

“© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

A Social Media Tax Data Warehouse to Manage the Underground Economy

Aaron Groulx
Faculty of Business and Information Technology
University of Ontario Institute of Technology
Oshawa, Canada
aaron.groulx@outlook.com

Carolyn McGregor
Faculty of Business and Information Technology
University of Ontario Institute of Technology
Oshawa, Canada
Faculty of Engineering and IT
University of Technology, Sydney
c.mcgregor@ieee.org

Abstract—*Social media can provide a wealth of information valuable to tax administrators in managing the underground economy. This paper proposes a data warehouse design to integrate social media data into tax analytics processes. The warehouse is designed to support modern tax administration strategies that encourage self-regulation and voluntary compliance by shaping public opinion, improving services and developing inclusive tax policies. The warehouse also incorporates the use of social media analytics to support tax evasion detection and enforcement activities such as compliance risk assessment, audits, inspections and investigations.*

Keywords—analytics, data warehousing, social media, social bots, underground economy

I. INTRODUCTION

The underground economy, also known as the black, grey, shadow or cash economy consists of unreported taxable business activity, informal employment and the illicit sale of specially taxed goods (typically excise and consumption taxes on alcohol, tobacco and gasoline). The underground economy is an important social issue as it diverts revenue from public goods and disrupts the fairness of the economic environment, while fostering corruption and other criminal activity.

This paper proposes a social media tax data warehouse architecture and related social media analytics techniques designed to assist tax authorities in managing the underground economy. The architecture and analytics techniques presented are applied to 5 functional areas of tax administration: (1) communications, (2) policy, (3) risk assessment, (4) audit and inspections and (5) investigations in order to reduce current and future underground economic participation at a minimal cost to both government and taxpayers. Reducing participation in the underground economy will increase government revenues through improved tax compliance, dampen the negative social impacts associated with the underground economy and promote legitimate business activity.

The proposed design is a multi-layer data warehouse architecture that facilitates the extraction, processing and storage of integrated social media and tax data, along with end user data analysis capabilities. The design combines elements from existing tax and social media data warehouse architectures and incorporates social media analytics methods

currently used in government and commercial applications. The social media tax data warehouse provides tax authorities with additional data to supplement existing datasets, novel data analysis methods to manage the underground economy and opportunities improve existing tax compliance processes.

This paper adopts strategies developed by the Organization for Economic Development and Co-Operation (OECD) [1] and the International Monetary Fund (IMF) [2] for managing the underground economy. These strategies take a holistic perspective of the tax system, encouraging self-regulation and facilitating voluntary compliance through taxpayer engagement, improving services and developing inclusive tax policies, while detection and enforcement provide a deterrent to participate in the underground economy.

II. BACKGROUND

A. Underground Economy

Individuals participate in the underground economy in order to avoid the payment of taxes and minimize the cost of economic exchange. However there are a variety of economic, sociological, and psychological factors that contribute to the formation of the underground economy. And as each tax system faces challenges specific to its jurisdiction and compliance varies widely, there are a number of perspectives to be considered when developing strategies to manage it. The capabilities required to promote the strategies and associated compliance responses conveyed by these perspectives are included in the design proposed in this paper.

Green (2009), argues that the underground economy is caused by public misperception, confusion and a lack of understanding regarding the social costs of tax evasion [3]. Effective communications are essential to actively engage with citizens and shape public opinion and attitudes. Maintaining a positive public image of the tax system as being fair and just and portraying the payment of taxes as a social virtue and social norm can persuade individuals not to participate in the underground economy. Indeed, public service announcements have been shown to be effective at changing attitudes towards tax compliance [4] given the proper message design [5] [6].

Taxpayers, especially small businesses, may lack the resources to be tax compliant. They may also be wary of tax authorities and hesitant to request assistance. Lu, Huang, and

Lo found that the “perceived usefulness” and “perceived ease of use” of a tax system can affect filing rates and the acceptance of tax innovations and technology [7]. So proactively and visibly reducing barriers and costs through improved policy and program design can help encourage voluntary compliance.

Williams and Horodnic (2016) assert that the propensity to participate in the underground economy can be explained by the degree of asymmetry between society’s codified laws and its informal norms, values and beliefs. Baicu and Hapenciuc (2016) advocates for the taxpayer to play an active role in shaping tax policy; thus minimizing the gap between society’s formal and informal institutions [8]. Supporting this view, Lamberton, De Neve, and Norton (2014) found that eliciting taxpayer preferences can improve compliance rates and decrease participation in tax avoidance schemes [9].

Stalans, Kinsey, and Smith (1991) found that tax beliefs and sanctions are formed through the influence of an individual’s social network, such as interactions with families and co-workers[10]. Similarly, Brown and Drake (2014) found that low-tax paying firms employing tax avoidance strategies tend to cluster based upon the social networks of their board members [11]. Andrei, Comer, and Koehler’s (2014) research models the collective tax compliance behaviour of a population, studying the propagation of tax evasion behavior and the social network structures that influence tax compliance [12]. That work suggests that tax compliance attitudes and behavior spread through social networks via information sharing and imitation and focusing compliance efforts on particular individuals and networks based upon their social network topologies can improve overall tax honesty.

Allingham and Sandmo’s (1972) model regards decisions to underreport income as a rational economic choice balancing the tax savings derived from undeclared income against the probability of being caught and punished [13]. Therefore, the ability of the tax authority to accurately detect incidences of non-compliance and respond appropriately is an important feature of any underground economy management strategy. However, tax administration resources are limited. So tax authorities must maximize the impacts of their efforts within the given resource constraints.

B. Tax Analytics

The techniques commonly employed to analyze tax data are described below. These techniques are incorporated into the proposed architecture to support existing tax administration processes and leverage their use to analyze social media data.

1) Parametric Analysis

Parametric analysis uses explanatory variables or parameters (industry type, audit history, financial ratios compared to other firms in the industry, etc.) to model and predict the likelihood of non-compliance. Common parametric techniques include regression and discriminant analysis [14].

2) Data Mining

Data mining is the automated process of discovering meaningful patterns in large quantities of data. Common techniques include neural networks, clustering and regression

trees[1]. Data mining software can analyze tax data to discover previously unknown correlations that predict compliance risk. These correlations and patterns can be used to directly detect tax evasion and improve parametric models [14]. Provided with training data, typically historic examples of returns and audit results, data mining software can improve its compliance detection accuracy using a process called machine learning.

3) Network Analysis

Network analysis evaluates the compliance risk of groups of taxpayers based upon their relationships with other entities (people, businesses, tax receipts, documents, purchases and assets, etc.). These relationships can be segmented and explored as a visual representation of the network [15].

4) Data Matching

Data matching creates linkages between datasets, comparing and cross referencing information based on identifiers (name, phone number, birthday etc.) in order to find anomalous, inconsistent or otherwise interesting information [16]. It is commonly combined with network analysis to produce additional results.

C. Social Media Analytics

Social Media Analytics refers to the methods of gathering, storing and analyzing data from social media platforms. Social media analytics are used in a wide range of applications. They are employed in law enforcement investigations [17]. They are used in the military to influence public opinion and gather intelligence [18][19]. They have also been incorporated into insurance fraud detection processes [20], where input management tools integrate and organize unstructured social media data into case management systems [21]. Social media analytics are also used extensively in marketing as a customer engagement tool for brand management and product development, as well as capturing demographics [22] and psychographics (personality traits, attitudes, interests, opinions, and beliefs) for customer profiling. Stieglitz, Dang-Xuan, Bruns and Neuberger (2013) have identified three common social media analytic methods: (1) content analysis, (2) social network analysis, and (3) trend analysis [23]. These methods are supported by the architecture proposed in this paper.

1) Content Analysis

Content analysis is the automated process of analyzing unstructured content for meaning. Text analysis is the most common form of content analysis. However, the media being analyzed can include images, video, audio and geographic data. An important subfield of content analysis is sentiment analysis, which determines how individuals or groups feel about a topic and the strength of that sentiment.

2) Social Network Analysis

Social network analysis examines the relationships between entities participating on social media. This is typically analyzed through links created by the platform such as friends or followers, and also through less formal relationships such as comments, views and likes. The structure of these relationships and their topologies can identify the most influential individuals and groups in a network, degrees of separation, hierarchies and geographic proximity.

3) *Trend Analysis*

Trend analysis considers how ideas and concepts change over time and predict which topics will emerge and become popular or important. Trend analysis can identify and follow a topic as it moves through a social network, identifying the structures that impact a topic's emergence and probability of becoming an important or popular topic.

4) *Social Bots*

A set of related technologies are social media robots or social bots. A social bot is automated software that interacts with other entities on social media. Social bots can generate content and employ artificial intelligence to mimic human behavior; engaging in conversations, commenting on posts and answering questions. Several uses for social bots have been identified including content aggregation, responding to customer enquiries, creating an artificial audience, affecting online reputations, influencing opinions [24], orchestrating phishing attacks and distributing malware [25].

5) *Tax Data Warehousing*

Designs for tax data warehouses have been developed in work by Kishore, Anusha, Poornima, Anusha and Tejaswi (2012) [26] and Yang and Wu (2010) [27]. While these data warehouse designs provide basic tax data analysis functionality, they do not incorporate social media data analysis capabilities.

III. SOCIAL MEDIA TAX DATA WAREHOUSE ARCHITECTURE

The system architecture presented in figure 1 enriches existing tax and traditional data warehousing systems and incorporates elements from the University College London's Social-STORM social media analytics platform, as described in Batrinca and Treleaven (2015) [28]. System components that support the collection, storage and analysis of social media data and are not normally found in a traditional tax data warehouse are highlighted in peach. The analytic methods in this paper that can be applied to social media data and used in each of the tax administration functional areas are shown on the far right of the system diagram, highlighted in grey.

A. *Data Source Layer*

The data source layer includes the social media platforms and data brokers, along with the typical internal and external data sources for a tax data warehouse. Social media data is gathered from source systems in the data source layer for integration downstream. Extract, transform and load (ETL) processing occurs between the data source and integration layers. During ETL, data is retrieved, cleansed, formatted to be loaded into integration layer databases.

B. *Integration Layer*

Within the integration layer, data from source systems is merged and integrated into virtual or physical databases. The integration layer includes three major components, (1) the traditional tax data warehouse, (2) the social media data collector and (3) the social media data warehouse. Social media data is channeled through the social media data collector to the social media data warehouse, while the internal and

external tax and financial data are fed into the traditional tax data warehouse.

The social media data collector is comprised of extract transform and load (ETL) processes along with a suite of social media analysis services. Extract processes connect to social media and data brokers through application programming interfaces (APIs). Screen scraping and HTML and XML parsing are used to extract data from source systems when APIs are unavailable. Content analysis services, which leverage text, audio, video and geospatial analysis tools, analyze and identify content of interest. Trend and sentiment analysis services monitor social media based on keywords and other search criteria. Identification services determine online identities and connect them across social media platforms. Social network services identify and measure the relationships between individuals and groups. Social bots are directed through command and control processes.

Data collection and analysis service requests originate from end users and line-of-business tax applications and are typically brokered by the tax warehouse. The tax warehouse provides search criteria to the collector such as names, addresses, phone numbers and keywords. Results are returned to the social media warehouse either in a preprocessed form, to be used in standardized downstream processes or as raw data for manual and ad hoc processing in the analytic layer.

During the integration process, collector results are tagged with metadata to act as a search index, assist with data governance and preserve data lineage. Structured results are loaded into the primary social media data store while unstructured content (documents, images, video, etc.) is stored in an associated enterprise content management system (ECM).

If solely dedicated to serving tax applications, the social media data warehouse can be fully integrated and serve as a sub-component of the tax warehouse. Alternatively it can remain as an independent system, integrating with non-tax applications and serving multiple areas of the organization. Social media data can be integrated with tax data in the integration layer or can remain independent and be integrated downstream in the analytic layer.

C. *Analytic Layer*

Data from the integration layer is manipulated into meaningful information for use by end users and line-of-business applications in the analytical layer. The analytic layer includes software tools and techniques that can assist the end user in their analysis. This includes data manipulation and modeling tools, business intelligence, multidimensional tools, pattern recognition and statistical functions, as well as reporting and data visualisation tools.

Data is exposed to end users and line-of-business applications through a series of data marts. In this case, each tax functional area has a dedicated tax and social media data mart developed for their particular needs. Tax and social media data marts can be exposed separately or as an integrated offering. Social media data analysis is conducted using traditional tax analytics software and specialized social media analytics tools for content, trend and social network analysis.

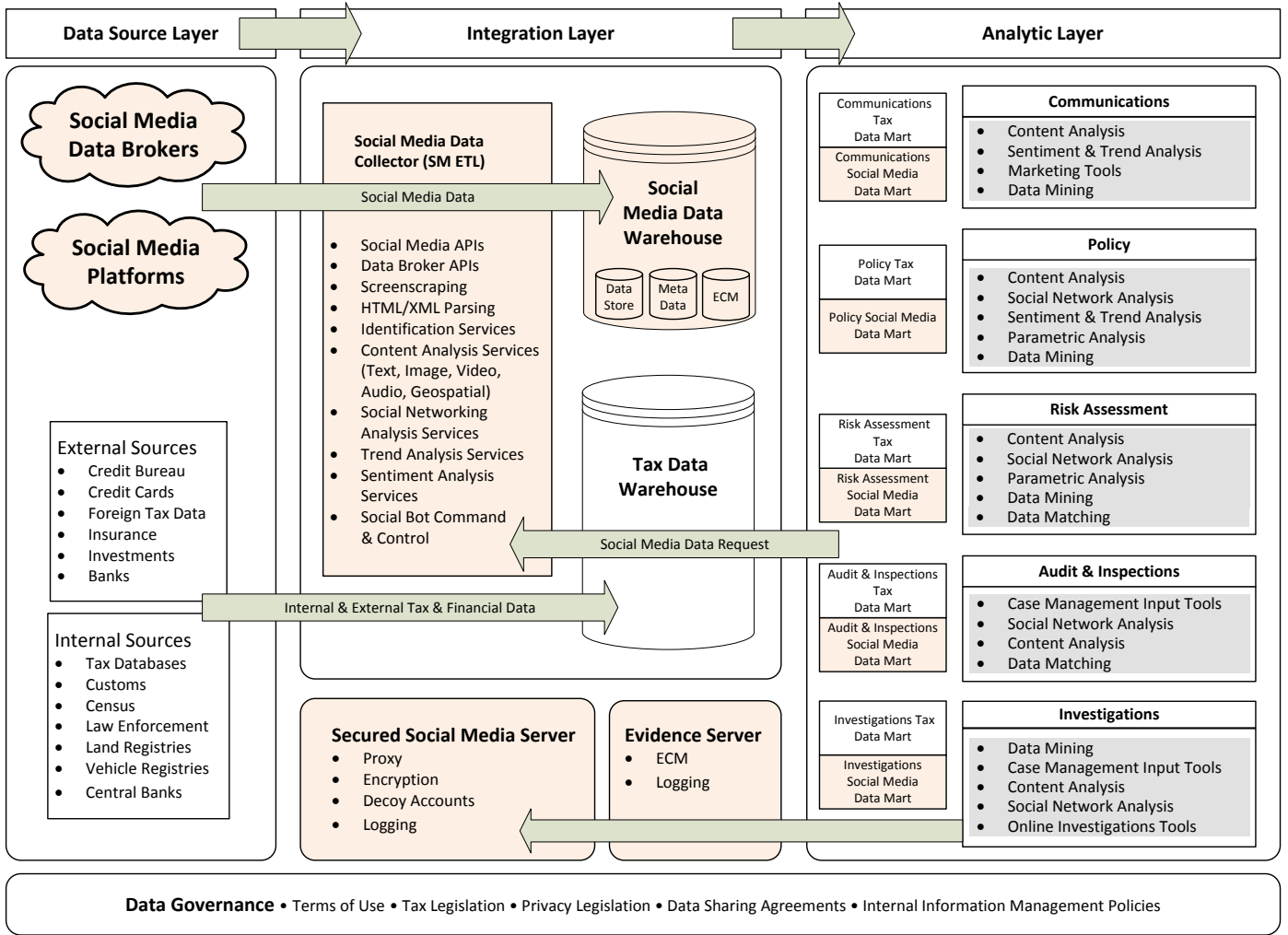


Figure 1: The Social Media Tax Data Warehouse

Case management input software, expedites manual analysis and data entry into case management systems and other line-of-business applications.

A dedicated tax investigations infrastructure is also included in the social media analytics suite of tools, consisting of a secured social media server and evidence server. The secured social media server provides internet proxy, encryption and decoy accounts to protect the identity of tax investigators. The evidence server provides a repository to store content found during the course of online investigations, as well as logging services to establish a chain of custody and avoid evidence tampering.

D. Data Governance

Encompassing the social-media-tax data warehouse are data governance considerations such as data sharing agreements, terms of use, internal information management policies and tax and privacy legislation. Data sharing agreements dictate the information external parties will provide and how it can be used internally. Social media services' terms of use further dictate how information can be collected and

used. Internal information management policies govern data storage, retention and destruction. Tax and privacy laws define the legal authority and limitations to collect, use and disclose data for tax administration.

IV. APPLICATIONS

Opportunities to improve existing tax administration processes by applying the social media tax data warehouse and related social media analytics techniques to each of the functional areas of communications, policy, risk assessment, audit and inspections, and investigations are discussed in the sections below.

A. Communications

1) Persuasive Messaging and Shaping Public Opinion

The social media tax data warehouse can be used to help influence attitudes towards tax compliance and persuade individuals not to participate in the underground economy. The collector scans social media, identifying individuals at high risk of non-compliance and those participating in industries subjected to increased tax scrutiny, based upon characteristics

in their social media profiles and posted content. Social media marketing tools in the analytic layer are employed to ensure advertisements are directed towards this target audience.

Advertisement messaging is automatically tailored to the individual, to maximize its impact, based on characteristics derived from their social media profile and tax data. Social bots, under the command of the collector, engage in conversations and post content on social media with the intention of shaping tax beliefs and behaviour. Bots create an artificial audience for themselves, increasing their perceived influence and bolstering the perception of voluntary compliance as a social norm.

Advertisement and bot conversation data are stored in the warehouse along with the social, economic and psychological characteristics of the taxpayer gathered from social media. The persuasiveness of advertising and bot messaging is improved through machine learning processes that measure the impact that various messages have on compliance rates and the outcomes of audits and inspections, while taking into consideration the characteristics of the taxpayer.

The social media collector's sentiment and trend analysis services also identify discontent online, allowing tax authorities to minimize damage to the image of the tax system through traditional communications channels, while bots dampen negative opinion on social media.

2) Taxpayer Engagement and Improving Services

Voluntary compliance can be promoted by achieving a higher level of customer service and simplifying filing processes, as this improves public confidence in the tax system and the perception that taxpayers are receiving value in return for their social contributions. Bots answer customer inquiries on social media, serving taxpayers more efficiently than manual processes; lowering costs and improving turnaround times. Content, sentiment and trend analysis services analyze social media and bot data stored in the warehouse, identifying common questions, tax compliance barriers and service disruptions. This information is used to improve services, public information campaigns and to direct targeted assistance towards those who are in need.

Trend analysis services are also used to identify neologisms (new slang and jargon terms) in industry, trade organizations and criminal organizations. By being able to speak the common language of their target audience, this information assists investigators in decoding messages captured during surveillance and improves communications messaging when engaging with industry and taxpayers. The neologisms are also used by content analysis services as keywords for detecting underground economy activity online.

B. Policy

1) Identifying and Measuring the Underground Economy

The social media tax data warehouse can be used to identify and measure active sectors of the underground economy. The collector's content analysis and geolocation services scour social media posts that originate in the tax jurisdiction, for information that indicates participation in the underground economy (such as advertisements for tax free

goods and services, discounts for cash payment methods, sales of contraband goods etc.). While traditional tax databases only include those businesses that have registered with the government, the social media data warehouse can identify and measure those operating outside of the official tax system. Results are compiled to better understand those industries, geographic locations and demographics comprising the underground economy. This information helps tax authorities determine the size of the underground economy and where to focus their efforts. Baseline measurements of underground economic activity on social media are taken before and after a tax initiative to estimate their impact on compliance. Bots are used to establish prices and gauge market demand of contraband goods in the black market by negotiating sales and purchases on social media. Tax authorities use this information to establish optimal tax rates for these goods that will erode the black market and maximize tax revenue.

2) Economic Behavioural Studies

The social media data warehouse provides a unique opportunity to study and develop policies that address the demographic, economic and psychographic factors that contribute to an individual's propensity to participate in the underground economy. Geolocation and identification services discover persons of interest online. Content and sentiment analysis services reveal a host of personal information that is not available in government databases. Bots engage in conversations to elicit missing data to complete demographic, psychographic and economic profiles. Social media information is merged with existing tax data and applied to tax compliance models.

Data mining is conducted to discover unknown personal characteristics, behaviours, motivations and contributing factors for underground economy participation. Parametric analysis is conducted to establish the relative influence these factors have on tax compliance within the jurisdiction, as well as to predict how taxpayers will react to various policy alternatives, propaganda and compliance activities. Social network and trend analysis is performed, analyzing group tax compliance dynamics to determine how tax compliance attitudes and behaviour spread through social networks and how the compliant and non-compliant tend to group or cluster together socially. Tax returns and audit results are fed back through social media based tax models to improve their accuracy.

C. Risk Assessment

1) Taxpayer Profiling

Information gathered by the social media collector and stored in the warehouse supplements data already present in government databases to create more expansive profiles of individual taxpayers. The OECD recommends the inclusion of taxpayer's behaviour-based social and psychological profiles when evaluating the risk of non-compliance [1]. When assessing the risk of businesses and other organizations, the social media data of company executives can be used as a proxy. Content and sentiment analysis services gather basic information such as phone numbers, addresses, and birthdates and along with non-trivial personal characteristics and social, economic and psychological data. A wealth indicator, which

estimates a taxpayer's affluence, is constructed by analyzing text, images and other postings and further refined with information from other internal government databases. These taxpayer profiles are stored, integrated with data from the tax warehouse and used in downstream risk assessment processes such as data matching, mining and parametric modeling

2) *Data Matching*

Data matching is performed, to detect differences between tax submissions stored in the tax data warehouse and online profiles in the social media warehouse. Segments of the tax data warehouse are matched directly against social media to uncover connections to other businesses operating under the same phone numbers, addresses and web domains. Inconsistencies may be indicative of compliance risk and warrant further scrutiny.

Identification and geolocation services determine who is offering goods and services on social media within the tax jurisdiction. These identities are compared against tax registration databases. Individuals who are operating outside the tax system would be at high risk for non-compliance and likely to be participating in the underground economy, making them promising targets for compliance activities.

3) *Data Mining & Parametric Modeling*

The integrated social media and tax database profiles of taxpayers are mined to find relationships between social media information, taxpayer submissions and compliance risk. Results from tax assessments, inspections, audits and investigations are used as training data for data mining software to determine the factors best correlated with compliance and non-compliance. Those factors that are highly correlated with compliance or non-compliance are used in parametric models to assess risk. Online postings by recipients of social benefits, who are at increased risk of non-compliance, are examined by content analysis services for evidence of employment in the underground economy and other sources of unreported income.

Personalized programs that maximize compliance and revenue are created for each taxpayer based upon their integrated tax and social media data profiles. The results of audits and other compliance activities are continually fed back into risk models as training data to improve their accuracy and determine the most effective compliance response.

4) *Social Network Analysis*

Information about taxpayers' social network connections are gathered by the collector, integrated with tax data and analyzed using social networking analysis services. Taxpayers heavily connected to known participants in the underground economy are presumably at an increased risk of non-compliance and more likely to be participating in the underground economy.

When performing analysis solely on tax data, the social network is relatively limited. By incorporating social media network data, the social network becomes much richer and larger. Used in conjunction with data mining and matching, social network analysis uncovers linkages and relationships between individuals and data that were previously unknown, detecting anomalous behaviour and clustering that is indicative

of individual and group compliance risk. Network visualization tools are used to display social networks, connections and their properties. Tax authorities target compliance activities toward particular social network nodes (influencers) to maximize the deterrence impact of their efforts.

D. *Audits and Inspections*

Audit case management systems send requests to the collector for audit subjects' social media data. Results are stored in the warehouse for further analysis. Verification of information during traditional audits is assisted by automated cross-referencing social media postings against tax submissions. Warehouse data is processed using data matching and social networking analysis to reveal non-arms-length transactions and relationships between individuals, holding companies and subsidiaries. Wealth indicators derived from social media content are compared against tax submissions to find discrepancies between lifestyles and stated earnings. This is done in conjunction with social network analysis to determine if individuals are hiding wealth through gifts to family members and other acquaintances.

Case management input tools compile results for manual analysis by auditors and inspectors and assist in entering social media data into case management systems, improving audit and inspections efficiency. Visualization tools display social networks, connections between taxpayers, businesses and their assets.

1) *Online Inspections and Audits*

Regulating tax compliance behavior online is particularly challenging due to due to online anonymity and ambiguity over tax locale [29]. Furthermore, the line between social media and online marketplaces has blurred. Many online marketplaces have taken the form of social networks (e.g. sharing economy, market networks), while many social media platforms offer business hosting and transaction processing services. The social media data warehouse can be used to help curtail the underground economy in cyberspace through online inspections and audits.

The collector gathers sales and inventory information from online platforms for goods and services offered by high volume users operating in the tax jurisdiction. This data is stored in the warehouse and compared against tax submissions to find evidence of underreporting. Taxpayer, trade and professional association databases are compared against the warehouse to identify unregistered operators and undeclared sources of income. Advertisements for the sale of specially tax goods are cross referenced against dealer listings to find unlicensed sellers of these goods. Identification and geolocation services determine the seller's identity and location for a follow-up inspection or referral for investigation.

Auditors and inspectors trigger warehouse processes that automatically generate applications to courts for seize and desist orders against sellers and notice and take down orders against social media providers hosting businesses advertising underground economy goods and services. Social bots are used to take offensive measures against particular egregious online retailers by mounting denial of service attacks and adversely affecting their online reputations.

E. Investigations

Tax investigations are typically large-scale, high-risk endeavors that are concerned with the illicit trade of specially taxed goods and egregious cases of tax evasion. The role of organized crime in illicit trade, along with links to more serious crimes such as human trafficking and terrorism has been acknowledged; with social media networks being implicated in the recruitment of customers and the sale of contraband goods [30]. Compared to fieldwork, using social media data for preliminary intelligence gathering and social network analysis is relatively inexpensive and low risk.

1) Intelligence Gathering

The collector scans social media, generating investigations leads by identifying advertisements for specially taxed goods and detecting gang and criminal organization symbology using image analysis. Leads are forwarded to investigators for manual evaluation.

Investigations case management systems send requests to the collector for information regarding intelligence targets. Data is extracted from social media, while data brokers provide archives of historical social media and ecommerce postings. Automated processes send requests to social media platforms for private information. Dedicated content analysis tools in the analytic layer assist the manual analysis of social media content. Investigators use proprietary online investigation tools to map and track suspects by geotags, reveal their private social network connections and verify online identities. Investigators use these investigation tools to locate subjects, prove timelines, uncover associations and memberships, verify employment, determine a subject's reactions and opinions to events and establish the times, locations and attendees of meetings, amongst many other potential uses.

Case management input tools integrate the warehouse, online investigations tools and evidence repositories to automate the entry of social media data and content into those systems. This improves caseload efficiencies by reducing input effort and maintaining the integrity of data.

2) Social Network Analysis & Data Mining

Social network analysis is performed, helping to establish investigation subjects' network of associates. Geospatial data is analyzed to locate warehouses and purchase locations of bulk specially taxed goods. Social network visualization tools and data mining assist investigators in analyzing connections to determine criminal organization kingpins and key members of contraband distribution networks.

3) Evidence & Prosecution

The secured social media server is used to engage with investigations targets on social media to make purchases and gather intelligence. Social media connections are protected by internet proxies and encryption services to foil the potentially sophisticated counterintelligence methods used by criminal organizations. Social bots create social media accounts with credible back-stories and networks of established online contacts for safe online communications. Bots are also used to coerce suspects into downloading malware that captures chat transcripts, passwords, friends lists and records audio and video from mobile devices.

The evidence server provides a data secure repository for data and content retrieved from social media in the course of investigations. Its centralised logging services establish an audit trail to demonstrate that evidence gathering procedures were followed, improving admissibility in legal proceedings.

V. CHALLENGES

Implementing a social media data warehouse and analytics program for taxation presents a number of challenges. A multitude of social media platforms exist, each with their own data access models and formats. The number of users and volume of content is overwhelming. The popularity of platforms ebb and flow, and the platforms themselves change quickly with new technology, standards and features. This is compounded with the difficulty of accurately interpreting unstructured content; making analysis a complex undertaking. Many of the applications proposed in this paper, especially those that rely upon image, audio and video analysis, are dependent upon technological advances before they are feasible.

A broad survey of social media for evidence of underground economy activity requires a glossary of terms, common phrases and specific keywords in order to minimize expensive manual reviews of data. Data analysis requires the development of a metadata language, metrics, and data model for the storage and retrieval of social media data. It also requires an infrastructure to support the collection and analysis of data. Given the effort required, the unknown return on investment and risks of data disclosure, reuse and misuse, there may not be strong executive and political support for this type of tax innovation.

Tax authorities can leverage existing analytics platforms to implement social media analysis services, thus lowering costs and utilizing existing technical skills and knowledge. Information gathering and analysis could be contracted to third parties. This can reduce the initial infrastructure costs and serve as a method to absorb and develop the skills and knowledge required to implement and support an internal social media analytics program. However using an external data broker raises a number of data governance issues such as data ownership and privacy concerns.

The reliability of social media data is suspect. It may be difficult to confidently identify individuals online and differentiate humans from social bots. Once taxpayers become aware their social media activity is being monitored, they may employ counterintelligence measures to keep their personal information private or mislead tax authorities. Furthermore, the information individuals post online is inconsistent and may not be trustworthy. It may be difficult to establish confidence in the predictive and explanatory power of social media data due to small sample sizes and missing variables. Social media data may not statistically represent the taxpayer population, as social media is typically used by younger users and the various platforms appeal to particular demographics. Those who publicly post their information may tend to have certain beliefs, attitudes and personality characteristics. Results from one tax jurisdiction may not be applicable to others due to a wide range of factors, including access to internet and social

media usage, cultural considerations, the composition of the underground economy, as well differences in tax, economic, social and political systems.

Many of the potential uses discussed in this paper for policy development and risk scoring are speculative in nature and have not been directly applied to taxation. The literature has established relationships between economic, sociological, and psychological factors and compliance risk. But the strength of these relationships across the population is unknown and requires further study to construct metrics and models that produce actionable insights. Similarly, the diffusion of compliance attitudes and behavior through social networks is mostly theoretical, based upon computer simulations and untested in a practical tax setting.

Implementing the proposed applications also raise a number of ethical and legal concerns. The manipulation of public opinion has the potential to be used for political interference. Mass surveillance of social media could be considered an invasion of privacy by government; at odds with privacy legislation and viewed as a cost borne by taxpayers. Similarly, using social network analysis for initial risk assessment could be considered unfair, as it assumes guilt based upon association. In some tax jurisdictions it is considered unethical and unlawful to discriminate against people based upon characteristics such as age, race and gender and may not be used for assessing compliance risk.

VI. CONCLUSIONS

The social media tax data warehouse offers a number of opportunities to improve existing tax administration process and manage the underground economy. The warehouse could be extended to related applications such as detecting tax fraud rings, smuggling, offshore business arrangements, money laundering and other financial crimes. It could also be extended to other functional areas of government administration such as tax collections and benefits administration.

There are a number of technical, political, ethical and legal challenges that must be met before the full range of capabilities proposed in this paper can be realized. However, it is likely that the analysis of social media data will become increasingly important and commonplace in tax administration as more of social and economic activity goes online.

REFERENCES

- [1] OECD, "Compliance Risk Management: Managing and Improving Tax Compliance," 2004.
- [2] B. Russell, "Revenue Administration: Managing the Shadow Economy," 2010.
- [3] S. P. Green, "What Is Wrong With Tax Evasion?," *Houst. Bus. Tax J.*, no. October 2008, pp. 220–233, 2009.
- [4] M. L. Roberts, "An Experimental Approach to Changing Taxpayers' Attitudes Towards Fairness and compliance via Television," *J. Am. Tax. Assoc.*, vol. 16, no. 1, p. 67, 1994.
- [5] D. Onu and L. Oats, "'Paying tax is part of life': Social norms and social influence in tax communication," *J. Econ. Behav. Organ.*, vol. 124, no. 2015, pp. 29–42, 2015.
- [6] M. Hallsworth, J. A. List, R. D. Metcalfe, and I. Vlaev, "The behavioralist as tax collector: Using natural field experiments to enhance tax compliance," *J. Public Econ.*, vol. 148, pp. 14–31, 2017.
- [7] C. Lu, S. Huang, and P. Lo, "An empirical study of on-line tax filing acceptance model: Integrating TAM and TPB," *African J. Bus. Manag.*, vol. 4, no. May, pp. 800–810, 2010.
- [8] C. Baicu and C. V. Hapenciuc, "Model of Choices, Institutions and Direct Democracy Quasi-Economic Factors of the Influence of Underground Economy," *Ecoforum*, vol. 5, no. 2, pp. 213–218, 2016.
- [9] C. Lambertson, J. E. De Neve, and M. I. Norton, "Eliciting Taxpayer Preferences Increases Tax Compliance," 2014.
- [10] L. J. Stalans, K. A. Kinsey, and K. W. Smith, "Listening to Different Voices: Formation of Sanction Beliefs and Taxpaying Norms," *J. Appl. Soc. Psychol.*, vol. 21, no. 2, pp. 119–138, 1991.
- [11] J. L. Brown and K. D. Drake, "Network Ties Among Low-Tax Firms," *Account. Rev.*, vol. 89, no. 2, pp. 483–510, 2014.
- [12] A. L. Andrei, K. Comer, and M. Koehler, "An agent-based model of network effects on tax compliance and evasion," *J. Econ. Psychol.*, vol. 40, pp. 119–133, 2014.
- [13] M. G. Allingham and A. Sandmo, "Income tax evasion: a theoretical analysis," *J. Public Econ.*, vol. 1, no. 3–4, pp. 323–338, 1972.
- [14] C. Vellutini, "Risk-Based Audits: Assessing the Risks," in *Risk-Based Tax Audits: Approaches and Country Experiences*, M. Sultan Khwaja, R. Awasthi, and J. Loeprick, Eds. The World Bank, 2011.
- [15] M. Hainey, "Building and Integrating Databases for Risk Profiles in the United Kingdom," in *Risk-Based Tax Audits: Approaches and Country Experiences*, M. Sultan Khwaja, R. Awasthi, and J. Loeprick, Eds. The World Bank, 2011, pp. 65–70.
- [16] Australian National Audit Office, "The Australian Taxation Office's Use of Data Matching and Analytics in Tax Administration," 2008.
- [17] LexisNexis® Risk Solutions, "Survey of Law Enforcement Personnel and Their Use of Social Media," 2014.
- [18] J. Fitsanakis and M. S. Bolden, "Social Networking as a Paradigm Shift in Tactical Intelligence Collection," 2012.
- [19] S. D. Omand, J. Bartlett, and C. Miller, "Introducing Social Media Intelligence (SOCMINT)," *Intell. Natl. Secur.*, vol. 27, no. 6, pp. 801–823, 2012.
- [20] "Social Intelligence to Provide Social Media Analytics Engine to the National Insurance Crime Bureau to Combat Workers' Compensation Fraud," *Insurance Weekly News*, no. December, p. 103, 05-Sep-2014.
- [21] R. Verma and S. R. Mani, "Using Analytics for Insurance Fraud Detection," *Finsights*, pp. 76–84, Jan-2013.
- [22] W. W. Moe and D. A. Schweidel, "Opportunities for Innovation in Social Media Analytics," *J. Prod. Innov. Manag.*, vol. 34, no. 5, pp. 697–702, 2017.
- [23] S. Stieglitz, L. Dang-Xuan, A. Bruns, and C. Neuberger, "Social Media Analytics An Interdisciplinary Approach and Its Implications for Information Systems," *Bus. Inf. Syst. Eng.*, pp. 89–96, 2013.
- [24] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," 2016.
- [25] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "Design and Analysis of a Social Botnet," *Comput. Networks*, vol. 57, pp. 556–578, 2012.
- [26] G. D. K. Kishore, B. Anusha, G. Poornima, G. Anusha, and K. Tejaswi, "Decision Making and Reduction Techniques for Tax Revenue using Data Warehouse," *Int. J. Eng. Trends Technol.*, vol. 3, no. 1, pp. 24–31, 2012.
- [27] G. Yang and C. Wu, "Information Analysis and Decision System for Tax Revenue Based on Data Warehouse," *Comput. Appl. Syst. Model. (ICCSM)*, 2010 Int. Conf., vol. 6, no. Iccasm, 2010.
- [28] B. Batrinca and P. C. Treleven, "Social media analytics: a survey of techniques, tools and platforms," *AI Soc.*, pp. 89–116, 2015.
- [29] M. Keen, J. Toro, K. Baer, V. Perry, J. Norregaard, J. Ueda, J. Brondolo, D. Cleary, E. Hutton, O. Luca, E. Rojas, M. Thackray, and P. Wingender, "Current Challenges in Revenue Mobilization: Improving Tax Compliance," *Staff Rep.*, no. April, pp. 1–79, 2015.
- [30] OECD Publishing, "Illicit Trade: Converging Criminal Networks," Paris, 2016.