

SWIN
BUR
NE

SWINBURNE
UNIVERSITY OF
TECHNOLOGY

NOSSDAV'09

Rapid Identification of Skype Traffic Flows

Philip Branch, Amiel Heyde, Grenville Armitage

Centre for Advanced Internet Architectures,
Swinburne University of Technology,
Melbourne Australia

Outline

- Why we are interested in rapid identification of Skype Traffic
- Skype overview
- Using Machine Learning to identify Skype traffic
- Results
- Future work

Skype

- A popular, proprietary VoIP application (also IM and video)
- Operators claim over 400 million users
- Skype is peer-to-peer
- Communication is encrypted
- Details of its protocol structure are secret
 - Lots of obfuscation but some sleuthing has identified how it probably works

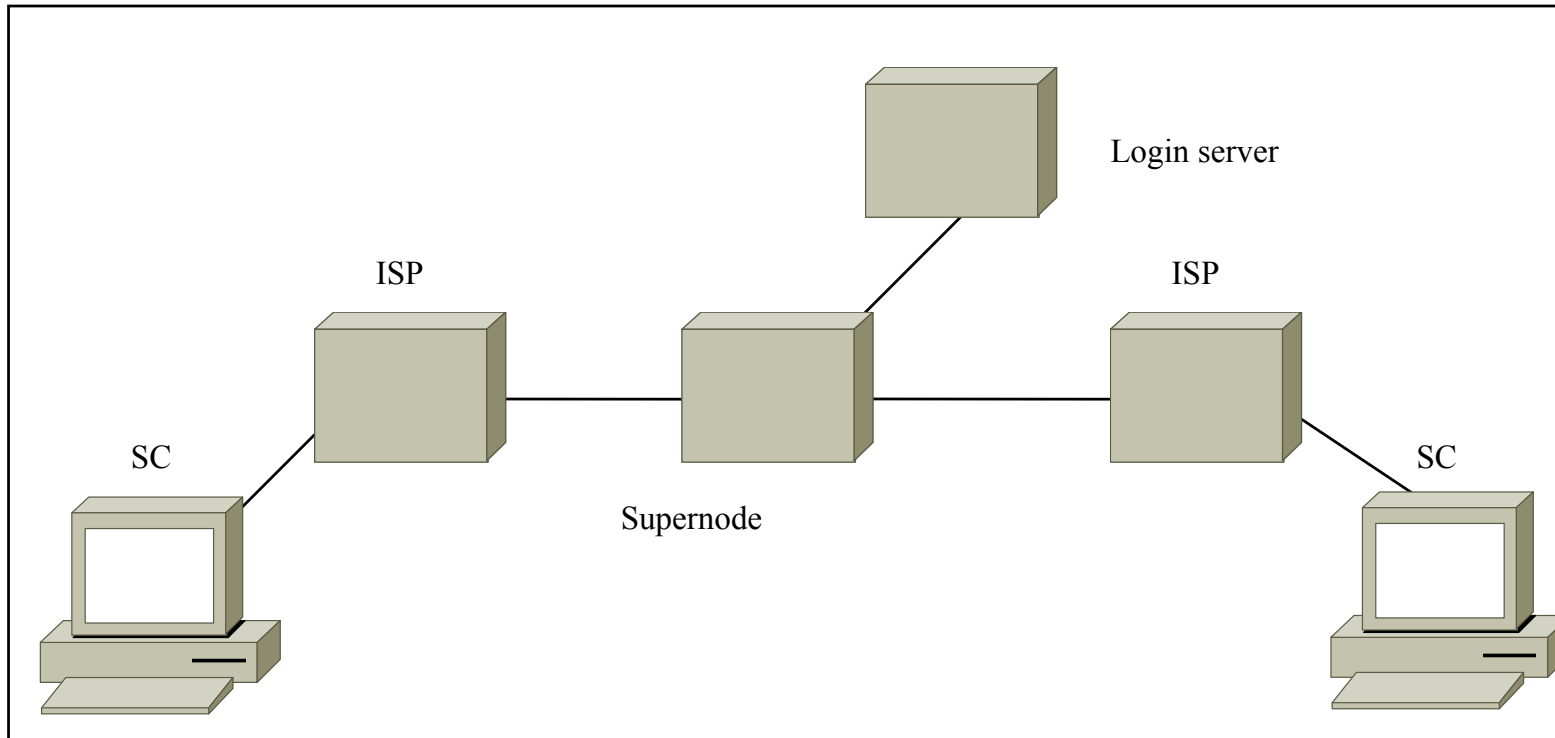
Skype

- Can use a number of codecs but default is Sinusoidal Voice Over Packet Coder (SVOPC)
 - Variable bit rate coder that adapts bit rate to whoever is talking and available bandwidth
- Skype attempts to use UDP for voice transport
 - Variant of STUN protocol used where NAT intervenes
 - But if UDP not possible will use TCP as a transport protocol
 - Uses ‘supernodes’ as relays

Skype signaling

- Supernodes
 - Any host with a public IP address and adequate free bandwidth
 - Used in call set up and as a TCP relay if UDP cannot be used as a transport protocol
- Call set up is via supernodes
 - Client connects to a supernode
 - Supernode relays authentication information to and from a Skype login server

Skype signaling



Why we are interested in rapid identification of Skype Traffic

- Lawful Interception
 - In most jurisdictions Law Enforcement Agencies can request communications be intercepted and delivered to them
 - “Wiretapping”, “Phonetapping”
- Telecommunications companies are interested (and concerned!) about the effect of Skype on their business
 - Questions about the extent of use of Skype
 - 400 million users? A lot of lost revenue
 - Carriers need to know how much revenue they are losing

Why we are interested in rapid identification of Skype Traffic

- Identifying Skype traffic means the ISP can treat it differently to other traffic
 - In a good way:
 - give it priority over non-realtime traffic in the network
 - In a bad way:
 - to discourage use of it, give it lower priority than other traffic
- Security applications
 - Skype can (is? might?) be used to control a botnet
 - Network of zombies used for DDoS or spam
- Regardless of motivation there is a lot of interest in identifying Skype flows

Lawful Interception

- Our interest driven by Interception obligations of ISPs
 - ISPs are obliged to intercept communications to or from a person of interest if instructed by a Law Enforcement Agency (LEA) to do so
- Law Enforcement Agencies particularly concerned with voice interception
 - Email and http can look at email relays or http logs (or proxy logs)
 - Voice much more ephemeral
- Would like to make a decision on whether or not to intercept in real time based on whether voice or not

Lawful Interception

- Skype is particularly difficult to intercept
 - Linkage of identity to IP address
 - Messy DHCP and RADIUS interactions
 - Peer-to-peer
 - No server that can act as an Interception Access Point
 - Login information not contained in any one server and probably not located within the LEA's jurisdiction
 - Some speculation that US government and Skype have an agreement regarding access to login information
 - Communication is encrypted
 - Not as big a problem as it may seem. Most intercepts require meta-data only. Who is communicating with whom, rather than the content of their communication.

Our research question

- Skype identification has applications for LI, Security, Traffic management and economics of voice networks
- Can we identify Skype VoIP traffic quickly AND Can we identify Skype traffic based on observing only part of a flow?
 - Goal is to identify Skype traffic flows based on its traffic characteristics such as packet length and interarrival times
 - “Quickly” means a few seconds
 - Do not want to be reliant on capturing particular parts of a flow
 - Do not want to have to see the entire flow

Identifying Skype traffic in realtime

- Skype traffic characteristics vary greatly over the life of a flow
 - Signaling traffic quite different to voice traffic
 - Initial period of additive increase where Skype appears to be finding capacity limits of channel
 - Silent and active periods
- We have found that multiple attributes are most effective in identifying a Skype flow
 - Require a fusion of multiple attributes to identify Skype traffic
 - An ideal application of machine learning

Machine learning

- Goal of machine learning is to produce a traffic classifier
 - Based on some attributes classify the traffic flow into either “Skype” or “non-Skype”
- Classifier can be a supervised or unsupervised
 - Unsupervised classifier maps similar objects to the same region of vector space
 - Supervised classifier is presented with multiple examples and learns to map certain combinations of features to specific classes

Supervised machine learning

- Many different types of supervised classifiers
 - Decision trees, Neural networks, Bayesian...
- Constructing a classifier
 - Identify characteristics (features) that might be suitable for classifying the traffic
 - Train the classifier to associate certain combinations of features with a particular class of flows
 - Once trained test the classifier on unseen flows

Performance measures of classifiers

- Subject to two types of errors
 - False positives
 - Flows incorrectly identified as Skype flows
 - False Negatives
 - Flows incorrectly identified as not Skype flows
- Two common measures are recall and precision

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

Experimental method

- Each flow was segmented using a sliding window into a number of subflows corresponding to a certain duration (the window size)
- A classifier was produced for each window size and its performance tested
 - We trained with one data set and tested with the other and then swapped.
- We made use of the WEKA package integrated into the NetAI package

Experimental method

- We used a rule based classifier constructed using the C4.3 algorithm
- Skype data was obtained from 18 calls across multi-hop public networks. Approximately 7 hours, 57 MB of Skype traffic, approximately 710,000 packets
- Non Skype data was obtained from the University of Twente's traces (503 MB and 752,000 flows) and our own Online game traces (~100 MB)
- Data was captured using tcpdump, and then edited into WEKA format

Sample WEKA file

```
@relation large-packet-lengths

@attribute minimum      {integer}
@attribute maximum      {integer}
@attribute median        {integer}
@attribute lower-quartile {integer}
@attribute upper-quartile {integer}
@attribute mean          {real}
@attribute stdev         {real}

@attribute skype-traffic {yes, no}

@data
35, 105, 78, 40, 90, 43, 12, no
32, 110, 77, 43, 91, 44, 3, yes
```

Best features for identifying Skype

- Characteristic packet lengths
 - Skype packets of length less than 90 bytes are restricted to a small range of values
 - 44, 45, 51, 58, 60, 65, 73, 74 bytes
 - For each window we determined a statistic related to characteristic packet lengths
 - $CPLP_{0.75/90}$ which is the percentage of packets less than length 90 with match one of the characteristic packet lengths
 - We (somewhat arbitrarily) chose packet lengths which occurred 0.75% of the time or more to be “Characteristic”
 - Occur frequently
 - A reliable indicator

Best features for identifying Skype

- Interarrival time
 - Skype client generates packets at intervals that are a multiple of 16 milliseconds
 - 16, 32, 48 or 64 milliseconds
 - Transitions between interarrival times occur frequently in Skype
 - Consequently cannot present the classifier with an average value of interarrival time since transitions between
 - Look for the largest sequence of contiguous time intervals within the window and use that to train the classifier

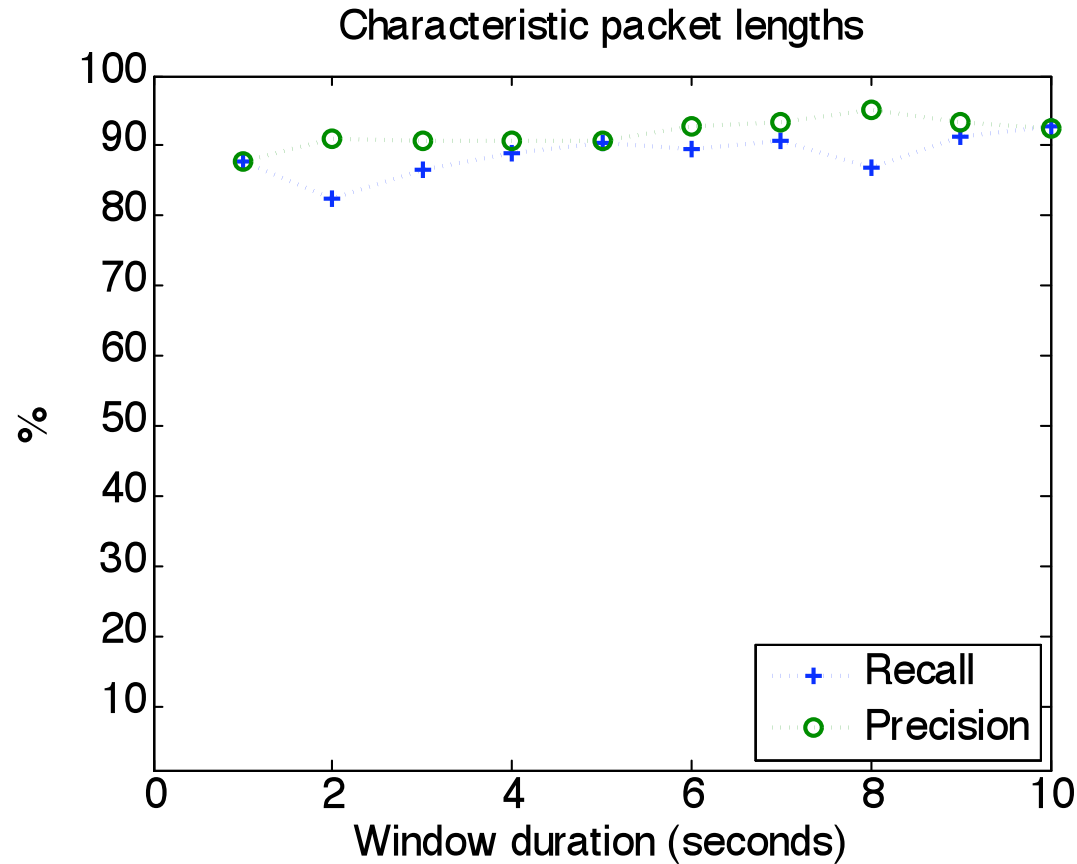
Best features for identifying Skype

- Large packet statistics
 - Used statistics calculated across the whole window
 - Calculated on packets in the window with length greater than or equal to 90 bytes
 - Statistics used
 - Minimum, Maximum, Median, Lower quartile, Upper quartile, Mean, Standard deviation

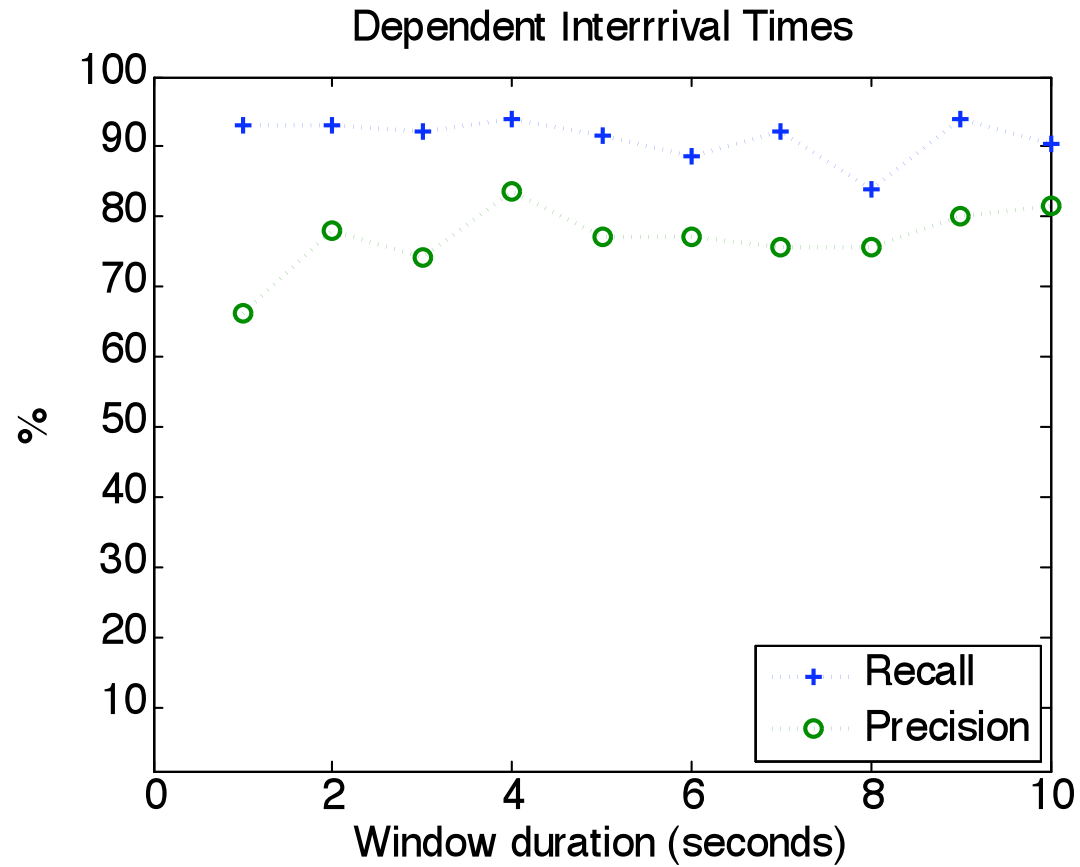
Results – Individual features

- Characteristic packet lengths most effective for very short windows of 1 to 4 seconds
- Large packet statistics most effective for longer windows of greater than 5 seconds
 - Need a reasonable length of time to obtain accurate statistics
- Inter-arrival times by themselves have very good recall but quite low precision
 - Lots of false positives
 - Not necessarily a bad thing. Depends on application. False positives often more acceptable than false negatives.

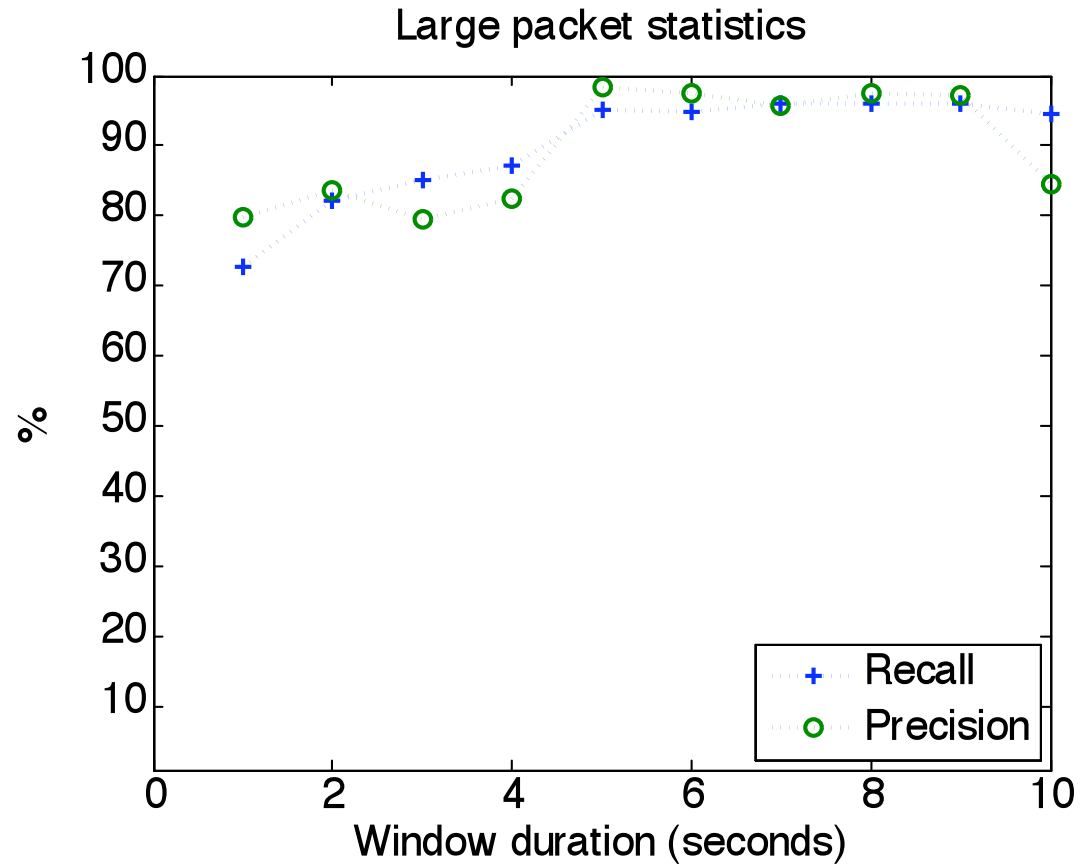
Results – Characteristic packet lengths



Results – Interarrival times



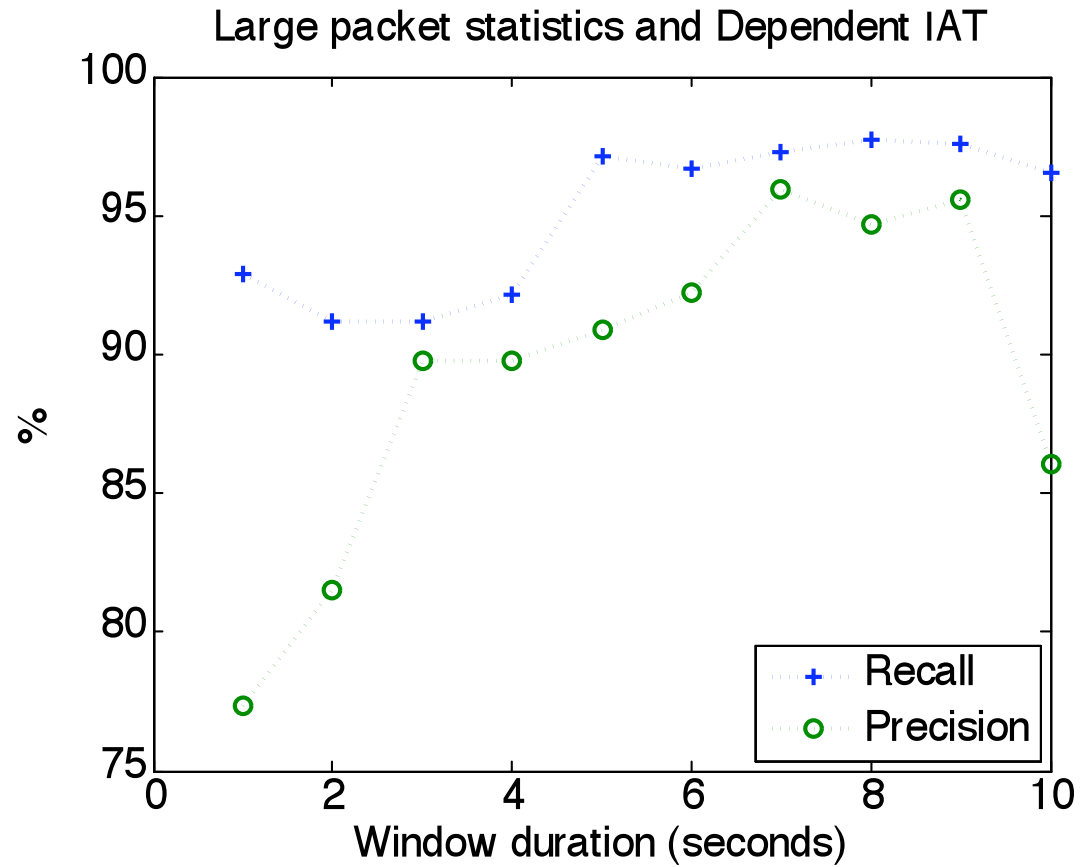
Results – Large packet statistics



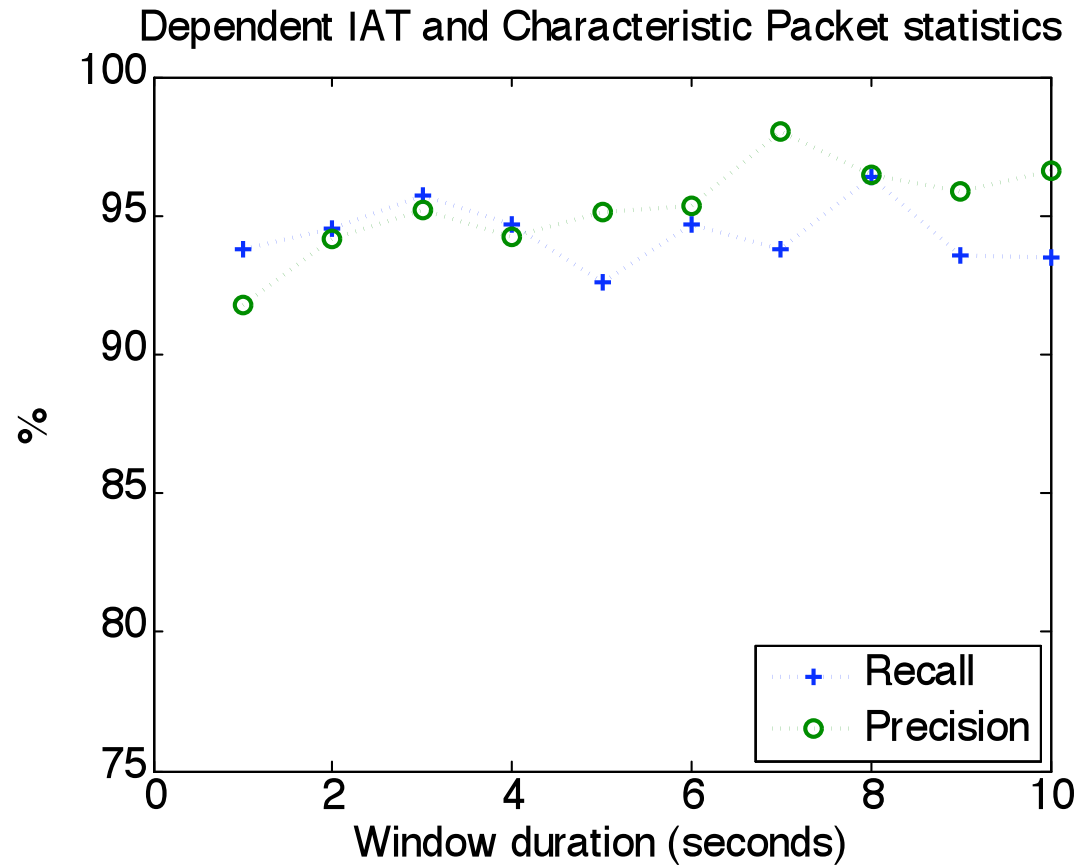
Results – Groupings of features

- Pairs of features
 - For each window we presented two features
 - All paired features performed better than single features (as expected)
 - Most effective was large packet statistics and characteristic packet lengths
- All three features
 - Recall and precision around 99% for windows greater than 5 seconds

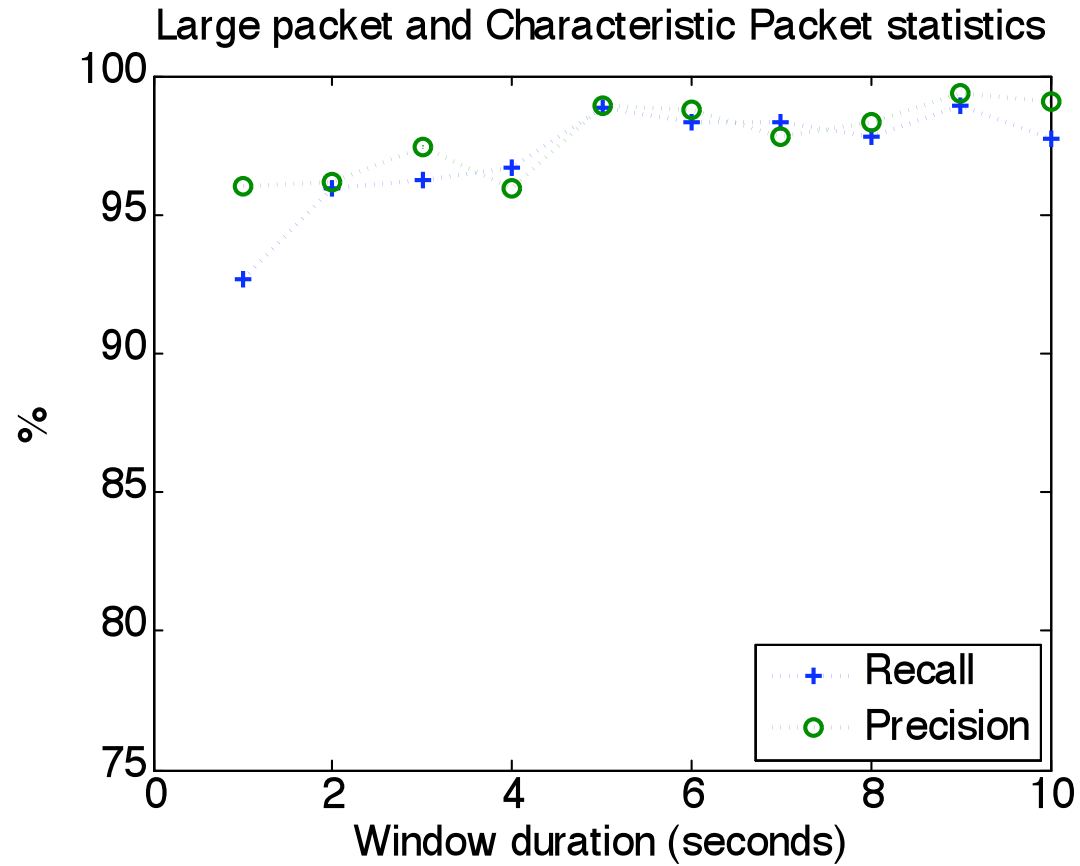
Results



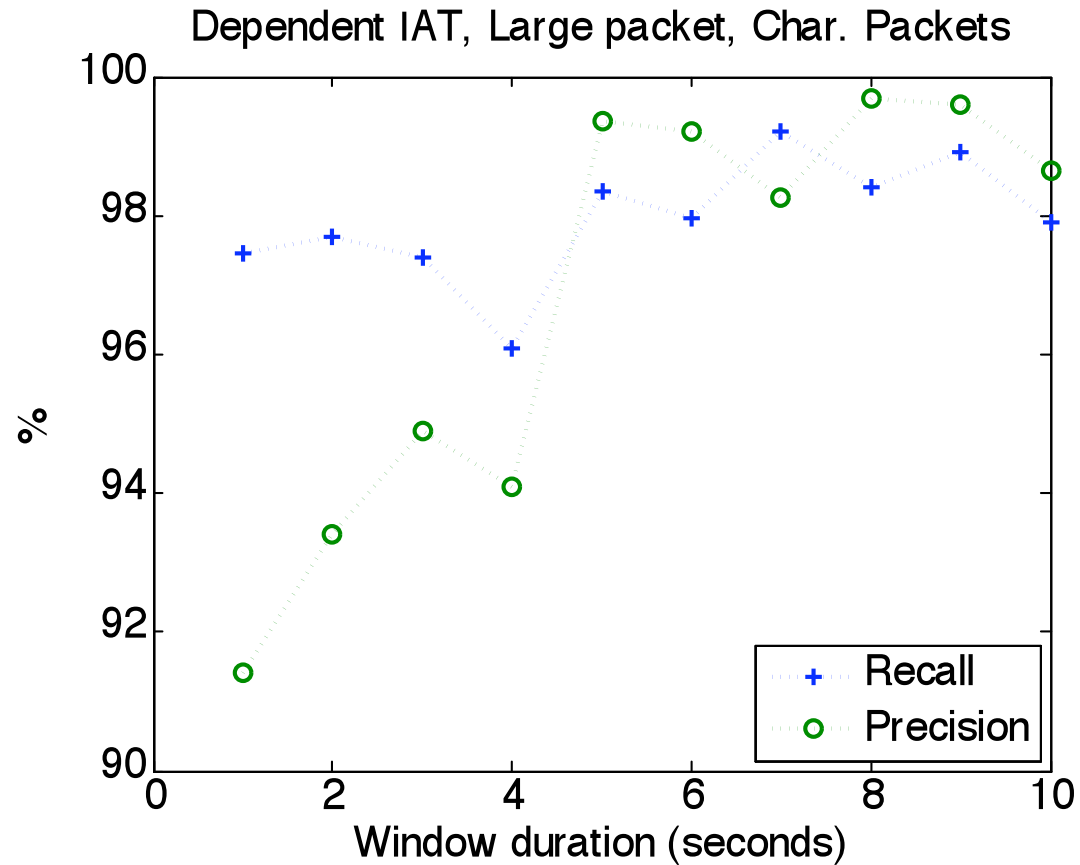
Results



Results



Results – all three attributes



Conclusion and future work

- Machine learning can be used to identify Skype traffic accurately and quickly
 - 98 to 99% accuracy within 5 seconds
- Future work
 - Lots of scope for improving feature selection
 - Quite a few arbitrary cut-offs (eg definition of “characteristic” packet length)
 - Robustness.
 - Works well with SVOPC. What about other codecs?
 - Detect whether or not Skype really is used as a Botnet control network
 - Other supervised classifiers
 - Integrate classifiers into LI systems