IMT School for Advanced Studies, Lucca

Lucca, Italy

The financial value of data

PhD Program in Computer, Decision and System Science

XXVIII Cycle

By

Giuseppe Pappalardo

2016

The dissertation of Giuseppe Pappalardo is approved.

Program Coordinator: Prof. Rocco De Nicola, IMT School for Advanced Studies Lucca

Supervisor: Guido Caldarelli, IMT School for Advanced Studies Lucca

The dissertation of Giuseppe Pappalardo has been reviewed by:

Prof. Tiziana Di Matteo, University College London, King's College London

Prof. Claudio J. Tessone, Department of Business Administration - Network Science, University of Zurich

IMT School for Advanced Studies, Lucca

2016

Contents

List of Figures

List of Tables

Chapter 2 shows an explorative analysis which relies on Facebook social data as a source of microeconomic information. All data presented are part of the data framework I developed, described in **Chapter 1**.
Chapter 3 is an extended version of the article "The Accounting Network: how financial institutions react to systemic crisis", co-authored with Alessandro Chessa, Andrea Flori, Fabio Pammolli and Michelangelo Puliga, published on PlosONE (`http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0162855`). Here, light was shed on the banks' response during the crisis using information from balance sheets. I used the data framework in order to retrieve balance sheets data and perform the analysis.
Chapter 4 is an extended version of the paper "Blockchain Inefficiency: The Bitcoin Peers Network", co-authored with Guido Caldarelli and Tomaso Aste, presented at the 2 Workshop of Peer-to-peer Financial Systems, 8-9 September, 2016 (University College London). In this work I developed a modified Bitcoin client capable of collecting data exchanged by peers on the Bitcoin Network. I used the data received by the client for two main purposes. The first one is to validate the results published by Decker [33] about blocks dynamics. The second goal is to study the dynamics of transactions. Transactions were analysed for the first time and their mechanism of inclusion in the Blockchain was shown to be responsible for the efficiency/inefficiency of the Bitcoin payment system.

Vita

October 17, 1984 Born, Acireale (Catania),Italy

Current Appointments

2016 Research Associate at University College London, Department of Computer Science

2013 PhD Candidate in Computer, Decision and System Science at IMT School for Advanced Studies Lucca, XXVIII Cycle

Research Visit

2015-2016 Visiting Researcher at University College London, Department of Computer Science (9 months)

Higher Education

2012 MSc in Automation Engineering and Control of Complex Systems, University of Catania

Thesis: Safety automotive application design using single and multicore architecture following ISO-26262: Hardware versus Software Redundancy.

Supervisor: Prof. R. Caponetto

Advisor: Eng. R. Martorana

2008 BSc in Computer Engineering, University of Catania

Thesis: Development of a Software for administration of low-cpu architectures

Supervisor: Prof. M. Malgeri

Publications

1. Michelangelo Puliga, Andrea Flori, Giuseppe Pappalardo, Alessandro Chessa, Fabio Pammolli, "The Accounting Network: how nancial institutions react to systemic crisis" *(submitted to PlosONE)*

Conferences and Posters

1. Stefano Battiston, Guido Caldarelli, Alessandro Chessa, Andrea Flori, Fabio Pammolli, Giuseppe Pappalardo and Michelangelo Puliga, "Estimating the performance and the systemic risk of a wide set of companies using financial statement data in a complex network framework" *(ECCS conference, 2014)*

2. Tharsis T. P. Souza, Giuseppe Pappalardo, Soong M. Kang, Guido Caldarelli, and Tomaso Aste, "Multiplex Structure of Social Media and Financial Networks" *(IC2S2 conference, 2016)*

3. Giuseppe Pappalardo, Guido Caldarelli, Tomaso Aste, "Blockchain inefficiency: The Bitcoin Peers Network" *(2nd Workshop Peer-to-peer Financial Systems, 2016)*

Abstract

During last years data are growing up in term of size and importance. Several applications start to focus on using different kind of data for describing several phenomena or, if possible, trying to predict them. Thanks to social networks become easy to gather data from users since they are directly providing them. Using Facebook every day, the average user provide a clear profile about what he likes, which cities he visit, his own favourite places and also about people who is sharing part of its life. Despite Facebook at the beginning do not provide any advertisement, sells any goods or asked money to its own user, on 2004 it was evaluated 10 billions dollars and today its own value is more than 300 billion dollars. This huge amount of money suggests that probably there is an "inner value" for who is able to know or use these kind of information. The role of Network Theory in the study of the financial crisis has been widely spotted in the latest years. It has been shown how the network topology and the dynamics running on top of it can trigger the outbreak of large systemic crisis. In this thesis are presented some application of data with the aim to exploit their own financial value using a network perspective. First, an explorative analysis on geo-localized data from Facebook for economic estimation is showed. Moreover, the network from financial statements covering a large database of worldwide banks is introduced, showing some features emerging during last global financial crisis of mid-2007. Finally, the Bitcoin network is investigated, measuring how long blocks and transactions require to propagate through peers, introducing an efficiency measure of the Bitcoin payment system.

Introduction

In the last years Internet has become a huge source of information and data that span in different fields. Nowadays finding information related to companies such as income statements, balance sheets and price series is easy, but also, thanks to social networks, it is possible to find also data related to common users and their habits, including for instance which places they visited, brands they like and their own topics of interest. Data derived from user activities, especially, are growing up in size every day and scientists are studying how it is possible to exploit this knowledge opening to new applications, such as forecast economical indicators or profiling models for users. Recent applications of social media data, collected from Twitter, IMDb and data gathered from online newspaper, were used to forecast the box-office revenue of movies after their release [13; 64; 73]. Aggregated information of Tweets, or volumes of queries submitted by users to search engines, become also accessible, i.e. through Google Trend, allowing to see what are the most popular trend queries in a certain geographic area, focusing also on user age or on a specific reference time. These data were analysed in relationship to several economics and financial features, such as unemployment [16; 53], sales [16] or private consumption [32; 61]. All these studies shows that it is possible to track present outcomes or to predict future behaviours within an acceptable success rate, using social or economic/financial data [45] and this leads researchers to use data to track equities and stock returns [47] or to provide new insight or pattern in advance [54]. As an example, evidences of correlation were found between transaction volume of com-

panies and social data, such as the number times these companies were cited by the Financial Times [8], or considering volume of web search queries [28] and Twitter events [58]. Using social media data in order to predict consumptions is driven by the fact that people are leaving their opinions and reviews about products and this lead itself to generate a trend for the market. The full potential domain applications of using data to forecasting sales events has not been totally explored, also due to the fact that relationships between data and products or events cannot hold for those which received less attentions on social media [37]. The use of Network Theory also grows up in fields such as economics and finance, as need of better understanding relationships among entities, such as banks, companies or users. Since financial institutions could be more or less interconnected, according to their investment strategy, and they can suffer by contagion or cascade effects. Cascade dynamics happens when a failure of a single or group of nodes can be propagated to its neighbours and this is a critical factor for financial instability [11; 14; 15]. This scenario is common when institutions are too-connected-to-fail , reducing in this way their individual risk but gaining exposure to external cascade dynamics [6; 29; 43; 70] importing distress from their neighbours [66]. Given that, one possible solution can be introducing regulatory mechanisms on a collective level such that regulating each bank as a function of its individual risk, but considering also other correlated to it [6]. This is coherent with the fact that if the banks have only few links they are more robust against the default of a single institution [29]. Financial systems present a robust-yet-fragile tendency and even if the contagion probability may be low, when a default occurs the effects can be widespread [43; 44]. Default cascades in financial systems were introduced by Eisenberg and Noe [50] which defined the problem, concluding that the loss induced by bank default is an estimation of how the bank is important for the network's stability. This work is considered a standard on financial networks and adopted or extended [44; 46]. Main drivers for system cascades in interbank systems consist on network topology, liquidity, capital ratio and centrality of the bank who spreads the crisis [60]. Studies on how these factors matters can be found on [7; 22; 30].

When the market is illiquid, topology has an important effect but in general there is not one dominant topology [60]. In general capital helps banks, increasing their survival probability and also improving the performances during crisis [22].

Taking into account the previous literature, the aim of this work is to introduce a new data analysis workspace, which uses a network perspective and put in relationship different kind of social/news data with economical/financial/social data, having the power of making analysis or simulations useful to identify new general pattern, correlation or causality useful to data mining and forecasting applications.

The thesis is structured as follow:

- **Chapter 1** describes the data workspace created and presents all its available data, and also data used in the next chapters.
- **Chapter 2** contains an explorative analysis with the aim of putting in relationship Facebook data provided by the framework within US Census Data, in order to nowcast economic indicators.
- **Chapter 3** presents the work "The Accounting Network: how Financial institutions react to systemic crisis", which uses the framework created on the previous chapter to retrieve data from Bloomberg and does the analysis.
- **Chapter 4** reports the study "Blockchain Inefficiency: The Bitcoin Peers Network".

Chapter 1 describes the architecture of the data framework built. Here it will be presented all the data sources used to collect the data. The aim of the framework is to provide an environment which allows to analyse and simulate different kind of models for financial, news and social data. The framework provided all the data used in the following chapters and it was wrote implementing Bloomberg API, Facebook Graph API, US Census API, Google API and the Bitcoin protocol. Moreover, it will define the datasets used in the following chapters.

Chapter 2 illustrate a explorative analysis with the aim of studying Facebook Places Data for nowcasting applications. Currently, economical indicators are built using surveys. This kind of indicators are very important in order to estimate economy's trend of a certain region but they have the drawback of being available at least within a 3 months time delay. Specifically, I will investigate whether Facebook data can be used as proxy for defining economic indicators, or if they can be paired with actual indicators in order to increase their quality and to anticipate possible trends. To achieve that, dataset from Facebook will be paired with data from the United States Census Bureau (US Census).

On **Chapter 3**, Financial Data provided by the previous data framework, specifically Balancesheet Data for a set of Tickers belonging to the Banks Sector from Bloomberg, are used to introduce the concept of Accounting Network. Even if market data are highly representative of investors perception, the novelty idea behind this study is that, during periods of distressed market conditions, they could be dis-informative if considered alone. Since the availability of balance sheets for each ticker is different, we defined an index, called quality ratio, in order to see how the network stability change in relationship to the input dataset. After selecting and cleaning the dataset, it will be presented some analysis on the reference period from 2001 to 2013. Networks measures during the reference period will be compared with classical economic indicators from literature, and a community detection algorithm will be applied to identify which communities emerge from the network. Finally the principal component analysis is applied to understand which financial variables characterize each cluster. Some correlations for the reference period are discovered among network measures and economic indicators.

Chapter 4 is a study of the Bitcoin [55] peers network which consists of about 5 thousand client connected each other, using a customized versions of the Bitcoin Client and obtaining different performance in terms of data shared with the network. The topic is novelty since distributed

ledgers become popular only during last couple of years. Also, the reference literature is provided only by Decker [33] and Coinscope [10]. Decker analysed only the blocks propagation mechanism and defined a model able to investigate Fork events. Coinscope [10] found the influential nodes on the Bitcoin network, but his work is not repeatable due to some security changes on the Bitcoin client [5]. The present work studies the Bitcoin network, investigating if Decker's results are still valid today regarding block propagations, and introducing the concept of efficiency of the whole Bitcoin payments system, and of the Blockchain itself. Bitcoin users know that, after including a transaction in the Blockchain, it is necessary to wait until 6 confirmations (six new block on top of the one where the transaction is included) in order to consider the amount spendable. Moreover the Blockchain in practice is not able to process more than 6 transactions per second (in theoretical conditions). This research wants to show the network ability on processing transactions and to inspect the presence of any factor which may be exploited in order to gain a competitive advantage on the other users.

Chapter 1

Data collection

1.1 Introduction

This chapter describes the raw data collected from the sources in order to built the data framework. All the different API implemented, such as Bloomberg Terminal API, Yahoo! Finance and Facebook Graph API are explained and also are showed how data were collected from the Bitcoin Network. All the datasets described will be used in the following chapters. Facebook data represents the basis of two future work. The first one aims at studying whether and how geo-localized data from Facebook Places (likes, checkins and talking about counts) allow to discover universal patterns which enable to measure economic activities or to nowcast local economies for a certain area, introduced as an explorative analysis on Chapter 2. The second work aims at forecasting product sales with user interactions data on Facebook Pages related to selected brands. Balance sheets will be used in Chapter 3 to study relationships between economic indicators and network measures during the period from 2001 to 2013, including the 2008 crisis. Bitcoin and Blockchain Data will be studied on Chapter 4. These technologies are pretty recent and are attracting a huge amount of scientists, developers and practitioner even if the current academic literature about them is very poor.

1.2 Financial Data

Financial Data were downloaded using two main sources: Yahoo! Finance and Bloomberg API. Both were implemented using Python as programming language. Yahoo! Finance API allow the user to download data for free from several different Yahoo! services (i.e Finance! or Yahoo! Answer) or third party applications[1] (i.e. Stack Overflow, Spotify and Steam). The number of requests that can be submitted to the service is limited to 2000/hour (if the request is sent as anonymous user) or 20000/hour (using the OAuth Authentication and an API Key). Compared with yahoo! Bloomberg offers a huge amount of financial data and news, but requires to pay an expensive subscription. The framework built is able to process and analyse balance sheets and equity price series for a specified ticker.

1.2.1 Project 1: The Accounting Network

The dataset analysed covers the set of banks provided by *Bloomberg* which were active (i.e. with traded instruments) at the end of the first quarter of 2014. Although quarterly information is available, it was focused on annual balance sheets and income statements for accounting standard reasons, as different countries can have different obligations in terms of the provision of quarterly financial statements and this can lead to a mismatch and a poor variables coverage. Data are collected during the reference period from 2001 to end of 2013.
As regards financial statements data, they were selected large set of variables among those available in *Bloomberg* and related to the current regulatory framework [25]. The analysis focus on proxies for banking business models (see e.g. [31; 71]). In particular, balance sheet data provide a year-by-year picture of stock variables in terms of assets and liabilities for different instruments and maturities, while income statement data describe annual economic performances by partitioning profits and losses according to banking activities ranging for instance from interests to fees.

[1] A full list of available services made available by the API could be found on [72]

Since national regulations allow firms to fix a different end of fiscal year, the "end of year" definition was extended and the relative financial statements according to a window in the range between three months before and after the end of the solar year. Solving overlapping issues in variables definitions, as well as the base currency choice, constitute the first step in the data pre-processing procedure. Firstly, total and subtotal measures were discarded (as they are redundant measures), and secondly US dollars is chosen as currency base, thus facilitating banks comparisons.

Working with financial statements data often leads to limitations in data coverage and completeness. Therefore, the starting point of our analysis is represented by the selection of a stable set of banks in terms of data availability during the sample period. In particular, banks might change the composition of their financial statements or they might be excluded by the *Bloomberg* provider due to several reasons, such as for instance a new regulation or a change in the bank's economic activities. This, in turn, might cause *missing values* for some variables or lack of financial statements for several banks in certain years. In order to limit the impact of these issues on our findings, it would be explained the methodology used in order to measure the coverage of available variables for each bank in the reference period. The *Quality Ratios (QRs)* is defined as the proportion of available and usable variables V_{OK} over the maximum of all possible ones V_{ALL} in the sample period: $QR = V_{OK}/V_{ALL}$. The tuning of this indicator, combined with two more filters on the frequency of financial reporting, provides a stable set of banks identified by their QR. The two additional criteria are: a minimum number of financial statements of ten out of thirteen possible fiscal years and a maximum gap period between two consecutive annual reports equal to seven hundred days. Once selected those banks that report almost continuously their financial statements, they were studied according to their respective QR. Actually, individual QRs, as empirically computed on the entire perimeter, lie in the range between 0.3 (low accuracy/coverage) and 0.8 (high accuracy/coverage). Interestingly, many measures computed on the sets of banks obtained by fixing the QR do not seem to be significantly af-

fected by its choice (except, as expected, for high QRs, where the size of the sample reduces significantly). There are few available banks which hold a greater values of the QR parameter, since only few of them have a large set of variables present in many of their financial statements. As the estimates are stable in a reasonable QR range, in this work it is analysed the set arising from the case of QR = 0.5 that, even if arbitrary, still represents a good compromise between the accuracy of the estimate and the size of the sample (see Figure 12).

1.3 Facebook Graph API

The framework built is able to retrieve Facebook Data using the Facebook Graph API[39]. On the following it will be described how retrieve data from Pages entities related to social counters (i.e. check in, likes) and user interactions (posts and comments).

1.3.1 Project 2: Facebook as micro-economic data source

The aim of this work is to show the possibility to nowcast economical indicators through social networks data. Data used are collected from the US Census, as micro-economic source, and from Facebook. We select a relevant geographic sample consisting in more than 100 counties and about 5700 tract. We used this geographic sample in order to retrieved all the Facebook pages located inside the specified tract area, understanding how pages category are distributed among each tract and recording their everyday actions.

The dataset is built selecting two random counties from each State of the United States [2] in order to obtain a dataset that includes information

[2]Regads Alaska were excluded North Slope, Northwest Arctic, Yukon-Koyukuk, Nome, Wade Hampton, Bethel, Dillingham, Aleutians West, Aleutians East Denali, Fairbanks North Star, Kenai Peninsula, Lake and Peninsula, Matanuska-Susitna, Skagway-Yakutat-Angoon, Southeast Fairbanks and Valdez-Cordova because their size is not comparable with others counties.

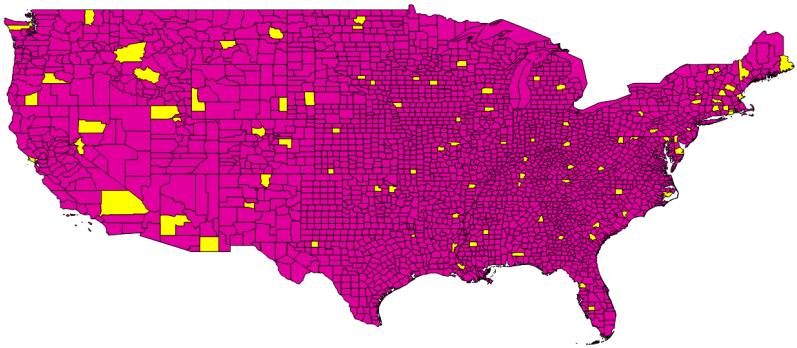

Figure 1: America's sample. In yellow are represented the selected counties.

related to Places from 101 counties of the United States (see tables 1 and 2 for the full list).

For each Place, Facebook Graph API provides two kind of information, some static attributes and some counters. In order to keep the data "general purpose", the statical information selected (if available) for each place are represented by:

- id - The Page ID.
- about - Information about the Page.
- attire - Dress code of the business. Applicable to Restaurants or Nightlife. Can be one among Casual, Dressy or Unspecified.
- category - The Page's main category. e.g. Product/Service, Computers/Technology.
- category_list - The Page's sub-categories.
- description - The description of the Page.
- founded - When the company was founded. Applicable to Companies.

Table 1: List of counties selected.

State	Counties
Alabama	Cleburne, Escambia
Alaska	Anchorage, Ketchikan Gateway
Arizona	Maricopa, Cochise
Arkansas	Lincoln, Greene
California	Marin, San Bernardino
Colorado	Elbert, Grand
Connecticut	Hartford, Windham
Delaware	Sussex, New Castle
District of Columbia	District of Columbia
Florida	Citrus, Desoto
Georgia	Bulloch, Macon
Hawaii	Kalawao, Honolulu
Idaho	Custer, Idaho
Illinois	Lee, Cumberland
Indiana	Perry, Johnson
Iowa	Henry, Monona
Kansas	Cherokee, Seward
Kentucky	Breathitt, Logan
Louisiana	Iberville, Catahoula
Maine	Washington, Oxford
Maryland	Harford, Allegany
Massachusetts	Essex, Hampshire
Michigan	Lake, Saginaw
Minnesota	Freeborn, Watonwan
Mississippi	Leake, Copiah
Missouri	Cooper, Audrain
Montana	Dawson, Musselshell
Nebraska	Sheridan, Red Willow
Nevada	Pershing, Lyon
New Hampshire	Cheshire, Merrimack

Table 2: List of counties selected.

State	Counties
New Jersey	Mercer, Hudson
New Mexico	Valencia, Taos
New York	Greene, Schenectady
North Carolina	Randolph, Hyde
North Dakota	Ramsey, Eddy
Ohio	Highland, Hardin
Oklahoma	Logan, Creek
Oregon	Deschutes, Jackson
Pennsylvania	Cumberland, Cameron
Rhode Island	Kent, Washington
South Carolina	Barnwell, Marion
South Dakota	Hanson, Jerauld
Tennessee	Claiborne, Cheatham
Texas	Gregg, Midland
Utah	Davis, Box Elder
Vermont	Lamoille, Chittenden
Virginia	Hopewell, Lynchburg
Washington	Jefferson, Ferry
West Virginia	Mercer, Preston
Wisconsin	Chippewa, Dane
Wyoming	Lincoln, Goshen

- general_info - General information provided by the Page.
- global_brand_page_name - If the Page is in a Global Pages structure and the viewer has a role on the Page. This is the Page name with the country code(s) that redirects to it appended in parentheses. If the Page is not part of a Global Pages structure or if the viewer does not have a role on the Page, this is simply the Page name.
- global_brand_parent_page - If the Page is in a Global Pages structure, this is the brand's global (parent) Page.
- hours - Indicates the opening hours for this location.
- is_permanently_closed - For businesses that are no longer operating.
- is_unclaimed - Indicates whether the Page is unclaimed.
- is_verified - Pages with large numbers of followers can have the authenticity of their identity manually verified by Facebook. This field indicates whether the page is verified in this way.
- location - The location of this place. Applicable to all Places.
- name - The name of the Page. This field can only be updated when the Page has less than 200 fans.
- parking - Information about the parking available at a place.
- price_range - Price range of the business. Applicable to Restaurants or Nightlife.
- restaurant_services - Services the restaurant provides. Applicable to Restaurants.
- restaurant_specialties - The restaurant's specialties. Applicable to Restaurants.

Since these information usually do not change over time, they need to be collected only the first time and eventually requires periodical checks for any updates.
The Counters instead are dynamic, and they can change every time users interact with the place, therefore, in order to achieve daily statistics, they are recorder every day. Counters available are:

- talking_about_count - The number of people talking about the Page.
- were_here_count - The number of visits to the Page's location. If the Page setting Show map, check-ins and star ratings on the Page are disabled, then this value will also be disabled.
- likes - The number of users who like the Page. For Global Brand Pages this is the count for all pages across the brand.
- checkins - Number of checkins at a place represented by a Page.

Starting from the counters it will be possible to built time series related to a specific place or within a group of places relative to the same category. In order to collect data from Facebook it has been used GIS data provided by GADM database of Global Administrative Areas [42] regarding all the counties in the United States. For each county it is possible to get some information such as the coordinates of the box that includes the county and the points needed for defining its shape. After choosing the counties list for the sample, it was defined a fixed number of points on unit area density in order to guarantee that each county's number of points is proportional to their geographic area. Having multiple points for each county makes redundancy on the number of requests but it was observed to be a necessary step since, after some tests with the Facebook Graph API, the number of entities returned by a request is strictly sensitive by the coordinates of the point itself (this effect was noticed using Facebook Graph API v2.2). Hence, it was necessary to include more points of the neighbours area in order to get more accurate results from requests. Besides, there were some problems with the Facebook Graph API depending on the coordinates of the points in the request. To solve these problems, it was defined that an area of 1.7 squared degrees

Bachelor's degree or higher, percent of persons age 25+, 2009-2013	46.6%	31.9%
Veterans, 2009-2013	119,707	582,265
Mean travel time to work (minutes), workers age 16+, 2009-2013	27.0	25.7
Housing units, 2013	868,345	2,926,217
Homeownership rate, 2009-2013	58.2%	63.2%
Housing units in multi-unit structures, percent, 2009-2013	38.2%	25.6%
Median value of owner-occupied housing units, 2009-2013	$377,300	$262,100
Households, 2009-2013	802,606	2,629,126
Persons per household, 2009-2013	2.42	2.54
Per capita money income in past 12 months (2013 dollars), 2009-2013	$39,911	$30,742
Median household income, 2009-2013	$71,811	$59,478
Persons below poverty level, percent, 2009-2013	11.5%	13.4%
Business QuickFacts	**King County**	**Washington**
Private nonfarm establishments, 2012	63,841	175,553[1]
Private nonfarm employment, 2012	1,017,348	2,361,697[1]
Private nonfarm employment, percent change, 2011-2012	-1.3%	0.3%[1]
Nonemployer establishments, 2012	150,301	412,542
Total number of firms, 2007	196,686	551,340
Black-owned firms, percent, 2007	3.3%	S
American Indian- and Alaska Native-owned firms, percent, 2007	0.9%	1.2%
Asian-owned firms, percent, 2007	10.9%	6.8%
Native Hawaiian and Other Pacific Islander-owned firms, percent, 2007	0.3%	0.2%
Hispanic-owned firms, percent, 2007	2.8%	3.2%
Women-owned firms, percent, 2007	29.0%	28.7%
Manufacturers shipments, 2007 ($1000)	37,390,762	112,053,283
Merchant wholesaler sales, 2007 ($1000)	41,042,685	76,790,966
Retail sales, 2007 ($1000)	37,153,888	92,968,519
Retail sales per capita, 2007	$20,002	$14,380
Accommodation and food services sales, 2007 ($1000)	5,478,918	12,389,422
Building permits, 2012	11,614	28,118
Geography QuickFacts	**King County**	**Washington**
Land area in square miles, 2010	2,115.57	66,455.52
Persons per square mile, 2010	912.9	101.2
FIPS Code	033	53
Metropolitan or Micropolitan Statistical Area	Seattle-Tacoma-Bellevue, WA Metro Area	

Figure 2: U.S Census statistics for the King county.

(obtained by the product of latitude and longitude) contains 1000 points. For the sample selected it means 30520 points to request through the API. This number is higher compared to the Facebook daily rate limits. In order to avoid this limit can lead problems that could cause the ban of the user from the developer platform, each point was checked and classified by its result that could be:

- Done: a request for a point which return data.
- Disabled: a request for a point which does not provide any data.
- Error: a request for a point which, despite it is valid, provides an unknown error (probably due to Facebook server's availability or with the availability of the page/points itself at the current time)

After this points classification, for each day all the Disabled points were neglected, obtaining a set of 18295 points and saving more than 12 thousand requests in a day. Indeed, since each county has not the same geographical size, they should include a different number of points, so for two neighbours points, their results may contain a common data part. These facts could be exploited in two ways:

1. Use the pages' id provided by each point to solve a minimization problem in order to have the same number of pages with the minimum number of requests.

2. Order the point requests with respect to counties in order to request two neighbours points with a time delay and not in series. This will be useful for obtaining more updated information for each page which is included in more than one point result.

It was chosen the second approach, in order to increase temporal information about pages from the set of about 18 thousand points.

1.3.2 Project 3: Facebook to forecast brand sales

It is also possible to analyse the messaging activities pages related to brands have with users in order to forecast sales. Each brand on Facebook may have a variable number of pages, according to its social media presence, popularity or marketing strategy adopted by the firm. Starting from a list of famous firms, for each of them, it was looked for a set of related Page entities from Facebook in order to downloaded all the available information. A page on Facebook can be created by users or automatically generated by the social network itself after a third-user action (i.e. users can made a check in and, in this way, if there are no any Page related to the Place, Facebook will automatically create it). Furthermore, users can create "fake Pages" in order gain traffic/visitors pretending to be a popular brand. Through the Search function of the API there are several information available, as it possible to see [4] for each page are available the following information:

- Category indicating which category the page is related to (i.e. Clothing, Local Business, Community etc).

- is_unclaimed if it is setted to *True* means that the Page was self generated by Facebook and it does not belong to any user. A user can claim the possession of the page in future, in this case it will be automatically set to *False*.

- global_brand_root_id When more pages belong to the same brand or page it is possible to find here the page root.
- name The main name showed by the page.
- website The website of the page, if available.
- brand The brand we used for the search.
- global_brand_parent_root similar to *global_brand_root_id*, showing the name of the root page.
- is_verified Some pages can have the authenticity of their identity manually verified by Facebook. In this case this field is set to *True*

In order to isolate "fake Pages" we grouped them using this fields obtaining, as an example, groups of Pages which share the same website or the page root and we neglected all the pages who has not any information in common. After this first check we manually checked again the pages to keep, as showed in table 3 Actually Facebook Graph API Don't allow to gather Data from self-generated Pages, and try to remove fake pages once reported. Data retrieved through the API have a hierarchical structure, as showed on 3. To built the time series we extract from each post all the comments, obtaining the following fields dataset:

- brand Indicating the brand which the page is referring to.
- page Id of the page on Facebook.
- id Id of the post. The format Facebook uses for ids is *pageid_postid*. Comments have the same id of the parent post.
- type indicate if the record refer to a post or a comment.
- created_time timestamp of the post or comment.
- updated_time timestamp of the last modify, if any.
- from Id belonging to who create the post or comment. It could refer to an user or to another page (pages could leave comments).

- message_lenght number of character of the message.
- place_id Id of the Facebook Page registered to the place where the user wrote the post, if available.
- place_latitude latitude coordinates of the user who left the post, if available
- place_longitude longitude coordinates of the user who left the post, if available
- users_likes Count of users who liked the post or the comment.
- pages_likes Count of pages which liked the post or the comment.
- total_likes_count total number of likes for the post or comment equal to *users_likes* + *pages_likes*

1.4 Bitcoin Data

In order to collect Bitcoin data it is necessary to develop a custom Bitcoin client which will join to the network and store the data. In order to work well, the customized client has to implement at least the following messages:

- getaddr This message is used to request a list of known node to a connected peer and will issue an "addr" message as response.
- addr Used to send peers to neighbours after a new connection or as reply of a "getaddr" request.
- ping Message used to check if the connection is still alive.
- pong reply to a "ping" message.
- inv This message is sent by a client in order to announce new blocks/peers/tran

Table 3: Number of Pages found and selected on Facebook for each Brand.

Brand	Total number of Pages	Selected Pages
Abercrombie	432	30
Adidas	493	89
Calin Klein	406	25
Chico's	447	202
Crocs	479	40
Gamestop	384	4
HomeDepot	421	10
kenneth cole	325	7
LaCoste	513	25
Levi	504	15
Loreal	504	101
Mattel	510	21
Meijer	417	50
newbalance	260	28
Nike	501	66
nissen	476	17
petco	366	17
Puma	510	30
radioshack	423	31
Reebok	427	114
Spanx	91	6
Speedo	471	36
Timberland	443	47
Tommy Hilfiger	411	37

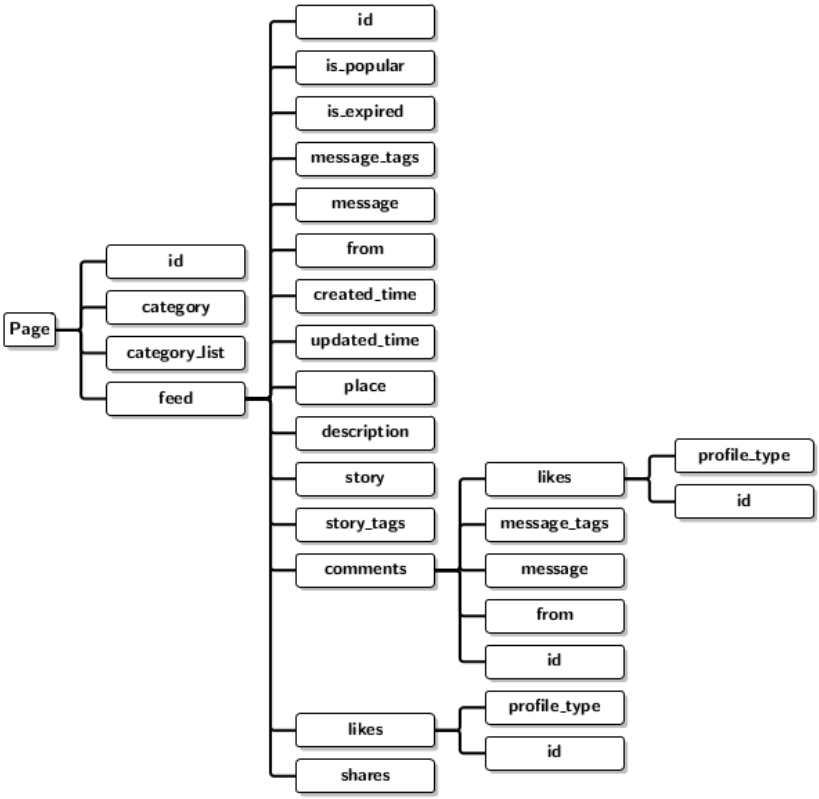

Figure 3: Structure of the Data retrieved using Facebook Graph API for the entity Page.

Implementing these messages and establish a connection with each reachable node it is possible to retrieve almost all the activities generated by the networks. Since the amount of data propagated is very high, it was stored only transactions and blocks ids. Establishing more than one connection for each node has the drawback that it may interfere with the network's dynamic. Regarding blocks propagated the dataset consist of the following fields:

- Address the ip address of the node
- block hash the block hash propagated by the node
- timestamp an integer corresponding to date and time when the block was received by the client.
- Height the index of the block in case it was included in the Blockchain.
- Blockdate the timestamp of in case it was included in the Blockchain.

1.4.1 Project 4: The Bitcoin Peer Network

We listened the Bitcoin network activity during the period from Wed, 04 May 2016 01:20:45 GMT to Wed, 11 May 2016 18:44:58 GMT. We collected 592GB of data in a period of 1209 valid blocks (from block height 410119 to 411327) mined during the listening time window. The most part of data regard transactions "inv" messages (589 GB) while the remaining is related to blocks "inv" messages. During the investigation time we found 12424 nodes, from which 12168 relay transactions, 11532 relay blocks and 11549 relay both. Due to the large amount of data received, we decided to classify blocks and transactions as follow:

- Blocks

  Mined During Listening Block (MDLB) - This set identify all the blocks which were included on the Blockchain during the listening period and propagated by the peers before the next block was discovered. There are 1209 blocks discovered by 530 source nodes and spread through 11179 destination nodes. The maximum number of blocks discovered by a single node during the listening time is 86.

  Echo Blocks (EB) - This set identifies all the blocks already included in the Blockchain and propagated in delay. There were propagated 406457 echo blocks, from 6938 nodes. These Data were not analysed.

Fork Block (FB) - This set identifies all the blocks not included in the Blockchain, propagated by the peers with a valid hash (below the proof-of-work threshold). There were 34 fork Blocks. These data were not analysed.

Invalid Block (IB) - This set identifies all the blocks not included in the Blockchain, propagated by the peers and having a hash above the proof-of-work threshold (so they should be discarded by the peers). There were 51103 Invalid Blocks transmitted by 23 nodes. These data were not analysed.

- Transactions

  Blockchain Transaction (BT) - Valid Transaction, included in a Blockchains Block and propagated before the block it is included is discovered and propagated through the network. We received 1744899 Blockchain Transaction, from which 1725508 were included in a block during the listening time and 19391 were included in a block after the recording time. Since it is not possible to know for how long a transaction is spread on the network we decided to discard the transactions received for the first time during the first and the last block and also those which were received after they were included into a block. We also don't analysed transaction with Locktime setted (about 5 thousands), the final set of transactions analysed count 64994 which were generated by 2518 nodes.

  Echo Transaction (ET) - Valid Transaction, included in a Block but propagated in delay. We received 12425 echo transactions. These data were not analysed.

  Invalid Transaction (IT) - Transaction not valid for some reasons. We received 62889 Invalid transactions. These Data were not analysed.

On table 4 it is possible to see some statistics regarding blocks of the Mined During Listening set, making a comparison between the time of each block and the time when a block was discovered from the network.

	Listening Time	Blockchain Time
Minimum Time	-5.48 seconds	-558 seconds
Maximum Time	4650.09 seconds	4642 seconds
Medium Time	550.05 seconds	550.05 seconds
Variance	550.11 seconds	550.30 seconds
Percentile 50%	383.25 seconds	384 seconds

Table 4: This table show some time statistics related to Mined During Listening Blocks set, comparing timestamp wrote on each block within the time reported inside the Blockchain. The time on the Blockchain can be wrong since a miner could vary the timestamp if the nonce don't converge to a valid proof-of-work block. The minimum time is negative due to a Fork event. During the monitored period we observed that the minimum time required to a block to be mined is about 2 minutes, while the maximum time is 77 minutes. Also, the medium time for discovering a block is about 9 minutes and the 50% percentile is about 6 minutes.

Table 5 and 6 show respectively the protocols and the Bitcoin client used by the nodes of the network.

Table 5: Bitcoin Protocol version used by nodes

Protocol	number of clients
70012	6655
70002	3013
N/A	771
7000	1153
70013	88
70010	78
80001	71
70011	68
80000	24
99999	4
50400	2
70014	2
60000	1
70003	1
60002	1
80002	1

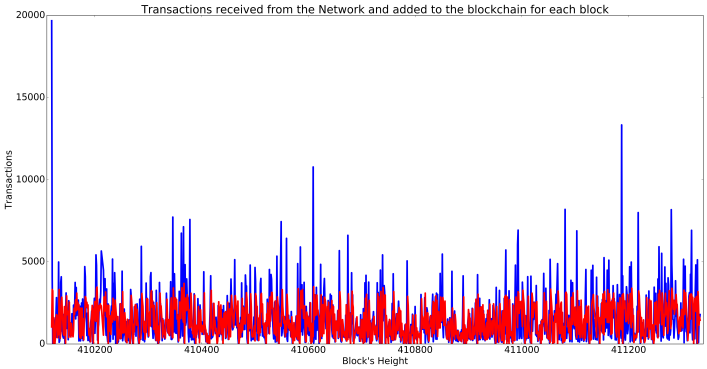

Figure 4: This figure compare the number of Transactions received by our client per block (in blue) and the number of transactions included in the Blockchain during the same block mining time (in red).

Table 6: Bitcoin Client Software used by nodes

Color	Bitcoin Software and Version	Number of clients
firebrick	Classic:0.12.0	2969
indianred	Satoshi:0.12.1	1790
hotpink	Satoshi:0.12.0	1691
gold	Satoshi:0.11.2	1323
lightsalmon	N/A	771
mistyrose	Satoshi:0.11.0	368
greenyellow	Satoshi:0.10.2	233
fuchsia	Satoshi:0.11.1	226
indigo	Classic:0.11.2	184
lightcoral	Satoshi:0.12.99	144
darkolivegreen	Satoshi:0.11.2(bitcore)	142
dimgray	Satoshi:0.9.3	122
ivory	Satoshi:0.10.0	116
skyblue	BTCC:0.12.1	93
lightseagreen	Satoshi:0.8.6	72
papayawhip	Satoshi:0.9.1	71
burlywood	BitcoinUnlimited:0.12.0(EB16; AD4)	70
blanchedalmond	Satoshi:0.10.1	67
pink	Satoshi:0.9.2.1	56
white	Satoshi:0.8.5	42
darkslategray	Bitcoin XT:0.11.0D	29
orangered	Bitcoin XT:0.11.0	26
peru	Bitcoin XT:0.11.0E(Linux; x86_64)/&22',	22
brown	Satoshi:0.8.1	20
seashell	Bitcoin XT:0.11.0E(Windows; x86_64)	19
lightskyblue	Satoshi:0.11.99	18
gray	Satoshi:0.12.0/Knots:20160225	18
olive	Bitcoin XT:0.11.0C	17
darkseagreen	BitcoinUnlimited:0.11.2	12
mediumorchid	Satoshi:0.10.4	12
lawngreen	Satoshi:0.8.3	12
crimson	btcwire:0.4.0/btcd:0.12.0	11
palegreen	Classic:0.12.1	10
antiquewhite	Satoshi:0.11.2/ljr:20151118	10
tomato	Satoshi:0.9.5	8
yellow	Statoshi:0.12.99	8

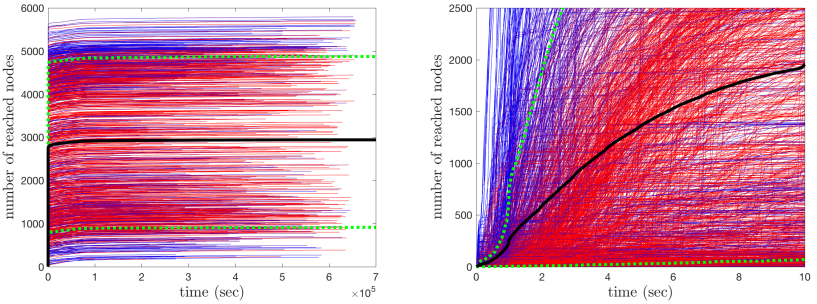

Figure 5: Number of nodes reached by a new valid block before a following block is discovered (left). The color are associated with the size of the block with blue being smaller and red larger. The black line is the average of all the observation and the two dashed green lines are respectively the 10% (lower) and 90% (upper) percentiles. The right plot is a detail of the initial propagation within the first 10 seconds.

Chapter 2

Facebook as microeconomic Data Source

2.1 Introduction

With the increasing and extensive adoption of web 2.0 and Social Networks, the Internet has become a very rich source of different kinds of freely and publicly accessible data and contents. Nowadays it is possible to access data describing financial and economical statistics, and, more importantly, data pertaining to news and social data. Examples of the latter are represented by web queries and visited places which are picking up interests among researchers interested in developing a better understanding of users' habits and building profiling models. The web surfing experience for a user is not constrained anymore or limited to the mere retrieval of information, but it also produces in itself a great deal of information. This kind of data about user activity ranges from shopping preferences to mobility habits, musical tastes and smartphone applications. It is not surprising that, given the amount of information they can reveal, these data are used by companies like Google or Netflix to forecast user preferences and to suggest potentially interesting contents, increasing in the revenues from their distribution. Recently, scientist have also exploited the idea of social media fingerprint, defined

by all the users activities, using it also in study related to unemployment [52], consumption indicators [62] and new car sales trend [17]. The aim of this work is to study whether and how geo-localized data from Facebook Places (likes, checkins and talking about counts) allow to discover universal patterns which enable to measure economic activities or to nowcast local economies for a certain area. Economical Information about consumer spending, employment and mobility regarding a geographic area are very important in order to estimate the economic activities but probably their main drawback is represented by their time delay which is usually greater than three months. Specifically, I will investigate whether social data can be used as economic indicators in order to forecast the economics activity of a certain county/city/tract or if this data could be paired with actual indicators in order to increase their quality and to anticipate possible trends. To achieve that, dataset from Facebook will be paired with data from the United States Census Bureau (US Census). This chapter contains an explorative analysis on Facebook data available on the data framework in order to show its potential for economics purposes.

2.2 Methods

The aim of this work is to investigate the possibility of nowcasting economical indicators through social networks data. Data used are collected from the US Census, as micro-economic source, and from Facebook. A relevant geographic sample consisting in more than 100 counties and about 5700 tract is selected as described on chapter 1. This geographic sample is used in order to retrieve all the Facebook Pages located inside the specified area. This allow also to understand how pages categories are distributed among each city or area in order to define patterns and recording their everyday actions. The dataset includes two random counties from each State of the United States [1] in order to contains in-

[1] Regads Alaska were excluded North Slope, Northwest Arctic, Yukon-Koyukuk, Nome, Wade Hampton, Bethel, Dillingham, Aleutians West, Aleutians East Denali, Fairbanks North Star, Kenai Peninsula, Lake and Peninsula, Matanuska-Susitna, Skagway-Yakutat-Angoon, Southeast Fairbanks and Valdez-Cordova because their size is not comparable

formation related to Places from 101 counties of the United States (see tables 1 and 2 for the full list). Starting from the counters is possible to built time series related to a specific place or within a group of places relative to the same category. All these data collected through Facebook will be studied in pair with the U.S Census data [68] or with any other economic or financial dataset which provide a geographic locations. As it was already discussed, U.S. Census provides economical and demographical data related to U.S. territory. Data from U.S. Census are obtained through surveys and they suffer of an important delay, with a minimum of three months. Due to that nowcasting studies become very important for economist in order to estimate now actual indicators that will be obtained in the future. This is a preliminary work, showing only information about the Facebook dataset. Future developments will focus on analyse if there are any relationships between Facebook data and Economical/Financial data.

2.3 Results and Discussion

Data were collected in a period starting on Feb 23, 2015 and stopped on July 1. Data collected during this temporal window contains more than 454 thousands places on Facebook belonging to the selected counties set. For each place it is possible to obtain information related to its own category or sector and also the counters increments due by users interactions. The first application of the framework could be to use Facebook data as a measure of the urban growth of a specific county, and relating it with the economical data obtained by Census data. Economic Indicators are very useful for drawing a picture of the overall economic activities of a geographical area. They are basically sentiment indicators made up through household surveys, so they present some drawbacks, such as temporal delay and high costs. The time delay and the high costs depend on all the operations behind organization of surveys. Usually economic indicators come out with several months of delay. In U.S. this delay is generally near to three month. Hence, when they become available, they

with others counties.

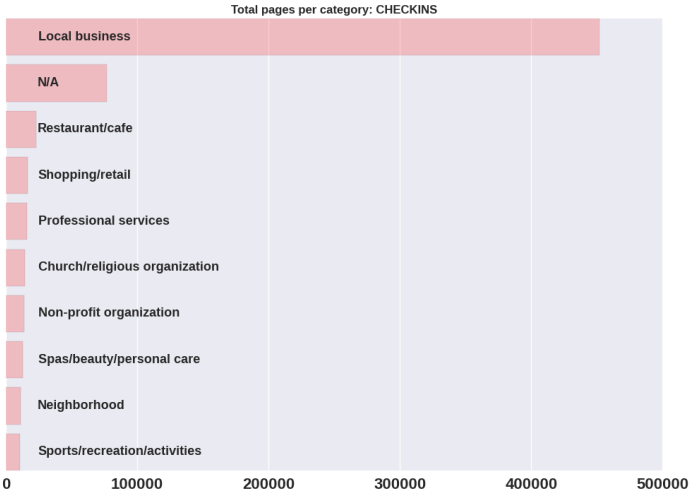

Figure 6: Total number of checkins grouped by pages' category

refer to three month in the past and not to the actual economic situation. For this reason, the field of forecasting current unobserved consumption (nowcast) is growing at a higher peace. Furthermore, there is no guarantee that these indicators represent a true measure of the regions' economy, first because they are the sentiment of householders and also since they don't include shadow economic activities. The advantages of using Facebook data are principally related to their nature: public, with a high resolution and available in real time. Moreover, as already described, Facebook Places provide a lot of useful information like their classifications, the price range and, of course, their geographical location. The pages' counters allow to built time series for a Place or grouping them by sector. The main hypothesis is that social actions flows on private consumption business are related to economic indexes provided by the U.S. Census Bureau. Following this rationale, it will be also possible to obtain similar economic indexes behind counties that present analogous patterns or overlap of shops distribution and social activities. Facebook

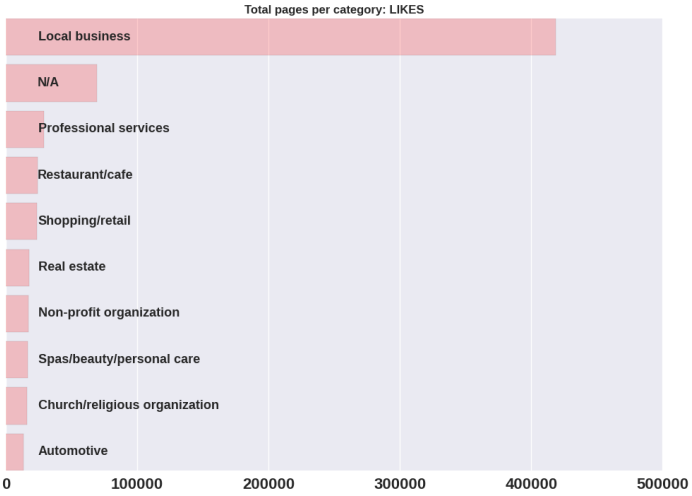

Figure 7: Total number of likes grouped by pages' category

Search API provides also information related to the page and some counters whose value represents the total interactions with that place at the actual time. Through the use of time series it may be possible to evaluate the interaction rate with the palaces during a certain period for a specific region. By comparing these trends it may be possible for example to identify tourist areas and whether there are peaks in a certain period. Furthermore, analysing the statistics with respect to the average level of activities inside a county, in some cases it should made possible to define the source of the mobility flow. It could be also possible to couple Facebook flows with others social geolocalized stream, such as Twitter and made comparisons among them. Furthermore, as discussed in [63] estimating shadow economic activities is crucial for countries however, it is very difficult to get accurate information about the phenomena because involved individuals don't wont be identified. Shadow economic activities are in part absorbed by regular businesses, this because, part of the shadow cash is spent in regular stores or shops. Other examples

involve money laundering where a regular business is involved in producing false invoices in order to justify an amount of money obtained by crime activities with the Internal Revenue Service. In these cases, for example, it could be identified an high online activity in a non tourist or poor area, or vice versa, that does not correspond to the economic expected indexes of the area. Facebook Data may be useful to track a profile for each county and may also enable to extract some patterns in order to define the economies' eco-systems. Under certain parameters, such as population or demographics extension, these patterns could result coherent or suspect compared to the characteristics of the territory. Lets think for example to a pattern for a tourist area. It is expected a spread of restaurants and hotels, and also a big online activity among them while, a business area may presents different patters because hotels and restaurants may be less diffuse. All these kind of networks could be compared among them and, also, against the U.S. Census data in order to discover if there are some anomalies or some frequent patterns/structures. Social data collected through Facebook for same sectors could define a pictures of the actual economy, especially regarding places or firms that have invested in their brand to make it a "status symbol". Considering this kind of places, users usually posts their recent activities on the social network in order to celebrate their new purchase or a dinner in their loved place, such as restaurants, luxury shops and so forth. On the contrary for some other shop categories is not common to have an online activity but they could have registered a Facebook page in order to promote their business. In this case it is still possible to know the business participation in the area, and try to get some structural patterns. As future works it can be useful to evaluate the post of the pages and also the users' posts. The former can be considered as a measure of the marketing investment on Facebook, in terms of money if the page has an advertisement campaign whereas the post leaved by users could be analysed with sentiment analysis in order to understand the individuals feeling with respect to the brand/shop/page. In figures 6 and 7 are showed the total amount of checkins and likes, respectively, for the ten main categories belonging to the data downloaded on the reference period. As it is possible to see,

on the figures, Local Business, Restaurants/Caf and Shopping Retails are the categories with highest checkin rate while Local Business, Professional Services and Restaurant/Caf have the highest likes rate for the reference period. Different geographical areas could have several shop categories distribution and also could present other categories with highest rate. Tourist area could get a peak on shops and restaurants compared to industrials area which could instead have factories. Figure 8 show

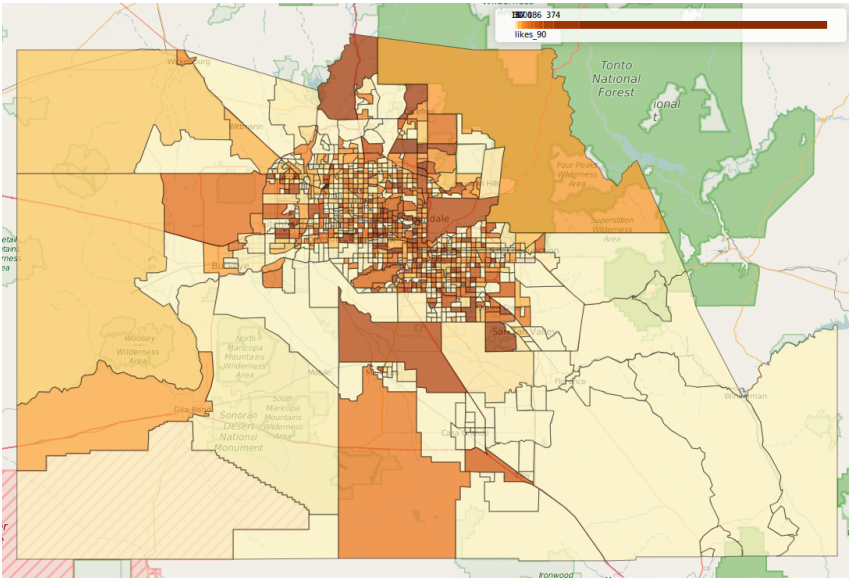

Figure 8: Likes rate for tracts inside Maricopa county (Arizona) belonging to Social Business category

the likes rate for Pages belonging to Social Business category for Maricopa county in Arizona. Each tract group an area with inside about 8 thousands people. It means that a smaller tract has an higher density of population compared with one bigger. The color on map represent the 90th percentile of all the page rates, belonging to Social Business category. Analogous figures are also available for District of Columbia9), Honolulu10 and Harford11. Analysing this kind of structure for each county could be useful for several reasons. The first is to analyse com-

petitors to better understand whether and how to invest on a new business activity in that area. These data could be very useful in order to understand the actual situation of the area.

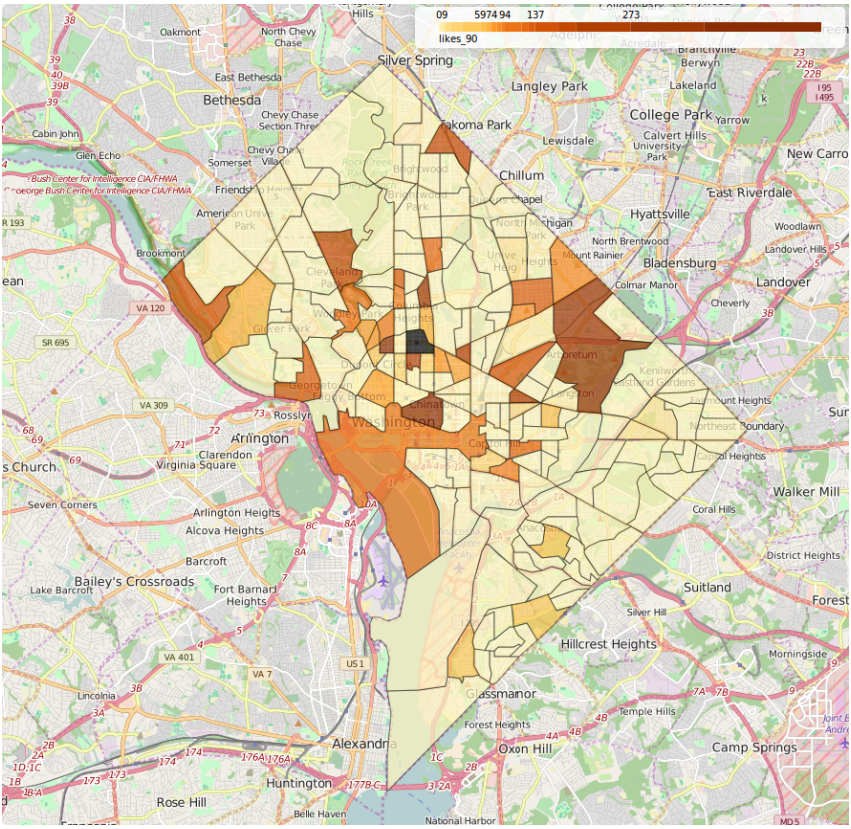

Figure 9: Likes rate for tracts in District of Columbia belonging to Social Business category

In particular, it may be possible to identify, for a specific category, if the area is already saturated or the demand is not or partial satisfied. This kind of study could be also applied to an existent business activity which wants to estimate the environment in order to understood its potential grow rate. Could be also useful to track people mobility, because, in case

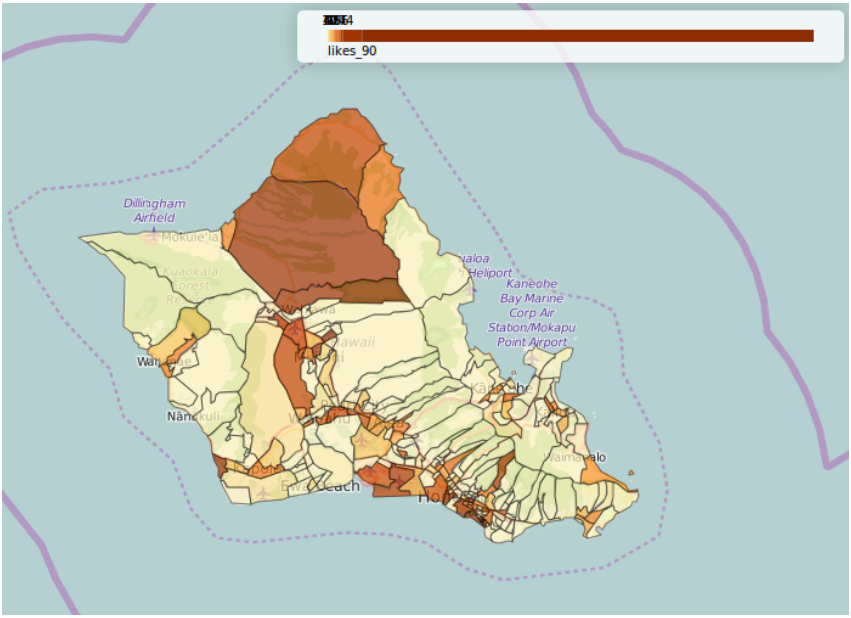

Figure 10: Likes rate for tracts in Honolulu (Hawaii) belonging to Social Business category

of lack of same categories of shops people are obligated to search in the near counties. This means that, according with the business structure of an area, if there is lack of variety, it may be possible to observe several mobility patterns of the population of that area to some nearby places.

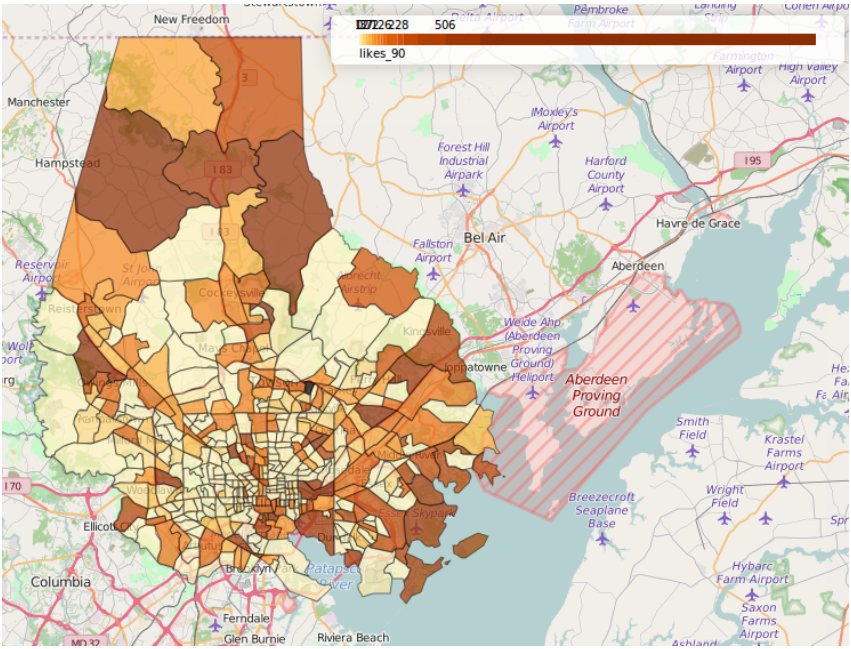

Figure 11: Likes rate for tracts in Harford (Maryland) belonging to Social Business category

Chapter 3

The Accounting Network

3.1 Introduction

Network Theory has been used to establish how contagion, through a variety of channels (mutual exposures, social networks of board members, moral hazard from permissive regulations, financial instruments like swaps and derivatives, etc.), triggered the outbreak of the 2007-08 crisis. Scholars suggest that financial systems may affect positively economic development and its stability [19; 20; 51], although they may represent a source of distress which leads to bank failures and currency crises, or greater contraction for those sectors that depend more on external finance during banking crisis [34; 59]. As a response to the recent financial turmoil, the banking sector has been affected by a substantial reorganization [24]. For instance, as highlighted by the European Central Bank for the Euro area *the main findings reflect the efforts by banks to rationalize banking businesses, pressure to cut costs, and the deleveraging process that the banking sector has been undergoing since the start of the financial crisis in 2008*[38]. This implies that market pressure and regulatory amendments induce banks to reduce their levels of debt, through cost containment and stricter capital requirements. In addition, a gradual improvement in bank capital positions aims to enhance the capacity of the system to absorb shocks arising from financial and economic distresses. This limits

the risk of spillover effects from the financial sector to the real economy and put the financial system in a better condition to reap the benefits of economic recovery. In particular, as the financial boom turned to a bust, banks' stability deteriorated abruptly and the economy entered a *balance sheet recession*, which depressed spending levels through a reduction in consumption by households and investments by firms. Therefore, although at an uneven pace across regulations, the need to strengthen fundamentals has influenced the banking sector, and differences in banks' portfolio allocations, financial performances, and capitalizations might be interpreted as the combined results of policy decisions and sectoral responses to changes in the regulatory framework (see e.g. [9; 36]).
This work relates to the literature on banking development and performance evaluation during the recent crisis (see e.g. [7; 22; 30]). It is considered a large data set of worldwide banks retrieved from *Bloomberg*, focusing on financial statements spanning from 2001 to 2013. It will be introduced a network based on similarities between banks' financial statement compositions (hereinafter *Accounting Network*). Due to data limitations, the reference sample is restricted to banks for which a continuum and stable set of variables is available for the entire period. The introduction of a methodology (*Quality Ratios*) to measure banks' data coverage aims to prevent that missing values for some variables or lack of annual financial statements for some banks affect the overall picture. Then, it would be exploited the maximum amount of available information from financial statements without further reducing the set of variables through an arbitrary selection of the financial statements fields. This choice aims to avoid any selection bias. Moreover, total assets (as a proxy for size) for each bank is applied to normalize banks' financial statements measures to prevent the emergence of "size effects" as the sizes of institutions are spanning for various orders of magnitude.
The introduction of Accounting Networks establish a bridge between the external perspective arising from market data and the internal one based on banking activities indicators. It is studied how Accounting Networks can be exploited to provide a description of the banking system during the crisis. This part sheds light on whether banks under different regula-

tory frameworks and diversification degrees have reacted to the crisis by strengthening their business peculiarities or by converging towards similar practices ([21; 35]). The main assumption of this work is that market data alone, although highly representative of investors' perception of the banking sector, might be dis-informative during periods of distressed market conditions. This, in turn, stimulates a broader exploitation of the information on banking activities, thus pointing to a more comprehensive investigation which takes into account also the internal perspective arising from financial statements data. In addition, the use of accounting data allows a partition of business activities where banks are involved in, providing therefore an approximation of the state of the system related to several potential channels through which the financial distress might propagate. This is appealing also for regulators, since authorities are interested in a wide set of economic indicators in order to prevent the systemic relevance of financial institutions and they introduce specific requirements and constraints which affect directly financial statements measures. For these reasons, enriching the debate on financial stability by means of the Accounting Networks might give new clues about the resilience of the banking system.

Another important result is the possibility of getting a neutral partition of banks in "network communities" (i.e. clusters) that results from the analysis of the network through community detection algorithms like the *Louvain* modularity maximization. The results indicates that regional communities evolve in time and the crisis has a clear role in weakening geographically determined structures. Furthermore, it is focused on proxies for leverage, size and performance in order to understand if these variables have played a key role among the set of economic measures usually applied to classify banks (see e.g. [27; 49]). Hence, it would aim to answer the question whether the collapse of financial markets has weakened these relationships, limiting therefore the power of traditional indicators to identify clusters of homogeneous banks. Correlation diagrams applied to show how network variables are related to economic measures suggest a turning point in correspondence of the outbreak of the crisis, which influenced the role of proxies for leverage, size or per-

formance to group similar banks. This preliminary results motivated the last section, where by means of Principal Component Analysis it is investigated which economic features are more likely to characterise the heterogeneity of the communities before, during and after the collapse of 2007-08.
The remaining part of the work discusses open issues and future lines of research, such as open questions on how to improve the building of the Accounting Networks. In particular, the effectiveness of this approach can be enhanced by means of a careful variable selection based on the best financial practices applied in the evaluation of the financial statements structures. In addition, a more accurate normalization of the variables and caring about national regulations may increase the usefulness of the methodology. Furthermore, matrix filtering techniques and missing data reconstruction for financial statements information can enhance the extraction of meaningful clusters. Then, more advanced and focused tools could be conceived to analyse banks evolution towards similar business configurations or, alternatively, their divergent patterns as a response to changing market conditions.

3.2 Methods

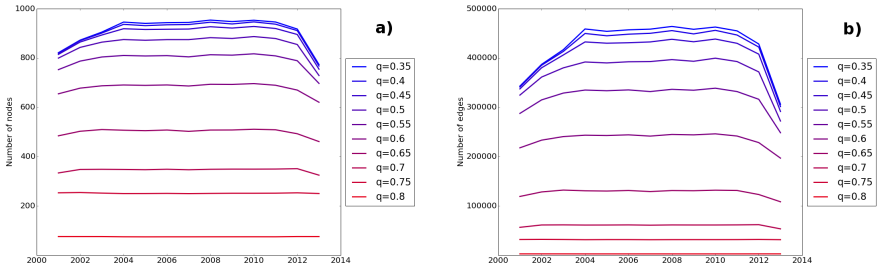

Figure 12: This picture shows the number of nodes and edges along the sample period for different QR values. It is clear to see how for small values of the Quality Ratio parameter the curves belong to a stricter range.

3.2.1 Accounting networks

For every year a vector of financial statement variables is assigned to each bank and used to compute the cosine similarities between pairs of banks/nodes. Here the intuition is that the most similar banks (as from their financial statements) must stay closer in the network and form a cluster. Then, the measure "cosine similarity" is transformed into a metrics (as triangular inequality must hold, the square root is used). The definition is the following: the cosine of the angle between each pair of vectors with the dot product is computed and then it is applied the simple transformation $w_{i,j} = 1 - \sqrt{1 - C_{i,j}^2}$, where $w_{i,j} \in [0, 1]$ and $C_{i,j}$ is the cosine similarity between *i* and *j*. In network terms $w_{i,j}$ is the weight. This transformation (see [69] for an introduction to similarity measures and relative metrics) is used to move from the cosine similarities defined in the space [-1,1] to weights in the interval [0,1]. With this transformation the more two nodes are similar (or anti-similar) the larger is the weight, while a weight of 0 is assigned to a pair of nodes having totally dissimilar financial statements (actually, in our networks cosine similarities range mainly between 0 and +1).
In addition, before the computation of the metrics, it is necessary to take care of the size distribution of banks, as it spans over several orders of magnitude. To avoid a bias toward large institutions for each bank, all variables in its vector were divided by the respective total assets in such a way that the attributes of the vector refer to economic and financial *ratios*. This operation ensures that clusters will be formed by banks with similar business activities regardless their sizes.
An important methodological choice of this study is the "neutral" approach used for the selection of the variables within the financial statements. A part from removing related and redundant measures (total and subtotals), all the available information were used, applying the same weight to each variable in the vectors. This agnostic approach is in line with the goal, introducing the concept of Accounting Network, although practitioners can give a different importance to each variable of the financial statement. In this perspective it is expected that the relevant in-

formation will emerge in a bottom up process, as a spontaneous feature selection carried by the methodology. Finally, a confidence level (95%) is introduced during the link formation. By using a Montecarlo sampling test, if the cosine similarity is statistically significant with 95% of confidence the link is chosen, otherwise discarded. As a result of this filtering procedure, it was observed that the networks tend to be very dense and almost complete. The most of the information is carried by the weights of the links and less by the simple topology (degrees and other structural features).

3.2.2 Community detection

A classical method to investigate the structure of a network is the search of communities, i.e. regions of the network with larger *internal* links density. Intuitively, these regions are formed by clusters of nodes with higher degrees or, for weighted networks, with larger strengths. Several methods were proposed to find network communities without imposing a priori the number of communities but letting them emerging from the network itself. Among others we cite the optimization of the modularity that is a measure of how much the link structure differs from the random network where links are assigned with uniform probability and internal communities are not present (a part from fluctuation). For weighted networks, the modularity is defined by the following formula:

$$Q_w = \frac{1}{2W} \cdot \sum_{ij} \left(w_{ij} - \frac{s_i s_j}{2W} \right) \delta(c_i, c_j) \tag{3.1}$$

where $s_i = \sum_j w_{ij}$ and $s_j = \sum_i w_{ij}$ are the strengths (sum of weights) of the nodes i and j respectively, $W = \sum_{ij} w_{ij}$ is the total sum of the weights and the function $\delta(c_i, c_j)$ is equal to 1 if (i, j) belong to the same community or 0 if they are members of different communities. The maximum modularity value is 1 (an ideal case for which the communities are isolated) and can also take negative values. The 0 value coincides with a single partition that will correspond to the whole graph. A negative value means that there is no particular advantage in separating the nodes in that particular clusters and so there is not community structure

whatsoever.
To study the presence of communities it is often necessary to prune the network cutting the links if their weight is below a certain threshold. In this case only the links formed by nodes having a large similarity/weight w of their financial statement vectors were considered. The procedure of pruning can be guided by the use of the tools related to the community detection methodology [41]. In particular, working with the modularity optimization function [56], with the Louvain technique [26], it is possible to look at the *significance* associated to the threshold (as in [67]), where the modularity is introduced as a parameter to check for the best resolved community structure. This parameter is used to help finding a reasonable pruning threshold range of values for the networks. A rule of thumb in this process is indeed avoiding network fragmentation, i.e. keeping the graph connected while removing not significant links. Extensive tests are done, computing quality/significance of the partitions (looking at the modularity parameter) using different pruning thresholds (i.e. removing the links having a low weight), determining a range of weights thresholds ($0.35 < w_{i,j} < 0.5$) that helps to prune the original networks to an optimal level. In this interval, communities are stable and the interpretation of each region can be seen as a result of the financial statement similarities across banks in different countries.

3.2.3 Network measures vs. Economic indicators

Comparisons among network measures and economic indicators are provided to describe the correlation between nodes' network topology and economic behaviour. These features are studied by means of extensive linear correlation tests (Pearson correlation) for the overall set of banks for each year and it is verified the significance of the estimates by means of parametric tests. These estimates are based on the filtered networks, which are themselves based on the significance and the quality of the community detection algorithm. This analysis shows how nodes' network properties (e.g. *Strength* or *Clustering Coefficient*) are associated to basic economic indicators (e.g. *Return on Assets*, *Total Assets* and *Total*

Debts to Total Assets), thus showing whether nodes' topological properties are positively or negatively related to certain economic features and how these relationships have weakened or reinforced during the crisis.
Clustering coefficient is a measure of the local tendency of the nodes to form small regions of fully connected nodes, it is an average measure of the local clustering coefficient (actual number of triangles centered in each node over the total). Return on assets (ROA) is the net income over total assets and is a measure of the bank performance. Total debts to total assets is an indicator of the leverage of the bank and it is computed as the ratio between debts and its size (measured by total assets).

3.2.4 Principal Component Analysis

Once communities are identified, the further step attempt to describe which financial statement variables are more likely to characterise these clusters. In order to facilitate comparability, it is focused on those indicators more popular within the set of variables utilised to compute the cosine similarities (i.e. those indicators appearing with larger frequency in the entire dataset). In fact the inclusion of very poorly represented measures across different banks would have made the comparisons less effective with potential biases related to e.g. different regulations frameworks or geographical memberships. Hence, since the interest on disentangle potential similarities/peculiarities across different communities, it is preferred to rely on common and well-diffused measures of banking activities among those present in banks' financial statements. In addition, this set is enriched by means of indicators such as ratios (e.g. *Return on cap* and *Total debts to total assets*) and aggregated measures (e.g. *Total assets*). Community detection identifies four main clusters, whose constituents are more numerous and stable in time. For the sake of conciseness, the *Result* section will focus mainly on these communities. In particular, for each year it is described by means of Principal Component Analysis (PCA) which economic features are more (less) able to contribute to the explained variability of communities' members.
PCA is a multivariate technique that analyses observations described by

several inter-correlated variables. PCA extracts the important information from the data and expresses it as a set of new orthogonal variables (principal components). In our exercise, since measures present different ranges of dispersion (e.g. by construction some ratios are bounded) we rely on a scaled version of PCA; finally, we consider only principal components with eigenvalues greater than 1 (in almost all cases they correspond to the first 3 components). Then, it is computed the proportion of the variance of each original economic measure that can be explained by the selected principal components. This, in turn, leads to a ranking of the original economic measures in terms of their ability to describe a certain community's variability. In particular, since we are interested in how the onset of financial crisis has affected the banking system, we split this analysis in three periods: from 2001 to 2006 (before the crisis), from 2007 to 2009 (the onset of the crisis), and from 2010 to 2013 (after the breakdown of the markets). For each period, communities are characterised by the top and the bottom three measures, thus analysing how these ranks have evolved over time and across communities.

3.3 Results

This section shows how Accounting Networks represent a complementary technique to traditional financial networks for the study of the banking system.

While financial networks reflect the view from the market, related to e.g. the pairwise correlations of stock prices, Accounting Networks capture the effects of business decisions on financial statements measures and on business models of different institutions. An "ideal" investigation of the financial system would involve also a detailed analysis of the money flows among companies, which determine the so called "mutual exposures" (an important contagion channel). Unfortunately, these high granular and detailed data are usually not available. However, financial statements provide an aggregated view of mutual exposures and obligations for different maturities and types of instrument. This is an important point in favour of Accounting Networks as they report summarised

information for e.g. phenomena occurring with different time scales and contractual terms, as opposite to the financial networks that rely only on homogeneous (daily or intraday) market data.

3.3.1 Community Detection Results

This sub-section focus the attention on the bottom up clusterization of the network from the application of the community detection algorithm and on the presence of geographical structures arising when each bank is labelled with its country. Therefore, whether banks belonging to different countries (as a proxy for different regulations and/or level playing fields) have shown the tendency to be part of separate or, alternatively, common clusters they were verified, by analysing communities' evolution over time, whether the crisis influenced these configurations. In particular, the community detection analysis on Accounting Networks shows these main results.
It exhibits the presence over time of a clear community representing US banks and another one composed by Japanese banks, although for both regions there is also an additional smaller second group quite persistent in time. By contrast, it is not possible to identify a single and an unambiguous European community, since banks belonging to European countries seem to be likely to form national or sub-regional communities or to be included in a vast and geographically heterogeneous cluster (hereinafter the *Mixed* community). In addition, Asian banks are fragmented in several sub-regions where, in particular, the Arab and the Indian-Pakistan groups emerge. Therefore, the detection of communities within Accounting Networks reveals the presence of two homogeneous clusters corresponding to US and Japanese banks surrounded by a more diversified cloud of banks belonging to different countries; remarkably, European banks are not able to clusterise together in a single community, while it persists over time a certain level of separation based also on national borders. Hence, an interesting contribution of the Accounting Networks points to the presence of a large and geographi-

cally heterogeneous community, which can be related to the fact that the globally established regulatory framework might have indeed accelerated the tendency of banking activities of different countries to converge into more uniform banking practices. This is shown for instance in Figure 13 where it is also possible to observe that the outbreak of financial markets contributed to make the Mixed community more cohesive; furthermore, although still representing separate communities, both US and JP clusters result topologically closer to the Mixed community after the breakdown of 2007-08, thus supporting the interpretation of a gradual convergence of different areas into more similar patterns. In addition, the application of the community detection on Accounting Networks allows to identify even small communities, such as those related to African or Scandinavian banks. This represents a quite promising aspect of the methodology, since it ensures the detection of local reliable communities although the approach taken so far is eminently agnostic.
It is not simple to explain the reasons behind the emergence and evolution of these communities, however it is possible to advance some intuitions based on the impact of globally recognized accounting standards [40], the establishment of supranational supervisory and regulatory authorities, and on the role of the harmonization process of banking practices which have been implemented through e.g. the various Basel regulations [25]. These contributions point to a common level playing field, which might have facilitated the emergence of a large and geographically heterogeneous community and its increasing topological proximity to both US and JP clusters. However, latter communities highlight the persistence of regional peculiarities. In Japan a deregulation process, known as the 'Japanese Big Bang', was formulated during the 1990s to transform the traditional bank-centered system into a market-centered financial system characterised by more transparent and liberalised financial markets [48]. In fact, peculiar features of Japanese banking sector were the over-reliance on intermediated bank lending, the absence of a sufficient corporate bond market and a marginal role for non-bank financial institutions, whose main consequences were an abundance of non-performing loans, excess in liquidity, scarce investments and low banks

profitability (see e.g. [18]). Although this program was intended to cover the period 1996-2001, the goals have not yet been achieved and policy makers' continuing reform efforts to remove past practices by market participants confirm the slowing convergence of the Japanese regulatory framework to a capital-market based financial system [12]. Thus, the presence of the JP community which gradually tends to the Mixed cluster is in line with evidences from the Japanese financial sector reforms aimed to change its reliance on indirect finance into a system of direct finance related to capital markets. Furthermore, it is remarkable the presence of a US community quite stable over time, which seems to be progressively attracted by the Mixed cluster. The US financial system presents peculiar features compared to other geographical areas. It is characterised by a relatively greater role of capital market-based intermediation, a higher importance of the 'shadow banking system', and differences in the accounting standards [38]. The impact of non-bank financial intermediation relates to the use of originate-to-distribute lending models, which determine the direct issuance of asset-backed securities and the transfers of loans to government-sponsored enterprises (GSEs, e.g. Fannie Mae and Freddie Mac). Financial innovation played a key role and the increasing use of securitisation explains the low percentage of loans to households on banks' balance sheets [38]. In addition, the US 'shadow banking system' is highly dependent on the presence of finance companies, money market funds, hedge funds and investment funds, which influenced the growth of total assets in the US financial sector during the last decades [57; 65]. The presence of a distinct community is probably also due to differences in accounting standards which mainly involve the treatment of derivatives positions between the US Generally Accepted Accounting Principles (US GAAP) and the International Financial Reporting Standards (IFRSs). In particular, US GAAP allows to report the net value of derivative positions with the same counterparty under the presence of a single master agreement, thus impacting on the size representation of balance sheets items. However, in Figure 13 we observe that the US community (similarly to the JP community) is gradually approaching the Mixed community, and the consequences of the

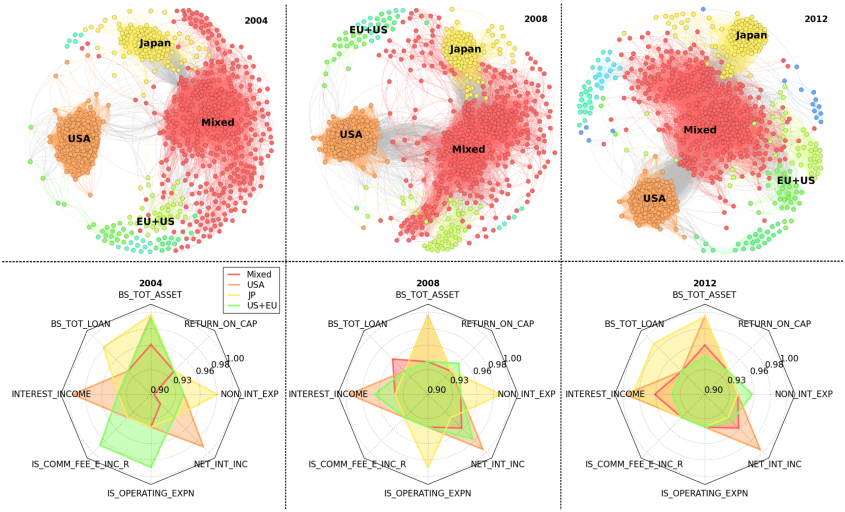

Figure 13: In the upper panels it is shown the Community Structure for the three periods. The impact of the financial down-turn of 2007-08 seems to be reflected more heavily after the crisis, with the emergence of many sub-region communities as a response against the deteriorated market conditions. In the lower panel the most important financial statements components by the PCA analysis.

breakdown of 2007-08 seem to have enhanced this behaviour. Among the possible several reasons, it is worthwhile to consider the impacts of the reform on the OTC derivatives market (embedded in the Dodd-Frank Act) and the Basel III new banking regulation, which may have facilitated similarities among US institutions and their peers in the Mixed cluster.

3.3.2 Relationships between Economic Indicators and Network Properties

In this Section is provided a preliminary investigation of the relationships between banks' economic indicators and their network properties. In order to characterise banks, three common proxies are considered for their classification: *Return on Assets* (for the *Performance*), *Total Assets* (for the *Size*) and *Total Debts to Total Assets* (for the *Leverage*). Then, comparisons are presented against two basic network measures: the *Strength* and the *Clustering Coefficient*. For each year from 2001 to 2013, some insights for these relationships are provided by estimating for the overall sample the correlations between banks' economic indicators and network measures. As explained in the Method, in this exercise is considered the network filtered according to the quality/significance of the *Louvain* community detection algorithm. This helps to assess the significance of results. Below, are showed some examples to discuss how these relationships have evolved over time.

In particular, it is investigated whether once the effects of the crisis have spread throughout the financial sector, the capacity of traditional economic indicators (e.g. leverage, size, performance) to group banks could result undermined. For instance, the onset of the financial crisis clearly affects the relationships between *Total Debts to Total Assets* and network properties. Although the correlation between *Strength* and *Total Debts to Total Assets* remains negative during the entire sample period, the breakdown of financial markets seems to further enhance this effect for subsequent years (Figure 14, plot on the left). Thus, this relationship

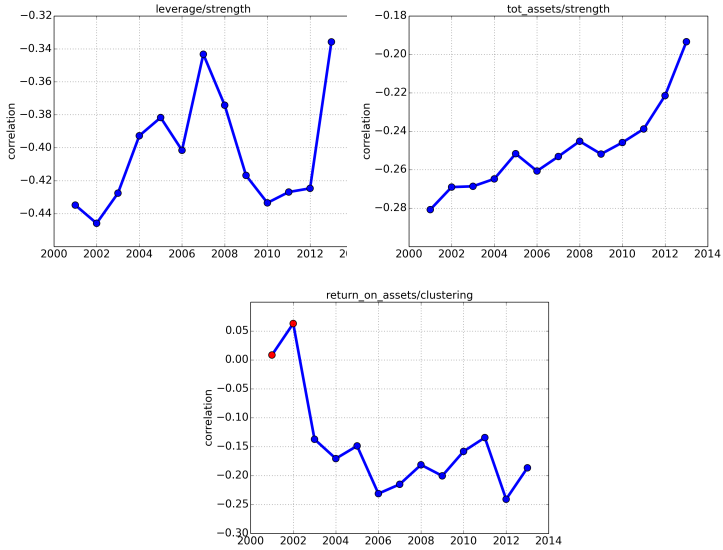

Figure 14: In these plots we present the correlations between banks' Strength versus the Total Debts to Total Assets (Leverage) (plot on the left), Strength versus Total Assets (Size) (plot on the middle) and Clustering Coefficient versus Return on Assets (Performance) (plot on the right). The correlation is computed across the years 2001-13. It is clear the effect of the financial crisis across the outbreak of 2007-08. Red points stand for no-significant estimates, while blue points refer to significant estimates.

suggests that, after the onset of the crisis, the use of leverage became on average more anti-correlated to the *Strength*. This implies that banks that are more dissimilar in terms of their financial statements (i.e. with lower values of *Strength*) are those that turned out to be less capitalised (i.e. with higher values of *Total Debts to Total Assets*). Furthermore, one might be interested in understanding the role played by the *Size* which represents a typical indicator utilised to classify banks. The correlation between *Strength* and *Total Assets* is almost flat and negative even after the collapse of 2007-08, but it shows an increasing trend in the recent period (Figure 14, plot on the middle). Hence, it seems that after the outbreak of the crisis the *Size* became less correlated to the similarity among

banks, as estimates pointing sharply to zero seem to suggest. Finally , it is analysed the relationship between *Performance* and network properties (Figure 14, plot on the right). In particular, in order to mimic how the presence/absence of more connected groups of banks is related to economic results it is considered the *Clustering Coefficient* for determining the level of structure in the system. Although poorly statistically significant in the early 2000s, correlations with *Return on Assets* exhibit a decreasing pattern before the onset of the crisis and then remain negative although slightly erratic. The negative relationship between *Clustering Coefficient* and *Return on Assets* seems to suggest that the presence of well connected areas in the network (nodes with higher clustering coefficients) do not foster economic performance.

These basic examples suggest that a clear investigation on the relationships between economic indicators and network properties might be not always conclusive. Moreover, it is considered the entire set of banks, there might be some cases where estimates are poorly significant. Still, some remarkable effects arise from this investigation strategy and preliminary results point to a turning point in the correlations across the outbreak of the financial crisis. In particular, diagrams confirm that leverage is an useful indicator for differentiating banks, hence deviations to a lower capitalization are associated to increasing dissimilarity with the rest of the system and the impact of the crisis suggests a reinforcement in this relationship. By contrast, it seems that size does not contribute too much on the similarity between banks after the breakdown of 2007-08, while it played a greater role before and during the crisis. Finally, the relationship between performance and the structure of the system is less clear and prevents straightforward conclusions.

The identification of economic features potentially able to characterise specific portions of the system is addressed in the next sub-section.

Community	Top Measures 2001-06	Values 2001-06	Top Measures 2007-09	Values 2007-09	Top Measures 2010-13	Values 2010-13
C0	BS_TOT_ASSET	0.9695	BS_TOT_LOAN	0.9588	INTEREST_INCOME	0.9635
C0	BS_TOT_LOAN	0.9257	NET_INT_INC	0.9529	NET_INT_INC	0.9620
C0	INTEREST_INCOME	0.9228	BS_TOT_ASSET	0.9492	BS_TOT_ASSET	0.9537
C1	INTEREST_INCOME	0.9955	BS_TOT_ASSET	0.9964	INTEREST_INCOME	0.9968
C1	BS_TOT_ASSET	0.9951	INTEREST_INCOME	0.9957	NET_INT_INC	0.9954
C1	NET_INT_INC	0.9917	NET_INT_INC	0.9933	BS_TOT_ASSET	0.9953
C2	BS_TOT_ASSET	0.9935	BS_TOT_ASSET	0.9927	BS_TOT_ASSET	0.9943
C2	BS_TOT_LOAN	0.9854	NON_INT_EXP	0.9886	NON_INT_EXP	0.9877
C2	INTEREST_INCOME	0.9770	IS_OPERATING_EXPN	0.9883	INTEREST_INCOME	0.9876
C3	BS_TOT_ASSET	0.9817	INTEREST_INCOME	0.9678	NON_INT_EXP	0.9671
C3	IS_COMM_AND_FEE_EARN_INC_REO	0.9803	NON_INT_EXP	0.9670	NET_INT_INC	0.9621
C3	NON_INT_EXP	0.9800	IS_OPERATING_EXPN	0.9624	IS_OPERATING_EXPN	0.9564
Community	Bottom Measures 2001-06	Values 2001-06	Bottom Measures 2007-09	Values 2007-09	Bottom Measures 2010-13	Values 2010-13
C0	BS_LT_BORROW	0.7196	BS_LT_BORROW	0.6723	BS_LT_BORROW	0.7185
C0	BS_SH_CAP_AND_APIC	0.7131	BS_ST_BORROW	0.6511	TOT_DEBT_TO_TOT_ASSET	0.6659
C0	TOT_DEBT_TO_TOT_ASSET	0.4386	TOT_DEBT_TO_TOT_ASSET	0.5360	BS_ST_BORROW	0.6537
C1	RETURN_ON_ASSET	0.7006	BS_LT_INVEST	0.5799	BS_SH_CAP_AND_APIC	0.8151
C1	INTERBANKING_ASSETS	0.4987	INTERBANKING_ASSETS	0.5724	BS_LT_INVEST	0.7075
C1	TOT_DEBT_TO_TOT_ASSET	0.4941	TOT_DEBT_TO_TOT_ASSET	0.1366	TOT_DEBT_TO_TOT_ASSET	0.1762
C2	INTERBANKING_ASSETS	0.8878	RETURN_ON_CAP	0.7911	BS_ST_BORROW	0.8892
C2	TOT_DEBT_TO_TOT_ASSET	0.6980	BS_SH_CAP_AND_APIC	0.7245	TOT_DEBT_TO_TOT_ASSET	0.7574
C2	BS_SH_CAP_AND_APIC	0.5491	TOT_DEBT_TO_TOT_ASSET	0.5702	RETURN_ON_CAP	0.7498
C3	BS_SH_CAP_AND_APIC	0.8124	BS_CASH_NEAR_CASH_ITEM	0.8004	BS_CASH_NEAR_CASH_ITEM	0.7045
C3	RETURN_ON_ASSET	0.8007	BS_NON_PERFORM_ASSET	0.7707	IS_INT_EXPENSES	0.6626
C3	TOT_DEBT_TO_TOT_ASSET	0.6345	TOT_DEBT_TO_TOT_ASSET	0.5311	TOT_DEBT_TO_TOT_ASSET	0.6519

Table 7: First table shows the sets of top three contributors for each community, while the second table shows the bottom three contributors. Values represent the contributions of original measures to the explained variances. Rankings refer to averaged values along each sub-period: 2001-06, 2007-09 and 2010-13.

3.3.3 PCA results

Community detection shows the presence of three large clusters (Mixed, US, and JP) and an additional quite stable and persistent but smaller community (mostly US+EU banks). In this Section is provided a way to describe how these communities can be represented in terms of economic features (see Figure 13). Given the multi-dimensionality of the set of measures utilised to build the networks, a Principal Component Analysis approach is adopted to identify those measures which contribute more (less) to the explained variance within each community. For the sake of simplicity, the ranking of the top (bottom) three measures is proposed for each community during the following intervals: pre-crisis ($2001 - 2006$), crisis ($2007 - 2009$), and post-crisis ($2010 - 2013$). In particular, for each year the contribution of the original measures is computed to explained variance; then, it is considered the average for each sub-period and determined the rankings based on the mean period values. Below, the community with a mixed geographical composition is named as *C0*, while referring to the communities with a prevalence of US, JP and European plus US banks as *C1*, *C2* and *C3*, respectively.
This representation allows to compare communities' features over time and across different groups. For instance, it is observed that *Total Assets* and *Interest Income* are quite frequent among top measures contributors, while *Total Debts to Total Assets* is recurrent among measures in the bottom rankings. This is not surprising given banks heterogeneity in terms of the size (*Total Assets*) and the economic results (*Interest Income*) distributions, in contrast with the tight constraints on leverage (*Total Debts to Total Assets*) due to regulatory requirements. By focusing on the top rankings *C0* and *C1* have fairly stable top contributors, while communities *C2* and *C3* are more affected by the wave of financial turmoil. Furthermore, bottom rankings seem to be on average only slightly influenced by the choice of different sub-periods. In addition, differences between mean values among the set of top three and the set of bottom three contributors are quite stable over time with only few exceptions, while the middle part of the distribution of measures' contributions is in general quite

sparse. One might be interested in how the outbreak of financial crisis have influenced these rankings. Top composition of *C1* is unaffected by the 2007-08 financial breakdown, while *C0* is only partially modified by the onset of the crisis (*Interest Income* is replaced by *Net Interest Income*). Conversely, *C2* presents a quite different configuration during the crisis sub-period when it exhibits a relevant role for expenses measures (i.e. Non Interest Expenses and Operating Expenses). Similarly, income statement measures become more relevant among top contributors also within the *C3* community. Interestingly, community *C0*, which is characterised by a mixed geographical composition, and the US community (*C1*) reach identical top contributors after the outbreak of 2007-08, while the JP community (*C2*), which shows the same top contributors as community *C0* in the first sub-period, seems to react differently during the crisis, although in the third sub-period it shows again top contributors similar to *C0* (and to *C1*). By contrast, community *C3* seems to present a peculiar pattern over time.

Therefore, the crisis sub-period coincides with remarkable differences in the top contributors, while the recent sub-period points to a renewed tendency to get similar contributors for a wider set of banks (*C0* and *C1*, and partially *C2*). This seems to be in line with the above discussion on community detection results, where there is highlighted a gradual proximity between clusters over time. Hence, these results suggest that heterogeneity within clusters is driven by similar economic measures after the crisis, although specific differences persist. This is the case for instance of loans, which are not present among top contributors in the US community while they are in the top ranking of both the Mixed and the JP community (as expected according to the above discussion). Also , the crisis seems to suggest an increasing importance of income statement measures in terms of contribution to the explained variance within communities. The breakdown of financial markets affected banks' results and this justifies the high level of heterogeneity expressed by income statements indicators. This can also be related to the impact of the crisis on financial statement measures and on the different ways banks update their balance sheet structures compared to the recognition of economic

results as reported in the income statements items. Similar comparisons can involve also the bottom three measures, but for conciseness we omit this part.

3.4 Discussion

In this project, banks' financial statements are used to depict the banking system. The main contribution is represented by the introduction of a methodology to exploit balance sheets and income statements data to construct Accounting Networks. Some relationships between economic indicators and network properties are shown, which might provide some new useful insights for banking classification practices. Having depicted some effects of the recent financial crisis by using a simple framework is an encouraging sign for further extensions. We rely on "neutral" and "naive" techniques to build the Accounting Networks. In particular, among common approaches usually applied to describe similarities concepts, it is adopted one of the basic method, i.e. the cosine similarity. Future works can exploit more advanced methodologies. Moreover, our selection of variables utilised to compute cosine similarities assumes that each component has the same importance. This is quite a naive hypothesis, which could be enriched by measures discrimination based on economic literature and/or practitioners practices. Finally, for accounting reasons our study is limited on annual financial statements, while a more detailed description of the system might easily involve the use of quarterly data. Despite these simplifying assumptions, the approach has the merit of introducing a novelty in the debate on banking networks, and future improvements in the directions outlined above will enforce Accounting Networks' ability to describe the evolution of banking systems.

Chapter 4

The Bitcoin Peers Network

4.1 Introduction

Behind Bitcoin [55], the most popular cryptographic currency, there are users distributed all over the world who, in a voluntary way or for profit, participate to a network where transactions are announced, verified and eventually inserted into blocks of a distributed ledger known as Blockchain. When a transaction is executed it is announced by broadcasting it to the network where peers contribute to spread the transactions received from other peers sending them to their own contacts. Peers also validate transactions which are gathered into blocks which, approximately every 10min are sealed cryptographically through the so-called proof-of-work (consisting of finding at random a valid hash associated with the block) and then get broadcasted to the network to be approved and inserted into the Blockchain in chronological order. Blocks are validated with a majority voting following the original Nakamoto's principle "one CPU, one vote" [55]. Currently[1] there are around 6000 peers that participate to this process and in each block are included in average between 1 and 1.7 thousand transactions, counting for about 100-170 transactions per minute which move a capital of 152 Bitcoins per minute equivalent to 91,787 USD. The mechanism of peer validation of blocks by majority is

[1] September, 2016

considered the most interesting innovation introduced by Bitcoin and it solves several issues related to trust and machine synchronisation that are otherwise hard to manage in a distributed system operating between untrustful peers. However, this peer-to-peer distributed design makes the system inefficient and hard to scale. It is therefore very important to identify the current inefficiencies in the Bitcoin network and understand if they are intrinsic consequences of the system design or instead if they are the result of the present setting of the Bitcoin network and can therefore be improved either within Bitcoin itself or in other future distributed systems. In this work we measure how the Bitcoin network works by monitoring transactions and blocks exchanged among peers during a period of 7 days from 04/05/2016 and 11/05/2016 and then by recording the time when transactions observed during that period have been inserted in the Blockchain during a following period of a few months.

4.1.1 Blockchain

The Blockchain is a distributed database which keeps track of all payments made using the Bitcoin currency. A payment is called "transaction" and involves one or more input Bitcoin addresses who are sending some funds to one or more "output" addresses. Transactions are included inside blocks by special peers called "Miners" which participate to the solution, by brute force, of a cryptographic puzzle, the proof-of-work. After a miner creates a block, he will try to seal it cryptographically with a hash produced from the block and a random part. If the number is by chances smaller than a threshold imposed by the proof-of-work then it is considered "valid" and it can start to be spread among the network. When a peer receives a new block, it should verify if the block is valid. In order to do that, it has to verify whether the hash of the block fulfills the proof-of-work requirements. After that, the peer has to verify also each transaction included inside the block. If the whole block and all the transactions are verified, it accepts the new block as valid and starts propagating it through the network (and if the peer is a miner, also

it will start to discover the next block on top of it). If the block is not valid, or at least one transaction inside the block is invalid, the block will be discarded.

4.1.2 Communication protocol

All Bitcoin clients are connected to each other in a peer to peer network. Consequently, there are no central servers or authorities. Each node individually decides how to contribute to the network by choosing which services to provide, such as relaying transactions, storing a copy of the Blockchain or using their own computational power for mining. A node who wants to join the network for the first time needs to connect to some well known peers called "seeds". Seed nodes provide a partial list of nodes joined to the network. This list does not depend on the geographic location of clients, and all the clients included are chosen randomly and can contain up to one thousand nodes. After retrieving the peers list, a node starts to choose peers until it reaches its default maximum number of connection (usually from 8 to 126 established connections, the number of connections may vary according to the configuration of the Bitcoin client used and depending on the network settings of the client itself). Once connected to the network, the node is able to send and receive messages from all the others connected nodes, such as Blocks, transactions and new peers joined to the network. All these messages have to respect the rules set up by the Bitcoin Protocol [1], which consists of a set of messages used by clients to enable communication among peers. There are several customizations of the Bitcoin client but all of them have to respect the rules provided by the protocol. In order to monitor the Bitcoin network I wrote a customized client able to recursively establish a connection with each reachable node, requesting its known peers list and again trying to connect to them and retriev their list. To accomplish this goal, the client does not need to implement the whole protocol, but only a reduced set of messages:

- getaddr, addr - The "getaddr" message is used to request a list of known peers from a node. The node will issue an "addr" message

as response, containing up to one thousand known nodes. "addr" messages are also sent automatically to each connected nodes when the client establishes a connection with a new node.

- inv - The "Inventory" message is sent by a client when it discovers new blocks or transactions in order to spread them to the network. In the same "inv" message it is possible to have blocks and transactions together.

The addr messages are required in order to connect to all reachable peers and to new discovered peers once they join the network. Once connected, the client stores all the inventory messages received in the form:

timestamp address hashcode

where:

- timestamp is a 64bit integer representing the time and date when the inv message is received.
- address is the ip address of the nodes (which can belong to ipv4, ipv6 or tor networks).
- hashcode is the hashing string corresponding to a block or to a transaction.

The client establishes only one connection to each peer and it does not do any getdata request in order not to add load to the network. This approach has the drawback that each peer can close the connection every time without sending any alert. When it happens, the client suddenly tries to establish a new connection but it will miss the information shared with other peers during the time the connection was down.

4.2 Related Work

In the last few years there has been some interest in the study of the Bitcoin network with notable contributions from Decker [33] and Coinscope

[10]. Also, Bitnodes [2] is an online service which provides a snapshot of all reachable peers on the networks and some statistics related to the type of the client (i.e. protocol version used, last block stored and ip-geolocalization). Since all the data are provided as a list of online clients it is not possible to understand how the peers are connected to each other or how data are propagated among them. The approach used to discover peers on the Bitcoin network is to send recursively "getaddr" message to each reachable node in order to get back part of their known nodes list. Coinscope uses the same approach in order to discover clients, also introduced an algorithm, named "AddressProbe" which was able to track how peers were connected. Discovering connections was possible because each client keeps the timestamp of a peer updated in the mempool after each data exchange so, until Bitcoin Core 0.10.1 [5], every time a client replied back to a peers list it was also sending their updated timestamps. At the beginning, the mechanism for updating the timestamp was the following: if a node exchanges some messages with a peer, it keeps its own timestamp on the database updated. If, instead, a node discovers some new nodes through another peer, it applies a 2 hours penalty on the timestamp before storing the address into its own peer database. Based on this mechanism it was possible to guess the connections [23] of a peer just by retrieving several times the known peers list and sorting all the records in chronological order. This kind of network topology inference makes use of behaviour specific to Bitcoin Core prior to version 0.10.1. Biryukov [23] [5], and showed that reconstructing the peers network could be used to make an attack on Bitcoin Core clients. In order to avoid the possibility of such attack, the software was modified and now each client does not update the timestamp of a connected client every time they send or receive data.

After the last update on the client we noticed that, for an active connection, the timestamp is updated only when the connection drops or each 24 hours (in case the connection is still alive). Other cases are still the same as described on [10]. Decker [33] studied the data propagation rate. His idea was to establish a connection with each node and measure the time at which each block or transaction was received. In this way, with-

out knowing how the peers are connected, he was able to measure how long a block or a transactions takes to propagate on the network. Here, following Decker's methodology, the aim is to identify the appearance of block and transactions in the network and to measure the propagation dynamics in the network and time they take to be included within the Blockchain.

4.3 Methods

The Bitcoin network groups on average six thousands heterogeneous reachable peers with different computational capabilities and distributed around the world, connected by "random" links. In this work I am using data propagated through the peers in order to reconstruct all the information related to the network. This is done to achieve a better understanding on peers, focusing on:

- How they are characterized
- How they interact with each other

Data exchanged on the network consist of coordinating signals (i.e. announcing new blocks or transactions) and data messages (blocks, addresses and transactions). The data were collected joining on the network as a normal node and trying to establish a connection with each peer address discovered and waiting for "inv" messages for both blocks and transactions. During the listening period of 10 days, I found more than 12 thousands unique peers, 8969 belonging to ipv4 network, 3332 belonging to ipv6 network and 124 belonging to Tor network, with an average of 5-7 thousands client connected at the same time. This amount of peers is consistent with the amount reported by Bitnodes [2]. Surprisingly more than 126 thousands different blocks were received (instead of about 1200), some of them valid but "old", where the oldest of them was included into the Blockchain more than 3 years ago. Only one connection was established with each reachable node in order not to interfere with the network load.

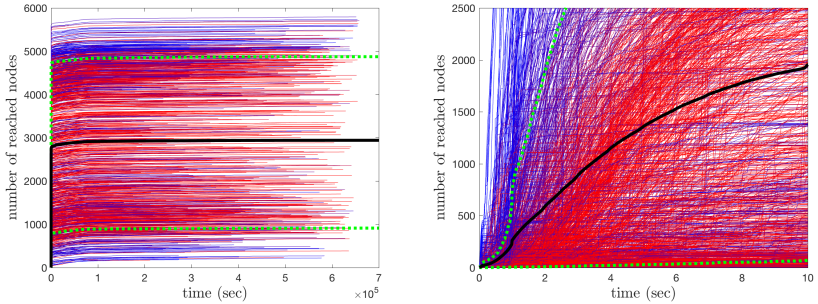

Figure 15: Number of nodes reached by a new valid block before a following block is discovered (left). The color is associated with the size of the block with blue being smaller and red larger. The black line is the average of all the observations and the two dashed green lines are respectively the 10% (lower) and 90% (upper) percentiles. The right plot is a detail of the initial propagation within the first 10 seconds.

4.4 Results and Discussion

Here I only report results about the MDLB set for blocks (mined during the listening time) and BT set for transactions. Our client established a connection with each reachable peer into the the network and waited for "inv" messages sent by them. The client to collect the data was written in Go programming languages[3], in order to exploit its multi threading native management. We established only one connection with each reachable node in order not to interfere with the network's behaviour. This is because most clients accepts only 8 connections from peers and it is not possible to measure or estimate the number of active connections held by each client. Each client has the possibility to drop the connection at any time without warning the peer. This means that if the connection is lost before the peer is propagating a new block (on average every 10 minutes), the node will not send anymore the block after that the connection is recovered. We collected 592GB of data in a period of 1208 valid blocks (from block height 410119 to 411327) mined during the listening time window. The most part of the data regard transactions of "inv" messages (589 GB), while the remaining is related to blocks of

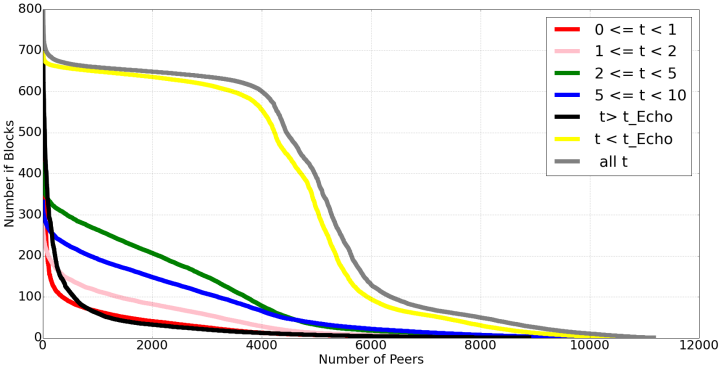

Figure 16: This figure shows the number of Blocks received from Peers in a defined time window. The red line groups how many Blocks (y axis) are received by nodes (x axis) in 1 second after the first propagation ($t < 1$ second).

"inv" messages. During the investigation we received a large amount of blocks and transactions, and we decided to classify them, as described in Chapter 1.

4.4.1 Blocks

In figure 15 a cumulative block propagation plot is shown. The curves have different length due to the fact that each block is discovered after a different time. The number of nodes depends on the number of connections established during the block propagation time. Table 4 shows some time statistics related to the time required by miners in order to discover a new block. Even though the data collected for each block are propagated by a different number of peers, preliminary results show that the propagation time seems to be quite stable on the network as showed in figure 16.

4.4.2 Transactions

All transactions id received from clients are recorded together with the client address and the receiving time. Figure 17 shows the received transactions rate per hour for IT set (in red) and BT plus ET set (in blue). During the listening time, a total of 1820212 Transactions were received. Among them, 1722696 Transactions have been included in the Blockchain during the period until Sat, 09 Jul 2016 10:52:38 GMT. The total number of Transactions included in the Blockchain during the monitoring time is 1723962. So, the client do not received 1266 Transactions which were included in the Blockchain, 1208 of them correspond to the transaction zero of each block (that do not need to be sent over to the network by default). In figure 18 it is possible to see, for each block, how many transaction were received by our client (in blue) and how many of them were included in the Blockchain (in red) during the corresponding block time. A peak of about 19674 transaction were received during the first listening block time, followed by two other peaks, one of 13334 transactions near block height 411200 and the last one of 10775 transactions near block 410600.

As it is possible to see in figure 18, the number of transactions included in each block of the Blockchain is very low compared to the number of transactions propagated by the network. In this context, the time and the number of blocks required by a transaction to be included in the Blockchain were computed. As it is possible to see in figure 19 and 20, transactions are included in the Blockchain with a clear delay, as a transaction may wait even 10 Blocks before becoming part of the Blockchain, corresponding to almost 2 hours on average. Furthermore, after a TX is included in the Blockchain it is good practice to wait for 6 confirmations before considering the amount received spendable for another payment. According to table 4, during the observation time we observed that the median time of issue of a new block is about 6 minutes, the mean time is about 10 minutes and the worst time is about 77 minutes.

Considering the mean time, this scenario leads to a total time of about 4 hours, starting from the time when the transactions is created until the

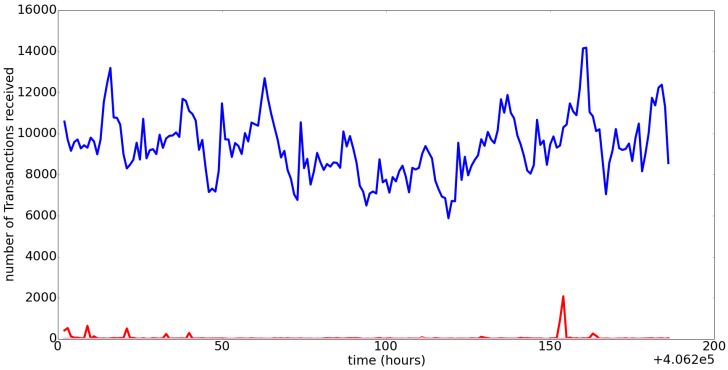

Figure 17: Number of Transactions per hour received during the listening time. The Blue line represent the transactions included in the Blockchain during or after the listening time (BT+ET). The red line represent the invalid transactions (IT).

time the new owner of the funds gains full possess of them. In figure 20 the time required (in terms of number of blocks) by a transaction to be included in the BC is compared to the fees paid by the transaction. As it can be seen, fees play a roles on the timing, but there is a sort of inefficiency since, as shown in figure20, there are some high fee transactions, included in the BC after 50 blocks from their first propagation on the network.

The time interval between the first time a transaction is observed in the network and the time when it is included in the Blockchain is shown in figure 19, where the distribution of such time intervals is measured in seconds and in number of blocks. It is possible to observe a decreasing behaviour,which is compared with the best-fitting exponential decay $n \sim exp(-t/\Delta)$, shown as a red line. The coefficient Δ is the characteristic time and it was measured to be equal, respectively, to 2758 seconds and 4.1 blocks. However, it is possible to observe from figure 19 that the exponential decay law is not well followed by the empirical time distribution, that tends to have a larger proportion of fast transaction and a

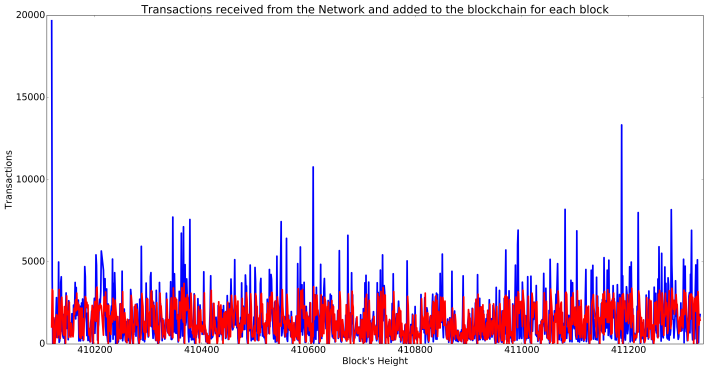

Figure 18: This figure compare the number of Transactions received by our client per block (in blue) and the number of transactions included in the Blockchain during the same block mining time (in red).

larger proportion of low transactions. Indeed, it results that 43% of the transactions are still not included in the Blockchain 1h after the first time they were observed and, remarkably, 20% of the transactions are still not included after 30 days, revealing therefore an unexpected inefficiency in the system. This statistics is reported in figure 22. If, instead, the fraction of transferred value that is included in the Blockchain after a given amount of time (figure 23) is chosen as measure, it is possible to note that the process is still rather slow but most of the value is included within 3h (93%) and after 30 days only 0.1% is still to be included. This is caused by the fact that the tail of long waiting transaction is mostly populated by transactions containing very small amounts as indicated by figure 24.

It is verified that fees (computed as the difference between total value inserted in the transaction minus the total value paid) play a minor role on the time a transaction takes to be included in the Blockchain. This is reported in figure 25, where it is possible to observe that some transactions associated with high fees have very long waiting times and, vice-versa, transactions with small fees are processed quite rapidly.

The major factor affecting waiting times appears instead to be the

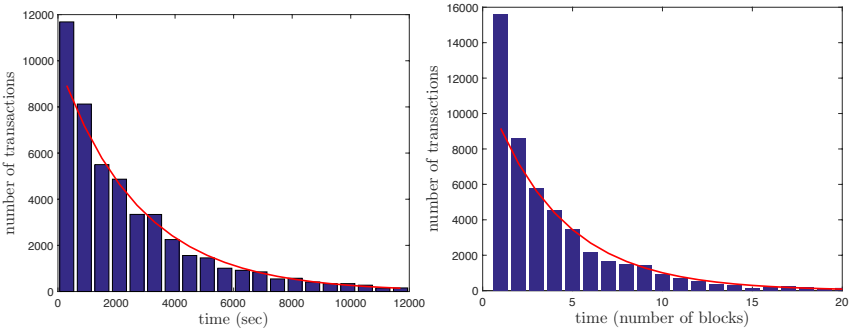

Figure 19: Distribution of time intervals between the first time a transaction is observed in the network and the time in which it is included into a valid block. The left plot reports time in seconds and the right plot reports time in number of blocks (approx 10min each). The red line are best fits with exponential decay law.

peers who first spread the transaction on the network.

4.5 Conclusion

In this work I showed some statistics related to the Bitcoin network. The propagation time required by Blocks seems quite stable and related to node features such us the Bitcoin client used. Also, the mechanism according to which transactions are included within blocks is pretty inefficient, compared with the actual transaction rate propagated on the network. This leads to a big number of transactions that are included in the BC with a delay of several hours up to some days after the transaction is issued, despite being already known by the network. Even if the actual block size is kept lower than 1MB, some peers include more transactions than allowed by the block capacity, guaranteeing a high transaction rate to Bitcoin users. Other miners instead create new blocks containing only few transactions, even if there is still space available inside the block, and other transactions are spread across the network, decreasing the transaction rate. There are also some cases in which miners leave the blocks empty, putting only the zero transaction inside a block, without con-

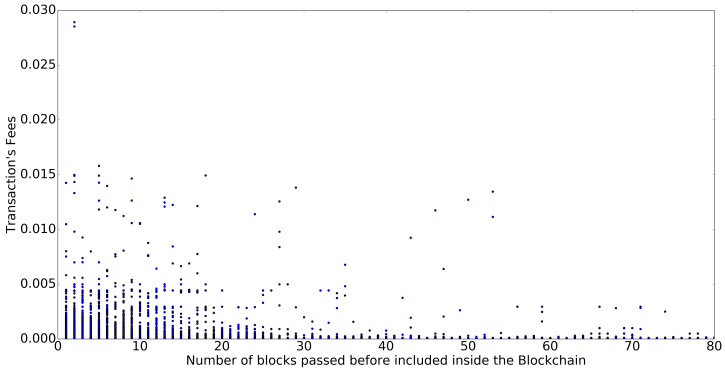

Figure 20: Number of blocks required by transactions to be included in the BC versus Fees earned by the miner.

tributing at all to the network, but yet obtaining the reward. This should suggest some improvements to the protocol, particularly on rules about how blocks are created. Indeed, if this kind of behaviour is adopted by the majority of miners it could lead to the freezing of the whole payment system.

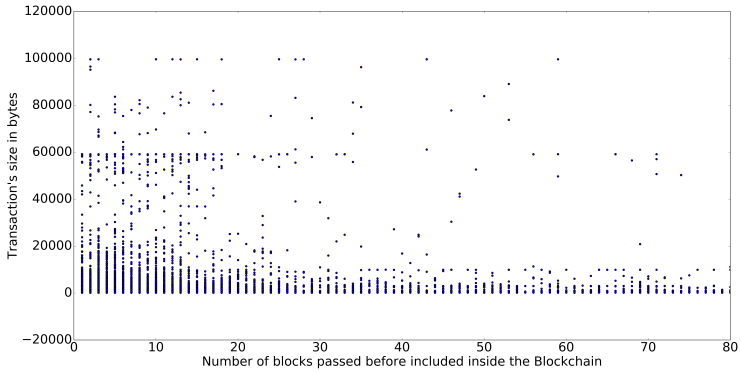

Figure 21: Number of blocks required by transactions to be included in the BC versus the size (in bytes) of the transaction.

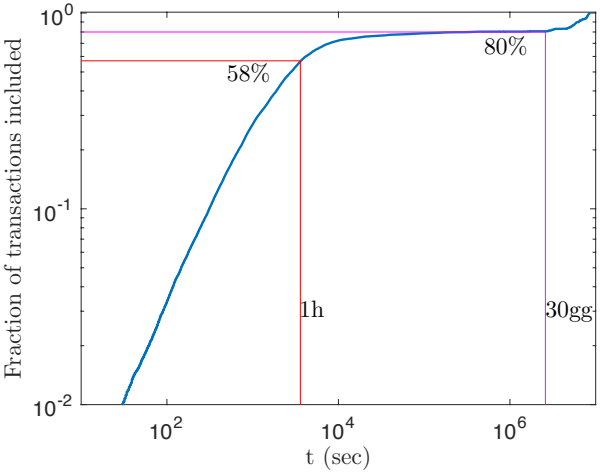

Figure 22: Fraction of Transactions included in the Blockchain after a given amount of time (seconds, x-axis) from first observation in the network. The two vertical lines mark 1h and 30days.

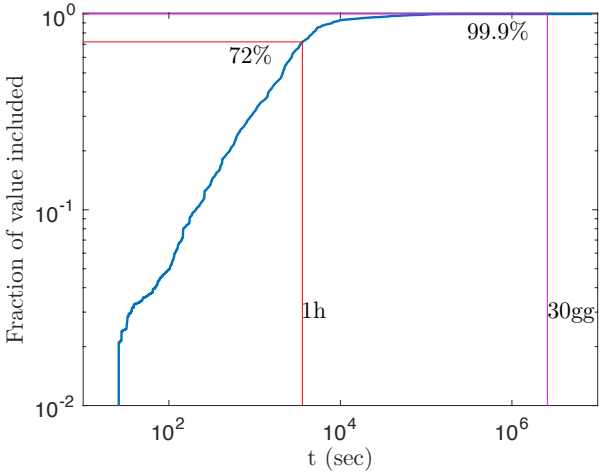

Figure 23: Fraction of transferred value included in the Blockchain after a given amount of time (seconds, x-axis) from first observation in the network.

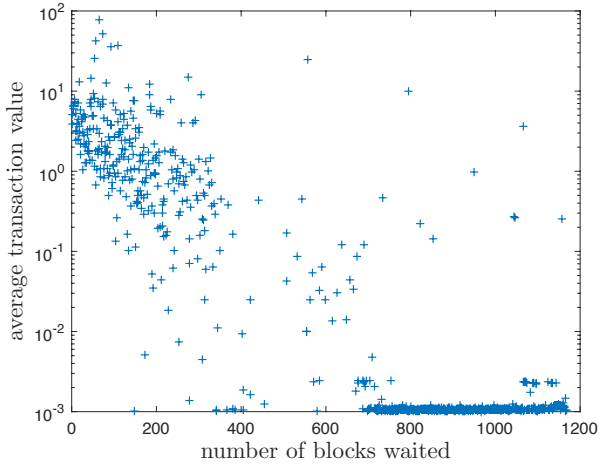

Figure 24: Average value of the transaction vs. waiting time in blocks numbers.

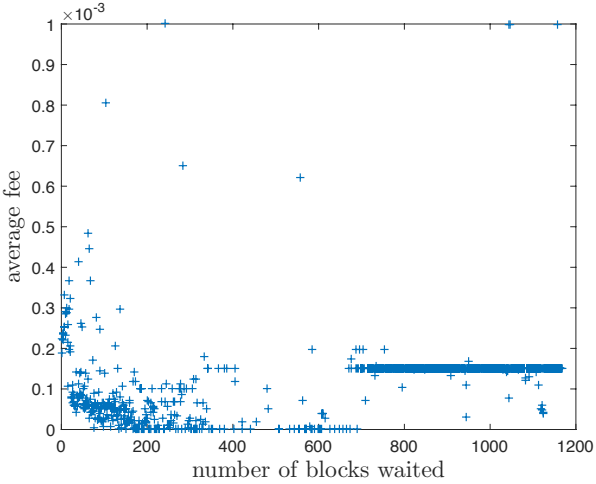

Figure 25: Average fee vs. waiting time in blocks numbers.

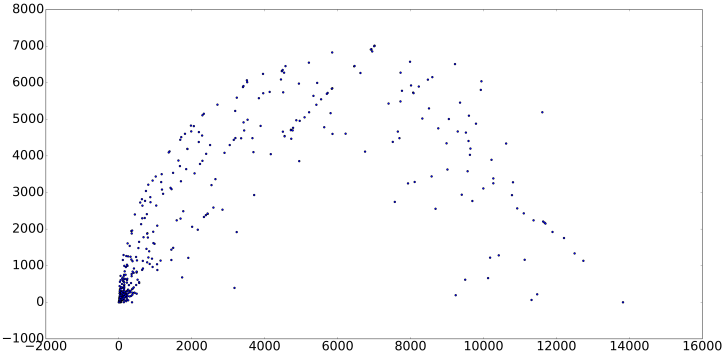

Figure 26: Mean and standard deviations of times required by a transactions to be included in the BC when issued by a certain peer.

Conclusion

For this dissertation, I developed a framework to retrieve, analyze and simulate data from several kinds of online sources by taking a network perspective. The aim of this work is to introduce a new data analysis workspace capable of using several tools on data coming from social media and financial sources in order to identify new patterns, correlations or causalities useful to data mining and forecasting applications.
In Chapter 1 I described the data I have been working with, and I showed some dataset examples created through the framework. All the studies proposed aim at finding a financial or inner value within the data. The explorative analysis on Chapter 2 shows how to use Facebook as Microeconomic data source. Wheareas it does not contain useful evidence for forecasting applications, it still represents a promising research project due to the data availability and cheap costs required. These indicators are very important in order to estimate economic trends and drive investments on specific sectors of a country/city. Application fields for future research range from nowcasting to mobility, ranging to measuring urban growth or developing models for shadow economy detection. As introduced in Chapter 3, within particular conditions data may be misinformative. In such cases introducing a network approach can shed light on hidden features of the global system, as it was done introducing the Accounting Network where networks features were compared with classical economic indicators from literature, considering as reference the period from 2001 to 2013. Here balance sheet data were used as a tool to build a network in order to study how banks acted during the recent 2007

crisis, finding some analogies between economic indicators and network features. Future works can take into account more advanced methodologies, which can allow also to differentiate each financial variable according to the case at hand. Finally, oi Chapter 4 I analysed the Bitcoin Peer network. Even though it is not known how peers are linked to each other, some important findings emerge from the data propagated by clients to the network. The rate of transactions included inside the Blockchain is kept very low by miners, even considering the higher transaction rate spread by peers. This leads to a global inefficiency of the system, due to policies adopted by several miners. Future works can suggest to introduce a permanent monitoring of the status of the network and to make some changes on the rules adopted to create new blocks, in order to discourage bad practices by miners which can freeze the global Bitcoin payment system.

References

[1] Bitcoin protocol documentation. https://en.bitcoin.it/wiki/Protocol_documentation. 59

[2] Bitnodes is currently being developed to estimate the size of the bitcoin network by finding all the reachable nodes in the network. https://bitnodes.21.co/. 61, 62

[3] The go programming language. https://golang.org. 63

[4] Graph api reference, fields description. https://developers.facebook.com/docs/graph-api/reference/user. 16

[5] Guessing bitcoin's p2p connections. https://jonasnick.github.io/blog/2015/03/06/guessing-bitcoins-p2p-connections/. 5, 61

[6] Viral Acharya. A theory of systemic risk and design of prudential bank regulation. *Journal of Financial Stability*, 5(3):224–255, 2009. 2

[7] T. Adrian and H.S. Shin. Liquidity and leverage. *Journal of Financial Intermediation*, 19:418–437, 2008. 2, 38

[8] Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Scientific Reports*, Volume 3:Article number 3578, December 2013. 2

[9] F. Allen, X. Gu, and O. Kowalewski. Financial crisis, structure and reform. *Journal of Banking & Finance*, 36:2960–2973, 2012. 38

[10] Miller Andrew, Litton James, Pachulski Andrew, Gupta Neal, Levin Dave, Spring Neil, and Bhattacharjee Bobby. Discovering bitcoins public topology and influential nodes. 5, 61

[11] Haldane Andrew G. Speech by mr andrew g haldane, executive direct or, financial stability, bank of england, at the financial student association, amsterdam, 28 april 2009. http://www.bis.org/review/r090505e.pdf. 2

[12] B. E. Aronson. A reassessment of japan's big bang financial regulatory reform. *IMES Discussion Paper Series*, (Discussion Paper No. 2011-E-19), 2011. 48

[13] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. *CoRR*, 2010. 1

[14] Per Bak and Kim Sneppen. Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Lett.*, 71:4083–4086, Dec 1993. 2

[15] Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality. *Phys. Rev. A*, 38:364–374, Jul 1988. 2

[16] Nuno Barreira, Pedro Godinho, and Paulo Melo. Nowcasting unemployment rate and new car sales in south-western europe with google trends. *NETNOMICS: Economic Research and Electronic Networking*, 14(3):129–165, 2013. 1

[17] Nuno Barreira, Pedro Godinho, and Paulo Melo. Nowcasting unemployment rate and new car sales in south-western europe with google trends. 14(3):129–165, 2013. 28

[18] J.A. Batten and P.G. Szilagyi. Why japan needs to develop its corporate bond market. *International Journal of the Economics of Business*, 10(1):pp. 83–108, 2003. 48

[19] T. Beck. The econometrics of finance and growth. *In Palgrave Handbook of Econometrics*, 2:1180–1211, 2009. 37

[20] T. Beck. The role of finance in economic development: Benefits, risks, and politics. *European Banking Center Discussion Paper*, (2011-038), 2011. 37

[21] A. Beltratti and R.M. Stultz. The credit crisis around the globe: Why did some banks perform better? *Journal of Financial Economics*, 105(Issue 1):1–17, 2012. 39

[22] A.N. Berger and C.H.S. Bouwman. How does capital affect bank performance during financial crises? *Journal of Financial Economics*, 109:146–176, 2013. 2, 3, 38

[23] Alex Biryukov, Dmitry Khovratovich, and Ivan Pustogarov. Deanonymisation of clients in bitcoin P2P network. *CoRR*, abs/1405.7418, 2014. 61

[24] BIS. *Bank of International Settlements 84th Annual Report (June 2014)*. 37

[25] BIS. Basel 3: A global regulatory framework for more resilient banks and banking systems. 2011. 7, 47

[26] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008+, July 2008. 43

[27] A. Blundell-Wignall and C. Roulet. Business models of banks, leverage and the distance-to-default. *OECD Journal: Financial Market Trends*, Vol. 2012/2, 2013. 39

[28] Ilaria Bordino, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen, and Ingmar Weber. Web Search Queries Can Predict Stock Market Volumes. *PLoS ONE*, 7(7):e40014–, July 2012. 2

[29] Michael Boss, Helmut Elsinger, Martin Summer, and Stefan Thurner. An Empirical Analysis of the Network Structure of the Austrian Interbank Market. *Financial Stability Report*, (7):77–87, 2004. 2

[30] M.K. Brunnermeier. Deciphering the liquidity and credit crunch 2007-2008. *Journal of Economic Perspectives*, 23:77–100, 2009. 2, 38

[31] C. W. Calomiris and D. Nissim. Crisis-related shifts in the market valuation of banking activities. *Journal of Financial Intermediation*, 23:400–435, 2014. 7

[32] HYUNYOUNG CHOI and HAL VARIAN. Predicting the present with google trends. *Economic Record*, 88:2–9, 2012. 1

[33] Christian Decker and Roger Wattenhofer. Information Propagation in the Bitcoin Network. In *13th IEEE International Conference on Peer-to-Peer Computing (P2P), Trento, Italy*, September 2013. xiv, 5, 60, 61

[34] Rajan R. Dell'Ariccia G., Detragiache E. The real effect of banking crises. *Journal of Financial Intermediation*, 17:89–112, 2008. 37

[35] A. Demirgüc-Kunt and Huizinga H.P. Bank activity and funding strategies: The impact on risk and return. *Journal of Financial Economics*, 98(Issue 3):626–650, 2010. 39

[36] D.W. Diamond and R.G. Rajan. The credit crisis: conjectures about causes and remedies. *American Economic Review*, 99:606–610, 2009. 38

[37] Remco M. Dijkman, Panagiotis G. Ipeirotis, Freek Aertsen, and Roy van Helden. Using twitter to predict sales: A case study. *CoRR*, abs/1503.04599, 2015. 2

[38] ECB. *Banking Structures Report (November 2013)*. 37, 48

[39] Facebook for developers. `http://developers.facebook.com`. 9

[40] FASB. Comparability in international accounting standards: A brief history. *Financial Accounting Standards Board*. 47

[41] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010. 43

[42] Gadm database of global administrative areas. `http://www.gadm.org/`. 14

[43] Prasanna Gai and Sujit Kapadia. Contagion in financial networks. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 466(2120):2401–2423, 2010. 2

[44] Prasanna Gai and Sujit Kapadia. Contagion in financial networks. Bank of England working papers 383, Bank of England, 2010. 2

[45] Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, October 2010. 1

[46] G. Andrew Haldane. *Rethinking the financial network*, pages 243–278. Springer Fachmedien Wiesbaden, Wiesbaden, 2013. 2

[47] Martin Hentschel and Omar Alonso. Follow the money: A study of cashtags on twitter. *First Monday*, 19(8), 2014. 1

[48] T. Hoshi and A. Kashyap. The japanese banking crisis: Where did it come from and how will it end? *NBER Working Paper Series*, 1999. 47

[49] H. Huizinga and L. Laeven. Bank valuation and accounting discretion during a financial crisis. *Journal of Financial Economics*, 106:614–634, 2012. 39

[50] Thomas H. Noe Larry Eisenberg. Systemic risk in financial systems. *Management Science*, 47(2):236–249, 2001. 2

[51] R. Levine. Finance and growth: Theory and evidence. *In Handbook of Economic Growth*, pages 865–934, 2005. 37

[52] Alejandro Llorente, Manuel Garc'ia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment, November 2014. 28

[53] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PLoS ONE*, 10(5):1–13, 05 2015. 1

[54] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific Reports*, 3, May 2013. 1

[55] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2009. 4, 57

[56] M E J Newman and M Girvan. Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 69(2 Pt 2):16, 2004. 43

[57] Z. Pozsar, T. Adrian, A. Ashcraft, and H. Boesky. Shadow banking. *Federal Reserve Bank of New York Staff Reports*, 2012. 48

[58] Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, Miha Grar, and Igor Mozeti. The effects of twitter sentiment on stock price returns. *PloS one*, 9(10):e0138441–1–e0138441–21, 2015. 2

[59] C. M. Reinhart and K. S. Rogoff. This time is different: Eight centuries of financial folly. 2009. 37

[60] Tarik Roukny, Hugues Bersini, Hugues Pirotte, Guido Caldarelli, and Stefano Battiston. Default cascades in complex networks: Topology and systemic risk. *Scientific Reports*, 3, September 2013. 2, 3

[61] Torsten Schmidt and Simeon Vosen. A monthly consumption indicator for germany based on internet search query data. Ruhr Economic Papers 208, RWI - Leibniz-Institut fr Wirtschaftsforschung, Ruhr-University Bochum, TU Dortmund University, University of Duisburg-Essen, 2010. 1

[62] Torsten Schmidt and Simeon Vosen. A monthly consumption indicator for germany based on internet search query data. Technical Report 0208, Rheinisch-Westflisches Institut fr Wirtschaftsforschung - Ruhr-Universitt Bochum - Universitt Dortmund, Universitt Duisburg-Essen, 2010. 28

[63] Friedrich Schneider, Andreas Buehn, and Claudio E. Montenegro. Shadow economies all over the world : new estimates for 162 countries from 1999 to 2007. Policy Research Working Paper Series 5356, The World Bank, Jun 2010. 31

[64] Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert Syst. Appl.*, 30(2):243–254, February 2006. 1

[65] H. S. Shin. Global banking glut and loan risk premium. *IMF Economic Review*, 60(2):155–192, 2012. 48

[66] Joseph E. Stiglitz. Risk and global economic architecture: Why full financial integration may be undesirable. *American Economic Review*, 100(2):388–92, May 2010. 2

[67] V. A. Traag, G. Krings, and P. Van Dooren. Significant scales in community structure. *Scientific Reports*, 3:2930 EP –, Oct 2013. Article. 43

[68] United states census bureau. `www.census.gov`. Accessed: 2015-01-30. 29

[69] Stijn van Dongen and Anton J. Enright. Metric distances derived from cosine similarity and pearson and spearman correlations. *CoRR*, abs/1208.3145, 2012. 41

[70] Stefania Vitali, James B. Glattfelder, and Stefano Battiston. The network of global corporate control. *PLoS ONE*, 6(10):1–6, 10 2011. 2

[71] Altunbas Y., Manganelli S., and David Marques-Ibanez. Bank risk during the financial crisis. do business model matter? *European Central Bank*, 2011. 7

[72] Yahoo! developer console. `https://developer.yahoo.com/yql/console/`. 7

[73] Wenbin Zhang and Steven Skiena. Improving movie gross prediction through news analysis. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 301–304, Washington, DC, USA, 2009. IEEE Computer Society. 1