Big Data Security Practices

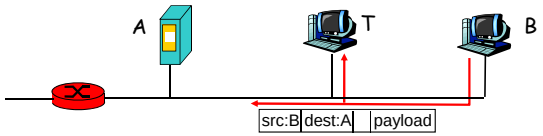
Kai Shen

Recap of Data Security Principles

- Cryptography for data confidentiality
  - Symmetric key cryptography (DES, AES, Stream Cypher)
  - Public key cryptography (RSA), computationally more expensive, not directly applied on big data
- Data integrity through digital signature
  - Signed using private key, verified using public key
  - Only applied on a data digest to save overhead
- Security costs for big data
  - Encrypting/decrypting 1TB data in 128-bit AES takes 7 hours
  - Hashing 1TB data in SHA-1 take 45 minutes
  - Public-key-crypto (RSA) would be much slower, but never directly applied on large datasets

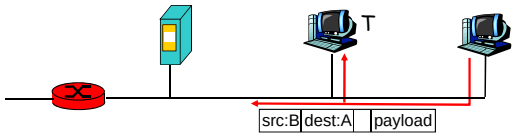
Security Threat: Network Sniffing

- Collect sensitive data while it is in transit over network even though you are not the intended receiver
- For instance, promiscuous network adapters can read all packets passing by a broadcast media (e.g. shared-link Ethernet)



- Can learn data communication pattern and read all unencrypted data (e.g. passwords)
- Tools (tcpdump, Wireshark) to parse network protocols and extract data

Security Threat: Network Sniffing



- Solutions:
  - Encrypt all data packets (what encryption?)
  - One host per segment in switched Ethernet, not really a broadcast media

WiFi Data Sniffing

- Wireless communications such as WiFi are true broadcast media
- Earlier WiFi encryptions are fast but have vulnerabilities
  - Reused key attack for stream cypher
  - Nonrandomness in RC4 ("A Practical Attack on Broadcast RC4" by Mantin and Shamir):
    - higher-than-normal chance for the second word to be zero
    - second word of the original data has a higher chance to be unchanged in the cypher text

    ⇨ easier to know if a code is from RC4, easier to break the code
- Both WEP and WPA use the RC4 stream cypher (WEP is also vulnerable to reused key attack). WPA2 instead uses …?

Data Security over the Internet

- SSL (Secure Sockets Layer) / TLS (Transport Layer Security): security service to any TCP-based applications
  - used for remote terminal access (SSH).
  - used for data copy (SCP).
  - used between Web browsers, servers for e-commerce (https).
  - used between IMAP clients and servers.
- Security services:
  - data confidentiality by encryption using a symmetric session key (AES), key encrypted with server's public key (RSA).
  - source authentication & data integrity by signed message digests.

Data Security over the Internet

- IPsec: Data security for all Internet data traffic
- Like SSL/TLS:
  - data confidentiality by encryption using a symmetric session key
  - source authentication & data integrity by signed message digests
- Done in a way that is compatible with basic IP routing functions
  - Some (manageable) changes at the boundary between private and public networks
- (Virtual Private Network) VPN:

Data Sniffing in Peer-to-Peer Networks

- Data sniffing when you "are" an intended receiver, but data isn't used in the intended way
- Peer-to-peer networks (Gnutella, BitTorrent, …)
  - share data (files, music, movies, …)
  - nodes exchange/pass search queries in the network
- You can pose as a legitimate node in a peer-to-peer network
  - collect and analyze queries
  - learn emerging trends, interests etc.
  - or find out who are stealing my copyrighted materials (RIAA, MPAA)
  - no encryption to worry about!

Data Security on Resource Constrained Environment

- Field device storage can be easily compromised
- Encryption expensive on resource-constrained devices
  - On a low-power Atom processor, AES runs an order of magnitude slower than on an Intel Xeon
- Low-cost (low-power) security in the field?

Work by Zhuan Chen

Privacy in Big Data Processing

- Web companies (Google, Facebook, Twitter, Microsoft, Amazon, your cable modem provider, …) possess large amount of user data, they process the data for a living
  - Big data processing for aggregate statistics
  - But also processing that targets individuals (targeted ads)
- Implicit contract:
  - Users understand that their privacy is compromised, but such privacy compromise is taken with care by the companies (that should not lead to harms to the users)
  - Reputation is everything

Data Anonymization

- Privacy preserving rule: Restricted accesses to data with identity (like what?)

[Sun Jan 6 00:00:00 2002] [165.14.6.134] 'Uncut Dragon Ball Z Pictures', 0.4528sec, 2750bytes
[Sun Jan 6 00:00:00 2002] [165.14.6.130] 'website for Brighton middle school', 1.9862sec, 9700bytes

- A segment of real trace that I got from … for research

```
… …
907248395    11f7de97af05332e    14796  0.130  200   /s?q=1dc6cef5d2cc7f12&st=10&n=10
907248395    a090c07d9757b3df    14780  0.274  200   /s?q=b7ebccde01659d6b&st=0&n=10
907248395    19e41ae549307787    11800  0.280  200   /s?q=19270c765681a379&st=40&n=10
907248396    b95d3f46b3e6cf34    14420  0.124  200   /s?q=8566df34c2b43765&st=20&n=10
907248396    c3b26d79f99b1def    14143  2.317  200   /s?q=e7b51a03c73df83a&st=0&n=10
907248396    4e5b8f3546583cd2    10495  0.431  200   /s?q=002b98719ccac147&st=0&n=10
… …
```


Your Government Likes Big Data Too

- Big data analysis helps national security. OK, but how does the government collect data?
  - Normal data crawling and sniffing (things you can do in a class project)
  - Coerce web companies to supply data
  - Buy data from them

http://www.nytimes.com/2013/11/07/us/cia-is-said-to-pay-att-for-call-data.html

  - Steal from them!

http://www.nytimes.com/2013/11/01/technology/angry-over-us-surveillance-tech-giants-bolster-defenses.html

Your Footprint in Big Data Era

- Phone call records
- Web searches
- Social network activities, tweets
- Accesses to web sites like online retailers
- All your web access history at your local ISP or your company's web proxy server (our department, I believe, doesn't maintain one)
- WiFi signals with weak encryption
- Your EZPass toll records
- Your smart phone is recording a ton of data (have you checked the privacy setting of your phone?)
- … …

Protect Yourself in Big Data Era

- Be knowledgeable and support the right use of big data
- Intelligent use of data encryption
  - Encryption protects data, but not the identity of a communication point
- Tor, The Onion Router, anonymity network
  http://en.wikipedia.org/wiki/Tor_(anonymity_network)
  - Tor is limited to protect the identity of a communication point, encryption is still needed to protect data