UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Construcción de una base de datos de tipo grafo para Subasta Inversa Electrónica

Mara Sofía Fortuny Moya

Carrera de Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniera en Ciencias de la Computación

Quito, 23 de mayo de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Construcción de una base de datos de tipo grafo para Subasta Inversa Electrónica

Mara Sofía Fortuny Moya

Nombre del profesor, Título académico **Daniel Riofrío, Ph.D.**

Quito, 23 de mayo de 2023

© DERECHOS DE AUTOR

Nombres y apellidos: Mara Sofía Fortuny Moya

Código: 00205125

Cédula de identidad: 1717324162

Lugar y fecha: Quito, 23 de mayo de 2023

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

En respuesta a la gran influencia que tiene la contratación pública en el mercado económico de Ecuador, la ley ecuatoriana fomenta el uso de la Subasta Inversa Electrónica "SIE" para promover la competencia entre proveedores que ofrecen bienes y servicios al Estado. Sin embargo, en condiciones de subasta, compradores y proveedores se las han ingeniado para violar las normas y principios que se han establecido para mitigar los riesgos de corrupción que existen en la contratación pública. Dichas infracciones suelen dejar rastros en los datos que son capturados por los medios electrónicos a través de los cuales se llevan a cabo las subastas inversas. Sin embargo, cuando esta información se publica, se extrae y se almacena en una base de datos SQL, los usuarios pierden la capacidad de visualizar patrones y anomalías en las relaciones que existen entre los objetos que componen este dominio. Por ese lado, hemos identificado la necesidad de transformar la forma predeterminada en que almacenamos los datos de contratación pública y, para ello, hemos construido un grafo para la Subasta Inversa Electrónica. En el modelo de datos, los procesos SIE que han sido adjudicados por las Entidades de la red se agrupan según su Mercado. A su vez, cada Contrato agrupa a los Proveedores que participaron en el proceso de subasta, y está vinculado -a través de una arista diferente- al participante que resultó adjudicatario del Contrato. Todos estos proveedores (ganadores o no) están vinculados a las Ofertas que ofrecieron durante un proceso de subasta y, en su caso, a los Accionistas que son propietarios parciales de dichas compañías. El presente trabajo describe el proceso que permitió la creación de estos nodos y relaciones en Neo4j, gestor de bases de datos de tipo grafo que, por ahora, almacena aproximadamente 365 mil objetos y 787 mil conexiones que pertenecen a la red de SIE en el Ecuador desde 2008-03-24 hasta 2022-09-15.

Palabras clave: contratación pública, subasta inversa electrónica, base de datos, grafo, detección de anomalías

ABSTRACT

In response to the great influence that public procurement has on Ecuador's economic market, the Ecuadorian law encourages the use of Electronic Reverse Auction (also known as Subasta Inversa Electrónica, SIE, in Spanish) to promote competition among suppliers that offer goods and services to the State. Yet, under auction conditions, buyers and suppliers have managed to find ways to violate the rules and principles that have been set out to mitigate the corruption risks that exist in public procurement. Such violations tend to leave behind a trail of evidence in the data that is captured by the electronic means through which reverse auctions are carried out. However, when this information is published, scraped, and stored within an SQL database, users lose the ability to visualize patterns and anomalies among the relationships that exist between the objects that make up this overtly relationship-centered domain. On that side, we've identified the need to transform the default way we store procurement data and, for this, we've built a graph for Electronic Reverse Auction. In the data model, SIE processes that have been awarded by the network's Entities are grouped according to their Market. In turn, each Contract groups the Suppliers that participated in the auction process, and is linked – through a different edge – to the participant that was awarded the Contract. All these suppliers (winners or not) are linked to the Bids they offered during an auction process and, if applicable, to the Stakeholders that are partial owners of such companies. The present paper describes the process that allowed for the creation of these nodes and relationships in Neo4j, a graph database management system that, as for now, stores approximately 365 thousand objects and 787 thousand connections that belong in the Ecuador's Electronic Reverse Auction network from 2008-03-24 to 2022-09-15.

Key words: public procurement, electronic reverse auction, graph database, anomaly detection

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

1. INTRODUCCIÓN

De todos los tipos de contratación pública que hay en el Ecuador, la Subasta Inversa Electrónica "SIE" es la modalidad con mayor monto adjudicado en el país. Para hacerse una idea, en el 2021 se publicaron 200.597 procedimientos en el Sistema Oficial de Contratación del Estado "SOCE", de los cuales 17.498 (el 8%) se llevaron a cabo en la forma de Subasta Inversa Electrónica. Se estima este pequeño porcentaje de procedimientos causaron la adjudicación de un total de 1.407,2 millones de dólares. En contraste, en ese mismo año, la modalidad de Catálogo Electrónico se aplicó en el 85,4% de los procesos de compra pública, pero el monto adjudicado a través de este tipo de proceso fue de 462,5 millones de dólares, aproximadamente un 33% de la cantidad de dinero que movió la Subasta Inversa Electrónica durante el 2021 (Servicio Nacional de Contratación Pública, 2020). Por ende, aunque SIE no es la modalidad preferida por las entidades contratantes, la mayor cantidad del gasto público destinado a la compra pública se concentra en los procesos de este tipo. Esto la convierte en objeto de interés para quienes están liderando iniciativas anticorrupción en el contexto de compras públicas, una actividad que en el Ecuador movió 5.320,5 millones de dólares en el 2021, "representando el 16,59% del Presupuesto General del Estado (PGE) y el 5,05% del Producto Interno Bruto (PIB)" (Servicio Nacional de Contratación Pública, 2020) .

En respuesta a la gran influencia que las contrataciones públicas tienen sobre el mercado económico del país, se han adoptado modalidades de compra que son dinámicas y que promueven competencia entre los proveedores que ofertan bienes y servicios al Estado. En los procesos de Subasta Inversa Electrónica, particularmente, se realiza un proceso análogo a las subastas que conocemos, en donde se realiza la venta pública de bienes o servicios al mejor postor, es decir, a quien está dispuesto a ofrecer el precio más alto entre un pool de compradores interesados. Sin embargo, los procesos SIE son subastas inversas en el sentido de que, en vez

de realizarse una venta, se realiza una compra al proveedor que ofrece el precio más bajo entre un pool de vendedores que desean conseguir la adjudicación de un contrato. Este proceso, como tal, se lleva a cabo a través de medios electrónicos, en el Sistema Oficial de Contratación Pública del Ecuador "SOCE". En él, se habilitan espacios para que los proveedores participantes pujen hacia la baja del precio ofertado durante un tiempo limitado. Todas las ofertas emitidas son capturadas por el portal, y transparentadas al público general una vez que la entidad contratante ha realizado la adjudicación del contrato al ganador de la puja (Servicio Nacional de Contratación Pública, 2023).

De esa manera, a través de una dinámica que evita la contratación directa en virtud de impulsar la competencia entre proveedores, se busca mitigar algunos de los riesgos de corrupción que existen en la compra pública (i.e. favoritismo). Sin embargo, se ha demostrado que, aún en condiciones de subasta, se puede corromper la integridad del procedimiento. Por ejemplo, la Oficina de Naciones Unidas contra la Droga y el Delito reporta que, durante los procesos de puja, los ofertantes pueden realizar prácticas coordinadas para distorsionar la competencia, o para inflar de forma artificial los precios de los bienes o servicios (2020). A estos casos de confabulación se los conoce como carteles y son reconocidos por la Organización para la Cooperación y el Desarrollo Económicos "OECD" como acciones que violan severamente los principios de competencia que han sido establecidos para promover eficiencia, impulsar la oferta y conseguir precios justificablemente bajos a cambio de bienes y servicios que concuerdan con las necesidades de las entidades contratantes (2016).

El problema que actualmente se enfrenta en el marco de lo que se ha descrito es la detección de dichos carteles. Esto se debe a que, debido a su naturaleza corrupta, estos actos de confabulación se llevan a cabo de forma encubierta y, en muchos casos, pretenden simular o falsear el comportamiento competitivo (Sampford, Shacklock, Connors, & Galtung, 2006).

Afortunadamente, los carteles sí dejan algunos trazos, pero dichos indicios se manifiestan en datasets tan grandes que no pueden ser procesados manualmente (Aarvik, 2019). Por ende, los investigadores en el campo han recurrido al uso de data mining e inteligencia artificial para identificar indicios de manipulación de ofertas en datos de puja y, a través de estas herramientas, han logrado automatizar la detección de estos comportamientos (ver Estado del Arte). Sin embargo, en el presente contexto, ninguno de los trabajos de investigación ha aplicado técnicas de detección de anomalías basadas en grafos. De hecho, en el estado del arte relacionado con la detección de anomalías e indicios de riesgos de corrupción en subastas inversas, la tendencia es utilizar tablas de datos relacionales, y buscar anomalías en los valores de licitación de los datos. Esto significa que se ha hecho muy poco para preservar la estructura de grafo que es inherente a los datos de este dominio, y para detectar relaciones anómalas dentro de la información estructural que ahí se almacena.

En respuesta a este problema (Levenon & Kumalesh, 2017), la presente investigación se centra en la construcción de una base de datos de tipo grafo para procesos de Subasta Inversa Electrónica. En el trabajo, se busca reunir la información necesaria para generar un grafo que relacione mercados, entidades contratantes, proveedores, accionistas y procesos de contratación. La visión detrás de este trabajo es establecer un marco para la detección de esquemas de corrupción y anomalías que se manifiestan naturalmente en grafos. A largo plazo, no sólo se espera que esta base de datos almacene datos sobre SIE en su forma natural, sino que resuelva problemas de grafos de forma más eficiente y en menos tiempo que utilizando consultas a bases de datos relacionales (Vukotic & Watt, 2014). Mientras tanto, se espera hacer uso de esta herramienta para hacer una detección semi-automática de indicios de riesgos de corrupción para los cuales todavía no existe un algoritmo que automatice su detección.

El presente artículo describe la metodología que se aplicó para reunir la información necesaria sobre los nodos y relaciones que conforman la red de Subasta Inversa Electrónica, y los pasos que se tomaron para transferir los datos extraídos a Neo4j, un sistema gestor de bases de datos de tipo grafo.

2. ESTADO DEL ARTE

La detección de anomalías en datos de contratación pública no es un tema de investigación muy reciente. En la década de 1980, Robert H. Porter y J. Douglas Zona acogieron el reto de detectar comportamientos anómalos en las licitaciones de contratos de construcción de autopistas estatales en Long Island. A falta de modelos computacionales de aprendizaje automático, los autores esbozaron un procedimiento econométrico de prueba que estima el comportamiento competitivo de las licitaciones en el mercado y detecta las anomalías que se apartan de la tendencia prevista (1993). De este modo, el modelo propuesto identifica ofertas sospechosas que no se ajustan a lo que matemáticamente se considera un comportamiento competitivo.

De manera similar, en 2014, dos investigadores japoneses – Kawai y Nakabayashi – buscaron un método sencillo para detectar colusión en subastas de contratación pública de Japón. Para ello, desarrollaron un artículo en el que, inspirándose en el diseño de regresión discontinua, los autores analizaron las tendencias entre los licitadores más bajos y los segundos más bajos que habían en cada una de las dos rondas de ofertas que suelen tener lugar en las subastas utilizadas en proyectos de construcción pública. Como resultado, desarrollaron una prueba estadística que detecta comportamientos colusorios en las subastas en las que la persistencia de la identidad del licitador más bajo señala la presencia de un ganador designado y, por tanto, de un anillo de ofertas (Kawai & Nakabayashi, 2014).

En ese mismo año, Ting Sun y Leonardo J. Sales se lanzaron a hacer algo más complicado en el marco de detección de anomalías en datos de contratación pública. En su artículo, ellos compartieron su experiencia con la aplicación de redes neuronales para predecir irregularidades en un dataset de procesos de subasta. En este caso, en lugar de aplicar pruebas econométricas sobre la información de licitaciones, estos investigadores aplicaron dos redes neuronales artificiales, una tradicional y otra profunda, para detectar irregularidades en los datos que

describen las características de los contratistas que participaron en procesos de subasta de Brasil. Como resultado, el rendimiento predictivo de la red neuronal profunda resultante superó a la de otros dos algoritmos (regresión logística y análisis de función discriminante) que se emplearon para identificar las mismas irregularidades en la data (Sun & Sales, 2014). Por lo que, a través de esta investigación, Sun y Sales abogaron a favor del uso de datos no estructurados y métodos avanzados de aprendizaje automático para hacer detección de anomalías en datasets de contratación pública.

Sin embargo, ningún autor de la literatura mencionó la aplicación de grafos para realizar identificación de comportamientos irregulares o anticompetitivos en datos sobre subasta inversa. De hecho, en el artículo "Detección de fraude: Una revisión bibliográfica sistemática de los enfoques de detección de anomalías basados en grafos", los autores revelan que, entre el 2007 y el 2018, la mayoría de los estudios que aplicaron técnicas de GBAD (*graph-based anomaly detection*) lo hicieron para hacer detección de fraude en datasets de seguros, telecomunicaciones, bancos y tarjetas de crédito, es decir, en datos de la industria de FinTech (Pourhabibi, Ong, Kam & Boo, 2020). Como tal, no hubo mención de la aplicación de estas técnicas para detectar anomalías en datasets de compra pública, y mucho menos en subasta inversa.

Por ese lado, el estudio realizado permitió conocer el estado del arte de los métodos que, hasta la fecha, se han aplicado para identificar comportamientos sospechosos y anticompetitivos en información estructurada y no estructurada de subasta inversa. Y, como resultado de este diagnóstico, se encontró un déficit en la investigación que se ha hecho para representar datos de contratación pública en un grafo, y encontrar anomalías en los nodos y relaciones que componen esta estructura. Por ello, las siguientes secciones describen el trabajo que se realizó para avanzar dentro de esta materia en el contexto local de Ecuador.

3. METODOLOGIA

A continuación, se describe la metodología que se aplicó para reunir y complementar los datos de Subasta Inversa Electrónica que están disponibles en el portal oficial de compras públicas del Ecuador, más los pasos que se tomaron para levantar una base de datos de tipo grafo a partir de la información extraída.

3.1. Fuentes de datos

3.1.1. Sistema Oficial de Contratación Pública del Ecuador

De forma general, los datos sobre procesos históricos y actuales de Subasta Inversa Electrónica están disponibles en el Sistema Oficial de Contratación Pública del Ecuador "SOCE". Pero, los resultados de subasta – la información de las ofertas emitidas durante el proceso de licitación – solo se publican en este portal a través de un módulo que se habilita una vez que un proceso SIE avanza hacia las etapas de adjudicación y ejecución del contrato (Servicio Nacional de Contratación Pública, 2023). En el módulo habilitado, se publican los detalles de la puja, incluyendo: el nombre de los proveedores que participaron en la puja, los valores de las pujas realizadas por cada uno de los proveedores y la hora a la que emitieron esos valores.

Adicional a esto, la sección contiene un resumen sobre la puja, en donde se publican los resultados del análisis de vinculaciones que se lleva a cabo por cada participante del proceso. A esta información se agrega el origen (nacional o internacional) de cada proveedor, y el tipo de empresa (micro, pequeña, etc).

Y, como la adjudicación del contrato acontece previamente a la ejecución del mismo, el nombre del ganador de la puja también está incluido en este módulo. Las entidades contratantes están obligadas a justificar el por qué lo seleccionaron, por lo que el portal también incluye la razón y fecha de la adjudicación.

Desafortunadamente, la información que se habilita en el módulo de "Resultados de Subasta" es insuficiente para detectar esquemas complejos de corrupción. Por ejemplo, dado el nombre de una empresa, es imposible saber si sus integrantes están vinculados a otro proveedor participante. Tampoco hay cómo conocer el año de constitución de una empresa, ni su razón social. Por ende, se complementó el dataset con la información que está disponible en la Superintendencia de Compañías.

3.1.2. Superintendencia de Compañías

La Superintendencia de Compañías, Valores y Seguros es el organismo que está a cargo de vigilar y controlar la organización, actividades, funcionamiento, disolución y liquidación de las compañías. Y, entre todos los servicios que ofrece a la ciudadanía, consta un buscador de compañías que, a través de un portal web, transparenta información acerca de todas las razones sociales que están registradas en su base de datos (International Center for Journalists, 2020).

Dado el RUC, nombre o expediente de una compañía que está registrada en la Superintendencia de Compañías, el portal habilita información de tipo general (por ejemplo, la fecha de constitución, el tipo de compañía, su ubicación, el capital a la fecha, la actividad económica, etc) y listas para mostrar los administradores actuales, los administradores anteriores, los accionistas, y la documentación legal de la empresa.

De manera que, para complementar el dataset original del proyecto, se extrajo una lista de los proveedores que han participado en las pujas de los procesos seleccionados y, por cada uno de ellos, se buscó su información en la Superintendencia de Compañías. Y, para restringir un poco el alcance de esta investigación, no se descargaron documentos legales ni información contable.

3.2. Extracción de datos

En la sección anterior, se describieron dos fuentes de datos que, en el Ecuador, ofrecen información oficial con respecto a procesos de Subasta Inversa Electrónica, y con respecto a las compañías que operan legalmente en el país. Estos datos están únicamente disponibles a través de los portales web del SOCE, y de la Superintendencia de Compañías. Un usuario común no tiene acceso a las bases de datos que consolidan la información que se hace pública a través de esos sitios web, por lo que la extracción de la data de este proyecto ocurrió a través de bots que ejecutan tareas de web crawling y web scraping.

3.2.1. Web Crawler para el Sistema Oficial de Contratación Pública del Ecuador "SOCE"

Previamente a la ejecución de esta investigación, se desarrolló un web crawler que descarga, de forma masiva, los datos y archivos que se publican en el portal del SOCE para procesos de SIE que están en proceso de ejecución, o que fueron ejecutados en el pasado. El paper, "Towards Smart Citizen Control in Public Procurement: Ecuador's Case Study", describe la metodología que se aplicó para crear ese bot, y almacenar la información descargada por él (Fortuny, Guerrero, Riofrío & Simon, 2023).

En la base de datos correspondiente, se detectó un total de 447.783 procesos de SIE descargados. Sin embargo, no todos estos procesos tienen información de pujas ya que no todos ellos fueron extraídos mientras cursaban las etapas de ejecución de contrato y registro de contratos. Solamente en estas fases se habilita el módulo de 'Resultados de Subasta', por lo que, si un proceso no se encuentra en alguna de estas dos etapas cuando su información está siendo descargada, se pierde la ventana de oportunidad de extraer su bidding data.

Debido a esto, solo se encontraron 66.753 procesos de SIE con información en la sección de 'Resultados de Subasta'. Aproximadamente 40 mil de ellos fueron descartados debido a que, en su mayoría, eran procesos de SIE que terminaron en negociación. Esto sucede de forma lícita cuando una entidad contratante publica un proceso de subasta inversa electrónica y se dan las siguientes condiciones (Servicio Nacional de Contratación Pública, 2023):

1. Varios proveedores envían sus ofertas técnicas, pero la entidad contratante descalifica a todas excepto a una, entonces, se procede a realizar una negociación con el proveedor calificado.
2. Solo un proveedor envía su oferta técnica, y éste clasifica la etapa de calificación técnica, entonces, la entidad contratante procede a negociar con el proveedor calificado.

Sin embargo, esta investigación trabaja exclusivamente sobre datos de pujas por lo que no se consideran los procesos de tipo de SIE en cuyo ciclo consta la etapa de negociación en vez de subasta. En ese sentido, nos quedamos con 27.719 procesos SIE en cuyo proceso de subasta participaron mínimo 2 participantes calificados.

3.2.2. Web Scraper para la Superintendencia de Compañías

En esos 27.719 procesos, se identificaron 14.706 proveedores únicos. Es decir, al reunir la información sobre todos los participantes que pujaron en los procesos seleccionados, se identificaron casi 15 mil nombres de empresas y personas que participaron en uno o más procesos de subasta inversa electrónica.

Esta lista de nombres de proveedores únicos fue recibida como input en un programa que fue desarrollado para realizar tareas de extracción automática (*web scraping* en inglés) en el buscador de compañías de la Superintendencia. Este bot basado en Selenium fue implementado

para este proyecto en específico, y el flujo de su operación se encuentra diagramado en la Figura 1.

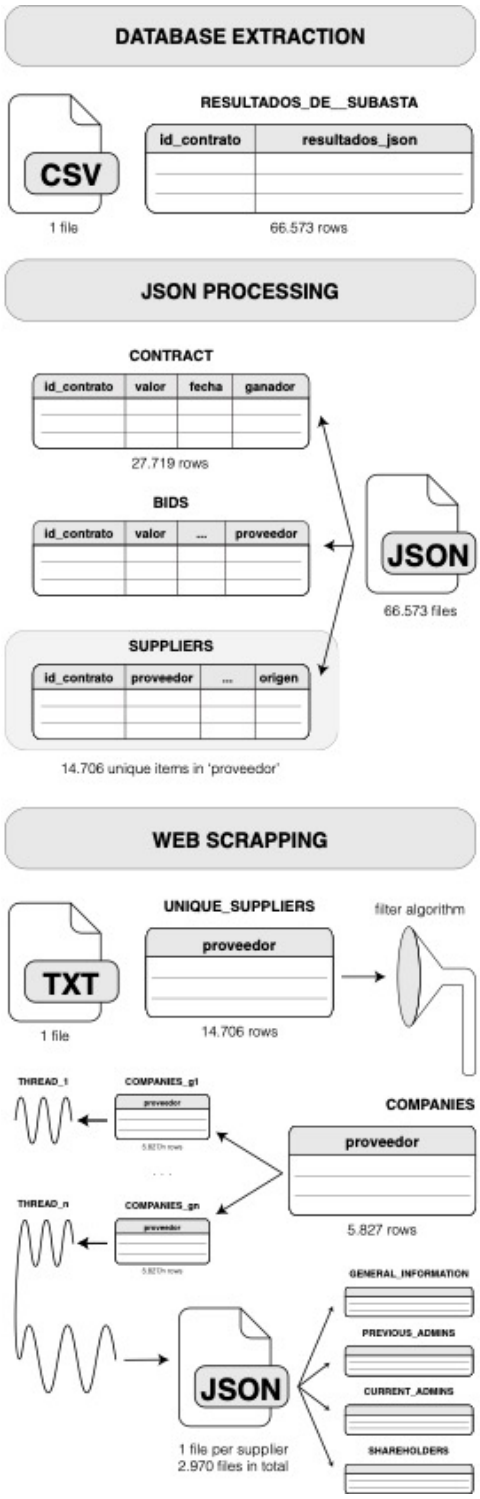

Figura 1: Flujo de datos que ingresan a y salen del web scraper de la Superintendencia de Compañías (en inglés)

En la figura se observa que, apenas recibe la lista de nombres, el programa utiliza un algoritmo sencillo para filtrar solo los nombres de compañías. Estos son strings que contienen abreviaturas tales como SA (sociedad anónima, por sus iniciales en español), CIA (compañía), EP (empresa pública), y que contienen palabras que no corresponden a nombres o apellidos de personas.

De manera que, después de aplicar este filtro a la lista de 14.706 nombres de proveedores únicos, el programa retuvo 5.827 nombres de potenciales compañías, y el resto se exportó a un archivo de texto. A continuación, se generaron *n* subgrupos de nombres de compañías con el fin de asignar estos recursos a *n* hilos de ejecución. Con estos hilos, el programa realizó *n* búsquedas simultáneas en la página de la Superintendencia de Compañías. Y, en cada una de estas búsquedas, se introdujo el nombre de una de las compañías de la lista en el buscador y, al existir una correspondencia exacta entre palabra ingresada y uno de los nombres del portal, se habilitó el acceso a la información de la compañía ingresada.

A través de este proceso, se logró detectar un total de 2.857 nombres de potenciales compañías que no fueron reconocidos por el sistema de la Superintendencia. En este grupo de palabras, se encuentran nombres de empresas que en el portal del SOCE no fueron escritos tal cual están registrados de forma oficial en Superintendencia de Compañías (ver Figura 1). Asimismo, encontramos nombres de personas que fueron incorrectamente filtrados por el programa, y nombres de fundaciones y asociaciones que no constan como compañías en el sitio. Sea cual sea el motivo, el programa no consiguió acceso a la información de esos 2.857 nombres de potenciales compañías, pero procedió con la extracción de datos asociados a los 2.970 nombres restantes.

En este proceso de extracción, cada hilo ejecutó funciones para recolectar la información contenida en los módulos que, por cada compañía, se habilitaban tras atravesar de forma exitosa

su página de búsqueda. De los 13 módulos habilitados, el proceso de extracción automática se llevó a cabo exclusivamente sobre las secciones de: Información General, Administradores Actuales, Administradores Anteriores y Accionistas.

Posteriormente, los datos descargados se exportaron en formato JSON a una carpeta local. Se exportó un JSON por cada una de las 2.970 empresas identificadas como válidas por el programa, y se procedió a utilizar estos datos en las fases posteriores de este proyecto.

3.3. Preprocesamiento de datos

3.3.1. Descomposición de los JSONs

Para poder construir una base de datos de tipo grafo a partir de la información descargada, hubo que desagregar los JSONs usando scripts basados en Python. Esto, como resultado, produjo la formación de las tablas que se observan en la Figura 1. Particularmente, los JSONs que se extrajeron del SOCE se descompusieron en tres tablas – Contratos, Pujas y Proveedores –, mientras que los JSONs que se obtuvieron de la Superintendencia de Compañías se desagregaron en 4 tablas – Información General, Administradores Actuales, Accionistas, y Administradores Anteriores –.

Como paso extra, se complementó la información de estas tablas a partir de la extracción de datos adicionales que estaban contenidos en la base de datos asociada al portal del SOCE. En este caso, resulta que la información general del contrato no está incluida en el módulo de 'Resultados de Subasta'. En dicho módulo, por ejemplo, no aparece el nombre de la entidad que adjudicó el contrato al ganador de la puja, ni tampoco la categoría (localmente conocido como CPC nivel 5) de los productos o servicios que fueron comprados a través del proceso de compra pública. Estos datos se desplegaron en un módulo aparte – en la página principal del contrato – y, por ende, se los tuvo que sacar de sus respectivas tablas SQL usando operaciones

de tipo *join*. Como resultado, se obtuvo un nuevo archivo CSV que contiene 40.577 filas de datos, una por cada contrato que aparece en la tabla Contratos. Estos nuevos datos no aparecen en la Figura 1, pero sus variables se mencionan en la Figura 2 (bajo el nodo Contrato).

3.3.2. Modelo de datos

Ahora, para poder convertir esta data en un grafo, hubo que conectar temporalmente la información de estas tablas a través de un esquema relacional. Para ello, primero identificamos los tipos de nodos y los tipos de relaciones (también conocidos como aristas) que modelan la información que ha sido capturada en las etapas previas de este trabajo, véase la Figura 2.

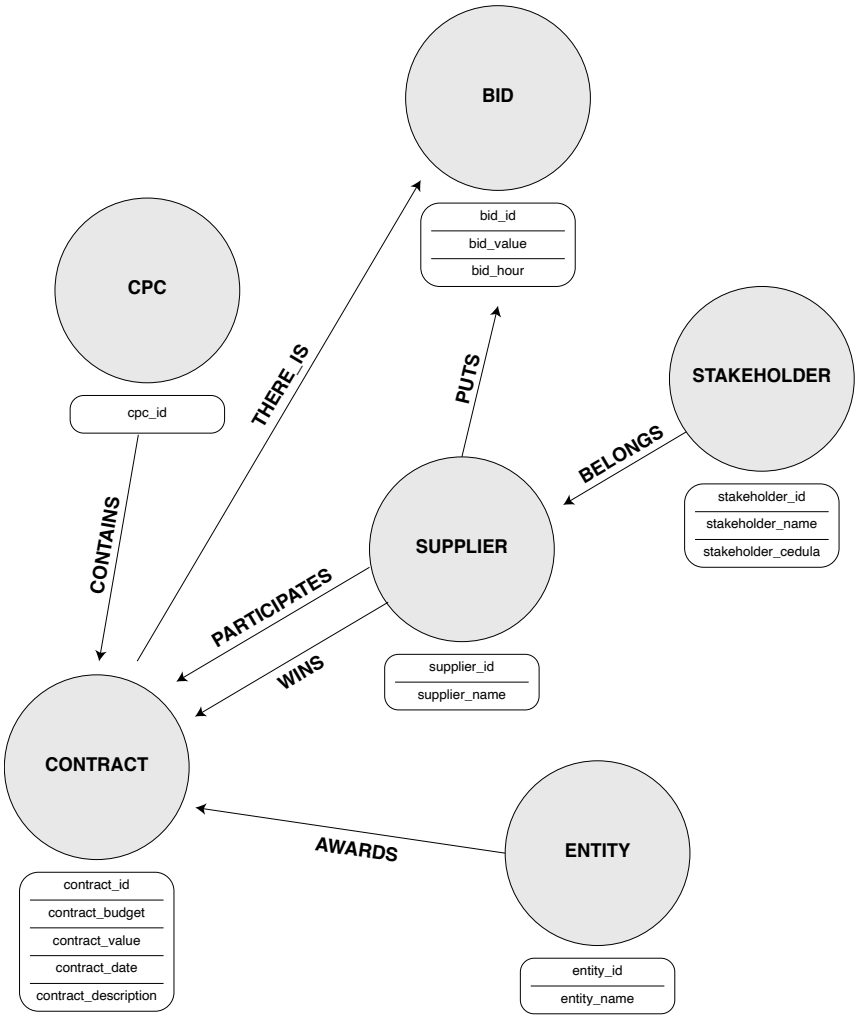

Figura 2: Modelo de datos del grafo que identifica a los principales actores de la Subasta Inversa Electrónica, y los relaciona (en inglés)

En esta figura, consta un modelo de grafo en donde los procesos de Subasta Inversa Electrónica que han sido adjudicados por las Entidades de la red se agrupan según el código del Mercado, también conocido como Clasificador Central de Productos "CPC". A su vez, cada Contrato agrupa a los Proveedores que participaron en el proceso de subasta, y se enlaza – a través de una arista diferente – al participante consiguió la adjudicación del contrato. Para haber llegado a ese punto, cada uno de los proveedores participantes debió haber emitido al menos una Puja en el sistema electrónico, por lo que esta información se captura en un nodo aparte que está asociado al proveedor. Por último, asumimos la presencia de Accionistas en esta red y conectamos cada uno de ellos a las compañías – nodos de tipo Proveedor – en las que tienen acciones según la información descargada desde la Superintendencia de Compañías. Cabe destacar que, para este primer experimento, se ha omitido la información de los administradores actuales y anteriores de cada compañía del grafo.

3.3.3. Tablas de transición

Una vez diagramado este modelo, se procedió a crear tablas que, para empezar, asignaban un ID numérico incremental a cada uno de los nodos del grafo, y registraban las propiedades asociadas con cada uno de ellos (ver las variables que aparecen debajo de los nodos de la Figura 2). Una vez realizada esta asignación de Ids, se construyeron las tablas que, mediante un esquema relacional, conectaban los nodos del grafo.

CONTRACT

id	budget	value	date	description
1	USD 7,900.00	USD 6,500.00	10/12/22	. . .
2	USD 204,00.00	USD 155,354.00	05/01/23	. . .
3	USD 45,00.00	USD 30,300.00	27/11/22	. . .

(a)

SUPPLIER

id	name
15	ECUACCESORIOS S.A.
98	SYNERGY TEAM CIA. LTDA.
126	ECUATRAN S.A.

(b)

WINS

supplier_id	contract_id
15	1
98	2
126	3

(c)

Figura 3: Ejemplo de las tablas de transición que se generaron como paso previo a la construcción de la base de datos de tipo grafo (en inglés)

Por ejemplo, en la Figura 3 tenemos una muestra de contratos en la tabla (a), y una muestra de compañías en la tabla (b). Estas tablas contienen la información vital de cada uno de los nodos de tipo Contrato y Proveedor. Pero, resulta que, en la vida real, la empresa Ecuaccesorios S.A. fue quien ganó la adjudicación del contrato con ID 1. Por ello, en la tabla (c) se muestra la información vital de cada uno de los nodos de tipo Contrato y Proveedor. Por ello, en la tabla (c) se crea este vínculo Contrato-Proveedor usando solamente los Ids asignados en las tablas previas. En ese sentido, esta última tabla de la Figura 3 es la que contiene la información de las aristas que se van a formar entre los nodos que están relacionados por medio de un vínculo de tipo "Ganador en".

Este proceso se llevó a cabo para cada uno de los nodos y aristas que están presentes en el modelo del grafo (véase la Figura 2). Como resultado, esta etapa de preprocesamiento produjo todos los archivos de datos que consiguientemente sirvieron para levantar el grafo para SIE.

3.4. Almacenamiento de datos

3.4.1. Sistema de gestión de bases de datos de tipo grafo

De hecho, para almacenar toda esta información en una estructura de grafo, tuvimos que exportar los ficheros de datos a un sistema de gestión de bases de datos de tipo grafo. Pero, para ello, tuvimos que primeramente escoger un software entre las varias tecnologías de almacenamiento no relacional (también conocidas como NoSQL) que actualmente se ofrecen en el mercado. Neo4j, AllegroGraph y OrientDB fueron algunas de las opciones que encontramos entre estas tecnologías NoSQL (Vukotic & Watt, 2014), y a la final decidimos trabajar con Neo4j.

Esto se debe a que, aunque las tres opciones mencionadas cumplen con ser bases de datos que almacenan la información en términos de nodos, aristas y atributos, Neo4j tiene una

característica que le pone por encima de todas las tecnologías NoSQL: es una base de datos compatible con ACID (Neo4j, s.f). Esto significa que la presencia de cuatro propiedades – atomicidad, consistencia, aislamiento y durabilidad – garantiza un comportamiento adecuado, oportuno, seguro y real en las bases de datos Neo4j al momento de hacer transacciones (Vukotic & Watt, 2014). Esto es particularmente importante cuando se está buscando adoptar nuevas tecnologías de almacenamiento en un contexto de la vida real como el de compras públicas, en donde es importante que la base de datos se comporte de manera confiable en todo momento y bajo cualquier situación. Neo4j es la única tecnología NoSQL que ofrece estas garantías, por lo que es la herramienta que se escogió para este proyecto.

3.4.2. Construcción de grafo en Neo4j

Como tecnología NoSQL, Neo4j no utiliza SQL como lenguaje de consulta predeterminado. De hecho, esta herramienta utiliza Cypher, un lenguaje de consulta declarativo para grafos. Por ello, como parte de este trabajo de investigación, hubo que familiarse con la sintaxis de este lenguaje, y los primeros resultados de este aprendizaje se expresaron en la forma de comandos que se utilizaron para importar las tablas de datos a Neo4j, véase la Figura 4.

(a)

```
LOAD CSV WITH HEADERS FROM
"file:///table_supplier.csv" AS csvLine

CREATE (supplier:Supplier
       {id: toInteger(csvLine.supplier_id),
        nombre: csvLine.supplier_name})
```

(b)

```
:auto
 LOAD CSV WITH HEADERS FROM
“file:///table_winners.csv” AS csvLine

CALL {
 WITH csvLine
 MATCH (contract:Contract
       {id: toInteger(csvLine.contract_id)}),

       (supplier:Supplier
       {id: toInteger(csvLine.supplier_id)})
 CREATE (supplier)-[:WINNER_IN
 {contract: csvLine.contract_id}]->(contract)
}

IN TRANSACTIONS OF 5000 ROWS
```

Figura 4: Muestra de los comandos Cypher que se utilizaron para construir los nodos y las atistas de base de datos de Neo4j (en inglés)

En el primer bloque de código de esta figura, tenemos una instrucción que se usó para importar la data que estaba contenida en la tabla "table_supplier.csv". En ese fichero CSV estaban almacenados todos los Ids y nombres de los Proveedores, tal cual se mostró anteriormente en la tabla (b) de la Figura 3. Y, para cargar estos datos al grafo de Neo4j, se utilizó el comando LOAD CSV para importar el archivo, y el comando CREATE para formar los nodos de tipo Proveedor, con sus respectivas propiedades. Al correr esta línea de código en el Neo4j Browser (un shell interactiva para comandos Cypher), el programa añadió 14.706 etiquetas, creó 14.706 nodos, y estableció 29.412 propiedades. Esto es consistente con las cifras reportadas en la Figura 1, en donde se registró la presencia de 14.706 proveedores únicos. Este proceso se replicó para el resto de los nodos del modelo del grafo, por lo que el esqueleto del comando (a) de la Figura 4 sirvió para armar todos los nodos del grafo.

Ahora, para formar las aristas de la base de datos, se usó un comando similar al que aparece en el segundo bloque de código de la Figura 4. En él, se asume que el grafo ya contiene nodos de tipo Proveedor y Contrato, por lo que procede a importar la tabla relacional que conecta a los contratos con el proveedor que ganó el proceso de Subasta Inversa Electrónica. Para ello, se utiliza nuevamente el comando LOAD CSV, y la tabla importada se emplea dentro de un comando MATCH. En él, se hace una conexión entre dos nodos – un Contrato y un Proveedor – en función de los Ids que aparecen emparejados en cada fila del dato importado. Y, luego, con un comando CREATE, se forma una relación de tipo "Ganador en" entre los nodos en cuestión. La línea "IN TRANSACTIONS OF 5000 ROWS" se utiliza al final de este bloque de código con el fin de limitar el número de filas que Neo4j debe procesar a la vez antes de consignar los cambios en la base de datos. En caso contrario, el Neo4j Browser no acepta la consulta para evitar *timeouts* que pueden producirse al ejecutar transacciones de larga duración.

3.5. Flujo de datos

Los pasos descritos en las secciones anteriores fueron incluidos dentro de un *pipeline* que automatiza la creación del grafo para procesos de Subasta Inversa Electrónica.

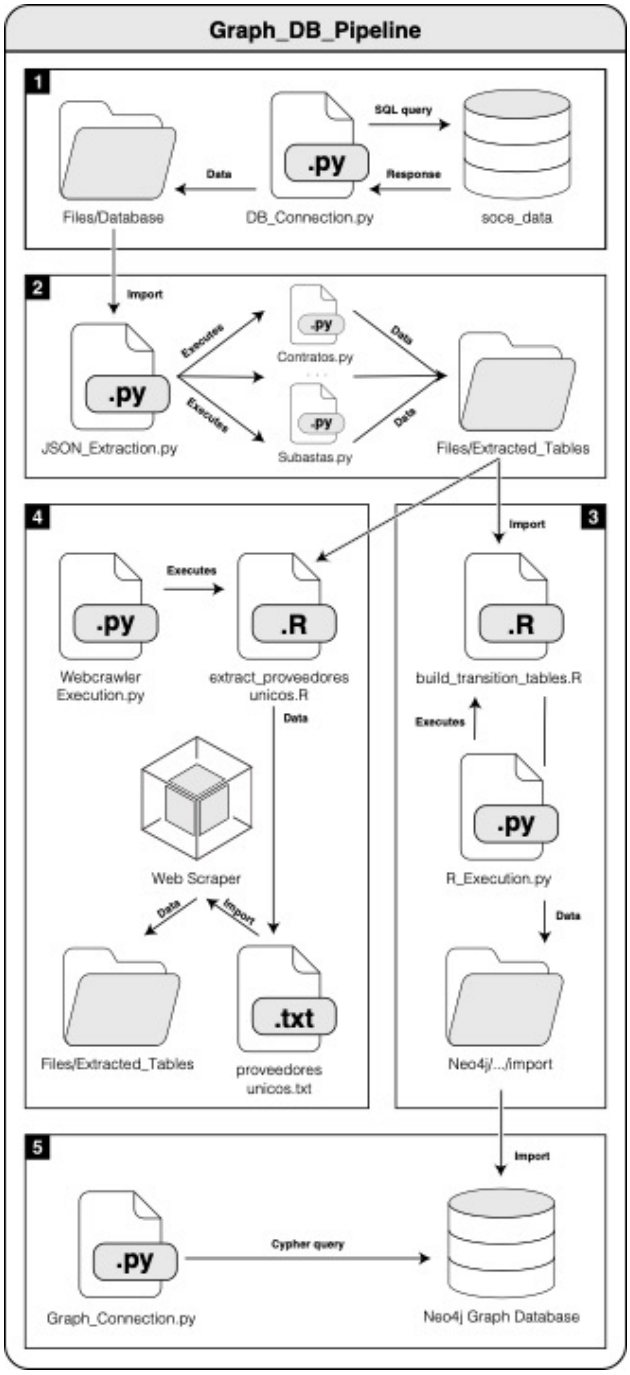

Figura 5: Cadena de procesos que se ejecutan en secuencia para construir automáticamente y desde cero la base de datos en Neo4j (en inglés)

Como se muestra en la Figura 5, el Graph_DB_Pipeline es una serie de 5 procesos principales a través de los cuales los datos del proyecto fluyen desde sus fuentes – desde la base de datos relacional que contiene información del portal SOCE, y desde el portal de la Superintendencia de Compañías – hasta su destino, una base de datos local de Neo4j.

Para ello, en las primeras etapas de este pipeline, un script de Python ejecuta una consulta SQL dentro de la base de datos del proyecto (también conocida como 'soce_data') para extraer filas de datos que almacenan contratos de tipo SIE en donde se ha llevado a cabo una subasta o una negociación. Los resultados de este query se guardan en formato CSV en una carpeta local. Esta información avanza hacia la siguiente etapa, en donde otro script de Python ejecuta una serie de códigos que desagregan los JSONs que están contenidos en las columnas del archivo descargado, y organizan la información desagregada en la forma de tablas CSV. Los resultados de esta etapa se exportan a otra carpeta local, cuyos contenidos se importan en los procesos 3 y 4 de este pipeline. Por un lado, en el tercer paso de este flujo, un script de Python ejecuta un script basado en R que construye las tablas con los Ids y las propiedades de todos los nodos del grafo, más las tablas de transición que mapean las relaciones que hay entre dichos nodos. Los resultados de este proceso son almacenados dentro de la ubicación por defecto de la carpeta de tipo *import* de la base de datos de Neo4j. Sin embargo, estos datos no están completos ya que faltan los nodos de tipo Accionista y las relaciones que los conectan con las compañías a las que están vinculados. Por ello, de forma asíncrona, el proceso 4 se encarga de extraer una lista de proveedores únicos de las tablas extraídas en el paso 2. Esta lista ingresa como input al web scraper que, desde el portal de la Superintendencia, extrae la información de los accionistas de las compañías enlistadas. La data descargada es entonces exportada a un archivo CSV que se coloca dentro de la carpeta local de tablas extraídas. Y, a través de la lista de pasos que se describieron para el proceso 3, se logra transformar estos datos en tablas de transición para posibilitar la adición de los nodos Accionista al grafo. Para hacer esto, en la fase final de este

pipeline, se corre un Python script que usa la librería ‘neo4j’ para interactuar con la base de datos de Neo4j sin hacer uso del *shell* de la aplicación, es decir, del Neo4j Browser. En particular, se utiliza un driver para ejecutar transacciones basadas en Cypher en el servidor, y crear nodos y edges a partir de los archivos CSV que han sido transferidos a la carpeta de importación de la base de datos.

De esa manera, se ha logrado construir un grafo Neo4j para contratos de Subasta Inversa Electrónica a través de un solo script que orquestra todos los procesos que protagonizan el *pipeline*. Esto, como tal, es clave para posibilitar el crecimiento del grafo ya que, hasta la fecha, en el Ecuador siguen vigentes los procesos de compra pública de este tipo, y el SOCE continúa publicando su información en el portal. Esto significa que la información del la base de datos SQL, ‘soce_data’, está en aumento, y el grafo necesita crecer para reflejar las adiciones que se realizan a sus fuentes de datos.

4. RESULTADOS Y DISCUSION

Por ahora, un grafo Neo4j totalmente conectado almacena datos relativos a los actores – entidades, proveedores y accionistas – que, entre 2008-03-24 y 2022-09-15, participaron de forma directa o indirecta en contratos de Subasta Inversa Electrónica para intercambiar bienes y servicios de diversas categorías. De hecho, la base de datos resultante contiene aproximadamente 367 mil nodos y 787 mil relaciones. Las tablas (a) y (b) de la Figura 6 muestran en detalle la composición de este grafo.

NODOS

	Bid	Contract	Stakeholder	Supplier	Entity	CPC
#	305.366	27.719	15.950	14.706	2.107	922
%	83,25	7,56	4,35	4,01	0,57	0,25

(a)

RELACIONES

	Puts	There_is	Participates	Belongs	Contains	Wins	Awards
#	305.366	305.366	58.983	33.722	27.719	27.719	27.719
%	38,82	38,82	7,50	4,29	3,52	3,52	3,52

(b)

Figura 6: Distribución por número y porcentaje de los nodos y relaciones que componen el grafo de Neo4j (en inglés)

En estas tablas, se observa que el 83% de los nodos y el 77% de los edges almacenan información sobre pujas que se realizaron como parte de los procesos de subasta que se llevaron a cabo para adjudicar cada uno de los 27.719 contratos que hay en la red. Recorrer todos estos datos en un esquema relacional resultaría computacionalmente costoso porque se sabe por hecho que las operaciones de tipo *join* calculan el producto cartesiano antes de descartar los resultados irrelevantes (Robison, Webber & Eifrem, 2015). De manera que, al momento de manejar datasets voluminosos en una tecnología SQL, el rendimiento de la base de de datos se vería afectado de forma polinomial, por lo menos en un segundo grado (Vukotic & Watt, 2014). Mientras que, como una base de datos de grafos está equipada con adyacencia sin índices (*index-free adjacency* en inglés), se puede navegar rápidamente a través de las relaciones que hay en los datos, independientemente del tamaño total del dataset (Robison, Webber & Eifrem,

2015). Por ese lado, los beneficios relacionados al rendimiento de esta tecnología se aprovecharán al máximo al analizar información sobre pujas en el grafo resultante.

Mientras tanto, desde ya podemos sacarle provecho al grafo para hacer detección visual de patrones en los nodos y edges que lo componen. A modo de ejemplo, se usó la información almacenada en la base de datos de Neo4j para computar un indicador de concentración de mercado a nivel de entidad (OECD, 2018):

$$PD_i = 100 - \left[\left(\sum_{j=1}^{n} PU_j\right) * \frac{100}{n}\right]$$

PD_i: Número de proveedores que concentran mercado en los últimos n contratos adjudicados por la unidad de observación i (entidad contratante).

PU_j: Proveedor que aparece una única vez entre quienes han sido adjudicatarios de los últimos *n* contratos de la unidad de observación i (entidad contratante).

Ecuación 1: Indicador de concentración de mercado a nivel de entidad

El indicador de la Ecuación 1 computa el porcentaje de proveedores que consiguieron la adjudicación de 2 o más contratos por parte de una entidad. Para poder calcularlo, se corrió el Cypher query (a) de la Figura 7 para obtener el total de contratos que adjudicó cada una de las entidades del grafo (la variable *n* de la Ecuación 1).

(a)

```
MATCH (e:Entity)-[:AWARDS]->(c:Contract)

WITH  e, collect(c.id) AS contracts_ids,
      count{(e)-[:AWARDS]-(:Contract)}
      AS total_awarded_contracts

ORDER BY total_awarded_contracts DESC
RETURN e.name, total_awarded_contracts;
```

(b)

```
MATCH (e:Entity)-[:AWARDS]->
      (c:Contract)<-[:WINS]-(s:Supplier)

WITH e, s, collect(c.id) AS contracts_ids,
     count{(e)-[:AWARDS]-(:Contract)-
     [:WINS]-(s)} AS num_awarded_contracts

ORDER BY e.name
WHERE num_awarded_contracts = 1
RETURN e.name, num_awarded_contracts;
```

Figura 7: Comandos Cypher que corrieron sobre la base de datos de Neo4j para extraer las variables necesarias para computar el indicador de la Ecuación 1 (en inglés)

Y, por otro lado, se usó el query (b) de dicha figura para contabilizar el número de contratos que cada una de las entidades adjudicó a los proveedores que estuvieron en el extremo receptor de sus adjudicaciones. Mediante este query, se retuvieron únicamente las entradas en donde la relación Entidad-Proveedor aparece una única vez, pero no se pudo contabilizar el número de instancias retenidas por cada entidad. En respuesta a esto, los resultados de cada uno de estos comandos fueron exportados en tablas CSV, e importados dentro de un R script que, en base a la tabla obtenida con (b), computó, por cada entidad, el total de proveedores que aparecieron una única vez entre quienes fueron adjudicatarios de sus *n* contratos. De esa manera, se obtuvo el numerador de la Ecuación 1 y, con esto, se pudo finalmente hacer el cálculo para obtener una medida de cuánto mercado concentran las entidades del grafo (siendo 100% el puntaje que indica la presencia de un actor que ha adjudicado todos sus contratos a un solo proveedor).

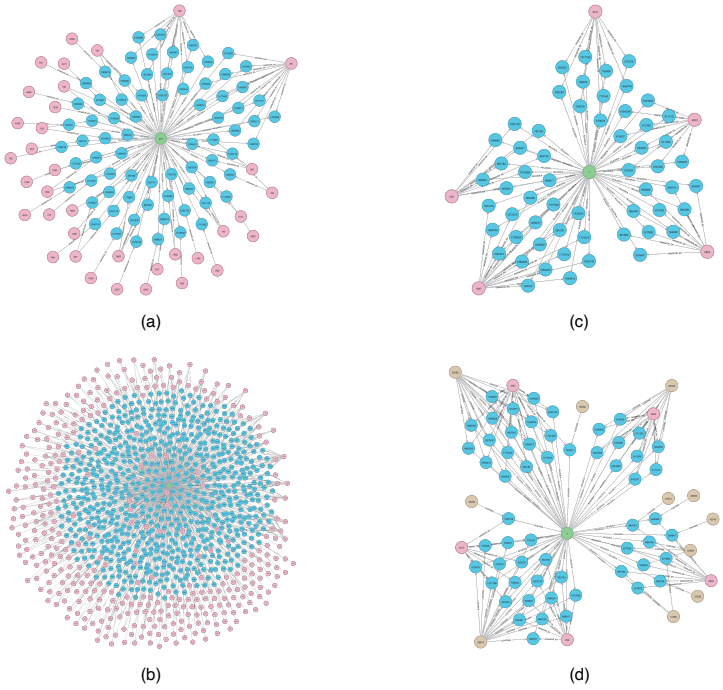

Figura 8: Grafos que muestran los contratos (nodos azules) adjudicados por las entidades A y B (nodos verdes), los proveedores (nodos rosa) que han ganado esos contratos, y los mercados (nodos cafés) que agrupan a los contratos adjudicados.

Las capturas (a) y (b) de la Figura 8 muestran cómo se ven los grafos de dos entidades- A y B- que obtuvieron un resultado de 68,00% y 67,63% en el indicador calculado. En este caso, de los porcentajes obtenidos en paso anterior, se seleccionaron dos entidades que obtuvieron puntajes altos y similares, pese a tener diferentes tamaños (siendo A un nodo que adjudicó 75 contratos, y B un nodo que adjudicó 726 contratos en la red de SIE). Los grafos 8.a y 8.b muestran estas diferencias en tamaño. En particular, los contratos adjudicados por cada una de las entidades se muestran como puntos azules que rodean al nodo verde que representa al actor contratante, mientras que los proveedores que consiguieron la adjudicación de esos contratos se visualizan como puntos rosados. Y, a pesar de las diferencias en tamaño, la concentración de mercado está presente en la forma de un patrón visual que se localiza en el extremo superior derecho de cada uno de los grafos. En él, se observa que el área circular de las figuras se empieza a hundir debido a que en esta región hay menos proveedores en la circunferencia del grafo, y más relaciones de tipo "Ganador en" vinculadas a los pocos nodos rosados que están ahí. En ese sentido, es posible detectar la presencia de indicios de concentración de mercado en el grafo de SIE a partir de la forma en que se visualizan las relaciones Entidad-Contrato-Ganador en el grafo.

Y, sin un script que haga cálculos, también se puede localizar a los proveedores que protagonizan dicha concentración de mercado. Esto está explícito en el grafo (a) de la Figura 8, en donde es muy obvio que dos nodos rosados fueron ganadores en más de tres procesos de subasta que fueron conducidos por la entidad A. Sin embargo, esta misma tarea- la de detectar manualmente a los proveedores que más mercado concentraron – se complica en el grafo (b) de la misma figura debido a su tamaño. Por suerte, se puede alterar el query de Cypher que produjo la visualización de dicho grafo para que muestre solamente los nodos asociados a los proveedores que recibieron más de 8 contratos por parte de la entidad B. Esta alteración genera

el grafo (c) de la Figura 8, en donde están destacadas los top 5 proveedores que más ganaron en los procesos de subasta que fueron conducidos por la entidad B.

La interacción con estos grafos en el Neo4j Browser permite visualizar en tiempo real los atributos asociados a sus nodos y sus relaciones. Esto es particularmente útil cuando se desea buscar más patrones de forma manual en los datos de la red. A manera de ejemplo, si volvemos a alterar el query del grafo 8.c para que agrupe los contratos según su CPC (i.e. categoría de los bienes/servicios), obtenemos el grafo (d) de la Figura 8. En él, observamos la adición de nodos cafés al grafo anterior y de relaciones de tipo CPC-Contrato. Como resultado de esto, podemos observar que dos proveedores comparten el mercado que está representado como nodo café en la parte inferior izquierda del grafo. También, se ve que los dos proveedores que están en la parte superior del grafo concentran cada uno un mercado, mientras que el que está en la región inferior derecha parece haber recibido contratos para bienes/servicios de diversas categorías. De usar el Neo4j Browser para interactuar con estos nuevos nodos y visualizar sus propiedades, se podría localizar más anomalías en el grafo si es que, por ejemplo, las descripciones de los CPCs que están vinculados a los proveedores no corresponden con sus razones sociales.

Lo que sí, la herramienta de Neo4j tiene una limitación que no permite enriquecer el análisis que se está realizando. En este caso, resulta que no se puede dibujar los nodos en proporción a uno de sus atributos. Esto, como tal, impide que visualicemos el tamaño de cada uno de los contratos en función a los montos adjudicados. De ser esto posible, no solo podríamos detectar patrones de concentración de mercado en base al número de contratos que adjudicó una entidad a un proveedor, sino, también en base a la cantidad de dinero que se transfirió entre estos actores del grafo. Esto, como tal, hubiera ayudado a resaltar anomalías escondidas en los atributos de los nodos de la base de datos.

De todas maneras, la presencia de esta limitación no impidió que se pueda realizar un análisis de concentración de mercado en los datos del grafo de Neo4j. Y, por ese lado, en esta sección se pudo demostrar las capacidades que ofrece esta herramienta al momento de hacer detección semi-automática de patrones de riesgos de corrupción en los nodos y relaciones que conforman la red de Subasta Inversa Electrónica. De hecho, el análisis realizado ayudó a probar que, en ausencia de algoritmos que recorren la data en busca de indicios específicos de corrupción, se puede usar el grafo y Cypher queries para buscar anomalías que todavía no se pueden localizar de forma automática.

5. CONCLUSIONES

El presente trabajo de investigación alcanzó su objetivo, el de construir un grafo para procesos de Subasta Inversa Electrónica, a través de un flujo de pasos que consiguió transformar y complementar los datos que fueron originalmente almacenados en tecnología SQL para fines del proyecto mencionado (Fortuny, Guerrero, Riofrío & Simon, 2023). Y, aunque esta investigación no propone ninguna metodología para calcular anomalías ni indicios de riesgos de corrupción en el grafo Neo4j, en la sección Resultados y Discusiones se demostró que, con la ayuda de consultas Cypher, se pueden localizar subgrafos para detectar, manualmente, patrones que son explícitos en la forma en que se visualizan sus nodos y relaciones, o que subyacen en los atributos de sus componentes. En este sentido, ya podemos empezar a utilizar la base de datos de tipo grafo para identificar semiautomáticamente riesgos de corrupción en los contratos SIE que la componen. Esto significa que, en manos de expertos conocedores de la contratación pública en Ecuador, esta herramienta puede ser utilizada para agilizar el proceso de detección visual de comportamientos anómalos o anticompetitivos en los datos.

Agilizar esto en un esquema relacional no hubiera sido posible, ni conveniente, porque la Subasta Inversa Electrónica, al igual que las redes sociales, es un dominio centrado en las relaciones. De manera que, cuando “intentamos encajar los datos en tablas y columnas relacionales, y normalizar y renormalizar su estructura, hacemos que la información parezca completamente distinta de lo que intenta representar” (Robison, Webber & Eifrem, 2015). Y, cuando esto ocurre, encontrar patrones y anomalías en los datos pasa a depender de algoritmos que realizan automáticamente operaciones entre tablas para revelar los vínculos que están implícitos en la base de datos. Afortunadamente, utilizando una base de datos de Neo4j, se ha logrado preservar la estructura gráfica natural de los datos SIE, es decir, las relaciones entre los nodos que forman parte de la información allí almacenada.

Esto, a su vez, no sólo facilita enormemente la visualización y localización manual de posibles indicios de riesgos de corrupción en los datos, sino que también mejora significativamente el rendimiento de las consultas que atraviesan el grafo en busca de información que en una base de datos SQL se buscaría mediante uniones entre tablas. Esta es una de las ventajas reales de utilizar una tecnología de grafos no SQL para modelar la información de la Subasta Inversa Electrónica. Y, como se mencionó anteriormente en la discusión de los resultados, esta ventaja comenzará a manifestarse una vez que comencemos a utilizar la herramienta para analizar los datos de las pujas, que constituyen el 83% de los nodos y el 77% de las aristas del grafo resultante. Para ello, en futuros trabajos, deberíamos investigar formas de utilizar técnicas de detección de anomalías basadas en grafos (GBAD), o heurísticas de búsqueda en grafos para identificar o medir, de forma automática y premeditada, comportamientos anómalos o anticompetitivos en la base de datos que se ha recopilado para los fines de este trabajo de investigación. En ese sentido, en el futuro, debemos ahondar en el campo de la ciencia de redes (*network science* en inglés) para construir indicadores de riesgo de corrupción basados en grafos para los procesos de contratación pública que se llevan a cabo en la modalidad de Subasta Inversa Electrónica.

Hasta entonces, debemos asumir el reto de mantener actualizada la base de datos de grafos. Esto debido a que el webcrawler del SOCE aún no está al día con la descarga de los procesos históricos de SIE, y en el futuro descargará los datos de las contrataciones que, bajo esta modalidad, se publiquen y adjudiquen en el marco de la contratación pública ecuatoriana. En respuesta a esto, es importante adaptar el pipeline de la Figura 6 para que pueda realizar actualizaciones sobre la base de datos de tipo grafo sin corromper la integridad de los datos ya almacenados en ella, y de los datos que se irán agregando a partir de las actualizaciones realizadas sobre la base de datos SQL original. En este sentido, para permitir el crecimiento del grafo a lo largo del tiempo, la estrategia de actualización debe superar los retos de evitar la

duplicación de información y garantizar la coherencia al generar nuevas conexiones entre los nodos existentes.

Con esto, la herramienta que ha sido creada en este trabajo de investigación no sólo será útil para responsabilizar a las entidades, proveedores y actores que exhibieron comportamientos potencialmente corruptos y anticompetitivos en el pasado. Sino que, si se utiliza de forma oportuna, este grafo en Neo4j permitirá vigilar, de forma más visual, las relaciones que se van formando entre los actores, contratos y mercados que componen la cada vez más extensa red de la Subasta Inversa Electrónica.

Y, para concluir, el presente trabajo es un resultado de haber aplicado el ciclo de diseño a un problema de Ciencias de la Computación (Levenon & Kumalesh, 2017). En él, se identificó que la información sobre Subasta Inversa Electrónica no está almacenada en el esquema más óptimo y que, en investigaciones pasadas, tampoco se ha hecho mucho para representar datos de contratación pública en tecnologías no SQL. Ante este diagnóstico, surgió la necesidad de identificar los nodos y relaciones que están implícitos en las tablas que guardan los datos sobre procesos SIE, y componer un grafo en Neo4j para almacenar esta información dentro del nuevo modelo de datos. Este trabajo de investigación describe la metodología que se aplicó para realizar esta transformación y, por ende, para solventar el problema identificado. A largo plazo, se espera hacer uso de la base de datos resultante en el campo de detección de anomalías y fraude usando Inteligencia Artificial. Este, como tal, es el caso de uso #20 del estándar ISO/IEC TR 24030 *Information technology — Artificial intelligence (AI) — Use case.* Por ese lado, la investigación realizada se enmarca dentro de los estándares para inteligencia artificial (ISO/IEC, 2021).

REFERENCIAS BIBLIOGRÁFICAS

Aarvik, P. (2019). *Artificial Intelligence – a promising anti-corruption tool in development settings?* Obtenido el 18 de agosto 2009 de https://www.u4.no/publications/artificial-intelligence-a-promising-anti-corruption-tool-in-development-settings

Fortuny, M., Guerrero, E., Riofrío, D., & Simon, F. (2023). Towards Smart Citizen Control in Public Procurement: Ecuador's Case Study. *2023 Ninth International Conference on Edemocracy & Egovernment (ICEDEG)*, 1–6. Doi:10.1109/ICEDEG58167.2023.10121991

International Center for Journalists. (2020). *Rastrear empresas en las Américas*. Obtenido el 6 de febrero de 2023 de https://labmedia.org/wp-content/uploads/docs/ES/Manual%20para%20rastrear%20empresas%20en%20las%20Ame%Efricas.pdf

ISO/IEC. (2021). ISO/IEC TR 24030: Information technology — Artificial intelligence (AI) — Use cases. Obtenido el 19 de mayo de 2023 de https://cdn.standards.iteh.ai/samples/77610/ebcb38166ad94fb0963e2ade7ea35f83/ISO-IEC-TR-24030-2021.pdf

Kawai, K., & Nakabayashi, J. (2014). Detecting Large-Scale Collusion in Procurement Auctions. *Journal of Political Economy*, *130*(5), 1364–1411. doi:10.1086/718913

Levenon, K., & Kumalesh, M. (2017). Design Process in Computer Science. *Available at SSRN 2998444*.

Neo4j. (n.d.). *What is a Graph Database?* Obtenido el 26 de marzo de 2023 de https://neo4j.com/developer/graph-database/

OECD. (2016). Report on implementing the OECD Recommendation on Fighting Bid Rigging in Public Procurement. Obtenido el 6 de febrero de 2023 de https://www.oecd.org/gov/public-procurement/recommendation/

(2018). *Market Concentration: Issues paper by the Secretariat*. Obtenido el 26 de marzo de 2023 de https://one.oecd.org/document/DAF/COMP/WD(2018)46/en/pdf

Porter, R. H., & Zona, J. D. (1993). Detection of Bid Rigging in Procurement Auctions. *Journal of Political Economy*, *101*(3), 518–538. Obtenido el 6 de febrero de 2023 de http://www.jstor.org/stable/2138774

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, *133*, 113303. doi:10.1016/j.dss.2020.113303

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph Databases*. Sebastopol: O'Reilly Media, Inc.

Sampford, C., Shacklock, A., Connors, C., & Galtung, F. (2006). Measuring Corruption. Http://Lst-Iiep. Iiep-Unesco. Org/Cgi-Bin/Wwwi32. Exe/[In=epidoc1. in]/?T2000=028634/(100).

Servicio Nacional de Contratación Pública. (2023). *Manual de Usuario: Subasta Inversa Electrónica*. Obtenido el 6 de febrero de 2023 de https://portal.compraspublicas.gob.ec/sercop/wp-content/uploads/downloads/2020/07/MANUAL-DE-USUARIO-SOCE-SUBASTA-INVERSA-ENTIDADES-comprimido.pdf

(2020). *Rendición de Cuentas en la Contratación Pública*. Obtenido el 6 de febrero de 2023 de https://bit.ly/3BTY98S

Sun, T., & Sales, L. J. (2018). Predicting Public Procurement Irregularity: An Application of Neural Networks. *Journal of Emerging Technologies in Accounting*, *15*(1), 141–154. doi:10.2308/jeta-52086

United Nations Office on Drugs and Crime. (2020). *Addressing Corruption Risks in Procurement*. Obtenido el 6 de febrero de 2023 de https://www.unodc.org/roseap//en/what-we-do/anti-corruption/topics/2020/corruption-risks-procurement.html

Vukotic, A., & Watt, N. (2014). *Neo4j in Action*. New York: Manning Publications.