
CPSC/ELEN 689 (Topics in NetSec)
(Spring 2004)

Part III

Lecture 2: Traffic Analysis in Detail

Traffic Analysis in Detail: The Menu

- Information Theory Background
 - [I.S. Moskowitz and M.H. Kang, "Covert Channels -- Here to Stay?", COMPASS'94]
 - Attacking **ssh**.
 - [D. X. Song, D. Wagner, and X. Tian, "Timing Analysis of Keystrokes and Timing Attacks on SSH", Usenix'01]
 - Covert Channels despite Countermeasures.
 - [I.S. Moskowitz, R.E. Newman, D.P. Crepeau, A.R. Miller, "A Detailed Mathematical Analysis of a Class of Covert Channels Arising in Certain Anonymizing Networks"]
 - Breaking Anonymity.
 - [Y. Zhu, X. Fu, B. Graham, R. Bettati and W. Zhao "On Flow Correlation Attacks and Countermeasures in Mix Networks", submitted to PET'04]
-

1. Information Theory Background

I.S. Moskowitz and M.H. Kang,
"Covert Channels -- Here to Stay?",
COMPASS'94

Covert Channels

- A Covert Channel is a communication channel that exists contrary to design.
 - We want to utilize or prevent their capacity.
-

Information Theory Background

- Communicating = Encoding + Transmitting + Receiving + Decoding
 - Shannon's **Channel Capacity** gives an upper limit on the rate at which messages can be communicated, as a function of
 - Error tolerance
 - Noise that affects transmission of signal
 - Inputs from the transmitter to the channel are called Input Symbols.
 - Outputs from the receiver before decoding are called Output Symbols.
 - Discrete Channel: input/output alphabets are discrete.
 - Memoryless Channel: Each transmission is independent of the past transmission and is time independent.
-

Information Theory Background (2)

- **Entropy:** $H(X)$ is expected value of “surprise” (or “information”) in random variable X .
- For particular value of x , surprise is $-\log P(x)$
 - If x_i happens with certainty, surprise is zero
 - If x_i never happens, surprise is “infinite”
- Base two is applied to use bit as unit.

$$H(X) = -\sum_i P(x_i) \log P(x_i)$$

Information Theory Background (3)

- Noiseless channel: Information at output = information at input.
- Over noisy channel, information sent is reduced. (If the channel noise is great, this reduces the surprise of seeing any one symbol over another.)
- This is modeled by conditional entropy $H(X|Y)$

$$H(X|Y) = \sum H(X|y_j)P(y_j) = -\sum P(y_j)P(x_i|y_j)\log P(x_i|y_j)$$

- Examples: $H(X|X) = 0$ $H(X|Y) = H(X)$ if X and Y independent
-

Information Theory Background (4)

- For a discrete memoryless channel, noise is expressed by conditional probabilities $p_{ij} = P(y_j | x_i)$
- Distributions for X and the noise determine $H(X)$ and $H(X|Y)$, respectively.
- Mutual information $I(X, Y)$ [bits per channel usage] measures how much information is actually sent from input X to receiver Y .

$$I(X, Y) = H(X) - H(X|Y)$$

- Transmitter cannot do anything about the noise. However, she send different symbols with different frequencies (I.e. different distributions for X). This affects the amount of information sent to receiver.
- C is the maximum amount of information [bits per channel usage] that can be sent over discrete memoryless channel:

$$C = \max I(X, Y)$$

2. Attacking SSH

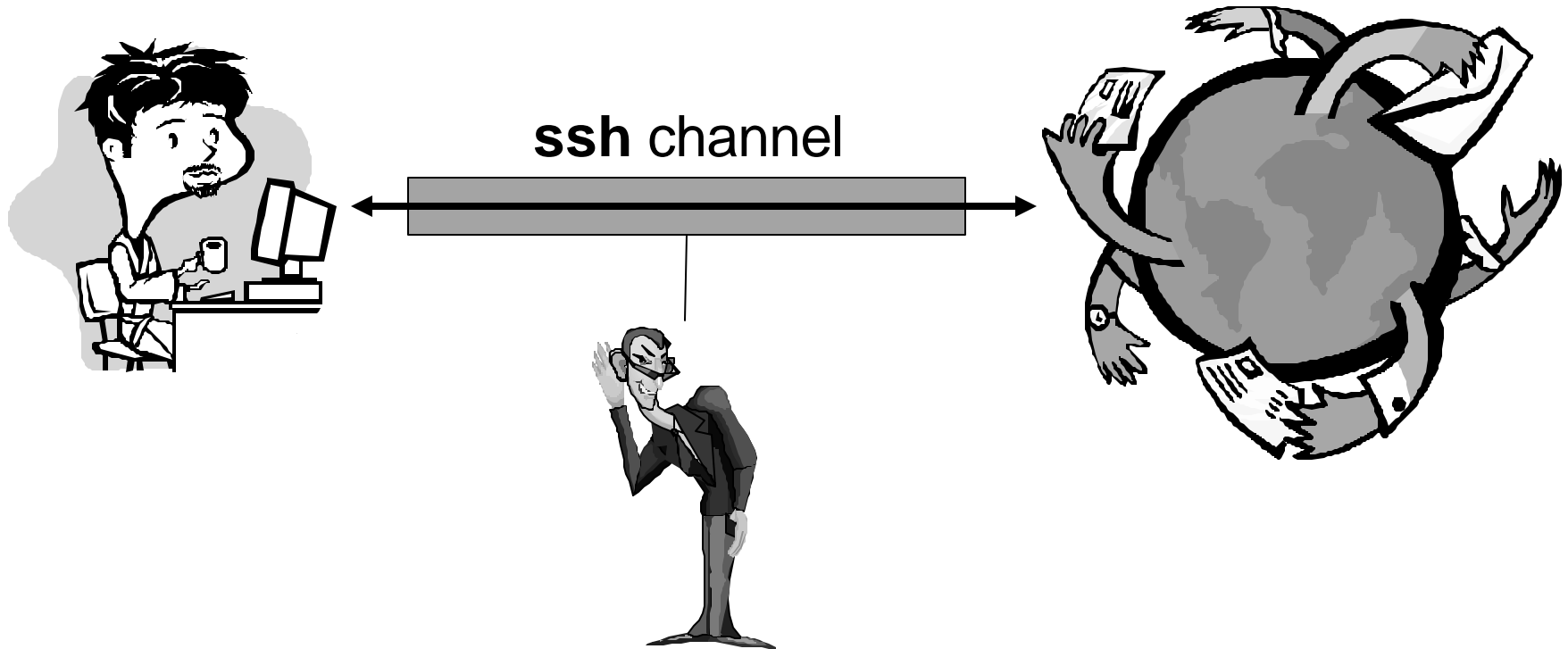
D. X. Song, D. Wagner, and X. Tian

“Timing Analysis of Keystrokes and Timing Attacks
on SSH”

Usenix'01

Timing Analysis of Interactive Applications

Example: Attacking **ssh**. [D. Wagner et al. "Timing Analysis of Keystrokes and Timing Attacks on SSH", Usenix'01]

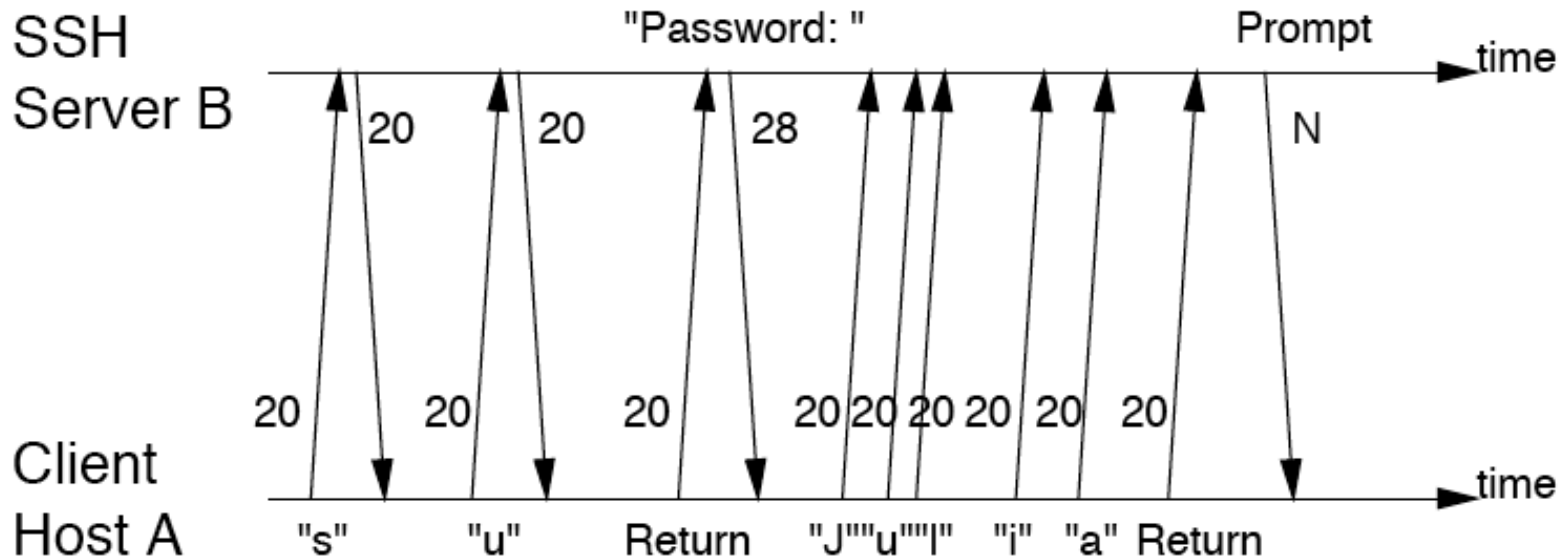


Eavesdropping SSH

- SSH leaks information in two ways:
 - Transmitted packets are padded only to an eight-byte boundary.
 - In interactive mode, every keystroke that a user types is sent to the remote machine in a separate IP packet.
 - Attacks on SSH to infer content of passwords
 - Traffic Signature Attack: Specific commands (e.g. su) may cause specific packet pattern on network (signature).
 - Multi-User Attack: User on remote machine can monitor (e.g. using ps or top) commands invoked by SSH.
 - Nested-SSH Attack: Password batching cannot tunnel through SSH channels.
-

Traffic Signature Attack for SSH

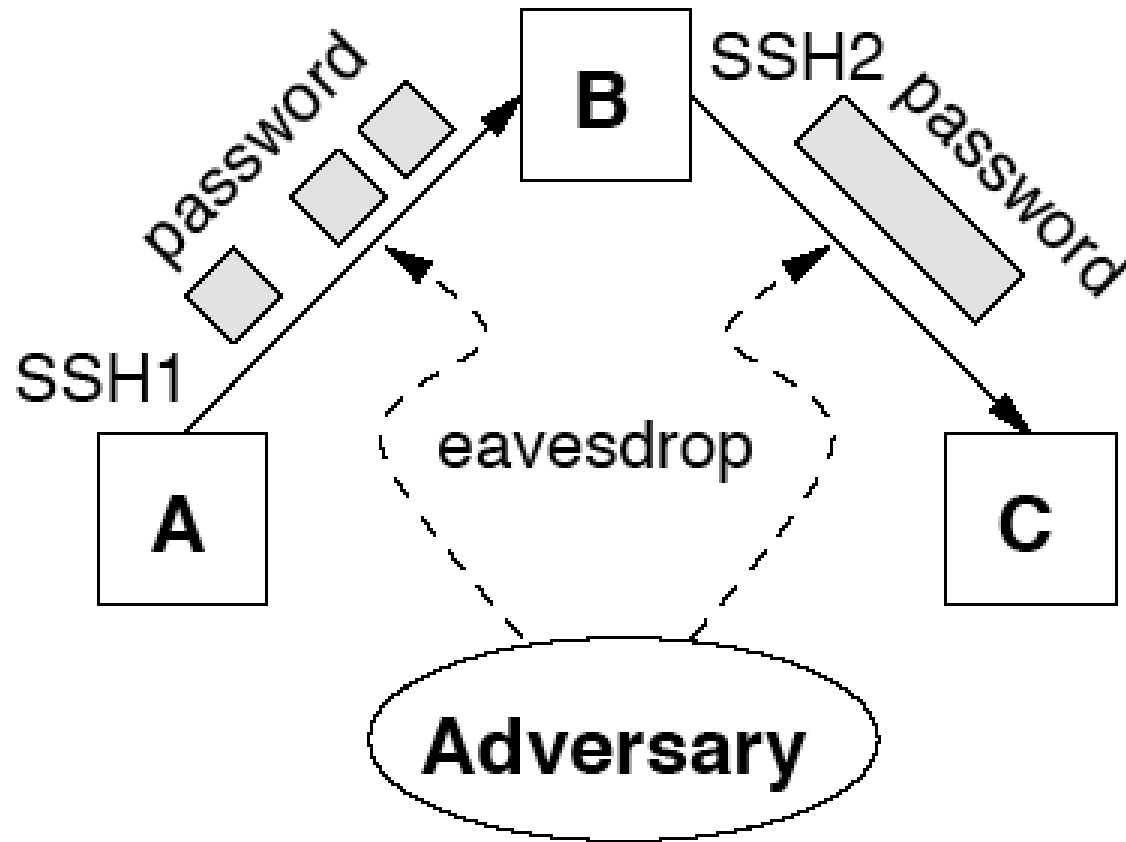
© D. Wagner



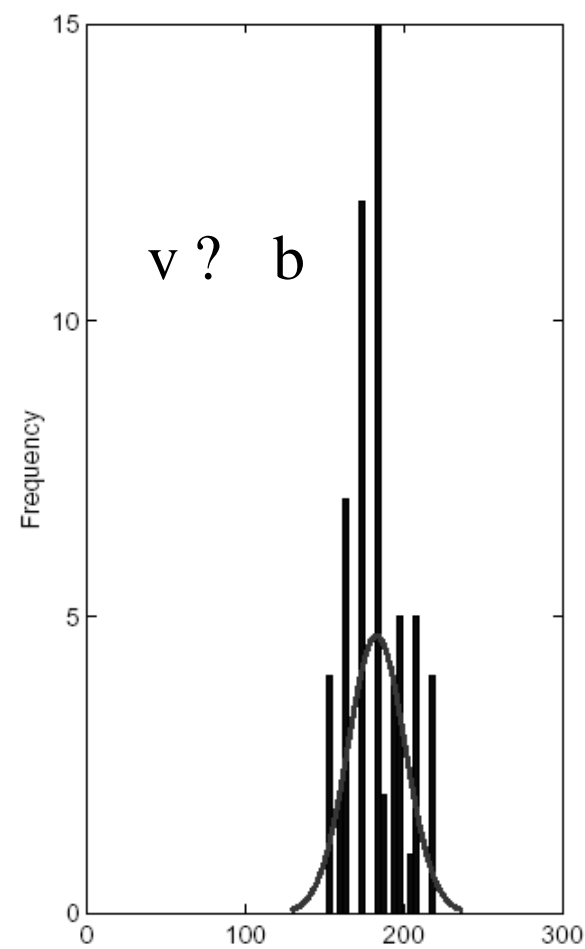
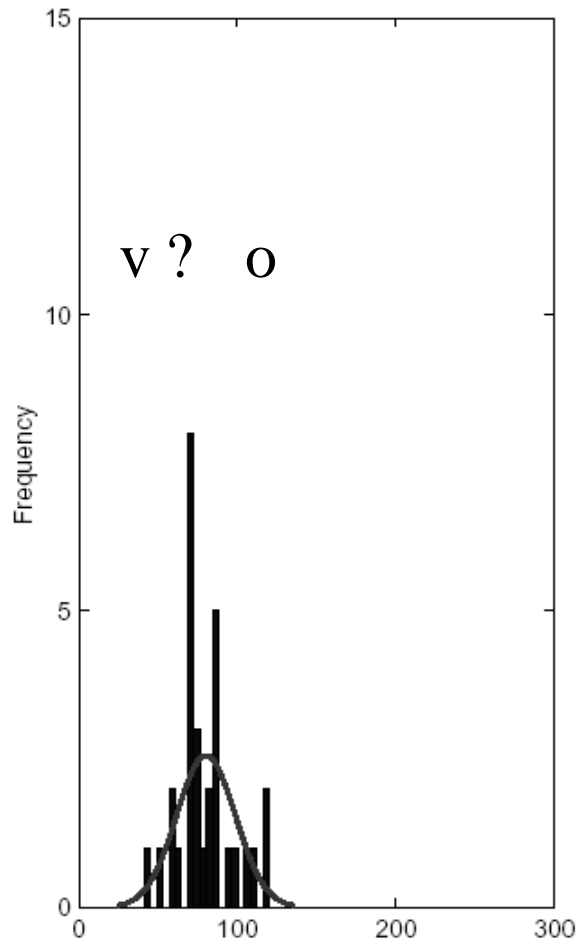
- By checking traffic against this "su" signature, attacker can identify when command "su" is used.
- This works for other applications as well (e.g. PGP)

Nested SSH Attack

© D. Wagner



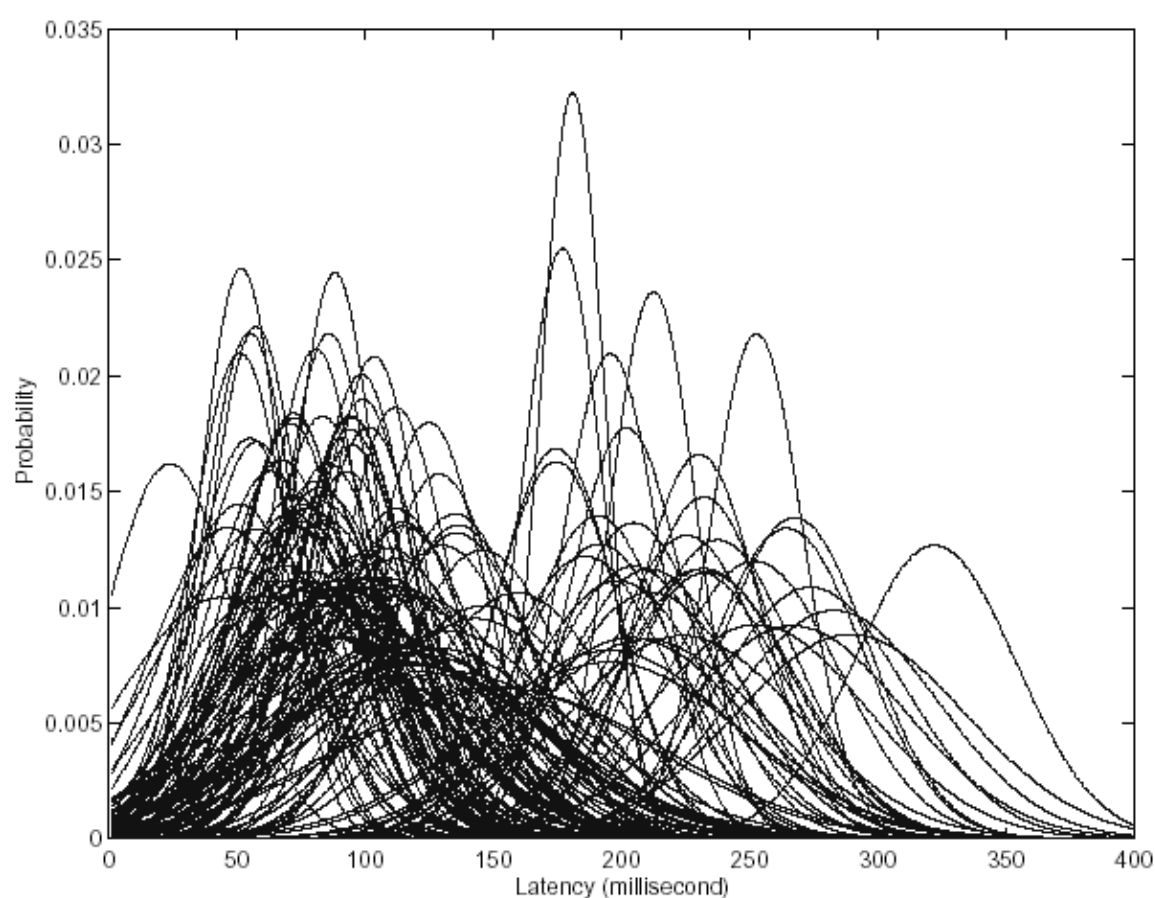
Character-Pair Delays



© D. Wagner

Measured delay between characters.

Character-Pair Delay Distributions



© D. Wagner

Estimated Gaussian delay distributions of character pairs collected from a user.

Information Gain

(here we repeat intro)

- Upper bound on how much information attacker can extract from timing information.
- Entropy of probability distributions of character pairs:

$$H_0(Q) = -\sum_q P(q) \log P(q)$$

- If attacker knows latency y_0 between the two keystrokes, the estimated entropy becomes

- Information gain induced by observation of latency y_0 is $H_1(Q | Y = y_0) = -\sum_q P(q | y_0) \log P(q | y_0)$ where $P(q | y_0) = \frac{P(y_0 | q)P(q)}{\sum_q P(y_0 | q)P(q)}$

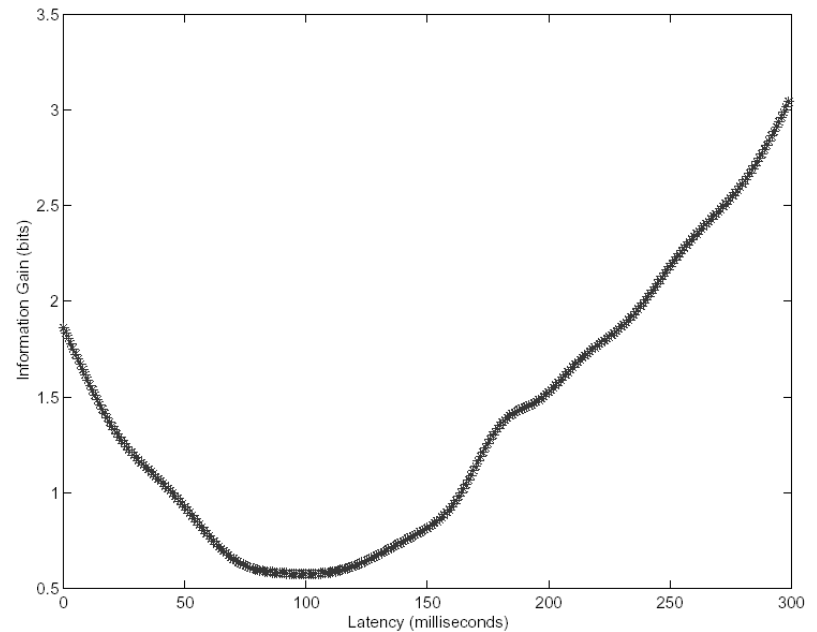
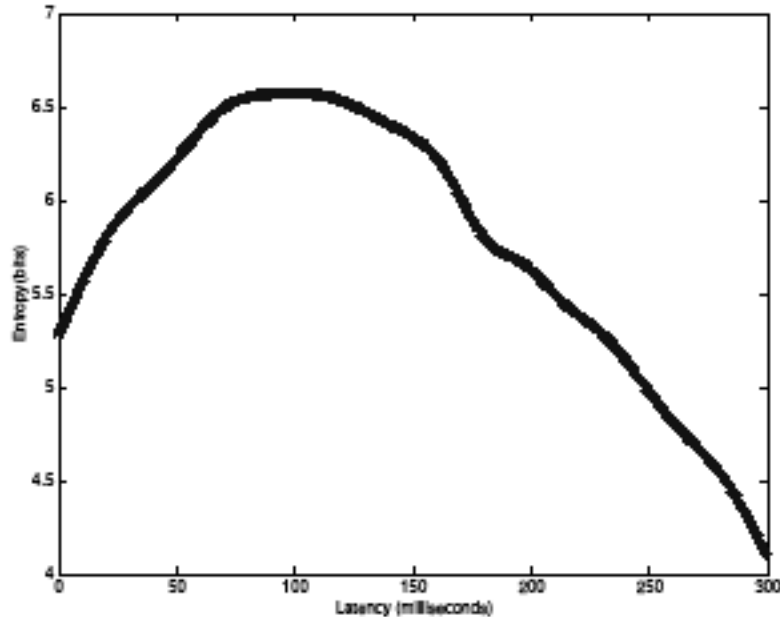
$$I(Q, Y) = H_0(Q) - H_1(Q | Y) = H_0(Q) - \int P(y_0) H_1(Q | Y = y_0) dy_0$$

where

$$P(y_0) = \sum_q P(y_0 | q)P(q)$$

Information Content of Keystroke Data

© D. Wagner



Entropy of Character Pairs
Given Latency of Observation

Information Gain

Numerical computation gives $I(Q, Y) = 1.2$ for character pairs selected uniformly at random.

This means 1.2 bits are leaked per character pair.

Inferring Character Sequences

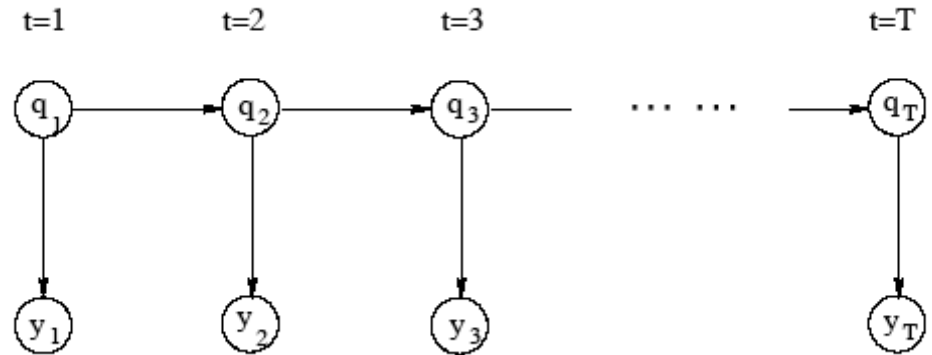
- Overview:
 - Model relationship of latencies and character sequences as Hidden Markov Model
 - Use n-Viterbi algorithm to output n most likely candidate character sequences.

Hidden Markov Models

- Markov Model: Description of finite-state stochastic memoryless process.
 - Memoryless: Probability of transition from current state to other state depends only on current state (not on any prior state).
 - Hidden Markov Model (HMM): The current state of the process cannot be directly observed.
 - Instead, some outputs from the state are observed.
 - Probability distribution of possible outputs given the state is dependent only on the state.
 - Information about the prior path of the process can be inferred from sequence of observed outputs of the states.
-

HMM and Character Sequences

- Character sequence
 K_0, \dots, K_T
- Character pairs
 q_1, \dots, q_T , where $q_t = (K_{t-1}, K_t)$
- Observed latencies
 y_1, \dots, y_T



- Modeling of typing of character sequence as HMM relies on two assumptions:
 - Probability of transition from current state to another only dependent on current state. (Why would this be a problem?)
 - Probability distribution of latency observation only dependent on current character pair and not on previous characters in sequence. (Why would this be a problem?)
-

Determining the Character Sequence

- Given: observation vector $\mathbf{y} = (y_1, y_2, \dots, y_T)$
 - We want: real character sequence that user has typed.
 - For each possible character sequence $\mathbf{q} = (q_1, q_2, \dots, q_T)$ we compute $P(\mathbf{q}/\mathbf{y})$, i.e. how likely \mathbf{q} is given the observation \mathbf{y} .

 - Which is the most likely sequence?
 - Which sequence \mathbf{q}^* has highest value for $P(\mathbf{q}/\mathbf{y})$ for all possible \mathbf{q} for given \mathbf{y} ?

 - Naive approach searches exhaustively: $O(|Q|^T)$
 - Viterbi Algorithm with Dynamic Programming: $O(|Q|^2T)$
-

The Viterbi Algorithm

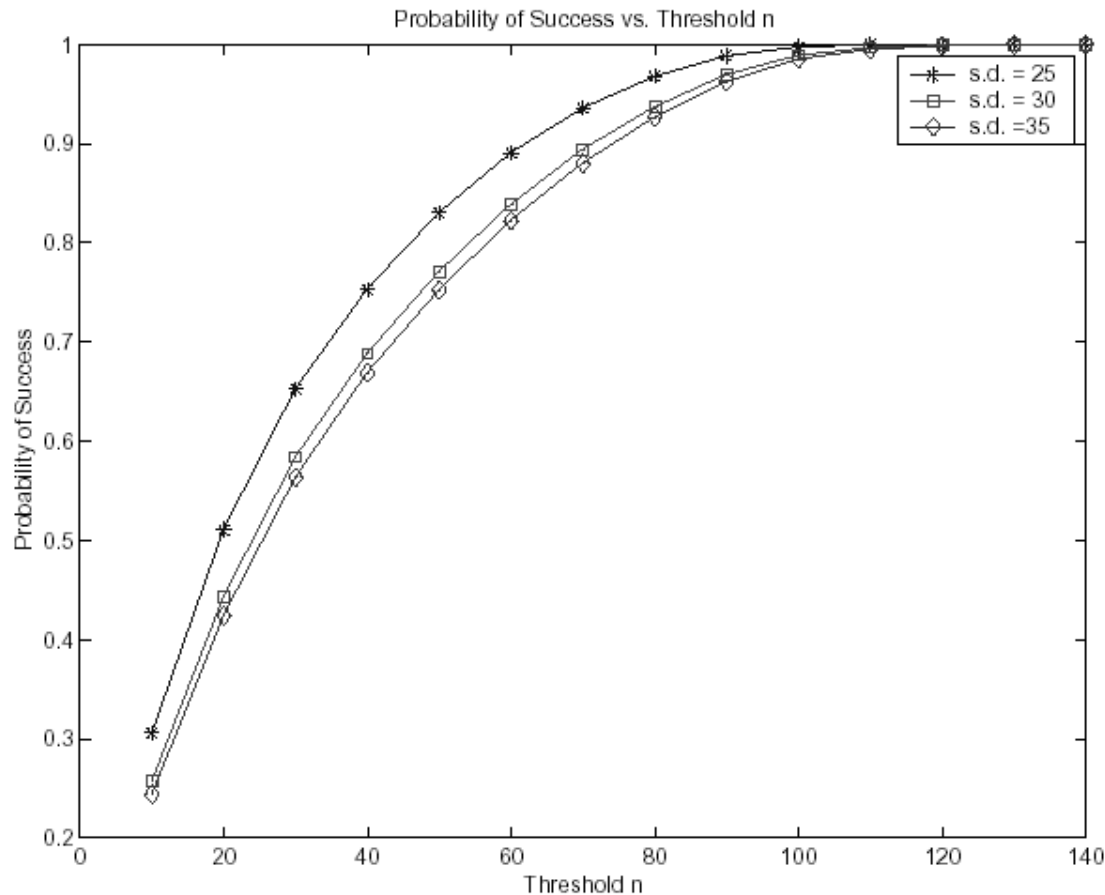
- Given an observation (y_1, y_2, \dots, y_T) of HMM.
 - Compute the most likely sequence (q_1, q_2, \dots, q_t) that generated the observation for each $t = 1, 2, \dots, T$.
 - Let $S(q_t)$ be the most likely sequence at time t that ends with state q_t , with posterior probability $V(q_t)$.
 - Viterbi Algorithm:
 - Start with $S(q_1) = q_1$ and $V(q_1) = P(q_1|y_1)$
 - Compute $V(q_t) = \max_{q_{t-1}} P(y_t|q_t)P(q_t|q_{t-1})V(q_{t-1})$
 - Let q_{t-1} be state that maximized $V(q_t)$
 - Define $S(q_t) := S(q_{t-1})|q_t$
-

n-Viterbi for Highly Overlapping Observation Distributions

- Latency distributions overlap.
 - Therefore, most likely sequence may well not be the correct sequence.
 - Extend Viterbi algorithm to return n most likely sequences
 - n -Viterbi Algorithm:
 - Start with $S^n(q_1) = q_1$ and $V^n(q_1) = P(q_1|y_1)$
 - Compute $V^n(q_t) = \text{nmax} \{ P(y_t|q_t)P(q_t|q_{t-1})v : q_{t-1} \in Q, v \in V^n(q_{t-1}) \}$
 - nmax denotes the set of n largest values.
 - $S^n(q_t)$ is the set of n highest-probability sequences corresponding to choice of $V^n(q_t)$.
-

Effectiveness of Keystroke Analysis Attack

© D. Wagner



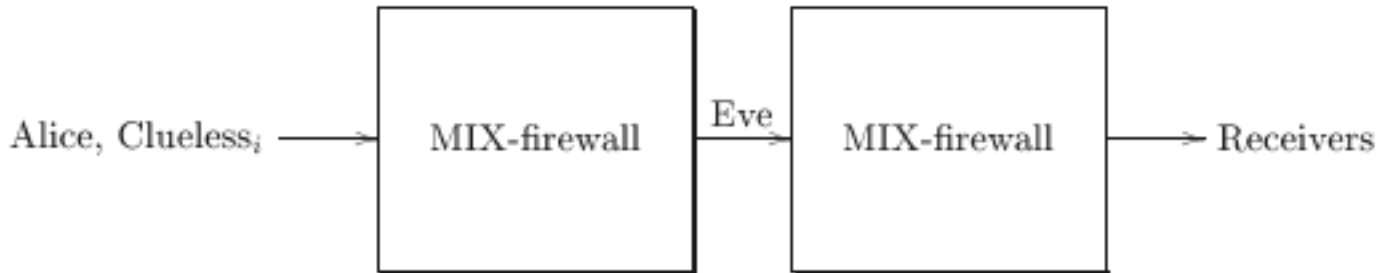
Probability of success in breaking password within n attempts.

3. Covert Channels Despite Countermeasures

I.S. Moskowitz, R.E. Newman, D.P. Crepeau, A.R. Miller

“A Detailed Mathematical Analysis of a Class of Covert Channels Arising in Certain Anonymizing Networks”

I dealized Batching Mixes



- Mixes batch at fixed intervals.
 - Alice and Clueless_i send at most one packet during each interval.
 - Eve can only count messages during each interval, and cannot distinguish between Alice's and Clueless_i's packets.
 - Alice attempts to signal Eve by transmitting to one of the receivers or by not transmitting at all.
 - Clueless_i is also transmitting, without regard to Alice.
-

Analysis of the Covert Channel

- What is the most information that Alice can send to Eve in this manner?
 - This covert channel is discrete, memoryless, with noise (Clueless_i's input randomly affects the output).
 - Simple Case: Alice is alone ($N = 0$)
 - Alice is only transmitter.
 - Alice sends S (**silent**) by not sending a message, and T (**transmit**) by sending a message.
 - Eve receives either $e_0=0$ (Alice silent) or $e_1=1$ (Alice transmits)
 - There is no noise on channel
 - What is the capacity of the channel?
-

Channel Capacity with $N=0$

- Reminders:

- **Channel Capacity** $C = \max_x I(X, Y)$
- $I(X, Y) = H(X) - H(X|Y)$
- $I(X, Y) = I(Y, X)$

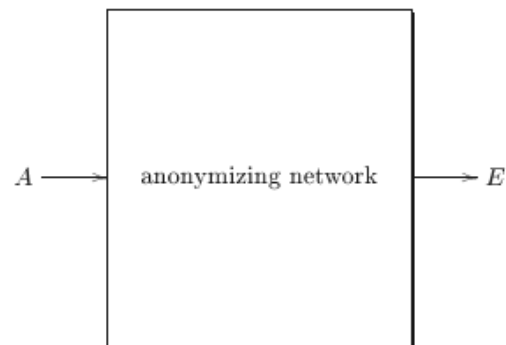
- Define:

- **E** = Distribution for observations by Eve
- **A** = Distribution for symbols sent by Alice
- $x = P(A = S)$

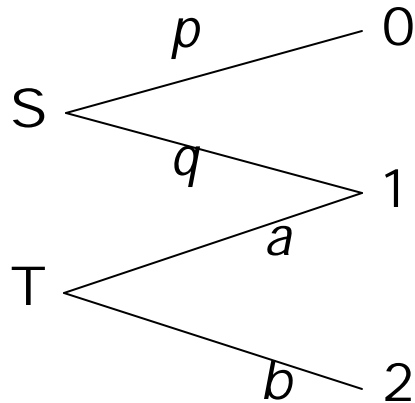
- No Noise:

$$\begin{aligned} I(E, A) &= H(E) - H(E|A) = H(E) - H(E|E) = H(E) \\ &= -x \log x - (1-x) \log(1-x) \end{aligned}$$

- This is maximized for $x=0.5 \rightarrow C = 1.0$



Channel Capacity for $N = 1$



Channel Transition Diagram

$$M_{2,1} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 \end{matrix} \\ \begin{matrix} S \\ T \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ p & q & 0 \\ 0 & a & b \end{pmatrix} \end{matrix}$$

Channel Matrix

$$M_{2,1}[i,j] = P(E=j | A=i)$$

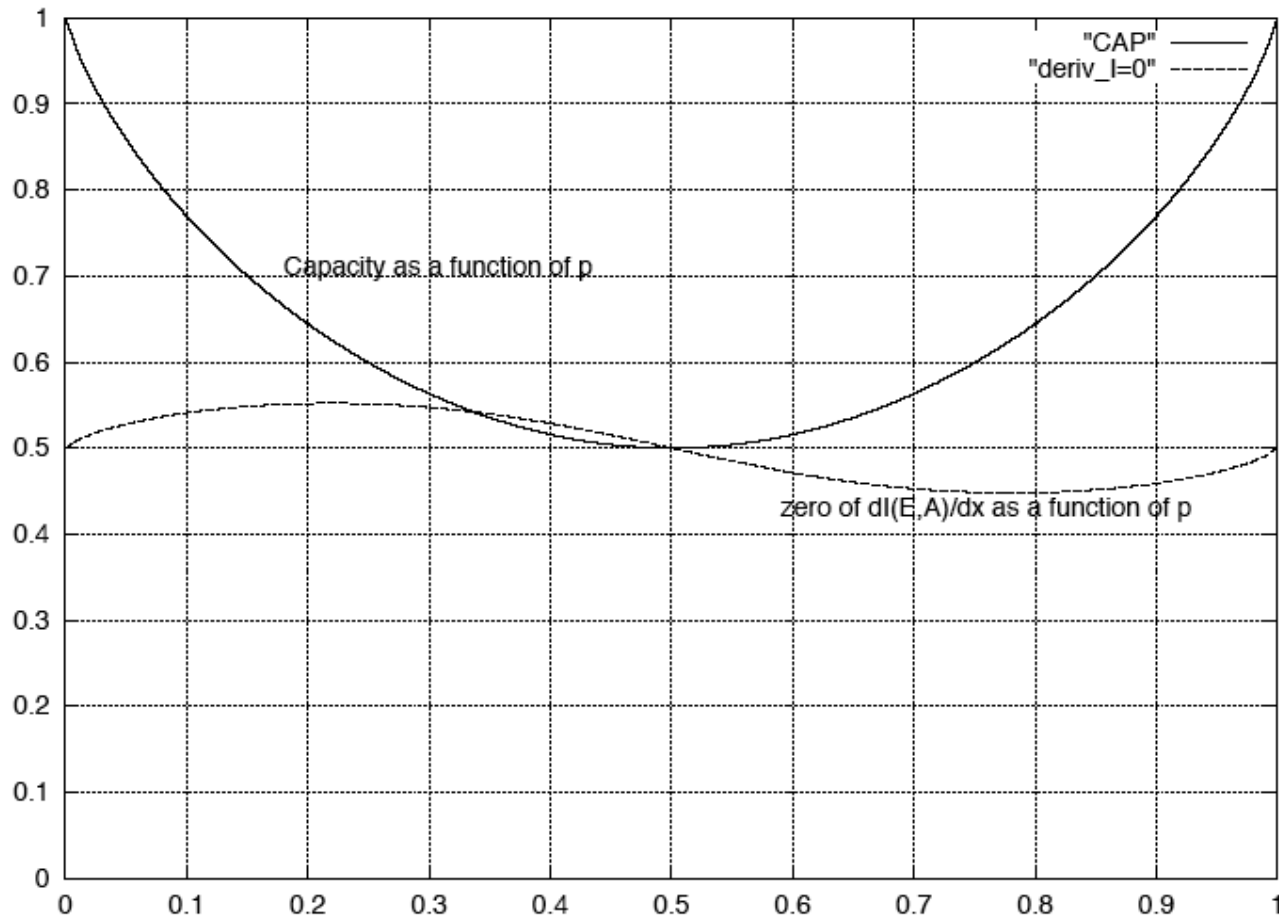
- It can be shown that $p = a$ and $q = b$.
- Simplified Channel Matrix:

$$M_{2,1} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 \end{matrix} \\ \begin{matrix} S \\ T \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ p & q & 0 \\ 0 & p & q \end{pmatrix} \end{matrix}$$

Channel Capacity for $N = 1$ (2)

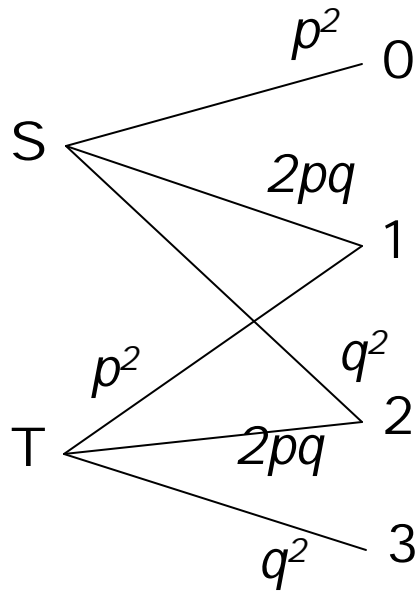
- $C = \max_x \{H(E) - H(E|A)\}$
 - $H(E|A)$ can be derived from channel matrix $M_{2,1}$
 - *(At this point we move to white board)*
-

Channel Capacity for $N = 1$ (3)



Covert Channel Capacity as Function of p , and x value that maximizes $I(E,A)$

Channel Capacity for $N = 2$

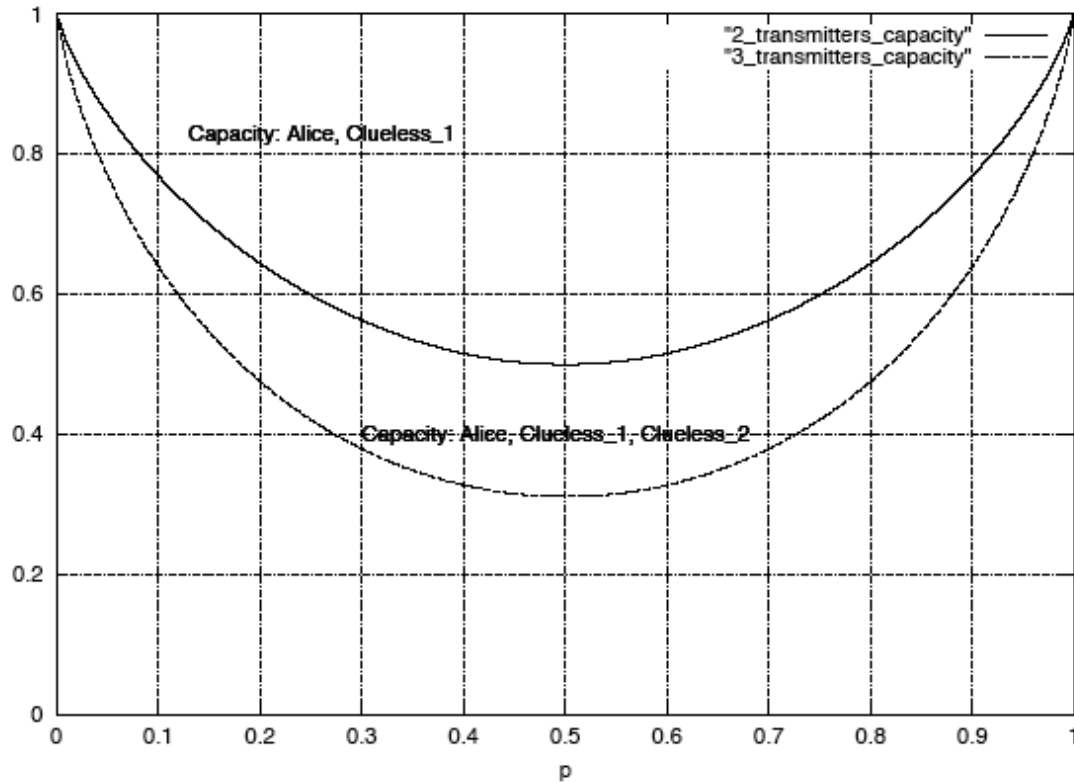


Channel Transition Diagram

$$M_{2,2} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 \end{matrix} \\ \begin{matrix} S \\ T \end{matrix} & \begin{pmatrix} p^2 & 2pq & q^2 & 0 \\ 0 & p^2 & 2pq & q^2 \end{pmatrix} \end{matrix}$$

Channel Matrix

Channel Capacity for $N=2$ (2)



Covert Channel Capacity for 2 and 3 Transmitters