Un modelo de autómatas de aprendizaje celulares para el problema de detección de comunidades en una red de investigadores en Colombia

Ana María Barajas Otálora, Laura Jimena Castiblanco Ramírez

Trabajo de investigación para optar el título de Ingeniera Industrial

Director

Henry Lamos Díaz

Doctor en Física- Matemática

Codirector

Hugo Ernesto Martínez Ardila

Doctor en Ingeniería

Universidad Industrial de Santander

Facultad de Ingenierías Físico mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2020

Dedicatoria

A mi madre y a mi padre, quienes lucharon para proporcionarme la maravillosa oportunidad de estudiar, a quienes debo todo lo que soy y a quienes procuraré llenar de orgullo el resto de mi vida.

A mi abuelo, quien siempre creyó en mí y quien con confianza y orgullo me llamaba "ingenierita" desde mi primer semestre.

A mi hermana, a quien siempre he querido darle el mejor ejemplo y quien ha sido mi mejor amiga siempre.

A Andrés, quien con su amor me ha acompañado y apoyado en los momentos más difíciles.

Ana María Barajas Otálora

Dedicatoria

A Dios por guiarme en cada momento de la vida y darme la sabiduría para afrontar las oportunidades y los obstáculos que se me han presentado.

A mis padres Emma y Jorge quienes me dieron la vida y han sido un ejemplo a seguir, por sus enseñanzas, su confianza, su cariño y por ser siempre mi mayor motivación.

A mis hermanas Katha y Camila por apoyarme incondicionalmente en cada paso que doy y quienes comparten con alegría y orgullo cada logro de mi vida.

A mi abuelita Margarita y mis tíos maternos por confiar en mí y apoyarme especialmente en los momentos más difíciles.

A mis amigos Miguel, Johana, María Paula y Oscar por animarme, acompañarme y aconsejarme durante todo el proceso.

Laura Jimena Castiblanco

Agradecimientos

A nuestro director Henry Lamos Díaz, por creer en nuestras capacidades para realizar este proyecto, por apoyarnos y aconsejarnos durante las dificultades y por inspirarnos a desafiar nuestras capacidades y dar nuestro mejor esfuerzo.

Al grupo Ópalo, quienes más que un grupo son una familia, por escucharnos, aconsejarnos y animarnos en cada paso del camino, por transmitirnos su pasión por la investigación y por ser más que compañeros unos grandes amigos.

A nuestras familias por ser nuestro principal apoyo, inspiración y motivación para alcanzar la excelencia y ser grandes profesionales.

A nuestra alma máter, el lugar que por años nos vio crecer, nos llenó de oportunidades para triunfar y ser mejores y cuyo nombre prometemos llevar siempre con orgullo.

Contenido

Lista de tablas

Lista de figuras

Lista de apéndices

(Ver apéndices adjuntos en el CD o en la Base de Datos de la Biblioteca UIS)

Resumen

Título: Un modelo de autómatas de aprendizaje celulares para el problema de detección de comunidades en una red de investigadores en Colombia[*]

Autores: Ana María Barajas Otálora, Laura Jimena Castiblanco Ramírez[**]

Palabras Clave: Detección comunidades, Análisis redes sociales, Autómatas aprendizaje celular, Aprendizaje reforzado, Colaboración científica.

Descripción:

En pleno auge de la cuarta revolución industrial los países se enfrentan al desafío de fortalecer la ciencia y tecnología para ser sociedades competitivas capaces de adaptarse a los cambios que atraviesa la industria y sociedad. Con el fin de fortalecer la ciencia a través de la colaboración científica, ha crecido el interés dentro del área de Análisis de Redes Sociales (ARS) por estudiar las redes de investigación.

El presente trabajo de investigación se centra en dar solución al problema de detección de comunidades en una red de investigación en el área de Ingeniería Industrial en Colombia. La detección de comunidades es un tema que ha ganado relevancia en los últimos años, ya que permite encontrar los subgrupos que se forman dentro de una red social y así extraer patrones útiles que permitan entender cómo evolucionan los individuos y las comunidades que forman.

La detección de comunidades en la red social de investigación se realiza mediante la implementación de un algoritmo basado en un modelo de autómatas de aprendizaje celular (CLA-Net). Una vez revelada la estructura de comunidad de la red se aplica la teoría de grafos para realizar la validación de las comunidades encontradas y para analizar otras propiedades de la red social y sus patrones de colaboración científica. Esto permitirá conocer el estado de la colaboración científica en la red de investigación, los investigadores más influyentes dentro de cada comunidad y dentro la red, los investigadores que actúan como puentes y permiten el flujo de información y conocimiento, entre otras propiedades interesantes de la red.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico-mecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Díaz. Codirector: Hugo Ernesto Martínez Ardila.

Abstract

Title: A cellular learning automata-based algorithm for community detection in a research network in Colombia*

Authors: Ana Maria Barajas Otálora, Laura Jimena Castiblanco Ramírez**

Keywords: Community detection, Social Network Analysis, Cellular learning automata, Reinforcement learning, Scientific collaboration.

Description:

Given the global hype of the fourth industrial revolution, countries face the challenge to strengthen science and technology to be more competitive and capable of adapting to the changes that the industry and society are going through. In order to strengthen science through scientific collaboration, the area of Social Network Analysis (SNA) has drawn its attention to research networks, since the study of those networks in particular leads to a better understanding of scientific collaboration.

This research work focuses on solving the community detection problem for an Industrial Engineering research network in Colombia. Community detection is a topic that has gained attention in recent years, since it allows to discover the clusters within a social network and thus extract useful patterns to understand how individuals within communities evolve and make relationships.

Community detection in the research network is carried out by implementing a cellular learning automata-based algorithm (CLA-Net). Once the community structure of the network has been revealed, graph theory is used to analyze the obtained communities and other properties of the social network as well as its patterns of scientific collaboration. This analysis provides some useful insights into the state of scientific collaboration of the research network, the most influential researchers within each community, researchers who act as bridges and allow the flow of information and knowledge, among other interesting features of the network.

* Bachelor's thesis

** Physical-mechanical Engineering Faculty. School of Industrial and Business Studies. Supervisor: Henry Lamos Díaz. Co-supervisor: Hugo Ernesto Martínez Ardila.

Introducción

El conocimiento es el bien más valioso de la sociedad, pues es fuente de desarrollo y progreso para la sociedad, en la actualidad con la gran cantidad de datos que se genera a diario el verdadero reto consiste en concentrarse en los datos relevantes y generar conocimiento útil a partir de estos.

La comunidad científica como principal generadora de conocimiento ha mostrado una mayor preocupación por entender cómo lograr una mayor difusión de su conocimiento y cómo aumentar su productividad científica, pues la investigación requiere grandes esfuerzos y en muchas ocasiones no genera el impacto esperado. Esta preocupación ha motivado el trabajo de diferentes autores en el área de las redes sociales.

El análisis de redes sociales (ARS) es un área de estudio que ha ganado suma importancia en los últimos años pues genera conocimiento sobre la forma en que las personas se relacionan y cómo fluye a través de ellas la información, esto resulta especialmente relevante para distintas áreas de aplicación como el marketing, las finanzas, la sociología, la psicología, entre otras. Sin embargo, en años recientes la preocupación se ha enfocado también en un problema en particular, la detección de comunidades. Este problema se centra en estudiar la estructura de las comunidades al interior de las redes sociales, pues una de sus propiedades es que comúnmente se tiende a tener asociaciones más fuertes con otros individuos con quienes existe alguna característica en común. El estudio de estos subgrupos con relaciones más fuertes permite optimizar diferentes características en las redes, encontrando aquellos individuos influenciadores, aquellos que conectan comunidades y otros que cumplen funciones de control.

Para la comunidad científica es de interés conocer cómo se comportan las redes de colaboración científica, cómo se asocian internamente sus comunidades y cómo encontrar aquellos investigadores que logran una mayor difusión del conocimiento o que son portadores de él. Trabajos recientes se han realizado al respecto con diferentes redes científicas evidenciando así un creciente interés por entender mejor las redes de investigación (Yudhoatmojo & Samuar, 2017., Ahmed et al., 2018., Horta et al. 2018).

En este trabajo se pretende abordar el problema de detección de comunidades en una red de investigadores de Colombia utilizando un modelo de autómatas de aprendizaje celular, pues ha mostrado superioridad en su eficiencia en comparación a otros modelos de la literatura, gracias a la incorporación de elementos de aprendizaje reforzado e interacción celular que prometen mejores resultados (Zhao et al., 2015). Además, se busca analizar las propiedades de la red social y los patrones de colaboración científica en la red.

Tabla 1.

Cumplimiento de los objetivos del proyecto

Objetivo	**Cumplimiento**
Realizar una revisión de la literatura sobre el problema de detección de comunidades en redes sociales y los modelos utilizados.	Capítulo 2
Construir el isomorfismo entre la red de investigadores y el modelo de autómatas de aprendizaje celular (CLA).	Capítulo 4
Modelar un modelo de autómatas de aprendizaje celulares a partir del grafo que representa la red.	Capítulo 4
Validar el modelo por medio de simulación computacional en una red de investigadores colombianos de acuerdo a sus co-publicaciones.	Capítulo 5, 6
Elaborar un artículo de carácter publicable sobre los resultados obtenidos en la investigación.	Apéndice G

1. Generalidades del proyecto

1.1 Planteamiento del problema

En la actualidad la información y el conocimiento se han convertido en elementos clave para lograr competitividad y desarrollo en la sociedad. El conocimiento permite mejorar considerablemente la toma de decisiones reduciendo la incertidumbre y proveyendo a individuos y organizaciones de una capacidad de respuesta más ágil ante los cambios. La comunidad científica, como principal generadora de conocimiento enfrenta entonces un gran reto para lograr una mayor productividad científica, especialmente universidades y otras instituciones de países en desarrollo, cuyas capacidades de investigación son más bajas y cuya generación de conocimiento es vital para dar respuesta a los desafíos y necesidades de desarrollo de la sociedad.

La difusión del conocimiento se ha convertido en un tema importante para la comunidad científica pues es un elemento clave para hacer que el conocimiento generado logre mayor visibilidad en la comunidad científica y consecuentemente un mayor impacto de desarrollo para la sociedad. Se ha encontrado que existe correlación entre la colaboración científica y el desempeño y productividad científica de las instituciones e investigadores (He et al., 2009, Lee y Bozeman, 2005). La colaboración científica aumenta la difusión y visibilidad de los trabajos de investigación y permite a investigadores o centros de investigación más incipientes fortalecerse cuando colaboran con investigadores o centros de investigación más influyentes (Aldieri et al., 2018). Cuando los investigadores no poseen todas las competencias para llevar a cabo una investigación, la colaboración aparece como una solución para asociar investigadores con diferentes habilidades que se integran para lograr grandes avances científicos (Beaver, 2001). Además, permite agregar originalidad a los trabajos ya que se pueden incluir autores de diferentes disciplinas que agregan

perspectivas diferentes a la investigación (Katz y Martin, 1997, Rigby y Edler, 2005). De ahí que en la actualidad se esté estudiando las comunidades al interior de las redes científicas para comprender mejor su funcionamiento e identificar la forma para mejorar el flujo de conocimiento a través de ellas.

Las redes de investigación atraviesan grandes cambios debido a la necesidad de realizar colaboración científica globalizada para poder dar respuesta a los problemas que enfrenta el mundo, pues son cada vez más grandes y complejos. Así mismo surge la necesidad de fortalecer la colaboración a nivel nacional, regional y local para responder a los desafíos que tiene cada lugar en particular. Es así como el Análisis de Redes Sociales (ARS) se convierte en una buena alternativa para estudiar las redes de colaboración científica, ya que estas pueden ser representadas a través de nodos y enlaces, aplicando la teoría de grafos, con el fin de lograr la comprensión de la estructura de las comunidades dentro de la red y sus propiedades, pues conocer su estructura permite extraer patrones útiles para entender la dinámica y funcionamiento de un grupo de individuos o instituciones. Para esto es necesario identificar la forma en la que se asocian los individuos dentro de la red, ya que generalmente hay individuos más interconectados entre ellos (con relaciones más fuertes) y a su vez están escasamente conectados a otros individuos, este fenómeno se conoce como “comunidad”.

El proceso de encontrar cómo se conforman las comunidades se denomina “Detección de comunidades”. Este problema dentro del Análisis de Redes Sociales ha sido abordado con algoritmos clásicos como el Genético (Pizzuti, 2011), Aglomerativo (Rahman, 2013), Colonia de hormigas (Ji et al., 2013), entre otros. Sin embargo, el mundo genera redes cada vez más grandes

y complejas, lo que ha motivado la aparición de nuevos métodos de solución con mayor capacidad y eficiencia que incorporan elementos de aprendizaje automático e inteligencia artificial.

1.2 Justificación

La colaboración y productividad científica en Colombia se han convertido en temas relevantes en los últimos años debido a la necesidad de evaluar la capacidad que tienen las personas para generar conocimiento e innovación como país, pues esta capacidad determina la competitividad del país frente a los grandes cambios sociales e industriales que atraviesa el mundo. Son grandes, pero no suficientes los avances en temas de colaboración y productividad científica, pues aunque ha aumentado la colaboración a nivel internacional y el volumen de publicaciones anuales en revistas (Maz, Jiménez y Villarraga, 2016), la investigación en el país sigue siendo baja en comparación con otros países líderes, por esto se siguen aumentando los esfuerzos desde la política pública para incentivar la investigación en el país.

En pleno desarrollo de la cuarta revolución industrial, la ciencia y tecnología son dos pilares necesarios para hacer de Colombia un país competitivo. Por lo tanto, el fortalecimiento de la ciencia y tecnología se ha convertido en un objetivo de la política pública. El ministerio de Ciencia y Tecnología de Colombia declara como su misión la labor de articular la política pública para la generación de conocimiento, innovación y competitividad (Minciencias, s.f.).

Con el fin de evaluar el impacto de las estrategias para incentivar la investigación en el país, es necesario conocer el estado de la colaboración y productividad científica en las redes de investigación en Colombia, además de estudiar las propiedades y principales patrones de estas, todo esto es posible a través del análisis de redes sociales ya que dicho análisis permite entender la

evolución de las redes de investigación y determinar la efectividad de las políticas públicas que buscan fortalecer la ciencia en el país.

El presente trabajo busca analizar una red de investigadores en el área de Ingeniería Industrial en Colombia usando un algoritmo basado en un modelo de autómatas de aprendizaje celular para estudiar la estructura de comunidad de la red y aplicando herramientas de la teoría de grafos para hacer el análisis de la red social, esto con el fin de conocer los subgrupos existentes dentro de la red, así como los patrones de colaboración científica y otras propiedades que permitan entender las dinámicas de colaboración científica y las debilidades y desafíos que enfrentan las comunidades de investigación en Ingeniería Industrial en el país. Dicho análisis permitirá hacerse una idea del estado actual de la colaboración y productividad científica en el área de Ingeniería Industrial y podría ser utilizado como referencia para identificar autores e instituciones que trabajan en las mismas líneas de investigación y con quienes se podrían desarrollar futuras colaboraciones.

1.3 Objetivos

1.3.1 Objetivo general

Implementar un modelo de autómatas de aprendizaje celular para el problema de detección de comunidades en una red de investigadores en Colombia.

1.3.2 Objetivos específicos

- Realizar una revisión de la literatura sobre el problema de detección de comunidades en redes sociales y los modelos utilizados.
- Construir el isomorfismo entre la red de investigadores y el modelo de autómatas de aprendizaje celular (CLA).

- Modelar unos autómatas de aprendizaje celulares a partir del grafo que representa la red.
- Validar el modelo por medio de simulación computacional en una red de investigadores colombianos de acuerdo a sus co-publicaciones.
- Elaborar un artículo de carácter publicable sobre los resultados obtenidos en la investigación.

1.4 Metodología

Para el desarrollo de esta investigación se hizo uso de la metodología KDD propuesta por Fayyad et al. (1996), la cual consiste en descubrir conocimiento a partir de base de datos y se ejecuta en siete fases principales como se puede ver en la siguiente figura:

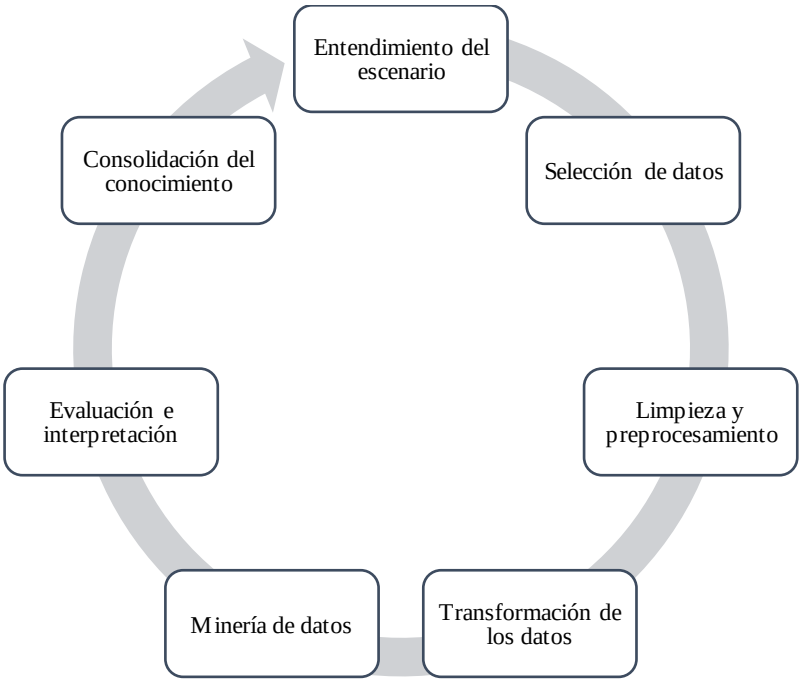

Figura 1. Principales fases de la metodología KDD.

Como se observa en la figura 1, el desarrollo de esta investigación se fundamenta en la metodología KDD, la cual está compuesta por las fases de Entendimiento del escenario, Selección

de datos, Limpieza y preprocesamiento, Transformación de los datos, Minería de datos, Evaluación e interpretación y finalmente la fase de Consolidación del conocimiento que se desarrolla transversalmente a la ejecución de las fases anteriores.

Inicialmente, en los Capítulos 2 y 3 se lleva a cabo la primera fase –*Entendimiento del escenario*- en la cual se hace una revisión Bibliográfica que está compuesta por un análisis bibliométrico y preliminar para comprender el estado y evolución de la literatura existente sobre el problema de detección de comunidades y la teoría necesaria para comprender el análisis de redes sociales.

En el Capítulo 4 inicialmente se desarrollan las fases de –*Selección, Limpieza y preprocesamiento y transformación de los datos*- en las que se hace uso del formato de hojas de vida electrónica de Colciencias- CvLAC para recopilar información de los autores de artículos publicados en el área de Ingeniería Industrial. Una vez obtenida dicha información, se procede a realizar la construcción de la red de investigadores en Colombia.

En el capítulo 4 también se lleva a cabo el proceso de selección y diseño del algoritmo el cual hace parte de la fase de – M*inería de datos* – en éste se presenta el algoritmo basado en un modelo de autómatas de Aprendizaje Celular. El proceso de implementación que también corresponde a la fase de Minería de datos se presenta en el capítulo 5 en el cual se implementa el algoritmo CLA-Net en la red de investigadores construida para el desarrollo de esta investigación.

En el Capítulo 6 se realiza la fase de –*Evaluación e interpretación*- en ésta se hace un análisis cualitativo y cuantitativo de la red social propuesta.

2. Revisión de la literatura

Con el propósito de comprender el estado y evolución para el problema de detección de comunidades en el análisis de redes complejas, se realiza una revisión de la literatura comprendida por dos fases.

Fase 1. Se establece una primera ecuación de búsqueda utilizando las palabras clave más relevantes del artículo "A cellular learning automata-based algorithm for detecting community structure in complex networks" propuesto por Zhao et al., ya que es el artículo base para realizar esta investigación. Además, se tiene en cuenta que el fin de este trabajo es detectar comunidades en una red de investigadores, por lo cual se habla de una red de tipo social. La búsqueda se realiza en tres de las bases de datos disponibles en la Universidad Industrial de Santander-UIS (Scopus, Web of Science y Science Direct), para así obtener una idea de la literatura existente. En la tabla 2 se presenta la ecuación utilizada en cada una de las bases de datos.

Tabla 2.

Ecuaciones de búsqueda Fase 1

Base de Datos	**Ecuación de Búsqueda**
Web of Science	TEMA: ("community detection" AND "social network")
Scopus	TITLE-ABS-KEY ("community detection" AND "social network")
Science Direct	Title, abstract, keywords: ("community detection" AND "social network")

La búsqueda arroja 366, 2132 y 226 resultados en las bases de datos Web of Science, Scopus y Science Direct respectivamente; sin embargo, se observa que se relacionan una gran cantidad de

artículos que no aportan a la comprensión del problema en el contexto de esta investigación por lo que se plantea una nueva ecuación.

Fase 2. Se realizan dos cambios, el primero consiste en incluir el término "algorithm" a la frase "community detection", ya que los artículos obtenidos en la Fase 1 mencionan el problema, pero no se enfocan en las técnicas utilizadas para dar solución al mismo. Por otra parte, se establece que la búsqueda respecto a las redes sociales incluya el término "analysis" ya que "Social Network Analysis" (SNA) es el estudio de las estructuras sociales por medio de redes y teoría de grafos, tema que abarca el problema de detección de comunidades y permite incluir resultados que estudian la detección de comunidades en el mismo contexto de esta investigación. En la Tabla 3 se observa la ecuación con los ajustes realizados.

Tabla 3.

Ecuaciones de búsqueda Fase 2

Base de Datos	**Ecuación de Búsqueda**
Web of Science	TEMA ((("community detection algorithm") AND ("social network analysis")))
Scopus	TITLE-ABS-KEY ((("community detection algorithm") AND ("social network analysis")))
Science Direct	Title, abstract, keywords(("community detection algorithm") AND ("social network analysis"))

Los resultados obtenidos de la anterior ecuación de búsqueda son: Web of Science 15, Scopus 114 y Science Direct 22.

Fase 2.1. Debido a la colección multidisciplinar que engloban las bases de datos, los resultados obtenidos contienen aplicaciones del problema en otras disciplinas que no resultan relevantes para

la presente investigación. Por este motivo se decide refinar los resultados aplicando un filtro en las revistas, en el cual se excluyen aquellas revistas cuyas disciplinas o ejes temáticos no tienen relación con el campo de análisis de redes sociales que se pretende estudiar en esta investigación. Además, se filtró por tipo de documento para encontrar únicamente artículos. En la tabla 4 se observan los resultados obtenidos y las revistas incluidas en la búsqueda que se realizó en cada una de las bases de datos.

Tabla 4.

Revistas incluidas para refinar la búsqueda

Base de datos	Revistas incluidas	Resultados
Web of Science	Computer Science Artificial Intelligence, Mathematics Interdisciplinary Applications, Computer Science Information Systems, Engineering Industrial, Operations Research Management Science, Computer Science Theory Methods, Physics Mathematical, Computer Science Interdisciplinary Applications, Physics Multidisciplinary, Behavioral Sciences, Information Science Library Science.	10
Scopus	Computer Science, Mathematics, Engineering, Social Sciences, Decision Sciences, Business, Management and Accounting, Arts and Humanities, Economics, Econometrics and Finance.	41
Science Direct	Future Generation Computer Systems, Information Sciences, Procedia Computer Science, Computers & Industrial Engineering, Knowledge-Based Systems Data & Knowledge Engineering, Social Networks	13

Los sesenta y cuatro (64) artículos obtenidos, son seleccionados para el análisis bibliométrico utilizando el software Vantage Point. Por medio de este se pretende organizar mejor

los datos de los artículos seleccionados, identificar duplicados en las bases de datos y encontrar tendencias que permitan justificar la importancia del tema de investigación, por lo cual se procede a hacer el análisis bibliométrico.

2.1 Análisis bibliométrico

Al procesar los metadatos para los sesenta y cuatro artículos en el software Vantage Point, se encuentra que existen algunos duplicados en las bases de datos, por lo tanto, el número de artículos se redujo a cuarenta y siete.

2.1.1 Año de publicación. Inicialmente se analiza la cantidad de publicaciones por año para el tema de investigación como se observa en la figura 2.

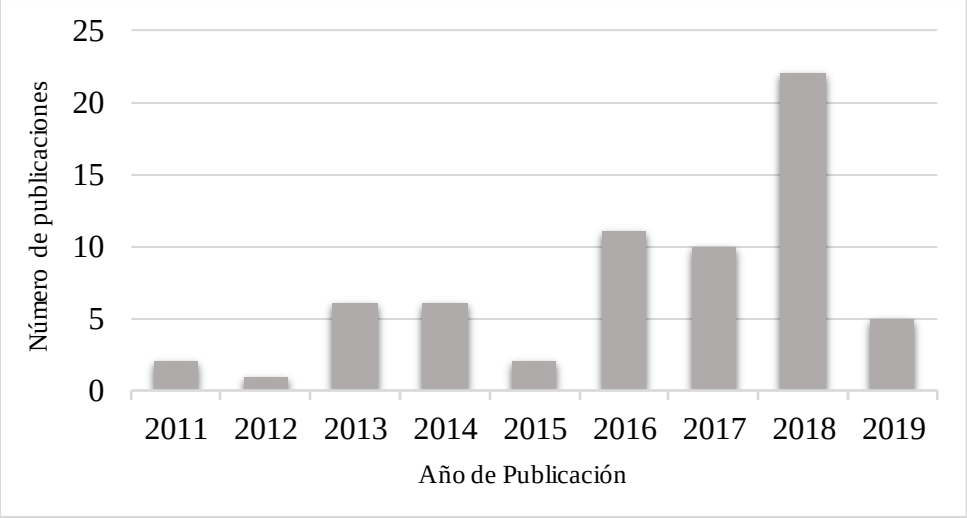

Figura 2. Publicaciones por año.

En la figura 2 se observa que en los años 2016, 2017 y 2018 se realizaron once, diez y veintidós publicaciones respectivamente, lo cual es significativo respecto al año 2011 en el cual solo se realizaron dos publicaciones. En el año 2018 se presenta el mayor número de publicaciones hasta

el momento lo cual indica que el tema está tomando mayor importancia y que resulta relevante investigar sobre él.

2.1.2 País de publicación. Se hace un análisis de los países en los que se ha investigado este tema con el fin de tener mayor idea del contexto en el que surge la necesidad e interés de investigar el mismo, además se pretende conocer los países que tienen mayor avance en esta área de investigación.

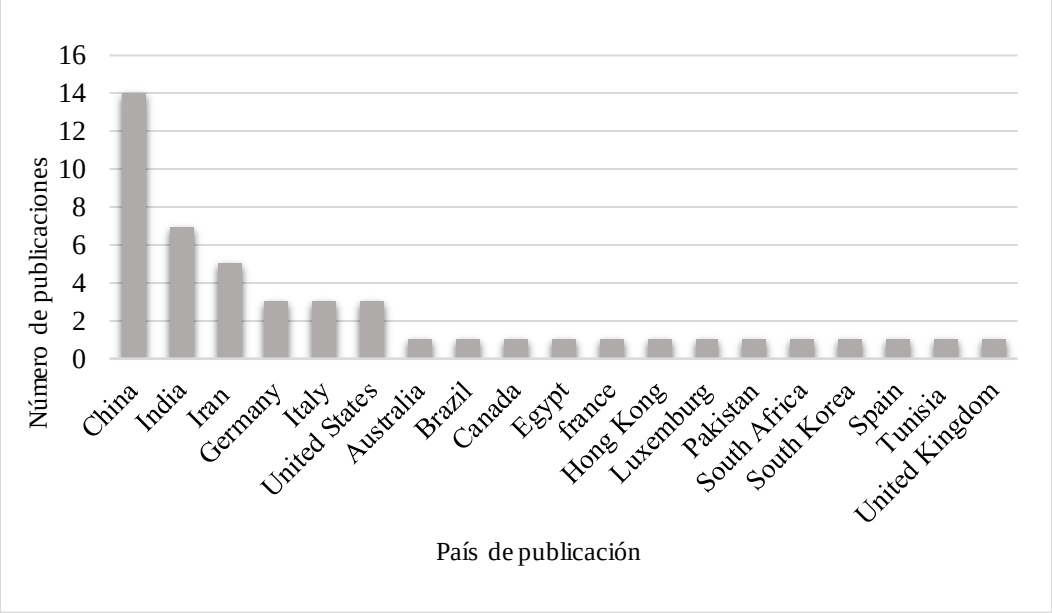

Figura 3. Publicaciones por país.

Los resultados de la figura 3 evidencian que de los diecinueve países donde se han realizado publicaciones, China, India e Irán representan el 54% de las publicaciones hechas. Lo cual indica que en esta zona geográfica se concentran la mayoría de los avances logrados en este campo.

2.1.3 Área de investigación. Resulta de gran interés conocer las diversas áreas de investigación y aplicaciones que se desarrollan a partir del problema de detección de comunidades.

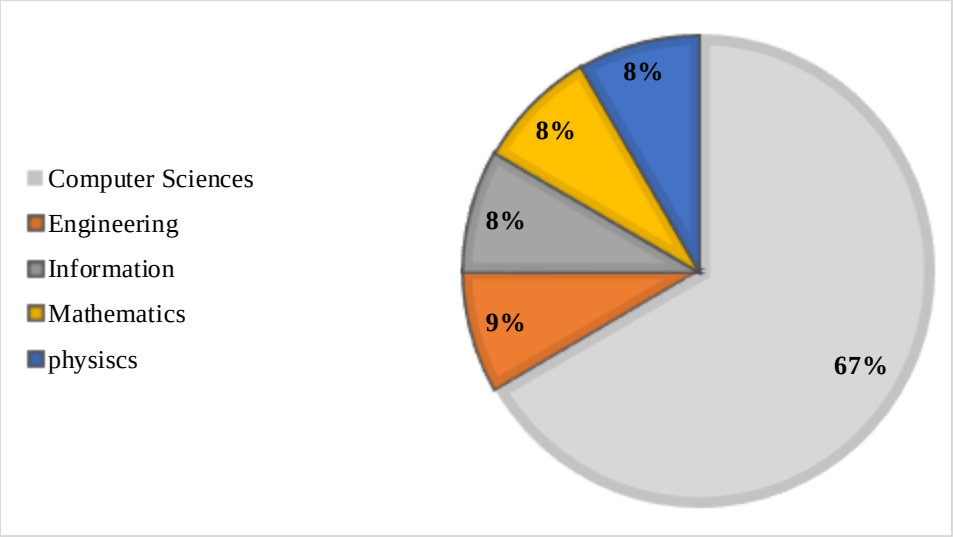

Figura 4. Porcentaje de participación por área de investigación.

Se puede observar en la figura 4 que el área de investigación que tiene mayor participación en este tema de estudio es Ciencias de la Computación, lo cual se debe a que las redes dentro de esta área son un concepto fundamental para entender la interacción de los sistemas. Es por esto que resulta de gran utilidad estudiar la estructura de las redes a través de diferentes enfoques como la detección de comunidades.

2.1.4 Palabras clave. Por último, se hace una revisión de las palabras clave más frecuentes incluidas en los artículos con el fin de ver la relevancia de los términos utilizados en la ecuación de búsqueda y revisar si se excluyó algún término que también sea relevante para la investigación. En la figura 5 se observan los términos más mencionados en los artículos seleccionados.

La nube de palabras muestra que las dos palabras más frecuentes son las mismas utilizadas en la ecuación de búsqueda, adicionalmente surge el término "Complex networks", el cual no se tuvo

en cuenta para formular la ecuación dado que hace referencia al tamaño de la red y no a las características que se pretenden identificar mediante la detección de comunidades.

Figura 5. Nube de palabras clave usadas por los autores. Adaptado de Vantage Point (2019).

Con el anterior análisis se verifica la idoneidad de la ecuación de búsqueda y la relevancia que tiene este tema de investigación. Finalmente se lee el resumen de los cuarenta y siete (47) artículos obtenidos con la ecuación de búsqueda, de los cuales se seleccionan dieciocho artículos para realizar el análisis preliminar de la literatura, ya que abordan el tema de detección de comunidades en redes sociales desde diversos enfoques y para diferentes contextos como el de este trabajo, "Redes Científicas". Además, se añaden otros trabajos que se encuentran como referencia en el artículo base de la investigación y que se consideran relevantes para entender la evolución del estudio del problema de detección de comunidades.

2.2 Análisis preliminar de la literatura

La representación de diferentes redes del mundo real en las áreas de la Biología, Sociología, Tecnología e Información ha sido realizada con el uso de grafos que son con un conjunto de nodos interconectados a través de enlaces. Paralelamente este tipo de representación se ha extendido a

otras áreas de estudios más recientes como el análisis de redes sociales. Según Wasserman & Faust (1994) una "red social" es un conjunto de actores interconectados por relaciones que han sido creadas a través de interacciones. Estas redes han sido objeto de estudio en las ciencias sociales por más de 50 años debido al interés de identificar patrones de interacción humana además de obtener información útil y características importantes para entender la organización y funcionamiento de la sociedad, pueden ser modeladas utilizando grafos en donde cada entidad se representa por un nodo y la interacción (relación) entre las entidades se representa por un enlace (Fortunato, 2010). En una red social, las personas se dividen en grupos de acuerdo a líneas de interés, ocupación, edad, entre otras, es por esto que Newman & Girvan (2004) denomina como "comunidad" a un grupo de vértices (nodos) que tienen una alta densidad de arcos (enlaces) con otros vértices, pero que existe una baja densidad de arcos entre grupos de vértices. Según Fortunato (2010), las comunidades, también llamadas clústeres o módulos, comparten propiedades comunes o juegan roles similares dentro de un grafo.

Los primeros estudios sobre las propiedades de las redes sociales estuvieron enfocados a propiedades como "Small-world", que dice que incluso si dos personas tienen una baja probabilidad de conocerse, la probabilidad de que tengan un conocido en común es alta, esto es lo que informalmente se conoce como "El mundo es un pañuelo". El primero en llevar a cabo experimentos para estudiar este fenómeno fue un psicólogo de la universidad de Stanford. Stanley Milgram (1967) en sus experimentos hizo un descubrimiento sobre la existencia de cierta estructura matemática en la sociedad por lo que surgió una pregunta interesante, dada una persona X y una persona Y escogidas aleatoriamente, ¿es posible saber a cuántos conocidos está la una de la otra?, en sus experimentos encontró que estamos a seis conexiones (círculos) de cualquier otra persona.

Posteriormente se acuñó el término de "Six-degree separation" a este hallazgo y comenzó a estudiarse como una propiedad de grafos (Watts & Strogatz, 1998).

Así mismo, se ha estudiado la propiedad de "Power-law degree distributions" que afirma que existe una distribución de los grados de un grafo que obedece a una ley de potencia y no sucede aleatoriamente como otros pensaban (Faloutsos et al., 1999) y la propiedad "Network transitivity" que en términos de redes sociales se puede explicar como "El amigo de tu amigo es tu amigo", pues se refiere a que en una tupla de nodos, si el nodo i está conectado a j y j está conectado a k, entonces i está conectado a k (Wasserman & Faust, 1994). Sin embargo, en los últimos años se incrementó el interés por estudiar otra propiedad de las comunidades conocida como "estructura de comunidad" la cual consiste en entender cómo están conformadas las comunidades al interior de una red social. Surge entonces el problema de detección de comunidades en el análisis de redes sociales.

La meta de la detección de comunidades es encontrar la mejor partición posible (Comunidades) de los nodos que pertenecen a una red. Además, es de gran importancia identificar la posición estructural de los nodos dentro de las comunidades, de tal forma que los nodos que se encuentran en la posición central de cada comunidad tengan una importante función de control y estabilidad dentro del grupo, mientras que los nodos que permanecen en las fronteras de las comunidades juegan un rol de mediación que permite mantener relaciones e intercambios entre diferentes comunidades (Fortunato, 2010).

Diversos tipos de algoritmos se han propuesto a lo largo del tiempo con el fin de identificar las comunidades presentes en una red. Girvan & Newman (2002) proponen un enfoque para detección de comunidades basados en "Edge betweenness" una medida de centralidad e influencia de los

nodos en las redes. El algoritmo consiste en revelar la estructura de comunidad al remover los enlaces que conectan las comunidades. Sin embargo, no es adecuado usar estos algoritmos en situaciones donde no se conocen las comunidades de antemano, por esta razón surge la necesidad de crear una medida de bondad de las comunidades encontradas por los algoritmos de manera que se garantice la calidad de las particiones encontradas. La función de calidad más popular adoptada en la literatura es la "Modularidad" (Fortunato, 2010).

La modularidad se define como una medida de calidad de las particiones encontradas en una red, se calcula como la diferencia entre la fracción de enlaces en la red que conecta nodos dentro de la misma comunidad y el valor esperado de la misma fracción con igual cantidad de comunidades, pero conexiones aleatorias entre los nodos. Un mayor valor de Q indica una mejor partición de las comunidades encontradas (Girvan & Newman, 2002).

Desde que se introdujo la función de modularidad se ha propuesto una gran variedad de algoritmos que pretenden dar solución al problema de detección de comunidades como un problema de optimización de la función de modularidad. Sin embargo, este ha mostrado ser un problema NP-Hard debido a la complejidad y gran tamaño de las redes actuales (Zhao et al., 2015), por lo cual los autores se han centrado en proponer algoritmos que den como resultado conjuntos de comunidades con altos valores de modularidad mientras mejoran la eficiencia y costo computacional.

Dado que el algoritmo propuesto por Girvan & Newman (2002) presenta una gran desventaja por la alta demanda computacional al calcular el "Edge betweenness", Newman (2004) propone el algoritmo "Fast algorithm for detecting community structure in networks" de tipo aglomerativo basado en la modularidad, el cual inicialmente asigna un nodo a cada comunidad y progresivamente

une parejas de nodos para formar comunidades. La función de modularidad se calcula a medida que se van formando las comunidades, lo cual es más sencillo y permite encontrar la estructura óptima de la comunidad.

Un algoritmo divisivo basado en optimización extrema es propuesto para optimizar la modularidad, en este se busca optimizar una función global (modularidad) mejorando una variable local extrema que mide las contribuciones de los nodos a la función de modularidad, el procedimiento consiste en dividir aleatoriamente en dos particiones la red y calcular la modularidad, en el siguiente paso se intercambian entre las dos el nodo con el valor más bajo de variable local extrema y se recalcula la modularidad. En el siguiente paso se realiza el mismo procedimiento, pero al interior de cada partición hasta obtener un valor máximo de modularidad (Duch, J. & Arenas, A., 2005).

Un algoritmo de propagación rápida de etiquetas también fue empleado para resolver este problema, en donde cada nodo tiene asignada una etiqueta al inicio del algoritmo y conforme avanza los nodos van adoptando las etiquetas de sus vecinos hasta que aquellos nodos que están más densamente conectados logran consenso en una etiqueta. Este algoritmo tiene grandes ventajas ya que no requiere una función predefinida para optimizar ni información previa sobre las comunidades, evitando el problema de límite de resolución y obteniendo resultados rápidamente en comparación con otros algoritmos de la literatura (Raghavan et al., 2007).

También los algoritmos genéticos han sido utilizados para abordar este tipo de problema. Tasgin et al. (2007) proponen un algoritmo genético para detectar comunidades optimizando la modularidad. Este algoritmo consiste en generar un conjunto de cromosomas candidatos que corresponden a posibles soluciones del problema, estos candidatos son sometidos a variaciones con

el uso de operadores inspirados en la genética como son los de cruce (combinación de segmentos de dos soluciones candidatas) y los de mutación (cambiar aleatoriamente un segmento de la solución). Los candidatos obtenidos en cada generación son evaluados con una función de aptitud que en este caso corresponde a la modularidad. La población inicial comúnmente se genera de forma aleatoria; sin embargo, en este caso se emplea una heurística simple de manera que la solución converja más rápido. A partir de la población inicial se calcula la función de aptitud, se organizan de mayor a menor las soluciones y se guardan los candidatos en las primeras B posiciones. Luego, se genera cruces de estos, se aplica mutaciones a algunos de los candidatos generando una nueva población, se guarda esta nueva población junto con la guardada previamente y se repite este ciclo para la siguiente generación por un número de veces que es establecido arbitrariamente.

El uso de la modularidad como una función de optimización trajo consigo una gran debilidad en el proceso de detección de comunidades, pues produjo lo que se conoce en la literatura como "problema del límite de resolución". Este problema consiste en la incapacidad de los algoritmos para detectar las comunidades inferiores a un determinado tamaño con el valor máximo de la modularidad, ya que se generan agrupamientos falsos compuestos de comunidades verdaderas que por su tamaño pequeño no se logran detectar (Fortunato & Barthélemy, 2007). Después de que se conoció este problema, algunos autores comenzaron a proponer alternativas que consisten en plantear el problema de optimización como un problema de optimización multi-objetivo o considerar nuevas funciones de calidad para superar el problema de límite de resolución. Es por esto que Clara Pizzuti (2008) propone un algoritmo genético llamado GA-Net, el cual usa la representación de adyacencia basada en la localización. Este algoritmo consiste en un conjunto de genes que de acuerdo a unos operadores de variación (Mutación o cruce uniforme), pueden tomar

valores de alelos y mediante un proceso de decodificación se identifica los nodos que pertenecen a un mismo componente para poderlos asignar a una comunidad. Para llevar a cabo este procedimiento, la autora incluye una nueva métrica denominada "community score" la cual proporciona una medida global de la partición de la red en comunidades. Además, se introduce la noción de "individuo seguro" la cual evita que los operadores de variación y el algoritmo genético sean actualizados innecesariamente generando un alto costo computacional.

Más adelante Clara Pizzuti (2011), propone el algoritmo MOGA-Net, un algoritmo multi-objetivo mediante el uso de la teoría de optimización de Pareto, la cual logra soluciones óptimas además de permitir la identificación de propiedades de las comunidades en varios niveles de resolución. El primer objetivo consiste en maximizar los enlaces por lo cual se emplea el concepto de "community score" para medir la calidad de la división de las comunidades en la red. Por otro lado está el objetivo de minimizar los enlaces externos, para este se introduce el concepto de "aptitud" en el cual los nodos pertenecen a una comunidad y al pasar de iteración en iteración la suma de esta función va siendo mayor. A medida que la aptitud aumenta se dice que el número de enlaces es minimizado.

Como se ha visto anteriormente, la mayoría de algoritmos se enfocan en mecanismos de búsqueda para el proceso de la detección de comunidades tales como la optimización global y la optimización local. Los algoritmos de optimización global emplean una función objetivo para evaluar la calidad de la modularidad de la red e intentan encontrar el resultado en el espacio completo de solución. Por el contrario, para llevar a cabo la optimización local se utiliza en algunos casos reglas heurísticas como "Edge betweeness", "Random walk", entre otras, es por esto que algunos algoritmos buscan mejorar los resultados de agrupamiento y mejorar la eficiencia de cada

uno de estos enfoques por separado (Ji et al., 2013). El algoritmo "Ant colony clustering algorithm" usa una nueva función de aptitud para tener una mejor percepción del ambiente local y permitir un mejor movimiento de los nodos entre las comunidades. Los resultados muestran que al combinar la percepción del ambiente local y la información global se logra obtener una mejor calidad de las comunidades encontradas.

Dado el amplio rango de aplicaciones en las que se puede encontrar la detección de comunidades, recientemente han surgido nuevas técnicas de solución que incorporan aprendizaje reforzado. Han sido propuestos varios modelos que usan autómatas de aprendizaje para detectar comunidades. Un conjunto de autómatas de aprendizaje distribuidos es utilizado en forma de red para permitir que los autómatas de aprendizaje interactúen entre sí y encuentren nodos que están densamente conectados utilizando aprendizaje cooperativo. Cada autómata de aprendizaje se activa asincrónicamente y selecciona un nodo vecino como acción, hasta que asigna a la comunidad actual un número mayor de enlaces que los nodos del resto de comunidades. Esto se repite hasta que todos los nodos son asignados a alguna comunidad y se computa la función de corte que mide la calidad las particiones y de la cual se actualiza el vector probabilidad de acción. (Khomami et al., 2017).

Nuevos modelos que utilizan autómatas celulares y autómatas de aprendizaje han comenzado a combinarse dando lugar a modelos más robustos y eficientes como el modelo de autómatas de aprendizaje celular (CLA), el cual incorpora autómatas de aprendizaje en sus celdas. Un modelo CLA ha sido utilizado para la detección de comunidades. El modelo es un CLA irregular, abierto y sincrónico que simultáneamente activa sus autómatas de aprendizaje en cada celda y elige como acción un nodo vecino. De acuerdo con dicha elección se constituyen las comunidades y se evalúan usando la función de modularidad, que gracias a la interacción local y global de los autómatas de

aprendizaje soluciona efectivamente el problema de límite de resolución. En cada iteración se evalúan las comunidades obtenidas y se actualiza el vector de probabilidad de acción de acuerdo al desempeño de las comunidades en la modularidad y la restricción de Raghavan (Zhao et al., 2015).

Más recientemente se ha propuesto un modelo CLA irregular, asincrónico que utiliza árboles de expansión parcial para disminuir el tamaño de la red y detecta las comunidades de los árboles reduciendo así el costo computacional. En este modelo, cada autómata de aprendizaje se activa solo cuando ha sido seleccionado como una acción de otro autómata de aprendizaje. Además, es añadido a la comunidad si satisface dos condiciones del grado interno y externo de sus nodos. Una vez encontrado el conjunto de comunidades se computa la función de conductancia y se actualiza el vector probabilidad de los autómatas de aprendizaje. Como criterio de parada se propone una función de entropía o un número arbitrario de iteración máxima (Khomami et al., 2017).

El interés por aplicar el problema de detección de comunidades en redes científicas ha crecido en los últimos años, es por esto que se han realizado trabajos para analizar las redes de colaboración científica, redes de citación y redes de coautoría, de manera que sea posible identificar los intereses de los autores, caracterizar grupos de investigadores dentro de una comunidad específica y entender las relaciones y colaboraciones que usan ellos en su proceso de investigación. Un algoritmo Link Rank ha sido empleado para detectar comunidades en una red de citación, los autores encontraron problemas en la implementación del algoritmo debido a limitaciones en el hardware que utilizaron por lo que debieron hacer modificaciones a las condiciones de la corrida del algoritmo. Analizaron las comunidades obtenidas y observaron el impacto del número de iteraciones en el tamaño de las comunidades. Exitosamente encontraron que un 70% de los nodos de las comunidades compartían

un tema de interés, en cuanto el 30% restante variaba en temáticas (Yudhoatmojo & Samuar, 2017). Además, surge el problema de identificar a aquellos investigadores que mantienen la conectividad dentro de la red, difunden la información y enlazan grupos de investigación para descubrir intereses de búsqueda. Horta et al. (2018) proponen el algoritmo NETSCAN que considera las características multidisciplinarias de los investigadores, permitiendo la identificación de los investigadores que están en más de un área de actividad y que hacen parte de dos o más comunidades científicas. Además de esto, el autor propone una función de subagrupamiento la cual tiene el objetivo de identificar subgrupos de comunidades de búsqueda. El algoritmo propuesto realiza un análisis topológico a una red social científica extraída de la base de datos DBLP. Este análisis demuestra que la topología de la red tiene una baja conectividad global y algunos nodos centralizados, que conllevan a una alta conectividad entre los nodos, permitiendo así identificar los grupos de investigación mejor definidos, los miembros de cada grupo y los investigadores que tienen mayor influencia sobre sus respectivas comunidades.

Otras áreas dentro de la comunidad que han demostrado un gran interés por el análisis de comunidades de autores han sido el área de relaciones públicas y el área de Administración Pública. Estas buscan demostrar que los autores alrededor del mundo tienen un fuerte campo de conocimiento y de colaboración que los relaciona. Diversos grupos de autores han sido identificados y posicionados basados en medidas de centralidad tales como “Edge betweenness”, “Degree”, “Page rank” y “Closeness”. Para este caso, Ahmed et al. (2018) proponen una metodología que consiste en analizar mediante el uso de medidas de centralidad y de organización una base de datos recolectada de Microsoft Academic Graph (MAG). Lo anterior, con el fin de determinar los patrones de relación entre los individuos, equipos, grupos, sociedades, consejos de organización y otras entidades. Para visualizar la centralidad y las comunidades de autores se

utilizaron los softwares Gephi y R. Finalmente los valores de centralidad para diferentes autores reflejan patrones de colaboración y tendencias de ocurrencia a lo largo de 16 años, también se identificaron los autores que tienen mayor influencia como colaboradores de investigación y portadores de conocimiento dentro de una comunidad.

Aunque se observa un gran interés en la literatura de detectar comunidades en redes de investigación, se observa que hasta ahora no se han estudiado este tipo de redes con modelos de autómatas de aprendizaje celular. Además, las redes de co-autoría o co-citación estudiadas anteriormente no han estado enfocadas a una región como Colombia, por lo cual resulta de interés para esta investigación realizar un análisis a una red de investigadores de Colombia, utilizando un modelo de autómatas de aprendizaje celular para la detección de comunidades

3. Marco teórico

3.1 Análisis de redes sociales

El estudio de las redes sociales es un campo de gran desarrollo dentro de las ciencias sociales, la psicología y otras disciplinas que investigan el comportamiento que presentan los seres humanos al relacionarse con otros.

El análisis de redes sociales ARS por su parte, permite comprender y evaluar de manera cuantitativa una gran diversidad de fenómenos presentes en las relaciones establecidas entre diferentes individuos, por lo tanto, el ARS es considerado como el estudio de la estructura social (Hawe, Webster & Shiell, 2004).

Según (Ávila & Madariaga, 2012), existen cuatro mecanismos principales que permiten valorar las particularidades de las redes sociales, los cuales son: Propiedades generales de la red, análisis de características posicionales de los actores o nodos, identificación de sub-agrupaciones dentro de la red y generación de recursos analíticos para comprender este tipo de estructuras.

Con el fin de abordar el problema de detección de comunidades en una red social compleja y teniendo en cuenta los mecanismos propuestos por Ávila & Madariaga (2012), la teoría necesaria para el desarrollo de esta investigación se presentará en dos partes. En la primera parte se muestran los principios de la teoría de grafos, las métricas de centralidad y las propiedades de las redes sociales, las cuales permiten comprender la estructura de la red, las características posicionales de los autores y la integración de miembros de la red respectivamente. En la segunda parte se presenta la teoría relevante para entender el modelo de autómatas de aprendizaje celular ya que será de gran utilidad para el desarrollo del algoritmo propuesto.

3.1.1 Teoría de grafos. Según Herrero (2000), La teoría de grafos es considerada como un área de conocimiento esencial para el análisis de redes sociales, ya que facilita la forma de estudiar las redes y sus estructuras.

A continuación, se definirán algunos conceptos y características de los grafos que se consideran de gran interés para llevar a cabo el análisis de la red social de investigadores.

3.1.1.1 Conceptos

-*Nodo*: Los nodos o vértices son elementos que representan personas, organizaciones, ciudades, etc. Los cuales están interconectados entre sí por una serie de líneas dentro de los grafos.

-*Arista*: Una arista o enlace representa las conexiones que existen entre dos o más nodos y se representa mediante una línea. Un enlace es dirigido si su recorrido va en una sola dirección y no dirigido sí su recorrido es en ambas direcciones.

-*Grafo*: Un grafo es una colección de vértices unidos por aristas o enlaces que representan la conexión entre ellos. Cuando en un grafo existen subgrupos de vértices que al no estar interconectados permiten la ruptura del grafo en dos o más subgrafos, se dice que son grafos no conexos, mientras que sí existe una trayectoria o camino para ir de un vértice a otro para cualquier par de vértices del grafo, se considera que es un grafo conexo.

-*Isomorfismo*: Se dice que un grafo es isomorfo a otro si existe un vértice correspondiente asociado a cada vértice del grafo inicial y las conexiones entre sus vértices se mantienen.

3.1.1.2 Características de un grafo

-*Grado de un nodo*: El grado de un nodo representa la cantidad de enlaces conectados directamente a él. Esta medida representa la conectividad de un vértice, por lo tanto, si el grado es igual a cero quiere decir que el nodo se encuentra aislado de los demás.

-*Densidad*: La densidad de un grafo representa el número total de conexiones o enlaces presentes en una red respecto al total de enlaces posibles. Cuando el valor de la densidad se acerca a 1, se dice que el número de enlaces en la red es similar al número máximo de enlaces.

-*Distancia*: Indica la cantidad de enlaces existentes entre un par de nodos al realizar el recorrido por la ruta más corta.

-*Accesibilidad*: Medida que determina si los vértices de un grafo están relacionados directa o indirectamente a todos los demás vértices de la red, se dice que un vértice no es accesible si no tiene al menos un enlace que lo conecte al resto del grafo.

3.1.2 Métricas de centralidad. En el análisis de redes sociales y la teoría de grafos las medidas de centralidad son una herramienta que permite evaluar la importancia relativa de un vértice en una red. Según Freeman (1978), el concepto de centralidad permite entender el proceso de comunicación y de relacionamiento en pequeños grupos al identificar los nodos de mayor influencia en la estructura de red. Se dice que los actores que ocupan una posición más central en la red, tienen mayor facilidad de acceso a la información y mayor capacidad para controlar el flujo de información.

Existen cuatro medidas que son principalmente usadas para entender el comportamiento de la red e identificar los actores claves que la componen así:

3.1.2.1 *Centralidad de grado.* Es considerada una medida radial de volumen ya que toma un nodo como punto de referencia. La centralidad de grado indica desde una perspectiva local las conexiones directas existentes que tiene un autor con otros autores de la red. Por lo tanto, un autor que posee un alto grado de centralidad tiene el privilegio respecto a otros autores de satisfacer con mayor facilidad las necesidades que demande un sector de la red. La centralidad de grado equivale al valor del grado total de incidencia directa que tiene cada nodo.

$$grado(i) = \sum_j Mat\_ady[i,j] \qquad (1)$$

Donde $Mat\_ady[i,j]$ equivale a 1, si el nodo i esta directamente conectado al nodo j

Para poder comparar esta medida en redes con diferente número de nodos se hace uso de la centralidad de nodo normalizada al dividirla por $n-1$, donde n representa la cantidad de nodos de la red estudiada.

3.1.2.2 ***Centralidad de intermediación.*** La centralidad de intermediación también conocida como "betweenness" representa la ubicación de los nodos en relación a que tanto un punto actúa de intermediario con otro en la red (Hanneman & Riddle, 2005). Un autor tendrá mayor centralidad de intermediación en cuanto más autores necesiten pasar por él para hacer sus conexiones indirectas por los caminos más cortos. Los autores con un valor alto de intermediación ocupan roles críticos en la estructura de la red, ya que pueden ser puentes locales para regular el flujo de contenidos y recursos de información.

Existen dos tipos de centralidad de intermediación, una basada en la frecuencia con la que aparece un nodo en un camino entre dos nodos (trayectoria mínima) y otra basada en la frecuencia con la que se presenta un enlace en medio de una trayectoria mínima. La centralidad de intermediación también se ocupa de evaluar la intermediación de los puntos centralmente periféricos y globalmente periféricos. El valor de betweenness puede ser calculado así:

$$Betweenness(v) = \sum_{i \neq v \in V} \sum_{j \neq v \in V} \delta_{ij} \qquad (2)$$

Donde δ_{ij} es el número de rutas de mínima distancia que unen a los nodos i y j

3.1.2.3 ***Centralidad de cercanía.*** El grado de centralidad de cercanía o "closeness", es una medida que indica la capacidad que tiene un nodo para acceder al resto de nodos de la red. Inicialmente se realiza la sumatoria de los caminos de cada nodo al resto de la red lo cual es conocido como "lejanía", luego para conocer el grado de cercanía de un nodo en particular se

calcula el valor inverso de la lejanía. Un mayor valor de grado de centralidad de cercanía, indica que existen un mínimo número de caminos entre el nodo evaluado y el resto de nodos en la red.

$$Cercania(i) = \sum_j \frac{1}{dij} \qquad (3)$$

Donde dij es la distancia o número de caminos para llegar desde el nodo i hasta el nodo j.

3.1.2.4 ***Centralidad de vector propio.*** Como medida radial de volumen, la centralidad de vector propio es una medida que permite identificar los nodos con un mayor valor de influencia en la red, es decir son nodos que están conectados a muchos nodos que a su vez están bien conectados entre sí. Al analizar una red social es de gran importancia reconocer los actores que poseen un mayor valor de influencia, ya que permiten difundir información, divulgar conocimiento e incluso propagar enfermedades.

A diferencia de la medida de centralidad de grado, donde todos sus nodos tienen el mismo peso, en la centralidad de vector propio los nodos tienen diferentes pesos debidos a sus conexiones, por lo tanto, no se tiene en cuenta la cantidad de las conexiones sino la calidad de las mismas.

3.1.3 Propiedades de las redes sociales. Existen algunas propiedades de las redes sociales que han sido objeto de interés en sus análisis a lo largo de los años; algunas de estas propiedades son: "Small-world", "Ley de potencia en la distribución de los grados", "Transitividad" y "Detección de comunidades". A continuación, se definen:

3.1.3.1 ***Small world.*** Esta propiedad determina los grados de separación entre dos nodos aleatorios de la red social, en otras palabras, a qué distancia se encuentran dos autores escogidos aleatoriamente de la red. Esta información resulta pertinente ya que una red donde esta distancia es pequeña, pone en contacto con más facilidad autores que nunca han trabajado juntos y presenta

mayor difusión y propagación de los productos de investigación de los autores, así como resulta más fácil conocer otros autores a través de coautores y colaborar con ellos en el futuro. Milgram (1967) encontró que en promedio esta separación es de seis grados para una red social, desde entonces ha sido una preocupación en el análisis de redes sociales comprobar dicha afirmación.

3.1.3.2 ***Power law degree distribution.*** Esta propiedad estudiada por Faloutsos et al. (1999) demuestra que la distribución de los grados de los nodos en una red sigue una ley de potencia, más específicamente, la probabilidad de grado crece de forma no-lineal, por lo que se representa usualmente con distribuciones de probabilidad como la exponencial, esto quiere decir que los valores atípicos o extremos de la distribución tienen mayor probabilidad de ocurrir que en distribuciones donde las probabilidades se concentran alrededor de la media y los valores en las colas no llegan a estar tan alejados de la media.

3.1.3.3 ***Transitividad.*** En el análisis de redes sociales, la transitividad es una medida que indica la probabilidad de que los vértices adyacentes a un vértice estén también conectados, es decir que, en un conjunto de 3 autores si el autor 1 ha publicado en coautoría con el autor 2, y el autor 2 ha publicado con el autor 3, entonces existe una probabilidad de que el autor 1 realice una publicación con el autor 3 la cual es conocida como transitividad.

Existen dos clases de transitividad: local y global. La transitividad local mide la relación entre los triángulos conectados al vértice y los triples centros que existen sobre el vértice, mientras que la transitividad global mide la relación de los triángulos y triples conexiones, pero a nivel del grafo.

3.1.3.4 ***Detección de comunidades.*** Una red puede ser representada como un grafo $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ el cual tiene vértices $\boldsymbol{V}$ y enlaces $\boldsymbol{E}$. El grafo G tiene una matriz de adyacencia A, el elemento

$\boldsymbol{A_{ij}}$ de la matriz es igual a 1 cuando el nodo $\boldsymbol{i}$ y el nodo $\boldsymbol{j}$ son vecinos, es decir, están conectados en la red, de lo contrario el valor de $\boldsymbol{A_{ij}}$ será 0.

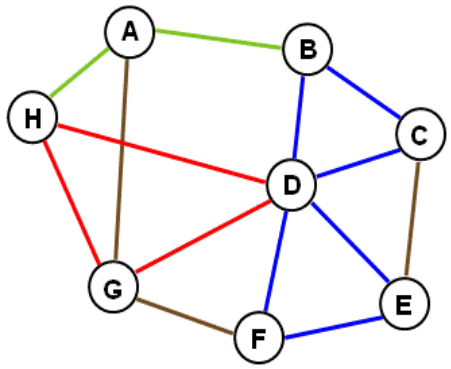

Figura 6. Grafo de 8 nodos y 14 enlaces. Adaptado de: https://upload.wikimedia.org/wikipedia/commons/e/e7/Graph_cycle.gif.

El número total de enlaces se representa por: $|V| = n$

3.1.3.4.1 Comunidad. Se define como comunidad a un subgrafo C de un grafo G, en el cual el número de enlaces al interior de la comunidad C tienen una mayor conexión, mientras que presenta una escasa conexión con otros nodos que pertenecen a otras comunidades. En la figura 7 se puede observar una red social con tres comunidades identificadas con diferentes colores.

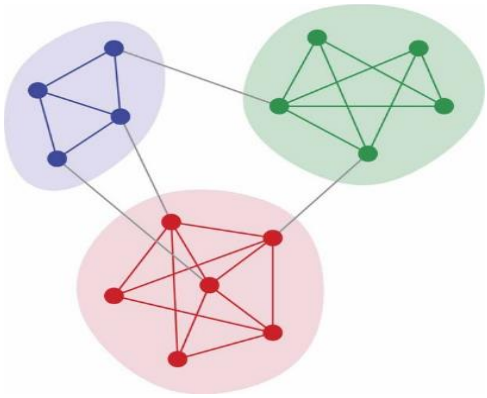

Figura 7. Grafo de una red con 3 comunidades. Adaptado de: http://i.stack.imgur.com/f6EFE.jpg

3.1.3.4.2 *Grado de un nodo.* El grado interno k_v^{int} y externo k_v^{ext} de un vértice $v \in C$, se define como el número de enlaces que conectan a v de otros vértices de C o al resto del grafo respectivamente.

El grado interno k_c^{int} de C es la suma de los grados de sus vértices internos, de igual forma el grado externo k_c^{ext} de C es la suma de los grados de sus vértices externos.

El grado total k^C de una comunidad C, es la suma de los grados de los vértices de C, donde:

$$K_c = k_c^{int} + k_c^{ext} \qquad (4)$$

3.1.3.4.3 *Densidad de comunidad.* La densidad intra- clúster $\delta_{int}(C)$ del subgrafo C, se define como la proporción entre el número de enlaces internos de C y el número total de enlaces posibles en la red.

$$\delta_{int}(C) = \frac{k_c^{int}}{\frac{n_c(n_c-1)}{2}} \qquad (5)$$

Similarmente, la densidad inter-clúster $\delta_{ext}(C)$ es la proporción entre el número de enlaces que hay entre los vértices de C y el resto del grafo y el máximo número de enlaces inter-clúster posibles.

$$\delta_{ext}(C) = \frac{k_c^{ext}}{n_c(n-n_c)} \qquad (6)$$

3.1.3.4.4 *Comunidad fuerte.* Es un subgrafo en el cual cada uno de sus vértices tiene una mayor probabilidad de estar conectado a cada vértice del subgrafo que a cualquier otro vértice de la red.

3.1.3.4.5 *Comunidad débil.* Es un subgrafo tal que la probabilidad promedio de que un vértice se conecte con otros miembros del grupo exceda la probabilidad promedio de que un vértice del grupo se conecte con cualquier otro grupo.

3.1.3.4.6 *Edge-betweenness.* Se define como el número de caminos más cortos entre parejas de nodos que pasan a través de i.

3.1.3.4.7 *Modularidad.* La modularidad Q es una medida de calidad que permite evaluar un conjunto de comunidades obtenidas por el algoritmo a partir de una estructura de la red, es la más utilizada en la literatura. Se calcula así:

$$Q = \frac{1}{2m} \sum_{C \epsilon P} \sum_{vi\, vj \epsilon C} [A_{ij} - \frac{k_i k_j}{2m}] \qquad (7)$$

Donde A es la matriz de adyacencia y A_{ij} toma un valor de uno (1) si existe un enlace entre el nodo v_i y el nodo v_j y toma un valor de cero (0) si no existe ningún enlace entre los nodos, por lo tanto el grado del nodo v_i es $k_i = \sum_j A_{ij}$ y m es el número total de enlaces en la red.

3.1.3.4.8 *Complejidad computacional.* La complejidad computacional de un algoritmo se estima como la cantidad de recursos requeridos por el algoritmo para realizar una tarea. Esto involucra tanto el número de pasos necesarios como el número de unidades de memoria que se requieren simultáneamente para correr la computación.

3.2 Autómatas de aprendizaje celular

Debido a que los modelos de autómatas de aprendizaje celular son una combinación de autómata celular y autómatas de aprendizaje se procede a definir primero ambos modelos.

3.2.1 Autómatas celulares. Los autómatas celulares son idealizaciones matemáticas de sistemas físicos en los que el espacio y el tiempo son discretos y las cantidades físicas toman valores finitos discretos. Los autómatas celulares consisten en una rejilla rectangular uniforme, usualmente de tamaño finito, con una variable discreta en cada celda. Cada celda tiene un estado que está especificado por cada variable en cada celda.

El autómata celular evoluciona a pasos discretos, en donde su variable cambia de valor dependiendo del valor en el paso anterior de las variables de otras celdas que son consideradas "vecinas". Esta vecindad se conforma usualmente por la celda en cuestión y todas las celdas inmediatamente adyacentes, sin embargo, dependiendo de la aplicación la vecindad puede tener un significado diferente.

Los autómatas celulares sincrónicos actualizan el valor de todas las variables en sus celdas sincrónicamente de acuerdo con un conjunto establecido de reglas locales usando un reloj interno. (Wolfram, 2018). Algunas características importantes de los autómatas celulares son:

3.2.1.1 ***Estados.*** Los estados son un conjunto finito de valores que puede tomar la variable dentro de la celda para cada paso en el tiempo.

3.2.1.2 ***Vecindad.*** Son las celdas que se consideran vecinas de la celda seleccionada y sus estados en el tiempo anterior afectan el nuevo estado de la celda seleccionada.

3.2.1.3 ***Reglas locales.*** Es un patrón determinístico que establece cuál será el estado de la celda (valor de la variable) en el paso actual de acuerdo con el estado previo de la celda y su vecindad.

3.2.1.4 ***Definición matemática.*** Un autómata celular de dimensión d es una estructura: $\boldsymbol{A} = (\boldsymbol{Z_d}, \boldsymbol{\Phi}, \boldsymbol{N}, \boldsymbol{F})$. En donde:

(i) Z_d es una rejilla de d-tuplas de números enteros. Cada celda en la rejilla de dimensión d dimensional Z_d, está representada por una d-tupla $(z_1, z_2, \ldots, z_d)$.

(ii) $\Phi = \{1, \ldots, m\}$ es el conjunto finito de estados que pueden tomar las celdas.

(iii) $\underline{N} = \{\underline{x}_1, \underline{x}_2, \ldots \underline{x_m}\} = \{\underline{x}_1, \underline{x}_2, \ldots \underline{x_m}\} = \{\underline{x}_1, \underline{x}_2, \ldots \underline{x_m}\}$ es un subconjunto finito de Z_d. llamado vector de vecindad, donde $x_i \in Z_d$. El vector vecindad determina la posición relativa de las celdas de la rejilla de la vecindad desde una celda dada u en la rejilla Z_d.

Los vecinos de una celda u particular son el conjunto $\{u + \underline{x}_i \mid i = 1, 2, \ldots, \underline{m}\}$ se asume que existe una función $\underline{N}(u)$ que mapea una celda u al conjunto de sus vecinos, eso es

$$\underline{N}(u) = (u + \underline{x}_1, u + \underline{x}_2, .., u + \underline{x_m}) \qquad (8)$$

(iv) $F: \Phi m \rightarrow \Phi$ es la regla loca del autómata celular. Computa el nuevo estado para cada celda del estado actual de sus vecinos.

En la figura 8 se observa un ejemplo clásico de autómata celular, consiste en la regla 30, corresponde a uno de los conocidos "autómatas celulares de Wolfram".

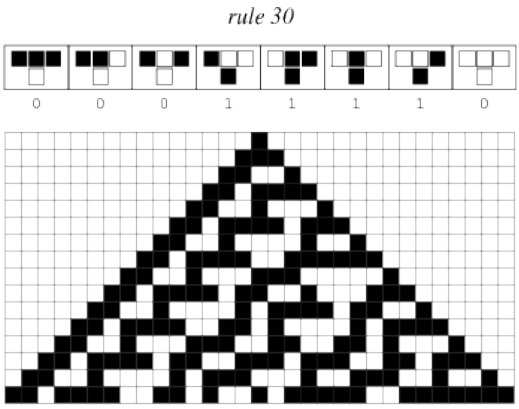

Figura 8. Regla 30. Ejemplo de autómata celular conocido como "regla 30". Adaptado de http://mathworld.wolfram.com/images/eps-gif/ElementaryCARule030_700.gif

3.2.2 Autómatas de aprendizaje. El aprendizaje en autómatas de aprendizaje ha sido estudiado usando el paradigma de un autómata que opera en un ambiente aleatorio desconocido. En palabras simples, un autómata de aprendizaje es una entidad que posee un conjunto finito de acciones para elegir en cada etapa, su elección (acción) depende de un vector de probabilidad de acción. Para cada acción escogida por el autómata, el ambiente envía una señal de refuerzo en cada etapa, y evoluciona hasta un comportamiento final deseado.

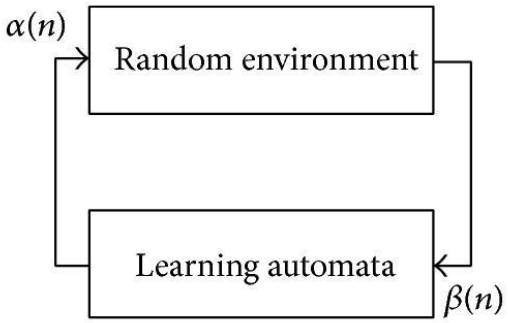

Figura 9. Autómata de aprendizaje. Adaptado de: https://file.scirp.org/Html/27-7400960/eb3e4b5a-fe77-47f7-b836-69c73a63ff8c.jpg

Una clase de autómata de aprendizaje es un autómata de aprendizaje con estructura variable y se representa mediante una tripleta $\{\beta, \alpha, T\}$ en donde β es un conjunto de entradas, α es el conjunto

de acciones y T es un algoritmo de aprendizaje. El algoritmo de aprendizaje es una relación recurrente que se usa para modificar el vector de probabilidad de acción p. En la literatura se han reportado varios algoritmos de aprendizaje. Uno de los más usados es el algoritmo lineal de recompensa y castigo (Linear reward-penalty algorithm).

Sea a_i la acción escogida en el tiempo k como una muestra de la distribución de probabilidad $p(k)$.

Cuando $\beta(k) = 0$

$$p_j(k+1) = \{p_j(k) + a \times \left[1 - P_j(k)\right] \qquad (9)$$

$$i = j \; p_j(k) - a \times p_j(k) \qquad i \neq j \qquad (10)$$

Cuando $\beta(k) = 1$

$$p_j(k+1) = \{p_j(k) \times [1-b] \qquad (11)$$

$$i = j \; \frac{b}{r-1} + p_j(k)(1-b) \qquad i \neq j \qquad (12)$$

Adicionalmente existen otros algoritmos de aprendizaje como el algoritmo CPRP, este algoritmo ha mostrado aumentar la rapidez de convergencia comparado con el algoritmo LRP que fue previamente mencionado en esta sección (Thathachar & Sastry, 2002). A continuación, se describe el algoritmo de aprendizaje:

El algoritmo de aprendizaje CPRP es un algoritmo que persigue la acción que se ha estimado actualmente como la acción óptima. Esa acción óptima se determina con el vector estimador $\widehat{D}$, el cual es calculado así:

$$\widehat{D}_i(t) = \frac{W_i(t)}{Z_i(t)} \;\; i = 1, 2, \ldots, r \qquad (13)$$

Donde $Z_i(t)$ es el número de veces que la acción a_i ha sido escogida hasta el ciclo actual y $W_i(t)$ es el número de veces que la acción a_i ha sido recompensada hasta el ciclo t. La acción con un mayor valor de $\widehat{D}_i(t)$ es estimada como la acción óptima actual para el ciclo t. Durante cada ciclo el algoritmo CPRP primero escoge una acción de sus acciones disponibles de acuerdo con el vector probabilidad de acción. Después, si la acción escogida es recompensada o penalizada, el algoritmo solo aumenta la probabilidad de la acción óptima actual a_m de acuerdo con las siguientes ecuaciones:

$$p_j(t+1) = \begin{cases} p_j(t) + \alpha\left[1 - P_j(t)\right] & j = m \\ j = m\ (1-\alpha)p_j(t) & j \neq m \end{cases} \qquad (14)$$

Donde α es el parámetro de recompensa

Los autómatas de aprendizaje han sido usados en una gran cantidad de aplicaciones como solución de problemas NP-Complete, asignación de capacidad, ingeniería de redes neuronales y redes celulares. (Beigy et al., 2004).

3.2.3 Autómata de aprendizaje celular. Como ya se ha mencionado previamente, un autómata de aprendizaje celular es una combinación de autómatas celulares y autómatas de aprendizaje, sin embargo, a continuación, se hará la introducción formal de este tipo de modelo.

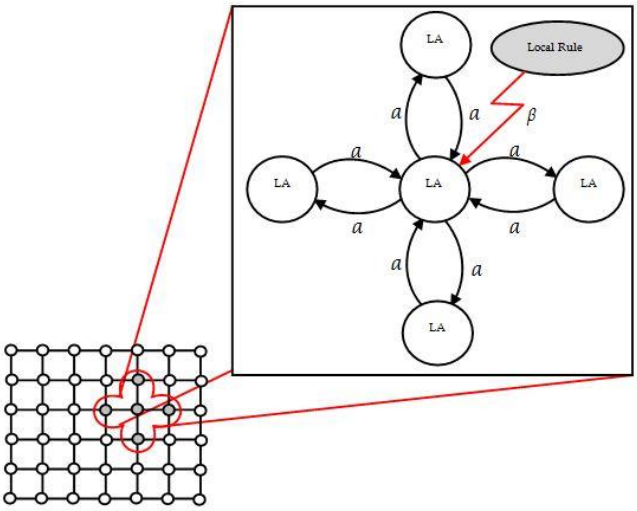

Figura 10. Autómata de aprendizaje celular. Adaptado de: https://www.intechopen.com/media/chapter/44535/media/image29.jpeg

La idea básica de un CLA, el cual es una subclase de los autómatas celulares estocásticos, es usar los autómatas de aprendizaje para ajustar la probabilidad de transición de estado del autómata celular estocástico. Los CLA se pueden clasificar en sincrónicos y asincrónicos. En sincrónicos, todas las celdas se sincronizan con un reloj global y se ejecutan al mismo tiempo. En este modelo el espacio está discretizado en celdas. Las acciones de cada autómata de aprendizaje corresponden a valores discretos de la correspondiente variable de control. Cada celda elige una acción cuando su correspondiente autómata de aprendizaje se activa, esta acción es elegida basada en el vector probabilidad de acción. Esas acciones son aplicadas y el estado de todo el autómata celular se actualiza. El ambiente pasa una señal de refuerzo a los autómatas de aprendizaje de las celdas activadas. Dependiendo de la señal el autómata de aprendizaje actualizará su vector de probabilidad de acción. Ese proceso continúa hasta que el estado de terminación se alcanza.

Un CLA de d-dimensión tiene la siguiente estructura: $A = (Z_d, \Phi, A, N, F)$, en donde:

-Z_d es una rejilla de d-tuplas de números enteros.

-Φ es un conjunto finito de estados.

-A es el conjunto de LA (autómatas de aprendizaje) donde cada uno está asignado a una celda del CLA.

-$\underline{N} = \{\underline{x}_1, \underline{x}_2, \ldots \underline{x_m}\}$ es un subconjunto finito de Z_d llamado vector de vecindad, donde $x_i \in Z_d$.

-F: $\Phi m \rightarrow \beta$ donde β es el conjunto de valores que la señal de refuerzo puede tomar. Computa la señal de refuerzo para cada LA. es la regla local del autómata de aprendizaje celular.

Se considera un CLA con n celdas y la función de vecindad $\underline{N} = \{\underline{x}_1, \underline{x}_2, \ldots \underline{x_m}\}$. Un LA, A_i, se asocia a cada celda i(para $i = 1, 2, \ldots, n$) del CLA, con un conjunto finito de acciones α_i donde la cardinalidad de a_i es m_i. El estado del CLA se representa por $P = \left(P_1', P_{2,\ldots}', p_n'\right)'$, donde $p_i = \left(p_{i1}, \ldots, p_{im_i}\right)'$ es el vector de probabilidad de acción de A_i.

3.2.3.1 ***Autómata de aprendizaje celular irregular (ICLA).*** Es un autómata de aprendizaje celular cuya estructura es irregular y no corresponde a la clásica forma rectangular de los autómatas celulares, este ha sido utilizado en detección de comunidades ya que permite hacer una correcta representación del grafo de la red social. (Esnaashari & Meybodi 2007). En la figura 11 se observa una representación visual de un autómata de aprendizaje celular irregular y la relación de sus celdas.

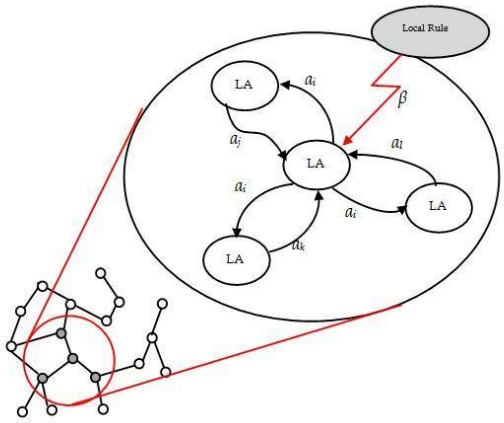

Figura 11. CLA irregular. Adaptado de: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQNl6GS6LEGrOGjr_nsspfkV2lBERVP9rYc5cm45OYaTugmX7CN

4. Preparación de datos y diseño del algoritmo

4.1 Fuentes de información

A continuación, se describe el proceso de recolección de la información, así como las fuentes utilizadas, la construcción de la red social y la preparación a la que se sometieron los datos antes del diseño e implementación del algoritmo.

4.1.1 Recolección de la información

El proceso de recolección de la información consistió inicialmente en buscar una lista de investigadores que han publicado en el área de Ingeniería industrial en Colombia. Para esto se utilizó el listado actual de pares evaluadores reconocidos del Sistema Nacional de Ciencia y Tecnología publicado por Minciencias, pues es una fuente de información accesible con investigadores colombianos reconocidos en Colombia, de la cual es posible seleccionar la muestra de autores para construir la red. La lista utilizada fue obtenida al aplicar los siguientes filtros:

Gran Área: Ingeniería y Tecnología >> **Área:** Otras Ingenierías y Tecnologías >> **Disciplina**: Ingeniería Industrial.

La lista consta de 396 autores reconocidos como pares evaluadores para la disciplina de Ingeniería Industrial, de los cuales se seleccionó una muestra de 100 autores a consultar con el fin de conservar un tamaño de red similar a las redes probadas en la literatura. Posteriormente, se utilizó el formato de hoja de vida electrónica de Colciencias CvLAC para recopilar información de los autores y de los artículos que cada autor del listado de pares ha publicado. En esta búsqueda individual se extrajo la información de los artículos relacionados en: "Producción bibliográfica >> Artículos", de tal forma que, en la construcción de la red, la relación entre autores se presente al haber publicado al menos un artículo científico juntos. En este proceso de consulta se excluyeron 17 investigadores que tienen su perfil de CvLAC oculto o no presentan información de sus coautores en los artículos publicados.

De manera paralela se consolido la información de los pares evaluadores a partir de los cuales se construyó la red social, así como de otros investigadores que hacen parte de la red. Para esto se incluyó los atributos de categoría, institución a la que pertenecen y sus respectivas líneas de investigación de tal forma que la información permita evaluar más adelante las comunidades obtenidas.

4.1.2 Construcción de la red social de investigación

A partir de los artículos seleccionados, se asume que los nodos de la red representan a cada uno de los autores y los enlaces la relación de coautoría, es decir, existirán enlaces entre aquellos autores que han publicado juntos al menos un artículo dentro del total de artículos seleccionados para el

análisis. Por otra parte, se verifica que el grafo sea conexo, quiere decir que no existan nodos aislados dentro del mismo ocasionando la partición del grafo en varios subgrafos.

Una vez se consolidó la base de datos, se encontraron 1161 artículos, los coautores de estos artículos que no se encontraban en la lista inicial de investigadores entraron a hacer parte de la red por lo que finalmente la red de investigadores consolidada fue de 966 nodos y 2357 arcos. La red se construyó usando el software estadístico R Studio ® mediante el paquete "igraph", un paquete para el análisis de redes sociales, con este se generó el grafo de la red social, lista de arcos, matriz de adyacencia, entre otros. Para la visualización del grafo se utilizó Gephi, una aplicación útil para la visualización de datos en forma de grafos. En la figura 12, se ilustra el grafo de la red social.

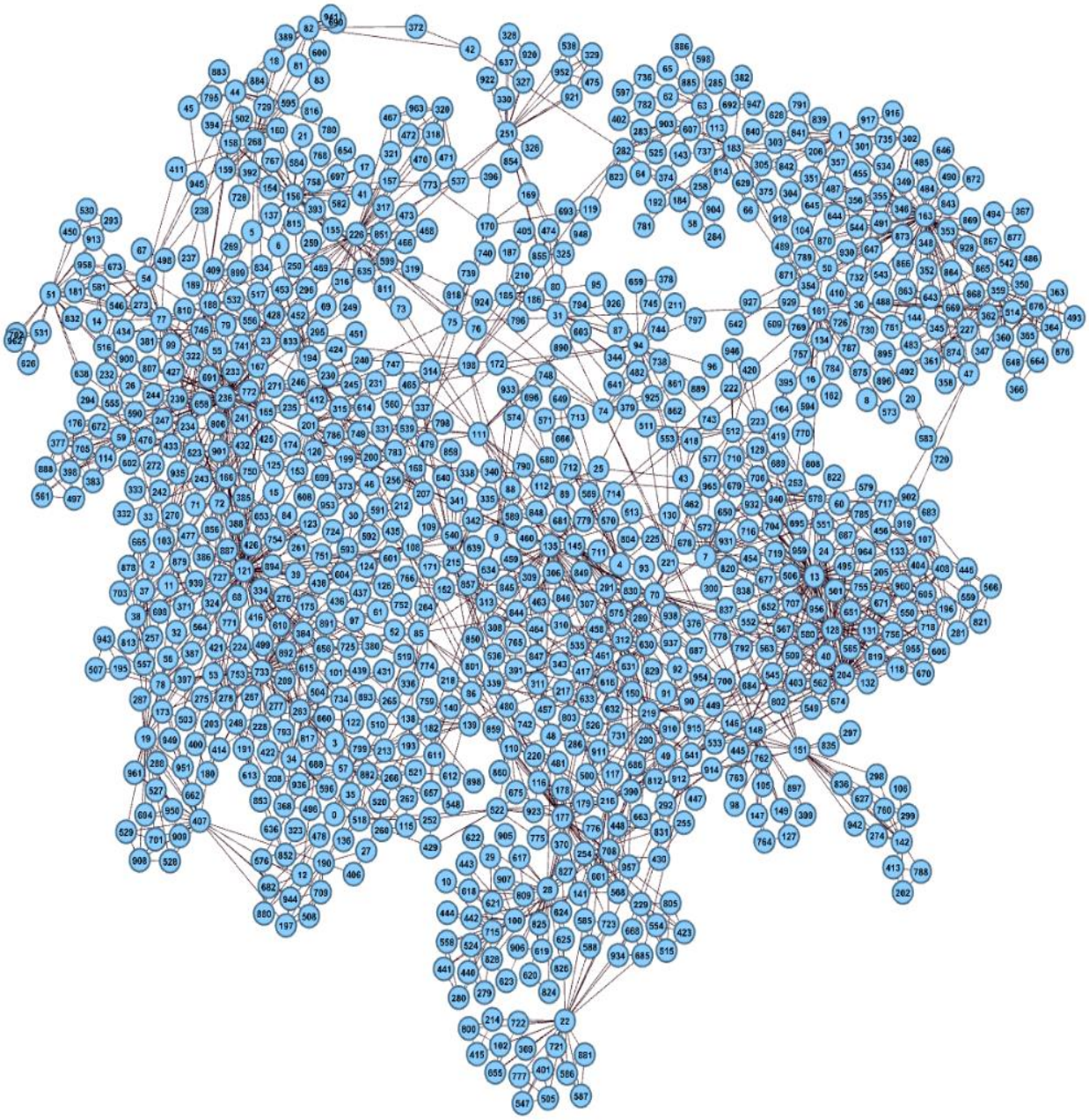

Figura 12. Grafo de la red social. Generado en Gephi 9.0.2 bajo licencia de desarrollo común.

4.2 Preparación de los datos

Una vez obtenido el grafo final de la red de investigadores se construyó la matriz de adyacencia del grafo, para esto se debe hacer una transformación que implica numerizar las relaciones entre

autores, asignando "1" a las parejas de autores que son coautores en al menos un artículo y "0" a quienes no son coautores en ningún artículo dentro de la base de datos de artículos consolidada. Así mismo, los autores se identificaron con un número que representa al nodo dentro de la red. La matriz de adyacencia es la entrada principal del algoritmo de Autómatas de Aprendizaje Celular; a partir de ésta, el algoritmo determina las posibles acciones que tiene cada autómata de aprendizaje en cada nodo y genera las probabilidades de elegir cada acción.

4.3 Diseño del algoritmo basado en un modelo de autómatas de aprendizaje celular

El algoritmo basado en un modelo de autómatas de aprendizaje celular es un algoritmo que ha sido aplicado a diferentes problemas de optimización (Meybodi & Kharazmi, 2004). Los autómatas de aprendizaje (LA) son objetos capaces de aprender a tomar decisiones por medio de una señal de refuerzo que obtienen de un ambiente aleatorio, estos objetos emplean aprendizaje reforzado para mejorar su toma de decisiones a medida que son castigados o recompensados tras ver el impacto de sus decisiones en una función que se pretende optimizar. El modelo de Autómatas de Aprendizaje Celular (CLA) es de tipo celular debido a que cada objeto se encuentra asignado a una celda dentro de una rejilla, en donde los autómatas de aprendizaje son capaces de interactuar entre sí de forma local y global permitiendo encontrar la estructura de comunidad óptima (Zhao et. Al, 2015).

El algoritmo CLA-Net es un algoritmo compuesto por un autómata celular irregular sincrónico (CA) en forma de grafo en donde cada celda corresponde a un nodo del grafo de la red social. A su vez cada nodo tiene asignado un autómata de aprendizaje (LA) que se encarga de controlar los estados de cada celda del autómata celular y la función de transición que determina la evolución

del mismo a través del tiempo. En la figura 13 se observa un ejemplo de la estructura de un CLA irregular en forma de grafo:

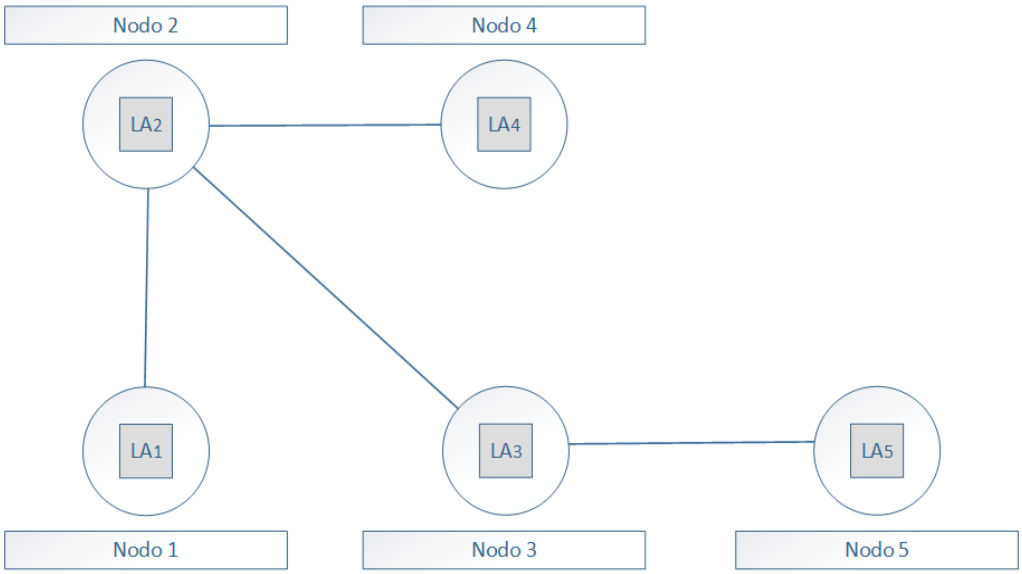

Figura 13. Ejemplo de un ICLA en forma de grafo.

Usando aprendizaje reforzado, el autómata de aprendizaje dentro de cada nodo es quien elige su acción del conjunto de acciones posibles en cada ciclo, esta acción es al mismo tiempo el estado de la celda del autómata celular. En este caso la acción del LA consiste en elegir alguno de sus nodos vecinos, el vecino elegido es el estado de la celda del autómata celular, para la solución esto implica que ambos nodos harán parte de una misma comunidad. La acción tomada por todos los nodos y los nodos en sí mismos son en esencia una lista de arcos que revelan una posible solución, es decir, una estructura de comunidad.

La estructura de comunidad obtenida se evalúa teniendo en cuenta dos condiciones que deben cumplirse: la función de optimización (modularidad) del ciclo t debe ser mayor o igual a la mejor modularidad obtenida hasta el momento y la desigualdad de Raghavan, una reconocida condición en la literatura que indica una estructura de comunidad fuerte, densas conexiones internas y

conexiones externas más débiles. El autómata de aprendizaje de cada nodo será recompensado si cumple ambas condiciones, de lo contrario será castigado.

Las probabilidades de elegir cada acción en el ciclo t, a medida que el autómata de aprendizaje LA_i es castigado o recompensado, aumentan para la acción óptima y disminuyen para el resto. La acción óptima es aquella que tiene un mayor índice de relación entre las veces que se ha recompensado y las veces que ha sido elegida, esta relación se denomina "Estimador". De esta manera y después de un cierto número de ciclos, los autómatas de aprendizaje comienzan a mostrar mayor preferencia por una de sus acciones hasta que el algoritmo converge a una estructura de comunidad.

La solución inicialmente corresponde a un vector solución $S(t)$ el cual en la posición i que representa al nodo i, tiene el valor del nodo vecino j que eligió; es por esto que la posición dentro del vector y el valor en la posición representan la lista de arcos que permite formar las comunidades. Sin embargo, con el fin de revelar las comunidades formadas es necesario decodificar este vector en un vector de comunidades (Communities vector) $C(t)$ que indica en la posición i el número de la comunidad a la que el nodo i pertenece, para la decodificación se utiliza la heurística "Depth-First Search". A continuación, se ilustra un ejemplo de esta decodificación:

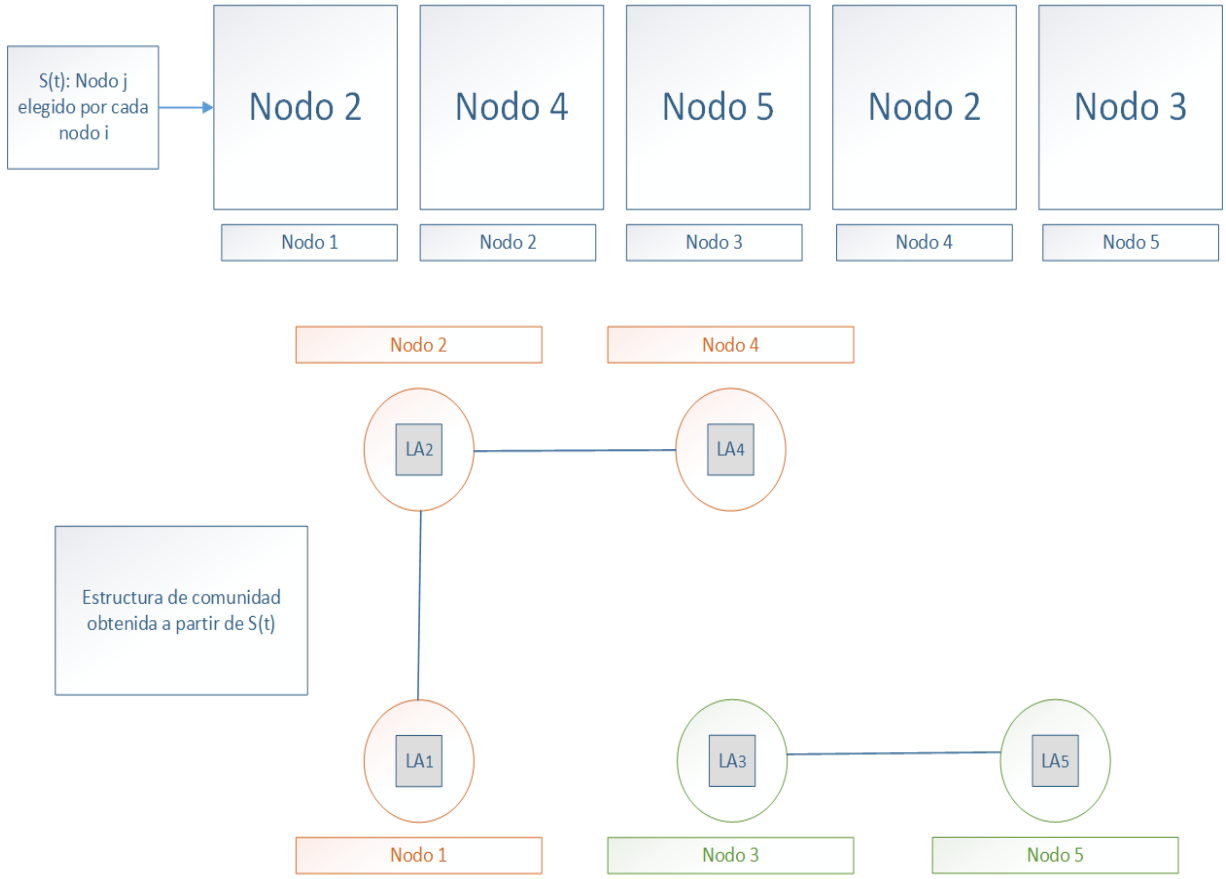

Figura 14. Decodificación del vector solución al vector de comunidades.

En el ejemplo anterior se observan solo los arcos o enlaces que indicó el vector solución y finalmente esto conduce a obtener dos comunidades, la naranja y la verde. Los nodos naranjas conforman la comunidad 1 y los verdes conforman la comunidad 2, estas comunidades se representan a través de un vector de comunidades como el siguiente: $C(t) = (1, 1, 2, 1, 2)$.

Una vez revelada la estructura de comunidad se repite la evaluación de las comunidades y cuando se obtiene la misma estructura de comunidad por un determinado número de ciclos consecutivos, el algoritmo se detiene.

4.3.1 Pseudocódigo del algoritmo CLA-Net

Entrada:

-$Anxn$: la matriz de adyacencia de la red social $G = (V, E)$, en donde V es el conjunto de nodos y E es el conjunto de arcos enlazando los nodos de la red. $A_{ij} = 1$ *c*uando los nodos i y j están conectados, $A_{ij} = 0$ de lo contrario.

Parámetros:

-α: parámetro de recompensa que actualiza el vector de probabilidad de acción, donde $0 < a < 1$

-*Min convergencia*: parámetro de convergencia que establece el número mínimo de veces que el algoritmo debe obtener un mismo conjunto de comunidades en varios ciclos para adoptarlo como solución y detener el algoritmo.

Variables:

-ri: número de acciones que puede elegir el autómata de aprendizaje LA_i en el nodo i, corresponde al grado del nodo i.

-$Wij(t)$: número de veces que la acción j elegida por i ha sido recompensada hasta el ciclo t. $1 < i < n$ y $1 < j < ri$.

-$Zij(t)$: número de veces que la acción j ha sido elegida por i hasta el ciclo t. $1 < i < n$ y $1 < j < ri$.

-$Qbest$: la modularidad más alta que se ha obtenido en los ciclos pasados.

Inicialización:

Paso 1: Se usa probabilidad uniforme para elegir la acción de cada LA_i. $Pij = 1/ri$, para $1 < i < n$ y $1 < j < ri$.

Paso 2: Se inicializa $Wij(t)$, $Zij(t)$ y $Qbest$, escogiendo la acción de cada LA_i con probabilidades uniformes por un pequeño número de veces, con $1 < i < n$ y $1 < j < ri$.

Ciclo:

Paso 3: Cada LA_i elige una acción j de acuerdo al vector de probabilidad $Pij(t)$, se guardan las acciones elegidas por cada LAi en $ai(t)$.

Paso 4: Se genera el vector solución $S(t)$ a partir de $ai(t)$.

Paso 5: Se transfiere el vector solución $S(t)$ al vector comunidades $C(t)$ usando Depth-First Search para decodificarlo.

Paso 6: Se calcula la modularidad Q para la comunidad $C(t)$ obtenida en el ciclo actual t.

Paso 7: Se evalúan las dos condiciones para recompensar o castigar al LA_i de acuerdo a su acción elegida en el ciclo t.

Compara la modularidad actual con la mejor modularidad obtenida

$$Q(t) > Qbest \qquad (15)$$

Se verifica el cumplimiento de la desigualdad de Raghavan.

$$Ki(Ci(t)) > Ki(C'), \forall C' \neq C_i(t) \qquad (16)$$

Si las ecuaciones 15 y 16 se cumplen, el LA_i se recompensa, de lo contrario se castiga, es decir:

$$\begin{cases} Bi\,(t) = 0 \\ Bi\,(t) = 1 \end{cases} \qquad (17)$$

Paso 8: Se actualiza $Qbest$ así:

$$Qbest = Max(Q(t), Qbest) \qquad (18)$$

Paso 9: Una vez obtenido $ai(t)$ se calcula el $Wij(t)$ y $Zij(t)$ correspondientes, así:

$$\begin{cases} w_{ij}(t) = w_{ij}(t-1) + \left(1 - \beta_i(t)\right) \\ z_{ij}(t) = z_{ij}(t-1) + 1 \end{cases} \qquad (19)$$

Paso 10: Se estima la acción óptima actual (mayor $\hat{D}_i$) de los LA_i usando:

$$\hat{D}_i(t) = \frac{w_i(t)}{z_i(t)}\, i = 1, 2, 3, \ldots, r \qquad (20)$$

Paso 11: Se actualiza el Pi de cada LA_i de acuerdo a:

$$P_j(t+1) = \begin{cases} P_j(t) + \alpha\left(1 - P_j(t)\right) & j = m \\ P_j(t)(1-\alpha) & j \neq m \end{cases} \qquad (21)$$

Se repite, hasta que las comunidades obtenidas no cambien por un mínimo de ciclos (parámetro de convergencia).

Resultado:

-Vector solución $S(t)$

-Vector de comunidad $C(t)$

-Estructura de comunidad de la red social (grafo)

5. Implementación del algoritmo en una red de investigadores en Colombia

El algoritmo CLA fue programado usando el lenguaje de programación R en el software R Studio ®. La primera parte del código corresponde a un algoritmo que procesa los datos, construye la red social y permite obtener la matriz de adyacencia de la red, la segunda parte corresponde al modelo de autómatas de aprendizaje celular y la tercera parte corresponde a un algoritmo que permite analizar las comunidades obtenidas por el algoritmo, así como la red en general.

5.1 Ajuste de parámetros

Para el algoritmo CLA-Net se tiene en cuenta dos importantes parámetros que deben ajustarse antes de inicializarlo. Estos dos parámetros son α y β, α es el parámetro de recompensa que define el aumento de las probabilidades para la acción óptima en cada iteración, mientras que β es el parámetro de convergencia, el cual rige al criterio de parada, β es el número mínimo de ciclos consecutivos en los que el algoritmo debe obtener la misma estructura de comunidad para detenerse. En la literatura no se menciona un valor apropiado de estos parámetros para usar, por esta razón, se realizó un análisis de varianza con el fin de determinar el efecto de estos dos parámetros en el valor de la función de modularidad de la estructura de comunidad obtenida.

Se diseñó un experimento con dos factores y dos niveles para cada factor. Se eligieron 0.2 y 0.6 como niveles para el parámetro α ya que requieren un número de iteraciones aceptable para converger, valores por fuera de este rango tienden a tomar tiempos más largos de convergencia. Así mismo, el parámetro de convergencia para más de 6 ciclos consecutivos demanda un gran número de iteraciones que complica la realización del experimento y un valor inferior a 3 no

permite tener certeza sobre la estabilidad de la solución obtenida. Para simplificar el experimento se decidió no usar réplicas, por esto se estudiarán solo los efectos principales de los factores A y B. Además, se escogió una semilla que para los parámetros escogidos mostrara convergencias relativamente rápidas. Esta semilla podría variar para otro conjunto de datos de red social, de manera arbitraria se estableció 1 como la semilla para esta red en particular. En la tabla 5 se observan los valores de la variable respuesta modularidad para las distintas combinaciones de los niveles en cada parámetro.

Tabla 5.

Resultados de modularidad para realizar diseño de experimentos

	α = 0.2	**α = 0.6**
β = 3	0.753904	0.710706
β = 5	0.754027	0.710706

Usando la prueba Fisher y la suma de cuadrados para cada factor como estimadores se realizó la prueba de hipótesis, se obtuvieron los siguientes resultados:

Tabla 6.

Prueba Fisher para el efecto de los factores de recompensa y convergencia

		GL	**MC**	**Fo**	**F**	**Valor-p**	**Nivel significancia**
SCA	0.001871	1	0.001871	494780.71	161.4	0.001	0.05
SCB	3.78E-09	1	3.78E-09	0.9999998	161.4	0.5	0.05
SCT	0.001871	3	0.000624				
SCE	3.78E-09	1	3.78E-09				

Ya que Fo > F para el factor A, se rechaza la hipótesis nula de que todas las medias de los niveles del factor A son iguales. Es decir, se concluye que el factor *A*, correspondiente al parámetro de recompensa, tiene un efecto significativo sobre la modularidad. Por otro lado, para el factor *B* la prueba F arroja como resultado Fo < F, esto quiere decir que no se rechaza la hipótesis nula, es decir, se acepta que todas las medias de los niveles del factor *B* son iguales, por tanto, no es posible afirmar que el factor de mínima convergencia tenga un efecto significativo sobre la modularidad. En la figura 15 se puede observar el efecto de cada factor en sus dos niveles:

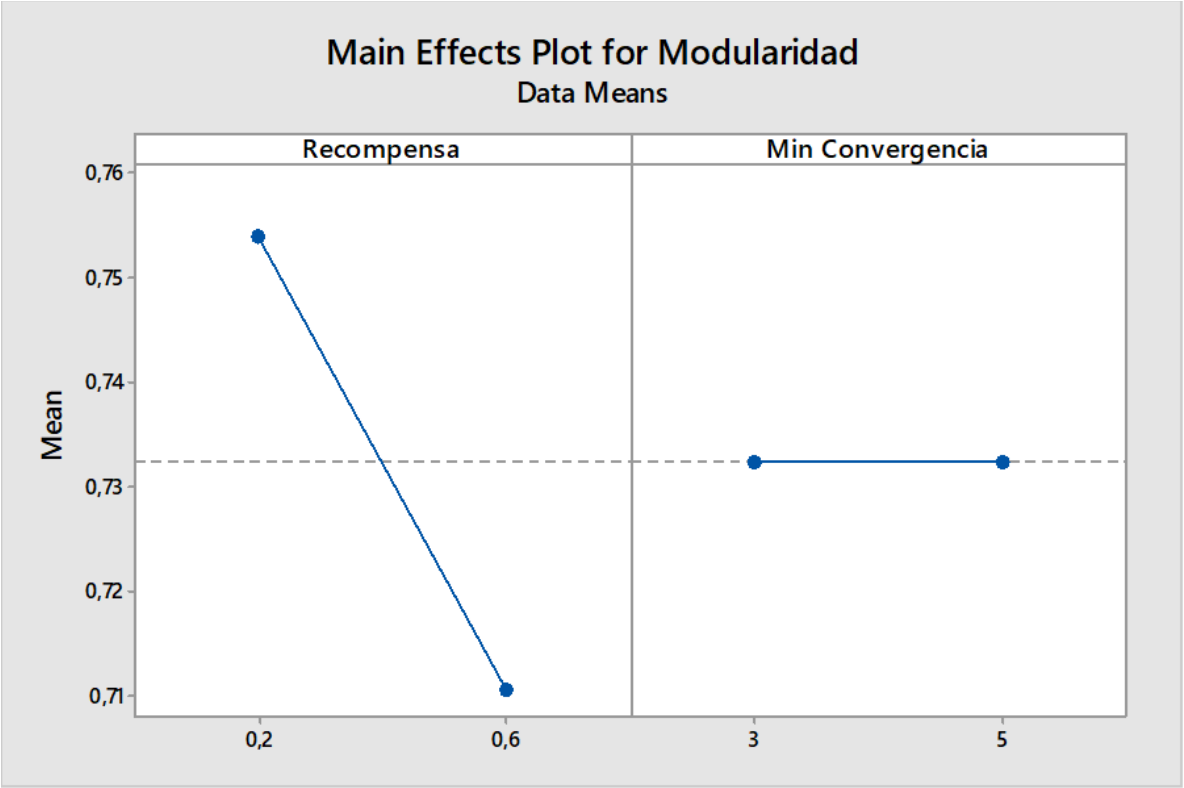

Figura 15. Gráfica de los efectos de los factores principales. Generado en el software estadístico Minitab.

Como se observa en la figura 15 el nivel de factor de recompensa que tiene mayor efecto sobre la modularidad es 0.2, por este motivo se eligió este parámetro, para el factor *B* no se evidencia diferencia en los dos niveles, sin embargo, se decidió elegir 5, de esta manera el algoritmo se

detendría después de mostrar estabilidad en la solución obtenida aunque no tuviera efecto real en la modularidad.

El tamaño de la red social y las características propias de la misma requieren una selección de parámetros particular para la red a analizar, por esto es importante seleccionar una buena semilla y analizar los efectos de los parámetros en la modularidad, de manera que se establezca una combinación apropiada de ellos que no requiera grandes tiempos de convergencia ni sacrifique la calidad de la solución obtenida. En general una apropiada combinación de parámetros y selección de semilla permite la implementación de este algoritmo en redes sociales del orden de 1000 nodos con tiempos de convergencia por debajo de los 30 minutos y obteniendo soluciones con valores de modularidad relativamente altos. Esto juega un papel importante en la estructura de comunidad obtenida para la red y de la calidad de la misma dependerá la precisión de los análisis de la red social y sus comunidades que se efectúan por lo general después de la detección de comunidades.

Se decidió usar un parámetro de recompensa de 0.2 para la semilla establecida que fue 1, ya que este parámetro ofrecía la mejor modularidad en un tiempo de convergencia aceptable. Este parámetro es importante ya que determina la calidad de las comunidades obtenidas y regula el tiempo de convergencia, además el parámetro de mínima convergencia se estableció en 5, es decir, el algoritmo debía obtener la misma solución 6 veces consecutivas para detenerse. El algoritmo se implementó en un computador con memoria RAM de 8GB y procesador Intel Core i5 a 3.2 GHz.

5.2 Resultados

Una vez seleccionados los parámetros anteriores se implementó el algoritmo en la red social de investigadores. En la tabla 7 se observan las métricas de la corrida:

Tabla 7.

Métricas al implementar el algoritmo CLA-Net

Iteraciones	**Q**	**Qmáx**	**Comunidades**
248	0.754027	0.754027	63

Ya que la modularidad obtenida para la estructura de comunidad se encuentra más cerca de 1 que de 0, se evidencia la existencia de una estructura de comunidad dentro de la red social y esto indica soluciones de buena calidad. Es así que un alto valor de modularidad indica que las comunidades obtenidas son significativas y no mero resultado del azar. Además, para los parámetros escogidos el algoritmo convergió en el valor más alto de modularidad alcanzado en las 248 iteraciones, lo cual refleja la buena selección de los parámetros. En la figura 16 se observa la red social con las 63 comunidades en diferentes colores:

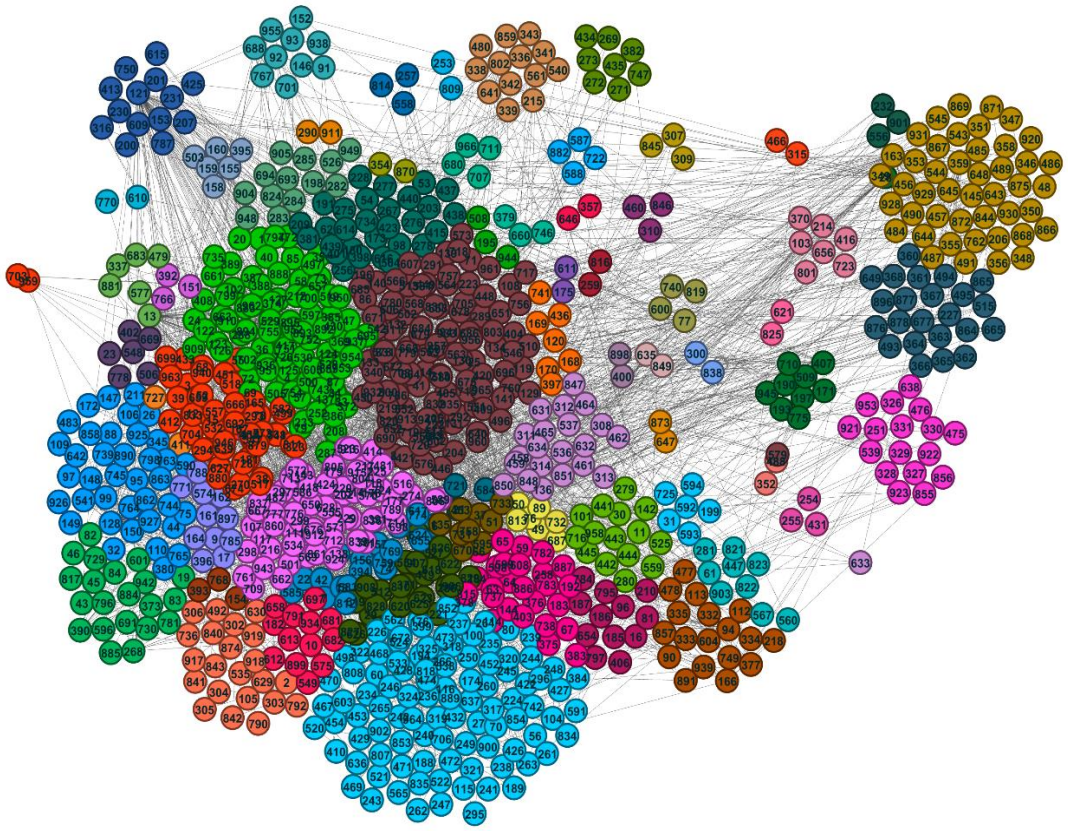

Figura 16. Red social con las 63 comunidades detectadas identificadas por colores. Generada en Gephi 9.0.2 bajo licencia de desarrollo común.

6. Análisis de la red social

El análisis de la red de investigadores, que incluye la detección de comunidades y la validación de éstas, es de gran importancia para comprender el estado y madurez de la red de colaboración científica. En esta sección se estudiarán las propiedades generales de la red social de investigadores, se analizarán algunas métricas de centralidad que ofrecen información interesante sobre los actores más relevantes de la red, se validarán las comunidades encontradas usando el algoritmo CLA-Net y finalmente se estudiarán los patrones de colaboración presentes en la red de colaboración científica. Ávila & Madariaga (2012) afirman que, con el fin de evaluar las particularidades de las

redes, el análisis de redes sociales ARS ha desarrollado cuatro mecanismos principales, los cuales consisten en definir las propiedades generales que tiene la red, analizar las características posicionales de los actores o nodos, identificar subagrupaciones dentro de la red y generar recursos analíticos para comprender este tipo de estructuras. Por lo tanto, se emplearán dichos mecanismos para evaluar la red de investigadores en el área de Ingeniería Industrial.

Tabla 8.

Propiedades estadísticas de la red social.

Propiedad	**Valor**
Número de autores	966
Número de artículos	1161
Artículos promedio por autor	3.90
Autores promedio por artículo	3.22
Colaboradores promedio	4.88
Componente más grande	966

En la tabla 8 se observan algunas propiedades estadísticas de la red social que permiten tener una idea general de la red de investigación previo al análisis.

6.1 Análisis de centralidad de la red de investigadores en Colombia

Mediante el uso de las métricas de centralidad, se realiza un análisis de importancia de los nodos pertenecientes a la red social de investigadores propuesta, pues se sabe que la importancia de un autor depende de su posición dentro de la red, además estas medidas reflejan la dependencia de la red debida a un individuo.

Para llevar a cabo este análisis de centralidad, se tienen en cuenta la totalidad de nodos (966), ya que todos los nodos están conectados en la red. Desde una perspectiva local se determina el valor del grado de centralidad ya que es una medida de influencia de un nodo respecto a sus vecinos más cercanos, mientras que en el ámbito global se determinan la centralidad de cercanía y de vector propio, valores que permiten identificar los nodos que están mejor ubicados para influir en toda la red de una forma muy rápida.

Otra de las medidas es la centralidad de intermediación, la cual puede analizarse tanto a nivel local como global, sin embargo, en este análisis se realiza de forma global con el fin de identificar los investigadores que permiten que el grafo tenga una mejor conectividad.

6.1.1 Grado de centralidad. En la red se determina inicialmente el valor del *Grado de Centralidad* en donde se encuentra que el autor Jairo Rafael Montoya Torres, Investigador Senior reconocido por Minciencias como par evaluador en el área de Ingeniería de producción y miembro de la universidad de la Sabana es el autor considerado de mayor importancia a nivel local con un grado de centralidad de 86, lo cual indica que a pesar de no pertenecer al área de conocimiento de Ingeniería Industrial sino de Ingeniería de Producción, se ha reunido para realizar colaboración científica con un gran número de autores pertenecientes a la red, entre ellos: Rene Alejandro Amaya, Elyn Solano Charris, Guisselle Adriana García Llinas, Diana Ramírez, Andrés Felipe Muñoz, Fernando Rafael González, los cuales en su mayoría han aportado a las líneas de investigación de Gestión de Operaciones y Logística. Esto refleja la similitud en varias de las líneas de investigación a las que aportan tanto los autores relacionados a Ingeniería Industrial como a Ingeniería de Producción. En la figura 17 se presentan los autores que obtuvieron los mayores valores de centralidad:

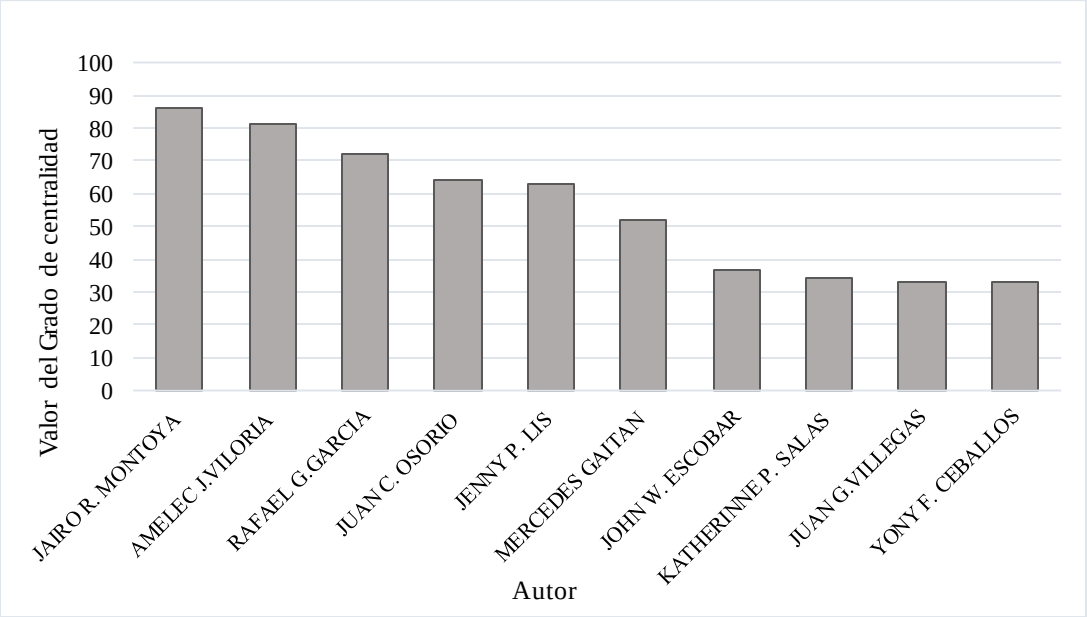

Figura 17. Grado de centralidad. Representación del valor de centralidad de los diez valores más altos para los autores de la red de investigación propuesta.

Autores como Amelec Jesús Viloria Silva y Rafael Guillermo García Cáceres con grados de centralidad de 81 y 72 respectivamente, también son considerados como autores de gran influencia a nivel local, pero en el área de Mejoramiento de procesos productivos, además se observa que estos autores agrupan en sus comunidades al mayor número de autores generando las comunidades de mayor tamaño en la red.

6.1.2 Centralidad de intermediación. De los 966 autores que conforman la red de investigación, tan solo el 26.4% son considerados como autores "puente", esto quiere decir que sirven como intermediarios para que otras parejas de investigadores en la red se puedan conectar, el 73.6% restante de los autores no sirven como punto de conexión para lograr la interacción entre autores. En la figura 18 se puede observar que la mayor concentración de autores tiene valores bajos de centralidad de intermediación, por lo que solo unos pocos son considerados puentes de conexión.

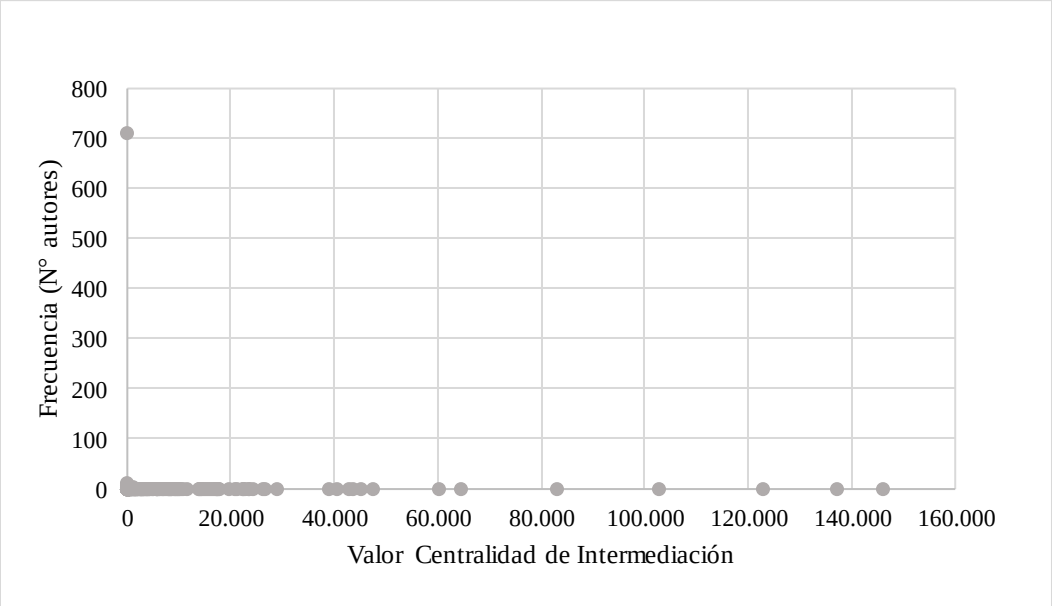

Figura 18. Distribución de frecuencia del valor de centralidad de intermediación. Indica los valores de la cantidad de conexiones que permite un nodo particular al actuar como "puente de conexión".

El investigador John Willmer Escobar Velásquez docente de la Universidad del Valle es el autor que se considera como el mejor "puente" de conexión entre otros autores de la red con un valor de centralidad de intermediación de 145 902 y un grado de 37. A pesar de que este autor no está dentro de los 5 autores con mayor valor de incidencia, si permite que se establezca el 15% de las posibles conexiones de la red.

Investigadores como Rafael Guillermo García Cáceres, Jairo Rafael Montoya, Dionicio Neira Rodado, Carlos Julio Vidal Holguín y Amelec Jesús Viloria Silva, también son considerados como intermediadores, ya que dentro de la red cumplen la función de ser puentes de conexión en temas de colaboración científica. Así mismo, este grupo de autores es indispensable para garantizar que la red tenga una buena conectividad y no este dividida en subgrafos.

6.1.3 Centralidad de cercanía. En temas de colaboración científica es de gran importancia determinar aquellos investigadores que se convierten en influenciadores para la

producción de conocimiento intelectual (Artículos Científicos) por su cercanía a otros autores de la red , es por esto que al analizar la red de investigadores propuesta se identifica que a nivel global de red los autores Rafael Guillermo García Cáceres, Dionicio Neira Rodado y Jairo Rafael Montoya, con valores de centralidad de cercanía de 0.34, 0.32 y 0.32 respectivamente, se encuentran más cerca de los demás investigadores de la red, lo cual quiere decir que al existir una menor cantidad de caminos entre cada uno de ellos y los demás autores de la red, les permite difundir con mayor facilidad y rapidez la información de un tema de investigación de interés en un área de conocimiento y a la vez son receptores de la información producida por otros investigadores de la red.

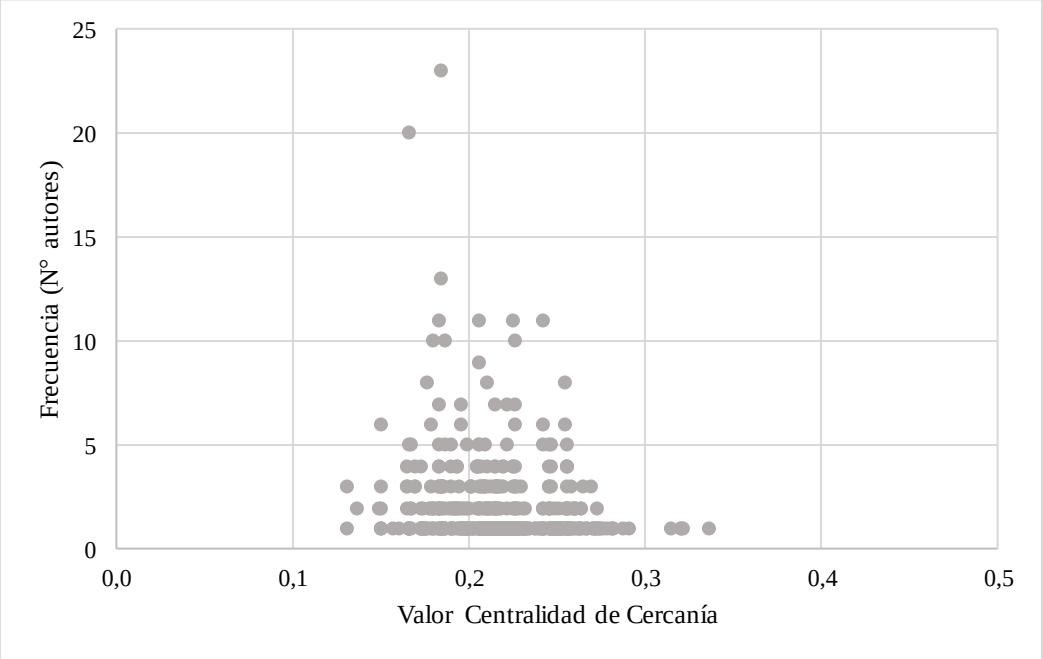

Figura 19. Distribución de frecuencia del valor de centralidad de cercanía. Indica la relación de cercanía de cada autor a los demás nodos de la red.

En la figura 19, se observa que los valores de cercanía de los autores a los demás miembros de la red oscilan entre 0.13 y 0.34, lo cual quiere decir que a pesar de que la información se puede

difundir de un lado a otro existe un número de caminos relativamente alto para poderla transmitir, sin embargo, más adelante se analizará este fenómeno de forma detallada con la propiedad "Small-world".

6.1.4 Centralidad de Vector Propio. El autor Amelec Jesus Viloria Silva miembro de la Universidad Sergio Arboleda es considerado como un autor influyente en la red de investigadores propuesta por la calidad de sus conexiones, pues al calcular el valor de centralidad de vector propio dicho autor obtuvo un valor de 1, por lo cual este investigador podría ser una persona idónea para liderar temas de investigación, tomar decisiones y portar información para generar conocimiento al interior de la red estudiada.

Con esta medida no solo se ratifica la importancia del autor en la red, sino que se refleja que es un autor que además de haber colaborado con su conocimiento a un gran número de investigadores (81), los investigadores con los que ha publicado también se encuentran muy bien relacionados entre sí, evidenciando que las conexiones generadas son de calidad.

A continuación, se observa la comunidad donde se encuentra el autor Amelec Silva (nodo 14) y otros autores con los que tiene conexiones, los cuales a su vez se encuentran relacionados entre sí. El tamaño de los nodos indica el valor del grado del nodo:

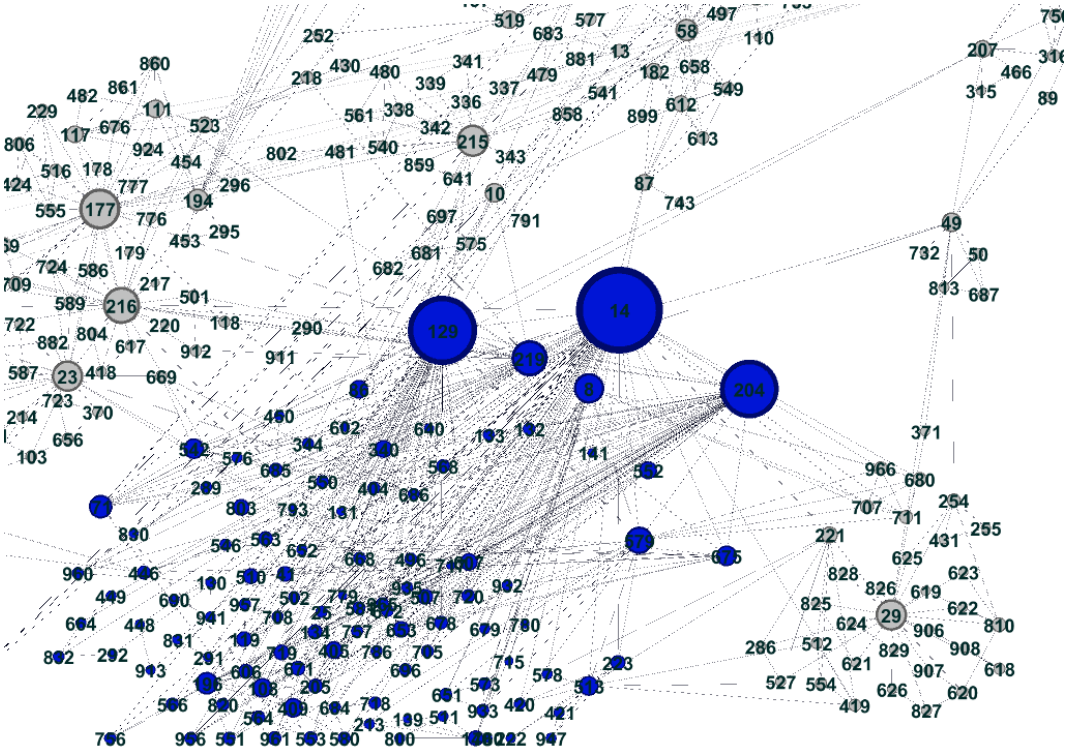

Figura 20. Representación del nodo con valor mayor de centralidad de vector propio y colaboradores con mayor grado dentro de su comunidad. Generada en Gephi 9.0.2 bajo licencia de desarrollo común.

El análisis de centralidad permite identificar los actores principales de la red, quienes ocupan roles importantes pues tienen la responsabilidad de mantenerla conectada y de regular el flujo de información y conocimiento. Ya que los actores principales identificados en el análisis de centralidad son investigadores Senior en su mayoría, se podría afirmar que la categoría de los investigadores tiene relación con los roles que estos ocupan en la red y que los autores que obtuvieron valores altos de centralidad gozan de un mayor reconocimiento como investigadores en la red.

6.2 Propiedades generales de la red social

Existen algunas propiedades de las redes sociales que han sido objeto de interés en el análisis de redes sociales a lo largo de los años, algunas de estas propiedades son: “Small-world”, ley de

potencia en la distribución de los grados y la transitividad. Estos conceptos fueron introducidos anteriormente y a continuación se analizan estas propiedades para la red de investigadores en Colombia.

6.2.1 Small-world problem. Para determinar los grados de separación promedio entre dos individuos se calcularon los caminos más cortos entre cada par de nodos de la red de investigadores, para obtener el valor promedio se dividió la suma de la distancia más corta para cada par de nodos entre el número total de autores en la red. En promedio una pareja de autores de la red está a 5.036 grados de separación, la propiedad de "Small-world" dice que en promedio existen 6 grados de separación entre dos personas cualquiera en una red social, por lo que se puede evidenciar esta propiedad de manera aproximada en la red de investigadores de Colombia.

6.2.2 Distribución de ley de potencia de los grados. Para probar que la distribución de los grados de la red de investigadores en Colombia sigue una ley de potencia, como otras redes sociales, se utilizó una función incorporada en el paquete "igraph" de R Studio ®. Esta función se basa en el trabajo de Clauset, Shalizi y Newman (2009) quienes determinaron un conjunto de técnicas para comprobar que un conjunto de datos empíricos del grado de los nodos en una red social sigue una ley de potencia. El valor-p obtenido para la prueba de Kolmogorov-Smirnov fue de 0.15, al ser mayor que el nivel de significancia establecido de 0.05, se acepta la hipótesis nula de que la frecuencia de los grados de los autores de la red se distribuye según una ley de potencia con exponente de 2.64. Con esto se comprueba la propiedad de la ley de potencia para la distribución del grado de los nodos en la red social de investigadores. En el contexto de la red estudiada esto quiere decir que son unos pocos autores en la red quienes hacen la mayor contribución en el número de colaboradores por autor, evidenciando que hay un pequeño grupo de

autores que son más influyentes y están más conectados con la red. Este fenómeno se discutirá también desde la perspectiva de la bibliometría en el análisis de patrones de colaboración científica.

6.2.3 Transitividad. La transitividad de la red, es decir, la probabilidad de que dos coautores de un autor sean coautores entre ellos es de 0.2049, este valor parece un poco bajo, ya que se espera que los coautores de un mismo autor trabajen en líneas de investigación similares o pertenezcan a la misma institución y por tanto podrían colaborar también. En algunos casos se evidencia que los investigadores senior tienden a trabajar por separado con diferentes coautores y posteriormente realizan un trabajo conjunto con coautores previos, fortaleciendo así la colaboración dentro de su comunidad. Los investigadores senior juegan un rol importante dentro de sus comunidades ya que pueden incentivar la colaboración entre otros investigadores menos experimentados con los que han trabajado en el pasado mejorando la conectividad y resiliencia de la red de investigación. La baja transitividad de la red, más que un problema, supone un desafío para aprovechar el potencial de colaboración que existe en la red y en las comunidades para los autores que tienen coautores en común, pero no se conocen o han trabajado juntos.

6.2.4 Detección de comunidades. La existencia de estructura de comunidad dentro de una red social resulta uno de los principales intereses en el análisis de redes sociales, ya que las comunidades que se forman no son producto del azar y por ello existen atributos en común entre quienes deciden asociarse, los cuales permiten además identificar cómo estos grupos interactúan y evolucionan en el tiempo. Por esto se intentó identificar los atributos similares para algunas de las comunidades obtenidas dentro de la red de investigadores, ya que al encontrar dichos atributos es posible validar la significancia de la comunidad formada por el algoritmo. Para el caso de una red de colaboración científica estos atributos proveen información útil sobre la forma en que se

organizan para colaborar los autores y esto permite concluir sobre el estado de la colaboración científica en la red.

Tras la implementación del algoritmo se detectaron 63 comunidades, con el fin de validar las comunidades obtenidas, se eligieron cinco comunidades al azar y se analizaron los atributos para cada uno de los miembros de la comunidad con el fin de encontrar atributos similares que expliquen la razón de su fuerte asociación. A continuación, se detalla el análisis de las comunidades seleccionadas.

Comunidad 58: Esta comunidad está formada por dos miembros, Koray Dogan (411) y Marc Goetschalckx (727), como se observa en la figura 21.

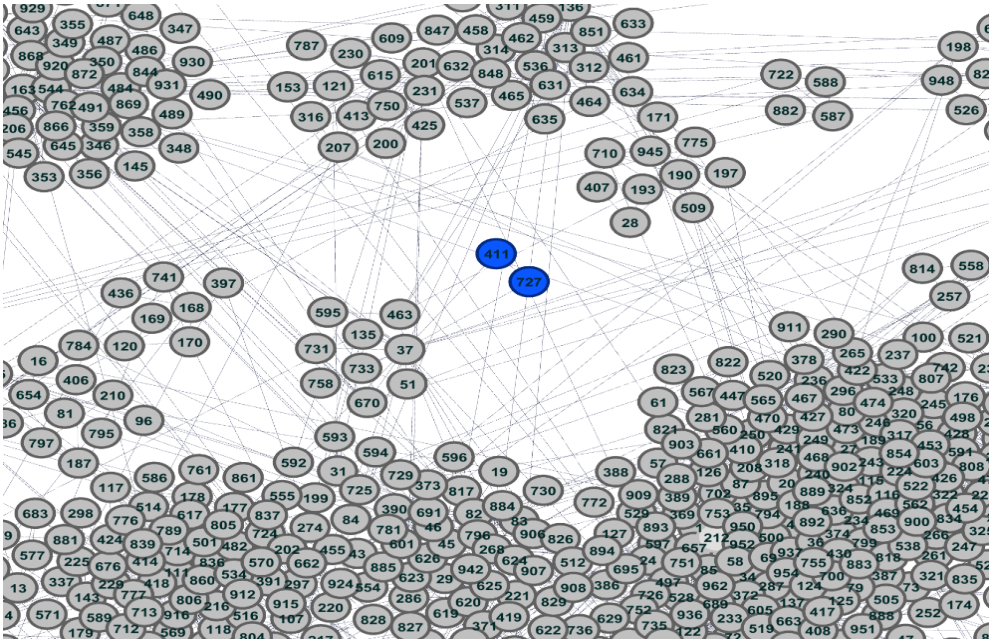

Figura 21. Comunidad 58 de color azul. Generada en Gephi 9.0.2 bajo licencia de desarrollo común.

Estos dos autores no se encuentran en la lista de pares evaluadores de Minciencias y tampoco cuentan con un perfil público de CvLAC. Se encontró que en su perfil de LinkedIn registran una

afiliación con Georgia Institute of Technology. Estos autores hacen parte de la red ya que colaboraron con el Investigador Asociado Carlos Julio Vidal Holguín quien realizó sus estudios de maestría y doctorado en Georgia Institute of Technology, sin embargo, el par evaluador fue incluido en una comunidad diferente. Esto se puede explicar al analizar a cada uno de los autores. Los dos autores asignados a la comunidad son extranjeros, por este motivo sus conexiones con la red son débiles, pues no presentan conexiones con otros miembros de la red en la base de datos. Por otro lado, el investigador Carlos Julio Vidal Holguín es un autor que tiene fuertes conexiones en su comunidad de investigación, el algoritmo detectó que los autores extranjeros solo tienen relación con un autor de la comunidad y, por ende, decidió asignarlos a otra comunidad. Es posible entonces pensar que los demás autores extranjeros dentro de la red que registran pocas conexiones con el resto de la red serán asignados a comunidades separadas, como sucedió en este caso.

Comunidad 22: Esta comunidad está formada por seis autores, como se observa en la figura 22.

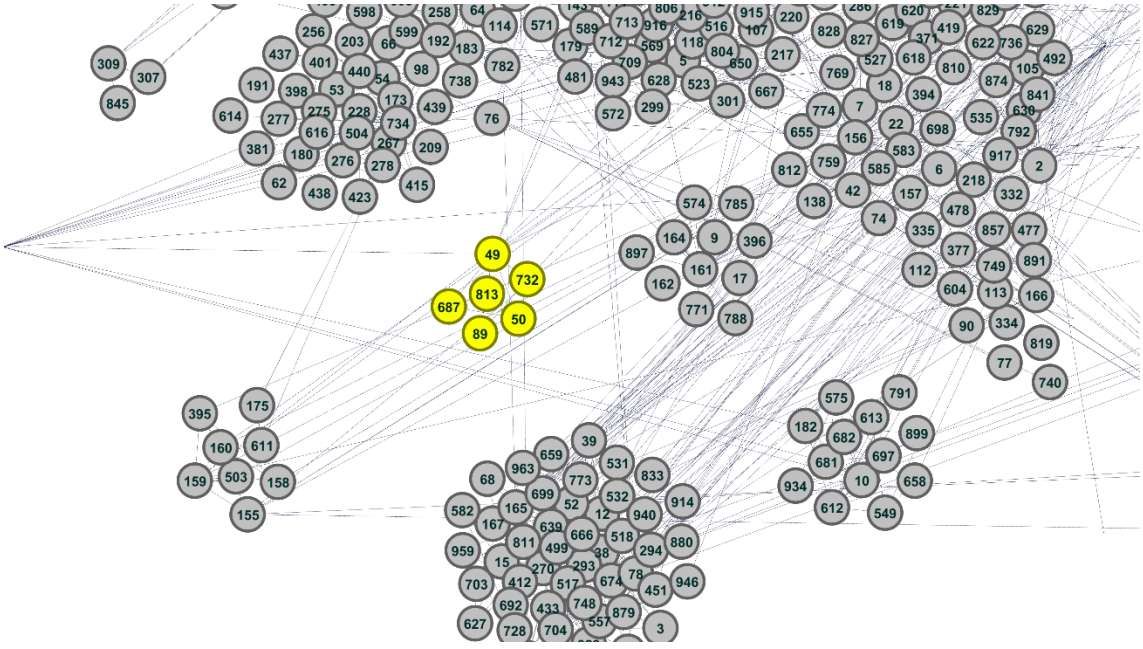

Figura 22. Comunidad 22 de color amarillo. Generada en Gephi 9.0.2 bajo licencia de desarrollo común.

De estos autores solo uno es un par evaluador reconocido por Minciencias, el investigador Asociado Daniel Alfonso Mendoza Casseres. Los demás autores según la información en su CvLAC ya no se encuentran investigando activamente. Por consiguiente, el par evaluador es el líder de la comunidad, pues es quien mantiene a su comunidad conectada y además el nodo fronterizo, es decir el autor que conecta a la comunidad con otras comunidades. Esta comunidad es el reflejo de una colaboración científica que ocurrió durante 2015 y 2016. Tiempo en el que los cinco coautores se encontraban terminando sus estudios de pregrado. Por lo que se presume que el par evaluador era su profesor y dirigió sus trabajos de investigación. No es preciso asignar el par evaluador a esta comunidad ya que este ha colaborado con otros pares de la red de forma consistente en los últimos años. Aquí se evidencia un aspecto importante a la hora de representar redes de colaboración científica, la ventana de tiempo de las colaboraciones puede afectar el desempeño en la detección de las comunidades, ya que una comunidad puede reflejar una red de colaboración del pasado como en este caso. Si se quisiera tener una mirada de la colaboración en el largo plazo sería recomendable usar pesos en la red para definir la fuerza de las relaciones de colaboración y que se reflejen las comunidades de colaboración que han sido más fuertes a lo largo del tiempo. Dado que el algoritmo que se pretendía implementar es para redes sociales sin pesos y que no se superponen no se consideraron dichos elementos a la hora de construir la red de investigadores, sin embargo, sería pertinente considerar estos elementos en el futuro al analizar una red de colaboración científica en Colombia, pues permitiría reflejar mejor en las comunidades obtenidas la colaboración recurrente entre investigadores.

Comunidad 9: Esta comunidad está formada por quince autores, en la figura 23 se observa la comunidad identificada con color verde en la red.

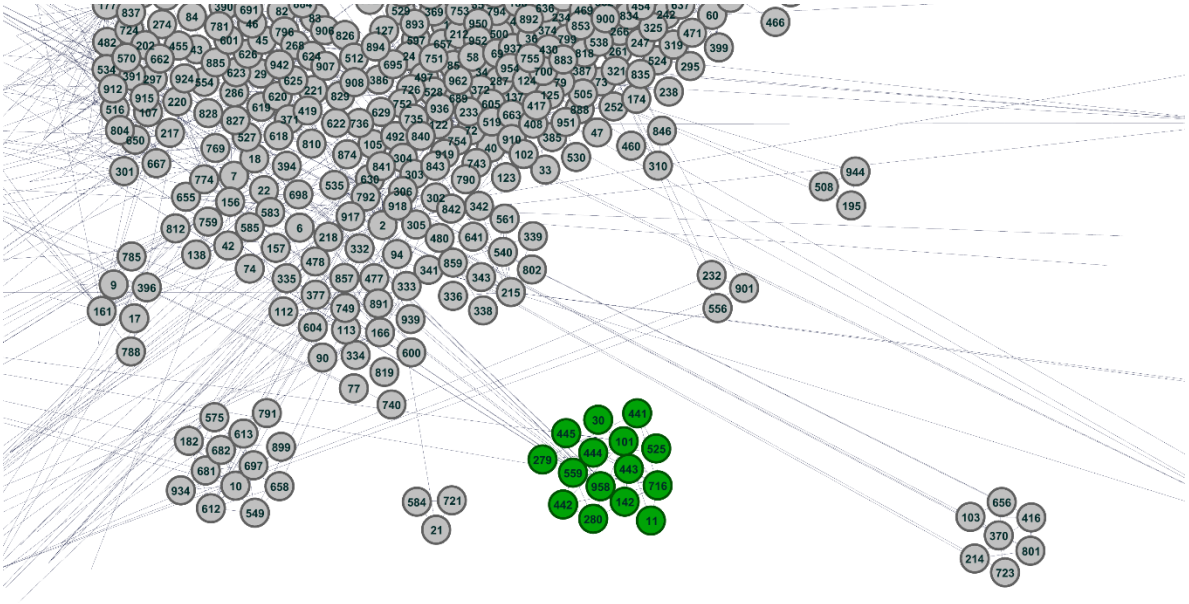

Figura 23. Comunidad 9 de color verde. Generada en Gephi 9.0.2 bajo licencia de desarrollo común.

Es una comunidad bastante particular ya que está formada por investigadores en diferentes áreas, hay investigadores de Ingeniería Mecánica, Ingeniería Química, Ingeniería Eléctrica e Ingeniería Industrial. El nodo líder de esta comunidad es la Investigadora Junior Yulineth Del Carmen Cárdenas Escorcia, quien es el único par evaluador de la comunidad cuya área de actuación es Ingeniería Industrial. Por otro lado, los nodos fronterizos son los autores Samir Francisco Umaña Ibañez y Jorge Iván Silva Ortega. Estos autores presentan el mayor número de conexiones con otras comunidades, por ende, son los mejores autores para transmitir información entre su comunidad y otras comunidades, así como incentivar la colaboración entre autores de su comunidad y otras comunidades de la red.

En esta comunidad se evidencia colaboración interdisciplinaria, ya que la línea de investigación en común es eficiencia energética y cada ingeniería puede contribuir desde distintas perspectivas a este tema de investigación. Se evidencia además que todos los autores tienen afiliación con

instituciones universitarias de la costa colombiana, por lo que existe también colaboración interinstitucional entre la Universidad del Atlántico y Corporación Universidad de la Costa, solo se encontró una autora extranjera en esta comunidad, la autora es cubana y es investigadora reconocida por Minciencias.

Comunidad 37: Esta comunidad está formada por tres autores, Cedrick Beler, Majda Lachhab, Thierry Coudert, como se observa en la figura 24.

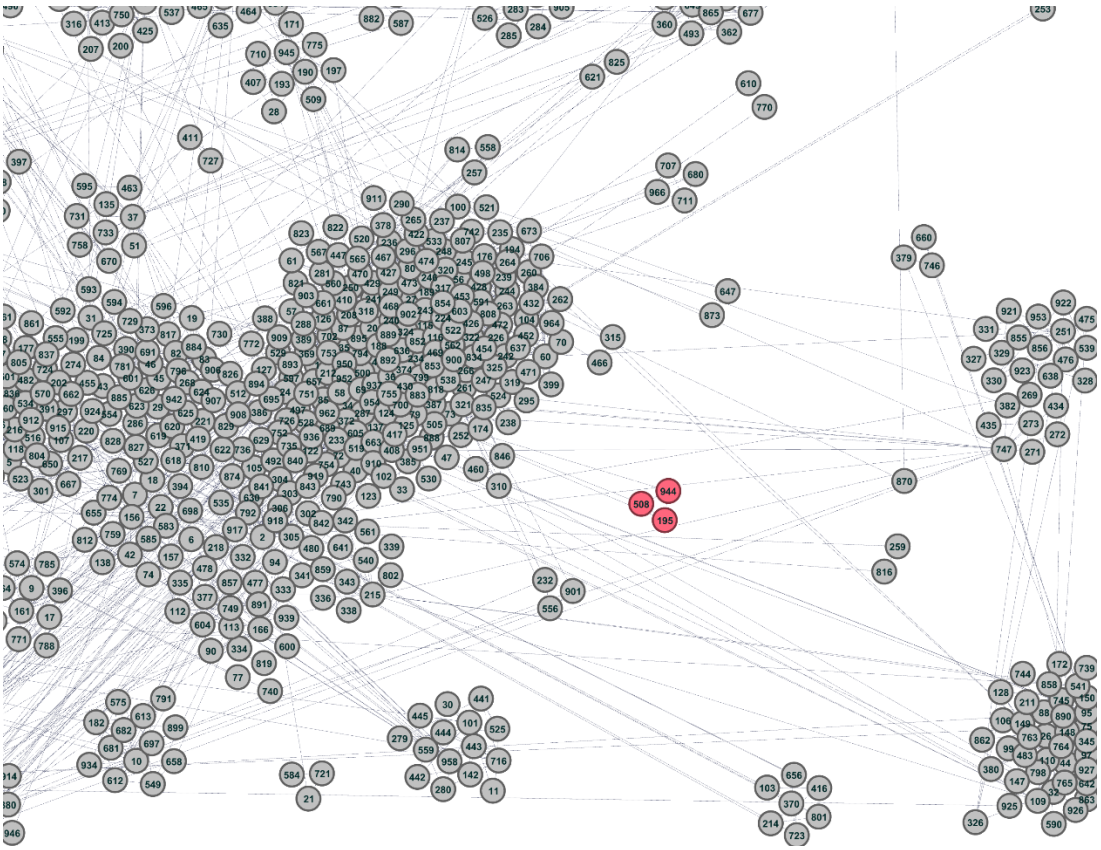

Figura 24. Comunidad 37 de color rosado. Generada en Gephi 9.0.2 bajo licencia de desarrollo común.

Estos tres autores son extranjeros y colaboraron con la investigadora Elyn Lizeth Solano Charris, sin embargo, fueron asignados a una comunidad separada ya que no presentan conexiones fuertes con el resto de la red. El par evaluador que colaboró con estos autores fue asignado a otra

comunidad pues presenta mayor colaboración con otros investigadores colombianos, aquí se evidencia un comportamiento importante del algoritmo cuando identifica individuos con escasas conexiones dentro de la red, este decide aislarlos en una nueva comunidad ya que no los considera parte de las comunidades existentes, de hecho en el resto de la red se evidencian otras comunidades de pequeño tamaño como esta que posiblemente corresponden a autores externos que no se asignaron a las comunidades de Colombia como sucedió anteriormente con la comunidad 58.

Comunidad 23: Esta es una comunidad de investigación con 29 autores como se observa en la figura 25.

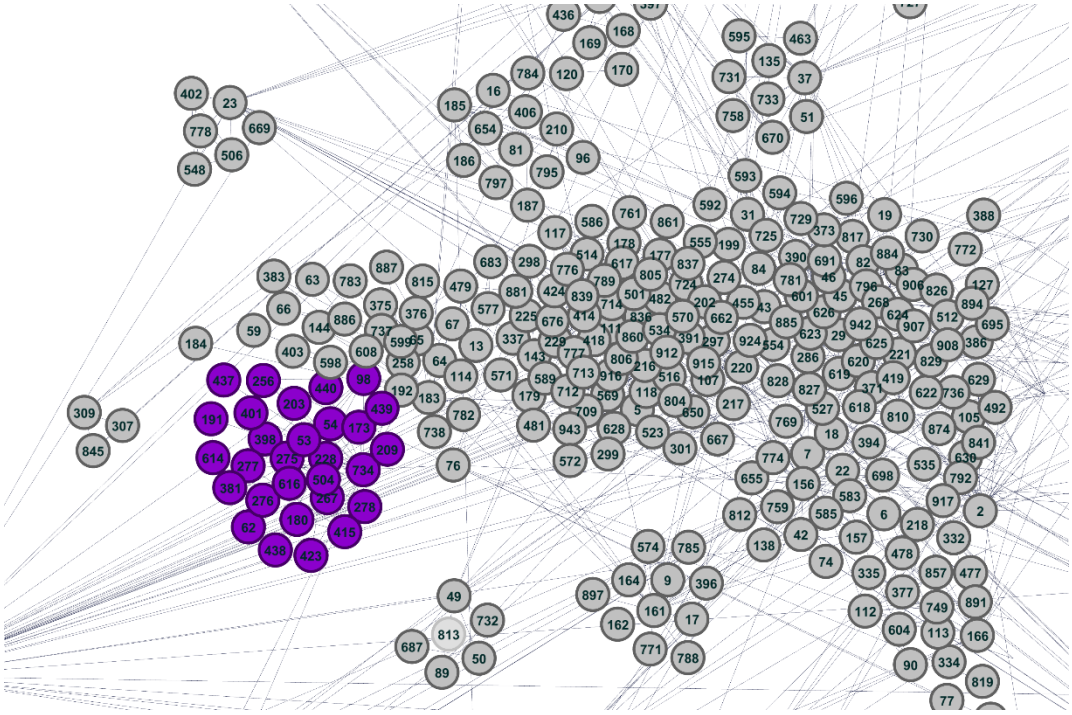

Figura 25. Comunidad 23 de color morado. Generada en Gephi 9.0.2 bajo licencia de desarrollo común.

Esta es una comunidad de investigación en simulación y dinámica de sistemas, en la comunidad se encuentran 9 pares evaluadores, quienes tienen estudios de Ingeniería de Sistemas, sin embargo, el líder de esta comunidad es el Investigador Asociado Yony Fernando Ceballos, quien pertenece

al grupo de investigación Ingeniería y Sociedad de la Universidad de Antioquia. En este grupo el autor colabora con otros investigadores con estudios de Ingeniería Industrial, por este motivo este investigador tiene gran conexión con la red de ingeniería industrial y actúa también como el nodo fronterizo de la comunidad, siendo el mejor autor para transmitir información y generar colaboraciones entre autores de su comunidad y otras comunidades.

Se observa que en la comunidad hay dos tipos de colaboración científica, una es producto de la colaboración entre el investigador Yony Fernando Ceballos y estudiantes de pregrado de la Universidad de Antioquia que realizan su trabajo de grado bajo su dirección y el otro tipo de colaboración que se evidencia es la colaboración entre pares evaluadores, en donde se encontró que todos los investigadores tienen alguna relación con la Universidad Nacional de Colombia Sede Medellín, algunos pares realizaron estudios de pregrado, maestría o doctorado durante el mismo periodo que los realizó el investigador Yony Fernando Ceballos. Por este motivo se presume que la colaboración con otros pares de la red se da principalmente entre autores que fueron compañeros de maestría o doctorado.

Esta comunidad permite observar un fenómeno presente a lo largo de la red, las comunidades tendieron a organizarse en torno a un investigador líder en la línea de investigación y un gran número de las colaboraciones en las comunidades son producto de tesis de pregrado y maestría. Esto es natural ya que una gran parte de los investigadores, realizan también actividades de docencia y por tanto colaboran constantemente con estudiantes en grupos de investigación.

Del anterior análisis se concluye que las comunidades de investigación tienden a dividirse por líneas de investigación dentro del área de Ingeniería Industrial. En la tabla 9 se puede observar el nodo líder de cada comunidad.

Tabla 9.

Líderes investigadores en cada una de las comunidades

Comunidad	Miembros	Líder	Línea de investigación
1	85	Jairo Rafael Montoya Torres	Optimización de operaciones logísticas y de producción
2	22	Diego Fernando Manotas Duque	Ingeniería financiera y gestión de riesgo
3	48	Juan Pablo Caballero Villalobos	Programación de la producción
4	67	Katherinne Paola Salas Navarro	Gestión industrial
5	18	Juan Guillermo Villegas Ramirez	Investigación de operaciones
6	110	Amelec Jesus Viloria Silva	Mejoramiento en la cadena de suministro
7	11	Juan Jose Bravo Bastidas	Optimización de cadenas de abastecimiento
8	13	Alfonso Rafael Romero Conrado	Diseño experimental
9	15	Yulineth Del Carmen Cardenas Escorcia	Sistemas de gestión de energía
10	6	Fabian Sanchez Sanchez	Producción y logística
11	12	Lindsay Alvarez Pomar	Productividad y competitividad de la industria colombiana
12	21	Nubia Milena Velasco Rodriguez	Logística de transporte
13	3	Andres Felipe Osorio Muriel	Optimización de cadenas de suministro
14	6	Andres Felipe Porto Solano	Competitividad e Innovación
15	38	German Mendez Giraldo	Dinámica del sistema
16	99	Rafael Guillermo Garcia Caceres	Logística de transporte
17	10	Luis Alfredo Paipa Galeano	Mejoramiento continuo de procesos
18	24	Adel Alfonso Mendoza Mendoza	Analisis envolvente de datos
19	6	Carlos Alberto Castro Zuluaga	Gestión de producción
20	9	Carlos Julio Vidal Holguin	Logística y producción
21	45	Juan Carlos Osorio Gomez	Modelamiento de sistemas productivos

22	6	Daniel Alfonso Mendoza Casseres	Optimización y simulación
23	29	Yony Fernando Ceballos	Simulación y dinámica de sistemas
24	26	Leonardo Rivera Cadavid	Sistemas avanzados de manufactura
25	9	Alba Ligia Lopez Rodriguez	Productividad y competitividad de la industria colombiana
26	5	Elena Valentina Gutierrez Gutierrez	Gestión de sistemas de salud
27	18	Jorge Andres Alvarado Valencia	Simulación y logística
28	10	Gabriel Mauricio Zambrano Rey	Inteligencia artificial
29	7	Bertha Ines Villalobos Toro	Gestión industrial
30	7	Juan Sebastian Jaen Posada	Dinámica de sistemas
31	14	Mario Cesar Velez Gallego	Gestión de producción
32	24	John Willmer Escobar Velasquez	Algoritmos metaheuristicos para problemas de localización y ruteo.
33	3	Edgar Ojeda Camargo	Gestión tecnológica
34	3	Juan Esteban Muriel Villegas	Investigación de operaciones
35	6	Caroline Prodhon	Logística de transporte
36	2	Karen Rocio Martinez Lancheros	Dinámica de sistemas
37	3	Cedrick Beler	Sistemas de información
38	13	Helga Patricia Bermeo Andrade	Gestión de la innovación y la tecnología
39	13	Miguel Angel Ortiz Barrios	Gestión industrial
40	23	Pablo Cesar Manyoma Velasquez	Gestión de operaciones
41	3	Maria Alejandra Martinez Uribe	Planeación de la cadena de abastecimiento
42	17	Johanna Trujillo Diaz	Cadena de abastecimiento y transportes
43	2	Ricardo Alberto Perez	Gestión de la tecnologia y la innovación
44	3	Roberto Jose Herrera Acosta	Control estadístico de la calidad
45	3	Daniele Ferone	Métodos de optimización
46	2	Andres Mauricio Villegas	Logística de aprovisionamiento

47	8	William Javier Guerrero Rueda	Logística de transporte
48	2	Eliana Pardo Mora	Desarrollo empresarial
49	2	Alfredo Figueroa Pastrana	Gestión de la calidad
50	3	Jhon Jairo Santa Chavez	Matemática aplicada
51	3	Jose Adalberto Soto Mejia	Análisis de medidas de eficiencia y productividad
52	2	Diana Ochoa	Simulación de inventarios
53	2	Diana Carolina Arango	Modelamiento de sistemas productivos
54	2	Luz Carime Urbano Guerrero	Modelamiento de sistemas productivos
55	2	Cristian Andres Parra Calderon	Gestión del riesgo
56	3	Cristhian David Pinto Anaya	Estrategias de administración
57	2	Laura Maria Obando Bobadilla	Competitividad empresarial
58	2	Koray Dogan	Optimización de cadenas de abastecimiento
59	4	Astelio De Jesus Silvera Sarmiento	Gestión e innovación empresarial
60	2	Mario Fernando Acosta Rios	Enseñanza de la investigación de operaciones
61	2	Allison Zarate Arroyo	Eficiencia en la calidad del servicio
62	2	Maria Alejandra Coronado	Sistemas productivos
63	4	Carlos Albeiro Pacheco Bustos	Producción y manejo de residuos

Dentro de las distintas comunidades se puede evidenciar colaboración inter-institucional entre universidades de Colombia y con universidades del exterior, sin embargo, por lo general los autores tienen estudios en una universidad en común, es decir, los autores pueden tener afiliaciones a universidades diferentes en la actualidad, pero sus colaboraciones se deben, en la mayoría de los casos, a que realizaron estudios de pregrado o posgrado en una misma universidad, lugar y tiempo en el que probablemente inició su relación de colaboración.

Existe entonces aún un gran desafío por incentivar la colaboración entre comunidades que tienen líneas de investigación similares, pues esto llevaría a unificar las comunidades de una determinada línea de investigación y fortalecer la productividad científica a través de la cooperación entre investigadores con distintas trayectorias y temas de experticia. En la siguiente sección se estudiarán otros patrones de colaboración para la red de investigación.

6.3 Patrones de colaboración científica

Uno de los principales motivos por los que las redes científicas son especialmente populares en el análisis de redes sociales es debido al interés de analizar patrones de colaboración entre investigadores pertenecientes a determinadas comunidades científicas y con esto entender cómo aumentar la productividad científica, el impacto de los productos de investigación y la difusión de conocimiento en la red. A continuación, se analizan algunas propiedades interesantes de la red de investigación y sus comunidades, así como sus patrones de colaboración científica.

6.3.1 Colaboradores promedio de un autor en la red. Para determinar el valor promedio de coautores para un autor de la red, se calculó la razón entre la sumatoria del grado de todos los nodos de la red que es 4714 y el número total de nodos que componen la red, 966, con lo que se obtuvo un valor de 4.88. Por lo tanto se puede afirmar que un investigador de la red se relaciona en promedio con aproximadamente 5 investigadores para producir artículos para el área de Ingeniería Industrial. Es necesario subrayar que se evidencia la colaboración con los mismos autores en diferentes artículos, por lo que en general un autor trabaja de forma reiterada con cinco autores.

A continuación, se presenta la distribución de frecuencia del número de colaboradores por autor (grado del nodo) en la red.

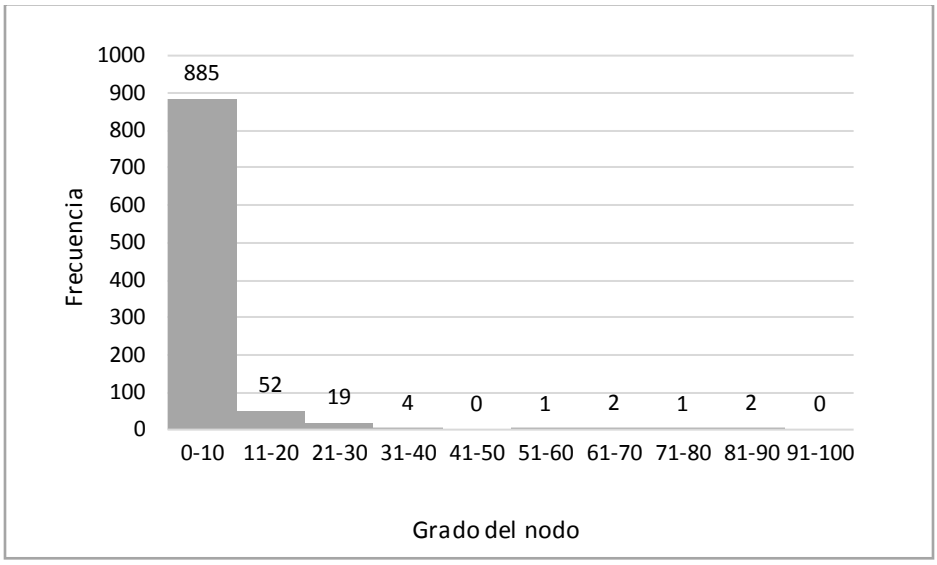

Figura 26. Histograma del grado de los nodos de la red social. Representa la distribución del valor de incidencia de los nodos en la red.

Como se observa en la figura 26, un 91.6% de los autores tienen entre 1 y 10 coautores. En la red se evidencia que 85 autores tienen el menor número de coautores por autor de la red, ya que estos han publicado en coautoría solo con un autor. Muchos de estos autores son estudiantes de pregrado que publicaron un artículo junto con un profesor como resultado de su trabajo de investigación para optar por el título de pregrado, así mismo se encontraron autores extranjeros que no pertenecen a la red de investigación de Colombia, pero trabajaron con algún par evaluador de la red, por lo que tienen un único coautor.

6.3.2 Co-publicaciones promedio de un autor en la red. El número promedio de co-publicaciones por autor para la red de investigadores estudiada es de 3.9. El autor con mayor número de co-publicaciones es también el investigador Senior Jairo Rafael Montoya Torres. Este

autor registra el mayor número de colaboradores y publicaciones en la red, esto da un indicio de su influencia en el área de Ingeniería Industrial.

El número mínimo de co-publicaciones para un autor de la red es 1, en este caso son autores que han trabajo con un par evaluador alguna vez, pero no tienen relación con otros pares evaluadores o miembros de la red además del par evaluador y los coautores de la co-publicación. Debido a que solo se buscó la información de los pares evaluadores, la visión de la red con respecto a esos autores podría ser incompleta, sin embargo, dichos autores son de baja influencia para la red ya que en su mayoría son estudiantes de pregrado que no cuentan con perfil de CvLAC o autores extranjeros.

Un importante patrón que se evidencia en las redes de colaboración científica es que la distribución de frecuencia del número de publicaciones por autor sigue una distribución de cola larga. (Lotka, 1926). A continuación, se ilustra la distribución de frecuencia de las publicaciones por autor en la red de investigadores de Colombia:

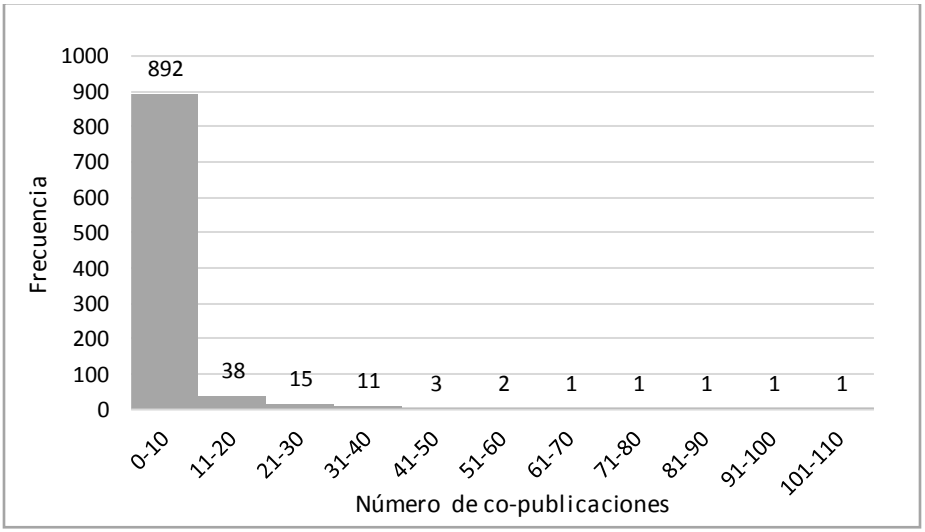

Figura 27. Distribución de frecuencia de las co-publicaciones por autor. Representa el número de co-publicaciones realizadas por los autores en la red.

Como se observa en la Figura 27, este patrón también está presente en la red de investigación de pares evaluadores reconocidos por Minciencias. Indicando que en las redes de colaboración normalmente un pequeño número de autores tienen el mayor número de co-publicaciones, esto al igual que el número de co-autores por autor, pudiera ser explicado de manera diferente para cada red en particular. Para el caso de la presente red se observó que los autores con mayor número de coautores y publicaciones son directores de grupos de investigación, por lo que se puede presumir que tienen participación en un gran número de los trabajos que se desarrollan dentro de los grupos debido a su rol.

En bibliometría existe una famosa ley propuesta por Lotka (1926), el autor afirma que, en las redes científicas, sin importar la disciplina o área, unos pocos autores publican una gran parte del total de publicaciones, esto quiere decir, en cada red un pequeño grupo de autores son quienes hacen la mayor parte de las contribuciones al área. Es necesario recalcar que en bibliometría, se acepta que la productividad científica de los investigadores sea medida usando el número de sus publicaciones, claramente alguien puede discrepar ya que es difícil garantizar que una mayor cantidad de publicaciones equivalen a un mayor impacto o contribución a la ciencia pues esto realmente depende de la calidad. Sin embargo, partiendo de esta premisa, la ley de Lotka establece que es posible predecir para una red cuántos autores han publicado un determinado número de artículos, ya que la distribución de frecuencia del número de publicaciones por autor sigue una distribución exponencial, esto ya se corroboró anteriormente cuando se ilustró la distribución de co-publicaciones para la red de investigadores colombianos. De acuerdo a Lotka (1926) el número de autores, An, que publican n trabajos sobre una materia es inversamente proporcional al cuadrado de n, la ley se formula así:

$$A_n = \frac{A_1}{n^2} \quad (22)$$

A continuación, se usa la ley de Lotka para verificar si es posible predecir el número de autores que publican determinado número de artículos en la red de investigadores de Colombia.

Tabla 10.

Número de autores según ley de Lotka

Número de artículos	**Predicción número de autores que publican los artículos**	**Número real de autores**
1	-	568
2	$A_2 = 568/2^2 = 142$	135
3	$A_3 = 568/3^2 = 63$	64
4	$A_4 = 568/4^2 = 35.5$	43
5	$A_5 = 568/5^2 = 22.72$	26
10	$A_{10} = 568/10^2 = 6$	6
54	$A_{54} = 568/54^2 = 0.195$	1
75	$A_{75} = 568/75^2 = 0.101$	0
109	$A_{109} = 568/109^2 = 0.048$	1

Como se puede evidenciar la ley ofrece estimados buenos del número de autores que existen hasta determinado n, sin embargo, cuando n se hace más grande el estimado comienza a ser cercano a 0 y si la cola de la distribución es muy larga es probable que aún en valores de n altos se encuentren autores y la fórmula indique valores cercanos a cero con esa cantidad de artículos, como es el caso del autor con 109 artículos en la red de investigadores.

A pesar de esto, la ley de Lotka permite destacar un importante patrón en las redes de colaboración científica, la mayor parte de las contribuciones a la productividad científica de la red

son hechas por un pequeño número de autores, esto es muy relevante en bibliometría, ya que significa que solo es necesario leer una pequeña parte de los autores de un área para conocer la mayoría de los avances y logros científicos que hay en ella. Esta ley está relacionada con el principio de Pareto en economía, evidenciando que es un fenómeno presente en distintas áreas, incluyendo la bibliometría.

6.3.3 Distancia típica entre dos investigadores. La distancia típica entre dos investigadores para la red es de 5.04. Esto quiere decir que dos autores en la red pueden contactarse a través 5 intermediarios que son su coautor y los coautores de sus coautores. Este número muestra que no existe una gran separación entre los autores y por tanto es fácil contactar a otro autor con el que nunca se ha trabajado para generar relaciones de colaboración en el futuro si así se desea.

6.3.4 Cadenas de referidos. La distancia mínima entre dos autores, que son los conocidos en común a través de los cuales se podría establecer contacto con un autor con el que nunca se ha trabajado es también otro patrón de interés, a esto se le conoce como "cadena de referidos".

Kautz, H., Selman, B., y Shah, M. (1997) afirman que las cadenas de referidos pueden ser herramientas de búsqueda útiles para acceder a información confiable sobre una temática que no dominamos directamente de un experto. Las referencias ofrecen juicios de calidad, ya que conocen al referido y al usar dicha persona como intermediario hay una mayor probabilidad de recibir respuesta. Esto no sucede cuando se usan otras herramientas de búsqueda, como motores de búsqueda web.

En el mundo de la colaboración científica esto es de gran importancia, ya que en ocasiones cuando se están formando los equipos para llevar a cabo un proyecto se requieren equipos interdisciplinarios con personas que tengan fortalezas en distintos temas de investigación, la cadena de referidos permite identificar a qué colaboradores o conocidos se deben contactar para establecer contacto con un experto que se necesita en el proyecto o para identificar a través de un coautor otro autor que cumple con el perfil necesario para el proyecto.

Este es solo un ejemplo de las muchas situaciones en las que la cadena de referidos podría ser útil, el análisis de redes sociales es el que hace posible localizar a otros autores e identificar el mejor camino para contactarlos. Es sencillo encontrar el camino más corto entre dos autores utilizando un algoritmo como Breadth-First Search.

Para la red de investigadores de Colombia dos autores fueron elegidos aleatoriamente y se construyó el camino más corto entre estos. Los autores seleccionados fueron el nodo 35 y 868. En la figura 28 se observa la cadena de referidos entre los dos nodos seleccionados:

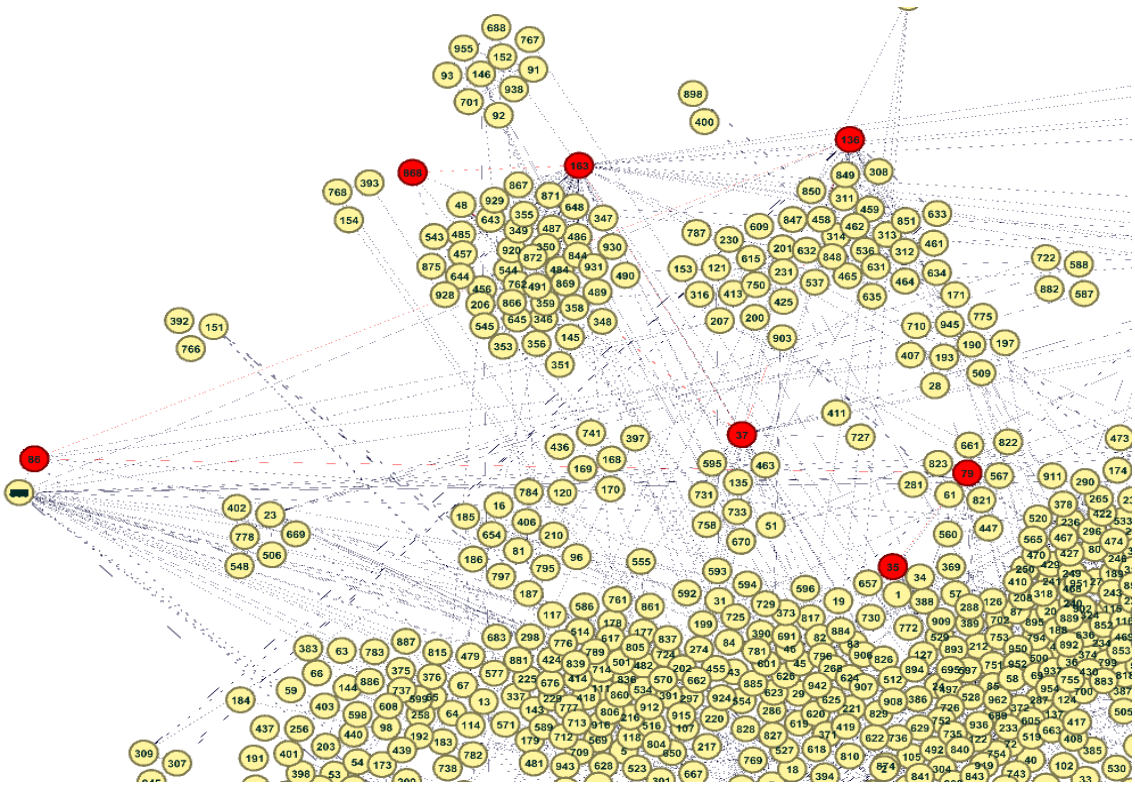

Figura 28. Cadena de referidos entre el autor Carlos Daniel Paternina y María Isabel Hernández Santibáñez. Generado en Gephi 9.0.2 bajo licencia de desarrollo común

El nodo 35 corresponde al autor Carlos Daniel Paternina Arboleda y el nodo 868 a la autora María Isabel Hernández Santibáñez. El camino más corto entre los dos investigadores es de 6, esto significa que el autor Carlos Daniel Paternina requiere de 5 intermediarios para contactar a la autora María Isabel Hernández Santibáñez.

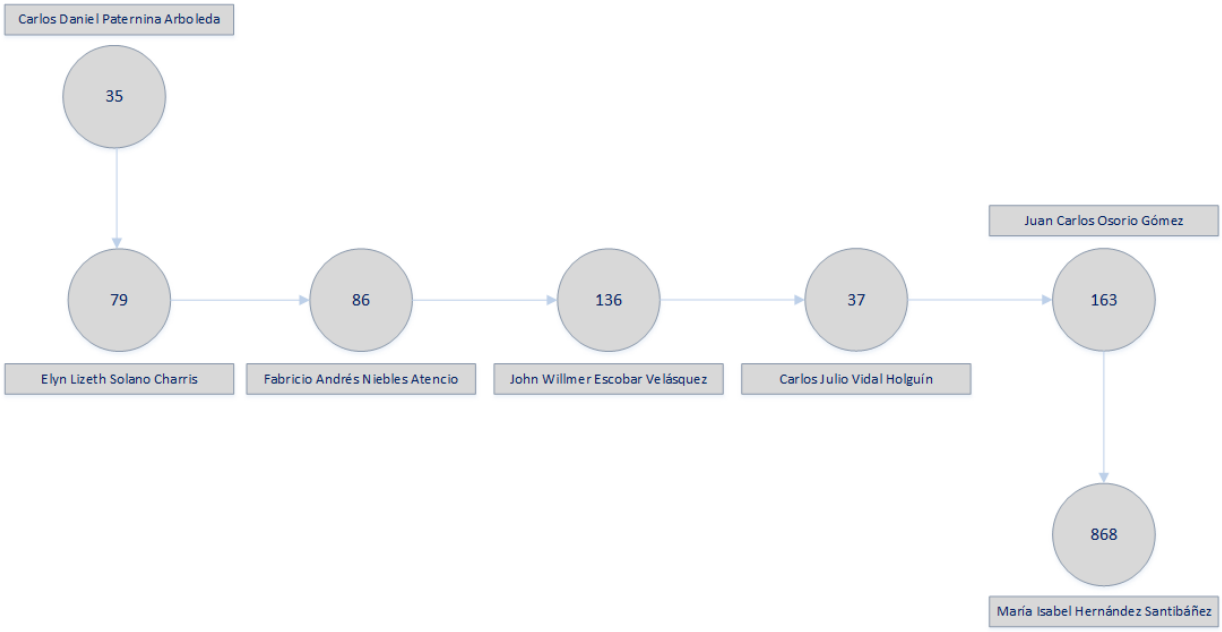

Figura 29. Cadena de referidos entre el nodo 35 y el nodo 868.

En la figura 29 se observa el camino que debe recorrer el autor para contactar a la autora María Isabel Hernández, este resultado es consecuente con la distancia media entre dos autores que se había calculado anteriormente, además se destaca que la distancia es pequeña considerando el tamaño de la red y la distancia real que existe entre ambos autores por su ubicación geográfica. El autor Carlos Daniel Paternina pertenece a la comunidad 1 y se encuentra fuertemente conectado con otros pares evaluadores con afiliación a la Universidad del Norte y la autora María Isabel Hernández pertenece a la comunidad 21, esta autora no tiene perfil en CvLAC por lo que se presume no se encuentra investigando activamente, sin embargo, hace parte de la red pues colaboró con el par evaluador Juan Carlos Osorio Gómez, ambos registran afiliaciones con la Universidad del Valle.

Los anteriores patrones son de gran utilidad pues permiten entender de qué manera es posible aumentar la colaboración científica entre las distintas comunidades, el análisis de redes sociales es

una herramienta que permite manipular y utilizar los datos de las redes sociales para optimizar determinados aspectos de la red, para este caso concreto la colaboración entre las comunidades y autores de la red, así como la difusión de la información y el conocimiento entre comunidades de investigación con líneas afines. Los patrones pueden ser usados para evaluar las capacidades de la red para colaborar y generar conocimiento, pues permiten comprender el estado actual y hacer seguimiento de la evolución de la colaboración científica a través del tiempo.

7. Validación del modelo

Con el fin de validar el algoritmo empleado para la detección de comunidades y aumentar la confianza en las comunidades obtenidas y el análisis desarrollado a cada una de ellas se implementó el algoritmo en una red social real de benchmark utilizada en la literatura conocida como "Zachary's Karate Club".

7.1 Club de Karate de Zachary

Esta es una conocida red social utilizada en la literatura para la validación de modelos de detección de comunidades. El uso de redes del mundo real es siempre preferido ya que son las redes para las cuales se pretenden implementar los modelos, sin embargo, no siempre resulta sencillo encontrar redes sociales con estructuras de comunidad conocidas que permitan validar las comunidades obtenidas. Para el caso de esta red los miembros corresponden a un club de Karate de una universidad que, tras una disputa entre el administrador del club y el instructor, se divide en dos nuevos clubes. Los nodos 1 y 34 corresponden al instructor y al administrador respectivamente, por ende, cada uno se encuentra en una comunidad diferente.

En la figura 30 se observa la partición de la red social en dos comunidades tras la separación del club de karate:

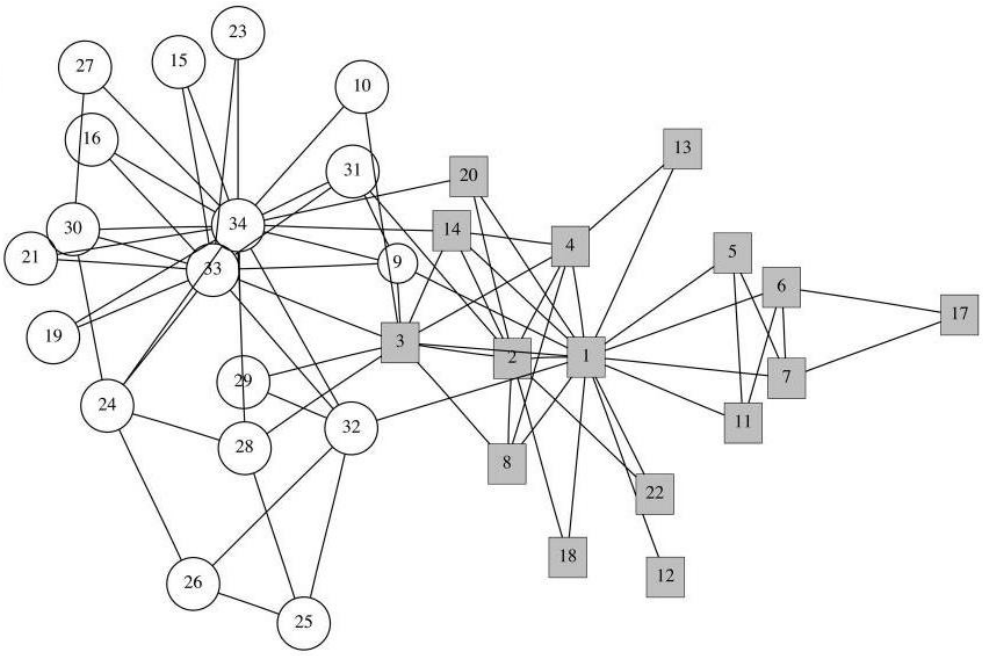

Figura 30. Partición de la red social tras la separación del club. Adapatado de: Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821–7826. doi:10.1073/pnas.122653799

Usando el algoritmo CLA-Net se detectaron las comunidades para la red social del club de karate de Zachary, esta red social tiene 34 nodos y 78 arcos. Se establecieron los siguientes parámetros para la implementación:

Tabla 11.

Parámetros usados en la red social de Karate de Zachary

Inicialización	**Recompensa**	**Convergencia mínima**
17	0.1	5

Tras la implementación se obtuvo una estructura de comunidad con 2 comunidades, al revisar los miembros en cada comunidad se encontró que el algoritmo solo erró al asignar el nodo 10. A continuación se muestran las dos comunidades obtenidas por el algoritmo:

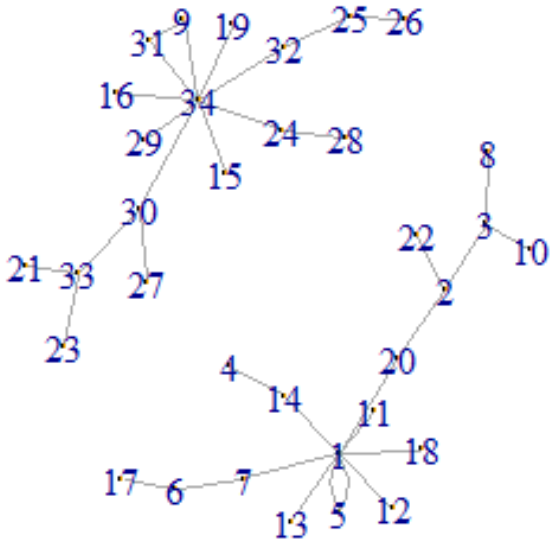

Figura 31. Estructura de comunidad obtenida con el algoritmo CLA-Net. Adaptado de Software Estadístico R Studio ®.

Como se puede observar en la figura 31 el desempeño del algoritmo fue bastante bueno, formó dos comunidades como se esperaba y el único nodo que no asignó de forma correcta fue el nodo 10, revisando la red social original se encontró que era razonable que el algoritmo tuviera dificultad asignando ese nodo, ya que dicho nodo en particular tenía un enlace con un nodo en cada comunidad, por lo que el problema era difuso y no tenía ningún criterio adicional que le permitiera discernir entre las dos comunidades. El resultado de la implementación en una red conocida como esta permite tener mayor confianza en los resultados obtenidos tras la implementación del algoritmo en la red de investigadores de Colombia.

8. Conclusiones

El algoritmo CLA-Net permitió encontrar 63 comunidades de investigación en el área de Ingeniería Industrial con una modularidad de 0.754 tras su implementación. Una red del orden de los 1000 nodos requirió aproximadamente 20 minutos para converger y adicionalmente se observó que el modelo resolvió el problema de límite de resolución ya que el algoritmo detectó incluso las comunidades más pequeñas.

El ajuste de los parámetros del algoritmo CLA-Net permitió obtener una solución de buena calidad en un tiempo de convergencia razonable. En el ajuste se evidenció que solo el parámetro de recompensa tenía un efecto en el valor de la modularidad y que a un nivel de 0.2 proporcionaba una modularidad más alta. Estos parámetros deben ser ajustados para cada red en particular y se debe explorar diferentes semillas hasta encontrar una que permita acelerar la convergencia del algoritmo, pues esto puede ahorrar esfuerzos computacionales.

En la validación de las comunidades se encontró que en la red los investigadores que trabajan en una misma línea de investigación formaron comunidad. Además, aproximadamente un 35% de las colaboraciones se dieron entre profesores y estudiantes que desarrollaban un trabajo de investigación bajo su dirección y un 65% entre investigadores de la misma línea de investigación.

Los investigadores que ocupan los roles de puentes e influenciadores en la red son investigadores de categoría Senior en su mayoría, esto podría indicar que los autores con roles más importantes son quienes cuentan con un mayor reconocimiento como investigadores.

La red de investigación de Ingeniería Industrial tiene un gran potencial de colaboración, ya que se evidencia que existen comunidades con líneas de investigación afines que podrían colaborar juntas, pero se encuentran segmentadas por institución o ubicación geográfica. Por otro lado, la red solo tiene un 0.02% de sus posibles conexiones, aun cuando no se espera que en una red de investigación todos los autores hayan colaborado los unos con los otros, este valor es bajo y evidencia que ni siquiera la gran mayoría de los autores con líneas de investigación afines lo han hecho.

El algoritmo CLA-Net y la teoría de grafos son buenas herramientas para analizar la colaboración científica en una red social de investigación, ya que permitieron comprender cómo se distribuyen las comunidades de investigación en el área de Ingeniería Industrial en Colombia, así como extraer patrones de dichas asociaciones, conocer los investigadores más influyentes de la red, los investigadores que controlan el flujo de información y conocimiento, los investigadores que conectan a las diferentes comunidades y los miembros que lideran los procesos de investigación dentro de las diferentes comunidades.

Este trabajo puede usarse como una guía para ser replicado en otras redes de investigación del país en el análisis de redes de semilleros y grupos de investigación de las Instituciones de Educación Superior-IES, redes de investigación del sector privado y en general de las redes de investigación por área a nivel nacional. Este análisis puede soportar los procesos de evaluación y generación de estrategias para el aumento de la colaboración científica. Esto permitirá aprovechar las diferentes habilidades y experticias de los investigadores del país en favor de fortalecer los procesos de investigación, la ciencia y la consolidación de la sociedad del conocimiento en Colombia.

9. Recomendaciones

El uso de una base de datos con información exhaustiva sobre las publicaciones en coautoría de los investigadores reconocidos por Minciencias en determinada área a través del tiempo permitiría obtener una visión completa de la red de investigación y del panorama de colaboración científica en el país. En muchos casos la información registrada en el perfil de CvLAC de los autores está incompleta o contiene errores, lo que distorsiona el análisis y disminuye la confiabilidad de los resultados.

El acceso a una base de datos con la información concerniente a todos los investigadores reconocidos por Minciencias permitiría analizar y validar de manera más fácil y precisa las comunidades detectadas extrayendo patrones de colaboración a partir de los atributos similares entre los miembros de la comunidad.

La programación e implementación del algoritmo CLA-Net para la red de investigadores se llevó a cabo usando el lenguaje de programación de R en R Studio ®. Sin embargo, el diseño del algoritmo se puede mejorar, en el trabajo se realizó una programación adecuada dadas las habilidades de las autoras en la programación; sin embargo, la programación por expertos puede hacerse de manera más eficiente, lo cual llevaría a reducir aún más los tiempos de convergencia del algoritmo.

Un algoritmo basado en un modelo de autómatas de aprendizaje celular que permita la detección de comunidades en redes sociales con pesos en sus enlaces permitiría una mejor representación de las redes de colaboración científica, pues reflejaría las colaboraciones recurrentes entre los pares de investigadores.

Referencias Bibliográficas

Ahmed, A. Khan, M. Usman, M. Saleem, K. (2018). Analysis of coautorship network in political science using centrality measures. *International Journal of Advance Computer Science and Applications*. Vol 9. N°10.

Aldieri, L., Kotsemir, M., & Vinci, C. (2018), The impact of research collaboration on academic performance: An empirical analysis for some European countries, *Socio-Economic Planning Sciences*, Vol 62, pág 13-30.

Ávila, J.H., Madariaga, C. (2012). Redes Sociales y análisis de redes- Aplicaciones en el contexto comunitario y virtual. *Corporación Universitaria Reformada*. Cap. 4 Pág. 97-120.

Beaver, D. D. B. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, *52*(3), 365–377. https://doi.org/10.1023/A:1014254214337

Beigy, H. Meybody, M. (2004). A Mathematical Framework for Cellular Learning Automata. *Advances in Complex Systems*, Vol. 7, pág 295–319.

Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4), 661-703.

Duch, J. & Arenas, A. (2005). Community detection in complex networks using Extremal Optimization. *Physical Review E 72*.

Esnaashari, M. Meybodi, M.R. (2007). Irregular cellular learning automata and its application to clustering in sensor networks, in: Proceedings of the 15th Conference on Electrical Engineering (ICEE). *Telecommunication Research Center*. Tehran, Iran.

Faloutsos, M., Faloutsos, P., Faloutsos, C. (1999). On Power-Law Relationships of the Internet Topology. *ACM SIGCOMM Computer Communication Review*. 29.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

Freeman, L. (1978). Centrality in social networks: Conceptual clarification. *Social Networks*, pag. 215-239.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, Vol 486, pág.75-174.

Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the national academy of sciences*, 104(1), 36-41.

Girvan, M., Newman M.E.J. (2002). Community structure in social and biological networks. *Proc. Natl.* Acad. Sci. USA, 99 (12), pp. 7821-7826.

Hanneman, R. A., & Riddle, M. (2003). Introduction to social network methods. 2005. *University of California*, Riverside.

Hawe, P., Webster, C., & Shiell, A. (2004). A glossary of terms for navigating the field of social network analysis. *Journal of Epidemiology & Community Health*, 58(12), 971-975.

He, Z., Geng, S., Campbell, C. (2009). Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. *Research Policy*, 38, 306–317.

Herrero, R. (2000). La terminología del análisis de redes. Problemas de definición y traducción. *Política y Sociedad*, 33, 199-206.

Horta, V. Stroele, V. Braga, R. David, J. Campos, F. (2018). Analyzing scientific context of researchers and communities by using and semantic technologies. *Future Generation Computer Systems*, Vol 89. pág 584-605.

Ji, J., Song, X., Liu, C., & Zhang, X. (2013). Ant colony clustering with fitness perception and pheromone diffusion for community detection in complex networks. *Physica A: Statistical Mechanics and Its Applications*, *392*(15), 3260–3272. https://doi.org/10.1016/j.physa.2013.04.001

Katz, J. S., & Martin, B. R. (1997). *What is Research Collaboration ? 7333*(October), 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kautz, H., Selman, B., & Shah, M. (1997). The hidden web. *AI Magazine*, 18(2), 28–36. https://doi.org/https://doi.org/10.1609/aimag.v18i2.1291

Khomami, M., Rezvanian, A., Meybodi, M. (2017). A new cellular learning automata-based algorithm for community detection in complex social networks. *Journal of Computational Science*. Vol 24. pág 413-426.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35, 673–702.

Lotka, A.J. (1926) The Frequency Distribution of Scientific Productivity. *Journal of the Washington Academy of Sciences*, 16, 317-323

Maz, A., Jiménez, N., & Villarraga, M., (2016). Colombian scientific production indexed in Scielo: A bibliometric analysis. *Revista Interamericana de Bibliotecologia*, 39(2), 111–119. https://doi.org/10.17533/udea.rib.v39n2a03

Meybodi, M. R., & Kharazmi, M. R. (2004). Application of cellular learning automata to image processing. *Journal of Amirkabir*, 14(56A), 1101-1126.

Milgram, S. (1967). The small world problem. *Psychology Today*. 67. pág. 61–67.

Minciencias. (s.f.). Recuperado de: https://minciencias.gov.co/ministerio/vision-y-mision

Newman, M. Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev*. E69. 026113.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review*. 69 (066133).

Pizzuti, C. (2008). GA-Net: a genetic algorithm for community detection in social networks. In: Parallel Problem Solving from Nature (PPSN). *Springer Berlin*. pág 1081–1090.

Pizzuti, C. (2011). A multiobjective genetic algorithm to find communities in complex networks. *IEEE Transactions on Evolutionary Computation*, 16(3), 418-430.

Plataforma Scienti. Servicio de Información de Evaluadores Pares Reconocidos del SNCtel. Recuperado de: https://scienti.minciencias.gov.co/ciencia-war/busquedaPares.do [Consultado 3 de enero de 2020].

Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3), 036106.

Rahman, M.S, Ngom, A. (2013). A fast-agglomerative community detection method for protein complex discovery in protein interaction networks. *Springer*. pág. 1-12.

Rigby, J., & Edler, J. (2005). Peering inside research networks: Some observations on the effect of the intensity of collaboration on the variability of research quality. *Research Policy*, *34*(6), 784–794. https://doi.org/10.1016/j.respol.2005.02.004

Tasgin, M., Herdagdelen, A., & Bingol, H. (2007). Community detection in complex networks using genetic algorithms. *arXiv preprint arXiv*:0711.0491.

Thathachar, M. & Sastry, P. (2002). Varieties of Learning Automata: An Overview. *IEEE transactions on systems*, man, and cybernetics—part b: cybernetics, vol. 32, número 6.

Wasserman, S. & Faust, K. (1994). Social Network Analysis. *Cambridge University Press*. Cambridge, UK. 1994.

Watts, D. & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*. 393. pág 440–442.

Wolfram, S. (2018). Cellular automata and complexity: collected papers. *CRC Press*.

Yudhoatmojo, S. B., & Samuar, M. A. (2017). Community Detection on Citation Network of DBLP Data Sample Set Using LinkRank Algorithm. *Procedia Computer Science*, *124*, 29–37. https://doi.org/10.1016/j.procs.2017.12.126

Zhao, Y., Jiang, W., Li, S., Ma, Y., Su, G., Lin, X. A cellular learning automata-based algorithm for detecting community structure in complex networks. *Neurocomputing* 151 (2015) 1216–1226.