Advanced Internet Systems
CSE 454
Daniel Weld

To do
- Add picture of original MT
- Add greg little or casting words
- Discussion included qualifications, contracts,

CSE 454 Overview

Transfer of Confidential Data

You (client)  Merchant (server)

I want to make a purchase.

Here is my RSA public key E and a cert and a "nonce".

My Credit Card and your nonce are

E("6543 2345 6789 8765", nonce).

Cryptography
- Symmetric + asymmetric ciphers
- Stream + block ciphers; 1-way hash
- Z=Ya mod N

DNS, HTTP, HTML
- Get, put, post
- Cookies, log file analysis

DNS
- Hierarchical namespace
- Susceptible to DOS attacks
- Recent news: Google Public DNS

HTTP
- Get, put, post
- Cookies, log file analysis

HTML
- Link extraction
Transfer of Confidential Data

You (client) Merchant (server)

I want to make a purchase.

What is your Credit Card Number?

My Credit Card is 6543 2345 6789 8765.

But the Internet provides no privacy.

Is there any way to protect my data from prying eyes at intermediate nodes?

Symmetric Encryption

- If the user has a pre-existing relationship with the merchant, they may have a shared secret key $K$ – known only to the two parties.

- User encrypts private data with key $K$.

- Merchant decrypts data with key $K$.

Asymmetric Encryption

- What if the user and merchant have no prior relationship?

- Asymmetric encryption allows me to encrypt a message for a recipient without knowledge of the recipient’s decryption key.

The Fundamental Equation

$E = mc^2$

$Z = Y^X \mod N$
The Fundamental Equation

\[ Z = Y^X \mod N \]

When \( Z \) is unknown, it can be efficiently computed.

When \( X \) is unknown, the problem is known as the **discrete logarithm** and is generally believed to be hard to solve.

When \( Y \) is unknown, the problem is known as **discrete root finding** and is generally believed to be hard to solve without the factorization of \( N \).

The problem is not well-studied for the case when \( N \) is unknown.

How to compute \( Y^X \mod N \)

Compute \( Y^X \) and then reduce \( \mod N \).

- If \( X \), \( Y \), and \( N \) each are 1,000-bit integers, \( Y^X \) consists of \( \sim 2^{1010} \) bits.
- Since there are roughly \( 2^{250} \) particles in the universe, storage is a problem.

How to compute \( Y^X \mod N \)

- Repeatedly multiplying by \( Y \) (followed each time by a reduction modulo \( N \)) \( X \) times solves the storage problem.

- However, we would need to perform \( \sim 2^{900} \) 32-bit multiplications per second to complete the computation before the sun burns out.
How to compute $Y^X \mod N$

Multiplication by Repeated Doubling

To compute $X \cdot Y$,
compute $Y, 2Y, 4Y, 8Y, 16Y, ...$
and sum up those values dictated by the binary representation of $X$.

Example: $26Y = 2Y + 8Y + 16Y$.

Exponentiation by Repeated Squaring

To compute $Y^X$,
compute $Y, Y^2, Y^4, Y^8, Y^{16}, ...$
and multiply those values dictated by the binary representation of $X$.

Example: $Y^{26} = Y^2 \cdot Y^8 \cdot Y^{16}$.

How to compute $Y^X \mod N$

We can now perform a 1,000-bit modular exponentiation using $\sim 1,500$ 1,000-bit modular multiplications.

- 1,000 squarings: $y, y^2, y^4, ..., y^{2^{1000}}$
- $\sim 500$ “ordinary” multiplications

The Fundamental Equation

$Z = Y^X \mod N$

When $Y$ is unknown, the problem is known as discrete root finding and is generally believed to be hard to solve ... without the factorization of $N$.

RSA Encryption/Decryption

- Select two large primes $p$ and $q$.
- Publish the product $N = pq$.
- The exponent $X$ is typically fixed at 65537.

- Encrypt message $Y$ as $E(Y) = Y^X \mod N$.
- Decrypt ciphertext $Z$ as $D(Z) = Z^{1/X} \mod N$.

- Note $D(E(Y)) = (Y^X)^{1/X} \mod N = Y$.

RSA Signatures and Verification

- Not only is $D(E(Y)) = (Y^X)^{1/X} \mod N = Y$,
  but also $E(D(Y)) = (Y^{1/X})^X \mod N = Y$.

- To form a signature of message $Y$,
  create $S = D(Y) = Y^{1/X} \mod N$.

- To verify the signature, check that
  $E(S) = S^X \mod N$ matches $Y$. 
Transfer of Confidential Data

You (client) Merchant (server)

I want to make a purchase.
What is your Credit Card Number?
My Credit Card is 6543 2345 6789 8765.

Slides by Josh Benaloh

Transfer of Confidential Data

You (client) Merchant (server)

I want to make a purchase.
Here is my RSA public key E.
My Credit Card is E(6543 2345 6789 8765).

Slides by Josh Benaloh

Intermediary Attack

You (client) Intermediary Merchant (server)

I want to make a purchase.
My public key is Ė.
É(CC#)

I want to make a purchase.
My public key is E.
E(CC#)

Slides by Josh Benaloh

Digital Certificates

“Alice’s public modulus is
Nₐ = 331490324840…”
-- signed …
someone you trust.

Slides by Josh Benaloh

Transfer of Confidential Data

You (client) Merchant (server)

I want to make a purchase.
Here is my RSA public key E and a cert.
My Credit Card is E(6543 2345 6789 8765).

Slides by Josh Benaloh

Replay Attack

You (client) Merchant (server)

I want to make a purchase.
Here is my RSA public key E and a cert.
My Credit Card is E(6543 2345 6789 8765).

Later …
Eavesdropper Merchant (server)

I want to make a different purchase.
Here is my RSA public key E and a cert.
My Credit Card is E(6543 2345 6789 8765).

Slides by Josh Benaloh
Transfer of Confidential Data

You (client)  Merchant (server)

I want to make a purchase. Here is my RSA public key $E$ and a cert and a “nonce”. My Credit Card and your nonce are $E(\text{“6543 2345 6789 8765”}, \text{nonce})$.

SSL/PCT/TLS History

- 1994: Secure Sockets Layer (SSL) V2.0
- 1995: Private Communication Technology (PCT) V1.0
- 1996: Secure Sockets Layer (SSL) V3.0
- 1997: Private Communication Technology (PCT) V4.0
- 1999: Transport Layer Security (TLS) V1.0

SSL/PCT/TLS

You (client)  Merchant (server)

Let’s talk securely. Here are the protocols and ciphers I understand. I choose this protocol and these ciphers. Here is my public key, a cert, a nonce, etc.

Using your public key, I’ve encrypted a random symmetric key (and your nonce).

SSL/TLS

All subsequent secure messages are sent using the symmetric key and a keyed hash for message authentication.

Symmetric Ciphers

Private-key (symmetric) ciphers are usually divided into two classes.

- Block ciphers
- Stream ciphers

Block Ciphers

DES, AES, RC2, RC5, etc.
Block Cipher Modes

Electronic Code Book (ECB) Encryption:

Electronic Code Book (ECB) Decryption:

Block Cipher Integrity

With ECB mode, identical blocks will have identical encryptions.

This can enable replay attacks as well as re-orderings of data. Even a passive observer may obtain statistical data.

Cipher Block Chaining (CBC) Encryption:

Cipher Block Chaining (CBC) Decryption:

Stream Ciphers

RC4, SEAL, etc.

- Use the key as a seed to a pseudo-random number-generator (PRNG).
- Take the stream of output bits from the PRNG and XOR it with the plaintext to form the ciphertext.
Stream Cipher Encryption

Plaintext: ****************************
PRNG(seed): ****************************
Ciphertext: ****************************

Stream Cipher Integrity

- It is easy for an adversary (even one who can't decrypt the ciphertext) to alter the plaintext in a known way.

Bob to Bob's Bank:
Please transfer $0,000,002.00 to the account of my good friend Alice.

One-Way Hash Functions

- MD4, MD5, SHA-1, SHA-256, etc.

A one-way hash function is a function

\[ H : \{0,1\}^* \rightarrow \{0,1\}^k \quad \text{(typically} \quad k \text{ is 128 or 160)} \]

such that, given an input value \( x \), one can't find \( x' \neq x \) such that \( H(x) = H(x') \).

One-Way Hash Functions

- Non-invertability: given \( y \), it's difficult to find any \( x \) such that \( H(x) = y \).

- Collision-intractability: one cannot find a pair of values \( x', \neq x \) such that \( H(x) = H(x') \).
One-Way Hash Functions

- When using a stream cipher, a hash of the message can be appended to ensure integrity. [Message Authentication Code]
- When forming a digital signature, the signature need only be applied to a hash of the message. [Message Digest]

Slides by Josh Benaloh

Structure of Mercator Spider

1. Remove URL from queue
2. Simulate network protocols & REP
3. Read w/ RewindInputStream (RIS)
4. Has document been seen before? (checksums and fingerprints)
5. Extract links
6. Download new URL?
7. Has URL been seen before?
8. Add URL to frontier

Common Types of Clusters

- Simple Web Farm
- Search Engine Cluster

Inktomi (2001) Supports programs (not users) Persistent data is partitioned across servers:

\[ \text{⇑ capacity, but ⇓ data loss if server fails} \]

From: Brewer Lessons from Giant-Scale Services

CSE 454 Overview

- Search Engines
  - HTTP, HTML, Scaling & Crawling
  - Cryptography & Security

The Precision / Recall Tradeoff

- Precision $\frac{tp}{tp + fp}$
  - Proportion of selected items that are correct
- Recall $\frac{tp}{tp + fn}$
  - Proportion of target items that were selected
- Precision-Recall curve
  - Shows tradeoff

The Precision-Recall Tradeoff Diagram

- Correct Tuples
- Tuples returned by System

The Precision-Recall Tradeoff

- Vector Space Representation
  - Dot Product as Similarity Metric
- TF-IDF for Computing Weights
  - $w_i = f(i,j) \times \log(N/n_i)$
  - Where $q = \cdots \text{word} \cdots$
  - $N = |\text{docs}|$ $n_i = |\text{docs with word} i|$

- How Process Efficiently?

Copyright © Wald 2002-2007
Thinking about Efficiency

- Clock cycle: 2 GHz
  - Typically completes 2 instructions / cycle
  - ~10 cycles / instruction, but pipelining & parallel execution
  - Thus: 4 billion instructions / sec
- Disk access: 1-10ms
  - Depends on seek distance, published average is 5ms
  - Thus perform 200 seeks / sec
  - (And we are ignoring rotation and transfer times)
- Disk is 20 Million times slower !!!

Inverted Files for Multiple Documents

```
LEXICON

WORD  DOCS  OCCUR  POS 1  POS 2  ...

jezebel 20
jezer 3
jezerit 1
jeziah 1
jeziel 1
jeziah 1
jezoar 1
jezrahiah 1
jezreel 39
jezoar
...

OCCURRENCE INDEX

DOCID  OCCUR  POS 1  POS 2  ...

566 3 203 245 287
...
```

- One method. Alta Vista uses alternative

How Inverted Files are Created

Crawler → Repository → Scan → Forward Index → Sort → Inverted File List → NF Scan → Sort → Sorted Index

AltaVista

- Basic Framework
  - Flat 64-bit address space
  - Index Stream Readers: Loc, Next, Seek, Prev
  - Constraints
    - Let E be ISR for word enddoc
    - Constraints for conjunction a AND b
      - prev(E) ≤ loc(A)
      - loc(A) ≤ loc(E)
      - prev(E) ≤ loc(B)
      - loc(B) ≤ loc(E)

Beyond Size - Search User Interfaces

- Specialized Search
- Suggestions
- Spelling Correction

50-80% Page Views = Revisits

<table>
<thead>
<tr>
<th>Cluster Group</th>
<th>Name</th>
<th>Shape</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Sources</td>
<td>F1</td>
<td></td>
<td>Pen &amp; paper; small &amp; local</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td></td>
<td>Pen &amp; paper, local &amp; regional</td>
</tr>
<tr>
<td></td>
<td>F3</td>
<td></td>
<td>Small &amp; local &amp; regional</td>
</tr>
<tr>
<td></td>
<td>F4</td>
<td></td>
<td>Small &amp; local &amp; regional</td>
</tr>
<tr>
<td>Medium Sources</td>
<td>M1</td>
<td></td>
<td>Popular sources, local</td>
</tr>
<tr>
<td></td>
<td>M2</td>
<td></td>
<td>Popular sources, local</td>
</tr>
<tr>
<td>Short Stories</td>
<td>S1</td>
<td></td>
<td>Easy stories, local &amp; regional</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td></td>
<td>Easy stories, local &amp; regional</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td></td>
<td>Easy stories, local &amp; regional</td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td></td>
<td>Easy stories, local &amp; regional</td>
</tr>
<tr>
<td>Hybrid Sources</td>
<td>H1</td>
<td></td>
<td>Popular text &amp; local &amp; regional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Popular text &amp; local &amp; regional</td>
</tr>
</tbody>
</table>
Revisitation

Resonance

CSE 454 Overview

A-B testing

Nine Changes in Site Above

Landscape of IE Techniques: Models

Any of these models can be used to capture words, formatting or both.

Slides from Cohen & McCallum
IE with Hidden Markov Models

Given a sequence of observations:
Yesterday Pedro Domingos spoke this example sentence.
and a trained HMM:

Find the most likely state sequence: (Viterbi) \( \arg \max_i P_i \alpha_{t,i} \)

Any words said to be generated by the designated “person name” state extract as a person name:

Person name: Pedro Domingos

The Forward Algorithm

\( \alpha(T) \) is the unknown state sequence

Finding the most likely state sequence is the Viterbi algorithm

\[ \alpha_{t,i} = \max_{j} \left\{ \alpha_{t-1,j} \cdot P_{trans} \cdot P_{obs}(y_t) \right\} \]

Time: \( O(K^2T) \)
Space: \( O(KT) \)

Linear in length of sequence

Terminating Viterbi

How did we compute \( \delta_i \)?

Max \( \delta_{T-1}(i) \cdot P_{trans} \cdot P_{obs} \)

Now backchain to find final sequence

What is Open Information Extraction?

<table>
<thead>
<tr>
<th>Traditional IE</th>
<th>Open IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Corpus - Domain-Independent Methods</td>
</tr>
<tr>
<td>Relations</td>
<td>Specified In Advance</td>
</tr>
<tr>
<td>Complexity</td>
<td>( O(D \cdot R) )</td>
</tr>
<tr>
<td>Documents, Relations</td>
<td>( D ) documents, ( R ) relations</td>
</tr>
</tbody>
</table>

Self-Supervised Learning from Wikipedia

Ben is living in Paris.

CSE 454 Overview

<table>
<thead>
<tr>
<th>Human Comp</th>
<th>Cool Ults (Zoetrope &amp; Revisiting)</th>
<th>Adverts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open IE</td>
<td>Parsing &amp; POS Tags</td>
<td>Search Engines</td>
</tr>
<tr>
<td>Information Extraction</td>
<td>Web Tables</td>
<td>Cryptography &amp; Security</td>
</tr>
<tr>
<td>Supervised Learning</td>
<td>HTTP, HTML, Scaling &amp; Crawling</td>
<td></td>
</tr>
</tbody>
</table>
Self-Supervised Learning from Wikipedia

[Wu et al. CIKM'07]

How Motivate People to Help?

- Money
- Fun
- Altruism
- Esteem
- Self-Interest

Altruism

Self-Esteem

How Motivate People to Help?

- Pay them...
The term silver dollar is often used for any large white metal coin issued by the United States with a face value of one dollar; although purists insist that a dollar is not silver unless it contains some of that metal.

In our experiment we ask for 10 annotations each of the full 30 word pairs, at an offered price of $0.02 for each set of 30 annotations (or, equivalently, at the rate of 1500 annotations per USD). The most surprising aspect of this study was the speed with which it was completed; the task of 300 annotations was completed by 10 annotators in less than 11 minutes … 1724 annotations / hour.

Who are those Turkers?
Complex Jobs

- Casting Words
- TurkIt

Iterative Improvement

A partial view of a pocket calculator together with some coins and a pen.

Iterative Improvement

A close-up photograph of the following items:

- A Casio multi-function, solar-powered scientific calculator.
- A blue ballpoint pen with a blue rubber grip and the tip extended.
- Six British coins: two of £1 value, three of 20p value and one of 1p value.
- Seems to be a theme illustration for a brochure or document cover treating finance - probably personal finance.

Motivating People

- Money
- Fun
ACCESSIBILITY
LESS THAN 10% OF THE WEB IS ACCESSIBLE TO THE VISUALLY IMPAIRED
REASON: MOST IMAGES DON'T HAVE A CAPTION

LABELING IMAGES WITH WORDS
FACE
MAN
SUPER SEXY
STILL A COMPLETELY OPEN PROBLEM

DESIDERATA
A METHOD THAT CAN LABEL ALL IMAGES ON THE WEB FAST AND CHEAP

THE ESP GAME
TWO-PLAYER ONLINE GAME
PARTNERS DON'T KNOW EACH OTHER AND CAN'T COMMUNICATE
OBJECT OF THE GAME: TYPE THE SAME WORD
THE ONLY THING IN COMMON IS AN IMAGE

THE ESP GAME
PLAYER 1
GUESSING: CAR
GUESSING: HAT
GUESSING: KID
SUCCESS!
YOU AGREE ON CAR

PLAYER 2
GUESSING: BOY
GUESSING: CAR
SUCCESS!
YOU AGREE ON CAR

© 2004 Carnegie Mellon University, all rights reserved. Patent Pending.
THE ESP GAME IS FUN
3.2 MILLION LABELS WITH 22,000 PLAYERS
MANY PEOPLE PLAY OVER 20 HOURS A WEEK

LABELING THE ENTIRE WEB
5000 PEOPLE PLAYING SIMULTANEOUSLY CAN LABEL ALL IMAGES ON GOOGLE IN 30 DAYS!
INDIVIDUAL GAMES IN YAHOO! AND MSN AVERAGE OVER 10,000 PLAYERS AT A TIME

9 BILLION MAN-HOURS OF SOLITAIRE WERE PLAYED IN 2003
EMPIRE STATE BUILDING
7 MILLION MAN-HOURS
(6.8 HOURS OF SOLITAIRE)
PANAMA CANAL
20 MILLION MAN-HOURS
(LESS THAN A DAY OF SOLITAIRE)

GWAP
- Problem?

Motivating Vision
Next-Generation Search = Information Extraction + Ontology + Inference

Next-Generation Search
- Information Extraction
  - <Einstein, Born-In, Germany>
  - <Einstein, ISA, Physicist>
  - <Einstein, Lectured-At, IAS>
  - <IAS, In, New-Jersey>
  - <New-Jersey, In, United-States>
- Ontology
  - Physicist(x) \rightarrow Scientist(x)
- Inference
  - Einstein = Einstein