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
Abril 4 de 2020

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 este tipo de datos.

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 cuentan con varios métodos: 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.

Para este proyecto se utilizarán los datos abiertos proporcionados por el Gobierno de Colombia a través de `gov.co` y el SECOP II. Con estos datos se construirá una base de datos de grafos. Se espera que posteriormente, al utilizar algoritmos de grafos, se puedan evidenciar prácticas corruptas.

2. Marco teórico

2.1. Conceptos de contratación pública

Para conformar un bosquejo sobre el procesos 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 jusa retribución.

Gran parte de los conceptos que conforman la contratación pública provienen de un ámbito normativo. En particular, la Ley 80 de 1993 [4] 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 Sistema Electrónico para la Contratación Pública (SECOP).

2.2. 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 [5], es un motor de almacenamiento de datos que combina las estructuras básicas de grafos (nodos y arcos) con tecnología de persistencia y un lenguaje de consulta.

Modelar los datos es la primera parte del proceso, el procesamiento de los datos 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 [6].

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*.

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

La búsqueda de caminos resulta fundamental en la analítica de grafos. Aquí, la búsqueda de la ruta más corta es tal vez la actividad más común. La centralidad consiste en comprender qué nodos son los más importantes en una red. La detección de comunidad se soporta en concepto de *conectividad* que permite encontrar núcleos o centros dentro de la red.

3. Estado del arte

Son múltiples los ejemplos de la utilización de grafos par resolver problemas en los que el contexto del problema es complejo y la relación entre los elementos no es inmediata.

En trabajos como el de Branting [7] 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 colocación geoespacial.

Muchos de estos algoritmos han sido aplicados con éxito en el análisi de redes sociales. En trabajos como el de Naik [8] 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 [9] en donde se analizan patrones de conectividad en *Twitter* con el fin de obtener métricas de influencia dentro de la red social.

Simperl [10] 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 [11] y [12] construyo también un grafo de conocimiento con datos abiertos provenientes de la contratación estatal en 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 multilinguistica.

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 la competición en los procesos de aprovisionamiento para insumos de la salud pública. [13]: 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.

4. Planteamiento del problema

Según Serrano [14], 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

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.

5.2. Objetivos específicos

- 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 [15]. Por consiguiente, el trabajo comprenderá las siguientes fases:

- Entendimiento del negocio
  En esta etapa se estudiará con la profundidad necesaria los procesos de contratación 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 26 de julio de 2021 hasta el 20 de noviembre de 2021.

Actividad	No. de semanas
Entendimiento del negocio	3
Entendimiento de los datos	3
Preparación de los datos	3
Modelamiento	3
Evaluación	2
Despliegue	2

8. Presupuesto

Recurso	Fuente de financiación	Costo (COP)	Observaciones
Acceso a Internet	Propia	$110.000	
Equipo de cómputo	Propia	$0	Adquirido
Neo4j Enterprise 2.2.5 for Developers	Propia	$0	Gratis con previo registro y uso no comercial.
Intérprete de Python 3.9	Propia	$0	Licencia Open Source compatible con GPL.

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

[5] 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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.