FIRST YEAR PROGRESS REPORT

EU FET OPEN PROJECT FOC
255987

Deliverable D1.1

"Preliminary Report of Activities in WP1 - Data Collection and Consolidation"

WP Leader : Felix Reed-Tsochas

Objectives

Set up an European database of financial networks including: a) institution financial profile, b) financial linkages between institutions, c) most important indicators relevant to the credit market, and d) country-level socio-economic indicators. The activity of this work package is organized in the following Tasks:

T1.1 Balance Sheet Data
T1.2 Financial Exposure Data
T1.3 Credit Market Indicators Data
T1.4 HSBC Dataset
T1.5 Query Logs
T1.6 Consolidation, Consistency and Sharing

Progress Overview

We have created a master database to house data from across the consortium and have begun to populate it with the financial data discussed in tasks T1.1 through T1.5. The completion of the database – originally scheduled for month 12 – is now scheduled for month 18. This delay was caused by difficulties in finding a suitable candidate to manage the project within the UOXF.MQ node. Now that the position has been filled (in month 9), WP1 is progressing as scheduled in the updated grant agreement. In addition to the data discussed in the original grant agreement, we have sourced new data for the project – some of which we hope to integrate into the database.

See below for details of specific progress on each of the tasks.

Infrastructure

The data is currently housed in several SQL databases that are in different physical locations. Care has been and will be taken in cleaning and consolidating the data as it is placed within the master database structure – for example, making sure that the data are

uniform and share the same names across different tables. As discussed in D3.1, the databases are made available to the collaborative platform through web services with the JSON file format for data exchange. Data can be made private or public by appropriately configuring web services for each database.

T1.1 Balance Sheet Data ((Planned duration Month 6-24)
The goal of task T1.1 is to collect financial balance sheet data that can be used in WP3-4. These data fulfils *objective a)* of WP1. The activity of this is delayed with respect to the plan. However, two relevant data sets have been collected by ETHZ.

1) FED Discount Window database. This is a unique database containing detailed information relevant both to balance sheets of institutions and their liability exposure relation. As a result of a legal action pursued by Bloomberg, the FED was forced to disclose, last spring, detailed information on all loans granted to financial institutions in distress during 2 years around the peak of the crisis. The information was released in a big PDF document of about 7'000 pages reporting near to 90'000 transactions with several payment mechanisms and a varying format. The release raised a lively discussion on the media. However, data is hard to analyse systematically due to its heterogeneous text format. With a significant effort and combination of skills in computer science and finance the document was parsed and data were transformed in a relational database. Table 1 summarizes some basic statistics on the data set.

Total transactions number	90324
Total Volume of transactions	48'742 B$
Average amount of a single transaction	54 M$
Minimum transaction amount	1000$
Maximum transaction amount	61 B$
Maximum daily transaction number	2232 (5th May 2009)
Maximum daily transaction Volume	1890 B$ (27th April 2009)
Total number of financial institutions (not counting for companies with multiple names)	1967
First 10 companies for Volume Loans (not counting for expired loans just the total nominal values of the loans)	DEPFA BK PLC NY BR, BANK OF NY MELLON, DEXIA CREDIT LOCAL NY BR, JP MORGAN CHASE BROK, NORINCHUKIN BK NY BR, AIG, BANK OF NY, ARAB BKG CORP NY BR, US CENTRAL FCU, PARK NB

Table 1. FED Discount Window Dataset. Basic Statistics

The dataset provides, on a daily basis, the amount of liabilities towards FED of all institutions involved, along with the maturity of the loans. With some elaboration one can then infer the total daily exposure of FED towards each institution and overall. This

dataset has been used in Task4.1 (Analysis of Bank Markets) of WP4in order to make a first simple analysis of the structure of FED exposure and its evolution in time.
There has been a significant debate, since the release, about the disclosure this information, its meaning in terms of systemic risk and the right of US tax payers to know how much money and to whom it was given. In June 2011, we were probably the first to have this data parsed into a database and analysed. However, we did not have enough resources to push faster for publishing at least a report on our work. A working paper will only be available on the FOC website by the week of the first year review meeting. Remarkably, in the last days, Bloomberg has released a widget on their website to analyse this data, while the Accounting Agency GAO has released a document analysing the data from an accounting point of view. ETHZ is happy to share this dataset with other nodes of FOC who want to collaborate on analysing the data.

2) FDIC dataset. The Federal Deposit Insurance Corporation (FDIC) website provides various types of data mainly for the US, in text format. ETHZ has gathered and transformed into a database a dataset of financial information on US institutions. This includes total assets, cash flows, return on assets, net income as well as the geographical area of influence of each bank (state and county). Moreover, ETHZ has gathered and cross-referenced with this dataset, the list of default banks in US, including the date and the total assets at the moment of the default. This dataset has been used in WP3 to produce examples of risk evolution maps, available in the FOC platform. ETHZ is happy to share this dataset with other nodes of FOC who want to collaborate on the analyses.

T1.2 Financial Exposure Data (Planned duration Month 6-24)
The goal of task T1.2 is to collect financial exposure data for firms – individual banks and corporations – and also for different economic sectors across various countries. This data fulfils *objectives b)* and *d)* of WP1. As expected this data is very difficult to gather due its confidential nature and its strategic value. So far, we have gathered the following datasets.

e-Mid Dataset. The CITY node has access to data on interbank lending from e-MID, which is an European interbank market. We are currently working out the terms for sharing of this data.

The ETH node has provided to this task the following datasets:
a) ORBIS Dataset. This database includes information on ownerships ties between a large number of companies which could be used as a proxy of financial exposure.
b) FED Discount Window dataset. As described earlier it provides both balance sheet information and exposure information. Unfortunately, this is limited to the loans granted by the FED to the institutions in need. Therefore, the network structure is trivially a star.

c) CDS network dataset. The dataset contains the historical time series of the (credit default swaps) CDS prices on major US financial companies. This dataset is used in T4.2 (Interdependence and Trend Reinforcement) of WP4 in order to estimate the dependence between any two institutions based on the co-movements in their CDS price time series. We have obtained snapshots of networks that can be used as an estimate of the network of exposure among institutions.

Finally the UOXF.MQ node has downloaded data that could be used to estimate exposures among institutions. This is included in:
- the Bank Holding Company Performance Report,
- the International Swaps and Derivatives Association Surveys and Markets Statistics,
- the Bank for International Settlements Survey sand Markets Statistics,
- the Depository Trust & Clearing Corporation Warehouse Reports, and
- the Euro Area Accounts Dataset.

We are currently working with various members of the consortium to decide how best to integrate this data into the master database.

T1.3 Credit Market Indicators Data (Planned duration Month 6-30)
The goal of task T1.3 is to collect credit market indicators data for a variety of markets and countries. This data fulfils *objectives c)* and *d)* of WP1.

Federal Reserve Bank of St. Louis dataset. The UOXF.MQ node has collected data on interest rates that are made publicly available by the Federal Reserve Bank of St. Louis and has incorporated this data into the master database. The data types are too numerous to list individually, but include rates on different qualities and maturities of US corporate debt, different maturities of US government debt, different qualities of foreign corporate debt, spreads between rates of different qualities, corporate paper rates, US federal funds rates, different maturity swap rates, US treasury inflation indexed rates, Eurodollar deposit rates, certificate of deposit rates, etc. Depending on data type, the frequency of observations are anywhere from daily to annually.

RMI Dataset. The ETH node has collected from the NUS Risk Management Institute of University of Singapore some data on the default probability of several thousands of companies (financial and not) from several countries in the world. These data are directly related to the credit market data because the credit quality is strongly dependent on the costs to insure a company against a possible default. ETHZ is happy to share this data in a collaboration. This data could be used to study the spread of systemic risk among firms and across countries.

T1.4 HSBC Dataset (Planned duration Month 6-24)

The goal of task T1.4 is to interface with HSBC, using the high-quality financial datasets they maintain. The UOXF.MQ node has had regular meetings with HSBC over the past 12 months. From these meetings, we have gained a better understanding of the proprietary data housed at HSBC and have started collaborating on several research projects. HSBC maintains a high quality database of most foreign exchange currency pairs starting in 1992, and we are currently using this data to search for early warning signs of interbank stress. HSBC also maintains data on major economic news announcements from a large number of countries. Using this data, we are analyzing how prices respond to unanticipated news announcements so that we can better understand how economic shocks affect price and volatility. This data cannot be shared with the Consortium. Under quite strict rules collaborations of the FOC nodes are possible via the UOXF.MQ node.

T1.5 Query Logs (Planned duration Month 1-24)

The activity of task T1.5 consists in creating a dataset of anonymised and aggregate query data extracted from the logs of the Yahoo! Search engine. This data has been and will be used in Task T4.5 of WP4 in order to try and detect in the querying activity of web users possible precursors of financial trends. To this date, the BM node has compared the Yahoo! query-log and the trading volumes of a set of companies traded in the NASDAQ (National Association of Securities Dealers Automated Quotation) stock market. The query-log data analysed is a segment of the Yahoo! US search-engine log, spanning a time interval of one year, from May 1st, 2010, to April 30th, 2011. The query-log stores information about all the actions performed by users during their interactions with the search engine. For instance we have the queries they submitted and the result pages they were returned, as well as the specific documents they decided to click on. For every user query we have a timestamp representing the exact time point of its submission. We have used this temporal information to aggregate search volumes at different levels of granularity. We have computed the query volumes by extracting and aggregating on a daily basis two different types of queries for each traded company: (1) all the queries whose text contains the company ticker (i.e., “YHOO” for Yahoo!) as a distinct word; (2) all the queries whose text is exactly matching the company names, after removing the legal ending (“Incorporated” or “Corporation” or “Limited”, and all their possible abbreviations). Furthermore, every action is also annotated with a cookie, representing the user who submitted the query. These cookies allow us to track the activity of a single user during a time window of a month. We discarded information about user identity when computing the aggregated search volumes for the stocks considered, but used this information to compute the user volumes by counting the daily number of distinct users who made at least one search related to one company. As the query data involves personal information, it is available only on-site and is unlikely to be included in the master database.

T1.6 Consolidation, Consistency and Sharing

Task T1.6 considers the overall consolidation, consistency and sharing of data across the consortium. Some data will be accessible to other nodes if used on-site and subject to controls of the data vendor, some data will be available to other nodes when analysis is done in collaboration with the node holding the data, and some data will be accessible to all within the consortium. The consolidation and consistency of the data across datasets will start in the second year. A preliminary summary of the accessibility of data is shown in *Table 1* below.

	Available to Others On-Site	**Available to Others In Collaboration**	**Freely Available to All**
FED Discount Window		X	
FDIC Dataset		X	
e-MID Dataset		X	
ORBIS Dataset		X	
CDS Network Dataset		X	
St Louis Fed Dataset			X
RMI Dataset		X	
HSBC Dataset	X		
Yahoo Query Dataset	X		

Table 1: Preliminary summary of the accessibility of data (subject to change as we continue discussions with various nodes and data vendors).

Supplemental Datasets

In addition to the data discussed in tasks T1.1 through T1.5, we have sourced additional financial data that can potentially benefit the project, but which was not originally planned in WP1.

The UNIVPM node has obtained **Japanese bank-firm credit network** data and also **Brazilian and Italian credit network data**.
The UOXF.MQ node has been in discussions with the South African Reserve Bank about obtaining **credit network data from South Africa**.
The CNR node is working on data of country trading at the level of single product. These data are currently publicly available in different datasets. In order to be used they need to be aggregated and cured. Even if this kind of analysis was not present in the original project we are more and more convinced that this kind of analysis can provide a further indicator of fragility/solidity of national economies.
Finally the BM node now has access to **Twitter data** that supplements the Yahoo query data discussed in T1.5. We are currently working out the terms for sharing of these supplemental datasets.