Data Analytics Governance and Ethics

Dr. Flavio Villanustre, CISSP
VP, Technology
26 February 2014

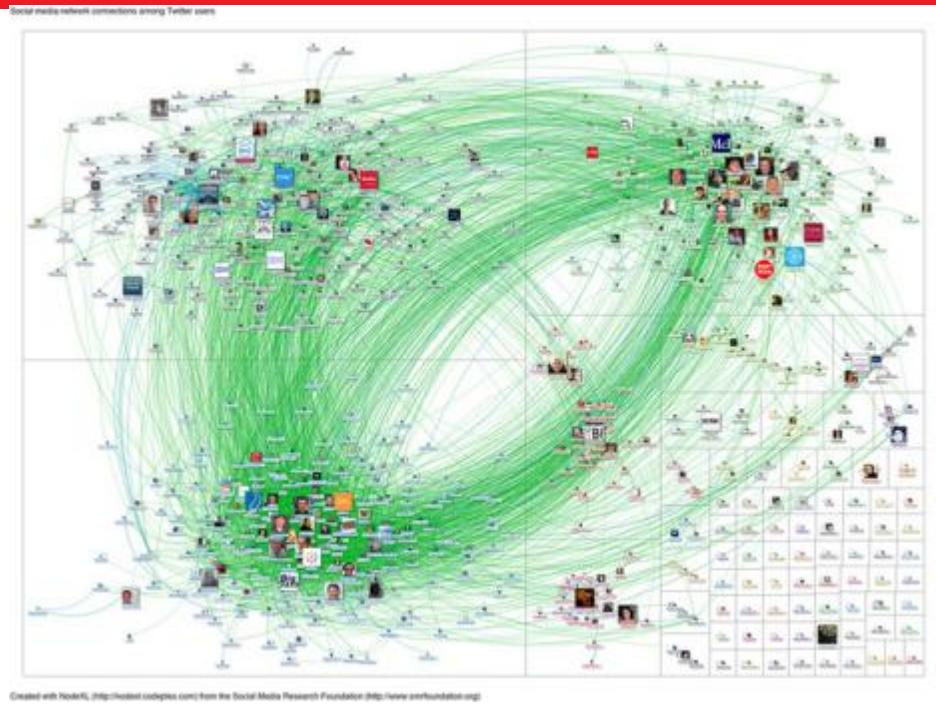
LexisNexis®

Risk Solutions

What is Big Data?

- Defined by its four dimensions:
  - Volume
  - Velocity
  - Variety
  - Value

- Driven by the proliferation of social media, sensors, the Internet of Things and more



- Open source distributed data intensive platforms (for example, the open source LexisNexis HPCC Systems platform and Hadoop) help its adoption

- Consumer oriented services, such as recommendation systems and search engines, made it popular

LexisNexis Risk Solutions

- Provider of data and analytics-based solutions
- Helps clients across all industries assess, predict and manage risk associated with the people and companies they do business with
- Leading positions in insurance, financial, legal and collections
- Solutions fueled by the most comprehensive database of public records information in the U.S., including 37 billion public records and contributory databases
- Products delivered by state of the art open and proprietary technology and analytics, designed for speed and scalability, in processing today's Big Data challenges
- Built on the LexisNexis 40-year reputation as a trusted custodian of essential information
- Designed and developed its own Big Data analytical platform (the HPCC Systems Big Data Platform) over the past 15 years.

Big Data challenges

Technical challenges:

- Regular integration of tens of thousands of data sources (ingestion, parsing, cleansing, normalization, standardization and linking)
- Data is usually dirty, contain errors and duplicates and has missing information
- Unique identifiers/keys are non-existent
- Entities (individuals/companies) are not directly observable
- Relationships between attributes and entities are non-obvious, and neither are relationships across entities

Other challenges:

- Security – Least privilege, roles based access controls, accountability
- Privacy – Data protection frameworks differ from country to country
- Legal and regulatory requirements – GLBA, DPPA, FCRA, HIPAA, SOX, PCI, etc.
- Ethical – Just because it can be done, should it?

Security and Privacy

- Security
  - Keeping the bad guys out
  - Making sure the good guys are good and stay good
  - Preventing mistakes
  - Disposing of unnecessary/expired data
  - Enforcing “least privilege” and “need to know” basis

- Privacy
  - Statistics safer than aggregates
  - Aggregates safer than tokenized samples
  - Tokenized samples safer than individuals
  - Don’t underestimate the power of de-anonymization
  - Mistakes in privacy are irreversible

- Security <> Privacy

Data security concepts

Big Data brings new challenges

a. Additional data sources can greatly increase the information: data + data can be > 2 * information
b. Public clouds offer scalability but introduce risks
c. Distributed data stores can make it harder to enforce access controls
d. Since interdisciplinary teams may be required, more people needs access to the data repository

Established practices may not be completely effective

a. Tokenization only goes so far
b. Encryption at rest just protects against misplaced hardware
c. Tracking data provenance can be difficult
d. Enforcing data access controls can quickly get unwieldy
e. Conveying policy information across multiple systems can be challenging

Be wary of “tokenization in a box”

- Tokenized dataset * Tokenized dataset ~= identifiable data
- The problem of eliminating inference is NP-complete
- Several examples in the last decade:
  - The “Netflix case”
  - The “Hospital discharge data case”
  - The “AOL case”

Effective controls for Big Data

- Track data provenance, permissible use and expiration through metadata (data labels and RCM)
- Enforce fine granular access controls through code/data encapsulation (data access wrappers)
- Utilize homogeneous platforms that allow for end-to-end policy preservation and enforcement
- Deploy (and properly configure) Network and host based Data Loss Prevention
- A comprehensive data governance process is king
- Always use proven controls: although differential privacy, private information retrieval and homomorphic encryption show promise, they still need to mature to become mainstream
- Destroy data as soon as it's no longer needed

Always keep in mind that…

- Access to the hardware ~= Access to the data
  - Encryption at rest only mitigates the risk for the isolated hard drive, but NOT if the decryption key goes with it
  - Compromise of a running system is NOT mitigated by encryption of data at rest
- Virtualization may increase efficiency, but…
  - In virtual environments, s/he who has access to your VMM/Hypervisor, also has access to your data
  - Cross-VM side-channel attacks are not just theoretical

Ethical Big Data Analytics: a possible framework

- Clarity on practices
- Simplicity on settings
- Privacy by design
- Exchange of value

Remember

- Data exhibits its own form of **quantum entanglement**: once a copy is compromised, all other copies instantly are
- **The closer to the source** the [filtering, grouping, tokenization] is applied, **the lower the risk**
- Do you want to be cool or creepy? The choice is yours!

Useful Links

- LexisNexis Risk Solutions: http://lexisnexis.com/risk
- LexisNexis Open Source HPCC Systems Platform: http://hpccsystems.com
- Cross-VM side-channel attacks: http://blog.cryptographyengineering.com/2012/10/attack-of-week-cross-vm-timing-attacks.html
- K-Anonymity: a model for protecting privacy: http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf
- Robust de-anonymization of Large Sparse Datasets (the Netflix case): http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
- Broken Promises of Privacy: Responding to the surprising failure of anonymization: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006
- Tamper detection and Relationship Context Metadata: http://blogs.gartner.com/ian-glazer/2011/08/19/follow-up-from-catalyst-2011-tamper-detection-and-relationship-context-metadata/

Questions?

Email: info@hpccsystems.com