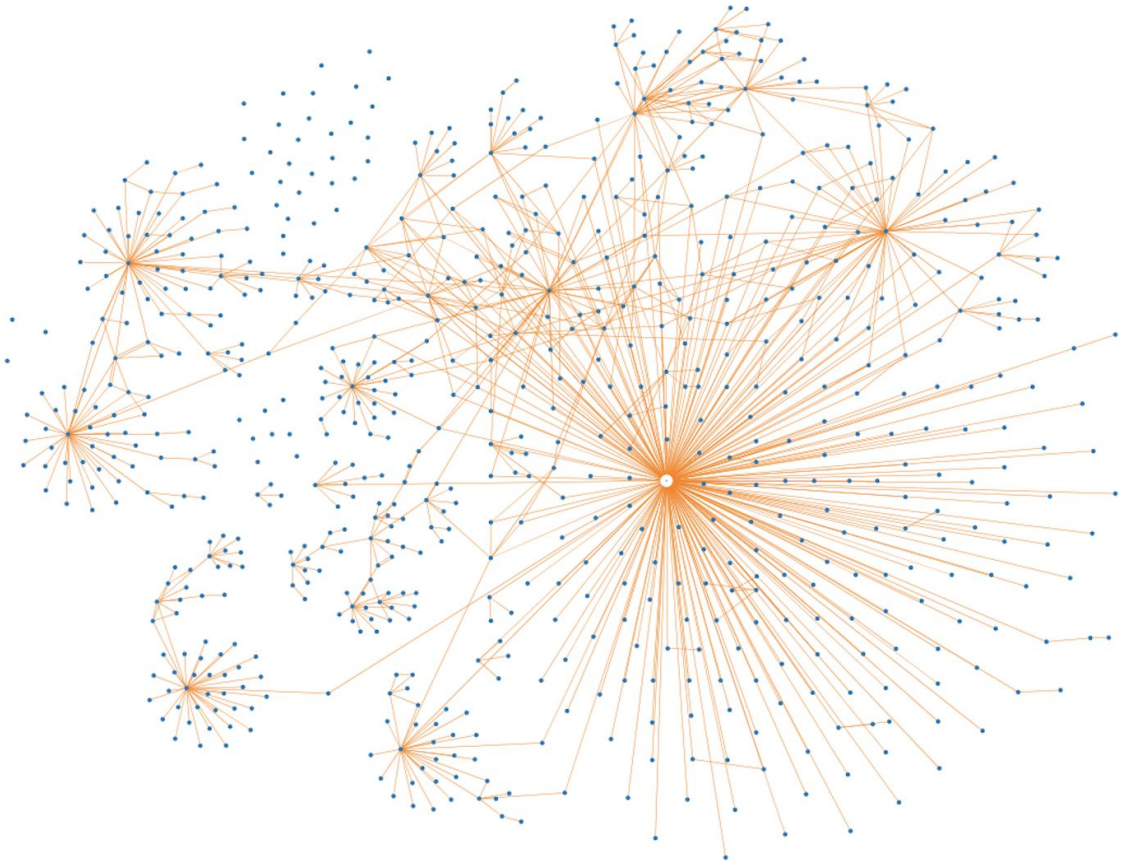
# The Integration of the Legal Entity Identifier (LEI) and National Information Center (NIC) Databases Using TigerGraph

Uttam Rao and Tommy Colitsas



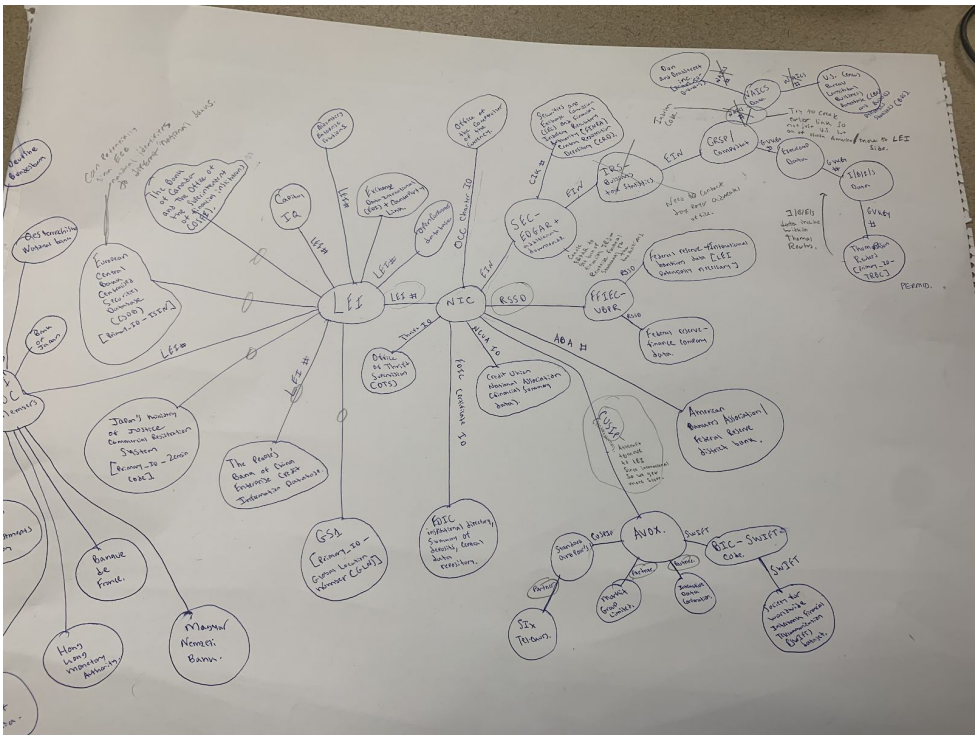
Our Goal (By: Thomas Colitsas):

When the Lehman Brothers failed in 2008, its counterparties struggled to assess their total exposure. Financial regulators were also unclear about the consequences of a Lehman failure in part because no industry-wide standards existed for identifying and linking financial data representing entities or instruments. Standards are needed to produce high-quality data. And high-quality data are essential for effective risk management for financial companies, especially to assess their connections and exposures to other firms and regulatory oversight.

To mitigate the impacts of recession and bankruptcy in large financial entities, companies and individuals require a means to understand more accurately the risks associated with these events. The means to make this happen already exists in the form of data, and the importance of high quality data is already highlighted across the industry. The good news is the data exists. Across thousands of global databases run by central banks, private institutions, and government agencies invaluable attributes can be found in relational databases. The bad news is that the data is scattered across these thousands of global databases, which makes it extremely difficult for economists to reach an understanding of how it all connects.

Our goal began rather simply, as a proof of concept. We worked to integrate the National Information Center (NIC) and Legal Entity Identifier (GLEIF) datasets. We did so using TigerGraph, a complete, distributed, parallel graph computing platform supporting web-scale data analytics in real-time. TigerGraph includes ideas such as MapReduce, Massively Parallel Processing, and fast data compression/decompression. The software delivers with speed, scalability, and deep exploration/querying capability. We worked to use TigerGraph to integrate said relational financial databases into a graph database, and in the process produce the highest granularity possible for our data.

Although a few challenges arose from a technical standpoint (as Uttam will discuss), we decided to continue with the project, and eventually expanded to see what other domestic and international financial datasets could be linked through common identifiers.




This map is a preliminary investigation of the links between some of these critical datasets. Listed below are a few key components/pitfalls of the model that should be investigated prior to continued research:

- The LEI Regulatory Oversight Committee (ROC) is an important resource. It's members consist of 71 international government organisations (along with 19 additional observers), most of whom have important datasets in respect to their economies. These members and observers have pledged to coordinate and oversee a worldwide framework of legal entity identification. It offers some strategic research advantages over the GLEIF as, again, many of its members maintain critical datasets.
- NAICS is an industry classification code and CANNOT be used as an identifier. Thus Dun and Bradstreet and LBD/BDS should be linked to more central private datasets that use other common identifiers (potentially need for pattern matching).
- The IRS Business Tax Statistics is, for our purposes, unusable. This is because even under the Freedom of Information Act constituents still are not allowed access to business tax statistics for individual financial entities, instead they are only given access to metadata/analytics. Analysis could potentially be done to recognize financial statements to tax positions to create an additional attribute for entities, but that requires advanced datasets and additional software.
- There may exist more efficient ways to link the data that improve granularity, however generating these pathways would require software analysis as some datasets are not equipped with data dictionaries and others have entries that are incomplete.

Technological Roadblocks (By: Uttam Rao):

From a technical standpoint we faced issues in loading, integrating, and updating the datasets, the majority of them being limitations of TigerGraph.

- Loading - TigerGraph only supports loading the vertices of a graph from CSV or JSON formatted files and only supports CSV for loading edges. The vast majority of databases discussed above (including the GLEIF and NIC) use XML or XBRL format. Although both the GLEIF and NIC also provided data in CSV format, a quick look at it revealed many gaps, errors, and loss of attributes that did not exist in the XML files. To get around this issue we used the Global IDs software to read the XML file into a relational database, and output them as CSV files that could then be loaded into TigerGraph. TigerGraph also does not support defining schemas loading data from anywhere other than the command line. This makes it extremely difficult to define large schemas or edit them in case of adding attributes.
- Integration - The map drawn above shows our plan to integrate the various datasets using the common identifiers between them. While during the research phase we thought this would be simple, we quickly realized that not all the datasets actually collected the information that was reported as collected. For example, even though the LEI number was one of the first listed attributes in the NIC dataset, there were LEI numbers listed for only about 10,000 entities. Although this allowed us to expand and add information to our graph, in the scope of more than a million vertices, this progress was insignificant. We explored inexact matching using regular expressions based on company name and address, but ultimately ditched the idea due to uncertainty. Even in the cases when a common identifier between two datasets existed, it was not possible to integrate the two datasets just using TigerGraph. Unlike Neo4j which allows you to combine two graphs based on a common attribute and updates the schema automatically, the datasets have to be combined in CSV format and the schema has to be redefined in TigerGraph. A whole new graph has to be created to integrate the datasets, which has to be done in the command line. We solved this problem using python and command line tools (csvkit) to combine the various CSV files into the proper format for loading.
- Updating - Both the GLEIF and NIC update their datasets twice a day. While TigerGraph does support a Kafka loader which consumes data in a Kafka cluster and automatically loads it into TigerGraph, we did not have experience with this and had to figure out a different way to update the graph. We did this by writing a simple python program which uses wget to replace the source file and reload the data into the graph. However, if an attribute is added to the dataset the program will crash as the schema will have to be changed. Looking into this problem online revealed that the same problem occurs when using a Kafka loader.

- TigerGraph vs. Neo4j - Neo4j has been around for more than a decade longer than TigerGraph and as a result has many more connectors and APIs, supports more languages than TigerGraph, and can deal with data in many different formats. Both Cypher and GSQL, Neo4j and TigerGraph's query languages, are similar and easy to learn. TigerGraph, however, is known for its speed, with Neo4j achieving nowhere near the same performance. TigerGraph is known for its fast graph traversal which allows for greater analytical capabilities. While Neo4j times out at a 4-hop query, TigerGraph is still fast at 10-hops.

Legal/Bureaucratic Roadblocks (By: Thomas Colitsas):

Although many of our issues arose in the technical hemisphere, there were also many specific legal and bureaucratic issues as well. A project of this nature requires interaction with different government and private institutions while working to gain access to datasets. Here I will briefly list a few important things to understand in relation to this topic.

- The Freedom of Information Act is an important piece of legislation to understand in attempts to gain access to higher level domestic datasets. There exist 9 major exceptions to this act that work in conjunction with the protective legislation of individual institutions (such as the IRS 26 U.S. Code § 6103 related to confidentiality and disclosure of returns and return information). Bellow I will list those 9 major exceptions and how they typically apply to financial datasets.
    - National Defense - *there may be instances in which financial regulatory agencies have records containing classified information that would be withheld.*
    - Internal personnel rules and practices - *Any information related solely to the internal personnel rules and practices of the Board. This exemption may also be used to withhold internal policies (e.g., security procedures) whose disclosure might lead to circumvention of those policies.* ***Many US financial regulatory agencies use this exception to withhold datasets.***
    - Statutory exemption - *Any information specifically exempted from disclosure by statute (other than 5 USC 552b), if the statute (A) requires that the matters be withheld from the public in such a manner as to leave no discretion on the issue or (B) establishes particular criteria for withholding or refers to particular types of matters to be withheld. Examples include grand jury materials and currency transaction and suspicious activity reports.*
    - Trade secrets; commercial or financial information -*Any matter that is a trade secret or that constitutes commercial or financial information obtained from a person and that is privileged or confidential. In the context of requests for applications and application-related materials, the exempt information often includes business plans, pro forma financial information, nonpublic portions of*

*transactional agreements, and descriptions of due diligence procedures and findings. Other information, such as copies of individual loan files obtained during an examination, and voluntarily submitted proprietary information, may fall within the scope of this exemption.*

- Inter- or intra-agency memorandums- *Information contained in inter- or intra-agency memorandums or letters which would not be available by law to a party (other than an agency) in litigation with an agency. Often does not apply for our purposes.*
- Personnel and medical files - *Any information contained in personnel and medical files and similar files the disclosure of which would constitute a clearly unwarranted invasion of personal privacy. Financial data withheld under this exemption includes the names and/or personal addresses of shareholders holding less than 10 percent of the shares of a bank or bank holding company, completed interagency biographical and financial reports, nonpublic portions of employment or non-competition agreements, and other pieces of personal information that vary from agency to agency.*
- Information compiled for law enforcement purposes - *Information usually withheld under this exemption includes investigatory records related to pending or potential enforcement actions (7)(A); details on individuals who are targets of or witnesses to pending or completed investigations (7)(C); and materials reflecting the different financial regulators procedures for conducting investigations (7)(E). However the exact application of this exception vary between agencies.*
- Examination, inspection, operating, or condition reports, and confidential supervisory information - *Any matter that is contained in or related to examination, operating, or condition reports prepared by, on behalf of, or for the use of an agency responsible for the regulation or supervision of financial institutions, including a state financial institution supervisory agency. This exemption often includes examination reports, examination-related correspondence, examiners' work-papers, and audit plans.* ***Many US financial regulatory agencies use this exception to withhold datasets.***
- Geological and geophysical information - *Not often used.*

- An additional roadblock is the bureaucracy and complicated international legal knowledge necessary for dealing with central banks. We contacted multiple central banks, and although the specific policy for each country was different the results remained generally consistent. Researchers working with the data must ensure that their calculation results contain no data that can be traced back to individual observation units or statistical units such as banks, (non-financial) corporations, individuals or households. Different disclosure control regulations exist for research results with regard to data

confidentiality. Even central banks like Bundesbank that are LEI ROC members still have these regulations in place, making it very hard for us to use the data in the context of our project.

- Furthermore, many large public databases run by private corporations, such as OpenCorporates, have specific regulations set for access to their data. A condition of access in these cases is that the use is for a public purpose, for journalism, for academic research, etc..
- Lastly, many of the key datasets are private (Thomson Reuters, Dun and Bradstreet, Avox, etc) and should be purchased inorder to undergo a complete version of our project. These datasets not only offer key attributes and identifiers but also often have the most complete data dictionaries and are thus the most efficient to work with.