Identifying Relevant Patterns in a Large Graph of Open Data: A Semantic Exploration of the Panama Papers

Antoine Vion
Aix-Marseille University, CNRS, LEST
Aix-en-Provence, France

Abstract

The following essay examines the Panama Papers from a sociological perspective, using a method based on analytic induction through graph mining. By way of an introduction, it provides a general overview of the chosen approach. A series of case studies concerning tax evasion practices as detailed in the Panama Papers help to elucidate the kind of query processing that might accompany the practical implementation of the proposed method, before the final section gives a brief summary and touches upon a methodological issue that has yet to be resolved.

1 Introduction

Identifying relevant patterns in a large graph of open data is a complex, exploratory process – a statement that finds ample support in the following essay, which examines the Panama Papers from a sociological perspective, employing a method based on analytic induction through graph mining. From a user's perspective, data visualization is a crucial step on the path to data-driven discoveries (Riche, 2015). In the field of visual analytics, graphic

In: Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.): Graph Technologies in the Humanities - Proceedings 2020, published at http://ceur-ws.org
This long paper is based on research presented at "Graph Technologies in the Humanities 2019" (January 18-19, Academy of Sciences and Literature | Mainz, Germany).

visual interfaces tease knowledge out of the data and reveal ways in which this data can be tested, refined, and shared (Pike et al., 2009).

In an effort to better understand how the process of analytic induction through graph mining might function in practice, this essay explores open data knowledge graphs (KGs) with the help of DERIVO's SemSpect tool (Liebig et al., 2017).[1] Running on a Neo4J graph database, SemSpect allows researchers to explore KGs in an intuitive manner. Moreover, the tool's usage of the GraphScale system (Liebig et al., 2017) permits the addition of an abstraction layer to the KG, which enables fast reasoning and high-performance querying (Glimm et al., 2014). With SemSpect, scholars can embark on a data-driven exploration of the KG in question with a mere click of their mouse – no blind queries are required. In contrast to other systems, SemSpect's aggregated representation allows users to investigate highly complex KGs, making it possible to understand their structure without having to engage with the details of the queries applied. At the same time, the tool empowers researchers to group parts of the KG into clusters and categories, and to re-use them in new queries, which in turn allows the model of the KG to be progressively refined.

In the following sections of this paper, I will provide a general overview of analytic induction (Section 2); discuss the kind of query processing that was used to examine a series of case studies on tax evasion practices, as revealed in the Panama Papers (Section 3); present a selection of data-driven discoveries, which will illustrate the benefits of this kind of query processing (Section 4); and present an outlook on further research based on analytic induction (Section 5).

2 Applying the Method of Analytic Induction to Tax Evasion Patterns

Analytic induction (Znaniecki, 1934) is a well-established research strategy in the social sciences. A researcher begins by studying a small number of cases of a particular phenomenon with the intention of finding a set of common denominators. The information gathered is used to draw up a hypothesis, which is then tested on additional cases (Robinson, 1951). If any of the new cases do not verify the hypothesis, either the hypothesis is reformulated to match the features of all the cases studied so far, or the original definition of the type of phenomenon to be explained is altered on the grounds that it does not represent a causally homogeneous category (Lewis-Beck et al., 2003). Further cases are investigated until no more irregularities appear.

[1] http://www.semspect.de

The method of analytic induction highlights a number of complex challenges associated with graph mining and semantic analysis (Table 1). First and foremost, it underscores the importance of refining and developing the categories that are initially used to define a particular social phenomenon; computationally speaking, this corresponds to ontology refinement and iterative semantic processing. Even more significant, perhaps, is the fact that the primary goal of analytic induction is to identify cases (Ragin and H., 1992) that can be used for comparative research: when translated to the field of query processing in web data, such an approach requires a whole new set of techniques, case-based reasoning search for similarities being a case in point (Mottin et al., 2019).

Analytical induction		Graph mining
Case building Building a singular complexion of properties Looking for interesting relations	=>	**Pattern in the graph:** Fixed number of objects Required characteristics of each object Required relations between the objects
Added value	=>	Deep understanding of the data while building hypotheses
Frequency	=>	Interesting patterns do not have lots of occurrences but are frequent enough to capture general features

Table 1: Analytic induction and graph mining

3 Accounting Heterotopias: Retrieving Relevant Patterns as a Methodological Challenge

Using the method outlined in the previous section, we examined the open data package commonly referred to as the Panama Papers with a view towards gathering any relevant information that could help to shed light on the tax evasion techniques that were being used.

Typically, the tax optimization practices of large companies are overseen by external treasury management providers and include methods such as over-invoicing; transferring expenses to subsidiaries located in tax havens; 'forgetting' sub-subsidiaries in the accounting consolidation; organizing white sales; and laundering money on a large scale through the purchase of raw materials, and on a small scale through cheques or prepaid card systems.

All of these practices function as "black holes of power" (Lascoumes and Lorrain, 2007) insofar as intermediation is protected by the institutionalization of secrecy in certain jurisdictions. The Lux Leaks scandal, in which Luxembourg's tax rulings were shown to have provided an unfair advantage to over 340 companies worldwide, is a clear example of the high price paid by whistleblowers from major audit firms who dare to reveal their companies' schemes to the general public. In 2016, the European Directive on the Protection of Trade Secrets further enhanced the secrecy of accounting data. As a result, it has become very difficult to provide proof of offshoring through accounting – while the annual reports of major corporations are a matter of public record via their consolidated statements, those of offshore subsidiaries are no longer available.

In the absence of open transaction data, the information provided by the Panama Papers is predominantly topological. Suzanne Roberts (1994) has shown how the geography of offshore financial flows contributes to the construction of fictitious spaces. In order to understand this phenomenon, it is necessary to keep in mind that any accounting entry, whatever its final form, assumes as a basic principle that the value recorded in one book is related to a duplicate value in another book. Normally, this practice of double-entry bookkeeping can serve as a basis for reconstructing transactions and, by extension, the networks of economic agents who have recorded the transactions in question. However, the various methods of fiscal optimization now in widespread use combine to create what could be called an accounting heterotopia; a term inspired by Michel Foucault's concept of spaces that suspend, neutralize, or reverse their relationships to other locations (Foucault, 1984). In effect, tax havens offer a means for constructing an accounting heterotopia by permitting the registration of fictitious duplicates referring to accounts drawn up in places where they escape any jurisdictional control. What the Panama Papers allow us to do is to track down these duplicates.

3.1 Offshoring Tricks

Historically, law firms and corporate service providers such as Mossack Fonseca, who were heavily implicated in the Panama Papers scandal, have used certain strategies to help companies play with the spatiality of jurisdictions in their relations with supervisory institutions. What these strategies have in common is that they are not codified, but rather based on tacit knowledge. We have therefore examined graph data from the Panama Papers with a view towards patterns that are characteristic of such implicit practices.

The online databases in question match the definition of what is commonly called Open Data: "A piece of data is open if anyone is free to use,

reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike." The progressive aggregation of such databases means that they become increasingly linked (Kitchin, 2014) in a process of data massification. The main technical problem associated with such a mass of data is the concomitant proliferation of duplicates, since the data is not subject to semantic cleansing. If, for example, a company name has been entered into the Evasion Professionals database as [Name] Limited and [Name] Ltd, the same company will appear in two separate instances.

The Evasion Professionals database is an aggregate of the Offshore Leaks database. The architecture of the database consists of officers (operators identified as set-up operators), intermediaries (intermediaries working on the files), entities (natural or legal persons on whose behalf the file is processed), and addresses (addresses linked to the three types of predefined entities); companies are classified according to operational criteria (active, inactive, dissolved, relocated, redomiciled, etc.). The content available online does not, however, give access to all of the extracted data: in order to protect themselves from cumbersome and costly legal proceedings, the International Consortium of Investigative Journalists (ICIJ) and the participating hackers have left out information concerning the amounts and dates of transactions.

The absence of dates means that fine temporal processing is not available, whether it be genealogical, archaeological, or sequential. As a result, the attempt to represent the complex social temporalities involved in the various transactions recorded in the data produces a synchronic fiction. In fact, computation without dates makes it almost impossible to carry out a pragmatic analysis of the operations of record capture, circulation, and manipulation, or any form of matching of editing sequences forming the 'careers' of clients (Abbott, 1999), which raises the question of the usefulness of uploading data with such obvious technical and legal limitations.

In the case of the Panama Papers, the ordering of the data corresponds neither to the logic of entirely raw and unstructured data, nor to semantically consistent object classes or ontologies.[2] It does, however, appear semi-structured for the purposes of statistical analysis, whether it be simple descriptive statistics or network statistics, and the ICIJ website provides a graph-based network analysis software that appears to be intended for precisely that purpose: the examples made available to users online show networks of co-affiliation located at the same tax address, and the various affiliations of operators identified as fraudsters.

[2]In computing, ontologies are structured sets of terms and concepts that are used to define the meaning of a given information field, either via the metadata of a namespace or the elements defined by a stabilized knowledge domain.

In combination, the openness of the available data, its sheer volume, and the kind of investigative tools that are provided on the ICIJ's website exert considerable influence on how the mechanisms of tax evasion can be examined.

3.2 The Limitations of ICJI's Collaborative Exploration Tools

Sociologists investigating the phenomenon of tax evasion face a number of methodological challenges. When conducting exploratory research of open data, two types of problems immediately become apparent: the first concerns the analytical limitations of the tools that are made available to the researcher, as in the case of the ones provided on the ICIJ website; the second has to do with the tools used for alternative research strategies.

The ICIJ gives interested users the option of downloading network processing software with which they can generate graphs comparable to those created by information services companies to visualize the affiliations of company directors or other persons of interest. However, the ICIJ's choice to focus on what is commonly referred to as self-centered networks or affiliation networks has some rather unfortunate consequences. Generally speaking, working with self-centered networks involves a simple query based on filters and dictionaries of proper names (celebrities, companies, etc.), which makes it possible to quickly retrieve information concerning the participation of a given person or company. In the case of Mossack Fonseca, the immediate targets of ICIJ's efforts were the British and Icelandic prime ministers, David Cameron and Sigmundur Davíð Gunnlaugsson, both of whom resigned in the wake of the Panama Papers scandal.[3]

By conjuring up visible networks, the ICIJ succeeded in arousing media interest in what are ultimately invisible practices. However, by choosing to focus on the relationship between a particular personality and the intermediaries of an optimization firm like Mossack Fonseca, the ICIJ participated in a logic of shaming the latter's clients, as opposed to conducting an in-depth analysis of the firm's practices. Self-centered or affiliation networks are simply not conducive to exposing the creative bookkeeping techniques of such firms, which often duplicate, conceal, or anonymize company names – in the final analysis, assessing the phenomenon of tax evasion based on dyadic relationships, such as those between officers and intermediaries, or entities and officers and/or intermediaries, is an inadequate response to the challenges at hand.

[3] The news that Cameron held shares in an offshore company together with his father broke in the middle of the Brexit campaign, while Gunnlaugsson's involvement in offshore dealings came to light shortly after he had signed a military agreement with the US.

3.3 Visual Analytics

One way to address the problems outlined above is to draw on software solutions from the domain of visual analytics, which can be defined as a science of analytical reasoning (Ribarsky et al., 2009) supported by interactive visual interfaces that help to overcome problems of data size and complexity (Dill et al., 2012). Visual analytics is a multi-disciplinary field of research that leverages recent findings in areas such as visualization, data mining, data management, data fusion, statistics, and cognitive science (Kielman et al., 2009). As pointed out in the introduction, the ability to visualize data is an essential prerequisite for data-driven discoveries. In the course of our own investigation into the Panama Papers, we used GraphScale and SemSpect software developed by DERIVO – two excellent tools that make intuitive graph exploration possible. The main advantage of GraphScale and SemSpect lies in the fact that they exploit a level of ontological abstraction that is automatically constructed from a semantic graph by means of graph mining in order to enable complex and traceable queries.

4 Three Key Skills: Occulting, Partitioning, Porting

By proceeding in an inductive way from basic data to the detailed analysis of complex files, our systematic exploration of the database enabled us to identify three recurring tax evasion techniques in the Panama Papers.

4.1 Occulting

As discussed in the previous section, the specific nature of the data contained in the Panama Papers demands a suitable approach to processing it, and great care must be taken in regard to the inferences that are being drawn.

After conducting multiple tests, it became clear that the only reliable method for engaging with the data at hand – incomplete as it is – was to reconstruct the chains of tax packages starting with their associated addresses. The registered companies are letterbox companies that have been set up for the very purpose of bypassing legal restrictions, and they typically carry more or less creative names. But if one manages to move past this ruse, it becomes possible to reconstruct chains of letterboxes, to follow circuits of evasion, and to expose the accounting heterotopias that are made possible by the geographical location of the financial vehicle corporations (FVCs) involved in the scheme.

With this goal in mind, the query logic of searching for co-affiliations of companies at specific addresses appears far more promising than the analysis of egocentric networks. Indeed, the mass of data uploaded by the ICIJ is

rendered practically useless if it is not reconfigured under an ontological database model which allows the user to define object classes and inferences according to comparable semantic properties or certain logical properties.

The need for such an ontological database model is easily demonstrated. For example, the ICIJ site provides a graphical representation of the relationships that can be reconstructed based on the identification of the various companies registered to the same hotel suite in the Seychelles (Figure 1).

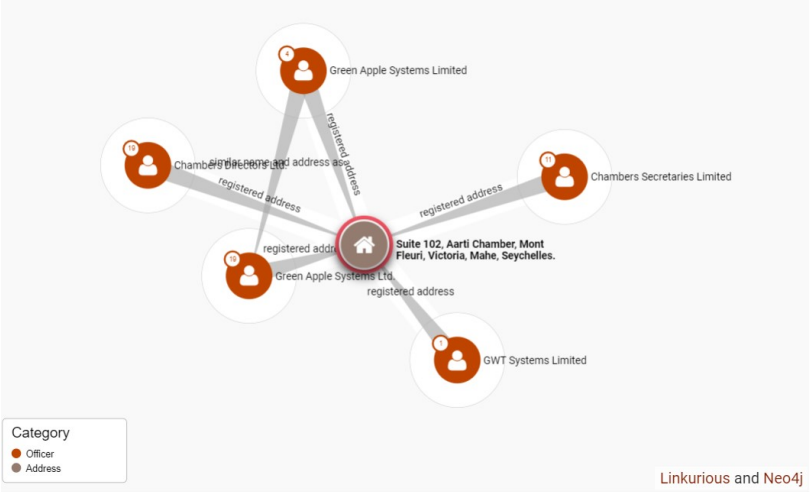

Figure 1: Dyadic ties with suite 102 Aarti Chambers. Source: ICU, `https://offshoreleaks.icij.org/nodes/233584` (January 2017)

Here we see how an unmodeled database can be used to trace the relationships between the hotel's managers and the companies Green Apple System Limited/Ltd and GWT Systems Limited by relying on nothing more than an address (Suite 102, Aarti Chambers, Mont Fleuri, Victoria, Mahé, Seychelles).

As Table 2 shows, the details of this address are subject to considerable variation in the database, with changes falling under three broad categories: punctuation, capitalization, and (deliberate) spelling errors. Without a semantic approximation model, one would have to multiply the queries to search for other possible occurrences of the same suite under different names. However, thanks to the modeling we developed in cooperation with DERIVO, we were able to identify not one but twelve occurrences of the same suite (Table 3). Moreover, we found other suites with multiple occurrences at the same residence, as well as in other locations throughout the Indian Ocean (Seychelles, Mauritius, etc.).

Since the data is not time-stamped, it is difficult to infer whether or not the relationships between all of these companies were simultaneous – it ap-

102; Aarti Chambers; Mont Fleluri; Victoria; Mahe; Seychelles
102; Aarti Chambers; Mont Fleuri; Victoria; Mahe; Republic of Seychelles
102; AARTI CHAMBERS; MONT FLEURI; VICTORIA; MAHE SEYCHELLES
102; Aarti Chambers; Mont Fleluri; Victoria; Male; Seychelles
102; AARTI CHAMBERS; MONT FLEURI; VICTORIA MAHE SEYCHELLES
102; AARTI CHAMBERS; MONT FLEURI; VICTORIA; MAHE; SEYCHELLES
102; Aarti Chamnbers; Mont Fleuri; Victoria; Mahe; Seychelles
102 AARTI CHAMBER; MONT FLEURI; VICTORIA; MAHE; SEYCHELLES
102 Aarti Chamnbers; Mont Fleuri; Victoria; Mahe; Seychelles
102 Aarti Chambers; Mont Gleuri; Victoria; Mahe; Seychelles
102 AARTI CHAMBERS MONT FLEURI VICTORIS; MAHE; SEYCHELLES
Suite 102, Aarti Chamber; Mont Fleuri; Victoria; Mahe; Seychelles.

Table 2: Typographic variations of the Aarti Chambers address

pears possible, and indeed likely, that we are dealing with a series of successive domiciliations. But in any case, what becomes evident here is an important mechanism of tax evasion, namely the use of concealment procedures within protected databases in accordance with Hervé Falciani's (Falciani, 2015) description of the IT strategies employed by HSBC to disperse data within the organization for reasons of secrecy. This insight became possible because our modeling allowed us not only to detect simple dyadic relationships, but to establish sets of companies linked to the same registration address.

Were we to use ICIJ's approach to analyze the same data, we would find that a total of eighty-one addresses remain after the subtraction of four duplicates, all of which appear as separate entities. What becomes obvious here is that while ICIJ managed to extract a large amount of information concerning the mechanisms of tax evasion (and, as we have seen, generated impressively large numbers in the process), the knowledge it produced is actually much more superficial than it appears. In other words, it is quite clear that the cognitive strategy employed to convert information extracted from the data into knowledge (Boisot and Canals, 2004) was not based on a firm grasp of the nature of that data, even though this understanding is absolutely essential. So-called 'open' data is by no means immune to such problems: if anything, the very term itself may create a dangerous illusion of accessibility, when in fact the mere availability of information is meaningless in the absence of a viable analytical strategy.

4.2 Partitioning

As the previous example has shown, the uncovering of tax avoidance structures entails identifying the location nodes of FVCs and using these nodes to reconstruct corporate networks based on criteria such as company owners, intermediaries, and assembly operators. Instead of relying on the soft-

Lettering of the address	No. of companies	No. of doublets
102; AARTI CHAMBERS MONT FLEURI; VICTORIA MAHE SEYCHELLES	67	3
Suite 102, Aarti Chamber, Mont Fleuri, Victoria, Mahe, Seychelles.	4	1
102; Aarti Chambers; Mont Fleluri; Victoria; Mahe; Seychelles	4	
102 Aarti Chambers; Mont Fleur; Victoria; Mahe; Seychelles	1	
102; AARTI CHMAMBES MONT FLEURI; VICTORIA; MAHE; SEYCHELLES	1	
102 Aarti Chambers; Mont Gleuri; Victoria; Mahe; Seychelles	1	
102 AARTI CHAMBERS MONT FLEURI VICTORIS MAHE SEYCHELLES	1	
102; Aarti Chambers; Mont Fleuri; Victoria; Male; Seychelles	1	
102; AARTI CHAMBERS; MONT FLEURI; VICTORIA; MAHE SEYCHELLES	1	
102; Aarti Chambers; Mont Fleuri; Victoria; Mahe; Republic of Seychelles	1	
102 AARTI CHAMBER; MONT FLEURI; VICTORIA; MAHE; SEYCHELLES	1	
102; Aarti Chamnbers; Mont Fleuri; Victoria; Mahe; Seychelles	2	

Table 3: Number of companies linked with the diverse occurences of [Suite 102, Aarti Chambers in MOnt Fleuri] (N=85)

ware made available online by ICIJ with all of its limitations, I would argue that it is much more expedient to investigate evasion circuits on the basis of relational chains.

The analysis of networks in terms of relational chains, first proposed by Stanley Milgram (1967), was picked up by Mark Granovetter (2018) in his study on employment in the US, which examines the chains of contact that need to be mobilized in order to gain access to information on career opportunities and to request assistance with the application process.

Let us consider an example of such a chain: the employer knows A, who knows B, who knows C, who knows the potential recruit; B passes the latter's name along to A together with his assessment of C's qualification to give a recommendation, and the likelihood of his giving a frank opinion; A in turn passes this information on to the employer, along with his own assessment of B. Of course, chains of this kind become more and more impractical the longer they are – even the three-step recommendation chain from our example has a slightly implausible ring to it. But Milgram's work holds out hope that short chains could in fact be the norm: his investigation showed that randomly chosen pairs of Americans (one from Massachusetts and the other from Nebraska) need a mere six to eight links on average in order to connect them with one another (Granovetter, 2018, 137).

In the case of the Panama Papers, the relational chains are not always chains of operators in the strictest sense of the word, since it is not uncommon for the final recipient to be the initial customer at the end of a perfectly circular array of intermediaries. This poses several problems if we are to conduct a true network analysis.

The first obstacle is gaining access to the governance structures of companies residing in tax havens, as managers often enter account lines in shell companies that are nothing more than mailboxes hosted by intermediaries. This raises the question of which sources can be used to trace these financial flows – after all, the concerned parties have no incentive whatsoever to cooperate. And what of the various other challenges associated with modeling these complex and sometimes flat-out paradoxical relationships? Can one be connected to a fictitious double of oneself via an intermediary one has never met? And if so, to what extent is that double actually fictitious? Are there no formal obligations to be fulfilled (signature, face-to-face procedures, etc.), no practical precautions to be taken to facilitate eventual liquidation?

For the purposes of our investigation, the lack of complete accounting data means that carefully planned qualitative explorations are necessary to understand the structure of the relational chains being examined. One particularly instructive case involves a number of companies linked to Portcullis,

an assembly operator that is registered at 113 different addresses in 9 countries (Table 4).

These addresses are in turn linked to 1,735 officers, i.e. professional assembly operators (Table 6). The majority of these officers are located in tax havens, but some are based in European countries that have not traditionally fulfilled such a role, including Italy. The two Italian officers are registered at 18 different addresses in their home country and are linked to an additional 22 officers, one of whom is a legal entity called Sharecorp. When we proceed to examine the number of beneficiaries of this company, we discover no less than 1,610 further entities, distributed geographically over several tax havens (Table 5).

We can thus regress to infinity. What we have uncovered here, then, is an intricate partitioning system that organizes transnational address sharing in the form of concentric circles; an arrangement that is difficult if not impossible to penetrate for all those who do not have access to the database of participating law firms, most of which are based in the British Virgin Islands. Graph-based visualization, however, can help to facilitate the search for an original pattern. Beginning with what appears to be a kind of statistical anomaly, it becomes possible to identify forms of organization that would go unnoticed by classical detection algorithms – spotting their relevance is not so much a matter of statistical knowledge of graph structures, but a question of familiarity with the object of inquiry.

4.3 Porting

Our analysis revealed two characteristic traits of the assemblies in question: high-intensity backrest turnover and pronounced diversity of the carrying vehicles. The use of shipping companies and shipping addresses seems to be especially favored (Table 6).

The available data is testament to the massive scale on which asset management companies have been incorporated (72,720 in the Panama Papers and a further 4,260 in the Offshore Leaks database). Figure 2 shows the distribution of these companies by country.

Of course, a chart based on the organizing principle of companies per country has significant limitations, as it does not reflect the volume of assets carried. Nevertheless, it does show that the extensive use of shipping companies is favored by Western companies and their traditional tax havens, whereas it is much less prevalent in Asia (with the notable exception of Hong Kong, which has traditionally served as the financial interface between major Western banks and Asian companies).

The 76,980 companies in question own a total of 35,345 entities, of which

Country	Number of adresses
British Virgin Islands	55
China	1
Cook Islands	14
Hong Kong	10
Malaysia	1
Samoa	20
Seychelles	1
Singapore	10
Taiwan	1

Table 4: Registration of the Porticullis Co.

Bahamas	4
British Virgin Islands	1537
Cayman Islands	31
Germany	1
Hong Kong	7
Indonesia	7
Malaysia	3
Mauritius	2
Russia	2
Samoa	5
Seychelles	7
Singapore	150
Taiwan	1
Thailand	1
U.S. Virgin Islands	

Table 5: Geographical distribution of entities linked with the Sharecorp Co.

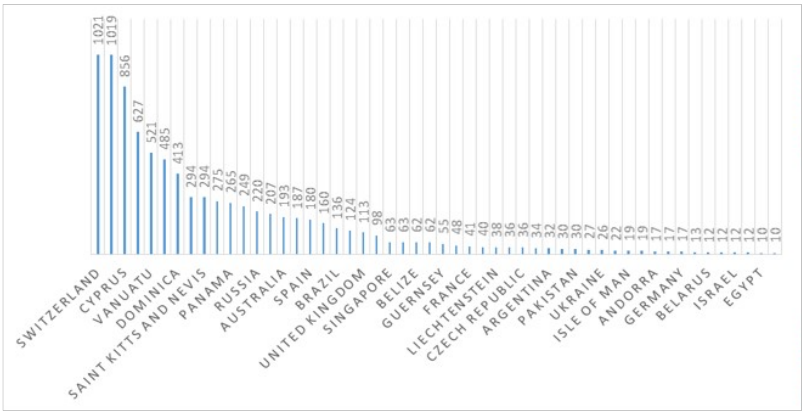

Figure 2: Geographical distribution of shipping companies.(Countries with less than 10 companies have been removed from the histogram.)

31,549 were set up by Mossack Fonseca, and the rest by Portcullis. The status profile of these entities shown in Table 7 illustrates the intensity of turnover.

With companies registered as an international business company (IBC) in a tax haven, the situation is straightforward: they are considered to be classic instruments of tax evasion. Other vehicles, such as the porting of addresses for consulting firms, blur the line between tax evasion and money laundering. Indeed, the scheme of using a multitude of 'shipping addresses' is very much in line with the findings of Fabian Teichmann (2017), who conducted an interview survey with agents mentioned in the Panama Papers. In his study on money laundering methods, Teichmann demonstrated that, in addition to more 'traditional' methods (buying gold, jewellery, rough diamonds, antiques, and paintings; arranging cash deals or deals with foreign exchange offices; etc.), organizational arrangements such as real estate projects, over-invoicing by consulting firms, certain types of mergers and acquisitions, and banking transactions (particularly in Dubai) play an important role.

To cite just one example, a relational chain based on the addresses connected to a specific porting address alerted us to complex financial dealings between Russia, Dubai, and the United Kingdom. After researching identities and companies, we found that one of the individuals (subsequently referred to as operator X) involved in the scheme was already being prosecuted in the United Kingdom and Moldova for money laundering on a scale of several billion euros.

Country	Number of offices
Australia	4
Bahrain	1
British Virgin Islands	1313
Canada	1
China	4
Cook Islands	47
Djibouti	1
France	1
Germany	1
Greece	1
Guernsey	1
Hong Kong	5
India	7
Indonesia	28
Italy	2
Japan	1
Kazakhstan	1
Kenya	2
Malaysia	9
Mauritius	1
Myanmar	1
Philippines	1
Russia	3
Samoa	366
Seychelles	1
Singapore	25
South Africa	2
South Korea	1
Spain	1
Sri Lanka	2
Switzerland	1
Taiwan	10
Thailand	4
Turkey	2
United Arab Emirates	2
United Kingdom	2
United States 1	

Table 6: Addresses of Porticullis offices

Status	**Number**
Defaulted	12,134
Active Entity	9,346
Dissolved	5,271
Changed agent	3,041
Struck / Defunct / Deregistered	1,734
Inactivated	1,321
Resignated as agent	1,039
Dead	348
Relocated in new jurisdiction	323
In transition	170
Discontinued	165
Transferred Out	168
Shelf company	94
Bad debt account	52
Liquidated	74
Not to be renewed / In deregistration	31
Shelf company not possible to sell	12
In liquidation	10
Unregistered	5
Change in administration pending	5
Redmoicited	2
Trash company	1
Profile	**Number**
Standard International Company	2,889
Standard Company under IBC* Act	840
Business Company Limited by Shares	24
Nominee Only Entity	5
Bahamas IBC*	3
Turks	3
Geographical location of the entities carried	**Number**
Entity in Switzerland	14,118
Entity in Luxembourg	4,532
Entity in British Virgin Islands	3,261
Entity in Hong Kong	1,337
Entity in USA	39
Entity in Germany	39
Entity in France	29

Table 7: Status of the 35,345 entities held by the holding companies (out of a Semspect simple query)

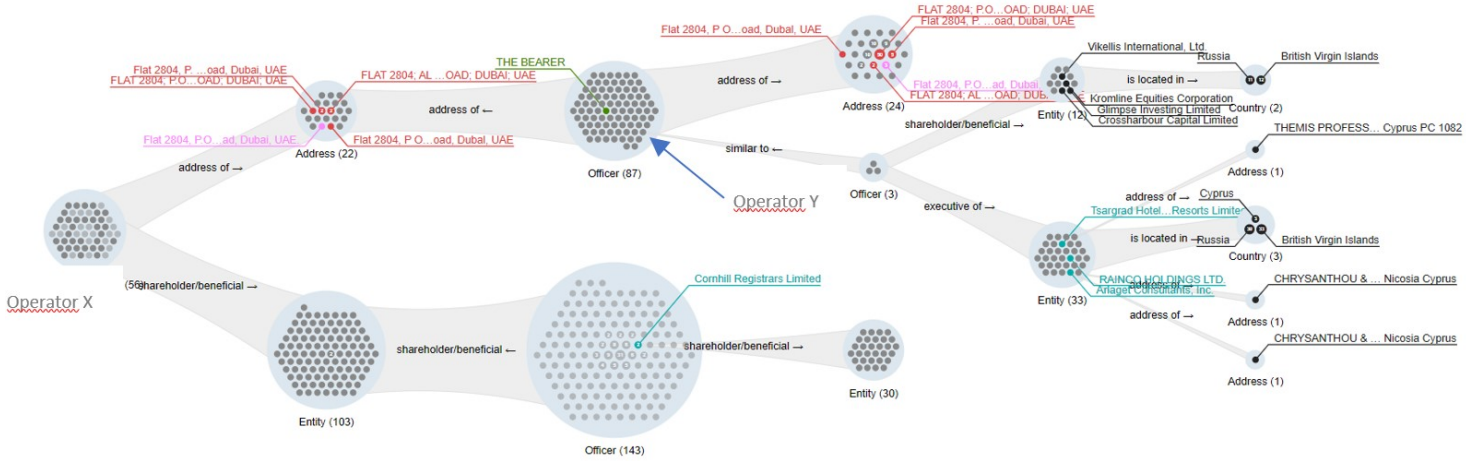

Figure 3: Relational chain based on the addresses connected to a specific porting address alerted us to complex financial dealings between Russia, Dubai, and the United Kingdom.

Yet the connected addresses revealed arrangements that went well beyond the indictments brought in this case: in the Mossack database, operator X appears as the manager of 103 companies located in ten different countries: 82 in the British Virgin Islands, 70 in Russia, 9 in Cyprus, 7 in Samoa, 6 in Ukraine, 4 in the United Kingdom, 3 in the US, 2 in the Seychelles, 1 in Hong Kong, and 1 in the United Arab Emirates.

Operator X as an individual is domiciled in a series of addresses in Dubai that are artificially distinguished according to the method described above. Some of these addresses are shared by the anonymous directors of Cyprus-domiciled carrier companies. After a query of the operators sharing the address of the apartment in Dubai, it became apparent that one of them, a British national (Y), is one of several British shareholders in six companies domiciled in Russia and Cyprus, neither of which is owned or managed by X. However, when we explored X's own activities in Cyprus, we found that X manages other companies in the country. In turn, these companies are all mediated by a Russian consultancy firm registered at an address in Cyprus – and one of the directors of this firm is Y, the person who shares X's address in Dubai (Figure 3).

The chain thus operates as follows: the British partners in Russia, linked to the Dubai address via their business partner Y, manage the business relations of Russian and foreign companies established in Russia with other companies in which Y is involved via a consultancy company registered in Cyprus. The management of Cypriot companies from Dubai in turn enables the establishment of holding companies that carry out transactions via Mossack Fonseca in the British Virgin Islands.

What makes this chain so interesting is that it is connected, via its main operator, to a large-scale Russian money laundering network. In this case, 19 British companies have already been prosecuted for money laundering to the tune of 20 billion pounds. The schemes described above may be legal, and the presumption of innocence must be respected, but their discovery illustrates the fragility of the boundaries between legally permissible financial optimization and money laundering – some consulting firms such as Mossack Fonseca have certainly been complicit in large-scale money laundering operations, as the recently leaked FinCEN files show.

5 Summary and Outlook

The questions and findings that have arisen over the course of our project are indicative of a much broader need for tools that enable the investigative exploration of open data in the humanities and social sciences (HSS) – a technological infrastructure we refer to as an Investigation Support System (ISS).

We believe that two key building blocks are needed to establish such an ISS: visual analytics and knowledge graph data mining.

Visual analytics integrates new theoretical approaches and computational tools with innovative interactive techniques and visual representations to enable human-information discourse. In our case, where visual analytics is applied to knowledge graphs or semantic graphs constituted from open data, DERIVO's GraphScale and SemSpect tools have proven to be ideally suited to the task at hand.

When it comes to knowledge graph mining, relational data mining (RDM) is an especially promising approach. Unlike traditional data mining algorithms, which look for patterns in a single table (propositional patterns), RDM algorithms look for patterns among multiple tables (relational patterns). Davide Mottin's groundbreaking work in the field allows researchers to apply the sociological method of analytic induction to substantial bodies of open data. By constructing exemplar patterns, (Mottin et al., 2014), examining their properties (Mottin et al., 2017), and systematizing this kind of exploration (Mottin et al., 2019), scholars are able to expand their queries intuitively, which produces a more informative (full) query that can retrieve more detailed and relevant answers (Lissandrini et al., 2020).

And yet, one significant methodological challenge remains. Graph pattern mining aims at identifying structures that appear frequently in large graphs, under the assumption that frequency signifies importance. Several measures of frequency have been proposed that respect the *a priori* property, which is essential for an efficient search of the patterns. This property states that the number of appearances of a pattern in a graph cannot be larger than the frequency of any of its sub-patterns. In real life, however, there are many graphs with weighted nodes and/or edges, in which case it would be clearly sensible for the importance (score) of a pattern to be determined not only by the number of its appearances, but also by the weights on the nodes/edges of those appearances.

Removing this obstacle will require both a rigorous methodology of frequency scoring, which Mottin et al. are in the process of developing, and suitable methods for case-building, which GraphScale and SemSpect already permit. In light of this, SemSpect's plugging with Neo4j may well prove to have been the decisive step towards the large-scale application of analytic induction to open data.

Acknowledgements

The author would like to thank Thorsten Liebig, co-founder and CEO of DERIVO, for modeling the initial semantic graph of the ICIJ's online data-

base, and Vincent Vialard, senior engineer at DERIVO and co-presenter of this paper at the 2019 Graph Technologies conference, for his collaboration and astute assessment of the research process.

References

Abbott, A. (1999). What Do Cases Do? Some Notes on Activity in Sociological Analysis. In Ragin, C. and H., B., editors, *What Is a Case? Exploring the Foundations of Social Inquiry*, pages 53–82. Cambridge University Press, Cambridge, NY.

Boisot, M. and Canals, A. (2004). Data, Information and Knowledge: Have We Got It Right? *Journal of Evolutionary Economics*, 14(1):43–67, DOI: 10.1007/s00191-003-0181-9.

Dill, J., Earnshaw, R., Kasik, D., Vince, J., et al., editors (2012). *Expanding the Frontiers of Visual Analytics and Visualization*. Springer, London, DOI: 10.1007/978-1-4471-2804-5.

Falciani, H. (2015). *Séisme sur la planète finance: au coeur du scandale HSBC*. La Découverte, Paris.

Foucault, M. (1984). Dits et écrits, Des espaces autres. *Architecture, Mouvement, Continuité*, 5:46–49.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., et al. (2014). HermiT: an OWL 2 reasoner. *Journal of Automated Reasoning*, 53(3):245–269.

Granovetter, M. (2018). *Getting a Job: A Study of Contacts and Careers*. University of Chicago Press, Chicago, IL.

Kielman, J., Thomas, J., and May, R. (2009). Foundations and Frontiers in Visual Analytics. *Information Visualization*, 8(4):239–246, DOI: 10.1057/ivs.2009.25.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage, London.

Lascoumes, P. and Lorrain, D. (2007). Trous noirs du pouvoir. Les intermédiaires de l'action publique. Introduction. *Sociologie du travail*, 49(1):1–9, DOI: 10.4000/sdt.20509.

Lewis-Beck, M., Bryman, A. E., and Liao, T. F. (2003). *The Sage Encyclopedia of Social Science Research Methods*. Sage, London.

Liebig, T., Vialard, V., and Opitz, M. (2017). Connecting the Dots in Million-Nodes Knowledge Graphs With Semspect. In Nikitina, N., Song, D., Fokoue, A., and Haase, P., editors, *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks*, volume 1963 of *CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-1963/.

Lissandrini, M., Mottin, D., Palpanas, T., and Velegrakis, Y. (2020). Graph-Query Suggestions for Knowledge Graph Exploration. In *WWW '20: Proceedings of The Web Conference 2020*, pages 2549–2555, New York, NY. Association for Computing Machinery, DOI: 10.1145/3366423.3380005.

Milgram, S. (1967). The Small World Problem. *Psychology today*, 2(1):60–67.

Mottin, D., Lissandrini, M., Velegrakis, Y., and Palpanas, T. (2014). Exemplar Queries: Give Me an Example of What You Need. *Proceedings of the VLDB Endowment*, 7(5):365–376, DOI: 10.14778/2732269.2732273.

Mottin, D., Lissandrini, M., Velegrakis, Y., and Palpanas, T. (2017). New trends on exploratory methods for data analytics. *Proceedings of the VLDB Endowment*, 10(12):1977–1980.

Mottin, D., Lissandrini, M., Velegrakis, Y., and Palpanas, T. (2019). Exploring the Data Wilderness through Examples. In *Proceedings of the 2019 International Conference on Management of Data*, pages 2031–2035.

Pike, W. A., Stasko, J., Chang, R., and O'connell, T. A. (2009). The Science of Interaction. *Information Visualization*, 8(4):263–274, DOI: 10.1057/ivs.2009.22.

Ragin, C. and H., B., editors (1992). *What Is a Case? Exploring the Foundations of Social Inquiry*. Cambridge University Press, Cambridge, NY.

Ribarsky, W., Fisher, B., and Pottenger, W. M. (2009). Science of Analytical Reasoning. *Information Visualization*, 8(4):254–262, DOI: 10.1057/ivs.2009.28.

Riche, N. H. (2015). Data-Driven Discoveries: Pushing Visualization Research Further. *IEEE Computer Graphics and Applications*, 35(3):42–43, DOI: 10.1109/MCG.2015.54.

Roberts, S. (1994). Fictitious Capital, Fictitious Spaces: The Geography of Offshore Financial Flows. In Corbridge, S., Thrift, N., and Martin, R., editors, *Money, Power and Space*, pages 91–115. Blackwell, Oxford.

Robinson, W. S. (1951). The Logical Structure of Analytic Induction. *American Sociological Review*, 16(6):812–818, DOI: 10.2307/2087508.

Teichmann, F. M. J. (2017). Twelve Methods of Money Laundering. *Journal of Money Laundering Control*, 20(2):130–137, DOI: 10.1108/JMLC-05-2016-0018.

Znaniecki, F. (1934). *The method of sociology*. Rinehart & Company, Inc., New York.