The Past and Future of Flow Analysis

John McHugh

Canada Research Chair in Privacy and Security

Faculty of Computer Science

Dalhousie University

Halifax, NS, Canada

My-last-name at cs dot dal dot ca

Greetings from Canada

The anomaly of keynotes

- All of a sudden, I'm being asked to give keynotes at various conferences and workshops.

- I suspect that it has something to do with advancing age, approaching senility and the desire of the organizers to give the jet lagged participants an extra hour of sleep on the first day.

- Nonetheless, it gives me a chance to express opinions without having to back them up with facts that have to pass muster with reviewers.

Aphorisms

- (1) Pay attention to details. (2) Don't make stuff up.

  - Roy Maxion

- "She got the goldmine, I got the shaft"

  - Song by Jerry Reed

- Your mileage may vary

  - From the standard EPA disclamer

# Outline - 4 Themes in search of an author

- Parkinson's law

- Is it good for anything else?

- "Look homeward angel"

- Cast your net broadly

Parkinson's Law

- Parkinson's Law
  - Prof. Cyril Northcote Parkinson , 1958
  - **"WORK EXPANDS SO AS TO FILL THE TIME AVAILABLE FOR ITS COMPLETION"**
- The corollary is obvious, the answers are not.
  - Buy stock in your favorite disk vendor
  - Figure out how to break the law
  - Lets look a bit at the later

Distance and precision

- Look at the contact line distribution
- Are there regions where we can usefully abstract away the details?
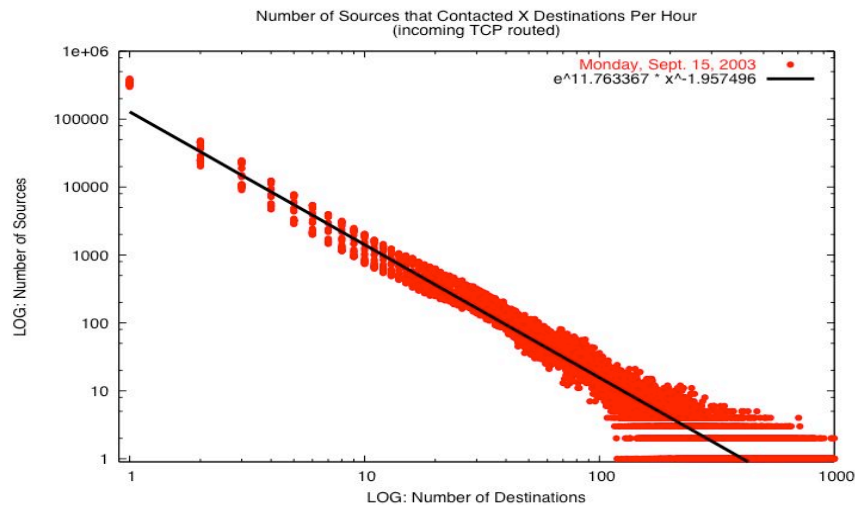
Infrequent and super frequent contact

- We see that there are a lot of sources that appear once per hour.

- There are a handfull that appear a lot.

- If we could represent these more efficiently, lots of space could be saved

- The underlying assumption is that there is not a single mechanism at work here, but the results of multiple mechanisms and they can be treated separately.

Singletons

- We haven't really looked too closely at these since there are millions per hour, but.

  - The majority appear to form a small number of groups based on protocol, ports, flag combinations, etc.

  - I suspect that most are addressed to unoccupied addresses

  - Looking back at data that is months or years old, do you really care what the individual target was or when in the hour the flow occurred?

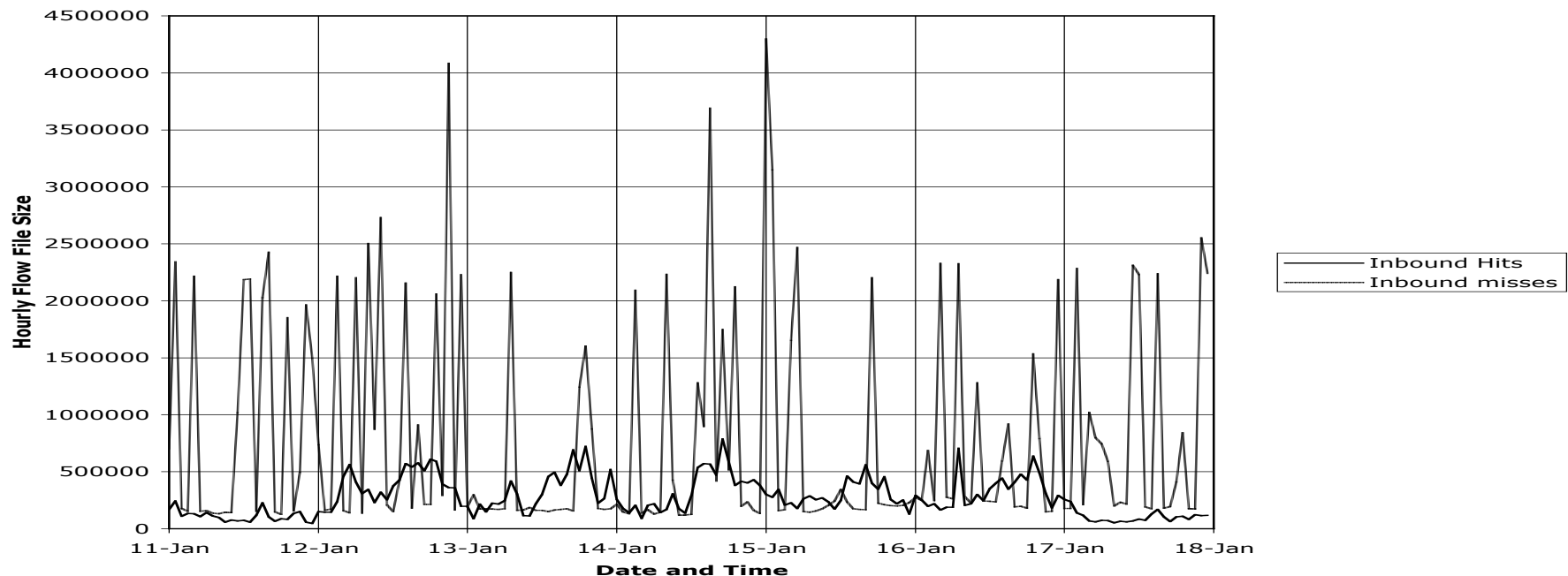 - If not, some lossy compressions are obvious

More singletons

- Suppose we look at the one/hours over a week
  - Again, the majority are never seen again
  - This reinforces the earlier discussion
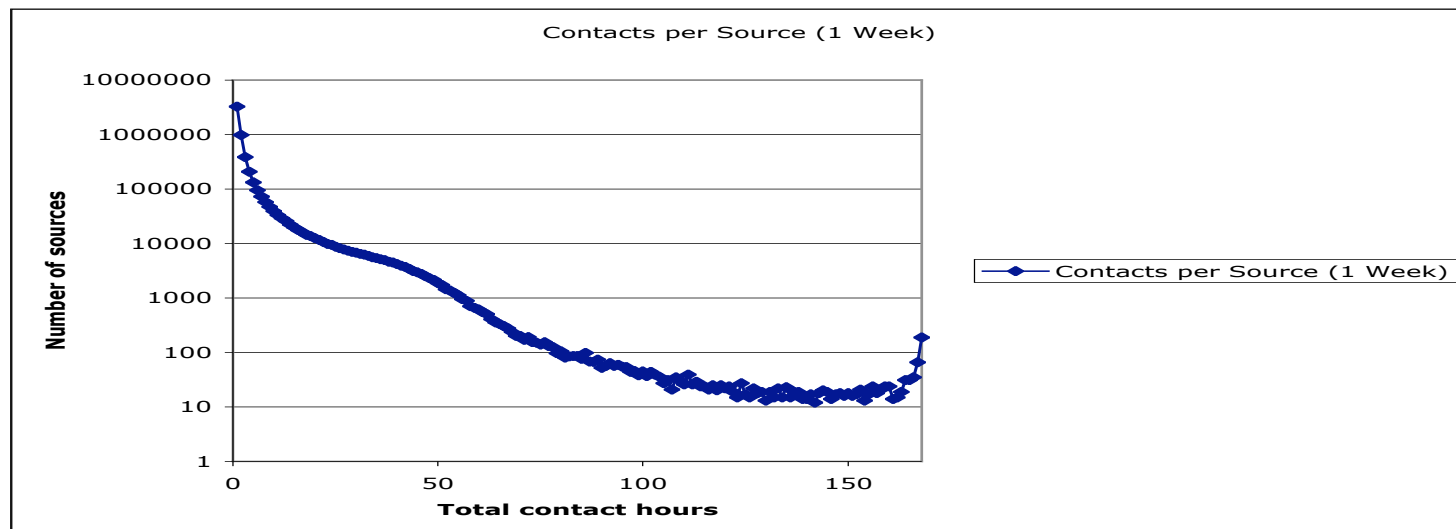
Hyperactive sources

- Most of these are high volume scanners.
- Most of the scans do not find targets

Typical scan

- Looking at a single spike, one often finds things like:

```
            sIP|dPort|pro|packets|   bytes|   flags|
AA.BB.x.y|    80| 6 |        2|      88| S      |  96842
AA.BB.x.y|    80| 6 |        1|      44| S      |  20233
AAA.BBB.0.0/16 : 51744 hosts in 228 /24s and 1785 /27s.
```

- Again, there are obvious compressions that trade precision for volume

- Note that backscatter is another source of hyperactivity, but similar processing applies

Other tricks

- I suspect that treating low levels of activity and high levels of activity will reduce the volume of the remaining data by 75 to 95 % (YMMV).

- At that point, further processing of the main stream might be useful
  - Session reconstruction comes to mind
  - Aggregation of similar sessions
  - Etc.

When is not a flow, a flow?

- Reduced to its essentials, the SiLK paradigm is a way of thinking about intermingled time series that are glued together by common factors.
  - Think of the paradigm as a way of reasoning about large aggregations of event data.
  - Convert the data into pseudo flow records and have at it.
  - The essentials are some sort of anchors (IP addresses), possibly some volumes (1 is a perfectly good default), and maybe some other stuff (tags, flags, ports, etc.)

Packets are obvious

- For a DARPA project I worked on at CERT, it was extremely useful to apply SiLK style filtering to packet data.

- My vision was "rwpfilter", a program that would bring the power of rwfilter to packet data and extract pass/fail files containing both packets and degenerate (1/pkt) flows. For packets,it would be possible to filter payload with regular expressions

- A prototype was built, but it has been an uphill battle to bring this into the main stream. I have a MS student who is going to try, and the latest rwptoflow starts to be a building block.

Other issues

- DHCP
    - Especially in wireless networks
- NAT
- Sensor identification issues, etc.
- Inside to outside distinction and grouping
    - In -> Out
    - Out -> In
    - In -> In
    - Out -> Out ???

Crawdad / predict / etc.

- Dartmouth has a repository of 4 months of pkt headers from their wireless network
  - 160GB compressed / 18 sensors
  - Anonymized (badly)
  - 17000 MAC addresses also badly anonymized
- Converted to hourly flows in early 2006 and used in my course.
- Students were able to identify a number of interesting things including several worm infestations, but …
- DHS may bring the predict archive on line soon
  - This is a potential source of data, but … YMMV

Other sources

- With a text to flow program and a few text processing scripts, the possibilities are interesting.
  - Log data from firewalls, IDS, etc.
  - Pilot study involving data from a large managed services company looked promising and could produce "top N" reports easily.

Generalize, Generalize, Generalize

- Extend SiLK tools for filtering, set and bag production, etc. to all scalar fields in the archive format.

  - Sets of sensors make sense

  - So do sets of flags, etc.

  - In a strange way, so do sets of times or durations.

  - I experimented with Bloom filters for detecting connection level service activity (SIP, DIP, service)

Plagiarize, Plagiarize, Plagiarize,
Only, please, call it research

- Support set and bag structures, along with hash trees, Bloom filters, perfect hashes, etc. at a library level for specialized analyses.

- The dynamic library concept allows a lot of creative use of the concepts. Make it easy on the researcher analyst to use the bits and pieces of the tools

  - Internals documentation?

  - Builders guide?

  - Skeleton programs?

Know thy network!

- Much of the SiLK work at CERT has been focused on border data from a large network.

- There is growing evidence that continuous monitoring at the small enterprise level is useful for all sorts of things, including security, but also for provisioning.

- You will see two papers later in the workshop on one such effort performed by Ron McLeod.

- Even if there are no large scale compromises, such monitoring often provides interesting insights
  - Why do both of the cases in which we monitored a cable modem show 95% ARP?

More Data, More Data …

- The sources represented at this meeting are largely big institutional networks.

- Several people have recently suggested a co-op approach to data collection and sharing.

  - Eurecom has a honeypot co-op. Contribute by running their honeypot system and you get access to all the reported data.

  - Anyone for a flow co-op from your home DSL / cable modem drop?

  - There are a number of collectors and ideas for better ones …

Thanks

- Tim for inviting me.

- SLK for seizing the moment

- Tom Longstaff for creating the environment that made it all possible

- Mike for keeping the spirit alive in trying times

- All the developers for their support, understanding and willingness to take suggestions.

- QUESTIONS …

If it ain't broke - Don't IPFIX it.