Modelamiento y análisis de datos de contratación estatal en Colombia a través de grafos

Jesid Mauricio Mejía Castro

Maestría en Ingeniería y Analítica de Datos

Facultad de Ciencias Naturales e Ingeniería
Universidad de Bogotá Jorge Tadeo Lozano
Bogotá, Colombia
Octubre 19 de 2021

Índice

1. Introducción

La contratación pública en Colombia es uno de los mecanismos más importantes a través de los cuales el Estado colombiano adquiere obras de infraestructura, servicios y consultorías con el fin de atender las necesidades de las instituciones públicas [1].

Una de las razones que motiva esta investigación es la atención especial que la ciudadanía ha puesto en los procesos de contratación pública. Buena parte de este interés proviene de la percepción del aumento de estos fenómenos en los ámbitos local y nacional. Según Betancourt [2], la corrupción rampante en este tipo de contratos obedece al hecho de que no se han tomado medidas suficientes en la legislación, pero también a una deficiente administración del riesgo en los procesos de contratación.

Para tratar de mitigar este flagelo, al menos desde una perspectiva académica que permita señalar el camino a proyectos posteriores, se propone la utilización de una base de datos basada en grafos para modelar los datos provenientes de este campo.

Los grafos son objetos matemáticos compuestos de *nodos* y *arcos*. Aunque suelen presentarse con cierto nivel de abstracción, estos objetos tienen mucha utilidad al momento de modelar y analizar datos [3]. Los algoritmos de grafos son un subconjunto de las herramientas utilizadas en la analítica de grafos. Allí se cuenta con varias alternativas: consultar los datos del grafo, utilizar estadística básica, explorar visualmente el grafo o incorporar grafos en tareas de aprendizaje automático. Se escoge este tipo de modelo de datos debido al potencial que tiene para exhibir comportamientos no evidentes.

Siguiendo de carca el trabajo de Fernandes [4], encontramos que los datos modelados con grafos tienen una ventaja con respecto a aquellos almacenados bajo el modelo relacional. En particular, resulta altamente conveniente cuando los conjuntos de datos exhiben alta escalabilidad y las relaciones entre los datos es compleja de manera que se requiere mayor flexibilidad al momento de analizar dichas relaciones. Este tipo de tecnología ha resultado ser disruptiva en múltiples áreas como la gestión de cadenas de abastecimiento, sistemas de recomendación para comercio en línea, seguridad, detección de fraudes, entre otras.

Para este proyecto se utilizarán los datos abiertos proporcionados por el Gobierno de Colombia a través de `datos.gov.co` y el Sistema Electrónico de Contratación Pública (SECOP). En particular, se tendrán en cuenta los siguientes conjuntos de datos:

- Proveedores registrados: información básica de los proveedores registrados en SECOP II.
- Contratos electrónicos: información de los contratos registrados en SECOP II desde su lanzamiento.
- Procesos de contratación: registro de los procesos de compra, sean o no adjudicados, hechos en la plataforma SECOP II desde su lanzamiento.
- Adiciones: adiciones hechas a los contratos firmados en la plataforma SECOP II.

Dadas las ventajas que ofrece un modelo de grafos para detectar relaciones ocultas en los datos se utilizará una base de datos de este tipo. Como resultado, se desea obtener un modelo de datos que permita realizar consultas con alta profundidad y detecte de manera eficaz patrones de corrupción y fraude.

2. Marco teórico

2.1. Conceptos de contratación pública

Para conformar un bosquejo sobre los procesos de contratación estatal, se seguirá de cerca el trabajo de Angarita [1]. En un contrato público intervienen dos partes: en un lado está el *oferente* o *contratista*; mientras que su contraparte se denomina *entidad pública* o *contratante*. El primero ofrece servicios o bienes a cambio de una remuneración económica, mientras que el segundo establece las reglas bajo las cuales se determinará la relación teniendo en cuenta que lo público tiene prioridad sobre lo privado.

Desde el siglo XVIII, prácticamente con el nacimiento de la república, se viene regulando la contratación estatal en Colombia. Desde ese entonces era clara la idea de proteger el patrimonio público sin desconocer el derecho que tienen los particulares a una justa retribución.

Gran parte de los conceptos que conforman la contratación pública provienen de un ámbito normativo. En particular, la Ley 80 de 1993 [5] define los fines de la contratación estatal de la siguiente manera:

> ARTÍCULO 3o. Los servidores públicos tendrán en consideración que al celebrar contratos y con la ejecución de los mismos, las entidades buscan el cumplimiento de los fines estatales, la continua y eficiente prestación de los servicios públicos y la efectividad de los derechos e intereses de los administrados que colaboran con ellas en la consecución de dichos fines.
>
> Los particulares, por su parte, tendrán en cuenta al celebrar y ejecutar contratos con las entidades estatales que, además de la obtención de utilidades cuya protección garantiza el Estado, colaboran con ellas en el logro de sus fines y cumplen una función social que, como tal, implica obligaciones.

Con el fin de hacer más transparente la política pública, el país adoptó programas como el Sistema de Información para la Contratación Estatal (SICE) y el SECOP.

2.2. Sistema Electrónico de Contratación Pública (SECOP II)

El SECOP II es un sistema transaccional donde se registran los proveedores y las instituciones del estado. Las cuentas asociadas a las Entidades Estatales pueden crear, evalúar y adjudicar Procesos de Contratación. Los Porveedores pueden presentar ofertas y seguir el proceso de selección en tiempo real. En la introducción se listaron los conjuntos de datos a utilizar en este proyecto. En el Cuadro 1 se reseñan algunas de las características básicas de estos datos.

Conjunto de datos	Número de columnas	Número de filas	Fecha de creación	Fecha última actualización	Detalle de columnas
Proveedores Registrados	11	678K	30 de septiembre de 2019	6 de junio de 2021	Anexos, Cuadro 4
Contratos Electrónicos	67	1.05M	30 de septiembre de 2019	4 de junio de 2021	Anexos, Cuadro 5
Procesos de Contratación	58	1.03M	30 de septiembre de 2019	4 de junio de 2021	Anexos, Cuadro 6
Adiciones	5	1.04M	30 de septiembre de 2019	6 de junio de 2021	Anexos, Cuadro 7

Cuadro 1: Conjuntos de datos

2.3. Teoría de grafos

De acuerdo a Needham [3], la historia de los grafos comienza en 1736 cuando Leonhard Euler resuelve el problema de los "Siete Puentes de Königsberg". Aquel problema preguntaba si era posible recorrer cuatro áreas de una ciudad conectadas por siete puentes si solo se cruzaba cada puente una sola vez.

Aunque los grafos tienen un origen matemático, son una forma fidedigna y práctica de modelar datos. Un grafo se compone de dos tipos de objetos: *nodos* y *arcos*. Se pueden pensar los nodos como sustantivos en una frase y se pueden imaginar los arcos como los verbos que dan contexto a los nodos. Esta idea resulta útil al momento de modelar datos a través de grafos.

Una *base de datos de grafos*, según Bechberger [6], es un motor de almacenamiento de datos que combina las estructuras básicas de grafos (nodos y arcos) con un mecanismo de persistencia y un lenguaje de consulta.

Modelar los datos es la primera parte del proceso, el procesamiento de estos permitirá revelar aquello que no es tan obvio. La analítica de grafos es el uso de algoritmos de grafos para analizar datos conectados. Existen varios métodos: consultas, estadísticas básicas, exploración visual del grafo o tareas de aprendizaje automático, véase [7].

Los grafos pueden tomar múltiples formas:

- Redes aleatorias: tienen distribuciones promedio, no tienen estructura o patrón jerárquico.
- Redes de mundo pequeño: altamente densas con longitudes de arco pequeñas.
- Redes de escala libre: redes altamente distribuidas.

En un grafo *no dirigido*, los arcos se consideran bidireccionales. En un grafo *dirigido*, los arcos poseen una dirección específica. Los arcos que apuntan a un nodo se les denomina *enlaces de entrada*, mientras que aquellos que se originan desde un nodo se denominan *enlaces de salida*.

2.4. Algoritmos de grafos

Recurriendo nuevamente a Needham [3] y Wiese [8], pueden agruparse los algoritmos de grafos en tres categorías: búsqueda de caminos, cálculo de centralidad y detección de comunidad.

2.4.1. Búsqueda de caminos

Los algoritmos de búsqueda en grafos exploran la red con el objetivo de realizar búsquedas explícitas o descubrimientos generales. Los algoritmos fundamentales para recorrer un grafo son la Búsqueda en Profundidad (*Depth First Search*) y la Búsqueda en Anchura (*Breadth First Search*); estos son a menudo utilizados como primer paso en otros tipos de análisis. En particular, resultan importantes los siguientes algoritmos de búsqueda de caminos:

- Camino más corto con dos variaciones (A* y Yen): encuentra la camino más corto entre dos nodos

- Camino más corto dos-a-dos y camino más corto desde una fuente: encuentran la ruta más corta entro todas las parejas de nodos o desde un nodo dado.
- Árbol de mínima expansión: encuentra la estructura de árbol conectada con el menor costo de nodos visitados desde un nodo dado.
- Caminata aleatoria: es un paso útil de preprocesamiento en algoritmos de aprendizaje automático y otros procedimientos.

2.4.2. Cálculo de centralidad

Los algoritmos de centralidad son utilizados para entender el rol de nodos particulares en el grafo y su impacto en la red. Son útiles para identificar nodos importantes y entender dinámicas de grupos tales como credibilidad, accesibilidad, velocidad de propagación y puentes entre grupos. Los algoritmos más importantes de este tipo son:

- Grado de centralidad: es una métrica base para evaluar *conectitud*.
- Centralidad de cercanía: mide variaciones para grupos desconectados.
- Centralidad de interposición: encuentra puntos de control, incluyendo una alternativa para la aproximación
- *PageRank*: muestra la influencia general de un nodo, incluyendo una opciones de personalización.

2.4.3. Detección de comunidad

La formación de comunidades es común en todo tipo de redes e identificarlas es esencial para evaluar el comportamiento de grupos y fenómenos emergentes. El principio general al encontrar comunidades es que sus miembros tendrán más relaciones dentro del grupo que los nodos fuera del grupo. Los algoritmos más representativos de detección de comunidades son los siguientes:

- Conteo de triángulos y coeficiente de agrupamiento: se utilizan para calcular la densidad en las relaciones
- Componentes fuertemente conectados y componentes conectados: encuentran grupos conectados
- Propagación de etiquetas: infiere rápidamente los grupos basándose en las etiquetas de los nodos.
- Modularidad de Louvain: busca jerarquías y grupos de calidad.

2.5. Bases de datos basadas en grafos

Para la siguiente parte, se ha seguido de cerca el trabajo de Fernandes [4]. Las principales bases de datos activas orientadas a grafos son las siguientes:

- **AllegroGraph**: es una base de datos RDF (*Resource Persistence Framework*) orientada a grafos muy utilizada en proyectos comerciales, *open-source* y de defensa. Está caracterizada por un uso eficiente de memoria al combinarlo con almacenamiento en disco. Algunas de sus ventajas competitivas son las siguientes:
  - Soporte de consultas *ad hoc* a través de SPARQL, PROLOG y lenguajes como JavaScript.
  - Índices quíntuples ordenados para campos primarios y no primarios.
  - Visualización de grafos a través de `Gruff`.
  - Recuperación completa y rápida.
- **ArangoDB**: es un sistema de base de datos multimodelo desarrollado por triAGENS GmbH. Los datos pueden ser almacenados como parejas de llave/valor, documentos o grafos y pueden ser todos accedidos a través de un solo lenguaje de consultas denominado AQL (*ArangoDB Query Language*). Sus más importantes ventajas competitivas son las siguientes:
  - Manejo de múltiples modelos de datos con un solo lenguaje de consultas.
  - API HTTP para gestionar bases de datos.
  - Multiarquitectura
  - Complejidad operacional reducida
  - Consistencia de datos fuerte
- **InfiniteGraph**: es una base de datos distribuida implementada en Java con su núcleo en C++ y desarrollada por Objectivity. Algunas de sus ventajas competitivas son:
  - API/Protocolos: Java (núcleo en C++)
  - modelo de grafo multipropiedad
  - respaldo en línea
  - Procesamiento multihilo
- **Neo4j**: es una base de datos de grafos *open-source* implementada en Java. Sus desarrolladores la describen como una base de datos totalmente transaccional y un motor de Java persistente donde se pueden almacenar estructuras en forma de grafos en vez de tablas. Algunas de sus características más importantes son:
  - Esquema flexible
  - Escalabilidad y confiabilidad
  - Lenguaje de consultas Cypher
  - API HTTP para gestionar la base de datos
  - Soporte de indices con Apache Lucence

- **OrientDB**: es un sistema gestor de bases de datos *open-source* y multimodelo que soporta modelos de datos en documentos, grafos, llave/valor y objetos. Algunas de sus ventajas son las siguientes:
    - Soporte de lenguaje SQL
    - Soporte de tecnologías web (HTTP, protocolo RESTful, bibliotecas JSON)
    - Distribuido con soporte de replicación multimaestro.
    - Manipulación de la base de datos a través de Java

En el estudio realizado por Fernandes [4], se resumen las características de estas bases de datos en el Cuadro 2 donde cada fila representa la característica más importante al elegir el software. En esta comparación se utiliza una escala de 0 a 4 donde el grado 4 significa que la característica ha sido bien implementada y 1 significa que la característica no ha sido bien implementada y debe mejorarse. El grado 0 significa que la característica no es soportada por el software.

	AllegroGraph	**ArangoDB**	**InfiniteGraph**	**Neo4j**	**OrientDB**
Esquema Flexible	1	3	3	4	3
Lenguaje de consulta	3	3	3	4	3
Sharding	3	3	0	0	3
Respaldo	3	2	3	4	3
Multimodelo	4	4	2	2	4
Multiarquitectura	3	4	3	4	3
Escalabilidad	3	4	3	4	3
Cloud Ready	3	3	4	4	3
Total	23	26	21	26	25

Cuadro 2: Resumen de las principales bases de datos orientadas a grafos

Para este proyecto se escoge Neo4J no solo por el análisis de Fernandes [4], sino por su facilidad de uso, su abundante documentación y su interesante lenguaje de consultas.

3. Estado del arte

Son múltiples los ejemplos de la utilización de grafos para resolver problemas en los que el contexto es complejo y la relación entre los elementos no es evidente.

La investigación de los Papeles de Pandora en 2021 [9] es considerada la “alianza periodística más grande de la historia”. Esta filtración representó varios retos en relación con el tamaño de los datos y la variedad de sus formatos (véase la Figura 1). Se trata de un esfuerzo que involucró a 600 periodistas de 150 medios provenientes de 117 países. Los 2.94 terabytes de datos filtrados se encontraban en diferentes formatos: como documentos, imágenes, correos electrónicos, hojas de cálculo, entre otros. Solo el 4 % de los datos era estructurado. En casos más complejos el ICIJ utilizó aprendizaje automático para separar ciertos tipos de documentos. Una vez estructurados los datos se utilizaron las plataformas Neo4j y Linkurious para generar visualizaciones y hacer la información accesible a búsquedas. Esto permitió a los reporteros trazar conexiones entre personas y compañías a través del mundo.

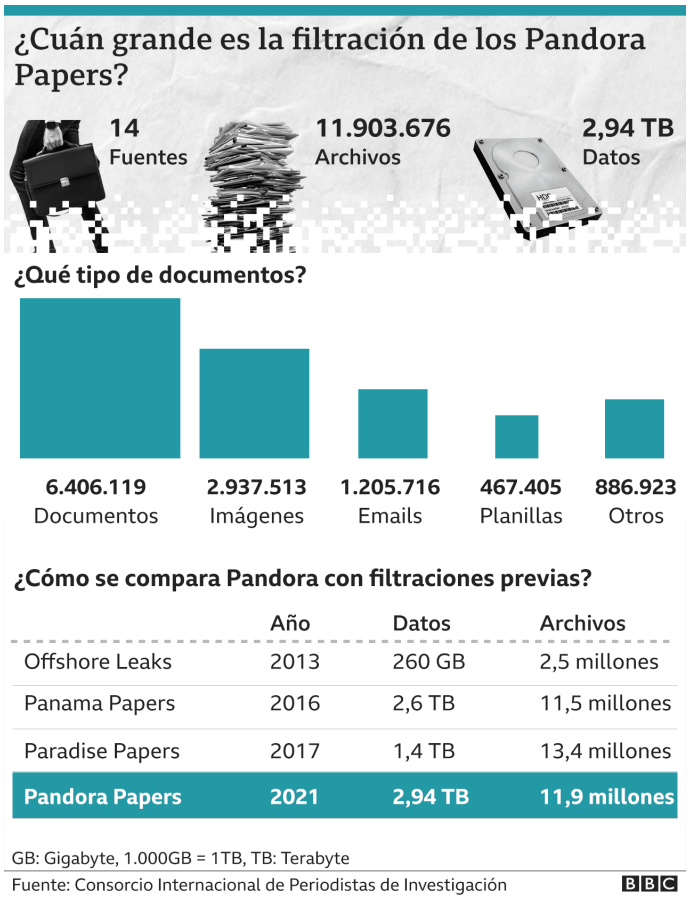

Figura 1: Los Papeles de Pandora y el tamaños de los datos. Tomado de `bbc.com` [10]

En 2016, una filtración similar había ocurrido en los llamados Papeles de Panamá. En esta ocasión, se filtraron 11.5 millones de documentos (2.6 terabytes) con información financiera de más de 2000 entidades *offshore* [11]. En primera instancia, se formó un equipo de datos que limpio y anotó los datos no estructurados. Posteriormente, se introdujeron los datos en una base de datos de grafos. Finalmente, una red de periodistas expertos generó *insights* financieros y fiscales a partir de los nodos y arcos del grafo.

En trabajos como el de Branting [12] se utilizó la analítica de grafos con el fin de estimar los fraudes de los proveedores de salud. Para ello estos investigadores se basaron en dos grupos de algoritmos. Un primer grupo calcula la similaridad en el comportamiento para separa a los proveedores fraudulentos de los no fraudulentos. Un segundo grupo de algoritmos estimó la propagación del riesgo de fraude a través de ubicación geoespacial.

Muchos de estos algoritmos han sido aplicados con éxito en el análisis de redes sociales. En trabajos como el de Naik [13] se utilizó satisfactoriamente la analítica de datos para resumir los *tweets* de manera que los usuarios puedan comprender y decidir a quien seguir dentro de la red social. Un artículo relacionado es el de Drakopoulos [14] en donde se analizan patrones de *conectividad* en *Twitter* con el fin de obtener métricas de influencia dentro de la red social.

Simperl [15] planteó en 2020 una plataforma de contratación pública para la Unión Europea basada en grafos con el fin de que las partes en el proceso pudieran tomar decisiones.

Soylu en [16] y [17] construyó también un grafo de conocimiento con datos abiertos provenientes de los datos abiertos de la Unión Europea. En el primer caso, su estrategia se basa fuertemente en la integración de datos, pues estos provienen de múltiples fuentes con diferentes formatos dependiendo del país. En el segundo caso, con mucho más detalle técnico, se agregan herramientas de *front-end* para la detección de anomalías y la búsqueda multilinguística.

En otros artículos se ha utilizado la analítica de grafos para contemplar aspectos de la contratación pública en los estados. Por ejemplo, se ha utilizado para analizar las competencias en los procesos de aprovisionamiento para insumos de la salud pública. [18]: a través de estas técnicas se trató de identificar oligopolios con el fin de que el sector público pudiera intervenir en la regulación en problemas de competición.

En la tesis de maestría de Herrera [19] se realizó un estudio basado en la premisa de que la detección de corrupción y de los riesgos de corrupción está circunscrita en el ámbito de la detección de fraudes. Este afirma que el análisis de procesos de contratación como eventos discretos es insuficiente par capturar la actividad de las redes de empresas y servidores públicos que participan en actividades ilícitas. Allí se propuso un modelo de grafos capaz de modelar las redes de contratación con tres objetivos claros: (1) representar intuitivamente las alertas que tienen una naturaleza relacional, (2) describir actores y comunidades con actividad sospechosa en las redes de contratación pública, y (3) implementar un modelo de aprendizaje automático que prediga si un contrato ha sido corrompido o no. Este trabajó concluyó que existe n potencial real en la identificación de casos de corrupción en la contratación pública.

En 2020, el trabajo de Carneiro [20] demuestra que, aunque los esquemas de corrupción se han hecho complejos, también ha progresado la tecnología para detectarlos. En este articulo se presenta un modelo para la detección de fraudes en la contratación publica del gobierno de Portugal. Además de una base de datos orientada a grafos se incluyó un motor de reglas legales para enriquecer estos. También se construyó con una interfaz gráfica de usuario para que estos tomaran decisiones de manera ágil.

En 2017, van Erven et al [21] utilizaron bases de datos basadas en grafos para recolectar evidencia de fraude en el gobierno de Brasil en procesos de contratación. Allí se sigue una aproximación similar a la de este proyecto que consiste en modelar los datos de contratación a través de Neo4j y realizar consultas con Cypher para detectar anomalías.

El trabajo de Swords [22] explica como identificar patrones en los datos de los proveedores y los procesos de contratación pública. Esta prueba de concepto nuevamente exhibe las posibilidades del uso de bases de datos orientadas a grafos, pues muestra y explora pa-

trones interesantes al momento de recuperar información utilizando análisis de centralidad y detección de comunidades. Este trabajo compara dos modelos con diferente granularidad y concluye que existe un incremento diferente en las velocidades de consulta en la medida en que aumenta el costo de almacenamiento.

Finalmente, el Cuadro 3 muestra un análisis de los trabajos mencionados, las técnicas utilizadas y los principales hallazgos.

4. Planteamiento del problema

Según Serrano [23], la corrupción es un flagelo que ha tocado a todas las civilizaciones del mundo en algún momento de su historia. Las consecuencias de la corrupción incluyen el aumento de la ineficiencia administrativa que a su vez puede incluir la baja calidad en los bienes y servicios prestados. Además, reduce el presupuesto estatal, lo que hace menos productivo el gasto público.

Para Betancourt [2], uno de los principales problemas es la lentitud al momento de identificar el fenómeno. El trabajo acá propuesto trata de brindar una alternativa en la identificación de estos fenómenos al proporcionar una herramienta con la capacidad de encontrar relaciones complejas en los datos proporcionados por el SECOP.

5. Objetivos

5.1. Objetivo general

Explorar técnicas de analítica de grafos para identificar irregularidades en datos provenientes de la contratación pública en Colombia.

5.2. Objetivos específicos

- Proporcionar una herramienta alternativa para el modelo de datos de la contratación pública que permita identificar de manera más intuitiva las prácticas corruptas o actividades inusuales a través de la analítica de grafos.
- Diseñar un esquema de base de datos basada en grafos capaz de almacenar la información procesada del SECOP.
- Construir una base de datos de grafos a partir de los datos proporcionados por el SECOP.
- Utilizar algoritmos de grafos con el fin de identificar participantes con relaciones inusuales en los contratos públicos.

6. Metodología

Este trabajo, al ser un proyecto de minería de datos, se alineará con la metodología CRISP-DM [24]. Por consiguiente, el trabajo comprenderá las siguientes fases:

- Entendimiento del negocio
  En esta etapa se estudiará con la profundidad necesaria los procesos de contratación

Trabajo	Conjunto de datos	Técnica	Resultados
Papeles de Pandora	Múltiples fuentes (2.9 TB de información)	*Keyword Search*, aprendizaje automático, entre otras técnicas.	El trabajo expone los tratos y negocios de numerosos políticos en todo el mundo (Rusia, Reino Unido, Chile, Pakistán, Filipinas, Azerbaiyán, etc.) Al igual que en 2016, pone nuevamente en el centro del debate las consideraciones morales y legales sobre la retención de grandes sumas de dinero en entidades *offshore*.
Papeles de Panamá	Múltiples fuentes (2.6 TB de información)	*PageRank*, *Betweeness centrality*, *closeness centrality*	Desencadena un debate público mundial sobre los paraísos fiscales. Varias celebridades, políticos, deportistas, entre otros personajes, son públicamente cuestionados por la gestión de su impuestos y el traslado de sus riquezas.
Naik	Twitter	Detección de comunidades	La técnica indicada permite agrupar *tweets* similares para diferentes eventos noticiosos.
Simperl	Proyecto *TheyBuyForYou*	Grafos semánticos de conocimiento	El proyecto se encuentra en construcción.
Soylu	OpenOpps (contratación pública de múltiples fuentes) y datos abiertos de varios países y otras entidades reguladoras	Grafos semánticos de conocimiento, aprendizaje automático, detección de anomalías, entre otros.	Los problemas de calidad de datos representaron un reto. No obstante, se logra habilitar un tipo de analítica avanzado que con otros modelos de datos sería imposible obtener.
Herrera	*Open Contracting Data Standard* (millones de registros asociados a contratos públicos)	Aprendizaje automático	Los resultados sugieren que el modelo propuesto tiene un alto potencial de identificar casos reales de corrupción y permitiría pasar de una estrategia reactiva a una proactiva.
Carneiro	Conjuntos de datos públicos	Detección de fraudes	El sistema se encuentra en construcción
Van Erven	Datos abiertos del gobierno de Brazil	Árboles de mínima expansión	Se detectaron escenarios con compañías altamente centrales que representan riesgo de corrupción.
Swords	*Tenders Electronics Daily* (datos abiertos de contratación de la Unión Europea)	Centralidad y detección de comunidades	Se detecta incremento en el rendimiento asociado al procesamiento de consultas. Se detectaron entidades con alta influencia en el conjunto de datos.

Cuadro 3: Análisis de trabajos similares

pública en Colombia. Esta exploración conceptual estará alineada con los objetivos del proyecto. El resultado de esta etapa será traducir el conocimiento en términos de un problema de minería de datos.

- Entendimiento de los datos
  En este punto se realizará una recolección inicial de datos desde el SECOP a través de `datos.gov.co` y se llevarán a cabo actividades relacionadas con el entendimiento de los mismo teniendo como referencia el conocimiento adquirido sobre contratación pública.

- Preparación de los datos
  En esta fase se realizarán a cabo la transformación de los datos recopilados desde el SECOP hacia una base de datos basada en grafos lista para analizar.

- Modelamiento
  Una vez preparado el conjunto de datos, se utilizarán algunos de los algoritmos descritos en el Marco Teórico con el fin de encontrar patrones ocultos en la información orientados a identificar participantes sospechosos en los procesos de contratación.

- Evaluación
  Para evaluar la efectividad del modelo, se recopilarán datos en los que previamente se hayan identificado patrones de corrupción con el fin de comparar los resultados.

- Despliegue
  Los resultados del modelo se prepararán para ser mostrados a través de herramientas de visualización de grafos.

7. Cronograma de trabajo

El siguiente cronograma de trabajo se concibe para llevarse a cabo desde el 24 de enero de 2022 hasta el 27 de mayo de 2022.

Actividades	Meses	1				2				3				4			
	Semanas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Revisión Bibliográfica																	
Comprensión del negocio																	
Comprensión de los datos																	
Preparación de los datos																	
Modelado																	
Evaluación																	
Despliegue																	
Elaboración de Informe Final																	

8. Presupuesto

Bajo las condiciones de un proyecto empresarial hipotético, se debería disponer por lo menos de una célula compuesta de un ingeniero de calidad de datos, un ingeniero de datos, un científico de datos y un documentador. El presupuesto mostrado a continuación asume una célula como esta con salarios promedios a 2021 en pesos colombianos y un porcentaje de dedicación. Con estos datos, se estima que el valor aproximado del proyecto estaría alrededor de los 23 millones de pesos.

Recurso	**Costo por hora**	**Porcentaje de dedicación**	**Total costo en proyecto**
Calidad de datos	$13.846	60 %	$5.316.864
Ingeniería de datos	$24.615	30 %	$4.726.080
Analítica de datos	$24.952	30 %	$4.790.784
Documentación	$12.923	100 %	$8.270.720
Total	**$76.336**		**$23.104.448**

9. Anexos

Nombre de la columna	Descripción	Tipo
Nombre	Nombre del proveedor como se registro en SECOP II	Texto simple
NIT	Número de identificación con el que figura el proveedor en SECOP II	Texto simple
Tipo de empresa	Tipo de empresa que declara el proveedor al registrarse	Texto simple
¿Es PYME?	Determina si el proveedor se registró como pequeña empresa	Texto simple
Ubicación	Ubicación geográfica de la empresa, de acuerdo al registro del proveedor	Texto simple
Fecha Creación	Fecha en la que se hizo el primer registro del proveedor	Fecha y hora
País	País de origen del Proveedor	Texto simple
Departamento	En caso de Ser un proveedor colombiano, indica el departamento al que corresponde la ubicación principal del Proveedor	Texto simple
Municipio	En caso de Ser un proveedor colombiano, indica el Municipio al que corresponde la ubicación principal del Proveedor	Texto simple
Código Categoría Principal	Codigo UNSPSC principal del proveedor	Texto simple
Descripción Categoría Principal	Descripción UNSPSC principal del proveedor	Texto simple

Cuadro 4: Columnas del conjunto de datos *Proveedor*.

Nombre de la columna	Descripción	Tipo
Nombre de entidad	Nombre de la entidad del estado que publica el contrato	Texto simple
NIT Entidad	NIT de la entidad del estado que publica el contrato	Número
Departamento	Departamento en el cual se registró la entidad del estado que publica el contrato	Texto simple
Ciudad	Ciudad en el cual se registró la entidad del estado que publica el contrato	Texto simple
Localización	Ubicación completa de la entidad del estado que publica el contrato	Texto simple
Orden	Orden entidad del estado que publica el contrato	Texto simple
Sector	Sector entidad del estado que publica el contrato	Texto simple
Rama	Rama del estado de la entidad que publica el contrato	Texto simple
Entidad Centralizada	Define si la entidad es descentralizada o centralizada	Texto simple
Proceso de Compra	Identificador del procesos de compra publicado	Texto simple
ID Contrato	Identificador del proceso de compra publicado	Texto simple
Referencia del Contrato	Identificador del contrato firmado, generado por la entidad del estado	Texto simple
Estado Contrato	Estado del contrato, frente a su ejecución, firma o liquidación	Texto simple
Codigo de Categoria Principal	Codigo UNSPSC de la categoría principal para el contrato	Texto simple
Descripción del Proceso	Descripción del objeto del proceso de compra	Texto simple
Tipo de Contrato	Tipo de contrato de acuerdo a su marco jurídico	Texto simple
Modalidad de Contratación	Modalidad de contratación de acuerdo al modelo de selección	Texto simple
Justificación Modalidad de Contratación	Justificación de la modalidad, el escenario bajo el cual se toma la decisión de definir una u otra modalidad de contratación	Texto simple
Fecha de Firma	Fecha en que fue firmado digitalmente el contrato	Fecha y hora
Fecha de Inicio del Contrato	Fecha de inicio de las responsabilidades contractuales	Fecha y hora
Fecha de Fin del Contrato	Fecha de fin de las responsabilidades contractuales	Fecha y hora
Fecha de Inicio de Ejecución	Fecha de inicio de la ejecución de las actividades del contrato	Fecha y hora
Fecha de Fin de Ejecución	Fecha de fin de la ejecución de las actividades del contrato	Fecha y hora
Condiciones de Entrega	Condiciones bajo las cuales se entrega el producto o servicio	Texto simple
TipoDocProveedor	Tipo de documento del proveedor adjudicado	Texto simple
Documento Proveedor	Número de documento del proveedor adjudicado	Texto simple
Proveedor Adjudicado	Nombre del proveedor adjudicado	Texto simple
Es grupo	Determina el proveedor es un grupo de entidades, existe un conjunto de datos de CCE que contiene la conformación de los grupos	Texto simple
Es Pyme	Determina si la empresa es una Pyme	Texto simple
Habilita Pago Adelantado	Determina si el contrato tiene habilitada la opción de pago de adelantos	Texto simple
Liquidación	Determina si el contrato ha sido liquidado	Liquidación
Obligaciones Ambiental	Determina si el contrato tiene compromisos de cumplimiento a obligaciones ambientales	Texto simple
Obligaciones Postconsumo	Determina si el contrato tiene compromisos de cumplimiento a obligaciones posteriores a la entrega del producto o prestación del servicio	Texto simple
Reversión	Determina si el contrato ha sido reversado	Texto simple
Valor del Contrato	Valor total del contrato	Número
Valor de pago adelantado	Valor del pago por adelantado	Número
Valor Facturado	Valor Facturado a la fecha	Número
Valor Pendiente de Pago	Valor Pendiente de Pago a la fecha	Número
Valor Pagado	Valor Pagado a la fecha	Número
Valor Amortizado	Valor Amortizado a la fecha	Número
Valor Pendiente de Amortización	Valor Pendiente de Amortizacion a la fecha	Número
Valor Pendiente de Ejecución	Valor Pendiente de Ejecucion a la fecha	Número
Estado BPIN	Estado de asignación del código del Banco de Proyectos de Inversión	Texto simple
Código BPIN	Código asociado al Banco de Proyectos de Inversión	Texto simple
Año BPIN	Año de asignación del código del Banco de Proyectos de Inversión	Texto simple
Saldo CDP	Saldo del CDP asignado al proceso y al contrato	Número
Saldo Vigencia	Saldo actual para la vigencia del CDP asignado al proceso y al contrato	Número
EsPostConflicto	Determina si el proceso está asociado a algún evento de acuerdo de paz	Texto simple
URLProceso	URL del proceso de compra en la plataforma SECOP II	URL
Destino Gasto	Destino del gasto, a nivel presupuestal	Texto simple
Origen de los Recursos	Origen de los Recursos, a nivel presupuestal	Texto simple
Dias Adicionados	Número de días en que el contrato ha sido adicionado	Número
Puntos del Acuerdo	En caso de ser un proceso que da cumplimiento a compromisos en el acuerdo de paz, determina a qué puntos da conformidad	Texto simple
Pilares del Acuerdo	En caso de ser un proceso derivado de compromisos del acuerdo de paz, define el pilar de acuerdo de paz al que corresponde	Texto simple
Nombre Representante Legal	Nombre del Representante legal de la empresa proveedora	Texto simple
Nacionalidad Representante Legal	Nacionalidad del representante legal de la empresa proveedora	Texto simple
Tipo de Identificación Represente Legal	Tipo de identificación del representante legal de la empresa proveedora	Texto simple
Identificación Representante Legal	Número de identificación del representante legal	Texto simple
Genero Representante Legal	Número de identificación del representante legal	Texto simple
Presupuesto General de la Nación - PGN	Valor de origen de los recursos que corresponde al Presupuesto General de la Nación – PGN	Número
Sistema General de Participaciones	Valor de origen de los recursos que corresponde al Sistema General de Participaciones	Número
Sistema General de Regalías	Valor de origen de los recursos que corresponde al Sistema General de Regalías	Número
Recursos Propios (Alcaldía, Gobernaciones y Resguardos Indígenas)	Valor de origen de los recursos que corresponden a Recursos Propios (Alcaldías, Gobernaciones y Resguardos Indígenas)	Número
Recursos de Crédito	Valor de origen de los recursos que corresponde a Recursos de Crédito	Número
Recursos Propios	Valor de origen de los recursos que corresponde a Recursos Propios	Número
Ultima Actualización	Fecha de actualización del registro	Fecha y hora
Código Entidad	Código generado en la plataforma para la entidad	Texto simple

Cuadro 5: Columnas del conjunto de datos *Contratos Electrónicos*.

Nombre de la columna	Descripción	Tipo
Entidad	Nombre de la Entidad que publica el proceso de compra pública	Texto simple
Nit Entidad	NIT de la Entidad que publicó el proceso	Texto simple
Departamento Entidad	Departamento en el cual está registrada la entidad	Texto simple
Ciudad Entidad	Ciudad en la cual está registrada la entidad	Texto simple
OrdenEntidad	Orden de la Entidad (Nacional, Regional)	Texto simple
Entidad Centralizada	Identifica si la entidad es o no centralizada	Texto simple
ID del Proceso	Identificador Único del Proceso, valor generado por la plataforma	Texto simple
Referencia del Proceso	Identificador del Proceso, valor generado por la Entidad	Texto simple
PCI	Codigo de Unidad - Sub Unidad Contratación	Texto simple
ID del Portafolio	Identificador del Portafolio al cual corresponde el proceso de compra	Texto simple
Nombre del Procedimiento	Nombre dado al proceso de compra por la Entidad	Texto simple
Descripción del Procedimiento	Primera definición de las características principales del proceso	Texto simple
Fase	Fase en la que actualmente se encuentra el proceso	Texto simple
Fecha de Publicacion del Proceso	Fecha de la publicación inicial del proceso de compra	Fecha y hora
Fecha de Ultima Publicación	Fecha de la última publicación hecha para el proceso de compra	Fecha y hora
Fecha de Publicacion (Fase Planeacion Precalificacion)	Fecha de publicación, dentro del proceso, de la fase de Planeación en Precalificación	Fecha y hora
Fecha de Publicacion (Fase Seleccion Precalificacion)	Fecha de publicación, dentro del proceso, de la fase de Selección en Precalificación	Fecha y hora
Fecha de Publicación (Manifestación de Interés)	Fecha de publicación, dentro del proceso, de la fase de Manifestación de Interés	Fecha y hora
Fecha de Publicación (Fase Borrador)	Fecha de publicación, dentro del proceso, de la fase Borrador	Fecha y hora
Fecha de Publicación (Fase Selección)	Fecha de publicación, dentro del proceso, de la fase Selección	Fecha y hora
Precio Base	Precio Base, proyectado, del proceso de Compra	Número
Modalidad de Contratación	Modalidad de selección bajo la cual se desarrolla el proceso de Compra	Texto simple
Justificación Modalidad de Contratación	En caso de requerirse, Justificación para la modalidad de selección elegida para el proceso de compra	Texto simple
Duración	Valor de la Duración estimada del proceso de compra pública	Número
Unidad de Duración	Unidad que aplica a la Duración estimada del proceso de compra pública	Texto simple
Fecha de Recepción de Respuestas	Fecha asignada para la recepción de respuestas por parte de los proveedores, dentro del proceso de compra	Fecha y hora
Fecha de Apertura de Respuesta	Fecha Estimada para la Apertura de las respuestas	Fecha y hora
Fecha de Apertura Efectiva	Fecha Real para la Apertura de las respuestas	Fecha y hora
Ciudad de la Unidad de Contratación	Cuidad en la que aparece registrada la unidad de contratación de la Entidad	Texto simple
Nombre de la Unidad de Contratación	Nombre de la unidad de contratación de la Entidad	Texto simple
Proveedores Invitados	Número de Proveedores invitados a participar del proceso, en total	Número
Proveedores con Invitación Directa	Proveedores con Invitación a participar hecha de forma directa	Número
Visualizaciones del Procedimiento	Número de Visualizaciones hechas a través de la herramienta, del Proceso de Compra	Número
Proveedores que Manifestaron Interés	Proveedores que Manifestaron Interés en el proceso a través de la plataforma	Número
Respuestas al Procedimiento	Respuestas hechas al procedimiento, tanto de proveedores como de la misma entidad	Número
Respuestas Externas	Número de Respuestas hechas por entes externos	Número
Conteo de Respuestas a Ofertas	Número de Respuestas hechas de forma directa en las ofertas	Número
Proveedores Únicos con Respuestas	Proveedores Únicos que han redactado respuestas en el proceso	Número
Numero de Lotes	Número de lotes de artículos solicitados dentro del proceso	Número
Estado del Procedimiento	Estado actual de desarrollo del procedimiento de compra pública	Texto simple
ID Estado del Procedimiento	Identificador del Estado del procedimiento	Número
Adjudicado	Determina si el proceso fue adjudicado	Texto simple
ID Adjudicación	Identificador de la adjudicación	Texto simple
CodigoProveedor	Código, en la plataforma, del proveedor adjudicado	Texto simple
Departamento Proveedor	Departamento en el que está registrado el proveedor adjudicado	Texto simple
Ciudad Proveedor	Ciudad en la que está registrado el proveedor adjudicado	Texto simple
Fecha Adjudicación	Fecha en la que se hizo la adjudicación del proceso para el proveedor seleccionado	Fecha y hora
Valor Total Adjudicación	Valor total Adjudicado	Número
Nombre del Adjudicador	Nombre del Usuario que ejecutó la acción de adjudicación	Texto simple
Nombre del Proveedor Adjudicado	Nombre del Proveedor Adjudicado	Texto simple
NIT del Proveedor Adjudicado	NIT del Proveedor Adjudicado	Texto simple
Código Principal de Categoria	Código UNSPSC de la categoría principal del producto o servicio adquirido en proceso de compra	Texto simple
Estado de Apertura del Proceso	Estado actual de Apertura de información del proceso	Texto simple
Tipo de Contrato	Tipo de Contrato definido para el proceso de compra	Texto simple
Subtipo de Contrato	Subtipo de Contrato definido para el proceso de compra	Texto simple
Categorias Adicionales	Identificador de las categorías UNSPSC adicionales, incluidas en el producto o servicio adquirido en proceso de compra	Texto simple
URLProceso	URL, en la plataforma, en la que se puede consultar el proceso de compra	URL del sitio web
Codigo Entidad	Identificador único de la Entidad en la plataforma	Texto simple

Cuadro 6: Columnas del conjunto de datos *Procesos de Contratación*.

Nombre de la columna	Descripción	Tipo
Identificador	Identificador único del evento de modificación	Texto simple
ID_Contrato	Identificador del contrato al que corresponde la modificación	Texto simple
Tipo	Tipo de modificación, de acuerdo al impacto que tiene sobre el contrato	Texto simple
Descripción	Descripción detallada de la justificación de la adición o modificación	Texto simple
FechaRegistro	Fecha en que se hace la modificación	Fecha y hora

Cuadro 7: Columnas del conjunto de datos *Adiciones*.

Referencias

[1] R. D. Angarita, L. A. R. Carvajalino, and M. M. D. Bueno, "Características del sistema de contratación estatal en Colombia," *HIPOTESIS LIBRE*, no. 11, 2018.

[2] J. S. Betancourt Cortes *et al.*, "El fenómeno de la corrupción en los procesos de licitación pública en contratación estatal en Colombia," 2018.

[3] M. Needham and A. Hodler, *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. O'Reilly Media, 2019.

[4] D. Fernandes and J. Bernardino, "Graph databases comparison: Allegrograph, arangodb, infinitegraph, neo4j, and orientdb." in *DATA*, 2018, pp. 373–380.

[5] C. de la República de Colombia, "Ley 80 de 1993," 1993.

[6] D. Bechberger and J. Perryman, *Graph Databases in Action*, ser. In Action. Manning Publications, 2020. [Online]. Available: https://books.google.com.co/books?id=kWIFEAAAQBAJ

[7] I. Robinson, J. Webber, and E. Eifrem, *Graph databases: new opportunities for connected data*. O'Reilly Media, Inc., 2013.

[8] L. Wiese, "Data analytics with graph algorithms–a hands-on tutorial with neo4j," *BTW 2019–Workshopband*, 2019.

[9] ICIJ, "Pandora papers: An offshore data tsunami." [Online]. Available: https://www.icij.org/investigations/pandora-papers/about-pandora-papers-leak-dataset/

[10] BBC, "Pandora papers: guía simple para entender una de las mayores filtraciones de la historia con 12 millones de documentos divulgados." [Online]. Available: https://www.bbc.com/mundo/noticias-internacional-58784755

[11] M. Mukhopadhyay and K. Ghosh, "Panama papers: How data science fought corruption," *Available at SSRN 3644821*, 2020.

[12] L. K. Branting, F. Reeder, J. Gold, and T. Champney, "Graph analytics for healthcare fraud risk estimation," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, aug 2016.

[13] A. P. Naik and S. Bojewar, "Tweet analytics and tweet summarization using graph mining," in *2017 international conference of electronics, communication and aerospace technology (ICECA)*, vol. 1. IEEE, 2017, pp. 17–21.

[14] G. Drakopoulos, A. Kanavos, P. Mylonas, and S. Sioutas, "Defining and evaluating twitter influence metrics: a higher-order approach in neo4j," *Social Network Analysis and Mining*, vol. 7, no. 1, pp. 1–14, 2017.

[15] E. Simperl, O. Corcho, M. Grobelnik, D. Roman, A. Soylu, M. J. F. Ruíz, S. Gatti, C. Taggart, U. S. Klima, A. F. Uliana *et al.*, "Towards a knowledge graph based platform for public procurement," in *Research Conference on Metadata and Semantics Research*. Springer, 2018, pp. 317–323.

[16] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, F. Martinez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl *et al.*, "Integrating and analysing public procurement data through a knowledge graph: A demonstration in a nutshell," in *Proceedings of ISWC*, 2020.

[17] A. Soylu, O. Corcho, B. Elvesæter, C. Badenes-Olmedo, F. Y. Martínez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl *et al.*, "Enhancing public procurement in the european union through constructing and exploiting an integrated knowledge graph," in *International Semantic Web Conference*. Springer, 2020, pp. 430–446.

[18] I. Fountoukidis, I. E. Antoniou, and N. C. Varsakelis, "Analyzing the competition in public procurement procedures using graph analytics," *Social and Economic Challenges and Regional Development*, p. 107, 2021.

[19] D. J. H. Murillo, "Using social network analysis in open contracting data to detect corruption and collusion risks."

[20] D. Carneiro, P. Veloso, A. Ventura, G. Palumbo, and J. Costa, "Network analysis for fraud detection in portuguese public procurement," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2020, pp. 390–401.

[21] G. C. van Erven, M. Holanda, and R. N. Carvalho, "Detecting evidence of fraud in the brazilian government using graph databases," in *World conference on information systems and technologies*. Springer, 2017, pp. 464–473.

[22] M. Swords, "Finding patterns in procurements and tenders using a graph database," 2019.

[23] A. Serrano Cuervo *et al.*, "Corrupción en la contratación pública en Colombia," 2014.

[24] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Springer-Verlag London, UK, 2000.