□) /
mainframe (○)
Edge – phone line

1977

"The first demonstration of the inter network"



Node – IMP

Clique – network

Edge – phone line/satellite link/radio link

So what's the shape of the Internet again?

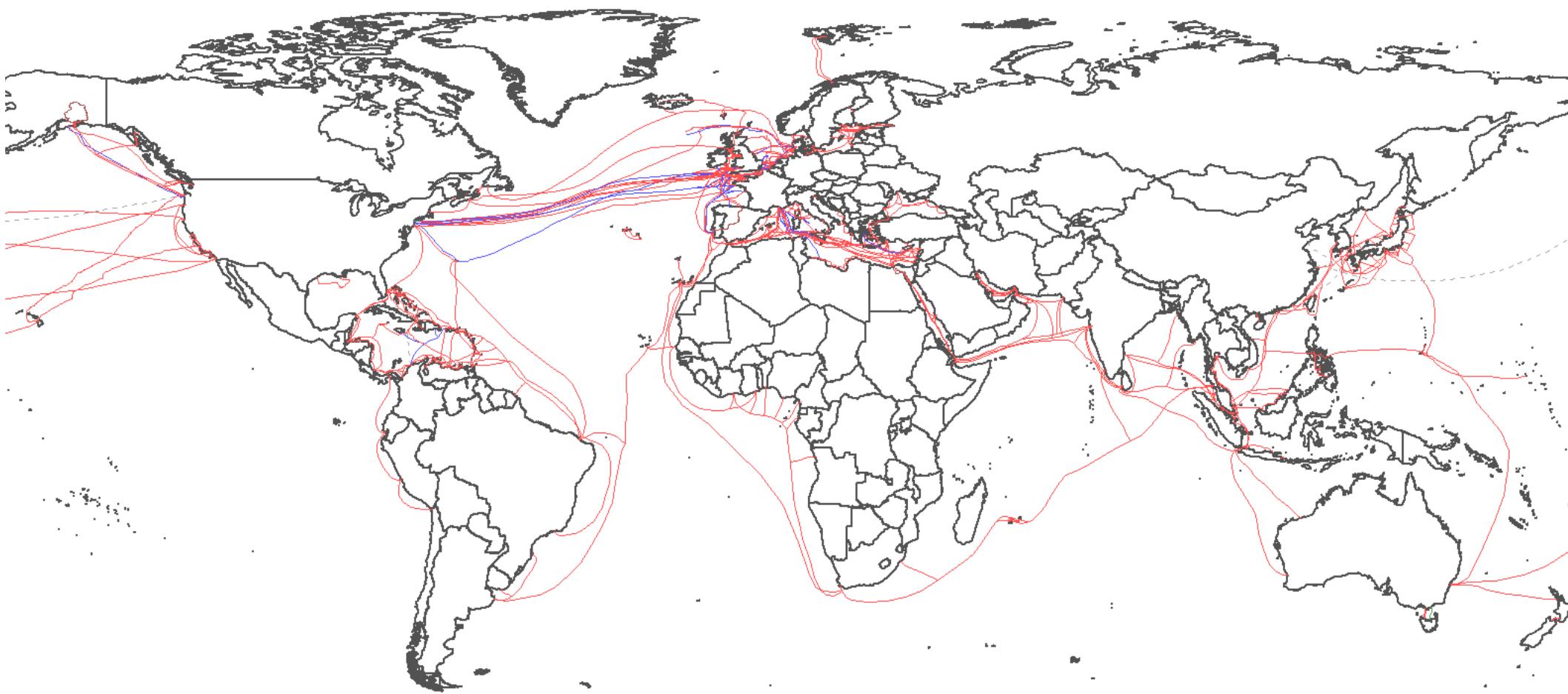
Network of Networks?

Mapping the physical layer

- First layer of the Internet Stack, the only layer that refers to hardware
- The nodes are routers, switches, and modems, the edges are cables and satellite, cellular or wireless links
- Hierarchical dynamics: large commercial, academic or governmental internet service providers (ISPs) distribute connectivity downstream
- ISP networks are interconnected at Internet Exchange Points (IXPs) via high-bandwidth fiber optic cable
- Locations of most fiber routes and IXPs are public data, so the “backbone” of the physical infrastructure of the Internet can be mapped

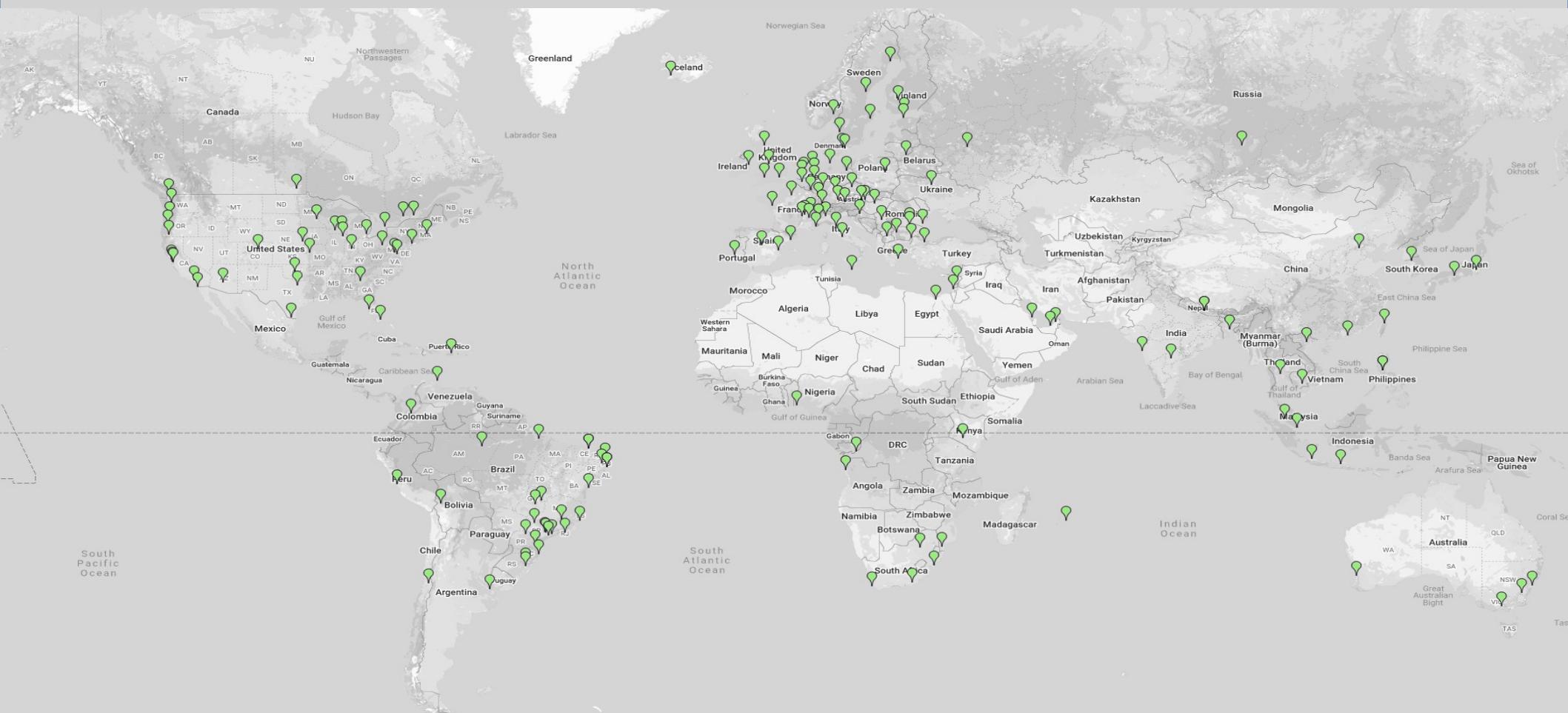
2007

Underwater cables map



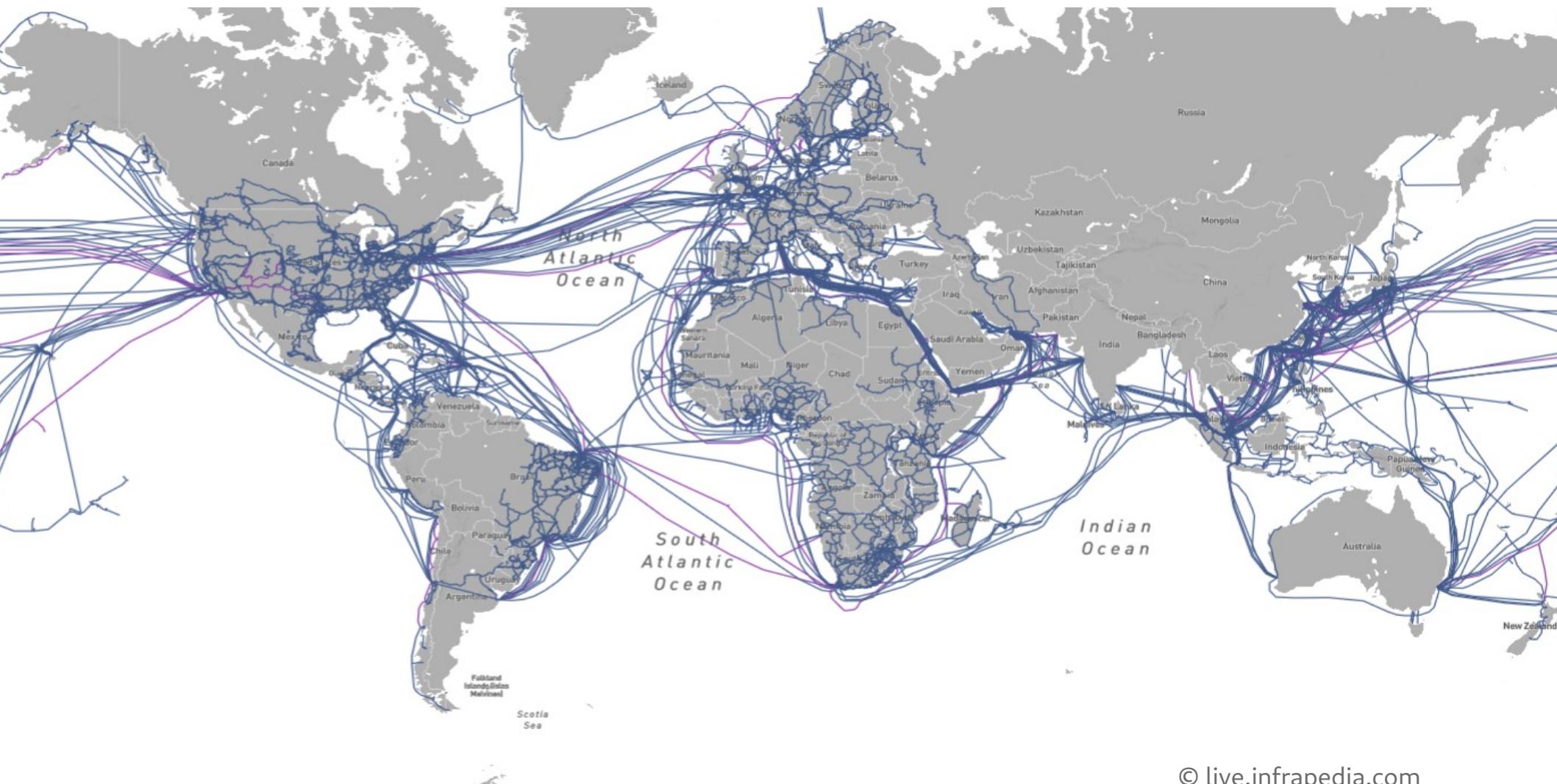
2018

Internet Exchange Points (IXP) map



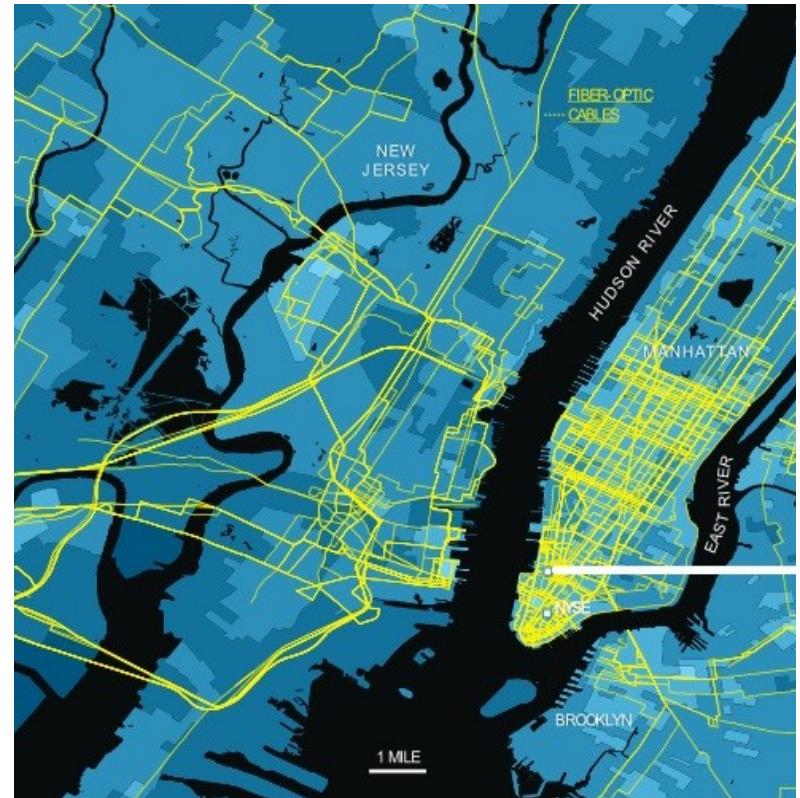
2020

Internet physical infrastructure map



Dynamics of the physical layer

- No actual routing at the physical layer, just physical connectivity
- Bandwidth and traffic volume can illustrate physical layer dynamics
- Most fiber routes are underwater but maps exist for some terrestrial routes
- Disaster statistics: watch out for ships, hungry sharks and old Georgian ladies digging for scrap copper!



Manhattan fiber routes © geo-tel.com

Woman who cut internet to Georgia and Armenia 'had never heard of web'

A 75-year-old woman arrested for single-handedly cutting off the internet in Georgia and Armenia has tearfully insisted she is innocent and had never heard of the internet.



Hayastan Shakarian has been arrested for single-handedly cutting off the Internet in Georgia and Armenia on March 28 Photo: AFP

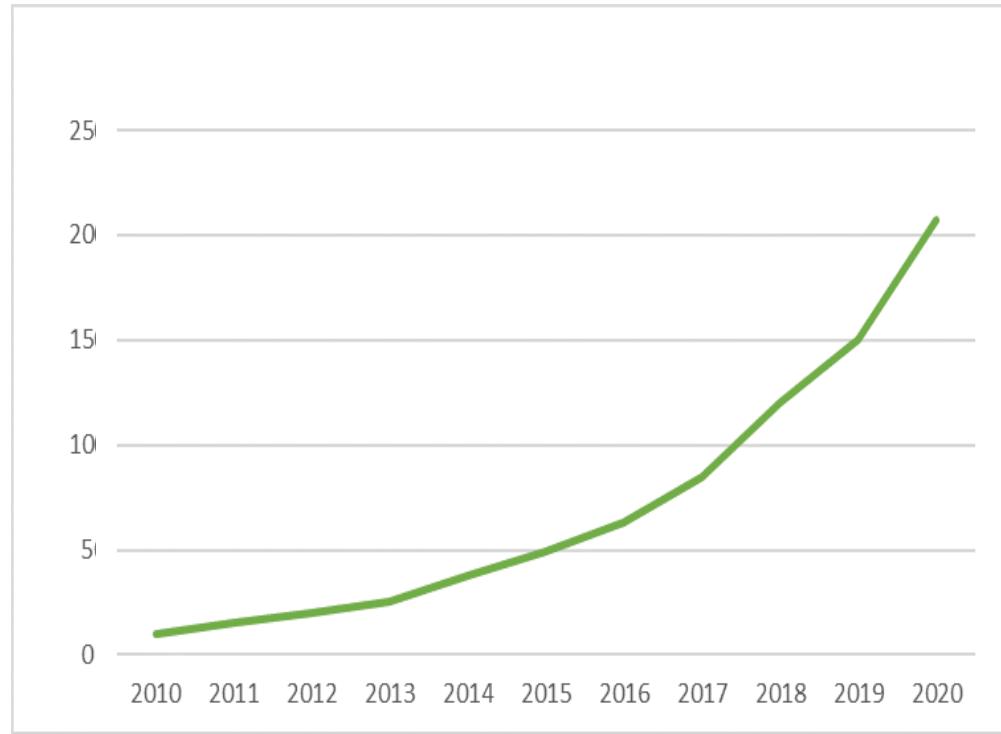
Europe news»

- [France news](#)
- [German news](#)
- [Italy news](#)
- [Spanish news](#)
- [Russian news](#)
- [European Union](#)

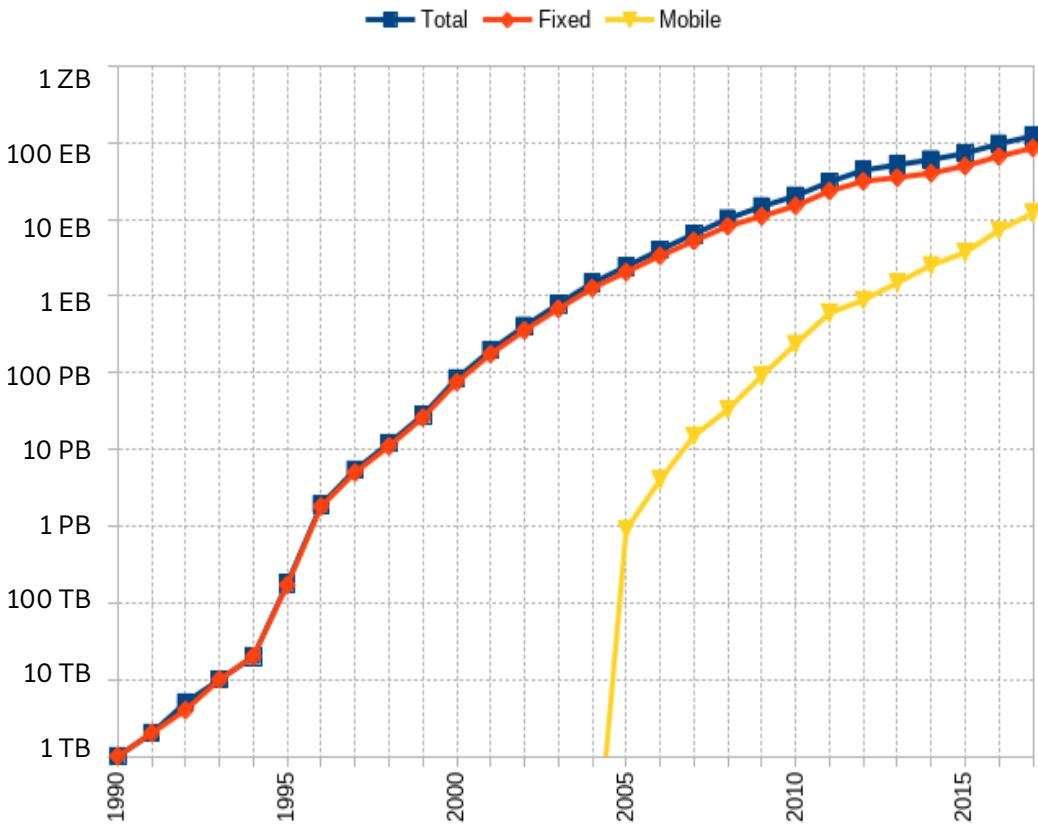
Temporal evolution

- 1842 – first telegraph underwater line, 1859 – first transatlantic cable
- Now >95% of global Internet traffic goes through submarine cable
- Terrestrial and aerial routing points added on existing routes – new local ISPs, local networks, WiFi and BlueTooth hubs
- Millions of new devices connect to the internet every year, the network periphery grows exponentially (first PCs, then phones, now Internet of things)
- Edholm's law: Internet bandwidth in telecommunication networks doubles every 18 months
- Enter the zettabyte era

Billions of connected devices



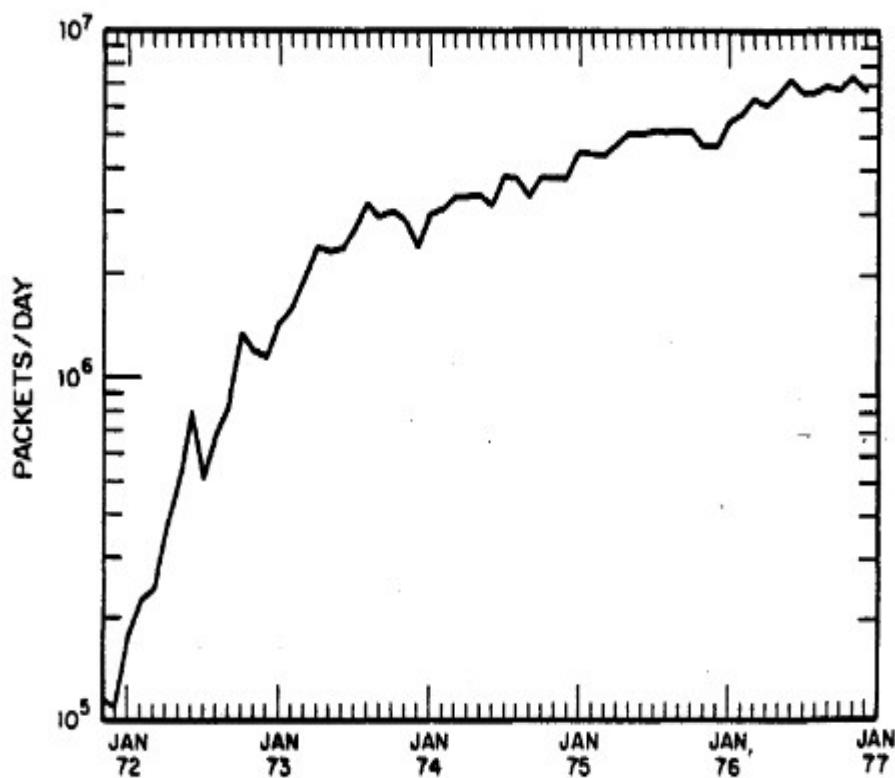
Monthly Internet Traffic (log)



img © Internet Systems Consortium, [link]

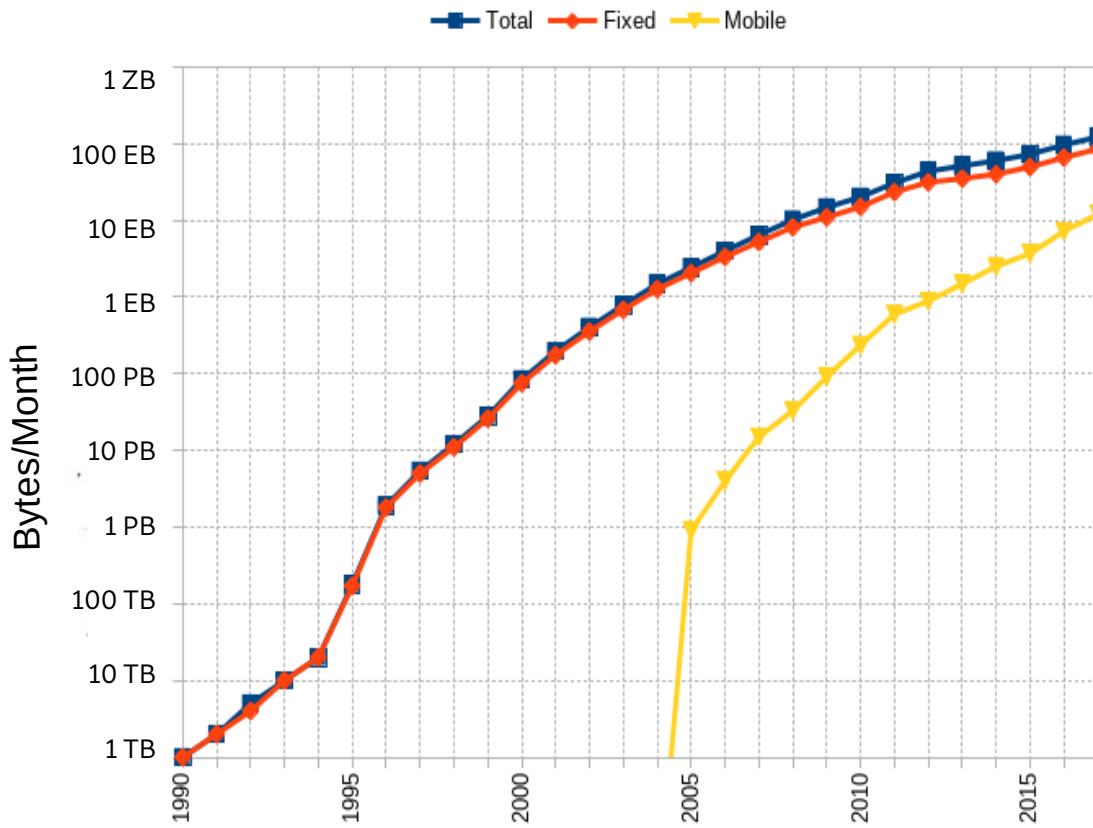
data © Cisco traffic projections, [link]

ARPANET HOST INTERNODE TRAFFIC



Img © DARPA, ARPAnet completion report [link]

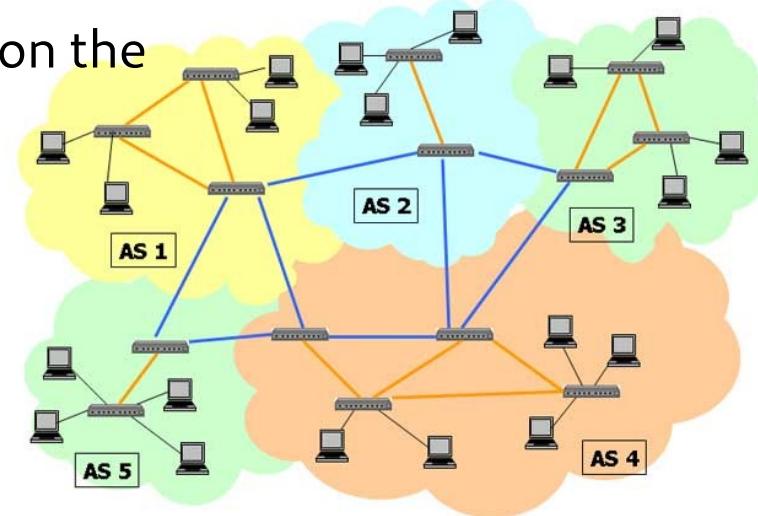
Monthly Internet Traffic (log)



data © Cisco traffic projections, [link]

The network layer

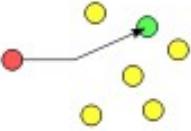
- At the network layer, each node is identified by its IP address
- A host is a node at the edge of the network that participates in user applications
- The network layer's core is formed of Autonomous Systems (AS), blocks of connected IP addresses inter-connected by core routers
- ASs are managed by upstream ISPs and hosted on the physical infrastructure of fiber cable and IXPs
- Packets travel through networks from router to router until their destination
- Internet backbone: the principal data routes between core routers / ASs



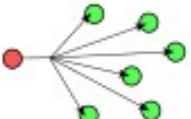
Mapping the network layer

- In this layer, nodes are hosts or routers and edges are routing paths
- Routing is the process of directing network packets from source to destination host through intermediate network nodes
- Can map geographically or topologically
- Idea: send packets to unknown hosts from a known host
- Geolocation: infer distance from latency of packets sent to the unknown host and use triangulation to deduce its location
- Tracerouting: observe the route packets take to the unknown host to deduce network topology

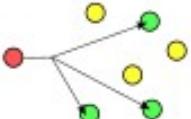
Unicast



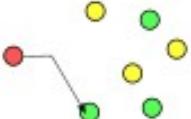
Broadcast



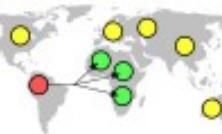
Multicast



Anycast

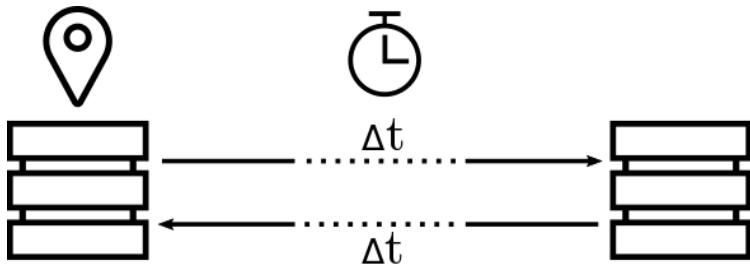


Geocast

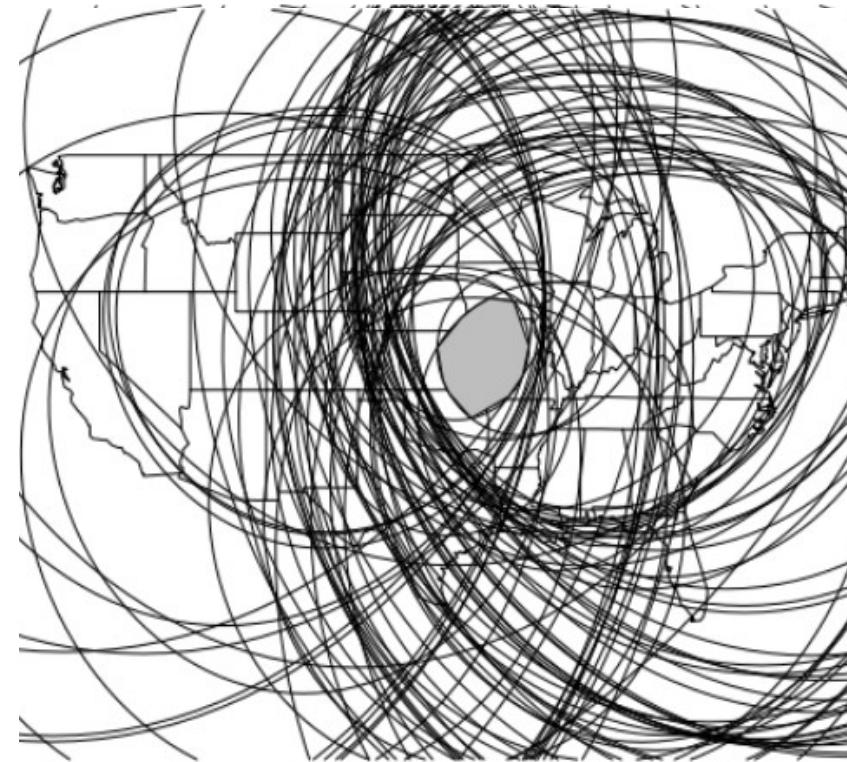


Geolocation

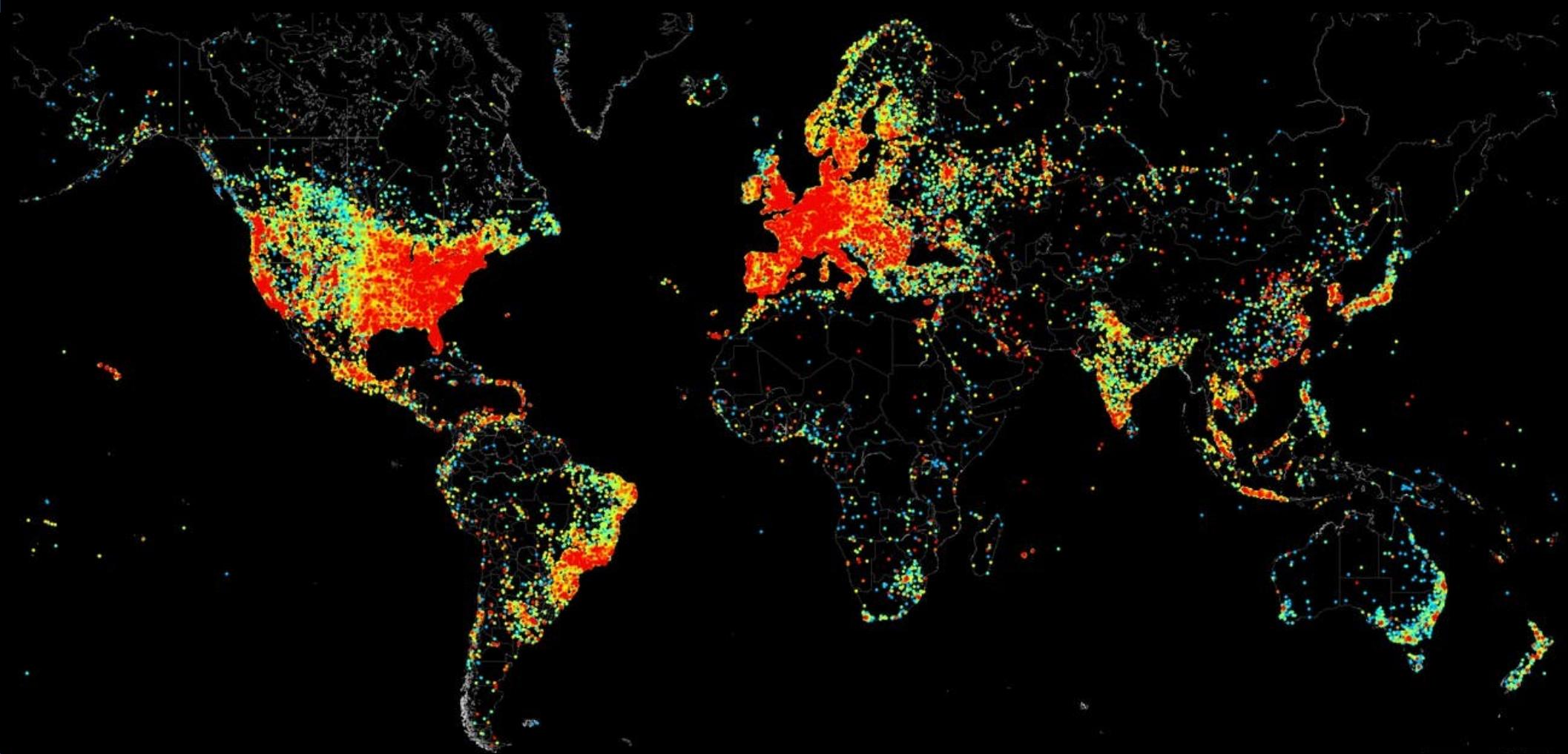
- From some known hosts A_i send a message to an unknown host B and estimate the distance from the delay in receiving a reply

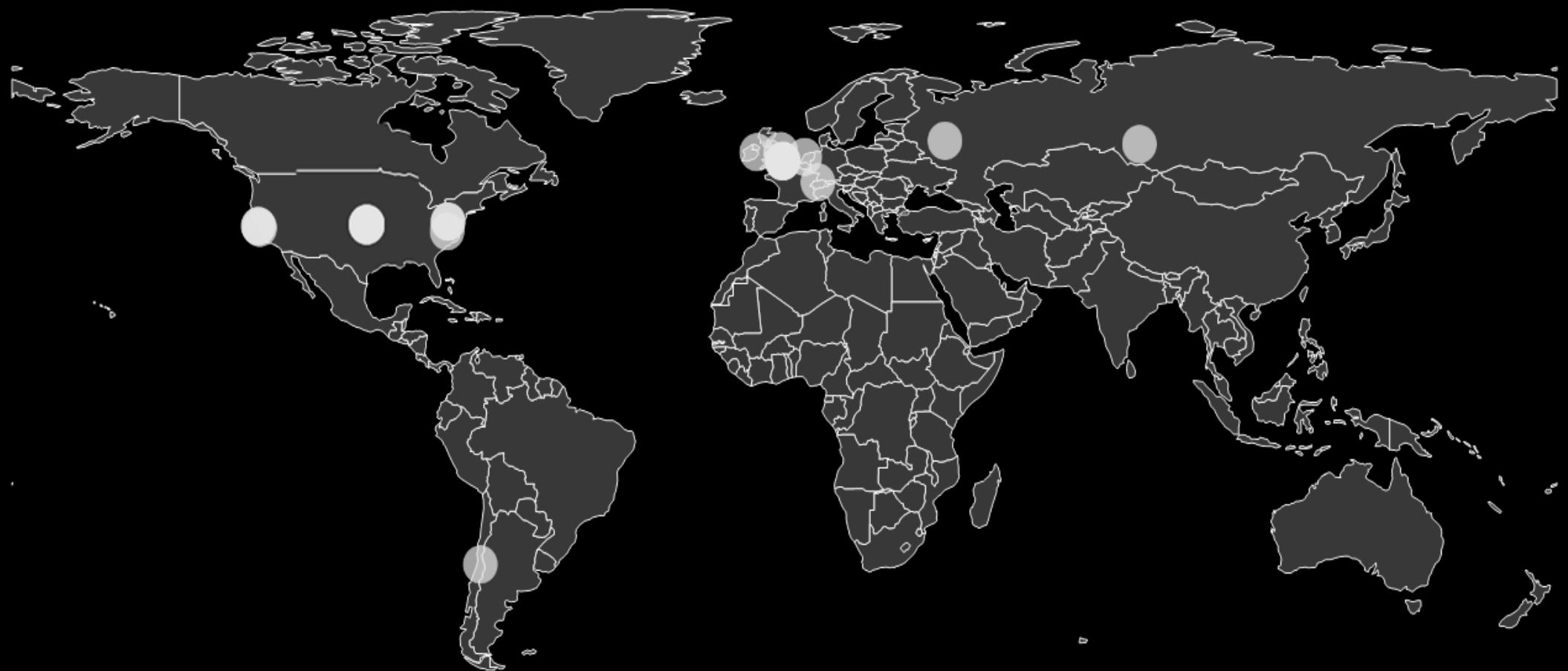


- For each A_i draw a circle centered at its known location with the estimated distance as radius and take the intersection as the area where B must be located



2016



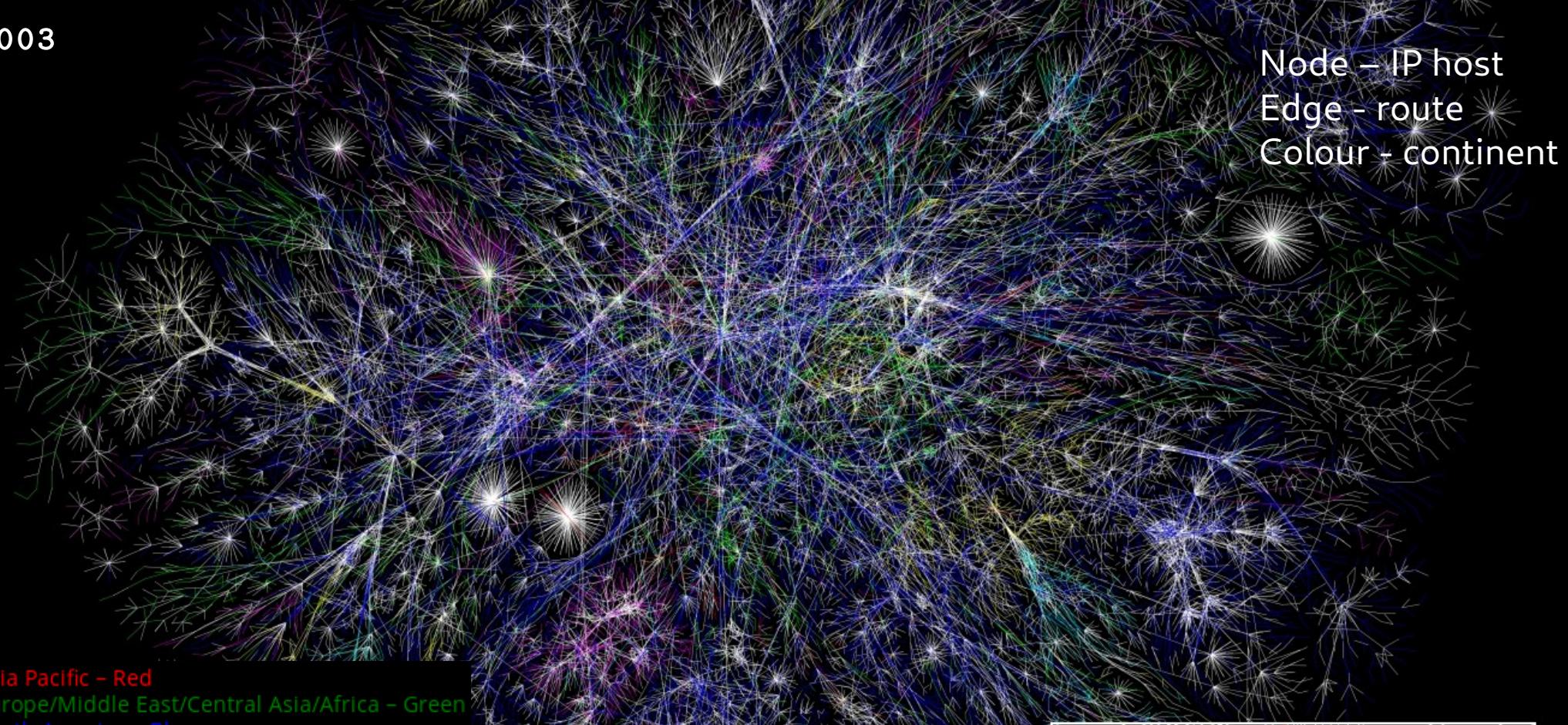


Tracerouting

- **traceroute** records all intermediate nodes in a packet's route from source to destination
- A packet may travel through at least an AS or ISP network to reach destination, i.e. each packet needs to go through the network's core
- AS path inference: IP addresses are allocated contiguously in blocks, and the AS routing tables are public, so a routing map of the core can be built from them
- ASs introduce routing hierarchy: connections can be inter-AS or intra-AS
- This hierarchy suggests a long-tailed degree distribution, with a connective core of fewer very high degree nodes, but the majority of nodes are at the network's periphery

2003

Node – IP host
Edge – route
Colour – continent



Asia Pacific – Red

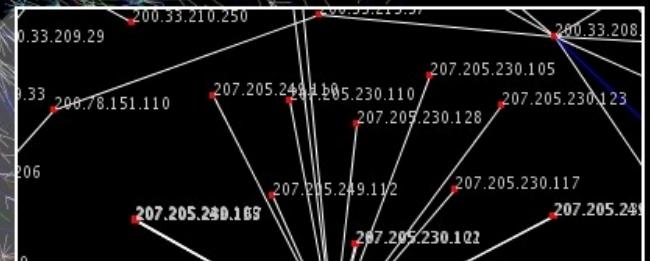
Europe/Middle East/Central Asia/Africa – Green

North America – Blue

Latin American and Caribbean – Yellow

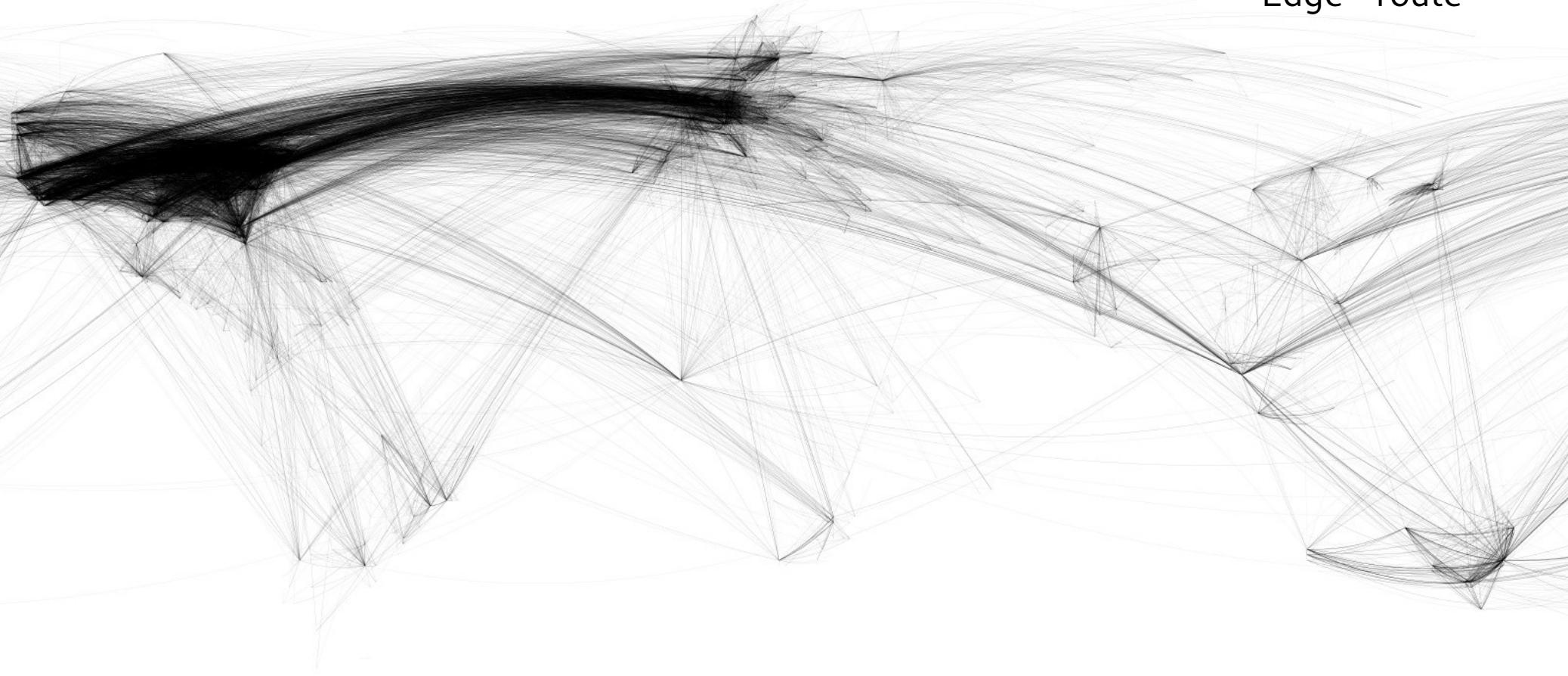
RFC1918 IP Addresses – Cyan

Unknown – White

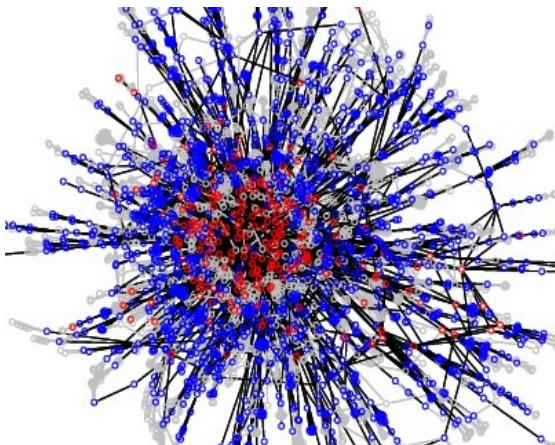


2002

Node – cities
Edge - route



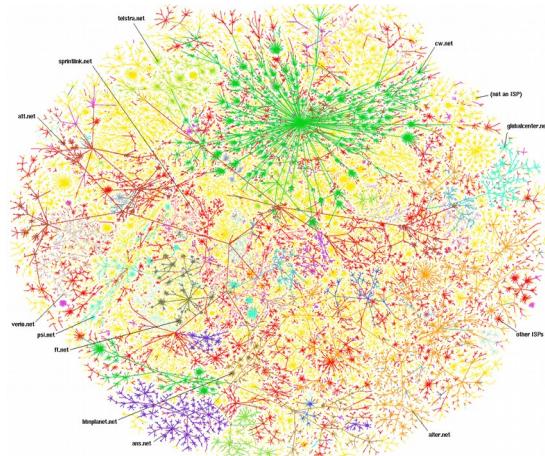
More maps



Mercator (1998)
Router Map

Govindan, R., Reddy, A.,
Information Sciences
Institute

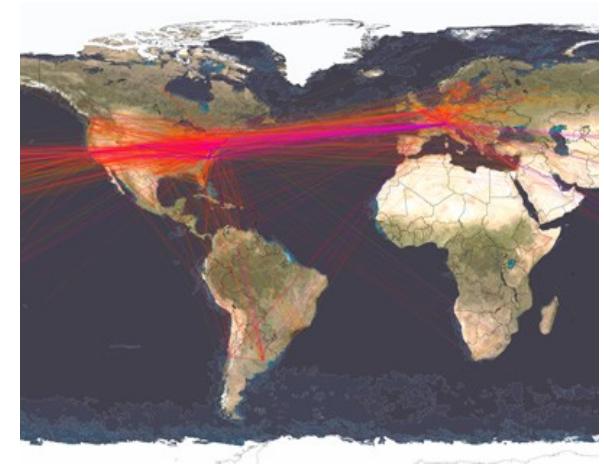
[link]



Internet Map (1998)
Map of major ISP networks

Cheswick, W., Bell Labs
Burch, H., Carnegie Mellon
University

[link]



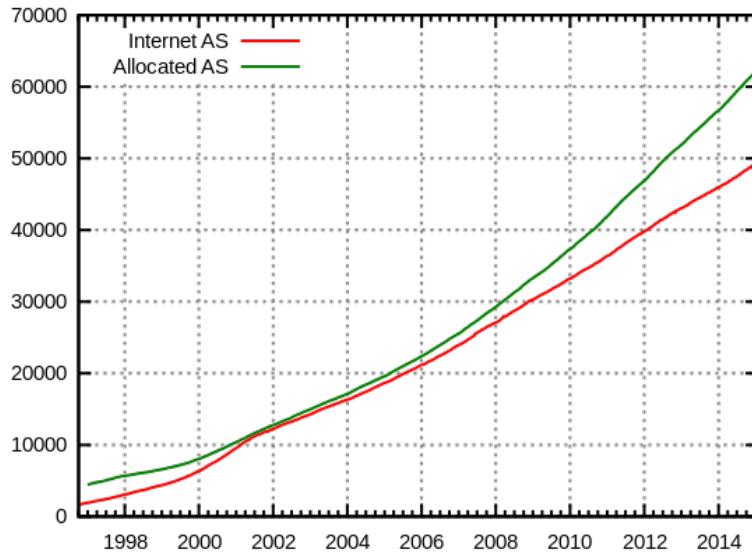
Rootzmap (2002)
AS geographical map

Bourcier, P.
Data from CAIDA

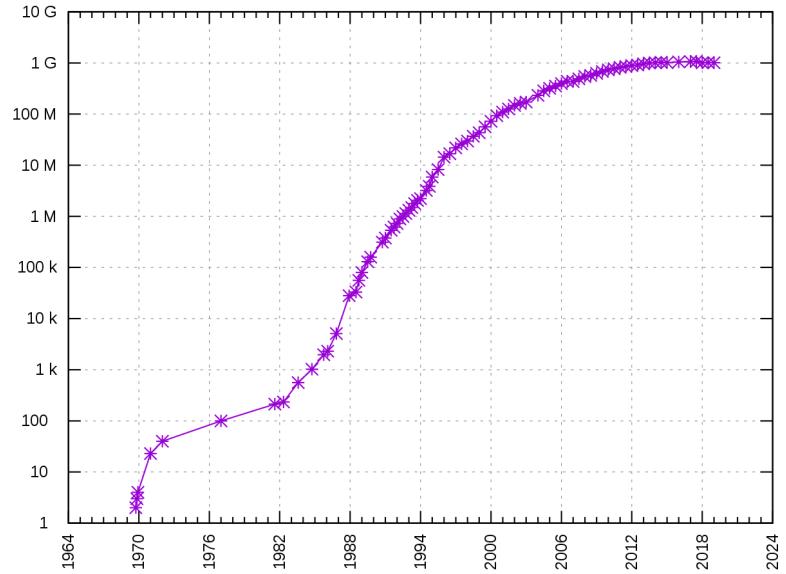
[link]

Temporal evolution

- Number of ASs, size of total routing table from all ASs grows linearly
- IPv4 addresses exhausted in 2011, so IP allocation has slowed down
- The core and the size of the routing table continue to grow unaffected



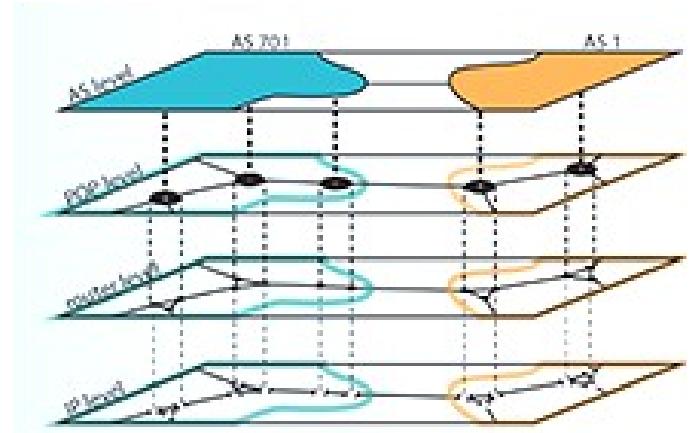
Allocation of AS numbers © Wikipedia



IP hosts on the Internet (log) © Internet Systems Consortium

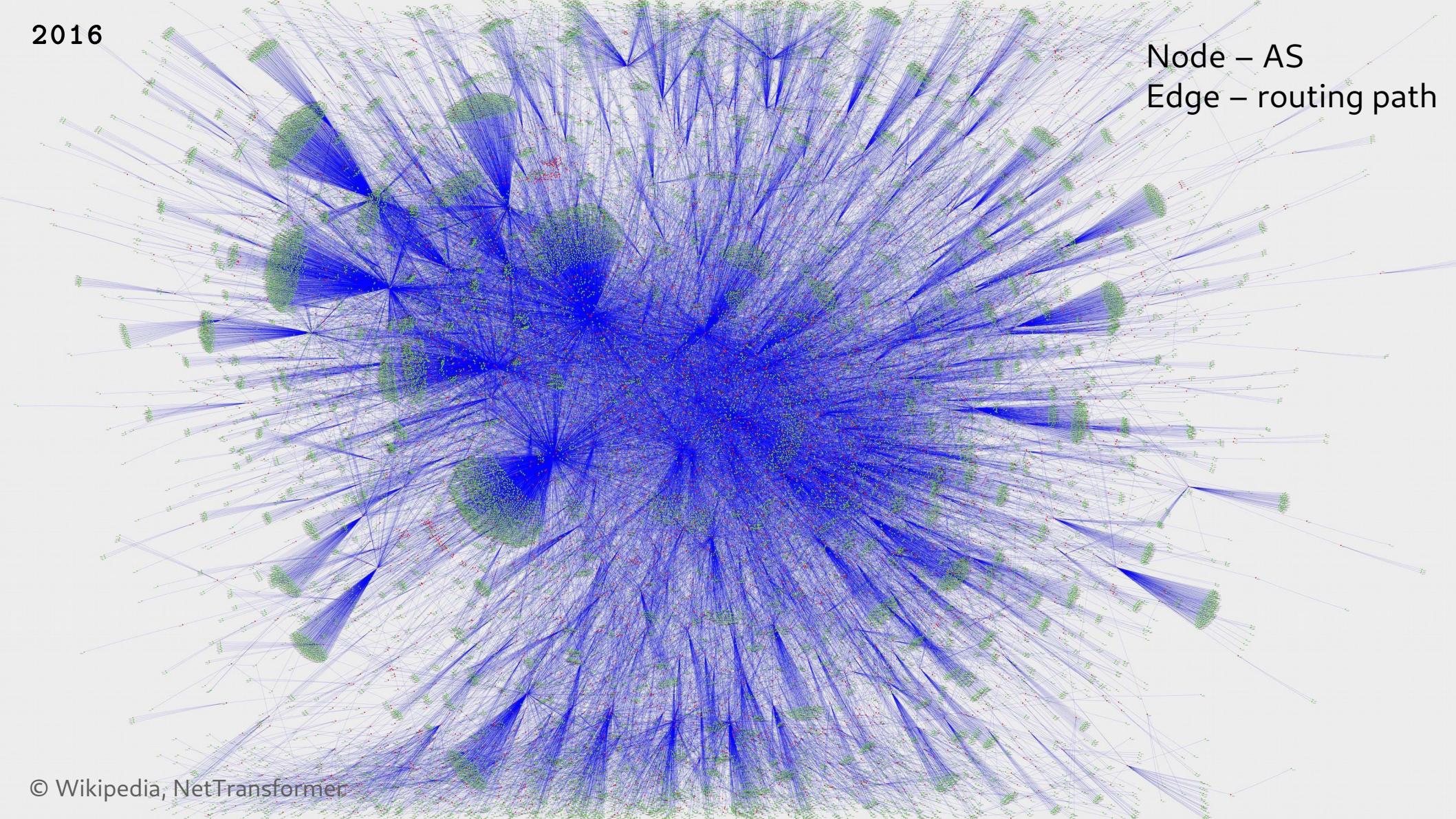
Mapping the core

- Center for Applied Internet Data Analysis (CAIDA) combines above techniques to create a yearly map of the ASs
- 4-layer hierarchy: IP level, router level, Points of Presence level (routers with known geographical location), AS level
- Concerned with internetwork topology analysis
- Further hierarchies: multiple tiers of AS
- An AS's customer cone represents all the ASs downstream from it, i.e. that directly or indirectly pay the AS to connect to the internet

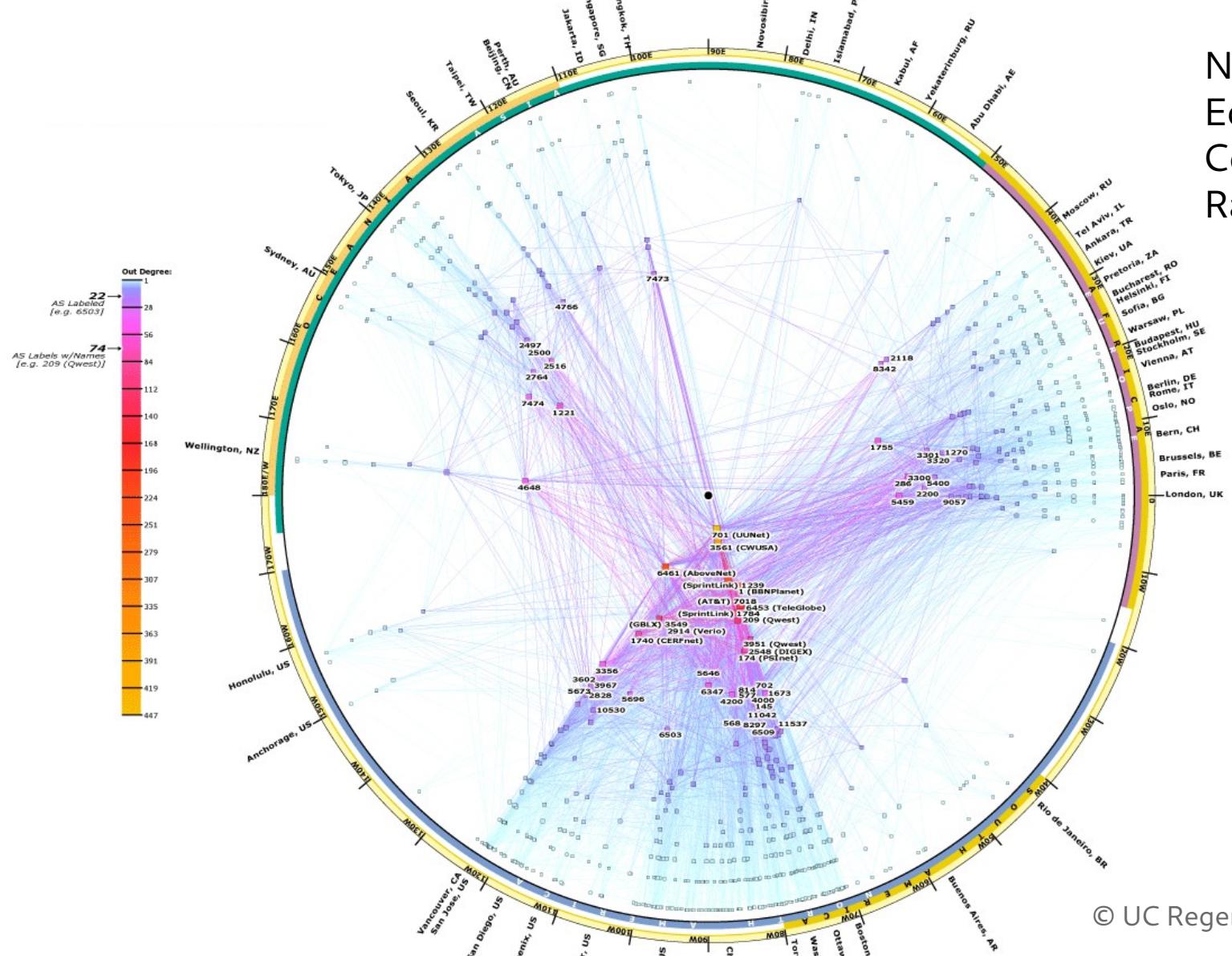


2016

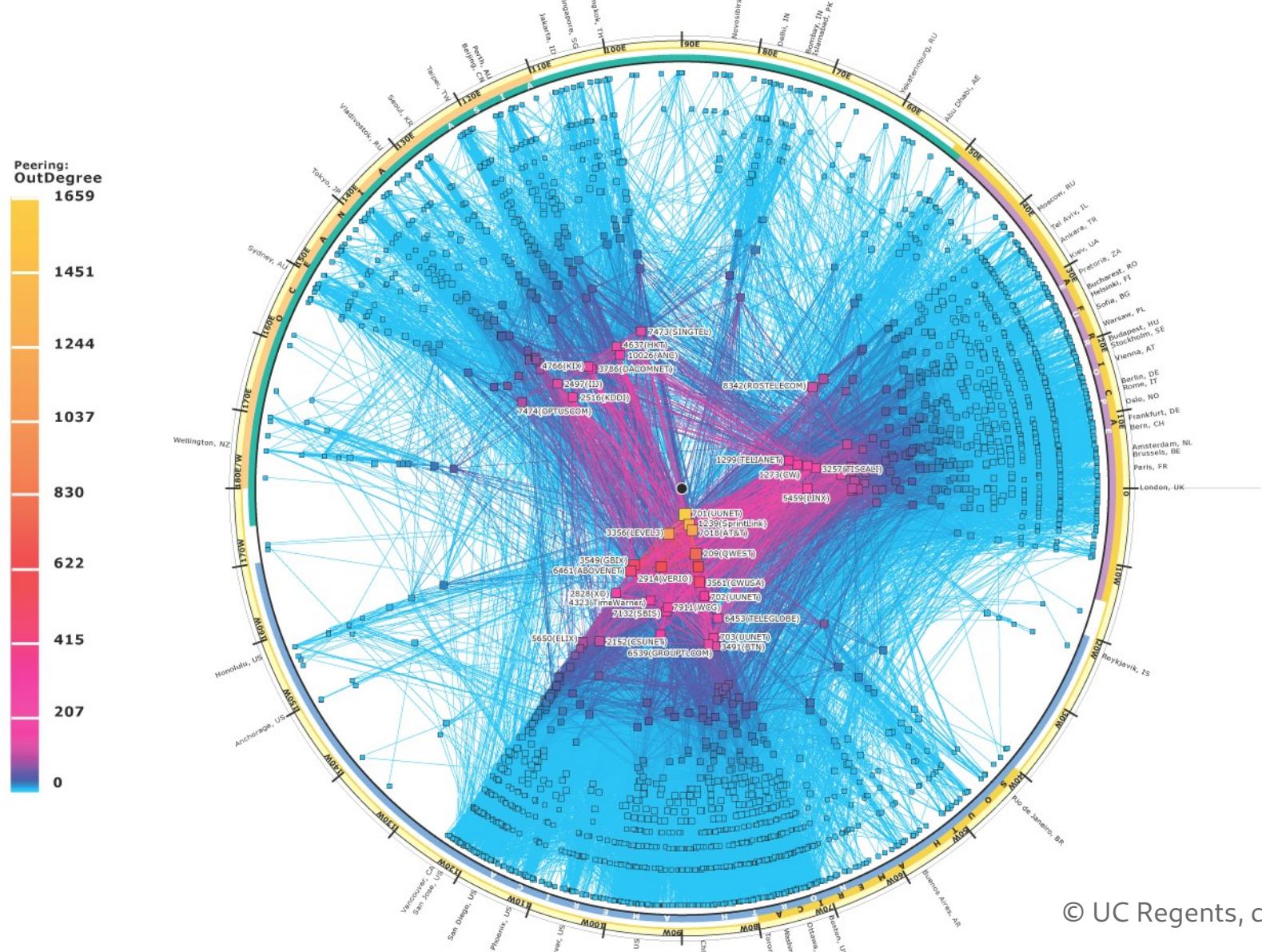
Node – AS
Edge – routing path



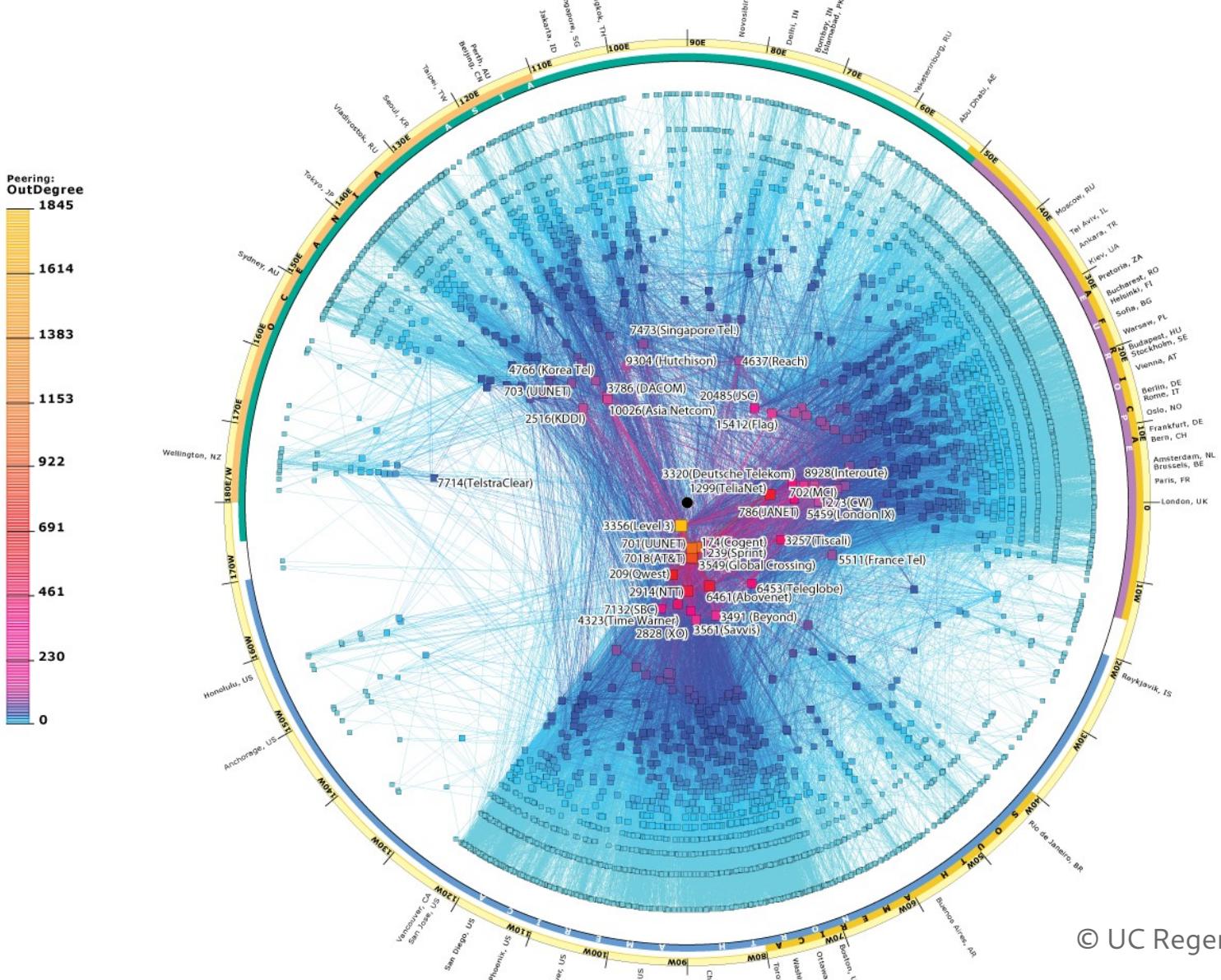
2000



2005



2008



2012

Neighbors (degree)

4132 861

3718 774

3305 688

2892 602

2479 516

2066 430

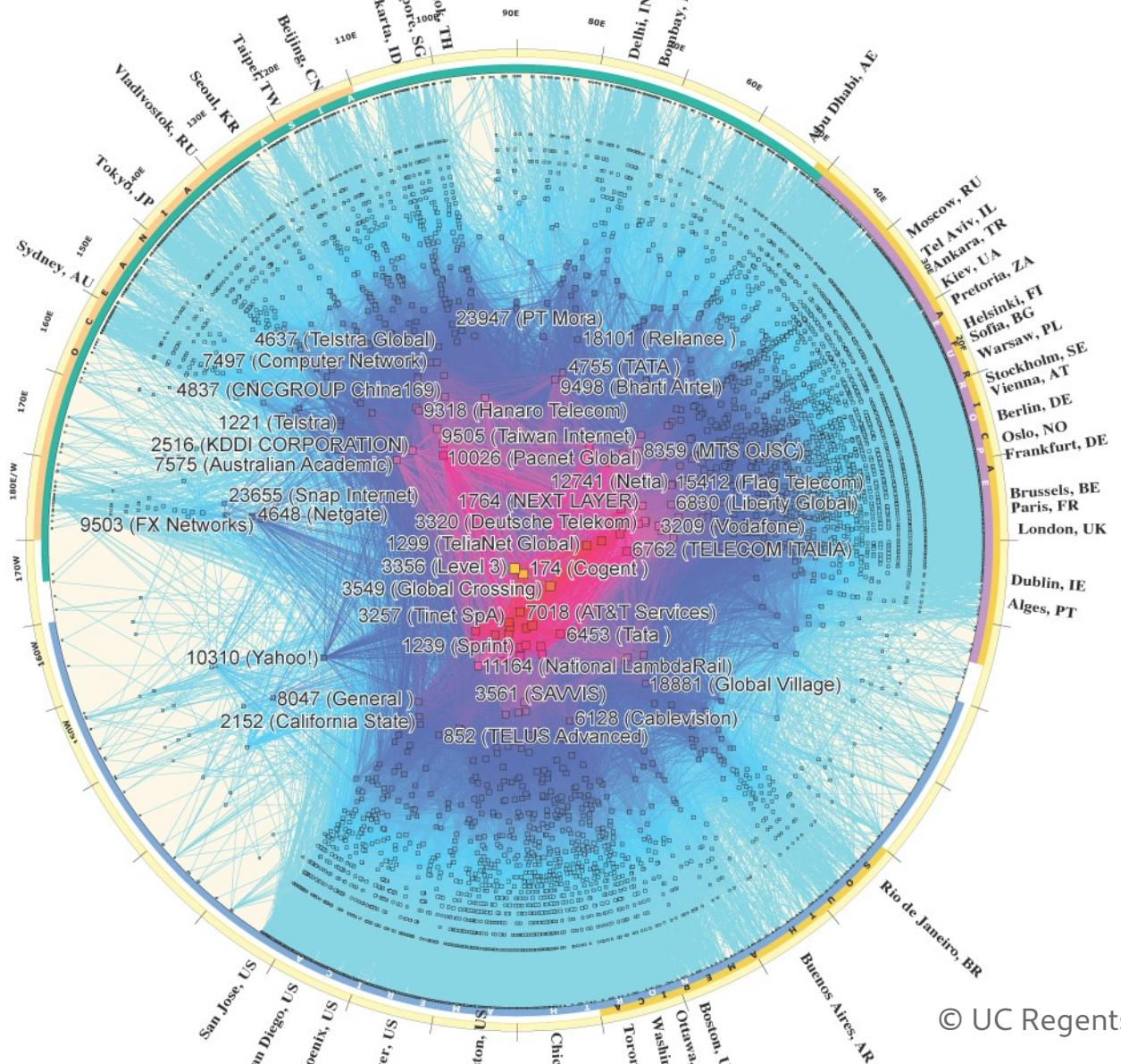
1652 344

1239 258

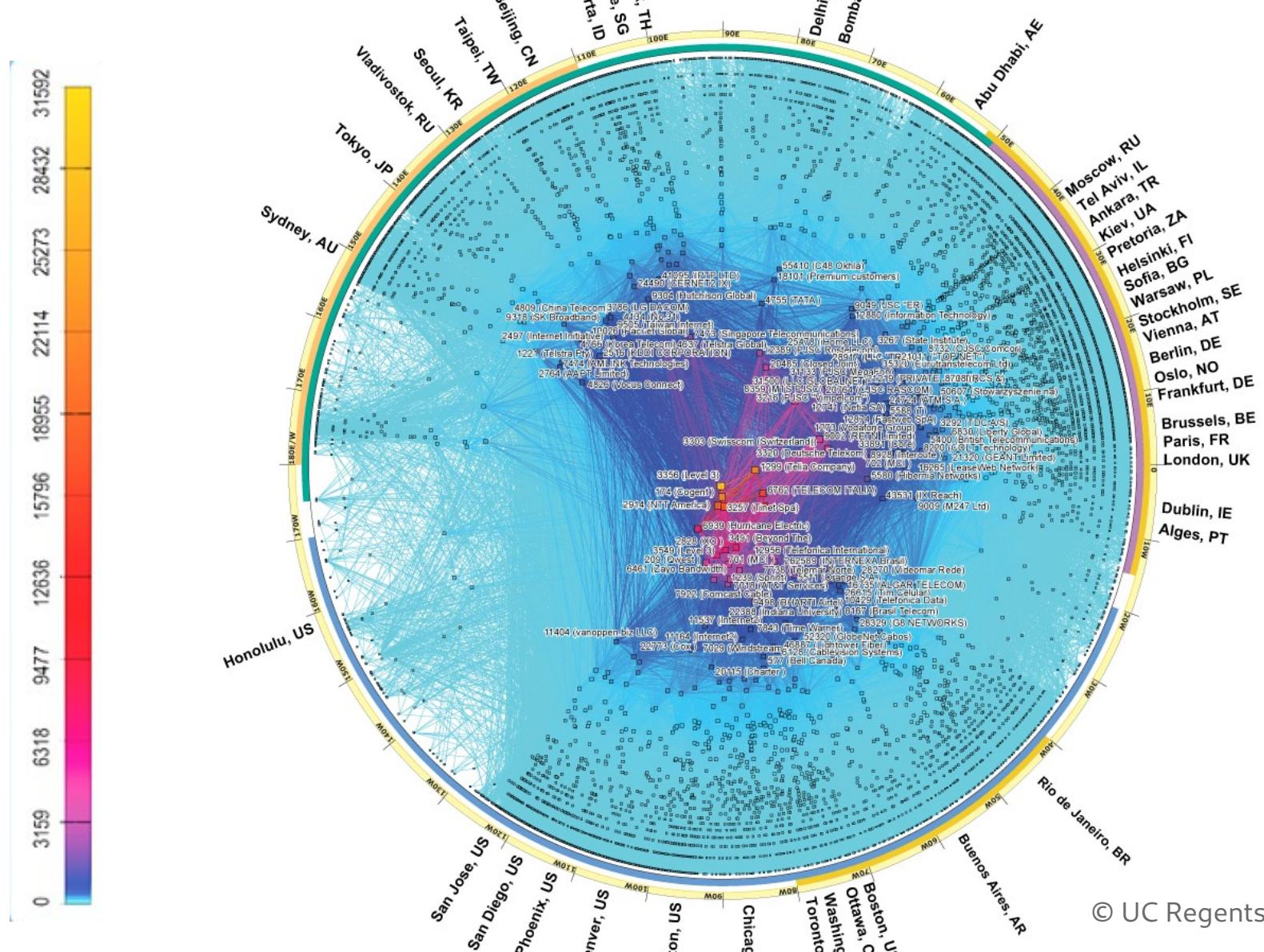
826 172

413 86

0 0

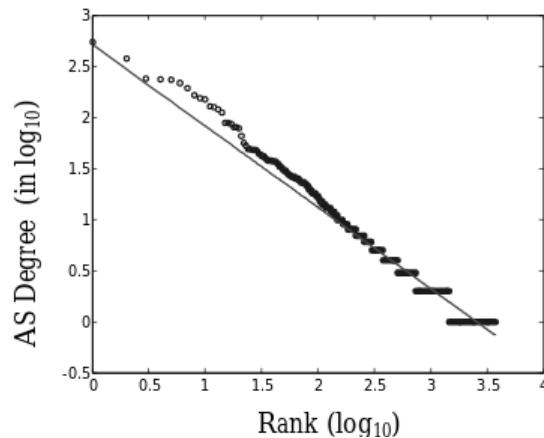
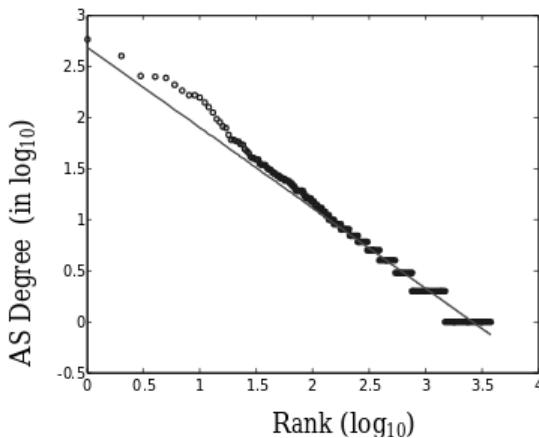


2017

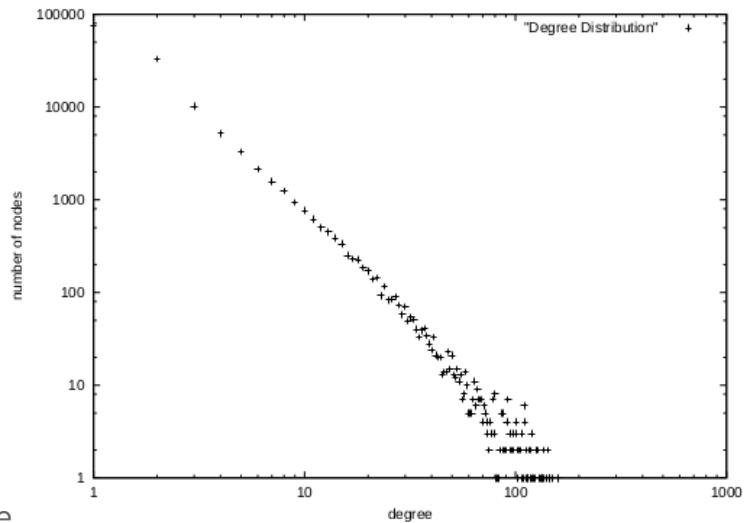


Network properties

- The network layer has asymmetric path length
- Degree distribution of routing network and degree distribution of ASs appears to be a power law?



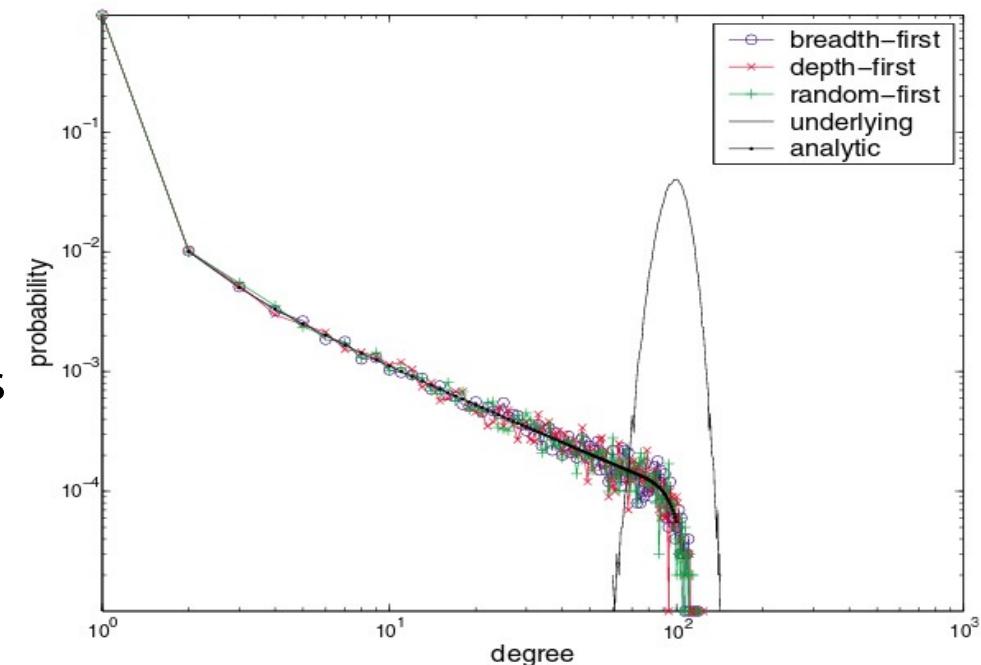
Degree distributions of a traceroute map of the Internet versus a map of the routing tables embedded in ASs (Border Gateway Protocol tables) © Amini et al



Mercator Router map © Govindan et al

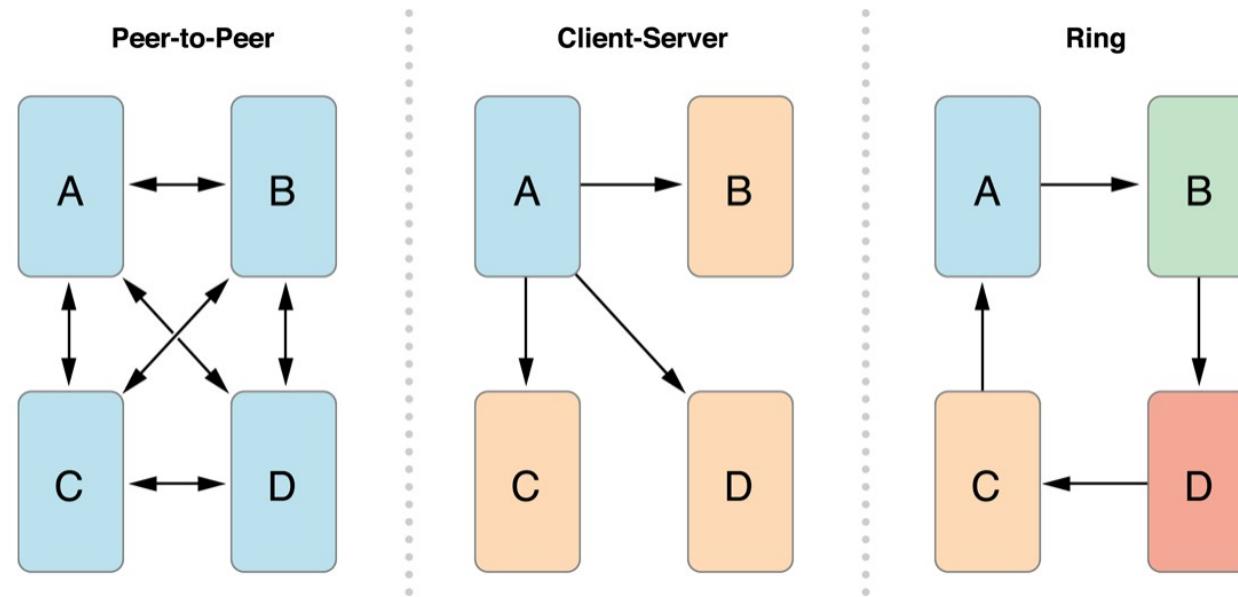
Conundrums

- These models don't use the whole internet, but sample it
- Most models generated from single-source, all-destinations, shortest-path trees. These trees only sample a fraction of the network's edges
- Lakhina et al: traceroute sampling is inherently biased, the internet maps do not reflect actual topology
- Clauset et al: spanning tree sampling of a large random network always manifests power law-like distributions regardless of spanning tree algorithm used



The Application layer

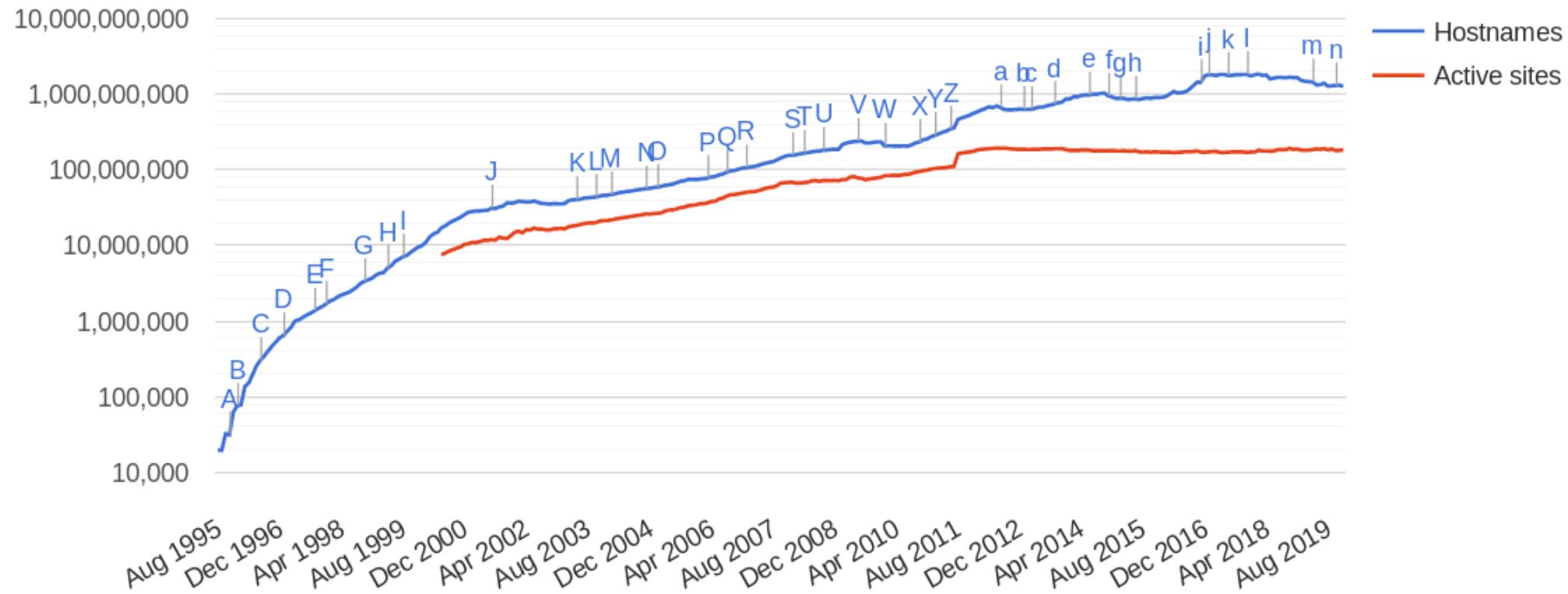
- User interaction happens at the application layer, and complex communities have emerged around protocols or platforms
- Various network topologies arise at this layer, depending on the nature of protocols



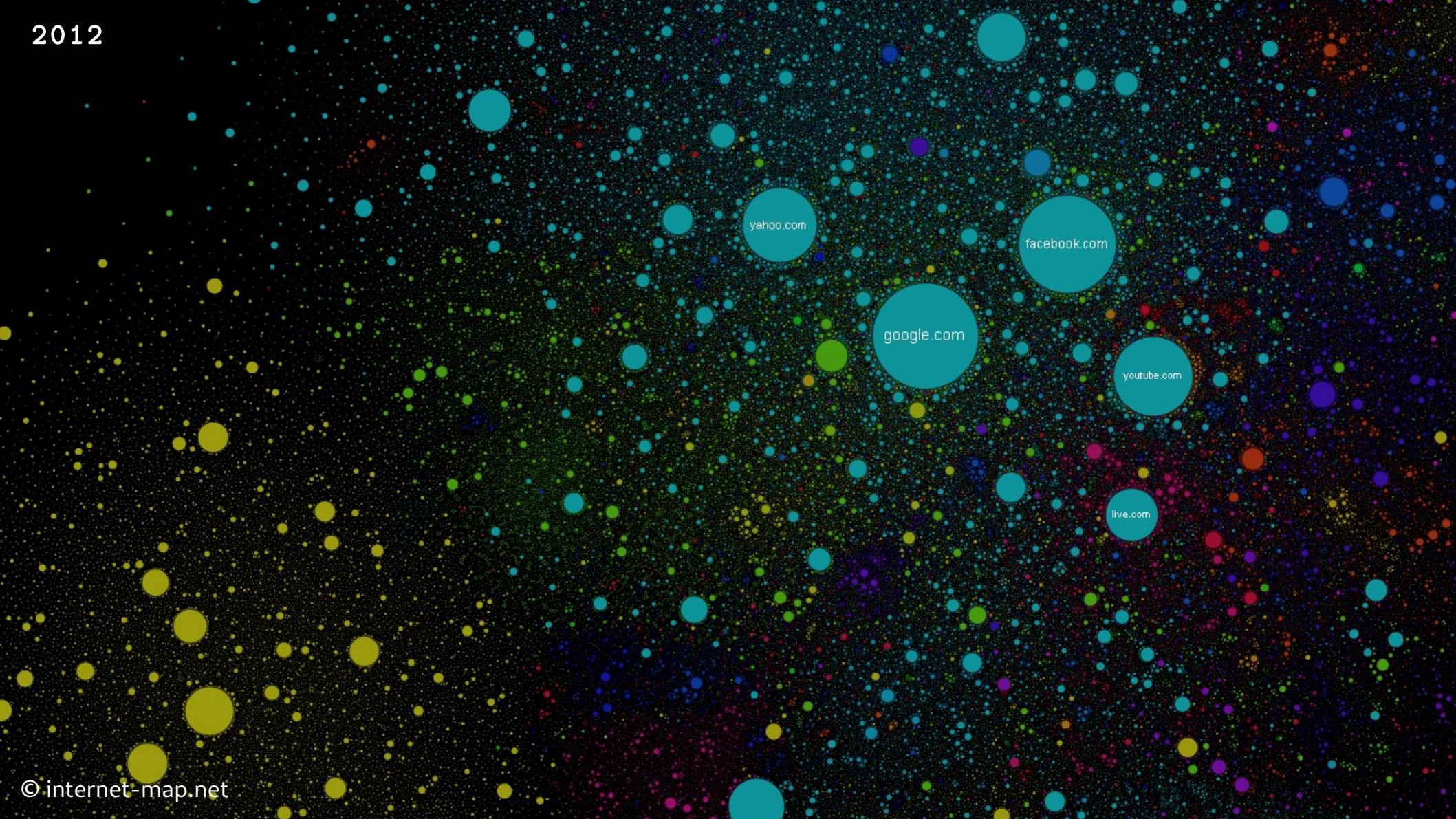
Mapping the World Wide Web

- The World Wide Web is not a layer of the internet, but an information system that exists within the application layer
- Nodes are web resources, identified by a universal resource link (URL), and edges are hyperlinks as per HTTP standards
- Hierarchical: a collection of interlinked resources with a common theme and domain is a website, a collection of websites or web profiles with the same domain can be contained in a content creation platform or social network
- Understanding the topology of the Web allows improving search engines (Google's PageRank uses node in-degree to rank all pages in the visible web that use the search keyword)

Total number of websites (logarithmic scale)

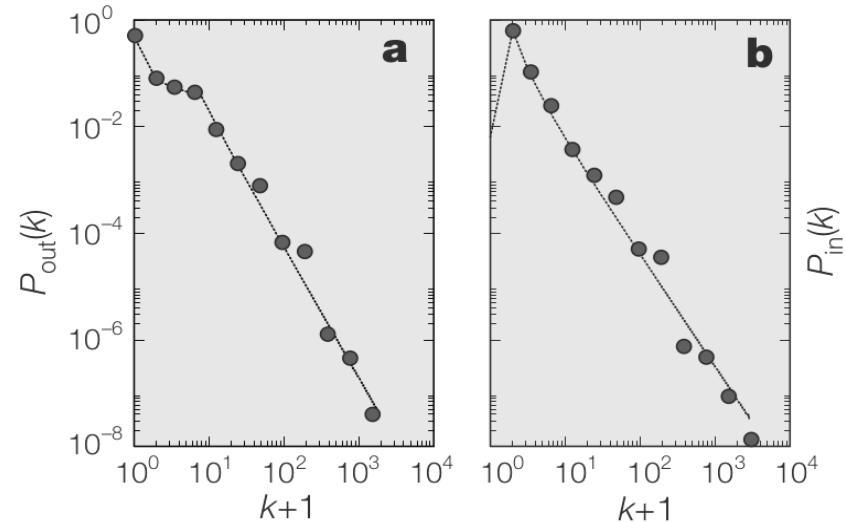


2012



Scale-free hypertext network?

- 1999, degree distributions of hyperlinks on the nd.edu domain are power laws of exponent -2.5 (out) and -2.1 (in)
- Average shortest path size grows linearly as function of system size, claims network is small-world
- But the paper only analysed 300K pages and 1.5M links on a single domain
- Internet at the time was 7 million host-names and at least 100M pages
- Later research agrees on exponents around -2.3, but only for nodes of high degree (>1000) – power law with cutout or inverse exponential?

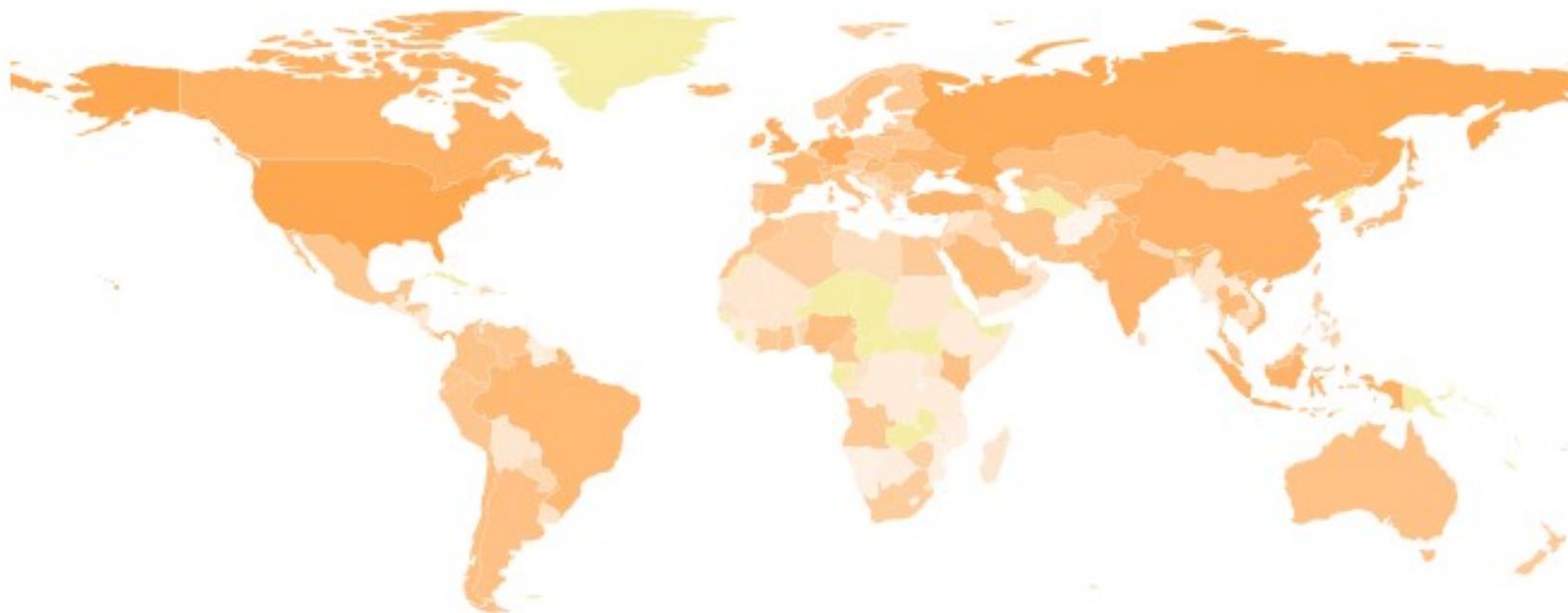


Networks of fraud

- More networks emerge at the application layer from fraudulent use
- Large-scale attacks: specific ISPs and domain providers are more resilient to content takedowns. Fraudsters aggregate around these platforms to rent servers or register hundreds of typo-squatting domains
- Phishing economy: a small number of developers create a phishing site (phishing kit), which is deployed by novice users. Hierarchical identity theft?
- Worm spreading: malicious software copies itself to new hosts by randomly connecting to other hosts, or via the local network, or via USB/file share. Epidemic model?
- Botnets: some worms “recruit” infected hosts by placing a backdoor. This allows the hacker to remote control the infected host for e.g. denial of service attacks

2020

Cyberattack map

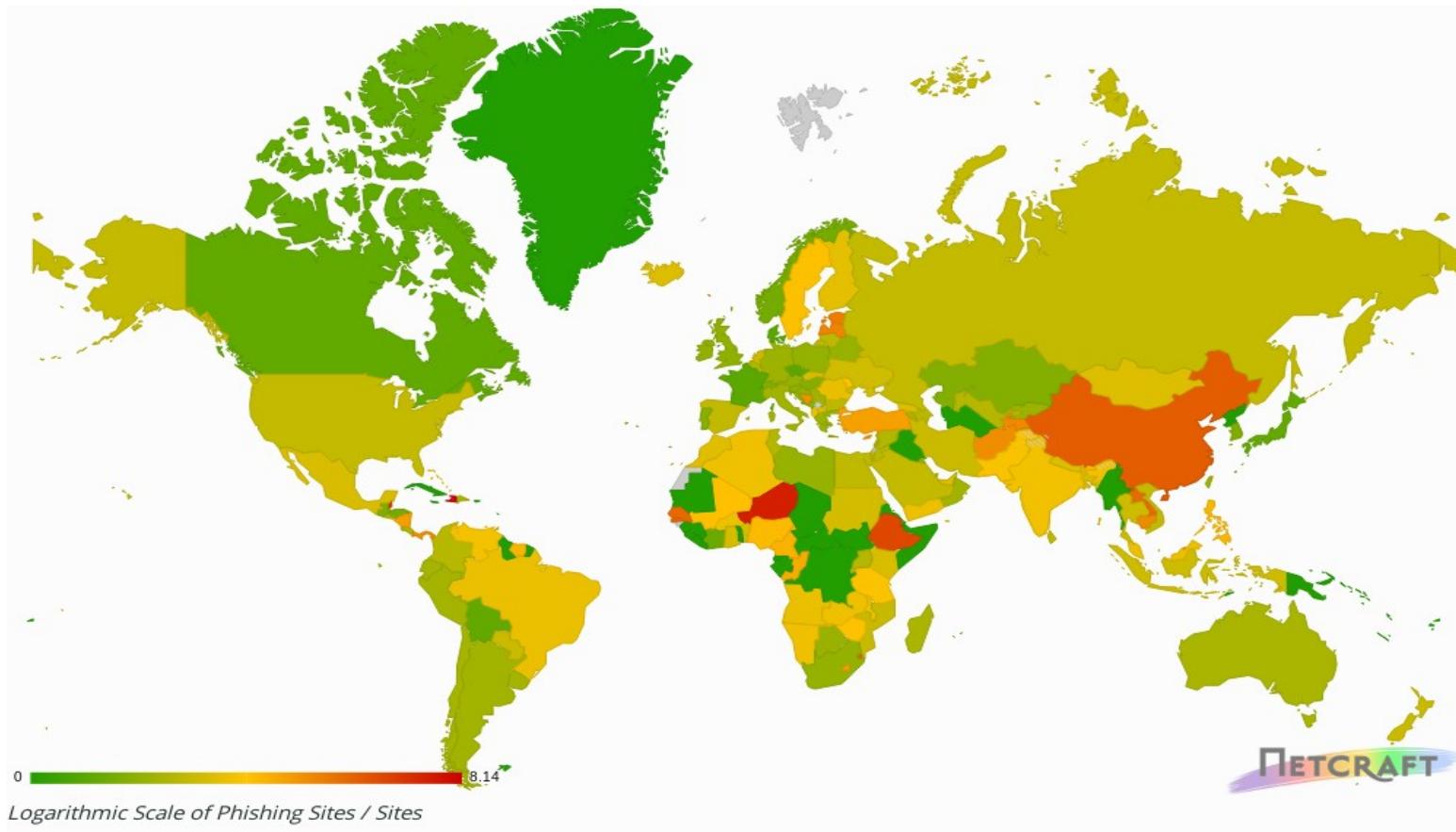


Attacks per Hour



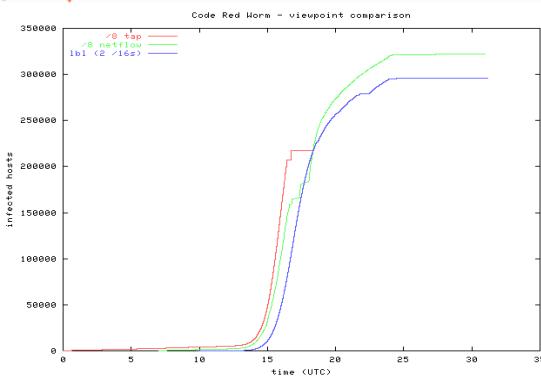
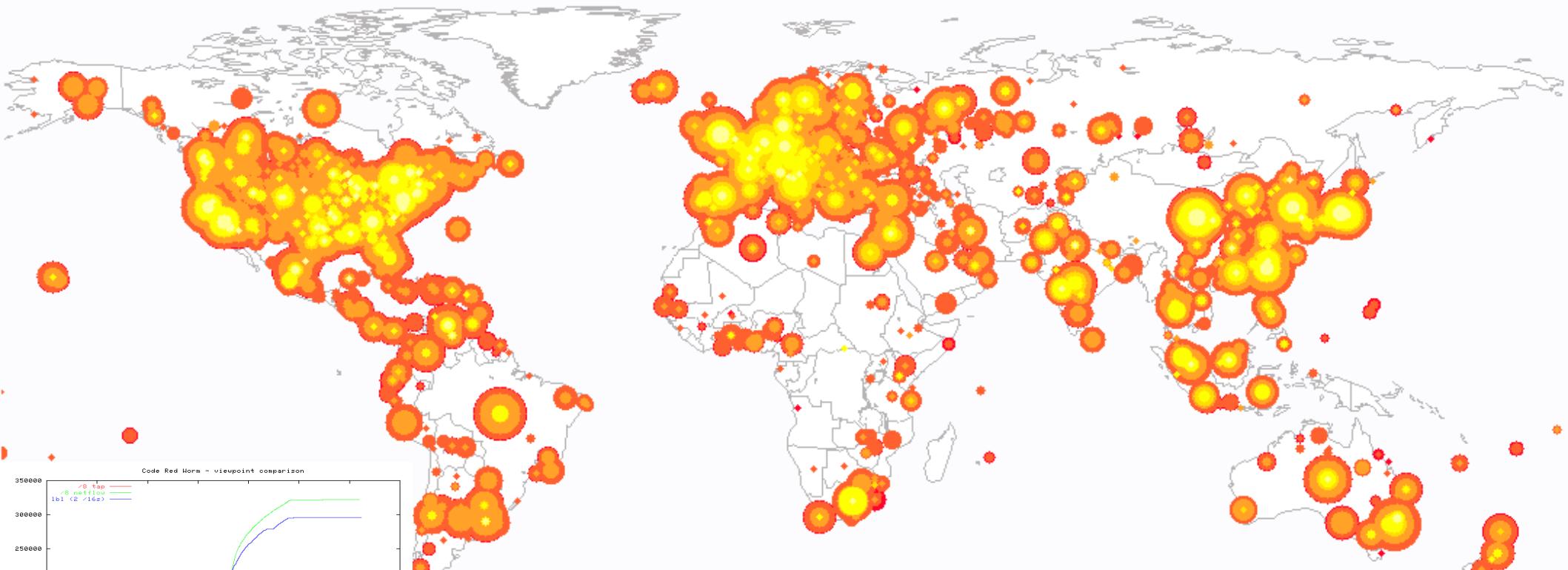
2020

Phishing map



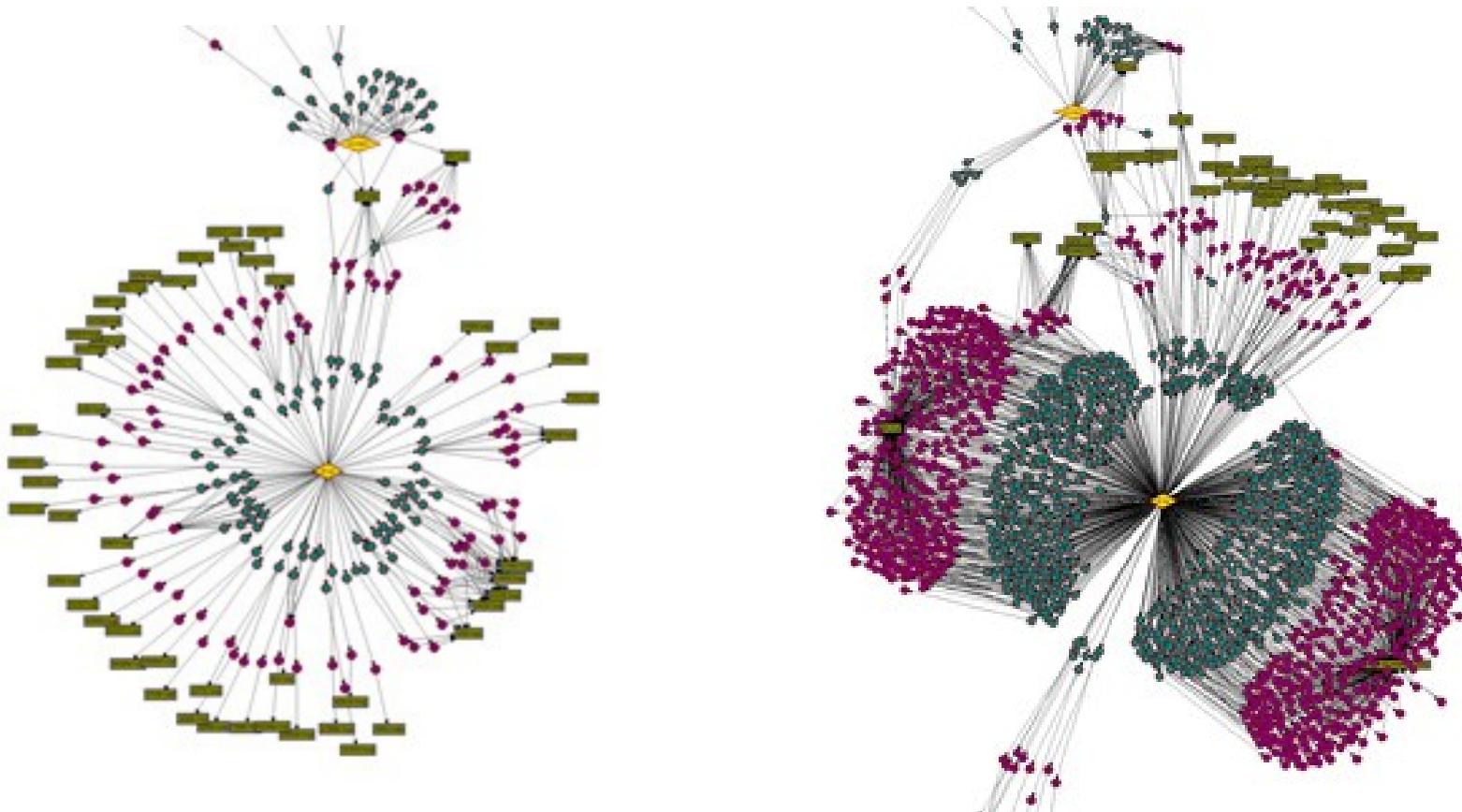
2001

Worm infection map



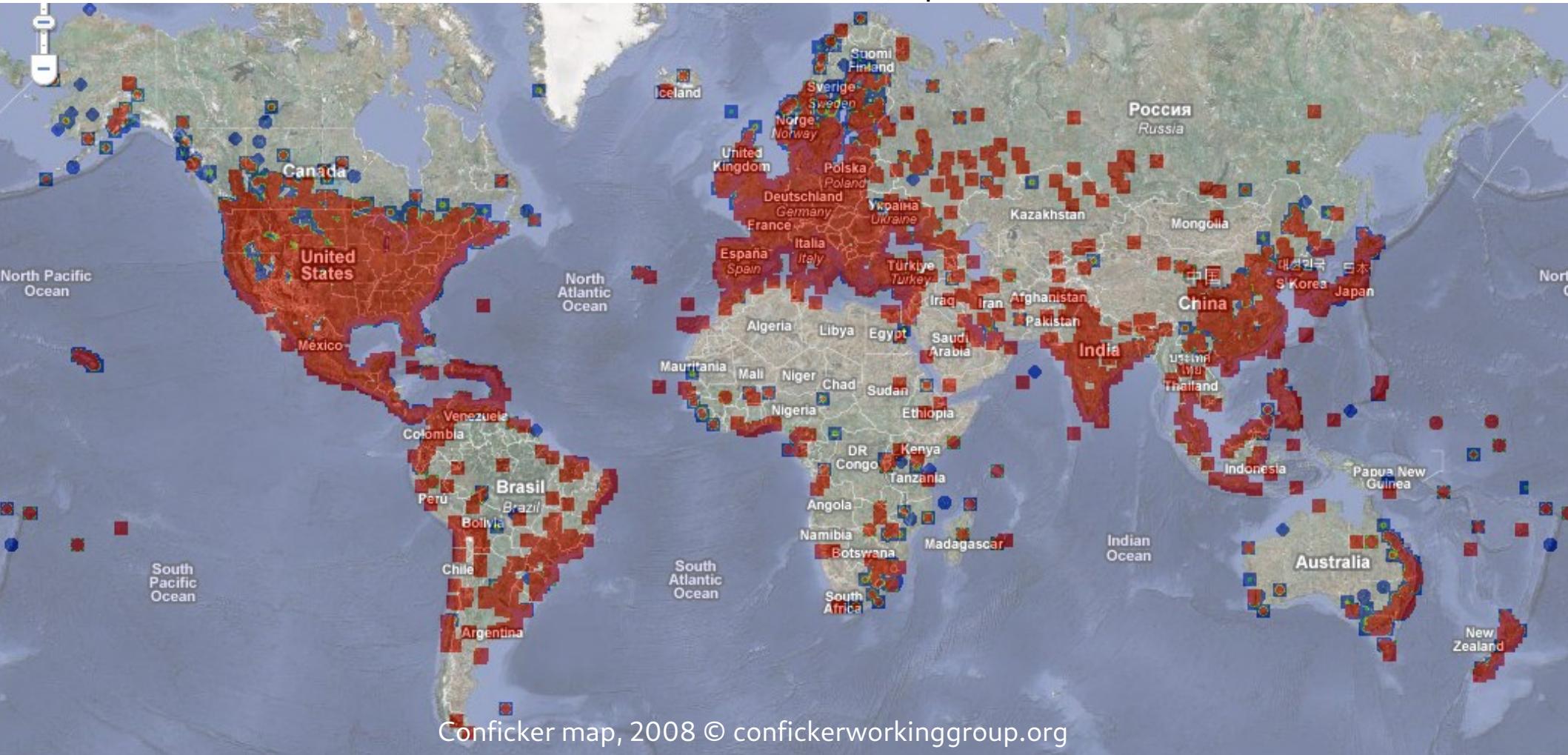
spreading of the Code Red worm, 2001 © caida.org

Botnet dynamics



SQL-Slammer worm spreading in honeypot network, 2003 © Ed Blanchfield

Botnet map

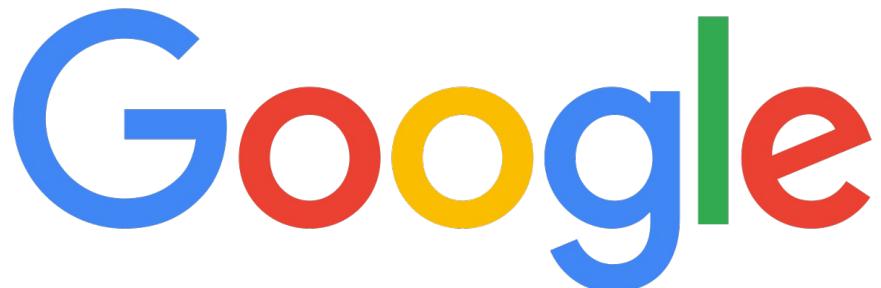


Networks of fraud

- Analysing cybercrime behaviour can provide insights about the underlying security of countries and segments of the internet backbone
- UK: hosting 5.4% of global phishing attacks in 2016. Today less than 2%
- Botnets: all bots connect to a small number of servers to download updates or instructions (Command & Control). Analysing the botnet's network dynamics helps detecting these C&C servers and taking them down.
- Domain switches: some worms disable themselves as soon as e.g. a domain is registered – WannaCry, 2019
- Critical update patches are usually released quickly after the first infections, but patch adoption rate is extremely slow
- Serious political consequences: Russian web brigades, John Podesta hack, but also freedom protests in totalitarian countries

Ownership

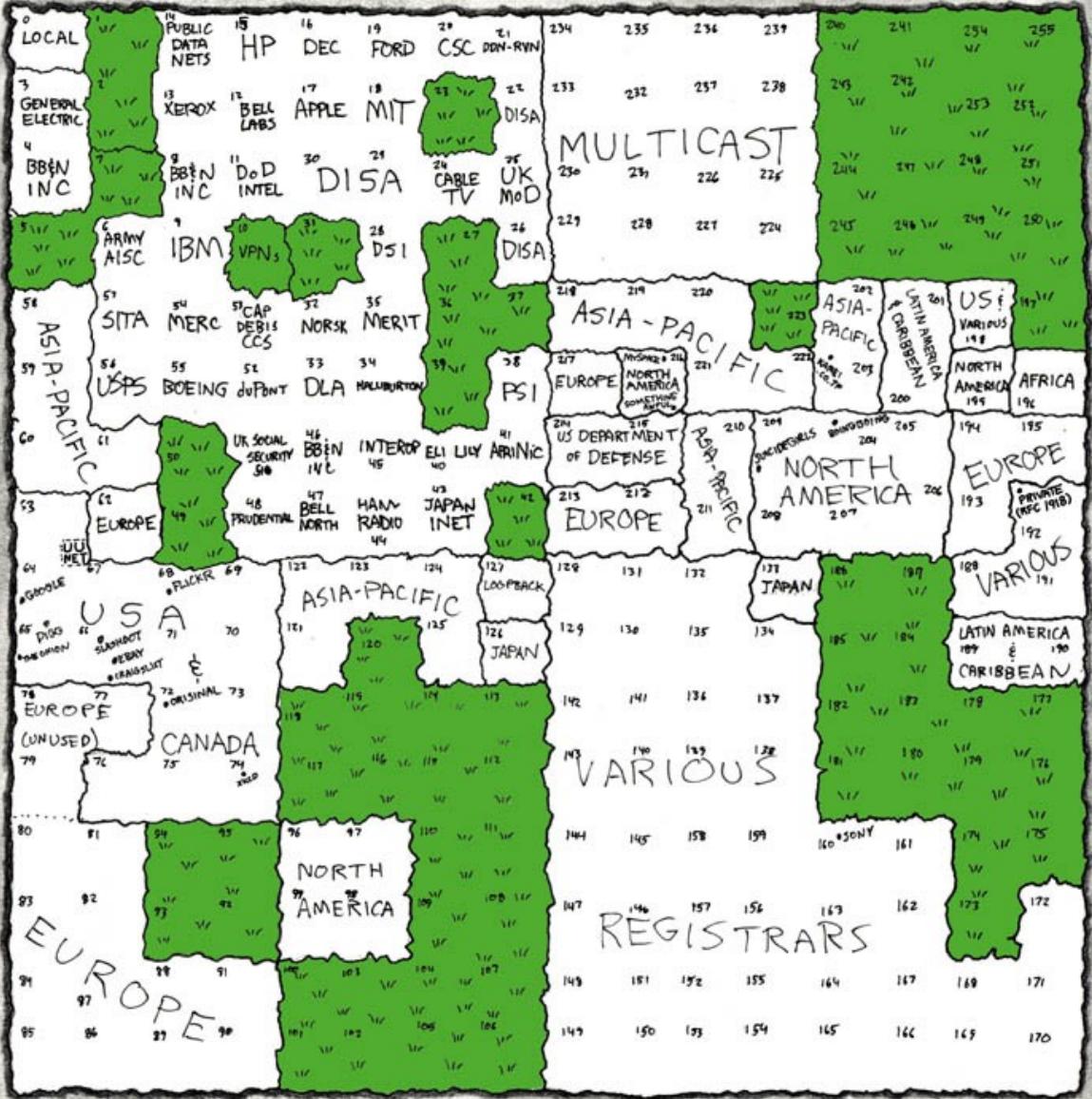
Who owns the internet?



Who owns the internet?

- Upstream ISPs manage most of the physical infrastructure, cables and routers, but also ASs
- Most IXPs managed by non-profit organizations, as they must be exchange points between different networks owned by different ISPs
- Regional Internet Registries (RIRs) manage the network layer and distribute IP addresses to new ASs as infrastructure extends
- Before the RIRs, almost a quarter of IP addresses pre-sold to governments or corporations
- Internet Registrars: sell IP addresses to the public

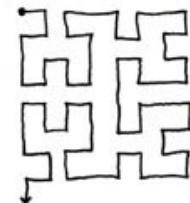


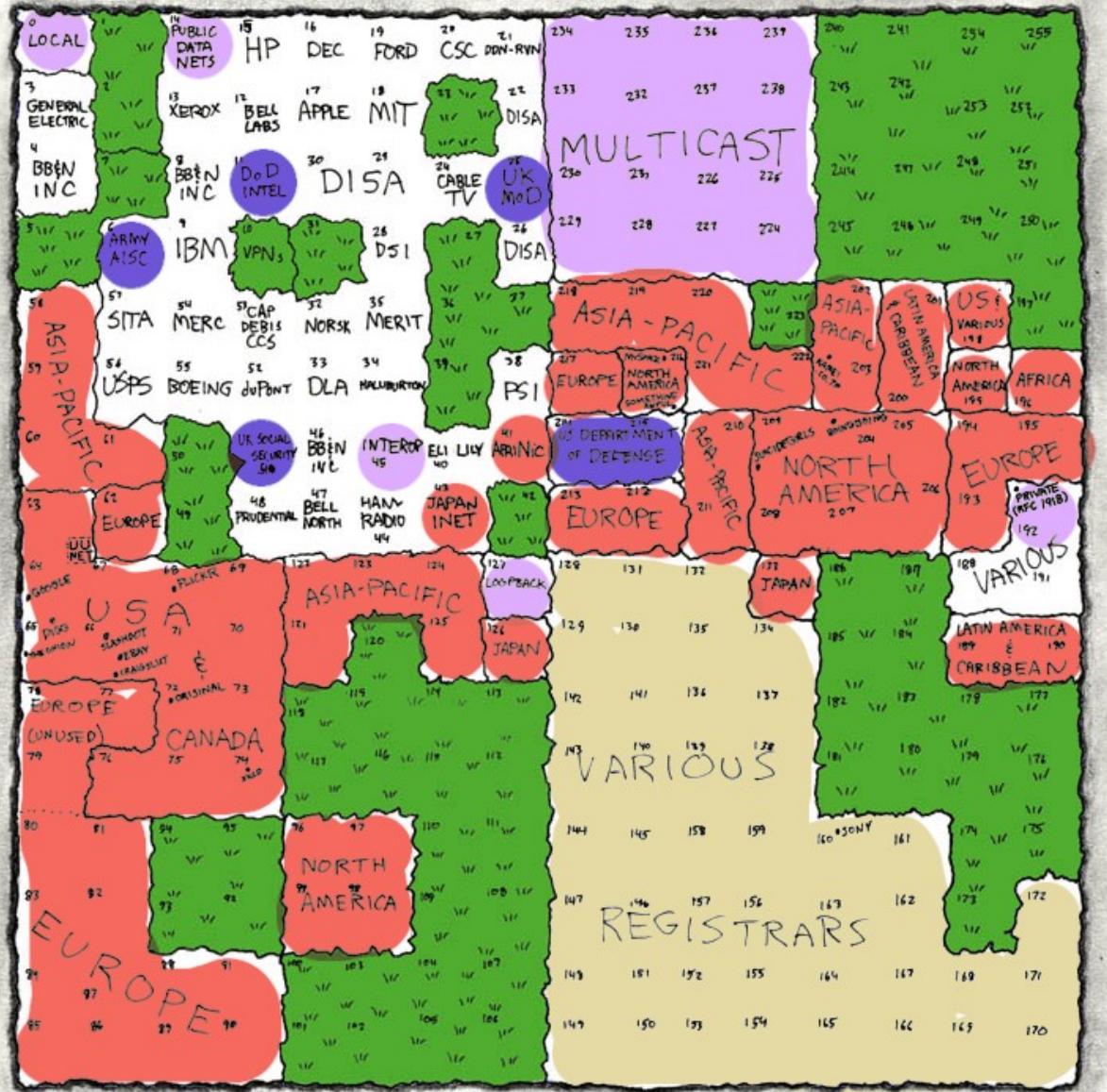


Node – Class A segment
Colour – allocated/
unallocated

Hillbert Curve Map

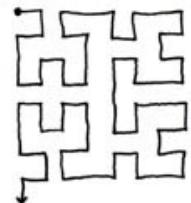
0 1 14 15 16 19 →
3 2 13 12 17 18
4 7 8 11
5 6 9 10





- Public/network use
- Registrable
- RIR
- Unallocated (in 2006)
- Military
- Corporation

0 1 14 15 16 19 →
3 2 13 12 17 18
4 7 8 11
5 6 9 10



Who owns the internet?

- The Internet Society: open, nonprofit, develops Internet standards
- The Internet Engineering Task Force (IETF): open, non-profit, a group of committees and working groups that maintain the Internet's architecture and stability.
- The Internet Corporation for Assigned Names and Numbers (ICANN): private nonprofit corporation, manages DNS. Makes sure every domain name links to the correct IP address. Not controlled by government.
- Domain registrars: provide domain names to the public. Governmental or corporate.
- Net neutrality: no protocol or content should be given priority in transmission

Who owns the internet?

- Google: 90% of search advertising, YouTube: 60% of all streaming-audio business but pay for only 11% of the total streaming-audio revenues artists receive, >50% of websites on the Internet use Google Analytics
- Facebook: 80% of mobile social traffic, claims IP of user content
- Amazon: 75% per cent of e-book sales
- GoDaddy+Amazon+Google: host ~40% of all websites
- Net neutrality?
- Difference between piracy and Google Books?
- No more self-hosted, or even self-built websites – instead, profiles on a centralised platform

Surveillance

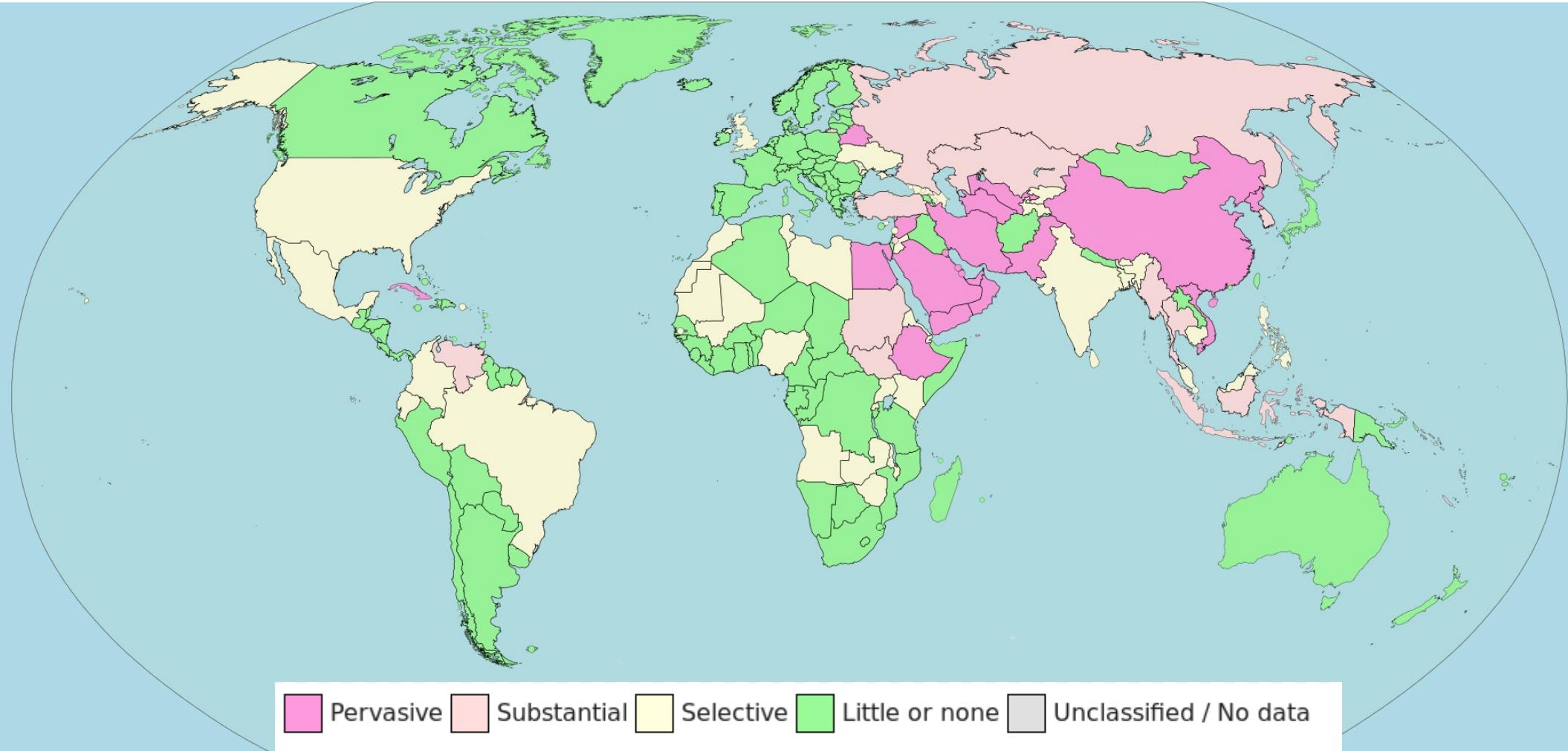
- Government surveillance of ISP activity (e.g. German govware, UK ISPs to store 1 year worth of internet activity etc.)
- Backdoors in communication systems by largest software and telecom providers
- Bad security of protocols and software allows interception and data theft
- Five Eyes, Six eyes, Nine eyes, Fourteen eyes: global superpowers share intelligence gathered from mass surveillance
- Bonus: tech companies selling users' data for advertisement purposes



Censorship

- Governmental censorship: Great Firewall of China, Turkey bans Twitter, Iran, Egypt
- Cybercrime and counter-terrorism laws used to crack down on assembly and expression online in Middle Eastern countries
- Censorship on social platforms: community law twisted for political interest
- Google favours certain search results regardless of in-degree rank
- Tik-Tok blocks videos about alternative communities, disabled people, even human rights

Censorship map



[Legend] Pervasive [Legend] Substantial [Legend] Selective [Legend] Little or none [Legend] Unclassified / No data

Wrap-up

How do we better map the internet?

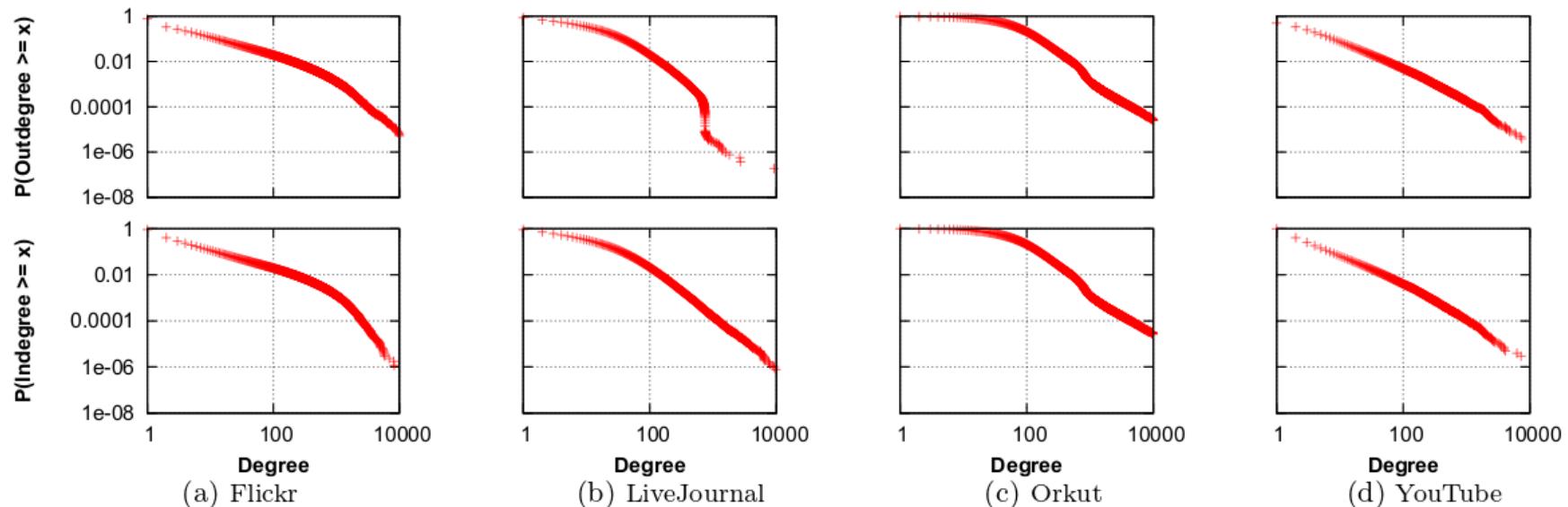
- Aggregate bigger, more consistent datasets, but we must keep in mind privacy
- Demand more transparency from ASs and ISPs
- Be aware of the inherent sampling bias of shortest-path routing algorithms when trying find a representative subset of the network layer
- Consider different distributions for low-degree and high-degree nodes, for example
- Correlate with population data, known infrastructure and geography
- Probing methods are inaccurate when there are not enough known hosts, so a distributed network for collecting and computing such statistics?

Other things I wanted to talk about

- Better systems at the application layer: Freenet, Tor, Interplanetary File System, Packet Radio, blockchains, BitTorrent
- Social networks: are they small world? Scale free?

Other things I wanted to talk about

- Better systems at the application layer: Freenet, Tor, Interplanetary File System, Packet Radio, blockchains, BitTorrent
- Social networks: are they small world? Scale free?



Conclusion

**If you torture your data long enough,
it's going to tell you exactly what you want to hear.**

Thank you!

FIGHT FOR A FREE INTERNET



truth