Delimitación de comunidades de marcas de productos en datos transaccionales para el negocio de venta al por menor utilizando Minería de Grafos

Trabajo final presentado por:
Oscar Marcel Brito Lillo

Trabajo de titulación para optar al título de:
Ingeniero Estadístico

Profesores guía:
Dr. Harvey Rosas

Valparaíso, Chile, 26 de Enero de 2018

Índice general

Resumen

En este trabajo, se buscó determinar distintas comunidades de marcas de productos sobre las bases de datos transaccionales de una cadena de supermercados y una de tiendas por departamentos, las cuales pertenecen al negocio de venta al por menor. Los datos representan la venta diaria de cada una de estos negocios en determinados periodos de tiempo.

Previo al análisis de la información, se realizó la administración de los datos con el fin de reorganizarlos de tal manera que esto permitiera encontrar los posibles conjuntos de elementos frecuentes. Luego, se calcularon los filtros utilizando el principio de los tres bordes más pesados para la construcción de los grafos y posterior detección de comunidades de marcas de productos utilizando el algoritmo de Newman y Girvan.

Palabras claves: datos transaccionales; conjunto de elementos frecuentes; principio de los tres bordes más pesados; grafos; comunidades; algoritmo de Newman y Girvan.

Capítulo 1

Introducción

En la actualidad, un tema relevante para muchas empresas de venta al por menor es poder manejar y transformar los grandes volúmenes de datos que contienen una gran cantidad información valiosa que pueda ser utilizada en diferentes áreas de negocio, tales como marketing, abastecimiento, planificación, entre otras. Estos pueden estar orientados a la información detallada del cliente o sobre las ventas diarias que se efectúan en el mercado.

Los datos transaccionales representan la información detallada de las ventas que se realizan diariamente y son una fuente inimaginable de información. Con estos datos se pueden desarrollar diferentes estudios que pueden estar enfocados en recopilar información de los clientes o del negocio. Algunos de los estudios más comunes son la visualización del comportamiento de los clientes mediante sus compras, la segmentación de clientes, la predicción de la venta de uno o más productos en un periodo de tiempo determinado, entre otros. Los datos transaccionales utilizados para este trabajo pertenecen a una cadena de supermercados y a una de tiendas por departamentos.

En el presente trabajo de titulación, se plantea que mediante el uso de grafos es posible visualizar datos transaccionales agrupados para determinar comunidades de marcas de productos en el negocio de venta al por menor con el fin de obtener información valiosa para el negocio.

El concepto de grafos para muchos es instalado por Leonardo Euler en el siglo XVIII en la ciudad de Königsberg situada junto al río Pregel en la Prusia Oriental.

En la actualidad Kaliningrado (Toranzos, 1976). Estos se definen como un conjunto de vértices o nodos que están conectados a través de bordes y estos permiten visualizar relaciones binarias. La construcción de los grafos se realizó utilizando la técnica de los conjuntos de elementos frecuentes, la cual está enfocada en determinar los pares, tríos o grupos numerosos de productos que son adquiridos de manera simultánea y en reiteradas ocasiones.

Las comunidades son particiones que se realizan en el interior de un grafo y se basan en el principio que debe existir una mayor cantidad de bordes en el interior de una comunidad que bordes entre comunidades (Fortunato, 2010). Existen distintos algoritmos que permiten realizar esta partición y en este trabajo se utilizó el algoritmo de Newman y Girvan (Newman y Girvan, 2004), el cual es un algoritmo basado principalmente en el cálculo de la intermediación de bordes.

El principal objetivo de este trabajo es encontrar comunidades de marcas de productos utilizando grafos sobre datos transaccionales, de modo que se obtenga información de utilidad para el negocio. A fin de cumplir el propósito establecido de este trabajo, se debe caracterizar detalladamente los datos transaccionales y la jerarquía de productos de venta al por menor, analizar las agrupaciones de productos de venta al por menor y su relación con la respectiva marca para la detección de conjuntos de elementos frecuentes, explorar sobre la teoría de grafos, visualización y construcción de estos y finalmente examinar algoritmos para la determinación de comunidades de productos.

En una primera instancia, los resultados obtenidos para la cadena de supermercados no fueron los esperados, puesto que en un caso específico no se hallaron las comunidades como era deseable, por lo tanto se optó por realizar un estudio complementario que permitiera determinar qué estaba sucediendo en los datos. Luego de esto se detectaron nuevamente las comunidades de marcas de productos, pero con una modificación en los datos, lo que tuvo buenos resultados.

Las programas con los cuales se contó para realizar este trabajo son el software estadístico RStudio y Spark versión 1.6, este una herramienta que permite usar diferentes lenguajes como Python, R, Java y Scala, siendo este último el lenguaje escogido para este trabajo.

Capítulo 2

Nociones preliminares

En este capítulo se exhiben algunas definiciones que se deben tener presentes para este trabajo de titulación. Se introducirán algunos temas de datos transaccionales, grafos y sus características, además de algoritmos de comunidades y otras medidas que fueron de gran utilidad en este trabajo.

2.1 Datos

Para cumplir los objetivos de este proyecto de titulación, se necesitó la información que contiene la venta diaria de las tiendas y la estructura de los diferentes niveles de los productos. Es por esto, que los datos transaccionales y la jerarquía de productos, tanto de supermercado como de tienda por departamentos, fueron definidos como los datos de entrada necesarios para este trabajo.

2.1.1 Datos transaccionales

Los datos transaccionales entregan la información detallada de la venta diaria, por ejemplo, las visitas de cada cliente o el monto de los productos entre otras cosas, estos son definidos como un grupo de productos (P) y transacciones (T). El primero se especifica como:

$$P = p_1, p_2, \ldots, p_n \quad , \tag{2.1}$$

donde cada p_i representa un artículo o producto. Por otro lado, las transacciones se pueden explicar como la venta de uno o más artículos que adquieren los clientes en una tienda determinada (Videla y Ríos, 2014).

Las variables de la principal fuente de datos en ambas tiendas en este caso es definida de la siguiente manera:

- ID_CLIENTE : Identificador de cliente.
- ID_LOCAL : Identificador de local.
- FECHA : Fecha en que se ejecutó la transacción.
- HORA : Hora en que se ejecutó la transacción.
- N_CAJA : Número de la caja donde el producto fue vendido el producto.
- ID_PRODUCTO : Identificador de producto.

Sin tomar en cuenta el identificador del producto y observando las otras cinco variables mencionadas anteriormente, se puede identificar la compra de un cliente en particular dentro de las transacciones, lo que desde ahora se nombrará como boleta, cuyas características principales son que existen en gran cantidad en el interior de las transacciones y pueden contener uno o más productos.

2.1.2 Jerarquía de productos

Es necesario llevar un orden de los productos en las tiendas de venta al por menor, una manera útil es clasificándolos según la función que cumpla, las características que posea o sus atributos, entre otros aspectos.

La jerarquía o maestro de productos nos proporciona la información de cada producto y puede variar dependiendo de la tienda de venta al por menor que se esté observando. Pérez y Pérez (2006) muestran una jerarquía de productos estándar que se compone por los siguientes siete niveles (ver Figura 2.1):

1. Familia de necesidades: da origen a la idea del producto.
2. Familia de producto: agrupa todos los productos que pueden satisfacer una necesidad fundamental.
3. Clase del producto: conjunto de productos que pertenecen a la misma familia, por que cumplen funciones muy similares.

4. Línea de producto: están dentro de una misma clase, se enfocan hacia los mismos clientes y fluctúan en los mismos rangos de precio.

5. Tipo del producto: son aquellos que dentro de una misma línea se encuentran en una o varias formas de producto.

6. Marca: es una característica asociada al producto para poder identificarlo.

7. Articulo: es el nivel más bajo de la jerarquía, siendo la unidad distinguible dentro de la marca o línea.

Figura 2.1: Jerarquía de productos

En el caso de las tiendas que se estudiaron en este trabajo de titulación, la jerarquía de productos consta de tan sólo cinco niveles. Estos son: la sección, la clase, la subclase, la marca y el producto o artículo. Bajo estas condiciones la estructura del maestro de productos utilizado es la siguiente:

- ID_PRODUCTO : Identificador de producto.

- ID_MARCA : Identificador de la marca.

- DESC_MARCA : Descripción de la marca.

- ID_SUBCLASE : Identificador de la subclase.

- DESC_SUBCLASE : Descripción de la subclase.

- ID_CLASE : Identificador de la clase.

- DESC_CLASE : Descripción de la clase.

- ID_SECCION : Identificador de la sección.

- DESC_SECCION : Descripción de la sección.

2.2 Teoría de Grafos

Es necesario destacar que cada una de las variables están relacionadas al producto. También, a modo de comparación, entre la jerarquía que se utilizó y la definida por Pérez y Pérez (2006), se puede decir que la sección sería el nivel de familia de productos, mientras que la subclase se equipara a la línea de productos.

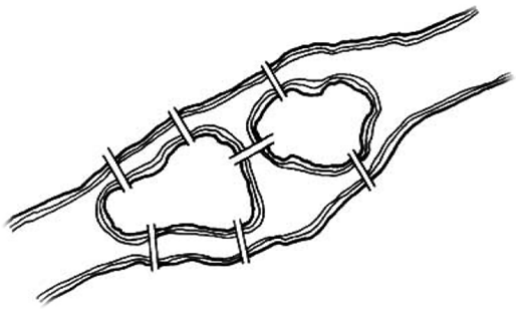

Figura 2.2: Los puentes de la ciudad de Königsberg (Toranzos, 1976).

Para muchos el concepto de grafos es instalado por Leonardo Euler en el siglo XVIII en la ciudad de Königsberg situada junto al río Pregel en la Prusia Oriental, en la actualidad Kaliningrado. Esta es famosa por sus puentes, siete en total, que unen el resto de la cuidad con sus dos islas como se puede apreciar en la Figura 2.2 (Toranzos, 1976).

En sus paseos dominicales por la ciudad este matemático transitaba frecuentemente por estos puentes, es ahí donde su mente lógica y ordenada se preguntó si era posible planear su próximo paseo de manera que saliendo de su casa cruce los siente puentes, una sola vez cada uno, antes de regresar a ella (Toranzos, 1976).

Es así como Euler expresó el problema de los puentes de Königsberg basado el mapa de la ciudad (Figura 2.2) a través de un multígrafo (Figura 2.3), donde cada sector terrestre de la ciudad lo denotó como un nodo o vértice y los puentes serían las conexiones llamadas bordes o aristas, de esta manera se planteó el problema que hace referencia a si es posible encontrar una ruta que recorra todas las aristas del grafo sólo una vez y que vuelva al punto de partida (Menéndez, 1998).

2.2.1 Definición

Los grafos son un conjunto de vértices que están conectados a través de bordes que permiten mostrar relaciones binarias. Se pueden considerar como diagramas o dibujos (representación diagramática). Por otra parte, se pueden ver geométricamente como puntos en el espacio, los cuales están unidos entre sí mediante líneas (Menéndez, 1998).

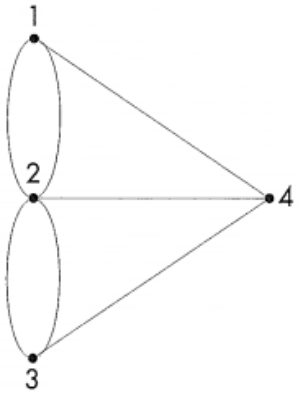

Figura 2.3: Representación geométrica del multígrafo realizado por Leonardo Euler, el cual representala disposición de los siete puentes de la ciudad de könisberg (Menéndez, 1998).

Formalmente de manera algebraica, un grafo G se define como un par ordenado $G = (V, E) = (V(G), E(G))$, donde V es un conjunto no vacío conocido como vértices o nodos y E es un conjunto de pares no ordenados de elementos distintos de V, denominados aristas o bordes (Menéndez, 1998).

Se dice que un grafo es dirigido u orientado, si sus aristas tienen una dirección definida, estas se pueden representar a través de un par ordenado (v_i, v_j) donde el primer elemento es el nodo de origen y el segundo el de destino, por lo tanto, se puede decir que la arista va desde v_i hasta v_j, a estos también se les conoce como digrafos. Si las aristas no indican dirección, el grafo es no dirigido y por lo tanto, cada par ordenado de (v_i, v_j) se puede representar sin importar el orden (Caicedo, Wagner y Méndez, 2010), es decir:

$$(v_i, v_j) = (v_j, v_i) \tag{2.2}$$

Otra característica de un grafo hace referencia a si sus aristas o bordes tiene algún peso o valor definido, si esto sucede se denota como grafo ponderado o etiquetado (Martínez, 2011).

Si se desea profundizar para tener un mayor conocimiento sobre los grafos, en general se recomienda estudiar la literatura propuesta por Menédez (1998) y Caicedo et al. (2010), además del trabajo realizado por Martínez (2011).

2.2.2 Representación matricial

Los grafos pueden ser representados de manera matricial y la matriz de adyacencia es una esas. Esto conlleva una gran ventaja puesto que, para las matrices ha sido desarrollada toda una teoría que permite la manipulación de estas para extraer cierta información característica de un grafo (Menéndez, 1998).

Sea $G = (V, E)$ un grafo de V vértices. La matriz de adyacencia, que será denotada por M, es una matriz de $M_{v \times v}$ valores, donde $M(i, j)$ toma el valor 1 si y sólo si existe una arista desde el nodo i al j (Caicedo, Wagner y Méndez, 2010).

$$M(i,j) = \begin{cases} 1, & \text{si existe el arco } (i,j) \\ 0, & \text{en caso contrario} \end{cases} \tag{2.3}$$

Algunas características de la matriz de adyacencia son:

- Las filas y las columnas representan los nodos del grafo.
- En un grafo no dirigido, la matriz de adyacencia siempre es simétrica. En el caso contrario no necesariamente lo será.
- Los valores de los elementos $M(i, j)$ corresponderán a los pesos de las aristas del grafo, si es que este los posee.
- La diagonal principal de esta matriz serán sólo ceros si el grafo no tiene bucles (aristas que apunten a un mismo nodo) en sus nodos.

Para más información ver Menéndez (1998) y Caicedo et al. (2010).

$$M(i,j) = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

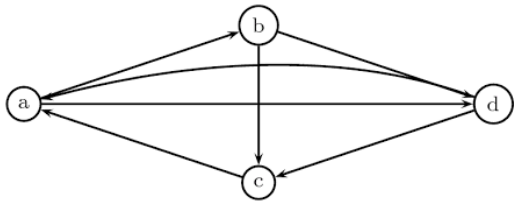

Figura 2.4: Matriz de adyacencia de un grafo dirigido sin pesos en sus aristas (Caicedo, Wagner y Méndez, 2010).

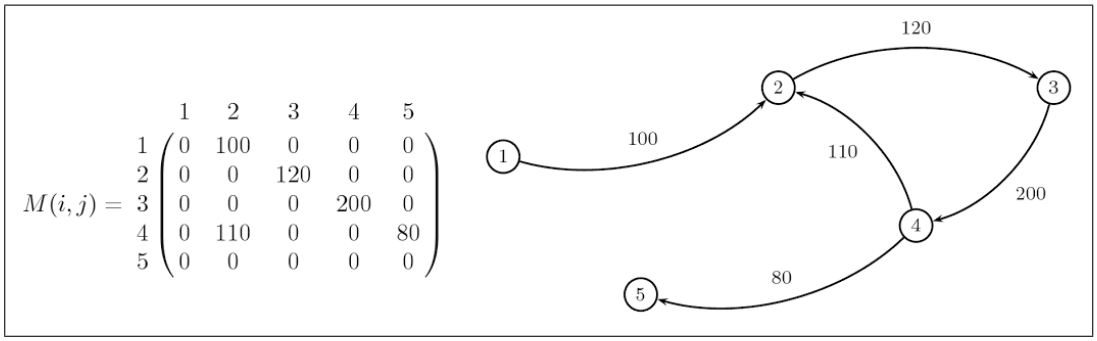

Figura 2.5: Matriz de adyacencia de un grafo dirigido con pesos en sus aristas (Caicedo, Wagner y Méndez, 2010).

2.3 Algoritmo de detección de comunidades

Las comunidades son particiones que se realizan dentro de un grafo. Fortunato (2010)expresa la idea de que debe haber más bordes dentro de una comunidad que bordes que vinculen a esta con el resto del grafo. En cambio, Martínez (2011) entrega una definición más formal en donde precisa que las comunidades dentro un grafo son la división de nodos en subgrupos dentro de los cuales las conexiones entre nodos son muy numerosas, pero no así las conexiones externas que son bastante escasas. A continuación, se presentan dos algoritmos que permiten detectar comunidades en grafos o redes, estos serán de gran ayuda para entender lo realizado durante este trabajo. También, se muestran algunas medidas que utilizan estos algoritmos.

2.3.1 Algoritmo de Newman y Girvan

Newman y Girvan (2004), presentan un algoritmo basado principalmente en el cálculo de la intermediación de bordes para la detección de comunidades dentro de un grafo, este debe tener como características principales ser no dirigido y no tener peso en sus bordes. Aseguran que existen dos características centrales que distinguen a su algoritmo de aquellos que los han precedido. Primero, indican que su algoritmo es de separación y no de aglomeración, centrándose en la búsqueda de bordes con mayor nivel en la medida intermediación, donde ésta favorece a los bordes entre comunidades y no aquellos que se encuentran dentro de ellas. La segunda característica, es la inclusión de un nuevo paso en el cálculo del algoritmo.

Calcular la medida de intermediación para cada borde del grafo y retirar aquel con mayor puntaje, son los primeros pasos que se deben seguir del algoritmo, luego

se integra una nueva etapa que consiste en volver a calcular la intermediación ya sin el borde retirado. Por lo tanto, el algoritmo para la detección de comunidades se compone de 4 pasos y se define de la siguiente forma:

Paso 1: Calcular las puntuaciones de intermediación para todos los bordes de la red.

Paso 2: Encontrar el borde con la puntuación más alta y quitarlo de la red.

Paso 3: Volver a calcular intermediación para todos los bordes restantes.

Paso 4: Repita desde el paso dos.

Este algoritmo finaliza una vez que se eliminan la totalidad de los bordes, es decir, en la última iteración quedarán tantas comunidades como nodos existan en el grafo. Por otro lado, en cada iteración se debe calcular la medida de modularidad sobre las comunidades, con el fin de lograr determinar cuál es la mejor partición en el interior del grafo. Esta será aquella que tenga la máxima modularidad entre todas las iteraciones.

La etapa de recalcular la medida de intermediación es trascendental para el funcionamiento de este método y para obtener resultados satisfactorios, según indican Newman y Girvan (2004).

2.3.2 Maximización de la modularidad

Blondel et al. (2008) proponen un método para la detección de comunidades en grandes redes basado en la optimización de la medida de modularidad, la cual usualmente es utilizada para medir la calidad de las particiones de un grafo, en este caso el algoritmo la emplea como parte fundamental para encontrar las posibles comunidades que puedan existir dentro de una red. Este algoritmo consta de dos fases que se repiten de forma iterativa, si se tiene un grafo con N cantidad de nodos, entonces el algoritmo se define de la siguiente manera:

Fase 1: La primera fase consiste en asignar cada nodo de la red como una comunidad diferente, es decir, el número de comunidades será igual a la cantidad de nodos que contenga la red, una vez listo esto se calcula la modularidad, aunque es esperable que no sea muy buena medida por ahora. Luego el nodo i es situado con el vecino j más cercano, se calcula y evalúa la modularidad que se lograría al posicionar la comunidad i en j, si este cambio hace que la medida de modularidad aumente, la posible unión se mantiene. Este procedimiento se aplica repetidamente en todos los nodos hasta que la modularidad llegue a su máximo.

Fase 2: La segunda fase consiste en la construcción de un nuevo grafo, donde todas las posibles uniones serán definidas como los nuevos nodos y siguiendo la lógica de la primera fase como las nuevas comunidades. Si el grafo es ponderado (con peso en sus bordes), se sumaran los pesos de los bordes que unen los a nodos correspondientes a las nuevas comunidades. Luego es posible aplicar nuevamente la primera fase.

Una vez terminada la segunda fase se debe volver a realizar el primer paso, cuando la modularidad de la siguiente iteración no es mejor que la anterior se da por terminado el proceso.

2.3.3 Medida de intermediación

Brandes (2008), se refiere a la intermediación como una de las medidas más prominentes de la centralidad, a su vez este es un concepto central para el análisis de redes o grafos. Newman y Girvan (2004), nombran distintas medidas de intermediación, aunque para efectos de este trabajo solo se profundizará sobre la medida de intermediación de bordes, la cual se basa en el cálculo de las distancias geodésicas o caminos más cortos.

Si se define un grafo $G = (V, E)$ donde V es un conjunto de nodos y E los bordes que unen a estos, además se tiene que s y t son dos elementos que pertenecen a V, en donde s es el nodo fuente y t el objetivo. Se puede definir la intermediación de bordes o "edges betweenness" como la suma de la división entre el número de caminos mas cortos que pasan por algún vértice v distinto de s y t y el número de caminos mas cortos entre s y t, expresado con la siguiente fórmula (Brandes, 2008):

$$C_B(\nu) = \sum_{s,t \in \nu} \frac{\sigma(s, t|v)}{\sigma(s, t)} \quad , \tag{2.4}$$

donde $\sigma(s, t)$ es el número de caminos más cortos entre s y t y $\sigma(s, t|v)$ el número de caminos más cortos que pasan por algún vértice v distinto de s y t.

Para conocer en mayor detalle y ampliar los conocimientos sobre esta u otras medidas de intermediación se recomienda revisar los artículos propuestos por Newman y Girvan (2004), Brandes (2008) o Freeman (1977).

2.3.4 Medida de modularidad

Cada vez que se obtienen comunidades dentro de un grafo, queda la incertidumbre de saber qué tan bien fue particionada la red. Intuitivamente se podría decir que, si un grafo con varias comunidades y cada una de estas tiene una gran cantidad de bordes o enlaces internos, entonces probablemente representen una buena partición. Pero para responder de mejor manera a esta duda, existe una medida llamada modularidad que mide la calidad de las comunidades que se encuentran en un grafo. Raeder y Chawla (2009), muestran que la modularidad Q de un conjunto de comunidades se define como:

$$Q = \sum_i (e_{ii} - a_i^2) \quad , \tag{2.5}$$

donde e_{ii} es la fracción de los bordes que unen vértices en la comunidad i a otros vértices dentro de i, por otro lado a_i es la fracción de bordes que son puntos finales en la comunidad i.

El máximo valor de Q que se puede obtener es 1, esto indica una muy buena estructura de comunidades, mientras que el valor mínimo es 0 y apunta a que no se realizó ningún tipo de partición a la red. En la práctica un valor aceptable de modularidad varía aproximadamente en el intervalo de 0.3 a 0.7, los valores más altos son muy poco comunes (Newman y Girvan, 2004).

Capítulo 3

Metodología

La metodología utilizada para este trabajo de titulación, permitió determinar las diferentes comunidades de marcas de productos que existen en un conjunto de datos transaccionales en un periodo de tiempo definido y consta de diferentes etapas; en la primera se realizó un trabajo sobre los datos con el fin de prepararlos para las siguientes etapas. Una vez realizado lo mencionado anteriormente, se determinaron los conjuntos de elementos frecuentes; en la tercera etapa se buscó generar diferentes filtros para eliminar los bordes no esenciales entre nodos, posteriormente en la cuarta fase, se construyeron los grafos y por último se detectaron las comunidades de marcas productos.

Las tres primeras etapas se realizaron con la herramienta de Spark 1.6 utilizando el lenguaje de programación llamado Scala, mientras que para la construcción de grafos y detección de comunidades se trabajó en el software estadístico RStudio utilizando el paquete de igraph.

3.1 Administración de datos

Dentro de supermercados y tiendas por departamentos existen alrededor de 50 y 55 secciones diferentes, entre las cuales no se tomaron en cuenta algunas que están asociadas a procesos internos, y otras en las cuales no se realizaron ventas en el periodo de tiempo definido para los experimentos.

Por motivos de negocio, no siempre tiene sentido asociar algunos productos, como

por ejemplo, los de la sección de carnicería con la de autos, bajo esta condición y con la idea de optimizar la detección de comunidades se determinó dividir en dos grupos las diferentes secciones de supermercados (Tablas 3.1 y 3.2). Paralelamente, en tiendas por departamentos no fue necesario realizar una división de las secciones, puesto que no existen estos problemas de manera tan evidente.

Posteriormente, se definieron las secciones de los grupos A y B. En el primero se encuentran secciones similares en cuanto a compras o características de consumo, por el ejemplo, abarrotes, licores, bebidas, entre otras. En el segundo, aparecen secciones que son poco habituales o de menor consumo en términos de venta como por ejemplo jardín o tecnología.

Grupo A				
carnicería	vino	galletas/golosinas	pastelería	cocktail
quesería	pastas	cerdo/cordero	cafetería	licores
congelados	almacén	botillería/gaseosas	pescadería	pollos

Tabla 3.1: Secciones definidas para el grupo A

Grupo B				
automotor	mascotas	telefónica	accesorios de jardín	promociones
electrohogar	jardín	piletas/parrillas	tecnología	electrónica
ferretería	muebles	mesa/terraza	electrodomésticos	decoración

Tabla 3.2: Secciones definidas para el grupo B

Las transacciones de supermercados y tiendas por departamentos contienen los registros de compra de cada cliente, cada registro hace referencia a un producto. El objetivo principal de la administración de datos es identificar las distintas boletas que existen dentro los datos transaccionales y a su vez, las diferentes marcas entre los productos que se encuentren dentro de estas.

Con el objetivo de hallar las distintas marcas dentro de las boletas que se encuentran en las transacciones, se debió agregar la información proveniente de la tabla de jerarquía de productos, en donde se añadieron las variables de ID_MARCA e ID_SECCION a los datos transaccionales, la primera variable se utilizó para determinar cuál es la marca asociada al producto, mientras que la segunda solamente ayudó para dividir las transacciones en los grupos A y B antes mencionados, para el caso de supermercados. Una vez que se tienen las diferentes marcas de los productos y los grupos ya construidos, se deben identificar las distintas boletas que existen en los datos transaccionales utilizando la llave de la boleta definida anteriormente. Efectuada esta acción, se debe realizar una agrupación a los datos

en donde se dejan de lado los productos manteniendo solo las marcas de manera única, ya que dentro de una misma boleta pueden haber productos que tengan la misma marca. Por último, se eliminan todas aquellas boletas que contengan solamente una marca asociada a los productos, puesto que estas no son de utilidad para los procedimientos posteriores.

Con el fin de ejemplificar de mejor manera lo definido anteriormente, se simularon algunas transacciones que se muestran en la siguiente tabla:

ID_PRODUCTO	ID_CLIENTE	ID_LOCAL	FECHA	HORA	N_CAJA	ID_MARCA
Producto_1	Cliente_1	Local_1	21-07-2017	16:22	Caja_1	Marca_1
Producto_2	Cliente_1	Local_1	21-07-2017	16:22	Caja_1	Marca_1
Producto_3	Cliente_1	Local_1	21-07-2017	16:22	Caja_1	Marca_1
Producto_4	Cliente_2	Local_2	21-05-2017	18:33	Caja_2	Marca_2
Producto_5	Cliente_2	Local_2	21-05-2017	18:33	Caja_2	Marca_2
Producto_6	Cliente_2	Local_2	21-05-2017	18:33	Caja_2	Marca_3
Producto_7	Cliente_3	Local_3	21-04-2017	21:22	Caja_1	Marca_1
Producto_8	Cliente_3	Local_3	21-04-2017	21:22	Caja_1	Marca_1
Producto_9	Cliente_3	Local_3	22-04-2017	21:22	Caja_1	Marca_3
Producto_10	Cliente_3	Local_3	23-04-2017	21:22	Caja_1	Marca_3

Tabla 3.3: Ejemplo de transacciones.

Se puede ver que en la Tabla 3.3, las transacciones corresponden a diez productos distintos de tres clientes diferentes para dos locales, según las variables que identifican una compra hay tres boletas distintas dentro de los datos, ya que existen tres combinaciones únicas de estas cinco variables, como se puede apreciar en la Tabla 3.4.

ID	ID_PRODUCTO	ID_CLIENTE	ID_LOCAL	FECHA	HORA	N_CAJA	ID_MARCA
1	Producto_1	Cliente_1	Local_1	21-07-2017	16:22	Caja_1	Marca_1
1	Producto_2	Cliente_1	Local_1	21-07-2017	16:22	Caja_1	Marca_1
1	Producto_3	Cliente_1	Local_1	21-07-2017	16:22	Caja_1	Marca_1
2	Producto_4	Cliente_2	Local_2	21-05-2017	18:33	Caja_2	Marca_2
2	Producto_5	Cliente_2	Local_2	21-05-2017	18:33	Caja_2	Marca_2
2	Producto_6	Cliente_2	Local_2	21-05-2017	18:33	Caja_2	Marca_3
3	Producto_7	Cliente_3	Local_3	21-04-2017	21:22	Caja_1	Marca_1
3	Producto_8	Cliente_3	Local_3	21-04-2017	21:22	Caja_1	Marca_1
3	Producto_9	Cliente_3	Local_3	22-04-2017	21:22	Caja_1	Marca_3
3	Producto_10	Cliente_3	Local_3	23-04-2017	21:22	Caja_1	Marca_3

Tabla 3.4: Ejemplo de transacciones con boletas identificadas.

En la Tabla 3.4 se observa que hay productos dentro una misma boleta (columna número uno) que son de la misma marca, al realizar la agrupación dejando de

lado el identificador del producto, se logra ver las marcas que contienen cada boleta, como se muestra en la Tabla 3.5:

ID	ID_CLIENTE	ID_LOCAL	FECHA	HORA	N_CAJA	ID_MARCA
1	Cliente_1	Local_1	21-07-2017	16:22	Caja_1	Marca_1
2	Cliente_2	Local_2	21-05-2017	18:33	Caja_2	Marca_2
2	Cliente_2	Local_2	21-05-2017	18:33	Caja_2	Marca_3
3	Cliente_3	Local_3	21-04-2017	21:22	Caja_1	Marca_1
3	Cliente_3	Local_3	21-04-2017	21:22	Caja_1	Marca_3

Tabla 3.5: Ejemplo en donde se identifican las marcas dentro de una boleta.

3.2 Conjunto de elementos frecuentes

El descubrimiento de elementos frecuentes es una de las técnicas de minería de datos, la cual se enfoca en determinar los pares, tríos o grupos numerosos de productos que son adquiridos en el tiempo de manera simultánea y en reiteradas ocasiones. Los conjuntos de elementos más frecuentes o FI (por sus siglas en Inglés: Frecuent Itemsets) son vistos regularmente en el uso de las reglas de asociación, esta última utiliza como parte principal los conjuntos de elementos frecuente en la obtención de sus reglas, además de las medidas de soporte y confiabilidad.

Rajaraman y Ullman (2012), plantean que los conjuntos de elementos que aparecen en una cantidad determinada de boletas se definen frecuentes. De manera formal, se asume que hay un número s denotado como el umbral de la regla. Si I es un conjunto de elementos, se dirá frecuente si en la cantidad de boletas que aparece es mayor o igual al umbral s definido anteriormente.

Para efectos de este trabajo en el cálculo de los conjuntos de elementos frecuentes, tanto para supermercados como para tiendas por departamentos, se tomó la tabla resultante de la administración de datos (Tabla 3.5) en la cual se tienen identificadas las distintas boletas dentro de los datos transaccionales y las marcas que contienen estas. Además, se definió el umbral mínimo s que debe cumplir cada regla, posteriormente se calculan todas las combinaciones de pares de productos dentro de las distintas boletas que cumplan con umbral s definido anteriormente.

Con el propósito de entender de mejor manera el trabajo que se realizó, más adelante se presenta el siguiente ejemplo, en donde se puede apreciar cómo se produce la obtención de los conjuntos de elementos frecuentes. En un local de ventas al por menor, se tienen los siguientes datos transaccionales (Tabla 3.6)

en los cuales existen cuatro boletas distintas con diferentes productos, como se puede ver en la siguiente tabla:

ID_BOLETA	**PRODUCTOS**
1	Carne - Bebida - Papas fritas - Maní
2	Pan - Carne - Bebida - Cerveza
3	Tallarines - Salsa de tomate - Carne - Pan
4	Salsa de tomate - Tallarines - Papas fritas - Bebida - Carne

Tabla 3.6: Transacciones que contienen boletas y productos distintos.

Desde las transacciones de la Tabla 3.6 se pueden apreciar las siguientes reglas:

1. Carne - Bebida
2. Tallarines - Salsa de tomate
3. Pan - Cerveza
4. Papas fritas - Carne
5. Salsa de tomate - Papas fritas

La frecuencia con que se repiten cada una de estas reglas en las ventas es de tres, dos, dos, uno y uno respectivamente. Para que estos conjuntos de elementos sean frecuentes tienen que cumplir con la condición de que la frecuencia de compra sea mayor o igual al umbral de compra establecido. Si se define un umbral s igual a dos, se puede observar que *Pan - Cerveza* y *Salsa de tomate - Papas fritas* no cumplen con la condición, por lo tanto, no son frecuentes. Mientras que *Carne - Bebida*, *Fideos - Salsa de tomate* y *Papas fritas - Carne* si lo son, ya que cumple con el umbral de compras como se puede apreciar en la Tabla 3.7.

Conjunto de artículos	**Frecuencia**	**Umbral(s)**	**Resultado**
Carne - Bebida	3	2	Frecuente
Fideos - Salsa de tomate	2	2	Frecuente
Papas Fritas - Carne	2	2	Frecuente
Pan - Cerveza	1	2	No frecuente
Salsa de Tomate - Papas Fritas	1	2	No frecuente

Tabla 3.7: Conjunto de elementos con un umbral definido s=2, donde se muestran si son frecuentes o no.

Es necesario recalcar que las reglas definidas anteriormente son sólo algunas escogidas para caracterizar el ejemplo, de ningún modo son todas las reglas existentes.

3.3 Filtro para la construcción de los grafos

Para poder visualizar de mejor manera las posibles comunidades existentes dentro de los grafos, se hace necesario eliminar bordes poco influyentes o no esenciales que existen en estas redes. Videla y Ríos (2014), presentan un método que es realizable en dos pasos. El primero es calcular el límite superior de tres bordes más pesados o *tthet* (por sus siglas en Inglés: top three heavy edges thresholds), para ello se debe escoger los tres bordes con mayor peso dentro del grafo y calcular el promedio, entonces el *tthet* es igual a:

$$tthet = \frac{E_1max + E_2max + E_3max}{3} \quad , \tag{3.1}$$

donde E_1max, E_2max, E_3max son los bordes con mayor peso respectivamente.

Si se filtra directamente por el peso obtenido en el cálculo de *tthet* dentro del grafo correspondiente, solo uno o dos bordes cumplirían esta condición. El segundo paso consiste en generar filtros que eliminen bordes y nodos de manera gradual, para obtener una mejor visualización de los grafos y una buena calidad de comunidades. La idea es ir calculando una proporción del valor de *tthet*, la cual comienza desde el 5 %, 10 % y así sucesivamente hasta llegar al 100 % de este valor, por ende se obtendrá un total de veinte filtros. Esto se puede apreciar en la Fórmula 4.2.

$$filtros = \{\, 0,05 \times tthet,\, 0,10 \times tthet,\, \ldots\, ,\, 0,95 \times tthet,\, 1 \times tthet \,\} \tag{3.2}$$

Videla y Ríos (2014), presentan la idea de filtrar los grafos una vez que están construidos. Bajo la metodología establecida para este trabajo, la aplicación de los filtros es un poco distinta, puesto que estos son aplicados antes de la construcción de los grafos.

Tanto para supermercados y tiendas por departamentos el procedimiento de calcular la métrica de *tthet* fue realizada sobre la tabla en la cual se tienen los conjuntos de elementos frecuentes, esta fue ordenada de mayor a menor para encontrar las tres reglas con mayor frecuencia. Luego, se calculó la media de los tres bordes más pesados y de esta, se fueron calculando las distintas proporciones, obteniendo un total de veinte filtros distintos, como se especificó anteriormente. Una vez calculados, se fue filtrando la tabla de los conjuntos de elementos frecuentes por cada uno de los filtros, sin embargo, esto no quiere decir que se obtendrían veinte tablas, ya que estos se aplicaron de forma gradual hasta que quedaba una cierta cantidad de registros suficientes para la creación de los grafos y la posterior detección de comunidades.

3.4 Construcción de los grafos

La construcción de los grafos, tanto para supermercados como para tiendas por departamentos, se realizó bajo el mismo procedimiento. Luego de aplicar los distintos filtros a las reglas obtenidas en los conjuntos de elementos frecuentes, se utilizaron las tablas resultantes de este proceso, con el fin de construir los grafos correspondientes.

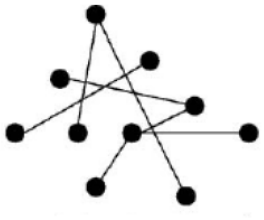

Figura 3.1: Representación de un grafo (Martínez, 2011).

Los grafos construidos son no dirigidos y no ponderados, es decir, las características de estos, para ambas tiendas, son que sus bordes no tienen dirección definida ni tampoco poseen peso alguno.

3.5 Detección de comunidades

Dentro de un grafo, la detección de comunidades es un problema que se fue generando en el tiempo bajo la necesidad de poder extraer la valiosa información que contienen estas redes que pueden ser aplicadas en diferentes áreas de negocio.

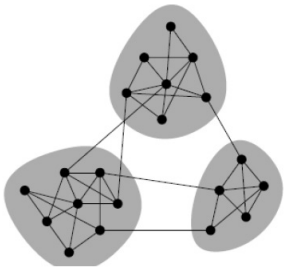

Figura 3.2: Representación de comunidades en un grafo (Martínez, 2011).

La detección de comunidades, ya sea supermercados o tiendas por departamentos se realizó bajo el mis procedimiento y es la última etapa en la cual se consolidan todos los pasos realizados anteriormente. A los grafos que fueron construidos con los distintos filtros en el paso anterior, se les aplicó el algoritmo de Newman y Girvan definido en la Sección 2.3 del Capítulo 2. Posteriormente, a las comunidades encontradas se les calculó la medida de modularidad, la cual ayudará a determinar qué tan buenas son las particiones efectuadas por el algoritmo.

Cabe destacar que la cantidad de comunidades que se hallen dentro de un grafo no se puede determinar de antemano, este valor solo se conocerá una vez aplicado el algoritmo.

Capítulo 4

Resultados

En este capítulo se mostrarán los resultados obtenidos en las cadenas de supermercados y tiendas por departamentos al aplicar la metodología señalada en el Capítulo 3. Se detallarán la cantidad de boletas que se encontraron de la administración de datos, las reglas obtenidas y el umbral mínimo s utilizado. Se podrán visualizar los distintos grafos construidos y las comunidades encontradas con los distintos filtros calculados.

4.1 Datos utilizados

Como se mencionó en el Capítulo 2, los datos utilizados para la detección de comunidades de marcas de productos corresponden a una cadena de supermercados y tiendas por departamentos. En el caso de supermercados se seleccionaron datos correspondientes a las ventas de la semana número diecinueve, cuyas fechas van entre el 9 y 15 de mayo de 2016. En esta semana se cuenta con 13.841.928 transacciones de diferentes clientes, donde se pueden encontrar 53.088 productos distintos de 3.914 marcas (Tabla 4.1).

Cantidad de transacciones	**Fecha inicial**	**Fecha final**	**Productos distintos**	**Marca distintas**	**Secciones distintas**
13.841.928	09-05-2016	15-05-2016	53.088	3.914	52

Tabla 4.1: Resumen del conjunto de datos para la cadena de supermercados.

Para tiendas por departamentos los datos utilizados corresponden a las ventas entre el 6 y 19 de Julio, correspondientes a las semanas número veintiocho y veintinueve del 2015. Estas contienen 4.709.194 transacciones, en donde se pueden apreciar 134.860 productos distintos de 1.093 marcas diferentes como se refleja en la Tabla 4.2.

Cantidad de transacciones	**Fecha inicial**	**Fecha final**	**Productos distintos**	**Marca distintas**	**Secciones distintas**
4.709.194	06-07-2015	19-07-2015	134.860	1.093	63

Tabla 4.2: Resumen del conjunto de datos para la cadena de tiendas por departamentos.

4.2 Resultados de supermercados

Al aplicar los distintos métodos y algoritmos descritos en el Capítulo 3 sobre los conjuntos de datos de los grupos A y B pertenecientes a supermercados se encontraron dos escenarios diametralmente opuestos en cuanto a resultados. Con respecto a las transacciones del grupo A de supermercados se hallaron un total de 659.703 boletas distintas con diversas marcas, sobre estas se calcularon los conjuntos de elementos frecuentes en los cuales se hallaron 5.062 reglas de pares de marcas, cuyo umbral mínimo s utilizado fue igual a 659, este corresponde al 0,1 % de la cantidad total de boletas que fueron encontradas.

Posteriormente, al calcular el límite superior de los tres bordes más pesados, que en este caso son las cantidades de las tres reglas con mayor frecuencia dentro de las boletas, se obtuvo lo siguiente:

$$tthet = \frac{39.900 + 38.390 + 33.469}{3} = 37.253 \quad , \tag{4.1}$$

este valor es utilizado para calcular los diferentes filtros, que serán una proporción de esta constante al ir multiplicando el $tthet$ por valores entre 0,05 y 1, que representan desde el 5 % al 100 %, como se puede ver en la siguiente tabla:

Proporción	tthet	Filtros
0,05	37.253	1.862,7
0,10	37.253	3.725,3
0,15	37.253	5.588,0
0,20	37.253	7.450,6
0,25	37.253	9.313,3
0,30	37.253	11.175,9
0,35	37.253	13.038,6
0,40	37.253	14.901,2
0,45	37.253	16.763,9
0,50	37.253	18.626,5
0,55	37.253	20.489,2
0,60	37.253	22.351,8
0,65	37.253	24.214,5
0,70	37.253	26.077,1
0,75	37.253	27.939,8
0,80	37.253	29.802,4
0,85	37.253	31.665,1
0,90	37.253	33.527,7
0,95	37.253	35.390,4
1,00	37.253	37.253,0

Tabla 4.3: Filtros calculados para el grupo A de supermercados.

Los filtros calculados anteriormente son aplicados sobre las reglas para dejar fuera las relaciones con menor peso. Desde ahora en adelante, los filtros serán denominados por el porcentaje del *tthet* que lo representa y no por el valor del resultado expuesto en la Tabla 4.3. Este mismo criterio será utilizado en los casos de grupo B de supermercados y tiendas por departamentos.

En la Tabla 4.4, se puede apreciar cómo a medida que aumenta el porcentaje de los filtros disminuye la cantidad de reglas o bordes no esenciales.

Filtros	**Cantidad de Reglas**	**Cantidad de Reglas Filtradas**
Sin Filtro	5.062	0
5 %	1.687	3.375
10 %	649	4.413
15 %	340	4.722
20 %	196	4.866
25 %	127	4.935
30 %	80	4.982
35 %	51	5.011
40 %	40	5.022

Tabla 4.4: Cantidad de reglas por cada filtro en el grupo A de supermercados.

Cabe destacar que los filtros fueron aplicados hasta que se mantuvo una cierta cantidad de reglas suficientes que permitiera generar los grafos pertinentes y, la posterior detección de comunidades de marcas de productos.

El primer grafo construido es con la totalidad de las reglas, es decir, sin aplicar ningún filtro. Posteriormente se crearon los siguientes grafos utilizando los diferentes filtros aplicados sobre las reglas, por lo tanto, se crearon nueve grafos para el grupo A de supermercados.

En las Figuras 4.1, 4.2 y 4.3 se pueden ver algunos grafos con distintos filtros, además se percibe cómo van desapareciendo los bordes y nodos cuando los valores de los filtros van aumentando.

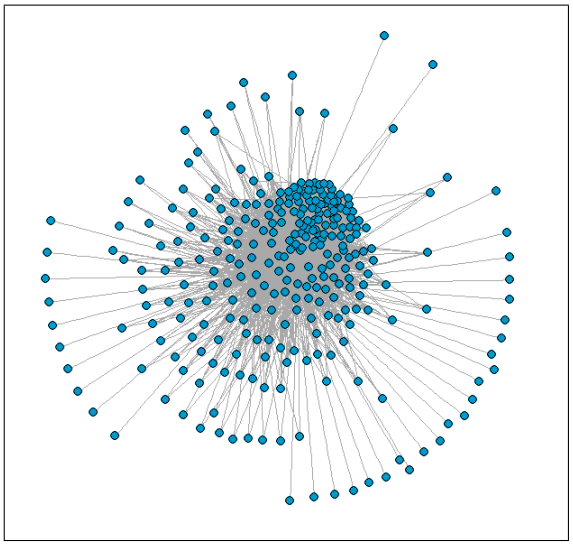

Figura 4.1: Grafo sin filtro para el grupo A de supermercados.

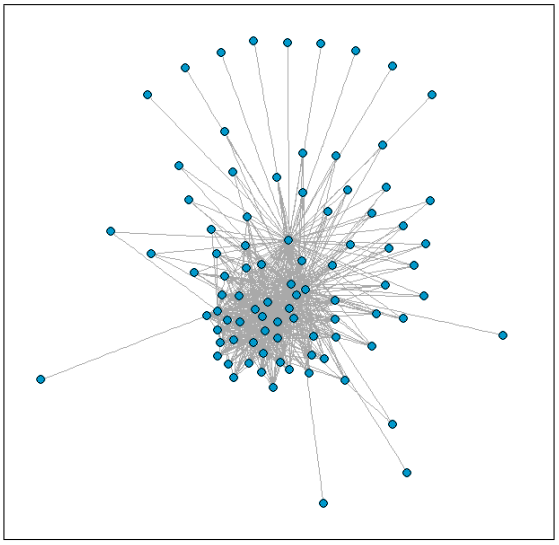

Figura 4.2: Grafo con filtro de 10 % para el grupo A de supermercados.

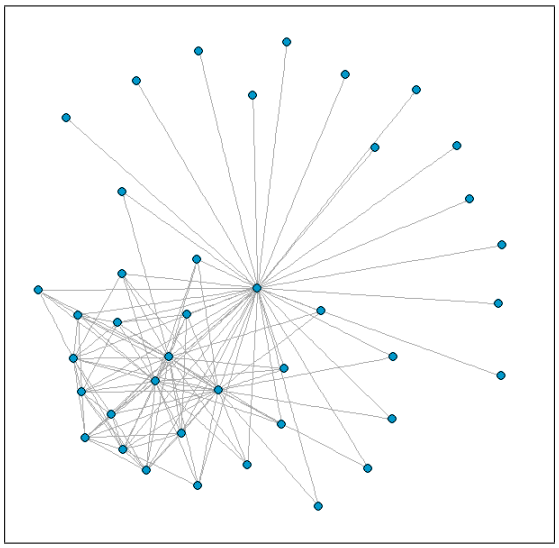

Figura 4.3: Grafo con filtro de 25 % para el grupo A de supermercados.

Al aplicar el algoritmo de Newman y Girvan para la detección de comunidades no se obtuvieron los resultados esperados en los diferentes grafos construidos, puesto que no se lograron determinar comunidades de productos, tal como se puede apreciar las Figuras 4.4 y 4.5.

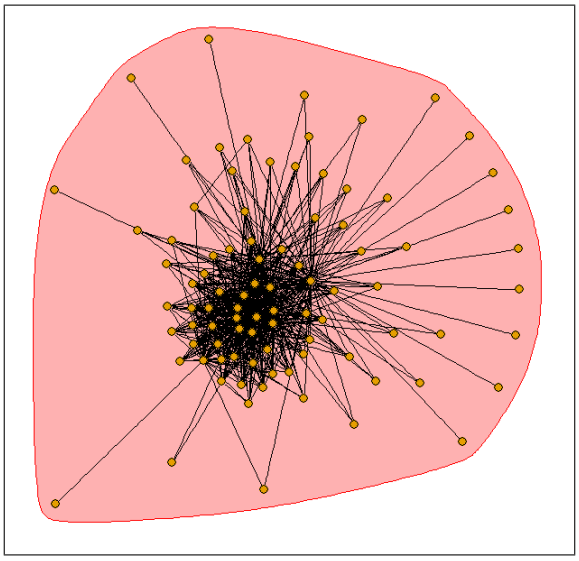

Figura 4.4: Comunidades con filtro de 10 % para el grupo A de supermercados.

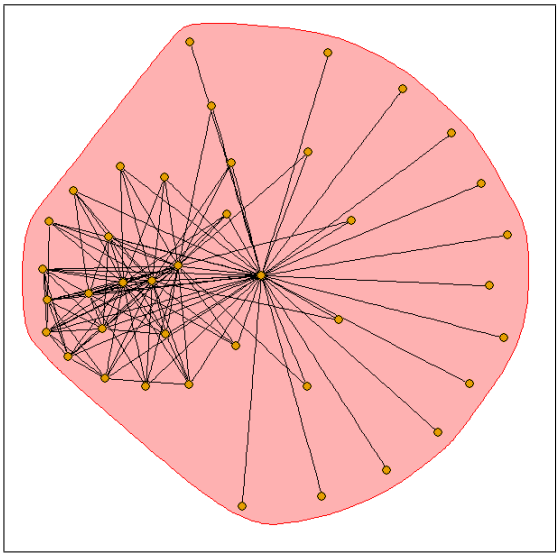

Figura 4.5: Comunidades con filtro de 25 % para el grupo A de supermercados.

En relación al grupo B de supermercados, se encontraron 57.550 boletas distintas, sobre las cuales se determinaron 310 reglas de marcas con un umbral mínimo $s = 57$, al igual que para el grupo A corresponde al 0,1 % de la cantidad total de boletas.

El cálculo del límite superior de los tres bordes más pesados dio como resultado lo siguiente:

$$tthet = \frac{1.767 + 1.516 + 1.432}{3} = 1.571,7 \quad , \tag{4.2}$$

De la misma manera que en el caso anterior, el valor de *tthet* es multiplicado por valores entre 0,05 y 1 para crear filtros que vayan aumentando de forma gradual (Tabla 4.5).

Porcentaje	tthet	Filtros
0,05	1.571,7	78,6
0,10	1.571,7	157,2
0,15	1.571,7	235,8
0,20	1.571,7	314,3
0,25	1.571,7	392,9
0,30	1.571,7	471,5
0,35	1.571,7	550,1
0,40	1.571,7	628,7
0,45	1.571,7	707,3
0,50	1.571,7	785,8
0,55	1.571,7	864,4
0,60	1.571,7	943,0
0,65	1.571,7	1.021,6
0,70	1.571,7	1.100,2
0,75	1.571,7	1.178,8
0,80	1.571,7	1.257,3
0,85	1.571,7	1.335,9
0,90	1.571,7	1.414,5
0,95	1.571,7	1.493,1
1,00	1.571,7	1.571,7

Tabla 4.5: Filtros calculados para el grupo B de supermercados.

En la Tabla 4.6 se observa la cantidad de reglas que se tiene por cada filtro. Además, se percibe la cantidad de reglas eliminadas o filtradas por cada uno de ellos, y cómo a medida que aumentan los porcentajes también aumenta la cantidad de reglas filtradas.

Filtros	Cantidad de Reglas	Cantidad de Reglas Filtradas
Sin Filtro	310	0
5 %	234	76
10 %	113	197
15 %	75	235
20 %	48	262
25 %	33	277

Tabla 4.6: Cantidad de reglas por cada filtro en el grupo B de supermercados.

En primera instancia se genera un grafo sin filtros, el cual contiene la totalidad de las reglas, para luego elaborar los demás grafos con los distintos filtros calculados en la tabla anterior. En total se crearon siete grafos. En este caso se pueden ver en las Figuras 4.6 y 4.7 el grafo sin filtro y con un filtro de 10 % respectivamente.

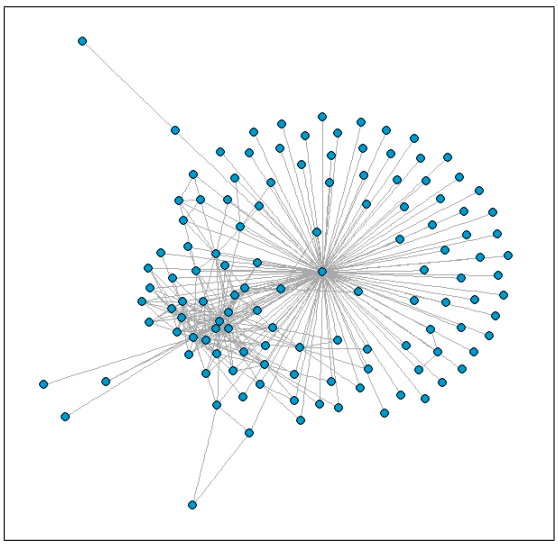

Figura 4.6: Grafo sin Filtro para el grupo B de supermercados.

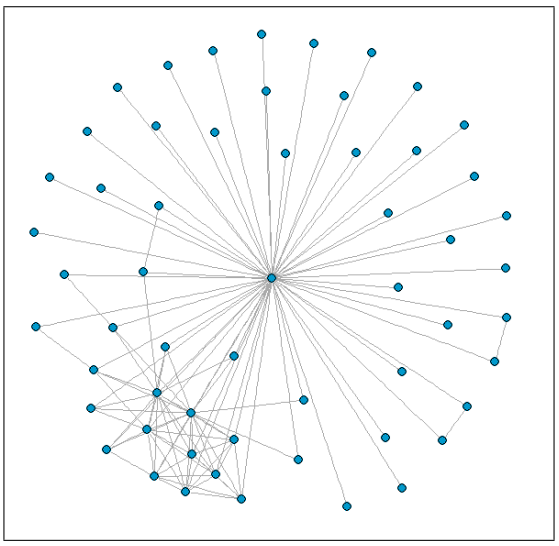

Figura 4.7: Grafo con Filtro de 10 % para el grupo B de supermercados.

Aplicando el algoritmo para la detección de comunidades sobre los distintos grafos construidos se encontraron resultados positivos, obteniendo diferentes comunidades de marcas de productos, las cuales se pueden apreciar en las Figuras 4.8, 4.9 y 4.10.

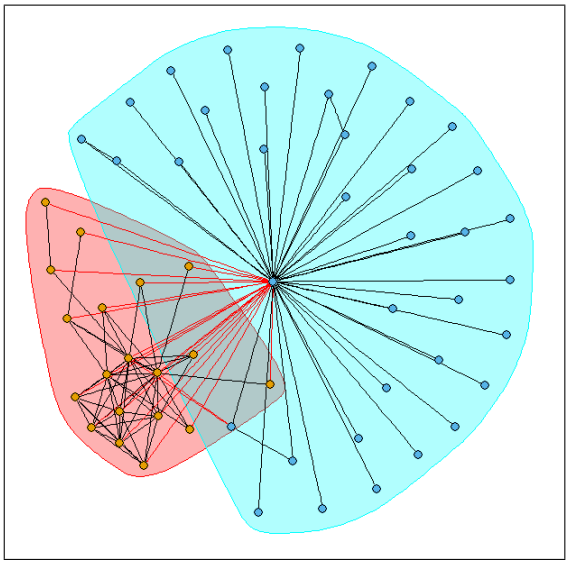

Figura 4.8: Comunidades con filtro de 10 % para el grupo B de supermercados.

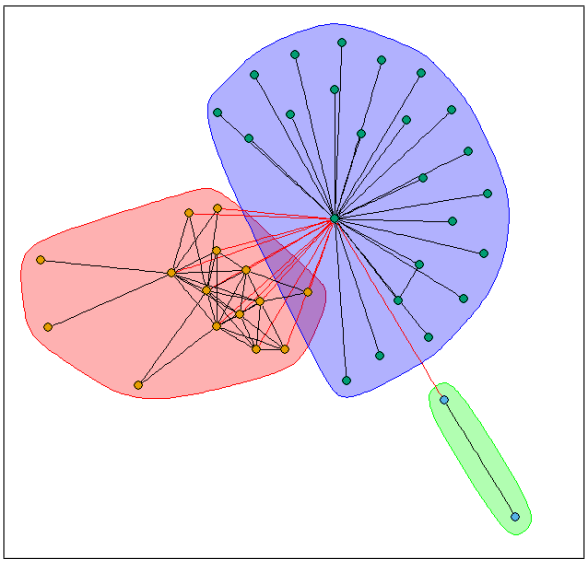

Figura 4.9: Comunidades con filtro de 15 % para el grupo B de supermercados.

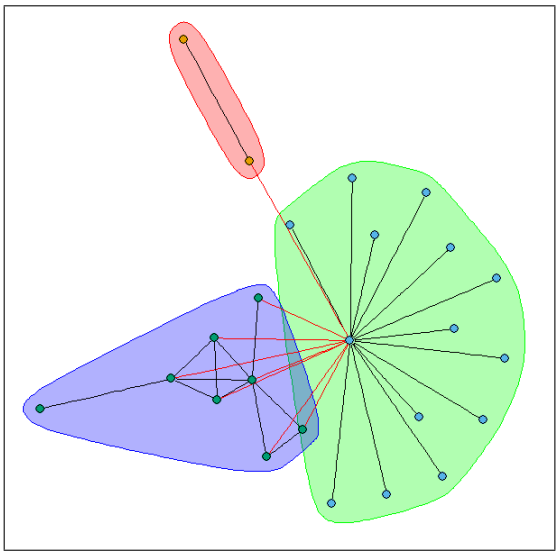

Figura 4.10: Comunidades con filtro de 25 % para el grupo B de supermercados.

Para evaluar la calidad de las comunidades descubiertas, se calcularon las modularidades por cada filtro, los resultados de esta medida se pueden ver en la Tabla 4.7:

Filtros	Modularidad
Sin Filtro	0,178
5 %	0,167
10 %	0,217
15 %	0,245
20 %	0,253
25 %	0,210

Tabla 4.7: Modularidad de cada filtro para el grupo B de supermercados.

Se puede apreciar que las modularidades de las comunidades no son muy buenas, a pesar que a medida que aumentan los porcentajes de los filtros, estas también lo hacen.

Antes de profundizar en las comunidades de marcas de productos, es necesario aclarar que por motivos de confidencialidad de los datos no se pueden presentar las tablas de las comunidades con sus respectivas marcas, es por lo que se realizó una clasificación de estas dependiendo del rubro al cual esté asociada.

En las Tablas 4.8, 4.9 y 4.10 se presentan las comunidades con los diferentes filtros, en estas se muestran las diferentes categorías y la frecuencia de cada una de estas que componen las distintas comunidades. Al mirar en detalle las categorías de las marcas al interior de las comunidades, se aprecia cómo a medida que cambian los filtros se van modificando los elementos de estas.

Con un filtro del 10 % (Tabla 4.8) en la comunidad 1, se encuentran principalmente marcas asociadas a productos para mascotas, tales como accesorios, juguetes, alimentos, entre otros, pero además se puede ver una categoría de energía la cual corresponden a baterías y/o pilas. Paralelamente, en la comunidad 2 se aprecia una mayor diversidad de categorías, tales como productos para decoración de interiores, electrodomésticos, electrónica, automóvil, entre otras.

Comunidad	Categoría	Cantidad de marcas
1	Mascota	15
1	Energía	2
1	Hogar	1
2	Hogar	7
2	Decohogar	5
2	Electrodomésticos	4
2	Automóvil	3
2	Ferretería Hogar y Automóvil	3
2	Infantil	3
2	Multimedia	3
2	Artículos Cocktail	3
2	Electrónica	2
2	Energía	2
2	Paquetería	2

Tabla 4.8: Comunidades con filtro de 10 % para el grupo B de supermercados.

Al comparar las comunidades anteriores con las descubiertas al utilizar un filtro del 15 % (Tabla 4.9), se produjeron algunas modificaciones. Lo primero que se observó es que ahora se determinaron tres comunidades; en la comunidad 1 se encuentra solamente la categoría de mascota, puesto que las dos marcas asociadas a la categoría de energía se trasladaron a la comunidad 3 junto con otras marcas asociadas a esta. Por otro lado, en la comunidad 2 se encuentran las categorías de hogar y multimedia, en esta última los productos que la componen son principalmente de audio y video.

Comunidad	Categoría	Cantidad de marcas
1	Mascota	14
2	Hogar	1
2	Multimedia	1
3	Decohogar	5
3	Electrodomésticos	4
3	Energía	4
3	Hogar	4
3	Ferretería Hogar y Automóvil	2
3	Automóvil	1
3	Electrónica	1
3	Mascota	1
3	Paquetería	1
3	Artículos Cocktail	1

Tabla 4.9: Comunidades con filtro de 15 % para el grupo B de supermercados.

En el caso del filtro de 25 %, las marcas cuya categoría es la de mascota, se dividieron en dos comunidades, 1 y 3, como se puede apreciar en la Tabla 4.10. Esto último da indicios de que las 2 marcas de la primera comunidad tienen un lazo mayor entre ellas que con las demás que fueron apartadas en otra.

Comunidad	Categoría	Cantidad de marcas
1	Mascota	2
2	Decohogar	4
2	Energía	3
2	Electrodomésticos	2
2	Mascota	2
2	Ferretería Hogar y Autómovil	1
2	Paquetería	1
2	Hogar	1
3	Mascota	7

Tabla 4.10: Comunidades con filtro de 25 % para el grupo B de supermercados.

Es relevante entender que cuando cambian las categorías de una comunidad a otra debido al porcentaje de filtro aplicado, no implica que necesariamente sigue conteniendo las mismas marcas en su interior.

Los resultados obtenidos en el grupo A de supermercados no fueron los esperados,

puesto que no se encontraron comunidades con los datos trabajados, mientras que en el grupo B sí se lograron determinar. Bajo este escenario, se tomó la decisión de realizar el mismo procedimiento con otro conjunto de datos correspondientes a una cadena de tiendas por departamentos.

4.3 Resultados de tiendas por departamentos

En tiendas por departamentos se hallaron 268.273 boletas en el interior de las transacciones, se calcularon 317 reglas utilizando un umbral mínimo correspondiente al 0,1 % de la cantidad total de boletas, al igual que en ambos grupos de supermercado, es decir, $s = 268$.

Al realizar el cálculo del límite superior de los tres bordes más pesados, se obtuvo lo siguiente:

$$tthet = \frac{6.105 + 5.440 + 4.579}{3} = 5.374,7 \quad , \tag{4.3}$$

De igual manera que en los casos anteriores, los filtros se calcularon en base a una proporción del *tthet*. Ver Tabla 4.12.

Filtros	**Cantidad de Reglas**	**Cantidad de Reglas Filtradas**
Sin Filtro	317	0
5 %	317	0
10 %	137	180
15 %	88	229
20 %	58	259
25 %	44	273
30 %	37	280

Tabla 4.11: Cantidad de reglas por cada filtro en tiendas por departamentos.

Porcentaje	tthet	Filtros
0,05	5.374,7	268,7
0,10	5.374,7	537,5
0,15	5.374,7	806,2
0,20	5.374,7	1.074,9
0,25	5.374,7	1.343,7
0,30	5.374,7	1.612,4
0,35	5.374,7	1.881,1
0,40	5.374,7	2.149,9
0,45	5.374,7	2.418,6
0,50	5.374,7	2.687,3
0,55	5.374,7	2.956,1
0,60	5.374,7	3.224,8
0,65	5.374,7	3.493,5
0,70	5.374,7	3.762,3
0,75	5.374,7	4.031,0
0,80	5.374,7	4.299,7
0,85	5.374,7	4.568,5
0,90	5.374,7	4.837,2
0,95	5.374,7	5.105,9
1,00	5.374,7	5.374,7

Tabla 4.12: Filtros calculados para tiendas por departamentos.

La Tabla 4.11 muestra el porcentaje de filtro que fue aplicado, además la cantidad de reglas que contiene cada filtro y las que fueron eliminadas. En este caso particular, al aplicar el filtro del 5 % no se elimina ninguna regla, por lo tanto, los grafo construidos sin ningún tipo de filtro y con un filtro del 5 % serán totalmente iguales. La cantidad de grafos construidos son siete en total, que van desde el filtro de 5 % hasta 30 % más el grafo natural sin filtro.

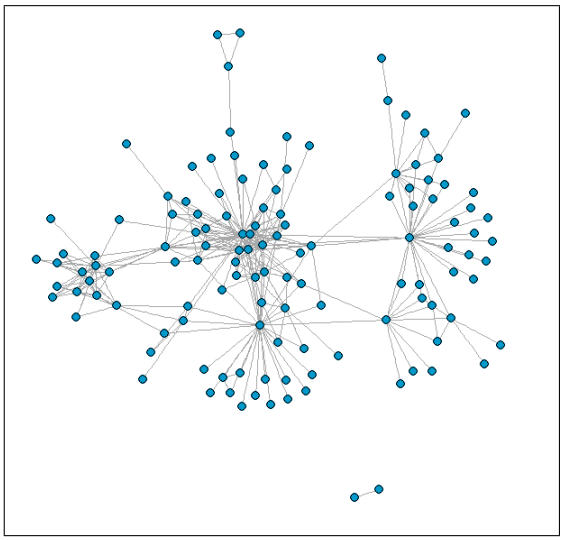

Figura 4.11: Grafo sin filtro para tiendas por departamentos.

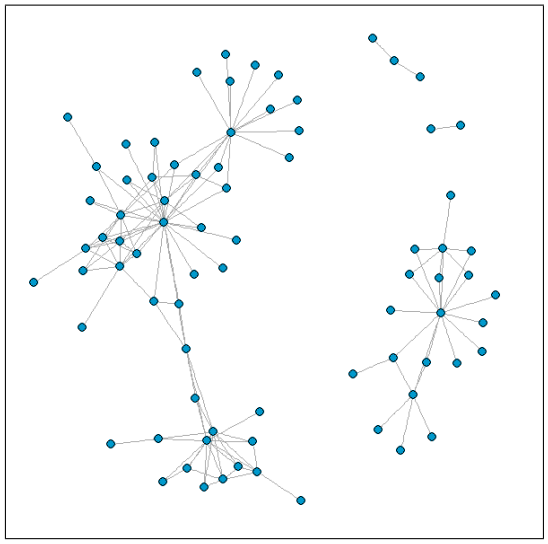

Figura 4.12: Grafo con filtro de 10 % para tiendas por departamentos.

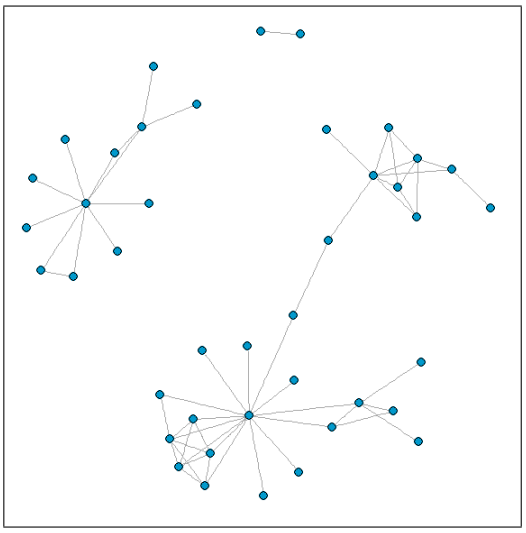

Figura 4.13: Grafo con Filtro de 20 % para tiendas por departamentos.

En la Figura 4.11 se muestra un grafo construido sin filtro, en este se puede apreciar cómo los nodos están agrupados de manera natural, visualizando así de antemano cuáles son las posibles comunidades que puedan existir en el interior del grafo, lo mismo se puede ver en la Figura 4.12 y 4.13.

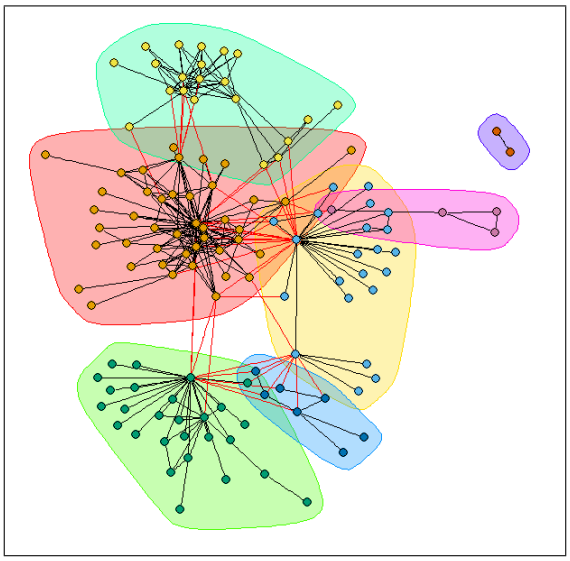

Figura 4.14: Comunidades sin filtro para tiendas por departamentos.

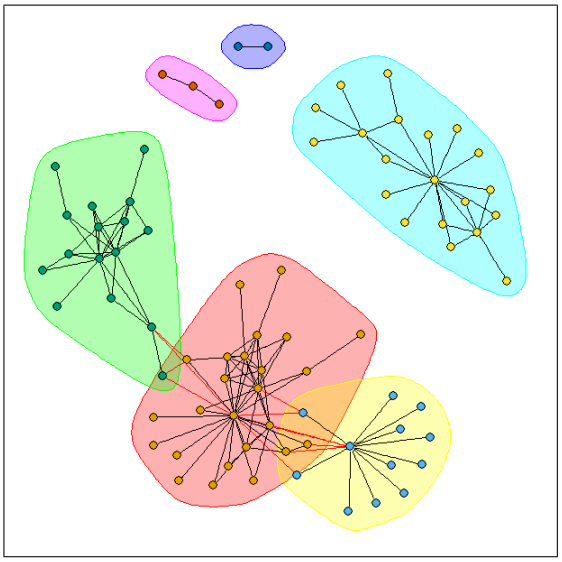

Figura 4.15: Comunidades con filtro de 10 % para tiendas por departamentos.

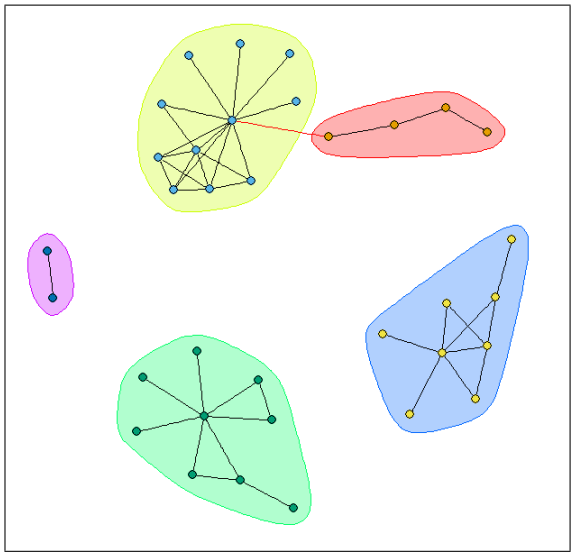

Figura 4.16: Comunidades con filtro de 25 % para tiendas por departamentos.

Las comunidades descubiertas en la cadena de tiendas por departamentos están muy bien definidas (Figuras 4.14, 4.15 y 4.16), así lo demuestran los resultados que se obtuvieron al calcular la modularidad. Los valores resultantes de esta medida

son bastante buenos y fluctúan entre 0,53 y 0,65, dependiendo del filtro ocupado. Al igual que en el caso del Grupo B de supermercados, las modularidades van aumentando a medida que aumenta el porcentaje de los filtros (Tabla 4.13).

Filtros	**Modularidad**
Sin Filtro	0,531
5 %	0,531
10 %	0,607
15 %	0,634
20 %	0,647
25 %	0,647
30 %	0,637

Tabla 4.13: Modularidad de cada filtro para tiendas por departamentos.

Al analizar detenidamente las comunidades descubiertas en tiendas por departamentos, se observa que existen marcas que a pesar de que están asociadas a la misma categoría no se agruparon en una misma comunidad. Esto puede deberse a diferentes factores, uno de ellos es la gran variedad de productos que contienen estas marcas, otros son el precio y/o la calidad de las marcas asociadas a los productos.

En la Tabla 4.14 se puede percibir lo mencionado en el párrafo anterior, ya que las comunidades 1 y 4 están compuestas por categorías cuyas marcas son de vestuario ya sea infantil, juvenil o formal. Una categoría que diferencia a estas dos comunidades es que en la 4 está la mayor cantidad de marcas deportivas, salvo 2 que están solas en la comunidad 5. Además, en la comunidad 1 existen categorías como Carteras y Belleza, que consta de productos como maquillaje, cremas y demás, que se pueden unir porque aparecen marcas que en general se pueden llevar en conjunto con las demás categorías. Esto también se ve reflejado en las otras comunidades, por ejemplo, en la comunidad 2 se observan en su mayoría las categorías de Juguetes y Vestuario Infantil o en la comunidad 3 que incluye marcas afiliadas a las categorías de Hogar, Electrónica y Electrodomésticos, las cuales en, ambos casos, tienen sentido que las categorías pertenezcan a una misma comunidad, lo que lleva a que tengan una fuerte relación.

Comunidad	Categoría	Cantidad de marcas
1	Ropa Interior	10
1	Vestuario Juvenil	10
1	Vestuario y Calzado	7
1	Vestuario	6
1	Carteras	2
1	Vestuario Formal	2
1	Vestuario Infantil	2
1	Belleza	1
2	Vestuario Infantil y Juguetería	6
2	Juguetería	5
2	Vestuario Infantil	5
2	Hogar	3
2	Electrónica y Electrodomésticos	1
2	Ropa Interior	1
2	Vestuario Juvenil	1
3	Electrodomésticos	10
3	Electrónica	5
3	Electrónica y Electrodomésticos	5
3	Hogar	3
3	Juguetería	1
4	Vestuario y Accesorios Deportivos	6
4	Vestuario y Calzado	5
4	Vestuario Formal	4
4	Vestuario Infantil	2
4	Juguetería	1
4	Ropa Interior	1
4	Vestuario Infantil y Juguetería	1
4	Vestuario Juvenil	1
5	Hogar	7
6	Vestuario y Accesorios Deportivos	2
7	Belleza	4

Tabla 4.14: Comunidades sin filtro para tiendas por departamentos.

En comparación con la Tabla 4.14, en la Tabla 4.15 que tiene las comunidades con un filtro del 10 %, se observa que la cantidad de comunidades disminuyó de 7 a 6 y en ellas también la cantidad de marcas de 120 a 74. Por otro lado, las comunidades se mantuvieron, y no se nota gran variación, en cuanto a las categorías incluidas en

ellas, pero sí en el número que identifica la comunidad. En cuanto a la comunidad 1, se percibe que hay marcas asociadas a Vestuario, al igual que en la 3, en cambio la 2 posee mayoritariamente categorías de Vestuario Infantil y Juguetería, y en la 4 hay categorías que tienen relación con Hogar, Electrónica y Electrodomésticos, la 5 y 6 tienen categorías de Belleza y de Vestuario y Accesorios Deportivos respectivamente.

Comunidad	**Categoría**	**Cantidad de marcas**
1	Vestuario y Calzado	7
1	Vestuario Juvenil	6
1	Ropa Interior	3
1	Carteras	2
1	Vestuario	2
1	Vestuario Formal	2
1	Vestuario Infantil	1
2	Vestuario Infantil	5
2	Vestuario Infantil y Juguetería	4
2	Juguetería	1
2	Ropa Interior	1
2	Vestuario	1
3	Vestuario y Calzado	4
3	Ropa Interior	3
3	Vestuario Formal	3
3	Vestuario Infantil	2
3	Vestuario y Accesorios Deportivos	2
3	Juguetería	1
3	Vestuario Juvenil	1
4	Hogar	9
4	Electrodomésticos	4
4	Electrónica y Electrodomésticos	4
4	Electrónica	1
5	Belleza	2
6	Vestuario y Accesorios Deportivos	3

Tabla 4.15: Comunidades con filtro de 10 % para tiendas por departamentos.

La Tabla 4.16 representa a las comunidades con un filtro del 25 %, en este caso las comunidades sufrieron cambios evidentes en comparación con las Tablas 1 y 2. A primera vista se ve que la cantidad de categorías en las comunidades disminuyeron considerablemente, ya que en general las comunidades se reorganizan de un filtro

a otro. A causa de esto, se adicionó una nueva comunidad de Vestuario. Las comunidades 1, 2 y 4 se basan en Vestuario, pero con diferentes marcas; la 3 está constituida por Hogar, Electrónica y Electrodomésticos y la 5 consta de Vestuario y Accesorios Deportivos. Cabe destacar que las marcas de la categoría de Vestuario y Accesorios Deportivos se mantienen en una comunidad, lo que no necesariamente implica que contenga las mismas que en las Tablas 4.14 y 4.15.

Comunidad	**Categoría**	**Cantidad de marcas**
1	Vestuario Infantil	3
1	Vestuario Juvenil	1
2	Vestuario y Calzado	5
2	Vestuario Juvenil	3
2	Ropa Interior	1
2	Vestuario	1
2	Vestuario Formal	1
3	Hogar	5
3	Electrónica y Electrodomésticos	2
4	Vestuario y Calzado	4
4	Vestuario Infantil	2
4	Vestuario Formal	1
4	Vestuario Juvenil	1
5	Vestuario y Accesorios Deportivos	2

Tabla 4.16: Comunidades con filtro de 25 % para tiendas por departamentos.

4.4 Análisis de Resultados

Con el fin de comprender por qué en los datos pertenecientes al grupo A de supermercados no se encontraron comunidades, mientras que en el grupo B de este mismo y en tiendas por departamentos los resultados fueron los esperados, se realizó un estudio donde se calculó la cantidad de boletas con igual cantidad de marcas, es decir, se contaron las boletas con 2 marcas, luego las que tenían 3 y así sucesivamente.

Al mirar los resultados de este estudio (Figuras 4.17, 4.18 y 4.19) se puede ver que existen hasta 89 marcas distintas dentro de una boleta en el grupo A de supermercados. Aunque este sólo es un caso que contrasta radicalmente con el grupo B y tiendas por departamentos, las cuales cuentan con un máximo de 13 y 17 marcas diferentes en el interior de una boleta, respectivamente. Por otro lado, en el grupo A se puede apreciar que son bastante más los casos en los que

se tiene una gran cantidad de boletas con varias marcas, por ejemplo existen 559 boletas con 31 marcas distintas, mientras que en el grupo B y tiendas por departamentos la barrera de las 500 boletas se encuentra en las 5 y 7 marcas distintas, respectivamente. Esto muestra la gran cantidad de marcas distintas que los clientes llevan de manera conjunta en una sola boleta en el Grupo A.

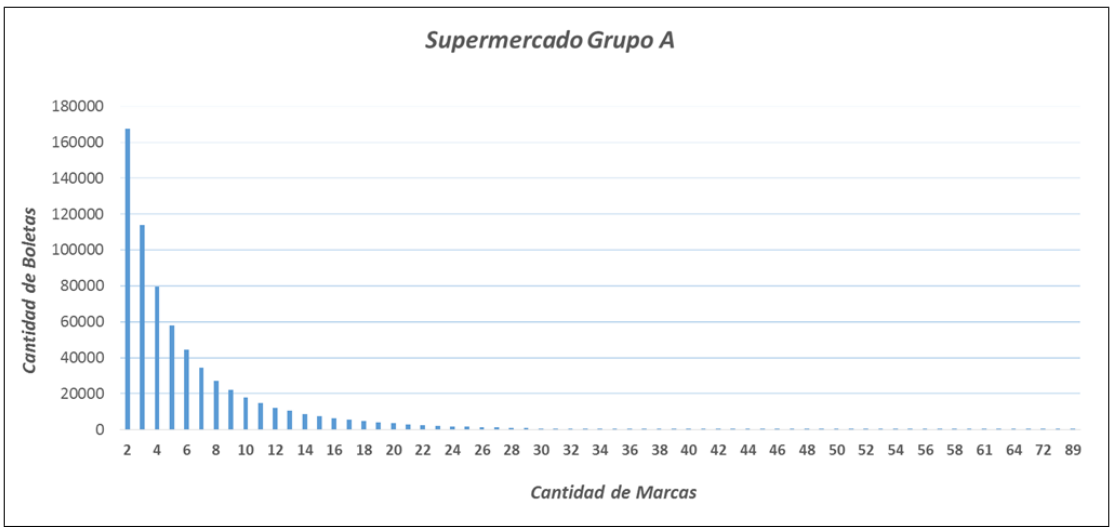

Figura 4.17: Cantidad de boletas con igual número de marcas en el grupo A de supermercados.

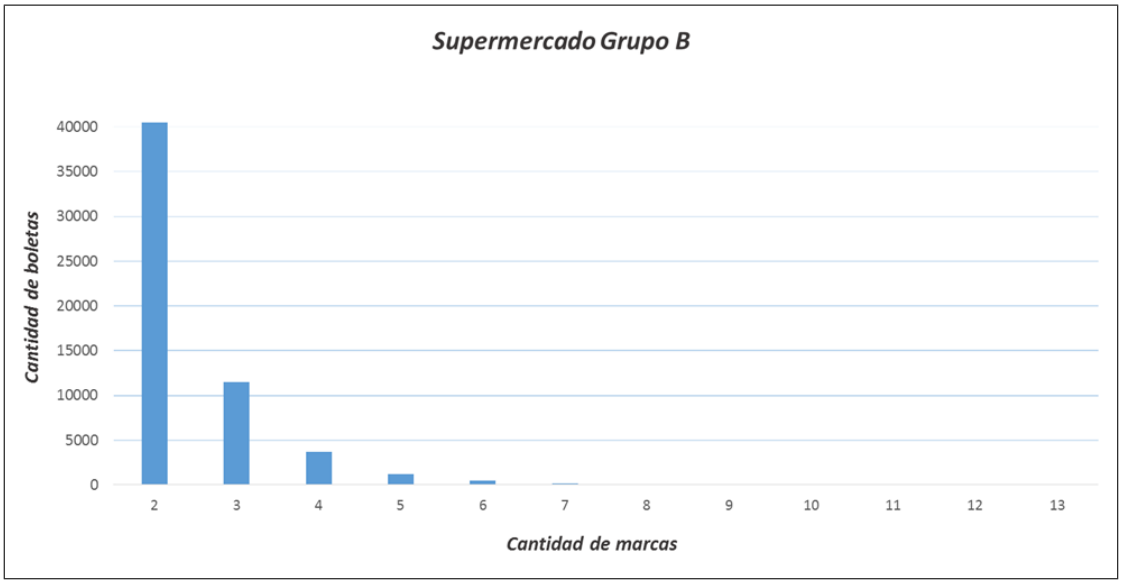

Figura 4.18: Cantidad de boletas con igual número de marcas en el grupo B de supermercados.

Figura 4.19: Cantidad de boletas con igual número de marcas en tiendas por departamentos.

Debido a lo que se observó sobre el grupo A, se decidió tomar en cuenta sólo las boletas que contenían de 2, 3 y 4 marcas distintas para replicar el trabajo realizado, con el fin de verificar si efectivamente esto afectó directamente al momento de buscar las comunidades.

4.5 Resultados de supermercados grupo A versión 2

En el nuevo grupo A de la cadena de supermercados, se hallaron 350.278 boletas con 2, 3 y 4 marcas distintas que están en el interior de las transacciones estudiadas. Se obtuvieron 4.573 reglas en el proceso del cálculo de los conjuntos de elementos frecuentes con un umbral mínimo correspondiente al 0,01 % de la cantidad total de boletas, es decir, $s = 35$.

Al realizar el cálculo del límite superior de los tres bordes más pesados en este caso, se obtuvo lo siguiente:

$$tthet = \frac{5.227 + 3.488 + 3.029}{3} = 3.914,7 \quad , \tag{4.4}$$

A continuación, se presentan los filtros obtenidos al multiplicar el valor del *tthet* por porcentajes que van aumentando de manera gradual del mismo modo que en los casos anteriores (Tabla 4.17).

Porcentaje	tthet	Filtros
0,05	3.914,7	195,7
0,10	3.914,7	391,5
0,15	3.914,7	587,2
0,20	3.914,7	782,9
0,25	3.914,7	978,7
0,30	3.914,7	1.174,4
0,35	3.914,7	1.370,1
0,40	3.914,7	1.565,9
0,45	3.914,7	1.761,6
0,50	3.914,7	1.957,3
0,55	3.914,7	2.153,1
0,60	3.914,7	2.348,8
0,65	3.914,7	2.544,5
0,70	3.914,7	2.740,3
0,75	3.914,7	2.936,0
0,80	3.914,7	3.131,7
0,85	3.914,7	3.327,5
0,90	3.914,7	3.523,2
0,95	3.914,7	3.718,9
1,00	3.914,7	3.914,7

Tabla 4.17: Filtros calculados para el grupo A versión 2 de supermercados.

En la Tabla 4.18 se observa el porcentaje de filtro que fue aplicado, además la cantidad de reglas que contiene cada filtro y las que fueron eliminadas. Sobre la base de esto, se generaron 7 grafos en total, que van desde el filtro de 5 % hasta 30 % más el grafo natural sin filtro.

Filtros	Cantidad de Reglas	Cantidad de Reglas Filtradas
Sin Filtro	4.573	0
5 %	627	3.946
10 %	230	4.343
15 %	125	4.448
20 %	74	4.499
25 %	44	4.529
30 %	28	4.545

Tabla 4.18: Cantidad de reglas por cada filtro en el grupo A versión 2 de supermercados.

En las Figuras 4.20, 4.21 y 4.22 se exponen distintos grafos construídos con diferentes porcentajes de filtros aplicados en los datos. En la Figura 4.20, se tiene la gráfica de los datos cuando se les aplicó un filtro del 10 %, en el que no se observan comunidades tan definidas, ya que los nodos no logran separarse lo suficiente, no así en las Figuras 4.21 y 4.22 en donde a medida que se aumentan los filtros, se notan cada vez más las posibles comunidades.

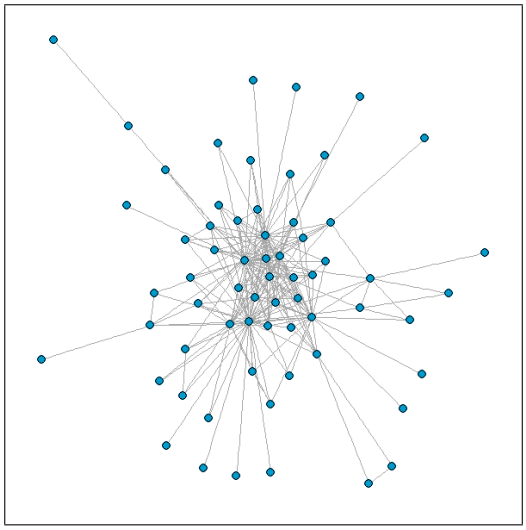

Figura 4.20: Grafo con filtro de 10 % para el grupo A versión 2 de supermercados.

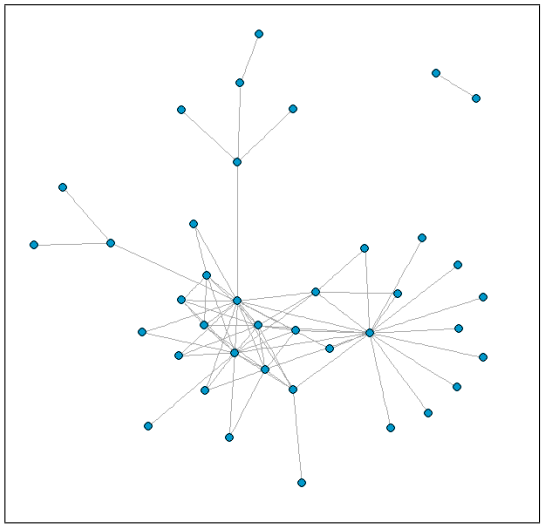

Figura 4.21: Grafo con filtro de 20 % para el grupo A versión 2 de supermercados.

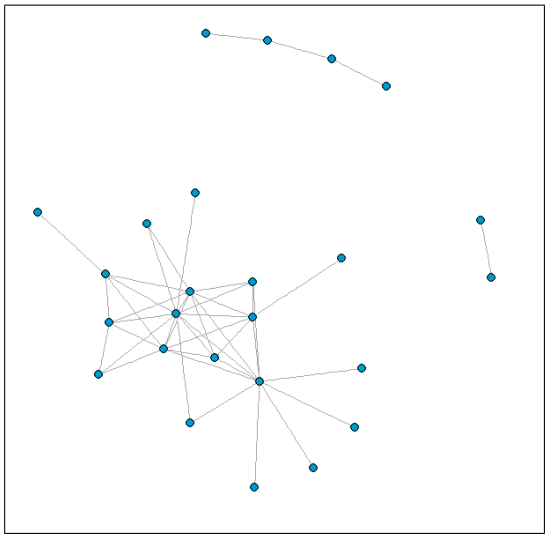

Figura 4.22: Grafo con filtro de 25 % para el grupo A versión 2 de supermercados.

Los resultados obtenidos al aplicar el algoritmo para determinar comunidades son positivos, en las Figuras 4.23, 4.24 y 4.25 se muestran las comunidades descubiertas para el grupo A de supermercados al replicar el trabajo realizado eliminando las boletas con más de 4 marcas distintas, lo que demuestra que la gran cantidad de boletas con una abultada cantidad de marcas afectó de manera negativa en la detección de comunidades.

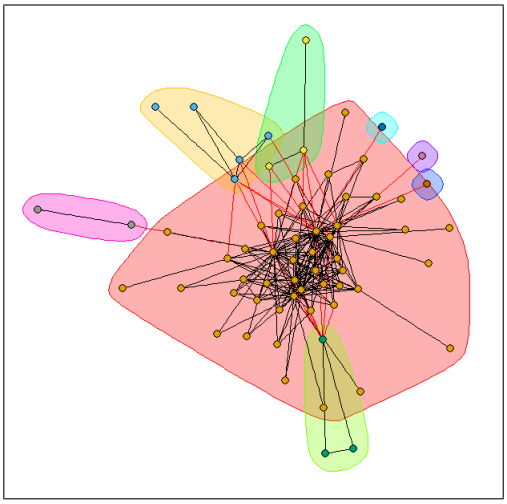

Figura 4.23: Comunidades con filtro de 10 % para el grupo A versión 2 de supermercados.

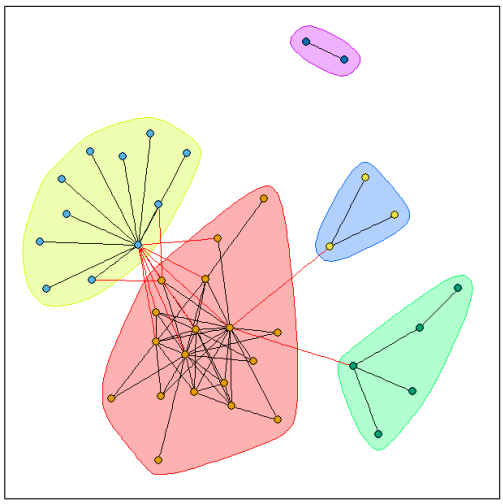

Figura 4.24: Comunidades con filtro de 20 % para el grupo A versión 2 de supermercados.

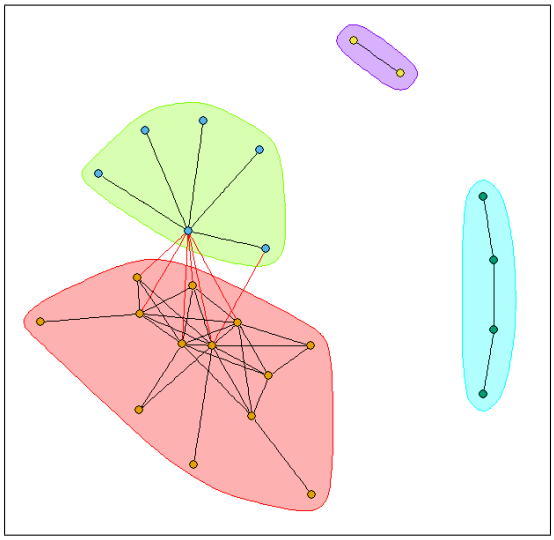

Figura 4.25: Comunidades con filtro de 25 % para el grupo A versión 2 de supermercados.

Las comunidades descubiertas en el nuevo grupo A no son muy buenas, sobre todo en los casos en donde se aplicó un porcentaje de filtro menor al 15 %, esto se puede apreciar por las modularidades calculadas (Tabla 4.19). Por otro lado, con los porcentajes de filtros entre el 20 y 30 % se obtiene una modularidad que casi alcanza un límite aceptable para un conjunto de comunidades.

Filtros	**Modularidad**
Sin Filtro	0,011
5 %	0,057
10 %	0,097
15 %	0,146
20 %	0,267
25 %	0,290
30 %	0,305

Tabla 4.19: Modularidad de cada filtro para el grupo A versión 2 de supermercados.

Al analizar con mayor detención las categorías que componen las diferentes comunidades obtenidas al utilizar un filtro del 10 % (Tabla 4.20), se aprecia que en la comunidad 1 existen una gran cantidad de marcas que pertenecen a diferentes categorías, al mirar estas en detalle es fácil ver la coherencia que tiene la comunidad, ya que son categorías que los clientes en general compran en conjunto,

como por ejemplo los jugos y bebidas con galletas y golosinas o con productos de cocktail como papas fritas, maní, entre otras. En las demás sólo hay una categoría por comunidad, aunque con varias marcas.

Comunidad	**Categoría**	**Cantidad de marcas**
1	Galletas y golosinas	11
1	Almacén	9
1	Jugos y Bebidas	9
1	Carnicería	7
1	Lácteos	6
1	Cocktail	3
1	Congelados	2
2	Licorería	5
3	Galletas y golosinas	3
4	Almacén	3
5	Almacén	1
6	Almacén	1
7	Jugos y Bebidas	1
8	Cocktail	2

Tabla 4.20: Comunidades con filtro de 10 % para el grupo A versión 2 de supermercados.

En la Tabla 4.21, que expone el estudio con un filtro de 20 %, se define un total de 5 comunidades. Como primera observación se tiene que, a pesar de que existen categorías que se repiten en las distintas comunidades, estas representan diferentes marcas. Por otro lado, está la comunidad 1 que no varía mucho de un filtro a otro, lo que sí se puede ver es que en esta disminuyó la cantidad de marcas por categorías. En cuanto a la comunidad 2, la cantidad de categorías asociadas a esta aumentó, ya que en la Tabla 4.20 la comunidad 2 se componía sólo por Licorería, en la Tabla 4.21 esta categoría está en la comunidad 3.

Comunidad	Categoría	Cantidad de marcas
1	Jugos y Bebidas	5
1	Galletas y golosinas	4
1	Almacén	3
1	Cocktail	3
1	Carnicería	1
1	Lácteos	1
2	Carnicería	4
2	Lácteos	3
2	Almacén	2
2	Congelados	2
3	Licorería	5
4	Galletas y golosinas	3
5	Almacén	1
5	Lácteos	1

Tabla 4.21: Comunidades con filtro de 20 % para el grupo A versión 2 de supermercados.

En el caso del filtro del 25 % (Tabla 4.22), la comunidad 1 permanece similar a los filtros anteriormente detallados en las Tablas 4.20 y 4.21. Desde otra perspectiva, también se puede afirmar que esta tiene categorías con una alta relación entre sí, lo mismo se observa en la comunidad que consta con la categoría de Licorería, la cual mantuvo la cantidad de marcas en casi todos los filtros. Por otra parte, la categoría Almacén, que está asociada a productos como fideos, azúcar, arroz, entre otros, disminuye de manera considerable y ya no está presente en casi todas las comunidades, como pasaba en los filtros anteriores.

Comunidad	Categoría	Cantidad de marcas
1	Cocktail	3
1	Galletas y golosinas	3
1	Jugos y Bebidas	3
1	Almacén	2
1	Lácteos	1
2	Carnicería	4
2	Congelados	1
2	Lácteos	1
3	Licorería	4
4	Galletas y golosinas	2

Tabla 4.22: Comunidades con filtro de 25 % para el grupo A versión 2 de supermercados.

Capítulo 5

Conclusiones

En este trabajo de titulación se utilizaron diferentes métodos para lograr detectar comunidades de marcas de productos en datos transaccionales, tales como conjuntos de elementos frecuentes, algoritmos de detección de comunidades y grafos, entre otros; estos fueron implementados en los software de programación estadística Spark versión 1.6 y RStudio.

Por otro lado, se caracterizaron detalladamente los datos transaccionales y la jerarquía de productos como se puede apreciar en el Capítulo 2, en donde se analizaron las agrupaciones de productos de venta al por menor y su relación con la respectiva marca para la detección de conjuntos de elementos frecuentes. Asimismo, se indagó sobre la teoría de grafos, visualización y la construcción de estos, además de reconocer distintos algoritmos para la determinación de comunidades de marcas de productos.

En primera instancia la idea de este trabajo fue estudiar los datos transaccionales correspondientes a una cadena de supermercados, los cuales se dividieron en dos grupos. Los resultados obtenidos bajo la metodología definida son diametralmente opuestos, en el grupo A de supermercados no se obtuvieron comunidades de marcas de productos, mientras que en el grupo B los resultados fueron positivos para este estudio. Posteriormente, se agregó un nuevo conjunto de datos pertenecientes una cadena de tiendas por departamentos, en esta oportunidad se consiguieron resultados positivos e incluso mejor que en el grupo B de supermercado. No obstante, aún estaba la interrogante de por qué en el grupo A de

supermercados no se lograron los resultados esperados, por lo que fue necesario evaluar el comportamiento de los datos transaccionales en base a la cantidad de boletas con igual cantidad de marcas de estos tres grupos. Con los resultados de este análisis, se decidió replicar el estudio para el grupo A, pero esta vez utilizando sólo boletas con 2, 3 y 4 marcas distintas en donde ahora sí se lograron encontrar comunidades de marcas de productos.

En base al trabajo realizado, se puede concluir que efectivamente mediante el uso grafos es posible visualizar datos transaccionales agrupados para determinar comunidades de marcas de productos en el negocio de venta al por menor, siendo fundamental una buena administración de datos como se puede ver reflejado en el caso del grupo A de supermercados.

En el caso de tiendas por departamentos las comunidades obtenidas están muy bien definidas, lo que se ve reflejado en las modularidades ya que estas son realmente muy altas. En los grupos B y A versión 2 de supermercados, las comunidades pueden que no sean del todo buenas, pero con un porcentaje de filtro de 25 % y 30 % mejoran bastante y así lo refleja la medida de modularidad.

El tener una cantidad limitada de marcas dentro de cada boleta, facilita el proceso de hallar las posibles comunidades de marcas de productos que puedan existir en los datos. Esto lleva a concluir que este estudio es más factible de aplicar en un negocio cómo el de tiendas por departamentos cuya cantidad productos y marcas distintas que los clientes adquieren en una misma boleta es menor al de supermercados.

Capítulo 6

Referencias Bibliografía

- Blondel, V., Guillaume, J. L., Lambiotte, R. y Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), 2-10.
- Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2), 136-145.
- Caicedo, A., Wagner G. y Méndez, R. (2010). *Introducción a la Teoría de Grafos*. Quindío, Armenia: ELIZCOM.
- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1) 35-41.
- Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3), 75-174.
- Martínez, N. (2011). *Análisis, comparativa y visualización de redes sociales on-line representadas como grafos*(Tesis de pregrado). Universitat Pompeu Fabra, España.
- Meghanathan, N. (2016). A greedy algorithm for neighborhood overlap-based community detection. *Algorithms*, 9(1), 11-15.
- Menéndez, A. (1998). Una breve introducción a la teoría de grafos. *SUMA*, 28, 11-26.

- Newman, M. E. y Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2), 1-16.
- Pérez, D. y Perez, I. (2006). El Producto Concepto y Desarrollo. *España: Escuela de Negocios EOI*, 7-11.
- Raeder, T. y Chawla, N. V. (2009). Modeling a store?s product space as a social network. *In Proceedings of International Conference on Advances in Social Network Analysis and Mining*. IEEE Computer Society, Washington DC, USA, 164-169.
- Rajaraman, A. y Ullman, J. (2011). *Mining of massive datasets*. New York, USA: Cambridge University Press.
- Toranzos, F. (1976). *Introducción a la Teoría de Grafos*. Buenos Aires, Argentina: Oea.
- Videla, I. y Ríos, S. (2014). Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications*, 41(4), 1928-1936.

Capítulo 7

Anexos

7.1 Grafos

7.1.1 Supermercado grupo B

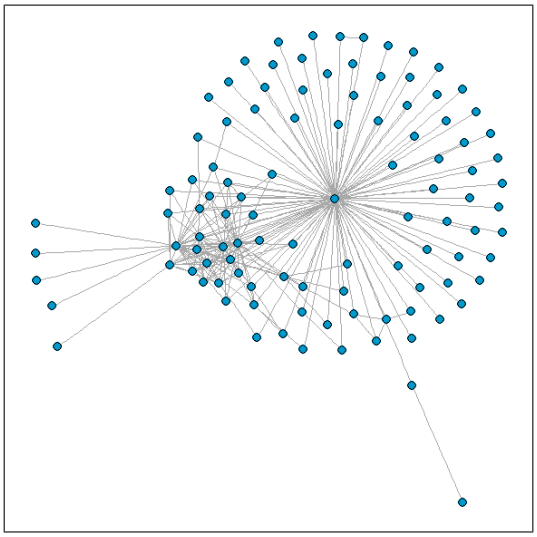

Figura 7.1: Grafo con filtro de 5 % para el grupo B de supermercados.

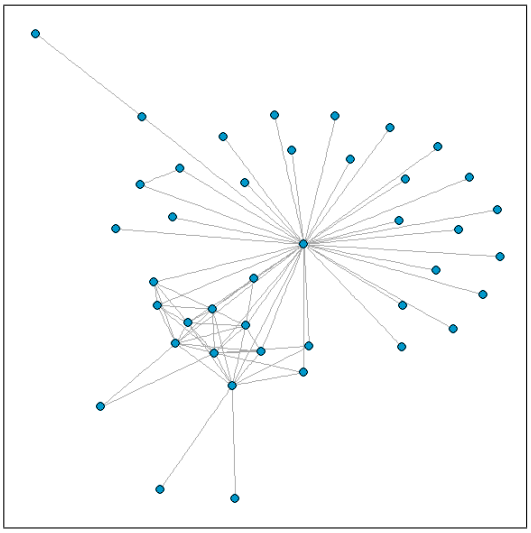

Figura 7.2: Grafo con filtro de 15 % para el grupo B de supermercados.

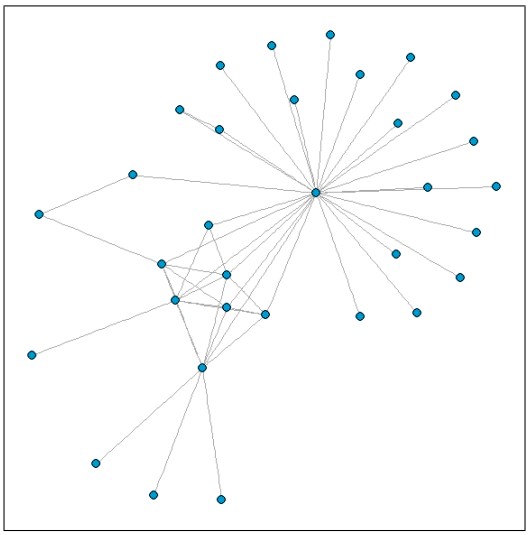

Figura 7.3: Grafo con filtro de 20 % para el grupo B de supermercados.

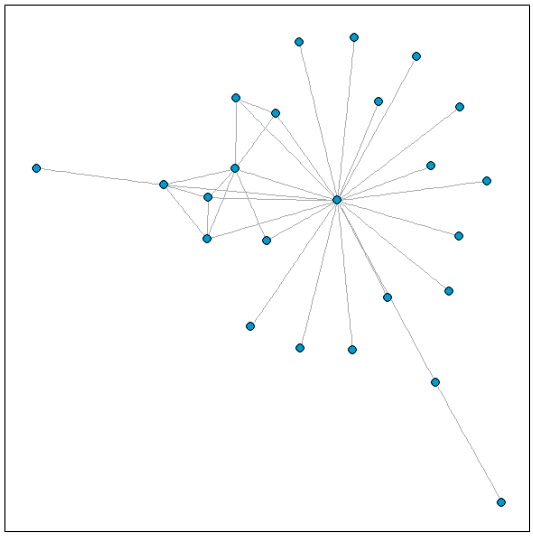

Figura 7.4: Grafo con filtro de 25 % para el grupo B de supermercados.

7.1.2 Tienda por departamentos

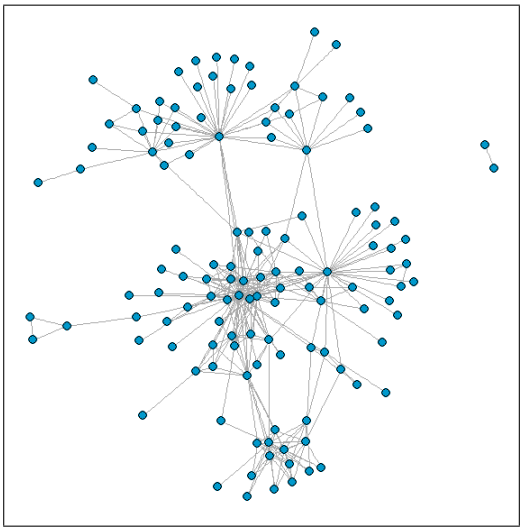

Figura 7.5: Grafo con filtro de 5 % para tiendas por departamentos.

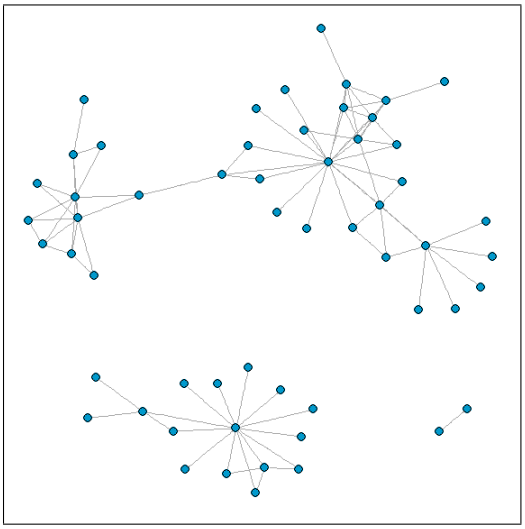

Figura 7.6: Grafo con filtro de 15 % para tiendas por departamentos.

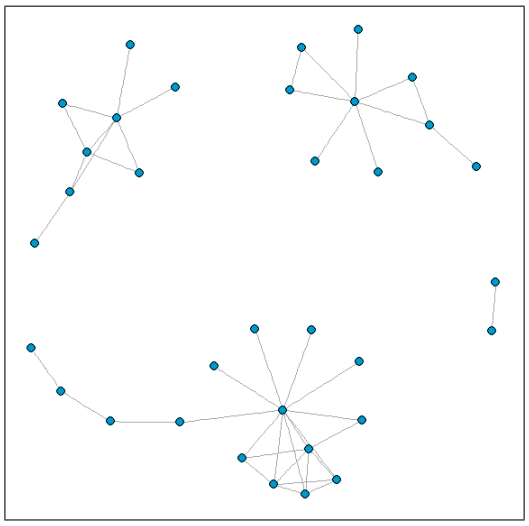

Figura 7.7: Grafo con filtro de 25 % para tiendas por departamentos.

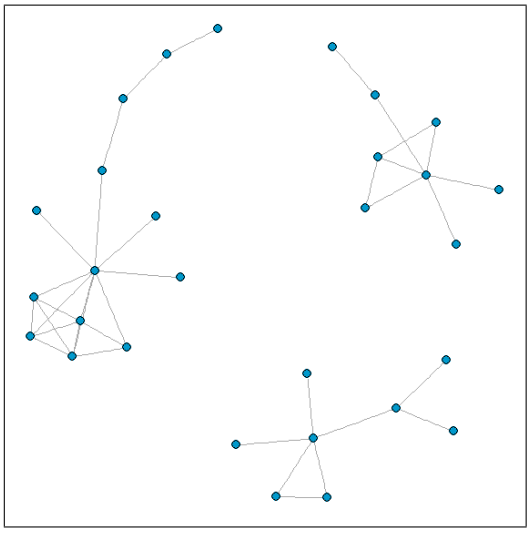

Figura 7.8: Grafo con filtro de 30 % para tiendas por departamentos.

7.1.3 Supermercado grupo A versión 2

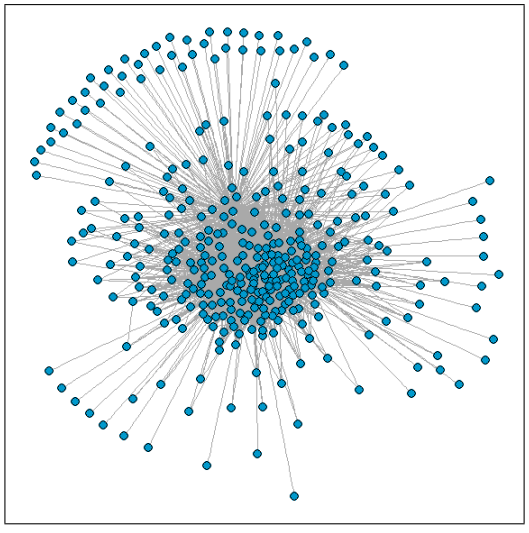

Figura 7.9: Grafo sin filtro para el grupo A versión 2 de supermercados.

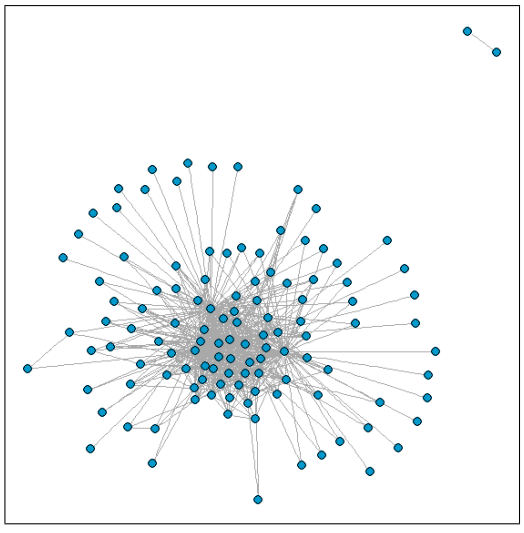

Figura 7.10: Grafo con filtro de 5 % para el grupo A versión 2 de supermercados.

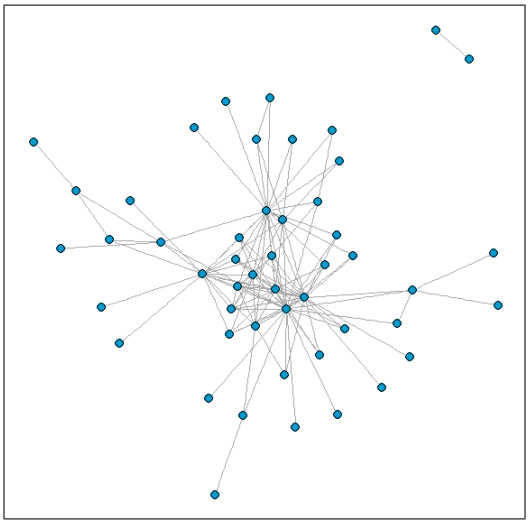

Figura 7.11: Grafo con filtro de 15 % para el grupo A versión 2 de supermercados.

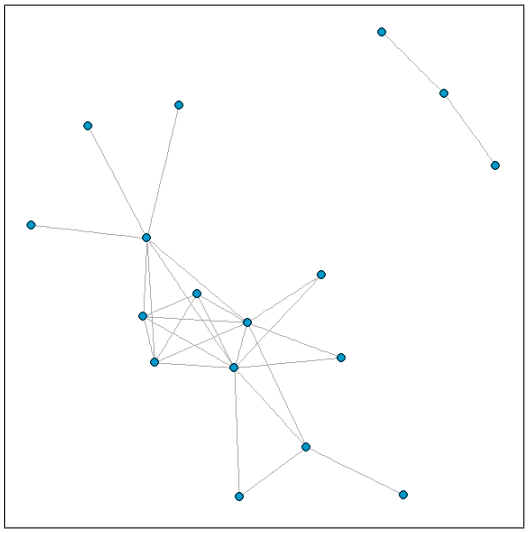

Figura 7.12: Grafo con filtro de 30 % para el grupo A versión 2 de supermercados.

7.2 Comunidades

7.2.1 Supermercado grupo B

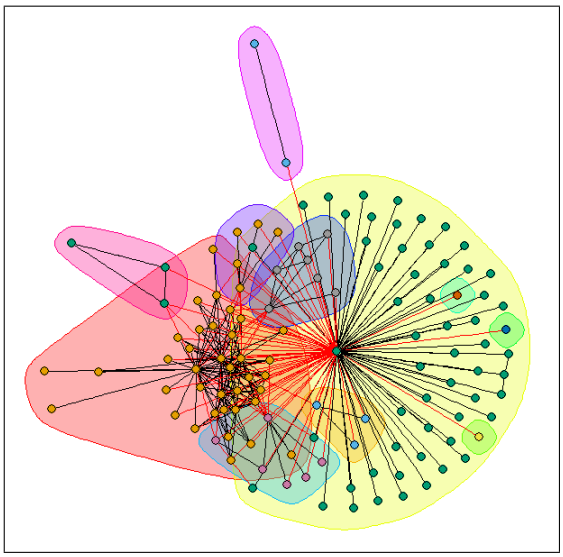

Figura 7.13: Comunidades sin filtro para el grupo B de supermercados.

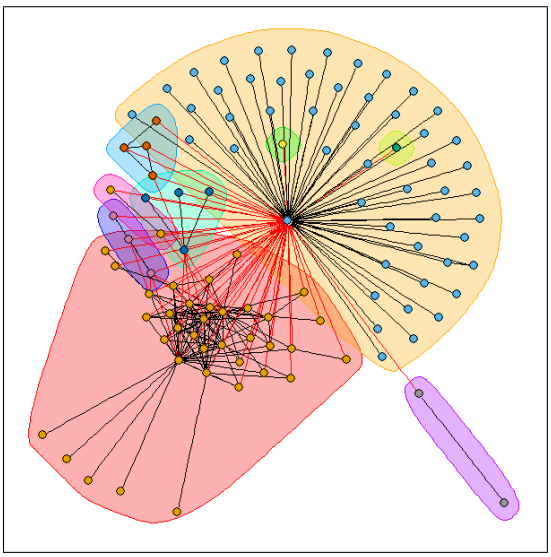

Figura 7.14: Comunidades con filtro de 5 % para el grupo B de supermercados.

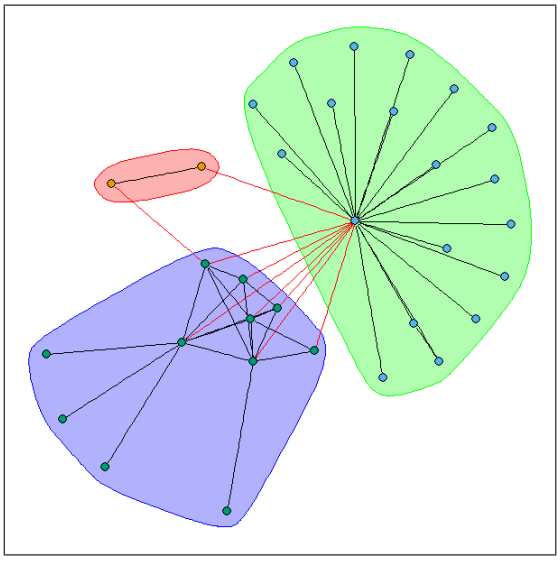

Figura 7.15: Comunidades con filtro de 20 % para el grupo B de supermercados.

7.2.2 Tienda por departamentos

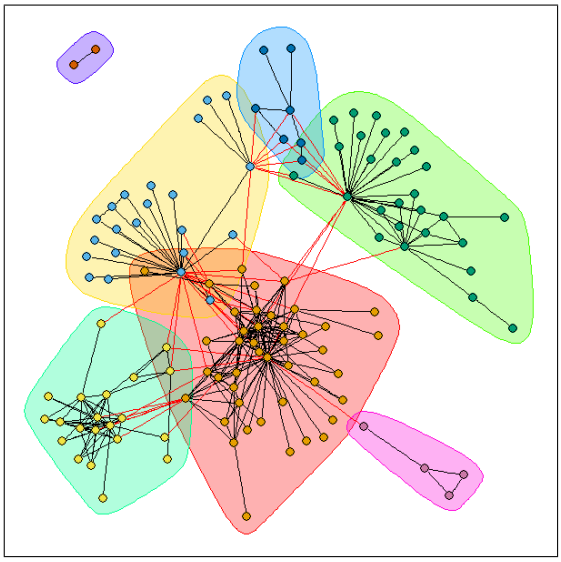

Figura 7.16: Comunidades con filtro de 5 % para tiendas por departamentos.

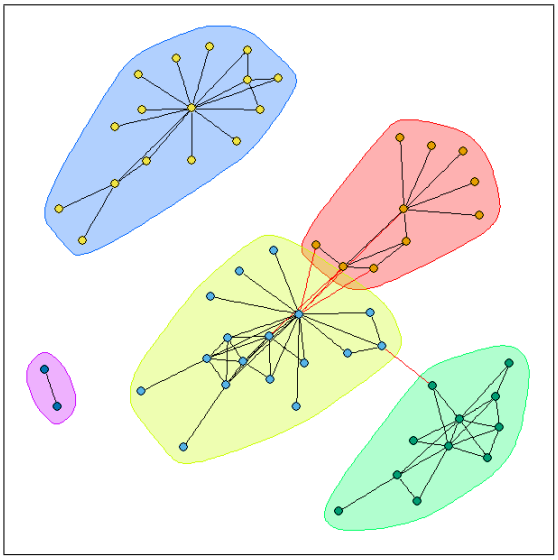

Figura 7.17: Comunidades con filtro de 15 % para tiendas por departamentos.

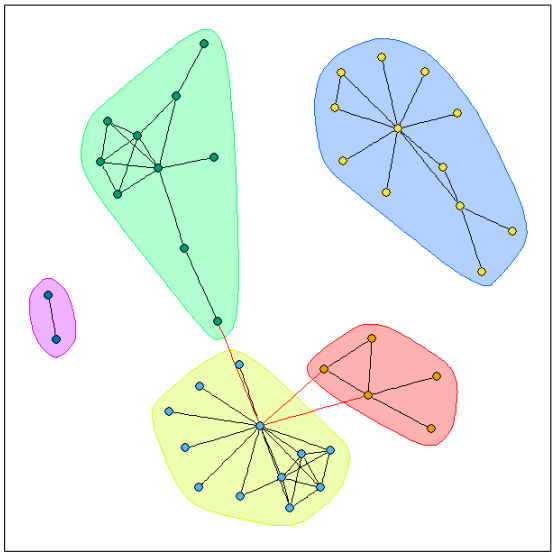

Figura 7.18: Comunidades de tienda por departamentos con filtro 20 %.

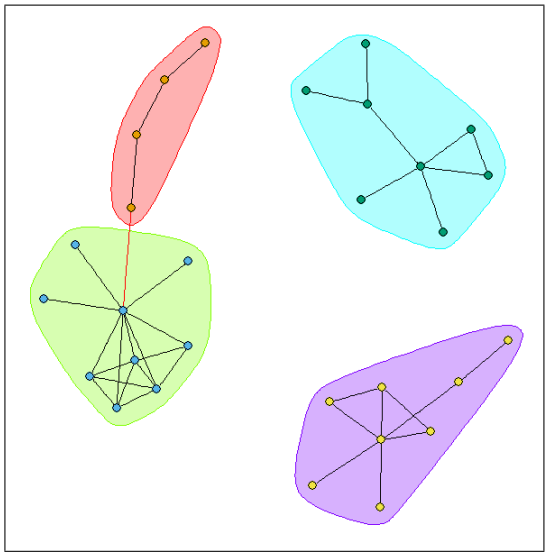

Figura 7.19: Comunidades con filtro de 30 % para tiendas por departamentos.

7.2.3 Supermercado grupo A versión 2

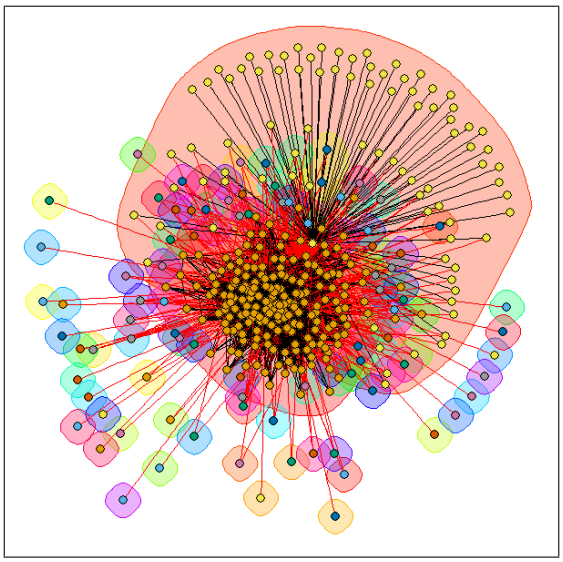

Figura 7.20: Comunidades sin filtro para el grupo A versión 2 de supermercados.

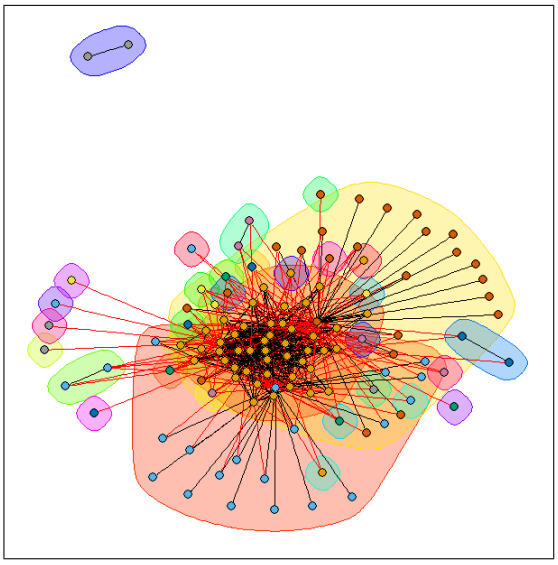

Figura 7.21: Comunidades con filtro de 5 % para el grupo A versión 2 de supermercados.

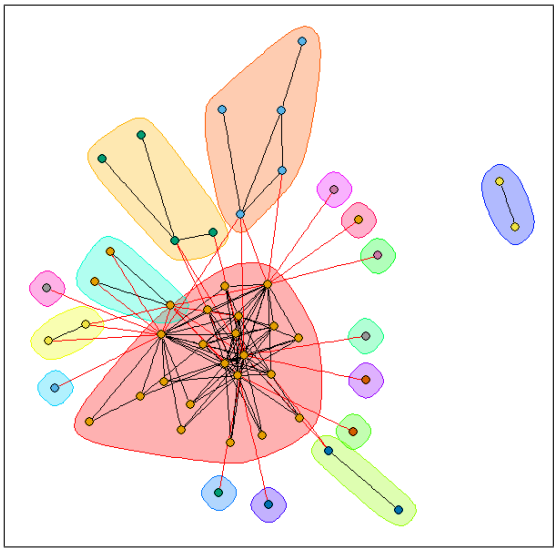

Figura 7.22: Comunidades con filtro de 15 % para el grupo A versión 2 de supermercados.

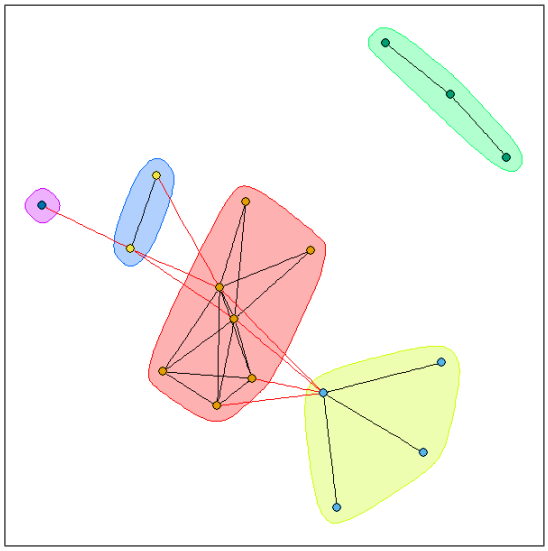

Figura 7.23: Comunidades con filtro de 30 % para el grupo A versión 2 de supermercados.