# Interdisciplinary Approaches to Financial Stability

CENTER ON FINANCE, LAW, AND POLICY
UNIVERSITY OF MICHIGAN

OFR OFFICE OF FINANCIAL RESEARCH

## Panel 3: Sharing Financial Data – Challenges in a Networked World
## Thursday, October 22, 2015 at 2:00 p.m.
## Hutchins Hall 100

*Moderator*:
Mark Flood, Office of Financial Research

*Panelists:*
Doug Burdick, IBM Research
John Clippinger, MIT Media Lab/ID3
H.V. Jagadish, University of Michigan
David Saul, State Street

### 1. The Benefits of Large-scale Data Integration

Public financial data is becoming increasingly available in large volumes (e.g., SEC EDGAR, Federal Reserve FRED, MSRB EMMA), with a significant portion available as unstructured text. Big Data infrastructures are becoming increasingly sophisticated, and now have the capability to perform extraction and integration across large, heterogeneous, unstructured datasets to identify entities of interest and their relationships. This "knowledge graph" of entities and their relationships serves as a sophisticated data-driven model of a financial ecosystem, enabling complex systems level "meta-modeling" techniques used in climate modeling and systems biology over financial and economic data.

The following examples illustrate use-cases where large-scale extraction and integration from relevant public data enabled creation of new applications enable analysis both at a granular and system-wide level

- Counterparty Risk Analysis from SEC and FDIC filings [5, 6]
    - Extract and integrate information for 3,000 publicly-traded financial institutions and 30,000 company insiders (officers / directors) from over 2 million unstructured regulatory filings with SEC and FDIC between 2005 – present.
    - For each company, populate material information about the company, such as officers / directors, institutional holdings, material events (SEC Form 8-K), and financial metrics (SEC 10-K / 10Q).

- Identify relationships between public companies, including common insiders (e.g., companies sharing an officer / director), subsidiaries, mergers / acquisitions, and syndicated loans between companies.
- For company insiders, populate information about employment history and reported transactions.

There are two interesting system-wide analyses this knowledge graph of financial companies supports. First, as described in [6], co-lending relationships among banks allow identification of which banks are most central to the lending network from public data, and, hence, pose the greatest risk to the financial system. Second, the common insider relationships allow for discovery of social networks among officers and directors at public companies.

- Modeling Residential Mortgage Supply Chain [7]
  - Mortgage backed security (MBS) prospectus are publically available SEC filings. Extract from each MBS prospectus filing information about MBS contract, including: participant institutions with their roles (e.g., issuer, servicer, depositor), payout structure of MBS contracts ("waterfall structure"), amount, and issue date.
  - Resolve identified participants across MBS contracts to build a MBS participant graph, which records for each participant institution which MBS contracts it has a role in.

The resulting knowledge graph of the mortgage supply market enables analysis of the entire MBS market structure. This graph also enables identifying risk level of market participants and which participants are central to the MBS market.

The increased availability of public financial data and automated processing capabilities of Big Data systems are creating additional opportunities to leverage economic value of financial data.

- As described above, the availability of public financial data and sophistication of Big Data systems enable creation of complex data-driven models of financial eco-systems.
- Data is no longer necessarily just "exhaust" (i.e., something created as a byproduct of a process, such as regulatory compliance), and may have economic value when merged or integrated with other publically available financial data.
- Leverage value by sharing data in appropriate manner: data marketplaces, partnerships, controlled release, etc.

**The Importance of Big Data**

- Big Data is important. It matters at least as much to the finance sector as to any other sector of our economy.
  - Can be required for regulatory compliance.
  - Can be required for enterprise risk management.
  - Can deliver profits!!
- A central prerequisite for unlocking the value of Big Data is integrating data from multiple sources. This integration can be challenging.
  - Standards are crucial.

- Whether or not you want to share information today, assume that you will at some point in the future.
    - Hugely impacts economics of information sharing.
- Control sharing through explicit control – can share with regulators, for example, but not competitors. Can also share selectively with vendors, clients, …
    - Intentional inability to share is not a winning strategy.
- What can you share? What is competitive value you should not share?
    - Rules are generally overly restrictive – always the safer choice when in doubt
- Data mining from public sources can reveal more than we expect.
    - E.g. Korea http://www.nkeconwatch.com/north-korea-uncovered-google-earth/
    - These sources are not utilized enough today for the most part.
- Must invoke the power of the intellectual community, and not just a favored few.
    - Must find means to publish and share in ways that are not competitively harmful: through adding up, adding noise, or adding delay.

## 2. The Sharing Economy - Organizations: The End of Silos and Rise of P2P Networks

The typical current conversation about "privacy" and security is inherently misguided it is based on a pre-Internet= mobile Net - IOT world in which data and computation were not pervasive and autonomous. There is the presumption that we can and should stop the flow of data:

Classic Regulatory Framework

- Do not collect = do not see
- Opt out = digital pariah
- Notification and consent = denial of service attack
- Do not track = do not know-learn
- Do not share = digital ghetto

Harms and Duties come with the use of the specific use of the data and chain of custody

New Data Realities:

- We are immersed in Data: Data is Water New Reality for all
- Continuous Sharing - Exchange is essential
- Data as asset classes
- API Economy
- Security by Design - Architecture of distributed - autonomous - authorities

- Control at The Edge - P2P
- Self-configuring and self-correcting
- Decentralized Public and Private Keys
- Chain of custody for data assets

### 3. Sharing cyber threat information

Would sharing cyber threat information among companies help to reduce the number of damaging incidents and lessen their impact? Can threat sharing programs and standards be devised that protect the participants from legal, privacy and financial risks?

The Verizon 2015 Data Breach Investigations Report (DBIR), conducted with contributions from 70 global organizations, is generally considered as the most accurate world-wide source of cyber threats, vulnerabilities, and security incidents. In 2015 Verizon introduced information on the impact of cyber incidents and a new model for estimating their costs. Verizon estimates that the cost of a breach involving 10 million records is between $2.1 million and $5.2 million but could go as high as $73.9 million.

Some highlights quoted from the 2015 Verizon DBIR report:

- Victim Demographics - The top three industries affected are the same as previous years: Public, Information and Financial Services.
- Breach Trends – In 70% of the attacks where we know the motive for the attack, there's a secondary victim. Unfortunately, the proportion of breaches discovered within days still falls well below that of time to compromise.
- Indicators of Compromise – 75% of attacks spread from Victim 0 to Victim 1 within one day (24 hours).
- Phishing – The reality is that you don't have time on your side when it comes to detecting and reacting to phishing events.
- Malware – Half of organizations discovered malware events during 35 or fewer days in 2014.
- Impact – The forecasted average loss for a breach of 1,000 records is between $52,000 and $87,000.

Activities

As evidenced by the Verizon DBIR and confirmed from other sources earlier detection of cyber incidents will lessen both their spread and impact. Efforts to encourage sharing of cyber threats have mainly centered on industry groups like the Financial Services Information Sharing and Analysis Center (FS-ISAC). Other industry ISACs share among themselves and the Nation Council of ISACs provides cross-industry consolidation. Regional groups like the Advanced Cyber Security Center (ACSC) in New England have developed local trust mechanisms to confidently share threat profiles.

In late 2014 the National Institute of Standards and Technology (NIST) issued a Guide to Cyber Threat Information Sharing (Draft). On the legislative front, the proposed National Cybersecurity Protection Advancement Act (NCPA) protects companies from customer lawsuits after they voluntarily share cyber threat information with each other and with government agencies.

National Institute of Standards and Technology (NIST)

NIST recommendations include eight best practices that organizations should take to share cyber threat information:

- Organizations should perform an inventory of the information they have and how it could be shared.
- Share threat intelligence with partners and learn from each other.
- Use open standards to foster interoperability.
- Enhance internal findings with outside information sources.
- Use a life cycle approach to develop defenses at all stages of an attack.
- Commit the necessary resources and training to an ongoing cyber program.
- Protect sensitive information through awareness and controls.
- Build the infrastructure needed to monitor and control cyber security, including timely patching of vulnerabilities.

Standards

NIST's third best practice emphasizes the importance of open standards. Verizon uses and promotes the VERIS Framework. The Vocabulary for Event Recording and Incident Sharing (VERIS) is a set of metrics designed to provide a common language for describing security incidents in a structured and repeatable manner. The DHS Office of Cybersecurity and Communications are leading efforts to automate and structure operational cyber security information with technical specifications designed to enable automated information sharing.

- TAXII™ defines a set of services and message exchanges that enable sharing of actionable cyber threat information. TAXII defines concepts, protocols, and message exchanges.
- STIX™ is an effort to define and develop a standardized language to represent structured cyber threat information. TAXII is the transport mechanism for cyber threat information represented as STIXII.
- CybOX™ is a standardized schema for the specification, capture, characterization and communication of events or stateful properties that are observable in the operational domain.
- Other standards relevant to sharing cyber threats include:

- CAPEC™ is a comprehensive dictionary and classification taxonomy of known attacks that can be used by analysts, developers, testers and educators to advance community understanding and enhance defenses.
- MAEC™ is a standardized language for encoding and communicating information about malware based on attributes such as behaviors, artifacts and attack patterns. MAEC aims to improve human-to-human, human-to-tool, tool-to-tool and tool-to-human communication about malware for faster development of countermeasures.

## Sharing Barriers

Countering the benefits of sharing cyber threats there are several obstacles that need to be overcome. They have a common thread of trust. How can an organization ensure that it can trust the information it receives from other companies and government sources? In return, can organizations trust that the information they share about their experiences will be protected and not used against them? A worst case scenario would be to share information with a would-be attacker who would now have an advantage.

The various threat sharing standards all include methods for strong authentication to increase the confidence level that information is actually coming from the represented party. Intermediaries like the FS-ISAC and the ACSC provide a trusted third party to both authenticate and anonymize cyber threat data. Legislation, like NCPA, is also needed to indemnify organizations against legal actions.

## Collaboration

The benefits of sharing cyber threat information clearly outweigh the disadvantages of non-sharing. The malefactors share their tools and techniques with great success. It is time for governments and companies to band together to combat these threats.

**Sharing cyber threat information**

| Point | Counterpoint |
|---|---|
| The number, breadth and sophistication of data breaches are increasing. The reporting of data breaches has expanded (e.g. Verizon DBIR). | Organizations are being forced to increase spending on cyber defense and reporting, taking resources away from new development. |
| Sharing cyber threat information among companies helps to shorten discovery times and accelerate remediation times of attacks. Speed is critical as the time between infiltration and attack decreases. | Adversaries already share code and data. The latest attack methodologies are readily available on the Internet. For some time most attacks have been so called "zero day" which largely defeats the value of sharing cyber threat intelligence. |
| Most data breaches are initiated by a human action like clicking on an email link. "Phishing" works. Users are unaware of when they activate malware imbedded in images. | User education, including SPAM testing, can significantly reduce the number of incidents. But it is not 100% effective. Hiding code in images can be activated without clicking. |
| SPAM filters are largely effective (>90%) at filtering out the most common unsafe emails. | To be most effective SPAM filters block a number of valid emails. |
| Other cyber defense tools and processes are improving. Defense in Depth coordinates different safety measures. 95% of all malware attacks workstations and their local drive designations. | Organizations have don't know where and when an attack will occur. They have to monitor everywhere. Attackers only have to succeed one time. |
| Traditional tools recognize malware by profiles and signatures. | Modern malware morphs itself to avoid signature detection. |
| The time between an intrusion and its consequences can often be measured in months. | Attackers can be patient and use the time to probe for weaknesses. |
| The point of attack and the point of exploitation are different. | Correlating event logs across different systems is a complex data problem. |
| Diligent patching provides protection against known threats. | Zero day attacks are more frequent and occur too quickly to be patched. |
| Industry sharing groups like the Financial Services Information Sharing and Analysis Center (FS-ISAC) have helped to curtail the spread of malware. | Cross-industry sharing is limited to date. Groups like the Advanced Cyber Security Center (ACSC) only work in their own geographic area. |
| The National Institute of Standards and Technology (NIST) issued a draft Guide to Cyber Threat Information Sharing with best practices. | Only the very largest organizations have the resources and skills to implement the NIST recommendations. |

| | |
|---|---|
| Standard data formats for sharing of cyber threats have been promulgated (TAXII™ and STIX™). Government and industry are starting to work together to promote these standards. | Adoption of these standards has been very sparse and uneven. Broader public/private partnership agreements need to be established to describe the actual use of TAXII and STIX. |
| Sharing cyber threat information potentially violates national security and privacy laws as well as the intellectual property and proprietary interests of organizations. | The proposed National Cybersecurity Protection Advancement Act (NCPA) protects companies from customer lawsuits after they voluntarily share cyber threat information with each other and with government agencies. |
| New secure operating system and browser technologies are being developed that don't have the vulnerabilities of existing software. Many effective anti-virus and anti-SPAM products already exist for earlier operating system version. | Moving legacy applications to new platforms is a hugely complex effort. Often needed documentation is unavailable. Many companies don't take advantage of products already available. |
| New disruptive technologies have the potential to totally change the way in which financial transactions operate. | These approaches will require massive changes not just to IT systems but to business processes as well. |

## 4. Sharing Financial Data and the Economics of Information

The interplay between information, transparency, and the structure of markets and institutions is complex [1]. In the 19th century, for example, the "bucket shops" of New York and Chicago — unlicensed, off-exchange gambling houses — threatened the franchise of the stock exchanges by operating as unsavory free riders on the price-discovery process. The exchanges fought back, ultimately establishing property rights in their price data, allowing them to starve the bucket shops of information [2]. Bond ratings and similar credit scorecards involve a lossy projection of detailed data on borrower creditworthiness into a finite set of ratings grades, a process that involves an intentional discarding of information [3].

More generally, the economic forces include:

- The industrial organization (the numbers, types, and sizes of participants) of the financial services industry is intimately bound up with problems of information.
    - Asymmetric information can convey competitive (dis)advantages, affecting the optimal scale of financial firms. Conversely, size can generate opportunities to manufacture advantageous information.
- Asymmetries arise naturally in financial markets, creating a need for regulation.
    - For example, counterparty networks of contractual exposures exhibit "endogenous myopia": Each dealer can see its own counterparties, but cannot peer any deeper into the network. There is a role for impartial supervisors to help monitor the system as a whole.

    - However, because the regulatory system is adapted to the institutional structure of firms and markets, the datasets available for sharing among individual regulatory agencies are idiosyncratic and constrained.

- Transparency can often create value by alleviating coordination problems.
    - Financial information typically exhibits the non-rival characteristic of a public good, and can be made non-excludable via publication.
    - We should therefore expect information to be under-produced in a competitive equilibrium – again creating a role for welfare-enhancing intervention by regulators.
    - A primary channel by which information revelation enhances coordination is by reducing information asymmetries. This can enhance liquidity – but it can also drive liquidity away in extreme cases.
- Successful disclosure regimes require careful attention to the particulars of the case: who is disclosing what to whom, and why.
    - Simplistic "reveal everything" disclosures can backfire by discouraging market participation or overwhelming recipients with superfluous detail.
- Fully common knowledge can generate benefits beyond those created by simple symmetric information.
    - In particular, information sharing among regulators should facilitate mutually beneficial coordination.
- Successful sharing requires good data standards. Both senders and recipients must agree on formats and semantics.
    - Data sharing requirements are state-dependent. The level of transparency (i.e., detail) needed will be much higher during a crisis or after a failure that in the routine course of events. Standards are the key to successful sharing in emergencies.

## References


[1] M. Flood, J. Katz, S. Ong, & A. Smith. (2013). Cryptography and the Economics of Supervisory Information: Balancing Transparency and Confidentiality. *Office of Financial Research Working Paper #0011*. September.

[2] J. Mulherin, J. Netter, & J. Overdahl. (1991). Prices are property: the organization of financial exchanges from a transaction cost perspective. *J. of Law and Economics,* 591-644.

[3] B. Holmström. (2012). The nature of liquidity provision: When ignorance is bliss. *Presidential Address, Econometric Society, ASSA Meetings, Chicago*. http://economics.mit.edu/files/7500 and http://www.youtube.com/watch?feature=player_embedded&v=4yQO512B-_A.

[4] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, & C. Shahabi. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7), 86-94.

[5] D. Burdick, A. Evfimievski, R. Krishnamurthy, N. Lewis, L. Popa, S. Rickards, and P. Williams. (2014). Financial Analytics from Public Data. In *Proceedings of the International Workshop on Data Science for Macro-Modeling* (DSMM'14).

[6] D. Burdick, M. A. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, S. Vaithyanathan, S. R. Das. (2011). Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study. *IEEE Data Engineering Bulletin*, 34(3): 60-67.

[7] D. Burdick, M. Franklin, P. Issler, R. Krishnamurthy, L. Popa, L. Raschid, R. Stanton, and N. Wallace. (2014). Data Science Challenges in Real Estate Asset and Capital Markets. In: *Proceedings of the International Workshop on Data Science for Macro-Modeling* (DSMM'14).