Privacy Issues in Big Data Processing in light of Data Breaches

S. Srinivasan, Ph.D.
Texas Southern University
Houston, Texas
Outline

• Big Data processing
• Steps to protect stored data
• Five major Data Breaches since 2013
• Threat to privacy due to breaches
• Ways to protect privacy
• Suggested Best Practices
Big Data Processing

• Big Data encompasses data gathered from multiple sources
• Benefit is derived when data from multiple sources are combined
• Data comes in structured and unstructured format
• Automated data processing is essential to derive benefits
• Speed of knowledge gathering from data is essential to benefit
• Example:
  ➢ Customer shares with friend interest in a jazz concert
  ➢ Customer searches reveal interest in jazz
  ➢ Customer tweets about a recent jazz news
Big Data Processing

• We have successfully evolved from:
  ➢ Agricultural age
  ➢ Industrial age
  ➢ Information age
  ➢ Knowledge age

• Moving towards the true goal of internet – being connected
• Much of our interaction today involves “connected to documents”
• This navigation is facilitated by hypertext
• Big Data enables knowledge derivation by combining data from multiple sources
Steps to Protect Stored Data

• One of the key aspects of Big Data is centralization of knowledge from disparate data
• This requires protecting such disparate data
• Such data contain:
  ➢ Personally Identifiable Information (PII)
  ➢ PHI (Protected Health Information)
  ➢ PCI (Payment Card Industry) data
• Different types of data require different levels of protection
• By law, all three types of data identified above require higher level of protection and violations are punished severely
Steps to Protect Stored Data

• PII data includes name, address, phone, SSN, DoB, email
• One way to protect PII is to collect less
• For example, instead of DoB, age might be enough
• Instead of address, zip code might be enough
• Know the application for which data is required and collect accordingly
• Encryption is the most common form of protection
• Use of encryption/decryption is time consuming and costly
• For example, HTTPS fully encrypts all data, which is redundant
Steps to Protect Stored Data

• Data protection could be accomplished using format preserving encryption whereby other applications used with that data would work properly
• Using format preserving encryption is expensive
• Another form of protection is by access control
• Access to data can be granted based on need
• Monitor all access through logs
• Monitor all changes made to the data by users
Steps to Protect Stored Data

• Two common access control requirements are:
  • Provide users with the minimal amount of access needed to perform job
  • User with privileges to approve is not the executor of the policy

• Role-based access would free up the time of IT administrators to focus on security threats

• Behavioral analytics should be used with access control because some insiders may abuse their privileges

• Sensitive information should be protected with greater access control

• Access control matrix could be a tool to accomplish this

• Use of alarms is common to alert specific people when suspicious activity is detected
Major Data Breaches

• We examine five major data breaches since 2013 to understand how breaches occur and how to prevent them

• Major data breaches considered are:
  ➢ Target in 2013
  ➢ Home Depot in 2014
  ➢ Chase Bank in 2014
  ➢ Anthem in 2015
  ➢ Office of Personnel Management in 2015
Target

- Data breach occurred in 2013 during peak holiday period
- 70 million credit card data were stolen through POS attack
- Hacker gained access to Target’s network first through the stolen credentials of a HVAC vendor from PA
- HVAC vendor was not the main supplier of HVAC services to Target but did electronic billing and Project Management
- Hackers exploited the vulnerabilities in Target’s network beyond initial limited access
- Target installed the powerful Mandiant technology by FireEye but did not monitor the alerts generated by the software
- Shows policy implementation failures within Target’s Information Systems Network
Home Depot

• Data breach occurred in 2014
• 56 million credit card data were stolen
• Hackers exploited vulnerabilities in the self-service POS terminals in stores
• Hackers entered the network similar to Target breach
• Data breach was detected by third parties prior to Home Depot acknowledging it
• Organization did not take data security seriously
Chase Bank

• Data breach occurred in 2014
• 76 million households and 7 million businesses were affected by the stolen data
• Only non-financial data was stolen
• Hackers gained entry to corporate network by stealing credentials of employees
• Data breach was detected first by third parties
• One lesson learned from this attack was that two-factor authentication would provide greater protection from data breaches
• 3 perpetrators from Israel were prosecuted in 2015 for this attack
Anthem

- Anthem is a US Insurance carrier operating in almost all states under different brands, often Blue Cross & Blue Shield
- Hack is suspected to be a state sponsored attack
- Compromised 78.8 million records in 2015
- People’s names, social security numbers, addresses were stolen but no credit card data
- Anthem did not encrypt stored data
Office of Personnel Management

- Data breach occurred in 2015
- Compromised records related to over 25 million individuals
- PII information was stolen
- Attributed to State-sponsored attack from abroad
- Information stolen was highly sensitive because the purpose of data collection was for security clearance
- OPM was slow to notify affected individuals
- Caused the Director of OPM to resign
Office of Personnel Management

• Hackers gained access to OPM network with stolen credentials in May 2014
• Started exfiltration of data pertaining to 20 million security clearance applicants in July 2014
• Rest of the data related to other family members of applicants
• Foreign spy agency could use the data to identify CIA agents working under diplomatic cover
• OPM does the investigation for security clearance applicants but the clearance is granted by the agency such as DoD requesting it
• Security clearances are not transferable between agencies
Causes for Data Breach

- Weak access control
- Non-use of two factor authentication
- Centralized data storage attracts hackers
- Use of untested third party Apps
- Lack of policy enforcement
- Lack of sharing information about a breach with other businesses
- Signature-based intrusion detection is outmoded in today’s attack scenarios
- Available tools are expensive to implement
Infographic on 2015 Breaches

Data Breaches

Data records lost or stolen in first six months of 2015

245,919,393

Only 4% of breaches were “Secure Breaches” where encryption was used and the stolen data was rendered useless.

1,358,671 records lost or stolen every day
56,611 records every hour
943 records every minute
16 records every second

Number of Breach Incidents by Type

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of Incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity Theft</td>
<td>472</td>
</tr>
<tr>
<td>Financial Access</td>
<td>111</td>
</tr>
<tr>
<td>External Data</td>
<td>119</td>
</tr>
<tr>
<td>Account Access</td>
<td>120</td>
</tr>
<tr>
<td>Malicious Insider</td>
<td>107</td>
</tr>
<tr>
<td>Malicious Outsider</td>
<td>62%</td>
</tr>
<tr>
<td>Hacktivist</td>
<td>12%</td>
</tr>
<tr>
<td>State Sponsored</td>
<td>2%</td>
</tr>
<tr>
<td>Accidental Loss</td>
<td>22%</td>
</tr>
</tbody>
</table>

Number of Breach Incidents by Source

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of Incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malicious Outsider</td>
<td>888</td>
</tr>
<tr>
<td>Accidental Loss</td>
<td>117</td>
</tr>
<tr>
<td>Hacktivist</td>
<td>19</td>
</tr>
<tr>
<td>State Sponsored</td>
<td>17</td>
</tr>
<tr>
<td>Malicious Insider</td>
<td>107</td>
</tr>
</tbody>
</table>

Source: Gemalto
Infographic on 2015 Breaches

Data Records Lost/Stolen by Industry

- **Healthcare**: 34% (84,426,356 records)
- **Government**: 31% (72,191,899 records)
- **Technology**: 15% (27,459,526 records)
- **Retail**: 8% (18,603,074 records)
- **Education**: 6% (15,708,330 records)
- **Other**: 5% (13,992,016 records)
- **Financial**: <1% (48,133 records)

Breach by Region

- **North America**: 707 incidents (89%)
  - United States: 611 incidents
  - Canada: 31
  - Mexico: 1
- **Europe**: 94 incidents (11%)
  - United Kingdom: 63 incidents
  - France: 9
  - Germany: 3
  - Italy: 1
  - Russia: 2
  - Switzerland: 2
  - Spain: 1
  - Portugal: 1
  - Poland: 1
- **South America**: 2 incidents (<1%)
  - Brazil: 1
  - Columbia: 1
- **Africa**: 2 incidents (<1%)
  - South Africa: 2 incidents
- **Asia/Pacific**: 63 incidents (7%)
## Per-record Cost of Data Breach by Industry

<table>
<thead>
<tr>
<th>Industry</th>
<th>Per-record cost for breach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare</td>
<td>$363</td>
</tr>
<tr>
<td>Financial</td>
<td>$215</td>
</tr>
<tr>
<td>Retail</td>
<td>$165</td>
</tr>
<tr>
<td>Industrial</td>
<td>$155</td>
</tr>
<tr>
<td>Consumer</td>
<td>$135</td>
</tr>
<tr>
<td>Energy</td>
<td>$132</td>
</tr>
<tr>
<td>Technology</td>
<td>$127</td>
</tr>
<tr>
<td>Public sector</td>
<td>$68</td>
</tr>
</tbody>
</table>

Source: IBM
# Data Breach Statistics

<table>
<thead>
<tr>
<th>Region</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td>76%</td>
</tr>
<tr>
<td>Europe</td>
<td>12%</td>
</tr>
<tr>
<td>Asia-Pacific</td>
<td>8%</td>
</tr>
</tbody>
</table>

## # of Breaches in 2014 in Europe

<table>
<thead>
<tr>
<th>Country</th>
<th># of Breaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>United Kingdom</td>
<td>117</td>
</tr>
<tr>
<td>France</td>
<td>9</td>
</tr>
<tr>
<td>Germany</td>
<td>7</td>
</tr>
<tr>
<td>Netherlands</td>
<td>6</td>
</tr>
<tr>
<td>Belgium</td>
<td>5</td>
</tr>
<tr>
<td>Italy</td>
<td>3</td>
</tr>
</tbody>
</table>

Source: Gemalto
Threat to Privacy due to Breaches

• Breaches expose customer PII data
• Credit monitoring for an year or so does not really solve the question of a hacker having customer SSN
• Many businesses are slow to notify customers in the event of a breach
• There is no single federal law requiring notification, instead there are 47 different state laws on this (AL, NM, SD)
• Exclusions apply for encrypted data
• See the problem with Ashley Madison breach because their data was encrypted. It is not a US company.
Ways to protect privacy

• Users of online systems should verify secure storage capabilities of systems
• Limit disclosure of sensitive data
• Verify why a site wants a particular type of sensitive data
  • Is there a need for DoB while age would do?
  • Is there a need for address while zip code would do?
• As a site developer, be aware of US federal requirements to protect sensitive data (HIPAA, SOX, GLBA, etc.). Similar constraints apply to Europe.
• Know what level of detailed data would be needed and what is the intended use for data
• Do not collect data first and then decide how to use it
Ways to protect privacy

• Storage costs are very low
• Cloud provides ability to store large volumes of data for long
• Have a policy to purge data from storage regularly
• Common data retention requirements are:
  • Email: 1 year
  • Personnel Info.: 7 years after termination
  • Medical/benefits: 6 years after plan year
  • Hiring Records: 2 years after hiring decision
• When data is retained there is cost for backup and recovery
Suggested Best Practices

• Evaluate the implications of any privacy protection policies developed to see if customers would find it acceptable

• Before collecting data or images, consider the privacy implications

• Even large companies like Google have to pay a penalty when image collected should not be made public (think of obfuscation now in Google Earth images)

• Businesses have obligations to its members when they change their Privacy Policy without adequate time or thought (note Facebook’s Privacy Policy change in 2012 to share information with affiliates)
Suggested Best Practices

• Have a Privacy Council in a business to develop Privacy Policies, with members from multiple departments
• Commitment of C-Suite to business Privacy Policies is essential
• Use searchable encryption such as Format Preserving encryption
• Technology is capable of many things and so the goal of Privacy should be by Design
• When collecting personal data, choose the opt-in approach as opposed to collecting all with option for customer to opt-out
Suggested Best Practices

• After a data breach, the hacker has more believable data to use to extract the information they need from a customer
• Alert customers as soon as a breach is detected
• To regain customer trust, communicate well with customers by directing them to resources such as the one from FTC to protect from identity theft
• To prevent data breach in the first place, segregate sensitive data to a secure place on the network
Summary

• Protecting personal privacy requires many people, including users
• Encryption is a common method to protect privacy and stored data
• Major breaches have occurred and more will occur
• Centralized storage attracts hackers and so greater care needs to be taken in data collection and storage
• Best protection is to share only essential data with others
• Keep privacy as default in policy development
References

• PHI http://security.stackexchange.com/questions/14376/can-phi-be-hipaa-compliant-on-a-cloud

• PII http://www.gsa.gov/portal/content/104256

• Kirk Hearth, ”Building a Privacy Program: A Practitioner’s Guide”, 2011, IAPP publication

• Ann Cavoukian, “Privacy By Design”, Toronto, Canada
Thank you!

Questions?

srinis@tsu.edu