P2P Application Recognition Based on Signaling Activity

Chen-Chi Wu[1], Kuan-Ta Chen[2], Yu-Chun Chang[1], Chin-Laung Lei[1]

[1]Department of Electrical Engineering, National Taiwan University

[2]Institute of Information Science, Academia Sinica

Talk Outline

- Introduction
- Fundamentals of our scheme
- Methodology
- Performance evaluation
- Conclusion

Introduction

- P2P traffic constitutes a substantial volume of Internet traffic
- Accurately identify P2P applications from the network traffic is important
  - Network management, capacity planning, etc.
- Conventional approaches: port numbers or payload signatures
  - Dynamic ports, encrypted payload

Fundamentals of Our Scheme

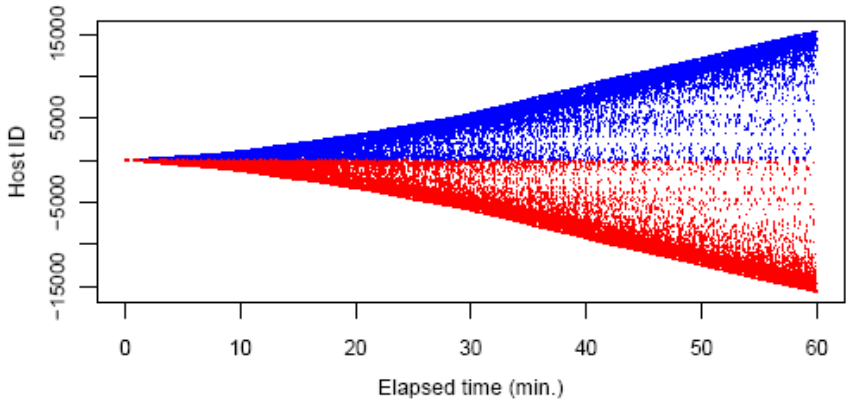
- P2P applications generate two types of traffic
  - Data transfer traffic
    - File-sharing or file-redistribution
  - Signaling traffic
    - File information refreshment, peer discovery, control information exchange, etc.
- Signaling activity is regulated by the underlying P2P protocol
  - Each P2P application may have a unique characteristic

Fundamentals of Our Scheme

- Verify our conjecture
  - Compare the *signaling activity patterns* of BitTorrent, eMule, and Skype
- Traffic data
  - Capture the traffic of 3 hosts that execute BitTorrent, eMule, or Skype
  - Assume packets with payload size smaller than 100 bytes are signaling packets

Signaling Activity Patterns

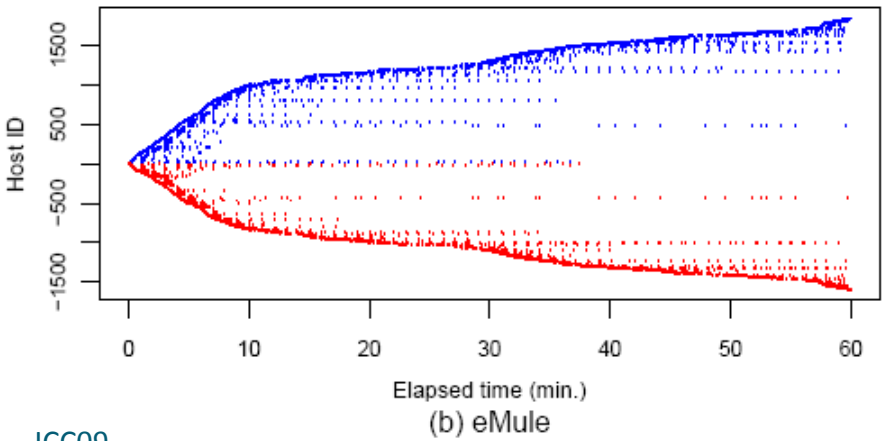
- Assign id to hosts that were contacted by the monitored host based on the order in which they are observed
- BitTorrent
  - Intensive exchange of signaling packets
  - The BitTorrent client progressively discovers new hosts



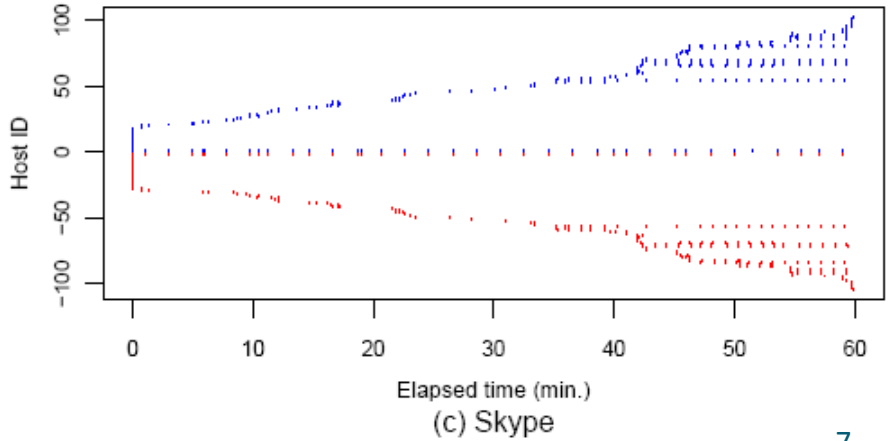
(a) BitTorrent

Signaling Activity Patterns

- eMule
  - The number of hosts increases rapidly in the first 10 minutes but increases slowly thereafter
- Skype
  - Most of signaling packets belong to the probe traffic



(b) eMule



(c) Skype

Proposed Scheme

- Identify P2P applications running on hosts based on the signaling behavior

- How to characterize signaling traffic?

Signaling Behavior Characterization

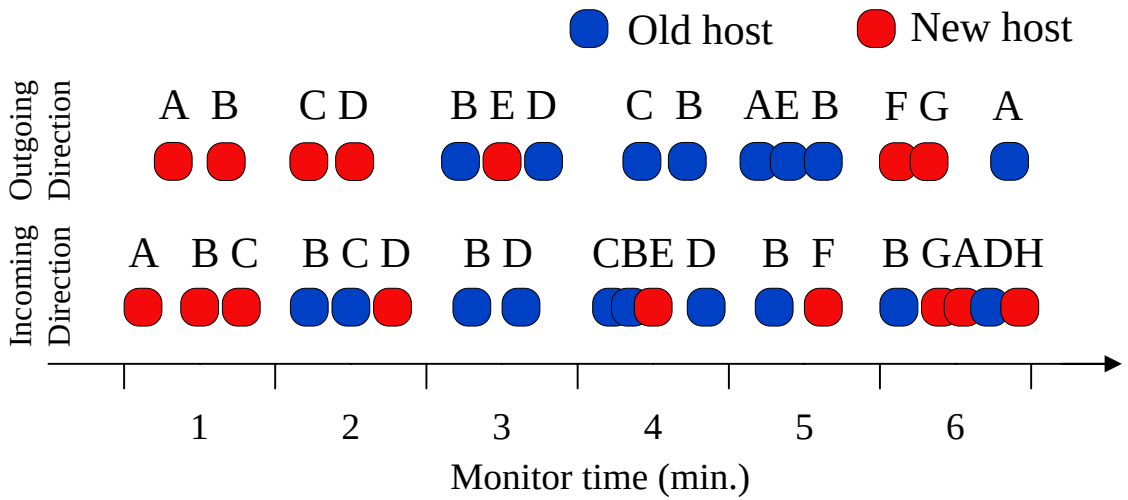
- Keep track of signaling packets of a monitored host for a period of time
- Count the number of hosts contacted and the number of packets sent and received every minute
- Classify hosts contacted with the monitored host into 2 types
  - Sending/receiving packets within 5 minutes => *old host*
  - Otherwise => *new host*
- Characterize the signaling behavior on two levels
  - Host level: based on the number of new or old hosts
  - Message level: based on the number of new or old packets

Signaling Behavior Features

Host level
Ratio of new / old hosts
Growth rate of new / old hosts
Correlation coefficient between the number of new and old hosts
Message level
Ratio of new / old packets
Growth rate of new / old packets
Correlation coefficient between the number of new and old packets

Example

- Host level - ratio of new hosts
  - Keep track of hosts contacted with the monitored host
  - Incoming direction in the 6th min.: B and D are old hosts; A, G, and H are new hosts
  - Ratio of new hosts in the 6th min. => 3/5



Identifier Design

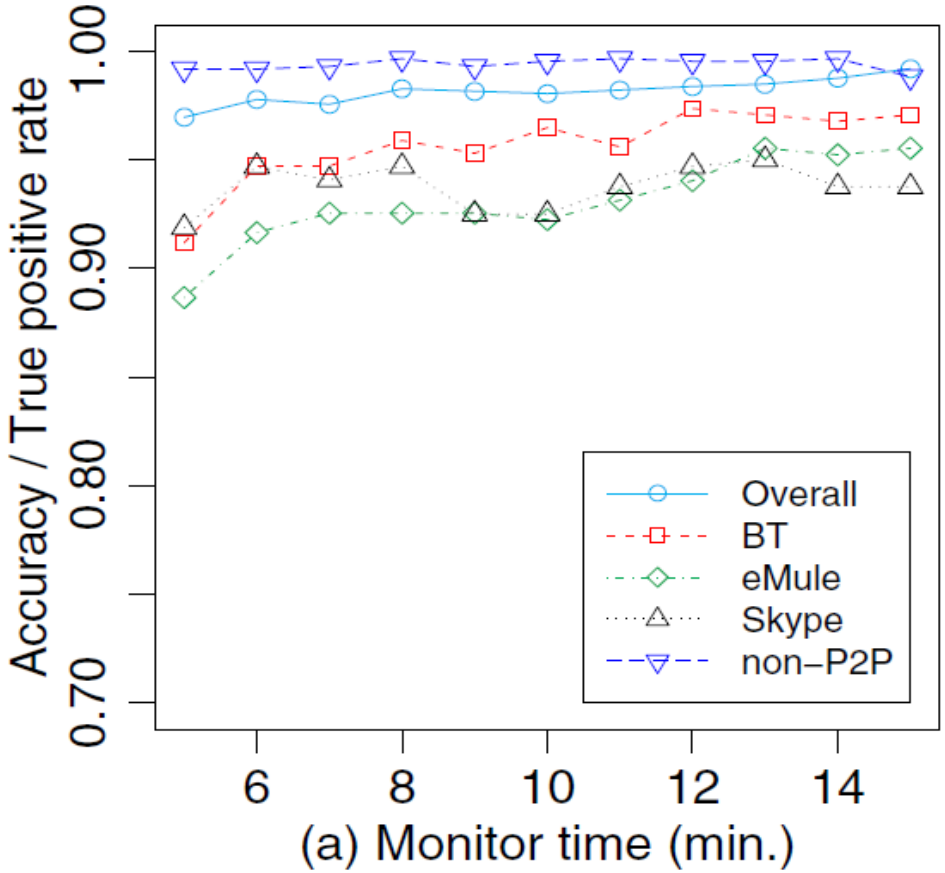
- Adopt support vector machine (SVM)
- Training phase
- Derive features from each training data
- Label each training data with the name of P2P applications
- Train the SVM classifier
- Identification phase
- Derive features from a signaling packet stream
- Use the trained classifier to determine the P2P application

Traffic Data

Category	Hosts	Packets
BitTorrent	110,711	104,722,150
eMule	42,377	36,716,588
Skype	61,777	34,076,328
World of Warcraft	218	2,528,359
TELNET	362	21,118,522
HTTP	4,448	28,264,360

Performance Evaluation

- 10-fold cross validation



Conclusion

- Summary
  - Identify distinct P2P applications without examining payload
  - Characterize signaling behavior possessed by P2P applications
- Future work
  - Consider the case that a host launches multiple P2P applications
  - Short flows?

Thank you for your attention!