



Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento de Conocimiento

Tesis de Maestría

Diseño de una herramienta para minería de redes de coautoría. Estudio del caso de la Revista Argentina de Antropología Biológica

Tesista: Ing. Julio Augusto Spairani

Directora de Tesis: Dra. María Gabriela Russo

Co-directora: Dra. María Elena Buemi

Buenos Aires, Noviembre de 2020

Resumen

El presente trabajo aborda el problema de análisis de redes de coautoría dentro de una revista científica local en un periodo de 20 años desde su primera edición, que al momento de escritura del trabajo no figuraba en la red scopus o web of science. Se extrajeron los datos vía 'minado web' (web mining) de la página oficial de la revista, se consolidó una base de datos y finalmente se desarrolló una herramienta de software para poder realizar análisis de coautoría con dichos datos. La herramienta es una aplicación web que consta de tres secciones principales 'análisis preliminar de los datos' que permite explorar de forma resumida los datos con los que se va a realizar el análisis; un 'análisis de periodos' donde se pueden seleccionar uno o varios periodos a analizar, que incluye visualización con navegación interactiva de la red, análisis de estructura del grafo, análisis de mundo pequeño, análisis de libre escala, búsqueda de comunidades, breve análisis de títulos vía minado de texto (text mining); y una última sección de 'análisis temporal' que tiene foco en mostrar las variaciones de la red de forma gráfica e interactiva entre los periodos de análisis. Finalmente se realiza el análisis de la revista, utilizando la herramienta para contestar las hipótesis planteadas en el capítulo inicial.

Palabras clave: redes de coautoría; minado web; minado de texto; análisis de componentes; análisis de comunidades; aplicación web interactiva; análisis temporal; mundo pequeño; libre escala; desambiguación de autores; R; shiny; open source.

Abstract

This work tackles coauthorship networks in a local scientific journal in a time lapse of 20 years since its first edition, at writing time the network does not appear on scopus or web of science. Data was gathered via web mining of the journal official website, making transformations to consolidate a database and finally develop a software tool to do the coauthorship network analysis over that data. The software is a web application composed by 3 sections: an 'exploratory data analysis' to know which data is used to make the coauthorship network; a 'yearly analysis' where you can select one or several years to aggregate data; this includes interactive network visualization, graph structure analysis, small world analysis, free scale analysis, identifying communities, brief title analysis via text mining ; and last section of temporal analysis which focuses on variations of network over time that can be explored in an interactive way. Finally we use the web application for the journal analysis and answer the hypothesis and some questions.

Keywords: coauthorship networks; web mining; text mining; component analysis; community analysis; interactive web application; temporal analysis; small world; free scale; author disambiguation; R; Shiny; opensource

Agradecimientos

A Nancy, Máximo, Emma, familia, directores, maestros, amigos, compañeros y muchas otras personas que me dieron apoyo, soporte y ayuda en este largo trayecto.

Indice

Resumen	3
Abstract	4
Agradecimientos	5
Indice	6
Capítulo 1: Introducción	10
Introducción	10
Organización de la tesis	11
Características de la Antropología Biológica	12
Características de la RAAB	13
Conceptos clave	14
Redes y Grafos	14
Redes de Coautoría	15
Fuerza de colaboración	16
Grado	17
Intermediación	17
Componente	18
Mundo Pequeño	19
Antecedentes sobre análisis de coautorías	19
Objetivos	20
Objetivo 1: Desarrollar una herramienta para el análisis de redes de coautoría	20
Objetivo 2: Analizar redes de coautorías de la RAAB	20
Hipótesis	20
Resumen del capítulo	22
Capítulo 2: Generación del conjunto de datos	23
Introducción	23
Materiales y métodos	26
Herramientas de trabajo	26
Introducción al proceso	26
Análisis de la estructura del sitio	27
Descripción de los datos de interés	28
Adquisición de los datos	29
Parseo de los archivos	30
Parseo de los archivos HTML	30
Parseo de los archivos bibTex	30
Parseo de los archivos PDF	30
Generación de la Base	31

Resultados intermedios	31
Generación de la base relacional	31
Estructura de la base final	31
Generación de los grafos	33
Descripción de las propiedades de los nodos y aristas	34
Propiedades de los Nodos	34
Propiedades de las Aristas	35
Resumen del capítulo	36
Capítulo 3: Creación de la Herramienta para análisis	37
Introducción	37
Materiales y métodos	37
Estructura Interna de la aplicación	37
Herramientas de trabajo	37
Respecto de la estructura de la aplicación	38
Disponibilidad local, en la red, en internet	39
Estructura de la interfaz gráfica	39
Sección 1: Consideraciones	40
Datos Crudos	41
Composición	42
Resumen	43
Preguntas	44
Sección 2: Análisis Estático	46
Configuración	46
Visualización	48
Artículos asociados	50
Estructura de la red	51
Métricas de la red	51
Métricas de los nodos	53
Detalles de los componentes	56
Comparación de modelos	59
Validación de Mundo Pequeño	63
Validación de Libre Escala	64
Comunidades	66
Sección 3: Análisis temporal	70
Animación del grafo de manera acumulada	71
Medidas de cada periodo	72
Medidas Acumuladas a cada periodo	72
Ranking de autores en el tiempo	73
Ranking de autores en el tiempo para medidas acumuladas	74
Resumen del capítulo	74

Capítulo 4: Análisis de la Red	75
Introducción	75
Hipótesis 1: La cantidad de autores aumenta en el tiempo debido al crecimiento de la disciplina en el país, lo que se refleja en un aumento en la cantidad de nodos en la red	78
Autores en artículos por año	83
Hipótesis 2: Los autores fundadores se refuerzan en el tiempo, lo cual puede ser medido por grado, intermediación y fuerza de colaboración	84
Hipótesis 3: Los componentes más grandes absorben a los más pequeños a lo largo del tiempo, dado que autores periféricos pasan a trabajar en conjunto con autores principales	89
Verificación de absorción de componentes en la red	91
Relación con actividades de investigación	94
Búsqueda de comunidades	95
Explicación de la métrica de evaluación de algoritmos: Modularidad	97
Explicación del algoritmo seleccionado: Optimización de modularidad multi-nivel	98
Búsqueda de sentido en los grupos obtenidos	99
Explicación de criterio de búsqueda: verificación de comunidades según títulos publicados.	100
Hipótesis 4: La red agregada total (desde 1996 a 2016) es de mundo pequeño puesto que es una característica presente en la mayoría de las redes de coautorías científicas	106
Mundo pequeño	106
Libre Escala	107
Caso aplicado	107
Mundo pequeño para la red agregada total	107
Libre escala para la red total	109
Mundo pequeño para el periodo 1996 - 2001	110
Libre escala 1996 - 2001	112
Mundo pequeño para el periodo 2002 - 2006	113
Libre escala 2002 - 2006	115
Mundo pequeño para el periodo 2007 - 2011	116
Libre escala 2007 - 2011	117
Mundo pequeño para el periodo 2012 - 2016	119
Libre escala 2012 - 2016	120
Resumen de análisis de las propiedades de mundo pequeño y libre escala	122
Comparación con otras redes de coautoría	123
Resumen de capítulo	124
Capítulo 5: Conclusiones	126
Resumen	126
Productos asociados	126
Consideraciones sobre el conjunto de datos y la variación de estructura del sitio en el tiempo	127
Trabajos futuros	127
Actualización de la base de datos	127

Aspectos técnicos	127
Proyección sobre artículos	128
Análisis de textos y tópicos en el tiempo	128
Análisis de conectividad / puentes	128
Bibliografía	130
ANEXO: software y librerías utilizadas	134
ANEXO: Detalle de algoritmos para búsqueda de comunidades	138
ANEXO: Detalle procesamiento bibTEX	139
Problemas de transformación y encoding bibtex	139
Problemas autores	139
Problemas con el volumen de la revista	140
Problema Abstracts Vacios en bibtex	141
ANEXO: Detalle consideraciones PDF	143
Obtención de los pdf	143
Parseo de los pdf	143
ANEXO: Detalle estandarización autores	144
Revisión y validación con expertos	144
Iteraciones de clustering	145
Resultado de Revisión y validación con expertos	150

Capítulo 1: Introducción

Introducción

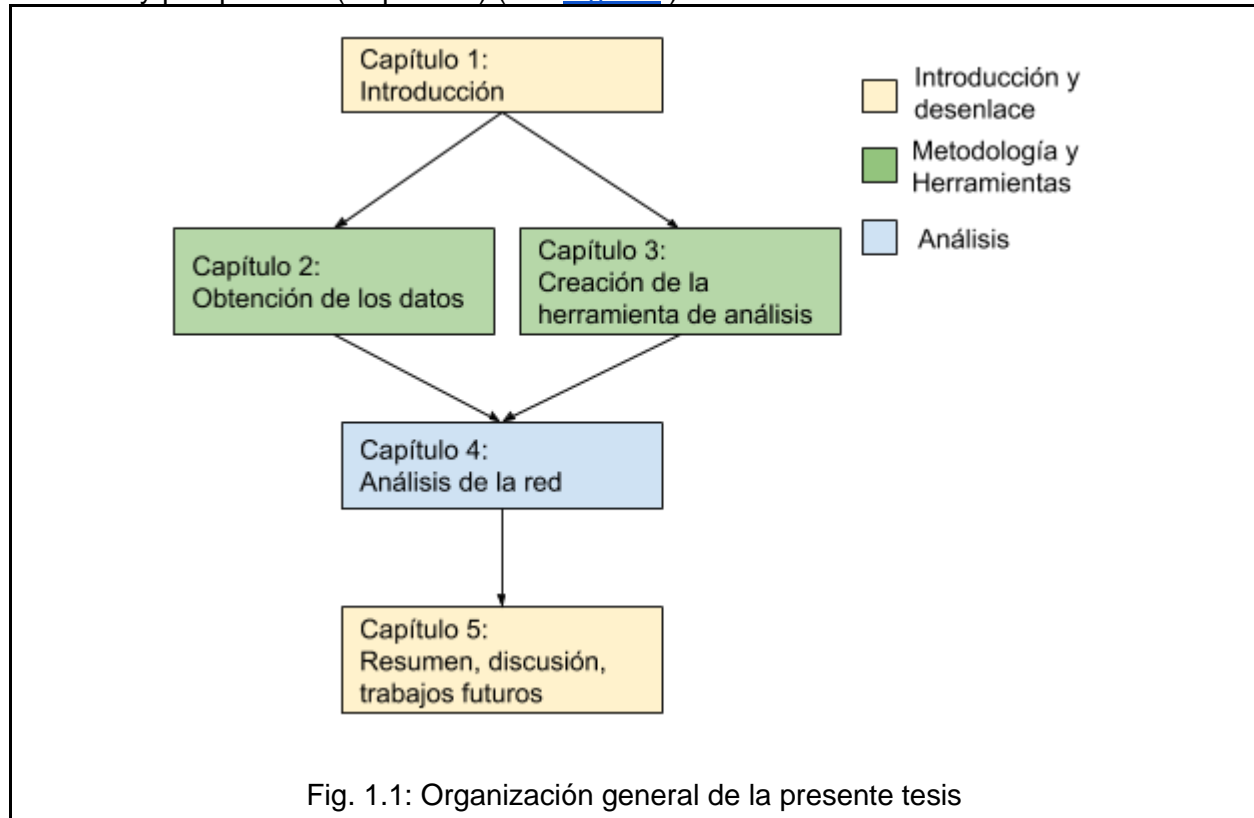
Las redes de coautoría son una herramienta útil para analizar patrones de colaboración científica. Estos patrones pueden ser estudiados dentro de una disciplina, región geográfica o revista científica, así como en forma comparativa ([Kumar 2015](#)). Al analizar redes de coautoría en una disciplina particular, las características relacionadas al grado de cohesión o fragmentación, el tamaño de los nodos y componentes o el patrón de conexión entre nodos, entre otras, brindan información sobre el diálogo colaborativo entre los distintos grupos de investigación que la conforman y la “importancia” relativa de cada uno en el contexto particular.

En esta tesis se analizó una red de coautorías de la Revista Argentina de Antropología Biológica (RAAB), como ejemplo de análisis dentro de una disciplina particular y a nivel local (Argentina). La RAAB es el órgano oficial de comunicación científica de la Asociación de Antropología Biológica Argentina (AABA), por lo que el análisis de coautorías brindó una oportunidad para el estudio de las dinámicas de colaboración dentro de esta disciplina a nivel nacional. En particular, resultó de interés analizar la estructura de la red como reflejo de las interacciones entre distintos grupos de investigación y obtener una aproximación al estudio del crecimiento y desarrollo de una disciplina de origen relativamente reciente en Argentina.

Para ello, en esta tesis se propuso el desarrollo de una herramienta de visualización y análisis de la red, a fin de facilitar el estudio de este caso en particular, así como también su aplicación a otros similares. Siguiendo lo propuesto por [Morel, Serruya, et al. \(2009\)](#), se analizó la red a fin de evaluar el incremento en el número de publicaciones, grupos de investigación y colaboraciones entre los grupos, así como el cambio en estas dinámicas a lo largo del tiempo.

Organización de la tesis

La presente tesis está organizada en cinco capítulos que abarcan las siguientes temáticas: introducción (Capítulo 1), metodología (Capítulos 2 y 3), resultados y análisis (Capítulo 4), y discusión y perspectivas (Capítulo 5) (ver [Fig. 1.1](#)).



En el Capítulo 1 tratamos el contexto general de la presente tesis y una breve justificación de lo realizado en los capítulos siguientes. Se presentan los objetivos e hipótesis subyacentes a este trabajo.

En el Capítulo 2 se especifica cómo se obtuvieron los datos, qué transformaciones realizamos sobre ellos, y finalmente cómo dimos origen a la base de datos que se utilizó para el desarrollo de la aplicación.

El tercer capítulo cuenta la estructura de la aplicación a nivel *software*, las secciones que componen la aplicación, así como el fundamento de estas funcionalidades que componen cada sección.

En el Capítulo 4 realizamos el análisis del caso específico de la RAAB usando la base de datos creada en el Capítulo 2, la herramienta creada en el Capítulo 3, a fin de contrastar las hipótesis planteadas inicialmente. Además, se ejemplifica cómo con la herramienta desarrollada se pueden realizar con facilidad varias tareas comunes en el análisis de redes de coautoría.

Finalmente en el Capítulo 5, último capítulo, resumimos las tareas realizadas, y se discute el proceso, los resultados obtenidos y los potenciales trabajos futuros.

Características de la Antropología Biológica

La Antropología Biológica puede definirse como “el estudio de la evolución y variabilidad biológica de los humanos en el presente y en el pasado” ([Madrigal y González-José, 2016:5](#)). Dentro de esta disciplina se estudian procesos y fenómenos de las poblaciones humanas desde tres ejes o puntos de vista principales: evolutivo, trans-cultural y biocultural ([Madrigal y González-José, 2016](#)). Como ejemplo de estos enfoques, [Madrigal y González-José \(2016:6\)](#) presentan el caso particular del parto humano:

Por lo tanto, aún si estamos estudiando algo tan clínico como, por ejemplo, el parto, lo vemos desde estos tres puntos de vista:

- 1. **Evolutivo:** ¿Cuál es el papel de la pelvis en la locomoción y en el parto? ¿Cuándo en nuestra historia evolutiva los humanos empezamos a caminar de manera bípeda, así afectando la pelvis? ¿Cómo paren los primates no humanos?*
- 2. **Transcultural:** ¿De qué manera las familias y los asistentes médicos ayudan a las madres que están pariendo en culturas alrededor del mundo? ¿Es posible que la manera de parir típica de la biomedicina no sea la mejor para la mujer?*
- 3. **Biocultural:** ¿Cómo es que el estrés de estar pariendo en un hospital afecta el éxito del parto para mujeres que están aisladas de su familia? ¿Podríamos mejorar la salud de madre y bebé incorporando prácticas culturales de la mujer que está pariendo?*

En otras palabras, los antropólogos biológicos estudian un hecho biológico como una enfermedad o un gen, incorporando información acerca del pasado evolutivo de dicha enfermedad o gen, incorporando información acerca de la manifestación biológica de la enfermedad o gen en diferentes culturas-ecosistemas, y estudiando cómo la enfermedad o gen afecta la cultura donde se manifiesta, y cómo la cultura afecta la enfermedad o gen.

La Antropología Biológica como disciplina constituye un campo interesante para analizar redes de colaboración científica puesto que, como se desprende del ejemplo anterior, la articulación de diferentes enfoques puede dar lugar al trabajo transversal entre distintas líneas de investigación. En este marco, el análisis de coautorías en la Revista Argentina de Antropología Biológica (RAAB) brindaría una primera aproximación al estudio de las colaboraciones dentro de esta disciplina a nivel nacional.

Características de la RAAB

La Asociación de Antropología Biológica Argentina (AABA) comienza a gestarse hacia fines de la década de 1980. Una de las necesidades existentes era crear una revista que documentara la producción científica dentro de la disciplina en nuestro país. De esta forma, uno de los objetivos de la AABA fue la creación de la “Revista Argentina de Antropología Biológica” (RAAB) la cual constituye el órgano oficial de difusión de la Asociación ([Cesani, Sardi, et al. 2013](#))

En la [página web de la RAAB](#) puede encontrarse una descripción de sus características generales: [<https://revistas.unlp.edu.ar/raab/about>]

Enfoque y alcance

La Revista Argentina de Antropología Biológica publica estudios originales y de significación en los campos de la antropología biológica: evolución y variación humana, adaptación, genética, crecimiento y desarrollo, antropología forense, experimentación, morfología y primatología. Además, publica temas de antropología social, arqueología, psicología evolutiva, geografía humana, demografía, ecología, epidemiología, museología, conservación, historia y filosofía de la ciencia, medicina, entre otros, siempre que los contenidos tengan una vinculación explícita con algún área en particular de la antropología biológica.

Sobre La RAAB

Es propiedad de la AABA. Se comenzó a publicar en 1996 con el objetivo de difundir estudios originales en los campos de la Antropología Biológica y temas afines. Actualmente se publican dos números por año que incluyen artículos científicos y de revisión, comentarios bibliográficos, noticias breves, resúmenes de tesis de postgrado, cartas al editor, reportes técnicos y comunicaciones breves sobre diversos temas de interés en la especialidad. Los idiomas de publicación son español, portugués e inglés. La recepción de manuscritos está abierta todo el año y se publican en orden de aceptación. El proceso de evaluación de los manuscritos sigue la modalidad de revisión por pares.

Indizada en

Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal LATINDEX (Folio N° 7579, versión papel; Folio N° 20444, versión on-line)

Servicio de Difusión de la Creación Intelectual de la UNLP (SeDiCI)

HOLLIS Catalog, Harvard University (012566584)

Núcleo Básico de Revistas Científicas Argentinas (Resolución N° 1855/13 CONICET)

SciELO Argentina <http://www.scielo.org.ar/>

Redalyc. Red de Revistas Científicas de América Latina y el Caribe, España y Portugal

Respecto de la estructura de la revista, generalmente su publicación es en uno o varios volúmenes anuales. A estos volúmenes a lo largo de la tesis también los llamamos “ediciones”. Cada volumen tiene artículos publicados, estos pueden ir en diferentes categorías. El foco de esta tesis está centrado en los de tipo “Trabajo Original”

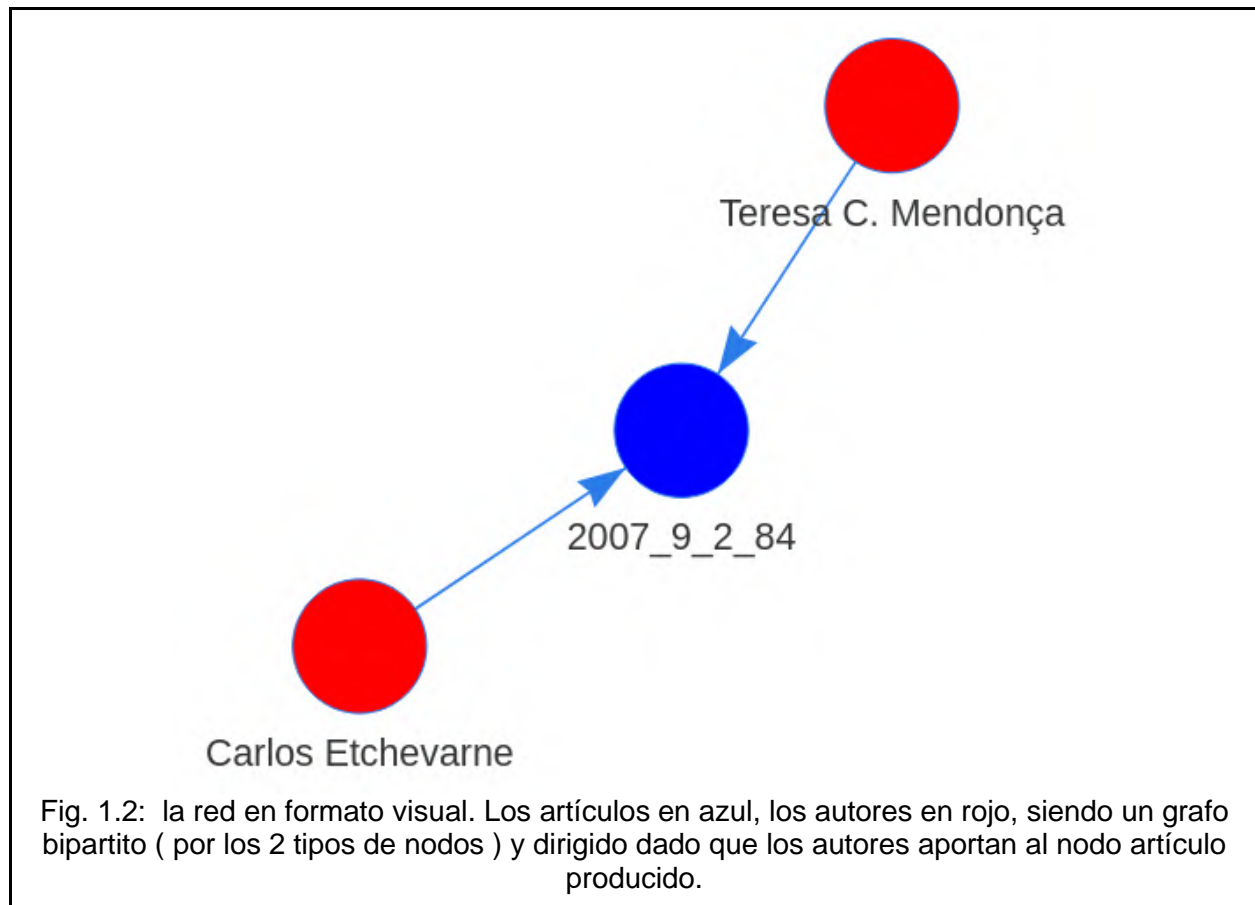
Conceptos clave

Redes y Grafos

En palabras de [Barabási \(2016\)](#), traducidas y adaptadas por el autor de esta tesis:

Si queremos entender sistemas complejos, primero necesitamos saber cómo sus componentes interactúan entre ellos. En otras palabras, necesitamos un mapa de cómo están conectados. Una red es un catálogo de los componentes de un sistema, comúnmente llamados nodos o vértices, y las interacciones directas entre ellos, llamadas conexiones o aristas. Este tipo de representación de redes ofrece un lenguaje común para estudiar sistemas que pueden diferir enormemente en naturaleza, apariencia o alcance.

En nuestro caso la red está compuesta de "nodos autores", "nodos artículos" y las conexiones representan "autor escribió artículo"; este tipo de redes que tienen 2 tipos de nodos se llaman bipartitas. A lo largo de la tesis entraremos en detalles respecto de los grafos. Un ejemplo de la naturaleza de la red en formato visual se puede ver en la [Fig. 1.2](#).



Redes de Coautoría

Extraído del resumen de [Kumar \(2015\)](#) que a su vez cita a [Newman \(2001\)](#). Traducido y adaptado por el autor de la tesis:

Una red social es una representación de dos entidades o más conectadas entre ellas a través de algún tipo de relación. Si uno acota las redes sociales humanas para las cuales existen datos cuantitativos, quedan pocas: una de estas son las redes de los actores de películas en las que los nodos actores trabajan juntos en una película; y otra es la red de investigadores, donde los investigadores escriben un trabajo científico en conjunto ([Newman 2001](#)). De estas redes la que probablemente tiene más función social es la red de investigadores. A diferencia de la red de películas y actores, los investigadores eligen con quien hacer investigación y luego escriben los resultados del trabajo en la forma de un artículo científico con coautores. Estas colaboraciones dejan huellas digitales en la forma de bibliografía, que luego puede ser seguida y evaluada.

El texto de Newman fue escrito antes de que internet fuera adoptada masivamente. Hoy en día abundan muchos casos de redes, sin embargo queda investigar cuales existen con el nivel de calidad que tiene una red de escritura científica.

Las redes de coautoría se presentan al realizar la proyección de la red sobre los nodos autores, quedando los autores como nodos, y los nodos artículos son reemplazados para formar las relaciones, a las cuales se les da más peso según la cantidad de trabajos realizados en conjunto.

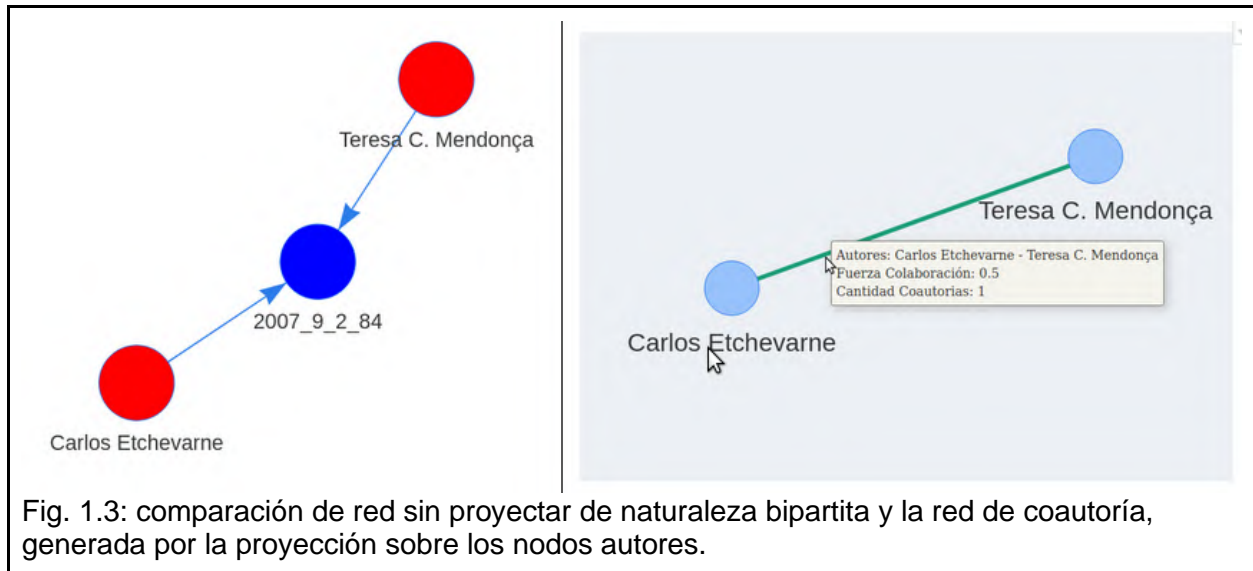
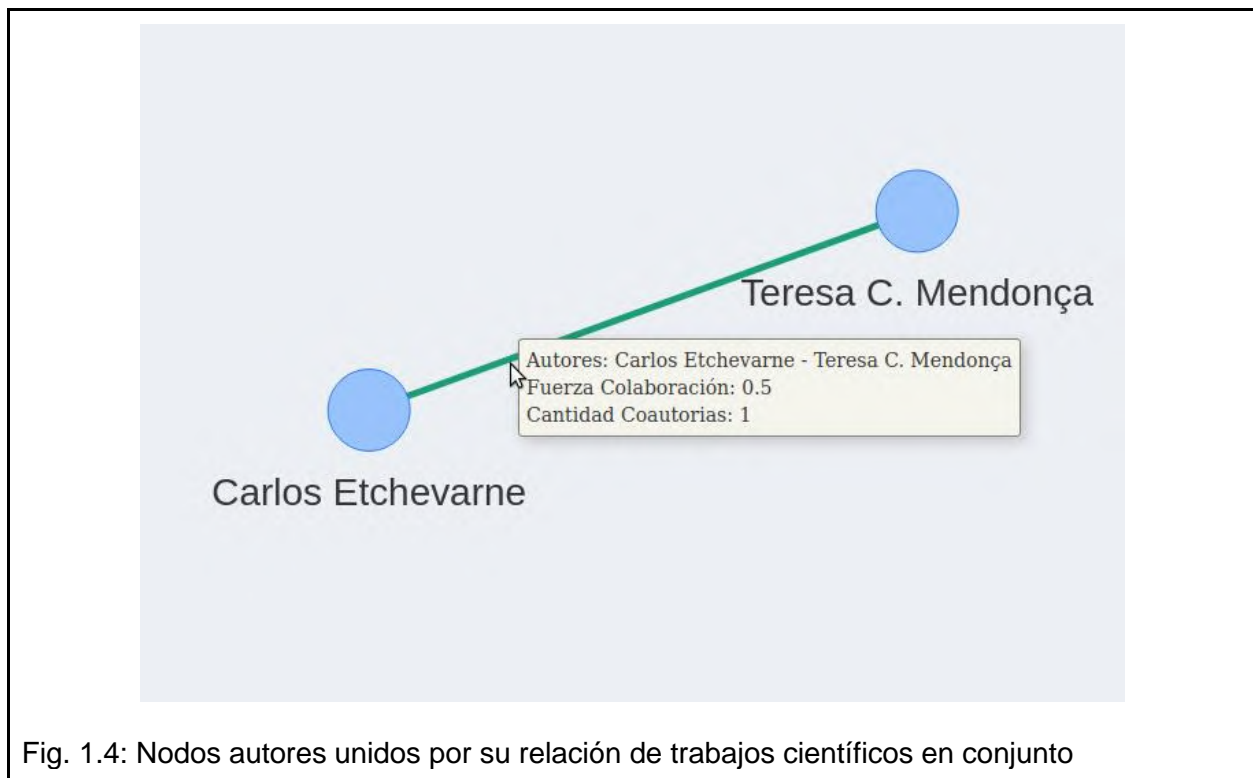


Fig. 1.3: comparación de red sin proyectar de naturaleza bipartita y la red de coautoría, generada por la proyección sobre los nodos autores.

Para entender la proyección podemos ver en paralelo la red sin proyectar y la red proyectada sobre los autores en la [Fig. 1.3](#)

Fuerza de colaboración

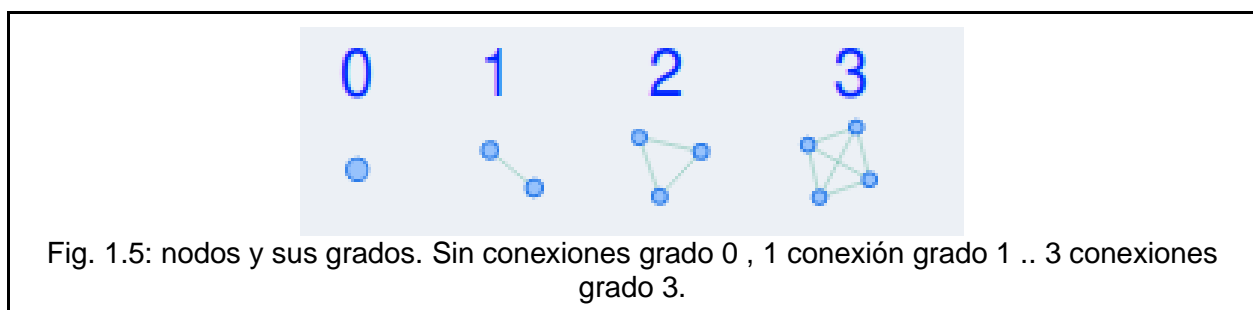
Es una Métrica descrita en el paper de [Newman \(2004\)](#), representa el artículo creado dividido la cantidad de autores que lo escriben, esta métrica plasma cómo se reparte el esfuerzo de la generación de un artículo realizando una cuenta simple como dividir el esfuerzo para realizar un artículo por su cantidad de autores. En la [Fig. 1.4](#) vemos que realizaron 1 trabajo en conjunto entre 2 autores, por lo tanto $1 \text{ artículo} / 2 \text{ autores} = 0.5$ de fuerza de colaboración.



En la [Fig. 4](#) se observan 2 nodos autores unidos por su relación. Cuando vemos los detalles de esta relación observamos que se tuvo en cuenta la medida llamada Fuerza de colaboración, del inglés *collaboration strength*.

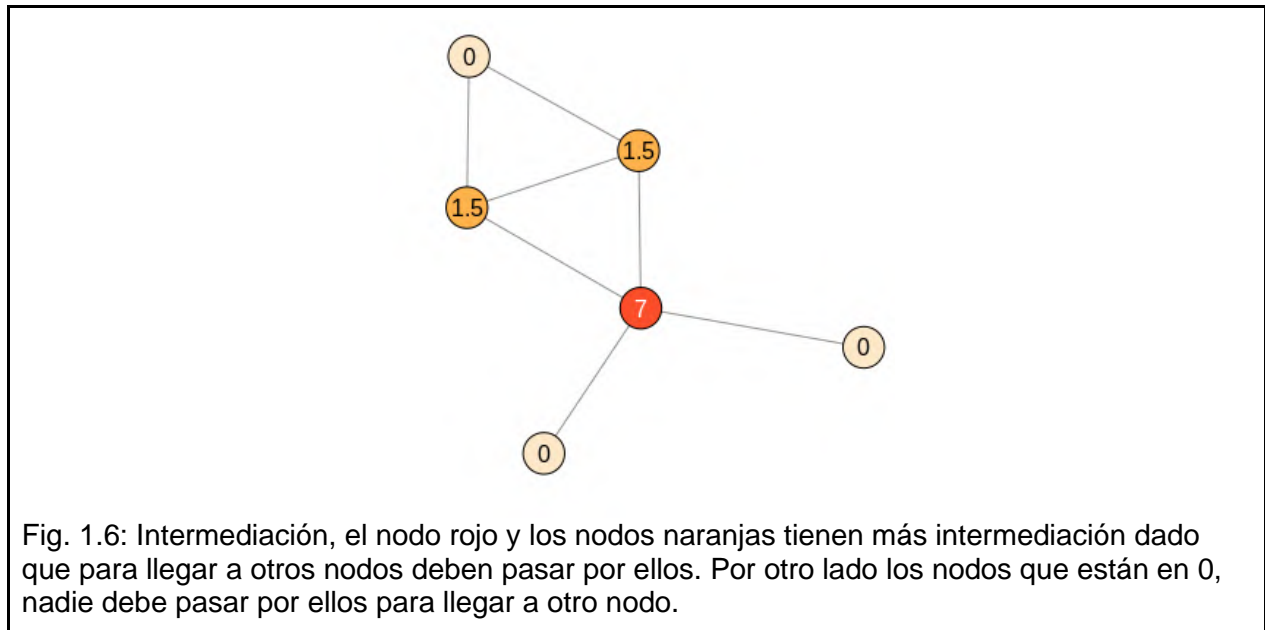
Grado

El grado del nodo es la cantidad de conexiones directas del nodo. Como ejemplo se puede ver la [Fig. 1.5](#).



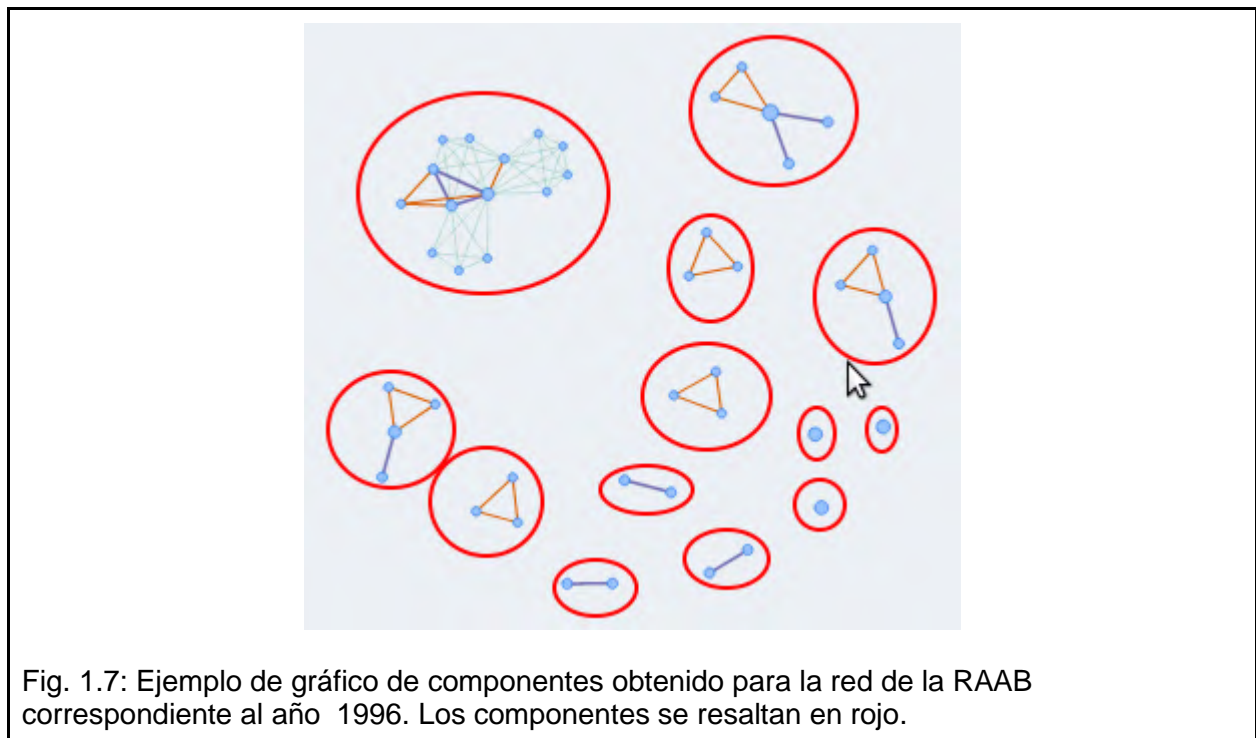
Intermediación

La intermediación representa la importancia de un nodo según su participación en caminos para llegar de un nodo a otro. Vemos un ejemplo de esta métrica en la [Fig. 1.6](#).



Componente

Un componente es una porción de red en donde todos los nodos están conectados directa o indirectamente por lo menos por una arista. El análisis de componentes da un vistazo general de la estructura de la red ([Fig. 1.7](#)) y provee información valiosa en relación a su nivel de fragmentación ([Morel, et al. 2009](#)).



Mundo Pequeño

El estudio de “mundo pequeño” está asociado a evaluar ciertas propiedades de la red que permiten que se pueda afirmar que las personas o nodos se conectan con otras a través de una cantidad pequeña de saltos, más allá de que se hallen muy separadas entre sí en la red.

Para comprender un poco mejor el estudio de “Small world”, surge de la frase en inglés ‘*My it’s a small world*’ / ‘*Que mundo pequeño*’ cuando dos personas conocen a otra y, por lo tanto, estarían interconectados a través de una tercera. Este estudio de [Milgram \(1967\)](#) consistió en enviar mensajes desde una persona en un punto A hasta otra en un punto B bajo ciertas reglas, entre ellas, que solo podían mandar el mensaje a personas conocidas inmediatas.

Para ganar una intuición gráfica podemos revisar la [Fig. 1.8](#) donde el autor en el círculo rojo desea contactar con el autor en el círculo azul, recordemos que las aristas que unen son trabajos realizados en conjunto.



Fig. 1.8: Si hablase el nodo del círculo rojo que desea contactarse con el nodo azul “*El coautor de mi trabajo original, tiene otro coautor que trabajó con nodo en círculo azul que nos podría presentar*”

Nos interesa revisar esta propiedad para la red de la RAAB dado que es una característica frecuente en el contexto de redes científicas según [Kumar \(2015\)](#).

Antecedentes sobre análisis de coautorías

El análisis de redes sociales es un tópico que existe desde aproximadamente la década de 1950, lo que se denominaba sociometría. Se hizo popular en la década de 1960 por el estudio de mundo pequeño de [Milgram \(1967\)](#), y luego volvió a tener popularidad durante la década de 1990 con el juego “6 grados de Kevin Bacon” que medía qué tan lejos alguien del círculo de actuación en Hollywood estaba del actor Kevin Bacon medido en interacciones por compartir películas o series.

En bibliometría, las redes no son algo nuevo, generalmente se estudiaron las redes de citación, en las que los nodos son los papers y las citas las aristas. Sin embargo, las redes de coautoría, presentan otro enfoque cercano a la comunidad científica más que al conocimiento, tal vez por

eso tuvieron menos atención anteriormente. A pesar de esto nos puede brindar mucha información ([Newman 2004](#))

En este trabajo utilizamos varios de los estudios mencionados en la revisión de [Kumar \(2015\)](#); como por ejemplo: desambiguación de nombres de autor, el estudio de mundo pequeño, la búsqueda de comunidades, la visualización de la red, tamaño del componente más grande, entre otros. Cada uno de estos es profundizado a lo largo del estudio.

Un enfoque deseado para esta tesis fue el punto de vista temporal dado que exploramos el periodo de 20 años desde el inicio de la revista, para ello nos basamos en parte del estudio realizado por [Koseoglu \(2016\)](#); que revisa temas como evolución temporal de una revista específica, evaluando métricas, visualizaciones y comparaciones.

Objetivos

El **objetivo general** de la presente tesis es realizar una aporte a la comprensión del desarrollo de la Antropología Biológica en nuestro país a través del análisis de una red de colaboración científica. Para ello, se plantearon dos objetivos específicos.

Objetivo 1: Desarrollar una herramienta para el análisis de redes de coautoría

Se planteó la creación de una herramienta para facilitar el análisis de redes de coautoría científica mediante la simplificación de tareas repetitivas, como por ejemplo: variación de parámetros para aplicar filtros y/o algoritmos, selección de diferentes nodos/aristas/grupos, aplicación de filtros a los datos, visualización general de la red desde diferentes puntos de vista, obtención de detalles de nodos y aristas, obtención de detalles de los datos, entre otras.

Dentro de este objetivo se buscó, además, facilitar la exploración de los datos mediante interacciones intuitivas brindadas a través de una interfaz web de acceso abierto para diferentes personas que puedan realizar el o los análisis. Los tres focos de la herramienta son un breve análisis de los datos que dan origen a la red, el análisis de la red para ciertos periodos seleccionados, y finalmente el análisis temporal de la red.

Objetivo 2: Analizar redes de coautorías de la RAAB

Utilizando la herramienta desarrollada, se propuso analizar el caso particular de la Revista Argentina de Antropología Biológica (RAAB) y nos interesarnos principalmente en cómo fue el comportamiento de la red en el tiempo.

Hipótesis

En congruencia con lo planteado en el objetivo 2, nos planteamos las siguientes hipótesis:

- Hipótesis 1: La cantidad de autores aumenta en el tiempo debido al crecimiento de la disciplina en el país, lo que se refleja en un aumento en la cantidad de nodos en la red.
- Hipótesis 2: Los autores fundadores se refuerzan en el tiempo, lo cual puede ser medido por grado, intermediación y fuerza de colaboración.
- Hipótesis 3: Los componentes más grandes absorben a los más pequeños a lo largo del tiempo, dado que autores periféricos pasan a trabajar en conjunto con autores principales.
- Hipótesis 4: La red agregada total (desde 1996 a 2016) es de mundo pequeño puesto que es una característica presente en la mayoría de las redes de coautorías científicas.

Estas hipótesis tienen sus fundamentos en:

Hipótesis 1: *“La cantidad de autores aumenta en el tiempo debido al crecimiento de la disciplina en el país, lo que se refleja en un aumento en la cantidad de nodos en la red”* : nos interesa verificar si la disciplina de Antropología Biológica crece en el país analizando la cantidad de autores a lo largo del tiempo en la Revista Argentina de Antropología Biológica. Acotamos el estudio a dicha publicación dado que buscamos representación de la materia a nivel local, entendemos que esta hipótesis puede ser expandida en publicaciones internacionales por autores locales y en otras revistas, lo cual queda fuera del alcance de la presente tesis.

Hipótesis 2: *“Los autores fundadores se refuerzan en el tiempo, lo cual puede ser medido por grado, intermediación y fuerza de colaboración.”*: nos interesa analizar si el avance de la revista depende de los autores fundadores, o si está impulsada por nuevos autores y proponemos métricas con las cuales poder evaluar esta actividad.

Hipótesis 3: *“Los componentes más grandes absorben a los más pequeños a lo largo del tiempo, dado que autores periféricos pasan a trabajar en conjunto con autores principales”* nos interesa para tratar de entender en mayor profundidad el comportamiento interno de la red revisando cómo se comportan los componentes. Para lograr esto analizamos la evolución temporal del tamaño del componente principal identificando momentos específicos de absorción de componentes.

Hipótesis 4: *“La red agregada total (desde 1996 a 2016) es de mundo pequeño puesto que es una característica presente en la mayoría de las redes de coautorías científicas”* nos interesa para comprar con puntos de vista externos. Realizamos validación de propiedades comunes que se dan en redes del mundo real como Mundo pequeño y Libre escala. De forma complementaria para complementar la vista hacia afuera, contrastamos con redes de coautoría existentes en trabajos previos.

Resumen del capítulo

Se puso en contexto el trabajo realizado, la estructura, se detallaron objetivos e hipótesis a tratar en esta tesis. En los siguientes capítulos desarrollaremos en profundidad lo planteado en los objetivos y verificaremos las hipótesis.

Capítulo 2: Generación del conjunto de datos

Introducción

La Revista Argentina de Antropología Biológica se encuentra en formato digital y consta de una página web en la que están disponibles todos los artículos publicados durante los 20 años considerados para este estudio. Es importante destacar que la utilización de datos públicos, como los artículos publicados en la revista, es un enfoque altamente favorable para la replicación de los resultados obtenidos, así como para la aplicación de la metodología en casos similares. La URL de la página es: <https://revistas.unlp.edu.ar/raab/>

Para realizar la extracción de datos se usó *web-scraping*. Para comprender mejor qué significa esta técnica a continuación se presenta un fragmento de [Saurkar et al.\(2018\)](#), adaptado al español y contextualizado para este caso por el autor de esta tesis:

La extracción de los datos de una página web de forma manual es una tarea exhaustiva y consumidora de tiempo. Debido a esa limitación, se crean las técnicas de web-scraping. Al usar técnicas de web-scraping el usuario puede extraer información disponible en múltiples sitios web en una sola base de datos u hoja de cálculo. La técnica de web-scraping es subdisciplina de la tecnología de minado de la red mining.

Web mining se encuentra entre las técnicas de Recuperación de la información, Extracción de la información y Minado de datos. Tanto la Recuperación como la Extracción de la información cumplen roles importantes en el seguimiento y obtención de información valiosa a partir de datos no estructurados, antes de que esa información esté en condiciones de ser procesada por herramientas y técnicas de Minado de datos.

Podemos decir que son una serie de pasos necesarios para poder disponer de los datos de forma adecuada para un posterior uso en gran parte automatizados. Para este caso específico se busca generar dos estructuras resultantes: una estructura de base de datos relacionales y otra en formato de grafos.

Acorde a [Bourgeois \(2014\)](#) una base de datos relacional es un sistema que organiza los datos en forma de tablas que están relacionadas. Cada tabla está compuesta por columnas o campos que definen la naturaleza de los datos guardados en esa tabla. Luego, un registro corresponde a una instancia particular de un conjunto de campos de una tabla, es decir, son los datos de cada elemento ([Fig. 2.1](#))

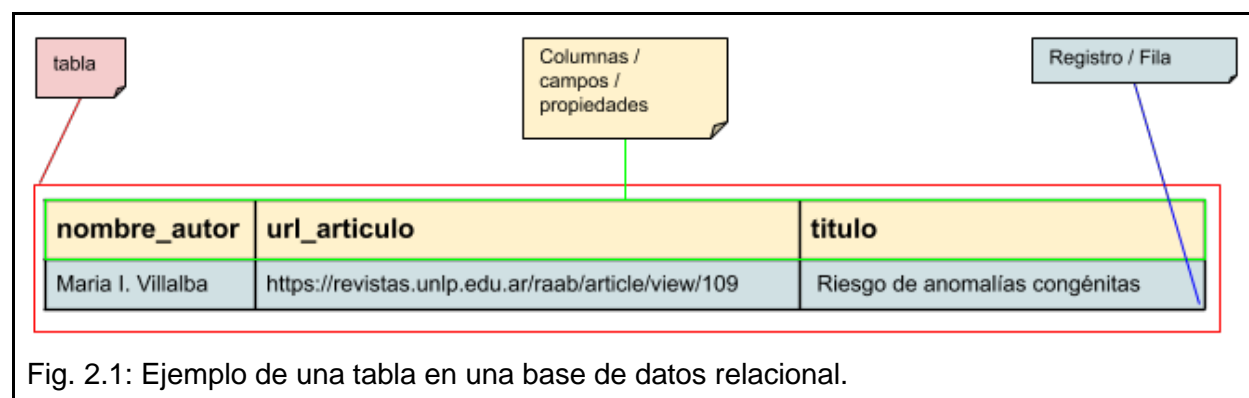
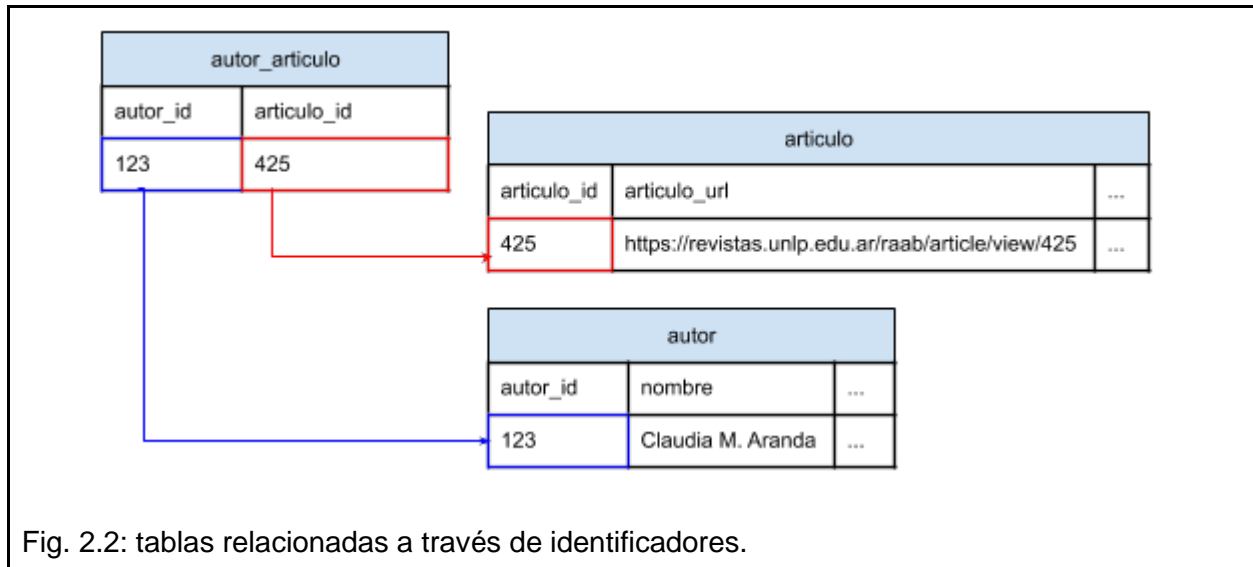


Fig. 2.1: Ejemplo de una tabla en una base de datos relacional.

Las relaciones entre las tablas se realizan a través de identificadores que permiten conectarlas (Fig. 2.2).



En esta tesis uno de los principales objetivos fue encontrar cómo están conectados los autores participantes de la revista, a través de los artículos publicados en ella. Para disponer de una estructura que ayude a comprender este sistema se utilizaron grafos. [Barabási \(2016\)](#) brinda una breve explicación al respecto, adaptada al español por el autor de esta tesis:

Si queremos entender sistemas complejos, primero necesitamos saber cómo sus componentes interactúan entre ellos. En otras palabras, necesitamos un mapa de cómo están conectados. Una red es un catálogo de los componentes de un sistema, comúnmente llamados nodos o vértices, y las interacciones directas entre ellos, llamadas conexiones o aristas. Este tipo de representación de redes ofrece un lenguaje común para estudiar sistemas que pueden diferir enormemente en naturaleza, apariencia o alcance.

Este mismo autor ofrece una consideración interesante respecto de la terminología utilizada en cuanto a redes o grafos [Barabási \(2016\)](#):

¿Redes o Grafos?

En la literatura científica se intercambia frecuentemente la terminología entre redes y grafos:

Redes	Grafos
Nodos	Vertices
Conexiones	Aristas

Sin embargo, hay una pequeña diferencia entre estas terminologías, el conjunto {red, nodo, conexión} se refiere a sistemas de la vida real, mientras que los términos {grafo, vértice, arista} se utilizan para la discusión de representaciones matemáticas de esas redes. Finalmente, esta distinción raramente es tomada en cuenta y se usan como sinónimos.

Por la naturaleza de las redes de autoría, usualmente tenemos dos tipos de nodos: autores y artículos. Sus conexiones son dirigidas y muestran la relación “autor escribe artículo” ([Fig. 2.3](#)).

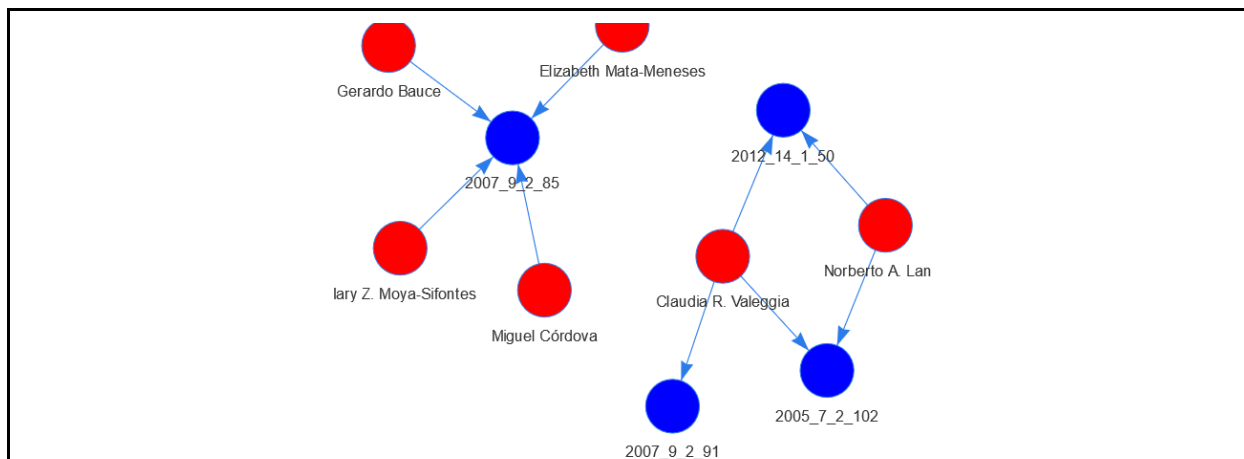
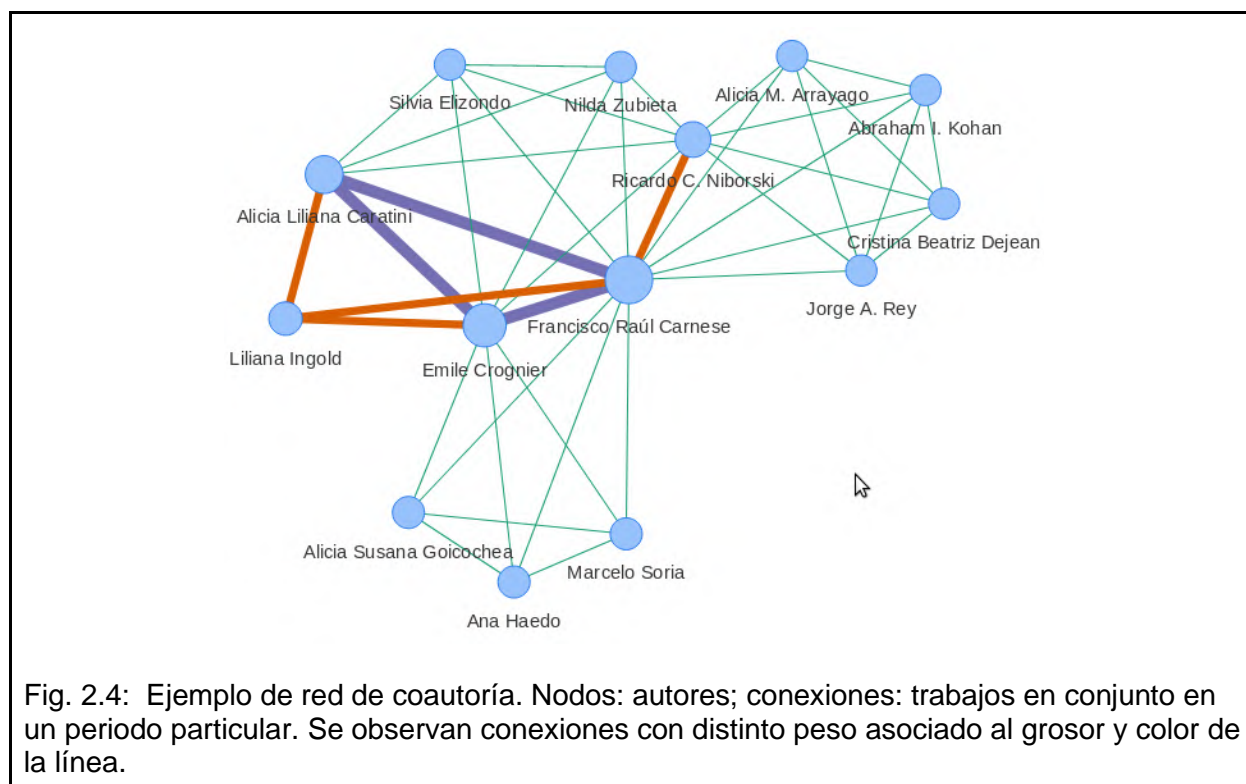


Fig. 2.3: ejemplo de red de autoría con dos tipos de nodos: autores y artículos, y su relación “autor escribe artículo”; Siendo los autores en rojo y los artículos en azul.

Al transformar dicha red en una de coautoría, como menciona [Newman \(2004\)](#): los nodos en la red representan científicos, y la línea entre ellos indica que fueron coautores de por lo menos un artículo durante el periodo analizado.

Para transformar una red de autoría en una de coautoría se realiza la proyección de la red sobre los nodos autores, transformando los nodos artículos en conexiones, que representan cuánto han trabajado en conjunto los autores conectados, y pasando de ser una red dirigida a una red no dirigida, cuyas conexiones tienen un peso acorde a la cantidad de artículos creados por ambos autores ([Fig. 2.4](#)).



Materiales y métodos

Herramientas de trabajo

La generación del conjunto de datos fue armada casi en su totalidad en *scripts* en Lenguaje de programación R con múltiples librerías. Para la sección del procesamiento de autores se utilizó la herramienta Openrefine.

El detalle de herramientas de software utilizadas se destacan en el anexo: [librerías utilizadas](#), donde se puede ver qué librerías se usaron para qué secciones y su correspondiente cita. Los detalles de uso de las herramientas se ven a lo largo de este capítulo.

Introducción al proceso

Se estudió la estructura del sitio web de la RAAB para identificar la información que pudiera ser de utilidad en el presente estudio. Una vez determinados qué datos se iban a extraer, se armaron los *scripts* para realizar *web-scraping*, luego para el procesamiento de los mismos, y finalmente para el almacenamiento en un formato más conveniente.

El foco del *web-scraping* de la página fue extraer los datos de los artículos y las ediciones de la revista (a veces referidos como volúmenes). Los mismos se encontraban en diferentes formatos, principalmente HTML, pero también se encontraron archivos PDF asociados a los artículos y anexos.

Una vez extraídos los archivos de Internet, se creó una estructura de base de datos relacional para almacenar el resultado de la extracción y, de esta forma, facilitar la consulta a los datos obtenidos.

Finalmente se transformó el resultado de la base relacional a un formato de grafos, para su posterior uso en la herramienta de análisis de coautoría facilitando la exploración de los datos.

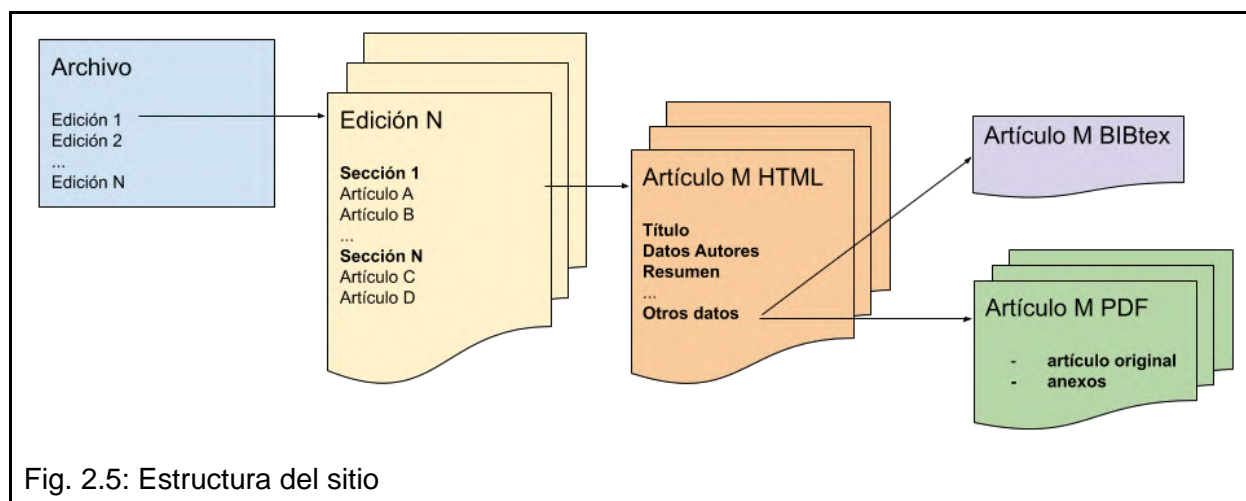
El proceso de adquisición de los datos quedó definido de la siguiente forma:

1. Análisis de estructura del sitio (código fuente de las páginas, navegación del sitio)
2. Adquisición de los datos
 - a. Descarga del archivo de ediciones
 - i. Generación de un índice de ediciones (tabla intermedia)
 - b. Descarga de cada edición
 - i. Generación de un índice de artículos (tabla intermedia)
 - c. Descarga de los artículos asociados a las ediciones
 - i. Descarga de HTML
 - ii. Descarga de PDF
 - iii. Descarga de bibtex
3. Parseo de los datos
 - a. Parseo de ediciones
 - b. Parseo de artículos
 - c. Parseo de bibtex
 - d. Parseo de PDF
4. Generación de la base
 - a. Generación de Resultados intermedios
 - b. Generación de la base relacional
 - c. Generación de los grafos

Análisis de la estructura del sitio

Se exploró la página de la revista para ver cómo es la estructura. Se identificó que existe una página “archivo” en la cual se tiene un listado de todas las ediciones de la revista. Para cada edición de la revista, existe una página en la que se publican datos de la edición y las secciones con los *links* a los artículos correspondientes, así como sus archivos PDF relacionados.

Para cada artículo existe una página en la que se publican datos del artículo junto con un resumen del mismo, los PDF asociados y las citas a los artículos con un *link* para la generación de citas en diferentes formatos ([Fig. 2.5](#)).



Descripción de los datos de interés

Se puede ver en la [Tabla 2.1](#) los datos extraídos, sus orígenes y la razón por la cual resultan de interés.

Tabla 2.1: Datos extraídos del sitio web de la RAAB.

Origen de datos	Razón de interés en el origen de datos	Datos extraídos
Página de archivo	Para poder navegar a las ediciones de la revista, en donde se extraen los datos considerados de interés para este estudio.	Listado de ediciones de la revista, y para cada edición: Nombre de edición, <i>link</i> a la edición
Páginas de cada edición	Para cada edición de la revista se encuentran los datos propios de la edición, secciones que la componen y links a cada artículo.	Fecha publicación edición, Nombre edición, Secciones de la edición: Nombre sección, Pertenencia de artículos a secciones; Artículos en la edición: Título de artículo, <i>Link</i> a artículo, <i>Link</i> a pdfs por artículo, Autores de artículo, Páginas de artículo.
Páginas de Artículos	Para cada artículo se encuentran los datos de interés detallados en la columna de datos extraídos.	Título del artículo, Nombre edición, Sección edición, Autores del artículo, Afiliación Autores (opcional), Resumen (opcional), Abstract (opcional), Pdfs asociados (opcional), Fecha de publicación del artículo, Forma de cita, doi (opcional), <i>Link</i> a artículo (esta página),
Páginas de citas (*1)	Estas páginas no están directamente visibles, pero se pueden extraer por el javascript del <i>link</i> de	author title journal

	<p>cambio de formato de cita; esto permite bajar un HTML con diferentes formatos de citas.</p> <p>Se decidió usar el formato bibtex dado que es muy amigable para la extracción de los datos; además de ser un estándar que se puede transformar a otros en caso de ser necesario.</p> <p>Para cada página de citas se puede extraer el texto de formato bibtex; pudiendo generar el archivo .bib a posterior.</p>	<p>volume</p> <p>number</p> <p>year</p> <p>Keywords (opcional)</p> <p>Abstract (opcional)</p> <p>issn</p> <p>pages</p> <p>url</p>
Archivos PDF	<p>Cada artículo, su texto completo se encuentra en formato PDF. Hay artículos que tienen más de un pdf dado que tienen anexos de información complementaria al principal.</p> <p>Si bien se bajan todos los PDF, todavía no se decide si se va a trabajar solo el principal o la información complementaria también.</p>	El texto completo del artículo

(*1) las páginas de citas ya no están disponibles desde 2018

Adquisición de los datos

La obtención de los datos se realizó siguiendo el plan de acción detallado en la [Fig. 2.6](#). El detalle de los pasos está en la [Tabla 2.2](#).

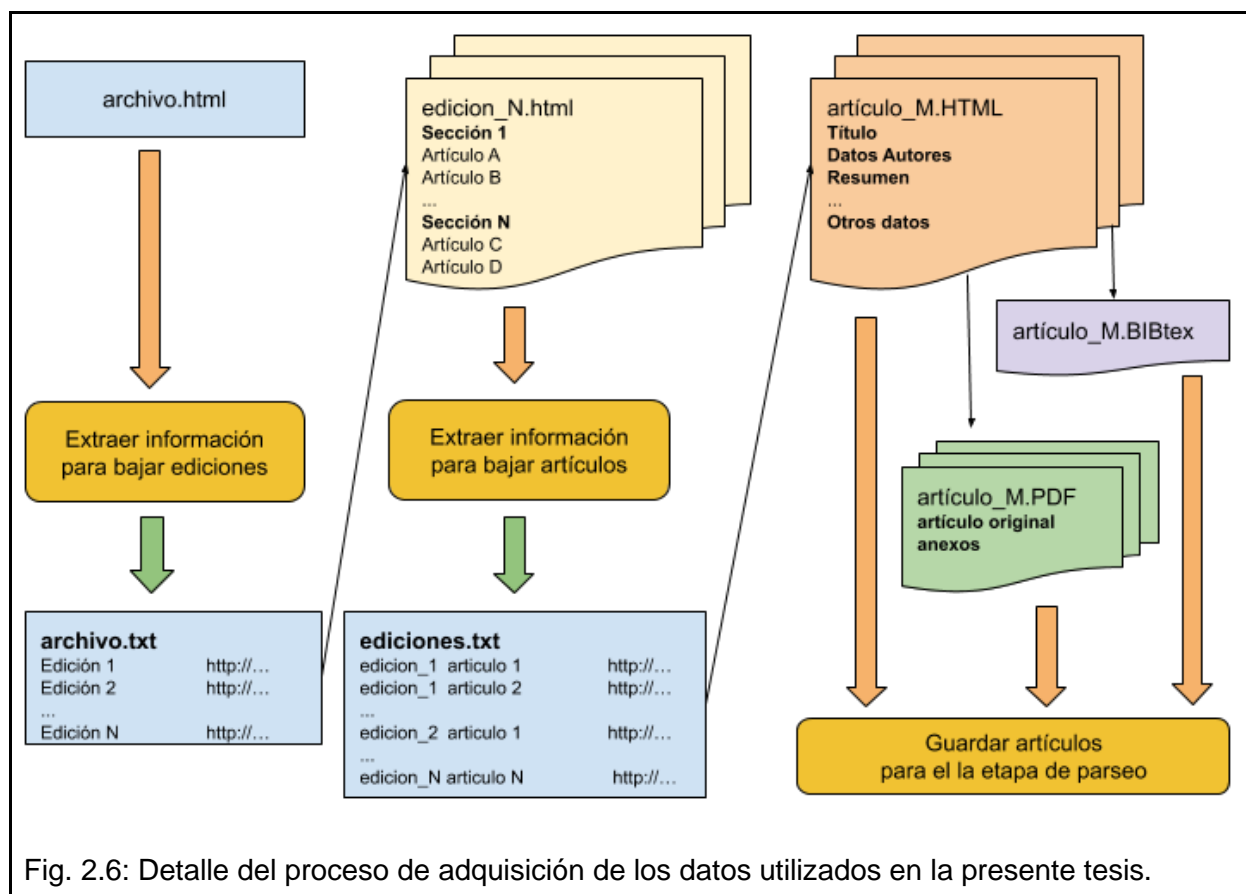


Fig. 2.6: Detalle del proceso de adquisición de los datos utilizados en la presente tesis.

Tabla 2.2: Detalle de las tareas del proceso de adquisición de datos

tarea	descripción	entrada	resultado
Descargar el archivo	Descargar la página HTML del archivo de ediciones de la revista	Url del archivo de ediciones de la revista	archivo.html
Generar un índice de ediciones	Obtiene del archivo las ediciones y la URL de cada una	archivo.html	indice_ediciones.txt
Descargar ediciones	Descargar la página HTML de cada edición.	indice_ediciones.txt	edicion_####.html
Generar un índice de artículos	Obtiene de cada edición, las url de los artículos.	indice_ediciones.txt	indice_articulos.txt
Descargar artículos (HTML) (PDF) (BIBTEX)	Descargar de cada artículo, los archivos HTML / PDF y BIBTEX.	indice_articulos.txt	articulo_####.html articulo_####.bibtex articulo_####.pdf articulo_####_anexo_### #.pdf

Parseo de los archivos

Se habla de *parseo* para referirse a la lectura de los archivos para su transformación desde su formato original a uno más conveniente para la extracción de los datos que ayude a generar la información necesaria para el estudio.

Parseo de los archivos HTML

Para los datos en archivos HTML se utilizó, en el lenguaje R, la [librería rvest](#) que facilita la extracción de este tipo de archivos.

Parseo de los archivos bibTex

Para los datos bibTex, se utilizó primero la [librería rvest](#) para extraer el HTML, luego se corrigieron errores y por último se procesaron con la [librería RefManager](#). Para más detalles ver [ANEXO: Detalle procesamiento bibTEX](#).

Parseo de los archivos PDF

Los archivos PDF debieron ser procesados para extraer los textos en un formato más flexible para la interpretación por una computadora. Por lo tanto, los textos se extraen de formato .pdf a texto plano.

Sin embargo, la particularidad del conjunto de datos en este caso fue que no todos los PDF tenían la misma estructura. Por ejemplo, desde el punto de vista del almacenamiento, algunos archivos son el resultado de un escaneo, mientras que otros son texto. Además, desde el punto de vista de la disposición (*Layout*) del texto, no todos tienen el mismo formato de texto, como espaciados, formato bicolumnar o simple, entre otros detalles. Para más detalle acerca de las dificultades relacionadas al trabajo con archivos PDF ver [ANEXO: Detalle consideraciones PDF](#).

Generación de la Base

Resultados intermedios

Luego del *parseo*, los datos se guardaron en tablas intermedias en *data.frames* de R y a su vez en archivos intermedios con formato de texto separado por tabulaciones. Generar archivos intermedios forma parte fundamental del proceso a fin de no repetir tareas y ahorrar tiempos de procesamiento en caso de querer ejecutar todo el procedimiento nuevamente.

Cada función intermedia fue configurada con parámetros por defecto, que pueden ser sobreescritos en caso de desearlo.

Generación de la base relacional

El armado de las tablas finales requirió de un extenso proceso de estandarización de las referencias a los autores (nombres y apellidos) y sus afiliaciones institucionales. En el caso de los autores se logró de manera exitosa, mientras que en el de las filiaciones surgieron dificultades, como la falta de datos disponibles en la revista, que imposibilitaron la estandarización y, por lo tanto, el uso de esa información. Para más detalle de este proceso se puede ver el anexo [ANEXO: Detalle estandarización autores](#)

Una vez finalizado el proceso de estandarización de autores, las tablas finales se guardaron en una base de datos relacional *SQLite*. Se optó por esta herramienta dado que es fácilmente instalable a través de la [librería RSQLite](#) y brinda la posibilidad de tener un formato de base relacional sin la necesidad de instalar programas externos a R. Asimismo, brinda la potencia del almacenamiento como base de datos relacional en un formato conveniente para su reproducibilidad.

Estructura de la base final

La estructura final fue una N:M, en donde N autores pueden participar en la creación de M artículos. Para representar esta estructura se contó con una tabla de artículos, una tabla de autores, y una tabla intermedia uniendo autores y artículos, en este caso denominada "autores_articulos" ([Fig. 2.7](#)). En la tabla intermedia existieron campos replicados para facilitar las consultas a la base ([Tabla 2.3](#)).

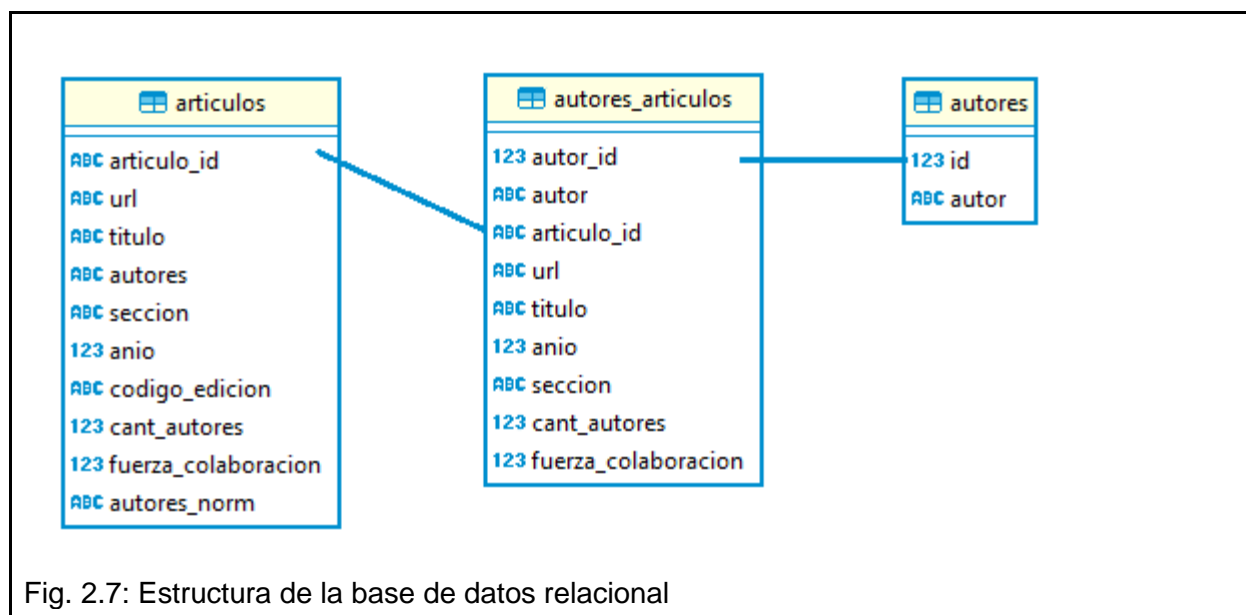


Tabla 2.3: Propiedades de las tablas que conforman la base de datos relacional.

tabla	campo	descripción	Ejemplos / Observaciones
articulos	articulo_id	Identificador de artículo	Ejemplo: 1090, extraído de la página.
articulos	url	Url del articulo	Ejemplo: https://revistas.unlp.edu.ar/raab/article/view/1090
articulos	titulo	Titulo del articulo	Ejemplo: Riesgo de anomalías congénitas en grupos étnicos de Sudamérica / Risk of congenital anomalies in ethnic groups from South America
articulos	autores	Autores del artículo separados por “,”	Ejemplo: Maria I. Villalba;Hebe Campaña;Sandra C. Scala;Mariela S. Pawluk;Jorge S. López-Camelo
articulos	seccion	Sección de la edición	Ejemplo: Trabajos Originales
articulos	anio	Año de publicación en la revista	Ejemplo: 2015
articulos	codigo_edicion	Código de edición	Ejemplo: 2015_17_1
articulos	cant_autores	Cantidad de autores que participaron en el artículo	Ejemplo: 5
articulos	fuerza_colaboracion	Fuerza de colaboración que otorga el artículo por haber participado en él.	Ejemplo: 0.2
articulos	autores_norm	Autores normalizados,	Ejemplo: Hebe Campaña;Jorge

		ordenados en orden alfabético.	S. López-Camelo;Maria I. Villalba;Mariela S. Pawluk;Sandra C. Scala
autores	id	Id del autor	Ejemplo: 1
autores	autor	Nombre del autor	Ejemplo: A. Benítez
autores_articulos	autor_id	Id del autor	1
autores_articulos	autor	Nombre del autor	Campo replicado de la tabla articulos, para facilitar consultas
autores_articulos	articulo_id	Id del articulo	1821
autores_articulos	url	Url del articulo	Campo replicado de la tabla articulos, para facilitar consultas
autores_articulos	titulo	Titulo del articulo	Campo replicado de la tabla articulos, para facilitar consultas
autores_articulos	anio	Año de publicación del artículo en la revista	Campo replicado de la tabla articulos, para facilitar consultas
autores_articulos	seccion	Sección del artículo en la edición	Campo replicado de la tabla articulos, para facilitar consultas
autores_articulos	cant_autores	Cantidad de autores en el artículo.	Campo replicado de la tabla articulos, para facilitar consultas
autores_articulos	fuerza_colaboracion	Fuerza de colaboración que otorga el artículo por haber participado en él.	Campo replicado de la tabla articulos, para facilitar consultas

Generación de los grafos

La transformación de la base de datos relacional a un grafo que facilite su visualización se realizó de la siguiente manera ([Fig. 2.8](#)):

1. Recepción de un parámetro de filtro para el grafo (ej.: periodos a analizar -ver [Capítulo 3](#))
2. Aplicación del filtro a la tabla “autores_articulos” para la preparación de un data.frame que representa los datos necesarios para crear el grafo.
3. Utilización de la [librería igraph](#), para la creación de un grafo bipartito de nodos autores y nodos artículos.
4. Utilización de la [librería igraph](#) para la realización de la proyección sobre los nodos autores, pasando a significar “*estos autores están juntos por haber realizado un artículo en conjunto*”
5. Adjudicación de propiedades a los nodos autores y las aristas relaciones de coautoría.

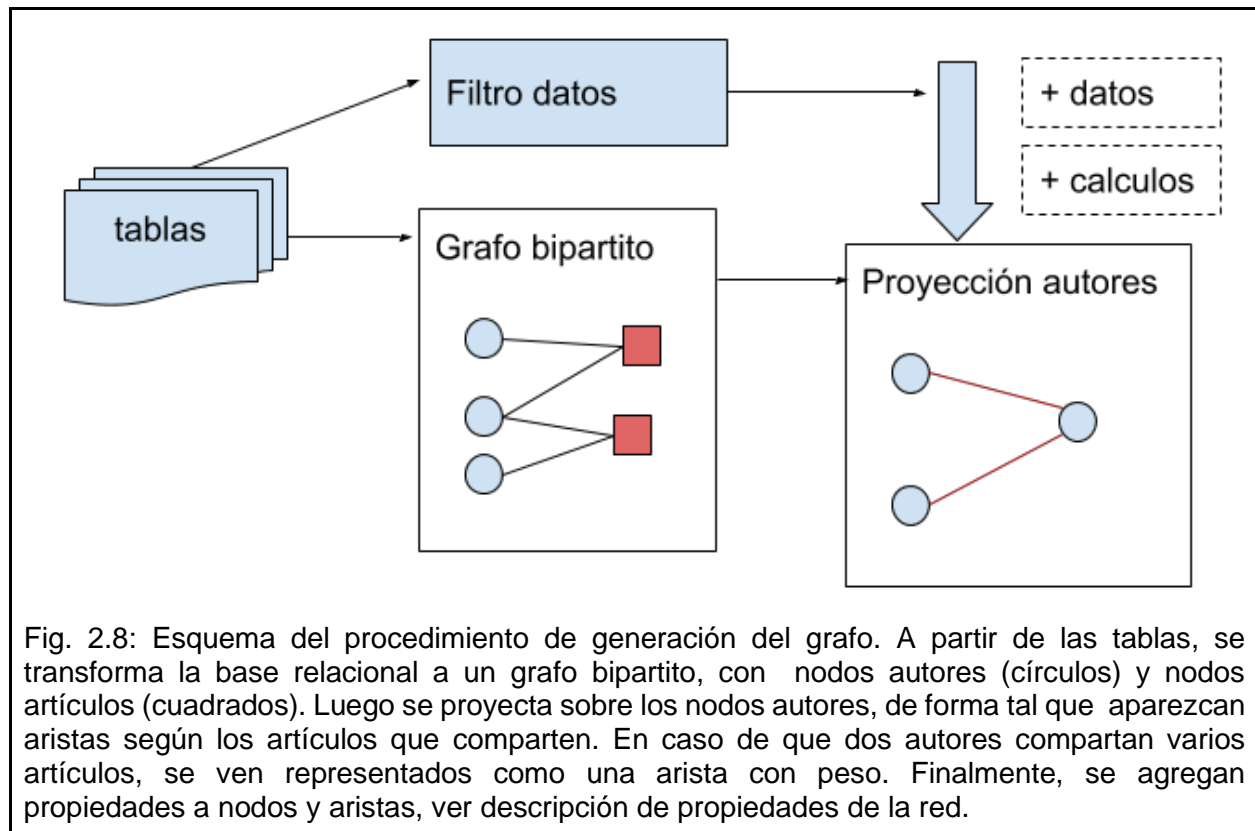


Fig. 2.8: Esquema del procedimiento de generación del grafo. A partir de las tablas, se transforma la base relacional a un grafo bipartito, con nodos autores (círculos) y nodos artículos (cuadrados). Luego se proyecta sobre los nodos autores, de forma tal que aparezcan aristas según los artículos que comparten. En caso de que dos autores compartan varios artículos, se ven representados como una arista con peso. Finalmente, se agregan propiedades a nodos y aristas, ver descripción de propiedades de la red.

Descripción de las propiedades de los nodos y aristas

En este apartado se detallan las propiedades disponibles para los nodos y las aristas. Recordemos que los nodos representan a los Autores, y las aristas representan la relación de coautoría entre los autores que unen. Cada uno de estos, tienen características propias como en el caso de autores pueden ser el nombre y apellido del autor, y en el caso de las aristas pueden ser los autores que la generan. En la [Tabla 2.4](#) podemos ver las propiedades de los autores y en la [Tabla 2.5](#) de las aristas.

Debemos recordar que la generación del grafo se realiza luego de aplicar filtros a los datos, por lo tanto algunos valores son calculados según los filtros aplicados. Para comprender mejor acerca de los filtros ver Capítulo 3 ([Capítulo 3, Estructura de la Interfaz gráfica, Sección 2](#))

Propiedades de los Nodos

Estas propiedades están disponibles para cada nodo autor ([Tabla 2.4](#)).

Tabla 2.4: Propiedades de los Nodos

nombre	descripción	ejemplo
name id	Identificador generado para poder seguir mejor los nodos y los cruces de los mismos.	"a0001"
id_old	Identificador inicial al generar el grafo	"4"
fuerza_colaboracion	Es la suma de fuerza de colaboración del autor en todos los artículos en los que participa. En el grafo actual	0.1666667
cant_autores	Cantidad de autores con los que se relaciona en el grafo actual	6
size	Tamaño que va a ser visualizado. Sirve para que se note la diferencia visual. La fórmula aplicada es: $10 + (\text{fuerza_colaboracion} * 10)$	11.66667
label	Etiqueta del nodo, con el nombre del autor. Se conservó la propiedad como "label" porque no siempre se puede configurar por parámetros a funciones. Como el label es default en muchos lados, se uso esto para simplificar.	"Abraham I. Kohan"
anio	Propiedad que dice el año de participación y cuántos artículos escribió	"1996[1]"

Propiedades de las Aristas

Estas propiedades están disponibles para cada arista que representa la relación de coautoría entre autor 1 y autor 2 ([Tabla 2.5](#)).

Tabla 2.5: Propiedades de las Aristas

nombre	descripción	ejemplo
id	Id generado para la relación	"a0001--a0016"
weight	Peso de la relación, resultante de la proyección del grafo bipartito sobre los nodos autores. Si hay 2 artículos compartidos , el peso es 2	1
color	Código hexadecimal del color a aplicar al edge en la visualización. Se genera a partir fuerza_colaboracion en 3 separaciones. (leves , medias , fuertes)	"#1B9E77"
width	Ancho de la arista en la visualización. Se genera a partir fuerza_colaboracion en 3 separaciones. (leves , medias , fuertes)	1
fuerza_colaboracion	La suma de fuerza de colaboración de los artículos en esa relación. Ej: Artículo 1: fuerza_colaboracion = 0.5 Artículo 2: fuerza_colaboracion = 0.33 Total: fuerza_colaboracion = 0.5+0.33 = 0.83	0.1666667
años	Años en los que fueron realizados los artículos de la relación y cantidad de artículos en esos años	"1996[1]" "2001[1];2005[1];2007[1]" "1996[2];2005[1];"
autores	Autores entre los cuales es la relación	"Abraham I. Kohan - Alicia M. Arrayago"
autor1_label	Nombre del autor 1	"Abraham I. Kohan"
autor2_label	Nombre del autor 2	"Alicia M. Arrayago"

Resumen del capítulo

En este capítulo revisamos las diferentes tareas realizadas para generar una base de datos a fin de poder utilizarla posteriormente en una herramienta de análisis de redes. Dichas tareas comprendieron la extracción desde una página web alojada en Internet, el procesamiento y estructuración de la información para el objetivo específico, así como el armado de documentación que acompañe las tareas realizadas.

En el próximo capítulo veremos cómo la información generada es explotada a través de una herramienta que simplifica el acceso y estudio de los datos para transformarlos en información.

Capítulo 3: Creación de la Herramienta para análisis

Introducción

Como parte del Objetivo 1, Desarrollo de una herramienta para análisis de redes de coautoría, el objetivo de centralizar tareas en una aplicación es disponibilizar tareas rutinarias de análisis de redes de coautoría a través de una interfaz web para facilitar su accesibilidad.

Algunas de las tareas contenidas en la aplicación son:

- visualización general de la red desde diferentes puntos de vista,
- selección de diferentes nodos/aristas/grupos,
- detalles de nodos y aristas, detalles de los datos, entre otras,
- filtros en los datos,
- variación de parámetros para filtros y/o algoritmos.

Se consideró clave facilitar la exploración de los datos mediante interacciones intuitivas brindadas a través de una interfaz web para facilitar el acceso a diferentes personas que realicen los análisis.

En este capítulo se abordarán temas de metodología del trabajo, estructura de la aplicación, y funcionalidades disponibles.

Materiales y métodos

Estructura Interna de la aplicación

Para disponibilizar la herramienta a más de una persona, se decidió que la mejor forma sería mediante una aplicación web. Se entiende por aplicación web a una herramienta de software informático que se encuentra alojada en un servidor, al cual se le hacen consultas a través de clientes generalmente navegadores de Internet, lo cual facilita su acceso por analistas o investigadores desde diferentes lugares en el mismo momento sin necesidad de instalar nuevo software más allá del navegador.

Herramientas de trabajo

La aplicación fue creada enteramente en Lenguaje R con múltiples librerías, que se destacan en el anexo: [librerías utilizadas](#), donde se puede ver cuáles librerías se usaron para qué secciones y su correspondiente cita. Entre ellas se destaca *Shiny*, que permite crear una aplicación web en el entorno de R, facilitando la reproducibilidad al reducir las herramientas utilizadas dado que no se utilizan otros lenguajes para poder crear la aplicación.

Para el uso de las herramientas de trabajo, y buenas prácticas, se usaron recomendaciones destacadas en [Wilson, et al. \(2017\)](#).

Para crear la aplicación se utilizó principalmente el entorno de desarrollo RStudio, dado que facilita considerablemente el desarrollo de aplicaciones y scripts con el lenguaje R.

Se armó la aplicación en forma de proyecto como un Rproject, porque ayuda a gestionar los archivos de código. Esta forma de trabajo resultó útil principalmente para gestionar referencias entre archivos y facilitar el uso de algunas librerías.

Para el versionado del código fuente se utilizó GIT. Si bien no se utilizó desde el inicio, fue necesario a partir de un momento en que el código fue demasiado extenso y se hizo inmanejable el versionado, al igual que, al cometer errores, volver para atrás podía llegar a ser prohibitivo. Para entender un poco más de que se trata GIT debemos entender mejor qué es un Sistema de control de versiones (*Version control system, VCS*), en palabras de [Blischak, et al. \(2016\)](#) traducidas por el autor de esta tesis:

Un sistema de control de versión permite realizar el seguimiento de cambios iterativos que vas armando en tu código. Por lo tanto, podés experimentar con nuevas ideas y siempre tener la opción de volver atrás a una versión específica del código que usaste para generar resultados particulares.

Utilizar un sistema de versiones da otras ventajas además de tener versiones del código, como poseer un orden en la forma de trabajar y, en caso de compartir remotamente, tener *backups* en otras máquinas, lo cual reduce el riesgo de pérdida del trabajo realizado.

Respecto de la estructura de la aplicación

La estructura del proyecto fue cambiando en el tiempo, desde ser un simple *script* pasando a varios *scripts* separando tareas, hasta consolidarse en un proyecto para la aplicación propiamente dicha. Luego se agregaron los conceptos de *Shiny modules* en base a que la aplicación creció considerablemente. La versión final de la aplicación tiene la estructura siguiente:

- proyecto
 - data (Para los datos a utilizar)
 - db_raab_grafos.sqlite
 - temporal_acumulado_data.Rdata
 - temporal_acumulado_dyn_data.Rdata
 - temporal_basico_data.Rdata
 - funcs (funciones y scripts de la aplicación)
 - imports.r (script que tiene todas las librerías a utilizar)
 - globals.r (script que tiene variables globales de la aplicación)
 - funciones.r (script que tiene funciones de la aplicación)
 - ui.r (Script que tiene la Interfaz de usuario)
 - server.r (Script que tiene el lado servidor de la aplicación)
 - shiny_module_estructura_modelos.r
 - (Módulo de Shiny: tiene UI y Server para la sección de modelos)
 - shiny_module_estructura_nodos.r
 - (Módulo de Shiny: tiene UI y Server para la sección de estructura de nodos, se recicla en varios lados de la aplicación)
 - shiny_module_exploratorio_tablas.r
 - (Módulo de Shiny: tiene UI y server para la parte del análisis exploratorio de la aplicación)
 - shiny_module_subgrafos.r
 - (Módulo de Shiny: tiene UI y server para la parte de análisis de subgrafos, se utiliza para la sección de componentes, se podría reciclar en comunidades.)
 - shiny_module_comunidades.r
 - (Módulo de Shiny: tiene ui y server utilizados en la parte de Análisis de comunidades)
 - shiny_module_articulos_text_mining.r

- (Módulo de Shiny: tiene ui y server utilizados en la parte de análisis de comunidades y de módulo subgrafos)
- test (pruebas)
 - múltiples archivos de pruebas de funciones
 - Que devuelvan lo esperado
 - múltiples archivos de prueba de concepto (probar si era viable realizar las funciones / aplicaciones para incluir en la app)
 - Pruebas de concepto y generación de archivos, para probar diversas librerías antes de agregar a la aplicación en sí.
- www (css para estilos personalizados en algunos lugares de la aplicación)
- app.r
 - Es el archivo principal de la aplicación cuando se arma un proyecto Shiny; en este caso en particular se hacen las invocaciones a todos los otros scripts que forman la aplicación y finalmente creación de la app Shiny.

Disponibilidad local, en la red, en internet

Inicialmente se trabajó de forma local, esto significa en la misma computadora donde se crean los scripts, esta forma de trabajo es necesaria para pruebas, sin embargo no es suficiente en caso de querer disponibilizar para otras personas.

Cómo revisa [Wojciechowski, et al. \(2015\)](#), existen varias formas de disponibilizar las aplicaciones desarrolladas con el paquete *Shiny*:

Las aplicaciones Shiny pueden ser compartidas fácilmente compartiendo los scripts ui.R y server.R con un colega que tenga instalado R y la librería Shiny. Sin embargo, no está limitado a esta forma. Las aplicaciones Shiny pueden ser disponibilizadas en servidores web haciéndolas accesibles para aquellos que no están familiarizados con, o no tienen instalado, R.

Hay 3 formas ofrecidas por RStudio para disponibilizar una aplicación web Shiny: Shinyapps.io, Shiny Server, y Shiny server Pro. Los usuarios van a ver estas opciones como más lentas que correrlos en sus propias máquinas, predominantemente dependientes en la velocidad y latencia de la conexión a internet. Cuando una aplicación Shiny se está actualizando se verán los output de la aplicación grisados, indicando qué procesamiento se está llevando a cabo.

Otras consideraciones respecto de los servicios mencionados, como shinyapps.io, es que sirven para hacer despliegues de aplicaciones *Shiny*. El servicio es pago, sin embargo brinda una categoría gratis que da cierta cantidad de horas de uso por mes.

Por este tema se consideró la opción de montar un servidor propio de *Shiny*; como plan de *backup* en caso de quedarse sin crédito y no poder pagar una categoría más alta.

Las 3 opciones implican cierto nivel de aprendizaje, requisitos para poder lograr el objetivo de disponibilizar las aplicaciones. Para ver más detalles acerca de sus ventajas y desventajas referirse al artículo de [Wojciechowski, et al. \(2015\)](#).

Estructura de la interfaz gráfica

La aplicación está dividida en 3 secciones principales

Sección 1: **consideraciones**

Sección 2: **análisis estático**

Sección 3: **análisis temporal**

Para poblar las pantallas de la herramienta de análisis de red de coautoría, se utilizaron las tablas finales de la base de datos relacional. El proceso de generación de grafos se realiza al aplicar ciertos filtros en la aplicación, utilizando los datos de las tablas finales. Además, se filtran estos datos para poder acceder al resultado de forma más eficiente según los criterios escogidos.

Para lograr armar el esqueleto de la interfaz gráfica se utilizaron las librerías [Shiny](#), [shinyDashboard](#) y [shinyDashboardPlus](#).

Sección 1: Consideraciones

En esta sección se encuentran destacadas algunas consideraciones respecto de los datos utilizados, un análisis exploratorio de los mismos, y algunas preguntas para cada tabla de datos.

Las consideraciones a tener en cuenta a la hora de usar la aplicación corresponden específicamente a que se tomaron los trabajos originales publicados por la RAAB durante el periodo entre 1996 y 2016. [Fig. 3.1](#).

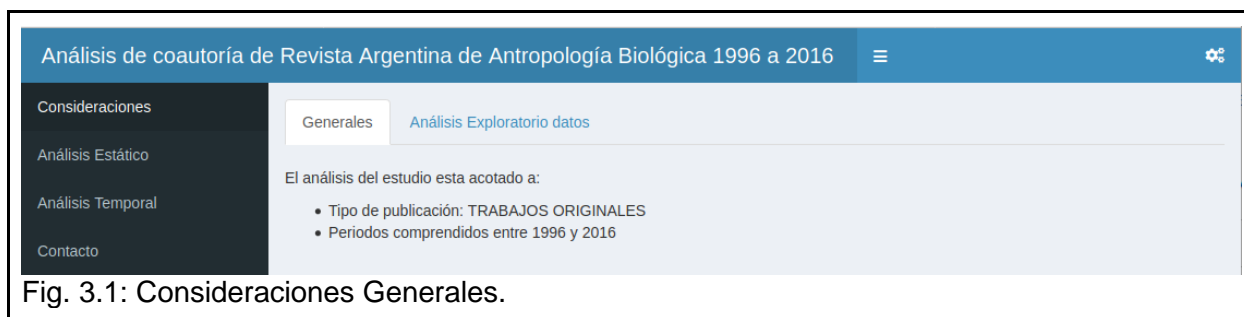


Fig. 3.1: Consideraciones Generales.

En la sección de **análisis exploratorio** se encuentra un “checkbox” ([Fig. 3.2](#) y [Fig. 3.3](#)) donde es posible seleccionar si se desea ver los datos disponibles considerando o no este acotamiento temporal y con respecto al tipo de trabajo propio de la presente tesis. Esta opción fue agregada para que quien realice un análisis utilizando la herramienta sepa que el conjunto de datos completo de la revista contiene más información. La utilidad principal se ve a la hora de evaluar la cantidad de artículos en otras secciones (este punto en particular se abordará más adelante).



Fig. 3.2: Análisis Exploratorio acotado al objetivo de la presente tesis



Fig. 3.3: Análisis Exploratorio expandido a la totalidad de la base de datos obtenida (es decir, hasta marzo 2018 y abarcando todas las secciones)

El análisis exploratorio que se puede ver en la [Fig. 3.4](#) se focaliza en el acceso a los datos crudos, la composición de los datos, un resumen estadístico y algunas preguntas básicas de interés. Analizaremos estos valores en el próximo capítulo, por el momento solo se muestran imágenes para familiarizarse con la estructura de la interfaz gráfica.

Datos Crudos

Refiere a la visualización de los datos en forma de tabla interactiva en la cual se pueden aplicar distintos filtros. Para lograr esta utilidad se utilizó la [librería DT](#) (por *data tables*). La herramienta permite bajar los datos, buscar por palabras que figuren en cualquier columna, entre otras.

articulo_id	url	titulo	autores	seccion	anio	codigo_edicion	cant_autores	fuerza_colaboracion	autores_norm
239	https://revistas.unlp.edu.ar/raab/article/view/239	Nuevas fuentes de evidencia para la paleoparasitología y la antropología biológica en Patagonia: resultados preliminares	Martín H. Fugassa, Armando Cicchino, Norma H. Sardella, Ricardo A. Guichón, Guillermo M. Denegrí, Adauto Araújo	Trabajos Originales	2007	2007_9_2	6	0.166666666666667	Adauto Araújo, Armando Cicchino, Guillermo M. Denegrí, Martín H. Fugassa, Norma H. Sardella, Ricardo A. Guichón
281	https://revistas.unlp.edu.ar/raab/article/view/281	Nuevos aportes a la paleoparasitología del sitio arqueológico (Nombre de Jesús), Cabo Virgenes, Argentina	Martín H. Fugassa, Ricardo A. Guichón	Trabajos Originales	2006	2006_8_1	2	0.5	Martín H. Fugassa, Ricardo A. Guichón

Mostrando registros del 1 al 2 de un total de 2 registros (filtrado de un total de 171 registros)

Anterior 1 Siguiendo

Fig. 3.4: Visualización de la sección “Datos crudos”. A modo de ejemplo, se muestra la aplicación de un filtro por palabra.

Composición

Esta opción muestra un análisis visual de cómo es la composición de las tablas, en cuanto a tipos de datos y cantidad de observaciones. Para lograr esta visualización se utilizó la librería [visdat](#) (Fig. 3.5).



Fig. 3.5: Vista de la composición de los tipos de datos por columnas de una tabla específica.

Resumen

Permite ver un resumen estadístico de la distribución de los datos en la tabla. Para lograr esta funcionalidad se utilizó la librería [skimr](#). Mediante esta opción se pueden ver las estadísticas que se utilizan para cada tipo de dato ([Fig. 3.6](#)).

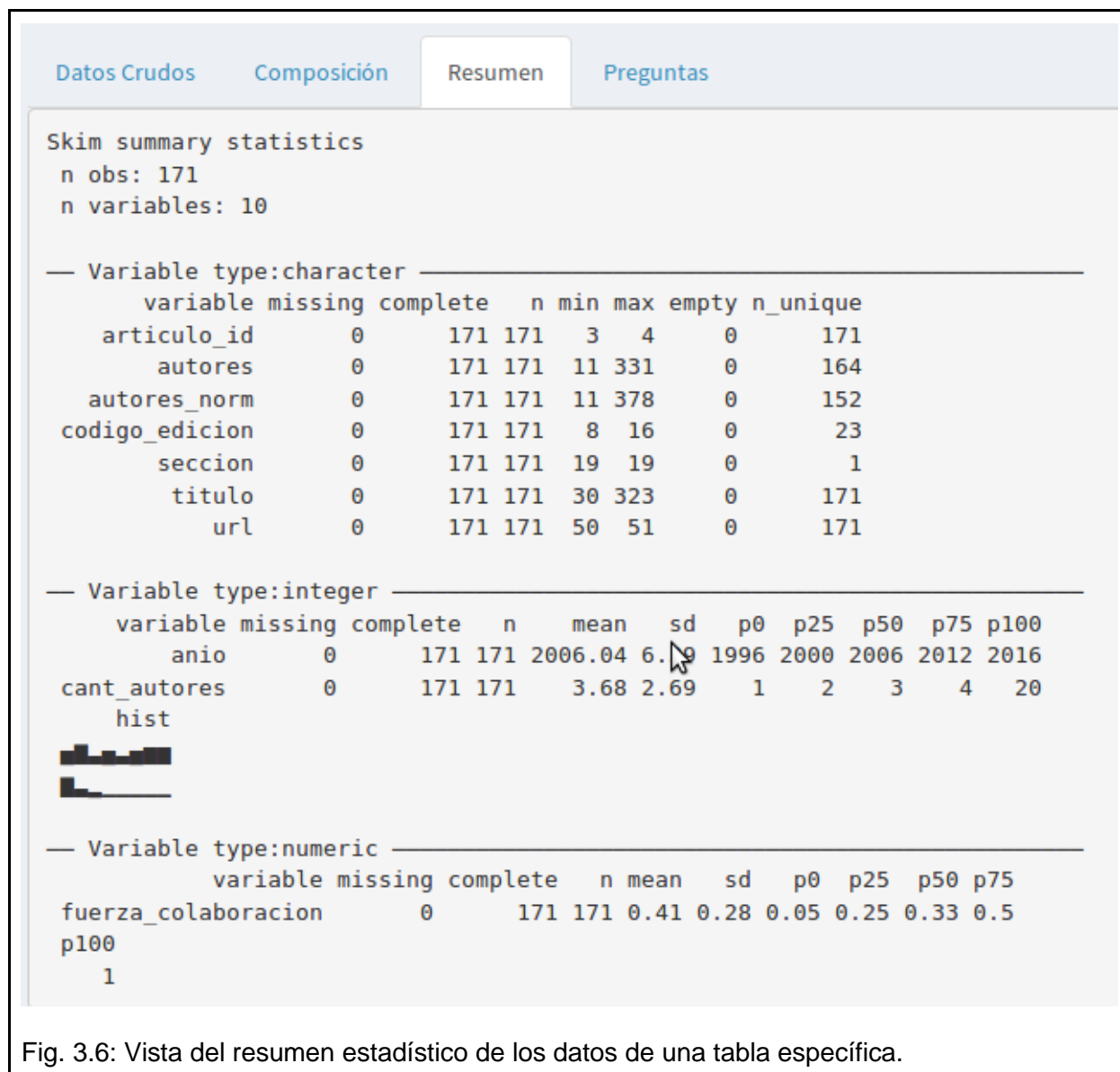


Fig. 3.6: Vista del resumen estadístico de los datos de una tabla específica.

Preguntas

Este ítem refiere a preguntas básicas que salieron al ver los datos de cada tabla. Estas preguntas son utilizadas a lo largo de la tesis, dado que se usan para familiarizarse con los datos. Para poder hacer los gráficos interactivos en la sección se utilizó la librería plotly ([Fig. 3.7](#)).

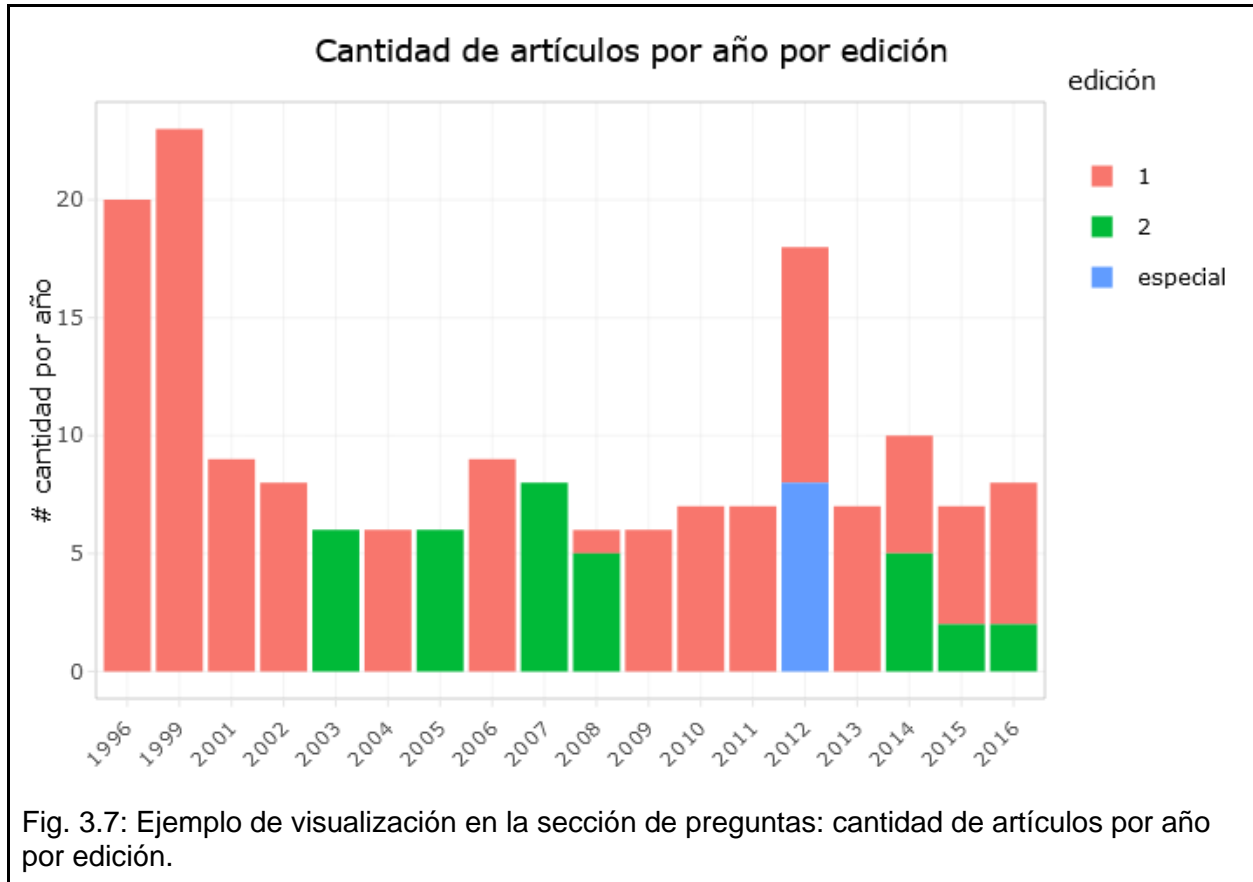


Fig. 3.7: Ejemplo de visualización en la sección de preguntas: cantidad de artículos por año por edición.

Las preguntas que se hicieron en este análisis están en la [Tabla 3.1](#).

Tabla 3.1: Preguntas Análisis Exploratorio.

Tabla	Pregunta
Autores	¿Cuántos autores hay en total de la base?
Artículos	¿Cuántos artículos hay en total de la base?
Artículos	Cantidad de artículos por año
Artículos	Cantidad de artículos por sección (tiene sentido cuando no se filtra por sólo Trabajos originales)
Artículos	Cantidad de artículos por año por edición
Autores - Artículos	¿Cuántas autorías hay en total de la base?
Autores - Artículos	¿Quiénes son los 20 autores con más artículos?
Autores - Artículos	¿Cuál es la frecuencia de artículos por autor?
Autores - Artículos	¿Cuáles son los 20 artículos que tienen más autores?
Autores - Artículos	¿Cuál es la frecuencia de autores por artículo?
Autores - Artículos	Cantidad de autores por artículo por periodo agrupados (1, 2 a 4, 5 a 7, 8 o +)

Con esto concluimos la sección 1 de **consideraciones**, el objetivo de esta sección es relacionarse inicialmente con los datos, ver cómo están estructurados, y una primera aproximación a cuáles son preguntas que se pueden contestar.

Luego de estar familiarizados con los datos se puede proceder tanto a la sección 2 como la 3 dependiendo de si se desea estudiar uno o varios períodos en particular, o analizar la evolución temporal de los datos, respectivamente.

Un ejemplo es explorar la evolución temporal, para luego focalizarse en algún periodo particular.

Sección 2: Análisis Estático

El objetivo de esta sección es poder configurar, filtrar, visualizar, agrupar y estudiar métricas de la red en detalle. Para realizar estas tareas esta sección se divide en varias subsecciones: Visualización de coautoría, Artículos asociados, Estructura de la red, Comparación con modelos y Comunidades. En la [Fig. 3.8](#) podemos ver la pantalla de Visualización de coautoría.



Fig. 3.8: Ejemplo de resultado obtenido en la sección de Análisis estático > Visualización de coautoría > Red.

Configuración

Para configurar el análisis a realizar, se debe hacer click sobre el engranaje como se ve en la [Fig. 3.9](#). En esta sección se configura la semilla (*seed*) para asegurar la reproducibilidad en partes aleatorias de las ejecuciones, como podrían ser algunas de las “disposiciones” de la red (*layouts*) o las ejecuciones de algunos algoritmos, por ejemplo, en la parte de Comunidades, las “caminatas aleatorias cortas” (*short random walks*).

Luego, la opción periodos permite seleccionar los periodos a analizar. Al hacerlo, se desprenden opciones de la red resultante en particular, como ser la disposición de la visualización de Red, si se desea seleccionar algún autor en particular, y la opción de búsqueda de comunidades en la red actual.

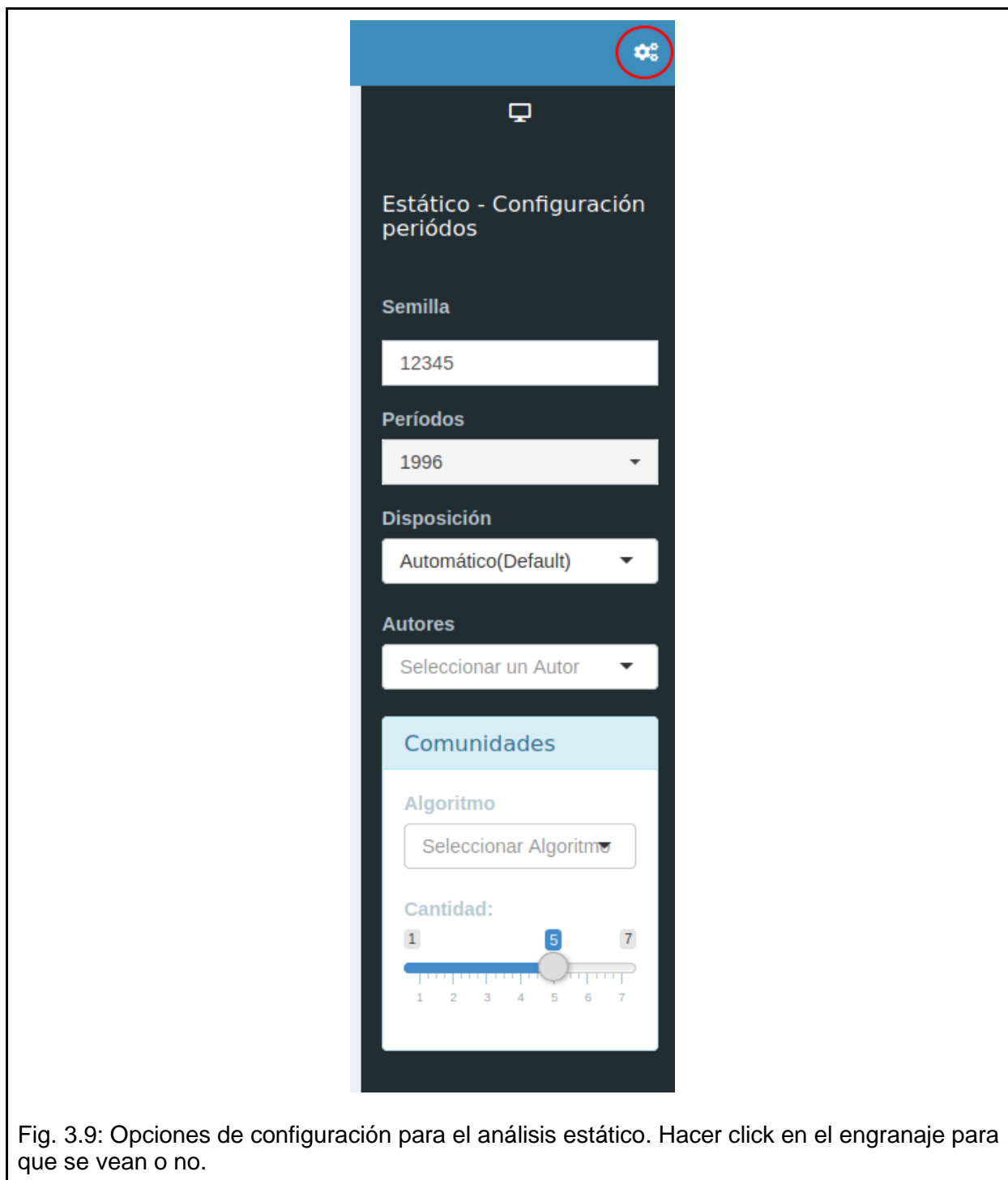


Fig. 3.9: Opciones de configuración para el análisis estático. Hacer click en el engranaje para que se vean o no.

Visualización

Utilizando la aplicación es posible obtener diferentes tipos de visualización del grafo: la vista de red, Biofabric y Mapa de calor (*heatmap*). Esta sección fue armada con controles de la [librería shiny](#).

La visualización en forma de red ([Fig. 3.10](#)) muestra a los Autores como nodos y su relación de coautoría en forma de vértices. El tamaño de los nodos representa la suma total de fuerza de coautoría del autor. El color de las aristas está basado en la suma de fuerza de colaboración de la relación y pueden seleccionarse tres categorías relativas al grafo actual (que representan baja, media o alta relación de coautoría) para filtrar la visualización de aristas. Esta visualización fue realizada con la librería *visnetwork*.

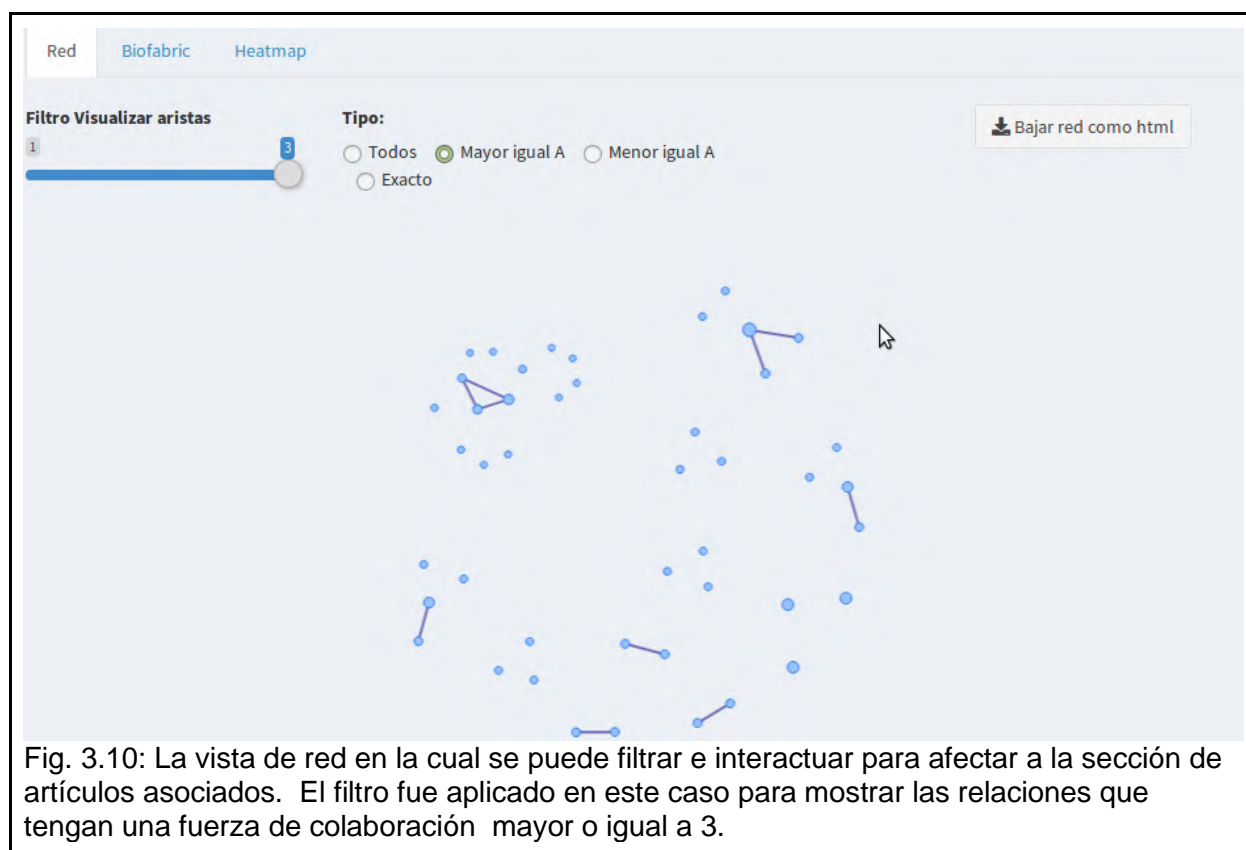
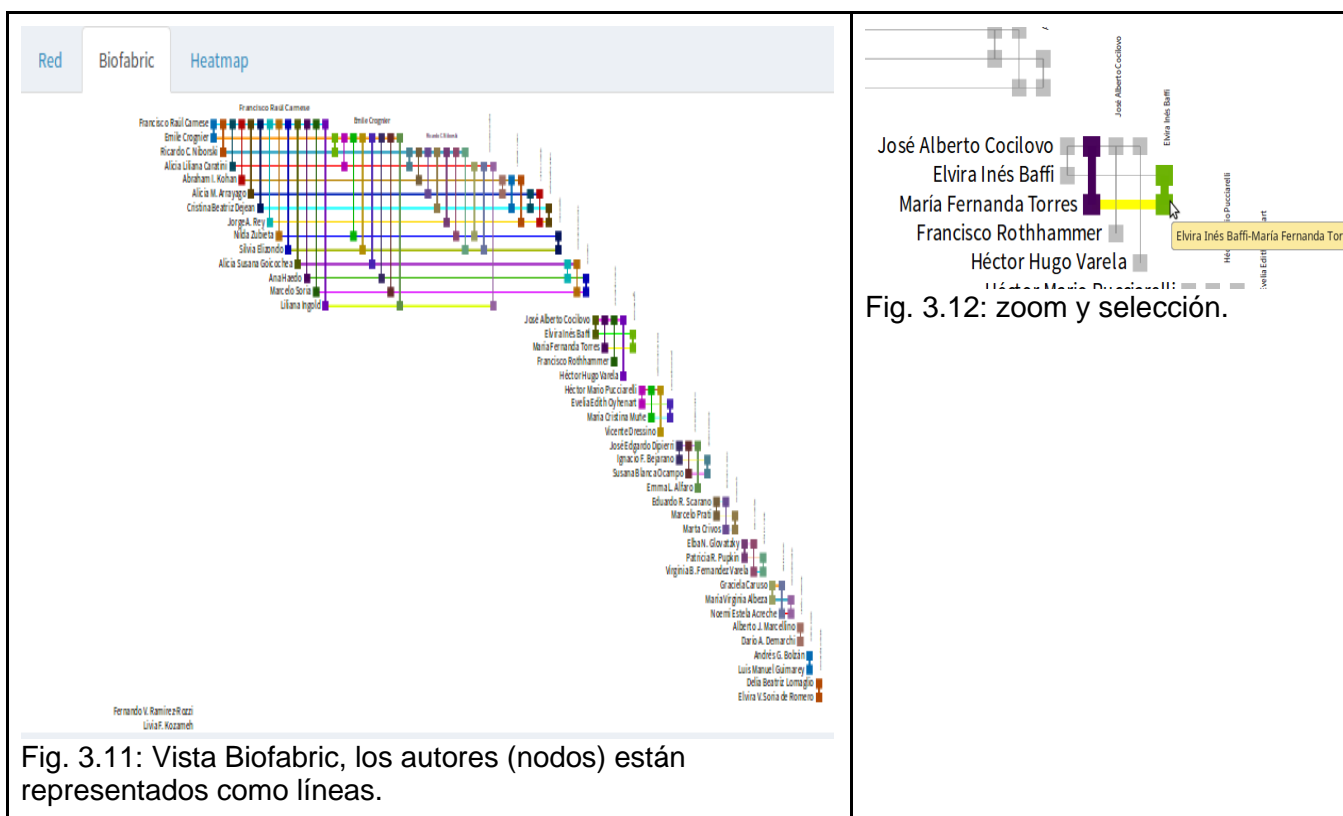
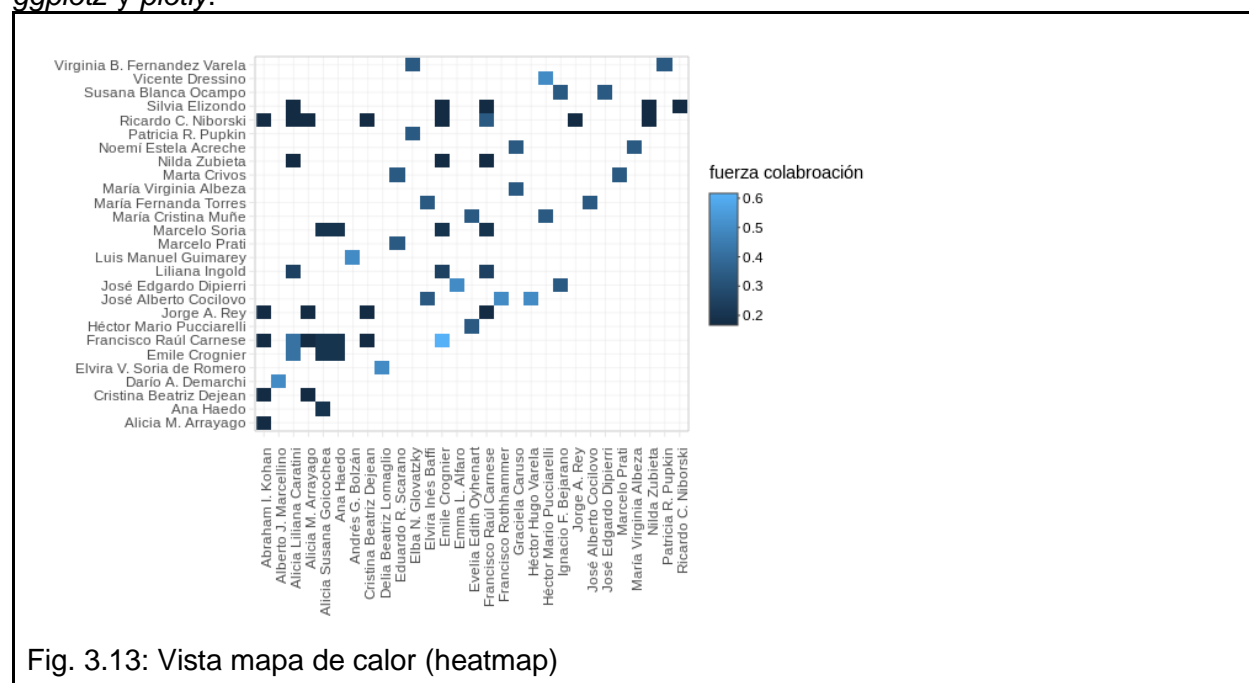


Fig. 3.10: La vista de red en la cual se puede filtrar e interactuar para afectar a la sección de artículos asociados. El filtro fue aplicado en este caso para mostrar las relaciones que tengan una fuerza de colaboración mayor o igual a 3.

La Visualización Biofabric ([Longabaugh 2012](#)), es otra opción interactiva para ver las relaciones entre autores ([Fig. 3.11](#) y [Fig. 3.12](#)). En esta se toman los nodos como líneas evitando, de esta forma, el problema de tener muchos vértices superponiéndose en forma desordenada. Esta sección fue armada con la adaptación de R disponible en la librería *RBioFabric*.



Finalmente, la visualización de tipo mapa de calor (*heatmap*) (Fig. 3.13) permite observar la fuerza de colaboración entre pares de autores. Esta adaptación fue realizada con las librerías *ggplot2* y *plotly*.



Artículos asociados

Esta sección está asociada a la visualización en formato de red. Al seleccionar un nodo Autor, se muestran los datos asociados a dicho Autor incluyendo las visualizaciones de subgrafo y heatmap. En el resumen se muestra quién es el Autor seleccionado, en qué periodos participó, la cantidad de artículos relacionados con dicho autor, la cantidad de autores asociados y la fuerza colaboración total para dicho autor ([Fig. 3.14](#)). Es importante considerar que cualquier métrica o valor tendrá aplicado el filtro correspondiente a los períodos seleccionados en la configuración de la sección de análisis estático. La realización de esta visualización está basada en el artículo de [Newman \(2004\)](#).

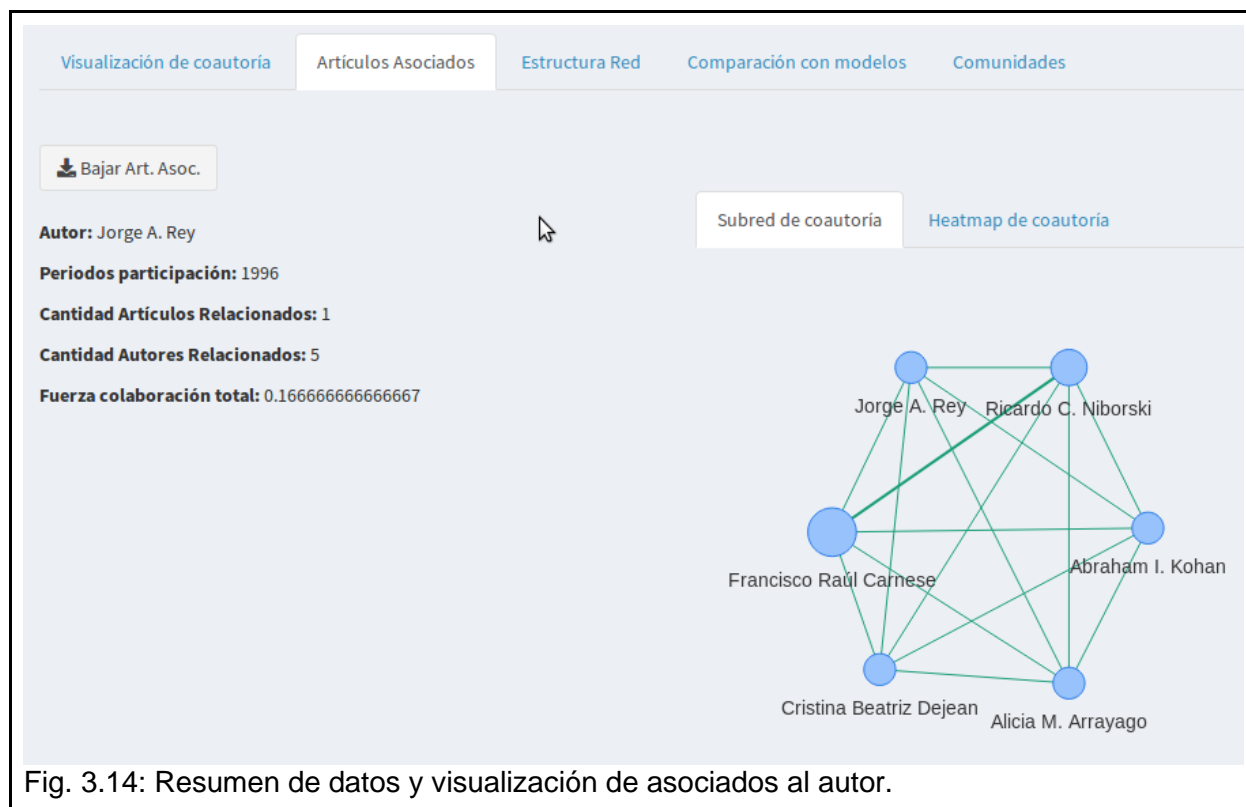


Fig. 3.14: Resumen de datos y visualización de asociados al autor.

En esta sección se muestra, además, un detalle de los datos asociados al autor en el período seleccionado ([Fig. 3.15](#)), con la intención de comprender mejor con qué datos específicos están asociadas las métricas y visualizaciones obtenidas.

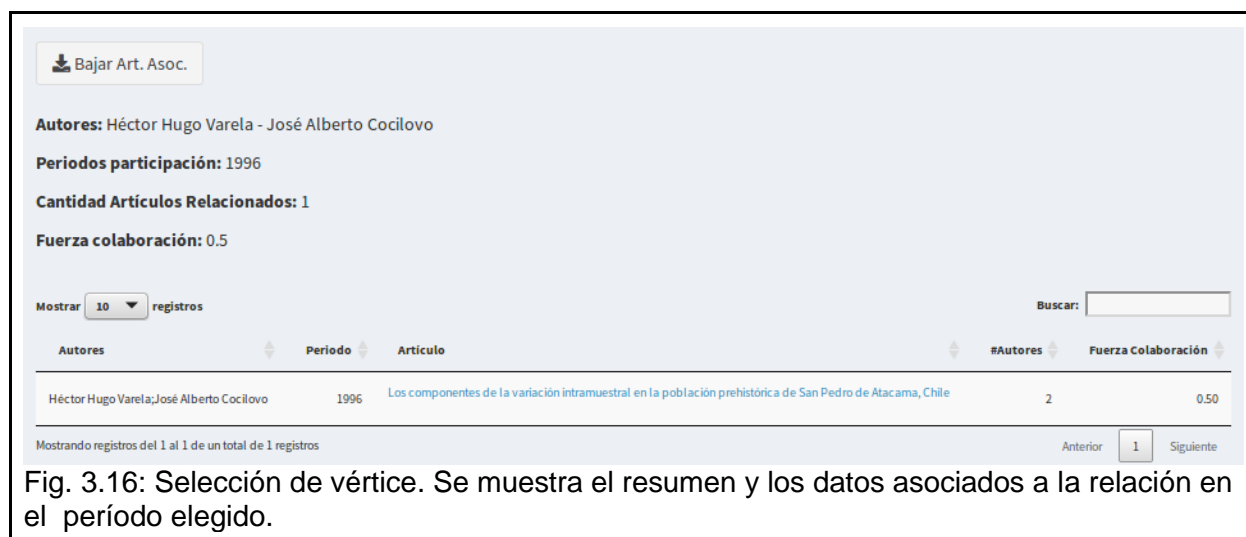
The screenshot shows the 'Selección de autor' (Author Selection) section. It displays a table of associated articles for the selected author. The table has the following columns: Autores, Período, Artículo, #Autores, and Fuerza colaboración. The data row shows the following information:

Autores	Período	Artículo	#Autores	Fuerza colaboración
Abraham I. Kohan; Alicia M. Arrayago; Cristina Beatriz Dejean; Francisco Raúl Carnese; Jorge A. Rey; Ricardo C. Niborski	1996	Perfil infectológico de poblaciones mapuche de Cerro Policia y Aguada Guzmán de la Provincia de Río Negro	6	0.17

At the bottom of the table, it indicates 'Mostrando registros del 1 al 1 de un total de 1 registros' (Showing records from 1 to 1 of a total of 1 records). Navigation buttons for 'Anterior' (Previous) and 'Siguiendo' (Next) are also present.

Fig. 3.15: Selección de autor. Se muestran los datos asociados al autor en el período seleccionado.

Por otro lado, al seleccionar un vértice se pueden ver los autores involucrados en la relación, los períodos de participación de esa relación, la cantidad de artículos relacionados y finalmente la fuerza colaboración asociada a esa relación (Fig. 3.16).



Estructura de la red

En esta sección se pueden obtener métricas asociadas a la red, los nodos y los componentes, que potencialmente sirvan para fundamentar conclusiones extraídas a partir de las visualizaciones.

La gran mayoría de las métricas se obtuvieron con la [librería igraph](#), exceptuando a la fuerza de colaboración, que fue calculada por el autor de esta tesis en base a los cálculos especificados por [Newman \(2004\)](#).

Métricas de la red

Esta sección dispone de diferentes métricas que se utilizan habitualmente en análisis de redes, que fueron seleccionadas en base a las propuestas por [Newman \(2004\)](#). Para este caso bajo estudio, se quitó el Promedio de colaboradores, dado que en la revista no se cuenta con información explícita para una correcta distinción entre el autor principal y los colaboradores sino que solamente figura el orden en la página de internet. Además se agregaron otras métricas como: el porcentaje que representa el *cliqué* más grande, la densidad de la red, la cantidad de autores en el componente más grande, la cantidad de componentes, la cantidad de contribuciones y la cantidad de relaciones ([Tabla 3.2](#)).

Tabla 3.2. Descripción de las métricas disponibles en la sección Métricas de la red.

Métrica	Descripción
# Aut.	Cantidad de autores en la red
# Art.	Cantidad de artículos en la red
# Contrib.	Cantidad de contribuciones en la red: la suma de las cantidades de autores por artículo.
# Rel.	Cantidad de relaciones en la red: por cada artículo en conjunto se genera una relación entre los diferentes autores. La suma por artículo de: $n_{\text{autores}} * (n_{\text{autores}} - 1) / 2$
# Media de Art. por Autor	Media de artículos por autor
# Media de Aut. por Art.	Media de autores por artículo
# de Comp.	Cantidad de componentes en la red. Un componente es un grupo de nodos conectados.
# Aut. en Comp. Grande	Cantidad de autores en el componente más grande, es decir, la mayor cantidad de autores conectados.
Comp. más Grande (%)	Porcentaje del componente más grande. (respecto del total de nodos en la red)
Densidad	Densidad de la red: La densidad de un grafo es la relación del número de aristas y el número de aristas posibles.
Dist. Media	Distancia media entre nodos conectados en la red.
Diámetro	Diámetro de la red: Es la geodésica más larga. La geodésica es la mínima longitud para conectar 2 puntos.
Cliqué más grande (%)	Cliqué más grande expresado en porcentaje $n_{\text{autores_clique_maximo}} / n_{\text{autores}} * 100$, recordemos que un clique es un subgrafo completo dentro de la red en cuestión. Completo significa que todos los que son parte de ese subgrafo están conectados entre sí. Esta medida representa que tanto de la red representa el máximo subgrafo completo
Tran.	Transitividad: es la probabilidad media que dos coautores también sean coautores con otro en común.
Asort.	Asortatividad: es el coeficiente de correlación de Pearson entre los grados de los nodos adyacentes en la red.

En la ([Fig. 3.17](#)) se presenta cómo se visualizan estas métricas en la aplicación. Estos datos se pueden exportar mediante el botón *Bajar est. red.* disponible arriba de la tabla.

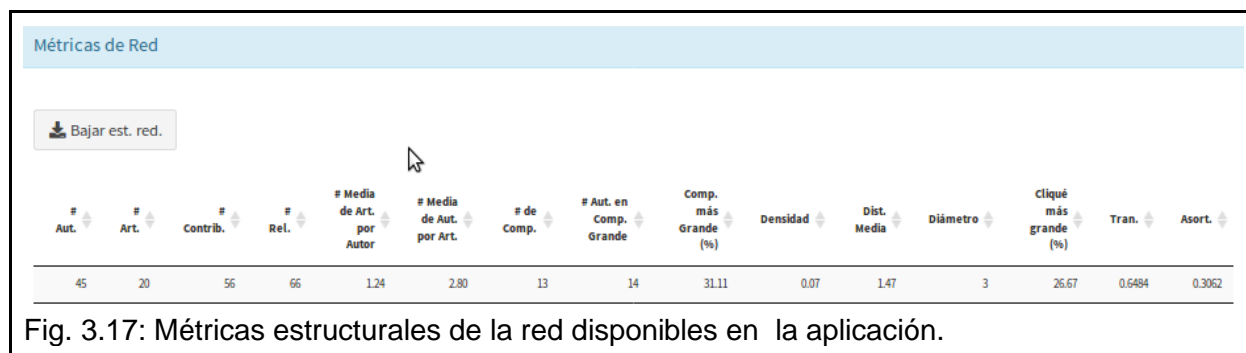


Fig. 3.17: Métricas estructurales de la red disponibles en la aplicación.

Métricas de los nodos

Esta sección incluye distintas “solapas” donde se puede acceder a los datos (es decir, las métricas propiamente dichas), la distribución de grado, el histograma de variables, y la densidad de variables. Para una revisión de las medidas dentro de las redes se puede consultar a [Newman \(2016\)](#) y a [Freeman \(1978\)](#). Las métricas que se incluyeron se describen en la [Tabla 3.3](#).

Tabla 3.3. Descripción de las métricas de los nodos.

Columna	Descripción
Autor	El autor representado por ese nodo.
Fuerza Colaboración	(<i>Collaboration Strength</i>) Suma total de la fuerza de colaboración por cada artículo del autor. Recordemos fuerza de colaboración = 1 artículo / cantidad de autores. Fuerza de colaboración total: suma de fuerza colaboración de cada artículo que participó el autor
Grado	(<i>Degree</i>) El grado del nodo: la cantidad de conexiones directas del nodo.
Intermediación	(<i>Betweenness</i>) La importancia de un nodo según su participación en caminos para llegar de un nodo a otro.
Centralidad Autovector	(<i>Eigen centrality</i>) Además de la cantidad de conexiones valora la calidad de las mismas. Si un nodo se conecta con otros nodos con gran centralidad de autovector, se posiciona mejor.
Cercanía	(<i>Closeness</i>) Es la media de la distancia geodésica entre un nodo y el resto de los nodos. La distancia medida en cantidad de nodos participantes de los caminos. Se recomienda utilizarla cuando todos los nodos están conectados.
Page Rank	Es una variante de centralidad autovector.
# Triángulos	(# <i>Triangles</i>) Cantidad de triángulos de los que participa el autor. Un triángulo son 3 nodos completamente conectados.

En la [Fig. 3.18](#) se muestra la disposición en la aplicación de la sección Métricas de Nodos. Para la sección de datos está disponible el cuadro de búsqueda que agiliza la búsqueda de algún autor en particular.



Fig. 3.18: Métricas de los nodos, sección Datos.

La distribución de grado es un gráfico ampliamente utilizado en el estudio de redes. Analizando la distribución se puede ver qué tipo de nodos predomina más dentro de una red en particular (Fig. 3.19).

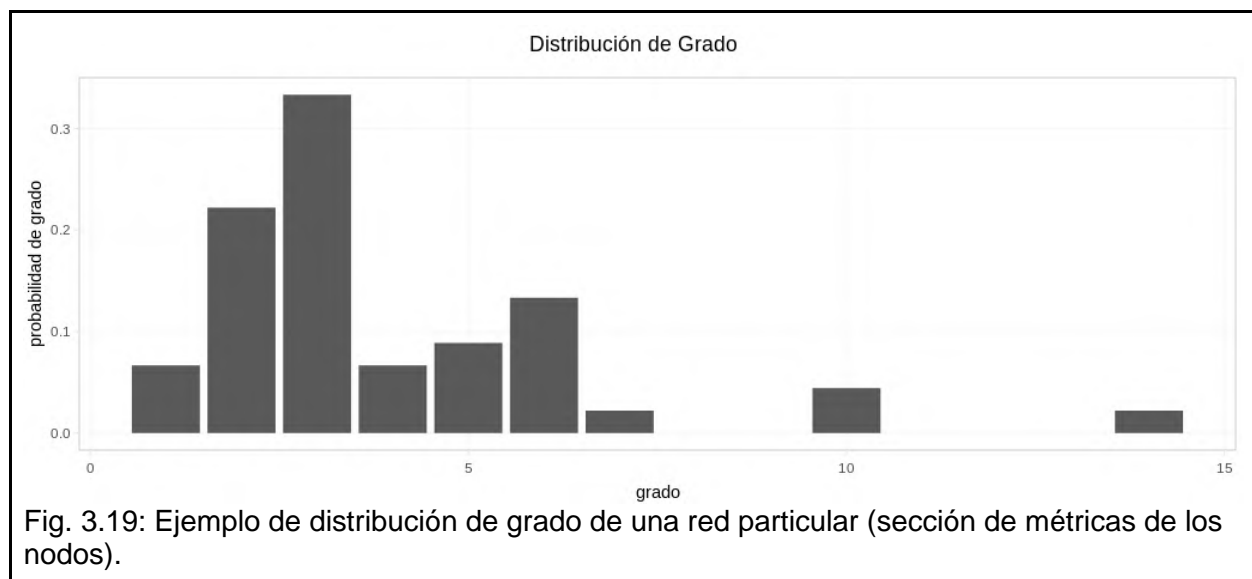
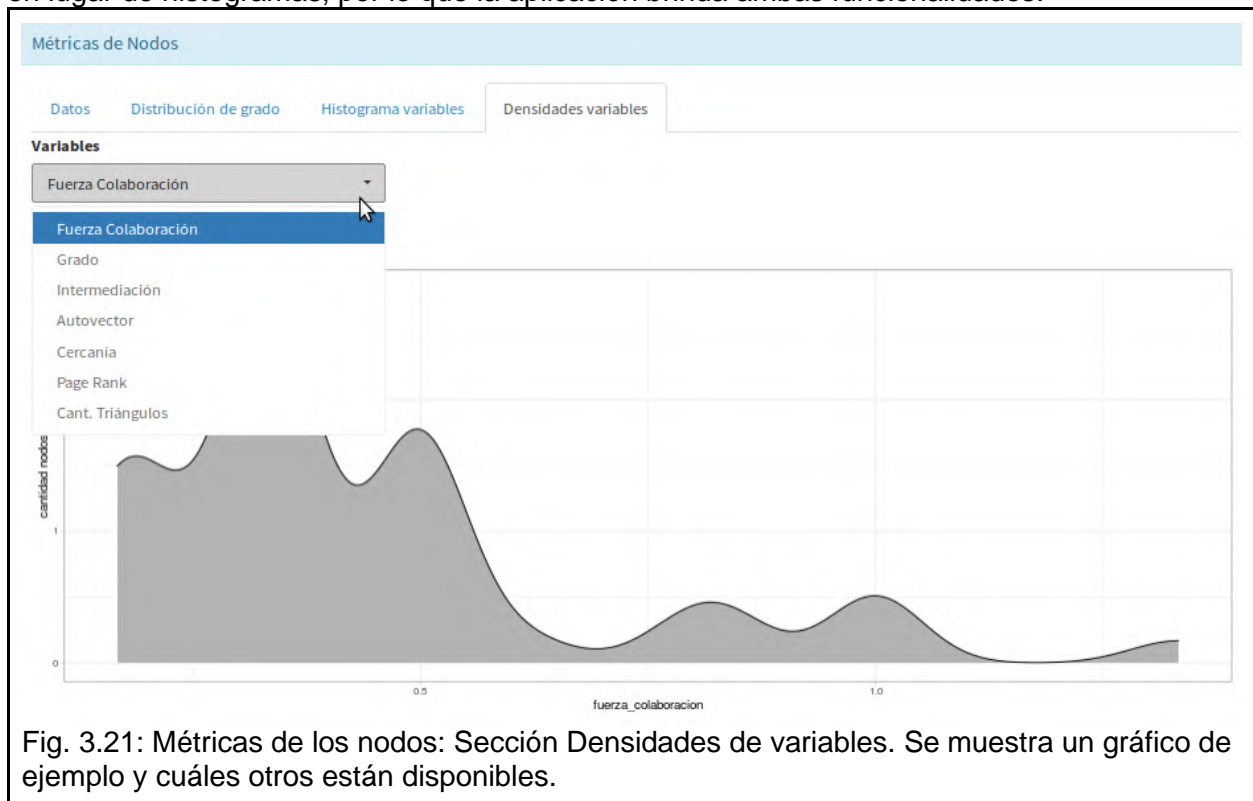


Fig. 3.19: Ejemplo de distribución de grado de una red particular (sección de métricas de los nodos).

En la sección Histograma variable (Fig. 3.20) podemos obtener los histogramas de las métricas que se mencionaron anteriormente para poder visualizar de forma gráfica los datos.



Estas métricas son muchas veces representadas en forma de gráfico de densidades ([Fig. 3.21](#)) en lugar de histogramas, por lo que la aplicación brinda ambas funcionalidades.



Detalles de los componentes

Un componente es una porción de red en donde todos los nodos están conectados directa o indirectamente por lo menos por una arista ([Morel et al. 2009](#)).

El análisis de componentes da un vistazo general de la estructura de la red ([Fig. 3.22](#)) y provee información valiosa en relación a su nivel de fragmentación ([Morel et al. 2009](#)).

Para extraer los componentes se utilizó la [librería igraph](#).

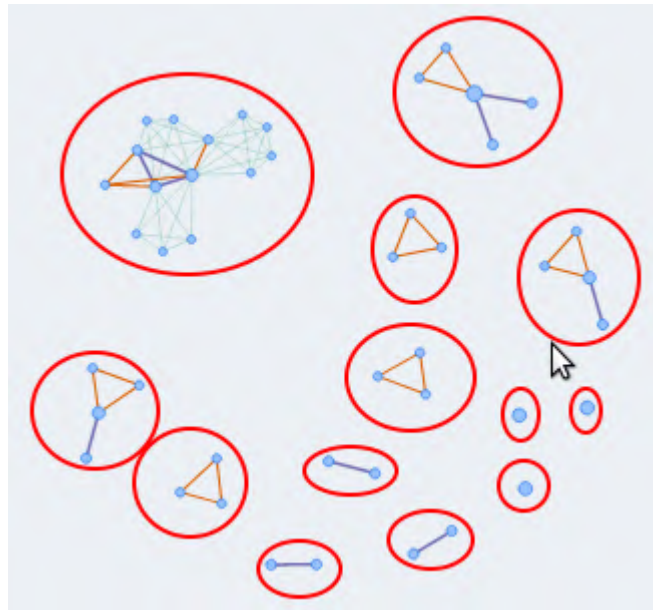


Fig. 3.22: Ejemplo de gráfico de componentes obtenido para la red de la RAAB correspondiente al año 1996. Los componentes se resaltan en rojo.

En esta sección pueden obtenerse las métricas de grafo para cada componente (Estructura subgrafo), los autores pertenecientes a cada componente, y detalles del componente seleccionado y sus nodos asociados (Detalle subgrafo seleccionado y Detalle nodos del subgrafo seleccionado). En este último caso puede verse qué artículos están relacionados a dicho componente, así como distintas métricas de los nodos correspondientes.

Las métricas que pueden obtenerse para cada componente ([Fig. 3.23](#)) son las mismas que las que fueron utilizadas para la red total ([Tabla 3.2](#)).

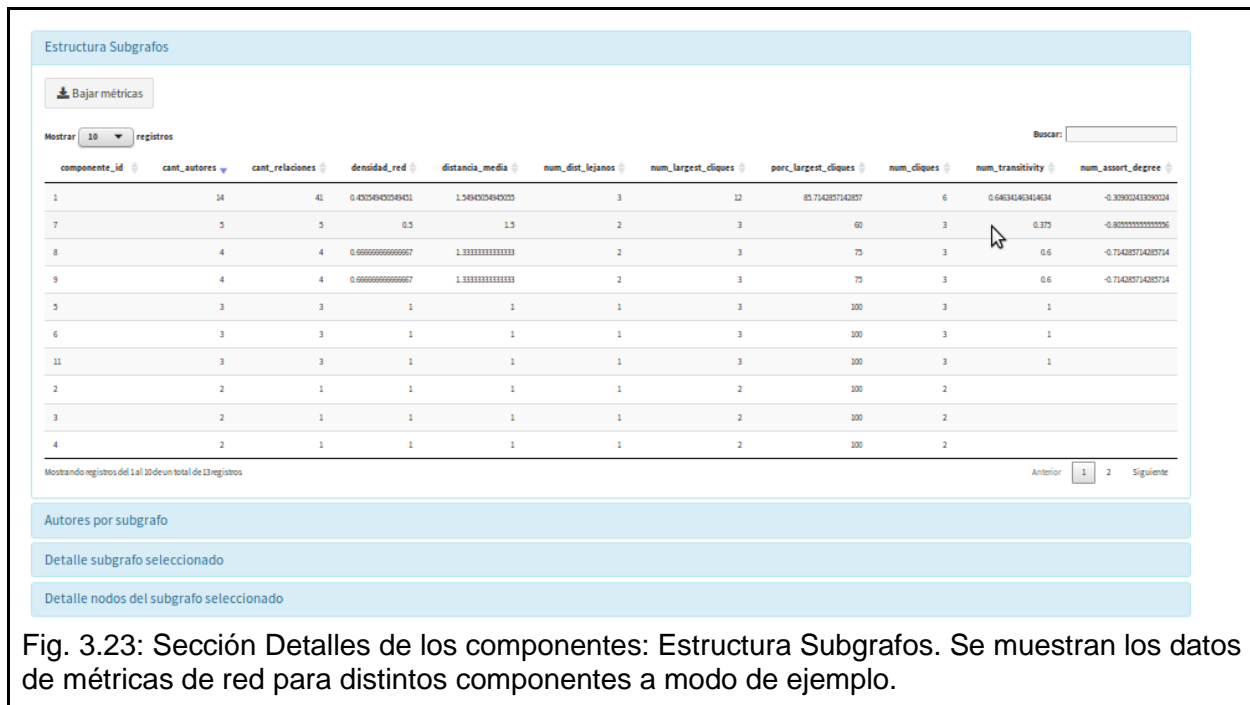


Fig. 3.23: Sección Detalles de los componentes: Estructura Subgrafos. Se muestran los datos de métricas de red para distintos componentes a modo de ejemplo.

En la subsección Autores por subgrafo se detalla la cantidad y los autores que participan de cada componente. Es posible realizar búsquedas por autores, ordenar por cantidades, y también exportar los datos (Fig. 3.24).

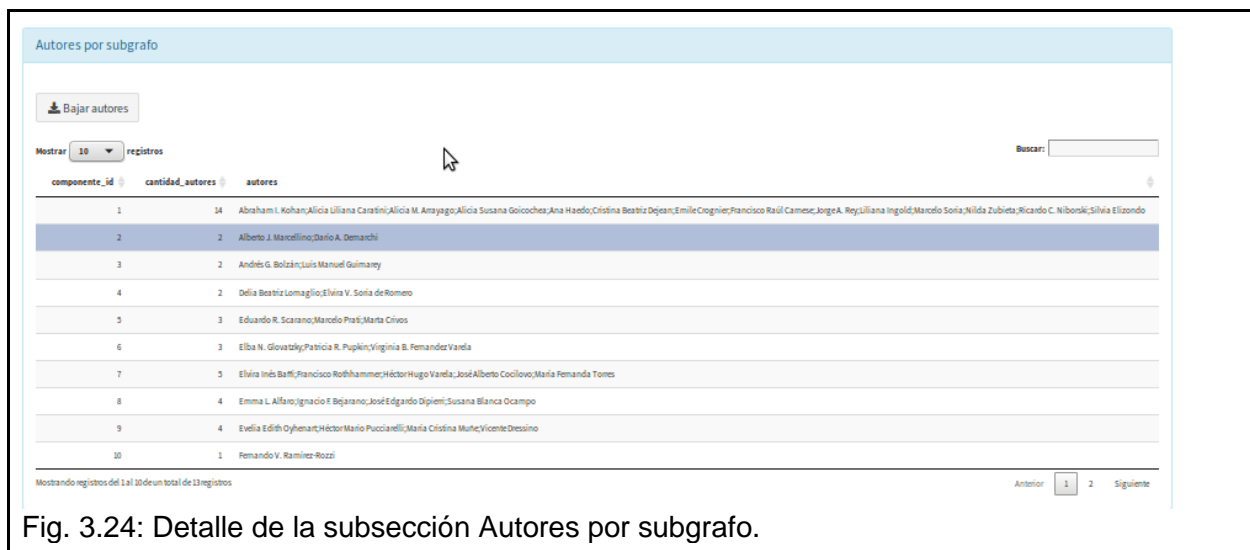


Fig. 3.24: Detalle de la subsección Autores por subgrafo.

En Detalle del subgrafo se encuentra el detalle de los artículos relacionados al componente seleccionado en la subsección Autores por subgrafo (Fig. 3.25). Es posible realizar búsquedas, que aplican a todas las columnas, y exportar los datos asociados.

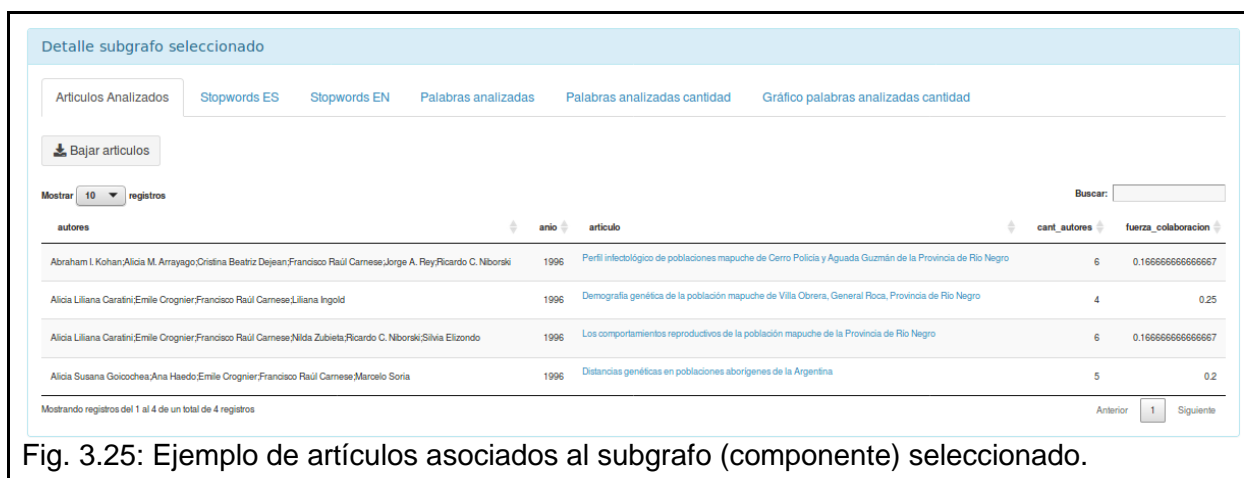


Fig. 3.25: Ejemplo de artículos asociados al subgrafo (componente) seleccionado.

También se cuenta con secciones relacionadas con análisis de texto. Puede verse, por ejemplo, la frecuencia de palabras en los títulos de los artículos asociados al componente seleccionado (Fig. 3.26).



Fig. 3.26: Ejemplo de cantidad de palabras en los artículos asociados al componente seleccionado.

Finalmente, se ofrece un detalle de los nodos del subgrafo o componente seleccionado (Fig. 3.27) con las mismas funcionalidades descritas anteriormente: métricas de nodos, distribución de grado, histograma de variables y densidad de variables.

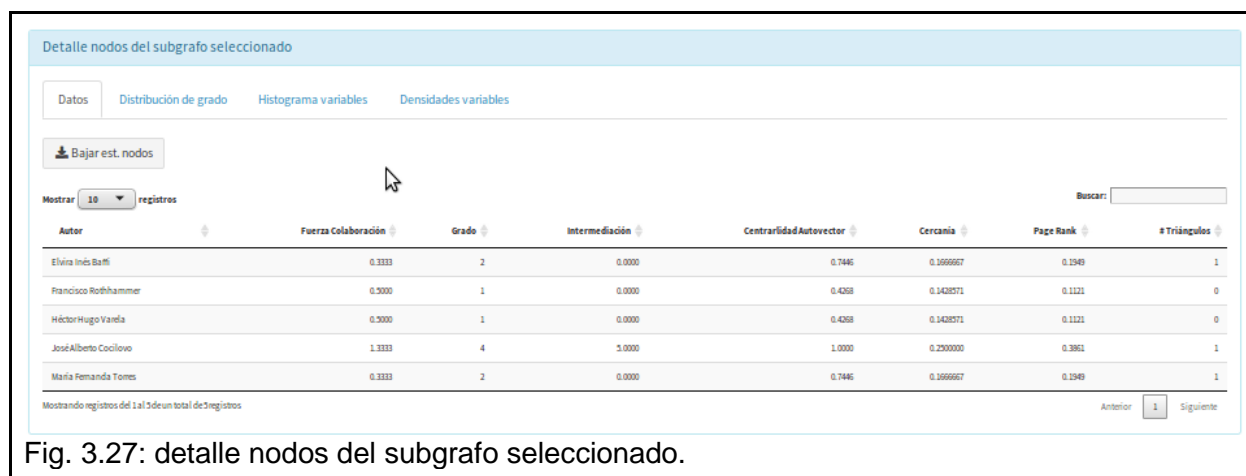


Fig. 3.27: detalle nodos del subgrafo seleccionado.

Comparación de modelos

En los análisis de redes es de utilidad realizar comparaciones con modelos. En palabras de [Kolaczyk-Csárdi \(2014\)](#):

En la práctica, los modelos de redes de grafos son usados para diferentes propósitos. Estos incluyen: 1) la prueba de “significancia” de ciertas características predefinidas en una determinada red, 2) el estudio de mecanismos propuestos para generar ciertas propiedades que se observan comúnmente en redes del mundo real (como distribuciones de grado amplias o efectos de mundo pequeño) o 3) la evaluación de potenciales factores predictivos de relaciones.

Esta subsección está compuesta por tres partes: Datos del modelo, Modelos y Verificaciones de propiedades. Para la creación de los grafos a partir de modelos se utilizó la [librería igraph](#).

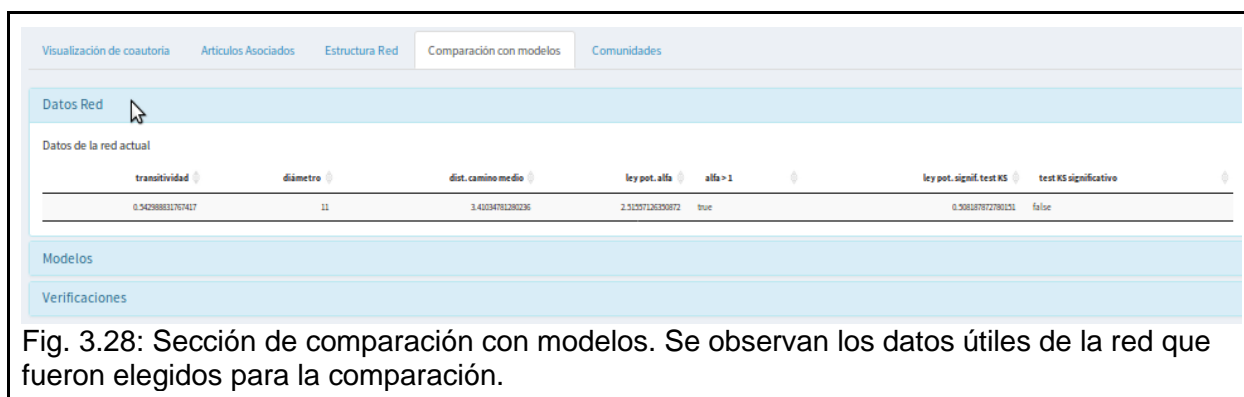
Para los estudios de Mundo Pequeño (*Small World*) y Libre escala (*Scale-Free*) se tienen en cuenta ciertas propiedades de la red, que se usan para hacer las comparaciones de la red con modelos. Las propiedades analizadas fueron las siguientes ([Tabla 3.4](#)):

Tabla 3.4. Propiedades de la red que pueden analizarse en esta subsección.

Propiedad	Razón de interés
Transitividad	De utilidad para diagnosticar mundo pequeño.
Diámetro	De utilidad para diagnosticar mundo pequeño.
Distancia camino medio	De utilidad para diagnosticar mundo pequeño.
Valor Alfa de la distribución de ley de potencia	De utilidad para diagnosticar Libre Escala. El exponente de la función para ajustar la distribución de ley de potencia
Verificación si alfa es mayor a 1	De utilidad para diagnosticar Libre Escala. Es un resultado de aplicar la función de ajuste a la ley de potencia según los parámetros.
Valor del <i>test</i> de Kolmogorov-Smirnov para ley de potencia	De utilidad para diagnosticar Libre Escala. Es un resultado de aplicar la función de ajuste a la ley de potencia según los parámetros.

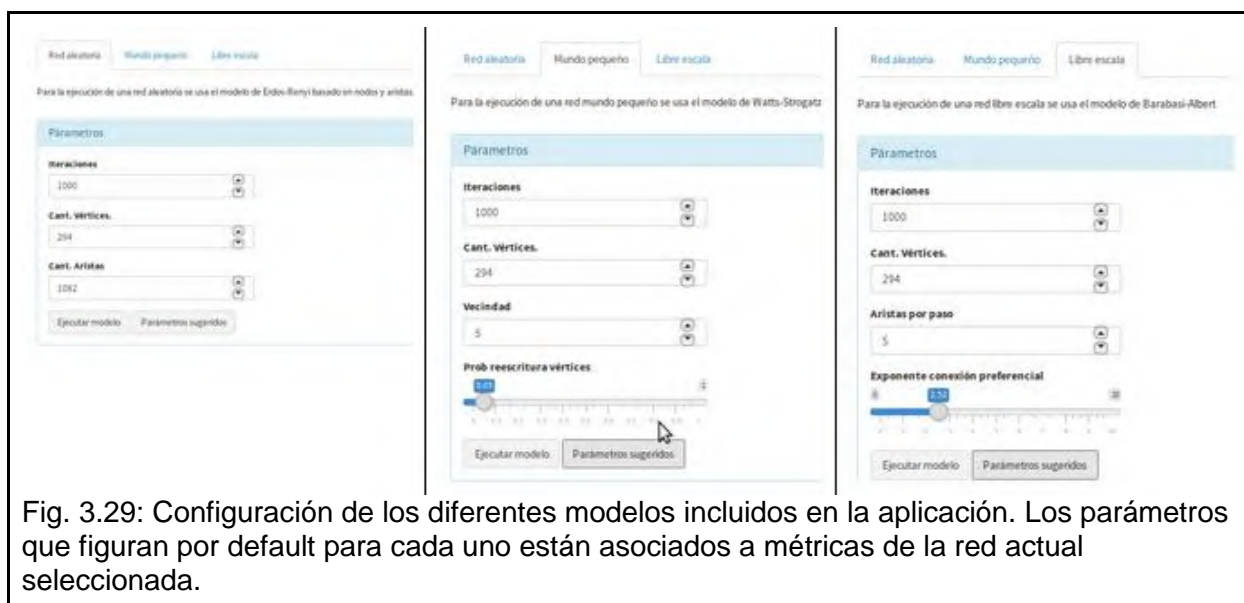
Significancia del test Kolmogorov-Smirnov	De utilidad para diagnosticar Libre Escala. Es un resultado de aplicar la función de ajuste a la ley de potencia según los parámetros.
---	--

La forma en la cual se puede visualizar esta tabla en la aplicación se muestra en la [Fig. 3.28](#)



En la solapa de modelos disponemos de tres modelos que se pueden simular para comparar con la red actual seleccionada previamente. Para cada uno de estos modelos existen distintas secciones: configuración de parámetros ([Fig. 3.29](#)), ejecución del modelo, histogramas para comparar, y los datos resultantes de la simulación.

Los modelos incluidos en la aplicación son el de Red Aleatoria según el modelo Erdos-Renyi configurado a través de nodos y aristas, Mundo pequeño de Watts-Strogatz y Libre escala de Barabasi-Albert. Para mayor detalle de estos modelos se recomienda leer el capítulo 5 de [Kolaczyk-Csárdi \(2014\)](#).



Luego de la ejecución del modelo podemos evaluarlo de forma visual a través de histogramas de los modelos en comparación con la red ([Fig. 3.30](#)). Para crear los histogramas se utilizó la librería ggplot.

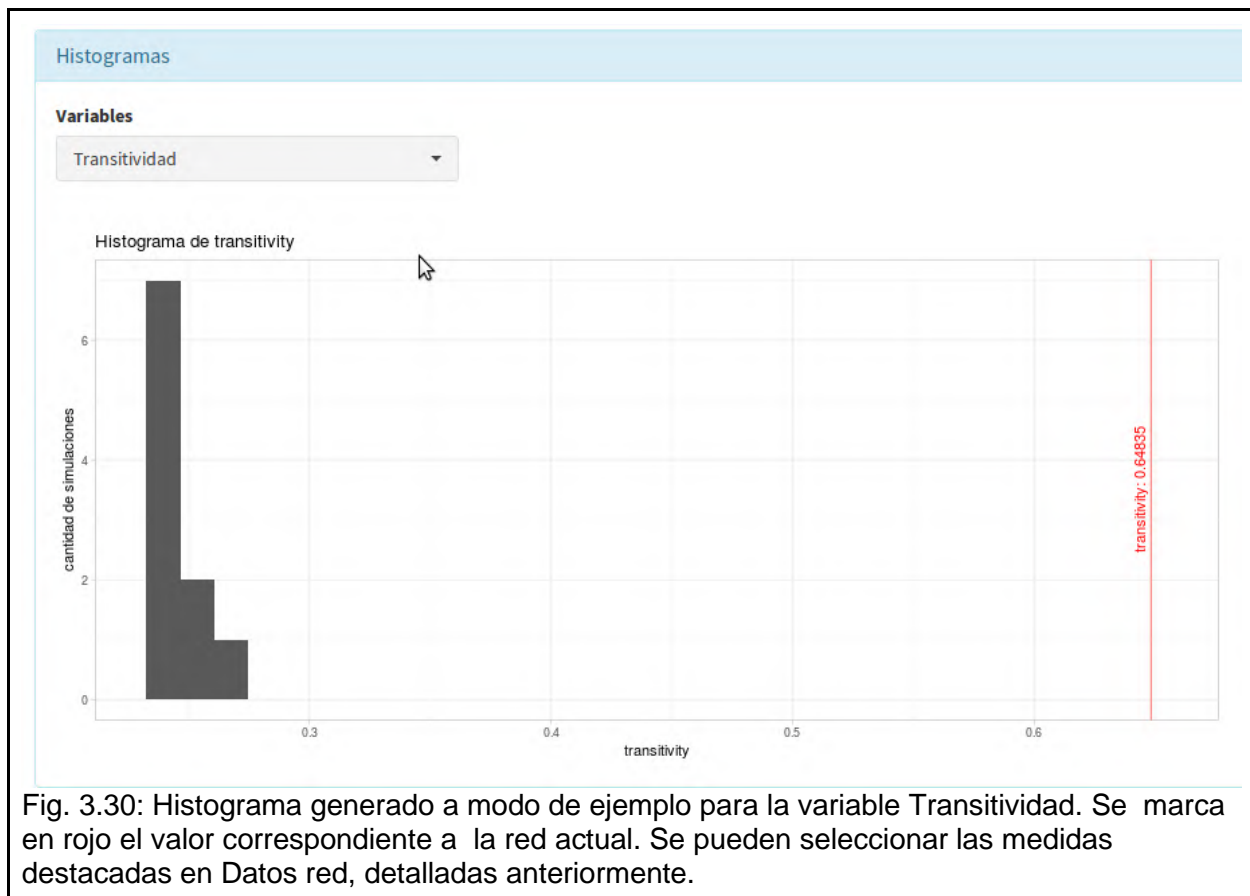


Fig. 3.30: Histograma generado a modo de ejemplo para la variable Transitividad. Se marca en rojo el valor correspondiente a la red actual. Se pueden seleccionar las medidas destacadas en Datos red, detalladas anteriormente.

Por otro lado, se pueden explorar los datos asociados a la estructura de la red de cada una de las iteraciones del modelo teniendo en cuenta las métricas presentadas anteriormente al describir las métricas de interés para la comparación con modelos (Fig. 3.31).

Datos simulación

Bajar simulación

Mostrar 10 registros

Buscar:

transitividad	diámetro	dist.camino medio	ley pot. alla	alla > 1	ley pot. signif. test KS	test KS significativo
0.030904210247997	3	3.07202477133362	12.617626742232	true	1	false
0.024743052912067	3	3.06171205683639	7.3247117613344	true	0.935770946230126	false
0.0293123886133033	6	3.05388063623225	15.7641281668902	true	1	false
0.0275299512676778	3	3.06125404223627	17.7504880226853	true	0.999996119000393	false
0.0234762672561297	3	3.06222384125487	7.68104207362733	true	0.984295216825982	false
0.0253068746021642	3	3.04622950133625	7.74273347676012	true	1	false
0.0248296603772985	6	3.06201388405101	16.89302477373102	true	1	false
0.02630910011321025	3	3.05738360590033	8.34119908278701	true	0.97235112950903	false
0.0279295082937797	6	3.06423884834178	11.2436041153762	true	1	false
0.026488427123838	3	3.05688403722257	5.9320548158025	true	0.943712120992908	false

Mostrando registros del 1 al 10 de un total de 1,000 registros

Anterior 1 2 3 4 5 ... 100 Siguiente

Fig. 3.31: Datos asociados a la simulación según un modelo en particular.

Finalmente en la parte de Verificaciones podemos contestar dos preguntas que se hacen frecuentemente en los análisis de redes: si la red es de Mundo Pequeño y si la red es de Libre Escala.

Validación de Mundo Pequeño

Para la validación de Mundo Pequeño utilizamos dos métodos de validación: el método 1 descrito en [Bialonski et al. \(2010\)](#) y el método de validación 2 realizado con la función *smallworldness* de la librería [qgraph](#) que se basa en lo planteado por [Humphries-Gurney \(2008\)](#). Visualmente, este análisis consta de tres paneles: dos de parámetros utilizados (uno por cada método de validación) y uno de la comparación de ambos métodos ([Fig. 3.32](#)).



Fig. 3.32: Paneles de criterios de evaluación de mundo pequeño.

La diferencia de valores que aparece en la comparación de métodos ([Fig. 3.32](#)) está asociada al método de generación de las redes aleatorias. En la validación 1 se usa el método de Erdos-Renyi implementado a través del método `sample_gnm` de la librería [igraph](#) que se basa en la publicación de [Erdős-Renyi \(1959\)](#); mientras que en la función de la librería [qgraph](#) se utiliza la función para generar los aleatorios de la librería [igraph](#) `degree.sequence.game` pasando el

método de generación “*simple.no.multiple*”. Para más información se puede ver la documentación de la librería [igraph](#).

Validación de Libre Escala

Según lo escrito por [Barabási \(2016\)](#):

Una red Libre Escala es una red en la cual su distribución de grado sigue la ley de potencia.

Para verificar que la distribución de grado de la red sigue la ley de potencia utilizamos la función `fit_power_law` de la librería [igraph](#), que se basa en lo planteado por [Clauset, et al. \(2009\)](#) y [Newman \(2005\)](#). Esta función recibe como parámetros los valores a ajustar a la ley de potencia, los utilizados dentro de la herramienta es una lista con el grado de todos los nodos de la red, y el tipo de implementación `plfit`.

En nuestro caso utilizamos como valores a ajustar los grados de los nodos de la red y la implementación `plfit` que utiliza máxima verosimilitud para determinar el alfa dado un parámetro *xmin*. Cuando no se brinda el parámetro *xmin*, como es nuestro caso, el algoritmo trata de buscar el valor óptimo para maximizar el p-valor de un test de Kolmogorov-Smirnov entre la distribución ajustada y la muestra original. La función utiliza el método de [Clauset, et al. \(2009\)](#) para calcular los parámetros de la distribución ajustada.

Al aplicar la función, se obtiene el valor alfa correspondiente y los resultados de la prueba de Kolmogorov-Smirnov. Acorde a [Barabási \(2016\)](#) para que la red analizada sea de Libre Escala el valor alfa debe ser mayor a 2. Luego puede verificarse la hipótesis de ajuste de la distribución de grado a la distribución de ley de potencia mediante la prueba de Kolmogorov-Smirnov.

Por lo tanto, si el p-valor de la prueba de Kolmogorov-Smirnov es mayor a 0.05 y asimismo el alfa es mayor a 2, la red es considerada de Libre escala. En la [Fig. 3.33](#) se observa la implementación de esta prueba en la aplicación.

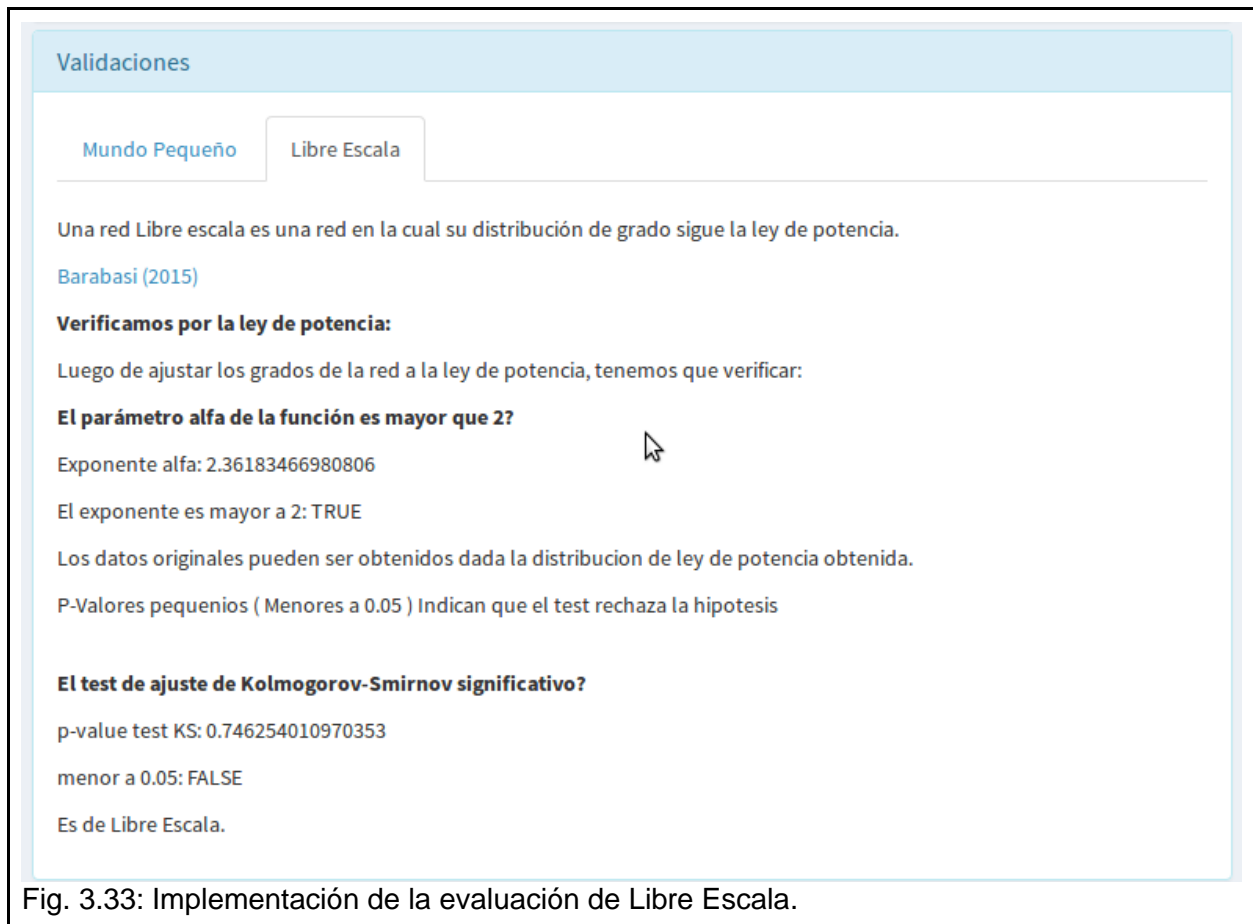


Fig. 3.33: Implementación de la evaluación de Libre Escala.

Comunidades

Explicado en palabras de [Savic et al. \(2015\)](#):

La estructura de comunidad es una propiedad típica de las redes sociales. Una comunidad (agrupamiento o módulo) es una parte de una red (grupo de nodos) donde las conexiones internas son más densas que las externas. Descubrir comunidades nos ayuda a entender la estructura de la red, a identificar subgrupos cohesivos y a dibujar un mapa entendible de la red.

La necesidad de búsqueda de comunidades está asociada a buscar patrones según ciertas propiedades de la red. Esta búsqueda de patrones se puede realizar de diferentes maneras, para lo cual existen distintos algoritmos.

Por lo tanto, para poder analizar comunidades de una red específica, primero se debe seleccionar el algoritmo con el cual se van a buscar dichas comunidades. Esta opción está disponible en el panel de configuración de la sección de Comunidades, así como la posibilidad de seleccionar cuántas comunidades se desean ver en la visualización de la red ([Fig. 3.34](#)). Para más detalles de los algoritmos disponibles ver el Anexo: Detalle de algoritmos para búsqueda de comunidades.

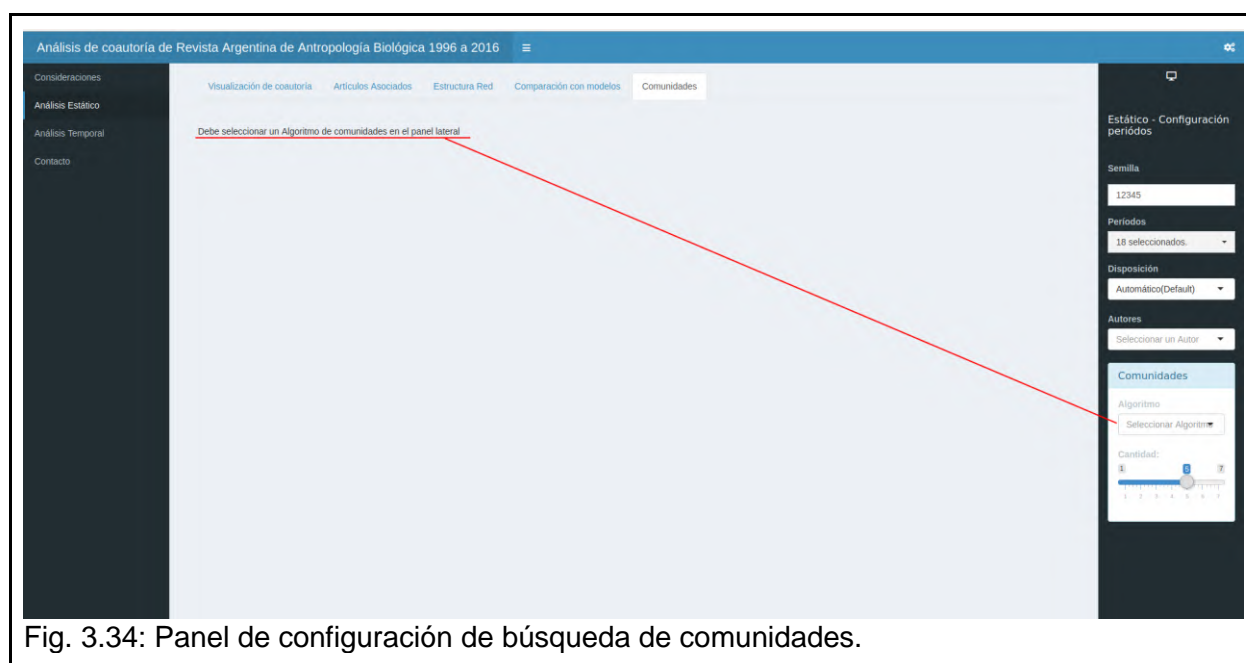


Fig. 3.34: Panel de configuración de búsqueda de comunidades.

Luego de seleccionar el algoritmo, en la sección de Visualización de coautoría se puede ver cómo están distribuidas la totalidad de las comunidades obtenidas ([Fig. 3.35](#)) .



Fig. 3.35: Ejemplo de selección de 5 comunidades en el Panel de configuración. Se observan las 5 comunidades destacadas en la red con diferentes colores.

Para complementar la visualización, se puede proceder con un análisis de la composición de los subgrafos resultantes de la búsqueda de comunidades, con ese objetivo se dispuso de la sección Comunidades ([Fig. 3.36](#)). Como detalle se describe qué algoritmo fue utilizado para la búsqueda, la cantidad total de comunidades y, para evaluar la relevancia de los resultados de los algoritmos, está disponible la medida de modularidad ([Newman-Girvan 2004](#)). Una vez seleccionado el algoritmo se puede ver en el panel de comunidades el valor de esta medida teniendo en cuenta y sin tener en cuenta el peso de las relaciones ([Fig. 3.36](#)). Cabe recordar que los pesos de las relaciones en este caso están dados por la fuerza de colaboración de cada relación.

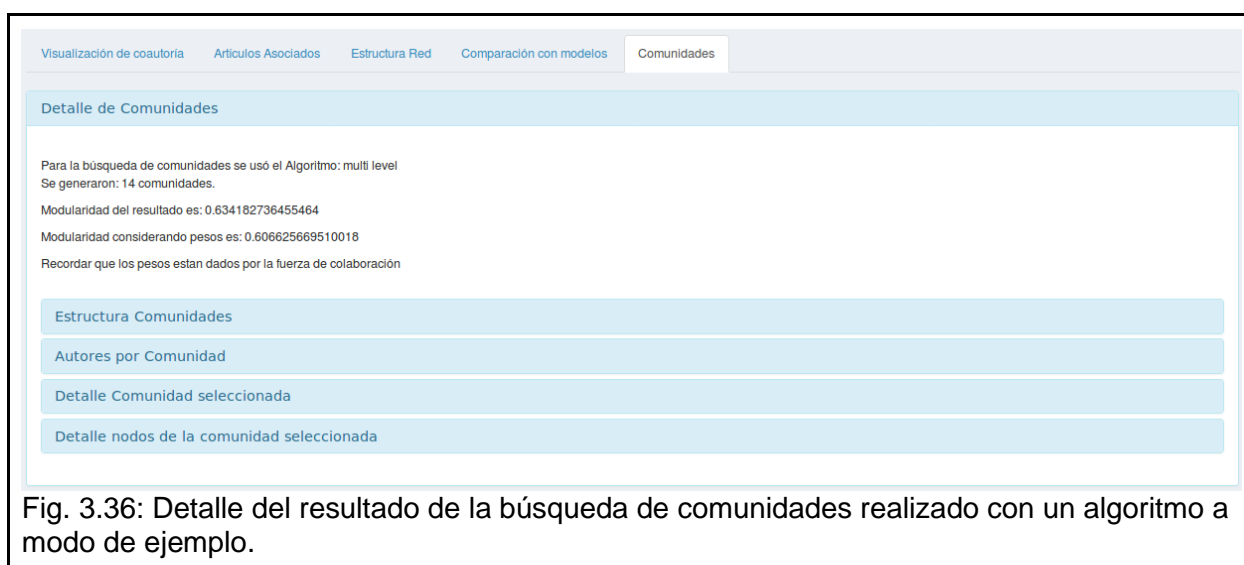


Fig. 3.36: Detalle del resultado de la búsqueda de comunidades realizado con un algoritmo a modo de ejemplo.

Además es posible realizar un análisis de la estructura de cada comunidad ([Fig. 3.37](#)), en el que se puede ver qué autores forman parte de ella y los artículos asociados. Esto es similar a lo disponible para los componentes de la red, solo que esta vez el subgrafo es cada comunidad en vez de cada componente.

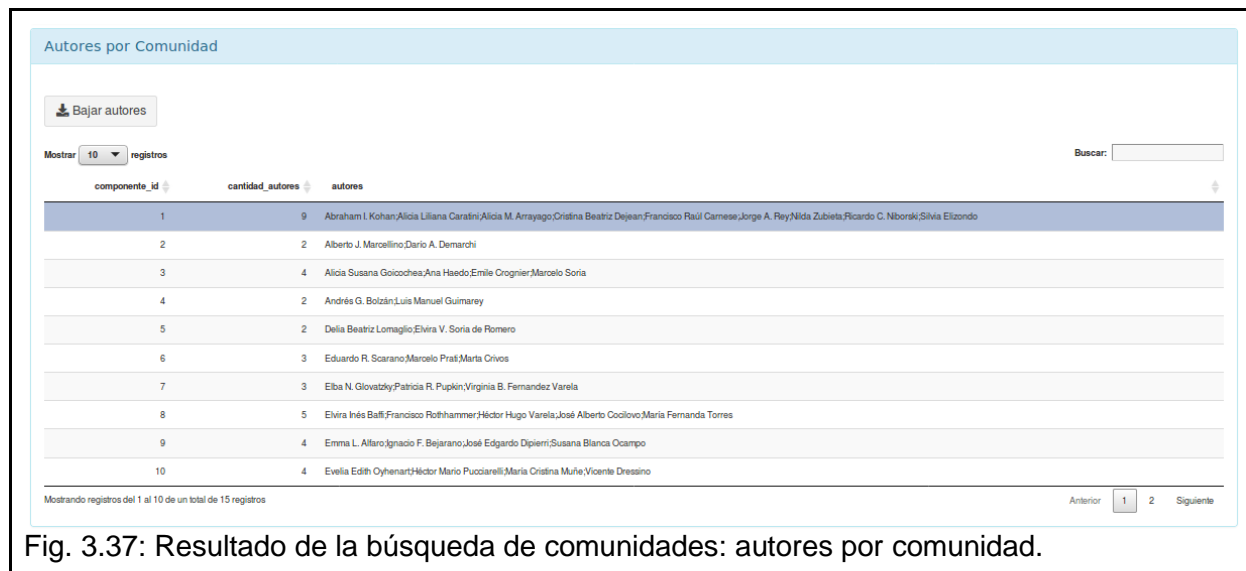


Fig. 3.37: Resultado de la búsqueda de comunidades: autores por comunidad.

Asimismo, es posible obtener las mismas métricas que las descritas para el análisis de componentes, solo que en este caso, los subgrafos son las comunidades en vez de los componentes ([Fig. 3.38](#)).



Fig. 3.38: Métricas de los subgrafos de comunidades

Luego de seleccionar una comunidad, en la tabla de autores por comunidad se dispone de los detalles de la misma, entre los que se encuentran los artículos asociados ([Fig. 3.39](#)).



Fig. 3.39: Resultado de la búsqueda de comunidades. Detalle de los artículos de la comunidad seleccionada.

De dichos artículos se pueden ver algunos resultados del análisis de texto realizado sobre los títulos de los artículos, tales como: las palabras más frecuentes de esa comunidad, la cantidad de palabras frecuentes y los filtros aplicados de *stop words* en español e inglés (dado que existen algunos títulos en dicho idioma) ([Fig. 3.40](#)).

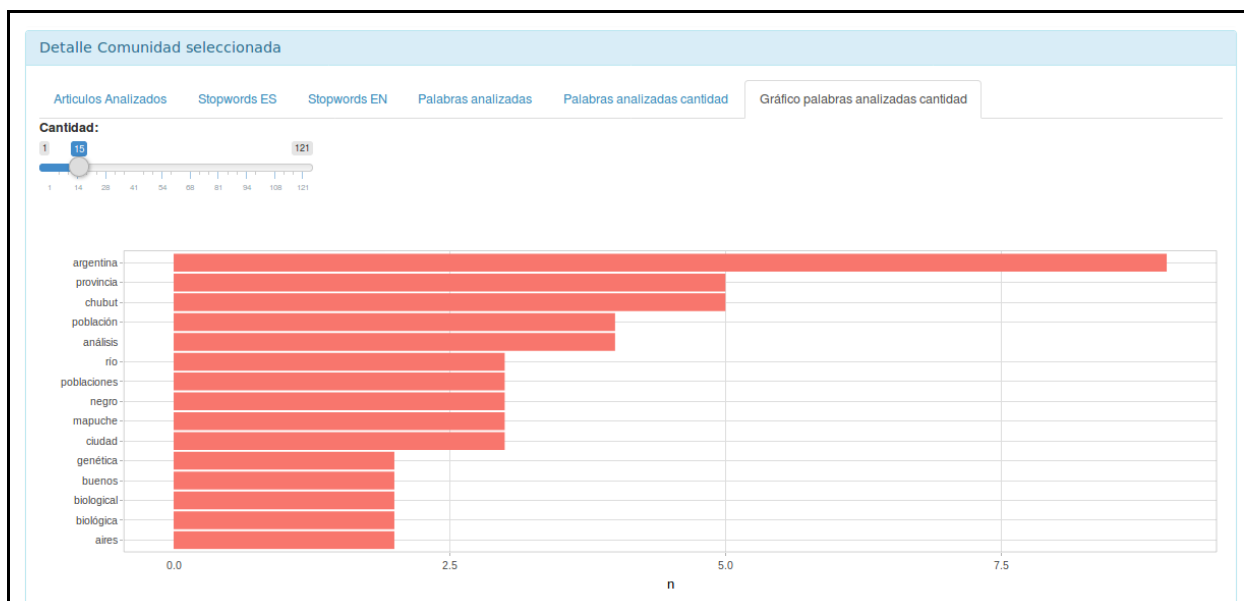


Fig. 3.40: Resultado de la búsqueda de comunidades. Detalle de palabras más frecuentes en los títulos de los artículos.

De manera similar a lo realizado para la sección de componentes, también se dispone del detalle de los nodos de la comunidad seleccionada ([Fig. 3.41](#)).

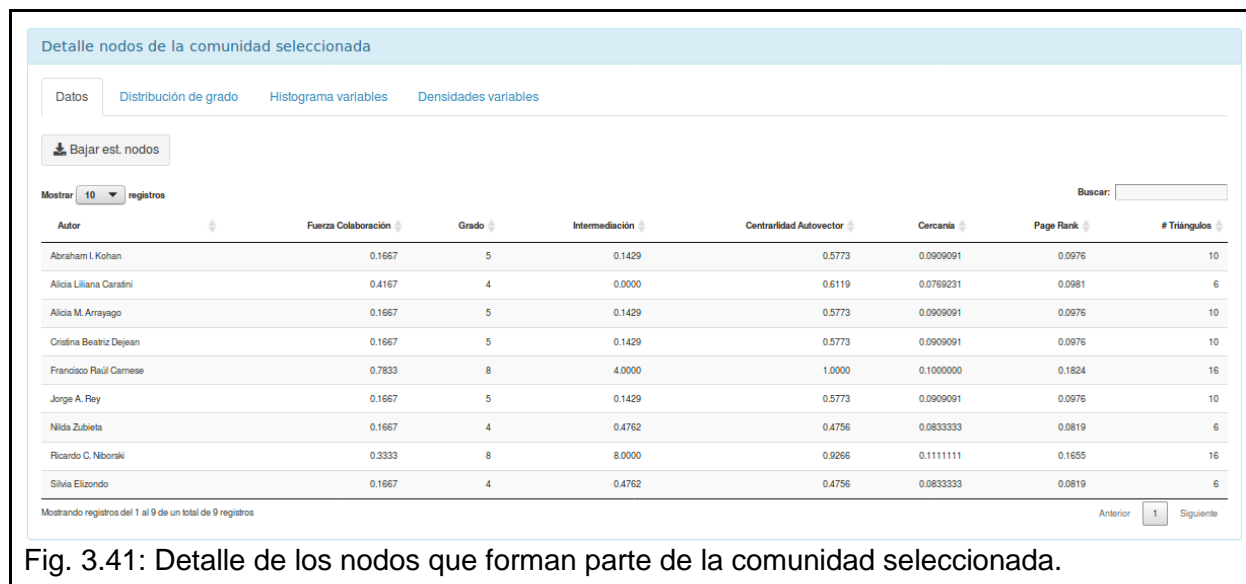


Fig. 3.41: Detalle de los nodos que forman parte de la comunidad seleccionada.

Sección 3: Análisis temporal

En esta sección se realizó un análisis temporal de la red desde su primera publicación en 1996 hasta el periodo 2016 puesto como cota por 20 años desde su primera edición.

Como vistazo inicial del **Análisis temporal** se planteó una visualización dinámica. Para su generación se utilizó el grafo acumulado desde el periodo 1996 como $T=0$ hasta 2016 como $T=18$. De esta manera, a medida que se avanza en los tiempos aparecen nuevos nodos y se modifican las relaciones.

Luego de tener una idea general de cómo se generaron las relaciones en el tiempo se pueden obtener diferentes medidas que sirven para analizar diversos aspectos de la red como por ejemplo qué nodos cumplen distintos roles. Dependiendo del objetivo, puede ponerse el foco en una o varias de estas medidas para contestar las preguntas de interés. La herramienta presenta dos enfoques de análisis de las medidas: de forma individual por periodo o de forma acumulada total. Esto nos permite analizar de forma general las relaciones de autoría dentro de la RAAB, de manera similar al trabajo realizado por [Koseoglu \(2016\)](#).

Finalmente la exploración de diferentes agregaciones temporales puede realizarse en la sección Análisis estático de periodos, mediante la selección de los periodos específicos de interés.

Animación del grafo de manera acumulada

Esta herramienta muestra cómo se generaron las relaciones entre los nodos y cómo fueron conectándose los diversos componentes a lo largo del tiempo. Esto ayuda a focalizarnos en el periodo de interés para posteriormente realizar un análisis más detallado en la sección **Análisis estático**.

En la animación se pueden ver los nodos en escala de azules según su fuerza de colaboración y los colores de las aristas según la fuerza de relación entre los autores (medida en fuerza de colaboración) (Fig. 3.42). A la derecha de la animación hay una tabla para ver las equivalencias entre periodos y años.

En el detalle de los nodos de la animación podemos ver quién es el autor, cuánta es la fuerza de colaboración acumulada, los periodos de participación y cantidad de artículos por periodo.

En el detalle de las aristas podemos ver los autores participantes de dicha relación, la fuerza de colaboración resultante de las coautorías de los autores, los periodos en los cuales participan, y cuántos artículos contribuyeron en esos periodos.

Para poder realizar esta animación se usaron los paquetes igraph, ndtv, intergraph, networkDynamic y network. Para más detalles ver el [anexo de librerías](#).

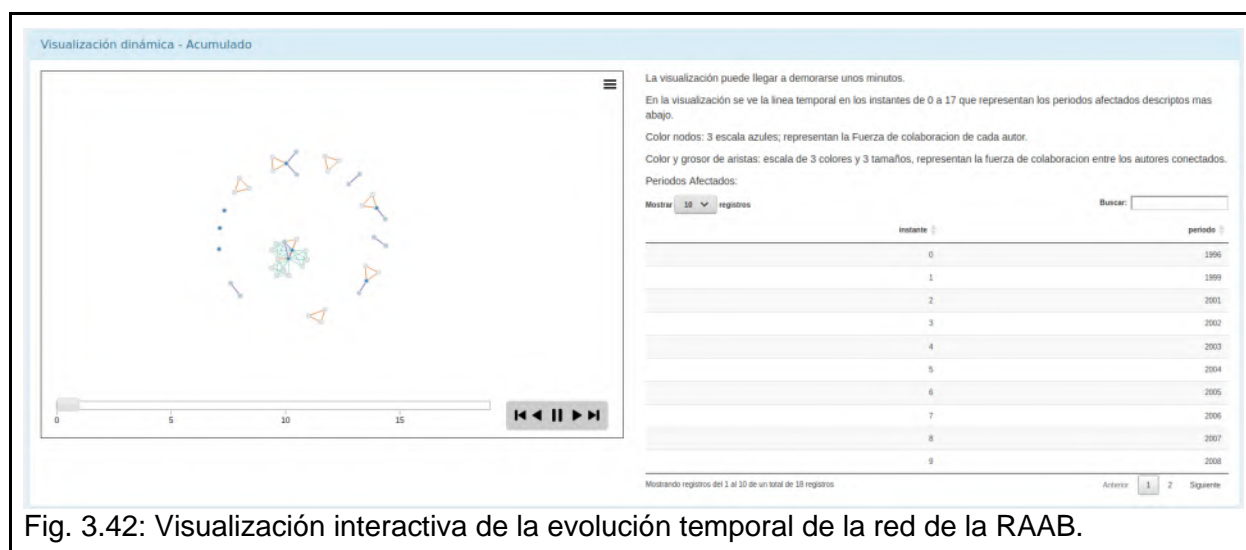


Fig. 3.42: Visualización interactiva de la evolución temporal de la red de la RAAB.

Medidas de cada periodo

Esta sección fue diseñada para ver las medidas en su evolución por año. Las medidas disponibles son las mismas detalladas para la sección Métricas de la red en Estructura de la red. (Fig. 3.43)

Los gráficos fueron realizados con la [librería ggplot](#) y se hicieron interactivos utilizando la [librería plotly](#).

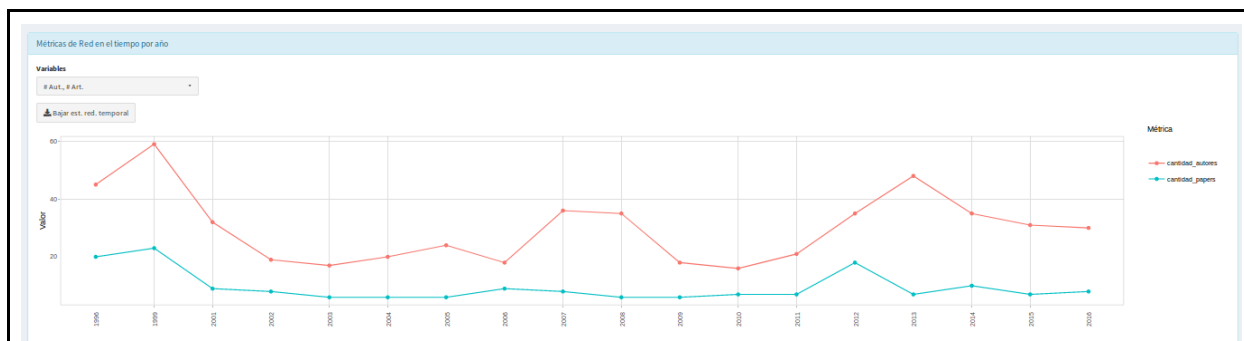


Fig. 3.43: Métricas de la red para cada instante anual. Se muestran como ejemplo la cantidad de autores y la cantidad de artículos.

Medidas Acumuladas a cada periodo

Dado que en cada nueva edición de la revista se fueron incorporando más trabajos, resultó de interés ver la interacción entre los viejos nodos y los nuevos nodos, cuándo aparecen los nuevos y qué tanto aporte realizan en el total de la revista. Las medidas que se incorporaron en la visualización acumulada en el tiempo son las mismas que las detalladas previamente para Estructura de la red > Métricas de la red (Fig. 3.44).

Los gráficos fueron realizados con la [librería ggplot](#) y se hicieron interactivos utilizando la [librería plotly](#).

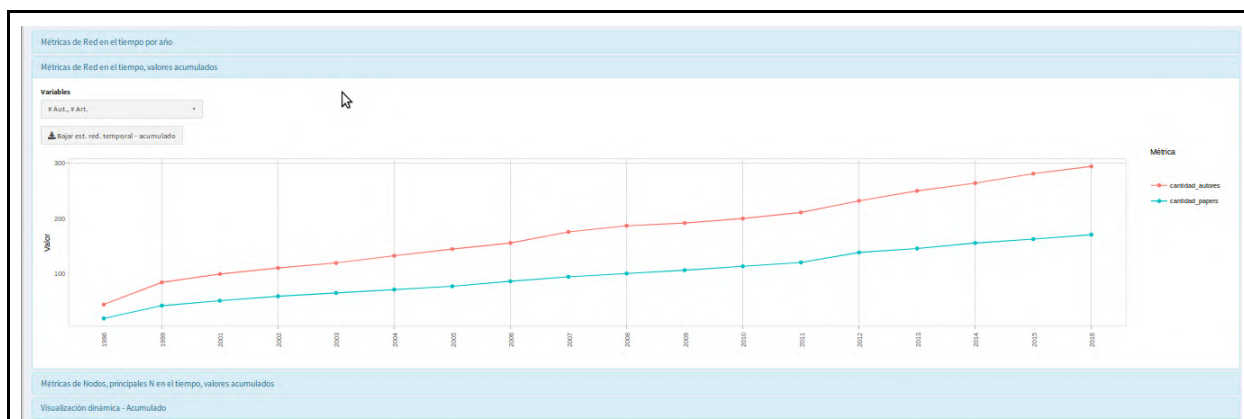


Fig. 3.44: Métricas acumuladas para cada instante anual. Se muestran como ejemplo la cantidad de autores y la cantidad de artículos.

Ranking de autores en el tiempo

En esta sección es posible comparar a los autores a lo largo del tiempo según alguna métrica elegida y una cantidad de autores por año especificada (N) ([Fig. 3.45](#)).

Se obtiene entonces un gráfico con las posiciones en el ranking de los N autores elegidos. Las líneas sirven para ver entre cuáles autores fueron variando las posiciones a lo largo del tiempo. Si se desea ver el valor de la métrica elegida, se puede visualizar en el detalle interactivo. Los empates entre autores se visualizan como múltiples puntos del mismo color en el mismo periodo, unidos por una línea vertical.

Las métricas disponibles son las mismas que las empleadas en la sección Estructura de nodos, especificadas anteriormente. Los gráficos fueron realizados con la [librería ggplot](#) y se hicieron interactivos utilizando la [librería plotly](#).

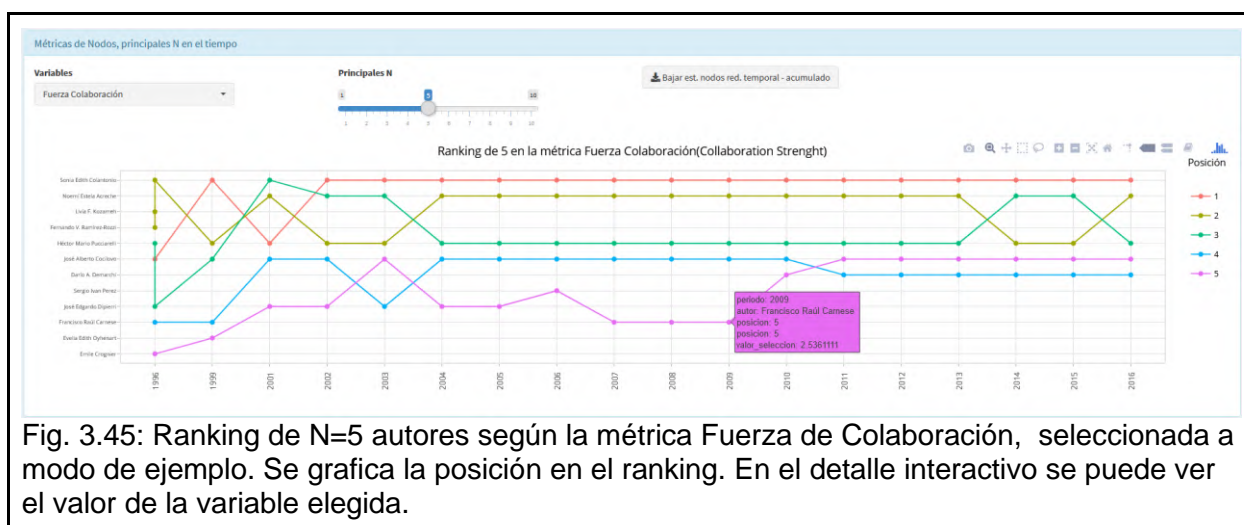


Fig. 3.45: Ranking de N=5 autores según la métrica Fuerza de Colaboración, seleccionada a modo de ejemplo. Se grafica la posición en el ranking. En el detalle interactivo se puede ver el valor de la variable elegida.

Ranking de autores en el tiempo para medidas acumuladas

Esta sección es similar a la anterior, solo que se visualiza el color de la métrica acumulada en el tiempo (Fig. 3.46). Al igual que en el caso anterior, se elige una métrica de interés y la cantidad de autores a visualizar (N). En caso de empate aparecen todos los autores que estén entre los primeros N.

Las métricas disponibles son las mismas que las empleadas en la sección Estructura de nodos, especificadas anteriormente. Los gráficos fueron realizados con la [librería ggplot](#) y se hicieron interactivos utilizando la [librería plotly](#).



Fig. 3.46: Ranking de N=5 autores según la métrica Fuerza de colaboración, seleccionada a modo de ejemplo. En este caso se destaca en gradiente de color el valor acumulado de la métrica.

Resumen del capítulo

En este capítulo se presentó la aplicación desarrollada y se describieron, para cada una de las secciones que la componen, su fundamentación, la metodología aplicada y las herramientas utilizadas en su desarrollo, su implementación con ejemplos y visualizaciones, así como también sus limitaciones.

En el próximo capítulo se usará esta herramienta para analizar el caso específico de la RAAB.

Capítulo 4: Análisis de la Red

Introducción

El periodo temporal analizado fue entre los años 1996 y 2016. Se optó por esta cota temporal como límite factible para el análisis en el marco de la presente tesis, y puesto que conforma un periodo de 20 años desde la primera publicación de la revista.

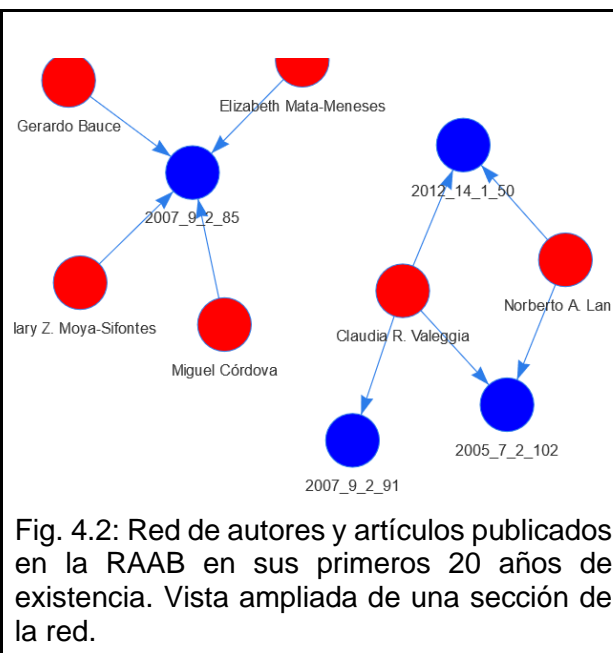
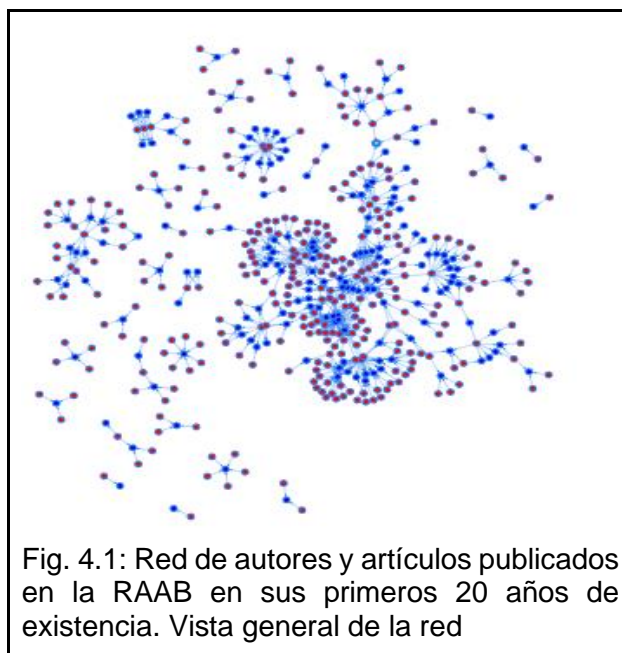
Se utilizaron sólo los Trabajos originales para el análisis de la red de coautoría de la revista, siguiendo a [Ordoñez et al. \(2009\)](#) quienes mencionan que uno de los criterios de calidad que hacen a una revista especializada es el componente investigativo y original de sus artículos.

En el presente capítulo analizamos la red en su totalidad (es decir, de 1996 a 2016) y luego por periodos acotados, a fin de evaluar posibles cambios en los patrones generales a lo largo del tiempo.

En el análisis contestamos las hipótesis de investigación planteadas en esta tesis así como también otras preguntas que fueron surgiendo al familiarizarnos con los datos.

Antes de comenzar con el análisis, cabe recordar algunas características de la red bajo estudio. En primer lugar, un autor puede ser autor de múltiples obras y una obra puede ser elaborada por múltiples autores. Esto genera una red de autoría compuesta por dos tipos diferentes de nodos, lo que se denomina un grafo bipartito. En el caso de la RAAB, analizada en este caso, los autores son los científicos y las obras analizadas son los trabajos originales ([Fig. 4.1](#)).

Además, observamos que la red es dirigida, dado que los autores generan las obras ([Fig. 4.2](#)).



Finalmente, para transformar esta red bipartita en una de coautoría utilizamos una proyección a la dimensión de los autores, quedándonos así con los nodos autores y generando aristas de “coautoría” o “trabajo conjunto”, a las cuales se les asignó un peso según la cantidad de trabajos involucrados. En este caso, observamos que la red pasó a ser no dirigida ([Fig. 4.3](#)), dado que la

obra generada es la relación y, en caso de existir múltiples relaciones, se simplificó en una arista con mayor peso. Es decir que a mayor cantidad de trabajos en conjunto de un dado par de autores, más peso tuvo la relación o arista que los unió. Para este análisis, no se tuvo en cuenta el orden de autores en cada trabajo realizado; es decir, todos se encontraron en la misma jerarquía.

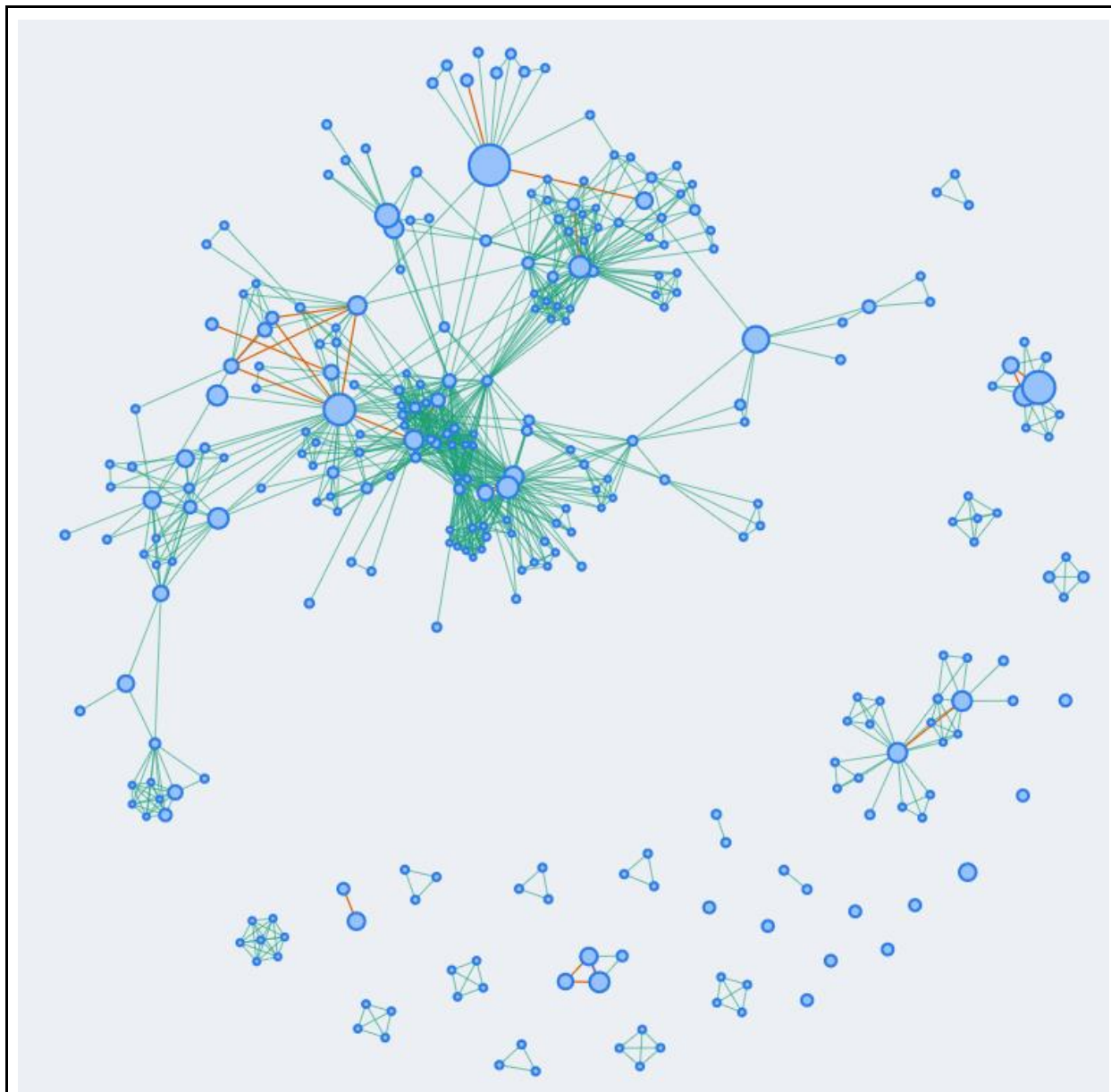


Fig. 4.3: Red proyectada para el periodo completo (1996-2016). Disposición: Fruchterman Reingold. El tamaño de los nodos (autores) es proporcional a su fuerza de colaboración y el color de las aristas varía según la fortaleza de la relación, también medida por fuerza de colaboración, de verde (menor valor) a naranja (mayor valor).

Algunas de las preguntas que surgieron al analizar este grafo ([Fig. 4.3](#)) fueron: ¿Los autores individuales aportan igualmente a la red que aquellos que participan en equipos de dos personas? ¿Cuál es el aporte de los equipos de más de dos personas?

Para abordar estas cuestiones se tuvieron en cuenta diferentes métricas. Cabe destacar una en particular para este tipo de redes: la **fuerza de colaboración**. Esta métrica se define como un artículo repartido en la cantidad de autores ($1 / N$ autores) ([Newman 2004](#)).

En la vista de la red de la [Fig. 4.3](#), se puede ver el tamaño de los autores según su fuerza de colaboración y, de la misma forma, el color de las aristas varía según la fortaleza de la relación medida por fuerza de colaboración.

Sin embargo, con esta métrica se destacan principalmente los autores que han publicado de manera individual por sobre los equipos, a veces interdisciplinarios, que pueden realizar otro tipo de aportes a la disciplina. Por lo tanto, en el caso de los nodos se utilizaron otras métricas de redes como el **Grado** de cada autor, que mide con cuántos otros autores se conecta directamente el autor, y la **Intermediación**, que mide la importancia de un autor según su participación en caminos para llegar de un autor a otro, generando equipos.

Como los autores pueden trabajar en forma conjunta o individual, o en equipos que no se conectan entre ellos, se espera encontrar una red inconexa, esto significa que no todos los nodos están conectados entre sí.

Cabe recordar finalmente el aspecto temporal del presente análisis puesto que puede decirse que la red pasó por varias etapas para llegar a su estadio final, en este caso el año 2016. Las distintas redes de coautoría de la RAAB pueden ser analizadas entonces por etapa/año o bien de forma agregada, es decir, teniendo en cuenta lo acumulado en etapas anteriores hasta la etapa seleccionada.

Teniendo presentes todos estos aspectos detallados previamente, a continuación se presentan los análisis realizados para poner a prueba las cuatro hipótesis de investigación planteadas en la presente tesis.

Hipótesis 1: La cantidad de autores aumenta en el tiempo debido al crecimiento de la disciplina en el país, lo que se refleja en un aumento en la cantidad de nodos en la red

La primera hipótesis que planteamos trata de verificar si existe un crecimiento de la disciplina Antropología Biológica a nivel local. Para constatar esta hipótesis elegimos analizar las publicaciones de la Revista de Antropología Biológica (RAAB), que es el principal elemento de difusión de la Asociación Argentina de Antropología Biológica (AAAB). Revisar los artículos publicados en la RAAB nos da una idea si se fomenta el crecimiento local, una red interna de Argentina, en contraste de solo la red externa internacional. Estimamos que en ambas redes hay desarrollo, este trabajo se focaliza en la red interna argentina tomando la RAAB como representante de la materia a nivel local.

Tomamos como métrica la cantidad de autores en el tiempo, sin embargo esta métrica se puede ver de diferentes formas, acumulada en cada instante temporal o individualizando cada instante. Estas aproximaciones pueden otorgar diferentes puntos de vista sobre el contexto de estudio.

Si analizamos la cantidad en cada instante temporal tendremos mayor cantidad de autores según potenciales eventos en la existencia de la revista, como puede ser su inicio o congresos. Se estima que estos eventos afectan la producción de trabajos originales, haciendo crecer o reduciendo la cantidad de trabajos originales, y por ende, la cantidad de autores participantes en la red de coautoría de trabajos originales de la RAAB. Este punto de vista nos ayuda a comprender mejor cómo se comportan las dinámicas asociadas a los eventos de mayor o menor producción. Tal vez para plantear nuevas preguntas como cuánto tiempo tardan en aparecer nuevos trabajos siendo disparados por interacción en congresos.

Si tomamos las cantidades acumuladas de autores, estimamos que crecen en el tiempo en caso de crecer la revista, o mantenerse sin incrementos en caso de no crecer. Crecer implica que nuevos autores aparecen en la red, que estos pueden conformar nuevos equipos, o tal vez se integran a equipos existentes. Al revisar la cantidad acumulada van a ser de nuestro interés los saltos que hay entre edición y edición viendo la variación que hay entre cada momento. Mayor variación entre periodos significa que nuevos autores participan de la red, menor variación significa que la red se mantiene por los autores existentes.

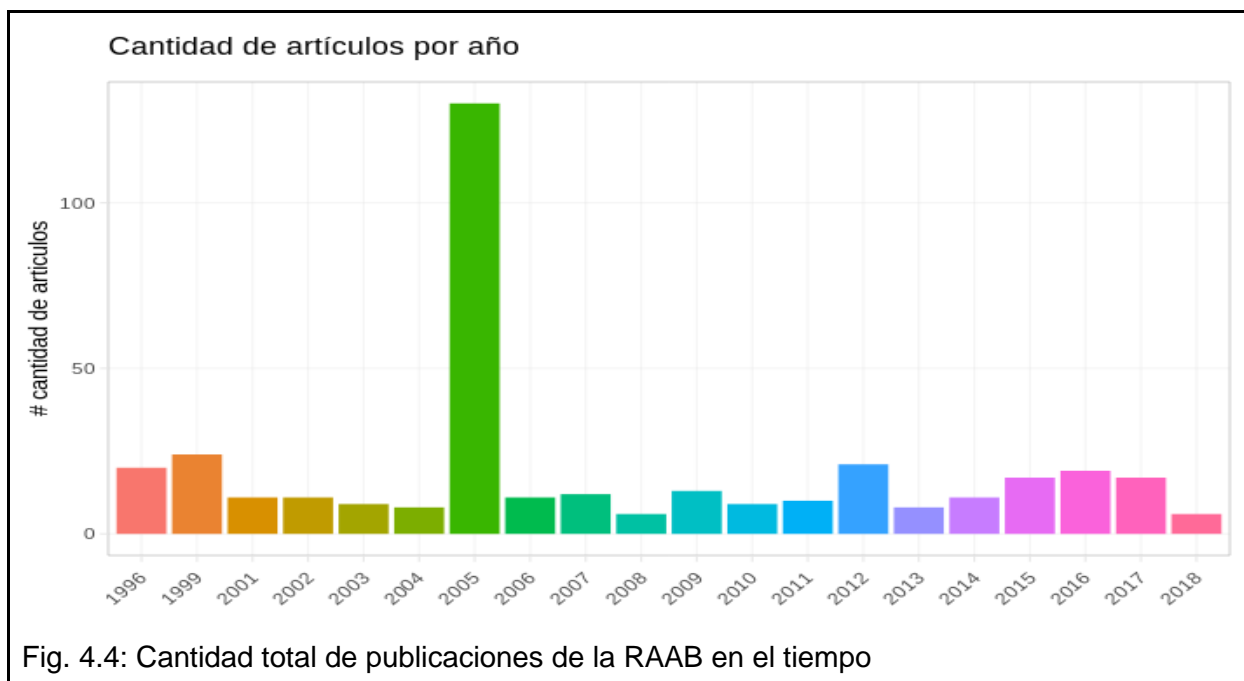
Para tener un mejor contexto, se evaluó la cantidad de autores y publicaciones por año desde la primera edición de la RAAB ([Tabla 4,1](#)).

Recordemos que estas medidas corresponden generalmente a dos ediciones de la revista por año, lo cual en cierto sentido es una restricción a la cantidad de artículos publicables por año y quizás también a la de autores. Además, en esta primera aproximación no solo se tuvieron en cuenta trabajos originales, sino también otro tipo de publicaciones ([Tabla 10](#)).

año		1996	1999	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018*
total		20	24	11	11	9	8	130	11	12	6	13	9	10	21	8	11	17	19	17	6
sección	Artículos de Revisión	0	1	1	0	0	0	0	1	0	0	0	0	2	1	0	0	0	1	1	0
	Cartas al Editor	0	0	0	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
	Comentarios Bibliográficos	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	1	2	1	0
	Comunicaciones Breves	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	Dossier	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	7	7	0
	Nota editorial	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0
	Notas Técnicas	0	0	0	0	0	0	0	0	0	0	6	0	1	0	0	0	0	0	0	0
	Obituario	0	0	0	1	0	1	1	1	0	0	1	0	0	1	0	0	1	0	0	0
	Resúmenes de Jornadas	0	0	1	0	2	0	123	0	2	0	0	0	0	0	0	0	0	0	0	0
sección	Trabajos Originales	20	23	9	8	6	6	6	9	8	6	6	7	7	18	7	10	7	8	7	6
edición	1	20	24	10	11	2	8	123	11	2	1	13	9	10	12	8	6	7	8	10	6
	2	0	0	1	0	7	0	7	0	10	5	0	0	0	0	0	5	10	11	7	0
	especial	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0

Tabla 4.1: Cantidad de publicaciones por año, discriminado en secciones y ediciones.* Al momento de la confección de la tabla, el año 2018 no estaba completo.

Al observar la cantidad total de publicaciones en la RAAB por año ([Fig. 4.4](#)) y discriminada en ediciones (primera, segunda y especial) ([Fig. 4.5](#)) pudo verse una gran disparidad para el año 2005.



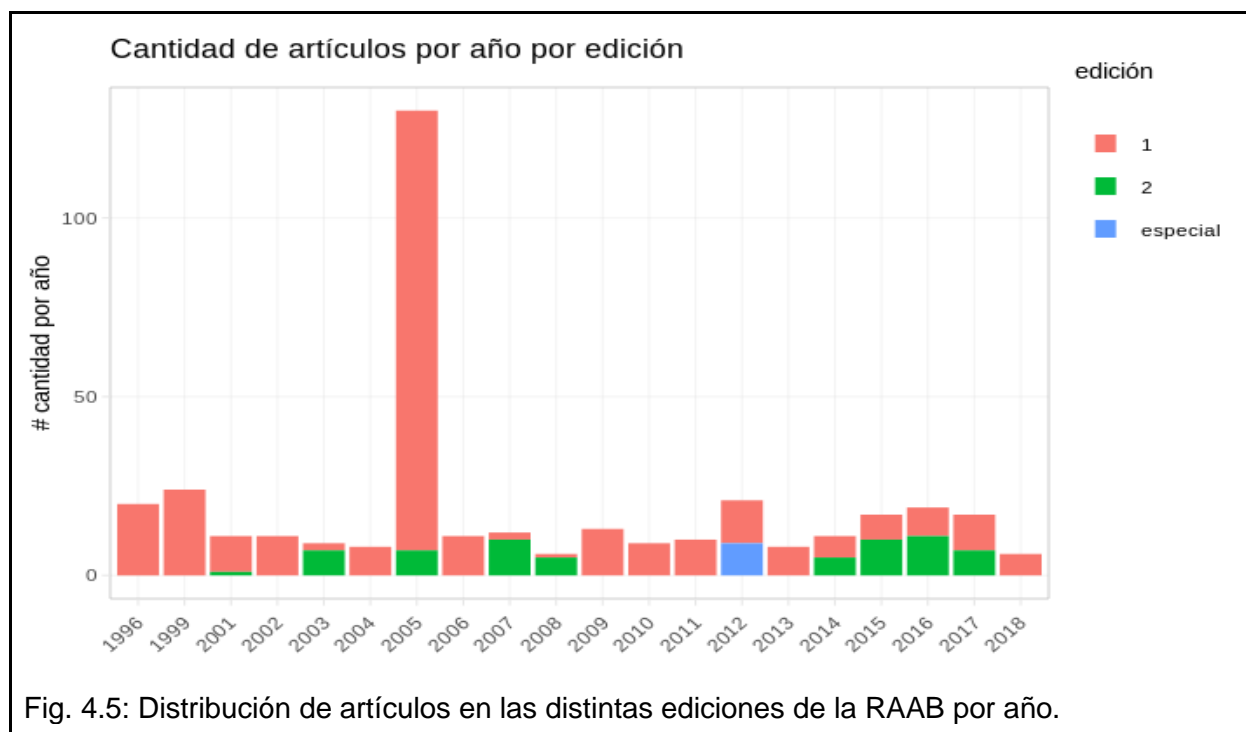
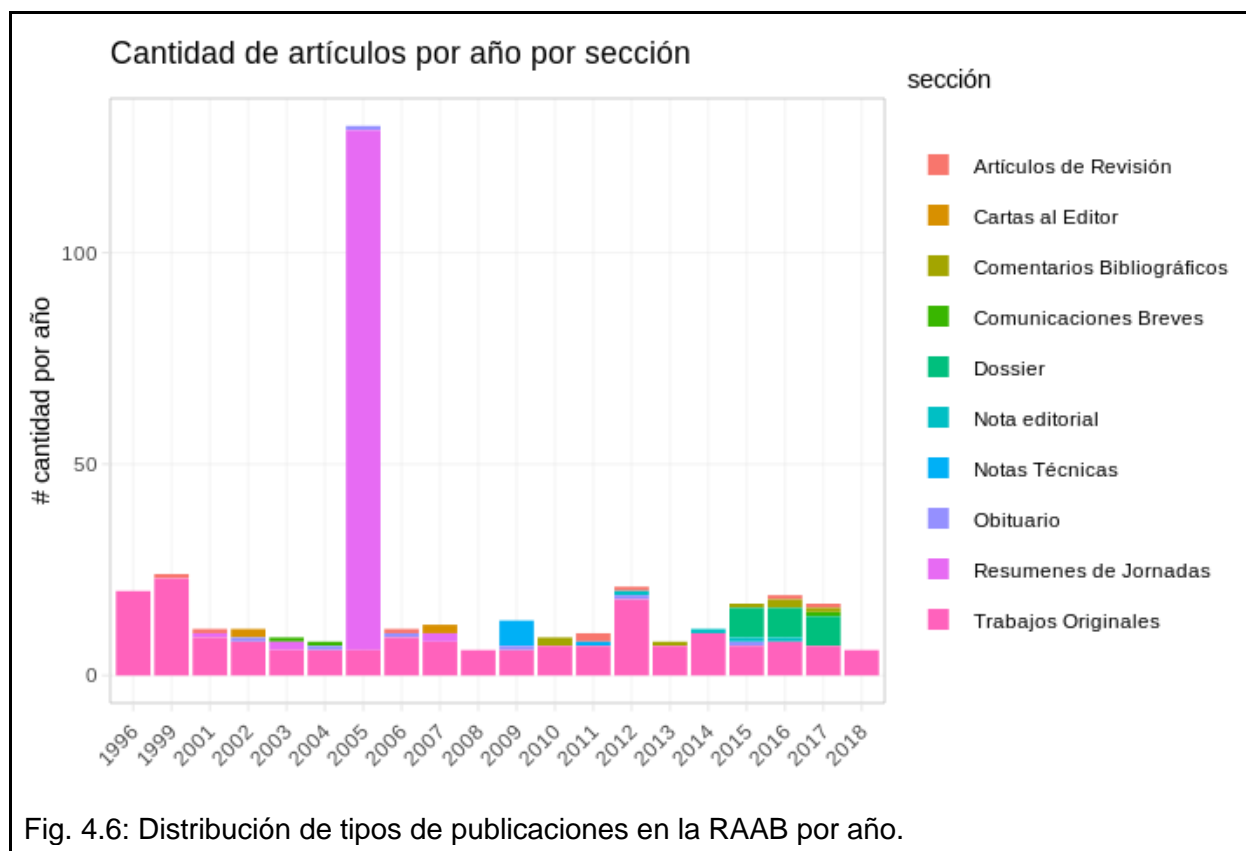
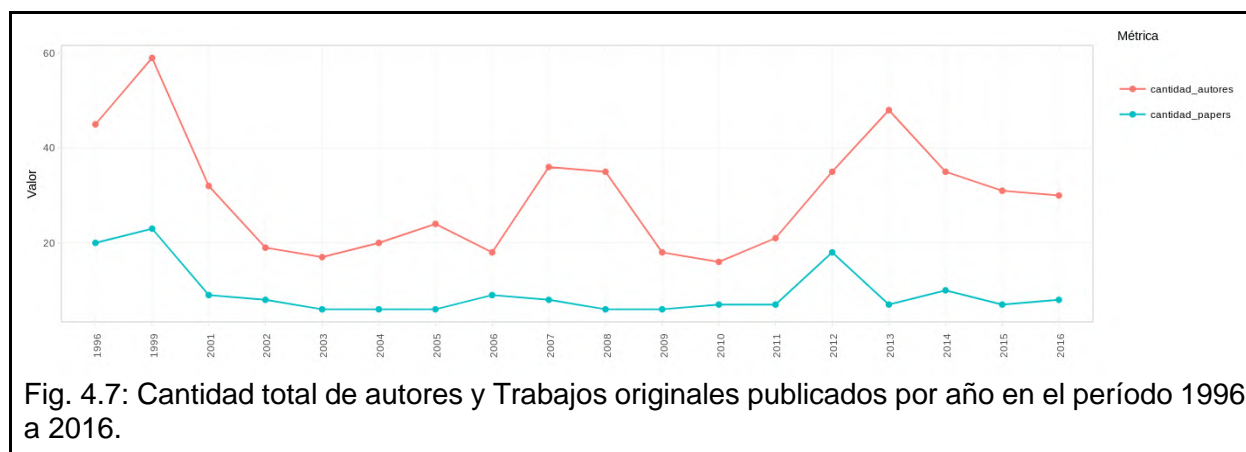


Fig. 4.5: Distribución de artículos en las distintas ediciones de la RAAB por año.

Cuando se evaluó la composición por secciones (es decir, tipo de publicación) se observó que esta diferencia en la cantidad total de publicaciones en el año 2005 se debe a la inclusión de los resúmenes de los trabajos presentados en las Jornadas Nacionales de Antropología Biológica ([Fig. 4.6](#)).



Esta contextualización inicial puso en evidencia la complejidad de incorporar distintos tipos de publicaciones al análisis. Por lo tanto, y retomando lo planteado originalmente, en esta tesis analizamos la cantidad de autores y artículos de tipo Trabajo original, en el período de tiempo acotado entre 1996 y 2016 ([Fig. 4.7](#)).



Observamos que en la etapa inicial la RAAB tuvo un fuerte impulso, tanto en cantidad de autores como artículos. Sin embargo, el patrón general no es el mismo para ambas medidas, lo cual amerita su análisis por separado.

Por un lado, con respecto al número total de Trabajos originales, se observó que a partir del 2001 se alcanzó un valor que corresponde a la media para la RAAB, y que se mantuvo sin grandes

variaciones en los años que siguieron ([Fig. 4.7](#)). Este patrón es de esperarse en revistas científicas puesto que, por lo general, el número de publicaciones que se admiten en cada edición está pautado previamente y se mantiene constante o sin mayores variaciones a lo largo de los años. En este sentido, el crecimiento de una revista no implica necesariamente un aumento en la cantidad de artículos publicados, sino que el mantenimiento de la cantidad de publicaciones a lo largo de los años también puede considerarse como síntoma de “buena salud”. Contrariamente, la disminución en la cantidad de artículos sí podría vincularse a períodos de menor actividad de la revista y/o de la disciplina asociada.

Distinto es lo que ocurrió con la cantidad de autores. En este caso, descartando el impulso inicial de los dos primeros años, se observó un aumento en líneas generales pero con la existencia de picos en determinados años, no necesariamente asociados a un incremento en el número de publicaciones ([Fig. 4.7](#)). Por ejemplo, en 2013 hubo un fuerte aumento en la cantidad de autores, quedando como segundo año con mayor cantidad de autores para trabajos originales, luego del impulso inicial. Esto podría estar asociado al incremento en la cantidad de artículos originales publicados en 2012, que corresponde al tercer lugar en la línea de vida de la RAAB para este tipo de publicaciones. Sin embargo, otro aumento en la cantidad de autores se observó en los años 2007 y 2008 pero sin asociarse a un incremento marcado en el número de publicaciones ([Fig. 4.7](#)). Este aspecto del análisis se profundizará en el siguiente apartado.

Ahora que comprendemos mejor el contexto anual de los artículos, y su punto de vista individual de cada instante queda revisar cómo se comportan los valores acumulados de la cantidad de autores a lo largo del tiempo ([Fig. 4.8](#)).

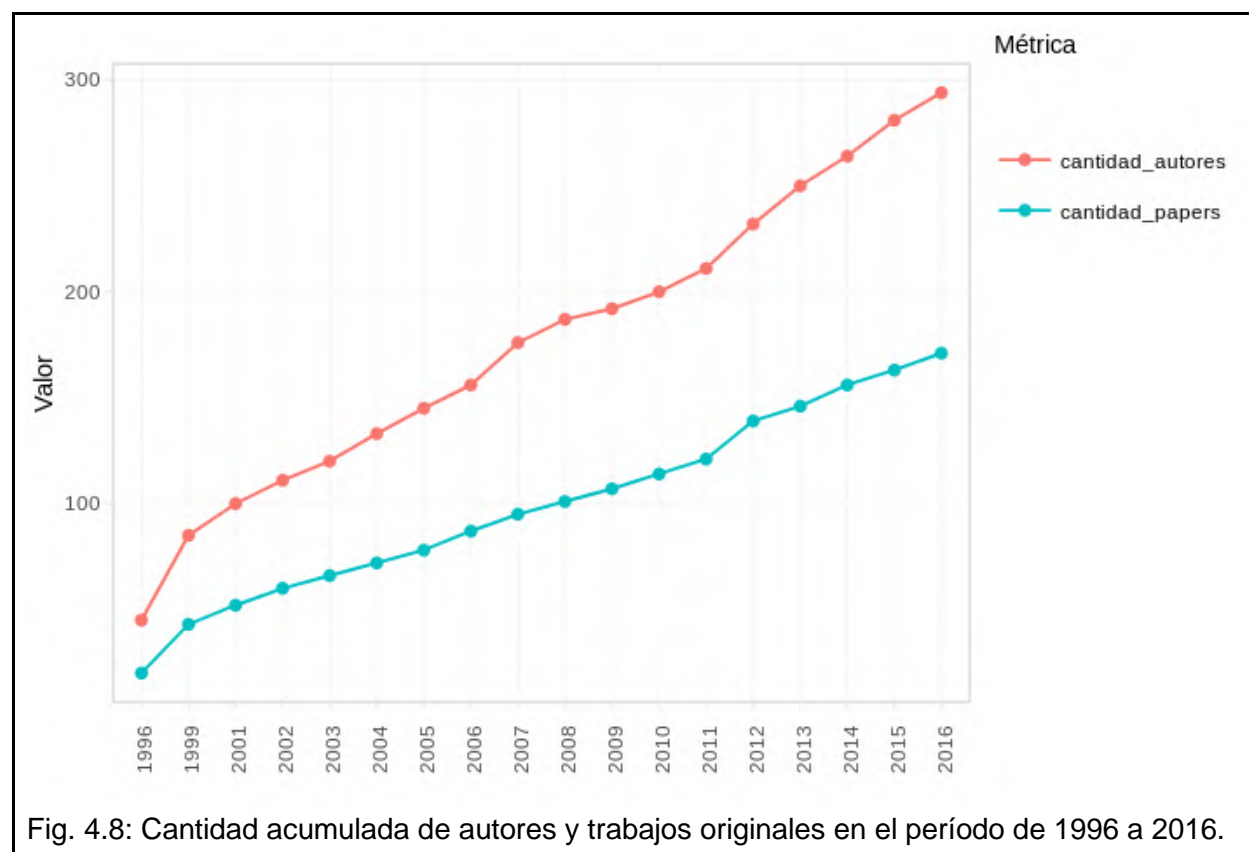


Fig. 4.8: Cantidad acumulada de autores y trabajos originales en el período de 1996 a 2016.

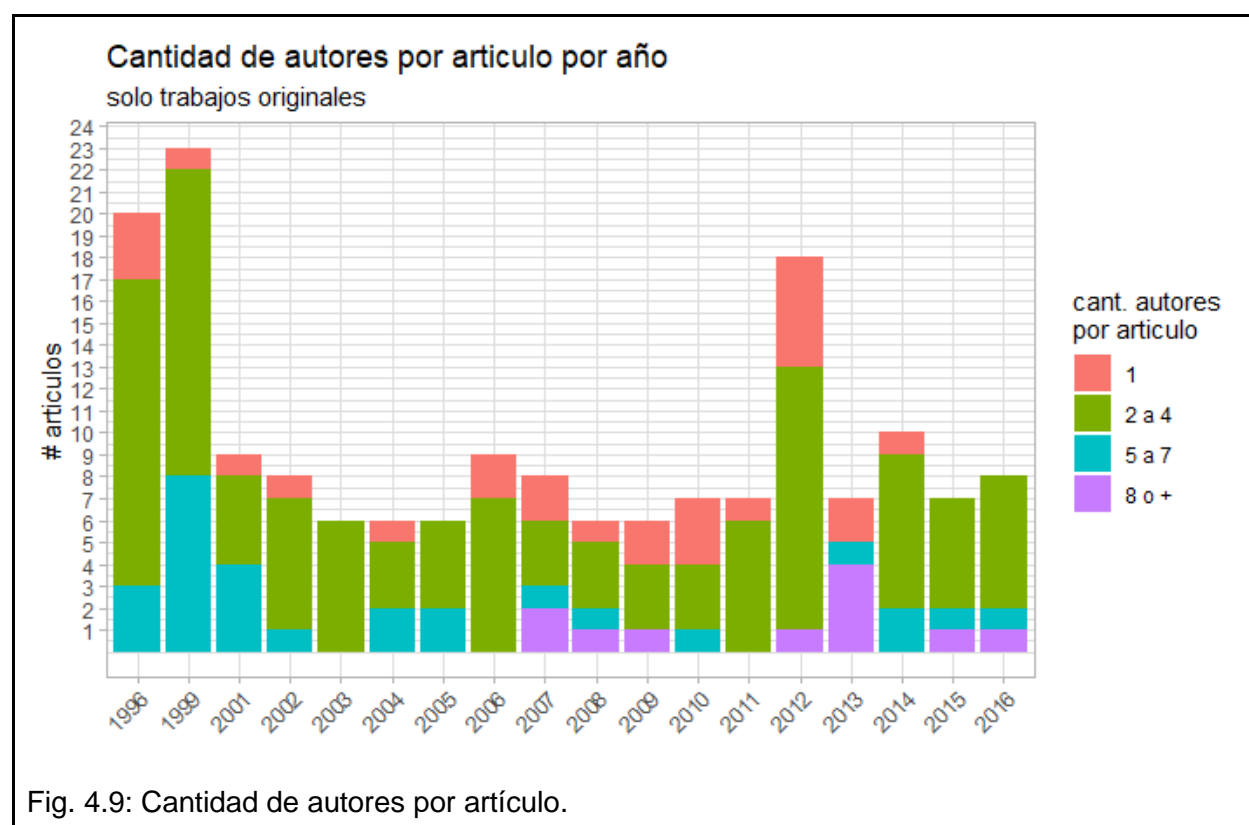
Podemos ver las diferencias mencionadas anteriormente acerca del ritmo de crecimiento en la cantidad de autores y de artículos. En particular, para la cantidad de autores vemos que existen

ciertos ciclos de cambios en la pendiente de crecimiento acumulado, asociados a los picos señalados para los años 1999 (impulso inicial), 2007 y 2012-2013. En todos los casos, estos incrementos en la pendiente parecen sucederse por períodos de aumento relativamente constante ([Fig. 4.8](#)). Observamos, además, que los aumentos en la cantidad de autores son mayores y más variables que en el número de artículos, cuyo incremento acumulado se mantiene relativamente constante por períodos mayores de tiempo. En función de esto, intuimos que estas diferencias podrían deberse a una mayor cantidad de artículos trabajados en conjunto, aspecto que se profundizará en el apartado siguiente.

Tomando en consideración todo lo anterior, podemos decir que no tenemos evidencias para rechazar la hipótesis 1 en cuanto al crecimiento de la cantidad de autores en la RAAB a lo largo del tiempo.

Autores en artículos por año

Elaboramos un poco más respecto de los múltiples autores y vemos que en diferentes momentos varían los patrones de cantidad de autores y cantidad de artículos, no son relaciones lineales, deben existir momentos donde se hacen más artículos con múltiples autores y momentos donde la faceta individual prevalece. Una pregunta interesante a realizar es si con el pasar del tiempo se reducen los artículos individuales y se fortalecen los artículos en equipo. Vemos la [Fig. 4.9](#) :



Vemos que en general predominan los trabajos originales que son realizados en equipo, entre 2 y 7 personas. A partir del 2007, vemos que se agregan equipos de 8 o más personas. En 2015 y 2016, vemos la ausencia de trabajos originales individuales durante más de 1 periodo. Valores correspondientes a la [Fig. 4.9](#) en la [Tabla 4.2](#).

año		1996	1999	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
# autores por artículo	1	3	2	3	2	3	2	45	4	5	1	4	5	2	6	3	1	5	4	3	0
	2 a 4	14	14	4	8	6	4	61	7	4	3	7	3	8	14	0	8	9	13	8	3
	5 a 7	3	8	4	1	0	2	22	0	1	1	1	1	0	0	1	2	1	1	6	1
	8 o +	0	0	0	0	0	0	2	0	2	1	1	0	0	1	4	0	2	1	0	2

Tabla 4.2: cantidad de artículos con cantidad de autores por año.

Hipótesis 2: Los autores fundadores se refuerzan en el tiempo, lo cual puede ser medido por grado, intermediación y fuerza de colaboración

Luego de ver el crecimiento de la revista en cuanto a cantidad de autores y artículos, interesó evaluar si ese aumento estuvo dado por nuevos autores o bien, si correspondió a un reforzamiento de los autores fundadores. La idea detrás de esta hipótesis es revisar si el avance de la revista depende de los autores fundadores, o si está impulsado por la incorporación de nuevos autores en grupos de trabajo en principio independientes de los fundadores.

Para ello, planteamos otra de nuestras hipótesis respecto de los autores participantes: los autores fundadores se refuerzan en el tiempo, lo cual puede ser medido por grado, intermediación y fuerza de colaboración. Recordemos que los fundadores son los autores que publicaron en la primera edición de la revista.

Cada una de estas métricas nos brindan formas de evaluar los nodos de la red desde diferentes puntos de vista. Por otro lado, evaluar la pregunta de si se refuerzan o no en el tiempo da diferentes visiones respecto de los roles que cumplen los autores dentro de la red. Estos roles pueden ser Concentrador (*Hub*), quién trabaja más con los demás y que genera nuevos equipos y es medido por el Grado; Autoridad (*Authorities*), que son generadores de contenido, es decir, quienes más aportes hacen en cantidad de trabajos originales, lo cual está relacionado con la Fuerza de Colaboración; y finalmente quienes generan puentes entre otros autores, quienes tal vez puedan generar nuevas tendencias con mayor facilidad, uniendo equipos con los que trabajaron alguna vez o, también, que su desaparición de la red puede desconectar equipos, asociados a la distancia de Intermediación. Para profundizar de estos roles se puede verificar [Kleinberg \(1999\)](#).

A fin de evaluar estas tres métricas a lo largo del tiempo, se clasificó a los autores para ordenarlos en un *ranking* de cinco puestos según cada una de ellas ([Fig. 4.10](#), [Fig. 4.11](#) y [Fig. 4.12](#)). Para interpretar correctamente las siguientes figuras es necesario comprender que las líneas de colores están asociadas a la posición en el ranking, esto ayuda a visualizar si los autores se mantienen o varían en el tiempo (eje X). Cuando vemos líneas rectas horizontales significa que ese autor se mantiene por un tiempo en la misma posición del ranking. Mientras que la ausencia de las mismas significa que hay mucha variación y que los autores no se mantienen estables en una misma posición a lo largo del tiempo. Por otro lado, las líneas verticales indican que se comparte el puesto del ranking entre varios autores.

De manera complementaria se presentan tablas, para verlo desde un punto de vista numérico (Tabla 4.3, [Tabla 4.4](#) y [Tabla 4.5](#)), se puede ver una tabla a lo largo del tiempo de los primeros 5 puestos. Se toman en cuenta 3 indicadores por cada puesto: **mantiene** con valor 0 para no y 1 para sí, respecto de si el autor se mantiene en el puesto respecto del tiempo anterior.

es_fundador, para representar si el autor estuvo en la primera edición de la revista o no, con valor 0 para representar que NO es de la primera edición, y valor 1 para representar que sí es de la primera edición. Finalmente la variable **valor**, que representa el valor obtenido de la métrica analizada para el puesto en ese año.

En la última columna se muestra **ratio_fundador** que representa del total de los años analizados, cuánto corresponde a si fue fundador o no. Para esta métrica, un valor de 1 equivale a 100 %, significa que en todos los años evaluados fue un autor fundador el que estuvo en esa posición. Si fuese menor significa que hubo momentos en los cuales el autor no fue fundador.

A continuación se muestran las figuras correspondientes a cada métrica seguidas por el detalle de cada una.

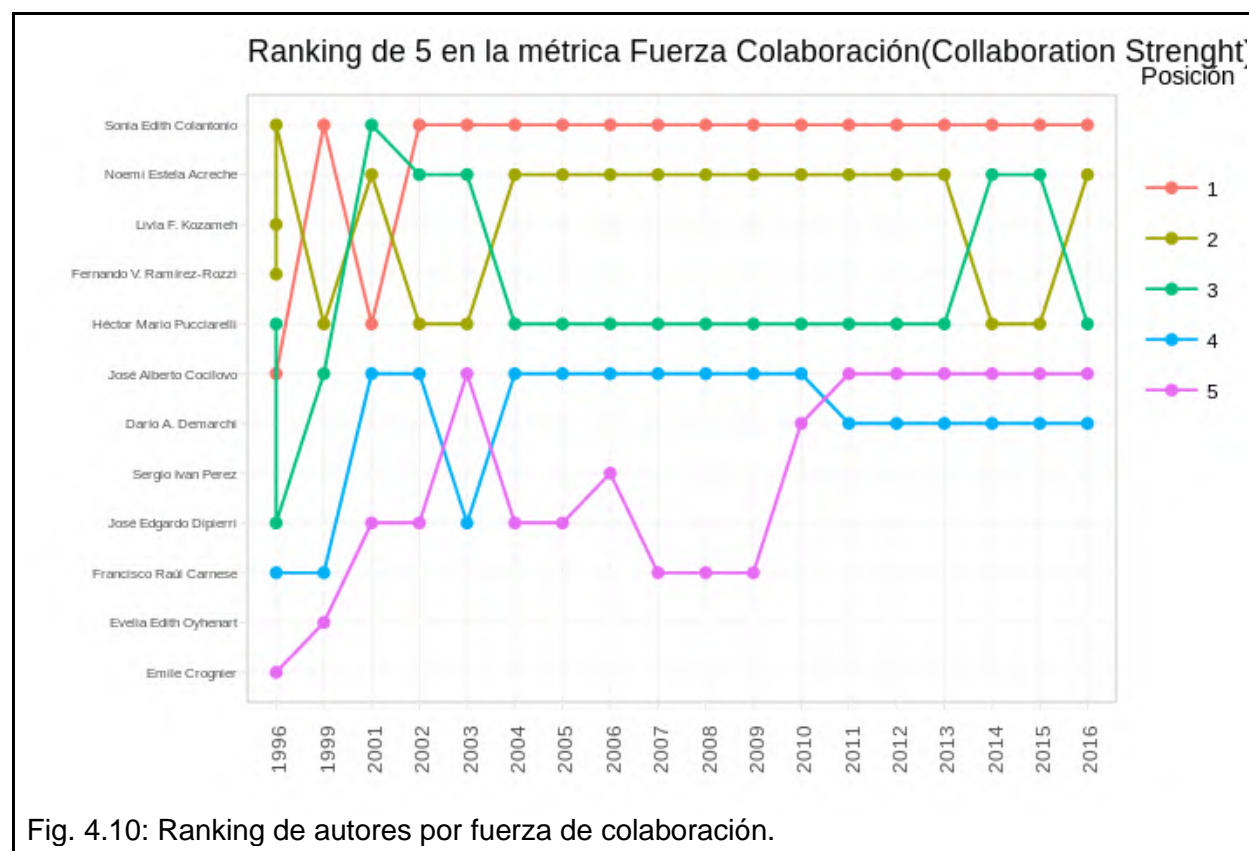


Fig. 4.10: Ranking de autores por fuerza de colaboración.

Recordemos que la métrica fuerza de colaboración es artículo / cantidad Autores. Esto significa que el ranking de la métrica representa quienes mejor están posicionados respecto a producción de contenido de trabajos originales. Si en un año se ven dos puntos ([Fig. 4.10](#)), significa que esos autores 'empataron' en cuanto a fuerza de colaboración. Vemos que los autores fundadores se refuerzan en el tiempo y compiten por sus posiciones en los primeros tres puestos. Los puestos 4 y 5 son más variables, pero a partir de 2011 se mantienen.

Por lo tanto, vemos que respecto a la métrica fuerza de colaboración, la generación de contenido se mantiene bastante estable a partir del 2002. Como análisis complementario quedaría

pendiente ver la variación interanual para evaluar cuánto crecen estos valores año a año. Los valores correspondientes a la [Fig. 4.10](#), están en la [Tabla 4.3](#).

métrica	#	variable	1996	1999	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	ratio fundador
Fuerza de colaboración	1	mantiene	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.00
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	2	mantiene	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1.00
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	3	mantiene	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1.00
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	4	mantiene	0	1	0	1	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1.00
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	5	mantiene	0	0	0	1	0	0	1	0	0	1	1	0	0	1	1	1	1	1	0.94
		es_fundador	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	
		valor	0.6167	1.3429	1.7262	1.7262	1.9167	1.9762	2.1429	2.3095	2.4111	2.4111	2.5361	2.5833	2.8333	2.8333	2.8333	2.8333	3.1667	3.1667	

Tabla 4.3: Valores del ranking de Fuerza de Colaboración. Análisis de si se mantiene y si es fundador.

Para Fuerza de Colaboración, se aprecia que salvo el quinto puesto del ranking, se mantienen los fundadores en el tiempo; y en el quinto puesto el 94% del tiempo son autores fundadores.

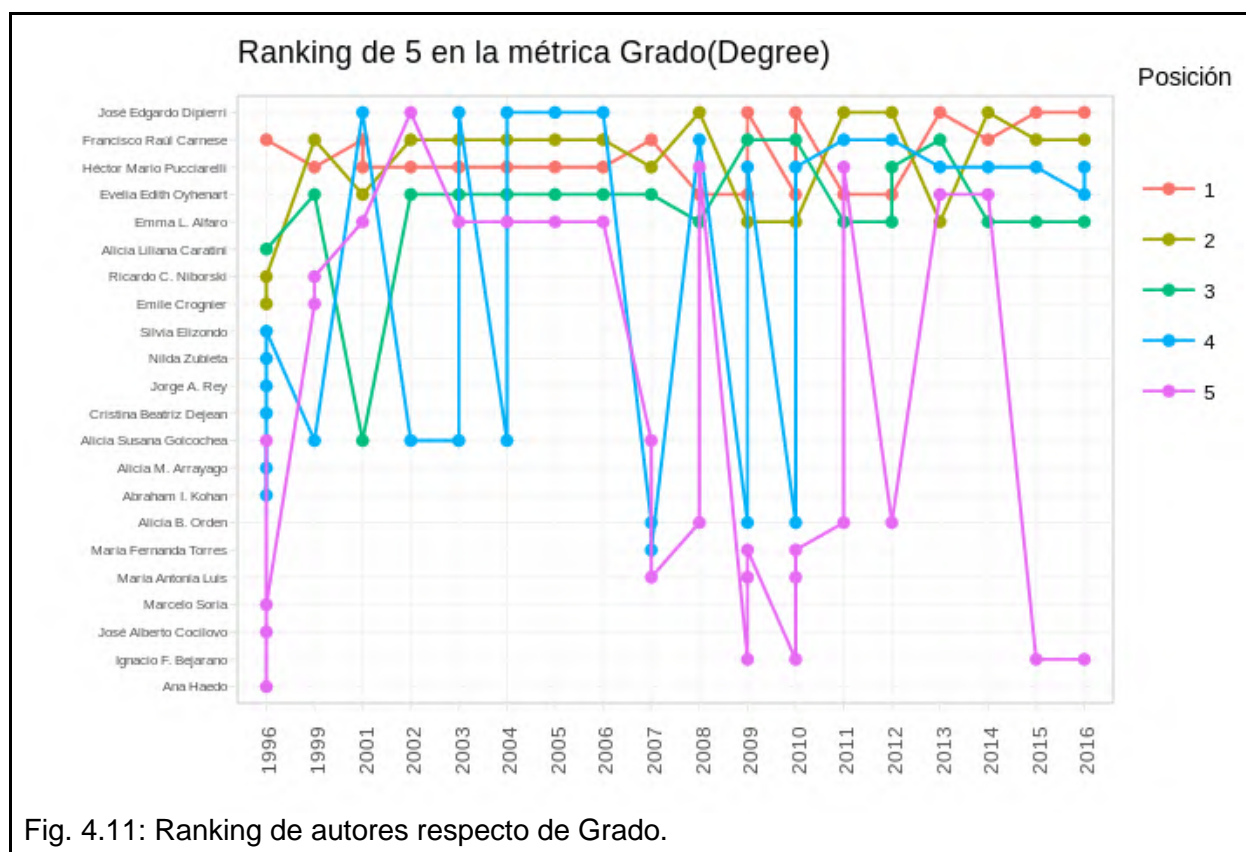


Fig. 4.11: Ranking de autores respecto de Grado.

En la [Fig. 4.11](#) se observa la métrica Grado que indica cuánto se conecta directamente este nodo con otros nodos. En el contexto de red de coautoría hace referencia a con cuántos otros autores publicó ese año. Este ranking es competitivo, anualmente vemos más empates que en el de fuerza de colaboración ([Fig. 4.10](#)). Con respecto a los autores fundadores, vemos que los

primeros cuatro puestos compiten bastante, entre 2002 y 2006 estuvieron estables, mientras que los puestos 4 y 5 tienden a variar más.

Es interesante ver cómo algunos nodos evolucionan en el tiempo. Por ejemplo, el primer nodo aparece en 2001 en el puesto 4, luego en 2002 baja al puesto 5, en 2003 vuelve al 4 y se mantiene hasta 2006. En 2007, no aparece. Luego en 2008 aparece con más fuerza, saltando al puesto 2, en 2009 y 2010, escaló hasta el primer puesto y en los años siguientes, se mantiene entre el puesto 1 y el 2. Esto significa que, en el contexto de la revista analizada, dicho autor es un gran generador de colaboraciones e integrador de nuevos nodos a la red. Lo cual tiene gran valor al impulsar el interés por la revista dentro de la disciplina en el país.

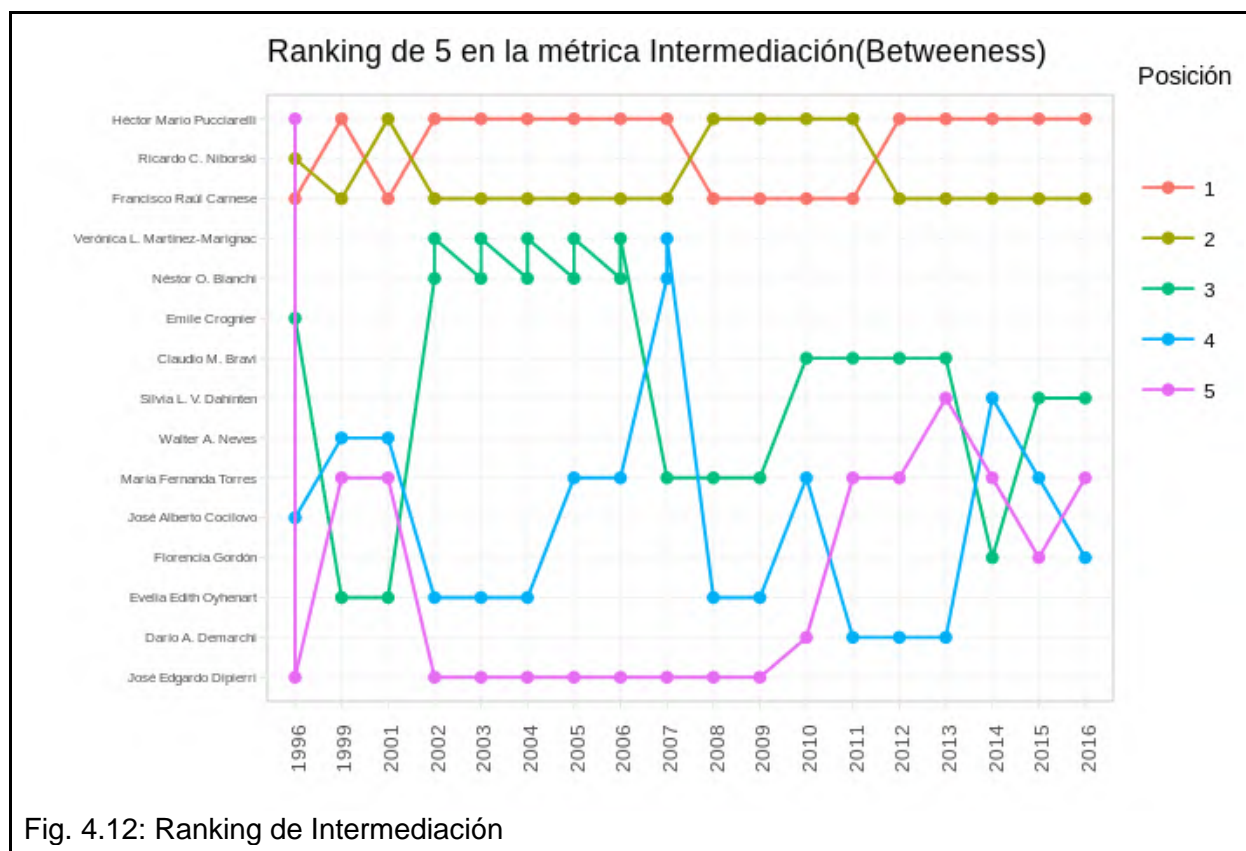
Como análisis complementario se podría incorporar la variación interanual para ver cuánto crecen estos valores año a año. Recordemos que el Grado es la cantidad de nuevos nexos que genera el autor (son nuevos porque no estamos considerando fortalecimiento de una relación por hacer más trabajos en conjunto, para eso está la Fuerza de Colaboración).

metrica	#	variable	1996	1999	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	ratio fundador	
grado	1	mantiene	0	0	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1	1.00	
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
		valor	13	23	24	27	27	27	27	27	31	34	34	34	36	36	43	44	49	49		
	2	mantiene	0	0	0	0	1	1	1	1	1	0	0	0	1	0	1	0	0	0	1	1.00
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		valor	9	21	15	24	24	24	26	26	27	33	33	33	34	34	42	43	44	44		
	3	mantiene	0	0	0	0	1	1	1	1	1	0	0	0	1	0	1	0	0	1	1	1.00
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		valor	6	13	13	15	15	15	18	18	23	32	32	32	33	33	39	42	42	42		
	4	mantiene	0	0	0	0	1	1	1	1	1	0	0	0	0	0	1	0	1	1	1	0.83
		es_fundador	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	
		valor	5	10	12	13	13	13	16	16	18	31	30	30	32	32	38	40	40	40		
	5	mantiene	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	0	1	0	1	0.94
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	
		valor	4	9	11	12	12	12	15	15	15	17	30	29	29	30	30	37	37	39	39	

Tabla 4.4: Valores del ranking de Grado. Análisis de si se mantiene y si es fundador.

Tabla 4.4: Valores del ranking de Grado. Análisis de si se mantiene y si es fundador.

Para Grado, se aprecia que salvo el cuarto y quinto puestos del ranking, se mantienen los fundadores en el tiempo; y en el cuarto puesto el 83% del tiempo son fundadores y el 94% del tiempo son fundadores para el quinto puesto ([Tabla 4.4](#)).



La métrica de Intermediación mide la importancia de un autor según su participación en caminos para llegar de un autor a otro. En la [Fig.4.12](#) vemos que hay más empates para esta métrica que en comparación a Fuerza de Colaboración, pero menos que en el caso de Grado. Los primeros dos puestos se mantienen bastante estables sin embargo los 3, 4, y 5, varían bastante. Respecto de los autores fundadores, se mantienen los primeros 2 puestos mientras que el resto cambia. Llama la atención como el puesto 5 se mantiene estable durante el periodo de 2002 a 2009.

métrica	#	variable	1996	1999	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	ratio fundador
Intermediación	1	mantiene	0	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1.00
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
		valor	26.500	658.500	736.018	1601.177	1656.177	1691.292	1853.668	1885.902	2272.349	1990.606	2087.649	2438.818	2522.497	2777.240	4124.605	6355.776	6614.722	7059.152	
	2	mantiene	0	0	0	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Intermediación		valor	12.533	591.979	721.052	1064.907	1085.463	1069.504	1297.290	1309.977	1777.675	1398.516	1452.230	2161.094	2225.381	2715.160	2663.442	3595.673	3728.957	3891.051	1.00
	3	mantiene	0	0	1	0	1	1	1	1	0	1	1	0	1	1	1	0	0	1	
		es_fundador	1	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	
		valor	10.333	210.500	259.614	381.683	411.163	418.163	542.122	550.622	723.950	916.141	1027.498	1368.792	1502.125	1985.458	2554.467	2223.364	2435.435	2503.777	
	4	mantiene	0	0	0	1	0	1	0	1	0	0	1	0	0	1	1	0	0	0	
Intermediación		es_fundador	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	0.72
		valor	5.000	158.667	180.000	291.614	293.614	300.500	482.103	540.090	635.622	727.105	758.956	1141.215	1236.000	1749.000	2184.000	2223.058	2357.307	2444.122	
	5	mantiene	0	0	1	0	1	1	1	1	1	1	1	0	0	1	0	0	0	0	
		es_fundador	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	
		valor	2.000	145.794	154.595	259.465	292.673	297.212	424.005	430.032	490.295	490.767	553.594	1096.000	1179.789	1252.436	2029.277	2011.820	2332.303	2401.666	

Tabla 4.5: Valores del ranking de intermediación. Análisis de si se mantiene y si es fundador.

Para Intermediación, el primero y segundo puesto son estables respecto de los fundadores. Sin embargo, el tercer puesto, tiene un 33% del tiempo autores fundadores. El cuarto puesto tiene autores fundadores un 72% del tiempo, y el quinto puesto tiene autores fundadores un 89% del tiempo ([Tabla 4.5](#)).

Vemos que en las tres figuras hay muchos nombres que varían según la métrica considerada, lo cual significa que diferentes autores han cumplido diferentes funciones en la RAAB dentro del período analizado. En los ranking que involucran Grado e Intermediación se comparten más autores, sin embargo en Fuerza de Colaboración varían.

Estas métricas son contextuales en cada tiempo de evaluación y respecto a la red en dicho momento, por lo tanto mantenerse entre los primeros 5 significa que se refuerzan en el tiempo, como puede ser validado en los números de cada métrica.

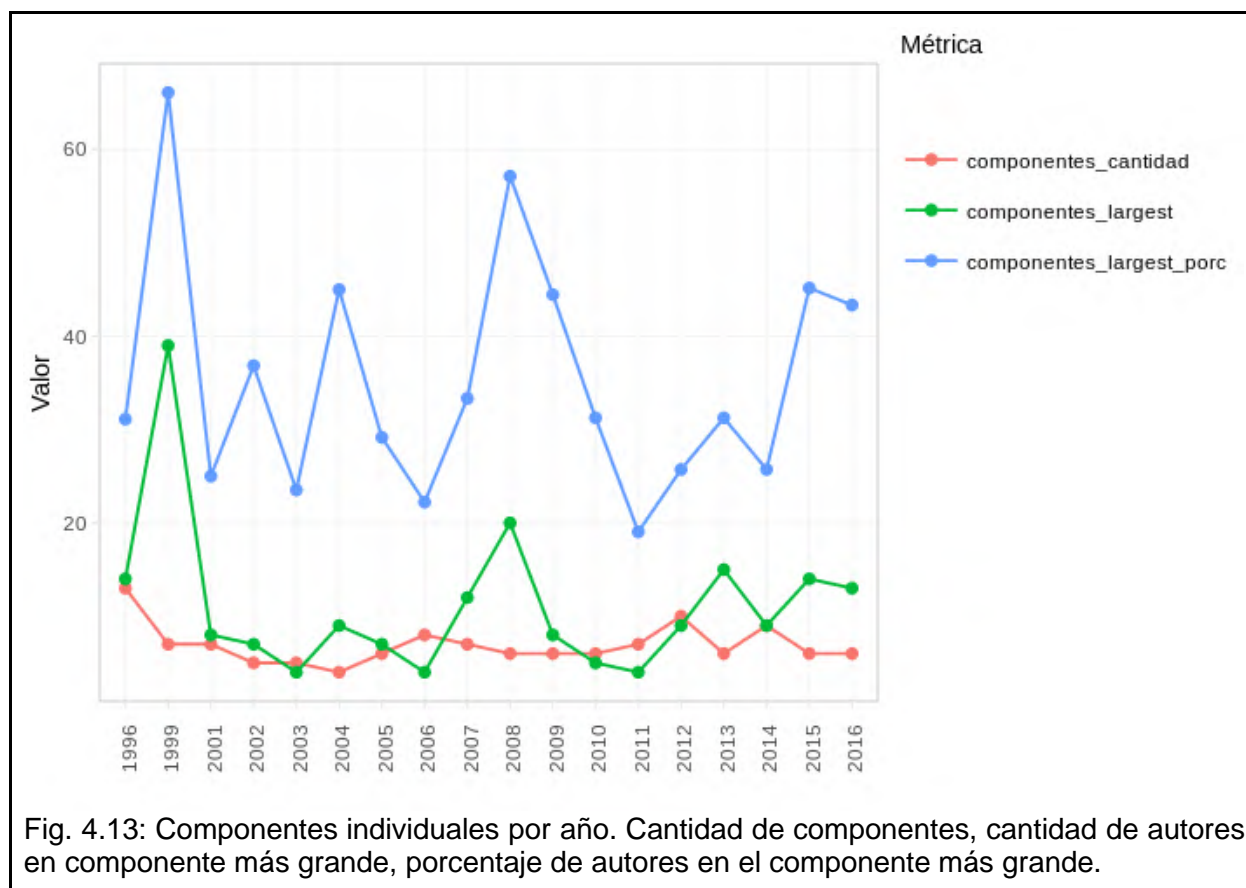
Respecto de la hipótesis acerca de que los fundadores se refuerzan en el tiempo, los resultados indican que no habría evidencias para rechazarla dado que mayormente se mantienen los autores fundadores entre los primeros 5 puestos de todas las métricas a lo largo del tiempo.

Hipótesis 3: Los componentes más grandes absorben a los más pequeños a lo largo del tiempo, dado que autores periféricos pasan a trabajar en conjunto con autores principales

Resultó de interés contestar esta hipótesis dado que representa la integración de individuos o equipos a la comunidad local mayormente conectada, que es el subgrafo representado por el componente mayor. Recordemos que un componente es un grupo de nodos que están conectados entre sí y está desconectado de otros grupos. Los autores individuales son componentes en sí mismos. Cuando existe más de un componente se califica a la red total como una red inconexa, lo cual es natural en redes de coautoría dado que dentro de una comunidad científica, existen autores individuales y grupos que todavía no realizaron colaboraciones a través de ninguno de sus autores, o que publican en otros lugares además de la revista analizada.

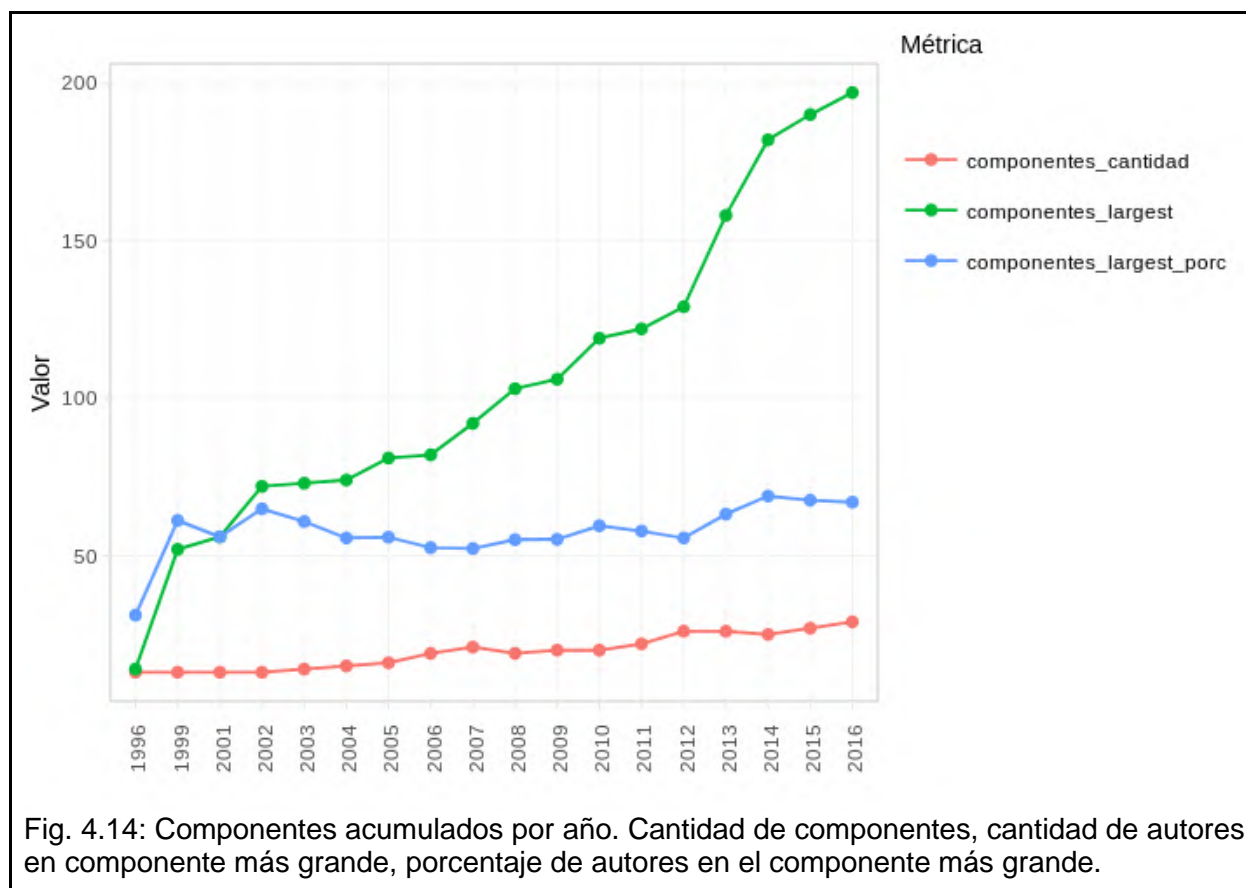
En cuanto al orden, esta hipótesis fue planteada en tercer lugar porque necesitamos un contexto previo para realizar este análisis; luego de ver la cantidad de autores por artículos por año, es interesante ver cómo se van generando los componentes, cómo se relacionan entre ellos y su integración en el tiempo.

Como vimos anteriormente, podemos realizar el análisis desde el punto de vista anual o agregado obteniendo diferente información respecto de su evolución temporal.



Desde un punto de vista anual ([Fig. 4.13](#)), vemos que en 1999 tuvo su pico de participación de autores involucrados en el componente principal para una misma edición. Observando ese año en el gráfico de autores por artículo ([Fig. 4.9](#)), vemos que es el año con mayor cantidad de trabajos originales presentados, de estos su mayoría artículos entre 2 y 7 autores. Volviendo a los componentes, vemos que hay un segundo pico en 2008. Si vemos el detalle de composición mencionado en la hipótesis 1, esto es resultado de un crecimiento en la cantidad de autores y cantidad de artículos, que podría estar relacionado con que en ese año se publicó un número especial. Se observa que fue el 3er número de mayor publicación en la historia de la RAAB, no se tenían esas cantidades desde los números iniciales del 1996 y 1999. En cuanto a cantidad de artículos es similar a la de los primeros números, solo que existe un artículo de 8 o más personas, más artículos individuales, y no hay artículos entre 5 y 7 personas.

Este punto de vista nos da una noción de cuando la red tiene incorporaciones de autores y cómo se comportan respecto de realizar trabajos en conjunto de manera anual. Por otro lado, el acumulado para ver cómo queda la red total agregada en su evolución temporal, nos brinda información acerca de cuándo se realizan absorciones importantes respecto del componente mayor.



A diferencia del gráfico anterior ([Fig. 4.13](#)) podemos ver los saltos de cantidad de autores en el componente mayor. Estos saltos significan que un componente grande fue absorbido por el componente mayor. Lo podemos ver en las transiciones de 1996 a 1999, de 2012 a 2013 y de 2013 a 2014 ([Fig. 4.14](#)).

Teniendo en cuenta las observaciones de 2012, se ve que alguno de los componentes originados, fue absorbido en el 2013 al realizar trabajos de colaboración. A continuación vamos a verificar estas supuestas absorciones a través de revisar las transiciones estructurales de los grafos de forma visual.

Verificación de absorción de componentes en la red

Veamos la red en estos instantes para verificar ([Fig. 4.15](#)). En la aplicación se pueden ver de forma dinámica las transiciones.

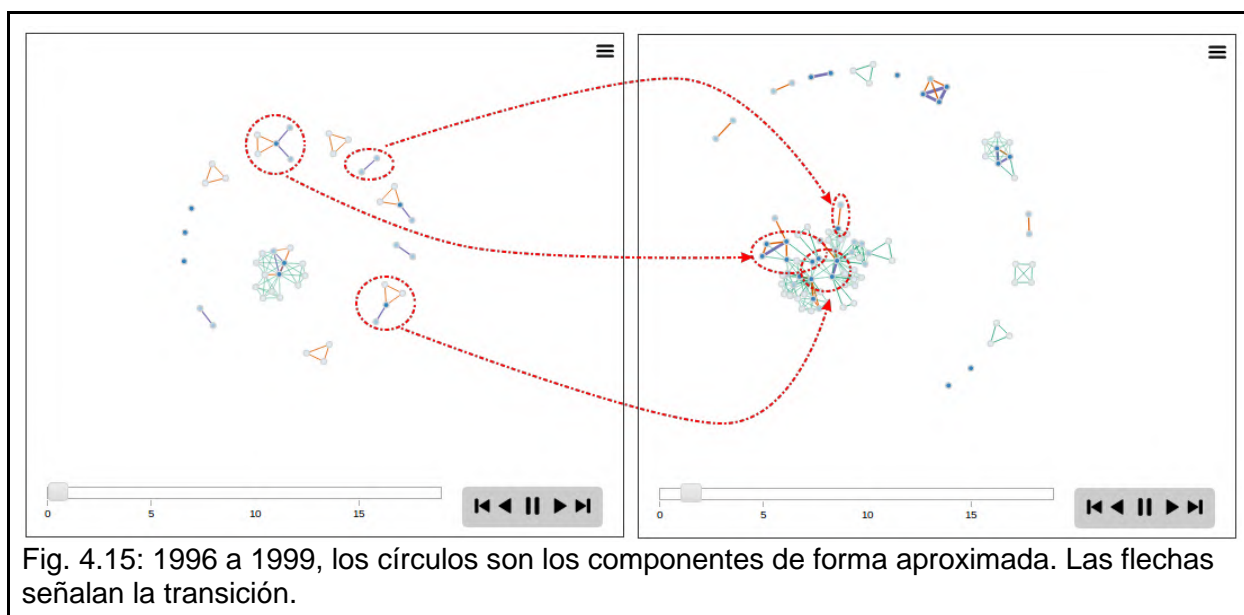


Fig. 4.15: 1996 a 1999, los círculos son los componentes de forma aproximada. Las flechas señalan la transición.

Vemos que en la primera transición el componente mayor absorbe 2 componentes muy importantes del tiempo anterior de 5 y 4 autores. Y un tercer componente de 2 nodos. Además aparecen nuevos nodos en el nuevo año haciendo que pase de 14 a 52 autores en 1999 (crecimiento de 38 autores).

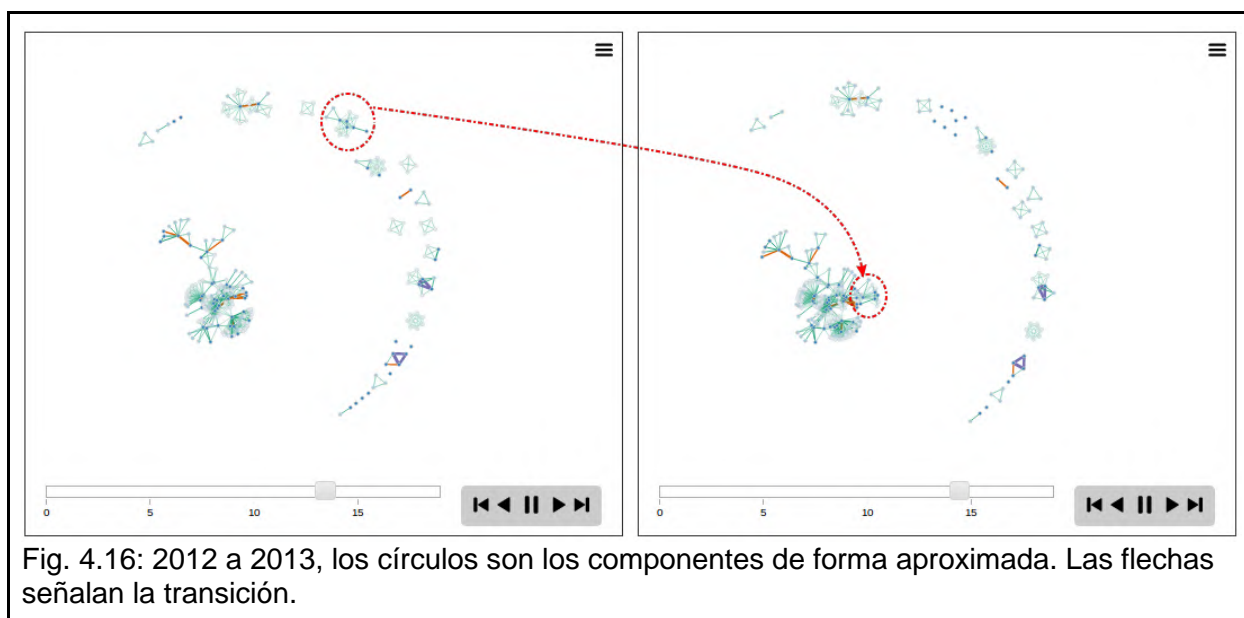
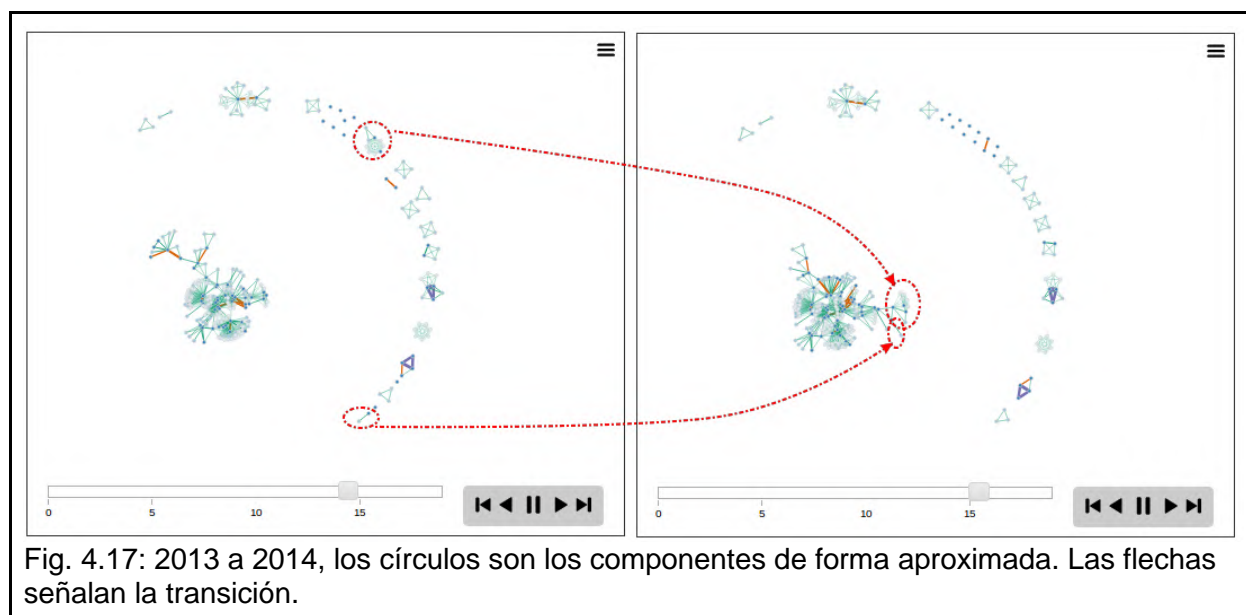


Fig. 4.16: 2012 a 2013, los círculos son los componentes de forma aproximada. Las flechas señalan la transición.

Para el período siguiente ([Fig. 4.16](#)) vemos que uno de los componentes más grandes (12 autores) es absorbido por el componente mayor. De forma complementaria se agregan directamente nuevos nodos al componente mayor pasando de 129 a 158 autores en 2013 haciendo un crecimiento de 29 autores.



Nuevamente, en el período de 2013 a 2014 ([Fig. 4.17](#)), vemos que uno de los componentes más grandes (9 autores) es absorbido por el componente mayor. De forma complementaria también absorbe un componente menor de 2 autores además se agregan directamente nuevos nodos al componente mayor pasando de 158 a 182 en 2014 haciendo un crecimiento de 24 autores.

Para terminar de revisar la hipótesis vamos a revisar numéricamente en la [Tabla 4.6](#).

Tabla 4.6: Cantidades de nodos Autores en componente mayor.

Cantidad total de nodos, Cantidad de nodos en componente mayor en tiempo inicial de comparación (t_0), Cantidad de nodos en componente mayor en tiempo final de comparación (t_1), porcentaje que representa en (t_1), Cantidad de nuevos nodos en el componente, Cantidad de nodos absorbidos por el componente entre t_0 y t_1 . Entendemos por absorbidos, nodos que ya existían en el la red, y pasaron a formar parte del componente mayor.

Periodo (t_0 a t_1)	Cantidad total t_1	Cantidad componente mayor t_0	Cantidad componente mayor t_1	Representa % Red (t_1)	Nuevos nodos	Nodos absorbidos	Total nuevos nodos en t_1
1996 a 1999	85	14	52	61.18	27	11	38
2012 a 2013	250	129	158	63.2	17	12	29
2013 a 2014	264	158	182	68.9	13	11	24

De esta forma los resultados son concordantes con la hipótesis de que los componentes periféricos pasan a ser parte del componente mayor dado que se generan nuevos trabajos entre los autores de los componentes. Como complemento podemos ver que de los nuevos nodos que se agregan a este componente mayor, una parte corresponde a nuevos nodos publicando en la revista, además de los nodos ya existentes.

Relación con actividades de investigación

De manera complementaria se intentó corroborar lo planteado por [Fatt et al. \(2010\)](#):

El componente gigante podría significar la principal actividad de investigación; mientras que los otros componentes pueden ser agrupamientos especializados o sub-comunidades.

Este análisis se realizó sobre la red total. Para ello, revisamos los títulos de los artículos creados por los 5 componentes con mayor cantidad de autores. Los resultados se pueden ver en la [Tabla 4.7](#).

Tabla 4.7: Artículos, autores, y palabras de mayor frecuencia en los títulos. Para los primeros 5 componentes.

componente	Cantidad autores	Cantidad artículos	Primeras 10 palabras por frecuencia en el grupo
1	197	114	argentina, poblaciones, análisis, provincia, cordoba, población, jujuy, influencia, populations, craneofacial
2	21	10	tierra, fuego, grande, regional, fagnano, río, argentina, resultados, isótopos, estables
3	9	11	Salta, argentina, genética, poblacional, tamaño, localidades, homogamia, genéticas, calchaquí, demografía
4	7	1	Rio, province, population, racial, polymorphisms, pinar, genetic, cuba
5	5	1	Risk, riesgo, grupos, etnicos, ethnic, congenitas, congenital, anomalies, anomalías, america

En las siguientes figuras vemos la frecuencias de las palabras obtenidas por componente para los primeros tres puesto que, los grupos 4 y 5 solo tienen publicado un artículo, lo que lleva a que las palabras que figuran son solo las que aparecen en dicho artículo ([Tabla 4.7](#)). Por lo tanto, estos últimos dos componentes quedaron fuera del análisis de frecuencia de palabras. Por otro lado, consideremos también que el componente principal tiene 114 artículos, mientras que el segundo tiene solo 10 artículos, y el tercero 11 lo cual repercute en las frecuencias de palabras de las figuras.

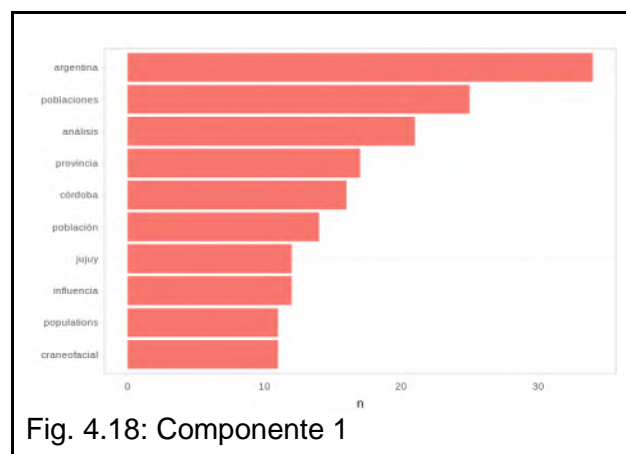


Fig. 4.18: Componente 1

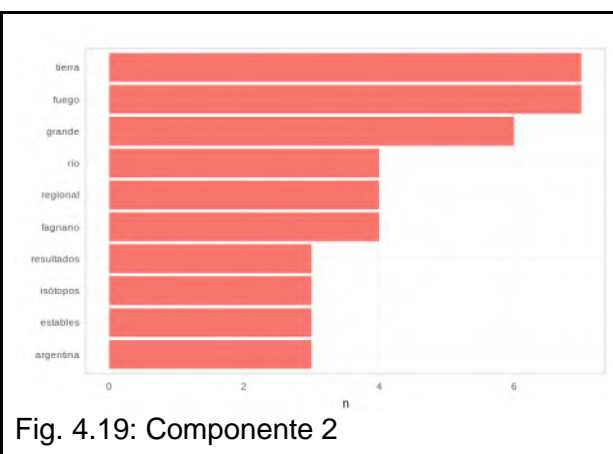


Fig. 4.19: Componente 2

Lo que llama la atención del 1er componente ([Fig. 4.18](#)) es que se destacan algunas localidades, como Córdoba y Jujuy, a pesar de ser un componente muy heterogéneo a nivel geográfico. Sin

embargo, la presencia del término “craneofacial”, estaría indicando que este componente tiene una relación estrecha con esa línea de investigación en varias regiones del país. Vemos también que aparecen palabras en inglés, como “populations”.

Con respecto al segundo componente ([Fig. 4.19](#)), puede verse que es bien local de Tierra del Fuego (“tierra del fuego”, “río grande”), aunque también parece asociado a una línea de investigación por la presencia de las palabras “isótopos” y “estables”.

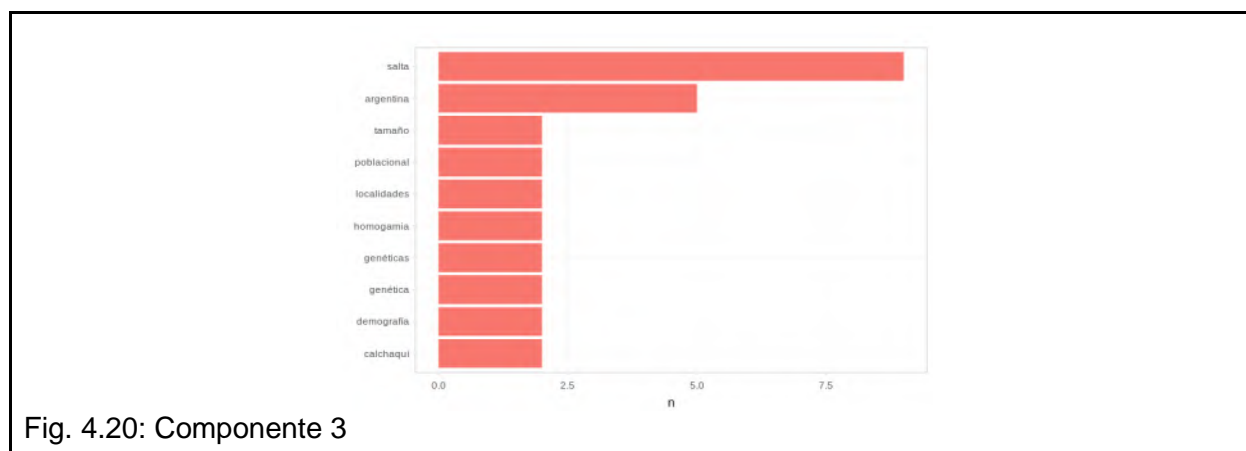


Fig. 4.20: Componente 3

En el componente 3 ([Fig. 4.20](#)) también vemos que se destaca otra localidad, Salta, haciendo que lo geográfico sea el factor común, mientras que parecería ser heterogéneo en términos de las líneas de investigación que incluye.

Vemos que los resultados obtenidos pueden interpretarse acorde al postulado de que el componente principal abarca la principal actividad de investigación, mientras que los secundarios abarcan temas más especializados ([Fatt et al. 2010](#)). Aún teniendo en cuenta la cota de 10 palabras en el análisis realizado, puede verse que, en términos geográficos, el primer componente es más heterogéneo que los dos más pequeños. Además, en relación a las actividades de investigación, es probable que la línea que surge del componente más grande (asociada a la palabra “craneofacial”) sea la más frecuente dentro de la disciplina a nivel nacional, lo cual repercute en una mayor cantidad de artículos publicados en la RAAB. Queda profundizar el enfoque para comprender con mayor detalle estos componentes así como su composición de temas, sobre todo porque el componente mayor contiene múltiples artículos que seguramente no fueron representados en las primeras 10 palabras. De la misma forma sería interesante, a futuro, revisar en el tiempo los diferentes componentes y sus temas para evaluar si mientras se va armando el componente gigante va absorbiendo las diferentes comunidades, enriqueciendo así los temas tratados a nivel revista. También interesa, a futuro, explorar estas temáticas en un análisis de texto más complejo, teniendo en cuenta, por ejemplo, los conceptos compuestos por varias palabras (como ‘genética poblacional’ o ‘tierra del fuego’).

Búsqueda de comunidades

Además del análisis de componentes que representa cómo se van uniendo a través de trabajos en conjunto, se realizó una búsqueda de comunidades resultantes según diferentes algoritmos. Estos algoritmos buscan según diferentes criterios que suelen ser útiles en la búsqueda de patrones en redes. Para determinar que algoritmo aplica mejor para la red estudiada se midieron a través del criterio de modularidad. De esta forma resta interpretar el resultado para el caso de

la RAAB, ver qué comunidades surgen y si tienen cierto sustento temático como ocurrió con los componentes.

En primer lugar, verificamos si encontramos agrupamientos bajo algún criterio determinado por los algoritmos disponibles y también lo complementamos con análisis de comunidades dentro de la red ([Fig. 4.21](#)).

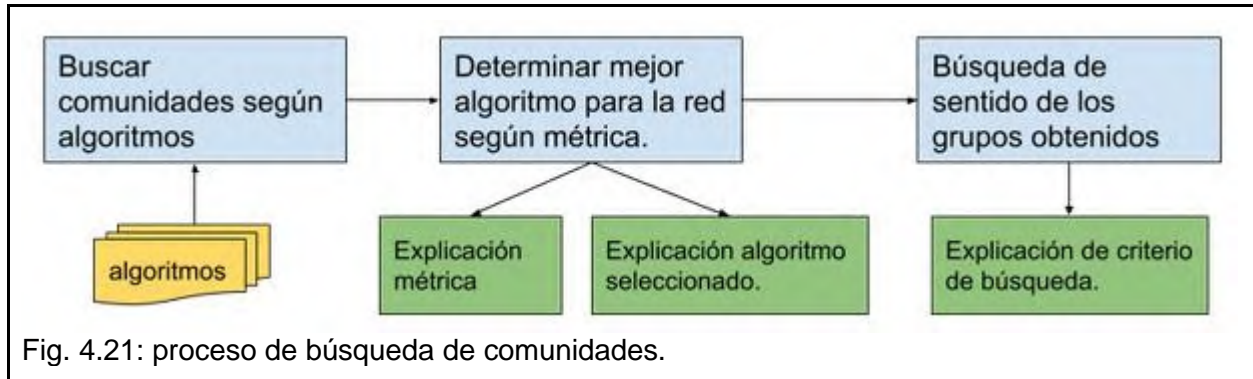


Fig. 4.21: proceso de búsqueda de comunidades.

Explicación de la métrica de evaluación de algoritmos: Modularidad

Para la selección del algoritmo utilizamos la métrica de modularidad. Esta métrica fue propuesta por [Newman-Girvan \(2004\)](#), la implementación utilizada es la que se encuentra en el el paquete `igraph`, que utiliza la versión de [Clauset, Newman y Moore \(2004\)](#).

Siguiendo a [Barabási \(2016\)](#) puede comprenderse mejor la métrica de la siguiente forma:

Para calcular la modularidad de una red con N nodos y conexiones L , particionados en N_c comunidades; y cada una de las comunidades tiene N_c nodos y L_c conexiones, donde c es el número de comunidad 1 a N .

Si el número de L_c es mayor que el número de L esperado para la secuencia de grado de la red, entonces los N_c , podrían ser parte de una comunidad. Entonces para cada comunidad se calcula la diferencia entre la matriz de conexión de la red actual (A_{ij}), y la matriz de conexión de los mismos nodos pero conectados aleatoriamente (p_{ij}), preservando el grado para los nodos.

$M_c = 1/2L * \sum_{(i,j) \in C_c} (A_{ij} - p_{ij})$	
<p> A_{ij} matriz de conexiones actuales, P_{ij} matriz conexiones aleatorias C comunidad dentro de comunidades L conexiones totales de la red $\frac{1}{2}$ porque es un grafo no dirigido. </p> <p> $M_c > 0$: cuanto mayor Modularidad ,mejor el particionamiento. $M_c == 0$: todos los nodos en 1 sola partición $M_c < 0$: cada nodo como una partición separada. </p>	

Por lo tanto, cuanto mayor sea la Modularidad significa que la configuración de conexiones actual es mejor que una conexión aleatoria y mejor que otras configuraciones.

En la [Tabla 4.8](#) podemos ver la comparación de los algoritmos de búsqueda de comunidades en la red agregada total (periodo 1996 a 2016).

Tabla 4.8: Comparación de los algoritmos utilizados para la búsqueda de comunidades.

Algoritmo Utilizado	Cantidad de comunidades	Valor modularidad	Valor modularidad con peso (peso dado por fuerza de colaboración)
Optimización de modularidad multi-nivel (Multi-level Optimization of Modularity) (Louvain)	37	0.692	0.7075

Intermediación de Aristas (Edge Betweenness)	43	0.6744	0.6596
Busqueda de comunidades Infomap (Infomap community finding)	46	0.6732	0.6935
Caminatas aleatorias cortas (Short Random Walks)	47	0.6725	0.6822
Autovector lider (Leading Eigenvector)	35	0.6551	0.6385
Etiquetas Propagates (Propagating Labels)	44	0.6459	0.642

En la [Tabla 4.8](#) podemos ver que el algoritmo mejor posicionado es *Optimización de modularidad multinivel*, sea con o sin considerar los pesos de las aristas. En el segundo puesto podemos ver que si no consideramos los pesos, *Intermediación de aristas* queda seleccionado, mientras que si consideramos los pesos, *Búsqueda de comunidades infomap* queda mejor ranqueado.

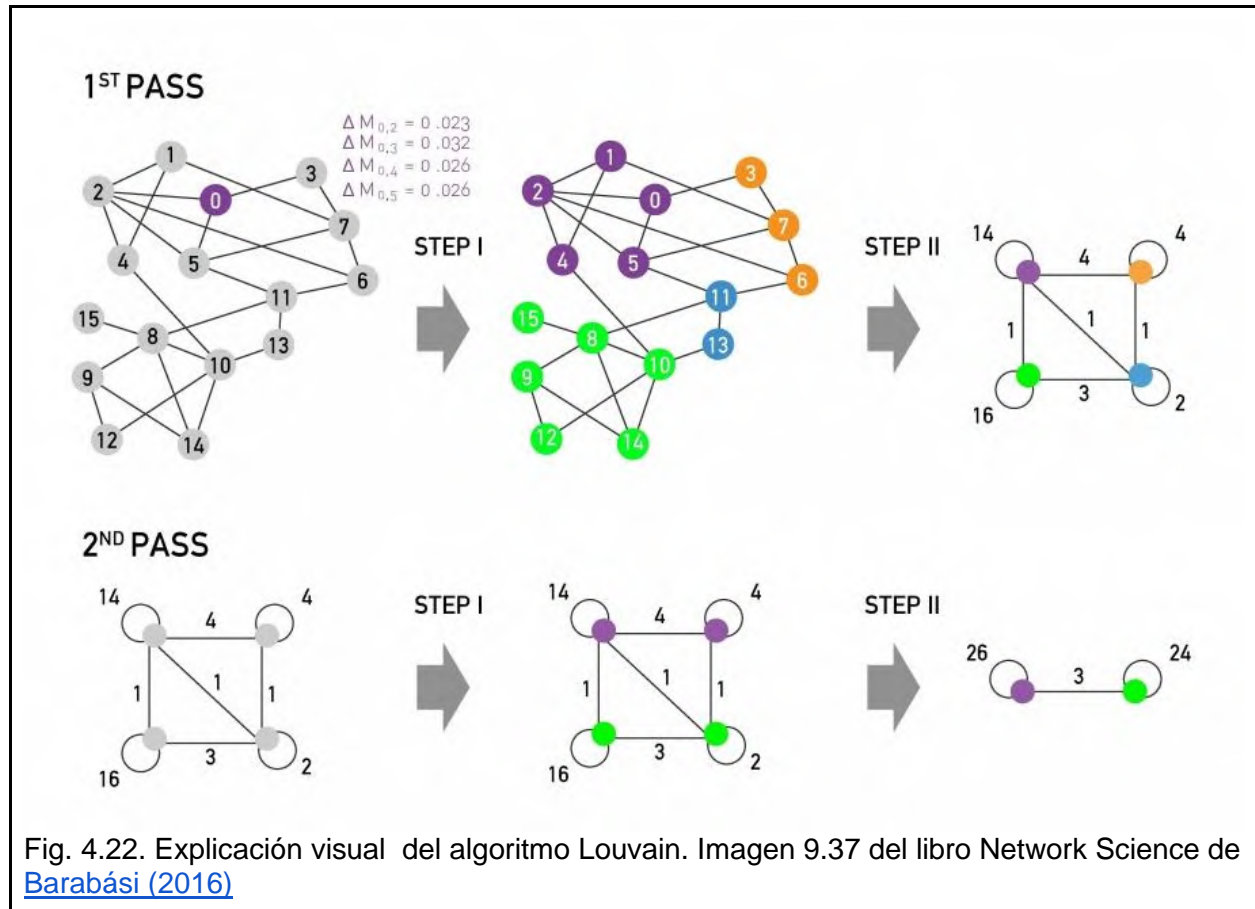
Explicación del algoritmo seleccionado: Optimización de modularidad multi-nivel

A continuación se detalla solamente cómo funciona el algoritmo ganador dentro de los evaluados (para más información de los algoritmos probados ver el [anexo de los algoritmos de comunidades](#)). Para comprender mejor cómo es que se realizan las divisiones con este algoritmo analizamos el trabajo de [Blondel \(2008\)](#) y la explicación del libro de [Barabási \(2016\)](#).

El algoritmo tiene N Pases y cada Pase debe cumplir 2 pasos. El paso 1 es de asignar a todos los nodos a una comunidad propia, y luego analizar cuánto gana en modularidad por incorporarse a una comunidad vecina. Si supera, se muda de comunidad. Este paso se repite hasta no poder mejorar la modularidad de forma local ([Fig. 4.22](#)).

Luego el paso 2 es crear un nuevo grafo que va agregar los nodos de cada comunidad, generando 1 nodo representando a cada comunidad. Las conexiones representan la cantidad de conexiones con las otras comunidades, y los bucles consigo misma representan todas las conexiones internas de la comunidad ([Fig. 4.22](#)).

Luego de esto se termina el Pase, arrancando un siguiente Pase en el que se repiten los pasos 1 y 2. El algoritmo termina cuando no se puede aumentar más la modularidad ([Fig. 4.22](#)).



Búsqueda de sentido en los grupos obtenidos

Estudiando los resultados del algoritmo de *Optimización de modularidad multinivel* podemos visualizar las primeras 5 comunidades ([Fig. 4.23](#)) en las cuales nos vamos a centrar para constatar un sentido más allá de métricas estructurales. Se trató de encontrar sentido a las comunidades según la información que se obtiene a partir de centrarse en sus datos, más específicamente, qué trabajos fueron elaborados por ellas.



Fig. 4.23: primeras 5 comunidades según el algoritmo de *Optimización de modularidad multinivel* (rojo [id comunidad: 10], azul [id comunidad: 37], verde [id comunidad: 23], violeta [id comunidad: 33], naranja [id comunidad: 18])

Explicación de criterio de búsqueda: verificación de comunidades según títulos publicados.

En este análisis nos centramos en las primeras 5 comunidades según cantidad de integrantes. Más allá de la búsqueda de la optimización de la modularidad del algoritmo, evaluamos si podíamos entender cómo están compuestas analizando los títulos de los artículos generados por las comunidades propuestas ([Tabla 4.9](#)). Para ello, realizamos un breve análisis de frecuencias de palabras en los artículos publicados, tomando las primeras 10 palabras de cada comunidad. La idea de estos análisis es tener un bosquejo de qué temas se tratan en dichas comunidades, sin embargo si se encuentra interés, debería servir de inicio para análisis más extensos.

Tabla 4.9: Artículos, autores, y palabras de mayor frecuencia en los títulos. Para las primeras 5 comunidades.

Comunidad	Cantidad autores	Cantidad artículos	Primeras 10 palabras por frecuencia en el grupo
-----------	------------------	--------------------	---

10	41	191	craneofacial (8), crecimiento (8), influencia (7), poblaciones (5), desarrollo (4), prenatal (4), argentina (3), contexto (3), estudio (3), growth (3)
37	37	183	jujuy (6), provincia (5), poblaciones (4), regional (4), argentina (3), distribución (3), peso (3), san (3), adultos (2), adults (2)
23	21	189	influencia (7), argentina (6), crecimiento (6), nutricional (5), población (5), socio (5), infante (4), aires (3), ambientes (3), craneofacial (3)
33	26	63	craneofacial (8), crecimiento (8), influencia (7), poblaciones (5), desarrollo (4), prenatal (4), argentina (3), contexto (3), estudio (3), growth (3)
18	21	45	fuego (7), tierra (7), grande (6), fagnano (4), regional (4), río (4), argentina (3), estables (3), isótopos (3), resultados (3)

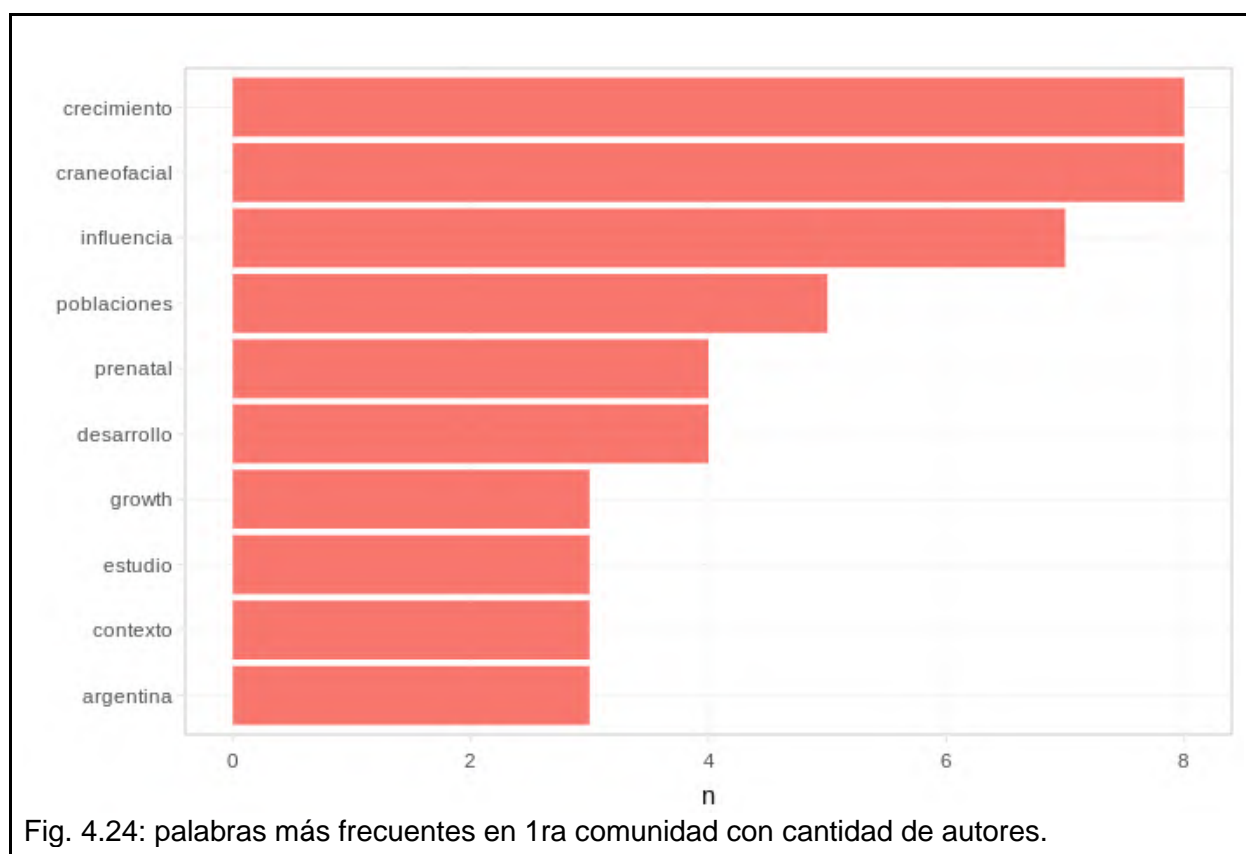
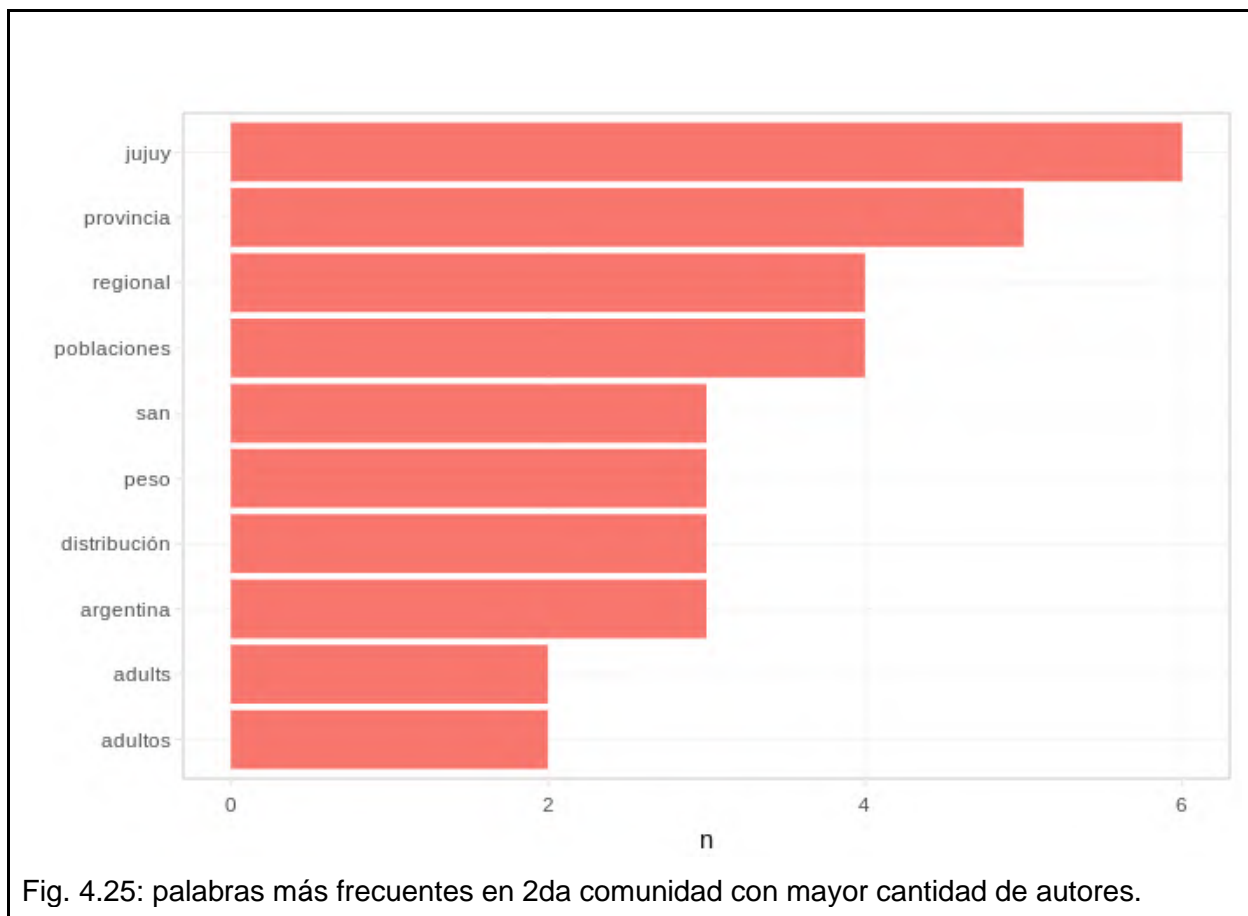
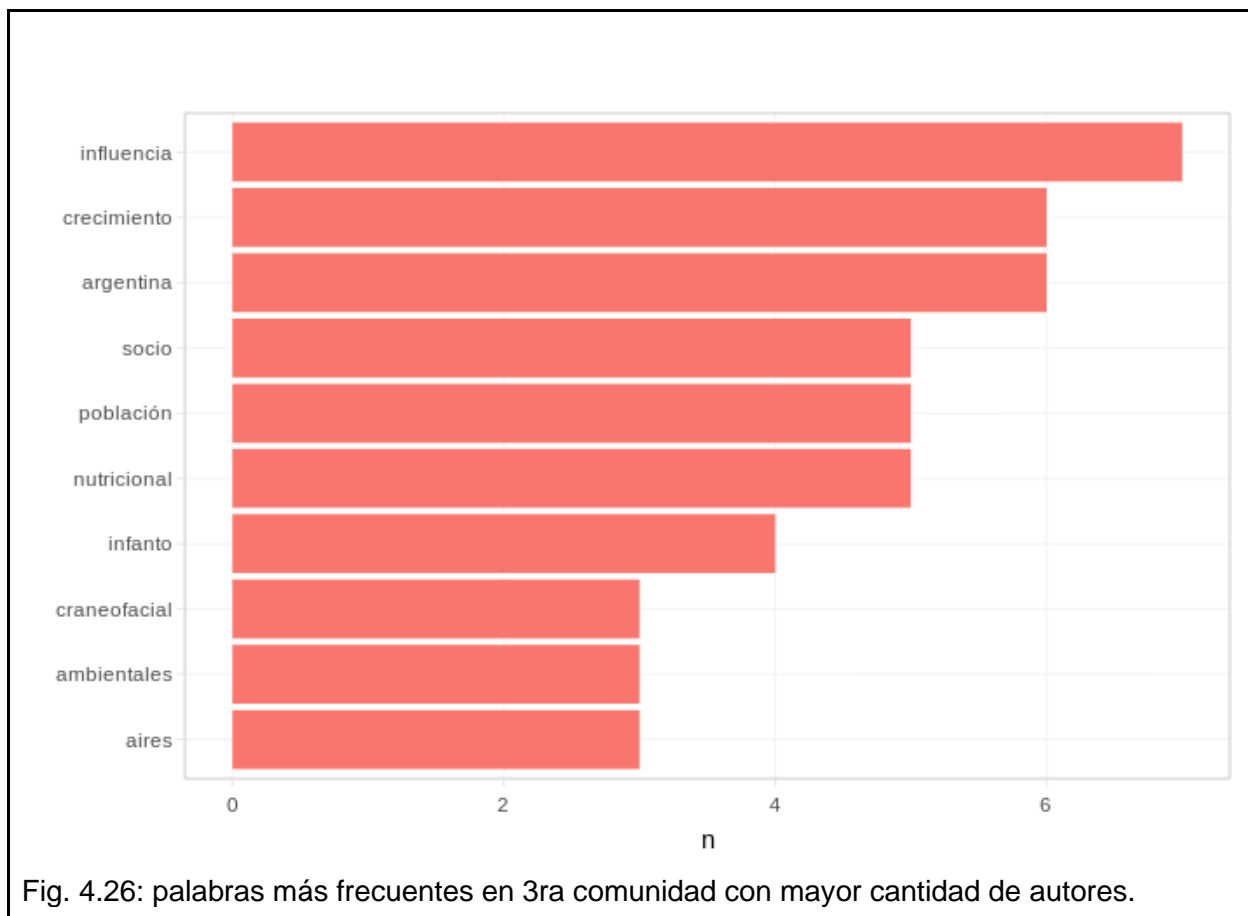


Fig. 4.24: palabras más frecuentes en 1ra comunidad con cantidad de autores.

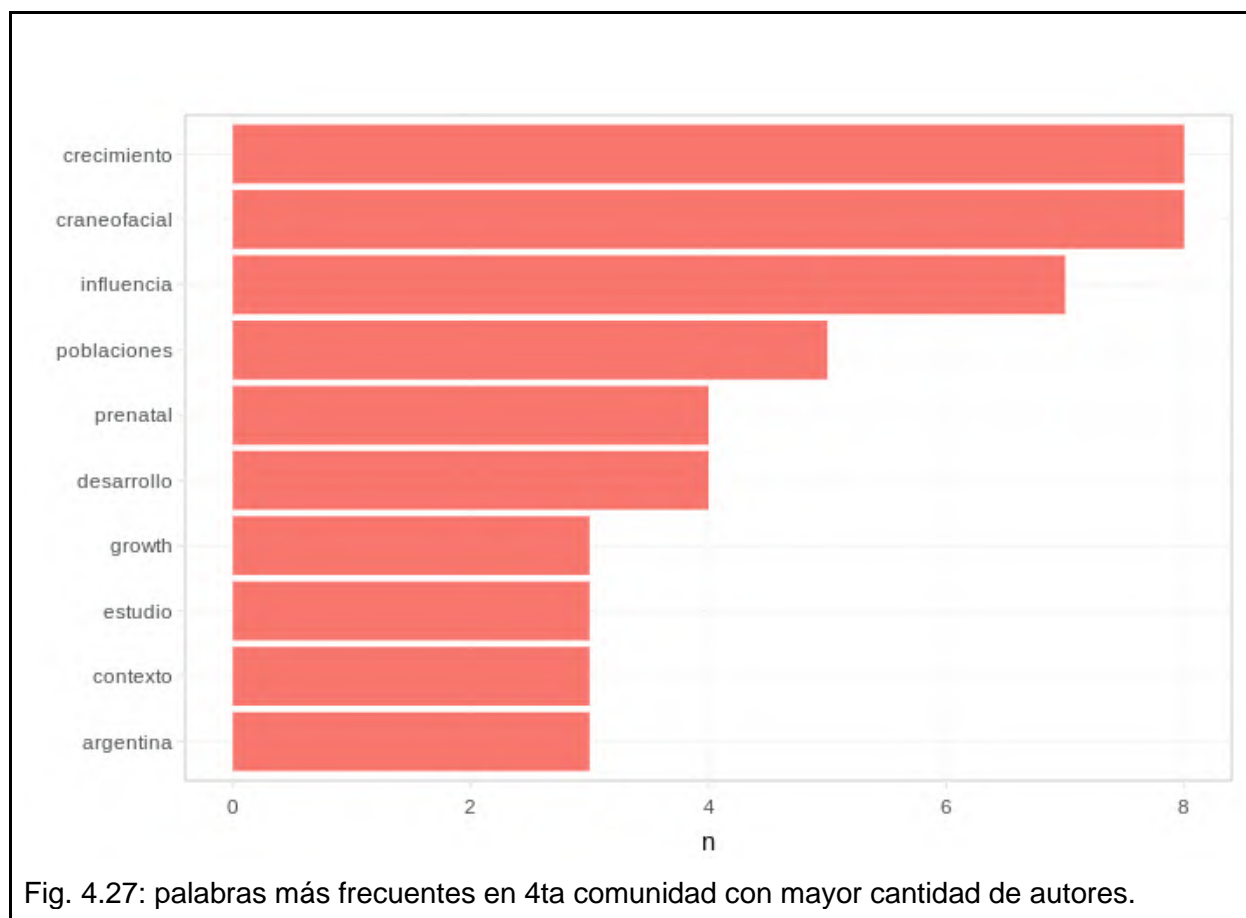
En la comunidad 1 ([Fig. 4.24](#)) se destaca como principal las palabras “craneofacial”, “crecimiento” e “influencia”, y llama la atención que aparecen palabras en inglés. Esto último puede deberse a que algunos artículos en la revista poseen sus títulos publicados tanto en castellano como inglés, cosa que se repite en tres de las cinco comunidades evaluadas. Respecto de una de las palabras que más se destacan, “influencia”, resultó difícil comprender la razón. Sin embargo, al leer los títulos de los artículos se observó que muchos de ellos arrancan con esa palabra (como ‘*influencia de X factor en crecimiento/nutrición de Y población*’), lo cual podría explicar su relativamente alta frecuencia.



En la comunidad 2 ([Fig. 4.25](#)), los nodos parecen estar vinculados por un componente de localidad, asociado a la provincia de Jujuy, y por estudios poblacionales sobre dicha provincia.



En la comunidad 3 se resaltan las palabras “influencia”, “argentina”, “crecimiento”, “nutricional” y “población” ([Fig. 4.26](#)). Es posible que esta comunidad esté relacionada con el tema análisis de poblaciones en Argentina considerando diversos factores que afectan el crecimiento.



En la comunidad 4 se observan tanto palabras en inglés como castellano y, nuevamente, se resalta la temática “craneofacial”, “crecimiento” y la palabra “influencia” ([Fig. 4.27](#)). Al leer algunos títulos se puede ver que los temas desarrollados por esta comunidad se asocian a cómo repercuten las dietas en el desarrollo humano. Cabe destacar que los temas asociados a las comunidades 3 y 4, según el análisis de las 10 palabras más frecuentes en los títulos, poseen similitudes por lo que queda para futuro realizar otro tipo de análisis de texto que permita profundizar e identificar mejor las líneas de investigación posiblemente asociadas a las comunidades.

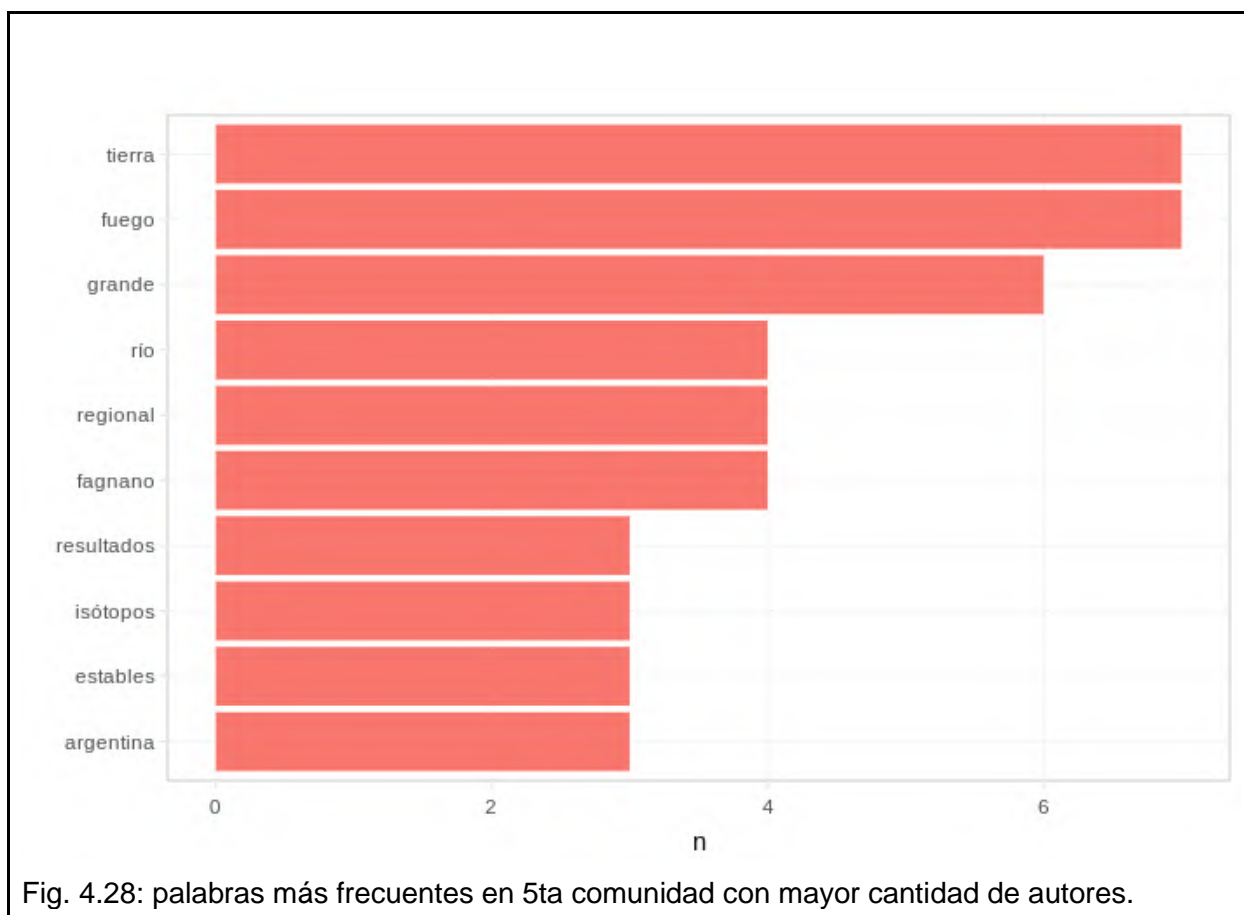


Fig. 4.28: palabras más frecuentes en 5ta comunidad con mayor cantidad de autores.

Finalmente, la comunidad 5 parece estar afectada principalmente por una componente de localidad asociada a la provincia Tierra del Fuego y, más específicamente, la localidad de Río Grande ([Fig. 4.28](#)). Se puede observar que la frecuencia de palabras de forma simple no resultó suficiente para este análisis, que se deben analizar compuestos de palabras, como ser “tierra del fuego” en su completitud en vez de cada palabra por separado.

Todo este breve análisis realizado sobre las comunidades halladas nos permitió confirmar que, a futuro, sería necesario implementar análisis de texto más complejos y detallados para poder identificar correctamente las líneas o actividades de investigación asociadas a cada comunidad. Sin embargo, a pesar de las limitaciones del análisis realizado, en muchos casos hemos podido asociar a las comunidades con diferentes temas de investigación y regiones geográficas del país.

Una vez concluido este análisis de la red “hacia adentro”, nos preguntamos cómo se comporta “hacia afuera”, es decir, al compararla con otras redes. Para lograr esta comparación en esta tesis realizamos dos enfoques: a través de modelos para verificar si la red tiene ciertas propiedades que se observan comúnmente en redes del mundo real, y comparando con otras redes de coautoría, de estudios previos.

Hipótesis 4: La red agregada total (desde 1996 a 2016) es de mundo pequeño puesto que es una característica presente en la mayoría de las redes de coautorías científicas

En los análisis de redes es de utilidad realizar comparaciones con modelos. En palabras de [Kolaczyk-Csárdi \(2014\)](#):

En la práctica, los modelos de redes de grafos son usados para diferentes propósitos. Estos incluyen: 1) la prueba de “significancia” de ciertas características predefinidas en una determinada red, 2) el estudio de mecanismos propuestos para generar ciertas propiedades que se observan comúnmente en redes del mundo real (como distribuciones de grado amplias o efectos de mundo pequeño) o 3) la evaluación de potenciales factores predictivos de relaciones.

En este caso nos vamos a centrar en el punto 2, el estudio de mecanismos propuestos para generar ciertas propiedades que se observan comúnmente en redes del mundo real (como distribuciones de grado amplias o efectos de mundo pequeño).

Mundo pequeño

El estudio de “mundo pequeño” está asociado a evaluar ciertas propiedades de la red que permiten que se pueda afirmar que las personas o nodos se conectan con otras a través de una cantidad pequeña de saltos, más allá de que se hallen muy separadas entre sí en la red.

Para comprender un poco mejor el estudio de “Small world”, surge de la frase en inglés ‘*My it’s a small world*’ / ‘*Que mundo pequeño*’ cuando dos personas conocen a otra y, por lo tanto, estarían interconectados a través de una tercera. Este estudio de [Milgram \(1967\)](#) consistió en enviar mensajes desde una persona en un punto A hasta otra en un punto B bajo ciertas reglas, entre ellas, que solo podían mandar el mensaje a personas conocidas inmediatas.

Con el pasar del tiempo se siguió analizando esta característica de las redes, en diferentes estudios, llegando a medidas para evaluar esta propiedad. Las propiedades clave son un alto coeficiente de *clustering*, también llamado transitividad, y un promedio bajo de camino más corto.

El coeficiente de *clustering* de un nodo mide hasta qué punto los vecinos de un nodo también están interconectados, mientras que el coeficiente de clustering de la red es el promedio del coeficiente de clustering de todos los nodos.

El promedio de camino más corto entre nodos, mide el camino más corto de un nodo A a un nodo B. Para esta métrica también se trabaja con el promedio para todos los nodos en la red.

Para poder evaluar estas propiedades usualmente se realizan comparaciones con otras redes de características similares. Como al momento de surgimiento de esta tesis fue difícil encontrar redes reales con dichas características, se procedió a realizar simulaciones de modelos aleatorios para poder comparar sus propiedades con las de la red analizada y así analizar si cumple o no con las características de “mundo pequeño”.

Libre Escala

En forma complementaria evaluamos la propiedad de libre escala, que se basa en analizar la distribución de grado de la red. Recordemos que el grado de cada nodo representa con cuántos otros nodos se conecta directamente el nodo analizado. Una red Libre escala es una red en la cual su distribución de grado sigue la ley de potencia ([Barabási 2016](#)), lo cual significa que la red tiene muchos nodos con pocas conexiones y pocos nodos con muchas conexiones.

Caso aplicado

En primer lugar, analizamos la red en su totalidad, de 1996 a 2016, para contrastar la hipótesis general y luego por periodos, para verificar si este comportamiento varió en el tiempo. Recordemos que los 20 años de la revista están repartidos en 18 años de publicaciones dado que en 1997, 1998 y 2000 no hubo ediciones.

Mundo pequeño para la red agregada total

La [Fig. 4.29](#) muestra los parámetros utilizados para realizar las simulaciones usando el modelo Erdos-Renyi para el caso de mundo pequeño de la red total entre 1996 y 2016.

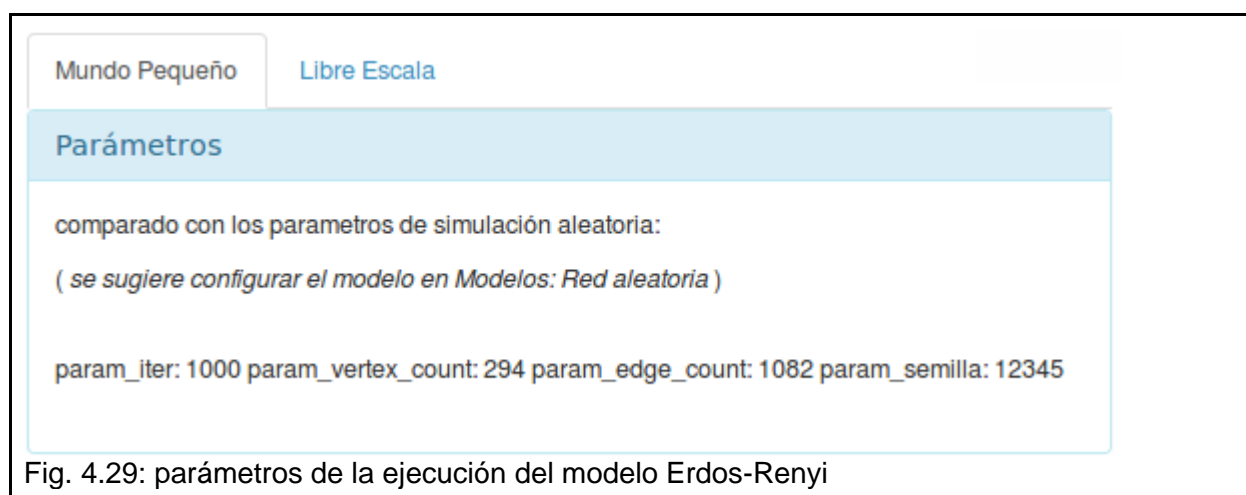
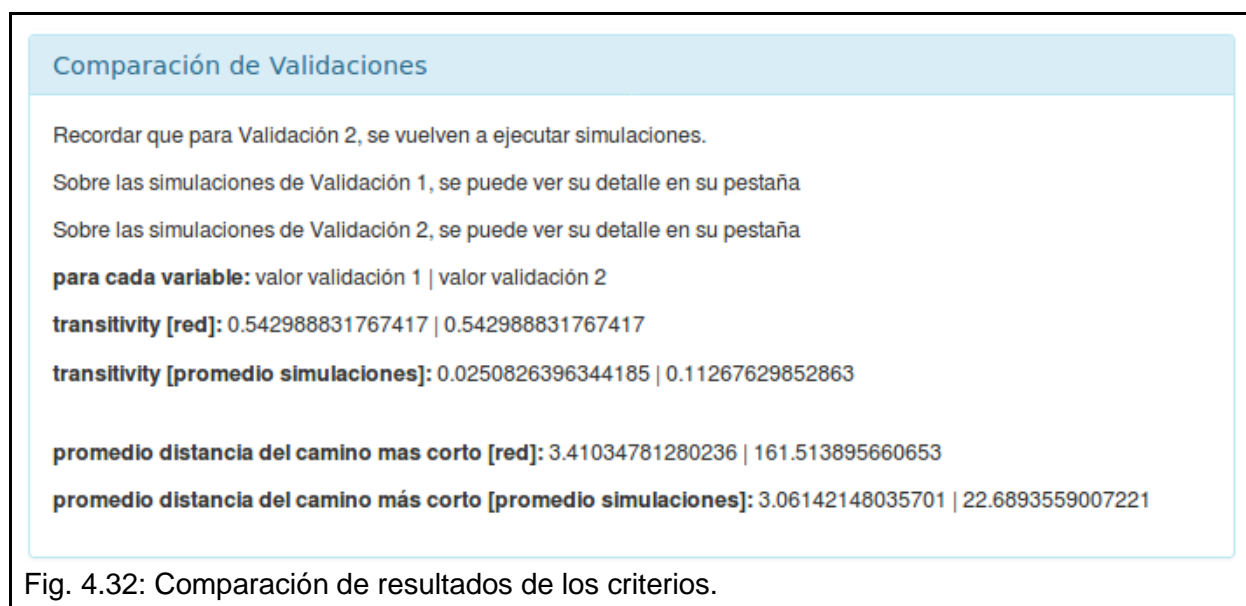
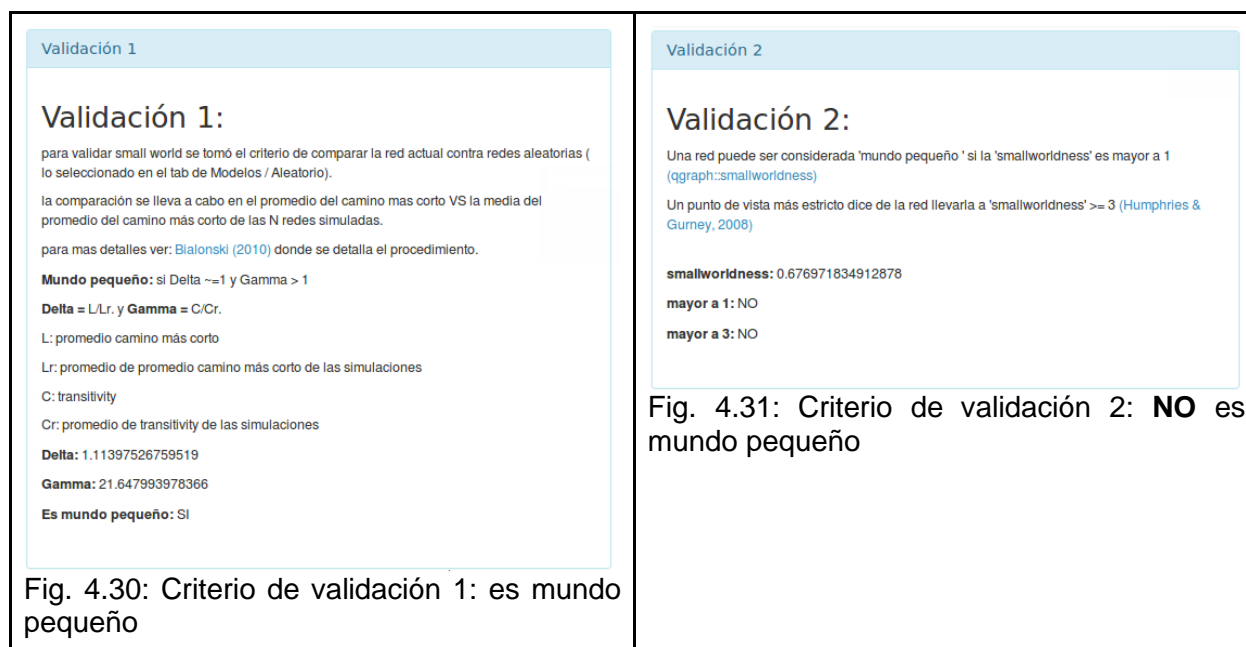


Fig. 4.29: parámetros de la ejecución del modelo Erdos-Renyi

La red total de los periodos desde 1996 hasta 2016, según el criterio de validación expuesto por [Bialonski et al. \(2010\)](#) es de mundo pequeño ([Fig. 4.30](#)). Sin embargo, para el criterio de validación utilizando la métrica *smallworldness*, no lo es ([Fig. 4.31](#)), tanto para una cota laxa de 1, como para una cota estricta de 3, según lo planteado por [Humphries-Gurney \(2008\)](#). Viendo los resultados comparados ([Fig. 92](#)), vemos que la generación de las simulaciones a través de la función *smallworldness* varía considerablemente con lo obtenido en las simulaciones creadas bajo el criterio de validación de [Bialonski et al. \(2010\)](#).



Esta diferencia de valores que aparece en la comparación de métodos, está asociada al método de generación de las redes aleatorias. En el criterio 1 ([Fig. 4.30](#)) se usa el método de Erdos-Renyi implementado en el método `sample_gnm` de la librería [igraph](#) que se basa en la publicación de [Erdős-Renyi \(1959\)](#), mientras que en el criterio 2 la función de la librería [qgraph](#) se utiliza la función para generar los aleatorios de la librería [igraph](#) `degree.sequence.game` pasando el método de generación `"simple.no.multiple"`.

Los parámetros para el modelo de Erdos-Renyi se pueden ver en la [Fig. 4.29](#). Los parámetros de configuración de la ejecución de `degree.sequence.game` son diferentes, el único configurado es la cantidad de iteraciones: se usó la misma cantidad que para el modelo configurado, en este caso 1000. El modelo de Erdos-Renyi se basa en la cantidad de nodos y de aristas del grafo

original, y se hacen 1000 iteraciones aleatorias para obtener los resultados. Por otro lado, *degree.sequence.game* se basa en generar un grafo aleatorio dada una lista de grados tomada del grafo original, en la implementación de la función *smallworldness*. Es importante aclarar que en la documentación de la función *degree.sequence.game*, se recomienda ver otras funciones por problemas de bucles que se generan al ejecutarla y, entre ellas, recomienda el modelo de Erdos-Renyi basado en probabilidades. Para más información se recomienda ver la documentación de la librería [igraph](#).

Como conclusión, podemos decir que la red de coautorías de la RAAB es de mundo pequeño, aunque hemos visto que depende de los diferentes enfoques utilizados para analizarla.

Libre escala para la red total

En los análisis realizados se observa claramente que la red de coautorías de la RAAB posee muchos nodos con pocas conexiones y pocos nodos con muchas conexiones ([Fig. 4.33](#) y [Fig. 4.34](#)). Confirmamos entonces que se da el fenómeno de libre escala.

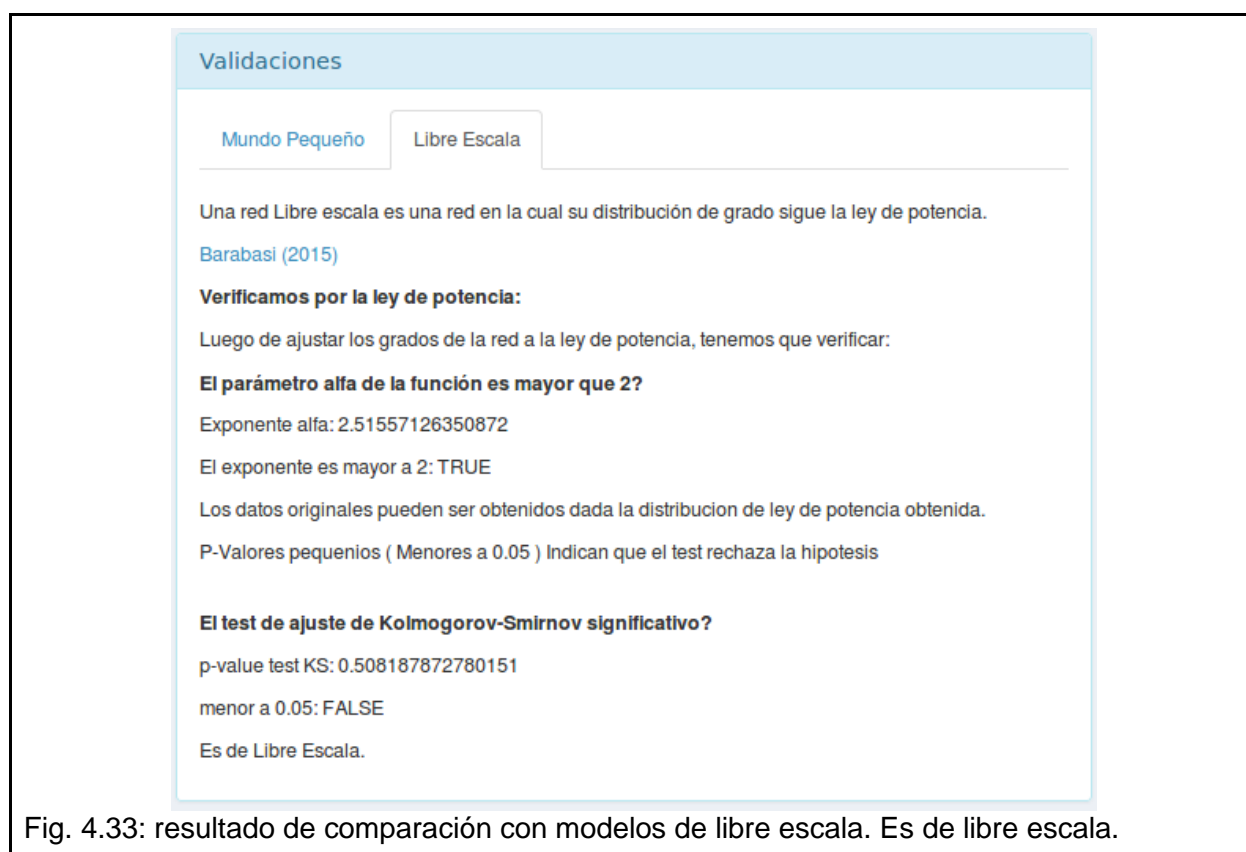
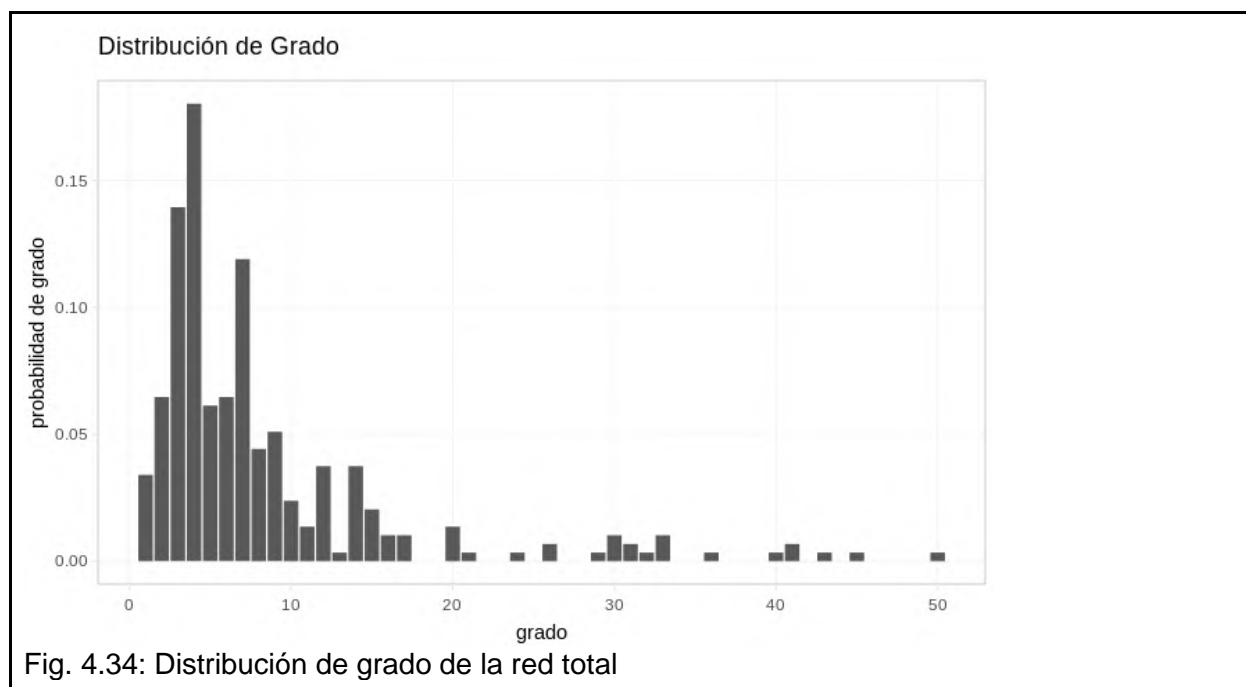


Fig. 4.33: resultado de comparación con modelos de libre escala. Es de libre escala.



Luego de realizar los análisis para la red total (período completo), evaluamos cómo se comportaba la red en periodos inferiores. Para ello, se utilizó la división especificada en la [Tabla 4.10](#). Recordemos que los 20 años de la revista están repartidos en 18 años de publicaciones dado que en 1997, 1998 y 2000 no hubo ediciones.

Tabla 4.10: División de años de la revista para el análisis.

Período	desde	hasta	incluye
1	1996	2001	1996,1999,2001
2	2002	2006	2002,2003,2004,2005,2006
3	2007	2011	2007,2008,2009,2010,2011
4	2012	2016	2012,2013,2014,2015,2016

Mundo pequeño para el periodo 1996 - 2001

Podemos ver que para el periodo 1996-2001, La [Fig. 4.35](#) muestra los parámetros utilizados para realizar las simulaciones usando el modelo modelo Erdos-Renyi para el caso de mundo pequeño. ([Fig. 4.35](#))

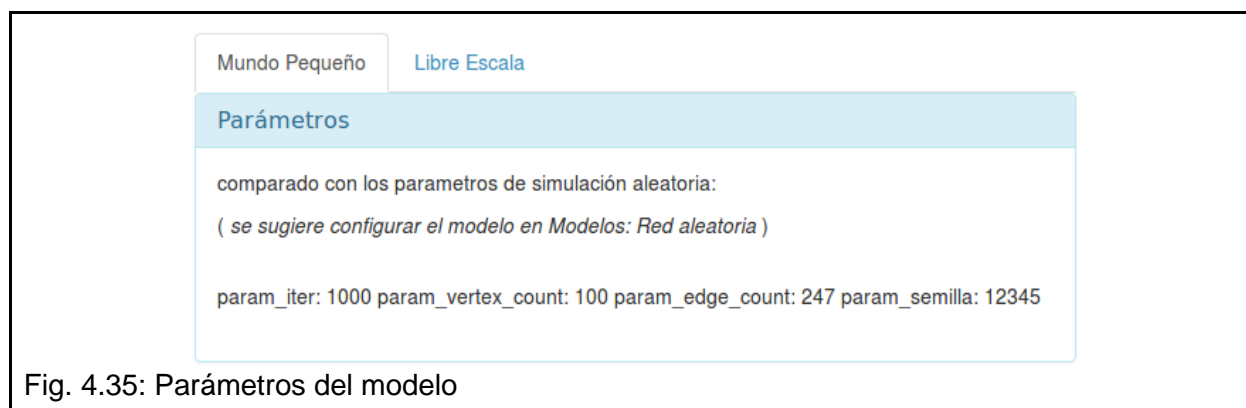


Fig. 4.35: Parámetros del modelo

Los resultados determinan que la red es de mundo pequeño según el criterio 1, sin embargo para el criterio 2 no, considerando tanto el umbral laxo como el estricto (Fig. 4.36 y Fig. 4.37).

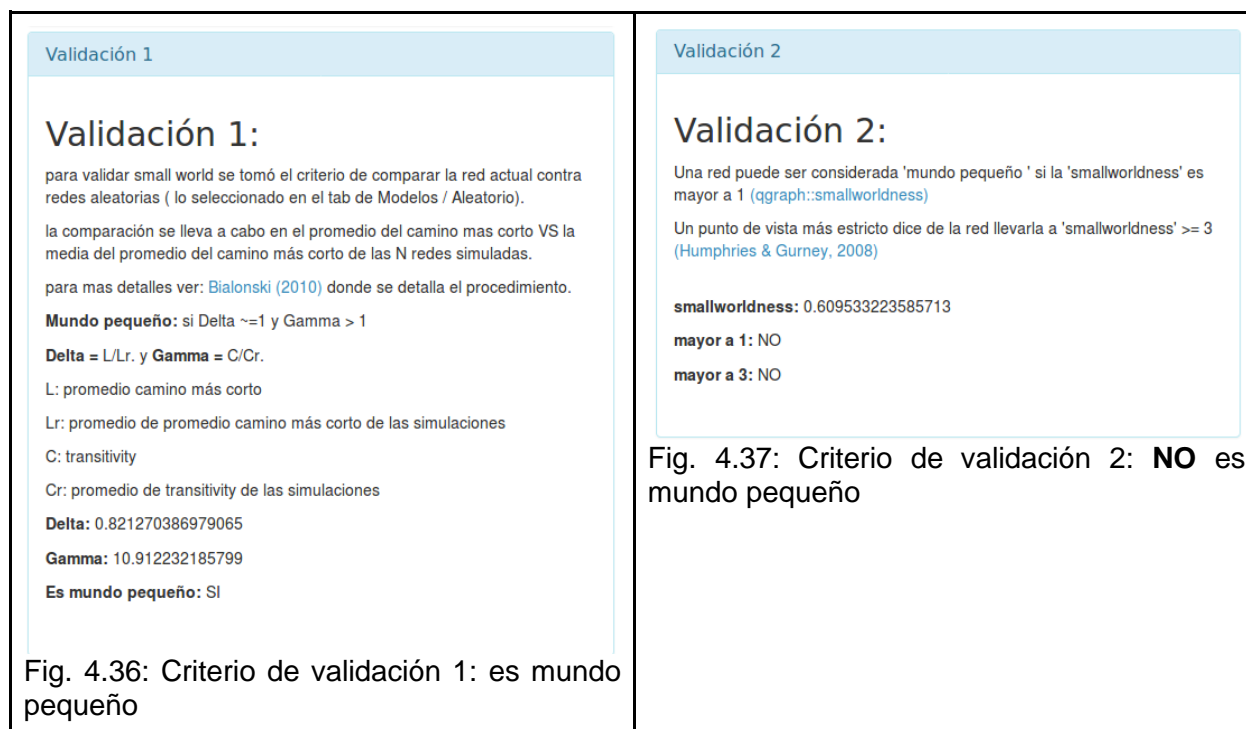


Fig. 4.36: Criterio de validación 1: es mundo pequeño

Fig. 4.37: Criterio de validación 2: **NO** es mundo pequeño

Como se mencionó anteriormente, puede evaluarse la diferencia entre ambos enfoques de validación de mundo pequeño (Fig. 98).

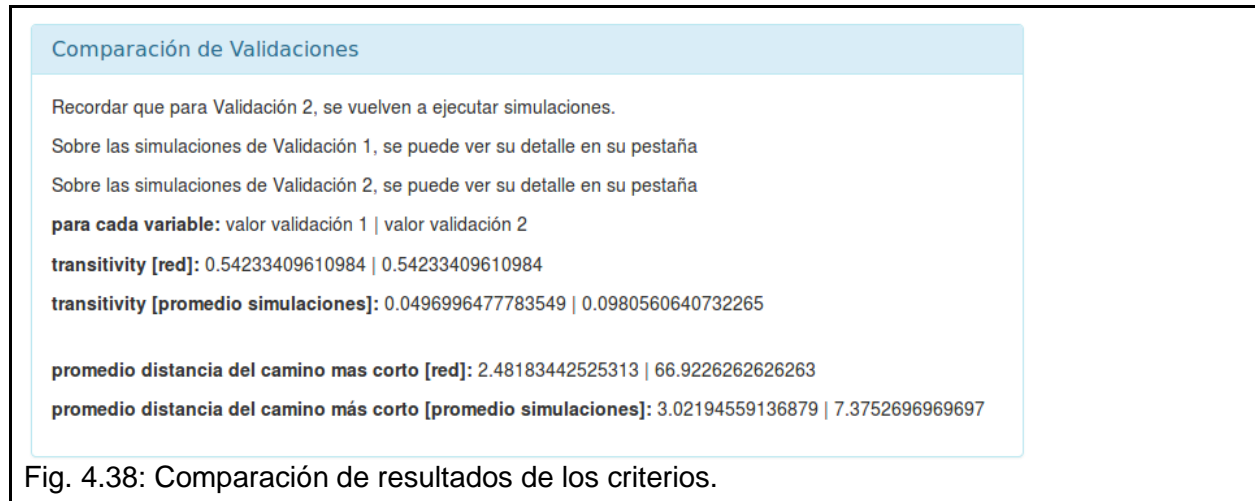


Fig. 4.38: Comparación de resultados de los criterios.

Libre escala 1996 - 2001

Podemos ver que para el periodo 1996-2001, la red es de libre escala ([Fig. 4.39](#)). Además, vemos que hay mayor probabilidad en grados menores a 10 y menor en grados mayores a 10 ([Fig. 4.40](#)).

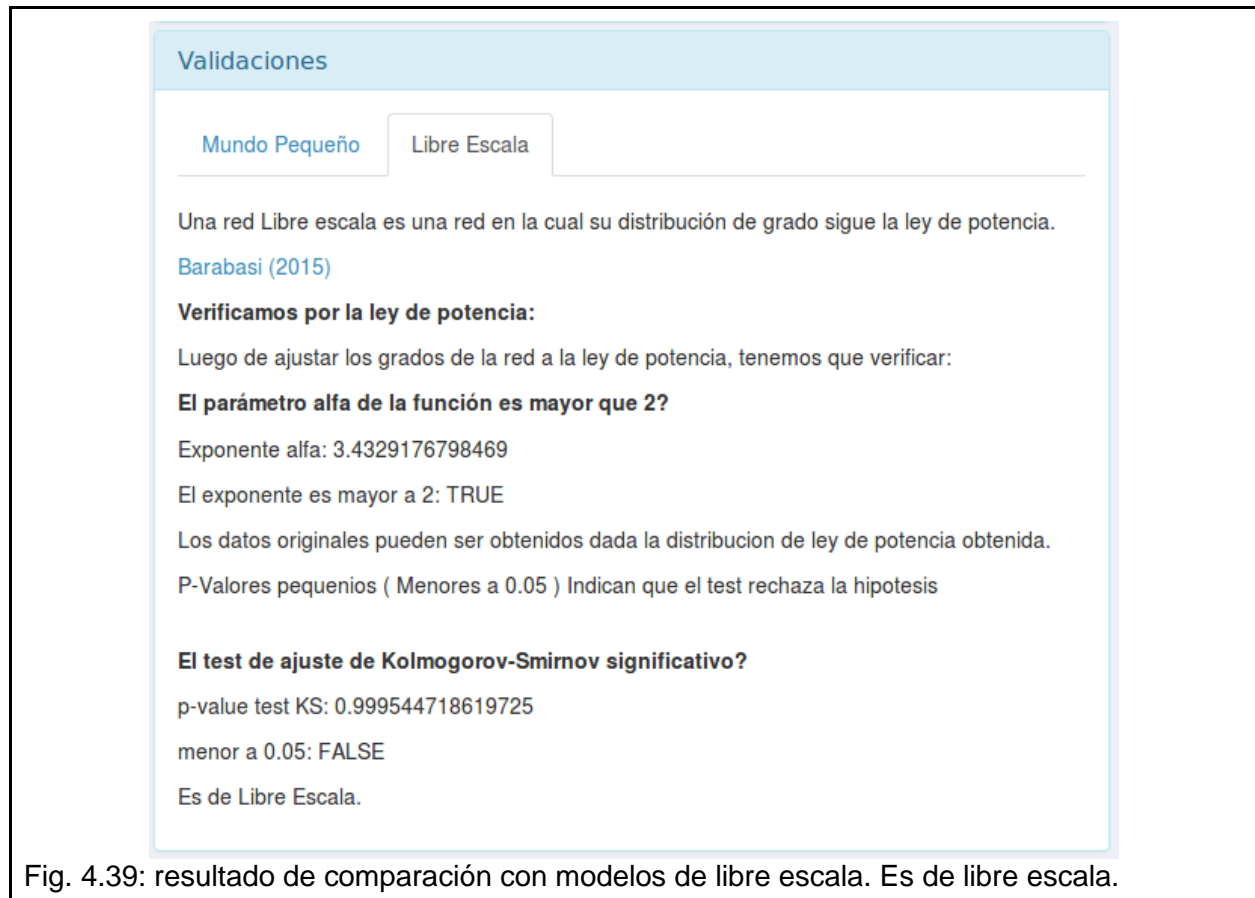
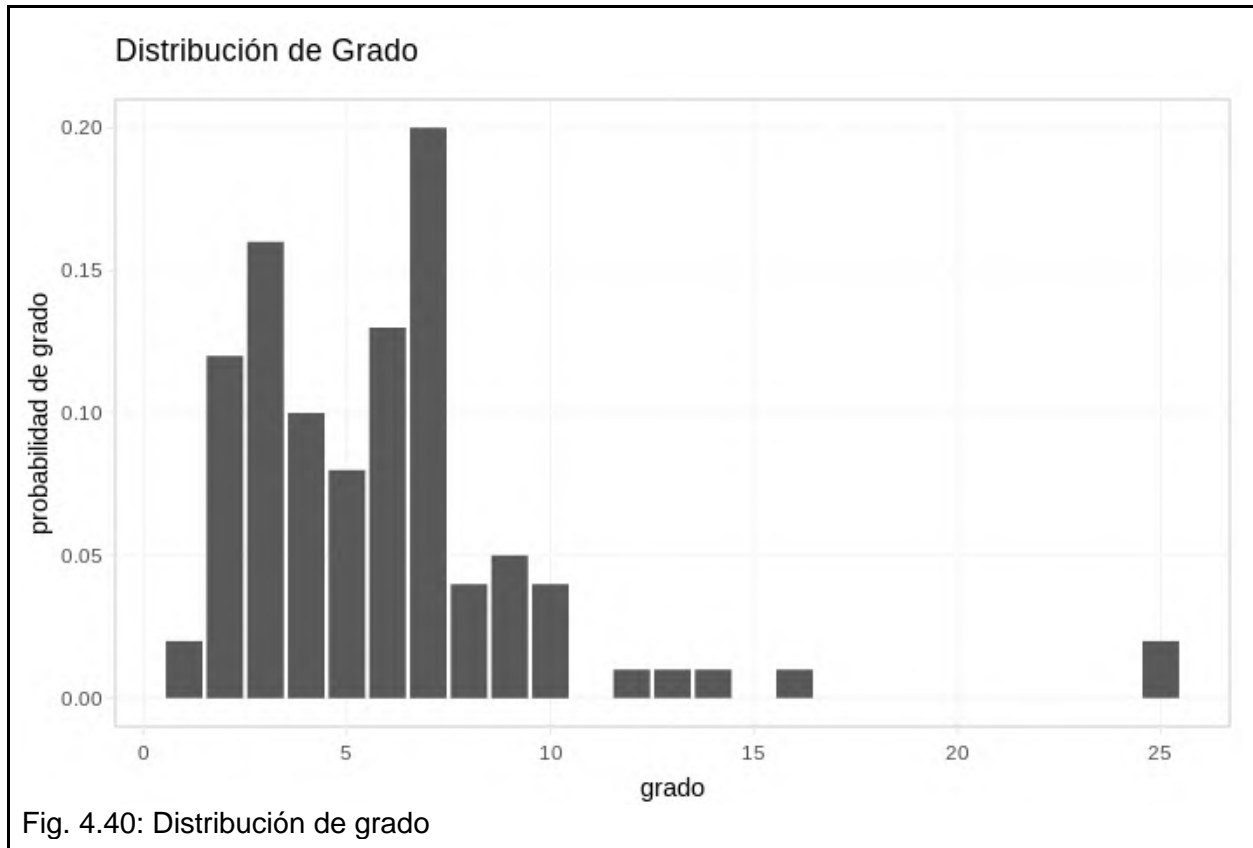
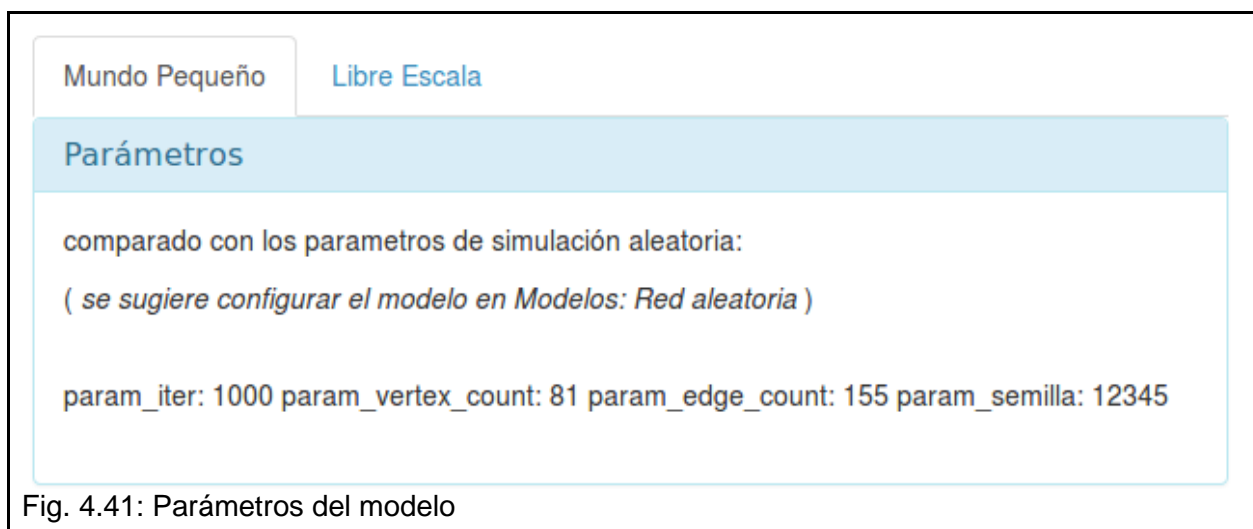


Fig. 4.39: resultado de comparación con modelos de libre escala. Es de libre escala.

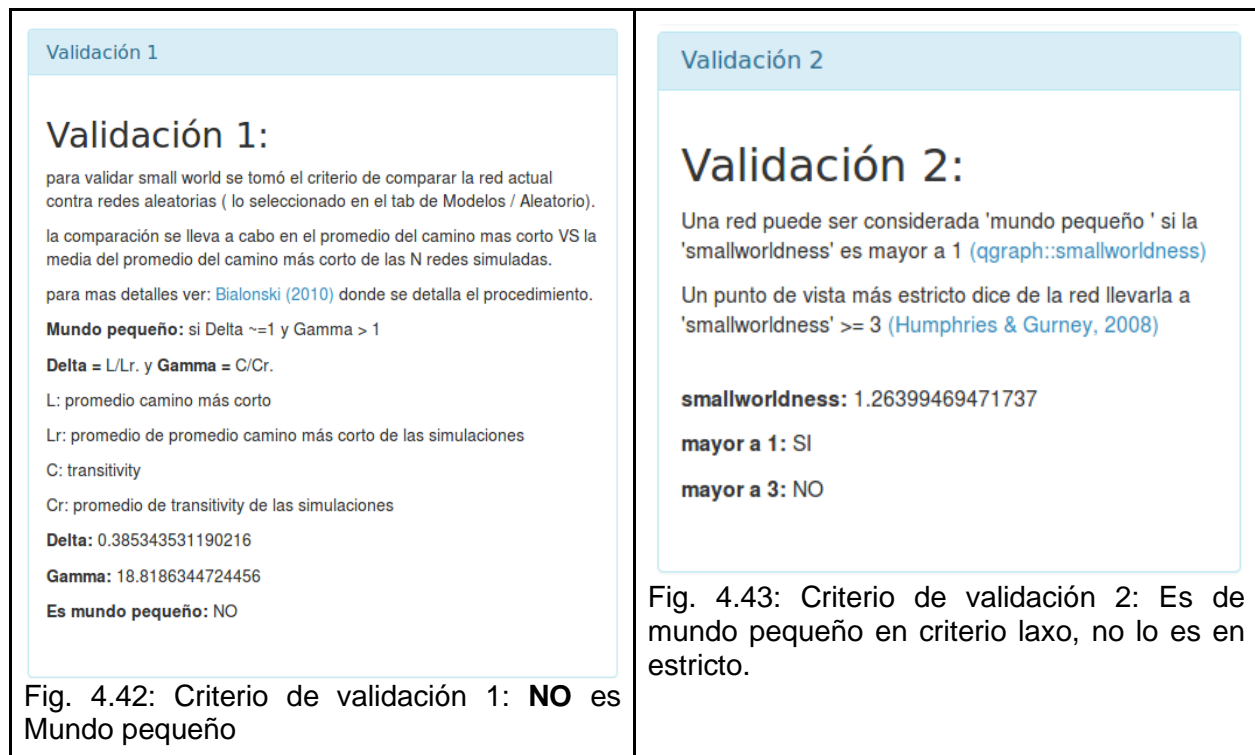


Mundo pequeño para el periodo 2002 - 2006

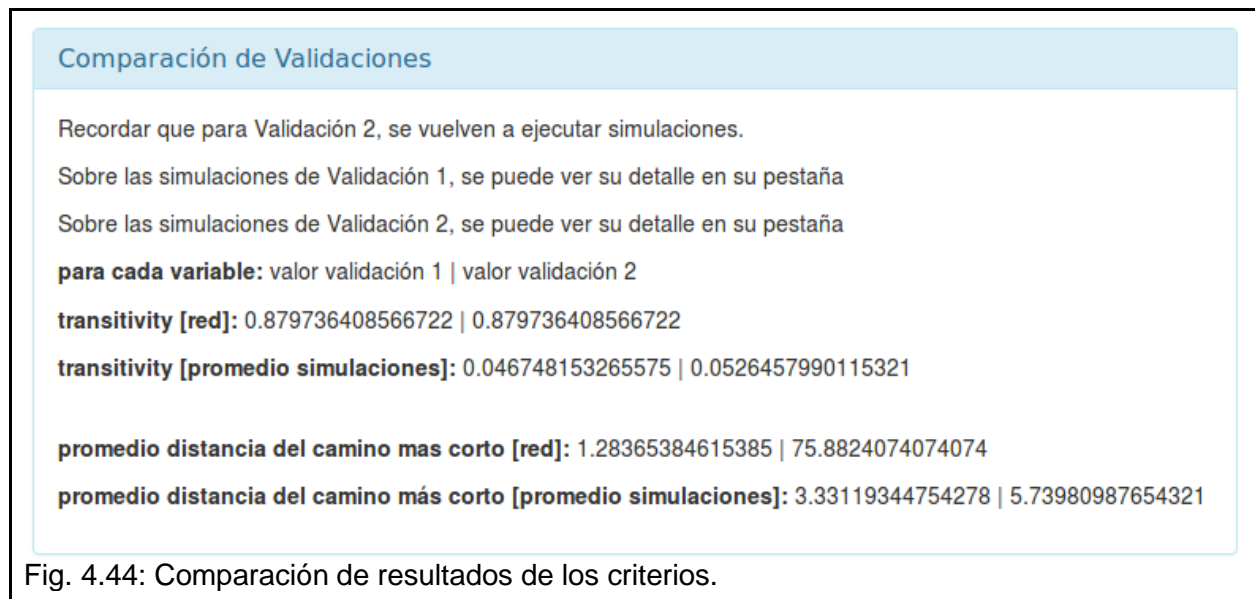
Los parámetros utilizados para esta ejecución se pueden verificar en la [Fig. 4.41](#).



Podemos ver que para el periodo 2002-2006, la red no es de mundo pequeño según el criterio 1 ([Fig. 4.42](#)). Sin embargo, para el criterio 2 sí cumple con esta característica bajo el criterio laxo pero no bajo el estricto ([Fig. 4.43](#)).



La diferencia entre ambos enfoques de validación de mundo pequeño se observa en la [Fig. 4.44](#).



Libre escala 2002 - 2006

Podemos ver que para el periodo 2002-2006, la red es de libre escala ([Fig. 4.45](#)), hay mayor probabilidad de grados bajos y bajas probabilidades de grados elevados. Sin embargo, en la distribución vemos que el grado 7 se destaca bastante ([Fig. 4.46](#)).

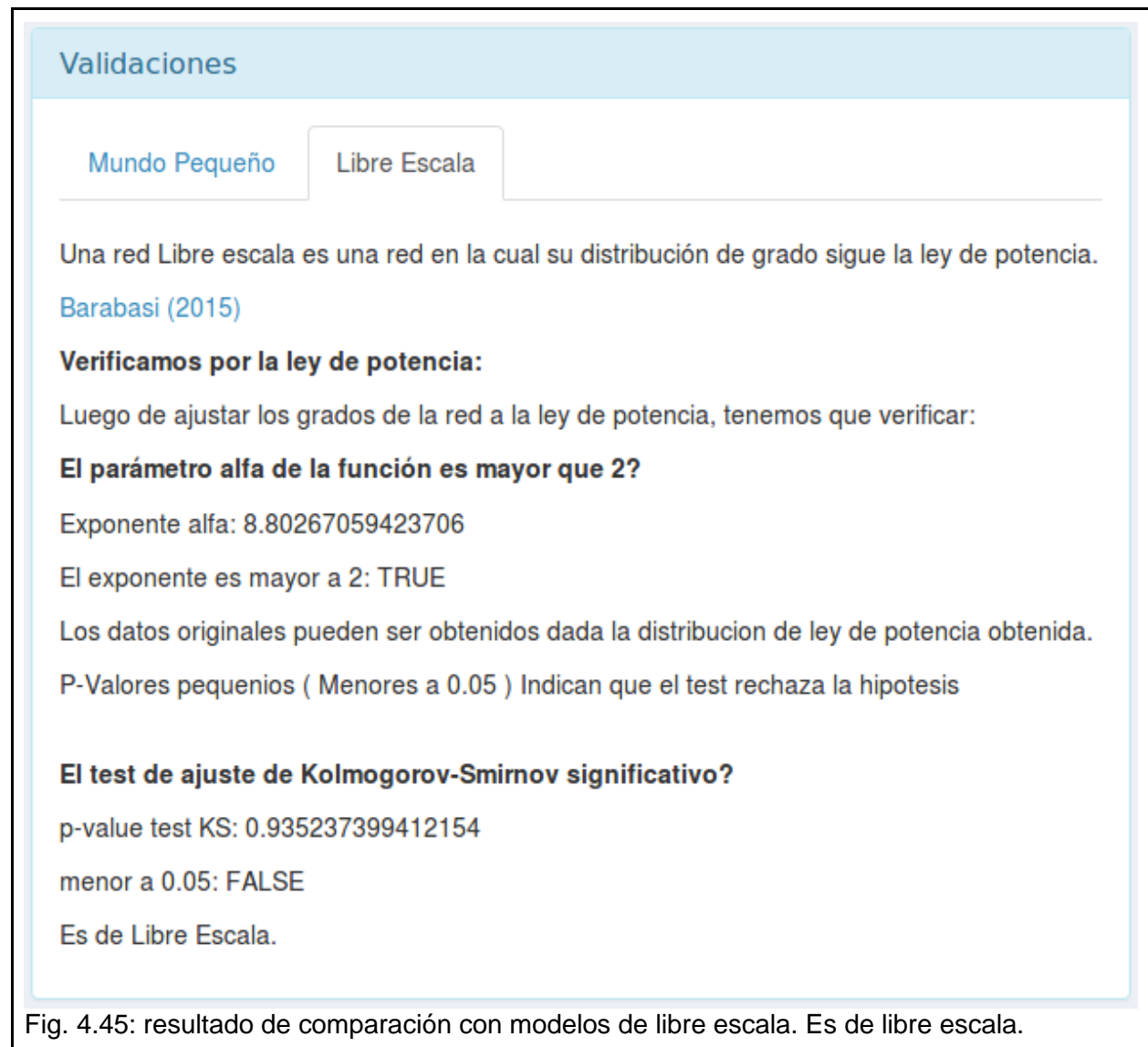
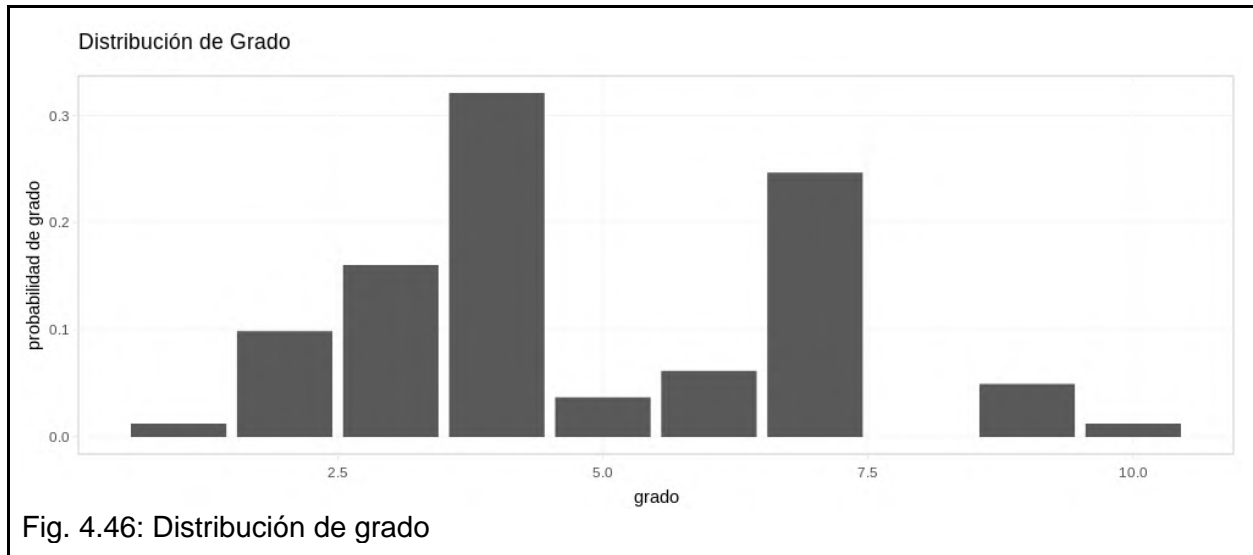
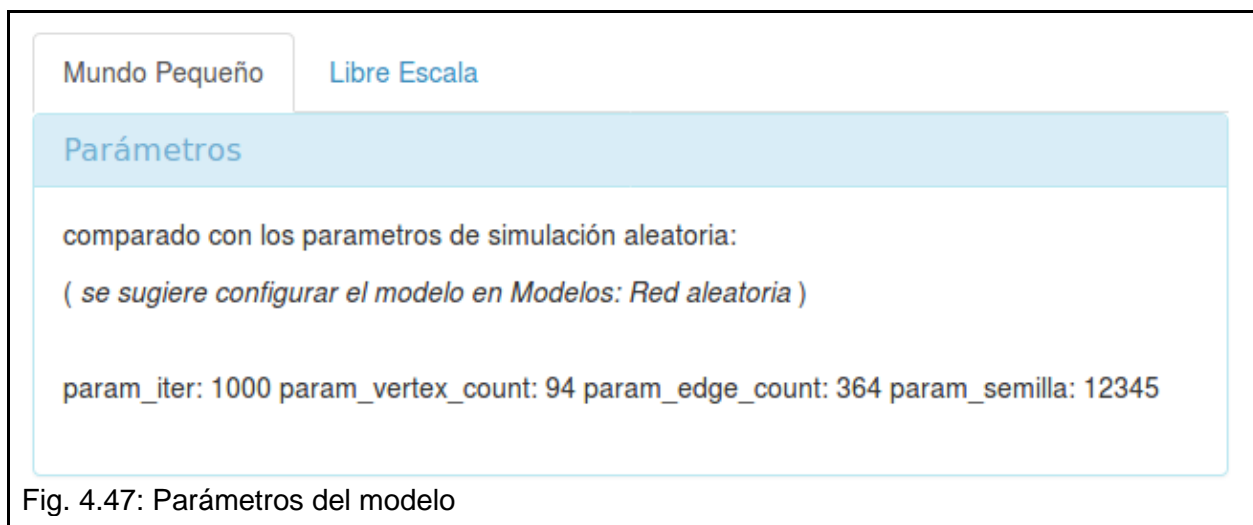


Fig. 4.45: resultado de comparación con modelos de libre escala. Es de libre escala.

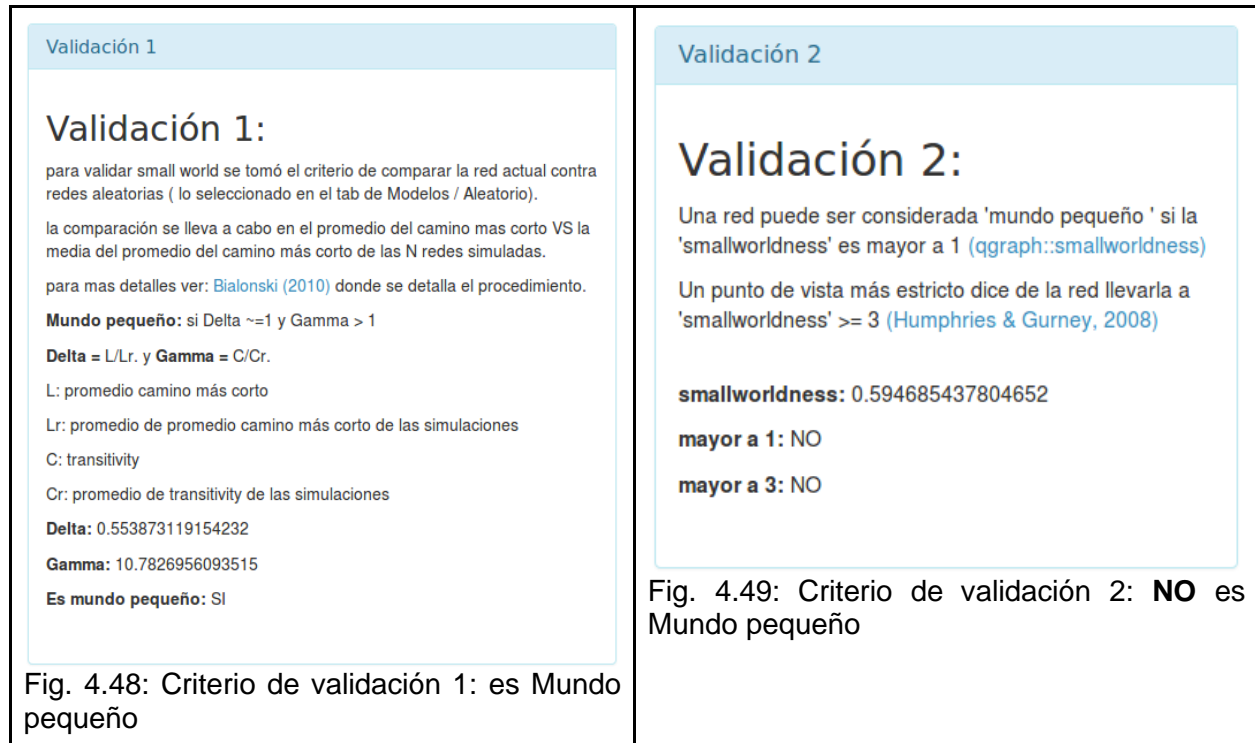


Mundo pequeño para el periodo 2007 - 2011

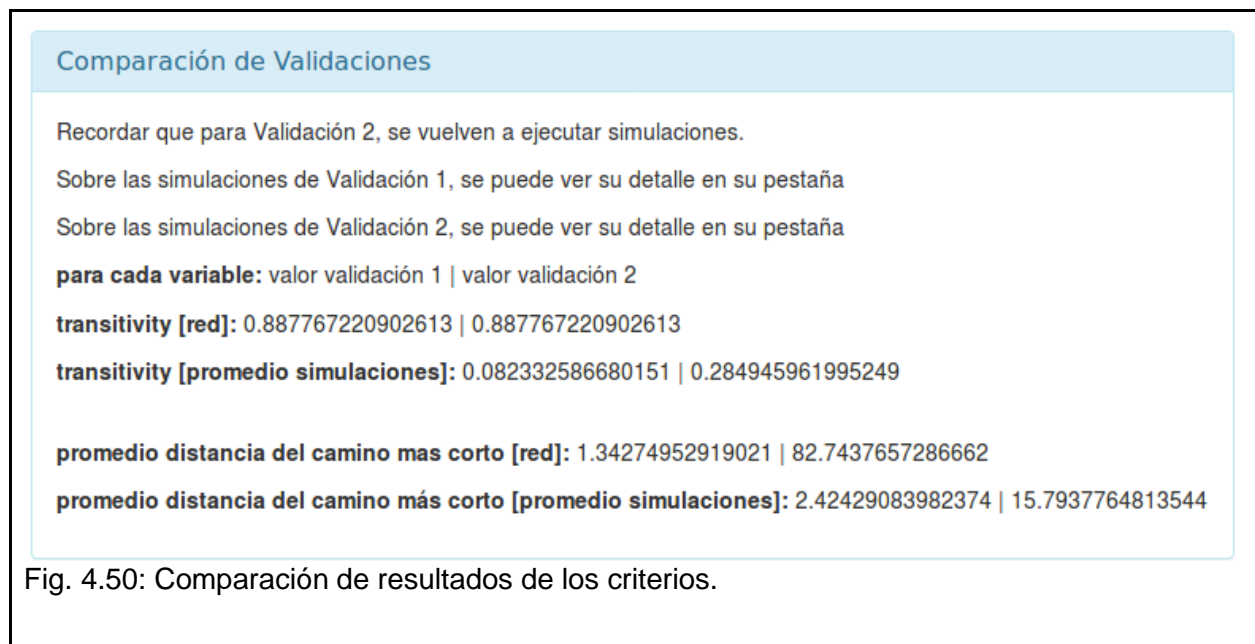
Los parámetros utilizados para esta ejecución se pueden verificar en la [Fig. 4.47](#).



Podemos ver que para el periodo 2007-2011, la red es de mundo pequeño para el criterio 1, sin embargo para el criterio 2 no, considerando tanto el umbral laxo como el estricto ([Fig. 4.48](#) y [Fig. 4.49](#)).

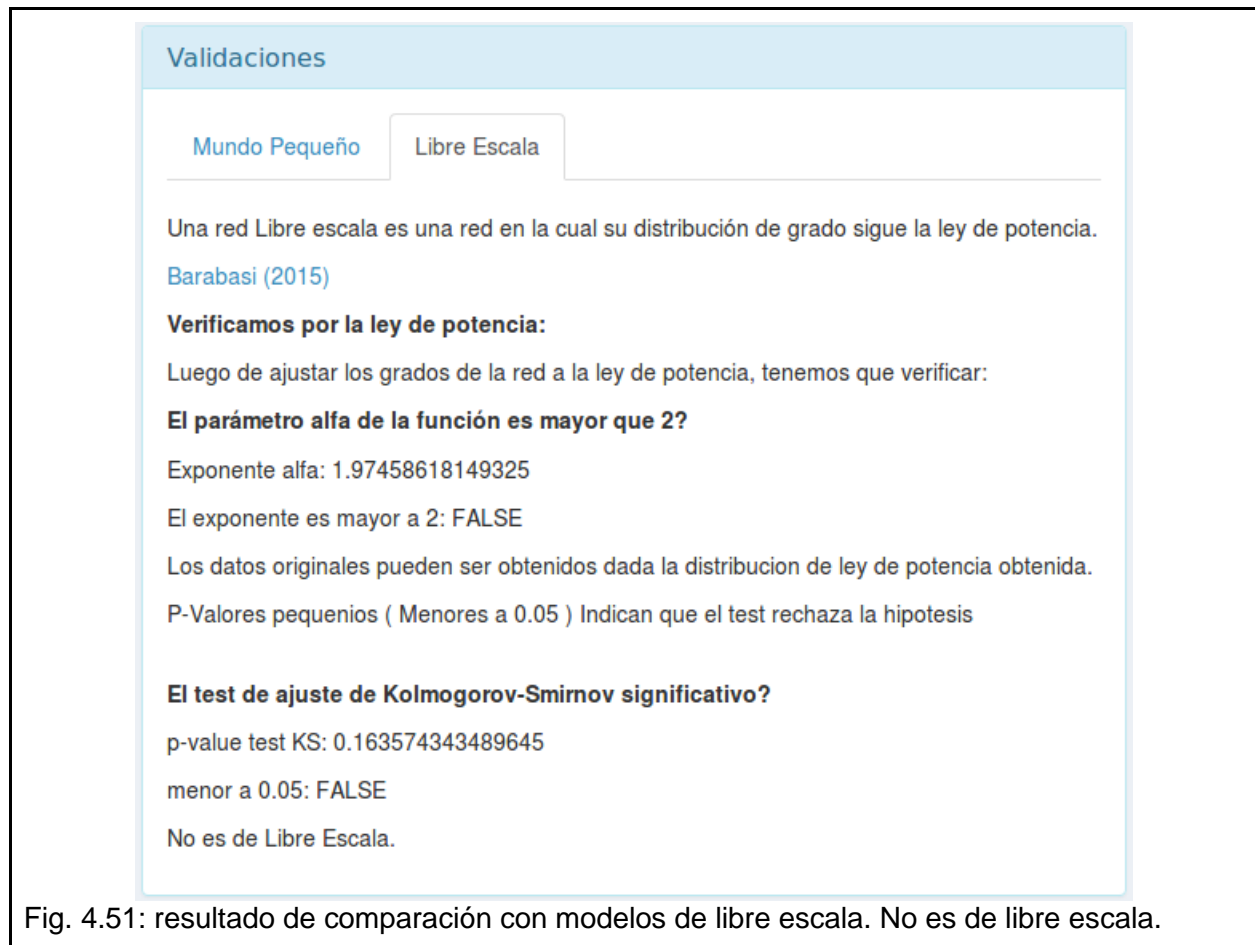


La diferencia entre ambos enfoques de validación de mundo pequeño se observa en la [Fig. 4.50](#).

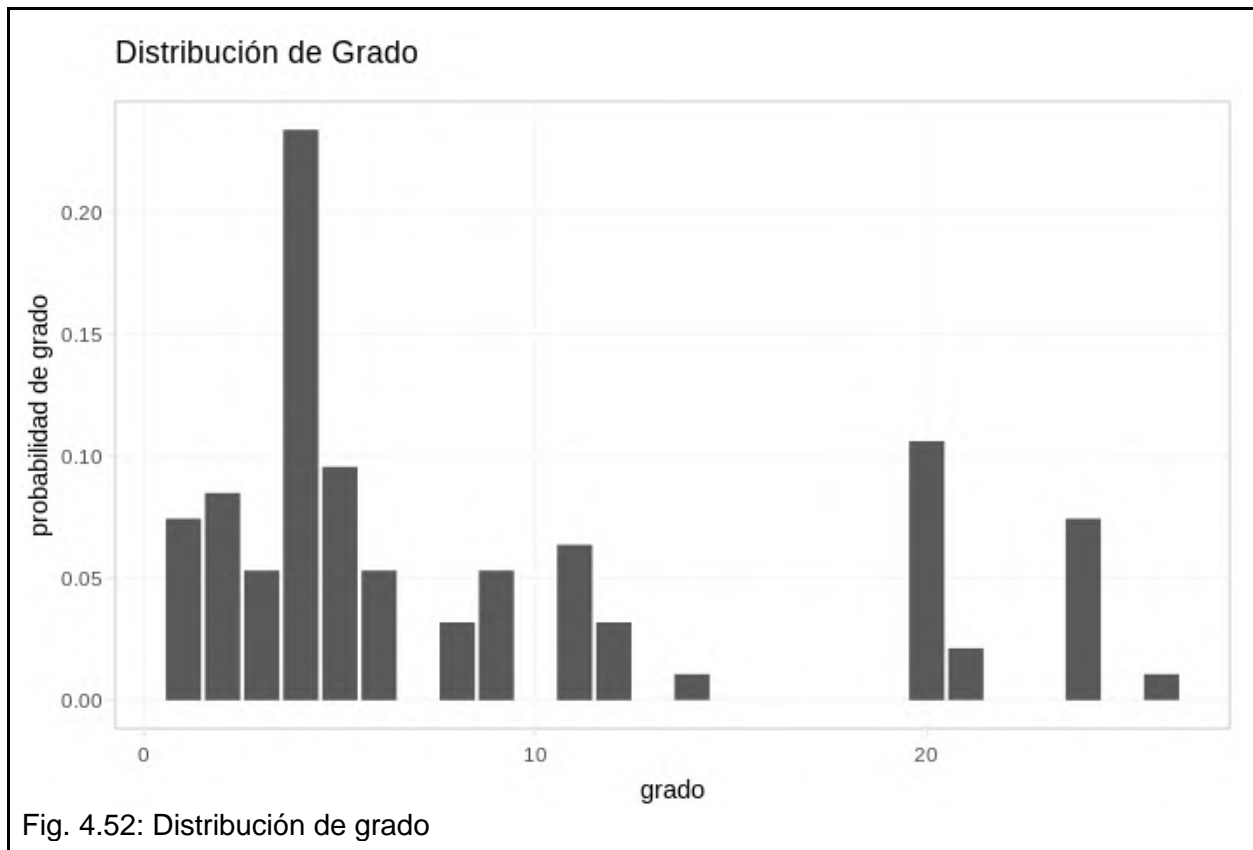


Libre escala 2007 - 2011

Para el periodo 2007 - 2011 vemos que la red no es de libre escala ([Fig. 111](#)).

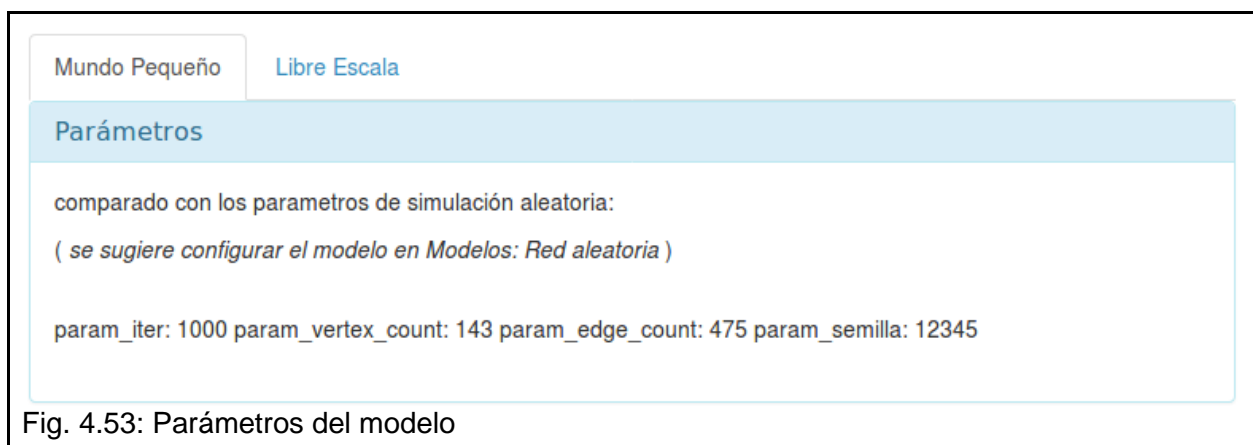


Teniendo en cuenta la distribución, esto lo vemos dado que en altos grados hay una probabilidad elevada, por ejemplo, como se observa en los grados 20 y 24 ([Fig. 4.52](#)).

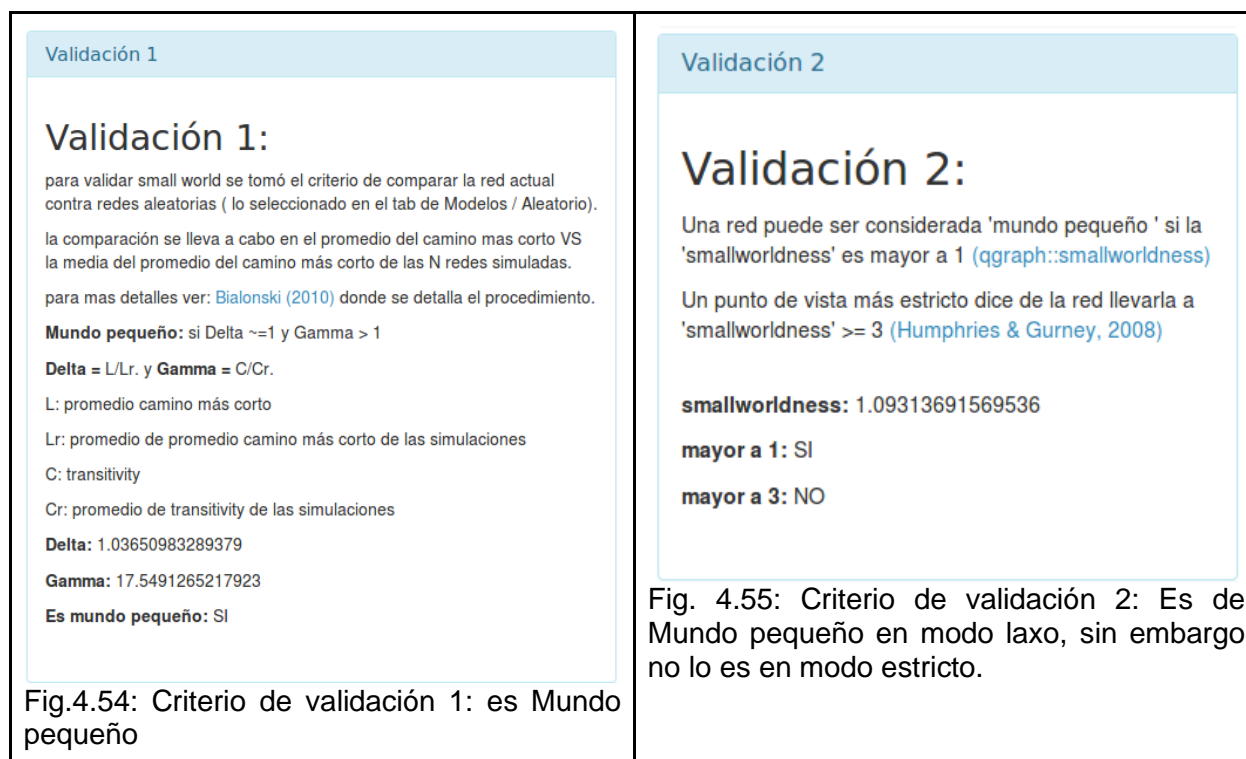


Mundo pequeño para el periodo 2012 - 2016

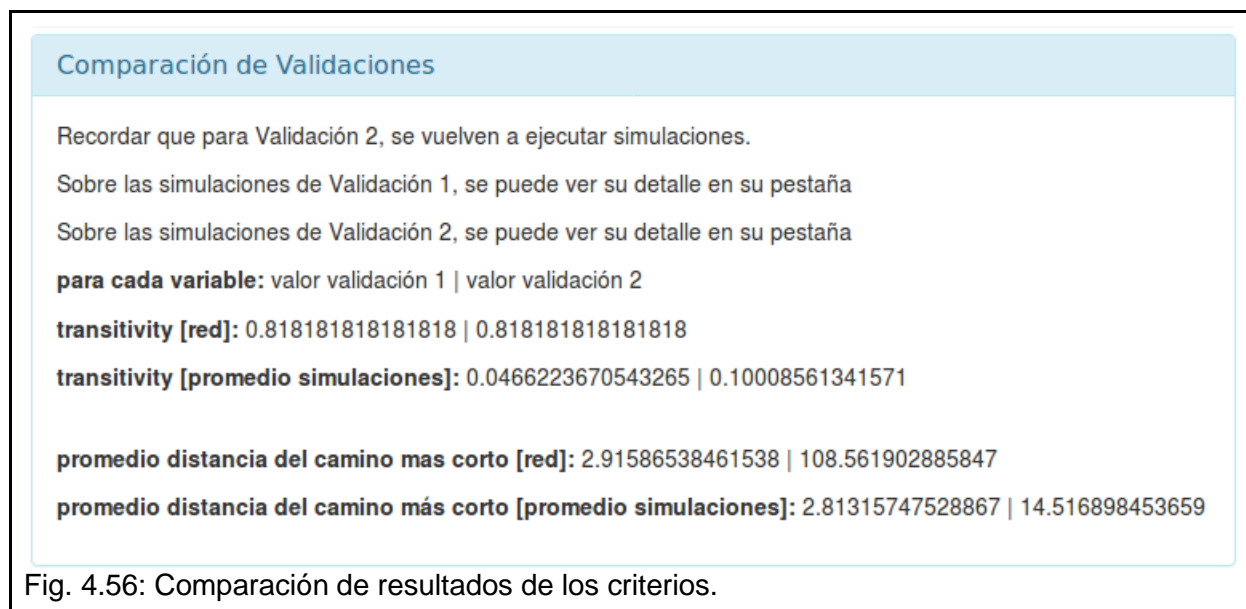
Los parámetros utilizados para esta ejecución se pueden verificar en la [Fig. 4.53](#).



Podemos ver que para el periodo 2012-2016, la red es de mundo pequeño para el criterio 1 ([Fig. 4.54](#)). Sin embargo, para el criterio 2 sí cumple esta característica para el umbral laxo pero no para el estricto ([Fig. 4.55](#)).



Nuevamente, la diferencia entre ambos enfoques de validación de mundo pequeño se pueden observar en la [Fig. 4.56](#).



Libre escala 2012 - 2016

En el periodo de 2012-2016, la red es de libre escala ([Fig. 4.57](#)). Sin embargo, en la distribución se observa que en los grados 14 y 17 hay probabilidad alta ([Fig. 4.58](#)).

Si vemos la referencia es de 0.05, y el máximo manejado para este periodo es de 0.20.

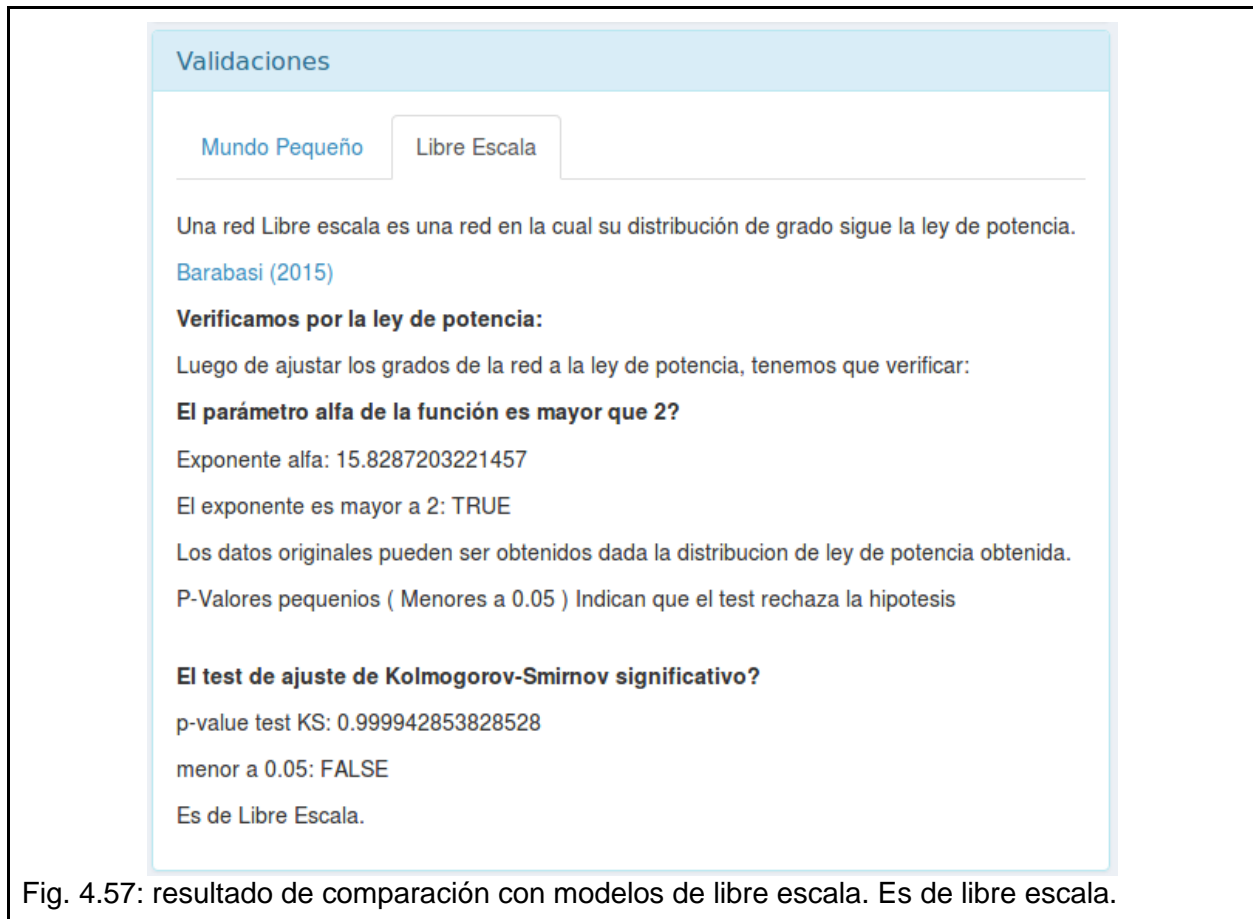
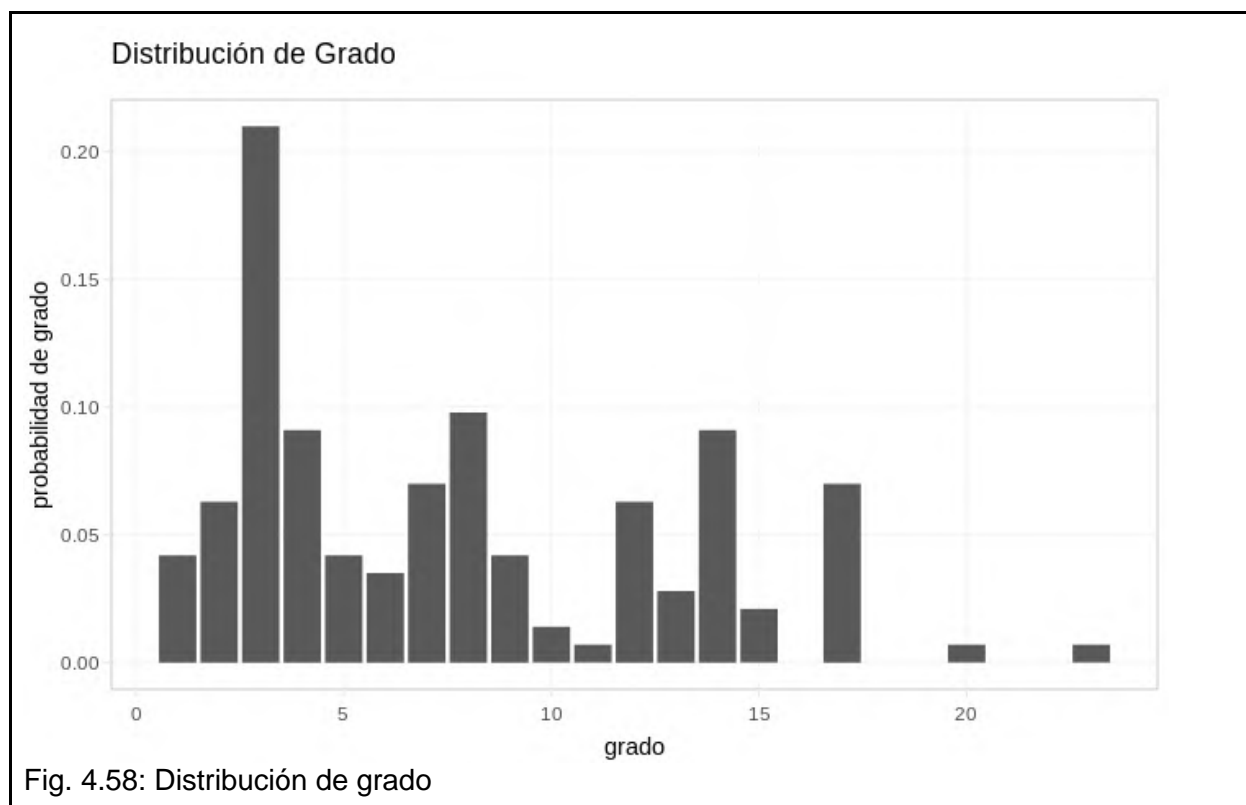


Fig. 4.57: resultado de comparación con modelos de libre escala. Es de libre escala.



Resumen de análisis de las propiedades de mundo pequeño y libre escala

El resumen de la comparación de las redes en el tiempo para las propiedades de mundo pequeño y libre escala se ve en [Tabla 4.11](#).

Tabla 4.11: Resumen de los análisis de Mundo pequeño y Libre Escala para todos los períodos considerados.

propiedad	total	1996-2001	2002-2006	2007-2011	2012-2016
Mundo pequeño criterio 1	SI	SI	NO	SI	SI
Mundo pequeño criterio 2: laxo	NO	NO	SI	NO	SI
Mundo pequeño criterio 2: estricto	NO	NO	NO	NO	NO
Libre escala	SI	SI	SI	NO	SI

Podemos ver que salvo el periodo 2007-2011, la red siempre se comporta como libre escala. Sin embargo, para el caso de mundo pequeño el comportamiento depende de los diferentes criterios de evaluación. Según el criterio 1, es de mundo pequeño excepto en el período 2002-2006. En cambio, para el criterio 2, nunca cumple con la característica de mundo pequeño bajo el umbral estricto mientras que, cuando se aplicó el umbral laxo, tenemos más variación y, para algunos períodos, si se comporta como mundo pequeño.

Luego de esta comparación de la red contra redes generadas según diversos modelos, el siguiente paso fue evaluar cómo se compara con otras redes de coautoría reales.

Comparación con otras redes de coautoría

Así como comparamos la red con modelos similares generados en base sus propiedades, lo más parecido que podríamos obtener en el mundo real son otras redes de coautoría de escritura científica. De esta forma, comparamos las métricas con las obtenidas en otros trabajos, tales como: (a) artículos por autor, (b) coeficiente de clustering (transitividad), (c) tamaño del componente principal, (d) componente principal representado en porcentaje, (e) distancia media. La siguiente Tabla fue obtenida del trabajo de [Koseoglu \(2016\)](#) y modificada para agregar el caso de la RAAB ([Tabla 4.12](#)).

Tabla 4.12: Métricas obtenidas para la RAAB y otras redes de coautorías de revistas científicas.

métricas	Biomedical (Newman, 2001)	Computer science (Newman, 2001)	Management and organization (Acedo et al.,2006a)	Tourism and hospitality (Yeet al., 2013)	Strategic Management (Koseoglu 2016)	Revista Argentina de Antropología Biológica (este estudio)
artículos por autor	6.4	2.6	2.04	1.1	0.88	2.14
autores por artículo	3.75	2.22	1.88	1.87	1.13	3.68
coeficiente de clustering (transitividad)	0.066	0.496	0.68	0.748	0.13	0.543
componente principal representado en porcentaje	92.6	57.2	45.4	59.3	69	67.01
distancia media	4.6	9.7	-	7.2	5.05	3.41
tamaño del componente principal	1395693	6396	4625	1376	296	197

Más allá de las métricas, hay que recordar que para hacer una mejor comparación se debe tratar de entender el contexto de las otras revistas con las cuales se compara, sin embargo este nivel de detalle quedará para trabajos futuros. Algunas preguntas que pueden ayudar a entender el contexto son las siguientes: ¿de qué temáticas tratan?, ¿se acercan más a lo técnico o a lo humano?, ¿cuál es la cota de artículos por edición ?, ¿con qué periodicidad se publica?, ¿publican de todo el mundo?, ¿publican solo artículos locales?, ¿en qué idioma publican?, ¿qué políticas de publicación tienen? Cada una de estas preguntas contestadas en el contexto de publicación del paper en el cual se ven los resultados. Y siempre que las cosas son “en el tiempo” estas preguntas pueden variar de tiempo T a tiempo T+1. Dicho esto, pasamos a comparar solamente valores específicos.

En cuanto a cantidad de autores RAAB se posiciona entre Computer science ([Newman, 2001](#)) y Management and organization ([Acedo et al.,2006a](#)). Esta métrica representa que en el tiempo analizado, los autores participan en promedio con 2.14 artículos a la revista para el caso de la RAAB.

Respecto de autores por artículo, se posiciona cerca de Biomedical ([Newman, 2001](#)). En cuanto a qué puede interpretarse, esta métrica indica qué tanto trabajan en conjunto los autores. Podemos ver que para el caso de RAAB trabajar en un trabajo original en equipo es algo común.

Al revisar transitividad, vemos que lo más próximo es Computer science ([Newman, 2001](#)). La transitividad es la probabilidad media de que dos coautores también sean coautores con otro en común. Esta métrica refuerza lo observado en el punto anterior de que los autores que publican en la RAAB tienden a trabajar en equipo.

Respecto de cuánto representa el componente principal en porcentaje de la red, vemos que lo más próximo es Strategic Management ([Koseoglu 2016](#)). En cuanto a la métrica podemos expresar que representa qué tan conectada está la red total. Cuanto más conectada esté significa que hay menos equipos aislados, por trabajar en trabajos originales en conjunto por lo menos alguna vez a lo largo de la historia de la revista.

Observamos que según la distancia media está más próxima a Biomedical ([Newman, 2001](#)). Recordemos que la distancia media es el promedio de la distancia entre nodos conectados en la red. Vemos que la RAAB es la que da valor más bajo respecto de las otras redes.

Finalmente para el tamaño del componente mayor observamos que la RAAB es la red más pequeña respecto a las demás.

Resumen de capítulo

En este capítulo nos focalizamos en verificar las hipótesis planteadas en el capítulo 1 a través de los datos obtenidos en el capítulo 2, utilizando la aplicación web generada en el capítulo 3.

La hipótesis 1 *“La cantidad de autores aumenta en el tiempo debido al crecimiento de la disciplina en el país, lo que se refleja en un aumento en la cantidad de nodos en la red”* nos interesó para verificar si la disciplina de Antropología Biológica crece en el país luego de verificar revisando la cantidad de autores a lo largo del tiempo en la Revista Argentina de Antropología Biológica, acotamos el estudio a dicha publicación dado que buscamos representación de la materia a nivel local, entendemos que esta hipótesis puede ser expandida en publicaciones internacionales por autores locales y en otras revistas, llegamos a que no podemos rechazar la hipótesis. De manera complementaria se analizó la cantidad de autores por artículo por año como potencial indicador del trabajo en equipo.

Luego de estar más familiarizados con los autores y artículos asociados a la RAAB, nos interesó revisar si el avance de la revista depende de los autores fundadores, o si está impulsada por nuevos autores. Para esto planteamos la hipótesis 2 *“Los autores fundadores se refuerzan en el tiempo, lo cual puede ser medido por grado, intermediación y fuerza de colaboración”* que luego de verificar a través del análisis de las métricas de grado, intermediación y fuerza de colaboración, llegamos a que no podemos rechazar la hipótesis y que la revista depende en gran medida de los autores participantes en las primeras ediciones.

Para tratar de entender en mayor profundidad el comportamiento interno de la red revisamos cómo se comportan los componentes llegando a plantear la hipótesis 3: *“Los componentes más grandes absorben a los más pequeños a lo largo del tiempo, dado que autores periféricos pasan a trabajar en conjunto con autores principales”* analizamos la evolución temporal del tamaño del

componente principal identificando momentos específicos de absorción de componentes. Luego detallamos qué componentes eran integrados al principal. Sin embargo a este enfoque le faltó el detalle de comprender qué estudios o temas fueron los que crearon puentes entre los componentes, y se decidió dejar para un futuro dado que podría ser de interés revisar con profundidad los temas tratados por estos textos y su relación con la creación de puentes entre componentes. De forma complementaria al análisis de componentes verificamos si existe relación con las actividades de investigación, para validar que el componente mayor es la rama principal de investigación, mientras que los otros componentes son ramas especializadas. Y complementamos con la búsqueda de comunidades, para dar otro enfoque a la búsqueda de la rama principal de investigación y ramas especializadas, ambas en una profundidad superficial analizando solo los títulos de los trabajos. Se podría profundizar analizando los textos completos en trabajos futuros.

Luego de tener un panorama interno de la revista, nos llamó la atención comprar con puntos de vista externos, inicialmente validación de propiedades comunes que se dan en redes del mundo real como Mundo pequeño y libre escala dando origen a la hipótesis 4: *“La red agregada total (desde 1996 a 2016) es de mundo pequeño puesto que es una característica presente en la mayoría de las redes de coautorías científicas”* que dependiendo del criterio adoptado puede rechazarse no. Finalmente, de forma complementaria realizamos la comparación con otras redes de coautoría, para complementar la vista hacia afuera, no solo quedarnos en los modelos de la hipótesis 4, sino constatar contra redes de coautoría existentes en trabajos previos.

Logrados los objetivos planteados en el capítulo 1, y con complementos a estos, damos por cerrado el capítulo 4 pasando al capítulo de cierre de la tesis.

Capítulo 5: Conclusiones

Resumen

En la presente tesis se evaluaron las características de la red de coautorías de la Revista Argentina de Antropología Biológica (RAAB), como caso de aplicación para el estudio de redes de coautorías mediante el desarrollo de una aplicación web que facilitó el análisis a través del cálculo automatizado de grafos, métricas y modelos, entre otras herramientas.

En el capítulo uno vimos qué es la RAAB y qué son las redes de coautoría, a modo de contexto y antecedentes. Además, se presentaron los objetivos, preguntas e hipótesis que se trataron en esta tesis.

En el capítulo dos vimos cómo se obtuvo el conjunto de datos. Definimos las fuentes de entrada, las transformaciones realizadas y las salidas de esas transformaciones, y también nos relacionamos con las herramientas para realizar estas tareas.

En el capítulo tres vimos cómo se realizó el proceso de armado de la herramienta para explorar los datos que nos ayudó a investigar las inquietudes iniciales de forma reproducible, interactiva y accesible por otros pares. Además, se detalló qué herramientas usamos para su creación y diferentes consideraciones de armado.

El capítulo cuatro se focalizó en utilizar esta herramienta para analizar el caso puntual de la RAAB. Este análisis se realizó contextualizando y armando un hilo conductor entre las preguntas y sus respuestas.

Finalmente en el capítulo cinco vemos un repaso de lo realizado, discutimos posibles aplicaciones, y planteamos algunos potenciales trabajos futuros.

Productos asociados

Esta tesis dio como resultado los siguientes productos que pueden ser de utilidad en otras tareas y en otros ámbitos:

- **Base de datos:** de los artículos de la RAAB, el link a donde están publicados y el trabajo de desambiguación de autores.
- **Aplicación web para análisis de redes de coautoría:** La aplicación en sí para realizar el análisis en la RAAB.
- **Código abierto:** El código fuente de la aplicación web para análisis de coautoría.
- **Proceso de armado de la Base de datos:** El detalle del proceso del armado de la base de datos.
- **Proceso de armado de la aplicación:** El detalle del proceso del armado de la aplicación.

Consideraciones sobre el conjunto de datos y la variación de estructura del sitio en el tiempo

Durante el desarrollo de la presente tesis hemos encontrado algunos desafíos y dificultades que vale la pena mencionar. Uno de ellos fue que los datos fueron obtenidos en 2017 pero a la hora de la escritura final de la tesis, la estructura de la página de la revista cambió y, por ejemplo, los archivos bibTex ya no son más accesibles.

Además, actualmente las referencias que tiene cada artículo son incluidas en el HTML. Sin embargo, en etapas tempranas de este estudio se descartó hacer análisis de referencias dado que, en aquel momento, estaban embebidas en los PDF y estos no eran accesibles de no ser por tareas manuales. A partir del cambio de estructura del sitio, las referencias de cada artículo están disponibles en la página de cada uno pero esto no era así al momento de extracción de los datos.

Trabajos futuros

Más allá de las ideas propuestas en la presente tesis, existe una amplia bibliografía en el campo de análisis de redes de coautorías y, tomadas en su conjunto, pueden expandirse a futuros trabajos, ya sean de desarrollo de herramientas para facilitar tareas o de aplicación de las técnicas para el soporte en la toma de decisiones fundamentadas en un mapa estratégico más complejo. Como menciona [Kumar \(2015\)](#), el análisis de redes de coautoría es una parte de un análisis que se debe complementar con otros estudios para lograr mejores resultados.

Finalmente el concepto de nodos que se asocian para generar un producto, y el conjunto de herramientas utilizadas para analizar este fenómeno puede abrir otras implementaciones en otras ramas diferentes de coautoría de artículos científicos.

A partir de los desafíos y los análisis realizados en la presente tesis, es posible sugerir varias líneas para explorar a futuro. A continuación mencionamos algunas de ellas.

Actualización de la base de datos

Un producto deseable podría ser expandir la base de datos y quitar la cota temporal. De esta forma, sería posible reutilizar la aplicación para expandir el análisis a más periodos de publicación en la RAAB.

También podría replicarse con la misma estructura para otras revistas, logrando que los análisis sean comparables a través del mismo proceso. Esto no es un trabajo trivial dado que una nueva base de datos implica el proceso de desambiguación de autores, la carga de datos en la aplicación, y luego sacarle la cota temporal en los lugares donde está filtrado. Sin embargo, el trabajo realizado en esta tesis agilizaría mucho la realización de estudios similares.

Aspectos técnicos

La aplicación puede mejorarse desde su estructura, procesos y performance. Así como también aspectos estéticos y de calidad de la herramienta, como puede ser navegabilidad, y sugerencias de trabajo para realizar análisis específicos. También pueden incluirse mejoras en evaluaciones y métricas. Por este motivo se optó por liberar el código fuente para que sea accesible y pueda ir evolucionando como comunidad para poder aceptar propuestas de interesados en el proyecto.

Proyección sobre artículos

Dada la red de naturaleza bipartita (autores - escriben - artículos), el trabajo actual se basó en la proyección sobre autores, pudiéndose leer: “*autor - trabaja con - autor*”. La otra proyección basada sobre los nodos artículos, que se puede leer como “*artículo - comparte autor con - artículo*”, representa un interesante análisis para tener en cuenta en trabajos futuros.

Análisis de textos y tópicos en el tiempo

En etapas muy tempranas del desarrollo de la presente tesis otro objetivo de interés era la realización de análisis de texto en base a los artículos publicados en la RAAB en el período estudiado. Sin embargo, se decidió no incluir estos análisis, dado que extenderían considerablemente los tiempos de trabajo y forman parte de otro tipo de hipótesis. Además, otro aspecto que ayudó a tomar la decisión es que al momento de escritura de la tesis se está desarrollando una “revolución” con respecto a la interpretación de imágenes, que avista que se obtendrán mejores resultados facilitando esta tarea en trabajos futuros.

Por lo tanto, consideramos que, para trabajos futuros, sería interesante realizar análisis textual de los PDF para la extracción de tópicos, evaluar su evolución en el tiempo y su relación con la red de coautoría. En particular, resulta de interés investigar tópicos a través de un análisis textual y verificar si esos tópicos resultantes cuadran con las descripciones de las líneas de investigación en Antropología Biológica planteadas por [Cesani et al.\(2013\)](#). Además, se podría validar si los grupos de coautores identificados en la red están asociados a los tópicos determinados.

Análisis de conectividad / puentes

Este análisis permite plantear varias preguntas estratégicas. Para entender mejor el contexto vamos a contextualizar lo que es un puente según [Barabási \(2016\)](#):

Si una red consiste de dos componentes una arista bien puesta puede conectar ambos componentes, transformando la red en una red conectada. Esas aristas se llaman Puentes. En general un puente es cualquier link que si se corta, desconecta la red.

La herramienta desarrollada en esta tesis ayuda a analizar mejor cómo se fueron dando los puentes en el tiempo, así como qué elementos generaron los mismos. También puede utilizarse para analizar componentes candidatos para incentivar a generar nuevos “Puentes” entre los componentes.

Algunas preguntas asociadas serían:

- ¿Existen posibilidades de generar estos puentes vía coautoría?
- ¿Qué tipo de estudios “podrían generar” dichos puentes?
- ¿Vale la pena tratar de unir estos componentes?
- ¿Qué beneficios se obtendrán de dichos puentes?
- ¿Cuánto cuesta crear esos puentes?

Así como se generan puentes para componentes, se puede tratar de fomentar la conectividad entre comunidades, fomentando nuevas áreas de investigación o complemento entre equipos. En este sentido, algunas preguntas que pueden surgir son:

- ¿Quiénes son los puentes entre las comunidades?
- ¿Qué trabajos se encuentran asociados a los puentes?
- ¿Qué temáticas incluyen estos trabajos?

Dentro del análisis de relaciones se puede evaluar el vínculo entre los autores en el tiempo, conocer quiénes están vinculados, qué temas tratan, en qué época, si existe una frecuencia, entre otras.

Bibliografía

- [1] AABA, Asoc. Antrop. Biolog. Arg. *Revista Argentina de Antropología Biológica*. URL: <https://revistas.unlp.edu.ar/raab/>. Accessed: 2020-01-06.
- [2] Barabási, A. L. (2016). *The power of network science, the beauty of network visualization*. Cambridge University Press. ISBN: 978-1107076266. URL: <http://networksciencebook.com/>.
- [3] Bialonski, S, M. Horstmann, and K. Lehnertz (2010). "From brain to earth and climate systems: Small-world interaction networks or not?" In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20.1, p. 013134. DOI: 10.1063/1.3360561. URL: <https://doi.org/10.1063/1.3360561>
- [4] Blischak, J. D, E. R. Davenport, and G. Wilson (2016). "A Quick Introduction to Version Control with Git and GitHub". In: *PLOS Computational Biology* 12.1, pp. 1-18. DOI: 10.1371/journal.pcbi.1004668. URL: <https://doi.org/10.1371/journal.pcbi.1004668>.
- [5] Blondel, V. D, J. Guillaume, R. Lambiotte, et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, p. P10008. DOI: 10.1088/1742-5468/2008/10/p10008. URL: <https://doi.org/10.1088/1742-5468/2008/10/p10008>.
- [6] Bourgeois, D. T. (2014). *Information Systems for Business and Beyond*. Open Textbook Challenge by the Saylor Academy. URL: <https://bus206.pressbooks.com/>.
- [7] Cesani Rossi, M. F, M. Sardi, S. E. Colantonio, et al. (2013). "Líneas de investigación actuales de la antropología biológica argentina/Current researches in Argentine biological anthropology". In: *Revista Argentina de Antropología Biológica* 16.1, pp. 31-37. URL: <https://revistas.unlp.edu.ar/raab/article/view/749>.
- [8] Clauset, A, M. E. J. Newman, and C. Moore (2004). "Finding community structure in very large networks". In: *Phys. Rev. E* 70 (6), p. 066111. DOI: 10.1103/PhysRevE.70.066111. URL: <https://link.aps.org/doi/10.1103/PhysRevE.70.066111>.
- [9] Clauset, A, C. R. Shalizi, and M. E. J. Newman (2009). "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4, pp. 661-703. DOI: 10.1137/070710111. URL: <https://doi.org/10.1137/070710111>.
- [10] ERDÖS, P. and A. RÉNYI (1959). "On random graphs I". In: *Publ. Math. Debrecen*.
- [11] Fatt, C. K, E. A. Ujum, and K. Ratnavelu (2010). "The structure of collaboration in the *Journal of Finance*". In: *Scientometrics* 85.3, pp. 849-860. ISSN: 1588-2861. DOI: 10.1007/s11192-010-0254-0. URL: <https://doi.org/10.1007/s11192-010-0254-0>.
- [12] Freeman, L. C. (1978). "Centrality in social networks conceptual clarification". In: *Social Networks* 1.3, pp. 215 - 239. ISSN: 0378-8733. DOI: [https://doi.org/10.1016/0378-8733\(78\)90021-7](https://doi.org/10.1016/0378-8733(78)90021-7). URL: <http://www.sciencedirect.com/science/article/pii/0378873378900217>.

- [13] Humphries, M. D. and K. Gurney (2008). "Network 'Small-World-Ness': A Quantitative Method for Determining Canonical Network Equivalence". In: *PLOS ONE* 3.4, pp. 1-10. DOI: 10.1371/journal.pone.0002051. URL: <https://doi.org/10.1371/journal.pone.0002051>.
- [14] Kleinberg, J. M. (1999). "Authoritative Sources in a Hyperlinked Environment". In: *J. ACM* 46.5, p. 604-632. ISSN: 0004-5411. DOI: 10.1145/324133.324140. URL: <https://doi.org/10.1145/324133.324140>.
- [15] Kolaczyk, E. D. and G. Csárdi (2020). *Statistical Analysis of Network Data with R*. En. Springer International Publishing. DOI: <https://doi.org/10.1007/978-3-030-44129-6>.
- [16] Koseoglu, M. A. (2016). "Growth and structure of authorship and co-authorship network in the strategic management realm: Evidence from the Strategic Management Journal". In: *BRQ Business Research Quarterly* 19.3, pp. 153 - 170. ISSN: 2340-9436. DOI: <https://doi.org/10.1016/j.brq.2016.02.001>. URL: <http://www.sciencedirect.com/science/article/pii/S2340943616300019>.
- [17] Kumar, S. (2015). "Co-authorship networks: a review of the literature". In: *Aslib Journal of Information Management* 67.1, pp. 55-73. ISSN: 2050-3806. DOI: 10.1108/AJIM-09-2014-0116. URL: <https://doi.org/10.1108/AJIM-09-2014-0116>.
- [18] Longabaugh, W. J. (2012). "Combing the hairball with BioFabric: a new approach for visualization of large networks". In: *BMC Bioinformatics* 13.1, p. 275. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-275. URL: <https://doi.org/10.1186/1471-2105-13-275>.
- [19] Madrigal, L. and J. R. González (2016). *Introducción a la antropología biológica*. Asociación Latinoamericana de Antropología Biológica (ALAB). URL: https://scholarcommons.usf.edu/islac_alab_antropologia/1.
- [20] Milgram, S. (1967). "The small world problem". In: *Psychol. Today* 2.1, pp. 60-67.
- [21] Morel, C. M, S. J. Serruya, G. O. Penna, et al. (2009). "Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases". In: *PLOS Neglected Tropical Diseases* 3.8, pp. 1-7. DOI: 10.1371/journal.pntd.0000501. URL: <https://doi.org/10.1371/journal.pntd.0000501>.
- [22] Newman, M. E. J. (2001). "The structure of scientific collaboration networks". In: *Proceedings of the National Academy of Sciences* 98.2, pp. 404-409. ISSN: 0027-8424. DOI: 10.1073/pnas.98.2.404. eprint: <https://www.pnas.org/content/98/2/404.full.pdf>. URL: <https://www.pnas.org/content/98/2/404>.
- [23] Newman, M. E. J. (2004). "Coauthorship networks and patterns of scientific collaboration". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5200-5205. ISSN: 0027-8424. DOI: 10.1073/pnas.0307545100. URL: https://www.pnas.org/content/101/suppl_1/5200.
- [24] Newman, M. E. J. (2005). "Power laws, Pareto distributions and Zipf's law". In: *Contemporary Physics* 46.5, pp. 323-351. DOI: 10.1080/00107510500052444. URL: <https://doi.org/10.1080/00107510500052444>.

- [25] Newman, M. E. J. (2006). "Finding community structure in networks using the eigenvectors of matrices". In: *Phys. Rev. E* 74 (3), p. 036104. DOI: 10.1103/PhysRevE.74.036104. URL: <https://link.aps.org/doi/10.1103/PhysRevE.74.036104>.
- [26] Newman, M. E. J. (2016). "Mathematics of Networks". In: *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan UK, pp. 1-8. ISBN: 978-1-349-95121-5. DOI: 10.1057/978-1-349-95121-5_2565-1. URL: https://doi.org/10.1057/978-1-349-95121-5_2565-1.
- [27] Newman, M. E. J. and M. Girvan (2004). "Finding and evaluating community structure in networks". In: *Phys. Rev. E* 69 (2), p. 026113. DOI: 10.1103/PhysRevE.69.026113. URL: <https://link.aps.org/doi/10.1103/PhysRevE.69.026113>.
- [28] Ordoñez, G, A. Hernández Badillo, C. Hernández, et al. (2009). "Análisis bibliométrico de la Revista de Economía Institucional en sus primeros diez años". In: *Revista de Economía Institucional* 11.20. URL: <https://revistas.uexternado.edu.co/index.php/ecoins/article/view/354>.
- [29] Pons, P. and M. Latapy (2005). Computing communities in large networks using random walks (long version). arXiv: physics/0512106 [physics.soc-ph].
- [30] Raghavan, U. N, R. Albert, and S. Kumara (2007). "Near linear time algorithm to detect community structures in large-scale networks". In: *Physical Review E* 76.3. ISSN: 1550-2376. DOI: 10.1103/physreve.76.036106. URL: <http://dx.doi.org/10.1103/PhysRevE.76.036106>.
- [31] Rosvall, M, D. Axelsson, and C. T. Bergstrom (2009). "The map equation". In: *The European Physical Journal Special Topics* 178.1, p. 13–23. ISSN: 1951-6401. DOI: 10.1140/epjst/e2010-01179-1. URL: <http://dx.doi.org/10.1140/epjst/e2010-01179-1>.
- [32] Saurkar, A. V., K. G. Pathare, and S. A. Gode (2018). *n Overview on Web Scraping Techniques and Tools*. URL: http://www.ijfrscce.org/download/browse/Volume_4/April_18_Volume_4_Issue_4/1524638955_25-04-2018.pdf.
- [33] Savić, M., M. Ivanović, M. Radovanović, et al. (2015). "Exploratory Analysis of Communities in Co-authorship Networks: A Case Study". In: *ICT Innovations 2014*. Springer International Publishing, pp. 55-64. DOI: https://doi.org/10.1007/978-3-319-09879-1_6. URL: https://link.springer.com/chapter/10.1007/978-3-319-09879-1_6.
- [34] Wilson, G, J. Bryan, K. Cranston, et al. (2017). "Good enough practices in scientific computing". In: *PLOS Computational Biology* 13.6, pp. 1-20. DOI: 10.1371/journal.pcbi.1005510. URL: <https://doi.org/10.1371/journal.pcbi.1005510>.
- [35] Wojciechowski, J, A. M. Hopkins, and R. N. Upton (2015). "Interactive Pharmacometric Applications Using R and the Shiny Package". Eng. In: *CPT: pharmacometrics & systems pharmacology* 4.3. 26225240[pmid], pp. e00021-e00021. ISSN: 2163-8306. DOI: 10.1002/psp4.21. URL: <https://pubmed.ncbi.nlm.nih.gov/26225240>.
- [36] Acedo, F. J., Barroso, C., Casanueva, C., & Galan, J. L. (2006). Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis*. *Journal of Management Studies*, 43(5), 957–983. doi:10.1111/J.1467-6486.2006.00625.X

ANEXO: software y librerías utilizadas

Dado que las librerías son parte esencial de cualquier sistema desarrollado hoy en día, se especifica qué librerías se utilizaron para cada sección de la aplicación y con qué propósito. Si como también sus citas correspondientes.

librería / software	sección	proposito	citas
R	Generación y aplicación	Lenguaje de programación	R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ .
tidyr	Generación y aplicación	Transformación datos	Hadley Wickham and Lionel Henry (2019). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.3. https://CRAN.R-project.org/package=tidyr
dplyr	Generación y aplicación	Transformación datos	Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.1. https://CRAN.R-project.org/package=dplyr
purrr	Generación y aplicación	Transformación datos	Lionel Henry and Hadley Wickham (2019). purrr: Functional Programming Tools. R package version 0.3.2. https://CRAN.R-project.org/package=purrr
stringr	Generación y aplicación	Transformación datos	Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr
stringi	Generación y aplicación	Transformación datos	Gagolewski M. and others (2019). R package stringi: Character string processing facilities. http://www.gagolewski.com/software/stringi/ .
lubridate	Generación y aplicación	Transformación datos	Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/ .
RSQLite	Generación y aplicación	Operaciones Base de datos	Kirill Müller, Hadley Wickham, David A. James and Seth Falcon (2018). RSQLite: 'SQLite' Interface for R. R package version 2.1.1. https://CRAN.R-project.org/package=RSQLite
DBI	Generación y aplicación	Operaciones Base de datos	R Special Interest Group on Databases (R-SIG-DB), Hadley Wickham and Kirill Müller (2018). DBI: R Database Interface. R package version 1.0.0. https://CRAN.R-project.org/package=DBI
igraph	Aplicación	Operaciones grafos	Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org
ggplot2	Aplicación	Graficar	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
qgraph	Aplicación	Operaciones grafos	Sacha Epskamp, Angelique O. J. Cramer, Lourens J. Waldorp, Verena D. Schmittmann, Denny Borsboom (2012). qgraph: Network Visualizations of Relationships in Psychometric Data. Journal of Statistical Software, 48(4), 1-18. URL http://www.jstatsoft.org/v48/i04/ .
RColorBrewer	Aplicación	Selección colores	Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer
visNetwork	Aplicación	Visualización grafos	Almende B.V., Benoît Thieurmél and Titouan Robert (2019). visNetwork: Network Visualization using 'vis.js' Library. R package version 2.0.6. https://CRAN.R-project.org/package=visNetwork
shiny	Aplicación	Funcionamiento aplicación	Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2019). shiny: Web Application Framework for R. R package

			version 1.3.2. https://CRAN.R-project.org/package=shiny
DT	Aplicación	Visualización y funcionalidades tablas interactivas	Yihui Xie, Joe Cheng and Xianying Tan (2019). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.6. https://CRAN.R-project.org/package=DT
shinyWidgets	Aplicación	Funcionamiento Aplicación	Victor Perrier, Fanny Meyer and David Granjon (2019). shinyWidgets: Custom Inputs Widgets for Shiny. R package version 0.4.8. https://CRAN.R-project.org/package=shinyWidgets
bsplus	Aplicación	Funcionamiento Aplicación	Ian Lyttle (2018). bsplus: Adds Functionality to the R Markdown + Shiny Bootstrap Framework. R package version 0.1.1. https://CRAN.R-project.org/package=bsplus
plotly	Aplicación	Gráficos interactivos	Carson Sievert (2018) plotly for R. https://plotly-r.com
shinydashboard	Aplicación	Funcionamiento Aplicación	Winston Chang and Barbara Borges Ribeiro (2018). shinydashboard: Create Dashboards with 'Shiny'. R package version 0.7.1. https://CRAN.R-project.org/package=shinydashboard
shinydashboardPlus	Aplicación	Funcionamiento Aplicación	David Granjon (2019). shinydashboardPlus: Add More 'AdminLTE2' Components to 'shinydashboard'. R package version 0.7.0. https://CRAN.R-project.org/package=shinydashboardPlus
ndtv	Aplicación	Animaciones grafos	Skye Bender-deMoll (2018). ndtv: Network Dynamic Temporal Visualizations. R package version 0.12.3. https://CRAN.R-project.org/package=ndtv
intergraph	Aplicación	Transformación grafos	Bojanowski, Michal (2015) intergraph: Coercion Routines for Network Data Objects. R package version 2.0-2. http://mbojan.github.io/intergraph
assertive	Aplicación	Pruebas de funciones	Richard Cotton (2016). assertive: Readable Check Functions to Ensure Code Integrity. R package version 0.3-5. https://CRAN.R-project.org/package=assertive
testthat	Aplicación	Pruebas de funciones	Hadley Wickham. testthat: Get Started with Testing. The R Journal, vol. 3, no. 1, pp. 5–10, 2011
igraphdata	Aplicación	Pruebas de funciones	Gabor Csardi (2015). igraphdata: A Collection of Network Data Sets for the 'igraph' Package. R package version 1.0.1. https://CRAN.R-project.org/package=igraphdata
forcats	Aplicación	Transformación datos	Hadley Wickham (2019). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.4.0. https://CRAN.R-project.org/package=forcats
visdat	Aplicación	Gráficos	Tierney N (2017). "visdat: Visualising Whole Data Frames." _JOSS_, *2*(16), 355. doi:10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355 >.
ggridges	Aplicación	Gráficos	Claus O. Wilke (2018). ggridges: Ridgeline Plots in 'ggplot2'. R package version 0.5.1. https://CRAN.R-project.org/package=ggridges
skimr	Aplicación	Cálculo, visualización estadísticas	Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu, Shannon Ellis and Michael Quinn (2019). skimr: Compact and Flexible Summaries of Data. R package version 1.0.5. https://CRAN.R-project.org/package=skimr
ggrepel	Aplicación	graficos	Kamil Slowikowski (2019). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.8.1. https://CRAN.R-project.org/package=ggrepel

Rcpp	compilacion	Compilación código	Dirk Eddebuettel and Romain Francois (2011). Rcpp: Seamless R and C++ Integration. Journal of Statistical Software, 40(8), 1-18. URL http://www.jstatsoft.org/v40/i08/ . Eddebuettel, Dirk (2013) Seamless R and C++ Integration with Rcpp. Springer, New York. ISBN 978-1-4614-6867-7. Dirk Eddebuettel and James Joseph Balamuta (2017). Extending R with C++: A Brief Introduction to Rcpp. PeerJ Preprints 5:e3188v1. URL https://doi.org/10.7287/peerj.preprints.3188v1 .
devtools	compilacion	Obtención código	Hadley Wickham, Jim Hester and Winston Chang (2019). devtools: Tools to Make Developing R Packages Easier. R package version 2.0.2. https://CRAN.R-project.org/package=devtools
htmlwidgets	Aplicación	Funcionamiento aplicación	Ramnath Vaidyanathan, Yihui Xie, JJ Allaire, Joe Cheng and Kenton Russell (2018). htmlwidgets: HTML Widgets for R. R package version 1.3. https://CRAN.R-project.org/package=htmlwidgets
jsonlite	Aplicación	Funcionamiento aplicación	Jeroen Ooms (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805 .
RBioFabric	Aplicación	Visualización	William Longabaugh (2016). RBioFabric: BioFabric network visualization tool. R package version 0.4. http://www.BioFabric.org/
OpenRefine	Generación	Limpieza datos autores.	OpenRefine. A free, open source, power tool for working with messy data. http://openrefine.org/
here	Generación y Aplicación	Acceso a archivos	Kirill Müller (2017). here: A Simpler Way to Find Your Files. R package version 0.1. https://CRAN.R-project.org/package=here
Rstudio	Generación y Aplicación	Programación	RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/ .
Git	Generación y Aplicación	Versionado	Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency. https://git-scm.com/
RefManageR	Generación	Extracción citas	McLean MW (2017). "RefManageR: Import and Manage BibTeX and BibLaTeX References in R." _The Journal of Open Source Software_. doi: 10.21105/joss.00338 (URL: https://doi.org/10.21105/joss.00338). McLean MW (2014). _Straightforward Bibliography Management in R Using the RefManager Package_. arXiv: 1403.2036 [cs.DL], <URL: https://arxiv.org/abs/1403.2036 >.
rvest	Generación	Extracción HTML	Hadley Wickham (2019). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.4. https://CRAN.R-project.org/package=rvest
pdftools	Generación	Extracción PDF	Jeroen Ooms (2019). pdftools: Text Extraction, Rendering and Converting of PDF Documents. R package version 2.2. https://CRAN.R-project.org/package=pdftools
tesseract	Generación	Extracción OCR	Jeroen Ooms (2018). tesseract: Open Source OCR Engine. R package version 4.0. https://CRAN.R-project.org/package=tesseract
network	Aplicación	grafos	Butts C (2015). _network: Classes for Relational Data_. The Statnet Project (<URL: http://www.statnet.org >). R package version 1.13.0.1, <URL: https://CRAN.R-project.org/package=network >. Butts C (2008). "network: a Package for Managing Relational Data in R." _Journal of Statistical Software_, "24"(2).

			<URL: http://www.jstatsoft.org/v24/i02/paper >.
networkDynamic	Aplicación	grafos	Carter T. Butts, Ayn Leslie-Cook, Pavel N. Krivitsky and Skye Bender-deMoll (2019). networkDynamic: Dynamic Extensions for Network Objects. R package version 0.10.0. https://CRAN.R-project.org/package=networkDynamic

ANEXO: Detalle de algoritmos para búsqueda de comunidades

Cada uno de los algoritmos disponibles representa diferentes formas de considerar la agrupación y su interpretación depende de que representen los nodos y las aristas. Los algoritmos utilizados están disponibles a través de la [librería igraph](#). Un breve resumen se puede ver en [Tabla 21](#).

Tabla 21: Resumen algoritmos de búsqueda de comunidades.

Función en Herramienta	Tipo Comunidad	Referencia de implementación	Descripción algoritmo
Cluster_edge_betweenness	Edge Betweenness	http://igraph.org/r/doc/cluster_edge_betweenness.html	Identifica la intermediación de las aristas, y luego las va quitando a las que mayor valor tienen, para ver cómo se van formando nuevos subgrafos. Para detalles ver Newman-Girvan (2004)
Cluster_label_prop	Propagating Labels	http://igraph.org/r/doc/cluster_label_prop.html	Cada nodo arranca con una etiqueta única, en el siguiente paso cada nodo adopta la etiqueta que la mayoría de sus vecinos tiene. Para más detalles ver Raghavan,Albert, et al. (2007)
Cluster_leading_eigen	Leading Eigenvector	http://igraph.org/r/doc/cluster_leading_eigen.html	Utiliza la matriz de modularidad, y luego utiliza los autovectores y autovalores de la misma. Para más detalles ver Newman (2006)
Cluster_louvain	Multi-level Optimization of Modularity	http://igraph.org/r/doc/cluster_louvain.html	Mezcla una forma jerárquica con la modularidad. Optimiza la modularidad primero localmente, cuando no quedan más vértices. Vuelve a arrancar tomando cada comunidad como un vértice. Termina cuando solo queda un vértice. Para más detalles ver Blondel Guillaume, et al.(2008)
Cluster_walktrap	Short Random Walks	http://igraph.org/r/doc/cluster_walktrap.html	La idea es que caminatas aleatorias cortas se quedan en la misma comunidad. Para más detalles ver Pons-Latapy (2005)
Cluster_infomap	Infomap community finding	http://igraph.org/r/doc/cluster_infomap.html	Encuentra estructuras de comunidad que minimizan el largo de la descripción de la trayectoria de un caminador aleatorio. Rosvall, Axelsson, et al.(2009)

ANEXO: Detalle procesamiento bibTEX

Los archivos bibtex fueron generados de la siguiente manera:

- 1) Se accedió al link de javascript generado por revista
- 2) Se guardó la página HTML que tenía un elemento textarea con el texto bibtex correspondiente
- 3) Se extrajo el texto bibtex del elemento textarea
- 4) Se guardó en un archivo bibtex
- 5) Se trató de parsear el archivo bibtex con la [librería RefManageR](#)
- 6) Se encontraron problemas generalmente relacionados a mala escritura del formato bibtex
 - a) Tema de escribir autor1 Y autor2, mientras que en bibtex es autor1 **AND** autor2.
- 7) Se encontraron problemas relacionados con el encoding de los archivos.
 - a) ASCII vs UTF8 (acentos, ñ, etc)
- 8) Se resolvieron ambos problemas
 - a) El primero identificando los casos de forma manual
 - b) El 2do utilizando sistema operativo Linux para realizar la transformación y guardar en UTF-8
- 9) Se encontraron inconsistencias en los datos presentados en el formato bibtex y lo que se encontraba presentado en los HTML de la revista.
 - a) Falta de dato volumen
 - b) Falta de abstract
- 10) Se terminó extrayendo la información que se consideraba complementaria y se almacenó en la base de datos de la revista.

Problemas de transformación y encoding bibtex

Detalles técnicos de los sistemas operativos utilizados:

Linux: Linux Mint 17.3 Rosa (basado en Ubuntu 14.04)

Windows: Windows 10 y Windows 7

Luego de resolver estos temas en Linux (con caracteres no ASCII), se pudo generar el archivo de texto y luego levantarlo sin problemas.

Sin embargo cuando utilizo una computadora con Sistema operativo Windows no se pudo levantar el archivo generado en Linux por problemas de BOM ya que en Windows por default toma los archivos con cierta configuración de R. Mientras el default de configuración Linux es UTF-8, por lo tanto hace más compatible los acentos y caracteres no ASCII

De esta forma se optó por hacer las tareas de parseo de archivos en Linux, y luego de tener generados los archivos bases, se pudo seguir trabajando en ambos sistemas operativos.

Problemas autores

Apellidos con apóstrofe.

<https://revistas.unlp.edu.ar/raab/article/view/229>

<https://revistas.unlp.edu.ar/raab/article/view/1513>

Gabriela L'Heureux

Manuel D. D'Angelo del Campo

Solución: reemplazar el apóstrofe por comilla simple.

Problemas con el volumen de la revista

Link bibtex	
https://revistas.unlp.edu.ar/raab/article/cite/612/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/425/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2541/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/554/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2505/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/458/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/1172/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/258/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2506/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/259/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/574/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2514/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/260/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/561/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2516/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/261/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/557/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2515/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/265/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/579/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2511/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/268/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2507/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/269/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2177/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/270/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/2513/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/271/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/272/BibtexCitationPlugin	Falta campo volumen

https://revistas.unlp.edu.ar/raab/article/cite/273/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/274/BibtexCitationPlugin	Falta campo volumen
https://revistas.unlp.edu.ar/raab/article/cite/275/BibtexCitationPlugin	Falta campo volumen

Problema Abstracts Vacios en bibtex

https://revistas.unlp.edu.ar/raab/article/cite/800/BibtexCitationPlugin	Comillas dobles y simples
https://revistas.unlp.edu.ar/raab/article/cite/173/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/2860/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/747/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/2223/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/2489/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/612/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1090/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/352/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/377/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1510/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/675/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/863/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/3047/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/874/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1172/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/3722/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1513/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1514/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1515/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/2516/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/3036/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/228/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/269/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/2177/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/272/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/274/BibtexCitationPlugin	

https://revistas.unlp.edu.ar/raab/article/cite/1706/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1709/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1710/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1720/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1721/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1722/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1730/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1735/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1745/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1746/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1747/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1748/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1813/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1818/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1826/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1829/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1831/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1837/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1839/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1840/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1842/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1852/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1861/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1804/BibtexCitationPlugin	
https://revistas.unlp.edu.ar/raab/article/cite/1869/BibtexCitationPlugin	

ANEXO: Detalle consideraciones PDF

Los archivos PDF al final no fueron considerados. Se eliminó la sección de text mining para tópicos, quedando para potenciales trabajos futuros. Sin embargo existieron las siguientes lecciones aprendidas:

Obtención de los pdf

Se pudieron bajar sin problema, se agregó una demora entre archivo y archivo, para no saturar el servidor de la revista.

Parseo de los pdf

Conversión de los pdf a texto; Existían diferentes formatos de los pdf;

Se probó:

- Bajar un artículo de la edición más antigua: es un escaneo
- Bajar un artículo de una edición intermedia: es un escaneo
- Bajar un artículo de una edición moderna: es un texto

Potenciales soluciones:

- OCR (*optical character recognition*) para los escaneos
- Extracción de texto para los pdf más modernos.

Extracción de texto (Se utilizó el lenguaje R , con la [librería pdftools](#))

- Errores con formato de más de 1 columna.
- Errores de extracción de texto por márgenes de página.
- Se decidió convertir cada hoja a una imagen y luego pasarlas por OCR.

Aplicación de OCR (Se utilizó el lenguaje R , con las [librerías pdftools y tesseract](#))

- Extracción de las imágenes de cada página a través de la [librería pdftools](#)
- Procesar cada imagen por artículo ejecutando OCR a través de la [librería tesseract](#), se aplicó el uso del motor de idioma en español.
- Compaginar todos los textos en un archivo txt final que contenga todos los resultados OCR de cada una de las páginas.

Con esto se logró superar considerablemente el problema de márgenes, sin embargo seguían existiendo problemas con los textos, y extracción de los datos de los mismos.

Se decidió no incluir la parte de text mining en la tesis, dado que sino se extendería considerablemente. Además dado que hoy en día (2017-2020) hay una revolución con la interpretación de imágenes, que tal vez en breve tengamos mejores resultados pudiendo realizar de forma más fácil en trabajos futuros.

ANEXO: Detalle estandarización autores

Se juntaron los conjuntos de datos de los orígenes de HTML , y los archivos de citas BIBtex. Se exportaron los datos resultantes a un archivo en formato texto separado por tabulaciones.

Este archivo se importó en la herramienta [OpenRefine](#) la cual resulta muy práctica a la hora de buscar similitudes entre registros de cada columna. La versión utilizada para realizar este trabajo fue OpenRefine 2.8.

Una vez cargado el archivo , se procedió a repetir la columna autor , y a la nueva ponerle de nombre “autor_original”. Esta columna se agrega para cumplir 2 funciones, una de copia de respaldo, en caso de que llegase a haber inconvenientes. Y la segunda, que es para mapear el valor original de la columna con el valor obtenido luego de procesar la columna “autor” repetidas veces. ([Fig. anx_est_aut_1](#))



Sobre la columna “autor” las tareas realizadas fueron para buscar similitudes a través de métodos de agrupamiento (*clustering*). A continuación se pueden ver los métodos que se ejecutaron, y en qué orden. Más allá de este resultado se exportó un archivo con las transformaciones realizadas; sin embargo este archivo no conserva los métodos de clustering usados, ni los parámetros utilizados; por lo cual se pasó a detallar en formato textual dentro de este anexo. (ver [Iteraciones de clustering](#))

Revisión y validación con expertos

Luego de depurar empíricamente con diferentes formas de agrupar, validamos los nombres y casos con los expertos del área. Surge una lista de autores a unificar, que no aparecieron en el análisis de clustering o que había que modificar. (ver [Tabla 22](#))

Se cambió el formato de excel enviado por los expertos a un formato más amigable para la automatización. Se armó una utilidad intermedia que dado un excel, transforma a el formato de importación del openrefine, que es un archivo con un formato específico. Gracias a la herramienta se pudieron distinguir errores que se pasaron, en la parte de la unificación de autores.

Iteraciones de clustering

Clustering - Iteración 1

Método y parámetros

Clustering > columna autor > método: keycollision > keying function: fingerprint

Criterio de selección

Empírico. Es la primera opción al seleccionar la opción de clustering en openrefine en la versión utilizada (version: 2.7).

Observación

Dado que son varios los cambios se procede a detallar reglas generales aplicadas, para ver detalles de cuáles fueron exactamente ver el archivo exportado de reglas del openrefine.

Resultado

Al realizar el primer clustering en nombres de autores se puede ver que existen varios Temas con escritura del nombre:

- Nombres que consideran el nombre intermedio (María Antonia Luis)
- Nombres que al intermedio lo ponen con una abreviación (María A. Luis)
- Abreviación del nombre intermedio con punto (María A. Luis)
- Abreviación del nombre intermedio sin punto (María A Luis)
- Nombres con acentos o caracteres especiales (María A. Luis)
- Nombres en los que se obvian los acentos o caracteres especiales. (Maria A. Luis)
- Nombres que cambian en el tiempo > ej: Cesani > Cesani Rossi

Clustering - Iteración 2

Método y parámetros

Clustering > columna autor > método: keycollision > keying function: fingerprint

Criterio de selección

Empírico. Volver a tratar de ejecutar el mismo método para ver que no existan más duplicados con los parámetros utilizados.

Resultado

Se trata de dar un segundo clustering bajo la misma metodología y no se encuentran resultados

Clustering - Iteración 3

Método y parámetros

Se procede a probar con otro tipo de configuración de clustering
método: keycollision > keying function: ngram-fingerprint > ngram-size: 2

Criterio de selección

Empírico. Dentro de las opciones de parámetros dentro del método de key collision es la 2da opción. El tamaño del engrama fue seleccionado porque es el que viene por defecto.

Observación

Dados que son varios los cambios se procede a detallar reglas generales aplicadas, para ver detalles de cuáles fueron exactamente ver el archivo exportado de reglas del openrefine.

Resultado

Se encuentran temas con:

- Más de una abreviación en nombres intermedios
- Apellidos compuestos con guiones
- Múltiples nombres y apellidos

Clustering - Iteración 4**Método y parámetros**

Se procede a ejecutar nuevamente bajo la misma metodología.

método: keycollision > keying function: ngram-fingerprint > ngram-size: 2

Criterio de selección

Empírico. Validación de lo aplicado en Iteración 3. Se espera que no haya resultados.

Observación

Dados que son varios los cambios se procede a detallar reglas generales aplicadas, para ver detalles de cuáles fueron exactamente ver el archivo exportado de reglas del openrefine.

Resultado

No se encuentran resultados.

Clustering - Iteraciones 5-9

Método y parámetros

Se procede a ejecutar nuevamente bajo la misma metodología , variando el tamaño del engrama con los valores 3,4,5,6,7

método: keycollision > keying function: ngram-fingerprint > ngram-size: 2

Criterio de selección

Empírico. Probando expandir el tamaño del engrama base para el clustering.

Resultado

No se encuentran resultados. Para los parámetros 3,4,5,6.

En el parámetro con valor 7 si, pero no se utilizaron las sugerencias planteadas, dado que afecta los nombres y apellidos cortos.

Clustering - Iteración 10**Método y parámetros**

Se procede a pasar a otra estrategia. Método: keycollision , keying function: metaphone3

Criterio de selección

Empírico. Probando otra función.

Resultado

Dando nuevos resultados esta vez relacionados con:

- Casos de nombres compuestos o apellidos compuestos (Fernando Ramírez Rozzi)
 - Para la elección del adecuado, generalmente se termina buscando en google o cuit online u otras páginas para verificar existencia de la persona y asociada al área; para ver perfiles asociados a la persona, queda constatar con los expertos
- Hay casos dudosos como pro ejemplo: Rolando González-José , vs Rolando González
 - Al ser apellidos tan comunes es difícil de realizar elecciones; a constatar con expertos.

Respecto de consideraciones con apellidos del estilo

- Cesani > Cesani Rossi ; que pueden asociar a cambio de estados civiles. Por lo tanto sigue siendo la misma persona, hay que determinar si estos cambios interesan , o si se desea cambiar el análisis de red.
- Hay que tener cuidado al aplicar las modificaciones del campo dado que por ejemplo: Francisco Di Fabbio Rocca, no es lo mismo que Francisco de Ávila ; hay que ser criteriosos con las elecciones.

Clustering - Iteración 11**Método y parámetros**

Se procede a validar la estrategia anterior. Método: keycollision , keying function: metaphone3

Criterio de selección

Empírico. Verificando método utilizado en iteración 10, para búsqueda de más resultados.

Resultado

Se volvió a dar una 2da corrida del algoritmo luego de las correcciones, el único caso es el de Francisco Di Fabbio Rocca y Francisco de Ávila ; que lo dejamos sin corrección. Dado que son personas diferentes.

Clustering - Iteración 12**Método y parámetros**

Se procede a pasar a otra estrategia. Method: keycollision , keying function: cologne-phonetic

Criterio de selección

Empírico. Siguiendo opción dentro de keycollision.

Resultado

- Surge el caso de : Cristina Dejean vs Cristian Dejean > se deja separado dado que tranquilamente pueden ser 2 personas diferentes por sexo , sin embargo puede ser un error de entrada de datos. (se considera poco probable que se pase ese tipo de error hasta nivel paper) y cristian tiene 2 papers vs cristina 13 , podría ser algún familiar. A cotejar con expertos
- Se siguen encontrando casos de apellidos compuestos
- Se siguen encontrando casos de nombres con iniciales

Clustering - Iteración 13**Método y parámetros**

Se procede a validar la estrategia anterior.

Metodo: keycollision , keying function: cologne-phonetic

Criterio de selección

Empírico. Verificando método utilizado en iteración 13, para búsqueda de más resultados.

Resultado

- Se volvió a ejecutar esa estrategia, y solo aparecieron los casos de dejean y otro que no cuadran los apellidos, solo los nombres.

Clustering - Iteración 14

Método y parámetros

Se procede con otra estrategia:

Method: nearest neighbour , distance function: levenshtein , Radius: 1.0, Block Chars: 6

Criterio de selección

Empírico. Siguiendo combinación dentro de la herramienta.

Resultado

- En los casos de :
 - Emma E. Alfaro(20 rows) / Emma L. Alfaro(14 rows)
 - María C. Barboza(11 rows) / María J. Barboza(7 rows)
 - No se si es la misma personas, hay que ver con los expertos.
 - No se hace ese cambio

Clustering - Iteración 15

Método y parámetros

Se procede con otra estrategia: aumentamos el radio a 2

Method: nearest neighbour , distance function: levenshtein , Radius: 2.0, Block Chars: 6

Criterio de selección

Empírico. Variando los parámetros del algoritmo probado en iteración anterior.

Resultado

- Generico, Aparecieron casos que eran de segundos nombres.

Clustering - Iteraciones 16,17,18,19,20,21,22

Método y parámetros

Se procede con otras estrategias de variar el block chars: entre 3 y 10

Method: nearest neighbour, distance function: levenshtein , Radius: 2.0, Block Chars: 3, 10.

Criterio de selección

Empírico. Variando los parámetros del algoritmo probado en iteración anterior.

Resultado

- Sin resultados favorables.

Clustering - Iteraciones 23

Método y parámetros

Se procede con otra estrategia de aumentar el radio nuevamente por 1 punto.

Method: nearest neighbour , distance function: levenshtein , Radius: 3.0, Block Chars: 6

Criterio de selección

Empírico. Variando los parámetros del algoritmo probado en iteración anterior.

Resultado

- Resultados genericos, Muchos resultados favorables.
- Se siguen encontrando casos de apellidos compuestos
- Se siguen encontrando casos de nombres con iniciales
- Hay que revisar muy cuidadosamente los resultados dado que hay mucha similitud entre algunos y son diferentes. Por ejemplo:
 - Jorge A. Suby vs Jorge A. Rey
 - María C. Barboza vs María J. Barboza
 - Queda validar con los expertos.

Clustering - Iteración 24**Método y parámetros**

Se procede a repetir la estrategia

Method: nearest neighbour , distance function: levenshtein , Radius: 3.0, Block Chars: 6

Criterio de selección

Empírico. Repetimos los parámetros anteriores para ver cómo afecta.

Resultado

- Se encontraron algunos casos como por ejemplo: Mariano C. Del Papa vs Mariano Del Papa
- Siguen existiendo casos con los que tener mucho cuidado dada la poca variabilidad pero evidente diferencia
 - Vanina fernandez, Vanesa fernandez
 - Queda validar con los expertos.

Clustering - Iteración 25**Método y parámetros**

Se procede con otra estrategia: para reducir los casos problemáticos y ver si se encuentran otros mejores, bajamos los Block Chars a 3.

Method: nearest neighbour , distance function: levenshtein , Radius: 3.0, Block Chars: 3

Criterio de selección

Empírico. Variando los parámetros del algoritmo probado en iteración anterior.

Resultado

- Resultados genericos.
- Se encontraron casos como el de Ana M. Inda vs Ana Inda
- Siguen existiendo casos con los que tener mucho cuidado dada la poca variabilidad pero evidente diferencia.
- Queda validar con los expertos.

Clustering - Iteración 27

Método y parámetros

Se procede con otra estrategia: para reducir los casos problemáticos y ver si se encuentran otros mejores, bajamos los Block Chars a 2.

Method: nearest neighbour , distance function: levenshtein , Radius: 3.0, Block Chars: 2

Criterio de selección

Empírico. Variando los parámetros del algoritmo probado en iteración anterior.

Resultado

- Siguen existiendo casos con los que tener mucho cuidado dada la poca variabilidad pero evidente diferencia.
- Queda validar con los expertos.

Clustering - Iteración 28

Método y parámetros

Se procede con otra estrategia, cambiar la función de distancia.

Method: nearest neighbour , distance function: PPM , Radius: 3.0, Block Chars: 3

Criterio de selección

Empírico. Variando la función de distancia.

Resultado

- Se encuentran muchos resultados sin embargo sin valor dado que no toma en cuenta los apellidos, agrupa casos de nombres muy comunes como por ejemplo: los nombres que comienzan con 'María X yyyy'
- Sirve mucho para identificar los segundos nombres de varias personas, como por ejemplo > cristina b dejean > cristina beatriz dejean

Clustering - Iteración 29

Método y parámetros

Se procede nuevamente con la estrategia

Method: nearest neighbour , distance function: PPM , Radius: 3.0, Block Chars: 3

Criterio de selección

Empírico. Validando la estrategia de iteración anterior.

Resultado

- Mismas observaciones anteriores, un caso a revisar específico:
 - Héctor Hugo Varela(14 rows) / Hugo H. Varela(8 rows)

Resultado de Revisión y validación con expertos

La [Tabla anx estandarizacion autores 1](#) tiene el mapeo de los autores revisados por expertos y su valor original.

Tabla anx_estandarizacion_autores_1: mapeo de nombres finales - nombres originales.

autor_por_expertos	autor_original
--------------------	----------------

Alberto S. Slepoy	Alberto S. Slepoy
Alberto S. Slepoy	Alberto Slepoy
Alejandro C. Tobisch	Alejandro C. Tobisch
Alejandro C. Tobisch	Alejandro Tobisch
Alfredo I. Prieto	Alfredo I. Prieto
Alfredo I. Prieto	Alfredo Prieto
Christine Keyser	Christine Keyser-Tracqui
Cristina Beatriz Dejean	Cristian Dejean
Cristian Marcelo Crespo	Cristian Crespo
Cristian Marcelo Crespo	Cristian Marcelo Crespo
Emma L. Alfaro	Emma Alfaro
Emma L. Alfaro	Emma E. Alfaro
Emma L. Alfaro	Emma L. Alfaro
Emma L. Alfaro	Emma L. Alfaro Gómez
Fernando V. Ramírez-Rozzi	Fernando Ramírez Rossi
Fernando V. Ramírez-Rozzi	Fernando V. Ramírez Rossi
Fernando V. Ramírez-Rozzi	Fernando Ramirez Rozzi
Fernando V. Ramírez-Rozzi	Fernando Ramírez-Rozzi
Fernando V. Ramírez-Rozzi	Fernando Ramírez Rozzi
Francisco A. Zangrando	Francisco A. Zangrando
Francisco A. Zangrando	A. Francisco Sangrando
Francisco Raúl Carnese	Francisco R. Carnese
Francisco Raúl Carnese	Francisco Carnese
Francisco Raúl Carnese	Francisco Carnese
Francisco Raúl Carnese	Francisco R Carnese
Francisco Raúl Carnese	Francisco R. Carnese
Francisco Raúl Carnese	Francisco Raúl Carnese
Jimena Barbeito-Andrés	Gimena Barbeito
Jimena Barbeito-Andrés	Jimena Barbeito-Andrés
Héctor Mario Pucciarelli	Héctor Mario Pucciarelli
Héctor Mario Pucciarelli	Hector M. Pucciarelli
Héctor Mario Pucciarelli	Héctor M. Pucciarelli

Héctor Mario Pucciarelli	Hector Pucciarelli
Héctor Mario Pucciarelli	Héctor Pucciarelli
Héctor Hugo Varela	Hugo H. Varela
Héctor Hugo Varela	Héctor H. Varela
José A. Castro Ocón	José A. Castro
José A. Castro Ocón	José A. Castro Ocón
Leandro Hernán Luna	Leandro Hernán Luna
Leandro Hernán Luna	Leandro H. Luna
Leandro Hernán Luna	Leandro Luna
Gabriela Lorena L'Heureux	Gabriela L. L'Heureux
Gabriela Lorena L'Heureux	Lorena L'Heureux
Luis Edurado Castro	Luis E. Castro
Luis Edurado Castro	Luis Castro
Luis Edurado Castro	Luis Edurado Castro
María Antonia Luis	María A. Luis
María Antonia Luis	María Luis
María Antonia Luis	María Antonia Luis
María Laura Bergel Sanchís	María Bergel Sanchís
María Laura Bergel Sanchís	María Laura Bergel Sanchís
María Laura Bergel Sanchís	María Laura Bergel
María Laura Bergel Sanchís	Maria L. Bergel
María Laura Bergel Sanchís	Maria Bergel
María Cristina Muñe	Maria C. Muñe
María Cristina Muñe	María C. Muñe
María Cristina Muñe	María Cristina Muñe
María Cristina Muñe	Maria Muñe
María Cristina Muñe	María Muñe
María Dolores Marrodán	Maria D. Madorran
María Dolores Marrodán	Maria Madorran
María Dolores Marrodán	María Dolores Marrodán
María Dolores Marrodán	María Marrodán
María Dolores Marrodán	M. Dolores Marrodán

María Dolores Marrodán	María D. Marrodán
María E. Díaz Sánchez	Maria Díaz
María E. Díaz Sánchez	María E. Díaz Sánchez
María Eugenia Luna	Maria E. Luna
María Eugenia Luna	María Eugenia Luna
María Florencia Cesani	Maria Cesani
María Florencia Cesani	María Cesani
María Florencia Cesani	Maria Cesani Rossi
María Florencia Cesani	María Cesani Rossi
María Florencia Cesani	Maria F. Cesani
María Florencia Cesani	María F Cesani
María Florencia Cesani	María F. Cesani
María Florencia Cesani	Maria F. Cesani Rossi
María Florencia Cesani	María F. Cesani Rossi
María Florencia Cesani	María Florencia Cesani
Marta Ruiz	Marta Ruiz
María Fernanda Torres	María F. Torres
María Fernanda Torres	María Torres
María Fernanda Torres	María Fernanda Torres
María Fucini	María C. Fucini
María Fucini	María Fucini
María Isabel González	María González
María Isabel González	María Isabel González
María Laura Fuchs	María I. Fuchs
María Laura Fuchs	María Laura Fuchs
María Lorena Zonta	María L. Zonta
María Lorena Zonta	Lorena Zonta
María Lorena Zonta	María L Zonta
María Lorena Zonta	Maria L. Zonta
María Lorena Zonta	María Zonta
María Lorena Zonta	Maria Zonta
María M. Ramón	María M. Ramón

María M. Ramón	María Ramon
María M. Reca	María M. Reca
María M. Reca	María Reca
María Soledad Mesa	María Mesa
María Soledad Mesa	María S. Mesa
María Soledad Mesa	María Soledad Mesa
Pedro Hidalgo Calcines	Pedro C. Hidalgo
Pedro Hidalgo Calcines	Pedro Hidalgo Calcines
Sergio Ivan Perez	Sergio I. Perez
Sergio Ivan Perez	Sergio Ivan Perez
Sergio Ivan Perez	Sergio Perez
Sergio Ivan Perez	S. Ivan Perez
Sergio Ivan Perez	S. Perez
Susana Blanca Ocampo	Susana B. Ocampo
Susana Blanca Ocampo	Susana Ocampo
Susana Blanca Ocampo	Susana Blanca Ocampo
María C. Barboza	María J. Barboza
María C. Barboza	María C. Barboza
María E. Díaz	María E. Díaz Sánchez
María E. Díaz	María E. Díaz
María Asunción Bordach	María A. Bordach
María Asunción Bordach	M. Asunción Bordach