A Hybrid Approach for Accurate Application Traffic Identification

Thesis Defence

December 21, 2005

Young J. Won
yjwon@postech.ac.kr

Distributed Processing & Network Management Lab.
Dept. of Computer Science and Engineering
POSTECH, Korea

POSTECH
DP&NM Lab

Table of Contents

Introduction (1/2)

- Traditional traffic (e.g. web, e-mail, ftp) to peer-to-peer & multimedia traffic.
  - According to our study, more than 60% of the total traffic belong to peer-to-peer file sharing.

- Traffic dynamic of the Internet's dominant applications: New Challenge to Application Traffic Identification.
  - Ephemeral ports (e.g., VLAN, Firewall, P2P file sharing)
  - Multiple sessions (e.g., Distributed file downloading from multiple peers)
  - HTTP encapsulation

Introduction (2/2)

❖ What is Application Traffic Identification?

A procedure that determines the origin application of traffic in the unit of packet or flow.

- Informative snapshot of Networks
- Prerequisite for QoS
- Billing policy for ISPs

❖ Problems?

The new traffic dynamic deteriorates the credibility of accuracy of the existing identification methodologies.

- Moore *et al* argue that port matching is no more accurate than 50% ~ 70% [Moore *et al*, PAM'05].
- No strong proof of accuracy.

Research Goal

- To provide an accurate application traffic identification method for real-world networks within some boundary of practicality.
  - Exhaustive search of application information.
    (e.g. signature, port).

- Accuracy is the No. 1 factor to consider.
  - Verification of the identification results.

Related Work

Category	Identification Method	Accuracy	Vulnerable to payload encryption	Cost of Operation	Exhaustive Searching	Applicability
Session-based Approach	Well-known Port Matching [IANA]	Medium	No	Low	Yes	Practical
	Session Behavior Modeling [Karagiannis, SIGCOMM'05]	Low	No	Medium	No	Experimental
	Port + Session Behavior (e.g. FRM) [Kim, ETRI'05]	Medium	No	Low	Yes (Port Information)	Practical
Content-based Approach	Protocol Inspection [Kang, DSOM''03]	Medium	Yes	High	Yes	Practical but very limited
	Protocol Matching [Ethereal]	Medium /High	Yes	High	Yes	Practical
	Signature Matching [Sen, WWW'04]	Medium /High	Yes	High	Yes	Practical
Constraint-based Approach	Supervised Machine Learning [Moore, SIGMETRICS'05]	Unknown	No	Unknown	No	Experimental
	Statistical Signature-based [Roughan, IMC'04]	Unknown	No	Unknown	No	Experimental
	HMM Profiling [Charles, ACM Com. Secu.'04]	Unknown	No	Unknown	No	Experimental
Hybrid	Signature Matching + Session (flow) Pattern	High	Yes	Medium	Yes	Practical

Hybrid Approach - Concept

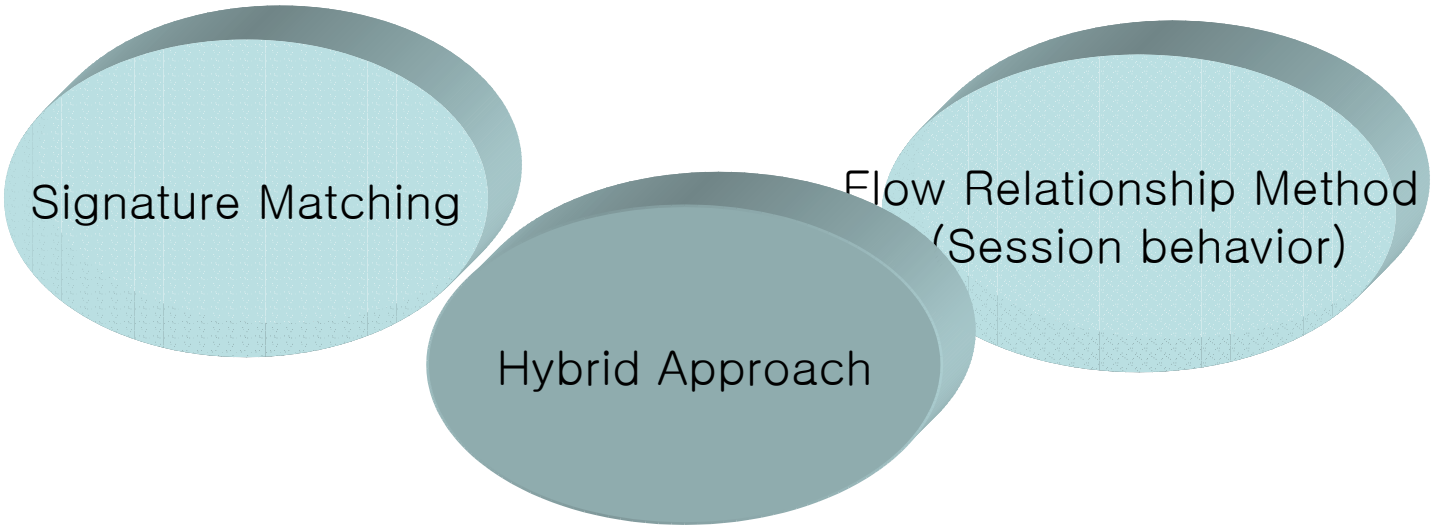
Accuracy of Signature Matching

+

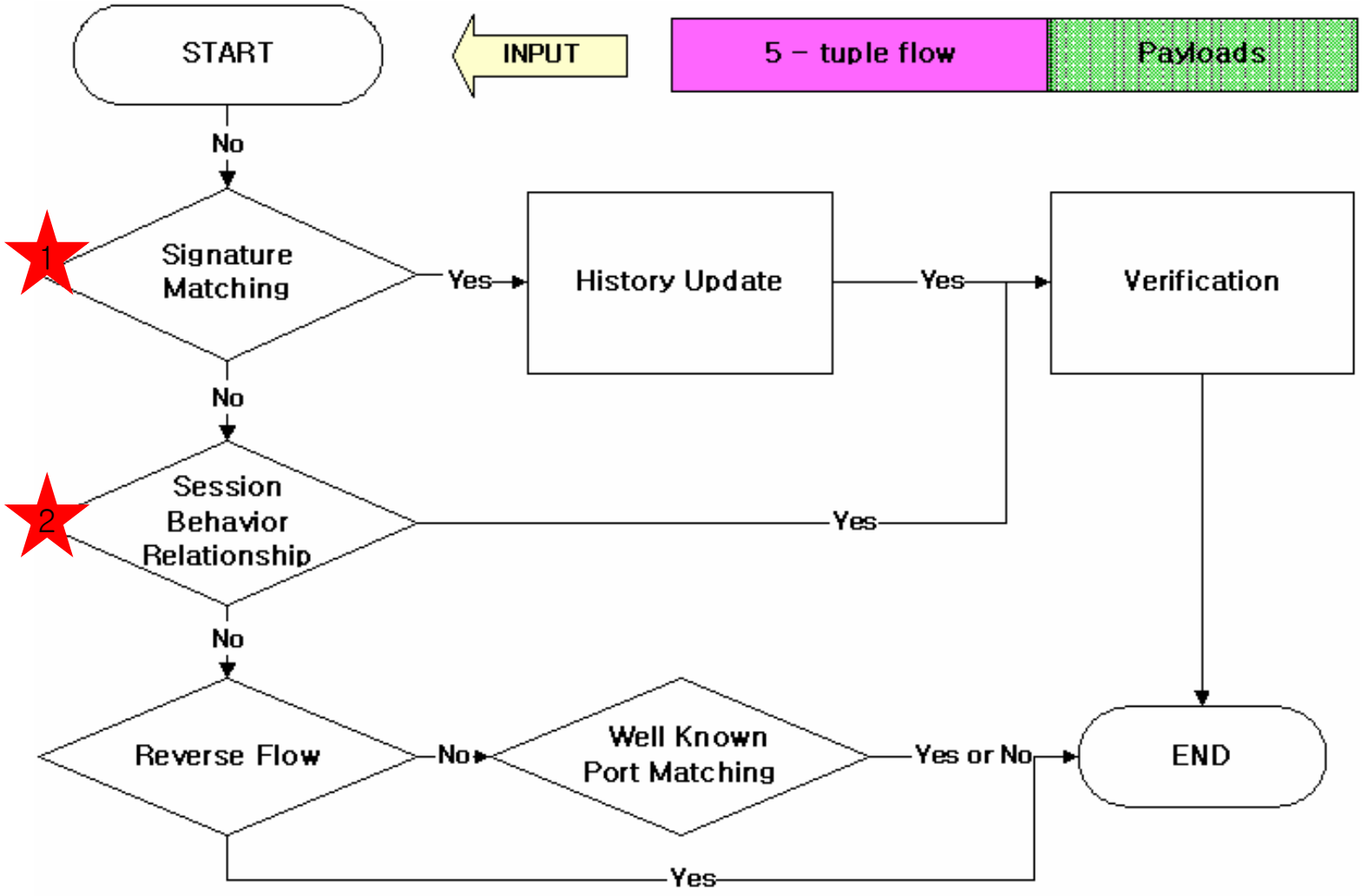
Capability of Session Behavior Relationship (FRM)

||

Accurate & Efficient Method



Hybrid Approach – Identification Procedure



Hybrid Approach - Assumptions

1. Packets occurring in the close time interval (< 1 min) and sharing the same 5-tuple (source IP & port, destination IP & port, protocol) are originated from the single application.

2. Reverse packets (displacement of 5-tuple information, protocol must be the same) in the close interval (< 1 min) belong to the same application.

3. Packets occurring in the close time interval (< 1 min) and sharing the same source (or destination) IP and port are originated from the single application.

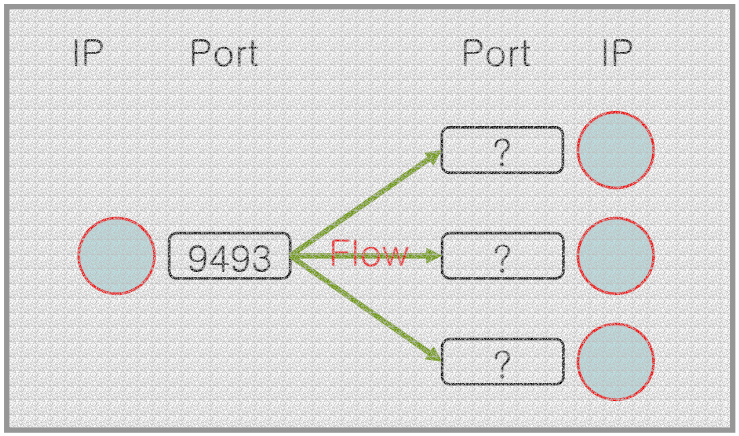
4. For limited applications (e.g. passive ftp), packets belonging to the multiple sessions between the two distinct hosts (IP) are originated from the single application.

Signature Matching on Flow

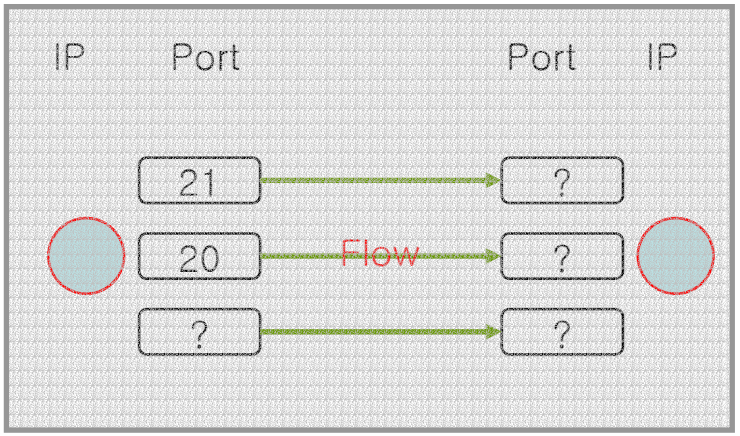
- Signature
  - A static, unique, and distinguishable, pattern of hexadecimal digits or specific strings in the payload of packet.

- Priority-based Signature Matching on limited packet samples.
  1. Application with a longer matching signature has higher priority.
  2. Application with multi-positioned signature has higher priority.
  3. ‘HTTP’ protocol signature has the lowest priority.

Session Behavior Relationship



a) {IP, Port} Relation



b) {IP, IP} Relation

- Search for undetermined flows satisfying these relations.
  - Relationship mapping between determined and undetermined flows.
- Default timeout value for both relations is 2 minutes.
  - Update the tables in History Update
- Selectively using the conditions in FRM.

Well-known Port Matching

- Berkeley Port Allocation Scheme
  - Most Unix and non-Unix TCP/IP implementations
  - Port 0 to 1023 are privileged.
  - ssh (22), telnet (23), mail (25), and more.

- For some cases, port matching is a still effective identification method – 'Suspected' traffic.
  - We cannot guarantee 100% accuracy.
  - Web (TCP 80, 8080), PD_Club_box (TCP 19101), and more.

Verification (1/2)

- Port Dependency Ratio (PDR) - How much traffic is bounded to a particular port number in the pool of flows which are identified via signature matching.
  - PDR of major ports (Zipf like distribution) and PDR Distributions

Interpretations:

1. Classification probability(%) of the suspected amount traffic.
   - Estimating how accurate the suspected amount traffic.

2. Indicator of how accurate the found signatures are for some applications.
   - Allocating the same port or ports in the close range repeatedly
   - High PDR

Verification(2/2)

❖ Example

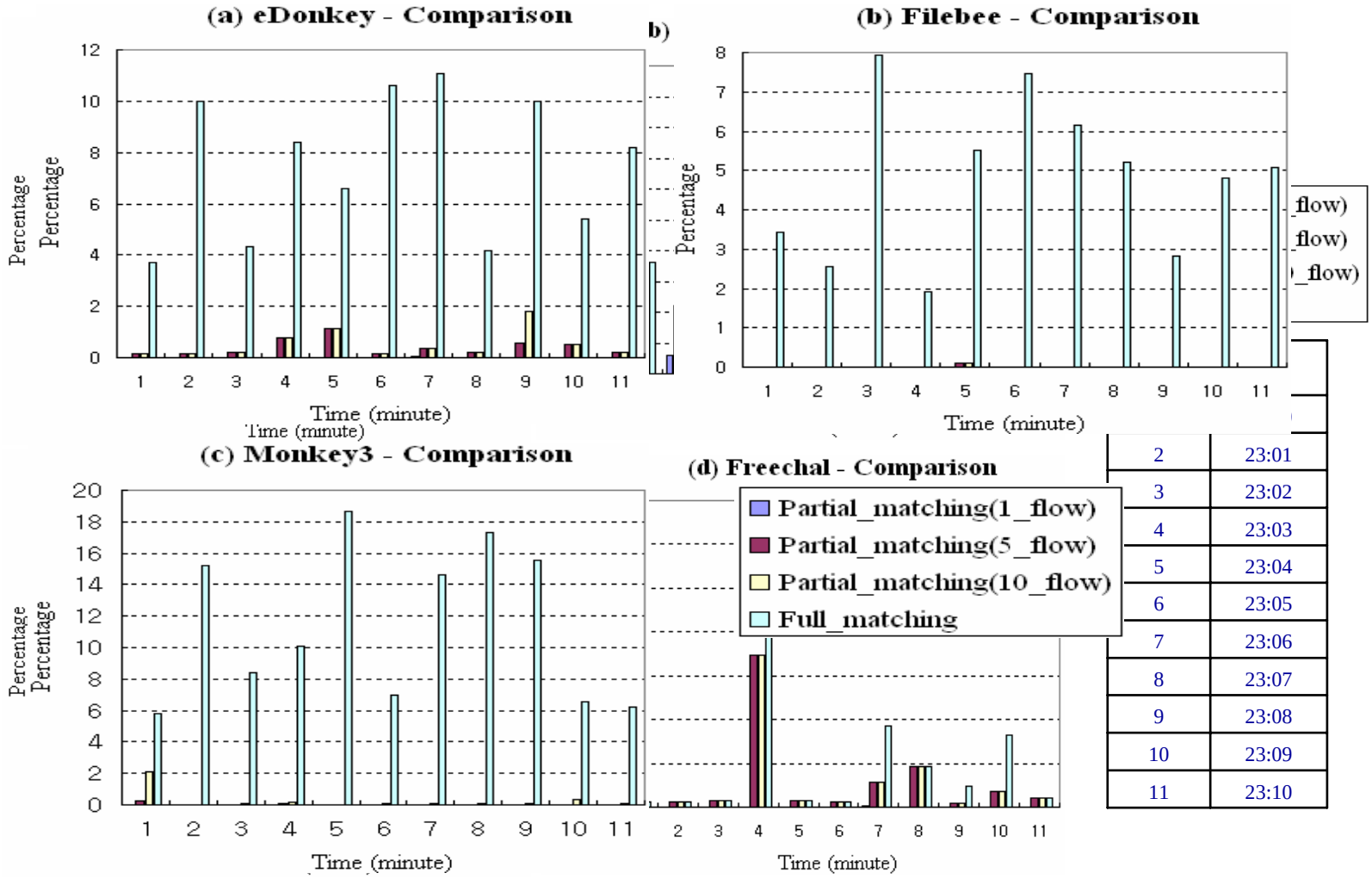
Application	Major Port	PDR
Web	80	90 ~ 100%
eDonkey	4662	70 ~ 90%
Freechal	9493	100%
Monkey3	8008	90 ~ 98%
BitTorrent	N/A	N/A
PD Club Box	150xx	10 ~ 80%

Full Matching vs. Partial Matching (1/3)

- Application signatures are more reliable when they are present in initial few packets of flow or first few bytes of payload.

- Comparison test - Identification ratios of the following 4 scenarios:
  - Signature matching on the first packet payload of flow (Partial Matching)
  - Signature matching on the first five packet payloads of flow
  - Signature matching on the first ten packet payloads of flow
  - Signature matching on every packet of flow (Full Matching)

Full Matching vs. Partial Matching (2/3)



2	23:01
3	23:02
4	23:03
5	23:04
6	23:05
7	23:06
8	23:07
9	23:08
10	23:09
11	23:10

Full Matching vs. Partial Matching (3/3)

		eDonkey		File bee		Monkey3	
Index	Time	Determined Byte – MB(%)	False Positive Ratio	Determined Byte – MB(%)	False Positive Ratio	Determined Byte – MB(%)	False Positive Ratio
1	23:00	50 (3.7%)	81% +	47 (3.44%)	78% +	78 (5.78%)	62% +
2	23:01	135 (10%)	74% +	34 (2.54%)	55% +	206 (15.19%)	35% +
3	23:02	60 (4.33%)	46% +	111 (7.93%)	74% +	118 (8.44%)	59% +
4	23:03	110 (8.42%)	72 % +	24 (1.91%)	29% +	132 (10.12%)	17% +
5	23:04	82 (6.6%)	37 % +	68 (5.5%)	55% +	233 (18.66%)	unknown
6	23:05	136 (10.6%)	79 % +	95 (7.45%)	unknown	89 (6.98%)	45% +
7	23:06	136 (11.1%)	76 % +	75 (6.14%)	unknown	180 (14.62%)	56% +
8	23:07	49 (4.16%)	25 % +	62 (5.22%)	43% +	206 (17.31%)	74% +
9	23:08	118 (10%)	69 % +	33 (2.84%)	90% +	184 (15.51%)	48% +
10	23:09	62 (5.39%)	68 % +	56 (4.82%)	80% +	76 (6.53%)	46% +

- Matching signature on every packet is unnecessary and could generate unreliable identification results.
- First five packets scheme for Validation process.

VALIDATION

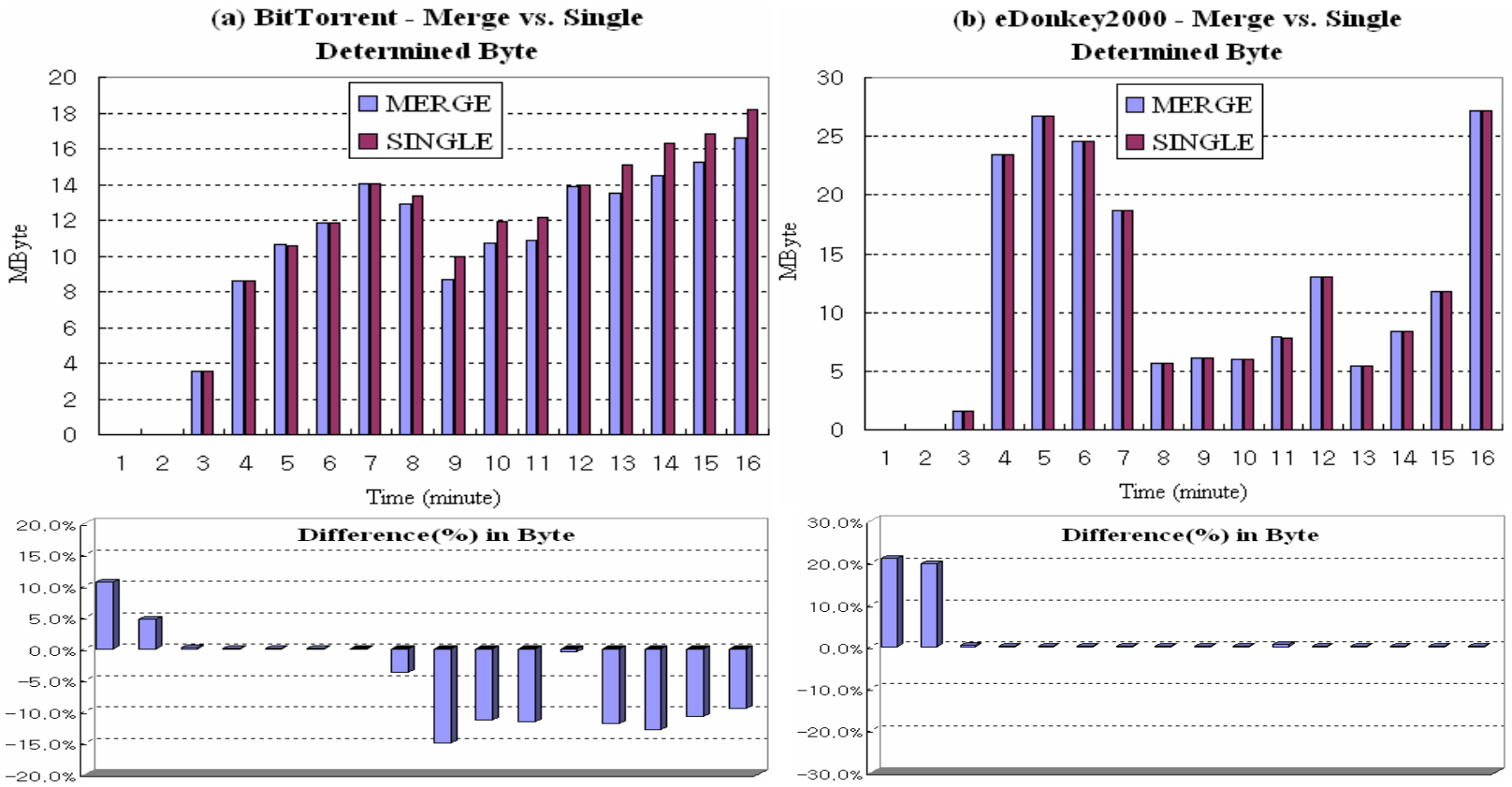
Isolated Traffic vs. Synthetic Traffic Mix (1/3)

- Obtain a suitable traffic data set.
  - BitTorrent, eDonkey2000, Freechal, KaZaA, Monkey3, MSN, and PD_Club_Box.

- Isolated Traffic - Each application traffic trace of 16 minutes while running on a single host independently, referring to Single.
  - Fully aware of all the applications that occupy the traffic.

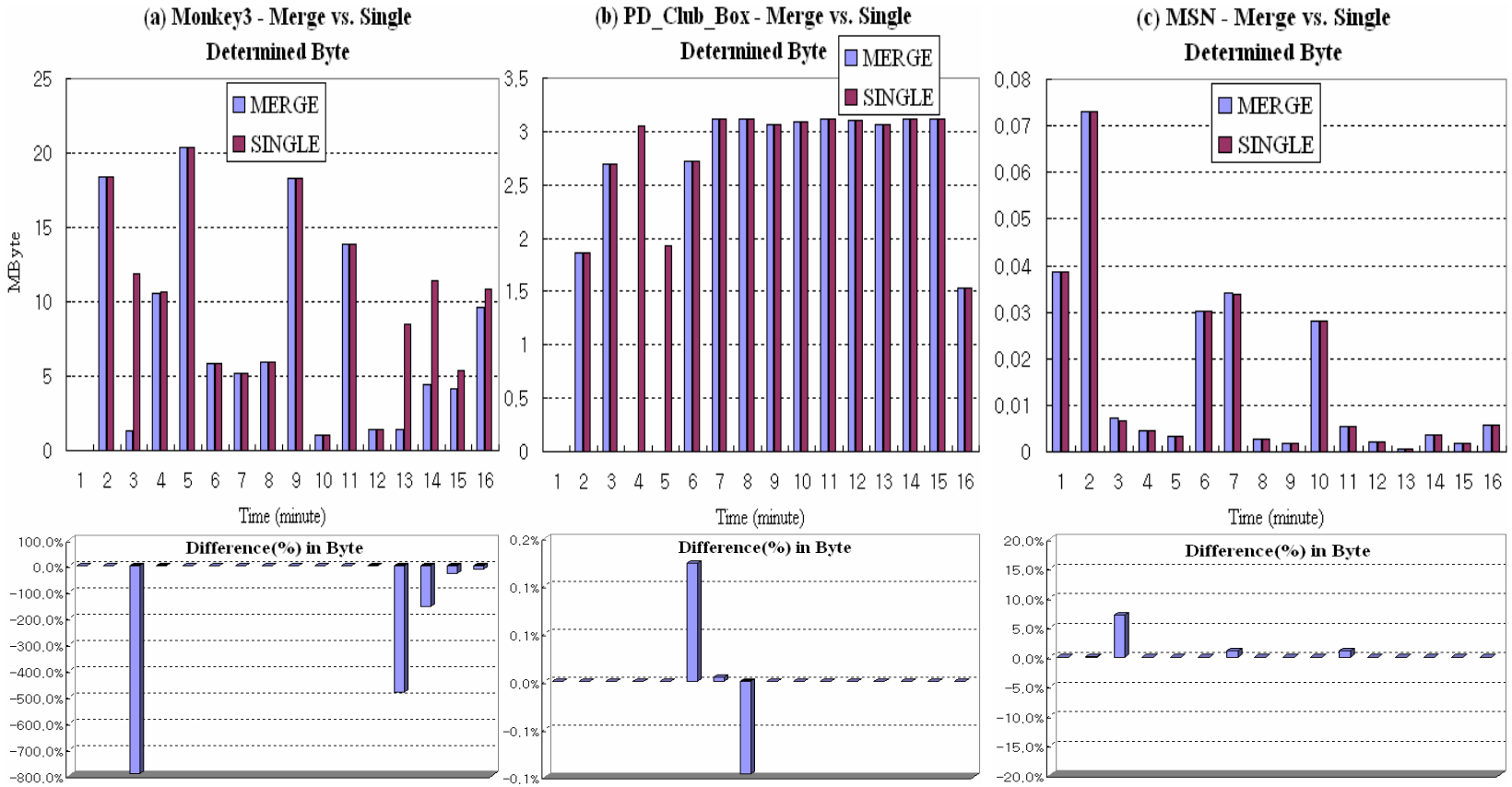
- Synthetic Traffic Mix – Shuffle all the traffic traces of these seven applications, referring to Merge.

Isolated Traffic vs. Synthetic Traffic Mix (2/3)



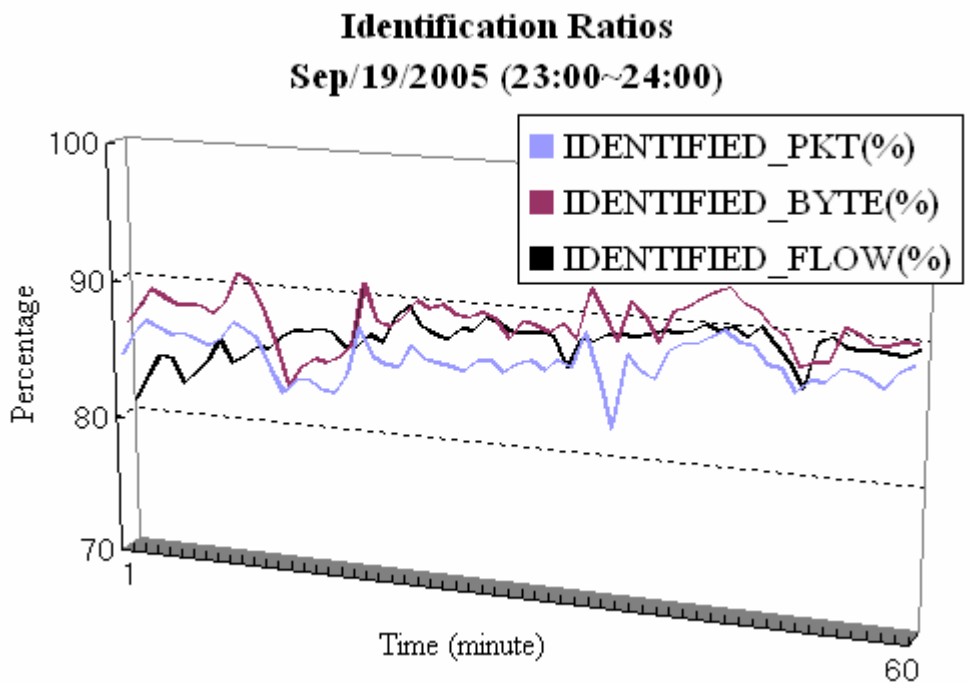
- + (-) difference percentage means large (less) byte counts in 'Merge' data set.
  - These missing or gaining amounts of traffic are negligible.
  - It is likewise for Freechal and KaZaA.

Isolated Traffic vs. Synthetic Traffic Mix (3/3)



- The sudden peaks in the difference indicate the unclear distinction of the traffic between the applications and web.
  - Is our identification wrong? NO!

Overall – Real Traffic Mix (1/6)



Type	Application	Byte (%)
	Web	20~40 %
P2P	eDonkey, monkey3, freechal, pd_club_box, co_file, File bee, and more	50+ %
Messenger	NateOn, MSN messenger	< 1 %
Others	ftp, mail, MS_dir_service, idisk, NetBios, and more	< 10 %
Unknowns	?	< 10 %

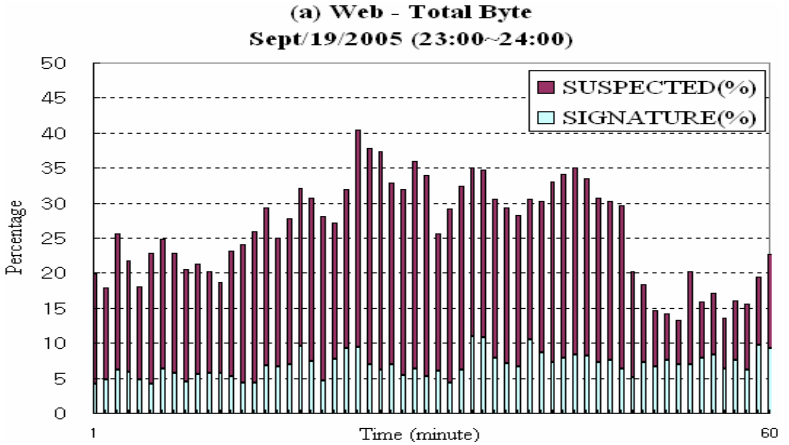
- The sample traffic is collected at the Internet junction of POSTECH, for one hour (23:00 ~ 23:59, Sept.19.2005).
  - 1 Giga-bit Ethernet link (Avg. 200 Mbps, 130M packets, 83 GB)

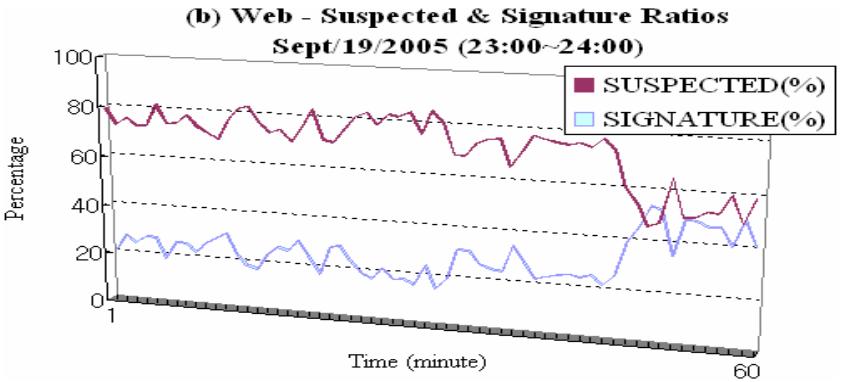
Overall – Real Traffic Mix (2/6)

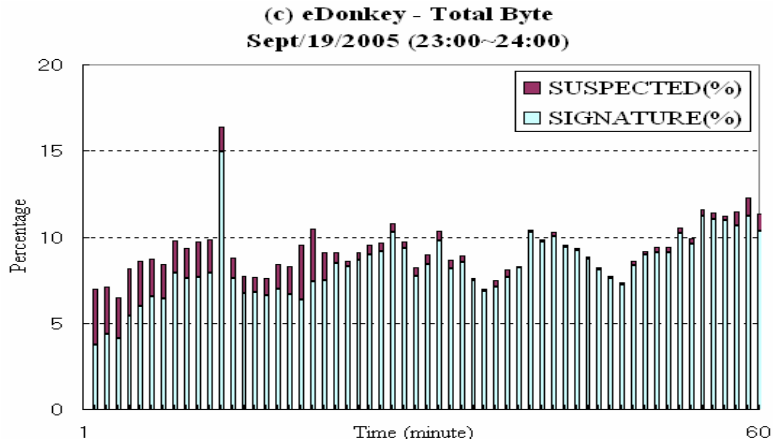
Type	Application		Byte (%)	
	FRM	Hybrid Approach	FRM	Hybrid Approach
Web	-	-	20 ~ 30 %	20 ~ 40%
P2P	eDonkey, freechal, bitTorrent, pd_club_box, and more	eDonkey, monkey3, freechal, bitTorrent, pd_club_box, file bee, and more	60 %	50+ %
Messenger	NateOn, MSN messenger and more	NateOn, MSN messenger and more	1 %	1 %
Others	ftp, mail, MS_dir_service, idisk, NetBios, and more	ftp, mail, MS_dir_service, idisk, NetBios, and more	15 %	10 %
Unknowns	unknown	unknown	5 %	10 %

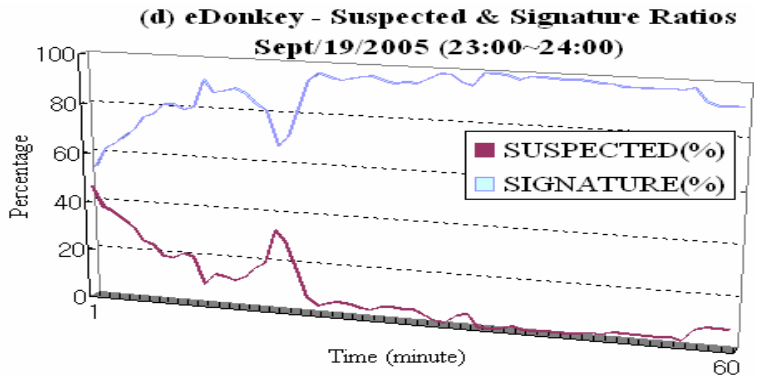
- Newly detected applications: e.g.) Monkey3, File bee.
- Difference? More unknown in hybrid approach; Proof of identified traffic.

Overall - Real Traffic Mix (3/6)



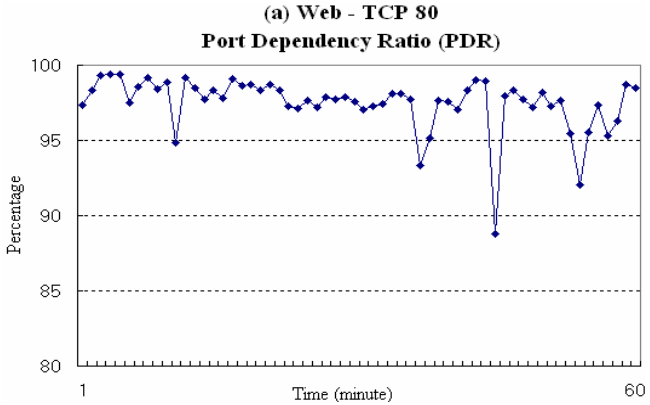


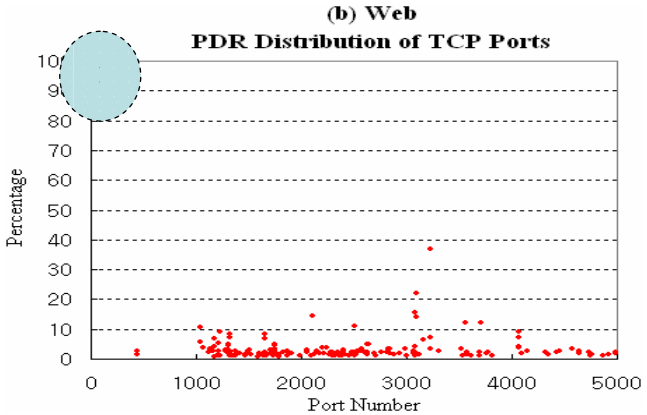


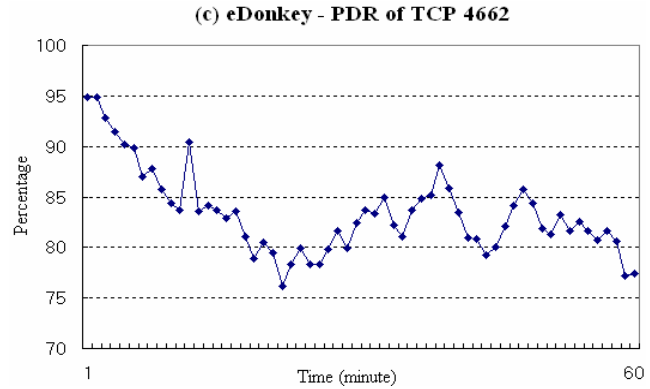


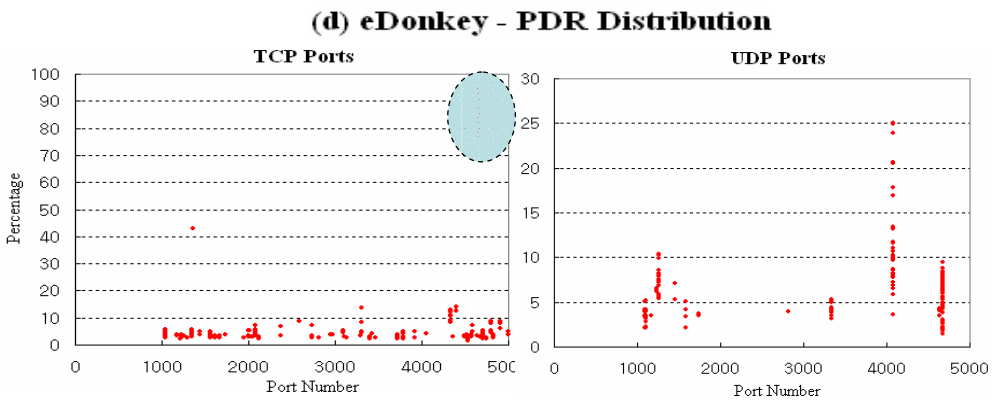
- The ratio of suspected traffic amount to the total identified traffic amount is declining in both.
  - eDonkey traffic can guarantee 100% accuracy when the ratio reaching down to 0.

Accuracy Analysis - Real Traffic Mix (4/6)



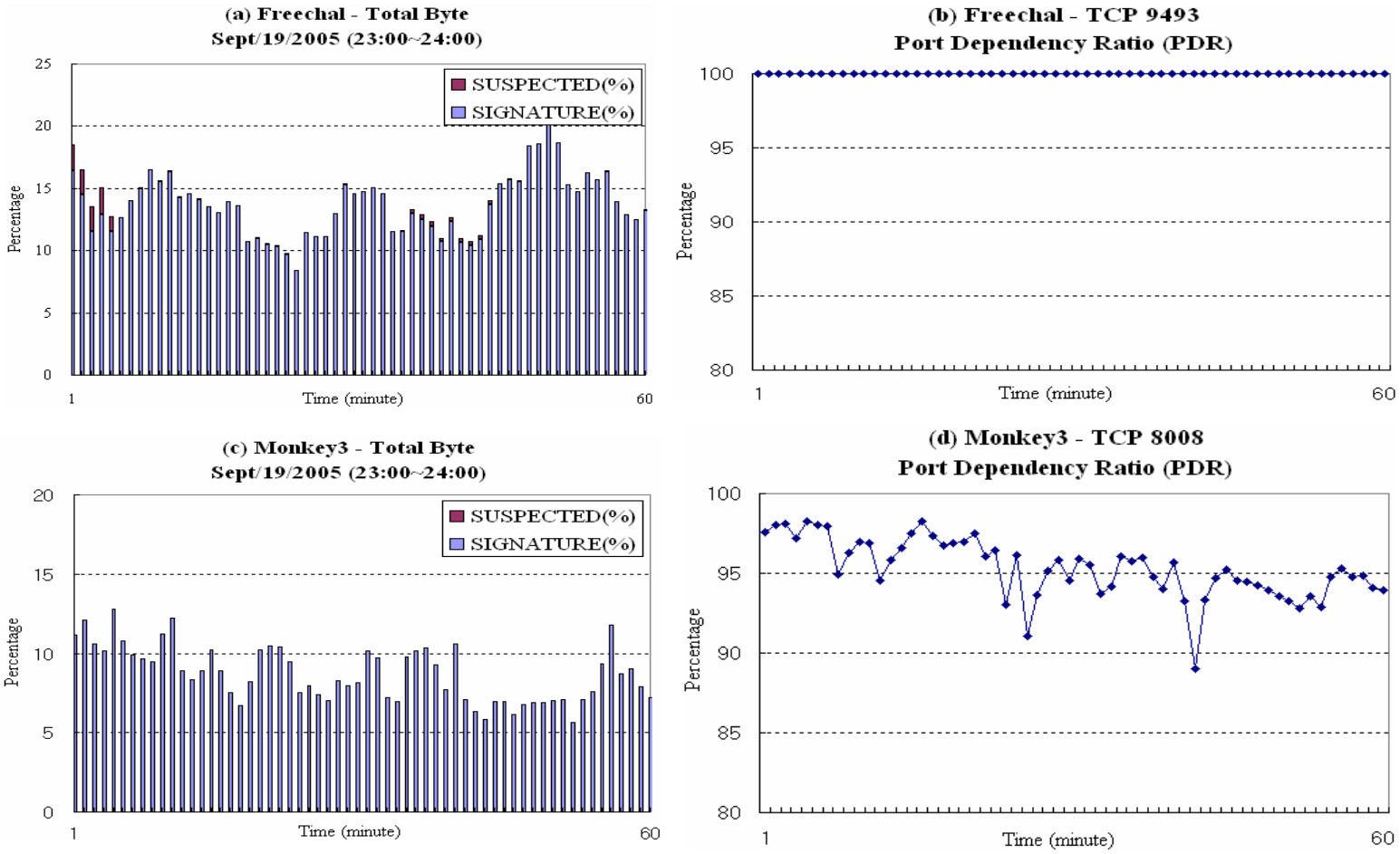






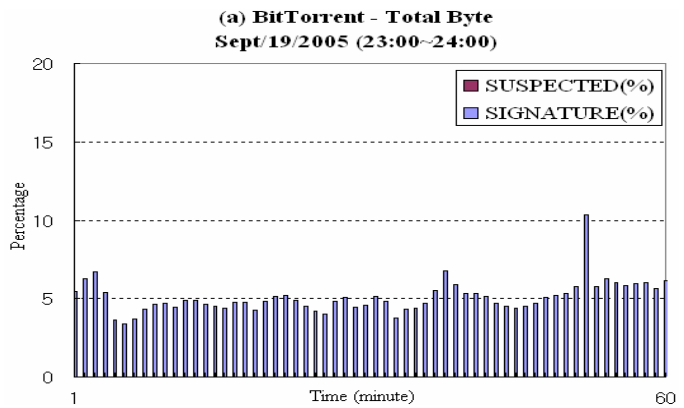
- The IANA's port listing for web is reliable
  - The PDR of TCP port 80 reaches up 99%.
- Frequent allocations of TCP 4662 for eDonkey
  - The PDR of TCP 4662 is high (70% ~ 90%)

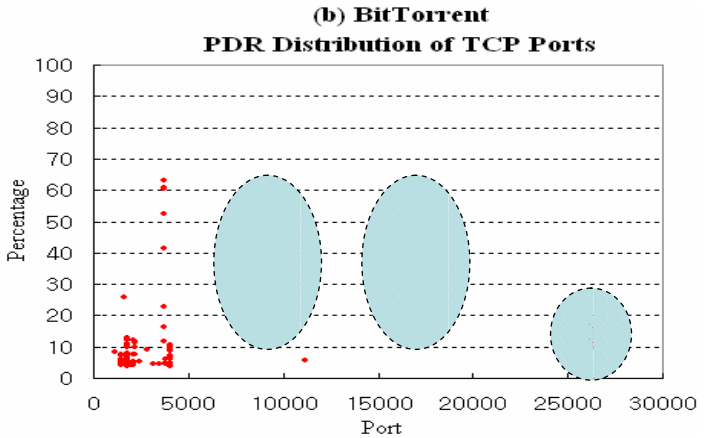
Accuracy Analysis - Real Traffic Mix (5/6)

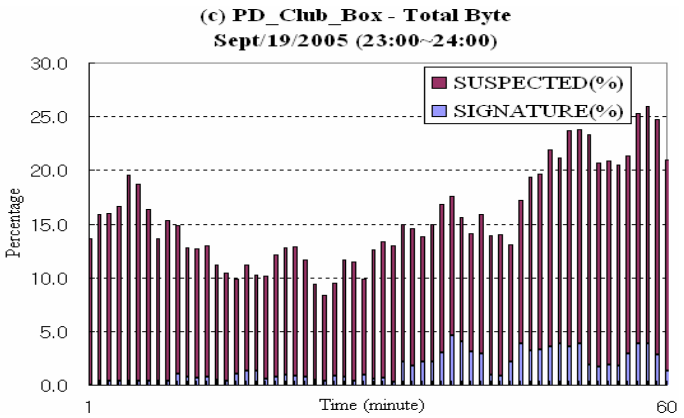


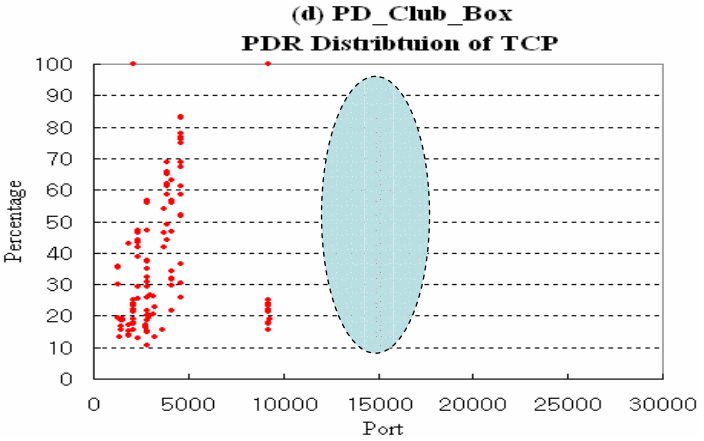
- Freechal and Monkey3 are bounded to almost fixed port usage.

Accuracy Analysis - Real Traffic Mix (6/6)









- Sporadic port allocations by BitTorrent (3 shaded areas)
- PD Box traffic is bounded to TCP ports within the range of 150xx.

Summary & Contributions

- Essential survey and categorization of application traffic identification algorithms.

- Proposed a hybrid approach of the existing identification techniques.
  - Priority-based signature matching on limited packets.
  - Session behavior mapping in flow-level.

- An analysis for identification accuracy.
  - The analysis consists of how to obtain the appropriate data set for preliminary accuracy testing and interpret the accuracy measure variable, PDR.

Future Work

❖ Automation of Signature Generation

- A strong requirement for efficient updates of signatures.
- Toward building the self-learning application traffic identification system.

❖ Near Real-time Deployment of the proposed algorithm

- Providing informative snapshot of networks.
- QoS provisioning.

❖ Real-time Visualization of Session Patterns

- Graph matching.
- Early detection of known as well as unusual patterns.

감사합니다