ReedSmith

The business of relationships.℠

Understanding and Defending Big Data

Presented to International Association of Privacy Professionals
Princeton, New Jersey KnowledgeNet
December 5, 2012

Mark S. Melodia, mmelodia@reedsmith.com, 609-520-6015
Paul Bond, pbond@reedsmith.com, 609-520-6393

Understanding and Defending Big Data

- What is “Big Data”?
- Why is it under attack?
- What has the result been to date?

What is “Big Data”?

- Roughly 90% of all the digital data in the world has been created in the last two years.
- These data sets are:
  - very large, even by modern standards,
  - diverse, encompassing many data elements, often relating to different topics, and
  - unstructured – frequently, the data sets are really data dumps.

What is “Big Data”?

- Databases often includes non-traditional types of information.
- Data is collected and analyzed in closer-to-real-time than ever before.
- Companies are adopting new tools and processes.

What is “Big Data”?

- Better analytics have made data sets more valuable.
- Investors demand granular details on what information assets are held – see Facebook IPO and other 2012 offerings.
- Per IBM’s *2012* report, *Analytics: The Real-World Use of Data*, 75% of companies use or intend to use Big Data solutions*.*

What is “Big Data”?

- For example, when consumers drive the Ford Focus, an electric car, they generate driving data used by:
  - Ford engineers to improve car design,
  - Utilities to decide on choosing locations for charging stations,
  - Traffic planners to speed up traffic flow.

What is “Big Data”?

- Other use cases:
  - Health insurance plans use big data to make better coverage decisions and improve patient outcomes
  - Epidemiologists can track the outbreak of diseases by monitoring online activity
  - Use of Big Data can greatly improve the effectiveness of fraud detection, acting in real-time and drawing on multiple channels
  - Studios track social media minute-by-minute as people react to movies playing on opening weekend.

What is “Big Data”?

- Other use cases:
  - Improving the effectiveness of power grids by monitoring of smart meters
  - Better tailoring advertisements to consumers:
    - Online
    - Over the phone (determining which upsell to try based on market segment information)
    - And, for example, in-store (changing digital displays based on facial recognition)
  - Better understanding customers and potential customers

Where Does Big Data Come From?

- Social media, online game worlds, and other self-expressive activity
- Other online activity that leaves a trail
- Offline activity that is tied into information streams
- Sensors and the “Internet of Things”.

Examples of Big Data Sources – Online, Intentionally Expressive Conduct

- People feed into Big Data by what they do and say in online communities
- Through social network aggregation platforms, there is a multiplier effect of communications
- Social networks and game worlds capture a large amount of diverse interaction.

Examples of Big Data Sources – Online, Intentional Use of Social and Game Networks

- Users don’t just speak, they also:
  - Say which brands they like
  - Show which ideas they agree with
  - Map out networks of friends and family
  - Indicate social status
  - Volunteer where they come from
  - Make purchases
  - Send gifts
  - Calendar important upcoming dates in their lives.

Examples of Big Data Sources – Online, Intentionally Expressive Conduct

- The wide variety of social media sites show the diversity of information captured, as well as how unstructured such information can be. Social media sites can be;
  - General Use (e.g., Facebook, MySpace, Google+)
  - Professional (e.g., LinkedIn)
  - Mostly text-based (e.g., Twitter)
  - Mostly visual (e.g., Pinterest, Instagram)
  - Location-based (e.g., FourSquare)
  - Question and answer (e.g., Quora)
  - Debate (e.g., Debate.org)
  - Online game worlds (e.g., World of Warcraft)

Examples of Big Data Sources – Online, Intentional Use of Other Expressive Outlets

- In addition to participating in social media, consumers are creating more substantive, personal content:
  - Multi-media (e.g., YouTube)
  - Extended commentary (e.g., Tumblr, Reddit)
  - Reviews (e.g., imdb, Amazon, goodreads, Yelp)
  - Blogging and news aggregation.

Examples of Big Data Sources – Conduct Not Intended to Be Expressive

- Internet searches (many of which are now performed through social media)
- Online activity:
  - creating information stored on server logs
  - generating information stored in cookies
  - resulting in the user's receipt of a pixel or web beacon

In Addition, People Engage in Online Conduct Not Intended to Be Expressive, But Which Is Telling

- Purchases online can result in the creation of device prints for security
- Purchases online that can be processed by payment card networks
- Internet browsing activity can be logged
- Pulling up pages with social media plug-ins may tell social media sites where you are
- Accessing communications over other digital channels may create a trail:
  - Cable viewing habits
  - Streaming of online music or other multi-media

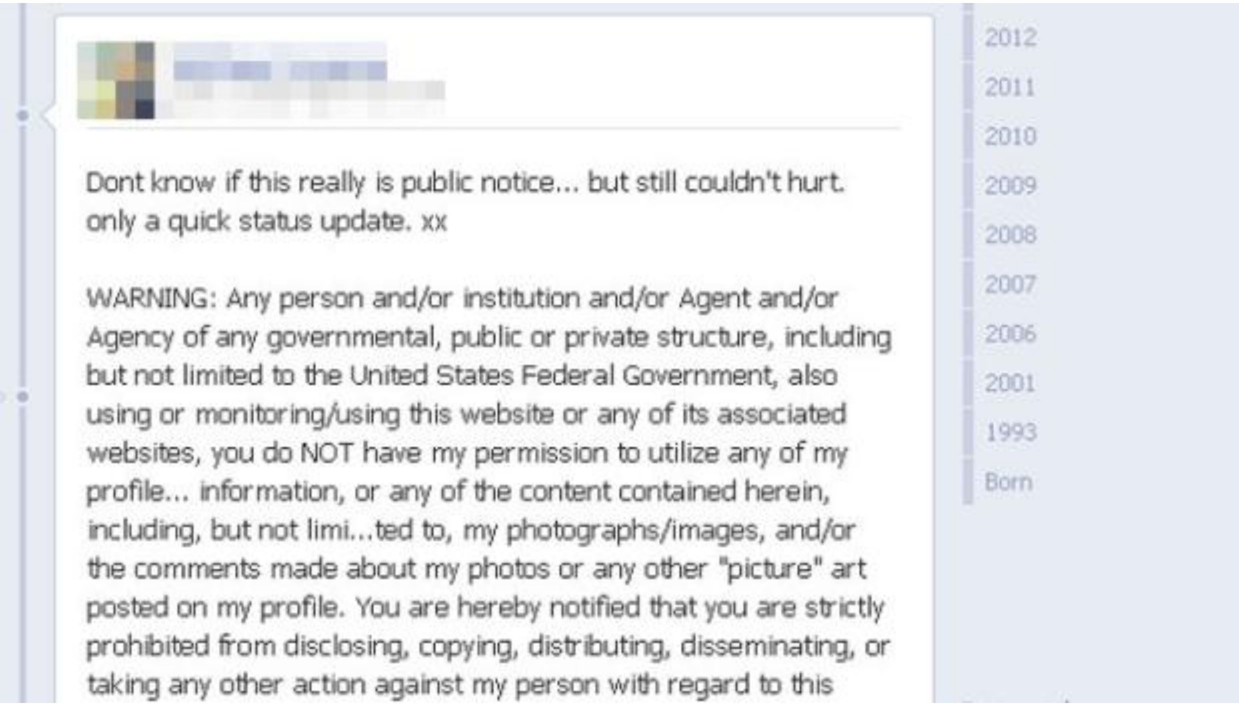
Offline Conduct Can Be Swept Into Big Data

- Location activity can be tracked with GPS-enabled devices
- Purchases with payment cards, or use of automatic payment devices
- Having a smart meter monitor energy consumption over time
- Filing court records or property records
- Walking past a monitor capable of facial recognition
- Driving past a monitor capable of license plate recognition
- Having an implanted device that sends information back to its manufacturer

Why is Big Data Under Attack?

- Anxiety
- Past failures of anonymization
- Concerns regarding how the profits of Big Data will be shared
- Additional concerns

Fear of Big Data And An Effort to Ward It Off



Dont know if this really is public notice... but still couldn't hurt. only a quick status update. xx

WARNING: Any person and/or institution and/or Agent and/or Agency of any governmental, public or private structure, including but not limited to the United States Federal Government, also using or monitoring/using this website or any of its associated websites, you do NOT have my permission to utilize any of my profile... information, or any of the content contained herein, including, but not limi...ted to, my photographs/images, and/or the comments made about my photos or any other "picture" art posted on my profile. You are hereby notified that you are strictly prohibited from disclosing, copying, distributing, disseminating, or taking any other action against my person with regard to this

Facebook Privacy Notice Hoax

- This notice is a hoax
- Every so often, it spreads like wildfire over Facebook
- Even after being debunked month after month
- People want to believe there is a simple solution.

Failures of Anonymization

- Privacy compliance concerns can chill adoption of Big Data solutions.
- Privacy focuses on “personal information”.
- So to get out of the privacy box, many companies rely on anonymization or psuedonomyziation.
- This approach offers significant comfort, but is not fail-safe.

Anonymization/Psuedonymization: Offers Significant Comfort in Principle

- GLBA excludes “aggregate information or blind data” from its privacy rules
- HIPAA has well-developed de-identification procedures
- In its report, “*Protecting Consumer Privacy in an Era of Rapid Change*,” the FTC also excluded such information:
- “as long as (1) a given data set is not reasonably identifiable, (2) the company publicly commits not to re-identify it, and (3) the company requires any downstream users of the data to keep it in de-identified form, that data will fall outside the scope of the framework.”

AOL Records Re-Identified Based On The Content of Searches

- Controls can fail.
- AOL released 20 million search records to the academic community in pseudonymous form.
- Journalists and researchers quickly re-identified individual users by name, using “tells” in searches.
- The release had already been mirrored all around the world, and could not be pulled back.
- This privacy event prompted a class action suit against AOL that lasted from 2006 to 2010.

Netflix Records Re-identified Based On Public Data At Other Film Websites

- Netflix released pseudonymous viewing records and reviews, as part of a contest to see who can improve search algorithms.
- Some academics compared the Netflix data with information on the Internet Movie Database.
- Since many movie lovers are on both, the researchers were able to deduce to the name of many Netflix users and see their viewing habits.
- Like AOL's release, the result was a class action against the releasing party, and a reduced public belief in psuedonymization.

Touched By A Flash Cookie: Tracking Litigation

- Adobe Flash video works, in part, but putting a Flash cookie onto the user's computer.
- That Flash cookie is not deleted when the user deletes browser cookies.
- Ad networks quickly realized the Flash cookie could be used to identify a device uniquely, and in a way that the user would be unlikely to address.
- Class actions followed against ad networks, resulting in settlements.
- Next, plaintiffs' attorneys sued both ad networks and brands who used them in the same action – resulting in a quick dismissal of the brands with prejudice.

Touched By A Flash Cookie: Tracking Litigation

- A rash of class action lawsuits have been filed against national brands in Missouri and Arkansas – without naming the ad networks or third party trackers as defendants at all.
- The Complaints allege generally that the defendants collect information wrongfully through Flash cookies, and share it with third parties.
- In this and other litigation, the plaintiff’s class action bar considers it irrelevant whether a name is connected with any cookie-generated information:
  - Plaintiffs assume any data subject will be re-identified in the end, by Big Data tools
  - And Plaintiffs assert that even a device-linked identifier is enough to trigger privacy protection requirements.

Opposition to Big Data From A Consumer Protection Point of View

- Many suits have been brought by groups of individuals who feel that companies shouldn’t profit from personal information.
- Those individuals have sought compensation for the value of their data as a marketing or analytics tool.
- So far, such suits for the value of personal information have not gained much ground, as described in the slides below.
- Despite the failure of such litigation, regulatory pressure continues unabated.

Additional Concerns: Web-lining

- The Bipartisan Congressional Privacy Caucus posed a pointed question to 12 supposed data brokers earlier this year, asking for:
  - Detailed information of what information each collected, stored, and use, potentially “extend[] far beyond rage, race, and sex”
  - Whether that information is used to “score” individuals “for purposes of offering access to education, healthcare, employment, and other economic opportunities”
  - The term the Caucus used for this scoring is “web-lining,” with explicit comparison to the prohibited, discriminatory practice of “redlining” that has caused so much litigation against mortgage companies and brokers.

Additional Concerns: Civil Liberties

- Governments and private industry routinely share information and cooperate with analytic projects
- There can be a negative public reaction to such projects, as evidenced by the costly NSA/Wiretap litigation
- In a recent Supreme Court case involving the use of GPS trackers, the Court noted that the device had sent more than 2,000 pages of locational data over 28 days, a level of surveillance that was not reasonable. *United States v. Antoine Jones*, 132 S. Ct. 945 (2012).
- The volume of information at issue may start to be one index of the reasonableness of government action.

Additional Concerns: Rights To Access And Duty To Destroy

- Because big data is large, diverse, and unstructured, it is correspondingly more difficult to search.
- Much of the data stored still consists of "icebergs" – blocks of data from which there is no practical way to draw intelligence.
- Yet, there is a growing push to give information subjects access rights to all information about them on demand.
- The right to access, joined with the nascent "right to be forgotten," will become all the more complex to accomplish in the Big Data era.

What Has The Result Been To Date?:

- For the most part, Courts have not stymied the collection, use, and re-use of information key to Big Data
- Objections that consumers have a right to more directly benefit from personal information have been largely unsuccessful.
- But litigation continues, and regulatory attention is at an all-time high.

In re Jetblue Airways Corp. Privacy Litigation, 379 F.Supp.2d 299 (E.D.N.Y. 2005).

- In general, information maintained by a company about an individual is not considered an economic asset of the individual.
- In *Jetblue*, airline passengers sued an airline for “unlawfully transferring their personal information …for use in a federally-funded study on military base security”. *Id*. at 303.
- The court found that there was “no support for the proposition that an individual passenger's personal information has or had any compensable value in the economy at large,” and hence plaintiffs stated no actual damages.

In re DoubleClick Inc. Privacy Litigation, 154 F.Supp.2d 497, 525 (S.D.N.Y. 2001).

- Courts have also resisted claims that consumers should be compensated for the value of information about their online activity.
- “[A]lthough demographic information is valued highly… the value of its collection has **never** been considered a economic loss to the subject.”
- “Demographic information is constantly collected on all consumers by marketers, mail-order catalogues and retailers. However, we are unaware of any court that has held the value of this collected information constitutes damage to consumers or unjust enrichment to collectors.”

CVS Caremark and *Walgreens* Cases

- There have even been suits about the resale of properly de-identified data sets, mainly involving pharmacy-level prescription data. These, too, have been rebuffed.
- “Under the circumstances the plaintiffs could have no reasonable expectation of being compensated for the information related to that transaction [filling a prescription], because that information carries with it no compensable value at the individual level.” *Steinberg v. CVS Caremark Corp*., 2012 WL 507807 at *9-10 (E.D. Pa. 2012).
- “the sale of de-identified prescription data does not carry a compensable value to consumers, and thus plaintiff had not shown that he was harmed by defendants’ actions.” *Todd Murphy v. Walgreen Corporation, et al*., docketed as 37-2011-00087162-CU-BT-CTL in the Superior Court of California, San Diego County (Minute Order of May 9, 2012).

Sorrell v. IMS Health Inc.

- United States Supreme Court case striking down a law designed to preclude certain analysis of Big Data sets
- Several states passed laws limiting the use of data about the prescription-writing habits of physicians.
- The laws prohibited companies like *IMS* from buying that information to derive actionable marketing intelligence for outreach to physicians.
- The Supreme Court found that these laws violated the First Amendment by limiting speech in a way that hinged on the content of the speech.
- In so doing, it gave short shrift to claims that outreach violated physician privacy.

However…

- Privacy class actions continue to be filed *en masse*:
  - Many of them presupposing that personal information has economic value
  - Many of them presupposing that all information collected will eventually be re-linked to a person
- The FTC continues to evidence an interest in more fully regulating “major platform providers” and “information brokers,” as well as casting a wider net with existing statutes like the FCRA and COPPA
- The actual implementation of Big Data analysis raises a host of legal questions – contracting, tax, cross-border transfer, insurance, anti-trust – as to which very little has yet been settled.