CE 817 - Advanced Network Security
Traffic Analysis

Lecture 17

Mehdi Kharrazi
Department of Computer Engineering
Sharif University of Technology

Acknowledgments: Some of the slides are fully or partially obtained from other sources. Reference is noted on the bottom of each slide, when the content is fully obtained from another source. Otherwise a full list of references is provided on the last slide.

Traffic Analysis

What is Traffic Analysis

- Making use of (merely) the traffic data of a communication to extract information. As opposed to ‘interception’ or ‘cryptanalysis’. What are traffic data?
  - Identities or call signs of communicating parties.
  - Time, duration or length of transmissions.
  - Location of emitter or receiver.
  - No content – it may be encrypted.
- A controversial starting point.
  - Diffie & Landau statement – ‘Privacy on the line’ on the politics of encryption.
  - “Traffic analysis, not cryptanalysis, is the backbone of communications intelligence.”
  - Could this become true on the Internet? Is it already?

[Danezis]

The military world – concrete attacks

- Naval and air operations: observing wireless communications
- Reconstruction of network structure of the German Air Force radio in 1941 by the British
- Identification of radio equipment
- Radio silence
- Morse ‘hand’
- Why is traffic analysis so valuable?
    - It provides lower quality information compared with cryptanalysis, but it is both easier and cheaper to extract and process.
    - Often used to perform ‘target selection’. ‘Economics of surveillance’

[Danezis]

The military world – defences

- Low probability of intercept and Low probability of direction finding communications.
  - Principle: make the adversary spend time or energy (power) to detect and jam.
  - Frequency hopping – modulate frequencies according to keys. Difficult to jam!
  - Spread spectrum – transform signal to high band low power. Difficult to detect (under the noise floor).
  - Burst communications – meteor scatter.
- Technologies used for civilian purposes: GSM (hopping), ADSL (SS), Cheap but reliable comms (meteor).
- Reference: Ross Anderson, Security Engineering.

And then came (not just) the Internet. . .

- Different environment – (not so) different players.
  - Not so hostile – commercial use, personal use, government, (critical infrastructure?), (military?)
  - A confederation of networks – different jurisdictions and security domains.
  - Different transport technologies: cable, wireless, satelite, ATM, ethernet, . . .
  - Common routing protocols – they expose traffic data.
  - New technologies: wireless, overlay networks, convergence with telephone – more opportunities for collecting traffic data.
  - Threats rapidly escalate – attack scripts!
- Use of encryption – NG Telephony and Internet: traffic analysis only option
- How to attack established security technologies? (Without making use of cryptanalysis or content)

[Danezis]

Can the Secure SHell (SSH) protect your privacy?

- SSH is used for secure remote login and file transfer. All data is encrypted and authenticated. What information can we extract about a password typed in a protected session?
  - Key observation: each key pressed is transmitted separately.

[Danezis]

Can the Secure SHell (SSH) protect your privacy?

- SSH is used for secure remote login and file transfer. All data is encrypted and authenticated. What information can we extract about a password typed in a protected session?
  - Key observation: each key pressed is transmitted separately.
  - Depending on the position of the key on the keyboard, different inter key timings.
  - Attack (Song et al.): observe the inter key timings (many times if you wish) – infer what keys have been pressed.
  - Result: reduce the entropy of password – fewer guesses required.
- Note that there is still variability across different people. Adds noise – but also opportunities (Rubin et al.)!

[Danezis]

Can the Secure SHell (SSH) protect your privacy?

- SSH is used for secure remote login and file transfer. All data is encrypted and authenticated. What information can we extract about a password typed in a protected session?
  - Key observation: each key pressed is transmitted separately.
  - Depending on the position of the key on the keyboard, different inter key timings.
  - Attack (Song et al.): observe the inter key timings (many times if you wish) – infer what keys have been pressed.
  - Result: reduce the entropy of password – fewer guesses required.
- Note that there is still variability across different people. Adds noise – but also opportunities (Rubin et al.)!
  - Monitor a user session and record the timings of key presses.
  - Use existing profiles to infer their identities according to the leaked timing.
- Can extract both information and identity from a 'secure' session.

[Danezis]

Do Secure Sockets (SSL/HTTPs) protect your privacy?

- SSL is used to ‘hide’ sensitive web information (HTTP encrypted and authenticated) – but does it hide everything? (Hintz et al, Simon et al.,. . .)
  - HTTP retrieves many resources per request (HTML page, style, images, . . .)
  - SSL does not disturb timing much – doesn’t hide length well.
  - Attack: profile the website using SSL. For each possible request make a list of retrieved resources and their lengths.
  - Observe the sequence of retrieved resource lengths of the victim – make a (good) guess about which page their correspond to.
- Do better if we observe a sequence of requests (Danezis).
  - Note that users are most likely to follow links on pages.
  - Try to guess not only one request but a sequence – can use hidden Markov models to do this efficiently.

[Danezis]

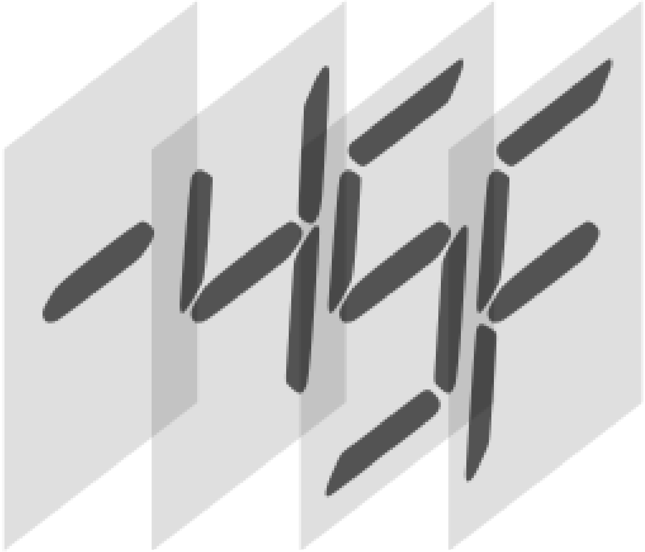
Can I guess which pages you visited before? (Without observing you!)

- Have you visited my competitor’s website before visiting mine?
  - Adversary is a hostile website that tries to determine browsing behavior.
  - Cannot directly observe the victim. The victim only makes one request to the hostile site.
  - Key observation: modern browsers have caches of pages visited – good for efficiency.
  - A resource in the cache will load much faster than if requested from the network.
  - Attack: embed in my website a sequence of pictures from my competitor’s site. Note how long it takes the browser to load these resources. Estimate if they were in the cache. Bingo!
  - Anonymizing proxies do not help! (Attack by Felten et al.)

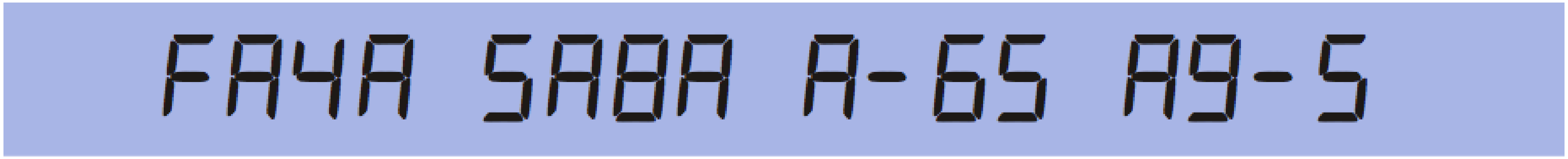
[Danezis]

Creative attacks!

- Can I find out what links you have visited?
    - Use javascript to get the color of the links on the page
        - Can find out if you have visited a specific link (This was fixed)!
    - Set the background color to that of the visited link color, and employ Captchas



4+5 = 9; 4+F = A; 5 + F = 6; 4 + 5 + F = 8



Identification – are two network hosts the same machine?

- How do I know if two different network addresses are the same machine? (CAIDA)
  - Key observation: clock crystals have a variable drift – sensitive to heat conditions.
  - If I can measure the time (i.e. TCP time, web, . . .) I can estimate the drift.
  - If over a period of time the drift matches it is the same machine.
  - Applications: estimating number of consolidated servers, honey pot detection.

[Danezis]

Identification – is one network host many machines?

- A single NAT gateway or firewall can ‘hide’ behind it many network hosts. How many? (Bellovin)
  - Need a way of differentiating different hosts from the traffic the gateway relays.
  - Key observation: many TCP/IP network stacks implement the IPID as a simple counter (Windows). Every time a packet is generated it is increased by one.
  - Attack: Observe all TCP/IP packets from the host, and plot their IPID numbers over time. Fit plausible straight lines – their number is the unique hosts.
- Field of network mapping and network measurements. Attack tools like nmap available and very sophisticated (indirect port scanning).

[Danezis]

Detecting stepping stones

- Traffic analysis can be used for intrusion detection (defence). Problem: I want my firewall to detect whether any host in my network is in fact compromised and used to relay attack streams.
  - Firewall observes all incoming and outgoing TCP/IP connections.
  - Their contents may well be encrypted (particularly if used by attacker).
  - Passive detection: Use inter-packet delays from incoming streams latency to find out relays.
  - Active detection: establish pseudo-random inter-packet delays (watermark stream) in incoming streams and try to detect them in outgoing.

[Danezis]

Location information

- Traffic data from cellular/GSM phones, WiFi base station registration can be mined. Results from early studies:
    - Pascual et al. studied WiFi access point data at HAL. Could infer talks/lectures attended by owner of machine. Could infer relationships by common patterns of movements.
    - MIT Reality Mining: 100 Media Lab staff and students were given mobile phones and traffic data was recorded. Could infer friends (Saturday 8pm), could infer status (entropy of location), could predict movements.
- Location data can be used to infer movements, relationships, status, . . . not just location!

[Danezis]

Traffic analysis resistance

- Over 20 years of research but only recently very active.
  - Anonymous communications – hide link between senders and receivers.
  - Location privacy – reduce the resolution of traffic data / linkability.
  - More generally: Privacy Enhancing Technologies

[Danezis]

Key policy issues: Traffic data retention

- What is traffic data retention
  - E.U. and G8 are thinking of traffic data retention.
  - Certain categories of traffic data kept for years.
  - To facilitate future investigations.
  - ISPs / Mobile providers to bear the cost of storage, access and security.
- What are the issues
  - Introducing a systemic risk of exposure to traffic analysis.
  - Covert communication networks can be established despite even the most stringent retention regimes.
  - Extent to which these data can be used for attack is understated.

[Danezis]

Conclusions

- Traffic analysis has been neglected for too long by the mainstream security community, despite being of vital importance when it comes to operational security.
- Lessons and paradigms from the military world teach us about techniques that can be used to attack civilian networks and security policies.
- The level of sophistication of the attacks, and defenses is advanced – established body of knowledge in the open community should not be ignored.
- Policy decisions that minimize exposure to attack even more important – the opposite is often observed.
- This, along with other network attackers, further fuels the deployment of covert communication networks.

[Danezis]

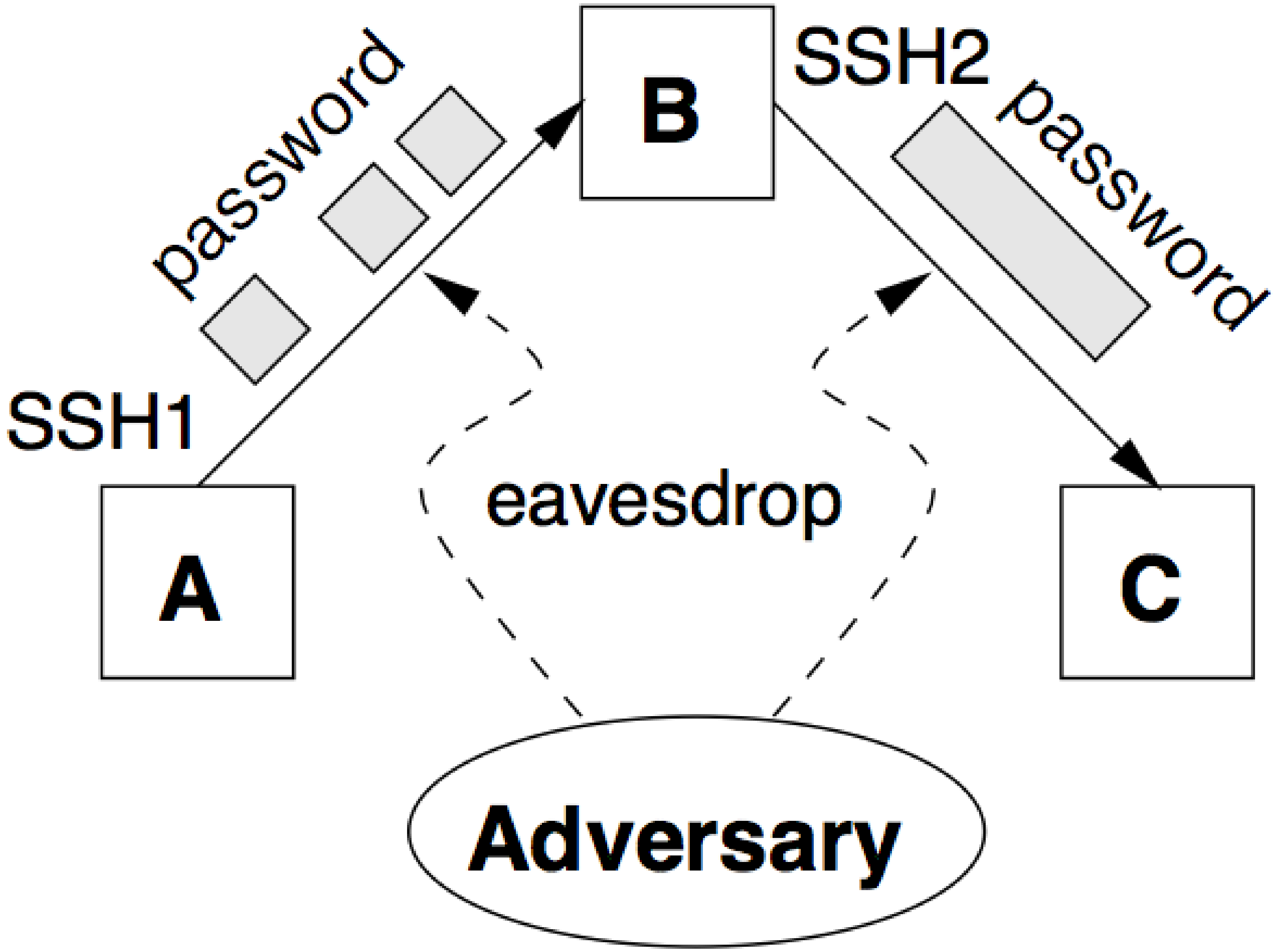
Timing Analysis of Keystrokes and Timing Attacks on SSH, D. Song, D. Wagner, and X. Tian,10th USENIX Security Symposium, 2001.

Secure Shell (SSH)

- Offers an encrypted channel and strong authentication.
- Replaces telnet, rlogin.
- Two seemingly minor weaknesses:
  - Padding: 1-8 bytes
    - Reveals approximate data size
  - Separate packet for each keystroke
    - Leaks timing information of user’s typing

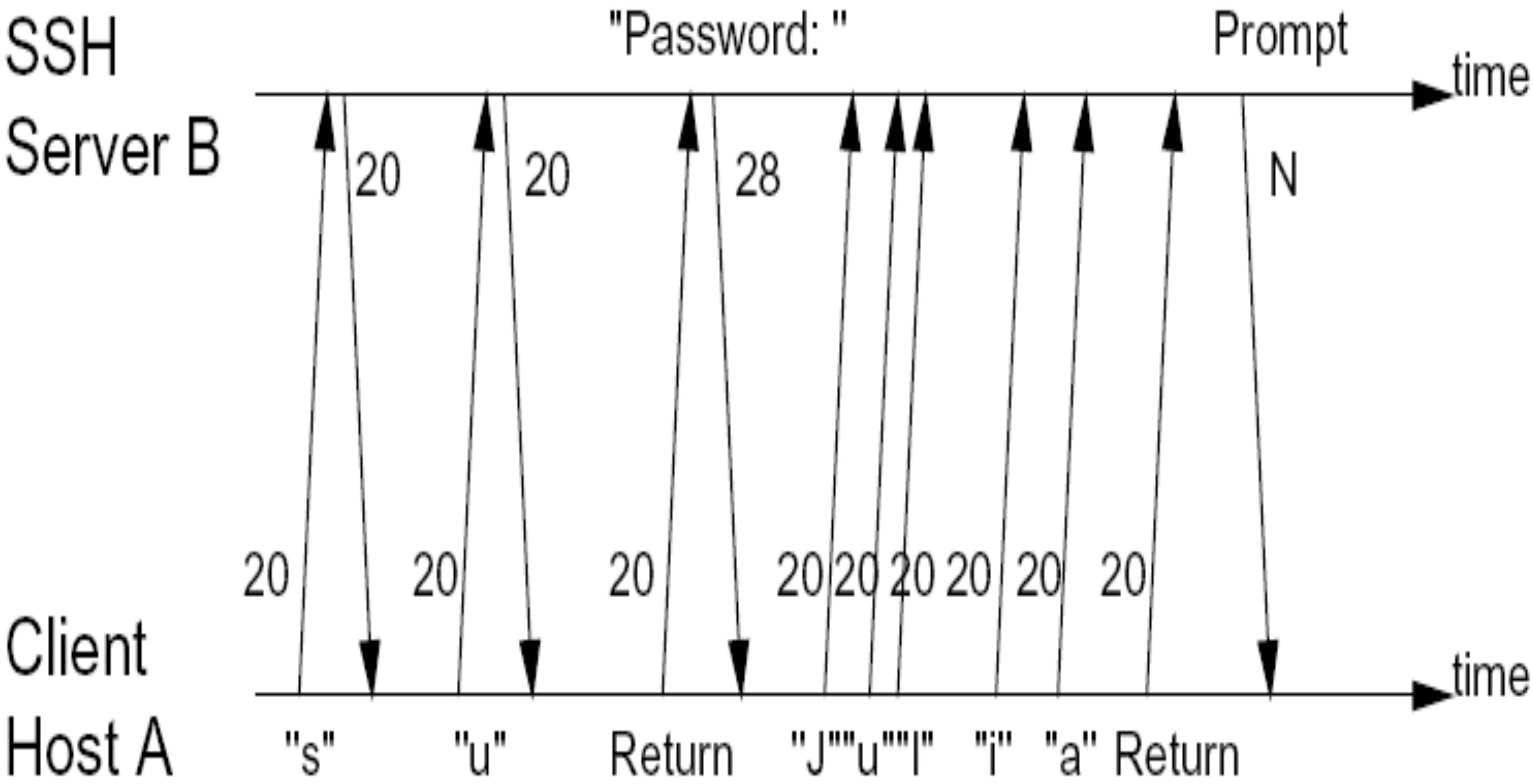
[Peng]

SSH Nested Attack



[Song]

Traffic Signature Attack



[Song]

What is the central idea ?

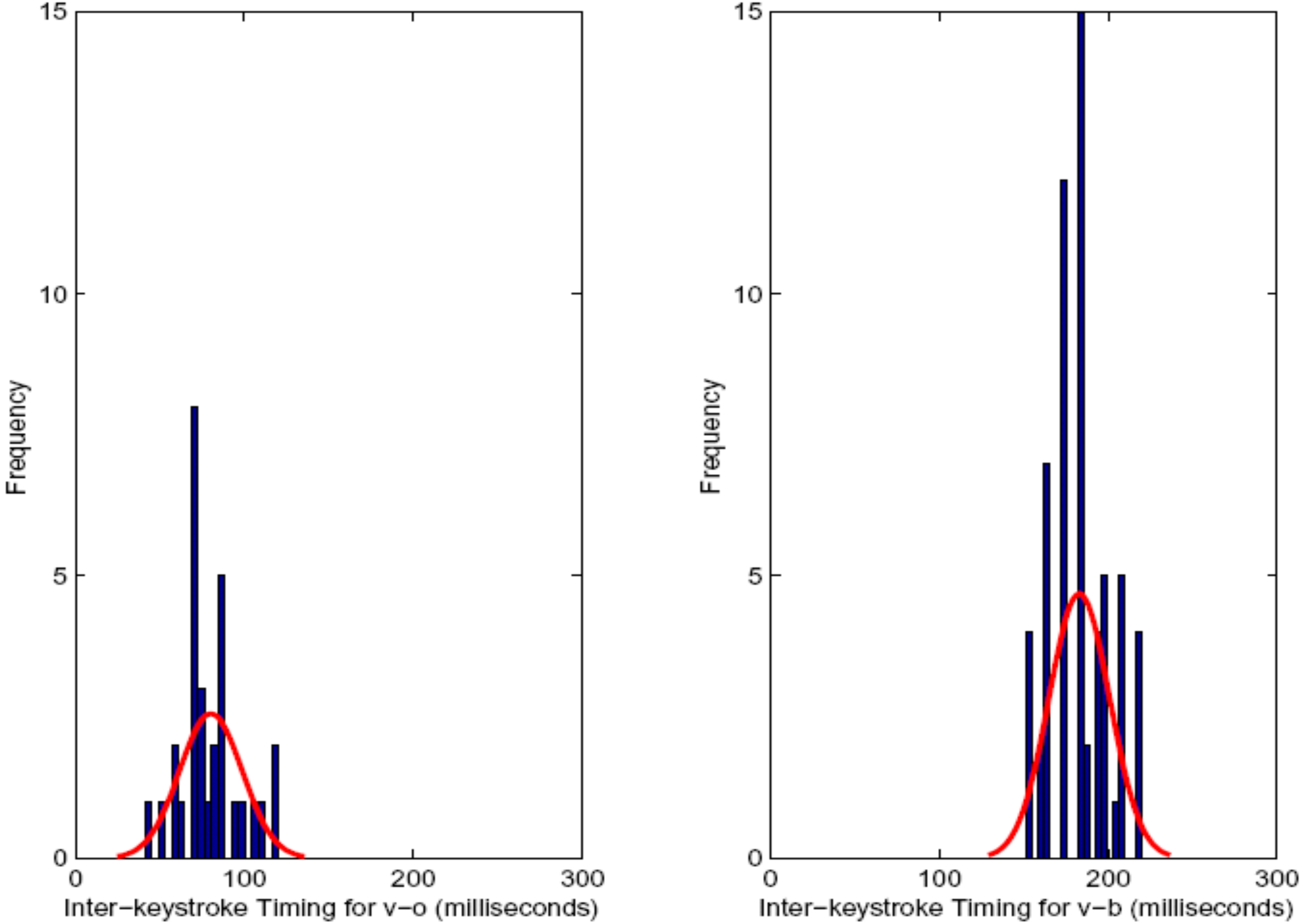
- Exploit SSH Weaknesses
- Obtain Inter-Keystroke Timing (Latency)
- Infer User Password
  - Collect user typing statistics
  - Build a Hidden Markov Model and train it using the data
  - Recommend passwords based on latency data

[Peng]

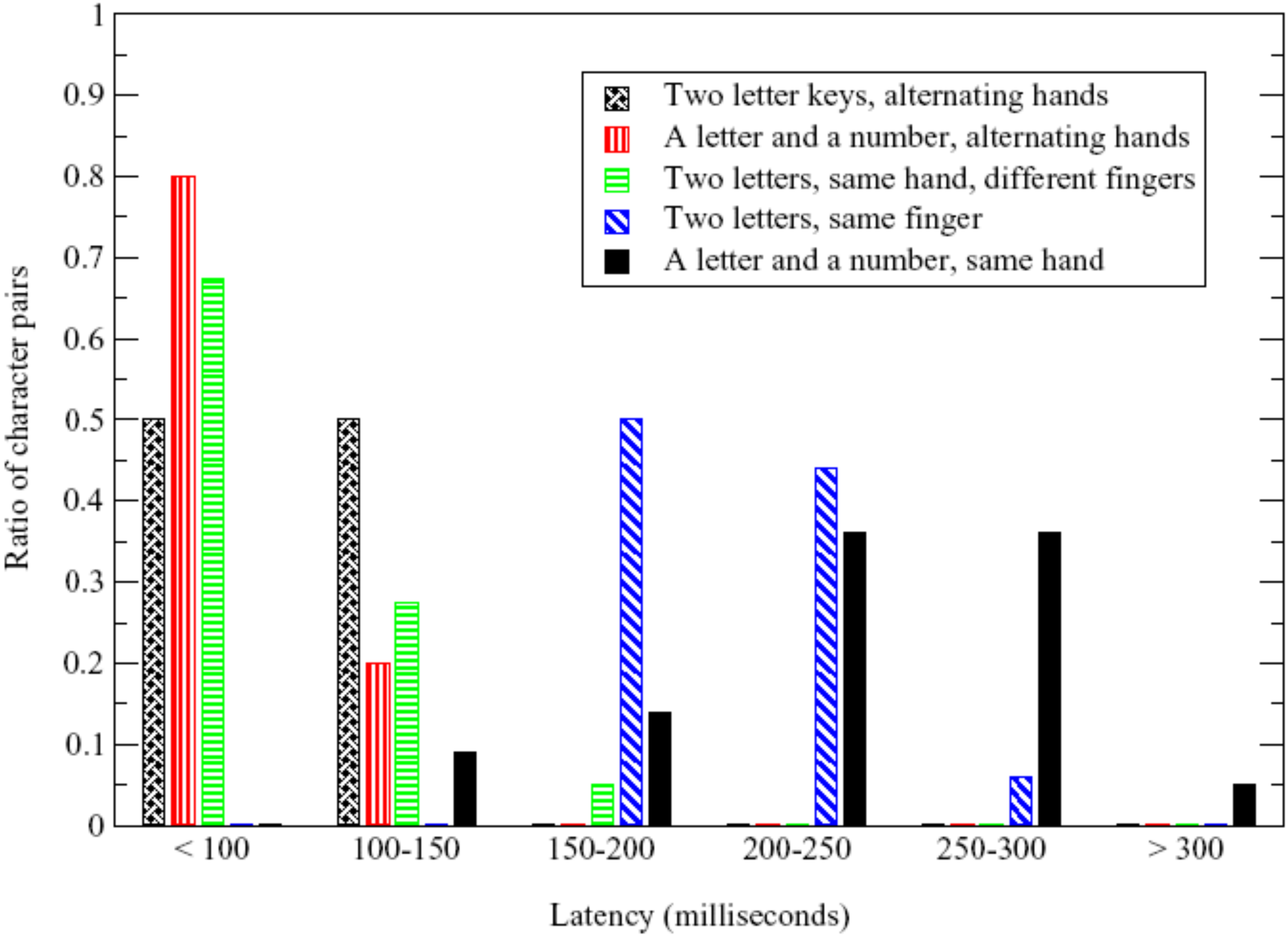
How Are Training Data Collected?

- Pick a pair of characters, e.g. (“v”, “o”)
- Ask users to type the pair for 30-40 times
- Collect latency information
- Repeat for every different pair of characters

[Peng]

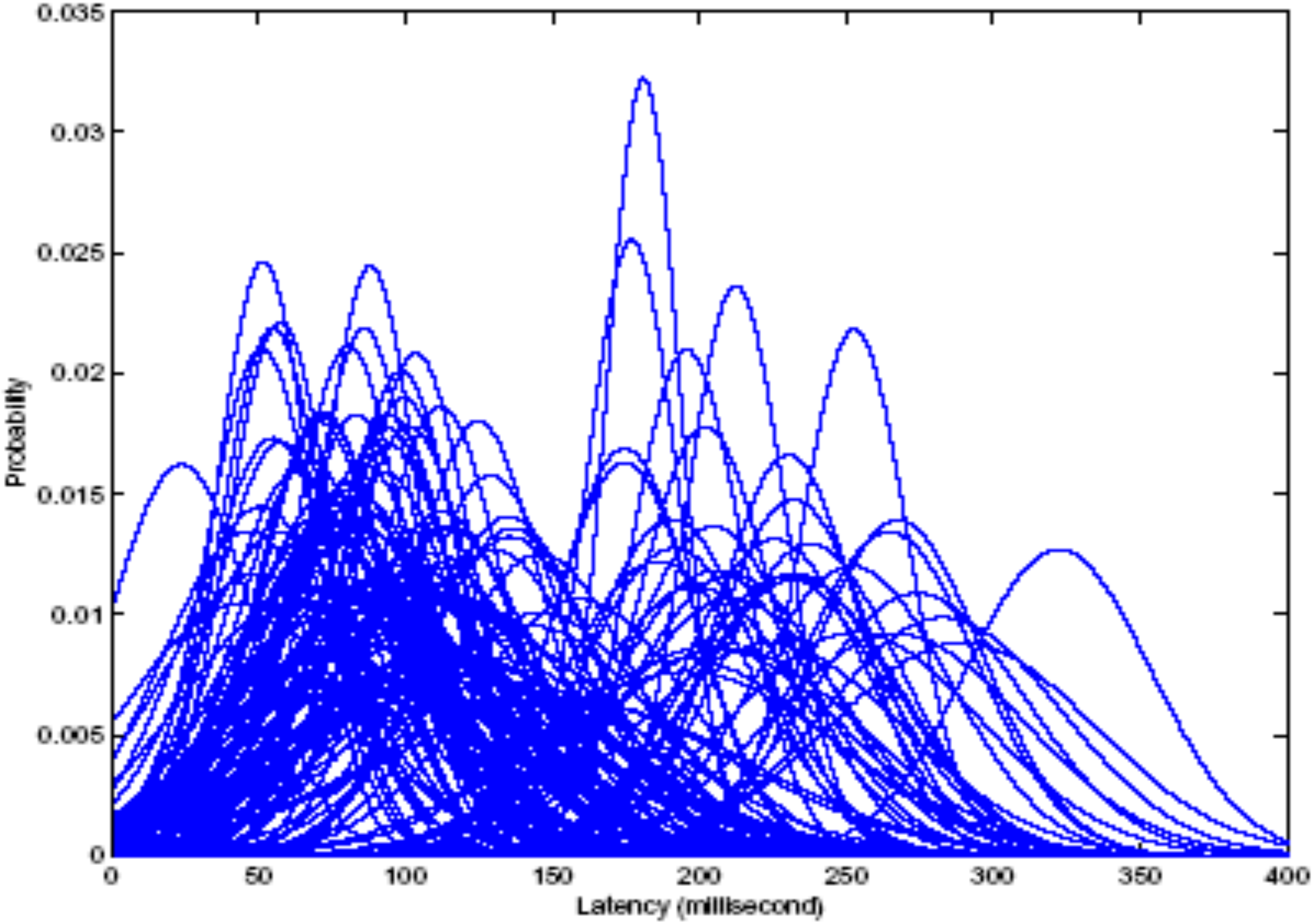

15
10
5
0
Frequency
0
100
200
300
Inter−keystroke Timing for v−o (milliseconds)
15
10
5
0
Frequency
0
100
200
300
Inter−keystroke Timing for v−b (milliseconds)

[Song]


Ratio of character pairs
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Two letter keys, alternating hands
A letter and a number, alternating hands
Two letters, same hand, different fingers
Two letters, same finger
A letter and a number, same hand
< 100
100-150
150-200
200-250
250-300
> 300
Latency (milliseconds)

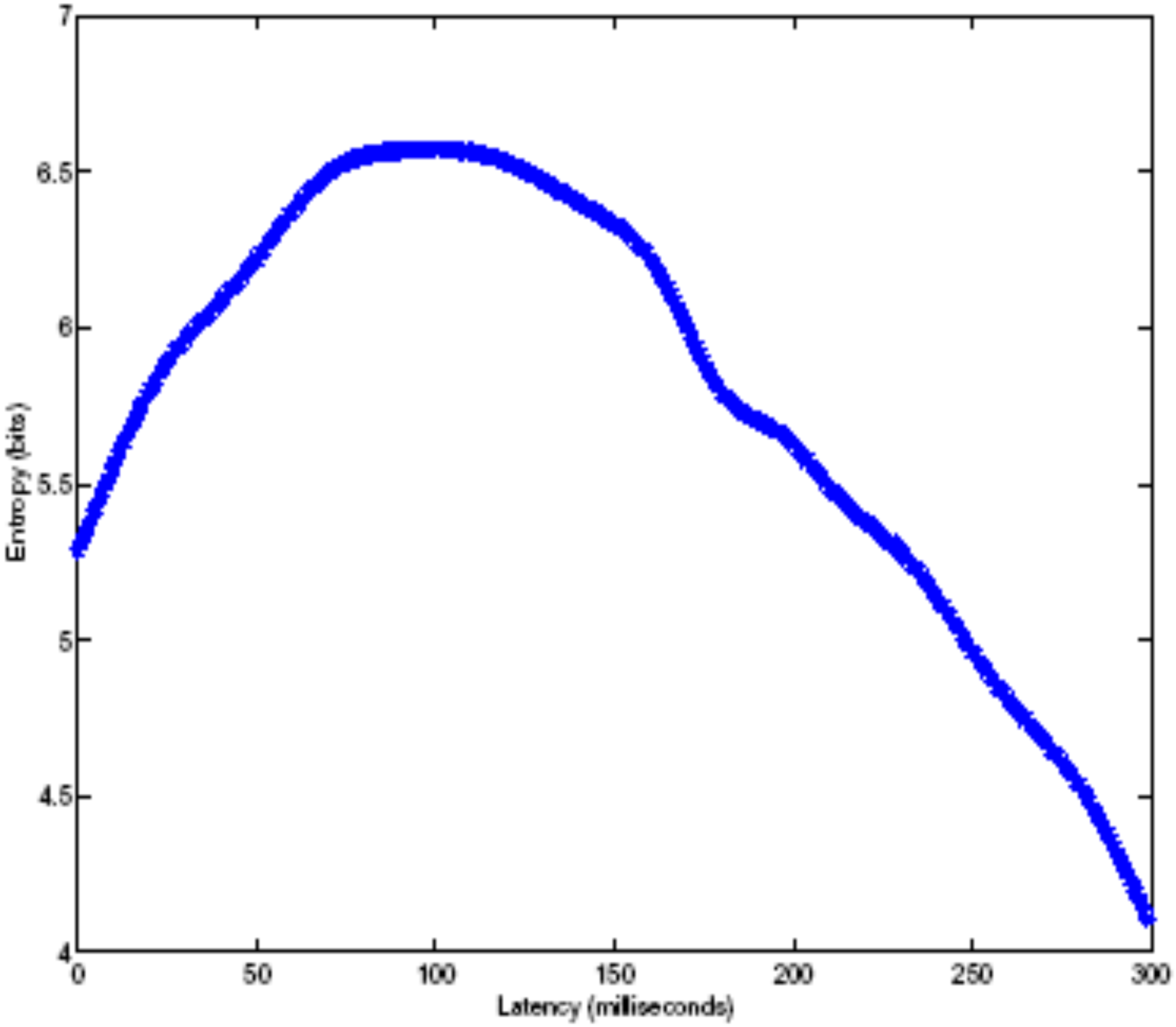
[Song]

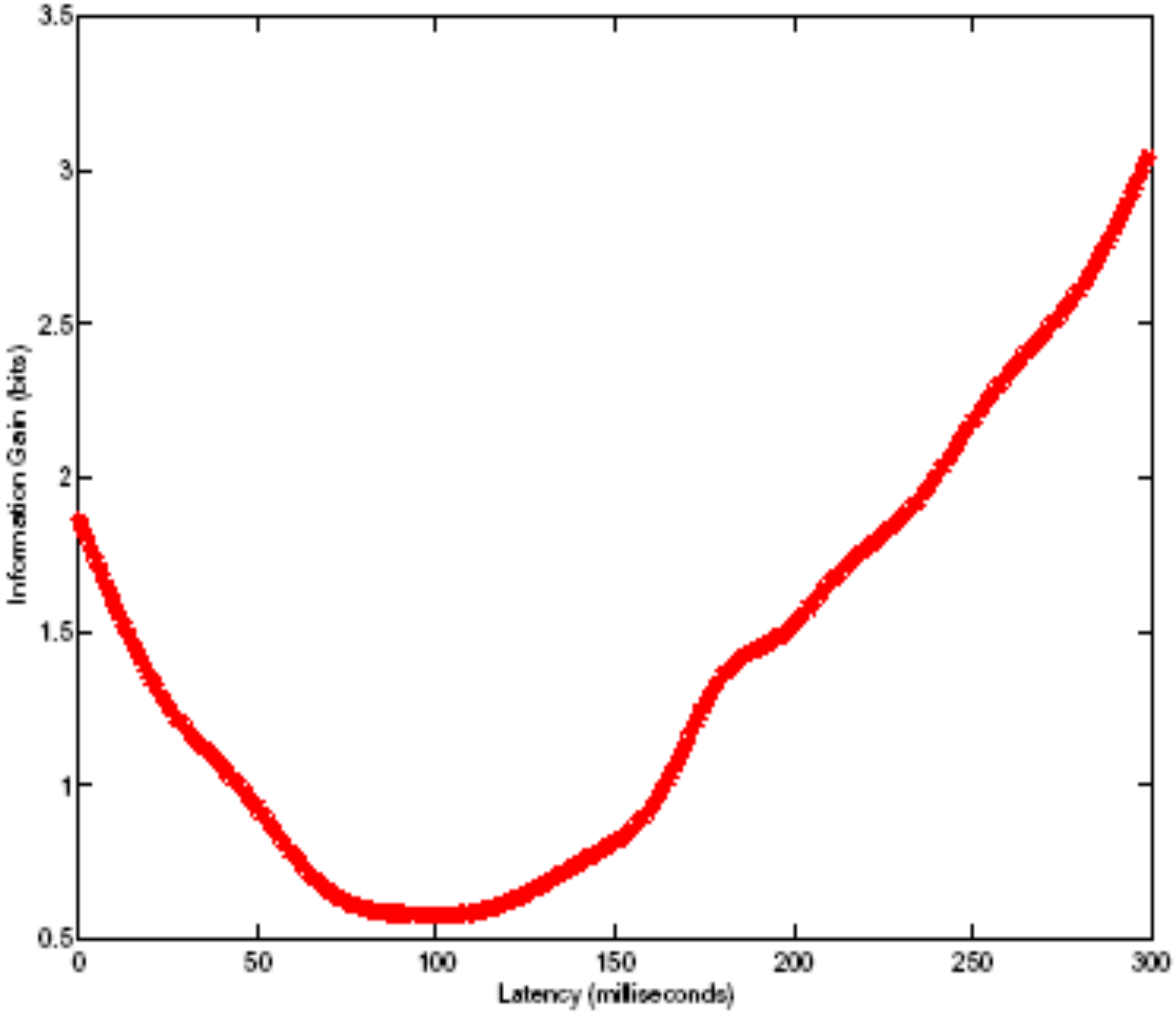
Estimated Gaussian Distributions of All Character Pairs



[Song]

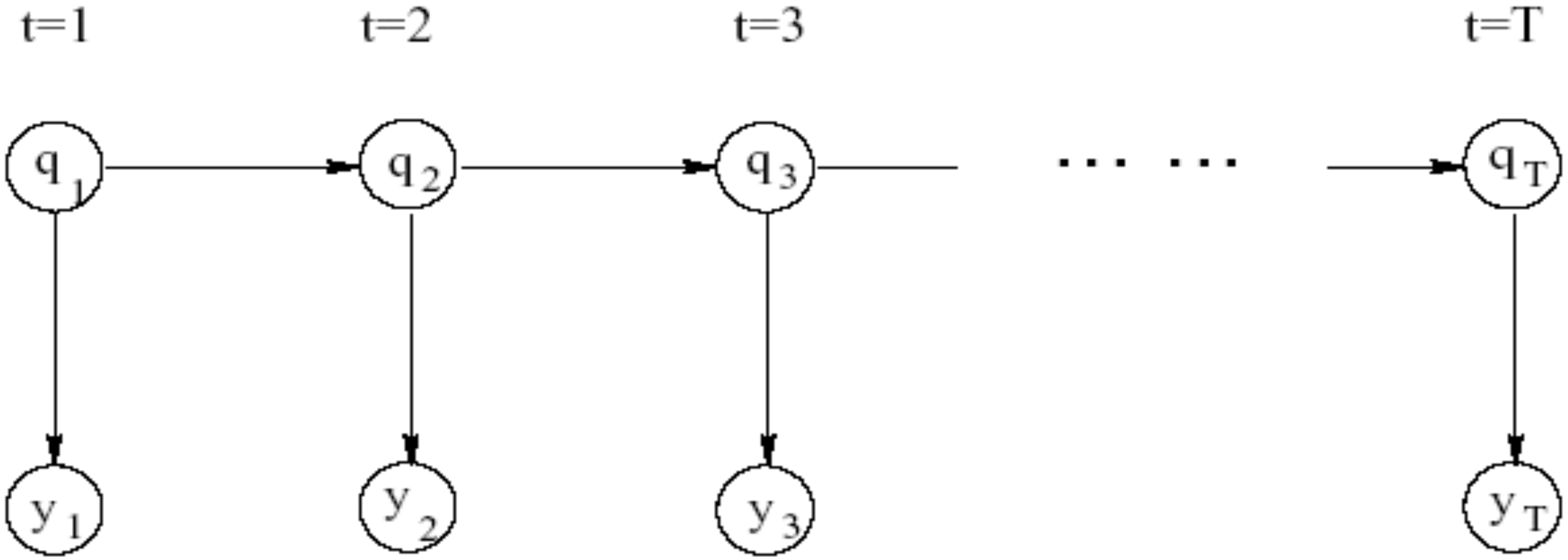
Entropy and Information Gain





[Song]

Inference Algorithm based on HMM

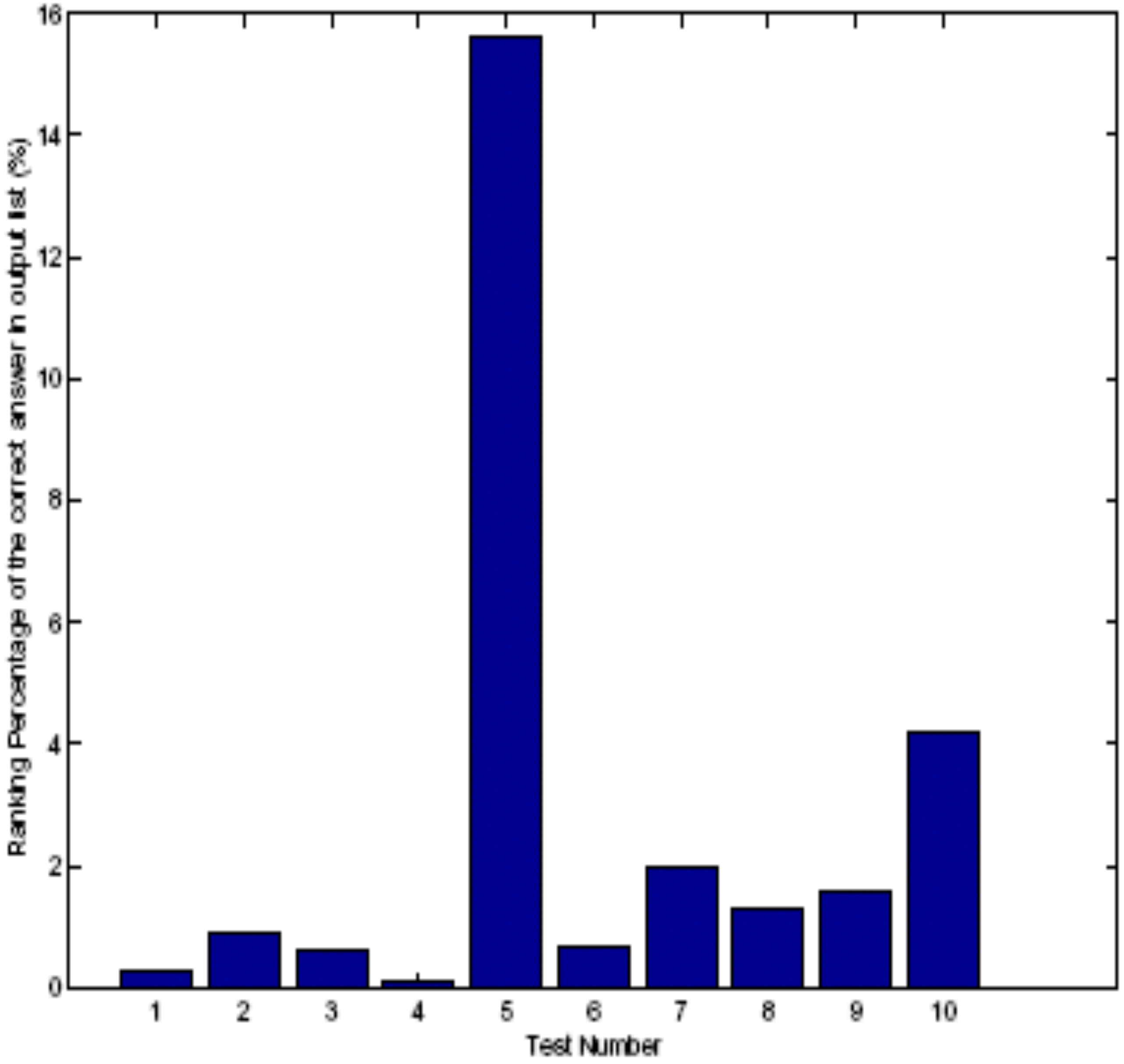


- y = (y1, y2, …, yT): sequence of latencies
- q = (q1, q2, …, qT): sequence of character pairs
- Calculate Pr(q|y): likelihood of the two
- Pr(q|y) essentially gives a ranking for each possible character sequence q

[Peng]

Performance results

- 10 tests all with length 8
- On average the real password is located within top 2.7% of the list.
- Half of the time the password will be in the top 1% of the list.



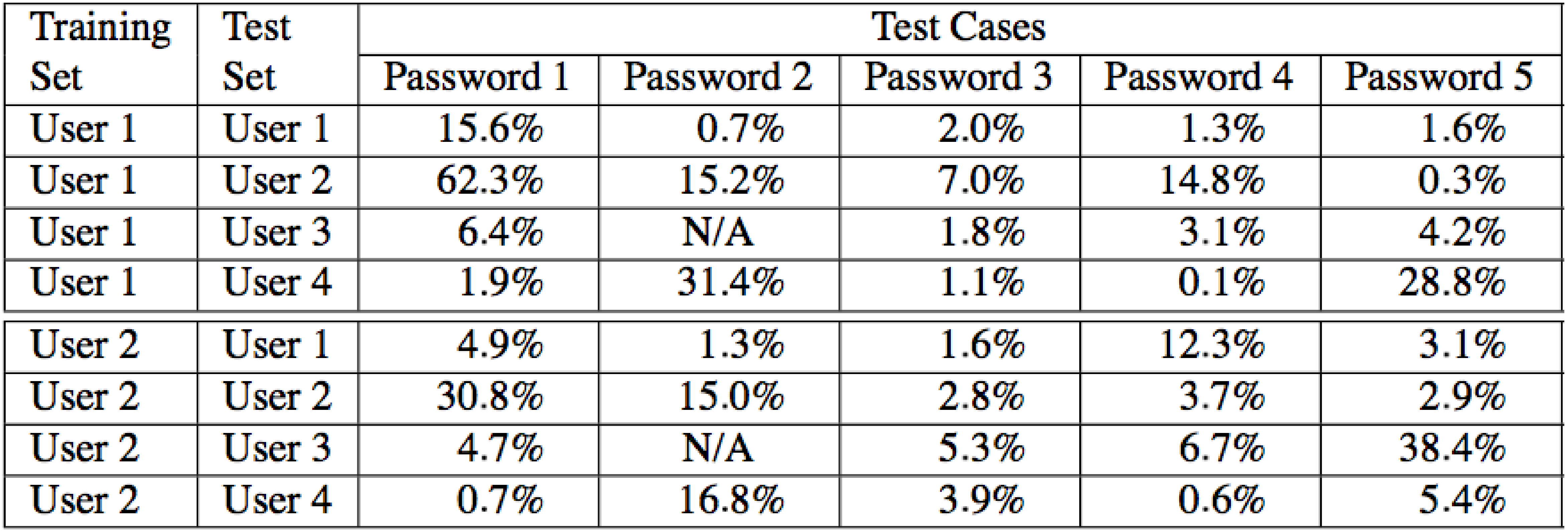
[Peng]

Difference in user typing patterns

- 75% of the time the latencies are the same.
- Typing statistics have a large component in common.
- Attack does NOT need typing statistics from the victim !

[Peng]

Success Rate for Password Inference with Multiple Users



Training Set	Test Set	Test Cases				
		Password 1	Password 2	Password 3	Password 4	Password 5
User 1	User 1	15.6%	0.7%	2.0%	1.3%	1.6%
User 1	User 2	62.3%	15.2%	7.0%	14.8%	0.3%
User 1	User 3	6.4%	N/A	1.8%	3.1%	4.2%
User 1	User 4	1.9%	31.4%	1.1%	0.1%	28.8%
User 2	User 1	4.9%	1.3%	1.6%	12.3%	3.1%
User 2	User 2	30.8%	15.0%	2.8%	3.7%	2.9%
User 2	User 3	4.7%	N/A	5.3%	6.7%	38.4%
User 2	User 4	0.7%	16.8%	3.9%	0.6%	5.4%

[Song]

Countermeasures

- Let the server return dummy packets when it receives keystroke packets from the client.
- Let the client randomly delay sending keystroke packets.
- Let the client send keystroke packets at a constant rate.

[Peng]

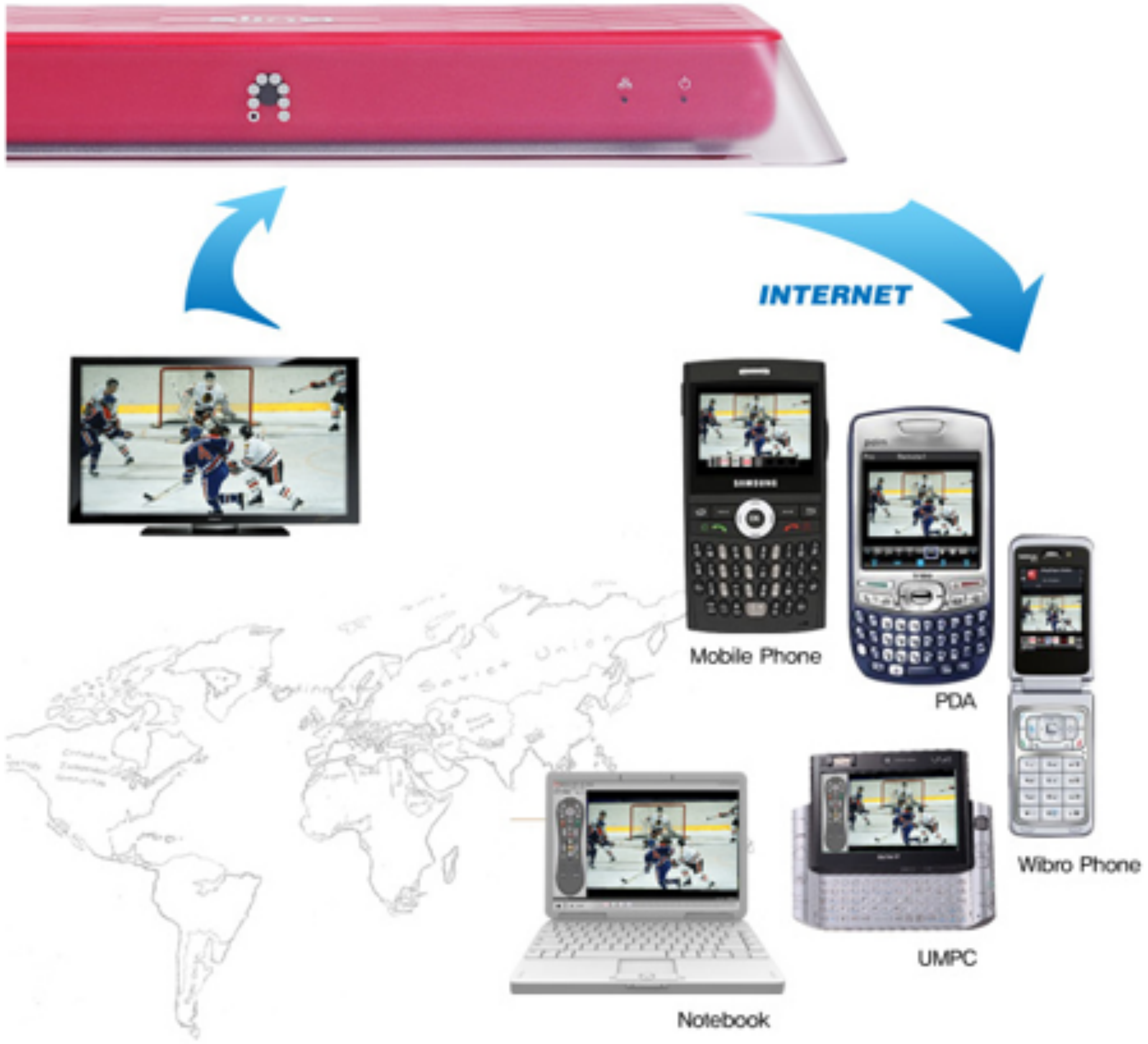
Devices That Tell On You: Privacy Trends in Consumer Ubiquitous Computing, T. Scott Saponas , Jonathan Lester, Carl Hartung, Sameer Agarwal , Tadayoshi Kohno, 16th USENIX Security Symposium, 2007.

Contents of Table

- 3 types of devices are studied in this work
  - Wireless multimedia environments
    - Commercial product (Sling box pro)
    - Information leakage
  - Devices that we have on our persons all the time
    - Commercial product (Nike+iPod Sports kit)
    - Lack of location privacy
  - Devices promoting social activity
    - Commercial product (Microsoft Zune)
- We will cover only the first type in this cass

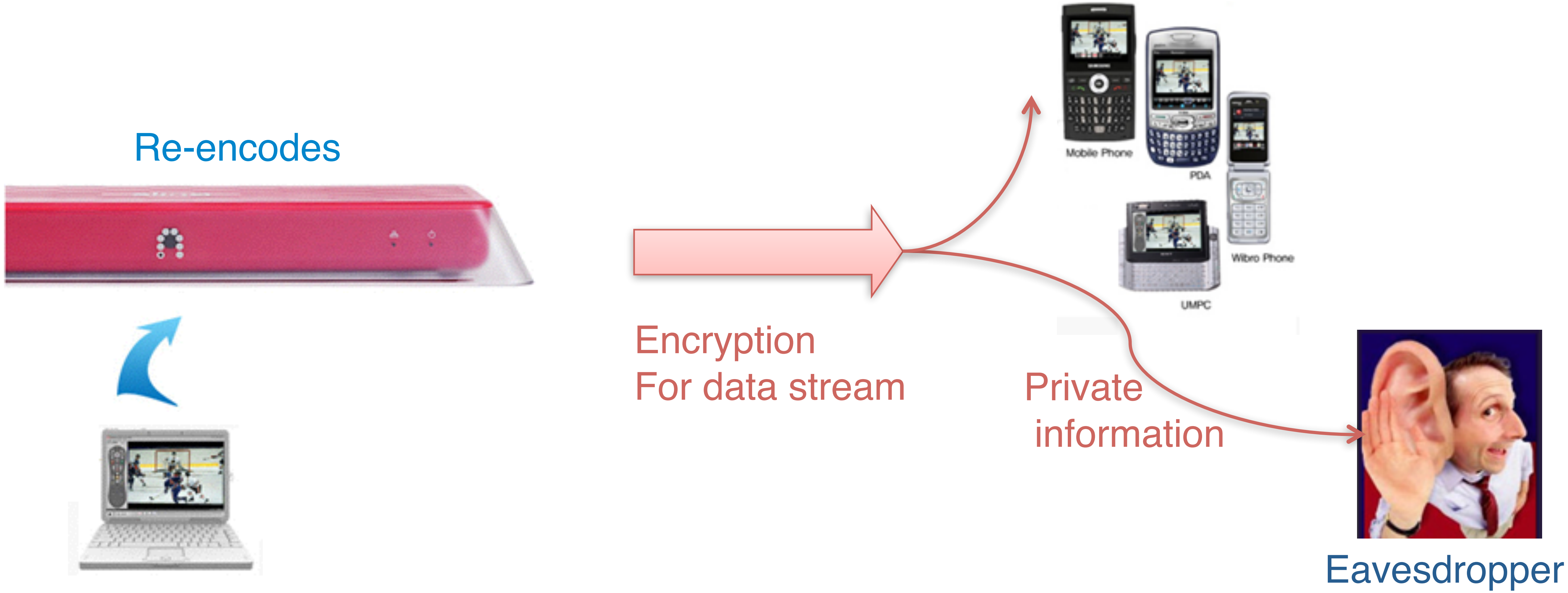
Wireless multimedia environments

- The Slingbox Pro is a networked video streaming device built by Sling Media, Inc.
- It allows users to remotely view (sling) the contents of their TV over the Internet.



[Lee]

Information leakage


Re-encodes

Encryption
For data stream

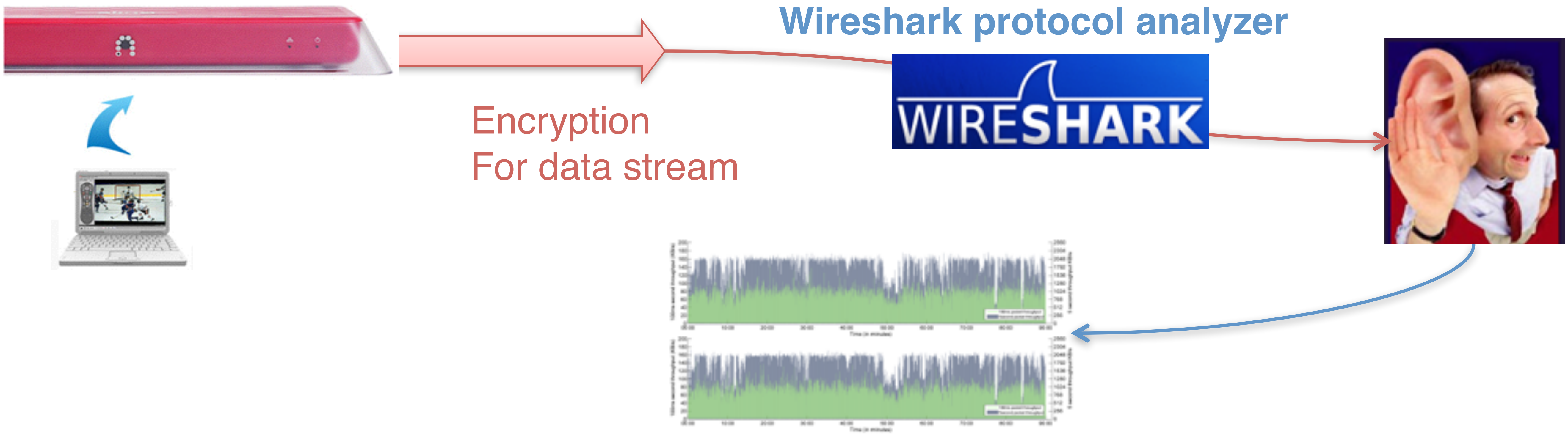
Private
information

Eavesdropper

- Re-encodes the video stream using a variable bit-rate encoder.
- Provides encryption for its data stream regardless of any transport encryption like WPA.
- Private information could be potentially sensitive if the content is illegal (i.e. copyright material)

[Lee]

Eavesdropping algorithms

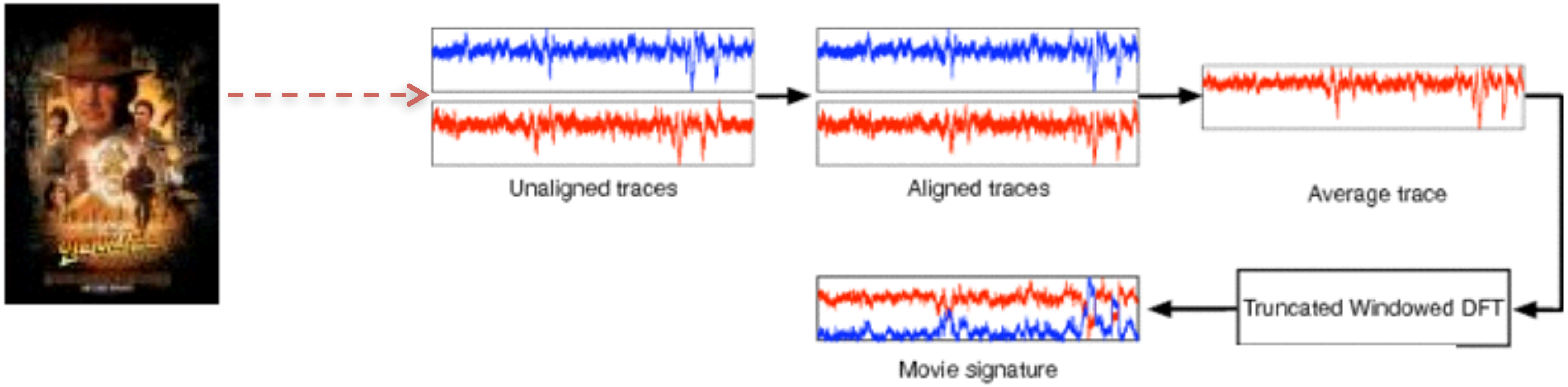


100-millisecond throughput traces

- Using Wireshark protocol analyzer to capture all of the Slingbox encrypted packets to file.
- Use these 100-millisecond throughput traces as the basis for the eavesdropping analysis.

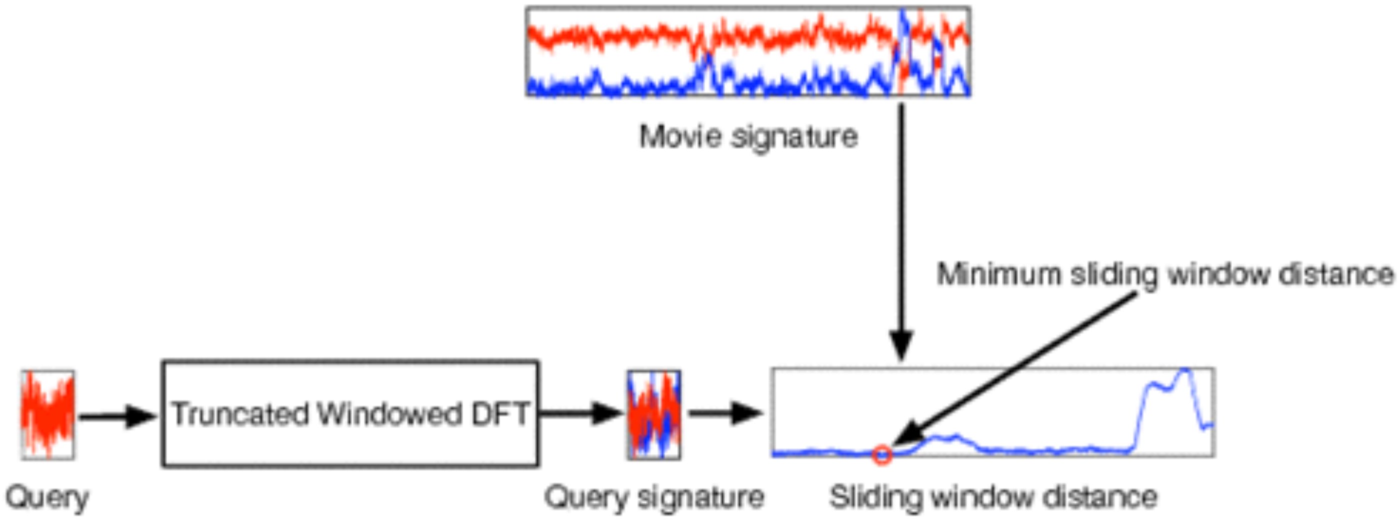
[Lee]

Building a database of movie signatures



- The raw throughput traces corresponding to a movie are aligned and averaged to produce a single composite trace.
- A windowed Fourier transform is performed on the single composite.
- Database of movie signatures is constructed in this manner.

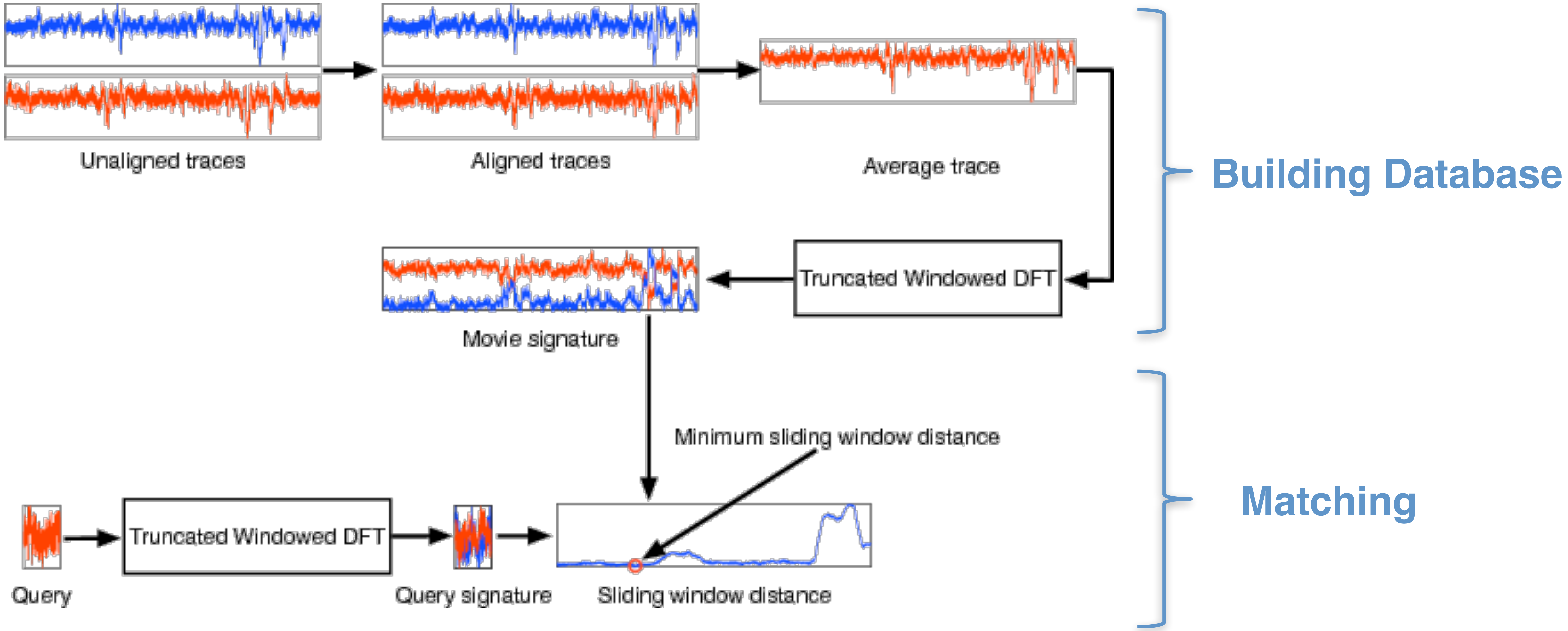
Matching a Query Trace to the Database.



- A query trace is transformed similarly into a signature.
- The minimum sliding window distance between the movie signatures and the query signature is calculated.
- The movie with the minimum distance is declared a match.

[Lee]

Eavesdropping algorithms

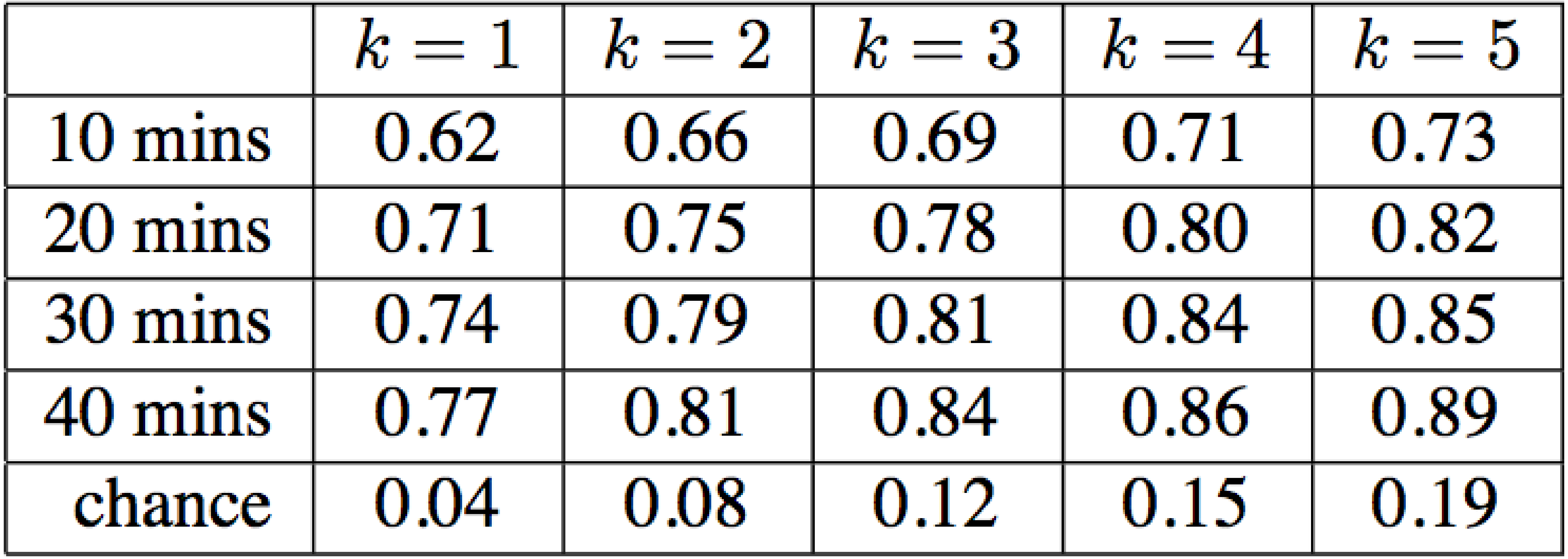


- Building a Database of Reference Traces.
- Matching a Query Trace to the Database.

[Lee]

Experimental Results

- Results obtained over experimental runs on 26 movies



	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
10 mins	0.62	0.66	0.69	0.71	0.73
20 mins	0.71	0.75	0.78	0.80	0.82
30 mins	0.74	0.79	0.81	0.84	0.85
40 mins	0.77	0.81	0.84	0.86	0.89
chance	0.04	0.08	0.12	0.15	0.19

[Saponas]

Information leakage

- Slingbox results provide further evidence that encryption alone cannot fully conceal the contents of encrypted data.
- The implications of results that an adversary might be able to infer information about what videos a user is watching.

[Lee]

Acknowledgments/References

- [Danezis] George Danezis, Introducing Traffic Analysis: Attacks, Defenses and Public Policy Issues. Invited Talk. Santa's Crypto Get-together. Prague, December 2005.
- [Peng] Rui Peng, CDA 6938 Special topic: Research in Computer and Network Security, University of Central Florida, Spring 2007.
- [Song] Timing Analysis of Keystrokes and Timing Attacks on SSH, D. Song, D. Wagner, and X. Tian,10th USENIX Security Symposium, 2001.
- [Saponas] Devices That Tell On You: Privacy Trends in Consumer Ubiquitous Computing, T. Scott Saponas , Jonathan Lester, Carl Hartung, Sameer Agarwal , Tadayoshi Kohno, 16th USENIX Security Symposium, 2007.
- [Lee] Jaejun Lee, http://dreamkorea.tistory.com/attachment/jk74.ppt , May 2008.