Basic Statistical Analysis of Traffic Data

Richard Clegg

Networks and Nonlinear Dynamics Group,

Department of Mathematics,

University of York

Slides prepared using the Prosper package and LaTeX

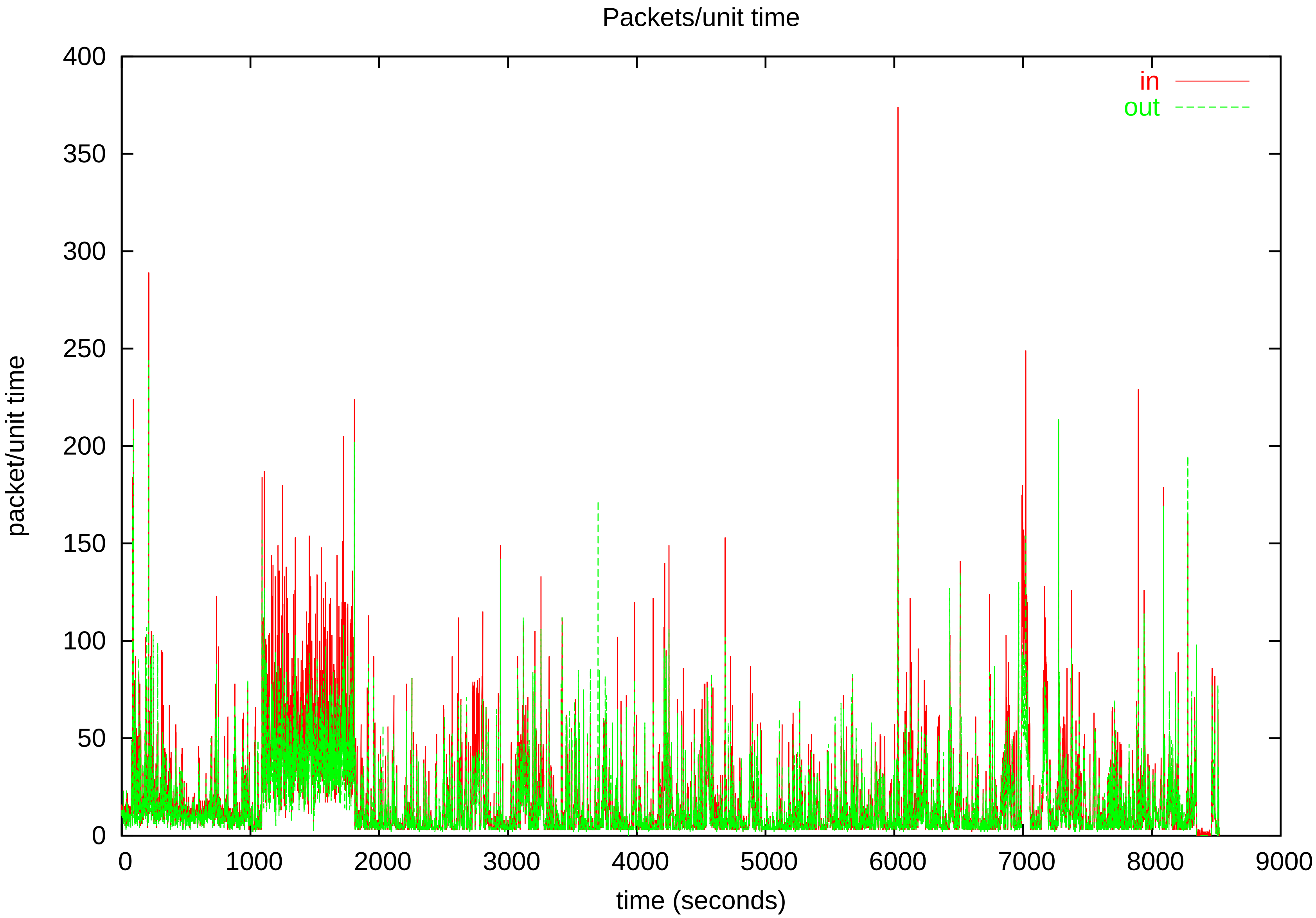
Initial Analysis of York Traffic Data

- Data collection from 11:45:09 – 14:07:14 Fri May 3 2002 (UTC) inbound and outbound stats from tcpdump on the main campus → JANET link.
- TCPdump (www.tcpdump.org) captures packet headers — this short trace is 24Mb and contains only connections to and from the electronics dept.
- There are lots of things that we can do with TCPdump — this presentation shows some of them.
- Our data is 296,333 packets 171,796,087 bytes (only 163.83MB — a quiet day).
- Naturally, these may be atypical figures. ICMP mainly blocked at firewall as are many common realtime applications (networked games)

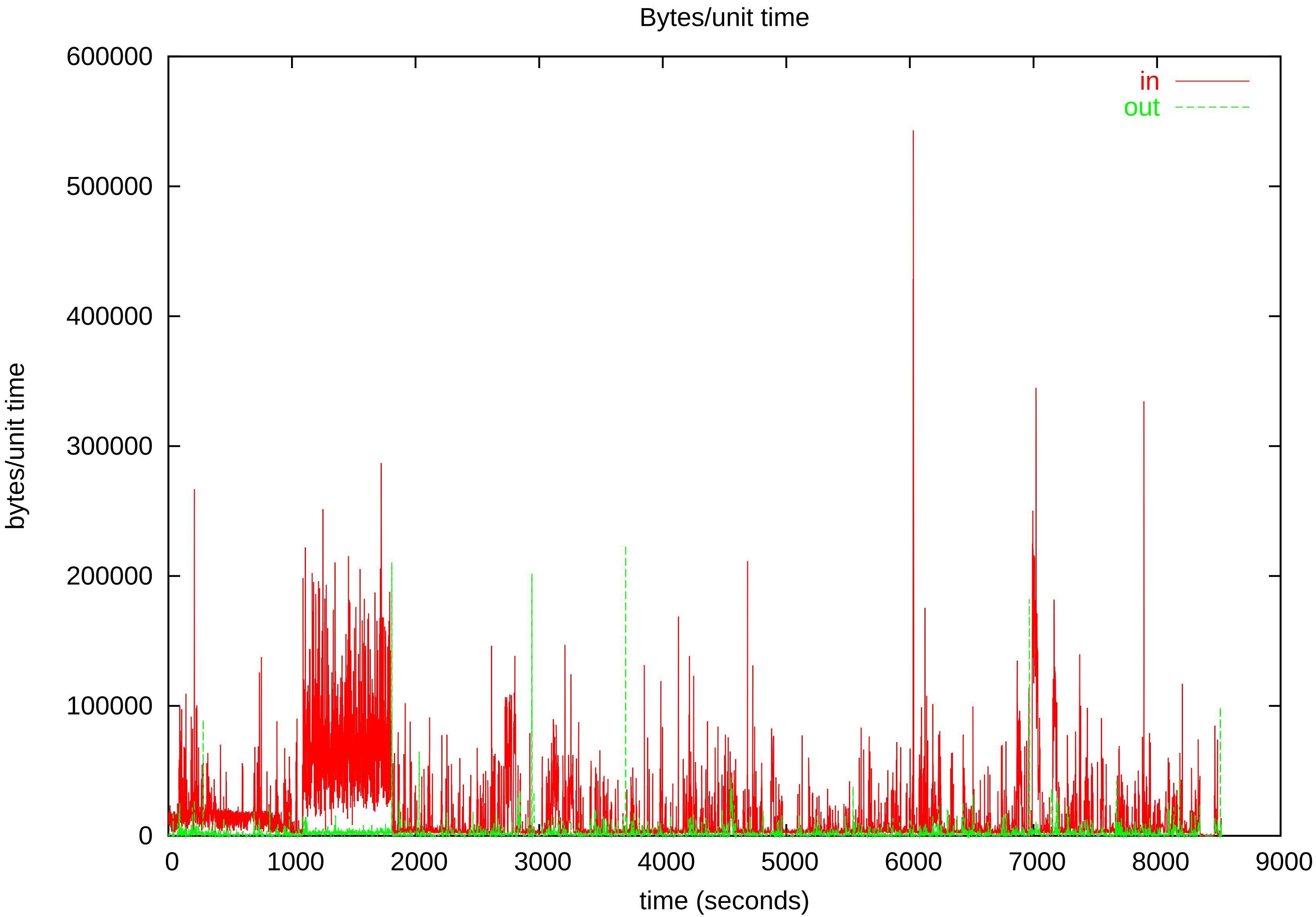
Gross Data Analysis

- 296,333 packets 171,796,087 bytes — Average packet length 579 bytes
- This can be broken into:
  1. 288,768 packets (97.4%) , 170,915,809 bytes (99.48%) TCP
  2. 7273 packets (2.5%), 855478 bytes (0.49%) UDP
  3. 292 packets (0.1%), 24800 bytes (0.03%) ICMP
- Or alternatively:
  1. 161,077 packets (54.35%) 155,978,063 bytes (90.79%) incoming — average packet length 968 bytes
  2. 135,256 packets (45.64%) 15,818,024 bytes (9.20%) outgoing — average packet length 116 bytes

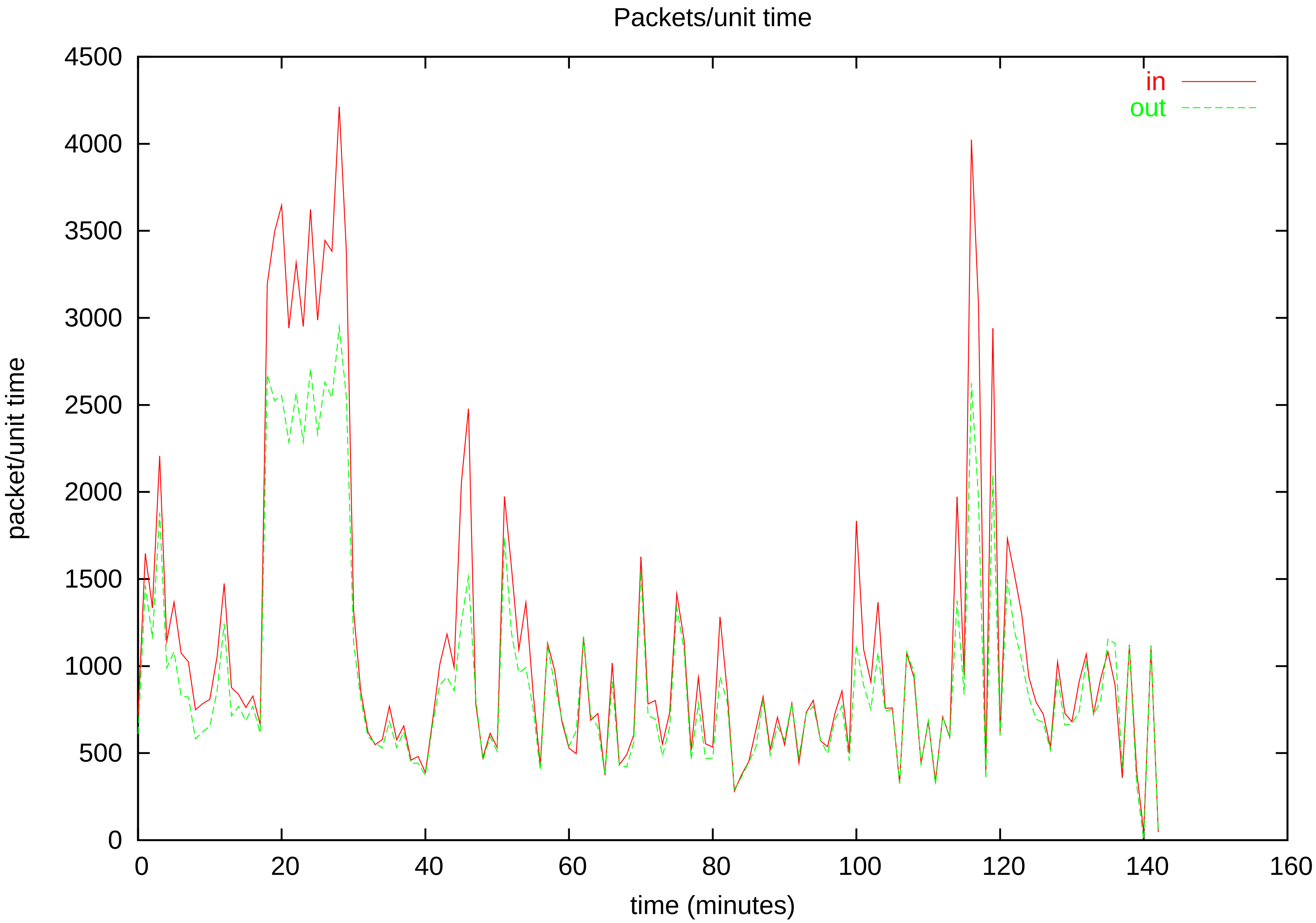
Time Domain — pkts/unit time (seconds)



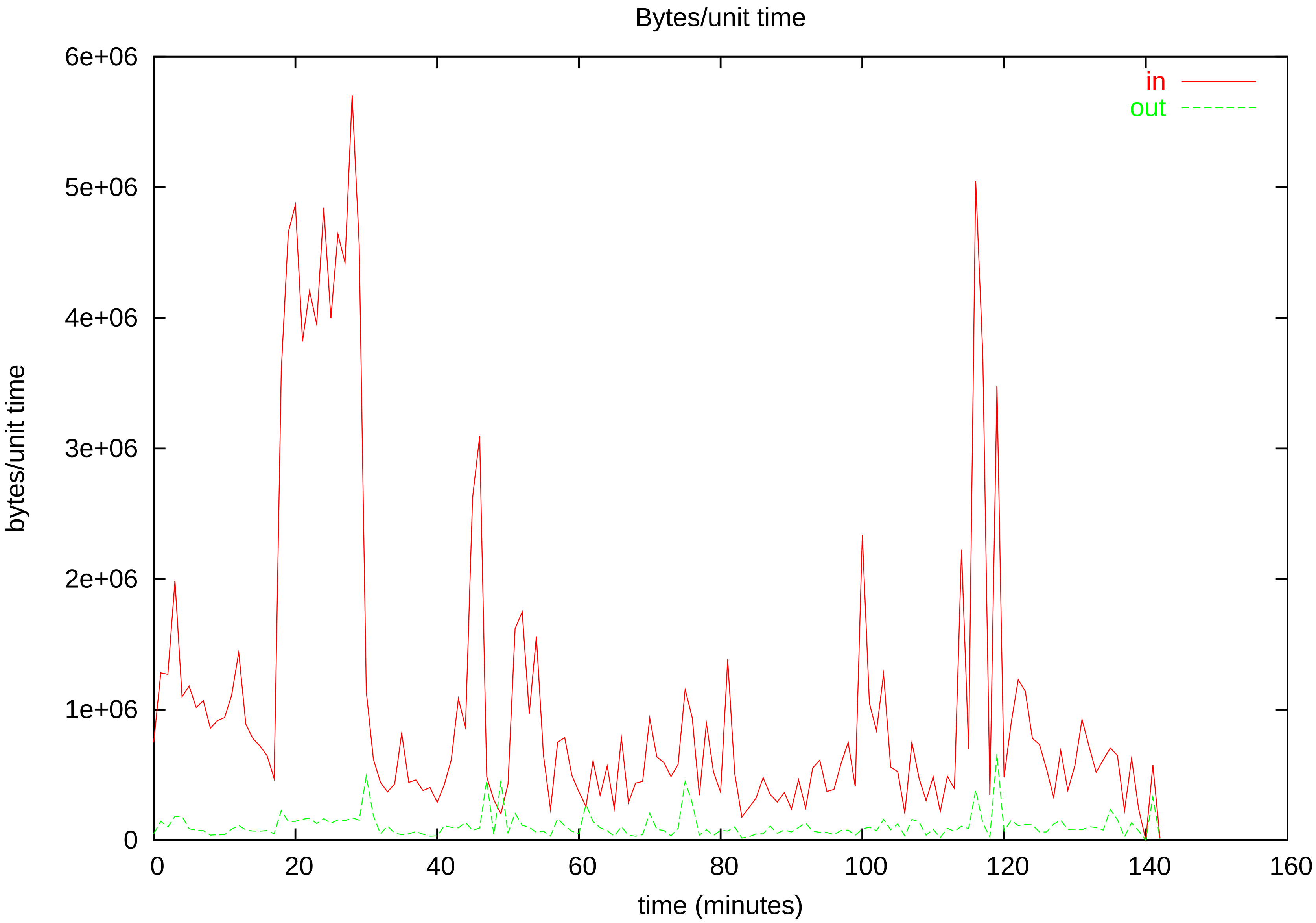
Time Domain — bytes/unit time(seconds)



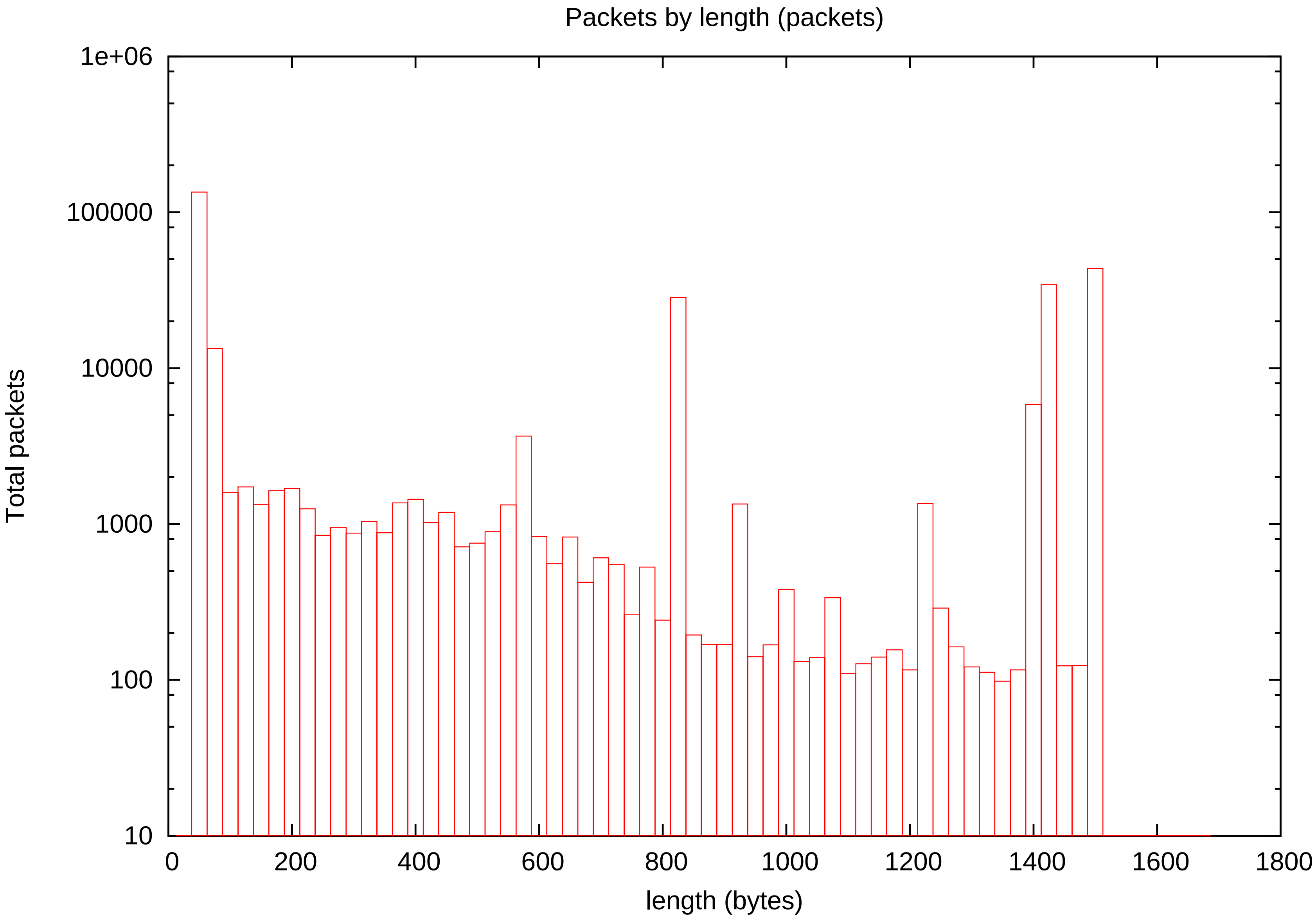
Time Domain — pkts/unit time(minutes)



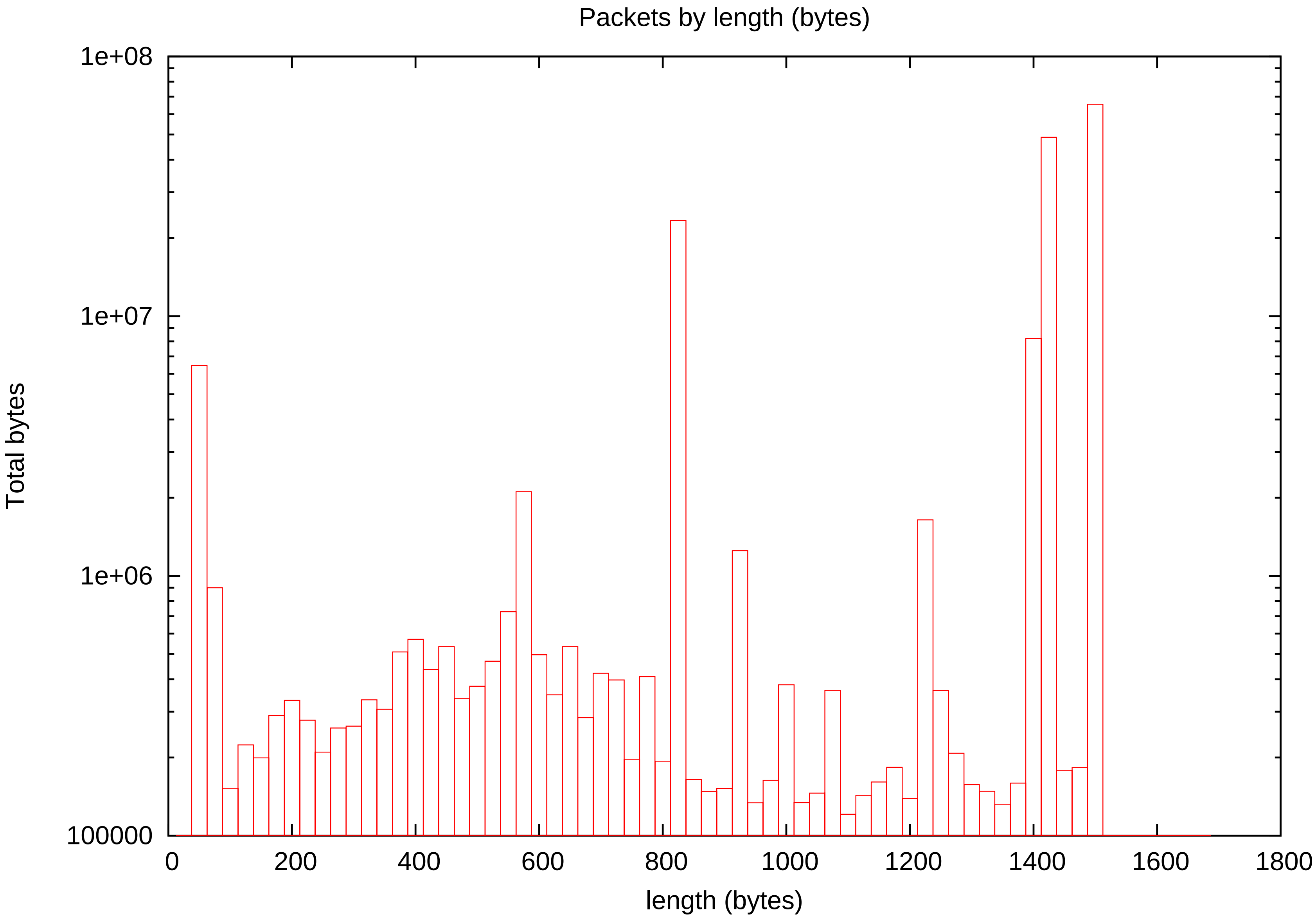
Time Domain — bytes/unit time(minutes)



Packet lengths — by no of packets



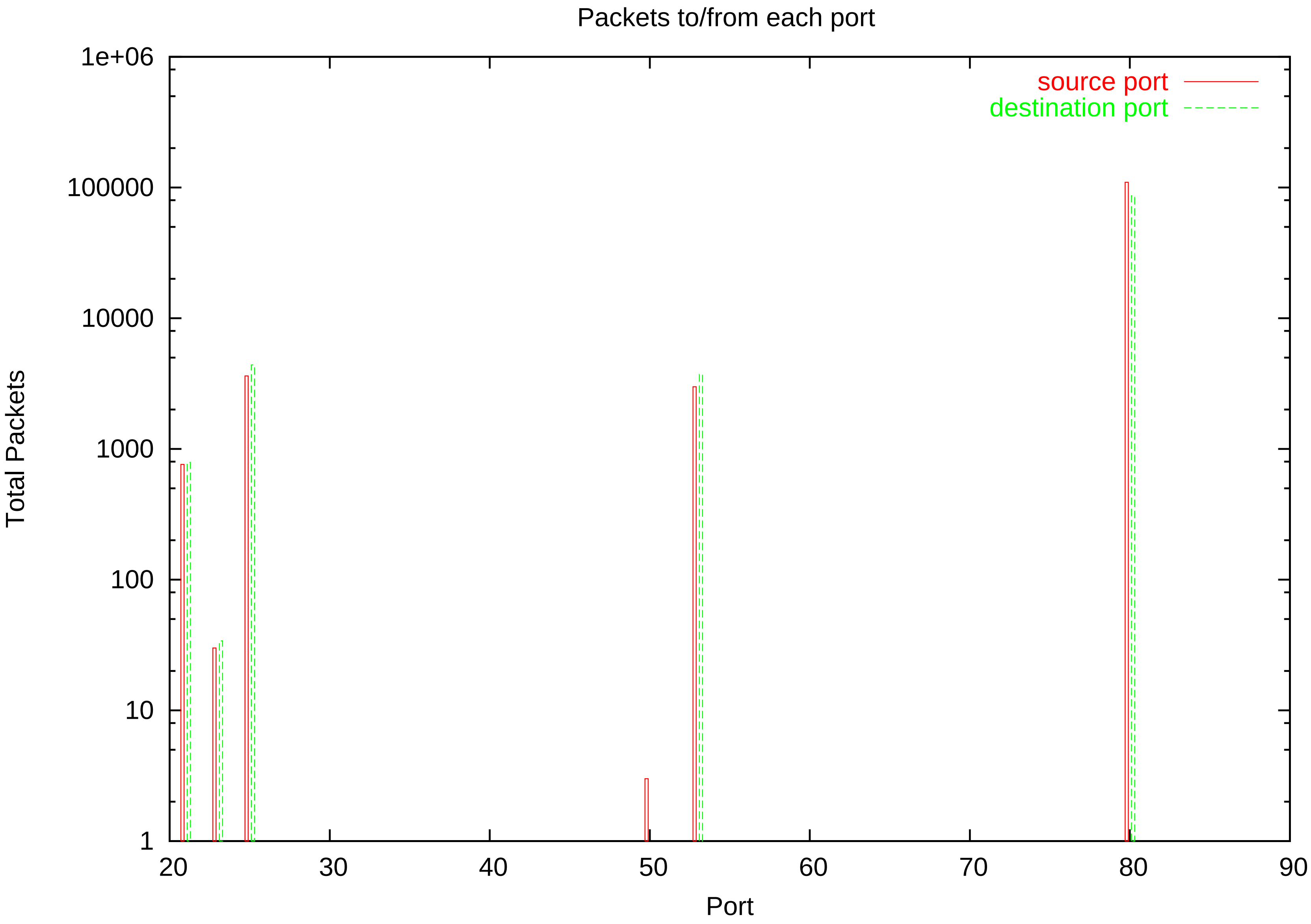
Packet lengths — by no of bytes



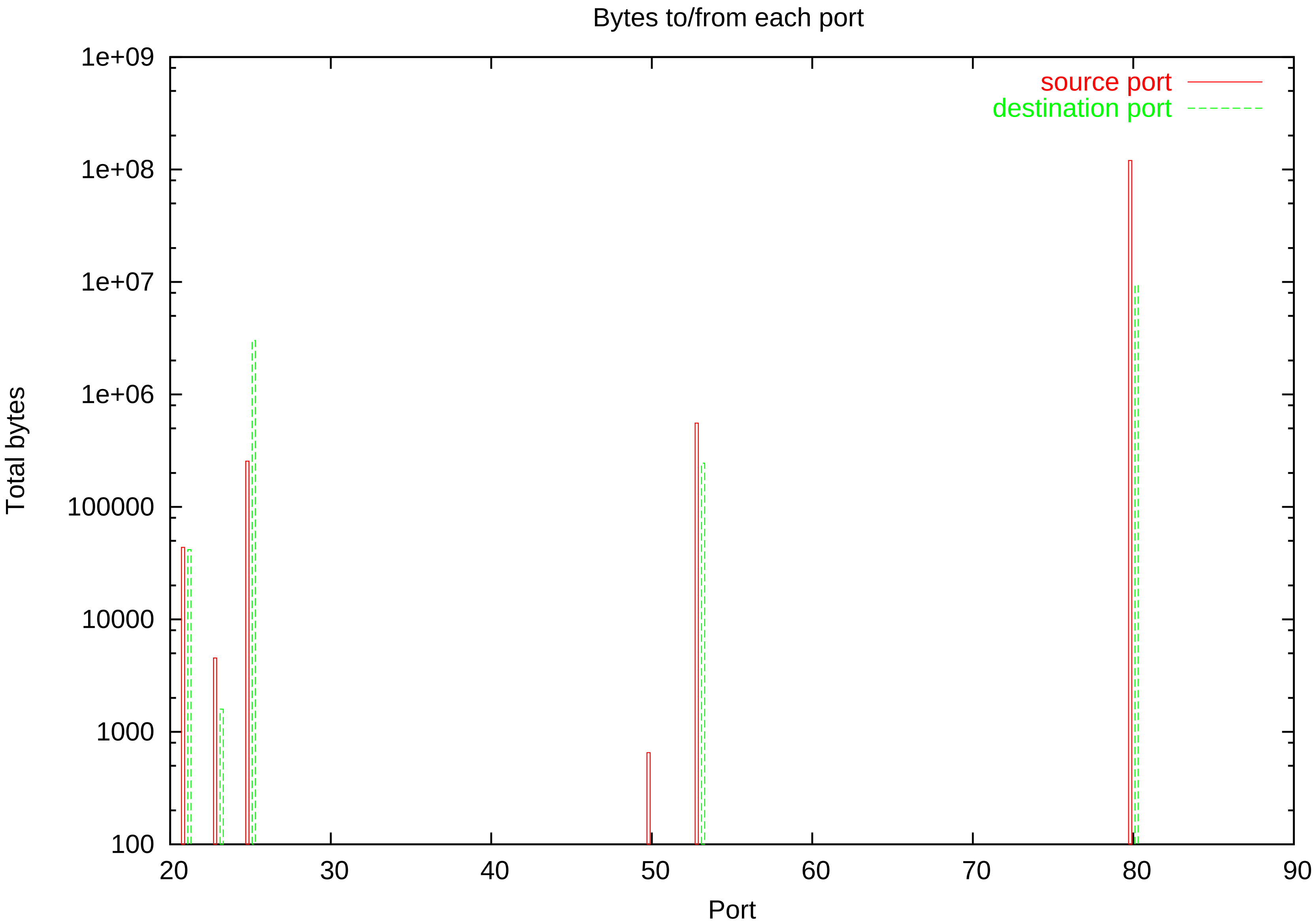
Disaggregating by port

- An obvious way to break this down is by port number.
- The following graphs show the main ports to which traffic was attracted
- Minor ports:
  - 23 (telnet)
  - 50 (remote mail check?)
  - 21 (ftp)
- Major ports:
  - 25 (SMTP)
  - 53 (Domain Nameserver?)
  - 80 (http) (by far the largest)
- (Only ports < 128 checked so far)

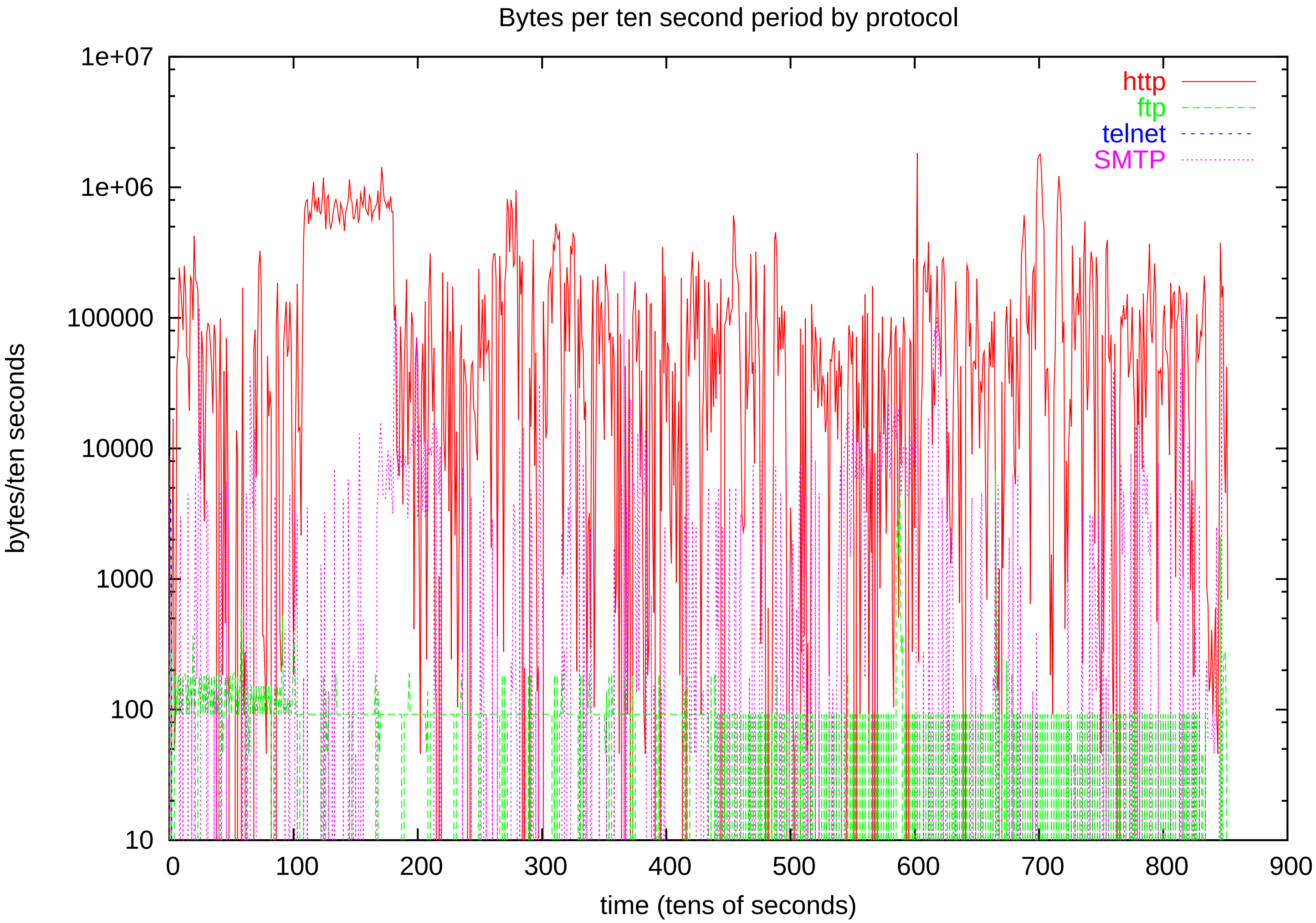
Total Packets Seen by Port



Total Bytes Seen by Port



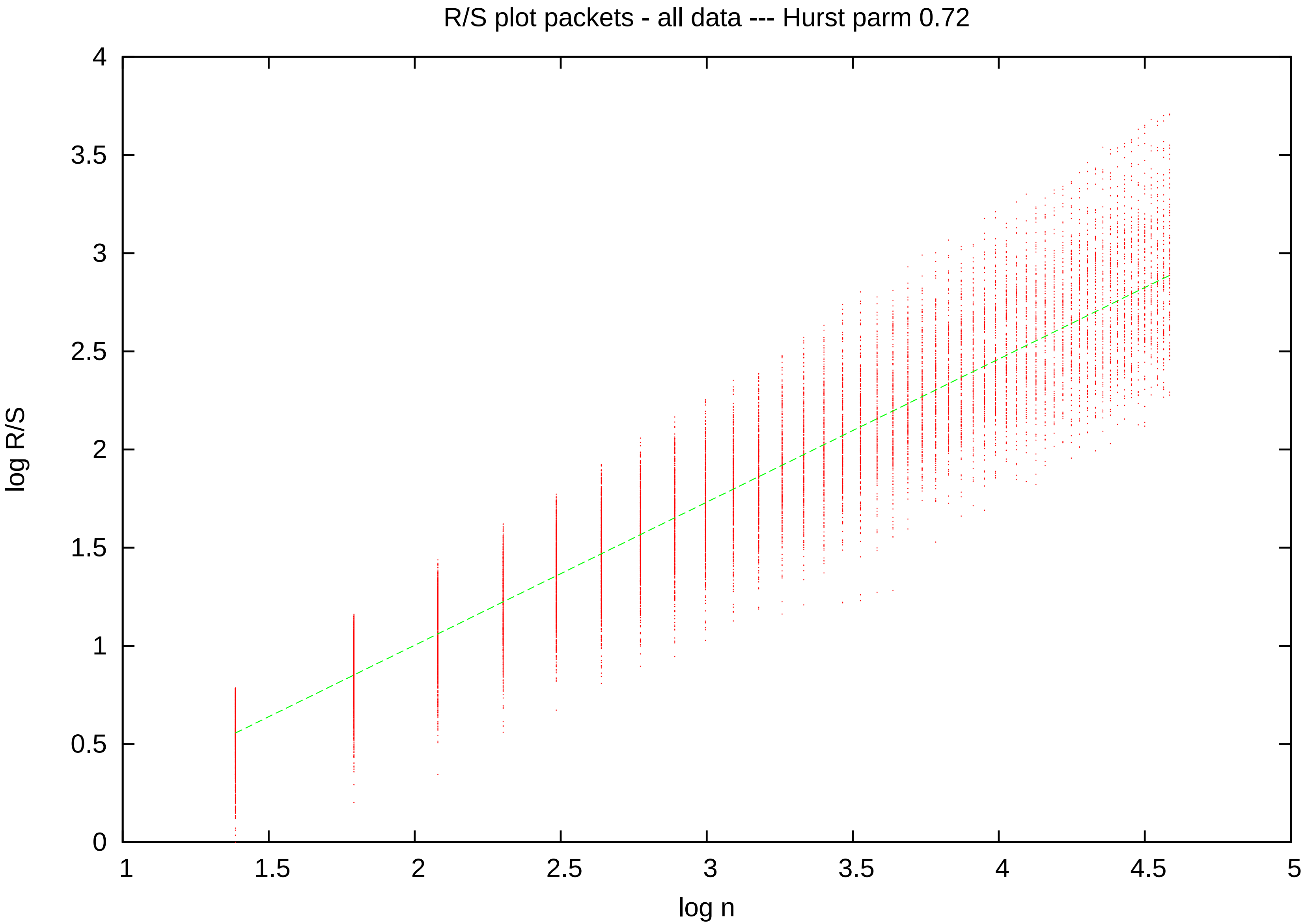
Total Bytes Seen by Port



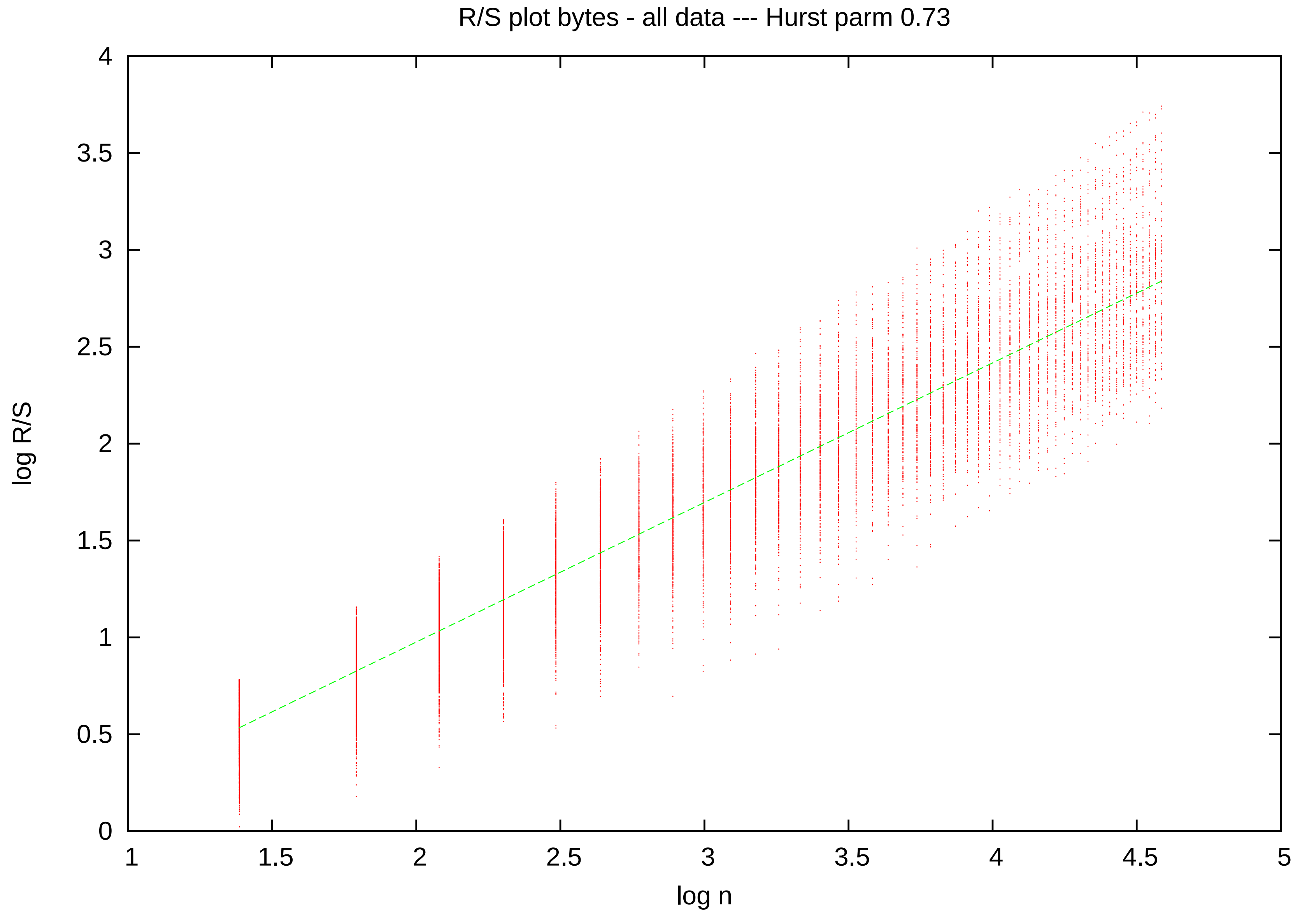
R/S plots

- The R/S plot has been used to estimate some Hurst parameters for the data.
- It is hoped that more sophisticated measures can be used later.
- For now the method is to take the standard R/S with series sizes from 4 to 64
- A plot of log n versus log R/S is given as usual.
- A least squares estimate of the Hurst parameter is given.
- We can also try disaggregating this data somewhat.

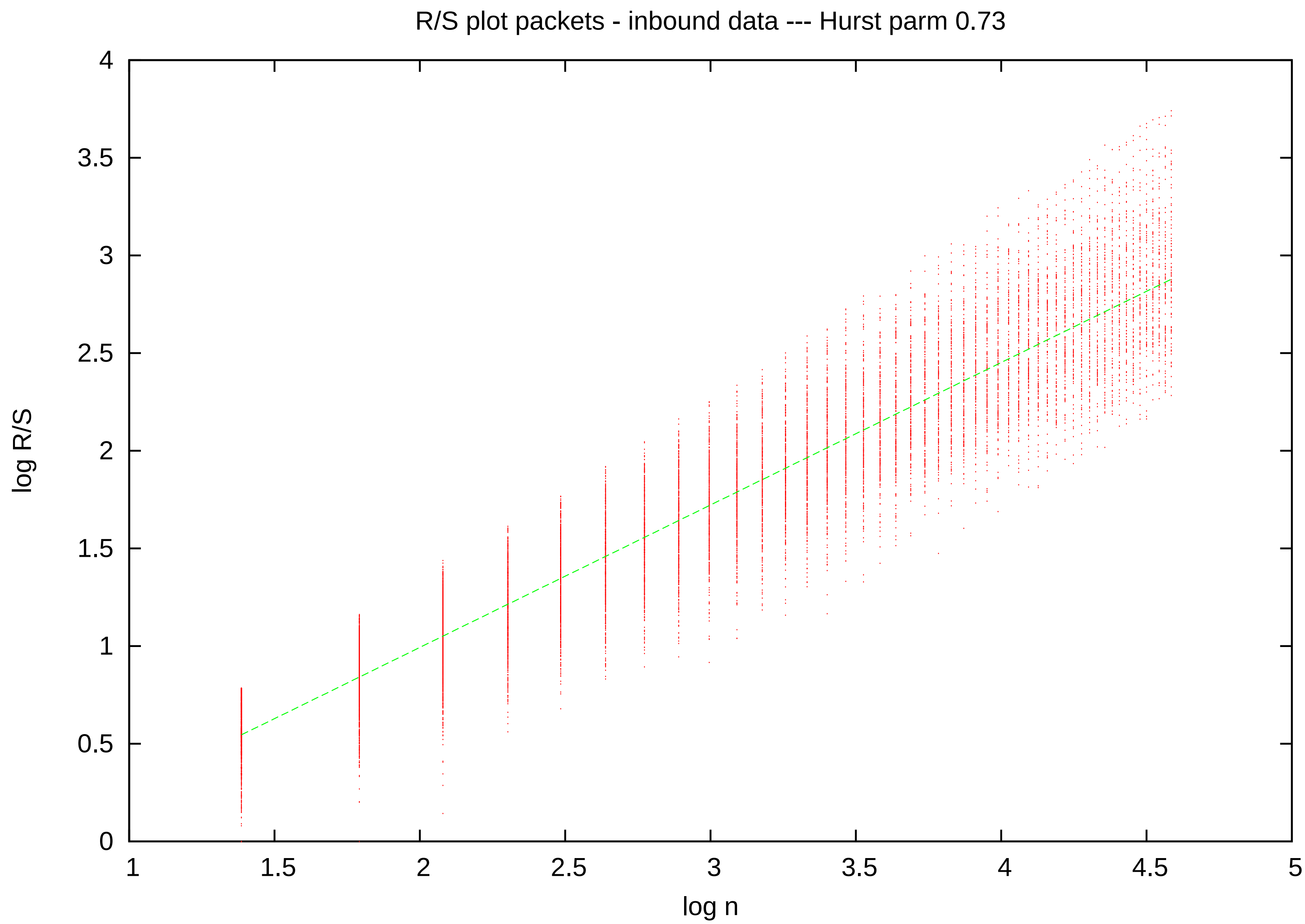
R/S data — All packets



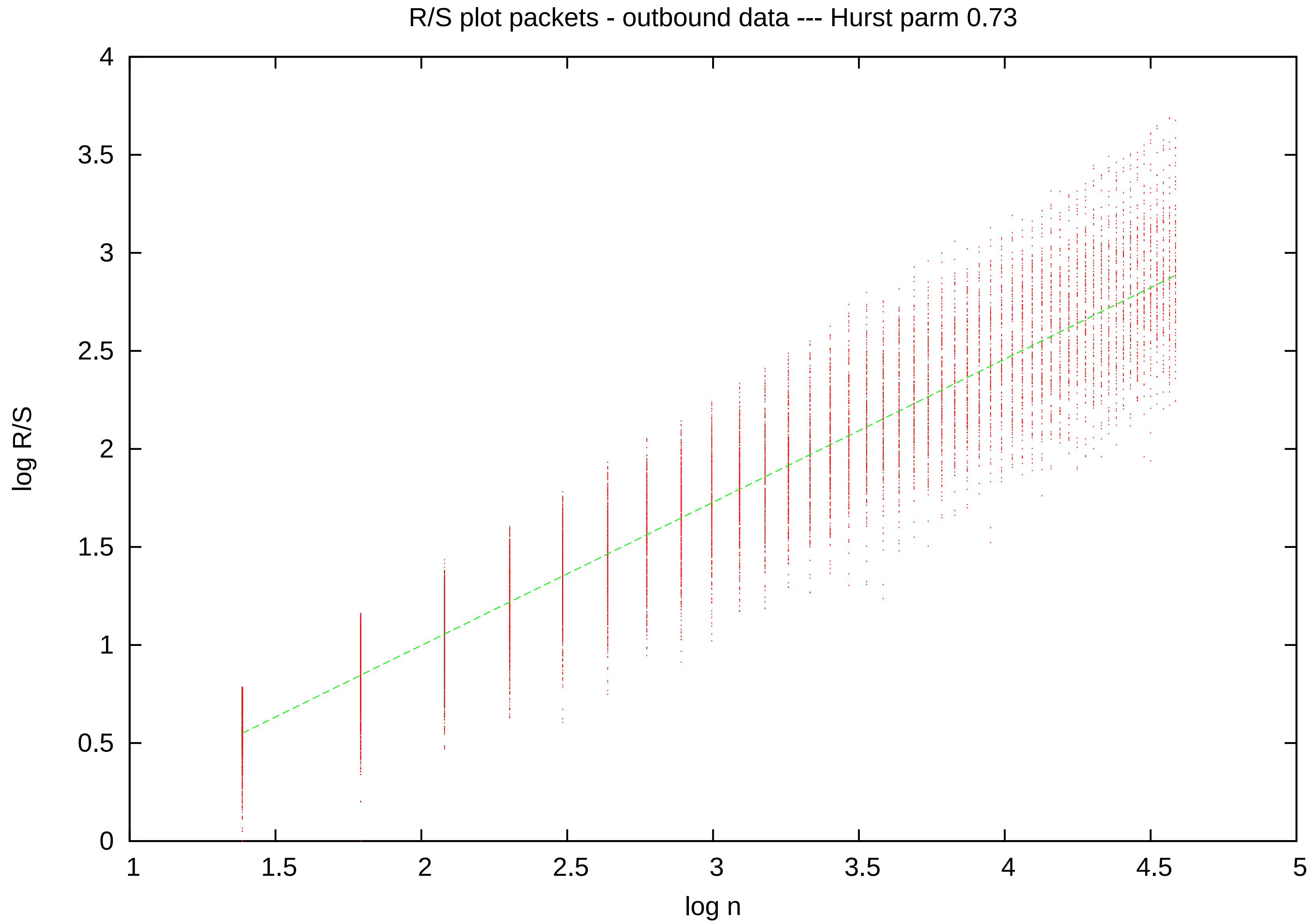
R/S data — All bytes



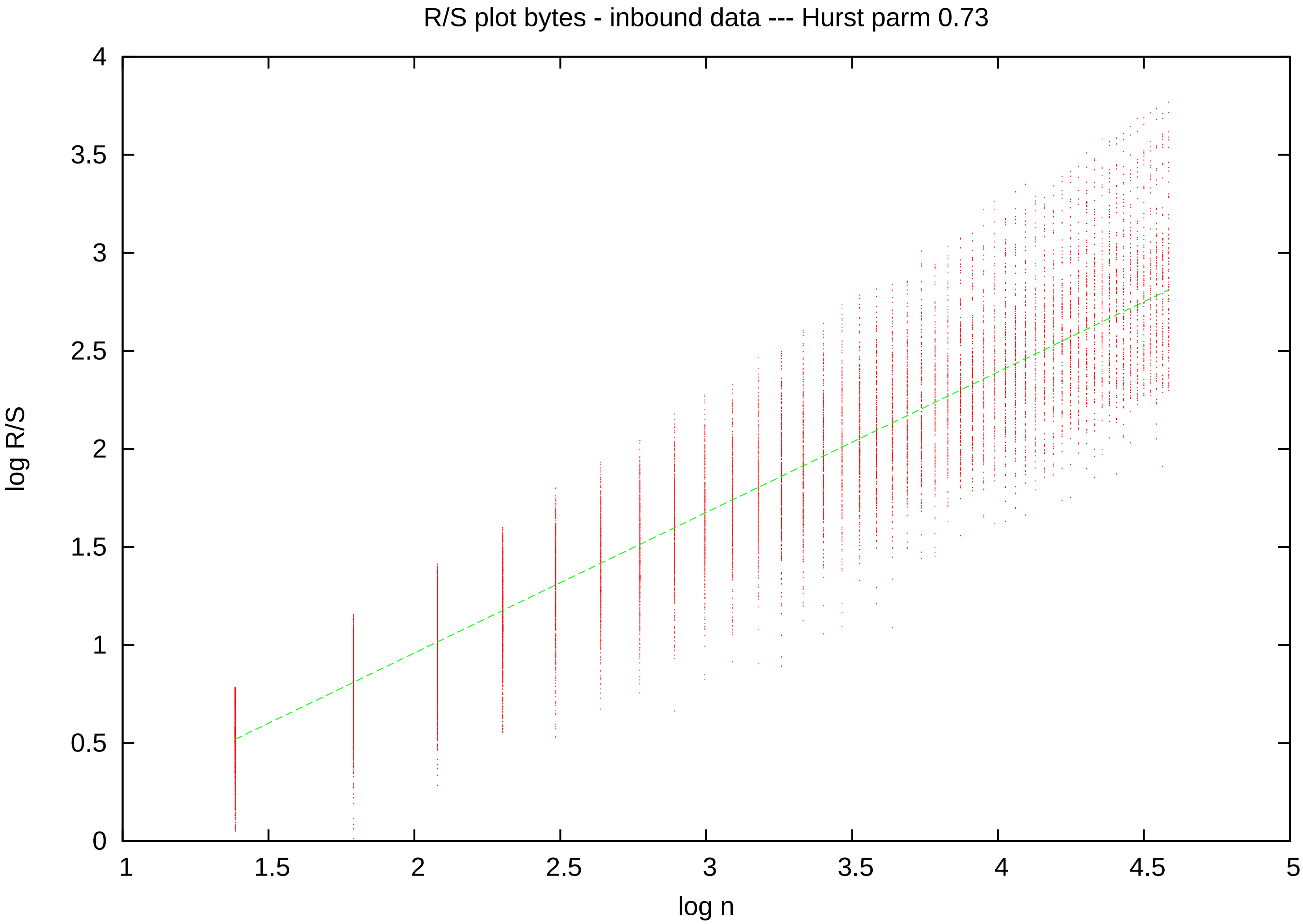
R/S data — Inbound packets



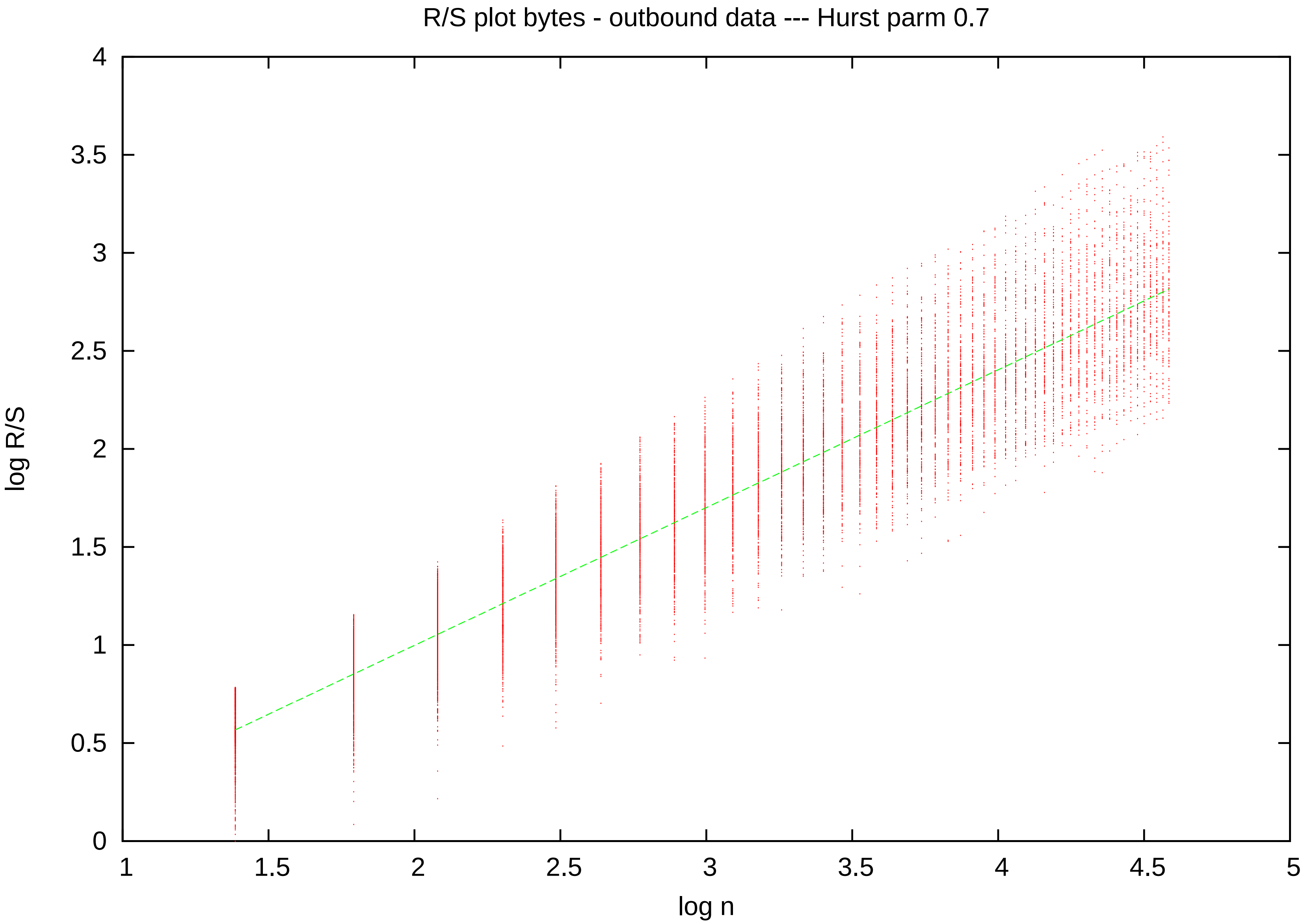
R/S data — Outbound packets



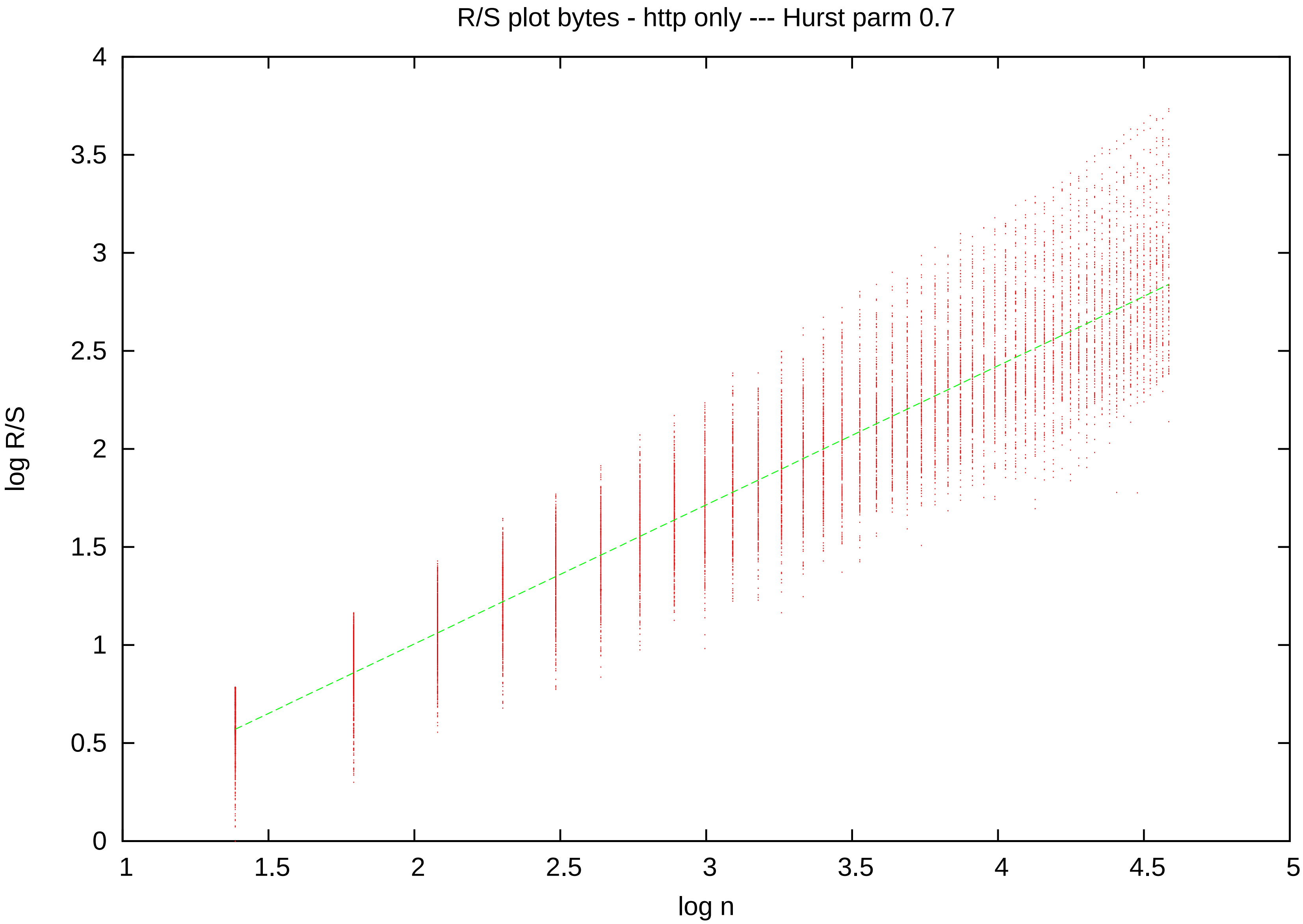
R/S data — Inbound bytes



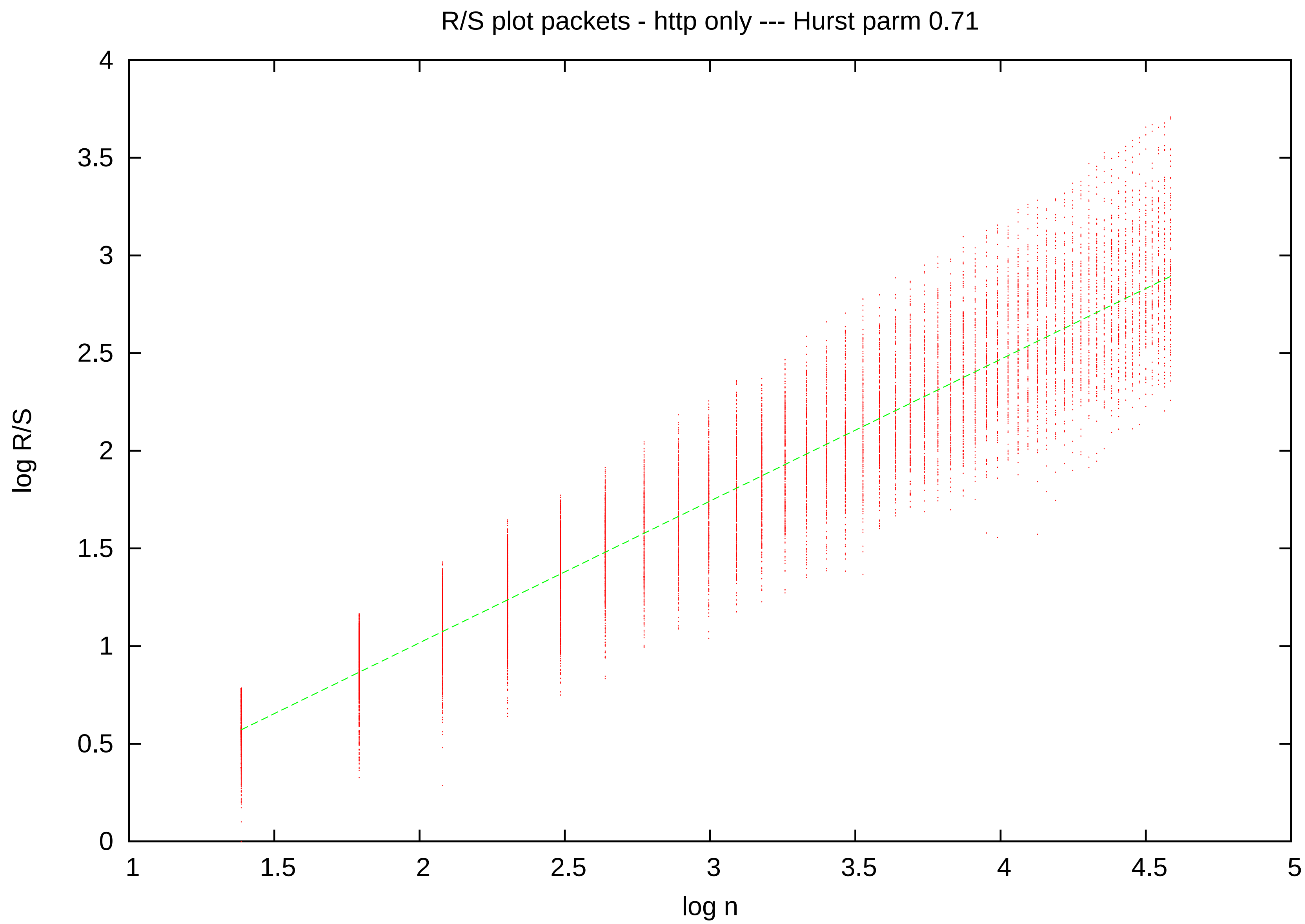
R/S data — Outbound bytes



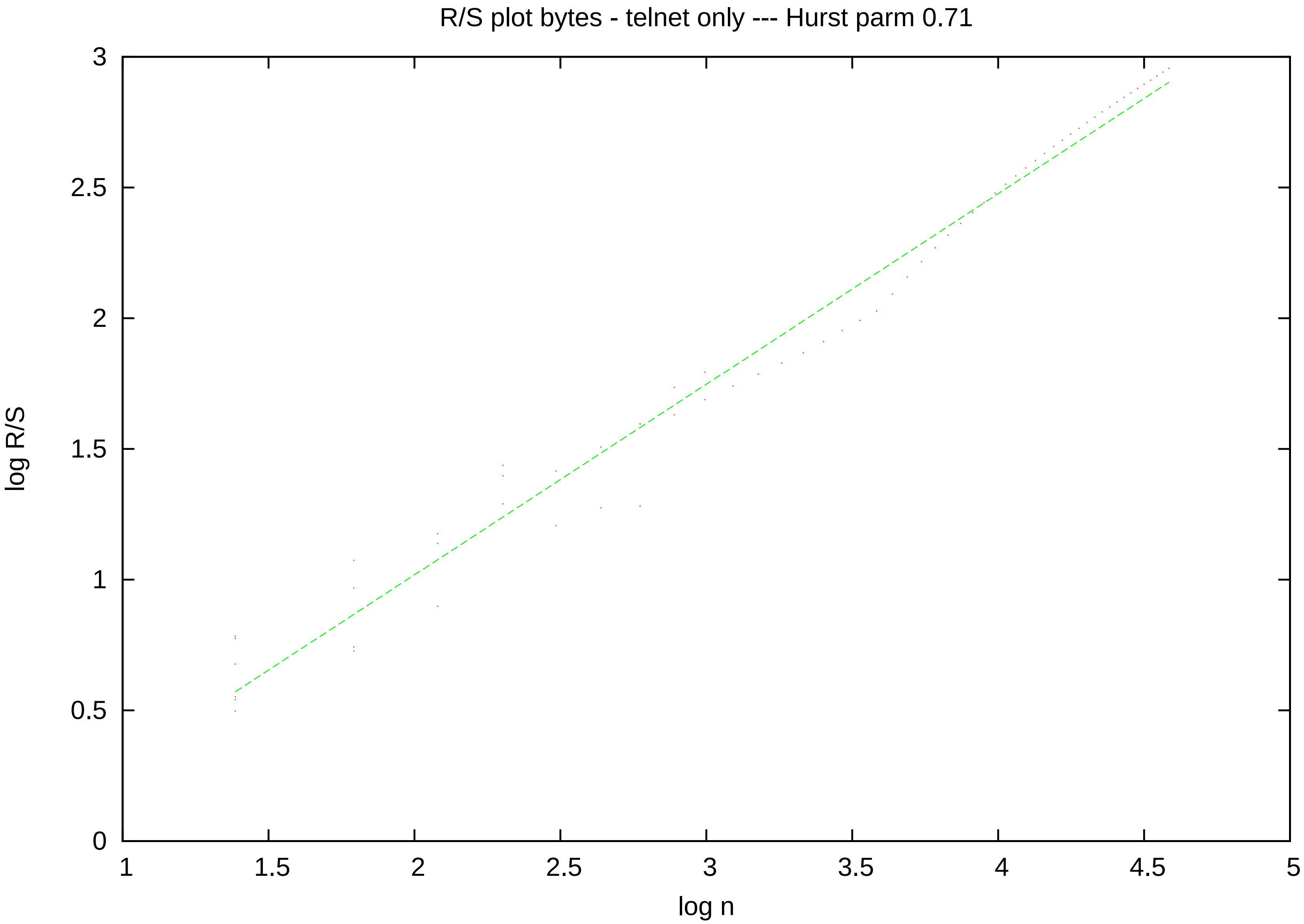
R/S data — http bytes



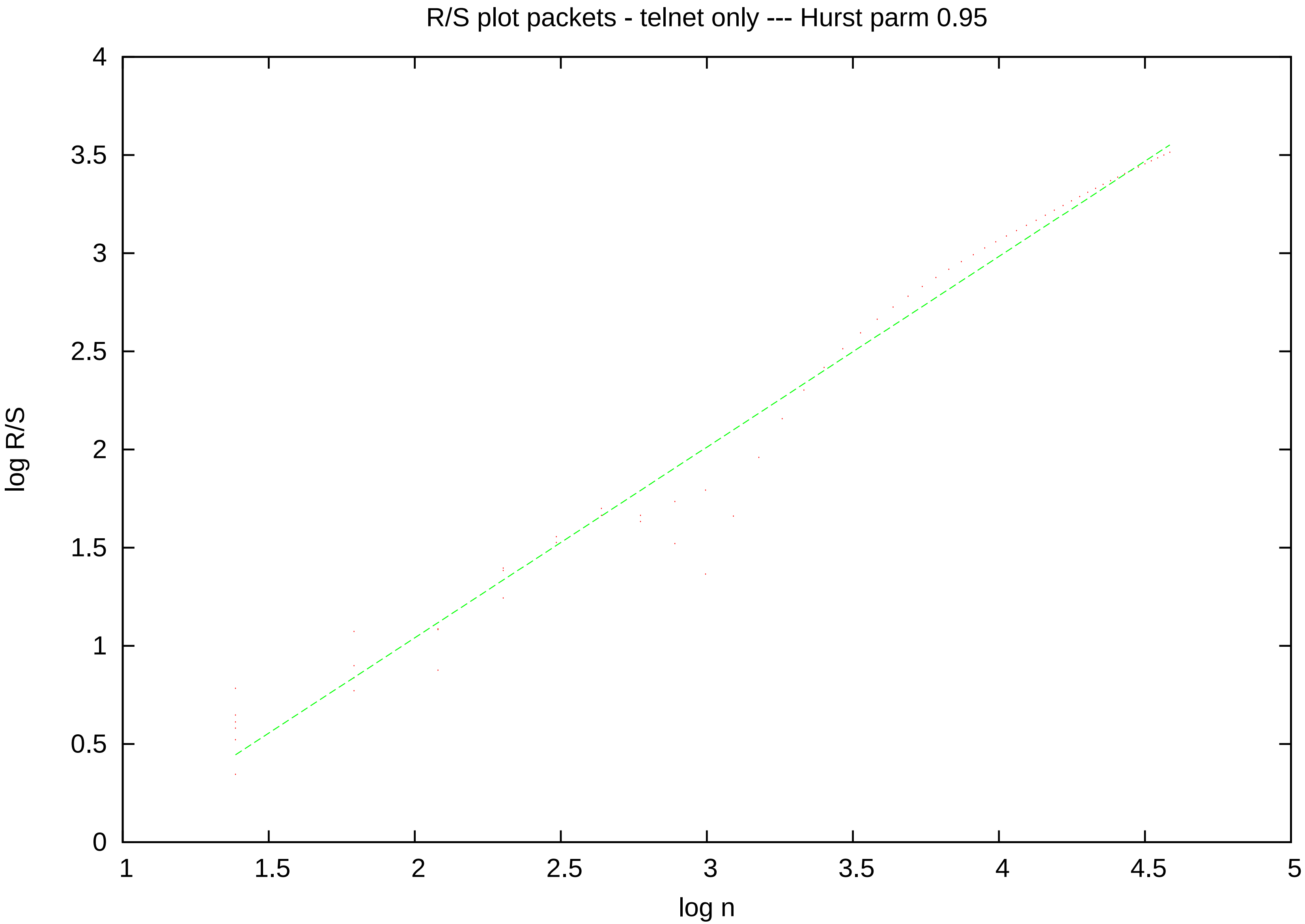
R/S data — http packets



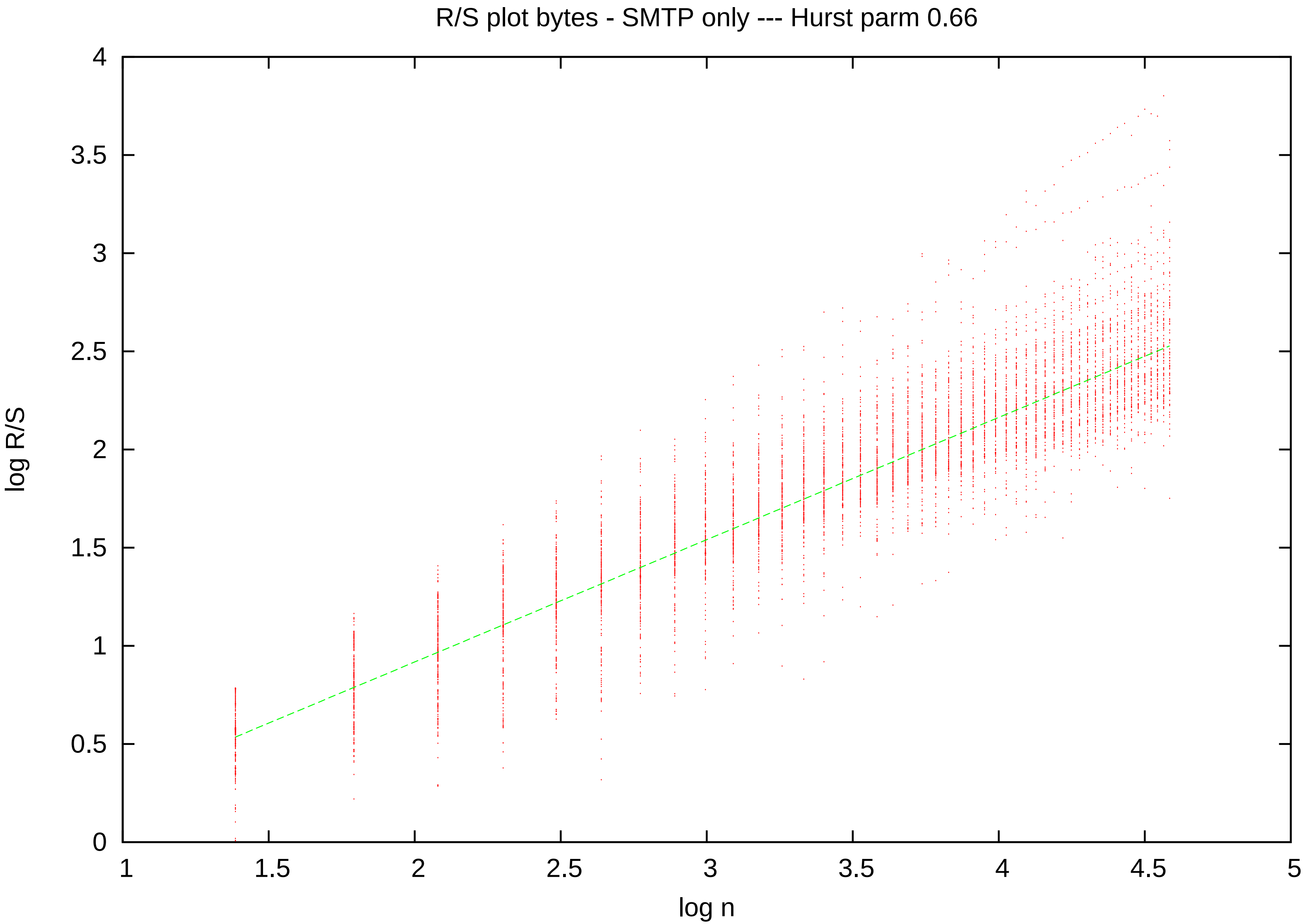
R/S data — telnet bytes



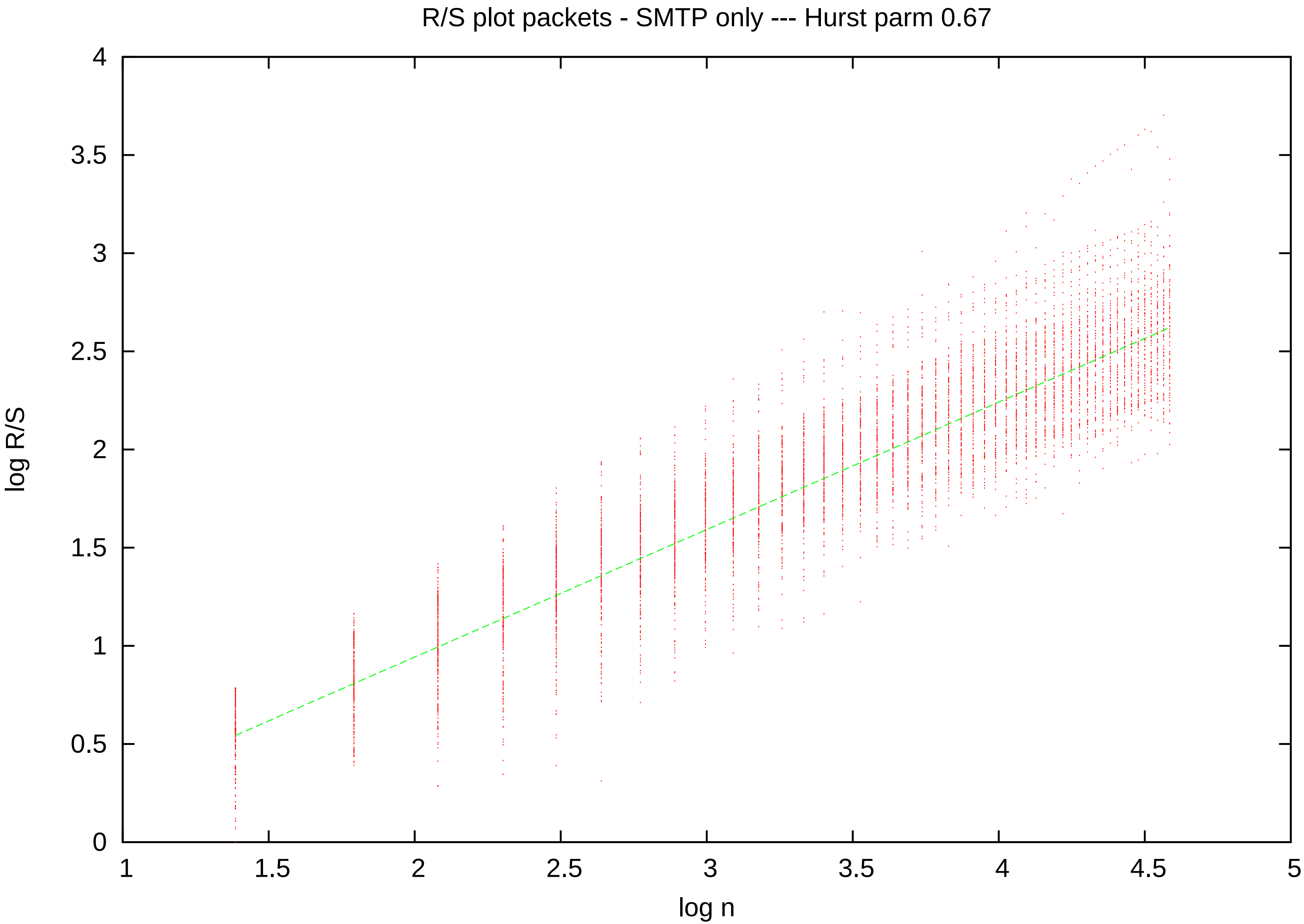
R/S data — telnet packets



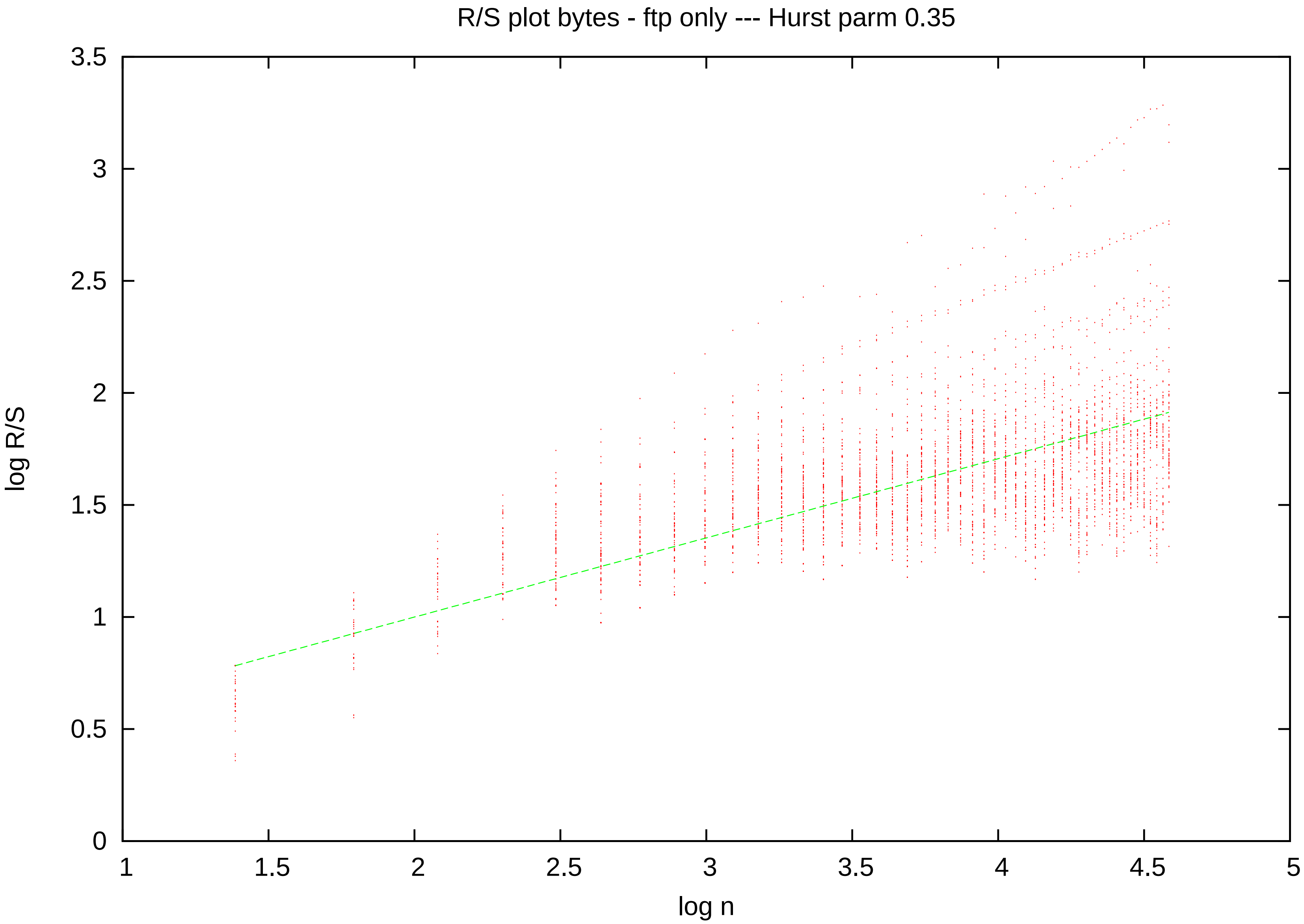
R/S data — SMTP bytes



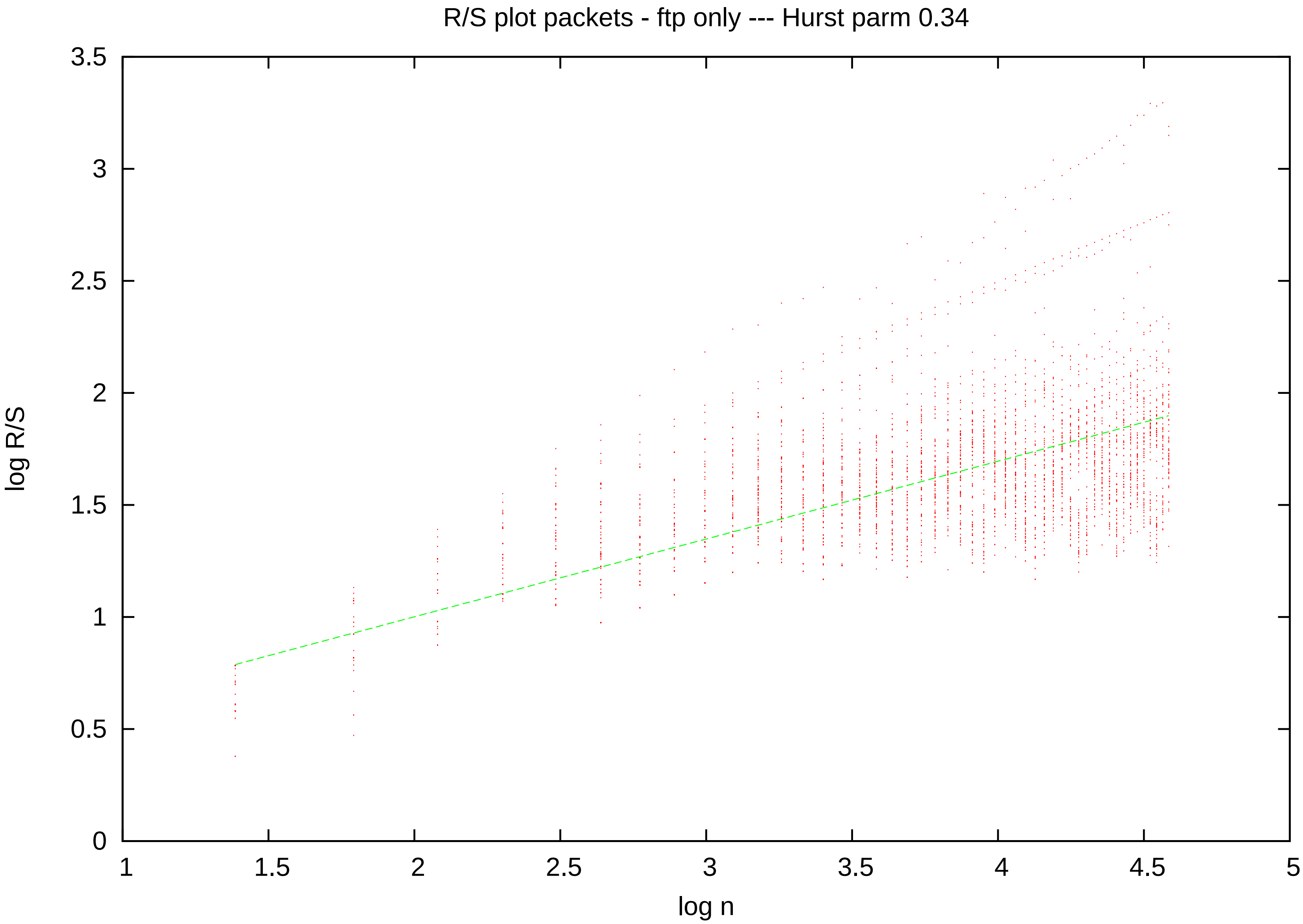
R/S data — SMTP packets



R/S data — ftp bytes



R/S data — ftp packets



Conclusions

- Automated tools have been created (in perl) which allow the simple creation of this type of graph from tcpdump data.
- The data analysis presented here is, obviously, preliminary and much remains to be done
- This data set is only partial — a larger data set is on the way soon.
- There are also important caveats about taking only part of a time series and performing standard time-series tests on it.
- Most of the results show little that is unexpected.
- However, the ftp results seems to be extremely unusual.