Time, Pollution and Maps

Michael Collins, CERT/NetSA

How’re we doing?

The basic cost: time

- Time to analyze
- Time to verify
- Time to retrack when we make mistakes

Basic success:

- In time t, x *things* happen
  - We understand > *x* in time *t*: good!
  - We understand < *x* in time *t:* bad!
- We’re probably at <<x right now

My (*work*) flow

Hypothesis

Look For Data

Remove Worms

Eliminate Proxies

Eliminate Scans

Restore Data I *thought* Was a Proxy

Discover exciting, but unrelated, new way the network "works"

Possibly get to original problem

CERT

Why Flow?

Ultimate cost: time

- Time = (storage) space

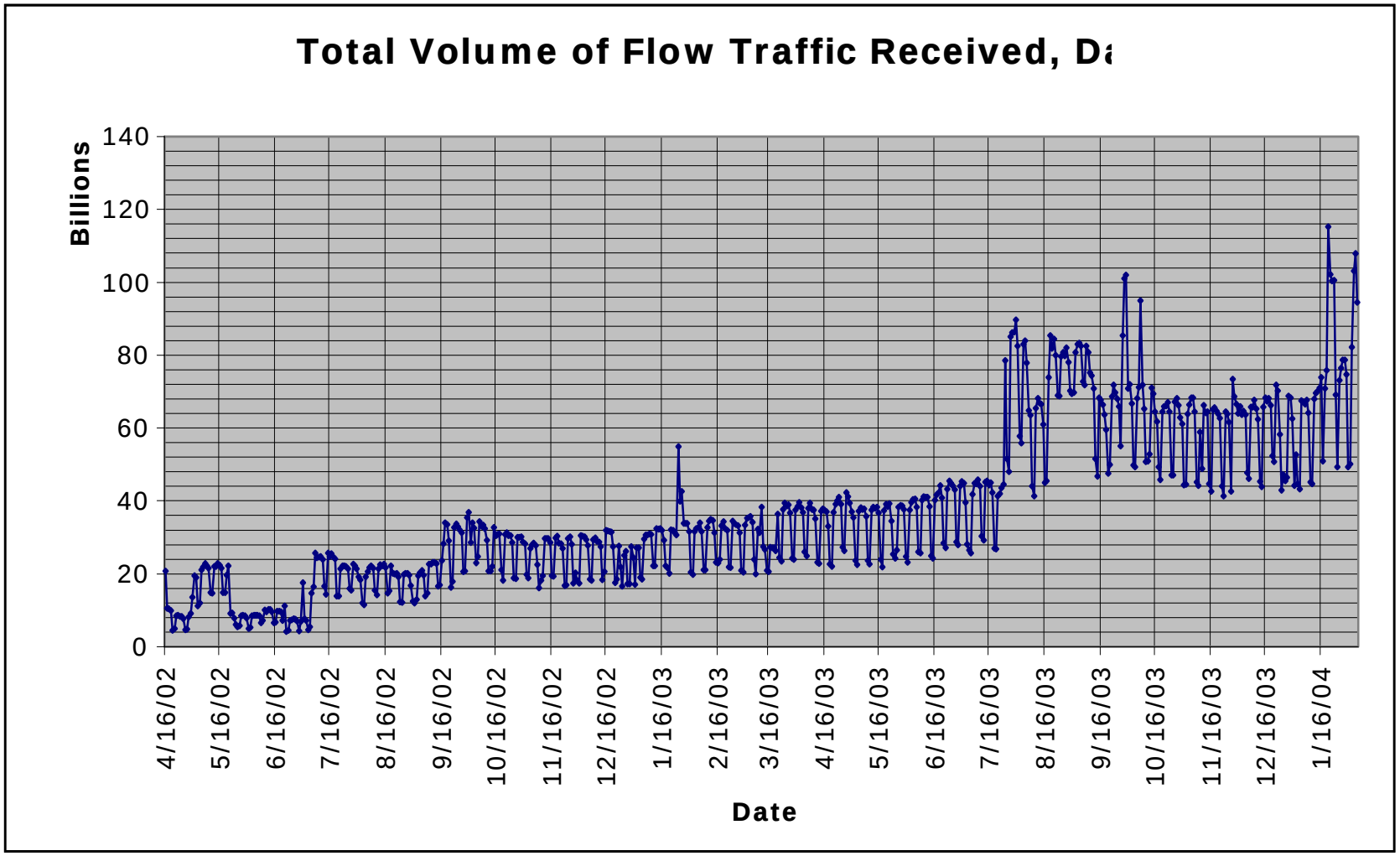
Basic issue - bang for the buck

- Catastrophe - the internet is regularly reconfigured, traffic volumes suddenly shift
- Pollution - approximately 70-80% of the TCP flows we see are not legitimate sessions

Flow is manageable where pure payload generally isn't

- I am looking at effectively *random* collections of packets
- Flow is the highest value information from a random collection of packets

Still have a basic problem



CERT

Manageable Additions

Adding additional flow information costs us

- Expression = field size = performance
- Additional data on disk should allows us to understand more *things*

Certain additions are going to come whether we like it or not

- IPv6
- Sasser

CERT

Expanding Flow Analysis

Fundamental Goal: *What's up?*

Secondary Goal: Don't break the bank

- Context
- Grouping
- Expansion

Context

Preserving knowledge of what's *on* the network

- Trickler
- Mapping - DNS, BGP, ICMP, etc.

Shouldn't have to repeatedly do ad hoc discovery

Maps *should* be smaller

Grouping

Annotating multiple flows together as one event

- Scan detection
- BitTorrent Distribution
- Websurfing

Don’t reconstruct this on a per-query basis

Expansion

Expand to *increase distinguishability*

- Increased time precision
- Some payload information

Try not to expand in order to identify *specific things*

- We *will* be attacked, any specific attack *implementation* is therefore of limited value

Concrete Suggestions

Heterogenous Splits:

- Full ICMP
- Short events
- Characteristics of payload
- Protocol validation

Conclusions

Our primary currency is time

- Time to access
- Time for backtracking
- Time for figuring out what the heck is going on

Time is equivalent to space

- Data on disk governs how long it takes to read information
- 10 billion events/day is about 2 DVDs/byte