CI-P: Developing the Next Generation of Community Financial CyberInfrastructure for Monitoring and Modeling of Financial Eco-Systems and for Managing Systemic Risk

Louiqa Raschid University of Maryland

Keywords:
Financial cyberinfrastructure; financial information management; financial ontologies; knowledge representation; information integration; visual analytics; network analysis; human language technology; information quality metrics; Legal Entity Identifier (LEI).

The Great Recession of 2008 and the continuing reverberations around debt and deficit in the Eurozone have highlighted significant limitations in monitoring and modeling national and global financial eco-system(s). In consequence, regulators are unable to forge knowledgeable and prudent policies, analysts are uncertain of the quality of their risk estimations, researchers are stymied in their ability to model markets and to predict behavior and outcomes, and firms may experience costly trading errors due to the use of sub-optimal risk management metrics.

While there is considerable activity today in developing more sophisticated models of financial eco-systems and in developing more advanced regulatory tools, *all such work must be driven and informed by data*. Unfortunately, current financial cyberinfrastructure severely restrict the availability of data to market participants, regulators and researchers. These limitations commence with constraints on the data collection authority of regulators. They are exacerbated by the lack (or low acceptance) of ontologies and standards and protocols within the financial industry. Beyond these limitations is the inherent challenge of dealing with the complexity of financial information and meeting the diverse and sophisticated analyses required to model heterogeneous eco-systems.

Advanced computing technology can help to address many of these challenges and can be used to develop the next generation of community financial infrastructure. Two workshops that were co-organized by PI Raschid, in 2010 and 2012, brought together a diverse community of academic researchers, regulators and practitioners. They articulated the range of multi-disciplinary research challenges and highlighted the urgent need for *community financial infrastructure*. An important challenge is the need to develop computational research frameworks, models and methods, in the spirit of past efforts to identify computational grand challenges in data intensive domains including the biomedical sciences, healthcare, climate change, etc. *The next generation of community financial cyberinfrastructure must provide a platform that can transform our current patchwork of approaches to monitoring and regulating systemic risk and must provide a unifying framework to identify computational grand challenges for improved financial monitoring and risk management.*

The **intellectual merits** of developing community financial cyberinfrastructure for monitoring and modeling financial eco-systems are based upon the following elements:

- A blueprint for developing community infrastructure that builds synergy among multi-disciplinary needs and opportunities and academic disciplines.
- A detailed specification of the infrastructure including datasets, annotations, ontologies, tools, metrics, ground truth, benchmarks and use cases.
- A framework that can articulate each computational research challenge and link it to the community infrastructure resources and testbed(s) that is envisioned through this proposed effort.

The **broader impacts** of the next generation of community financial cyberinfrastructure are significant. Regulators will not be as blind-sided during future crises. There will be increasing synergy from applying computational technology, BIGDATA and Linked Data, and social media, to address difficult modeling and monitoring problems in financial eco-systems. This may result in improved tools for regulators, as well as fundamentally new designs of market mechanisms, recommendations, ratings, etc. On the educational frontier, the planned community financial cyberinfrastructure will nurture a new generation of multi-disciplinary scholars, at all levels, who will blend computational solutions with theories, models and methodologies from finance, economics, mathematics and statistics. An advisory committee of researchers from finance, economics and mathematics and representatives of the financial industry has been identified. The vision and implementation plan for community financial cyberinfrastructure will be developed by a steering committee of computational researchers and representatives from the software industry. Both committees will ensure strong community input in developing the full CRI CI proposal.

1. Introduction

Recent events including the Great Recession of 2008 and the continuing debt and deficit challenges in the Eurozone have highlighted significant limitations in modeling national and global financial eco-system(s). This includes the lack of financial cyberinfrastructure to ingest and process numerous streams of financial transactions, as well as the accompanying data streams of economic activity, in real time. Also absent are open standards and shared semantics so that this data can be used to populate models of individual markets, financial networks and the interconnected eco-systems representing national or global financial systems. The limitations have been exhaustively described in [Cerutti et al 2012; Engle and Weidman 2010; IMF and FSB Report 2010]. There is an urgent need to develop computational research frameworks, models and methods, in the spirit of computational grand challenges in data intensive domains such as the biomedical sciences, healthcare, climate change, etc. The next generation of community financial cyberinfrastructure must provide a platform that can transform our current patchwork of approaches to monitoring and regulating systemic risk. The following *grand challenge* scenarios exemplify new tools and methods for regulators to deal with cataclysmic events:

- The ability to track financial products end-to-end along their supply chain. An extreme example is the mortgage supply chain, including sub-prime mortgage products, the asset backed securities into which individual mortgages were pooled, and finally the complex derivatives that were used to hedge bets against the securities. This lack of infrastructure continues to create problems in financial markets, the US housing market, and the courts, as state attorneys general struggle with robo-signed documents and improper and potentially illegal foreclosures.
- The ability to produce a "heat map" of our financial system transactions and accompanying economic activities, very much like a global weather map, so that one can identify financial weather patterns, pinpoint areas of high activity or vulnerabilities based on topology, warfare, political uncertainty, etc.
- Models of the global financial marketplaces and their interconnections, or the multi-party network of legal entities (financial institutions) that participate in complex financial contracts, as well as the network of relationships among them. Such models will provide the capability to run large-scale simulations to understand how these systems will perform under stress. We note that federal regulators in 2008 had to make expensive and drastic policy decisions about bailouts and stimulus spending, without real-time access to such models or simulation results.
- A significant amount of human activity is captured in new media – social media and social networks, as well as in traditional media – newswire, large document collections, etc. These resources can be a proxy for financial markets and can capture many aspects of human behavior including sentiment, persuasion, etc. Such knowledge can be extracted and mined to create more sophisticated models of financial markets. We note that there have been many recent successes in combining human language technologies, machine learning and data/text mining, e.g., in computational social dynamics or socio-computing in the humanities and the social sciences.

1.1 Community Description, the Need for Community Infrastructure and Broader Impact

Two workshops were co-organized by PI Raschid, in 2010 and 2012, and brought together a diverse community of academic researchers, regulators and practitioners, from the following disciplines:

- Computer science and information science (data management and data mining; visual analytics; information retrieval; human language technologies; machine learning; knowledge representation and reasoning; semantic Web; BIGDATA).
- Finance (financial informatics, risk management, and financial engineering) and financial accounting.
- Mathematics, economics and operations research related to financial information modeling.

The consensus of the community was that there was a *significant deficit* in computational and mathematical modeling and reasoning, as well as a dearth of best practices for standards and ontologies, data sharing protocols, quality metrics, etc. Hence, all interested actors have been unable to ingest market information in a timely manner, and to determine what information might be missing. Broader impacts of the planned community financial cyberinfrastructure include the following:

- The academic community will have access to community resources required to examine and analyze actual market operations and behavior.

- Regulators, analysts, and the financial press will reach a better understanding of capital market operations to forge knowledgeable and prudent financial policy.
- Business analysts will have increased confidence in their internal risk and accounting numbers.

Further, there will be increasing synergy from applying computational technology, BIGDATA and Linked Data, and social networks and social media, to address difficult modeling and monitoring problems in financial eco-systems. This may result in improved tools for regulators to monitor financial systems as well as fundamentally new designs of market mechanisms, new ways to reach consumers, new ways to exploit the wisdom of the crowds to review and rate financial products, to make recommendations, etc.

Broader multi-disciplinary educational impacts will be discussed in a later section.

The financial industry has historically been a leader in utilizing and driving advances in computational methods, and it is one of the largest consumers and producers of BIGDATA. Nevertheless, the industry does not have a history of making appropriate datasets available as community infrastructure for research. A key reason is that information asymmetry is a prime advantage in a financial trade. 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 organizational alignment or battles over priorities. The last decade has seen the emergence of a patchwork of standards and protocols such as SWIFT (bank-to-bank message transfer) and FIX (Financial Information eXchange is a messaging standard for the real-time electronic exchange of securities transactions). These have been developed as standalone protocols for specific purposes and lack a shared semantics, e.g., a shared controlled vocabulary or ontology.

The Office of Financial Research (OFR) has a mandate under the Dodd-Frank Act of 2010 to collect all required data inputs for managing systemic risk. However, requirements to ensure the privacy and confidentiality of fully identified data, and the need to provide a continuous audit of secure access to the data, behind a firewall, naturally lead to constraints that limit the ability of the OFR to make the acquired data widely available to the public. In some cases, the OFR may even be unable to share data collected through its authority even though portions of such data may already be made available to the public through some other possibly unauthorized channel. *The community infrastructure development activities envisioned in this proposal are therefore a valuable complement to the data collection authority and activities of the OFR.* Further, a potential outcome of developing community infrastructure may be improved methods for data de-identification and protocols to allow for greater data sharing by the OFR in the future.

PI Raschid made significant efforts to include women and members of under-represented groups in the 2 prior NSF workshops and will continue these efforts. She will reach out to current doctoral students and recent graduates of the KPMG Foundation Minority Doctoral Fellowship Program in Accounting to participate in the planned infrastructure. The financial recession had a disproportionate impact on many minority communities and the Consumer Finance Protection Bureau has made special efforts to connect to these communities. Where appropriate the planned infrastructure will support these efforts, e.g., through developing tools for social media monitoring and recommendation, as will be discussed in a later section.

1.2 Vision and Architecture for Community Infrastructure Development

We focus on the challenge of managing systemic risk in this CRI-CI-P (planning) document. The vision for exploiting BIGDATA, e.g., real time streams of all financial transactions, other signals of economic activity, social networks and social media data streams, Linked Data, etc. will be explored more fully in developing the implementation plan of the full CRI-CI proposal.

Financial data for systemic risk management can be classified as follows:

- 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 legal 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 new contracts (transactions) and the accumulated financial exposure on an entity's books (positions).
- Prices and related data: Transaction prices and data used in the valuation of positions, development of models and scenarios, and the measurement of micro-prudential and macro-prudential exposures.

The vision for developing community financial cyberinfrastructure will explore multiple approaches to accommodate a diversity of requirements. One approach is to start with a *seed collection* of highly curated data objects, and to exploit public or private collections, utilizing text extraction and human language technologies, to enhance and enrich the seed dataset. A vastly different approach would apply *scalable methods* from network analysis, machine learning, information retrieval, semantic Web, Linked Data, etc., to create large interlinked and annotated collections, with varying levels of completeness and quality. There is also a significant need to apply *knowledge representation and reasoning* methods to financial contracts so yet another approach will rely on combining methods for machine readable contracts, formal logics and reasoning, etc. We briefly comment on the datasets, tools, ontologies, metrics, metadata, user cases and a variety of artifacts that comprise community financial cyberinfrastructure. Details of some exemplars are provided in a later section.

DATASETS

- Ground truth datasets a la the TDT4 that has been used for topic detection human evaluation [TDT2004]. These datasets will be used to specific metrics, determine performance baselines, etc.
- Starter or seed datasets that have been manually curated and enriched, e.g., MIDAS collection from IBM [Hernandez et al 2012] or the Hoberg SEC collection [Ball et al 2012]; details will be provided in a later section.
- Large representative collections, e.g., for sampling, de-identification, etc. There are multiple portals that can provide such collections, e.g., the SEC/EDGAR portal.

TOOLS/ONTOLOGIES/METRICS/METADATA

- The Financial Industry Business Ontology (FIBO) includes a semantic model of concepts, their relationships and abstractions, as well as an operational ontology that targets pragmatic operational implementations. For example, using a semantic reasoner, representations in W3C RDF OWL and the FIBO, one can implement an end-to-end application to extract data from a spreadsheet and to classify undifferentiated financial swaps into their real asset classes.
- Karsha: The Smith School of Business and the Lanka Software Foundation have incubated the Karsha FOSS project to develop a recommendation tool and document search engine with respect to the Financial Industry Business Ontology (FIBO) [Karsha DASS]; details will be provided in a later section.
- Metadata, namespaces and RDF schemas, quality metrics, etc. will be developed in cooperation with/in alignment with the recommendations of the Financial Stability Oversight Council (FSOC) Standing Committee on Data.

USE CASES / SIMULATION SCENARIOS / CICI and LEI/ TESTBEDS/OTHER ARTIFACTS

- The proposed Legal Entity (LEI) Identifier and its precursor the CFTC Interim Compliant Identifier (CICI) comprise an important first step in providing a standard to uniquely identify each participant and to (partially) capture relationships among participants. The CICI has been structured to satisfy ISO 17442. The 20 digit LEI code, is expected to be identical to that of the CICI for those firms, which received a CICI identifier [ISO 17442 LEI].
- Workflows around the reporting of financial trades are not well documented. They are designed with a focus on an after-the-crisis mindset. The 2008 crisis highlighted the urgency for more proactive approaches to monitoring and modeling financial eco-systems. The use case scenarios from the 2010 Workshop Report [Flood et al 2010] will be developed as a resource to identify data quality metrics and data gaps and to measure the benefit of the reported data.
- The planned infrastructure will include a variety of tools and testbeds. An exemplar agent-based simulation testbed for automated trading [Wah et al 2012] is discussed in a later section.

BEYOND SYSTEMIC RISK

The 2008 financial crisis increased the focus on systemic risk. At the same time, there is a vast eco-system of financial markets and regulatory agencies and SIFIs (systemically important financial institutions) that interacts with the consumer and businesses. Our vision of shared infrastructure will embrace some of these eco-systems.

- GSE: Privately held corporations with public purposes created by the U.S. Congress to reduce the cost of capital for certain borrowing sectors of the economy. Examples of GSEs include the Federal Home Loan Bank, Federal Home Loan Mortgage Corporation (Freddie Mac), Federal Farm Credit Bank and the Resolution Funding Corporation.
- CFPB (Consumer Financial Protection Bureau); students loans; credit card debt; housing loans; "Know Before You Owe" campaign. The CFPB was also set up by the Dodd-Frank Act of 2010, and has taken a lead in using social media to educate the public on mortgage products, credit card debt, student loans, etc.

ACCESS AND DISSEMINATION

There are several examples of community infrastructure, portals, model organism databases, etc., that have been sponsored by the NSF and the NIH. Exemplars include the UCI Machine Learning Repository [Frank and Asuncion] and WormBase [Harris et al 2010]. We will follow best practices from both the computer science and bioinformatics communities to identify a plan for access and dissemination, and data management best practices and protocols. Every effort will be made to use open standards and protocols and to make all resources available to the public.

2. Planning Process and Budget for Full Proposal

Following the example of the two very successful NSF sponsored workshops in 2010 and 2012, the community infrastructure planning process will include a series of working group meetings and a workshop to determine specifications, to develop a blueprint, and to identify an implementation plan.

The working group meetings will be held in conjunction with appropriate conferences, for e.g., a tutorial on knowledge representation and contractual reasoning will be held in conjunction with the International Semantic Web Conference 2012; this would have been an ideal venue for a working group meeting as well. Other potential venues include VLDB, ICDE, SIGIR, SIGMOD, SIGKDD, etc. The workshop will be held at the University of Maryland.

The Financial Stability Oversight Council (FSOC) has a Standing Functional Committee on Data. The committee will support a coordinated approach to information sharing and provide direction to, and request data from, the Office of Financial Research (OFR). Additionally, the committee will work with

the OFR on data standardization efforts. PI Raschid will apply due diligence to align the planned community infrastructure with the strategic objectives identified by the Standing Committee on Data.

An **advisory committee** of finance, economics and mathematics researchers has been assembled to provide domain expertise. All members of the committee have participated in one or both of the NSF workshops. Members of the Financial Research Advisory Council being formed by the OFR will be invited to join the advisory committee to provide an additional avenue for alignment of data and tool related strategic objectives of the planned community infrastructure.

A **steering committee** of computer science researchers will develop the specifications, blueprint and implementation plan for community financial cyberinfrastructure and a subset will be PIs for a collaborative CRI CI proposal that will target an October 2013 submission. **Industry partners** have also been identified to participate in the planning process and their role in developing the infrastructure will be determined during this process.

Given the large number of potential participants in the planning process, a selected subset of letters of interest have been included in the proposal. PI Raschid invited researchers from the OFR to participate in the advisory and steering committees and requested a letter of interest to be included in the proposal. She was advised by OFR Legal Counsel that both participation and endorsement by the OFR in a proposal to another government agency was disallowed.

PI and coPIs

Louiqa Raschid	University of Maryland	PI; Data management.
Amol Deshpande	University of Maryland	co-PI; Data management.
Hal Daume	University of Maryland	co-PI; Human language technologies.
Doug Oard	University of Maryland	co-PI; Human language technologies.
Amitabh Varshney	University of Maryland	co-PI; Scientific visualization.

Advisory Committee

- Lewis Alexander, Chief U.S. Economist, Nomura. Formerly Counselor to the Secretary of the Treasury.
- Richard Anderson, Economist, Federal Reserve Bank of St. Louis.
- Mike Atkin, CEO, Enterprise Data Management Council.
- Andrei Kirilenko, Professor of the Practice of Finance at the Sloan School of Management, Massachusetts Institute of Technology. Formerly Chief Economist, CFTC.
- John Bottega, Chief Data Officer, Bank of America.
- Michael Bennett, Head of Semantic Technologies, Enterprise Data Management Council.
- Albert "Pete" Kyle, Charles E. Smith Professor of Finance at the Smith School of Business, University of Maryland.
- Joe Langsam, former Managing Director, Morgan Stanley.
- Andrew Lo, Charles E. and Susan T. Harris Professor at the Sloan School of Management, Massachusetts Institute of Technology.
- David Newman, Vice President for Enterprise Architecture, Wells Fargo.
- Chester Spatt, Pamela R. and Kenneth B. Dunn Professor at the Tepper School of Business, Carnegie Mellon University.
- Nancy Wallace, Lislie and Roslyn Payne Professor at the Haas School of Business, University of California, Berkeley.

Steering Committee

Elisa Bertino	Purdue University	Data management; cybersecurity.
Andrea Cali	University College of London	KR; formal reasoning.
Michael Franklin	University of California Berkeley	BIGDATA; data management.
Juliana Freire	NYU	Data management; provenance.

Johannes Gehrke	Cornell	Data management.
Lise Getoor	University of Maryland	Machine learning.
Georg Gottlob	Oxford University	KR; formal reasoning.
Gerard Hoberg	University of Maryland	Finance
Eduardo Hovy	CMU	Human language technologies
Vagelis Hristidis	University of California Riverside	Data management; social media.
H.V. Jagadish	University of Michigan	BIGDATA; data management.
Brad Malin	Vanderbilt University	Bioinformatics; privacy.
Philip Resnik	University of Maryland	Human language technologies.
Ben Shneiderman	University of Maryland	Visual analytics.
Michael Wellman	University of Michigan	AI; agent based modeling.

Industry Partners
IBM
Yahoo!
Enterprise Data Management Council
Wells Fargo (not confirmed)
Bank of America (not confirmed)
Note: Some industry partners have expressed an interest but have not committed any resources until we can provide more details about the implementation plan of the CRI full proposal.

Budget for Full Proposal and Long Term Maintenance

Initial funding for the planning process and funds for developing community financial cyberinfrastructure over a two year period will be requested from the NSF CRI-CI-P (this proposal) and the CRI-CI program (full proposal). The budget for the CRI-CI proposal is expected to be in the range of $2 Million dollars. It is expected that the OFR will provide some public access to additional data and software resources during this period. Industry sponsors including IBM and a SIFI Ithat is regarded as leading the financial industry with data management best practices) have agreed to provide a range of computational resources and expertise.

Over the long term, individual researchers and collaborative teams will incorporate these resources into their research agenda, thus providing a path for long-term support and maintenance of the community financial cybinfrastructure, as well as ensuring the development of tools, metrics and use cases. There has already been some progress along these lines. The Sloan Foundation recently announced a research program around the Legal Entity Identifier and have invited a team (Raschid, Langsam, Jagadish and Kyle) to submit a proposal (based on a pre-proposal). We expect several such funded efforts to communicate towards community infrastructure development.

3. From Individual Resources to Community Infrastructure

3.1 Big Picture

There will never be entirely clean, accurate, complete, and timely data for monitoring and modeling financial eco-systems and management systemic risk. The picture has improved recently, and there is more data now than before, including new standards such as the CFTC Interim Compliant Identifier (CICI), and the Legal Entity Identifier (LEI), for identifying participants (counterparties) to contracts. Complete end-to-end provenance is probably never going to be available given the complexity of financial contracts, and the potential need to track some contracts over decades. In many cases, analysts will have to deal with aggregated, anonymized data. Data will continue to have missing pieces and lack of provenance. Given this situation, the financial analytics community should strive to get better, more complete data but should also develop capabilities to deal with partial, less pristine data. Confidence levels and data quality and uncertainty metrics need to be developed and then evaluated through analyses and simulations. In this way, data in a range from complete and accurate to varying levels of incompleteness and uncertainty can be handled within the same framework. Some insights from the 2010 and 2012 workshops are as follows:

- More robust predictive analytics approaches and processes must be developed—ones that take into account uncertainty and confidence ranges, among other things.
- Hidden networks: Because of incomplete data, parts of the network that should be linked are not, or links may be uncertain (e.g., who are the leaders and who are the followers in a social network?).
- Improved analytical approaches will pay unexpected dividends. For example, statistical analyses can reveal correlations showing a wider network of who is at risk, even without direct connections.
- Metadata, provenance trails, information quality metrics and assessment protocols will play a key role in determining data quality gaps as well as the cost/benefit of financial cyberinfrastructure.

In this section, we first discuss some key aspects of computational research challenges. We then provide exemplars of community financial cyberinfrastructure.

3.2 Computational Research Challenges

3.2.1 Topological Descriptive Analysis

As a starting point of working with network- and graph-type data, it would be useful to employ topological analysis and related techniques commonly used in other applications, such as social networks or large scale electrical grid networks. These techniques are relatively well-developed—for example, methods such as topology of nodes and links, degree distributions, k-cores and centrality measures can be applied. However, extra emphasis should be put on link structures and identifying the meaning of links observed. Links should be appropriately weighted by their importance in the application. In addition, highly scalable approaches can be developed by coupling topological analysis with clustering methods (so that topologically coherent networks with substructures abstracted out can be created, for example).

3.2.2 Evolution and Temporal Analysis

Time needs to be incorporated as a first-order concept around which efficient and effective temporal structures can be created. This is a general need across financial analytics. There has been work on time-dependence, but this work has focused on shorter time periods or simpler, less detailed time structure. Now much more detailed temporal data are available (for example, large collections of financial transaction data that stretch over years or social media data, such as Twitter, that can be collected minute by minute). Other fields, such as GIS, have realized that temporal structure needs to be addressed as a research challenge. It is worth looking at events and event structures as a general, scalable approach to providing temporal structuring over a range of time scales.

In financial analysis, evolving structures can be a key to developing understanding and response. Transactional or social networks can evolve topologically, for example. In addition, nodes may change character over time and changes having to do with the strength or presence of links may occur. Emerging risk is a main area of interest for financial institutions and for government agencies. This can involve longer term trends that are only apparent when there is enough data over time, or it can involve events that affect underlying assumptions about the stability of financial instruments or transactions. In the worst case, these events may lead to unexpected cascading effects. Event and other temporal signatures can be made hierarchical, which makes the temporal structure scalable and also supports efficient, meaningful drill-down.

Evolving structures, relations, and trends in numerical variables are important. This is true for evolving networks but also for other types of data (e.g., other data associated with emerging risk).

With respect to evolving structure and in other ways, temporal analysis and temporal structuring is important. There has not been much done on the latter.

Modeling the evolution over time of organizations and contracts and interconnections within a global financial marketplace is important. Large scale event triggered simulations can be done to understand how these systems will perform under stress.

3.2.3 Dimension Reduction and Scalability

Inevitably, financial networks, social networks, and other types of relevant data are multi-dimensional. They can even be high dimensional involving hundreds or more dimensions. There could be hundreds of customer categories multiplied times many different types of financial products, for example. In addition to get a complete picture of emerging risk, it may be necessary to combine data from different sources, which will further raise dimensionality. (For example, financial data and social network data could point to the same phenomena such as housing prices and mortgage default rates in a region.) Dimension reduction methods need to be developed and applied to reduce the high dimensional space to a manageable number of dimensions for deeper analysis. Techniques such as clustering, multidimensional scaling, principal component analysis, and other can be applied. However, the key is to produce a dimension reduction that is understandable in terms of financial analysis; often-existing techniques produce mathematically transformed dimensions that are hard to understand in practical terms. Appropriate interactive visualization techniques can provide important elucidation here.

Scalability has been discussed throughout this document, but a special aspect should be mentioned here. In addition, to scalability dealing with *growing amounts* of information, there is scalability that deals with *growing complexity* of information. High dimensional spaces are an example where this second type of scalability needs to be considered. Another example would be complex processes, especially if they require more complex reasoning to understand and manage them (for example, complex and many-faceted financial transactions). Scalable techniques in this sense are important, too.

3.2.4 Large Scale Simulation

Based on what was discussed during the workshop, there appears to be a need for more comprehensive, real-world simulation approaches. According to our assumption above, there will always be an uncertainty and range of possibilities with varying confidence in the data. These aspects should be brought along in the predictive analytics and simulations that are applied. In addition, one cannot know beforehand just what situation may arise. To plan effectively for a range of possible situations, one should run an ensemble of simulations over a range of inputs. The range of uncertainties in the inputs also requires an ensemble of simulations. In addition, real world situations cannot usually be simulated using a single model. A potential financial crisis may stem from several factors and thus involve several interacting models. The general need for ensembles of simulations involving multiple interacting models is starting to be worked on in other fields. For example, severe coastal storm simulations that involve models for storm surge, hurricane winds, heavy rainfall, flooding, and people movement are being considered. In the area of the stability and resiliency of multiple interconnected, critical infrastructures

(e.g., electricity, water, gas, transportation, telecommunication), work on system of systems approaches are farther along.

Ensembles of interacting simulations (which may need to be run hundreds of times to cover a range of inputs) may be computationally daunting and expensive. It is unclear how much funding would be available for larger scale computations, so computational feasibility is an issue that needs to be addressed. But many financial simulations are substantially less complex than, say, high resolution physical simulations (e.g., storm surge models). In addition, there are system of systems approaches that concentrate on the interactions among the different models rather than on the realism of the individual models. (Critical infrastructure ensemble approaches that look for cascading effects are like this.) These can produce quite useful results without being that expensive computationally. In addition, borrowing from research in computational methods can pay off here. Often simulations can be parallelized or otherwise made significantly faster with appropriate computational approaches. The bottom line is that the simulation approach must be valid for real-world situations and must take into account the uncertainty in inputs and underlying assumptions. Otherwise the results may be misleading and, even worse, give a false sense of security.

3.2.5 Visual Analytics and Automation

The problems described above in predictive analytics, financial transactions, emerging risk, and other areas are complex and often large scale. Problems such as these require insertion of human reasoning, interpretation, and decision-making at just the right time. The analysis cannot be fully automated yet automation is needed so that the human analyst is not overwhelmed or the process made too expensive because of the need to apply large amounts of expensive human resources. Visual analytics (VA) provides a set of approaches that reserve for the human the aspects that the human is good at and for the computer those aspects at which it excels. VA then provides an interactive interface where the human and computer can work in collaboration. In a VA framework, automated techniques, often data-driven, are integrated with interactive visualization. For example, machine learning techniques that assess unstructured text or multimedia collections or even streaming content from social media or online news and blogs have been integrated into the VA framework. Various methods involving event detection, other temporal signatures, and the identification of evolving networks have been or could be integrated. Interactive visualization then provides to the analyst contextual overview, exploration, and the ability to discover interesting behavior or insights. Events or trends associated with emerging risks, for example, could be identified, given meaning, annotated, and then followed in greater detail. Predictive models or simulations could be assessed or even steered through the interactive interface. There has already been a fair amount of work done applying VA to financial analysis and this could be the basis for even deeper work. Collaborations with the VA community to more clearly describe the problems discussed in this workshop and the analytic approaches that could be used would be quite worthwhile.

3.2.6 Metadata, Quality and Provenance

There are many reasons for poor data quality in financial systems including incompleteness or error in the source(s) of data, errors in data integration, and fraud. 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. Other 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. Integration 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.

Besides the important issues of accounting systems and model formulae, 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 – 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.

3.2.7 De-identification and Data Privacy and Confidentiality

The 2010 and 2012 NSF workshops had a primary focus on knowledge representation, data management and visual analytics. During the 2012 workshop, the importance of cybersecurity, de-identification, privacy and confidentiality came up in many contexts. Confidentiality of financial data was also addressed at a workshop on data confidentiality that was organized at the Pennsylvania State University in March 2012. The steering committee for community financial cyberinfrastructure includes experts in these areas (Bertino and Malin and Jagadish). De-identification to enable the sharing of suitably aggregated data has been studied in a diversity of domains and there are well-understood solutions. Extensions include the de-identification of geo-spatial data or data obtained from mobile devices and sensors with geo-tags (Domingo-Ferrer et al 2010). Also of interest are longitudinal time-series collections. Malin is an expert on the de-identification of electronic medical records and the sharing of genomic data. He has also addressed the challenges of de-identification of longitudinal electronic medical records (Tamersoy et al 2012).

The de-identification of data from financial markets presents some special problems. The first challenge is that market strategies are typically built upon the ability to exploit information asymmetry whereas the de-identification of data is to support greater sharing, which can in turn lead to a decrease in information asymmetry. Another challenge is that several markets are concentrated with most of the trades occurring among a small number of financial institutions. In these situations, the intuitive meaning of de-identification, as well as the level of disaggregated information sharing, may have to be addressed. For example, one may need to partition the trades of a single institution to create a random number of participants in the market. There are also special challenges introduced by the nature of the data. The longitudinal (temporal) data records would typically represent events, e.g., trades associated with a specific financial contract. Each financial contract would be associated with 2 or more counterparties. Hence there are longitudinal records that are overlaid over a network of participating financial institutions. The participating counterparties themselves may change over time. De-identification would therefore have to consider the risks of disclosure of private data around a financial contract and the private data around the participating financial institutions. It would further have to consider the risk of disclosure of some private data that described the combination of the contract and the counterparties.

3.3 Exemplars of Community Infrastructure

3.3.1 Knowledge Extraction and Network Creation using Midas (IBM Research)

There is a significant amount of unstructured document content around publicly traded companies such as company filings made with regulatory agencies and news data sources. While this information is of crucial interest to regulators, investors, financial analysts and bankers, accessing the wealth of structured entity and relationship information buried in unstructured text is a non-trivial task. The Midas project at IBM Research addresses this problem by creating comprehensive views of publicly traded companies and related entities (people and companies) based on analysis of public data.

As an example, [Burdick et. al., 2011, Hernandez et al 2012] describes how by analyzing regulatory filings, a variety of counterparty relationships (e.g., lending, investment, ownership and insider) are built

across financial entities integrated across SEC and FDIC filings (annual reports, proxy statements, current reports, insider reports and FDIC Call Reports). A major step towards providing such insights is the aggregation of fine-grained data or facts from hundreds of thousands of documents into a set of clean, unified entities (e.g., companies, key people, loans, securities) and their relationships. They start from a document-centric archive, as provided by the SEC and FDIC, and build a concept-centric repository for the financial domain that enables sophisticated structured analysis. By focusing on high-quality financial data sources and by combining three complementary technology components – information extraction, information integration, and scalable infrastructure – Midas can provide valuable insights about financial institutions either at the whole system level (i.e., systemic analysis) or at the individual company level.

For instance, co-lending relationships extracted and aggregated from SEC text filings can be used to construct a network of major financial institutions. Centrality computations on this network enable the identification of critical hub banks for monitoring systemic risk. Financial analysts or regulators can further drill down into individual companies and visualize aggregated financial data as well as relationships with other companies or people. For example, centrality computation shows that a few major banks (J. P. Morgan Chase & Co, Citigroup Inc, Bank of America) are critical hubs in the network, as they have high connectivity to all the important components in the network. Hence, their systemic risk is high. While the results are intuitively as expected, they demonstrate that a data-driven analysis can lead to accurate results even by employing a few key relationships (in this case, just co-lending).

The second type of application is the drill-down inside the individual aggregated entities. For example, if Citigroup is identified as a critical hub in the global network, regulators may wish to drill down into the various aspects related to Citigroup, as follows:

- The list of key executives or insiders (either officers or directors), with their full employment history (including the movement across companies).
- The transactions (e.g., stock buys or sells) that insiders make, and the general trends of such insider transactions. As an example, having more buys than sells in a year may indicate either a strong company or simply that the market is at a low point.
- The relationships (of a given company) to other companies; this includes identifying subsidiaries of a company, institutional holdings in other companies, potential competitors, etc.

Midas provides tools and algorithms for the various unstructured analytic stages (e.g., text analytics, entity resolution and integration, and relationship identification) involved in building the entity and relationship views from multiple unstructured data sources. These analytics can be used to maintain the entity and relationship views on a continuous basis in a scalable manner (using the Hadoop infrastructure); the resulting entity and relationship views can then be used in conjunction with internal structured data sources, for building improved statistical models (e.g., for systemic risk analysis) or for monitoring events in a near real-time manner.

3.3.2 Language, Intent, Semantics - Modeling and Prediction from SEC Filings

Firm disclosures to the SEC EDGAR database constitute a highly informative and voluminous data repository available to researchers interested in determinants and explanations of underlying firm policies, performance, relationships, and business activities. These filings are required by law, are often filed on a periodic basis, and provide extensive detail that is largely untapped by researchers due to the time complexity of numerically and manually quantifying/coding its content, as needed for economic analysis. These filings, by nature of their being required, are both comprehensive in their coverage as well as semi-uniform regarding the issues discussed in firm disclosures. New research in this area taps this resource to understanding product market links, how securities are priced, and how firms differ in their corporate finance policies.

[Ball, Hoberg and Maksimovic 2012] presents some recent research on extracting a set of variables from the "Management's Discussion and Analysis" (MD&A) section of the 10-K filings. They utilized text

extraction software from Meta Heuristica LLC, to parse subsections, with a particular focus on the "Capitalization and Liquidity" subsection. They extract sentences concerning financial liquidity and intentions regarding capital market interactions. The approach leverages an empirical ontology to map phrases to concepts. Several variables are binary, e.g., which firm/year filing state that they may have to delay their investments, or that they are interested in issuing debt or equity.
There are many advantages to exploiting such resources and techniques. First, they obtain information for virtually all of the Compustat universe directly from firm's own disclosures. The variables have the advantage of low ambiguity due to direct textual context, and do not rely on ad-hoc aggregations of accounting variables. They can query the text for additional data regarding important related questions, akin to using a survey but without the problem of low response rates. For example, they can identify whether a firm is expressing concerns about issuing equity or debt in connection with an investment delay, or whether constraints seem to materialize following competition shocks or low demand shocks. Finally, the methodology is transparent, consistent, and reproducible.

3.3.3 Social Media Modeling and Prediction

A large number of social posts – in blogs, Twitter, LinkedIn, and so on – discuss financial matters, ranging from stock prices to macroeconomic analyses. A first challenge is to detect which posts and users share useful and relevant content. For example, in Twitter, most posts about the Apple stock use hashtags $AAPL or #AAPL. Preliminary studies [Ruiz et al 2012] have shown that only selecting Twitter posts that contain one of these two hashtags lead to a stronger correlation with the stock activity than employing more sophisticated information filtering methods to select relevant tweets. However, not much work has studied how public policies are correlated with activity in social networks, or how to select relevant posts for this purpose. Another challenge is how to link users' sentiment in social media back to policy decisions. How can we select relevant posts, and are traditional sentiment analysis tools adequate to classify public opinion? Further, how can we detect orchestrated social media activity that aims at influencing public opinion or confuse social media analysis tools? Influence pattern analysis can be used to identify natural progression of data in social media, and build classifiers accordingly. Government social media accounts can be viewed as the information sources, along with major world financial organization and institutions.

3.3.4 Assessing Information Quality in the pre-CICI and post-CICI/LEI Eras

The lack of unique and potentially immutable identifiers to represent legal entities (organizations) and financial instruments is a major impediment to information sharing and improving information quality. Addressing this issue correctly can single-handedly resolve many data quality issues around systemic risk. For example, CUSIP was developed to identify securities, but it is proprietary, and a fee-per-usage model has been developed around it. The proprietary nature of the CUSIP prevents federal agencies from sharing information that is linked to a CUSIP, leading to a major barrier to quality improvement. Post the passage of Dodd Frank, the CFTC wrote several rules around the adoption of a CICI (CFTC Interim Compliant Identifier). It is a precursor to an ISO standard - the Legal Entity Identifier (LEI).

Consider the following three scenarios/eras:
Current status: Company X (Morgan Stanley) maintains an internal database of entity identifiers and organizational hierarchies.
Short term future: CICI is widely deployed so that (public) financial contracts can be *marked up* using the CICI. Marked up means that if the same entity is a counterparty on several contracts, these contracts can be easily retrieved in response to a query against this entity.
Some future (ideal) state: LEIs are widely deployed.

Next consider the types of queries of interest to a federal regulator:
(1) A federal regulator asks Company X (Morgan Stanley) to report on its complete exposure to Company Y.
(2) A federal regulator asks Company X (Morgan Stanley) to report on its assessment of risk with respect to some position that X holds that involves an exposure to Company Y.

We must develop tools and datasets to answer the queries above as well as to address some of the following interesting research questions:
What information advantage does Company X (which has full knowledge of its inventory and positions) have over the federal regulator (which has full knowledge of the LEI database as well as confidential information reported historically by company and other institutions.
Conversely, what information advantage does the federal regulator have over Company X?

3.3.5 Karsha Annotation Recommendation and Markup Tool Using the Financial Industry Business Ontology (FIBO)

Karsha is a markup and recommendation tool to curate a repository of financial documents. Annotation can be done using the Financial Industry Business Ontology (FIBO) as well as other financial ontologies or thesauri. Raschid and colleagues are developing a sample repository comprising a collection of bond prospectus (corporate and municipal bonds) and their supplemental information. Karsha constructs a (Lucene) index over sections of the document (indexing the keywords within sentences). It uses Okapi cosine keyword based similarity [] to compare the sections (sentences) of the document with definitions for FIBO ontology terms and chooses/recommends the Top K terms. We focus on the FIBO since it provides an excellent set of definitions for each FIBO term. Karsha is already producing excellent initial results in providing Top K recommendations of FIBO terms using unsupervised methods, *without the use of training data or semi-supervised methods to tune the recommendation system.*

Potential use cases include the following:

- Rank and retrieve documents using FIBO search terms.
- Cluster documents to better understand the contents of a repository.
- Compare pairs of documents for similarities as well as gaps or dissimilarity.

Karsha can be extended to include sentence understanding so that one can answer more refined questions such as *Which of these instruments in this repository is likely to be impacted by a fluctuation of the price of crude oil futures*?

3.3.6 House Price Indices and Mortgage Valuation

The 2008 financial crisis, and the prominent role played by mortgages throughout, has emphasized the critical importance of modeling borrower default in valuing mortgages and mortgage-related securities, and has caused many investors to question both the safety of mortgage-related assets and thereliability of the ratings awarded by the rating agencies. Modeling mortgage default requires accurate estimates of both the current market value of a home and the distribution of its possible values in the future (including, at a minimum, its volatility). What makes this harder than, say, estimating the current price and volatility of a stock in the S&P 500 index is that houses trade far less frequently, and we can only directly observe their price when they trade. Otherwise, we have to rely on indirect measures of value, such as real-estate price indices.

In the United States, there are two dominant single-family residential house-price indices (HPI) used for estimating housing returns and for mortgage valuation: the repeat-sales indices of S&P Case-Shiller and of the Federal Housing Finance Agency (FHFA).The Case-Shiller family of indices includes twenty monthly metropolitan regional indices, two composite indices, and a quarterly national index that tracks an aggregate of the nine U.S. Census divisions. The FHFA family of indices provides quarterly estimates of housing prices for three hundred and eighty one metropolitan areas in the U.S. plus monthly aggregate U.S. and Census Division indices. However, despite the common econometric methodology that underlie the two indices, the S&P Case-Shiller indices and the HPI often do not agree. Despite their dominance in the U.S., no other country uses repeat-sales house-price indices, which have several significant shortcomings with regard to default modeling and mortgage pricing, as follows:
Sample size: A given house only enters the sample if it has transacted at least twice. Any house that has sold only once during the sample period will thus be excluded completely, as will all new houses.

Volatility: While changes in these indices are often used to estimate house-price volatility, this will almost always result in a significant under-estimate (certainly of the volatility of an individual house, which is what matters for pricing), because of the smoothing that goes into creating the index, combined with the fact that the index is attempting to measure returns on a somewhat diversified portfolio of real estate.
Sample selection: Since housing is heterogeneous and trading is infrequent, observed average transaction prices may be uninformative measures of actual supply and demand conditions.
Non-Constant Quality and Quantity of Housing: The houses that enter the sample will in general be of various types (e.g., two-bedroom versus three-bedroom, one- versus two-story, etc.), and will enter the sample randomly.
Local Markets: It is well known that house prices capitalize externalities created by nearby properties, by local neighborhood characteristics (such as schools and other public services), by the physical geography of their locations (such as their elevation, soil and weather characteristics), and by revitalization programs.
The underlying goal is to integrate a variety of heterogeneous data sources including land use maps, real estate purchase and rental datasets, energy consumption rates, etc. to determine new estimation methods for housing price index construction.

3.3.7 Development of a Simulation Testbed/ Strategies for Automated Trading

The rapid rise of automated trading - the use of quantitative algorithms to automate the process of buying or selling in a market - has led to the development of various speed-reliant trading strategies over the past two decades. One of the more controversial types of automated trading is high-frequency trading (HFT), characterized by large numbers of small orders in compressed periods, with positions held for extremely short durations. [Wah and Wellman 2012] studies the effect of latency arbitrage on allocative efficiency and liquidity in financial markets. They propose a simple model for latency arbitrage in which a single security is traded on two exchanges, with aggregate information available to regular traders only after some delay. The arbitrageur profits from market fragmentation by reaping the surplus when the two markets diverge due to this latency in cross-market communication. Using an agent-based approach, they simulate the interactions between high-frequency and zero-intelligence trading agents at the millisecond level, and evaluate allocative efficiency and market liquidity arising from the simulated order streams. The study indicates both detrimental effects of latency arbitrage and potential benefits of alternative market designs. Future work to be described in the community infrastructure proposal would comprise a testbed that would cover a range of options for market microstructure (including multiple exchanges, innovative designs), algorithmic trading strategies, background traders, and competitive configurations including multiple algorithmic traders. One particularly interesting avenue for research is the effect of widely available data and ubiquitous machine learning on financial market stability.

4. Education:

This project will develop infrastructure and a research framework to nurture a new generation of multi-disciplinary student scholars, at all levels, which will blend computational solutions with theories, models and methodologies from finance, economics, mathematics and statistics. 12 doctoral students in Business and Computer Science (advisees of members of the Steering Committee and Advisory Committee or the PI and co-PI) participated in the 2012 Workshop. Doctoral students will be involved in the planning process as well.

Raschid has offered a 1-credit course BMGT 499B on Next Generation Financial Cyberinfrastructure to Business School Information Systems majors in Spring 2012 and will offer the course in Spring 2013. She will also develop material for a 3-credit course and a research practicum targeting computer science undergraduate students (possibly with a double major in mathematics of economics or statistics). She also plans to develop the curriculum for a Master's degree in Financial Cyberinfrastructure. This will be a joint offering between the Computer Science department, the Smith School of Business and the applied mathematics/ scientific computing programs.

The community financial cybinfrastructure to be developed in this project will be key to these courses and will also inspire curriculum development of additional courses at the University of Maryland and other universities (via members of the Steering Committee).

Raschid used blended learning in the design of BMGT 499B. Students viewed PowerPoint presentations and videotapes of lectures on their own time. Classroom sessions reviewed that material and there were multiple guest lecturers, many of who participated via video-conferencing. A MOOC (massive open online course) will be developed as a method to familiarize students at all levels with the planned infrastructure.

PI Raschid has a strong record of mentoring over a dozen women doctoral students and post-doctoral research scientists and minority students and will continue these activities. Five of Co-PI Oard's graduate students and two of his undergraduate research students have been female. Daume currently advises three female Ph.D. students and one female undergraduate.

The University of Maryland, College Park, has an outstanding record of educating minorities at both undergraduate and graduate levels: In undergraduate degrees awarded to African Americans, the University ranks first among the U.S. News & World Report Top 25 Public Universities. In addition the University has earned high rankings in the granting degrees to Hispanic and Asian Americans at all levels with six graduate programs ranked in the top 10, and 13 in the top 20 programs nationally for awarding degrees to minorities.

PI Raschid has engaged two African American undergraduate women students in open source software projects and the PIs will actively recruit and engage minority REU students to be involved in developing community financial infrastructure by advertising at student chapter meetings of Women in Engineering Society and Society of Black Engineers.

5. Prior NSF Support

The PIs on this proposal and the members of the steering committee are the recipients of numerous NSF awards, all of which have some impact on the planned infrastructure. We focus on the most relevant awards.

Raschid was PI on IIS 1237476 Workshop on Next Generation Financial Cyberinfrastructure (2012) and IIS 1033927 Workshop on Knowledge Representation and Information Management for Financial Risk Management (2010). These two workshops created the multi-disciplinary committee that has articulated the need for community financial cyberinfrastructure and brought together the steering committee of computational researchers who will develop a blueprint and implementation plan. PI Raschid identified members of under-represented groups when developing the participant database for the two workshops. She will use this database to identify activities for the implementation plan. An example to emulate is the effort by the Consumer Financial Protection Bureau to determine the impact of predatory practices across a variety of communities.

Maryland has completed the related NSF-supported research infrastructure project on High Performance and Visualization Cluster for Research in Coupled Computational Steering and Visualization for Large Scale Applications, 2004 - 2009, $1,114,751 (NSF CISE RI grant 0403313). Co-PI Varshney was the PI on this proposal. The goal of this project was to build a high-performance computing and visualization cluster that takes advantage of the synergies afforded by coupling central processing units (CPUs), GPUs, displays, and storage. This resulted in the design, development, and deployment of the Chimera Cluster and Visualization Facility (http://chimera.umiacs.umd.edu). The Chimera infrastructure is being used to support a broad program of computing research that revolves around understanding, augmenting, and leveraging the power of heterogeneous vector computing enabled by commodity GPU co-processors. We have made significant progress in several interdisciplinary areas including high-energy physics, computational biology, computational radiology, scientific computing, and computer vision. This has resulted in participation of over 20 faculty drawn from four colleges on our campus. Over a 100 graduate and undergraduate students have received training on the cluster and today we have one of the most vibrant research and education communities on the East Coast engaged in general-purpose GPU-CPU computing. NVIDIA recognized this by recently awarding us the distinction of the NVIDIA CUDA Center of Excellence. This has resulted in our establishing several new funded research programs on CPU-GPU computing with NSF, DOE, ARL, AFOSR, DARPA, NIH, and NASA. This work has resulted in over 30 publications.

References

Adamic, L., Brunetti, C., Harris, J. and Kirilenko, A., "Trading Networks," Available at SSRN: http://dx.doi.org/10.2139/ssrn.1361184

Baader, F., I. Horrocks, and U. Sattler, 2004, "Description Logics," in: Handbook on Ontologies, S. Staab and R. Studer, eds., Springer Verlag, Berlin, pp. 3-28.

Ball, C., Hoberg, G, abd Maksimovic, V., "Redefining Financial Constraints: A Text-Based Analysis," University of Maryland Technical Report, March 2012.013

Bennett, M., 2010. "Enterprise Data Management Council Semantics Repository, " Internet resource http://www.hypercube.co.uk/edmcouncil/.

Bernstein, P., 2003, "Applying Model Management to Classical Meta Data Problems," Proceedings of the First Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, California, January 5-8, 2003.

Bernstein, P., A. Levy, and R. Pottinger, 2000, "A Vision for Management of Complex Models," Technical Report MSR-TR-2000-53, Microsoft Research, Redmond.

Borgida, A., M. Lenzerini, and R. Rosati, 2002, "Description Logics for Data Bases," in: Description Logic Handbook, F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, eds., Cambridge University Press, pp. 472-94.

Brammertz, Willi and Mendelowitz, Allan, 2010, "Regulatory Revolution: The Great Data Challenge," Risk Professional, 52-26.

Burdick, D., Hernández, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I., Vaithyanathan, S. and Das, S., "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study," IEEE Data Engineering Bullrtin, Volume 34, Number 3, pages 60-67, 2011.

Cerutti, E., Claessens, S. and McGuire, P., "Systemic Risks in Global Banking: What Can Available Data Tell Us and What More Data Are Needed?" Bank of International Settlements 376, April 2012.

Cohen-Cole, E., Kirilenko, A. and Patacchini, E., "Financial Networks and the Propagation of Systemic Risk," in Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press (forthcoming).

Committee to Establish the National Institute of Finance (CE-NIF), 2009, "Data Requirements and Feasibility for Systemic Risk Oversight," technical report, http://www.ce-nif.org/images/docs/ce-nif-generated/nif_datarequirementsandfeasibility_final.pdf.

Davis Polk, Client NewsFlash, "CFTC Begins Implementation of Mandatory Clearing of Swaps," July 30, 2012.

Demystifying Legal Entity Identifiers, http://www.dtcc.com/downloads/news/CiCi_Report.pdf

Domingo-Ferrer, J., Sramka, M. and Trujillo-Rasua, R. "Privacy-Preserving Publication of Trajectories Using Microaggregation," Proceedings of the Workshop on Security and Privacy in GIS and LBS, pages 25-33, 2010.

Engle, Robert F. and Weidman, Scott, 2010, Technical Capabilities Necessary for Regulation of Systemic Financial Risk: Summary of a Workshop, National Research Council of the National Academies, National Academies Press, Washington, DC, http://www.nap.edu/catalog.php?record_id=12841.

Farmer, J. Doyne, 2010, "Networks and Systemic Risks", Video, Institute for New Economic Thinking, Kings College, Cambridge.

Federal Register, Vol. 77, No. 9, Friday, January 13, 2012, Rules and Regulations, pp. 2136-2224

Federal Register, Vol. 77, No. 100, Wednesday, May 23, 2012, Rules and Regulations, pp 30596-30764

Federal Register, Vol 77, No. 113, Tuyesday, June 12, 2012, Rules and Regulations, pp. 35200-35239

Federal Register, Vol. 77, No 162, Tuesday, August 21, 2012, Proposed Rules, pp 50425-50443

Financial Stability Board, "Technical Features of the Legal Entity Identifier (LEI), March 7, 2012.

Flood, M., A. Kyle, and L. Raschid, 2010, "Workshop on Knowledge Representation and Information Management for Financial Risk Management," Internet resource:, http://www.nsf-fiw.umiacs.umd.edu/index.html.

Flood, M. and Mendelowitz, A. and Nichols, B., "Monitoring Financial Stability in a Complex World," in Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press (forthcoming).

Flood, M., Jagadish, H., Kyle, A., Olken, F. and Raschid, L., "Using Data for Systemic Financial Risk Management," Proceedings of the Conference on Innovations in Data Systems Research (CIDR2011), pages 144-147, 2011.

Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press (forthcoming).

Fouque, Jean-Pierre, Sun, Li-Shisen; "Systemic Risk Illustrated", Handbook on Systemic Risk, edited by Jean-Pierre Fouque and Joseph A Langsam, Cambridge University Press (forthcoming).

FpML, 2004, FpML Financial product Markup Language 4.0 Recommendation, Internet resource: http://www.fpml.org/spec/latest.php.
Frank, A. and Asuncion, A., UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
Garnier, Josselin, Papanicolaou, George, Yang, Tzu-Wei;" Diversification In Financial Networks May Increase Systemic Risk," Handbook on Systemic Risk, edited by Jean-Pierre Fouque and Joseph A Langsam, Cambridge University Press (forthcoming).
Harris, T. et al, "WormBase: A Comprehensive Resource for Nematode Research," Nucleic Acids Research, volume 38, pages 463-467, 2010.
Hernandez, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L, Stanoi, I., Vaithyanathan, S. and Das, S., "Unleashing the Power of Public Data for Financial Risk Measurement, Regulation, and Governance," IBM Technical Report, 2012.
Hunty, J., Stanton, R. and Wallace, N., 2011, "The End of Mortgage Securitization? Electronic Registration as a Threat to Bankruptcy Remoteness, " Technical Report, University of California, Berkeley, 2011.
International Standard ISO 17442, Financial Services – Legal Entity Identifier (LEI).
Jaffee, D., Stanton, R. and Wallace, N., 2011, "Energy Efficiency and Commercial Mortgage Valuation, " Technical Report, University of California, Berkeley, 2011.
Jaffee, D., Stanton, R. and Wallace, N., 2011, "Energy Factors, Leasing Structure and the market Price of Office Buildings in the U.S., " Technical Report, University of California, Berkeley, 2011.
Jagadish, H., "Data for Systemic Risk," in Fouque, J. and Langsam, J. "Systemic Risk Illustrated", Handbook on Systemic Risk, Cambridge University Press (forthcoming).
Karsha DASS. "Dcoument Annotation and Semantic Search," Internet resource: https://wiki.umiacs.umd.edu/clip/ngfci/index.php/KarshaDASS
PWC, "A Closer Look –The Dodd-Frank Wall Street Reform and Consumer Protection Act; Impact on Swap Data Reporting" June 2011.
Raschid, L., "Fiscal Policy, Governance, Citizenry and Financial Indicators: Modeling through the Lens of Social Media, University of Maryland Technical Report, May 2012.
Ruiz, E., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A.,"Correlating Financial Time Series with Micro-Blogging Activity," ACM International Conference on Web Search and Data Mining (WSDM) 2012.
Tamersoy, A., Loukides, G., Nergiz, M., Saygin, Y. and Malin, B. "Anonymization of Longitudinal Electronic Medical Records,'' IEEE Transactions on Information Technology in Biomedicine, volume 16, pages 413-423, 2012.
TDT2004 Workshop Presentations and System Description Papers. Internet resource: http://www.itl.nist.gov/iad/mig//tests/tdt/
The Financial Crisis and Information Gaps: A Report to the G-20 Finance Ministers and Central Bank Governors. Working paper by the IMF Staff and FSB Secretariat, 2009.
Wah, E. and Wellman, M., "Latency Arbitrage, Market Fragmentation and Efficiency: An Agent-based Model," University of Michigan Technical Report, October 2012.
Workshop on Data Confidentiality, March 2012. Internet resource: http://stability.psu.edu/policy-corner

Supplementary Documents: In the Supplementary Documents Section, provide a list of PIs, Co-PIs, Senior Personnel, paid Consultants, Collaborators and Postdocs to be involved in the project. This list should be numbered and include (in this order) Full name, Organization(s), and Role in the project, with each item separated by a semi-colon. Each person listed should start a new numbered line. For example:

Louiqa Raschid;	University of Maryland;	PI
Amol Deshpande;	University of Maryland;	co-PI
Hal Daume;	University of Maryland;	co-PI
Doug Oard;	University of Maryland;	co-PI
Amitabh Varshney;	University of Maryland;	co-PI

Elisa Bertino;	Purdue University;	Steering Committee (SC)
Andrea Cali;	University College of London;	SC
Michael Franklin;	University of California Berkeley;	SC
Juliana Freire;	New York University;	SC
Johannes Gehrke;	Cornell University;	SC
Lise Getoor;	University of Maryland;	SC
Georg Gottlob;	Oxford University;	SC
Gerard Hoberg;	University of Maryland;	SC
Eduardo Hovy;	Carnegie Mellon University;	SC
Vagelis Hristidis;	University of California Riverside;	SC
H.V. Jagadish;	University of Michigan;	SC
Brad Malin;	Vanderbilt University;	SC
Philip Resnik;	University of Maryland;	SC
Ben Shneiderman;	University of Maryland;	SC
Michael Wellman;	University of Michigan;	SC

4. Data Management Plan

All material that will be generated (datasets, metadata, use cases, software, ontologies, metadata) will be made available under an appropriate license that allows full access to the research community and for educational purposes. A key part of the proposal will be a detailed implementation plan that addresses methods for documentation, maintenance and dissemination of the community financial cyberinfrastructure.

There are several examples of community infrastructure, portals, model organism databases, etc., that have been sponsored by the NSF and the NIH. Exemplars include the UCI Machine Learning Repository [Frank and Asuncion] and WormBase [Harris et al 2010]. We will follow best practices from both the computer science and bioinformatics communities to identify a plan for access and dissemination, and a data management best practices protocol. Every effort will be made to use open standards and protocols and to make all resources available to the public.