Traffic Classification

a practical perspective

Matthew Roughan
<matthew.roughan@adelaide.edu.au>

School of Mathematical Sciences
University of Adelaide

Data is key

THE UNIVERSITY OF ADELAIDE AUSTRALIA

- Data is the key understanding traffic
- Data is the key to good models
- Data is the key to prediction/planning, anomaly detection, traffic engineering, ...
- Good data is hard to get

A few problems

THE UNIVERSITY OF ADELAIDE AUSTRALIA

You want to model an application

- how do you select this application from the traffic

You want to detect policy violations

- detecting a particular application
- maybe it's hidden somehow

Class of Service (CoS)

- operators don't always know what's on their network
  - mergers and acquisitions
  - global companies; local IT dept.s
  - easy proliferation of applications
- how can they provide differential CoS

Classification

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Why not use TCP port

- ports not defined for all app.s (e.g. Kazaa)
- application uses non-standard port (e.g. port 8080)
- incorrect implementations (e.g. old `bind`)
- ambiguities in port registrations (e.g. port 888)
- dynamically allocated ports
- firewall hopping (e.g. use of port 80)
- security attacks violate port conventions

Other methods

- specific signatures:
  - e.g. look for a particular string in the application header
  - e.g. particular malformation of packet
- statistical/machine learning/data mining
  - use features of flows to classify them
  - e.g. average packet size, flow length, ...

Pros/Cons

- specific signatures:
  - can be 100% accurate
  - but must know signature
    - requires pre-knowledge of traffic/attack
    - often manually intensive
- statistical/machine learning/data mining
  - machine-learning can be automated
  - machine-learning can detect previously unknown attacks
  - typically, give up some accuracy

. . . but wait

- What do we mean by an application anyway?
  - e.g. `ssh`
    - I can use this for interactive session
    - or bulk-data transfer `scp`
    - both use the same port, and often the same software
  - e.g. Lotus Notes
    - used for email (bulk data)
    - used for database transactions (interactive)
- There are different **use cases**
- Does this matter?
  - for detecting policy violations, or CoS it does

. . . but wait

- Sometimes, different software is used for the same purpose
  - e.g. `POP` and `IMAP`
  - e.g. `http` and `https`
- Does this matter?
  - for detecting policy violations, or CoS it does

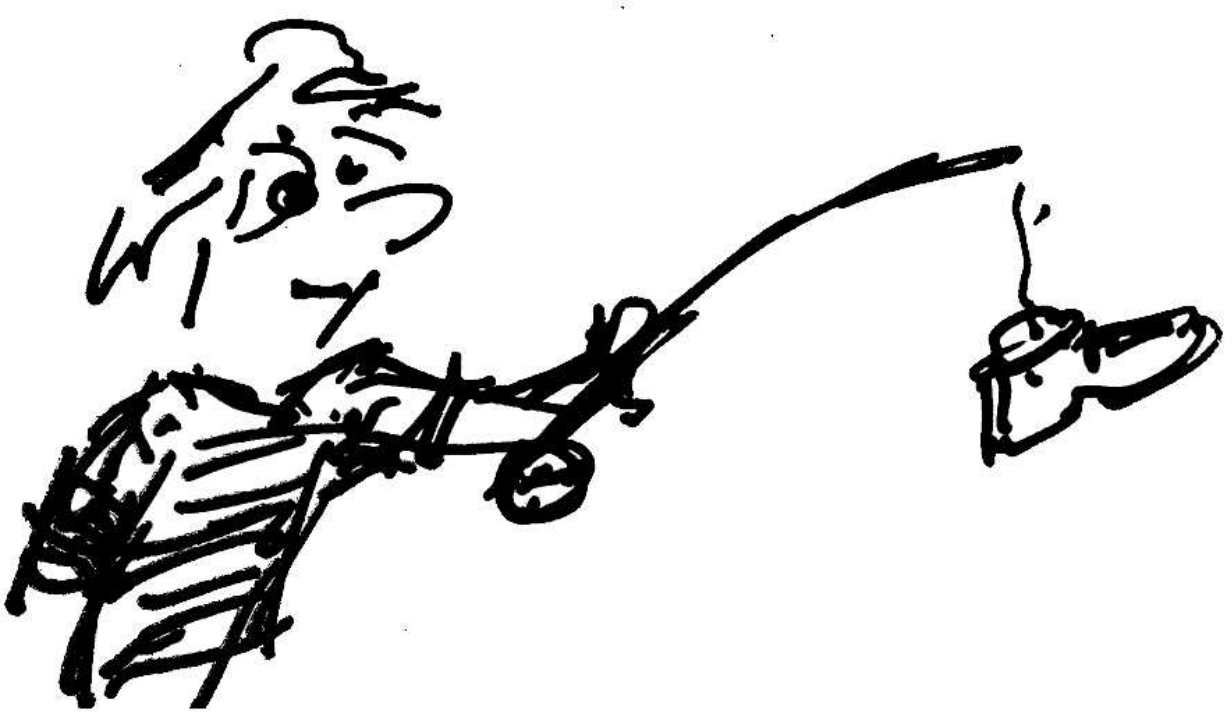
Practical focus

THE UNIVERSITY OF ADELAIDE AUSTRALIA

- it depends on what you are trying to do
  - look for policy violations
    - policy: no interactive logins, but `scp` is OK
    - how could I detect this being violated?
  - CoS mapping
    - interactive services get low-delay, bulk data gets high-throughput
    - don't care about individual applications
    - do care about use cases
- I would say that measuring success of an approach can't be based on how well it gets the "application" right
- the success depends on what you **really** need to do

Fishing can be a waste of time

THE UNIVERSITY OF ADELAIDE AUSTRALIA



Ah! Now I know the true distribution of shoe types in the river Torrens!

Conclusion

- I would like to see results more focussed on the practical outcomes
- classifications that are useful can have their performance quantified by how well they succeed, not just an "accuracy metric"
- and of course, we need more annotated public datasets