Unleashing the Power of Public Data for Financial Risk Measurement, Regulation, and Governance

Mauricio A. Hernández, Howard Ho, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana R. Stanoi, Shivakumar Vaithyanathan, Sanjiv Das*

IBM Research – Almaden
{mauricio,ho,lucian,shiv}_at_almaden.ibm.com
{gkoutri,rajase,irs}_at_us.ibm.com

*Finance Department
Leavey School of Business
Santa Clara University
srdas_at_scu.edu

ABSTRACT

We present Midas, a system that uses complex data processing to extract and aggregate facts from a large collection of structured and unstructured documents into a set of unified, clean entities and relationships. Midas focuses on data for financial companies and is based on periodic filings with the U.S. Securities and Exchange Commission (SEC) and Federal Deposit Insurance Corporation (FDIC). We show that, by using data aggregated by Midas, we can provide valuable insights about financial institutions either at the whole system level or at the individual company level. To illustrate, we show how co-lending relationships that are extracted and aggregated from SEC text filings can be used to construct a network of the major financial institutions. Centrality computations on this network enable us to identify 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 (e.g., officers or directors). The key technology components that we implemented in Midas and that enable the above applications are: information extraction, entity resolution, mapping and fusion, all on top of a scalable infrastructure based on Hadoop.

1. INTRODUCTION

During the last few years, we have observed an explosion in the number and variety of public data sources that are available on the web: research papers and citations data (e.g., Cora, Citeseer, DBLP), online movie databases (e.g., IMDB), etc. While many of these sources have been used and studied in recent years by computer science papers, there are, however, other types of public data covering additional domains. Two such significant domains are the business/financial domain and the government/regulatory domain. Examples of business/financial data include company filings with regulatory bodies such as SEC and FDIC, security market (e.g., stock, fund, option) trading data, and news articles, analyst reports, etc. Examples of government data include US federal government spending data, earmarks data, congress data, census data, etc. Yet another domain of significant importance is healthcare.

Public data sources tend to be distributed over multiple web sites, and their contents vary from unstructured (or text) to semi-structured (html, XML, csv) and structured (e.g., tables). In this paper, we will focus on business data sources in the financial domain, with particular emphasis on the filings that companies are required to submit periodically to SEC and FDIC. This allows us to access high-quality (i.e., fresh and post-audit) content that is often cleaner and more complete than community-contributed data sources, e.g., Wikipedia. Nevertheless, even though highly regulated, the SEC and FDIC data still poses challenges in that a large number of filings are in text. Thus, to extract and integrate key concepts from SEC filings, information extraction technology becomes a crucial part in the overall data flow.

In this paper, we present our experience with building and applying Midas, a system that unleashes the value of information archived by SEC and FDIC, by extracting, conceptualizing, integrating, and aggregating data from semi-structured or text filings. We show that, by focusing on high-quality financial data sources and by combining three complementary technology components – information extraction, information integration, and scalable infrastructure – we can provide valuable insights about financial institutions either at the whole system level (i.e., systemic analysis) or at the individual company level. 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. In other words, we start from a document-centric archive, as provided by SEC and FDIC, and build a concept-centric repository (a "Web of Concepts" [10]) for the financial domain that enables sophisticated structured analysis.

We exhibit two types of financial applications that can be built on top of our consolidated data. First, we show how we can construct a network of the major financial institutions where the relationships are based on their aggregated lending and co-lending activities. By employing centrality computation, we show 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 show that our 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 applica-

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

tion 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. To this extent, we provide multiple aggregated views that include:

- 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 based on movement of executives, as well as companies that are related via lending/borrowing activities.

These views foster tracking senior executives, and company interrelationships, etc., that are key components of monitoring corporate governance in financial institutions.

Midas employs a number of scalable technology components to achieve the desired level of integration. All components can process large number of documents and run as map/reduce jobs on top of Hadoop. One component is in charge of information extraction from unstructured sources and is based on SystemT [14]. This component includes high-level rules (expressed in AQL, the SystemT language) to extract structured data from unstructured text. The rest of the components are in charge of the structured information integration. Essentially, these components map and merge the extracted data into a pre-defined schema (e.g., Person). An *entity resolution* component helps identify references to the same real-world entity across the multiple input documents. All these components are implemented in Jaql [3], a high-level general language that compiles data transformations as Hadoop jobs.

This paper is organized as follows. Section 2 details some of the complex analysis that Midas enables. Section 3 explains the components in the Midas integration flow and Section 4 describes the public data sources that we used. Section 5 then explains how we programmed Midas to extract and integrate data from these public data sources. We conclude in Section 6 with an outlook of other applications that can benefit from Midas technology.

2. MIDAS: THE APPLICATIONS

In this section, we discuss the types of financial applications that the data aggregated by Midas enables. We group these applications into two types (one systemic, and one at the individual company level).

2.1 Systemic Risk Analysis

"Systemic" effects have emerged as a leading concern of economic regulators in the past few years since the financial crisis began in 2007/2008. Recessionary conditions result, of course, in the failure of individual financial institutions, but systemic risk is primarily concerned with the domino effect of one financial institution's failure triggering a string of failures in other financial institutions. The growing interconnectedness of business and financial institutions has heightened the need for measures and analytics for systemic risk measurement. The literature on techniques and metrics for assessing and managing systemic risk is nascent, and several risk measures are being proposed in this domain—see [5]. The need for systemic analysis, in addition to the analysis of individual institutions, is a growing focus of risk managers and regulators.

We define "systemic analysis" as the measurement and analysis of relationships across entities with a view to understanding the impact of these relationships on the system as a whole. The failure of a major player in a market that causes the failure/weakness of other players is an example of a systemic effect, such as that experienced with the bankruptcy of Lehman Brothers on September 15, 2008.[1]

A major challenge that makes systemic analysis harder to undertake is that it requires *most* or *all* of the data in the system—if a proper analysis of system-wide effects is to be carried out, then the data must represent the entire system. Thus, high-quality information extraction and integration that spans the entire system is critical.

Current approaches to systemic risk have used data that is easily available across the system, i.e., stock return correlations data [2, 1, 5, 15]. These papers stop short of undertaking a formal network analysis.

Midas enables enhancing the current work in finance in the following major way. By using unstructured or semi-structured public data archived by SEC and FDIC, the nature of data that is available for systemic analysis is greatly expanded. For example, in the illustrative application in this paper, we use co-lending relationships to construct networks of relationships between banks, and then use network analysis to determine which banks pose the greatest risk to the financial system. No more will researchers in finance have to only rely on the few standard (and proprietary) data sets on stock prices that are in current use.

Co-lending Systemic Risk. Using the data provided in the SEC/FDIC filings, we construct a network of connections between financial firms based on their co-investment in loans made to other corporations or financial institutions. For example, if five banks made a joint loan, we obtain all pairwise relations and score each of them to be equal to an instance of co-lending by the pair. These relationships are modeled as an undirected network with the banks as nodes, and the edges are the total count of pairwise co-lending, aggregated across all loans. These relationships may be represented in a lending adjacency matrix $\mathbf{L} \equiv \{L_{ij}\}, i, j = 1 \ldots N$, where N is the total number of financial institutions. Given that the network graph is undirected, this matrix is symmetric about its diagonal, and we set the diagonal to be zero, i.e., ignore self-loops.

We define the total lending impact on the system for each bank as $x_i, i = 1 \ldots N$. The failure of any bank i will impact the lending system by the partial withdrawal of lending support for other banks as well. Any one bank's failure will directly impact the co-lending activity of all banks it is connected with, and will also indirectly impact the banks that are connected to the ones it is directly connected with. Therefore, even if a bank has very few co-lending relationships itself, it may impact the entire system if it is connected to a few major lenders. Since the matrix $\mathbf{L}$ represents the pairwise connectedness of all banks, we may write the impact of bank i on the system as the following equa-

[1]This filing was the largest bankruptcy in the history of the U.S. financial markets.

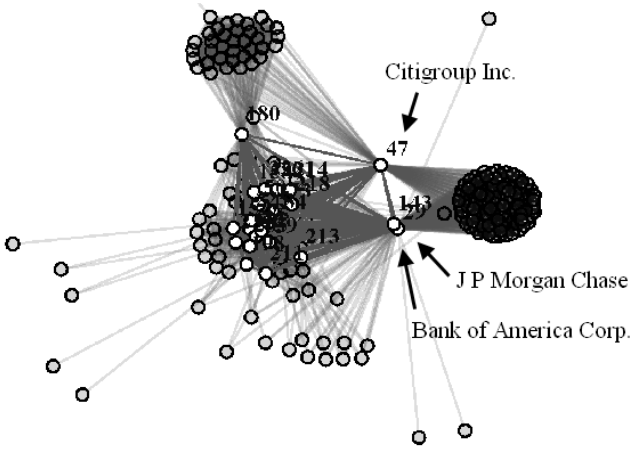

Figure 1: Co-lending network for 2005.

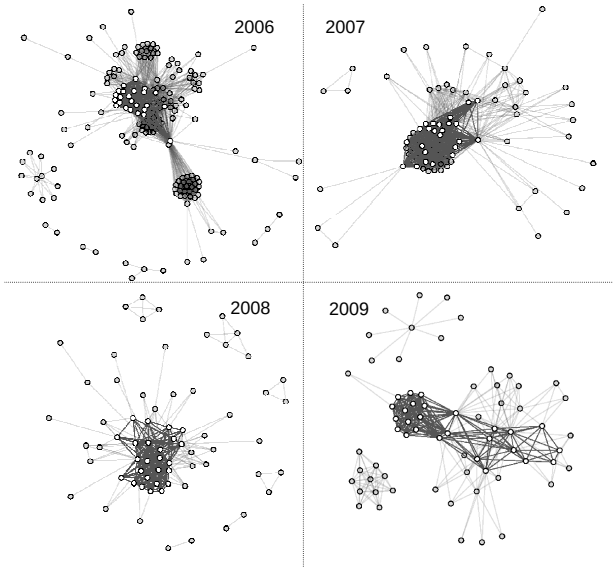

Figure 2: Co-lending networks for 2006–2009.

tion: $x_i = \sum_{j=1}^{N} L_{ij} x_j, \forall i$. This may be compactly represented as $\mathbf{x} = \mathbf{L} \cdot \mathbf{x}$, where $\mathbf{x} = [x_1, x_2, \ldots, x_N]' \in R^{N \times 1}$ and $\mathbf{L} \in R^{N \times N}$. We pre-multiply the left-hand-side of the equation above by a scalar λ to get $\lambda\ \mathbf{x} = \mathbf{L} \cdot \mathbf{x}$, i.e., an eigensystem. The principal eigenvector in this system gives the loadings of each bank on the main eigenvalue and represents the influence of each bank on the lending network. This is known as the "centrality" vector in the sociology literature [6] and delivers a measure of the systemic effect a single bank may have on the lending system. Federal regulators may use the centrality scores of all banks to rank banks in terms of their risk contribution to the entire system and determine the best allocation of supervisory attention.

The data we use comprises a sample of loans filings made by financial institutions with the SEC. Our data covers a period of five years, from 2005–2009. We look at loans between financial institutions only. Examples of included loans are 364-day bridge loans, longer term credit arrangements, Libor notes, etc. The number of loans each year is not as large as evidenced in the overnight market, and these loans are largely "co-loans", i.e., loans where several lenders jointly lend to a borrower. By examining the network of co-lenders, we may determine which ones are more critical, and we may then examine how the failure of a critical lender might damage the entire co-lending system. This offers a measure of systemic risk that is based directly on an interconnected lending mechanism, unlike indirect measures of systemic risk based on correlations of stock returns ([1]; [2]; [5]; [15]). A future extension of this analysis will look at loan amounts, whereas the current analysis is based on loan counts for which robust data is available.

After constructing the adjacency matrix representing co-lending activity, we removed all edges with weights less than 2, to eliminate banks that are minimally active in taking on lending risk with other banks. (This threshold level may be varied as required by a regulator.) We then removed all nodes that have no edges.

An example of the resulting co-lending network is presented in Figure 1 for 2005. We see that there are three large components of co-lenders, and three hub banks, with connections to the large components. There are also satellite co-lenders. In order to determine which banks in the network are most likely to contribute to systemic failure, we compute the normalized eigenvalue centrality score described previously, and report this for the top 25 banks. These are presented in Table 1. The three nodes with the highest centrality are seen to be critical hubs in the network—these are J.P. Morgan (node 143), Bank of America (node 29), and Citigroup (node 47). They are bridges between all banks, and contribute highly to systemic risk.

Figure 2 shows how the network evolves in the four years after 2005. Comparing 2006 with 2005 (Figure 1), we see that there still are disjointed large components connected by a few central nodes. From 2007 onwards, as the financial crisis begins to take hold, co-lending activity diminished markedly. Also, all high centrality banks tend to cluster into a single large giant component in the latter years.

We also compute a metric of *fragility* for the network as a whole, i.e., how quickly will the failure of any bank trigger failures across the network by expanding ripples across neighborhoods? One such metric of systemic risk is the expected degree of neighboring nodes averaged across all nodes—derived in [13], page 190, this is equal to $E(d^2)/E(d) \equiv R$, where d stands for the degree of a node. Neighborhoods are expected to expand when $R \geq 2$. We compute this for each year in our sample (Table 1). The ratio is highest just before the crisis—and then dissipates as banks take on less risk through the crisis. The diameter of the co-lending graph becomes marginally smaller as the network shrinks over time. This framework may be extended to other metrics of systemic risk to develop a systemic risk management system for regulators.

2.2 Drill-Down into Individual Entities

In this section we describe additional views that Midas provides centered around individual entities. For example, once a company such as Citigroup Inc. has been identified as a critical hub for the financial system, a regulator may want to dive deeper into various aspects that define Citigroup: its relationships with other companies (subsidiaries, competitors, investments, borrowers, etc.), its key executives (officers and directors, over the years), or aggregated financial data (loans, size of institutional investments, etc.).

For each view that we describe, we briefly mention the

Table 1: Summary statistics and the top 25 banks ordered on eigenvalue centrality for 2005.

Year	#Colending banks	#Coloans	Colending pairs	$R = E(d^2)/E(d)$	Diam.
2005	241	75	10997	137.91	5
2006	171	95	4420	172.45	5
2007	85	49	1793	73.62	4
2008	69	84	681	68.14	4
2009	69	42	598	35.35	4

(Year = 2005)

Node #	Financial Institution	Normalized Centrality
143	J P Morgan Chase & Co.	1.000
29	Bank of America Corp.	0.926
47	Citigroup Inc.	0.639
85	Deutsche Bank Ag New York Branch	0.636
225	Wachovia Bank NA	0.617
235	The Bank of New York	0.573
134	Hsbc Bank USA	0.530
39	Barclays Bank Plc	0.530
152	Keycorp	0.524
241	The Royal Bank of Scotland Plc	0.523
6	Abn Amro Bank N.V.	0.448
173	Merrill Lynch Bank USA	0.374
198	PNC Financial Services Group Inc	0.372
180	Morgan Stanley	0.362
42	Bnp Paribas	0.337
205	Royal Bank of Canada	0.289
236	The Bank of Nova Scotia	0.289
218	U.S. Bank NA	0.284
50	Calyon New York Branch	0.273
158	Lehman Brothers Bank Fsb	0.270
213	Sumitomo Mitsui Banking	0.236
214	Suntrust Banks Inc	0.232
221	UBS Loan Finance Llc	0.221
211	State Street Corp	0.210
228	Wells Fargo Bank NA	0.198

type of source documents from where the data is aggregated. The actual details and challenges regarding the various analysis stages will be described in subsequent sections.

2.2.1 Company Relationships

Figure 3 shows Citigroup's relationships with other companies through investment, lending and ownership relationships. For each relationship type, we show up to five representative companies, and also indicate the total count of related companies. The relationship types are:

- **Banking subsidiaries** : Citigroup has four banking subsidiaries registered with the FDIC. This information was obtained by integrating data from SEC and FDIC.
- **Subsidiaries** : An exhaustive list of Citigroup's global subsidiaries, as reported in their latest annual report (typically in text or html format).
- **5% Beneficial Ownership** : Securities in which Citigroup has more than 5% ownership based on analysis of SC-13D and SC-13G text filings made by Citigroup and its subsidiaries.
- **Overlapping board members/officers** : Key officer and board membership information is extracted from annual reports, proxy statements, current reports and insider transactions (text, html and xml formats).
- **Institutional Holdings** : Securities in which Citigroup has invested more than $10 million based on analysis of 13F text filings.

While the company relationship graph provides a birds-eye view of Citigroup's key relationships, additional details on individual relationships are available as described next.

2.2.2 Insider Analysis

Understanding management structure of companies and relationships across companies through common officers and board of directors is relevant in firm dynamics and corporate governance. Connected firms appear to end up merging

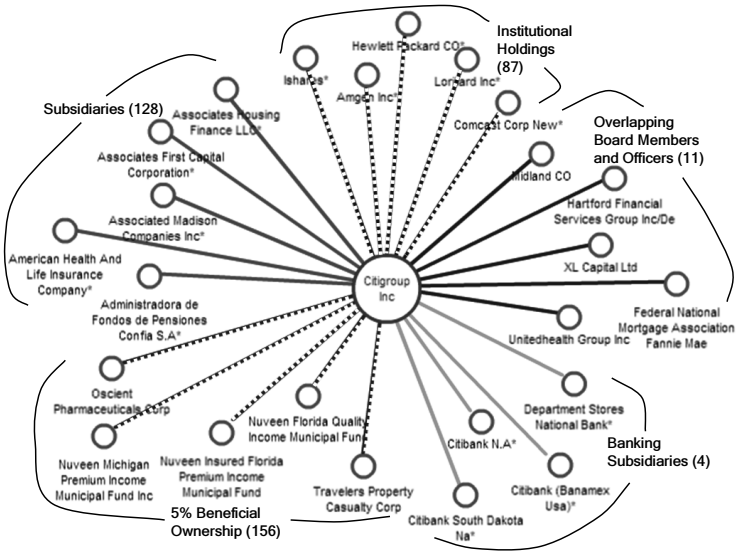

Figure 3: Companies related to Citigroup.

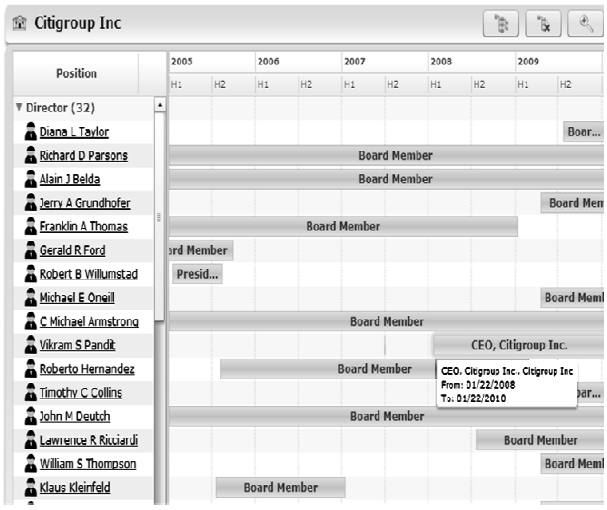

Figure 4: Key people for Citigroup.

more [7]. Understanding post-merger management structures based on earlier connections between the managers of the merged firms is also being studied [12]. To enable such analysis, Midas exposes detailed employment history and trading information for insiders (i.e., key officers and directors) of individual companies.

Employment History: Figure 4 shows some of the key officers and directors associated with Citigroup over the last several years. For each related key person, the various positions (s)he held in Citigroup along with the corresponding time periods are displayed in the figure. This profile is built by aggregating data from individual employment records present in annual reports, proxy statements, current reports and insider reports.

Insider Holdings: Figure 5 shows the current holdings of Citigroup securities (stocks and options) by the company's insiders. Each stacked bar represents the security holdings for an officer or director of Citigroup, broken down by type of holding. We show common stock, derivatives and other securities separately, with common stock further classified by whether ownership is direct or indirect (through trusts,

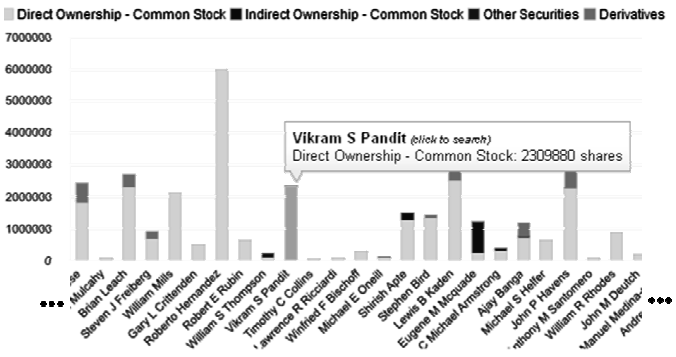

Figure 5: Insider Holdings for Citigroup.

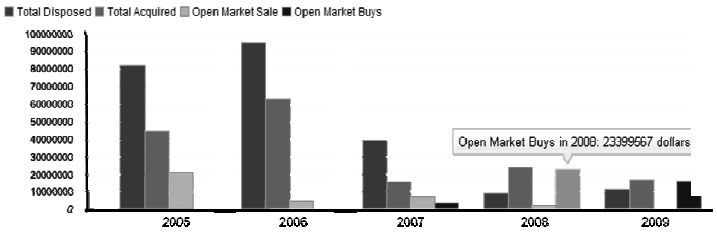

Figure 6: Insider transactions trend for Citigroup.

401K or family members).

Insider Transactions: Figure 6 presents a summary of insider transactions (buys and sells) of Citigroup securities from 2005-2009. A further breakdown of open market transactions compared with total transactions is provided. In general an open market purchase is a stronger indication of an insider's confidence. Observe that while in 2005 and 2006 there were a lot of sells of stock, in 2008 and 2009 there are not only more buys than sells, but the purchases are mostly on the open market, a very strong indication of confidence. This year so far there are more sells then buys, indicating that the trend has again reversed.

2.2.3 *Lending Exposure Analysis*

Figure 7 (top) shows a list of recent loans issued by Citigroup, either directly or through its subsidiaries. For each loan, the chart shows Citigroup's commitments to various borrowers, as compared to other co-lenders. This information has been extracted from the SEC filings made by the borrowers, where the loan documents were filed as part of their annual and current reports.

For any particular loan, additional details on the commitments made by all the lenders involved in that loan are displayed in the lower part of the figure. In this example, it shows details of an 800 million dollar loan to Charles Schwab corporation made jointly by 12 banks, including Citibank National Association, a subsidiary of Citigroup.

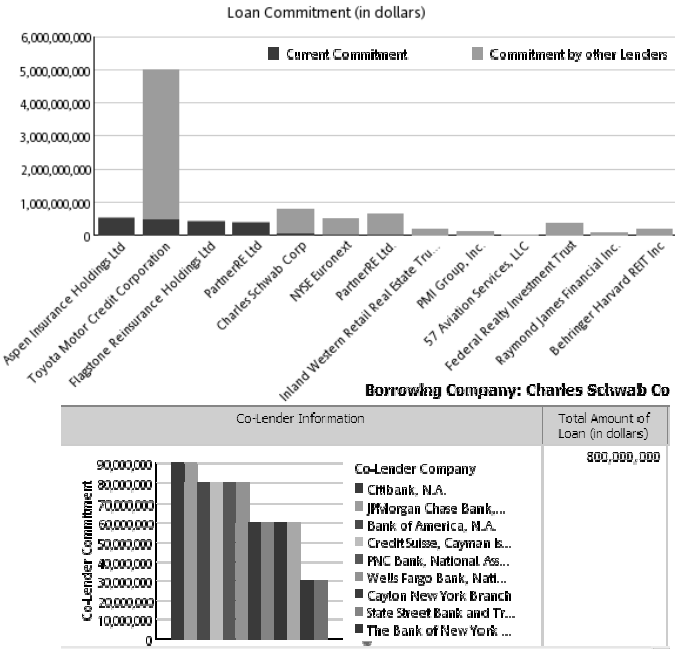

Figure 7: Lending activity for Citigroup.

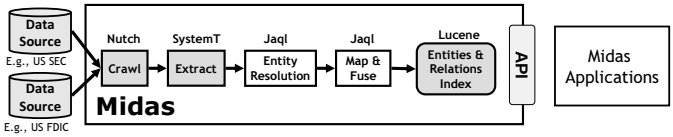

Figure 8: The Midas Data Flow

3. MIDAS OVERVIEW

We now give an overview of Midas, our system for extracting and integrating information from heterogeneous data sources. Figure 8 shows, at a high-level, the Midas data flow. Midas can take as input data from multiple sources and represented in different data formats. As output, Midas produces sets of integrated and cleansed objects and relationships between those objects which are then used by applications like the ones described in the previous section.

Input data sources can be large (Peta-bytes of information) with new incremental updates arriving daily. All operators in the Midas data flow must be capable to process large amounts of data efficiently and should scale well with increasing data sizes. To address these challenges, Midas operators are designed to run on top of Hadoop and are compiled into sequences of map/reduce jobs. For instance, the **Crawl** operator uses Nutch to retrieve input data documents. Nutch jobs are compiled into Hadoop jobs and executed in parallel. The **Extract** operator use SystemT [14] to annotate each document retrieved by **Crawl**. This operator is trivially parallelizable with Hadoop. However, the other operators (**Entity Resolution**, **Map & Fuse**) require complex data transformation whose parallel and distributed execution plan might not be trivial. To address this challenge, all instances of these operators are currently implemented using Jaql [3], a general-purpose language for data transformations. Jaql uses JSON as its data model and features a compiler that creates efficient map/reduce (Hadoop) jobs. Jaql runs the compiled jobs directly on our Hadoop cluster. Moreover, Jaql is implemented in Java and allowing many customizable extensions to be implemented in Java (e.g., user-defined functions) and seamlessly used at runtime. The Midas architecture is inspired, in part, by our Content Analytics Platform [4].

Crawl is in charge of retrieving data directly from public data sources and storing it in our local file system. Instances of **Crawl** are implemented using Nutch, a widely used open-

source crawler(http://nutch.apache.org/). To improve performance, we run Nutch as Hadoop jobs and parallelize the fetching of documents.

Extract is in charge of annotating unstructured data. Here, we leverage a large library of previously existing information extraction modules (annotators) implemented on top of SystemT [8]. SystemT is a rule-based information extraction system developed at IBM Research that makes information extraction orders of magnitude more scalable and easy to use. The system is built around AQL, a declarative rule language with a familiar SQL-like syntax. Rule developers focus on what to extract, with SystemT's cost-based optimizer determining the most efficient execution plan for the annotator. SystemT can deliver an order of magnitude higher annotation throughput compared to a state-of-the-art grammar-based IE system [8] and high-quality annotators can be built for individual domains that deliver accuracy matching or outperforming the best published results [9]. AQL rules are applied to each input document and produce a stream of annotated objects. For example, if we apply name extraction rules to the input data, we obtain structured objects that contain: 1) the raw text of the document and 2) the list of names extracted from the raw text (plus some meta-data such as the text location of each name).

Entity Resolution identifies and links annotated objects that correspond to the same real-world entity. Typically, the data required to build a single entity (e.g., a company) appears fragmented across several documents and spread over time. Recognizing that separate mentions refer to the same entity requires complex and domain-dependent analysis in which exact matching of values may not work. For instance, names of companies and people may not appear spelled the same in all documents and the documents might not explicitly contain a key to identify the company or person. Entity Resolution, which appears in the literature under other names (Record Linkage, Record Matching, Merge/Purge, De-duplication) [11], is often solved with methods that score fuzzy matches between two or more candidate records and use statistical weights to determine when these records indeed represent the same entity. Other methods explicitly use rules to express when two or more candidate records match. Our current implementation of Midas uses this latter approach and we implemented the matching rules in Jaql.

Map & Fuse transforms annotated (and possibly linked) data into a set of objects and relationships between those objects. All necessary queries to join and map the source data into the expected target schema(s) are implemented on top of this operator. The resulting queries, which are currently implemented in Jaql, must group, aggregate, and merge data into the proper, potentially nested, output schema. Since data is collected from multiple sources, duplicated values for certain fields are inevitable and must be dealt with in this stage. This data fusion step determines which of these multiple values survives and becomes the final value for the attribute. In certain cases, the data values must be merged into one consistent new value. For example, when the input set of values for a particular attribute represent time periods, we might need to compute the enclosing time period from all the valid time periods in the input set.

4. PUBLIC DATA SOURCES

Our financial application uses documents from two government data sources: the US Securities and Exchange Commission (SEC) and the US Federal Deposit Insurance Corporation (FDIC). The SEC regulates all security transactions in the US and the FDIC regulates banking institutions.

4.1 The SEC data

Public companies in the US (and key people related to these companies) are required to regularly report certain transactions with the SEC. The SEC maintains a repository of these *fillings*, organized by year and company[2]. Depending on the kind of transaction reported, public entities use different *forms* to report these regulated transactions. In some cases, forms are XML documents and, thus, contain some structured data items. In many other cases forms are filed as raw English text or as HTML documents. The SEC electronic repository contains filings going back to 1993 and currently contains over 9,000,000 filings covering about 17,000 companies and about 250,000 individual[3]. New filings are added daily and all data in the repository can be accessed via ftp.

There are many kinds of forms filed with the SEC[4] but we are only interested in those about the financial health of companies, insider transactions, and investments. We now describe the forms we used in our analysis to give a flavor of the data heterogeneity challenges we faced.

Insider Transactions (Forms 3, 4, and 5). Forms 3, 4, and 5 are XML forms that report any transaction involving securities of public company and key officer, director, or any party with at least a 10% stake on the company. These reports are filed by the company itself on behalf of the insider who is often a person but can also be another company. Form 3 is used to report when an insider is granted securities related to the company, Form 4 is used to report a transaction of such securities, and Form 5 is used annually to report all current insiders. Each form contains a common header section that provides the name of the insider, its role within the company (whether it is a key officer, directory, or a 10% owner), the name of the company, and, importantly, the *cik* for both the person and the company. The *cik* (Central Index Key) is a unique identifier provided by the SEC to every person and company that files data with the SEC. Since Forms 3/4/5 provides identifying information for both companies and key people (and due to its regulatory nature are expected to be correct), we use these forms to seed and initially populate our company and key people entities.

Financial Data and Company Status (Forms DEF 14A, 10-K, 10-Q and 8-K). Detailed information about the companies is found in a number of separate fillings. Proxy statements (Form DEF 14A) contain information for shareholders about the financial health of the company and the biographies of many key officers and directors. Much of this information is also found in the company's annual report (Form 10-K). Together, Forms 10-K and DEF 14A provide detailed business and financial information about the company including key merger and acquisitions, changes of officers and directors, business directions, key financial tables (e.g., balance sheet and income statements), executive compensation, and loan agreements. Companies must also provide quarterly updates to all shareholders, which are filed using Form 10-Q. Finally, Form 8-K is used to report signifi-

[2]http://www.sec.gov/edgar/searchedgar/webusers.htm
[3]Not all these companies or person are currently active.
[4]See http://www.sec.gov/info/edgar/forms/edgform.pdf for a complete list of all forms types.

cant events occurring in the middle of quarters. These events include mergers and acquisitions, changes of key officers or directors, offerings of equity/debt, bankruptcy, and entering material definitive agreements. All these forms contain a header that identify the company filing the form (including its *cik*). The content of the report is, however, English text formatted with HTML. Some of the financial tables are now reported in XBRL (XML, see http://xbrl.org/), but this is a recent requirement and many legacy filings in the repository contain this data in HTML tables.

Institutional Investment (Forms 13F, SC 13D and SC 13G). Companies report quarterly their ownership of securities in other companies. Form 13F, the institutional investment report, states each security owed by the reporting company, including the number of shares and the kind of share, in fixed-length column table format. However, the table representation varies from filer to filer making the task of identifying the columns and values a challenge. Form SC 13D and SC 13G are used to report 5% owners of securities related to the filing company.

4.2 The FDIC data

US banking institutions are required to report their financial health to the FDIC on a quarterly basis. These reports are very structured and are filed in XBRL format. In many cases, banks are subsidiaries of the public *holding company* which reports with the SEC. That is, often the parent company of a bank reports its results with the SEC while at the same time detailed information about the bank is submitted separately with the FDIC. All data in the FDIC repository can be accessed using a published web-service[5].

5. MIDAS INTEGRATION FLOW

We now give concrete details of the Midas flow that integrates information related to financial companies. We start by discussing the process that crawls all the forms related to the financial companies. We then discuss in Section 5.2 the initial construction of a reference or core set of company and people entities from insider reports (Forms 3/4/5). Since these forms are in XML and contain structured and relatively clean data, the resulting core set of entities forms the backbone of the rest of the integration flow. In Section 5.3, we detail how further information from a myriad of unstructured forms is extracted, linked and fused to the core set of entities. The final result is a set of entities with rich relationships, including detailed employment histories of key people, lending/co-lending relationships among companies, and all the other relationships we discussed in Section 2.2.

5.1 Crawling Data

The SEC contains data about *all* public companies in the US filing since 1993. We, however, are only interested in "financial" companies. Further, to avoid having too many stale entities in our data set, we restrict our crawl to documents no more than five years old (i.e., 2005-2010). Fortunately, the SEC publishes an index of all filings in the repository that we use to decide if a document is relevant. This index, which is updated daily, contains the *cik* and name of the filing company, the type of form filed (3/4/5, 10-K, etc.), and the ftp url to the actual document.

To determine if a company is a financial company, we preprocessed a large number of 10-K reports for *all* companies filing with the SEC for a period of 2 years. On each 10-K form, companies report their "Standard Industrial Classification (SIC) Code", an industry-wide numeric classification code. Roughly, entities reporting an SIC code in the [6000-6999] range are considered financial companies[6]. Using the SIC codes, extracted a "master" list of 3,366 financial companies ciks.

Given this master cik list, a range of dates (2005-2010), and a list form type we want, we filter the daily SEC document index and identify the ftp urls we need. The list of ftp urls forms a "seed" list that is fed into Nutch for crawling. In contrast to traditional web-crawling, our target documents do not change over time. The filings are never replaced with new updated versions and, thus, Nutch does not need to revisit previously crawled pages. Moreover, the seed list contains all the documents we want to crawl and Nutch does not need to parse the crawled documents to find more links.

Crawling data from the FDIC does not require filtering by industry code since, by definition, all banks are financial institutions. The FDIC publishes a web-service that allows downloading of the current financial report of a particular bank. Our crawler is in a web-service client that regularly downloads the most recent reports for all active banks.

We currently have a repository with close to 1,000,000 SEC documents related to financial companies and 77,000 FDIC reports for active banks. The SEC imposes some limits on crawlers (e.g., we could only run the crawler overnight) and it took several months to bootstrap the system with data covering several years. We now run the SEC and FDIC crawler monthly to catch up with recent filings.

5.2 Constructing Core Entities

We now discuss the initial construction and aggregation of company and key people entities from the XML files that correspond to insider reports (Forms 3/4/5).

Extraction of records from XML forms. We use Jaql to extract (and convert to JSON) the relevant facts from XML Forms 3/4/5. Each of these facts states the relationship, as of a given reporting date, between a company and a key officer or director. The relevant attributes for the company are: the SEC key (or cik) of the company, the company name and address, the company stock symbol. The relevant attributes for the person are: the SEC key or cik, name, an attribute identifying whether the person is an officer or a director, and the title of the person (i.e., "CEO", "Executive VP", "CFO", etc) if an officer. Other important attributes include the reporting date, a document id, a list of transactions (e.g., stock buys or sells, exercise of options) that the person has executed in the reporting period, and a list of current holdings that the person has with the company.

Aggregation of company and people entities. In this step, we process all the facts that were extracted from XML forms and group them by company cik. Each group forms the skeleton for a company entity. The important attributes and relationships for a company are aggregated from the group of records with the given company cik. As an example of important attributes of a company, we aggregate the set of all officers of a company such as Citigroup Inc. This aggregation is with respect to all the forms 3/4/5 that Citigroup Inc. has filed over the five years. Additional fusion

[5]See https://cdr.ffiec.gov/public/.

[6]See http://www.sec.gov/info/edgar/siccodes.htm.

must be done so that each officer appears only once in the list. Furthermore, for each officer, we aggregate all the positions that the respective person has held with the company. As an example, a person such as Sallie Krawcheck will result in one occurrence within the list of officers of Citigroup, where this occurrence contains the list of all the positions held by Sallie Krawcheck with Citigroup (e.g., CFO, CEO of Global Wealth Management). Since positions are strings that vary across forms, normalization code is used to identify and fuse the "same" position. Finally, each position is associated with a set of dates, corresponding to all the filings that report that position. The earliest and the latest date in this set of dates is used to define the time span of the position (assuming continuous employment). The end result of this analysis is exemplified in Figure 4.

To give a quantitative feel for the above processing, there are about 400,000 facts that are aggregated. Roughly, this number corresponds to the number of forms 3/4/5 that were filed over the five-year period by all the financial companies. These 400,000 facts result in about 2,500 company entities, each with a rich structure containing officers with their position timelines (within the company), directors (with similar timelines), and also containing an aggregation of transactions and holdings (to be discussed shortly).

A separate but similar processing generates, from the same 400,000 facts, an inverted view where people are the top-level entities. We generate about 32,000 people entities, corresponding to the officers or directors that have worked for the 2,500 financial companies.Like a company, each person entity is also a complex object with nested attributes such as employment history, which spans, in general, multiple companies. For example, a person such as Sallie Krawcheck will have an employment history spanning both Citigroup Inc. (where she served as CFO and then CEO of Global Wealth Management) and Bank of America (which she joined later as President of Global Wealth and Investment Banking).

Fusion of insider transactions and holdings. The aggregation of the transaction and holding data over the collection of forms 3/4/5 requires a detailed temporal and numerical analysis. First, we need to ensure that we group together securities of the same type. In general, there are multiple types of securities (derivatives or non derivatives), types of ownership (direct or indirect), and types of transactions (acquired, disposed, granted, open market purchase, etc.). The various values for such types are reported in text and have variations (e.g., "Common Stock" vs. "Class A stock" vs. "Common shares"). In order to avoid double counting of transactions and to report only the most recent holding amount for each type, we developed normalization code for types of securities and for types of ownership. Subsequent processing summarizes, for each company entity and for each year, the total amount of transactions of certain type (e.g., open market purchases) that company insiders executed in that year. The results of such aggregation were shown earlier in Figure 6. Similar processing retains, for each person entity, the current (i.e., the most recent) holding that the person has with a given company, for each type of securities (Figure 5).

5.3 Incorporating Data from Unstructured Forms

We now discuss the processing involved in the extraction and fusion of new facts from unstructured data into the core entities. The new facts, which are extracted from either

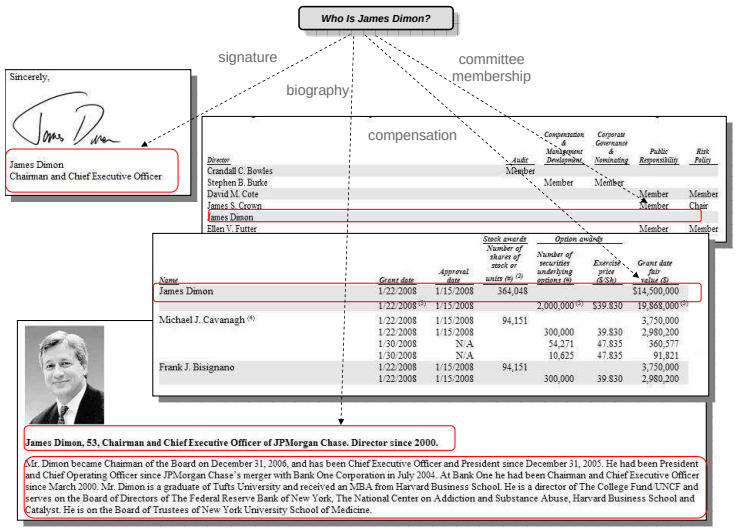

Figure 9: Employment information in various filings

text or tables, describe new attributes or relationships, and typically mention a company or a person by name without, necessarily, a key. Thus, before the new information can be fused into the existing data, entity resolution is needed to perform the linkage from the entity mentions to the actual entities in the core set.

5.3.1 *Example 1 : Enriching Employment History*

In addition to the insider reports, information about a person's association with a company is present in a wide variety of less structured filings, as illustrated in Figure 9. This information ranges from point-in-time facts (when an officer/director signs a document) to complete biographies that provide the employment history of a person. To extract and correctly fuse all the needed pieces of information, we must address several challenges.

Extract. Employment history records need to be extracted from various contexts such as biographies, signatures, job change announcements, and committee membership and compensation data. These records are typically of the form (person name, position, company name, start date, end date) for each position mentioned in the text. However, not all of the attribute values may be present or extracted successfully. For instance, the expected output from the biography in Figure 9 would include (James Dimon, Chairman, JP Morgan Chase, –, –), (James Dimon, Chief Executive Officer, JP Morgan Chase, –, –), (James Dimon, Director, JP Morgan Chase, 2000, –) and (Mr. Dimon, Chairman, unknown, "December 31, 2006", –). Using biographies as an example, we illustrate some of the challenges we encounter in extracting employment records from unstructured documents.

Identifying the beginning of a biography. Biographies typically appear in annual reports and proxy statements, as short paragraphs within very large HTML documents (100's KBs to 10s MBs) and within HTML tables, where individual employment facts may be formatted in different ways. For instance, a position with a long title may span multiple rows while the corresponding person's name may align with one of these rows, depending on the desired visual layout.

Past positions are expressed differently. For instance, a set of positions may be linked with a single organization (Chair-

man and Chief Executive Officer of JP Morgan Chase) or multiple positions may be associated with a single start date (Chief Executive Officer and President since 12/31/2005).

Anaphora resolution. Individual sentences may refer to an individual via a partial name (e.g., "Mr. Dimon") or by using pronouns (e.g., "he"). Sometime the name of a related individual may be mentioned in the biography.

Based on 10 random samples of all DEF 14A filings, our biographies annotator obtains 87% precision and 49% recall for extracting key people's names, and 91% precision and 51% recall for extracting the correct block of biographies.

Entity Resolution. As mentioned, the attributes extracted for biographies include the name of the person, the name of the filer company (also the cik, since this is associated with the filing entity) and the biography text itself. However, information in biographies does not contain a cik for the person and we need entity resolution to link each extracted biography record to a person cik.

Entity resolution is an iterative process requiring a complex and domain-dependent analysis that requires understanding the data, writing and tuning entity resolution rules, and evaluating the resulting precision (are all matches correct?) and recall (did we miss any matches and why?). In the process of matching people mentioned in biographies to the actual people entities, we faced the following challenges:

No standardization in entity names. People names come in different formats (e.g. "John A. Thain" vs. "Thain John" vs. "Mr. Thain", or "Murphy David J" vs. "Murphy David James III"). Hence, exact name matching will only find some matches and we need approximate name matching functions to resolve more biographies. On the other hand, two people with similar names (even when working for the same company) may be in fact two different people. For example, "Murphy David J" and "Murphy David James III" are two different people. *To tackle this challenge*, we designed specialized person name normalization and matching functions that cater for variations in names, suffixes such as "Jr.", 'II", and allow matching names at varying precision levels. We iterated through our data and entity resolution results several times in order to fine-tune our functions.

Achieving high precision. To improve precision beyond just the use of name matching, we observed that for a biography record, we typically know the cik of the company (since it is the filing entity). As a result, we were able to develop matching rules that exploit such contextual information. In particular, the rules narrow the scope of matching to only consider the people entities that are already known to be officers or directors of the filing company (as computed from Forms 3/4/5).

Improving recall. To improve recall, in general, one needs multiple entity resolution rules. For example, there are cases where the filer company is not in the employment history of a person (based on Forms 3/4/5). To account for such case, we had to include other, more relaxed rules that were based just on name matching. Having multiple rules, we prioritized them so that weaker matches are kept only when we do not have any matches based on stronger evidence. For instance, if we matched a "Thain John A" mentioned in a biography to both a "John A. Thain" and a "Thain John" in key people, via two different rules, we will only keep the first match since it is based on a rule that matches first/lastname *and* middlename initial.

Our initial matching rules achieved a 82.29% recall, that is, 82.29% of 23,195 biographies were matched to a person cik. At the end of the tuning process, we raised that to 97.38%. We measured precision by sampling our data, and we found it to be close to 100%.

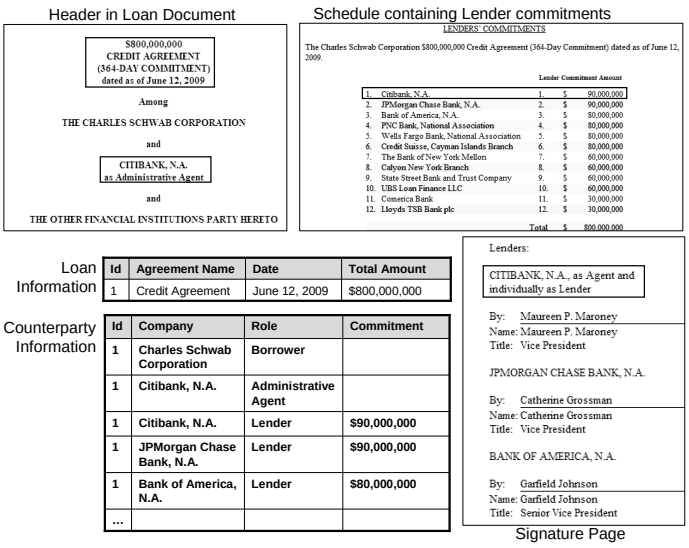

Figure 10: Loan document and extracted data

5.3.2 *Example 2 : Lending Exposure Analysis*

Figure 10 shows portions of a loan document filed by Charles Schwab Corporation with the SEC. This loan document is a complex 70 page HTML document, that contains key information about the loan such as the loan amount, date the agreement was signed, the companies involved in various capacities and the commitment made by individual lenders. As shown in the figure, this information is spread across different portions of the document such as the header at the beginning of the loan document, signature page and schedules containing lender commitments. The following analysis steps are performed on the loan data.

Extract. We first identify documents that describe loan agreements. Additional rules extract basic loan information from the header portion of these documents, which may appear either in a tabular form (as shown in this example) or as a paragraph in free-flowing text. The names and roles of the various counterparties involved in the loan are identified from three portions of the loan — header, signature and commitment table. Finally, the dollar amounts committed by individual lenders are extracted from commitment tables that typically appear in html tables in the document. Additional details about the name and role of officers who signed the loan document on behalf of different companies are also extracted. Portions of the extracted data for loan and counterparty information are shown in the figure.

Entity Resolution. Each extracted fact contains one or more company and person names, whose real-world identity needs to be resolved to facilitate aggregating facts from all loan documents. For example, for identifying lenders, we faced the following challenges.

Company name variations and subsidiaries. Company names may be written in various forms, for example, "Citibank, N.A.", and "CitiBank National Association". In addition, companies have subsidiaries; for example both "Citigroup Global Markets, Inc" and "CitiBank National Association"

are subsidiaries of Citigroup Inc. We need to be able to say when two company names refer to the same company and when one is a subsidiary of the other. To determine the unique identity of each lender, we built special normalization functions for company names and rules that compare the names of lenders with the names of all companies filing with the SEC and FDIC, and the names of all of their subsidiaries (extracted from the annual reports).

Measuring recall is another challenge because 1) we could indeed fail to resolve a company that is a lender, or 2) a company mentioned in a loan document does not file with SEC or it is not a lender. Unfortunately, in the latter case, we do not have the role of each company we extract from loan documents. We sampled 60 companies from our list of companies extracted from loan documents; 17% of them were resolved and they were all correct (i.e., achieving 100% precision); and 12.69% were not resolved but these contained errors from information extraction. Hence, our entity resolution rules are robust and do not propagate errors generated in the previous phase. 26.9% were companies that do not file with SEC hence, we do not resolve them. Finally, 42.8% were not resolved and included companies that are borrowers or institutions with no lending capacity.

Constructing the co-lending matrix. Based on the information extracted from loan documents, we were able to construct, for each year, a co-lending network where the nodes are lenders and an edge between two nodes counts the total number of loans where the two entities are co-lenders. One of the challenges in building a meaningful network is to generate a single node per company, since in the source data, a lender can appear under multiple names. For example, "Citibank" and "Citicorp USA" must be fused into the same entity ("Citigroup Inc.", which is the parent company). Entity resolution enables us to perform such identification. Once the nodes are correctly fused, subsequent processing computes the aggregated count of loans for each pair of nodes. The resulting co-lending matrix forms the basis for the systemic risk analysis described in Section 2.1.

6. OTHER BUSINESS APPLICATIONS

We conclude this paper with a description of some business applications that can exploit the consolidated public data from Midas, enhanced with more unstructured public data such as blogs, message boards, news feeds, etc.

Risk Measurement: In Section 2 we showed that financial institution systemic risk metrics may be developed from an analysis of the network of bank co-lending relationships. Measures such as centrality will help identify banks that are critical in the lending system. Community detection in lender networks will uncover groups of lenders that are critical to the system.

Generating non-return based data: Most public data is not available in structured data sets, nor is it widely available in numerical form. Text discussion on message boards, blogs, news forums, etc., can be used to uncover connectedness between firms and banks. Hence, construction of new data sets for meeting analysis or regulatory goals is an important application. Midas has already demonstrated several use cases in this domain.

Analyzing Organization Structure: Relationships between CEOs and management officers of firms are now being shown to be relevant in firm dynamics and corporate governance. Connected firms appear to end up merging more [7]. Understanding post-merger management structures based on earlier connections between the managers of the merged firms is also being studied [12].

Supporting Regulation: Large-scale data integration for decision-making and regulation is a growing field. In finance, the establishment of the National Institute of Finance (NIF) under the auspices of the Office of Financial Research (OFR), proposed in the Restoring American Financial Stability Act[7], has been tasked with setting up a systemic risk data warehouse for just this purpose. Technologies such as Midas are therefore extremely timely and may be deployed by the OFR.

Trading: Developing statistical arbitrage signals for convergence trading and high-frequency trading. This will be based on extracting signals from news feeds, blogs, message boards, and other public opinion forums.

7. REFERENCES

[1] V. Acharya, L. Pedersen, T. Philippon, and M. Richardson. Measuring Systemic Risk. SSRN: http://ssrn.com/abstract=1573171, 2010.

[2] T. Adrian and M. Brunnermeier. CoVaR. http://www.princeton.edu/∼markus/research/papers/CoVaR.pdf, Princeton University, 2009.

[3] K. Beyer and V. Ercegovac. Jaql: A Query Language for JSON, 2009. http://code.google.com/p/jaql/.

[4] K. S. Beyer, V. Ercegovac, R. Krishnamurthy, S. Raghavan, J. Rao, F. Reiss, E. J. Shekita, D. E. Simmen, S. Tata, S. Vaithyanathan, and H. Zhu. Towards a Scalable Enterprise Content Analytics Platform. *IEEE Data Eng. Bull.*, 32(1):28–35, 2009.

[5] M. Billio, M. Getmansky, A. Lo, and L. Pelizzon. Econometric Measures of Systemic Risk in the Finance and Insurance Sectors. SSRN: http://ssrn.com/abstract=1648023, 2010.

[6] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[7] Y. Cai and M. Sevilir. Board Connections and M&A Transactions. SSRN: http://ssrn.com/abstract=1491064, 2009.

[8] L. Chiticariu, R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, and S. Vaithyanathan. SystemT: An Algebraic Approach to Declarative Information Extraction. In *ACL*, 2010.

[9] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan. Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. In *EMNLP*, 2010.

[10] N. N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu. A Web of Concepts. In *PODS*, pages 1–12, 2009.

[11] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16, 2007.

[12] B.-H. Hwang and S. Kim. It pays to have friends. *Journal of Financial Economics*, 93(1):138–158, 2009.

[13] M. Jackson. *Social and Economic Networks*. Princeton University Press, NJ, 2009.

[14] R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan, and H. Zhu. SystemT: A System for Declarative Information Extraction. *SIGMOD Record*, 37(4):7–13, 2008.

[15] M. Kritzman, Y. Li, S. Page, and R. Rigobon. Principal components as a measure of systemic risk. SSRN: http://ssrn.com/abstract=1582687, 2010.

[7]See `http://www.ce-nif.org/`