Hacienda Pública Española/Review of Public Economics 239-(4/2021): 159-197
© 2021, Instituto de Estudios Fiscales
https://dx.doi.org/10.7866/HPE-RPE.21.4.5

Use of Social Network Analysis for Tax Control in Spain[*]

IGNACIO GONZÁLEZ GARCÍA[**]
ALFONSO MATEOS[***]
Universidad Politécnica de Madrid

Received: May, 2020
Accepted: January, 2021

Abstract

The Spanish Tax Agency is an experienced user of big data and has now deployed social network analysis (SNA) tools. SNA tools have led to a qualitative leap in such wide-ranging areas as tax collection, enforcement, control of ultra-high-net-worth individuals, and money laundering. This paper presents a comprehensive overview of the different lines of research, strategies and results of nine projects over the last five years, including the lessons learned.

We present the best practices in pattern discovery, the tools developed for the control of big fortunes and the strategy developed to create a bridge between expert knowledge and SNA technologies. We highlight the results of investigating interposed entities used to channel personal remuneration, complex corporate structures, and opaque companies.

Keywords: Enforcement, Net worth, SNA, Social networks, Pregel, Fraud.

JEL Classification: D85, H26, D31, E21.

1. Introduction

The massive use of data has transformed the most advanced tax agencies into data intelligent tax administrations (PwC, 2018). The introduction of social networks analysis (SNA) tools (Lismont *et al*., 2018) is currently leading to a new qualitative leap.

Spanish Tax Agency (Agencia Estatal de Administración Tributaria, AEAT) annual tax control plans (PCT) have accounted for this new reality. In 2017, one goal of the plan was to investigate the "use of interposed entities to channel personal remuneration with a significant

[*] *Acknowledgements:* This research was supported by the Ministry of Economy and Competitiveness project MTM2017-86875-C3-3-R.

[**] ORCID ID: 0000-0002-6323-677X.

[***] ORCID ID: 0000-0003-4764-6047.

reduction of tax levels" (AEAT, 2017, p. 6, 611). In 2018, the Spanish Tax Agency announced the creation of the High Value Assets Control Office. The media described this change as follows: "New technologies, new information collection systems, like Form 720 for declaring foreign assets, and new information sources (CRS, FATCA, etc.) led to the need for strategy redesign" (Serraler, 2018). This called for the use of new technologies: a new tool was designed to analyze participations, highlighting all the paths through the corporate relations between two taxpayers". The 2019 PCT noted that "Sophisticated analysis mechanisms have to be used to investigate and understand complex financial and corporate structures, and this has led to a major effort in developing specific software tools". Section A.2 of the 2020 PCT (AEAT, 2020, p. 8, 187) focused on wealth analysis. It announced the creation of the Central High Value Assets Control Coordination Unit and the use of specific computer tools, especially for identifying opaque companies).

Public administrations have shown growing interest in the use of big data (Hagen *et al.*, 2019; Vydra and Klievink, 2019) and data analytics (Rukanova *et al.*, 2020). Many techniques that were initially designed to combat financial fraud (Glancy and Yadav, 2011; West and Bhattacharya, 2016; Phua *et al.*, 2010), insurance fraud (Ngai *et al.*, 2011), and credit card fraud (Bhattacharyya *et al.*, 2011) have been adapted to specific tax problems (Bonchi *et al.*, 2020, Pourhabibi *et al.*, 2020).

Throughout the 20th century, the detection of suspicious cases was based on crossing data declared by taxpayers with data imputed on forms declared by other providers of information. A typical case in Spain was the use of Form 347[1] for VAT control. This strategy was gradually rounded out with the use of multivariate statistical techniques (CIAT, 2020; Castellón and Velásquez, 2013). Most of the advanced tax administrations have published the results of significant projects: VAT fraud (Vanhoeyveld *et al.*, 2019), credit misuse in Italy (Agenzia delle entrate, 2007; Andini *et al.*, 2018; Basta *et al.*, 2009), the use of data mining and neural networks for tax control in Finland (Titan, 2012), the use of classification trees in the USA (Murthy, 1998), control of Social Security fraud in Belgium (Van Vlasselaer *et al.*, 2018; Baesen *et al.*, 2015), and the control of direct taxes in Brazil (Matos *et al.*, 2016).

Over the last decade, data scientists have developed new supervised machine learning methods and combined anomaly detection with community detection (Vasudevan *et al.*, 2009) that tax specialists must adapt to the tax field. There is a huge amount of data at the disposal of analysts, and the size of samples used to train algorithms (*learners*) appears to be adequate. In Spain, there are almost 800,000 items by year: in 2015 the numbers were 484,090 items for income tax, 47,053 for corporate tax, 120,699 for VAT and other for customs duties an excises. All these items appear to be trustworthy, because they are the formal and revised outcomes of a process documented in 72 types of reports (termed *actas tributarias* in Spanish). However, this is a very complex problem, and machine-learning techniques need to be adapted.

The tax control problem boils down to two activities: a) selection, that is, the identification of communities of high-risk taxpayers for each tax, and b) inspection of taxpayers. These tasks are performed by different groups of specialists at the Spanish Tax Agency. Inspectors audit the taxpayers that they are assigned in the selection process. Therefore, it is possible to develop tools to support either the process of selection or the process of inspection or, preferably, both. The selection process takes many factors into account: the strategic objectives included in the PCT, candidate background, volume of operations and the fact that some types of companies, like banks and specific sectors, must be controlled. On legal and cost-related grounds, it is not possible to adopt an experimental method. Companies cannot be inspected just because a research group concludes that it would be interesting to assess, for instance, the relationship between level of fraud and some interesting structural parameters like betweenness or inclusion in *n*-cliques, which is the case of the Electra project (see Section 3.2). In the inspection process, the inspector requires the support of the most advanced technology. The control of groups of companies is a human-resource-intensive endeavor. It could use attribute-aware methods, using data related to taxpayers (nodes) or edges (relations) (Wasserman and Faust, 1994) or combine both approaches (Bothorel *et al.*, 2015). We opted to use attributed networks, networks that contain information related to nodes and edges (Chunaev, 2020), because we regard the rich information contained in the tax database as extremely valuable.

This paper describes the many different research lines and research and development effort carried out in the 2015-2020 period in order to highlight the opportunities that SNA techniques offer for reducing non-compliance. In particular, the research lines presented here were inspired by two central ideas:

a) Social network analytics can be used to discover fraud patterns and to facilitate tax enforcement.

  a.1) *Hypothesis 1*. Algorithms can be adapted to the complexity level of the tax control problem.

  SNA techniques originally identified communities and patterns using relationships like friendship or popularity, but many other relationships have to be considered in the tax field. The Spanish Tax Agency uses 49 relationship types, including family, commercial and legal relationships.

  a.2) *Hypothesis 2*. Pattern detection methods are applicable.

  The Spanish Tax Agency uses rule-based expert systems to detect suspicious taxpayers through anomaly detection. The hypothesis was that we could develop non-supervised methods to detect patterns of fraud.

  a.3) *Hypothesis 3*. A SNA approach combined with graphic tools can generate highly interpretable fraud control methods.

  State-of-the art artificial and machine learning tools, like adversarial or dual attention networks, are black box methods. The Spanish Tax Agency has developed neural networks in the past, which, despite their excellent performance, were not used because tax inspectors require interpretable solutions. We have tested the possibility of combining multivariate analysis methods, machine learning, SNA and graphic tools to provide a user-friendly interface for inspectors.

b) State-of-the-art SNA methods can be used to provide tax agencies with much more precise knowledge of the economic reality.

The research over the period reported in this article can be framed within these four strands of the literature[2]:

1. *Community identification* (*Electra project*). Technically, identifying communities in a network is finding node clusters in the graph. The aim in 2016 was to detect communities posing a tax risk using the graph of commercial relationships, identifying groups that had a lot of commercial relationships with each other and only a few with third parties (customers or suppliers). It was implemented by means of a modified version of algorithms based on the *k*-core concept (González and Mateos, 2018b).

2. *Community detection*. These methods allow to reveal meaningful subgraphs (El-Moussaoui *et al.*, 2019). For example, carrousel fraud investigations, being one of the most aggressive forms of VAT fraud. In this case, the objective was to detect communities suspicious of committing this specific type of fraud based on the fraud pattern identified in a previous phase (community identification). The difference between this concept and community identification and between their methods and algorithms has been explained in *Detecting and Identifying Communities in Dynamic and Complex Networks: Definition and Survey* (Vasudevan and Deo, 2018).

3. *Atypical topologies in business relationships*. Unexplained patterns, such as loops of transactions, originating in specific types of companies.

4. *Projects with specific objectives*:

   4.1. Detection of the relationships between companies in a particular sector (telephony service providers) and groups of high-risk taxpayers.

   4.2. Detection of potential false self-employed. The objective was to detect frauds whereby people that are on the company payroll are obliged to accept registration under the Special Scheme for Self-Employed Workers instead of being registered by their employer under the General Social Security Scheme. The company avoids the social security coverage payments, and the worker forfeits many rights. This situation can be detected by looking at the relationships declared on Tax Agency Form 190, that is, if the taxpayer's situation has changed from the previous year and has only one edge in the graph of commercial relationships (the taxpayer has only one customer or the taxpayer's only customers from a group of companies).

4.3. Risks involved in tax payment deferrals implementing an evolution of the ideas developed for consumer credit supervised machine learning (Kruppa *et al.*, 2013; Bao *et al.*, 2019) using the DEMATEL method.

4.4. Taxpayer net worth:

4.4.1. Calculation of wealth and wealth distribution in Spain (Mas, 2020). Detection of interposed companies.

All the hypothesis were confirmed. The lessons learned and the products developed have modelled a new tax control framework.

The rest of the article is organized as follows. Section 2 sets out the problem definition, objectives, state of the art and methodology. Section 3 reports the results and Section 4 outlines the conclusions.

2. Problem definition and strategy

All the research described in this article has a common purpose: tax enforcement. The International Monetary Fund (Crivelli *et al.*, 2015) has estimated the tax gap, that is, the difference between potential tax revenues and what is actually collected, at 600 billion dollars annually. The Organization for Economic Cooperation and Development (OECD) sets the interval at 100 to 200 billion US dollars, associated with a loss of corporate tax revenue ranging from 4% to 10%. The magnitude of this loss in the European Union (Murphy, 2012) is € 825,000 million.

Tax gap, that is the difference between the total amount of taxes collected and the total tax revenues that would be collected under full tax compliance and shadow economy are related. The General Council of Economists reviewed fifteen methods used by Spanish economists to estimate the shadow economy in Spain (Vaquero *et al.*, 2016) and suggested figures ranging from 3%, through 6%, to 15.4% of the country's GDP, that can be compared with the estimations of Friedrich Schneider that is considered the international reference in the field (Scheneider and Enste, 2000; Schneider *et al.*, 2010). The document of the Council also analyzes the VAT-gap in Spain, that had estimated (Murphy, 2012) in € 72,709 M, and using a more precise estimation of the shadow economy estimated the loss of revenue for in Spain, in a broad sense and for all the taxes at € 25,648 million. (Consejo General de Economistas, 2017, p. 75). Other estimations based on microdata for specific taxes, show a 44.34 per cent gap in wealth tax and 41.26 per cent in the inheritance and gift tax (Duran-Cabré *et al.*, 2019).

2.1. Background, state of the art and research

There are 72.5 million taxpayers on the Spanish Tax Agency (AEAT) census of taxpayers, including legal entities and non-residents. The Spanish Tax Agency is staffed by 24,939 persons (2018), of whom 7,485 work in tax management, 4,894 in inspection (19.2%), 4,119 in collection, 3,644 in customs, 1,674 in information technology and 3,123 in other departments (economic administration, human resources, etc.). A total of 21,707 (87,04%) staff are employed at regional branches.

The State Tax Agency conducts *extensive* and *intensive* control actions. Every year hundreds of thousands of extensive controls (direct taxes, customs and excises) are performed because they are low cost, dissuasive (as a result of their immediacy), and do not generate a lot of litigation. A third of these audits target personal income tax (PIT). The process is based on a rule system and supported by the tools characteristic of an advanced electronic administration. *Intensive controls* are performed by inspectors, following a guarantee-based, human-resource-intensive process that requires other types of tools to efficiently select taxpayers and support the inspector during the audit. In 2018 there were 1.58 M of controls (2018), which generated a total result of € 14,984 M. Intensive controls involved 26,984 taxpayers and concluded with 60,445 reports or assessments amounting to 5,378 million euros (AEAT, 2018).

Naive approaches to fraud control regard the taxpayer selection phase as the challenge. This used to be the problem in the past, but this is no longer the case. In the case of the AEAT, a settlement on a quota was reached in on average 85.75% of the cases over the last 25 years for all taxes on which an intensive control was conducted. Basta *et al*. (2009) gives a similar figure, 85.29%, for other countries. The fact is that inspectors start out in the knowledge that, using the contextual information provided in the selection process, there will be a positive outcome in 85% of cases. There is a critical issue with regard to the use of machine learning in tax inspection: the labeled sample that is used for algorithm training. The selected taxpayers are not a random sample of the population. There is bias, since inspected taxpayers were selected based on experts' decisions and not on random sampling.

In the 1990s, the AEAT developed the SERENE neural network system, with four layers and 243 neurons to select candidates for auditing VAT. Its results were impressive: in a 98.38% of the cases, was found a fraud greater than € 600, in 0,81% of the cases a fraud inferior to €600, and only in a 0.81% of the cases VAT fraud was not detected. The problem was the underlying black box technique: the system identified that there was something wrong with the selected candidate but did not provide the inspector with any information about the fraud type. As a result, the tool was abandoned, because, all things considered, less certainty but more information was considered to be a better option.

2.1.1. *Strategic objectives*

The research reported in this paper was conducted by different task forces, each with a specific objective classified as follows:

a) Adapt machine learning techniques to the specific AEAT Inspection Department control objectives (Items 1, 2, 3, and 4.1, 4.2, 4.3 described in Section 1).

b) Develop a new approach to achieve three tactical objectives: *i*) estimation of the net worth of each taxpayer; *ii*) knowledge of taxpayer direct and indirect economic structure; *iii*) provision of income and corporate tax inspectors with insight (4.4). This line of work opened up an unexplored approach for the control of big fortunes with three mainstays.

b.1) *Estimation of taxpayer current net worth*. In the field of direct taxation, a previous, necessary but not sufficient step in the control process is to assess taxpayer tax risk (using parametric or non-parametric tools). This assessment uses an estimation of taxpayer net worth or income and their components. This is not easy: there are many different types of assets, their values, if declared, may not be updated, and some of the assets may be owned indirectly through legal entities. The interest in current taxpayer net worth, which goes back a long way, has increased due to the relevance of the UHNWI (ultra-high-net-worth individuals) economy. Interest was originally academic (Pow, 2011; Hay and Muller, 2012; Beaverstock and Faulconbridge, 2013). However, after the 2016 publication of OECD data, some institutions began to investigate shadow economies (OECD, 2002; Schneider and Enste, 2000; Schneider *et al*., 2010) and big fortunes and Spanish Tax Agency set up the Central High Value Assets Control Coordination Unit (see Section 1).

The number and wealth of the UHNWI shapes the tail of the distribution of wealth, which is important from a number of viewpoints. Inequality influences growth and investment (Persson and Tabellini, 1994; Alesina and Rodrick, 1994) and predicts sociological change.

Paul Krugman's 2007 lecture at the Agenda for Shared Prosperity symposium, titled *The Disappearance of the Middle Class in the United States* aroused interest in the study of the reality of other societies (Hernández, 2014; Credit Suisse, 2015; Luque, 2015). In Spain, a report (Brindusa *et al*., 2018) used data from the Bank of Spain's Household Financial Survey estimated the distribution of the wealth. Traditional academic and social interest in inequality has increased since the publication of Piketty's *Capital in the 21st Century*. This has encouraged the use of more detailed data.

From 2017 onwards, the AEAT established, as a key objective, the control of UHNWI. Inspectors detected the importance of *indirect wealth*, owned by unlisted companies, and a task force, combining the expertise of tax inspectors, data mining specialists and data scientists, was set up to tackle the problem. The approach was to calculate the net worth of each taxpayer, load the detailed data in a warehouse, and connect this information to the graphic tools that inspectors use.

b.2) *Knowledge of business network structures*. Tax inspectors at the AEAT use a tool, called TESEO (EUROSOCIAL, 2016), to visualize and analyze the graph of relationships between taxpayers. After this tool had been built, interest switched from graphical interfaces to the next step: community detection, pattern discovery and community identification (Pourhabibi *et al*., 2020; Furth, 2010; Alhajj and Rokne, 2018). In the tax field, the problem is complex because there are many different types of relationships between companies –juridical, equity participation, commercial, corporate governance–, not all of which have to be taken into account in each and every problem. The techniques must detect patterns in big (70 million nodes, 350 million edges) digraphs (with directed edges), multigraphs (with dozens of relationships in each case), and attributed graphs (using many node and edge data). A new approach had to be created.

b.3) *Provision of inspectors with insight*. After the neural networks experience (SERENE) in the 1990s, it was decided that the priority was to develop systems that could provide inspectors with insight. It was assumed that information like "this company should be inspected because it is in the first percentile when listing the Page Rank of the global network" was useless to the inspectors.

2.2. Data

Table 1 shows a non-exhaustive list of examples of the numbers of each type of tax identification number (TIN) and the number of processed tax returns. The data are available in the AEAT annual reports. We used information from 68 tax return forms from a total of 117 different periodic informative declarations and annual returns related to issues as wide-ranging as gaming tax or childcare expenses.

Table 1
CENSUS OF INFORMATION USED (2018)

Types of TIN in the census	Number
Resident associations	1,055,382
Resident or non-resident foreign nationals	10,220,999
Natural persons	52,200,983
Limited companies	483,131
Business partnerships	3,802
Limited partnerships	613
Cooperative societies	105,902
Limited liability companies	3,233,969
Civil law partnerships (>2008)	432,646
Others	4,002,470
TOTAL	72,525,029

Information available	Number
Tax debts and loans (F. 181 and F. 196)	139,537,599
Current accounts	94,775,840
Financial assets	59,554,618

Forms submitted	Number
Personal income tax returns	19,988,643
Corporate tax (F. 200)	1,529,787
Wealth (F. 714)	207,225

Note: Own elaboration with: a) Form 181. Information Return. Loans, credit facilities and other financial operations pertaining to real estate; b) Form 196. Annual summary of withholdings and payments on account on income from movable capital and income obtained from accounts in all types of financial institutions, c) Form 200. Corporation tax and non-resident income tax and d) Form 714. Wealth Tax. Tax return and income document.

The forms include returns, self-assessments or imputations, i.e.: information about third parties, and they indicate all kinds of relationships. Some are commercial relationships, such as "payments made by A to B", which are contained in VAT Form 347, others are natural relationships, such as "being the mother of" declared on the Income Tax Form 100, others are legal relationships, such as "being a partner of a company" or "company X participates in Y" reported on Corporate Tax Form 200.

2.3. Methodology

The following steps were taken successively throughout the research: a) identification of the necessary local and global data in each project; b) creation of networks; c) creation of algorithms, coding, calculation and attribution of the values of local and global variables to the nodes; d) data processing, and e) results analysis.

a) *Identification of the data*. The prioritized objectives selected by the Inspection Department are transmitted for planning purposes to the IT Department in the last quarter of the year. A task force with two subgroups is set up in due time for each project. One subgroup is composed of data scientists from the IT Department specializing in different technologies (data mining, machine learning, etc.), and the members of the other subgroup are inspectors, mostly from the National Fraud Investigation Office, also specializing in a specific field, such as corporate tax, VAT, etc. From the point of view of SNA, there are four attributed data types: *i*) data declared by taxpayers about themselves; *ii*) data imputed to taxpayers by third parties; *iii*) local calculated data, such as the difference between taxes from the sales declared by taxpayers and the sum of all taxes imputed by their buyers; *iv*) global calculated data, such as the wealth imputed to taxpayers as a result of all their holdings in all the companies in the network. The data that must be used in each project are selected by tax specialists from many tens of thousands of variables. Data scientists suggest structural measures that could be used in the process.

b) *Network creation*. The networks can be easily set up by software tools because all the information items are associated with a tax identification number (TIN). Each network node corresponds to a unique tax identifier (TIN), and both the nodes and the relationships are qualified by the attributes necessary to solve each problem.

c) *Algorithm development*. Algorithms are designed and executed on a cluster of State Tax Agency servers. Most were developed and coded using Python.

d) *Data processing*. It is carried out in the Data Processing Center of the State Tax Agency.

e) *Results Analysis*. The results are analyzed by the above task force, often over a period of many months, refining the results and developing proofs of concept.

After the final conclusions, the Inspection Department decides if the developed technology should be applied, whether the tools should be distributed, and, if so, any changes that should be included in the protocol for taxpayer selection.

2.3.1. *Wealth and social networks*

Taxpayer wealth can be divided into two parts: *direct wealth* (car, house, current accounts, listed company shares, etc.), and *indirect wealth* (participation in unlisted companies). This indirect wealth could be distributed across many companies, some of them possibly based in tax havens, or hidden in companies set up by family members. Although each company duly pays its own corporate tax, some of the companies could have been set up to avoid income tax payments or provide the groundwork for money laundering mechanisms. Nowadays, tax control cannot be confined to the isolated control of companies and individual taxpayers, it should cover all information items, considering the entire social network surrounding the taxpayer.

One of our central ideas was that state-of-the-art SNA methods can be used to provide tax agencies with much more precise knowledge of the economic reality (Section 1). We implemented this idea using an original approach, and there is, to the best of our knowledge, no comparable reference in the literature.

There are natural persons, in our case taxpayers, that are related to each other in the present (e. g., parents and children) or in the past (e. g., former spouses). The inspectors generally need to know: a) the current family network; b) the network of companies partially owned by the taxpayers; c) data declared by the taxpayer, and d) data imputed to the taxpayer.

In the toy example of Figure 1, the analyzed natural person is owner of A, which has a direct participation in company B, which has shares in C and D. Therefore, we consider that the net worth of the natural person should be calculated by adding the taxpayer's personal direct net worth to the market value of his or her participations in these companies.

Direct net worth is calculated by deducting from the value of the taxpayer's assets, his or her liabilities, that is, amounts owed to other natural persons (Form 181), the State Tax Agency, and other institutions (Form 196).

We used Form 200, which is used to declare corporate tax in Spain, to calculate the taxpayer's indirect net worth.

2.3.2. *Calculation of net worth*

A company filling Form 200 must declare the tax identification codes (TIC) of any persons that have economic rights in a company (partners) (Form 200, p. 24), and the participation that the company filing the corporate tax form has in other companies (Form 200, p. 2). This includes tax identification number, percentage share, nominal value of the participation and book value of the company filing the form. For calculating indirect net worth, we consider four relationships, all of which are declared on Form 200: "*x* is a partner of *y*", "*y* has as a partner *x*", "company *x* participates in company *y*", and "company *y* is participated by company *x*". It is compulsory to declare these data on Form 200 only if the established threshold is exceeded (1% for listed companies and 5% for unlisted companies). The studies have shown that the effect of this threshold is not relevant for our purposes.

To calculate indirect net worth are needed:

- Percentage company participation in other companies. The model in Figure 1 shows connections between companies that form paths. In some cases, the path contains only one edge of the graph, an arc, which means that the distance between nodes is 1. In other cases, such as the path between A and D, the length of the shortest path is 2. There is more than one path between two companies or between a person and a company or between companies. The percentage participation between A and C (70.8%) is calculated by iteratively adding up all the different paths and amounts: AC (30%), (ABC = 80% × 50% = 40%), ABDC (80% × 10% × 10% = 0.8%). The percentage participation of a natural person in a company is calculated as above and by extension. In this case, we use one additional edge (100%) that connects the natural person with A.

- Value of participations in unlisted companies. They should be updated according to the real market value. As there is more than one path between companies, it is crucial to avoid duplications during network processing.

2.3.2.1. *Calculation of the updated value of items of company assets and liabilities*

We inventoried the available sources of information (Forms 100, 200, 720, 174, 196, 714, 720, 189, etc.), and information received from other bodies such as the Cadastre and from the Ministry of Agriculture, Fisheries and Food. We identified the boxes and codes that describe each type of wealth on each form. For example, Form 714 includes "savings c/c deposits, financial and other accounts" (Code E), and "public debt and obligations" (Code F1).

A decision was made on the best of the available values for most items. For example, it was decided that the value of a taxpayer's urban property is, in each case, the maximum of three available values: a) the value declared on Form 714 (Boxes A, A1, C, D and M), the cadastral value of the property plus the value declared on Form 720, or the market value according to a calculator provided by the Cadastre.

To calculate indirect wealth, the State Tax Agency has to:

1) Update declared values that have become obsolete. For example, real estate often appears on the declared balance sheets at purchase prices and is depreciated at different rates. To calculate the market price of real estate, we have chosen the greater of: *i*) the cadastral value; *ii*) the value used by public administrations for the settlement of other taxes; *iii*) the known monetary value or the agreed upon amount.

2) Process corporate group information taking into consideration accounting and corporate tax standards. To calculate market values, we compared: *i*) the value on the balance sheet; *ii*) the updated value, if applicable, declared on Form 200; *iii*) the value derived from the consolidated declaration of the group (Form 220).

2.3.2.2. *Detail of the method used to calculate net worth*

We have access to the each company accounting data and balance sheet stated on Form 200. Returning to Figure 1, in the bottom, the natural person has a 50% participation in company A, that has 100%, in B and 50% in C. We should take care not to unduly accumulate the values of the companies more than once along the path, because we have to cross node A twice to accumulate the natural person's participation in B and C. We solved the problem using two types of data:

- Percentage participation in other companies.
- Linked and unlinked values. We split the net worth of each company into two parts, calculating the "linked value" (LV), that is, the value of the participation in other companies and the non-linked value (NLV), that is featured in gray in the graph.

We describe how the two types of data items are used to address the duplication problem in more detail using a toy example in Figure 2, with three natural persons and six companies {N, A, B, C, D, E}.

The main aims are to estimate taxpayer net worth and to estimate the value of the participation of each taxpayer, for example natural person 2, in each company.

We calculate direct wealth from tax returns submitted by natural person 2 and declared by others, and indirect wealth by accumulating the market value of participations in unlisted companies {A, B, C, D}, located downstream in the flow of the participation relationship.

In this example, natural person 2 only holds shares in A. We have explained that the relationship between A and natural person 1 and natural person 2 is declared by A on page 2 of Form 200. The relationship between A and {B, C} is declared by A on p. 24 of Form 200 and includes the percentage participation and its value. This information is crossed with the information provided by B and C.

We could estimate indirect net worth by considering A as a parent and B, C, D as subsidiaries, gathering information about the related parties according to Regulation (EC) No. 1606/2002 of the European Parliament and of the Council of 19 July 2002 on the application of international accounting standards. However, this does not suffice for our purposes because, not only do we need to know total declared net worth, but, for pattern detection purposes, we also need information about each company's updated and current network at market prices.

The procedure is to:

a) Create the network combining all the declared relationships "participates" and "be participated" (one company might not be a declarant).

b) Accumulate, for each company (node), total assets and total liabilities applying the rules explained in Section 2.3.2.1. At the end of this process, each company's net worth is divided into its two components (linked and unlinked assets).

c) Calculate the indirect percentages of participation from the declared data (direct participations). In this case, natural person 2 has a direct participation in A and indirect participations in {B, C, D}. The total percentage participation in each and every case is calculated, as explained above, by adding the percentage along all the paths. The participation of A in D is ABD (0.5×0.25)+ACD (0.5×0.25)=0.25.

d) Accumulate the values (without duplications), for example, the value of the participation of A in D. The Figure 2 shows the process applied to assets, but the process for liabilities is the same. The total value of the participation of A in D is calculated using

$$NW_{Indirect} = \sum_c \sum_p \prod_{l=1}^{l=L} \%PAR_l * NLNW_l \tag{1}$$

The net indirect worth of the participation of a taxpayer in a company is the sum for the *c* companies, in which has a partnership, and for all the *p* paths that connects himself with a company, of the product of % of participations (*PAR*) times the non-linked net worth (*NLNW*) of these participations in the *l* companies that form a path. There are actually two problems: *i*) be sure that the % of participation, of ownership is not duplicated and *ii*) be sure that, even if (1) holds, net worth value of the nodes are not partially duplicated by considering assets in the parent company and subsidiaries. To avoid this problem, we used only the non-linked balance sheet items.

We explain how we have solved these two problems:

Figure 3 illustrates with a real case the solution offered to the first problem above described. Company names have been concealed on the grounds of confidentiality, but they are referenced by several digits of their tax identification number. The graph shows the relations between the companies, using directed arcs to represent participations. There are two horizontal panels illustrating the same information. They reflect the real situation, including companies that are not relevant for this explanation but form part of the real case.

An algorithm, a standard open source Pregel implementation, starts from each node and builds the chains of participated companies:

a) It sends a message to the next node in the path

b) It obtains the percentage participation (90%)

c) It advances to the next node and calculates the product of 90% x 79.91%, which equals 71.91%, and so on, until, after the sixth iteration, the initial node's percentage participation in X..3484, which is 6.619%, is calculated.

The use of the Pregel algorithm assures that each path is traversed and counted once and only once.

The non-linked assets (referred to in Spanish as "no vinculados") are calculated using data from Forms 200 and 220, including the balance sheet items. The rules used to calculate the value of non-linked assets were provided by a group of inspectors specialized in accounting and company group analysis. A close approximation to the linked value can be calculated from the balance sheet by deducting the value declared under items 00118, 00153, 00160 referring to assets and items 00223, 00238, 00243 referring to group and partner companies from the company assets, taking into consideration Art. 5 of Law 16/2007 (legislation on accounting matters) and Art 47.3 in fine of the Code of Commerce. It is actually rather more complex because the analysis is not confined to single companies, and tax groups also have to be accounted for. A company could be part of a group meeting the conditions of Art. 5 of Law 16/2007 above and be either dominant or not dominant. These complex situations are reflected in the information provided by companies on Forms 220 and 220. The system applies specific rules in each case, using data taken from Form 200 only in some simple cases, information provided on Forms 220 and 200, when information is declared by the dominant company, or, in some cases, adding the sum of the values entered under Form 714 codes (G3, G4, and H2) that provide information about the value of shares and participations in unlisted companies.

Using these two tools we apply the idea shown in Figure 2.

2.3.3. *Discovery of latent relationships*

A tax administration must deal with fraudulent business successions, concealment of capital gains and money laundering. To hide the relationship with their assets, fraudsters can get their families to cooperate and use interposed companies. The aim in this case is to detect with SNA tools latent relationships. This is a different, and more speculative, type of problem that the estimation of net worth.

When inspectors investigate financial crimes, they are looking for typologies: cases of people who sell their properties to their relatives before they are seized by the authorities, or "poor" people who participate, through intermediary companies, in corporations that own luxury cars or yachts, or who have their permanent residence in mansions registered in the name of corporations or in tax havens. In many cases, in these descriptions, inspectors use non-observable variables like dominance or control.

Figure 4 shows an individual A and his or her relationships. This individual has a direct participation in four companies {B, C, H, D} and participates directly or indirectly in seven companies. The participations and their values can be calculated with the technique described in the previous section. Since the relationships "to be a partner of" and "to be an administrator of" are known, it is in the inspector's interest to attribute this information to the network to investigate fraud.

Figure 4 shows the difference in results using either one or both relationships (partnership plus administration). When, at the expert's suggestion, the effects of "to be administrator working on behalf of A" are regarded as a 100% participation, A is found to *control* due to the value of his participation {B, C, H, F}and to *dominate* the group formed by {B, C, H, E, G}.

We show the reason with an example. A cannot command the votes of D due that has only owns a 40 % of the shares and has control over H. Because A cannot command D, he or she does not have a domain in F, even though he or she is entitled to receive 55% of the dividends. A has "control" by participation ($1 \times 25\% + 0.4 \times 75\% = 55\%$) but not domain.

We explain some cases of use of SNA concepts in the discovery of latent relationships.

Case 1. Opaque networks

An "opaque network" is a group of companies united by a structure of complex social relations created by a small number of individuals to hide assets acquired through crime. They are the main target of money laundering investigations. In the past, there was a tendency to use typologies in the fight against money laundering (FATF, 2006) but some efforts using the SNA concepts have emerged. An example is the use of the concept of belief propagation (Yedidia *et al*., 2003).

All the companies have dozens of juridical relationships and hundreds of commercial relationships. In most cases, expanding the network around the taxpayer by a three-path diameter unveils hundreds of companies. It is very valuable for the inspector to be able to detect privileged paths in these complex structures. Criminal network analysis focuses on mining useful and reliable structural patterns from such networks. (Brigth *et al*., 2012; Xu and Chen. 2005). The rationale applied in our approach is that, in a money laundering scheme, the beneficial owner has economic influence over the nodes of the network built for this purpose, and it is possible to trace the "flow of influence".

From the algorithmic point of view, the calculus of flow is a well-studied problem, and many algorithms output maximal flow (Goldeberg and Tarjan, 1998). Many studies have used the concepts of source and sink nodes in financial problems (Woods, 2017). We used the minimum cut algorithm (Papadimitriou and Steiglitz, 1998) that outputs the solution transmitting the greatest influence, giving money laundering investigators the opportunity to trace the simplest explanation compatible with the detected evidence and adapted some algorithms (Kiraly and Kovacs, 2012) to calculate influence without duplication and output the maximal flow of control between a person and a company, as well as the flow along each path, even when they share edges.

The State Tax Agency used this approach to investigate opaque networks created with companies with of less than 10,000 euros set up by people accused of money laundering related to drug trafficking. This project proved the hypothesis. It is, in practice, possible to adapt SNA algorithms to deal with latent or calculated complex relationships.

Case 2. Treatment of large family fortunes

The above algorithms, designed to ascertain the net worth of individuals, can be applied to calculate its distribution across the members of a family using instead of commercial relationships the familiar relationships (declared or inferred). This is necessary to investigate cases of corruption and money laundering or the wealth of the members of families with large fortunes that have established companies in different countries. The analysis covers at least two networks (the extended family network and the network of unlisted companies) and many types of relationships because different members play different roles. Some are mere owners, whereas others are administrators or managers.

Figure 5 contains an example of the implementation of the strategy using the extended network. It contains information about individuals who are close to each other, such as P1 and P2, who could be minors, children of ex-partners, or other related people. The possibility of node contraction provides more accurate reporting of the relationship between the individuals who actually control the assets and their estates. If only the relationship "to be a partner of" is used, the only relevant information is that P1 is the majority shareholder in C. If, on the contrary, family relationships are accumulated, we find that the mother (P1) and the daughter (P2) together could control all four companies.

Case 3. Risk of granting tax payment deferrals

Spain's General Tax Law allows for the deferral of tax debts under certain conditions. A company's ability to pay depends on its cash flow and ability to generate resources, which are local parameters, but also on the effect that the company may suffer from the bankruptcy or suspension of payments of companies with which it has commercial relations (Eslam *et al*., 2011; Bouyich, 2017). We used the DEMATEL method, studied at length in the literature on risk analysis. It was used to analyze the overall risk of 1,577 taxpayers who had requested deferrals.

3. Results[3]

We show the results for each of the research objectives.

3.1. Construction of the extended familiar network

In Table 2, we report the results of the process of recursive processing of family relationships. We have inferred relationships from tax returns and by processing historical data and inheritance tax returns.

State Tax Agency has increased the number of relationships to which has access for the purpose of tackling the above problems by 195 %, from 87 million to 257 million. This result opens up the possibility of efficiently using SNA to investigate financial crimes when members of the family are used as straw men.

Table 2
NUMBER OF DECLARED AND INFERRED RELATIONSHIPS

Declared		Inferred	
Spouse	21,843,117	Calculated	137,745,421
Ascendant	29,124,847		
Descendant	36,649,985	Deducted	32,629,192
	87,617,949	INFERRED	**170,374,613**
EXTENDED TOTAL	TOTAL		**257,992,562**

Source: Own elaboration with AEAT data.

3.2. Community identification

Community identification in graphs is based on old techniques (Zachary, 1977), which originally used non-directed graphs. Over the last decade have appeared new tools to treat digraphs (Papadopoulos *et al*., 2012).

In the Electra project (2016), included in Category 1 explained in Section 1, the aim was to detect suspicious communities using network structural characteristics. Hypothesis 2 stated that it should be possible to use non-supervised methods to detect fraudulent structures. One example of a suspicious community is a group of highly interconnected nodes that have only a few links to others. Using this concept in VAT control, the idea was to detect communities of companies that are highly connected by juridical relationships (González and Mateos, 2018c).

In this case, the concept applied was *k*-core, a classical approach in the study of cohesion in SNA (Borgatti and Everett, 2006) which is useful for identifying highly connected groups. For each group (subnetwork), the edges of the group and the edges that connect the group with other companies can be counted. The ratio of these two variables is a risk factor, and the high-risk groups can be selected for modeling. Figure 6 includes one such representation. In this 3D plot, the z-coordinate is the number of companies in an identified group, and the *slope* of each peak is the measure of its degree of isolation, where the ratio between intra-group edges is a risk factor.

Figure 6 (b) illustrates a real case of an identified network of companies specialized in importing electronic equipment from Asia, which, after inspection, was found to have committed VAT fraud.

The hypothesis was confirmed. The concept of cohesion can be used for community identification in tax investigations.

Although it is impossible to quantify the net benefit of these tools because comparison with the control of a “similar” group without their use is out of the question the mere fact that it provides the inspector charged with controlling one of the companies of the community with the total network pattern, the attributed data, and the graphical interface to investigate the network, clearly increases his or her efficiency.

3.3. Community detection

The “carousel fraud” also known as Missing Trade Intra-Community (MTIC fraud), consists on the use of zero-rated intra-Community supplies and VAT deduction on local transactions. Multiple traders create a chain transaction to generate the right to deduct on domestic purchases follow by zero-rated intra-Community supplies. Typically, many traders in the chain does not know about the fraud and only some of them disappear once they have effectively received the VAT amounts ‘incurred’. The detection of these structures is a complex pattern identification problem because not all these transactions have the same structure. As at November 2018, calculations estimating the annual costs of the fraud in. U.E range from € 20 billion up to more than € 100 billion (depending on methodology adopted) (FISCALIS, 2018).

In SNA analytics betweenness centrality is a measurement that describes how important a person is as a link between different networks and we tested and confirmed the hypothesis that was a good predictor in the detection of organizers of carousel fraud. We created an "index of risk" based on the number of paths in the graph that crossed a taxpayer connecting groups detected by the algorithm as suspicious.

The system was used by the National Office for Investigation of Fraud in the investigation of Operation Sith by the ONIF that detected 29 companies that detected a group of fraudsters in the period 2014 a 2016 (Europa Pres, 2017).

These results show that it is possible to systematically use community identification and community detection in the tax field, using structural network parameters, optimization techniques and attributed networks to provide inspectors with added value.

3.4. Detection of atypical typologies in business relationships

Figure 7 shows several examples of results obtained in Category 3 investigations detecting anomalous patterns:

a) *Commercial rings*. The aim was to detect rings of companies that buy from and sell to each other. In some cases, this unusual behavior is for tax purposes, whereas, in others, they may be just moving stock or trying to simulate an unreal business volume. In the commercial ring illustrated in Figure 7, detected with data from Form 347 ten companies set up a ring that forwarded freight of almost equal value.

b) *Daisy*. The commercial daisy shown in the Figure 7 shows a group of six companies detected using commercial data from Form 347. The company in the central node entered into many transactions with the five companies placed around it. Merchandise was sold to and (some months later) bought from the central node at identical prices (without profit). Figure 7 (b) shows the position of this core structure in the community when commercial and juridical relationships and Figure 7 (c) illustrates that algorithms detected a singular structure formed by combinations of rings. In the year 2015, an investigation into 2012 taxes, detected 112 rings and 38 fraudulent daisy chains.

c) *Flow of dominance*. In Figure 7, we show the detection of the flow of influence. Blue dots are natural persons and red dots are legal entities. In this case, the Customs Surveillance Service studied the degree of economic influence of some nodes over others in money laundering investigations. Figure 7 e) shows the degree of influence of the nodes labelled 9, 13, 15 and 17 calculated using the simplex algorithm explained above. Each panel in the figure represents the nodes on the abscissa and the level of influence on the ordinate. The abscissa (node 9) illustrates that only four of the network nodes are relevant, and the ordinate shows the relative value of the dominance.

3.5. Research and Development of new techniques with specific objectives

In this subsection, we will resume the results of original lines of research that use new strategies or algorithms.

3.5.1. *Detection of behavioral anomalies*

Fraud occurs when companies intentionally go bankrupt in order to avoid paying taxes. A new/existing company with (partly) the same structure is founded afterwards and continues the activities of the former company. Van Vlasselaer *et al*. (2016) proposed a method to detect these schemes. We used the idea and created an algorithm to detect companies interposed between an individual that has a big tax debt (> € 10,000) and an individual who is a wealth taxpayer (> € 1,000,000) with a path whose length is less than a specified threshold. A total of 16,358 suspicious paths were detected in 2015, as well as 7,911 rackets connecting the debtors with rich taxpayers.

3.5.2. *Selection of taxpayers for control*

Although the results of existing taxpayer selection systems for VAT fraud control in Spain, where the impact of the fraud is estimated at 2%, are excellent, a new methodology was developed. The new system is based on what is referred to as Bayesian dialysis. The idea is to partition the variable space used in the selection process, taking into account the *density* of the experience, that is, estimate the past results of inspectors in each *box* of the space, the size of the sample, the level of confidence, and use Bayesian statistics to find a new optimum for the next selection process, focusing the new controls on the most promising regions.

The system improves fraudster detection accuracy by 14.01%, raising the average fraudster detection rate in inspected groups from 82.28% to 96.09%. In addition, the fraud detection confidence interval at the 0.025 level improves by 14.03%, which is a qualitative leap in control effectiveness, which reached the 95.51% threshold (González and Mateos, 2020).

3.5.3. *Research related to wealth distribution*

This research should be classed in a new category, because:

a) It is a new approach leading to a new level of accuracy in the control of big fortunes. Therefore, it cannot be benchmarked against other practices.

b) It implies the combined work of three types of specialists: tax and accounting experts, big data analysts and programmers, and data scientists.

c) The results can be used with many purposes

3.5.3.1. *Obtention of the total wealth*

The total value of the Spaniard's net worth cannot be deduced from wealth tax and not only due to the non-compliance, that Durán-Cabré *et al*. (2019) estimated in a 44.3%. In 2017, only 221,437 taxpayers were liable for wealth tax. They declared 0.6 billion euros.

Wealth and wealth distribution were estimated (Brindusa *et al*., 2018) using data from the *Survey of Household Finances*, which is an official survey undertaken by the Bank of Spain included in the National Statistics Plan (BDE, 2017).

Only in a few cases have wealth and wealth distribution been estimated using tax data. Saez and Zucman (2016) studied inequality in the distribution of wealth in the United States using tax data that complement existing data on household wealth (Picketty *et al*., 2018; Saez, 2018; CBO, 2018). We set out to, for the first time, calculate, rather than estimate, and publish total direct and indirect net worth of a country's population based on detailed tax data.

Table 3 shows that the net value of estimated wealth was 4,543 billion euros obtained accumulating data from 41,726 M of taxpayers with net worth distinct from zero (DIS 0). Using tax items, it is possible, in the light of liabilities, to distinguish individuals with positive (NW > 0) or negative (NW < 0) net worth.

Table 3
NET WORTH AND ASSETS (2015)

	DIS 0	NW ≥ 0	NW < 0
Σ (M €)	4,543,210	4,660,118	-116,908
μ	107,648	123,531	-26,459
#	41,726,478	37,308,022	4,418,456

Source: Own elaboration (calculated in February 2020).

We highlight: *i*) the high number of taxpayers with negative net worth (4,418,456 taxpayers), due, in most cases, to real estate mortgages; *ii*) the existence of 9,787,947 individuals with zero net worth (basically minors without any data), calculated by subtracting individuals who have positive or negative net worth (41,726,478) from the total number of registered individuals (51,514,425).

These data could be compared with information from the Survey of Household Finances (Bank of Spain, 2017) a highly complex and specialized survey conducted by the Bank of Spain since 2002 in cooperation with the Spanish Tax Agency and the National Institute of Statistics. It takes as reference the Survey of Consumer Finances (USA) and the Survey on Household Income and Wealth (Italy). This new approach differs in two key respects:

a) Granularity. The approach reported here uses the data of 37,308,032 taxpayers instead of samples used in the Survey (5,962 in 2005 and 6,413 in 2017).

b) Assets. The Survey of Household Finances does not include indirect wealth, that is, the net worth of participation in non-listed companies. On technical grounds, the

Survey of Household Finances over-represents some intervals in the sample using data from Form 714 (Wealth Tax). A random sample of 5,962 households should include only 52 wealth tax declarants instead of the 536 selected in the 2005 Survey of Household Finances. This is the best option with the data used, with a standard error of 1.6 (Hospido, 2010), but our results show (see Tables 4 and 5) that indirect wealth, which is the most important wealth component in the €10 M to €100M interval, does not correlate with total wealth.

3.5.3.2. *Components of the total wealth*

Table 4 shows the breakdown of the direct and indirect wealth of the 37.3 million individuals (listed in the second column of Table 4, who have positive net worth).

Table 4
DISTRIBUTION OF WEALTH AND ITS COMPONENTS

	Individuals	Net Worth	Assets	Liabilities	Indirect
I	26,218,699	690,009	982,852	306,614	7,625
II	10,659,077	2,572,521	2,881,096	329,594	81,947
III	414,825	889,979	1,068,992	184,812	230,274
IV	14,929	321,008	415,275	94,300	191,296
V	492	135,154	187,089	51,935	98,17
Σ	37,308,022	4,608,696	5,535,259	966,627	609,313

Source: Own elaboration (data revised in February 2020 referring to 2015).

To clarify the distribution, it is divided into five sections: Net worth (I) from 0 to € 100,000; (II) from 100,000 to 1,000,000; (III) from 1,000,000 to 10,000,000; (IV) from 10,000,000 to 100,000,000; (V) over € 100,000,000. Table 4 contains six columns: 1) Taxpayer category; 2) Number of individuals; 3) Net worth; 4) Assets; 5) Liabilities, including include AEAT debts and third-party obligations; 6) Net worth of participations (of over 5%) in unlisted companies.

Table 5 shows the enormous importance of indirect wealth in relation with net worth. While indirect net worth accounts for 0.78% of the wealth of most individuals, it represents 52.47% for the wealthiest group.

Table 5
RATIOS

	Real estate	Accounts	Financial	Others	Indirect
I	75.35	14.83	4.45	4.59	0.78
II	70.46	13.9	7.59	5.2	2.84
III	41.25	6.67	15.06	15.48	21.54
IV	9.98	3.03	18.27	22.66	46.07
V	1.44	1.4	16.8	27.88	52.47
Σ	58.82	11.43	9.59	9.15	11.01

Reports about wealth distribution (OXFAM, 2018) had estimated the percentage of the wealth owned by the richest 1%, 5% and 10% of the population, providing an intuitive measurement of inequality across the distribution. We have improved the estimation, showing that the richest 1% owns 29%, rather than 25.1%, of wealth, as estimated so far (Brindusa *et al.*, 2018), the richest 5% accounts for 48%, and the richest 10% increases its share of wealth from 53.8% to 60.33%.

In the knowledge that the studies carried out on wealth in Spain, such as the one mentioned by the Bank of Spain and other organizations, are of high quality and are very useful insofar as they are comparable with other countries, we have not pursued our analysis any further. However, as more and more researchers use such detailed tax data, the accuracy of this alternative approach to household wealth distribution will improve, and it will be possible to gain more detailed knowledge, and a better control, of the behavior of UHNWI.

Thanks to this detailed study, we have been able to quantify, for what is, to the best of our knowledge, the first time, the total wealth of the population is Spain, the distribution by intervals of wealth, and the indirect wealth and obtained a better estimation, with a different method and microdata of the inequality. We believe that other techniques are prone to underestimate because they don't consider, at least in detail, the 11.01% of indirect wealth (Table 5).

3.5.3.3. *Results that have been used in other types of studies (Econophysics)*

In 1987, Nobel laureate Ken Arrow organized the conference *Economics as a complex evolving system* with two physicists (Phil Anderson and David Pines), promoting the idea of applying complex systems theory to economics. Specialists have tried to solve problems like excess volatility and improve the results of the Dynamic Stochastic General Equilibrium Model (Bouchaud, 2019, p. 2), as well as patterns observed in income and wealth distribution (Mantegna and Stanley, 2001; Chaterjee, 2005). Chakrabarti *et al.* (2013) reviewed these studies in detail.

Atkinson (2016) said: "The upper tail of the income distribution has long been a source of fascination for economists... for a recent review see Benhabib and Bisin (2016)". There are many empirical studies that have tried to confirm the empirical law proposed by Pareto (Yakovenko *et al.*, 2009; Santarelli and Thurick, 2006). They tried to fit the observed distributions to a number of theoretical curves: *i*) two-parameter curves (log-normal, gamma, Weibull); *ii*) three-parameter curves (generalized gamma, Singh-Maddala and Dagum), and *iii*) five-parameter curves (generalized beta) (Bandourian *et al.*, 2003). Economists have a fondness for log-normal distribution curves (Montroll and Schlesinger, 1982), whereas statisticians and scholars researching complex systems proposing analogies between such economic processes and temperature and entropy, that is, physical methods, prefer gamma distributions.

These researchers have to deal with the fact that the tax administration does not publish detailed tax data. As already mentioned, indirect wealth is not homogenously distributed, and, without the knowledge of the distribution of this component, econophysicists are using a biased wealth distribution as an empirical reference in their distorted models. As a contribution to this field (González and Mateos, 2018c) published data from the research described in Section 3.1.3. As a collateral result of this research, we found that fitting the curve that relates the number of taxpayers to their wealth in a Ln–Ln diagram (a standard procedure in these wealth distribution studies), the ratio between an individual's total wealth and the distance between the individual and his or her wealth appeared to be adjusted according to

$$R = 2e^{-\frac{2}{3}d} \qquad R^2 > 99\% \tag{2}$$

The presence of integers in the coefficients and of a rational number in the exponent moved us to consider the possibility of there being some kind of *natural law*, in the social network topology emerging from the decisions by company founders. This opens up the possibility of considering the use of this empirical law in the same way that Benford's law was used in the past to detect fake invoices. The curves representing indirect net worth through interposed companies can be adjusted by:

$$R_M = Ae^{-\frac{B}{3}d} \tag{3}$$

This is an incidental result, and its value requires further evaluation.

The relevant fact is that the detailed data obtained are useful beyond the field of inspection and that using microdata, the fitting of the observed distribution can be explained as the effect of a mixture of subpopulations (entrepreneurs and non-entrepreneurs).

3.5.3.4. *Community discovery using SNA concepts*

This investigation combines many of the lessons learned, uses the data obtained with other purposes. We believe that proves that SNA concepts provide and additional value, even a qualitative leap in the control of tax fraud.

The research question was: Is it possible to detect patterns of fraud in attributed networks using SNA concepts?

In this case, we used the concept of *ego networks* (Akoglu *et al.*, 2010). *Ego* refers to a focal node connected to others in a network. We will explain our use of the concept of egonet with a real case. In Figure 8, many companies are connected in a network. When a taxpayer, for example O, is selected for inspection, the inspector must decide the type or relationships that are useful for the specific type of control (commercial relationships, imports, juridical relationships, etc.).

With this input, the system should have the capacity to expand step by step giving the first level of companies connected (A, B, C), and letting the inspector filter the output with rules (some nodes, as telephonic companies or banks, should be eliminated in most of the cases because, on the contrary, the expansion of the network contains many unnecessary relationships). In the example, we have extended this first step until level three, obtaining the path between O and “1”. We consider this subnetwork “egonet of O”. The system created lets to treat the problem downstream or to obtain which ones are the owners of those related with “1”.

We reproduce the rationale of this investigation using Figure 9 as a toy example.

State Tax Agency studied the 2,548,480 *ego networks* (2017) whose focal points were people who had participations in companies. 2,334,274 of these individuals were sole owners or participated directly in companies and 214,206 participated through intermediaries in companies at distance greater that one.

A total of 1,427,455 companies filed a corporate tax return (F 200) in 2017. Of these 262,083 (18.36%) had participations in others as is the case of R or S in Figure 9. Some natural persons directly control companies and/or participate in others. NP4, NP5 and NP6 directly participate in companies. NP3 participates in H, and indirectly in another one at distance 2.

Table 6 summarizes the data on the relationships of participation at each distance. As a particular case, 8,879 different paths were found between legal persons at a distance of 3. We observe the existence of 14 cases in which one legal person owned 100% of the property of another after having interposed five companies. As we were looking for “hidden assets”, we thought that was useful to study the length of the path that connects natural persons and their assets. The results are shown in Table 7.

Table 6
PARTICIPATION RELATIONSHIPS BETWEEN LEGAL PERSONS (2017)

D	>0	>10	>25	>50%	>75	>90	100
1	395,313	172,355	147,421	113,015	75,132	67,219	49,179
2	210,249	34,936	24,889	15,360	9,865	8,219	5,314
3	136,932	8,879	5,398	2,594	1,795	1,431	886
4	82,457	2,552	1,405	711	22	317	203
5	62,684	798	400	197	110	80	51
6	38,275	252	122	57	35	24	14
7	29,755	78	36	17	15	8	5
8	18,513	119	14	7	6	2	1
9	12,745	2	4	1	1	0	0

Source: Own elaboration.

The first column of Table 1 shows the distance between the taxpayers and the companies in which they hold shares. The data are organized in three blocks according to the total percentage participation. Each row states the number of identified individuals (NP), the number of companies in which they participate (MS), the number of companies that are different from each other (#EM). We find, for example, that 1,101,172 people directly owned more than 50% of a company at distance 1. The rationale of this approach is that the activities of the companies placed in the bottom right-hand corner of the tables are suspicious, because, unless there are economic grounds, the benefit for natural persons of interposing so many companies between themselves and their assets is unclear.

We obtained that 2,182 = (1,624 + 293 + 61 + 19 + 160 + 18 + 5 + 2) companies have the "suspicious" trait of being "controlled" by a natural person at distance 4, with at least three interposed companies, open a line of research that could improve selection and provide inspectors with clues.

Table 7
DISTANCES TO PARTICIPATED LIMITED COMPANIES AND LIMITED PARTNERSHIPS

D	All			PR > 50%		PR = 100	
	NP	MS	#EM	NP	#EM	NP	#EM
1	1,889,919	2,604,892	1,253,581	1.101,172	982,378	289,600	329,144
2	182,170	414,644	139,015	27,305	45,136	4,939	6,977
3	46,275	183,087	43,933	3,100	7,147	374	862
4	14,812	87,985	17,881	505	1,624	50	160
5	5,442	45,367	8,825	91	293	9	18
6	2,264	28,850	4,714	20	61	2	5
7	998	19,537	2,760	3	19	1	2

Source: Own elaboration (2017).

Table 8 shows the results (2017) using instead of relationship of shareholding the *dominance* concept used by experts. The first column (D) is an indicator of distance. The second (NP) is the number of individuals who are identified as having a dominance relationship at this distance. The third (LE) is the number of legal entities with which they established a relationship. Columns 4 and 5 show the increase in the taxpayers (number of times) regarded as having a participation according to the dominance relationship specified by the experts. The last column (NP Fam) shows the additional increase if the network of extended family relationships is connected. As usual, the companies in the bottom right-hand corner are considered riskier, and this information is used in the selection process, also providing inspectors with insight into the inspected companies.

Table 8
DOMINANCE RELATIONSHIPS (2017)

D	NP	LE	(NP)	(LE)	NP Fam
1	3,078,885	5,671,535	30.43	4.55	1,113,610
2	310,174	394,187	11.36	7.87	125,844
3	112,082	171,013	36.16	22.06	46,665
4	56,501	75,109	111.88	42.36	24,147
5	13,365	22,035	146.87	67.59	5,869
6	3,949	5,106	197.45	72.94	1,701
7	1,080	1,228	360	53.39	486
8	675	692			306
9	1	2			1

Source: Own elaboration with AEAT data.

We were extremely surprised (Table 9) to find that the mean value of net worth and the mean value of the number of workers could increase with the distance. After filtering the network, leaving only companies with net worth of less than 75 million euros, that is, having eliminated firms listed as banks and oil companies, the phenomenon is accentuated. It seemed a logical conclusion that the creation (for example) by 998 taxpayers of 2,760 companies with a mean net worth of € 13 M and a mean of 140,9 workers, numbers much bigger that those of the companies directly owned, interposing six companies could be only intentional and that should be studied.

Table 9
DISTRIBUTION OF WEALTH AND ITS COMPONENTS BY SECTIONS (2015)

D	All	< 75 M €	
	Net Worth	Workers	Net Worth
1	414,108	10.28	434,212
2	1,862,322	35.5	1,837,556
3	3,343,302	63.33	1,963,885
4	6,366,615	103	2,342,837
5	9,466,406	137.98	2,603,259
6	13,107,250	140.93	2,656,064
7	12,921,629	123.03	2,735,448

Source: Own elaboration with AEAT data.

After a tax analysis, specially of the taxpayer that owned a 100% of two companies at a distance 7, we concluded that only in some specific cases, wind power companies distributed across several regions of the country, might this make some economic sense. In the other cases, the natural persons that build this type of long structures and accumulate assets at these distances are suspicious and should be investigated.

4. Concluding remarks

The review outlines the lessons learned from our real-world experiences. We have concluded that social network analytics can be used efficiently to discover and control fraud patterns and that state-of-the-art SNA methods can be used to provide tax agencies with much more precise knowledge of the economic reality.

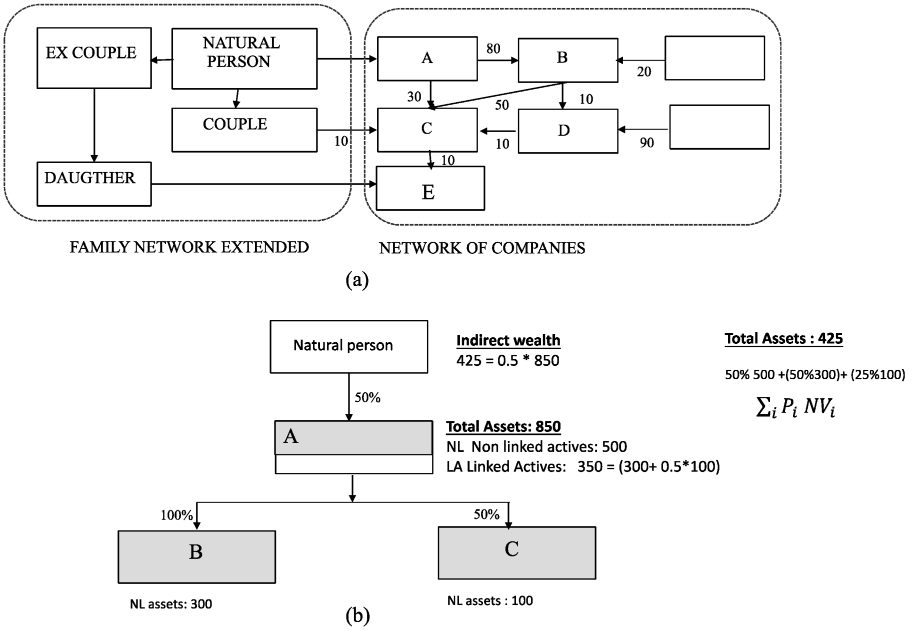
We have discussed cases of use in tax control: non-supervised community identification applied to VAT, community detection applied to VAT, control of deferrals, control of big fortunes, in criminal investigation as in the case of pattern detection in the investigation of money laundering and the benefits associated to this detailed knowledge with other purposes. The methods and tools described now constitute a state-of-the-art framework for the High Value Assets Control Office.

The initial hypotheses were confirmed. Tax inspectors nowadays use concepts and technologies to detect new forms of fraud and to understand how to most efficiently select companies for tax control. Combined with machine learning techniques, they provide a new frontier for tax administrations.

To conclude our review, we propose a possible direction for future research. Embedded intelligent agents are used in tax reporting and will be used by tax administrations to offer taxpayers pre-filled forms in a near future. The combination of SNA methods with artificial intelligence concepts like intelligent agents could be the essential components of a new framework for risk analysis and for the provision of advanced tax agency services.

Annex

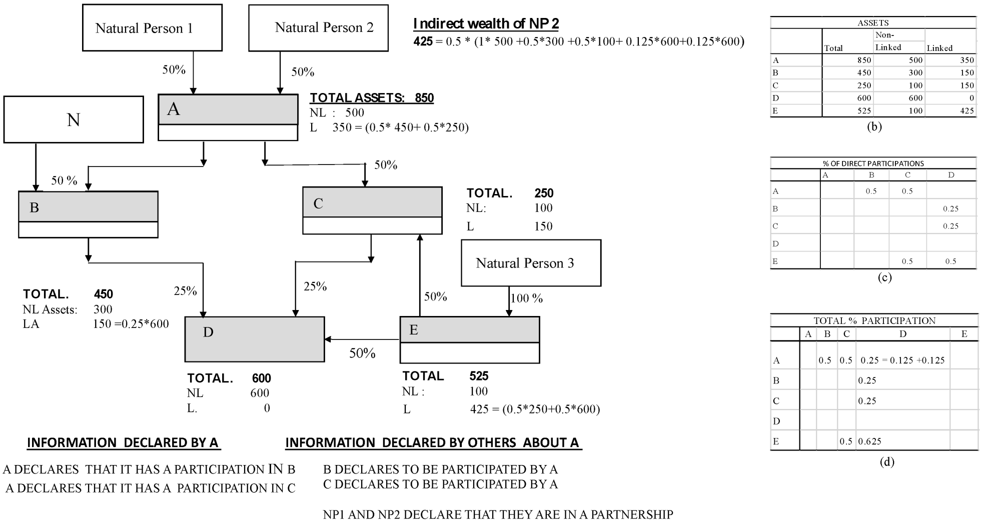
Figure 1
RELATIONSHIPS BETWEEN INDIVIDUALS AND BETWEEN COMPANIES



(a)

(b)

Figure 2
CALCULUS OF THE INDIRECT NET WORTH


Natural Person 1
Natural Person 2
Indirect wealth of NP 2
425 = 0.5 * (1* 500 +0.5*300 +0.5*100+ 0.125*600+0.125*600)
50%
50%
N
A
TOTAL ASSETS: 850
NL : 500
L 350 = (0.5* 450+ 0.5*250)
50 %
50%
B
C
TOTAL. 250
NL: 100
L 150
Natural Person 3
TOTAL. 450
NL Assets: 300
LA 150 =0.25*600
25%
25%
50%
100 %
D
E
50%
TOTAL. 600
NL 600
L. 0
TOTAL 525
NL : 100
L 425 = (0.5*250+0.5*600)
INFORMATION DECLARED BY A
INFORMATION DECLARED BY OTHERS ABOUT A
A DECLARES THAT IT HAS A PARTICIPATION IN B
A DECLARES THAT IT HAS A PARTICIPATION IN C
B DECLARES TO BE PARTICIPATED BY A
C DECLARES TO BE PARTICIPATED BY A
NP1 AND NP2 DECLARE THAT THEY ARE IN A PARTNERSHIP

ASSETS			
	Total	Non- Linked	Linked
A	850	500	350
B	450	300	150
C	250	100	150
D	600	600	0
E	525	100	425

(b)

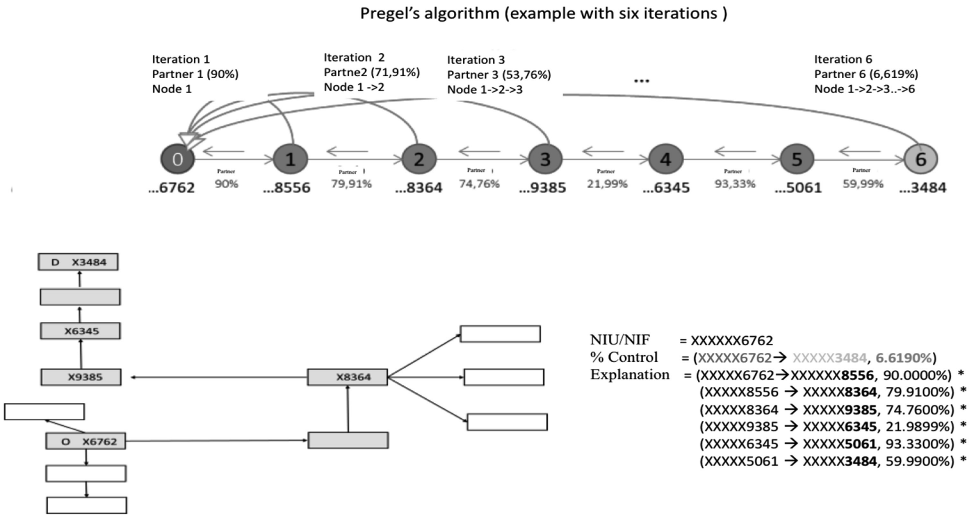
% OF DIRECT PARTICIPATIONS				
	A	B	C	D
A		0.5	0.5	
B				0.25
C				0.25
D				
E			0.5	0.5

(c)

TOTAL % PARTICIPATION					
	A	B	C	D	E
A		0.5	0.5	0.25 = 0.125 +0.125	
B				0.25	
C				0.25	
D					
E			0.5	0.625	

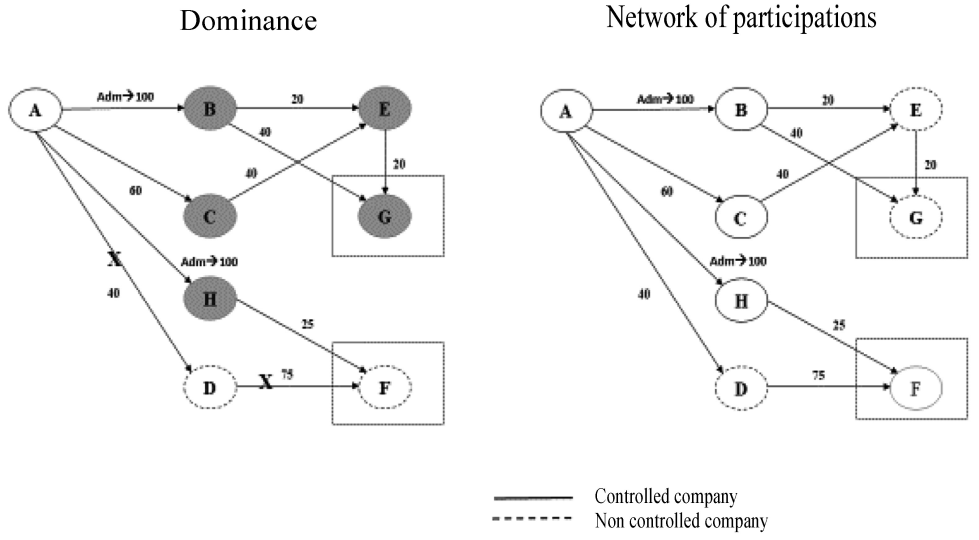
(d)

Figure 3
USE OF THE PREGEL ALGORITHM TO CALCULATE COMPANY PARTICIPATIONS



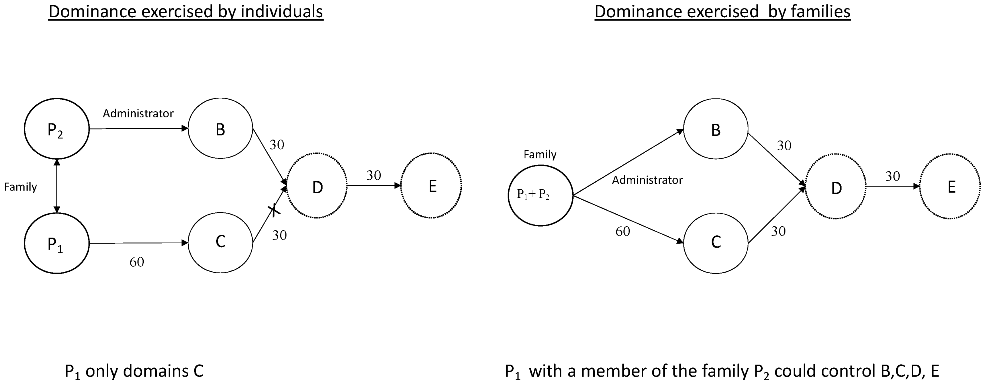
NIU/NIF = XXXXXX6762
% Control = (XXXXX6762→ XXXXX3484, 6.6190%)
Explanation = (XXXXX6762→XXXXXX**8556**, 90.0000%) *
(XXXXX8556 → XXXXX**8364**, 79.9100%) *
(XXXXX8364 → XXXXX**9385**, 74.7600%) *
(XXXXX9385 → XXXXX**6345**, 21.9899%) *
(XXXXX6345 → XXXXX**5061**, 93.3300%) *
(XXXXX5061 → XXXXX**3484**, 59.9900%) *

Figure 4
CONTROL BY PARTICIPATION AND CONTROL BY DOMINANCE


Dominance

Network of participations

Figure 5
USE OF EXTENDED FAMILY RELATIONSHIPS AS AN EXAMPLE OF INFERRED RELATIONSHIPS

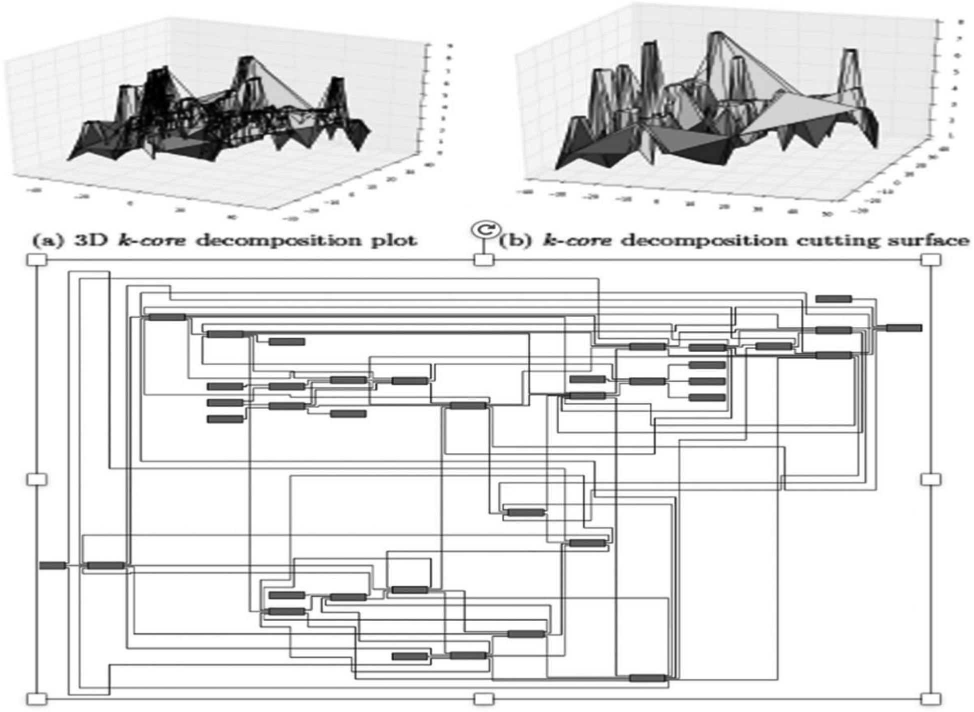

Dominance exercised by individuals

Dominance exercised by families

P1 only domains C

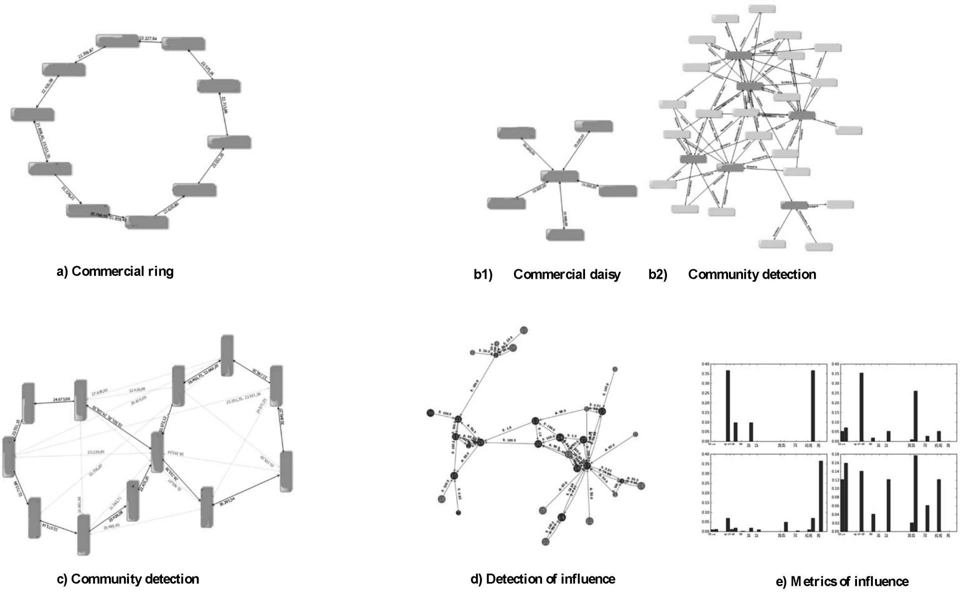
P1 with a member of the family P2 could control B,C,D, E

Figure 6
COMMUNITY DISCOVERY. AN EXAMPLE IN VAT FRAUD


(a) 3D *k-core* decomposition plot

(b) *k-core* decomposition cutting surface

Figure 7
COMMUNITY IDENTIFICATION AND COMMUNITY DETECTION


a) Commercial ring

b1) Commercial daisy b2) Community detection

c) Community detection

d) Detection of influence

e) Metrics of influence

Figure 8
USE OF THE CONCEPT OF EGONET

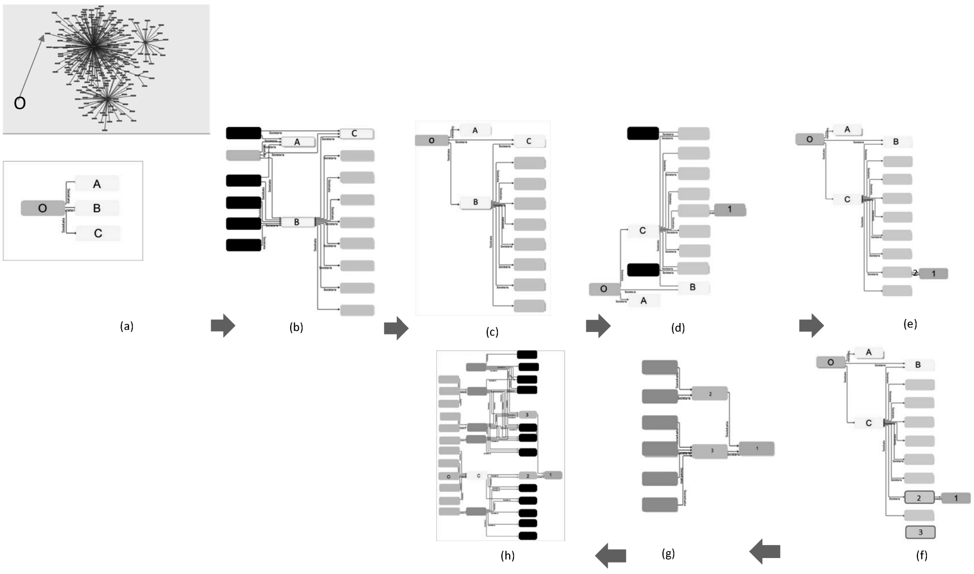
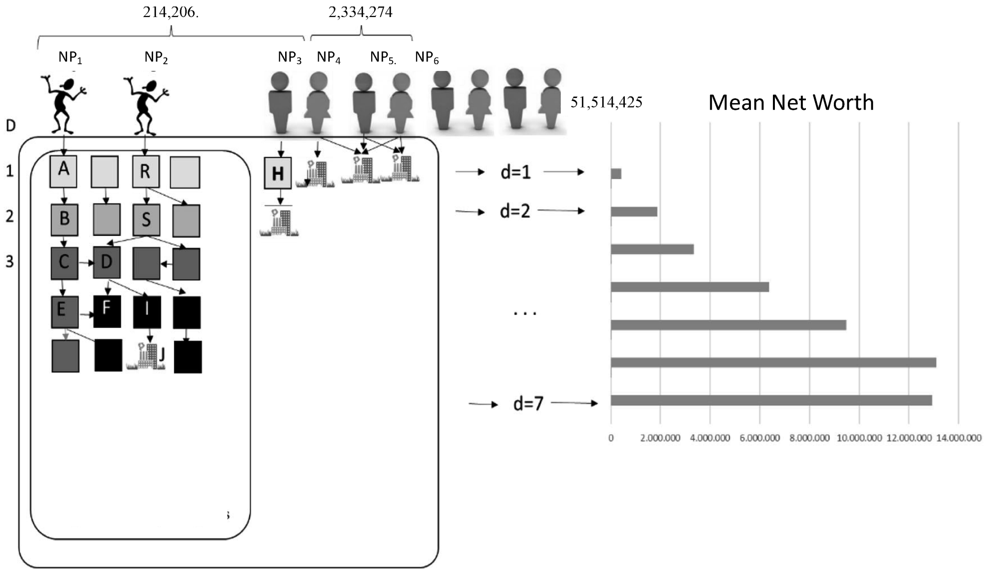


Figure 9
PROVIDING INSIGHT FOR CORPORATE TAX AUDITORS



Notes

[1] Form 347 and all informative tax return forms are described in full detail at www.aeat.es. https://www.agenciatributaria.gob.es/AEAT.sede/Inicio/Procedimientos_y_Servicios/Impuestos_y_Tasas/Declaraciones_Informativas/Declaraciones_Informativas.shtml.

[2] Although this research used AEAT data, it AEAT is academic research, and the results have not been reviewed by the AEAT. We therefore accept full responsibility for the conclusions.

[3] The results are the fruit of our academic research and do not necessarily represent or match the criteria of the State Tax Agency.

References

AEAT (2017), "Resolution of January 19th, 2017, of the General Directorate of the State Tax Agency approving the general guidelines of the 2017Annual Tax and Customs Plan", *Official State Gazette* (28-01-2020): 6, 611. All resolutions are available on the AEAT website www.aeat.es.

AEAT (2018), "Resolution of January 8th, 2017, of the General Directorate of the State Tax Agency approving the general guidelines of the 2017Annual Tax and Customs Plan", *Official State Gazette* (23-01-2020).

AEAT (2019), "Resolution of 11st January 2019 of the General Directorate of the State Agency for Tax Administration approving the general guidelines of the 2019 Annual Tax and Customs Control Plan", *Official State Gazette* (17-01-2020).

AEAT (2020), "Resolution of 21st January 2020 of the General Directorate of the State Agency for Tax Administration approving the general guidelines of the 2020 Annual Tax and Customs Control Plan", *Official State Gazette* (28-01-2020).

Agenzia delle entrate, https://www.agenziaentrate.gov.it/portale/documents/20143/234267/Analisi+2007+risk+analysis_risk_analysis_tax_evasion.pdf/c37eed28-601f-d836-6fb4-275cb36ab65d.

Akoglu L., McGlohon M. and Faloutsos, C. (2010), "OddBall: spotting anomalies in weighted graphs", in: *PAKDD, Hyderabad*, 410-421.

Alesina, A. and Rodrik, D. (1994), "Distributive politics and economic growth" *Quarterly Journal of Economics*, 109: 465-490.

Alhajj. R. and Rokne, J. (Editors), (2018), "Encyclopedia of Social Networks: Analysis and Mining. Second edition", *Tang Jie entry "Inferring social Ties"*, 1067-1075.

Andini, M., Ciani, E., Blasio, G., D'Ignazio, A. and Salvestrini, V., (2018), "Targeting with machine learning: An application to a tax rebate program in Italy", *Journal of Economic Behavior & Organization*, 156: 86-102.

Atkinson, A. B. (2016), "Pareto and the upper tail of the income distribution in the UK: 1799 to the present", Centre for Analysis of Social Exclusion London School of Economics, Paper prepared for a *special issue of Economica in honor of Frank Cowell.*

Baesen, B., Van Vlasselaer, V. and Verbecke, W. (2015), "Fraud analytics using predictive, and social network techniques. A guide to data science for fraud detection", Wiley.

Bandourian R., McDonald, J. B. and Turkey, R. S. (2003), "A comparison of parametric models of income distribution across countries and overtime", *Estadística*, 55: 135-152.

Bao, W., Lianju, N. and Yue, K. (2019), "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment", *Expert Systems with Applications*, 128: 301-315.

Basta, S., Fassetti, F., Guarascio, M., Manco, G., Giannotti, F., Pedreschi, D., Spinsanti, L., Papi, G. and Pisani, S. (2009), "High quality true-positive prediction for fiscal fraud detection", in: *2009 IEEE International Conference on Data Mining Workshops*: 7-12.

Bank of Spain (2017a), "Survey of Household Finances". Available online at https://www.bde.es/bde/en/areas/estadis/estadisticas-por/encuestas-hogar/relacionados /Encuesta_Financi/.

Bank of Spain, (2017b), "Encuesta financiera de las familias (EFF) 2017: Métodos resultados y cambios desde 2014". Available at: https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/InformesBoletinesRevistas/ArticulosAnaliticos/19/T4/descargar/Fich/be1904-art38.pdf.

Beaverstock, J. and Faulconbridge, J. (2013), "Wealth segmentation and the mobilities of the super-rich: A conceptual framework", in: T. Birchnell and J. Caletrio, ed., *Elite Mobilities, 1st ed.*, Routledge.

Benhabib, J. and Bisin, A. (2016), "Skewed wealth distributions: Theory and empirics", *NBER Working Paper,* 21924, National Bureau of Economic Research, Cambridge, Mass.

Bhattacharyya, S., Sanjeev, J. and Westland, C. (2011), "Data mining for credit card fraud: A comparative study", *Decision Support Systems*, 50(3): 602-613.

Bonchi, F., Giannotti, G., Mainetto, D. and Pedreschi (1999), "Using data mining techniques in fiscal fraud detection", in: M. Mohania and A. M. Tjoa (Eds.), *Data Warehousing and Knowledge Discovery: First International Conference*, DaWaK'99 Florence, Italy, August 30-September 1. Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, 1999: 369-376.

Borgatti S. P. and Everett, M. G. (2006), "A graph theoretic perspective on centrality", *Soc Netw*, 28: 466-484.

Bothorel, Cecile, Cruz, Juan David, Magnani, Matteo and Micenková, Barbora (2015), "Clustering attributed graphs: Models, measures and methods", *Netw. Sci*., 3(3): 408-444.

Bouchaud, J. P. (2019), "Econophysics: Still fringe after 30 years?", https://arxiv.org/abs/1901.03691v1.

Bouyich, C. (2017), *Social network analysis for fraud detection*, TFM. UPM. ETSII.

Bright, D. A, Hughes, C. E. and Chalmers, J. (2012), "Illuminating dark networks: a social network analysis of an Australian drug trafficking syndicate", *Crime Law Soc. Chang*, 57(2): 151-176 .

Brindusa, A., Basso, H., Bover, O., Casado, J. M., Hospido, L., Izquierdo, M., Kataryniuk, I., Lacuesta, A., Montero, J. M. and Vozmediano, E. (2018), "Inequality of income, consumption and wealth in Spain", *Documents* No. 1806, Bank of Spain.

Castellón, P. and Velásquez, J. D. (2013), "Characterization and detection of taxpayers with false invoices using data mining techniques", *Expert Systems with Applications*, 40: 1427-1436.

Chakrabarti, B., Chakraborti, A., Chakrabarty, S. and Chatterjee, A. (2013), *Econophysics of Income and Wealth Distribution*, U. K, Cambridge University Press.

Chatterjee (2005), *Econophysics of Wealth Distributions*, (Ed.: A. Chatterjee, S. *et al.*), Italy, Milan, Springer-Verlag.

Chunaev, P. (2020), "Community detection in node-attributed social networks: A survey", *Computer Science Review*, 37.

CIAT (2020), “Auditing with technological support, methods techniques and the experience of the Tax Administration of Spain”, in *ICT as a Strategic Tool to Leapfrog the Efficiency of Tax Administrations*, CIAT, Bill & Melinda Gates Foundation, Panamá.

Congressional Budget Office (2018), “The Distribution of Household Income, 2015”, November 8, https://www.cbo.gov/publication/54646.

Consejo General de Economistas (2017), “Reflexiones sobre el fraude fiscal y el problema de las estimaciones: 20 propuestas para reducirlo”, *Estudios del Consejo General de Economistas*, June, available online at: www.reaf-regaf.economistas.es.

Credit Suisse (2015), *Global Wealth Report*. Electronic resource available: https.//publications.credit-suisse.com/tasks/render/file/?fileID=F2425415-DCA7-80B8-EAD89AF9341D47E.

Crivelli, E., Mooij, A. R. and Keen M. (2015), “Base Erosion, Profit Shifting and Developing Countries”, *IMF Working Paper* 15/118, Fondo Monetario Internacional, Washington, DC.

Durán-Cabré, J. M., Esteller Moré, A., Mas-Montserrat, M. and Salvadori, L. (2019), “The tax gap as a public management instrument: application to wealth”, *Applied Economic Analysis*, 27(81): 207-225.

El-Moussaqui, M., Agouti, T., Tikniouine, A. and El Adnani, M. (2019), “A comprehensive literature review on community detection: Approaches and applications”, *Procedia Computer Science*, 151: 295-302.

Eslam, E., Pourdarab, S. and Nadali, A. (2011), “Credit Risk Assessment of Bank Customers using DEMTAEL and Fuzzy Expert Systems”, *International Conference on Economics and Finance Research IPEDR*, vol. 4, IASCSIT Press;, Singapore.

EUROSOCIAL (2016), “Herramientas de Análisis de Información de la AEAT” https://es.slideshare.net/EUROsocial-II/herramientas-de-anlisis-de-la-informacin-de-la-aeat-zujar-agencia-estatal-de-administracion-tributaria-aeat-espaa.

Europa Press (2017), “Hacienda desmantela una trama de IVA que ha defraudado más de 25 millones de euros”. Available at http://www.hoy.es/nacional/201706/01/hacienda-desmantela-trama-20170601140141-ntrc-rc.html#ns_campaign=rrsshoy&ns_mchannel=hoy-es&ns_source=fb&ns_linkname=nacional_14.

FATF (2006), “APG Trade Based Money Laundering Report”. Downloadable at http://www.fatf-gafi.org/publications/methodsandtrends/?hf=10&b=0&s=desc(fatf_releasedate).

FISCALIS (2018), *The concept of Tax Gaps Report III: MTIC Fraud Gap estimation methodologies*, Directorate-General for Taxation and Customs Union. Available at: https://ec.europa.eu/taxation_customs/sites/taxation/files/tax_gaps_report_mtic_fraud_gap_estimation_methodologies.pdf.

Furth, B., (2010), *Handbook of Social Network Technologies and Applications*, Springer.

Glancy, F. H. and Yadav, S. B. (2011), “A computational model for financial reporting fraud detection”, *Decision Support Systems*, 50(3): 595-601.

Goldberg. A. V. and Tarjan, R. E. (1988), “A new approach to the maximum flow problem”, *Journal of the ACM*, 35(4): 921. DOI: 10.1145/48014.61051.

González, I. and Mateos, A. (2018a), “Social Network Analysis tools in the Fight Against Fiscal Fraud and Money Laundering”, *Proceedings of the 15TH International Conference on Modelling Decisions for Artificial Intelligence (MDAI 2018).*

González, I. and Mateos, A. (2018b), “K-graph and highly community detection in graphs”, *Proceedings of the 15TH International Conference on Modelling Decisions for Artificial Intelligence (MDAI 2018).*

González, I. and Mateos, A. (2018c), “The distribution of wealth. Deconstructed Pareto, Reconstructed Gibrat”, *Journal of Applied Economics*, pp. 22-40.

González, I. and Mateos A. (2020), “Bayesian Dialysis of the Evidence in Fraud Detection”, *Proceedings of the ConferenceDECON 2020*, Springer.

Hagen, L., Keller, T. E, Yerden, X. and Luna-Reyes L. P. (2019), “Open data visualizations and analytics as tools for policy-making”, *Government Information Quarterly*, 36(4).

Hay, I. and Muller, S. (2012), “That tiny, stratospheric apex that owns most of the world”, *Exploring geographies of the super-rich. Geographical Research*, 50(1): 75-88.

Hernández, E. (2014), *El fin de la clase media*, Madrid: clave intellectual.

Hospido, L. (2010), “La encuesta financiera de las familias (EFF): La experiencia española y el proyecto europeo”. Día Mundial de la Estadística. Available at https://www.sgapeio.es/descargas/dia Mundial201010/LauraHospido.pdf.

Kiraly, Z. and Kovacs, P. (2012). “Efficient implementation of minimum-cost flow algorithms”, *Acta Universitatis Sapientiae*, *Informatica*, 4(1): 67118.

Kruppa, J., Schwarz, A., Arminger, G. and Ziegler, A. (2013), “Consumer credit risk: Individual probability estimates using machine learning”, *Expert Systems with Applications*, 40(3): 5, 125-51, 31.

Lismont, J., Cardinaels, E., Bryuynseels, L., Groote, S., Baesens, B., Llemahieu, W. and Vanthienen, J. (2018), “Predicting tax avoidance by means of social network analytics”, *Decision Support Systems*, 108, 13, Universidad Complutense de Madrid.

Luque, V. (2015), “A propósito de Piketty: evolución de la desigualdad en España”, *Papeles de Europa*, 28(1): 86-115.

Mantegna, R. N. and Stanley, H. E. (2001), *An introduction to Econophysics: Correlations and Complexity Finance*, Cambridge University Press, Cambridge.

Mas, M. (2020), *Essays on Wealth Taxation, Avoidance and Evasion among the Rich*, PhD Thesis, Universitat de Barcelona, Facultat d’ Economia i Empresa. Available at https://www.tdx.cat/handle/10803/668654#page=1.

Matos, T., Macedo, J. A., Lettich, F., Monteiro, J. M., Renso, Ch., Perego, R. and Nardini, F. M. (2020), “Leveraging feature selection to detect potential tax fraudsters”, *Expert Systems with Applications*, 145.

Montroll, E. W. and Schelesinger, M. F. (1982), “On 1/f noise and other distributions with long tails”, *Proceedings of the National Academy of Sciences USA*, 79: 3380-3383.

Murphy, R. (2012), “Closing the European Tax Gap. A report for Group of the Progressive Alliance of Socialists & Democrats in the European Parliament”, *Tax Research LLP, Norfolk*, (www.taxresearch.org.uk/blog), Reg. number OC316294.

Murphy, R. and Petersen, H. (2018), *Minding the tax gap at the heart of macroeconomic policy*, London: HYPERLINK “http://Coffers.eu” Coffers.eu *Working paper*.

Murthy, S. K. (1998), “Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey”, *Data Mining and Knowledge Discovery*, 2: 345-389.

Ngai, E., Hu, Y., Wong, Y., Chen,Y. and Sun, X. (2011), “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature”, *Decis. Support Syst*. 50(3): 559-569.

OECD (2002), *Measuring the Non-Observed Economy. A Handbook*, OECD, Paris.

OECD (2016), *Pensions Markets In Focus*, OECD. Available at: http://www.oecd.org/daf/fin/private-pensions/Pension-Markets-in-Focus-2016.pdf.

OXFAM (2018), “Fact or Fiction? Economic Recovery, in the Hands of a Minority”. Available online at: https://www.diarioabierto.es/wp-content/uploads/2018/01/Report_DavosEnglish_Spanish_NonDenominational.pdf.

Papadimitriou, Ch. H. and Steiglitz, K. (1998), “The Max Flow, Min Cut Theorem. Combinatorial Optimization: Algorithms and Complexity”, *Dover*: 120-128. ISBN 0-486-40258-4.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A. and Spyridonos, P. (2012), “Community detection in social media”, *Data Mining Knowledge Discov*, 24(3): 515-554.

Persson, T. and Tabellini, G. (1994), “Is inequality harmful for growth? Theory and evidence”, *American Economic Review*, 48: 600-621.

Phua, C., Lee, V., Smith, K. and Gayler, R. (2010), “A comprehensive survey of data mining-based fraud detection research”, arXiv:1009.6119.

Piketty, T. (2013), *Capital in the Twenty-First Century*, Harvard University Press.

Piketty, T., Saez, E. and Zucman, G. (2018), “Distributional National Accounts: Methods and Estimates for the United States”, *Quarterly Journal of Economics*, 133.

Pourhabibi, T., Ong, K. L., Kam, B. H. and Ling, Y. (2020), “Fraud detection: A systematic literature review of graph-based anomaly detection approaches”, *Decision Support Systems*, 133.

Pow, C. (2011), “Living it up: Super-rich enclave and transnational elite urbanism in Singapore” *Geoforu*: 382-393.

PwC (2018), “The Data Intelligent Tax Administration Meeting the challenges of Big Tax Data and Analytics”, https://www.pwc.nl/nl/assets/documents/the-data-intelligent-tax-administration-whitepaper.pdf.

Rukanova, B., Tan, Yao-Hua, Slegt, M., Molenhuis, M., Ben van Rijnsoever and Migeotte, J. (2020), “Identifying the value of data analytics in the context of government supervision: Insights from the customs domain”, *Government Information Quarterly*, 101496, ISSN 0740-624X.

Saez, E. (2018), “Striking it Richer: The Evolution of Top Incomes in the United States,” University of California, https://eml.berkeley.edu/~saez/saez-UStopincomes-2017.pdf.

Saez, E. and Zucman, G. (2016), “Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data,” *Quarterly Journal of Economics*, 131(2).

Santarelli, E. and Thurick, R. (2006), “Gibrat’s Law: An overview of the Empiric Literature”. DOI: 10.1007/0-387-32314-7_3.

Schneider, F. and Enste, D. (2000), “Shadow economies: Size, causes and consequences”, *Journal of Economic Literature*, XXXVIII: 77-114.

Schneider, F., Buehn, A. and Montenegro, C. (2010), “New Estimates for the Shadow Economies all over the World”, *International Economic Journal*, 24(4): 443-461.

Serraler, M. (2018), “The Treasury will investigate with ‘big data’ the assets of more than 10 million euros”, *Expansion*, December 26th.

Titan, Abo Akademy University (2012), *Online Tähtinen*, 40-53, pp. 65-67.

Vaquero, A., Lago, S. and Fernández, X. (2016), “Economía Sumergida y Fraude Fiscal en España: Un Análisis de la Evidencia Empírica”, (https://researchgate.net/publication/296332354).

Van Vlasselaer, V., Eliassi-Rad, T., Akoglue, L., Snoeck, M. and Baesens, B. (2017), “GOTCHA! Network-based Fraud Detection for Social Security Fraud” *2 Article submitted to Management Science*, manuscript no. MS-14- 0232. https://pdfs.semanticscholar.org/1151/7fe04965c01b34b679985bc626608224a7fd.pdf.

Vanhoeyveld, J., Martens, D. and Peeters. B. (2019), “Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue”, *Pattern Analysis and Applications*, London-New York, Springer, 2019: 1433-7541.

Vasudevan, M., Balakrishnan, H. and Deo, N. (2009), “Community discovery algorithms: an overview”, *Congressus Numerantium*, 196: 127-142.

Vasudevan, M., Balakrishnan, H. and Deo, N. (2009), “Community discovery algorithms: an overview”, in Allajh, R. and Rokne, J. (Eds), *Encyclopedia of Social Network Analysis and Data Mining* in J., Sc. Ed.

Vasudevan, M. and Deo, N. (2018), “Detecting and Identifying Communities in Dynamic and and Complex Networks: Definition and Survey”, *Encyclopedia of Social Network Analysis and Mining 2018*.

Vydra, S. and Klievink, B. (2019), “Techno-optimism and policy-pessimism in the public sector big data debate”, *Government Information Quarterly*, 36(4).

Wasserman, S. and Faust, K. (1994), *Social Network Analysis: methods and applications*, Cambridge University Press, Cambridge/New York.

West, J. and Bhattacharya, M. (2016), “Intelligent financial fraud detection: A comprehensive review, *Comput. ‘I&’ Secur. 57 (Supplement C)*: 47-66.

Woods, A. J. (2017), “Applications of Flow Network Models in Finance”, *Electronic Theses and Dissertations*, 1645. https://digitalcommons.georgiasouthern.edu/etd/1645.

Xu, J. and Chen, H. (2005) “Criminal network analysis and visualization”, *Commun ACM*, 48(6):101-107.

Yakovenko, V. M., Barkley, and Rosser, J. (2009), “Colloquium: statistical mechanics of money wealth and income”, *Review of Modern Physics*, 81: 703-1725.

Yedidia, J. S., Freeman, W. T. and Weiss, Y. (2003), “Understanding belief propagation and its generalizations”, in: *Exploring artificial intelligence in the new millennium*, vol. 8, Morgan Kaufmann, Amsterdam: 236-239.

Zachary, W. W. (1977), “An information flow model for conflict and fission in small groups”, *Journal of Anthropological Research*, 33: 452-473.

Resumen

La Agencia Estatal de Administración Tributaria de España es un usuario experimentado de las técnicas de *big data* y ha comenzado a implantar herramientas de análisis de redes sociales (SNA). El uso de conceptos y herramientas SNA ha dado lugar a un salto cualitativo en áreas tan diversas como el control y la recaudación tributarios, el control de las grandes fortunas y el blanqueo de capitales. Este artículo presenta un panorama completo de las diferentes líneas de investigación, estrategias y resultados de nueve proyectos durante los últimos cinco años, incluidas las lecciones aprendidas.

Presentamos las mejores prácticas en el descubrimiento de patrones, las herramientas desarrolladas para el control de grandes fortunas y la estrategia desarrollada para crear un puente entre el conocimiento experto y las tecnologías SNA. Destacamos los resultados obtenidos en la investigación de entidades interpuestas utilizadas para ocultar las rentas y en la detección de estructuras corporativas complejas y empresas opacas.

Palabras clave: cumplimiento, patrimonio neto, SNA, redes sociales, Pregel, fraude.

Clasificación JEL: D85, H26, D31, E21.