

DCNI.DMAS.063.21

Cd. México, a 27 de Septiembre de 2021.

Asunto: Proyecto terminal.

A QUIEN CORRESPONDA

Por medio de este conducto, me permito informar que la **DRA. ALICIA MONTSERRAT ALVARADO GONZÁLEZ**, académico del Departamento de Matemáticas Aplicadas y Sistemas, dirigió durante los trimestres 19P, 19O y 20O de manera satisfactoria los Proyectos Terminales I, II y III de la alumna López Díaz Sandra Lucero, con matrícula 2163071923, de la Licenciatura en Ingeniería en Computación de esta Unidad. Los proyectos antes señalados iniciaron el 25 de noviembre de 2019 y finalizaron el 08 de marzo de 2021, culminando con la entrega del trabajo escrito:

Desarrollo de un visualizador de grafos para el análisis de colaboraciones en el área de matemáticas discretas y combinatoria.

Se extiende la presente constancia para los fines que mejor convengan a los interesados.

Atentamente

“Casa abierta al tiempo”



Dr. Julián Alberto Fresán Figueroa

Jefe del Departamento de Matemáticas Aplicadas y Sistemas

Unidad Cuajimalpa

Departamento de Matemáticas Aplicadas y Sistemas

Torre III, 7º. piso. Avenida Vasco de Quiroga 4871, Colonia Santa Fe Cuajimalpa

Delegación Cuajimalpa de Morelos, México, D.F., C.P. 05348

Tel. 5814-6500 ext. 3805

www.cua.uam.mx

División de Ciencias Naturales e Ingeniería.

Unidad Cuajimalpa

Departamento de Matemáticas Aplicadas y Sistemas

Torre III, 7º. piso. Avenida Vasco de Quiroga 4871, Colonia Santa Fe Cuajimalpa

Delegación Cuajimalpa de Morelos, México, D.F., C.P. 05348

Tel. 5814-6500 ext. 3805

www.cua.uam.mx



UNIVERSIDAD AUTÓNOMA METROPOLITANA

CIENCIAS NATURALES E INGENIERÍA

DESARROLLO DE UN VISUALIZADOR DE GRAFOS
PARA EL ANÁLISIS DE COLABORACIONES EN EL
ÁREA DE MATEMÁTICAS DISCRETAS Y
COMBINATORIA

T E S I S

PARA OBTENER EL TÍTULO DE:

INGENIERA EN COMPUTACIÓN

P R E S E N T A :

SANDRA LUCERO LÓPEZ DÍAZ

TUTOR

DRA. MIKA OLSEN

DRA. ALICIA MONTSERRAT ALVARADO GONZÁLEZ

CIUDAD DE MÉXICO, CDMX, 2021



Dedicatoria

*–A mi familia por haberme apoyado incondicionalmente,
a Fer por alentarme a continuar con mis estudios
cuando parecía que me iba a rendir
y en especial a mi papá.*

Agradecimientos

A la Universidad Autónoma Metropolitana unidad Cuajimalpa por permitirme ser parte de la comunidad universitaria, aquí fue donde recibí todos los conocimientos necesarios para poder llevar acabo este proyecto.

A mi familia por haber sido mi apoyo a lo largo de toda mi carrera universitaria y a lo largo de mi vida.

A mis asesoras, la Dra. Mika Olsen y la Dra. Alicia Monserrat Alvarado, por sus enseñanzas, ideas y conocimientos compartidos.

A Monica por explicarme cosas que no entendía y siempre hacerse un espacio para poder ayudarme.

A Fernando por el apoyo que me ha brindado, estuvo a mi lado inclusive en los momentos y situaciones más difíciles.

A Daniel por todas esas enseñanzas a lo largo de mi vida y las platicas tan necesarias.

A todas las personas especiales que me acompañaron en esta etapa, aportando a mi formación tanto profesional y como ser humano.

Resumen

En este proyecto se pretende analizar y visualizar las relaciones de colaboración que existe entre la comunidad de matematicas discretas en los últimos 10 años. Se tomo ese periodo para que se tenga una representación más significativa de la red.

Se eligió un software llamado Gephi [1] que nos ayudará a realizar todo el proceso de generar el grafo que representa la red y para hacer el análisis del mismo, además, se contará con una base de datos en la cual se almacenan todos los datos necesarios para hacer este análisis. Para tener toda la información usamos el lenguaje de programación Python, el cual nos ayudará a realizar la extracción de los datos de la pagina web *American Mathematical Society*, *Mathscinet* [2], también usamos una base de datos hecha en *MongoDB*. Cabe destacar que los datos se van a almacenar en MongoDB y después serán exportados para hacer uso de una base de datos en *MySQL*. Es importante mencionar que se tuvo que hacer varios experimentos a lo largo de este proceso para tener resultados positivos.

Índice general

Agradecimientos	II
Resumen	III
1. Introducción	1
1.1. Objetivo	3
1.1.1. Objetivo general	3
1.1.2. Objetivos específicos	3
2. Marco Teórico	5
2.1. Definiciones y Medidas	5
2.2. Modelo de Redes Complejas	9
3. Metodología	11
3.1. Algoritmo de extracción de datos	12
3.1.1. Restricciones de selección de la información	13
3.2. Generación de la base de datos	13
3.3. Generación y visualización del grafo	23

<i>ÍNDICE GENERAL</i>	v
4. Experimento	24
4.1. Generación y visualización del grafo	24
4.2. Generación de consultas	27
5. Conclusiones	39
6. Trabajo Futuro	40

Índice de figuras

2.1. Camino Geodésico.	8
3.1. Diagrama UML.	12
3.2. Diagrama Entidad Relación (ER).	15
4.1. Contenido del archivo de vértices en Excel.	25
4.2. Contenido del archivo de aristas en Excel.	26
4.3. Grafo generado a partir de los archivos de vértices y aristas de Excel.	27
4.4. Ejemplo de autores con una base de datos introducidos uno a uno.	28
4.5. Relación de un autor con su correspondiente artículo.	29
4.6. Consulta relacionada a algoritmo 2.	30
4.7. Consulta relacionada a algoritmo 3.	31
4.8. Conteo total de colaboraciones entre dos vértices relacionado a algoritmo 4.	32
4.9. Resultado del Grafo de la red del experimento usando la base de datos.	34
4.10. Grafo correspondiente a las colaboraciones con peso en las aristas.	35
4.11. Grafo correspondiente a las colaboraciones entre mujeres y hombres.	36

4.12. Grafo correspondiente a las colaboraciones entre mujeres.	37
4.13. Grafo correspondiente a las colaboraciones entre hombres.	37

Índice de tablas

3.1. Tabla Autor.	16
3.2. Tabla Adscripción.	16
3.3. Tabla Institutos.	17
3.4. Tabla Estado.	17
3.5. Tabla País.	18
3.6. Tabla Artículo.	19
3.7. Tabla Revista.	19
3.8. Tabla RevistaIndicesQ.	20
3.9. Tabla Categoría.	20
3.10. Tabla RevistaIndices.	21
3.11. Tabla ArtículoKeywords.	21
3.12. Tabla ArtículoKeywords.	22

Capítulo 1

Introducción

La comunidad de matemáticas discretas y combinatoria, esta conformada por, al menos, una tercera parte del genero femenino y dos terceras partes pertenecen al genero masculino. Se puede notar que dicha comunidad tiene más hombres que mujeres. A lo largo del tiempo nos hemos expuesto a casos de discriminación, ya que muchas de las veces el genero masculino se siente superior al genero femenino. En algunos casos prefieren darles puestos de trabajo a los hombres por prejuicios o porque asumen que los hombres son más capaces que las mujeres.

Es importante mencionar que este proyecto es una extensión a la tesis de *Narda Cordero Michel* con título *Seis años de colaboración en el coloquio de gráficas*, en ese proyecto se estudia la red de colaboración del Coloquio Víctor Neumann-Lara de teoría de las gráficas, combinatoria y sus aplicaciones, que se llevaron a cabo del año 2013 al 2018.

Es por ello que se desarrollará una herramienta de software para visualizar grafos que nos ayuden a identificar la tendencia de colaboraciones entre investigadores e

investigadoras en la comunidad antes mencionada, y a contestar preguntas específicas como: las colaboraciones que se han hecho entre las mujeres, entre hombres y entre ambos géneros; la dinámica de las colaboraciones entre los que trabajan con gente de la misma institución y la gente de otras instituciones; y los temas que se trabajan en cada comunidad.

Este análisis sólo se llevará a cabo con la comunidad de investigadores en matemáticas discretas y combinatoria de México. Para que el grafo tenga una representación más significativa, se contemplará el periodo de los últimos 10 años y esta información será extraída de la página web *American Mathematical Society, Mathscinet* [2].

Cabe mencionar que a lo largo de esta investigación se encontró con otros proyectos similares a lo que se está haciendo: *VOSviewer* y *Connected Papers*.

VOSviewer es una herramienta de software que nos permite construir y visualizar grafos [3]. Estos pueden ser, por ejemplo, revistas, investigadores o publicaciones individuales, además, se pueden construir sobre la base de citas, acoplamiento bibliográfico, co-cita o relaciones de coautoría. Este software nos permite construir grafos desde cero y a partir de los datos antes mencionados, es importante mencionar que para hacer el grafo utiliza Web of Science, Scopus, Dimensions y PubMed. Tiene tres tipos de visualizaciones, uno de ellos es el *zoom y desplazamiento*: los grafos se pueden explorar a detalle utilizando esta función, además, contiene un algoritmo en el que evita que las etiquetas se encimen. El segundo es *visualizaciones de densidad y superposición*: este tipo nos proporciona una descripción general de las principales áreas del grafo. El tercero nos permite hacer *capturas* de alta calidad y estas se pueden

guardar en diferentes formatos de archivos gráficos.

Connected Papers es una herramienta visual que ayuda a los investigadores a encontrar documentos relevantes dependiendo su campo de trabajo [4]. Este selecciona las conexiones más fuertes con el documento de origen, se basa en co-cita y acoplamiento bibliográfico. Se dice que si dos artículos tienen citas y referencias muy similares, estos tienen una mayor probabilidad de que se trate un tema relacionado. Además, construye un grafo dirigido a la fuerza con el fin de distribuir los artículos, agrupando los similares y separando los menos similares. La base de datos que se utiliza en este software está conectada al *Semantic Scholar Paper Corpus* (con licencia de ODC-BY).

1.1. Objetivo

1.1.1. Objetivo general

Desarrollar un visualizador de grafos para analizar las colaboraciones del área de matemáticas discretas y combinatoria de México, generando una red de colaboraciones asociada a las publicaciones del área que aparecen en .

1.1.2. Objetivos específicos

Es importante notar que este proyecto es un complemento del proyecto de servicio social desarrollado por Esteban Díaz Pérez (aún no reportado), el cuál consistió en un software de extracción de datos de una página web (comúnmente conocido como

Scraper) y la generación de una base de datos para almacenar la información extraída de la página web bajo análisis. En aras de la integridad de este documento, se explicarán ambos desarrollos.

Además, para llevar acabo este proyecto se necesitaron varios conocimientos previos y una serie de pruebas.

- Introducción a los grafos.
- Estudio del Software *Gephi*.
- Implementación de un ejemplo con archivos de Excel para ver el funcionamiento de *Gephi*.
- Implementación de un ejemplo con la base de datos que será utilizada y exportada a *Gephi*, se realizaran consultas para la visualización del grafo resultante.

Capítulo 2

Marco Teórico

2.1. Definiciones y Medidas

En esta sección se utilizaron las definiciones del libro de *Mark. E. J. Newman* [5].

Una red guarda la información de lo que representa, es decir, nos muestra que representa cada vértice y la relación que existe entre ellos, mientras un grafo consta sólo de vértices y aristas, es decir, aunque sea la representación de una red esta no guarda la información de que representa cada vértice. Es más una gráfica no tiene porque representar una red.

Un *grafo* es denotado por $G = (V, A)$, consiste en dos conjuntos V y A , el primero es conocido como *vértices* y el segundo como *aristas*, los últimos representan la relación entre los vértices. Se dice que el *grafo* es *dirigido* cuando se hace una distinción de orden dando una dirección a la arista, cuando no se menciona la palabra dirigida, se dice que es simplemente un *grafo*.

Un *subgrafo* de un grafo es en sí mismo un grafo y está contenido en el grafo original, su conjunto de vértices es un subconjunto de vértices del grafo original, de la misma manera pasa con las aristas. Si un subgrafo contiene a todos los vértices del grafo, diremos que es un subgrafo generador.

Una *ruta* es una sucesión de vértices en la que cualesquiera dos vértices consecutivos son adyacentes, es decir, es un recorrido que nos permite llegar de un vértice a otro a través de las aristas del grafo, pasando por ciertos vértices intermediarios. La *longitud* de una ruta es la cantidad de aristas que atraviesa, es decir, la cantidad de pasos que da, si se pasa por una arista más de una vez esta se contara tantas veces como es atravesada por dicha ruta. Un *ciclo* es una sucesión de aristas adyacentes, donde no se recorre dos veces el mismo vértice y donde se regresa al punto inicial.

Un grafo es *conexo* si cada par de vértices está conectado por un camino, es decir, si para cualquier par de vértices (a, b) , existe al menos un camino posible desde a hacia b . Un grafo es *disconexo* si dos o más de sus vértices no están conectados por un camino. Una *componente* de un grafo es un subgrafo inducido en el que dos vértices cualesquiera están conectados entre sí por caminos y no están conectados a vértices adicionales en el resto del grafo.

Se le llama un *conjunto de corte* a un subconjunto de vértices que al ser removidos del grafo se obtiene un nuevo grafo que tiene más componentes que el original. Se llama *vértice de corte* a un vértice que, al ser removido de un grafo, produce un grafo con un número mayor de componentes conexas y se llama puente a una arista que al ser removida de una gráfica produce una nueva gráfica con exactamente una

componente conexa más.

La *centralidad* [6] en un grafo puede ser entendida como una medida que posee un vértice, una arista, un conjunto de vértices ó aristas dentro de un grafo, ayuda a determinar el impacto que causa dentro del grafo o de la red del que forma parte. La *centralidad de grado* mide el número de aristas que inciden en un vértice ó bien al que pertenece un vértice. Al vértice que tiene mayor cantidad de aristas se le conoce como "*hub*". Los *hubs* son vértices cuyo grado es extremadamente grande. La *centralidad de vector propio* mide la influencia directa de un vértice dentro de un grafo, los vértices que poseen un valor alto son aquellos vértices que tienen muchas aristas, por ejemplo, si tenemos a un autor relacionado con muchos autores, diremos que ese autor tiene una influencia muy alta. La *centralidad PageRank* [7] pone en consideración la calidad de las relaciones, está definida para grafos dirigidos pero se usó una adaptación para el caso no dirigido, ya que el grafo que se estudia es no dirigido. La manera de adaptar la definición de dicha centralidad para el caso no dirigido, es considerar que una arista entre dos vértices es equivalente a tener el par de arcos dirigidos entre esos vértices en las direcciones posibles. Esta medida fue introducida por Google en el año 1999 para determinar en una búsqueda de internet las páginas más importantes de acuerdo al tema y mostrar los resultados de una manera jerárquica.

Un *camino geodésico* es la longitud del camino mas corto, por ejemplo, en el grafo de la figura 2.1 se puede ver que la distancia entre el vértice 1 y el vértice 6 es 2, ya que los vértices 1 y 6 no son adyacentes y hay un camino $(1, 2, 6)$ de longitud 2 entre los vértices 2 y 6, también se puede considerar el camino $(1, 3, 6)$. Observe que el

camino $(1, 2, 4, 5, 6)$ es un camino entre 1 y 6 , pero no es un camino geodésico ya que tiene longitud 4 .

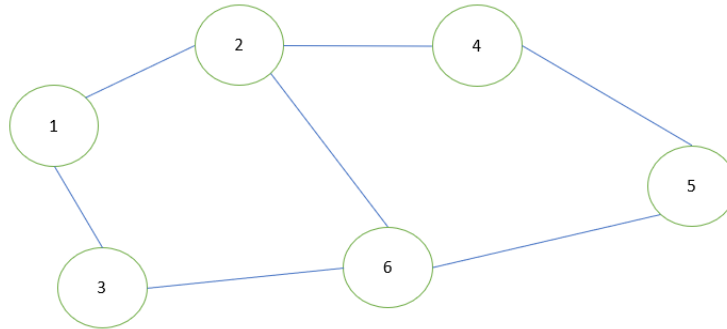


Figura 2.1: Camino Geodésico.

La *centralidad intermedia* de un vértice es el número de caminos geodésicos distintos que pasan a través de él, donde dos o más de ellas pueden tener vértices en común.

Se le llama *clique o clan* a un subconjunto de vértices tal que para cada pareja de vértices ahí es adyacente y es máximo con esta propiedad, es decir, si se agrega otro vértice al conjunto habrá una pareja de vértices no adyacentes.

En un grafo se puede identificar un tipo de subestructuras llamadas *comunidades*, estos son subgrafos en los que los vértices están muy conectados entre sí y tienen pocas aristas con otros vértices fuera del subgrafo. La *modularidad* es una medida de la porción con la que dos vértices similares se conectan.

2.2. Modelo de Redes Complejas

En esta sección se presentan tres modelos de redes que se han desarrollado de manera teórica, fueron consultados en [8], con el fin de identificar a cuál de ellos se asemeja el grafo.

1. Modelo de Redes Aleatorias (ER)

Un *grafo aleatorio* es un modelo de red en el cual algunos parámetros quedan fijos de entrada mientras que otros son aleatorios. El inicio de la teoría sobre las redes aleatorias se debe a los matemáticos *Erdős* y *Rényi*, siendo conocidas como modelos ER, esto se debe a las iniciales de sus autores. En este tipo de modelo van aumentando el número de aristas entre un número de vértices fijo y en los que no se conoce de qué manera se generan las nuevas aristas, se plantean desde el principio de que un par de vértices se puede conectar de manera aleatoria con probabilidad p .

2. Modelo de Redes de Mundo Pequeño (WS)

A diferencia del modelo ER, que dos vértices tengan un vecino común será significativo para determinar si están o no conectados. Esto significa que, si dos personas están conectadas a una tercera, la posible relación entre ellas es completamente independiente de la que tienen con su vecino. Este modelo revela que la sociedad es una red social con forma de mundo pequeño, están interconectados entre sí con una estructura de red, los vértices son las personas y las aristas, las relaciones entre ellos. La hipótesis que se plantea es que el grafo de

las colaboraciones entre investigadores e investigadoras de la comunidad de matemáticas discretas y combinatoria, es un modelo de mundo pequeño. Se hace esta suposición puesto que, en las colaboraciones muchas veces sucede que se trabaja con los colegas de nuestros colegas. Es por ello que el comportamiento de las aristas se asemeja a una estructura circular con triángulos y las aristas que cruzan de un lado a otro representan las relaciones imprevisibles.

3. Modelo de Libre Escala (BA)

Este modelo tiene la particularidad de que las aristas están distribuidos de forma muy dispareja. Además, están compuestos por pocos vértices altamente conectados y muchos vértices tienen muy pocas aristas. El modelo mejor conocido de este tipo es el de *enlace preferencial* se debe a Albert-László Barabási y Réka Albert [9]. En este modelo el grafo crece, se agregan vértices uno a uno y cada vértice se conecta con un conjunto de vértices ya existentes, elegidos con ciertas condiciones. Las conexiones serán no dirigidas y el número de aristas con los que se conecta cada vértice nuevo es c . La manera de elegir a sus vecinos es al azar, donde cada vértice existente es elegido con una probabilidad proporcional al grado que tiene hasta ese momento, lo cual propicia la aparición de hubs. Es importante mencionar que se agregan vértices y también las aristas, pero nunca se quitan.

Capítulo 3

Metodología

En esta sección explicaremos la metodología que se siguió para desarrollar una herramienta de software para visualizar grafos que nos ayuden a identificar la tendencia de colaboraciones entre la comunidad de matemáticas discretas, y a contestar preguntas específicas como es la estructura de la red de las colaboraciones que se han hecho sólo entre las mujeres, entre mujeres y hombres y sólo entre hombres.

El esquema general es el siguiente, y se resume en el diagrama 3.1:

1. Algoritmo de extracción de datos para recopilar la información de la página web bajo análisis.
2. Generación de la base de datos para almacenar la información extraída.
3. Generación de consultas sobre la base de datos que permitan contestar a las preguntas específicas ya mencionadas.
4. Visualización de los grafos.

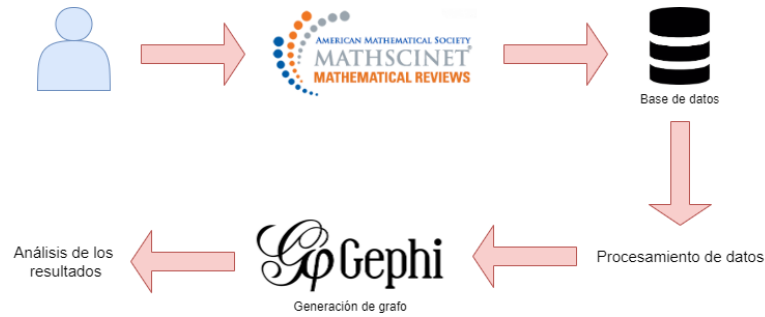


Figura 3.1: Diagrama UML.

3.1. Algoritmo de extracción de datos

En la extracción de datos basó en el desarrollo de un *Scraper*, el cual es un método que utiliza unas líneas de código, normalmente en Python, con el fin de conseguir información de páginas web. Estos programas normalmente imitan la forma que tiene un usuario de navegar en la red y recopilan la información indicada en el algoritmo desarrollado.

Para tomar dicha información se requirió de una página web llamada *American Mathematical Society, Mathscinet* [2] y para guardar toda la información recopilada se hizo uso de una base de datos en MongoDB y MySQL. Además, para la extracción de datos se usó el lenguaje de programación Python y un software llamado Gephi [1], el cual nos ayudará a analizar nuestra red de colaboraciones.

Como ya se ha mencionado, se utilizó el lenguaje de programación *Python*, con el objetivo de realizar la extracción de datos de la página web. Este nos sirve para sacar de alguna manera más rápida la información y almacenarla, evitando un esfuerzo

mayor, es decir, se evita el introducir dato por dato a la base de datos.

El algoritmo propuesto es el siguiente:

Algorithm 1 Extracción

Inicio

- 1: Ingresar a la página web de Mathscinet
- 2: Introducir usuario y contraseña
- 3: Conectar con la base de datos
- 4: Ingresar un id del autor

Fin

3.1.1. Restricciones de selección de la información

Los grafos que serán mostrados representarán sólo los artículos que tienen colaboraciones con otros autores, es decir, no se mostrarán los artículos que tengan un solo autor. Para que se tenga una representación más significativa de los autores que tienen distintas colaboraciones en diferentes artículos, se decidió tomar el periodo de los últimos 10 años. Para construir el grafo, se comenzó formando un grafo en el que se pone un vértice por cada autor. Cabe mencionar que se tomaron todos los datos de la pagina web antes mencionada, pero con el algoritmo de extracción de datos sólo se seleccionaran los autores de México.

3.2. Generación de la base de datos

Para construir la base de datos, se necesita de un diagrama *entidad-relación*. El diagrama entidad-relación es una herramienta para el modelado de datos que permite representar las entidades relevantes de un sistema de información, así como

sus interrelaciones y propiedades [10].

En la figura 3.2 se muestra el diagrama entidad-relación.

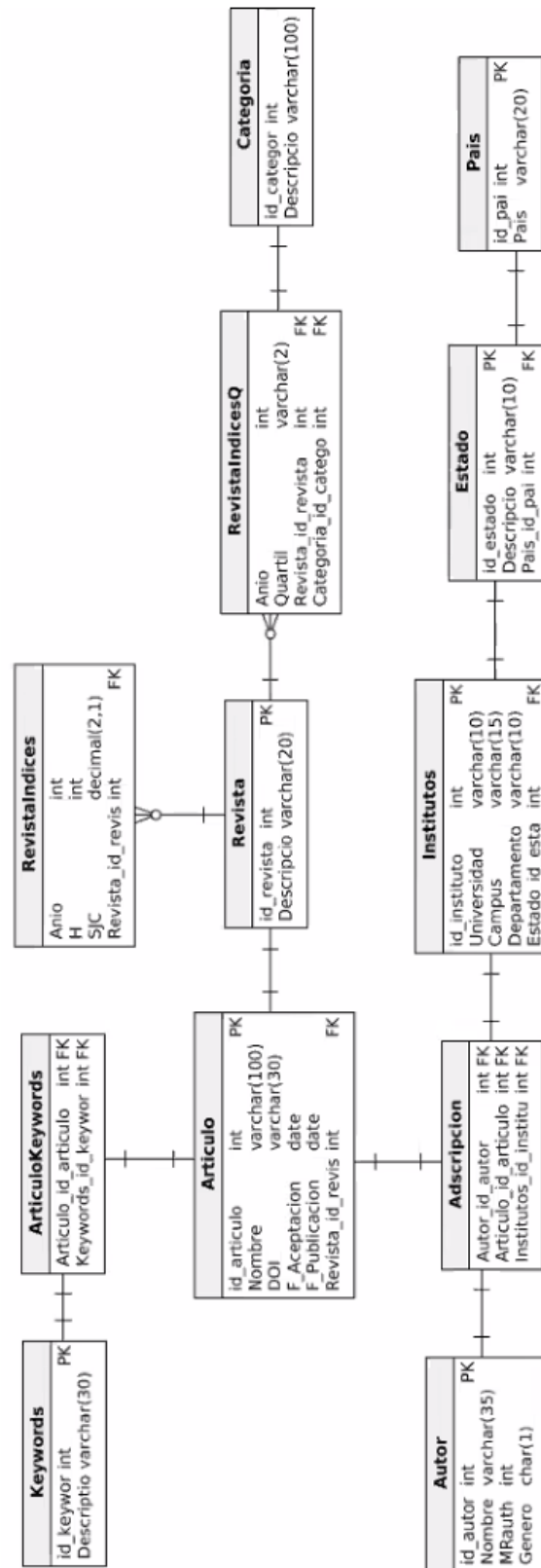


Figura 3.2: Diagrama Entidad Relación (ER).

A continuación se describirán las tablas de la figura 3.2.

1. Tabla Autor

Esta tabla contendrá información de cada uno de los autores.

Campo	Tipo	Descripción
id_autor	int(11)	Corresponde a un único ID por autor.
Nombre	varchar(35)	Corresponde al nombre del autor.
MRauthor	int(11)	Corresponde a un único número por cada autor.
Genero	char(1)	Corresponde al genero de cada autor.

Tabla 3.1: Tabla Autor.

2. Tabla Adscripción

Esta tabla será la encargada de relacionar un Autor, con un Artículo y su Institución.

Campo	Tipo	Descripción
Autor_id_autor	int(11)	Corresponde al ID del autor de la tabla Autor.
Articulo_id_articulo	int(11)	Corresponde al ID del artículo con el que trabajo un autor.
Institutos_id_instituto	int(11)	Corresponde al ID del instituto del autor.

Tabla 3.2: Tabla Adscripción.

3. Tabla Institutos

Esta tabla contiene información del Instituto de cada autor.

Campo	Tipo	Descripción
id_instituto	int(11)	Corresponde al ID de un instituto, este dependerá de cada autor.
Universidad	varchar(10)	Corresponde al nombre de la universidad de cada autor.
Campus	varchar(15)	Corresponde al campus en el que estudio cada autor.
Departamento	varchar(10)	Corresponde al departamento que pertenece cada autor.
Estado_id_estado	int(11)	Corresponde al estado en el que esta el instituto de cada autor.

Tabla 3.3: Tabla Institutos.

4. Tabla Estado

Esta tabla tiene información de la ubicación del instituto de cada autor.

Campo	Tipo	Descripción
id_estado	int(11)	Corresponde al ID del estado donde se encuentra el instituto de cada autor.
Descripcion	varchar(10)	Corresponde al nombre del estado.
Pais_id_pais	int(11)	Corresponde al ID del país donde se ubica el instituto.

Tabla 3.4: Tabla Estado.

5. Tabla País

Esta tabla tiene información del país donde se ubica el instituto de cada autor.

Campo	Tipo	Descripción
id_pais	int(11)	Corresponde al ID del país donde se encuentra un instituto de cada autor.
Pais	varchar(20)	Corresponde al nombre del país.

Tabla 3.5: Tabla País.

6. Tabla Artículo

Esta tabla contiene información de cada artículo.

Campo	Tipo	Descripción
id_articulo	int(11)	Corresponde al ID único de cada artículo.
Nombre	varchar(100)	Corresponde al nombre de cada artículo.
DOI	varchar(30)	Corresponde a un identificador único y permanente.
F_Aceptacion	date	Corresponde a la fecha en el que fue aceptado el artículo.
F_Publicacion	date	Corresponde a la fecha en la que fue publicado el artículo.
Revista_id_revista	int(11)	Corresponde a un ID para la revista en la que fue publicado el artículo.

Tabla 3.6: Tabla Articulo.

7. Tabla Revista

Esta tabla contiene información de cada revista.

Campo	Tipo	Descripción
id_revista	int(11)	Corresponde al ID único de cada revista.
Descripcion	varchar(20)	Corresponde al nombre de la revista.

Tabla 3.7: Tabla Revista.

8. Tabla RevistaIndicesQ

Esta tabla contendrá información de cada uno de las revistas.

Campo	Tipo	Descripción
Anio	int(11)	Corresponde al año de la revista.
Quartil	varchar(2)	Corresponde a un identificador para cada revista, en matematicas existen tres tipos de cuartiles los cuales son: $Q1$, $Q2$ y $Q3$.
Revista_id_revista	int(11)	Corresponde a un único número por cada revista.
Categoria_id_categoria	int(11)	Corresponde a un número único de categoría a la que pertenece la revista.

Tabla 3.8: Tabla RevistaIndicesQ.

9. Tabla Categoría

Esta tabla contiene información de cada categoría.

Campo	Tipo	Descripción
id_categoria	int(11)	Corresponde al ID único de cada categoría para una revista.
Descripcion	varchar(100)	Corresponde al nombre de la categoría para una revista.

Tabla 3.9: Tabla Categoria.

10. Tabla RevistaIndices

Esta tabla contendrá información de cada uno de las revistas.

Campo	Tipo	Descripción
Anio	int(11)	Corresponde al año de la revista.
H	int(11)	Corresponde a un único índice
SJC	decimal(2,1)	Corresponde a un indicador de prestigio independiente del tamaño que clasifica las revistas. según su "prestigio medio por artículo"
Revista_id_revista	int(11)	Corresponde a un único número por cada revista.

Tabla 3.10: Tabla RevistaIndices.

11. Tabla Keywords

Esta tabla contiene números únicos para poder identificar de una manera más rápida los temas que se trabajan en cada comunidad.

Campo	Tipo	Descripción
id_keyword	int(11)	Corresponde al número único de un keyword.
Description	int(30)	Corresponde al nombre del keyword.

Tabla 3.11: Tabla ArtículoKeywords.

12. Tabla ArtículoKeywords

Esta tabla relaciona la información de la tabla *Keywords* y *Articulo*.

Campo	Tipo	Descripción
Articulo_id_articulo	int(11)	Corresponde al ID del artículo.
Keywords_id_keyword	int(11)	Corresponde a un número único que nos ayudará a saber que tipo de tema se trabajan en cada comunidad.

Tabla 3.12: Tabla ArticuloKeywords.

Es importante mencionar que en la figura 3.2 se utilizan las claves primarias y foráneas. Una *clave primaria* es un campo único que no se puede repetir, solo existe una clave primaria por tabla. Y una *clave foránea* es uno campo de un tabla que hace referencia al campo de la clave primaria de otra tabla, es decir, indica como están relacionadas las tablas. Los datos en los campos de ambas deben coincidir, aunque los nombres no sean los mismos [11]. Al tener estos conceptos se puede entender las relaciones de las tablas.

La tabla *Autor* y *Adscripcion* tienen la *relación uno a uno*, es decir, un registro de una tabla se asocia a *uno y solo un* registro de otra tabla. Por ejemplo, la tabla *Autor* tiene como clave primaria al *id_autor* y esa clave será utilizada en la tabla *Adscripcion* como foránea ya que esa tabla será la encargada de relacionar el *Autor_id_autor*, *Articulo_id_articulo* de la tabla *Articulo* y el *Instituto_id_instituto* de la tabla *Institutos*, como se menciono anteriormente las claves foráneas deben coincidir con la primaria aunque no tengan los mismos nombres.

Otro ejemplo sería en la tabla *Revista*, tiene una relación *uno a muchos* con la

tabla *RevistaIndicesQ*, es decir, un registro de una tabla se puede asociar a uno o varios registros de otra tabla, para este caso un registro de la tabla *Revista* se puede asociar a uno o varios registros de la tabla *RevistaIndicesQ*.

Existen diferentes tipos de restricciones, en nuestra base de datos se utiliza *not null* para evitar insertar valores nulos en la columna especificada, considerándolo entonces como un valor no aceptado para esa columna. Esto significa que se debe proporcionar un valor válido.

3.3. Generación y visualización del grafo

Anteriormente se menciono sobre el software que se utilizará durante todo este proceso, *Gephi* es una software que se usa para hacer análisis y visualización de grafos [1]. Permite importar un archivo de datos e incluso una base de datos. Además, el usuario tiene la oportunidad de interactuar con la representación del grafo, puede ponerle diferentes colores, interactuar con la representación, agregar diferentes filtros, cambiar el tamaño de la letra, entre otros [12]. Su objetivo es permitirle al usuario hacer hipótesis y descubrir patrones, además, es una herramienta de gran utilidad ya que con interfaces interactivas facilita el razonamiento. Se considera un software para el Análisis Exploratorio de Datos.

Capítulo 4

Experimento

El objetivo de los experimentos es, por un lado, identificar la mejor manera de exportar los datos extraídos de la página web bajo análisis a *Gephi*, y por otro lado, generar las consultas para responder algunas preguntas específicas.

4.1. Generación y visualización del grafo

Para hacer el grafo con archivos de Excel, se requiere un archivo para almacenar los vértices y otro para almacenar las aristas. El archivo de *vértices* incluye dos columnas con etiquetas *id* y *label*, respectivamente. En la figura 4.1 se muestra la estructura de dicho archivo.

	A	B
1	id	label
2		1 monica
3		2 sandra
4		3 adela
5		4 cesar
6		5 daniel
7		6 fernando
8		7 arely
9		8 alonso

Figura 4.1: Contenido del archivo de vértices en Excel.

Por otro lado, el archivo de *aristas* incluye dos columnas cuyas etiquetas son *source* y *target*, que indican la relación entre dos vértices (véase la figura 4.2).

	A	B
1	source	target
2	1	3
3	1	4
4	2	3
5	2	1
6	3	5
7	4	5
8	4	6
9	5	2
10	6	7
11	6	8
12	7	1
13	7	5
14	8	2

Figura 4.2: Contenido del archivo de aristas en Excel.

Ambos archivos se importaron a *Gephi* para comprobar que se podía construir el grafo. En la figura 4.2 se muestra el grafo generado.

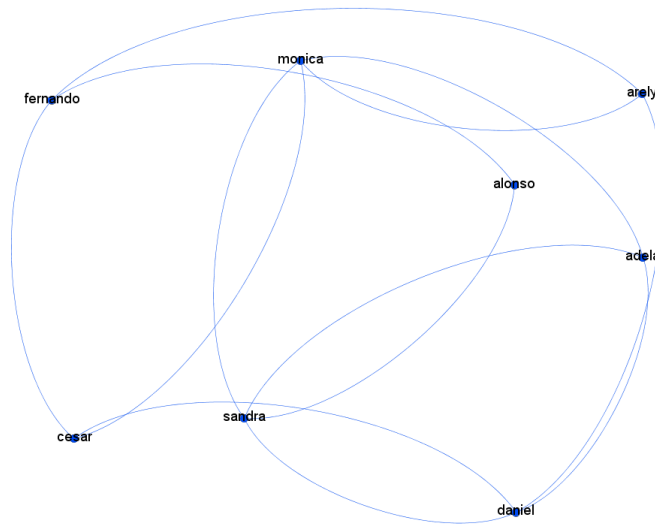


Figura 4.3: Grafo generado a partir de los archivos de vértices y aristas de Excel.

Se consideraba introducir todos los datos de la página web a la base de datos, después exportarlos a los archivos de Excel correspondientes y mediante esos archivos hacer el grafo. Pero, con el paso del tiempo y el estudio de Gephi, nos dimos cuenta de que también podíamos importar la base de datos directamente de *MySQL*, así que no fue necesario generar los archivos en Excel.

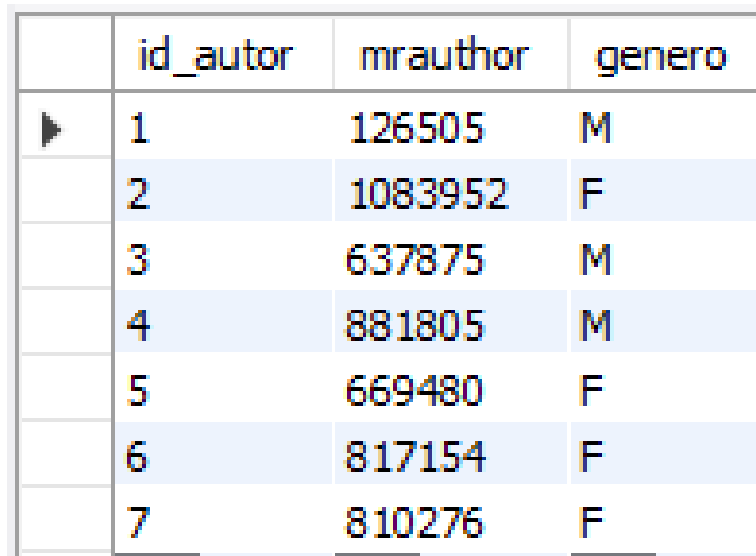
4.2. Generación de consultas

Anteriormente se mencionó cual era el objetivo de este proyecto, se pretende responder a tres preguntas esenciales. Una de ellas es visualizar las colaboraciones entre

mujeres, entre hombres y de forma global (mixto). La segunda es ver quien tiene más colaboraciones, ya sea los que trabajan con gente de la misma institución o con gente de otras. Y la tercera es visualizar que temas se trabajan en cada comunidad.

Al momento de importar la base de datos a Gephi, nos encontramos con otros errores, para empezar se tenía que tener una tabla con la relación de autores que trabajan con un mismo artículo.

Se utilizará un ejemplo para que el proceso sea más claro, usaremos las herramientas *Workbench* [13] y *Gephi*, es importante mencionar que los *datos almacenados* que aparecen en este ejemplo fueron introducidos uno a uno mediante código *MySQL*. La figura 4.4 contiene una lista de autores con su respectivo “ID”.

A screenshot of a database table with four columns: 'id_autor', 'mrauthor', and 'genero'. The table contains seven rows of data. The first row is highlighted with a blue background. The second row is highlighted with a light blue background. The third row is highlighted with a light blue background. The fourth row is highlighted with a light blue background. The fifth row is highlighted with a light blue background. The sixth row is highlighted with a light blue background. The seventh row is highlighted with a light blue background. The table is displayed in a window with a title bar and a scroll bar.

	id_autor	mrauthor	genero
▶	1	126505	M
	2	1083952	F
	3	637875	M
	4	881805	M
	5	669480	F
	6	817154	F
	7	810276	F

Figura 4.4: Ejemplo de autores con una base de datos introducidos uno a uno.

En la figura 4.4 tenemos toda la información de cada autor, además, necesitamos ver qué artículo esta relacionado con cada autor, la figura 4.5 contiene la información

antes mencionada.

	Autor_id_autor	Articulo_id_articulo	Institutos_id_instituto
▶	1	4089442	1
	2	4089442	2
	3	3432436	3
	2	3432436	2
	4	3855020	4
	5	3855020	5
	1	3855020	1
	6	3513769	6
	5	3513769	5
	6	4012858	6
	5	4012858	1
	5	4012858	1
	7	3962015	1
	6	3962015	6

Figura 4.5: Relación de un autor con su correspondiente artículo.

Para tener el resultado de la relación de autores que trabajan con un mismo artículo, se tuvieron que construir dos tablas temporales (véase en algoritmo 2 y 3). La primera tabla se construyó con una condición *WHERE* en donde se condicionó a que los autores fueran *diferentes* y que tuvieran un *mismo articulo*. Se consideró v que la relación de autores fuera *diferente* ya que si se ponía que fuera *igual* no nos salía las relaciones correctas, los resultados podemos verlos en la figura 4.6.

	autor_id_autor	Articulo_id_articulo	id_autor
▶	3	3432436	1
	2	3432436	1
	6	3513769	1
	5	3513769	1
	4	3855020	1
	5	3855020	1
	7	3962015	1
	6	3962015	1
	6	4012858	1
	5	4012858	1
	5	4012858	1
	2	4089442	1
	3	3432436	2
	6	3513769	2
	5	3513769	2
	4	3855020	2
	5	3855020	2
	1	3855020	2
	7	3962015	2
	6	3962015	2
	6	4012858	2
	5	4012858	2
	5	4012858	2
	1	4089442	2
	2	3432436	3
	6	3513769	3
	5	3513769	3
	4	3855020	3

	autor_id_autor	Articulo_id_articulo	id_autor
	5	3855020	3
	1	3855020	3
	7	3962015	3
	6	3962015	3
	6	4012858	3
	5	4012858	3
	5	4012858	3
	1	4089442	3
	2	4089442	3
	3	3432436	4
	2	3432436	4
	6	3513769	4
	5	3513769	4
	5	3855020	4
	1	3855020	4
	7	3962015	4
	6	3962015	4
	6	4012858	4
	5	4012858	4
	5	4012858	4
	1	4089442	4
	2	4089442	4
	3	3432436	5
	2	3432436	5
	6	3513769	5
	4	3855020	5
	1	3855020	5

	autor_id_autor	Articulo_id_articulo	id_autor
	7	3962015	5
	6	3962015	5
	6	4012858	5
	1	4089442	5
	2	4089442	5
	3	3432436	6
	2	3432436	6
	5	3513769	6
	4	3855020	6
	5	3855020	6
	1	3855020	6
	7	3962015	6
	5	4012858	6
	5	4012858	6
	1	4089442	6
	2	4089442	6
	3	3432436	7
	2	3432436	7
	6	3513769	7
	5	3513769	7
	4	3855020	7
	5	3855020	7
	1	3855020	7
	6	3962015	7
	6	4012858	7
	5	4012858	7
	5	4012858	7
	1	4089442	7
	2	4089442	7

Figura 4.6: Consulta relacionada a algoritmo 2.

Así que, la tabla de la figura 4.6 se construyó de la siguiente manera:

Sobre la base de datos que permitan contestar a las preguntas específicas ya mencionadas.

Algorithm 2 Consulta utilizando la condición diferente

```
CREATE TABLE temporal SELECT ad.autor_id_autor, ad.Articulo_id_articulo,
au.id_autor, ar.nombre FROM adscripcion AS ad, autor AS au,
articulo AS ar WHERE ad.Autor_id_autor!=au.id_autor AND
ar.id_articulo=ad.Articulo_id_articulo;
```

En la segunda tabla se utilizaron los datos que tiene la primera, la condición del *WHERE* fue diferente en este caso ya se utilizó que los autores fueran iguales y también los artículos, estos resultados podemos verlos en la figura 4.7, además, podemos notar que las relaciones fueron menos comparando la figura 4.6.

	Autor_id_autor	id_autor	Articulo_id_articulo
►	4	1	3855020
	5	1	3855020
	2	1	4089442
	3	2	3432436
	1	2	4089442
	2	3	3432436
	5	4	3855020
	1	4	3855020
	6	5	3513769
	4	5	3855020
	1	5	3855020
	6	5	4012858
	5	6	3513769
	7	6	3962015
	5	6	4012858
	6	7	3962015

Figura 4.7: Consulta relacionada a algoritmo 3.

La tabla de la figura 4.7 fue construida de la siguiente manera:

Algorithm 3 Consulta utilizando la condición igual

```
CREATE TABLE temporal_gephi SELECT t.Autor_id_autor, t.id_autor,
t.Articulo_id_articulo FROM temporal AS t, adscripcion AS d WHERE
t.id_autor=d.autor_id_autor AND t.articulo_id_articulo=d.articulo_id_articulo;
```

El siguiente paso es identificar las *colaboraciones* que existen entre los autores. Para ello, se creó una nueva tabla temporal que permitirá contar las veces que los autores participaron en un mismo artículo. Cabe destacar que la columna `autor_id_autor` está ordenada de forma ascendente, es decir, va con una numeración del menor al mayor. Es por ello que se utilizó una condición *WHERE* en donde se condicionará a que muestre un `autor_id_autor` *menor* al `id_autor`, es decir, en la tabla creada

en la figura 4.7 existía la relación, por ejemplo, *autor 6* con *autor 5* en un artículo *4012858* y viceversa, en esta tabla solo nos aparecerá la relación *autor 5* con *autor 6* y su respectivo *peso* para la arista, es decir, la suma total de las colaboraciones que hicieron esos *dos autores*. Los resultados podemos verlos en la figura 4.8.

	autor_id_autor	id_autor	conteo
▶	1	2	1
	1	4	1
	1	5	1
	2	3	1
	4	5	1
	5	6	2
	6	7	1

Figura 4.8: Conteo total de colaboraciones entre dos vértices relacionado a algoritmo 4.

La tabla de la figura 4.8 se construyó de la siguiente manera:

Algorithm 4 Consulta para contar las colaboraciones

```
CREATE TABLE conteo SELECT autor_id_autor, id_autor,
COUNT(articulo_id_articulo) AS conteo , articulo_id_articulo FROM tem-
poral_gephi WHERE autor_id_autor<id_autor GROUP BY autor_id_autor,
id_autor;
```

Si queremos verificar que este resultado de colaboraciones entre esos dos autores es correcto, debemos regresar a la figura 4.7 y buscar en las dos primeras columnas los autores con “ID” 5 y 6, se observa que aparece dos veces. La primera relación es 6 y 5 con un artículo *4012858* y esos autores también trabajan con el artículo *3513769*,

es por ello que en nuestro conteo de colaboraciones en los autores antes mencionados aparece un 2.

Para poder ver el grafo en *Gephi*, se requiere que se le ingresen dos consultas. Una de ellas fue la de los *vértices* en la cual lleva datos importantes como es el *id del autor*, *nombre*, *género* e *instituto*, cabe destacar que *Gephi* reconoce el nombre definido de sus columnas, por ejemplo, si al “*ID*” de nuestro autor le poníamos una etiqueta como “*Código*”, al momento de generar el grafo no nos mantendrá el número asignado para cada autor, es decir, si nosotros poníamos en la base de datos al autor *Juan* como “*código 1*”, *Gephi* no respeta ese número ya que esa columna no tiene el mismo nombre, al cambiarle el nombre de la columna por “*ID*”, *Gephi* ya respetaba dichos datos. Esto quiere decir que la consulta quedo de la manera:

```
CREATE VIEW nodos AS SELECT a.id_autor AS id, a.nombre AS label,
a.MRauthor, a.genero, ad.institutos_id_instituto FROM Autor AS a, Adscrip-
cion AS ad, Institutos AS i WHERE (a.id_autor=ad.autor_id_autor) AND
(ad.institutos_id_instituto=i.id_instituto) GROUP BY a.id_autor;
```

Se utiliza la palabra reservada **as** en *Workbench* para poder renombrar una etiqueta, esto se tenía que hacer debido a que *Gephi* reconoce sus propias etiquetas y no hubiera problema al hacer el grafo. En la consulta anterior ya se tiene que debe mostrar cada *vértice*, ahora se tiene que hacer la segunda consulta que son las *aristas*. Para las aristas se debe mostrar la relación de los autores, es decir, se pondrá una arista por cada par de vértices que trabajaron juntos en diferentes artículos. Además, se mostrará el *peso* y un *keyword*, este nos ayudará al análisis del grafo ya que se quiere revisar que tipo de temas trabajan en cada comunidad. Por lo tanto la consulta quedo de la siguiente manera:

```
CREATE VIEW aristas AS SELECT r.autor_id_autor AS source, r.id_autor AS
target, r.conteo AS weight, r.keywords_id_keyword FROM restriccion AS r;
```

Al tener nuestras dos consultas listas, se importan a *Gephi* y el grafo que resulta es el siguiente figura 4.9

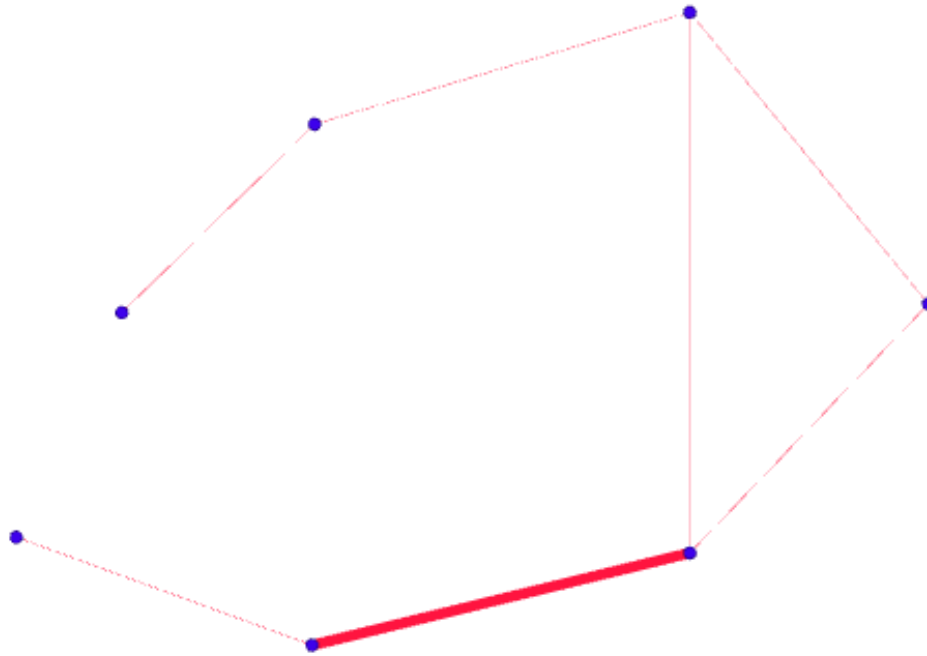


Figura 4.9: Resultado del Grafo de la red del experimento usando la base de datos.

Se puede observar que en el grafo anterior los *vértices* son de *color azul* y las *aristas* de *color rojo*. Si queremos saber cuales son las participaciones que tienen entre dos vértices, debemos activar una etiqueta en Gephi para que este nos las muestre. En la figura 4.10 notamos que nos aparece una *arista* más *gruesa* comparando con las demás, esto se debe a que entre esos dos vértices existen más colaboraciones, es decir, esos dos autores trabajan más en diferentes artículos. Para esta figura 4.10 también se activo una etiqueta de los vértices para que nos muestre el *ID* que corresponde

a cada autor, esto se puede comprobar con la figura 4.4 y en la figura 4.8 se puede comprobar la cantidad de colaboraciones que existe entre dos vértices.

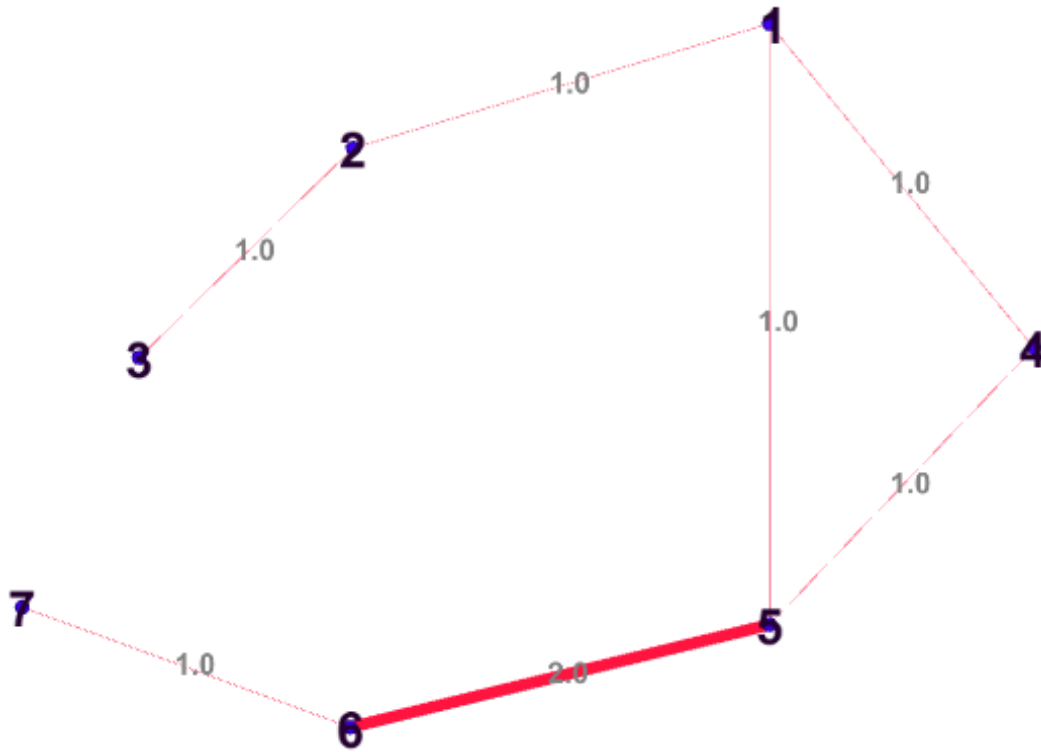


Figura 4.10: Grafo correspondiente a las colaboraciones con peso en las aristas.

Gephi también puede poner los vértices dependiendo del género de cada autor en la figura 4.11 notaremos que los vértices de color *azul* pertenecen al *género masculino* y los vértices *rosas* al *género femenino*.

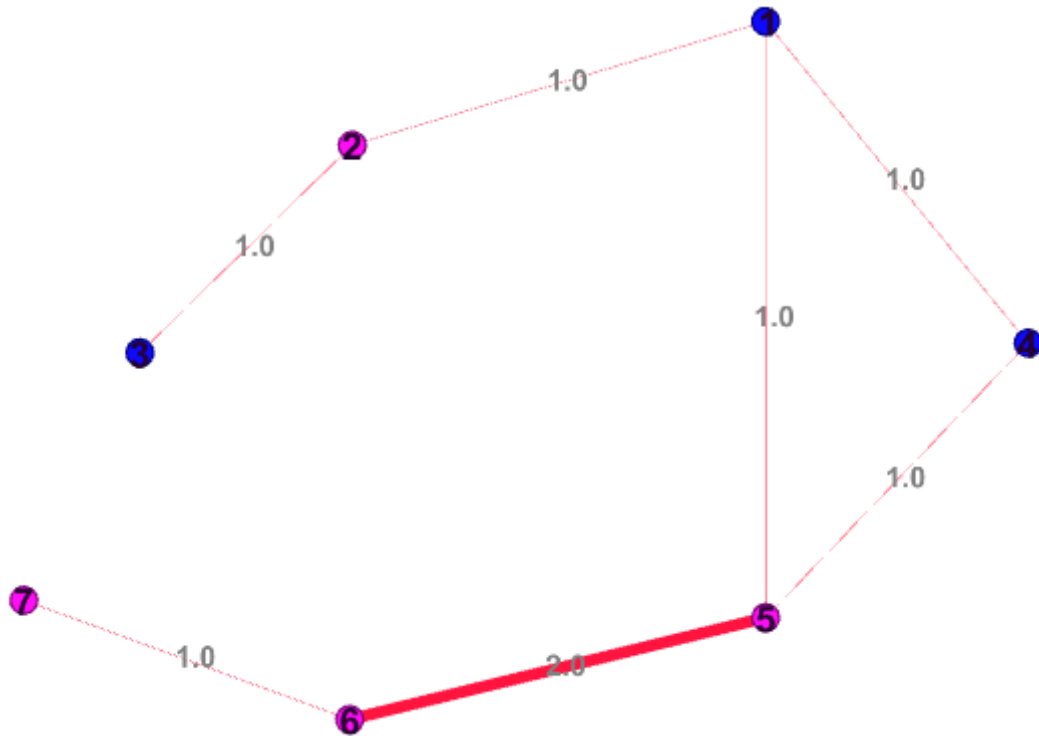


Figura 4.11: Grafo correspondiente a las colaboraciones entre mujeres y hombres.

Además, podemos poner ciertos filtros, como sabemos necesitamos ver las colaboraciones de las mujeres, hombres y de forma global. La forma global la ponemos ver en la figura 4.11. En la figura 4.12 se mostrará las colaboraciones de las *mujeres*.

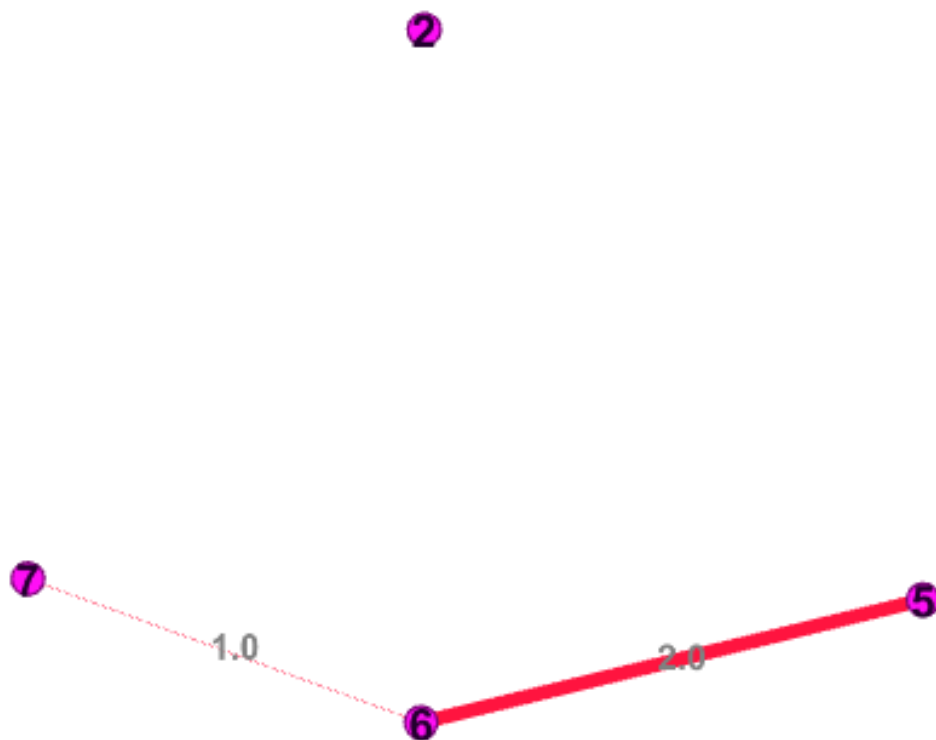


Figura 4.12: Grafo correspondiente a las colaboraciones entre mujeres.

Y en la figura 4.13 se muestran las colaboraciones de los *hombres*.

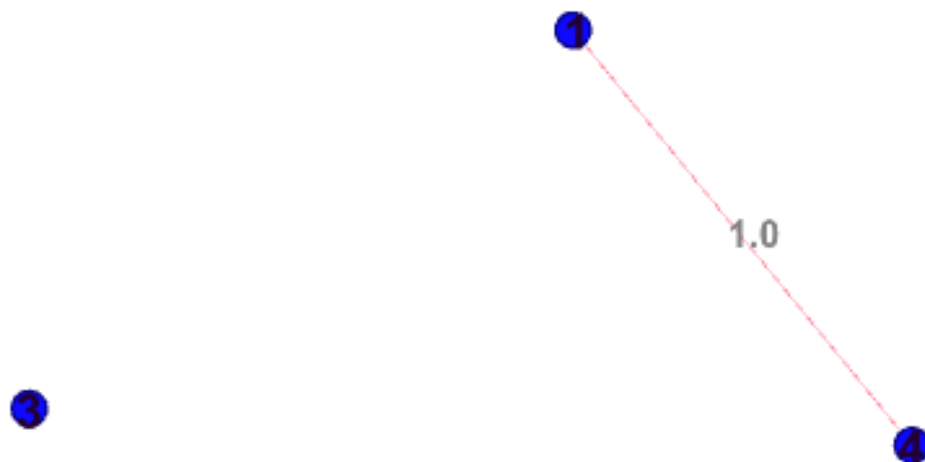


Figura 4.13: Grafo correspondiente a las colaboraciones entre hombres.

Con los conocimientos adquiridos podemos comenzar a realizar nuestro análisis de las colaboraciones de los últimos 10 años de la comunidad de matematicas discretas y combinatoria.

Capítulo 5

Conclusiones

Durante todo este proceso se aprendieron los componentes esenciales que caracterizan un grafo, además, del software que se utilizó para la generación del mismo. Gephi es una herramienta de software bastante interesante, ya que permite realizar de dos formas diferentes el grafo, una de ellas es mediante los archivos de Excel y la otra es con una base de datos. Además, permite interactuar con el grafo de diversas maneras, por ejemplo, añadiendo colores, modificar el tamaño de las etiquetas tanto de los vértices como de las aristas, agregando ciertos filtros, entre otros.

Con Gephi se elaboraron algunos experimentos tanto con archivos de Excel como con una base de datos, los cuales nos mostraron resultados exitosos (véase en la sección 4.1). Por lo tanto se concluyó la parte de la generación de un grafo, sin embargo, no se pudo analizar las colaboraciones de los últimos 10 años ya que no contamos con la base de datos.

Capítulo 6

Trabajo Futuro

Es importante mencionar que el programa para la extracción de datos quedo inconcluso y se tiene contemplado realizar un *Servicio Social*, en donde se propone concluir tanto la base de datos como el análisis y visualización del grafo. A lo largo de este proyecto se logró contestar una de las preguntas, la cual consta de ver la participación de las mujeres, hombres y mixto. Es importante mencionar que las preguntas restantes quedan pendientes.

Bibliografía

- [1] Gephi, <https://gephi.org/> (2020/02/20).
- [2] American mathematical society, mathscinet, <https://bidi.uam.mx:9284/mathscinet/search.html> (2020/12/2).
- [3] Vosviewer, <https://www.vosviewer.com/> (2021/01/03).
- [4] Connected papers, <https://www.connectedpapers.com/> (2021/01/03).
- [5] M. E. J. Newman, Networks: An Introduction, Oxford University Press, 2010.
- [6] Centralidad, <https://www.grapheveryWHERE.com/centralidad/> (2021/02/19).
- [7] Page rank, <https://www.grapheveryWHERE.com/page-rank/> (2021/02/19).
- [8] A. Barrat, M. Barthélemy, A. Vespignani, Dynamical processes on complex networks, Cambridge university press, 2008.
- [9] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, science 286 (5439) (1999) 509–512.

- [10] Qué es un diagrama entidad-relación, <https://www.lucidchart.com/pages/es/que-es-un-diagrama-entidad-relacion> (2021/01/03).
- [11] Claves primarias y foráneas, https://www.ibm.com/support/knowledgecenter/es/SS9UM9_9.1.2/com.ibm.datatools dimensional.ui.doc/topics/c_dm_primary-foreignkeys.html (2021/02/27).
- [12] Práctica de redes sociales con gephi, <https://pagines.uab.cat/joseluismolina/sites/pagines.uab.cat/joseluismolina/files/PR%C3%81CTICA%20DE%20REDES%20SOCIALES%20GEPHI.pdf> (2020/11/10).
- [13] Workbench, <https://www.mysql.com/products/workbench/> (2020/10/15).