Using Data for Systemic Financial Risk Management

Mark Flood
University of Maryland
mdflood@rhsmith.umd.edu

H. V. Jagadish
University of Michigan
jag@umich.edu

Albert Kyle
University of Maryland
akyle@rhsmith.umd.edu

Frank Olken
LBL
frankolken@acm.org

Louiqa Raschid
University of Maryland
lraschid@umiacs.umd.edu

ABSTRACT

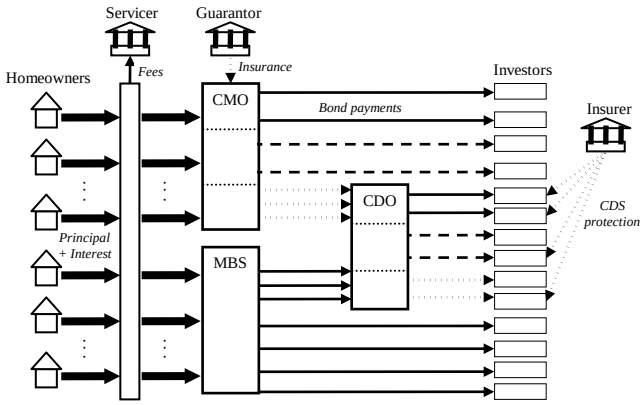
The recent financial collapse has laid bare the inadequacies of the information infrastructure supporting the US financial system. Technical challenges around large-scale data systems interact with significant economic forces involving innovation, transparency, confidentiality, complexity, and organizational change, to create a very difficult problem. The post-crisis reform legislation has created a unique opportunity to rebuild financial risk management on a solid foundation of information management principles. This should help reduce operating costs and operational risk. More importantly, it will support both the monitoring and the containment of financial risk on a previously unprecedented scale. These objectives will pose several information management challenges, including issues of knowledge representation, information quality, data integration, and presentation. This paper presents a vision of an information-rich financial risk management system, and a research agenda to facilitate its realization.

1. INTRODUCTION

The banking crisis that erupted in 2008 underscored the need for reductions in systemic financial risk. [4] While there were myriad interacting causes, a central theme in the crisis was the proliferation of complex financial products that overwhelmed the system's capacity for appropriately diligent analysis of the risks involved. In many cases, the information available about products and counterparties was minimal. In response, the Dodd-Frank Wall Street Reform Act – among its many regulatory changes – has created an Office of Financial Research (OFR) with the mandate to establish a sound data management infrastructure for systemic-risk monitoring. The OFR will contain a Federal Financial Data Center (FFDC) to manage data for the new agency.

For many years, both financial firms and their regulators have been hampered by a state of "data anarchy," despite (or perhaps because of) the enormous volumes of mission-critical data the industry handles daily. The widespread use of PCs has dispersed access, ownership, and control of data throughout the firm, to create multiple, overlapping data silos. The result is disparate, inconsistent and inaccurate information. Furthermore, cross-institutional barriers often inhibited regulators from obtaining the data that could have resulted in recognizing systemic risk early on, and thus, potentially preventing a timely response to emergent failures.

CIDR'11, Jan. 2011, Asilomar, CA, USA.



To be more concrete, consider the highly stylized example of home mortgage payments passing through the securitization chain depicted in the figure above. Homeowners submit principal and interest payments to a servicing bank, which collects a fee, passing the bulk on to securitization pools (trusts) that own the mortgages. A "pass-through" mortgage-backed security (MBS) pro-rates the payments to its bondholders, thus providing diversification benefits. A collateralized mortgage obligation (CMO) does the same, but structures the payments into prioritized tranches targeted to specific credit-risk and maturity clienteles (dashed lines indicate contingent cash flows). MBSs and CMOs typically have some credit-support, here from a third-party guarantor. Some of the MBS and CMO bonds are held directly by investors, but others are pooled into a collateralized debt obligation (CDO), which re-tranches the cash flows again to focus the credit and maturity exposures further. In the figure, some of the CDO investors have purchased additional credit protection, in the form of credit default swaps (CDS).

This greatly simplified depiction elides any number of important intricacies, such as the loan origination process, fixed vs. floating interest rates, homeowners' prepayment and curtailment options, the choice of funding sources to hedge interest-rate risk, tax and accounting treatment, government guarantees, the role of ratings agencies, portfolio management for the CDO pool, etc., etc. Nonetheless, the figure is already quite complicated. As most of us know, each mortgage loan is itself a complex legal contract, but the securitization (MBS, CMO and CDO) agreements are

typically much more involved still, with prospectuses and other offering documents that run on for hundreds of pages of legalese.

For an investor or investment manager, a key problem is determining the nature and magnitude of the financial risks she is exposed to through her various contracts. [3] A special challenge for the regulator is to determine the risks to the financial system as a whole, taking into account the correlations and dependencies between defaults, prepayments, movements in interest rates and other prices, institutional leverage, liquidity shocks, etc. Achieving the vision of computing risk through multiple counter-parties to complex contracts presents significant information management challenges. [5] In Section 2 we provide some background on the financial information system, and in Section 3 we discuss the information management challenges.

2. FINANCIAL SYSTEMS BACKGROUND

Systemic risk monitoring is inherently complex. Financial institutions acquire information from hundreds of sources, including prospectuses, term sheets, corporate filings, tender offers, proxy statements, research reports and corporate actions. They load the data into master files and databases providing access to prices, rates, descriptive data, identifiers, classifications, and credit information. They use this information to derive yields, valuations, variances, trends and correlations. They feed raw data streams and derived data into pricing models, calculation engines and analytical processes. These data are linked to accounting, trade execution, clearing, settlement, valuation, portfolio management analysts, regulators and market authorities. Further complicating these daunting technical challenges, there are also strong incentives for many market participants to restrict transparency around risks [7].

The data quality gap in finance is an evolutionary outcome of years of mergers and internal realignments, exacerbated by business silos and inflexible IT architectures. Difficulties in unraveling and reconnecting systems, processes, and organizations – while maintaining continuity of business – have made the problem intractable. Instead, data are typically managed on an ad-hoc, manual and reactive basis. Workflow is ill defined, and data reside in unconnected databases and spreadsheets with multiple formats and inconsistent definitions. Integration remains point-to-point and occurs tactically in response to emergencies. Many firms still lack an executive owner of data content and have no governance structure to address funding challenges, organizational alignment or battles over priorities.

Financial risk and information managers are gradually recognizing the concepts of metadata management, precise data definitions based on ontologies, and semantic models and knowledge representation as essential strategic objectives. [1] These same concerns apply with equal urgency to the financial regulators tasked with understanding individual firms and the overall system. A sound data infrastructure and open standards are necessary, both for effective regulation and for coordinating industry efforts. The history of patchwork standards and partial implementations demonstrates the tremendous obstacles to consensus over shared standards in the absence of a disinterested central authority.

The CDO depicted in the figure provides a simple example of the scale of the problem. A CDO might pool bonds from scores of MBSs, each of which pools hundreds of mortgages, entailing a total of many thousands of pages of contractual language. All of this legalese must be implemented in computer and information systems for each of the hundreds of participants in the pipeline.

While risk monitoring is not a precise science, there is a general consensus that systemic risk monitoring should consider at least [6]:

- forward-looking risk sensitivities to stressful events (e.g., what would a 1% rise in yields mean for my portfolio?);
- margins, leverage, and capital for individual participants (e.g., how large a liquidity shock could I absorb before defaulting?);
- the contractual interconnectedness of investors and firms (e.g., if Lehman Bros. fails, how will that propagate to me?)
- concentration of exposures, relative to market liquidity (e.g., how many banks are deeply exposed to California real estate?)

The OFR/FFDC will need the following types of information [2]:

- Financial instrument reference data: information on the legal and contractual structure of financial instruments, such as prospectuses or master agreements, including data about the issuing entity and its adjustments based on corporate actions;
- Legal entity reference data: identifying and descriptive information, such as legal names and charter types, for financial entities that participate in financial transactions, or that are otherwise referenced in financial instruments;
- Positions and transactions data: terms and conditions for both new contracts (transactions) and the accumulated financial exposure on an entity's books (positions);
- Prices and related data: transaction prices and related data used in the valuation of positions, development of models and scenarios, and the measurement of micro-prudential and macro-prudential exposures.

Together, these data can resolve the fundamental questions of who (i.e., which specific legal entity) is obligated to pay how much to whom, on which future dates, and under what contingencies. Based on this, one can assess both firm-wide and system-wide risk, and gains insights on the risks to consumers posed by particular financial products and practices.

3. RESEARCH CHALLENGES

To determine systemic risk from the large volume of complex and heterogeneous data describing the financial system, regulators (and industry participants) must understand ownership hierarchies, and counterparty and supply-chain relationships. They must keep up with financial innovation, corporate actions, and micro- and macro-level events occurring continually among thousands of entities around the world. New regulations will expose additional data for analysis. All of this must be analyzed and distilled to measures of systemic risk. This section briefly discusses the associated data management research challenges, including issues of knowledge representation, data integration, information quality, metadata and change management, data presentation, security, privacy and trust.

3.1 Data Representation and Complex Models

The notion of *risk* is central in finance, and it poses some fundamental questions. Where is risk intrinsically? Does it pertain to an analytical model or to the instrument itself? How is systemic

risk best represented? Risk often varies with time, and its evaluation typically is based on multiple sources of time-varying data. How should one best ensure the timeliness of risk evaluation and also indicate staleness of data sources, given that there are many of these?

Many data sources will have risk(s) associated with them. As one merges multiple data sources and computes derived information, these notions of risk are propagated. For example, a bank may track a measure of default risk for each mortgage it writes. These individual risk numbers are informative, but are even more valuable when aggregated to provide a default risk metric for the full portfolio. Risk measures may also be combined across risk dimensions (e.g., borrower creditworthiness and geographic concentration) to provide a richer risk picture. Unfortunately, aggregating risk measures is usually more difficult than simply adding them up – considerable additional information is needed. The challenge is to derive minimal risk representations with enough information for downstream derivation. The difficulty is that what information is needed may depend on the risk models used, which in turn may depend on what data are available, creating a chicken-and-egg problem.

Putting risk aside, financial information systems must have an adequate representation for many complex entities. For example, formulae themselves (e.g., a rule for calculating payouts under particular contingencies) should often be treated as data. It is straightforward to treat a formula as a simple string of text characters. Ideally, this formula-as-data would have more sophisticated handling, so that parts of formulae can be recognized, queried, and even computed. Actually manipulating formulae would be useful, but this may require more self-modification than most current database systems allow.

Another complex entity of interest is the accounting system. Even formal financial reporting rules frequently allow significant discretion in how positions and activities are treated, leading to large discrepancies in reported values. For example, internal transfer pricing schemes work to report profits in a firm's lowest-tax jurisdiction. Working with these issues requires at least that the accounting system be indicated as metadata. Queries on accounting system used are likely, and metadata query facilities may be needed.

Automated reasoning with complex contracts requires that the contracts be stated in a machine manipulable form. It appears that current knowledge representation techniques may be able to get us close to where we need to be in this regard.

3.2 Data Integration

Financial information management and data sharing confront the standard problems of data integration one would expect, given the multiplicity and heterogeneity of data sources. Data integration has been extensively studied, with many partial solutions already in place, and much progress over the past several decades. We believe financial systems are yet another important context and motivation for this line of work.

While the OFR/FFDC may have regulatory powers to force some standardization across sources, we nonetheless expect considerable heterogeneity. For example, regulators now have broad fiat authority to require fixed tags for certain data types, but an unsettled research question is which data should be tagged. Unless there is a standard ontology providing shared definitions and semantics, comparing financial data across multiple institutions will continue to be a challenge.

Consider primary keys or identifiers (such as CUSIP codes for North American securities). Currently, there are multiple competing numbering schemes in many markets; in some cases, a single identifier might even be reused for several instruments. Other markets may have very limited identifier coverage; for example, a CDO owner in the figure above would likely have trouble identifying all of the specific mortgages underlying his security. The research need is to: (a) determine which objects should have identifiers; (b) specify techniques to track identifiers across contractual netting and novation, and across corporate mergers and separations; (c) cross-reference different identifier standards; and (d) include checksums in identifiers to catch data entry errors. A further challenge is to define a protocol for the evolution of identifiers. Financial data sharing at the instance level may be simple on the one hand because much of the data appears as numeric streams. However, without precise agreement on the definition of terms or formulae used, comparing simple numeric values or other features of the data without access to metadata may be meaningless, or introduce confusion and error. Also, the OFR cannot collect everything, so triage based on the usefulness of the data is needed.

3.3 Data Quality

There are at least three distinct reasons for poor data quality in financial systems: incompleteness or error in the source(s) of data; errors in data integration; and fraud. We deal with each in turn.

One might expect some data sources, such as trade data, to be reasonably complete. However, "trade breaks" (i.e., cancelled transactions) due to un-reconcilable discrepancies in transaction details are painfully common. Others data sources, such as company data, are naturally incomplete or subject to interpretation. Yet other data represent estimates of aggregates, such as macroeconomic data. It may be possible to characterize the incompleteness and possible error in many data sources, but it is an open question how to record and reflect this in downstream computation. Furthermore, data quality may be measured and corrected at different levels, including the application level.

Given the large number and the variety of data sources, errors in data integration are to be expected. It is likely that integration will occur on an automated, best-efforts basis, with human correction applied to fix some, but probably not all of the errors. A research issue is to characterize aspects of the integration process most likely to affect derived results, so that scarce human effort can be devoted to checking the most critical areas.

There are strong incentives for fraud in financial systems, and many individual firms currently use fraud detection software. Integrated data from multiple sources should increase the opportunities to detect fraud, through comparison and reconciliation of discrepancies between data sources. Many large-scale frauds (e.g., the Madoff and Barings scandals) have required the entry of fictitious contracts into trading systems; since every contract has at least two counterparties, a simple check for the existence of the other side of the deal could have revealed the crimes. There is also a need for an automated protocol when a problem is detected – often one may want additional proof of fraudulent activity to avoid alerting the fraudsters prematurely.

3.4 Streaming, Change Management and Performance

Many data sources (e.g., high-frequency trading) produce large volumes of data. Furthermore, in some instances, rapid response is required, the "flash crash" of May 2010 being an obvious example. Time-stamp granularity is a concern for fast moving phenomena. Streaming data techniques are likely needed.

Many data series are recorded and published right away, but this timeliness implies that many data sources will only show estimates when first made public. For example, government economic indicators are typically revised as new information is revealed, frequently with multiple restatements. When revisions are published, procedures should exist to trigger an update of derived data that were based on the original numbers.

3.5 Metadata Management: Ontologies, Open Standards and Provenance

We have discussed above the importance of noting accounting systems and model formulae. Besides these, there is a host of other relevant metadata that must be recorded adequately, and folded into derivations where needed. For example, many historical series on corporate information should be merger-adjusted, just as equity prices must be adjusted for stock splits and dividends. In addition to metadata on what is measured, it is also important to track who is performing the measurement and how they are doing it to understand the reliability of derived results. In other words, extensive provenance management is required. Banks today already use audit trails, and the technology to do this is the natural place from which to build a full-fledged provenance recording and management system.

3.6 Data Presentation

Even with all of the above technologies in place, systemic risk will not reduce to a single global number. This is equally true for many other derived results of interest. Rather, these results will at best be derived as a function of various model assumptions and inputs. In many cases, there may be no closed form derivation at all – rather all we may be able to do is to simulate under specified conditions and obtain results thereby. In other words, the decision-maker cannot be given a single number that is easy to understand. Rather, there is a range of numbers, with complex dependence on multiple factors. Under such circumstances, the manner of data presentation becomes very important. A poorly designed decision "dashboard" may be worse than having no standard presentation at all. Research is required into the most effective presentation of complex data and the results derived from them.

3.7 Security, Privacy and Trust

The need for security at the technical level and trust at the organizational level are keys to achieving the goals of the OFR. In addition to the expected challenges, the open sharing of metadata and ontologies may not always be possible due to perceived competitive advantage associated with such knowledge. Important parts of the financial industry are cloaked in secrecy. One challenge will to develop an appropriate set of property and privacy rights to delineate between public and confidential financial information. Another research issue will be the design of a physical and information security infrastructure for the OFR to maintain confidentiality where required.

4. CONCLUSIONS

This paper identifies some of the key reasons for data anarchy in the financial industry, including multiple heterogeneous silos, the data quality gap, lack of standards and the inherent complexity and uncertainty involved. In response to this and other shortcomings, the Dodd-Frank Wall Street Reform Act has created an Office of Financial Research (OFR) with a mandate to establish a sound data-management infrastructure for systemic-risk monitoring. The new OFR includes a Federal Financial Data Center to manage data for the new agency. Achieving acceptable and successful solutions for meeting risk monitoring objectives will present several information management challenges. These are briefly discussed here, and include knowledge representation, information quality, data integration, metadata management, change management, presentation, security, privacy and trust.

5. ACKNOWLEDGMENTS

A workshop on financial information management was partially supported by the National Science Foundation grant IIS1033927 and the Pew Charitable Trusts in July 2010. The material in this paper is derived from discussions at this workshop.

6. REFERENCES

[1] Bennett, M., 2010. *Enterprise Data Management Council Semantics Repository*. Internet resource (downloaded 26 Sep 2010). DOI= http://www.hypercube.co.uk/edmcouncil/.

[2] The Committee to Establish the National Institute of Finance (CE-NIF), 2009. *Data Requirements and Feasibility for Systemic Risk Oversight*. Technical report (30 Nov 2009). DOI= http://www.ce-nif.org/images/docs/ce-nif-generated/nif_datarequirementsandfeasibility_final.pdf.

[3] Duffie, D., 2010. The Failure Mechanics of Dealer Banks, *J. of Econ. Perspectives*. 24, 1 (Winter, 2010) 51-72. DOI= http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.24.1.51.

[4] Engle, R. and Weidman, S. 2010. *Technical Capabilities Necessary for Systemic Risk Regulation: Summary of a Workshop*. (Washington DC, USA, Nov. 3, 2009). National Academies Press, Washington DC. DOI= http://www.nap.edu/catalog.php?record_id=12841.

[5] Haldane, A. 2009. *Why Banks Failed the Stress Test*. Speech, Bank of England (9-10 February 2009). DOI= http://www.bankofengland.co.uk/publications/speeches/2009/speech374.pdf.

[6] Lo, Andrew W., 2009. Regulatory Reform in the Wake of the Financial Crisis of 2007–2008, *J. of Financial Econ. Policy*, 1, 1 (2009) 4-43. DOI= http://www.emeraldinsight.com/journals.htm?issn=1757-6385&volume=1&issue=1.

[7] Rowe, D., 2009. Fostering Opacity, *Risk Magazine* (July 2009). DOI= http://davidmrowe.com/riskmag.php