

Network Flow Analysis

Mark Meiss

Presentation for NaN-Group

October 4, 2004

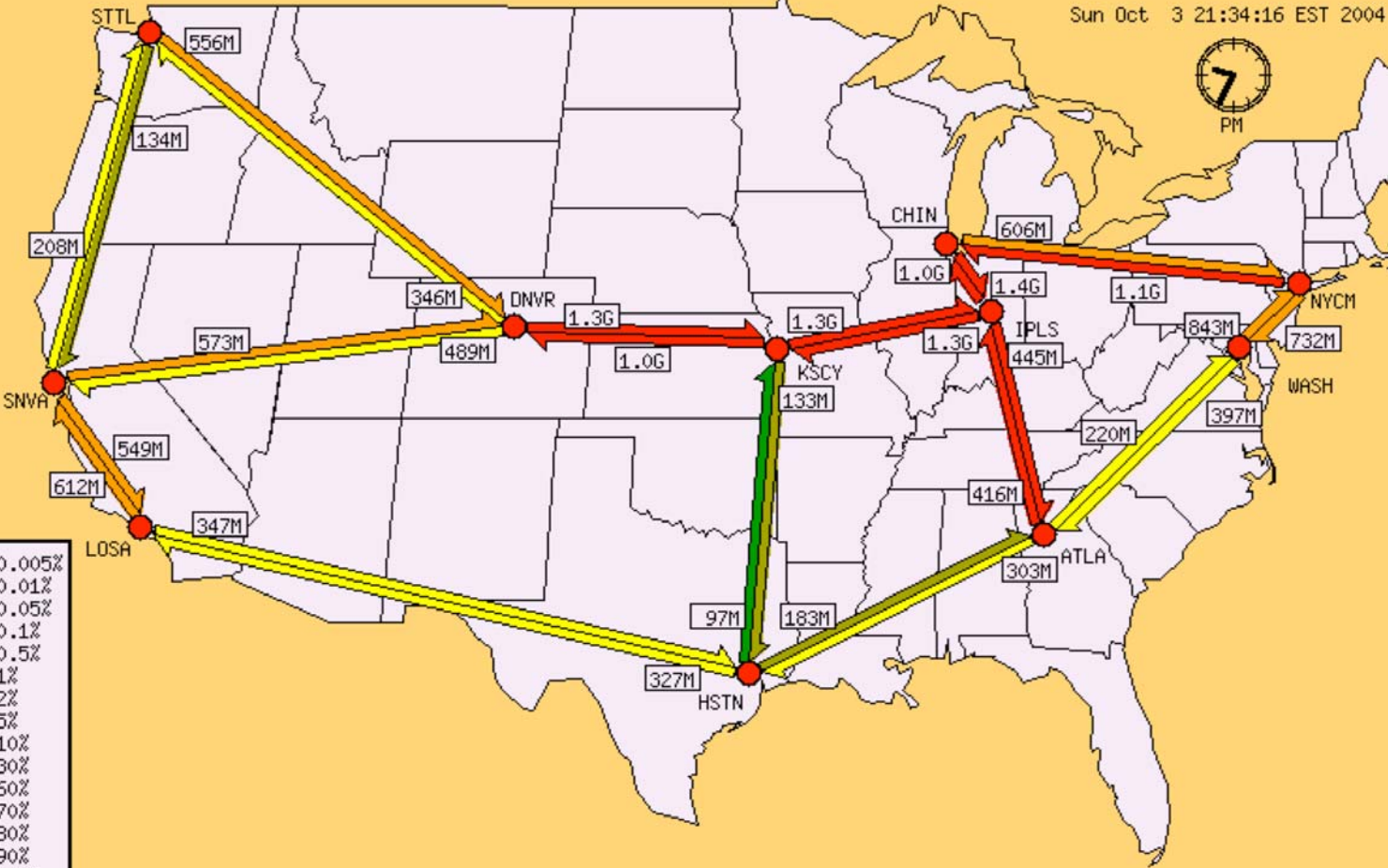
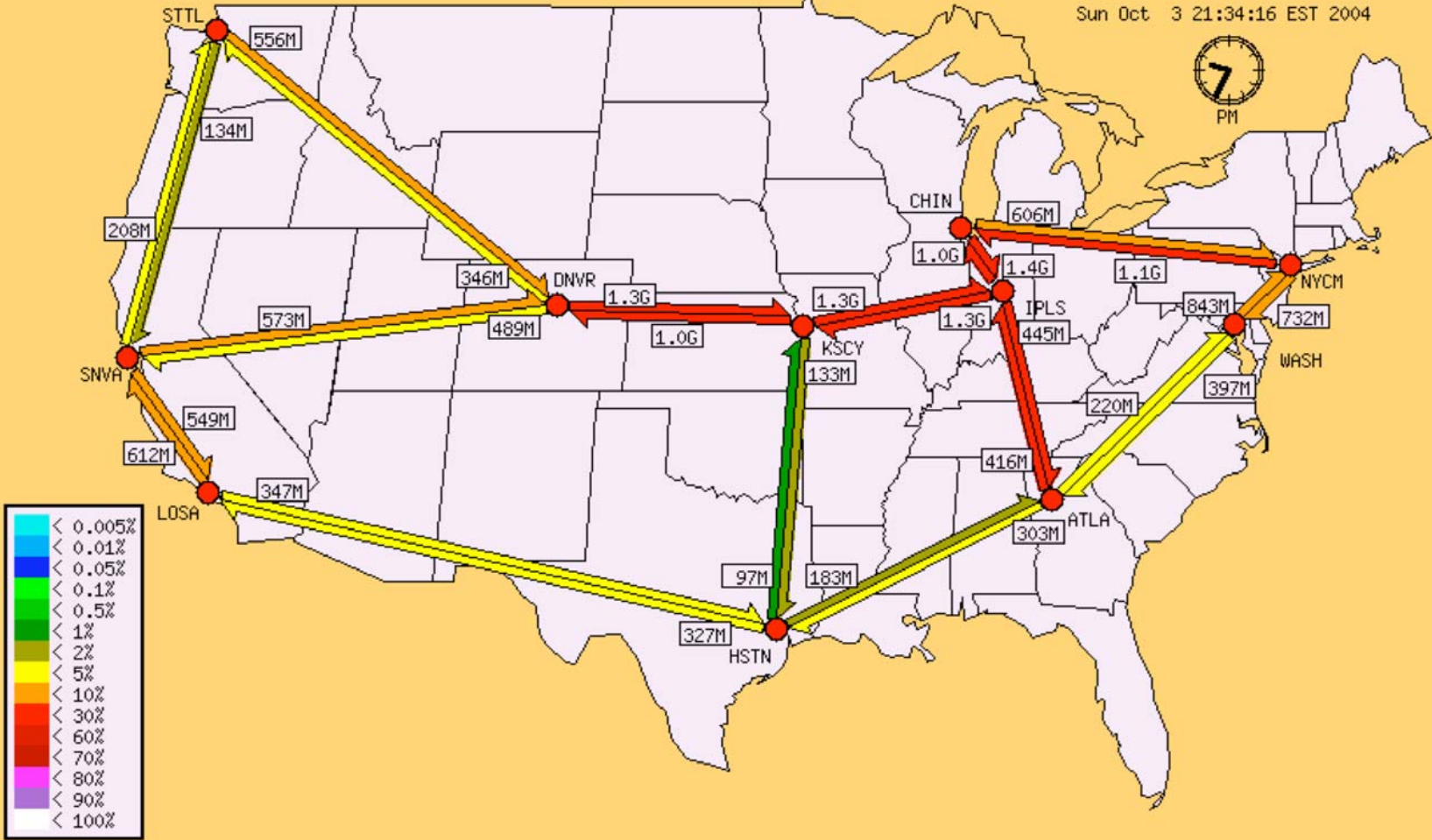
Overview

- Data description
 - The Internet2 (Abilene) data network
 - Netflow traffic data
- Data collection
- Data analysis
 - Techniques
 - Preliminary results
- Future work

What is Abilene?

- Internet2 (Abilene) is a nationwide high-speed data network for research and higher education.
 - Network backbone runs at 10 Gbps
 - Over 220 member institutions
 - Peers with over 40 other research networks
- Abilene uses the same protocols as Internet1 but only carries academic traffic.
 - This is like the old NSFnet or vBNS

Sun Oct 3 21:34:16 EST 2004



Why is Abilene Interesting?

- The Abilene network is a *transit network*.
 - It includes both international and domestic traffic.
 - It offers a good view of server networks.
 - Commercial transit networks do not share traffic data.
- The Abilene network is *uncongested*.
 - Statistics will not be biased by packet loss.
- The Abilene network contains *students*.
 - Students are unconcerned about niceties of law.
 - There is a lot of peer-to-peer and “grey” traffic.

What is "Netflow"?

- In the early 1990's, Cisco introduced a new network router architecture.
- The "line cards" in their new routers contained a hardware hash table for current network connections.
- Somebody got the bright idea of sending entries from the table onto the network before clearing them from the hash table.

What is a Network Flow?

- A *network flow* consists of one or more packets sent from a *source (IP, port)* to a *destination (IP, port)* using a certain *transport protocol* during some time interval.
- Example:
 - Source: 156.56.103.1, port 80
 - Dest.: 149.159.250.21, port 6132
 - Protocol: TCP
 - Packets: 20
- The above network flow would be typical for a Web connection.

Wait a Minute!

- Don't TCP connections involve two-way communication?
 - Yes, so every TCP connection is actually *two* flows from the point of view of Netflow.
- UDP and ICMP are stateless, so how can they be aggregated into flows?
 - We assume that packets with matching 5-tuples during some period of time are part of the same flow.
- Isn't it hard for a router to keep up with this?
 - Yes, so most modern routers *sample* the flow data at a ratio of about 100:1.

How is Netflow transmitted?

- Most modern routers support the “Netflow v5” format for representing flows.
 - This includes a variety of additional information about each flow.
- The router uses UDP to send packets containing between 1 and 30 flow records to a management workstation.
 - (In this case, the management workstation is sitting on my desk.)

Netflow-v5 Header Format

version number	# of flows in packet	
router uptime (ms)		
export time (sec. since 1970-01-01 00:00:00 UTC)		
export time (ns)		
sequence number		
engine type	engine ID	[padding]

Netflow-v5 Flow Record Format

source IP address			
destination IP address			
IP address of next-hop router			
SNMP ifIndex (in)		SNMP ifIndex (out)	
total number of packets			
total number of octets			
router uptime at start of flow (ms)			
router uptime at end of flow (ms)			
source port		destination port	
[padding]	TCP flags	protocol	ToS
source AS		destination AS	
source mask	dest. mask	[padding]	

How Much Data is There?

- The Abilene routers generate between 700,000,000 and 800,000,000 flows per day.
 - At 48 bytes per record, that amounts to around 35 GB of data.
 - Flows come in at a rate of about 3.4 Mbps.
 - Data compresses at a ratio of about 2.8:1.
- Most existing tools can't handle this volume of data.

What's the Motivation?

- Okay, so I'm storing egregious amounts of data and making my hard drive whimper...
what for?

Flow Data as a Behavioral Network

- Think of a single flow as defining an *edge* from a *source node* to a *destination node*.
- The resulting network describes the Internet *as it's actually being used*.
 - Many possible biases are eliminated.
 - A lot of dynamic information is included.
- Most structural analysis of the Internet has (necessarily) focused on its *physical* structure.
- Imagine a Google based on data about where people actually go!

Behavioral Anomaly Detection

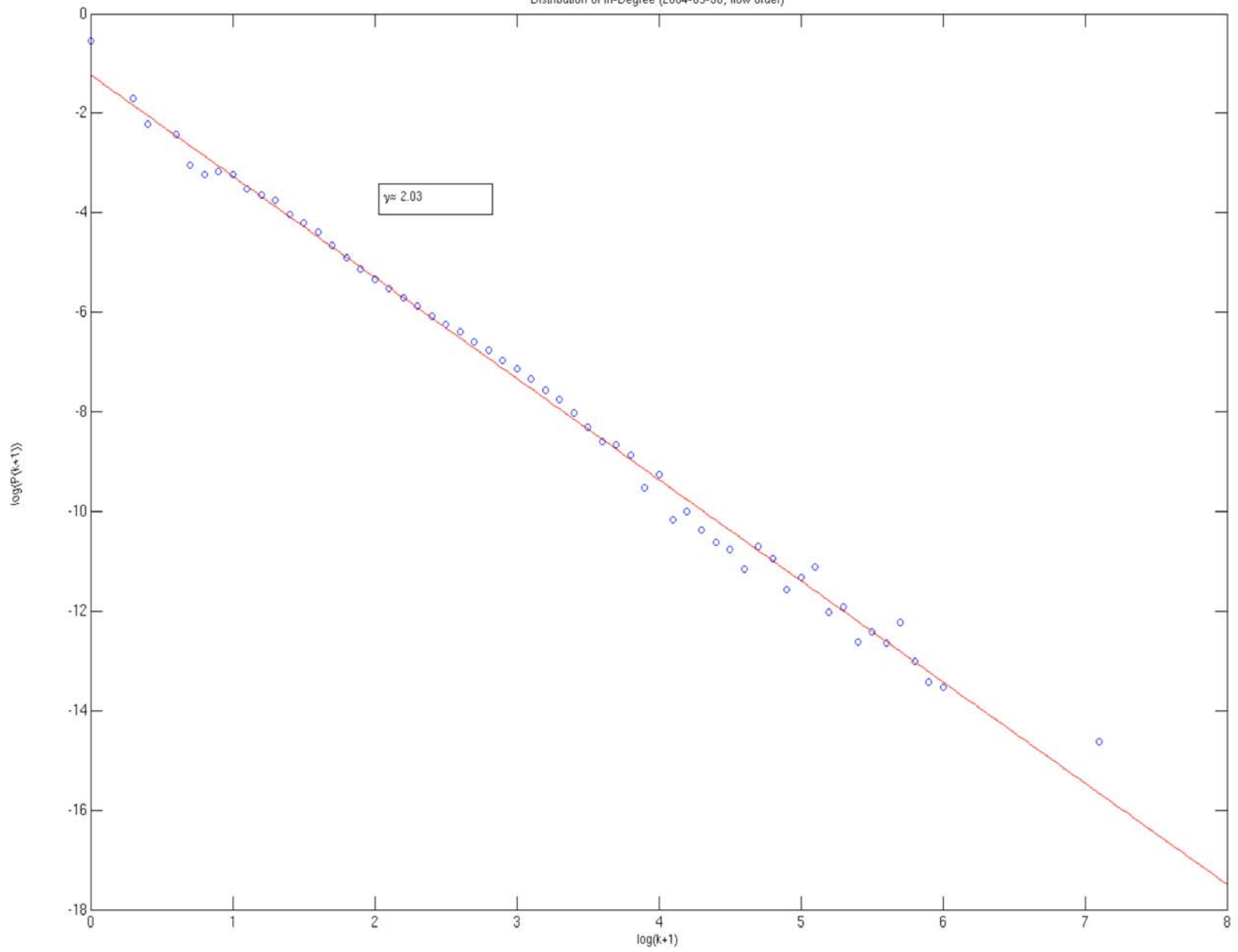
- My main interest is in recognizing different types of behavior based on flow data.
 - Can I determine whether a port is running a peer-to-peer application?
 - Can I see the spread of a new worm across the network?
 - Can I determine what kind of behavior is the prelude to an attack?
 - Can I find new peer-to-peer applications before the word is out?

Preliminary Results

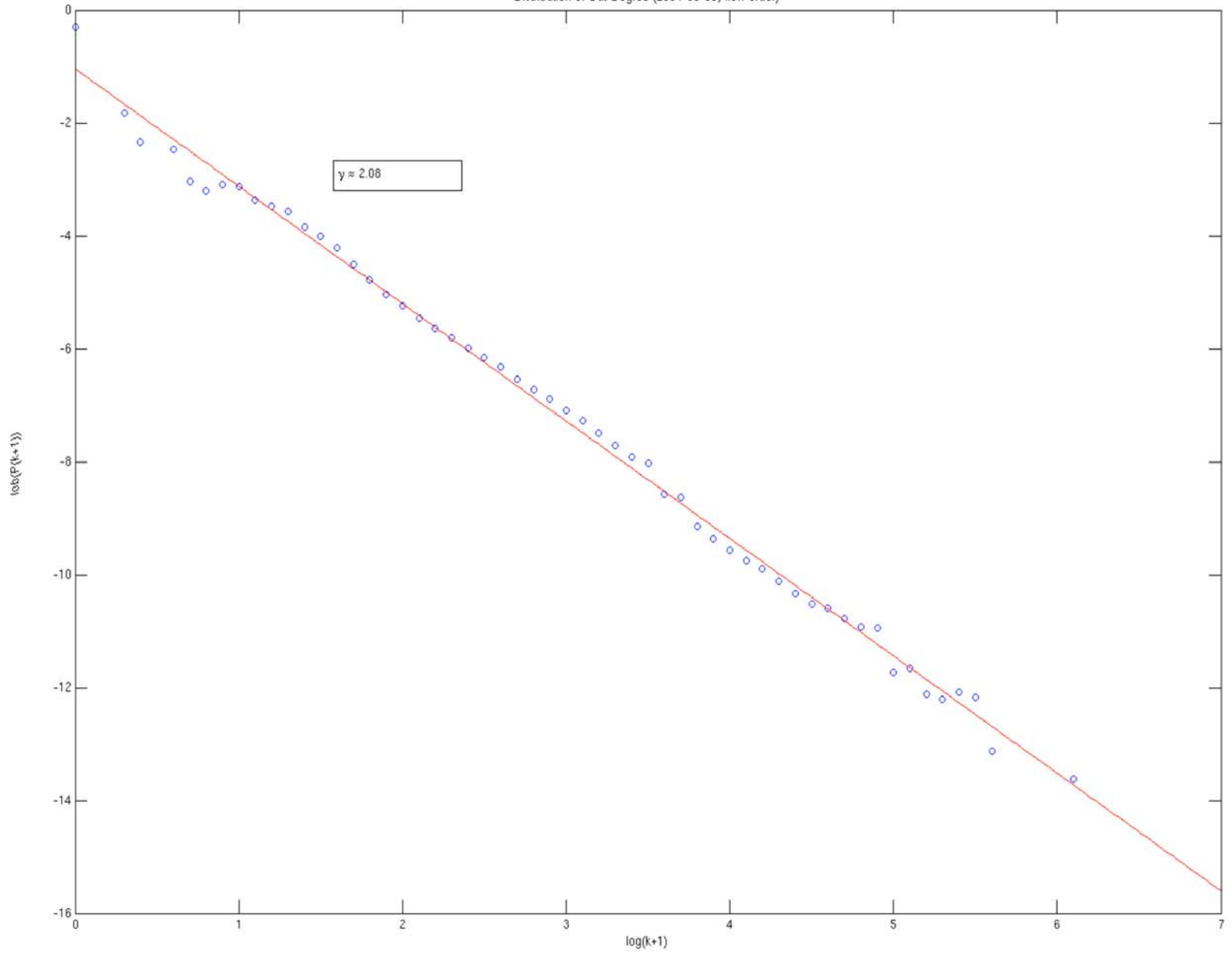
- I wish this section had more, but I'm really just getting off the ground...
- The size of data has been a major challenge.
 - The network formed by a day of flow data has about 29.7 million nodes and 128 million edges.
 - Just finding a way of converting a set of captured flows to a sparse matrix representation has been difficult.

Degree Distribution

Distribution of In-Degree (2004-09-30, flow order)



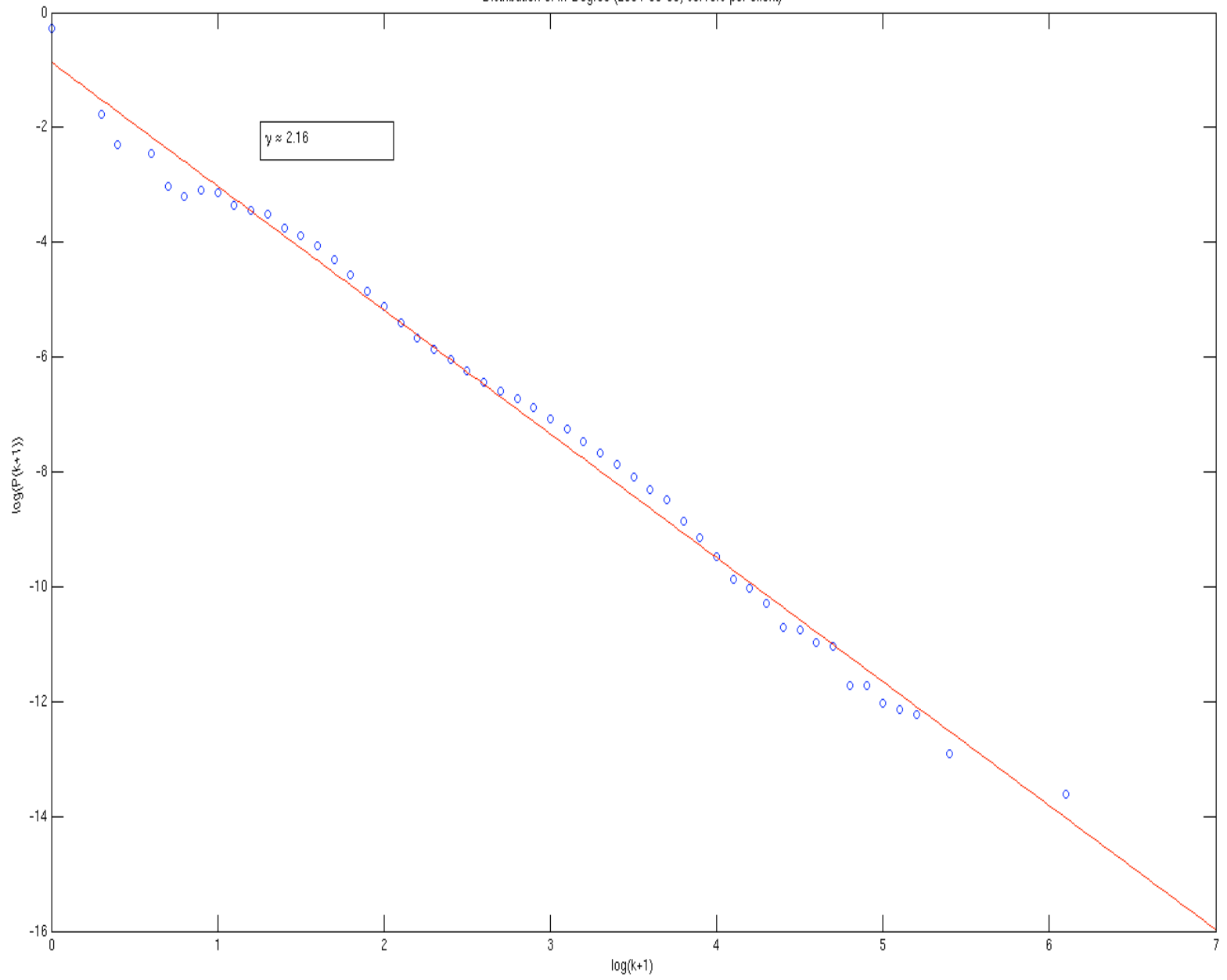
Distribution of Out-Degree (2004-09-30, flow order)



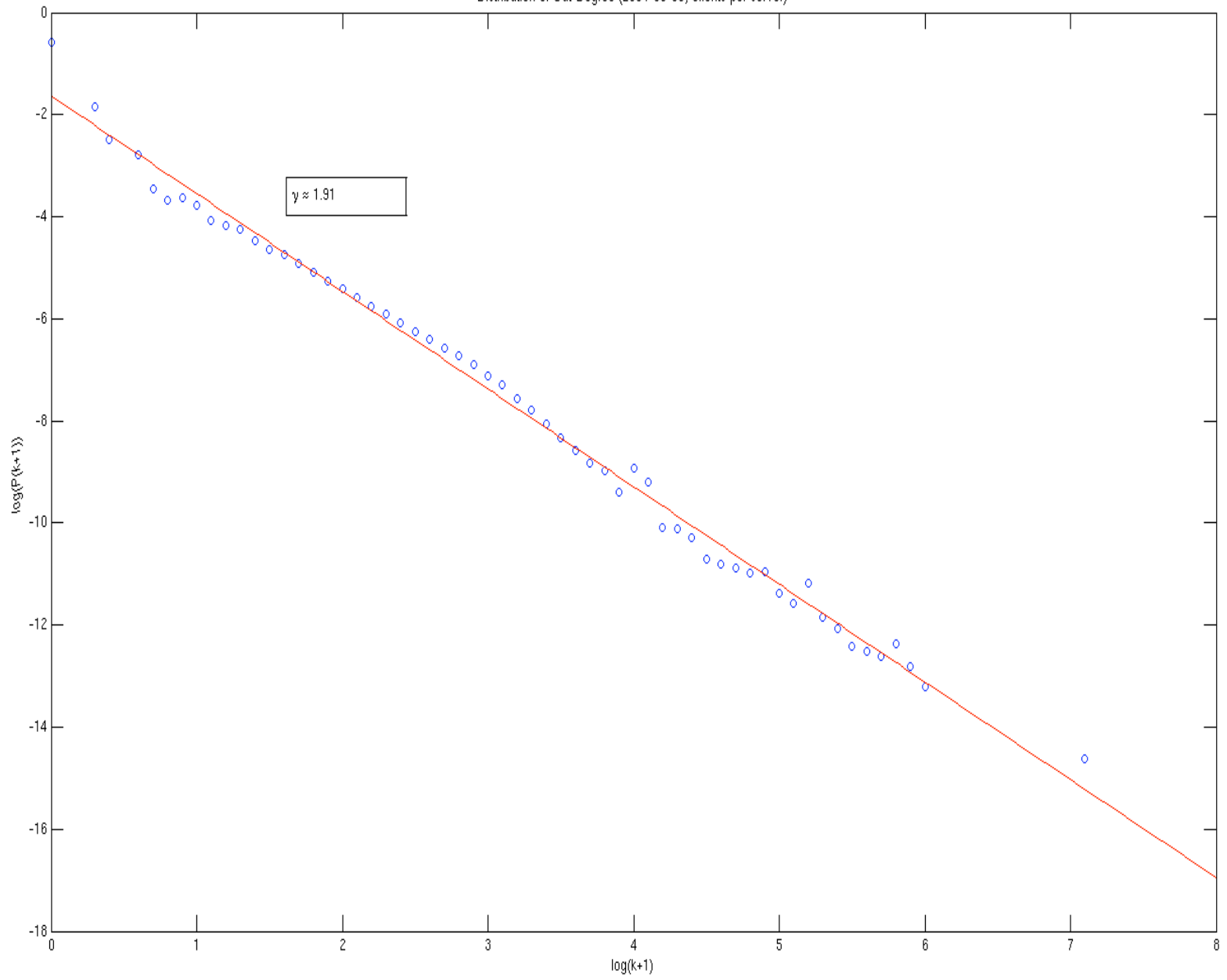
Determining Clients and Servers

- Every network connection involves two hosts:
 - The *client* is the system that *initiates* the connection.
 - The *server* is the system that *accepts* the connection.
- Because of sampling, we're as likely to see the client-to-server side as the server-to-client side.
 - This makes the direction basically meaningless.
- We can *guess* which is which using the port information.
 - The *more common* port number indicates the *server*.
 - The *less common* port number indicates the *client*.

Distribution of In-Degree (2004-09-30, servers per client)



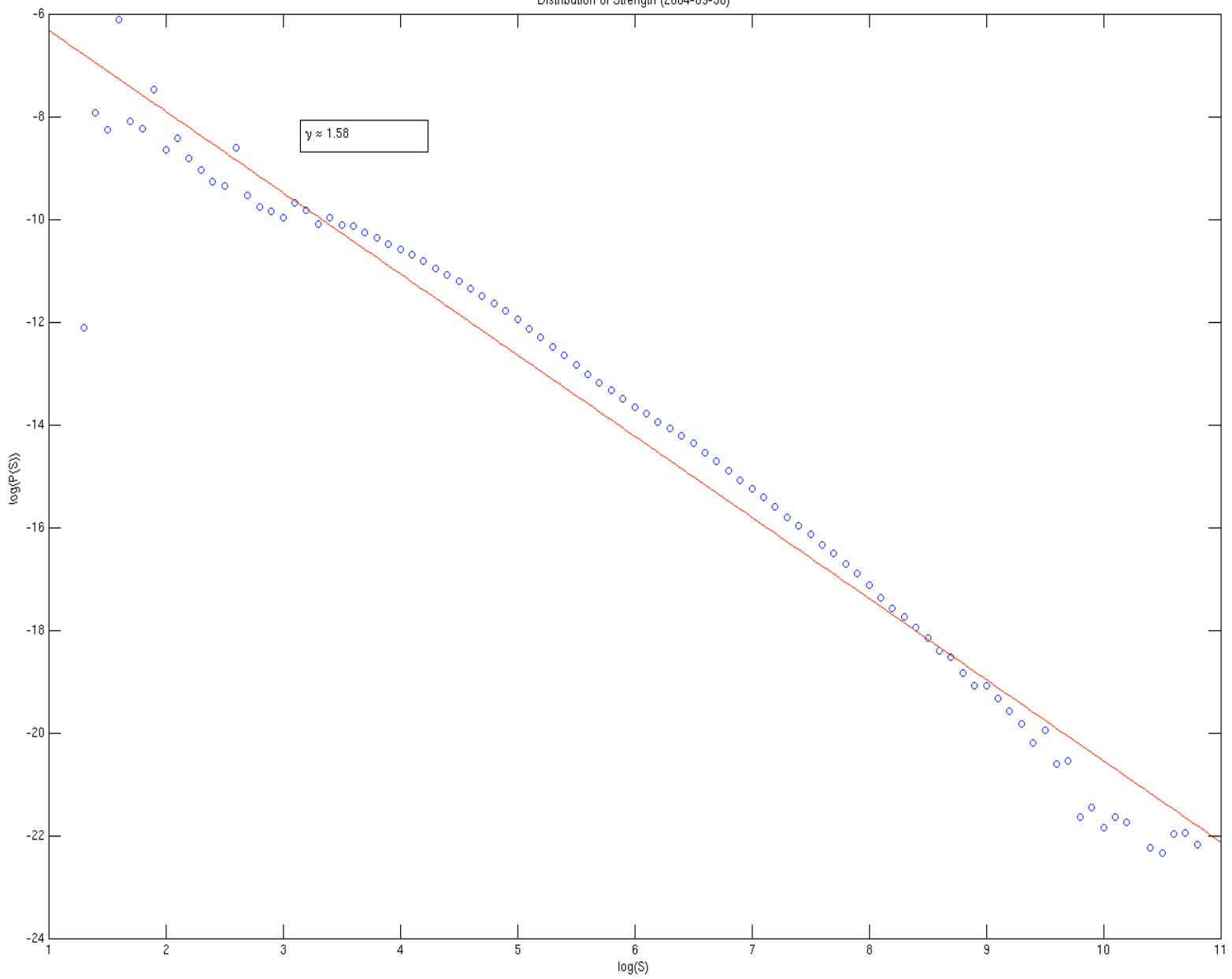
Distribution of Out-Degree (2004-09-30, clients per server)

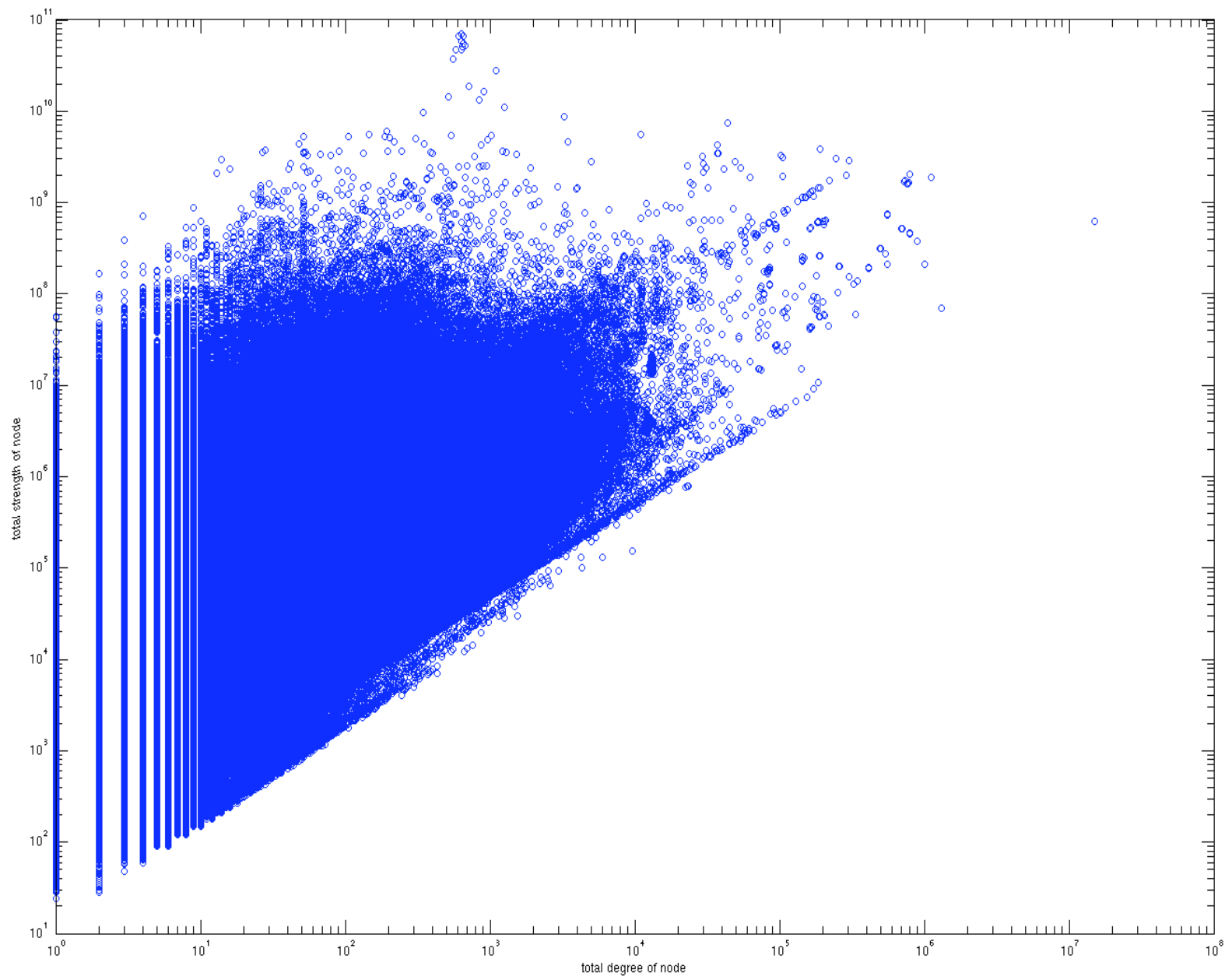


Strength Distribution

- This is the distribution of the total number of octets in and out of each node.
- Special problem for client/server version of the network
 - If we direct all flows from server to client, what do we do when we only have a volume for the opposite direction?
 - For now, I treat the network as being *undirected* for studying strength.

Distribution of Strength (2004-09-30)

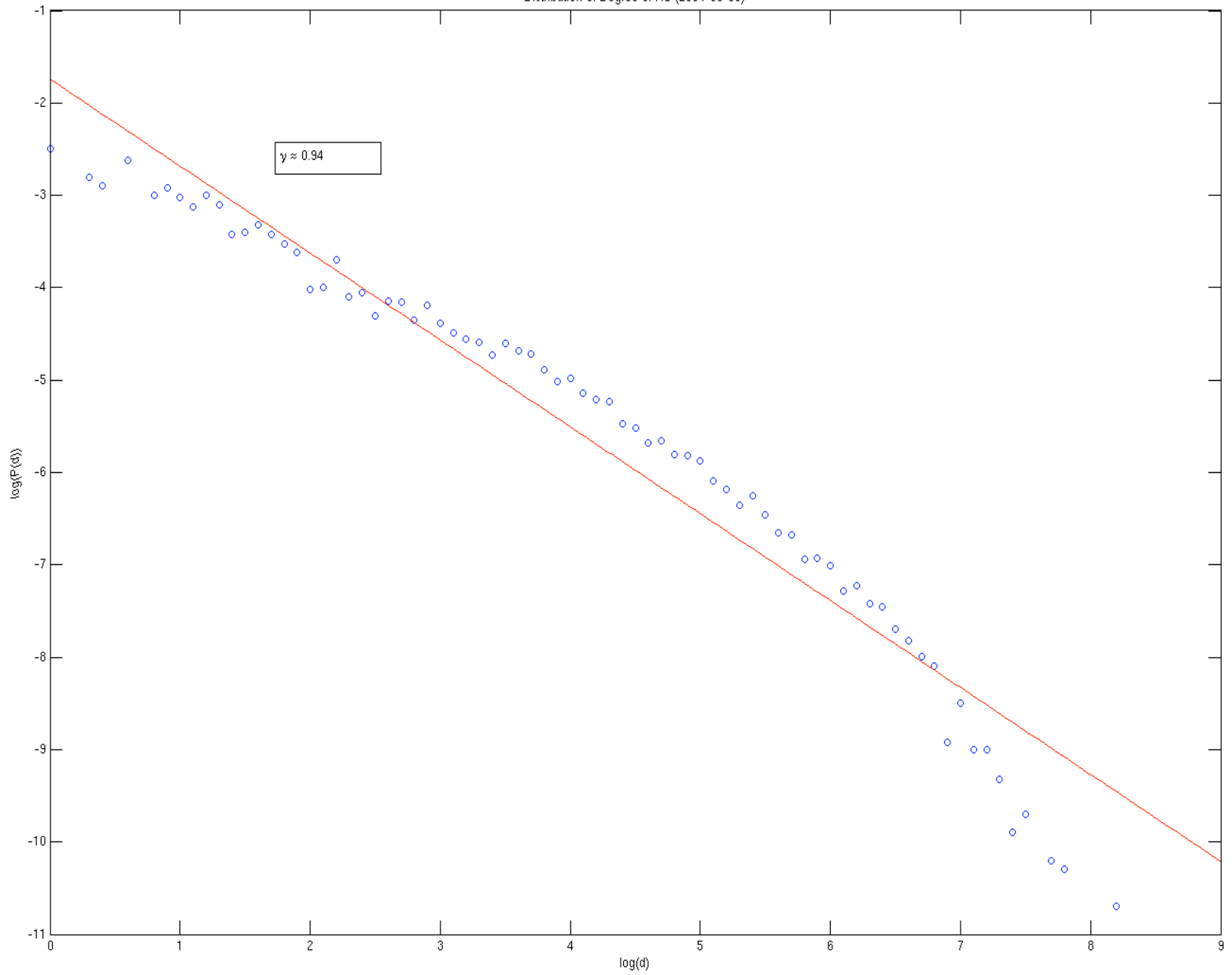




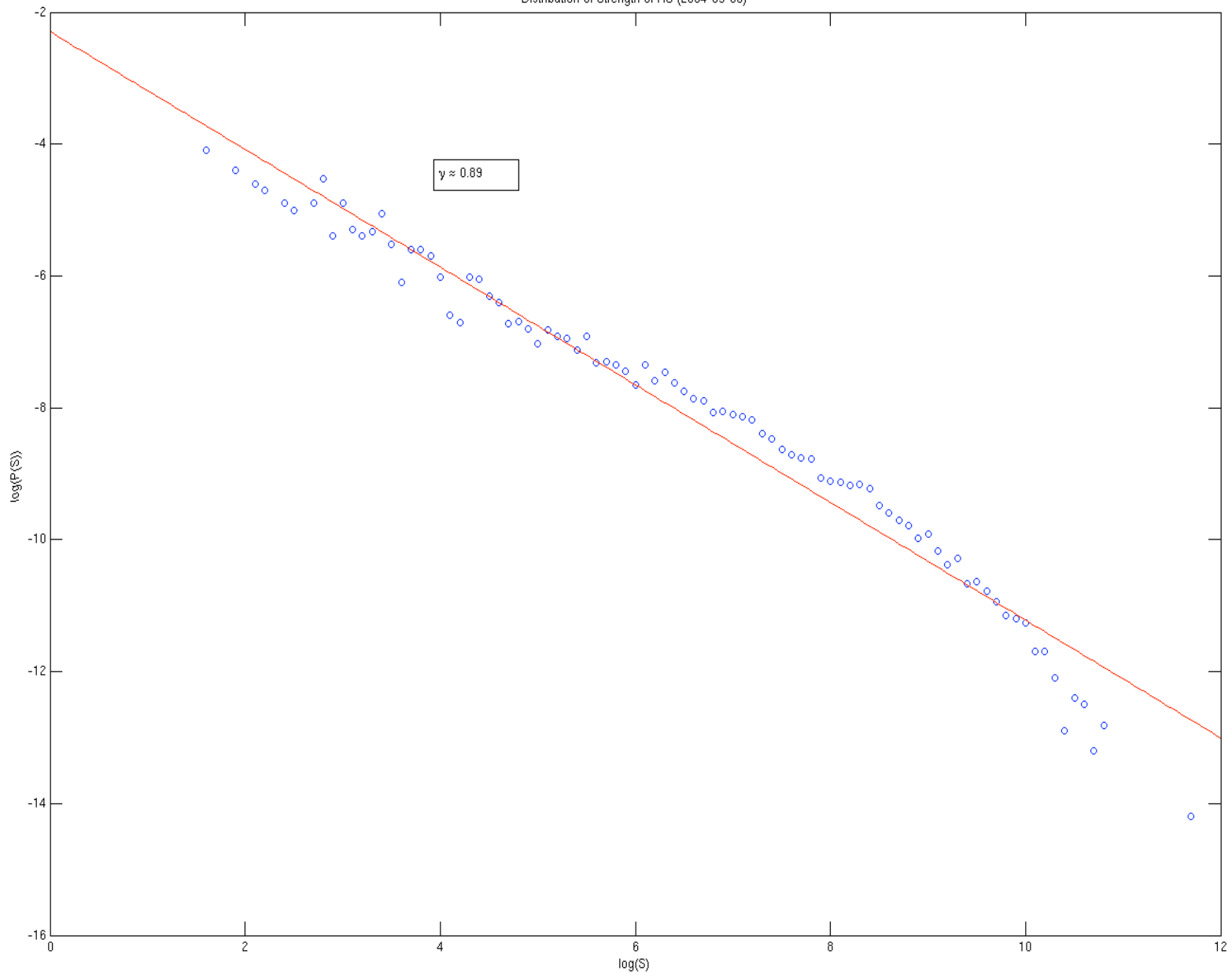
AS Numbers

- An “autonomous system” is the basic building block of the Internet.
 - An AS is responsible for its own interior routing.
 - An AS is usually a large organization.
 - For example, IU has its own AS, as does AT&T.

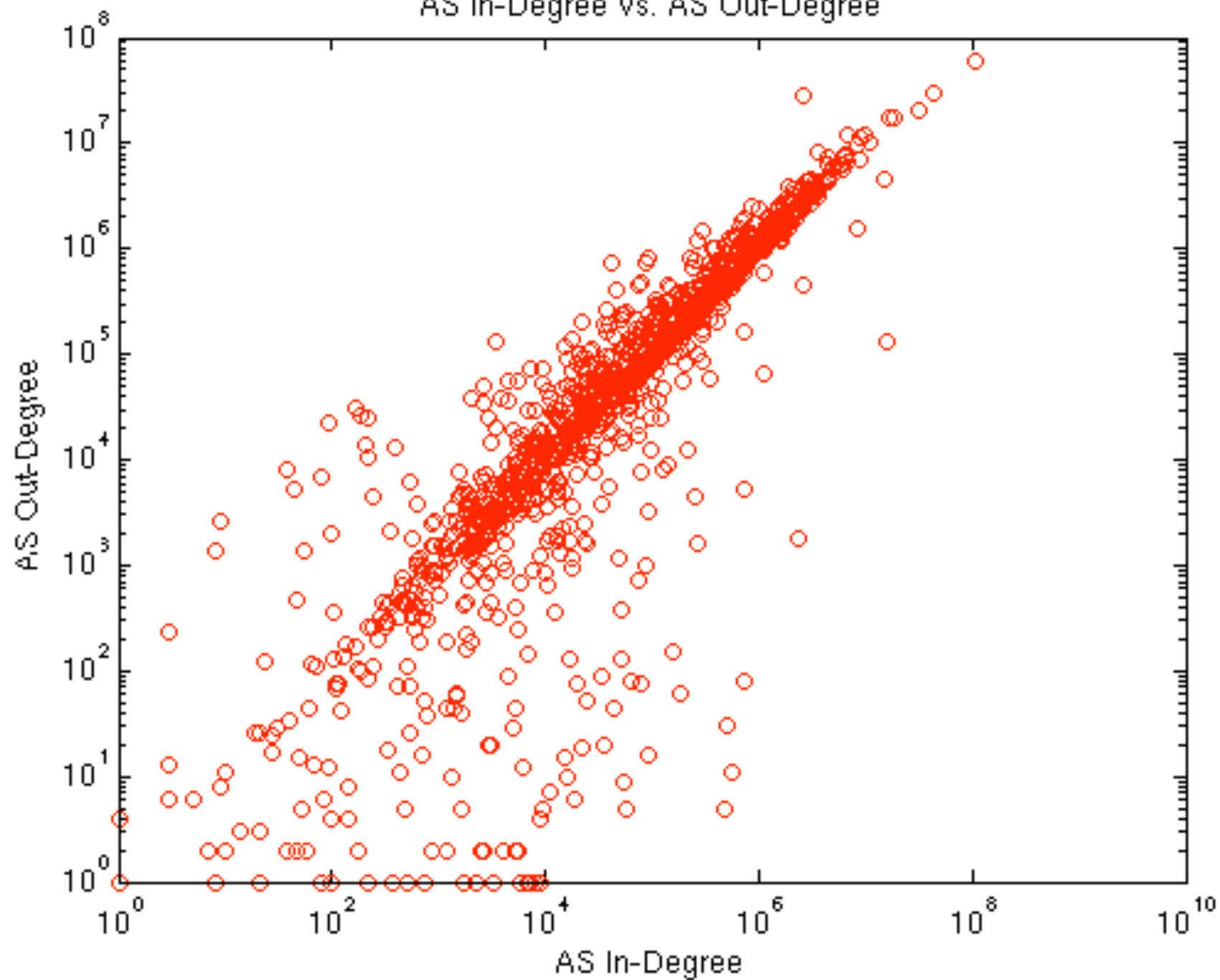
Distribution of Degree of AS (2004-09-30)



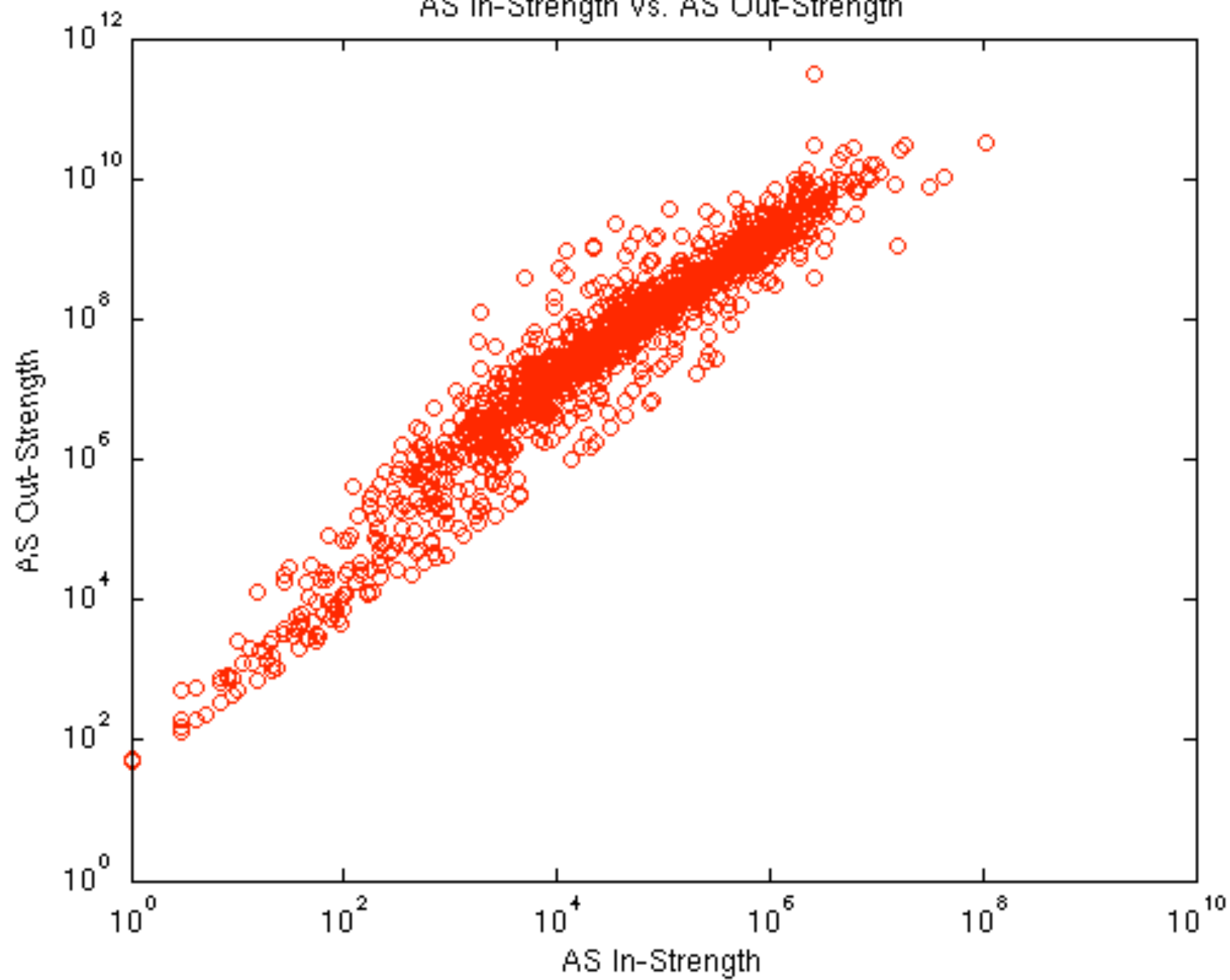
Distribution of Strength of AS (2004-09-30)



AS In-Degree vs. AS Out-Degree



AS In-Strength vs. AS Out-Strength



Top 10 ASes on Internet2

■ By degree

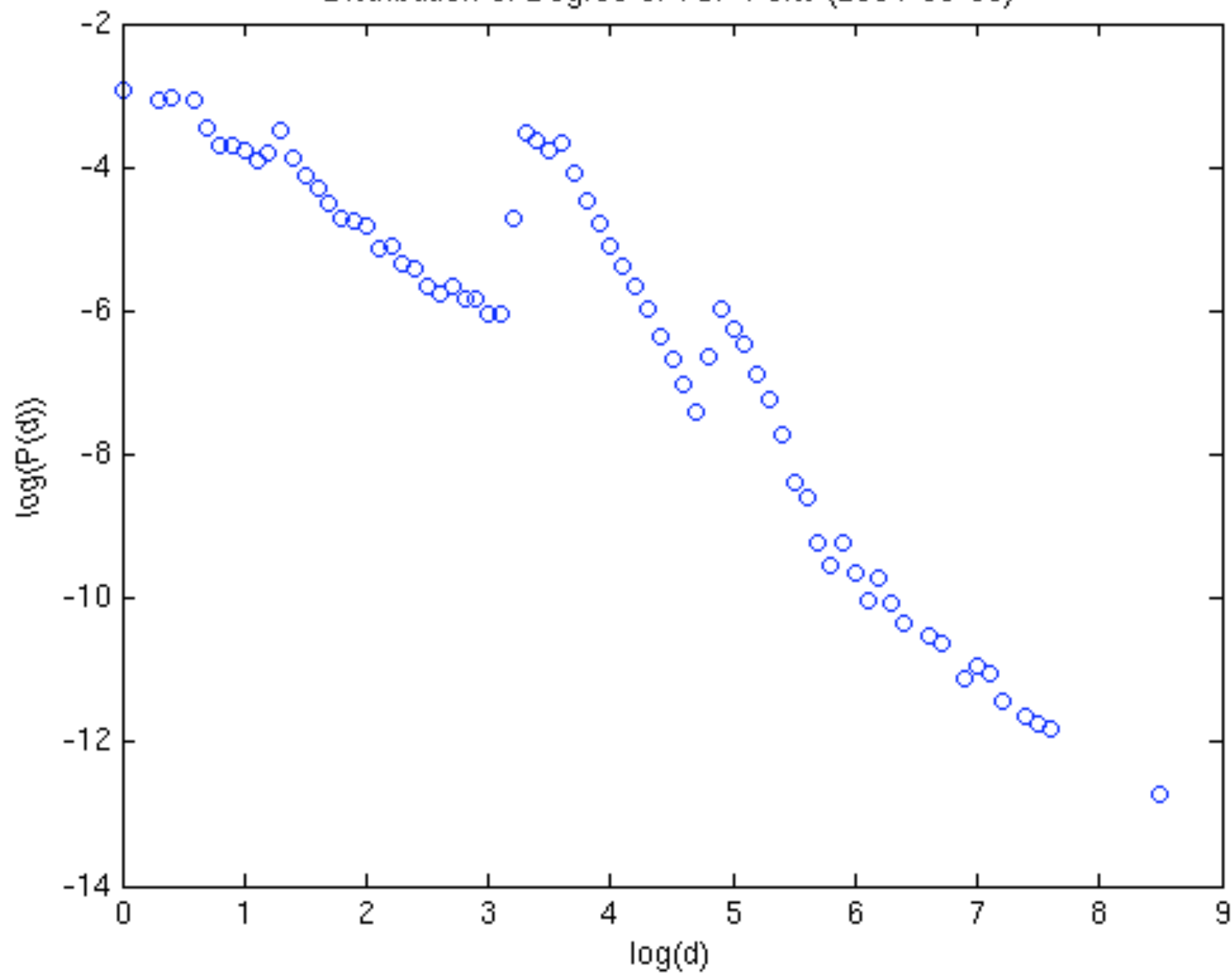
1. Hotmail
2. Microsoft
3. Microsoft-Europe
4. North Carolina (NCREN)
5. Michigan (MERIT)
6. University of Washington
7. MIT
8. UC-Berkeley
9. UMass
10. China (CERNET)

■ By strength

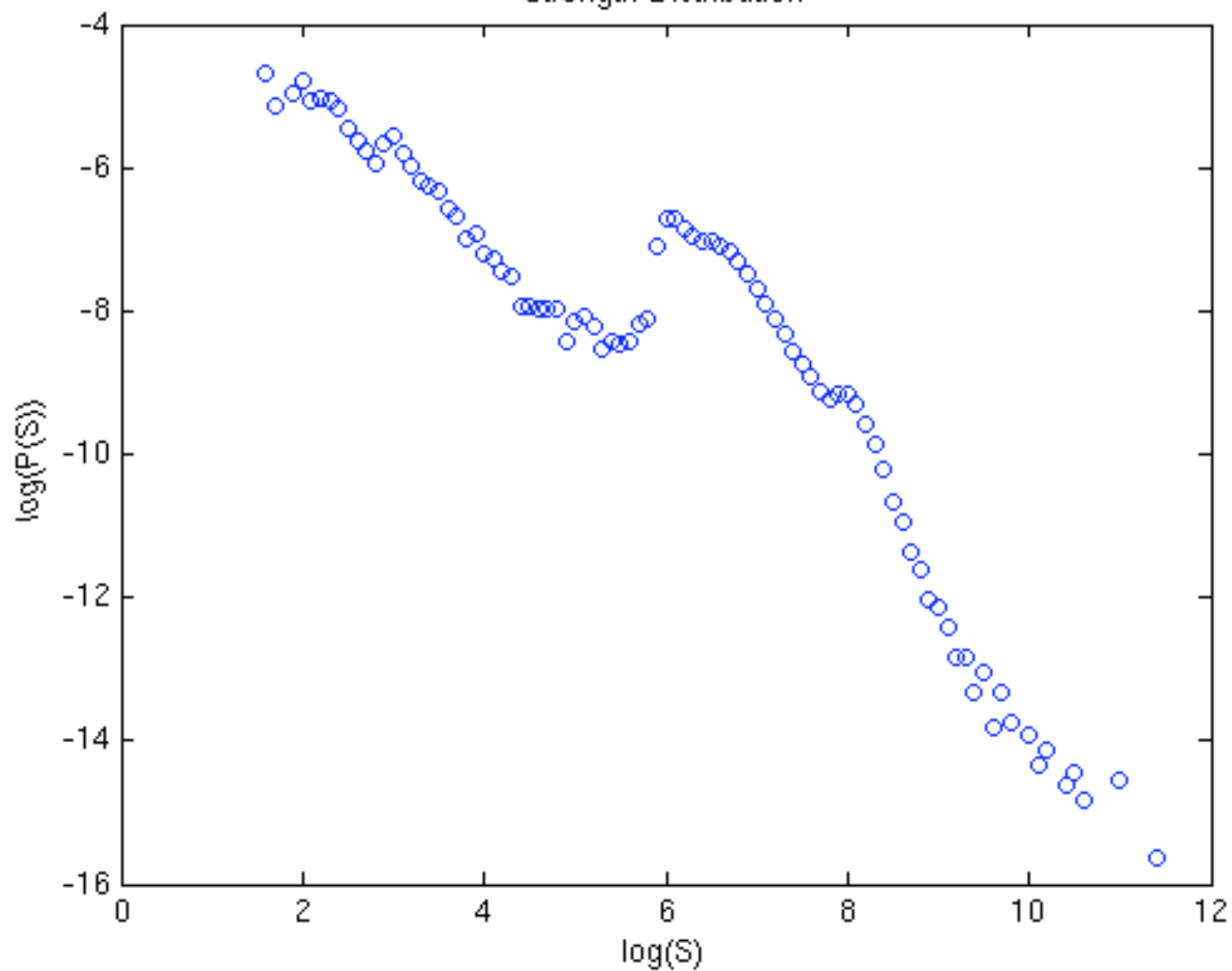
1. Abilene
2. University of Oregon
3. Hotmail
4. Microsoft
5. North Carolina (NCREN)
6. UCSD
7. UCLA
8. Michigan (MERIT)
9. University of Washington
10. UMass

TCP Ports

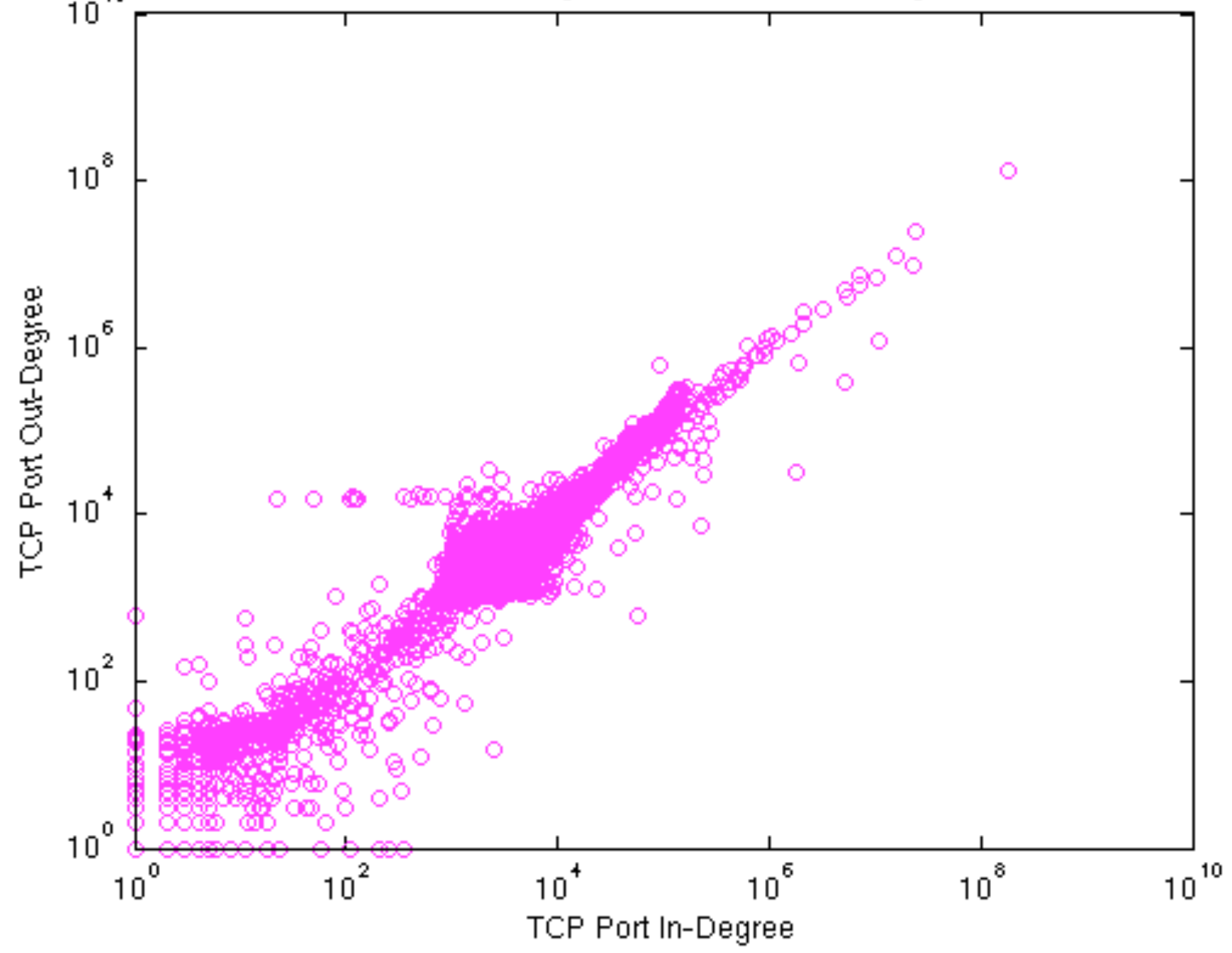
Distribution of Degree of TCP Ports (2004-09-30)



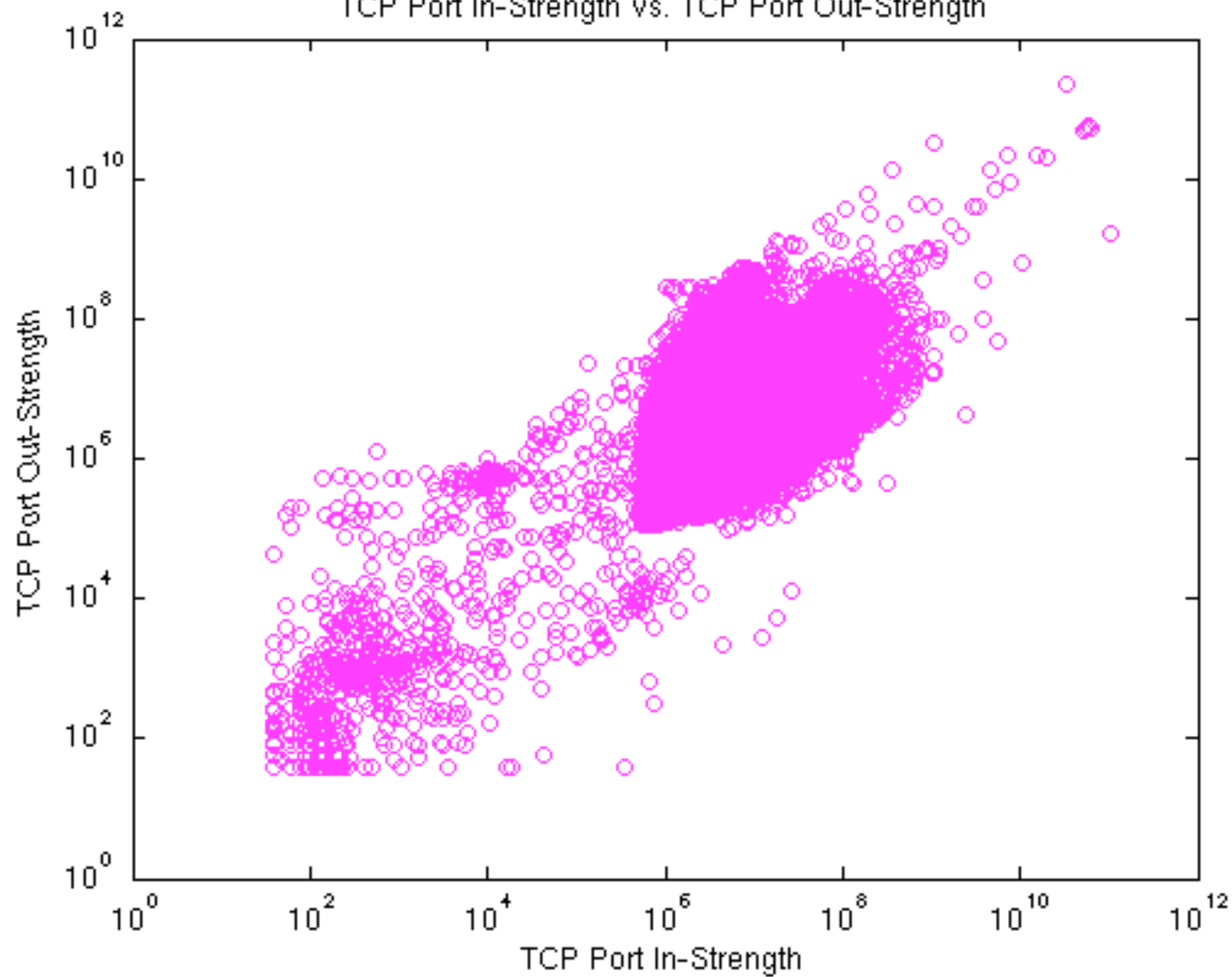
Strength Distribution



TCP Port In-Degree vs. TCP Port Out-Degree



TCP Port In-Strength vs. TCP Port Out-Strength



Top 10 TCP Ports on Internet2

■ By degree

1. Web
2. Gnutella
3. MS Messenger
4. SQL Server
5. Web (Encrypted)
6. Gnutella
7. Mail
8. Web Tunneling (8082)
9. BitTorrent
10. Usenet

■ By strength

1. Web
2. iperf
3. iperf
4. Usenet
5. RTP (Streaming)
6. iperf
7. SSH
8. BitTorrent
9. Port 388 ?!?
10. FTP

Where Do I Go Next?

- Start to look at the dynamics of the network.
- Focus on individual ports.
- Examine clustering coefficients.
- Attempt to filter out spoofed traffic.
- Consider the server-only and client-only networks.
 - This will involve treating flows as edges in a *bipartite* graph.
- Cluster nodes, ASes, and ports.

Thank You!

- Any thoughts, questions, comments, complaints, or observations are all welcome!