CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

Unidad Zacatenco

Indicadores Cienciométricos y Modelos Matemáticos en Bibliometría

Tesis que presenta

Juan Antonio Pichardo Corpus

para obtener el Grado de
Doctor en Ciencias

en la Especialidad de
Desarrollo Científico y Tecnológico para la Sociedad

Directores de Tesis: Dr Jesús Guillermo Contreras Nuño
Dr José Antonio Stephan de la Peña Mena

Ciudad de México Agosto de 2016

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología por la beca otorgada durante el doctorado.

Índice general

Resumen

Esta tesis presenta algunos indicadores bibliométricos que trastocan la cienciometría y con potenciales más allá de ésta, como la política científica. Se llevan a cabo análisis que van de lo más general e internacional a niveles más locales para el caso mexicano.

Primero se presenta un indicador basado en los errores publicados por revistas científicas en las principales bases de datos como Scopus y Web of Science (WoS), se identifican las áreas de investigación con proporciones de errores y se encuentra una correlación entre el factor de impacto y la proporción de errores en revistas de física, ingeniería, matemáticas y multidisciplinarias.

Enseguida se propone un modelo matemático para estudiar el impacto o la influencia de un artículo científico, partiendo de la red de citación y de su matriz de adyacencia, se presenta una medida basada en la exponencial de la matriz. La medida se calcula en varias redes de citación y se compara con otras medidas usuales, permitiendo identificar correlaciones y posibles usos de la propuesta en conjunto con otra u otras medidas.

Luego, en un análisis más local, con datos correspondientes a México en tres áreas: física, matemáticas y química, se analiza la evolución de estas áreas por medio de las redes de citación correspondientes. Se identifican tendencias y modelos útiles, como el cambio de fase para que emerja una componente gigante en la red o el ajuste de una distribución a las citas, así como una comparación entre las citas internas o locales y las externas o globales.

Finalmente se analiza el caso del científico Mario Molina, estudiando la evolución de la carrera del investigador dentro del ámbito científico por medio de sus redes de coautoria y el impacto de su investigación fuera de la academia. Se muestran las implicaciones de la obtención del premio y Nobel en sus interacciones dentro y fuera del ámbito académico. Y se propone una caracterización propia de la historia de las ciencias.

Abstract

This thesis presents some bibliometric indicators that disrupt scientometrics and with potential beyond this, as scientific policy. Analysis were performed ranging from general data (international) to a more specific level like the Mexican case.

We study the frequency of errata in the scientific literature and focus on the 2000-2014 period for a more detailed view of its characteristics. First we compare the frequency of errata in different fields, then we examine in detail three leading journals and correlate the quality of a journal, using the impact factor as a metric, with the amount of errata it publishes. We find that errata could be useful as an indicator to highlight differences in the practice of science across areas of knowledge. We also argue that errata could yield relevant additional information to assess the performance of journals.

A mathematical model is also proposed to study the impact or influence of a paper, based on the citation network and its adjacency matrix, a measure based on the exponential of the matrix is presented. The measure is calculated on different citation networks and compared to other usual measures in order to identify correlations and possible uses of the proposal, jointly with another or others measures.

It is analyzed the evolution of a citation network in physics, mathematics and chemistry, in papers with at least one author affiliated to a Mexican institution. Trends and useful models were identified, such as phase change for the emergence of a giant component in the network or a power law or log-norm fits for the data, then a comparison between local and global citations is also made.

Finally it is analyzed the Mario Molina case, by studing the evolution of his coauthors networks, searching for the scientific impact of Molina's in an out the academy, like the Nobel Prize and Mario Molina Center. It was made a model for Molina's scientific carrer based on history of science.

Introducción

El análisis de los datos relacionados con la producción científica dio origen a la llamada ciencia de la ciencia o cineciometría y de manera casi paralela a la bibliometría. Data de la década de los años 60 del siglo pasado, con los trabajos de Derek de Solla Price [1] y Eugene Garfield [2], dando pie a la creación del Instituto para la Información Científica (Institute for Scientific Information, ISI), actualmente parte de la empresa Thomson Reuters. El ISI, en ese momento, desarrolló un catálogo de las principales revistas de ciencia (SCI) y creó el llamado Factor de Impacto (FI) de una revista, dando origen a una de las medidas más discutidas en la actualidad.

La investigación en esta tesis se centra en el análisis de datos obtenidos de la producción científica en varias vertientes. En el capítulo 1 se da una breve descripción de los tipos de medidas que se han desarrollado en torno a la evaluación de la ciencia; de esta manera se contextualiza la investigación realizada planteando los objetivos del trabajo.

En el capítulo 2 se analizan los errores en la ciencia. Partiendo de la idea que los errores honestos son parte del proceso científico y que la corrección de errores se hace en artículos llamados, entre otros nombres, erratum y pese a que la capacidad de autocorrección es un componente crucial de la ciencia, hay relativamente poca investigación sobre la cienciometría de las erratas. En ese sentido se estudia la frecuencia de erratas en la literatura científica, centrándose en el período 2000-2014 para tener una vista más detallada de algunas tendencias. En primer lugar se compara la frecuencia de erratas en diferentes campos; a continuación se examinan en detalle tres revistas de corriente principal y, por último, se correlaciona la calidad de una revista, usando el factor de impacto como una métrica, con la cantidad de erratas que se publica.

En el capítulo 3 se plantea y desarrolla una propuesta para medir la influencia de un artículo en la red de citación. Se hace una revisión de los conceptos básicos para el análisis de redes así como de las medidas asociadas a un nodo en una red y a partir de ello se presenta la medida con base en la exponencial de la matriz de adyacencia de la gráfica (red). También se aplica la medida a la red de citación generada entre los artículos de las revistas de la American Physical Society y se compara con otras medidas haciendo una análisis de correlación, motivando una discusión matemática de las medidas y generando una interpretación muy sugerente de las correlaciones. Así, se muestra la pertinencia de la medida propuesta como un complemento a las

medidas usadas normalmente.

En el capítulo 4 se hace un análisis de redes de citación entre artículos con al menos un coautor con adscripción a una institución mexicana. Las redes se construyen para tres áreas: física, matemáticas y química, en cada una se generan redes en el tiempo para periodos de 5 años acumulados, desde 1970 hasta 2015. Los datos con que se generan las redes son de la Web of Science. Se hacen análisis de distribuciones de citas encontrando diferencias evidentes entre cada área, así como una comparación de las citas locales contra las globales y finalmente se calcula la influencia en estas redes. Teniendo así un acercamiento a la ciencia mexicana desde una perspectiva poco estudiada en el país y con potencial para la toma de decisiones en la política científica del país.

El capítulo (cap. 5) es una mirada a un caso de estudio centrado en un científico mexicano ganador del Premio Nobel de Química, Mario Molina, y la relación de la investigación del científico con la sociedad fuera del ámbito académico. Se muestran las redes de coautoría con base en los artículos de Mario Molina registrados en la Web of Science, para distintos intervalos de tiempo, mostrando el impacto de ganar el Premio Nobel y como una manera de contextualizar los análisis posteriores. Como eje de análisis se toma el artículo por el cual ganó el Premio Nobel y su relación con la pérdida de la capa del ozono. Se revisa cómo se produjo esta investigación y algunas de sus consecuencias en otros sectores de la sociedad, fuera del ámbito académico-científico. A través de esta mirada particular, se caracteriza la relación ciencia-sociedad mediante tres ejes de análisis: infraestructura, liderazgo y flexibilidad. Partiendo de esta caracterización se reflexiona sobre las oportunidades para Latinoamérica como la interacción entre gobierno, academia e industria, además de la inversión en ciencia y tecnología y los proyectos a largo plazo.

Capítulo 1

Planteamiento de la investigación

1.1. Sobre la medición de la actividad científica

Sólo como referencia, por bibliometría se entienden los datos asociados a las publicaciones científicas, y las estadísticas derivadas de ellos [3]. La cienciometría hace uso de la bibliometría con datos de otras áreas que permean la actividad científica y pueden ser determinantes en su desarrollo, tales como: número de investigadores, su distribución geográfica o por especialidad, fuentes de financiamiento, repercusión, etc. [4].

De acuerdo con la empresa Thomson Reuters [5] las medidas relacionadas con la productividad científica se pueden englobar en siete tipos:

1. Productividad
2. Reconocimiento total o influencia
3. Reconocimiento indirecto o influencia indirecta
4. Eficiencia
5. Impacto relativo o evaluación comparativa
6. Especialización
7. Análisis de tendencias

El conteo de artículos es la principal medida de productividad, el conteo de citas o el índice h son ejemplos del tipo 2, el conteo de citas de segunda generación es un ejemplo del tipo 3, el promedio de citas por artículo y el factor de impacto de una revista son ejemplos del tipo 4, el número de artículos citados sobre los no citados es un ejemplo de las medidas de tipo 5, los indicadores de colaboración son ejemplos del tipo 6, finalmente las series de tiempo son ejemplos del tipo 7 [5].

La investigación toca varios de las medidas mencionadas y otras de distinta índole. En principio se partirá de algo similar al conteo de citas de segunda generación, aunque se espera indagar en las redes de citación y las tendencias de algunos indicadores particulares.

A continuación se realiza una descripción más detallada de algunas medidas.

1.1.1. Para una revista

El FI de una revista es un número que se calcula a partir de las citas recibidas por los artículos publicados en la revista en un periodo de dos años. Por ejemplo para calcular el FI de una revistas en 2013 se cuentan las citas recibidas por los artículos publicados en el periodo 2011-2012 y se divide entre el número de artículos publicados en la revista en el mismo periodo.

En [3] critican fuertemente la comparación entre revistas por medio del FI, ya que las revistas pertenecen a ciertos campos disciplinares, y en cada disciplina se tienen formas de citación específicas; por ejemplo, en el año 2000 el promedio de citas de un artículo de matemáticas era de 1 mientras que ciencias de la vida superaba las 6.

En algunos casos se extrapola el FI hacia un artículo o incluso hacia un científico; entonces si un científico publica en una revista con alto impacto, su artículo sería un buen artículo y también el mismo científico. Sin embargo, en promedio, el 90 % de las citas de una revista se deben al 10 % de los artículos [5].

Aunado a ello para algunas áreas del conocimiento como matemáticas puede pasar que en el periodo de dos años los artículos sólo consigan el 10 % de las citas totales [3], entonces el FI no considera el 90 % de las citas.

Por supuesto, el FI puede ser útil si se toman las precauciones debidas para evitar comparaciones sin sentido o llevar ese número a una descontextualización y tomar decisiones con un margen de error grande.

Una alternativa al FI es el eigenfactor (http://www.eigenfactor.org/) que calcula un factor de impacto basado en el algoritmo Page Rank de Google [6]. En este enfoque una revista se considera influyente si es citada a menudo por otras revistas consideradas influyentes; además, considera el promedio de tiempo que un lector tiene que seguir hacia atrás la cadena de citas de los artículos de la revista. El algoritmo hace uso de la teoría de matrices y redes, de ahí un poco el nombre de la organización ya que hace uso de los eigenvalores o valores propios de la matriz de adyacencia que se genera.

Uno de los resultados de esta tesis es un indicador con base en las erratas; sin embargo, no es tan particular como el FI, más bien se presenta una clasificación entre áreas y se dan ejemplos de algunas revistas.

1.1.2. Para un artículo

Uno de los números con mayor impacto en la evaluación de las publicaciones de un científico es el conteo de citas y las citas promedio. Son varias las críticas

al respecto, ya que las citas a un artículo pueden darse por una gran variedad de razones, no necesariamente asociadas a la calidad del artículo. Entre esas razones se encuentran: el reconocimiento de una deuda intelectual, una cita retórica, la claridad de la exposición, la visibilidad del propio artículo, entre otros [3, 7].

Una de las formas para ir más allá en el impacto de un artículo científico es analizar las redes de citación entre artículos científicos. Para ello se han propuesto varias medidas, que se pueden obtener de la red de manera casi inmediata, como la cercanía o la intermediación, otras basadas en los valores y vectores propios de la matriz de adyacencia de la red [8]. También se han propuesto rankings basados en el Page Rank [9, 10], aunque no son claras las diferencias en la práctica con la obtención de los valores propios. De la Peña [7] considera las propiedades globales de la red de citación de artículos científicos para generar una función de impacto en un conjunto de artículos. Luego, esa función proporcionará una medida del impacto de un artículo, esta idea se ha puesto en práctica en redes complejas generando funciones basadas en la exponencial de una matriz [11–14], pero no en redes de citación.

En esta tesis se presenta una propuesta de investigación similar: uno de los objetivos centrales es medir el impacto o la influencia de un artículo científico en la red de citación usando modelos matemáticos asociados al análisis de redes.

1.1.3. Para un científico

La principal forma de cuantificar la productividad de un científico es la cantidad de artículos que escribe por año y luego el impacto de esa producción que se mide a través del número de citas que reciben los artículos o el promedio de las citas por artículo. Se sabe que el significado de las citas no es simple y las estadísticas basadas en ellas no son tan "objetivas" como algunos afirman [3].

Una medida que intenta cuantificar producción e impacto en un solo número es el índice h [15] con el objetivo de caracterizar la producción científica de un investigador. Un científico tiene índice h si tiene h artículos con al menos h citas cada uno.

La facilidad para calcular el índice h ha favorecido su aceptación, en efecto, basta con ordenar las publicaciones de un autor de mayor a menor frecuencia de citación y dónde coincide el número de citas con el número de artículo se tiene el índice h; actualmente, en la WoS o en Scopus ese número se obtiene directamente en la búsqueda de algún autor en su base de datos.

El índice h tuvo una gran influencia en las publicaciones sobre bibliometría y/o cienciometría [16]; hay quienes lo consideran una buena medida [17] otros hacen críticas [18, 19] y algunos más lo han complementado o extendido [16].

En [3] se menciona que la mayoría de trabajos sobre la confiablidad del índice h

están basados en la correlación con otras medidas, lo cual no es nada especial, ya que todas las variables son función de un mismo fenómeno básico: las publicaciones, y proponen realizar estudios más profundos al respecto.

Parte del trabajo en esta tesis analiza las redes de coautoría del científico Mario Molina y algunos datos asociados a su productividad, pero más que dar una medida para el científico se revisa su carrera científica y su impacto en la academia, pero sobre todo, fuera de ella; tomando como eje de análisis el problema del daño a la capa del ozono.

1.1.4. Análsis de tendencias

Un tema recurrente en la literatura de los últimos años, sobre índices o medidas bibliométricas, es su capacidad predictiva, aunque en general se ha encontrado que es muy limitada.

En [20] se pone a prueba la hipótesis de que el conteo de citas es un indicador fiable de éxito en el futuro y compara 10 medidas distintas. La conclusión más contundente es que el número de citas a trabajos futuros es muy difícil de predecir.

En [21] intentan medir el impacto científico en el largo plazo, también comparan varias medidas. Muestran que el FI es un predictor pobre de futuras citas para un artículo. El número de citas tampoco es un buen predictor, ya que para artículos con la misma cantidad de citas en 5 años tienen impactos muy diferentes en el largo plazo. De manera similar el índice h.

Posiblemente la limitación en la predicción está relacionada con las características inherentes a las medidas mismas, toda vez que intentan dar cuenta de un fenómeno complejo que es difícil reducir a un número.

Otro de los elementos de esta tesis es analizar tendencias, pero más que para predecir, para entender parte del desarrollo de algunos campos de investigación en México, utilizando las redes de citación de física, matemáticas y química.

1.1.5. Análisis espacial

Una de las tendencias en varias investigaciones sobre cienciometría o bibliometría son los análisis espaciales; en [22] se hace una revisión sobre el estado del arte de ello, se propone una agenda de investigación que se ha venido siguiendo en torno a tres componentes centrales: la distribución espacial de la investigación y las citas, la existencia de sesgos espaciales de colaboración, citas y movilidad, y el impacto de las citas nacionales frente a las colaboraciones internacionales. En [22] también se propone el concepto de “proximidad” como un marco único para combinar hipótesis desde diversas posturas teóricas.

Uno de los acercamientos hacia el análisis espacial combinado con factores económicos y de crecimiento de un país se hace en [23]; se presenta un análisis sistemático de las redes de citación y colaboración entre ciudades y países, mediante la asignación de artículos para las ubicaciones geográficas de las afiliaciones de los autores. En [24] se hace un análisis de cómo cambia la producción y el impacto de ésta, medido en citas, con base en el movimiento de un científico a nivel geográfico de acuerdo a su institución de adscripción.

En síntesis, hay varias medidas que intentan cuantificar la calidad del trabajo científico de un investigador; en algunos casos esos índices se podrían utilizar en conjunto con otras formas de medir o analizar la ciencia en su conjunto y el desempeño de un científico en particular. Para el caso de la política científica, a nivel micro (individual) se pueden identificar tres usos de estos indicadores: a nivel del reconocimiento de una carrera (p. ej. otorgar un premio), otorgamiento de plazas y apoyo de proyectos; estos a nivel individual o micro. Hay otros usos a nivel meso (institucional) o macro (país) para evaluar programas de desarrollo científico [25, 26].

Cada uno de los usos o propósitos requiere diferente capacidad de predicción: para el reconocimiento ninguna; para el otorgamiento de plazas a largo plazo, para apoyo de proyectos a corto o mediano plazo, para evaluaciones institucionales y nacionales a corto, mediano y largo plazo.

1.1.6. Las redes como modelo

Ya se mencionó que una manera de cuantificar el impacto de un artículo científico es a partir de la red de citación correspondiente; sin embargo, las redes son un modelo de análisis que va más allá de las redes de citación, ya que se pueden construir redes de coautoría entre científicos, redes de citación entre revistas, redes entre instituciones, entre países, etc. Y no sólo para obtener una medida sino para indagar sobre las interacciones a partir de tales redes; en esa dirección, en esta investigación se usan de las redes de citación. A continuación se presentan algunas generalidades sobre las redes y en particular su uso en en el análisis de la producción científica.

Las redes se pueden estudiar en teoría de gráficas, como un objeto matemático llamado gráfica. Una gráfica $G = (V, E)$ consiste de un conjunto no vacío y finito de vértices V y un conjunto E de pares de vértices distintos, llamados aristas. En términos de redes, normalmente a los vértices se les llama nodos y a las aristas enlaces. En física normalmente se refieren a las redes que se estudian como "redes complejas" (complex networks) termino acuñado desde el estudio de los sistemas dinámicos y/o complejos [27], también se habla de la ciencia de las redes (network science) [28].

Una de las investigaciones pioneras en usar redes para analizar la actividad científica fue [1] haciendo un análisis de las redes de citación entre artículos científicos y las revistas donde se publicaban. Mostró que un 35 % de los artículos no tenían citas y que otro 45 % sólo tenía una cita. Porcentajes que sorpresivamente siguen presentes en muchas de las redes de citación recientes [29].

En teoría de gráficas, uno de los modelos más estudiados es el de Erdös-Renyi [30], el cual se desarrolló para gráficas aleatorias, se denota por modelo *ER*. Sin embargo la mayoría de las redes reales con grandes volúmenes de datos, como las redes de citación, no se comportan como el modelo *ER*, no son aleatorias; sino libres de escala [28].

Una característica importante de las redes reales es que tienen la propiedad de "mundo pequeño" también conocida como "seis grados de separación", debido a que si dos personas están conectadas en el mundo se puede llegar de una a la otra pasando, en promedio, por seis personas [31] lo cual está estrechamente relacionado con el coeficiente de clustering o de agrupamiento que mide el grado en el que los nodos de una red tienden a agruparse.

En [32] se estudiaron las características de redes de coautoría de científicos, se encontró que las redes tenían la propiedad de mundo pequeño y que el coeficiente de agrupamiento era grande, a diferencia de la redes tipo ER.

En [33] se mostró que las redes de coautoría entre matemáticos y neurocientíficos correspondientes al periodo 1991-1998 eran redes libres de escala y que la evolución de las redes está determinada por el "apego preferencial"; esto significa que un autor con muchos coautores tendrá más coautores en el futuro que uno con menos coautores, en general esta es una característica de las redes libres de escala y el modelo de Barabási-Albert [34]. También en [33] se encontró que el grado promedio de los nodos se incrementa con el tiempo y que la separación de los nodos decrece, explicando la propiedad de mundo pequeño. Inversamente, el coeficiente de clustering o de agrupamiento decae con el tiempo, esto está relacionado con la probabilidad de que dos autores escribieran un artículo conjunto.

Otra vertiente en el análisis de redes es sobre la distribución de probabilidad asociada al grado de un nodo. Una característica de las redes libres de escala implica que su distribución se puede modelar como una ley de potencia. Particularmente en las redes de citación una buena parte de la discusión se ha centrado entre si la distribución de citas es una log-norm, ley de potencia, exponencial, ley de potencia con corte exponencial, entre otras [28, 29, 35–37].

En [29] se mostró que el mejor ajuste para las citas entre 353,268 artículos de la familia *PhysicalReview* era una log-norm.

En [36] se mostró que el ajuste de ley de potencia era moderado para las citas

recibidas desde su publicación y hasta junio de 1997 por 415 229 artículos publicados en 1981, listados en el Science Citation Index; además, ese ajuste no era estadísticamente mejor que el de log-norm.

En [37] se concluye que la hipótesis de ley de potencia sólo es válida para ciertas áreas del conocimiento, y que sólo es mejor que otros ajustes para física.

1.2. Sobre la investigación

Después de la breve revisión de algunas medidas sobre la producción científica y partiendo del modelo de redes, se tiene el contexto donde se inserta esta investigación. Desde la perspectiva del análisis meramente estadístico y descriptivo hasta los modelos mas elaborados para construir indicadores o medidas asociadas a un artículo, científico, institución, etc. Se presenta una investigación que parte de análisis muy generales y llega a otros muy particulares.

La conveniencia de analizar la productividad de la actividad científica, tanto internacional como nacional, no sólo permite entender interacciones entre científicos y/o sus productos sino contribuir a una compresión más integral del desarrollo de la ciencia y abonar a una toma de decisiones centrada en la calidad y no reemplazarla por números, sólo por que se pueden medir; sino entender las limitaciones y bondades de tales medidas.

El análisis de datos de las publicaciones científicas y su uso en el diseño o apoyo a la política científica, es un tema de discusión actual e importante; porque de ello puede depender la asignación recursos para proyectos, infraestructura, el otorgamiento de reconocimientos o la incorporación de un científico a una institución.

También permite entender parte del funcionamiento de la ciencia, diferenciado prácticas de colaboración, citación, de revisión, de corrección, entre otras. Identificando cambios de paradigmas y/o el impacto de estos. En ese sentido se pueden evitar investigaciones que pueden carecer de los elementos necesarios para ser útil o verdadera, como se plantea en [38, 39].

Además, los desarrollos en la comprensión de las redes de coautoría y/o citación han dotado de nuevos conceptos, técnicas y herramientas a la ciencia de las redes, en ese sentido, el diseño de análisis complementarios es útil no sólo para la bibliometría [28].

Sobre el caso mexicano, las investigaciones que se han realizado están en el contexto de las coautorias y de los impactos medidos en citas [40–42], así como de algunos patrones de producción [43, 44] entre otras. Sin embargo, sobre las redes de citación entre artículos con al menos un autor con adscripción a una institución mexicana, no hay investigaciones, y en ese sentido es importante tener un análisis de la evolución

de tales redes. Primero, como memoria histórica de la ciencia mexicana, y en segundo lugar como una herramienta de análisis para los tomadores de decisiones, tanto científicos como burócratas o políticos.

La caracterización de casos de éxito, como el del científico Mario Molina, es importante para entender los elementos que debiera considerar una política científica que buscara tener resultados similares.

1.3. Objetivos

En el marco del análisis de datos de la producción científica, y la ciencia misma, se analizaron cuatro vertientes.

La primera a partir de datos poco explorados hasta el momento, basados en los errores (erratas) que se comenten en las publicaciones científicas. La segunda es a través de las redes de citación, identificando un modelo para medir la influencia de un artículo científico en la red correspondiente. La tercera, a partir del desarrollo científico de algunas áreas viendo la evolución de las redes de citación para el caso mexicano. La cuarta a partir de la interacción entre la ciencia y la sociedad, tomando como ejemplo al científico Mario Molina y analizar sus interacciones desde el punto de vista de las redes de coautoría y el impacto fuera del ámbito académico.

Así, se plantearon los siguientes objetivos.

1.3.1. Objetivos específicos I

Analizar la proporción de errores en la producción científica mundial teniendo como propósitos:

1. Identificar tendencias en la frecuencia de errores por áreas del conocimiento y en algunas revistas en particular.

2. Establecer similitudes y diferencias entres las áreas a partir de los errores.

3. Comparar la proporción de errores en revistas con el factor de impacto.

4. Generar un indicador cienciométrico basado en la proporción de errores.

1.3.2. Objetivos específicos II

Se propone un modelo para analizar las principales medidas asociadas a un artículo en una red de citación, teniendo como propósitos:

1. Diseñar una medida que capture información distinta y relevante de la influencia o el impacto de un artículo en una red de citación.

2. Comparar las principales medidas asociadas a un nodo en una red con el modelo.

3. Aplicar y analizar la propuesta en una red real representativa a nivel mundial.

1.3.3. Objetivos específicos III

Se generaron y analizaron redes de citación entre artículos con al menos un autor adscrito a una institución mexicana, teniendo como propósitos:

1. Identificar patrones en la evolución de las redes de citación en el tiempo.

2. Realizar ajustes a las distribuciones de citas.

3. Comparar la citación interna contra la externa.

4. Analizar campos distintos: física, matemáticas y química.

1.3.4. Objetivos específicos IV

A partir del caso Mario Molina, se generaron y analizaron sus redes de coautoría en el tiempo así como datos relacionados con la actividad extra académica del científico, teniendo como propósitos:

1. Identificar cambios en sus colaboraciones a partir de la obtención del Premio Nobel.

2. Entrelazar la actividad extra académica con la obtención del Premio Nobel.

3. Caracterizar la relación ciencia-sociedad que ha desarrollado el científico.

Capítulo 2

Erratas en la ciencia

Este capítulo está basado en el artículo «Errata as a Scientometric Indicator» que ha sido enviado a publicación [45].

Las razones para estudiar las erratas son varias: desde el punto de vista bibliométrico o cienciométrico, las erratas son un objeto de estudio, el proceso de corrección es propio de la ciencia y en particular de las publicaciones científicas; hay pocas publicaciones al respecto y se puede arrojar información sobre el quehacer científico relacionado con este tópico.

Los errores honestos son parte del proceso científico. La corrección de errores publicados se realiza en artículos científicos llamados, entre otros nombres, *erratum* o *errata*.

La máxima *Errare humanum est, sed perseverare diabolicum* (Errar es humano, pero perseverar en el error es diabólico), atribuida a Seneca, ha sido acogida por la ciencia. Un artículo publicado es una invitación a una discusión abierta y una verificación cruzada de ideas y resultados. La cuidadosa exposición de las suposiciones y las razones para usarlas desencadenan la propuesta de nuevas ideas en un proceso continuo de correcciones y mejoras (véase [46] y sus referencias). En este contexto, las teorías pueden ser sustituidas en algunos casos por otras más generales, sin perder su validez para los casos particulares que fueron creadas, y por lo tanto sin ser consideradas necesariamente como erróneas. Algo similar ocurre con los datos experimentales, que pueden ser reemplazados por datos más precisos y no ser considerados erróneos, incluso si los nuevos datos permiten extraer conclusiones diferentes que con los anteriores.

Para que este proceso sea efectivo, los errores tienen que ser evitados tanto como sea posible, y cuando ocurren tienen que ser corregidos. Desde el punto de vista bibliográfico, este último paso se lleva a cabo mediante la publicación de *erratas*, dedicadas específicamente a corregir errores en publicaciones anteriores [47].

A pesar de que ésta es una parte fundamental en el proceso de publicación científica, no hay mucha investigación cienciométrica sobre este tema. La mayoría de los estudios se han realizado sobre el papel de las retracciones en la literatura médica más que en erratas. Las retracciones son artículos que han sido retirados de las revistas

debido a errores insalvables ya sean errores honestos (sin intención) o premeditados (como un engaño).

Una manera de diferenciar los errores honestos de los premeditados en las retracciones, es revisando si el artículo fue retirado por uno de los autores o por el editor de la revista u otros expertos. En el primer caso se consideraría un error honesto y en el segundo no. Esto se aplica a los casos de los hermanos Bodganoff [48] y de Schön [49], en ambos los errores fueron premeditados, salvo que Schön admitió haber falsificado los datos y los hermanos Bodganoff siguen defendiendo su tesis.

Las investigaciones previas han mostrado que los artículos retirados se siguen citando [50], que menos de la mitad de los artículos retirados son debidos a errores honestos o sin intención [51] y que las erratas en las revistas con alto impacto están correlacionadas (0.86) con el factor de impacto (FI) [52].

En este marco se analizan las erratas, primero desde una perspectiva global, para luego centrarse en unas áreas en particular, luego se profundiza más para analizar las erratas en algunas revistas. En general se identifican tendencias y una posible caracterización de los campos científicos a partir de las erratas.

2.1. Erratas a nivel global

Para tener una visión general del uso de erratas en los campos del conocimiento a través del tiempo, se han utilizado las 27 áreas definidas por la base de datos de Scopus (ver apéndice B). Una ventaja de este enfoque es que el número de artículos publicados y las erratas son suficientes para considerar los resultados independientes de fluctuaciones temporales. En la tabla 2.1 se resumen algunos aspectos de estos datos, mostrando el número de erratas y el porcentaje de erratas en relación con el número total de publicaciones en cuatro periodos: antes de 1900, la primera y segunda mitades de el siglo XX, y de 2000 a 2014.

Una primera observación es que a pesar del gran aumento en el número de artículos publicados desde 1900 hasta la actualidad, el porcentaje de erratas no ha cambiado mucho con el tiempo. En el último período que se muestra en la tabla 2.1, el más bajo (más alto) porcentaje de erratas es 0.11 % (1.46 %) se encuentra en COMP (MULT), mientras que la mayoría de las otras áreas están entre 0.2 y 1.0 %. No obstante, la tabla 2.1 también muestra que, dentro de estos porcentajes, la tasa de erratas cambia significativamente con el tiempo. De la primera a la segunda mitad del siglo pasado todas menos dos de las áreas redujeron significativamente la publicación de erratas. Este comportamiento contrasta con lo que se observa en este siglo: la mayor variación en un área determinada, a partir de la comparación de las dos últimas columnas de la tabla, se encuentra en MEDI, donde el porcentaje de erratas aumentó en más de

	Año de	<1900		1900-1949		1950-1999		2000-2014	
Área	inicio	Errata	%	Errata	%	Errata	%	Errata	%
AGRI	1857	137	1.54	463	1.31	4916	0.45	14333	0.68
ARTS	1857	1088	0.57	1408	0.92	561	0.37	2770	0.29
PHAR	1857	1116	0.59	1472	0.95	6607	0.61	8108	0.79
SOCI	1858	1041	0.56	1396	0.87	1702	0.27	6637	0.32
MEDI	1860	404	0.39	1256	0.28	24606	0.24	66770	0.80
BIOC	1863	118	0.91	637	0.87	21186	0.64	34785	0.99
CHEM	1875	311	2.23	1283	1.31	13646	0.83	14508	0.64
MATH	1875	125	1.50	470	1.41	3791	0.61	6077	0.37
ENGI	1897	4	0.80	261	1.07	3851	0.10	9360	0.16
PHYS	1910			642	1.45	18065	0.86	19024	0.56
ENVI	1920			108	0.70	3668	0.35	6820	0.52
IMMU	1920			136	1.34	4973	0.57	9638	1.06
MULT	1920			225	0.29	4255	1.23	4293	1.46
DECI	1930			48	0.65	1411	0.80	754	0.32
EART	1930			49	1.00	2270	0.23	6422	0.49
MATE	1930			86	3.14	3638	0.30	8724	0.34
CENG	1940			37	0.65	2507	0.39	4754	0.37
DENT	1950					270	0.18	894	0.48
NEUR	1950					4038	0.58	7221	0.92
PSYC	1950					1439	0.32	3627	0.57
ECON	1952					425	0.35	1468	0.32
HEAL	1955					1365	0.43	3183	0.79
ENER	1956					708	0.22	1581	0.20
COMP	1960					824	0.17	3277	0.11
BUSI	1970					167	0.11	1518	0.21
NURS	1970					689	0.33	3163	0.67
VETE	1970					420	0.23	1641	0.58

Tabla 2.1: **Número de erratas y porcentaje de ellas respecto al total del número de publicaciones.** Hay cuatro intervalos de tiempo para las 27 áreas. Las columnas están ordenadas cronológicamente iniciando con las áreas que tienen erratas desde el siglo XIX.

un factor de 3, yendo desde 0.24 % a 0.80 %. Tres áreas han permanecido casi constantes en estos dos períodos de tiempo: CENG, ECON, ENER. Del resto, 18 áreas aumentaron el porcentaje de erratas y sólo 6 lo redujeron.

Sería interesante entender detalladamente todas estas variaciones, pero dada la cantidad de datos y la variedad de campos involucrados, es una tarea enorme. Por tanto, el trabajo se concentró en el último período (2000-2014). Hay varias razones para esta selección. Una es que refleja las prácticas actuales. Otra es que la información es más accesible a través de búsquedas en la web.

La figura 2.1 muestra la evolución temporal de las 27 áreas de acuerdo a Scopus.

La cantidad de erratas y su tasa es diferente en cada área, pero la evolución en el tiempo parece ser constante en la mayoría de las áreas. Para cuantificar la posible existencia de una tendencia aplicamos un test de Cox-Stuart [53] con un nivel de significación del 5 % de los datos de cada área (ver apéndice). Se encontró que sólo siete de las áreas muestran una tendencia acuerdo con esta prueba. En seis de ellos la tasa de erratas crece (ARTS, BIOC, BUSI, CENG, ENER, MATE), en una disminuye (MULT). Aunque AGRI muestra un fuerte crecimiento en los últimos dos años no puede ser identificada con una tendencia utilizando esta prueba, pero es visualmente sugerente como se muestra en la figura 2.1.

A partir de los datos presentados en esta sección se tiene la impresión de que el comportamiento de erratas a través de los campos de conocimiento es lo suficientemente similar como para ser considerado un mismo fenómeno, pero al mismo tiempo muestra un nivel de detalle y variación a través de las áreas para ser una herramienta potencialmente valiosa y comprender algunas prácticas específicas al hacer investigación en diferentes disciplinas.

Para dar un primer paso hacia el uso de erratas como indicador cienciométrico, se separaron las áreas en cuartiles, de acuerdo a la frecuencia con la que aparecen las erratas en un área determinada. Esto se muestra en la figura 2.1. Algunos patrones son visibles. Por ejemplo, las áreas relacionadas con la medicina pueblan los cuartiles superiores (Q1 y Q2), es decir, tienen junto con el MULT el mayor porcentaje de erratas. Por el contrario, las áreas relacionadas con las ciencias sociales, como ART, BUSI, SOCI, tienden a aparecer en el cuartil más bajo (Q4); tienen el porcentaje más bajo de erratas. En el mismo cuartil también se encuentran COMP y ENGI. Una pregunta natural es si hay algo común en las prácticas de estas áreas que los colocan en Q4. Resulta que de acuerdo a [54] estas áreas tienen una gran contribución de artículos en memorias de congresos (proceedings papers) respecto al número total de publicaciones. Para probar si esto podría causar la ubicación de estas áreas en el cuartil más bajo, se seleccionaron las revistas que publican memorias de congresos y se buscó la cantidad de erratas en ellas. Resultó que para este tipo de artículos casi nunca se publican erratas. Esto es válido para todas las áreas. En quince de ellas practicamente no hay erratas en memorias de congresos. El área con el mayor porcentaje de erratas en proceedings es PHYS con 0.026 % en el período 2000-2014. Este es un factor de 20 menos que para el total de erratas en artículos. Desde este punto de vista se puede explicar, al menos parcialmente, que las áreas con un gran porcentaje de memorias de congresos en sus publicaciones, tendrán un porcentaje bajo de erratas.

Las observaciones anteriores muestran que la información sobre erratas se puede utilizar para poner de relieve un patrón diferente en la práctica de la ciencia a través

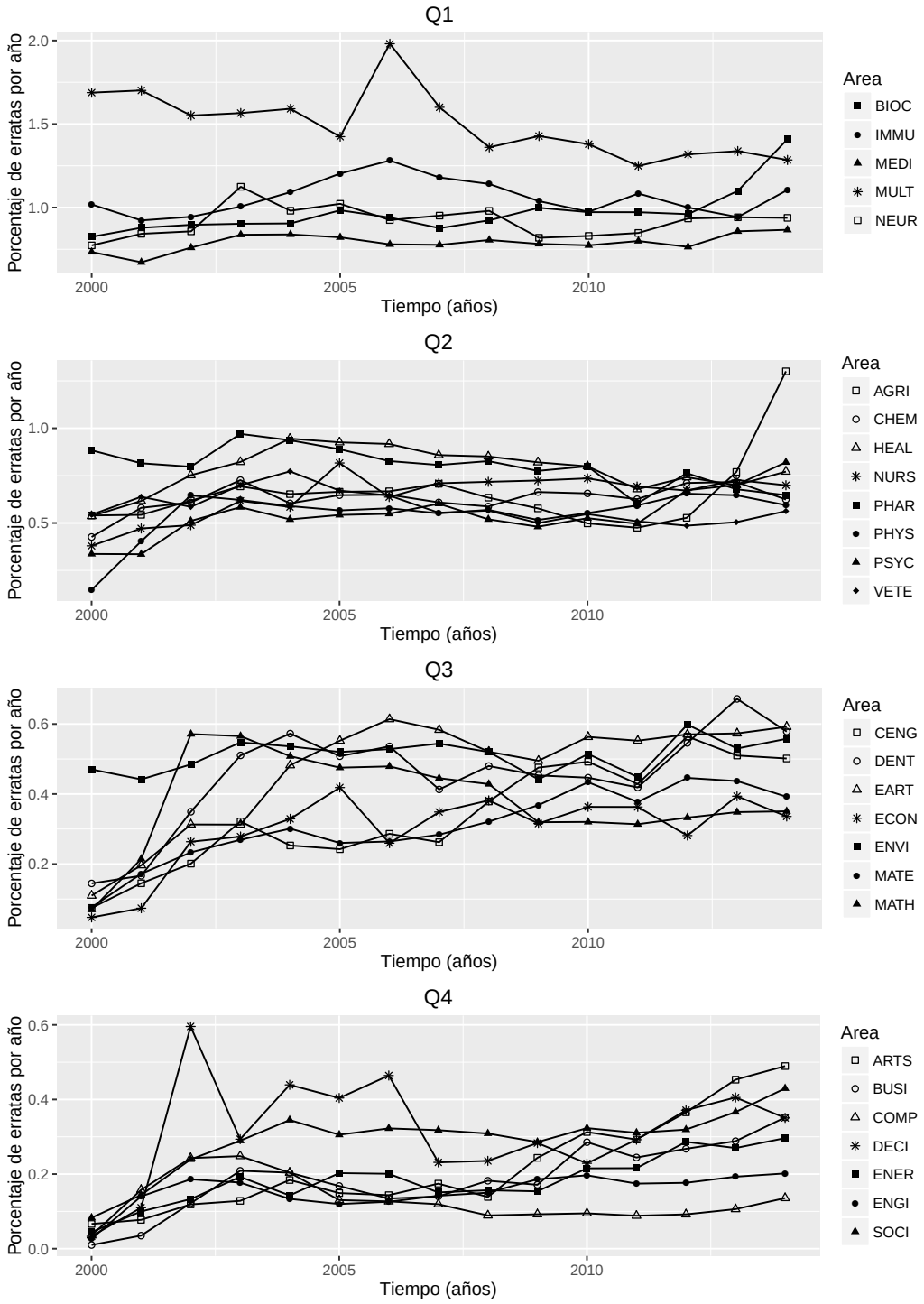


Figura 2.1: **Evolución temporal del porcentaje de erratas en cada área por año.** Corresponde al periodo 2000-2014 de acuerdo con los datos de Scopus. Las áreas se muestran en cuartiles del porcentaje de erratas en este periodo ordenadas del mayor (Q1) al menor (Q4) cuartil.

de las áreas. Antes de buscar nuevas formas de utilizar las erratas como indicador cienciométrico, se analizó si el comportamiento observado está relacionado con la

base de datos elegida. Con esta finalidad se compararon los resultados de Scopus con los de la Web of Science (WoS) para una área de cada cuartil. Se debe tener en cuenta que las categorías definidas en WoS no corresponden exactamente a las de Scopus, por lo que se esperan algunas diferencias en esta comparación. Teniendo esto en cuenta, las tendencias generales cualitativas son similares. Esto se muestra en la figura 2.2, donde hay una comparación entre las dos bases de datos en el tiempo en intervalos de un año para cuatro áreas. Es interesante notar que en MULT hay una marcada diferencia entre las dos bases de datos. Esta diferencia está asociada a que la revista PLOS ONE fue incluida en 2014 en la categoría de MULT en WoS, pero no está presente en la categoría correspondiente de Scopus. Una vez que se entiende el origen de la mayor discrepancia, esta comparación da cierta confianza en que los patrones observados en los datos son un producto de las diferentes prácticas en cada área y no un artefacto de la base de datos elegida.

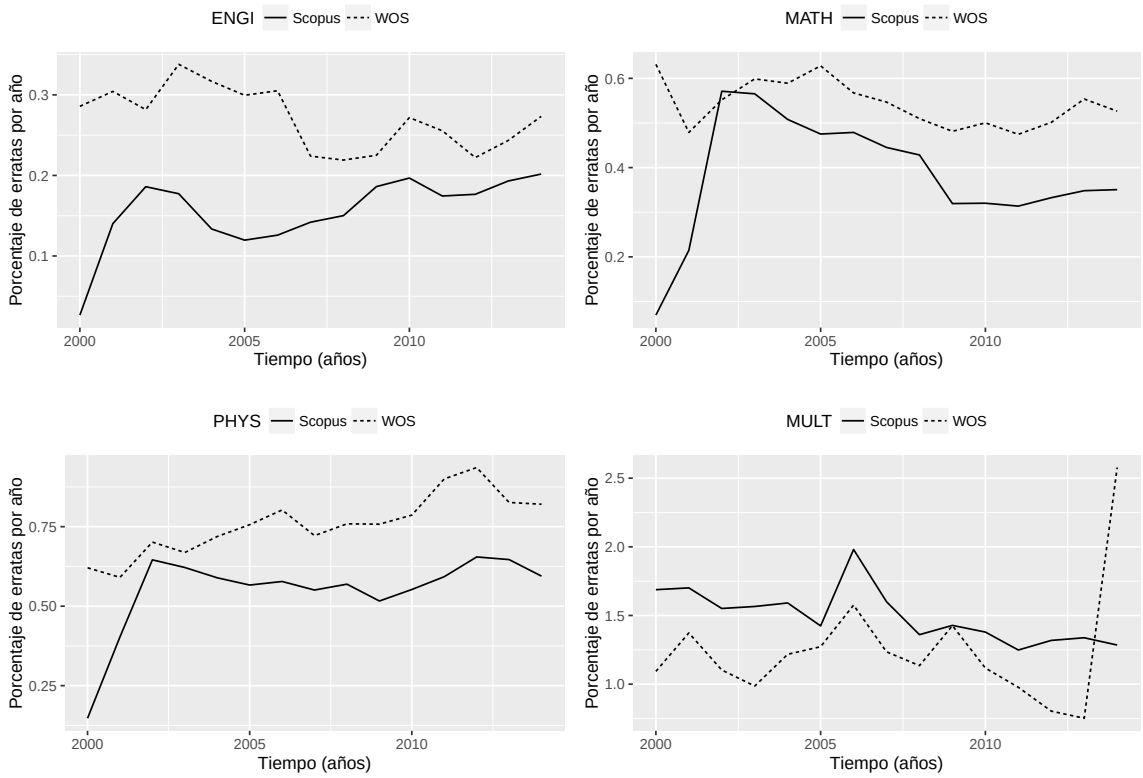


Figura 2.2: **Comparación del porcentaje de erratas en una área dada por año.** Para el periodo 2000-2014 de acuerdo a Scopus y WoS. Nótese que la definición de una área dada no es la misma en Scopus que en WoS.

2.2. Tres casos particulares

Mirando hacia atrás en la tabla 2.1 y figura 2.1, es claro que MULT ha tenido el mayor porcentaje de erratas en los últimos 15 años. Por otra parte, ya se mencionó que el aumento observado en la figura 2.2 se debe a una revista. Esto llevó a mirar más cerca el comportamiento de las revistas que se pueden considerar en cierto sentido multidisciplinarias. Se examinaron tres casos: Physical Review Letters, especializada en física, pero considerada multidisciplinaria dentro de este campo; Nature, una de las revistas científicas multidisciplinarias más famosas y PLOS ONE, la fuente de la tendencia abrupta en la figura 2.2.

2.2.1. Erratas en Physical Review Letters

La American Physical Society (APS) publica varias revistas científicas. Esto ofrece una situación ideal para este estudio, ya que se espera que las normas sean homogéneas en toda la familia de publicaciones de APS por lo que las variaciones entre las diferentes revistas pueden considerarse un reflejo del uso de las erratas en física.

APS tiene varias palabras clave para separar los artículos en general de artículos sobre errores. Estas incluyen *Comment, Reply, Publisher's note, Erratum* y *Retraction*. Las dos primeras no encajan en lo que se está discutiendo aquí como erratas, mientras que las tres últimas categorías sí. El caso de *Publisher's note* cubre errores tipográficos durante el proceso de publicación, mientras que *Errata* y *Retracción* son los errores cometidos por el (los) autor(es).

La tabla 2.2 muestra los datos de 2000 a 2014 para las diferentes revistas publicadas por APS. (Los detalles sobre la adquisición de los datos se dan en el apéndice B.) Physical Review Letters (PRL) es la revista insignia de APS publica artículos en todos los campos de la física. Reviews of Modern Physics (RMP) está especializada en artículos de revisión y Physical Review de A a E así como los aceleradores y haces (PRSTAB) cubren áreas específicas de la física.

El nivel de retracciones están muy por debajo de una por mil en todas las revistas y dado el tamaño de las muestras, no hay suficientes casos como para concluir algo sólido. Por otra parte, como se ha dicho antes, en este estudio interesa lo que se llamó errores honestos y por lo tanto se concentró en el tipo *Erratum*, que son concedidas a petición del autor (es). Hay una variación en el nivel de publisher's notes y erratas, se encuentran en rangos de 0.68-2.62 y 1.01-2.48, respectivamente. A excepción de RMP el número de erratas es mayor que el número de publisher's notes en cada revista. Ambos, el porcentaje de publisher's notes y erratas son superiores a la media del área (PHYS) para este periodo (0.56 %), como se muestra en

Revista	All	Publisher's note (%)	Erratum (%)	Retraction (%)
PRL	53692	528 (0.98)	900 (1.68)	11 (0.020)
RMP	611	16 (2.62)	11 (1.80)	-
PRA	34275	276 (0.80)	498 (1.45)	3 (0.009)
PRB	83050	595 (0.72)	1018 (1.22)	15 (0.018)
PRC	14717	149 (1.01)	286 (1.94)	1 (0.007)
PRD	39926	273 (0.68)	637 (1.59)	2 (0.005)
PRE	36668	303 (0.82)	372 (1.01)	5 (0.014)
PRSTAB	2139	30 (1.40)	53 (2.48)	-

Tabla 2.2: **Número de publicaciones y la fracción de ellas que son publisher's notes, errata and retractions.** La información corresponde a diferentes revistas publicadas por la APS. Los datos corresponden al periodo 2000-2014.

la tabla 2.1. Para todas las publicaciones de APS, el porcentaje de erratas es mayor que el promedio en PHYS. En el caso de PRL la diferencia es un factor de tres. Esta diferencia no puede explicarse por la presencia de memorias de congresos en PHYS descritas en la sección global. Y parece indicar que una de las revistas con los más altos estándares en cuanto a calidad en los artículos aceptados, tiene más erratas que el promedio del área.

La figura 2.3 muestra el tiempo transcurrido entre la publicación de un artículo y la errata asociada. En el período de 15 años que se examina, el número de erratas al año, 60 en promedio, es bastante estable. Una gran parte de las erratas se publicó el mismo año de la publicación original (40 % durante el período mostrado en la figura) y sólo en unos pocos casos las erratas vienen después de varios años. Parece que en esta revista (PRL) el mecanismo de autocorrección de la Ciencia funciona en la mayoría de los casos relativamente rápido.

2.2.2. Erratas en Nature

Nature es una revista multidisciplinaria con una larga tradición en la publicación de artículos de alta calidad. Estos artículos son de muchos tipos; sólo alrededor de un tercio de ellos son, estrictamente hablando, artículos de investigación. Esta revista tiene cinco categorías para publicar errores en los artículos previamente publicados [55]: *corrections* son errores en artículos que no son revisados por pares, *erratum* es una notificación de un error importante cometido por la revista, *corrigendum* es una notificación de un importante error cometido por el autor (es), *retraction* es una notificación de resultados no válidos y *addendum* es una notificación de una adición debida a la revisión por pares. En la tabla 2.3 se reporta el número de cada una de estas categorías por año en el período 2000-2014. Esta tabla también contiene el

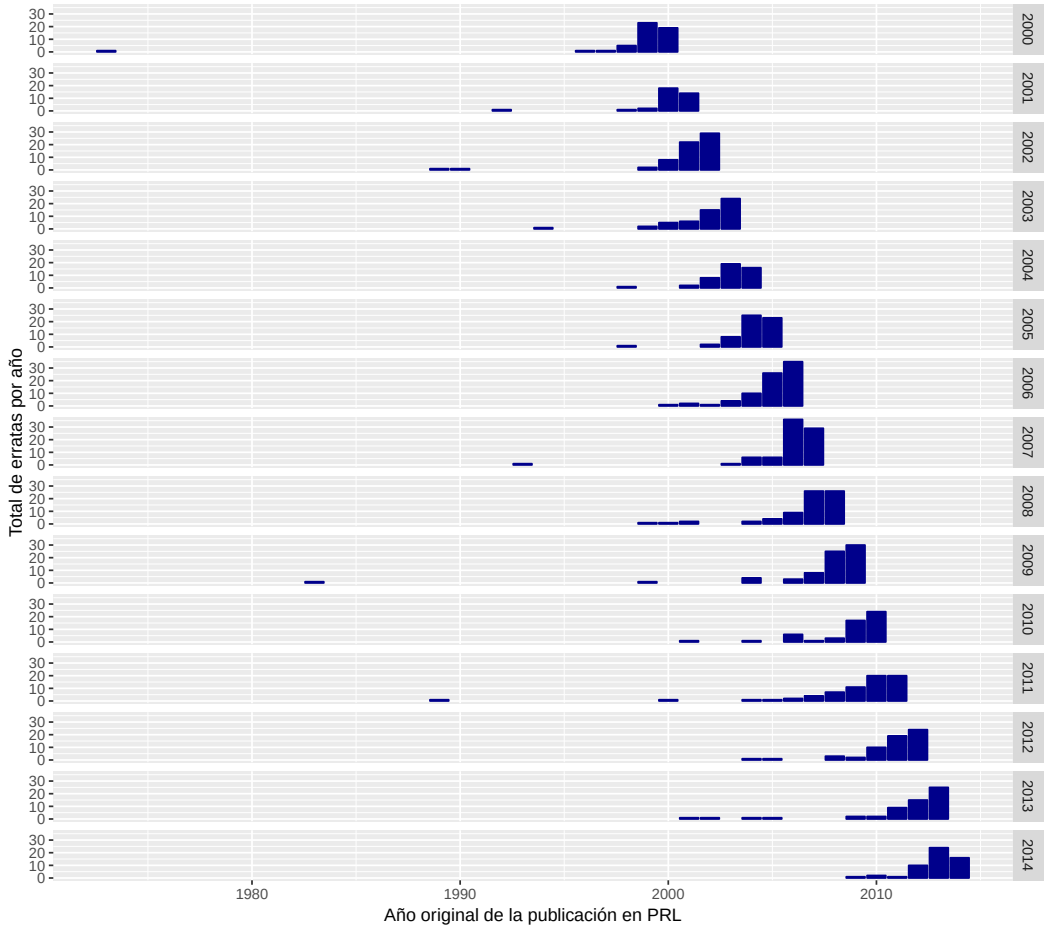


Figura 2.3: **Tiempo transcurrido entre la publicación de un artículo y la publicación de la errata correspondiente en PRL.** Periodo 2000-2014 considerando las erratas de cada año.

número total de artículos de investigación.

Hay algunas fluctuaciones en las cifras reportadas en la tabla 2.3. La más notable es el gran número de retracciones en el 2003, que se deben, directa e indirectamente, a la reiterada mala conducta científica de Jan Hendrik Schön [56]. También se reduce sustancialmente el número total de artículos de investigación a partir de 2006. La relación de corrección de errores con el número total de artículos de investigación 2006-2014 crece de 0.024 a 0.044. Estas cifras son significativamente más grandes que la media del área MULT como se mostró en la tabla 2.1 . La figura 2.4 muestra el tiempo transcurrido entre la publicación de un artículo y la publicación de la errata correspondiente en Nature durante el período 2000-2014. El comportamiento es similar al mostrado para PRL. La mayoría de los errores se encuentran bastante rápido, poco después de la publicación original, mientras que sólo en raras ocasiones un error queda sin ser publicado durante varios años.

Año	Erratum	Corrigendum	Retraction	Addendum	Correction	Research
2000	38	10	1	3	11	1282
2001	40	27	1	4	16	1158
2002	18	16	1	2	14	1054
2003	14	21	11	6	14	1069
2004	9	24	2	2	13	1031
2005	16	18	3	2	18	1217
2006	2	24	5	0	40	996
2007	5	18	2	3	31	760
2008	6	19	4	1	12	909
2009	13	24	1	3	13	749
2010	9	27	4	2	17	823
2011	9	34	1	4	16	797
2012	9	30	1	2	18	848
2013	10	38	6	0	14	846
2014	6	37	8	3	14	844

Tabla 2.3: **Errata en Nature.** Número de errores en cada categoría, y total de artículos de investigación publicados cada año por Nature en el periodo 2000-2014.

2.2.3. Errata en PLOS ONE

PLOS ONE es una revista de acceso abierto, que tiene un buen factor de impacto. La inclusión de sus artículos en las grandes bases de datos es bastante reciente además que la propia revista apenas cumplirá 10 años de existencia. En Scopus hay información desde 2011 y en WoS a partir de 2014. Por lo que se pudo evaluar, no hay manera fácil de distinguir en esta revista entre los errores en el proceso de redacción y las realizadas por el autor(es) utilizando las herramientas de búsqueda estándar disponibles en estas dos bases de datos. La tabla 2.4 muestra el número de publicaciones totales y erratas para esta revista. La figura 2.5 muestra el tiempo transcurrido entre la publicación de un artículo y la publicación de la errata correspondiente en PLOS ONE.

En este caso el número de retracciones varía significativamente en los últimos años, aunque por el mismo tiempo de vida de la revista los datos abarcan un periodo corto y es complicado extraer conclusiones firmes. El porcentaje de erratas reportado desde 2013 es significativamente mayor que los correspondientes en Nature o PRL. Por último, el tiempo transcurrido hasta que una errata aparece muestra una tendencia similar a la de las dos revistas analizadas anteriormente, sólo se muestran dos años debido a la dificultad para obtener los datos desde la página web de la revista y los que se muestran corresponden a los que se encontraron en WoS.

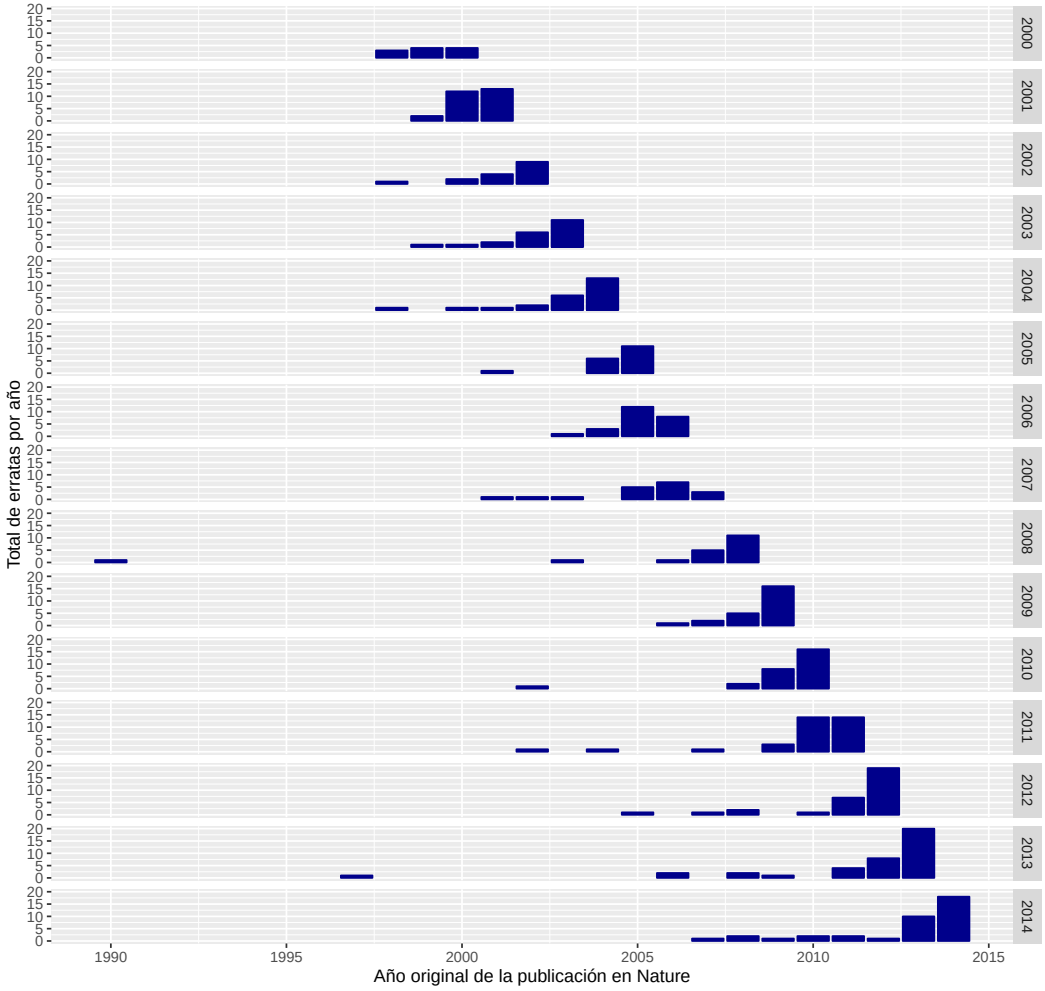


Figura 2.4: Tiempo transcurrido entre la publicación de un artículo y la publicación de la errata correspondiente en Nature para el periodo 2000-2014 por erratas de cada año.

Año	Total	Erratum	Erratum (%)	Retraction
2015	29815	1681	5.64 %	3
2014	31883	1805	5.66 %	12
2013	32987	1454	4.41 %	9
2012	24111	635	2.63 %	12
2011	14046	246	1.75 %	3
2010	6924	173	2.50 %	1
2009	4538	133	2.93 %	0
2008	2819	103	3.65 %	0
2007	1258	3	0.24 %	0
2006	140	0	0.00 %	0

Tabla 2.4: **Número de publicaciones, erratas y porcentaje de erratas en PLOS ONE.**

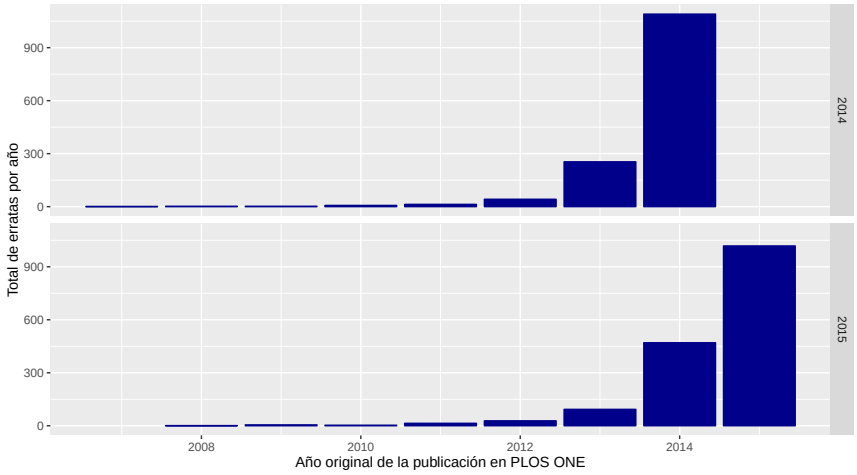


Figura 2.5: **Tiempo transcurrido entre la publicación de un artículo y la publicación de la errata correspondiente en PLOS ONE.**

2.3. Revistas y erratas

Los resultados de los dos apartados anteriores, a saber, que una revista puede cambiar el promedio de un área completa, y que algunas de las revistas más reconocidas tienen un porcentaje de erratas más grande que el promedio en un área sugieren una mirada más cercana desde la perspectiva de las revistas. Para esta parte del estudio se usaron los datos de WoS (ver apéndice B.1).

Se obtuvieron el número total de publicaciones y el número de erratas para cada revista en un área determinada para el período 2000-2014. Se extrajeron los datos para las siguientes áreas: ingeniería, matemáticas, física y multidisciplinario. Se debe tener en cuenta que una revista puede estar en más de un área al mismo tiempo. El número de revistas activas en el período mencionado en cada una de estas áreas fue 21114, 6627, 7228 y 764, respectivamente. Sorprendentemente para nosotros, el número de revistas con al menos una errata fueron 937, 555, 393 y 81 para un porcentaje de revistas sin una sola errata del 95.6 %, 91.6 %, 94.6 % y 89.4 %.

En el apéndice B.1 se muestran resultados globales de la búsqueda así como datos particulares de algunas revistas por área. A continuación se presentan los principales resultados por área.

2.3.1. Erratas en Ingeniería

En esta área un gran número de publicaciones se clasifican como *proceedings paper* o artículos en memorías de congresos (1,770,250 de 3,659,652, el 48.4 %). Como se mencionó antes prácticamente no hay erratas en relación con este tipo de publicaciones. Esto abre la pregunta, ¿por qué casi la mitad de los artículos en esta

área no tienen erratas? ¿Hay otro mecanismo que se encargue de corregir posibles errores introducidos en este gran conjunto de publicaciones? Aunque se trata de preguntas interesantes, van más allá del ámbito de acción de este estudio, por lo que se trabajó sólo en aquellas revistas que no tienen ningún artículo etiquetado como *proceeding paper* en WoS.

Sólo 6,293 de las 21,114 revistas publican artículos distintos a *proceeding paper*. De las 937 revistas con al menos una errata sólo cinco están en la categoría *Conference Title*, por lo que el 85 % de los 6293 no tienen publicaciones etiquetadas como erratum (correction) en WoS.

De estas 6293 revistas, se revisaron aquellas con al menos 100 artículos en el período 2000-2014. Hay 1283. De éstas, 369 no tienen erratas y 36 tienen menos de 0.1 % de los artículos clasificados como erratum.

Se observa que de las 10 revistas con mayor número de publicaciones y sin erratas y las 10 revistas con mayor número de publicaciones en el grupo con al menos una errata, pero menos de 0.1 % de erratas: 12 se encuentran en el cuarto cuartil de acuerdo con el factor de impacto (clasificación de 2014), 7 están en el tercer y segundo cuartil y sólo una está en el cuartil superior del FI.

2.3.2. Erratas en Matemáticas

Se siguieron los mismos pasos para analizar el área de Matemáticas. En este caso, sólo el 17 % de los artículos son *proceedings paper* y 4,172 (63 %) de las revistas no tienen artículos de este tipo, sólo una revista se encuentra es del tipo *Conference Title*. De las 4,172, sólo 718 revistas tienen más de 100 artículos en el período 2000-2014 y 175 de estas (24.4 %) no tienen erratas. Una vez más se seleccionaron las 10 revistas con la mayor cantidad de publicaciones y sin erratas; hubo 8 revistas con menos de un 0.1 % de erratas. De estas 18 revistas, 16 estaban en la mitad inferior de la clasificación del FI y sólo 2 están en el segundo cuartil.

2.3.3. Erratas en Física

Con el mismo análisis en el área de Física se obtuvieron los siguientes resultados: 28.3 % de los artículos son *proceedings paper*, 2,570 revistas no tienen ningún artículo de este tipo (sólo tres revistas están en *Conference Title*), y de éstas, 462 tienen más de 100 publicaciones en el período citado. En estas 462, 77 revistas no tienen erratas. Hay 8 revistas con más de 100 publicaciones, con al menos una errata y menos del 0.1 % de erratas; seis de ellas en la mitad inferior de la clasificación del FI y dos en el segundo cuartil. De las 10 revistas con más publicaciones y sin erratas, hay seis en

la mitad inferior de la clasificación del FI, dos en el segundo cuartil, una en el cuartil superior y otra sin factor de impacto.

2.3.4. Erratas en Multidisciplinario

En esta área, sólo el 5.2 % de las publicaciones son *proceedings paper* y 427 revistas no tienen ningún artículo de este tipo (sólo 17 revistas están en *Conference Title*). De ellas, sólo el 19 % tiene al menos una errata. De estas 427 revistas, sólo 34 tienen más de 100 publicaciones. De las 50 revistas con mas publicaciones, sólo tres tienen erratas con una frecuencia inferior a 0.1 %. Cada una tiene más de 1000 publicaciones, una está en el segundo cuartil, otra en el tercero y una en el cuarto. En las 50 revistas hay 11 revistas sin erratas. Ocho de ellas se encuentran en la mitad inferior del FI.

2.4. Comentarios generales sobre las erratas

Es preocupante la ausencia de erratas en los *proceedings paper*, dada la gran cantidad de artículos de este tipo. Hay un gran número de revistas con menos de 100 publicaciones en un periodo de 15 años, y de ellas, una buena fracción tampoco tiene erratas. De esas revistas con más de 100 publicaciones y sin erratas la mayoría están en los cuartiles bajos del factor de impacto, mostrando ausencia de estándares de calidad en su campo.

El análisis de PRL, Nature y PLOS ONE, indica que la mayoría de las erratas son publicadas rápidamente después de la publicación original, lo que sugiere que el proceso de autocorrección de la ciencia funciona, y en condiciones adecuadas el proceso es muy rápido. Esas revistas dejan ver que en un proceso editorial con altos estándares de revisión, poco más del 2 % de los artículos tendrán erratas.

A partir de este estudio se muestra la gran diferencia entre las bases de datos para delimitar las erratas, sobre todo entre las revistas y la WoS o Scopus. En ese sentido se aboga por la mejora en la identificación de erratas, separando las cometidas por los autores y las propias del proceso editorial.

Por supuesto hace falta investigación para entender porque en algunas revistas o campos las erratas son menos frecuentes o incluso inexistentes.

Capítulo 3

Sobre la influencia de un artículo en la red de citación

3.1. Introducción

Como se mencionó en el capítulo 1, una de las áreas de estudio con más interés en los últimos años es la relacionada con las redes (también llamadas gráficas), debido a la diversidad de contextos donde hay fenómenos que se pueden modelar mediante una gráfica o red [27, 28]: desde circuitos eléctricos en ingeniería, la interacción de genes o proteínas en biología, el estudio de las interacciones entre los bancos a nivel mundial en finanzas, en salud pública (el estudio de la propagación de enfermedades infecciosas en una población), en general en sistemas complejos vistos como redes complejas y por supuesto las redes sociales, entre otros campos del conocimiento [57].

Medir la importancia de un nodo en una red está condicionada a la definición de importancia que cada persona y/o en cada contexto se asigne. Intuitivamente se podría pensar que basta con contar las conexiones (aristas) de cada nodo con el resto de los nodos en la red y el que tenga más conexiones es el más importante, sin embargo hay muchas maneras de considerar tal importancia. En [58] se consideran medidas relacionadas con el grado (cantidad de conexiones) y con la distancia entre los nodos como la cercanía (closeness) y la intermediación (betweenness) vistas como medidas de centralidad en la gráfica asociada a la red en cuestión. Una de las más usadas en los últimos años es la asociada al valor propio máximo de la matriz de adyacencia de la gráfica, que en matemáticas cae dentro de la teoría espectral de gráficas [59] aunque en redes sociales ha sido explorada desde la perspectiva de [60] como un caso particular de su propuesta más general para medir la influencia o el impacto de un nodo en una red, también la investigación de [61] ha tenido una buena acogida en la comunidad de redes sociales y en general sobre redes.

Con el advenimiento de la Web se propusieron varias formas de clasificar el impacto de un artículo, como el Pange Rank [9, 10] que en cierto sentido es una variación de la centralidad por vector propio.

Una propuesta en la dirección de los planteamientos de [60] la desarrolla [62], planteando la centralidad por subgráfica, utilizando la idea de contar caminos que

inician y terminan en un vértice, para darle un valor al nodo en la gráfica. En [11] se plantea el concepto de comunicabilidad entre dos nodos a partir de la exponencial de la matriz de adyacencia de la gráfica. Luego en [12] se considera la comunicabilidad total como una medida de centralidad. Partiendo de estas investigaciones se propone medir la influencia de un artículo en una red de citación y se compara la propuesta con algunas de las medidas más comunes.

3.2. Conceptos básicos

A continuación se mencionan algunas definiciones importantes para el análisis.

Una gráfica $H = (V_H, E_H)$ es una subgráfica de G si $V_H \subseteq V$ y $E_H \subseteq E$. $F = (V_F, E_F)$ es una subgráfica inducida de G por V_F si $V_F \subseteq V$ y $E_F = \{(u, v) | u, v \in V_F, (u, v) \in E\}$.

Una gráfica dirigida (digráfica) $\vec{G} = (V, E)$ consiste de un conjunto no vacío y finito de vértices V y un conjunto E de pares ordenados de vértices, llamados aristas. A diferencia de G, en $\vec{G}$, al ser los pares ordenados implica que $(u, v) \neq (v, u)$.

Dos vértices $u, v \in V$ son adyacentes si están unidos por una arista de E, también se les llama vecinos. De manera similar dos aristas e_1, e_2 son adyacentes si tienen un vértice común. Un vértice $v \in V$ es incidente en una arista $e \in E$ si $v \in e$.

Un concepto importante en el análisis de redes es la de grado de un vértice. En G, k_i es el grado del vértice i definido como el número de aristas incidentes en i. En $\vec{G}$ hay dos tipos de grado para un vértice, grado de entrada (k^{in}) y grado de salida (k^{out}), el grado de entrada es igual al total de aristas que apuntan o que llegan a i y el grado salida es el total de aristas que salen de i.

Otro concepto muy importante en una red, es el de camino. Para definir camino primero se introduce lo que es una caminata y un recorrido. Una caminata es una secuencia alternada $\{v_0, e_1, v_1, e_2, ..., v_{l-1}, e_l, v_l\}$ donde $e_i = (v_{i-1}, v_i)$ y la longitud de la caminata es l. Un recorrido es una caminata sin repetir aristas y un camino es un recorrido sin repetir vértices. Un circuito es un recorrido en el que el vértice de inicio y final son el mismo, un camino con esta misma propiedad se llama ciclo. Para digráficas se extienden los conceptos de manera natural a caminata dirigida, recorrido dirigido, camino dirigido, etc.

Un vértice $v \in G$ se dice alcanzable desde otro vértice u si existe una caminata de u a v. La gráfica se dice conexa o conectada si cada vértice es alcanzable por cualquier otro. Una componente C de G, es una subgráfica conexa de G, si C es la subgráfica con la mayor cantidad de nodos y esta cantidad es una porción significativa del total de nodos, se suele llamar componente gigante.

Una digráfica $\vec{G}$ se dice débilmente conexa si la gráfica subyacente G (es decir

quitando la dirección de las aristas) es conexa. Es fuertemente conexa si cada vértice $v \in \vec{G}$ es alcanzable por cada $u \in \vec{G}$, y viceversa, mediante un camino dirigido.

3.2.1. Red de citación

Dado un conjunto de artículos S la red de citación $\vec{G}(S)$ o símplemente $\vec{G}$ es una digráfica con vértices en S y las aristas o flechas $a \to b$ implican que b cita a a o equivalente, si a está en la lista de referencias de b. Escribimos *n(S)* para el número de vértices y *m(S)* para el número de aristas en $\vec{G}(S)$.

Algunas propiedades importantes de las redes de citación son:

- Todas las aristas en las redes de citación apuntan hacia adelante en el tiempo, con la excepción de artículos que aparece simultáneamente en el tiempo y cuya lista de citación implica la presencia de ciclos. En ese sentido decimos que una red de citación es casi acíclica.
- Los vértices y las aristas agregados a las redes de citación son permanentes, no pueden ser removidos posteriormente.
- La parte ya existente de una red es casi estática, sólo los nodos frontales cambian.
- Un área de investigación es esencialmente cerrada respecto a las citas.

De lo anterior se tiene la siguiente proposición.

Proposición 1 *En una área dada, el promedio de citas de los artículos es igual al tamaño promedio de las listas de referencias.*

Prueba 1 *Sea k(S) el promedio de citas para artículos en S. Sea l(S) el tamaño promedio de la lista de referencias de artículos en S. Entonces k(S) es igual a*

$$\frac{1}{n(S)} \sum_a |a^-| = \frac{m(S)}{n(S)} = \frac{1}{n(S)} \sum_b |b^+|$$

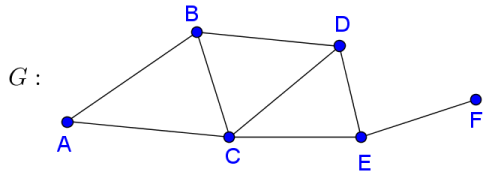
que es igual a l(S), donde a^- denota el conjunto de vértices iniciales de las aristas $a \to b$ y b^+ el conjunto de vértices finales. ■

La Proposición 1 es importante porque permite visualizar claramente que las variaciones en los promedios de citas de las distintas áreas de investigación se deben sólo a costumbres propias del área en cuanto al tamaño de las listas de referencia, es decir, a qué tantas referencias se consideran necesarias. Y como uno de los objetivos de la tesis es desarrollar una medida para identificar los artículos más influyentes en una área dada y dar sentido a esa medida, entonces se debe normalizar para que tengan sentido las comparaciones.

3.2.2. Matrices

A toda gráfica se le puede asociar una matriz de adyacencia y esto es lo que permite realizar la mayoría de los cálculos que se hacen sobre la gráfica. A continuación algunas definiciones.

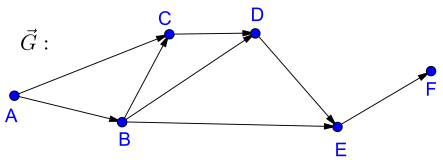
Dada una gráfica no dirigida G, su matriz de adyacencia se denota como $A(G) = (a_{ij})$ o simplemente A si es claro a qué gráfica se refiere. Si hay una arista entre i y j entonces $a_{ij} = 1$ sino $a_{ij} = 0$. En la tabla 3.1 se muestra un ejemplo.



$$A(G) = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Tabla 3.1: Ejemplo de gráfica y su matriz de adyacencia asociada.

Dada una gráfica dirigida $\vec{G}$, su matriz de adyacencia se denota como $\vec{A}(\vec{G}) = (a_{ij})$ o simplemente $\vec{A}$ si es claro a qué gráfica se refiere. Si hay una arista de i a j entonces $a_{ij} = 1$ sino $a_{ij} = 0$. En la tabla 3.2 se muestra un ejemplo de esto.



$$\vec{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Tabla 3.2: Ejemplo de gráfica dirigida y su matriz de adyacencia asociada.

Varias medidas de centralidad en gráficas están relacionadas con la distancia entre dos vértices, la cual está definida como la longitud l del camino más corto entre esos vértices, comúnmente se le llama distancia geodésica. La distancia de u a v se denota por $d(u, v)$. La excentricidad es la distancia desde un nodo al nodo más alejado de él en la red, y la excentricidad de mayor longitud es el diámetro de la gráfica.

El número de caminos de longitud k que van de i a j es igual al elemento en la posición (i, j) de la matriz A^k.

Las potencias de $\vec{A}$ son precisamente $\vec{A}^k = a_{ij}^k$. Siguiendo con los mismos ejemplos de las tablas 3.1 y 3.2 se tiene:

$$A^2 = \begin{pmatrix} 2 & 1 & 1 & 2 & 1 & 0 \\ 1 & 3 & 2 & 1 & 2 & 0 \\ 1 & 2 & 4 & 2 & 1 & 1 \\ 2 & 1 & 2 & 3 & 1 & 1 \\ 1 & 2 & 1 & 1 & 3 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}, \vec{A}^3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Una matriz que juega un papel muy importante en algunas medidas y en particular en la propuesta que se hace, es la exponencial de una matriz o la matriz exponencial e^A. Aquí A es la matriz de adyacencia de una gráfica dirigida o no dirigida de $n \times n$. e^A es una matriz de $n \times n$ definida por:

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \frac{A^4}{4!} + ... = \sum_{k=0}^{\infty} \frac{A^k}{k!} \tag{3.1}$$

Es claro que e^A cuenta todos los caminos de todas las longitudes, pero los caminos más largos son pesados o ponderados por el inverso del factorial, en ese sentido los caminos más largos tienen un peso mucho menor que los caminos más cortos.

Otros conceptos claves en el análisis de la centralidad son los valores y vectores propios de una matriz.

Dada una matriz $A = (a_{ij})$ de tamaño $n \times n$, un vector u de tamaño n es vector propio de A si existe un número real r tal que $Au = ru$, a r se le llama valor propio.

Los valores propios de A (es decir las raíces de $|\lambda I - A|$) se conocen como el espectro de A, denotado por $Sp(A) = \{\lambda_1, \ldots, \lambda_n\} = Sp(G)$, es decir, el espectro de la gráfica.

En caso de que la matriz sea no negativa (como las matrices de adyacencia de las gráficas o digraficas), esto es, $a_{ij} \geq 0$ para todas las entradas i, j, entonces hay valores propios $r \geq 0$. El máximo de los valores propios:

$$r(A) = max\{r : 0 \leq r \in Sp(A)\}$$

se llama radio espectral de A. El vector propio p correspondiente a $r(A)$, es el vector de Perron.

La existencia del radio espectral la asegura el Teorema de Perron.

Teorema 1 *Sea A una matriz no negativa y sea r su radio espectral, entonces:*

- *r es valor propio de A;*

- *existe un vector propio u de A tal que $Au = ru$ y todas las entradas de u son no negativas;*

- *si A es la matriz de adyacencia de una gráfica conexa (o digráfica fuertemente conexa), entonces valen:*

  - *r es raíz simple del polinomio característico de A y*
  - *el vector u tiene todas sus entradas > 0.*

Si $p = u$, la centralidad por vector propio se obtiene del vector de Perron, es decir la centralidad del nodo i es $p(i)$.

En la siguiente subsección se definen algunas medidas, de las más usuales, asociadas a la centralidad de un nodo.

3.2.3. Medidas más comunes asociadas a un nodo en una red

Se ha mencionado que hay muchas formas de medir la importancia de un nodo en una red dependiendo cómo se defina tal importancia y esta está asociada al tipo de información que se desea obtener. La más sencilla es la asociada con el grado, y simplemente cuenta el número de enlaces que tiene un nodo en la red, sin embargo esta información es limitada cuando se quiere ver qué tanto un nodo sirve como conector entre otros nodos o qué tan cerca está del resto de los nodos o cómo diferenciar un nodo de otro con el mismo número de enlaces, etc. Para ello se han desarrollado varias medidas, normalmente llamadas de centralidad, entre las más importantes están las que se describen a continuación.

- Cercanía o closeness [6] de un nodo v es el recíproco de la suma de las distancias desde el nodo v hasta los $n-1$ nodos restantes. Se normaliza por $n-1$, la suma de las mínimas distancias. La cercanía de v es:
  $C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$

- Intermediación o betwenness [6] de un nodo v es la suma de la fracción de todos los pares de geodésicas que pasan a través de v. Se denota por
  $B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$.
  Donde V es el conjunto de nodos, $\sigma(s,t)$ es el número de geodésicas de s a t y $\sigma(s,t|v)$ es el número de geodésicas de s a t que pasan por v.

- Espectral o de vector propio [61] calcula la centralidad de un nodo con base en el grado de sus vecinos. La centralidad espectral del nodo i es $p(i)$.

- Centralidad de Katz [6] es muy parecida a la espectral, ya que también se basa en el grado de los vecinos, sólo que más general, la centralidad del nodo i es:
  $x_i = \alpha \sum_j A_{ij} x_j + \beta$. Donde A es la matriz de adyacencia de la gráfica con valor propio λ. El parámetro β controla la centralidad inicial y $\alpha < \frac{1}{\lambda_{max}}$

- Page Rank [6]. Esta medida también se puede ver como un caso particular de la centralidad de Katz, la diferencia es que pesa todos los links o aristas por el grado de salida de cada nodo:
  $x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$
  donde k_j^{out} es el grado de salida del nodo j.

- Los hubs y las authorities sólo están definidas para redes dirigidas [6]. La centralidad de hub del vértice i es proporcional a las authorities, esto es:
  $x_i = \alpha \sum_j A_{ij} y_i$
  Y la authority del nodo i es proporcional a los hubs:
  $y_i = \alpha \sum_j A_{ij} x_i$

- Comunicabilidad [11]. En principio es una medida entre dos nodos, pero se ha extendido como una medida de centralidad para un nodo. Se basa en la suma de caminos cerrados de diferentes longitudes iniciando en un nodo u y terminando en el mismo nodo u. Se puede calcular usando la descomposición espectral de la matriz de adycacencia, así la comunicabilidad entre u y v es:
  $C(u,v) = \sum_{j=1}^{n} \phi_j(u)\phi_j(v)e^{\lambda_j}$. Donde $\phi_j(u)$ es el $u - esimo$ elemento del $j - esimo$ vector orthonormal de la matriz asociada con el valor propio λ_j. Esto es equivalente a
  $C(u,v) = (e^A)_{uv}$
  Luego la centralidad por comunicabilidad del nodo u se calcula como:
  $SC(u) = (e^A)_{uu}$.
  Esta también se conoce como centralidad de subgráfica [62].

- Comunicabilidad total [12]. Es muy parecida a la de comunicabilidad, sólo que en este caso se considera la suma sobre todos los caminos de un nodo hacia el resto de la red, la centralidad del nodo i es:
  $C(i) = \sum_j^n (e^A)_{ij}$

En el caso de la de comunicabilidad normalmente los algoritmos [63] están diseñados para trabajar con gráficas no dirigidas aunque no es difícil implementarlos para gráficas dirigidas como lo hacemos nosotros.

3.3. Propuesta

La propuesta está basada en la idea de medir la contribución de citas añejas, de segunda, tercera, cuarta generación que tiene un nodo en la red de citación, para ello se usa la exponencial de la matriz de adyacencia en un modelo más general que el de las redes de citación.

Sea la entrada $e^{\vec{A}}_{i,j}$ el tiempo de recorrido de un camino aleatorio de i a j. Entonces un camino invierte, en promedio:

$$\frac{(e^{\vec{A}})_{ij}}{w(\vec{G})} \tag{3.2}$$

unidades de tiempo para ir de i a j, donde $w(\vec{G}) = \sum_{i,j}^{n} (e^{\vec{A}})_{i,j}$ es el peso total de los caminos en $\vec{G}$.

Para $\beta > 0$, se considera que un camino asociado a la matriz $\beta\vec{A}$ invierte β unidades de tiempo en cada arista de la gráfica $\vec{G}$.

Fijando $s \in \vec{G}$. Sea $\vec{G}_s$ la gráfica formada por los n_s sucesores de $s \in \vec{G}$ (incluyendo el vértice s). La gráfica $\vec{G}_s$ tiene a s como el único origen.

Siguiendo a [11], se considera la comunicabilidad de i a j para cualesquiera par de nodos $i, j \geq s$ en $\vec{G}_s$ donde el tiempo de espera en cada arista es $\beta > 0$, como el número real:

$$\vec{c}_{ij}(\beta) = (e^{\beta\vec{A}_s})_{ij} \tag{3.3}$$

Y se plantea la influencia de s en $\vec{G}$ como:

$$I_s(\beta) = \frac{1}{w(\vec{G})} \sum_{k \geq s} \vec{c}_{sk}(\beta) \tag{3.4}$$

Por definición $0 \leq I_i(\beta) \leq 1$ y $\sum_{i=1}^{n} I_i(\beta) = 1, \forall i \in \vec{G}$. Es importante notar que una $I_i(\beta)$ cercana a cero indica una influencia baja y una $I_i(\beta)$ cercana a uno indica una influencia alta de i.

Por supuesto, para el análisis de redes de citación o de casi cualquier red real, la medida está bien definida porque $w(\vec{G}) \neq 0$, a menos que la red estuviera completamente desconectada.

La propuesta es muy similar al planteamiento en [12] como comunicabilidad total, salvo la definición de los tiempos de espera (aunque en las redes de citación serán siempre iguales a uno, $\beta = 1$) y que en [12] no consideran gráficas débilmente conexas, como es el caso de las redes de citación y es uno de los principales objetos de estudio

de esta tesis.

En la siguiente sección se presentan los métodos generales y los datos de la red donde se midió la influencia.

3.4. Datos y Métodos

3.4.1. Datos

Se trabajó con la red de citación de la American Physical Society, la cual corresponde al periodo 1893-2013.

La red tiene 543,744 nodos y 6,040,030 aristas. Es una red dirigida (digráfica) casi acíclica, aunque tiene una buena cantidad de ciclos (4973) no son significativos ya que la mayoría son ciclos de longitud dos o tres y muchos corresponden a subciclos de algunos más grande. Un ejemplo de un ciclo es el caso de "Theory of fermion exchange in massive quantum electrodynamics at high energy" (PhysRevD.13.369) publicado en seis partes, lo que origina ciclos de longitud seis ya que hay citas entre cada parte y subciclos de esos más grandes como se muestra en la figura 3.1, en la imagen no se aprecia con tanta claridad pero hay un ciclo que inicia en PhysRevD.13.484 y termina en PhysRevD.13.484: PhysRevD.13.484→ PhysRevD.13.424 → PhysRevD.13.508 → PhysRevD.13.369 → PhysRevD.13.379 → PhysRevD.13.395 → PhysRevD.13.484. Pero entre esos seis nodos hay 35 ciclos.

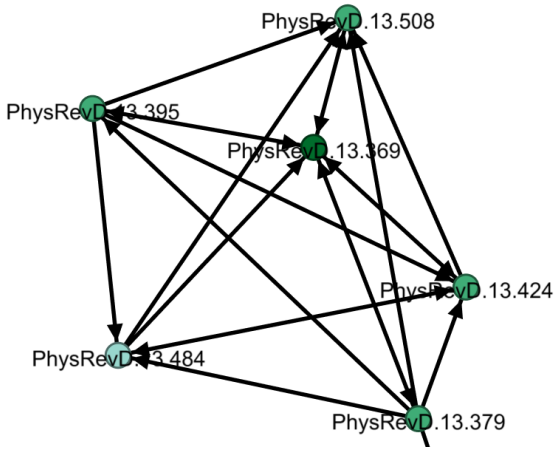


Figura 3.1: Ejemplo de ciclos en la red de APS entre artículos de PRD.

3.4.2. Métodos

Se obtuvieron subredes correspondientes a algunas revistas que componen el conjunto de publicaciones de APS, cabe recordar que la primera publicación de la APS

fue *Physical Review Series I (PRSI)* , de 1893 a 1912. En 1913 cambia su nombre a *Physical Review (PR)* y lo mantiene hasta 1969, en 1970 se divide en *Physical Review A (PRA)*, *Physical Review B (PRB)*, *Physical Review C (PRC)*, *Physical Review D (PRD)* y *Physical Review E (PRE)*. Antes de la división, en 1929 se crea *Review of Modern Physics (RMP)* y en 1958 *Physical Review Letters (PRL)*. En 1998 se crea *Physical Review Special Topics – Accelerators and Beams (PRSTAB)* y en enero de 2016 fue cambiado el nombre a Physical Review Accelerators and Beams (PRAB), en los datos con que se trabajó tiene el nombre corto PRSTAB. En 2005 surge *Physical Review Special Topics—Physics Education Research (PRSTPER)* y en enero de 2016 cambia de nombre a *Physical Review Physics Education Research (PRPER)*, en los datos con que se trabajó aparece como PRSTPER. En 2011 nace *Physical Review X (PRX)*. También hay datos de *Physical Review Focus* que nació en 1998 pero en 2011 cambió a *Physics*.

En la tabla 3.3 se muestra la distribución de las citas entre las diferentes revistas de la APS

Revista	Citas	Porcentaje
PR	635560	10.522
RMP	158621	2.626
PRL	1965514	32.541
PRA	548225	9.075
PRB	1510387	25.006
PRC	268675	4.448
PRD	724383	11.993
PRE	220851	3.656
PRSTAB	4948	0.081
PRSTPER	605	0.010
Physics	1375	0.022
PRX	886	0.014

Tabla 3.3: Distribución de citas por revista en APS en orden cronológico.

En las subredes correspondientes sólo se consideran las citas entre artículos de la revista, para tener consistencia y evitar muchos artículos sin citas, ya que en algunas revistas la cantidad de citas que reciben de otras es muy grande, como es el caso de RMP en la que el 96.96 % de las citas son externas (de otras revistas de la familia APS) así como PRL con 77.33 % de citas externas, lo cual es esperado dadas las características de las revistas, pero por esta razón estas no se consideran por separado, tampoco se consideraron Physics (97 % externas) ni PRX (95 % externas). Aunque en PR el 61.82 % de las citas son externas, si la consideramos como subred porque nos permite tener un corte a 1969.

Para calcular la exponencial de la matriz (e^A) se parte de la acción de una matriz sobre un vector [64], ya que realmente no interesan todas las entradas de la matriz sino la suma de cada renglón, se usaron los algoritmos implementados en las librerias de scipy sparse linealg de Python.

Para la red completa se comparó la I con las medidas más usuales, para ello se calculó el coeficiente de correlación de Pearson (ρ) entre ellas. En Python se obtuvo el grado de salida de cada nodo, que no es otra cosa más que las citas que tiene cada nodo y los Hubs. La centralidad por vector propio (Vp) y el page rank (Pr) se calcularon en Gephi [65]. Cabe señalar que para realizar los cálculos en Gephi se debe voltear la dirección de las aristas de lo contrario no tiene sentido la comparación con la I.

Para las subredes se también se calcularon la cercanía (Cl) y la intermediación (Be) también en Gephi. En todos los casos, los cálculos fueron sobre la componente gigante (Gc) de cada red.

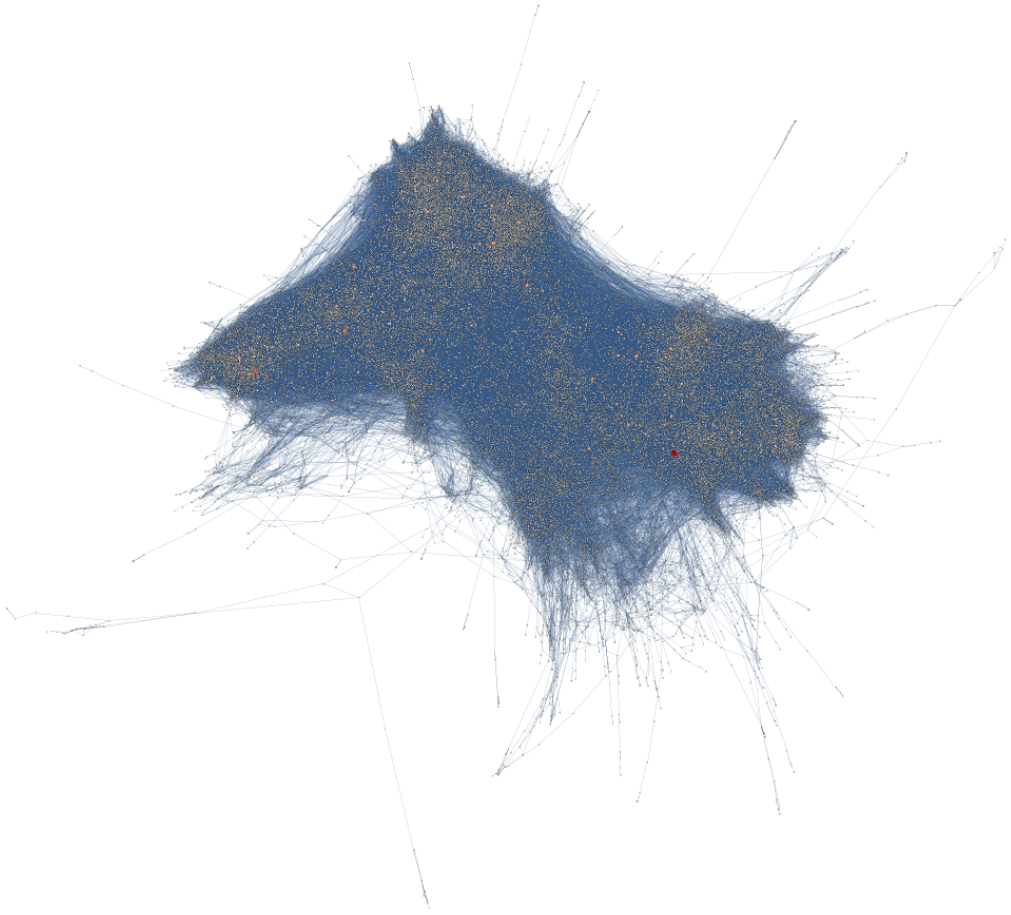


Figura 3.2: Componente gigante de PR. Con 43,896 nodos y 242,276 aristas.

En la figura 3.2 se muestra una imagen de la componente gigante de PR, si se compara con la figura 3.3 que corresponde a la componente gigante de PRB, se

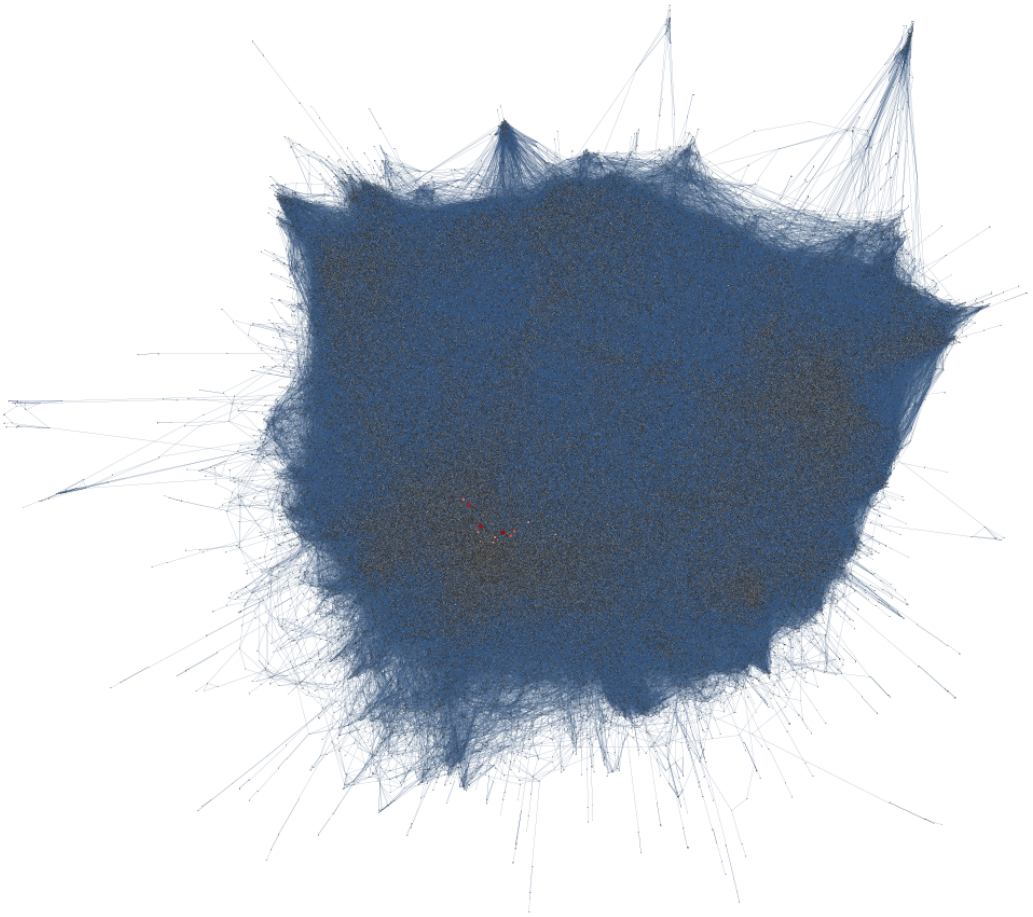


Figura 3.3: Componente gigante de PRB. Con 157,738 nodos y 1,214,262 aristas.

aprecia el impacto de la Proposición 1, ya que es claro como se modifica el promedio de citas, mientras que en PR es de 5.5, en PRB es de 7.7, visualmente es la diferencia en la densidad de cada red. Sólo se menciona como una muestra, ya que realmente el análisis visual de redes tan grandes no arroja mucha información, pero se alcanzan a apreciar estas características.

Sobre los cálculos en el apéndice C.1 se muestra un ejemplo del código en Python usado para realizar algunos cálculos.

3.5. Influencia en la red de APS

La Gc de la APS, en adelante GcAPS, tiene 530,681 nodos y 6,039,451 aristas, con un promedio de 11.38 citas por artículo. Sobre GcAPS se calculó la $I_i(\beta)$ así como el resto de las medidas mencionadas en los métodos. En GcAPS hay 459,360 nodos con al menos una cita, el 13.43 % no tienen citas. En la tabla 3.4 se muestra la matriz de correlaciones (ρ), para el conjunto de nodos con al menos una cita.

De acuerdo con la tabla 3.4 la I tiene una correlación medianamente alta con Pr

Medidas	Cn	I	H	Vp	Pr
Cn	1.00	0.39	0.60	0.90	0.47
I	0.39	1.00	0.17	0.60	0.66
H	0.60	0.17	1.00	0.50	0.17
Vp	0.90	0.60	0.50	1.00	0.55
Pr	0.47	0.66	0.17	0.55	1.00

Tabla 3.4: Matriz de correlaciones entre las medidas en GcAPS. Cn son las citas normalizadas ($k^{out}/(n-1)$) por el total de citas posibles en la red (n-1). I es la influencia. H son los hubs. Vp es la centralidad por vector propio. Pr es el page rank.

(66 %) y con Vp (60 %). Es interesante notar que Vp tiene la correlación más alta con Cn (90 %) y con el resto de las medidas. La I tiene la correlación más baja con Cn (39 %). En la figura 3.4 se muestran los diagramas de dispersión entre las medidas mencionadas.

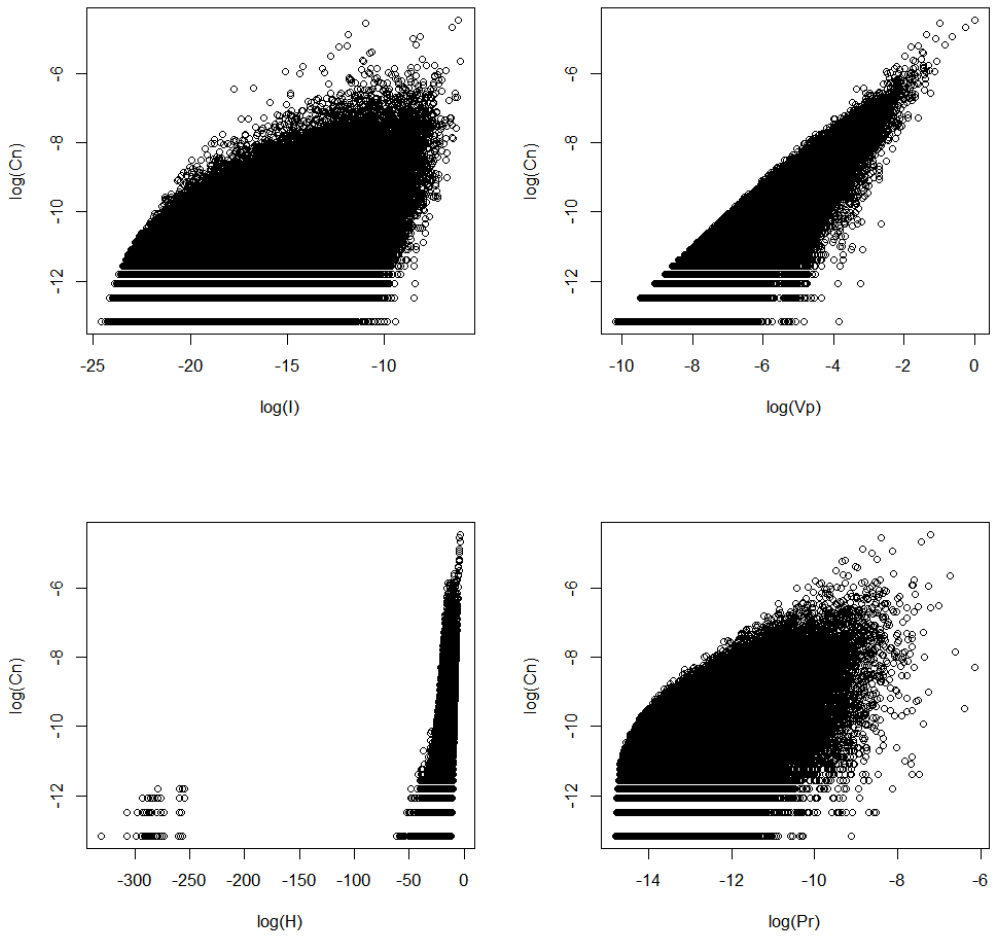


Figura 3.4: Diagramas de dispersión entre las medidas en GcAPS en escala log-log.

En la figura 3.4 se aprecia que la I posiciona varios nodos de manera distinta que el conteo de citas o que el vector propio, hay artículos con muchas citas a los que no les corresponden valores altos de influencia, esto indica que la I diferencia nodos (artículos) que el Vp o el Pr no.

3.5.1. Physical Review

Para PR, la gráfica se designa por GPR y la componente gigante por $GcPR$. GPR tiene 44,444 nodos (artículos) y 242,647 aristas (citas). GcPR tiene 43,896 nodos y 242,276 aristas, con un promedio de citas de 5.51 por artículo, hay 35 011 artículos con al menos una cita, lo que implica que el 20.24 % no tiene citas, aunque hay que recordar que sólo se cuentan las citas hasta 1969 y el 60 % de las citas de PR vienen de otras revistas. En la tabla 3.5 se muestra la matriz de correlaciones para el conjunto con al menos una cita.

Medidas	Cn	I	H	Cl	Be	Vp	Pr
Cn	1.00	0.44	0.33	0.13	0.58	0.68	0.60
I	0.44	1.00	0.038	0.09	0.32	0.90	0.64
H	0.33	0.038	1.00	−0.003	0.19	0.12	0.15
Cl	0.13	0.09	−0.003	1.00	0.11	0.14	0.15
Be	0.58	0.32	0.19	0.11	1.00	0.46	0.36
Vp	0.68	0.90	0.12	0.12	0.46	1.00	0.70
Pr	0.60	0.64	0.15	0.15	0.36	0.70	1.00

Tabla 3.5: Matriz de correlaciones entre las medidas en GcPR. Cl es la cercanía, Be es la intermediación.

En la tabla 3.5 sobresale la correlación entre la I y Vp (89 %), Vp es nuevamente la medida más correlacionada con el resto, aunque en este caso la correlación con Cn es más baja que en todo APS. Otro aspecto a destacar es que H es la segunda medida menos correlacionada con Cn (33 %) mientras que en todo APS es la segunda mejor correlacionada con Cn.

En la figura 3.5 se muestran los diagramas de dispersión entre las medidas más correlacionadas con las citas. De manera similar a GcAPS en GcPR también hay artículos con muchas citas y una influencia baja.

3.5.2. PRA

Para PRA la gráfica se denota por $GPRA$ y la componente gigante por $GcPRA$. En GPRA hay 62,423 nodos (artículos) y 373,116 aristas (citas), 5.97 citas en promedio. GcPRB tiene 61,363 nodos y 372,314 aristas, hay 49,766 con al menos una

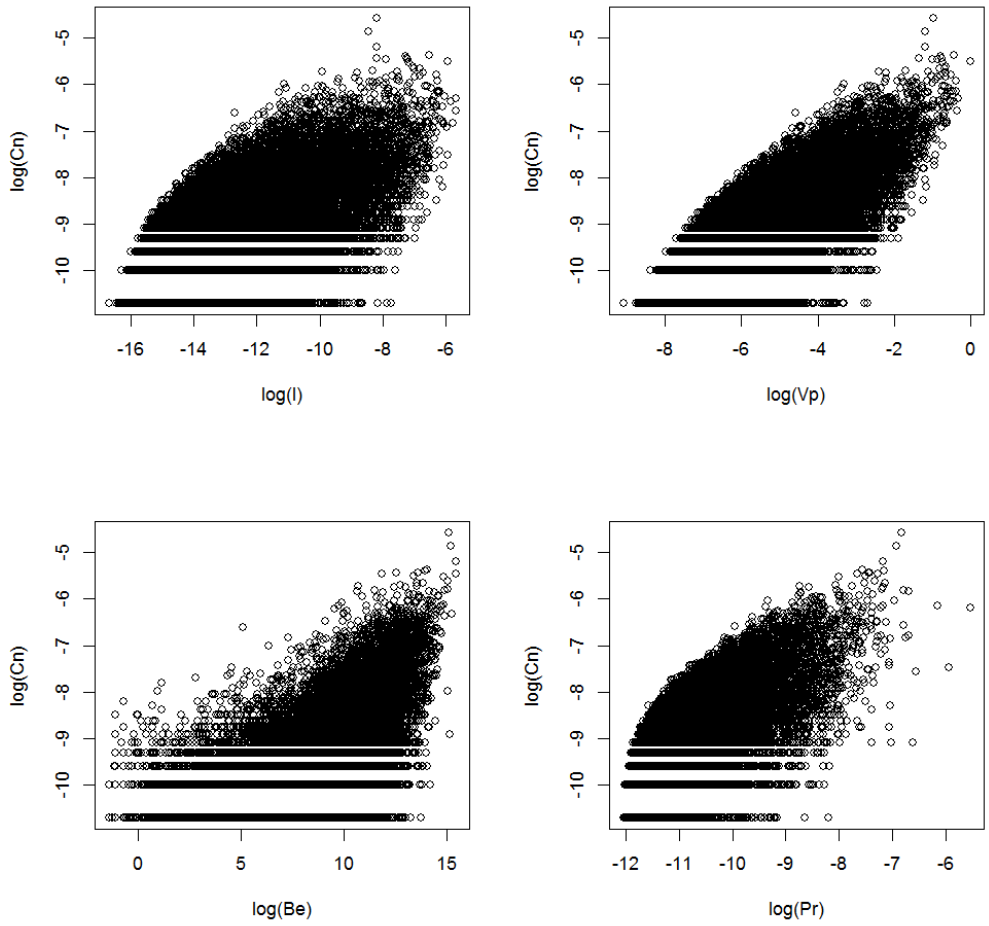


Figura 3.5: Diagrama de dispersión entre las medidas más correlacionadas con las citas en GcPR en escala log-log.

cita lo que implica que el 18.89 % no tiene citas. En la tabla 3.6 se muestra la matriz de correlaciones entre el conjunto con al menos una cita.

Medidas	Cn	I	H	Vp	Pr	Cl	Be
Cn	1.00	0.65	0.55	0.86	0.59	0.18	0.32
I	0.65	1.00	0.37	0.91	0.78	0.11	0.20
H	0.55	0.37	1.00	0.55	0.30	-0.003	0.083
Vp	0.86	0.91	0.55	1.00	0.74	0.15	0.27
Pr	0.59	0.78	0.30	0.74	1.00	0.13	0.14
Cl	0.18	0.11	-0.003	0.15	0.13	1.00	0.119
Be	0.32	0.20	0.083	0.27	0.14	0.119	1.00

Tabla 3.6: Matriz de correlaciones entre las medidas en GcPRA.

De acuerdo con la tabla 3.6 las citas están altamente correlacionadas con Vp

(86 %) seguidas de la influencia (65 %), en esta red (GcPRA) H es la tercera medida más correlacionada con Cn. Nuevamente se aprecia la influencia es útil para diferenciar donde el vector propio no lo hace como se muestra en la gráfica 3.6

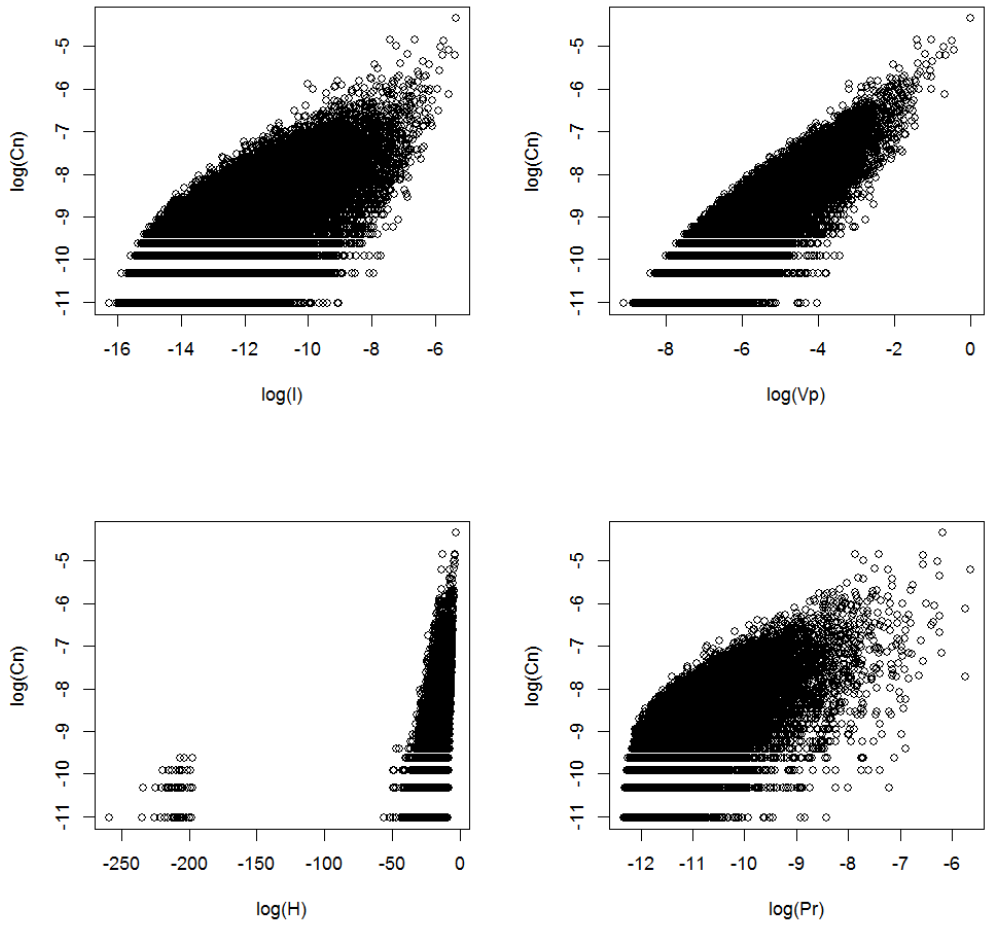


Figura 3.6: Diagrama de dispersión entre las medidas más correlacionadas con las citas en GcPRA en escala log-log.

3.5.3. PRB

Para PRB la gráfica se denota por $GPRB$ y la componente gigante por $GcPRB$. GPRB tiene 158,156 nodos (artículos) y 1,214,534 aristas (citas), dando 7.67 citas en promedio. GcPRB tiene 157,738 nodos y 1,214,262 aristas, hay 132,118 con al menos una cita lo que implica que el 16.24 % no tiene citas. En la tabla 3.7 se muestra la matriz de correlaciones entre el conjunto con al menos una cita.

En la tabla 3.7 sobresale que la correlación entre H y Cn es la segunda más alta, de manera similar a GcAPS. En este caso la I es la segunda menos correlacionada

Medidas	Cn	I	H	Vp	Pr	Cl	Be
Cn	1.00	0.330	0.775	0.837	0.414	0.078	0.45
I	0.330	1.00	0.119	0.664	0.519	0.009	0.13
H	0.775	0.119	1.00	0.513	0.175	-0.004	0.30
Vp	0.837	0.664	0.513	1.000	0.525	0.057	0.39
Pr	0.414	0.519	0.175	0.525	1.00	0.055	0.15
Cl	0.078	0.009	-0.004	0.057	0.055	1.00	0.08
Be	0.45	0.13	0.30	0.39	0.15	0.08	1.00

Tabla 3.7: Matriz de correlaciones entre las medidas en GcPRB.

con Cn y Vp vuelve a ser la medida más correlacionada con el resto. Además la I no tiene una correlación tan alta con Pr y Vp (51 % y 66 %). En la figura 3.7 se muestran los diagramas de dispersión entre algunas medidas para el conjunto de artículos con al menos una cita.

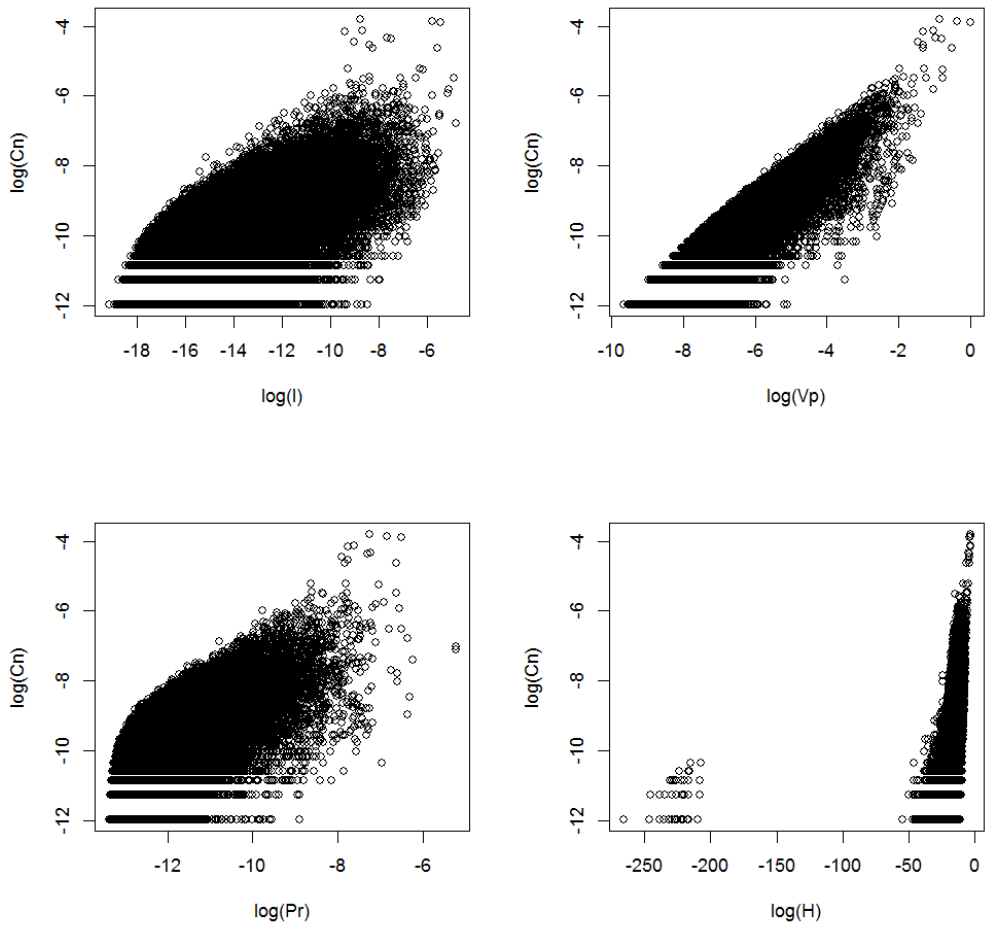


Figura 3.7: Diagrama de dispersión entre las medidas más correlacionadas con las citas en GcPRB en escala log-log.

En la figura 3.7 se aprecia que aunque la influencia tiene una correlación alta con *Vp* y *Pr* la manera de ordenar los nodos es distinta en tanto que se pueden diferenciar nodos con pocas citas, además a diferencia de *Pr* para valores bajos de *I* corresponden valores bajos de *Cn*.

3.5.4. PRC

Para PRC la gráfica se denota por *GPRC* y la componente gigante por *GcPRC*. GPRC tiene 33,457 nodos (artículos) y 225,796 aristas (citas), dando 6.74 citas en promedio. GcPRC tiene 33,245 nodos y 225,645 aristas, hay 28,702 con al menos una cita lo que implica que el 13.66 % no tiene citas. En la tabla 3.8 se muestra la matriz de correlaciones entre el conjunto con al menos una cita.

Medidas	*Cn*	*I*	*Pr*	*Vp*	*Cl*	*Be*	*H*
Cn	1.000	0.602	0.482	0.830	0.053	0.466	0.570
I	0.602	1.000	0.645	0.899	0.033	0.219	0.335
Pr	0.482	0.645	1.000	0.586	0.096	0.141	0.172
Vp	0.830	0.899	0.586	1.000	0.041	0.375	0.531
Cl	0.053	0.033	0.096	0.041	1.000	0.032	−0.019
Be	0.466	0.219	0.141	0.375	0.032	1.000	0.314
H	0.570	0.335	0.172	0.531	−0.019	0.314	1.000

Tabla 3.8: Matriz de correlaciones entre las medidas en GcPRC.

En la tabla 3.8 sobresale que la correlación entre *I* y *Cn* es la segunda más alta y *Vp* vuelve a ser la medida más correlacionada con el resto, 83 % con *Cn* y 89 % con *I*. En la figura 3.8 se muestran los diagramas de dispersión de las medidas más correlacionadas con las citas.

3.5.5. PRD

Para PRD la gráfica se denota por *GPRD* y la componente gigante por *GcPRD*. GPRD tiene 68,188 nodos (artículos) y 618,087 aristas (citas). GcPRD tiene 67,758 nodos y 617,760 aristas, con 9.11 citas en promedio. Hay 56,938 artículos con al menos una cita, lo que implica que el 15.96 % de los nodos no tienen citas. En la tabla 3.9 se muestra la matriz de correlaciones para el conjunto con al menos una cita.

Nuevamente la *I* tiene una correlación alta con *Vp* (82 %) y ésta es la medida más correlacionada con el resto, además *H* es la segunda más correlacionada con las citas. En la fig 3.9 se muestra el diagrama de dispersión entre las medidas más correlacionadas con las citas en el conjunto con al menos una cita.

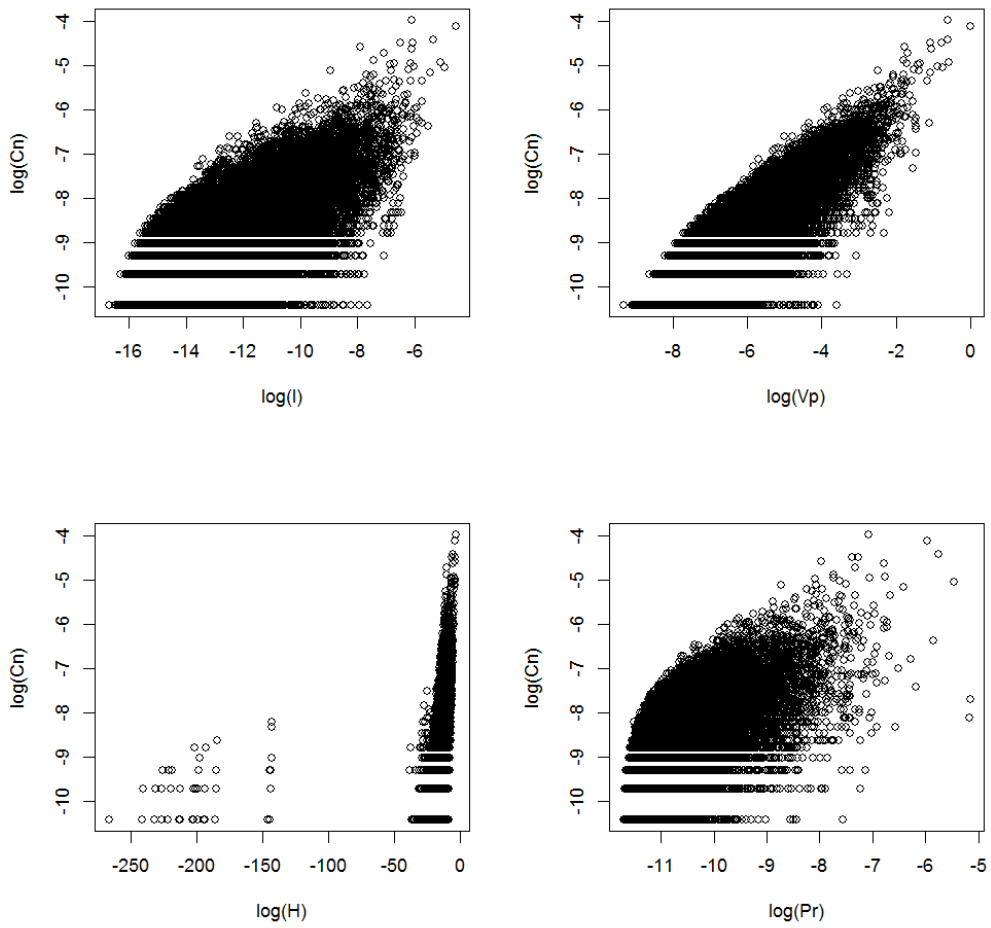


Figura 3.8: Diagrama de dispersión entre las medidas más correlacionadas con las citas en GcPRC en escala log-log.

Medidas	*Cl*	*Be*	*Cn*	*I*	*H*	*Vp*	*Pr*
Cl	1.00	0.069	0.069	0.020	0.003	0.049	0.017
Be	0.069	1.00	0.418	0.172	0.288	0.352	0.129
Cn	0.069	0.418	1.00	0.519	0.588	0.842	0.482
I	0.020	0.172	0.519	1.00	0.445	0.826	0.501
H	0.003	0.288	0.588	0.445	1.00	0.604	0.235
Vp	0.049	0.352	0.842	0.826	0.604	1.00	0.536
Pr	0.017	0.129	0.482	0.501	0.235	0.536	1.00

Tabla 3.9: Matriz de correlaciones entre las medidas en GcPRD.

En la figura 3.9 se aprecia que artículos con muchas citas pueden tener una influencia baja así como *Vp* baja. Además sobresalen los Hubs que en este caso se distribuyen diferente respecto a las citas.

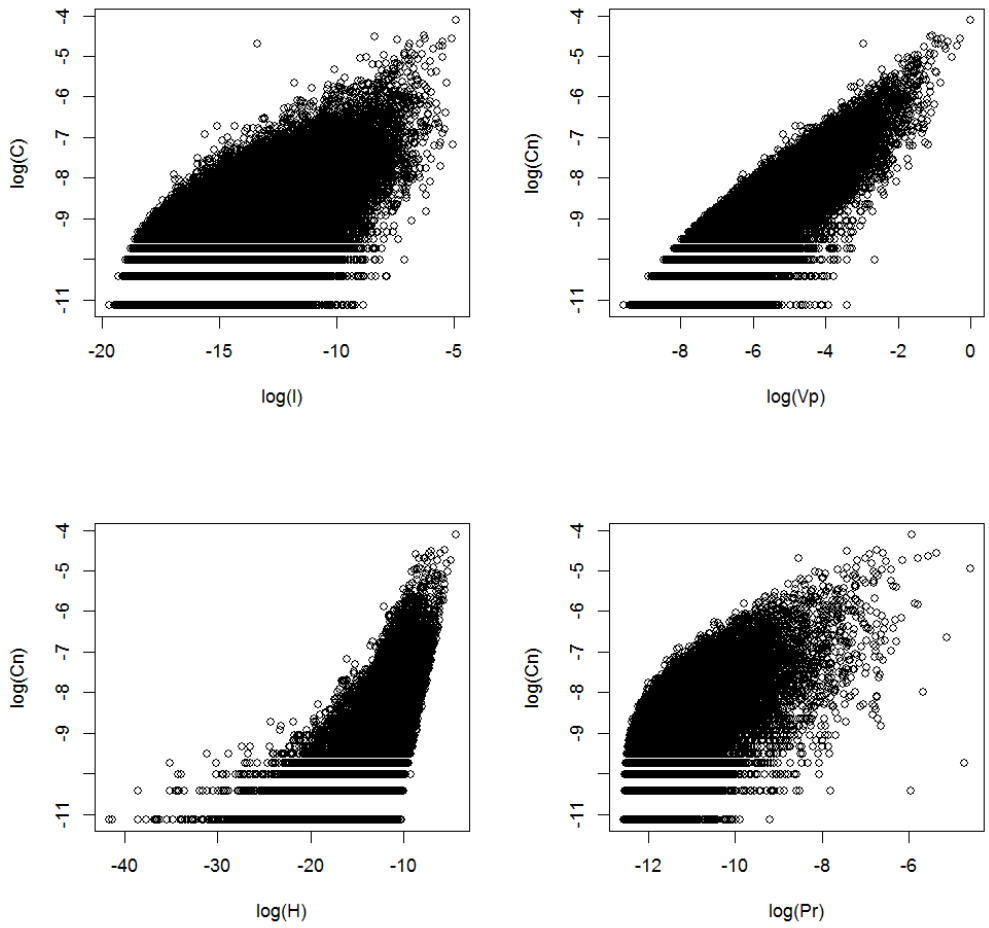


Figura 3.9: Diagrama de dispersión entre las medidas más correlacionadas con las citas en GcPRD en escala log-log.

3.5.6. PRE

Para PRE la gráfica se denota por *GPRE* y la componente gigante por *GcPRE*. GPRE tiene 42,513 nodos (artículos) y 160,068 aristas (citas). GcPRD tiene 41,135 nodos y 159,066 aristas, con 3.86 citas en promedio. Hay 31,147 artículos con al menos una cita, lo que implica que el 24.28 % de los nodos no tienen citas. En la tabla 3.8 se muestra la matriz de correlaciones para el conjunto con al menos una cita.

Nuevamente la I tiene una correlación alta con Vp (99 %) y ésta es la medida más correlacionada con el resto, además H es la cuarta más correlacionada con las citas en contraste con PRD donde es la segunda. En la figura 3.10 se muestra el diagrama de dispersión entre las medidas más correlacionadas con las citas en el conjunto con al menos una cita.

Medidas	*Cn*	*I*	*H*	*Vp*	*Pr*	*Cl*	*Be*
Cn	1.00	0.70	0.41	0.77	0.66	0.33	0.29
I	0.70	1.00	0.45	0.99	0.77	0.24	0.16
H	0.41	0.45	1.00	0.47	0.30	0.027	0.14
Vp	0.77	0.99	0.47	1.00	0.77	0.27	0.19
Pr	0.66	0.77	0.30	0.77	1.00	0.31	0.10
Cl	0.33	0.24	0.027	0.27	0.31	1.00	0.14
Be	0.29	0.16	0.14	0.19	0.10	0.14	1.00

Tabla 3.10: Matriz de correlaciones entre las medidas en GcPRE.

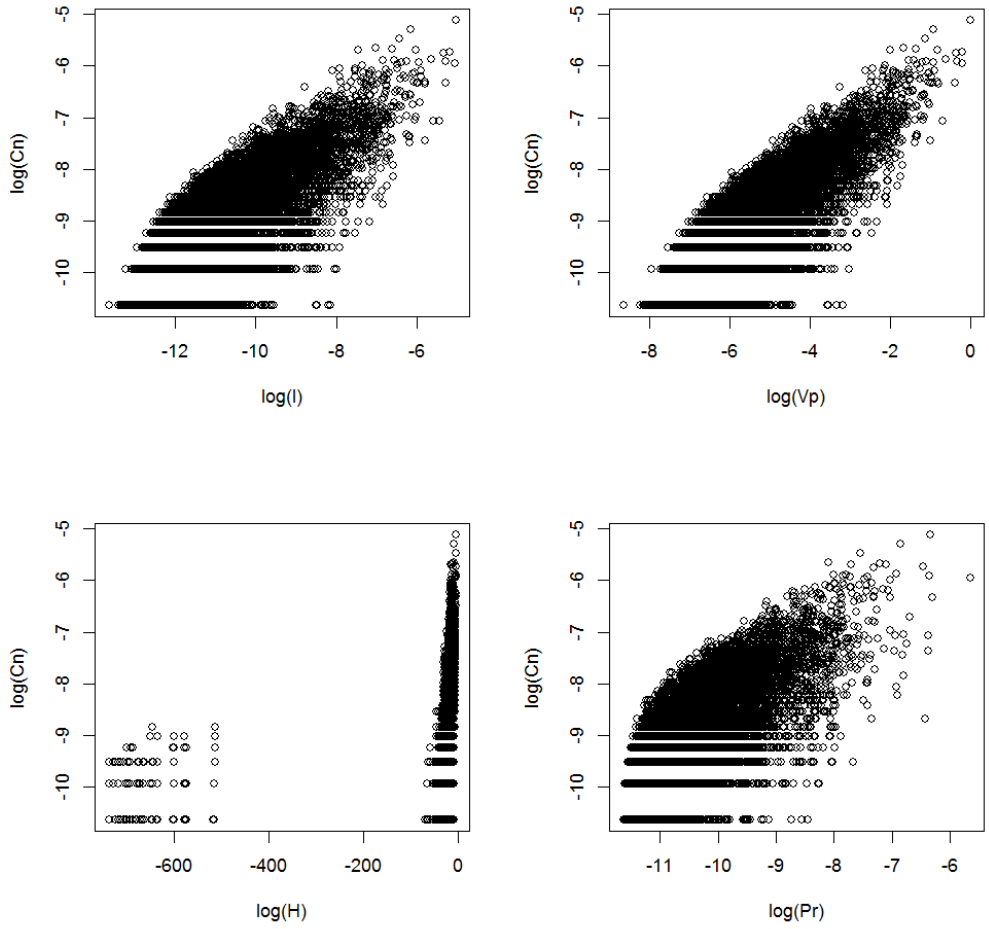


Figura 3.10: Diagrama de dispersión entre las medidas más correlacionadas con las citas en GcPRE en escala log-log.

Visualmente es muy sugerente la alta correlación entre *Cn*, *I*, *Vp* y *Pr*. En la siguiente sección ahondaremos en este punto.

3.5.7. PRSTAB

Para PRSTAB la gráfica se denota por *GPRSTAB* y la componente gigante por *GcPRSTAB*. GPRSTAB tiene 1,654 nodos (artículos) y 3,705 aristas (citas). GcPRSTAB tiene 1,463 nodos y 3,483 aristas, con 2.38 citas en promedio. Hay 1,036 artículos con al menos una cita, lo que implica que el 29.18 % de los nodos no tienen citas. En la tabla 3.11 se muestra la matriz de correlaciones para el conjunto con al menos una cita.

Medidas	*Cn*	*I*	*H*	*Pr*	*Vp*	*Cl*	*Be*
Cn	1.00	0.85	0.58	0.74	0.71	0.30	0.41
I	0.85	1.00	0.73	0.80	0.97	0.38	0.28
H	0.58	0.73	1.00	0.47	0.69	0.10	0.26
Pr	0.74	0.80	0.47	1.00	0.77	0.40	0.20
Vp	0.71	0.97	0.69	0.77	1.00	0.36	0.20
Cl	0.30	0.38	0.10	0.40	0.36	1.00	0.14
Be	0.41	0.28	0.26	0.20	0.20	0.14	1.00

Tabla 3.11: Matriz de correlaciones entre las medidas en GcPRSTAB.

En la tabla 3.11 se observa que la *I* tiene la correlación más alta con *Cn* (85 %) y en esta ocasión *I* es la medida más correlacionada con el resto, nuevamente *H* es la cuarta más correlacionada con *Cn* de manera similar a PRE. En la figura 3.11 se muestra el diagrama de dispersión entre las medidas más correlacionadas con las citas en el conjunto con al menos una cita.

De acuerdo con la figura 3.11 en este caso es aún más sugerente la alta correlación entre *Cn*, *I*, *Vp* y *Pr*. Por supuesto la cantidad de datos es mucho menor en PRSTAB que las revistas anteriores.

3.5.8. PRSTPER

Esta sección finaliza con la última subred de APS que corresponde a PRSTPER y es la red más pequeña pero se incluye porque es la única revista que no tiene citas de otras revistas, por supuesto se entiende que sea debido a que los datos son de los primeros 8 años de la revista.

Para PRSTPER la gráfica se denota por *GPRSTPER* y la componente gigante por *GcPRSTPER*. GPRSTPER tiene 221 nodos (artículos) y 605 aristas (citas). GcPRSTPER tiene 217 nodos y 603 aristas, con 2.77 citas en promedio. Hay 145 artículos con al menos una cita, lo que implica que el 33.17 % de los nodos no tienen citas. En la tabla 3.12 se muestra la matriz de correlaciones para el conjunto con al menos una cita.

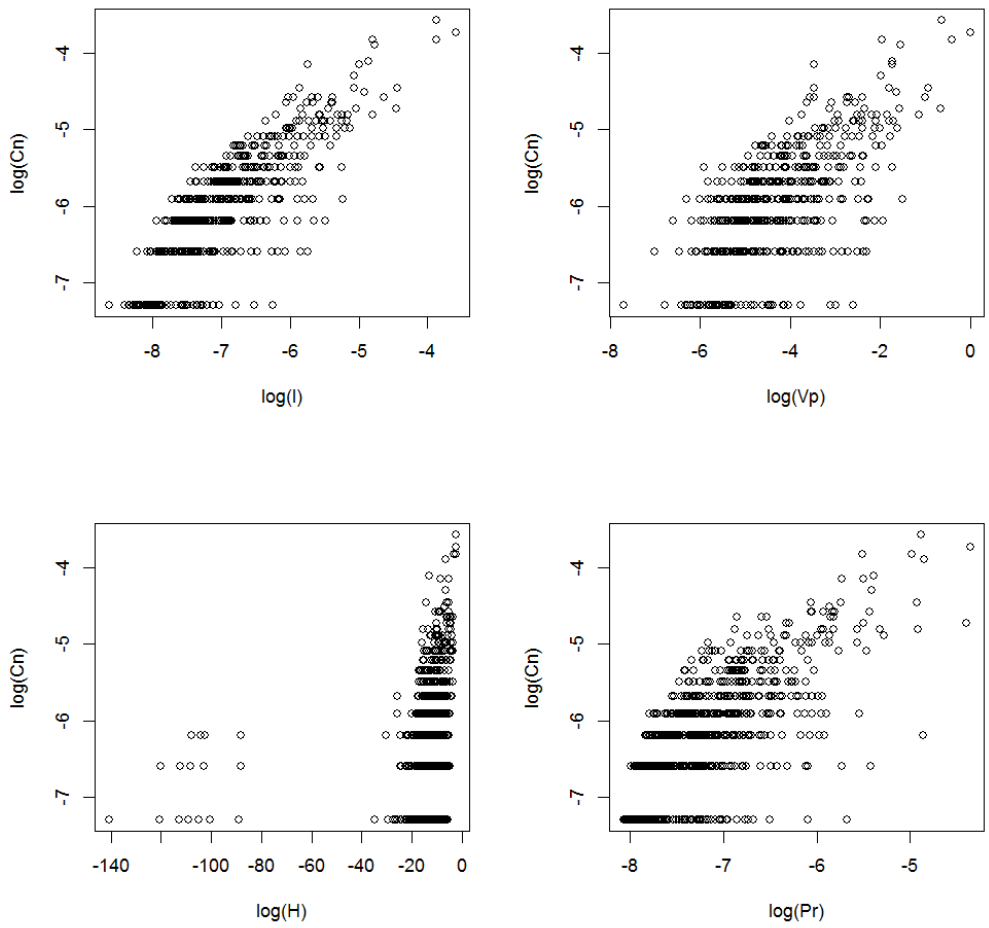


Figura 3.11: Diagrama de dispersión entre las medidas más correlacionadas con las citas en GcPRSTAB en escala log-log.

Medidas	Cn	I	H	Pr	Vp	Cl	Be
Cn	1.00	0.931	0.873	0.798	0.725	0.303	0.065
I	0.931	1.00	0.799	0.848	0.902	0.452	0.032
H	0.873	0.799	1.00	0.689	0.545	0.169	0.007
Pr	0.798	0.848	0.689	1.000	0.817	0.427	-0.044
Vp	0.725	0.902	0.545	0.817	1.00	0.534	0.001
Cl	0.303	0.452	0.169	0.427	0.534	1.00	0.056
Be	0.065	0.032	0.007	-0.044	0.001	0.056	1.00

Tabla 3.12: Matriz de correlaciones entre las medidas en GcPRSTPER.

En la tabla 3.12 se observa que la I tiene la correlación más alta con Cn (93 %) y nuevamente I es la medida más correlacionada con el resto, de manera similar a PRSTAB. En la figura 3.12 se muestra el diagrama de dispersión entre las medidas más correlacionadas con las citas en el conjunto con al menos una cita.

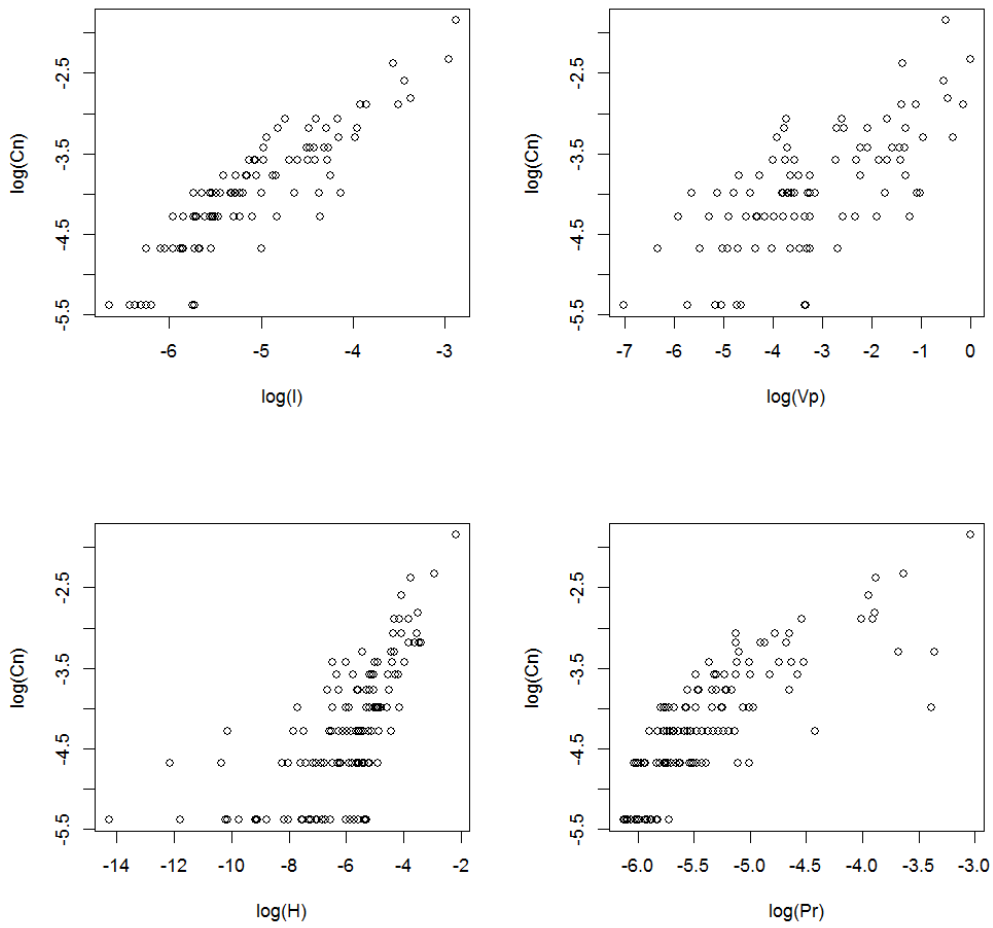


Figura 3.12: Diagrama de dispersión entre las medidas más correlacionadas con las citas en GcPRSTPER en escala log-log.

Como se ha mostrado en prácticamente todas las redes de citación analizadas la propuesta de medir la influencia de un artículo a través de la influencia (ecuación 3.4) es útil para diferenciar en muchos casos donde Vp no alcanza a diferenciar. También se ha puesto en evidencia la alta correlación entre la I y Vp, en la siguiente sección se ahonda en ello, pero en principio Vp es la medida más correlacionada con el resto, mostrando la potencia de asociar un invariante de la gráfica como lo es el valor propio y el vector propio correspondiente, por supuesto, se entiende que Pr esencialmente es una variación de Vp, y en parte H, pero no es tan clara la correlación con I.

3.6. Análisis de las correlaciones

De acuerdo con los resultados de la sección anterior Vp es una de las principales medidas en la red de citación, en ese sentido conviene analizar más detalladamente

esta medida y su relación teórica con *I*.

Como se mencionó en 3.2.3 *Vp* se basa en el grado de los nodos vecinos. Si x_i es la centralidad del nodo i, entonces se busca que x_i sea proporcional a la suma de las centralidades de los vecinos de i, esto es:

$$\lambda x_i = \sum x_k \tag{3.5}$$

tal que k es vecino de i.

La ecuación 3.5 se puede expresar como:

$$\lambda x = Ax \tag{3.6}$$

Donde A es la matriz de adyacencia de la gráfica y x es un vector cuyas componentes son x_i. Una condición necesaria y suficiente para la existencia de una solución no trivial es $|\lambda I - A| = 0$, es decir, los posibles factores de proporcionalidad λ son idénticos a los valores propios de A.

De la ecuación 3.5 se puede relacionar el hecho de contar caminos (influencia) con la medición de la centralidad de un nodo a partir de sus vecinos. Pero se puede dar una prueba formal de cotas para la influencia y como se asemeja a *Vp* o a *Cn* en el caso de las redes de citación.

Proposición 2 (Cota para la influencia) *Sea $\vec{G}$ débilmente conexa, G la gráfica subyacente de $\vec{G}$, A la matriz de adyacencia de G y $\vec{A}$ la de $\vec{G}$, $I_i(\beta)$ la influencia de i en $\vec{G}$ y $TC_i(\beta)$ la comunicabilidad total de i en G. Entonces:*

$$I_i(\beta) \leq \frac{TC_i(\beta)}{w(\vec{G})}$$

Prueba 2 *Es casi inmediato ver la cota, como $I_i(\beta) = \frac{1}{w(\vec{G})} \sum_{k \geq i} \vec{c}_{ik}(\beta)$, y $TC_i(\beta) = \sum_j (e^{\beta A})_{ij}$ entonces sólo es necesario comparar los numeradores y es claro que:*

$$\sum_k (e^{\beta \vec{A}})_{ik} \leq \sum_j (e^{\beta A})_{ij}. \blacksquare$$

En el contexto de las redes de citación y en general de redes reales, los casos de las igualdades no tienen demasiado interés, aparecen en casos extremos como cuando la red tiene un sólo nodo.

Una relación entre la influencia con *Vp* y *Cn* se desprende del siguiente teorema:

Teorema 2 *Sea $\vec{G}$ débilmente conexa, G la gráfica subyacente de $\vec{G}$, $I_i(\beta)$ la influencia de i en $\vec{G}$ y $I(\beta)$ el vector de las influencias, entonces:*

1. *Cuando $\beta \to 0+$ la clasificación producida por $I(\beta)$ es menor o igual a Cn, el vector de la centralidad de grado en G.*

2. *Cuando $\beta \to \infty$ la clasificación producida por $I(\beta)$ es menor o igual a Vp, el vector de la centralidad por vector propio en G.*

Prueba 3 *La prueba sigue de la Proposición 2 y de un teorema de [13] que se enuncia a continuación como lema:*

Lema 1 *Sea G conexa, $TC_i(\beta)$ la comunicabilidad total del nodo i y $TC(\beta)$ el vector de las comunicabilidades totales, entonces:*

1. *Cuando $\beta \to 0+$ la clasificación producida por $TC(\beta)$ converge a la que produce Cn, el vector de las centralidades de grado.*

2. *Cuando $\beta \to \infty$ la clasificación producida por $TC(\beta)$ converge a la que produce Vp, el vector de las centralidades por vector propio.*

Luego por un resultado básico de cálculo se sabe que si $g(x) \leq h(x)$ y $\lim_{x\to a} h(x) = L$ entonces $\lim_{x\to a} g(x) \leq L$

En este caso, por la proposición 2 se tiene la cota y del lema 1 se tiene el límite, y por lo tanto la demostración. ■

El Teorema 2 dice qué tanto la influencia de los nodos se parece a la centralidad de grado o de vector propio, pero de la gráfica subyacente de G no de $\vec{G}$, sin embargo es un acercamiento formal a la intuición de porque las correlaciones son tan altas entre estas medidas. En [13] se demuestra que el resultado del Lema 1 es válido para redes dirigidas fuertemente conexas, las redes de citación en general son débilmente conexas por tanto hasta ahora sólo se tienen aproximaciones.

De acuerdo con [13], otro acercamiento para entender las correlaciones es a través de la brecha espectral, b_s (spectral gap), ésta es la diferencia entre los dos valores propios más grandes de la matriz A:$|\lambda_1 - \lambda_2|$. En caso de que sean complejos, la brecha espectral es la diferencia entre los módulos de los dos valores propios más grandes $|\lambda_1| - |\lambda_1|$. Si la brecha espectral es grande entonces $I(\beta), SC(\beta), TC(\beta)$ tienden a la centralidad por vector propio, sin embargo, para las redes de la APS que se trabajaron no es posible diferenciar con este método las correlaciones entre $I(\beta)$ y Vp como se muestra en la tabla 3.13.

Como se ve en la tabla 3.13 la brecha de *GcPRB* es la mitad de la brecha de *GcAPS* y sin embargo las correlaciones son casi las mismas, 0.66 y 0.6 respectivamente, en este sentido es que la brecha espectral no permite diferenciar, además si observa la brecha de *GcPRE, GcPRSTAB, GcPRSTPER*, en los tres casos es cero porque los valores propios son los mismos, por lo que I y Vp deberían ser muy diferentes, pero no es el caso.

Red	λ_1	λ_2	Brecha espectral	$\rho_{I,Vp}$
GcAPS	8.4	7.7	0.69	0.60
GcPR	3.0	2.4	0.60	0.90
GcPRA	2.4	2.3	0.11	0.91
GcPRB	3.1	2.8	0.32	0.66
GcPRC	3.4	2.4	1.00	0.89
GcPRD	3.0	2.9	0.11	0.82
GcPRE	2.0	2.0	0.00	0.99
GcPRSTAB	1.0	1.0	0.00	0.97
GcPRSTPER	1.0	1.0	0.00	0.90

Tabla 3.13: Brecha espectral de las redes de APS comparada con la correlación (ρ) entre I y Vp.

Cabe señalar que en [66] se considera la comunicabilidad total como una alternativa para encontrar Hubs, lo cual muestra la correlación con esta medida.

De acuerdo con el Teorema 2 y que en las redes de citación $\beta = 1$ se pueden explicar las correlaciones altas entre el grado o citas (Cn) y las otras medidas (I,Vp,Pr,H). Esto permite entender las correlaciones, sin embargo el uso de una u otra medida dependerá del contexto, en este caso es claro que la influencia permite diferenciar en escalas más pequeñas y esa es una ventaja para clasificar y diferenciar entre dos nodos con el mismo grado, además de la definición misma de la medida.

En particular para las redes de citación, después del análisis con los datos de APS, lo más adecuado sería trabajar en dos dimensiones usando la influencia y las citas para clasificar a los artículos con más impacto en la red de citación. En general para cualquier red, cuando se desea diferenciar la importancia de dos nodos con el mismo grado se puede usar cualquiera de las medidas mencionadas en la subsección 3.2.3, dependiendo del uso que se le quiera dar a la clasificación.

Capítulo 4

Redes de citación en México

Como se mencionó en el capítulo 1 la importancia de analizar la productividad científica de un país es fundamental para identificar cambios y/o mejoras en la política científica; además permite integrar información a las bases de datos, como el Atlas de la Ciencia Mexicana, y entender mejor la evolución de la ciencia, en este caso de México.

Hasta donde se tiene información, las redes de citación mexicanas (aquellas que se generan entre artículos con al menos un autor con adscripción a una institución mexicana), han sido poco exploradas, tanto desde la perspectiva matemática y/o estadística como desde la perspectiva social. En esa dirección y a partir de las herramientas utilizadas para analizar la influencia en una red de citación (capítulo 3), se realizó un estudio de las características de las redes de citación mexicanas en tres áreas: física, matemáticas y química. Además se analizó la evolución de estas redes en el tiempo, desde 1970 a 2015, así como el cociente de citas locales sobre citas globales. Los resultados se presentan en este capítulo.

4.1. Datos y Métodos

4.1.1. Selección de Datos

Los datos fueron obtenidos de la Web of Science, para ello se realizó una búsqueda por área y luego se filtró por país considerando las publicaciones hasta 2015.

Como notación, se considera como conjunto universal U, el total de publicaciones registradas en la colección principal de la WoS, U_m es el conjunto de todas las publicaciones de matemáticas en U, U_f las de física y U_q las de química. El conjunto de publicaciones con al menos un autor con adscripción a una institución mexicana se denota por $M \subset U$, $M_m \subset M$ es el conjunto donde están las publicaciones del área de matemáticas M_f es el de física y M_q el de química. Por supuesto $M_m \subset U_m, M_f \subset U_f, M_q \subset U_q$. Aunque en la WoS las publicaciones están por área, en general sucede que $M_m \cap M_f \cap M_q \neq \emptyset$ ya que una publicación puede pertenecer a dos, tres o más conjuntos, dependiendo de la revista.

En la tabla 4.1 se muestran los datos de las publicaciones totales de cada área en la WoS, así como el periodo de tiempo que abarcan.

Área	Publicaciones	Periodo de tiempo
M_f	30291	1939-2015
M_m	8573	1973-2015
M_q	19 150	1912-2015

Tabla 4.1: Resumen de los datos descargados de WoS para México.

Aunque en M_f y M_q hay datos anteriores a 1970 sólo se consideran los datos a partir de 1970 ya que en el caso de M_f sólo hay cuatro publicaciones antes de 1970 y en M_q sólo hay tres publicaciones antes de 1970. Realmente hay más artículos anteriores a 1970, pero en la WoS sólo aparecen esos con dirección en México.

Para tener las citas por año de cada artículo dentro de U, se generó un informe de citas en WoS. En el caso de M_m como son menos de 10,000 registros se puede hacer automáticamente, para M_f y $M_q$se dividieron en subconjuntos de 5000 y de esa manera se obtuvo la información global.

4.1.2. Métodos

Para las tres áreas se dividieron los datos en periodos de 5 años, con ello se generaron los conjuntos de datos acumulados de cada 5 años, con esos conjuntos se hicieron las redes de citación para cada intervalo de tiempo. Para hacer las redes se uso Sci2 [67] y para la visualización y las estadísticas Gephi [65], R [68] y NetworX [63] en Python.

Una vez que se tenía cada red se obtuvieron las citas locales entre esas publicaciones hasta el año en cuestión, se añadieron las citas externas generadas por la WoS para comparar entre la citación interna y la externa.

Después se obtuvieron las estadísticas descriptivas de cada red en el tiempo: componente gigante, diámetro, longitud media de camino, entre otras, para todas estas medidas se utilizó Gephi.

El siguiente paso fue ajustar un modelo a la distribución de citas (grado de salida) de las redes de cada área con los datos totales, tanto para las citas locales como las globales. Por citas locales (Cl) se entienden aquellas que se otorgan entre elementos de M_f, M_m o M_q. Las citas globales (Cg) son aquellas que se otorgan entre elementos de U.

Para identificar si la distribución tenía una cola pesada y seguía una ley de potencia u otra distribución, como log-norm o exponencial se usaron los métodos propuestos por [36] que se pueden resumir como:

- Estimar los parámetros de la distribución: si es una ley de potencia el x_{min} y γ (potencia), si es una log norm el x_{min}, μ (la media) y σ (la desviación).

- Se calculan la bondad de ajuste entre los datos y la distribución supuesta. Si el resultado es un $p - value$ mayor o igual que 0.1, entonces la distribución en cuestión se considera una hipótesis plausible para los datos, de no ser así se rechaza. Para ello se generan conjuntos de datos (al menos 100) con la supuesta distribución y se comparan las diferencias entre los datos reales y los generados sintéticamente.

- Si más de una distribución es plausible ($p - value > 0.1$) se comparan las distribuciones para determinar cuál es un mejor ajuste usando la razón de verosimilitud.

Finalmente se calculó la influencia de cada artículo en la red de citación local y se comparó con las citas.

4.2. Análisis de redes

Se sigue la misma notación que en el capítulo 3 y se considera red de citación como una digráfica donde los artículos son los vértices y las aristas las citas. Entonces si un artículo B cita a un artículo A se denota como (A,B), visualmente $A \to B$.

4.2.1. Redes de citación en el tiempo

A continuación se presentan los resultados del cálculo de nodos (N), aristas (L), nodos en la componente (N_G), la fracción de N_G sobre al total (N_G/N), nodos aislados (N_a) y el grado promedio de salida ($\langle k \rangle$) o citas promedio. Para cada área están los datos desde 1970 hasta 2015 en periodos de tiempo acumulados de 5 años. La tabla 4.2 corresponde a M_f, la tabla 4.3 a M_m y la tabla 4.4 a M_q.

De acuerdo con las tablas 4.2,4.3 y 4.4, M_f y M_q tienen $\langle k \rangle > 1$ en los últimos años, mientras que M_m no, esta observación es fundamental en el análisis de las gráficas o redes correspondientes ya que permite identificar cuando aparece una componente gigante en una red.

En el modelo de Erdös-Renyi (E-R) [28] se establece que en $\langle k \rangle = 1$ se tiene un punto crítico para el surgimiento de una componente gigante, antes de este valor no hay tal componente; a esto se le llama un régimen subcrítico. Cuando $\langle k \rangle > 1$ se le llama un régimen supercrítico. Cuando $\langle k \rangle >> ln(N)$, prácticamente todos los nodos están conectados, casi se tiene una red conexa.

Interval	N	L	N_a	N_G	N_G/N	$\langle k \rangle$
1970-1975	184	50	123	13	0.071	0.27
1970-1980	750	618	322	40	0.053	0.82
1970-1985	1531	1673	582	173	0.113	1.09
1970-1990	2428	2986	809	455	0.187	1.23
1970-1995	4544	5618	1526	1449	0.319	1.24
1970-2000	8808	11881	2791	3636	0.413	1.35
1970-2005	14816	22560	4535	7271	0.491	1.52
1970-2010	21952	37636	6089	12007	0.547	1.71
1970-2015	30150	58510	7879	17942	0.595	1.94

Tabla 4.2: Datos para M_f por intervalo de tiempo acumulado.

Interval	N	L	N_a	N_G	N_G/N	$\langle k \rangle$
1970-1975	31	3	25	2	0.065	0.097
1970-1980	123	28	88	6	0.049	0.228
1970-1985	293	82	192	9	0.031	0.280
1970-1990	562	211	330	23	0.041	0.375
1970-1995	997	489	558	44	0.044	0.490
1970-2000	1998	1133	1080	73	0.037	0.567
1970-2005	3610	2444	1699	110	0.030	0.677
1970-2010	5898	4660	2590	170	0.029	0.790
1970-2015	8568	8119	3391	282	0.033	0.948

Tabla 4.3: Datos para M_m por intervalo de tiempo acumulado.

Interval	N	L	N_a	N_G	N_G/N	$\langle k \rangle$
1970-1975	119	36	81	15	0.126	0.30
1970-1980	480	151	332	21	0.044	0.31
1970-1985	928	402	556	50	0.054	0.43
1970-1990	1825	942	1097	87	0.048	0.52
1970-1995	3174	2086	1757	218	0.069	0.66
1970-2000	5420	4717	2662	623	0.115	0.87
1970-2005	8726	9332	3728	2529	0.290	1.07
1970-2010	13283	17482	5046	4932	0.371	1.14
1970-2015	19098	28778	6594	9140	0.479	1.51

Tabla 4.4: Datos para M_q por intervalo de tiempo acumulado.

Mientras el grado promedio no es mayor que 1, los nodos no se organizan en una red identificable, podría ser una red tipo E-R, libre de escala o de cualquier otro tipo. De acuerdo con [28] las redes libres de escala están en el régimen supercrítico.

En la figura 4.1 se muestra la comparación del $\langle k \rangle$ contra N_G/N, es decir, cómo varía la proporción de nodos en la componente mayor respecto al grado promedio en

cada una de las áreas, se observa que en $\langle k \rangle > 1$ empieza a surgir la N_G en M_f y M_q y es claro cómo en M_m no se ha llegado al punto crítico.

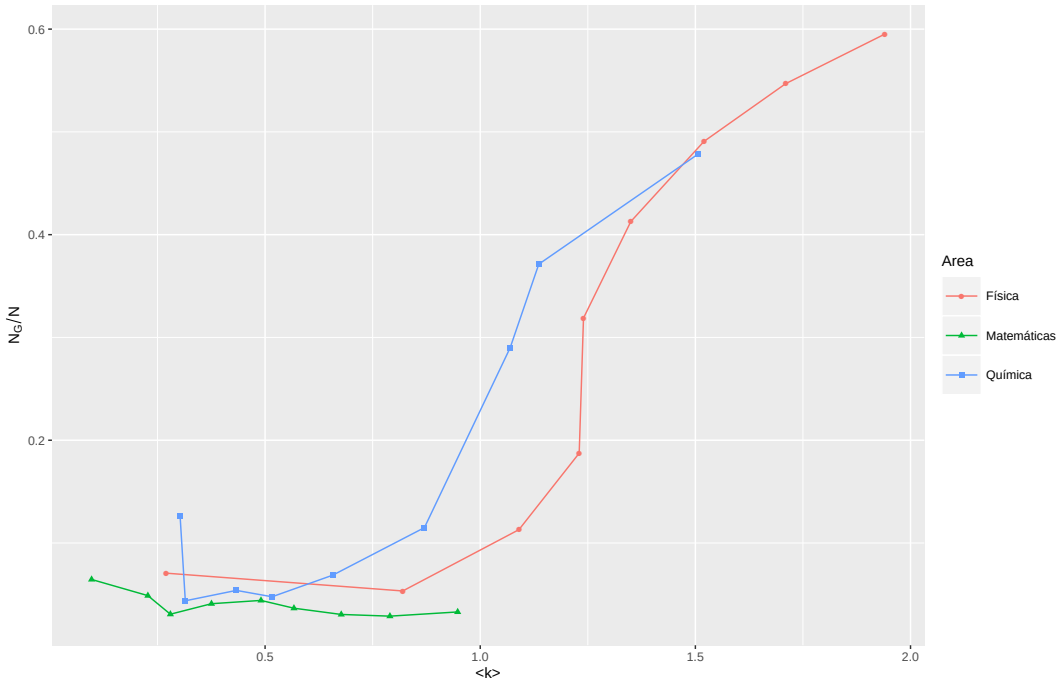


Figura 4.1: Comparación entre el grado promedio de citas y la proporción de nodos en la componente mayor por cada área.

La ausencia de una componente gigante en M_m implica que no hay "hubs"(nodos con grados significativamente mayores que el resto) y por tanto la red no es libre de escala como se verá en los ajustes a la distribución de citas en la siguiente subsección, aunque el $\langle k \rangle$ y N_G permiten hacer análisis iniciales.

La implicación de que M_m no sea libre de escala tiene connotaciones similares a las de otras redes reales donde no se presenta este comportamiento como: algunas en ciencias de materiales donde cada nodo tiene exactamente el mismo grado (k), la red neuronal del gusano C.elegans o la red eléctrica que consiste de generadores e interruptores conectados por lineas de transmisión, en estos tres ejemplos la propiedad de ser libres de escala está ausente porque los sistemas limitan el numero de conexiones que un nodo puede tener, lo cual restringe el tamaño máximo de los hubs [28].

En el caso de M_m dicha restricción podría estar asociada al tamaño mismo de la red, ya que es menos de la mitad de M_q y menos de un tercio de M_f, sin embargo M_f tuvo un $\langle k \rangle > 1$ cuando tenía 1531 nodos y M_q cuando tenía 8726 nodos; entonces el tamaño de la red no es la razón, lo cual implica que en efecto hay algún mecanismo extra en M_m que impide un $\langle k \rangle > 1$, entonces puede ser que por características de la disciplina misma se "restringa" el número de citas, ya que en matemáticas los

artículos tardan más tiempo en tener impacto, además, los promedios de citas son en general pequeños comparados con el resto de las disciplinas, prueba de ello son los factores de impacto (FI) en revistas de matemáticas donde el FI más alto es 7.36 y el que le sigue es de 4.17 y el factor de impacto medio es de 0.84 lo cual evidencia que el promedio de citas en general es bajo en U_m.

De acuerdo con los datos del $\langle k \rangle$ de M_m en la tabla 4.3, para 2020 se podría observar el surgimiento de una componente gigante y por tanto habrá cambiado la interacción entre los matemáticos mexicanos conectándose las áreas que por ahora están dispersas.

Las hipótesis expuestas para explicar las diferencias entre estas tres áreas, pueden servir como herramientas de política científica para identificar tanto un área que interactuá poco como las propuestas potenciales para impulsar su interacción. Una herramienta especifica en esta dirección es el proyecto análisis topológico de datos (http://atd.cimat.mx/) que involucra áreas que en hasta el momento no se citan mutuamente como son: la topología algebraica, la probabilidad, la estadística y la computación. Pero sólo en unos años se podrá contrastar esto.

4.2.2. Distribución de grado

En el análisis de redes una pregunta central es cuál es la distribución de grado, por ejemplo, [28] postula que la mayoría de las redes reales siguen leyes de potencia, debido a que surgen hubs con grados muy altos respecto al resto de los nodos, lo que permite observar distribuciones con colas pesadas, específicamente leyes de potencia. De acuerdo con [28] el mecanismo para que surjan los hubs es debido al "preferent attachment" que esencialmente consiste en que los nodos con más citas seguirán acumulando más citas de los nodos que se agreguen a la red.

Aquí se comparan dos distribuciones, ley de potencia y log-norm, ya que las implicaciones de una u otra son de interés para el contexto de las citas. La ley de potencia implica un comportamiento libre de escala, donde la media no es útil para determinar la probabilidad de un nodo elegido al azar, ya que la desviación no sirve para escalarla, teóricamente puede no existir. Por otra parte, una log-norm implica que los nodos con citas mucho mayor que la media, estos son valores atípicos (outliers) y tales nodos no son representativos, y la media y desviación de las citas son útiles.

La distribución de ley de potencia es del tipo $p(x) = x^{-\gamma}$, si se obtienen logaritmos se representa como $ln(P(x)) = -\gamma ln(x)$ así en escala log-log la ley de potencia se ve como una recta, donde la pendiente es γ.

En general, los ajustes a ley de potencia se hacen a partir de un valor dado, llamado x_{min} entonces se suele decir que el ajuste es en la cola de la distribución

como se observa en las figuras 4.2,4.3 y 4.4.

En el caso de la log-norm se tiene $p(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}$, donde μ es la media y σ la desviación estándar. De manera similar a la ley de potencia, se ajusta a partir de un x_{min}

Sólo para el caso de las citas locales se considera la distribución exponencial $e^{-\lambda x}$ porque las distribuciones de M_m y M_q caen más pronto que la ley de potencia y la log-norm, lo cual prácticamente las descarta, y se tienen redes que no son libres de escala.

En la figura 4.2 se muestra la distribución de grado acumulada de las citas en M_f, en las citas globales se tienen 21 802 registros con al menos una cita y en las citas locales 14 268. En ambos casos se incluyen ajustes, de ley de potencia y log-norm.

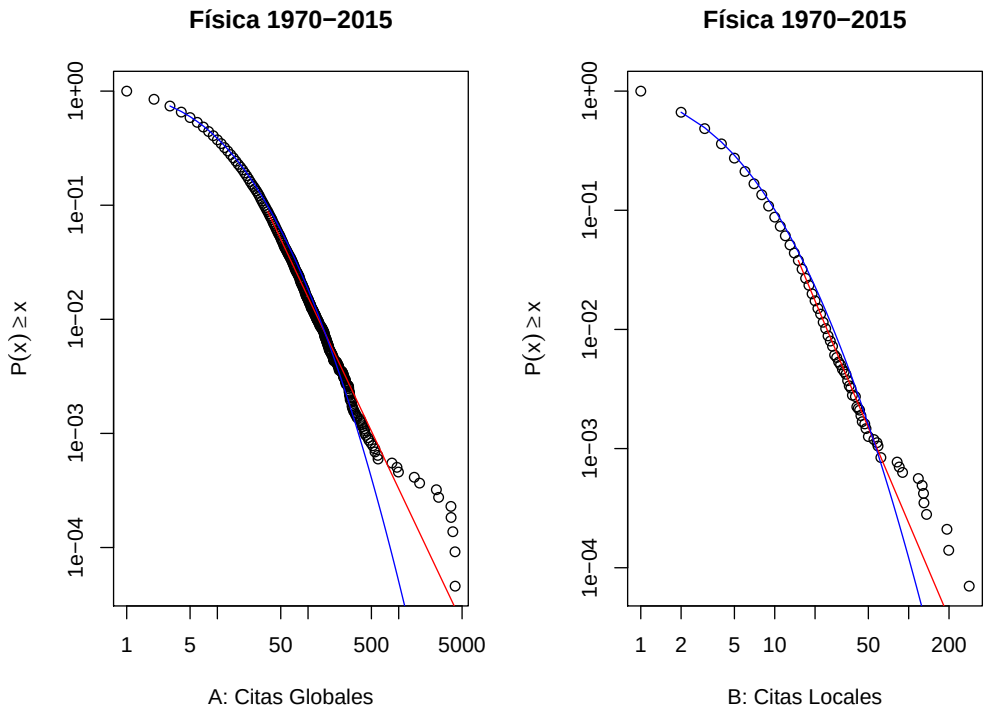


Figura 4.2: Ajustes para física. En rojo el ajuste de ley de potencia con $\gamma = 2.68$ para A y $\gamma = 3.64$ para B. En azul el ajuste log-norm con $\mu = 1.88, \sigma = 1.29$ para A y $\mu = 0.98, \sigma = 0.97$ para B

En el caso de M_f hay 17 artículos con 500 citas o más y 9 con más de 1000 citas, de estos últimos sólo un artículo no forma parte de las grandes colaboraciones en altas energías.

En la figura 4.3 se muestra la distribución de grado acumulada de las citas de publicaciones de matemáticas con al menos una cita, en las citas globales se tienen 5,398 registros y en las citas locales 3,102.

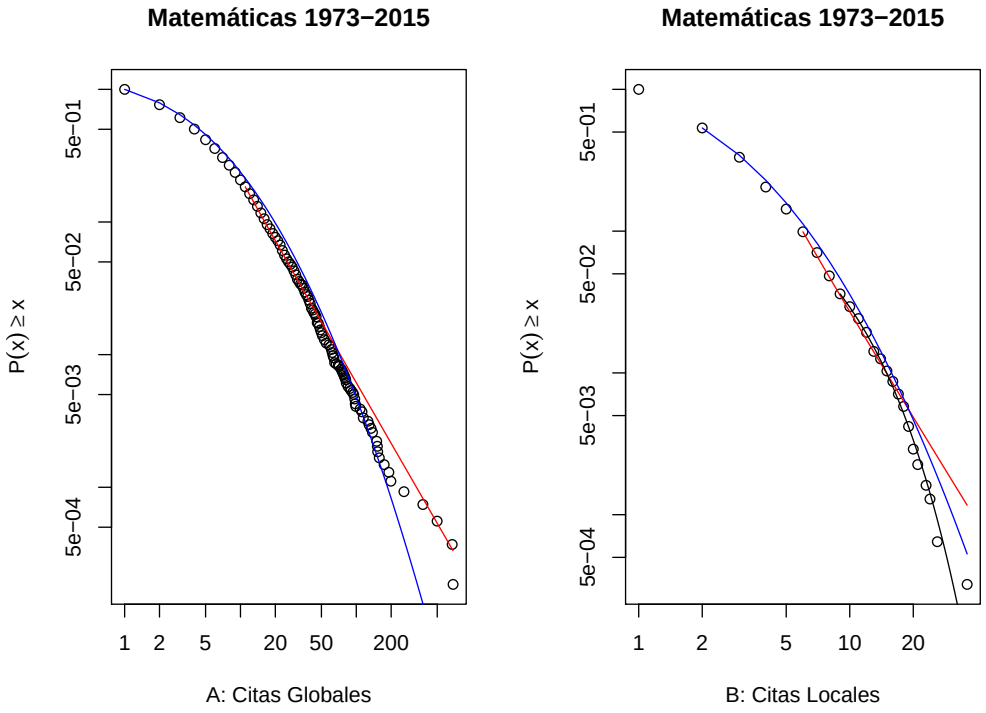


Figura 4.3: Ajustes para matemáticas. En rojo el ajuste de ley de potencia con $\gamma = 2.51$ para A y $\gamma = 3.39$ para B. En azul el ajuste log-norm con $\mu = 1.11, \sigma = 1.29$ para A y $\mu = 0.61, \sigma = 0.88$ para B. Sólo para B se muestra en negro el ajuste exponencial con $\lambda = 0.21$.

En la figura 4.4 se muestra la distribución de grado acumulada de las citas de publicaciones de química con al menos una cita, en las citas globales se tienen 14,756 registros y en las citas locales 8,285.

En la tabla 4.5 se muestran los resultados de los ajustes y las simulaciones para las redes con el total de los datos, es decir de 1970 a 2015. Los $p-value$ aparecen en la columna p, si $p \geq 0.1$ son significativos y se muestran en negritas. Para los ajustes se usaron los métodos propuestos por [36] implementados en R por [69].

Red	Ley de potencia			log-norm			
	γ	x_{min}	p	μ	σ	x_{min}	p
M_f Global	2.67 ± 0.01	38 ± 1.6	**0.16**	1.87 ± 0.02	1.29 ± 0.005	5 ± 2	0.02
M_f Local	3.64 ± 0.05	15.5 ± 0.5	**0.19**	0.98 ± 0.03	0.97 ± 0.005	2.4 ± 0.2	**0.76**
M_m Global	2.51 ± 0.01	12 ± 1	0.01	1.10 ± 0.02	1.29 ± 0.01	1.5 ± 0.5	**0.1**
M_m Local	3.40 ± 0.5	6.3 ± 0.2	0.01	0.5 ± 0.1	0.9 ± 0.02	2.2 ± 0.2	0.03
M_q Global	3.14 ± 0.4	55 ± 5	**0.68**	2.20 ± 0.02	1.1 ± 0.02	7 ± 1	**0.12**
M_q Local	3.47 ± 0.3	10 ± 1	0.0	1.08 ± 0.03	0.86 ± 0.02	4 ± 0.5	**0.28**

Tabla 4.5: Resultados de los ajustes para cada área.

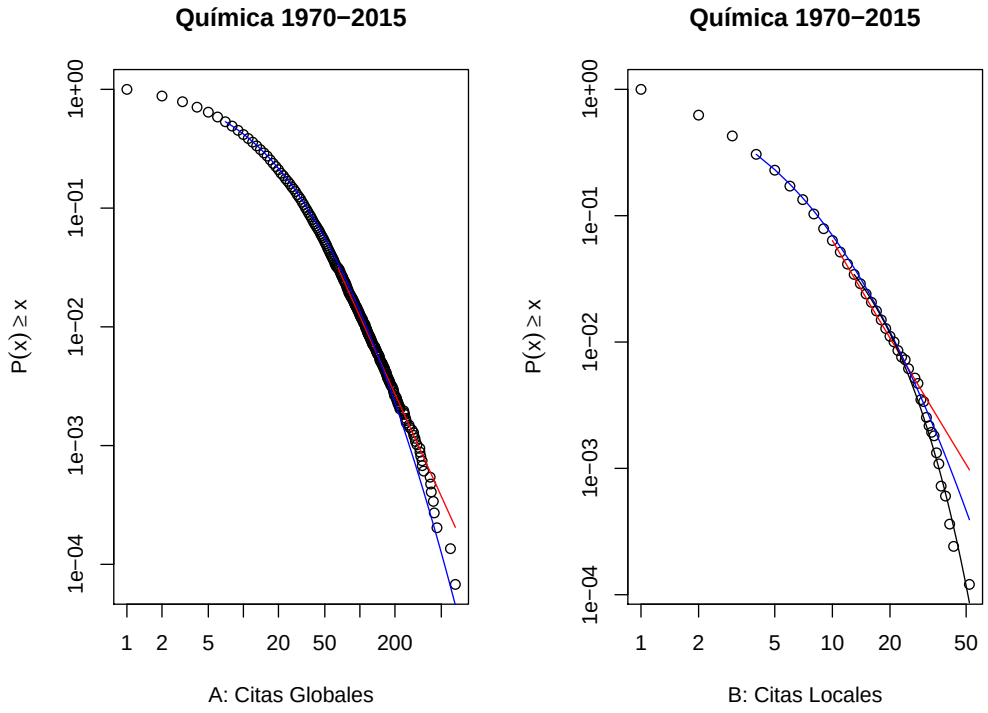


Figura 4.4: Ajustes para química. En rojo el ajuste de ley de potencia con $\gamma = 3.16$ para A y $\gamma = 3.47$ para B. En azul el ajuste log-norm con $\mu = 2.22, \sigma = 1.1$ para A y $\mu = 1.14, \sigma = 0.84$ para B. Para B, se muestra en negro el ajuste exponencial con $\lambda = 0.15$.

Como para M_m local no es significativo el ajuste ley de potencia ni log-norm se realizó un ajuste exponencial, los resultados de la simulación son: $\lambda = 0.24 \pm 0.04$, $x_{min} = 8 \pm 2$ y un $p - value = 0.88$.

De acuerdo con la tabla 4.5 en M_q local es significativa la log-norm, pero de acuerdo con la figura 4.4 los datos caen más rápido, es decir, la exponencial ajusta mejor la caída y con base en [28] es suficiente con que la caída sea más rápida que ley de potencia o log-norm para determinar que la red no es libre de escala. Los resultados de la simulación con el ajuste exponencial son: $\lambda = 0.155 \pm 0.01$, $x_{min} = 13 \pm 2$ y un $p - value = 0.14$.

Los resultados del ajuste para M_q local son muy representativos porque indican que aunque la red tiene una componente gigante, los nodos con grados altos son valores atípicos y no representativos de la red, por tanto la red no es libre de escala y su comportamiento es similar al caso de M_m sólo que M_q está en un nivel de desarrollo más avanzado desde el punto de vista estructural de una red libre de escala; o posiblemente nunca será libre de escala, teniendo algún mecanismo que impida la aparición de los hubs.

En el caso de las citas locales para M_f y las citas globales para M_q fue necesario

hacer un comparativo entre dos distribuciones. Para poder comparar dos distribuciones se tiene que fijar el mismo x_{min} para ambas. Cabe señalar que en esta parte la interpretación del $p - value$ es la tradicional, valores cercanos a cero son favorables y cercanos a 1 desfavorables, en ese sentido se considera significativo si es menor a 0.1.

Para M_f local, se fijó $x_{min} = 15$, considerando el mínimo para el ajuste de ley de potencia, el resultado del test no es concluyente ya que se obtiene una razón de verosimilitud de 1.12 a favor de la ley de potencia pero con un $p - value = 0.86$. Sin embargo, cuando se fija en 2, el mínimo para el ajuste de log-norm, el resultado es concluyente a favor de la log-norm con un razón de verosimilitud igual a -22.3 y un $p - value$ de $1.61e - 110$, es decir, prácticamente cero.

En el caso de M_q global, se fijó $x_{min} = 65$, considerando el mínimo para el ajuste de ley de potencia, el resultado del test no es concluyente ya que se obtiene una razón de verosimilitud de -1.04 a favor de la log-norm, pero con $p - value = 0.148$. Sin embargo cuando se fija en 7, el mínimo para el ajuste de log-norm, el resultado es concluyente a favor de la log-norm con un razón de verosimilitud igual a -19 y un $p - value$ de $1.23e - 80$, es decir, prácticamente cero.

Por tanto, se tiene que el ajuste log-norm es mejor que el de ley de potencia, salvo en un caso que es el de las citas globales para M_f que sobresale por las grandes colaboraciones. Sin ellas, la distribución de citas globales sería muy parecida a M_q. Pero con base en [28], estas comparaciones, aunque estadísticamente bien hechas tampoco son concluyentes para afirmar, por ejemplo, que M_q global no es libre de escala. Muchas veces se puede ajustar una distribución como log-norm por su parecido con la ley de potencia en varios ordenes de magnitud, pero no hay un modelo en redes que prediga una la log-norm en la distribución, además en muchos casos la distribución de grado no es sólo una ley de potencia sino una mezcla con otras distribuciones en los grados cercanos a cero.

Una pregunta importante es qué tanta relación hay entre las citas locales y las globales. A continuación se aborda esta pregunta.

4.2.3. Citas locales vs globales

Se analizó si hay correlación entre las citas locales y globales. El coeficiente de correlación (ρ) entre citas globales y locales para física es de 0.6221, para matemáticas es de 0.2664 y para química es de 0.4642. Estos resultados son similares a la proporción N_G/N de cada área mostrado en la subsección 4.2.1, por supuesto no en números absolutos pero sí en orden, ya que física es el más alto, seguido de química y al final matemáticas.

En la figura 4.5 se muestra la comparación entre citas globales y locales en física, en la parte A se observa que no hay una tendencia clara entre los datos, en la parte B de la figura se muesta que hay artículos que no tienen ninguna cita en la red local pero que a nivel global tienen más de 100. El artículo con la mayor cantidad de citas locales (283) tiene mucho menos citas globales que el segundo artículo con más citas locales (199), el primero tiene apenas el 10.35 % de las citas globales del segundo.

Un caso que sobresale es el décimo tercer artículo con más citas globales (592), ya que sólo tiene tres citas locales, sin embargo al buscar el artículo en la WoS sí tiene más citas en M (14), pero las otras 11 citas son del área de astronomía y astrofísica.

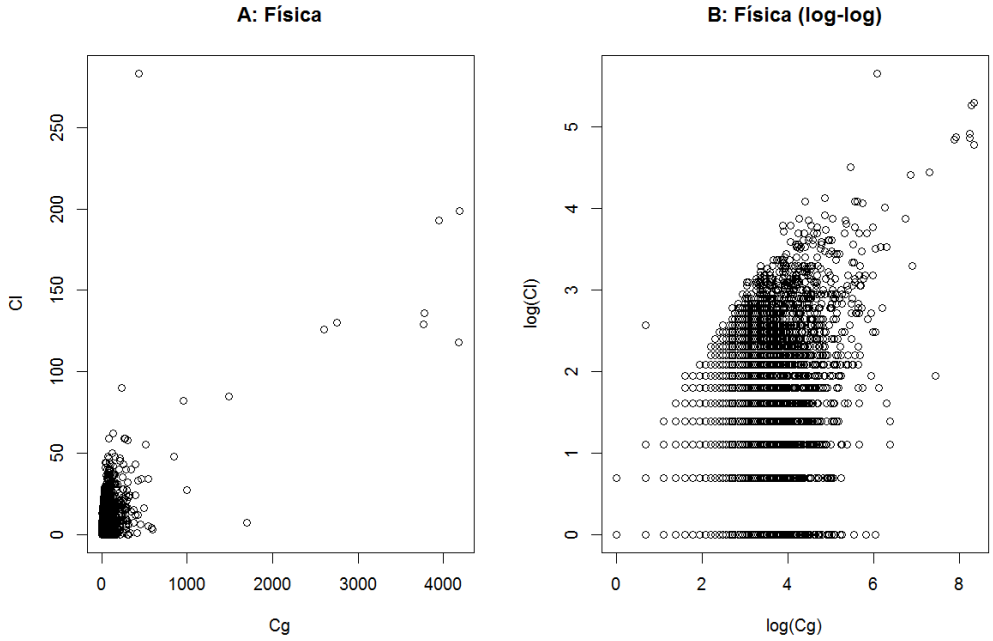


Figura 4.5: Comparación entre citas globales y citas locales en física. En A están los datos sin escalar y en B los datos en escala log-log.

En la figura 4.6 se muestra la comparación entre citas globales y locales en matemáticas, de manera similar a física en la parte A se observa que no hay una tendencia clara entre los datos, en la parte B de la figura también aparecen valores altos para citas globales en artículos sin citas locales, en este caso es más marcado que en física, por ejemplo, el artículo más citado tiene 685 citas globales con sólo 3 citas locales.

Indagando un poco sobre el artículo con más citas globales, fue publicado en Mathematics of Computation en 1980, esto hace suponer que podría tener más citas en el área de cómputo, para corroborarlo se hizo una búsqueda en WoS y sólo tiene 7 citas en M, entonces se indagó sobre el autor del artículo (Jorge Nocedal) y se encontró que sólo estuvo laborando en México durante tres años, se doctoró en 1978

fuera de México, regresó en ese mismo año y salió a laborar fuera del país en 1981.

El segundo caso con más citas globales (673) sólo tiene cuatro citas locales, fue publicado en COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING en 2002, haciendo una búsqueda similar al caso anterior, tiene más citas locales en U_m (52) pero en áreas más bien de cómputo.

El tercer artículo con más citas globales (499) fue publicado en INTERNATIONAL STATISTICAL REVIEW en 1987 y sólo tiene una cita local, en este caso es claro que el artículo es de un área considerada de matemáticas, sin embargo el autor (Carlos Jarque) del artículo tiene una carrera en el ámbito de la administración púbica en México más que en la academia, lo cual hace plantearse cuestionamientos importantes sobre la visibilidad y repercusión del trabajo científico, intuitivamente se pensaría que una figura pública con artículos importantes en revistas internacionales sería muy visible en el ámbito nacional y como consecuencia se esperarían más citas locales, sin embargo en este caso no es así.

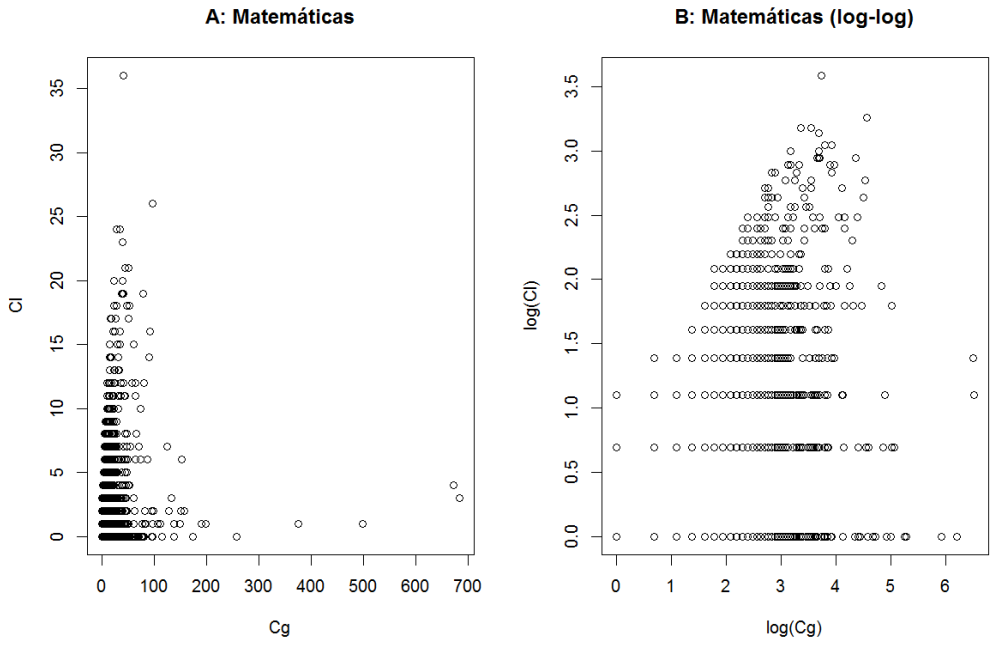


Figura 4.6: Comparación entre citas globales y citas locales en matemáticas. En A están los datos sin escalar y en B los datos en escala log-log.

En la figura 4.7 se muestra la comparación entre citas globales y locales en química, de manera similar a física y matemáticas en la parte A se observa que no hay una tendencia clara entre los datos, en la parte B de la figura también aparecen valores altos para citas globales en artículos sin citas locales.

Un caso interesante es el quinto artículo con más citas globales (427) ya que no tiene ninguna cita local, el artículo fue publicado en ADVANCED MATERIALS en

2001, por el título de la revista se pensó que tal vez tenía más citas en otras áreas en México, sin embargo al realizar una búsqueda en WoS sólo tiene cuatro citas en M.

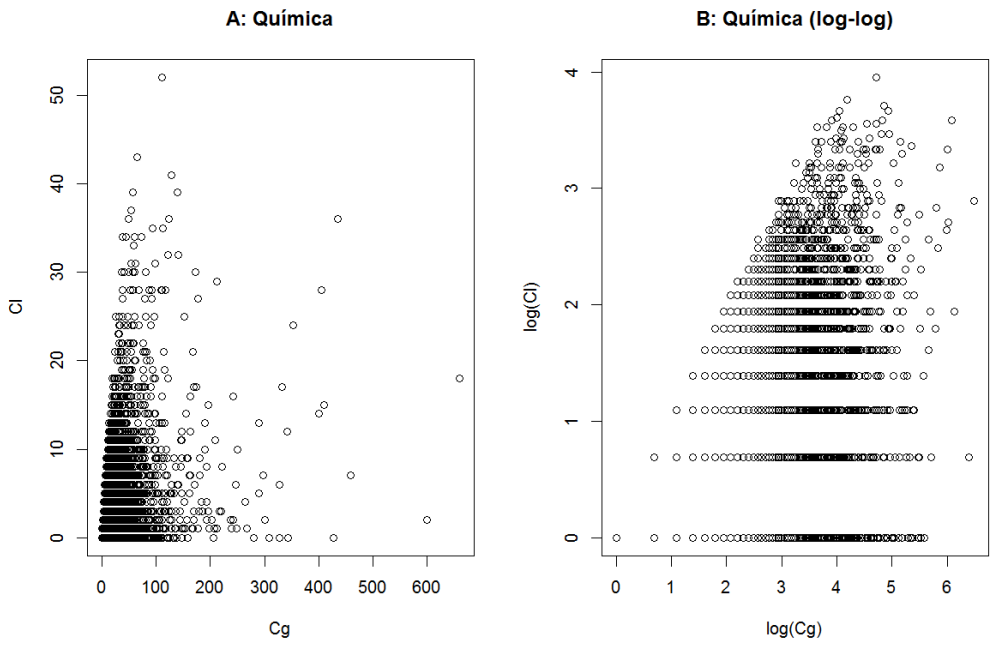


Figura 4.7: Comparación entre citas globales y citas locales en química. En A están los datos sin escalar y en B los datos en escala log-log.

Los resultados del análisis de correlación y de las gráficas correspondientes indican que un número alto de citas globales no garantiza un número alto en citas locales y viceversa.

En la figura 4.8 se observa la razón de citas locales (CL) sobre las globales (CG) en el tiempo para las tres áreas, los datos no son acumulados, corresponden a intervalos de tiempo de 5 años, desde 1970 hasta 2015. Es claro que de 1990 al 2000 hay una alza en la razón de citas, y de 2000 a 2015 una caída completa en el caso de química y hasta 2010 para física y matemáticas.

Las variaciones observadas en la figura 4.8 dan muestra de la gran diferencia entre las citas locales y globales de cada red, pero el aumento y la baja en la proporción de citas se dan por diferencias en las tasas de cambio tanto de las citas locales como en las globales, es decir, no sólo se debe a qué se cite más en lo global y menos en lo local o viceversa sino es un proceso combinado, pero que es claro que la razón de cambio disminuye para las citas locales.

Como se ve en la figura 4.9, de 1990 a 1995 hay un aumento notable en la razón de cambio de las citas locales, este se da en las citas globales cinco años después, de 1995 a 2000, a partir de ahí las pendientes en las citas globales son mucho más

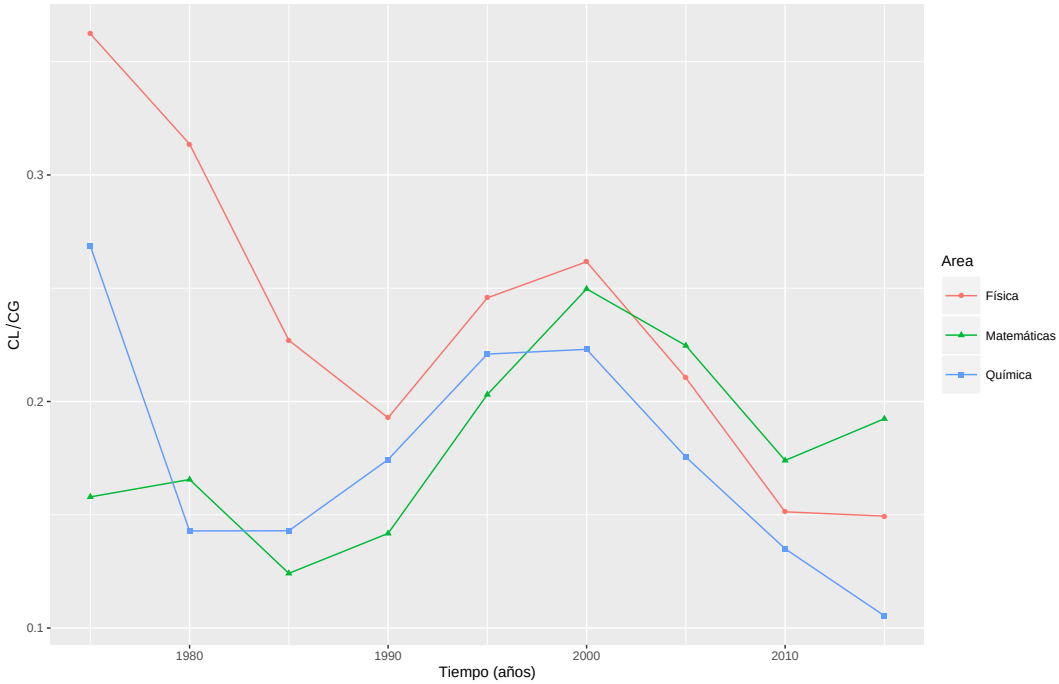


Figura 4.8: Comparación entre la razón de citas locales (CL) sobre las globales (CG) en el tiempo. Los datos no son acumulados, corresponden a intervalos de tiempo de 5 años, desde 1970 hasta 2015.

pronunciadas hasta 2010 para las tres áreas, en el caso de química continua hasta 2015.

De 2010 a 2015 la tasa de variación es mayor en las citas locales que en las globales para matemáticas, pero la tasa de cambio en las locales sigue cayendo desde periodos anteriores. En ese sentido el cambió en la gráfica no se debe a qué se tuvieron más citas en lo local sino que se tuvieron menos en lo global.

En física se mantiene la razón de citación, pero aquí sí se debe a que aumenta la tasa de cambio en las locales y disminuye la tasa en las globales, pero apenas para evitar que la proporción siga cayendo como se observaba en la figura 4.8.

Las diferencias entre las citas locales y globales permiten identificar cambios en el comportamiento de las áreas, y por tanto de los científicos, ya que la disminución en la tasa de citas locales implica al menos una de tres cosas: la búsqueda de visibilidad global enmarcada en la publicación de revistas internacionales, la colaboración internacional como es el caso de la física de altas energías, el cambio en las política científica.

Al respecto de la política científica, cabe señalar que en 1984 se creó el Sistema Nacional de Investigadores (SNI) y en ese momento se comenzaron a contar las citas otorgadas a los artículos de un autor, como una de los factores de evaluación para su ingreso al SNI; esto coincide con las similitudes observadas en la figura 4.8

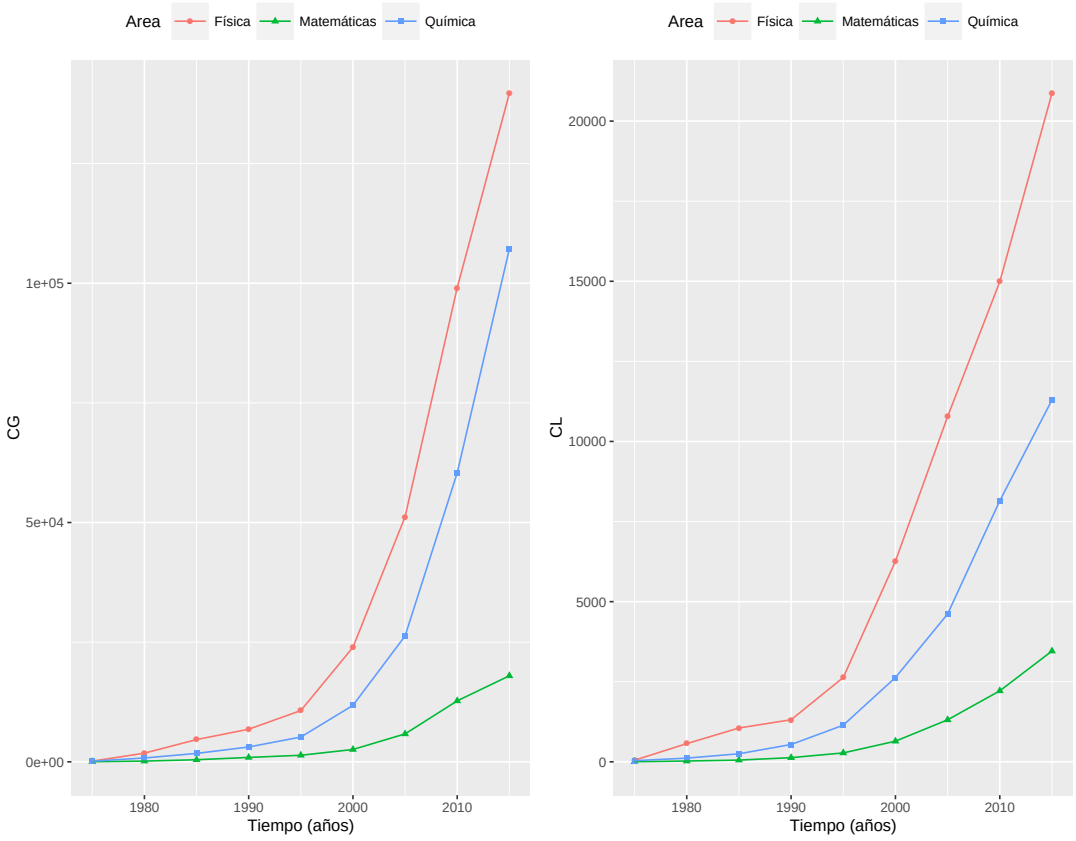


Figura 4.9: Crecimiento en el tiempo de citas globales (CG, izquierda) y citas locales (CL, derecha). Los datos no son acumulados, corresponden a intervalos de tiempo de 5 años, desde 1970 hasta 2015.

cuando sube la fracción de citas internas vs externas. Sin embargo, con el tiempo se decidió diferenciar entre citas tipo A y citas tipo B, donde las de tipo A son aquellas que se otorgan por artículos donde no participa el autor del artículo citado ni algún coautor, las citas tipo A son las que se toman en cuenta actualmente. Este cambio podría ser la causa del decremento en la fracción de citas internas vs externas de los últimos 10 o 15 años como se ve en la figura 4.8.

4.2.4. Influencia en las redes mexicanas

A partir de los resultados del capítulo 3 se calculó la influencia en las redes mexicanas.

En las figuras 4.10, 4.11 y 4.12 se muestran los diagramas de dispersión entre la influencia y las citas normalizadas (Cn) para las componentes gigantes de las redes de M_f, M_m y M_q. La componente gigante se denota por GcF para M_f, GcM para M_f y GcQ para M_f.

De acuerdo con la figura 4.10 es claro que la influencia clasifica de manera muy

distinta artículos con muchas citas, por ejemplo los tres artículos con más citas, en el mismo orden de magnitud, tienen distintos ordenes de magnitud en la influencia. La correlación entre la influencia y las citas es de $\rho = 0.78$. Cabe señalar que tanto la figura 4.10 como la correlación corresponden a los nodos con al menos una cita, en el caso de física son 11,860.

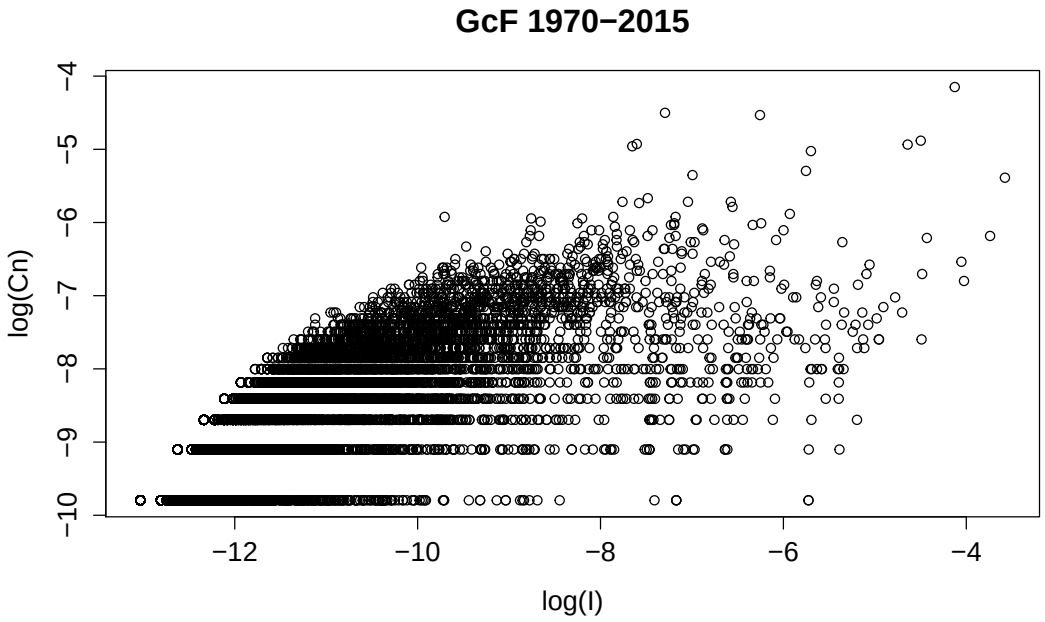


Figura 4.10: Comparación de la influencia con las citas en GcF en escala log-log.

En el caso de la influencia en matemáticas (figura 4.11) también se observa que hay artículos con el mismo orden de magnitud en las citas pero distintos ordenes de magnitud en la influencia. En este caso la correlación es de $\rho = 0.82$. Los nodos con al menos una cita son 178.

En el caso de la influencia en química (figura 4.12) es mucho más marcada la diferencia entre los artículos con la mayor cantidad de citas y la influencia, incluso la correlación es más baja ($\rho = 0.53$). En este caso los nodos con al menos una cita son 6,385.

A partir de los resultados se pueden diferenciar citas con el mismo orden de magnitud por medio de la influencia, aunque en este contexto se debe ser muy cuidadoso con las interpretaciones, pero en principio se tiene una herramienta a partir de la cual indagar a qué se deben las diferencias y cuáles son las implicaciones. En otros contextos como la web, la diferencia entre un orden de magnitud y otro puede ser fundamental para optimizar la búsqueda de información.

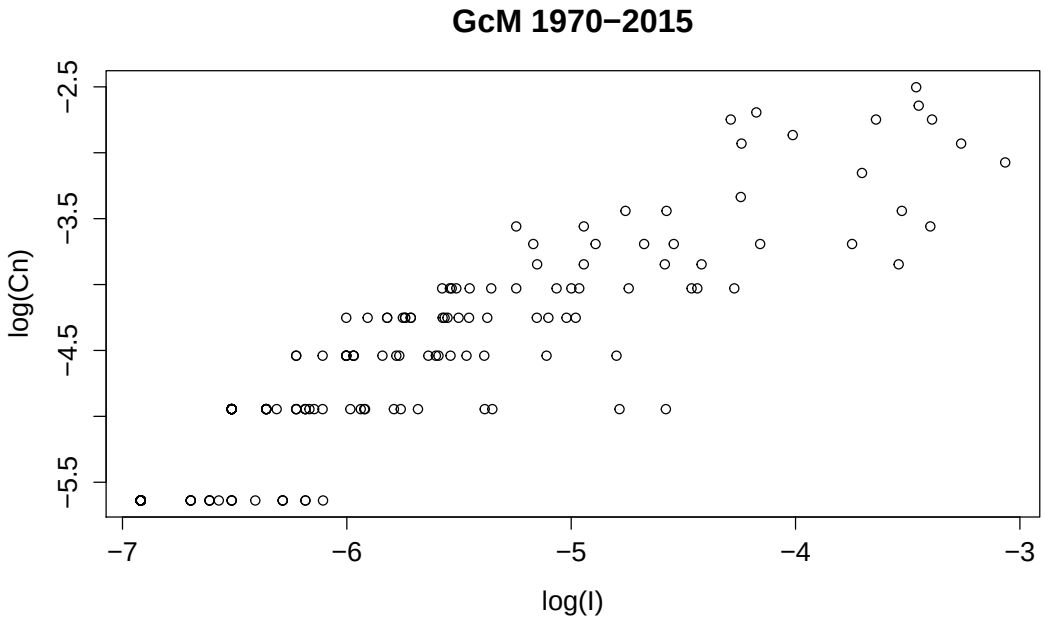


Figura 4.11: Comparación de la influencia con las citas en GcM en escala log-log.

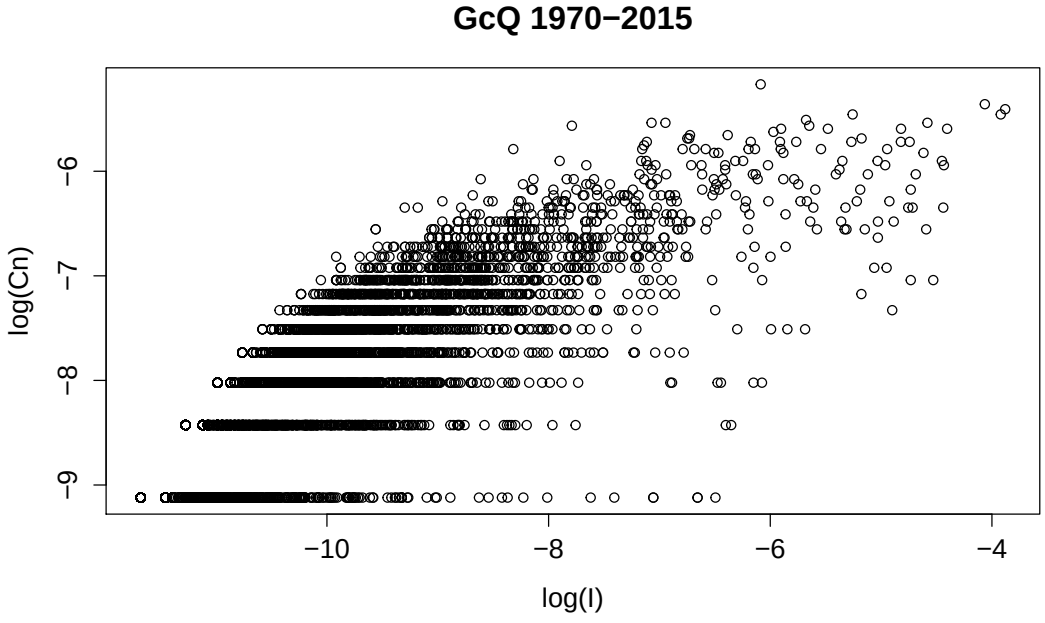


Figura 4.12: Comparación de la influencia con las citas en GcQ en escala log-log.

Capítulo 5

El impacto de la ciencia en la sociedad: el caso Molina

Hasta ahora se han mostrado casos globales (erratas y redes APS) y casos locales (redes mexicanas), de indicadores y modelos en cienciometría; en este capítulo se revisa la trayectoria del científico mexicano Mario Molina, ganador del Premio Nobel de Química, como un caso específico con implicaciones socio-medio ambientales evidentes y relevantes.

En esa dirección, se hicieron las redes de coautoría del científico con base en los artículos publicados desde 1965 hasta 2015, donde se encuentran cambios sobresalientes en dos o tres etapas de su carrera (incluyendo la obtención del Nobel). Luego se da una caracterización de la interacción del científico fuera del ámbito académico con un problema ambiental (el daño de la capa de ozono) estrechamente relacionado con sus investigaciones que a la postre le valieron la distinción del Nobel.

La parte de la caracterización está basada en el artículo «Mario Molina y la saga del ozono: ejemplo de vinculación ciencia-sociedad» [70].

5.1. Datos y Métodos

5.1.1. Datos

Los datos para las redes de coautoría fueron obtenidos de la WoS, se hizo una búsqueda por autor en todo el tiempo. El artículo más viejo data de 1965 y los más recientes de 2016, pero para tener años completos se cortó a 2015. Después de una búsqueda y los filtros necesarios se obtuvieron 206 artículos para el periodo 1965-2015.

5.1.2. Métodos

Para realizar las redes de coautoría se usó Sci2 [67]. Se hicieron redes en periodos de 5 años (acumulados) desde 1965 hasta 2015. En esta parte no se analizan las redes tan a fondo, más bien sirven como elementos de los cambios en la carrera del

científico desde el punto de vista de sus interacciones académicas.

Para analizar la interacción fuera del ámbito académico-científico, se parte de [70] donde se presenta un análisis en paralelo entre la carrera de Molina y el artículo de 1974 [71] que alertó sobre un problema ambiental, la pérdida de la capa del ozono. Se revisa cómo se produjo esta investigación y algunas de sus consecuencias en otros sectores de la sociedad.

Partiendo de [70] y del modelo del rosetón o rosácea de Latour [72] se caracteriza la relación fuera del ámbito académico-científico.

5.2. Redes de coautoría de Molina

Mario Molina estudió Ingeniería Química en la UNAM, al concluir su la licenciatura en 1965 publicó su primer artículo. En 1972 se doctoró por la Universidad de California en Berkeley y en 1974 inicia su colaboración con Rowland, con quién a la postre sería coganador del Premio Nobel de Química en 1995.

En la figura 5.1 se pueden ver varios aspectos interesantes a lo largo de la carrera de Molina. En principio es clara la influencia del Premio Nobel, tanto en la cantidad de publicaciones como en los coautores, aunque en este último rubro es mucho más marcada.

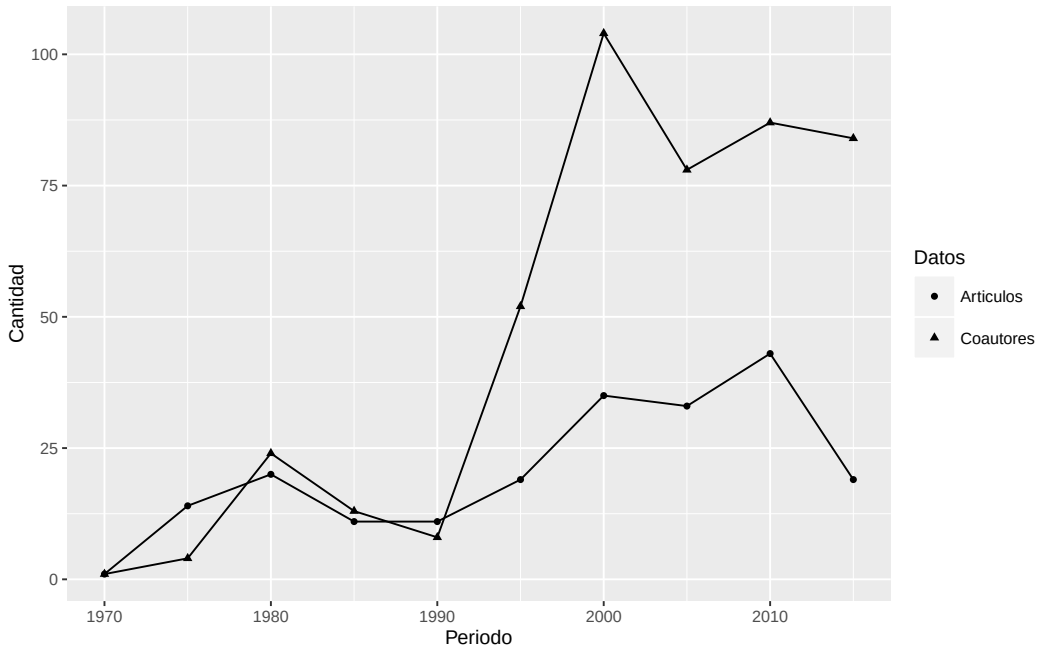


Figura 5.1: Artículos y coautores en periodos de tiempo de 5 años. Los datos no son acumulados.

Otro punto a destacar de la figura 5.1 es que hasta 1990 la cantidad de artículos

era muy similar a la cantidad de coautores en cada periodo. Además también se observa una caída entre 1980 y 1990 la cual seguramente se debe a que de 1982 a 1989 no estuvo en la academia sino en el Jet Propulsion Laboratory. También es interesante que de 2010 a 2015 su producción cae, pero sus coautorías no, mantiene casi la misma cantidad, probablemente una consecuencia del reconocimiento adquirido.

Para el periodo 2005-2010 el alza en la productividad está ligada a la creación del centro Mario Molina [73] ya que en 2006 Molina publica 11 artículos relacionados con la calidad del aire en la ciudad de México.

En la figura 5.2 se muestran las citas por periodos de 5 años desde 1975 hasta 2015. Nuevamente se observa el efecto de haber estado fuera de la academia, pero una observación muy importantes es que las citas no parecen aumentar drásticamente con la obtención del premio Nobel, como se ve en la misma gráfica pero en la subfigura, donde las citas están por cada año, el aumento en las citas venía dándose de manera independiente desde unos años antes. Aunque a partir del año 2005 sí hay un aumento notable en la cantidad de citas, muy similar al aumento de las publicaciones, aunque el aumento en publicaciones no es del doble y el aumento de citas casi lo es. Es prudente notar que el número total de citas recibidas hasta 2015 es de 12,859.

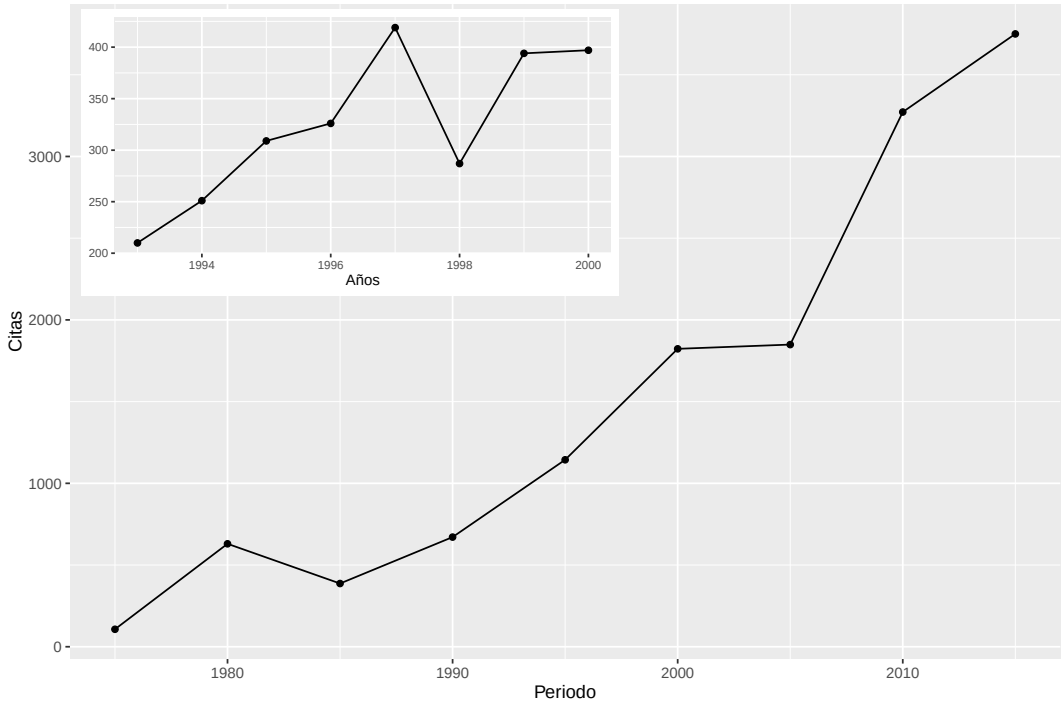
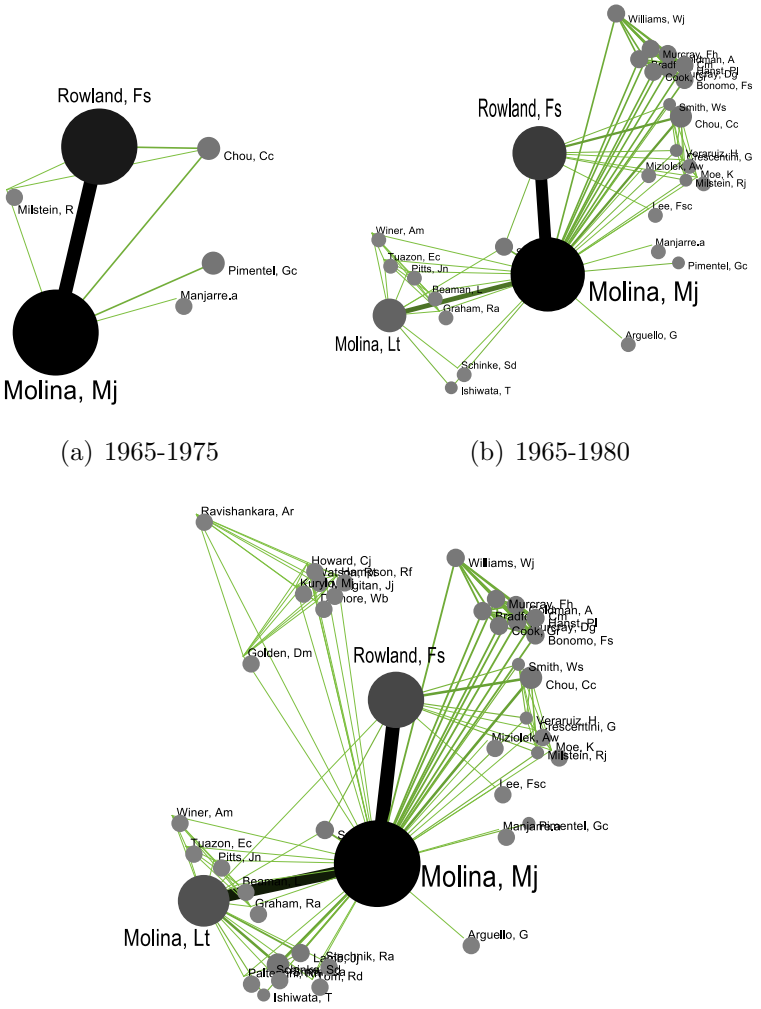


Figura 5.2: Citas recibidas por los artículos de Molina en periodos de 5 años.

A partir de este vistazo inicial surgen preguntas sobre cuáles son los coautores más importantes de Molina, en qué artículos colaboró para el incremento de los coautores. A continuación se presentan las redes de coautoría en intervalos de tiempo acumulado

de 5 años, esto permite observar el incremento de coautores en la red de un periodo a otro periodo.



(a) 1965-1975 (b) 1965-1980

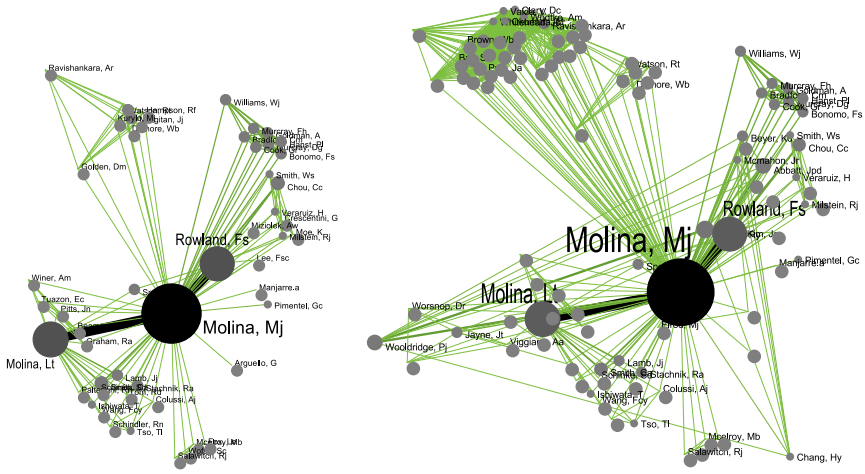
(c) 1965-1985

Figura 5.3: Redes de coautoría de Mario Molina de 1965 a 1985 en periodos de 5 años, salvo la primera por la poca cantidad de nodos entre 65 y 70.

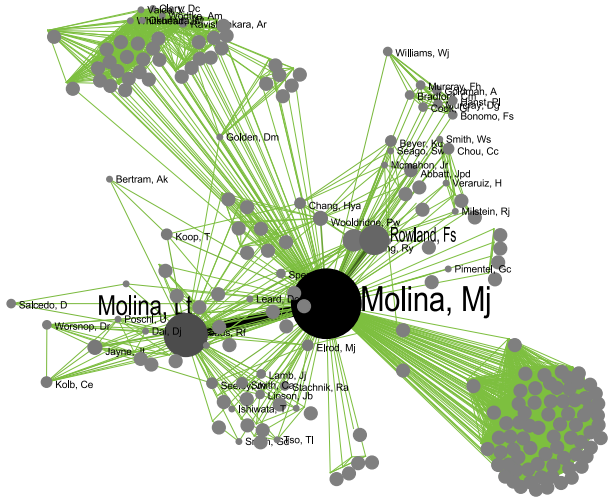
En la figura 5.3 se muestra la evolución para tres periodos de tiempo. En todas las redes cada nodo representa un autor, cada línea entre nodos al menos una colaboración, el grosor de la línea representa la frecuencia de colaboración entre los autores (en cuantos artículos son coautores), el tamaño del nodo representa la cantidad de artículos en la que cada autor aparece como coautor. Evidentemente en todas las redes Molina es el nodo más grande (Molina, Mj), sin embargo el resto de los nodos van cambiando de tamaño así como el grosor del enlace entre ellos.

En las redes (a) y (b) de la figura 5.3 Rowland es el coautor principal de Molina, además del segundo autor con más publicaciones, lo cual se muestra por medio del tamaño del nodo y del grosor de la línea entre ellos. Pero en la red (c) Molina Lt

(colaboradora y esposa) tiene el mismo peso que Rowland.



(a) 1965-1990 (b) 1965-1995



(c) 1965-2000

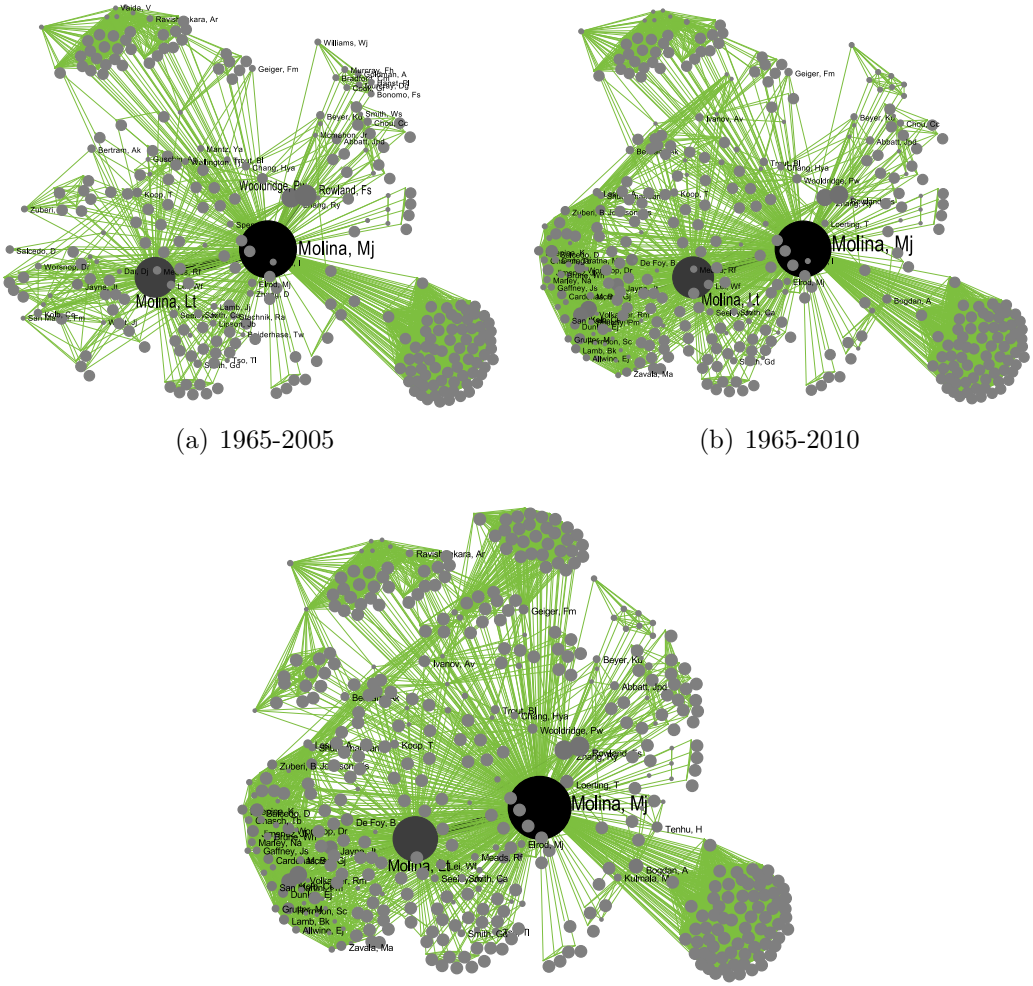
Figura 5.4: Redes de coautoría de Mario Molina de 1990 a 2000 en periodos de 5 años acumulados desde 1965.

En la figura 5.4 se muestra la evolución para los siguientes tres periodos de tiempo. Es claro que de 1985 a 1990 no hay casi ningún cambio en la red, salvo unas cuantas incorporaciones ya que como se mostró en la figura 5.1 entre 1985 y 1990 disminuyó la cantidad de coautores, respecto al periodo anterior. Sin embargo para 1995 es evidente el aumento de los coautores en el grupo de arriba a la izquierda de Molina, en 1990 el autor Ravishankara, Ar tenía sólo un artículo con 8 coautores, incluyendo a Molina, sin embargo para 1995 tiene tres artículos y sus coautores suben a 41, que son precisamente el grupo señalado.

En la red (c) de la figura 5.4 hay un conjunto grande de coautores nuevos en la parte inferior-derecha de la red. Este conjunto se genera de un artículo-carta

donde un grupo de científicos reconocidos piden se reconsidere la postura sobre los experimentos con células madre, que habían sido vetados en Estados Unidos. Molina firmó la carta y este uno de los primeros ejemplos donde se ve la actividad de Molina fuera de su ámbito de investigación (aunque actividades similares se remontan al problema con el agujero en la capa del ozono), ya que la carta no era de un tema de su área de estudio, sin embargo al ser un ganador del premio Nobel dio un respaldo importante a la petición, que a la postre influyó en que el veto fuera eliminado.

En la figura 5.4 también sobresale que para el año 2000 Molina Lt ya ha superado a Rowland como la coautora con más trabajos en conjunto con Molina.



(a) 1965-2005

(b) 1965-2010

(c) 1965-2015

Figura 5.5: Redes de coautoría de Mario Molina de 2005 a 2015 en periodos de 5 años acumulados desde 1965.

En la figura 5.5 se muestran las últimas tres redes. En (b) se observa la incorporación de un conjunto de coautores a la izquierda inferior de Molina, se aprecia como la red se va volviendo más densa cada vez y finalmente para el 2015 (c) surge otro

conjunto en la parte central superior.

Para indagar más sobre la interacción de Mario Molina con el ámbito extra académico, en la siguiente sección se presenta una revisión en paralelo entre la carrera de Molina y el artículo de 1974 [71] que alertó sobre la pérdida de la capa del ozono y algunas implicaciones sociales.

5.3. Molina fuera del ámbito académico

El artículo publicado en 1974 [71] por Rowland y Molina alertó sobre la pérdida de la capa del ozono y tuvo repercusiones a niveles inesperados fuera del ámbito académico, desde participaciones de los autores en el congreso de los estados unidos hasta apariciones en revistas de espectáculos, pasando por fuertes debates con empresas del sector privado y culminando con la obtención del premio Nobel y la recuperación de la capa del ozono.

Una de las primeras acciones importantes en la interacción fuera del ámbito académico de Molina sucedió cuando en conjunto con Rowland llamaron a ruedas de prensa y dieron entrevistas a muchos medios, algunos alejados de las revistas científicas (un ejemplo es [74]). De esa manera lograron atraer la atención de los oficinas gubernamentales encargadas de regular productos e industrias [75].

Uno de los resultados de las ruedas de prensa fue que el Congreso de Estados Unidos convocó audiencias sobre el tema –en las que participaron Molina y Rowland, entre otros– y tomó dos acciones concretas: creó el comité sobre la Modificación Inadvertida de la Estratosfera (IMOS) para darle seguimiento a estos problemas potenciales y solicitó a la Academia Nacional de Ciencias (NAS) que estudiara el problema. Ambos concluyeron que el argumento parecía ser válido y que era necesario hacer más estudios y eventualmente regular el uso de estos compuestos [76].

En la misma década de los años 70's se publicaron 324 historias acerca de la capa de ozono, la mayoría de ellas en medios de comunicación masiva [77]. Muchas funcionaron como medio educativo, enfatizando que al usar productos ordinarios tales como los aerosoles, entonces muy populares.

El primer triunfo visible del activismo para proteger la capa del ozono fue la regulación de aerosoles en 1976, luego en 1978 se prohíbe el uso de CFCs (clorofluorocarbonos) en aerosoles en Canadá, Estados Unidos, Noruega y Suecia.

En los años que van desde la publicación del artículo de 1974 hasta 1985, Molina publicó un par de docenas de artículos relacionados con el problema del ozono. En particular el de Rowland y Molina [78] se refuta un anuncio pagado por Du Pont, el mayor productor a nivel mundial de CFCs en ese momento, aparecido en varios periódicos y revistas científicas, incluida Science. En él la compañía Du Pont critica

el artículo de 1974 de Rowland y Molina.

En 1985 se publica un artículo en Nature [79] donde reportaron pérdidas del 30 % de la capa del ozono sobre la Antártida durante el mes de octubre con respecto a los valores medidos en octubre a mediados de los 1950s. A esto se le llamo el agujero del ozono. Farman et al. [79] sugirieron que el culpable del agujero eran los CFCs y que el clima agudizaba el problema.

En marzo de 1986 una expedición llamada NOZE realizó mediciones y concluyó que sospechaban que procesos químicos causaban el agujero. Lo vago del mensaje no fue bien acogido por la comunidad científica, ni por la industria, pero fue suficiente para reanudar discusiones internacionales en búsqueda de un tratado que incluyera controles. En agosto de 1987 se realizó una segunda expedición a la Antártida, llamada AAOE, organizada por la NASA y otras agencias en la que participaron más de un centenar de científicos. Poco antes de que los resultados de AAOE estuvieran disponibles, representantes gubernamentales se reunieron para la negociación final de un tratado del ozono. La reunión fue en septiembre de 1987 en Montreal, donde 43 países firmaron el Protocolo de Montreal [80], que requería una reducción del 50 % en la producción mundial de CFC para el año 2000 y, más importante, incluía el acuerdo de modificar el protocolo conforme se fueran conociendo nuevos resultados científicos.

Como se mencionó en la sección anterior la obtención del Premio Nobel dio otra dimensión a la figura de Mario Molina, le permitió adentrarse en otros temas ambientales y relacionarse a otro nivel con la ciencia y sobre todo con la política tanto en Estados Unidos, como en México y en Latinoamérica. En 1997 publica un primer artículo acerca de las condiciones atmosféricas de la Ciudad de México; años más tarde le seguirían otros. En 2004 se funda el Centro Mario Molina (CMM) en la Ciudad de México, con una agenda flexible para hacer frente a problemas ambientales: “Su propósito es encontrar soluciones prácticas, realistas y de fondo a los problemas relacionados con la protección del medio ambiente, el uso de la energía y la prevención del cambio climático, a fin de fomentar el desarrollo sustentable” [73].

Desde su fundación ha sido muy activo, habiendo firmado ya cuatro convenios con el Conacyt (la agencia financiadora de la ciencia en México). El más reciente, firmado el 15 de mayo de 2013 incluye 8 líneas de investigación y 17 proyectos en el área ambiental.

En paralelo con la creación del CCM en México, se creó en el mismo año, 2004, otro Centro Mario Molina en Chile (http://www.cmmolina.cl/). Este centro tiene una misión similar al de México y de la misma manera participa activamente en la investigación de problemas ambientales y en la definición de la política científica para tratarlos.

En 2014 la Asociación Mundial Meteorológica [81] publicó un reporte donde se indica que la capa del ozono que protege el planeta está en vías de recuperarse en las próximas décadas. Este logro, debido a la aplicación del Protocolo de Montreal, es visto como una historia de éxito de la vinculación entre ciencia y tecnología [82].

5.4. Molina y el modelo de Latour

A partir de la breve revisión de la sección anterior son evidentes varios puntos de encuentro entre el científico y el mundo extra-académico. Un acercamiento desde la historia de la ciencia para analizar la carrera de Molina, con una visión integral, es un modelo propuesto por Latour [72] que pretende dar cuenta de la evolución de cómo un área de investigación se gesta y logra convertirse en un campo alto impacto dentro y fuera del ámbito académico. A continuación se hace una correspondencia entre los elementos del modelo y la carrera de Molina.

Latour propone un modelo basado en 5 elementos (círculos o cónicas) que interactúan, como se muestra en la figura 5.6. El círculo azul al centro simboliza al científico, a la cónica 1 le llama la *movilización del mundo*, a la cónica 2 *autonomía*, la cónica 3 representa las *alianzas*, a la cónica 4 le llama las *representaciones* y a la cónica 5 los *vínculos y elementos vinculantes*.

La *movilización del mundo* trata sobre dotar al mundo de movilidad, de encauzarlo hacia los puntos controvertidos, en el caso de Molina esto se ve reflejado en el artículo de 1974 [71] y su activismo en diversos medios, que ya se han mencionado.

La *autonomía* está relacionada con lograr que un tema de investigación o un área completa se vuelva de interés para un grupo de investigadores que se especializan en ello y se forman nuevos investigadores, esto se refleja en Molina por medio de sus coautores que fueron creciendo en el tiempo y de los proyectos en el centro Mario Molina.

Las *alianzas*, por su parte, son lazos que se crean para obtener los recursos necesarios para realizar la investigación, y de acuerdo con [72] las más importantes son: el estado, el ejército, la industria y el sistema de enseñanza. En este sentido, inicialmente Molina no tuvo una buena alianza con la industria pero sí con el estado, sobre todo con el congreso y la NASA, que al mismo tiempo podría considerarse parte del ejército; el sistema de enseñanza se involucró relativamente rápido, una vez que el congreso de Estados Unidos tomó algunas medidas como las que se mencionan a continuación.

En 1972 el gobierno de Estados Unidos aprobó el proyecto del transbordador espacial propuesto por la NASA. Algunos estudios sugirieron en 1973 que el transbordador depositaría cloro en la atmósfera [83], y era claro que la pregunta de la capa

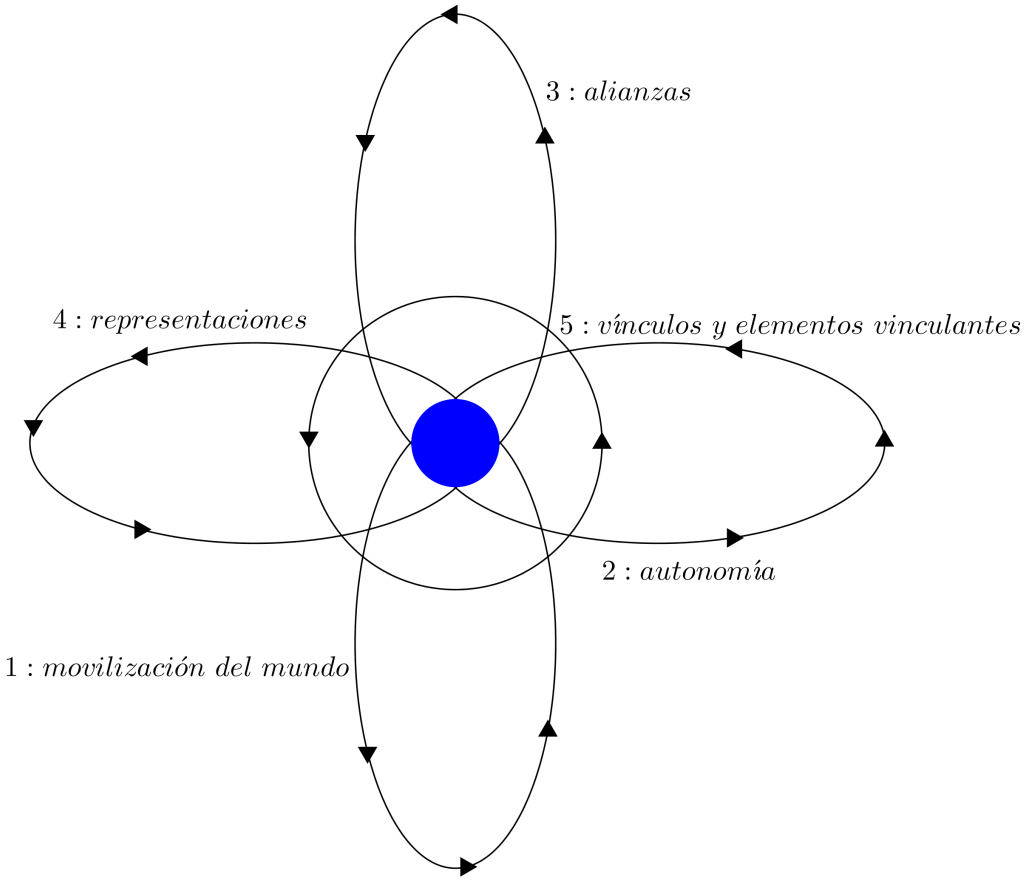


Figura 5.6: Representación del modelo de Latour.

del ozono surgiría. Para la NASA el programa del transbordador espacial era crucial. El revuelo creado por el artículo de Molina y Rowland forzó al gobierno a considerar un nuevo programa de investigación sobre la estratósfera y la NASA aprovechó la coyuntura para convertirse en líder de esta área. En 1975 el Congreso comisionó a la NASA para que se hiciera cargo de la investigación necesaria. Como respuesta, la NASA creó la Oficina de Investigación de la Atmósfera Alta (UARO), encargada de estudiar y monitorear la estratósfera. Dado que la presión política por el efecto de los CFCs no abatía, en 1977 el Congreso ordenó a la NASA a emitir reportes bianuales acerca de la situación de la capa del ozono.

Las *representaciones* son aquellas que hace el público (en general) de los resultados de investigación, en el caso del daño a la capa del ozono ya se mencionó el gran impacto que tuvieron las intervenciones de Molina y Rowland en medios de comunicación de diversa índole, en ese sentido el público acogió bien las propuestas de la ciencia y esta es una de las grandes labores de Molina.

Finalmente el círculo 5, de los *vínculos y elementos vinculantes* habla de cómo mantener movilizados simultáneamente los cuatro anteriores ya que de esa manera la

unión entre los cuatro perdurará más allá del científico en cuestión. Nuevamente, Molina da muestra de estos vínculos cuando mantiene los coautores con menos artículos o cuando la cantidad de citas crece independientemente de obtener el premio Nobel, con la creación del centro Mario Molina, entre otras más que se han mencionado.

5.5. Comentarios finales sobre el caso Molina

Entre el problema del ozono y la carrera científica de Mario Molina, hay hechos que claramente fueron fundamentales y que en general están presentes en el desarrollo histórico de las ciencia, como se mostró usando el modelo de Latour y que también se podrían resumir en: infraestructura, liderazgo y flexibilidad.

Todos los actores en las alianzas son necesarios para el desarrollo de la ciencia, aunque muchas veces no sean evidentes, pero al menos el gobierno, industria, medios de comunicación y academia sí son notables en el caso de Molina.

En Latinoamérica, las alianzas pueden mejorarse mucho, pero también las representaciones. Es importante apoyar la creación de nuevos centros de investigación y enseñanza, incluyendo zonas geográficas donde actualmente no existen. Es importante dotarlos de equipo de punta, pero también de jóvenes brillantes y de temas interesantes estudiados con el más alto estándar científico. La industria podría aprovechar a aquellos estudiantes bien preparados que decidan dejar el mundo académico, para innovar en sus procesos y productos como respuesta a los avances científicos; de la misma manera como tuvo que reaccionar la industria que usaba CFC para crear nuevas tecnologías que sustituyeran su uso.

Conclusiones

Los resultados de la investigación se han incluido en cada capítulo: fue posible generar un indicador novedoso a partir de los errores, que tiene utilidad a nivel mundial dentro de la bibliometría o cienciometría. Se encontró que las erratas pueden ser útiles para resaltar las diferencias en la práctica de la ciencia, a través de las áreas del conocimiento. También que las erratas pueden dar información adicional pertinente para evaluar el desempeño de las revistas. Y se propone mejorar las herramientas para clasificar y encontrar erratas, en las bases de datos de las propias revistas así como en Scopus y Web of Science. Hay varias preguntas al respecto de las erratas: porqué hay tan bajos niveles de erratas en áreas con gran cantidad de revistas con memorias de conferencias, cómo identificar quién o quienes encuentran los errores más comúnmente, si los autores, editores o los lectores; a qué se deben las diferencias entre áreas del conocimiento.

Ahondando en la última pregunta, una de las diferencias se puede trazar desde la concepción misma de las áreas de estudio; las más teóricas y menos experimentales en efecto tienen un porcentaje menor de erratas, como serían matemáticas y cómputación, así como una parte de la física, de la economía y de la ingeniería. Sin embargo, el porcentaje de erratas en física es tres veces mayor que en ingeniería, como ya se mencionó; la explicación puede venir de los artículos de conferencias que en ingeniería son más de la mitad del total de sus publicaciones. Esto muestra que hay características que se pueden identificar fácilmente, pero otras que requieren de trabajo en esta área para poder encontrarlas; y de esa manera comprender mejor las diferencias.

La medición de la influencia de un artículo de investigación no es trivial y la propuesta que se planteó está bien sustentada teóricamente; se mostró su potencial analizando redes de citación internacionales y locales. Del análisis con otras medidas se clarifican varias interrogantes sobre las correlaciones entre ellas. Por supuesto, no se pretende dar una receta pero sí un acercamiento distinto hacia el problema de medir la influencia o el impacto de un artículo en la red de citación. Un trabajo futuro sería tomar casos de estudio de algunos artículos y profundizar en la interpretación de las métricas y qué implicaciones tienen en una clasificación.

Además, respecto a la medida propuesta, los análisis de las correlaciones permiten identificar que no hay mucha diferencia en los resultados de las mediciones, a nivel general. Pero si es necesario diferenciar nodos con el mismo grado, entonces se puede

usar distintas medidas para ello, pero si se usa $I_(S)$ se tendrá una medida donde es claro lo que se está midiendo y a qué se debe la diferenciación; además, está menos sesgada que los Hubs, el Page Rank o la intermediación y permite diferenciar a escalas más pequeñas que la centralidad por vector propio. También es claro que su implementación no es costosa, desde el punto de vista de cómputo, ya que la red de APS es suficientemente grande y los cálculos se hicieron en minutos.

En las redes de citación mexicanas, se mostró que el análisis de éstas con métodos estadísticos, ya establecidos, es útil para identificar cómo se ha dado el desarrollo de la ciencia en las áreas analizadas y las grandes diferencias entre ellas. Además permite observar dónde se tienen terrenos fértiles para desarrollar proyectos novedosos en México, como el caso del proyecto TDA. También es importante considerar las distribuciones de citas para entender las interacciones entre los científicos, desde el punto de vista de la citación. Como se vió, en química sí hay una componente gigante, pero la red no es libre de escala; esto indica que al menos en la red de citación no hay evidencia de artículos arriba del promedio, que impacten la ciencia mexicana; en química en partiular y de manera similar para matemáticas donde ni siquiera hay una componente gigante.

La comparación de las citas locales contra las globales permitió ver cambios abruptos en la razón entre ellas en el tiempo, mostrando la posible influencia de la política científica del CONACYT, como en el caso de las convocatorias del SNI donde se diferencian las citas, dependiendo de si son autocitas o no. Esto lleva a una implicación importante: si la tasa de citación interna disminuye debido a una política científica, entonces esas citas no eran necesarias, es decir eran prescindibles. Esto es, si un autor considera que debe citar su propio artículo por necesidades científicas seguiría teniendo esas autocitas, o seguiría citando a los colegas coautores, no importaría que tales citas no cuenten para el SNI.

En esa dirección, los resultados de la comparación entre las citas podrían indicar que muchas citas que se otorgaban poco tenían que ver con el proceso científico en sí y respondían a una política científica, y cuando tal política cambia esos patrones también lo hacen.

También es importante considerar los casos particulares donde hay una gran diferencia entre las citas locales y globales, ya que una pregunta que surge es: ¿qué es más importante para la ciencia de México? un científico que es muy citado en el extranjero pero en el país no, o uno que es muy citado dentro del país pero fuera de él no, se podría pensar que lo mejor sería tener científicos con trabajos que tuvieran citas internacionales y nacionales, pero es claro que eso no está sucediendo en México, al menos para las tres áreas analizadas. La otra vertiente es: qué se hace para que sucedan ambas.

Una extensión natural de las redes de citación en México, es hacer el análisis para el resto de las áreas del conocimiento.

El caso Mario Molina es una buena práctica de lo que muchas veces se busca con la cienciometría: dar perspectivas novedosas sobre el quehacer científico y su relación con la sociedad fuera del ámbito científico, además de la influencia de los galardones en la interacción de un científico con sus pares y con la vida extra académica. La caracterización inicial en: infraestructura, liderazgo y flexibilidad; también se puede encontrar en modelos de historia de la ciencia, como el que se mostró.

Identificar estos elementos, para que se den resultados como el de Molina, es fundamental para México si se piensa en tratar de cosechar casos de éxito como ese. Siendo mexicano y estudiando hasta la licenciatura en México, si hubieran existido condiciones similares a las que encontró en Estados Unidos, posiblemente él u otros habrían tenido repercusiones de la misma índole, tanto en el ámbito científico, como en el ámbito social fuera de la academia. Pensando en ello, sería pertinente realizar el análisis de cuáles elementos vinculantes han permitido a Molina movilizar el resto de los recursos.

Finalmente, hacer hincapié en que esta tesis es una clara interacción entre diversas disciplinas, con resultados de impacto a corto y mediano plazo, lo cual es uno de los propósitos del programa de doctorado.

Apéndice A

Métodos Generales

En términos metodológicos este trabajo es de corte cuantitativo con implicaciones de orden cualitativo, en general cualquier trabajo sobre el análisis de la ciencia tiene esos dos componentes, ya que siempre incluyen análisis matemáticos que van desde los muy someros hasta los más especializados, pasando por un trabajo arduo en las bases de datos y finalmente dando una interpretación de todo ello [84, 85].

Esencialmente los trabajos relacionados con los mapas o redes científicas se componen de ocho etapas [84, 85]:

1. Recopilación de la información
2. Preprocesamiento
3. Extracción de la red
4. Normalización
5. Creación de la red
6. Análisis
7. Visualización
8. Interpretación

En esta tesis se desarrollaron esas etapas durante el trabajo, aunque no sólo se hizo análisis de redes, el proceso funciona también para el análisis de erratas salvo que las etapas 3, 4 y 5 se cambian por la síntesis de información, caracterización y gráficas de tendencias e histogramas.

Uno de los aportes más importantes está en la etapa del análisis ya que se construyo un modelo y se generó un indicador.

De la segunda a la cuarta etapa esencialmente se trata de un manejo de datos.

A.1. Fuentes de datos

Los datos que se usaron en la investigación provienen de tres fuentes: Scopus, Web of Science (WoS) y páginas web de revistas (American Physical Society, Nature y PLOS ONE).

En el caso de Scopus, WoS, Nature, PLOS ONE y la parte de los datos de APS para analizar los errores todos los datos fueron descargados directamente de Internet.

Para el análisis de las redes de citación internacionales los datos fueron solicitados directamente a la APS.

Para las redes de citación nacionales y las redes de coautoría de Mario Molina los datos se descargaron de la WoS.

Apéndice B

Datos sobre erratas

Por cada una de las 27 áreas (tabla B.1) en que Scopus divide las publicaciones científicas se descargó el total de publicaciones de cada área por año, el total de revistas del área, el total de erratas por año y el total de revistas con erratas.

Tabla B.1: Áreas del conocimiento y sus claves asociadas usadas por la base de datos Scopus.

Área	Clave
Agricultural and Biological Sciences	AGRI
Arts and Humanities	ARTS
Biochemistry, Genetics and Molecular Biology	BIOC
Business, Management and Accounting	BUSI
Chemical Engineering	CENG
Chemistry	CHEM
Computer Science	COMP
Decision Sciences	DECI
Dentistry	DENT
Earth and Planetary Sciences	EART
Economics, Econometrics and Finance	ECON
Energy	ENER
Engineering	ENGI
Environmental Science	ENVI
Health Professions	HEAL
Immunology and Microbiology	IMMU
Materials Science	MATE
Mathematics	MATH
Medicine	MEDI
Multidisciplinary	MULT
Neuroscience	NEUR
Nursing	NURS
Pharmacology, Toxicology and Pharmaceutics	PHAR
Physics and Astronomy	PHYS
Psychology	PSYC
Social Sciences	SOCI
Veterinary	VETE

De la WoS, por cada área, ENGI, MATH, PHYS y MULT, se descargó la misma información que en el caso de Scopus, pero sólo para el periodo 2000-2014.

Además, por cada una de las cuatro áreas anteriores se obtuvo la información de revistas con erratas (correcciones en WoS) y sin erratas.

Los datos de la página web de cada revista para el periodo 2000-2014 se obtuvieron como se indica a continuación. De la página de la American Physical Society Journals se obtuvo la información de las revistas de la APS así como los datos en particular de Physcial Review Letters (PRL) con respecto a los tipo de erratas: *Comment, Reply,* ***Publisher's note, Erratum*** y ***Retraction***.

De manera similar para Nature usando búsqueda avanzada se obtuvo la información correspondiente a: *corrections,* ***erratum, corrigendum, retraction*** y *addendum*.

De la página web de PLOS ONE se decargó la información correspondiente a: *corrections* y *retractions*.

B.1. Datos de Journals en WoS

Se realizaron búsquedas en la WoS por área de investigación (SU) para tres áreas: ENGI, MATH y PHYS. Y para MULT por categoría de web of science (WC). Aunque en WoS no se tiene la misma nomenclatura que en Scopus para las áreas o categorías aquí se considera la nomenclatura de Scopus sólo por practicidad. Para cada área o categoría se seleccionó el periodo 2000-2014 y se descargaron los datos correspondientes al total de publicaciones y de correcciones por año. Las correcciones son el equivalente a erratas en Scopus, en adelante se usa erratas independientemente de si se habla de WoS o Scopus. También se obtuvo por cada área la información de los journals respecto al total de publicaciones en el periodo seleccionado, así como las erratas por año. El resumen de los datos se muestra en la tabla B.2.

Área	Total de publicaciones (T_p)	Erratas (E)	% E/T_p
ENGI	3659652	9612	0.26 %
MATH	786265	4210	0.53 %
MULT	427705	6798	1.28 %
PHYS	2123387	16316	0.76 %

Tabla B.2: Resumen de datos de WoS en el periodo 2000-2014.

Por cada área se obtuvo la información de cuantos journals tienen erratas, en la tabla B.3 se muestra el porcentaje de journals con erratas de cada área.

A continuación se describen características de journals de cada área, en caso de existir se incluye también el factor de impacto (IF) de 2014.

Área	Total de journals (T_j)	Journlas con Erratas (E_j)	% E_j/T_j
ENGI	21114	937	4.43 %
MATH	6627	555	8.37 %
MULT	764	81	10.58 %
PHYS	7228	393	5.43 %

Tabla B.3: Porcentaje de Journals con erratas en cada área.

B.1.1. Erratas en Journals de ENGI

Como se vio, en ENGI hay más del 95 % de journals que no tiene erratas, entonces se indagó en cuáles sí hay y en cuáles no. En la tabla B.4 se muestran los 25 Journals con más publicaciones en ENGI para el periodo 2000-2014, el total de publicaciones y el total de erratas de cada journal en ese periodo.

Como se puede apreciar en la tabla B.4 prácticamente la mitad de los Journals no tienen erratas, sin embargo los 12 sin erratas corresponden a proceedings, algo común en el área. Para tener una perspectiva diferente se redujeron los journals a aquello que no eran proceedings, de los 3659652 de resultados iniciales 1770250 corresponden a PROCEEDINGS PAPER, estos se descartaron y con esta disminución el total de Journals pasó de 21114 a 6293, entonces hay 85.11 % de journals sin erratas ya que el total de erratas sigue siendo el mismo.

De estos 6293 Journals se seleccionaron los 50 journals con más publicaciones, en la tabla B.5 están los primeros 15 con el total de publicaciones, las erratas, el porcentaje de erratas, el factor de impacto (IF) y el cuartil en el que se encuentra el journal de acuerdo al IF (los datos del IF corresponden a 2014). En los casos donde el journal pertenece a más de una categoría se mencionan los cuartiles correspondientes como «CHEMICAL ENGINEERING NEWS» que pertenece a dos categorías, entonces en la columna cuartil dice «Q4 (2)» lo que significa que en las dos categorías está en el cuarto cuartil. Otro caso es «MATERIALS SCIENCE AND ENGINEERING A STRUCTURAL MATERIALS PROPERTIES MICROSTRUCTURE AND PROCESSING» que está en tres categorías, en la columna cuartil dice «Q1 (2) y Q2 (1)» lo que implica que en dos categorías se encuentra en el primer cuartil y en una en el segundo, el caso como «PROFESSIONAL ENGINEERING» que tiene «Q4 (2012)» significa que no hay datos para el IF en 2014 pero que hasta 2012 estaba en el cuarto cuartil.

En la tabla B.5 hay dos journals que no tienen erratas (PROFESSIONAL ENGINEERING y OIL GAS JOURNAL) y se encuentran en Q4, en total en el Top 50 hay 3 journals sin erratas y están en Q4. Además hay 5 journals con menos del 0.1 % de erratas, cuatro se encuentran en Q4 y uno en Q2. En consecuencia para el Top 50 de ENGI sin proceedings la ausencia de erratas o un porcentaje bajo de éstas sirve

Journals	Total	Erra
ADVANCED MATERIALS RESEARCH	65172	N/A
APPLIED MECHANICS AND MATERIALS	58535	N/A
PROCEEDINGS OF THE SOCIETY OF PHOTO OPTICAL INSTRUMENTATION ENGINEERS SPIE	47300	N/A
CHEMICAL ENGINEERING NEWS	42038	227
PROCEEDINGS OF SPIE	34010	N/A
JOURNAL OF ALLOYS AND COMPOUNDS	25581	65
ENVIRONMENTAL SCIENCE TECHNOLOGY	21103	277
INDUSTRIAL ENGINEERING CHEMISTRY RESEARCH	18661	118
MATERIALS SCIENCE AND ENGINEERING A STRUCTURAL MATERIALS PROPERTIES MICROSTRUCTURE AND PROCESSING	18600	88
INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING ICASSP	17799	N/A
ELECTRONICS LETTERS	15613	174
IEEE TRANSACTIONS ON MAGNETICS	14920	50
PROFESSIONAL ENGINEERING	14247	N/A
OIL GAS JOURNAL	14216	78
AIP CONFERENCE PROCEEDINGS	13404	N/A
MATERIALS SCIENCE FORUM	13289	N/A
JOURNAL OF HAZARDOUS MATERIALS	12452	61
KEY ENGINEERING MATERIALS	12331	N/A
IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY CONFERENCE PROCEEDINGS	12315	N/A
TRANSPORTATION RESEARCH RECORD	11661	N/A
MICROWAVE AND OPTICAL TECHNOLOGY LETTERS	11087	130
WATER SCIENCE AND TECHNOLOGY	10942	27
RARE METAL MATERIALS AND ENGINEERING	10761	N/A
BIOMATERIALS	10625	40
IEEE TRANSACTIONS ON APPLIED SUPERCONDUCTIVITY	10099	20

Tabla B.4: 25 Journals con más publicaciones en ENGI.

como indicador del impacto de un journal. Por supuesto lo inverso no se cumple, ya que «CHEMICAL ENGINEERING NEWS» esta en Q4 pero tiene un porcentaje de erratas mayor que «JOURNAL OF ALLOYS AND COMPOUNDS» y éste está en Q1.

Journal	Total	Erratas	%Erratas	IF	Cuartil
CHEMICAL ENGINEERING NEWS	42024	227	0.54 %	0.27	Q4 (2)
JOURNAL OF ALLOYS AND COMPOUNDS	21300	65	0.31 %	3.00	Q1 (2) y Q2 (1)
ENVIRONMENTAL SCIENCE TECHNOLOGY	21092	277	1.31 %	5.33	Q1 (2)
INDUSTRIAL ENGINEERING CHEMISTRY RESEARCH	17957	118	0.66 %	2.59	Q1
ELECTRONICS LETTERS	15613	174	1.11 %	0.93	Q3
MATERIALS SCIENCE AND ENGINEERING A STRUCTURAL MATERIALS PROPERTIES MICROSTRUCTURE AND PROCESSING	14410	88	0.61 %	2.57	Q1 (2) y Q2 (1)
PROFESSIONAL ENGINEERING	14247	N/A	N/A	N/A	Q4 (2012)
OIL GAS JOURNAL	14170	78	0.55 %	0.07	Q4 (2)
JOURNAL OF HAZARDOUS MATERIALS	12101	61	0.50 %	4.53	Q1 (3)
MICROWAVE AND OPTICAL TECHNOLOGY LETTERS	10999	130	1.18 %	0.57	Q4 (2)
BIOMATERIALS	10607	40	0.38 %	8.56	Q1
RARE METAL MATERIALS AND ENGINEERING	10333	N/A	N/A	0.19	Q4
IEEE PHOTONICS TECHNOLOGY LETTERS	9881	66	0.67 %	2.11	Q1 (1) y Q2 (2)
JOURNAL OF MEMBRANE SCIENCE	8831	57	0.65 %	5.06	Q1 (2)
INTERNATIONAL JOURNAL OF HEAT AND MASS TRANSFER	8812	49	0.56 %	2.38	Q1 (3)

Tabla B.5: 15 Journals con más publicaciones en ENGI sin contar proceedings.

De los 6293 Journals que no son proceedings hay 1283 con al menos 100 publicaciones, de estos 369 no tienen erratas y 36 tienen menos del 0.1 %. De los 369 que no tienen erratas se buscó en qué cuartil se encuentran los 10 con más publicaciones: seis están en Q4, dos en Q3 y dos en Q3 y Q2. De manera similar se seleccionaron los 10 journals con más publicaciones con menos del 0.1 % de erratas: seis están en Q4, dos en Q3 uno en Q2 y sorpresivamente uno en Q1. En la tabla B.6 están los 20 Journals con más publicaciones sin erratas o con un porcentaje menor a 0.1 %.

Un dato interesante a destacar es que de los journals en la tabla B.6 seis ya no aparecen en la medición del IF de 2014, y hasta el año en el que aparecieron

Journal	Total	Erra	%Erra	Cuartil
PROFESSIONAL ENGINEERING	14247	N/A	N/A	Q4 (2012)
RARE METAL MATERIALS AND ENGINEERING	10333	N/A	N/A	Q4
TRANSPORTATION RESEARCH RECORD	7235	N/A	N/A	Q4
PRZEGLAD ELEKTROTECHNICZNY	4712	N/A	N/A	Q4 (2011)
AIRCRAFT ENGINEERING AND AEROSPACE TECHNOLOGY	3718	N/A	N/A	Q4
ACTA METALLURGICA SINICA	3349	N/A	N/A	Q3
CHINESE JOURNAL OF CATALYSIS	3075	N/A	N/A	Q2 (2) Q3 (1)
PROGRESS IN ELECTROMAGNETICS RESEARCH PIER	2621	N/A	N/A	Q2 (1)y Q3 (2)
JOURNAL OF CENTRAL SOUTH UNIVERSITY OF TECHNOLOGY	2302	N/A	N/A	Q3
PIPELINE GAS JOURNAL	2121	N/A	N/A	Q4 (2003)
AVIATION WEEK SPACE TECHNOLOGY	7693	4	0.05 %	Q4 (2003)
TCE	5982	2	0.03 %	Q4 (2011)
EDN	5974	1	0.02 %	Q4
MECHANICAL ENGINEERING	5847	3	0.05 %	Q4
TRANSACTIONS OF NONFERROUS METALS SOCIETY OF CHINA	5414	1	0.02 %	Q2
NAVAL ARCHITECT	4790	1	0.02 %	Q4 (2011)
REVISTA DE CHIMIE	3764	3	0.08 %	Q3
QUANTUM ELECTRONICS	3183	3	0.09 %	Q3 y Q4
JOURNAL OF TISSUE ENGINEERING AND REGENERATIVE MEDICINE	3178	2	0.06 %	Q1 (4)
R D MAGAZINE	2933	1	0.03 %	Q4

Tabla B.6: 20 Journals en ENGI sin erratas o con menos del 0.1 %.

estaban en Q4. A partir de los resultados en ENGI, la información de las erratas es un indicador complementario del impacto de una revista y en un sentido más global de la calidad de la revista.

B.1.2. Erratas en Journals de MATH

En MATH sólo el 8.37 % de los Journals tienen erratas, en la tabla B.7 se muestran los 25 journals con más publicaciones así como el total de erratas. En este caso sólo hay tres que son proceedings, procediendo de manera similar que en ENGI, se excluyen los proceedings quedando 651647 publicaciones de las 786265 iniciales con 4172 journals de los 6627 iniciales. El porcentaje de journals sin erratas es de 86.7 %.

En el TOP 50 de MATH sin proceedings sólo hay un journal sin erratas y está en Q3 y Q2, hay tres con menos de 0.1 % uno en Q4, uno en Q4 y Q3 y uno en Q3 y Q2. De manera similar a ENGI, en MATH la ausencia de erratas o porcentajes bajos son un indicador del impacto del artículo en el Top 50 sin proceedings paper.

En MATH hay 718 journals con más de 100 publicaciones, de estos, 175 no tienen de erratas. De los 10 journals con más publicaciones que no tienen erratas se tiene que tres están en Q4, dos en Q4 y Q3, tres en Q3, dos en Q3 y Q2 y uno sin información de IF. Para el caso de journals con menos del 0.1 % de erratas sólo hay 8, de éstos uno está en Q4, dos en Q4 y Q3, tres en Q3 y Q2 y dos en Q2. En la tabla B.8 se muestran los 18 journals sin erratas o con un porcentaje menor a 0.1 % de los 718 con más de 100publicaciones.

De manera global, en MATH es más conservador el efecto de la ausencia o porcentajes bajos de erratas en el impacto de los journals. Sin embargo predominan los journals en Q3 y Q4, de los 18 journals en B.8 hay 18 apariciones entre Q3 y Q4, sólo 7 de Q2 y ninguna de Q1, dando un porcentaje de 72 % para Q3 y Q4. Hay dos journals para los que no hay IF en 2014 uno con IF en 2011 y otro en ningún año.

B.1.3. Erratas en Journals de MULT

En la tabla B.9 se muestran los 25 journals con más publicaciones de MULT así como el total de erratas, de manera similar a MATH hay pocos proceedings, en este caso sólo 1, aunque tiene publicaciones que no son proceedings.

Excluyendo los proceedins se tienen 405296 registros de los 427705 iniciales, quedando 427 journals de los 765, teniendo así un porcentaje de journals con erratas de 18.97 %.

En el top 50 de journals en MULT hay sólo tres con menos del 0.1 % de erratas: uno en Q4, uno en Q3 y Q2 y uno en Q2. Realmente son los únicos tres journals con erratas que tienen menos del 0.1 % en MULT y los tres tienen más de 1000 publicaciones. En el Top 50 Hay 11 journals sin erratas: cuatro en Q4, uno en Q3 y Q4, dos en Q3, uno en Q3 y Q2, uno en Q2 y dos en Q1. Cada journal sin erratas tiene más de 500 publicaciones. En la tabla B.10 están los tres journals con menos del 0.1 % de erratas y los 11 journals sin erratas.

Journal	Total	Erratas
APPLIED MATHEMATICS AND COMPUTATION	13758	67
JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS	11937	94
LECTURE NOTES IN COMPUTER SCIENCE	9982	N/A
BIOINFORMATICS	9029	95
AIP CONFERENCE PROCEEDINGS	8198	N/A
NONLINEAR ANALYSIS THEORY METHODS APPLICATIONS	8019	47
PROCEEDINGS OF THE AMERICAN MATHEMATICAL SOCIETY	6932	50
JOURNAL OF ALGEBRA	6754	85
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS	6680	28
DISCRETE MATHEMATICS	6382	31
LINEAR ALGEBRA AND ITS APPLICATIONS	6240	50
CHAOS SOLITONS FRACTALS	6181	38
COMPUTERS MATHEMATICS WITH APPLICATIONS	6139	43
MATHEMATICAL PROBLEMS IN ENGINEERING	5767	7
INTERNATIONAL JOURNAL OF QUANTUM CHEMISTRY	5346	51
COMMUNICATIONS IN ALGEBRA	5039	52
STATISTICS IN MEDICINE	4827	74
COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING	4386	23
STATISTICS PROBABILITY LETTERS	4242	37
INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING	4083	12
MATHEMATICAL AND COMPUTER MODELLING	4048	17
JOURNAL OF STATISTICAL PLANNING AND INFERENCE	3934	26
INTERNATIONAL JOURNAL OF BIFURCATION AND CHAOS	3889	N/A
JOURNAL OF DIFFERENTIAL EQUATIONS	3879	40
APPLIED MATHEMATICAL MODELLING	3867	27

Tabla B.7: 25 Journals con más publicaciones en MATH.

Los registros relevantes de la tabla B.10 son los dos journals en Q1 sin erratas uno sólo tiene ese registro hasta 2010 y el otro (PeerJ) apenas está en WoS desde 2013 y es un journal open acces que tiene retroalimentación a las publicaciones en

Journal	Total	Erra	%Erra	Cuartil
INTERNATIONAL JOURNAL OF BIFURCATION AND CHAOS	3572	N/A	N/A	Q3 y Q2
ELECTRONIC JOURNAL OF COMBINATORICS	2111	N/A	N/A	Q4 y Q3
SCIENCE IN CHINA SERIES A MATHEMATICS	1036	N/A	N/A	Q3 (2011)
PROCEEDINGS OF THE STEKLOV INSTITUTE OF MATHEMATICS	938	N/A	N/A	Q4 (2)
UTILITAS MATHEMATICA	923	N/A	N/A	Q4 (2)
CHINESE ANNALS OF MATHEMATICS SERIES B	828	N/A	N/A	Q3
JOURNAL OF COMPUTATIONAL MATHEMATICS	813	N/A	N/A	Q3 (2)
METRIKA	681	N/A	N/A	Q4
FILOMAT	662	N/A	N/A	Q3 y Q2
SPRINGER SERIES IN OPTIMIZATION AND ITS APPLICATIONS	623	N/A	N/A	N/A
ELECTRONIC JOURNAL OF DIFFERENTIAL EQUATIONS	2466	1	0.04 %	Q4 y Q3
LECTURE NOTES IN MATHEMATICS	2419	1	0.04 %	Q4
APPLIED MATHEMATICS AND MECHANICS ENGLISH EDITION	2409	2	0.08 %	Q3 y Q2
ACTA MATHEMATICA SINICA ENGLISH SERIES	2252	1	0.04 %	Q4 y Q3
DISCRETE DYNAMICS IN NATURE AND SOCIETY	1463	1	0.07 %	Q3 y Q2
ADVANCES IN DIFFERENCE EQUATIONS	1416	1	0.07 %	Q3 y Q2
APPLIED NUMERICAL MATHEMATICS	1216	1	0.08 %	Q2
MATHEMATICAL INEQUALITIES APPLICATIONS	1085	1	0.09 %	Q2

Tabla B.8: 18 Journals en MATH sin erratas o con menos del 0.1 %.

Journal	Total	Erra	%Erra
PLOS ONE	115435	1427	1.24 %
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	55301	1505	2.72 %
NATURE	40057	1074	2.68 %
SCIENCE	39464	1328	3.37 %
NEW SCIENTIST	33596	273	0.81 %
ANNALS OF THE NEW YORK ACADEMY OF SCIENCES	13806	38	0.28 %
CURRENT SCIENCE	11118	68	0.61 %
CHINESE SCIENCE BULLETIN	8529	8	0.09 %
SCIENTIST	7815	54	0.69 %
SCIENTIFIC REPORTS	7577	159	2.10 %
SCIENTIFIC AMERICAN	6772	90	1.33 %
NATURE COMMUNICATIONS	5782	90	1.56 %
SCIENTIFIC WORLD JOURNAL	5766	14	0.24 %
INTERNATIONAL JOURNAL OF BIFURCATION AND CHAOS	3889	N/A	N/A
TECHNOLOGY REVIEW	3383	21	0.62 %
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES	3078	16	0.52 %
R D MAGAZINE	2933	1	0.03 %
JOVE JOURNAL OF VISUALIZED EXPERIMENTS	2924	24	0.82 %
PROCEEDINGS OF THE ROYAL SOCIETY A MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES	2809	33	1.17 %
ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING	2272	6	0.26 %
SOUTH AFRICAN JOURNAL OF SCIENCE	2015	2	0.10 %
AMERICAN SCIENTIST	1984	53	2.67 %
JOURNAL OF THE ROYAL SOCIETY INTERFACE	1941	21	1.08 %
COMPTES RENDUS DE L ACADEMIE BULGARE DES SCIENCES	1813	4	0.22 %
ISSUES IN SCIENCE AND TECHNOLOGY	1805	2	0.11 %

Tabla B.9: 25 Journals con más publicaciones en MULT.

Journal	Total	Erra	%Erra	Cuartil
CHINESE SCIENCE BULLETIN	8508	8	0.1 %	Q2
R D MAGAZINE	2933	1	0.0 %	Q4
DISCRETE DYNAMICS IN NATURE AND SOCIETY	1463	1	0.1 %	Q3 y Q2
INTERNATIONAL JOURNAL OF BIFURCATION AND CHAOS	3572	N/A	N/A	Q3 y Q2
THESCIENTIFICWORLDJOURNAL	1366	N/A	N/A	Q1 (2010)
SAINS MALAYSIANA	1262	N/A	N/A	Q3
SCIENTIFIC RESEARCH AND ESSAYS	987	N/A	N/A	Q3 (2010)
DEFENCE SCIENCE JOURNAL	953	N/A	N/A	Q4
HERALD OF THE RUSSIAN ACADEMY OF SCIENCES	925	N/A	N/A	Q4
ADVANCED SCIENCE LETTERS	860	N/A	N/A	Q2 (2010)
PEERJ	703	N/A	N/A	Q1
JOHNS HOPKINS APL TECHNICAL DIGEST	570	N/A	N/A	Q4 (2013)
JOURNAL OF ZHEJIANG UNIVERSITY SCIENCE A	561	N/A	N/A	Q3 y Q4
ANTHROPOLOGIST	560	N/A	N/A	Q4

Tabla B.10: Journals con menos de 0.1 % de erratas y sin erratas en MULT.

la propia página web (https://peerj.com/). Finalmente de los 11 journals sin erratas cuatro sólo tienen calculado el IF antes de 2014.

B.1.4. Erratas en Journals de PHYS

En la tabla B.11 se muestran los 25 journals con más publicaciones de PHYS, el total de erratas y el porcentaje de erratas de cada journal. Los cuatro registros sin erratas corresponden a proceedings, en este caso sólo 1, aunque tiene publicaciones que no son proceedings.

Excluyendo los proceedings paper quedan 1521398 publicaciones con 2570 journals teniendo 15.29 % de journals con erratas. En el Top 50 de estos journals sólo hay dos journals con menos del 0.1 % de erratas, uno en Q3 y uno en Q2. Y no hay journals sin erratas en el Top 50.

De los 2570 journasl hay 462 con más de 100 publicaciones, de éstos hay 8 journals con menos de 0.1 % de erratas: tres en Q4, uno en Q4 y Q3, dos en Q3 y dos en Q2. También aquí en PHYS predominan los journals en Q4 y no hay ninguno en Q1. Los 10 journals con más publicaciones de los 462 que no tienen erratas hay: tres en Q4, uno en Q4 y Q3, uno en Q3, uno en Q3 y Q2, uno en Q2, uno en Q2 y Q1, uno en Q1 y uno sin clasificación. En la tabla B.12 se muestran los 18 journals con menos de 0.1 % de erratas y los 10 sin erratas.

En la tabla B.12 se observa que el journal sin erratas en Q1 es «COMMUNICATIONS IN COMPUTATIONAL PHYSICS» lo cual es congruente con los datos globales obtenidos de Scopus, ya que el área de COMP (Computation) tiene 0.11 % de erratas en promedio para el periodo 2000-2014.

Journal	Total	Erratas	%Erra
PHYSICAL REVIEW B	83030	1610	1.94 %
AIP CONFERENCE PROCEEDINGS	73229	N/A	N/A
APPLIED PHYSICS LETTERS	67587	873	1.29 %
JOURNAL OF APPLIED PHYSICS	55614	501	0.90 %
PHYSICAL REVIEW LETTERS	53687	1425	2.65 %
JOURNAL OF CHEMICAL PHYSICS	40540	748	1.85 %
PHYSICAL REVIEW D	40076	917	2.29 %
PHYSICAL REVIEW E	36654	675	1.84 %
PHYSICAL REVIEW A	34354	785	2.29 %
JOURNAL OF PHYSICS CONFERENCE SERIES	29983	N/A	N/A
PROCEEDINGS OF SPIE	25631	N/A	N/A
PROCEEDINGS OF THE SOCIETY OF PHOTO OPTICAL INSTRUMENTATION ENGINEERS SPIE	25390	N/A	N/A
JOURNAL OF PHYSICAL CHEMISTRY A	23659	261	1.10 %
APPLIED SURFACE SCIENCE	22158	80	0.36 %
NUCLEAR INSTRUMENTS METHODS IN PHYSICS RESEARCH SECTION A ACCELERATORS SPECTROMETERS DETECTORS AND ASSOCIATED EQUIPMENT	21954	120	0.55 %
THIN SOLID FILMS	20823	77	0.37 %
PHYSICAL CHEMISTRY CHEMICAL PHYSICS	19326	173	0.90 %
JOURNAL OF PHYSICS CONDENSED MATTER	19254	139	0.72 %
CHEMICAL PHYSICS LETTERS	19188	164	0.85 %
JOURNAL OF HIGH ENERGY PHYSICS	18268	121	0.66 %
ACTA PHYSICA SINICA	16964	3	0.02 %
JOURNAL OF MAGNETISM AND MAGNETIC MATERIALS	16550	45	0.27 %
JOURNAL OF CRYSTAL GROWTH	15364	61	0.40 %
MATERIALS LETTERS	15151	47	0.31 %
PHYSICS LETTERS B	15030	248	1.65 %

Tabla B.11: 25 Journals con más publicaciones en PHYS.

Journal	Total	Erra	%Erra	Cuartil
PROGRESS IN ELECTROMAGNETICS RESEARCH PIER	2621	N/A	N/A	Q3
HIGH ENERGY PHYSICS AND NUCLEAR PHYSICS CHINESE EDITION	2090	N/A	N/A	Q4 (2009)
PROBLEMS OF ATOMIC SCIENCE AND TECHNOLOGY	2074	N/A	N/A	Q4
SENSOR LETTERS	1486	N/A	N/A	Q4 (2013)
INTERNATIONAL JOURNAL OF PHOTOENERGY	1334	N/A	N/A	Q3 y Q2
LECTURE NOTES IN PHYSICS	1095	N/A	N/A	N/A
SCIENCE OF ADVANCED MATERIALS	1040	N/A	N/A	Q2 y Q1
SYMMETRY INTEGRABILITY AND GEOMETRY METHODS AND APPLICATIONS	941	N/A	N/A	Q2
MICRO	890	N/A	N/A	Q4 y Q3
COMMUNICATIONS IN COMPUTATIONAL PHYSICS	876	N/A	N/A	Q1
ACTA PHYSICA SINICA	16964	3	0.018 %	Q3
CHINESE PHYSICS B	7522	4	0.053 %	Q2
TECHNICAL PHYSICS LETTERS	4989	1	0.020 %	Q4
QUANTUM ELECTRONICS	3183	3	0.094 %	Q4
CHINESE PHYSICS C	1799	1	0.056 %	Q3 y Q4
JAPANESE JOURNAL OF APPLIED PHYSICS PART 2 LETTERS	1746	1	0.057 %	Q3
EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS	1554	1	0.064 %	Q2
JOURNAL OF SURFACE INVESTIGATION X RAY SYNCHROTRON AND NEUTRON TECHNIQUES	1039	1	0.096 %	Q4

Tabla B.12: Journals sin erratas y con menos de 0.1 % de erratas en PHYS.

Apéndice C

Manejo de datos

La limpieza o normalización de la base de datos suele ser un trabajo exhaustivo, hay varias investigaciones que mencionan la metodología que siguieron para el tratamiento de bases de datos bibliométricas [24]. En esta investigación fue dependiendo de cada uno de los procesos de análisis, aunque buena parte de ese trabajo se hizo manualmente desde las páginas web de las revistas como en el caso de la información de las erratas de Nature.

Una herramienta muy útil para analizar los datos desde la WoS es "Science of Science (Sci2)" [67] que desarrolló la School of Library and Information Science y el Cyberinfrastructure for Network Science Center de la Universidad de Indiana. Esta herramienta es de código abierto lo que implica que se pueden modificar los algoritmos y adecuarlos a las necesidades propias. Sólo se uso software de acceso abierto para hacer los cálculos, gráfica y análisis necesarios, además de Sci2 se uso Gephi [65], R [68] y Python.

Todos los cálculos se hicieron en una computadora con un procesador i5-3350 a 3.1 GHz con 8 GB de memoria RAM en un sistema operativo de 64 bits. Cabe señalar que el caso de la red de APS algunos cálculos tardaron más de un día y en total fueron necesarias varias semanas de cómputo para generar todos los resultados.

En las siguientes dos secciones hay ejemplos de código en Python que se usó para calcular la influencia y otras medidas de la red completa de APS y en R que se usó para los ajustes.

C.1. En Python

APS_2013

June 12, 2016

In [1]:
```
#primero se leen los paquetes necesarios
import networkx as nx
import numpy as np
import scipy as sc
```

In [3]:
```
#luego se leen los datos de la red
G_APS = nx.read_edgelist('APS13.csv', delimiter=',', nodetype=str,create_using=nx.DiGraph())
```

In [5]:
```
# regresa la cantidad de nodos en la red
len (nx.nodes(G_APS))
```

Out[5]: 531478

In [6]:
```
# regresa la cantidad de aristas en la red
len (nx.edges(G_APS))
```

Out[6]: 6039993

In [14]:
```
# se obtiene la componente gigante de la red
Gc_APS = max(nx.weakly_connected_component_subgraphs(G_APS), key=len)
M_APS=nx.adjacency_matrix(Gc_APS)
```

In [8]:
```
Nodos_GcAPS = len (nx.nodes(Gc_APS))
```

In [9]:
```
# se calcula la influencia
aps=np.ones((Nodos_GcAPS,1))
TC_MAPS=sc.sparse.linalg.expm_multiply(M_APS,aps)
STC_MAPS=TC_MAPS.sum(axis=0)
P_GAPS=TC_MAPS/STC_MAPS
```

In [10]:
```
# diccionario para asociar la influencia a cada nodo
I_APS={}
for x in range(Nodos_GcAPS):
            I_APS.setdefault(nx.nodes(Gc_APS)[x],P_GAPS[x,0])
```

In [11]:
```
# calcula el grado de salida de cada nodo normalizado por n-1
Deg_GcAPS=nx.out_degree_centrality(Gc_APS)
```

In [12]:
```
# calcula el grado de salida de cada nodo sin normalizar
Deg2_GcAPS=Gc_APS.out_degree()
```

In [20]:
```
# calcula los hubs y las authorities de cada nodo
h,a=nx.hits_scipy(Gc_APS,max_iter=1000, tol=1e-06, normalized=True)
```

In [21]:
```
# calcula el page rank de cada nodo
pr = nx.pagerank_scipy(Gc_APS)
```

In [22]:
```
# se exportan los resultados
resultsAPS2 = [(k,Deg_GcAPS[k],Deg2_GcAPS[k],I_APS[k],h[k],a[k],pr[k])
               for k in Gc_APS]
f = open('Medidas2_APS.txt','w')
for item in resultsAPS2:
    f.write('\t'.join(map(str,item)))
    f.write('\n')
f.close()
```

In [ ]:

C.2. En R

El siguiente script de R fue utilizado para hacer el ajuste de la distribución de citas en la red de física de México, incluyendo las simulaciones necesarias.

```
Fis2015=read.csv("RedFis2015dos.csv")
FGc15=FGC15$globalcitationcount

FGC15_pl=displ$new(FGc15)
est_FGC15_pl=estimate_xmin(FGC15_pl)
FGC15_pl$setXmin(est_FGC15_pl)
bs_FGC15=bootstrap_p(FGC15_pl,no_of_sims = 100, threads = 4)
plot(bs_FGC15)

FGC15_ln=dislnorm$new(FGc15)
est_FGC15_ln=estimate_xmin(FGC15_ln)
FGC15_ln$setXmin(est_FGC15_ln)
bs_FGC15_ln=bootstrap_p(FGC15_ln,no_of_sims = 100, threads = 4)
plot(bs_FGC15_ln)

FLC15=subset(Fis2015, localcitationcount>0,select=localcitationcount)
FLc15=FLC15$localcitationcount

FLC15_pl=displ$new(FLc15)
est_FLC15_pl=estimate_xmin(FLC15_pl)
FLC15_pl$setXmin(est_FLC15_pl)
bs_FLC15=bootstrap_p(FLC15_pl,no_of_sims = 100, threads = 4)
plot(bs_FLC15)

FLC15_ln=dislnorm$new(FLc15)
est_FLC15_ln=estimate_xmin(FLC15_ln)
FLC15_ln$setXmin(est_FLC15_ln)
bs_FLC15_ln=bootstrap_p(FLC15_ln,no_of_sims = 100, threads = 4)
plot(bs_FLC15_ln)

#Para comparar log-norm con power law
FLC15_ln2=dislnorm$new(FLc15)
FLC15_ln2$setXmin(15)
```

```
est_FLC15_ln2=estimate_pars(FLC15_ln2)
FLC15_ln2$setPars(est_FLC15_ln2)
compFLC15=compare_distributions(FLC15_pl,FLC15_ln2)
plot(compFLC15)

FLC15_pl2=displ$new(FLc15)
FLC15_pl2$setXmin(2)
est_FLC15_pl2=estimate_pars(FLC15_pl2)
FLC15_pl2$setPars(est_FLC15_pl2)

compFLC15_2=compare_distributions(FLC15_pl2,FLC15_ln)
plot(compFLC15_2)
#termina comparacion
```

Bibliografía

[1] D.J. De Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[2] Eugene Garfield. From the science of science to scientometrics visualizing the history of science with histcite software. *Journal of Informetrics*, 3(3):173 – 179, 2009. Science of Science: Conceptualizations and Models of Science.

[3] Robert Adler, John Ewing, and Peter Taylor. Citation statistics. *Statist. Sci.*, 24(1):1–14, 02 2009.

[4] M.A Pérez. Usos y abusos de la cienciometría. *CINVESTAV*, 25(1):29–33, 2006.

[5] *Whitepaper Using Bibliometrics: A Guide to Evaluating Research Performance with Citation Data*. Thomson Reuters, 2008.

[6] Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

[7] José A. de la Peña. Impact functions on the citation network of scientific articles. *Journal of Informetrics*, 5(4):565 – 573, 2011.

[8] Saikou Y. Diallo, Christopher J. Lynch, Ross Gore, and Jose J. Padilla. Identifying key papers within a journal via network centrality measures. *Scientometrics*, 107(3):1005–1020, 2016.

[9] Dylan Walker, Huafeng Xie, Koon-Kiu Yan, and Sergei Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007.

[10] Jianlin Zhou, An Zeng, Ying Fan, and Zengru Di. Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, 106(2):805–816, 2016.

[11] Ernesto Estrada and Naomichi Hatano. Communicability in complex networks. *Physical Review E*, 77(3):6111–6123, 2008.

[12] Michele Benzi and Christine Klymko. Total communicability as a centrality measure. *Journal of Complex Networks*, 1(2):124–149, 2013.

[13] Christine Klymko. *Centrality and Communicability Measures in Complex Networks: Analysis and Algorithms*. PhD thesis, Emory University, 2013.

[14] Ernesto Estrada, José A de la Peña, and Naomichi Hatano. Walk entropies in graphs. *Linear Algebra and its Applications*, 443:235–244, 2014.

[15] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[16] Leo Egghe. The hirsch index and related impact measures. *Annual Review of Information Science and Technology*, 44(1):65–114, 2010.

[17] Lutz Bornmann and Hans-Dieter Daniel. Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3):391–392, 2005.

[18] Sune Lehmann, Andrew D. Jackson, and Benny E. Lautrup. Measures for measures. *NATURE*, 444(7122):1003–1004, DEC 21 2006.

[19] P. Vinkler. Eminence of scientists in the light of the h-index and other scientometric indicators. *Journal of Information Science*, 33(4):481–491, 2007. cited By 56.

[20] A. Mazloumian. Predicting scholars' scientific impact. *PLoS ONE*, 7(11), 2012.

[21] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[22] K. Frenken, S. Hardeman, and J. Hoekman. Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, 3(3):222–232, 2009.

[23] R.K. Pan, K. Kaski, and S. Fortunato. World citation and collaboration networks: Uncovering the role of geography in science. *Scientific Reports*, 2, 2012.

[24] P. Deville, D. Wang, R. Sinatra, C. Song, V.D. Blondel, and A.-L. Barabási. Career on the move: Geography, stratification, and scientific impact. *Scientific Reports*, 4, 2014.

[25] D. Ubfal and A. Maffioli. The impact of funding on research collaboration: Evidence from a developing country. *Research Policy*, 40(9):1269–1279, 2011.

[26] J.M. Benavente, G. Crespi, L. Figal Garone, and A. Maffioli. The impact of national research funds: A regression discontinuity approach to the chilean fondecyt. *Research Policy*, 41(8):1461–1475, 2012.

[27] Eric D Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*. Springer, 2014.

[28] AL Barabasi. *Network Science,*. Cambridge University Press, 2016.

[29] S. Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58(6):49–54, 2005.

[30] P. Erdős and A Rényi. On the evolution of random graphs. In *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*, pages 17–61, 1960.

[31] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[32] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[33] A.L Barabási, H Jeong, Z Néda, E Ravasz, A Schubert, and T Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3–4):590 – 614, 2002.

[34] AL Barabasi and R Albert. Emergence of scaling in random networks. *SCIENCE*, 286(5439):509–512, 1999.

[35] S. Redner. How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B*, 4(2):131–134, 1998. cited By 691.

[36] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. cited By 1576.

[37] M. Brzezinski. Power laws in citation distributions: evidence from scopus. *Scientometrics*, 103(1):213–228, 2015. cited By 0.

[38] John P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8), 08 2005.

[39] John P. A. Ioannidis. Why most clinical research is not useful. *PLoS Med*, 13(6):1–10, 06 2016.

[40] Claudia Gonzalez-Brambila and Francisco M. Veloso. The determinants of research output and impact: A study of mexican researchers. *Research Policy*, 36(7):1035 – 1051, 2007.

[41] The impact of network embeddedness on research output. *Research Policy*, 42(9):1555 – 1567, 2013.

[42] Claudia N. Gonzalez-Brambila, Leonardo Reyes-Gonzalez, Francisco Veloso, and Miguel Angel Perez-Angón. The scientific impact of developing nations. *PLoS ONE*, 11:1–14, 03 2016.

[43] Ma. Elena Luna-Morales, Francisco Collazo-Reyes, Jane M. Russell, and Miguel Ángel Pérez-Angón. Early patterns of scientific production by mexican researchers in mainstream journals, 1900–1950. *Journal of the American Society for Information Science and Technology*, 60(7):1337–1348, 2009.

[44] F. Collazo-Reyes, M. E. Luna-Morales, J. M. Russell, and M. A. Pérez-Angón. Publication and citation patterns of latin american & caribbean journals in the sci and ssci from 1995 to 2004. *Scientometrics*, 75(1):145–161, 2008.

[45] J.A. Pichardo-Corpus, J.G. Contreras, and J.A. de la Peña. Errata as a scientometric indicator. *Enviado a PLOS ONE*, 2016.

[46] Ilkka Niiniluoto. Scientific progress. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition, 2015.

[47] Editorial. Correction or retraction? *Nature*, 444:123–124, 2006.

[48] D. Butler. Theses spark twin dilemma for physicists. *Nature*, 420(6911):5, 2002.

[49] M.R. Beasley and Bell Labs Innovations. *Report of the Investigation Committee on the Possibility of Scientific Misconduct in the Work of Hendrik Schön and Coauthors*. Bell Labs, 2002.

[50] J. M. Budd, M. Sievert, T. R. Schultz, and C Scoville. Effects of article retraction on citation and practice in medicine. *Bulletin of the Medical Library Association*, 87(4):437–443, 1999.

[51] Elizabeth Wager and Peter Williams. Why and how do journals retract articles? an analysis of medline retractions 1988-2008. *Journal of Medical Ethics*, 37(9):567–570, 2011.

[52] Paul J. Hauptman, Eric S. Armbrecht, John T. Chibnall, Camelia Guild, Jeremy P. Timm, and Michael W. Rich. Errata in medical publications. *The American Journal of Medicine*, 127(8):779–785.e1, 2014.

[53] P. Sprent and N. Smeeton. *Applied Nonparametric Statistical Methods,*. Chapman & Hall/CRC Texts in Statistical Science, CRC Press., 4 edition, 2007.

[54] Scopus. Content coverage guide. `https://www.elsevier.com/solutions/scopus/content`, January 2016.

[55] Nature. Correction and retraction policy. `http://www.nature.com/authors/policies/corrections.html`, 2016. [Online; accessed 20-March-2016].

[56] Editorial. Retractions' realities. *Nature*, 422:1, 2003.

[57] Eric Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

[58] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[59] D.M. Cvetković, M. Doob, and H. Sachs. *Spectra of graphs: theory and application*. Pure and applied mathematics. Academic Press, 1980.

[60] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[61] Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, pages 1170–1182, 1987.

[62] Ernesto Estrada and Juan A Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):6103–6112, 2005.

[63] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August 2008.

[64] Awad H Al-Mohy and Nicholas J Higham. Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM journal on scientific computing*, 33(2):488–511, 2011.

[65] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.

[66] Michele Benzi, Ernesto Estrada, and Christine Klymko. Ranking hubs and authorities using matrix functions. *Linear Algebra and its Applications*, 438(5):2447–2474, 2013.

[67] Sci2 Team. Science of science (sci2) tool. *Indiana University and SciTech Strategies*, 2009.

[68] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[69] Colin S. Gillespie. Fitting heavy tailed distributions: The powerlaw package. *Journal of Statistical Software*, 64(2), 2015.

[70] J.G. Contreras, D.J. Álvarez, and J.A. Pichardo-Corpus. Mario molina and the saga of ozone example of linking science and society [mario molina y la saga del ozono: Ejemplo de vinculación ciencia-sociedad]. *Andamios*, 12(29):15–32, 2015.

[71] MJ MOLINA and FS ROWLAND. STRATOSPHERIC SINK FOR CHLOROFLUOROMETHANES - CHLORINE ATOMIC-CATALYSED DESTRUCTION OF OZONE. *NATURE*, 249(5460):810–812, 1974.

[72] Latour B. Joliot: punto de encuentro de la historia y de la física. In M. Serres, editor, *Historia de las Ciencias*, pages 553–573. Cátedra, 1998.

[73] Centro Mario Molina. `http://centromariomolina.org/`, 3 de Febrero de 2015, 2015.

[74] S.J. Diamond. Why ban aerosol sprays? a noted chemist tells how they endanger the ozone—and us. *People weekly*, 6(16):45–51, 1976.

[75] F. S. Rowland and Mario J. Molina. *The Cfc-Ozone Puzzle: Environmental Science in the Global Arena*. National Council for Science and the Environment, Washington D. C., 2001.

[76] A.M.P.S.P.A. Chemistry and N.R.C.P.A. Chemistry. *Halocarbons, Effects on Stratospheric Ozone*. National Academy of Sciences, 1976.

[77] S.O. Andersen, K.M. Sarma, and L. Sinclair. *Protecting the Ozone Layer: The United Nations History*. Taylor & Francis, 2002.

[78] FS ROWLAND and MJ MOLINA. OZONE QUESTION. *SCIENCE*, 190(4219):1038–1039, 1975.

[79] JC FARMAN, BG GARDINER, and JD SHANKLIN. LARGE LOSSES OF TOTAL OZONE IN ANTARCTICA REVEAL SEASONAL CLOX/NOX INTERACTION. *NATURE*, 315(6016):207–210, 1985.

[80] UNEP. *Montreal Protocol on Substances That Deplete the Ozone Layer*. United Nations Environment Programme, 1987.

[81] WHO. *Assessment for Decision-Makers: Scientific Assessment of Ozone Depletion: 2014*. World Meteorological Organization, 2014.

[82] Cass R Sunstein. Of montreal and kyoto: a tale of two protocols. *Harv. Envtl. L. Rev.*, 31:1, 2007.

[83] R. J. Cicerone, D. H. Stedman, R. S. Stolarski, A. N. Dingle, and R. A. Cellarius. *Assessment of Possible Environmental Effects of Space Shuttle Operations*. NASA; United States, 1973.

[84] M.J. Cobo, A.G. Lõpez-Herrera, E. Herrera-Viedma, and F. Herrera. Scimat: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, 63(8):1609–1630, 2012.

[85] David Pendlebury. White paper using bibliometrics in evaluating research. 2010.