GUEST LECTURE
CS475
Computer and Network Security

AYLIN CALISKAN-ISLAM
PHD CANDIDATE AT PSAL-DEXEL UNIVERSITY
ADVISOR: DR. RACHEL GREENSTADT
MARCH 6, 2014

CREDITS TO
DR. RACHEL GREENSTADT
DR. ALESSANDRO ACQUISTI
DR. DAWN SONG
AND OTHERS
Security Review of Digital Privacy

FIRST PART OF LECTURE
What is privacy?

- Hard to define
  - Data concealment
  - A right “to be left alone”
  - Freedom
  - The ability to control the information released about you
Privacy in a Digital World

- How does it change?
“It is surprising how recently changes in law and technology have been permitted to undermine sanctuaries of privacy that Americans have long taken for granted.”

“Suddenly, shopping that had once seemed anonymous was being archived in personally identifiable dossiers”

NYT Magazine, April 30, 2000

“In the background, advertising services are building profiles of where people browse, what they buy, how they think, and who they are.”

“For about 9 cents, some medical data sites will sell you your neighbor's history of urinary tract infections. It will get worse.”

BusinessWeek, March 20, 2000

“Most commercial healthcare Web sites lure consumers with free medical information, then sell data on them to third parties in ways that threaten the consumer's privacy.”

Mark Smith, February 1, 2000
And then...

- “The overall B2C market opportunity should reach $450BN in transaction volume by 2004.”
- Actually...

Sources: Forrester Research and IDC, circa 2001
How the market reacted

Economic challenges pushed merchants to more restrictive policies.

*This policy may change from time to time so please check back periodically.*

- Yahoo privacy policy circa 2001
Stakeholders

- Individuals
- Businesses
- Governments
- Other groups
Digital Surveillance

- Who is the adversary?
Who is the data revealed to?

- Some faceless company?
- The government?
- The Internet?
- Friends/Family?
- Acquaintances/colleagues/employers?
Threats or what could possibly go wrong?

- Identity Fraud/Theft
  - Information actually used for harm
- Discrimination - social or economic
- Conformity pressure
Privacy vs. Security

- When is there a tradeoff?
- When are they the same?
“You pay for content or services with anonymous electronic cash. You connect to content and service providers with an anonymizing mixnet. You authenticate yourself with anonymous credential schemes or zero-knowledge identification protocols. You download content via private information retrieval or oblivious transfer. You use secure function evaluation when interacting with services that require some information.”
- [Feigenbaum, Sander, Freedman, Shostack]
A protocol for customer Alice, merchant Bob and the bank

- Alice withdraws $1 digital bill from the bank. This can be a message signed by bank’s RSA private key saying: serialNoXXX - $1
- Alice gives the cash to Bob.
- Bob deposits the cash to the bank.

2 problems:
- Not anonymous: the bank keeps track of Alice’s transactions.
- Double spending: Alice can duplicate the cash and give it to merchant Carol.
One of the key building blocks to achieve anonymity is a blind signature. Recall that RSA signatures require the signer to compute \( m^d \mod N \), where \((d, N)\) is the private key and \((e, N)\) is the public key.

A blind RSA signature is carried out as follows:

- Alice sends Bob \( s = (r^e m) \mod N \) where \( r \) is a random number \( \mod N \).
- Bob computes \( t = s^d \mod N \) and sends the result to Alice.
- Alice computes \( t/r \mod N = m^d \mod N \).

The point is that \( t = s^d = (r^e m)^d = rm^d \mod N \). A blind signature allows Alice to obtain Bob’s signature on a message of her choice, without Bob having any idea what the message being signed is.
Alice and bank create an anonymous $1 bill signed by the bank:

In this scheme, a valid $1 bill is a pair \((x, y)\) where \(y = f(x)^d \mod N\). Here \(f()\) is a one-way function (ideally a one-way hash function). \((e, N)\) is the bank’s public key and \((d, N)\) is the private key.

- To withdraw the $1 bill, Alice picks \(x\), computes \(f(x)\) and runs the blind signature protocol with the bank on \(f(x)\). i.e. Alice picks a random \(r \mod N\), sends the bank \(s = r^e f(x) \mod N\). The bank sends back \(t = s^d \mod N\), and Alice recovers \(y = f(x)^d = t/r \mod N\).

- To pay Bob $1, Alice sends him \((x, y)\).

- When Bob later deposits \((x, y)\), the bank checks that \(y^e = f(x)\), and that \((x, y)\) is not on its list of previously deposited bills.
What is needed for ecash?
Some examples:

1) Unlinkability
2) Unforgeability
3) Double spending problem
4) Divisibility
5) Anonimity revocation
Anonymous credentials (insurance cards, student IDs, etc) - can use digital signatures for this too
  
  - Brands generalized with certificate scheme

What if, instead of providing a SSN or ID number, you provided a zero-knowledge proof that you know the private key related to some public key that identifies you?

Mix-nets - Batch and mix messages to provide anonymity (high latency)
Technology Rundown

- Private Information Retrieval/Oblivious Transfer: Bob has database of n elements, Alice pays to access 1 item and should not get more, Bob should not know which item Alice accessed.

- Secure Function Evaluation - Alice and Bob want to compute some function, but keep the inputs private (classically, which one is richer?)

- Both of these can be done, but not always efficiently - take crypto class to learn more.
What are the obstacles?

- To these identity management technologies?
Types of Privacy Enhancements: Anonymity

- Anonymity (unlinkability) - Data is not linked to an identity
  - Location anonymity (Tor, mixes)
  - Data anonymity - “we anonymized the data before releasing it”
    - Netflix/census data/etc
    - k-Anonymity
Types of Privacy Enhancements:
Policy

- Policies that protect information
  - Internal access control measures
    - Data tagged with XACML or EPAL
    - Agreements with partners
    - Internal Auditing (Google example)
    - Regulatory Compliance (HIPAA?)
# Privacy vs P3P Policy

<table>
<thead>
<tr>
<th>Privacy policy</th>
<th>P3P policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designed to be read by a human</td>
<td>Designed to be read by a computer</td>
</tr>
<tr>
<td>Can contain fuzzy language with “wiggle room”</td>
<td>Mostly multiple choice - sites must place themselves in one “bucket” or another</td>
</tr>
<tr>
<td>Can include as much or as little information as a site wants</td>
<td>Must include disclosures in every required area</td>
</tr>
<tr>
<td>Easy to provide detailed explanations</td>
<td>Limited ability to provide detailed explanations</td>
</tr>
<tr>
<td>Sometimes difficult for users to determine boundaries of what it applies to and when it might change</td>
<td>Precisely scoped</td>
</tr>
<tr>
<td>Web site controls presentation</td>
<td>User agent controls presentation</td>
</tr>
</tbody>
</table>
The Privacy Paradox

Why do we have great privacy enhancing technologies... that almost nobody uses?

Why do so many people claim to be concerned about privacy... and then do little to protect it?
Privacy and Economics

- Will anyone buy privacy?
  - Maybe...we buy curtains/blinds
Difficulties in privacy economics

- Asymmetric information
  - Individual does not know how, how often, for how long information will be used
  - Intrusions invisible and ubiquitous
  - Externalities and moral hazard

- Ex-post
  - Value uncertainty
  - Keeps on affecting individual after transaction
Difficulties in privacy economics

- Context-dependent (states of the world)
  - Anonymity sets (how many people could I be confused with?)
  - Sweeney (2002) 87% Americans uniquely identified by gender, birth year, and zip code from the Personal Genome Project.
- The more parties that use the good (personal information) the higher risks for original data owner
- Different individuals value the same piece of information differently
  - Market for personal information is not necessarily the same as a market for privacy
Privacy trade-offs

- **Protect**
  - Immediate cost (or loss of immediate benefit)
  - Future (uncertain) benefits

- **Do not protect**
  - Immediate benefits
  - Future (uncertain) costs
Why is this Problematic?

- Incomplete information
- Bounded rationality/Behavioral distortions
  - Complacency towards large risks
  - Inability to handle prolonged accumulation of small risks
  - Coherent arbitrariness
  - Hyperbolic discounting
5 MINUTE BREAK
The Underground: Miscreant Activity in the Internet

SECOND PART OF LECTURE
Introduction

- The second part of the lecture focuses on underground miscreant activity in the Internet.
- Different components of underground miscreant activities are covered.
- The contents are linked to the references at the end.
- The state of underground economy research is evaluated.
- Open research problems in the field are identified.
Definition of Cyber-crime

European Commission’s proposed definition:

1) If traditional crime such as fraud and forgery are committed over information systems and electronic communications networks, it is considered cyber-crime.

2) Publishing illegal content such as child sexual abuse or racial hatred material over electronic media falls under the category of cybercrime.

3) If the crime is unique to electronic networks such as attacks against information systems, denial of service or hacking, it is considered cyber-crime.
“Measuring the cost of cybercrime” [4]

- In 2009, AT&T’s Ed Amoroso testified before the US Congress that global cyber-crime profits topped $1 trillion, which was 1.6% of the world GDP that year.

- Anderson et al. show that existing cyber-crime loss estimates are very large, there are flaws in existing reports and it is unclear what is being measured.

- Anderson et al. came up with a framework differentiating cyber-crime from other crime and for analyzing the costs.
Types of Cyber-crime

1) ‘Genuine’ Cyber-crime
   Online banking fraud, fake anti-virus, scams, infringing pharmaceuticals, replica goods, copyright infringement

2) ‘Transitional’ Cyber-crime
   Payment card fraud, identity theft, phone fraud

3) Traditional Crime Becoming ‘Cyber’
   Welfare fraud, tax fraud
Statistical Analysis of Underground Economy

- Defence costs
- Indirect losses
- Criminal revenue
- Direct losses
- Cybercrimes
- Supporting infrastructure

Cost to society
Cost of cyber-crime per citizen:
- Genuine cyber-crime cost a few tens of pounds/euros/dollars a year but most of them are indirect and defense costs.
- Transitional cyber-crime cost a few tens of pounds/euros/dollars a year.
- Traditional frauds becoming cyber cost a few hundred pounds/euros/dollars a year.

Each category earns a criminal a few tens of pence/cents per citizen. However, indirect and defense costs are roughly ten times the sum of revenue.

More investment in law enforcement can be especially valuable if it can reduce indirect costs and defense expenditures.
Actors in Underground Economy

1) Virus Writers
2) Website Masters/Crackers
3) Envelopes Stealers
4) Virtual Asset Stealers/Sellers
5) Affiliate Markets
6) Affiliate Programs
7) Suppliers
8) Players/Customers
9) Vulnerable Internet Users
“Studying malicious websites and the underground economy on the Chinese Web” [6]

- Zhuge et al. model individual actors in the Chinese underground black market and describe the interaction between Internet miscreants.
- They use a honeypot to automatically examine websites.
- China presents a unique ecosystem but their model can be extended to describe underground actors in the world.
- They show that 1.49% of the Chinese web contains malicious content.
Malicious Actors on the Chinese Web

[Diagram showing the process of web-based malware, from virus writers to victims.]
McCoy et al. present an empirical analysis of affiliate-affiliate program business model using “ground-truth” data sets including four years of raw transaction logs covering $185 million in sales.

Spamming attracts thousands of new customers for each affiliate program every week.

The market is not saturated with spam.

10% of the affiliates account for 80% of the total affiliate program revenue. Disrupting these particular affiliates will have disproportionate damage on the whole program.
“Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs” [5]

- 75% of orders and 80% of revenue in the SpamIT/GlavMed affiliate program come from ED pill sales.
- Consumers can purchase pills for 5% of the US price.
“Why do Nigerian scammers say they are from Nigeria?” [9]

- Herley analyzes profitable strategies for scams.
- Why are they being so unconvincing by saying that they are from Nigeria but not NJ or Philadelphia?
- The trade-off between the true-positives and false-positives is graphed as a ROC curve.
- Opportunity drops much faster than viable victim density.
- False positives may be used to intentionally erode attacker economics.
- At low densities, certain attacks pose no threat.
"Why do Nigerian scammers say they are from Nigeria?" [9]

\[ E\{R\} = (d \cdot t_p \cdot G - (1 - d) \cdot f_p \cdot C) \cdot N \]

- \( d \) = victim density
- \( t_p \) = true-positive rate
- \( f_p \) = false-positive rate
- \( G \) = net gain from success
- \( C \) = cost of failure
- \( N \) = population

Aylin Caliskan-Islam, Drexel University

3/6/2014
Sheng et al. present the results of a role-play survey instrument administered to 1001 Amazon Mechanical Turk respondents to study both the relationship between demographics and phishing susceptibility and the effectiveness of several anti-phishing educational materials.

Statistical analysis shows that women are more susceptible to phishing than men.

People between the ages 18-25 are more susceptible to phishing than other age groups.

Going over educational materials reduced users’ tendency to enter information into phishing webpages by 40% and also clicking some legitimate links.
Cyber-criminals communicate through the Internet to establish business relationships. They establish trust relationships, trade and process fraudulent transactions online. They either use IRC channels or Web forums. Out of 2,677 IRC channels, 4.7% of the chat rooms were identified as underground IRC channels. The results were obtained through a support vector machine classifying channel content as benign or underground activity.
“An analysis of underground forums” [1]

- Cyber-criminals maintain profiles and build social relationships similar to Facebook but with differing motivation.
- Motoyama et al. explain how users interact, how baseline reputation is established and how it changes over time by characterizing six forums.
- Forum users start as ‘newbies’ and are elevated to higher groups as they actively participate in the forums.
Using the private message (PM) and thread relationships in the forums, they performed social degree analysis on users.

Private messaging does not have much effect on reputation.

Group status influences how many private messages the users receive in the week after posting trading threads.

35 regular expressions were used to bin 13 common categories for banning reasons.

Rippers and multiple account creators get banned.
Infrastructure

- Underground infrastructure is required to support cyber-criminal activities.
- Affiliates handle advertising by spamming and search engine optimization.
- Spam has become a multi-million dollar business.
- Spam is generally distributed by a network of infected machines that form a botnet.
- Many spambots are infected through pay-per-install.
Botnets

- A set of compromised hosts controlled by a single authority is called a botnet.
- Once an attacker controls a botnet, the attacker can steal sensitive information from the compromised hosts, orchestrate DDoS attacks, send spam, poison the search engines and commit click fraud.
Gross et al. [19] present analysis of a large-scale botnet from the botmasters perspective, that highlights the intricacies involved in spam campaigns such as the quality of email address lists, the effectiveness of IP-based blacklisting, and the reliability of bots.

They obtained 2.35TB of data from 16 command-and-control servers used by the Pushdo/Cutwail botnet.

The Cutwail botnet had 121,336 unique online IPs per day and a total of 2,536,934 unique IPs.

38% of the bots were in India, 9% in Australia, 4% in Russia, 3% in Brazil and 3% in Turkey.
"The underground economy of spam: A botmaster’s perspective of coordinating large-scale spam campaigns" [7]
Spammer bots were only able to deliver 30.3% of the spam volume they sent out because of errors.

In one hour 12.8%, in two hours 29.6%, in three hours 46.4%, in six hours 75.3% and in eighteen hours 90% of the bots were blacklisted.

Reliability and quality of bots did not depend on geographic location.

One million e-mail addresses cost $25-$50.

Spam-as-a-service can be purchased for $100-$500 per million e-mails sent.
Gross et al. leveraged information about Torpig’s domain generation algorithm and C&C protocol to register domains that the infected bots would contact.

They took over the Torpig network for ten days and obtained information from 182,800 infected machines.

12 different build values for Torpig.

Torpig is offered as a “malware service” to third parties.

Torpig botmasters made $83K-$8.3M in ten days.

Passwords of botnet victims are weak.

28% of the victims reused passwords.
Levchenko et al. try to understand the end-to-end resource dependencies in the spam enterprise.

Spam URLs are crawled and pages are clustered according to content.

Illegal pharmaceuticals, replica goods and counterfeit software are three popular spam advertised products.

Targeted purchases are made from affiliate programs.

Replica goods ship from China, drugs from India, herbal supplies from New Zealand and China.

There are a handful of merchant banks authorizing or settling transactions for spam-advertised purchases.
“Measuring pay-per-install: The commoditization of malware distribution” [3]

- Caballero et al. perform a measurement study of the PPI market by infiltrating four PPI services.
- They develop ‘milkers’ to automatically download malware from PPI C&C servers. Then, they download, execute, and classify malware.
- In six months, they collected 1,065,895 client binaries of which 9,153 were distinct.
- 12 of the world’s top 20 malware families use PPI services for distribution.
- On average, samples were repacked every 11 days.
- PPI clients are given a geographic location choice.
Synthesis

- Cyber-crime has a negative effect on businesses, computers and every day lives of Internet users.
- Internet miscreants pose a great security threat and have a significant effect on the world’s economy.
- Understanding all the aspects of underground economy is crucial for dealing with cyber-crime.
- Underground economy has many evolving components, thus there are many open problems for security researchers.
Open problems in the area

- We looked at the ‘big picture of the underground’: The definition and types of cyber-crime, Measurement of underground economy, Actors in carrying out miscreant activities, Types of underground marketplaces, Infrastructure required for miscreant activity.

- Still, some research questions need a better answer or remain unanswered in this field.
Definition of Cyber-crime

- The definition of cyber-crime is still vague.
- There is no common understanding for types of cyber-crime.
- There is no clear distinction between types of cyber-crime.
- Defense costs for cyber-crime is too high.
- Costs of different types of cyber-crime cannot be calculated accurately.
- A systematic approach is needed to estimate losses.
Spamming Strategies

- Spam is adapting to new media and most research focuses on email spam. Different types of spam need to be analyzed.

- Finding out the optimum strategy for spamming might help authorities counteract the approach.

- Botnets are key contributors for spreading spam. There is no method to stop the operations of botnets in the long term, in a feasible way.
Internet Miscreants

- An automated analysis of underground market users in the IRC channels or Web forums has not been performed yet.

- Using machine learning to identify types of cyber-criminals and bad actors such as the banned users or multiple account holders will aid in analyzing the connection between individual miscreants and provide a better picture of the monetization process.

- Weak points might be discovered and attacked to interrupt monetization.
Cyber-crime Prevention

- Blacklisting and user education do not provide enough protection from Internet miscreants.
- Sophisticated malware cannot be detected by anti-virus programs or experienced users.
- Virus detection needs to improve.
- International laws and regulations are required to fight cyber-crime.
- Security researchers need to collaborate more with authorities to come up with better defense tactics.
My work on this field

- **Stylometric analysis of underground forums:**
  - We obtained publicly leaked forum sql dumps
  - AntiChat, BadHacke, BlackhatWorld, Carders and L33tCrew

- **We wanted to detect:**
  - Product information
  - Users and Doppelgängers

- **Challenges**
  - Micro-text with l33t-speak
  - Multi-lingual
  - Text contains product information
  - Users with multiple accounts
Identifying product information

- Product information has repeated patterns but conversation contains verbs.
Identifying product information

- If the repeated pattern does not contain a verb, it is product information.
- Remove product information in pre-processing to achieve higher accuracy in user classification.

Bankname: XX
CCNumber: XXXXXXX
CHolder: XX XXXX
CExpire: X / XXXX
CVV2: XX
Vorname: XX
Nachname: YY
Adresse: XXXX
Stadt: XXXX
PLZ: XXXX
Land: XX
Telefon: XXXXX-XXXX
E-mail: [email]victim@example.com[/email]
Geburtsdatum: XX / XX / XXXX

Tag parts of speech

Noun(NN): CD
Noun(NN): CD
Noun(NN): CD
Noun(NN): CD
Noun(NN): CD SYM CD
Noun(NN): CD
Noun(NN): CD
Noun(NN): CD
Noun(NN): CD
Noun(NN): CD
Noun(NN): CD
Noun(NN): CD SYM CD
Noun(NN): CD
Noun(NN): CD
Noun(NN): CD SYM CD

Find repeated patterns

Check for verb

Product
Identifying users

- We train the classifier with a feature-set that adapts to foreign language and l33t-speak.
References

Thank you!

Questions & Answers