Noticia 174/2020-EN

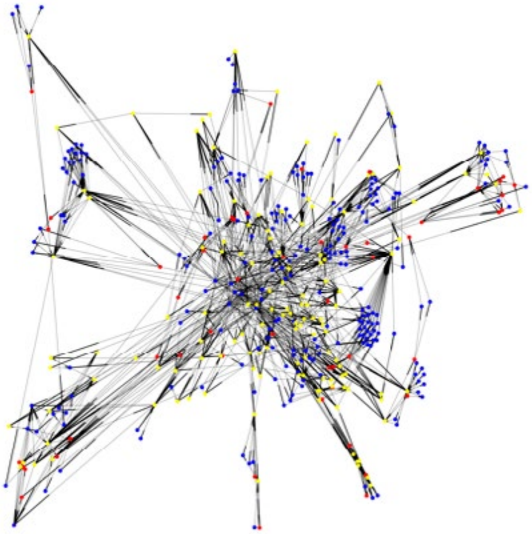
Artificial intelligence and network science against tax evasion

Laura Vargas-Parada

3 June 2020

A multidisciplinary team of researchers from the Centro de Ciencias de la Complejidad (C3) and the *Instituto de Física, Universidad Nacional Autónoma de México (UNAM)*, in collaboration with the Department of Network and Data Science at Central European University, Hungary, carried out an analysis using artificial intelligence and network science to estimate the amount of value added tax (VAT) evaded in Mexico by companies issuing electronic invoices that simulate operations that were never actually carried out.

"The main aim of the research was to develop methods to identify tax evaders and to estimate the amount of money lost due to evasion," explained Carlos Gershenson by email, Coordinator of C3's Computational Intelligence and Mathematical Modeling Program, Full Time Researcher at the *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM*, and co-author of the study.



Small network of anonymized RFCs (companies or individuals), where links represent invoices emitted by one node and received by another. Red nodes are known to sell invoices for tax evasion of those who receive them. Yellow ones are suspected to be part of their tax evasion network.

Authors estimate that evasion due to fraudulent digital tax receipts over the internet (called fraudulent CFDIs in Mexico) reached, during the period 2015 to 2018, an average of just over 60 billion pesos per year (2.5 billion USD), according to the report *VAT Evasion: Network Analysis* published at the web microsite of the local Tax Administration Service (called SAT), instance for which they carried out the analysis.

The researchers also found that the evading trend is increasing, having gone from 40 billion pesos in 2015 to 77 billion three years later, an increase of 93%. This evasion, according to investigators, comes from 7,677 federal taxpayer records (called RFCs) of potential tax evaders.

"This research has great merit because it uses artificial intelligence techniques to propose a public policy solution to a sensitive problem for countries with fiscal systems that have a structural weakness," said Roberto Ponce-López, Professor and Researcher at the *Tecnológico de Monterrey*, who was not involved in the study. "The authors used data science to tackle an endemic problem in developing countries. Identifying and contextualizing technological tools developed in industrialized countries to tackle local problems requires sensitivity and creativity."

According to Tomás Veloz, advisor on data science at the Office of the Comptroller General of the Republic of Chile and who did not participate in the study, this project allows, "with a very small investment, to solve a problem that has an important impact in the correct performance of the State and the trust of the people in the public administration."

Complex systems

For the analysis, scientists used methods developed in the study of complex systems that allow the analysis of a large amount of data. In this case, the analyzed data were all the digital tax vouchers issued in Mexico between January 2015 and December 2018, anonymized and aggregated by month.

Using algorithms based on artificial intelligence, researchers sought to reproduce the patterns in the activity of companies that had already been identified by SAT as Companies that carry out Simulated Operations (denoted EFOS) and, by means of statistical analysis of network theory, to "identify new patterns," Gershenson explained.

This way, the scientists developed a tool that not only allowed detecting patterns in the behavior of the issuance of tax receipts, identifying taxpayers with a similar behavior to tax evaders (and therefore suspicious), but also estimating how many resources they had evaded.

Regarding the methodology, Gerardo Iñiguez, Assistant Professor at Central European University, Visiting Researcher at Aalto University, Finland, and co-author of the study, explained in an electronic message: "We used two methods based on machine learning algorithms to detect possible EFOS not yet identified [by SAT] from a list of EFOS already detected."

In the first automated learning method, researchers used the characteristics of EFOS already detected as evaders to train a neural network. Once the neural network is trained, the same algorithm is applied to all RFCs in the study, finding the RFCs most similar to the EFOS already detected and, therefore, suspected of the same evasion activity.

"The second method works in a similar way," adds Iñiguez, a specialist in computational social science. In this case, the EFOS already detected are used to train a random forest, which reports the characteristics that any other RFC must have to be considered suspected of being an EFOS.

The learning of the two algorithms (neural networks and random forests) is a result from the accumulation of statistical information, which is what allows suspect EFOS to be identified. "If any RFC behaves statistically in a similar way or has characteristics similar to the EFOS already identified, then the algorithms detect it as an RFC suspected of being EFOS," explains Iñiguez, a physicist and Doctor in Computational Science.

By using two different methods researchers are able to contrast results and make their research more robust.

Limitations

For Iñiguez, one of the limitations of this research is that it does not allow estimating tax evasion in economic activity that leaves no online trace in SAT data, as is the case for the Mexican informal economy.

In Mexico, 56% of jobs are informal and generate 22% of the Gross Domestic Product (GDP), explained in an email Ponce-López, a specialist in combining spatial analysis tools with machine learning and large-scale databases to build information infrastructure. The problem with tax evasion and informality, he said, is that they limit the resources that the State could obtain to increase investments in health, education, infrastructure and social programs.

Another very relevant limitation, clarifies Iñiguez, is that "there is a possibility that the behavior of some honest taxpayers is statistically similar to that of the already identified EFOS, and therefore these taxpayers will be erroneously classified as suspicious EFOS".

For this reason, for Gershenson, a specialist in adaptive and evolutionary systems, "these types of tools do not replace SAT experts, but they can help them identify more potential evaders in less time, thus allowing them to optimize the use of public resources" and help them in their fight against tax evasion.

Ponce-López considers that a limitation faced by this type of study is "obtaining more records from the so-called "*factureros*" [that is, tax evaders] to increase the size of the tool's training base." The researcher believes that this could be solved as the availability of robustness and digital information quality improves.

Veloz, who is also Director of Mathematical Modeling at the Leo Apostel Center for Interdisciplinary Studies, Vrije Universiteit, Belgium, considers that the application of machine learning in classification tasks has some limitations such as the dependence of the method on the data, because the data can have multiple biases causing artificial intelligence methods to learn in a biased way what is a simulated tax operation. This would cause the system to be unable to detect certain types of simulated operations, or, as Iñiguez mentioned, could classify certain non-operations as suspicious. "For this you must work continually on improving the way artificial intelligence learns, and that can take a lot of work," explains the academic.

A further limitation, in Ponce-López's opinion, is the institutional capability of making the most out of this type of tool, since it can only be useful "to the extent that SAT [and the authorities] makes use of it", which requires important organizational and procedural changes. "Organizational change is usually an even more complicated challenge than the technological challenge," added the academic in an interview via email.

Ponce-León highlights, however, that the investigation is in itself a clear example of the degree of collaboration achieved between the authors and the personnel in charge of collecting and classifying the tax information of millions of taxpayers. "The analysis and specification of the different models illustrates an important degree of collaboration between the researchers and public officials. The latter represents a successful case of collaboration and organizational transformation."

Network theory to identify other evasion patterns

Statistical analysis of network theory involves defining a network where each node is an RFC and a link describes a tax receipt (called CFDI) between two taxpayers. The structure of the network indicates all of the economic activity between RFCs in Mexico, according to the SAT, month by month through several years.

By studying the statistical properties of the structure of this network (measures of connectivity, centrality, etc.) it is possible to identify the typical patterns of nodes and links found around EFOS already identified by SAT. Searching for other RFCs in the network that have a similar structure around them is what allows identifying the RFCs suspected of being EFOS, explained Iñiguez.

Researchers consider that, in the near future, in collaboration with SAT, they could carry out a more advanced analysis of network theory in order to detect "typical patterns of collaboration between taxpayers (called temporal motifs in network science) that are related to evasion".

A motif is a specific pattern between a small number of nodes (for example, 3 connections between 3 nodes: a triangle). A temporal motif is a motif that also has information on when the links (in this case transactions between RFCs) appear in time.

The idea is to look for these specific patterns that commonly appear around an EFOS (for example, a triangle or something more complicated) and look for other RFCs that have similar patterns in the network around them.

In this way, it is possible to specifically measure how EFOS interact with other RFCs, look for that same structure in other parts of the network, and thus identify in more detail the typical practices that promote tax evasion, such as the use of generic RFC codes and self-billing within evader circles. This will be essential to "create automatic detection and monitoring systems for suspicious EFOS," explains Iñiguez.

An important aspect of this research is that, once the system is trained, it is also possible to identify new evasion patterns. "Our analysis will allow us to identify (in the future) more sophisticated evasion methods. One of them is circularity, the next step in complexity of self-billing: instead of billing itself, an EFOS could bill another RFC, which bills another RFC, and so on several times until billing reaches the initial EFOS (forming a circle), which makes identifying tax evasion more difficult. This type of behavior or statistical pattern (in fact a motif) is something that can be discovered through network analysis tools like the ones we are using" explained Iñiguez.

Relevance

The relevance of this study is "giant," says Veloz. "From a historical point of view, the poor integration of science in public policy has been a major obstacle to the development of countries and the approach to social justice," wrote in an electronic message the also Director of the Foundation for the Interdisciplinary Development of Science, Technology and the Arts, in Santiago de Chile. "It is urgent to link science and the academic world with the processes of public policy in order to achieve more efficient results. For me, data science is key to allow such a link, because these tools can speed work tremendously, and even achieve results that no team of people ever could, since massive data processing is out of the scope of the information that we are capable of handling as human beings."

For the specialist in mathematical modeling, data science, and interdisciplinary studies, this type of analysis is one of the most "useful": the use of big data for automated classification of illegal conduct. For the Chilean academic, with the information provided by this tool, SAT staff will be able to gain efficiency in their work. Something known as business intelligence.

"With a relatively small investment, public agencies can team up with data scientists to conduct business intelligence and improve tremendously their processes. In the future, we hope that this will be established to put specialized knowledge at the service of State processes," he concluded.