Asociación Colombiana de Facultades de Ingeniería

Revista Educación en Ingeniería
http://www.educacioneningenieria.org

Algoritmo genético basado en coeficiente de agrupamiento para la detección de comunidades en red de docentes de la Universidad Industrial de Santander

David Nicolás Camelo García, Paola Carolina Suárez Suárez, Henry Lamos Díaz, David Esteban Puentes Garzón

Escuela de Estudios Industriales y Empresariales, Universidad Industrial de Santander, Bucaramanga, Colombia.
davidpw2898@gmail.com, paolacarolina061297@gmail.com,
hlamos@uis.edu.co,dpuentesgarzon@gmail.com

Resumen— Alrededor del mundo las academias son el centro de concentración y difusión de conocimiento más importante para la sociedad, mediante sus programas de formación profesional, se entrega constantemente a la sociedad motores de conocimiento de alta calidad, los cuales por medio de alianzas bien establecidas a través de la divulgación científica se logra dar solución en distintas formas a las diversas necesidades e inquietudes que abruman en la cotidianidad. En el presente documento se busca dar solución al problema de detección de comunidades (CD) por medio de un algoritmo genético basado en coeficiente de agrupamiento (CC-GA) a una red de colaboración de la Universidad Industrial de Santander conformada por docentes que han dirigido y codirigido trabajos de grado al interior del campus en programas diferentes a los que ellos se vinculan originalmente. Con ello se podrá establecer la condición de colaboración interdisciplinaria de la red, así como identificar los docentes más participativos en estas modalidades, entre otras características representativas de la red.

Palabra clave— algoritmo genético, detección de comunidades, coeficiente de agrupamiento.

Recibido: 1 de junio de 2021. Revisado: 12 de julio de 2021. Aceptado: 8 de agosto de 2021.

A clustering coefficient based genetic algorithm for community detection in a network of teachers from the Universidad Industrial de Santander

Abstract— Around the world, the academies are the most important concentration and diffusion center of knowledge for society. Through their professional training programs, they constantly provide to the society high quality knowledge engines, which through well-established alliances through scientific popularization, are able to provide solutions in different forms to the diverse needs and concerns that overwhelm in everyday life. In this research work, it seeks to solve the community detection problem (CD) through a clustering coefficient based genetic algorithm (CC-GA) to a collaborative network of the Universidad Industrial de Santander formed by teachers who have directed and co-directed work on campus in different programs to which they are originally linked. This will establish the condition of interdisciplinary collaboration of the network, as well as identify the most participatory teachers in these modalities, among other representative characteristics of the network.

Keywords— Genetic algorithm, community detection, clustering coefficient.

1 Introducción

Una empresa, una sociedad o una red social se arraiga en las relaciones entre empleados e individuos [1]; las redes sociales se encuentran involucradas en los diferentes campos de investigación, como la sociología, las matemáticas, la antropología, entre otras. El mundo actual evoluciona a gran velocidad y con ello los datos que este posee, por lo que se hacen necesarios modelos y algoritmos apropiados para trabajar tal magnitud de datos, con el objetivo de lograr una cuantificación de las características propias de una comunidad, y así extraer sus principales cualidades para su posterior análisis.

La Universidad Industrial de Santander (UIS) con su generación continua de información y datos como la formación de profesionales integrales mediante la difusión de conocimiento, construye procesos colaborativos por medio de la creación de proyectos interdisciplinarios, otorgándoles la profundización de competencias adquiridas en su etapa académica; se convierte en un objetivo clave de estudio para la detección de comunidades pues "es más probable que las personas comparta sus ideas con gente de la misma área, la misma universidad, o al menos que hablen el mismo idioma" [2].

Actualmente la Universidad Industrial de Santander (UIS) se encuentra en el top 7 en la clasificación de las mejores IES colombianas según indicadores de investigación del Ranking U-Sapiens 2020-1 como reflejo de la importancia de la investigación en la universidad y su aporte hacía la sociedad.

En ese sentido, durante años se ha tratado de entender los comportamientos científicos y de investigación para una difusión de conocimiento íntegro ante la comunidad universitaria y la comunidad en general. A través del tiempo el estudio de redes sociales (SN por sus siglas en inglés, Social Networks) ha sido útil para describir, distinguir e interpretar las conductas de estas; las redes de colaboración se han enfocado en hallar cómo se comportan las comunidades en líneas disciplinarias, metodologías de investigación, temáticas estudiadas dentro de una disciplina, entre otros.

Si bien, este tipo de redes se han estudiado desde los años 80, de acuerdo con Newman, (2001) en su trabajo "*The structure of scientific collaboration networks*" afirma que anterior a tal fecha no se habían realizado reconstrucciones detalladas de las mismas y a partir de allí su estudio retrata estructuras puntuales y demarcadas sobre las interrelaciones entre científicos, artículos, colaboraciones y subgrupos de trabajo en campos especializados.

Las conexiones interdisciplinarias permiten tener mayor perspectiva del conocimiento, técnicas, aprendizaje en conjunto y visión integral de las problemáticas, a su vez, este tipo de estudios dan paso a aplicaciones como "la red basada en citas bibliográficas" [2], que pueden optimizar los resultados de búsqueda en plataformas académicas, con el fin de que se recomiende literatura más valiosa sobre el tema requerido.

De acuerdo con ello, el presente trabajo pretende destacar aquellas áreas y tendencias de mayor impacto mediante el estudio de patrones de colaboración científica o colaboración de áreas de conocimiento entre proyectos realizados con base en el problema de Detección de Comunidades (CD), haciendo uso de un algoritmo genético basado en el coeficiente de agrupamiento, en adelante coeficiente de clustering (CC-GA) a una red de colaboración conformada por docentes de los distintos grupos de investigación de la Universidad Industrial de Santander mediante la dirección y codirección de trabajos de grado interdisciplinarios.

2 Revisión de literatura

Entender y predecir el comportamiento de distintos grupos sociales es un tema de interés para la ciencia, los negocios, la política, entre otros. Sin embargo, no es una tarea sencilla de realizar por medio de técnicas cotidianas de fácil obtención en el entorno debido al gran dinamismo de dichas redes. Esto ha dado vía libre para el desarrollo de investigaciones que traten de llegar a una solución óptima con base en distintos objetivos de caracterización bajo el estudio de redes sociales y redes complejas.

A través de la detección de comunidades se logra la caracterización de particularidades de las redes objeto de estudio, ya que presenta diversas aplicaciones; el funcionamiento de sistemas en una organización [3], en grupos de decisión [4], influencia de las personas entre sus redes [5], tipos de relaciones entre los grupos [6], así como gran variedad de empleos y usos de acuerdo al análisis deseado por el investigador.

Dado lo anterior, la siguiente revisión de literatura presenta la evolución del uso de diversos algoritmos y técnicas utilizadas en diferentes aplicaciones del CD. Esta temática fue estudiada inicialmente (de acuerdo con las citaciones realizadas en la mayoría de los artículos y la cronología de los mismos), por [7] mediante la aplicación de detección de estructuras comunitarias, por medio del algoritmo de Girvan y Newman (GN) utilizando el método conocido de agrupamiento jerárquico, aplicado a distintas "redes ficticias", comparando los resultados con otras técnicas tradicionales de menor complejidad. Estos autores demostraron que su algoritmo obtenía mejores respuestas en menor tiempo, gracias a que su algoritmo también comprende el concepto de modularidad, el cual es un aspecto que permite la evaluación de conexiones entre los nodos de la red, logrando la optimización de este y alcanzando mejores niveles de conexión *intrared* y *entrered.*

A raíz de ello, este mismo método de agrupamiento jerárquico, fue usado por [8] acoplándose al problema del agente viajero (TSP) otorgándole a tal problema un enfoque complementario, demostrando que además de encontrar la ruta más corta de viaje y menos costosa, todos los individuos visitados y atendidos por el "vendedor" presentan características de comportamiento de redes complejas que pueden caracterizarse por medio de conglomerados que representan características similares acorde a criterios de análisis de clúster realizados después del hallazgo de la ruta más corta del problema determinando a su vez comunidades entre los nodos pertenecientes a un mismo camino. Sin embargo, de acuerdo con [9] esta técnica presenta algunos inconvenientes, debido que el proceso de eliminación de aristas va cambiando constantemente el valor de modularidad y agrupación inicial, lo cual afecta de manera considerable el valor óptimo del mismo al finalizar la iteración, encontrando posiblemente valores indeseables en el resultado final.

A raíz de ello, con intención de optimizar los resultados obtenidos, al problema de CD se han introducido distintos tipos de algoritmos, los cuales también dependen "en gran medida de la topología de la red, ya sea dinámica o estática." [10]. Esto se debe a que cada algoritmo funciona de distinta forma en cada área de aplicación lo cual resume que, si en una red el algoritmo funciona bien, para otra red con características muy similares, la eficiencia de este puede no ser suficiente.

Dichos algoritmos se han separado en Algoritmos Evolutivos (EA) entre estos se presenta una extensión como lo son los Algoritmos Genéticos (GA). [11]–[14] aplicaron técnicas de EA para el problema de CD, implementando en varios operadores conocidos como operadores cruzados (crossover), los cuales funcionan "combinando dos individuos para producir una o dos descendencias. Por ello es un operador muy importante, ya que debe generar una descendencia con las mejores propiedades heredadas de sus padres, mientras que también debe mantener la diversidad de la población." [13]. Cabe mencionar que los autores previamente mencionados, aplicaron este algoritmo con base en la optimización multi-objetivo, teniendo como principio el concepto de modularidad aplicado por [7] testeado en la "red de colaboradores de del Instituto Santa Fe de Nuevo México, constituida por un centro de investigación interdisciplinario", a través de los siguientes criterios: "Bajo número de valores entre clústeres y un alto número de valor *intra cluster*" [14], donde el algoritmo determina una clasificación de la red en 2 tópicos; similitudes de tema de investigación y metodología aplicada, demostrando por qué en los departamentos interdisciplinarios se pueden encontrar conexiones frente a otros, e intuyendo un pronóstico sobre posibles futuros vínculos entre campos de conocimiento vagamente similares.

Por otro lado, los GA son algoritmos más utilizados para la optimización de problemas complejos, estos se adaptan completamente a la dificultad requerida en la CD. "Se basan en mecanismo con los procesos evolutivo como la reproducción, selección natural entre otros, pueden determinar automáticamente el número de clústeres en una red, lo que los hace útiles para redes del mundo real" [5]. Este mismo autor realizó una comparación entre diversos GA aplicándolos a redes sociales; Algoritmo Genético Multiobjetivo (MOGA-net), Densidad de Modularidad y Modularidad Rápida, usando Información Mutua Normalizada (NMI), la cual calcula la calidad de las particiones realizadas a través de estos métodos, concluyendo para este caso que el mejor nivel de NMI encontrado entre dichos métodos es alcanzado por el método de densidad de modularidad.

Adicional a ello la gran variedad de GA permite la integración de diversas técnicas de agrupamiento mejoradas, entre ellas la más conocida y usada en la investigación es el *Community Coefficient*-CC; un concepto implementado desde 1998 para determinar y comprobar la existencia de una propiedad común de las redes sociales reales, [15] ejemplifica tal característica en redes de colaboración, donde los centros de conocimiento podrían formar comunidades científicas importantes y a su vez conjuntos de investigadores que trabajen en subcampos especializados particulares. [16] utilizó dicha técnica con el fin de alcanzar una maximización adaptativa a redes ficticias, sin embargo, también concluye que con base en el algoritmo principal no es factible el uso de ellas sobre el análisis de redes o grafos con un tamaño reducido. Posteriormente [17] implemento la CC en medio de redes superpuestas, lo que da paso a encontrar no solo mejores valores de agrupamiento entre nodos, sino también la diversidad de conexiones que pueden tener éstos de acuerdo a características que no son comunes para todos los individuos pertenecientes a un grupo, esta característica de comunidades superpuestas mediante CC, se explica mejor a través del CD en Multicapas estudiados por [18] en "Differential Flattening: A Novel Framework for Community Detection in Multi-Layer Graphs" y por [19] en "*Community detection for multi-layer social network based on local random walk*" quienes determinan que cada capa representa una característica de una comunidad en especial. Por ejemplo: "Trabajo, Amigos, Facebook, Coautor y Comida" [19], sin embargo, como es de inferir, un individuo puede pertenecer a más de una comunidad al mismo tiempo, por ello estos autores evalúan el CC local, donde determina agrupaciones en una sola capa del problema y el CC multicapa, tratando de conectar en grupos a individuos existentes en todas las capas el mismo tiempo sin tener que separarlos de su capa principal.

Si bien el CC se puede usar como técnicas de evaluación y optimización extras al GA, [20] en su artículo "CC-GA: A *clustering coefficient based genetic algorithm for detecting communities in social networks*", como bien su nombre lo indica, integraron el CC a un GA (CC-GA), con el fin de que las iteraciones de mismo vayan evaluando todas las condiciones para alcanzar no solamente identificación de comunidades sino encontrar también grupo de gran cohesión entre distintos nodos de la red, esto les permite evaluar redes más grandes y complejas con niveles de eficiencia más altos. Con base en los resultados obtenidos por los autores el CC-GA alcanza mejores resultados en el análisis de redes sociales, donde en 6 de las 10 redes analizadas obtuvo un mayor valor de modularidad comparado contra otros algoritmos de GA; (TGA, LGA y MENSGA), en este mismo orden de ideas este algoritmo presenta la capacidad de crear conexiones con gran modularidad, alcanzando los valores máximos en menos cantidad de tiempo con respecto a los demás algoritmos.

En resumen, el CD es un problema que ha logrado el interés de investigación en distintas áreas de aplicación, el cual, si bien empezó a trabajarse desde 2002, se ha llevado mejoras a la solución de este problema por medio de gran diversidad de algoritmos desde el 2014, facilitando en gran medida las tareas del análisis de data. Se observa que entre los algoritmos de gran uso en tal problema están los GA, los cuales debido a su gran variedad de operadores logran ligarse completamente con las características propias de las redes sociales. Cabe destacar algunos aspectos importantes en la aplicación de la CC a dichos algoritmos, puesto que logran conseguir agrupaciones más concentradas durante la identificación de comunidades, con mejores valores de conexión en menor tiempo. Por ende, dado el gran uso de estos dos últimos casos mencionados, el presente trabajo llevará a cabo mediante la aplicación del CC a un GA para así fortalecer los conocimientos en las distintas áreas de los involucrados en los proyectos interdisciplinares, encontrando posibles focos de conocimiento asociado.

3 Metodología

En la Fig. 1 se representan los pasos implementados para el desarrollo del trabajo.

Figura 1. Metodología implementada en la indagación.
Fuente: Los autores.

3.1 Recolección de la información

La recolección de información parte de la familia de docentes adscritos a los grupos de investigación de las facultades de ingenierías fisicomecánicas (FIFM) y fisicoquímicas (FIFQ) de la universidad en cuestión, elegidas por su grado de representación de la información, de acuerdo con los criterios específicos para la elección óptima de los datos que implique tantas interacciones interdisciplinarias como sea posible.

Los datos utilizados se obtuvieron a partir del Sistema Nacional de Ciencia y Tecnología del Ministerio colombiano de ciencia, tecnología e innovación (Minciencias), se recolectaron los listados de los docentes vinculados a los grupos de investigación explorados en la página de la UIS en cifras, verificando que la información exportada estuviera dentro del periodo de vinculación vigente del docente en cada grupo, dando como resultado 19 grupos de investigación para FIFM y 11 para FIFQ.

Usando como criterio de selección los docentes que poseían colaboraciones en la dirección o codirección de proyectos de grado de pregrado con otras carreras distintas a la escuela de vinculación del grupo de origen, concluyendo para la Facultad de Ingenierías Fisicomecánicas (FIFM) 93 docentes con distintas colaboraciones interdisciplinarias mientras que para la Facultad de Ingenierías Fisicoquímicas (FIFQ) se encontraron un total de 80 docentes.

3.2 Preparación de los datos

Posterior a la recolección de los datos, fue pertinente la correcta preparación de estos, con el fin de evitar reprocesos y confusiones con el algoritmo, se eliminó la múltiple aparición de los docentes en el archivo base, ya que algunos de ellos se encontraban adscritos a más de un grupo de investigación, a la vez, también se consideró adecuada la asignación a cada docente de un número de identificación, y así establecer formalmente la red de colaboración de los docentes recolectados. De este mismo modo se logrará una corrida más eficiente del algoritmo, dado que de esta forma no tendría que reconocer cierta cantidad de caracteres, sino un solo número.

3.3 Construcción de la red de docentes

Con base en la información mencionada, se anexó la información en un fichero para plantear la red inicial con la librería Networkx de Python, donde un total de 166 docentes adscritos a cada uno de los grupos de investigación se extiende a 280 docentes mediante colaboración, definiendo como punto de partida a los docentes como los nodos de la red y los enlaces como la cantidad de colaboraciones; ello genera un total de 300 nodos con 347 enlaces, no obstante, debido a que muchos de los datos encontrados fueron directores "independientes", se encuentran 146 enlaces entre ellos mismos lo que deja un total de 154 con un docente distinto (verde), cabe mencionar que dentro estos 146 relaciones independientes 95 de ellos no colaboraron con algún otro docente (naranja). Ver Figura 2.

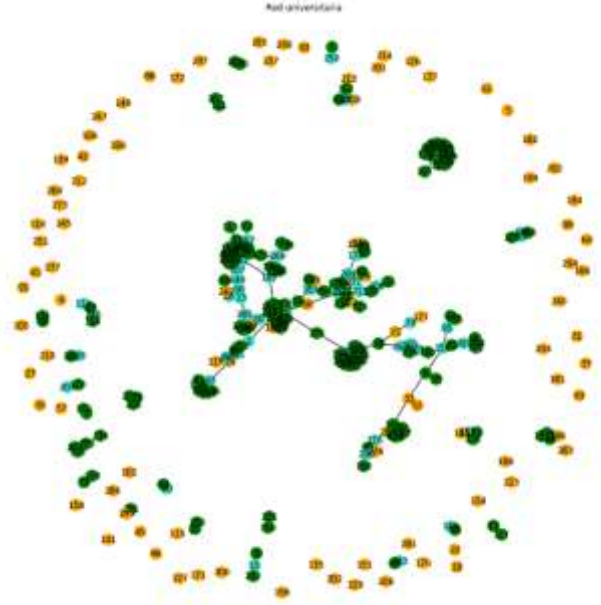

Figura 2. Red universitaria. Generada por librería Networkx y Matplotlib de Python.
Fuente: Los autores.

3.4 Detección de comunidades

Con base en la información obtenida durante la revisión de literatura, se da paso a la formulación del algoritmo a implementar en la presente investigación mediante la adaptación de elementos importantes del algoritmo planteado por [20] en "CC-GA: *A clustering coefficient based genetic algorithm for detecting communities in social networks*", el cual, como se ha mencionado anteriormente, tiene como objetivo aplicar un algoritmo genético partiendo del cálculo del CC en cada uno de los nodos con base en la ecuación 1. El coeficiente de clustering evalúa la tendencia de agrupamiento entre los nodos lo cual representa la probabilidad de conexión entre los vecinos un mismo nodo lo cual determina la eficiencia de las comunidades encontradas, variando entre 0 y 1 acorde a los enlaces entre sus vecinos dicho cálculo se realiza mediante la evaluación de grados de conexión entre un conjunto de vecinos de la estructura de red, si el nodo $\boldsymbol{v_i}$ tiene un único vecino $\boldsymbol{v_j}$ (es decir un valor de 1) entonces estos dos nodos están conectados, sin embargo si el nodo $\boldsymbol{v_i}$ no presenta ningún vecino (o valor de 0), dicho nodo se aislará en una comunidad separada para posterior análisis.

$$\boldsymbol{C_{vi}} = 2 * \frac{L_i}{K_i(K_i-1)} \quad (1)$$

Dónde: L_i corresponde al número total de enlaces entre vecinos del nodo V_i y $K_i$representa el grado del vértice *i.*

La utilización de la medida del CC tiene como fin disminuir criterios de partida del algoritmo, pues se entenderá que la creación de la población inicial se generará a partir de enlaces con alto nivel de conectividad según el coeficiente del vecindario cercano en cada nodo.

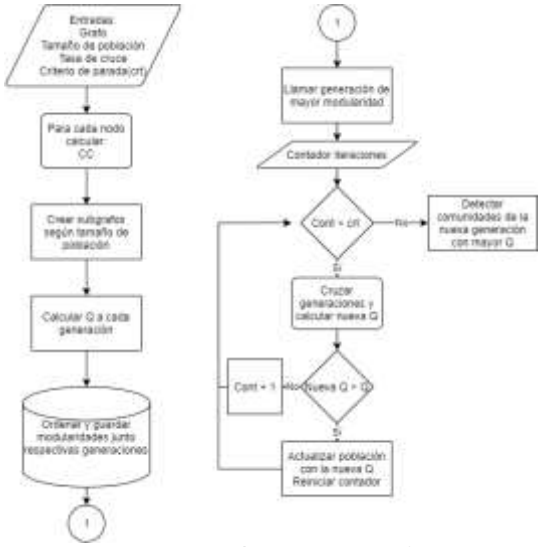

Figura 3. Flujograma del algoritmo genético basado en coeficiente de agrupamiento.
Fuente: Los autores.

Tal como se observa en la Fig. 3, el algoritmo se separa en 2 etapas fundamentales: la primera, una vez se ingresan parámetros de entrada del algoritmo, como la red a utilizar, la tasa de

población, tasa de cruce y el criterio de parada, que determina cuándo parar después de no encontrar mejorías de valor de aptitud. Se procede a calcular el CC en cada uno de los nodos existente en la red, con el fin de crear pequeños conjuntos de subgrafos o subredes con el mejor vecino en cada nodo, conformando así los cromosomas de la población inicial de trabajo.

Simultáneamente en este paso, a cada cromosoma creado se le calcula su valor de aptitud *fitness*, conocido en el trabajo de redes como modularidad (2), comparando los valores continuamente para así obtener el mejor valor como punto de partida del operador cruzado.

$$\boldsymbol{Q} = \sum_{\boldsymbol{c=1}}^{\boldsymbol{n}} [\frac{\boldsymbol{L_c}}{\boldsymbol{L}} - \left(\frac{\boldsymbol{K_c}}{\boldsymbol{2L}}\right)^{\boldsymbol{2}}] \quad (2)$$

Continuando ahora con la segunda etapa del algoritmo, se toma el mayor valor *fitness* como referencia, y se selecciona aleatoriamente cromosomas de la población inicial para proceder en parejas con el crossover, este operador está ligado a crear enlaces entre nodos para a partir de ella generar nuevas descendencias, a través de una cadena binaria, igual al tamaño de la red, con valores aleatorios de 0 o 1, escogiendo genes del primer progenitor cuando el valor de esta cadena es 1 y del segundo progenitor cuando es 0, garantizando descendencia de genes conectados en las nuevas generaciones como se muestra en la Fig. 4Figura *4*. Durante este cruce suceden 2 procesos más, el primero es el cálculo continuo de modularidad en las nuevas generaciones y el segundo es la mutación que tiene como intención lograr que comunidades pequeñas puedan incluirse en comunidades más grandes, siempre y cuando su vecindario lo permita, creando enlaces temporales entre nodos frontera con tal de crear cohesión en estos puntos de la red, evaluando continuamente la modularidad del grafo obtenido.

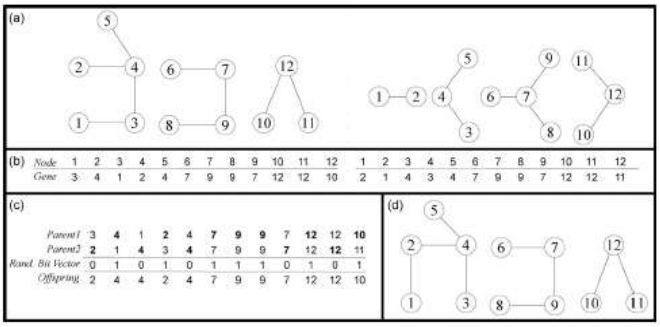

Figura 4. Crossover CC-GA.
Fuente: Tomado de [20].

Estas iteraciones ocurren con la intención de encontrar un valor de modularidad más alta que la referencia tomada al comienzo de la etapa 2, y se tomará como referencia el valor mayor encontrado en el transcurso de iteraciones, hasta el punto en que no se encuentre un mejor valor después de superar el criterio de parada propuesto al comienzo de la operación, pues, una vez se supera el límite se asume que no habrá mejor aptitud de cohesión en la red original y se procederá a determinar las particiones indicadas que conformarán las comunidades encontradas de la misma.

3.5 Análisis de comunidades

La detección de comunidades es una parte fundamental del análisis de redes sociales, ya que con ésta se puede obtener claridad acerca de los patrones de comportamiento en la creación de estructuras comunitarias, que no son características dadas al azar, sino definidas por algún tipo de interés o particularidad en común. Es por esto por lo que se procuró la búsqueda de comunidades en una red de colaboración de proyectos de grado dentro de la universidad, para de esta forma poder definir algunos atributos usuales y así diagnosticar las relaciones colaborativas con más fortaleza y sentar un precedente para posibles investigaciones futuras que pretendan impulsar la colaboración interdisciplinaria dentro del claustro. Para ello se han seleccionado cinco comunidades al azar, para comprobar la validez del algoritmo y asimismo analizar un poco más a fondo una pequeña muestra de la red creada.

4 Resultados

Una vez implementado el algoritmo CC-GA modificado para esta red, se alcanzó un valor de modularidad de 0.97364. La modularidad es una medida de aptitud en los problemas de CD, que se encarga de medir la calidad que tienen las conexiones encontradas en cada iteración del algoritmo medido entre -1 y 1, donde los valores negativos representan baja conectividad de la red mientras que los valores positivos una buena conectividad. [21] define la modularidad como "el número de bordes que existen dentro de los grupos menos el número esperado en una red equivalente con aristas colocados al azar", por lo tanto, la función determina un menor número de conexiones que las esperadas en la red. Es decir, se presenta entre 2 grupos una cantidad de "aristas pequeñas, y al interior de cada comunidad el equivalente es mayor, lo que concluye que la estructura comunitaria alcanza grandes características atractivas. En este caso al llegar a un resultado cercano al 1 se garantiza buena calidad en las estructuras comunitarias encontradas que comprueba la fuerza en los enlaces de la red, con un total de 133 comunidades encontradas después de 93 iteraciones.

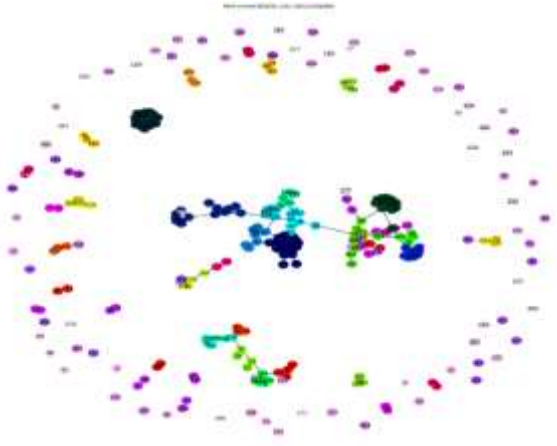

Figura 5. Red universitaria con comunidades obtenidas a partir del algoritmo genético basado en coeficiente de agrupamiento.
Fuente: Los autores.

En la Fig. 5Figura 5 se observa la red universitaria con sus comunidades mapeadas por colores, con el fin de poder diferenciar las comunidades más grandes, se agrupan de manera especial las que tienen 2 y 1 individuos en 4 y 6 conjuntos respectivamente, ya que representan más de la mitad del total de comunidades. Así las comunidades grandes son aquellas que toman colores tierra y azules, los intermedios colores fríos entre azules claros y verdes, las medianas colores cálidos mientras que los conjuntos especiales se distinguen entre morados y blancos.

Dentro de las 133 comunidades encontradas, se puede observar que la de mayor tamaño, cuenta con su líder representado por el nodo 71, adscrito al Grupo de Investigaciones en Corrosiones, donde gran parte de las colaboraciones fueron realizadas con la escuela de Ingeniería Química, seguido de Ingeniería Electrónica. De otro modo, es importante resaltar la existencia de comunidades con un solo individuo dentro de ellas, esto debido a la gran dispersión de los datos y la incapacidad del algoritmo para la creación de enlaces en estos casos específicos, cabe resaltar que los docentes presentes en estas comunidades individuales se encontraban adscritos a un grupo de investigación de una escuela diferente a la del proyecto en cuestión, algunos de ellos en el rol de codirectores, pero sin información suficiente en la hoja de vida electrónica para la creación del vínculo pertinente. En la Tabla 1 se muestra las comunidades que poseen más de un integrante, con los líderes de cada una, así como la cantidad de participantes en ella.

Tabla 1
Listado de comunidades más grandes

Comunidad	Tamaño comunidad
1	21
2	18
3	15
4	14
5	9
6	8
7	7
8	7
9	6
10	6
11	6
12	5
13	5
14	4
15	4
16	4
17	4
18	4
19	4
20	3
21	3
22	3
23	3
24	3
25	3

Fuente: Los autores.

Ahora realizando un análisis individual se han seleccionado cinco comunidades al azar, para comprobar la validez del algoritmo, y asimismo analizar un poco más a fondo una pequeña muestra de la red creada. A continuación, se presentan cada una de ellas.

Comunidad 16: Como se puede observar en la Fig. 6 esta comunidad se encuentra compuesta por cuatro docentes, identificada con el color magenta oscuro, además se puede notar que es una comunidad aislada de las demás ya que no posee un nodo fronterizo que la conecte con el resto.

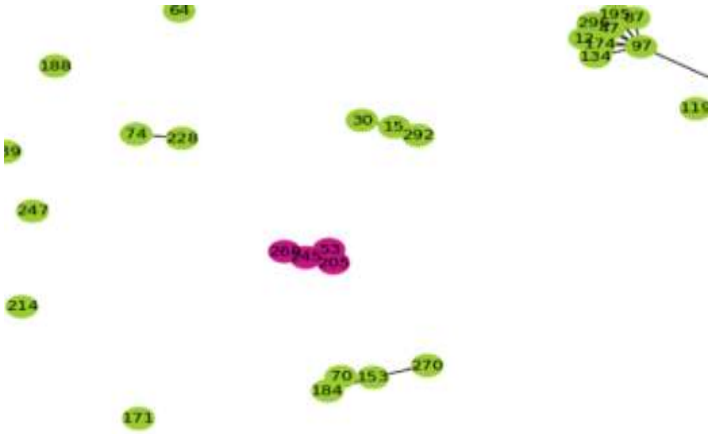

Figura 6. Comunidad 16. Generada con Matplotlib y Networkx de python.
Fuente: Los autores.

El líder de este grupo se consolida el nodo 53, docente adscrito a la Escuela de Eléctrica, Electrónica y Telecomunicaciones (E3T) con el grupo de investigación CEMOS, colaborando con otros tres docentes bajo el rol de director de proyectos para la escuela de Ingeniería de Sistemas. La docente representada por el nodo 245, al momento de realizar la colaboración con el profesor en mención, también se encontraba en vinculación con el grupo CEMOS en el año 2012, lo que nos lleva a visualizar la fuerza de la E3T en los proyectos de la escuela de Ingeniería de sistemas, dado que su dirección y codirección estuvieron a cargo de esta escuela específicamente en este proyecto. También es importante resaltar que el líder de la comunidad en el rango de 2014 a 2017 estuvo vinculado al grupo de investigación Cómputo Avanzado y a Gran Escala de Ingeniería de Sistemas, donde en este periodo realizó la dirección de 17 proyectos para la Escuela de Ingeniería Electrónica.

Esta colaboración interdisciplinaria es una de las más frecuentes y esperadas de la red, dado el campo de conocimiento de cada una de las escuelas mencionadas, su constante relación y complementariedad de conocimientos, aunque puede llegar a ser demasiado hermética y descuidar la posibilidad de la transmisión de conocimiento a otras comunidades y disciplinas.

Comunidad 3: Se encuentra formada por 15 docentes, y está demarcada por el color ladrillo como se puede apreciar en la Fig. 7.

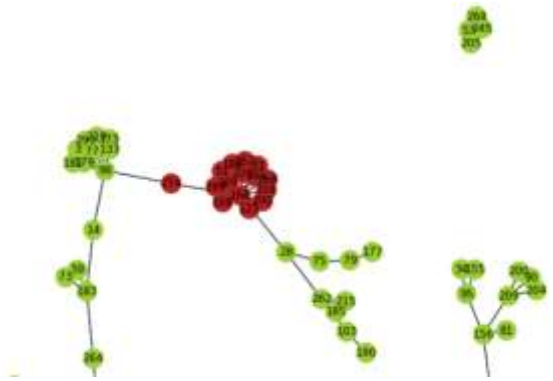

Figura 7. Comunidad 3. Generada con Matplotlib y Networkx de Python.
Fuente: Los autores.

Esta comunidad se encuentra ubicada en la facultad de Ingenierías Fisicoquímicas, teniendo como líder al docente identificado como el nodo 168, adscrito a los grupos *Energy and Other Non Renewable Resources Research Group* y Grupo de Investigación en Geología Básica y Aplicada de la escuela de Geología, dirigiendo en 58 proyectos de la escuela de Química, Ingeniería Química e Ingeniería Metalúrgica. Es bastante esperado este comportamiento, dado que el docente posee un pregrado en Química y el resto de su trayectoria gira en torno a esta disciplina, también es importante resaltar que, aunque es de las comunidades más grandes se encuentran los nodos muy aislados de los demás, y sus nodos fronterizos se encuentran distribuidos en el líder y la docente del nodo 274. Este último cumple una función importante en las colaboraciones interdisciplinarias, ya que es el conector entre una comunidad perteneciente a la FIFQ con otra de la FIFM, gracias a la colaboración de la escuela de Geología con Ingeniería Civil. A pesar de la cercanía existente entre los distintos conglomerados de nodos que se observan, estos no se agrupan dentro de la misma comunidad debido a la ausencia de colaboración entre los demás integrantes, pues tal como se observa, la comunidad mencionada y las comunidades del entorno se distribuyen a través del nodo líder.

Comunidad 131: Esta es una comunidad particular, dado que se encuentra compuesta por un solo individuo, el cual se puede apreciar de color en la Fig. 8 de color azul oscuro.

Figura 8. Comunidad 131. Generada con Matplotlib y Networkx de python.
Fuente: Los autores.

Este es un caso particular y que se generó con bastante frecuencia en la detección de comunidades de la red en cuestión; se trata de una estructura comunitaria individual generada por la dirección del docente 158 adscrito a Energy And Other Non Renewable Resources Research Group, con una validez en el grupo de investigación desde 2018 a la actualidad, de la escuela de Geología en colaboración con Ingeniería de Petróleos.

El docente cuenta con un pregrado en Ingeniería Civil, un doctorado en la Universidad del Sur de Carolina en Geología y una maestría y especialización en Geofísica en la Escuela de Minas de Colorado, aún con todo lo anterior, el algoritmo no identificó suficientes recursos por lo que generó conexiones débiles de esta comunidad, hasta el punto de dejarla totalmente aislada.

Comunidad 28: Formada por 2 individuos, identificados mediante el color naranja en la Fig. 9.

Figura 9. Comunidad 28. Generada con Matplotlib y Networkx python.
Fuente: Los autores.

Esta comunidad se encuentra formada por la relación entre el docente identificado por el nodo 131, Ingeniero Mecánico de la Universidad Industrial de Santander y PhD en Ingeniería Civil de la Universidad Politécnica de Cataluña y el nodo 267, Ingeniero Eléctrico y Electrónico de la UIS. La relación entre ellos fue generada mediante la dirección y codirección, respectivamente, de un proyecto de Ingeniería Electrónica bajo la dirección del grupo de investigación DICBOT de la escuela de Ingeniería Mecánica.

Dicho lo anterior, se evidencia un complemento de conocimientos interdisciplinarios, dado que el director del proyecto no posee estudios certificados específicos acerca del área dirigida, pero sí el codirector, aun así, la red al igual que el caso anterior no posee la suficiente fortaleza en los enlaces de la red, por lo que también se evidencia una comunidad aislada sin conexión con otras.

Comunidad 22: Formada por 3 individuos, identificados mediante el color gris en la Fig. 10.

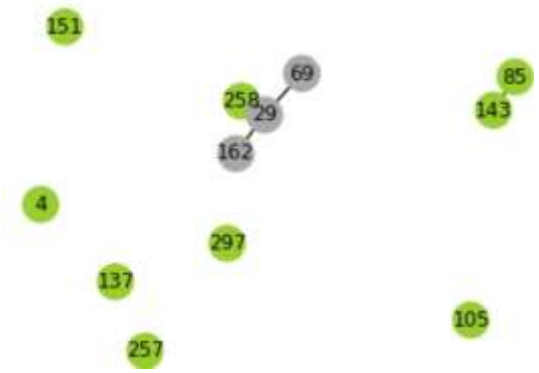

Figura 10. Comunidad 22. Generada con Matplotlib y Networkx python.
Fuente: Los autores.

Así como en la comunidad 16, esta comunidad representa la relación entre las escuelas de Ingeniería de Sistemas e Ingeniería Electrónica, teniendo como líder al nodo 29 Ingeniera Electrónica y PhD en Ingeniería Eléctrica, adscrita al grupo de investigación en diseño de algoritmos y procesamiento de datos multidimensionales de la escuela de Ingeniería de Sistemas, 69 y 162 también miembros de la comunidad e Ingenieros Electrónicos.

Este caso es particular, debido a que la líder de la comunidad no tiene algún estudio específico en la disciplina correspondiente al grupo de investigación, más, no son especialidades aisladas, dado que algunos proyectos de Ingeniería Electrónica pueden llegar a necesitar el soporte de la Ingeniería de Sistemas en cuestiones de programación y automatización de características internas del proyecto en cuestión.

4.1 Análisis general de la red

De forma general de toda la red, las comunidades de un solo individuo conforman el 64% de la cantidad de comunidades encontradas, este grupo se conforma por 85 docentes de los cuales, de la información recolectada suministrada en cada CvLAC, su modalidad de dirección fue independiente, no obstante, en este caso se redujeron 10 agrupaciones esperadas, pues acorde a lo mencionado en la red inicial se habían encontrado 95 docentes que únicamente realizaron colaboración independiente. Adicional a esto, también se conformaron parejas de colaboración con un total de 23 que abarcan el 17%, ambos conjuntos en total comprenden el 81% del total de comunidades definidas, aun así, en el restante 19% los integrantes de cada comunidad presentan patrones de relación y colaboración interesantes en cada conglomerado, ya sea por títulos académicos, experiencia de profesión que brindan enriquecimientos a los resultados alcanzados.

4.2 Análisis de longitud de ruta más corta

En cuanto a otras métricas de calidad dentro de la red, se encuentra la longitud de ruta más corta, que se define la distancia que se tiene que recorres por la red entre 2 nodos a través de los enlaces existentes entre los demás, para término de las redes de colaboración esta métrica se traduce como la cadena de referidos con el cual se busca relacionar 2 docentes o autores que no se conocen, buscando crear nuevas colaboraciones entre sí. De principio este concepto tiene pocos resultados en la red universitaria de estudio, dada la ausencia de conexiones entre nodos, la dispersión de individuos, comunidades con baja cantidad de individuos y comunidades con conexiones casi lineales. No obstante, en las zonas más centrales, como se ve en la Fig. 17, se evidencia la posibilidad de encontrar cadenas sobresalientes, como, por ejemplo, escogiendo el nodo 58, que es nodo frontera de la comunidad 7 y el nodo 173 perteneciente a la comunidad 5.

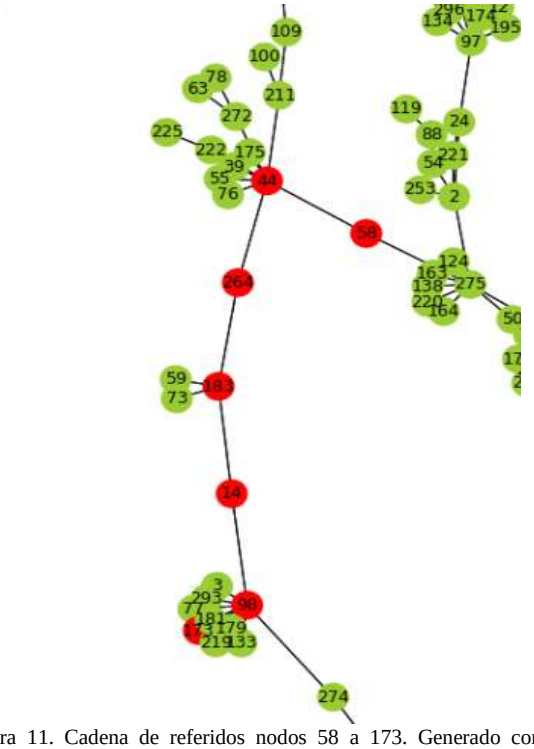

Figura 11. Cadena de referidos nodos 58 a 173. Generado con Matplotlib y Networkx de python.
Fuente: Los autores.

En la Fig. 11 se muestra la ruta de referidos en color rojo entre los nodos mencionados, que es necesario contactar con un total de 6 docentes para que el 58, ubicado en la parte superior, logre contactar al nodo 173 en la parte inferior. Durante esta cadena, se encuentran docentes adscritos a los programas de ingeniería industrial, diseño industrial, ingeniería civil, ingeniería mecánica y geología, lo cual constata la interdisciplinariedad de las distintas comunidades que se conjuntan en esta zona de la red. El nodo 58

hace referencia al docente que ha pertenecido desde el 2004 al grupo de investigación GIEMA de la escuela de ingeniería mecánica ha presentado fuertes relaciones con docentes de ingeniería industrial, diseño industrial e ingeniería de petróleos, así como trabajos dirigidos para ingeniería química. El segundo corresponde al docente magister en geología de la UIS, que a pesar de ello solo se encontró un trabajo codirigido en la misma área con un docente de ingeniería civil.

5 Conclusiones

En base a la revisión de literatura, el problema de detección de comunidades es ampliamente aceptado y utilizado para realizar estudios en las redes de colaboración, no obstante, hasta el momento la cantidad investigaciones registradas en esta temática puede llegar a ampliarse dada la importancia y utilidad de la aplicación de estos conceptos en los algoritmos.

El CC-GA modificado, logró detectar 133 comunidades con un valor de modularidad máximo de 0,9738, garantizando una estructura comunitaria de mejor calidad dada su cercanía con su máximo valor posible, siendo este el 1. El comportamiento general de las comunidades se basó en su gran mayoría en la colaboración entre escuelas de la misma facultad, así como docentes que no realizaron su pregrado en la escuela de origen del grupo de investigación al que se encuentran adscritos, pero si alguno de sus estudios complementarios, a excepción de algunos casos atípicos, donde no existía concordancia entre la finalidad del grupo y la información acerca de la preparación del docente.

Un 81,79% de las comunidades están compuestas por estructuras individuales aisladas, dada la debilidad de los enlaces para la creación de vínculos con otros nodos, y a la vez conectar con otras comunidades. Si bien, toda información fue exportada de la página oficial de Minciencias, es preciso resaltar que los resultados podrían variar con la existencia de información más detallada en tal plataforma por parte de los docentes vinculados a la red de investigación. Por otro lado, también se encuentran los trabajos de grado que solo cuentan con la característica de ser una colaboración interdisciplinaria mas no contemplan existencia de un nodo externo de relacionamiento.

En las 3 comunidades con mayor tamaño encontradas, sobresalen distintas escuelas, siendo la de mayor dimensión Ingeniería Metalúrgica, seguido de Ingeniería de Sistemas y Geología, donde se observan colaboraciones atípicas entre Ingeniería de Sistemas con Ingeniería Química a través del líder de la respectiva comunidad, mientras que en los demás casos se realizaron colaboraciones entre escuelas de la misma facultad.

Con este modelo de investigación es posible explorar y expandir distintas áreas, que para futuras investigaciones sería interesante estudiar y conocer los comportamientos de la comunidad científica universitaria del país de los distintos grupos de investigación, así como es la participación nacional e internacional de eventos científicos, colaboración y temáticas sobre la creación de patentes, diseño de softwares, etc., pues podría otorgar patrones característicos para indagación sobre nuevos campos investigativos.

Referencias

[1] D. Easley, “Volumen 26 del TCE número 5 Portada y reverso | Teoría econométrica | Cambridge Core,” 2010. [Online]. Available: https://www.cambridge.org/core/journals/econometric-theory/article/ect-volume-26-issue-5-cover-and-back-matter/FA59F3056051664CB2969F82CC6CE0C8. [Accessed: 21-Nov-2019].

[2] Q. Wang, W. Li, X. Zhang, and S. Lu, “Academic Paper Recommendation Based on Community Detection in Citation-Collaboration Networks,” *APWeb*, vol. 2, pp. 56–67, 2016.

[3] Y. Yang, P. G. Sun, X. Hu, and Z. J. Li, “Closed walks for community detection,” *Phys. A Stat. Mech. its Appl.*, vol. 397, no. 37, pp. 129–143, 2014.

[4] P. Bedi and C. Sharma, “Community detection in social networks,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 6, no. 3, pp. 115–135, 2016.

[5] S. Kaur, S. Singh, S. Kaushal, and A. K. Sangaiah, “Comparative analysis of quality metrics for community detection in social networks using genetic algorithm,” *Neural Netw. World*, vol. 26, no. 6, pp. 625–641, 2016.

[6] N. Girdhar and K. K. Bharadwaj, “Community Detection in Signed Social Networks Using Multiobjective Genetic Algorithm,” *J. Assoc. Inf. Sci. Technol.*, vol. 70, no. 8, pp. 788–804, 2019.

[7] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 12, pp. 7821–7826, 2002.

[8] Z. Jiang, J. Liu, and S. Wang, “Traveling salesman problems with PageRank Distance on complex networks reveal community structure,” *Phys. A Stat. Mech. its Appl.*, vol. 463, pp. 293–302, 2016.

[9] M. Arasteh and S. Alizadeh, “A fast divisive community detection algorithm based on edge degree betweenness centrality,” *Appl. Intell.*, vol. 49, no. 2, pp. 689–702, 2019.

[10] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, “Community detection in networks: A multidisciplinary review,” *J. Netw. Comput. Appl.*, vol. 108, no. September 2017, pp. 87–111, 2018.

[11] B. A. Attea, W. A. Hariz, and M. F. Abdulhalim, “Improving the performance of evolutionary multi-objective co-clustering models for community detection in complex social networks,” *Swarm Evol. Comput.*, vol. 26, pp. 137–156, 2016.

[12] F. Folino and C. Pizzuti, “An evolutionary multiobjective approach for community discovery in dynamic networks,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1838–1852, 2014.

[13] K. R. Žalik and B. Žalik, “Multi-objective evolutionary algorithm using problem-specific genetic operators for community detection in networks,” *Neural Comput. Appl.*, vol. 30, no. 9, pp. 2907–2920, 2018.

[14] S. Mishra, C. Hota, L. Kumar, and A. Nayak, “An Evolutionary GA-Based Approach for Community Detection in IoT,” *IEEE Access*, vol. 7, pp. 100512–100534, 2019.

[15] M. E. J. Newman, "The structure of scientific collaboration networks", *Struct. Dyn. Networks*, vol. 9781400841, pp. 221–226, 2001.

[16] M. C. V. Nascimento, "Community detection in networks via a spectral heuristic based on the clustering coefficient", *Discret. Appl. Math.*, vol. 176, pp. 89–99, 2014.

[17] X. Deng, J. Zhai, T. Lv, and L. Yin, "Efficient Vector Influence Clustering Coefficient Based Directed Community Detection Method," vol. 5, 2017.

[18] J. Kim, J. G. Lee, and S. Lim, "Differential flattening: A novel framework for community detection in multi-layer graphs", *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, 2016.

[19] X. M. Li, G. Xu, and M. Tang, "Community detection for multi-layer social network based on local random walk", *J. Vis. Commun. Image Represent.*, vol. 57, pp. 91–98, 2018.

[20] A. Said, R. A. Abbasi, O. Maqbool, A. Daud, and N. R. Aljohani, "CC-GA: A clustering coefficient based genetic algorithm for detecting communities in social networks", *Appl. Soft Comput. J.*, vol. 63, pp. 59–70, 2018.

[21] M. E. J. Newman, "Modularity and community structure in networks", *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

D. N. Camelo García es Ingeniero Industrial de la Universidad Industrial de Santander, Colombia (2021). Sus intereses de investigación incluyen: Big Data; Machine Learning; Supply Chain Management.

P. C. Suárez Suárez es Ingeniera Industrial de la Universidad Industrial de Santander, Colombia (2021). Sus intereses de investigación incluyen: Big Data; Machine Learning; Innovación en industrial 4.0.

H. Lamos Díaz actualmente es profesor Titular, imparte asignaturas en área de Estadística y Teoría de Optimización en la Escuela de Estudios Industriales y Empresariales (EEIE) de la Universidad Industrial de Santander (UIS). Tiene el título de Matemático de la Universidad de la Amistad de los Pueblos de Rusia, Magíster en Matemáticas y en Informática y PhD en Física y Matemáticas. Su línea de investigación es en logística humanitaria y ciencia de datos.
ORCID: 0000-0003-1778-9768

D. E. Puentes Garzón, Ingeniero Industrial (2016), Magíster en Ingeniería Industrial (2019) y estudiante de doctorado en ciencias de la computación (2021) en la Universidad Industrial de Santander. Sus intereses de investigación incluyen: analítica de datos; aprendizaje automático; inteligencia artificial y optimización.
ORCID: 0000-0001-8178-2339