UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"IDENTIFICACIÓN DE CONGLOMERADOS EN EL GRAFO DE COAUTORÍAS FORMADO POR LAS INSTITUCIONES PERUANAS CON INVESTIGACIÓN EN MEDICINA INDIZADA EN SCOPUS"

Presentada por:

LUCÍA MÁLAGA SABOGAL

TESIS PARA OPTAR EL GRADO DE MAESTRO

MAGISTER SCIENTIAE EN ESTADÍSTICA APLICADA

Lima - Perú

2017

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"IDENTIFICACIÓN DE CONGLOMERADOS EN EL GRAFO DE COAUTORÍAS FORMADO POR LAS INSTITUCIONES PERUANAS CON INVESTIGACIÓN EN MEDICINA INDIZADA"

TESIS PARA OPTAR EL GRADO DE
MAESTRO MAGISTER SCIENTIAE

Presentada por:
LUCÍA MÁLAGA SABOGAL

Sustentada y aprobada ante el siguiente jurado:

Mg. Jesús Salinas Flores **PRESIDENTE**	Mg. Sc. Clodomiro Miranda Villagómez **PATROCINADOR**
Mg. Sc. Jaime Porras Cerrón **MIEMBRO**	Mg. Sc. Carlos López de Castilla Vásquez **MIEMBRO**

AGRADECIMIENTO

A Francisco Sagasti, por la motivación constante para superarme.

A mi familia por su fe inquebrantable en mi trabajo.

A mi patrocinador por su confianza en la calidad de mi investigación.

A los miembros del jurado por dedicar su tiempo para revisar este texto.

ÍNDICE GENERAL

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

Acrónimos y siglas

Códigos de idiomas

de alemán

en inglés

es español

fr francés

it italiano

pt portugués

Títulos de revistas

Am J Trop Med Hyg	The American Journal of Tropical Medicine and Hygiene
Emerg Infect Dis	Emerging Infectious Diseases
Int J Tuberc Lung Dis	International Journal of Tuberculosis and Lung Disease
J Clin Microbiol	Journal of Clinical Microbiology
J Infect Dis	The Journal of Infectious Diseases
PLos Negl Trop Dis	PLoS Neglected Tropical Diseases
Rev Panam Salud Publica	Revista Panamericana de Salud Pública
Rev Per Med Exp Salud Publica	Revista Peruana de Medicina Experimental y Salud Pública

RESUMEN

Este estudio halló los grupos de investigación conformados por las instituciones peruanas con investigación en medicina indizada en Scopus en base a las coautorías. La información se descargó de Scopus en formato no estandarizado y se utilizó aprendizaje supervisado con k-medias y un conjunto de datos de entrenamiento, para la identificación de las instituciones involucradas. El procesamiento de los datos se hizo con R. Las instituciones identificadas se clasificaron en ocho categorías: universidades, institutos públicos de investigación, clínicas y hospitales, organismos y dependencias del gobierno nacional, organismos y dependencias del gobierno local, empresas, organizaciones internacionales con filiales en Perú, instituciones privadas sin fines de lucro; y dos sectores: público y privado. Posteriormente se identificó los conglomerados existentes utilizando la metodología de particionamiento jerárquico aglomerativo propuesta por Moore, Clauset y Newman e implementada en el paquete igraph en R. Se halló que las instituciones del sector salud tienden a colaborar con sus símiles pero que no existe relación entre el tipo y sector de la institución y los patrones de colaboración para otras instituciones.

Palabras clave: Análisis de redes, Conglomerados, Coautorías, Bibliometría, Cienciometría, Análisis de dominio.

ABSTRACT

This study found the research groups formed by Peruvian institutions with scientific papers published in Scopus, based on papers co-authorship. The information about authors institutional affiliation has been downloaded from Scopus in unstandardized format. K-means supervised learning and a training dataset have been used to identify and normalize the information about the institutions. R has been used for data preprocessing and processing. The identified institutions have been classified in eight kinds: universities, national research institutes, health institutions (clinics and hospitals), institutions that form part or are dependent from the national government, institutions that form part or are dependent from the local government, firms, international organizations that have delegations in Peru, other institutions (most of them non-governmental organizations). Those institutions have been classified also in two sectors: public and private. Once the institutions had been identified, clusters have been found using fast greedy hierarchic partitioning proposed by Moore, Clauset and Newman and implemented in the library igraph in R. The research found that health institutions prefer to write in co-authorship with other health institutions. In other cases, there isn't any kind of relationship between the kind of institution and its election of co-authors.

Keywords: Network analysis, Clusters, Co-authorship, Bibliometrics, Scientometrics, Domain analysis

I. INTRODUCCIÓN

La investigación científica se basa en la colaboración entre los investigadores. Por eso, es de especial importancia la colaboración entre las instituciones y entre los autores. Frecuentemente las investigaciones culminan con la publicación de un artículo en una revista científica. Sin embargo, el estudio de las publicaciones académicas en su papel de indicadores de producción enfrenta ciertas dificultades. Para comenzar, el Perú no cuenta con una base de datos que recopile la producción científica nacional (similar al Publindex colombiano) y las revistas peruanas en las bases de datos internacionales están sub-representadas. Aun tomando en cuenta estas limitaciones, si se desea realizar un estudio de la producción científica nacional a un costo razonable, la utilización de las bases de datos internacionales es la mejor opción. Las bases de datos más importantes que contienen información estructurada, y por ende adecuada para el análisis, son Scopus de la editorial Elsevier y Web of Science (WoS) de Thomson. Sin embargo, para ambas bases de datos, las disciplinas cubiertas en mayor grado son las ciencias de la vida y medicina, en detrimento de otras disciplinas. Para el caso de América Latina, que está infrarrepresentada en WoS, Scopus es una mejor opción por temas de cobertura.

Por otro lado, las diferentes áreas del conocimiento tienden a producir investigaciones y colaborar de diferentes maneras. Los estudios que producen las diferentes disciplinas no siempre tendrán como resultado un artículo académico en habla inglesa, pero son los artículos académicos en inglés los que están más presentes en las bases de datos utilizadas para el análisis. Las ciencias sociales, que con frecuencia tocan temas de interés local, tienden a producir investigación monográfica y tienen un índice de colaboraciones bajo. En matemática teórica se escribe poco, y rara vez en colaboración. En el caso peruano, la medicina es una disciplina que se presta al análisis bibliométrico y de colaboraciones pues es la que tiene mayor cantidad de documentos en Scopus. Además, es una disciplina en la que se tiende a escribir en colaboración, como tal, es la más interesante para analizar en cuanto a la presencia de colaboraciones interinstitucionales y de colegios invisibles.

Conocer estas colaboraciones nos permite saber un poco sobre las instituciones encargadas de la investigación en los sistemas de innovación. El concepto de «sistemas nacionales de innovación» (también llamados «sistemas de innovación» o «sistemas de innovación tecnológica») se consolida durante los decenios de 1980 y 1990, con trabajos de Freeman y Lundvall (Freeman 1993 [1987], Lundvall 1988, 1992, National Innovation Systems: A Comparative Analysis 1993). Este concepto reconoce la complejidad e importancia de diferentes actores involucrados en la innovación y, en términos sencillos, está constituido por «el conjunto de entidades privadas, públicas, académicas y de la sociedad civil involucradas en la creación, difusión y utilización del conocimiento y la tecnología; sus interrelaciones e interacciones, las estructuras institucionales y los incentivos y reglas del juego que las condicionan; y los beneficios y ventajas que generan en la producción de bienes y la provisión de servicios (Sagasti 2013: 44)».

En este conjunto de elementos las entidades productoras de conocimiento científico y tecnológico y las relaciones entre estas juegan un rol importante.

En esta tesis se analizará sólo un fragmento de esta compleja red de elementos e interrelaciones llamada sistema de innovación peruano, específicamente el más representativo en términos de volumen de documentos producidos en revistas indizadas, es decir la producción científica peruana del área de medicina. La investigación analizará las interrelaciones de coautoría de las instituciones peruanas que tengan al menos un documento indizado en Scopus[1] desde el 2005 hasta el 2015 y buscará identificar, en el grafo que representa estas coautorías, los grupos de investigación formados por las instituciones peruanas.

La información recuperada es de arriba-abajo (top-down), sin una entrevista a cada una de las instituciones sobre su investigación indizada en bases de datos internacionales. La recuperación se realizó seleccionando de Scopus los documentos con participación de por lo menos un autor afiliado a una institución peruana. Esto puede significar que existe alguna investigación publicada por investigadores asociados a ciertas instituciones, pero que no han firmado colocando la filiación correspondiente, también puede haberse obviado artículos que por errores de procesamiento de la base de datos no aparecieran con la afiliación de país

[1] No se tomó en cuenta la producción indizada en ISI Web of Science ya que esta en gran medida es repetitiva con la incluida en Scopus. Si bien pueden existir diferencias en los resultados finales estas serán más visibles en el caso de actores pequeños, con poca producción. No se encontró una fuente de información que compare la cobertura de ambas bases de datos para el Perú pero Lucio-Arias (2013) presentó un resultado de fuerte yuxtaposición en Colombia.

correspondiente a Perú. Por último, puede haber algún caso en el que por error de procesamiento humano en el momento de estandarizar las afiliaciones para su conteo alguna de estas no se haya identificado correctamente.

Las instituciones se clasifican por tipología institucional, es decir asignación sectorial, división basada en el manual de Frascati para indicadores de ciencia y tecnología (OECD 2002) y adaptada por Málaga (2014).

1.1. JUSTIFICACIÓN DE LA INVESTIGACIÓN

La presente investigación permitió identificar grupos no formales de investigación formados por las instituciones peruanas, las relaciones entre estos, y las centralidades de sus elementos. Esta información es de utilidad para la elaboración de los instrumentos de política y distribución de recursos, ya que los grupos pueden compartirlos. Así, los elementos de la red con mayor centralidad tendrían la mayor capacidad de compartir acervos, por lo que proporcionárselos (acompañados de los convenientes instrumentos de política para asegurar la distribución) sería una inversión adecuada. Se ha demostrado que la posición de un vértice en una comunidad puede afectar el papel o función que cumplen. Así por ejemplo, en los estudios de redes sociales se ha encontrado que los individuos que conectan a elementos que de otra manera estarían inconexos (intermedian), tienen un alto nivel de influencia sobre el flujo de la información entre los grupos (Granovetter 1973, Burt 1976, Freeman 1977).

Adicionalmente, se obtendrá información sobre la centralidad de los actores en los grupos de investigación, característica que puede compararse a la infraestructura de las instituciones, y en consecuencia plantear políticas para el aprovechamiento de estas estructuras.

Entre las tareas asociadas a la identificación de conglomerados está la elección de una metodología para delimitarlos (Fortunato 2010, Clauset et al. 2004, Kolaczyk 2009, Kolaczyk y Csárdi 2014) y la identificación de las instituciones de manera única. Aunque la segunda tarea para obtener buenos resultados debe hacerse de manera semi-supervisada, puede servirse de metodologías de conglomerados para facilitar el trabajo (Cuxac et al. 2013, Van der Loo 2014).

1.2. OBJETIVOS DE LA INVESTIGACIÓN

Objetivo general: Identificar los conglomerados en el grafo de coautorías de instituciones peruanas con investigación en medicina indizada en Scopus mediante particionamiento jerárquico.

Objetivos específicos:

- Utilizar métodos de clasificación para identificar de manera única las afiliaciones institucionales.
- Describir la red y los elementos que la componen.
- Describir el perfil de los conglomerados.

1.3. ALCANCES DE LA INVESTIGACIÓN

Esta es una investigación de carácter descriptivo ya que delinea los elementos de un sistema.

II. REVISIÓN DE LITERATURA

2.1. ELEMENTOS DEL SISTEMA DE INNOVACIÓN

El 23 de julio del año 2004 se promulgó la Ley Marco de Ciencia, Tecnología e Innovación Tecnológica (Ley N.° 28303) que establece en su artículo séptimo el Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica (SINACYT) compuesto por el conjunto de instituciones y personas naturales del país, dedicadas a la Investigación, Desarrollo e Innovación Tecnológica (I+D+I) en ciencia y tecnología y a su promoción. La ley anuncia de manera no limitativa los actores de este sistema:

a) «El Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC), como organismo rector del SINACYT.

b) El Fondo Nacional de Desarrollo de la Ciencia, Tecnología e Innovación Tecnológica (FONDECYT), para el fomento de los planes, programas y proyectos del SINACYT.

c) El Consejo Consultivo Nacional de Investigación y Desarrollo para la CTel [Ciencia, Tecnología, Innovación], (CONID), como órgano consultivo multidisciplinario e intersectorial del SINACYT.

d) Las instancias de los Gobiernos Regionales y Locales dedicadas a las actividades de CTel en sus respectivas jurisdicciones.

e) Las universidades públicas y privadas, sector empresarial, programas nacionales y especiales de CTel, instituciones e integrantes de la comunidad científica.

f) El Instituto Nacional de Defensa de la Competencia y de la Protección de la Propiedad Intelectual - INDECOPI, para la protección y difusión de los derechos intelectuales en CTel, y el registro y difusión de las normas técnicas y metrológicas.

g) Las comunidades campesinas y nativas, como espacios activos de preservación y difusión del conocimiento tradicional, cultural y folclórico del país (Ley 28303. Ley Marco de Ciencia, Tecnología e Innovación Tecnológica 2004)».

En suma, el SINACYT tendría una estructura bastante compleja ya que incluye también los Institutos públicos de investigación, las instancias del gobierno regional y local y los fondos temporales de la promoción de la CTI.

Sin embargo, la promulgación de una ley no es suficiente para crear un sistema de innovación operativo. Según Bazán y otros (2013: 157) estos son algunos de los elementos con los que un sistema nacional de innovación debería contar para lograr una actividad coordinada:

- «**Organizaciones generadoras de conocimiento** en el sistema educativo y de capacitación, así como aquellas dedicadas específicamente a la investigación científica y tecnológica;
- **Empresas productivas y de servicios que realizan innovaciones** incorporando tecnología y conocimiento en sus actividades, sea ya en forma individual u operando de manera conjunta en redes;
- **Organizaciones y entidades públicas, privadas o de la sociedad civil** que prestan servicios (información, normas, asistencia técnica, gestión tecnológica, asesoría financiera) a las unidades productivas y de servicios que realizan innovaciones;
- **Instituciones y agencias públicas que establecen políticas** para los sectores productivos y sociales, la ciencia y tecnología, y el marco de regulación, todas las cuales condicionan y afectan el proceso de innovación;
- **Entidades que proporcionan la infraestructura física** (transportes, telecomunicaciones, energía, agua y saneamiento) que constituye el soporte material para la innovación que realizan las unidades productivas y de servicios;
- **Entidades que ayudan a crear un ambiente favorable** para la ciencia, la tecnología y la innovación, realizando actividades tales como: proporcionar acceso al acervo mundial de conocimientos, promover y difundir la ciencia, y fomentar la toma de decisiones basadas en evidencias empíricas, así como medidas para garantizar la transparencia en el ejercicio de las funciones públicas y la actividad privada, y las prácticas democráticas».

Para esta investigación los objetos de análisis son las organizaciones generadoras de conocimiento. Estas organizaciones pueden agruparse según diferentes criterios. Un criterio muy importante es el administrativo y de financiamiento. Siguiendo esta división se tiene

por un lado las instituciones del sector público, financiadas por el Estado y las instituciones del sector privado, las cuales tienen financiamiento variado, en parte público, a través de concursos, en parte propio, en parte obtenido a través de organizaciones internacionales.

Cabe mencionar que en el manual de Frascati (OECD 2002), que da pautas para la recuperación de información estadística sobre actividades de I+D se sugiere seguir una división sectorial de las instituciones de acuerdo a su administración y financiamiento de la siguiente manera:

- empresas,
- administración pública,
- instituciones privadas sin fines de lucro,
- enseñanza superior,
- extranjero - organizaciones internacionales.

Sin embargo, tal división tiene la desventaja de asumir los hospitales como parte de los diferentes sectores dependiendo de su fin: lucro, enseñanza, o administración (dependencia de la administración pública o no). Al ser los hospitales y clínicas importantes centros productores de conocimiento, especialmente en medicina, en varios informes se ha decidido separar los hospitales y clínicas como un sector adicional (Salazar et al. 2011, Bravo 2006, Gómez et al. 2004, Moya et al. 2006, 2004).

2.1.1. INSTITUCIONES INVESTIGADORAS: IMPORTANCIA DE LAS UNIVERSIDADES

En el Perú las instituciones dedicadas a la investigación son sobre todo las universidades, acompañadas de los institutos públicos de investigación, las organizaciones no gubernamentales y las filiales en el Perú de organizaciones internacionales. Al contrario de lo que ocurre en los sistemas de innovación consolidados,[2] la investigación financiada por las empresas en nuestro país es mínima. En general en América Latina destaca el papel de las universidades en I+D, ya que, como explica la OEI «las universidades latinoamericanas ejecutan el 36,6 por ciento de la I+D regional» lo cual es muy superior al promedio de los

[2]En Estados Unidos el sector privado cubre el 70 por ciento del total de recursos destinados a la investigación (NSF 2012: cap. 4)

países de la OCDE (17,1%), Estados Unidos (14,3%) o la Unión Europea (UE-27) (22,1%) e incluso Sudáfrica (19.3%). En consecuencia, una estrategia para la ciencia y tecnología en el continente requiere potenciar las capacidades en I+D de los centros universitarios de excelencia y ampliar las oportunidades educativas para todos los sectores de la población (OEI 2014: 58).

Aunque en el Perú el mayor volumen de la investigación esté concentrado en las universidades, la tendencia a nivel mundial no siempre es esa, por ejemplo, ya en 1996 se destacaba la importancia de otros tipos de instituciones, principalmente hospitales y empresas industriales, en la producción científica del Reino Unido (Hicks y Katz 1996). Entre las universidades Moed (2006) diferencia por su volumen de investigación dos tipos: las universidades de investigación y las universidades de formación profesional; y por su grado de especialización también dos tipos: las universidades generalistas y las especializadas, aunque declaró que la frontera entre estas categorías es discutible. En el ranking realizado entre las universidades con más producción, y habiéndolas clasificado según su especialización o falta de ella, pudo determinar que la especialización de una universidad no implicaba que esta consiguiera un superior impacto que las universidades generalistas, salvo para las especialidades de: ciencias biológicas dedicadas a humanos, medicina clínica, biología molecular y bioquímica, y física.

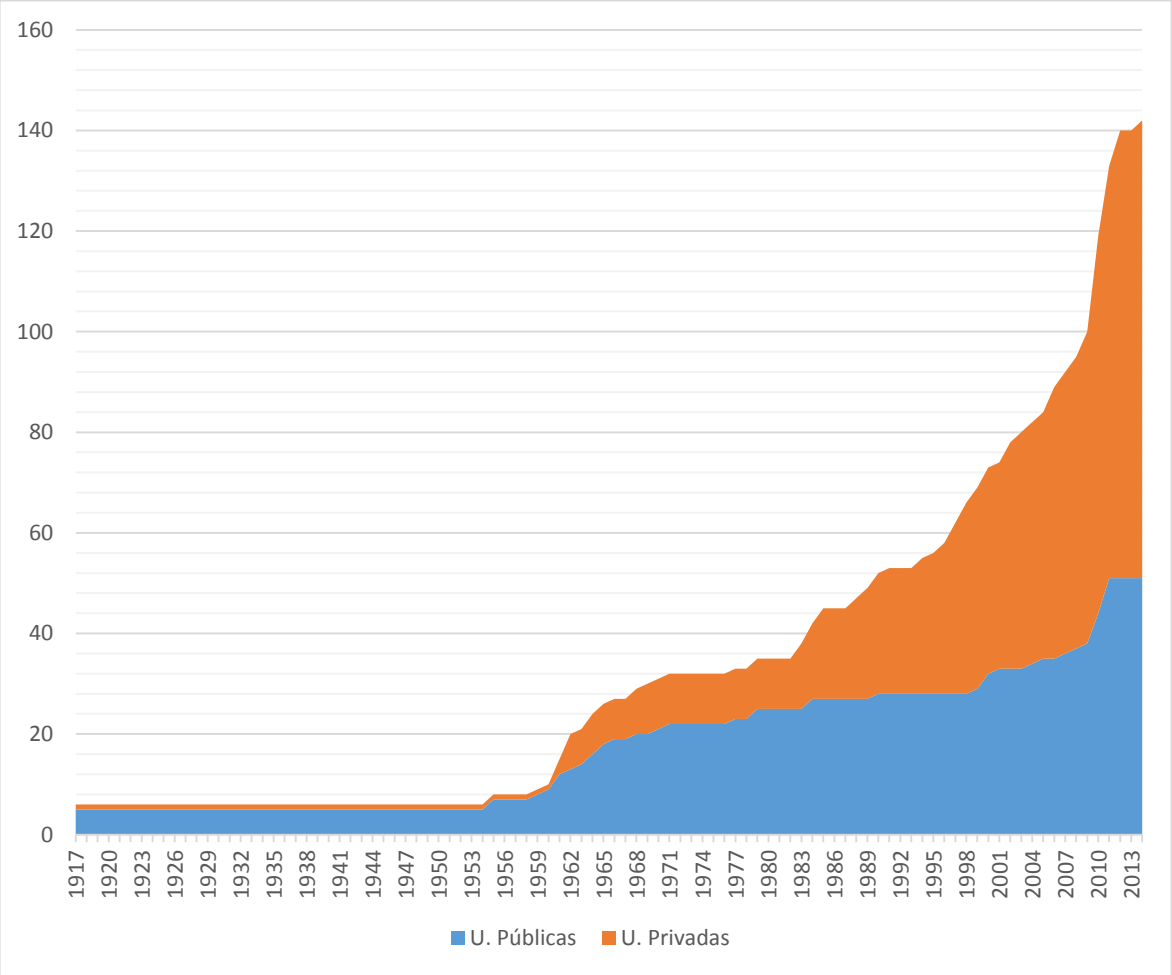
En el marco legal vigente del Perú existen tres tipos de universidades: las universidades públicas, las universidades privadas sin fines de lucro y las universidades privadas con fines de lucro – figura legal creada con el Decreto Legislativo 882 de 1996. La promulgación del DL 882 permitió un crecimiento acelerado de las universidades, que aumentaron de 56 en el año 1995, a 142 en diciembre del 2016 (ver figura 1).

Si bien el incremento de la oferta universitaria es parte de una tendencia mundial, en el caso del Perú el aumento de la cantidad no fue acompañado de mecanismos de control de calidad. Además, entre las universidades empresa existe el peligro (y en varios casos este se materializa) de ofrecer carreras que no requieren mayor equipamiento ni infraestructura, además de contratar a profesores a tiempo parcial o en el cargo de auxiliares, sin darles las facilidades para la investigación (MINEDU 2006). En la figura 2 puede verse que en el 2010 el 80 por ciento de los profesores de universidades privadas tenían estatus laboral de contratados, mientras en las públicas esta categoría la ostentaba sólo el 39 por ciento de los docentes.

Con la formación de la Superintendencia Nacional de Universidades se han dado los primeros, aunque aún tímidos, intentos de control sobre la calidad de la formación ofrecida. Si bien las acciones de esta institución pueden tildarse de reglamentaristas son un paso en dirección a la vigilancia de la calidad.

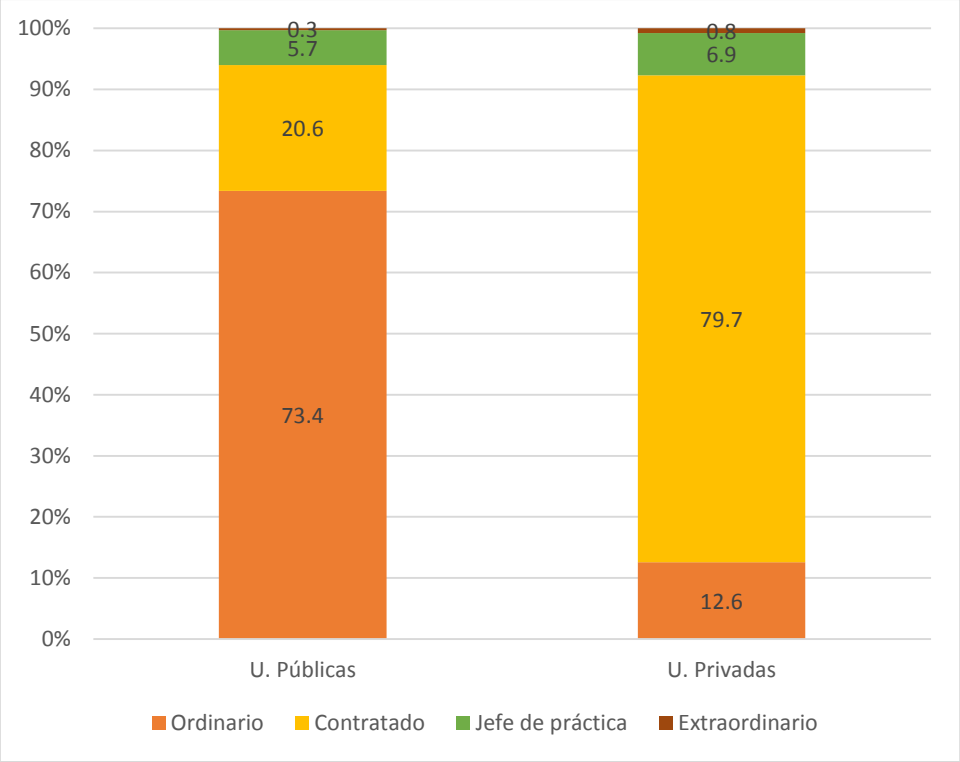
Aun así, en este momento en el Perú sólo existe una docena de universidades con actividades reales de investigación.

Figura 1: Perú. Número acumulado de universidades según sector público o privado (1917 – 2014).



FUENTE: SUNEDU (2017). Elaboración propia.

Figura 2: Perú. Porcentaje de docentes universitarios, por condición laboral, según tipo de universidad (2010).



FUENTE: INEI (2011). Elaboración propia

2.1.2. INSTITUTOS PÚBLICOS DE INVESTIGACIÓN

Otro de los actores del SINACYT son los institutos públicos de investigación (IPI). La información para esta sección procede principalmente del informe de Advancis sobre los IPIs, elaborado en el año 2012 para el Programa de Financiamiento de Ciencia y Tecnología (Lemola et al. 2012).

Los IPIs son organismos autónomos (salvo INICTEL) y forman parte de los siguientes sectores del poder ejecutivo:

- Sector Agricultura: Instituto Geofísico del Perú (IGP), Instituto Nacional de Innovación Agraria (INIA)
- Sector Defensa: Comisión Nacional de Investigación y Desarrollo Aeroespacial (CONIDA), Instituto Geográfico Nacional (IGN).
- Sector Producción: Instituto del Mar del Perú (IMARPE), Instituto Tecnológico Pesquero del Perú (ITP)

- Sector Salud: Instituto Nacional de Salud (INS)
- Sector Energía y Minas: Instituto Geológico Minero y Metalúrgico (INGEMMET), Instituto Peruano de Energía Nuclear (IPEN)
- Universidad Nacional de Ingeniería: Instituto Nacional de Investigación y Capacitación de Telecomunicaciones del Perú - INICTEL (Es el único instituto dependiente de una universidad)
- Multisectorial: Instituto de Investigaciones de la Amazonía Peruana (IIAP)

Los institutos públicos de investigación (IPI) peruanos forman un grupo heterogéneo, con recursos reducidos y limitaciones para la contratación de personal especializado y con funciones varias, entre las cuales la investigación no siempre es la más importante. Los IPIs son actores importantes en el sistema de investigación peruano, al emplear a 2779 personas. A continuación, puede verse un resumen de la situación de los IPIs al 2012.

Cuadro 1: Situación de los institutos públicos de investigación al 2012 (resumen)

Aspecto	*Situación*
Colaboración en investigación con entidades extranjera	A pesar de contar con varios acuerdos internacionales casi no se ejecuta colaboración en investigación con entidades extranjeras
Colaboración nacional	Los IPIs no colaboran entre ellos y colaboran sólo mínimamente con las universidades y empresas, esto provoca inversiones y actividades duplicadas o superpuestas, además de doble inversión en el equipamiento tecnológico.

Continuación

Aspecto	*Situación*
Marco legal	Los IPIs se establecen mediante diferentes instrumentos legales: leyes, decretos ley y decretos legislativos. Sus funciones son muy variadas y no siempre corresponden a una institución de investigación científica. Un ejemplo es el INIA que es la Autoridad del Sistema Agrícola de Innovación.
Funciones	a) Organización de investigación b) Organismo regulación y monitoreo c) Agencia de implementación pública d) Agencia de promoción y divulgación e) Coordinador público (entre autoridades) f) Agencia de apoyo técnico g) Organismo ambientalista y de conservación h) Organización de capacitación i) Representación del Perú ante la cooperación internacional j) Productor de bienes y servicios industriales k) Organización consultora y asesora
Relación con los ministerios del sector al que corresponden	La relación de los IPIs con los ministerios del sector al que corresponden se suele limitar a interacciones sobre temas presupuestarios y casi nunca incluyen una orientación estratégica. En consecuencia, las actividades de los IPIs no cumplen adecuadamente sus funciones dentro del sistema de innovación.

Continuación

Aspecto	*Situación*
Financiamiento	El financiamiento es insuficiente para cumplir con sus misiones (que no se limitan a la investigación), proveniente principalmente de fondos públicos (a excepción del ITP, que genera más del 60 por ciento de su presupuesto con ventas). Además, se destina en la mayor parte a costos administrativos. El financiamiento más positivo para investigación fue el proveniente del FINCyT y de INCAGRO.[3]
Recursos humanos	Las restricciones presupuestarias y la utilización de los contratos CAS han generado que la contratación de nuevos talentos esté a niveles muy bajos. Además, la imposibilidad de ascender al personal ha generado renuncias del personal más calificado. Mientras tanto el personal está envejeciendo, llegando a ser muy alto el promedio de edad de los investigadores CAP en el IPEN (55), IMARPE (53) e INIA (55). Sólo en el IGN el promedio de edad es inferior a 40 años. Además, la calificación del personal es baja y el personal administrativo es muchas veces superior en número al personal de investigación.
Propiedad intelectual	La protección y comercialización de los resultados de investigación es una práctica que aún falta introducir en los IPIs (sólo ITP tiene dos patentes nacionales).

[3] El Programa Incagro ya cerró

Continuación

Aspecto	*Situación*
Investigación	Las funciones de muchos IPIs son tan amplias que varios tienen poco que ver con la investigación en el sentido concreto de la palabra. Los IPIs se dedican principalmente a proveer servicios, a tareas administrativas y a la recolección, procesamiento y distribución de la información.
Aspectos regionales	Salvo IIAP todos los IPIs tienen su sede central en Lima, la presencia en las regiones es limitada.

FUENTE: Lemola et al. (2012). Elaboración propia.

Entre los IPIs el único con investigación importante en salud es el Instituto Nacional de Salud (INS). Sobre la situación de éste, el informe de Advancis (Lemola et al. 2012) detalla lo siguiente:

Financiamiento. (1) Aunque más del 90 por ciento del financiamiento es público existe la oportunidad de obtener fuentes de financiamiento externo de la venta de productos y servicios. (2) Las actividades reguladoras son generalmente gratuitas para el cliente.

Recursos humanos. (1) Personal: 900 empleados, cinco investigadores con doctorado, 20 investigadores con magister. (2) Hay muy pocos investigadores con PhD y no existen incentivos para desarrollar capacidades de doctorados. (3) En el Perú no existen programas de capacitación y calificación relevantes

Desempeño. (1) Resultados de investigaciones de buena calidad y alto efecto en campos científicos seleccionados. (2) Desempeño bastante bueno en un amplio número de responsabilidades donde no hay otros actores nacionales. (3) Las actividades de investigación y desarrollo son marginales, opacadas por las funciones públicas reguladoras y otras.

2.1.3. EL SISTEMA DE SALUD

Los establecimientos de salud son los «que realizan atención de salud en régimen ambulatorio o de internamiento, con fines de prevención, promoción, diagnóstico, tratamiento y rehabilitación, para mantener o restablecer el estado de salud de las personas (MINSA 2011: 5.1).»

Como explican Alcalde-Rabanal et al. (2011) las instituciones del sistema de salud del Perú se pueden dividir en dos sectores: el público y el privado. En la figura que sigue puede visualizarse los actores, los beneficiados y los financiamientos que a continuación explicamos.

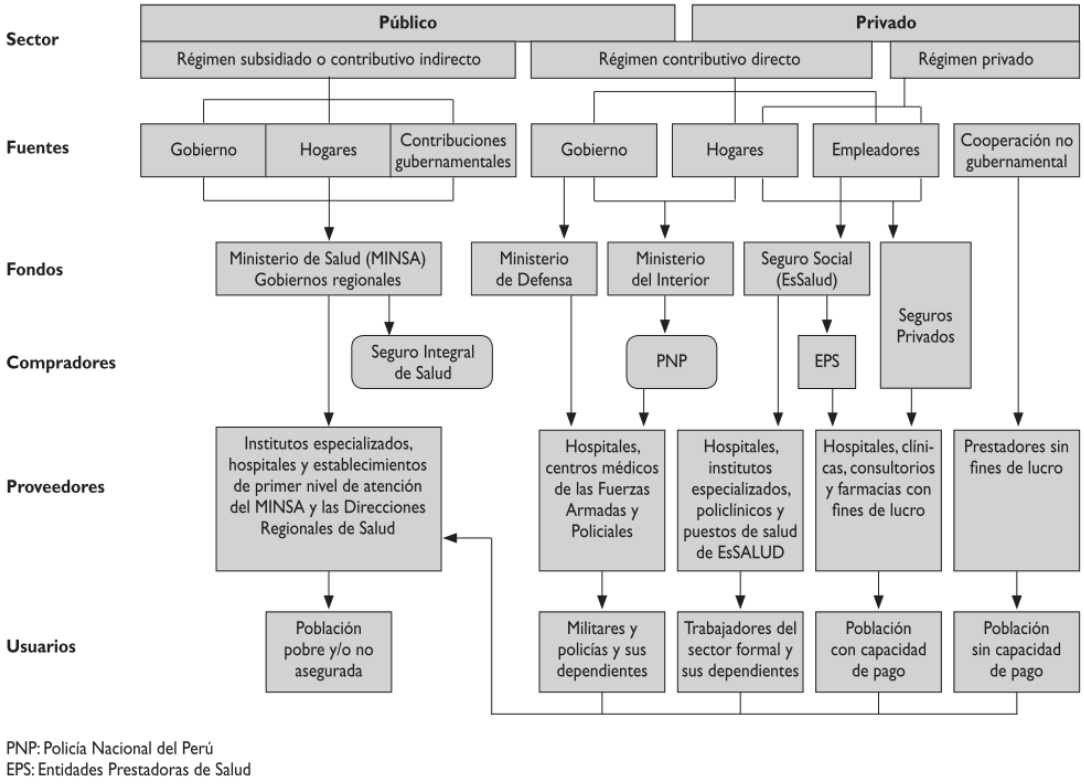
El sector *público* se divide en dos de acuerdo a su régimen contributivo:

- *El régimen subsidiado o contributivo indirecto.* En este régimen se otorga servicios a la población abierta[4] o asegurada al Seguro Integral de Salud (SIS). Los servicios de salud a la población abierta se dan a cambio del pago de una cuota de recuperación de montos variables sujetos a la discrecionalidad de las organizaciones. Los servicios de los afiliados al Seguro Integral de Salud (SIS), son pagados por esta institución para la población que vive en condiciones de pobreza y pobreza extrema. Ambos servicios se realizan a través de la red de establecimientos del Ministerio de Salud (MINSA): hospitales e institutos especializados.

- *El régimen contributivo directo* corresponde a la seguridad social y es cubierto con los aportes de los asegurados. Este se divide en dos subsistemas: (1) el seguro social con provisión tradicional (EsSalud) y (2) la provisión privada (EPS). EsSalud cuenta con instalaciones propias en las que ofrece servicios de salud. Adicionalmente, el sector privado le vende servicios de salud a EsSALUD a través de las Entidades Prestadoras de Salud (EPS). Adicionalmente a los dos subsistemas ya mencionados, el personal del ejército y de la policía y sus familias cuentan con un propio subsistema de salud integrado por hospitales y centros médicos de las Fuerzas Armadas y Policiales.

[4] Población que no tiene que cumplir ninguna condición en especial para ser atendida, salvo cubrir el costo de atención solicitado por la institución prestadora del servicio de salud.

Figura 3: El sistema de salud en el Perú



FUENTE: Alcalde-Rabanal et al. (2011: s244)

El sector *privado* se divide en dos de acuerdo a sus fines de lucro:

1. *El sector privado lucrativo*. Compuesto por: (1) Entidades prestadoras de Salud (EPS); (2) Aseguradoras privadas; (3) Clínicas privadas especializadas y no especializadas; (4) Centros médicos; (5) Policlínicos; (6) Consultorios médicos y odontológicos; (7) Laboratorios; (8) Servicios de diagnóstico por imágenes; (9) Establecimientos de salud de empresas; (10) Proveedores de medicina tradicional.

2. *El sector privado no lucrativo.* Alcalde-Rabanal et al. (2011) explican que la mayor parte de las instituciones no lucrativas prestan servicios de primer nivel (atención médica muy básica) o bien como acción secundaria (su actividad principal no es prestar servicios de salud).

Por otro lado, el Ministerio de Salud dispuso una categorización de los establecimientos de salud dependiendo de la complejidad de atención que brindan, que se detalla en el cuadro que sigue. Entre las categorías señaladas en el Cuadro 2, sólo los establecimientos de la categoría III-2, es decir los institutos especializados, tienen como prerrequisito para

pertenecer a esta categoría el tener una unidad especializada en actividades de investigación y docencia.

Cuadro 2: Categorías de establecimientos de salud en el Perú

Nivel de atención	Categoría	MINSA	EsSalud	PNP	FAP	Naval	Privado
1	I -1	Puesto de salud		Puesto sanitario		Enfermería Servicios de sanidad	Consultorio
	I -2	Puesto de salud con médico	Posta médica	Posta médica	Posta médica	Departamento de sanidad posta naval	Consultorio médico
	I -3	Centro de salud sin internamiento	Centro médico	Policlínico B	Departamento sanitario		Policlínico
	I -4	Centro de salud con internamiento	Policlínico			Policlínico naval	Centro médico
2	II -1	Hospital I	Hospital I	Policlínico A	Hospital zonal	Clínica naval	Clínica
	II-2	Hospital II	Hospital II	Hospital regional	Hospital regional		Clínica
	II-E	Hospital especializado					Clínica especializada
3	III-1	Hospital III	Hospital III y IV	Hospital nacional	Hospital Central FAP	Hospital Naval Buque Hospital	Clínica
	III-E	Hospital especializado					Clínica especializada
	III -2	Instituto especializado	Instituto				Instituto

Fuente: Salaverry y Cárdenas-Rojas (2009: 266). Adaptación propia a la Norma Técnica vigente (MINSA 2011).

2.1.4. ADMINISTRACIÓN PÚBLICA

Según la Ley del Procedimiento Administrativo General (Ley 27444. Ley del Procedimiento Administrativo General 2001: art. 1) comprenden la administración pública las siguientes instancias:

- «El Poder Ejecutivo, incluyendo Ministerios y Organismos Públicos Descentralizados;
- El Poder Legislativo;
- El Poder Judicial;
- Los Gobiernos Regionales;
- Los Gobiernos Locales;
- Los Organismos a los que la Constitución Política del Perú y las leyes confieren autonomía.

- Las demás entidades y organismos, proyectos y programas del Estado, cuyas actividades se realizan en virtud de potestades administrativas y, por tanto, se consideran sujetas a las normas comunes de derecho público, salvo mandato expreso de ley que las refiera a otro régimen; y
- Las personas jurídicas bajo el régimen privado que prestan servicios públicos o ejercen función administrativa, en virtud de concesión, delegación o autorización del Estado, conforme a la normativa de la materia.»

Por otro lado, el Ministerio de Economía y Finanzas (MEF) organiza las ejecuciones presupuestales de los diferentes niveles de la administración pública de la siguiente manera:

a) Gobierno Nacional. Incluye poder ejecutivo, legislativo y judicial.
b) Gobiernos Locales
c) Gobiernos Regionales

2.1.5. INSTITUCIONES PRIVADAS SIN FINES DE LUCRO

Según el Manual de Frascati (OECD 2002: 3.6.1) las instituciones privadas sin fines de lucro son:

- Las instituciones privadas sin fines lucro, que están fuera del mercado y al servicio de los hogares (es decir, del público).
- Los particulares y los hogares.

Por otro lado, el Código Civil Peruano contempla tres tipos de personas jurídicas sin fines de lucro: las asociaciones, las fundaciones y los comités. Estas se definen de la siguiente manera:

- La *asociación* es «una organización estable de personas naturales o jurídicas, o de ambas, que a través de una actividad común persigue un fin no lucrativo (Código Civil. DL No 295 1984: art. 80).»
- Las *fundación* es «una organización no lucrativa instituida mediante la afectación de uno o más bienes para la realización de objetivos de carácter religioso, asistencial, cultural u otros de interés social (Código Civil. DL No 295 1984: art. 99).»
- El *comité* es «la organización de personas naturales o jurídicas, o de ambas, dedicada a la recaudación pública de aportes destinados a una finalidad altruista (Código Civil. DL No 295 1984: art. 111).»

- Por último, otro tipo de personas que pueden asociarse con las personas jurídicas no lucrativas (aunque legalmente tienen otro estatus) son las *comunidades campesinas y nativas* que «son organizaciones tradicionales y estables de interés público, constituidas por personas naturales y cuyos fines se orientan al mejor aprovechamiento de su patrimonio, para beneficio general y equitativo de los comuneros, promoviendo su desarrollo integral (Código Civil. DL No 295 1984: art. 134).»

De estas cuatro categorías institucionales son principalmente las asociaciones las que producen investigación médica.

2.1.6. EMPRESAS

Según el Manual de Frascati esta categoría comprende:

- «Todas las empresas, organismos e instituciones cuya actividad principal consiste en la producción mercantil de bienes y servicios (exceptuando la enseñanza superior) para su venta al público, a un precio que corresponde al de la realidad económica;
- Las instituciones privadas sin fines de lucro, que están esencialmente al servicio de las empresas (OECD 2002: 3.4.1).»

En el Perú la Ley General de Sociedades, explica que «quienes constituyen la Sociedad convienen en aportar bienes o servicios para el ejercicio en común de actividades económicas (Ley 26887. Ley general de sociedades 1997: art. 1).» La ley especifica los tipos de sociedades, que son las siguientes:

- Sociedad anónima cerrada,
- Sociedad anónima abierta,
- Sociedad colectiva,
- Sociedad en comandita,
- Sociedad comercial de responsabilidad limitada,
- Sociedad civil ordinaria,
- Sociedad civil de responsabilidad limitada.

2.1.7. SECTOR EXTRANJERO

Según el Manual de Frascati esta categoría comprende:

- «Todas las instituciones e individuos situados fuera de las fronteras políticas de un país, excepto los vehículos, buques, aeronaves y satélites espaciales utilizados por instituciones nacionales y los terrenos de ensayo adquiridos por estas instituciones.
- Todas las organizaciones internacionales (excepto empresas) cuyas instalaciones y actividades están dentro de las fronteras de un país (OECD 2002: 3.8.1).»

2.2. K-VECINOS MÁS CERCANOS

La clasificación es la acción de ordenar o disponer por clases. En la clasificación de datos, existe una variable categórica objetivo (variable explicada), que es dividida en predeterminadas clases o categorías (por ejemplo: ingreso alto, ingreso medio e ingreso bajo). Utilizando alguno de los métodos de clasificación existentes se examina un conjunto de registros, cada uno de ellos con información sobre la clase a la que pertenece la variable y una serie de variables predictoras (cualitativas y/o cuantitativas). El algoritmo examinará un conjunto de datos de entrenamiento que contiene tanto los valores de las variables predictoras como las clases a las que pertenecen y de esa manera «aprenderá» qué combinaciones de variables están asociadas a la pertenencia a determinadas clases. Este es el conjunto de datos de entrenamiento. Basándose en la información de los datos de entrenamiento, se asignará clasificaciones a los registros nuevos. Existen numerosos métodos estadísticos para poder identificar la clase a la que pertenecerá un elemento en base a las características de éste. (Larose y Larose 2014)

Entre estos el método K-nn (vecinos más cercanos) es uno de los más sencillos, es no paramétrico y robusto frente a la existencia de outliers. Para su aplicación, como para cualquier método de clasificación supervisada, es necesario que:

a) Exista una muestra de entrenamiento que contenga un conjunto de observaciones para cada vector de variables explicativas $X_1 \dots X_p$ y la variable respuesta Y.

b) Exista una observación *C* que tiene asignado un vector de variables explicativas $X_1 \dots X_p$ y para la cual queremos pronosticar el valor de la variable cualitativa respuesta Y(clase).

El algoritmo consiste en: (1) la comparación de los valores de las variables explicativas para la observación *C* con los valores de estas variables para cada una de las observaciones en el conjunto de entrenamiento, (2) la selección de k (número definido de antemano) observaciones más cercanas a *C* del conjunto de entrenamiento, (3) promediar los valores (o encontrar la mediana) de la variable explicada para las observaciones escogidas lo cual nos permite obtener la predicción.

La definición de «la observación más cercana» en el punto 2 se reduce a la minimización de una métrica escogida, que mide la distancia entre los vectores de las variables explicativas de dos observaciones. Usualmente se utiliza la distancia euclidiana o de Mahalanobis.

Este método es adecuado cuando la dependencia entre las variables explicativas y la variable respuesta es compleja o atípica y por lo tanto difícil de modelar de manera clásica. En los casos en los que la dependencia entre las variables explicativas y la variable respuesta es lineal y el conjunto no contiene outliers los métodos clásicos pueden dar mejores resultados.

2.3. GRAFOS

Los grafos son una conceptualización que permite representar la interacción entre dos elementos. El primero en definir las relaciones entre los términos en forma de un grafo fue Leonherd Euler en el conocido artículo que analiza la problemática de los puentes de Königsberg (Euler 1741). La pregunta planteada era: «Dado el mapa de Königsberg, con el río Pregel dividiendo el plano en cuatro regiones distintas, que están unidas a través de siete puentes, ¿es posible dar un paseo comenzando desde cualquiera de estas regiones, pasando por todos los puentes, recorriendo sólo una vez cada uno, y regresando al mismo punto de partida?». Euler lo resolvió abstrayendo los elementos del problema al enfocarse sólo en las regiones terrestres y las conexiones entre estas. De esa manera generó un primer «grafo», una abstracción que le permitió responder que tal recorrido no es posible. Este fue el punto de partida para el desarrollo de la teoría de grafos que hoy, gracias al incremento de las capacidades computacionales que permiten analizar redes complejas con centenares y miles de elementos, ha adquirido especial fuerza. Siguiendo a Gross (Handbook of graph theory 2014) a continuación se listan las definiciones de los términos que se utilizarán más adelante.

Definición 1: Un grafo G es un par $(V; E)$, donde V y E son conjuntos, junto con una aplicación

$$\gamma_G : E \rightarrow \{\{u, v\}: u, v \in V\}$$

- Al conjunto V se le llama conjunto de vértices; al conjunto E: conjunto de lados o aristas, y a la aplicación γG: aplicación de incidencia. Cada arista tiene uno o dos vértices asociados a ella que se llaman puntos finales (endpoints).

Notación: V_G y E_G o (V(G) y E(G)) se utilizan para simbolizar los conjuntos de vértices y de aristas cuando G no es el único grafo que se toma en cuenta.

Definición 2: Si un vértice v es el punto final de una arista e, entonces se dice que v es incidente a e, y e es incidente a v.

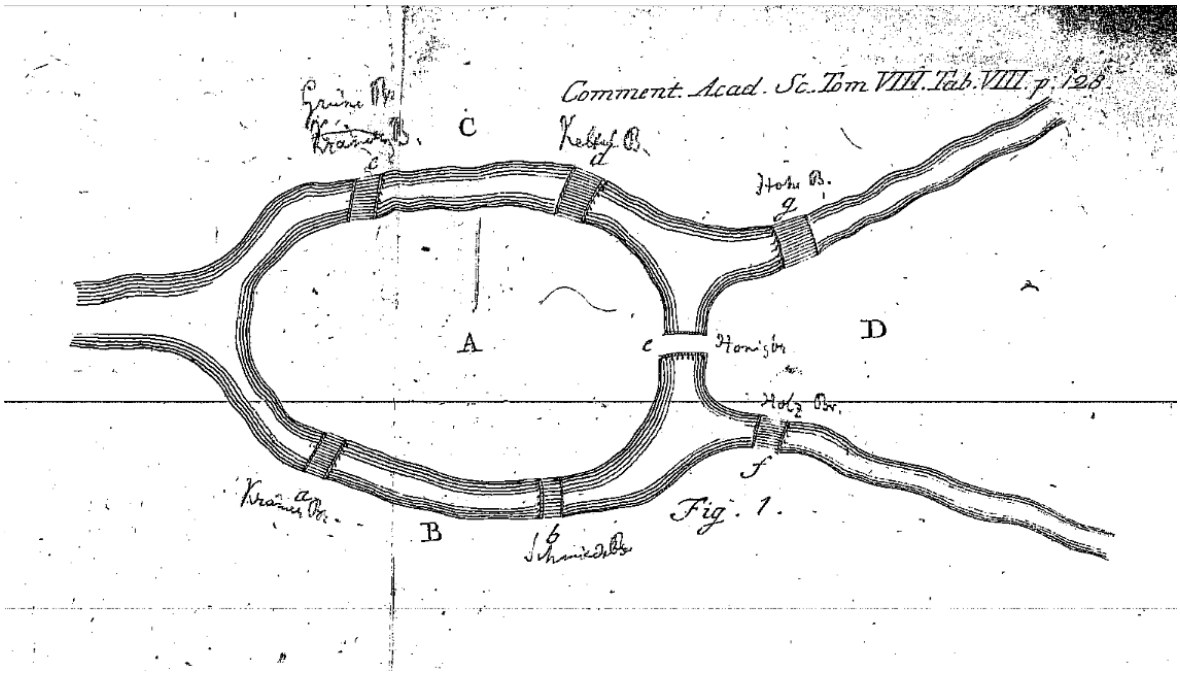
Definición 3: Un vértice u es adyacente al vértice v si los dos están unidos por una arista.

Definición 4: Dos vértices adyacentes son vecinos (neighbors) o conforman una vecindad.

Definición 5: Las aristas adyacentes son dos aristas que tienen un punto final en común.

Definición 6: Las aristas propiamente dichas son aristas que unen dos diferentes vértices.

Figura 4: El problema de los puentes Königsberg representado por Euler.



FUENTE: Euler (1741: 129)

Definición 7: Una arista múltiple (arista paralela, multi-arista) es una colección de una o más aristas que tienen puntos finales idénticos.

Definición 8: Entre los vértices existe adyacencia siempre y cuando haya por lo menos una arista entre ellos.

Definición 9: Un bucle es una arista que une un punto final a él mismo.

Definición 10: Un grafo simple es un grafo sin aristas múltiples o bucles

Definición 11: Un multigrafo es un grafo con aristas múltiples o bucles

Definición 12: Un camino es una secuencia de vértices dentro de un grafo tal que exista una arista entre cada vértice y el siguiente. Se dice que dos vértices están conectados si existe un camino que vaya de uno a otro.

Definición 13: Una componente conexa es un conjunto de vértices en el grafo para el cual, para cualquier par de vértices u y v en la componente, existe al menos un camino (una sucesión de vértices adyacentes que no repita vértices) de u a v.

No existe consenso sobre si un multigrafo puede considerarse grafo. En este texto se entiende por grafo a un grafo simple.

2.4. ANÁLISIS DE REDES

Una aplicación frecuente de los grafos es representar las interacciones entre los elementos en el análisis de redes (sociales, biológicas, de internet, etc.). En las redes se puede analizar la naturaleza de los componentes, y de las interacciones, pero, además de ello –y de eso se ocupa el estudio de grafos– el patrón de las interacciones. El patrón de las interacciones forma una red que puede tener efectos importantes sobre el comportamiento del sistema. Existen varios ejemplos: las redes sociales de soporte pueden mejorar el bienestar individual al proveer recursos sicológicos y físicos; los grupos de pares pueden aumentar o disminuir el riesgo de abuso de sustancias o influir sobre decisiones de uso de anticonceptivos; redes familiares y amicales pueden influenciar las prácticas dietéticas, los hábitos de ejercicio y otros comportamientos relacionados con la obesidad o el fumar; las redes formadas por las parejas sexuales pueden aumentar o disminuir el riesgo de contraer enfermedades de transmisión sexual; las redes de discusión entre colegas médicos pueden influenciar la implementación de protocolos o decisiones sobre prescripción de drogas novedosas. Además

de los aspectos sociales, las redes pueden también ofrecer acceso a recursos tangibles como asistencia financiera o transporte (O'Malley y Marsden 2008, Newman 2010).

Como explica Newman (2010), el grafo es una representación simplificada de esa estructura que recoge sólo información básica sobre los patrones de las conexiones. Se puede agregar información a los vértices y a las aristas como nombres y pesos, pero sigue siendo una representación reduccionista. La ventaja de esta simplificación es que abre paso a un campo de estudio que puede ser explorado desde el punto de vista de diferentes disciplinas, siempre y cuando los datos de interés sean posibles de representar en un grafo. Uno de los campos del análisis de redes es el análisis de redes sociales. Una red social es una red de personas o de grupos de personas, como instituciones. En el vocabulario de las redes sociales con frecuencia se hace referencia a los nodos como *actores,* y a las aristas como *enlaces*.

En la sociología, el estudio de estas interacciones, tal como se dan en el mundo real, tiene una larga tradición. Uno de los pioneros de esta disciplina fue el sicólogo Jacob Moreno que publicó en 1934 un estudio donde planteó el uso de sociogramas (hoy llamados redes sociales) y trazó los fundamentos de la sociometría – hoy llamada análisis de redes sociales (Moreno 1934). Uno de los estudios sociológicos más conocidos de redes sociales, ya que permitió predecir la separación de un grupo de karate, lo realizó el sociólogo Zachary. Tras una observación durante dos años de las interacciones de los miembros de un club de karate pudo detectar dos comunidades cercanas a los dos líderes del grupo. Después de un conflicto interno el grupo llegó a separarse y la topología del grafo le permitió asignar correctamente a los miembros del club (a excepción de uno) a cada una de las dos comunidades resultantes (Zachary 1977).

La más importante diferencia de los estudios estadísticos de redes sociales, con otros tipos de análisis estadístico es que se estudia la relación entre actores, no la relación entre variables (Hanneman y Riddle 2005). Son comunes dos tipos de diferentes modelos de redes: los basados en individuos y los basados en relaciones. En el primero el análisis se enfoca en el resultado a nivel individual y los datos sobre la red se utilizan para definir las variables explicativas. En el segundo, se modela las relaciones entre los individuos en una red, en esencia tratando a la red en sí como una variable multivariada dependiente en la que las aristas son sus elementos (O'Malley y Marsden 2008).

2.4.1. MEDIDAS LOCALES PARA EL ANÁLISIS DE REDES SOCIALES

Existen varias medidas que permiten describir las características de los vértices y el papel que estos juegan en el grafo. Estas son las medidas locales. Las diferentes medidas de centralidad con frecuencia están correlacionadas positivamente, pero no siempre es así, ya que recogen diferentes aspectos de prominencia en la red.

2.4.1.1. Grado

El grado de un vértice es el número de aristas incidentes en él. Con frecuencia se normaliza, en esos casos tenemos que:

$$k_i(normalizado) = \frac{k_i}{n-1} \quad (1)$$

Donde k_i es el grado del vértice *i* y *n* es el número de vértices en la componente conexa.

Existe una generalización del grado para los grafos ponderados, en esos casos el grado del vértice *i* es la suma de los pesos de las aristas incidentes, el nombre de esta medida es grado ponderado o fuerza (Kolaczyk y Csárdi 2014).

En estos casos se asume que los vértices con mayor grado son más centrales y por ende en una red social estos actores son más influyentes. Sin embargo, esta medida sólo mide la centralidad en referencia a los vecinos y no toma en cuenta las conexiones que ellos puedan o no tener (O'Malley y Marsden 2008, Newman 2010, Cordón 2012).

2.4.1.2. Cercanía

La cercanía (closeness centrality) es una medida pensada para capturar la cercanía de un actor a los otros actores.

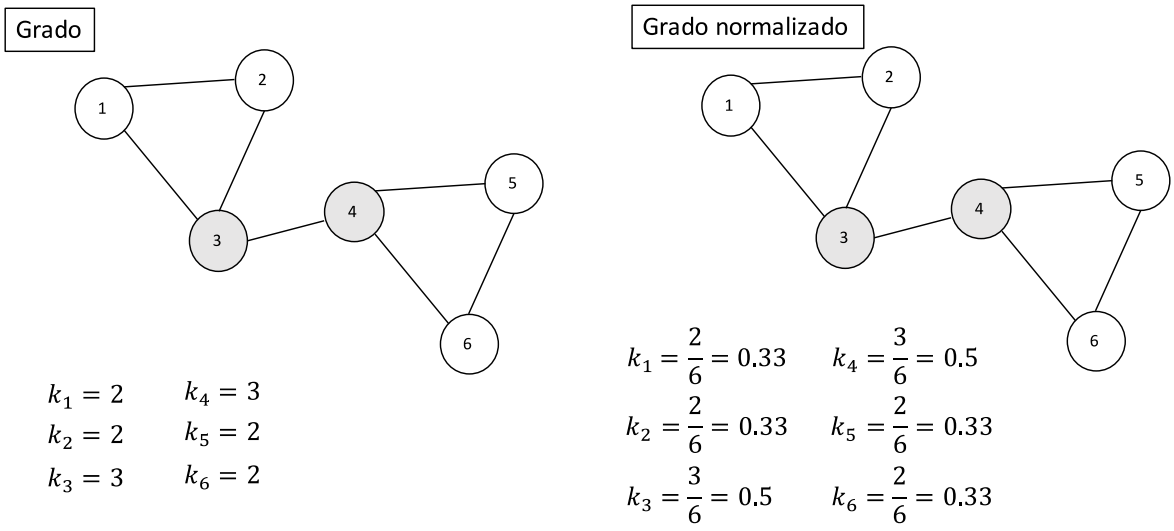
Tal como lo explica Newman (2010), si la $dist(v,u)$ es la longitud (número de aristas) del camino más corto entre v y u, entonces la distancia media entre v y u, promediada entre todos los vértices de la red menos el vértice mismo es

$$\ell_v = \frac{1}{N_v - 1}\sum_u dist(v,u) \quad (2)$$

Esta medida da valores pequeños para vértices que están separados de los otros por unos pocos vértices. Los vértices con valores bajos pueden tener mejor acceso a la información o

más influencia sobre otros elementos de la red. En redes sociales una persona cuyo ℓ_v es inferior al de otros actores de la misma red, puede encontrar que sus opiniones se propagan con más velocidad que las de otros.

Figura 5: Grados en un grafo.



FUENTE: Elaboración propia

Ya que la distancia media ℓ_v otorga valores bajos a los vértices más centrales y valores altos a los menos centrales la medida que se suele usar para el estudio de centralidad es su inversa, que se llama cercanía (closeness centrality) y que se define como sigue:

$$c_{Cl}(v) = \frac{1}{\ell_v} = \frac{N_v - 1}{\sum_{u \in V} \text{dist(v, u)}} \quad (3)$$

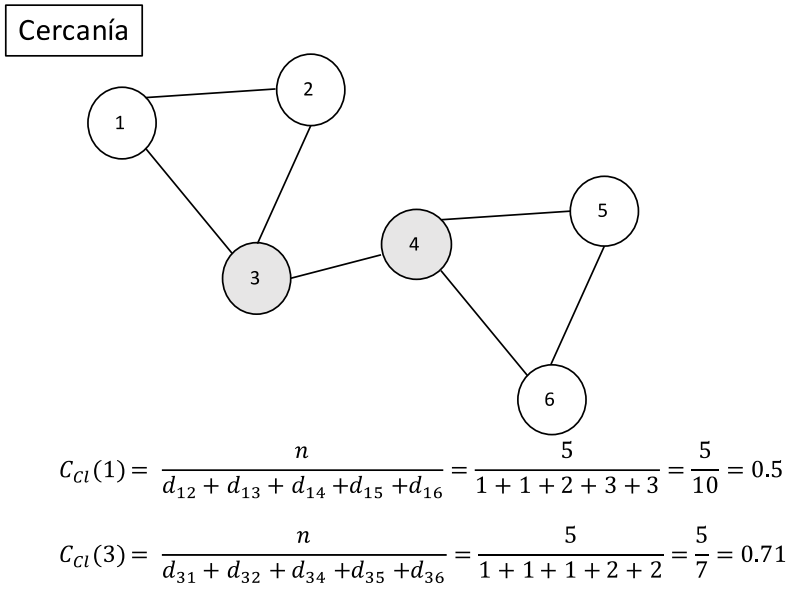
A veces se utiliza esta medida sin normalizar (es decir sin dividir entre el número de vértices), aunque esa cifra no es comparable entre diferentes grafos. En esos casos la cercanía se define como:

$$c_{Cl}(v) = \frac{1}{\sum_{u \in V} \text{dist(v, u)}} \quad (4)$$

La cercanía es una medida natural de la centralidad y se utiliza a menudo en estudios de las redes sociales. Pero tiene algunos problemas. El rango de valores que adquiere es pequeño ya que la distancia típica de un grafo aumenta logarítmicamente conforme crece la red. Esto significa que la ratio entre la menor distancia, que es 1, y la más grande, que es de orden $\log(n)$, es en sí mismo de orden $\log(n)$. En una red típica el rango de valores de $C_{Cl}(v)$ rara vez supera 5.

Otro de los problemas que enfrenta la cercanía es que se sirve de las distancias entre los vértices y como tal sólo puede calcularse sobre un grafo conexo.

Figura 6: Medidas de cercanía en un grafo.



FUENTE: Elaboración propia.

2.4.1.3. Intermediación

La intermediación (betweenes) es una medida pensada para capturar la correduría, es decir el número de veces que un actor se halla en posición intermedia en los caminos más cortos que unen dos vértices (O'Malley y Marsden 2008).

Los vértices con gran intermediación pueden tener mayor influencia, por ejemplo, en una red que transmite información, tendrían control sobre los mensajes que transfieren. Además, su remoción de la red tendría un fuerte impacto en la comunicación dentro del grafo, ya que

están posicionados en el mayor número de caminos. La comunicación en el mundo real no siempre se da siguiendo los caminos más cortos –que son los evaluados en la medida de intermediación– pero puede ser una buena aproximación al dominio que un vértice puede tener sobre el flujo de información (Newman 2010).

Como explican Kolaczyk y Csardi (2014), la medida de intermediación más comúnmente usada es

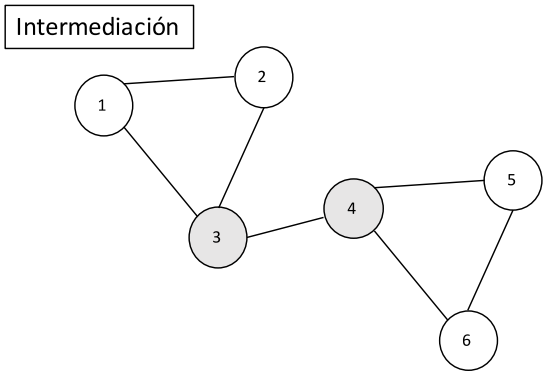
$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{5}$$

donde $\sigma(s,t|v)$ es el número de caminos más cortos entre *s* y *t* que pasan por v y $\sigma(s,t)$ es el número total de caminos más cortos entre *s* y *t* (independientemente si pasan o no por *v*). Cuando solo existe un camino más corto, $C_B(v)$ cuenta sólo el número de caminos más cortos que pasan por *v*. Newman (2010) explica que esta medida toma un rango de valores mucho más amplio que la cercanía, lo cual permite diferenciar mejor los vértices.

Una versión normalizada de la intermediación que permite reducir el rango de valores que esta puede tener entre 0 y 1 es

$$c_B(v)(normalizada) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t) \times \frac{(N_v - 1)(N_v - 2)}{2}} \tag{6}$$

Figura 7: Medidas de intermediación en un grafo.



$$C_B(3) = \frac{\sigma(1,2|3)}{\sigma(1,2)} + \frac{\sigma(1,3|3)}{\sigma(1,3)} + \frac{\sigma(1,4|3)}{\sigma(1,4)} + \frac{\sigma(1,5|3)}{\sigma(1,5)} + \frac{\sigma(1,6|3)}{\sigma(1,6)} + \frac{\sigma(2,3|3)}{\sigma(2,3)} + \frac{\sigma(2,4|3)}{\sigma(2,4)} + \frac{\sigma(2,5|3)}{\sigma(2,5)} + \frac{\sigma(2,6|3)}{\sigma(2,6)} + \frac{\sigma(3,4|3)}{\sigma(3,4)} + \frac{\sigma(3,5|3)}{\sigma(3,5)} + \frac{\sigma(3,6|3)}{\sigma(3,6)} + \frac{\sigma(4,5|3)}{\sigma(4,5)} + \frac{\sigma(4,6|3)}{\sigma(4,6)} + \frac{\sigma(5,6|3)}{\sigma(5,6)} = 0 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 = 11$$

$$C_B(1) = \frac{\sigma(1,2|1)}{\sigma(1,2)} + \frac{\sigma(1,3|1)}{\sigma(1,3)} + \frac{\sigma(1,4|1)}{\sigma(1,4)} + \frac{\sigma(1,5|1)}{\sigma(1,5)} + \frac{\sigma(1,6|1)}{\sigma(1,6)} + \frac{\sigma(2,3|1)}{\sigma(2,3)} + \frac{\sigma(2,4|1)}{\sigma(2,4)} + \frac{\sigma(2,5|1)}{\sigma(2,5)} + \frac{\sigma(2,6|1)}{\sigma(2,6)} + \frac{\sigma(3,4|1)}{\sigma(3,4)} + \frac{\sigma(3,5|1)}{\sigma(3,5)} + \frac{\sigma(3,6|1)}{\sigma(3,6)} + \frac{\sigma(4,5|1)}{\sigma(4,5)} + \frac{\sigma(4,6|1)}{\sigma(4,6)} + \frac{\sigma(5,6|1)}{\sigma(5,6)} = 1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 5$$

FUENTE: Elaboración propia.

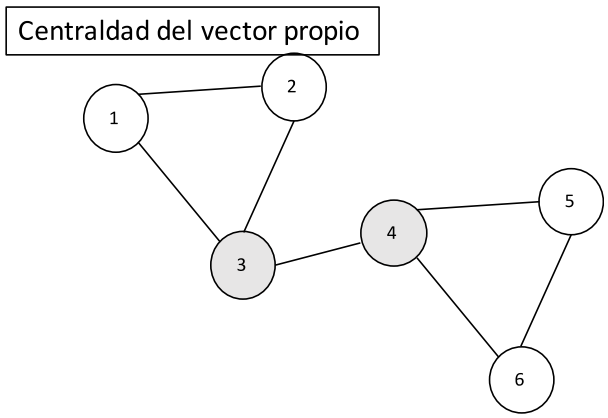
2.4.1.4. Centralidad de vector propio

Como explican Kolaczyk y Csardi (2014), otras medidas de centralidad se basan en la noción de «estatus» o «prestigio». Estas buscan capturar la idea que mientras más centrales son los vecinos de un vértice, más central es el vértice en sí. Estas medidas se basan en vectores propios (eigenvector) y entre las varias definiciones que existen el paquete de R utiliza esta:

$$c_{Ei}(v) = \alpha \sum_{\{u,v\} \in E} c_{Ei}(u) \qquad (7)$$

Donde el vector $\boldsymbol{c}_{Ei} = (c_{Ei}(1), \dots, c_{Ei}(N_v))^T$ es la solución al problema de valores propios de $\boldsymbol{A}\boldsymbol{c}_{Ei} = \alpha^{-1}\boldsymbol{c}_{Ei}$, donde **A** es la matriz de adyacencia del grafo G. En esta definición, que sigue la propuesta por Bonacich, el valor óptimo de α^{-1} es el máximo valor propio de **A** y por lo tanto $\boldsymbol{c}_{Ei}$ es el vector propio correspondiente. Por convención se toman en cuenta los valores absolutos de este resultado.

Figura 8: Medidas de centralidad del vector propio en un grafo.



FUENTE: Elaboración propia.

2.4.2. MEDIDAS GLOBALES PARA EL ANÁLISIS DE REDES SOCIALES

Diámetro. – La distancia entre vértices en un grafo se define como el largo (número de aristas) del camino más corto entre los dos nodos. Cuando no existe tal camino la distancia es igual al infinito. El valor de la distancia más larga en un grafo es su diámetro.

Distancia media. – La distancia media en un grafo es el promedio de las distancias.

Densidad. – La densidad de un grafo es la ratio del número de aristas entre el número máximo posible de aristas en el grafo:

$$\delta = \frac{\mathrm{n}_e}{\binom{\mathrm{n}_v}{2}} = \frac{\mathrm{n}_e}{\mathrm{n}_v(\mathrm{n}_v - 1)/2} \tag{8}$$

Donde n_e es el número de aristas en el grafo y n_v es el número de vértices en el grafo.

Coeficiente de clustering, coeficiente de agrupamiento o transitividad. – Mide la frecuencia relativa de los triángulos entre el número de tripletes (tres vértices con dos aristas):

$$Cl_T(\mathrm{G}) = \frac{3\tau_\Delta(G)}{\tau_3(G)} \tag{9}$$

2.4.2.1. Medidas de asortatividad

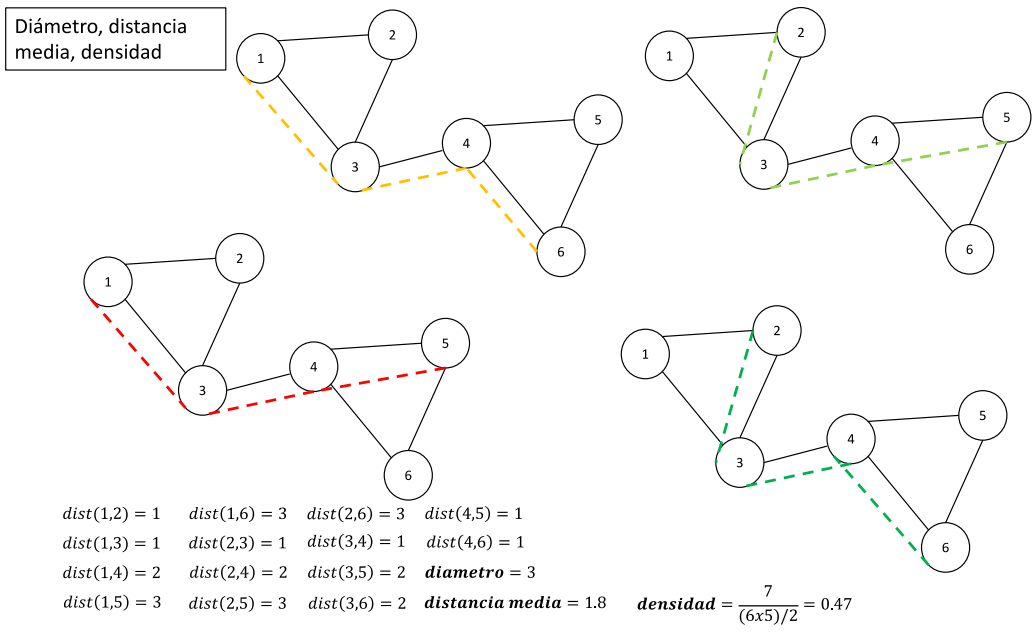
La asortatividad es la preferencia de los nodos de una red por unirse a otros que le son similares en alguna característica. En redes sociales con frecuencia se utiliza esta medida para analizar la tendencia de los vértices de cierto grado a ser vecinos de vértices con grado similar. El coeficiente de asortatividad es un coeficiente de correlación entre dos nodos. Los valores que adquiere pueden estar en un rango entre -1 y 1. Cuando el valor es 1 la red es totalmente asortativa, cuando es 0 la red no es asortativa y cuando es -1 la red es disortativa.

Las características de los nodos involucrados pueden ser categóricas, ordinales o continuas. Cuando se trata de variables categóricas, se asume que cada vértice en un grafo G puede clasificarse en una de las M categorías. En ese contexto el coeficiente de asortatividad se define como

$$r_a = \frac{\sum_i f_{ij} - \sum_i f_{i+} f_{+i}}{1 - \sum_i f_{i+} f_{+i}} \qquad (10)$$

En esta ecuación f_{ij} es la fracción de aristas en *G* que unen a un vértice en la categoría *i* con un vértice en la categoría *j* y f_{i+} y f_{+i} denota las sumas de las filas y columnas marginales de la matriz *f* resultante.

Figura 9: Medidas globales en un grafo.



FUENTE: Elaboración propia

2.5. LA BIBLIOMETRÍA Y LA CIENCIOMETRÍA

En la medición de las actividades de la ciencia nos topamos con dos conceptos: la cienciometría y la bibliometría. Según Spinak (1998: 142) la bibliometría comprende las siguientes actividades:

- «Aplicación de análisis estadísticos para estudiar las características del uso y creación de documentos.
- Estudio cuantitativo de la producción de documentos como se refleja en las bibliografías.

- Aplicación de métodos matemáticos y estadísticos al estudio del uso que se hace de los libros y otros soportes, dentro y entre los sistemas de bibliotecas.
- Estudio cuantitativo de las unidades físicas publicadas, o de las unidades bibliográficas, o de sus sustitutos.»

Mientras que la cienciometría, como explican (Arencibia y Moya 2008: 12), «no es más que la aplicación de técnicas bibliométricas al estudio de la actividad científica. Su alcance va más allá de las técnicas bibliométricas, puesto que puede ser empleada para examinar el desarrollo y las políticas científicas.» Esta definición no es la única ya que el objetivo de la cienciometría es el estudio de la actividad científica, donde las publicaciones académicas son sólo una de las posibles unidades de análisis. Según Bellis (2009) entre las unidades de análisis podría incluirse los recursos humanos, los instrumentos, la infraestructura, las inversiones económicas, y los resultados financieros.

Por otro lado, algunas leyes de bibliometría, desarrolladas por Lotka y Bradford encuentran una tendencia en la investigación científica, a acumular la producción en «manos» de pocos, como lo explica (Bellis 2009: xxiv-xxv, traducción propia):

«Entre 1920 y 1930, fueron publicados tres estudios clave en la historia de la disciplina, uno por Alfred Lotka sobre la distribución de la productividad científica, otro por Samuel Bradford sobre la dispersión de los artículos académicos entre las revistas especializadas, otro por George Zipf sobre las propiedades estadísticas de los corpus lingüísticos. Desde diferentes puntos de partida y perspectivas analíticas, los tres autores formalizaron un conjunto de regularidades – las 'leyes bibliométricas' – que estarían detrás de los procesos por los cuales cierto número de elementos (artículos académicos, textos, palabras), están relacionados a las fuentes que los generan (autores, revistas académicas, cuerpos lingüísticos). Su característica en común es una tendencia increíblemente estable a la concentración de los elementos en un número relativamente pequeño de fuentes. No era un secreto que existieran pocos científicos muy productivos en comparación al muy elevado número de autores de un solo texto, que la mayor parte de la literatura relevante para un campo de investigación sea generada por un número pequeño de revistas, y que pocas palabras se repitan con mucha más frecuencia que otras en el lenguaje hablado (y escrito), (…) las leyes de Lotka, Bradford y Zipf ofrecieron dar un acabado matemático a los elusivos patrones de comunicación.»

Torres (2009) diferencia tres etapas en el desarrollo de la bibliometría:

- *La fase de iniciación*, se dio aproximadamente desde 1917 hasta los años 1950's. Se inició con el primer recuento bibliográfico sobre anatomía, estudio en el cual se aplicó un análisis cuantitativo a la literatura sobre el tema desde 1543, hasta 1860. Este estudio por muchos es considerado la primera investigación bibliométrica ya que tuvo como objeto de estudio las publicaciones, como propósito la evaluación del performance, y el mapeo de las disciplinas científicas, y tuvo además las limitaciones propias de un estudio de análisis cuantitativo (Bellis 2009: 6-7). En esta época se elaboraron los modelos teóricos de la distribución de la producción bibliográfica de Lotka (1926) y Bradford (1985 [1934]), ya mencionados anteriormente. Además, desde 1923 se empezaron a utilizar los conteos de trabajos publicados para comparar la productividad científica de diversos países.

- *La fase de establecimiento*, se dio entre los años 1950's hasta los 1980's. Después de la segunda guerra mundial la guerra fría generaba una continua competencia por tener el mejor desenvolvimiento científico entre los dos bloques. Este contexto político favoreció el interés en la evaluación del desempeño científico. Así, en la década de los 1960's se acuño el término Ciencia de la Ciencia y se definió el término Bibliometría: una ciencia que mediante la contabilización y análisis de la comunicación escrita pretende conocer la naturaleza y desarrollo de una disciplina. En esa época se estableció también el término Ciencimetría o Cienciometría que «se ocupa de la evaluación cuantitativa y comparativa de la contribución al avance de conocimiento de científicos, grupos, instituciones y países. (...) los documentos publicados no son más que una de las varias unidades de análisis posibles [y] la cienciometría y bibliometría se yuxtaponen en gran medida.» (Bellis 2009: 3, traducción propia)

  En esta década se dio un importante impulso a los estudios bibliométricos para medir los resultados de las investigaciones, gracias, tanto al interés de los responsables de la planificación científica, como a la automatización de datos bibliográficos. Un hito de la época fue la Teoría de la Citación de Garfield quien también impulsó la publicación del Science Citation Index (SCI) y la formalización del Institute for Scientific Information (ISI), productor de este índice. Los índices producidos por el ISI se convirtieron en las fuentes de información sobre citaciones, permaneciendo

para muchas instituciones cómo únicas fuentes de este tipo de información hasta el día de hoy. Otro hito de esta fase fue la creación de la revista Scientometrics (1978) y la introducción del concepto de calidad en los estudios cuantitativos de la ciencia.

- *La fase de consolidación*, se da a partir de 1980 y sigue hasta la actualidad. La consolidación se da en el uso de canales de difusión (revistas, conferencias, listas de discusión) y en el desarrollo conceptual. Además, la tecnología informática aplicada a las bases de datos bibliográficas permite un procesamiento complejo de las fuentes de información. En esta fase se crea el Comité de Informetría (India), como parte de la ya desaparecida Federación Internacional de Documentación que data de 1980. En los años 1980s se populariza también la elaboración de extensas bibliografías. Además, desde 1995, se establecen los procedimientos de evaluación bibliométrica: el análisis bibliométrico de la producción científica y los mapas de la ciencia.

El análisis bibliométrico de la investigación científica evalúa variables científicas a partir de datos bibliográficos. Las bases de datos de uso generalizado para dicho procedimiento son las del ISI, que tuvo el monopolio de los análisis bibliométricos hasta el año 2004, año en que se introdujo al mercado la base de datos Scopus, de Elsevier, y se lanzó el servicio web Google Scholar. De las tres opciones disponibles en el momento los servicios de Elsevier y de Thomson ISI tienen características comparables, mientras que el servicio de Google Scholar tiene una cobertura mucho más amplia pero menos información o de menor calidad (por ejemplo, no cuenta con información institucional) de interés para el análisis bibliométrico. La medición se puede dar a diferentes niveles de agregación: países, instituciones, revistas e individuos.

Los mapas de la ciencia buscan representar los dominios científicos, mostrando las relaciones existentes entre los diferentes campos de estudio mediante el análisis de co-citas y de citas de las mismas revistas. Como explica Torres (2009: 40) «para la construcción de dichos mapas se utilizan una serie de técnicas (…) como el Análisis Clúster, Escalamiento Multidimensional (MDS) y el Análisis Factorial, Mapa Auto Organizativo que está basada en un tipo de red neuronal y el Análisis de Redes Sociales.» Existen numerosos estudios que aplican estas técnicas, se tiene a investigadores de la Universidad de Drexel (Programa CiteSpace) y de la Universidad de Granada –Grupo Scimago- Proyecto Atlas de la Ciencia Española (Moya et al. 2004, 2004, 2005, 2005, 2006). Algunos de estos proyectos pueden verse en la web como el Atlas de la Ciencia Española (SCImago 2006), el proyecto de

Boyack y Klavans de Information Esthetics del 2006 (Boyack et al. 2007) y un proyecto similar por los mismos autores The Better Maps (Boyack et al. 2013). A nivel de América Latina tenemos la tesis doctoral «La Visualización de la Información en el entorno de la Ciencia de la Información» sobre la visualización de la ciencia por la investigadora cubana Torres (2010).

En el análisis de la investigación científica, hace ya un tiempo que se utiliza el análisis de grafos como una herramienta que permite conocer las interacciones entre los investigadores. Tomando en cuenta el objeto de estudio de interés, las características de la investigación científica pueden analizarse con la ayuda del análisis de co-citas, coautorías, co-uso de palabras, co-enlaces y similares. Para conocer la intensidad de las colaboraciones científicas entre autores, instituciones o países es ya tradicional recurrir a las coautorías (Narin et al. 1991, Slone 1996, Cockburn y Henderson 1998, Glänzel 2001, Newman 2001, 2001, Cronin et al. 2003, Newman 2004, Moody 2004, Liben-Nowell y Kleinberg 2007). En este caso se traduce los elementos de la coautoría en un grafo de la siguiente manera: el autor es el nodo (dependiendo del nivel de análisis se puede considerar como autor a una persona, una institución o un país), existe una arista entre dos autores cuando comparten la autoría de un documento en conjunto. Los autores que colaboran entre ellos con más frecuencia que con otros forman conglomerados.

Como explica Arencibia (2010) este tipo de estudios se remonta a los años 1960, ya que en 1958, Michael Smith sugirió el conteo de los artículos en co-autoría como una medida de colaboración entre grupos de investigadores. En los años 1970 el tema fue desarrollado por la socióloga Diane Crane y las redes de colaboración fueron de interés de varios investigadores entre los 60s y 70s, quienes analizaban su estructura, surgimiento, la validez de la coautoría como medición, etc. Entre los más importantes estuvieron Warren Hagstrom, Derek J. de Solla Price, Donald Beaver y Stanley Milgram.

A partir del año 1971, en que se decide medir la colaboración científica como el conjunto de trabajos desarrollados por más de una persona, los estudios métricos que buscaban conocer las diferencias entre las estructuras y costumbres en diferentes especialidades de la investigación se hicieron muy frecuentes. Se encontró también que los artículos en colaboración tenían más posibilidades de obtener citas, especialmente cuando la co-autoría fuera internacional. En los años 1990, Katz, además de otras investigaciones, elaboró una clasificación de los tipos de colaboración científica a partir del nivel en el que se diera la colaboración (individuo, grupo, departamento, institución, sector, nación).

A partir del nuevo milenio, se comienza a utilizar las técnicas del Análisis de Redes Sociales (ARS), desarrolladas con anterioridad por autores como Barabási y Albert (1999), Borgatti y Wasserman, que dieron pie a las nuevas investigaciones como el estudio de Newman (2001), que analiza las coautorías recogidas en MEDLINE, Los Alamos e-Print Archive y NCSTRL. Newman sirviéndose del ARS estudia la estrechez de la colaboración, la agrupación de los investigadores en clústeres, y las diferencias en los patrones de colaboración en diferentes campos de estudio. El mismo autor realizó dos años después un recuento de las herramientas existentes en el campo de las redes sociales (Newman 2003). El análisis de co-autorías y co-citaciones recibe el nombre de indicadores relacionales de primera generación, en contraposición al análisis de la co-ocurrencia de palabras que recibe el nombre de indicadores relacionales de segunda generación (Torres 2007).

Actualmente, el análisis de las redes de investigación nos permite plantear y responder preguntas complejas, gracias a herramientas como el estudio sociológico, la investigación de redes científicas, así como la utilización de herramientas de diferentes campos de estudio para la visualización de las redes de co-citación y co-autorías. Tal desarrollo no sería posible sin la ayuda de las tecnologías de la información, cuya capacidad de procesamiento permite realizar tareas de análisis sobre enormes cantidades de datos (Chinchilla-Rodríguez et al. 2008). Así las últimas investigaciones del área se sirven de redes de citas, co-autorías, co-citas, de «bibliographic coupling networks», «co-word networks», y otras redes híbridas y heterogéneas, siendo la unidad de análisis un artículo científico, el cual es utilizado como base para niveles agregados de análisis como una revista, una institución, un país o un campo de estudio (Yan y Ding 2012).

2.5.1. ANÁLISIS DE DOMINIO

El análisis de dominio no es un indicador sino un *paradigma o perspectiva de análisis* vigente en la Ciencia de la Información. El marco teórico lo postularon los daneses Hjørland y Albrechtsen (1995: 400, traducción propia) que recopilaron las tendencias de las investigaciones existentes - como ellos mismos explicaron «un punto central es que la visión del análisis de dominio, que estamos tratando de formular aquí, está más o menos latente en muchas investigaciones contemporáneas en [Ciencias de la información], así como en contribuciones anteriores.» En el análisis de dominio se entiende la producción científica como producto de comunidades discursivas, en cuanto la pertenencia a determinadas comunidades implica formar parte de una manera de organizar el conocimiento, de ser parte de una estructura, de servirse de determinados patrones de cooperación, lenguaje y modos

de comunicación. En este paradigma los comportamientos anómalos de algún individuo no tendrían que afectar necesariamente la interpretación de los resultados de análisis de un grupo. Como lo resume Miguel (2008: 9) los aspectos más relevantes de este enfoque son:

- «una visión holística del conocimiento entendido como proceso y producto social y cultural;
- una concepción y enfoque metodológico social-colectivista, en detrimento del cognitivo-individualista;
- un intento por comprender las características tanto explícitas como implícitas del comportamiento de información y comunicación, y
- un análisis centrado en la comunicación científica, las publicaciones, las disciplinas y especialidades científicas, las estructuras de conocimiento y los paradigmas.»

Unos años después Hjørland (2002) propuso once herramientas o métodos para el análisis de dominio:

- guías bibliográficas,
- sistemas de clasificación especializados y tesauros,
- especialidades de recuperación e indización,
- estudios de usuarios,
- estudios bibliométricos,
- historia de la ciencia (especialmente en la parte relativa a la generación de disciplinas científicas),
- estudios de los documentos y tipologías documentales,
- estudios epistemológicos,
- estudios de terminología, lenguajes para propósitos específicos (LSP), bases de datos semánticas, análisis del discurso,
- estructura e instituciones de la comunicación científica,
- la cognición científica, el conocimiento experto, la inteligencia artificial.

Como puede apreciarse los estudios bibliométricos serían una de varias herramientas propuestas, aunque con gran potencial, especialmente cuando se complementan con los estudios de historia de la ciencia y la epistemología. La bibliometría permite «explorar los patrones de la comunicación científica, y las conexiones entre autores, artículos, revistas, disciplinas, paradigmas, en tanto exponentes de las comunidades de discurso que producen y difunden conocimientos científicos en forma de publicaciones (Miguel y Moya 2009: 2).»

Esta propuesta tuvo una buena recepción dentro de las ciencias de la información, en donde ya con anterioridad se habían generado y propuesto mapeos de co-citaciones. Se tienen estudios de análisis de dominio a nivel global (Boyack et al. 2007, Börner 2010, Boyack et al. 2013, SCImago 2006), de disciplinas (Miguel et al. 2007), de países (Chinchilla-Rodríguez 2005, Vargas-Quesada et al. 2008, Miguel 2008, Miguel et al. 2008, Moya et al. 2004, 2006), así como de sectores (Miguel et al. 2006) e instituciones (Moya et al. 2005). Los estudios de dominio se dan sobre un campo dinámico, ya que los paradigmas de la ciencia son dinámicos, y la colaboración existente entre las diferentes sub-disciplinas varía. Por ello es esencial tratar las diferentes dimensiones con las metodologías adecuadas para los análisis, considerando los diferentes discursos existentes.

2.6. CONGLOMERADOS

Existen varias investigaciones sobre la detección de conglomerados[5] a nivel computacional, la investigación de Fortunato (2010) recopila estas metodologías con énfasis en los modelos generados por los investigadores de ciencias físicas, mientras que Kolaczyk (2009) y Kolaczyk y Csárdi (2014) dan un enfoque metodológico más estadístico. En cuanto a países hispanohablantes son interesantes las tesis «Comunidades en grafos» (Carvajal 2006), y «Detección de comunidades en redes complejas» (Aldecoa 2013).

Newman y Girvan explican que las redes sociales tienden a tener una estructura comunitaria, y que pueden tener características a nivel de comunidades que no son típicas de toda la red (Newman 2006, Girvan y Newman 2002, Newman y Girvan 2004). La identificación de conglomerados en un grafo se realiza analizando la topología del mismo. No existe consenso sobre la metodología para realizarlo, pero Fortunato (2010) lo explica de la siguiente manera.

Se puede representar un grafo en una matriz de adyacencia A, definida de la siguiente manera:

$$A_{ij} = \begin{cases} 1, si\ i\ y\ j\ son\ adyacentes \\ 0, en\ los\ otros\ casos \end{cases} \quad (11)$$

En la presente investigación sólo se analiza los grafos no dirigidos, es decir aquellos donde $A_{ij}=A_{ji}$. En estos grafos la matriz de adyacencia tendrá una forma simétrica.

[5] En la terminología de análisis de redes los términos conglomerados y comunidades se utilizan indistintamente.

Comencemos con un subgrafo C de un grafo G, con $|C| = n_c$ y $|G| = n$ vértices, respectivamente. Se define el grado externo e interno del vértice v, $v \in C, k_v^{int}\ y\ k_v^{ext}$ como el número de aristas que unen v a otros vértices de C o al resto del grafo respectivamente.

Si $k_v^{ext} = 0$ el vértice tiene vecinos solo dentro de C, el cuál sería un buen conglomerado para v; en cambio si $k_v^{int} = 0$ el vértice es disjunto de C y en consecuencia debería ser asignado a un conglomerado diferente. El *grado interno* k_{int}^C de C es la suma del grado interno de sus vértices, de manera similar el *grado externo* de C k_{ext}^C es la suma de los grados externos de sus vértices. El grado total k^C es la suma de los grados de los vértices de C. Por definición $k^C = k_{int}^C + k_{ext}^C$.

Se define la densidad interna $\delta_{int}(C)$ del subgrafo C como la ratio entre el número de los vértices internos de C y todos los posibles vértices internos, es decir

$$\delta_{int}(C) = \frac{k_{int}^C}{n_c(n_c - 1)/2} \tag{12}$$

De manera similar, la densidad externa $\delta_{ext}(C)$ del subgrafo C como la ratio del número de aristas desde los vértices del subgrafo C al resto del grafo y el número máximo de vértices externos posible, es decir

$$\delta_{ext}(C) = \frac{k_{ext}^C}{n_c(n - n_c)} \tag{13}$$

Para que C constituya un conglomerado se espera que $\delta_{int}(C)$ sea mucho mayor que la densidad promedio $\delta(G)$ del grafo G que se obtiene de la división del número de vértices de G y el número máximo de posibles vértices *n(n-1)/2.* Al mismo tiempo $\delta_{ext}(C)$ debe ser mucho más pequeña que $\delta(G)$. Buscar la mejor relación entre $\delta_{ext}(C)\ y\ \delta(G)$ es el objetivo de los algoritmos de conglomerados. Una manera sencilla de realizarlo es maximizando le suma de las diferencias de $\delta_{int}(C) - \delta_{ext}(C)$ para todos los conglomerados.

El particionamiento, en líneas generales, se refiere a la identificación de estructuras de subgrupos naturales. Una partición $C = \{C_1, \ldots, C_k\}$ de un conjunto finito S es la descomposición de S en K conjuntos disjuntos, no vacíos C_k tales que $\cup_{k=1}^{K} C_k = S$. En el análisis de grafos el particionamiento constituye una herramienta útil para identificar de

manera no supervisada subconjuntos de vértices que demuestran «cohesión» en relación a las estructuras del grafo, es decir, la detección de comunidades.

A continuación, se ilustra el uso de un método bien establecido de identificación de grafos: el particionamiento jerárquico.

2.6.1. EL PARTICIONAMIENTO JERÁRQUICO

Tal como lo explican Kolaczyk y Csárdi (Kolaczyk 2009, Kolaczyk y Csárdi 2014) los métodos de particionamiento jerárquico son variaciones de los conceptos de conglomerados jerárquicos utilizados en el análisis de datos. Existen varias técnicas propuestas para la identificación de los conglomerados, las cuales difieren en los algoritmos utilizados para optimizar la calidad de los conglomerados y en las metodologías para evaluar la calidad de los conglomerados propuestos. Estos métodos emplean algoritmos voraces (greedy)[6] para buscar el espacio de todas las posibles particiones de C, mediante iterativas modificaciones de sucesivas particiones candidatas. Los métodos jerárquicos pueden clasificarse en (i) aglomerativos, basados en sucesivos aumentos de las particiones mediante un proceso de fusión, o (ii) divisivos, basados en sucesivas particiones mediante un proceso de división. En cada etapa, la partición vigente es modificada de una manera que minimice una medida de costo previamente especificada. En los métodos aglomerativos, se realiza la fusión menos costosa de dos particiones previamente existentes, mientras que, en el método divisivo, se realiza la división en dos menos costosa de alguna de las particiones existentes.

2.6.2. LA MODULARIDAD

La medida de costo incorporada en los métodos de conglomerados jerárquicos refleja el concepto de lo que se entiende como un subconjunto de vértices «cohesivo.» Hay varias medidas de costo propuestas. Entre las más populares está la de modularidad (Newman y Girvan 2004). Fortunato (2010) explica que la modularidad se basa en la idea, de que no se espera de un grafo al azar que tenga una estructura comunitaria, así que la posible existencia de conglomerados se puede encontrar comparando la densidad real de las aristas en el subgrafo, frente a la densidad que se esperaría de un grafo sin estructura comunitaria. Esta

[6] Un algoritmo voraz (también conocido como ávido, devorador o goloso) es aquel que, para resolver un determinado problema, sigue una heurística consistente en elegir la opción óptima en cada paso local con la esperanza de llegar a una solución general óptima. Este esquema algorítmico es el que menos dificultades plantea a la hora de diseñar y comprobar su funcionamiento. Normalmente se aplica a los problemas de optimización.

densidad esperada depende del modelo nulo, es decir de la copia del grafo original que mantenga algunas de sus propiedades estructurales, pero sin la estructura comunitaria.

Los primeros en definir la modularidad fueron Newman y Girvan (2004). Siguiendo su planteamiento Kolaczyk (2009) define la modularidad como sigue: sea $C = \{C_1, \ldots, C_k\}$ una partición candidata y $f_{ij} = f_{ij}(C)$ la fracción de las aristas en el gráfico original que conectan los vértices en C_i con los vértices en C_j. La modularidad de C es e

$$Q(C) = \sum_{k=1}^{K} |f_{kk}(C) - f_{kk}^{*}|^2 \qquad (14)$$

donde f_{kk}^{*} es el valor esperado de f_{kk} en un modelo al azar.

Con frecuencia, se define f_{kk}^{*} como $f_{k+}f_{+k}$, donde f_{k+} y f_{+k} son las sumas de f en la fila y columna k de la matriz $K \times K$ formada por las entradas f_{ij}. Para un grafo no dirigido, como el de esta investigación, esta matriz será simétrica, con $f_{k+} = f_{+k}$. Este modelo permite construir un grafo con la misma distribución de grado (degree distribution) que G, pero con las aristas distribuidas al azar, sin tomar en cuenta la topología de las particiones de C. Valores grandes de modularidad llevarán a suponer que C capta estructuras comunitarias no triviales, por encima de lo esperado al asignar las aristas al azar.

En un principio, la optimización de la modularidad en (14) requiere una búsqueda sobre todas las posibles particiones C, lo cual resulta muy costoso en grafos grandes o moderados.

La utilización de la maximización de la modularidad está asociada a ciertas asunciones (Good et al. 2010: 2):

- Las redes empíricas con estructura modular tienden a exhibir una partición óptima clara;
- las particiones de alta modularidad de una red empírica son estructuralmente similares a esta partición óptima y
- el Qmax estimado puede compararse significativamente entre las redes.

Además, en el caso de los grafos correspondientes a redes reales se ha encontrado que las particiones halladas mediante maximización de la modularidad, deben interpretarse cautelosamente, en algunos casos nos hallaremos frente a particiones degeneradas que no corresponden a la topología modular de la red. El artículo de Good resalta que esta limitación

no es única de la función de modularidad y que probablemente es compartida por otras técnicas de identificación de conglomerados (Good et al. 2010). Adicionalmente, se ha demostrado que la optimización de modularidad puede fallar en reconocer módulos más pequeños que una escala asociada al tamaño del grafo y al grado de interconectividad entre los módulos, aún en casos en los que los módulos están definidos sin ninguna ambigüedad, lo cual constituye un sesgo del método mismo. (Fortunato y Barthélemy 2007, Fortunato y Castellano 2012). Eso significa que el valor máximo de la función de modularidad no necesariamente coincide con la partición que correctamente identifica los módulos intuitivos o la partición intuitiva.

Good resume las limitaciones asociadas al uso de la modularidad de la siguiente manera:

- «La partición óptima puede no coincidir con la partición más intuitiva [límite de resolución mencionado entre otros por Fortunato y Barthélemy (2007)], [este es] un efecto impulsado principalmente por las consecuencias de suponer que la conectividad entre módulos sigue un modelo de grafo aleatorio.

- Normalmente hay un número exponencial de particiones alternativas estructuralmente diversas con modularidades muy cercanas al óptimo (el problema de degeneración). Este problema es más severo cuando se aplica a redes con estructura modular; ocurre para generalizaciones de la modularidad para grafos ponderados, dirigidos y/o bipartitos; y es probable que exista en muchas de las funciones de partición menos populares para la identificación de módulos.

- La máxima puntuación de modularidad Qmax depende del tamaño de la red n y del número de módulos k que contiene.» (Good et al. 2010: 12, trad. propia)

2.6.3. ALGORITMO AGLOMERATIVO DE CLAUSET

Entre los algoritmos que buscan identificar conglomerados ganó especial popularidad el algoritmo voraz, jerárquico y aglomerativo, propuesto por Clauset (Clauset et al. 2004) e implementado en *fastgreedy.community* en el paquete igraph de R (Csardi y Nepusz 2006). La lógica que sigue el algoritmo es la siguiente:

- Se define la modularidad Q de la partición analizada del grafo como el ratio del número de aristas dentro de la comunidad entre el número de aristas entre las comunidades, menos la ratio que se esperaría de una distribución al azar.

- Tomando en cuenta dos comunidades *i* y *j* dentro de una misma partición, se define ΔQ_{ij} como el aumento en la modularidad de la partición, que es resultado de la unión de las comunidades *i* y *j*. Mientras se cumpla que $\Delta Q_{ij} > 0$, aglomerar *i* y *j* mejorará la modularidad de la partición.

Esto lleva al autor a proponer un algoritmo voraz que sigue la siguiente secuencia:

- Comienza con cada nodo en su propia comunidad. Calcula ΔQ_{ij} para cada par de comunidades. Une cualesquier par de comunidades *i* y *j* tengan el mayor ΔQ_{ij}.
- Repite el paso.
- Se obtiene el máximo de modularidad cuando todos los ΔQ_{ij} tengan valores negativos, aunque se puede seguir realizando uniones (que es lo que sugiere Clauset).

En orden de seguir con la explicación del algoritmo es necesario hacer algunas definiciones. Se comenzará con una matriz de aumentos de modularidad en la que:

$$\Delta Q_{ij} = \begin{cases} \frac{1}{2m} - \frac{k_i k_j}{(2m)^2} & si\ i\ y\ j\ son\ adyacentes, \\ 0 & en\ otros\ casos, \end{cases} \tag{15}$$

Donde:

m – es el número de vértices en el grafo

k_i – Es el grado del nodo *i*

Además,

$$a_i = \frac{k_i}{2m} \tag{16}$$

para cada *i*. Esto corresponde a un grafo no ponderado, pero se puede hacer una generalización para grafos ponderados (Newman 2004).

En cuanto al almacenamiento de datos esto implica que se requiere recuperar los respectivos valores máximos de ΔQ_{ij} rápidamente. La solución que se propone es la que sigue:

- Para cada comunidad *i*, cada valor de ΔQ_{ik} se almacena en un árbol binario balanceado, donde k indiza el árbol; y en un montículo (heap), para facilitar el hallazgo de los máximos.
- El valor máximo de ΔQ_{ij} de cada comunidad es posteriormente almacenado en otro montículo para que el máximo en cada paso del proceso pueda encontrarse fácilmente.
- Cuando se fusiona las comunidades *i* y *j* (lo cual ocurre si esta fusión lleva al aumento de modularidad máximo), el árbol binario se modifica, tomando en cuenta que la comunidad fusionada (*k*) se colocará en el lugar de los valores de *j,* los valores modificados de ΔQ_{ij} serán:

(a) para *k* adyacentes a *i* y *j:*

$$\Delta Q'_{ik} = \Delta Q_{ik} + \Delta Q_{jk} \tag{17}$$

(b) para *k* adyacentes a *i,* pero no a *j:*

$$\Delta Q'_{ik} = \Delta Q_{ik} - 2a_j a_k \tag{18}$$

(c) para *k* adyacentes a *j,* pero no a *i:*

$$\Delta Q'_{ik} = \Delta Q_{jk} - 2a_i a_k \tag{19}$$

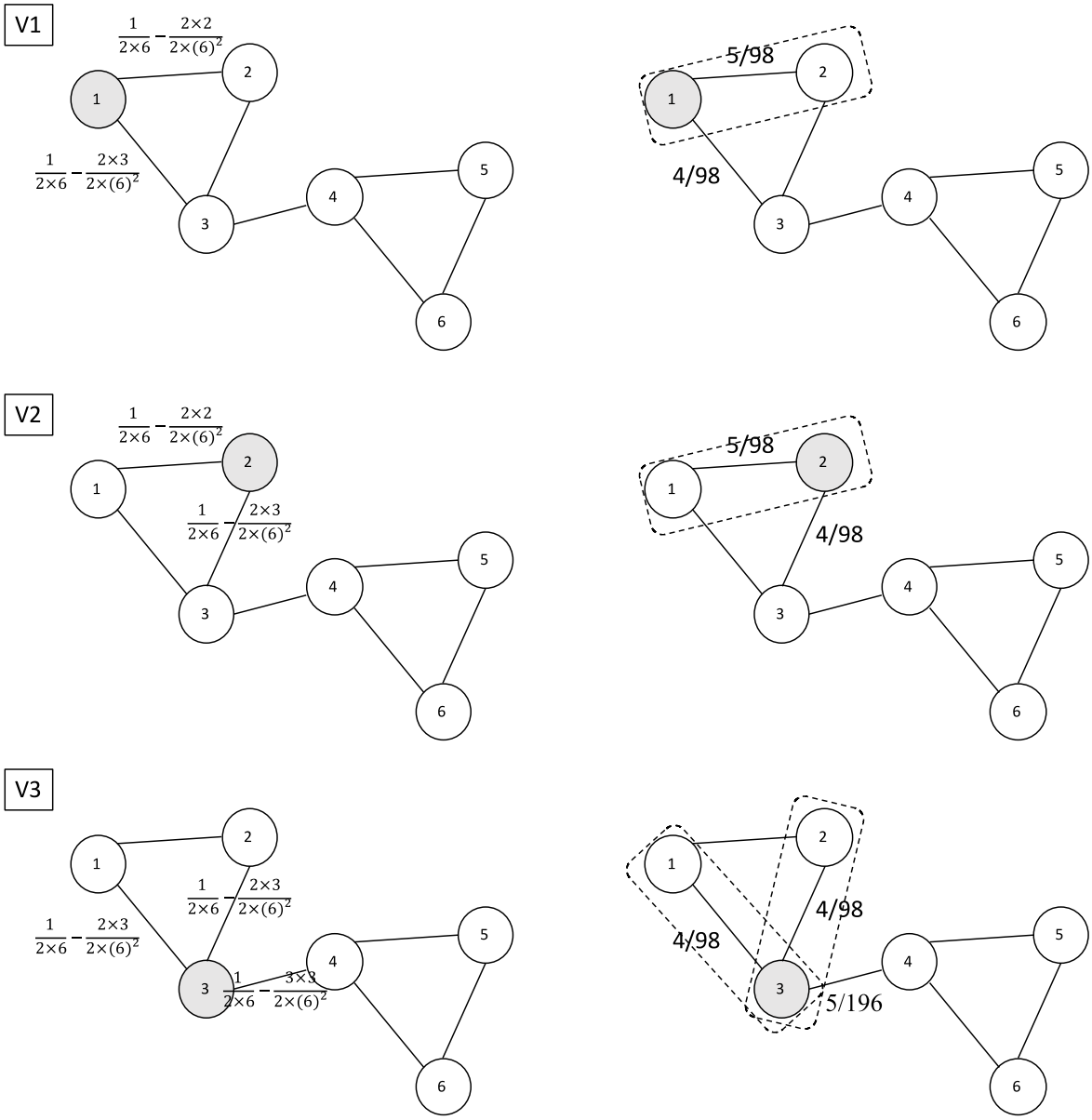
En la figura 10 puede verse el proceso de cálculo de modularidad para un grafo simple.

Figura 10: Cálculo de modularidad en un grafo.

$$\Delta Q_{ij} = \frac{1}{2m} - \frac{k_i k_j}{2m^2}$$

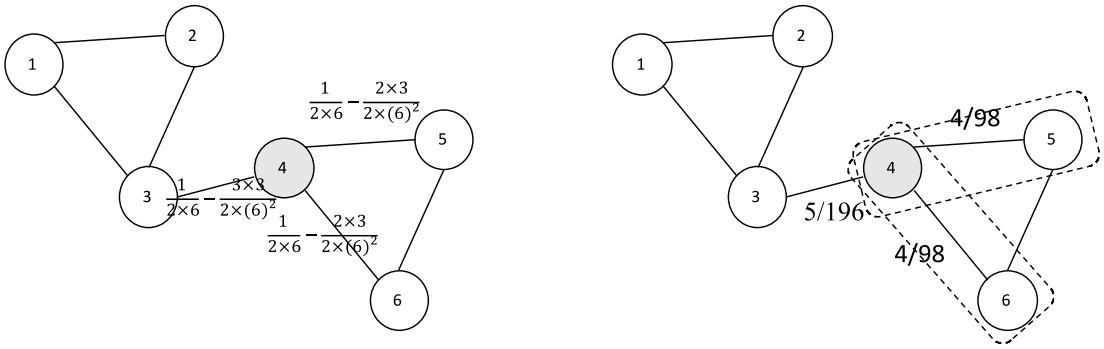
$m = 6$

$k_i = Grado\ del\ vértice\ i$

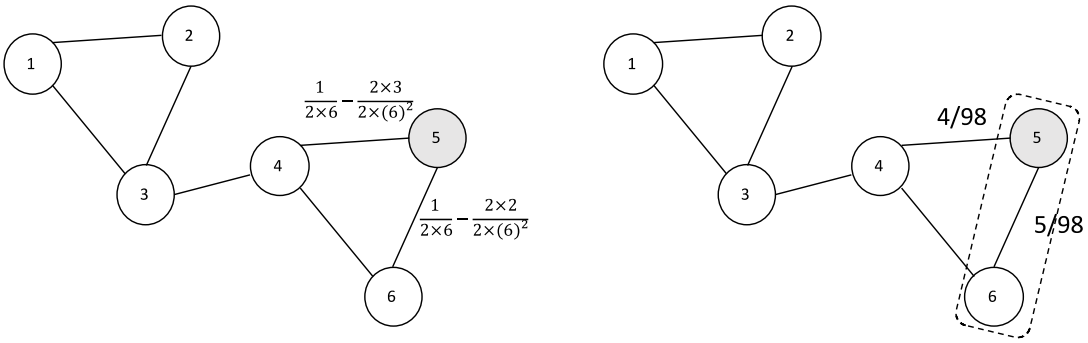


Continuación

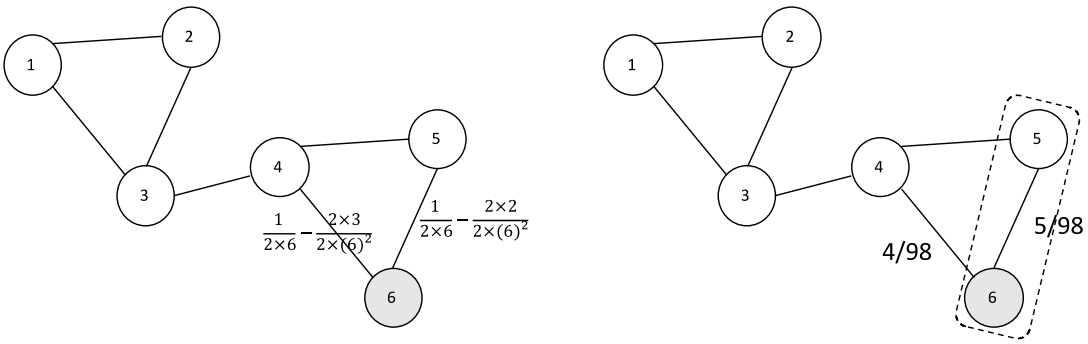
V4



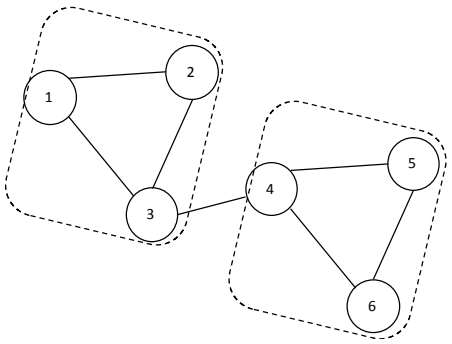
V5



V6



Conglomerados resultantes



FUENTE: Elaboración propia.

III. MATERIALES Y MÉTODOS

3.1. MATERIALES

Los materiales y equipos de los que se hizo uso en la presente tesis son los siguientes:

1. Una laptop marca Toshiba con un procesador Intel® core™ i7 CPU @2.5 GHz 2.5 GHz, con 12 GB de memoria RAM y un sistema operativo Windows 10 de 64 bits.
2. Un USB de 32GB de velocidad 3.0 en la modalidad readyboost para aumentar la memoria caché.
3. El programa estadístico R versión 3.2.5 para 64 bits y su interfaz RStudio 0.99.896 (R Core Team 2016, RStudio 2016)
4. Los paquetes de R: *ape, class, e1071, igraph, network, qdap, readr, RTextTools, sna, SnowballC, stringr, stringdist, tm.*

3.2. METODOLOGÍA DE LA INVESTIGACIÓN

3.2.1. TIPO DE LA INVESTIGACIÓN

El tipo de investigación es descriptivo y exploratorio, ya que se describen los datos con el objetivo de encontrar reglas de asociación entre las instituciones analizadas.

3.2.2. DISEÑO DE LA INVESTIGACIÓN

Es una investigación no experimental con diseño transversal debido a que se obtienen los datos de una etapa prolongada de tiempo, pero sólo se analizan para el periodo en conjunto.

3.2.3. IDENTIFICACIÓN DE LAS VARIABLES

Las fuentes de datos de esta investigación son los artículos con participación de autores peruanos indizados en Scopus publicados entre el 2000 y el 2015. Las variables son las interacciones que tiene el conjunto de instituciones que forman parte de la red.

3.2.4. POBLACIÓN

El primero en mencionar los problemas con el muestreo en redes fue Granovetter (1976) quien se interesó en la determinación del grado medio de los vértices sirviéndose de múltiples muestras. Sin embargo, si el objeto de investigación no es el grado medio de los vértices, sino alguna característica más compleja de la red, utilizar muestreo no es conveniente. La mayoría de redes reales son redes libres de escala (ver más sobre el concepto en Barabási (2016)), y las subredes de las redes libres de escala, no son a su vez libres de escala, es decir las características de la red no se mantienen cuando hay muestreo (Stumpf et al. 2005). Es por eso que en este trabajo el análisis se aplica sobre la población de interés.

Se trabajó con los datos sobre coautorías entre instituciones peruanas. Las fuentes de información fueron 5697 documentos producidos entre el 2000 y el 2015, indizados en la base de datos Scopus en la categoría medicina y que tuvieron participación de alguna institución con sede en el Perú. En la muestra sólo se tomó en cuenta a las instituciones peruanas y las interacciones entre ellas. Las instituciones peruanas en la población son 613. Las aristas en el multigrafo compuesto por estas son 6453.

Datos. Para el análisis, se realizó la descarga directa de la base de datos Scopus de todo el volumen de artículos publicados por autores pertenecientes a instituciones peruanas durante el período 2000-2015. Se buscó Perú en el campo «Affiliation country». La recuperación se realizó el 16 de junio del 2016. Los registros fueron descargados en tres bloques: del 2000 al 2004, del 2005 al 2010, del 2011 al 2013, del 2014 al 2015, en ficheros «.csv», para su procesamiento en R. Además, se descargó los datos de entrenamiento provenientes de un trabajo previo de Málaga (2014) y se les dio un procesamiento de limpieza.

3.2.5. METODOLOGÍA APLICADA

1. Análisis exploratorio de los datos descargados y una limpieza para utilizarlos como datos de validación.

2. Exploración usando métodos de conglomerados y de clasificación para escoger el más adecuado para el procesamiento de los datos. De manera no supervisada se exploró los resultados con el *q-grams* (n-gramas) del paquete stringdist y con k-medias. Los métodos supervisados que se utilizaron fueron: clasificador Máquinas de Soporte Vectorial, Random Forest, Bagging, K-Vecinos más cercanos con k=1,3,5,10.

3. Procesamiento de los datos de validación con el algoritmo de clasificación de k vecinos cercanos con k=1. Se hizo una revisión manual de una muestra de los resultados y se procedió a repetir la clasificación sobre los valores restantes. El proceso se repitió varias veces.

4. Análisis exploratorio del grafo resultante y eliminación de los nodos y componentes aislados.

5. Identificación de los conglomerados y descripción de sus características

IV. RESULTADOS Y DISCUSIÓN

4.1. ANÁLISIS EXPLORATORIO DE DATOS

Para el análisis, se realizó la descarga directa de la base de datos Scopus de todo el volumen de artículos publicados por autores pertenecientes a instituciones peruanas durante el período 2000-2015. Se buscó Perú en el campo «Affiliation country». La recuperación se realizó el 16 de junio del 2016. Los registros fueron descargados en tres bloques: del 2000 al 2004, del 2005 al 2010, del 2011 al 2013, del 2014 al 2015, en ficheros «.csv», para su procesamiento en R. Para entender bien el tipo de interrelaciones que se van a explicar en este documento es necesario entender el tipo de datos a partir de los cuales se extrae la información ya que la elección de la base de datos influye sobre el resultado final (ver entre otros Mongeon y Paul-Hus 2015).

El fichero csv descargado contiene 30 columnas (variables) y 5697 filas - cada una correspondiente a un documento. Las variables incluidas son:

- *Authors*. Este campo contiene una lista de autores del documento separados por comas. Ejemplo: «Carreazo N.Y., Bada C.A., Chalco J.P., Huicho L.»
- *Title.* Este campo contiene el título del documento. Ejemplo: «Audit of therapeutic interventions in inpatient children using two scores: Are they evidence-based in developing countries? »
- *Year*. Este campo contiene el año de publicación del documento
- *Source.title*. Este campo contiene el título de la revista o libro en la que se publicó el artículo o el capítulo de libro. En el caso de los libros este campo contiene el mismo valor que el campo *Title.* Ejemplo: «BMC Health Services Research»

- *Volume*. Este campo contiene el número de volumen de la revista. No siempre se llena, ya sea porque la revista no tiene número de volumen, ya sea porque el documento no se ha publicado en una revista.
- *Issue*. Número de la revista. No siempre se llena, ya sea porque la revista no tiene número, ya sea porque el documento no se ha publicado en una revista.
- *Art..No*. Número de artículo. Algunas revistas utilizan este campo para identificar al artículo de manera única.
- *Page.start.* Página en la que inicia el artículo. Sólo disponible en algunos casos.
- *Page.end.* Página en la que termina el artículo. Sólo disponible en algunos casos.
- *Page.count*. Extensión total del artículo. Sólo disponible en algunos casos.
- *Cited.by*. Número de veces que el artículo fue citado. Es una información que puede ser engañosa ya que los artículos más recientes son citados menos veces ya que estuvieron disponibles por un periodo más breve. Cuando no existe información de que el artículo haya sido citado este campo queda vacío.
- *DOI*. El identificador digital de objeto, conocido en inglés como digital object identifier y abreviado DOI, es un sistema parecido a los identificadores URI. Permite dar a las publicaciones científicas un número específico que cualquiera puede utilizar para localizar a través de la red el citado artículo. Aún no es de uso obligatorio por lo que no siempre existe información sobre este, especialmente para los textos más antiguos. Ejemplo: «10.1186/1472-6963-4-40».
- *Link*. Enlace URL que da acceso a los datos dentro de Scopus. Ejemplo: «https://www.scopus.com/inward/record.uri?eid=2-s2.0-14544294467&partnerID=40&md5=5f436a86fe45f71ebf6480913bdc0052»
- *Affiliations*. Afiliaciones con las que firmaron los autores del documento. Ejemplo: «Univ. Nacional Mayor de San Marcos, Instituto de Salud del Niño, Lima, LI 05, Peru».

- *Authors.with.affiliations.* Nombre de los autores junto a la afiliación correspondiente separados por puntos y comas. Ejemplo: «Carreazo, N.Y., Univ. Nacional Mayor de San Marcos, Instituto de Salud del Niño, Lima, LI 05, Peru; Bada, C.A., Univ. Nacional Mayor de San Marcos, Instituto de Salud del Niño, Lima, LI 05, Peru; Chalco, J.P., Univ. Nacional Mayor de San Marcos, Instituto de Salud del Niño, Lima, LI 05, Peru; Huicho, L., Univ. Nacional Mayor de San Marcos, Instituto de Salud del Niño, Lima, LI 05, Peru».

- *Abstract*. Resumen proporcionado por los autores. Sólo está disponible cuando los autores proveen la información correspondiente. Ejemplo: «Background: The evidence base of clinical interventions in paediatric hospitals of developing countries has not been formally assessed. We performed this study to determine the proportion of evidence-based therapeutic interventions in a paediatric referral hospital of a developing country. Methods: The medical records of 167 patients admitted in one-month period were revised. Primary diagnosis and primary therapeutic interventions were determined for each patient. A systematic search was performed to assess the level of evidence for each intervention. Therapeutic interventions were classified using the Ellis score and the Oxford Centre for Evidence Based Medicine Levels of Evidence. …»

- *Author.Keywords.* Palabras clave proporcionadas por los autores. Sólo está disponible cuando los autores proveen la información correspondiente.

- *Index.Keywords*. Palabras clave utilizando un lenguaje controlado asignadas manualmente por los procesadores de la base de datos. Ejemplo: «adrenalin; albendazole; amikacin; ampicillin; analgesic agent; antibiotic agent; antipyretic agent; beta 2 adrenergic receptor stimulating agent; captopril; ceftazidime; ceftriaxone; chloramphenicol; ciprofloxacin; clindamycin; corticosteroid; dexamethasone; diazepam; …»

- *Correspondence.Address*. La dirección de correspondencia proporcionada por el coordinador del proyecto. Incluye información sobre el coordinador, su afiliación y un correo electrónico. Ejemplo: «Huicho, L.; Univ. Nacional

Mayor de San Marcos, Instituto de Salud del Niño, Lima, LI 05, Peru; email: lhuicho@viabcp.com».

- *Editors*. Información sobre los editores. Sólo está disponible en nueve de los casos analizados.
- *Publisher*. Información sobre la editorial, para los datos analizados está disponible en el 62 por ciento de los casos.
- *ISSN*. International Standard Serial Number o Número Internacional Normalizado de Publicaciones Seriadas es un número internacional que permite identificar de manera única los diarios y las publicaciones periódicas. Ejemplo: «14726963».
- *ISBN*. El Estándar Internacional de Libros o Número Internacional Normalizado del Libro (International Standard Book Number) es un identificador único para libros, previsto para uso comercial.
- *CODEN*. Es un código alfanumérico de seis caracteres, aplicado con fines bibliográficos de acuerdo con la norma ASTM E250, que proporciona una identificación concisa, única e inequívoca de los títulos de publicaciones científicas periódicas y publicaciones no periódicas. Para los datos analizados está disponible en el 62 por ciento de los casos.
- *PubMed.ID*. Es un identificador único para hallar los documentos indizados en MEDLINE – base de datos de citaciones y resúmenes de artículos de investigación biomédica. Ejemplo: «15625006».
- *Language.of.Original.Document*. Lengua del documento. Ejemplo: «English».
- *Abbreviated.Source.Title*. Título abreviado de la fuente del documento (revista o libro). Ejemplo: «BMC Health Serv. Res.».
- *Document.Type.* Aunque la mayoría de documentos son artículos hay también otros tipos de textos.
- *Source*. Fuente de información. Esta variable no es informativa ya que todos los datos corresponden a Scopus.

- *EID*. Código de identificación único asignado por Scopus. Ejemplo: «2-s2.0-14544294467».

Sólo la columna correspondiente a afiliaciones (*affiliations*) se utilizará para este estudio, pero es necesario almacenar el resto para poder identificar cada documento de manera única, si se quiere agregar documentos a la base de datos en el futuro. Además de los campos anteriores, en Scopus existen campos con información sobre el financiamiento de los proyectos, pero estos, rara vez son llenados en el caso de los documentos firmados por las instituciones peruanas. Adicionalmente, existe también información sobre las referencias bibliográficas, pero se consideró que no eran de interés en este caso. En los anexos está el código R de las transformaciones a las que se sometió los datos para obtener el grafo final.

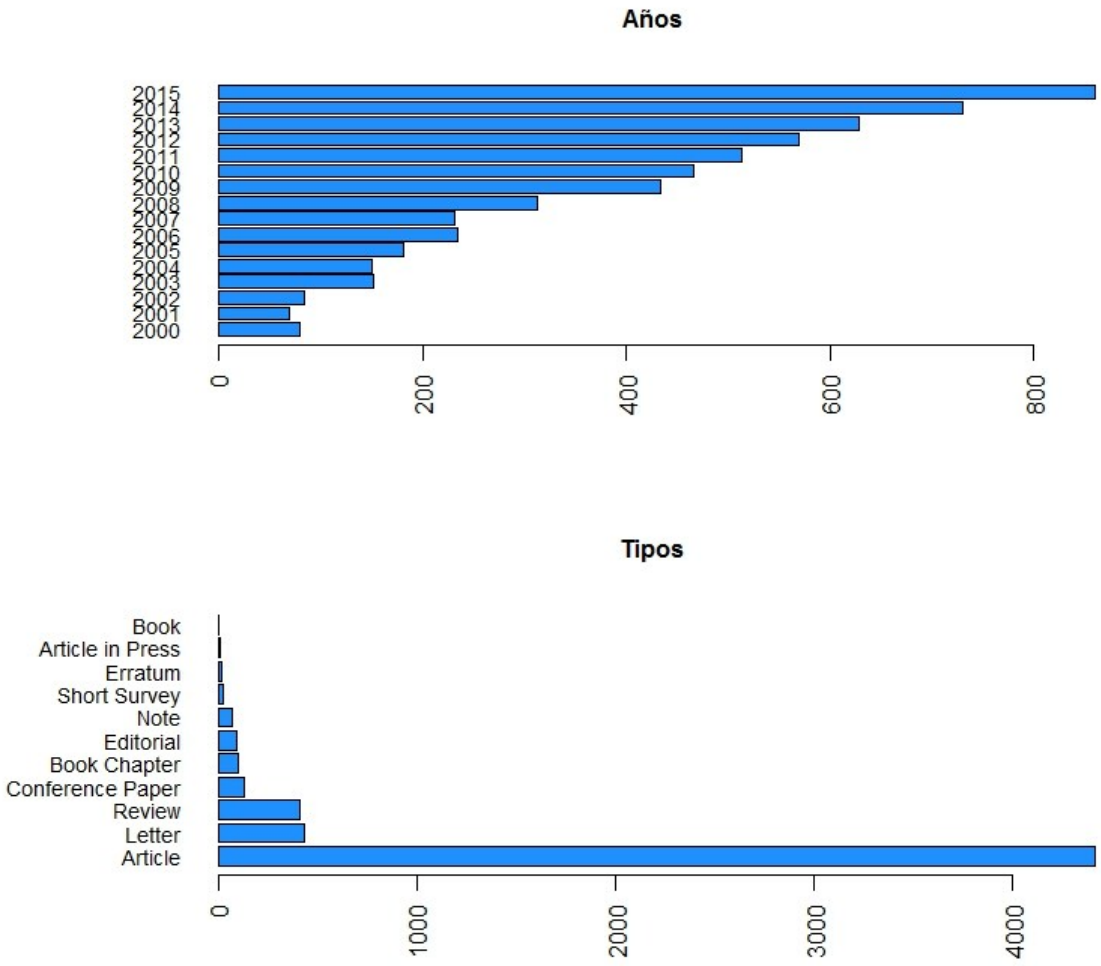
Si bien los artículos están fechados entre el 2000 y el 2015, el uso y la cobertura de Scopus ha ido creciendo en ese periodo por lo que la mayoría de documentos analizados proviene de los años 2011 a 2015. Además, Scopus (al igual que WOS) cubre principalmente artículos de investigación. Por eso, al igual que ocurre con la base de datos que sirve de fuente de información, la gran mayoría de los documentos analizados son artículos, seguidos muy de lejos por cartas y artículos de revisión, otros tipos de documentos constituyen un porcentaje insignificante (8 por ciento) del total (ver fig. 11).

Otra característica de la base de datos Scopus es la mayor cobertura de textos en inglés (sólo 21 por ciento de las revistas publicadas están en otros idiomas) (Scopus content coverage 2016). Los documentos peruanos no se alejan demasiado de esa tendencia, alrededor del 76 por ciento está en inglés. El idioma es reconocido por el sistema a partir de la descripción de la revista, por lo que hay cierta imprecisión en cuanto a los documentos individuales, pues hay revistas que aceptan documentos en hasta tres idiomas, en ese caso no es posible determinar en qué idioma en concreto se escribió el artículo analizado. La revista que más se repite entre los artículos es una revista que publica en español: la Revista Peruana en Medicina Experimental y Salud Pública del Instituto Nacional de Salud. Con 755 documentos presentes en el periodo de estudio, la presencia de esta revista es probablemente la razón por la que la tasa de documentos en inglés es algo inferior a la tasa de documentos en inglés en todo Scopus. Es además una muestra de lo importante que es contar con cobertura de revistas nacionales para obtener datos representativos (ver fig. 12).

Otro tema de interés son las citas, es decir con qué frecuencia son citados los documentos peruanos. Esta medición sólo tiene sentido si se compara con otras instituciones o países.

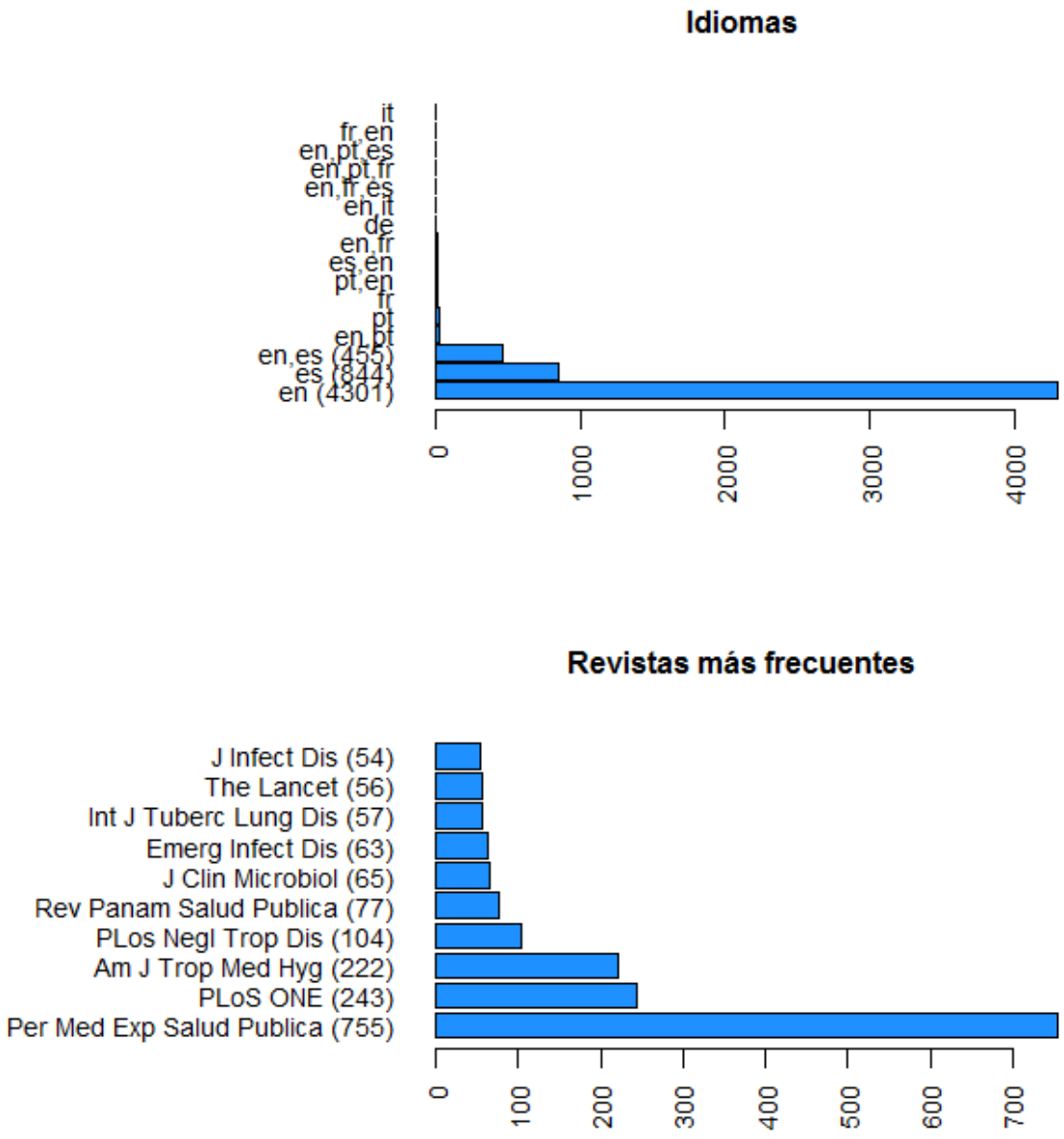
Perú tiene un promedio de citación en medicina por artículo bastante alto, por encima de la media mundial (Málaga 2014), pero esta medida debe tomarse con cautela pues el número de documentos de Perú no es grande y eso significa que unos cuantos outliers influyen significativamente sobre el promedio (ver fig. 13).

Figura 11: Cantidad de documentos analizados por año y tipo.



FUENTE: Elaboración propia con datos de Scopus.

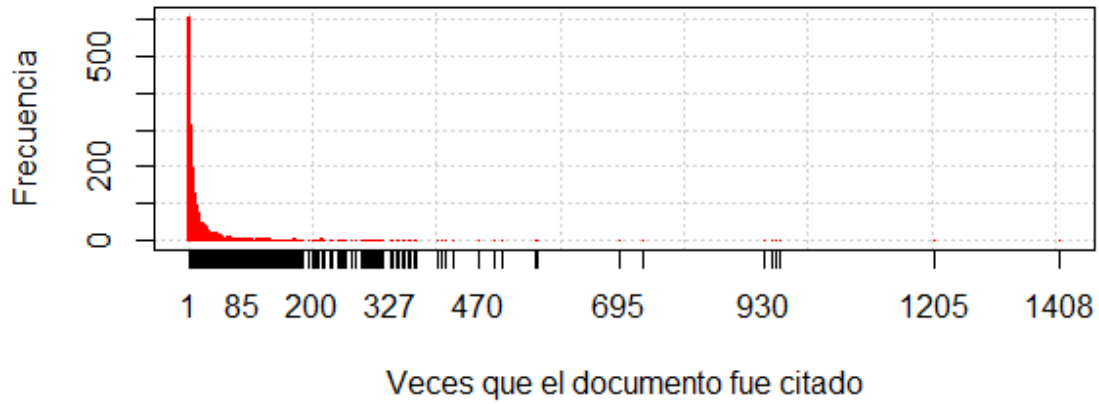
Figura 12: Idioma y revistas más frecuentes entre los documentos analizados



FUENTE: Elaboración propia con datos de Scopus.

Las siglas de los idiomas siguen la norma ISO 639-1. Las abreviaciones de las revistas siguen las normas de abreviación de la «List of Title Word Abbreviations» del International Standard Serial Number International Centre. Ver detalle en lista de acrónimos y siglas.

Figura 13: Distribución de las citaciones por documento

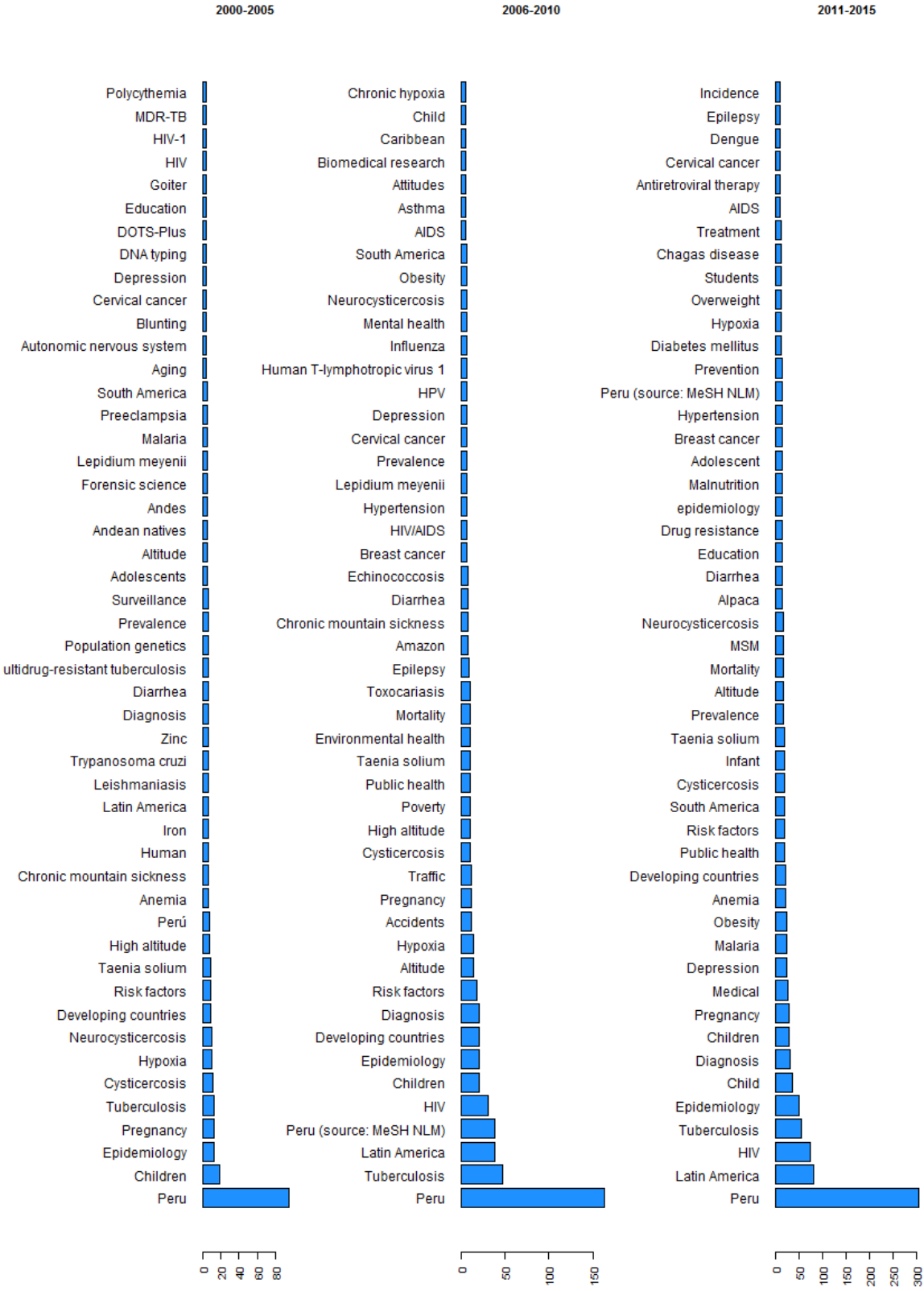


FUENTE: Elaboración propia con datos de Scopus.

Este análisis se limita a los documentos de medicina pues constituyen la parte más grande de lo producido por el Perú y, al describir resultados de investigación de una sola disciplina, tienen características similares que permiten que sean comparables. Eso no impide que existan importantes diferencias entre los documentos en cuanto a los contenidos. Para saber más sobre los temas que cubren es adecuado revisar las palabras clave. En Scopus existen dos tipos de palabras clave: (1) las proporcionadas por los autores, (2) las obtenidas mediante procesamiento manual de los documentos que se sirven de una serie de vocabularios controlados (Scopus content coverage 2016). Las palabras clave proporcionadas por autores son más informativas sobre temas específicos de los documentos, mientras que las palabras de índices permiten navegar más fácilmente a nivel de temas generales. Sin embargo, en el conjunto de documentos analizado hay varios documentos (2509) que no tienen palabras clave proporcionadas por autores, sea por que no existen, sea por que no fueron ingresadas en el sistema.

Como puede verse en la figura 14 el tema más frecuente de los documentos analizados es el Perú. Algunos temas específicos (palabras clave) que se repiten son: América Latina, Tuberculosis, VIH/SIDA, Cisticercosis, Altitud. Curiosamente las enfermedades tropicales no aparecen con tanta frecuencia, pero esto puede deberse al número de documentos sin palabras clave en el sistema.

Figura 14: Palabras clave asignadas por los autores más frecuentes entre los documentos (por quinquenios 2000 - 2015).



FUENTE: Elaboración propia con información de Scopus.

En conclusión, el conjunto de artículos que se utilizará para el análisis no se diferencia del total de documentos indizados en la base de datos Scopus. La presencia de documentos en español es ligeramente superior, pero los datos con los que se cuenta (un porcentaje proporcionado por Elsevier en el que no se especifica cómo contabilizan las revistas multilingües) no permiten hacer una comparación sobre lo significativo de esta diferencia.

4.2. EXPLORACIÓN DE MÉTODOS DE CONGLOMERADOS Y DE CLASIFICACIÓN PARA LA IDENTIFICACIÓN DE LAS INSTITUCIONES

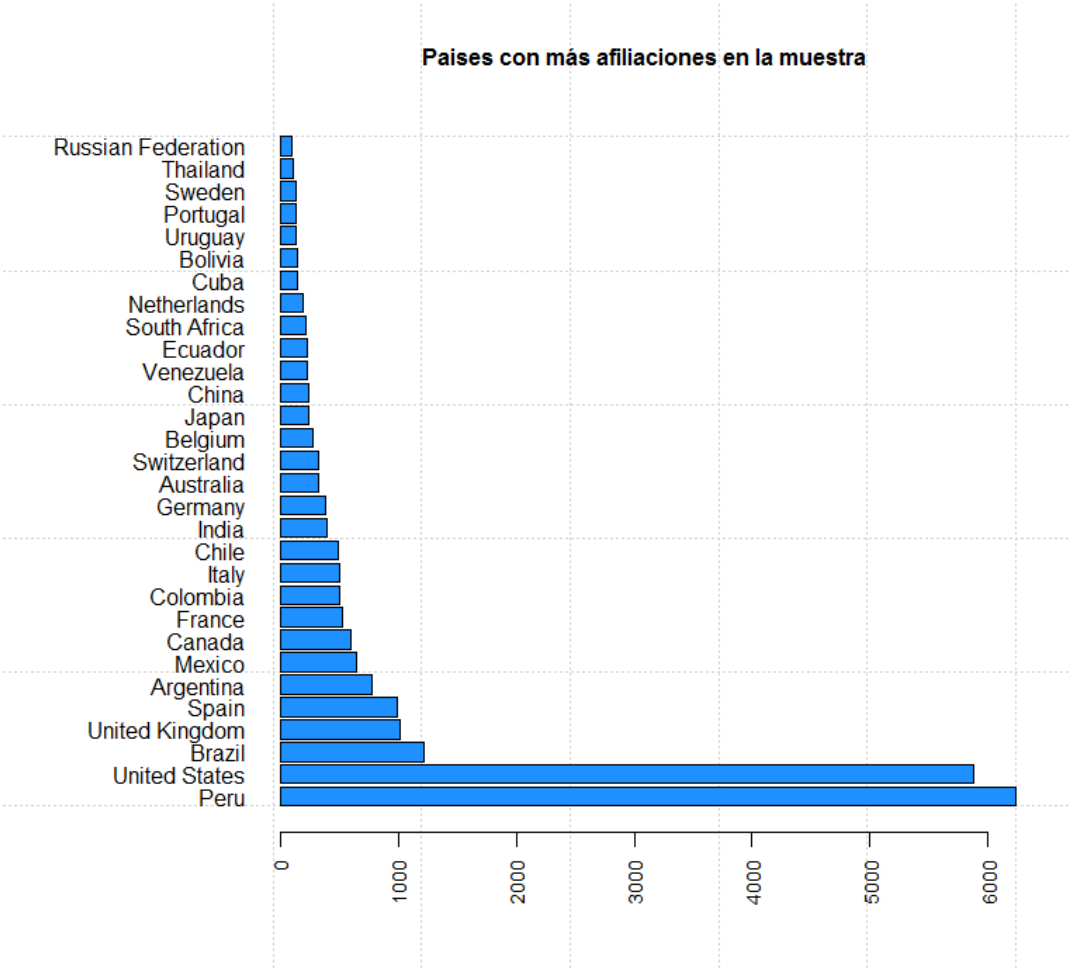
Como ya se mencionó, son 5697 los documentos que se recuperaron con la búsqueda de Perú en el campo de afiliación. Entre estos, son de interés sólo los que tienen más de una afiliación. Para identificarlos se revisó el campo de afiliaciones y se escogió sólo las filas con por lo menos un punto y coma, que es el signo que separa en los datos las afiliaciones. Descartando los documentos firmados por una sola institución se tiene 5073 documentos firmados por más de una institución. En algunos casos esto corresponde a diferentes dependencias de la misma institución (bucles) pero no es algo que sea de nuestro interés en esta etapa.

A partir de los 5073 documentos se identificó 35879 firmas de afiliaciones presentes en los artículos. Eliminando las afiliaciones redactadas exactamente de la misma manera se tiene como resultado 26524 maneras únicas en las que se han redactado las diferentes afiliaciones. Estos son los datos con la que se trabajará en adelante. Las afiliaciones de interés para este estudio son las asociadas a instituciones peruanas. En cada una de las afiliaciones el país debería ser el último elemento de la cadena de caracteres, antecedido por una coma. Sin embargo, los datos que se tienen a mano no están normalizados, por lo que hay 155 casos en los que la coma no existe. En total se corrigieron 167 países correspondientes a las instituciones. El país más frecuente entre las instituciones es, como era de esperarse, Perú (6236 casos), seguido muy de cerca por Estados Unidos (5882 casos). Los otros países siguen a estos dos muy de lejos (ver fig. 15).

Para este análisis se escogió solo los datos correspondientes a instituciones peruanas. La identificación de manera única de las instituciones en este corpus tiene que superar dos retos: (1) los datos son desbalanceados, al igual que lo es la producción científica por instituciones (y personas) (ver Lotka 1926); (2) aunque se cuenta con un conjunto de entrenamiento exhaustivo para los años 2009-2011, la producción científica es un campo dinámico e

instituciones que antes no habían publicado comienzan a hacerlo continuamente, por ello, no se tiene un marco de las instituciones, sólo una lista parcial.

Figura 15: Países de las afiliaciones institucionales.



FUENTE: Elaboración propia con datos de Scopus

Este gráfico representa la presencia de afiliaciones institucionales tal como se han escrito, sin una estandarización adicional. Esto significa que la mayoría de instituciones se cuenta más de una vez. Es suficiente que haya diferencia de una coma con las otras afiliaciones, para que las variantes se cuenten por separado.

La información sobre las instituciones, tal como la entrega Scopus, no está estandarizada y las instituciones no están identificadas de manera única. Sin embargo, el portal se sirve de una mezcla de métodos de conglomerados y de clasificación para realizar la identificación

para uso propio. Esta identificación no es la más precisa, aunque bastante acertada en el caso de las instituciones con mucha investigación. El problema es que la recuperación de las identificaciones realizadas por Scopus sólo puede darse a partir de una búsqueda institucional. Es decir, es necesario descargar la información sobre lo producido por cada una de las instituciones por separado y conocer con anterioridad cuáles son las instituciones existentes. Además, Scopus sólo identifica a las instituciones con más producción y no le interesan las instituciones «pequeñas» de menos de 10 documentos. Mientras menos haya escrito una institución, es más probable que la clasificación de Scopus tenga errores.

En este caso nos interesaba la identificación de instituciones pequeñas con poca investigación ya que en redes de colaboración, especialmente a nivel local, la fuerza de los *weak ties* es un aspecto importante de una red social y en particular de una red de coautorías científicas (Granovetter 1973, Easley y Kleinberg 2010, Pan y Saramäki 2012).

Se exploró métodos de conglomerados y de clasificación para escoger el más adecuado para el procesamiento de los datos.

Métodos no supervisados

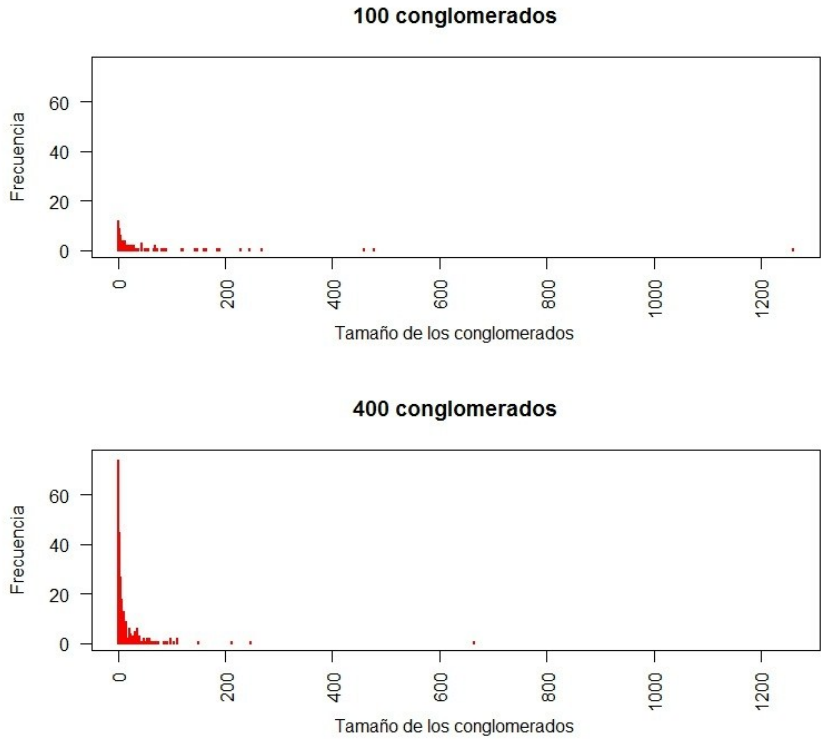
N-gramas (stringdist)**.** – Este método es el único de los utilizados que no se sirve de la transformación de la descripción de las afiliaciones en una matriz de documentos-textos (dtm). La lógica que sigue es la similitud entre cadenas de caracteres y el costo que en términos de transformación del texto original implicaría conseguir el texto con el que se compara el original (*Distances based on edit operations*). En n-gramas no se compara todo el texto sino bloques de cadenas de caracteres (Van der Loo 2014). Un n-grama es una subsecuencia de n elementos consecutivos en una secuencia dada. En este caso, por el número de datos a comparar, los n-gramas no pueden ser muy grandes, se comparó con n-gramas de tamaño cuatro. Es rápido y es un paquete adecuado para identificar semejanzas entre cadenas de caracteres relativamente cortas, en el caso de las afiliaciones el reconocimiento de similitudes es adecuado. Escogiendo un grupo al azar de los 100 conglomerados obtenidos se encontró que un solo conglomerado (de 30 elementos) incluía solo una institución. Cuando se trabajó con 400 conglomerados, un solo conglomerado (de 66 elementos) incluía dos instituciones. A pesar de todo, su aplicación se enfrenta a varias limitaciones:

- Existen instituciones con nombres semejantes – estas instituciones son agrupadas a causa de esa similitud.

- Es necesario indicar previamente el número de conglomerados a hallar.
- Sólo reconoce conglomerados para instituciones que se repiten frecuentemente en los datos.
- Agrupa elementos disímiles si tienen poca frecuencia en los datos.

Adicionalmente, como puede verse en la figura 16, conforme aumenta el número de conglomerados, el número de clústeres compuestos de un solo elemento crece rápidamente.

Figura 16: Resultados de la utilización del método de n-gramas para la identificación de instituciones



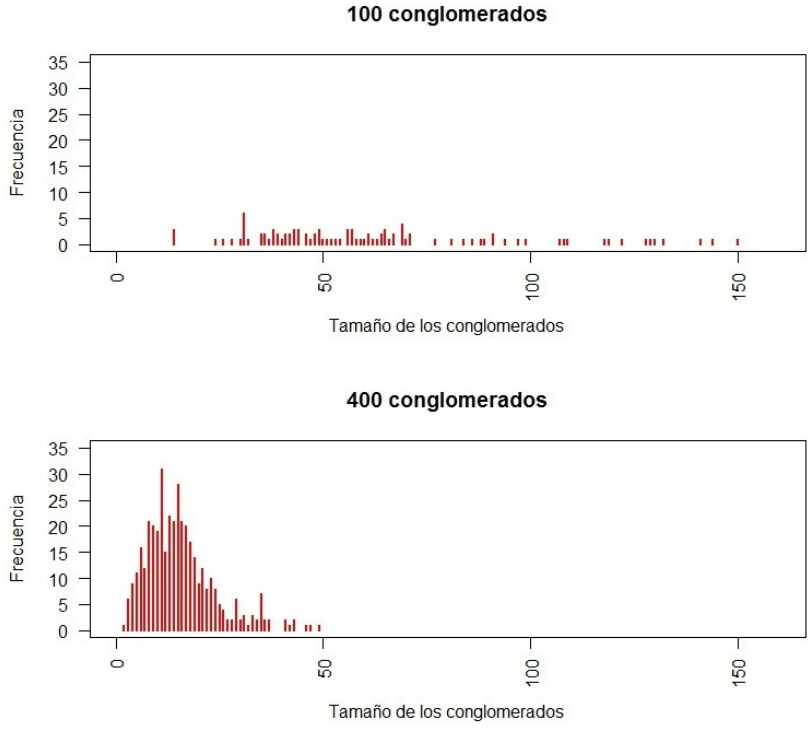
FUENTE: Elaboración propia con datos de Scopus.

K-medias. – El método de k-medias permite la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo más cercano a la media. La desventaja de este método es que es necesario introducir el número de conglomerados a encontrar antes de iniciar el proceso. Además, reconoce mejor los conglomerados para instituciones que se repiten frecuentemente en los datos.

Resultados. Se procedió a realizar una identificación de conglomerados exploratoria con 100 conglomerados (un número muy inferior al número de instituciones finalmente hallado). Se

puede ver que los conglomerados pequeños (de hasta 5 elementos) constituyen una minoría en el resultado (ver fig. 17). Esto significa que hay muchos elementos que aparecen pocas veces en los datos, que son de diferentes instituciones, pero terminan siendo agrupados en un solo conglomerado. Las afiliaciones que aparecen con más frecuencia en los datos son probablemente las que tienen mejor ajuste en los resultados. Cuando se aumenta el número de conglomerados a 400, esta tendencia disminuye, pero comienzan a «desmenuzarse» los grupos grandes. Escogiendo un grupo al azar de los conglomerados obtenidos con k=100 se encontró que un solo conglomerado (de 59 elementos) incluía por lo menos seis instituciones (todos hospitales). Escogiendo un grupo al azar de los conglomerados obtenidos con k=400, se encontró que un conglomerado (de 19 elementos) incluía siete diferentes instituciones (todas universidades). En suma, el método permite identificar instituciones similares, pero le falta la precisión de un método supervisado.

Figura 17: Resultados de la utilización del método de k-medias para la identificación de instituciones



FUENTE: Elaboración propia con datos de Scopus.

Métodos supervisados

De un trabajo anterior de Málaga (2014) se contaba con datos de entrenamiento exhaustivos para los años 2009 – 2011. El total de diferentes formas de afiliaciones incluidas en esos datos son 4191. Las afiliaciones identificadas en los datos de entrenamiento son 658. La desventaja de estos datos era la falta de consistencia en la codificación que generó ciertos errores en el resultado final. Se realizó una limpieza previa para eliminar, por lo menos en parte, los errores generados por la codificación. De manera exploratoria se trabajó con una muestra de 90 por ciento de los datos para entrenar y 10 por ciento de los datos para validación, para conocer cuál de los métodos supervisados sería más adecuado.

Support Vector Machine, Random Forest y Bagging. - Para la revisión de los métodos *Máquina de Soporte Vectorial, Random Forest,* y *Bagging,* se usó el paquete *RTextTools* (Jurka et al. 2014) que permite hacer la clasificación de textos con todas las medidas de precisión de la predicción.

Resultados. Los datos que se quiere clasificar son desbalanceados y los resultados no fueron satisfactorios, además de eso, el tiempo de procesamiento es excesivo (especialmente para Random Forest). En todos los casos la precisión y la exhaustividad tienen valores bajos. En cuanto a los tiempos de procesamiento, la utilización de las máquinas de soporte vectorial fue la mejor opción (ver Cuadro 3).

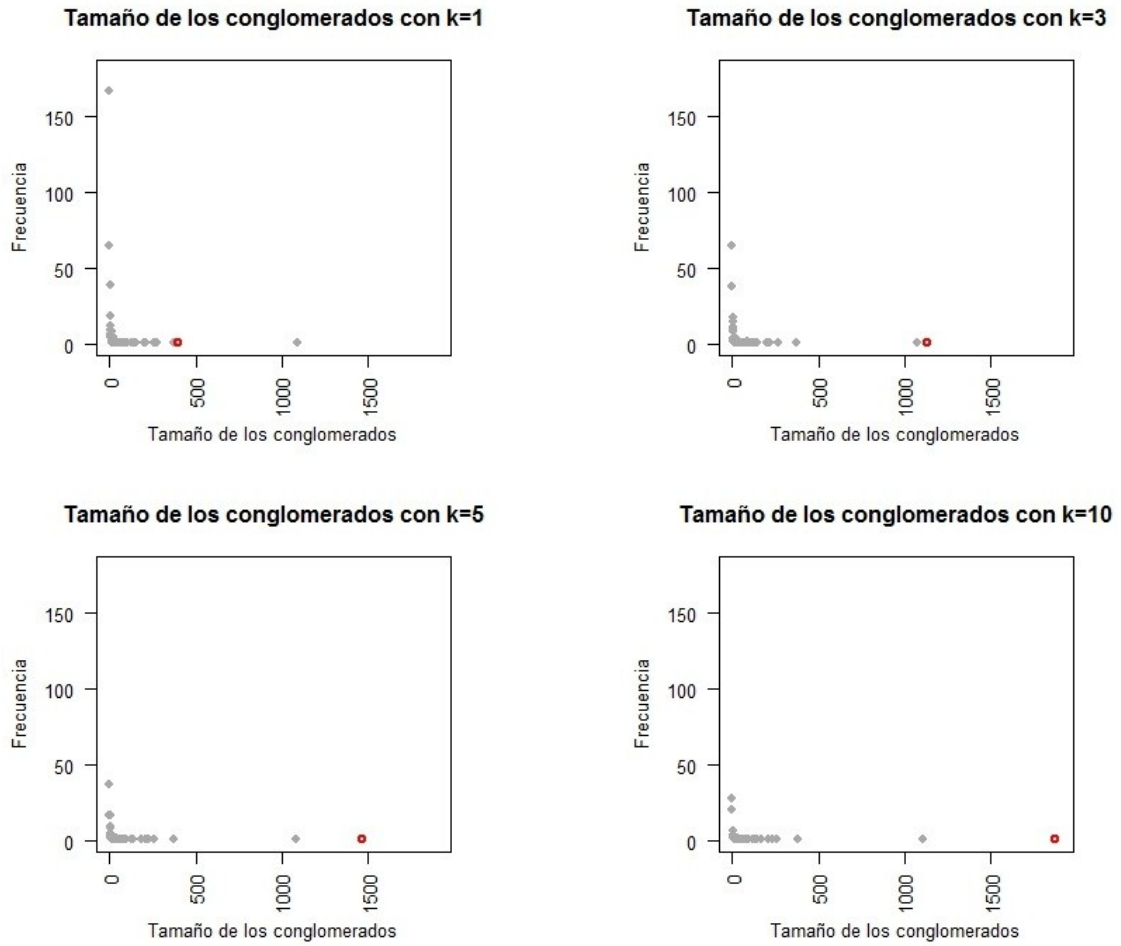
Cuadro 3: Resultados de la utilización de los métodos de conglomerados supervisados para la identificación de las instituciones

Método	*Validación cruzada (4F)*	*Precision*	*Recall*	*Fscore*	*Recall accuracy*	*Tiempo (CV–transcurrido)*	*Tiempo (CV–usuario)*
Support Vector Machine	0.3518	0.0437	0.0513	0.0448	0.00159	336.03	313.22
Random Forest	0.3762	0.1033	0.0803	0.0847	0.00265	4346.97	4345.21
Bagging	0.2976	0.2079	0.1901	0.1873	0.00106	619.98	638.69

FUENTE: Elaboración propia

K-vecinos más cercanos. – Después de tener resultados medianamente satisfactorios con Máquinas de Soporte Vectorial aplicadas a los datos de validación, se procedió a explorar la clasificación con k-vecinos más cercano con valores de 1, 3, 5, 10. Conforme aumenta el valor de k, aumenta el tamaño de los conglomerados, pero también aumenta el número de elementos que no pueden ser asignados (ver puntos rojos en figura 19). Entre estos, k-vecinos más cercanos con k=1 tiene la ventaja sobre otros métodos de clasificar elementos con muy pocas ocurrencias, que son muy frecuentes en los datos, y que ningún otro método pudo identificar. Además, es el método que permite clasificar el mayor número de elementos.

Figura 18: Resultados de la utilización del método de k-vecinos más cercanos sobre los datos de entrenamiento para la identificación de instituciones



FUENTE: Elaboración propia.

El algoritmo devuelve NA (puntos rojos) cuando no existe suficiente información para asignar la clasificación.

4.3. CLASIFICACIÓN CON K-VECINOS MÁS CERCANOS

Como se mencionó anteriormente: (1) la primera elección en cuanto al método de clasificación fue *support vector machines* que dio buenos resultados para los elementos a clasificar que más aparecen en los datos de entrenamiento, pero cuyo clasificador es tiempo-demandante y no encuentra los elementos «pequeños». En un primer intento el modelo sólo asignó los datos de validación a quince clases. En el proceso de encontrar el método de clasificación más adecuado, fragmentos de los datos de validación se fueron asignando manualmente a las clases correctas e incluidos en los datos de entrenamiento; (2) finalmente, se decidió usar k-medias con k=1, el único método que permitía recuperar la clasificación de ítems poco frecuentes.

Se procesó los datos de validación con el algoritmo de clasificación de k vecinos cercanos con k=1. Se hizo una revisión manual de una muestra de los resultados y se procedió a repetir la clasificación sobre los valores restantes. El proceso se repitió hasta completar la revisión de datos.

4.4. ANÁLISIS EXPLORATORIO DEL GRAFO

4.4.1. LOS ACTORES

Se toma en cuenta las instituciones que han publicado un documento en coautoría indizado en Scopus entre el 2000 y el 2015. El total de documentos analizados son 5073 artículos. En algunos casos la firma en el campo afiliación no es suficiente para identificar una institución de manera única (por ejemplo, sólo siglas, sólo una o dos palabras, sólo una dirección) o no corresponde a una institución. Cada uno de estos casos se consideró como una institución separada.

Se ha reconocido la presencia de coautorías a partir de la existencia de puntos y comas en el campo «afiliaciones» de los datos descargados. En algunos casos esto puede significar que la «co-autoría» que se está señalando en el campo es sólo entre diferentes dependencias de la misma institución. En ese caso la coautoría analizada constituiría un bucle y como tal no es tomada en cuenta para el análisis del grafo. Se han identificado 733 instituciones peruanas que han publicado un documento en coautoría en el periodo de análisis. De estas 733 instituciones, previo descarte de las coautorías en bucle (colaboraciones entre dependencia de la misma institución), 613 han escrito algún documento en coautoría con una institución

peruana en el periodo de análisis. Eso significa que el 83 por ciento de las instituciones peruanas que escriben en colaboración lo hacen (por lo menos en una fracción de sus publicaciones) con instituciones peruanas. El total de coautorías tomadas en cuenta es 2278.

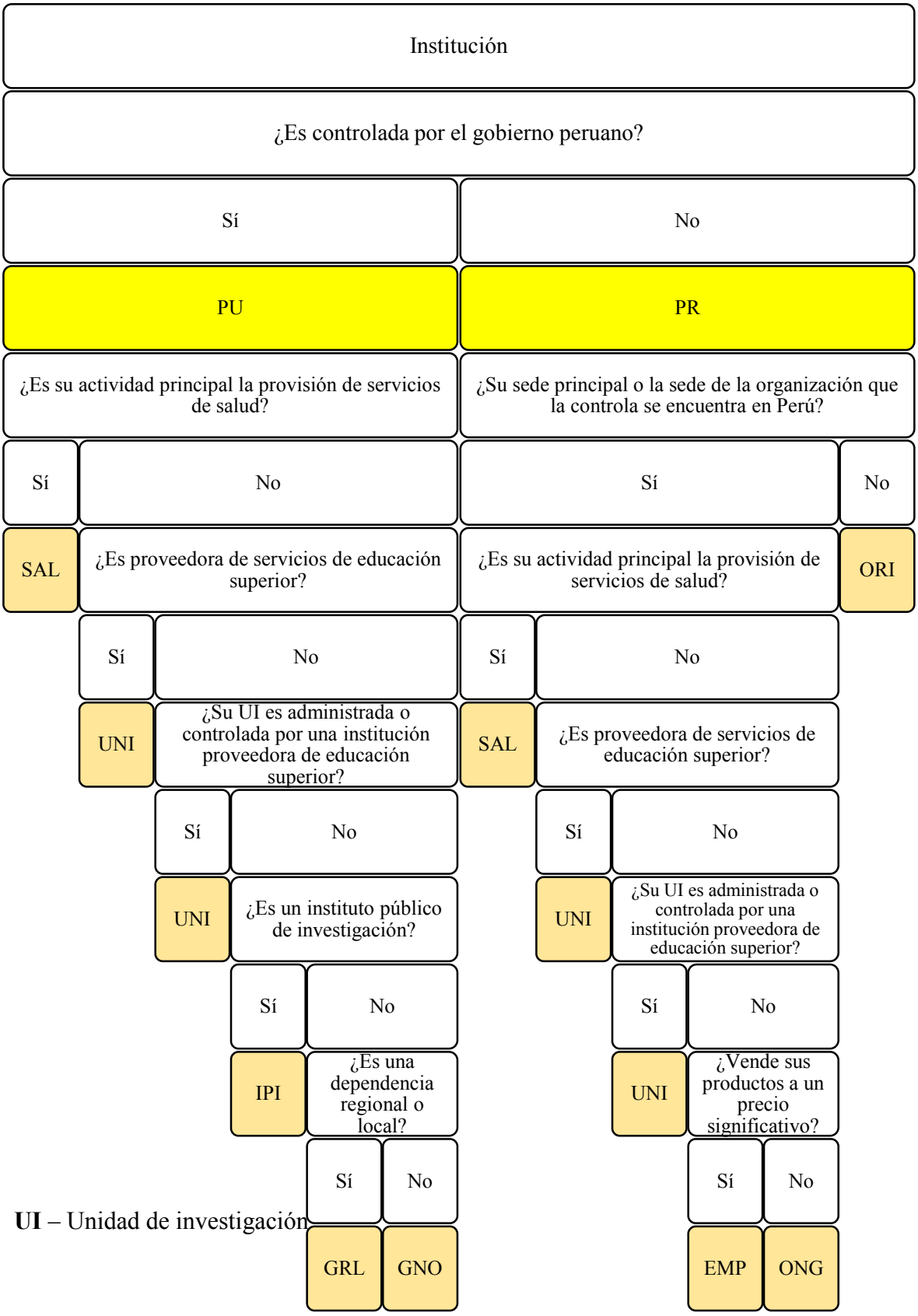
Las instituciones fueron clasificadas por sector (público o privado) y tipo (servicios de salud, universidad, instituto público de investigación, gobierno regional, gobierno local, empresa, instituciones privadas sin fines de lucro, organizaciones internacionales con sede en el Perú). Para la asignación de la clasificación se siguió los criterios del árbol de decisión en la figura 19 y los resultados se pueden ver en la figura 20.

Posteriormente se limitó el análisis al gran componente, que acumula el 93 por ciento de los vértices en el grafo. Se analizó la importancia que cumplen los diferentes nodos en la red. Se tomó en cuenta las siguientes medidas de centralidad: grado, grado ponderado, cercanía, intermediación, centralidad del vector propio. En todas las medidas de centralidad la Universidad Peruana Cayetano Heredia (UPCH) y la Universidad Nacional Mayor de San Marcos (UNMSM) tienen un papel predominante en la red. Los vértices más importantes en grado y grado ponderado son los dos ya mencionados y el Instituto Nacional de Salud (INS), el Ministerio de Salud (MINSA), la Unidad de Investigación de la Marina de Estados Unidos (NAMRU), la Universidad Peruana de Ciencias (UPC) y la Asociación Benéfica PRISMA (PRISMA). En estos casos, las instituciones son centrales en la red local, es decir son vecinas de varios vértices, independientemente si estos a su vez son o no centrales en la red (ver Cuadro 4).

La medida de cercanía refleja la posición central en la red, es decir qué tan central es un vértice en los diferentes caminos que unen a la red. En este grafo, los dos vértices ya mencionados son los que tienen un papel más central, seguidos de cerca (es una característica de esta medida que el rango de valores que abarca es pequeño) por la Universidad Nacional José Faustino Sánchez Carrión (UNJFS). Las siguientes instituciones en esta medida tienen valores de cercanía muy similares.

La medida de intermediación refleja la capacidad de correduría de un vértice, es decir se basa en el número de caminos más cortos que pasan por ese vértice. En este caso la intermediación más alta la tienen otra vez la UPCH y la UNMSM, pero son también importantes el INS y el MINSA. Además de estos cuatro, dos actores más cumplen un papel importante de intermediación: la UPC y el Hospital Rebagliati (HREBA).

Figura 19: Árbol de decisión para la asignación de categorías institucionales

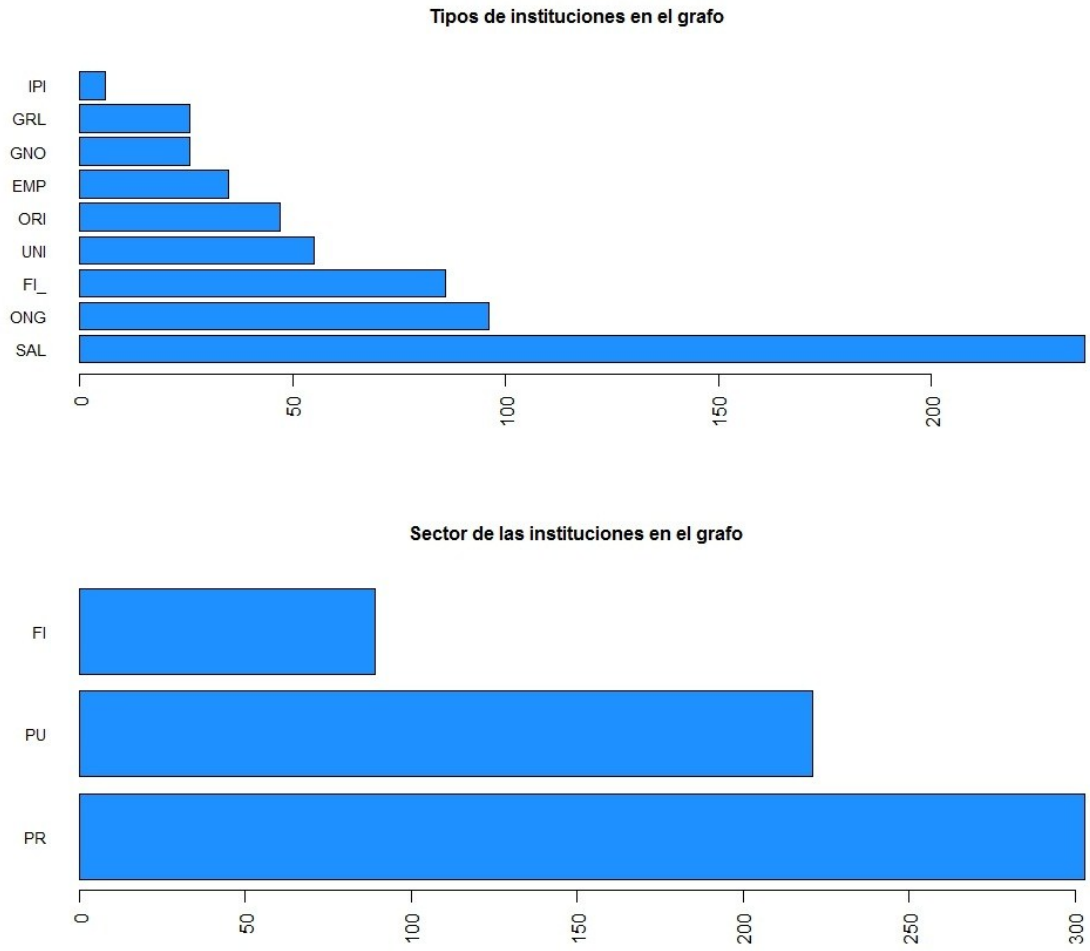


UI – Unidad de investigación

Fuente: Elaboración propia, adaptado de OECD (2015)

La medida de centralidad de vector propio mide la centralidad de un vértice en base a la centralidad de los vecinos. Se asume que mientras más importantes son los vecinos, más importante es el vértice. En esta medida, tal como en los casos anteriores, los vértices más centrales son la UPCH y la UNMSM. Le siguen un poco más de lejos PRISMA e INS. Finalmente, dos instituciones con una centralidad de vector propio alta, que no se ve reflejada en el mismo grado en otras medidas de centralidad, son el Hospital Nacional Cayetano Heredia (HNCH) y el Instituto Nacional de Ciencias Neurológicas (INCN).

Figura 20: Características de las instituciones involucradas en el grafo de coautorías



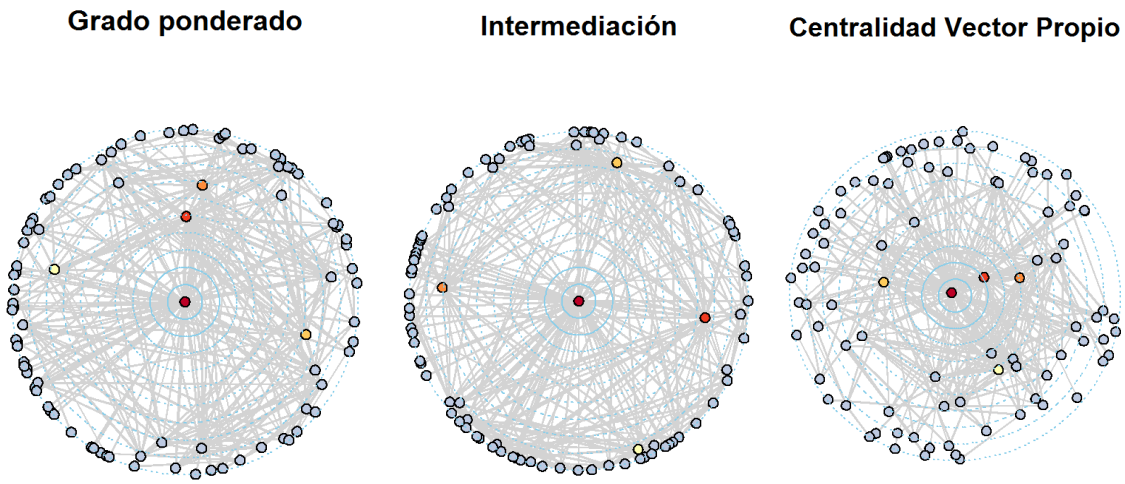
FUENTE: Elaboración propia

Cuadro 4: Vértices con mayor centralidad en el grafo

Vértice	*Grado*	*Grado ponderado*	*Cercanía*	*Intermediación*	*Centralidad de vector propio*
UPCH	252	1 853	0.42	48 439	1.00
UNMSM	193	1 021	0.41	36 430	0.68
INS	128	658	-	14 531	0.45
MINSA	108	457	-	14 781	-
UPC	80	-	-	9 892	-
NAMRU	83	522	-	-	-
PRISMA	-	406	-	-	0.47
UNJFS	-	-	0.40	-	-
HREBA	-	-	-	8 935	-
HNCH	-	-	-	-	0.42
INCN	-	-	-	-	0.42

FUENTE: Elaboración propia

Figura 21: Centralidades en el grafo de coautorías de instituciones peruanas con investigación en medicina en Scopus entre el 2000 y el 2015



FUENTE: Elaboración propia

4.4.2. LA REPRESENTACIÓN

Escoger la forma en la que se representan los elementos del grafo es una tarea compleja, por encima de unos cuantos vértices la representación gráfica de las interacciones puede dar resultados muy poco claros (ver fig. 22). Los algoritmos de distribución de los nodos más conocidos son:

(a) Circular: Distribuye los vértices de manera equidistante en un círculo. Utilizado con frecuencia para topologías de la red de comunicaciones. En este caso la representación circular es inadecuada para el grafo, no permite reconocer centralidades ni subgrafos densamente conectados.

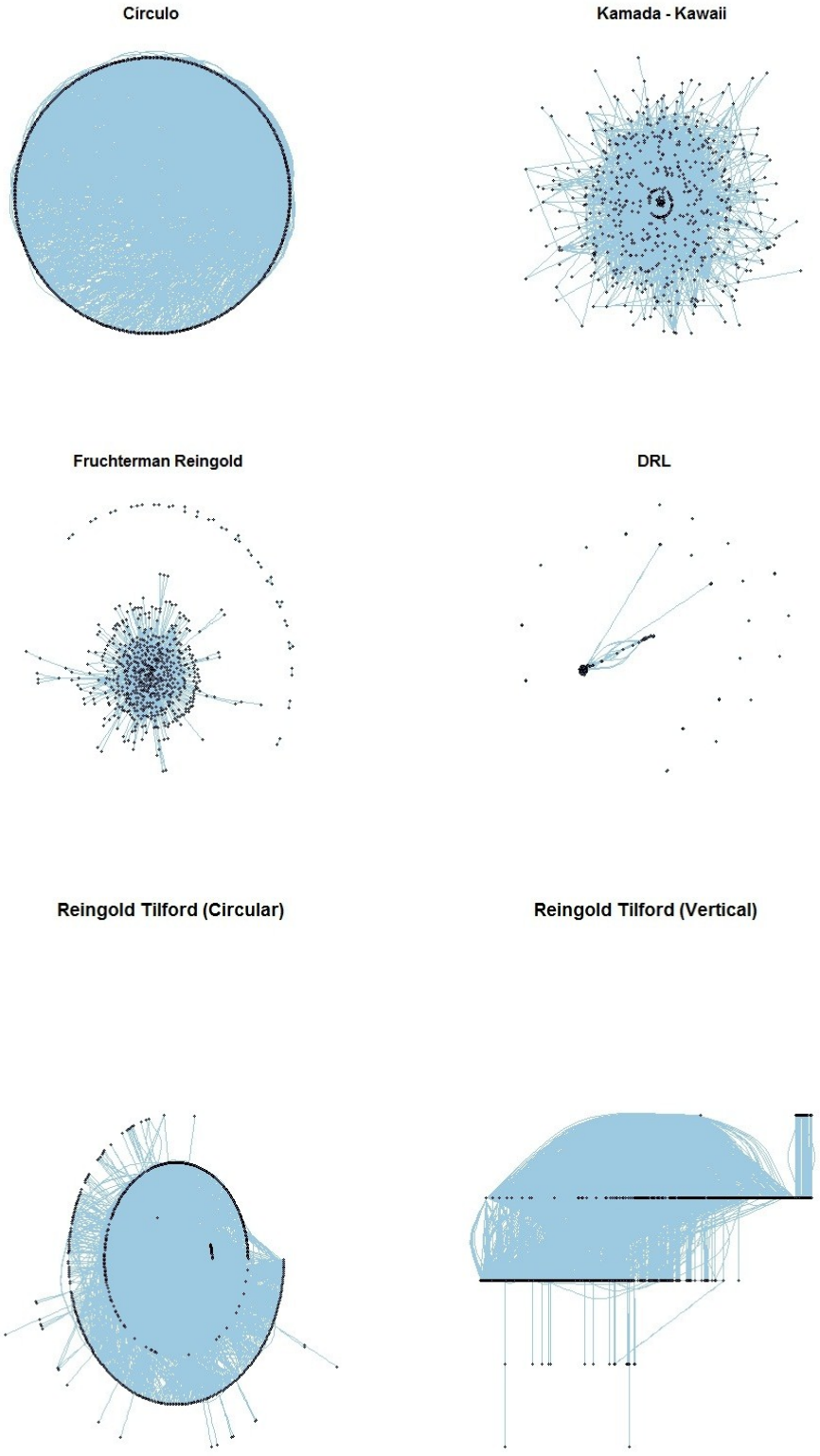
(b) Métodos dirigidos por fuerza: Estos métodos modelan el grafo como si fuese un sistema físico que está en equilibrio. Los elementos que componen estos métodos son el modelo físico y el algoritmo para buscar el equilibrio. Los nodos que son adyacentes se atraen y los que no lo son se repelen. Siguen esta lógica los métodos: (1) Kamada y Kawai, (2) Fruchterman Reingold y (3) DRL. DRL está optimizado para grafos muy grandes, pero podría no ser adecuado en este caso. Fruchterman Reingold pueden tener un desempeño pobre en el caso de grafos con muchos vértices, en este caso, al tratarse de un grafo intermedio, eso no es materia de preocupación (Brandenburg et al. 1995, Gibson et al. 2013, Jacomy et al. 2014).

Además de estos algoritmos, se tiene también distribuciones adecuadas para representar jerarquías, pero estas son más convenientes para grafos en forma de árbol y como puede verse para este análisis los resultados no son comprensibles a nivel visual.

El grafo resultante de los datos de origen no es un grafo simple pues las coautorías que se repiten constituyen multiaristas. El grafo se puede «simplificar» sumando las multiaristas y asignando ese atributo a las aristas como ponderaciones. El Cuadro 5 resume las características del grafo previas y posteriores a la transformación. La ponderación de las aristas es desigual, la gran mayoría sólo aparece una vez porque la coautoría entre los vértices se dio una sola vez (ver fig. 23).

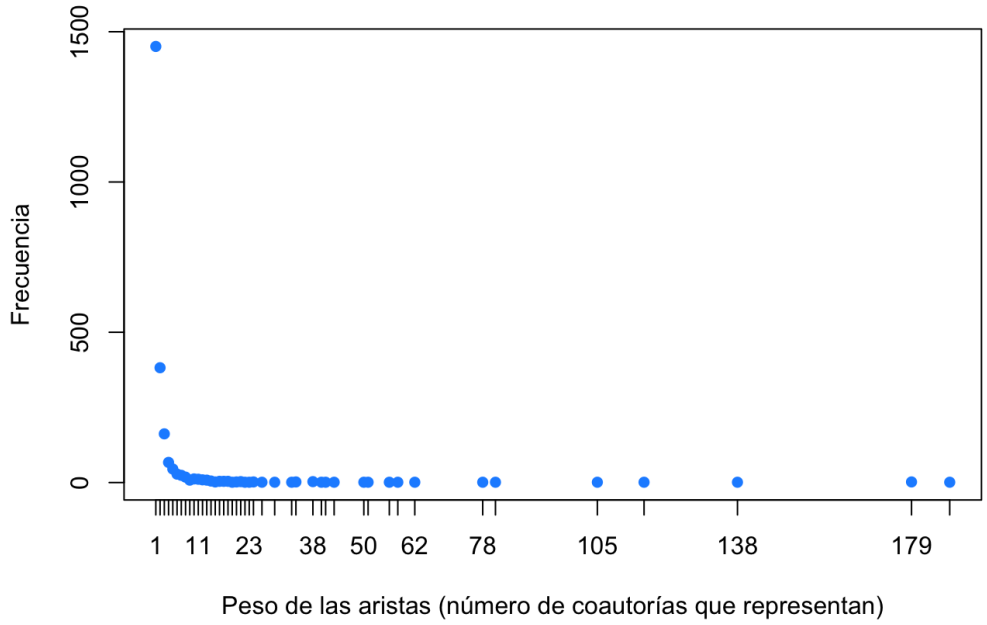
Además, podemos ver que el grafo no es conexo, es decir existen satélites - partes del grafo que no están conectadas al componente principal (gigante) a través de aristas. Los satélites se eliminan del grafo para una mejor visualización y para facilitar el análisis, los conglomerados formados por las instituciones se analizarán en base al gran componente es decir el grafo conexo en el que participan 613 vértices.

Figura 22: Representaciones del grafo de coautorías de instituciones peruanas con investigación en medicina en Scopus entre el 2000 y el 2015 utilizando diferentes algoritmos de distribución de los vértices.



Fuente: Elaboración propia

Figura 23: Distribución de la ponderación de las aristas en el grafo de coautorías de instituciones peruanas con investigación en medicina indizada en Scopus (2000-2015).



FUENTE: Elaboración propia

Cuadro 5: Características del grafo de coautorías.

Característica	**Multigrafo**	**Grafo ponderado**
Vértices	613	
Aristas	6453 (coautorías)	2278
Pesos de las aristas	No	Max 188, mín. 0
Grafo conexo	No	
Número de componentes conexos	21	

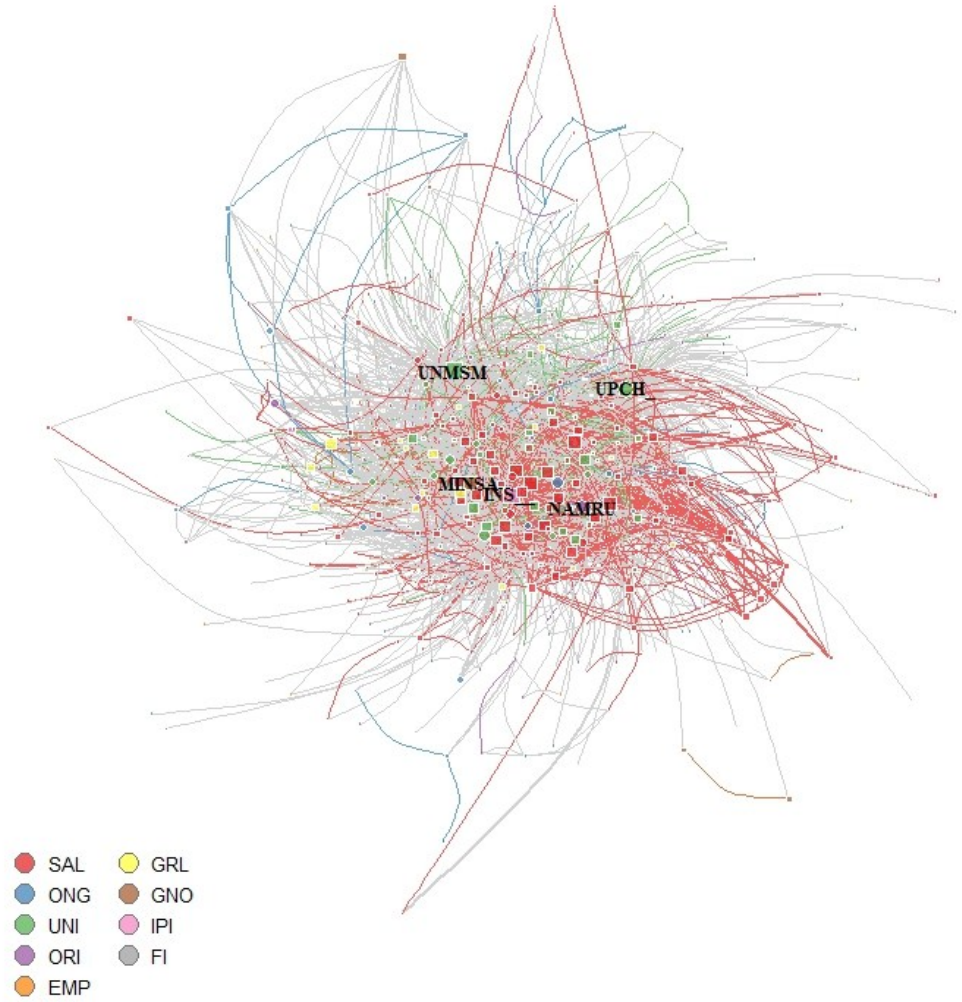
Continuación

Característica	*Multigrafo*	*Grafo simple (ponderado)*
Tamaño de los componentes	568, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2, 2, 3, 3, 2, 3, 2, 2	
Diámetro	8	
Atributos de los vértices	(1) Nombre ($*name)*; (2) Sector ($*sec*): privado (*PR),* público (*PU*). (3) Tipo ($*tip*): universidades (*UNI),* instituciones de atención en salud (*SAL*), organizaciones internacionales (*ORI*), empresas (*EMP),* administración pública y gobierno (*GNO y GRL),* institutos públicos de investigación (*IPI),* instituciones privadas sin fines de lucro que no se dedican a la atención en salud (*ONG*), información insuficiente (*FI*).	

FUENTE: Elaboración propia

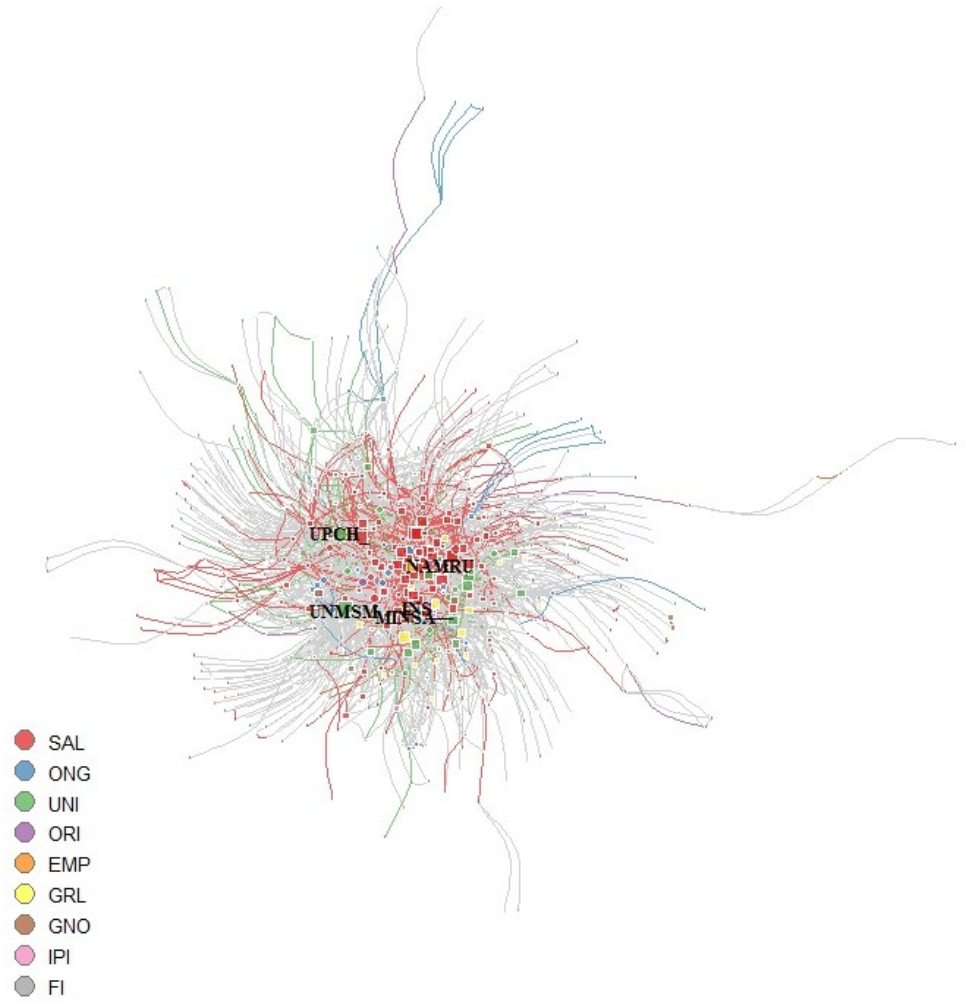
En las figuras a continuación se presenta el grafo mostrando la tipología institucional (por color), con diferentes distribuciones de los vértices. Queda claro que para el tipo de datos con los que se cuenta la distribución de los nodos con el algoritmo de Kamada y Kawai es el más adecuado, (para la comparación ver figuras 24 a 27). Aunque R ofrece muchas opciones para generar grafos, la distribución de los vértices no es su mejor atributo. Finalmente, como puede verse en la figura 28, una manera de hacer más comprensible la información que el grafo debe exponer es reducir los elementos que lo componen. En este caso se ha agrupado los vértices del mismo tipo en un solo vértice y las aristas entre estos, en una sola arista. Las dimensiones de cada uno corresponden a la frecuencia en el grafo (para los vértices) y a la ponderación acumulada (para las aristas).

Figura 24: Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución Kamada y Kawai



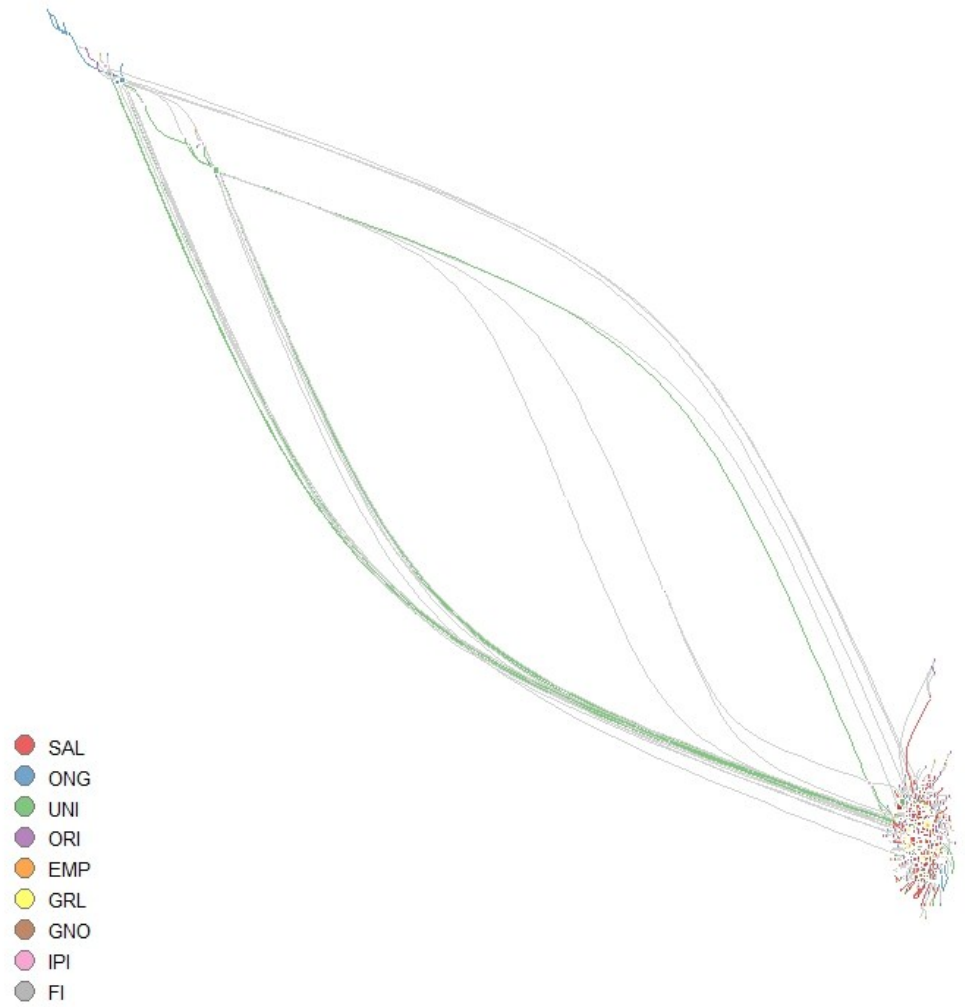
FUENTE: Elaboración propia con igraph.

Figura 25: Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución Fruchterman Reingold



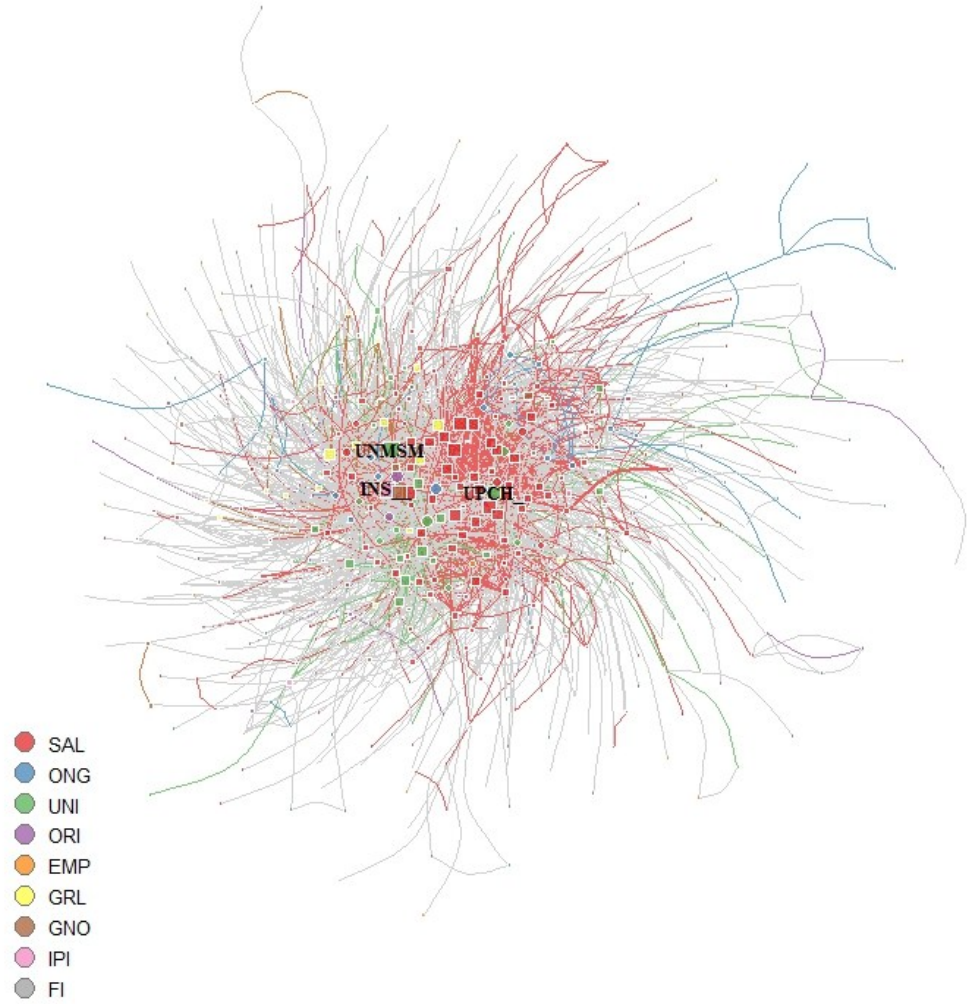
FUENTE: Elaboración propia con igraph.

Figura 26: Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución con DRL



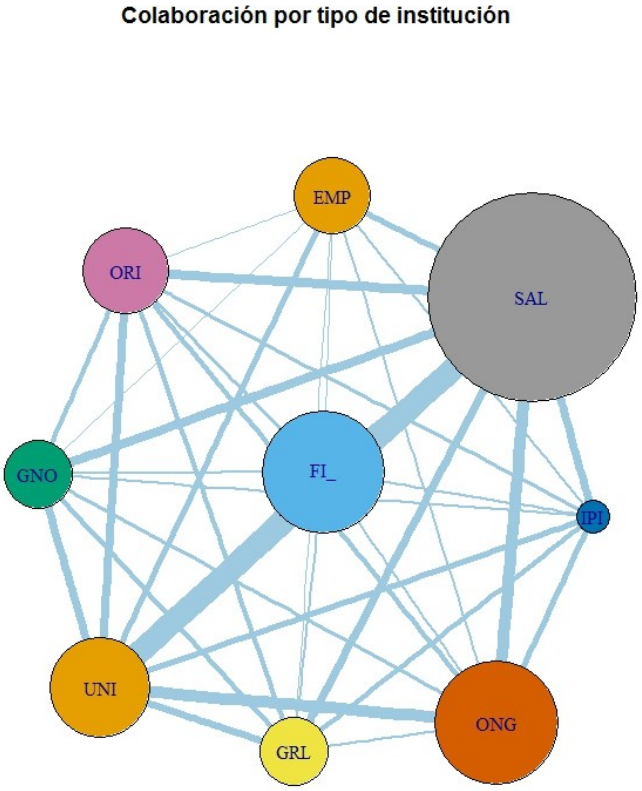
FUENTE: Elaboración propia con igraph

Figura 27: Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución con Large Graph Layout



FUENTE: Elaboración propia con igraph.

Figura 28: Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Tipos de instituciones.

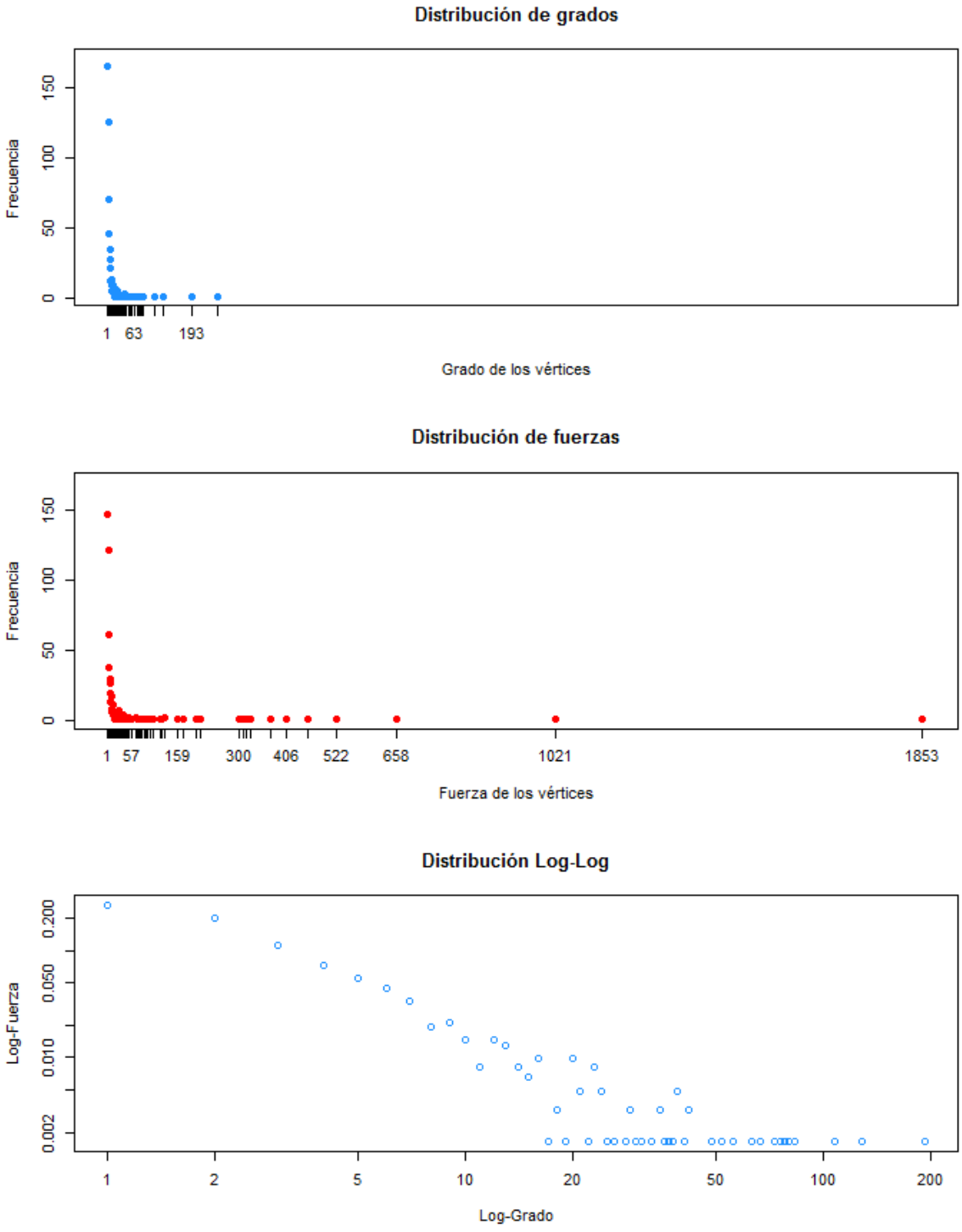


FUENTE: Elaboración propia con igraph.

4.4.3. CARACTERÍSTICAS DEL GRAFO

En el caso de los grafos ponderados, como el que estamos analizando, es interesante ver la distribución de la fuerza de los vértices que componen el grafo. Se entiende por fuerza de los vértices al grado ponderado de los vértices. Como puede verse en la fig. 29 la distribución de las fuerzas es altamente sesgada a la izquierda y el gráfico log-log se acerca a una recta, lo cual indica la posibilidad de una distribución *power law* (ley potencial). La característica de un grafo con una distribución ley potencial es que unos pocos vértices pueden ejercer gran influencia sobre el resto.

Figura 29: Distribución del grado y fuerza del grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015



FUENTE: Elaboración propia

Para analizar la cohesión de un grafo puede tomarse en cuenta los cliques, es decir los conjuntos de vértices tales que para todo par de vértices existe una arista que los conecta. En este caso los cliques máximos son seis, están compuestos por 13 vértices y constituyen conjuntos de colaboración particularmente densos (ver Cuadro 6). Como puede verse, entre estos cliques hay variedad de instituciones, pero las especializadas en atención de salud (principalmente hospitales) forman un conjunto de colaboraciones particularmente denso.

El grafo es poco denso, sólo existe el 1 por ciento de las aristas que podrían existir si fuese un grafo completo (con todas las aristas posibles). La transitividad o coeficiente de clustering es la medida en que dos pares de vértices tienden a tener una conexión cuando tienen un vecino en común. En este grafo la transitividad es: 0.17. Como comparación, en grafo al azar con características similares al analizados la transitividad estaría en un rango entre 0 y 0.12 La transitividad puede medirse también para los subgrafos, es decir para los grafos centrados en un vértice en particular. En este caso, la medida compararía la tendencia de los colaboradores del actor analizado a colaborar entre ellos. Para los nodos más importantes de esta red el coeficiente de clustering es muy bajo: UPCH – 0.04, UNMSM – 0.05, INS – 0.09, NAMRU – 0.15, MINSA- 0.1. En el caso de la ONG Prisma que tiene un grado ponderado de más de 400, el nivel transitividad es alto, por encima del nivel del grafo (0.26).

Las medidas de conectividad de la red permiten ver si el grafo se aproxima por sus características a un *pequeño mundo,* un grafo donde la distancia media es pequeña, pero el coeficiente clustering es relativamente alto. En este caso la distancia media es 2.76 y el diámetro es igual a 11, por lo que puede concluirse que el grafo cumple con las características de *pequeño mundo*.

Además, como puede verse en la figura 30, en este grafo los vértices con grado alto tienden a asociarse (ser vecinos) de vértices con grado bajo. El coeficiente de asortatividad por grado es de -0.26 lo cual significa que exista una correlación negativa para las relaciones de vértices con grado alto con sus pares.

4.5. IDENTIFICACIÓN DE LOS CONGLOMERADOS

Se utilizó el algoritmo fast.greedy.community del paquete R que identifica los conglomerados de manera aglomerativa, utilizando el algoritmo de Clauset (Clauset et al. 2004), tomando la modularidad como medida de costo. Se identificaron dieciséis conglomerados, tres con más de sesenta vértices, cuatro con más de veinte vértices, y nueve con menos de veinte vértices (ver figura 31).

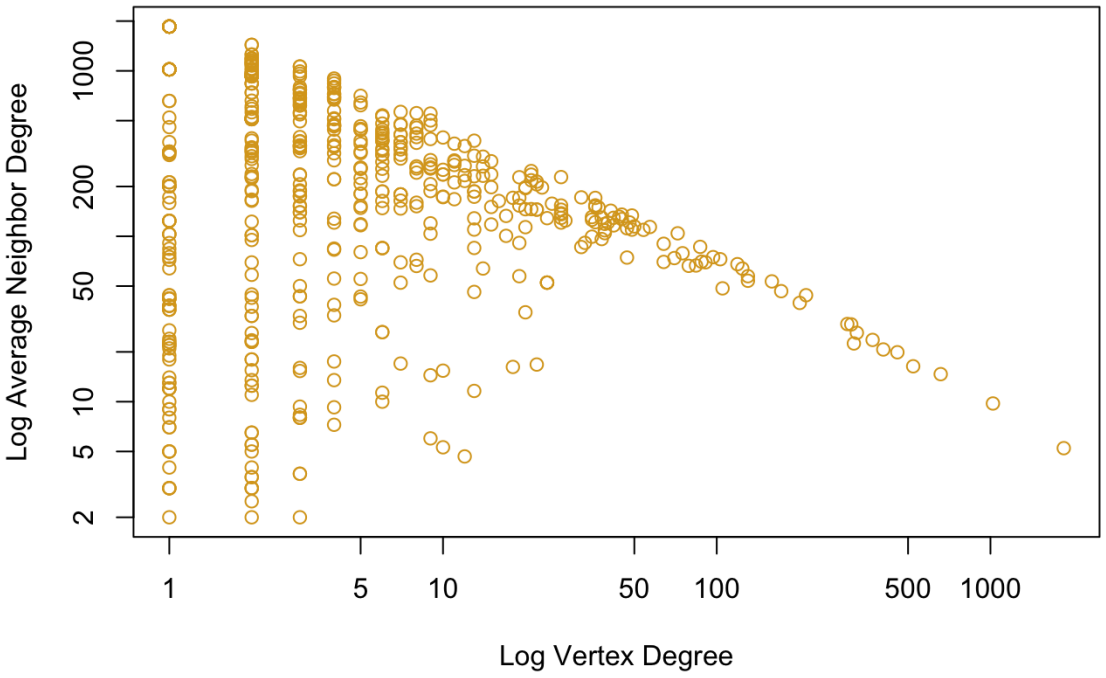
Cuadro 6: Cliques máximos en el grafo.

Tipo	*Vértice*	*Clique 1*	*Clique 2*	*Clique 3*	*Clique 4*	*Clique 5*	*Clique 6*
UNI	UPCH	X	X	X	X	X	
	UNMSM	X	X	X	X		
	UPC	X	X	X	X		
ORI	NAMRU	X	X	X	X		
ONG	PRISM	X	X	X	X		
IPI	INS__	X	X	X	X		
SAL	HNCH_	X	X	X	X	X	X
	HNDAC	X	X	X	X	X	
	INSN_	X	X	X	X	X	
	HREBA	X	X	X	X	X	
	HDOSM	X	X	X	X		X
	HLOAY	X	X	X			X
	HALME		X				
	HEPP			X		X	
	CSURU					X	
	CSWAN					X	
	HNDSB					X	
	HRCUS					X	
	HRGDV					X	
	HYANA					X	
	PMARE					X	
	CSANA						X
	H4ESH						X
	HASUL						X
	HCCHI						X
	HMOLI						X
	HNCAS						X
	HNHU						X
	HNSEB						X
	HRCH_						X
	HRDTR						X
GNO	MINSA				X		
GRL	DRSLI	X			X		

FUENTE: Elaboración propia

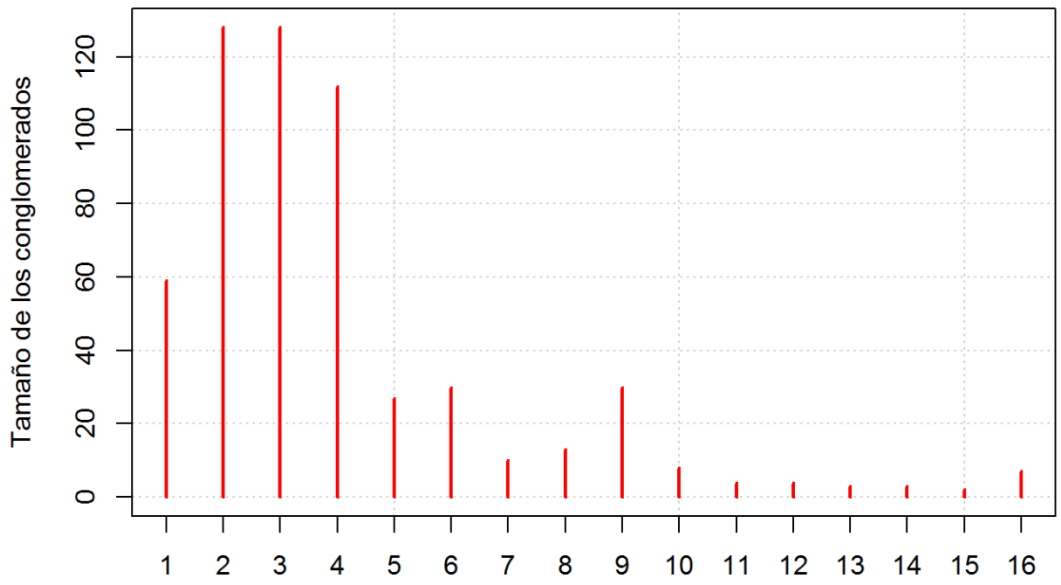
El valor de modularidad para el grafo es 0.28. Esto significa que la estructura del grafo se acerca más a una estructura comunitaria que un grafo al azar. El rango de valores posibles es de -0.5 a 1, por lo que 0.28 no es un valor alto. La robustez de la estructura comunitaria se puede visualizar en la figura 33. En este gráfico se muestran los resultados de mil simulaciones del número de comunidades identificadas cuando las aristas se distribuyen al azar, pero las características del grafo permanecen iguales a las del grafo estudiado. En un caso se simula con un grafo que tiene el mismo tamaño (número de vértices y número de aristas) y en el otro con uno que tiene la misma distribución de grados. Como puede verse, la mayoría de las veces el número de comunidades identificado en los grafos al azar es inferior al número de comunidades en el grafo estudiado, por lo que puede concluirse que el grafo estudiado no tiene una estructura fuertemente comunitaria.

Figura 30: Promedio del grado de los vecinos versus grado de los vértices (escala logarítmica) para las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015



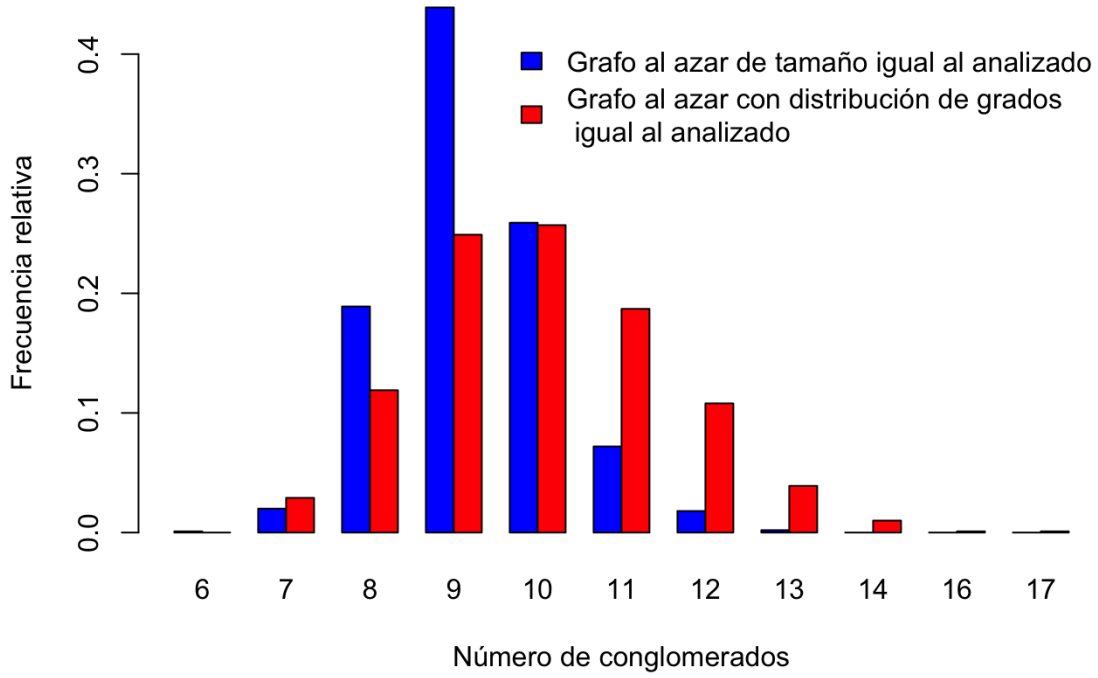
FUENTE Elaboración propia

Figura 31: Número de elementos que conforman los conglomerados hallados en el grafo.



FUENTE: Elaboración propia

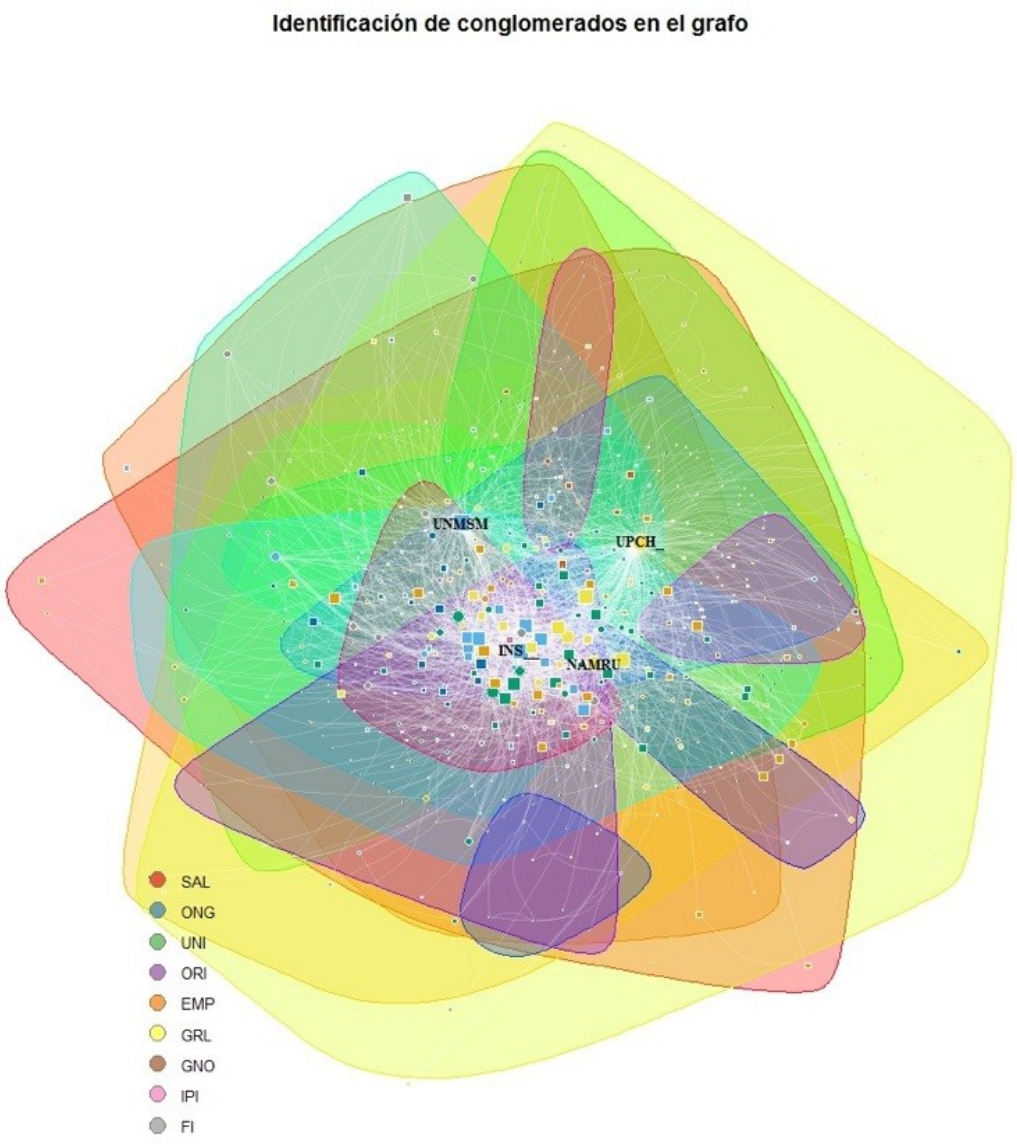
Figura 32: Simulación de las frecuencias relativas del número de comunidades identificadas en el grafo al azar.



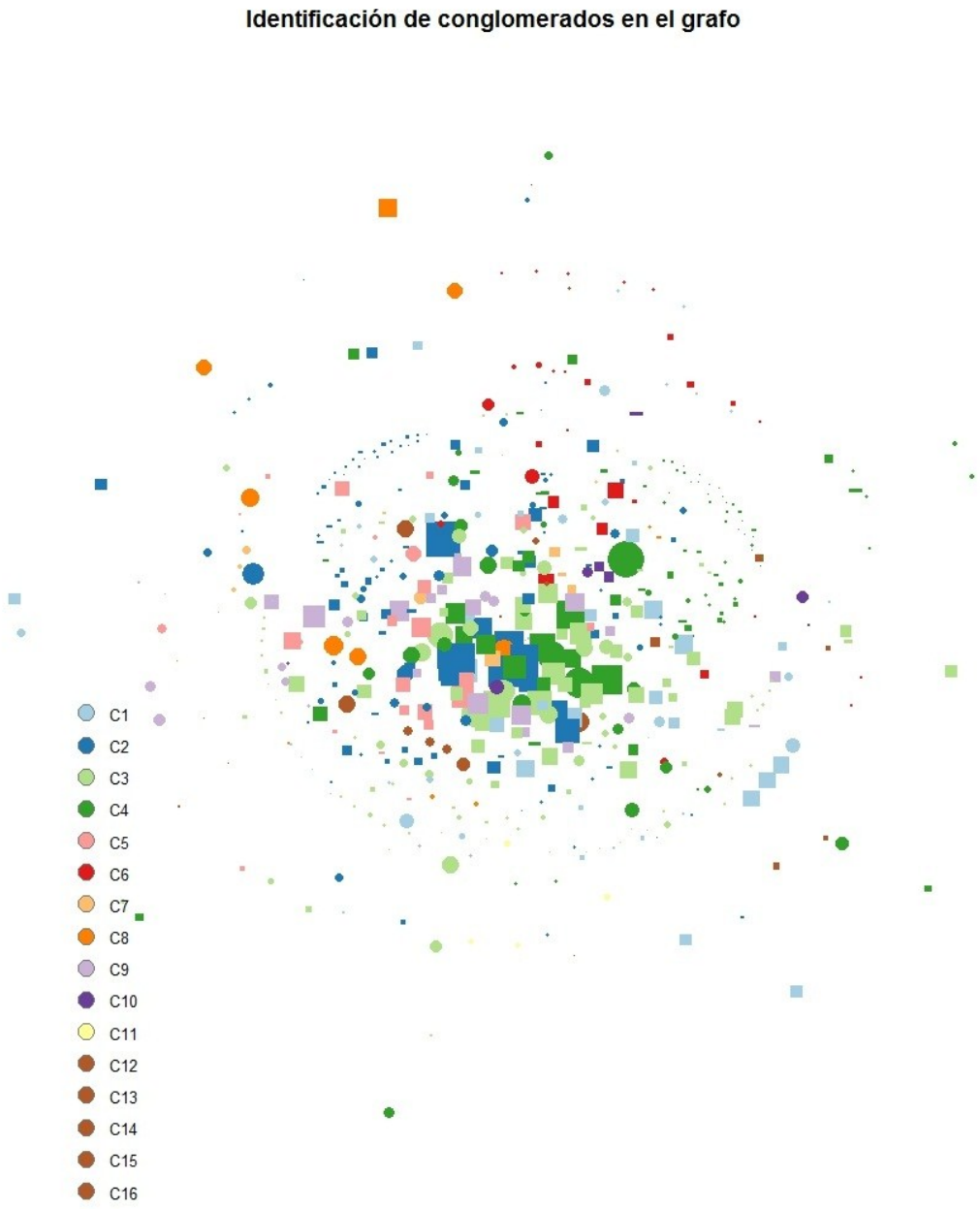
FUENTE: Elaboración propia

Como puede verse en las figuras 33 y 34, diferenciar los elementos de un conglomerado frente a los vértices pertenecientes a los otros no es fácil en el gráfico, tanto por el número de vértices en el grafo, como por el número de comunidades identificadas. Si se resume cada comunidad identificada como un solo vértice la visualización de las comunidades y de las colaboraciones entre estas se hace más clara.

Figura 33: Conglomerados en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015

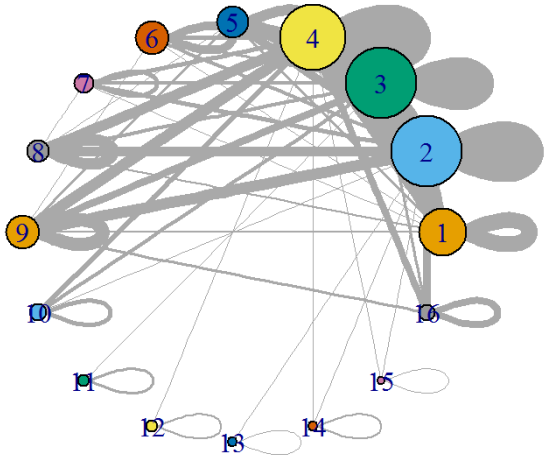


Continuación



FUENTE: Elaboración propia

Figura 34: Colaboración entre conglomerados en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.



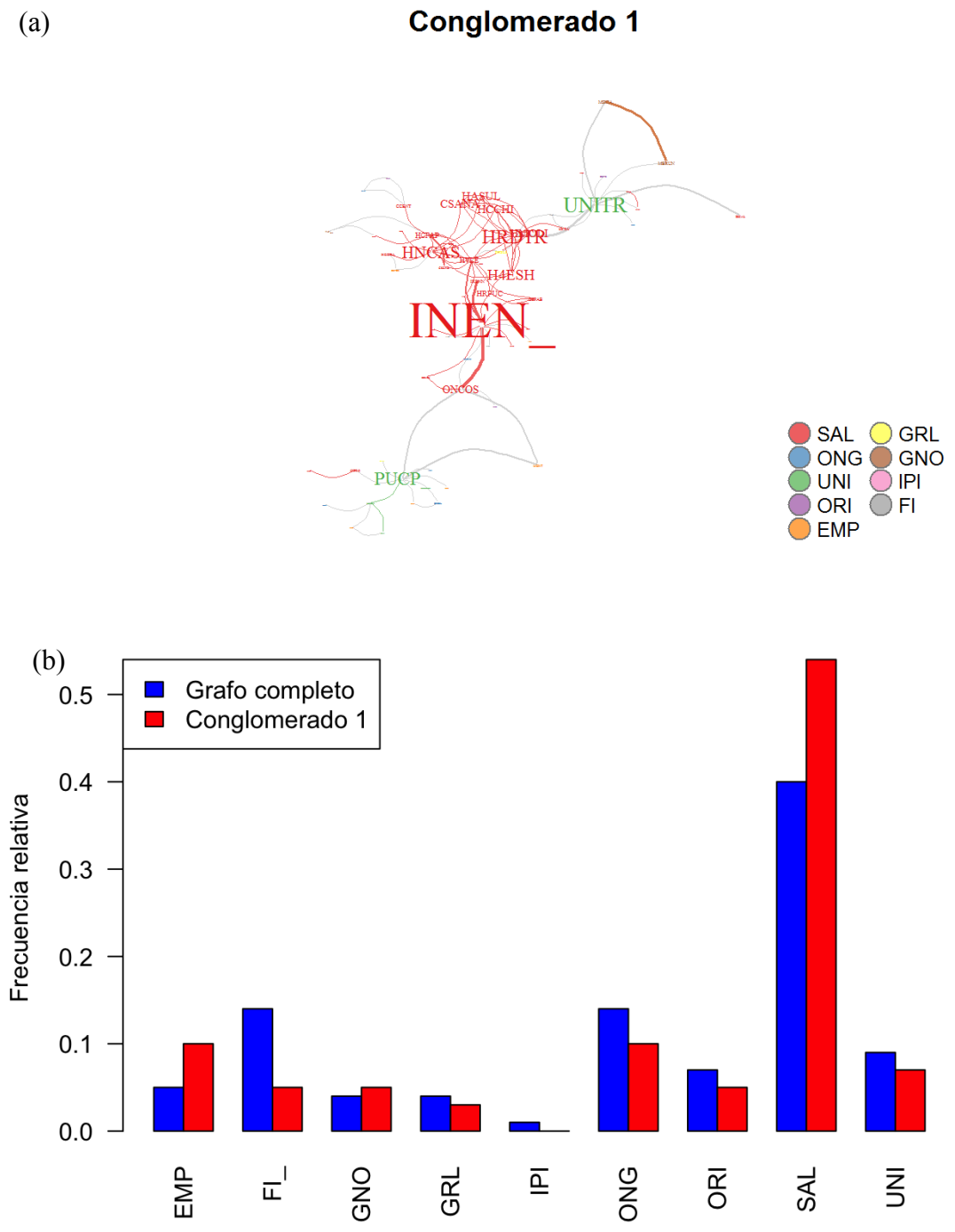
Este gráfico agrupa los conglomerados identificados – cada nodo corresponde a un conglomerado y su tamaño corresponde al tamaño del conglomerado. Los loops representan las coautorías que se dan dentro del mismo conglomerado.

FUENTE: Elaboración propia

4.5.1. CONGLOMERADO 1: INSTITUTO NACIONAL DE ENFERMEDADES NEOPLÁSICAS

El conglomerado uno está liderado por el Instituto Nacional de Enfermedades Neoplásicas (INEN). Las instituciones en este conglomerado que están en el primer decil en las tres medidas de centralidad (grado ponderado, intermediación, centralidad del vector propio) son INEN, Universidad Nacional de Trujillo (UNITRU), Pontificia Universidad Católica del Perú (PUCP), Hospital Carlos Alberto Seguín – Arequipa (HNCAS). Está compuesto por 59 instituciones, más de la mitad son instituciones especializadas en la atención en salud. En el grafo completo la proporción de instituciones del sector salud es de 40 por ciento, en este conglomerado, alcanza el 50 por ciento y es significativamente superior a la proporción en el grafo completo (ver figura 35). El detalle de las instituciones involucradas y sus medidas de centralidad puede verse en los anexos (Cuadro 8).

Figura 35: Conglomerado 1 (INEN) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.



FUENTE: Elaboración propia

(a) En el grafo el tamaño de las etiquetas es proporcional al grado ponderado de los vértices. (b) Comparación de la proporción de tipos de instituciones en el grafo versus proporción de instituciones en el conglomerado.

4.5.2. CONGLOMERADO 2: UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

El conglomerado dos está liderado por la Universidad Nacional Mayor de San Marcos (UNMSM). Las instituciones en este conglomerado que están en el primer cinco por ciento en las tres medidas de centralidad (grado ponderado, intermediación, centralidad del vector propio) son UNMSM, Instituto Nacional de Salud (INS), Ministerio de Salud (MINSA), y el Instituto Nacional de Salud del Niño (INSN). Está compuesto por 128 instituciones, casi el 40 por ciento son instituciones especializadas en la atención en salud, pero esta proporción no es diferente a la proporción de este tipo de instituciones en el grafo completo (ver figura 36). El detalle de las instituciones involucradas y sus medidas de centralidad puede verse en los anexos (Cuadro 9).

4.5.3. CONGLOMERADO 3: UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS

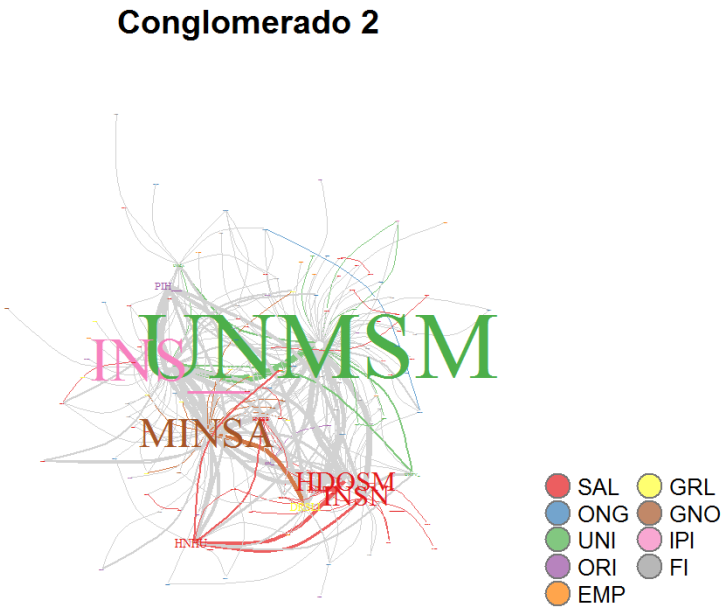
El conglomerado tres está liderado por la Universidad Peruana de Ciencias Aplicadas (UPC). Las instituciones en este conglomerado que están en el primer cinco por ciento en las tres medidas de centralidad (grado ponderado, intermediación, centralidad del vector propio) son UPC, Hospital Rebagliati (HREBA), Universidad San Martin de Porras (USMP), y el Hospital Almenara (HALME). Está compuesto por 128 instituciones, casi el 45 por ciento son instituciones especializadas en la atención de salud, es decir esta proporción es ligeramente superior a la proporción de este tipo de instituciones en el grafo completo (ver figura 37). El detalle de las instituciones involucradas y sus medidas de centralidad puede verse en los anexos (Cuadro 10).

4.5.4. CONGLOMERADO 4: UNIVERSIDAD PERUANA CAYETANO HEREDIA

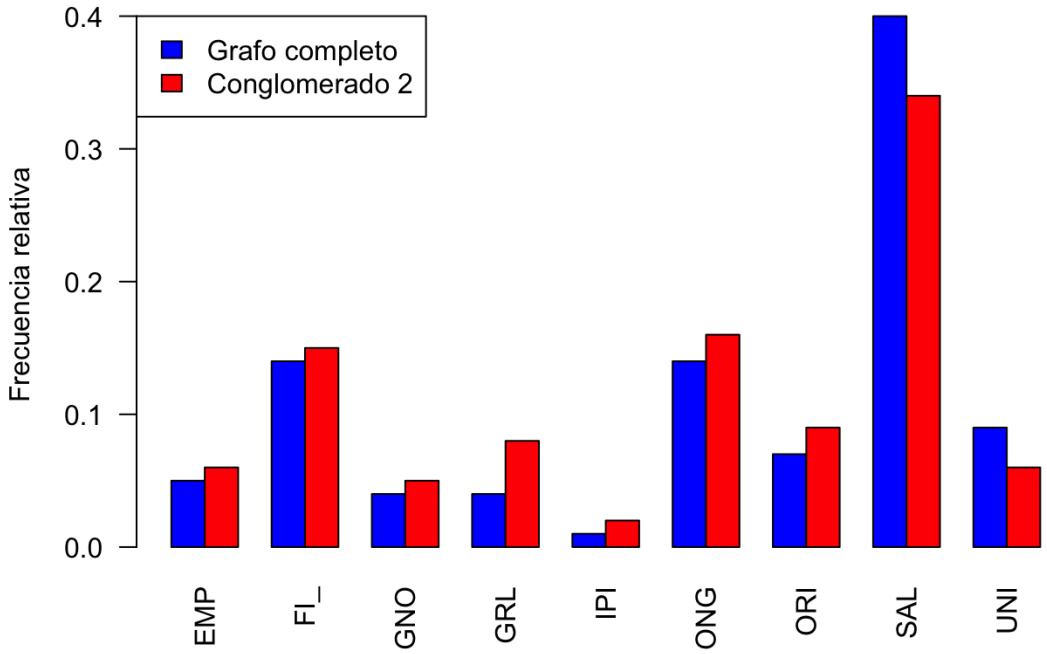
El conglomerado cuatro está liderado por la Universidad Peruana Cayetano Heredia (UPCH). Las instituciones en este conglomerado que están en el primer cinco por ciento en las tres medidas de centralidad (grado ponderado, intermediación, centralidad del vector propio) son UPCH, Naval Medical Research Unit Six (NAMRU), Asociación Benéfica Prisma (PRISMA), Hospital Nacional Cayetano Heredia (HNCH) y el Hospital Loayza (HLOAY). Está compuesto por 112 instituciones, casi el 28 por ciento son instituciones sobre las que falta información (firmas individuales), es decir esta proporción es significativamente superior a la proporción de este tipo de instituciones en el grafo completo (ver figura 38). El detalle de las instituciones involucradas y sus medidas de centralidad puede verse en los anexos (Cuadro 11).

Figura 36: Conglomerado 2 (UNMSM) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.

(a)



(b)

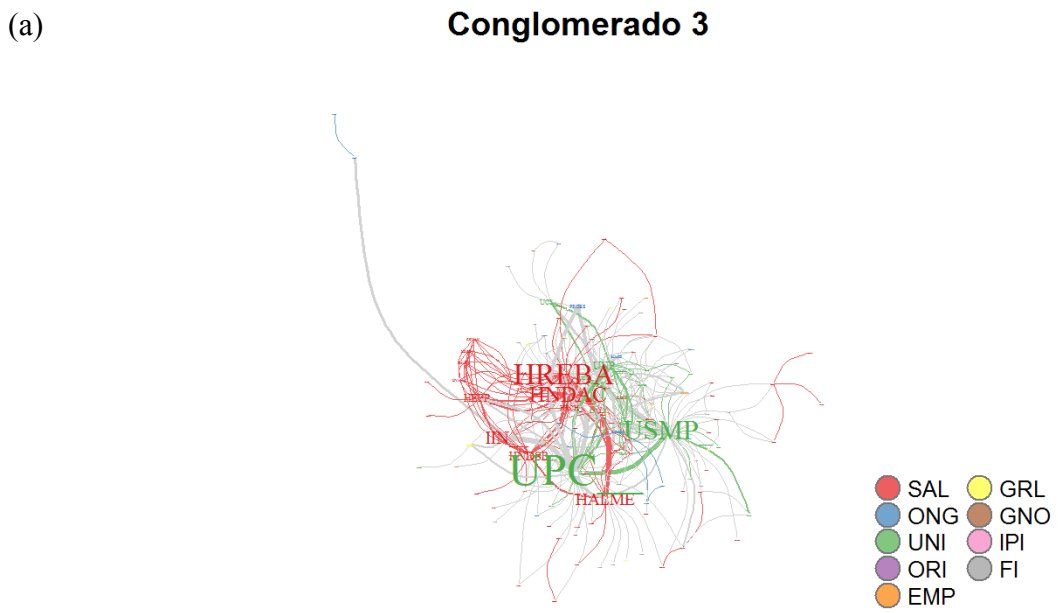


FUENTE: Elaboración propia

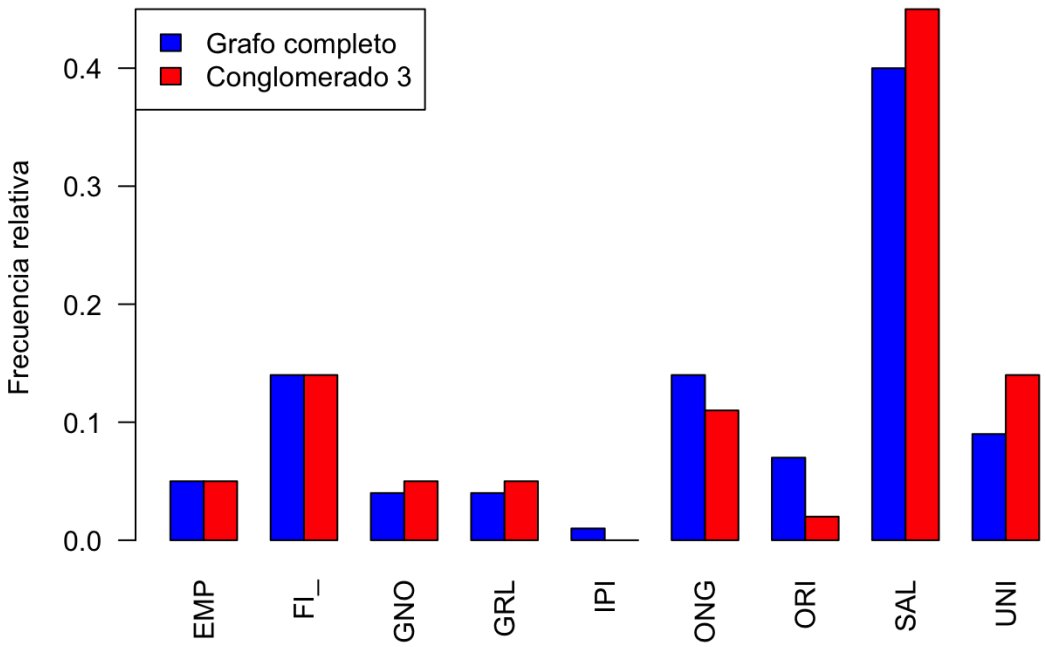
(a) En el grafo el tamaño de las etiquetas es proporcional al grado ponderado de los vértices. (b) Comparación de la proporción de tipos de instituciones en el grafo versus proporción de instituciones en el conglomerado.

Figura 37: Conglomerado 3 (UPC) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.

(a)



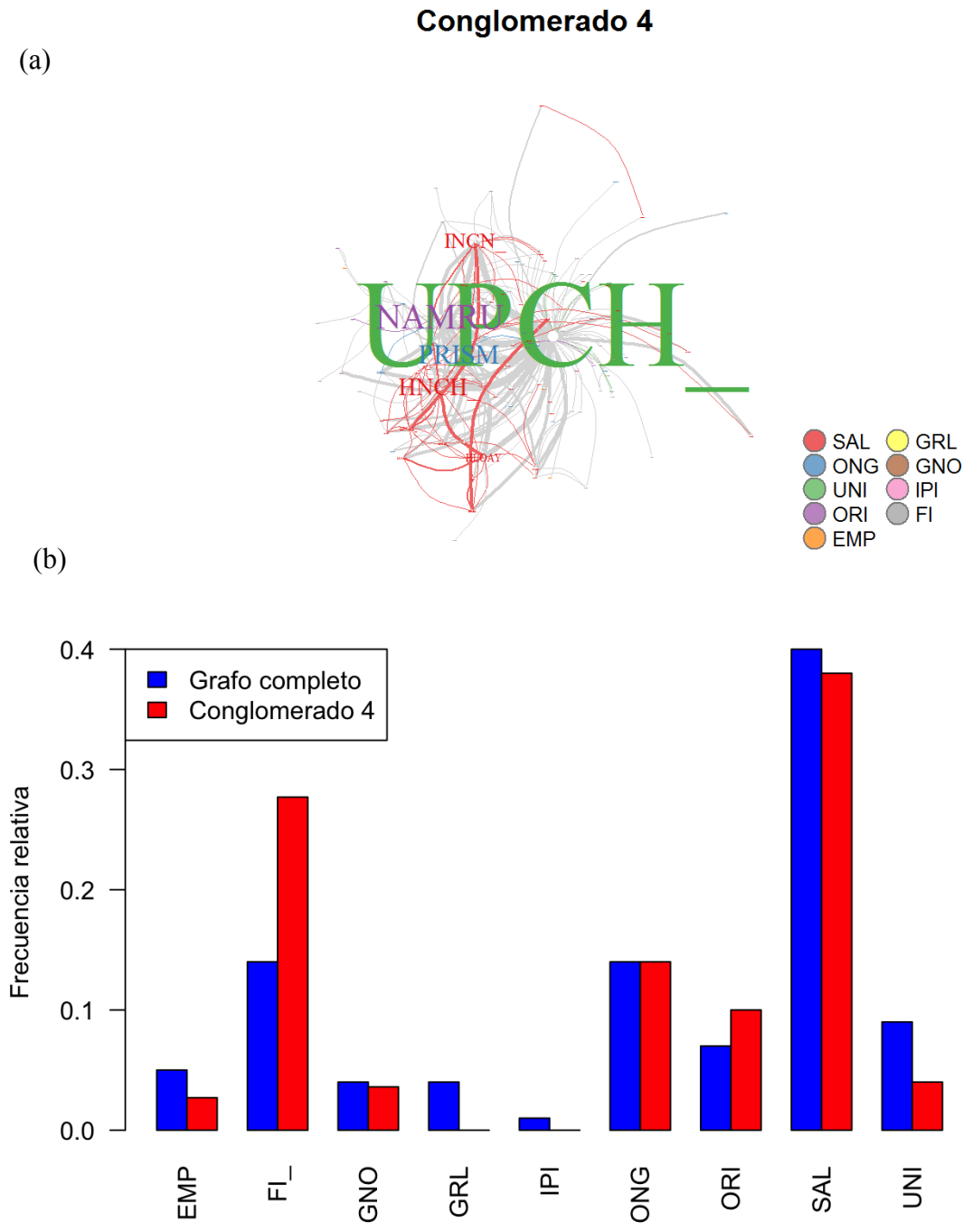
(b)



FUENTE: Elaboración propia

(a) En el grafo el tamaño de las etiquetas es proporcional al grado ponderado de los vértices. (b) Comparación de la proporción de tipos de instituciones en el grafo versus proporción de instituciones en el conglomerado.

Figura 38: Conglomerado 4 (UPCH) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.



FUENTE: Elaboración propia

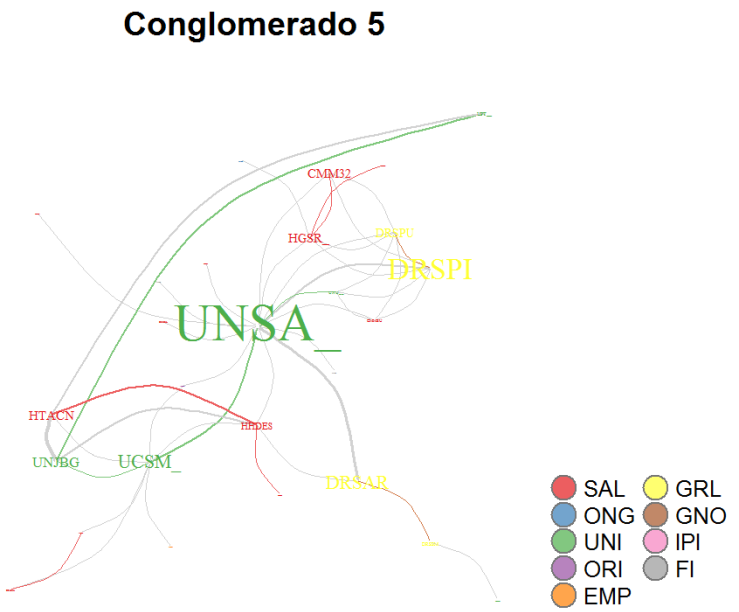
(a) En el grafo el tamaño de las etiquetas es proporcional al grado ponderado de los vértices. (b) Comparación de la proporción de tipos de instituciones en el grafo versus proporción de instituciones en el conglomerado.

Los conglomerados cinco, seis y nueve son más pequeños que los anteriores: su número de vértices va de veintisiete a treinta.

4.5.5. CONGLOMERADO 5: UNIVERSIDAD NACIONAL DE SAN AGUSTÍN

El conglomerado cinco está liderado por la Universidad Nacional de San Agustín (UNSA). Las instituciones en este conglomerado que están en el primer diez por ciento en las tres medidas de centralidad (grado ponderado, intermediación, centralidad del vector propio) son la UNSA y la Dirección Regional de Salud de Piura (DRSPI). Está compuesto por 27 instituciones, más del 70 por ciento son instituciones pertenecientes al sector público, y más del 20 por ciento son universidades, es decir esta proporción es significativamente superior a la proporción de este tipo de instituciones en el grafo completo. Además, casi el 45 por ciento son instituciones especializadas en la atención de salud, es decir esta proporción es ligeramente superior a la proporción de este tipo de instituciones en el grafo completo (ver figura 39 y 40). El detalle de las instituciones involucradas y sus medidas de centralidad puede verse en los anexos (Cuadro 12).

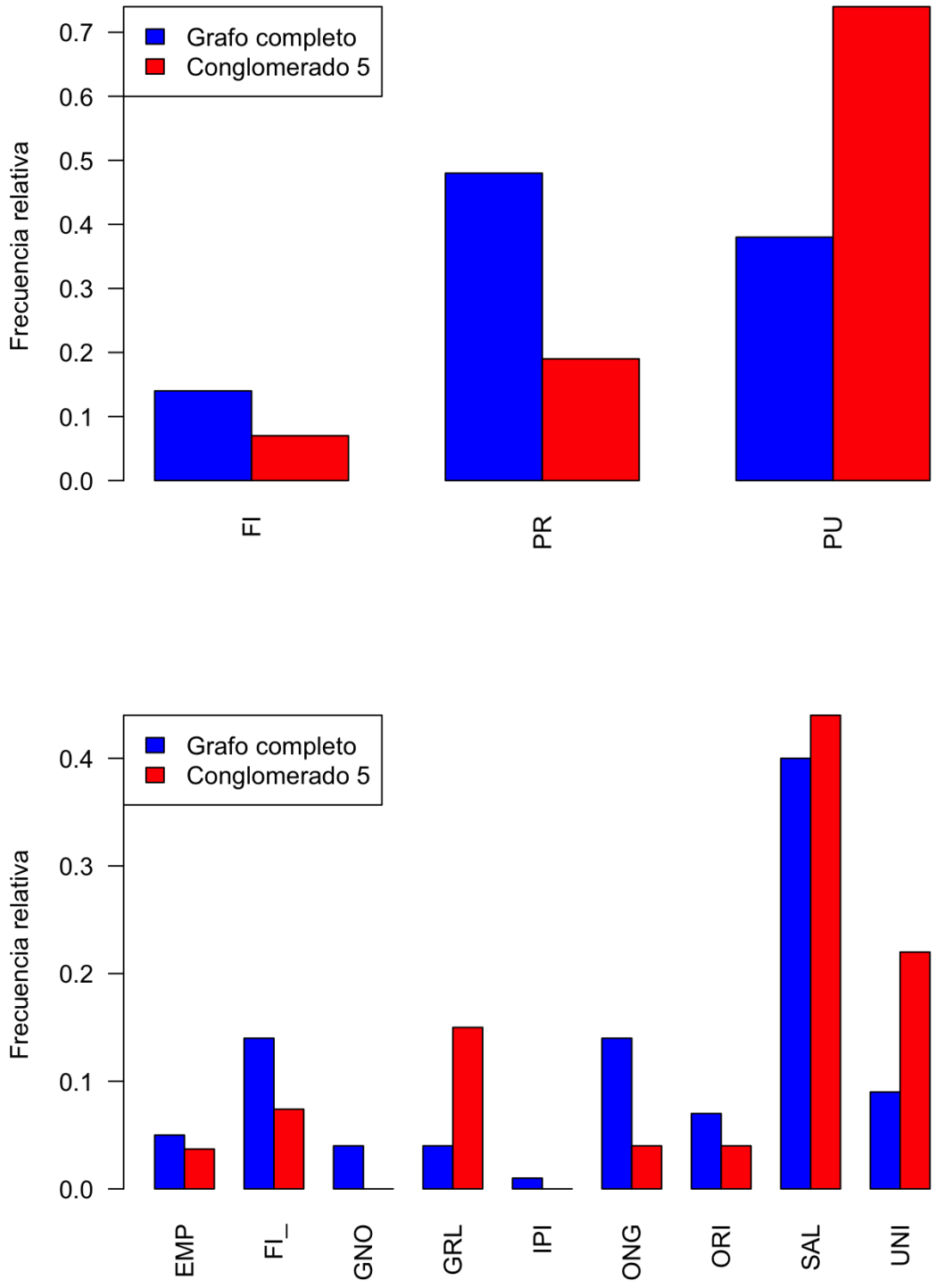
Figura 39: Conglomerado 5 (UNSA) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.



FUENTE: Elaboración propia

En el grafo el tamaño de las etiquetas es proporcional al grado ponderado de los vértices.

Figura 40: Proporción de tipos de instituciones en el conglomerado 5 (UNSA) versus en el grafo completo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.



FUENTE: Elaboración propia

4.5.6. CONGLOMERADO 6: CENTRO INTERNACIONAL DE LA PAPA

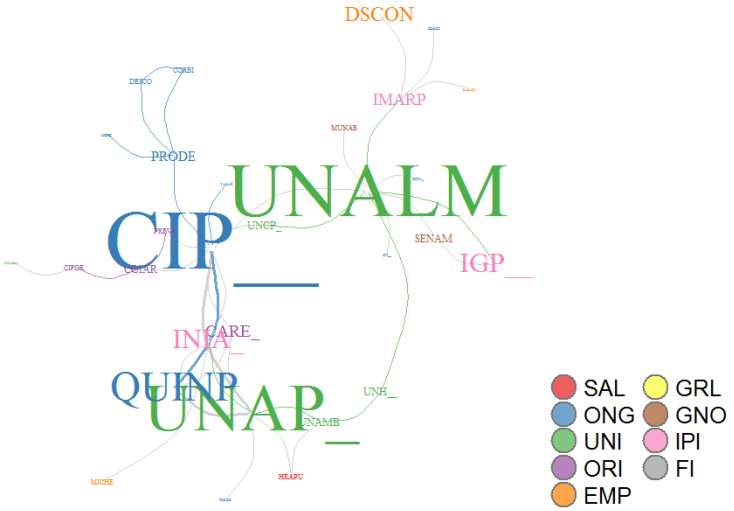
El conglomerado seis está liderado por el Centro Internacional de la Papa (CIP). Las instituciones en este conglomerado que están en el primer diez por ciento en las tres medidas de centralidad (grado ponderado, intermediación, centralidad del vector propio) son el CIP, la Universidad Nacional Agraria la Molina (UNALM), y la Universidad Nacional del Altiplano. Está compuesto por 30 instituciones, el 60 por ciento son instituciones pertenecientes al sector privado, más del 35 por ciento son instituciones privadas sin fines de lucro, y más del 20 por ciento son universidades, es decir esta proporción es significativamente superior a la proporción de este tipo de instituciones en el grafo completo. Además, menos del 5 por ciento son instituciones especializadas en la atención de salud, es decir esta proporción es significativamente inferior a la proporción de este tipo de instituciones en el grafo completo (ver figura 41). El detalle de las instituciones involucradas y sus medidas de centralidad puede verse en los anexos (Cuadro 13).

4.5.7. CONGLOMERADO 9: DIRECCIÓN REGIONAL DE SALUD DE LORETO

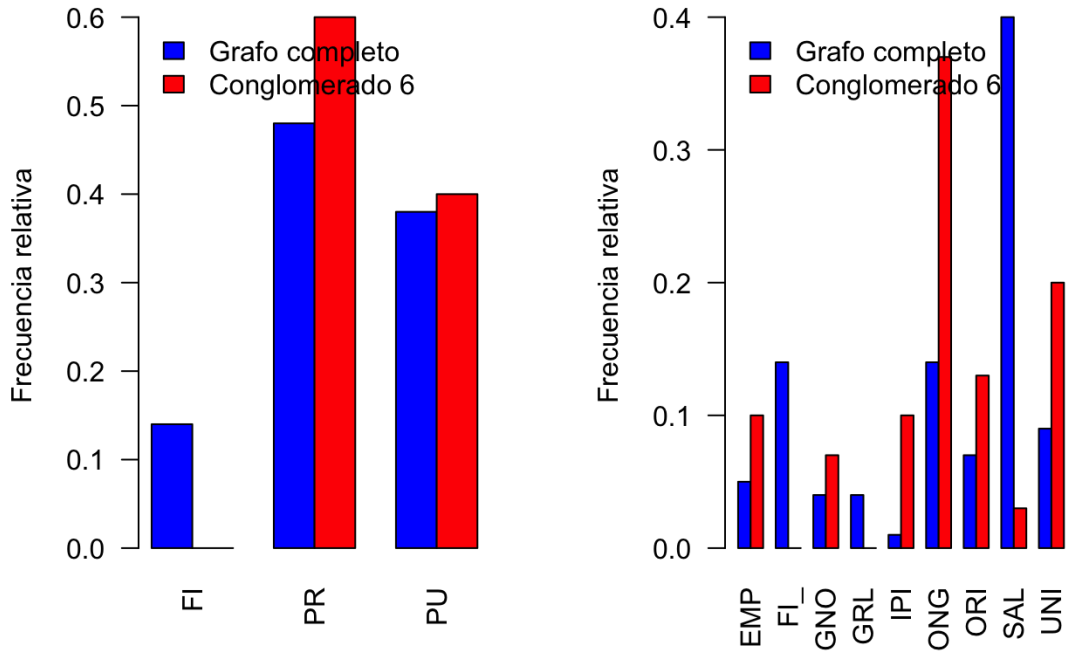
El conglomerado seis está liderado por la Dirección Regional de Salud de Loreto (DRSLO). Las instituciones en este conglomerado que están en el primer diez por ciento en las tres medidas de centralidad (grado ponderado, intermediación, centralidad del vector propio) son la Dirección Regional de Salud y la Universidad Nacional San Antonio Abad del Cusco (UNSAA). Está compuesto por 30 instituciones, el 60 por ciento son instituciones pertenecientes al sector público, es decir esta proporción es significativamente superior a la proporción de este tipo de instituciones en el grafo completo. Además, casi el 45 por ciento son instituciones especializadas en la atención de salud, es decir esta proporción es ligeramente superior a la proporción de este tipo de instituciones en el grafo completo (ver figura 42). El detalle de las instituciones involucradas y sus medidas de centralidad puede verse en los anexos (Cuadro 16).

Figura 41: Conglomerado 6 (CIP) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.

(a)



(b)

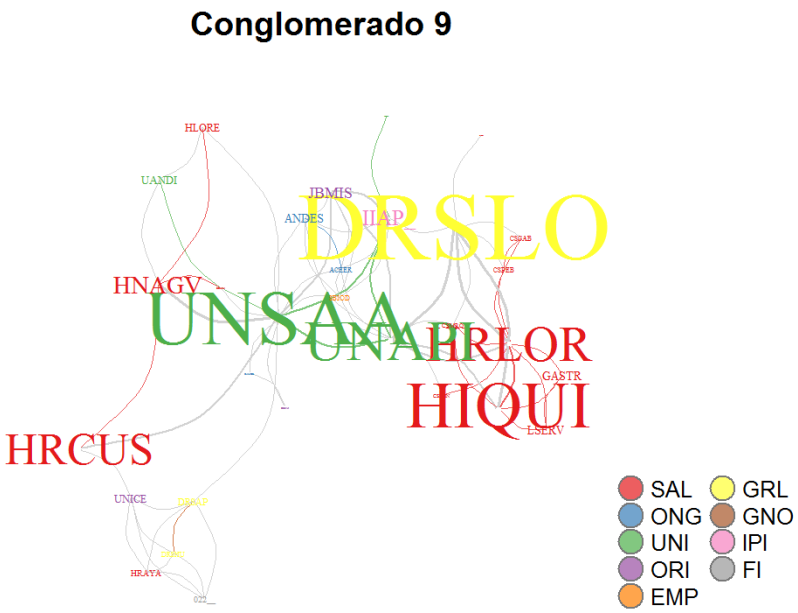


FUENTE: Elaboración propia

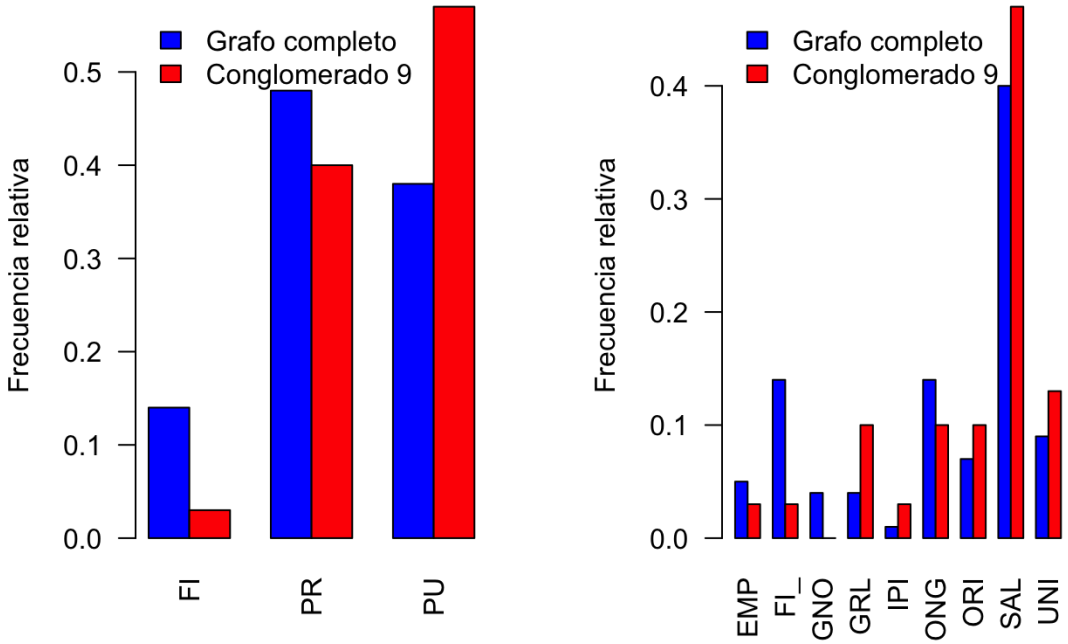
(a) En el grafo el tamaño de las etiquetas es proporcional al grado ponderado de los vértices.
(b) Comparación de la proporción de tipos de instituciones en el grafo versus proporción de instituciones en el conglomerado.

Figura 42: Conglomerado 9 (DRSLO) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.

(a)



(b)



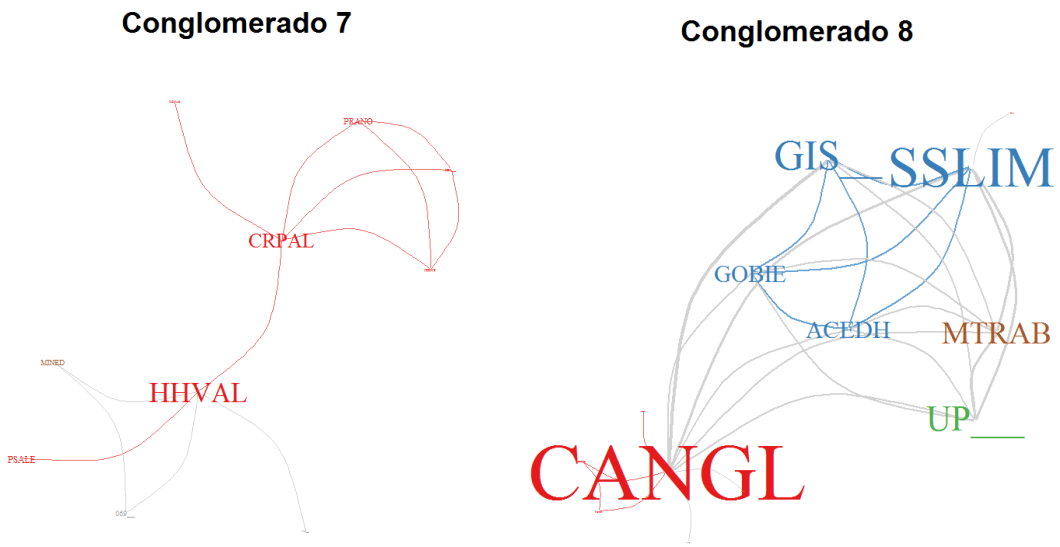
FUENTE: Elaboración propia

(a) En el grafo el tamaño de las etiquetas es proporcional al grado ponderado de los vértices.
(b) Comparación de la proporción de tipos de instituciones en el grafo versus proporción de instituciones en el conglomerado.

4.5.8. CONGLOMERADOS PEQUEÑOS

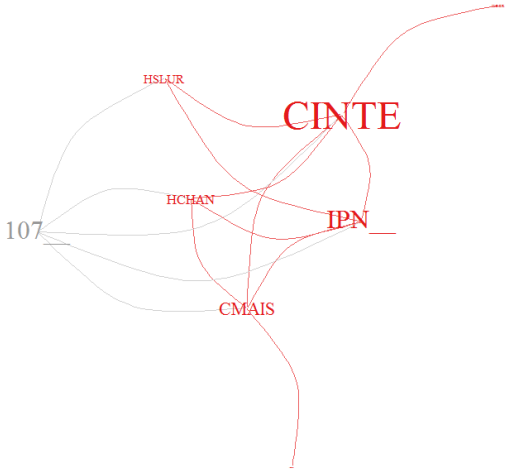
Los conglomerados 7, 8, 10, 11, 12, 13, 14, 15, 16 tienen menos de 20 vértices cada uno, el conglomerado 8 y el conglomerado 16 recogen algunos vértices que tienen un papel central en el grafo entero. En el 16 están: Impacta Salud y Educación (IMPAC) e Investigaciones Médicas en Salud (INMEN); en el 8 están: Clínica Anglo Americana (CANGL), Salud Sin Límites (SSLIM), Grupo de Investigación en Sueño (GIS), Universidad del Pacífico (UP). Los detalles de sus medidas de centralidad pueden verse en los anexos (Cuadros 14-15, 17-23).

Figura 43: Conglomerados pequeños en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.

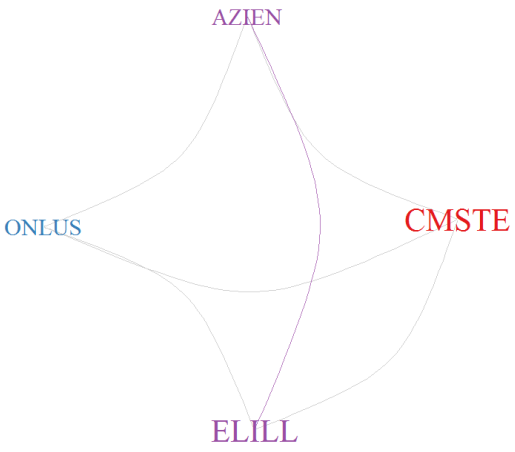


Continuación

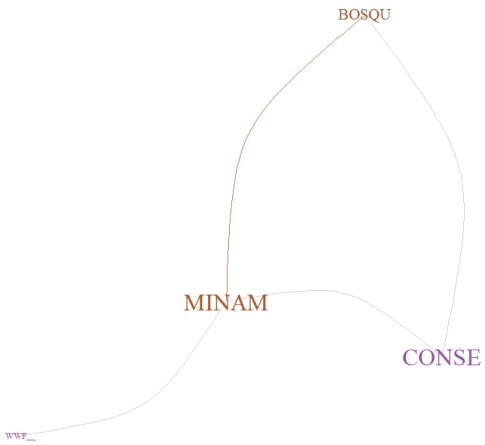
Conglomerado 10



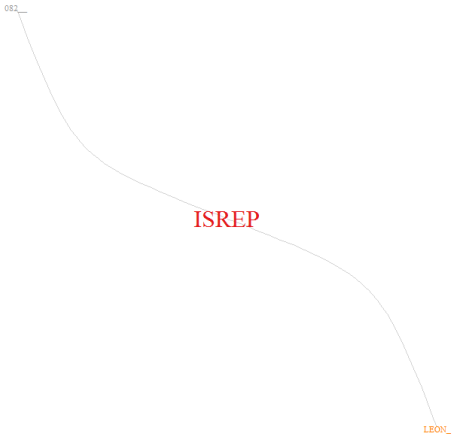
Conglomerado 11



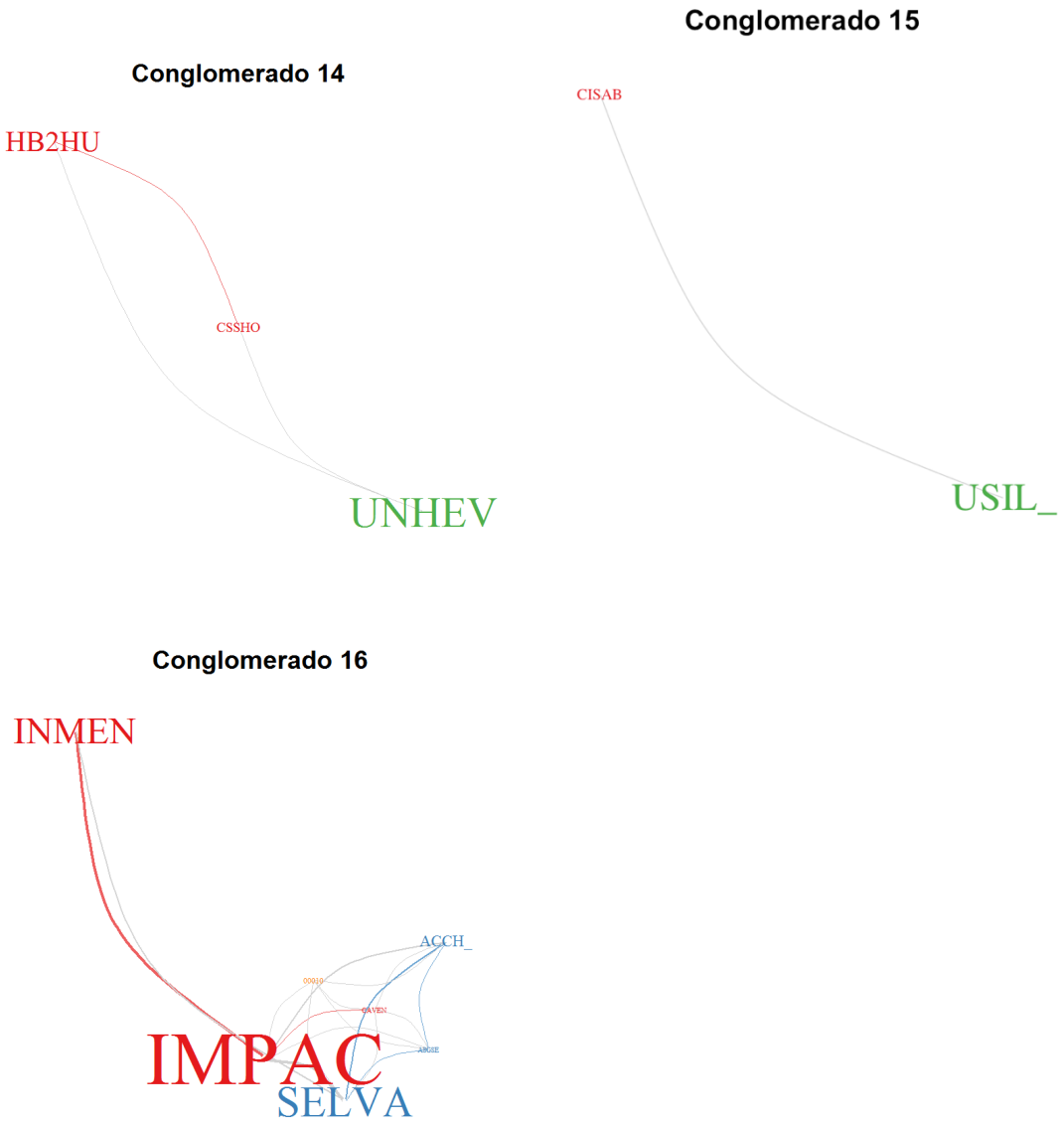
Conglomerado 12



Conglomerado 13



Continuación



FUENTE: Elaboración propia

4.6. ASORTATIVIDAD VS TIPOLOGÍA INSTITUCIONAL

Asortatividad. - El coeficiente de asortatividad (assortative mixing) mide el grado en el que vértices de características similares tienden a asociarse. Un valor de 1 indica asortatividad perfecta (sólo elementos del mismo tipo son adyacentes), un valor de -1 indica disasortatividad (sólo elementos de diferentes características se asocian). En este caso se ha

encontrado que no existe asortatividad entre ninguno de los tipos institucionales ni entre los sectores a excepción de las instituciones de atención en salud que son las que tienen una tendencia a ser vecinas de instituciones similares (ver cuadro).

Cuadro 7: Coeficiente de asortatividad de los elementos del grafo.

División		*Coeficiente*
Tipo	SAL	0.2203
	ONG	0.0814
	ORI	0.0580
	UNI	0.0210
	FI	0.0017
	EMP	-0.0188
Sector	PU	0.1317
	PR	0.0988
	FI	-0.0020

FUENTE: Elaboración propia

V. CONCLUSIONES

- Se utilizaron varios métodos de clasificación para identificar de manera única las afiliaciones institucionales. El más efectivo fue el método de vecino más cercano.
- Se describió la red y los elementos de esta, con lo cual:
    - Se clasificaron los tipos de instituciones involucradas en la producción de investigación sobre medicina. La mayoría son instituciones especializadas en atención de salud y ONGs.
    - Se graficó la red y los elementos de esta. Se encontró que la red tiene un diámetro igual a 11 y un coeficiente de clustering de 0.17, por encima del coeficiente de clustering de un grafo al azar, por lo que cumple con las características de pequeño mundo.
    - Se encontraron los vértices centrales en el grafo: la Universidad Peruana Cayetano Heredia y la Universidad Nacional Mayor de San Marcos son los actores más centrales de la red tanto por su importancia a nivel local (tienen muchos vecinos), como por su capacidad de intermediación (los caminos entre los diferentes vértices en la red con frecuencia tienen que pasar a través de ellos). Son también los vértices que se unen con más vértices que a su vez son centrales (centralidad de vector propio).
    - Sólo en el caso de las instituciones del sector salud existe tendencia a asociarse con instituciones del mismo tipo. En otros casos las instituciones no tienden a colaborar con ningún tipo de institución en particular.

- Se identificaron y describieron los conglomerados. De tal manera que:
    - Se identificaron 16 conglomerados en el grafo de coautorías de instituciones peruanas con investigación en medicina indizada en Scopus mediante particionamiento jerárquico aglomerativo.
    - Se halló que el grafo tiene una estructura comunitaria por encima de un grafo al azar (0.28) pero el número de conglomerados identificados es alto en comparación a un grafo al azar por lo que la estructura comunitaria identificada no es robusta.
    - Los conglomerados hallados en su mayoría unen diferentes tipos y sectores de instituciones.
    - El conglomerado 1, liderado por el Instituto Nacional de Enfermedades Neoplásicas, tiene una presencia de hospitales por encima de la proporción de hospitales presentes en el grafo.
    - El conglomerado 6, liderado por el Centro Internacional de la Papa, está compuesto principalmente por ONGs y organizaciones internacionales y casi no tiene participación de hospitales (sólo un hospital forma parte de esta comunidad).
    - El conglomerado 5, liderado por la Universidad Nacional de San Agustin, y el conglomerado 9, liderado por la Universidad Nacional San Antonio Abad del Cusco, tienen mayor participación del sector público.

VI. RECOMENDACIONES

- Para futuras investigaciones, si al investigador le interesa recuperar todas las afiliaciones institucionales, es conveniente utilizar para la clasificación el algoritmo de k-vecinos más cercanos.
- Tomando en cuenta que los dos principales conglomerados están liderados por la Universidad Peruana Cayetano Heredia y por la Universidad Nacional Mayor de San Marcos, la implicancia de políticas y recursos destinadas a estas dos instituciones y a que estas trabajen en colaboración, va a tener un gran impacto.
- Para el investigador que está interesado en la visualización de los grafos y quiere tener mucho control sobre el resultado final, es recomendable usar igraph de R, u otro paquete de R.
- Para el investigador que está interesado en un grafo visualmente claro, y quiere intervenir en la posición de los nodos, es recomendable usar el programa gephi.

VII. REFERENCIAS BIBLIOGRÁFICAS

Alcalde-Rabanal, JE; Lazo-González, O; Nigenda, G. 2011. Sistema de salud de Perú. Salud Pública de México 53: s243-s254.

Aldecoa, R. 2013. Detección de comunidades en redes complejas. Tesis Doctoral. Valencia, Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y Computación.

Arencibia, R. 2010. Visibilidad Internacional de la Ciencia y Educación Superior Cubanas: desafíos del estudio de la producción científica. Tesis Doctoral. Granada, La Habana, Universidad de Granada, Universidad de la Habana.

Arencibia, R; Moya, F de. 2008. La evaluación de la investigación científica: una aproximación teórica desde la cienciometría. ACIMED 17(4).

Barabási, A-L. 2016. The scale-free property. Network science. Cambridge, Cambridge University.

Barabási, A-L; Albert, R. 1999. Emergence of Scaling in Random Networks. Science 286(5439): 509-512.

Bazán, M; Sagasti, F; Cárdenas, R. 2013. Perú: avances y desafíos de los sistemas de innovación para el desarrollo inclusivo. Sistemas de innovación para un desarrollo inclusivo: la experiencia latinoamericana. México DF, Foro Consultivo Científico y Tecnológico (México); LALICS, p.155-180.

Bellis, N de. 2009. Bibliometrics and citation analysis from the Science citation index to cybermetrics. Lanham, Md., Scarecrow Press.

Börner, K. 2010. Atlas of science: visualizing what we know. Cambridge, Mass., MIT Press.

Boyack, K; Klavans, D; Paley, B. 2007. Relationship among scientific paradigms. United States, Information Esthetics

Boyack, K; Klavans, D; Small, H; Patek, M. 2013. Map of Science. Consultado 13 ene. 2013. Disponible en http://www.mapofscience.com/ (Map of Science).

Bradford, S. 1985. Sources of information on specific subjects 1934. Journal of Information Science 10(4): 176-180.

Brandenburg, FJ; Himsolt, M; Rohrer, C. 1995. An experimental comparison of force-directed and randomized graph drawing algorithms. Graph Drawing. Berlin, Springer, p.76-87, (Lecture Notes in Computer Science, no. 1027).

Bravo, Á. 2006. Análisis bibliométrico de la producción científica de México en ciencias agricolas a través de las bases de datos internacionales, «Agricola», «Agris», «Cab Abstracts», «Science Citation Index», «Social Science Citation Index» y «Tropag & Rural», en el periodo 1983-2002. Tesis Doctoral. Madrid, Universidad Carlos III de Madrid. Departamento de Biblioteconomía y Documentación.

Burt, RS. 1976. Positions in networks. Social forces 55(1): 93–122.

Carvajal, R. 2006. Comunidades en Grafos. Memoria para optar al título de Ingeniero Civil Matemático. Santiago de Chile, Universidad de Chile. Facultad de Ciencias Matemáticas y Físicas.

Chinchilla-Rodríguez, Z. 2005. Análisis del dominio científico español: 1995-2002 (ISI, Web of Science). Thesis. s.l., Universidad de Granada (Spain).

Prime Europe-Latin American Conference on Science and Innovation Policy 2008. Mexico City, 24-26 September. (2008, México DF). 2008. Inter-institutional scientific collaboration: an approach from social network analysis. Ed. Chinchilla-Rodríguez, Z; Moya, F de; Vargas-Quesada, B; Corera-Álvarez, E; Hassan-Montero, Y. México DF,

Clauset, A; Newman, ME; Moore, C. 2004. Finding community structure in very large networks (En arxiv: cond-mat/0408187). Physical Review E 70(6).

Cockburn, IM.; Henderson, RM. 1998. Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. Journal of Industrial Economics 46(2): 157-182.

Código Civil. DL No 295. 1984

Cordón, O. (2012) "Redes y sistemas complejos: cuarto curso del grado de ingeniería informática: tema 3: Redes Sociales Centralidad".

Cronin, B; Shaw, D; La Barre, K. 2003. A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. Journal of the American Society for Information Science and Technology 54(9): 855-871.

Csardi, G; Nepusz, T. 2006. The igraph software package for complex network research. InterJournal Complex Systems: 1695.

Cuxac, P; Lamirel, J-C; Bonvallot, V. 2013. Efficient supervised and semi-supervised approaches for affiliations disambiguation. Scientometrics 97(1): 47-58.

Easley, D; Kleinberg, J. 2010. Strong and weak ties. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. 1 edition New York, Cambridge University Press.

Euler, L. 1741. Solutio problematis ad geometriam situs pertinentis (En reimpreso en opera omnia series prima, vol. 7. pp. 1-10, 1766). Commentarii academiae scientiarum Petropolitanae 8: 128–140.

Fortunato, S. 2010. Community detection in graphs. Physics Reports 486(3-5): 75-174.

Fortunato, S; Barthélemy, M. 2007. Resolution limit in community detection. Proceedings of the National Academy of Sciences of the United States of America 104(1): 36-41.

Fortunato, S; Castellano, C. 2012. Community Structure in Graphs. Computational Complexity. New York, NY, Springer New York, p.490-512.

Freeman, C. 1993. La experiencia de Japón: el reto de la innovación. Caracas, Galac, xxiii, 200.

Freeman, LC. 1977. A set of measures of centrality based on betweenness. Sociometry 1977: 35–41.

Gibson, H; Faith, J; Vickers, P. 2013. A survey of two-dimensional graph layout techniques for information visualisation. Information Visualization 12(3-4): 324-357.

Girvan, M; Newman, ME. 2002. Community structure in social and biological networks. Proceedings of the national academy of sciences 99(12): 7821–7826.

Glänzel, W. b. 2001. Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies. Library Trends 50(3): 461-473.

Gómez, I; Fernández, MT; Bordons, M; Morillo, F. 2004. Proyecto de obtención de indicadores de producción científica y tecnológica de España (1996-2001). Madrid, Consejo Superior de Investigaciones Científicas.

Good, BH; de Montjoye, Y-A; Clauset, A. 2010. The performance of modularity maximization in practical contexts (En arxiv: 0910.0165). Physical Review E 81(4).

Granovetter, M. 1976. Network sampling: Some first steps. American Journal of Sociology 81(6): 1287–1303.

Granovetter, MS. 1973. The strength of weak ties. American journal of sociology 1973: 1360–1380.

Gross, JL; Yellen, J; Zhang, P. eds. 2014. Handbook of graph theory. 2 ed. Boca Raton London New York, CRC Press, a Chapman & Hall book, (Discrete mathematics and its aplications).

Hanneman, RA; Riddle, M. 2005. Chapter 18: Some statistical tools. Introduction to social network methods. Riverside, CA, University of California, Riverside.

Hicks, DM; Katz, JS. 1996. Where Is Science Going? Science, Technology, & Human Values 21(4): 379-406.

Hjørland, B. 2002. Domain analysis in information science: eleven approaches - traditional as well as innovative. Journal of Documentation 58(4): 422-462.

Hjørland, B; Albrechtsen, H. 1995. Toward a new horizon in information science: Domain-analysis. Journal of the American Society for Information Science 46(6): 400-425.

INEI (Instituto Nacional de Estadística e Informática, PE). 2011. II Censo Nacional Universitario 2010. Lima, INEI, ANR, 456.

Jacomy, M; Venturini, T; Heymann, S; Bastian, M. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE 9(6): e98679.

Jurka, TP; Collingwood, L; Boydstun, AE; Grossman, E; Atteveldt, W van. 2014. RTextTools: Automatic Text Classification via Supervised Learning

Kolaczyk, ED. 2009. Statistical analysis of network data: methods and models. Boston, Springer.

Kolaczyk, ED; Csárdi, G. 2014. Statistical Analysis of Network Data with R. New York, Springer, v.65, (Use R!).

Larose, DT; Larose, CD. 2014. Discovering knowledge in data: an introduction to data mining. 2 ed. Hoboken, Wiley, 316, (Wiley series on methods and applications in data mining).

Lemola, T; Halme, K; Viljamaa, K; Peña-Ratinen, C. 2012. Diagnóstico del desempeño y necesidades de los Institutos Públicos de Investigación y Desarrollo del Perú: informe final. Lima, Advancis, Finnish Innovation and Technology Group, 102.

Ley 26887. Ley general de sociedades. 1997

Ley 27444. Ley del Procedimiento Administrativo General. 2001

Ley 28303. Ley Marco de Ciencia, Tecnología e Innovación Tecnológica. 2004

Liben-Nowell, D.; Kleinberg, J. 2007. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58(7): 1019-1031.

Lotka, AJ. 1926. The frequency distribution of scientific productivity. Journal of Washington Academy Sciences 16(12).

Lucio-Arias, D. 2013. Cobertura Web of Science y Scopus, el caso de Colombia [Presentación]. IX Congreso Iberoamericano de Indicadores de Ciencia y Tecnología 2013.

Lundvall, B-Å. 1988. Innovation as an interactive process: from user-producer interactions to the national system of innovation. Technical Change and Economic Theory. London, Pinter, p.349–369.

_____. 1992. National systems of innovation: towards a theory of innovation and interactive learning. London, Pinter.

Málaga, L. 2014. Indicadores bibliométricos en medicina de las instituciones peruanas (2009-2011). Tesis de licenciatura. Lima, Universidad Nacional Mayor de San Marcos. 277 p.

Miguel, S. 2008. Aproximación cienciométrica al análisis y visualización del dominio científico argentino, 1990-2005. Tesis Doctoral. Madrid, Universidad de Granada. 657 p.

Miguel, S; Moya, F de. 2009. Aproximación cienciométrica al análisis y visualización del dominio científico argentino, 1990-2005. IV Encuentro de Jóvenes Investigadores (I Escuela Doctoral Iberoamericana) de Estudios Sociales y Políticos sobre la Ciencia y la Tecnología, Caracas, Venezuela, 21 al 24 de abril de 2009. Caracas, Instituto Venezolano de Investigaciones Científicas, p.2-32.

Miguel, S; Moya, F de; Herrero-Solana, V. 2006. Aproximación metodológica para la identificación del perfil y patrones de colaboración de dominios científicos universitarios. Revista española de Documentación Científica 29(1): 36-55.

_____. 2008. A new approach to institutional domain analysis: Multilevel research fronts structure. Scientometrics 74(3): 331-344.

Miguel, S; Moya, F; Herrero-Solana, V. 2007. El análisis de co-citas como método de investigación en Bibliotecología y Ciencia de la Información. Investigación bibliotecológica 21(43): 139-155.

MINEDU (Ministerio de educación, PE). 2006. La universidad en el Perú: razones para una reforma universitaria: informe 2006. Lima, MINEDU, (Cuadernos de reflexión y debate, no. VII).

MINSA (Ministerio de Salud, PE). 2011. Categorías de establecimientos del sector Salud: Norma técnica de salud NTS N° 021-MINSA 1 DGSP-V.03

Moed, HF. 2006. Bibliometric Rankings of World Universities. Leiden, Centre for Science and Technology Studies (CWTS) Leiden University.

Mongeon, P; Paul-Hus, A. 2015. The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics 106(1): 213-228.

Moody, J. 2004. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. American Sociological Review 69(2): 213-238.

Moreno, JL. 1934. Who shall survive? A new approach to the problem of human interrelations. Washington, Nervous and Mental Disease Pub. Co.

Moya, F; Chinchilla-Rodríguez, Z; Corera-Álvarez, E; Vargas-Quesada, B; Muñoz-Fernández, FJ; Herrero-Solana, V. 2005. Análisis de dominio institucional: la producción científica de la Universidad de Granada (SCI, 1991-99). Revista española de documentación científica 28(2): 170–195.

Moya, F de; Chinchilla-Rodríguez, Z; Corera-Álvarez, E; González, A; Muñoz-Fernández, FJ; Vargas-Quesada, B. (2006) "Resultados de investigación científica con visibilidad internacional del Principado de Asturias.".

Moya, F de; Herrero-Solana, V; Jiménez-Contreras, E. 2006. A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library and information science research. Journal of Information Science 32(1): 63–77.

Moya, F de; Herrero-Solana, V; Vargas-Quesada, B; Chinchilla-Rodríguez, Z; Corera-Álvarez, E; Muñoz-Fernández, F; Guerrero, V; Olmeda, C. 2004. Atlas de la ciencia española: propuesta de un sistema de información científica. Revista española de Documentación Científica 27(1): 11-29.

Moya, F de; Solís, F; Sánchez-Malo, F; Corera-Álvarez, E; Chinchilla-Rodríguez, Z; Hassan-Montero, Y; Herrero-Solana, V; Muñoz-Fernández, FJ; Navarrete, J; Ruiz, E; Vargas-Quesada, B. 2004. Indicadores científicos de la producción andaluza en biomedicina y ciencias de la salud (ISI, Web of Science, 1990-2002). Andalucía, Junta de Andalucía. Consejería de Salud.

Moya, F de; Vargas-Quesada, B; Chinchilla-Rodriguez, Z; Corera-Álvarez, E; Herrero-Solana, V; Munoz-Fernández, FJ. 2005. Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation. Information Processing & Management 41(6): 1520-1533.

Moya, F de; Vargas-Quesada, B; Chinchilla-Rodríguez, Z; Corera-Álvarez, E; Muñoz-Fernández, FJ; Herrero-Solana, V. 2005. Cocitación de clases y categorías: Proyecto Atlas de la Ciencia. El Estado de la Ciencia. Principales Indicadores de Ciencia y Tecnología Iberoamericanos/Interamericanos 2004. Buenos Aires, RICYT, p.1–18.

Moya, F de; Vargas-Quesada, B; Corera-Álvarez, E; Muñoz-Fernández, FJ; Herrero-Solana, V; González-Molina, A; Chinchilla-Rodríguez, Z. 2006. Visualización y análisis de la estructura científica española: ISI Web of science 1990-2005. El profesional de la información 15(4): 258–269.

Moya, F de; Vargas-Quesada, B; Herrero-Solana, V; Chinchilla-Rodríguez, Z; Corera-Álvarez, E; Munoz-Fernández, FJ. 2004. A new technique for building maps of large scientific domains based on the cocitation of classes and categories. Scientometrics 61(1): 129-145.

Narin, F; Stevens, K; Whitlow, ES. 1991. Scientific co-operation in Europe and the citation of multinationally authored papers. Scientometrics 21(3): 313-323.

Nelson, RR. ed. 1993. National Innovation Systems: A Comparative Analysis. New York, Oxford University Press, 541.

Newman, ME. 2001. Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 64(1 II): 016131/1-016131/8.

_____. 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality (En cited by 1016). Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 64(1 II): 016132/1-016132/7.

_____. 2001. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98(2): 404-409.

_____. 2003. The structure and function of complex networks. SIAM review 45(2): 167–256.

_____. 2004. Analysis of weighted networks. Physical Review E 70(5).

_____. 2004. Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America 101(SUPPL. 1): 5200-5205.

_____. 2006. Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74(3).

_____. 2010. Networks: an introduction. Oxford ; New York, Oxford University Press, 772.

Newman, ME; Girvan, M. 2004. Finding and evaluating community structure in networks. Physical review E 69(2): 26113.

NSF (National Science Foundation, US). 2012. National Science Board. Consultado 29 dic. 2016. Disponible en https://www.nsf.gov/statistics/seind12/ (Science and engineering indicators 2012).

OECD (Organisation for Economic Co-operation and Development). 2002. Manual de Frascati: medición de las actividades científicas y tecnológicas, propuesta de norma práctica para encuestas de investigación y desarrollo experimental. s.l., Fundación Española para la Ciencia y la Tecnología FECYT.

_____. 2015. Institutional sectors and classifications for R&D statistics. In OECD (Organisation for Economic Co-operation and Development). Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development. s.l., OECD Publishing, p.81-107.

OEI (Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura). 2014. Ciencia, tecnología e innovación para el desarrollo y la cohesión social: un programa iberoamericano en la década de los bicentenarios. Madrid, OEI.

O'Malley, AJ; Marsden, PV. 2008. The analysis of social networks. Health Services and Outcomes Research Methodology 8(4): 222-269.

Pan, RK; Saramäki, J. 2012. The strength of strong ties in scientific collaboration networks (En arxiv: 1106.5249). EPL (Europhysics Letters) 97(1): 18007.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

RStudio. 2016. RStudio

Sagasti, F. 2013. Ciencia, tecnología, innovación: políticas para América Latina. 2 ed. Lima/México, Fondo de Cultura Económica.

Salaverry, O; Cárdenas-Rojas, D. 2009. Establecimientos asistentciales del sector salud. Rev Peru Med Exp Salud Publica 26(2): 264-267.

Salazar, M; Lucio, J; Rivera, SC; Bernal, E; Ruiz, C; Usgame, G; Lucio-Arias, D; Daza Caicedo, S; Guerrero C., J; Guevara, A; Perea, GI; Cifuentes, F; García, M; Pérez, S; Sanchez, E; Observatorio Colombiano de Ciencia y Tecnología OCyT. 2011. Indicadores de ciencia y tecnología, Colombia 2011. Bogotá, Observatorio Colombiano de Ciencia y Tecnología OCyT, p.214.

SCImago Research Group. 2006. Atlas of Science. Consultado 9 ene. 2013. Disponible en http://www.atlasofscience.net/atlas_of_science.htm

Scopus content coverage. 2016. ene. 2016.

Slone, RM. 1996. Coauthors' contributions to major papers published in the AJR: Frequency of undeserved coauthorship. American Journal of Roentgenology 167(3): 571-579.

Spinak, E. 1998. Indicadores cienciométricos. Ci. Inf., Brasília 27(2): 141-148.

Stumpf, MPH; Wiuf, C; May, RM. 2005. Subnets of scale-free networks are not scale-free: Sampling properties of networks. Proceedings of the National Academy of Sciences 102(12): 4221-4224.

SUNEDU (Superintendencia Nacional de Educación Universitaria, PE). 2017. Universidades. Consultado 17 feb. 2017. Disponible en https://www.sunedu.gob.pe/universidades/ (Sunedu).

Torres, D. 2007. Diseño de un sistema de información y evaluación científica. Análisis ciencimétrico de la actividad investigadora de la Universidad de Navarra en el área de Ciencias de la Salud. 1999-2005. Tesis Doctoral. Granada, Universidad de Granada.

Torres, D. 2010. La Visualización de la Información en el entorno de la Ciencia de la Información. Tesis Doctoral. Granada, La Habana, Universidad de Granada, Universidad de la Habana.

Torres, JA. 2009. Desarrollo científico de las Ciencias Sociales en México; análisis bibliométrico del período 1997-2006: Social Science Citation Index (SSCI-ISI) y CiteSpace. Tesis Doctoral. s.l., Universidad de Granada.

Van der Loo, MP. 2014. The stringdist package for approximate string matching. The R Journal 6(1): 111-122.

Vargas-Quesada, B; de Moya-Anegón, F; Chinchilla-Rodríguez, Z; Corera-Álvarez, E; Guerrero-Bote, VP. 2008. Evolución de la estructura científica española: ISI Web of Science 1990-2005. El Profesional de la Informacion 17(1): 22-37.

Yan, E; Ding, Y. 2012. A framework of studying scholarly networks. Proceedings of 17th International Conference on Science and Technology Indicators. Montreal, Science-Metrix and OST, v.2, p.917-926.

Zachary, WW. 1977. An information flow model for conflict and fission in small groups. Journal of anthropological research 33(4): 452–473.

VIII. ANEXOS

8.1 Anexo 1: Conglomerados, sus componentes y medidas de centralidad

Cuadro 8: Vértices del conglomerado 1 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
*PU_SAL_INEN_*	103	4209.469452	5.96E-02
PU_UNI_UNITR	44	4481.45482	1.92E-02
*PR_UNI_PUCP_*	38	5633.147543	1.84E-02
PU_SAL_HNCAS	36	1815.47202	8.66E-03
PU_SAL_HRDTR	39	715.976282	8.62E-03
PU_SAL_H4ESH	32	214.296503	7.08E-03
PU_SAL_HEGRA	8	469.059185	5.68E-03
PR_SAL_CSPAB	9	44.587133	5.24E-03
PU_SAL_HCFAP	14	1612.3744	5.21E-03
PR_SAL_ESCAL	7	302.999519	4.36E-03
PU_SAL_HRPUC	15	522.751108	4.35E-03
PU_SAL_CSPRO	6	248.26323	4.23E-03
PR_SAL_CSANA	24	0	3.60E-03
PU_SAL_HCCHI	24	0	3.60E-03
PU_SAL_HMOLI	24	0	3.60E-03
PU_SAL_HASUL	24	0	3.60E-03
PU_SAL_IPOFT	3	455.418615	3.51E-03
PU_GRL_DRSLL	8	37.418694	3.50E-03
PR_SAL_CCENT	9	1130	3.02E-03
PU_GNO_FAP__	5	0	2.77E-03
PR_ORI_CEPIS	5	336.591027	2.54E-03
PR_UNI_ULIMA	5	1823.995172	2.05E-03
PU_SAL_CSCAN	2	113.695631	2.02E-03
PR_SAL_ONCOS	22	623.83427	1.75E-03
PR_ONG_ORDEL	5	212.972321	1.67E-03
PR_EMP_CETOX	4	149.244347	1.53E-03
PR_SAL_REPRO	2	63.770999	1.49E-03

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_EMP_ALTAI	3	0	1.41E-03
PU_SAL_HFAPP	3	335.589315	1.39E-03
*PU_SAL_HVLE_*	14	734.055677	1.34E-03
PU_SAL_HBTRU	7	384.940588	1.33E-03
PU_SAL_IRENN	9	0	7.72E-04
PR_ONG_LLCAN	5	208.90599	6.62E-04
PR_SAL_CDMOL	4	0	4.45E-04
PR_EMP_I3RES	2	1.894726	2.45E-04
PU_SAL_HIRAP	2	12.984317	1.23E-04
PR_EMP_MEINT	6	0	1.21E-04
FI_FI__015__	1	0	1.20E-04
FI_FI__099__	1	0	1.20E-04
PR_SAL_CRADI	1	0	1.20E-04
PU_GNO_MSICA	10	0	1.17E-04
PU_GNO_MBRUN	9	0	7.88E-05
FI_FI__086__	3	8.712733	5.86E-05
PR_SAL_CSGAB	6	0	5.16E-05
PU_SAL_CAPMT	2	0	3.87E-05
PU_SAL_CSPAM	2	0	3.87E-05
PR_ONG_CAO__	1	0	3.86E-05
PR_SAL_ESCAC	1	0	3.86E-05
PR_ONG_INDEA	1	0	3.69E-05
PR_EMP_CBBRE	1	0	3.69E-05
*PU_GRL_BMSI_*	1	0	3.69E-05
PU_SAL_HRD__	4	3.334127	3.07E-05
PR_SAL_CARDI	1	0	1.74E-05
PR_ONG_AHIGA	2	0	6.10E-06
*PR_ORI_REED_*	2	0	6.10E-06
PR_UNI_UNIFE	2	0	4.13E-06
PR_EMP_LEARN	2	0	4.13E-06
PR_ONG_ANEUR	1	0	4.12E-06
PR_ORI_DFARB	1	0	3.53E-06

FUENTE: Elaboración propia

Cuadro 9: Vértices del conglomerado 2 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PU_UNI_UNMSM	1021	36430.03543	6.83E-01
PU_IPI_INS__	658	14530.68034	4.47E-01
PU_GNO_MINSA	457	14781.41218	2.99E-01
*PU_SAL_INSN_*	310	4369.453428	2.31E-01
PU_SAL_HDOSM	300	2146.744662	2.01E-01
PU_GRL_DRSLI	119	2991.917096	8.70E-02
*PU_SAL_HNHU_*	130	1526.683333	7.75E-02
*PU_SAL_INMP_*	97	2138.488974	6.22E-02
PR_ORI_PIH__	105	51.155974	6.21E-02
PR_ORI_OPS__	36	655.116406	3.26E-02
*PU_UNI_UNFV_*	45	919.456022	2.06E-02
PU_SAL_HNSEB	57	1994.636989	1.91E-02
PU_SAL_ESSAL	32	997.544346	1.82E-02
PU_UNI_UNICA	36	2552.872369	1.69E-02
PR_ONG_AVANS	13	208.800716	1.20E-02
PU_GRL_DRSMA	13	749.681972	1.16E-02
PU_GRL_DRSUC	9	46.056341	9.94E-03
*PU_IPI_IPEN_*	7	566	8.80E-03
PR_ONG_ANMED	8	194.760755	7.93E-03
*PR_UNI_ESAN_*	9	908.578978	7.86E-03
PU_GRL_DRSTU	9	1026.000727	7.06E-03
PU_GNO_SENAS	10	2.8044	6.99E-03
PU_SAL_CSBAG	9	0	6.87E-03
PU_UNI_UNI__	6	575.785057	6.22E-03
*PR_ORI_UNPF_*	5	102.498898	5.93E-03
PU_UNI_UNSCH	6	238.083723	5.88E-03
FI_FI__075__	7	876.491086	5.70E-03
PR_SAL_IVIDA	7	162.583732	5.65E-03
PR_ONG_SPEIT	6	0	5.36E-03
PR_SAL_CSMON	9	289.457167	5.29E-03
PU_SAL_HCPNP	10	1589.315444	4.97E-03
PR_ONG_CIMEL	6	3.332057	4.57E-03
PU_SAL_HYURI	4	387.088877	4.34E-03
PR_UNI_UDAFF	4	332.702804	3.86E-03
FI_FI__002__	3	294.048629	3.85E-03
PR_ORI_USAID	5	0	3.82E-03

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_SAL_PSICO	3	201.492853	3.40E-03
FI_FI__016__	2	133.004887	3.39E-03
*PR_UNI_UIGV_*	2	133.004887	3.39E-03
PR_ONG_SPMED	6	41.268628	3.17E-03
PR_ONG_ALBIO	4	437.520786	3.09E-03
PR_EMP_PAGRO	4	5.48364	3.08E-03
PU_GRL_DRSIC	10	287.948087	3.07E-03
PU_GRL_DRSCU	6	421.297429	3.05E-03
PR_EMP_ASENS	2	121.686947	2.91E-03
PR_GRL_DRSTA	2	121.686947	2.91E-03
PR_ONG_PACIS	3	0	2.75E-03
PR_EMP_ADN__	2	0	2.75E-03
PR_EMP_INVET	2	0	2.75E-03
PR_ONG_ACH__	6	0	2.75E-03
FI_FI__074__	3	199.664985	2.64E-03
FI_FI__080__	3	161.023406	2.62E-03
PU_SAL_HSOLI	3	238.592822	2.61E-03
PR_ORI_GIUST	2	99.329331	2.61E-03
PU_SAL_CSSMA	2	51.726954	2.42E-03
PU_GRL_DRSAY	5	0.7	2.41E-03
FI_FI__076__	3	74.984751	2.33E-03
PR_EMP_00005	4	77.054326	2.31E-03
FI_FI__136__	3	175.646117	2.28E-03
PU_SAL_CSOLL	2	58.724628	2.27E-03
PR_ONG_CAAAP	2	58.724628	2.27E-03
PR_SAL_IPICM	2	58.724628	2.27E-03
PR_ONG_SPMMA	5	201.430145	2.17E-03
PU_GRL_DRSAM	3	214.810586	2.16E-03
PR_SAL_CSANT	6	64.414439	2.04E-03
FI_FI__045__	5	217.212638	2.03E-03
PU_SAL_HAPIC	2	0	2.02E-03
PR_SAL_CGONZ	4	56.899558	2.01E-03
PU_SAL_HMERC	3	90.580729	2.00E-03
PU_GNO_MINAG	2	51.003278	1.98E-03
PU_GNO_SERVI	2	51.003278	1.98E-03
PU_GNO_CONGR	2	0	1.80E-03
*PR_ONG_ADRA_*	2	0	1.80E-03
PU_SAL_HRICA	4	265.22282	1.78E-03

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
FI_FI__149__	2	31.252417	1.78E-03
PR_ORI_EMUSA	2	35.595412	1.73E-03
FI_FI__079__	3	48.294585	1.72E-03
FI_FI__109__	3	83.982545	1.61E-03
PU_SAL_CSHUA	3	37.683393	1.51E-03
FI_FI__010__	3	25.754066	1.44E-03
FI_FI__054__	3	25.754066	1.44E-03
FI_FI__067__	2	19.015647	1.41E-03
PR_ORI_CLIMA	2	85.562007	1.39E-03
FI_FI__137__	2	283	1.38E-03
PU_SAL_PSYAN	2	0	1.38E-03
PU_SAL_CSMOR	2	0	1.38E-03
PR_SAL_CLIMA	2	0	1.38E-03
PU_SAL_HDMER	2	0	1.38E-03
PR_SAL_IOOFT	1	0	1.37E-03
PR_EMP_YAVAC	1	0	1.37E-03
PR_SAL_IHEMA	1	0	1.37E-03
PU_SAL_PSPIC	1	0	1.37E-03
PR_ONG_FYUNK	1	0	1.37E-03
FI_FI__132__	1	0	1.37E-03
PR_SAL_HSMAR	1	0	1.37E-03
PR_SAL_SANNA	1	0	1.37E-03
PR_ONG_PLEYE	1	0	1.37E-03
FI_FI__042__	1	0	1.37E-03
*PR_ORI_IFEA_*	1	0	1.37E-03
PR_ONG_AMETR	1	0	1.37E-03
PR_EMP_PERUP	1	0	1.37E-03
PR_ONG_CONOP	1	0	1.37E-03
PU_SAL_HACAJ	1	0	1.37E-03
PU_SAL_HGSJP	1	0	1.37E-03
PU_SAL_HRDIC	8	15.166946	1.33E-03
PU_SAL_INCOR	2	26.166407	1.01E-03
PR_ORI_PRA__	2	10.201235	1.01E-03
PU_SAL_H3FTG	3	59.406108	9.40E-04
PU_GNO_PENCH	2	11.887921	9.36E-04
PU_GRL_DRSAN	1	0	9.00E-04
PR_ONG_SAIST	1	0	9.00E-04
PR_SAL_VESAL	2	12.69931	8.70E-04

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_ONG_VISIO	3	214.113379	7.65E-04
PU_SAL_CSCOL	3	64.666048	6.95E-04
PR_ONG_SCPED	4	7.311438	6.87E-04
PU_SAL_CSLUR	4	135.781453	6.45E-04
PR_SAL_ACESA	2	0	6.07E-04
PR_ORI_AISLC	1	0	6.02E-04
PR_SAL_INGEC	3	0	4.68E-04
PR_SAL_WINAY	1	0	4.65E-04
PR_ONG_APCIR	2	0	1.67E-04
FI_FI__ICA__	2	0	6.80E-05
PR_ONG_RMICA	1	0	3.40E-05
PU_SAL_HSMSI	1	0	3.40E-05
PR_ORI_PTATI	1	0	2.33E-05
*FI_ONG_PCAR_*	1	0	1.77E-05
PR_EMP_MINSU	1	0	1.58E-05
FI_FI__126__	1	0	1.00E-05

FUENTE: Elaboración propia

Cuadro 10: Vértices del conglomerado 3 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_UNI_UPC__	325	9892.390492	1.83E-01
PU_SAL_HREBA	212	8934.521595	1.00E-01
PU_SAL_HNDAC	159	2005.876418	9.23E-02
*PR_UNI_USMP_*	201	8628.223172	9.22E-02
PR_SAL_IIN__	130	1851.931746	8.39E-02
PU_SAL_HALME	124	3719.500822	8.29E-02
PU_SAL_HNDSB	88	735.291192	4.39E-02
*PU_SAL_HEPP_*	72	2313.779211	3.77E-02
PR_ONG_PROES	38	0	2.86E-02
PR_UNI_UCS__	48	1682.843764	2.65E-02
PU_UNI_UNP__	84	3507.520471	2.44E-02
PR_UNI_UCV__	36	2275.14085	1.50E-02
PR_UNI_URP__	42	3550.262868	1.31E-02

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
*PU_SAL_HRCH_*	50	1773.013841	1.19E-02
PU_GRL_DRSCA	20	56.252956	1.00E-02
PR_EMP_AMISS	19	295.213145	8.83E-03
*PR_UNI_UPAO_*	27	2689.767595	8.79E-03
PU_SAL_HYANA	27	76.966084	8.62E-03
PU_SAL_HRGDV	27	594.342957	7.91E-03
PU_SAL_PMARE	22	0	7.85E-03
PU_SAL_CSWAN	22	0	7.85E-03
PU_SAL_HVITA	9	0	7.70E-03
PR_ONG_COLME	23	3294.961944	7.25E-03
PR_ONG_SCMEP	22	1145.294028	7.22E-03
*PR_SAL_GPIN_*	11	81.134898	6.95E-03
PU_SAL_HNAAA	16	386.846111	6.01E-03
PU_UNI_UNJFS	21	2952.943447	5.95E-03
PU_SAL_PSRAT	7	40.941127	5.71E-03
PU_GNO_EJERC	25	3115.07206	5.40E-03
PU_SAL_HNOFT	12	2462.649923	5.28E-03
PU_UNI_UNC__	11	930.895676	4.93E-03
PU_SAL_HRCHI	13	22.437649	4.66E-03
PU_SAL_HHERV	8	0	4.58E-03
PR_ONG_GRADE	13	566	4.49E-03
FI_FI__021__	4	0	4.43E-03
PU_GNO_IMLMP	8	936.97747	4.05E-03
PU_GNO_EIEJE	8	882.649934	4.04E-03
PU_SAL_CSURU	12	1476.69458	3.97E-03
PR_SAL_LCROE	4	590.630562	3.94E-03
PU_SAL_ESIQU	11	368.282522	3.80E-03
FI_UNI_EEP__	10	549.668943	3.59E-03
*PR_UNI_USAT_*	5	488.717384	3.59E-03
PU_SAL_HRMNB	13	46.682675	3.50E-03
PR_UNI_UPSJB	5	5.896964	3.00E-03
PR_EMP_ABBOT	6	823.853468	2.76E-03
PU_GNO_MEF__	4	241.483874	2.67E-03
FI_FI__072__	3	230.95027	2.53E-03
PR_SAL_CMONT	3	0	2.38E-03
PU_SAL_HEJCU	13	393.07076	2.23E-03
FI_FI__146__	2	157.921075	2.21E-03
PR_GRL_DRSPA	3	110.179477	2.20E-03

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PU_SAL_HLAFR	3	110.179477	2.20E-03
PR_UNI_UCONT	7	264.88315	2.15E-03
PU_SAL_HCARR	2	31.243168	2.10E-03
FI_FI__037__	2	31.243168	2.10E-03
PU_SAL_HCAJA	8	417.977337	1.77E-03
PR_SAL_CTEZZ	5	435.805326	1.62E-03
PR_SAL_ARMON	13	15.110443	1.58E-03
PU_SAL_CSNAR	2	98.631382	1.56E-03
PR_EMP_NEWTR	5	55.082853	1.16E-03
FI_FI__019__	5	55.082853	1.16E-03
PR_ONG_MALLQ	4	48.813931	1.11E-03
PU_GRL_MUNHU	4	48.813931	1.11E-03
PU_GRL_CEIMP	2	0	1.07E-03
PU_UNI_UNPRG	20	389.999823	9.87E-04
PR_SAL_CLAME	5	65.925986	9.48E-04
PR_UNI_UPAGU	2	32.515053	9.20E-04
PR_SAL_CDELG	3	37.047112	8.22E-04
*PR_EMP_PATH_*	2	13.477288	7.70E-04
*PR_EMP_AMOR_*	4	0	7.59E-04
PU_SAL_CGNAV	3	0	7.37E-04
PR_ONG_AAIRE	3	0	7.06E-04
FI_FI__030__	2	25.461999	5.70E-04
PR_SAL_IBRAZ	6	45.425194	5.45E-04
PR_SAL_LCANT	6	45.425194	5.45E-04
PR_SAL_INPPA	6	45.425194	5.45E-04
PU_SAL_HESSU	2	2.502258	5.41E-04
*PR_SAL_IIEI_*	3	9.013576	5.38E-04
FI_FI__CHACH	2	9.013576	5.37E-04
PR_SAL_MACUL	2	25.500204	5.35E-04
PR_ONG_CENEP	3	0	5.07E-04
PR_SAL_GENMO	5	0	4.46E-04
PR_ONG_GCLIR	3	14.76892	4.05E-04
PR_SAL_IGLAU	5	0	3.93E-04
PR_UNI_UWIEN	2	0	3.76E-04
PR_SAL_CBAMB	2	283	3.70E-04
PR_SAL_IMM__	1	0	3.68E-04
PR_ORI_IBLCE	1	0	3.68E-04
PR_SAL_HBA__	1	0	3.68E-04

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PU_GRL_DRSLA	7	14.155069	2.74E-04
PU_SAL_ALIAD	2	0	2.52E-04
FI_FI__104__	1	0	2.02E-04
FI_FI__117__	1	0	2.02E-04
FI_FI__119__	3	310.346661	1.97E-04
FI_FI__085__	2	0	1.86E-04
PU_SAL_INR__	2	0	1.86E-04
FI_FI__024__	1	0	1.86E-04
PR_SAL_IKIRS	1	0	1.86E-04
*PR_SAL_CBCT_*	1	0	1.86E-04
FI_FI__101__	1	0	1.86E-04
*PU_UNI_UNAS_*	1	0	1.86E-04
PR_SAL_JUVEN	1	0	1.86E-04
PR_UNI_ILIBE	1	0	1.86E-04
FI_FI__143__	1	0	1.67E-04
PR_SAL_CJP__	1	0	1.67E-04
FI_FI__003__	1	0	1.67E-04
FI_FI__063__	1	0	1.67E-04
*PU_SAL_H1SR_*	5	65.122278	1.37E-04
PR_ONG_REIDE	4	124.803948	1.24E-04
PU_SAL_CSSJP	4	3.25482	1.01E-04
PR_SAL_IDAMO	1	0	7.59E-05
PU_GNO_IEAJS	2	0	5.36E-05
PR_SAL_CRSAN	2	0	4.94E-05
PR_EMP_INCAB	1	0	4.91E-05
PU_GRL_MUNPV	1	0	2.63E-05
FI_FI__147__	1	0	2.63E-05
PU_GNO_PNP__	1	0	2.63E-05
FI_FI__148__	1	0	2.63E-05
PR_ONG_MEDRE	2	0	1.46E-05
PR_ONG_CNRES	2	0	1.46E-05
PR_ONG_APSAL	1	0	1.46E-05
PR_ORI_PATHF	1	0	1.46E-05
PR_ONG_SCEMV	2	1.005435	1.41E-05
PU_SAL_HHUAC	1	0	1.20E-05
PR_SAL_FOFTN	1	0	1.06E-05
PR_SAL_DIVIN	1	0	1.06E-05
*PR_ONG_CIES_*	1	0	9.02E-06

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_SAL_BIOLI	1	0	8.15E-06

FUENTE: Elaboración propia

Cuadro 11: Vértices del conglomerado 4 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
*PR_UNI_UPCH_*	1853	48439.08055	1.00E+00
PR_ONG_PRISM	406	3325.86591	4.72E-01
*PU_SAL_HNCH_*	371	4493.10841	4.23E-01
*PU_SAL_INCN_*	317	1965.67847	4.21E-01
PR_ORI_NAMRU	522	7178.8548	4.13E-01
PU_SAL_HLOAY	172	2350.53322	1.27E-01
*PU_SAL_HNAS_*	87	1904.66709	6.21E-02
PR_ONG_INPPA	49	929.94549	6.07E-02
*PU_SAL_INSM_*	35	925.71927	4.34E-02
PR_SAL_PPJAP	39	484.78637	3.29E-02
*PU_SAL_HAMA_*	45	2480.39184	3.01E-02
*PU_SAL_CMST_*	49	1396.73583	2.88E-02
PR_SAL_VIALI	20	177.79456	2.65E-02
PR_SAL_CSFEL	33	498.45643	2.50E-02
PU_SAL_HAYUR	15	61.9335	2.25E-02
PR_SAL_EPICE	13	23.61539	2.10E-02
PR_ORI_PAMAF	12	125.12205	1.91E-02
PR_ORI_IRDyO	18	630.15896	1.78E-02
PR_ONG_CYSTI	15	48.90602	1.68E-02
PR_SAL_HYPNO	17	0	1.62E-02
*PR_SAL_IMDR_*	9	0	1.61E-02
PU_SAL_HMC__	27	2035.98696	1.54E-02
*PU_SAL_UNCH_*	9	10.14795	1.51E-02
PR_SAL_IPBMA	11	269.20232	1.16E-02
FI_FI__102__	11	0	1.10E-02
PR_SAL_CSBOR	20	60.56813	1.03E-02
PR_SAL_IMDER	6	566	1.01E-02
*FI_FI__LIMA_*	11	1553.06104	9.68E-03
PU_SAL_HACAR	7	0	8.81E-03

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_ORI_INTER	12	1285.71629	8.39E-03
PU_SAL_CEBAR	8	320.85357	8.25E-03
PU_GNO_MSPUR	4	0	8.05E-03
PR_SAL_RESOC	9	194.16341	7.89E-03
PU_SAL_HNSUR	14	648.4739	7.55E-03
PR_ONG_ANCIE	6	70.74823	6.98E-03
PR_ONG_SCEMC	5	0	6.89E-03
PU_SAL_HRHUA	6	0	6.81E-03
FI_FI__046__	6	107.02251	6.30E-03
PR_SAL_EMIND	3	0	6.03E-03
PR_ONG_CIDDE	3	0	6.03E-03
PR_ONG_CONAP	4	0	5.43E-03
*PU_SAL_UNAP_*	8	1135.34684	5.33E-03
PU_GNO_CONCY	7	165.10133	5.28E-03
PR_ORI_FOGAR	4	252.77419	5.18E-03
PR_SAL_PPJ__	4	210.13851	4.64E-03
FI_FI__110__	4	249.06436	4.25E-03
FI_FI__036__	3	195.49364	4.23E-03
FI_FI__151__	3	0	4.14E-03
PU_SAL_HMOYO	4	0	4.11E-03
FI_FI__032__	4	1.72391	4.11E-03
PR_ONG_MRAMO	2	0	4.02E-03
PR_ORI_NEMUS	2	0	4.02E-03
FI_FI__047__	4	113.20653	3.80E-03
FI_FI__048__	4	113.20653	3.80E-03
*PU_SAL_PMCH_*	3	113.20653	3.79E-03
FI_FI__111__	4	120.928	3.62E-03
PR_ONG_APDIA	3	114.61853	3.12E-03
PU_SAL_CSKEP	3	61.4115	3.03E-03
PR_ONG_FCH__	2	14.09953	2.96E-03
PR_EMP_NATCL	2	14.09953	2.96E-03
PR_SAL_CMSER	4	163.86377	2.91E-03
PU_SAL_CNSRE	2	59.6246	2.86E-03
PR_ORI_VOXIV	2	59.6246	2.86E-03
PR_SAL_CHSPA	3	589.39723	2.86E-03
PR_SAL_CEREM	2	23.39723	2.86E-03

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_ONG_COPRE	2	91.08785	2.84E-03
*PU_SAL_HSJD_*	2	91.08785	2.84E-03
PU_SAL_CSCHI	2	91.08785	2.84E-03
PU_GNO_MARIN	3	0	2.49E-03
FI_FI__128__	2	40.27949	2.27E-03
FI_FI__020__	3	137.8934	2.10E-03
FI_FI__066__	2	18.50753	2.10E-03
*PU_SAL_HEGB_*	2	18.50753	2.10E-03
FI_FI__004__	2	18.50753	2.10E-03
FI_FI__023__	2	39.89378	2.08E-03
PR_ORI_SOLID	2	155.81892	2.05E-03
FI_FI__078__	2	518.32502	2.03E-03
FI_FI__91__	2	0	2.02E-03
PR_ONG_MAGEN	2	0	2.02E-03
PR_ONG_IESSD	1	0	2.01E-03
FI_FI__065__	1	0	2.01E-03
FI_FI__145__	1	0	2.01E-03
PR_ORI_IASSC	1	0	2.01E-03
PU_SAL_HESCU	1	0	2.01E-03
FI_FI__105__	1	0	2.01E-03
FI_FI__039__	1	0	2.01E-03
PR_ONG_IEP__	1	0	2.01E-03
PU_SAL_HASJD	1	0	2.01E-03
PR_ONG_CEDRO	1	0	2.01E-03
PR_UNI_UAP__	1	0	2.01E-03
FI_FI__014__	1	0	2.01E-03
PU_GNO_MIDIS	1	0	2.01E-03
PR_ONG_REDTR	1	0	2.01E-03
PU_UNI_UNTUM	1	0	2.01E-03
PU_UNI_UNDAC	1	0	2.01E-03
FI_FI__034__	1	0	2.01E-03
PU_SAL_HLAME	1	0	2.01E-03
FI_FI__070__	1	0	2.01E-03
FI_FI__131__	1	0	2.01E-03
FI_FI__018__	1	0	2.01E-03
PR_SAL_GENET	1	0	2.01E-03

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_SAL_LSERS	1	0	2.01E-03
PR_SAL_GUADA	1	0	2.01E-03
FI_FI__133__	1	0	8.50E-04
FI_FI__025__	1	0	8.47E-04
PR_ORI_DODGE	1	0	8.30E-04
FI_FI__061__	1	0	2.56E-04
PR_ORI_UNOPS	2	0	1.95E-05
PR_EMP_HCCON	2	0	1.95E-05
FI_FI__134__	1	0	1.69E-05
FI_FI__124__	1	0	1.69E-05
PR_EMP_GSK__	1	0	5.75E-06

FUENTE: Elaboración propia

Cuadro 12: Vértices del conglomerado 5 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
*PU_UNI_UNSA_*	75	4009.796708	2.95E-02
PU_GRL_DRSAR	27	67.4308463	2.42E-02
PU_GRL_DRSPI	42	1404.224342	2.26E-02
*PR_UNI_UCSM_*	27	2088.037159	1.86E-02
PU_SAL_HHDES	14	915.7725573	6.47E-03
*PU_SAL_HGSR_*	19	392.0161889	5.87E-03
PU_SAL_CMM32	19	211.7412135	3.99E-03
PU_GRL_DRSPU	17	54.4968781	3.73E-03
PU_UNI_UNJBG	21	553.1096245	3.66E-03
PU_SAL_CSBEL	4	324.2977007	3.50E-03
PU_GRL_DRSSM	7	1097.044882	3.16E-03
PU_SAL_CMMSU	6	83.3837132	1.55E-03
PU_UNI_UNU__	6	83.3837132	1.55E-03
PU_SAL_HTACN	19	0.3560606	1.53E-03
PR_ONG_GIB__	3	49.970273	1.47E-03
FI_FI__095__	3	24.3104665	6.78E-04
PU_SAL_HGHUA	2	7.168069	6.13E-04
FI_FI__052__	2	0	4.64E-04
PR_ORI_IRNHD	3	12.441479	1.50E-04

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_UNI_UPT__	7	0	1.42E-04
PU_SAL_HNSCA	2	0	1.19E-04
PU_SAL_CSHUN	4	73.5271557	8.59E-05
PU_SAL_HHDEL	1	0	5.94E-05
PU_SAL_PSCOR	2	14.8122541	3.77E-05
PR_EMP_PLUSP	1	0	3.75E-05
PU_SAL_HGOYE	1	0	1.30E-05
PU_UNI_UNSAM	1	0	6.35E-06

FUENTE: Elaboración propia

Cuadro 13: Vértices del conglomerado 6 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PU_UNI_UNALM	22	4167.141621	6.29E-03
*PU_UNI_UNAP_*	19	1954.793093	5.66E-03
PR_ONG_CIP__	24	5885.871823	3.83E-03
*PU_IPI_INIA_*	8	1652.79613	3.41E-03
PR_EMP_DSCON	6	125.481872	2.40E-03
PU_IPI_IGP__	8	361.596958	2.16E-03
PU_GNO_MUNAB	2	0	2.02E-03
PU_IPI_IMARP	5	1205.838007	8.67E-04
*PR_ORI_CARE_*	5	827.50197	7.31E-04
PR_ONG_QUINP	12	2.714757	1.04E-04
PU_GNO_SENAM	3	5.480668	5.27E-05
*PU_UNI_UNCP_*	3	0	3.30E-05
PU_UNI_UNH__	3	4.908795	2.53E-05
PR_EMP_MICHE	2	0	1.37E-05
*PR_ONG_CCTA_*	1	0	1.27E-05
PR_ONG_ICT__	1	0	1.27E-05
PU_UNI_UNAMB	3	128.977108	1.15E-05
PU_SAL_HRAPU	2	0	1.14E-05
PR_ONG_FOROS	1	0	1.14E-05
PR_ORI_CGIAR	3	1130	7.71E-06
PR_ORI_PREVA	2	0	7.71E-06
PR_ONG_PRODE	4	1694	7.70E-06
PR_ONG_YANAP	1	0	7.70E-06

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_ONG_ECOCE	1	0	1.74E-06
PR_EMP_PACAD	1	0	1.74E-06
PR_ONG_DESCO	2	0	1.55E-08
PR_ONG_CORBI	2	0	1.55E-08
PR_ORI_CIFOR	2	566	1.55E-08
PR_ONG_CEPEC	1	0	1.55E-08
PU_UNI_UNAMA	1	0	3.12E-11

FUENTE: Elaboración propia

Cuadro 14: Vértices del conglomerado 7 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PU_SAL_HHVAL	21	1626.39318	1.42E-02
PR_SAL_CRPAL	15	2414.65427	8.21E-03
PR_SAL_PRANO	6	0	6.05E-03
PU_SAL_PSALE	7	486.13699	4.39E-03
PU_GNO_MINED	5	189.40901	2.41E-03
FI_FI__069__	5	189.40901	2.41E-03
PR_SAL_NOVAS	2	136.99096	2.03E-03
FI_FI__115__	2	13.66301	1.40E-03
PR_SAL_IOL__	3	0	2.88E-05
PR_SAL_IRENS	3	0	2.88E-05

FUENTE: Elaboración propia.

Cuadro 15: Vértices del conglomerado 8 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_SAL_CANGL	91	2442.80912	0.066585801
PR_ONG_SSLIM	64	614.06788	0.049467811
PR_ONG_GIS__	47	20.13337	0.03399972
PR_UNI_UP___	39	80.60695	0.019750883
PU_GNO_MTRAB	35	0	0.017008556
PR_ONG_ACEDH	28	0	0.014769794

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_ONG_GOBIE	27	0	0.012770408
FI_FI__051__	2	70.51007	0.002145386
PR_EMP_RIOTI	2	70.51007	0.002145386
PR_SAL_CESPE	3	14.42941	0.000503154
PR_SAL_PHYSI	3	14.42941	0.000503154
PR_SAL_CSCOC	1	0	0.000133934
PU_SAL_CSSJS	1	0	9.95021E-05

FUENTE: Elaboración propia.

Cuadro 16: Vértices del conglomerado 9 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PU_GRL_DRSLO	79	1479.631647	5.49E-02
PU_SAL_HIQUI	54	373.1322706	2.66E-02
PU_UNI_UNSAA	64	3445.309403	2.57E-02
PU_SAL_HRLOR	41	1099.864045	1.96E-02
PU_UNI_UNAPI	47	1385.230276	1.79E-02
PU_SAL_HRCUS	37	495.7746451	1.46E-02
PU_SAL_HNAGV	21	671.0147812	1.04E-02
PU_SAL_HLORE	9	0.3179894	5.02E-03
PR_SAL_GASTR	10	464.1434254	3.24E-03
PR_SAL_LSERV	10	464.1434254	3.24E-03
PU_SAL_CSPEB	6	321.6050263	3.03E-03
PU_SAL_CSCAB	6	321.6050263	3.03E-03
PU_GRL_DRSHU	6	819.2903237	2.62E-03
PU_SAL_CSSMN	4	120.3896088	2.35E-03
PR_ORI_UNICE	9	11.5205403	1.85E-03
PU_SAL_HRAYA	7	396.3736007	1.52E-03
PR_UNI_UANDI	9	104.6204704	1.24E-03
FI_FI__022__	7	0	1.22E-03
PU_SAL_CSMRO	5	189.5583458	1.07E-03
PU_GRL_DRSAP	8	163.866316	7.36E-04
*PU_IPI_IIAP_*	18	709.8138489	5.01E-04
PR_ORI_JBMIS	13	0	2.67E-04
PR_ONG_ANDES	10	0	1.79E-04

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PU_SAL_HJUNI	1	0	1.10E-04
PR_ONG_ACEER	6	521.3998503	8.98E-05
PR_EMP_SBIOD	6	521.3998503	8.98E-05
PR_SAL_CENIM	2	3.4598857	7.27E-05
PR_ONG_CAMDE	2	0	5.18E-05
PR_ORI_CESVI	2	0	5.18E-05
PR_UNI_UCP__	1	0	1.01E-06

FUENTE: Elaboración propia.

Cuadro 17: Vértices del conglomerado 10 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_SAL_CINTE	20	800.6636	1.01E-02
PR_SAL_IPN__	13	0.5000	6.17E-03
FI_FI__107__	13	0.5000	6.17E-03
PR_SAL_CMAIS	9	1238.6320	3.13E-03
PU_SAL_HCHAN	7	451.2834	3.10E-03
PU_SAL_HSLUR	6	451.2834	3.09E-03
PR_SAL_IMMIR	2	0.0000	5.13E-05
PR_SAL_ICRIO	1	0.0000	6.29E-06

FUENTE: Elaboración propia.

Cuadro 18: Vértices del conglomerado 11 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_ORI_ELILL	4	1166.2996	2.02E-04
PR_SAL_CMSTE	4	128.0006	1.63E-05
PR_ONG_ONLUS	3	0.0000	4.40E-07
PR_ORI_AZIEN	3	0.0000	4.40E-07

FUENTE: Elaboración propia.

Cuadro 19: Vértices del conglomerado 12 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_ORI_CONSE	3	1692	8.30E-04
PU_GNO_MINAM	3	566	1.67E-06
PU_GNO_BOSQU	2	0	1.67E-06
PR_ORI_WWF__	1	0	3.36E-09

FUENTE: Elaboración propia.

Cuadro 20: Vértices del conglomerado 13 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_SAL_ISREP	3	1131	1.25E-04
FI_FI__082__	1	0	2.52E-07
*PR_EMP_LEON_*	1	0	2.52E-07

FUENTE: Elaboración propia.

Cuadro 21: Vértices del conglomerado 14 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PU_UNI_UNHEV	6	709.188445	3.08E-03
PU_SAL_HB2HU	4	171.465339	2.02E-03
PU_SAL_CSSHU	2	2.922216	1.03E-05

FUENTE: Elaboración propia.

Cuadro 22: Vértices del conglomerado 15 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
*PR_UNI_USIL_*	5	879.4165	4.29E-03
PR_SAL_CISAB	2	0.0000	1.72E-05

FUENTE: Elaboración propia.

Cuadro 23: Vértices del conglomerado 16 y sus medidas de centralidad.

Institución	*Fuerza*	*Intermediación*	*Centralidad de vector propio*
PR_SAL_IMPAC	70	193.976324	0.039918815
PR_SAL_INMEN	35	25.894829	0.02192358
PR_ONG_SELVA	40	3.601564	0.015293792
*PR_ONG_ACCH_*	15	16.399724	0.003615561
PR_SAL_CAVEN	7	79.170676	0.001626381
PR_EMP_00010	7	79.170676	0.001626381
PR_ONG_ASGSE	7	79.170676	0.001626381

FUENTE: Elaboración propia.

8.2 Anexo 2: Comandos en R

8.2.1 Análisis exploratorio de datos

```
# Comandos correspondientes a 4.1.ANÁLISIS EXPLORATORIO DE DATOS

#Paso 1, cargar datos
scopus_2000a2004 <- read.csv("bd/scopus_peru_medi_2000a2004_rec16_06_16.csv",stri
ngsAsFactors = FALSE, encoding = "UTF-8")
scopus_2005a2010 <- read.csv("bd/scopus_peru_medi_2005a2010_rec16_06_16.csv",stri
ngsAsFactors = FALSE, encoding = "UTF-8")
scopus_2011a2013 <- read.csv("bd/scopus_peru_medi_2011a2013_rec16_06_16.csv",stri
ngsAsFactors = FALSE, encoding = "UTF-8")
scopus_2014a2015 <- read.csv("bd/scopus_peru_medi_2014a2015_rec16_06_16.csv",stri
ngsAsFactors = FALSE, encoding = "UTF-8")
scopus<-rbind(scopus_2000a2004,scopus_2005a2010,scopus_2011a2013,scopus_2014a2
015)

#Paso 2, estructura de los datos
str(scopus)

## 'data.frame':    5697 obs. of  30 variables:
##  $ Authors                : chr  "Carreazo N.Y., Bada C.A., Chalco J.P., Huicho L." "Arb
oleda M., De Guzman I.N., Ticona E., Morales G., Gloria E., Obregon P., Lora A., Ganiku
```

M., Adrianzen M., Falcon L." "Castro M., Martínez R." "Iannacone J., Ayala L." ...
$ Title : chr "Audit of therapeutic interventions in inpatient children usin g two scores: Are they evidence-based in developing countries?" "Surgical repair of the an omalous origin of the right pulmonary artery from the ascending aorta [Correção cirúrgica da origem an"| __truncated__ "Process of the development of Corynosoma obtuscens (Aca nthocephala: Polymorphidae) in Canis familiaris and its possible involve"| __truncated__ " Census of Ornithodoros amblus Chamberlin (Acarina: Argasidae) in Mazorca Island, Lima , Peru [Censo de Ornithodoros amblus Chamb"| __truncated__ ...
$ Year : int 2004 2004 2004 2004 2004 2004 2004 2004 2004 2004 ...
$ Source.title : chr "BMC Health Services Research" "Arquivos Brasileiros d e Cardiologia" "Parasitologia Latinoamericana" "Parasitologia Latinoamericana" ...
$ Volume : chr "4" "83" "59" "59" ...
$ Issue : chr "" "6" "1-2" "1-2" ...
$ Art..No. : chr NA NA NA NA ...
$ Page.start : chr "" "516" "26" "56" ...
$ Page.end : chr "" "521" "30" "60" ...
$ Page.count : int NA NA NA NA NA NA NA NA NA NA ...
$ Cited.by : int 4 3 3 1 7 13 16 2 25 50 ...
$ DOI : chr "10.1186/1472-6963-4-40" "" "" "" ...
$ Link : chr "https://www.scopus.com/inward/record.uri?eid=2-s2.0-14 544294467&partnerID=40&md5=5f436a86fe45f71ebf6480913bdc0052" "https://www.sco pus.com/inward/record.uri?eid=2-s2.0-19944428577&partnerID=40&md5=2eb9e4a331d8 004dd997a7cc66dc2cf7" "https://www.scopus.com/inward/record.uri?eid=2-s2.0-5544913 2499&partnerID=40&md5=5107a78d2c3dfabdb4185acfe33ad289" "https://www.scopus.c om/inward/record.uri?eid=2-s2.0-55449091147&partnerID=40&md5=5d27f3a334d513c6 79c03bea8e6225d8" ...
$ Affiliations : chr "Univ. Nacional Mayor de San Marcos, Instituto de Salud del Niño, Lima, LI 05, Peru" "Dpto. 301, Calle Trípoli, 350, Miraflores - Lima 18 - Lima, Peru" "Laboratorio de Parasitología de Fauna Silvestre, Fac. Ciencias Biológicas, UNMS M, AV. Venezuela cuadra 34 s/n, Lima, Peru" "Laboratorio de Ecofisiología Animal, Facul tad de Ciencias Naturales y Matemáticas, Universidad Nacional Federico Villarreal, Pe"| _ _truncated__ ...
$ Authors.with.affiliations : chr "Carreazo, N.Y., Univ. Nacional Mayor de San Mar cos, Instituto de Salud del Niño, Lima, LI 05, Peru; Bada, C.A., Univ. Nacional "| __trunca

ted__ "Arboleda, M.Dpto. 301, Calle Trípoli, 350, Miraflores - Lima 18 - Lima, Peru; De Guzman, I.N.; Ticona, E.; Morales, G.; Gloria,"| __truncated__ "Castro, M., Laboratorio de Parasitología de Fauna Silvestre, Fac. Ciencias Biológicas, UNMSM, AV. Venezuela cuadra 34 s/n, Lima"| __truncated__ "Iannacone, J., Laboratorio de Ecofisiología Animal, Facultad de Ciencias Naturales y Matemáticas, Universidad Nacional Federico"| __truncated__ ...

$ Abstract : chr "Background: The evidence base of clinical interventions in paediatric hospitals of developing countries has not been formally a"| __truncated__ "The anomalous origin of the right pulmonary artery (AORPA) from the ascending aorta is a rare congenital malformation. We descr"| __truncated__ "A report of the results obtained from an experimental infection of the domestic dog Canis familiaris, with Corynosoma obtuscens"| __truncated__ "The aim of the current research was to evaluate the tick Ornithodoros amblus Chamberlin in Mazorca Island, Lima, Peru and its r"| __truncated__ ...

$ Author.Keywords : chr "" "" "Canis familiaris; Corynosoma obtuscens; Experimental infection" "Guano seabirds; Island; Ornithodoros; Survey; Tick" ...

$ Index.Keywords : chr "adrenalin; albendazole; amikacin; ampicillin; analgesic agent; antibiotic agent; antipyretic agent; beta 2 adrenergic receptor "| __truncated__ "article; ascending aorta; case report; echocardiography; female; human; infant; patent ductus arteriosus; pulmonary artery" "" "" ...

$ Correspondence.Address : chr "Huicho, L.; Univ. Nacional Mayor de San Marcos, Instituto de Salud del Niño, Lima, LI 05, Peru; email: lhuicho@viabcp.com" "Arboleda, M.; Dpto. 301, Calle Trípoli, 350, Miraflores - Lima 18 - Lima, Peru; email: miguel_arboleda@hotmail.com" "Castro, M.; Laboratorio de Parasitología de Fauna Silvestre, Fac. Ciencias Biológicas, UNMSM, AV. Venezuela cuadra 34 s/n, Lima"| __truncated__ "Laboratorio de Ecofisiología Animal, Facultad de Ciencias Naturales y Matemáticas, Universidad Nacional Federico VillarrealPeru"| __truncated__ ...

$ Editors : chr "" "" "" "" ...

$ Publisher : chr "" "" "" "" ...

$ ISSN : chr "14726963" "0066782X" "07177704" "07177704" ...

$ ISBN : chr "" "" "" "" ...

$ CODEN : chr "" "ABCAA" "" "" ...

$ PubMed.ID : int 15625006 NA NA NA NA 15572646 15536343 15546030 15663172 15592270 ...

$ Language.of.Original.Document: chr "English" "Portuguese; English" "Spanish" "Sp

```
anish" ...
##  $ Abbreviated.Source.Title     : chr  "BMC Health Serv. Res." "Arq. Bras. Cardiol." "Pa
rasitol. Latinoam." "Parasitol. Latinoam." ...
##  $ Document.Type                : chr  "Article" "Article" "Article" "Article" ...
##  $ Source                       : chr  "Scopus" "Scopus" "Scopus" "Scopus" ...
##  $ EID                          : chr  "2-s2.0-14544294467" "2-s2.0-19944428577" "2-s2.0-554
49132499" "2-s2.0-55449091147" ...
```

- *FIGURA 11. Cantidad de documentos analizados por año y tipo.*

```
##############
#Figura 11
##################
#Años
#Tipos

tr<-table(scopus$Year)[-17]
nombres<-names(tr)
par(mfrow=c(2,1))
par(las=2) # make label text perpendicular to axis
par(mar=c(5,8,4,2)) # increase y-axis margin.

counts<-as.numeric(tr)
barplot(counts, main="Años",  cex.main = 1,
      horiz=TRUE, names.arg=nombres, cex.names=0.9, col="dodgerblue")

tr<-sort(table(scopus$Document.Type),decreasing = T)
nombres<-names(tr)
#Genero gráfico
counts<-as.numeric(tr)
barplot(counts, main="Tipos",   cex.main = 1,
      horiz=TRUE, names.arg=nombres, cex.names=0.9, col="dodgerblue")

(tr)
```

```
##
##         Article          Letter          Review Conference Paper
##            4416             435             410             128
##    Book Chapter       Editorial            Note     Short Survey
##              97              93              69              24
##         Erratum Article in Press            Book
##              12              11               2
```

```
#Proporción de documentos del tipo artículo, carta, de revisión
tipo.ALR<-length(scopus$Document.Type[scopus$Document.Type=="Article"  | scopus$
Document.Type=="Letter"  |  scopus$Document.Type=="Review" ])
tipo.total<-length(scopus$Document.Type)
tipo.ALR_<-tipo.total - tipo.ALR
tipo.ALR_/tipo.total
```

```
## [1] 0.07653151
```

- *FIGURA 12. Idioma y revistas más frecuentes entre los documentos analizados*

```
###############
#Figura 12
#################
#Idiomas
tr<-sort(table(scopus$Language.of.Original.Document),decreasing = T)
nombres<-names(tr)
par(mfrow=c(2,1))
par(las=2) # make label text perpendicular to axis
par(mar=c(5,12,4,2)) # increase y-axis margin.

counts<-as.numeric(tr)
sum(tr)
```

```
## [1] 5697
```

```
lan.lb.sort<-c("en (4301)","es (844)","en,es (455)","en,pt","pt","fr","pt,en",
              "es,en","en,fr","de","en,it","en,fr,es","en,pt,fr","en,pt,es","fr,en","it")
en<-(4301+29+8+4+2+1+1+1)/sum(tr)
en

## [1] 0.7630332

tot.en<- 4301+455+29+8+6+4+2+1+1+1+1#Cuenta también las "intersecciones"
es<-(844+455+6+1+1)/sum(tr)
es

## [1] 0.229419

tot.es<-844+455+6+1+1 #cuenta también las "intersecciones"

barplot(counts, main="Idiomas", cex.main = 1,
        horiz=TRUE, names.arg=lan.lb.sort, cex.names=0.9, col="dodgerblue",cex.axis = 0.9)

#Revistas
pr<-head(sort(table(scopus$Source.title),decreasing = T), 10)
lbls<-c("Rev Per Med Exp Salud Publica (755)","PLoS ONE (243)",
        "Am J Trop Med Hyg (222)","PLos Negl Trop Dis (104)",
        "Rev Panam Salud Publica (77)","J Clin Microbiol (65)",
        "Emerg Infect Dis (63)","Int J Tuberc Lung Dis (57)",
        "The Lancet (56)","J Infect Dis (54)")
barplot(as.numeric(pr), main="Revistas más frecuentes", cex.main = 1,
        horiz=TRUE, names.arg=lbls, cex.names=0.9, col="dodgerblue",cex.axis = 0.9)
```

- *FIGURA 13. Distribución de las citaciones por documento*

```
#############################################################################
#Figura 13
################################################
#Citaciones
```

```
plot.new()
dev.off()

##                                   null                                   device
##          1

plot(table(scopus$Cited.by),ylab="Frecuencia",xlab="Veces que el documento fue citado"
,
    col="red",panel.first=grid())
```

- *FIGURA 14. Palabras clave asignadas por los autores más frecuentes entre los documentos (por quinquenios 2000 - 2015).*

```
#############################################################################
#Figura 14
################################################
#Subgráficos, cada 5 años
#Palabras clave más frecuentes:
sc2011.2015<-scopus[scopus$Year==2011 | scopus$Year==2012 | scopus$Year==2013
                |scopus$Year==2014  | scopus$Year==2015,]
dim(sc2011.2015)

## [1] 3303   30

sc2006.2010<-scopus[scopus$Year==2006  |  scopus$Year==2007  |  scopus$Year==2008
                |scopus$Year==2009                  |              scopus$Year==2010,]
dim(sc2006.2010)

## [1] 1679   30

sc2000.2005<-scopus[scopus$Year==2000  |  scopus$Year==2001  |  scopus$Year==2002
                |scopus$Year==2003        |     scopus$Year==2004|     scopus$Year==2005,
                ]
dim(sc2000.2005)

## [1] 716  30
```

```
library(stringr)

#2000-2005
a.kw<-str_split(sc2000.2005$Author.Keywords,";                    ")
a.kw<-unlist(a.kw)
str(a.kw)

##  chr [1:2371] "" "" "Canis familiaris" "Corynosoma obtuscens" ...

tr<-head(sort(table(a.kw),decreasing                    =                    T),50)
nombres<-names(tr)
#Genero                                                                    gráfico
par(mfrow=c(1,3))
par(las=2)      #      make      label      text      perpendicular      to      axis
par(mar=c(5,14,4,2))          #          increase          y-axis          margin.

counts<-as.numeric(tr)
barplot(counts[-1],      main="2000-2005",      horiz=TRUE,      names.arg=nombres[-1],
      cex.names=1.1,      col="dodgerblue",      cex.main      =      1,      cex.axis      =      0.9)

#2006-2010
a.kw<-str_split(sc2006.2010$Author.Keywords,";                    ")
a.kw<-unlist(a.kw)
str(a.kw)

##  chr [1:5110] "Acromegaly" "Immunoassay" "OGTT" "" "" ...

tr<-head(sort(table(a.kw),decreasing                    =                    T),50)
nombres<-names(tr)
#Genero                                                                    gráfico
par(mar=c(5,10,4,2))          #          increase          y-axis          margin.

counts<-as.numeric(tr)
barplot(counts[-1],      main="2006-2010",      horiz=TRUE,      names.arg=nombres[-1],
      cex.names=1.1,      col="dodgerblue",      cex.main      =      1,      cex.axis      =      0.9)
```

```
#2011-2015
a.kw<-str_split(sc2011.2015$Author.Keywords,";                                    ")
a.kw<-unlist(a.kw)
str(a.kw)

##  chr [1:10753] "Antimicrobial resistance" "Autotransporters" ...

tr<-head(sort(table(a.kw),decreasing                    =                    T),50)
nombres<-names(tr)
#Genero                                                                    gráfico
par(mar=c(5,10,4,2))          #          increase          y-axis          margin.


counts<-as.numeric(tr)
barplot(counts[-1],     main="2011-2015",     horiz=TRUE,     names.arg=nombres[-1],
      cex.names=1.1, col="dodgerblue", cex.main = 1, cex.axis = 0.9)
```

8.2.2 Exploración de métodos de conglomerados y clasificación para la identificación de instituciones

```
# Comandos correspondientes a 4.2 Exploración de métodos de conglomerados y
clasificación para la identificación de instituciones

# 4.2. EXPLORACIÓN DE MÉTODOS DE CONGLOMERADOS Y DE CLASIFICACIÓN
PARA LA IDENTIFICACIÓN DE LAS INSTITUCIONES

#explorando datos
#recordar de no convertir el texto en factores

#
scopus_2000a2004 <- read.csv("bd/scopus_peru_medi_2000a2004_rec16_06_16.csv",stri
ngsAsFactors = FALSE, encoding = "UTF-8")
scopus_2005a2010 <- read.csv("bd/scopus_peru_medi_2005a2010_rec16_06_16.csv",stri
ngsAsFactors = FALSE, encoding = "UTF-8")
```

```
scopus_2011a2013 <- read.csv("bd/scopus_peru_medi_2011a2013_rec16_06_16.csv",stri
ngsAsFactors = FALSE, encoding = "UTF-8")
scopus_2014a2015 <- read.csv("bd/scopus_peru_medi_2014a2015_rec16_06_16.csv",stri
ngsAsFactors = FALSE, encoding = "UTF-8")
scopus<-rbind(scopus_2000a2004,scopus_2005a2010,scopus_2011a2013,scopus_2014a2
015)

############################
#creo dos data frame pare evitar el problema de factores vs carácteres
scopus_analiz_cuan<-data.frame(scopus$Year,scopus$Cited.by)
scopus_analiz_cual<-cbind(scopus$EID,scopus$Authors.with.affiliations,scopus$Affiliati
ons)
scopus_analiz_cual<-as.data.frame(scopus_analiz_cual, stringsAsFactors = FALSE)
#ahora puedo unir ambos tipos de datos
scopus_analiz<-cbind(scopus_analiz_cuan,scopus_analiz_cual)
summary(scopus_analiz)

##  scopus.Year   scopus.Cited.by       V1                 V2
## Min.   :2000   Min.   :  1.00   Length:5697        Length:5697
## 1st Qu.:2008   1st Qu.:  3.00   Class :character   Class :character
## Median :2011   Median :  7.00   Mode  :character   Mode  :character
## Mean   :2011   Mean   : 22.52
## 3rd Qu.:2014   3rd Qu.: 20.75
## Max.   :2016   Max.   :1408.00
## NA's   :1      NA's   :1339
##       V3
## Length:5697
## Class :character
## Mode  :character
##
##
##
##
```

```
####

#Escoger elementos con más de una afiliación (son los que sirven para el ARS)
#la función grepl me ayuda a encontrar si un caracter está presente en una cadena de cará
cteres, si tengo ; es un campo que tiene más de una afiliación
scopus_mult_aff<-scopus_analiz[grep(";", scopus_analiz[,5]), ]
head(scopus_mult_aff$V3)
```

```
## [1] "School of Public Health, Cayetano Heredia University, Southern Campus, Av Arm
endoriz 445, Lima 18, Peru; School of Public Health, Cayetano Heredia University, Lima,
Peru; Inst. de Investigacion Nutricional, Lima, Peru"
## [2] "Department of Pathology, Inst. Nac. de Enferm. Neoplasicas, Lima, Peru; Departm
ent of Neurosurgery, Inst. Nac. de Enferm. Neoplasicas, Lima, Peru; Departamento de Pato
logía, Inst. de Enfermedades Neoplasicas, Avenida Angamos Este 2520, San Borja (Lima 3
4), Lima, Peru"
## [3] "Inst. de Invest. de la Altura, Universidad Peruana Cayetano Heredia, Postal Office
1843, Lima, Peru; Dept. of Biol./Physiol. Sciences, Universidad Peruana Cayetano Heredia
, Postal Office 1843, Lima, Peru; Faculty of Vet. Med./Animal Sciences, Universidad Peru
ana Cayetano Heredia, Postal Office 1843, Lima, Peru"
## [4] "Inst. Med. Tropical A. von Humboldt, Universidad Peruana Cayetano Heredia, Av.
Honorio Delgado 430, Lima 100, Peru; Dept. de Enfermedades Infecciosas, Hospital Nacio
nal Cayetano Heredia, Lima, Peru; Department of International Health, Johns Hopkins Uni
versity, Bloomberg School of Public Health, Baltimore, MD 21205, United States"
## [5] "Hospital Dos de Mayo, Lima, Peru, Peru; Universidad Peruana Cayetano Heredia,
Lima, Peru, Peru; Univ. Washington Depts. Med. Ctr. A., Seattle, Wash, United States; U
WCenter for AIDS and STD, University of Washington, MS #359931, 325 9th Ave, Seattl
e, W., United States"
## [6] "Department of Dermatopathology, University Hospital of Liège, Liège, Belgium; D
epartment of Dermatology, Hospital Essalud, Cusco, Peru; Department of Dermatopatholo
gy, CHU Sart-Tilman, BE-4000 Liège, Belgium"
```

```
#separo los elementos de afiliaciones por punto y coma; esto me da una lista con la que es
muy dificil trabajar, por eso pongo la opción unlist
af_variat<-unlist(strsplit(scopus_mult_aff$V3, '\\; '))
head(af_variat)
```

```
## [1] "School of Public Health, Cayetano Heredia University, Southern Campus, Av Arm
endoriz 445, Lima 18, Peru"
## [2] "School of Public Health, Cayetano Heredia University, Lima, Peru"
## [3] "Inst. de Investigacion Nutricional, Lima, Peru"
## [4] "Department of Pathology, Inst. Nac. de Enferm. Neoplasicas, Lima, Peru"
## [5] "Department of Neurosurgery, Inst. Nac. de Enferm. Neoplasicas, Lima, Peru"
## [6] "Departamento de Patología, Inst. de Enfermedades Neoplasicas, Avenida Angamos
Este 2520, San Borja (Lima 34), Lima, Peru"
```

```
#Selecciono los elementos únicos
af_variat <- unique(af_variat)


#Procesamiento adicional de la columna que contiene las diferentes afiliaciones
library(stringr)
comas_por_pais<-str_count(af_variat, ",")
table(comas_por_pais)
```

```
## comas_por_pais
##    0    1    2    3    4    5    6    7    8    9   10   11
##  155 2013 6237 8169 6032 2789  862  211   45    8    2    1
```

```
#Esto me indica que hay filas que no tienen comas por lo que una separación basada en co
mas no es suficiente
af_variat_df<-data.frame(af_variat,comas_por_pais)
#utilizo gsub para obtener los países
af_variat_df$country<-gsub("^.*, ","",af_variat_df$af_variat)
#Corrigo casos en los que el país no se ha señalado correctamente
library(stringr)
af_variat_df$cleaned<-str_replace_all(af_variat_df$country, "Can. Prog. Genom. and Glo
bal Hlth.", "Canada")
af_variat_df$cleaned<-str_replace(af_variat_df$cleaned,"\\(","") #quito las comas porque
no funcionan bien con str_rplace
af_variat_df$cleaned<-str_replace(af_variat_df$cleaned,"\\)","") #quito las comas porque
no funcionan bien con str_rplace
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Centre for Public Law", "Fa
```

```
lta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "ã", "a")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Faculdade de Medicina de Sao do Rio Preto FAMERP", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Instituto de Ciencias Neurológicas", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Médico Patologista do HB de S.J.R.Preto", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Médico Hematologista e Hemoterapeuta do HB de S.J.R.Preto", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Movement Disord. U.", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Unidade de TMO HB-FUNFARME / CINTRANS", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Sections of Affective Disorders and Old Age Psychiatry", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "World Pychiatric Association", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "World Health Organization WHO", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "WHO", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "University of Yaounde 1", "Cameroon")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Université Toulouse 3", "France")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Universidad Peruana Cayetano Heredia", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Universidad Peruna Cayetano Heredia", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Universidad National Mayor de San Marcos", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Universidad Nacional Mayor de San Marcos", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "World Bank", "Falta") #inte
```

```
rnacional, no tengo datos sobre sede
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Wisconsin", "United States
")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Vellore", "India")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Universidad Nacional Fede
rico Villarreal", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Universidad de Concepción
", "Chile")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "UMR 152 Pharma-DEV", "
France")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "U.K.", "United Kingdom")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Trujillo", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Tropicales y Dermatológica
s", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Trasplantes de Órganos Y
Tejidos de EsSalud", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Trasplantes de Órganos Y
Tejidos", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Texas", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Sociedade Brasileira de He
matologia e Hemoterapia", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Sociedad Peruana de Obstet
ricia Y Ginecología", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Sociedad Científica Huacha
na de Estudiantes de Medicina", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "School of Medicine", "Falta
")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Sao Paulo", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Sanliurfa", "Turkey")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Seattle", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Sanatorio Quintar", "Argent
ina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Rio de Janeiro-RJ", "Brazil
")
```

```
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "República Dominicana", "Dominican Republic")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Pucallpa", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Pontificia Universidad Católica del Perú PUCP", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Policlinico Regional J. D. Perón", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Perth", "Australia")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "PA", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Nutrition", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Novasalud EPS", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "North Carolina", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "NM 87131", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "San Salvador", "El Salvador")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "New Hampshire", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "New Delhi", "India")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "NC", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Neonatología", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Math. and Info.", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Manila", "Philippines")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "M. C. Carranza Hospital Dr. Pedro Moguillansky", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Libyan Arab Jamahiriya", "Libya")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Laboratorio de Biología Molecular y Biotecnología - Instituto Nacional de Salud", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Laboratorio Clínica Radiológica del Sur S.A.", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Laboratorio Central de Salud Pública", "Peru")
```

```
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "La Habana", "Cuba")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "IPK", "Cuba")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Instituto de Medicina Tropical Pedro Kourí IPK", "Cuba")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "International Society for Affective Disorders", "United Kingdom")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Instituto Peruano De Parasitología Clínicay Experimental", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Instituto Nacional de Salud del Niño", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Instituto Nacional De Ciencias Neurológicas", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Institut de Recherche Pour le Développement", "Falta") #es internacional, Perú y Francia
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Indianapolis", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "IN 46202", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Illinois", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "International Society for Affective Disorders", "United Kingdom")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Htal. Zonal de Chos Malal Dr. Gregorio Alvarez", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital SAMIC - El Dorado", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Regional Pasteur", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Regional Dr. Sanguinetti", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Provincial Héroes de Malvinas", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Provincial Dr. E. Borzani", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Natalio Burd", "Argentina")
```

```
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Nacional Sur Este EsSalud del Cusco", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Materno Infantil A. Diego", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Luis Calvo Mackenna", "Chile")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Justo José de Urquiza", "Argentina")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Infantil Dr. H. Notti", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital G. Sayago", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Diego Paroissien", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital del Niño Jesús", "Spain")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital del Milagro", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital de Infecciosas Muñiz", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital de Enf. Infecciosas Dr. J. Lencinas", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital de Concarán", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital de Clínicas de Assunçao", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Central", "Falta")
#información insuficiente
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Centenario", "Argentina")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Perú", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Bogotá", "Colombia")
```

```
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "California", "United States"
)
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Nacional Guillerm
o Almenara", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Gonzaga", "Peru")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital Vall d'Hebron", "S
pain")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Instituto Nacional de Enfer
medades Neoplasticas I.N.E.N", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Lanús", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Lima", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Louisiana", "United States"
)
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Maryland", "United States"
)
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Massachusetts", "United St
ates")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "MD", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Peru Irigoyen", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Peru.", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "10 Andar", "Brazil") #En es
te caso
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Albuquerque", "United Stat
es")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Albuquerque NM", "United
States")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Advsy. Committee on Healt
h Research", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "American Foundation for A
IDS Research", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Antalaya", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "APO AP 96205-52", "Falta
```

```
")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "APO AP 96546", "Thailand
")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Área de Concentraçao Endo
crinologia UFRJ", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Área de Concentraçao Epid
emiologia UERJ", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Assistance and Research in
Infancy", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Barranquilla", "Colombia")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Caracas University Hospita
l", "Venezuela")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Hospital 4 de Junio", "Arge
ntina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "H.I.G.A. V. López Y Plane
s", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "H.I.G.A Prof. Dr. Rodolfo
Rossi", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "H.I.G.A Dr. Luis Güemes",
"Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "H.I.G.A Dr. A. Oñativia", "
Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Group Health's Behavioral
Health Service", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "GA 30333", "United States
")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Fundación Horizonte", "Bol
ivia")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Federación Latinoamerican
a de Sociedades de Obstetricia Y Ginecología", "Falta") #internacional
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Family Health International
", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Facultad de Medicina Univ
```

```
ersidad del Desarrollo", "Chile")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Instituto de Medicina Tropical Pedro Kourí Cuba", "Cuba")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Faculdade de Medicina UFMG", "Brazil")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "ETS y VIH/SIDA", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "entre escaleras 5-7", "Spain")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Dutch Royal Tropical Institute", "Netherlands")
#af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Dominica", "Dominican Republic") No puedo usarlo porque reemplaza valores en el mismo dominican republic
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "División Laboratorio Central", "Falta")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Department of Pediatrics", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Department of Microbiology", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Department of Haematology", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Department of Clinical Tropical Medicine", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Department of Biostatistics", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Departamentos de Urologia Oncologica y Anatomia Patologica Oncologica", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Departamento de Tocoginecologia/CAISM /FCM/UNICAMP", "Brazil")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Departamento de Cardiologia Pediátrica", "Falta")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "D.C", "United States")
```

```
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "CRAI-Norte. CUCAIBA", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Consultorios Clínica Mayo", "Falta") #Es internacional, no se sabe a qué sede se refiere
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Consortium for Biomedical Research in Epidemiology and Public Health CIBERESP", "Spain")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Colegio Bertolt Brecht", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Clínica San Borja", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Child Health and United States Research Initiative of the Global Forum for Health Research", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Centro para el Desarrollo Infantil Learn and Play", "Peru")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Centro de Informacion Y Educacion para la Prevencion del Abuso de Drogas - CEDRO", "Peru")

af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Centro de Especialidades Médico Ambulatorias CEMAR", "Argentina")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Centre Nationalde la Recherche Scientifique", "France")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "United States NM", "United States")
af_variat_df$cleaned<-str_replace_all(af_variat_df$cleaned, "Mexico DF", "Mexico")

#tengo un Sucre de Bolivia y otro de Colombia
af_variat_df$cleaned[2258]<-"Bolivia"
af_variat_df$cleaned[2263]<-"Colombia"

#En Georgia tengo que ver uno por uno si es país o Estado de Estados Unidos
af_variat_df$cleaned[8506]<-"United States"
af_variat_df$cleaned[20349]<-"United States"
af_variat_df$cleaned[21486]<-"United States"
```

```
#En Dominica tengo que reemplazarlo por Dominican Republic
af_variat_df$cleaned[20769]<-"Dominican Republic"
```

- *FIGURA 15: Países de las afiliaciones institucionales.*

```
#################################################################
#Figura 15
#########################################################
#Gráfico
tr<-head(sort(table(af_variat_df$cleaned),decreasing = T),30)

nombres<-names(tr)
#Genero gráfico
par(mfrow=c(1,1))
par(las=2) # make label text perpendicular to axis
par(mar=c(5,10,4,2)) # increase y-axis margin.

counts<-as.numeric(tr)
barplot(counts, main="Paises con más afiliaciones en la muestra", horiz=TRUE, names.arg=nombres,
        cex.names=1, col="dodgerblue", cex.main = 1, cex.axis = 0.9, panel.first=grid())
```

```
#########################################################
#Remover signos diacríticos
unwanted_array = list(    'Š'='S', 'š'='s', 'Ž'='Z', 'ž'='z', 'À'='A', 'Á'='A', 'Â'='A', 'Ã'='A', 'Ä'='A', 'Å'='A', 'Æ'='A', 'Ç'='C', 'È'='E', 'É'='E',
                    'Ê'='E', 'Ë'='E', 'Ì'='I', 'Í'='I', 'Î'='I', 'Ï'='I', 'Ñ'='N', 'Ò'='O', 'Ó'='O', 'Ô'='O', 'Õ'='O', 'Ö'='O', 'Ø'='O', 'Ù'='U',
                    'Ú'='U', 'Û'='U', 'Ü'='U', 'Ý'='Y', 'Þ'='B', 'ß'='Ss', 'à'='a', 'á'='a', 'â'='a', 'ã'='a', 'ä'='a', 'å'='a', 'æ'='a', 'ç'='c',
                    'è'='e', 'é'='e', 'ê'='e', 'ë'='e', 'ì'='i', 'í'='i', 'î'='i', 'ï'='i', 'ð'='o', 'ñ'='n', 'ò'='o', 'ó'='o', 'ô'='o', 'õ'='o',
```

```
                'ö'='o', 'ø'='o', 'ù'='u', 'ú'='u', 'û'='u', 'ý'='y', 'ý'='y', 'þ'='b', 'ÿ'='y' )
af_variat_df$af_clean<-chartr(paste(names(unwanted_array), collapse=''), paste(unwanted_array, collapse=''), af_variat_df$af_variat)



#Países en los datos
sort(table(af_variat_df$cleaned),decreasing = TRUE)

##
##                Peru United States
##                6236          5882
##              Brazil United Kingdom
##                1217          1014
##               Spain     Argentina
##                 991           774
##              Mexico        Canada
##                 648           593
##              France      Colombia
##                 527           505
##               Italy         Chile
##                 499           487
##               India       Germany
##                 400           380
##           Australia   Switzerland
##                 328           323
##             Belgium         Japan
##                 283           239
##               China     Venezuela
##                 237           232
##             Ecuador  South Africa
##                 226           222
##         Netherlands          Cuba
##                 190           143
##             Bolivia       Uruguay
```

```
##                 142                 137
##            Portugal              Sweden
##                 134                 133
##            Thailand Russian Federation
##                 114                  99
##       South Korea                Denmark
##                  97                  91
##             Nigeria          Costa Rica
##                  88                  87
##              Turkey              Israel
##                  81                  79
##              Norway              Poland
##                  79                  78
##             Austria            Paraguay
##                  74                  74
##               Kenya               Egypt
##                  69                  66
##         Philippines              Panama
##                  65                  61
##            Malaysia             Finland
##                  60                  57
##              Taiwan      Czech Republic
##                  57                  56
##              Greece            Pakistan
##                  56                  56
## Dominican Republic           Guatemala
##                  55                  55
##           Singapore                Iran
##                  54                  51
##              Uganda         New Zealand
##                  49                  45
##           Indonesia            Viet Nam
##                  44                  44
##          Bangladesh         Puerto Rico
```

##	42	42
##	Saudi Arabia	Romania
##	42	41
##	Ireland	Slovakia
##	39	36
##	El Salvador	Honduras
##	34	34
##	Ethiopia	Malawi
##	33	33
##	Tanzania	Ghana
##	33	31
##	Georgia	Hong Kong
##	29	29
##	Cameroon	Lebanon
##	27	27
##	Serbia	Bulgaria
##	27	26
##	Latvia	Morocco
##	25	24
##	Nepal	Nicaragua
##	24	24
##	Estonia	Haiti
##	23	23
##	Ukraine	Falta
##	22	21
##	Lithuania	Sri Lanka
##	21	21
##	Croatia	Tunisia
##	20	20
##	Cambodia	Hungary
##	19	19
##	Cote d'Ivoire	Rwanda
##	17	17
##	Zambia	Zimbabwe

```
##                 17                 17
##             Jordan            Algeria
##                 14                 13
##               Iraq         Macedonia
##                 13                 12
##     French Guiana              Sudan
##                 11                 11
##              Congo             Cyprus
##                 10                 10
##               Fiji            Iceland
##                 10                 10
##            Jamaica               Mali
##                 10                 10
##            Senegal             Gambia
##                 10                  9
##  Papua New Guinea              Qatar
##                  9                  9
##         Azerbaijan Bosnia and Herzegovina
##                  8                  8
##           Botswana         Mozambique
##                  8                  8
##           Slovenia United Arab Emirates
##                  8                  8
##              Benin         Luxembourg
##                  7                  7
##           Barbados             Belize
##                  6                  6
##      Guinea-Bissau         Madagascar
##                  6                  6
##           Palestine       Sierra Leone
##                  6                  6
##  Syrian Arab Republic        Afghanistan
##                  6                  5
##             Guinea         Kazakhstan
```

```
##                                   5                                   5
##                             Moldova                             Albania
##                                   5                                   4
##                        Burkina Faso                              Kuwait
##                                   4                                   4
##                           Kyrgyzstan                            Mongolia
##                                   4                                   4
##                                Oman                             Bahamas
##                                   4                                   3
##                             Belarus                             Myanmar
##                                   3                                   3
##                               Niger                                Togo
##                                   3                                   3
##                              Angola                             Armenia
##                                   2                                   2
##                             Bahrain                             Burundi
##                                   2                                   2
##          Central African Republic                              Guyana
##                                   2                                   2
##                               Macau                           Mauritius
##                                   2                                   2
##                             Namibia                 Netherlands Antilles
##                                   2                                   2
##                       New Caledonia                         Saint Lucia
##                                   2                                   2
##                               Samoa                          Seychelles
##                                   2                                   2
##                           Swaziland                 Trinidad and Tobago
##                                   2                                   2
##                          Uzbekistan                              Bhutan
##                                   2                                   1
##          Democratic Republic Congo                   Equatorial Guinea
##                                   1                                   1
## Federated States of Micronesia                               Greenland
```

```
##                                   1                                   1
##                          Guadeloupe                                Laos
##                                   1                                   1
##                             Lesotho                               Libya
##                                   1                                   1
##                               Malta                          Montenegro
##                                   1                                   1
##                 Solomon Islands                                Vanuatu
##                                   1                                   1
##                                                                   Yemen
##                   1

######################
#Guardar                                                          países
af_paises<-af_variat_df[,2:4]
############################33

af_variat_Peru <- paste((af_variat_df[ which(af_variat_df$cleaned=='Peru'), ])[,5])

#Lo guardo para usarlo después
#save(af_variat_Peru,file="bd/af_variat_Peru.Rda")

#Guardo tmb los que no son de Perú, para usarlo después
af_variat_no_Peru <- paste((af_variat_df[ which(af_variat_df$cleaned!='Peru'), ])[,5])
#save(af_variat_no_Peru,file="bd/af_variat_no_Peru.Rda")

#de la anterior manera evito que se convierta en factor
library(tm)

## Loading required package: NLP

#install.packages("SnowballC")
library(SnowballC)

#Creo una matriz de documentos-textos
```

```
corpus <- Corpus(VectorSource(af_variat_Peru))
corpus<-tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, function(x) removeWords(x, stopwords("spanish")))
corpus <- tm_map(corpus, stemDocument, language = "spanish")


####################Prueba1
m <- as.matrix(dtm)
#shorten rownames for display purposes>
rownames(m) <- paste(substring(rownames(m),1,3),rep("..",nrow(m)),substring(rownames(m), nchar(rownames(m))-12,nchar(rownames(m))-4))
d<-dist(m)


kfit <- kmeans(d, 100, nstart=1) #demora en procesar
kfit.400<-kmeans(d, 400, nstart=1) #demora en procesar

kclust<-cbind(kfit$cluster,kfit.400$cluster,af_variat_Peru)

#Resultados
dev.off()

## null device
##  1

par(mfrow=c(2,1))
```

- *FIGURA 17: Resultados de la utilización del método de k-medias para la identificación de instituciones*

```
###########

#Figura 17

###############
```

```
#Tamaño de los conglomerados k
af_clust<-cbind(af_variat_Peru,kfit$cluster)
str(af_clust)
af_clust<-data.frame(af_clust, stringsAsFactors = F)
sorted_af_clust_Peru<-af_clust[order(kfit$cluster),]
#Elementos del conglomerado 1
sorted_af_clust_Peru[sorted_af_clust_Peru$V2==1,1]
par(mfrow=c(2,1))
plot(table(table(sorted_af_clust_Peru[,2])), ylab="Frecuencia",
    xlab="Tamaño de los conglomerados", xaxt='n', xlim=c(0,160),
    main="100 conglomerados", ylim=c(0,35), col="red") #max x 150, max y=6
Axis(side=1, labels=T)
#Agrego con hclust 400
af_clust<-cbind(af_variat_Peru,kfit.400$cluster)
af_clust<-data.frame(af_clust, stringsAsFactors = F)
sorted_af_clust_Peru<-af_clust[order(kfit.400$cluster),]
#Elementos del conglomerado 1
sorted_af_clust_Peru[sorted_af_clust_Peru$V2==1,1]
plot(table(table(sorted_af_clust_Peru[,2])), ylab="Frecuencia",
    xlab="Tamaño de los conglomerados", xaxt='n', xlim=c(0,160),
    main="400 conglomerados", ylim=c(0,35), col="red")
Axis(side=1, labels=T)
```

##############

```
#Escoger un cluster al azar

azar.400<-round(runif(1,1,400),digits = 0)

azar.100<-round(runif(1,1,100),digits = 0)

#Elementos de un cluster al azar

kclust[kclust$V1==azar.100,]

kclust[kclust$V2==azar.400,]

########################################

#Usando stringdist,

#################################

#install.packages("stringdist")

library(stringdist)

d <- stringdistmatrix(tolower(af_variat_Peru), tolower(af_variat_Peru), method = "qgram", q=4)

###########
```

- *FIGURA 16: Resultados de la utilización del método de n-gramas para la identificación de instituciones*

```
#Figura 16

################

cl<-hclust(as.dist(d))

cl.400<-cutree(cl, 400)

cl.100<-cutree(cl,100)

############################3

#Tamaño de los conglomerados stringdist + grafico

af_clust<-cbind(af_variat_Peru,cl.100)
```

```
af_clust<-data.frame(af_clust, stringsAsFactors = F)

sorted_af_clust_Peru<-af_clust[order(cl.100),]

par(mfrow=c(2,1))

plot(table(table(sorted_af_clust_Peru[,2])), ylab="Frecuencia", xlab="Tamaño de los conglomerados", xaxt='n', xlim=c(0,1260), main="100 conglomerados", ylim=c(0,75), col="red")

Axis(side=1, labels=T)

#Escogo uno al azar

sorted_af_clust_Peru[sorted_af_clust_Peru$cl.100=="11",][,1]

#Agrego con hclust 400

af_clust<-cbind(af_variat_Peru,cl.400)

af_clust<-data.frame(af_clust, stringsAsFactors = F)

sorted_af_clust_Peru<-af_clust[order(cl.400),]

plot(table(table(sorted_af_clust_Peru[,2])), ylab="Frecuencia", xlab="Tamaño de los conglomerados", xaxt='n', xlim=c(0,1260), main="400 conglomerados", ylim=c(0,75), col="red")

Axis(side=1, labels=T)


#Escogo uno al azar

sorted_af_clust_Peru[sorted_af_clust_Peru$cl.400=="2",][,1]

########
```

- *CUADRO 3: Resultados de la utilización de los métodos de conglomerados supervisados para la identificación de las instituciones*

```
#Cuadro 3: Resultados de la utilización de los métodos de conglomerados supervisados para la identificación de las instituciones

########

#Exploración con trainset

af_train_antiguo <- read.csv2("bd/af_train_antiguo.csv", header=FALSE, stringsAsFactors = FALSE)

af_train_antiguo<- af_train_antiguo[sample(1:dim(af_train_antiguo)[1],round(((dim(af_train_antiguo)[1])/10)*10,0)),]

library(RTextTools)

library(e1071)

# build document term matrix

matrix= create_matrix(af_train_antiguo[,1], language="english", removeStopwords=FALSE, removeNumbers=FALSE, stemWords=TRUE)

#Probando con varios métodos de machine learning

# build the data to specify response variable, training set, testing set.

#First, to specify our data:

container = create_container(matrix, as.numeric(as.factor(af_train_antiguo[,2])), trainSize=1:(dim(af_train_antiguo)[1]/2), testSize=(dim(af_train_antiguo)[1]/2+1):dim(af_train_antiguo)[1],virgin=FALSE) #tuve que modificar a mitad/mitad porque no tenía suficiente memoria

#Second, to train the model with multiple machine learning algorithms:

models = train_models(container, algorithms=c("SVM","RF","BAGGING"))

#MAXENT no soporta más de 255 etiquetas

#SMV no me bota error

#RF no me bota error

#BAGGING no me bota error
```

```
#TREE no soporta más de 32 niveles de respuesta

#Now, we can classify the testing set using the trained models.

results = classify_models(container, models)

#How about the accuracy?

# accuracy table

head(table(as.numeric(as.factor(af_train_antiguo[tr_sample,                2])),
results[,"FORESTS_LABEL"]))

# recall accuracy

recall_accuracy(as.numeric(as.factor(af_train_antiguo[tr_sample,           2])),
results[,"FORESTS_LABEL"])

recall_accuracy(as.numeric(as.factor(af_train_antiguo[tr_sample,           2])),
results[,"BAGGING_LABEL"])

recall_accuracy(as.numeric(as.factor(af_train_antiguo[tr_sample,           2])),
results[,"SVM_LABEL"])

#To summarize the results (especially the validity) in a formal way:

# model summary

analytics = create_analytics(container, results)

summary(analytics)

head(analytics@document_summary)

analytics@ensemble_summary

#To cross validate the results:

N=4

#set.seed(2014)

cross_validate(container,N,"SVM")

system.time(cross_validate(container,N,"SVM"))

cross_validate(container,N,"RF")
```

```
system.time(cross_validate(container,N,"RF"))

cross_validate(container,N,"BAGGING")

system.time(cross_validate(container,N,"BAGGING"))
```

- *FIGURA 18. Resultados de la utilización del método de k-vecinos más cercanos sobre los datos de entrenamiento para la identificación de instituciones*

```
##########

#Figura 18: Resultados de la utilización del método de k-vecinos más cercanos sobre los datos de entrenamiento para la identificación de instituciones

############################

#si partiese desde 0, sólo para marco teórico

validacion<-af_variat_Peru

af_train_antiguo <- read.csv2("bd/af_train_peru_modf.csv", stringsAsFactors = FALSE, encoding = "ISO-8859-1")

af_train_antiguo <- af_train_antiguo[,1:2]

entrenamiento<-af_train_antiguo

names(entrenamiento)<-c("X1","X2")

########################

dataknn<-rbind(entrenamiento,cbind(X1=validacion,X2=rep("val",length(validacion))))

##install.packages("RTextTools")

##install.packages("ipred")

library(RTextTools)

#Creo la matriz

matrixknn= create_matrix(dataknn, language="english",

                removeStopwords=FALSE, removeNumbers=FALSE,
```

```
stemWords=TRUE) #probar tmb con stemWords=FALSE/TRUE

#la convierto en una matriz corriente

#me están sobrando columnas de información que no se utilizará pero por ahora no me preocupo por eso

matknn = as.matrix(matrixknn)

library(class)

kn10<-
paste(knn(matknn[1:dim(entrenamiento)[1],],matknn[(dim(entrenamiento)[1]+1):(dim(entrenamiento)[1]+length(validacion)),],entrenamiento[,2],k=10,l=1))

kn5<-
paste(knn(matknn[1:dim(entrenamiento)[1],],matknn[(dim(entrenamiento)[1]+1):(dim(entrenamiento)[1]+length(validacion)),],entrenamiento[,2],k=5,l=1))

kn3<-
paste(knn(matknn[1:dim(entrenamiento)[1],],matknn[(dim(entrenamiento)[1]+1):(dim(entrenamiento)[1]+length(validacion)),],entrenamiento[,2],k=3,l=1))

kn1<-
paste(knn(matknn[1:dim(entrenamiento)[1],],matknn[(dim(entrenamiento)[1]+1):(dim(entrenamiento)[1]+length(validacion)),],entrenamiento[,2],k=1,l=1))

#le doy un vistazo a los datos

k_predictions<-
data.frame(cbind(OR=paste(validacion),K1=paste(kn1)),K3=paste(kn3),K5=paste(kn5),K10=paste(kn10),comp1_5=(kn5 == kn3 & kn3 == kn1),stringsAsFactors = FALSE)

#save(k_predictions,file="bd/k_predictions.Rda")

load("bd/k_predictions.Rda")

colnames(k_predictions)<-c("OR","K1","K3","K5","K10","comp1_5")

head(sort(table(k_predictions$K1),decreasing = TRUE))

head(sort(table(k_predictions$K3),decreasing = TRUE))
```

```
head(sort(table(k_predictions$K5),decreasing = TRUE))

head(sort(table(k_predictions$K10),decreasing = TRUE))

################

#Gráfico

#k1

af_clust<-cbind(k_predictions$OR,k_predictions$K1)

str(af_clust)

af_clust<-data.frame(af_clust, stringsAsFactors = F)

sorted_af_clust_Peru<-af_clust[order(k_predictions$K1),]

sorted_af_clust_Peru[sorted_af_clust_Peru$X2==1,1]

par(mfrow=c(2,2))

plot(table(table(sorted_af_clust_Peru[,2])), ylab="Frecuencia",

    xlab="Tamaño de los conglomerados", xaxt='n', xlim=c(0,1900),

    main="Tamaño de los conglomerados con k=1", ylim=c(0,180),

    col=ifelse(names(table(table(sorted_af_clust_Peru[,2])))=="401", "red", "darkgray")

    ,  type="p",pch=ifelse(names(table(table(sorted_af_clust_Peru[,2])))=="401",  1,  16))
#max x 150, max y=6, #pch1 o

Axis(side=1, labels=T)

#k3

af_clust<-cbind(k_predictions$OR,k_predictions$K3)

str(af_clust)

af_clust<-data.frame(af_clust, stringsAsFactors = F)

sorted_af_clust_Peru<-af_clust[order(k_predictions$K3),]

sorted_af_clust_Peru[sorted_af_clust_Peru$X2==1,1]

plot(table(table(sorted_af_clust_Peru[,2])), ylab="Frecuencia",
```

```
    xlab="Tamaño de los conglomerados", xaxt='n', xlim=c(0,1900),
    main="Tamaño de los conglomerados con k=3", ylim=c(0,180),
    col=ifelse(names(table(table(sorted_af_clust_Peru[,2])))=="1129", "red", "darkgray")
    , type="p",pch=ifelse(names(table(table(sorted_af_clust_Peru[,2])))=="1129", 1, 16))
#max x 150, max y=6, #pch1 o
Axis(side=1, labels=T)
#k5
af_clust<-cbind(k_predictions$OR,k_predictions$K5)
str(af_clust)
af_clust<-data.frame(af_clust, stringsAsFactors = F)
sorted_af_clust_Peru<-af_clust[order(k_predictions$K5),]
sorted_af_clust_Peru[sorted_af_clust_Peru$X2==1,1]
plot(table(table(sorted_af_clust_Peru[,2])), ylab="Frecuencia",
    xlab="Tamaño de los conglomerados", xaxt='n', xlim=c(0,1900),
    main="Tamaño de los conglomerados con k=5", ylim=c(0,180),
    col=ifelse(names(table(table(sorted_af_clust_Peru[,2])))=="1467", "red", "darkgray")
    , type="p",pch=ifelse(names(table(table(sorted_af_clust_Peru[,2])))=="1467", 1, 16))
#max x 150, max y=6, #pch1 o
Axis(side=1, labels=T)
#k10
af_clust<-cbind(k_predictions$OR,k_predictions$K10)
str(af_clust)
af_clust<-data.frame(af_clust, stringsAsFactors = F)
sorted_af_clust_Peru<-af_clust[order(k_predictions$K10),]
sorted_af_clust_Peru[sorted_af_clust_Peru$X2==1,1]
```

```
plot(table(table(sorted_af_clust_Peru[,2])), ylab="Frecuencia",
    xlab="Tamaño de los conglomerados", xaxt='n', xlim=c(0,1900),
    main="Tamaño de los conglomerados con k=10", ylim=c(0,180),
    col=ifelse(names(table(table(sorted_af_clust_Peru[,2])))=="1872", "red", "darkgray")
    , type="p",pch=ifelse(names(table(table(sorted_af_clust_Peru[,2])))=="1872", 1, 16))
#max x 150, max y=6, #pch1 o
Axis(side=1, labels=T)
k_predictions_raw<-k_predictions #Ojo este archivo que estoy guardando incluye la predicción
```

8.2.3 Clasificación con k-vecinos más cercanos

```
# Comandos correspondientes a 4.3 Clasificación con k-vecinos más cercanos
#1paso, aunque la clasificación se hizo principalmente con k-vecinos más cercanos, en un principio se trabajó con máquina de soporte vectorial
library(RTextTools)
library(e1071)
#Datos a utilizar
af_train_antiguo <- read.csv2("bd/af_train_peru_modf.csv", stringsAsFactors = FALSE, encoding = "ISO-8859-1") #datos de entrenamiento
af_train_antiguo <- af_train_antiguo[,1:3]
load("bd/ af_variat_Peru.Rda")
new_valid_Peru<-af_variat_Peru
new_train_Peru<-af_train_antiguo
new_train_Peru_big<-af_train_antiguo
colnames(new_train_Peru_big) <- c("X1", "X2") #para rbind necesito que los nombres de las columnas coincidan
```

```
new_train_Peru_big<-rbind(new_train_Peru_big,new_train_Peru[,1:2])

entrenamiento<-new_train_Peru_big

validacion<-new_valid_Peru

#####

dtMatrix <- create_matrix(entrenamiento["X1"], language="english",

                removeStopwords=FALSE, removeNumbers=FALSE,
removePunctuation=TRUE,

                stemWords=FALSE, toLower = TRUE)

# Configure the training data

container <- create_container(dtMatrix, entrenamiento$X2,
trainSize=1:dim(entrenamiento)[1], virgin=FALSE)

# train a SVM Model

model <- train_model(container, "SVM", kernel="linear", cost=1)

predMatrix <- create_matrix(validacion, originalMatrix=dtMatrix)

#Hay un problema con la función (paquete tm), si sale error procedemos con el siguiente
paso

trace("create_matrix",edit=T)

#y cambiamos la "A" por una "a" en "Acronym"

#probamos la función nuevamente


# create the corresponding container

predSize = length(validacion);

predictionContainer <- create_container(predMatrix, labels=rep(0,predSize),
testSize=1:predSize, virgin=FALSE)

# predict

results <- classify_model(predictionContainer, model)
```

#Junto los originales con los valores predichos

SC_CL_SVM1<-
data.frame(cbind(paste(validacion),paste(results[,1]),results[,2]),stringsAsFactors = FALSE)

sort(table(SC_CL_SVM1[,2]),decreasing = TRUE)

#probar resultados

ordenarxprob<-SC_CL_SVM1[order(results[,2],results[,1],decreasing=TRUE),]

#guardar

#save(ordenarxprob,file="bd/ordenarxprob")

#revisar resultados, agregar cuarta columna para marcar como correctos los resultados que están bien, ejemplo:

ordenarxprob$X2[69]<-"PU_SAL_UNCH_,PR_UNI_UPCH_," *#Modifico clasificación*

ordenarxprob$X4[69]<-"ok" *#Agrego información sobre que está modificada y en orden*

#Ya que la información está ordenada por la probabilidad con la que un elemento pertenece a la categoría asignada, esto puede hacerse fácilmente para los elementos con mayor probabilidad de asignación correcta.

#Una vez revisado un número importante de clasificaciones,, se agrega esta información a los datos de entrenamiento.

#Por ejemplo:

agregaratrain_Peru <- ordenarxprob[ordenarxprob$X4 == "mod",][,1:2]

af_revisadas_Peru<-rbind(ordenarxprob[1:343,1:2],entrenamiento[,1:2])

#De esa manera se obtiene nuevos datos que pueden agregarse a los datos de entrenamiento y puede precisarse el modelo

############################

#Para asignar la clasificación con el método de K-vecinso más cercanos se sigue este modelo

```
################################

dataknn<-rbind(entrenamiento,cbind(X1=validacion,X2=rep("val",length(validacion))))

library(RTextTools)

#Creo la matriz

matrixknn= create_matrix(dataknn, language="english",

                removeStopwords=FALSE, removeNumbers=FALSE,

                stemWords=TRUE) #probar tmb con stemWords=FALSE/TRUE

#la convierto en una matriz corriente

#me están sobrando columnas de información que no se utilizará pero por ahora no me preocupo por eso

matknn = as.matrix(matrixknn)

library(class)

kn10<-
paste(knn(matknn[1:dim(entrenamiento)[1],],matknn[(dim(entrenamiento)[1]+1):(dim(entrenamiento)[1]+length(validacion)),],entrenamiento[,2],k=10,l=1))

kn5<-
paste(knn(matknn[1:dim(entrenamiento)[1],],matknn[(dim(entrenamiento)[1]+1):(dim(entrenamiento)[1]+length(validacion)),],entrenamiento[,2],k=5,l=1))

kn3<-
paste(knn(matknn[1:dim(entrenamiento)[1],],matknn[(dim(entrenamiento)[1]+1):(dim(entrenamiento)[1]+length(validacion)),],entrenamiento[,2],k=3,l=1))

kn1<-
paste(knn(matknn[1:dim(entrenamiento)[1],],matknn[(dim(entrenamiento)[1]+1):(dim(entrenamiento)[1]+length(validacion)),],entrenamiento[,2],k=1,l=1))

#le doy un vistazo a los datos
```

```
k_predictions<-
data.frame(cbind(OR=paste(validacion),K1=paste(kn1)),K3=paste(kn3),K5=paste(kn5),K10=paste(kn10),comp1_5=(kn5 == kn3 & kn3 == kn1),stringsAsFactors = FALSE)

colnames(k_predictions)<-c("OR","K1","K3","K5","K10","comp1_5")

head(sort(table(k_predictions$K1),decreasing = TRUE))

################

#cuando la fila correspondiente a K1 y la fila correspondiente a K10 son iguales probablemente la asignación es correcta

#para revisar en bloque una institución que tiene muchas ocurrencias puedo seleccionarla, por ejemplo selecciono  INS,

(k_predictions[k_predictions$K1 == "PU_IPI_INS__," & k_predictions$comp1_5 == "TRUE", ])[1:20,1]

indices<-c(which(k_predictions$K1 == "PU_IPI_INS__," & k_predictions$comp1_5 == "TRUE"))

#repito el proceso que realizé en máquinas de soporte vectorial, aumentando los datos de validación revisados a los datos de entrenamiento para precisar la clasificación.

#Para una revisión final, la utilización del paquete stringdist es una buena opción.


################

#Revisión con stringdist.

load("bd/dgr9.Rda") #Es el resultado de la clasificación y revision previa.

length(table(dgr9$X2))

##857

library(stringdist)

d <- stringdistmatrix(tolower(dgr9$X1), tolower(dgr9$X1), method = "qgram", q=4)

#El anterior proceso es BIEN largo, no tiene errores en Windows pero por temas de codificación resulta con errores en Mac
```

```
cl<-hclust(as.dist(d))

#Realizo una clasificación en 850 grupos

cl.850<-cutree(cl, 850)

dgr9<-cbind(dgr9,cl.850)

str(dgr9)

#ordenar por columnas

dgr9_sorted<-dgr9[with(dgr9, order(X2,cl.850)), ]

#Cuando pone solo ciudad y pais lo borro, por lo cual le asigno el código  RM

#Reemplazar valores, por ejemplo

dgr9_sorted[1:20,c(1:2,4)]

dgr9_sorted[1:4,2]<-c("FI_FI__001__,",
"FI_FI__002__,","FI_FI__003__,","FI_FI__004__,")

dgr9_sorted[5:6,2]<-c("FI_FI__005__,")

#Cuando la clasificación es correcta hago como en el ejemplo a continuación

dgr9_sorted[1:10,5]<-"ok"

#Como resultado de la revisión manual, en la etapa final cada clasificación debería estar asignada correctamente.
```

8.2.4 Análisis exploratorio del grafo

```
# Comandos correspondientes a 4.4 Análisis exploratorio del grafo

# Lo primero que es necesario hacer es convertir los datos existentes a un formato que pueda ser leído por igraph.


#1ro cargo los datos

load("bd/scopus.Rda") #son los datos de scopus en bruto, los mismos que usé en pasos anteriores.
```

```
dim(scopus)

##tengo 5697 filas/artículos y 30 variables que describen los artículos

#guardo solo las variables que en algún momento podrían interesarme

#creo dos data frame pare evitar el problema de factores vs carácteres

scopus_analiz_cuan<-data.frame(scopus$Year,scopus$Cited.by)

scopus_analiz_cual<-cbind(scopus$EID,scopus$X.U.FEFF.Authors,scopus$Affiliations)

scopus_analiz_cual<-as.data.frame(scopus_analiz_cual, stringsAsFactors = FALSE)

#ahora puedo unir ambos tipos de datos

scopus_analiz<-cbind(scopus_analiz_cuan,scopus_analiz_cual)

#Escoger elementos con más de una afiliación (son los que sirven para el ARS)

#la función grepl me ayuda a encontrar si un caracter está presente en una cadena de carácteres, si tengo ";" es un campo que tiene más de una afiliación

scop_maf<-scopus_analiz[grep(";", scopus_analiz[,5]), ]

#para facilitar trabajo cambio el nombre de las columnas

colnames(scop_maf)<-c("Year","Cited.by","EID","Aut","Af")


#Quiero modificar los caracteres en la columna 6 de tal manera que cada uno comience con un "; ", y termine con un ";"

#Es importante para no identificar como instituciones a cadenas cortas de caracteres incluidas en otras cadenas que identifican a la institucion.

#El resultado de esta modificación lo guardaré en una columna nueva.

scop_maf$Af2<-paste("; ",scop_maf$Af,";", sep ="")

####LIMPIEZA#########################

rm(scopus)

#########################################
```

```
#Cargo los datos de reconocimiento revisados

load("bd/dgr10_rev_simpl.Rda")

dgr<-dgr10_rev_simpl #para seguir usando el mismo código

load("bd/af_variat_no_Peru.Rda")

library(stringr)

#lo junto todo en un tercer data.frame

#sólo necesito tres columnas, (1) datos sin normalizar, (2) datos normalizados, (3) Pais

dgr<-dgr[,1:3]

names(dgr)<-c("X1","X2", "X3")

dgrall<-rbind(dgr,cbind(X1=af_variat_no_Peru,X2="",X3="noPeru"))

#Agrego los puntos y comas all comienzo y al final para que coincida

dgrall$X4<-paste("; ",dgrall$X1,";",sep="") #valores sin estandarizar

#agrego puntos y comas en los valores estandarizados (en caso contrario no encontrará puntos y comas para reconocer los campos)

dgrall$X2<-paste(";",dgrall$X2,";",sep="") #valores estandarizados

dgrall$X5<-paste(";",dgrall$X1,";",sep="") #valores sin estandarizar y sin espacio después del punto y coma

#En Af3 guardaré los registros modificados usando el reconocimiento #str_replace_all

#primero hago coincidir los datos sometiendolos al procedimiento previo de limpieza

#Remover signos diacríticos

unwanted_array = list(    'Š'='S', 'š'='s', 'Ž'='Z', 'ž'='z', 'À'='A', 'Á'='A', 'Â'='A', 'Ã'='A', 'Ä'='A',
'Å'='A', 'Æ'='A', 'Ç'='C', 'È'='E', 'É'='E',

                'Ê'='E', 'Ë'='E', 'Ì'='I', 'Í'='I', 'Î'='I', 'Ï'='I', 'Ñ'='N', 'Ò'='O', 'Ó'='O', 'Ô'='O',
'Õ'='O', 'Ö'='O', 'Ø'='O', 'Ù'='U',

                'Ú'='U', 'Û'='U', 'Ü'='U', 'Ý'='Y', 'Þ'='B', 'ß'='Ss', 'à'='a', 'á'='a', 'â'='a', 'ã'='a',
'ä'='a', 'å'='a', 'æ'='a', 'ç'='c',
```

'è'='e', 'é'='e', 'ê'='e', 'ë'='e', 'ì'='i', 'í'='i', 'î'='i', 'ï'='i', 'ð'='o', 'ñ'='n', 'ò'='o', 'ó'='o', 'ô'='o', 'õ'='o',

'ö'='o', 'ø'='o', 'ù'='u', 'ú'='u', 'û'='u', 'ý'='y', 'ý'='y', 'þ'='b', 'ÿ'='y')

scop_maf$Af3<-chartr(paste(names(unwanted_array), collapse=''), paste(unwanted_array, collapse=''), scop_maf$Af2)

#Remover los espacios sobrantes es algo nuevo, si no funciona lo borro

library(stringr)

library(qdap)

ahora necesito agregar a la lista todos los valores que no tienen una estandarización,

#es lo que he hecho en dgrall

table(dgrall$X2)

af_reem <- mgsub(dgrall$X4, dgrall$X2, scop_maf$Af3)

#esto funciona así mgsub(original, cadena de caracteres dentro del original a ser reemplazada, cadena de caracteres de reemplazo)

head(af_reem)

#esto funciona así mgsub(original, cadena de caracteres dentro del original a ser reemplazada, cadena de caracteres de reemplazo)

#A continuación corrigo algunos errores de tipeo que hay en los datos

af_reem1 <- mgsub(dgrall$X5, dgrall$X2, af_reem)

af_reem1 <-mgsub("; Department of Public Health, Institute of Tropical Medicine, Antwerp, Belgium;","","",af_reem1)

af_reem1 <-mgsub("; PATH, Seattle, WA, United States;","","",af_reem1)

af_reem1 <-mgsub("; Escuela de Terapia Fisica, Universidad Peruana de Ciencias Aplicadas, Villa (Chorrillos), Lima, Peru;","PR_UNI_UPC__","",af_reem1)

af_reem1 <-mgsub("; Centre on Aging, University of Manitoba, Winnipeg, MB, Canada;","","",af_reem1)

```
af_reem1 <-mgsub("; University of Southern California, Los Angeles, CA, United States;",";",af_reem1)

af_reem1 <-mgsub("; National Centre of Intercultural Health, National Institute of Health in Peru, Peru;","PU_IPI_INS__,",af_reem1)

af_reem1 <-mgsub("; Colegio Medico del Peru, Lima, Peru;","PR_ONG_COLME,",af_reem1)

af_reem1 <-mgsub("; Centro de Investigacion de Enfermedades Tropicales de la Marina de los EE, UU. (NAMRU-6), Iquitos, Peru;","PR_ORI_NAMRU,",af_reem1)

pr<-data.frame(af_reem1)

#En el siguiente, elimino todos los ;

af_reem1 <-mgsub(";","",af_reem1)

af_reem1<-gsub("\\s+", "", str_trim(af_reem1))

#después tengo que darle una revisión a los resultados

revisar<-unlist(strsplit(af_reem1, '\\,'))

sort(table(revisar),decreasing = TRUE)[1:20]

table(revisar)

head(sort(table(revisar)),10)

length(revisar)

#Reviso los resultados por artículo

table(sort(table(af_reem1),decreasing = TRUE))

#Creo una tabla con las afiliaciones por un lado y los datos del otro

scop_maf$AfPst<-af_reem1

#Guardo el resultado

#el de febrero

scop_maf10<-scop_maf

#save(scop_maf10,file="bd/scop_maf10.Rda")
```

```
load("bd/scop_maf10.Rda")

#Identifico las instituciones existentes
```

```
#Ahora

load("bd/scop_maf10.Rda")

scop_maf<-scop_maf10


############################

#Usar vector de carácteres separados por comas

spl_trans_to_edglist<-function(Wektor){

cols=list()


for (i in 1:length(Wektor)){ #

  af_fila1<-paste(Wektor[i])

af_fila1=unique(unlist(strsplit(af_fila1, '\\,'))) # splits the character strings into list with diff
erent vector for each line

#para evitar errores relacionados con no poder generar valores porque queda una sóla afi
liacion agrego la siguiente

#condicion - después eliminaré estas filas

if(length(af_fila1)==1) {

af_fila1 <- c(af_fila1,af_fila1)

}

af_fila1<-combn(af_fila1,2)

col1<-af_fila1[1,]

col2<-af_fila1[2,]

cols[[i]]<-data.frame(cbind(col1,col2),stringsAsFactors = FALSE)
```

```
}
return(cols)}


cols<-spl_trans_to_edglist(scop_maf$AfPst)
cols_adj<-do.call(rbind,cols)
head(cols_adj)
##          col1         col2
## 1 PR_UNI_UPCH_ PR_SAL_IIN__
## 2 PU_SAL_INEN_ PU_SAL_INEN_
## 3 PR_UNI_UPCH_ PR_UNI_UPCH_
## 4 PR_UNI_UPCH_ PU_SAL_HNCH_
## 5 PU_SAL_HDOSM PR_UNI_UPCH_
## 6 PU_SAL_HNAGV PU_SAL_HNAGV
str(cols_adj)
## 'data.frame':    9283 obs. of  2 variables:
##  $ col1: chr  "PR_UNI_UPCH_" "PU_SAL_INEN_" "PR_UNI_UPCH_" "PR_UNI_UPCH_" ...
##  $ col2: chr  "PR_SAL_IIN__" "PU_SAL_INEN_" "PR_UNI_UPCH_" "PU_SAL_HNCH_" ...
instituciones<-unique(c(cols_adj[,1],cols_adj[,2]))
str(instituciones)
##  chr [1:733] "PR_UNI_UPCH_" "PU_SAL_INEN_" "PU_SAL_HDOSM" ...
###
#Elimino los self-loops (bucles)
cols_adj$eq<-cols_adj$col1==cols_adj$col2
```

```
cols_adj<-cols_adj[cols_adj$eq==FALSE,]

cols_adj<-cols_adj[,1:2]

#cols_adj10<-cols_adj

#save(cols_adj10, file="bd/cols_adj10.Rda")
```

8.2.5 Generación y procesamiento de grafo

 # Comandos correspondientes a Generación de los grafos, en total se han generado cuatro archuvos: g – el multigrafo sin simplificar, wg – el grafo ponderado (simplificado), wg_p – la componente conexa del grafo simplificado, cl.wg_p – la componente conexa del grafo simplificado separada por conglomerados, wg_p_1 - la componente conexa del grafo simplificado (wg_p) sin las correspondientes a menos de 4 coautorías.

```
load("bd/cols_adj10.Rda")
edgtimes2<-cols_adj10

library(igraph)
##
## Attaching package: 'igraph'
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
## The following object is masked from 'package:base':
##
##     union

g=graph_from_data_frame(edgtimes2, directed=FALSE)
g
## IGRAPH UN-- 614 6475 --
## + attr: name (v/c), both (e/c)
## + edges (vertex names):
##  [1] PR_UNI_UPCH_--PR_SAL_IIN__ PR_UNI_UPCH_--PU_SAL_HNCH_
```

```
##  [3] PR_UNI_UPCH_--PU_SAL_HDOSM PR_UNI_UPCH_--PR_ONG_PRISM
##  [5] PR_ONG_PRISM--PU_GNO_MINSA PR_UNI_UPCH_--PU_GNO_MINSA
##  [7] PR_ORI_NAMRU--PR_SAL_IMPAC PR_ORI_NAMRU--PU_GNO_MINSA
##  [9] PR_SAL_IMPAC--PU_GNO_MINSA PR_UNI_UPCH_--PU_UNI_UNITR
## [11] PR_UNI_UPCH_--FI_FI__014__ PR_ORI_NAMRU--PU_IPI_INS__
## [13] PR_UNI_USMP_--FI_FI__101__ PR_UNI_UPCH_--PR_ONG_PRISM
## [15] PR_UNI_UPCH_--PU_SAL_HDOSM PR_UNI_UPCH_--PU_SAL_HAMA_
## + ... omitted several edges
#quitando RM
g<-g-vertices(c("RM"))
g$name <- "Red de coautorías medicina 2000-2015"

####
#Aquí resumo las características del multigrafo
g
## IGRAPH UN-- 613 6453 -- Red de coautorías medicina 2000-2015
## + attr: name (g/c), name (v/c), both (e/c)
## + edges (vertex names):
##  [1] PR_UNI_UPCH_--PR_SAL_IIN__ PR_UNI_UPCH_--PU_SAL_HNCH_
##  [3] PR_UNI_UPCH_--PU_SAL_HDOSM PR_UNI_UPCH_--PR_ONG_PRISM
##  [5] PR_ONG_PRISM--PU_GNO_MINSA PR_UNI_UPCH_--PU_GNO_MINSA
##  [7] PR_ORI_NAMRU--PR_SAL_IMPAC PR_ORI_NAMRU--PU_GNO_MINSA
##  [9] PR_SAL_IMPAC--PU_GNO_MINSA PR_UNI_UPCH_--PU_UNI_UNITR
## [11] PR_UNI_UPCH_--FI_FI__014__ PR_ORI_NAMRU--PU_IPI_INS__
## [13] PR_UNI_USMP_--FI_FI__101__ PR_UNI_UPCH_--PR_ONG_PRISM
## [15] PR_UNI_UPCH_--PU_SAL_HDOSM PR_UNI_UPCH_--PU_SAL_HAMA_
## + ... omitted several edges
#% de instituciones peruanas que colaboran con instituciones peruanas
length(V(g)$name)/733
## [1] 0.8362892
#Cuantos clusters
is.connected(g)
## [1] FALSE
#tamaño de los clusters
```

```
clusters(g)$csize
##  [1] 568   2   2   2   2   2   2   2   2   2   3   3   2   2   2   3   3
## [18]   2   3   2   2
diameter(g)
## [1] 8
#################

summary(g)
## IGRAPH UN-- 613 6453 -- Red de coautorías medicina 2000-2015
## + attr: name (g/c), name (v/c), both (e/c)
##

#Extrayendo sector y tipo
vertices<-V(g)$name
summary(vertices)
##    Length     Class      Mode
##       613 character character
vertices_sec<-substr(vertices,1,2)
vertices_tip<-substr(vertices,4,6)

#Agregar estas propiedades
V(g)$tip<-vertices_tip
V(g)$sec<-vertices_sec

#Propiedades de los vértices
list.vertex.attributes(g)
## [1] "name" "tip"  "sec"
#Veo si es un grafo simple o un multigrafo
is.simple(g)
## [1] FALSE
#debería darme "TRUE"
#plot(gm,edge.width=E(g)$weight/2)

#Guardo g - multigrafo
```

```
#save(g,file="graphs_data/g.Rda")

wg<-g
E(wg)$weight <- 1
wg <- simplify(wg)
is.simple(wg)
## [1] TRUE
str(V(wg)$name)
##  chr [1:613] "PR_UNI_UPCH_" "PU_SAL_HDOSM" "PR_ONG_PRISM" ...
str(V(g)$name)
##  chr [1:613] "PR_UNI_UPCH_" "PU_SAL_HDOSM" "PR_ONG_PRISM" ...
#Se mantuvieron los pesos?
table(E(wg)$weight)
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 1451  382  162   67   45   28   24   18    8   12   11    9    8    5    2
##   16   17   18   19   20   21   22   23   24   26   29   33   34   38   40
##    4    4    4    1    2    3    1    1    2    1    1    1    2    3    1
##   41   43   50   51   56   58   62   78   81  105  116  138  179  188
##    1    1    1    1    1    1    1    1    1    1    1    1    2    1
summary(wg)
## IGRAPH UNW- 613 2278 -- Red de coautorías medicina 2000-2015
## + attr: name (g/c), name (v/c), tip (v/c), sec (v/c), weight (e/n)
#Para identificar los conglomerados necesito que el grafo tenga un solo componente

is.connected(wg)
## [1] FALSE
#tamaño de los clusters
clusters(wg)$csize
##  [1] 568   2   2   2   2   2   2   2   2   2   3   3   2   2   2   3   3
## [18]   2   3   2   2
clusters(wg)$no
## [1] 21
#El diametro
```

```
diameter(wg, weights=NA)
## [1] 8
diameter(wg)
## [1] 11
#Guardo el grafo weighted
#save(wg,file="graphs_data/wg.Rda")

#Separo los grafos conexos
dg <- decompose.graph(wg)
#sólo tomo en cuenta al componente conexa principal
wg_p<-dg[[1]]

###
#agrego etiquetas
vertices<-V(wg_p)$name
vertices_lab<-substr(vertices,8,12)
V(wg_p)$label<-vertices_lab
############################3

#Comienzo con las etiquetas
#agregaré etiquetas
#Revisando propiedades
summary(wg_p)
## IGRAPH UNW- 568 2249 -- Red de coautorías medicina 2000-2015
## + attr: name (g/c), name (v/c), tip (v/c), sec (v/c), label (v/c),
## | weight (e/n)
#Probando con RGB
V(wg_p)[tip == "SAL"]$color <- "#e41a1c"
V(wg_p)[tip == "ONG"]$color <- "#377eb8"
V(wg_p)[tip == "UNI"]$color <- "#4daf4a"
V(wg_p)[tip == "ORI"]$color <- "#984ea3"
V(wg_p)[tip == "EMP"]$color <- "#ff7f00"
V(wg_p)[tip == "GRL"]$color <- "#ffff33"
V(wg_p)[tip == "GNO"]$color <- "#a65628"
```

```
V(wg_p)[tip == "IPI"]$color <- "#f781bf"
V(wg_p)[tip == "FI_"]$color <- "#999999"

# Sector indicado por forma
V(wg_p)[sec=="PR"]$shape <- "circle"
V(wg_p)[sec=="PU"]$shape <- "square"
V(wg_p)[sec=="FI"]$shape <- "rectangle"

#Vertex size proportional to degree
# Vertex area proportional to vertex strength
# (i.e., total weight of incident edges).
V(wg_p)$size <- log(graph.strength(wg_p))/2 #aquí uso escala logarítmica
V(wg_p)$size2 <- log(graph.strength(wg_p))/4

#Weight edges by weight
E(wg_p)$width <- log(E(wg_p)$weight) #aquí uso escala logarítmica
####
#Resumiendo propiedades
summary(wg_p)
## IGRAPH UNW- 568 2249 -- Red de coautorías medicina 2000-2015
## + attr: name (g/c), name (v/c), tip (v/c), sec (v/c), label (v/c),
## | color (v/c), shape (v/c), size (v/n), size2 (v/n), weight (e/n),
## | width (e/n)
# Color edges by within/between faction.
SAL<-V(wg_p)[tip == "SAL"]
ONG<-V(wg_p)[tip == "ONG"]
UNI<-V(wg_p)[tip == "UNI"]
ORI<-V(wg_p)[tip == "ORI"]
EMP<-V(wg_p)[tip == "EMP"]
GRL<-V(wg_p)[tip == "GRL"]
GNO<-V(wg_p)[tip == "GNO"]
IPI<-V(wg_p)[tip == "IPI"]
FI_<-V(wg_p)[tip == "FI_"]
```

```
#########################
#Si la máquina no soporta transparencias usar estos colores, para transparencias agrego
alpha=0.7
E(wg_p)$color <- rgb(210/255, 210/255, 210/255)
E(wg_p)[ SAL %--% SAL ]$color <- "#ed5e5f"  # e41a1c
E(wg_p)[ ONG %--% ONG ]$color <- "#69a3d2"  # 377eb8
E(wg_p)[ UNI %--% UNI ]$color <- "#80c87d"  # 4daf4a
E(wg_p)[ UNI %--% IPI ]$color <- "#80c87d"  # 4daf4a
E(wg_p)[ IPI %--% IPI ]$color <- "#80c87d"  # 4daf4a
E(wg_p)[ ORI %--% ORI ]$color <- "#b87dc1"  # 984ea3
E(wg_p)[ EMP %--% EMP ]$color <- "#ffa54d"  # ff7f00
E(wg_p)[ GRL %--% GNO ]$color <- "#d37a48"  # a65628
E(wg_p)[ GRL %--% GRL ]$color <- "#d37a48"  # a65628
E(wg_p)[ GNO %--% GNO ]$color <- "#d37a48"  # a65628
#posibles combinaciones de diferentes elementos

#revisar asignación de colores
table(E(wg_p)$color)
##
## #69a3d2 #80c87d #b87dc1 #D2D2D2 #d37a48 #ed5e5f
##      28     183      11    1433      42     552
#Veo el tamaño de los vertices decreciente
sort(V(wg_p)$size, decreasing = T)[1:5]
## [1] 3.762281 3.464269 3.244602 3.128834 3.062342
#Cuando quiero limitar etiquetas
V(wg_p)$label <-ifelse(V(wg_p)$size >= 3.05, vertices_lab, "")
par(bg="white")

#Un layout de ejemplo
#Escoger layout
lkk <- layout.kamada.kawai(wg_p)

plot(wg_p, layout=lkk, edge.curved=T, vertex.label.color= "black", vertex.label.font=2,
    vertex.label.cex=0.8, vertex.frame.color="#FFFFFF")
```

```
#legend(x="bottomleft", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI"),
pch=21,     col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)
#Para el marco no uso este título
title("Distribución con Kamada y Kawai")

###############
#identificación de conglomerados
cl.wg_p <- fastgreedy.community(wg_p)

length(cl.wg_p)
## [1] 16
V(wg_p)$community <- cl.wg_p$membership


#Guardo identificación de conglomerados
#save(cl.wg_p,file="graphs_data/cl.wg_p.Rda")

###################
#Gran componente reducido
#Escogo los vértices que cumplen cierta característica

wg_p
## IGRAPH UNW- 568 2249 -- Red de coautorías medicina 2000-2015
## + attr: name (g/c), name (v/c), tip (v/c), sec (v/c), label (v/c),
## | color (v/c), shape (v/c), size (v/n), size2 (v/n), community
## | (v/n), weight (e/n), width (e/n), color (e/c)
## + edges (vertex names):
##  [1] PR_UNI_UPCH_--PU_SAL_HDOSM PR_UNI_UPCH_--PR_ONG_PRISM
##  [3] PR_UNI_UPCH_--PR_ORI_NAMRU PR_UNI_UPCH_--PR_SAL_IMPAC
##  [5] PR_UNI_UPCH_--PU_UNI_UNITR PR_UNI_UPCH_--PR_UNI_USMP_
##  [7] PR_UNI_UPCH_--PU_SAL_HLOAY PR_UNI_UPCH_--PU_SAL_HNCH_
##  [9] PR_UNI_UPCH_--PU_UNI_UNMSM PR_UNI_UPCH_--PU_UNI_UNICA
## [11] PR_UNI_UPCH_--PR_ORI_OPS__ PR_UNI_UPCH_--PR_SAL_PPJAP
## + ... omitted several edges
```

```
V(wg_p)$grado<-degree(wg_p)
V(wg_p)$intermediacion<-betweenness(wg_p)
V(wg_p)$eigen_centr<-eigen_centrality(wg_p)$vector
V(wg_p)$fuerza<-strength(wg_p)

#Guardo el componente principal con membership y medidas de centralidad
#save(wg_p,file="graphs_data/wg_p.Rda")


#Genero el componente reducido para algunos gráficos el número de vértices hace la visualización poco clara
wg_p72<-wg_p-vertices(V(wg_p)[V(wg_p)$fuerza<25])
is.connected(wg_p72)
## [1] TRUE
wg_p72
## IGRAPH UNW- 72 739 -- Red de coautorías medicina 2000-2015
## + attr: name (g/c), name (v/c), tip (v/c), sec (v/c), label (v/c),
## | color (v/c), shape (v/c), size (v/n), size2 (v/n), community
## | (v/n), grado (v/n), intermediacion (v/n), eigen_centr (v/n),
## | fuerza (v/n), weight (e/n), width (e/n), color (e/c)
## + edges (vertex names):
##  [1] PR_UNI_UPCH_--PU_SAL_HDOSM PR_UNI_UPCH_--PR_ONG_PRISM
##  [3] PR_UNI_UPCH_--PR_ORI_NAMRU PR_UNI_UPCH_--PR_SAL_IMPAC
##  [5] PR_UNI_UPCH_--PU_UNI_UNITR PR_UNI_UPCH_--PR_UNI_USMP_
##  [7] PR_UNI_UPCH_--PU_SAL_HLOAY PR_UNI_UPCH_--PU_SAL_HNCH_
##  [9] PR_UNI_UPCH_--PU_UNI_UNMSM PR_UNI_UPCH_--PU_UNI_UNICA
## + ... omitted several edges
#save(wg_p72,file="graphs_data/wg_p72.Rda")

#El resultado no se visualiza bien con centralidad, eliminaré aristas <4
wg_p_1 <- delete_edges(wg, E(wg)[weight<4])

#Compruebo
```

```
is_connected(wg_p_1)
## [1] FALSE
wg_p_1<-decompose.graph(wg_p_1)[[1]]
#save(wg_p_1,file="graphs_data/wg_p_1.Rda")
```

8.2.6 Figuras 20 a 43.

- *FIGURA 20. Características de las instituciones involucradas en el grafo de coautorías*

```
load("graphs_data/g.Rda")
library(igraph)
all_institutions<-c((get.edgelist(g)[,1]),(get.edgelist(g)[,2]))
all_institutions<-unique(all_institutions)
all_institutions_sec<-substr(all_institutions,1,2)
all_institutions_tip<-substr(all_institutions,4,6)
par(mfrow=c(2,1))
par(las=2) # make label text perpendicular to axis
par(mar=c(5,5,4,2)) # increase y-axis margin.
tr<-sort(table(all_institutions_tip),decreasing = T)
counts<-as.numeric(tr)
nombres<-names(tr)
barplot(counts, main="Tipos de instituciones en el grafo",  cex.main = 1,
      horiz=TRUE, names.arg=nombres, cex.names=0.9, col="dodgerblue")
####
tr<-sort(table(all_institutions_sec),decreasing = T)
counts<-as.numeric(tr)
nombres<-names(tr)
```

```
barplot(counts, main="Sector de las instituciones en el grafo",  cex.main = 1,
        horiz=TRUE, names.arg=nombres, cex.names=0.9, col="dodgerblue")
```

- *FIGURA 21. Centralidades en el grafo de coautorías de instituciones peruanas con investigación en medicina en Scopus entre el 2000 y el 2015*

```
#cargo los grafos
load("graphs_data/wg_p_1.Rda")
library(igraph)
g <-wg_p_1
A <- get.adjacency(g, sparse=FALSE)
library(network)
g <- network::as.network.matrix(A, directed = F)
library(sna)
colores<-c("#bd0026",rep("#bdc9e1",1),"#ffffb2","#fecc5c",rep("#bdc9e1",5),"#f03b20",
           rep("#bdc9e1",6),"#fd8d3c",rep("#bdc9e1",75))
df<-data.frame(V(wg_p_1)$name, colores)
sna::gplot.target(g, strength(wg_p_1), main="Grado ponderado",
                  circ.lab = FALSE, circ.col="skyblue",
                  usearrows = FALSE,
                  vertex.col=colores,
                  edge.col="lightgray")
sna::gplot.target(g, betweenness(g), main="Intermediación",
                  circ.lab = FALSE, circ.col="skyblue",
                  usearrows = FALSE,
```

```
            vertex.col=colores,
            edge.col="lightgray")
sna::gplot.target(g, evcent(g), main="Centralidad Vector Propio",
            circ.lab = FALSE, circ.col="skyblue",
            usearrows = FALSE,
            vertex.col=colores,
            edge.col="lightgray")
```

- *FIGURA 22. Representaciones del grafo de coautorías de instituciones peruanas con investigación en medicina en Scopus entre el 2000 y el 2015 utilizando diferentes algoritmos de distribución de los vértices*

```
##########################
#guardar 1000 width, 1200 high
par(mfrow=c(1,2))
#Probando diferentes layouts
load("graphs_data/wg.Rda")
g<-wg
library(igraph)
igraph.options(vertex.size=2, vertex.label=NA,
            edge.arrow.size=0.3,edge.color=rgb(158,202,225,max=255),
            vertex.color=rgb(49,130,189,max=255))
plot(g, layout=layout.circle)
title("Círculo")
#En este caso se trata de mostrar las centralidades mediante la centralidad de los nodos
#Ejemplo con energy placement methods :#El método de Kamada y Kawai se basa en eso
```

```
plot(g, layout=layout.kamada.kawai)
title("Kamada - Kawaii")
plot(g, layout=layout.fruchterman.reingold)
title("Fruchterman Reingold")


plot(g, layout=layout.drl)
title("DRL")
#Reingold Tilford muestra jerarquías
par(mfrow=c(1, 2))
plot(g, layout=layout.reingold.tilford(g,circular=T))
title("Reingold Tilford (Circular)")
plot(g, layout=layout.reingold.tilford)
title("Reingold Tilford (Vertical)")
```

- *FIGURA 23. Distribución de la ponderación de las aristas en el grafo de coautorías de instituciones peruanas con investigación en medicina indizada en Scopus (2000-2015).*

```
plot(table(E(wg)$weight), col="dodgerblue",
    xlab="Peso de las aristas (número de coautorías que representan)", ylab="Frecuencia", main="", type="p", pch=16)
```

- *FIGURA 24. Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución Kamada y Kawai. / FIGURA 25. Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución Fruchterman Reingold. /*

FIGURA 26. Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución con DRL. / FIGURA 27. Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución con Large Graph Layout.

```
#FIGURA 24. Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Distribución Kamada y Kawai./ Figura 25. Distribución Fruchterman Reingold./ Figura 26. Distribución DRL./  Figura 27. Distribución Large Graph Layout

load("graphs_data/wg_p.Rda")


wg<-wg_p

library(igraph)

#Layout

set.seed(45)

lkk <- layout.kamada.kawai(wg)

set.seed(45)

lfr<-layout.fruchterman.reingold(wg)

set.seed(45)

ldrl <- layout.drl(wg)

set.seed(45)

llgl<-layout_with_lgl(wg,root="PR_UNI_UPCH_")


#Revisando propiedades

vertices<-V(wg)$name

vertices_lab<-substr(vertices,8,12)
```

```
#Lo siguiente está bien pero demora en procesar

#Originalmente era vertex.label.cex=0.8, lo estoy cambiando a 1.5 para guardarlo con anc
ho 1000px


#Cambio los colores para que las etiquetas se noten mejor, ojo hay maquinas que no sopor
tan transparencias,

#Además, dependiendo de la tarjeta gráfica los resultados de este código varían


#Tipo indicado por color

col1<-rgb(228/255,26/255,28/255, alpha = 0.7)

col2<-rgb(55/255,126/255,184/255, alpha = 0.7)

col3<-rgb(77/255,175/255,74/255, alpha = 0.7)

col4<-rgb(152/255,78/255,163/255, alpha = 0.7)

col5<-rgb(255/255,127/255,0/255, alpha = 0.7)

col6<-rgb(255/255,255/255,51/255, alpha = 0.7)

col7<-rgb(166/255,86/255,40/255, alpha = 0.7)

col8<-rgb(247/255,129/255,191/255, alpha = 0.7)

col9<-rgb(153/255,153/255,153/255, alpha = 0.7)


colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)


plot(wg, layout=lkk, edge.curved=T, vertex.label.color= "black", vertex.label.font=2,

    vertex.label.cex=0.8, vertex.frame.color="#FFFFFF")

legend(x="bottomleft", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI"),
pch=21,
```

```
      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)
#Para el marco no uso este título
#title("Distribución con Kamada y Kawai")

###################################3
plot(wg, layout=lfr, edge.curved=T, vertex.label.color= "black", vertex.label.font=2,
    vertex.label.cex=0.8, vertex.frame.color="#FFFFFF")
legend(x="bottomleft", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI"),
pch=21,
      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=1)
#No lo aplico en marco teórico
#title("Distribución con Fruchterman Reingold")
#################################
#La visualización no es tan clara, utilizo el algoritmo para grafos grandes
plot(wg, layout=ldrl, edge.curved=T, vertex.label=NA, vertex.frame.color="#FFFFFF") #C
ambié etiquetas porque no se pueden visualizar
legend(x="bottomleft", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI"),
pch=21,
      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=1)
#No lo aplico en marco teórico
#title("Distribución con DRL")

#################################
#La visualización no es tan clara, utilizo el algoritmo para grafos grandes
#Las etiquetas se sobreponen
```

```
V(wg)$label <-ifelse(V(wg)$size >= 3.2, vertices_lab, "")


plot(wg, layout=llgl, edge.curved=T, vertex.label.color= "black", vertex.label.font=2,
    vertex.label.cex=0.8, vertex.frame.color="#FFFFFF")
legend(x="bottomleft", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI"),
pch=21,
     col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=1)
#No lo aplico en marco teórico
#title("Distribución con Large Graph Layout")
```

- *FIGURA 28. Grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015. Tipos de instituciones.*

```
#FIGURA 28. Grafo de coautorías de las instituciones peruanas con investigación en medi
cina indizada en Scopus, 2000-2015. Tipos de instituciones
load("graphs_data/wg_p.Rda")


library(igraph)
par(new=TRUE)
#Ahora agrupo los elementos del mismo tipo
#bIEN LENTO
tipos.f <- as.factor(V(wg_p)$tip)
tipos.nums<-as.numeric(tipos.f)
wg_p.c <- contract.vertices(wg_p, tipos.nums)
E(wg_p.c)$weight <- 1
wg_p.c <- simplify(wg_p.c)
```

```
tipos.size <- as.vector(table(V(wg_p)$tip))

tipos.names<-names(table(V(wg_p)$tip))

plot(wg_p.c, vertex.size=5*sqrt(tipos.size),vertex.color=V(wg_p.c), edge.width=sqrt(E(wg_p.c)$weight),

    vertex.label=tipos.names, vertex.label.dist=0, edge.arrow.size=0) #después de vertex.size va , ¿de donde saca los datos?

title("Colaboración por tipo de institución")
```

- *FIGURA 29. Distribución del grado y fuerza del grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
par(mfrow=c(3,1))

#Degree distribution

plot(table(degree(wg)), col="dodgerblue",

    xlab="Grado de los vértices", ylab="Frecuencia", main="", xlim=c(1,1860),

    ylim=c(1,170), type="p", pch=16)


#si quero agregar límite en x agrego xlim=c(0, 50)

title("Distribución de grados")

#Strength distribution (weighted)

plot(table(graph.strength(wg)),col="red",xlab="Fuerza de los vértices", ylab="Frecuencia",

    main="", xlim=c(1,1860),

    ylim=c(1,170), type="p", pch=16)

title("Distribución de fuerzas")
```

```
#log log
d.gm <- degree(wg)
dd.gm <- degree.distribution(wg)
d <- 1:max(d.gm)-1
ind <- (dd.gm != 0)
plot(d[ind], dd.gm[ind], log="xy", col="dodgerblue", xlab=c("Log-Grado"),
    ylab=c("Log-Fuerza"), main="Distribución Log-Log")
```

- *FIGURA 30. Promedio del grado de los vecinos versus grado de los vértices (escala logarítmica) para las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
#Homophilia por fuerza? Usaré el gran componente
par(mfrow=c(1,1))
a.nn.deg.wg_p <- graph.knn(wg_p,V(wg_p))$knn
d.wg_p<-strength(wg_p)
plot(d.wg_p, a.nn.deg.wg_p, log="xy",
    col="goldenrod", xlab=c("Log Vertex Degree"),
    ylab=c("Log Average Neighbor Degree"))
```

- *FIGURA 31. Tamaño de los conglomerados hallados en el grafo de coautorías*

```
#se identificó 16 conglomerados
par(mfrow=c(1,1))
plot(sort(sizes(cl.wg_p)),ylab="Tamaño de los conglomerados", xlab="", col="red",panel.f
irst=grid())
```

- *FIGURA 32. Simulación de las frecuencias relativas del número de comunidades identificados en el grafo sin estructura comunitaria.*

```
#cargo los grafos

load("graphs_data/wg_p.Rda")

library(igraph)

nv <- vcount(wg_p)

ne <- ecount(wg_p)

degs <- degree(wg_p)

#Over 1000 trials,

ntrials <- 1000

#we then generate classical random graphs of this same order and size and, for each one, w
e use the same community detection algorithm to determine the number of communities.


num.comm.rg <- numeric(ntrials)

for(i in (1:ntrials)){

g.rg <- erdos.renyi.game(nv, ne, type="gnm")

c.rg <- fastgreedy.community(g.rg)

num.comm.rg[i] <- length(c.rg)

}

#Similarly, we do the same using generalized random graphs constrained to have the requi
red degree sequence.


num.comm.grg <- numeric(ntrials)

for(i in (1:ntrials)){
```

```
g.grg <- degree.sequence.game(degs, method="vl")

c.grg <- fastgreedy.community(g.grg)

num.comm.grg[i] <- length(c.grg)

}

#The results may be summarized and compared using side by side bar plots.

rslts <- c(num.comm.rg,num.comm.grg)

indx <- c(rep(0, ntrials), rep(1, ntrials))

counts <- table(indx, rslts)/ntrials

barplot(counts, beside=TRUE, col=c("blue", "red"),

xlab="Número de conglomerados",

ylab="Frecuencia relativa",

legend=c("Grafo al azar de tamaño igual al analizado", "Grafo al azar con distribución de g
rados igual al analizado"))
```

- *FIGURA 33. Conglomerados en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
#Figura 33. Conglomerados en el grafo de coautor´ias de las instituciones peruanas con in
vestigaci´ón en medicina indizada en Scopus, 2000-2015

load("graphs_data/wg_p.Rda")

load("graphs_data/cl.wg_p.Rda")


library(igraph)

#plot de conglomerados

#Layout

set.seed(45)
```

```
lkk <- layout.kamada.kawai(wg_p)


#Tipo indicado por color

col1<-rgb(228/255,26/255,28/255, alpha = 0.7)

col2<-rgb(55/255,126/255,184/255, alpha = 0.7)

col3<-rgb(77/255,175/255,74/255, alpha = 0.7)

col4<-rgb(152/255,78/255,163/255, alpha = 0.7)

col5<-rgb(255/255,127/255,0/255, alpha = 0.7)

col6<-rgb(255/255,255/255,51/255, alpha = 0.7)

col7<-rgb(166/255,86/255,40/255, alpha = 0.7)

col8<-rgb(247/255,129/255,191/255, alpha = 0.7)

col9<-rgb(153/255,153/255,153/255, alpha = 0.7)

colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)

plot(cl.wg_p,wg_p,edge.color = rgb(1,1,1,alpha = 0.2),edge.curved=T,layout=lkk, vertex.la
bel.dist=0,edge.curved=T, vertex.label.color= "black", vertex.label.font=2,

    vertex.label.cex=0.8, vertex.frame.color="#FFFFFF")

legend(x=-0.8, y=-0.5, c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI"),
pch=21,

     col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=1)

title("Identificación de conglomerados en el grafo")

colrs16<-c(rgb(166,206,227,maxColorValue = 255),rgb(31,120,180,maxColorValue = 255
),rgb(178,223,138,maxColorValue = 255),

        rgb(51,160,44,maxColorValue = 255),rgb(251,154,153,maxColorValue = 255),rgb(
227,26,28,maxColorValue = 255),

        rgb(253,191,111,maxColorValue = 255),rgb(255,127,0,maxColorValue = 255),rgb(
202,178,214,maxColorValue = 255),
```

```
        rgb(106,61,154,maxColorValue = 255),rgb(255,255,153,maxColorValue = 255),rgb
(177,89,40,maxColorValue = 255)

        ,rgb(177,89,40,maxColorValue = 255),rgb(177,89,40,maxColorValue = 255),rgb(17
7,89,40,maxColorValue = 255),

        rgb(177,89,40,maxColorValue = 255))

#Cambio el tamaño

V(wg_p)$size <- log(graph.strength(wg_p)) #aquí uso escala logarítmica

V(wg_p)$community <- cl.wg_p$membership

#plot de conglomerados, nodos por color

plot(wg_p,edge.color = rgb(1,1,1,alpha = 0.2),edge.curved=T,layout=lkk, vertex.label.dist=
0,edge.curved=T,

    vertex.label.color= "black", vertex.label=NA, vertex.frame.color=colrs16[V(wg_p)$co
mmunity], vertex.color=colrs16[V(wg_p)$community])

legend(x=-0.9, y=-0.1, c("C1","C2", "C3","C4","C5","C6","C7","C8","C9","C10","C11","
C12","C13","C14","C15","C16"), pch=21,

     col="#777777", pt.bg=colrs16, pt.cex=2, cex=.8, bty="n", ncol=1)

title("Identificación de conglomerados en el grafo")

#Ojo las coordenadas x y y osn especificas para guardar con ancho 1000
```

- *FIGURA 34. Colaboración entre conglomerados en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
tipos.size <- as.vector(table(V(wg_p)$community))

tipos.names<-names(table(V(wg_p)$community))

plot(cgs, vertex.size=3*sqrt(tipos.size),vertex.color=V(cgs), edge.width=sqrt(E(cgs)$weigh
t)*0.7,
```

```
    vertex.label=tipos.names, vertex.label.dist=0, edge.arrow.size=0,
    vertex.label=seq(1,12), layout=layout_in_circle)
#después de vertex.size va , ¿de donde saca los datos?
title("Colaboración entre conglomerados")
```

- *FIGURA 35. Conglomerado 1 (INEN) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")
load("graphs_data/wg_p.Rda")


#1er gráfico (A)


cl1=delete_vertices(wg_p, which(V(wg_p)$community!=1))
vertices_lab<-substr(V(cl1)$name,8,12)
V(cl1)$label <-vertices_lab


col1<-rgb(228/255,26/255,28/255, alpha = 0.7)
col2<-rgb(55/255,126/255,184/255, alpha = 0.7)
col3<-rgb(77/255,175/255,74/255, alpha = 0.7)
col4<-rgb(152/255,78/255,163/255, alpha = 0.7)
col5<-rgb(255/255,127/255,0/255, alpha = 0.7)
col6<-rgb(255/255,255/255,51/255, alpha = 0.7)
col7<-rgb(166/255,86/255,40/255, alpha = 0.7)
col8<-rgb(247/255,129/255,191/255, alpha = 0.7)
```

```
col9<-rgb(153/255,153/255,153/255, alpha = 0.7)

colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)


plot(cl1, layout=layout_components, edge.curved=T,

    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl1)$color,

    vertex.label.cex=(V(cl1)$fuerza)/50, vertex.size=(V(cl1)$fuerza)/50, vertex.color="#FF
FFFF")

legend(x="bottomright", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI")
, pch=21,

      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)

title("Conglomerado 1")

#2do gráfico (B)

############

#Instituciones por sector y tipo

######################

#De todo el grafo

all_institutions<-V(wg_p)$name

all_institutions_sec<-substr(all_institutions,1,2)

all_institutions_tip<-substr(all_institutions,4,6)


trt_all<-table(all_institutions_tip)

counts_all<-as.numeric(trt_all)

nombres_all<-names(trt_all)


#Instituciones involucradas en el congloemrado
```

```
all_institutions<-df$V.cl1..name

all_institutions_sec<-substr(all_institutions,1,2)

all_institutions_tip<-substr(all_institutions,4,6)


#Genero gráfico

par(mfrow=c(2,1))

par(las=2) # make label text perpendicular to axis

par(mar=c(5,5,4,2)) # increase y-axis margin.


trt<-table(all_institutions_tip)

counts<-as.numeric(trt)

nombres<-names(trt)

barplot(counts, main="Tipos de instituciones en el conglomerado",  cex.main = 1,

      horiz=TRUE, names.arg=nombres, cex.names=0.9, col="dodgerblue")


####

counts<-rbind(c(round(trt_all/ sum(trt_all),digits = 2)),

            c(round(trt/ sum(trt),digits = 2)[1:4],0,c(round(trt/ sum(trt),digits = 2)[5:8])))

barplot(counts, beside=TRUE, col=c("blue", "red"),

      ylab="Frecuencia relativa",

      legend=c("Grafo completo","Conglomerado 1"),

      args.legend = list(x = "topleft"))
```

- *FIGURA 36. Conglomerado 2 (UNMSM) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")

load("graphs_data/wg_p.Rda")


library(igraph)

cl2=delete_vertices(wg_p, which(V(wg_p)$community!=2))

vertices_lab<-substr(V(cl2)$name,8,12)

V(cl2)$label <-vertices_lab


#1er gráfico (A)


col1<-rgb(228/255,26/255,28/255, alpha = 0.7)

col2<-rgb(55/255,126/255,184/255, alpha = 0.7)

col3<-rgb(77/255,175/255,74/255, alpha = 0.7)

col4<-rgb(152/255,78/255,163/255, alpha = 0.7)

col5<-rgb(255/255,127/255,0/255, alpha = 0.7)

col6<-rgb(255/255,255/255,51/255, alpha = 0.7)

col7<-rgb(166/255,86/255,40/255, alpha = 0.7)

col8<-rgb(247/255,129/255,191/255, alpha = 0.7)

col9<-rgb(153/255,153/255,153/255, alpha = 0.7)

colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)


plot(cl2, layout=layout_components, edge.curved=T,
```

```
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl2)$color,
    vertex.label.cex=(V(cl2)$fuerza)/300, vertex.size=(V(cl2)$fuerza)/300, vertex.color="#FFFFFF")
legend(x="bottomright", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI"), pch=21,
      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)
title("Conglomerado 2")


#2do gráfico (B)
############
#Instituciones por sector y tipo
######################
#De todo el grafo
all_institutions<-V(wg_p)$name
all_institutions_tip<-substr(all_institutions,4,6)

trt_all<-table(all_institutions_tip)
counts_all<-as.numeric(trt_all)
nombres_all<-names(trt_all)

#Instituciones involucradas en el congloemrado
all_institutions<-df$V.cl2..name
all_institutions_tip<-substr(all_institutions,4,6)

#Genero gráfico
```

```
par(mfrow=c(2,1))
par(las=2) # make label text perpendicular to axis
par(mar=c(5,5,4,2)) # increase y-axis margin.

trt<-table(all_institutions_tip)
counts<-as.numeric(trt)
nombres<-names(trt)
counts<-rbind(c(round(trt_all/ sum(trt_all),digits = 2)),
              c(round(trt/ sum(trt),digits = 2)))
barplot(counts, beside=TRUE, col=c("blue", "red"),
        ylab="Frecuencia relativa",
        legend=c("Grafo completo","Conglomerado 2"),
        args.legend = list(x = "topleft"))
```

- *FIGURA 37. Conglomerado 3 (UPC) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")
load("graphs_data/wg_p.Rda")

library(igraph)
#1er gráfico (A)

cl3=delete_vertices(wg_p, which(V(wg_p)$community!=3))
vertices_lab<-substr(V(cl3)$name,8,12)
```

```
V(cl3)$label <-vertices_lab


col1<-rgb(228/255,26/255,28/255, alpha = 0.7)

col2<-rgb(55/255,126/255,184/255, alpha = 0.7)

col3<-rgb(77/255,175/255,74/255, alpha = 0.7)

col4<-rgb(152/255,78/255,163/255, alpha = 0.7)

col5<-rgb(255/255,127/255,0/255, alpha = 0.7)

col6<-rgb(255/255,255/255,51/255, alpha = 0.7)

col7<-rgb(166/255,86/255,40/255, alpha = 0.7)

col8<-rgb(247/255,129/255,191/255, alpha = 0.7)

col9<-rgb(153/255,153/255,153/255, alpha = 0.7)

colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)


plot(cl3, layout=layout_components, edge.curved=T,

    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl3)$color,

    vertex.label.cex=(V(cl3)$fuerza)/200, vertex.size=(V(cl3)$fuerza/200), vertex.color="#
FFFFFF")

legend(x="bottomright", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI")
, pch=21,

      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)

title("Conglomerado 3")

#2do gráfico (B)

############

#Instituciones por sector y tipo

######################
```

```
#De todo el grafo
all_institutions<-V(wg_p)$name
all_institutions_tip<-substr(all_institutions,4,6)

trt_all<-table(all_institutions_tip)
counts_all<-as.numeric(trt_all)
nombres_all<-names(trt_all)

#Instituciones involucradas en el congloemrado
all_institutions<-df$V.cl3..name
all_institutions_tip<-substr(all_institutions,4,6)

trt<-table(all_institutions_tip)
counts<-as.numeric(trt)
nombres<-names(trt)

counts<-rbind(c(round(trt_all/ sum(trt_all),digits = 2)),
              c(round(trt/ sum(trt),digits = 2)[1:4],0,c(round(trt/ sum(trt),digits = 2)[5:8])))
barplot(counts, beside=TRUE, col=c("blue", "red"),
        ylab="Frecuencia relativa",
        legend=c("Grafo completo","Conglomerado 3"),
        args.legend = list(x = "topleft"))
```

- *FIGURA 38. Conglomerado 4 (UPCH) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")

load("graphs_data/wg_p.Rda")


library(igraph)


cl4=delete_vertices(wg_p, which(V(wg_p)$community!=4))

vertices_lab<-substr(V(cl4)$name,8,12)

V(cl4)$label <-vertices_lab


#1er gráfico (A)

col1<-rgb(228/255,26/255,28/255, alpha = 0.7)

col2<-rgb(55/255,126/255,184/255, alpha = 0.7)

col3<-rgb(77/255,175/255,74/255, alpha = 0.7)

col4<-rgb(152/255,78/255,163/255, alpha = 0.7)

col5<-rgb(255/255,127/255,0/255, alpha = 0.7)

col6<-rgb(255/255,255/255,51/255, alpha = 0.7)

col7<-rgb(166/255,86/255,40/255, alpha = 0.7)

col8<-rgb(247/255,129/255,191/255, alpha = 0.7)

col9<-rgb(153/255,153/255,153/255, alpha = 0.7)

colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)


plot(cl4, layout=layout_components, edge.curved=T,
```

```
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl4)$color,
    vertex.label.cex=(V(cl4)$fuerza)/380, vertex.size=(V(cl4)$fuerza/400), vertex.color="#FFFFFF")
legend(x="bottomright", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI"), pch=21,
      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)
title("Conglomerado 4")

#2do gráfico (B)
############
#Instituciones por tipo
######################
#De todo el grafo
all_institutions<-V(wg_p)$name
all_institutions_tip<-substr(all_institutions,4,6)

trt_all<-table(all_institutions_tip)
counts_all<-as.numeric(trt_all)
nombres_all<-names(trt_all)

#Instituciones involucradas en el congloemrado
all_institutions<-df$V.cl4..name
all_institutions_tip<-substr(all_institutions,4,6)

trt<-table(all_institutions_tip)
```

```
counts<-as.numeric(trt)
nombres<-names(trt)
counts<-rbind(c(round(trt_all/ sum(trt_all),digits = 2)),
        c(round(trt/ sum(trt),digits = 3)[1:3],0,0,c(round(trt/ sum(trt),digits = 2)[4:7])))
barplot(counts, beside=TRUE, col=c("blue", "red"),
    ylab="Frecuencia relativa",
    legend=c("Grafo completo","Conglomerado 4"),
    args.legend = list(x = "topleft"))
```

- *FIGURA 39. Conglomerado 5 (UNSA) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")
load("graphs_data/wg_p.Rda")

library(igraph)
cl5=delete_vertices(wg_p, which(V(wg_p)$community!=5))
vertices_lab<-substr(V(cl5)$name,8,12)
V(cl5)$label <-vertices_lab

col1<-rgb(228/255,26/255,28/255, alpha = 0.7)
col2<-rgb(55/255,126/255,184/255, alpha = 0.7)
col3<-rgb(77/255,175/255,74/255, alpha = 0.7)
col4<-rgb(152/255,78/255,163/255, alpha = 0.7)
```

```
col5<-rgb(255/255,127/255,0/255, alpha = 0.7)

col6<-rgb(255/255,255/255,51/255, alpha = 0.7)

col7<-rgb(166/255,86/255,40/255, alpha = 0.7)

col8<-rgb(247/255,129/255,191/255, alpha = 0.7)

col9<-rgb(153/255,153/255,153/255, alpha = 0.7)

colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)


plot(cl5, layout=layout_components, edge.curved=T,

    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl5)$color,

    vertex.label.cex=(V(cl5)$fuerza)/40, vertex.size=(V(cl5)$fuerza/50), vertex.color="#FF
FFFF")

legend(x="bottomright", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI")
, pch=21,

     col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)

title("Conglomerado 5")
```

- *FIGURA 40. Proporción de tipos de instituciones en el conglomerado 5 (UNSA) versus en el grafo completo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")

load("graphs_data/wg_p.Rda")


library(igraph)

cl5=delete_vertices(wg_p, which(V(wg_p)$community!=5))

#######
```

```
############

#Instituciones por sector y tipo

######################

#De todo el grafo

all_institutions<-V(wg_p)$name

all_institutions_sec<-substr(all_institutions,1,2)

all_institutions_tip<-substr(all_institutions,4,6)


trt_all<-table(all_institutions_tip)

counts_all<-as.numeric(trt_all)

nombres_all<-names(trt_all)


trs_all<-table(all_institutions_sec)

counts_all<-as.numeric(trs_all)

nombres_all<-names(trs_all)



#Instituciones involucradas en el congloemrado

all_institutions<-df$V.cl5..name

all_institutions_sec<-substr(all_institutions,1,2)

all_institutions_tip<-substr(all_institutions,4,6)


#Genero gráfico

par(mfrow=c(2,1))

par(las=2) # make label text perpendicular to axis
```

```
par(mar=c(5,5,4,2)) # increase y-axis margin.


trt<-table(all_institutions_tip)

counts<-as.numeric(trt)

nombres<-names(trt)

####

trs<-table(all_institutions_sec)

counts<-as.numeric(trs)

nombres<-names(trs)

counts <- rbind(c(round(trs_all/ sum(trs_all),digits = 2)),

                c(round(trs/ sum(trs),digits = 2)))

barplot(counts, beside=TRUE, col=c("blue", "red"),

        ylab="Frecuencia relativa",

        legend=c("Grafo completo","Conglomerado 5"),

        args.legend = list(x = "topleft"))


counts<-rbind(c(round(trt_all/ sum(trt_all),digits = 2)),

              c(round(trt/ sum(trt),digits = 3)[1:2],0,c(round(trt/ sum(trt),digits = 2)[3]),0,c(roun
d(trt/ sum(trt),digits = 2)[4:7])))

barplot(counts, beside=TRUE, col=c("blue", "red"),

        ylab="Frecuencia relativa",

        legend=c("Grafo completo","Conglomerado 5"),

        args.legend = list(x = "topleft"))
```

- *FIGURA 41. Conglomerado 1 (CIP) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")

load("graphs_data/wg_p.Rda")


library(igraph)

cl6=delete_vertices(wg_p, which(V(wg_p)$community!=6))

vertices_lab<-substr(V(cl6)$name,8,12)

V(cl6)$label <-vertices_lab


#1er gráfico (A)

col1<-rgb(228/255,26/255,28/255, alpha = 0.7)

col2<-rgb(55/255,126/255,184/255, alpha = 0.7)

col3<-rgb(77/255,175/255,74/255, alpha = 0.7)

col4<-rgb(152/255,78/255,163/255, alpha = 0.7)

col5<-rgb(255/255,127/255,0/255, alpha = 0.7)

col6<-rgb(255/255,255/255,51/255, alpha = 0.7)

col7<-rgb(166/255,86/255,40/255, alpha = 0.7)

col8<-rgb(247/255,129/255,191/255, alpha = 0.7)

col9<-rgb(153/255,153/255,153/255, alpha = 0.7)

colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)


plot(cl6, layout=layout_components, edge.curved=T,

    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl6)$color,
```

```
    vertex.label.cex=(V(cl6)$fuerza)/8, vertex.size=(V(cl6)$fuerza/10), vertex.color="#FFF
FFF")
legend(x="bottomright", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI")
, pch=21,
      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)
title("Conglomerado 6")


#2do gráfico (B)
############
#Instituciones por sector y tipo
######################
#De todo el grafo
all_institutions<-V(wg_p)$name
all_institutions_sec<-substr(all_institutions,1,2)
all_institutions_tip<-substr(all_institutions,4,6)

trt_all<-table(all_institutions_tip)
counts_all<-as.numeric(trt_all)
nombres_all<-names(trt_all)

trs_all<-table(all_institutions_sec)
counts_all<-as.numeric(trs_all)
nombres_all<-names(trs_all)

#Instituciones involucradas en el congloemrado
```

```
all_institutions<-df$V.cl6..name

all_institutions_sec<-substr(all_institutions,1,2)

all_institutions_tip<-substr(all_institutions,4,6)


trt<-table(all_institutions_tip)

counts<-as.numeric(trt)

nombres<-names(trt)


####

trs<-table(all_institutions_sec)

counts<-as.numeric(trs)

nombres<-names(trs)


##Genero gráfico

par(mfrow=c(1,2))


counts <- rbind(c(round(trs_all/ sum(trs_all),digits = 2)),

                c(0,round(trs/ sum(trs),digits = 2)))

barplot(counts, beside=TRUE, col=c("blue", "red"),

        ylab="Frecuencia relativa",

        legend=c("Grafo completo","Conglomerado 6"),

        args.legend = list(x = "topleft", bty="n"))


counts<-rbind(c(round(trt_all/ sum(trt_all),digits = 2)),
```

```
        c(round(trt/ sum(trt),digits = 3)[1],0,c(round(trt/ sum(trt),digits = 2)[2]),0,c(round(
trt/ sum(trt),digits = 2)[3:7])))

barplot(counts, beside=TRUE, col=c("blue", "red"),

    ylab="Frecuencia relativa",

    legend=c("Grafo completo","Conglomerado 6"),

    args.legend = list(x = "topleft", bty="n"))
```

- *FIGURA 42. Conglomerado 9 (DRSLO) en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")

load("graphs_data/wg_p.Rda")


library(igraph)

cl9=delete_vertices(wg_p, which(V(wg_p)$community!=9))

vertices_lab<-substr(V(cl9)$name,8,12)

V(cl9)$label <-vertices_lab


#1er gráfico (A)

col1<-rgb(228/255,26/255,28/255, alpha = 0.7)

col2<-rgb(55/255,126/255,184/255, alpha = 0.7)

col3<-rgb(77/255,175/255,74/255, alpha = 0.7)

col4<-rgb(152/255,78/255,163/255, alpha = 0.7)

col5<-rgb(255/255,127/255,0/255, alpha = 0.7)

col6<-rgb(255/255,255/255,51/255, alpha = 0.7)
```

```
col7<-rgb(166/255,86/255,40/255, alpha = 0.7)

col8<-rgb(247/255,129/255,191/255, alpha = 0.7)

col9<-rgb(153/255,153/255,153/255, alpha = 0.7)

colrs<-c(col1,col2,col3,col4,col5,col6,col7,col8,col9)


plot(cl9, layout=layout_components, edge.curved=T,

    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl9)$color,

    vertex.label.cex=(V(cl9)$fuerza)/23, vertex.size=(V(cl9)$fuerza/25), vertex.color="#FF
FFFF")

legend(x="bottomright", c("SAL","ONG", "UNI","ORI","EMP","GRL","GNO","IPI","FI")
, pch=21,

      col="#777777", pt.bg=colrs, pt.cex=2, cex=.8, bty="n", ncol=2)

title("Conglomerado 9")


#2do gráfico (B)

###########

#Instituciones por sector y tipo

######################

#De todo el grafo

all_institutions<-V(wg_p)$name

all_institutions_sec<-substr(all_institutions,1,2)

all_institutions_tip<-substr(all_institutions,4,6)


trt_all<-table(all_institutions_tip)

counts_all<-as.numeric(trt_all)
```

```
nombres_all<-names(trt_all)


trs_all<-table(all_institutions_sec)
counts_all<-as.numeric(trs_all)
nombres_all<-names(trs_all)


#Instituciones involucradas en el congloemrado
all_institutions<-df$V.cl9..name
all_institutions_sec<-substr(all_institutions,1,2)
all_institutions_tip<-substr(all_institutions,4,6)


#Genero gráfico
par(mfrow=c(2,1))
par(las=2) # make label text perpendicular to axis
par(mar=c(5,5,4,2)) # increase y-axis margin.


trt<-table(all_institutions_tip)
counts<-as.numeric(trt)
nombres<-names(trt)
####
trs<-table(all_institutions_sec)
counts<-as.numeric(trs)
nombres<-names(trs)


##Genero gráfico
```

```
par(mfrow=c(1,2))
counts <- rbind(c(round(trs_all/ sum(trs_all),digits = 2)),
                c(round(trs/ sum(trs),digits = 2)))
barplot(counts, beside=TRUE, col=c("blue", "red"),
        ylab="Frecuencia relativa",
        legend=c("Grafo completo","Conglomerado 9"),
        args.legend = list(x = "topleft", bty="n"))

counts<-rbind(c(round(trt_all/ sum(trt_all),digits = 2)),
              c(round(trt/ sum(trt),digits = 2)[1:2],0,round(trt/ sum(trt),digits = 2)[3:8]))
barplot(counts, beside=TRUE, col=c("blue", "red"),
        ylab="Frecuencia relativa",
        legend=c("Grafo completo","Conglomerado 9"),
        args.legend = list(x = "topleft", bty="n"))
```

- *FIGURA 43. Conglomerados pequeños en el grafo de coautorías de las instituciones peruanas con investigación en medicina indizada en Scopus, 2000-2015.*

```
load("graphs_data/cl.wg_p.Rda")
load("graphs_data/wg_p.Rda")

library(igraph)

##Conglomerado 7
```

```
cl7=delete_vertices(wg_p, which(V(wg_p)$community!=7))
vertices_lab<-substr(V(cl7)$name,8,12)
V(cl7)$label <-vertices_lab
plot(cl7, layout=layout_components, edge.curved=T,
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl7)$color,
    vertex.label.cex=(V(cl7)$fuerza)/18, vertex.size=(V(cl7)$fuerza/20), vertex.color="#FFFFFF")
title("Conglomerado 7")

##Conglomerado 8
cl8=delete_vertices(wg_p, which(V(wg_p)$community!=8))
vertices_lab<-substr(V(cl8)$name,8,12)
V(cl8)$label <-vertices_lab
plot(cl8, layout=layout_components, edge.curved=T,
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl8)$color,
    vertex.label.cex=(V(cl8)$fuerza)/28, vertex.size=(V(cl8)$fuerza/30), vertex.color="#FFFFFF")
title("Conglomerado 8")

##Conglomerado 10
cl10=delete_vertices(wg_p, which(V(wg_p)$community!=10))
vertices_lab<-substr(V(cl10)$name,8,12)
V(cl10)$label <-vertices_lab
plot(cl10, layout=layout_components, edge.curved=T,
```

```
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl10)$color
,
    vertex.label.cex=(V(cl10)$fuerza)/13, vertex.size=(V(cl10)$fuerza/15), vertex.color="#
FFFFFF")
title("Conglomerado 10")


##Conglomerado 11
cl11=delete_vertices(wg_p, which(V(wg_p)$community!=11))
vertices_lab<-substr(V(cl11)$name,8,12)
V(cl11)$label <-vertices_lab
plot(cl11, layout=layout_components, edge.curved=T,
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl11)$color
,
    vertex.label.cex=(V(cl11)$fuerza/3), vertex.size=(V(cl11)$fuerza/4), vertex.color="#FF
FFFF")
title("Conglomerado 11")


##Conglomerado 12
cl12=delete_vertices(wg_p, which(V(wg_p)$community!=12))
vertices_lab<-substr(V(cl12)$name,8,12)
V(cl12)$label <-vertices_lab
plot(cl12, layout=layout_components, edge.curved=T,
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(cl12)$color
,
    vertex.label.cex=(V(cl12)$fuerza/3), vertex.size=(V(cl12)$fuerza/4), vertex.color="#FF
FFFF")
```

```
title("Conglomerado 12")


##Conglomerado 13
Cl13=delete_vertices(wg_p, which(V(wg_p)$community!=13))
vertices_lab<-substr(V(Cl13)$name,8,12)
V(Cl13)$label <-vertices_lab
plot(Cl13, layout=layout_components, edge.curved=T,
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(Cl13)$colo
r,
    vertex.label.cex=(V(Cl13)$fuerza/3), vertex.size=(V(Cl13)$fuerza/4), vertex.color="#F
FFFFF")
title("Conglomerado 13")


##Conglomerado 14
Cl14=delete_vertices(wg_p, which(V(wg_p)$community!=14))
vertices_lab<-substr(V(Cl14)$name,8,12)
V(Cl14)$label <-vertices_lab
plot(Cl14, layout=layout_components, edge.curved=T,
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(Cl14)$colo
r,
    vertex.label.cex=(V(Cl14)$fuerza/3), vertex.size=(V(Cl14)$fuerza/4), vertex.color="#F
FFFFF")
title("Conglomerado 14")


##Conglomerado 15
Cl15=delete_vertices(wg_p, which(V(wg_p)$community!=15))
```

```
vertices_lab<-substr(V(Cl15)$name,8,12)
V(Cl15)$label <-vertices_lab
plot(Cl15, layout=layout_components, edge.curved=T,
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(Cl15)$color,
    vertex.label.cex=(V(Cl15)$fuerza/3), vertex.size=(V(Cl15)$fuerza/4), vertex.color="#FFFFFF")
title("Conglomerado 15")

##Conglomerado 16
Cl16=delete_vertices(wg_p, which(V(wg_p)$community!=16))
vertices_lab<-substr(V(Cl16)$name,8,12)
V(Cl16)$label <-vertices_lab

plot(Cl16, layout=layout_components, edge.curved=T,
    vertex.frame.color="#FFFFFF", vertex.shape="circle", vertex.label.color=V(Cl16)$color,
    vertex.label.cex=(V(Cl16)$fuerza/20), vertex.size=(V(Cl16)$fuerza/20), vertex.color="#FFFFFF")
title("Conglomerado 16")
```

- **Cuadros**
- *CUADRO 4. Vértices con mayor centralidad en el grafo*

```
#cargo los grafos
load("graphs_data/g.Rda")
```

```
load("graphs_data/wg.Rda")

load("graphs_data/wg_p.Rda")

load("graphs_data/cl.wg_p.Rda")


###Grado

library(igraph)


#ya tengo fuerza, grado, intermediacion, eigen_centr, falta cercanía

V(wg_p)$cercania<-closeness(wg_p, normalized = T)


#Tabla resumen

df<-data.frame(V(wg_p)$name,V(wg_p)$grado,V(wg_p)$fuerza,V(wg_p)$cercania,V(wg_p)$intermediacion,V(wg_p)$eigen_centr)

df_sorted<-df[with(df, order(V.wg_p..grado,decreasing = T)), ]

df_sorted[1:6,1:2]

df_sorted<-df[with(df, order(V.wg_p..fuerza,decreasing = T)), ]

df_sorted[1:6,c(1,3)]

df_sorted<-df[with(df, order(V.wg_p..cercania,decreasing = T)), ]

df_sorted[1:6,c(1,4)]

df_sorted<-df[with(df, order(V.wg_p..intermediacion,decreasing = T)), ]

df_sorted[1:6,c(1,5)]

df_sorted<-df[with(df, order(V.wg_p..eigen_centr,decreasing = T)), ]

df_sorted[1:6,c(1,6)]
```

- *CUADRO 5. Características del grafo de coautorías*

```
load("graphs_data/g.Rda")

load("graphs_data/wg.Rda")

g

wg

table(E(wg)$weight)

is.connected(wg)

is.connected(g)

clusters(wg)$no

clusters(wg)$csize

diameter(wg, weights=NA)
```

- *CUADRO 6. Cliques máximos en el grafo.*

```
#cargo los grafos

load("graphs_data/wg_p.Rda")


###Intermediación aristas

library(igraph)

table(sapply(cliques(wg_p), length))

#568 nodos, 2249 edges, 6034 triángulos, 6 cliques de 13

#Los cliques máximos son

cl13<-cliques(wg_p)[sapply(cliques(wg_p), length) == 13]

cl13
```

- *CUADRO 7. Coeficiente de asortatividad de los elementos del grafo.*

```
#cargo los grafos
load("graphs_data/wg_p.Rda")


###Intermediación aristas
library(igraph)


assortativity.nominal(wg_p, (V(wg_p)$tip=="EMP")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$tip=="FI_")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$tip=="GNO_")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$tip=="GRL_")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$tip=="IPI_")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$tip=="ONG")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$tip=="ORI")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$tip=="SAL")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$tip=="UNI")+1,
                      directed=FALSE)
```

```
assortativity.nominal(wg_p, (V(wg_p)$sec=="FI")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$sec=="PR")+1,
                      directed=FALSE)
assortativity.nominal(wg_p, (V(wg_p)$sec=="PU")+1,
                      directed=FALSE)
assortativity.degree(wg_p)
```

- *CUADRO 8. Vértices del conglomerado 1 y sus medidas de centralidad.*

```
cl1=delete_vertices(wg_p, which(V(wg_p)$community!=1))
df<-data.frame(V(cl1)$name,V(cl1)$fuerza,V(cl1)$intermediacion,V(cl1)$eigen_centr)
df_sorted<-df[with(df, order( V.cl1..eigen_centr,V.cl1..fuerza,V.cl1..intermediacion,decrea
sing = T)), ]
df_sorted
```

- *CUADRO 9. Vértices del conglomerado 2 y sus medidas de centralidad.*

```
cl2=delete_vertices(wg_p, which(V(wg_p)$community!=2))
df<-data.frame(V(cl2)$name,V(cl2)$fuerza,V(cl2)$intermediacion,V(cl2)$eigen_centr)
df_sorted<-df[with(df, order( V.cl2..eigen_centr,V.cl2..fuerza,V.cl2..intermediacion,decrea
sing = T)), ]
df_sorted
```

- *CUADRO 10. Vértices del conglomerado 3 y sus medidas de centralidad.*

```
cl3=delete_vertices(wg_p, which(V(wg_p)$community!=3))
df<-data.frame(V(cl3)$name,V(cl3)$fuerza,V(cl3)$intermediacion,V(cl3)$eigen_centr)
df_sorted<-df[with(df, order( V.cl3..eigen_centr,V.cl3..fuerza,V.cl3..intermediacion,decrea
sing = T)), ]
df_sorted
```

- *CUADRO 11. Vértices del conglomerado 4 y sus medidas de centralidad.*

```
cl4=delete_vertices(wg_p, which(V(wg_p)$community!=4))
df<-data.frame(V(cl4)$name,V(cl4)$fuerza,V(cl4)$intermediacion,V(cl4)$eigen_centr)
df_sorted<-df[with(df, order( V.cl4..eigen_centr,V.cl4..fuerza,V.cl4..intermediacion,decrea
sing = T)), ]
df_sorted
```

- *CUADRO 12. Vértices del conglomerado 5 y sus medidas de centralidad.*

```
cl5=delete_vertices(wg_p, which(V(wg_p)$community!=5))
df<-data.frame(V(cl5)$name,V(cl5)$fuerza,V(cl5)$intermediacion,V(cl5)$eigen_centr)
df_sorted<-df[with(df, order( V.cl5..eigen_centr,V.cl5..fuerza,V.cl5..intermediacion,decrea
sing = T)), ]
df_sorted
```

- *CUADRO 13. Vértices del conglomerado 6 y sus medidas de centralidad.*

```
cl6=delete_vertices(wg_p, which(V(wg_p)$community!=6))
df<-data.frame(V(cl6)$name,V(cl6)$fuerza,V(cl6)$intermediacion,V(cl6)$eigen_centr)
```

```
df_sorted<-df[with(df, order( V.cl6..eigen_centr,V.cl6..fuerza,V.cl6..intermediacion,decrea
sing = T)), ]

df_sorted
```

- *CUADRO 14. Vértices del conglomerado 7 y sus medidas de centralidad.*

```
cl7=delete_vertices(wg_p, which(V(wg_p)$community!=7))

df<-data.frame(V(cl7)$name,V(cl7)$fuerza,V(cl7)$intermediacion,V(cl7)$eigen_centr)

df_sorted<-df[with(df, order( V.cl7..eigen_centr,V.cl7..fuerza,V.cl7..intermediacion,decrea
sing = T)), ]

df_sorted
```

- *CUADRO 15. Vértices del conglomerado 8 y sus medidas de centralidad.*

```
cl8=delete_vertices(wg_p, which(V(wg_p)$community!=8))

df<-data.frame(V(cl8)$name,V(cl8)$fuerza,V(cl8)$intermediacion,V(cl8)$eigen_centr)

df_sorted<-df[with(df, order( V.cl8..eigen_centr,V.cl8..fuerza,V.cl8..intermediacion,decrea
sing = T)), ]

df_sorted
```

- *CUADRO 16. Vértices del conglomerado 9 y sus medidas de centralidad.*

```
cl9=delete_vertices(wg_p, which(V(wg_p)$community!=9))

df<-data.frame(V(cl9)$name,V(cl9)$fuerza,V(cl9)$intermediacion,V(cl9)$eigen_centr)

df_sorted<-df[with(df, order( V.cl9..eigen_centr,V.cl9..fuerza,V.cl9..intermediacion,decrea
sing = T)), ]
```

```
df_sorted
```

- *CUADRO 17. Vértices del conglomerado 10 y sus medidas de centralidad.*

```
cl10=delete_vertices(wg_p, which(V(wg_p)$community!=10))

df<-data.frame(V(cl10)$name,V(cl10)$fuerza,V(cl10)$intermediacion,V(cl10)$eigen_centr)

df_sorted<-df[with(df, order( V.cl10..eigen_centr,V.cl10..fuerza,V.cl10..intermediacion,decreasing = T)), ]

df_sorted
```

- *CUADRO 18. Vértices del conglomerado 11 y sus medidas de centralidad.*

```
cl11=delete_vertices(wg_p, which(V(wg_p)$community!=11))

df<-data.frame(V(cl11)$name,V(cl11)$fuerza,V(cl11)$intermediacion,V(cl11)$eigen_centr)

df_sorted<-df[with(df, order( V.cl11..eigen_centr,V.cl11..fuerza,V.cl11..intermediacion,decreasing = T)), ]

df_sorted
```

- *CUADRO 19. Vértices del conglomerado 12 y sus medidas de centralidad.*

```
cl12=delete_vertices(wg_p, which(V(wg_p)$community!=12))

df<-data.frame(V(cl12)$name,V(cl12)$fuerza,V(cl12)$intermediacion,V(cl12)$eigen_centr)

df_sorted<-df[with(df, order( V.cl12..eigen_centr,V.cl12..fuerza,V.cl12..intermediacion,decreasing = T)), ]
```

```
df_sorted
```

- *CUADRO 20. Vértices del conglomerado 13 y sus medidas de centralidad.*

```
cl13=delete_vertices(wg_p, which(V(wg_p)$community!=13))

df<-data.frame(V(cl13)$name,V(cl13)$fuerza,V(cl13)$intermediacion,V(cl13)$eigen_centr)

df_sorted<-df[with(df, order( V.cl13..eigen_centr,V.cl13..fuerza,V.cl13..intermediacion,decreasing = T)), ]

df_sorted
```

- *CUADRO 21. Vértices del conglomerado 14 y sus medidas de centralidad.*

```
cl14=delete_vertices(wg_p, which(V(wg_p)$community!=14))

df<-data.frame(V(cl14)$name,V(cl14)$fuerza,V(cl14)$intermediacion,V(cl14)$eigen_centr)

df_sorted<-df[with(df, order( V.cl14..eigen_centr,V.cl14..fuerza,V.cl14..intermediacion,decreasing = T)), ]

df_sorted
```

- *CUADRO 22. Vértices del conglomerado 15 y sus medidas de centralidad.*

```
cl15=delete_vertices(wg_p, which(V(wg_p)$community!=15))

df<-data.frame(V(cl15)$name,V(cl15)$fuerza,V(cl15)$intermediacion,V(cl15)$eigen_centr)

df_sorted<-df[with(df, order( V.cl15..eigen_centr,V.cl15..fuerza,V.cl15..intermediacion,decreasing = T)), ]
```

```
df_sorted
```

- *CUADRO 23. Vértices del conglomerado 16 y sus medidas de centralidad.*

```
cl16=delete_vertices(wg_p, which(V(wg_p)$community!=16))

df<-data.frame(V(cl16)$name,V(cl16)$fuerza,V(cl16)$intermediacion,V(cl16)$eigen_centr)

df_sorted<-df[with(df, order( V.cl16..eigen_centr,V.cl16..fuerza,V.cl16..intermediacion,decreasing = T)), ]

df_sorted
```