DELIVERABLE

BIGPROD PILOTS: Write-up of first analysis

Summary

This deliverable reports on the work in progress analysis of the BIGPROD project data. The analysis focuses on reporting the first write-up of the novel webscraped data and reporting the sample of companies used in the project.

Deliverable Information

Deliverable number and name:	**Work in progress scientific publication**
Due date:	31st January 2021
Deliverable:	D11
Work Package:	WP4
Lead Partner for the Deliverable:	STRATH
Author:	Arho Suominen, Scott Cunningham, Arash Hajikhani and Cees van Beers
Reviewers:	Hugo Hollanders
Approved by:	Arho Suominen
Dissemination level:	Public
Version	v. 1.0 27th January 2021 v. 2.0 18th January 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822

Disclaimer

This document contains a description of the **BIGPROD** project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium coordinator for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of BIGPROD consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu.int/)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822.

Introduction

The ambition of the BIGPROD project is to develop an econometric model integrating “Big data” measures, which would be by size and volume, significantly beyond the state-of-the-art. Our objective is to build a sample of 160,000 - 200,000 European companies for which we will create “Big data” productivity measures. In creating this sample, we have used NACE codes to focus the sample on high-tech companies. The companies selected for the sample are associated primarily to the following NACE codes:

1. 20 - Manufacture of chemicals and chemical products,
2. 21 - Manufacture of basic pharmaceutical products and pharmaceutical preparations,
3. 254 - Manufacture of weapons and ammunition,
4. 26 - Manufacture of computer, electronic and optical products,
5. 27 - Manufacture of electrical equipment,
6. 28 - Manufacture of machinery and equipment nec,
7. 29 - Manufacture of motor vehicles, trailers and semi-trailers,
8. 303 - Manufacture of air and spacecraft and related machinery, and
9. 325 - Manufacture of medical and dental instruments and supplies

In addition, we have used a geographical boundary focusing on European Union and United Kingdom companies. We also required that the companies in the sample are "active" during the time of the search, so while we use historical data, companies that have ended operations by the search date are not included to the sample. Overall, the BIGPROD sample contains 183,161 companies.

The process used in BIGPROD project follows the The CRoss Industry Standard Process for Data Mining (CRISP-DM) process. CRISP-DM is a six-phase process to describe the data science life cycle. The process starts from problem understanding, what is the actual need to solve the question at hand. In the second phase, the process focuses on data understanding – answering questions like what type of data do we need to have and how do we need to process the data. Third, the process moves to data preparation that focuses on how we can organize the data for modeling. The fourth step in the CRISP-DM process focuses on modeling the data and selection of techniques needed to accomplish modeling. Finally, the process focuses on evaluating a deployment of the data. All of these steps are reflected as a part of the BIGPROD project in its work package structure.

Our current focus is on phase two and three, data understanding and data preparation. Due to the computational time needed to crawl our full sample, BIGPROD project uses a convenience sample of data to look at what data does the crawling produce and does it fit the project objectives. Also, we will use the convenience sample to understand how we organize the data for modeling. For the first analysis of the data reported here, we used a convenience sample of 14,341 observations. This sample is created by the web crawling process that continuously adds new data.

Descriptive Statistics of High-Tech Sample

The 14,341 observations in the convenience sample represent companies from 28 different countries. However, the convenience sample is heavily skewed, where in total ten countries represent 80% of the firms in the sample. Shown in detail in Table 1 the companies in the convenience sample are based on the largest economies in the sample area, such as the United Kingdom, Germany and France. This said, there are significant differences when compared to the overall industrial structure of the whole sample. For example, the convenience sample contains Finland from the Nordic countries, but excludes Sweden, a larger economy. This already highlights that we should not consider the convenience sample as a representative subsample of the whole sample envisioned by BIGPROD nor a reflection of the European Union and United Kingdom economic activity in the selected NACE codes.

Table 1 Number of companies in the convenience sample of 14,341 in different countries.

Two-letter code	Percent of total	Number	Country name
UK	26.4	3,642	United Kingdom
DE	18.0	2,579	Germany
IT	9.8	1,412	Italy
FR	6.4	913	France
NL	5.0	714	Netherlands
ES	4.7	671	Spain
CZ	2.8	404	Czech Republic
DK	2.8	403	Denmark
FI	2.5	365	Finland
PL	2.5	362	Poland
-	20.1	2,878	All others

Looking at the industrial structure of the convenience sample, we also see a skewed distribution, as a few NACE codes are disproportionately represented in the sample. Electronic products represent nearly one-third of the total. In addition, pharmaceutical and chemistry firms are included in the sample with over ten percent share. This proportionality probably says more about the industrial organisation of high technology than the significance of the respective technologies involved in the overall economy.

T

NACE prefix	Percent of Total	N	Name of NACE code
26	31,7	4,552	Manufacture of computer, electronic and optical products
28	19,4	2,783	Manufacture of machinery and equipment n.e.c.
20	13,,1	1,875	Manufacture of chemicals and chemical products
27	12,8	1,842	Manufacture of electrical equipment
21	10,2	1,470	Manufacture of basic pharmaceutical products and pharmaceutical preparations
29	6,0	867	Manufacture of motor vehicles, trailers, and semi-trailers

32	5,0	718	Other manufacturing
30	1,5	212	Manufacture of other transport equipment
25	0,2	24	Machinery of fabricates metal products, except machinery and equipment

able 2 Number of companies in the convenience sample of 14,341 in different NACE codes.

To look at the financial status of the selected companies, our descriptive analysis showed, as expected, skewed distribution of companies. We used the kernel density estimate (KDE) plot to visualize the distribution of observations in selected financial indicators. The KDE plot is a method for visualizing the distribution of observations, like a histogram. KDE visualizes the data using a continuous probability density curve. We identified similar density curves across variables in the convenience sample.

Figure 1 shows the KDE plot for the number of employees in the convenience sample. The frequency distribution of number of employees in the sample focuses on smaller companies. This said, we should note that the mean value of employees was 1,356 (s= 11,715, N= 14,341), which can be seen as relatively high. Seen in the figure in blue is the same distribution for the whole sample of 183,161.

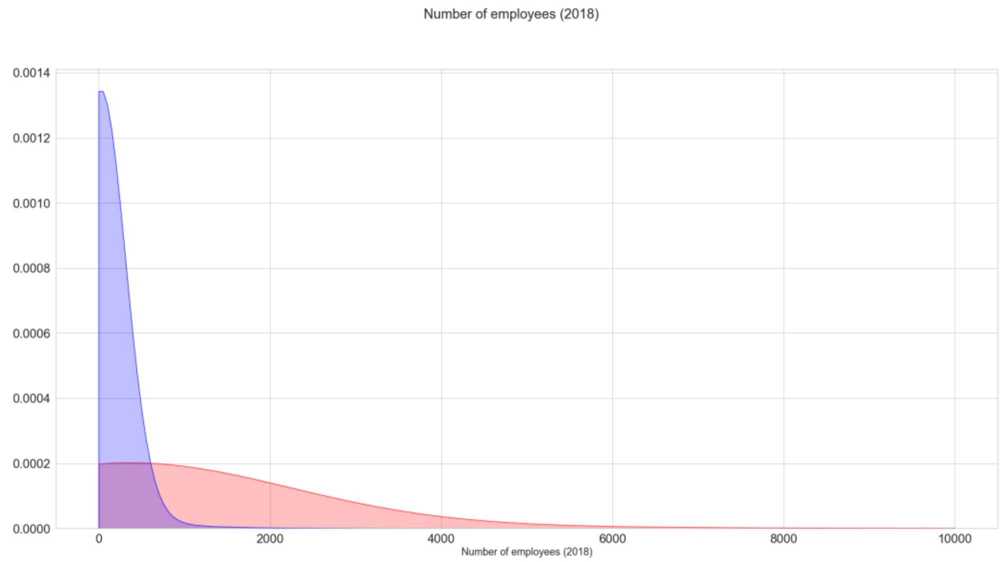


Figure 1 Density plot of number of employees in companies. Red distribution represents the convenience sample, blue the whole sample.

If we look at other financial indicators, we get a similar visualization. As seen in Figure 2, the operating revenue of the convenience sample has a larger portion of the companies in the lower operating revenue area. The mean value of the operating revenue of the convenience sample is 433,894 thousand euro (s=4,024 041, N=14,341). In estimating the revenues, we should note that the convenience sample contains the largest companies in the European Union and United Kingdom, such as Volkswagen AG and Daimler AG. Again, the figure contains the distribution for the whole sample in blue. This highlights the high operating revenue of the convenience sample.

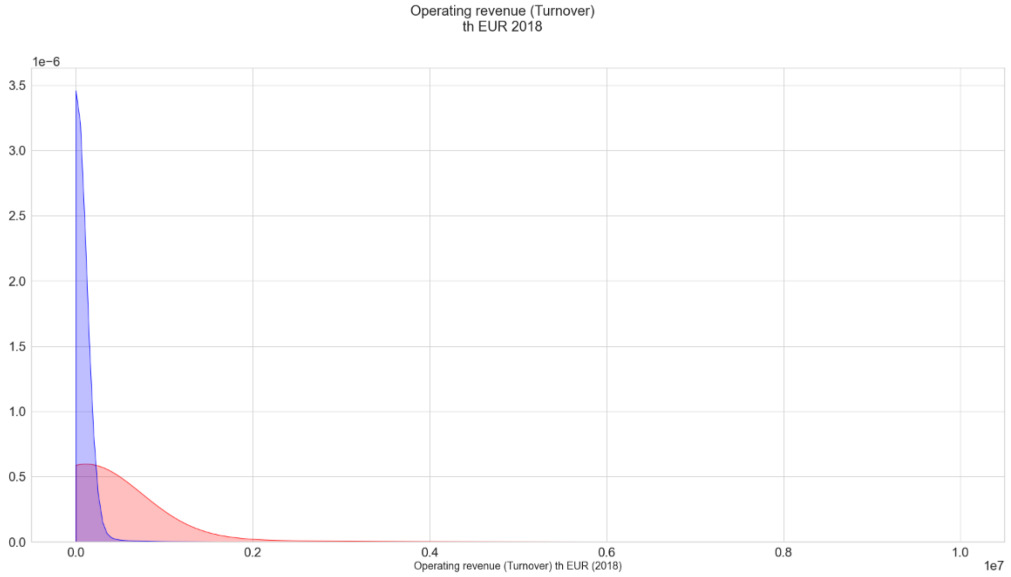


Figure 2 Density plot of Operating revenue in the sample companies. Red distribution represents the convenience sample, blue the whole sample.

To further reflect on the convenience sample, we looked at the total assets distribution and the return on equity (ROE). For assets, as seen in Figure 3 the companies mean value was 550,944 thousand euro (s=6,793,675, N=14,341). This aligns with the analysis made based on the operating revenues. If we focus on the ROE value, using profits before tax, the mean value is 17 (s=67, N 14,341). This is visualized in Figure 4. Similarly, in both Figure 3 and Figure 4 the blue area is for the whole sample.

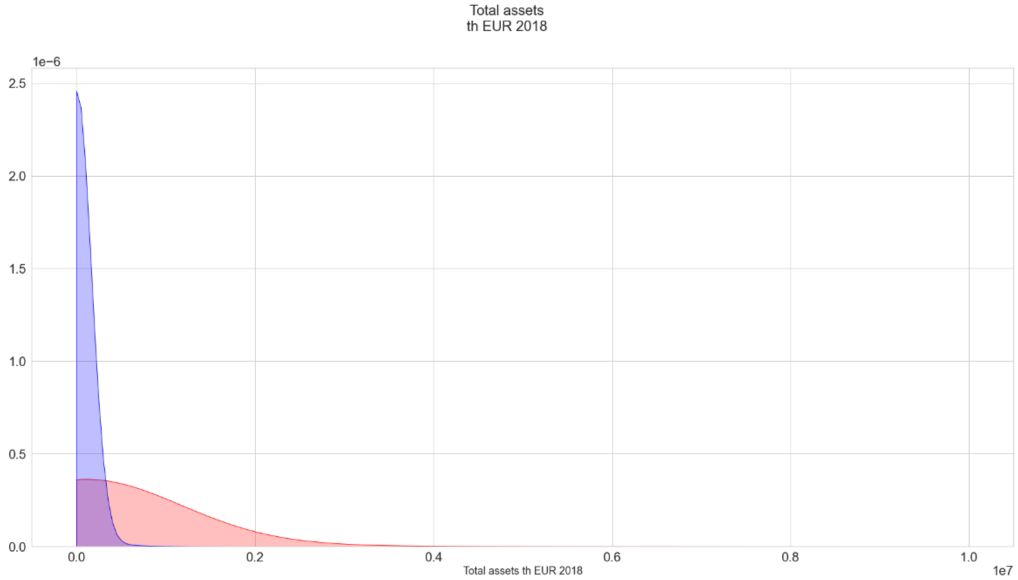


Figure 3 Density plot of the Total assets of the sample companies. Red distribution represents the convenience sample, blue the whole sample.

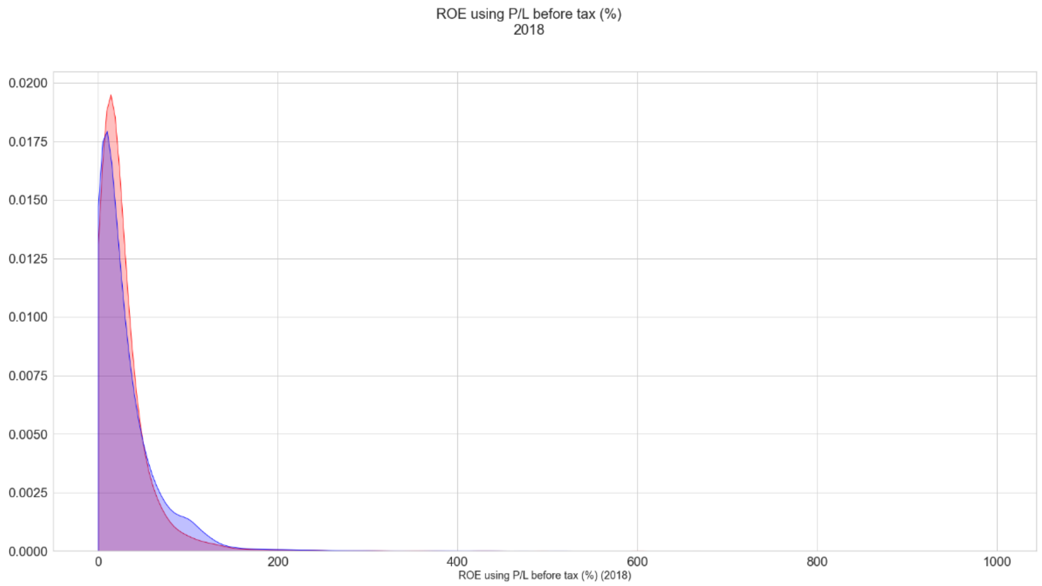
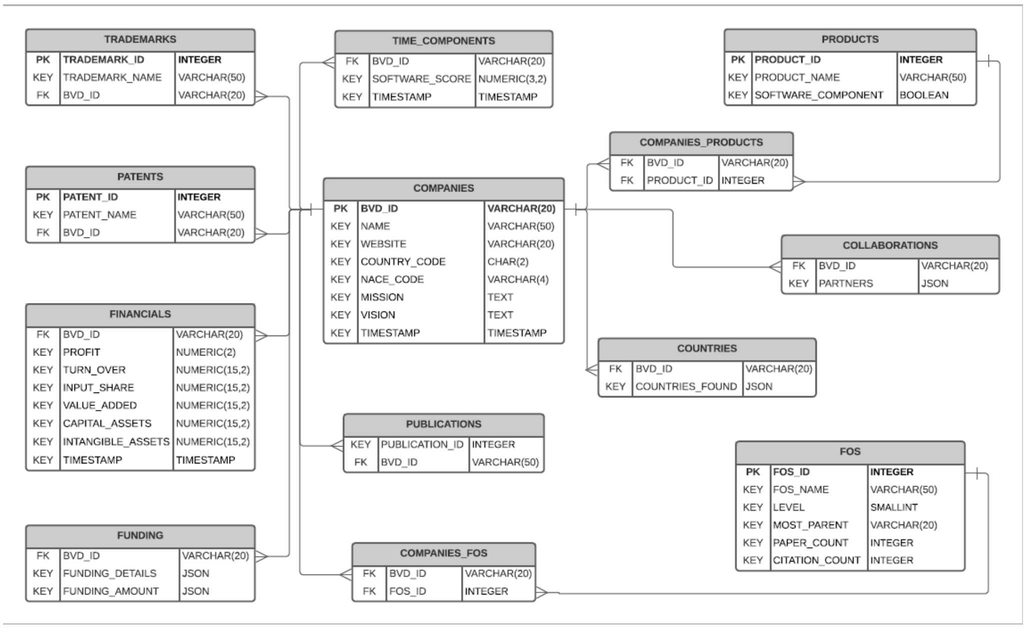


Figure 4 Density plot of the return on earning (ROE) of the sample companies. Red distribution represents the convenience sample, blue the whole sample.

Overall, it should be noted that the convenience sample is skewed towards large companies. This should be considered when moving forward in the analysis. It is clear that the larger companies can be expected to have a stronger online presence. Thus, the web scraping procedure for the convenience sample will probably yield a significant amount of information, in comparison to scraping smaller companies.



	TRADEMARKS	
PK	**TRADEMARK_ID**	**INTEGER**
KEY	TRADEMARK_NAME	VARCHAR(50)
FK	BVD_ID	VARCHAR(20)

	TIME_COMPONENTS	
FK	BVD_ID	VARCHAR(20)
KEY	SOFTWARE_SCORE	NUMERIC(3,2)
KEY	TIMESTAMP	TIMESTAMP

	PRODUCTS	
PK	**PRODUCT_ID**	**INTEGER**
KEY	PRODUCT_NAME	VARCHAR(50)
KEY	SOFTWARE_COMPONENT	BOOLEAN

	COMPANIES_PRODUCTS	
FK	BVD_ID	VARCHAR(20)
FK	PRODUCT_ID	INTEGER

	PATENTS	
PK	**PATENT_ID**	**INTEGER**
KEY	PATENT_NAME	VARCHAR(50)
FK	BVD_ID	VARCHAR(20)

	COMPANIES	
PK	**BVD_ID**	**VARCHAR(20)**
KEY	NAME	VARCHAR(50)
KEY	WEBSITE	VARCHAR(20)
KEY	COUNTRY_CODE	CHAR(2)
KEY	NACE_CODE	VARCHAR(4)
KEY	MISSION	TEXT
KEY	VISION	TEXT
KEY	TIMESTAMP	TIMESTAMP

	COLLABORATIONS	
FK	BVD_ID	VARCHAR(20)
KEY	PARTNERS	JSON

	FINANCIALS	
FK	BVD_ID	VARCHAR(20)
KEY	PROFIT	NUMERIC(2)
KEY	TURN_OVER	NUMERIC(15,2)
KEY	INPUT_SHARE	NUMERIC(15,2)
KEY	VALUE_ADDED	NUMERIC(15,2)
KEY	CAPITAL_ASSETS	NUMERIC(15,2)
KEY	INTANGIBLE_ASSETS	NUMERIC(15,2)
KEY	TIMESTAMP	TIMESTAMP

	COUNTRIES	
FK	BVD_ID	VARCHAR(20)
KEY	COUNTRIES_FOUND	JSON

	PUBLICATIONS	
KEY	PUBLICATION_ID	INTEGER
FK	BVD_ID	VARCHAR(50)

	FOS	
PK	**FOS_ID**	**INTEGER**
KEY	FOS_NAME	VARCHAR(50)
KEY	LEVEL	SMALLINT
KEY	MOST_PARENT	VARCHAR(20)
KEY	PAPER_COUNT	INTEGER
KEY	CITATION_COUNT	INTEGER

	FUNDING	
FK	BVD_ID	VARCHAR(20)
KEY	FUNDING_DETAILS	JSON
KEY	FUNDING_AMOUNT	JSON

	COMPANIES_FOS	
FK	BVD_ID	VARCHAR(20)
FK	FOS_ID	INTEGER

Figure 5 Database schema of the web scraped data used in BIGPROD project.

The scraping procedure follows the procedure described in the data platform report (Deliverable 8) from the BIGPROD project. The data platform is used to run the web scraping operation, but also to represent the scraped data in a convenient database format for further analysis. In the data platform, raw data is provided for analysis following the data structure in Figure 5.

In the convenience sample, we have used the data on the Field of Study (FOS), ISO and CE standard mentions, countries mentions and word index. FOS is based on Microsoft Research's open-source Microsoft Academic Graph[1] (MAG). We are using the MAG to assign a FOS'es which are the most similar to its vector of the website text using cosine distance. The ISO is a list of all mentioned ISO or CE standards on the website. Country mentions extracts all mentioned countries from the website.

To estimate the completeness of data we calculated the descriptive values for the variables. Seen in the below table, the data is almost complete for word index and FOS. Most of the sample also contains country mentions. For ISO codes, the data is only available for under 50 % of the sample.

Table 3 Descriptive statistics for the scraped variables.

	ISO & CE	Word index	Country mentions	FOS
count	6,236	14,332	12,067	1,4343
mean	2.9	13.4	11.4	72.3
std	3.9	2.2	18.3	30.4
min	1	1	1	0
25 %	1	11	2	51
50 %	2	15	5	85
75 %	3	15	12	98
max	98	15	188	100

Early analysis of the sample

Associated Web Pages

The following section provides descriptive details of the home page URLs associated with the high technology sample in our study. The data is associated with a moderate amount of duplication in the provided URLs. The question concerns the nature of this duplication and overlap, and the implications of this overlap for further web mining of the pages.

One step to explore this is to split the URLs first by "/," thereby identifying the root page and various subdirectories. The vast majority of URLs (98%) point to the root directory of the site. A second step is to split the root directory by "." indicating the various components of the web pages. When processed in this manner most URLs are in three parts -- for instance (www, panasonic, com). However, some URLs have up to five separate parts or components.

[1] Microsoft Academic: https://academic.microsoft.com/home

The most common variation is to drop off the "www" of the URL, which is not required by standard. Some URLs then proceed with the company name, but others prefer their own prefix. For instance, one company in the sample prefixes the URLs with the handle "new." A naive procedure for evaluating variation in the URLs is simply to calculate the information measure associated with each part of the directory structure. The information measure converts a discrete distribution of parts, and their frequency, to a single positive number which is limited by zero and asymptotically related to the total number of distinct firms in the sample. As a result, the measure converts greater variation into greater units of information (bits).

The results suggest that the largest source of variation is in the second part of the root directory -- the company name. This variation is even greater if an effort is made to sort out "www" leading prefixes from those which drop the prefix. Many companies provide a unified web presence for their various divisions, starting with a common prefix to the URL. Some of the most frequent exemplars of this are AkzoNobel (a Dutch painting and coatings company), IBM (the US headquartered information technology giant), and Faurecia (a French autoparts supplier). Roughly one in every four web pages indicated are subject to duplication in this part of the URL.

The fact that most of the variation remains here, and not deeper in the directory structure, is a positive indicator for the validity of the sample. We have attempted to create interesting graphics here, involving trace diagrams of increasing variation in the sample, but have not yet created figures that would be valuable for further analysis. Such a figure would communicate that most of the variation occurs at the second part of the URL (the company indicator), and that subsequent variation after this is comparatively trivial.

High-Tech Sample Codes per Company

Each of the high technology company websites were scraped and indexed using a pre-determined vocabulary. A weighted vector of vocabulary codes are used to assign field of study (FOS) to each of the websites. The frequency distribution of number of FOS codes assigned to a company can be seen in Figure 6.

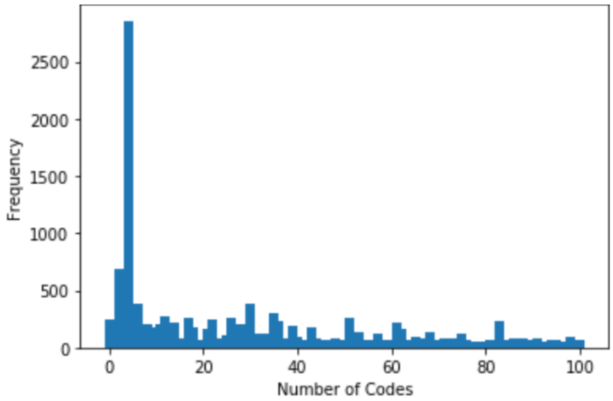


Figure 6 Frequency distribution of number of FOS codes assigned to a company

FOS codes per company have been truncated at 100, but these are sorted to keep the most significant. The figure demonstrates the highly skewed character of the data. Most websites are identified with very few codes (six or so) while others have many associated codes. It is not uncommon to have more than 50 assigned codes. This skewed data reflects two features of the data. The first is that the websites themselves vary in the volume of content provided. The second is that the websites differ in the proportion of recognizable text which can be mined from the data.

Annotated Graph of the FOS Network

Semantic techniques are used to create a graph of field of study (FOS). In this graph, seen in Figure 7, the nodes represent the FOS code, and the weighted arcs indicate the number of firms sharing an assigned FOS code. The most frequent FOS codes, numbering in the thousands, are considered here. Graph layout techniques enable us to attain a high-level survey of the FOS codes in a single graph. Two other analytical features add additional detail to this figure. First the most central FOS nodes to the network are rendered with a larger node size. Second multiple communities are identified in the data, where communities are defined as collections of nodes with more arcs directed into the community than out of the community.

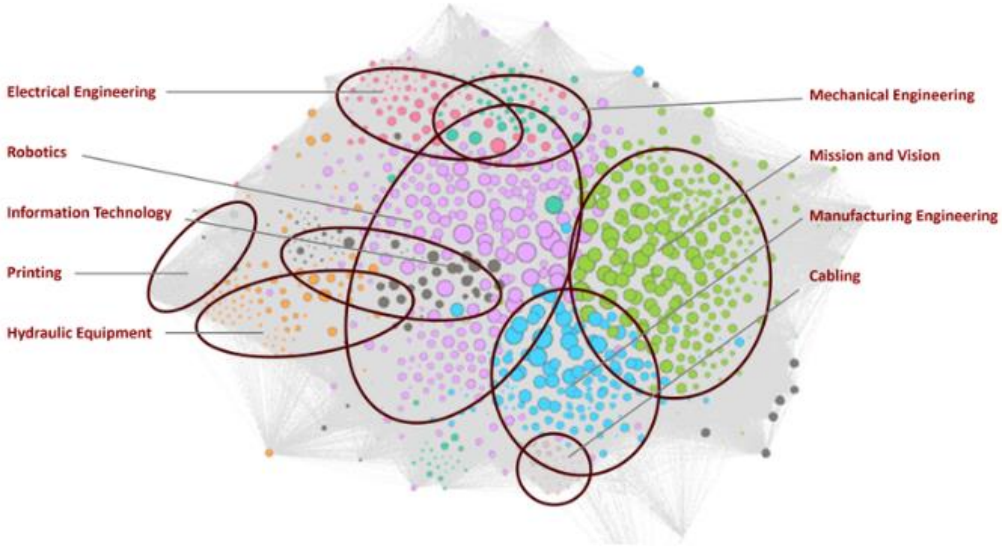


Figure 7 Network representation of the FOS codes, where nodes represent the FOS code, and the weighted arcs indicate the number of firms sharing an assigned FOS code.

A survey of each of the FOS codes, and their associated descriptive labels is performed, giving an indication of the major features of the data. Noted on the graph are the nine major communities detected in the graph. Two of these communities may be of specific significance for BIGPROD activities. These are the information technology and mission and vision communities. The associated web content may be a proxy for some of the firm intangibles the project aims to measure.

Networks of High-Tech Firms with Similar FOS

Figure 8 portrays the high-technology sample of the roughly 14,000 firms in terms of a network. Each vertex is a firm in the sample and each edge is a link formed by shared field of study codes with surrounding firms. The layout uses the Yufan-Hu algorithm which is appropriate and emphasizes the multi-level structure of the data. The vertex colours represent distinct communities discovered by the community detection algorithm. There are many communities identified by the algorithm; here we show just seven of the communities. The remainder of the nodes are in grey.

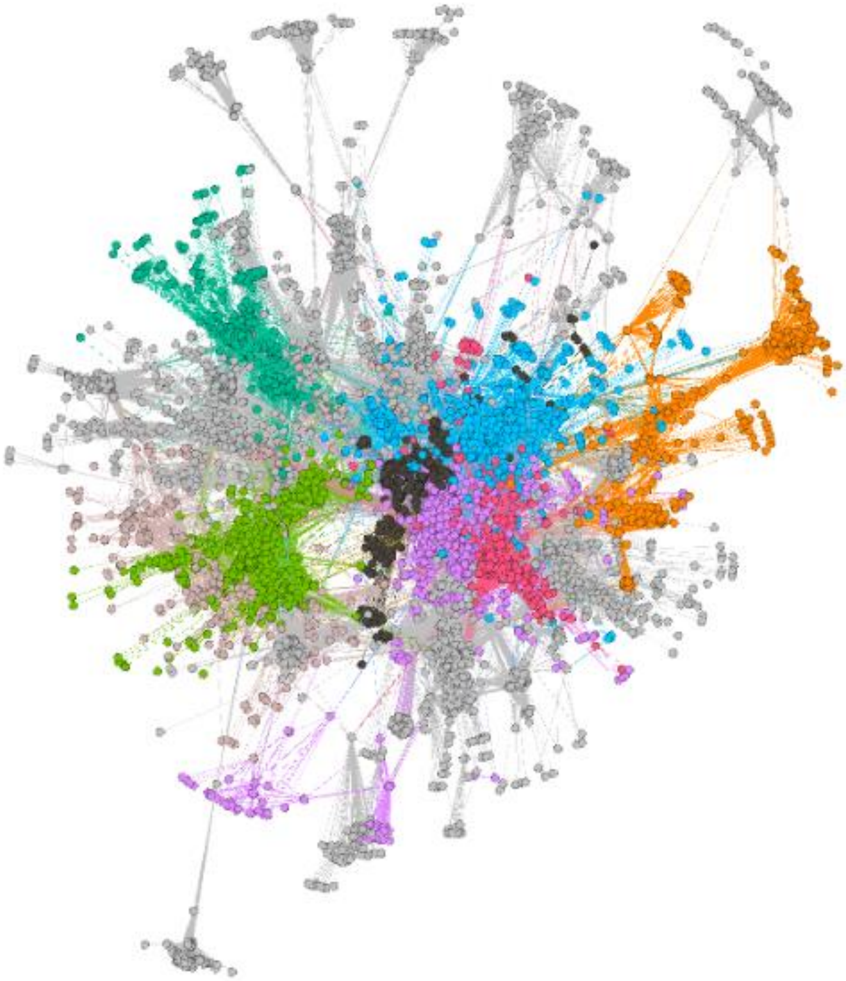


Figure 8 High-technology sample as a network. Each node is a firm in the sample and each edge is a link formed by shared field of study codes with surrounding firms.

Here are a few conjectures about this network. The structure of the network may represent industrial supply chains. Each of the communities may represent a high-technology industry. Nodes in the middle of network are systems integrators which require many skills sets for successful operation. These nodes may also represent multinational corporations in the form of holding companies that own multiple divisions.

Radiating outwards from the chain are the supply chain, with tier one, tier two and later participants in the chain. These are located at the periphery of the network. This may make sense given the specialty skills and investments needed to participate in the chain. Many of these chains seem pinched at the top of the chain, indicating relatively few participants. Several of the industries seem to have multiple distinct supply chains, such as the conglomerations in pink, red, orange and green. These conjectures can and will be tested with a more detailed annotation of the network, using the NACE code, the company names, and the company revenues.

Shown in Figure 9 is the same network structure but with the 2-digit NACE codes used to colour the nodes.

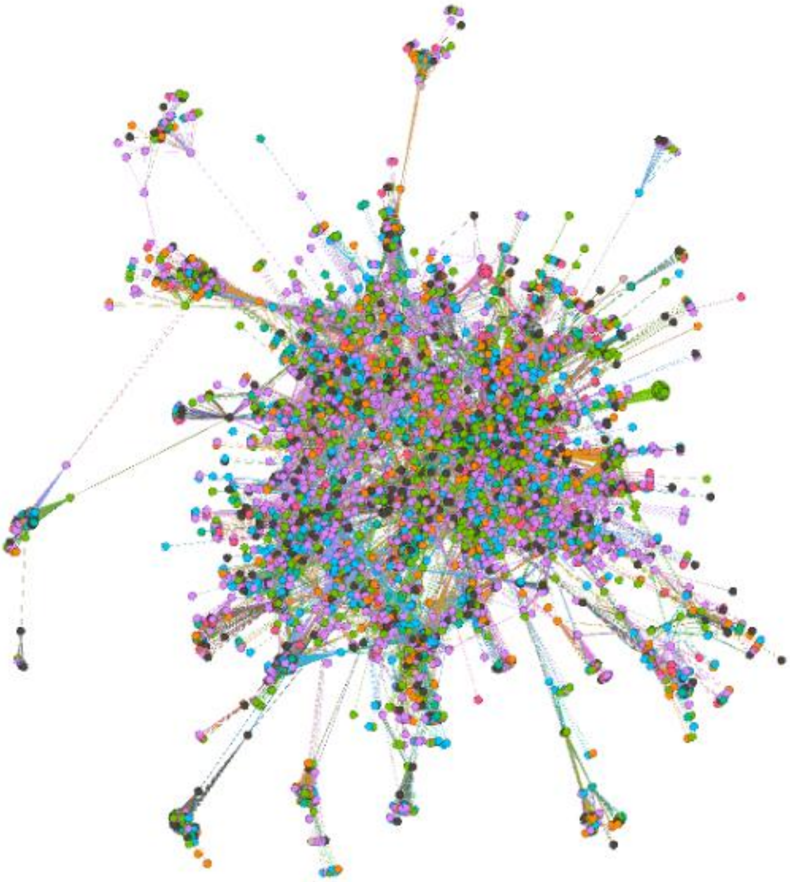


Figure 9 Figure 8 represented using 2-digit NACE codes to colour the network.

As can be seen the network communities do not readily correspond to industrial codes. There are many edges between nodes of the same industry, and these are distributed in a manner that suggest they form a superstructure for the network. Nonetheless these distinct industry networks are overlaid on top of one another, and the network as a whole is highly mixed on the industrial scale. Thus, the network structure does not correspond to our prior hypotheses.

This suggests that the most frequent 5,000 FOS codes used in this representation may represent core competencies for all firms, rather than sector specific or technologically specific knowledge. If industrial specific networks are desired this suggests a selection of FOS codes which can maximally predict or reproduce or assign NACE codes in the absence of labels. This model could be a form of categorical discriminant analysis. Alternatively, a mix of frequent FOS codes and industrially specific FOS codes may be used in the construction of this network.

An alternative hypothesis would be that geography matters. Firms that share physical locations also share sources of knowledge. Sources of knowledge may be represented by the NACE codes in this sample.

Further networks may also be built which show the financial attributes of the companies in the network. We still expect the position of the nodes in the network to be predictive of finance, where the central nodes should have more employees and higher turnover. This hypothesis is based upon theories of complex knowledge in the economy, where companies best able to gather and recombine sources of knowledge attain higher measures of value add. The centre of the network is where such mixing of disparate knowledge will most readily occur.

For more information, please contact

Dr. Arho Suominen (Consortium leader)
Tel. +358 50 5050 354
arho.suominen@vtt.fi

About BIGPROD

BIFPROD is a research project focusing on Big Data based analysis of productivity using webscraped data. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822.

The project partners in the project are Quantitative Science and Technology Studies team, Foresight-driven Business Strategies, 1) VTT Technical Research Centre of Finland, Competence Center Innovation and Knowledge Economy (Coordinator), 2) Fraunhofer ISI, Economics of Knowledge and Innovation team, 3) UNU-MERIT, Maastricht University, 4) Public Policy and Management Institute, 5) Economics of Technology and Innovations, Faculty of Technology, Policy and Management, 6) Delft University of Technology, Economics of Technology and Innovations, 7) Faculty of Technology, Policy and Management, Delft University of Technology

www.bigprod.eu