

Evaluación del riesgo sobre las dinámicas contractuales en Colombia: caso aplicado al Programa de Alimentación Escolar para el periodo 2014 – 2017 en la ciudad de Bogotá.

Introducción

La evaluación de la calidad de los contratos estatales es un tema muy discutido en la actualidad dado que representa una de las maneras más claras de empoderamiento de la población sobre la designación de los recursos que se le aportan al Estado (Martínez, 2014). En el contexto colombiano se han visto diversos casos de corrupción (Silva, 2015) (LA FM, 2018) asociados a la contratación pública que generan amplios huecos financieros en la nación, fomentando el detrimento de la calidad de vida de gran parte de la población del país. Chang y Chu (2006) afirman que *"la corrupción política es considerada uno de los problemas más destructivos, pero aún no resueltos que son comunes a la mayoría de las sociedades. Es importante destacar que la corrupción política representa una traición directa y brutal de la confianza depositada en las instituciones"*. Por esta razón –y contando con las herramientas tecnológicas gubernamentales¹ emergentes que permiten hoy en día crear metodologías para promover una mayor transparencia económica sobre los movimientos transaccionales de Colombia– se quiso generar un algoritmo de respuesta y evaluación que genere una alerta precisa sobre el riesgo de la designación de los recursos gubernamentales del país.

Para lograr esto se delimitó el contexto evaluado al conjunto de contratos asociados directamente a la implementación del Programa de Alimentación Escolar (PAE) en la ciudad de Bogotá. El PAE es un programa que se creó en 1926 en Boyacá y que consiste *"en el suministro organizado de un complemento nutricional con alimentos inocuos, a los niños, niñas y adolescentes matriculados en el sistema educativo público, y el desarrollo de un conjunto de acciones alimentarias, nutricionales, de salud y de formación [...] que contribuyen a mejorar el desempeño de los escolares y apoyar su vinculación y permanencia en el sistema educativo"* (Hoyos Catañeda, 2015). Este contexto se escogió ya que ha sido foco de investigación en varias regiones del país debido a su mal manejo por parte de las entidades estatales responsables de su ejecución (Colombia Compra Eficiente, 2014). Se fijó un panorama temporal de cuatro años a partir del 01 de enero de 2014 hasta el 31 de diciembre de 2017 y, además, se evaluaron las limitaciones asociadas a las fuentes de información disponibles para el proyecto, según la completitud y la consistencia de los datos aportados por cada una de ellas. En este trabajo de investigación

¹ "El gobierno electrónico se ve como la plataforma para lograr una mayor transparencia, lo que aumenta los niveles de responsabilidad y frenar la corrupción" (McGee & Gaventa, 2010).

se evaluaron las bases de datos aportadas por el Sistema Electrónico de Contratación Pública (SECOP I y SECOP II) y fueron consultadas de manera electrónica en los portales web que el Estado colombiano dispone para la consulta abierta y libre sobre las dinámicas contractuales del país. El SECOP, que fue creado mediante el decreto 2178 de 2006 “*es un sistema gerencial basado en una herramienta tecnológica de apoyo a la contratación pública que integra los sistemas de información actuales, facilitando información sobre estos procesos y entregando a las entidades estatales herramientas para modernizar y optimizar la gestión contractual*” (Ministerio de Comunicaciones, 2008). Más adelante se mencionará en qué medida se tuvieron en cuenta dichas bases de datos para la creación del índice de riesgo contractual.

Asimismo, se implementaron diversos tipos de análisis, algunos asociados a la rama investigativa de minería de datos con aprendizaje no supervisado, que permitieron desarrollar una metodología imparcial y objetiva en la cual el criterio personal del evaluador no pudiera jugar un rol arbitrario para el análisis de los resultados obtenidos. Según Ławrynowicz y Volker (2014) el aprendizaje no supervisado “*encuentra patrones ocultos o estructuras intrínsecas en los datos. Se utiliza para extraer inferencias de conjuntos de datos que consisten en datos de entrada sin respuestas etiquetadas*”. Este tipo de desarrollo investigativo es innovador en Colombia, debido a que son muchas las metodologías de calificación que se basan en el análisis bajo ponderadores, proponiendo “*recomendaciones demasiado simplificadas, que dejan de tener en cuenta escenarios sociopolíticos complejos*” (Martínez, 2014). Este tipo de análisis causa que la creación de un índice de calidad, riesgo o riesgo se fundamente en supuestos que modifican de manera importante su resultado final. A causa de implementar un nuevo sistema de riesgo contractual, este proyecto está sujeto a varias limitantes que serán tratadas más adelante.

El objetivo general del trabajo se basa en estudiar la relación existente entre los atributos cuantificables propios de cada contrato y la estructura de la red de contratación pública en Colombia para los años 2014, 2015, 2016 y 2017 en el PAE, específicamente en la ciudad de Bogotá. Paralelamente, se creará un índice de riesgo para los contratos estatales a partir de la información disponible en el SECOP, se evaluarán las asociaciones que puedan llegar a existir entre el índice de riesgo y la estructura de la red de contratación pública en el sector estudiado, y se creará una herramienta y una metodología de análisis sobre el riesgo de la contratación pública que pueda ser replicada en otros sectores estrechamente vinculados con las dinámicas contractuales de Colombia.

Métodos

La metodología de esta investigación se dividirá en tres etapas según el aporte teórico, científico y conceptual que cada una de ellas le brinda al estudio. Se tratarán temas como el proceso de recolección y limpieza de datos, la creación del índice de riesgo contractual y el análisis de redes bipartitas para el entendimiento de las dinámicas de contratación pública en el contexto mencionado.

Recolección y limpieza de datos

La recolección de información se basó fundamentalmente en el estudio de las bases de datos abiertas brindadas por el Estado colombiano, en las que se evidencian los movimientos contractuales “desde la planeación del contrato hasta su liquidación” (Colombia Compra Eficiente, 2017). Existen datos disponibles desde años previos al 2014, pero muchos de estos carecen de completitud en la información brindada (Anexo 8). Por dicha razón, se decidió centrar la recolección de datos desde el año 2014 hasta el año 2017. Se estudiaron varias fuentes de información empezando por el portal web de datos abiertos del país, tomando como componente principal las bases de datos del SECOP I y SECOP II. Estas bases de datos reflejan la información referente a cada uno de los contratos realizados por el Estado y cuentan con una interfaz en la que se expone de manera digitalizada la información contenida en cada uno de los documentos. De esta manera, fue posible acceder tanto a los archivos digitalizados de los contratos como a la información tabulada de los mismos, permitiendo la verificación de la veracidad y la consistencia entre ambas partes.

Al estudiar las bases de datos de diversos sectores e instituciones vinculados a escándalos de corrupción y estrechamente relacionados con las dinámicas de contratación pública en Colombia², fue posible observar que la consistencia entre los datos brindados por las bases de información y los contratos originales no coinciden en algunos de los casos. Dada la presencia de esta problemática, se decidió acotar el estudio a un sector con una menor cantidad de contratos asociados que permitiera a la investigación seguir siendo pertinente para el cumplimiento de los objetivos planteados. Por este motivo, se decidió no reducir el panorama temporal establecido desde un principio y evaluar las dinámicas de contratación pública presentes en el SECOP I y relacionadas con la implementación del PAE en la ciudad de Bogotá.

Se crearon filtros de búsqueda en la base de datos abierta escogida, extrayendo los movimientos contractuales que tenían relación directa con la implementación del PAE en la ciudad de Bogotá dentro del espectro temporal establecido: 484 contratos fueron seleccionados, agrupando un total aproximado de 2.400 documentos digitales.

Como se mencionó anteriormente, a pesar de contar con los recursos electrónicos libres brindados por el Estado colombiano, la información que se encontró (especialmente en el SECOP I) presentaba inconsistencias. Los archivos descargados de estos portales web incluían erratas en los campos tabulados, así como falta de homogeneidad en los formatos de sus columnas. Estas primeras inconsistencias no fueron categorizadas como errores propios de los contratos y se desarrolló un código en Python³ que lograra incorporar y

² Estudio realizado durante el segundo semestre de 2017 sobre las entidades: Instituto Nacional de Vías (INVIAS), Instituto Colombiano de Bienestar Familiar (ICBF), Ministerio de Agricultura y Desarrollo Rural (MINAGRICULTURA), Departamento Administrativo para la Prosperidad Social (DPS), Servicio Nacional de Aprendizaje (SENA).

³ Entregable # 1: herramienta de limpieza preliminar (aplicable únicamente en SECOP I).

estandarizar la información de cada uno de los movimientos contractuales para facilitar su posterior análisis.

Una vez procesada la información mediante la herramienta creada para la estandarización de datos, se empezó a verificar y validar cada uno de los contratos reportados en el SECOP teniendo como punto de comparación el contrato original disponible de forma digital. Se verificó que la información reportada por el SECOP correspondiera de manera exacta con la estipulada en los documentos originales, comparando el nombre de las entidades del Estado, de los contratistas y sus representantes legales (incluidos sus números de identificación); las fechas de firma, inicio y fin de cada contrato, así como los plazos con sus respectivas adiciones de tiempo y la cuantía del contrato, sus adiciones y su valor final. Este proceso llevó a que se verificaran alrededor de 1.500 documentos con el fin de contar con información totalmente válida para la investigación⁴.

Índice de riesgo contractual

El objetivo de crear un índice de riesgo para cada contrato es lograr generar un método que permita clasificarlo dentro de un rango de “normalidad” según el valor de sus atributos. La creación del índice se dividió en tres etapas que consistieron en categorizar, agrupar y clasificar los datos asociados a los valores de cada contrato: plazo de ejecución, porcentaje de adiciones sobre el plazo de ejecución, valor inicial del contrato y porcentaje de adiciones sobre el valor inicial del contrato. Las adiciones en el plazo y en el valor se trabajaron de manera porcentual ya que así se obtiene más información sobre el riesgo del contrato y se evidencia qué tan significativo es dicho atributo para cada uno y no para el conjunto de contratos estudiados.

Categorización

Antes de proceder con el tratamiento de los datos se buscó entender su comportamiento con el fin de establecer una metodología clara para la estandarización de las variables. Este procedimiento fue necesario dado que la unidad correspondiente a cada una de las variables no era numéricamente comparable con el resto (Tabla 1).

<i>Atributo</i>	<i>Unidad</i>
<i>Plazo</i>	<i>Días</i>
<i>Adición en plazo</i>	<i>Porcentaje</i>
<i>Valor</i>	<i>Dinero (COP)</i>
<i>Adición en valor</i>	<i>Porcentaje</i>

Tabla 1: Tipología de los atributos de los contratos

⁴ Entregable # 2: base de datos verificada, 484 contratos vinculados a la implementación del PAE.

Para la estandarización de los atributos se implementó un método utilizado comúnmente en este tipo de contextos. Éste consistió en normalizar los datos respecto a la desviación estándar y al valor medio asociados a cada atributo. A continuación se presenta la fórmula asociada (Jain & Bhandare, 2011):

$$\sigma_i = \sqrt{\frac{(\sum_{j=1}^C (x_{i,j} - \bar{x}_i)^2)}{C-1}} \quad \forall i \in A \quad \rightarrow \quad y_{i,j} = \frac{x_{i,j} - \mu_i}{\sigma_i} \quad \forall i \in A, j \in C$$

Tal que μ_i es el promedio de atributo $i \in \text{Atributo } (A)$, σ_i representa la desviación estándar del atributo $i \in \text{Atributos } (A)$, $x_{i,j}$ el valor del atributo $i \in \text{Atributos } (A)$ para el contrato $j \in \text{Contratos } (C)$.

En este sentido, el valor adoptado por la variable $y_{i,j}$ representa la cuantía estandarizada del atributo $i \in \text{Atributos } (A)$ para el contrato $j \in \text{Contratos } (C)$ sobre la cual se realizará el procedimiento de agrupamiento descrito en el siguiente apartado.

Agrupamiento

Una vez estandarizados los valores de los atributos se procedió al agrupamiento de los datos. El agrupamiento, o “*clustering*”, hace parte de la rama de minería de datos llamada “*clasificación no supervisada*” o “*aprendizaje no supervisado*” que segmenta un conjunto de datos en grupos que pueden ser disjuntos o no. En otras palabras, “*clustering es una tarea no supervisada, cuyo objetivo es agrupar un conjunto de objetos en clases de objetos similares. Un clúster es una colección de objetos que son similares entre sí dentro del mismo clúster y diferentes a los objetos de otros clústeres.*” (Ławrynowicz & Volker, 2014). A través de este análisis se realiza el agrupamiento mediante la implementación del algoritmo de clustering “*k-medias*”. Al aplicarlo se logró clasificar cada uno de los contratos en una clase específica donde sus homólogos comparten cualidades similares según sus atributos (Tabla 1).

En primer lugar, se aplicó el proceso denominado “*método de codo*” que brinda la información necesaria para determinar el número de grupos más adecuado para la segmentación de contratos (Kodinariya & Makwana, 2013). Este método varía el número de particiones minimizando la inercia general del sistema intervenido. La siguiente función representa el criterio de decisión para la creación de grupos:

$$\min \text{inercia} = \min \sum_{k \in P} \sum_{j \in C} \sum_{i \in A} \|x_{i,j,k} - \mu_{i,k}\|^2$$

s. a

$$\mu_{i,k} = \frac{\sum_{j \in C} x_{i,j,k}}{A} \quad \forall i \in A, k \in P$$

Donde $k \in \text{Particiones}$ es el número de grupos creados, $x_{i,j,k}$ el valor del atributo $i \in \text{Atributos}(A)$ del contrato $j \in \text{Contratos}(C)$ perteneciente a la partición $k \in \text{Particiones}(P)$ y donde $\mu_{i,k}$ representa la media del atributo $i \in \text{Atributos}$ dentro de la partición $k \in \text{Particiones}$. En el apartado de resultados se mostrará gráficamente de qué manera se determinó el número más adecuado de particiones para la muestra de contratos analizada, así como las características de cada una de estas agrupaciones.

Clasificación

Una vez realizada la agrupación de cada uno de los contratos según la minimización de la inercia total calculada, se determinó cuáles de los contratos podrían llegar a ser considerados como riesgosos dentro de su grupo. Calculando un intervalo de confianza ($\alpha = 5\%$) para la distancia euclidiana de cada una de las observaciones a su centroide correspondiente, se asumió que los contratos cuya distancia fuera mayor al límite superior del intervalo de confianza serían considerados como riesgosos. Una de las ventajas de realizar dicho procedimiento es que el sistema en sí mismo asigna la clasificación para los contratos según sus atributos y no según la categoría a la cual el evaluador considere que el contrato pertenece. El intervalo de confianza para cada clúster se calculó de la siguiente manera:

$$s_{j,k} = \sqrt{\sum_{i \in A} (x_{i,j,k} - \mu_{i,k})^2} \quad \forall j \in C, k \in P \rightarrow \bar{s}_k = \frac{\sum_{j \in C} s_{k,j}}{C} \quad \forall k \in P$$

$$\sigma_k = \sqrt{\frac{\sum_{j \in C} (s_{j,k} - \bar{s}_k)^2}{C-1}} \quad \forall k \in P \rightarrow Z_{k, \frac{\alpha}{2}} \cdot \frac{\sigma_k}{\sqrt{C}} \quad \forall k \in P$$

$$i.c_k = \left[\bar{s}_k \pm Z_{k, \frac{\alpha}{2}} \cdot \frac{\sigma_k}{\sqrt{C}} \right] \quad \forall k \in P$$

En este caso $s_{j,k}$ representa la distancia euclidiana del contrato $j \in \text{Contratos}(C)$ a su respectivo centroide perteneciente al clúster $k \in \text{Particiones}(P)$, \bar{s}_k el promedio de las distancias euclidianas de cada clúster $k \in \text{Particiones}(P)$, σ_k la desviación estándar muestral de las distancias euclidianas de cada clúster $k \in \text{Particiones}(P)$, $Z_{k, \frac{\alpha}{2}}$ el coeficiente de confianza del clúster $k \in \text{Particiones}(P)$ y, finalmente, $i.c_k$ el intervalo de confianza asociado a la distancia promedio \bar{s}_k para cada clúster $k \in \text{Particiones}(P)$. De esta manera, la categorización de un contrato estaría representada de la siguiente forma:

$$f_j = \begin{cases} 0 & \rightarrow s_{j,k} \leq \bar{s}_k + Z_{k, \frac{\alpha}{2}} \cdot \frac{\sigma_k}{\sqrt{C}} \\ 1 & \rightarrow \text{d.l.c} \end{cases}$$

Si $f_j = 0 \rightarrow \text{Contrato } j \text{ es normal}$

Si $f_j = 1 \rightarrow \text{Contrato } j \text{ es anormal}$

Donde f_j representa el atributo binario que toma valor de 0 cuando el contrato $j \in Contratos (C)$ se encuentra por debajo del límite superior del intervalo de confianza de su respectiva agrupación $k \in Particiones (P)$, y el valor de 1 (anormal) cuando su distancia euclidiana es mayor a dicho límite.

Análisis de redes

El análisis de redes tuvo como propósito, por un lado, lograr visualizar la estructura de la red de contratación pública estudiada y, por el otro, evaluar la posible existencia de una correlación entre el índice de riesgo creado y las dinámicas de contratación del PAE en la ciudad de Bogotá durante el periodo 2014 – 2017. El análisis se dividió en cuatro etapas en las que se explicarán los métodos y conceptos que se utilizaron para dar forma a esta investigación: red bipartita, proyecciones de una red bipartita, indicadores de centralidad de una red y homofilia.

Red bipartita

El contexto estudiado cuenta con cuatro categorías distintas de actores involucrados: entidades contratantes, representantes de las entidades contratantes⁵, contratistas y representantes de los contratistas⁶. Cuando se realiza un contrato, la entidad contratante, por medio de su representante, firma un acuerdo con el contratista (en caso de ser una persona natural) o con su representante (en caso de ser una persona jurídica). En este sentido, dos entidades o dos contratistas no pueden firmar un acuerdo entre ellos, generando que no sea posible crear una conexión clara entre contratistas o entre entidades. Para estos casos existe un tipo de modelamiento de redes denominado “*red bipartita*”. Este tipo de red es una “*representación natural en muchos sistemas complejos al describir relaciones entre dos tipos diferentes de actores o entidades*” (Cura & Vezub, 2012). En el marco de este estudio es posible realizar 6 combinaciones de redes bipartitas entre los cuatro tipos de actores presentes en cada contrato.

Un ejemplo de la situación explicada se muestra en la Ilustración 1, donde podríamos asumir que el primer conjunto de nodos (letras) y el segundo (números) corresponden a alguna combinación de los actores presentes en esta investigación.

⁵ Persona natural que representa de manera legal a una entidad o institución del Estado.

⁶ Persona natural que representa de manera legal a un persona jurídica o razón social.

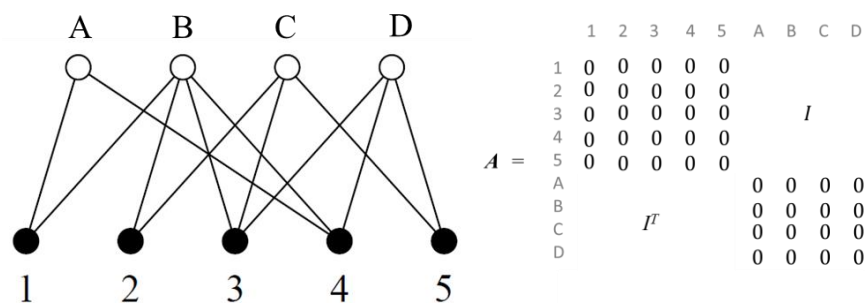


Ilustración 1: Ejemplo de red bipartita (izq.) con su respectiva matriz de adyacencia A (der.)⁷

Considerar la estructura de la red de contratación pública como el conjunto de combinaciones de una red bipartita permite que la visualización y el análisis de las dinámicas contractuales sean mucho más amplios, determinando la existencia o la no existencia de elementos correlacionales dentro de la estructura y la designación de los recursos pertenecientes al PAE.

Proyecciones de la red bipartita

La ventaja de estructurar la red de contratación estudiada como una red bipartita está en las proyecciones que es posible realizar. “A partir de un gráfico bipartito, siempre se obtendrán dos redes proyectadas” (Banerjee, Jenamani, & Pratihar, 2017). Las proyecciones son transformaciones de la red bipartita que logran conectar los actores de un mismo tipo: si ambos comparten una relación con un actor de tipo opuesto, se genera una conexión entre ellos⁸. Las proyecciones permiten realizar un análisis comparativo sobre cada uno de los tipos de actores que existen dentro de la red. La Ilustración 2 muestra las posibles proyecciones resultantes de una red bipartita.

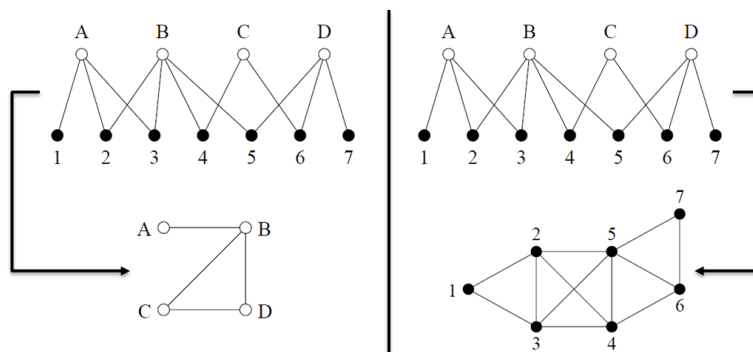


Ilustración 2: Proyecciones de la red bipartita respecto a los nodos “letra” y “número”⁹

⁷ http://materias.df.uba.ar/IRCBSa2016c2/files/2016/08/07_ConceptosBasicos.pptx

⁸ En el ejemplo de la Ilustración 2 el nodo C y D estarían conectados dado que comparten una conexión con el nodo 6.

⁹ http://materias.df.uba.ar/IRCBSa2016c2/files/2016/08/07_ConceptosBasicos.pptx

Se pueden realizar 12 proyecciones a partir de la estructura contractual presente en el SECOP dado que se tienen cuatro niveles de actores por cada contrato. Esta investigación decidió estudiar la proyección en la cual se descentraliza a mayor medida el número de contratos asociados a un actor y donde se puedan evidenciar las personas naturales detrás de los mismos. Esto quiere decir que se escogieron los dos conjuntos con mayor número de actores y sobre estos se realizó el posterior análisis. Por esta razón se evaluó, en primera instancia, la red bipartita que involucraba a los representantes de las entidades contratantes con los representantes de los contratistas, siendo la proyección sobre los representantes de los contratistas la que proporcionaría el mayor número de nodos y por tanto la mayor descentralización de la información contractual estudiada.

Indicadores de centralidad

Sobre una red, sea bipartita o no, se pueden realizar distintos tipos de análisis. Es posible estudiar la estructura de la proyección de la red de los representantes de los contratistas para entender si se encuentran densamente conectados entre sí, si la red está desconectada, si es posible identificar clústeres dentro de los nodos y conocer el número medio de conexiones entre los actores. La Tabla 2 muestra la utilidad de cada uno de los elementos mencionados frente al análisis sobre las dinámicas de contratación pública estudiadas.

Medida sobre la red	Fórmula	Rango	Interpretación
Densidad	$\frac{\sum_{i \in Rc} \sum_{j \in Rc j \neq i} r_{i,j}}{Rc \cdot (Rc - 1) / 2}$	[0,1]	Una alta densidad propone que los representantes de los contratistas tienen bastantes representantes de entidades en común.
Componentes conexos	Cn^*	[1, Rc]	En caso de ser mayor a 1 se entiende que la red está conformada por (n) componenetes no interconectados entre sí.
Modularidad	Md^{**}	[0,1]	Un grado muy alto propone que en la estructura de la red está compuesta significativamente por clústeres bastante definidos de representantes de contratistas.
Grado medio	$\frac{\sum_{i \in Rc} \sum_{j \in Rc j \neq i} r_{i,j}}{Rc}$	[0, Rc]	Un valor muy alto propone que, en promedio, los representantes de los contratistas comparten no sólo bastantes conexiones dentro de su clúster, sino que lo hacen con representantes de contratistas de otros clústeres de la red.

* El número de conexiones hace referencia al número de representantes de entidades que dos representantes de contratistas tienen en común.

Tabla 2: Medidas de centralidad evaluadas para las proyecciones de la red bipartita¹⁰

Donde $r_{i,j}$ toma el valor de 1 si existe una conexión entre los nodos $i, j \in \text{Representantes de Contratistas (Rc)}$ y de 0 en caso contrario. Estos indicadores de centralidad crearon un entendimiento general sobre el comportamiento de las dinámicas contractuales estudiadas y permitieron sacar conclusiones sobre aspectos puntuales propios de la asignación de recursos en la implementación del PAE en la ciudad de Bogotá durante los últimos cuatro años.

¹⁰ Cn^* : Robert Tarjan, Depth-First Search and Linear Graph Algorithms, in SIAM Journal on Computing 1 (2): 146–160 (1972).

Md^{**} : Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000.

Aparte de los indicadores generales de la red, es posible calcular ciertas medidas de centralidad sobre los nodos con el objetivo de conocer su importancia en las dinámicas relacionales de la red. En este sentido, se discutirán varias características propias de los nodos como se muestra en la siguiente tabla:

<i>Indicador del nodo</i>	<i>Rango</i>	<i>Interpretación</i>
<i>Grado</i>	$[0, R_c]$	<i>Número de representantes de entidad que un nodo comparte con los demás.</i>
<i>Grado con pesos</i>	$[0, R_c \cdot \max(\text{anormalidad})]$	<i>Número de representantes de entidad que un nodo comparte con los demás, teniendo en cuenta el grado de anormalidad del contrato realizado.</i>
<i>Fuerza</i>	$[0, R_c]$	<i>Número de veces que un representante de contratista comparte el mismo representante de entidad.</i>
<i>Closeness</i>	$[0, 1]$	<i>Grado promedio de cercanía de un representante de contratista a todos los demás representantes.</i>
<i>Betweenness</i>	$[0, 1]$	<i>Número de veces que el representante se encuentra en el camino más corto entre otros dos representantes.</i>
<i>Pagerank</i>	$[0, 1]$	<i>Probabilidad de que, partiendo de un representante de contratista al azar, se pase por el representante de contratista evaluado.</i>
<i>Clustering</i>	$[0, 1]$	<i>Grado en el cual un representante de contratista tiende a agruparse con otros representantes.</i>
<i>Eigencentrality</i>	$[0, 1]$	<i>Puntaje de un representante de contratista sobre su magnitud de número de conexiones y el número de conexiones que este tiene con nodos con alto grado de conexión.</i>

Tabla 3: Indicadores de centralidad para los nodos de la proyección bipartita

Las características propias de un nodo dentro de la red dan una idea más clara de su importancia y del rol que podría llegar a jugar en la estructuración de nuevas relaciones. Adicionalmente, gracias a estas características es posible comparar cada uno de los nodos con sus homólogos y con el resto de la red. Los impactos del análisis sobre las características de los nodos serán desarrollados en la sección de resultados.

Homofilia

Una vez se entiende la estructura de la red de contratación evaluada es necesario estudiar a mayor profundidad las posibles causas del comportamiento observado. Para esto existen varias herramientas de estudio, siendo una de las más representativas para este contexto la “homofilia”: “*tendencia de los individuos a formar asociaciones con individuos de antecedentes similares y componente estructural básico de gobierno de cualquier red social*” (Al-Qaheri & Banerjee, 2013). Esta, junto con su concepto antagónico “heterofilia”, logran determinar bajo un nivel de confianza establecido si las conexiones entre los actores de una red están asociadas a los atributos de dichos actores o, por el contrario, si son independientes a estas características. En otras palabras, el estudio propuesto busca verificar si las conexiones entre un conjunto de actores son aleatorias o si existe algún tipo de correlación entre ellas y alguno de los atributos de los involucrados.

Dado que los actores de la red que se quiso intervenir no contaban con un atributo propio¹¹, se decidió crear un atributo que los diferenciara de sus homólogos. En este sentido, cada uno de los representantes de los contratistas contaría con un atributo representado por el riesgo promedio de todos los contratos que hubiera realizado durante el periodo evaluado. El siguiente algoritmo muestra cómo se creó el atributo en mención:

$$\overline{q_r} = \frac{\sum_{j \in C} q_{r,j}}{g_r} \quad \forall r \in Rc$$

Donde $q_{r,j}$ representa el riesgo del contrato $j \in Contratos(C)$ realizado con el representante $r \in Representantes\ de\ Contratistas(Rc)$ y g_r , por su parte, representa el número de contratos asociados a dicho representante $r \in Representantes\ de\ Contratistas$. El grado de riesgo de cada representante de contratista sería entonces un número incluido dentro del rango $[0, 1]$: entre mayor sea dicho valor, se determina que el contratista realiza una mayor cantidad de contratos “riesgosos”¹² respecto a los categorizados como “normales”.

Una vez atribuido el grado de riesgo promedio para cada uno de los representantes, se procedió a clasificar dicho valor dentro de dos categorías (Tabla 4). La razón por la que se escogió 80% como la brecha entre ambas categorizaciones estuvo ligada al comportamiento de los datos en sí mismos y será explicada en la sección de resultados.

<i>Categoría Anormalidad promedio</i>	
<i>A</i>	<i>0% - 80%</i>
<i>B</i>	<i>mayor..</i>

Tabla 4: Categorización de riesgo promedio

Contando con la categorización del atributo de riesgo promedio para cada uno de los actores presentes en la red, fue posible comenzar el análisis de homofilia y heterofilia paralelamente. Con esto se buscó determinar si los representantes de los contratistas estaban creando más conexiones unos con otros a razón de su grado medio de riesgo o, por el contrario, si las dinámicas contractuales de los mismos carecían de un patrón aparente ligado a este atributo.

En primer lugar, se tomaron como punto de referencia algunos de los atributos de la red original¹³:

¹¹ Los atributos “plazo”, “valor”, “adiciones en el plazo”, “adiciones en el valor” y “calificación” son propios de los contratos mas no de los representantes de los contratistas.

¹² Los contratos riesgosos de este estudio son aquellos cuya distancia euclidiana a su centroide correspondiente es mayor al límite superior del intervalo de confianza creado para las distancias del clúster.

¹³ Proyección sobre “representantes de contratistas” de la red bipartita que involucra a “representantes de contratistas” y a “representantes de entidades”.

- 1) Proporción de nodos de categoría A y B sobre el total de nodos de la red
- 2) Número de conexiones entre nodos de la categoría A y B :
 - a. Número de conexiones ($A - A$)
 - b. Número de conexiones ($B - B$)
 - c. Número de conexiones ($B - A$)

Estos atributos permitieron que las réplicas realizadas sobre la red original fueran comparables con la misma y determinaran si en efecto la diferencia entre ellas recaía sobre el atributo categórico estudiado. Tomando elementos de la investigación realizada por Montes, Jimenez y Onnela, se realizaron 4.000 réplicas de la red original y se tomaron en cuenta únicamente las que cumplieron con las condiciones 1) y 2), dándole un margen de tolerancia del 10% a las conexiones entre los nodos de tipo A y B (numeral 2)).

Una vez obtenida la muestra final de las redes replicadas que cumplieron con las condiciones establecidas, se crearon tres listas de valores que guardaron en cada posición la siguiente división:

$$lista_{A-A} = \frac{O[A-A]}{R_i[A-A]} \quad lista_{B-B} = \frac{O[B-B]}{R_i[B-B]} \quad lista_{B-A} = \frac{O[B-A]}{R_i[B-A]}$$

Tal que $O[A - A]$ sea el número de conexiones entre los nodos (representantes) con atributo A de la red original, $R_i[A - A]$ el número de conexiones entre los nodos con atributo A de la red replicada $i \in \text{Réplicas Escogidas}$ y así sucesivamente para cada uno de los casos. Si la distribución de una de las listas generadas se comporta como una distribución normal con $media = 1$, quiere decir que la mayoría de las divisiones calculadas se aproximan a dicho valor. Esto propone que el tipo de conexión evaluada no genera un cambio significativo en la estructura de la red y, por consiguiente, no existiría evidencia estadística para asegurar que las dinámicas contractuales estarían siguiendo el patrón de conexión expresado por la lista: $lista_{A-A}$, $lista_{B-B}$ o $lista_{B-A}$ según sea el caso.

Para determinar entonces si existe o no evidencia estadística que logre afirmar si hay homofilia o heterofilia en la red evaluada, es necesario plantear las siguientes hipótesis:

$H_{0,i}$: la $lista_i$ sigue una distribución normal con $\mu = 1$

$H_{1,i}$: $\neg H_{0,i}$

La manera en la que se determinó la negación o la aprobación de la hipótesis nula para cada uno de los casos fue mediante la observación del $p - value$ asociado a cada uno de los procesos¹⁴: si el $p - value$ fuese menor al grado de significancia $\alpha = 5\%$, querría decir

¹⁴ “La estadística del p -value informará la probabilidad de que el caso observado se deba al azar, suponiendo que la hipótesis nula es verdadera.” (Figueiredo Filho, y otros, 2013).

que habría suficiente evidencia estadística para rechazar $H_{0,i}$ y, por tanto, afirmar que existe homofilia o heterofilia dependiendo del caso evaluado¹⁵.

Resultados

Recolección y limpieza de datos

Entre la recolección de la información contractual del PAE en la ciudad de Bogotá se encontraron distintas tipologías pertenecientes tanto al proceso en general como al contrato (Anexo 1).

Cada una de las combinaciones entre tipo de proceso y tipo de contrato presentes en la base de datos estudiada estaba definida tanto por su tipología como por sus atributos cuantificables. Dichos atributos, como se mencionó en la sección de métodos, hacían referencia a la temporalidad y a la cuantía de dinero involucradas en cada contrato. Se revisaron alrededor de 1.500 documentos digitales, verificando que toda la información allí presentada coincidiera con la reportada en el SECOP y se resaltó cada campo de los contratos que presentara inconsistencias respecto al documento original. Realizando la verificación de la información mencionada se decidió crear una tabla para determinar si un campo podía ser o no considerado como un error grave de tabulación.

Campo	Campo	Margen de tolerancia del 2%
ID de la entidad	Fecha de la firma del contrato	SI
Nombre de la entidad	Fecha de inicio del contrato	SI
ID del representante de la entidad	Fecha de fin del contrato	SI
Nombre del representante de la entidad	Plazo del contrato en días	SI
ID del contratista	Adiciones en días	SI
Nombre del contratista	Cuantía del contrato	NO
ID del representante del contratista	Adiciones en cuantía	NO
Nombre del representante del contratista	Valor total del contrato	NO

Tabla 5: Campos asociados a errores de tabulación

Lo anterior quiere decir que si un contrato presentaba algún tipo de error en cualquiera de los campos de la Tabla 5 (fuera del margen de tolerancia asignado según el caso), sería catalogado como un contrato con errores graves de tabulación. Por otro lado, si presentaba errores no estipulados en dicha tabla, o incluidos dentro del rango de tolerancia, no sería categorizado como un contrato con errores graves. Estos últimos fueron utilizados únicamente para realizar el análisis de inconsistencias de tabulación. Siguiendo esta metodología se encontró que el 76% de las fechas reportadas en los 484 contratos no eran consistentes, que el 36% de los plazos reportados no coincidían con lo fijado en los documentos originales (pero estaban dentro del rango de tolerancia) y que, entre todos los

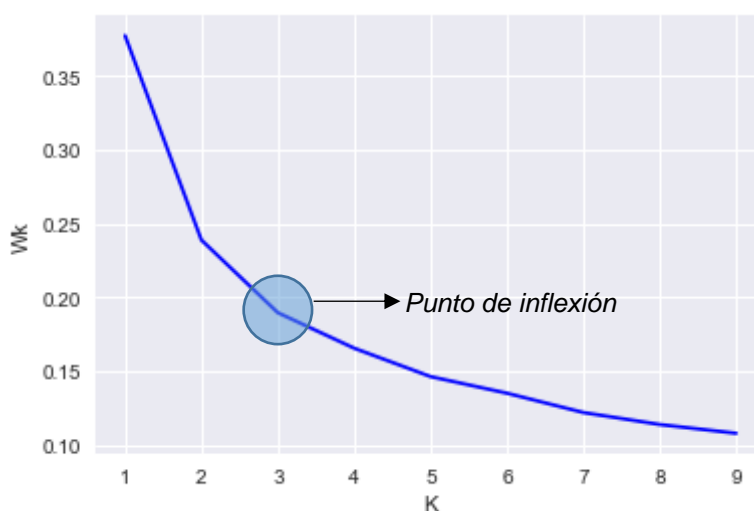
¹⁵ $(A - A)$ y $(B - B)$ son casos de homofilia, mientras que $(B - A)$ es un caso de heterofilia; todo esto si se rechaza la hipótesis nula en cada uno de los casos.

números de identificación y sus tipologías, el 22% de dichos datos presentaron inconsistencias (Anexo 3). Cabe resaltar que únicamente el 5.2% de los contratos reportados en el SECOP presentaron errores graves (22 contratos).

Índice de riesgo contractual

Agrupamiento

Como se mencionó en el apartado de métodos, se implementó el “*método de codo*” para identificar el número de grupos más adecuado para la segmentación los contratos según el valor de sus atributos. De esta manera, se quiso representar la inercia calculada a medida que se variaba el número de grupos (particiones) k dentro del rango $k = [1, \dots, 9]$ (Gráfica 1) (Kodinariya & Makwana, 2013).



Gráfica 1: Inercia (Wk) respecto a número de particiones (k)

Observando el comportamiento de la inercia total calculada, se pudo determinar que el número de particiones más adecuado para segmentar la muestra de contratos era de $k = 3$, donde cada una de las tres particiones creadas de manera no supervisada tuvo una característica que la diferenciaba de las demás¹⁶. Se realizó una regresión logística para cada uno de los clústeres creados buscando entender qué atributo o atributos jugaban un papel importante en la caracterización de cada clúster (Anexo 2). Un análisis de regresión logística “*estudia la asociación entre una variable dependiente categórica y un conjunto de variables independientes (explicativas)*” (NCSS Statistical Software, 2018). La siguiente tabla agrupa y resume los resultados de las regresiones logísticas obtenidas en cada caso, mostrando la relación que cada uno de los clústeres tiene respecto a los atributos de los contratos:

¹⁶ El “*método de codo*” propone escoger el número de particiones k correspondiente al punto de inflexión de la gráfica (k vs. *inercia*).

VARIABLES INFLUYENTES	PLAZO	VALOR	ADICION PLAZO	ADICION VALOR
CLUSTER # 0	+	-		-
CLUSTER # 1	-	-		+
CLUSTER # 2	-	+		

Tabla 7: Relación proporcional (+), relación inversamente proporcional (-)

Lo anterior quiere decir que, a mayor plazo, la probabilidad de que un contrato pertenezca al clúster # 0 aumentaría; a mayor valor, mayor la probabilidad de pertenecer al clúster # 1 y a mayor cantidad de adiciones en el valor, mayor la probabilidad de pertenecer al clúster # 2. El Anexo 5 expone las proporciones de cada atributo dentro del grupo creado, evidenciando las características propias de cada uno de los clústeres, dando cuenta de los resultados presentados en la Tabla 7.

INSTITUCIÓN	CLUSTER # 0	CLUSTER # 1	CLUSTER # 2
Secretaría de Educación	32%	54%	13%
Gobernación de Cundinamarca	38%	31%	31%
ICBF	24%	59%	18%
MEN	43%	22%	36%

Tabla 8: Porcentaje de contratos de cada institución pertenecientes ubicado en cada clúster

Como se evidencia en la Tabla 8, podemos afirmar entonces que algunas de las instituciones evaluadas tienen un mecanismo particular de contratación. Aquellas que presentan una forma de contratación más marcada son la Secretaría de Educación y el ICBF, dado que cuentan con más del 50% de sus contratos concentrados en uno de los clústeres creados por el algoritmo “*k-medias*”. Esto, por otra parte, nos permite entender que la forma en la que el Estado realiza sus contratos mediante las diferentes instituciones es sesgada y que depende en gran parte de la institución y el representante involucrados en el proceso.

Clasificación

Se estipuló anteriormente que la categorización de un contrato como “*riesgoso*” o “*normal*” dependería de si su distancia euclidiana a su centroide respectivo se encontraba fuera o dentro del límite superior del intervalo de confianza calculado.

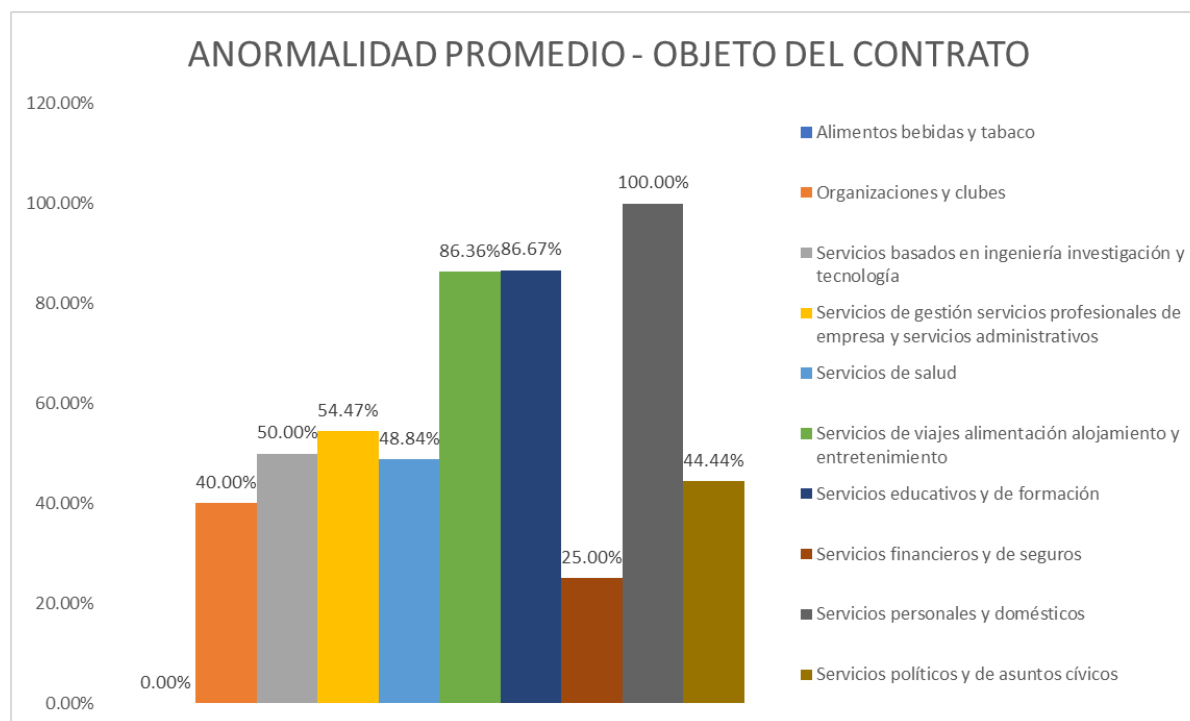
Clúster	Contratos	Límite superior (i.c.)	Nodos dentro del límite	Nodos dentro del límite (%)
0	181	0.212	108	59.7%
1	181	0.194	87	48.1%
2	122	0.269	15	12.3%

Tabla 6: Características de los clústeres creados

La Tabla 6 muestra el comportamiento de los clústeres que se crearon a partir de la implementación del algoritmo “*k-medias*”. En ella se evidencia que en 2 de los 3 grupos el

número de contratos considerados como “*normales*” oscilaba entre el 50 – 60%, mientras que en el tercero (clúster # 2) el porcentaje de contratos “*riesgosos*” aumentaba hasta casi un 88%, a pesar de que el límite superior de su intervalo de confianza era el mayor de todos los grupos.

A fin de entender lo que el índice de riesgo representa sobre la tipología de los contratos fue preciso analizar el promedio de riesgo para cada tipo de proceso, objeto del contrato y tipo de contrato. Para esto, se graficó el promedio de los riesgos en cada uno de los casos (Anexo 4). La Gráfica 2 muestra el riesgo promedio calculado para cada objeto del contrato, evidenciando que los asociados a “*servicios personales y domésticos*”, “*servicios financieros y de seguros*” y “*servicios educativos y de formación*” son los que cuentan con una riesgo promedio más alto.



Gráfica 2: Riesgo promedio calculada para cada objeto de contrato

Las tres tipologías resaltadas anteriormente acumulan el 7.2% de los contratos totales de la muestra. Por esta razón, y dado que el tamaño de la base de datos evaluada es tan reducido (Mindrila & Balentyne, 2012), no se propuso realizar un análisis sobre la posible correlación entre la tipología del contrato y su índice de riesgo.

Análisis de redes

Red bipartita

Tal como se mencionó anteriormente, era posible crear 6 redes bipartitas a raíz del contexto estudiado (Tabla 9). Sin embargo, como muchos de los contratistas eran personas jurídicas se crearon únicamente 3: ID (1), ID (3) e ID (5). En el Anexo 6 se encuentra la representación gráfica de las redes creadas.

Redes Bipartitas		
ID	Nodo tipo 1	Nodo tipo 2
1	Entidades	Representantes de entidades
2	Entidades	Contratistas
3	Entidades	Representantes de contratistas
4	Representantes de entidades	Contratistas
5	Representantes de entidades	Representantes de contratistas
6	Contratistas	Representantes de contratistas

Tabla 9: Combinaciones de redes bipartitas

La red bipartita estudiada se conformó por 484 conexiones (contratos) entre dos tipos de nodos: representantes de entidades (22) y representantes de contratistas (248) (Ilustración 3). En dicha red, el tamaño de los nodos representa el número de conexiones que cada uno tiene y su color representa la entidad con la que realiza la mayor parte de sus contratos. El grosor de los arcos representa el número de contratos realizados entre un par de nodos y el color de estos el promedio de su grado de riesgo (entre más oscuro, mayor riesgo promedio).

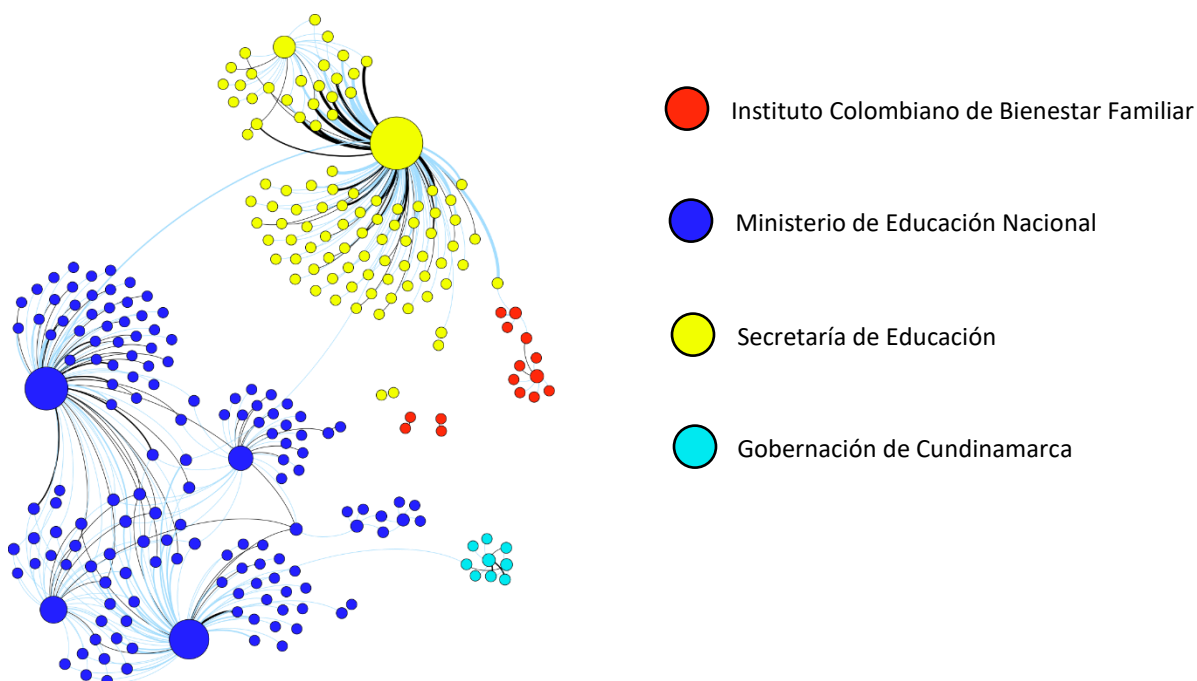


Ilustración 3: Red bipartita entre representantes de entidades y representantes de contratistas

Mediante esta red se puede decir que las entidades más representativas para la contratación del PAE en Bogotá son el Ministerio de Educación Nacional y la Secretaría de Educación. La primera cuenta con un mayor número de representantes asociado, mientras que la segunda cuenta con un mayor número de conexiones riesgosas con sus representantes. Aparte de esto, podemos observar que existe una conexión entre un representante del Ministerio de Educación y uno de la Secretaría de Educación, generando que la red de contratación deje de estar desconectada entre sí y abriendo las puertas a una investigación futura sobre este fenómeno.

Sobre esta red no se calcularon medidas de centralidad por su desconexión entre componentes y debido a que hay en ella dos tipos de nodos cuyas medidas no serían comparables entre sí. Sin embargo, cuando se realice la proyección correspondiente, el análisis de la red podrá ser consistente con el tipo de nodo evaluado y con sus conexiones.

Proyección de la red bipartita

La proyección de la red bipartita estudiada correspondió a la proyección sobre los nodos de los representantes de los contratistas, siendo las conexiones entre ellos los representantes de entidades que cada par tuviera en común. Esta red comprendió 248 nodos (representantes) y 2.658 conexiones entre sí (Ilustración 4). Este elevado número de conexiones se genera porque varios representantes de contratistas han realizado un contrato con el mismo representante de entidad, creando una conexión entre todos ellos.

El color del nodo representa la entidad con la cual cada uno de estos realiza los contratos¹⁷. En la mayoría de los casos dichos nodos tienen una relación particular con una de las cuatro entidades institucionales mencionadas, pero uno de ellos (nodo blanco) tiene una relación con dos de estas¹⁸. El tamaño del nodo hace referencia al número de conexiones que este tiene con los demás nodos de la red y el color de las aristas representa el grado de riesgo promedio entre los nodos: si un representante de contratista (*A*) realizó un contrato con un representante de entidad (*C*) con un *indicador de riesgo* = 1, y un representante de contratista (*B*) realizó un contrato con el mismo representante de entidad (*C*) con un *indicador de riesgo* = 2, la conexión que existirá entre *A* y *C* tendrá un *indicador de riesgo* = 1.5.

¹⁷ Los nodos representados con el color negro no tienen un promedio de riesgo contractual significativo.

¹⁸ Este comportamiento se evidencia en la red bipartita mostrada en el Anexo 7.

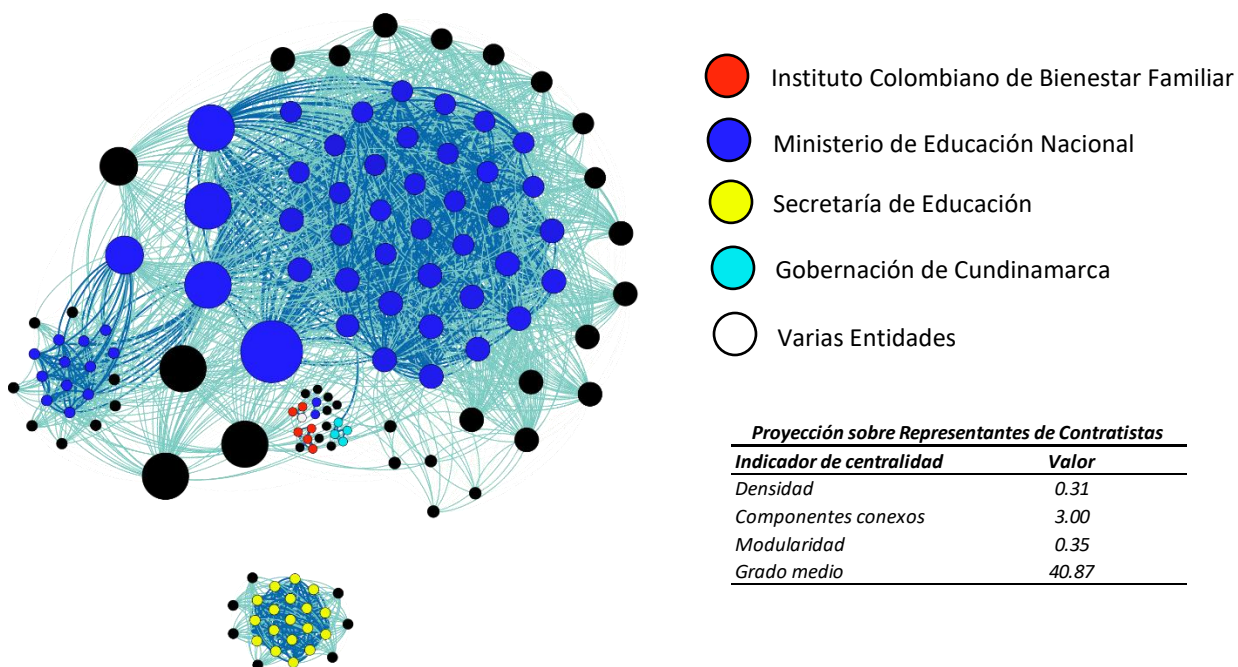
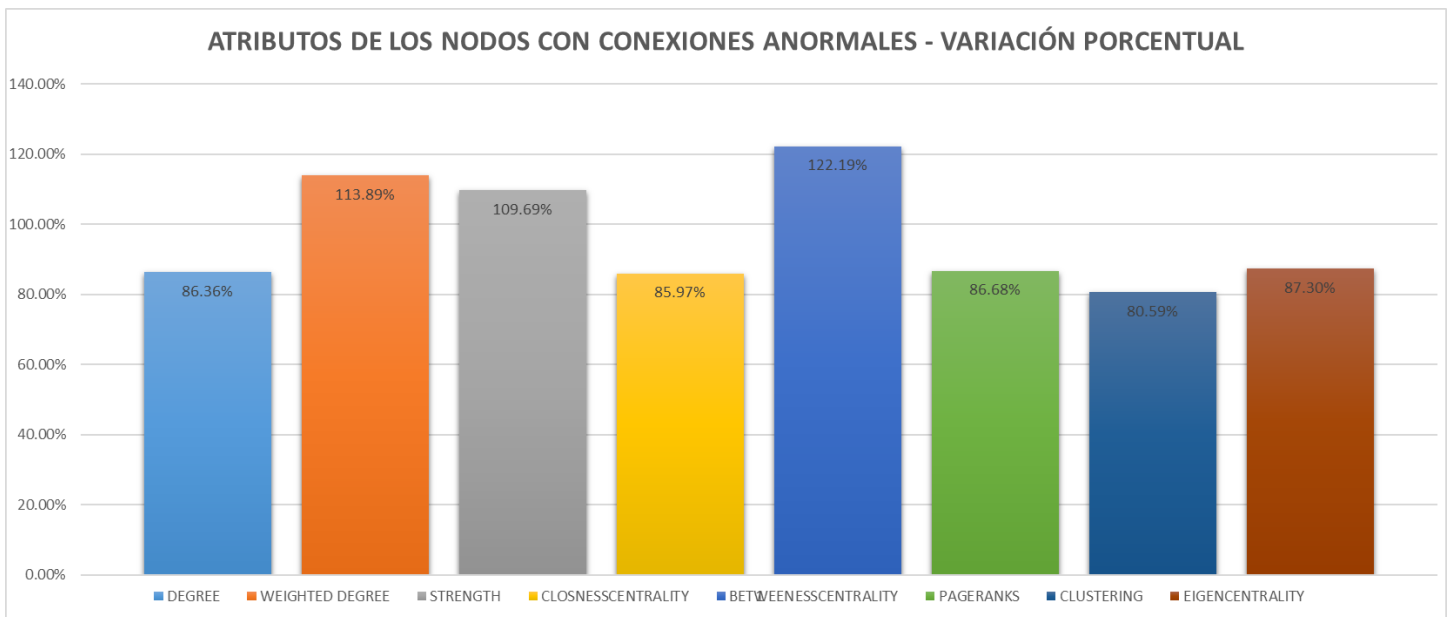


Ilustración 4: Proyección sobre los representantes de los contratistas de la red bipartita evaluada

Como se puede observar en la Ilustración 4, la red cuenta con 3 componentes conexos que se podrían atribuir a las diferentes entidades institucionales presentes en el estudio, donde dos de ellas (ICBF y Gobernación de Cundinamarca) estarían conectas entre sí, formando un solo grupo. Podemos observar que los nodos tienen en promedio 40.87 conexiones entre sí y que, a pesar de que la densidad de la red es medianamente baja (*densidad* = 31%), se encuentra altamente conectada. Visualmente se podría decir que los nodos más centrales de la red son aquellos que cuentan con un mayor número de conexiones “riesgosas”, lo cual se discutirá a continuación.

Indicadores de centralidad

Para realizar un análisis preciso sobre las cualidades de los nodos que tienen al menos una conexión más anormal que los demás (conexiones más oscuras), se extrajeron los nombres de los representantes de contratistas y sobre ellos se aplicaron cada uno de los indicadores concernientes (Tabla 3). Se resaltaron 85 nodos de los cuales se calculó el promedio de cada uno de sus indicadores y se comparó la variación porcentual respecto al promedio global de cada uno de ellos (Gráfica 3).



Gráfica 3: Variación porcentual de los atributos de los nodos encontrados con conexiones riesgosas

Es evidente que la variación porcentual de los atributos evaluados respecto a su promedio en la red es muy alta, tanto así que el coeficiente de clustering es el que cuenta con la variación más baja respecto a la media general del sistema y su valor es de 80.59%. Más adelante se discutirán estos resultados, tomando en cuenta los diferentes factores que podrían estar influenciando este comportamiento.

Homofilia

Al realizar el promedio de la calificación \bar{q}_r para cada uno de los nodos $r \in \text{Representantes de Contratistas } (R_c)$, se encontró que 104 (42%) de las observaciones estaban por encima del 80% de riesgo (Tabla 7). Se escogió este porcentaje como brecha para la categorización, porque de haber fijado una brecha de menor porcentaje no habría sido posible evaluar objetivamente el comportamiento de los nodos con altos grados de riesgo. Por otro lado, la determinación de una brecha mayor al 80% habría incluido nodos con un riesgo promedio demasiado alto como para poder considerarlos libres de riesgo contractual.

<i>Categoría</i>	<i>Anormalidad Promedio</i>	<i>Frecuencia</i>	<i>Porcentaje</i>
A	0 - 80%	144	58%
B	mayor..	104	42%

Tabla 7: Distribución de grado de riesgo promedio para los representantes de los contratistas

Habiendo categorizado los nodos según el rango de su riesgo promedio calculado (\bar{q}_r), se procedió a realizar el análisis de homofilia y heterofilia correspondiente. Los resultados, según lo muestra el Anexo 7, mostraron que existe evidencia estadística para afirmar (bajo

un nivel de significancia del 95%) que existe homofilia entre los nodos de categoría *B* y heterofilia entre los nodos de tipo *B* y tipo *A*. Sin embargo, no existe evidencia estadística para afirmar que hay presencia de homofilia entre los nodos de tipo *A*, a pesar de que conforman el 60% de la red. Lo anterior propone que los representantes de los contratistas con un alto grado de riesgo contractual tienden a generar conexiones con representantes de contratistas con igual grado de riesgo; también generan conexiones con representantes de contratistas con un grado bajo de riesgo. Por el contrario, se puede observar que los representantes de contratistas categorizados con un grado medio o bajo de riesgo contractual no tienden a conectarse entre sí. La interpretación de dichos resultados debe ser extrapolada al contexto original evaluado, es decir a la red bipartita mencionada.

Al observar estas dinámicas relacionales en la red bipartita se puede afirmar que existe evidencia estadística para proponer que los representantes de contratistas con un grado alto de riesgo contractual tienden a crear conexiones con representantes de entidades que hayan realizado contratos con representantes de contratistas con un grado de riesgo igual de alto. De la misma manera, es posible afirmar que los representantes de contratistas con un grado bajo de riesgo contractual tienden a crear conexiones con representantes de entidades que hayan realizado contratos con representantes de contratistas con un grado alto de riesgo.

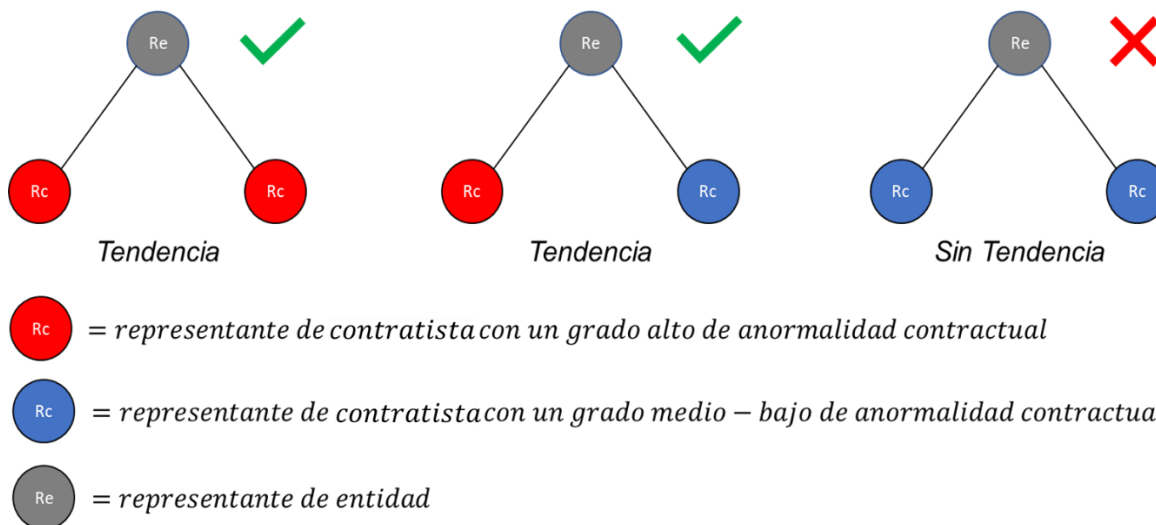


Ilustración 5: Explicación ilustrativa de las dinámicas de homofilia y heterofilia halladas en análisis

La Ilustración 5 muestra de manera gráfica cuáles fueron los resultados obtenidos por el análisis de homofilia y heterofilia, resaltando las interacciones en las cuales se encontró un patrón de conexión aparente entre los nodos.

Discusión

A partir de los resultados obtenidos en el segmento anterior, es posible realizar varias conclusiones empezando por la recolección de datos de la investigación. En teoría, las bases de datos donde se encuentra contenida la información contractual del país (SECOP) reflejan transparentemente las transacciones realizadas por el Estado colombiano y *“deben disponer de información confiable para la toma de decisiones de adquisiciones gubernamentales”* (Ministerio de Comunicaciones, 2008). Sin embargo, es preciso decir que no son una fuente de información consistente cuando el propósito de estudio se basa en la creación de un índice de calidad, riesgo u otro tipo de evaluación que tenga como objetivo entender las dinámicas contractuales del país. No obstante, hay que resaltar que la existencia del SECOP representa el primer paso hacia una economía más transparente, permitiendo la apertura de varias ramas de investigación en temas relacionados con corrupción, dinámicas contractuales y que se enfoquen en la creación de herramientas para la supervisión de la designación de los recursos gubernamentales.

Estas bases de datos del Estado brindan una gran cantidad de información que no es sencilla de procesar. Al tratar de unificarla y omitir los errores de tabulación presentes en los contratos extraídos, fue necesario recurrir a la estandarización y normalización de datos. Esto se hizo debido al tamaño de la muestra que, al ser tan reducida, cuenta con un alto grado de variabilidad entre sus datos. Se evaluaron contratos con valores que oscilaban en un rango de \$64.000.000.000 COP y cuyos plazos variaban entre los 12 y los 985 días. Por esta razón, es posible suponer entonces que, al existir un rango tan amplio de valores para los atributos mencionados dentro de una muestra de 484 contratos, la variabilidad entre sus características será significativa. Por ende, a menor cantidad de datos presentes en el estudio, mayor el grado de complejidad para su evaluación.

En lo concerniente a la agrupación de los datos, se debe mencionar que el algoritmo que se utilizó es dinámico pues se sirve de métodos heurísticos para la determinación de las particiones de los datos (Kanungo, y otros, 2004). Esto trae consigo implicaciones positivas y negativas simultáneamente. En primer lugar, el razonamiento de la metodología busca minimizar la inercia total del sistema y por tanto es posible que la partición generada ubique varios contratos demasiado lejos de su centroide; en otras palabras, que los considere como riesgosos. A medida que el número de contratos de la muestra aumenta, el modelo adquiere la capacidad de reducir sus errores (Mindrila & Balentyne, 2012), delimitando las particiones de manera más eficiente y precisa. Implementar un algoritmo de agrupación no supervisada independiza los datos y permite que encuentren la mejor manera de segmentarse dentro del sistema. De esta forma, se evita incluir ponderadores u otro tipo de metodologías que creen un entorno en el cual el evaluador tome partida y donde se abra la posibilidad de presentar resultados sesgados. En este sentido, uno de los aportes brindados por la investigación recae en la entrega de una metodología capaz de funcionar autónomamente, asignando un índice de riesgo que considera únicamente el

comportamiento de los datos y donde el rol del evaluador se limite al análisis de los resultados.

Según la categorización, agrupación y clasificación de los contratos de la muestra, se advierte que no más del 60% fue encontrado como “*normal*” (Tabla 6). Este resultado no busca proponer que la calidad de contratación pública frente a la implementación del PAE sea baja, puesto que la muestra escogida para el análisis no logra reproducir las dinámicas de designación de recursos del PAE en el territorio Nacional. Por dicho motivo, solo se afirma que existe un gran porcentaje de contratos que no se encuentra dentro de un intervalo de confianza calculado dependiendo del grupo al que pertenecen. Al contar con una muestra de 484 contratos, es preciso ser cuidadoso con la interpretación de los resultados que se obtengan puesto que la variabilidad de las distancias de los contratos a su centroide puede ser muy alta. A pesar de esta condición, se lograron poner en evidencia las tipologías de contratos con altos índices de riesgo (Gráfica 2): “*servicios personales y domésticos*”, “*servicios financieros y de seguros*” y “*servicios educativos y de formación*”. No obstante, estas tipologías solo logran almacenar un pequeño porcentaje de los contratos totales de la muestra y no sería correcto proponer la existencia de una correlación entre la tipología del contrato con su grado de riesgo.

A pesar de que el grado de riesgo de cada uno de los contratos no pudo ser relacionado con su tipología (dado el tamaño de la muestra), el análisis de redes proporcionó un acercamiento distinto a esta situación. Fue posible visualizar de diferentes maneras la estructura de contratación del PAE, reconociendo la existencia de personas que realizan contratos con más de una entidad a la vez. De esta manera, se pudo observar un caso donde dos entidades estatales realizaban un contrato entre ellas por medio de sus representantes. Por otra parte, se evidenció que la posición de los actores en la red sí tiene una influencia directa con el índice de riesgo calculado. Esto quiere decir que entre mayor sea el número de contratos realizados con un representante de entidad altamente conectado o central en la red, mayor la probabilidad de que dicho contrato sea categorizado como anormal. Sin embargo, el análisis propuesto en este estudio no evaluó los contratos de forma particular y por ende no se realizó una mayor profundización en varios de los hallazgos anteriormente mencionados.

Un elemento importante que surge de la investigación es el reconocimiento de patrones de conexión entre los entes institucionales y las personas con las que se realizan los procesos contractuales. Las dinámicas de contratación tienden a aumentar el riesgo conjunto de la red, evidenciando que existe un ciclo de contratación dado a la volatilidad de las características de los contratos. Esto, por una parte, quiere decir que los contratos tienden a suscribir adiciones no solo en tiempo sino en cuantía, y que su porcentaje tiende a variar más a medida que observamos personas que han realizado un gran número de contratos con el Estado. A raíz del análisis de la proyección sobre la red bipartita fue posible encontrar que la forma de contratación de las cuatro entidades no era independiente, entendiendo que existen personas que realizan contratos con una u otra a través del tiempo.

Conclusiones

Es posible concluir que, si bien el análisis no logra dar una idea general sobre la efectividad de la implementación del PAE en Colombia, sus hallazgos evidencian que es posible replicar la metodología sobre este y otros contextos a nivel nacional. Por una parte, se cumplió con uno de los objetivos generales, que establecía la creación de una metodología para el análisis de los riesgos en la contratación pública del país. Por otro lado, se logró hacer que dicha metodología se convirtiera en un algoritmo automatizado que itera sobre los contratos sin la necesidad de ser intervenida.

Sin embargo, el procedimiento está limitado por la verificación de los contratos, que se debe realizar previo al análisis. Es necesario denunciar la metodología utilizada para la tabulación de datos contractuales por parte del Estado, y se debe advertir a las futuras investigaciones sobre las precauciones necesarias a la hora de trabajar con una base de datos descargada desde el portal web del SECOP.

Se recomienda, paralelamente, que se trabajen bases de contratos más amplias (Mindrila & Balentyne, 2012) con el fin de reducir la alta dispersión de los datos por atributo y así poder ser agrupados de manera más eficiente. De igual manera, y al contar con una mayor cantidad de información para procesar, se deberían evaluar diferentes formas de estandarización de datos y lograr incluir la mejor de ellas en el proceso de automatización para la creación del índice de riesgo. Por último, se podría flexibilizar el modelo con el fin de crear una escala que tenga un mayor número de categorías de riesgo para evitar la polarización de los contratos como “*normales*” o “*riesgosos*”. La distancia entre el límite superior de cada intervalo de confianza a sus contratos riesgosos puede ser traducida en términos de su número de desviaciones estándares. De esta manera, sería posible crear una escala de medición que permita matizar el riesgo de cada contrato y ser más específico en la clasificación y en el análisis de resultados.

Como trabajo futuro se ha planteado replicar la metodología a otros sectores más grandes que incluyan la tipología de “*terrenos, edificios, estructuras y vías*”, realizando un análisis de riesgo contractual sobre entidades como el Instituto Nacional de Vías (INVIAS) y la Agencia Nacional de Infraestructura (ANI) que han sido foco de investigación en varias ocasiones por casos de corrupción (Judicial, 2017) (EL ESPECTADOR, 2011).

Bibliografía

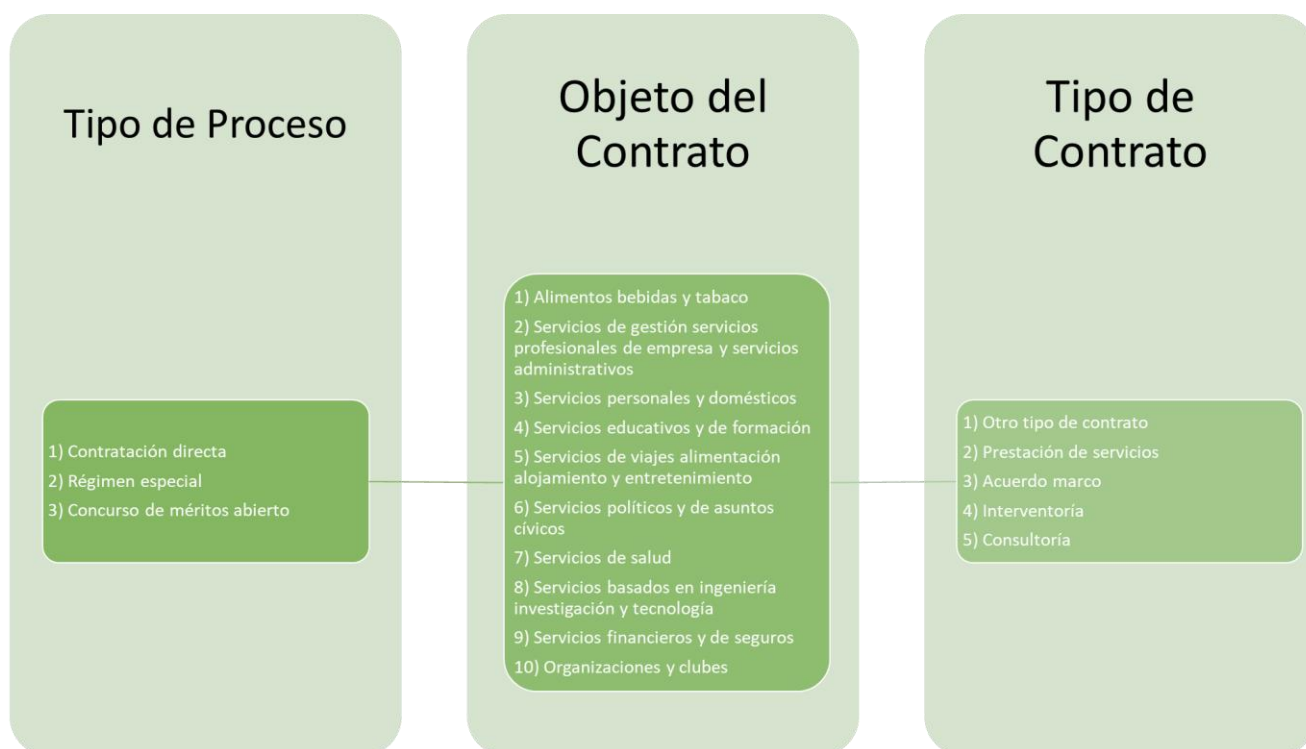
Al-Qaheri, H., & Banerjee, S. (2013). Measuring Homophily in Social Network: Identification of Flow of Inspiring Influence under New Vistas of Evolutionary Dynamics. *International Journal of Advanced Computer Science and Applications*, 14-24.

- Banerjee, S., Jenamani, M., & Pratihari, D. K. (2017). *Properties of a Projected Network of a Bipartite Network*. Kharagpur: Industrial and Systems Engineering at Indian Institute of Technology.
- Chang, E. C., & Chu, Y.-H. (2006). Corruption and Trust: Exceptionalism in Asian Democracies? *Journal of Politics*.
- Clavijo M., J. A., & Granada D., H. A. (2016). A Classification Technique With Categorical Variables. *Revista Ciencia en Desarrollo*, 15-20.
- Colombia Compra Eficiente. (2014). *Colombia Compra Eficiente*. Obtenido de <https://www.colombiacompra.gov.co/content/reduciendo-los-costos-de-transaccion-del-programa-de-alimentacion-escolar-en-bogota#main-content>
- Colombia Compra Eficiente. (19 de Enero de 2017). *SECOP*. Obtenido de <https://www.colombiacompra.gov.co/secop/secop-i>
- Cura, R. R., & Vezub, J. E. (2012). *Análisis de grafos bipartitos en función de múltiples proyecciones: aplicación a un caso de estudio histórico*. CENPAT - CONICET, UNPSJB.
- EL ESPECTADOR. (6 de Mayo de 2011). *EL ESPECTADOR*. Obtenido de <https://www.elespectador.com/content/denuncian-n%C3%B3mina-paralela-en-invias-por-m%C3%A1s-de-7-mil-millones>
- Fazekas, M., Tóth, I. J., & King, L. P. (Noviembre de 2013). Anatomy of grand corruption: A composite corruption risk index based on objective data. *Corruption Research Center Budapest*.
- Figueiredo Filho, D. B., Paranhos, R., Da Rocha, E. C., Batista, M., Da Silva, J. A., Wanderley D. Santos, M. L., & Marino, J. G. (2013). When is statistical significance not significant? *Brazilian Political Science Review*, 31-55.
- Hoyos Catañeda, I. M. (2015). *Informe sobre el seguimiento al programa de alimentación – pae a nivel nacional a 31 de diciembre 2015 de la procuraduría delegada para la defensa de los derechos de la infancia, la adolescencia y la familia*. Bogotá D.C.
- Jain, Y. K., & Bhandare, S. K. (2011). Min Max Normalization Based Data Perturbation Method for Privacy Protection. *International Journal of Computer & communication Technology*, 45-50.
- Judicial, R. (28 de Agosto de 2017). *EL ESPECTADOR*. Obtenido de <https://www.elespectador.com/noticias/judicial/odebrecht-sigue-por-ahora-con-la-ruta-del-sol-2-articulo-710408>
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2004). A local search approximation algorithm for k-means clustering. *Computational Geometry*, 89-112.
- Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 90-95.
- LA FM. (26 de Abril de 2018). *LA FM*. Obtenido de <https://www.lafm.com.co/colombia/sic-nuevo-cartel-de-la-contratacion-estatal-podria-seguir-contratando-con-el-estado>
- Ławrynówicz, A., & Volker, T. (2014). *Introducing Machine Learning*. Poznan: Institute of Computing Science, Poznan University of Technology.

- Martinez, T. M. (2014). Combating Corruption Through Transparency: How Have e-Government Investments Affected Levels of Accountability in Brazil? *ICEGOV2014*, 528-531.
- McGee, R., & Gaventa, J. (2010). *Review of Impact and Effectiveness of Transparency and Accountability Initiatives: Synthesis Report*.
- Mindrila, D., & Balentyne, P. (2012). The Basic Practice of Statistics. En D. Moore, W. I. Notz, & M. A. Fligner, *The Basic Practice of Statistics* (pág. Cap. 19).
- Ministerio de Comunicaciones. (2008). *Sistema Electrónico para la Contratación Pública – SECOP*. Bogotá, D.C.
- Montes, F., Jimenez, R. C., & Onnela, J.-P. (s.f.). Connected but Segregated: Social Networks in Rural Villages.
- Morales Moreno, L. M., Torres Páez, J. O., & Serrano Martínez, A. (2014). A Government IT-focused e-Procurement Strategy for Colombia. *ICEGOV2014*, 446-449.
- NCSS Statistical Software. (2018). *Logistic Regression*.
- Silva, M. (28 de Abril de 2015). *EL TIEMPO*. Obtenido de <http://www.eltiempo.com/archivo/documento/CMS-15648817>
- University of Pittsburgh. (2003). *Applied Statistical Methods*. Pittsburgh: University of Pittsburgh.

Anexos

Anexo 1: Tipología de los contratos



Anexo 2: Características de los clústeres

Regresión logística: Clúster # 0

	coef	std err	z	P> z	[0.025	0.975]
PLAZO (CLASE)	36.7911	5.630	6.535	0.000	25.757	47.825
VALOR (CLASE)	-6.1700	1.685	-3.662	0.000	-9.473	-2.867
A. TIEMPO (CLASE)	6.1203	11.650	0.525	0.599	-16.714	28.955
A. DINERO (CLASE)	-29.6928	12.913	-2.300	0.021	-55.001	-4.385

Regresión logística: Clúster # 1

	coef	std err	z	P> z	[0.025	0.975]
PLAZO (CLASE)	-68.5674	22.626	-3.030	0.002	-112.914	-24.221
VALOR (CLASE)	-87.8731	31.428	-2.796	0.005	-149.471	-26.275
A. TIEMPO (CLASE)	-42.8001	25.264	-1.694	0.090	-92.317	6.717
A. DINERO (CLASE)	128.8406	49.889	2.583	0.010	31.059	226.622

Regresión logística: Clúster # 2

	coef	std err	z	P> z	[0.025	0.975]
PLAZO (CLASE)	-12.0817	1.479	-8.171	0.000	-14.980	-9.184
VALOR (CLASE)	14.7587	1.668	8.847	0.000	11.489	18.028
A. TIEMPO (CLASE)	-2.9323	9.186	-0.319	0.750	-20.936	15.071
A. DINERO (CLASE)	-1.0066	9.283	-0.108	0.914	-19.201	17.188

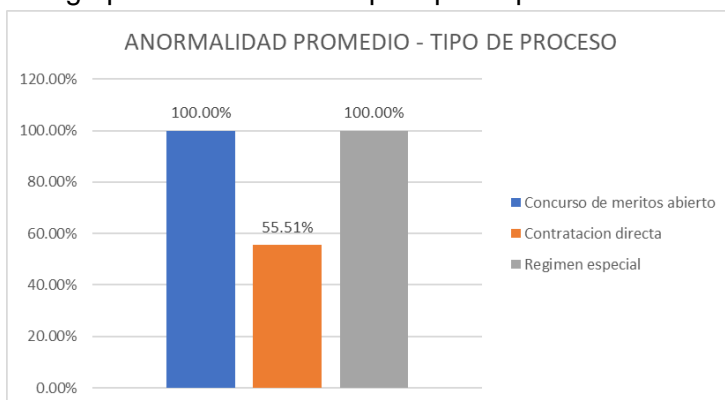
Anexo 3: Inconsistencias en los contratos

Representación gráfica de las inconsistencias encontradas en la verificación de información



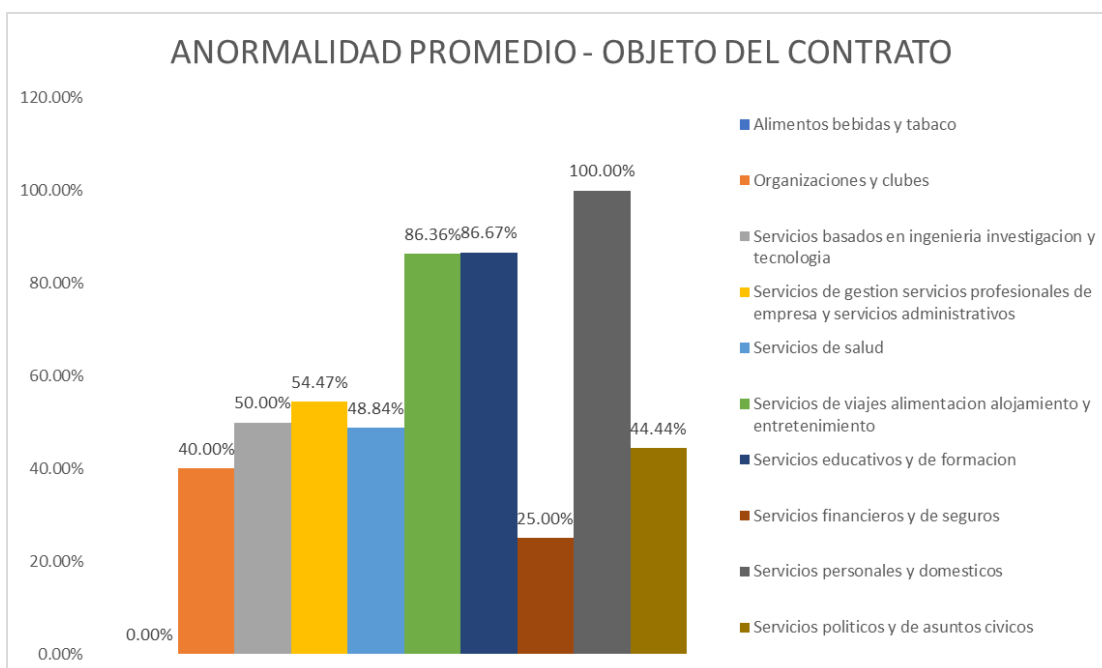
Anexo 4: Riesgo promedio por tipología de contrato

Riesgo promedio calculada por tipo de proceso contractual



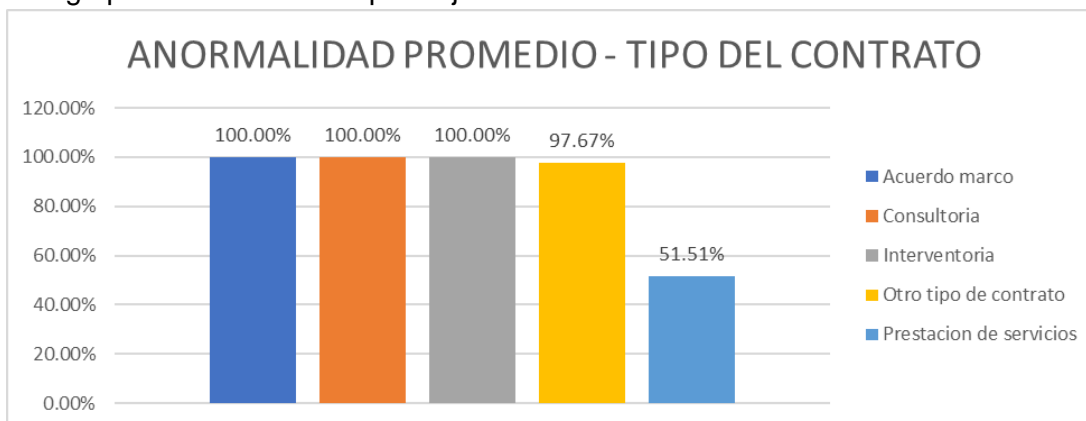
<i>Tipo del proceso</i>	<i>Anormalidad promedio</i>
<i>Concurso de méritos abierto</i>	<i>100.00%</i>
<i>Contratación directa</i>	<i>55.51%</i>
<i>Régimen especial</i>	<i>100.00%</i>

Riesgo promedio calculada por objeto del contrato



Objeto del contrato	Anormalidad promedio
Alimentos bebidas y tabaco	0.00%
Organizaciones y clubes	40.00%
Servicios basados en ingeniería investigación y tecnología	50.00%
Servicios de gestión servicios profesionales de empresa y servicios administrativos	54.47%
Servicios de salud	48.84%
Servicios de viajes alimentación alojamiento y entretenimiento	86.36%
Servicios educativos y de formación	86.67%
Servicios financieros y de seguros	25.00%
Servicios personales y domésticos	100.00%
Servicios políticos y de asuntos cívicos	44.44%

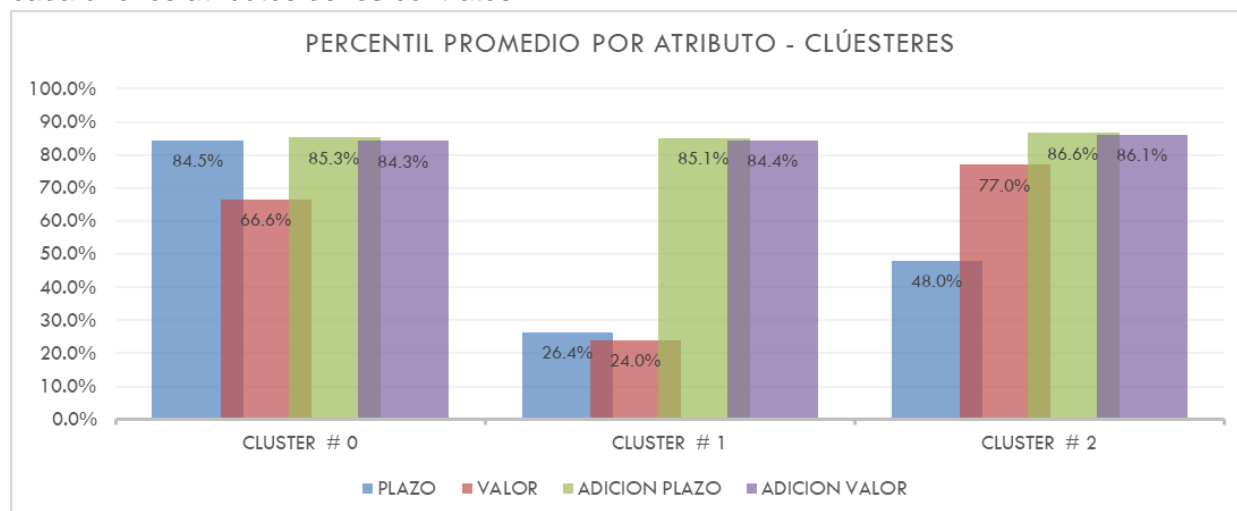
Riesgo promedio calculada por objeto del contrato



Tipo del contrato	Anormalidad promedio
Acuerdo marco	100.00%
Consultoria	100.00%
Interventoria	100.00%
Otro tipo de contrato	97.67%
Prestacion de servicios	51.51%

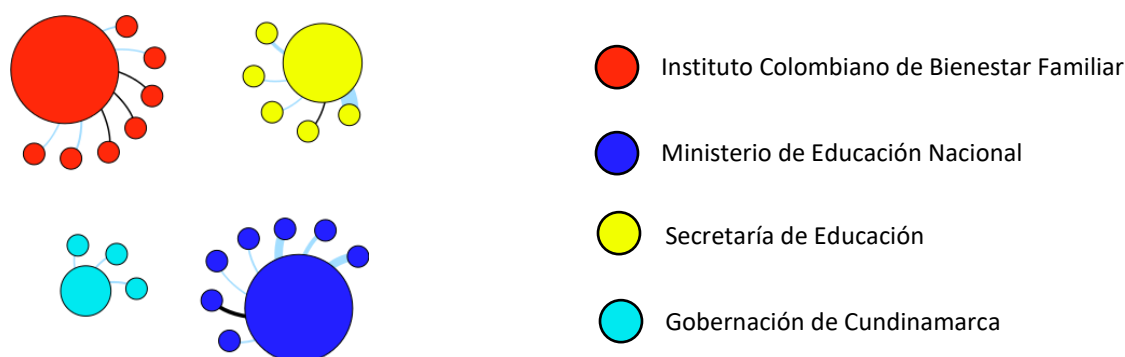
Anexo 5: Características de los clústeres

Características propias de cada clúster según el promedio del percentil acumulado de cada uno los atributos de los contratos



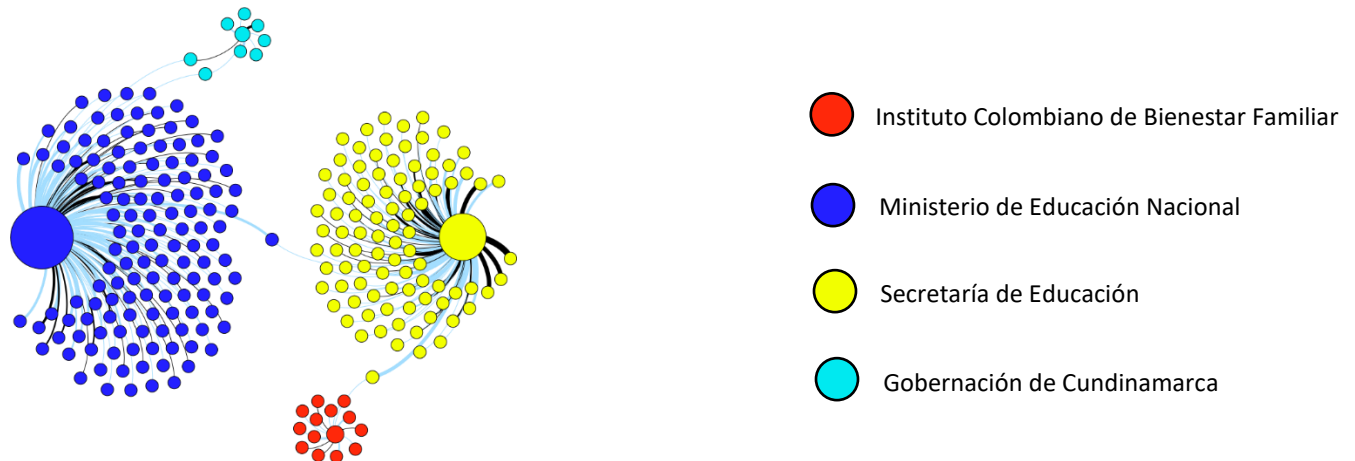
Anexo 6: Redes bipartitas

Red bipartita entre las entidades y los representantes de las entidades. El tamaño de los nodos representa el número de conexiones (contratos) del nodo frente al total de conexiones; el grosor de la conexión hace referencia al número de contratos en los cuales el representante actúa en nombre de la entidad; el color del nodo representa la entidad que está realizando los contratos; el color de la conexión muestra el riesgo promedio de los contratos realizados por la entidad con cada representante (entre más oscura la conexión, mayor el riesgo contractual promedio).

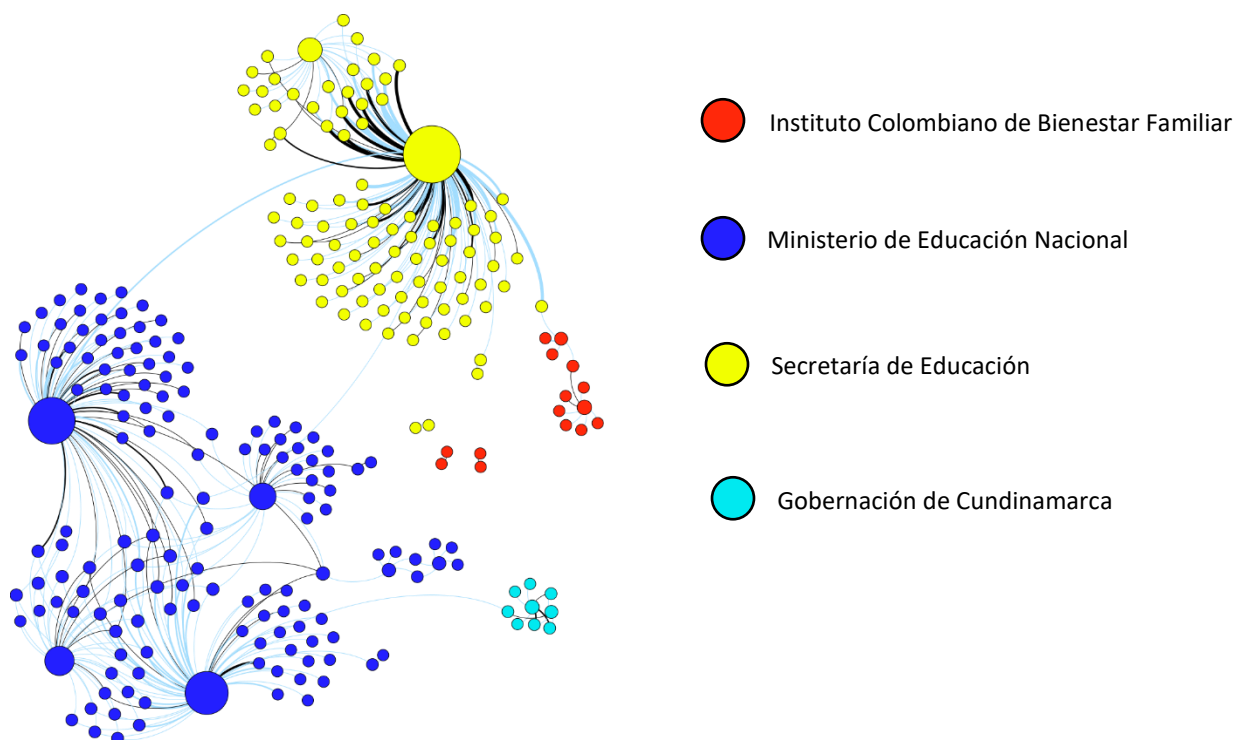


Red bipartita entre las entidades y los representantes de contratistas. El tamaño de los nodos representa el número de conexiones (contratos) del nodo frente al total de conexiones; el grosor de la conexión hace referencia al número de contratos en los cuales el representante realiza un acuerdo con la entidad; el color del nodo representa la entidad que está realizando los contratos; el color de la conexión muestra el riesgo promedio de los

contratos realizados por la entidad con cada representante (entre más oscura la conexión, mayor el riesgo contractual promedio). En este caso hay nodos que realizan contratos con más de una entidad.

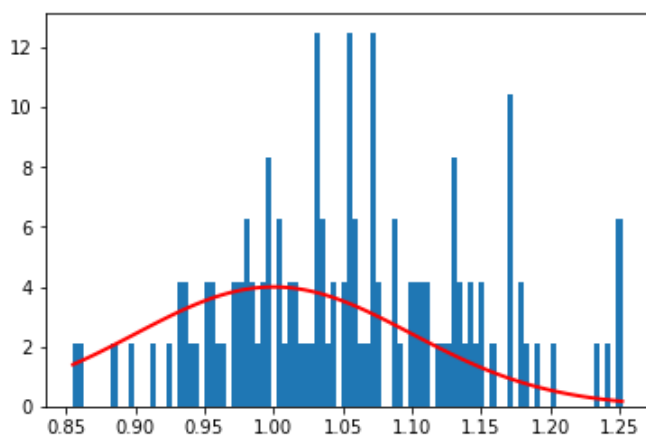


Red bipartita entre los representantes de las entidades y los representantes de los contratistas. El tamaño de los nodos representa el número de conexiones (contratos) del nodo frente al total de conexiones; el grosor de la conexión hace referencia al número de contratos en los cuales el representante realiza un acuerdo con el representante de las entidad; el color del nodo hace referencia al representante de la entidad que está realizando los contratos; el color de la conexión muestra el riesgo promedio de los contratos realizados por el representante de la entidad con cada representante de contratista (entre más oscura la conexión, mayor el riesgo contractual promedio).



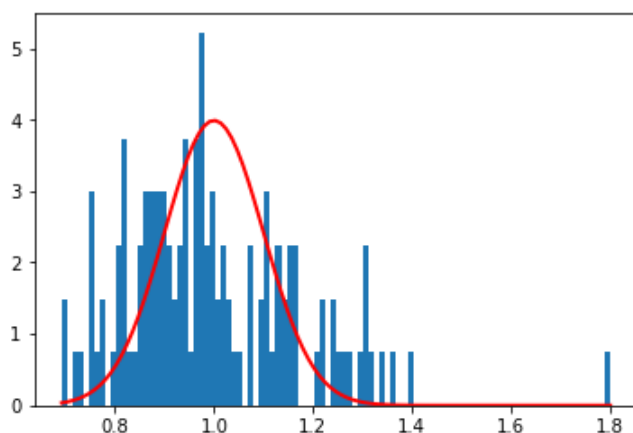
Anexo 7: Resultados para las pruebas de homofilia y heterofilia

Resultado para la prueba de homofilia sobre las conexiones entre los nodos de tipo A



Verificación de homofilia para el tipo de conexión $[A - A]$
 $pvalue = 0.6339$

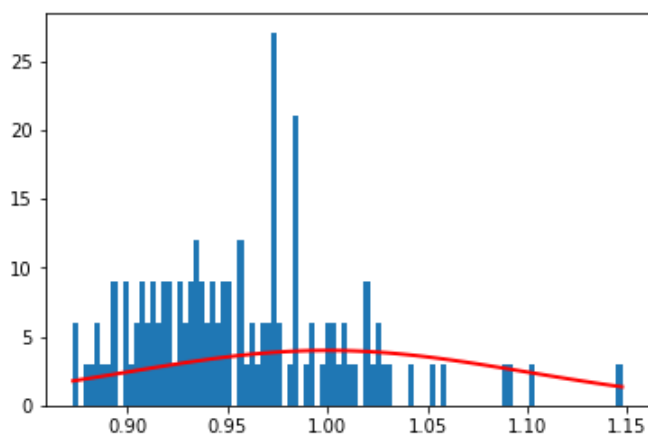
Resultado para la prueba de homofilia sobre las conexiones entre los nodos de tipo B



Verificación de homofilia para el tipo de conexión $[B - B]$

$pvalue = 0.0000$

Resultado para la prueba de heterofilia sobre las conexiones de los nodos de tipo B y A



Verificación de heterofilia para el tipo de conexión $[B - A]$

$pvalue = 0.0002$

[Anexo 8: Contratación histórica del SECOP \(2011 – 2016\)](#)

Tabla comparativa de la publicidad de la contratación en el SECOP (2011-2016)
(Colombia Compra Eficiente, 2017)

Total publicado	2011	2012	2013	2014	2015	2016
Valor (miles de millones de pesos)	\$33.093	\$50.880	\$73.507	\$94.534	\$121.255	\$83.747
Número de contratos	195.135	439.562	549.868	760.643	886.242	1.023.981
Nivel nacional	2011	2012	2013	2014	2015	2016
Valor (miles de millones de pesos)	\$16.607	\$30.764	\$40.334	\$24.028	\$69.266	\$42.343
Número de contratos	75.192	151.816	171.307	236.972	209.742	211.229
Nivel territorial	2011	2012	2013	2014	2015	2016
Valor (miles de millones de pesos)	\$16.485	\$20.115	\$33.173	\$67.709	\$51.989	\$41.404
Número de contratos	119.943	287.746	378.561	523.671	676.500	812.752

Fuente: Colombia Compra Eficiente con base en SECOP

Actualizado: 19 de enero de 2017