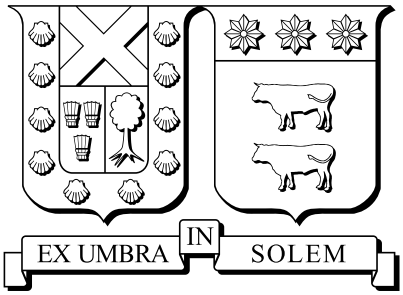
**UNIVERSIDAD TECNICA FEDERICO SANTA MARIA**
Departamento de Informática
Valparaíso - Chile




# El Espacio Investigación: Un Mapa de Producción de Ciencia

Tesis presentada para la obtención
del grado académico de

Doctor en Ingeniería Informática

**Miguel R. Guevara Albornoz**

**Comité de Evaluación**
Prof. Dr. Marcelo Mendoza Rocha (Guía, UTFSM)
Prof. Dr. César A. Hidalgo (Co-guía, MIT)
Prof. Dr. Andrés Moreira Wenzel (Correferente, UTFSM)
Prof. Dr. Sebastián Ríos (Evaluador externo, U. Chile)
Prof. Dr. Rodrigo Costas (Evaluador externo, CWTS, Leiden U.)
Prof. Dr. Claudio Moraga (Evaluador externo, Dortmund U.)

Octubre, 2016

**UNIVERSIDAD TECNICA FEDERICO SANTA MARIA**
Departamento de Informática
Valparaíso - Chile

Título de la tesis

# El Espacio Investigación: Un Mapa de Producción de Ciencia

Autor
**Miguel R. Guevara Albornoz**

Tesis enviada al Departamento de Informática de la Universidad Técnica Federico Santa María con los requerimientos para obtener el grado de Doctor en Ingeniería Informática.

| | |
|---|---|
| Dr. Marcelo Mendoza R.<br>(Guía) | _____ |
| Dr. César Hidalgo R.<br>(Co-guía) | _____ |
| Dr. Andrés Moreira W.<br>(Evaluador Interno) | _____ |
| Dr. Sebastián Ríos.<br>(Evaluador Externo Nacional) | _____ |
| Dr. Claudio Moraga.<br>(Evaluador Externo Internacional) | _____ |
| Dr. Rodrigo Costas.<br>(Evaluador Externo Internacional) | _____ |
| Dr. Hernán Astudillo.<br>(Presidente de la comisión) | _____ |

Octubre, 2016

a mi adorable familia, donde cada uno aporta la nota que conforma nuestra diaria sinfonía
Mariana, José Miguel, Antonia, Ignacio, Josefina y Paulina

# Agradecimientos

En estos casi seis años, son muchas las instituciones, proyectos y, principalmente, personas que han contribuido para que esta tesis llegue a su término. Procurando no omitir a nadie, abusaré en la extensión de estos agradecimientos.

Un especial agradecimiento a la Agrupación de Universidades Regionales de Chile AUR, institución que me otorgó la beca arancelaria durante todos mis estudios de doctorado. Esta beca gatilló mi ingreso al programa de doctorado.

Mi gratitud a las instituciones que facilitaron y aportaron con recursos para la realización de esta tesis. La Universidad de Playa Ancha, UPLA, me entregó el invaluable tiempo para iniciar, cursar y concluir mis estudios de doctorado. La Universidad Técnica Federico Santa María, UTFSM, el Departamento de Informática, DI y el programa de Doctorado, apoyaron mi participación en congresos internacionales donde esta tesis fue difundida además de proveer de un espacio grato y bien equipado para el trabajo diario. El Instituto Tecnológico de Massachusetts, MIT, participó fuertemente en el financiamiento para mis dos estadías en Estados Unidos, durante 2013 y 2014-2015. Además de apoyar este trabajo con infraestructura tecnológica de primer nivel.

Fueron fundamentales para el desarrollo de esta tesis, los proyectos de investigación: FONDECYT 11121435 de la Comisión Nacional de Ciencia y Tecnología Conicyt, Chile. MECESUP FSM1101 del Ministerio de Educación y Cultura del Gobierno de Chile. UPA01-1516 de la UPLA. También el Programa de Inicio a la Investigación Científica PIIC, convocatorias 2013 y 2015 de la UTFSM. Proyecto de colaboración internacional entre MIT y Chile, MISTI-CHILE Seed Fund, convocatoria 2013 de MIT.

Mi agradecimiento a mi profesor tutor y guía Marcelo Mendoza que me inició en el fabuloso mundo de la minería de la Web y quien vislumbró este tema de tesis como un trabajo atractivo y valioso para la Informática. Marcelo supo guiar esta tesis doctoral con especial dedicación además de darme la suficiente libertad para que yo pudiera hacer este trabajo tan de mis entrañas como me fuera posible. Además de enseñar-haciendo, Marcelo también fue de gran soporte en la logística para mis actividades de doctorado.

Gracias a César Hidalgo, mi profesor supervisor en el MediaLab del MIT quien supo guiar el desarrollo científico de esta tesis desde su mirada holística y compleja de la ciencia, la información, las artes y el mundo. César me brindó la oportunidad de salir del subdesarrollo científico acogiéndome en su grupo de investigación *MacroConnections*, en la *meca* de la

ingeniería y la innovación. Además César gestionó todo cuanto le fue posible para hacer de mis estadías en la zona de Boston, lo más cómodas y duraderas. Su atención vehemente a mi trabajo durante todo enero 2016, en el campamento de investigación Bits Bots and Behavior $B^3$, fue fundamental para llegar a buen puerto aunque ya estábamos en uno.

Mi respeto y agradecimiento al profesor Héctor Allende Olivares, que es de aquellos profesores que aún entienden la universidad como el espacio donde se discuten y reflexionan las ideas. Su presencia en el laboratorio de doctorado es la de mayor frecuencia y fue muy agradable poder conversar con él de las Máquinas de Aprendizaje, pero también —y cuanto más— de la realidad nacional, del Chile de antes, del Chile del futuro, de la institucionalidad universitaria, de la política, de la historia, de la filosofía de la ciencia, de la ética, y de un sinnúmero de tópicos que sería necesario un mapa de la ciencia para poder describirlos adecuadamente. Me siento afortunado de haber recibido su atención y respeto.

Mi postulación fue apoyada por todos los estamentos jerárquicos de la UPLA, a los cuales agradezco en las autoridades de la época: Mario Bruno, director del departamento de Computación e Informática; Daniel Valdivia, decano de la Facultad de Ingeniería; Daniel López, Vicerrector de Investigación y Postgrado; Patricio Sanhueza, rector de la Universidad. Para continuar este proceso, fue fundamental el apoyo permanente de las autoridades de la Facultad de Ingeniería, de la UPLA, los diferentes decanos: Manuel Contreras López y Verónica Meza, quienes me brindaron total apoyo en cada uno de sus respectivos períodos de gobierno.

Gracias a mis colegas del laboratorio de doctorado de la UTFSM, que fueron pieza central en la polinización de este trabajo de tesis. Su retroalimentación del día a día fue de gran aporte. Agradezco también los aportes y la convivencia diaria de mis co-laboratoristas del MediaLab en MIT. Después de quince años de haber escuchado de este bastión de la creatividad, compartir laboratorio con ellos fue una experiencia revitalizante al mismo tiempo que desafiante. En particular agradezco a quienes participaron cercanamente de este trabajo de tesis. Manuel Aristarán realizó un aporte invaluable en la obtención y almacenamiento de datos, además de iluminarme con su magistral manejo de la informática en general, haciendo con el computador lo que Claudio Arrau hacía con el piano. El bávaro Dominik Hartmann aportó la mirada cualitativa del trabajo científico y la amistad del día a día, la que ha permanecido en el tiempo a pesar de la distancia. Su incontinencia creativa fue un aliciente permanente. Hacia el final de este proceso, Christian Jara-Figueroa aportó con importante retroalimentación desde su sencilla genialidad.

En el ámbito personal, agradezco a mis padres Franklin Guevara y Gloria Lola Albornoz quienes residen en Quito-Ecuador. Mi padre fue visionario en adquirir —en pleno boom de los computadores personales— aquel Macintosh 512K800 que cambiaría mi vida para siempre. Mi madre fue el bastión del que supe llenarme de amor y paciencia para rodar toda la vida. A los dos les agradezco, además de la formación valórica, el haberme hecho partícipe del negocio familiar, donde aprendí desde muy pequeño el valor del trabajo y también del ciclo productivo de la imprenta. A mis 18 años, ahí auto-aprendí a diseñar pequeños libros (cuadernillos) por computador, lo que sería seguramente un hecho premonitorio a los *papers* que ahora escribo-diseño.

*Last but not least*, mi agradecimiento más sentido a mi familia, mi esposa Paulina Hurtado y nuestros hijos Josefina Abarca, Ignacio Abarca, Antonia Abarca, José Miguel Guevara y

Mariana Guevara, a quienes está dedicada esta tesis. Mi familia es el centro, guía y motor de cada paso que soy capaz de colocar delante del otro. Compartir nuestras vidas me hizo recordar que debo pensar en grande. Ellos fueron generosos en sacrificar el tiempo familiar necesario para que yo pudiera cursar y terminar este trabajo de tesis. Además, todos supieron acomodar sus vidas para acompañarme en esta travesía que también reconocieron suya. El orgullo que sienten al verme concluir este proceso es pequeño comparado al orgullo que yo siento al ser su papastro, papá y esposo. Con diligente amor, Paulina me motivó a iniciar, me apoyó para continuar, me empujó cuando retrocedía, me levantó cuando desfallecía y caminó conmigo de la mano hasta el final. Mi agradecimiento y redondo amor a su apoyo incondicional, atemporal y omnipresente.

Puerto de Valparaíso, Septiembre 2016.

# Abstract

In this thesis we propose a new map of science based on the productive paths of scholars. We call this network the *research space*. To create this map, we had to mine the web in order to curate a disambiguated dataset of individuals. With this dataset we were able to build the research space in which the nodes represent fields of research and the links represent how likely a scientist publishes in an area given that she published in other one. With the structure of the research space we evaluated the scientific production of individuals, institutions and countries. We also defined a methodology to evaluate the predictive power of maps of science in general and the research space in particular. With this methodology we found that the research space is a better predictor—in comparison with a map based on citation patterns—of the diversification and evolution of individuals and institutions. We also proved that our results hold when we change the classification of areas of science. Finally we built two applications, one for the scientific community and the other one for general public. These applications facilitate the analysis of the diversity of scientific production based on maps of science.

**Keywords:** maps of science, complex networks, scientometrics, scientific data, google scholar

# Resumen

En esta tesis proponemos un nuevo mapa de la ciencia basado en la trayectoria productiva de los investigadores, al que denominamos *Espacio Investigación* o *Research Space*. Para la creación de este mapa, debimos minar la web con la intención de construir un conjunto de datos desambigüado a nivel de individuos. Con este conjunto de datos, fuimos capaces de construir el espacio investigación en el que los nodos representan áreas de la ciencia y los enlaces representan cuán probable es que un científico publique en una determinada área de la ciencia, dado que ha publicado en otra área. Con la estructura del espacio investigación evaluamos la producción científica de individuos, instituciones y países. También definimos una metodología para la evaluación del poder predictivo de mapas de la ciencia en general y del espacio investigación en particular. Con esta metodología encontramos que el espacio investigación es un mejor descriptor —en comparación con un mapa basado en patrones de citación— de la diversificación y la evolución de individuos e instituciones a lo largo de áreas de la ciencia. También comprobamos que nuestros resultados son robustos a la hora de cambiar de clasificación de áreas de la ciencia. Finalmente construimos dos aplicaciones, una para la comunidad científica y otra para el público general. Estas aplicaciones tienen como objetivo facilitar el análisis de la diversidad de la producción científica basados en mapas de la ciencia.

**Palabras clave:** mapas de la ciencia, redes complejas, scienciometría, datos científicos, google scholar

# Índice general

# Índice de figuras

# Índice de tablas

Capítulo 1

# Introducción: De mapas y ciencia

La cartografía de la ciencia se ha venido practicando desde tiempos ancestrales. Visualizar lo que conocemos nos ayuda a entender dónde estamos ubicados en el universo del conocimiento y a dónde queremos —podemos— ir. En las últimos décadas, gracias al avance de la computación y los métodos de análisis de datos, se han realizado importantes contribuciones en esta área, principalmente en la creación automática de mapas de la ciencia, los mismos que están basados en grandes volúmenes de datos respecto de publicaciones científicas así como del proceso que las produce, la investigación. En este capítulo realizamos una introducción a los mapas de la ciencia desde sus orígenes ancestrales hasta los últimos desarrollos en esta área, para identificar con precisión cuál será el foco de nuestro estudio.

## 1.1. Motivación

La ausencia de conocimiento nos encadena a un mundo de tinieblas donde solo se perciben —y aceptan como reales— las sombras de la naturaleza, como lo explica Platón en su alegoría de la Caverna. Adicionalmente y con fervientes ejemplos a lo largo de la historia, podemos afirmar que la ignorancia crece salvajemente sobre el temor, mientras que el conocimiento se cultiva diligentemente sobre la fértil libertad. El uso del temor ha sido recurrente en nuestra historia como herramienta de protección —para la sobrevivencia— o de subyugación —para la conquista o la dominiación—.

El temor se avivó en los mapas de la tierra en la antigüedad. Esos mapas de la tierra incluían monstruosas serpientes marinas y temibles dragones en aquellos lugares aún desconocidos o no explorados por el hombre (ver la Carta Marina en la Figura 1.1). Entre los marineros y aventureros era conocida la frase *hic sunt dracones* (aquí hay dragones) (ver Figura 1.2). Claro, el temor, es parte de nuestro natural mecanismo de autodefensa en la búsqueda de la sobrevivencia, pero si nos invade —o nos es inculcado— más allá de ese instinto natural, entonces nos limita, nos coarta, nos impide conocer, nos aviva la idea de dragones en lugar de alentarnos a explorar la tierra indómita.


**Figura 1.1:** Carta marina. Dibujado por Olaus Magnus en el siglo XV. Nótese los dragones.

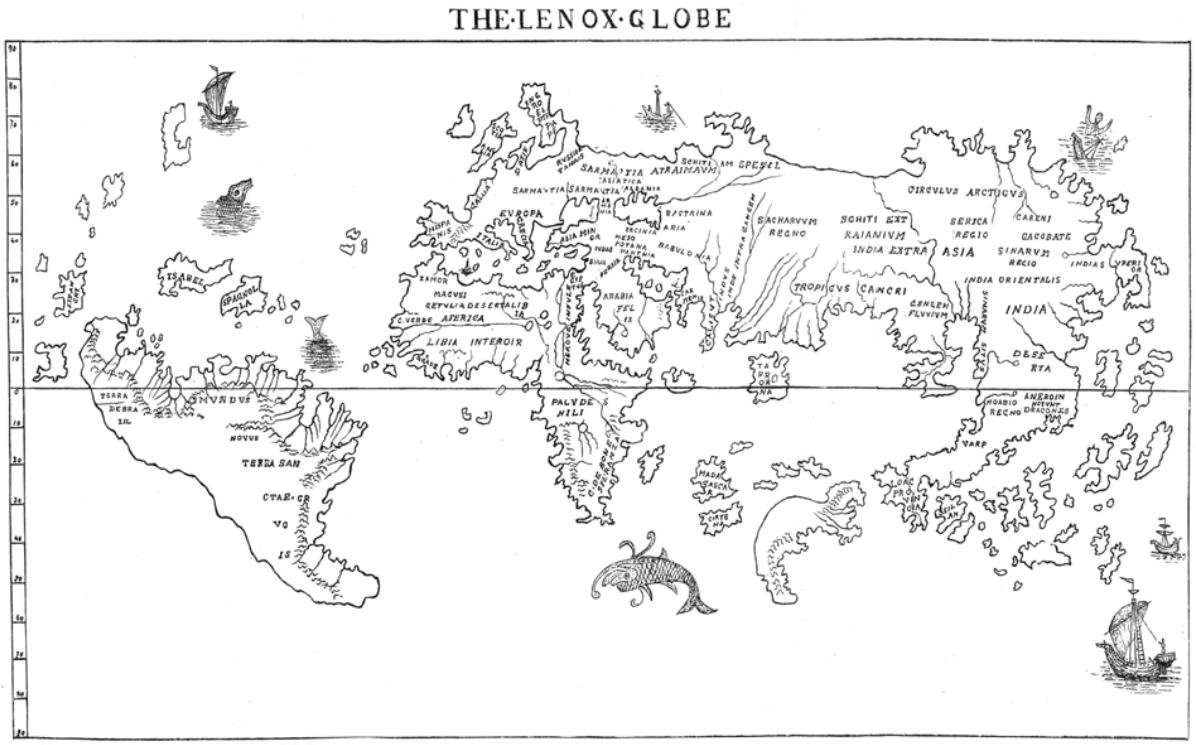



**Figura 1.2:** Mapa de Lenox que data del año 1510. Este mapa es famoso por ser el primer y único mapa de la tierra en incluir la frase "Aquí hay dragones" (hic sunt dracones) de forma explícita. Ver sudeste asiático.

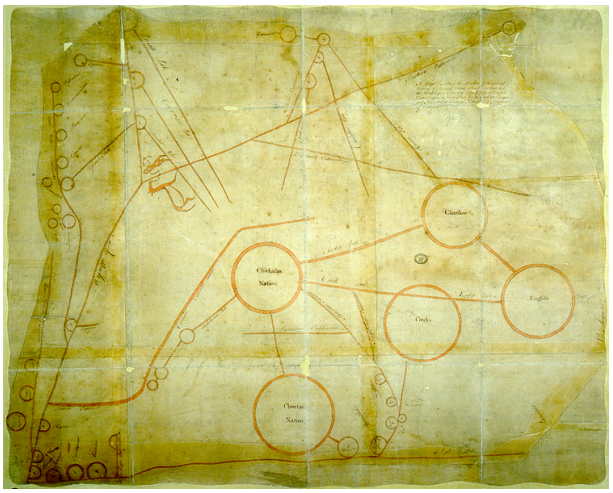
Esta metáfora motiva el trabajo que en este medio queda plasmado. Queremos aportar, desde la informática, con un nuevo mapa, no de la tierra, sino de la ciencia, pero que también cumpla el propósito de facilitar la exploración, a los principales navegantes de la ciencia: los artesanos del conocimiento y los hacedores de política institucional o pública. Se trata de ayudar a encontrar y despejar sus propios dragones, vale decir, planificar el viaje a la tierra indómita de la ciencia a la que aún no han arribado. En un contexto más general, podemos afirmar que la utilidad de los *mapas*, entendidos como aquellos que proveen una imagen completa de interrelaciones entre elementos, se ha puesto de manifiesto —de forma interdisciplinaria— en nuestra era digital, principalmente por la sobreabundancia de datos y el avance de las técnicas de análisis y visualización de datos. Un ejercicio rápido en la plataforma de búsqueda de documentos científicos, Google Scholar, nos indica que documentos con la palabra inglesa *mapping* se ha incrementado de 497K en el año 2000 a 801K en 2010, casi el doble de documentos. Esta es una breve muestra de que en el último tiempo, crece la tendencia de *mapear* fenómenos complejos. Una muestra del crecimiento de nuestra necesidad de mapear, ha sido recogida por Katy Börner de la Universidad de Indiana en Estados Unidos, quien ha publicado recientemente, dos inspiradores libros [Börner, 2010, 2015]. El primero recoge la historia y evolución de los Mapas de la Ciencia, mientras que el segundo, selecciona mapas de propósito general aplicados a distintos dominios del diario vivir. Este libro consigna en la portada la frase *cualquiera puede mapear*. En esta tesis proponemos un nuevo mapa de la ciencia basado en las capacidades productivas de los investigadores, al que denominamos *Research Space* (Espacio Investigación). Para la construcción de este mapa, debimos minar la web con la intención de construir un conjunto de datos desambigüado a nivel de individuos. Con este conjunto de datos, fuimos capaces de proponer una nueva técnica para la construcción de mapas de la ciencia, basados en datos de producción. También definimos una metodología para la evaluación del poder predictivo de mapas de la ciencia en general y avanzamos técnicas de visualización de mapas de la ciencia que no se habían explorado previamente. Finalmente dejamos sentadas las bases de una futura aplicación de uso masivo donde la utilización del Research Space propicie un verdadero impacto en la sociedad.

## 1.2. Mapas como redes

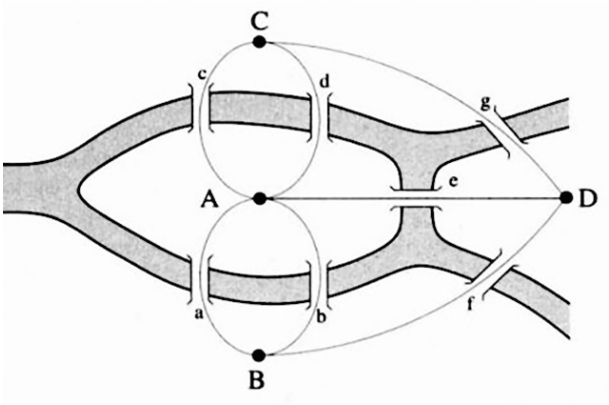
La necesidad de mayor expresión y significado, respecto de los mapas geográficos, produjo el aparecimiento de nuevos mapas que, siendo más abstractos, entregaran una visión esquemática de los actores y sus relaciones, cualquiera que estas fueran. Este fue el caso, por ejemplo, del pueblo nativo que habitó en América del Norte, los *Chickasaw*, quienes se vieron en la obligación de desplegar sus dotes diplomáticos para formar alianzas con los colonos ingleses además de estar a salvo de tribus vecinas y enemigas como los *Choctaw*, una nación superior en número de personas y además aliada a los colonos franceses [Clark, 2005, p. 132-133]. Entre 1723 y 1724, un cacique indio de los *Chickasaw* presentó a sus aliados ingleses un *mapa* que lejos de ser geográficamente correcto, sí proveía información relevante respecto de los diferentes autores involucrados en el conflicto —representados por los círculos—, así como de las relaciones que mantenían —representados por las líneas o enlaces—, tales como vías de conexión, relaciones de comercialización, o conexiones hostiles. Si bien, el mapa original

se perdió, una copia fue realizada por un colono inglés de la época, la misma que se presenta en la Figura 1.3a.

En el mismo siglo XVIII, aproximadamente diez años más tarde, en 1736 en San Petersburgo, Leonard Euler construía un mapa esquemático que también procuraba abstraerse de la geografía para analizar la posibilidad de visitar 4 espacios de tierra conectados por 7 puentes, iniciando y terminando el viaje en el mismo punto. Esta situación geográfica-urbanística, ocurría en la cercana localidad de Königsberg. Como afirma Barabási [2003, p. 11], en ese momento nace la *teoría de grafos*, puesto que Euler define matemáticamente las propiedades de los elementos presentes en su mapa (*grafo*), compuesto por *nodos* y *enlaces*. En el caso de los puentes de Königsberg, se trata de un grafo de 4 nodos y 7 enlaces (ver Figura 1.3b) y Euler logra determinar que mientras los nodos tengan un *grado* —cantidad de enlaces que conectan el nodo— impar, el problema no se puede resolver. Nótese que los nodos $B, C, D$ tienen grado 3, mientras que el nodo $A$ tiene grado 5.



**(a)** Mapa creado por el pueblo Chickasaw que habitó en lo que hoy se conoce como New Orleans, USA. El mapa data de 1722 ó 1723 y sirvió para esquematizar las relaciones entre naciones y actores de la época.




**(b)** Mapa de los puentes de Königsberg, creado por Leonard Euler en 1736, quien analiza el problema de visitar cada espacio de tierra sin atravesar por el mismo puente dos veces y además terminar en el mismo punto de partida.

**Figura 1.3:** Mapas que evolucionan de la geografía a los grafos.

Estos dos casos, ejemplifican también la distinción que realiza Hidalgo [2016] cuando analiza el desarrollo del área científica de las redes. Por un lado, la de comunidades —a futuro estudiadas por cientistas sociales—, de crear redes con enlaces provistos de contexto, mientras que por otro lado, la intención de cientistas naturales de crear redes como modelos generalizables, sin profundizar en la expresividad o en el por qué de esos enlaces, en última en el contexto.

En particular, el ámbito de los cientistas naturales, durante el siglo XIX e inicios del siglo XX, la teoría de grafos, había sido ampliamente explotada en sus aristas matemáticas más profundas, pero se encontraba limitada a contextos abstractos, propios de la matemática, y no se lograba apreciar su real aplicación al estudio de la naturaleza o la sociedad. Esta

teoría cobra un nuevo impulso a principios del siglo XX de la mano del matemático Paul Erdös, y a fines del siglo XX y principios del siglo XXI de los físicos Newman, Barabási y Vespignani [Barabási, 2003, 2016; Barrat et al., 2012; Newman, 2010; Newman et al., 2006], entre muchos otros, quienes desde la comunidad científica de los Sistemas Complejos, y gracias al avance de métodos computacionales y estadísticos modernos, popularizan el área de las *redes complejas*, como una disciplina que va más allá de los grafos y está orientada a analizar fenómenos complejos de problemas reales. La idea de conectividad, ha ayudado a la difusión del término red, así encontramos cuantiosos ejemplos de redes, como por ejemplo: red social, red de comunicaciones, red de información, red de citas, red de influencia, redes de poder, red de coautoría, red de productos, entre muchas otras.

En esta tesis, adheriremos a la distinción de nomenclatura propuesta por Kolaczyk [2010, p. 2] y utilizaremos principalmente el término *red* para la estructura macro en estudio, mientras que el término *grafo*, solamente para la representación matemática de esa red.

## 1.3. El ecosistema de la ciencia y cómo se mide

En el ecosistema de la ciencia, hasta que los experimentos, resultados y hallazgos científicos no se plasman en algún *medio* y se comunican, en la práctica, no existen y en consecuencia los creadores de éstos, los científicos o artesanos del conocimiento, también carecen de existencia en el mundo de la investigación. En esta sección analizamos los principales elementos que han permitido la diseminación y estudio de la ciencia.

### 1.3.1. El medio

El *medio* por excelencia para comunicar ciencia, es el escrito (*paper*) y cobran mayor importancia aquellos que son previamente validados por pares del área disciplinar (*peer review*). Los *papers* se aglutinan en revistas científicas (*journals*), actas de conferencias (*proceedings*), entre otros entes en los que sucede el devenir del conocimiento nuevo.

El aparecimiento de la prensa de imprimir o imprenta moderna de la mano de Johannes Gutenberg a fines del siglo XV, facilitó en el siglo XVII la publicación de los primeros *journals* en el mundo occidental, hecho ocurrido en 1665. Desde entonces la humanidad ha venido utilizando este medio y procedimiento, como principal artefacto de difusión científica. En Francia se trató del *Journal des Sçavans* (que más tarde se llamó *Journal des savants*) publicado por el escritor y abogado, Denis de Sallo, mientras que meses más tarde la *Royal Society of London* publicó el *journal Philosophical Transactions* (Figura 1.4), por encargo del rey de la época. Mientras que el *journal* francés estaba orientado a las humanidades, el *journal* inglés estaba orientado a las ciencias naturales. Después de 350 años, la comunicación escrita, se ha mantenido como el medio por excelencia para dar a conocer los hallazgos científicos. Aunque en la actualidad esta comunicación se pueda acceder también en formato digital y además pueda contar con material suplementario tipo multimedia, la comunicación escrita —*paper*—, sigue siendo el principal medio para comunicar ciencia.

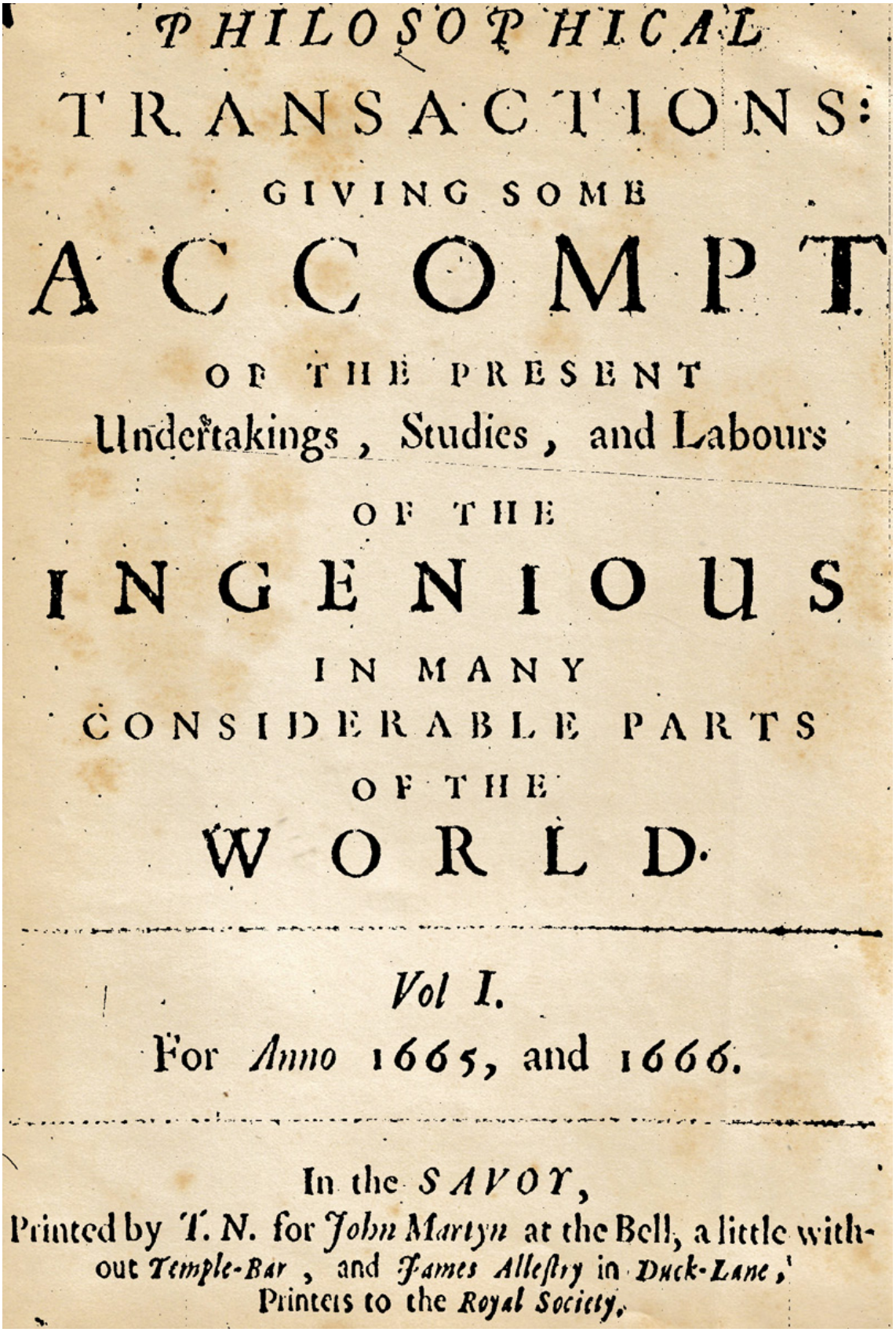

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane, Printers to the Royal Society.

**Figura 1.4:** Primera número de la revista Philosophical Transactions. 1665.

**Tabla 1.1:** Características de las principales bases de datos bibliográficas globales.

| B. Datos | Cobertura | Acceso | Publicador | Desde | *journals* (miles) | Conferencias (miles) |
|---|---|---|---|---|---|---|
| WoS | mundial | suscripción | Thomson Reuters | 1900 | 12 | 160 |
| Scopus | mundial | suscripción | Elsevier | 1999 | 21 | 6100 |
| SCImago | mundial | público | SCImagolab | 1999 | 21 | 6100 |
| SciELO | regional | público | Consorcios locales | 1997 | 1 | - |

### 1.3.2. La indexación y recuperación de información

Aunque la informática moderna pudo haber surgido en la Inglaterra victoriana (siglo XVIII) de la mano de Charles Babbage y Ada Lovelace, no es hasta el siglo XX cuando surge realmente y nos entrega la posibilidad de registrar datos de forma sistemática. Datos por ejemplo, de publicaciones científicas. Es así que se suman al ecosistema, bases de datos automáticas que cualifican journals, proceedings y otros, para elegir aquellos que cumplen con ciertos estándares de calidad y los indexan —almacenan, ordenan, clasifican— para hacerlos más accesibles, de forma pública o privada, para los consumidores del nuevo conocimiento. Junto con proveer facilidades de búsqueda, estas bases de datos también se encargan de mantener métricas de uso y relaciones entre *papers*, autores, journlas, instituciones y países, con el ánimo de facilitar la polinización de la ciencia y —en no pocos casos— lucrar de un rentable negocio, la venta de información.

Las bases de datos que indexan información de publicaciones científicas, se detallan a continuación, clasificándolas según su propósito.

#### 1.3.2.1. Globales

Estas fuentes de información disponen de datos de publicaciones académicas en todas las áreas de las ciencias, las ciencias sociales, las humanidades y las artes. En la Tabla 1.1 se presentan las principales características de las bases de datos más utilizadas. En general la comunidad científica, los países —como Chile— y las instituciones le han venido entregando mayor calidad a las revistas indexadas por Thomson Reuters™ en el servicio Web of Science™ (WoS). En la actualidad el *core collection* de este servicio incluye tres índices (bases de datos), que abarcan las ciencias, las ciencias sociales y las artes y humanidades. Estos índices son *Science Index Expanded™ (SCIE)*, *Social Science Index® (SSCI)* y *Arts & Humanities Citation Index® (AHCI).*

De forma muy cercana a WoS, se encuentra su principal competidor —también de pago— Scopus® de la editorial Elsevier. Asociado a Scopus® se encuentra el sitio público de difusión de rankings SCImago Journal & Country Ranking (SJR). En términos de bases de datos regionales SciELO es la base de datos de mayor difusión en Iberomérica, aunque hay quienes la critican por tener un importante sesgo hacia las revistas brasileñas, por ser ahí donde nació la idea original [EC3Metrics, 2014]. En otro ámbito, SciELO destaca por haber sido pionera en el contenido abierto de las publicaciones indexadas [Packer et al., 2014].

Desde el año 2014, la empresa Thomson Reuters™, también oferta el servicio del índice

SciELO Citation Index como medida para mejorar la cobertura de revistas regionales y en otros idiomas distintos al inglés, característica en tiene desventaja en comparación a su competidor Scopus [EC3Metrics, 2014].

Los datos de cobertura para confeccionar la Tabla 1.1 se obtuvieron de los respectivos sitios de WoS [Thomson Reuters, 2015] y de Scopus® [Elsevier, 2015], mientras que para SciELO del análisis comparativo de Falagas et al. [2008].

#### 1.3.2.2. Disciplinares

En un contexto más local, se pueden encontrar bases de datos de publicaciones orientadas a disciplinas o áreas específicas. Por no constituir el foco de este estudio, mencionamos brevemente las más conocidas en un afán de describir el ecosistema completo de la ciencia.

Para ciencias de la computación existe la base de datos Computer Science Bibliography (DBLP), ampliamente difundida por ser de libre acceso y disponibilizar de forma gratuita la información. En las ciencias biomédicas, lidera la indexación la base de datos Public domain information on the National Library of Medicine (PubMed), mientras que en las ciencias sociales, la base de datos de mayor uso es la Social Science Research Network (SSRN).

#### 1.3.2.3. Nacionales

Las bases de datos de producción científica a nivel nacional, usualmente están a cargo de las agencias de estado encargadas de gestionar los recursos y dinámica de la Investigación de forma local. Aunque no se conocen ejemplos de datos nacionales liberados por países, sí se pueden encontrar aplicaciones creadas *ad hoc* que entregan funcionalidades de análisis. Por ejemplo el *redsearchExplora* [1] de la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT). En esta aplicación se utiliza la base de datos de la institución que resume la información de publicaciones y proyectos financiados por el estado de Chile.

Otros datos disponibles, constituyen los reportes de productividad o de impacto que crean las naciones para analizar la política pública a futuro en esta materia. El Reino Unido, es un buen ejemplo en esta categoría. Este país, desde 1986, revisa y evalúa la productividad de sus universidades respecto de proyectos financiados por el estado [Sweeney, 2015]. En la última entrega se utilizó un método distinto a los tradicionales —frecuencia de *papers*—, para elaborar el informe denominado Research Excellence Framework REF [HEFCE, 2014] cuyos datos tabulados se disponibilizan para consulta en un sitio web [2]. Este conjunto de datos ha servido, por ejemplo, de base para sucesivos reportes, los mismos que también liberan datos agregados de producción científica de las instituciones nacionales del Reino Unido.

#### 1.3.2.4. Institucionales

Varias instituciones, principalmente universidades, se han visto en la necesidad de construir o implementar sistemas de información que les permitan recopilar, ordenar y disponibilizar sus propios registros de producción y financiamiento científico. Varias instituciones

[1] http://www.redsearch.cl/

[2] http://impact.ref.ac.uk/CaseStudies/search1.aspx

han adherido al software de código abierto VIVO [3], para realizar esta tarea. Aprovechando la instancia de haber curado su propio conjunto de datos para ingresarlo al sistema VIVO, varias instituciones han liberado públicamente ejemplos de sus datos en formato RDF, con la intención de permitir el benchmark y el análisis científico [4]. A pesar de lo anterior, aún no existe una base de datos pública o privada que entregue indicadores detallados de producción a nivel de instituciones de forma global. Solo son accesibles indicadores pre-calculados que se utilizan en rankings, como el ranking de Shangai [5], el THE (Times Higger Educational) ranking [6] o el mismo ranking de instituciones de SCImago [7].

#### 1.3.2.5. Individuales

Si bien, desde los años ochenta, las bases de datos a nivel de *papers* y *journals* lideraron el proceso de acceso a la información científica; a partir del año 2000, se han realizado importantes esfuerzos en construir bases de datos que incorporen información a nivel de usuario. A nivel de individuos, aún se encuentran desafíos pendientes, entre ellos el principal dice relación con la desambiguación de la identidad (única) de un autor, lo que conlleva a curar información de movilidad (geográfica y laboral), cambio de nombres y otros aspectos relacionados a la identidad. Recientemente, un breve análisis de los distintos tipos de bases de datos orientadas a autores, fue difundido por Lambert [2015]. Este estudio analiza en detalle las diferentes características que tienen —o se espera que tengan— este tipo de servicios. Entre las principales están aspectos relacionados a la identidad, la movilidad laboral, las conexiones sociales, citas ganadas y facilidad de acceso a los datos.

Los principales servicios con información de investigadores son Research Gate [Research Gate, 2015], Google Scholar [Google, 2015] y ORCID [ORCID, 2012].

Research Gate es considerado el "Facebook" de la investigación, por cuanto sus funcionalidades están pensadas principalmente en facilitar y acelerar la interacción entre investigadores. Research Gate además procura propiciar la discusión de temas científicos y la diseminación de trabajos entre autores. La principal crítica a este servicio se refiere a la velocidad en indexación de nueva información.

Google Scholar se muestra como una opción sólida, principalmente por el poderoso y ampliamente difundido motor de búsqueda que está detrás. En este servicio, los mismos autores deben realizar el proceso de seleccionar sus publicaciones de un conjunto de publicaciones posibles que Google detecta en el ciberespacio como publicaciones candidatas del autor o autora. Además Google Scholar mantiene información de citas e indicadores como el índice-H así como también incluye información de coautoría. Las principales críticas a este sistema son la incorporación de información que no ha sido validada previamente por la comunidad científica, como por ejemplo, archivos con diapositivas o documentos informales, lo que tiende a hacer que la cantidad de citas se incremente fuertemente. A pesar de esto, se ha logrado determinar que su nivel de citaciones está correlacionado con bases de datos más tradicionales

[3] http://vivoweb.org/
[4] http://datahub.io/dataset/vivo
[5] http://www.shanghairanking.com/
[6] https://www.timeshighereducation.com/world-university-rankings
[7] http://www.scimagoir.com/

[Falagas et al., 2008]. También se le objeta su carácter público, que hace que se puedan crear cuentas falsas o espurias así como la posibilidad de adulterar el número de citas con códigos automáticos [Labbé, 2010].

ORCID ha surgido como una iniciativa de varias instituciones, con la intención de poner fin a la ambigüedad de identidades (nombres) de manera transversal. Los usuarios pueden registrarse para obtener un identificador único y agregar a éste sus publicaciones, recorrido laboral, alias y proyectos ganados. Este servicio está siendo incorporado por varios *journals* que dan la posibilidad de indicar el (ORC)ID del autor para hacer el link unívoco con su identificación en el mundo de la investigación. También ofrecen la posibilidad de gestionar directamente con las instituciones el registro de sus académicos.

Debe notarse que estas bases de datos cobran especial relevancia para nuestro trabajo, por cuanto nos hemos visto en la necesidad de minar la web en la intención de construir un conjunto de datos de individuos e instituciones que nos permita abordar nuestra propuesta a cabalidad.

## 1.3.3. Clasificación de la ciencia

Con la intención de hacer más expedito el proceso de recuperación de información y más eficiente la gestión de los procesos asociados a la ciencia, es que, los *papers* y las revistas en las que se publican, se clasifican en distintos campos o categorías de la ciencia. Estas clasificaciones se elaboran de forma automática o en base a comités de expertos quienes atienden el llamado de alguna institución.

### 1.3.3.1. Clasificaciones de contenido

Este tipo de clasificaciones, basan su desarrollo en las discusiones temáticas de sub comités y generalmente se publican en informes técnicos que describen en detalle cada una de las categorías.

Por ejemplo, los denominados Field of Science (FOS) de la clasificación de la Organization for Economic Co-operation and Development (OECD), se desarrollaron en el contexto del manual de Frascati [OECD, 2002] cuya última revisión es del año 2002. Esta clasificación está dirigida a clasificar la Investigación y el Desarrollo en el sector público y su aplicación se ha extendido más allá de los países miembros de la OECD. Los FOS propuestos por el comité de expertos de la OECD, se incluyen como parte del manual Frascati y fueron actualizados por última vez en el año 2007 [OECD, 2007]. Esta clasificación jerárquica, incluye 6 áreas principales y 39 FOS.

Por otro lado, la clasificación *Australian-New Zeland Fields of Research* surge como respuesta a la necesidad de representar la *actividad* de hacer ciencia, esto es la investigación, y no las unidades administrativas que gestionan la ciencia [Pink and Bascand, 2008]. La última clasificación disponible es la de 2008. Es una clasificación de organización y notación jerárquica, posiblemente inspirada en las clasificaciones elaboradas para otros contextos, como el económico (por ejemplo, el Standard International Trade Classification (SITC)), puesto que aplica una lógica de niveles de acuerdo a pares de dígitos en los códigos. En total son 6 dígitos y cada dos dígitos representa uno de los tres niveles, denominados Division (2 dígitos),

Group (4 dígitos) y Field (6 dígitos). El mismo sitio del Instituto de Estadística de Australia, entrega tablas de vinculación, con otras clasificaciones, como la clasificación de la OECD.

#### 1.3.3.2. Clasificaciones semiautomáticas

Las empresas indexadoras de información científica, también han generado sus propias clasificaciones, en el contexto de hacer la recuperación de información más expedita. Estas clasificaciones se han generado de manera semiautomática, generalmente utilizando un método propio basado en patrones de citas y evaluado-revisado posteriormente de forma manual por expertos o por un comité editorial. Son clasificaciones semiautomáticas, la de WoS y la de Scopus®.

Por parte de Thomson Reuters™ se definen 251 *subjects* o categorías de la ciencia, que se pueden agrupar en 22 áreas. Esta clasificación no incluye una categoría Multidisciplinaria.

Por otro lado, Scopus® y su socio SCImago definen 310 categorías agrupadas en 27 áreas principales. Nótese que esta clasificación sí incluye una categoría Multidisciplinaria donde se asignan los *journals* que tienen esta característica, como *PNAS* o *Science*. Esta clasificación también permite que un *journal* se encuentre indexado en más de una categoría.

### 1.3.4. Las ciencias que estudian la ciencia

Junto con la información científica comunicada en cada *paper*, también van quedando huellas de datos (metadata) que es posible estudiar con el apoyo de métodos científicos, aparece entonces la ciencia de la ciencia o *Cienciometría*, cuyos orígenes datan de 1928, y su mayor aporte fundacional moderno se atribuye a Eugene Garfield, creador de los primeros índices de producción científica; y, Derek John de Solla Price, creador de la primera *red* de citas bibliográficas [Börner, 2010, p.52]. Disciplinas asociadas y relacionadas al esfuerzo de estudiar la ciencia con métodos científicos son la Bibliometría, la Bibliotecología y las Ciencias de la Información.

Los *papers*, al igual que las obras de arte y en última instancia, las personas, son únicos e irrepetibles y para construirlos, se hace necesario aunar esfuerzos colaborativos entre quienes han sabido conformar equipos que comunican sus hallazgos a través de estos vehículos de divulgación científica. Esta colaboración también se puede extrapolar a nivel de instituciones y países, en base a la filiación y procedencia de los autores.

#### 1.3.4.1. El aporte de la informática

El área 'nueva', de la Informática, es ciertamente de gran apoyo a la tarea moderna de estudiar la ciencia y tiene mucho que facilitar, aportar y decir en el ecosistema de la ciencia. Aplicaciones prácticas, en forma de herramientas para *plasmar* el trabajo de los artesanos de la ciencia, fueron propuestas por dos próceres de la computación: Donald Knuth (Turing Award 1974) creador de TEX y Leslie Lamport (Turing Award 2013) creador de LATEX y BibTEX. De igual forma, el otrora programador de la European Organization for Nuclear Research (CERN), Tim Berners-Lee, se propuso el desafío de mejorar la difusión de los documentos científicos que se producían en ese centro, lo que finalmente derivó en la creación de la World Wide Web (WWW). Apoyar la ciencia desde la informática, ha permitido también

expandir las fronteras del conocimiento de la informática. En un ejemplo reciente, al ganador del Turing Award 2014, Michael Stonebraker [CSAIL, 2015], se le reconoce entre otras cosas, sus aportes a la creación de bases de datos modernas *ad hoc* para albergar datos científicos.

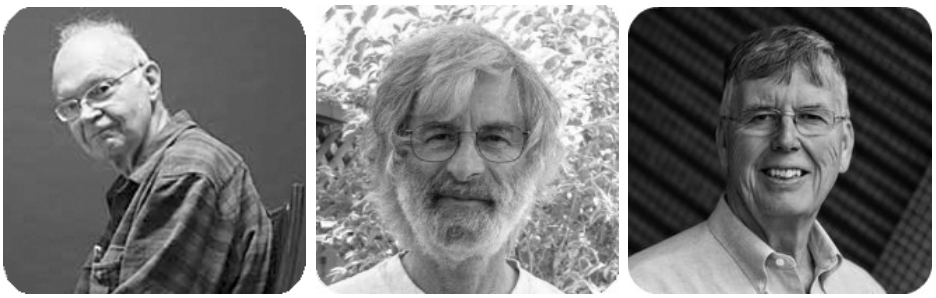


**Figura 1.5:** De izquierda a derecha, los ganadores del Turing Award: Donald Knuth (1974), Leslie Lamport (2013) y Michael Stonebraker (2014). Fuente: Sitio Turing Award

Es así como en la actualidad la informática continúa apoyando los esfuerzos de organización de los grandes volúmenes de datos que se pueden adquirir de las publicaciones científicas, así como también a su clasificación, detección de patrones, estudio de dinámicas evolutivas, medición y representación del conocimiento. Esta tesis, se enmarca en estos esfuerzos y aporta al crecimiento de la disciplina de la informática, con miras a impactar y beneficiar directamente la Cienciometría, que a su vez permita impactar en las políticas públicas y el desarrollo.

### 1.3.5. Medidas de desempeño científico

Entre las medidas de productividad y calidad científica propuestas en el ámbito cienciométrico, podemos encontrar por ejemplo, mediciones del impacto de *papers*, las mismas que procuran encontrar cómo el contenido de un artículo en particular, influencia una comunidad científica. Para journals, son ampliamente utilizadas, medidas como el número de citas o el *Factor de Impacto* propuesto por Garfield [1955].

Por otro lado, y a un nivel de granularidad distinto, también se procura develar la influencia o importancia de un determinado investigador en la comunidad científica. Medidas como el número de citas, también son aplicables a este nivel, aunque aquí surgen otros indicadores como por ejemplo, el *índice-H* propuesto por Hirsch [2005] o el índice de centralidad de un autor al interior de una comunidad científica, como el indicador *eigenfactor*$^{TM}$ [West et al., 2012].

En el último tiempo, otras métricas no basadas en citación han aparecido como oportunidades para evaluar otras aristas de impacto del trabajo científico. Tal vez la más publicitada es aquella desprendida de la difusión o atención que un trabajo científico ha tenido a través de la WWW, este índice es conocido como *altmetric* y mide la difusión en sitios de Noticias, redes sociales, entradas de Blogs, así como en conversaciones en redes sociales en línea como Twitter o Facebook [Altmetric, 2015]. Este índice también publica anualmente un ranking de los *papers* que han tenido mayor cobertura.

## 1.4. La ciencia y sus redes

Una forma de abordar el estudio del mencionado ecosistema de la ciencia, es precisamente a través del análisis de la estructura y dinámica de redes que sintetizan y abstraen la información de este ecosistema que germina principalmente en la publicación de *papers*. El acto de publicar, permite la construcción de un conjunto de redes basadas en la información que los *papers* contienen. En esta sección clasificamos y ejemplificamos varias de las redes más estudiadas y difundidas en la comunidad científica.

En ciencia, se pueden distinguir claramente dos tipos de redes: *redes de colaboración* y *redes de información*. En los primeros tipos de redes, los elementos principales —los nodos— son los productores de ciencia, sean estos individuos, instituciones o países; mientras que en el segundo grupo de redes, son entidades informativas como *papers*, *journals*, categorías o áreas de la ciencia.

### 1.4.1. Redes de colaboración

El acto de ser coautor de un *paper*, permite asumir que hay una relación de colaboración implícita entre los autores del mismo. Esta suposición facilita la creación de redes de colaboración a diferentes niveles de agregación de los productores de ciencia, esto es: autores, instituciones y países.

#### 1.4.1.1. Red de coautoría

La *red social* de coautoría que genera un *paper*, ha sido ampliamente analizada en la literatura tanto como estructura social y colaborativa [Adams, 2012; The Royal Society, 2011] cuanto más como interfaz para la medición del desempeño e importancia de los científicos [Janssen et al., 2006; Newman, 2004]. Para ilustrar una red de co-autoría, en la Figura 1.6 presentamos una red tipo *ego network* para un autor específico. Cada nodo representa un autor y el grosor de los enlaces representa la cantidad de *papers* publicados en conjunto por dos autores. También se ha aplicado un algoritmo de detección de comunidades, que permite identificar —con el uso de colores— las distintas comunidades en las que participa el autor (ego).

#### 1.4.1.2. Red interinstitucional

Si los autores se agregan por la institución a la que pertenecen se pueden crear redes de coautoría entre instituciones, las mismas que se pueden interpretar como redes de colaboración entre instituciones. A nivel global existen escasos estudios y aplicaciones que aborden este tipo de redes, por cuanto no existe una base de datos disponible que facilite esta tarea. Ortega and Aguillo [2013] realizan un breve análisis de este tipo de redes, basados en una muestra de Google Scholar.

Este tipo de redes, son más comunes de encontrar a nivel local, por cuanto las agencias estatales tienen información más detallada de las instituciones de cada país. Como ejemplo,

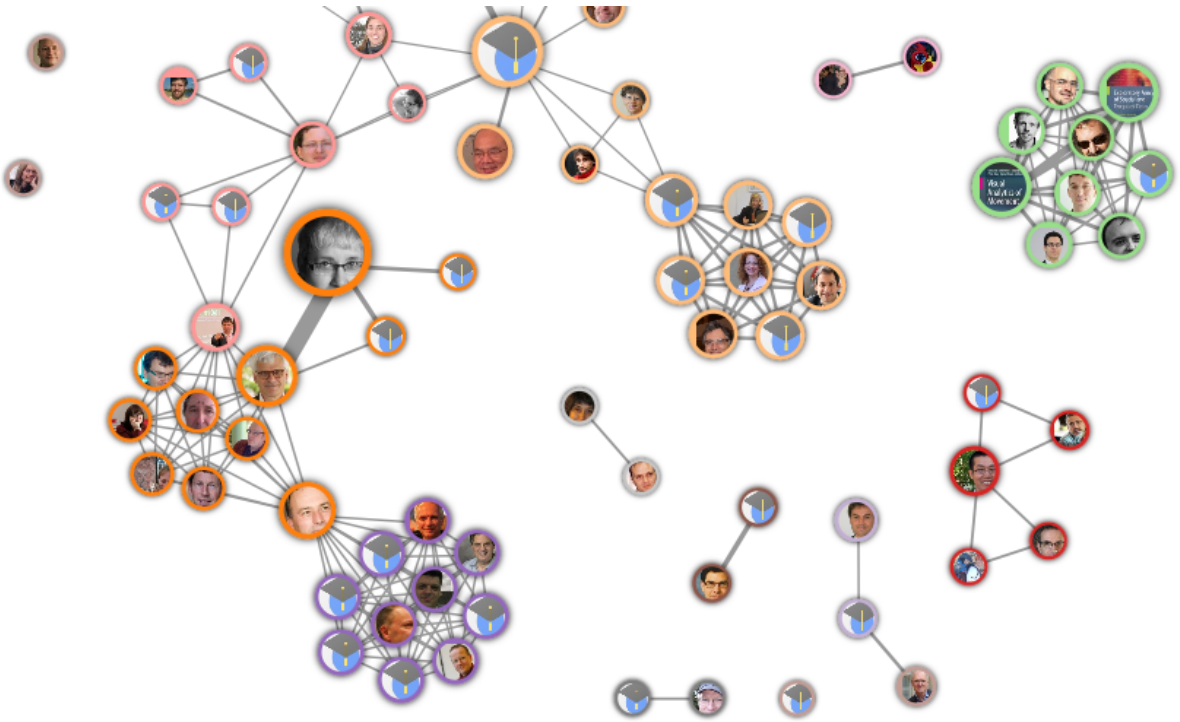
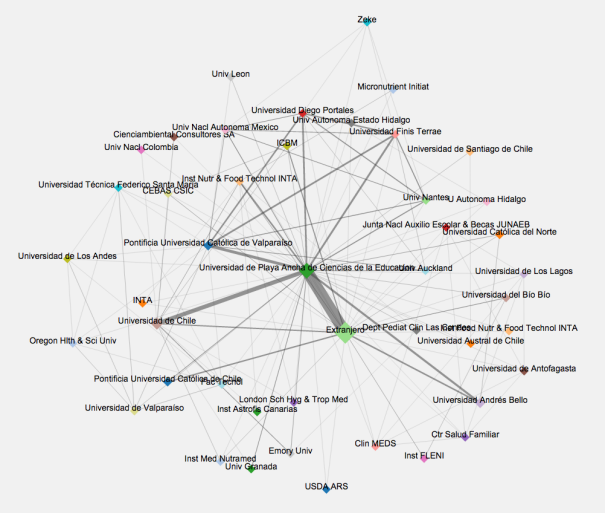
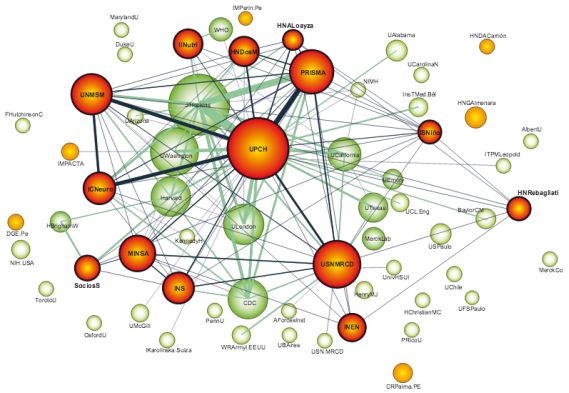


**Figura 1.6:** Ejemplo de red de autores. Los nodos representan autores y los enlaces representan la cantidad de artículos que han publicado en conjunto. Los colores indican las comunidades científicas detectadas automáticamente en base a coautoría.

utilizando la aplicación RedSearch[8] de CONICYT, en la Figura 1.7a hemos creado una red de colaboración entre instituciones que han publicado *papers* con autores afiliados a la Universidad de Playa Ancha (UPLA). Otro ejemplo a nivel local, son aquellos estudios bibliométricos acotados a un país y área disciplinar, como el realizado por Huamaní y Mayta-Tristán [Huamaní and Mayta-Tristán, 2010] que, entre otras cosas, analiza la colaboración de instituciones peruanas, tomando como fuente de datos los artículos indexados en la categoría *Clinical Medicine* de Thomson Reuters$^{TM}$ entre los años 2000 y 2009 (Figura 1.7b).




**(a)** Coautoría entre instituciones que colaboran con la UPLA en el período 2013-2014. Generado con RedSearch.cl de CONICYT.




**(b)** Colaboración nacional e internacional de instituciones peruanas en el período 2000-2009 de artículos registrados en el índice Clinical Medicine de Thomson Reuters$^{TM}$. Fuente [Huamaní and Mayta-Tristán, 2010].

**Figura 1.7:** Redes de colaboración entre instituciones. Los nodos representan instituciones y los enlaces representan 'colaboración' a través de coautorías entre instituciones.

### 1.4.1.3. Red de colaboración internacional

Si los coautores de un *paper*, se asignan a una institución y la institución al país, entonces es posible medir la colaboración entre instituciones y países, como en el caso de las redes de colaboración, incluidas en el informe de The Royal Society [2011]. La Figura 1.8 presenta una de las redes de colaboración global que se incluyen en este informe anual. En este tipo de red, se puede aprender, por ejemplo, qué países son los que actúan como clusters de colaboración científica, como este caso sucede con aquellos países etiquetados y que se encuentran hacia el centro de la red. Otros estudios también han sido propuestos en esta misma línea, como el realizado por [Leydesdorff et al., 2013b] en base al índice Science Index $^{TM}$(SCCI) de Thomson Reuters$^{TM}$ del año 2011.

[8] http://redsearch.cl/

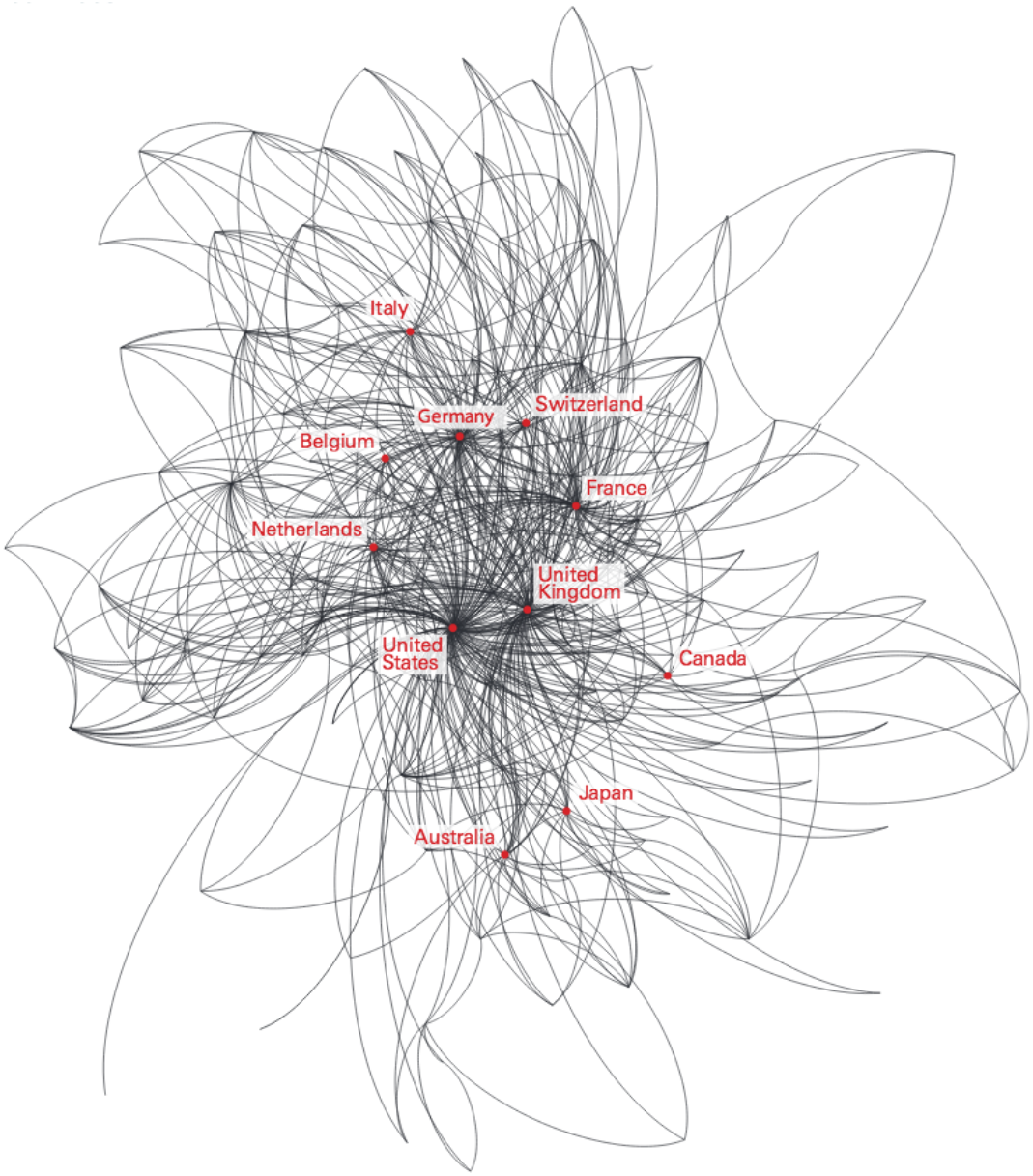



**Figura 1.8:** Ejemplo de una red de colaboración entre países para el período 2004 a 2008. Los nodos representan países y los enlaces, cantidad de *papers* con autores provenientes de dos países distintos. De esta red se puede aprender, por ejemplo, qué países actuan como clusters de colaboración. Fuente [The Royal Society, 2011, p. 51].

### 1.4.2. Redes de información

El acto de *citar* un *paper* desde otro, se constituye en el principal elemento para diferentes análisis, como por ejemplo, los indicadores que se mencionaron en la Sección 1.3. En términos de redes, se pueden crear *redes de información* a distintos niveles de agregación, esto es: *papers*, *journals*, categorías y áreas.

#### 1.4.2.1. Red de citación

A diferencia de una red de coautoría, la red de citación, es una red es dirigida, por cuanto se conoce de antemano qué *paper* cita (apunta) a cuál. En la Figura 1.9, se presenta un ejemplo de este tipo de red, en el que los nodos se han etiquetado con el nombre del primer autor del *paper* y además se incluye una escala temporal del año de publicación de los *papers*. El ejemplo está tomado del trabajo de Janssen et al. [2006]. En este tipo de red podemos identificar trabajos seminales en un área en particular, también aquellos trabajos que más citas han concitado en una comunidad de práctica. Nótese que esta red se puede extrapolar también a una red de citación entre autores, lo que la transformaría en una red de personas con intereses comunes.

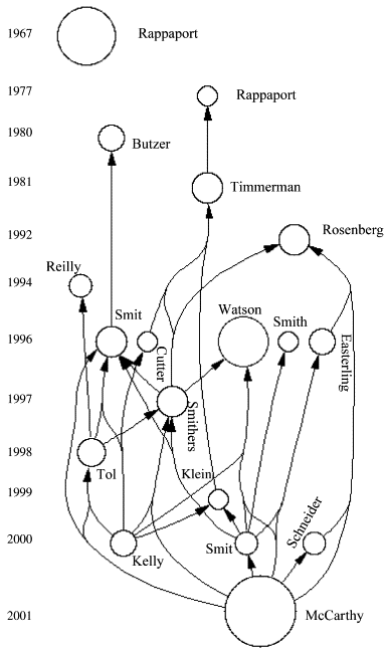



**Figura 1.9:** Ejemplo de red de citas entre *papers*. Los nodos representan *papers* que se han etiquetado con el primer autor. Los enlaces representan las citas de un *paper* a otro. El tamaño de los nodos es proporcional a la cantidad de citas recibidas por un *paper*. Fuente [Janssen et al., 2006].

El servicio WoS de Thomson Reuters$^{TM}$ también incluye una funcionalidad para visualizar las redes de citación entre *papers*. Sin embargo esta aplicación es muy poco eficiente y en la mayoría de los casos, la operación no se puede concluir por exceso de tiempo.

### 1.4.2.2. Red de journals

Si los *papers* son agregados en *journals*, podemos obtener una red de relaciones entre *journals*. En este tipo de redes, los nodos constituyen *journals* y los enlaces alguna medida de relación, generalmente inter-citación. Nótese que el termino *inter-citación*, es distinto del término *cita directa*, puesto que el primer método agrega citas entre *journals* en los dos sentidos y descarta la direccionalidad de las citas. Esto es, se calcula agregando citas entre *journals* en las dos direcciones y se omite la direccionalidad. La empresa Thomson Reuters™ a través del servicio Journal Citations Report® (JCR), en su versión en línea, provee de redes de *journals*, en las que el tamaño de los nodos es proporcional al Factor de Impacto (ver Glosario) y los enlaces representan intercitación entre *journals*. El ejemplo de la Figura 1.10 muestra los *journals* indexados en la categoría *Computer Science, Information Systems*. Esta red es útil para encontrar fuertes relaciones de intercitas entre *journals*. En la aplicación web, se pueden seleccionar los nodos para mejorar la visualización de los enlaces.

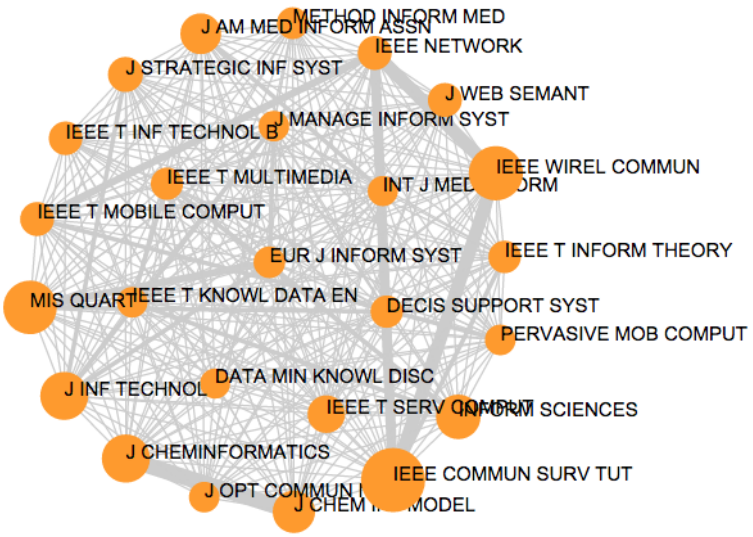



**Figura 1.10:** Ejemplo de una red de intercitación entre *journals*. Imagen creada utilizando la aplicación web del Journal Citations Report® (JCR) de la empresa Thomson Reuters™ para el año 2014. *Journals* incluidos en la categoría *Computer Science, Information Systems*. El tamaño de los nodos es proporcional al Factor de Impacto del *journal* para el año 2014.

### 1.4.2.3. Red de categorías y áreas de la ciencia

Si los *papers* o los *journals* son agregados en categorías y/o áreas, según cierta clasificación manual o automática, también podemos obtener una red de categorías o áreas de la ciencia. Al igual que la red anterior, en la Figura 1.11 se presenta un ejemplo de este tipo de red, que incluye las 22 áreas en las que la empresa Thomson Reuters™ clasifica los *journals* y por ende la ciencia.

Las redes incluidas en esta subcategoría, son las que denominaremos Mapas de la Ciencia y se describen a profundidad en la próxima sección.

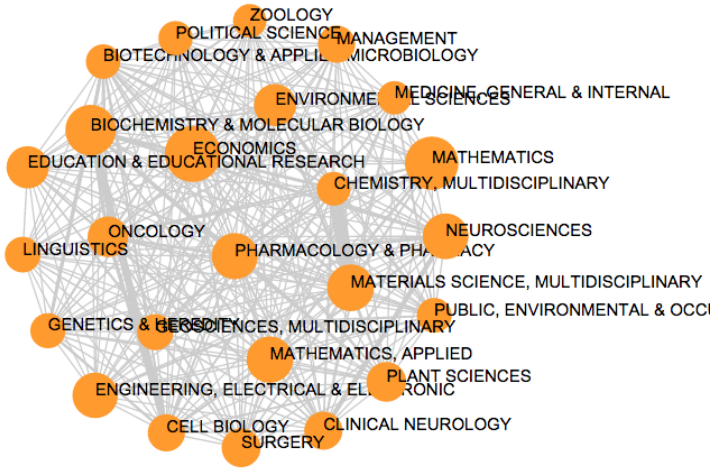



**Figura 1.11:** Ejemplo de una red entre áreas de la ciencia, según la clasificación de Thomson Reuters$^{TM}$. El tamaño de los nodos es proporcional a la cantidad de *journals* indexados en esa área en el año 2014. Imagen creada utilizando la aplicación web de la misma empresa.

## 1.5. Qué son los mapas de la ciencia

Las redes de información entre áreas de la ciencia —objeto de estudio de esta tesis—, constituyen un particular tipo de red basada en información científica. Los así denominados, *mapas de la ciencia*, se pueden rastrear hasta épocas ancestrales donde ya se proponían diagramas que mostraban la clasificación del conocimiento [Lima and Shneiderman, 2014, p. 28], como el tan difundido árbol de Porfirio creado en la antigua Grecia o el árbol de la ciencia de Ramón Llull (ver Figura 1.12) —quien para muchos es el padre de las ciencias de la información— que data del siglo XIII. Aquellos diagramas ya presentaban las características de una red, que hoy es la misma que se utiliza para elaborar mapas de la ciencia.

Al nivel de categorías, los Mapas de la Ciencia, son redes que relacionan campos de la ciencia entre sí. Estas relaciones se pueden interpretar como conexiones de similaridad basadas en una medida específica o señal. En la época moderna, trabajos seminales en esta idea de relacionar y representar áreas del conocimiento se deben a Moreno [1934] quien propone un primer diagrama de las áreas del conocimiento. Después, el primer denominado —textualmente— *Mapa de la Ciencia* es creado por Bernal [1939]. Desde entonces, un número importante de esfuerzos se han realizado con la intención de revelar las complejidades de la Ciencia.

Nótese que la palabra *mapa* hace énfasis, en la analogía con aquellos mapas que provee la cartografía (ver Sección 1.2), en la intención de dar fuerza a la idea de que un mapa ayuda a vislumbrar el territorio, planificar el viaje y tomar decisiones respecto a él.

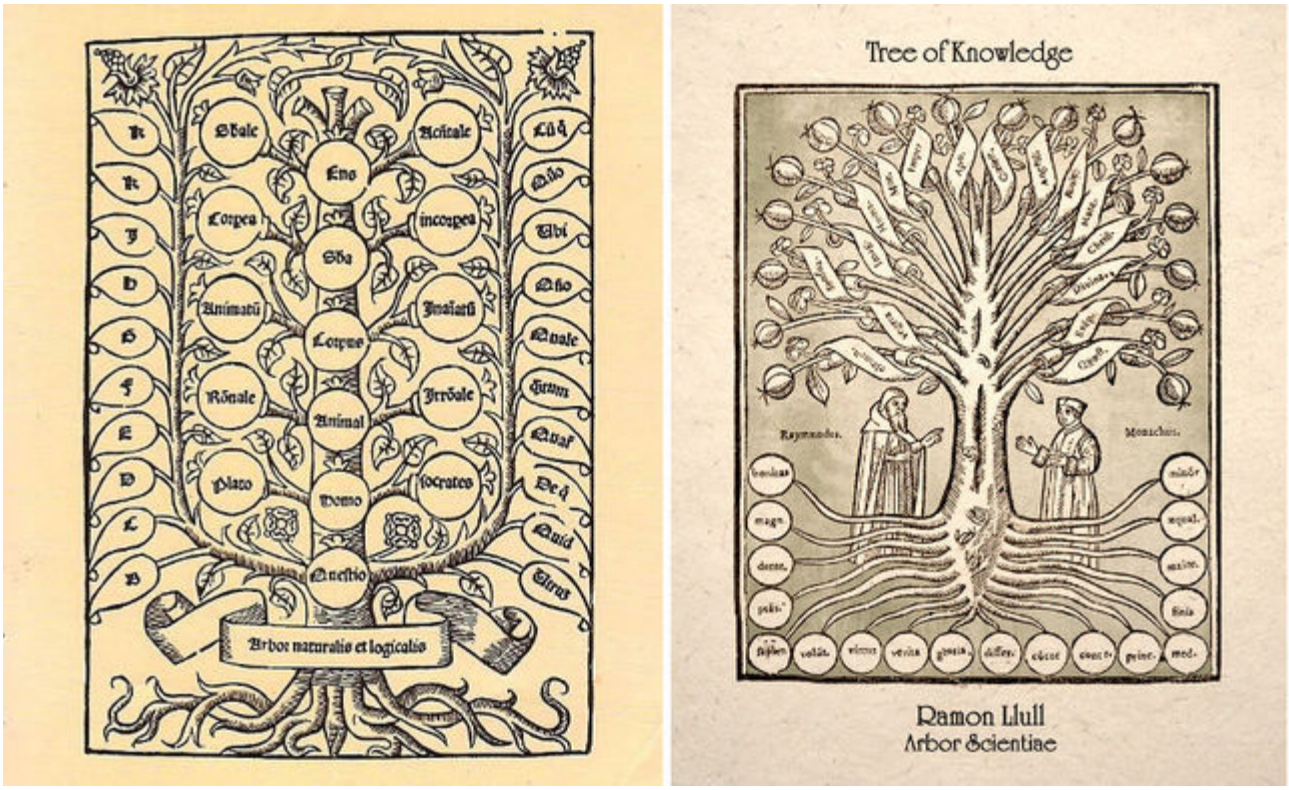



**Figura 1.12:** Representación del árbol de Porfirio (izq) y del árbol de la ciencia del catalán Ramon Llull, siglo XIII.

Una recopilación de la trayectoria y métodos utilizados para crear mapas de la ciencia se puede encontrar en el libro *Atlas of Science* de Börner [2010], y un análisis de las similitudes y diferencias de una veintena de mapas, se puede consultar en el *Consensus map of science* de Klavans and Boyack [2009].

Nótese que es justamente en esta área disciplinar, donde se enmarca nuestro trabajo, esto es, la generación automática de mapas de la ciencia.

### 1.5.1. Diferentes miradas, similares mapas

Según el objetivo o la intención de quienes mapean la ciencia, los mapas pueden tener diferentes formas y características pero en última instancia son de similar esencia. Por ejemplo, los profesionales de la información y bibliotecólogos crean estos mapas con la intención de organizar espacialmente el conocimiento, similar al ejercicio que se realiza cuando se desea clasificar los libros en los estantes de una biblioteca [Klavans and Boyack, 2009], se procura colocar más cerca disciplinas de mayor similitud, entendiendo similitud como el conocimiento que intercambian o que fluye entre disciplinas.

Los profesionales del mundo de los sistemas complejos —con ascendiente a la física—, crean mapas de la ciencia con la intención de representar las dinámicas o flujos de información que suceden entre áreas de la ciencia [Rosvall and Bergstrom, 2008].

Empresas consultoras, así como los investigadores y profesionales encargados de política pública crean estos mapas con el nombre de *Brain Scan* [Adams, 2015] o *Overlay Maps* [Leydesdorff et al., 2013a; Rafols et al., 2010; SciTech Strategies, 2012a], con el objetivo de visualizar el estado de desarrollo actual de los países e instituciones en relación al mapa de la ciencia.

Los cientistas sociales y estudios del comportamiento humano, elaboran mapas de la ciencia con la intención de analizar la interdisciplinariedad y diversidad de las áreas de la ciencia y de los hacedores del conocimiento [Leydesdorff et al., 2015; Rafols et al., 2010].

Los profesionales y científicos al interior de sus propias disciplinas, crean *mapas locales* de la ciencia o *mapas disciplinares*, para descubrir la estructura de relaciones intra-disciplinas y la dinámica de su comunidad [Pham et al., 2011].

Los artistas y medios de difusión, crean mapas lúdicos de la ciencia [Börner, 2010, p.174,188], o interpretan mapas ya creados utilizando metáforas como la de la Vía Láctea [Börner, 2010]. En resumen, nos encontramos en una época, como bien diría Katy Börner, donde "todos pueden mapear", pero que sin embargo, nos plantea nuevos desafíos tales como mejorar la precisión de estos mapas y la generación de aplicaciones prácticas en directo beneficio de todos los involucrados en el mundo de la investigación.

## 1.6. Planteamiento del problema

Como se ha visto en las secciones precedentes, en las últimas décadas se han creado un importante número de mapas de la ciencia. Sin embargo, este desarrollo aún adolece de tres principales inconvenientes que nos interesa abordar.

**Primero** Se han realizado pocos aportes en la creación de mapas en los que las relaciones sean distintas a las desprendidas del acto de citar. Vale decir, la mayoría de mapas de la ciencia están basados en citas y han profundizado mucho en cuanto a los flujos de información que se producen entre áreas de la ciencia. Estos mapas se construyen en base, por ejemplo, a las citas que se hacen de un área a otra, lo que no implica necesariamente que un científico sea capaz de publicar en las áreas en las que cita. Estos métodos también tienen la desventaja de que requieren un alto costo computacional por cuanto trabajan sobre la matriz de citas entre *papers* que es una matriz cuadrática de varios millones de registros. Es por esto que nuestra mirada está orientada a construir una red de información de áreas de la ciencia, basada en producción científica —no en citas—, la misma que refleje las capacidades de los científicos para producir ciencia en las diferentes categorías de la ciencia, más allá de los flujos de información entre áreas.

**Segundo** La disponibilidad de mapas de la ciencia se ha centrado principalmente en la creación gráfica de estos mapas, pero no en el análisis cuantitativo de su utilidad en tanto herramientas para la toma de decisiones. En general todos los mapas se han hecho disponibles en su componente gráfica, pero solo unos pocos (no más de 3) se han hecho disponibles en su componente cuantitativo (matriz de adyacencia), lo que hace ver la poca utilización —sin tomar en cuenta consultorías privadas— que se ha venido realizando de estos mapas en términos de herramientas para la toma de decisiones. En esta misma línea, los mapas han servido para analizar las características pasadas o actuales de los productores de ciencia, pero no para recomendar el desarrollo futuro. Por ejemplo, se han avanzado trabajos en términos de ubicar a los productores sobre el mapa [Rafols et al., 2010] o medir su diversidad considerando también la disparidad entre áreas de la ciencia [Guevara et al., 2016b; Rafols, 2014], sin embargo, aún no existen estudios, hasta donde hemos podido conocer, que evalúen la utilidad de estos mapas para apoyar la toma de decisiones. Esta funcionalidad ha quedado relegada a servicios de empresas privadas como SciTech Strategies [2012b], Digital Science [2016] o Science Metrix [2016].

**Tercero** La aplicación y el análisis descriptivo que se ha hecho en base a mapas de la ciencia, ha estado limitada en dos aspectos principales: los niveles de granularidad de análisis y la cantidad de productores analizados. La limitación de no contar con conjuntos de datos de producción científica a un alto nivel de resolución (granularidad), esto es, a nivel de individuos, ha impedido conducir análisis de la aplicación de mapas de la ciencia a nivel de personas e instituciones. En general, no existen estudios que, utilizando mapas de la ciencia, describan el comportamiento de los científicos y los estudios que existen a nivel de instituciones, se limitan a casos de estudio de unas pocas (no más de 5) instituciones [Rafols et al., 2010]. A nivel de países, aún no se han realizado estudios masivos que utilicen mapas de la ciencia, aunque sí existen estudios cuantitativos de la producción científica de los países con técnicas propias de la bibliometría o la economía ([Abramo and D'Angelo, 2014; Cimini et al., 2014; Harzing and Giroud, 2014]).

Nuestra propuesta pretende hacerse cargo de estos vacíos en la literatura de mapas de la ciencia, para lo que en las siguientes secciones detallaremos el planteamiento de esta

investigación.

### 1.6.1. Objetivos

Una vez que hemos determinado los principales vacíos en el desarrollo actual de los mapas de la ciencia, nos planteamos el siguiente objetivo general:

Determinar la estructura de una red de categorías de la ciencia, basada en las trayectorias productivas de los científicos.

Para cada nivel de granularidad de las distintas entidades productoras de ciencia ( individuos, instituciones y países), nos planteamos los siguientes objetivos:

- Describir la diversificación productiva de la entidad productora de ciencia.
- Evaluar si una red basada en trayectorias productivas es mejor descriptor de la diversificación y el desarrollo científico de la entidad productora de ciencia, en comparación con una red basada en patrones de citas.
- Evaluar si una red basada en trayectorias productivas es robusta al cambio de clasificación de áreas de la ciencia.

Entenderemos como *mejor descriptor* aquel mapa que entregue mejores resultados promedio en el valor del *área bajo la curva ROC* (siglas en inglés de Response Operative Characteristics) para la entidad productora de ciencia en estudio, al momento de evaluar cómo se diversifican en el tiempo a través de las áreas de la ciencia representadas en el mapa.

## 1.7. Hipótesis de investigación

Siendo que creemos que los mapas de la ciencia creados hasta el momento —principalmente basados en patrones de citación— son un excelente descriptor del flujo de información entre áreas de la ciencia [Klavans and Boyack, 2009; Rosvall and Bergstrom, 2008], y permiten clasificar el conocimiento eficientemente [Boyack and Klavans, 2014; Leydesdorff et al., 2016], también creemos que es posible crear una red que describa mejor cómo los productores de ciencia se diversifican a través de las áreas de investigación.

Es por lo anterior que nos planteamos la siguiente hipótesis de investigación: *la estructura de una red, basada en la trayectoria productiva (autorías) de los científicos será capaz de describir mejor cómo se diversifican y desarrollan en el tiempo los productores de ciencia (individuos, instituciones o países), en comparación con otra red basada en patrones de citas.*

A la red basada en trayectoria productiva (propósito de esta tesis) le llamaremos *espacio investigación* y para la comparación con una red basada en patrones de citas, utilizaremos el mapa UCSD (ver Sección 3.4).

Como medida de comparación calcularemos el promedio del área bajo la curva ROC, construida para cada productor de ciencia, a partir de la evaluación de la predicción de su diversificación en el tiempo. Realizaremos la predicción de las áreas a las que se va a diversificar (áreas nuevas) un productor en el futuro, construyendo un ranking de áreas recomendadas,

el mismo que está basado en la *densidad de activación* que es una medida de cuán cercanas y avecindadas se encuentran las áreas inactivas de las áreas ya activadas (ver Sección 5.1.5).

En base a estas definiciones, nuestra hipótesis investigativa conllevará las siguientes hipótesis alternativas[9] que pretendemos validar a lo largo de esta tesis:

$H_1$ El espacio investigación es mejor descriptor de la diversificación de *individuos* en áreas de la ciencia en comparación con el mapa UCSD.

$H_2$ El espacio investigación es mejor descriptor de la diversificación de *instituciones* en áreas de la ciencia en comparación con el mapa UCSD.

$H_3$ El espacio investigación es mejor descriptor de la diversificación de *países* en áreas de la ciencia en comparación con el mapa UCSD.

Con la intención de probar la robustez de nuestro método, también construiremos un espacio investigación en una clasificación de áreas de la ciencia distinta a la utilizada por el mapa UCSD. Para este fin, hemos elegido la clasificación SCImago y para efectos de comparación, debido a que no está disponible públicamente un mapa previamente creado en esta clasificación, hemos construido una red basada en la probabilidad de que un *journal* esté indexado en dos áreas al mismo tiempo (co-ocurrencia de *journals* en categorías). Lo que nos entrega una red en esa clasificación y que al mismo tiempo nos ayudará a descartar que el espacio investigación es dependiente de la clasificación de los *journals* en categorías. Denominamos a esta red, *mapa SCImago* y por lo tanto, evaluaremos también las siguientes hipótesis alternativas:

$H_4$ El espacio investigación es mejor descriptor de la diversificación de *individuos* en áreas de la ciencia en comparación con el mapa SCImago.

$H_5$ El espacio investigación es mejor descriptor de la diversificación de *instituciones* en áreas de la ciencia en comparación con el mapa SCImago.

$H_6$ El espacio investigación es mejor descriptor de la diversificación de *países* en áreas de la ciencia en comparación con el mapa SCImago.

En el ánimo de evaluar las capacidades descriptivas del espacio investigación, no solamente en cuanto a la aparición de nuevas áreas en el portafolio productivo de instituciones y países, sino también el nivel de desarrollo (productivo) de cada área, en comparación con otros productores de ciencia, realizaremos la misma evaluación (basada en densidad y curvas ROC) para otras dos transiciones: desde áreas *nacientes* hacia áreas *desarrolladas* y para áreas *intermedias* hacia áreas *desarrolladas*. La definición de áreas nacientes, intermedias y desarrolladas, se realizará calculando las ventajas comparativas Revealed Comparative Advantajes (RCA) para los datos de producción de instituciones y universidades. Los detalles de cada definición de estas variables se pueden consultar en la Sección 5.1.1.

Esta consideración, conlleva la formulación de las siguientes cuatro hipótesis alternativas:

[9]Para mejorar la comunicación del propósito de esta tesis, hemos omitido las correspondientes hipótesis nulas que se pueden entender como el opuesto a las hipótesis alternativas aquí planteadas.

$H_7$ El espacio investigación es mejor descriptor del desarrollo de *instituciones* en áreas de la ciencia en comparación con el mapa UCSD.

$H_8$ El espacio investigación es mejor descriptor del desarrollo de *países* en áreas de la ciencia en comparación con el mapa UCSD.

$H_9$ El espacio investigación es mejor descriptor del desarrollo de *instituciones* en áreas de la ciencia en comparación con el mapa SCImago.

$H_{10}$ El espacio investigación es mejor descriptor del desarrollo de *países* en áreas de la ciencia en comparación con el mapa SCImago.

Entenderemos como *desarrollo* de un área, la evaluación de las dos transiciones definidas previamente en base a la medida de RCA.

## 1.8. Dominio y alcance

Este trabajo se circunscribe en las siguientes áreas de la informática, según la taxonomía ACM Computing Classification System (CCS):

- Applied computing, Digital libraries and archives
- Human-centered computing, Scientific visualization
- Computing methodologies, Network science

Están fuera del alcance de este trabajo, mapas de la ciencia a otros niveles de agregación, como tópicos, *papers* o *journals*. También están fuera del alcance de este trabajo el análisis de comunidades científicas de personas, redes de cocitación o indicadores de calidad de *journals*, individuos, instituciones o países. Este trabajo no profundiza la teoría de la complejidad ni métodos asociados, tales como algoritmos de detección de comunidades o algoritmos de generación de redes.

Sin embargo, a lo largo de esta tesis, se hace referencia a varios conceptos incluidos en esas áreas, sin profundizar en ellos. Esto con la intención de proveer contexto al trabajo realizado en esta tesis.

### 1.8.1. Respecto de las imágenes y mapas

Debido a la naturaleza del trabajo abordado en esta tesis, es necesario indicar que aquellas Figuras que contienen mapas, no generados producto de esta tesis, se encuentran referenciadas adecuadamente para que el lector pueda acceder a la imagen original en alta resolución o a la aplicación que las generó. Respecto de los mapas generados producto de esta tesis, en la versión digital de este documento y debido a que estas imágenes son de tipo vectorial, se pueden escalar a voluntad para revisarlas en detalle. Sin embargo, para la versión impresa, también hemos incluido en los Anexos, imágenes impresas a color, en alta resolución y en formato doble carta.

## 1.9. Organización del documento

El documento se encuentra organizado de la siguiente forma. En el Capítulo 2 se entregan los conceptos preliminares necesarios para comprender los fundamentos de los mapas de la ciencia, las fuentes principales de datos y las técnicas utilizadas actualmente. El Capítulo 3 hace un recorrido por los principales mapas de la ciencia que se han construido con diferentes enfoques y conjuntos de datos (*datasets*) en el último tiempo. Nuestra propuesta de espacio investigación se encuentra detallada en el Capítulo 4, mientras que la evaluación cuantitativa que realizamos de nuestro mapa se encuentra en el Capítulo 5. El Capítulo 6 describe las principales aplicaciones desarrolladas en el marco de esta tesis. Finalmente, las conclusiones y el trabajo futuro se detallan en el Capítulo 7.

Capítulo 2

# Preliminares: Cómo se construyen mapas de la ciencia

Un mapa de la ciencia, consiste en una red compuesta por nodos y enlaces. En esta red, los nodos representan *entidades de la ciencia* y los enlaces, la relación, generalmente de similitud, entre estas entidades. En este capítulo analizamos el proceso y los métodos utilizados para construir mapas de la ciencia, así como la definición formal de este tipo de mapa.

## 2.1. Unidades fundamentales de los mapas de la ciencia

En un primer enfoque, las unidades fundamentales de un mapa de la ciencia —los nodos del mapa—, corresponden a entidades de la ciencia que contienen conocimiento científico. Éstas se pueden agregar a distintos niveles de granularidad como se ha introducido en el Capítulo 1. El nivel atómico de las entidades de la ciencia, constituyen las publicaciones científicas (*papers*), que se agregan en revistas científicas (*journals*), que se agregan en categorías, las que finalmente se pueden agregar en áreas de la ciencia. Para tener una idea de la dimensión de cada nivel, podemos decir que la cantidad de *papers* es de orden $10^6$, la de *journals* es orden $10^4$, la de categorías es de orden $10^2$ y la de áreas es de orden $10^1$.

Un segundo enfoque, proveniente de la minería de texto, caracteriza los *papers* por sus palabras, agrupa las palabras en tópicos y los tópicos en categorías. En este enfoque los nodos del mapa corresponden a tópicos y los enlaces representan similitudes entre textos. Estos mapas tienen como objetivo entregar más detalle y dinamismo a la red, esto a través de permitir que los *papers* se *autoagrupen* en función de su texto, sin que tenga que mediar una clasificación pre-establecida de las revistas científicas asignadas en áreas.

En esta tesis nos interesa trabajar con el primer enfoque y abordar un nivel de granularidad no tan específico (como tópicos) debido a su alta variabilidad a través del tiempo, ni tan general —como áreas— debido a que sería poco útil por su grueso nivel de agregación. Por tal motivo, hemos elegido trabajar a nivel de categoría, considerando que el dinamismo de este nivel, no es tan variable en función del tiempo, vale decir, se crean pocas categorías

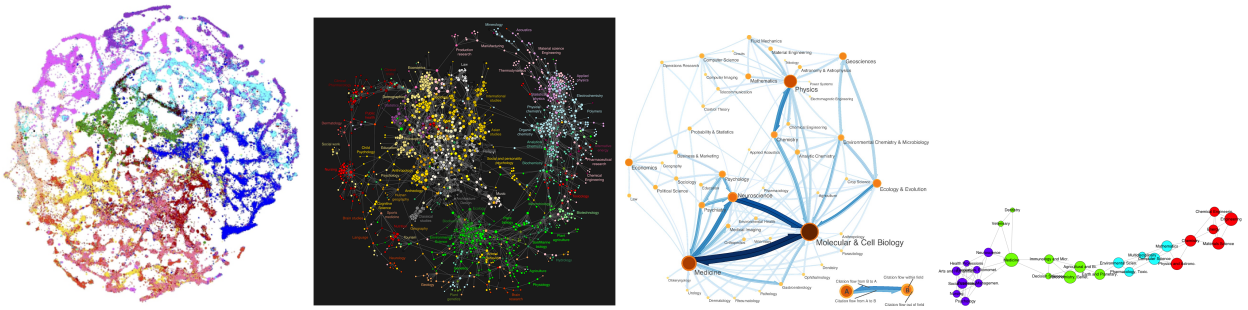



**Figura 2.1:** Mapas de la ciencia construidos a diferentes niveles de granularidad. Los niveles (nodos) son: *papers* [Boyack and Klavans, 2014], *journals* [Bollen et al., 2009], categorías [Rosvall and Bergstrom, 2008] y áreas de la ciencia [Guevara et al., 2016b].

nuevas cada año y a su vez, provee un nivel de detalle adecuado para la toma de decisiones a nivel macro.

En relación al enfoque elegido, diferentes estudios han propuesto mapas de la ciencia para cada nivel de granularidad. En la Figura 2.1 se muestra un mapa de ejemplo para cada nivel de las diferentes entidades de la ciencia, esto es, nivel *paper* [Boyack and Klavans, 2014], *journals* [Bollen et al., 2009], categorías [Rosvall and Bergstrom, 2008] y áreas de la ciencia [Guevara et al., 2016b]. Esta imagen nos permite dejar de manifiesto que han existido en la comunidad científica, intereses diversos en abordar cada nivel con distintas ópticas. Imágenes detalladas de alta resolución, se pueden encontrar en los trabajos citados.

## 2.2. Definición matemática de un mapa de la ciencia

Formalmente la representación matemática de un mapa o red, está dada por el grafo $G = (V, E)$ donde $V$ es el conjunto de vértices, también llamados nodos del grafo y, $E$ el conjunto de enlaces, también llamados arcos del grafo. En el caso de mapas de la ciencia, los vértices representan entidades de la ciencia, a diferentes niveles de granularidad como se han definido en la Sección 2.1 y que en muchos casos corresponden a la clasificación utilizada, como aquellas definidas en la Sección 1.3.3.

En cuanto a los enlaces del conjunto $E$, estos son pares del tipo $\{u, v\}$, para $u, v \in V$. Si estos pares son ordenados, esto es, $\{u, v\} \neq \{v, u\}$, se dice que el grafo es *dirigido*, y los arcos se representan por flechas donde $u$ corresponde a la cola del arco (origen), y $v$ corresponde a la cabeza (destino). La mayoría de mapas de la ciencia son no dirigidos. La cantidad de nodos $N_v = |V|$ y la cantidad de enlaces $N_e = |E|$ son conocidos como *orden* y *tamaño* del grafo, respectivamente.

La noción fundamental de *conectividad* de un grafo está dada por la *matriz de adyacencia* $\mathbf{A}$, la misma que es simétrica, binaria y cuadrada de $N_v \times N_v$. Los elementos de esta matriz se definen como

$$A_{ij} = \begin{cases} 1, & \text{si } \{i, j\} \in E, \\ 0, & \text{caso contrario,} \end{cases} \tag{2.1}$$

es decir, se colocan números uno en las celdas que conectan dos vértices y cero en caso contrario.

Si en lugar de una matriz binaria, se considera una matriz numérica con valores en los reales, en la que los números entregan información de la intensidad del enlace, comúnmente conocido como *peso* del enlace, se puede hablar de una *matriz de pesos*. En los mapas de la ciencia, los pesos representan generalmente, el valor de similitud o proximidad entre entidades de la ciencia. Vale decir, la matriz de pesos es conocida también como *matriz de similitud* a la que notaremos por $\Phi$, y sus valores dependerán de la medida utilizada para calcular la proximidad entre entidades. Una estructura de datos común para representar esta matriz es la *lista de enlaces* (*edge list*) la misma que consiste en una lista de tres columnas. Las dos primeras incluyen los pares de vértices que forman un enlace y la tercera columna incluye el valor de similitud (peso) de ese enlace. En R [R Core Team, 2014] es común utilizar la estructura de datos *dataframe* para almacenar estos datos.

Definir la estructura fundamental de un mapa de la ciencia, básicamente consistirá en definir la matriz de similitud $\Phi$ entre entidades de la ciencia.

Esta matriz, generalmente se calcula aplicando una medida de similitud o proximidad, sobre otra matriz $\mathbf{M}$, que contiene el indicador (bruto) de relación que se utilizará para definir el mapa. Las medidas de similitud pueden ser por ejemplo, *similitud coseno*, *índice Jaccard*, *probabilidad condicional*, entre otras (ver Sección 2.4). En tanto, los indicadores que definen la matriz $\mathbf{M}$ se analizan en detalle en la sección siguiente, por cuanto constituyen uno de los ejes centrales de nuestra propuesta.

## 2.3. Indicadores de relación utilizados para construir mapas de la ciencia

Los indicadores incluidos en la matriz $\mathbf{M}$, determinan la intensidad con que dos entidades de la ciencia —publicaciones, *journals*, categorías— se encuentran *más o menos lejos*. Waltman and van Eck [2012] denominan estas señales como indicadores de relación (*relatedness*) entre entidades de la ciencia. Estos indicadores pueden estar basados en citas-referencias, información compartida, archivos de registros (*logs*) o, en nuestro caso, productividad de los investigadores.

### 2.3.1. Basadas en citas

La señal preferida a lo largo de la historia de los mapas de la ciencia, es la que está basada en referencias y citas, que se puede interpretar o capturar de varias formas. Los diferentes métodos para capturar esta señal, son los siguientes:

- Co-citación (*co-citation*): Es la señal o relación que se produce entre dos *papers* que son citados dentro de un mismo *paper*. En definitiva se construye la red de todos contra todos, en base al total de referencias que incluye un *paper*.

- Citar-Citado (*cite*): Un *paper* puede citar a otro, lo que define la señal del *paper* que cita hacia el *paper* que es citado. También se conoce como citación directa.

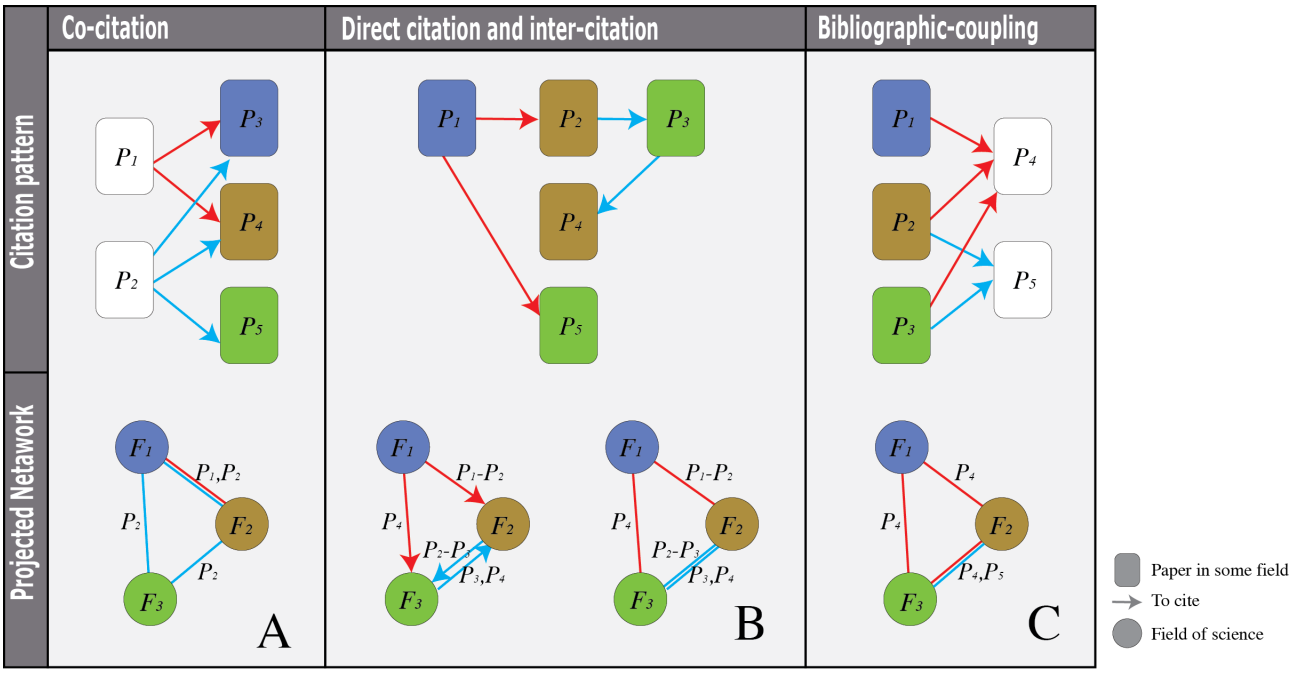



**Figura 2.2:** Esquema de cómo se capturan señales entre áreas de la ciencia, basados en referencias-citas entre *papers*. En la parte superior se representa el patrón de citación, mientras que en la parte inferior se presenta la red o mapa que se proyecta en base al método utilizado.

- Inter-citación (*intercite*): Considera y agrega tanto el acto de citar como el de ser citado. Es decir, relaciona dos elementos, por citas, sin importar la dirección de la cita. También es conocido como cita cruzada (*cross-citation*).
- Bibliografía emparejada (*Bibliographic coupling*): Relaciona dos *papers* que citan un mismo *paper*.

La Figura 2.2 presenta de forma esquemática los diferentes métodos basados en citas y la forma en que se construyen las relaciones o señales entre elementos de la ciencia.

Nótese que lo descrito anteriormente, puede ser extrapolado a nivel de *journals* y categorías, tomando, en lugar del *paper*, el *journal* al que este pertenece, o la categoría en la que está indexado. Así por ejemplo, utilizando el método de *intercitas* a nivel de *journals*, se puede construir una matriz $\mathbf{M}$ en la que sus elementos están definidos como

$$M_{ij} = C_{ij} + C_{ji}, \tag{2.2}$$

donde $C_{ij}$ es el número de citas del *journal* $i$ al *journal* $j$ y $C_{ji}$ la cantidad de citas en sentido contrario. Nótese además que $M_{ij} = M_{ji}$.

El cálculo computacional de la ecuación anterior, se facilita y optimiza si se computa $\mathbf{M}$ utilizando matrices. Para esto se puede aplicar la siguiente ecuación

$$\mathbf{M} = \mathbf{C} + \mathbf{C'}, \tag{2.3}$$

donde la matriz **C** incluye en filas y columnas los *journals*, mientras que las entradas de la matriz, corresponden a la cantidad de citas que realizan los *journals*, en las filas, a los *journals* en las columnas. En tanto, **C'** incluye el número de citas desde los *journals* en las columnas hacia los *journals* en las filas.

Los indicadores de relación, basados en referencias y citas, han sido ampliamente utilizados y estudiados para construir mapas de la ciencia y los diferentes métodos se han comparado según los propósitos específicos de los autores. Mientras que Boyack and Klavans [2010] señalan que para clusterizar publicaciones del área de biomédica, citación-directa, es menos exacto que co-citación o bibliografía emparejada. Por otro lado, Shibata et al. [2009] reportan que citación directa tiene mejor comportamiento para encontrar frentes de investigación, que co-citación o bibliografía emparejada.

Un reciente estudio de Börner et al. [2012b] revisa cuál de estos métodos es mejor a la hora de crear taxonomías para organizar la ciencia, basados en citas. Su conclusión es que la citación directa produce mejores resultados para este fin.

### 2.3.2. Basados en información mutua o co-ocurrencia

Otra mirada para capturar la señal que relaciona dos entidades de la ciencia, es aquella que propone medir la información compartida entre dos entidades de la ciencia. Esta técnica se aplica por ejemplo, al mapa Brain Scan [Adams, 2015], en el que se mide la cantidad de proyectos de investigación que se indexan en dos campos de investigación al mismo tiempo [HEFCE, 2014], entonces la matriz **M** incluye en sus filas y columnas campos de investigación, y los valores de sus celdas corresponden al número de proyectos indexados en las dos categorías. Esta técnica es, de alguna manera similar a la de Bibliografía emparejada, en el sentido de que dos entidades de la ciencia comparten la misma bibliografía.

En este ejemplo, la matriz **M** se puede obtener sobre los datos brutos, aplicando la siguiente ecuación:

$$\mathbf{M} = \mathbf{P}^{\intercal} \times \mathbf{P}, \tag{2.4}$$

donde **P** es la matriz binaria que contiene los datos brutos o matriz productiva, en este caso, de proyectos de investigación en las filas, y campos de investigación en las columnas. El valor de cada entrada $P_{ij}$ es 1 en caso de que el proyecto $i$ se encuentre indexado en el campo de investigación $j$ y es 0 en caso contrario. También se puede leer como: el proyecto $i$ se produce en el campo $j$. Nótese además que el producto matricial aplicado en la ecuación anterior, resulta en que la diagonal de **M** contiene la cantidad total de proyectos en cada campo de investigación.

Como veremos más adelante, este mismo cálculo matricial, puede ser extrapolado a cualquier otro indicador de información compartida, en nuestro caso, será aplicado a la cantidad de científicos que comparten las áreas de investigación.

### 2.3.3. Basados en archivos de registros o *logs*

Esta técnica está basada en minar o analizar los archivos de registros de aplicaciones informáticas que apoyan el proceso de búsqueda de información científica. Es una técnica novedosa en el campo de los mapas de la ciencia, pero bastante difundida en otras áreas como

Recuperación de Información (ver por ejemplo [Baeza-Yates et al., 2007]). Sus principales ventajas, constituyen el hecho de que analizan la ciencia, en el mismo lugar donde se está produciendo, esto es, en donde está sucediendo el ejercicio de investigar (los motores de búsqueda de información científica). Accediendo a los registros de herramientas de búsqueda y gestión de bibliografía como Mendeley, o los archivos de registros (logs) de bibliotecas institucionales, los practicantes de esta metodología [Bollen et al., 2009], son capaces de definir indicadores que permiten encontrar las relaciones de proximidad entre áreas de la ciencia.

Cada una de las técnicas descritas en esta sección, en última instancia, propiciarán redes en las que los enlaces son expresiones de diferentes fenómenos, por ejemplo: usando citas, el flujo de información; usando logs, la actividad investigativa; usando proyectos de investigación, el portafolio investigativo de un país.

En el Capítulo 4, referente a nuestra propuesta de mapa, abordaremos una nueva señal basada en capacidades productivas de los científicos.

## 2.4. Medidas de similitud

Hasta aquí se han determinado las formas en que se obtienen las matrices de indicadores $\mathbf{M}$ ya sea directamente o, basados en información productiva, como matrices de citas $\mathbf{C}$ o matrices de información mutua $\mathbf{P}$. El paso siguiente, para finalmente determinar la estructura básica de un mapa de la ciencia, es calcular la matriz de similaridad $\Phi$. Esto se realiza aplicando alguna medida de similitud a la matriz de indicadores o a las matrices productivas.

Existen una cantidad importante de medidas de similaridad que se pueden aplicar sobre matrices, las más usuales son: similaridad coseno, índice Jaccard y probabilidad condicional. La comunidad de mapas de la ciencia, también ha creado sus propias métricas de similaridad, dependiendo el tipo de datos con los que trabaja. Por ejemplo la medida *K50* que es una derivación de la similaridad coseno. Esta medida se propone en el trabajo seminal de Boyack et al. [2005] en el cual también se analizan las ventajas y desventajas de varias medidas de similaridad aplicables a datos basados en patrones de citas. Una compilación de medidas de similaridad también se puede revisar en la documentación del package para R, *proxy* [Meyer and Buchta, 2015]. A continuación se describen las medidas de mayor uso en la comunidad de mapas de la ciencia y también la medida de similaridad que se utilizará en este trabajo de tesis.

### 2.4.1. Similitud coseno

Considere una matriz productiva $\mathbf{P}$ de $M$ proyectos por $N$ campos de investigación (categorías), como la descrita en la ecuación 2.4. Para calcular la similitud coseno entre categorías, cada columna de la matriz se interpreta como un vector, cuyas $M$ componentes lo caracterizan. Si dos vectores (columnas) que representan categorías $\vec{V}(c_1)$ y $\vec{V}(c_2)$ son iguales, entonces el coseno del ángulo que forman ($0°$) es igual a 1. Si por otro lado, los dos vectores (categorías) son completamente dis-similares, entonces serán ortogonales (ángulo de $90°$) en el espacio vectorial, cuyo coseno es igual 0. Por consiguiente, valores altos de similitud coseno,

significarán valores altos de similitud entre categorías.

Las entradas de la matriz de similitudes $\Phi$ se calculan de la siguiente forma

$$\phi_{ij} = \frac{\vec{V}(c_i) \cdot \vec{V}(c_j)}{|\vec{V}(c_i)||\vec{V}(c_j)|}, \tag{2.5}$$

donde el numerador, representa el producto punto entre los dos vectores, y el denominador representa la multiplicación de normas. El producto punto de dos vectores es un escalar resultante de la suma de las multiplicaciones de las componentes. Por ejemplo el producto punto $\vec{x} \cdot \vec{y}$ de dos vectores, se define como $\sum_i x_i y_i$. La norma de todo $c$ se obtiene con $\sqrt{\sum_i \vec{V}_i^2(c)}$.

La similitud coseno también puede calcularse con una ecuación equivalente a la ecuación 2.5, en la que se calcula el producto punto de los vectores unitarios $\vec{v}(c_1)$ y $\vec{v}(c_2)$ correspondientes a los vectores $\vec{V}(c_1)$ y $\vec{V}(c_2)$. Generalizando, las entradas de la matriz $\Phi$ se calculan

$$\phi_{ij} = \vec{v}(c_i) \cdot \vec{v}(c_j), \tag{2.6}$$

donde cada vector unitario $\vec{v}(c) = \vec{V}(c)/|\vec{V}(c)|$. Matricialmente, $\Phi$ se puede computar con la siguiente ecuación

$$\Phi = \widehat{\mathbf{P}}^{\intercal} \times \widehat{\mathbf{P}} \tag{2.7}$$

donde $\widehat{\mathbf{P}}$ es la matriz de vectores unitarios, es decir, contiene los valores normalizados de $\mathbf{P}$.

Por otro lado, si previamente se ha calculado una matriz $\mathbf{M}$ con los productos punto de los vectores —no normalizados— incluidos en $\mathbf{P}$, como la matriz descrita en la ecuación 2.4; entonces, es posible calcular las entradas de la matriz $\Phi$, simplemente normalizando cada valor $m_{ij}$ de $\mathbf{M}$ por su correspondiente multiplicación de normas calculadas sobre la matriz original $\mathbf{P}$. Cada entrada de $\Phi$ se calcula con la siguiente ecuación

$$\phi_{ij} = \frac{m_{ij}}{|\vec{V}(c_i)||\vec{V}(c_j)|}. \tag{2.8}$$

En la intención de simplificar la similitud coseno, algunos autores como Boyack et al. [2005] utilizan en el denominador un factor de normalización más simple, que corresponde a la raíz cuadrada de la suma de las componentes del vector y no a su norma euclidiana. Esto es

$$\phi_{ij} = \frac{m_{ij}}{\sqrt{\sum_i c_{ij} \cdot \sum_j c_{ij}}}. \tag{2.9}$$

Sin embargo, en algunos casos es más factible contar con la matriz de productos punto $\mathbf{M}$ que con la matriz original de vectores $\mathbf{P}$, por lo que también se propone calcular el factor de normalización sobre la suma de elementos en cada columna $i, j$. Esto es

$$\phi_{ij} = \frac{m_{ij}}{\sqrt{\sum_i m_{ij} \cdot \sum_j m_{ij}}}. \tag{2.10}$$

Nótese que todas las matrices $\Phi$ definidas en esta sección son simétricas, por lo que, por ejemplo a la hora de construir la lista de enlaces de un mapa de la ciencia basado en estas matrices, se debe elegir solamente la matriz triangular superior ó la matriz triangular inferior. Adicionalmente, cabe mencionar, que la similitud coseno ha sido ampliamente utilizada en el área de Recuperación de Información donde interesa, por ejemplo, encontrar similitudes entre una consulta dada y los documentos en una colección. Más detalles de su definición y uso se pueden encontrar en el libro de Manning et al. [2008, p. 158].

### 2.4.2. K50

La medida K50, propuesta por Boyack et al. [2005] es una variación de la similitud coseno calculada según la ecuación 2.10. La variación consiste en restar a los valores $m_{ij}$ un valor esperado, entonces las entradas de la matriz $\Phi$ se calculan

$$\phi_{ij} = \frac{m_{ij} - E_{ij}}{\sqrt{S_i \cdot S_j}}, \tag{2.11}$$

donde el valor esperado $E_{ij} = \frac{S_i \cdot S_j}{SS - S_i}$, $S_i = \sum_i m_{ij}$ y $SS = \sum_i S_i$.

Como $E_{ij} \neq E_{ji}$, la matriz resultante $\Phi$ no es simétrica. Por lo que el proceso de construcción de un mapa de la ciencia, basado en este método, debe elegir entre el máximo o el mínimo de los valores $\phi_{ij}$ y $\phi_{ji}$ para la construcción de la lista de enlaces. Los practicantes de este método, usualmente eligen el máximo.

Esta medida ha sido utilizada en la creación de diferentes mapas de la ciencia. [Boyack, 2009; Börner et al., 2012b; Pham et al., 2011]

### 2.4.3. Probabilidad Condicional

Otra medida para definir similaridad, en base a matrices que describen entidades, es aquella basada en la probabilidad condicional. Por ejemplo, consideremos el caso en que las entidades son categorías de la ciencia. Si dos categorías $c_1$ y $c_2$ son muy similares, la probabilidad de que un evento ocurra en la categoría $c_1$ dado que ocurrió en la categoría $c_2$, debería ser muy alta. Formalmente, si consideramos las entradas $p$ de la matriz binaria binaria de producción $\mathbf{P}$ descrita en la ecuación 2.4; la probabilidad condicional de que un proyecto de investigación $k$ esté indexado en el campo de investigación $c_i$ y también en el campo $c_j$ , está dada por

$$\phi_{ij} = \frac{\sum_k p_{ki} p_{kj}}{\sum_i p_{ij}}. \tag{2.12}$$

Nótese que el numerador de la ecuación anterior, no es más que la información mutua entre los campos $i$ y $j$, vale decir, la cantidad de proyectos de investigación que están indexados en los dos campos al mismo tiempo. Esta información, es la misma entregada por la matriz $\mathbf{M}$ descrita en la ecuación 2.4, por lo que la ecuación anterior, se puede reescribir en función de $\mathbf{M}$, de la siguiente forma:

$$\phi_{ij} = \frac{m_{ij}}{\sum_i p_{ij}}. \tag{2.13}$$

Debe notarse también, que $\phi_{ij} \neq \phi_{ji}$ por lo que para la construcción de un mapa de la ciencia —no dirigido— basado en esta medida, se debe elegir entre el máximo o el mínimo de estos dos valores. Los practicantes de esta metodología, usualmente eligen el mínimo, dado que privilegian el peor caso posible, para no sobreevaluar los enlaces de la red. Sin embargo, bien se puede utilizar la matriz completa, considerando su direccionalidad.

Esta medida se ha utilizado ampliamente en redes producción económica [Hausmann and Hidalgo, 2013, p. 60] puesto que su principal ventaja radica en la facilidad para explicar cómo representa un fenómeno real, sin caer en tecnicismos como sucede, por ejemplo, con la similitud coseno.

## 2.5. Medidas de dis-similitud

En algunos casos, es conveniente trabajar con la noción contraria a similitud, esto es la dis-similitud, a menudo llamada distancia, aunque en estricto rigor, varias medidas de dis-similitud, no cumplen con la definición matemática formal de distancia.

### 2.5.1. Distancia Euclidiana

La noción más común de dis-similitud es la distancia Euclidiana, donde la distancia entre dos vectores $\vec{V}(c_i)$ y $\vec{V}(c_j)$ se calcula como la raíz cuadrada de la resta de las correspondientes componentes de cada vector (columna). Considere una matriz binaria de producción $\mathbf{P}$ de $M$ proyectos por $N$ campos de investigación (categorías), como la descrita en la ecuación 2.4. Para calcular la distancia Euclidiana de cada entrada de la matriz de distancias $\Phi'$ utilizamos la siguiente ecuación:

$$\phi'_{ij} = \sqrt{\sum_k \vec{V}_k(c_i) - \vec{V}_k(c_j)} \tag{2.14}$$

### 2.5.2. Transformaciones de dis-similitudes en similitudes

Una práctica común, es aplicar una transformación a las medidas de similitud, para convertirlas en medidas de dis-similitud. La noción básica es que ambas medidas deben ser opuestas, esto es, mientras una provee una idea de cercanía o proximidad, la otra debe entregar una noción de disparidad o de lejanía. Las dos transformaciones más comunes, dependiendo de la medida de similitud, son: restar la similitud de la unidad ($\Phi' = 1 - \Phi$), y, tomar el valor inverso ($\Phi' = 1/\Phi$). En el caso en que la medida de similitud no supere el valor de uno, la primera transformación será la más adecuada. Esto es, por ejemplo, el caso de la similitud coseno o la probabilidad condicional cuyos valores se encuentran en el rango $[0, 1]$.

Por ejemplo, la dis-similitud coseno $\phi'_{ij}$, entre dos vectores $\vec{V}(c_i)$ y $\vec{V}(c_j)$ de una matriz productiva $\mathbf{P}$, como la descrita en la ecuación 2.4, se puede adaptar de la ecuación de similitud coseno original 2.5, con la siguiente ecuación

$$\phi'_{ij} = 1 - \frac{\vec{V}(c_i) \cdot \vec{V}(c_j)}{|\vec{V}(c_i)||\vec{V}(c_j)|}. \tag{2.15}$$

## 2.6. Un ejemplo didáctico

En la intención de ejemplificar los conceptos y técnicas descritas en este capítulo, a continuación presentamos un ejemplo de cómo construir un sencillo mapa de la ciencia. Hemos elegido como fuente de datos, la producción científica total para el año 2013, de 10 países en las 27 áreas de la ciencia en las que Scopus® clasifica la información científica. Los datos fueron obtenidos y agregados desde el sitio público SCImago Journal & Country Ranking (SJR) [SCImago, 2007]. Con esta información se puede construir una matriz de producción **P** de 10 países por 27 áreas (ver Tabla 2.1). Este indicador (producción de *papers*) se puede utilizar para definir similitudes entre áreas.

| | Agricultural and Biological Sciences | Arts and Humanities | Biochemistry, Genetics and Molecular Biology | Business, Management and Accounting | Chemical Engineering | Chemistry | Computer Science | Decision Sciences | Dentistry | Earth and Planetary Sciences | Economics, Econometrics and Finance | Energy | Engineering | Environmental Science | Health Professions | Immunology and Microbiology | Materials Science | Mathematics | Medicine | Multidisciplinary | Neuroscience | Nursing | Pharmacology, Toxicology and Pharmaceutics | Physics and Astronomy | Psychology | Social Sciences | Veterinary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Argentina | 3507 | 477 | 2704 | 111 | 551 | 1219 | 638 | 53 | 39 | 1260 | 107 | 229 | 1136 | 1195 | 55 | 775 | 1083 | 728 | 3783 | 57 | 419 | 66 | 480 | 1542 | 186 | 867 | 224 |
| China | 35351 | 2011 | 69881 | 2927 | 29029 | 60479 | 62795 | 2523 | 663 | 30913 | 1845 | 24765 | 198239 | 31023 | 750 | 10528 | 98403 | 45498 | 72402 | 6004 | 5720 | 1029 | 17915 | 73797 | 1223 | 7051 | 1228 |
| Germany | 15603 | 4224 | 31131 | 3169 | 5915 | 17126 | 16129 | 1279 | 662 | 12552 | 2915 | 3863 | 22999 | 9241 | 1307 | 5345 | 21976 | 13347 | 55680 | 1198 | 7127 | 1527 | 4674 | 28139 | 4596 | 8160 | 1116 |
| Hungary | 1346 | 502 | 2032 | 147 | 347 | 1157 | 935 | 82 | 19 | 570 | 110 | 154 | 1417 | 589 | 51 | 332 | 1106 | 1272 | 2860 | 79 | 445 | 69 | 508 | 1566 | 268 | 800 | 176 |
| Iran | 4158 | 344 | 5824 | 482 | 2947 | 6536 | 4042 | 501 | 334 | 1951 | 157 | 2312 | 11424 | 3861 | 279 | 1332 | 7332 | 4186 | 11095 | 1057 | 577 | 274 | 1885 | 6180 | 170 | 900 | 517 |
| Mexico | 4164 | 459 | 2861 | 207 | 858 | 1895 | 2062 | 124 | 39 | 1395 | 177 | 599 | 2983 | 1980 | 54 | 788 | 2202 | 1925 | 4836 | 108 | 481 | 172 | 563 | 3102 | 332 | 1184 | 248 |
| Singapore | 870 | 333 | 3371 | 586 | 1178 | 2283 | 4102 | 252 | 47 | 389 | 373 | 736 | 5239 | 885 | 116 | 495 | 4137 | 1655 | 4434 | 175 | 464 | 188 | 431 | 2719 | 332 | 1301 | 13 |
| South Korea | 5825 | 766 | 15085 | 1101 | 5417 | 10030 | 9978 | 437 | 497 | 2519 | 630 | 2985 | 23239 | 4460 | 791 | 3024 | 19818 | 5490 | 23069 | 636 | 1804 | 1202 | 3464 | 12587 | 528 | 2269 | 419 |
| Spain | 11487 | 4177 | 13964 | 2223 | 3397 | 9562 | 11344 | 885 | 341 | 5476 | 1508 | 2665 | 12373 | 7453 | 932 | 2628 | 8754 | 7491 | 26882 | 384 | 2370 | 1284 | 2551 | 10589 | 2105 | 7868 | 739 |
| United States | 58949 | 28511 | 127654 | 16701 | 16831 | 40018 | 56514 | 5209 | 2322 | 42540 | 11578 | 14784 | 85753 | 38355 | 7758 | 22681 | 54186 | 38190 | 245790 | 5837 | 25449 | 14763 | 21860 | 61766 | 27289 | 64450 | 4041 |

**Tabla 2.1:** Matriz de producción de *glsplpaper* por área de la ciencia. Datos agregados desde SCImago Journal & Country Ranking (SJR) para el año 2013

En este ejemplo, las entidades de la ciencia corresponden a áreas (orden $10^1$), que es el mayor nivel de agregación de mapas de la ciencia (ver Sección 2.1). Considerando cada columna (área) como un vector, podemos construir un mapa de la ciencia entre áreas, utilizando la matriz de dis-similitudes $\Phi'$ obtenida aplicando la ecuación 2.15 de dis-similitud Coseno. La matriz $\Phi'$ se presenta en la tabla 2.2.

Finalmente, aplicando un filtro sobre la matriz de dis-similitudes $\Phi'$ y un algoritmo de representación en el espacio (layout) basado en atracción-repulsión, podemos graficar la red o mapa de la ciencia que se presenta en la Figura 2.3.

Nótese que en este primer ejemplo, aún cuando la información es pequeña y la agregación alta, es posible obtener un mapa de la ciencia, que es coherente con una evaluación cualitativa. Por ejemplo, se pueden apreciar cuatro comunidades, de izquierda a derecha, de Ciencias Sociales y Humanidades, Medicina y Biología, Ciencias ambientales y matemáticas, y finalmente, Ciencias naturales y aplicadas.

| | Agricultural and Biological Sciences | Arts and Humanities | Biochemistry, Genetics and Molecular Biology | Business, Management and Accounting | Chemical Engineering | Chemistry | Computer Science | Decision Sciences | Dentistry | Earth and Planetary Sciences | Economics, Econometrics and Finance | Energy | Engineering | Environmental Science | Health Professions | Immunology and Microbiology | Materials Science | Mathematics | Medicine | Multidisciplinary | Neuroscience | Nursing | Pharmacology, Toxicology and Pharmaceutics | Physics and Astronomy | Psychology | Social Sciences | Veterinary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Agricultural and Biological Sciences | 0.00 | 0.11 | 0.00 | 0.07 | 0.12 | 0.09 | 0.04 | 0.01 | 0.04 | 0.01 | 0.07 | 0.12 | 0.19 | 0.01 | 0.10 | 0.01 | 0.13 | 0.05 | 0.03 | 0.05 | 0.05 | 0.12 | 0.02 | 0.06 | 0.12 | 0.10 | 0.03 |
| Arts and Humanities | 0.11 | 0.00 | 0.09 | 0.01 | 0.43 | 0.38 | 0.27 | 0.07 | 0.05 | 0.15 | 0.01 | 0.42 | 0.53 | 0.17 | 0.00 | 0.07 | 0.45 | 0.29 | 0.03 | 0.26 | 0.02 | 0.00 | 0.18 | 0.30 | 0.00 | 0.00 | 0.04 |
| Biochemistry, Genetics and Molecular Biology | 0.00 | 0.09 | 0.00 | 0.05 | 0.14 | 0.11 | 0.05 | 0.00 | 0.03 | 0.01 | 0.06 | 0.13 | 0.21 | 0.02 | 0.08 | 0.00 | 0.15 | 0.06 | 0.02 | 0.05 | 0.04 | 0.09 | 0.02 | 0.07 | 0.10 | 0.08 | 0.03 |
| Business, Management and Accounting | 0.07 | 0.01 | 0.05 | 0.00 | 0.34 | 0.29 | 0.20 | 0.04 | 0.02 | 0.10 | 0.00 | 0.33 | 0.44 | 0.12 | 0.00 | 0.03 | 0.36 | 0.22 | 0.01 | 0.19 | 0.01 | 0.01 | 0.13 | 0.22 | 0.01 | 0.00 | 0.02 |
| Chemical Engineering | 0.12 | 0.43 | 0.14 | 0.34 | 0.00 | 0.00 | 0.02 | 0.16 | 0.26 | 0.09 | 0.35 | 0.00 | 0.01 | 0.07 | 0.40 | 0.17 | 0.00 | 0.02 | 0.26 | 0.03 | 0.31 | 0.43 | 0.06 | 0.02 | 0.45 | 0.40 | 0.25 |
| Chemistry | 0.09 | 0.38 | 0.11 | 0.29 | 0.00 | 0.00 | 0.01 | 0.13 | 0.22 | 0.06 | 0.30 | 0.01 | 0.03 | 0.05 | 0.35 | 0.14 | 0.01 | 0.01 | 0.22 | 0.03 | 0.26 | 0.38 | 0.04 | 0.01 | 0.40 | 0.35 | 0.21 |
| Computer Science | 0.04 | 0.27 | 0.05 | 0.20 | 0.02 | 0.01 | 0.00 | 0.07 | 0.15 | 0.03 | 0.21 | 0.02 | 0.06 | 0.01 | 0.25 | 0.08 | 0.03 | 0.00 | 0.14 | 0.01 | 0.18 | 0.27 | 0.01 | 0.01 | 0.29 | 0.25 | 0.14 |
| Decision Sciences | 0.01 | 0.07 | 0.00 | 0.04 | 0.16 | 0.13 | 0.07 | 0.00 | 0.02 | 0.02 | 0.04 | 0.16 | 0.24 | 0.03 | 0.06 | 0.00 | 0.18 | 0.08 | 0.01 | 0.07 | 0.03 | 0.08 | 0.03 | 0.08 | 0.09 | 0.06 | 0.02 |
| Dentistry | 0.04 | 0.05 | 0.03 | 0.02 | 0.26 | 0.22 | 0.15 | 0.02 | 0.00 | 0.07 | 0.02 | 0.26 | 0.36 | 0.08 | 0.03 | 0.02 | 0.27 | 0.16 | 0.01 | 0.14 | 0.02 | 0.05 | 0.09 | 0.16 | 0.05 | 0.04 | 0.01 |
| Earth and Planetary Sciences | 0.01 | 0.15 | 0.01 | 0.10 | 0.09 | 0.06 | 0.03 | 0.02 | 0.07 | 0.00 | 0.10 | 0.08 | 0.14 | 0.01 | 0.13 | 0.02 | 0.10 | 0.03 | 0.06 | 0.02 | 0.08 | 0.15 | 0.01 | 0.03 | 0.16 | 0.13 | 0.06 |
| Economics, Econometrics and Finance | 0.07 | 0.01 | 0.06 | 0.00 | 0.35 | 0.30 | 0.21 | 0.04 | 0.02 | 0.10 | 0.00 | 0.35 | 0.45 | 0.13 | 0.01 | 0.04 | 0.37 | 0.23 | 0.01 | 0.20 | 0.00 | 0.01 | 0.14 | 0.23 | 0.01 | 0.01 | 0.02 |
| Energy | 0.12 | 0.42 | 0.13 | 0.33 | 0.00 | 0.01 | 0.02 | 0.16 | 0.26 | 0.08 | 0.35 | 0.00 | 0.01 | 0.06 | 0.39 | 0.17 | 0.00 | 0.02 | 0.26 | 0.03 | 0.30 | 0.42 | 0.06 | 0.02 | 0.44 | 0.39 | 0.25 |
| Engineering | 0.19 | 0.53 | 0.21 | 0.44 | 0.01 | 0.03 | 0.06 | 0.24 | 0.36 | 0.14 | 0.45 | 0.01 | 0.00 | 0.12 | 0.50 | 0.25 | 0.01 | 0.05 | 0.35 | 0.07 | 0.40 | 0.53 | 0.11 | 0.06 | 0.55 | 0.50 | 0.35 |
| Environmental Science | 0.01 | 0.17 | 0.02 | 0.12 | 0.07 | 0.05 | 0.01 | 0.03 | 0.08 | 0.01 | 0.13 | 0.06 | 0.12 | 0.00 | 0.16 | 0.03 | 0.08 | 0.02 | 0.07 | 0.01 | 0.10 | 0.18 | 0.00 | 0.02 | 0.19 | 0.16 | 0.07 |
| Health Professions | 0.10 | 0.00 | 0.08 | 0.00 | 0.40 | 0.35 | 0.25 | 0.06 | 0.03 | 0.13 | 0.01 | 0.39 | 0.50 | 0.16 | 0.00 | 0.06 | 0.41 | 0.27 | 0.02 | 0.23 | 0.01 | 0.00 | 0.16 | 0.27 | 0.01 | 0.00 | 0.03 |
| Immunology and Microbiology | 0.01 | 0.07 | 0.00 | 0.03 | 0.17 | 0.14 | 0.08 | 0.00 | 0.02 | 0.02 | 0.04 | 0.17 | 0.25 | 0.03 | 0.06 | 0.00 | 0.18 | 0.09 | 0.01 | 0.07 | 0.02 | 0.07 | 0.03 | 0.09 | 0.08 | 0.06 | 0.02 |
| Materials Science | 0.13 | 0.45 | 0.15 | 0.36 | 0.00 | 0.01 | 0.03 | 0.18 | 0.27 | 0.10 | 0.37 | 0.00 | 0.01 | 0.08 | 0.41 | 0.18 | 0.00 | 0.02 | 0.27 | 0.04 | 0.32 | 0.44 | 0.07 | 0.02 | 0.47 | 0.42 | 0.27 |
| Mathematics | 0.05 | 0.29 | 0.06 | 0.22 | 0.02 | 0.01 | 0.00 | 0.08 | 0.16 | 0.03 | 0.23 | 0.02 | 0.05 | 0.02 | 0.27 | 0.09 | 0.02 | 0.00 | 0.16 | 0.01 | 0.19 | 0.30 | 0.02 | 0.00 | 0.31 | 0.27 | 0.15 |
| Medicine | 0.03 | 0.03 | 0.02 | 0.01 | 0.26 | 0.22 | 0.14 | 0.01 | 0.01 | 0.06 | 0.01 | 0.26 | 0.35 | 0.07 | 0.02 | 0.01 | 0.27 | 0.16 | 0.00 | 0.13 | 0.00 | 0.03 | 0.08 | 0.16 | 0.03 | 0.02 | 0.01 |
| Multidisciplinary | 0.05 | 0.26 | 0.05 | 0.19 | 0.03 | 0.03 | 0.01 | 0.07 | 0.14 | 0.02 | 0.20 | 0.03 | 0.07 | 0.01 | 0.23 | 0.07 | 0.04 | 0.01 | 0.13 | 0.00 | 0.17 | 0.25 | 0.01 | 0.02 | 0.27 | 0.23 | 0.13 |
| Neuroscience | 0.05 | 0.02 | 0.04 | 0.01 | 0.31 | 0.26 | 0.18 | 0.03 | 0.02 | 0.08 | 0.00 | 0.30 | 0.40 | 0.10 | 0.01 | 0.02 | 0.32 | 0.19 | 0.00 | 0.17 | 0.00 | 0.02 | 0.11 | 0.19 | 0.02 | 0.02 | 0.01 |
| Nursing | 0.12 | 0.00 | 0.09 | 0.01 | 0.43 | 0.38 | 0.27 | 0.08 | 0.05 | 0.15 | 0.01 | 0.42 | 0.53 | 0.18 | 0.00 | 0.07 | 0.44 | 0.30 | 0.03 | 0.25 | 0.02 | 0.00 | 0.18 | 0.30 | 0.00 | 0.00 | 0.05 |
| Pharmacology, Toxicology and Pharmaceutics | 0.02 | 0.18 | 0.02 | 0.13 | 0.06 | 0.04 | 0.01 | 0.03 | 0.09 | 0.01 | 0.14 | 0.06 | 0.11 | 0.00 | 0.16 | 0.03 | 0.07 | 0.02 | 0.08 | 0.01 | 0.11 | 0.18 | 0.00 | 0.02 | 0.20 | 0.16 | 0.08 |
| Physics and Astronomy | 0.06 | 0.30 | 0.07 | 0.22 | 0.02 | 0.01 | 0.01 | 0.08 | 0.16 | 0.03 | 0.23 | 0.02 | 0.06 | 0.02 | 0.27 | 0.09 | 0.02 | 0.00 | 0.16 | 0.02 | 0.19 | 0.30 | 0.02 | 0.00 | 0.32 | 0.28 | 0.15 |
| Psychology | 0.12 | 0.00 | 0.10 | 0.01 | 0.45 | 0.40 | 0.29 | 0.09 | 0.05 | 0.16 | 0.01 | 0.44 | 0.55 | 0.19 | 0.01 | 0.08 | 0.47 | 0.31 | 0.03 | 0.27 | 0.02 | 0.00 | 0.20 | 0.32 | 0.00 | 0.00 | 0.05 |
| Social Sciences | 0.10 | 0.00 | 0.08 | 0.00 | 0.40 | 0.35 | 0.25 | 0.06 | 0.04 | 0.13 | 0.01 | 0.39 | 0.50 | 0.16 | 0.00 | 0.06 | 0.42 | 0.27 | 0.02 | 0.23 | 0.02 | 0.00 | 0.16 | 0.28 | 0.00 | 0.00 | 0.03 |
| Veterinary | 0.03 | 0.04 | 0.03 | 0.02 | 0.25 | 0.21 | 0.14 | 0.02 | 0.01 | 0.06 | 0.02 | 0.25 | 0.35 | 0.07 | 0.03 | 0.02 | 0.27 | 0.15 | 0.01 | 0.13 | 0.01 | 0.05 | 0.08 | 0.15 | 0.05 | 0.03 | 0.00 |

**Tabla 2.2:** Matriz de dis-similitudes coseno entre áreas de la ciencia.

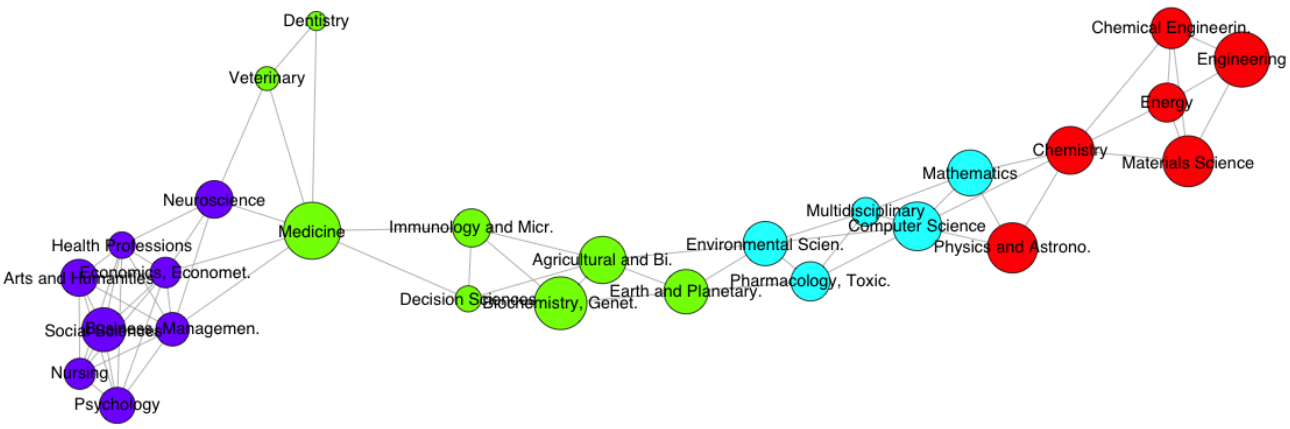



**Figura 2.3:** Mapa de la ciencia. Basado en datos de producción de 10 países en 27 áreas de la ciencia. Filtrados enlaces mayores a 0.015. Aplicado algoritmo de atracción-repulsión *Fruchterman-Reingold*. El tamaño de los nodos es proporcional al número de *papers* publicados en cada área. El color de los nodos está definido acorde al algoritmo de detección de comunidades *fastgreedy*.

Capítulo 3

# Estado del arte: Cartografía actual

En este capítulo revisamos los principales mapas que se han construido en las últimas décadas, haciendo énfasis en aquellos que han sido los más difundidos tanto en el mundo investigativo como en el mundo de las aplicaciones prácticas. Destacamos en cada uno los aspectos metodológicos, el tipo de datos utilizado, para finalmente encontrar sus similitudes y diferencias.

## 3.1. Mapas manuales

Los primeros mapas de la ciencia, se elaboraron de forma manual y derivaron de ser taxonomías creadas por expertos a mapas relacionados en función de cuánto conocimiento fluía de una ciencia a otra. Como hemos mencionado en la Sección 1.5, su existencia, puede ser rastreada tan atrás como a la antigua Grecia y la clasificación del árbol del filósofo Porfirio, que se reconoce como una de las primeras taxonomías de la ciencia; o hasta la época medieval y los árboles de la ciencia de Raymundo Llull, que representaban en sus raíces los conocimientos básicos necesarios para construir conocimiento en una determinada ciencia, en sus ramas los géneros de esa ciencia y en las hojas las especies, mientras que en los frutos las personas (ver Figura 1.12).

En la época moderna, un mapa digno de destacar es el creado por uno de los reconocidos padres de la bibliometría, Small [1973] quien proponía en la década de los 70 un primer mapa de co-citación para buscar similitudes entre documentos (ver Figura 3.1).

## 3.2. Mapas pioneros

Ya en la época contemporánea, comienzan a aparecer mapas construidos con el apoyo computacional. De esta época (década de los noventa) data este mapa de la ciencia pionero [Small, 1999] (ver Figura 3.2), creado por Herny Small, utiliza patrones de co-citación de

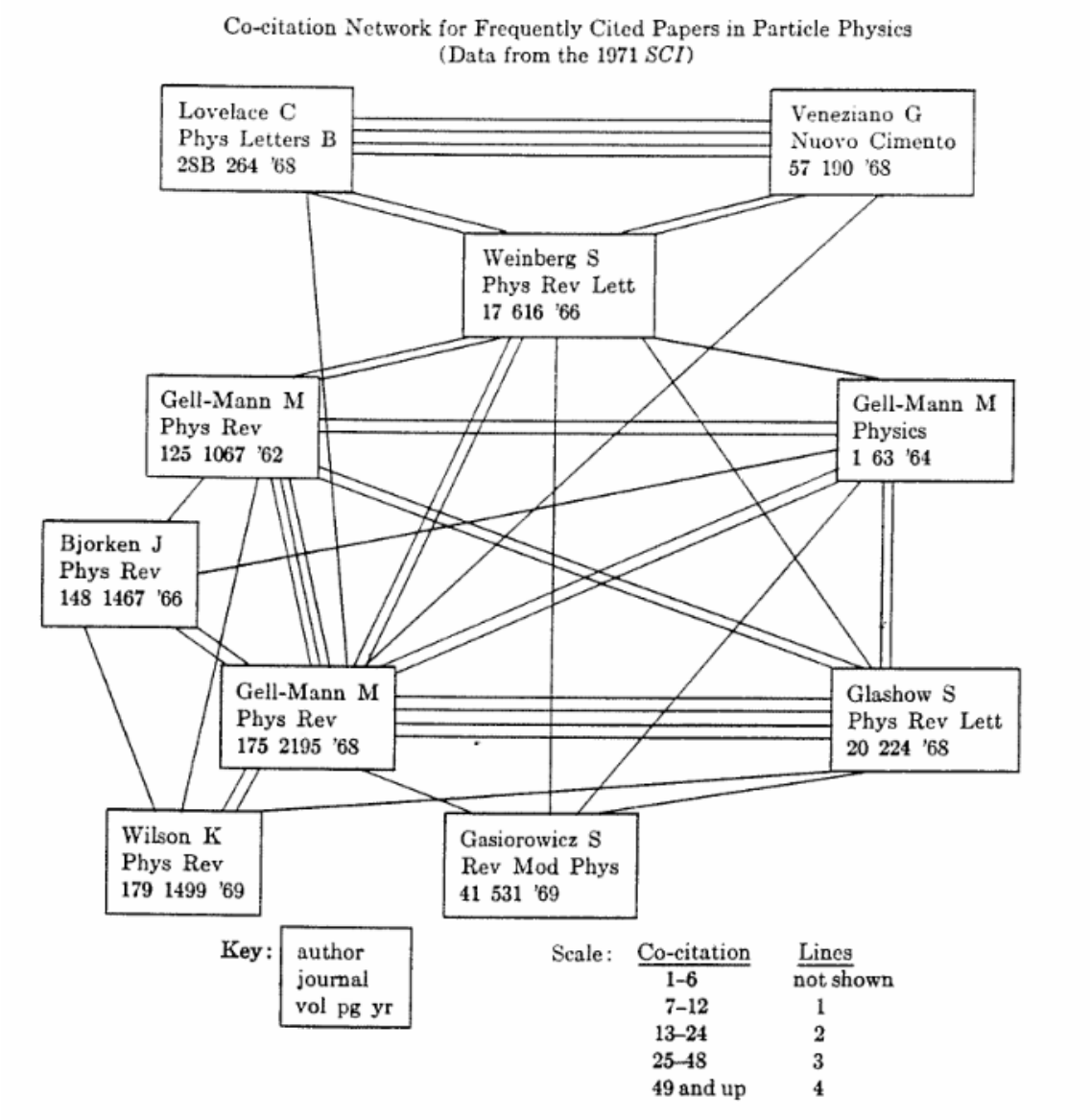
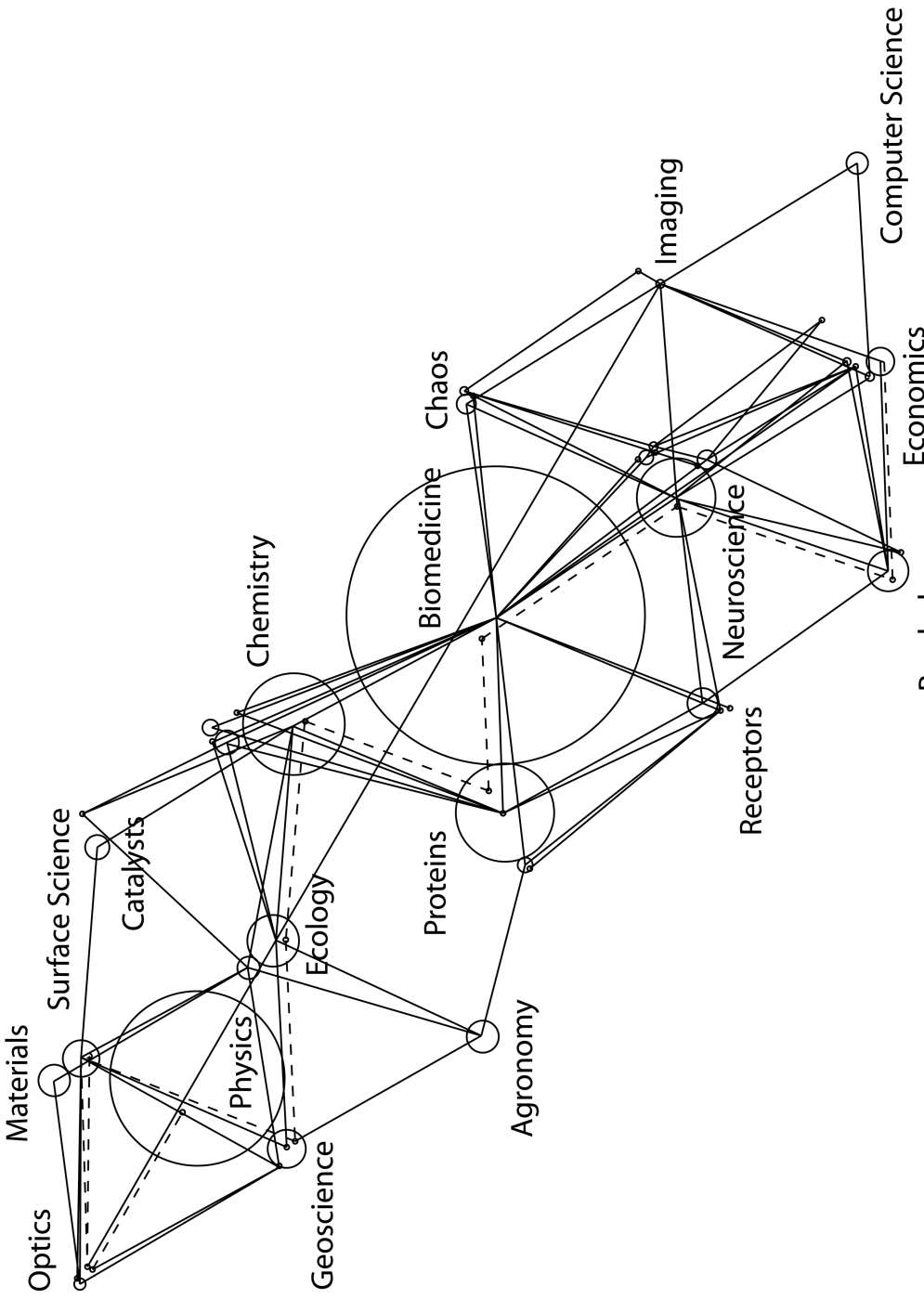



**Figura 3.1:** Mapa de cocitación entre *papers* del área de la física. Creado manualmente por Henry Small [1973].

una considerable cantidad de documentos así como un proceso automático para generarlo. Este mapa cristaliza las ideas —desarrolladas independientemente— por Small y Marshakova [Marshakova-Shaikevich, 1973] en los años setentas.




Börner, Katy. *Atlas of Science: Visualizing What We Know*. (2010). The MIT Press. Pg 32.

**Figura 3.2:** Mapa pionero en el sentido de ser el primer mapa construido automáticamente. Creado por Small [1999].

## 3.3. Backbone of Science

Este mapa [Boyack et al., 2005] (ver Figura 3.3) define y marca la pauta de cómo se iba a desarrollar esta área durante la década del 2000. En resumen, los autores evalúan dos formas de crear mapas, siempre basados en *journals*, estas son, co-citación e inter-citación (ver Sección 2.3). Basados en estas dos representaciones de la relación entre *journals*, evalúan varias medidas de proximidad, incluyendo correlación de Pearson, similitud coseno e índice Jaccard. Los autores utilizan como *gound truth* para evaluar las métricas propuestas, la clasificación de Thomson Reuters™ debido al hecho de que es una clasificación construida manualmente y bien puede entenderse como una clasificación elaborada por expertos. Para la representación de co-citación, encuentran que es mejor utilizar Jaccard, mientras que para la inter-citación obtienen mejores resultados con una modificación de la similitud coseno que ellos denominan K50. Después de crear un mapa de *journals*, aplican métodos de clusterización para reemplazar los clusters de *journals* por disciplinas y obtener un mapa a nivel de categorías de la ciencia. Un proceso manual es llevado a cabo para etiquetar los nodos que representan clusters de *journals*. Cuando se observa este mapa, se debe recordar que se miran clusters de *journals* en los nodos, y relaciones de intercitación en los enlaces. En este mapa, es interesante apreciar, por ejemplo, el grado de centralidad de Bioquímica, que es una disciplina muy conectada, lo que está muy vinculado con ser una disciplina particularmente interdisciplinaria. Este método y análisis, es depués extrapolado en otros mapas en los que los autores participan, tales como el mapa UCSD Börner et al. [2012a], o por otros autores, como el mapa de áreas de la computación elaborado por Pham et al. [2011].

## 3.4. El mapa UCSD

Este es uno de los mapas de la ciencia más difundidos. Es una red basada en Bibliographic Coupling que clusteriza *journals* en 554 categorías agrupadas en 13 áreas. El proyecto inicial fue liderado por la consultora SciTechnologies para la University of California, San Diego (UCSD) de donde heredó su nombre. Después de su entrega privada en 2007, fue actualizado en 2010 y 2009, actualizaciones que fueron comunicadas en un *paper* del año 2012 Börner et al. [2012b], año en que también se liberó de forma pública el conjunto de datos y las imágenes asociadas. Este mapa se presenta en la Figura 3.4.

El proceso de construcción del mapa incluyó datos de Institute for Scientific Information (ISI) WoS (2001-2005) y de Scopus® (2006-2009). Como señal entre *journals* utilizaron Bibliographic Coupling a través de las citas entre *journals*, pero también a través de Keywords, con lo que consiguen conformar 18 matrices que después se combinan bajo ciertos criterios en una sola matriz de señales. Estas matrices no incluyen la frecuencia cruda de Bibliographic Coupling, sino un valor normalizado, que tiene la intención de controlar por el tamaño de la frecuencia que tiene una distribución con alta varianza. Posteriormente, para encontrar la matriz de similaridades entre *journals*, aplican una versión mejorada de la similaridad coseno (K50) que no es más que la medida coseno menos el coseno esperado. Un dato relevante, es que aplican un filtro para eliminar similaridades no relevantes, o lo que ellos llaman la cola larga de las similaridades, vale decir, por cada *journal*, consideran solo un número reducido de

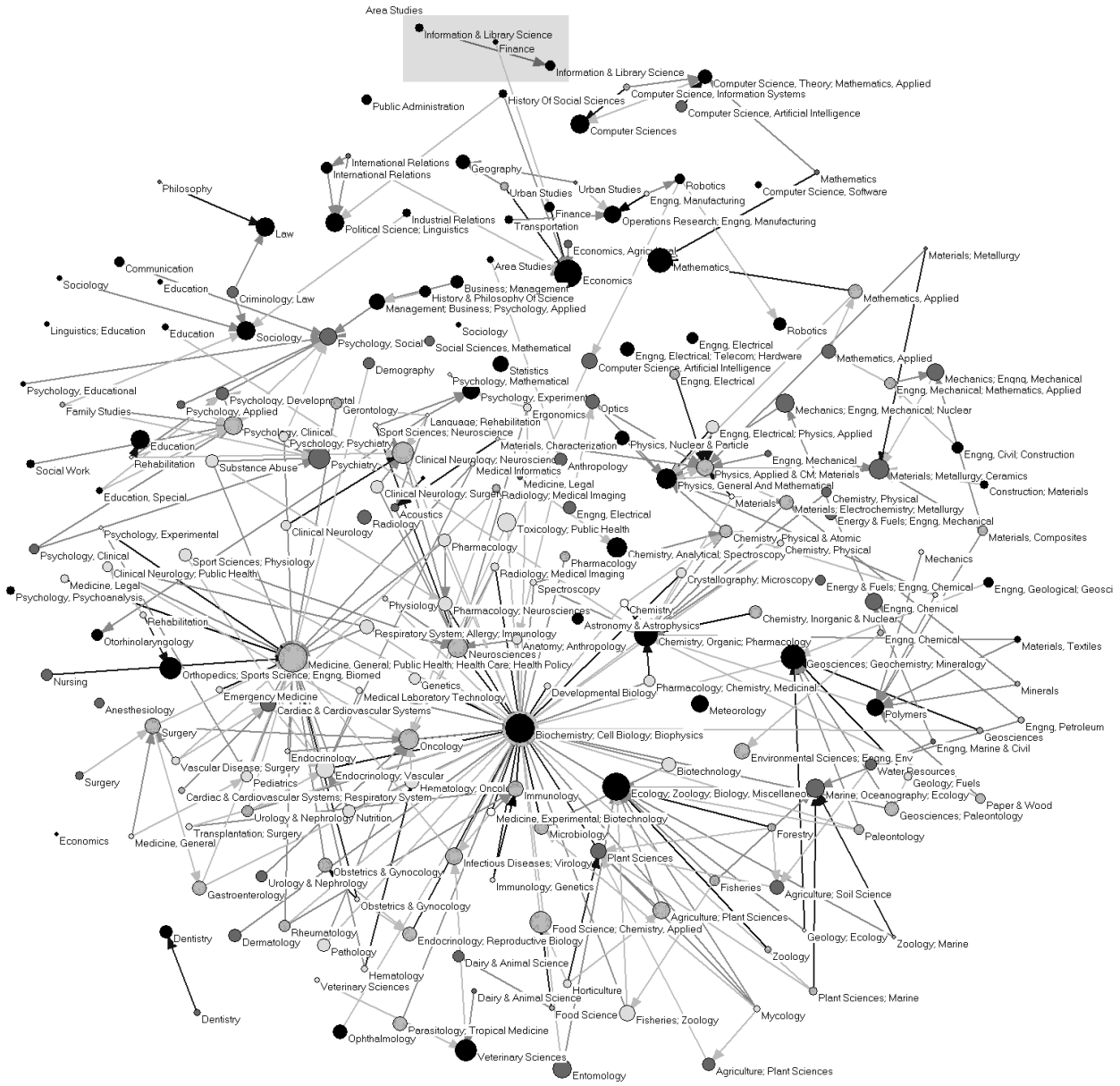



**Figura 3.3:** Mapa *The Backbone of Science* . Los nodos representan categorías de la ciencia y los enlaces representan similitudes basadas en patrones de citas. Fuente [Boyack et al., 2005].

similaridades fuertes con otros *journals*. Eligen entre 5 y 15 similaridades para cada *journal*, basados en el logaritmo de la suma de las citas. Posteriormente aplican un algoritmo de clustering basado en RandomWalk, OpenOrd, para obtener las 554 categorías, las mismas que fueron etiquetadas manualmente, al igual que las 13 áreas principales a las que se asignaron las 554 categorías para obtener un primer nivel de áreas de la ciencia.

Este mapa (3.4) está ampliamente difundido, porque se ha incluido como parte del software VIVO que es muy utilizado por las Instituciones de Educación Superior para gestionar su actividad investigativa, así como también el software Sci, que es un software de escritorio para analizar publicaciones científicas. Este mapa destaca también porque liberó sus datos tanto de matriz de conectividad como de la asignación de *journals* en categorías, lo que facilita la elaboración de nuevos estudios, como el que hemos abordado en esta tesis.

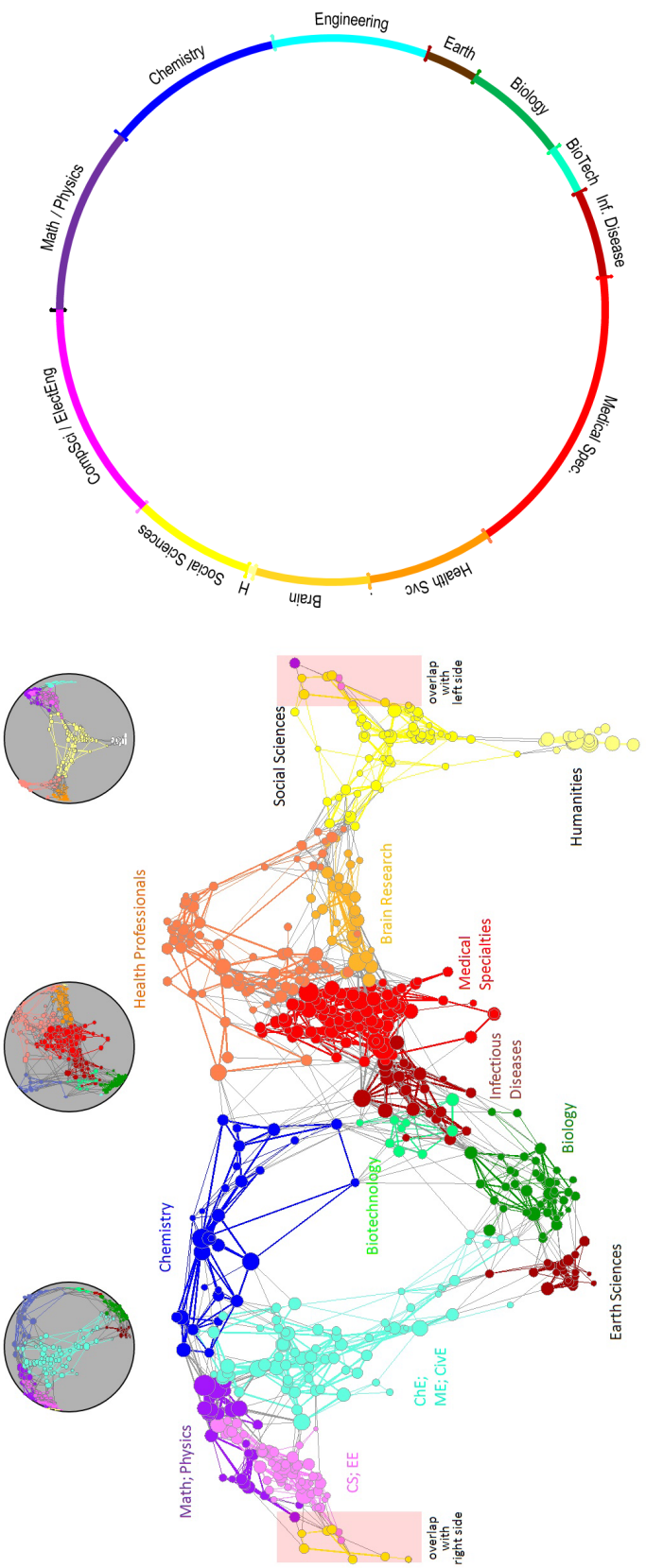



**Figura 3.4:** Mapa UCSD. Incluye 554 clusters de *journals* que se distribuyen en 13 grandes áreas. El mapa corresponde a una proyección en 2 dimensiones del mapa propuesto por los autores que es de tipo esférico (3 dimensiones). Fuente [Börner et al., 2012b].

## 3.5. El mapa clickstream

Este mapa es también conocido como el preferido por *journals* como Science. Está basado en logs de búsqueda de información científica [Bollen et al., 2009]. Una de las principales ventajas de este mapa, según los autores, es el hecho de que la mayoría de bases de datos de publicaciones científicas, están sesgadas en favor de áreas naturales y exactas, y en desmedro de áreas sociales o humanidades, que están menos presentes en esas bases de datos, por cuanto el conocimiento en esas áreas no se genera necesariamente a través de publicaciones en *journals*. La Figura 3.5 presenta el denominado mapa Clickstream.

## 3.6. El mapa basado en random walks

Si bien el trabajo donde fue publicado este mapa de la ciencia por parte de Rosvall and Bergstrom [2008] no tenía como objetivo construir un mapa de la ciencia, sino que ejemplificar el uso de un algoritmo de detección de comunidades en redes complejas; de todas maneras, resulta mandatorio mencionar este mapa, por cuanto es aquel que más explícitamente describe cómo los mapas basados en patrones de citación, son básicamente mapas de flujo de información o conocimiento entre áreas de la ciencia. Esto implica por ejemplo, que este mapa —es uno de los pocos— preserva la direccionalidad de los enlaces (ver Figura 3.6).

Las similitudes entre áreas están basadas en el cálculo de la probabilidad de que un *random walk* salte de un área hacia otra, sobre vías expresadas por patrones de citación, concretamente extraídos del Journal Citations Report® (JCR) de la empresa Thomson Reuters™ del año 2004. El enfoque de detección de comunidades se alcanza en un paso siguiente, cuando se aplica el algoritmo de códigos de Huffman [Huffman, 1952] —proveniente de las Ciencias de la Información— para asignar los nodos a las distintas comunidades.

Un aspecto relevante en el método utilizado, es que los autores filtran previamente *journals* que son altamente multidisciplinarios como PNAS, Science o Nature, argumentando que esos son *journals* que podrían incluir relaciones artificiales entre áreas. Tampoco se incluyeron autocitas.

Respecto del mapa creado, los autores describen que es una red más bien de tipo $U$ que de tipo *anillo*, y si se analiza en detalle se pueden encontrar interesantes coincidencias con el mapa creado por Leydesdorff and Rafols [2009] que utiliza otro método (ver Sección 3.7) para la construcción del mapa, aunque utiliza el mismo conjunto de datos del Journal Citations Report® (JCR), lo que probablemente estaría indicando que la estructura de la red obtenida está fuertemente asociada a los datos más que al método con el que se construye o también, que la clasificación de WoS tiene fuertes coincidencias con una clasificación de tipo *bottom-up* obtenida por el mapa descrito en esta sección.

## 3.7. El mapa de ISI subjects

Este mapa creado por Leydesdorff and Rafols [2009], utiliza datos de inter-citación extraida del Journal Citations Report® (JCR) de 2007, asignando previamente cada *journal*

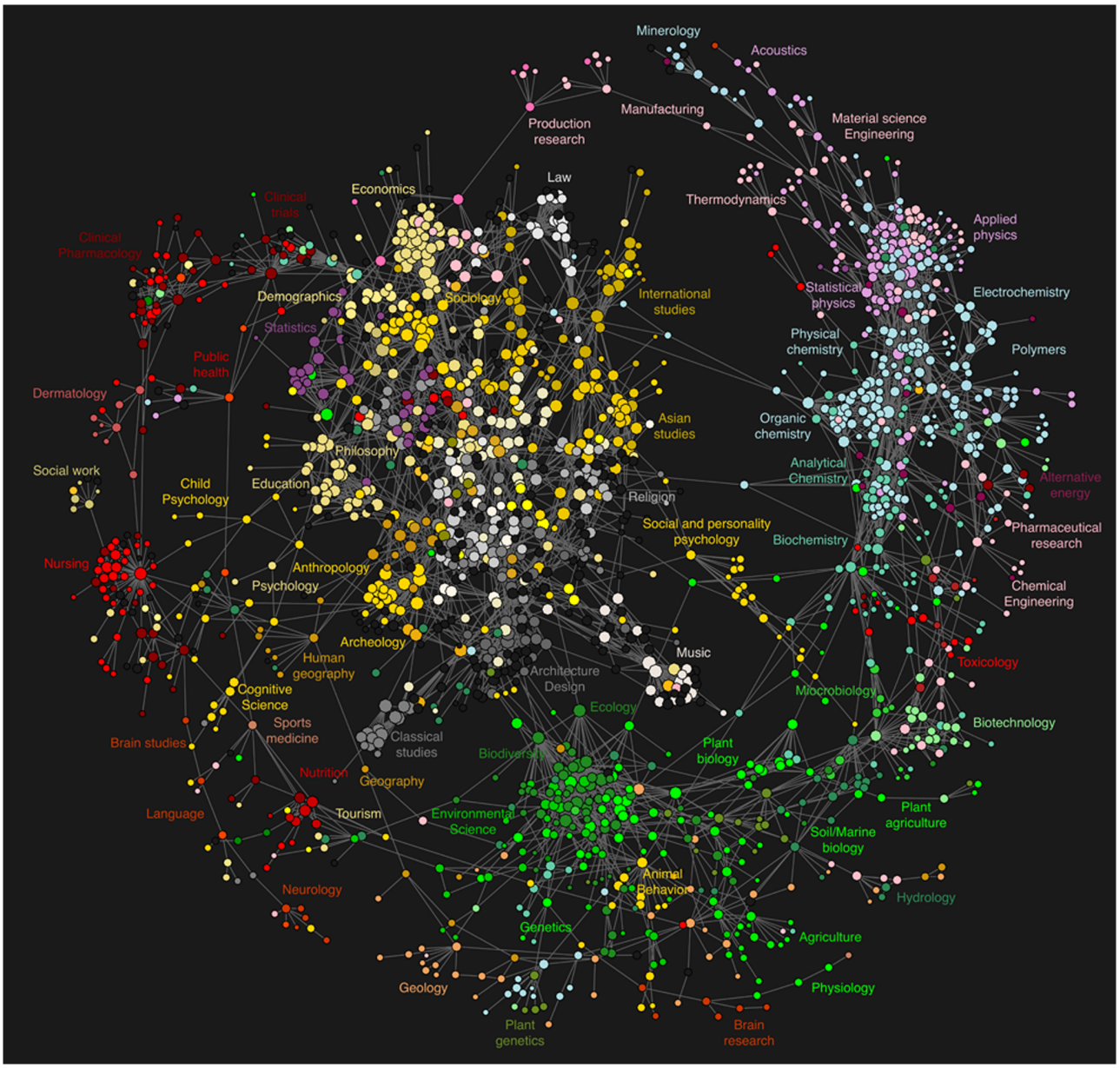



**Figura 3.5:** Mapa *clickstream*. Se contruyó utilizando registros (*logs*) de búsquedas de documentos científicos. Los nodos representan *journals* y los enlaces representan similitudes entre *journals*. Fuente [Bollen et al., 2009].

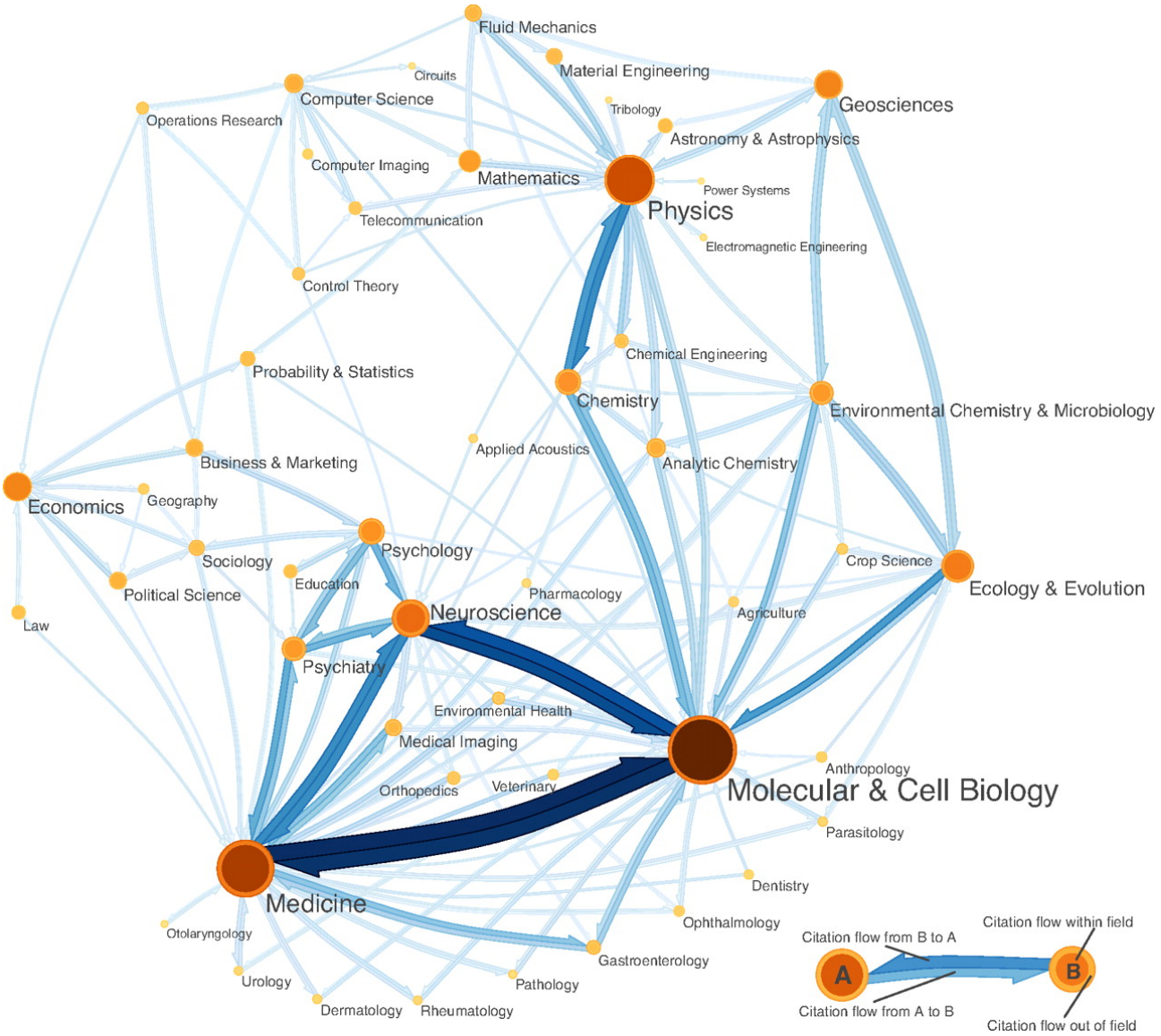



**Figura 3.6:** Mapa que utiliza co-citación entre *journals* y *random walks* como base para encontrar clusters de *journals*. Los nodos representan categorías y los enlaces flujos de citas entre categorías. Fuente [Rosvall and Bergstrom, 2008].

a la categoría en la que es indexado por Thomson Reuters™, de ahí su nombre. Este mapa incluye solamente los índices Science Index ™(SCCI) y Social Science Index® (SSCI), excluyendo del análisis el índice Arts & Humanities Citation Index® (AHCI). Los autores utilizan similitud coseno para determinar la proximidad entre categorías o subjects de ISI y filtran enlaces menores a un umbral de 0.15. Este mapa destaca por ser uno de los primeros mapas liberados en la clasificación de Thomson Reuters™ y también por liberar la matriz de conectividad completa, a diferencia por ejemplo del mapa UCSD (ver Sección 3.4) que solamente libera los datos de los enlaces ya filtrados.

Una de las discusiones a las que este trabajo, le dedica mayor énfasis, es a la *forma* más adecuada que debe tener la ciencia. Vale decir, si debiesen existir áreas más centrales que otras, incluso una topología centro-periferia, o si la ciencia tiene más bien una estructura tipo U o círculo no cerrado con un hoyo en el centro. Los autores argumentan en favor de la segunda opción, en base al mapa resultante que justamente se adscribe a esta estructura.

Este mapa es utilizado como base de al menos otros 5 estudios futuros. Tal vez, el de mayor relevancia hasta el momento, es el trabajo [Rafols et al., 2010] en el que se proponen los denominados *Overlay Maps* como herramienta útil para bibliotecarios y hacedores de política pública. Utilizaremos este tipo de metodología en la Sección 6.1.1.

## 3.8. El Science Brain Scan

Tomando la analogía de un *scanner de cerebro*, la empresa Digital Science, en conjunto con la universidad King's College London, han creado un mapa de la ciencia, sobre la base de la producción-impacto de los proyectos financiados por el Reino Unido. Los datos de base utilizados para construir el mapa corresponden a información textual, entregada por las Universidades del Reino Unido, respecto de cómo esos proyectos impactaron en la socioeconomía de ese país [Digital Science et al., 2015]. Los investigadores y consultores, aplicaron la técnica de *tópic models* para asignar cada proyecto a un tópico en particular, luego de lo cual asignaron los tópicos a un FoR, según la clasificación de 4 dígitos de la Australia-New Zealand Standard Research Classification [Pink and Bascand, 2008]. Con esta matriz de Casos vs FoR, son capaces de construir lo que ellos denominan un *Science BrainScan*, que presenta las relaciones de campos de investigación a través del número de casos que los vinculan [Adams, 2015]. En esta dirección[1], se puede consultar también un mapa a nivel de casos. Los clusters fueron construidos por similaridad de texto entre casos, aunque no se especifican detalles metodológicos de la técnica utilizada. Este mapa se presenta en la Figura 3.8.

[1] www.hefce.ac.uk/pubs/rereports/Year/2015/analysisREFimpact/

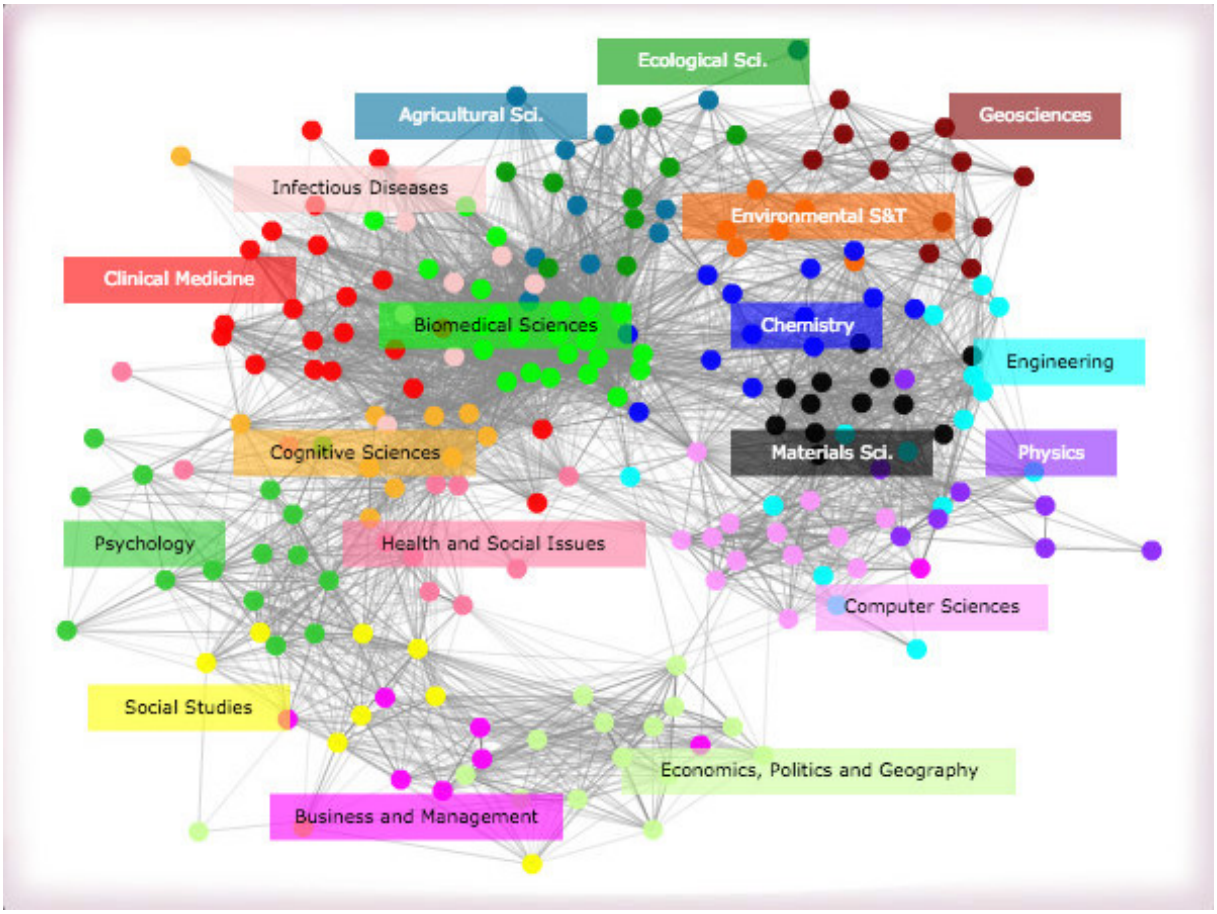



**Figura 3.7:** Mapa de categorías en Journal Citations Report® (JCR). Los nodos representan categorías en la clasificación de Thomson Reuters™ y los enlaces se calcularon utilizando similitud. Fuente [Leydesdorff and Rafols, 2009].

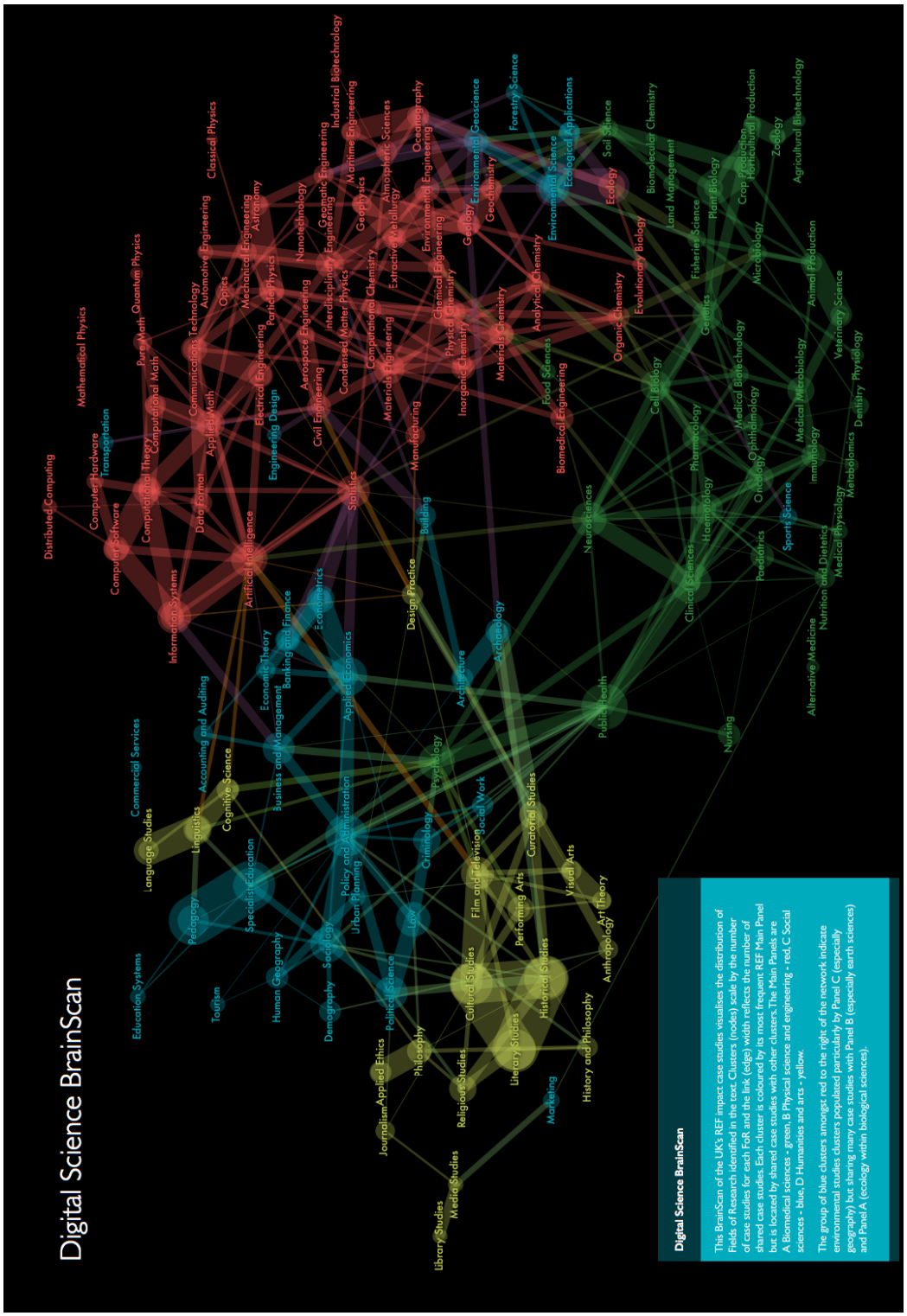



**Figura 3.8:** *Science Brain Scan.* Los nodos representan *FoR*. Fuente [Digital Science et al., 2015].

Este mapa constituye, aunque local al territorio inglés, uno de los más recientes mapas publicados y destaca por haber sido construido sobre la base de datos textuales entregados por los propios productores de ciencia, los misma datos que fueron revisados y filtrados por expertos en el área.

## 3.9. Mapas de journals

Aún cuando nuestro interés central se encuentra a nivel de categorías de la ciencia, existen algunos importantes ejemplos de mapas a nivel de *journals* que es necesario destacar por cuanto los métodos utilizados son extrapolables a mapas a nivel de categorías. En este ámbito, y aún cuando mapas a nivel de *journals* son algo inmanejables tanto visual como computacionalmente (orden de 20 mil nodos), existen algunos esfuerzos en hacer estas redes legibles y comprensibles. Mas aún algunos de los mapas presentados previamente, primero se implementaron como mapas de *journals* y luego se clusterizan para encontrar mapas de categorías de la ciencia, esto se conoce como el enfoque *bottom-up*.

### 3.9.1. The shape of Science

Este trabajo describe el proceso para obtener un mapa basado en *journals* indexados por Scopus® [Hassan-Montero et al., 2014]. Los autores hacen bastante énfasis en el tipo de forma y plantilla (*layout*) que se debe aplicar a una red de este tipo, principalmente citando trabajos en esta materia de Waltman and van Eck [2012]. Los autores descartan varios de los algoritmos de tipo *force directed* por considerarlos interesantes para propósitos estéticos pero poco útiles para representación de similaridades a través de ubicación en el espacio. Los autores también argumentan que la magnitud de este tipo de mapas hace innecesario la representación de los enlaces por cuanto sería ilegible además que la similitud entre *journals* viene dada por su ubicacion en el espacio. Este trabajo es la base del sitio web del mismo nombre (*Shape of Science*) que se encuentra en el URL http://www.scimagojr.com/shapeofscience/ que si bien es un sitio que requiere de mucha memoria del lado del cliente, incluye algunas funcionalidades interesantes como información de superposición de países y la posibilidad de realizar *zoom*. La Figura 3.9 presenta este mapa.

### 3.9.2. VOS Overlay Map

Este trabajo [Leydesdorff et al., 2015] viene a ser el equivalente mapa a nivel de *journals* del previamente publicado mapa a nivel de categorías de Thomson Reuters™ (ver Sección 3.7). Sin embargo, en este mapa, los autores utilizan datos de Scopus®. La medida utilizada es la similitud coseno y se consideró la señal de citación directa como fuente de datos. La representación está construida con el ampliamente difundido software VOSViewer [van Eck and Waltman, 2009] (ver Figura 3.10).

Además los autores entregan una herramienta de escritorio para la superposición de datos sobre el mapa, vale decir para que los usuarios puedan crear sus propios Overlay Maps de forma local.

## 3.10. Mapas de tópicos

Por completitud, con la temática de esta tesis, incluimos en esta sección, ejemplos de mapas, que si bien, no son globales, sí adscriben a la idea de mapear, como en los casos anteriores. Además los métodos que aplican son interesantes y necesarios de mencionar.*journals*

Una creciente tendencia a crear mapas, a un nivel de mayor detalle, ha tomado ventaja de las recientes técnicas de minería de texto, en particular de *Topic Modeling* [Blei et al., 2003]. Con esta técnica es posible, minar el texto de las publicaciones científicas, para encontrar tópicos y posteriormente vincular estos tópicos para definir un mapa de tópicos. Como a nivel de tópico los datos se incrementan sustantivamente, las propuestas actuales de mapas, están limitadas a una disciplina en particular o a un país específico, como medida para reducir la cantidad de texto a procesar.

### 3.10.1. Mapa de Ciencias de la Computación

Utilizando la base de datos DBLP, este proyecto de la universidad de Indiana [Fried and Kobourov, 2014], permite la creación interactiva de mapas del área de la computación. Los mapas propuestos están compuestos de dos submapas, un mapa de base (basemap) y un mapa de calor (heatmap), que se superpone al mapa base. En definitiva conlleva la misma idea de *mapas superpuestos* (*overlay maps*) que se propone en algunos mapas anteriores.

### 3.10.2. Mapa que relaciona tópicos y campos de investigación

Utilizando datos de WoS investigadores de Finlandia [Suominen and Toivanen, 2016], han realizado un análisis de la obtención de tópicos, versus una clasificación tradicional de campos de investigación (FOS de la OECD). Si bien la propuesta y fundamentación realizada son interesantes, por las mismas razones esgrimidas al inicio de esta sección, la validación de lo obtenido y la disponibilidad de los datos, son insuficientes e inexistentes. Los términos principales, asociados a los 60 tópicos obtenidos se liberaron como imágenes de nubes de texto, sin información cuantitativa, que pueda ser realmente utilizada. También la red obtenida, es una red bipartita de enlaces entre FOS y Tópicos, cuando lo necesario hubiese sido contar con una red entre FOS, y la matriz de similaridad entre ellos. Lo publicado, es solamente una imagen de este mapa y se muestra en la Figura 3.12. A pesar de las debilidades del estudio en comento, se menciona en esta tesis como una técnica que seguramente encontrará asidero en los próximos años para el análisis —en línea—de la evolución de la ciencia y el desarrollo de los productores.


**Figura 3.9:** Mapa la Forma de la Ciencia (*Shape of Science*). Basado en la información de Scopus publicada en SCImago. Fuente [Hassan-Montero et al., 2014].

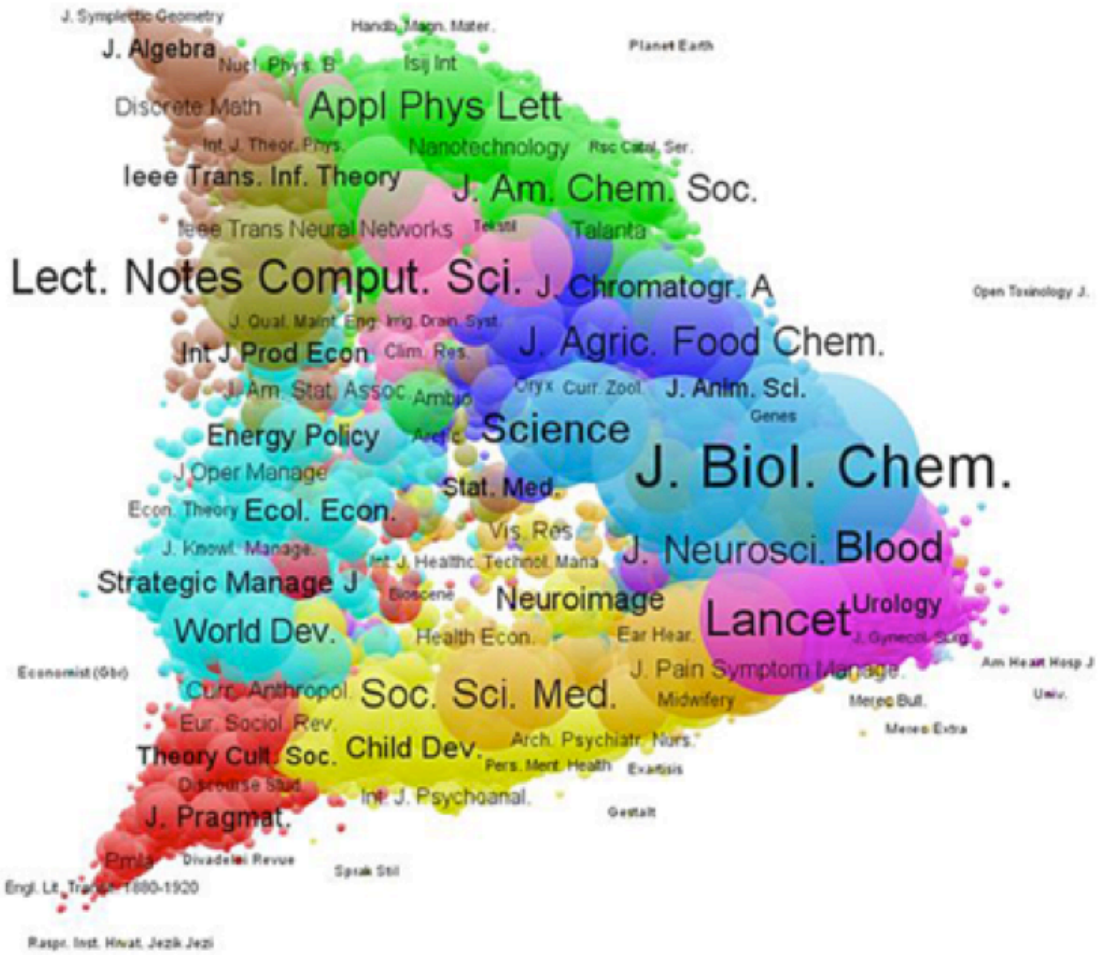



**Figura 3.10:** Mapa de *journals*, utilizando la interface VOSViewer. Fuente [Leydesdorff et al., 2015].

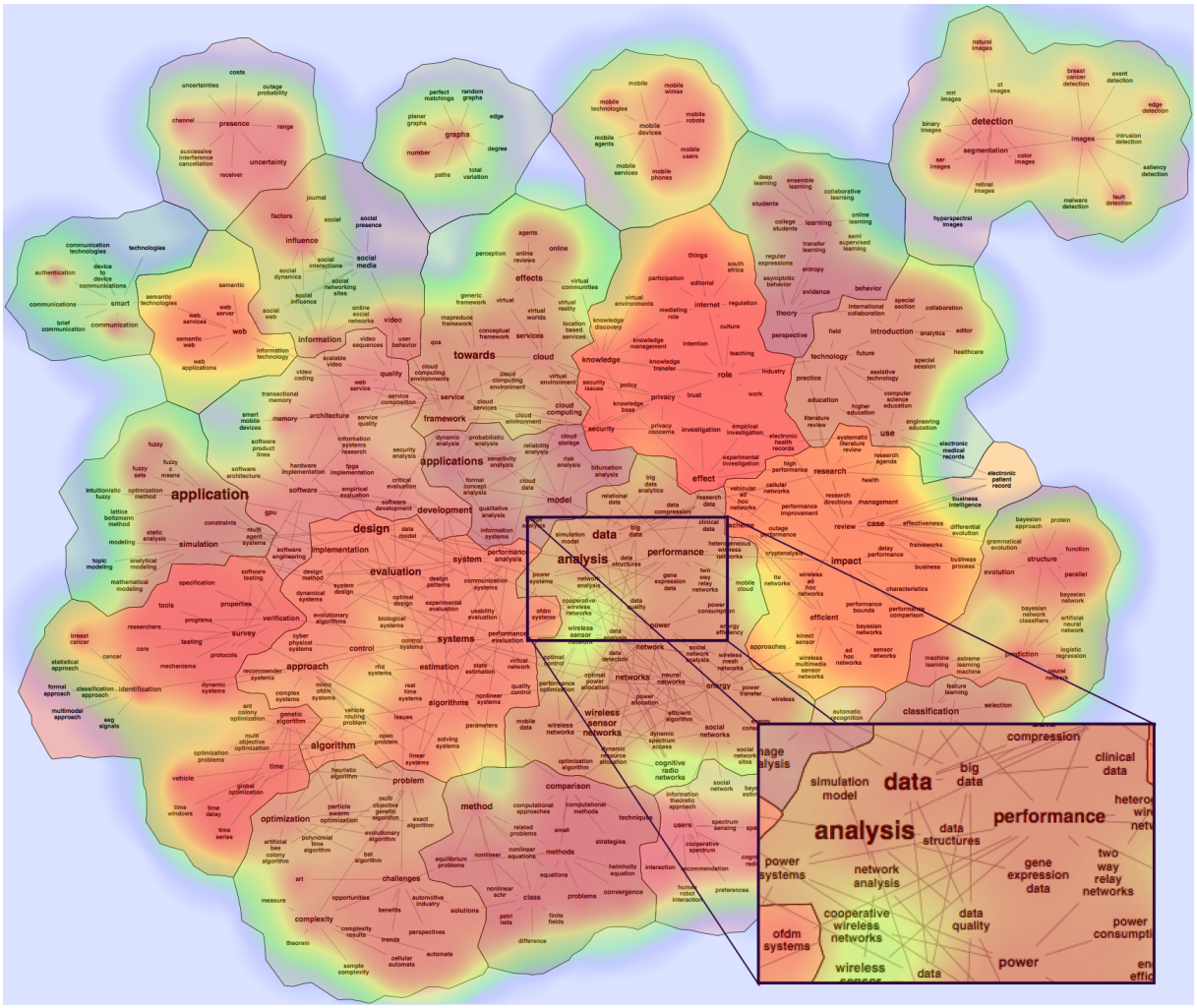



**Figura 3.11:** Mapa de las áreas de la computación. Mapa base incluye información de DBLP desde 1954 hasta 2013. El mapa de calor superpuesto incluye datos de 2013. Hemos incluido un detalle ampliado, para mejorar la comprensión de esta imagen que fue creada utilizando la aplicación web de [Fried and Kobourov, 2014].

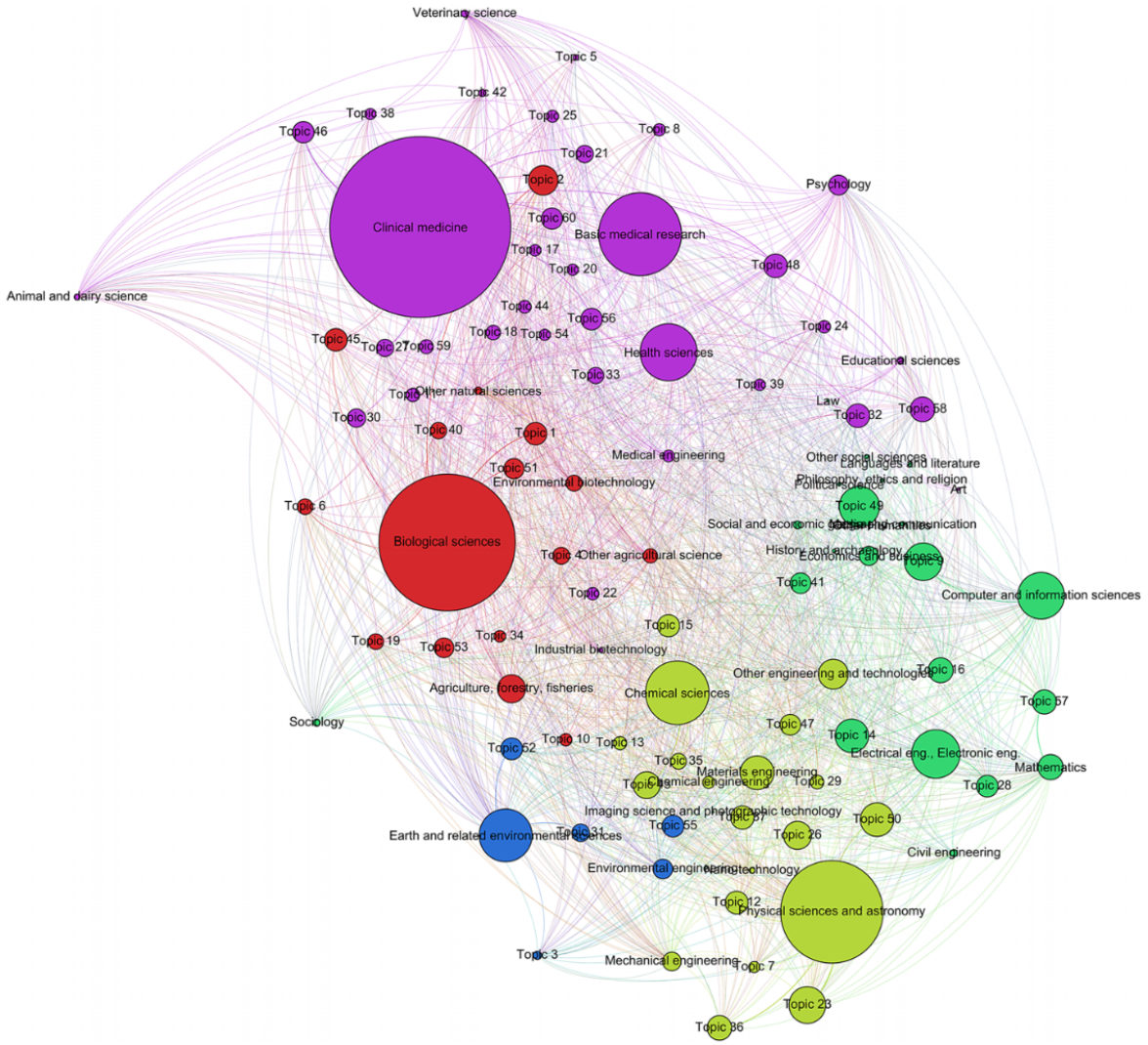



**Figura 3.12:** Mapa de tópicos y *Fields of Science* de la OECD. Fuente [Suominen and Toivanen, 2016].

Capítulo 4

# Propuesta: El Espacio Investigación

En este capítulo detallamos nuestra propuesta de mapa de la ciencia, al que denominamos *espacio investigación* (en inglés *Research Space RS*) y que está basado en las capacidades productivas de los autores. Primeramente describimos cómo construimos la señal para capturar las capacidades productivas de los individuos y posteriormente detallamos la metodología para calcular las similaridades entre las categorías de la ciencia utilizando dos clasificaciones, UCSD y SCimago.

## 4.1. Una nueva señal

Como ya hemos adelantado en los capítulos precedentes, los mapas de la ciencia, requieren de una señal que permita capturar su estructura (ver Sección 2.3). En nuestra propuesta, utilizamos una nueva señal que viene dada por las capacidades de los autores de producir en un campo de la ciencia o en otro. Esto es, esperamos que nuestra red, exprese en los enlaces que unen las áreas de la ciencia, las capacidades productivas de los científicos. En la Figura 4.1 presentamos un esquema que resume nuestra idea de enlaces productivos.

Es por esto también, que denominamos a nuestra red, *espacio investigación*, para hacer énfasis en el hecho de que está pensado en reflejar el acto de hacer ciencia (investigar) más que en el interés de develar cómo se relaciona la información científica con la propia información científica.

La idea fuerza que se encuentra detrás, es que los mapas actuales de la ciencia, básicamente son representaciones del flujo de información entre áreas, más que de la diversificación de los hacedores de ciencia a través de estas áreas. Esto queda muy bien ejemplificado, entreo otros, por el mapa de *Random Walks* que describe claramente los flujos de información entre áreas (ver Sección 3.6). Sin embargo, nosotros argumentamos que flujos de información entre *papers* (citas, co-citación, bibliografía emparejada), o entre categorías, no necesariamente representan las habilidades de los científicos de diversificarse (producir), en una u otra categoría. Esto es, un científico de la computación bien puede citar un *paper* en Neurofisiología

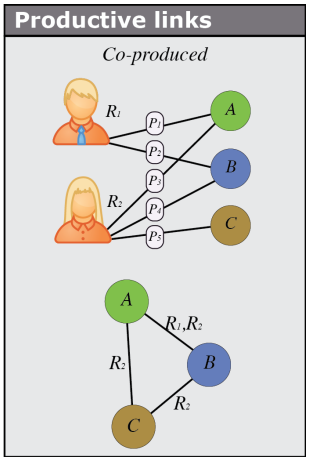



**Figura 4.1:** Esquema de cómo se capturan señales entre áreas basados en las capacidades productivas de los autores. La red que proyectamos y que se muestra en la imagen inferior, es un reflejo de las capacidades de los autores.

(como un *paper* de Redes Neuronales) pero no necesariamente será un futuro contribuidor del área de la Neurofisiología. Vale decir, citar un *paper* en un área específica no demuestra las capacidades del autor para producir en esa área. Por otro lado, ser autor de un *paper* en una área específica sí demuestra las capacidades de un científico para producir conocimiento en un área determinada, así como también, publicar en dos áreas a la vez, puede implicar que esas dos áreas requieren de similares capacidades en los hacedores de ciencia. Sin embargo para esta tarea, se requerirá de datos que no tengan ambigüedades en los nombres de los individuos.

## 4.2. Fuentes de datos

Como hemos argumentado previamente, para la construcción del espacio investigación, necesitamos una matriz de indicadores (señales) de producción científica entre usuarios y categorías de la ciencia. En esta sección describimos los diferentes tipos de datos utilizados y cómo construimos nuestro conjunto de datos de experimentación.

### 4.2.1. Producción científica de individuos

La primera etapa es contar con una fuente de datos a nivel de individuos, que no tenga inconvenientes de ambigüedad y que disponga de información a nivel global, esto es de las Ciencias Naturales y Exactas, las Ciencias Sociales y las Artes y Humanidades. Nuestra elección fue Google Scholar, por considerar (al año que se inició este proyecto de tesis, 2013) que se encontraba ampliamente difundido y que las desventajas que incluía se podían controlar o mitigar, tales como el caso de usuarios espurios o falsos (ver Sección 1.3.2.5).

Descargamos información de Google Scholar por el período de dos meses, para construir una base de datos con las características requeridas. Debido a que Google Scholar no provee de

una *Application Program Interface (API)* de acceso que facilite la obtención de datos, debimos descargar la información, utilizando técnicas de *web scraping* por un tiempo prolongado (app. dos meses) en atención a respetar los *Términos de Servicio (TOS)* con que Google ofrece esta información públicamente.

El *scraper* que construimos consta de tres módulos principales, un primer módulo, *Manager* es el encargado de navegar por las páginas de coautoría de los científicos y obtener los identificadores principales de usuarios que se agregan a una cola. El Algoritmo 1 presenta la estructura general del módulo Manager. Nótese que no disponemos a priori de la lista

**Algorithm 1** Algoritmo general del módulo Manager.

**Require:** $u_1, u_2, ..., u_n$, usuarios semilla
1: $Q = \{u_1, u_2, ..., u_n\}$, lista de identificadores de usuarios a visitar
2: $V = \emptyset$, lista de identificadores de usuarios ya indexados
3: **while** $Q \neq \emptyset$ **do**
4: $u \leftarrow pop(Q)$
5: Consultar la página $p$ de coautores de $u$ utilizando la URL de Google Scholar
6: $V \leftarrow V \cup \{u\}$
7: Parsear $p$ para extraer identificadores de coautores de $u$
8: $C \leftarrow$ lista de identificadores de coautores de $u$
9: **for all** $u' \in C$ **do**
10: **if** $u' \notin Q \wedge u' \notin V$ **then**
11: $Q \leftarrow Q \cup \{u'\}$
12: **end if**
13: **end for**
14: **end while**

completa de identificadores de Google Scholar por lo que el trabajo principal del módulo Manager, consiste en construir esta lista. Los identificadores para usuarios que define Google Scholar son strings de 12 caracteres, que pueden incluir tanto letras como símbolos. Por ejemplo, el identificador de Google Scholar de Katy Börner es `YirSp_cAAAAJ`. Conocido este identificador, podemos acceder a la página de sus coautores, utilizando el URL que resulta de concatenar el URL base de Google Scholar y el identificador del usuario (Línea 5). Lo que equivale a la siguiente instrucción:

```
url = 'https://scholar.google.cl/citations?user=' + 'YirSp_cAAAAJ'
```

La Figura 4.2 presenta un ejemplo de la página de coautores de un autor específico, la misma que se utiliza para alimentar la lista de científicos a descargar.

Una vez que se ha obtenido una lista general de usuarios, el módulo *Harvester* se encarga de visitar las páginas de publicaciones de cada autor en la lista $Q$ y procede a guardar a disco todas las páginas (HTML) con la información cruda de publicaciones de los autores y sus datos personales. Siguiendo con el mismo ejemplo, la Figura 4.3 presenta el tipo de página de publicaciones que se almacena en disco.

Finalmente, un tercer módulo *Gatherer* se encarga de leer todas las páginas almacenadas por *Harvester* en memoria secundaria para *parsear* y almacenar los datos requeridos en una base de datos relacional. Este módulo extrae y almacena información personal, indicadores

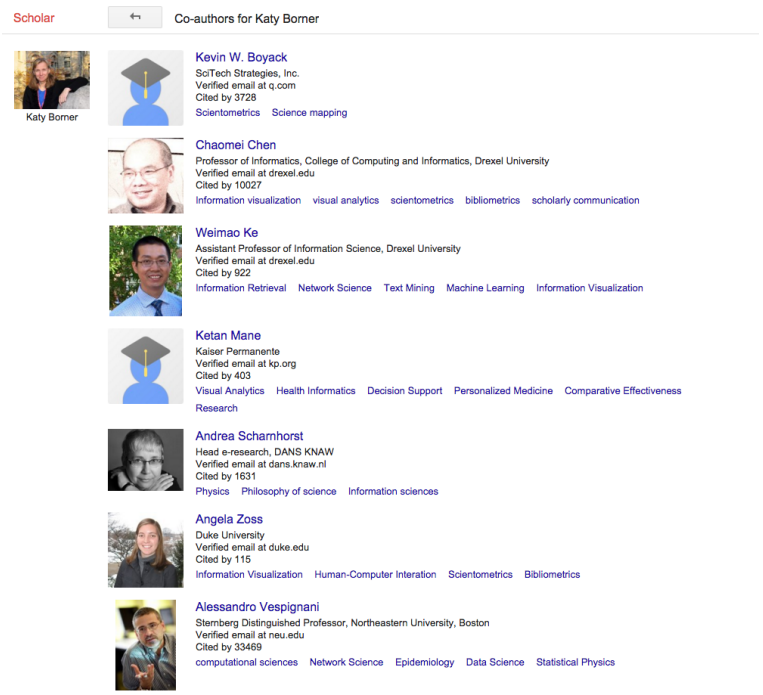



**Figura 4.2:** Página de coautores del usuario con identificador `YirSp_cAAAAJ`, Katy Börner.

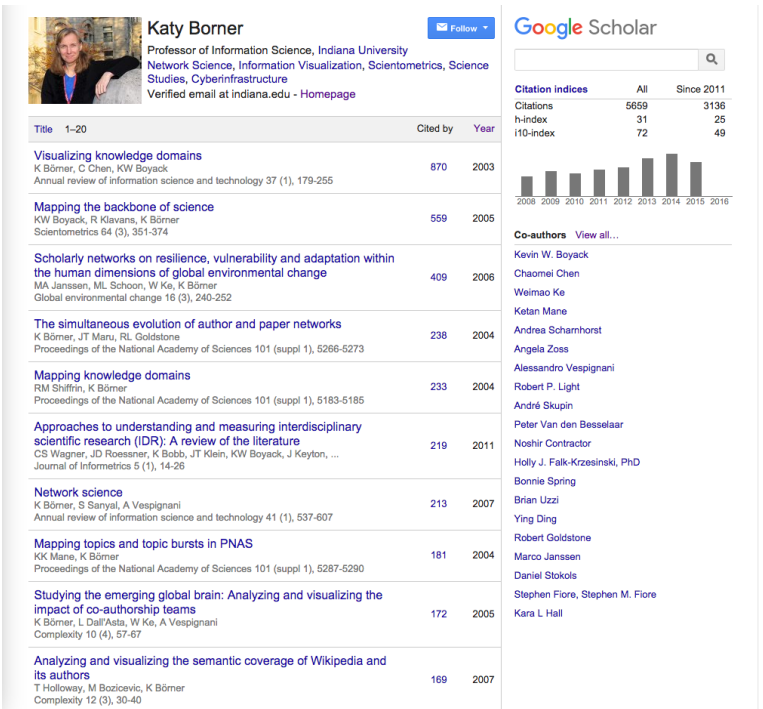



**Figura 4.3:** Página de publicaciones del usuario con identificador `YirSp_cAAAAJ`, Katy Börner.

de citación anual y, principalmente, información de las publicaciones del autor. Es necesario destacar dos campos extraídos, que serán de vital importancia en nuestro estudio. El primero es el nombre de los *journals* de las publicaciones, el que nos permitirá asignar la publicación a una categoría de la ciencia. Y el segundo, es el dominio del correo electrónico del usuario, lo que nos permitirá asignar el usuario a su institución de origen (Nótese que los identificadores únicos de Google Scholar para Instituciones, no aparecieron hasta el año 2016, cuando este trabajo de tesis ya se encontraba en etapa final).

Los nombres de los módulos, así como la idea básica detrás del crawler están adaptados del trabajo doctoral de Carlos Castillo [2004, p. 49-50].

### 4.2.2. Limpieza y dimensiones del conjunto de datos

Es importante notar que tanto la lista de coautores, como la lista de publicaciones, son *ofrecidas* por Google Scholar al usuario y no añadidas automáticamente a su cuenta. El usuario registrado en Google Scholar, debe validar la información de coautores y publicaciones ofrecidas por el motor de búsqueda, lo que permitirá contar —en la mayoría de los casos— con información revisada por —miles de— humanos, lo que en inglés se denomina *crowdsourcing*.

El conjunto de datos completo obtenido con el *scraper* incluye $12,445,334$ publicaciones y $358,947$ perfiles de usuarios. De este conjunto de datos, primero filtramos todas aquellas publicaciones que contenían información inconsistente en el campo fecha, como por ejemplo, aquellas que fechadas en 1900 ó en 2024. También descartamos publicaciones menores al año 1971. Este filtro nos entregó un total de $12,293,468$ entre los años 1971 y 2014. La distribución de publicaciones por año se puede revisar en la Figura 4.4donde se nota claramente una —conocida— tendencia creciente del número de publicaciones por año, lo que no sucede en los últimos años debido a la velocidad de indexación de Google Scholar de nuevas publicaciones así como debido a la demora en la publicación final de nuevos artículos por parte de las editoriales.

Nótese que los usuarios que no han realizado adecuadamente el trabajo de selección de sus publicaciones (para reducir ambigüedades producidas principalmente por nombres similares, que es crítica en el caso de nombres asiáticos), se pueden detectar fácilmente debido al alto número de publicaciones por año. Esto también es fácilmente detectable para perfiles falsos como el conocido caso de Ike Antkare [Labbé, 2010]. Para controlar este tipo de usuarios, miramos la distribución de número de publicaciones por autor por año (Figura 4.5) y decidimos filtrar aquellos usuarios con más de 50 publicaciones en algún año. Después de este filtro, conservamos un total de $319,049$ perfiles y sus correspondientes publicaciones.

### 4.2.3. Indexación de *journals* en categorías

Para asignar las publicaciones de un autor a una categoría de la ciencia, es necesario contar con una base de datos que indexe *journals* en categorías, como las clasificaciones introducidas en la Sección 1.3.3. Hemos utilizado, dos categorías disponibles: SCImago y UCSD (no debe confundirse el mapa UCSD con la clasificación de la ciencia UCSD que deriva del mismo mapa).

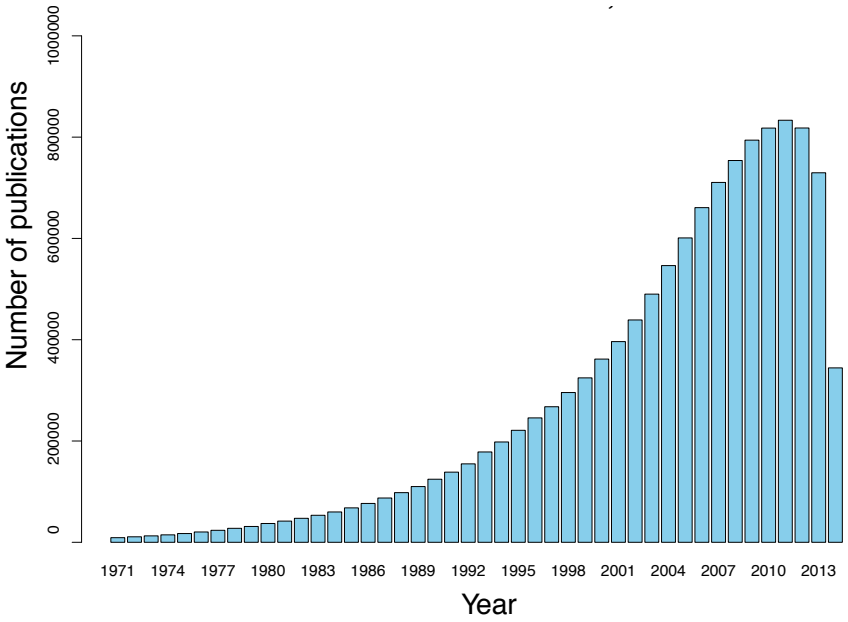



**Figura 4.4:** Distribución de número de publicaciones por año en el conjunto de datos.

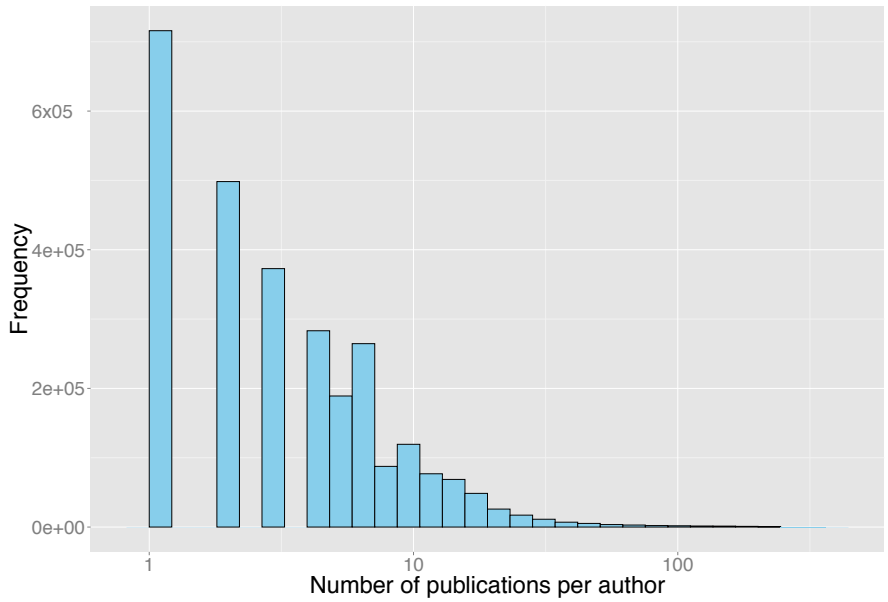



**Figura 4.5:** Distribución de número de publicaciones por autor por año.

#### 4.2.3.1. Clasificación UCSD

La clasificación UCSD, se considera una clasificación *bottom-up*, puesto que se ha definido en base a clusters de *journals* que se obtienen de analizar sus relaciones de citas [Börner et al., 2012b]. Esta clasificación incluye 554 categorías de la ciencia agregadas en 13 áreas principales. El conjunto de datos se encuentra disponible en varios formatos, en el sitio web[1] que se encuentra bajo licencia Creative Commons, Attribution-NonCommercial-ShareAlike 3.0.

El conjunto de datos disponible, entrega tablas de vinculación entre *journals* y categorías. También se entrega un factor fraccionario de asignación que indica la fracción con que un *journal* se encuentra indexado en una categoría. Por ejemplo un factor de 0.25 se puede interpretar como que el *journal* se encuentra indexado en 4 categorías, siendo la suma de factores igual a 1.

La tabla de indexación incluye tanto *journals* indexados en Scopus® como *journals* indexados en WoS.

#### 4.2.3.2. Clasificación SCImago

La clasificación SCImago incluye 236 categorías agrupadas en 27 áreas principales [Gómez-Núñez et al., 2011]. Si bien el conjunto de datos se entrega públicamente —no requiere suscripción— a través del sitio http://scimagojr.com/, no se encuentra disponible una tabla de indexación de *journals* en categorías. Es por ello que hemos construido recuperado los datos utilizando técnicas de web scraping. Debido a que el sitio web de SCImago se encuentra estructurado y los identificadores de categorías son fácilmente detectables, se ha utilizado un scraper simple programado en Python para descargar y almacenar la información de asignación de *journals* en categorías.

Después de vincular las publicaciones en el conjunto de datos con los *journals* en la clasificación, conseguimos descartar publicaciones que pertenecen a la denominada "literatura gris" como libros, conferencias no indexadas o editoriales. Con esta acción también descartamos publicaciones que no son académicas o publicaciones informales que pueden haber sido indexadas por Google Scholar. Después de esta vinculación, nuestro conjunto de datos está compuesto por $4,745,774$ publicaciones. La distribución de publicaciones por año se puede analizar en la Figura 4.6.

### 4.2.4. Vincular usuarios con instituciones y países

Para completar la construcción de nuestro conjunto de datos, utilizamos la información del dominio de la cuenta de correo de los usuarios en el conjunto de datos de Google Scholar para vincular cada usuario con la institución de la que depende laboralmente (se debe notar que la creación de identificadores para instituciones, por parte de Google Scholar se implementó en el año 2016). Para esto hemos accedido al listado público de instituciones en el ranking Webometrics de universidades y centros de investigación. Estos datos está disponibles a través del sitio en http://www.webometrics.info.

[1] http://sci.cns.iu.edu/ucsdmap/

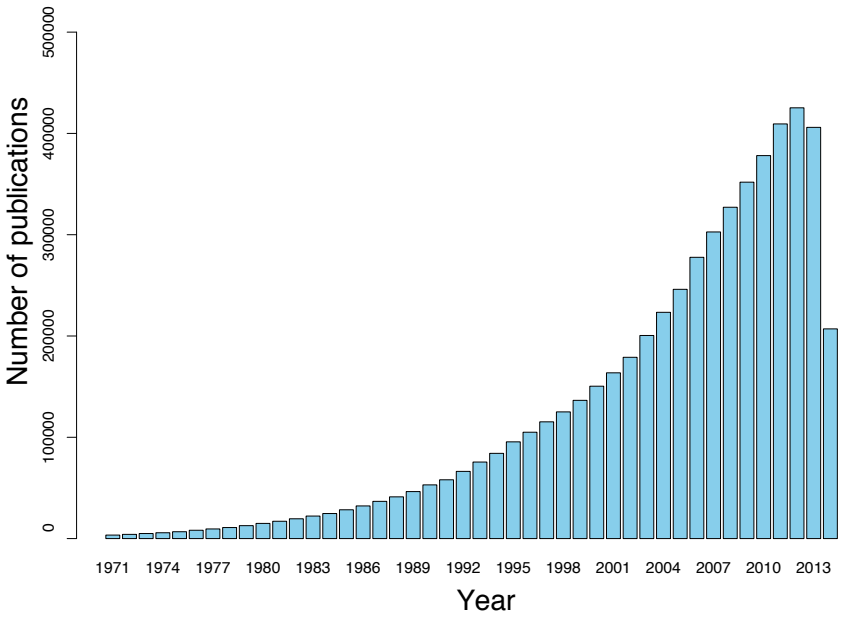



**Figura 4.6:** Distribución por año del número de publicaciones vinculadas a un *journal* en la clasificación SCImago.

Con este procedimiento podemos obtener indicadores a nivel de institución, los mismos que serán utilizados en la Sección 5.

Con la información a nivel de institución, es posible conducir otro proceso de agregación para agregar datos a nivel de países. Este conjunto de datos se utilizará también en la Sección 5 para la validación de nuestra hipótesis.

### 4.2.5. Vincular usuarios con categorías

Una vez que hemos obtenido datos de publicaciones por individuo (Google Scholar) por un lado y, datos de *journals* indexados en categorías (UCSD y SCImago) por otro; se hace necesario poder vincular los dos conjuntos de datos.

Para esto, hemos utilizado expresiones regulares para limpiar la cadena de texto que contiene el nombre del *journal* en el conjunto de datos de Google Scholar, el mismo que se puede vincular con la cadena de texto que contiene el nombre en las tablas de indexación de las respectivas clasificaciones. Hemos dejado en el conjunto de datos, solo aquellas publicaciones con las que fue posible realizar un *match* del 100 %. Nótese también que este procedimiento, descarta publicaciones no científicas incluidas en Google Scholar, debido a que solo incluye aquellas publicadas en *journals* y conferencias identificados en UCSD y SCImago.

Con esto, finalmente hemos podido componer un conjunto de datos que nos permite vincular usuarios con categorías, que es un primer aporte relevante de esta tesis, por cuanto no existen datos disponibles, que entreguen información de la diversidad productiva de los cien-

tíficos o de las instituciones. Solo se pueden encontrar sitios con información de la diversidad productiva de los países, como SCImago, por cuanto la tarea de extracción de información de países desde los *papers*, ha sido una tarea no tan complicada como la de la extracción de nombres o instituciones.

Para destacar el proceso de agregación y las bondades de nuestro conjunto de datos, en la Figura 4.7, presentamos una secuencia de treemaps que muestran cómo se van agregando los datos para un autor en particular, desde sus publicaciones hacia las categorías y hacia las áreas de la ciencia. Los treemaps nos ayudan a distinguir tanto los diferentes tipos de categorías, como la cantidad de publicaciones en cada categoría, la que se refleja en el valor de la superficie de cada caja.

## 4.3. Diversidad productiva

El conjunto de datos que hemos construido nos ha permitido estudiar nuestra propuesta de espacio investigación.

Un aspecto relevante de lo que proponemos, radica en el hecho de que al considerar la producción de ciencia de los individuos, estamos considerando también —implícitamente— su *diversidad* productiva, esto es, su capacidad para participar en equipos de investigación que son capaces de conseguir hacer ciencia en diversas categorías.

Para ilustrar esta idea, en la Figura 4.8, presentamos treemaps para la producción científica de dos individuos en nuestro conjunto de datos. En un análisis breve de diversidad, considerando solo cantidad de tipos (especies) distintas, podemos decir, que el primero, tiene mayor diversidad que el segundo, por cuanto ha sido capaz de publicar en 8 áreas de la ciencia, mientras que el segundo autor, ha tenido un comportamiento más concentrado en solo tres categorías. En nuestra propuesta, ambos autores están *señalando* que para producir en aquellas áreas se necesitan capacidades como las que ellos tienen. Simplificando nuestra idea de *enlace productivos*, cada autor, incrementa la fuerza de un enlace entre dos áreas de la ciencia en las que ha producido. Si recogemos esta señal para todos los autores en nuestro conjunto de datos, entonces podemos construir nuestro espacio investigación basado en capacidades productivas.

En la Sección 6.1, veremos además que los mapas de la ciencia, son útiles para cuantificar diversidad, tomando en consideración otras dimensiones como la *disparidad* entre especies (categorías de la ciencia).

## 4.4. Indicadores utilizados

Una vez que logramos componer el conjunto de datos bruto con información anual de las capacidades productivas de individuos en categorías de investigación, calculamos varios indicadores y agregamos los datos por intervalos de tiempo para la futura definición de la estructura del Espacio Investigación.

Para construir nuestro conjunto de datos de experimentación, debimos colocar nuestros datos, tanto en la clasificación UCSD como en la clasificación SCImago lo que nos permitirá realizar comparaciones con un mapa basado en patrones de citas (mapa UCSD) y también

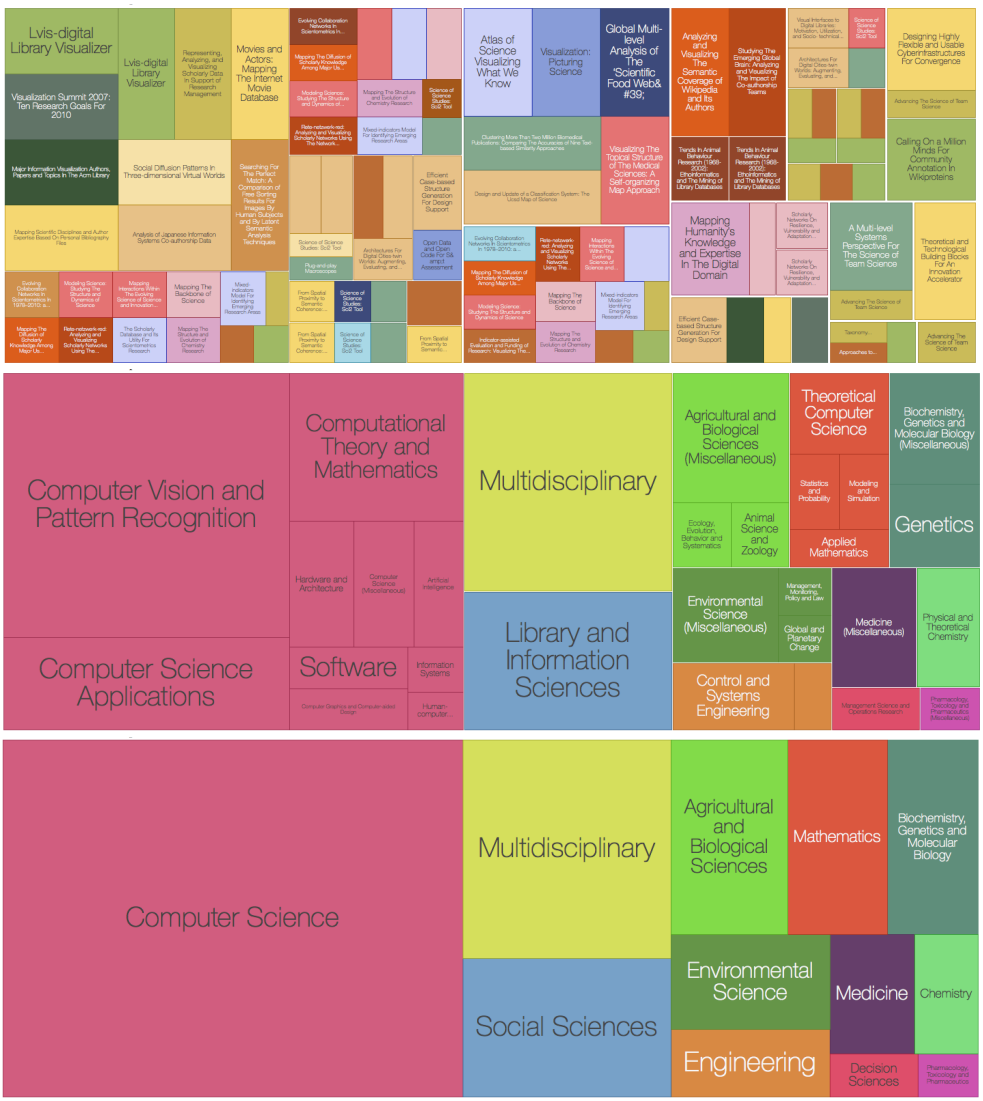



**Figura 4.7:** *Treemaps* que ilustran el proceso de agregación que se puede realizar con el conjunto de datos que hemos construido. De arriba hacia abajo, se muestran publicaciones, que se agregan en categorías que se agregan en áreas de la ciencia, según la clasificación SCImago. El color de las cajas en el treemap de *papers* es aleatorio, mientras que el color de las cajas en los otros dos treemaps, está asociado al área de la ciencia.

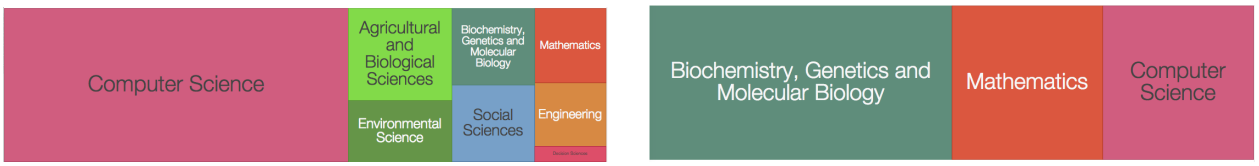



**Figura 4.8:** *Treemaps* para dos usuarios en nuestro conjunto de datos. El tamaño de las cajas es proporcional al número de publicaciones en cada categoría.

verificar la robustez de nuestra propuesta al cambiar a otra clasificación, en este caso SCImago.

Para cada clasificación, definimos el indicador de autoría ${}^{A}X_{ict}$ (*authorship*) como la cantidad de *papers* de un individuo $i$ publicados en *journals* indexados en una categoría $c$ en un determinado intervalo de tiempo $t$.

Nótese que el indicador bruto ${}^{A}X_{ict}$ de la cantidad de autorías de un individuo $i$ en una determinada categoría de la ciencia $c$, puede ser poco adecuado, considerando principalmente dos factores. Primero, el hecho de que el individuo en cuestión sea coautor de un *paper* con un alto número de coautores, implicará que su capacidad productiva para hacer ciencia en esa categoría, si bien existe, es muy probable que no sea tan fuerte, puesto que cuenta con otros muchos colaboradores. Segundo, el hecho de que el *journal* donde se publicó un *paper*, esté indexado en varias categorías implicará que las capacidades productivas del autor se están demostrando, a través de un solo *paper*, en varias categorías de la ciencia a la vez. Este hecho también puede generar una señal ruidosa, sobre todo con *journals* multidisciplinarios que están asignados a más de una categoría.

Tomando estas dos situaciones en cuenta, hemos definido tres indicadores adicionales en nuestro conjunto de datos. El primero ${}^{WA}X_{ict}$ (*weighted authorship*), es la autoría ponderada según el número de coautores en un *paper*. Esto es:

$$ {}^{WA}X_{ict} = \sum_{p} \frac{1}{n_p}, \tag{4.1} $$

donde $n_p$ es el número total de autores del *paper* $p$, además $p$ pertenece al conjunto de publicaciones de ese autor en la categoría $c$.

Como segundo indicador y para tomar en cuenta la múltiple indexación de *journals* en categorías, también consideramos una ponderación $m_p$ correspondiente a la cantidad de categorías en las que el *journal* en que está publicado $p$ se encuentra indexado. Así definimos el indicador ${}^{JF}X_{ic}$ (*journal fractional*) de autoría fraccionaria según el *journal* como:

$$ {}^{JF}X_{ict} = \sum_{p} \frac{1}{m_p}. \tag{4.2} $$

Finalmente, definimos un indicador ${}^{WAJF}X_{ict}$ que pondera tanto por la cantidad de coautores $n_p$ como por la cantidad de categorías $m_p$ en las que está indexado el *journal* donde fue publicado $p$. Este indicador se calcula con la siguiente ecuación:

$$ {}^{WAJF}X_{ict} = \sum_{p} \frac{1}{n_p \cdot m_p}. \tag{4.3} $$

## 4.5. Estructura del Espacio Investigación

En base a los indicadores descritos en la sección anterior, procedimos a realizar nuestra exploración, agregando los datos en diferentes intervalos de tiempo y considerando los tres indicadores. Esta exploración consistió en calcular la matriz de proximidades que se describe en la sección siguiente y en visualizar la red obtenida.

### 4.5.1. Medidas de similitud

Las medidas de similitud consideradas fueron información mutua (o co-ocurrencia) y probabilidad condicional. La elección fue realizada después de evaluar cualitativamente los mapas obtenidos y considerando la ventaja de contar con una medida de similitud que sea de fácil comunicación.

Para cada experimento definimos un intervalo de tiempo $t$ en el que se agregó y calculó el indicador elegido. Así pudimos definir matrices binarias de autoría o producción $\mathbf{P}$, entre autores y categorías de la ciencia, para el indicador utilizado. Las entradas de la matriz $\mathbf{P}$ se definen con la siguiente ecuación:

$$P_{ic} = \begin{cases} 1, & \text{si } X_{ict} > k, \\ 0, & \text{caso contrario,} \end{cases} \tag{4.4}$$

donde $k$ es el parámetro que hemos definido para filtrar la fuerza del indicador $X$. La intención del parámetro $k$ es considerar solo aquellas señales que sean, efectivamente, indicadores de las capacidades productivas del autor $i$ en la categoría $c$. En nuestro caso, $k$ se fijó en 0.1.

Una vez definida la matriz $\mathbf{P}$, podemos calcular las similitudes basadas en información mutua y probabilidad condicional, siguiendo los lineamientos de lo definido en la Sección 2.4.

La matriz de información mutua, se calcula con el producto matricial de $\mathbf{P}$, con su traspuesta. Este producto se define en la siguiente ecuación:

$$M = P^{\intercal} \times P \tag{4.5}$$

En el caso particular en que se desee utilizar la información mutua (cantidad de autores que publican en ambas áreas a la vez) como medida de similitud entre áreas, entonces las entradas de la matriz de similitud $\Phi$, basada en información mutua, corresponden a la entradas $M_{ij}$ de la matriz de información mutua $\mathbf{M}$:

$$\phi_{ij} = M_{ij}. \tag{4.6}$$

En este caso, la información que comparten las categorías $i$ y $j$ son el número de autores que publican en ambas categorías según el indicador de fuerza de publicación.

La matriz de similitud basada en probabilidad condicional, se calcula utilizando la información mutua entre dos categorías $M_{ij}$ pero dividiendo el valor de información mutua por el valor total de una de la segunda categoría $\sum_s P_{sj}$, puesto que la probabilidad condicional no es necesariamente simétrica.

Las entradas de la ecuación de similitud, se pueden calcular utilizando la siguiente ecuación:

$$\phi_{ij} = \frac{M_{ij}}{\sum_s P_{sj}} \tag{4.7}$$

En este caso, las entradas de la matriz $\Phi$ representan cuán probable es que un autor publique en dos áreas al mismo tiempo. Por definición de probabilidad los valores de estas entradas estarán en el rango $[0, 1]$. Nótese que el grafo que se desprende de la matriz $\Phi$ es

un clique, esto es un grafo en el que todos los nodos están conectados entre sí y la matriz resultante es asimétrica, por tanto el grafo es dirigido.

Para los cálculos futuros, utilizamos la matriz a-simétrica, en tanto para las visualizaciones consideramos solamente un enlace entre cada par de áreas de la ciencia. En este último caso, elegiremos el máximo valor entre las dos posibles probabilidades para cada enlace entre dos categorías.

### 4.5.2. Visualización

Una vez que construimos la matriz de similitudes $\Phi$ para el indicador utilizado, en un intervalo de tiempo determinado, procedemos a visualizar la red obtenida.

Debido a que la matriz de similitud es no-simétrica, para efectos de visualización elegimos el valor máximo posible entre dos nodos. Adicionalmente, como la red resultante es un clique, para visualizar la red de manera más comprensible, filtramos enlaces, dejando primeramente los enlaces del *Árbol Recubridor Mínimo (MST)* y posteriormente agregando enlaces de mayor similitud hasta obtener un grado promedio deseado, a esta última visualización la denominamos *red de enlaces fuertes*.

El Árbol Recubridor Mínimo (MST) se obtiene, ordenando los enlaces de forma descendente y agregando enlaces si y solo sí, el grafo resultante es un árbol. Esta técnica, ha sido ampliamente estudiada en series de tiempo financieras (ver por ejemplo [Kim and Wilhelm, 2013; Tumminello et al., 2007]) y en mapas de la ciencia se ha utilizado en el mapa denominado *Scientogram* propuesto por Moya-Anegón et al. [2004] en el que se utiliza para analizar las conexiones fuertes entre áreas de la ciencia.

El Árbol Recubridor Mínimo (MST) es una técnica que además define una estructura inicial para nuestra red, que también se puede evaluar cualitativamente (ver Figura 4.9 ). En nuestro ejemplo, la categoría con enlaces más fuertes es *Neurology*. Es interesante de visualizar por ejemplo el subgrafo formado por el nodo *Diabetes Metabolism* hasta *Food Chemistry*, que pasa por categorías como *Obesity* y *Nutrition*. Otra ventaja del Árbol Recubridor Mínimo (MST) es que produce una componente conexa del grafo conectando todos los nodos del grafo sin dejar islas. La cantidad de enlaces de un MST es igual a la cantidad de nodos menos uno, es decir, la cantidad suficiente de enlaces para conectar todos los nodos. El MST nos permite apreciar en la parte inicial de su estructura que los enlaces más fuertes se encuentran en el área de Neurology Una vez filtrados los enlaces, aplicamos un algoritmo de layout basado en atracción-repulsión, que nos permita organizar los enlaces de una forma más comprensible.

Respecto del color de los nodos, estos se definirán de acuerdo al área preestablecida en el que se agregan las categorías (nodos) y que viene definida por la clasificación, o según la comunidad a la que pertenezcan, de acuerdo a algún algoritmo de detección automática de comunidades.

Respecto del tamaño de los nodos, este será proporcional al grado del nodo, o a la sumatoria de las autorías en esa categoría de la ciencia.

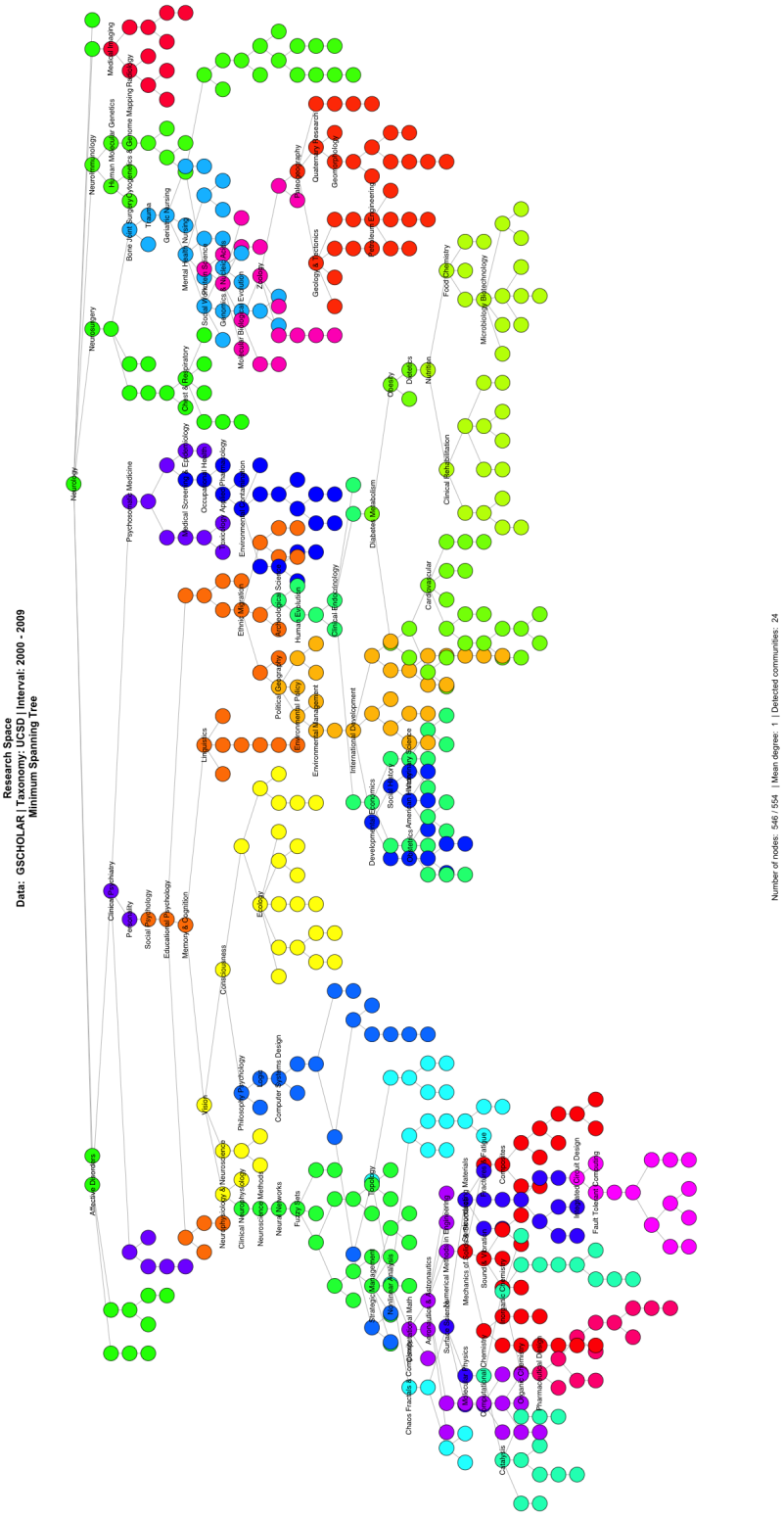



**Figura 4.9:** Ejemplo de Árbol Recubridor Mínimo (MST). Se ha utilizado un *layout* de tipo jerárquico para hacer énfasis en la estructura de este tipo de red. Los colores corresponden a comunidades detectadas automáticamente. Una imagen impresa de alta resolución se puede consultar en el Anexo A.

### 4.5.3. Selección de espacio investigación

Para elegir una sola configuración en la construcción del espacio investigación, exploramos varias combinaciones de configuraciones posibles. Las posibles combinaciones se definen a través de las siguientes variables y características:

- Indicador elegido: se debe elegir entre cuatro posibles valores de $X_{ict}$.
- Filtro $k$ de presencia mínima: se debe definir el valor del parámetro $k$ que servirá para obtener la matriz binaria $\mathbf{P}$ de la presencia del investigador en cada área (ver ecuación 4.4).
- Intervalo de tiempo a agregar. Se dispone de datos entre 1971 y 2014.

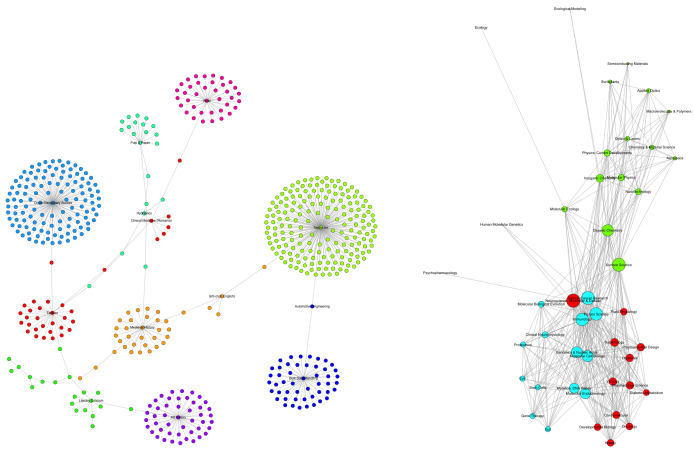
Para la selección final del espacio investigación, conducimos un proceso exploratorio, donde variamos cada uno de los ítems descritos anteriormente. Construimos un total de 135 espacios de investigación. Para evaluar estos espacios de investigación, calculamos la matriz de proximidad y graficamos tanto el MST como diferentes versiones de la red (variando su grado promedio). Inspeccionamos estas redes de forma visual, considerando aspectos como la aparición de áreas tipo *hub*, la conformación de comunidades y la distribución de las áreas de investigación en el espacio.

En la Figura 4.10 presentamos un ejemplo que compara tanto el MST como la red con enlaces fuertes, para los tres indicadores descritos en la Sección 4.4. En esta figura se aprecia con claridad que tanto el indicador de autoría ponderada (Figura 4.10a), como el de *journal* fraccionario (Figura 4.10b), tienden a crear áreas tipo *hub* o el fenómeno de nodos centrales y nodos satélite, lo que se detecta a través de los gráficos de MST (izquierda). Además estos dos indicadores producen espacios de investigación en los que la estructura de comunidades es poco clara, detectándose, en la red con enlaces fuertes (derecha), un máximo de tres comunidades en los dos indicadores. Por otro lado, con el indicador que pondera tanto autoría como asignación del *journal* (Figura 4.10c) se aprecia todo lo contrario, esto es una estructura compuesta de muy pocos *hubs* y una detección de comunidades (en total 8) mucho más clara y concordante con las áreas que agrupan a cada nodo o categoría.

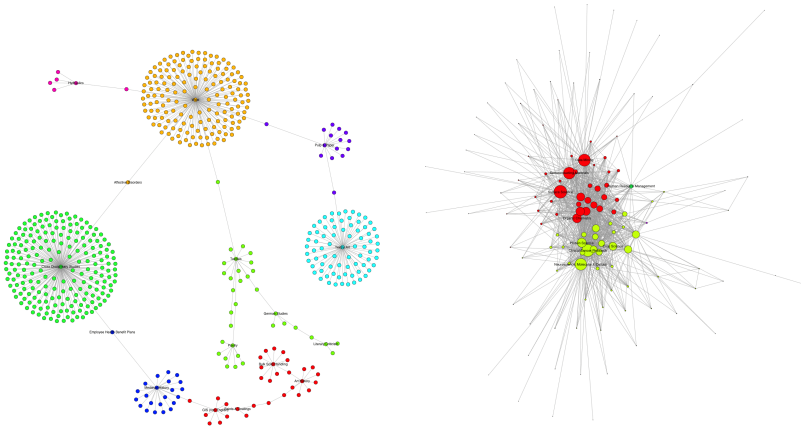
El detalle de las 135 las configuraciones que hemos probado para la construcción del espacio investigación, se puede encontrar en el material digital disponible en el siguiente enlace https://www.dropbox.com/sh/bkjmvkl1zosx9xm/AAB6TS4f0q1dAjHtmk3qfO2ia?dl=0

Luego de este análisis exploratorio, hemos definido la siguiente configuración para los mapas que utilizaremos en lo que resta de este documento:

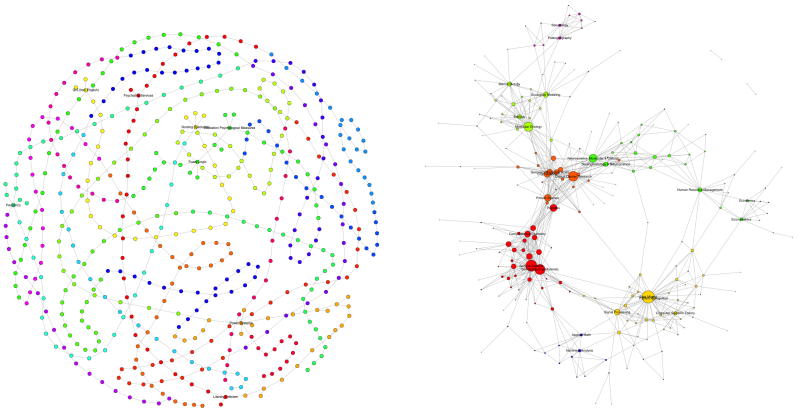
- Indicador elegido: Hemos elegido el indicador ${}^{WAJF}X_{ict}$ que pondera tanto la autoría como la asignación del *journal* (ver ecuación 4.2), al que notaremos de forma más simple como $X_{ict}$ y que denominaremos indicador de la *fuerza de autoría* de un científico en una categoría de la ciencia.
- Filtro $k$ de presencia mínima: La elección del filtro a través del parámetro $k$, nos permite considerar solamente aquellas señales o indicadores fuertes que representan las capacidades productivas de los autores. Sin embargo un filtro $k$ muy elevado puede



**(a)** Indicador autoría ponderada. Filtro $k = 0.25$



**(b)** Indicador *journal* fraccionario. Filtro $k = 0.25$



**(c)** Indicador autoría y *journal* ponderado. Filtro $k = 0.25$

**Figura 4.10:** Comparación entre indicadores. En cada subfigura se presenta una visualización del MST (izquierda) y de la red de enlaces fuertes (derecha). Datos agregados entre los años 2000 y 2009. Los colores representan comunidades detectadas automáticamente. La matriz de proximidad se ha obtenido en base a información mutua. Para la red de enlaces fuertes se han agregado enlaces con valores altos, hasta obtener un grado promedio de aproximadamente 17.

considerar muy pocos datos o solamente usuarios muy consolidados en cada categoría. Sintonizamos $k$ en 0.1, lo que permitió limpiar de nuestro conjunto de datos de señales ruidosas de *journals* indexados en muchas categorías o de autores que publican con muchos otros autores.

Para ejemplificar este valor, consideremos dos situaciones hipotéticas de autores cuya producción en una determinada categoría, se considerarían como 0 en la matriz $\mathbf{P}$. Por ejemplo, consideremos la publicación de un *paper* en conjunto con otros 9 coautores ($n = 10$) en un *journal* indexado en una sola categoría ($m = 1$) (ver ecuación 4.2). Este valor de $k = 0.1$ también se puede ejemplificar como 1 *paper* publicado como autor único ($n = 1$) en un *journal* indexado en diez categorías ($m = 10$). En definitiva, hemos considerado que 0.1 es un valor adecuado para filtrar el ruido producido por hiperautorías o por múltiples asignaciones de *journals* en categorías.

- Intervalo de tiempo a agregar: Hemos procurado elegir la mayor cantidad de datos históricos disponible en nuestro conjunto de datos, pero también hemos reservado suficientes datos para realizar la evaluación de nuestra propuesta. Por este motivo, para la construcción de nuestro espacio investigación, hemos fijado $t$ como el intervalo de años entre 1971 y 2010.

Con la configuración señalada previamente, construimos dos mapas o espacio investigación. El primero en clasificación UCSD y el segundo en clasificación SCImago. Hemos graficado cada uno de estos espacio investigación en las figuras 4.11 (UCSD) y 4.12 (SCimago), en las que, para facilitar la visualización, hemos utilizado el valor máximo de las dos probabilidades entre cada par de áreas. Además hemos conservado los enlaces que pertenecen al MST más los enlaces que tienen una probabilidad mayor al 25 %.

Imágenes en alta resolución al igual que las matrices de proximidad se pueden encontrar en los siguientes enlaces:

- Espacio investigación en clasificación UCSD:
  https://www.dropbox.com/sh/x6srmzmo856hc28/AABAZBKQLn9nltAkXz1b5Gtza?dl=0
- Espacio investigación en clasificación SCIMAGO:
  https://www.dropbox.com/sh/lh4p2ic2tv5qqmu/AACaxmrxjtcLRhIT6ctFt828a?dl=0

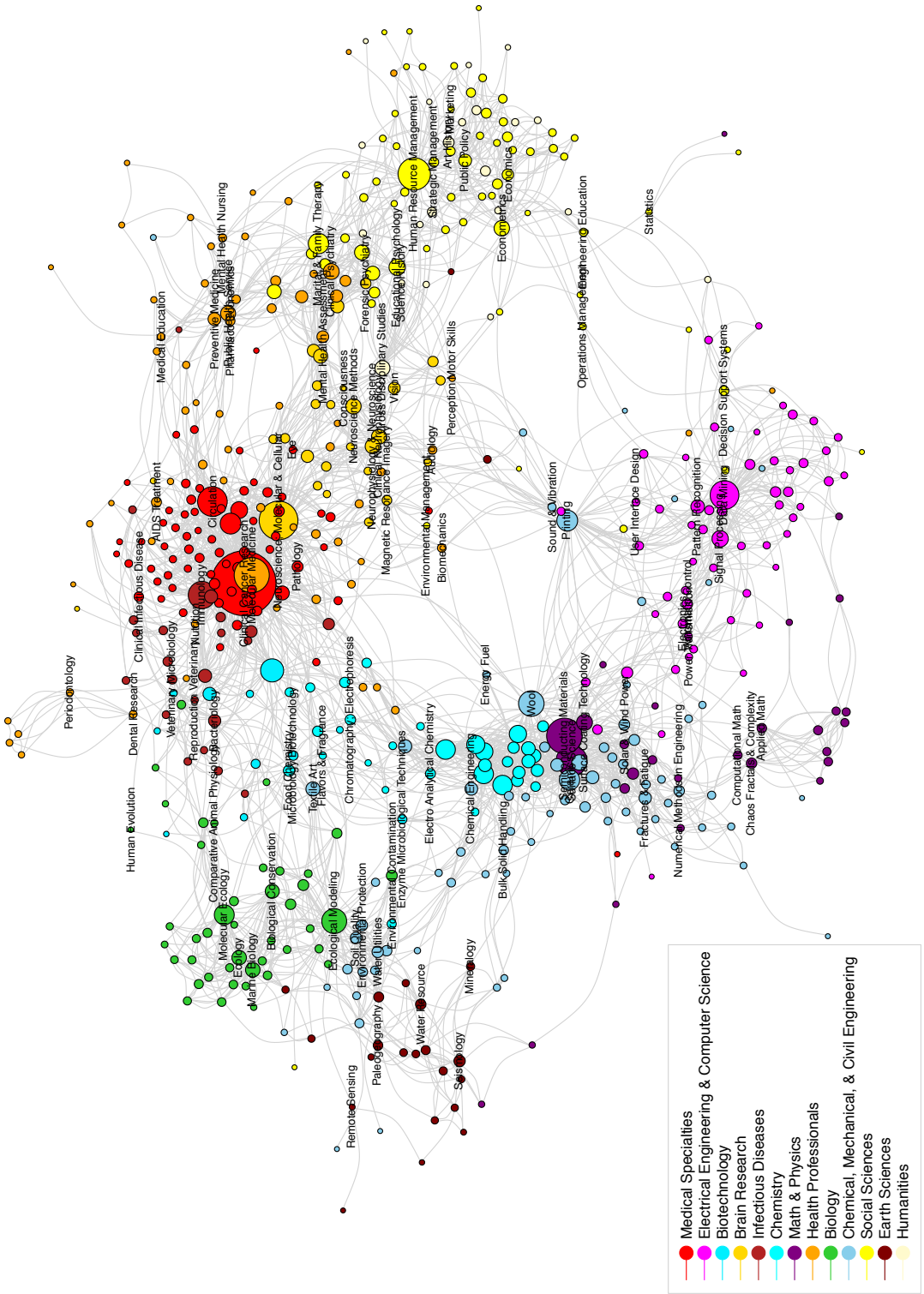



**Figura 4.11:** Espacio Investigación, basado en datos de autoría entre los años 1971 y 2010. Los nodos representan categorías de la ciencia según la clasificación UCSD. Los colores de los nodos corresponden a los colores originales de las trece áreas de la ciencia sugeridas por la clasificación UCSD. Los tamaños de los nodos son proporcionales a su grado. Los enlaces representan la probabilidad condicional de que un autor publique en ambas áreas a la vez. Se filtraron menores a 0.21. Se puede consultar una imagen impresa en alta resolución en el Anexo B.

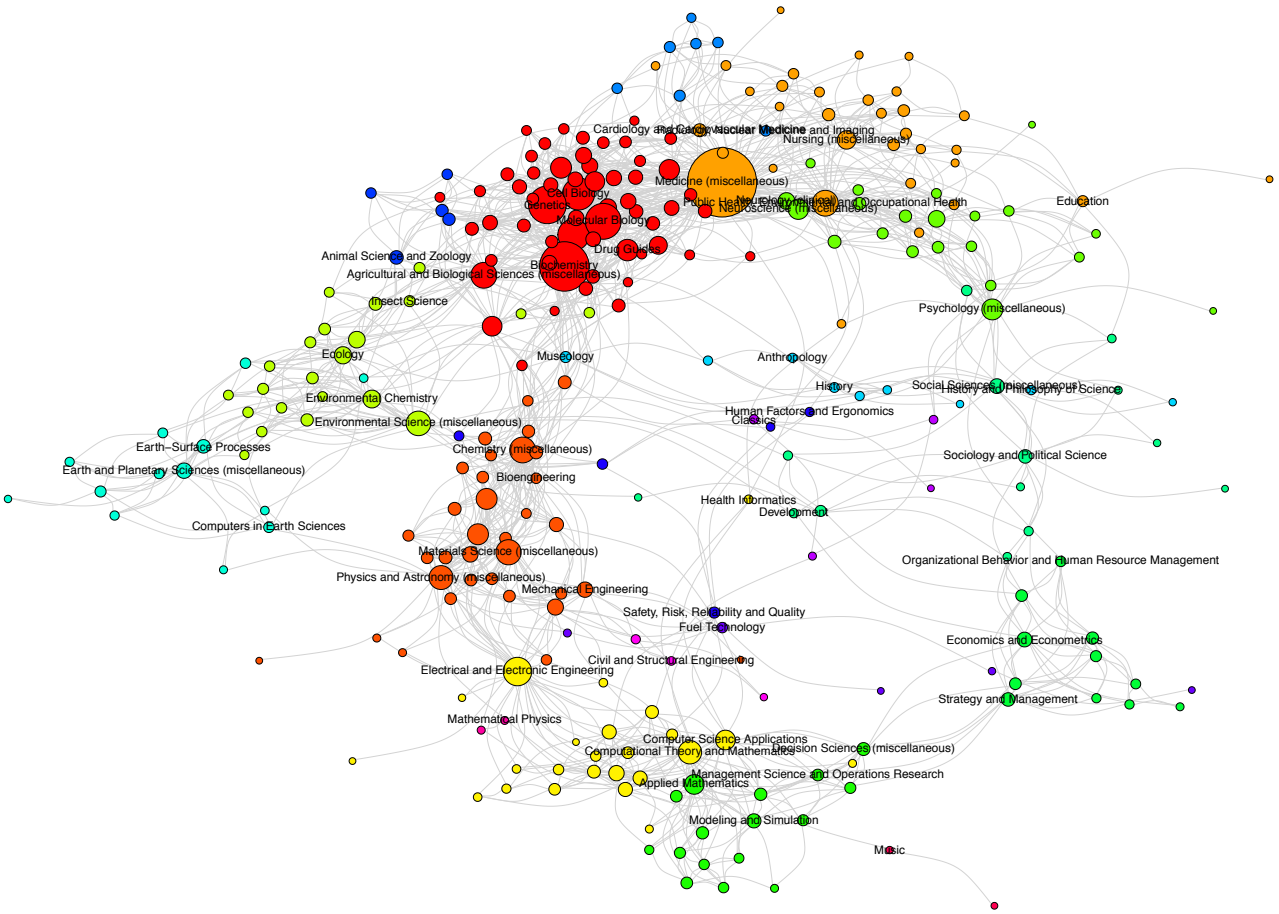



**Figura 4.12:** Espacio Investigación, basado en datos de autoría entre los años 1971 y 2010. Los nodos representan categorías de la ciencia según la clasificación SCImago. Los colores de los nodos representan 28 comunidades detectadas automáticamente por el algoritmo *Fastgreedy*. Los tamaños de los nodos son proporcionales al grado. Los enlaces representan la probabilidad condicional de que un autor publique en ambas áreas a la vez. Se filtraron enlaces menores a 0.383. Se puede consultar una imagen impresa en alta resolución en el Anexo C.

Capítulo 5

# Evaluación y resultados

En este capítulo definimos un método cuantitativo para realizar la evaluación de nuestra propuesta de espacio investigación desde una mirada productiva. Para esto, definiremos estados, transiciones y una me*papers*dida de desempeño, la que nos permitirá comparar nuestro mapa con otros mapas creados con un enfoque distinto.

## 5.1. Evaluación de mapas de la ciencia

La evaluación de mapas de la ciencia, usualmente conlleva una verificación cualitativa de su estructura, vale decir, analizar la cercanía, posicionamiento y conformación de las comunidades del mapa. La utilidad de un determinado mapa, por ejemplo, se valida con la evaluación cualitativa de productores conocidos. Es decir, se construye el mapa de superposición para algún productor específico conocido, por ejemplo, individuo o institución, luego de lo cual se verifica que sus áreas de desarrollo (su territorio) sea consistente con el conocimiento previo que tiene el evaluador respecto del mapa del productor.

En algunos casos esta evaluación cualitativa de la estructura, puede incluir por ejemplo, compararlo con algún otro mapa, generalmente el Mapa de Consenso de la Ciencia [Klavans and Boyack, 2009].

Otro tipo de evaluación puede consistir en evaluar la capacidad del mapa para ayudar a clasificar ciencia, vale decir, qué tan buen clasificador —en áreas de la ciencia— es el mapa, de nuevos *papers* que se van produciendo [Boyack and Klavans, 2014].

Como el espacio investigación está orientado a la producción de ciencia —y no a su clasificación—, nuestra mirada de evaluación del mapa, está orientada a medir qué tan bueno es el mapa para ilustrar dónde están ubicados los productores de ciencia y hacia dónde se puede recomendar que vayan en el futuro. Este camino de evaluación, aún no ha sido explorado en la literatura de mapas de la ciencia y constituye también uno de los aportes destacables de este trabajo de tesis.

### 5.1.1. Estados y transiciones

Localizar la posición actual de un productor de ciencia, sobre un mapa, tiene como objetivo analizar las futuras áreas de producción, análisis que puede demarcar las políticas públicas de una institución, una región o un país; mientras que a nivel personal, las decisiones investigativas de un individuo.

La validación de mapas de la ciencia, ha transcurrido por la utilidad que éstos pueden tener como herramientas de análisis de las capacidades actuales de los productores de ciencia, pero aún no se han enfocado en las áreas que se pueden desarrollar a futuro.

En esta tesis, nos interesa avanzar la evaluación de mapas de la ciencia a una evaluación cuantitativa que permita considerar también su utilidad para futuras inversiones y desarrollo científico. Con esto en mente, la evaluación propuesta para nuestro espacio investigación, se centra en medir qué tan buena es la red propuesta, para recomendar nodos (categorías) de futura diversificación o desarrollo, en el contexto global y también en comparación con otros productores. Para esto, debemos definir previamente, diferentes estados y transiciones para los productores de ciencia. Estos estados y transiciones, se definirán de manera distinta, dependiendo del nivel de granularidad del productor en estudio, por ejemplo individuos, instituciones, ciudades y países.

En relación a considerar la *diversificación* de un productor a lo largo de las áreas de la ciencia, definiremos los estados generales: Activo (A) e Inactivo (I). Activo implica algún nivel de producción en esa categoría, esto es, que el productor en estudio es coautor de alguna publicación en esa categoría (se debe recordar que se han filtrado contribuciones menores. Ver Ecuación 4.4 ), mientras que Inactivo indica que no existe producción en ese nodo (categoría). A mayor activación de nodos en el tiempo, mayor diversificación del productor. Debe notarse que el concepto de diversificación que pretendemos evaluar, está relacionado con las Hipótesis $H_1$ a $H_6$ (ver Sección 1.7) es decir, si el espacio investigación supera a sus competidores en términos de predecir con mayor precisión las áreas de futura diversificación tanto para individuos, como para instituciones y países.

La transición que nos permitirá analizar la diversificación de los productores, la llamaremos *Activación* y la representaremos como O $\rightarrow$ A.

### 5.1.2. Desarrollo y ventajas comparativas RCA

Tanto a nivel de países como de instituciones, interesa analizar el conjunto de datos de producción científica con medidas que sean independientes del tamaño (o normalizadas por tamaño) y también en comparación con los otros productores de ciencia. En esta línea, en el área económica, es común analizar la producción mundial, utilizando la medidad de Ventajas Comparativas RCA [Balassa, 1965]. En los últimos años, esta idea de productividad en el contexto de todos los productores, se ha aplicado a producción científica para analizar el *output* de los países [Abramo and D'Angelo, 2014; Chen and Chen, 2015; Elhorst and Zigova, 2014; Harzing and Giroud, 2014].

El valor de RCA para un determinado productor $i$ en una categoría $c$ es una relación entre la proporción de producción que esa categoría representa para el productor ($S_{ic}$), con la proporción que esa categoría representa para el mundo ($C_c$). Utilizando la matriz de autoría

agregada por institución o por país, podemos calcular las entradas $R_{i,c}$ de la matriz de RCAs $\mathbf{R}$, con la siguiente ecuación:

$$R_{ic} = S_{ic}/C_c, \tag{5.1}$$

Donde $S_{ic}$ es la proporción de la categoría $c$ para el productor $i$, que se puede calcular con $X_{ic}/\sum_c X_{ic}$ y, $C_c$ es la proporción que la categoría $c$ representa para el mundo, y se puede calcular con $\sum_i X_{ic}/\sum_i \sum_c X_{ic}$.

Debe notarse que un valor de RCA mayor a uno, implica una ventaja comparativa de un país en esa categoría, en relación a otros países. Esto se entiende de forma similar para instituciones.

Por ejemplo, si para un país, la producción total de autorías en *Machine Learning* equivale al 20 % de su producción total, y a su vez, *Machine Learning* equivale al 10 % de la producción mundial, entonces se dice que ese país tiene ventajas comparativas en esa categoría en relación a los otros países debido a que $0.2/0.1 > 1$. En el área de Bibliometría, donde este índice se conoce como Índice de Actividad, se dice que el país en estudio es más *activo* en esa área, que el resto de sus competidores.

En relación a analizar el desarrollo de un área para productores del tipo institución o país, definiremos para todo nodo Activo, los estados Naciente (N), Intermedio (I) y Desarrollado (D). Estos estados se definirán según la medida de ventajas comparativas RCA de la siguiente forma:

- Naciente: $RCA \leq 0.5$
- Intermedio: $0.5 < RCA \leq 1$
- Desarrollado: $RCA > 1$

Para evaluar los niveles de desarrollo analizaremos las dos transiciones desde nodos Nacientes e Intermedios que cambian su estado a Desarrollados. Esto es N $\rightarrow$ D y I $\rightarrow$ D. La primera transición determina saltos más grandes en relación a la producción de un país o institución en comparación a sus pares, mientras que la segunda transición representa cambios más moderados, pero no menos importantes.

En la próxima sección, analizaremos en detalle cómo visualizar y utilizar mapas superpuestos como procedimiento previo a la predicción de diversificación y desarrollo.

### 5.1.3. Mapas superpuestos: ¿dónde estoy?

Toda vez que ya hemos construido nuestro espacio investigación y que hemos definido estados para el nivel de producción, podemos responder a la pregunta *¿dónde estoy?*, vale decir, cuáles son las áreas de la ciencia donde un determinado productor tiene mayores ventajas comparativas, así como cuáles son aquellas áreas en proceso de desarrollo o aún no activadas.

Para esto, podemos superponer datos de producción de una entidad productora de ciencia (individuos, instituciones o paises), sobre la estructura del espacio investigación. Esta superposición la realizamos asignando colores a cada uno de los estados. Utilizando una escala cromática de calor, definimos *blanco* para las áreas inactivas (*inactive*), *amarillo* para

las áreas nacientes (*nascent*), *naranjo* para las áreas intermedias *intermediate* y *rojo* para aquellas áreas desarrolladas (*developed*).

A modo de ejemplo, en las figuras 5.1 y 5.2 se pueden analizar y comparar los espacios investigación de India y Holanda. En estos mapas se han etiquetado solamente aquellas áreas desarrolladas y que tienen un alto valor de centralidad para el productor.

En una comparación a nivel macro, se puede observar que para el período analizado, India cuenta con 95 categorías de la ciencia (nodos) aún no activos, mientras que Holanda solamente tiene 34 categorías no activas. En más detalle, se puede notar que India presenta mayores ventajas en áreas de la Física, la Ingeniería y las Ciencias Ambientales, mientras que Holanda se ha desarrollado más en áreas de las Ciencias Sociales, la Neurociencia y el área de la Salud.

Los mapas de la ciencia superpuestos cobraron fuerza en la última década gracias al trabajo seminal de Rafols et al. [2010], aunque en redes de complejidad económica se popularizaron desde el año 2007 [Hausmann and Hidalgo, 2013; Hidalgo et al., 2007]. En esta tesis, hemos avanzado este tipo de visualización, agregando semántica a los nodos a través del uso de colores que diferencian los niveles de desarrollo. Generamos de manera automática, los mapas superpuestos para dos intervalos de tiempo y para todos los productores en nuestro conjunto de datos. Una muestra de estos mapas para instituciones se puede consultar en el Apéndice D, mientras que para países en el Apéndice E.

Identificar las categorías del mapa en las que una entidad produce ciencia, permite también evaluar su *diversidad* productiva. Para facilitar esta tarea, construimos un paquete de software que detallamos en la Sección 6.1, donde se abordan diferentes tipos de medidas de diversidad.

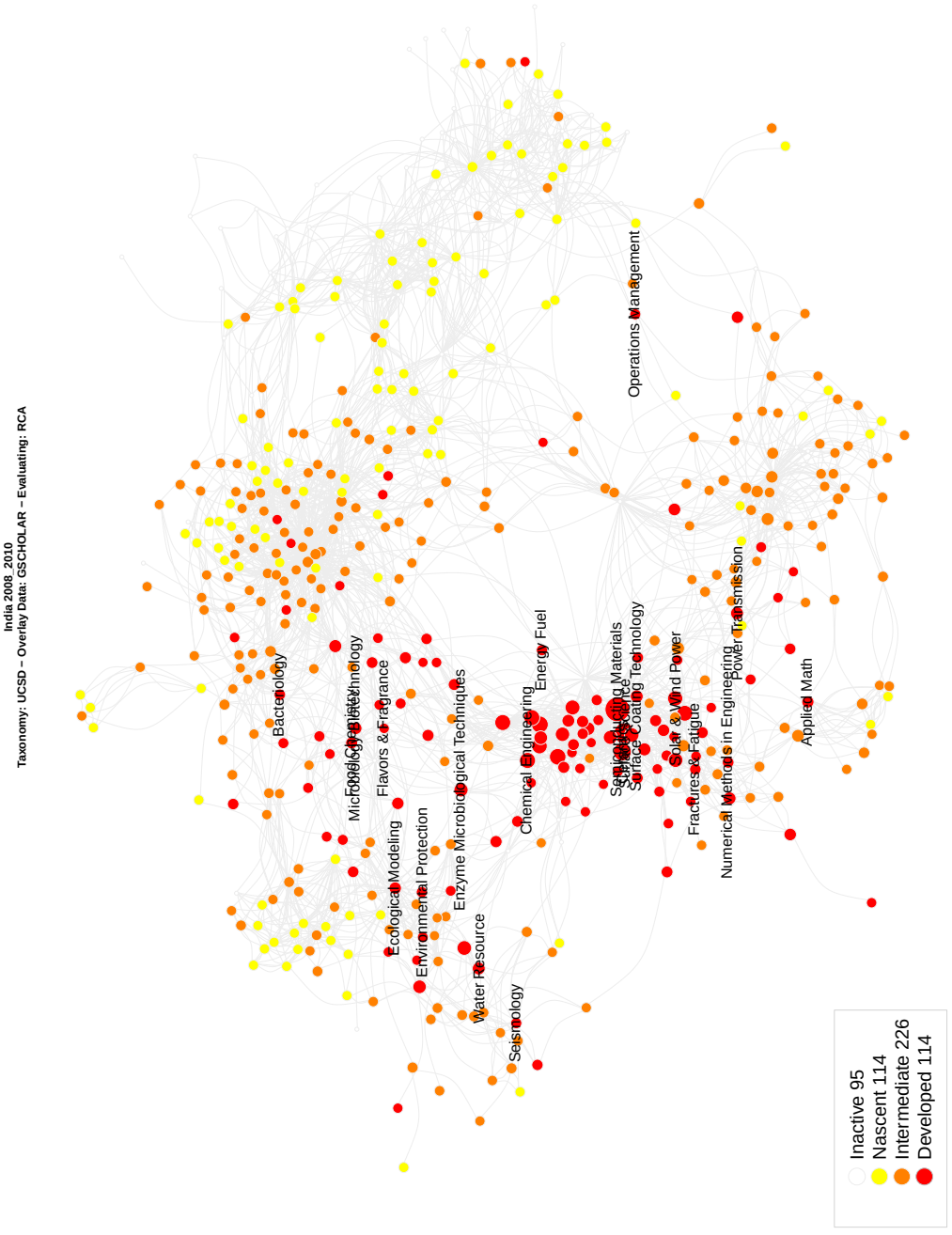



**Figura 5.1:** Mapa de ventajas comparativas de India en el intervalo de tiempo 2008-2010.


**Figura 5.2:** Mapa de ventajas comparativas de Holanda en el intervalo de tiempo 2008-2010.

### 5.1.4. Recomendaciones: ¿a dónde puedo ir?

De los mapas superpuestos, que nos presentan el estado actual de dónde se encuentra un productor de ciencia, podemos aprender que los productores tienden a desarrollarse en áreas cercanas a aquellas que ya se encuentran desarrolladas, algo que es intuitivo y que ha sido demostrado en otras redes, como el *espacio producto* [Hidalgo et al., 2007]. Por ejemplo, si consideramos el espacio investigación de Taiwan, veremos que en el intervalo de tiempo 2008-2010 aún le falta un nodo por desarrollar en las cercanías de la categoría *Sufrace Couting Technology*, este nodo se encuentra rodeado por nodos ya desarrollados (ver Figura 5.3a) y sucede que en el siguiente intervalo de tiempo, este nodo se activa, vale decir Taiwan desarrolla ventajas comparativas en esa área de la ciencia (ver Figura 5.3b).

Esta observación conlleva a la idea fuerza de que es más probable que un nodo inactivo en un período de tiempo $t-1$, se active en un período siguiente $t$, en la medida en que se encuentre avecindado por un alto número de nodos ya activados. Dicho de otra forma, es más probable que los productores tiendan a desarrollarse en áreas cercanas a las áreas en las que ya se encuentran desarrollados.

Definiendo un método cuantitativo adecuado, podremos hacer predicciones respecto de qué áreas tienen mayor probabilidad de activarse o desarrollarse en el futuro, esto es, una lista ordenada o *ranking* de áreas. Basados en estas predicciones se pueden abordar recomendaciones para cada productor. Sin embargo una *recomendación* en el sentido estricto, deberá abordar un estudio mucho más completo, por ejemplo, debe incluir necesariamente otros aspectos de un productor particular en estudio, como su portafolio de productos, su desarrollo de nuevos inventos (patentes) o las políticas a largo plazo y la visión estratégica del productor. En esta tesis, nos abocaremos a predecir la lista ordenada de áreas a desarrollar en un tiempo futuro, como posibles recomendaciones de áreas a potenciar, sin embargo, como hemos dicho, una recomendación acabada requeriría un estudio profundo de cada productor.

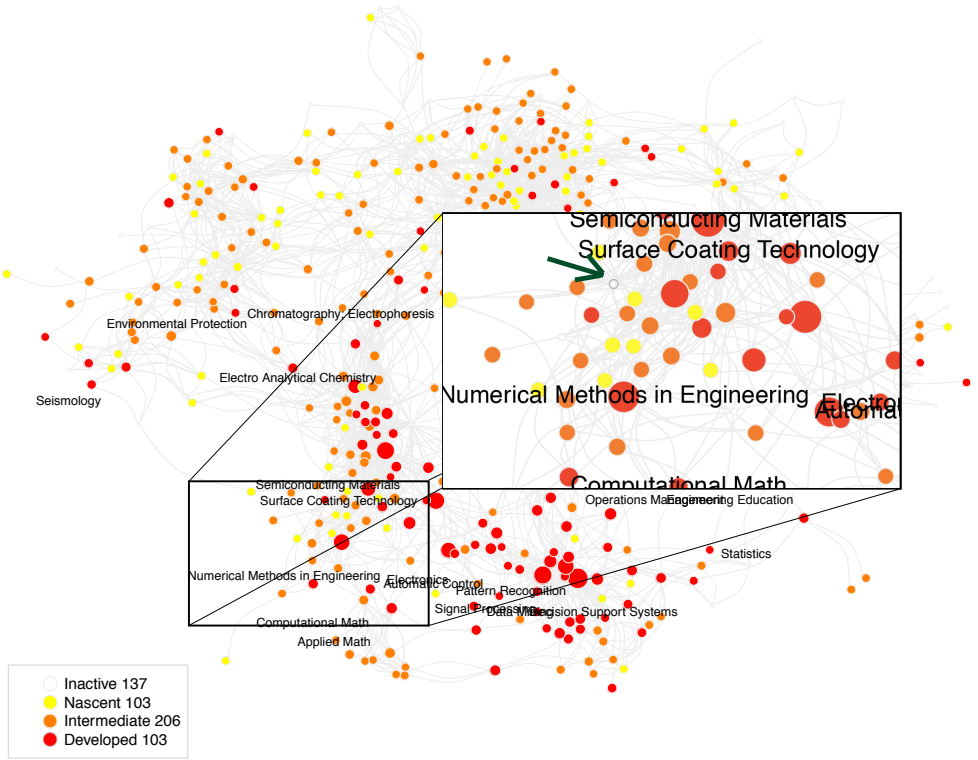
En la próxima sección, proponemos un método para poder evaluar la noción de cercanía con nodos activos.

### 5.1.5. Predicción de transiciones

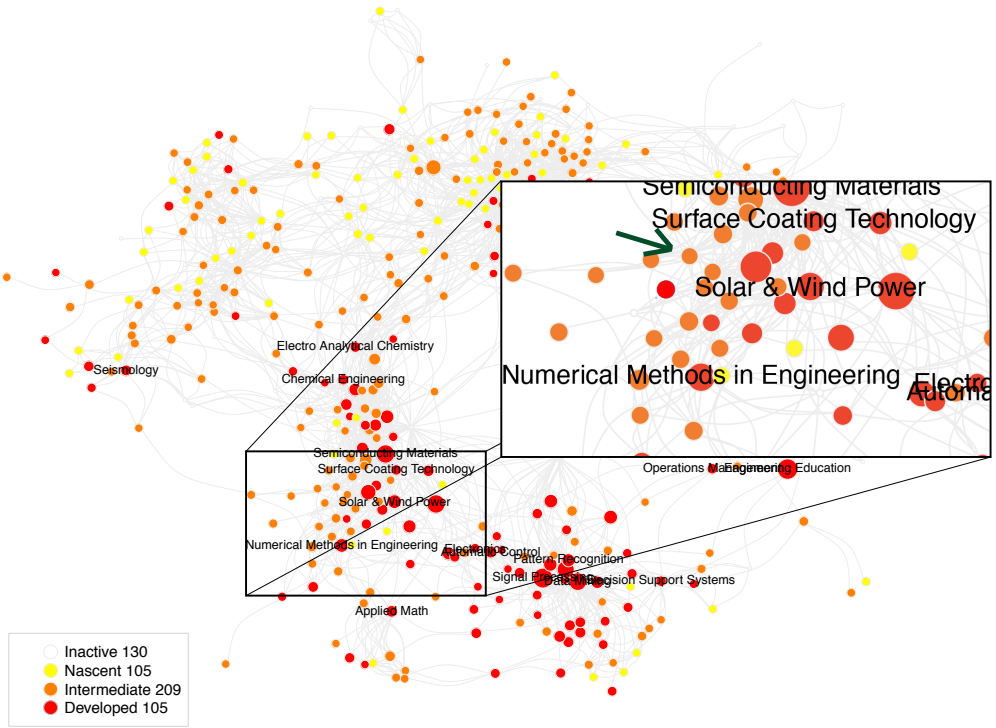
A la representación gráfica que es de gran aporte cualitativo, se suma la información *cuantitativa* que obtenemos al sobreponer datos, vale decir la posibilidad de analizar la dinámica temporal productiva de los actores de la ciencia.

Si dividimos el conjunto de datos en más de un intervalo de tiempo, podemos utilizar los datos del primer intervalo para recomendar nodos que se pueden desarrollar a futuro y utilizar los datos en el segundo intervalo de tiempo para evaluar la calidad de nuestra recomendación.

La medida propuesta para priorizar (*rankear*) las categorías de la ciencia (nodos) que se desarrollarán a futuro, se denomina *densidad activa* y es una medida que para cada nodo Inactivo, evalúa qué tan conectado se encuentra con nodos Activos. Esta medida se ha aplicado en otros estudios de redes de producción, como por ejemplo en la red *espacio producto* [Hidalgo et al., 2007]. La densidad activa $\omega_c$ para una categoría —no activa— $c$ se calcula como la razón entre los enlaces activos —que conectan el nodo con nodos activos— con el total de enlaces que conectan el nodo en estudio con otros nodos, es decir su grado. La densidad




**(a)** 2008-2010




**(b)** 2011-2013

**Figura 5.3:** Espacio investigación con acercamiento para Taiwan en áreas relacionadas a Materiales y Energía. Nótese la activación (de un intervalo de tiempo a otro), del nodo al que apunta la flecha.

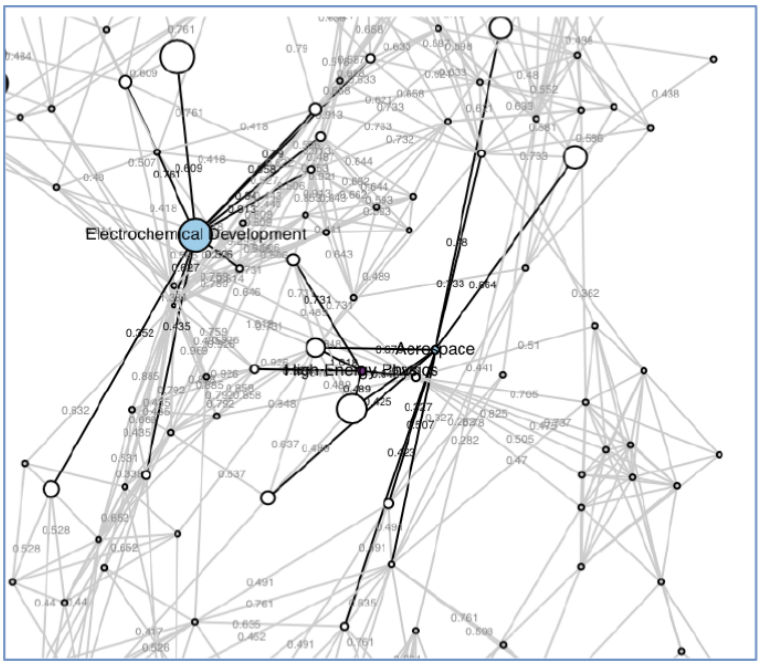



**Figura 5.4:** Zoom al mapa de la ciencia de un productor aleatorio, que muestra los valores de similitud para los enlaces y destaca en blanco los nodos *oportunidad* y en colores los nodos ya *desarrollados*. El tamaño de los nodos blancos es proporcional al valor de su densidad activa (ver Ecuación 5.2).

$\omega$ para un nodo inactivo $c$, se puede calcular con la siguiente ecuación:

$$\omega_c = \frac{\sum_j A_j \phi_{cj}}{\sum_j \phi_{cj}}, \tag{5.2}$$

donde $A_j$ es 1 en caso de que el nodo vecino $j$ se encuentre activo y es 0 en caso contrario.

Nótese que $\omega_c$ es igual a uno, en caso de que el nodo (Inactivo) en evaluación $c$, se encuentra conectado solamente con nodos Activos. Mientras que $\omega_c$ es cero en caso de que el nodo $c$ se encuentra conectado solamente con nodos Inactivos.

Nótese también que este análisis se puede conducir para las otras dos transiciones que hemos definido en la Sección 5.1.1, es decir, para evaluar la *oportunidad* de nodos Nacientes e Intermedios de convertirse en Desarrollados. Para esta última transición, en la Figura 5.4 que contiene un mapa superpuesto de ejemplo, destacamos en color blanco y borde negro, aquellos nodos en evaluación o nodos *oportunidad*.

Para destacar la idea de cómo se calcula la densidad, en la Figura 5.4 presentamos un acercamiento a un mapa de la ciencia, que incluye los valores de similitud de los enlaces y donde los nodos de color blanco representan nodos Oportunidad y su tamaño representa el valor de densidad. Debe notarse como la densidad de nodos oportunidad cercanos a nodos ya desarrollados es mayor (círculo de mayor diámetro) en relación a aquellos que se encuentran poco conectados a nodos desarrollados.

### 5.1.6. Evaluación basada en curvas ROC

Para cada transición, la predicción basada en ranking, propone el listado de todas las áreas en el primer estado o estado actual (por ejemplo Inactivo) y que pueden obtener, en el futuro, el segundo estado o estado deseado (por ejemplo Activo). Un buen resultado será aquel que presente en las primeras posiciones del ranking, las áreas que efectivamente alcanzaron el estado deseado en el futuro.

Formalmente esta evaluación de la calidad de la predicción, se puede abordar con el uso de curvas ROC (por sus siglas en inglés Receiver Operating Characteristic). Las curvas ROC grafican en el eje Y la tasa de Verdaderos Positivos (VP) y en el eje X la tasa de Falsos Positivos (FP). La curva ROC se encuentra en el intervalo de 0 a 1, tanto para el eje X como para el eje Y. Las curvas ROC son ampliamente utilizadas en clasificación binaria y en esta tesis se utilizan para evaluar el ranking de áreas recomendadas basado en la medida de densidad especificada en la sección anterior.

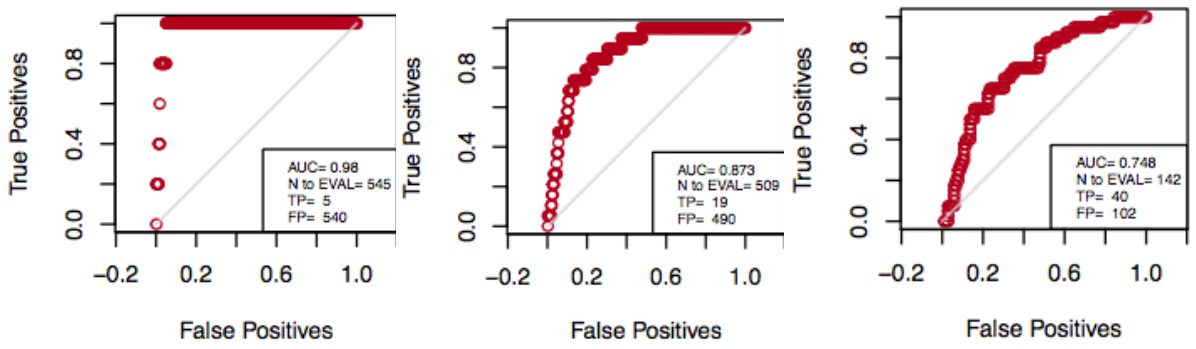
En nuestro caso, cada vez que un área recomendada en el intervalo de tiempo actual, alcance el estado deseado en el intervalo de tiempo siguiente, se contará como un Verdadero Positivo. Se contará como un Falso Positivo en caso contrario. Por ejemplo, en la transición de diversificación, esto es, la evaluación de áreas en estado Inactivo que pasan a estado Activo, se considerarán como verdaderos positivos, todas aquellas áreas que se activaron en el segundo intervalo de tiempo y se considerarán como Falsos Positivos aquellas que permanecieron Inactivas. El predictor perfecto, será aquel que incluya todos los VP en las primeras posiciones del ranking recomendado. Esto es, que el gráfico suba rápidamente por el eje Y, en lugar de avanzar por el eje X.

Una vez que se han ubicado todos los puntos en el gráfico de la curva ROC, se calcula el área bajo la curva, AUC (por sus siglas en inglés *Area Under the Curve*) para obtener un valor que dé cuenta de qué tan buena fue la clasificación para el productor (individuo, institución o país) en evaluación. Un valor de 0.5 para el área bajo la curva ROC, indica un rendimiento comparable a aleatorio (todos los puntos sobre la diagonal), valores menores a 0.5 implican un rendimiento menor a aleatorio y valores por sobre 0.5 se aceptan como mejores que aleatorio.

En la Figura 5.5 presentamos tres ejemplos de curvas ROC (con buenos resultados) para productores de tipo individuo, institución y país. Evaluamos la transición de Inactivo hacia Activo, utilizando la medida de densidad calculada sobre la matriz de proximidad del espacio investigación. Nótese que en el ejemplo, se ilustra también, lo que parece ser lógico, y es el hecho de que la cantidad de áreas a las que se diversifica un autor, es menor que las áreas a las que se diversifica una institución, que a su vez, es menor que la cantidad de áreas a las que se diversifica un país. La Figura 5.5 presenta también, algo que no parece tan obvio, y es el hecho de que el valor del área bajo la curva a mayor resolución del productor (individuos) es mucho mejor que a menor nivel de resolución (instituciones y países.)

### 5.1.7. Diseño experimental

Para conducir la evaluación, utilizamos dos intervalos de tiempo, el primero de 2008 a 2010 y el segundo de 2011 a 2013. Nótese que los datos utilizados para evaluar la predicción,




**(a)** Científico: Andre Anders **(b)** Institución: Universität Mannheim **(c)** País: Taiwan

**Figura 5.5:** Ejemplo de curvas ROC que evalúan la recomendación de áreas que se activarán en el futuro, basado en el espacio investigación. La predicción se realiza respecto de áreas en estado Inactivo en el período 2008-2010 y la evaluación se realiza sobre el estado de las mismas áreas en el período 2011-2013. Se consideran Verdaderos Positivos, aquellas áreas que cambiaron su estado a Activo, y Falsos Positivos en caso contrario. Cada subfigura incluye el valor del área bajo la curva AUC, cantidad de áreas inactivas N_toEval, cantidad de verdaderos positivos TP y cantidad de falsos positivos FP.

no han sido previamente utilizada ni para la construcción del espacio investigación ni para el cálculo de las medidas de densidad. Adicionalmente, se eligió una ventana temporal de tres años, por considerar que es un tiempo suficiente para individuos, instituciones y países para generar nueva productividad científica o para avanzar de un área del conocimiento a otra.

Utilizando los datos del primer intervalo de tiempo, calculamos las densidades correspondientes a las tres transiciones (Inactivo-Activo, Naciente-Desarrollado e Intermedio-Desarrollado). Este cálculo se realizó sobre cuatro mapas: dos espacio investigación y dos mapas de comparación, estos mapas son, el mapa UCSD y el mapa Scimago.

Las matrices de proximidad utilizadas para los dos espacio investigación en evaluación, son las matrices obtenidas del proceso de construcción del espacio investigación, descrito en la Sección 4.5. Un espacio investigación se encuentra en clasificación UCSD y el otro en clasificación SCImago.

El mapa de comparación, basado en citas, UCSD, se obtuvo del sitio web referido en [Börner et al., 2012b].

Como en la literatura, no existe un mapa (liberado) en clasificación SCImago, creamos un mapa para comparación, utilizando la probabilidad condicional de que un *glsjournal* se encuentre indexado en dos categorías a la vez. Hemos utilizado a propósito la misma medida de similaridad que en el espacio investigación, para favorecer el contraste del tipo de datos utilizado en la construcción de cada mapa. Denominamos a este mapa de comparación, mapa SCImago.

Para que la experimentación se pueda llevar a cabo, esto es, obtener una cantidad suficiente de transiciones en el caso de cada tipo de productor (individuos, instituciones y países), hemos incluido solo productores en los que su nivel de producción en el tiempo elegido, cumple con

la siguiente desigualdad:

$$\sum_{t=2008}^{t=2013} \sum_{c} X_{ict} > b * 6, \tag{5.3}$$

con $b = 3$ para individuos, y $b = 30$ para instituciones y países.

Una vez que obtenemos la curva ROC para cada productor en cada mapa y por cada transición, obtenemos también las áreas bajo la curva AUC en cada caso. Con estos datos, podemos reportar los resultados que se presentan y discuten en la sección siguiente.

Para mayor detalle, en el Apéndice, hemos incluido la comparación de curvas ROC entre el espacio investigación y los mapas UCSD y SCIMAGO. Hemos incluido un número limitado de individuos (apéndices F y I ), instituciones (apéndices G y J) y países (apéndices H y K).

## 5.2. Resultados

### 5.2.1. Resultados en comparación con el mapa UCSD

Primeramente, presentamos los resultados obtenidos de la comparación entre el espacio investigación (RS por sus siglas en inglés Research Space) y el mapa UCSD que está basado en patrones de citas. Para esto, en la Figura 5.6, se muestran boxplots para cada transición evaluada (Inactivo-Activo, Naciente-Desarrollado e Intermedio-Desarrollado) y para cada tipo de productor en estudio: individuos, instituciones y países. Nótese que en el caso de individuos, solamente evaluamos diversificación, esto es, la transición de Inactivo a Activo.

En términos de diversificación, esto es transición de áreas Inactivas hacia áreas Activas, a nivel de individuos, el espacio investigación presenta mayor valor promedio (0.87) respecto del mapa UCSD (0.77). Lo mismo que sucede a nivel de Instituciones, 0.71 versus 0.69 aunque con menor diferencia (ver detalles en Tablas 5.1, 5.2 y 5.3). La diferencia es significativa para los casos de individuos e instituciones. No así para el caso de países donde la diferencia es inexistente. En la Tabla 5.4 se presenta el detalle del análisis ANOVA realizado para los valores de AUC en los tres niveles de granularidad estudiados en cuanto a diversificación.

Con estos resultados podemos rechazar las hipótesis nulas asociadas a las hipótesis alternativas $H_1$ y $H_2$, esto es, podemos afirmar que el espacio investigación es un mejor descriptor de la diversificación tanto de individuos como de instituciones que el mapa UCSD.

A la luz de estos resultados, también podemos afirmar que no existe evidencia suficiente para indicar que el espacio investigación es un mejor descriptor de la diversificación de países (hipótesis $H_3$).

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.42 | 0.86 | 0.92 | 0.90 | 0.96 | 1.00 | 879.00 |
| UCSD | 0.25 | 0.71 | 0.81 | 0.80 | 0.92 | 1.00 | 1119.00 |

**Tabla 5.1:** Individuos. Transición de Inactivo a Activo. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa UCSD. Tamaño de la muestra: 4850.

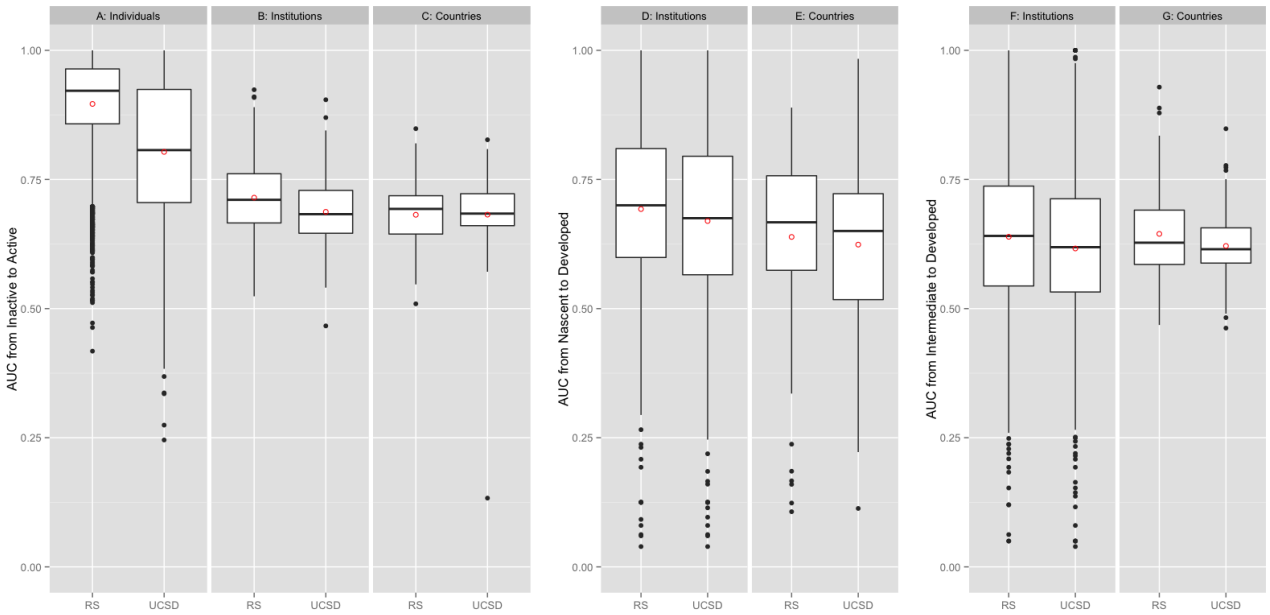



**Figura 5.6:** Resultados de valores de área bajo la curva ROC, para la evaluación de tres transiciones: de Inactivo a Activo, de Naciente a Desarrollado y de Intermedio a Desarrollado. La primera transición se evalúa para individuos, instituciones y países, mientras que las otras dos transiciones se evalúan para instituciones y países. Cada boxplot presenta en la línea horizontal el valor de la mediana y, en el círculo rojo, el valor del promedio.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| RS | 0.52 | 0.67 | 0.71 | 0.71 | 0.76 | 0.92 |
| UCSD | 0.47 | 0.65 | 0.68 | 0.69 | 0.73 | 0.90 |

**Tabla 5.2:** Instituciones. Transición de Inactivo a Activo. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa UCSD. Tamaño de la muestra: 730.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| RS | 0.51 | 0.64 | 0.69 | 0.68 | 0.72 | 0.85 |
| UCSD | 0.13 | 0.66 | 0.68 | 0.68 | 0.72 | 0.83 |

**Tabla 5.3:** Países. Transición de Inactivo a Activo. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa UCSD. Tamaño de la muestra: 77.

En términos de desarrollo, esto es para las transiciones de áreas Nacientes y áreas Intermedias, hacia áreas Desarrolladas, nuevamente el espacio investigación, presenta mejores resultados promedio a nivel de Instituciones que son además significativos, tanto en la transición de Nacientes a Desarrollados, (0.69 para el RS versus 0.67 para el mapa UCSD (Tablas 5.5 y 5.6), valor-p $<0.05$), como en la transición de Intermedios a Desarrollados (0.64 versus 0.62 (Tablas 5.8 y 5.9), valor-p $<0.01$). Esta diferencia mejor y significativa, no se produce a nivel de países (Tablas 5.6 y 5.9) en ninguna de las dos transiciones que describen desarrollo. Los resultados de los tests de significancia ANOVA se pueden revisar en las Tablas 5.7 y 5.10.

Estos resultados nos permiten rechazar la hipótesis nula asociada a la hipótesis alternativa $H_7$, vale decir, afirmar que el espacio investigación es un mejor descriptor del desarrollo de áreas científicas a nivel de instituciones, en comparación con el mapa UCSD. Sin embargo, no existe evidencia suficiente para afirmar que el espacio investigación también es un mejor descriptor de desarrollo a nivel de países (hipótesis $H_8$).

### 5.2.2. Resultados en comparación con el mapa SCImago

La segunda parte de nuestros resultados, presenta la comparación realizada con el mapa SCImago, que se realiza con el objetivo de verificar la robustez de nuestro método, al compararlo con otra red en una clasificación distinta, en este caso el mapa SCImago.

Para esto, en la Figura 5.7, se muestran boxplots para cada transición evaluada (Inactivo-Activo, Naciente-Desarrollado e Intermedio-Desarrollado) y para cada tipo de productor en estudio: individuos, instituciones y países. Nótese que, al igual que en la evaluación anterior, en el caso de individuos, solamente evaluamos diversificación, por lo que solo se presenta la transición de Inactivo a Activo.

En términos de diversificación, esto es transición de áreas Inactivas hacia áreas Activas, a nivel de individuos, el espacio investigación presenta mayor valor promedio (0.87) respecto

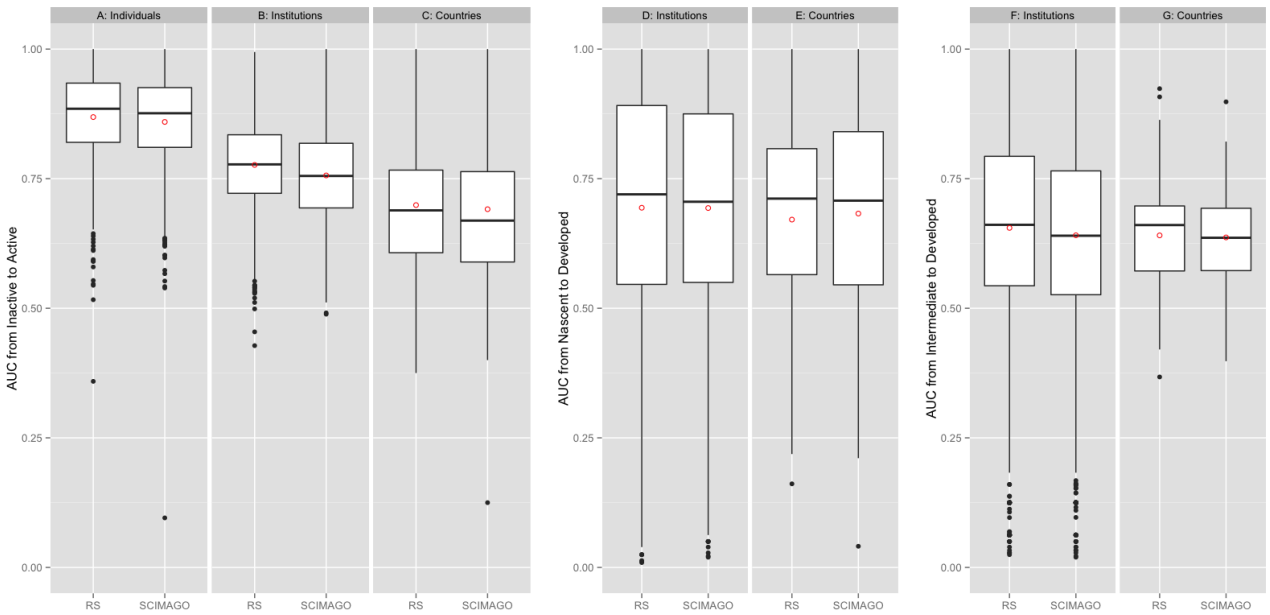



**Figura 5.7:** Comparación entre el espacio investigación y el mapa SCimago. Se comparan los resultados de valores de área bajo la curva ROC, para la evaluación de tres transiciones: de Inactivo a Activo, de Naciente a Desarrollado y de Intermedio a Desarrollado. La primera transición se evalúa para individuos, instituciones y países, mientras que las otras dos transiciones se evalúan para instituciones y países. Cada boxplot presenta en la línea horizontal el valor de la mediana y, en el círculo rojo, el valor del promedio.

| | Individuos | Residuos | Instituciones | Residuos | Países | Residuos |
|---|---|---|---|---|---|---|
| Df | 1.0000 | 7700.0000 | 1.0000 | 1458.0000 | 1.0000 | 152.0000 |
| Sum Sq | 16.5904 | 100.1688 | 0.2747 | 5.8308 | 0.0000 | 0.7952 |
| Mean Sq | 16.5904 | 0.0130 | 0.2747 | 0.0040 | 0.0000 | 0.0052 |
| F value | 1275.3046 | | 68.6886 | | 0.0005 | |
| Pr(>F) | 0.0000 | | 0.0000 | | 0.9824 | |

**Tabla 5.4:** Resultados del test ANOVA para la transición de Inactivo a Activo. Comparación de poblaciones entre el espacio investigación y el mapa UCSD. El valor de significancia aumenta a medida que el valor-p (Pr>F) es menor que 0.05, 0.01, 0.001.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.04 | 0.60 | 0.70 | 0.69 | 0.81 | 1.00 | 78.00 |
| UCSD | 0.04 | 0.57 | 0.68 | 0.67 | 0.79 | 1.00 | 77.00 |

**Tabla 5.5:** Instituciones. Transición de Naciente a Desarrollado. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa UCSD. Tamaño de la muestra: 730.

del mapa SCImago (0.86). Lo mismo que sucede a nivel de Instituciones, 0.78 versus 0.76 (ver detalles en Tablas 5.11, 5.12 y 5.13). La diferencia es significativa para los casos de individuos (valor-p <0.01) e instituciones (valor-p <0.01). No así para el caso de países donde la diferencia no es significativa, aún cuando el promedio del espacio investigación es mayor al promedio del mapa SCImago (0.70 versus 0.69). En la Tabla 5.14 se presenta el detalle del análisis ANOVA realizado para los valores de AUC en los tres niveles de granularidad estudiados en cuanto a diversificación.

Estos resultados nos permiten rechazar las hipótesis nulas asociadas las hipótesis alternativas $H_4$ y $H_5$, es decir, podemos afirmar que el espacio investigación, también es un mejor descriptor de la diversificación científica de individuos (hipótesis $H_5$) y de instituciones (hipótesis $H_6$). También podemos indicar que no existe evidencia suficiente para afirmar que el espacio investigación es mejor descriptor de la diversificación a nivel de países, en comparación con el mapa SCImago (hipótesis $H_7$).

En términos de desarrollo, esto es para las transiciones de áreas Nacientes y áreas Intermedias, hacia áreas Desarrolladas, el espacio investigación, presenta iguales resultados promedio a nivel de Instituciones en la transición de Nacientes a Desarrollados (0.69). En la transición de Intermedios a Desarrollados el espacio investigación (0.66) supera al mapa SCImago $(0,64)$, siendo esta diferencia significativa (valor-p <0.05). A nivel de países no existe diferencia significativa en ninguna de las dos transiciones que describen desarrollo. Las tablas 5.15 y 5.16 presentan el detalle de estos estadísticos descriptivos para las transiciones de naciente a desarrollado, mientras que las tablas Tablas 5.18 y 5.19 presentan los estadísticos descriptivos para las transiciones de Intermedios a Desarrollados. Los resultados de los tests de significancia ANOVA se pueden revisar en las Tablas 5.17 y 5.20.

Los resultados que hemos presentado, para las dos transiciones que componen el análisis

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.11 | 0.57 | 0.67 | 0.64 | 0.76 | 0.89 | 5.00 |
| UCSD | 0.11 | 0.52 | 0.65 | 0.62 | 0.72 | 0.98 | 5.00 |

**Tabla 5.6:** Paises. Transición de Naciente a Desarrollado. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa UCSD. Tamaño de la muestra: 77.

| | Instituciones | Residuos | Paises | Residuos |
|---|---|---|---|---|
| Df | 1.0000 | 1303.0000 | 1.0000 | 142.0000 |
| Sum Sq | 0.1741 | 40.3047 | 0.0079 | 4.2733 |
| Mean Sq | 0.1741 | 0.0309 | 0.0079 | 0.0301 |
| F value | 5.6290 | | 0.2636 | |
| Pr(>F) | 0.0178 | | 0.6085 | |

**Tabla 5.7:** Resultados del test ANOVA para la transición de Naciente a Desarrollado. Comparación de poblaciones entre el espacio investigación y el mapa UCSD . El valor de significancia aumenta a medida que el valor-p (Pr>F) es menor que 0.05, 0.01, 0.001.

del desarrollo científico de instituciones como de países, no son evidencia suficiente para rechazar las hipótesis nulas asociadas a las hipótesis alternativas $H_9$ y $H_{10}$, vale decir, no podemos afirmar que el espacio investigación es un mejor descriptor del desarrollo tanto de instituciones como de países, en comparación con el mapa SCimago.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.05 | 0.54 | 0.64 | 0.64 | 0.74 | 1.00 | 31.00 |
| UCSD | 0.04 | 0.53 | 0.62 | 0.62 | 0.71 | 1.00 | 31.00 |

**Tabla 5.8:** Instituciones. Transición de Intermedio a Desarrollado. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa UCSD. Tamaño de la muestra: 730.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.47 | 0.59 | 0.63 | 0.64 | 0.69 | 0.93 | 1.00 |
| UCSD | 0.46 | 0.59 | 0.61 | 0.62 | 0.66 | 0.85 | 1.00 |

**Tabla 5.9:** Paises. Transición de Intermedio a Desarrollado. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa UCSD. Tamaño de la muestra: 77.

| | Instituciones | Residuos | Paises | Residuos |
|---|---|---|---|---|
| Df | 1.0000 | 1396.0000 | 1.0000 | 150.0000 |
| Sum Sq | 0.1787 | 35.7562 | 0.0209 | 0.9382 |
| Mean Sq | 0.1787 | 0.0256 | 0.0209 | 0.0063 |
| F value | 6.9782 | | 3.3338 | |
| Pr(>F) | 0.0083 | | 0.0699 | |

**Tabla 5.10:** Resultados del test ANOVA para la transición de Intermedio a Desarrollado. Comparación de poblaciones entre el espacio investigación y el mapa UCSD . El valor de significancia aumenta a medida que el valor-p (Pr>F) es menor que 0.05, 0.01, 0.001.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.36 | 0.82 | 0.88 | 0.87 | 0.93 | 1.00 | 24.00 |
| SCIMAGO | 0.10 | 0.81 | 0.88 | 0.86 | 0.93 | 1.00 | 30.00 |

**Tabla 5.11:** Individuos. Transición de Inactivo a Activo. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa SCIMAGO. Tamaño de la muestra: 900.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| RS | 0.43 | 0.72 | 0.78 | 0.78 | 0.83 | 0.99 |
| SCIMAGO | 0.49 | 0.69 | 0.76 | 0.76 | 0.82 | 1.00 |

**Tabla 5.12:** Instituciones. Transición de Inactivo a Activo. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa SCIMAGO. Tamaño de la muestra: 2587.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.38 | 0.61 | 0.69 | 0.70 | 0.77 | 1.00 | 8.00 |
| SCIMAGO | 0.12 | 0.59 | 0.67 | 0.69 | 0.76 | 1.00 | 8.00 |

**Tabla 5.13:** Países. Transición de Inactivo a Activo. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa SCIMAGO. Tamaño de la muestra: 123.

| | Individuos | Residuos | Instituciones | Residuos | Países | Residuos |
|---|---|---|---|---|---|---|
| Df | 1.0000 | 1744.0000 | 1.0000 | 5172.0000 | 1.0000 | 228.0000 |
| Sum Sq | 0.0385 | 13.8240 | 0.5306 | 37.0968 | 0.0036 | 4.8587 |
| Mean Sq | 0.0385 | 0.0079 | 0.5306 | 0.0072 | 0.0036 | 0.0213 |
| F value | 4.8547 | | 73.9777 | | 0.1696 | |
| Pr(>F) | 0.0277 | | 0.0000 | | 0.6809 | |

**Tabla 5.14:** Resultados del test ANOVA para la transición de Inactivo a Activo. Comparación de poblaciones entre el espacio investigación y el mapa SCIMAGO. El valor de significancia aumenta a medida que el valor-p (Pr>F) es menor que 0.05, 0.01, 0.001.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.01 | 0.55 | 0.72 | 0.69 | 0.89 | 1.00 | 1378.00 |
| SCIMAGO | 0.02 | 0.55 | 0.71 | 0.69 | 0.88 | 1.00 | 1380.00 |

**Tabla 5.15:** Instituciones. Transición de Naciente a Desarrollado. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa SCIMAGO. Tamaño de la muestra: 2587.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.16 | 0.56 | 0.71 | 0.67 | 0.81 | 1.00 | 27.00 |
| SCIMAGO | 0.04 | 0.55 | 0.71 | 0.68 | 0.84 | 1.00 | 27.00 |

**Tabla 5.16:** Paises. Transición de Naciente a Desarrollado. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa SCIMAGO. Tamaño de la muestra: 123.

| | Instituciones | Residuos | Paises | Residuos |
|---|---|---|---|---|
| Df | 1.0000 | 2414.0000 | 1.0000 | 190.0000 |
| Sum Sq | 0.0002 | 130.9452 | 0.0066 | 7.7416 |
| Mean Sq | 0.0002 | 0.0542 | 0.0066 | 0.0407 |
| F value | 0.0034 | | 0.1613 | |
| Pr(>F) | 0.9538 | | 0.6884 | |

**Tabla 5.17:** Resultados del test ANOVA para la transición de Naciente a Desarrollado. Comparación de poblaciones entre el espacio investigación y el mapa SCIMAGO . El valor de significancia aumenta a medida que el valor-p (Pr>F) es menor que 0.05, 0.01, 0.001.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| RS | 0.02 | 0.54 | 0.66 | 0.66 | 0.79 | 1.00 | 647.00 |
| SCIMAGO | 0.02 | 0.53 | 0.64 | 0.64 | 0.77 | 1.00 | 668.00 |

**Tabla 5.18:** Instituciones. Transición de Intermedio a Desarrollado. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa SCIMAGO. Tamaño de la muestra: 2587.

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| RS | 0.37 | 0.57 | 0.66 | 0.64 | 0.70 | 0.92 |
| SCIMAGO | 0.40 | 0.57 | 0.64 | 0.64 | 0.69 | 0.90 |

**Tabla 5.19:** Paises. Transición de Intermedio a Desarrollado. Estadísticos descriptivos para valores bajo la curva ROC. Comparación entre el espacio investigación RS y el mapa SCIMAGO. Tamaño de la muestra: 123.

| | Instituciones | Residuos | Paises | Residuos |
|---|---|---|---|---|
| Df | 1.0000 | 3857.0000 | 1.0000 | 244.0000 |
| Sum Sq | 0.1951 | 164.1275 | 0.0009 | 2.3244 |
| Mean Sq | 0.1951 | 0.0426 | 0.0009 | 0.0095 |
| F value | 4.5853 | | 0.0991 | |
| Pr(>F) | 0.0323 | | 0.7531 | |

**Tabla 5.20:** Resultados del test ANOVA para la transición de Intermedio a Desarrollado. Comparación de poblaciones entre el espacio investigación y el mapa SCIMAGO . El valor de significancia aumenta a medida que el valor-p (Pr>F) es menor que 0.05, 0.01, 0.001.

Capítulo 6

# Aplicaciones: diverse y Opus

De los resultados obtenidos en esta tesis, se desprenden dos importantes aplicaciones que hemos facilitado o desarrollado. La primera radica en facilitar la medición de la diversidad en términos de producción científica, tanto de medidas simples, como de medidas más complejas en las que se utiliza el espacio investigación. Para esto, hemos construido y publicado el paquete para R denominado *diverse*. La segunda aplicación, constituye una aplicación web, denominada *Opus*, que permite hacer más comprensible y accesible al público general la diversidad productiva tanto de individuos, instituciones y países, utilizando para ello un conjunto de visualizaciones de datos, entre ellas, la red del espacio investigación. En este capítulo detallamos las dos aplicaciones y cómo se encuentran relacionadas con este trabajo de tesis.

## 6.1. *diverse*: Midiendo diversidad

Debido a que hemos relevado la importancia de contar con una herramienta que facilite la medición de las distintos aspectos de la diversidad y el manejo de datos, hemos construido un paquete para R [R Core Team, 2014] al que denominamos *diverse* [Guevara et al., 2015] y que provee una interfaz de fácil uso para trabajar con altos volúmenes de datos y calcular medidas de diversidad que son aplicables, no solo a Bibliometría sino a otras áreas de la ciencia. El paquete *diverse* incluye tres conceptos básicos de diversidad, estos son *variedad*, *balance* y *disparidad*. El paquete *diverse* también facilita el cálculo de RCA y otras medidas de normalización de datos. *diverse* está disponible en el repositorio de The Comprehensive R Archive Network (CRAN) y la versión de desarrollo se encuentra disponible en el siguiente repositorio de GitHub github.com/mguevara/diverse.

### 6.1.1. Mapas superpuestos

Para destacar la relevancia de la diversidad y cómo ésta se vincula con los mapas de la ciencia, es necesario describir primeramente los mapas —superpuestos— de la ciencia.

Los mapas superpuestos grafican la posición actual del productor sobre el mapa de la ciencia. Rafols et al. [2010] fueron pioneros en construir estos mapas superpuestos (*overlay maps*), como técnica para describir las características de un conjunto reducido de instituciones y proponer esta herramienta como instrumento para la política pública o institucional en investigación. En los últimos años se ha utilizado esta técnica en mapas de la ciencia a diferentes niveles de granularidad, por ejemplo a nivel de tópicos [Fried and Kobourov, 2014] o a nivel de *journals* [Chen and Leydesdorff, 2014; Leydesdorff et al., 2015].

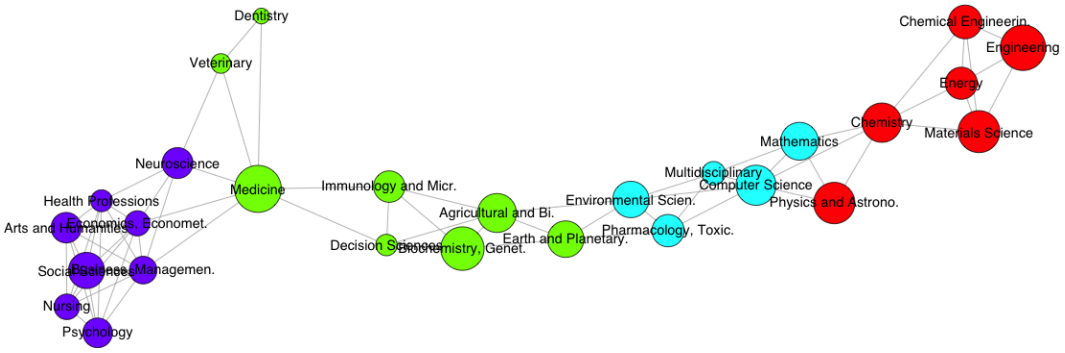
Si bien la forma más simple de superponer datos a un mapa de base, es activar/desactivar nodos, como en [Chen and Leydesdorff, 2014; Leydesdorff et al., 2015; Rafols et al., 2010], también se puede dotar de mayor semántica al mapa, agregando por ejemplo la intensidad de publicación, como en [Fried and Kobourov, 2014], que utiliza mapas de calor para hacer referencia a la cantidad de producción en cada área del mapa.

En esta tesis, hemos avanzado la creación de mapas superpuestos, utilizando la medida de ventajas comparativas RCA (ver Sección 5.1.2) que nos permiten sobreponer no solo la presencia-ausencia o los niveles de producción en cada nodo, sino también las ventajas comparativas en relación a los otros productores. Esto cobra sentido, cuando la mayoría de países producen *papers* en la mayoría de áreas de la ciencia y requerimos de mejores indicadores que nos permitan encontrar ventajas entre ellos.

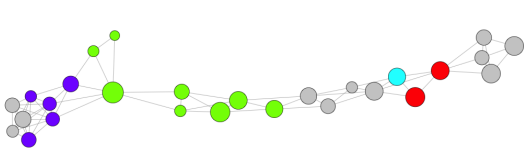
La idea más básica del uso de RCA en mapas superpuestos, consiste en activar aquellos nodos con ventajas comparativas (denominados en esta tesis como Desarrollados) y desactivar aquellas áreas de la ciencia, sin ventajas comparativas (denominados en esta tesis como Nacientes). La Figura 6.1, presenta un ejemplo para dos países (en el contexto de 10), utilizando un sencillo espacio investigación a nivel de áreas, cuya construcción y estructura se introdujo en la Sección 2.6.

Analizando los mapas superpuestos de los países representados en la Imagen 6.1, se puede conducir un análisis cualitativo y notar por ejemplo, la concentración de China en las áreas de ciencias básicas y aplicadas, mientras que Alemania se encuentra más diversificada a lo largo del mapa, explotando áreas desde la Psicología, la Neurociencia y la Medicina, pasando por las ciencias ambientales y biológicas, hasta llegar a las áreas básicas como la Física o la Matemática.

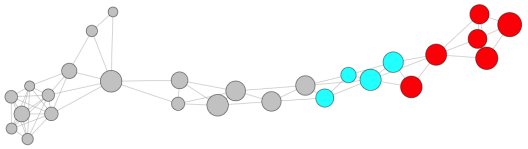
Este análisis, es en última instancia una descripción de la *diversidad* de los productores que gracias a las similitudes que determinan los mapas de la ciencia —como el espacio investigación— se ha podido profundizar para conducir estudios con medidas de diversidad más complejas, como los liderados por Ismael Rafols [2014] quien ha dirigido varios trabajos [Chavarro et al., 2014; Leydresdof and Rafols, 2011; Rafols et al., 2010] en esta línea donde se evalúan las diferentes dimensiones de la diversidad en el contexto de producción en ciencia y tecnología. Estas medidas se encuentran incluidas en *diverse* y se detallarán en las próximas secciones.




**(a)** Mapa base de similitudes entre áreas de la ciencia.



**(b)** Mapa superpuesto para Alemania.



**(c)** Mapa superpuesto para China.

**Figura 6.1:** Ejemplo de superposición de datos de RCA, a un espacio investigación de base. En 6.1b y 6.1c el tamaño de los nodos es proporcional a la producción de autoría para cada país. Los colores representan comunidades de áreas similares y los nodos en gris, identifican areas con ventajas comparativas (RCA) menores a 1.

### 6.1.2. Datos de entrada

El paquete *diverse* acepta como datos de entrada, tanto objetos tipo `dataframe` o tipo `matrix`. En el caso de recibir un `dataframe` se espera que el conjunto de datos incluya tres columnas, en este orden: productor, categoría, valor. En el caso de utilizar un objeto tipo `matrix` se espera que las filas incluyan las entidades productoras, las categorías se encuentren en las columnas y que las celdas sean los valores numéricos que indican el valor de producción de cada productor en cada categoría. El paquete también facilita la importación y lectura de datos de diferentes fuentes y en diferentes formatos, para esto se puede utilizar la función `read_data()` que utiliza los parámetros `path` para indicar la ruta hacia el archivo externo a importar.

En *diverse* se ha incluido un conjunto de datos de producción científica de ejemplo, con el fin de facilitar su descripción. Estos datos han sido descargados y agregados del sitio SCImago, y se puede acceder a través del dataframe `scidat`. Este conjunto de datos incluye 10 países y su producción en las 27 áreas de la ciencia definidas por SCImago. Las características de este conjunto de datos se pueden obtener en R utilizando `str(scidat)` lo que devolvería la siguiente información:

```
num [1:10, 1:27] 3507 35351 15603 1346 4158 ...
- attr(*, "dimnames")=List of 2
  ..$ : chr [1:10] "Argentina" "China" "Germany" "Hungary" ...
  ..$ : chr [1:27] "Agricultural and Biological Sciences" "Arts and ...
    Humanities"...
```

### 6.1.3. Normalización de datos

Un concepto central en esta tesis ha sido la aplicación de la medida RCA para medir desarrollo científico en el contexto de los otros productores. Tanto esta medida, como su normalización entre -1 y 1, conocida en Bibliometría, como Índice de Actividad, se han incluido en *diverse* como parte del proceso de normalización de datos, esto es, el proceso previo a medir diversidad. Este proceso es importante puesto que ayuda, por ejemplo, a profundizar las diferencias entre productores, situación que es muy necesaria, sobre todo cuando todos los productores producen en todas las categorías, como en el caso de producción científica y en concreto del conjunto de datos `scidat`.

La función `values()` facilita la normalización de datos. En esta función el parámetro `norm` se puede definir a `'p'`, `'rca'` o `'ai'`, para elegir respectivamente, proporciones, RCA e Indice de Actividad. También el parámetro `filter` permite elegir un valor como umbral para descartar observaciones bajo ese umbral. Finalmente el parámetro `binary`, realizará el proceso de binarización de la matriz.

Para relevar las importantes diferencias que produce la normalización de datos y las funcionalidades de *diverse*, el siguiente código calcula y grafica diferentes normalizaciones para el conjunto de datos `scidat`:

```
1  library(pheatmap)
2  colfunc <- colorRampPalette(c("deepskyblue4", "deepskyblue", "cyan"))
3  plot_mat <- function(data)
4  pheatmap(data, colfunc(100), cluster_rows = FALSE, cluster_cols = FALSE)
5
6  col_l <- names(sort(colSums(values(scidat)))) #order
7  row_l <- names(sort(rowSums(values(scidat)), decreasing = TRUE))
8  plot_mat(values(scidat)[row_l,col_l])
9  plot_mat(values(scidat, norm = 'p')[row_l,col_l])
10 plot_mat(values(scidat, norm = 'rca')[row_l,col_l])
11 plot_mat(values(scidat, norm = 'rca', filter = 1)[row_l,col_l])
```

El resultado del código anterior se presenta en la Figura 6.2 donde se puede apreciar claramente cómo el uso de normalizaciones internas —como proporciones— o en el contexto global —como RCA—, relevan la importancia o actividad de un área al interior de un país, o en comparación con otros países.

### 6.1.4. Midiendo diversidad

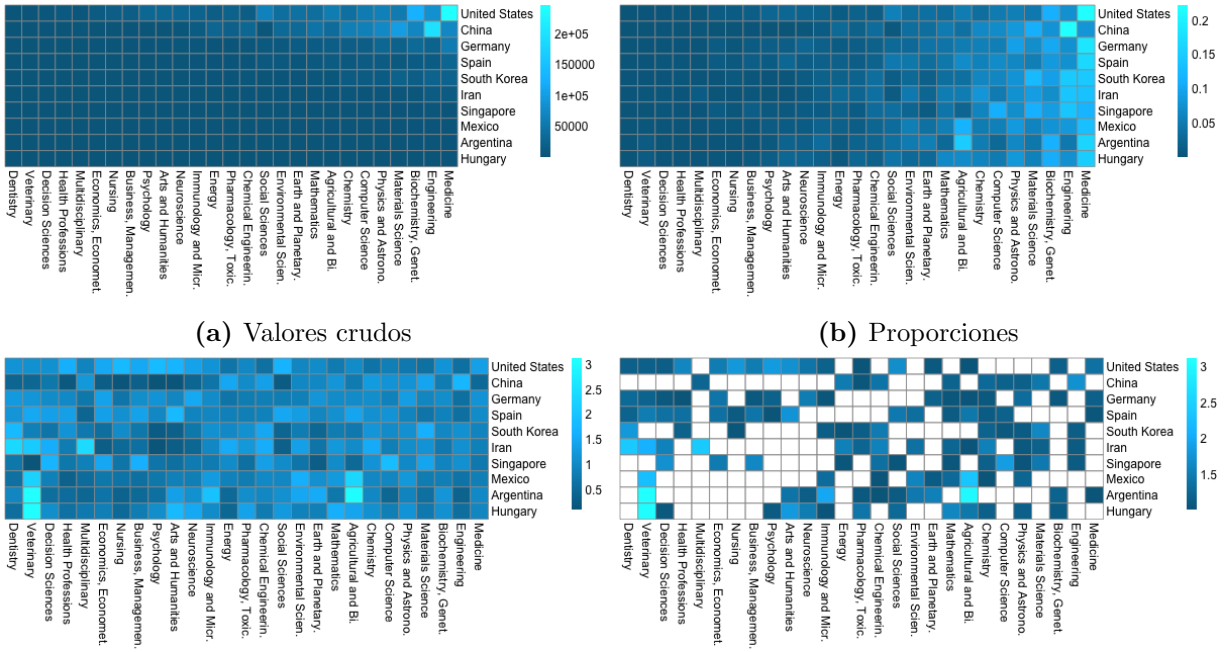
Según Stirling [2007] la diversidad se puede caracterizar en tres diferentes dimensiones, estas son: *variedad*, *balance* y *disparidad*. Siendo todo lo demás igual, cada arista mide y caracteriza aspectos diferentes. Por ejemplo, la variedad indica cuántos elementos de diferentes tipos posee —en nuestro caso, publica— la entidad productora. El balance indica, cuánto de cada tipo se está produciendo. Y la disparidad indica qué tan diferentes son entre sí los elementos que se están produciendo.

En la Tabla 6.1 presentamos todas las medidas de diversidad que se pueden computar con *diverse*. Las medidas se han ordenado desde las más simples hasta las más complejas.

La función que accede a cada una de estas medidas, es la función `diversity()` la misma que utiliza los parámetros `data` y `type` para indicar el conjunto de datos a procesar y el tipo

| ID | Medida | Fórmula | Referencia |
|---|---|---|---|
| v | Variety | $v = \sum_i (p_i^0)$ | |
| hhi | Herfindahl–Hirschman Index | $HHI = \sum_i (p_i^2)$ | Rhoades [1993] |
| b, gs | Blau Index, Gini-Simpson | $B = 1 - \sum_i (p_i^2) = 1 - HHI$ | Blau [1977]; Gini [1912] |
| s | Simpson | $D_S = \sum_i n_i(n_i - 1)/N_t(N_t - 1)$ | Simpson [1949] |
| bp | Berger-Parker | $D_{BP} = \max_i (p_i)$ | Berger and Parker [1970] |
| e | Shannon Entropy | $H = -\sum_i (p_i \log p_i)$ | Shannon [1948] |
| ev | Pielou Evenness | $J = -\sum_i (p_i \log p_i) / \log v$ | Pielou [1970] |
| re | Rényi-Entropy | ${}^qH = (1 - q)^{-1} \log (\sum_i p_i^q)$ | Rényi [1961] |
| td | True Diversity | ${}^qD_{TD} = (\sum_i p_i^q)^{1/(1-q)}$ | Jost [2006] |
| d | Disparity | $DIS = \sum_{ij} d_{ij}/N$ | |
| rao | Rao | $D_{RAO} = \sum_{ij} d_{ij} p_i p_j$ | Rao [1982] |
| rs | Rao-Stirling | $\Delta = \sum_{ij} {d_{ij}}^{\alpha} (p_i p_j)^{\beta}$ | Stirling [2007] |

**Tabla 6.1:** Resumen de medidas incluidas en el paquete *diverse*. El primer bloque de medidas están asociadas principalmente a las dimensiones de variedad y balance. El segundo bloque de medidas presenta aquellas que se encuentran asociadas con la dimensión de disparidad. $C$ es el conjunto de categorías presentes en la entidad $e$. $i, j \in C$. $i \neq j$. $n_i$ es el valor y $p_i$ la proporción de la categoría $i$ en la entidad $e$. $v = n(C)$ es el número de categorías presentes en la entidad —la variedad. $N_t = \sum n_i$. *log* es el logaritmo, usualmente natural. $q, \alpha, \beta \geqslant 0$. Para la medida de *True Diversity*, cuando $q = 1$ la ecuación se indefine por lo que una aproximación es calculada.




**(a)** Valores crudos

**(b)** Proporciones

**(c)** RCA

**(d)** RCA filtrados

**Figura 6.2:** Matrices de calor. Mientras más iluminada la celda, más alto es el valor. El color blanco representa celdas vacías.

de medida que se quiere calcular. El parámetro `data` acepta cualquier objeto de dato que tenga las características descritas en la Sección 6.1.2, mientras que el parámetro `type` puede recibir una cadena de caracteres o un arreglo de cadenas de caracteres, con los identificadores ID de las medidas a computar. Estos identificadores son los que se incluyen en la columna ID de la Tabla 6.1.

### 6.1.5. Variedad

La variedad mide cuántos tipos distintos produce la entidad. En nuestro caso, en cuántas áreas de la ciencia publica cada país. Aunque la variedad se puede calcular con la función `diversity()`, el paquete *diverse* incluye una función propia para acceder a esta medida, que entrega los datos ordenados en forma descendiente. Nuevamente, para que este análisis tenga sentido en nuestro conjunto de datos, se requiere de un proceso de normalización previo. En el siguiente código se presenta un ejemplo de esta funcionalidad aplicada sobre el conjunto de datos `scidat` que previamente se normaliza utilizando RCAs:

```
1 scidat_rca_fil <- values(data = scidat, norm = 'rca', filter = 1)
2 variety(scidat_rca_fil)
3                variety
4 United States       17
5 Germany             16
6 ...
7 China               10
8 Mexico               9
```

Nótese que esta simple medida de diversidad, ya da cuenta de la diversificación productiva de países y sus diferentes niveles de desarrollo.

### 6.1.6. Balance

El balance es el concepto que mide la diversidad en torno a cuánto se produce de cada categoría. Por ejemplo, supongamos el caso hipotético de dos países que solo producen en dos categorías de la ciencia $C_1$ y $C_2$. El primer país produce 99 % en $C_1$ y 1 % en $C_2$, mientras que el segundo produce 60 % en $C_1$ y 40 % en $C_2$. Ciertamente, la variedad de ambos productores es igual a 2, pero sin embargo el primer país se puede considerar menos diverso que el segundo, por cuanto la cantidad que produce en $C_2$ es mínima.

Este concepto, se captura a través de diferentes medidas que utilizan como base la medida más simple de balance, esto es la proporción o la probabilidad que es la noción más básica de cómo se distribuyen los datos. En *diverse* se calcula con la función `balance()`. Estas medidas se incluyen en el segundo bloque de la Tabla 6.1 y entre ellas destaca la Entropía de Shannon o su versión normalizada, conocida también como Evenness de Pielou.

En el siguiente ejemplo, se calculan las proporciones para el conjunto de datos `scidat`:

```
1  balance(scidat)
2               Agricultural and Biological Sciences Arts and Humanities...
3 Argentina                                0.14929122           0.020305649
4 China                                    0.03954286           0.002249461
5 Germany                                  0.05183721           0.014033223
6 ...
7 Spain                                    0.07115690           0.025874672
8 United States                            0.05171968           0.025014498
```

En el siguiente ejemplo presentamos varias medidas asociadas al concepto de balance, utilizando para ello la funcionalidad que incluye *diverse* para calcular más de una medida a la vez:

```
1  diversity(data = scidat, type = c('e','gs','ev'))
2                entropy gini.simpson  evenness
3 Argentina     2.761676    0.9154321 0.8379287
4 China         2.610961    0.9006212 0.7921997
5 Germany       2.809112    0.9185981 0.8523214
6 ...
7 Spain         2.878559    0.9276299 0.8733924
8 United States 2.828207    0.9134401 0.8581151
```

### 6.1.7. Disparidad

La característica de disparidad, mide cuán distintas son las categorías que se encuentra produciendo una entidad. Por ejemplo áreas como "Ingeniería" y "Ciencia de los Materiales" se encuentran más cercanas entre sí, al igual que "Ciencias Sociales" con "Artes y Humanidades". La disparidad será mayor, cuando mayor sean las diferencias entre categorías que se producen.

Este concepto está íntimamente vinculado con lo que hemos venido presentando a lo largo de esta tesis, con la salvedad de que el espacio investigación está basado en similitudes,

mientras que las medidas de disparidad deberán estar basadas en dis-similitudes o diferencias. Antes de obtener cualquier medida de disparidad, se debe contar con una adecuada matriz de dis-similitudes o distancias, para esto, se pueden aplicar cualquiera de las medidas descritas en la Sección 2.4 y otras como el Índice Jaccard. *diverse* incluye esta funcionalidad para la obtención de la matriz de dis-similitudes a través de la función `dis_categories()` la misma que en el parámetro `method` recibe el método solicitado para calcular las dis-similitudes, como 'coseno' o 'jaccard'.

En el siguiente ejemplo, calculamos la matriz de dis-similitudes para el conjunto de datos `scidat` y también graficamos el espacio investigación que se muestra en la Figura 6.1a para lo que hemos filtrado enlaces mayores a 0.015.

```
adj <- dis_categories(data = scidat, method = 'cosine')
adj[adj > 0.015] <- 0  #filter

library(igraph)
g <- graph.adjacency(adjmatrix = adj, mode = 'undirected', weighted = TRUE)
totals <- colSums(values(scidat))
V(g)$size = log(totals[match(V(g)$name, names(totals))], base = 2) -9
fc <- fastgreedy.community(g); colors <- rainbow(max(membership(fc)))
V(g)$color = colors[membership(fc)]
set.seed(67); g$layout <- layout.fruchterman.reingold(g)
plot.igraph(g,  vertex.label.cex = 0.9, vertex.label.font = 0,
        vertex.label.family = 'Helvetica', vertex.label.color='black', asp = ...
           FALSE)
```

En el paquete *diverse* el cálculo de disparidad se realiza a través de la función `disparity()` y requiere de una matriz de dis-similitud como la construida en el ejemplo anterior. Esta matriz es calculada por *diverse* o puede ser provista por el usuario en el parámetro `dis`. Esta matriz, por ejemplo puede obtenerse a partir de la matriz de similitudes del espacio investigación que hemos desarrollado en esta tesis.

El siguiente ejemplo ilustra el uso de esta función la que devuelve como resultado la suma y el promedio de las disparidades entre las categorías en las que el productor tiene actividad.

```
  scidat_rca_fil <- values(scidat, norm = 'rca', filter = 1)
disparity(scidat_rca_fil)
               disparity.sum disparity.mean
Argentina          121.12704       0.3450913
China               54.86895       0.1563218
...
Spain              147.35440       0.4198131
United States      190.86552       0.5437764
```

### 6.1.8. Medidas completas de diversidad

Finalmente, hemos considerado medidas más complejas de diversidad que incluyen en su cálculo los tres conceptos principales ya discutidos, estos son: variedad, balance y disparidad. La media más utilizada en este aspecto es la propuesta por Stirling [2007] y que es el producto de las proporciones por la distancia entre cada par de categorías (ver Tabla 6.1). Además

esta medida incluye dos parámetros que permiten otorgar importancia relativa tanto a la distancia entre categorías, como al producto de sus probabilidades.

En el siguiente ejemplo, computamos esta medida para el conjunto de datos `scidat`, otorgando mayor relevancia a la distancia entre categorías ($\alpha = 0.6$) y reduciendo la importancia de la multiplicación de probabilidades entre categorías ($\beta = 0.4$).

```
diversity(data=scidat, type='rs', method='cosine', alpha = 0.6, beta = 0.4)
              rao.stirling
Argentina         4.316475
China             3.731211
Germany           4.697477
Hungary           4.700062
Iran              4.355324
Mexico            4.505785
Singapore         4.566474
South Korea       4.381553
Spain             4.898211
United States     4.899825
```

La importancia de esta medida, que se puede calcular gracias a la noción de distancia entre categorías —como la que hemos aportado en esta tesis—, es que permite sintetizar en un número, aquel análisis que de otra forma se puede realizar solo cualitativamente (como el realizado en la Sección 6.1.1). El valor de la diversidad Rao-Stirling, recoge la noción de cuántas áreas se producen, la cantidad de producción en cada área y también las diferencias entre las áreas producidas.

## 6.2. *OPUS*: Visualizando la diversificación, colaboración y desarrollo científico

La materia prima con la que se llevó a cabo esta tesis, esto es, el conjunto de datos de publicaciones científicas, desambiguado a nivel de individuos y su conexión con áreas de la ciencia, ha suscitado el interés de científicos y agencias de estado, debido a la inexistencia de conjuntos de datos de similares caracterísiticas y sus posibilidades, como la investigación que se ha llevado a cabo en esta tesis. Es por esto, que en el marco de este trabajo de tesis, se ha propuesto la creación de la aplicación *OPUS* —palabra en latín que significa *obra*—, que permitirá visualizar la producción científica de individuos, instituciones y países, tanto en términos de publicaciones, cuanto en términos de colaboración, diversificación y desarrollo por las distintas áreas de la ciencia.

El desarrollo de la aplicación se encuentra a cargo del grupo *macroconnections* del MediaLab en el MIT y se espera su lanzamiento en el segundo semestre del año 2016 en el URL http://opus.media.mit.edu/. La versión actual (0.2), incluye las siguientes secciones operativas a nivel de individuos:

Mitchel Resnick

Massachusetts Institute of Technology

Mitchel Resnick started publishing in the year 1980, and to this date has produced 114 papers. As of today, Mitchel Resnick's h-index is 51.

Mitchel Resnick's main fields are Computer Science, Social Sciences, Engineering

**Figura 6.3:** Sección biografía de la aplicación OPUS. Incluye una mini biografía generada automáticamente del conjunto de datos.

### 6.2.1. Biografía

La primera sección de OPUS incluye la fotografía del científico, su nombre, la filiación actual y una breve biografía que describe desde cuándo ha comenzado su actividad científica y los valores asociados a cantidad de publicaciones y rendimiento en cuanto a citas conseguidas. También se describen los principales campos de investigación en los que trabaja. La Figura 6.3 presenta esta sección para el científico Mitchel Resnick, conocido por ser uno de los creadores de Scratch y que actualmente dirige el grupo *Longlife Kindergarten* en el MediaLab. Nótese que la minibiografía tiene como objetivo, además de entregar información al usuario, el de optimizar la indexación por parte de los motores de búsqueda como Google.

### 6.2.2. Publicaciones y citas

La sección de Publicaciones y Citas incluye dos subsecciones. En la primera se presentan las publicaciones del científico utilizando una visualización tipo *treemap* en la que el color de las cajas indica el área de la ciencia en la que está asignada cada publicación, mientras que el tamaño es proporcional al número de citas conseguidas en esa publicación, lo que da cuenta de cuán exitosa (entendiendo éxito como el número de citas) ha sido cada publicación. En la Figura 6.4 se presenta la primera parte de esta sección para nuestro científico de ejemplo. Nótese que la mayoría (más del 50 %) de publicaciones son en el área de computación (color rojo) y otra buena parte son en el área de Educación (más del 25 %). Destacan justamente las publicaciones en el área de Scratch.

La segunda parte de la sección de publicaciones incluye gráficos de barras anuales tanto para publicaciones como para citas. Además estos gráficos se encuentran vinculados funcionalmente entre sí y con el treemap, lo que permite ir filtrando por año (ver Figura 6.5). Esta última funcionalidad facilita el análisis de cómo se ha ido diversificando un científico a través de las áreas de investigación y a lo largo de los años. En la Figura 6.6 se presenta la diversificación de Mitchel Resnick entre 1980 y 2014. Nótese como este científico en particular, es también un gran contribuidor de la denominada "literatura gris" (color gris), esto es, capítulos de libros, artículos de opinión, libros, entre otros documentos que no son artículos científicos. La diversificación de Resnick va ocurriendo desde la Economía (color morado) hacia áreas


**Figura 6.4:** Primera parte de la sección Publicaciones de la aplicación OPUS. Incluye una visualización tipo treemap donde cada caja representa una publicación del científico. Los colores representan el área de la ciencia donde está indexada la publicación y el tamaño de cada caja representa el número de citas ganadas.

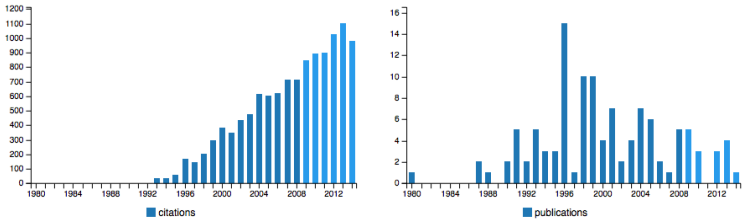



**Figura 6.5:** Segunda parte de la sección Publicaciones de la aplicación OPUS. Incluye dos gráficos de barras en el tiempo. Uno para publicaciones y otro para citas. En este ejemplo se han elegido (color azul claro) solamente los años entre 2009 y 2013.

de ciencias de la Computación (color rojo), educación (color Celeste), las humanidades y las artes (color naranjo). En los últimos años, su presencia más fuerte es en ciencias de la computación.

### 6.2.3. Colaboración

La colaboración —rescatada a través de la relación de coautoría en *papers*— se representa en Opus con una visualización tipo red, en la que los nodos corresponden a científicos -—registrados en Google Scholar— y los enlaces indican la cantidad de *papers* compartidos entre autores. En la red que se presenta en la Figura 6.7 se presenta la red de colaboración de nuestro científico de ejemplo. Esta visualización permite además identificar el país de procedencia de los colaboradores y detectar comunidades de coautoría que se representan por el color del círculo de los nodos. En este caso, Resnick es coautor de tres comunidades que también se encuentran vinculadas entre sí, estas son las comunidades de color celeste, naranjo y celeste claro. También se aprecia que la mayor colaboración es con autores procedentes de Estados Unidos, pero que también incluye autores de Canadá, Israel o Reino Unido.

Una funcionalidad adicional es que se puede elegir el país de colaboración para visualizar los autores asociados a ese país. Esto se presenta en las imágenes de la Figura 6.8 donde


**Figura 6.6:** Diversificación en el tiempo del científico Mitchel Resnick a través de las áreas de la ciencia. Cada panel representa un treemap para las publicaciones en la ventana de tiempo elegida. Se han tomado el período de tiempo comprendido entre 1980 y 2013 agregando el conjunto de datos cada 5 años (contabilizando solo años que tienen publicaciones) y moviendo la ventana de tiempo cada tres años. Los colores representan áreas de la ciencia. Rojo para Computación, celeste para Educación y naranjo para Humanidades.

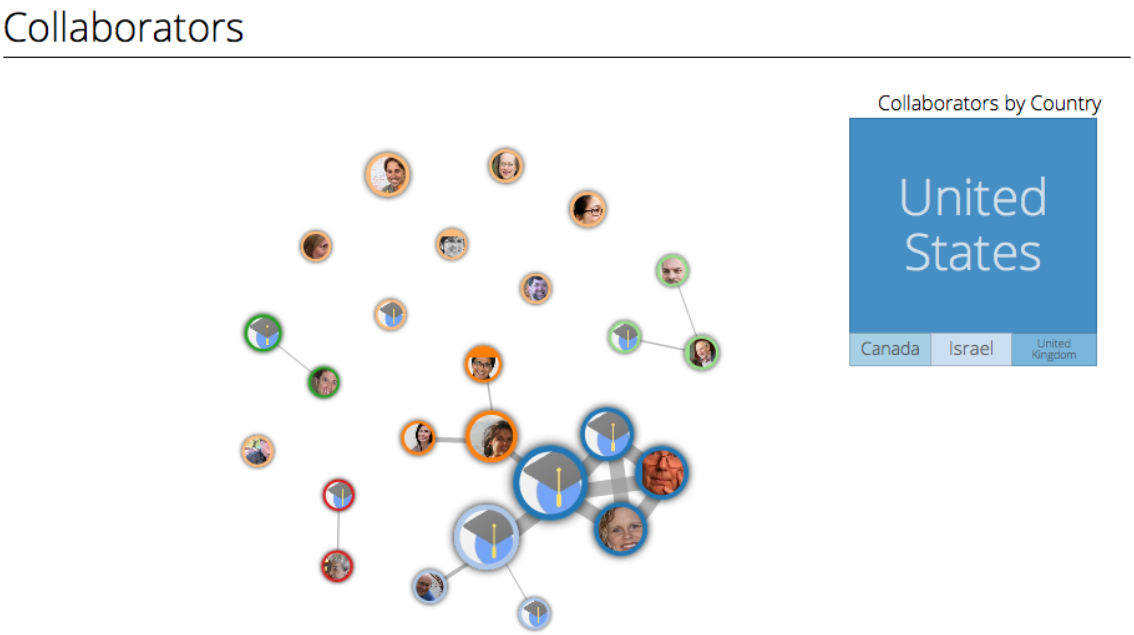



**Figura 6.7:** Red de colaboración incluida en la aplicación OPUS. Los nodos representan coautores mientras que los enlaces representan cantidad de *papers* que han publicado juntos. Los colores de los nodos representan comundiades de coautoría detectadas por el algoritmo *FastGreedy*. El tamaño de los nodos es proporcional a la cantidad de *papers* publicados con el autor "ego" que no se grafica en la red. Acompaña a la red, un treemap que da cuenta de la cantidad de coautores por país.

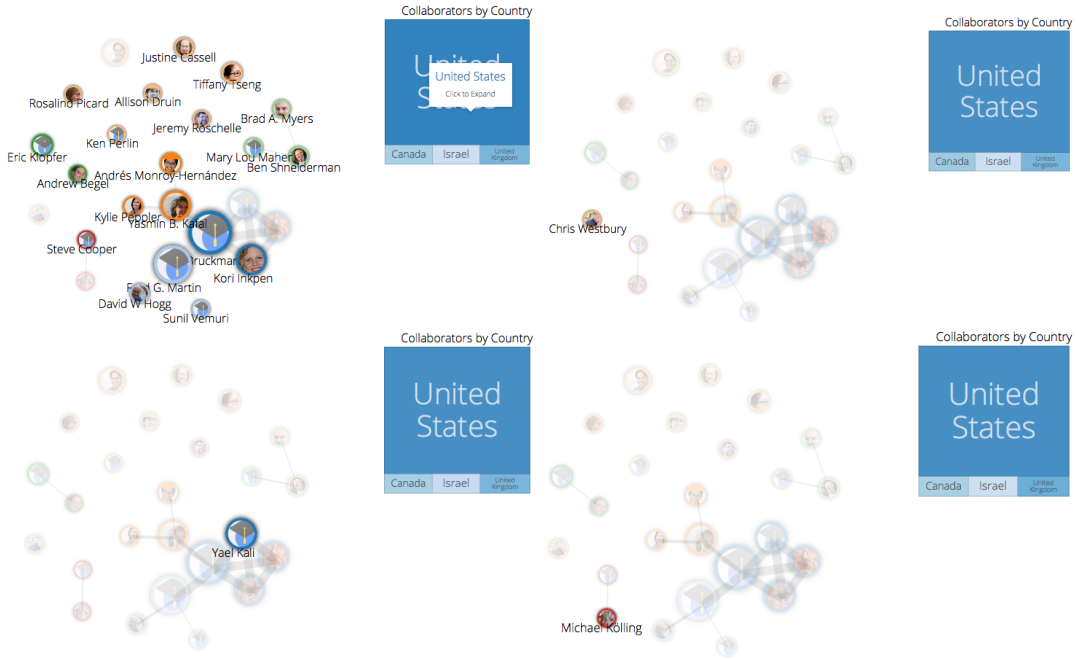



**Figura 6.8:** Red de colaboración incluida en la aplicación OPUS filtrada por país. Se presenta en orden por filas, los países Estados Unidos, Canadá, Israel y Reino Unido.

al pasar el mouse sobre el treemap de colaboración internacional se destacan los autores asociados a cada país.

Adicionalmente, si se dispone de información de la filiación de cada autor, al hacer click sobre el país, se baja un nivel de granularidad hasta la cantidad de autores por Institución. También se puede pasar el mouse por cada institución para destacar la filiación de los autores adscritos a esa institución (ver Figura 6.9).

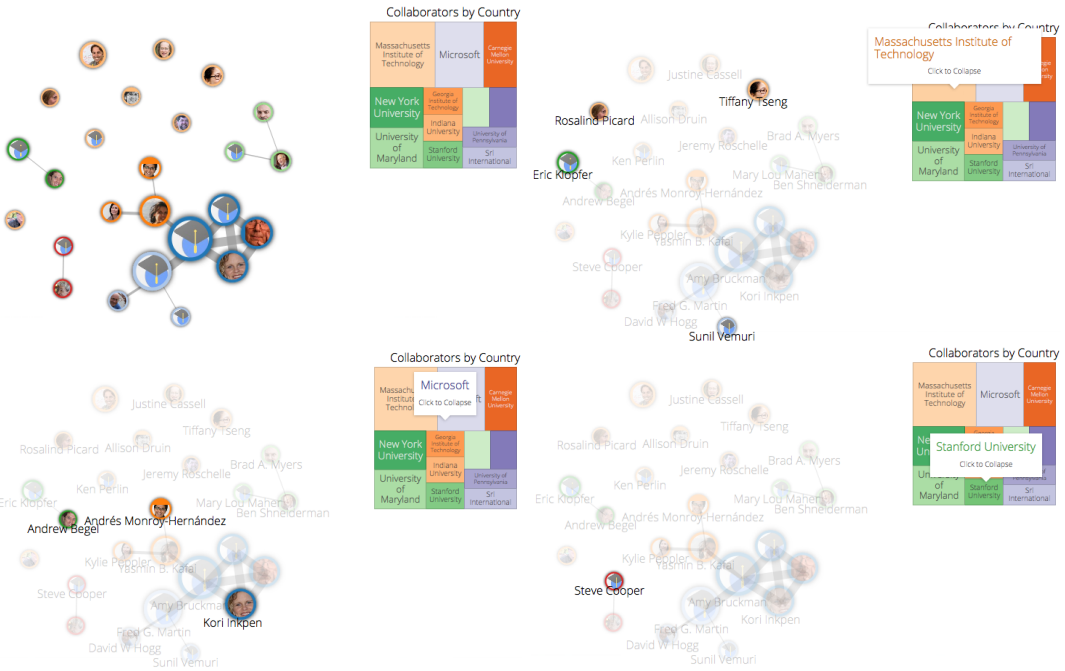



**Figura 6.9:** Red de colaboración con detalle de instituciones a nivel de país (superior izquierda) para los coautores de Mitchel Resnick. Las figuras restantes se encuentran filtradas por institución: MIT, Microsoft y Stanford.

Capítulo 7

# Conclusiones

En este capítulo revisamos cómo se cumplieron o no las afirmaciones que planteamos guiados por nuestra hipótesis investigativa. También comentamos los principales hallazgos de esta investigación para finalmente abordar líneas de trabajo futuro.

## 7.1. Análisis de resultados

El objetivo principal de esta investigación fue determinar la estructura de una red de categorías de la ciencia basada en las capacidades productivas de los científicos. Este objetivo fue satisfactoriamente cumplido y para lograrlo fue necesario la construcción de un nuevo conjunto de datos que permitiera identificar la trayectoria científica de individuos a nivel de áreas de la ciencia. En el Capítulo 4 se detallan los pasos seguidos y las redes contruídas, tanto en clasificación UCSD como en clasificación SCImago.

En base a este objetivo principal, nos planteamos la hipótesis de investigación que propone que un mapa basado en trayectorias productivas es mejor descriptor de la diversificación y el desarrollo de áreas de la ciencia, en comparación con una red basada en patrones de citación.

Logramos validar esta hipótesis a nivel de individuos e instituciones tanto para diversificación (aparecimiento de nuevas áreas) como para desarrollo (incremento de ventajas comparativas). Sin embargo a nivel de países esta hipótesis no se cumple.

En la intención de verificar la robustez de nuestro método, también nos propusimos evaluar el espacio investigación en una clasificación distinta (SCImago), en la que también se obtuvieron resultados comparables a los obtenidos con la clasificación UCSD. Además, los valores obtenidos en clasificación SCImago lograron ser mejores que otra red en esa clasificación (basada en co-ocurrencia) a la hora de medir diversificación de individuos e instituciones. Esto no se cumplió en el caso del desarrollo de países, donde los resultados no lograron ser significativamente mejores que la red en comparación.

En las secciones siguientes se detallarán los principales hallazgos suscitados a partir de la verificación de nuestra hipótesis. Los resultados principales de esta tesis se comunicaron en

la revista Scientometrics [Guevara et al., 2016a].

## 7.2. Principales hallazgos

### 7.2.1. Mejor descriptor que mapas basados en citas

En concordancia a lo que habíamos planteado al inicio de esta tesis, el espacio investigación resultó ser una mejor red para evaluar la diversificación y el desarrollo de productores de ciencia. Sin embargo, esta afirmación se cumplió solamente a nivel de individuos e instituciones y no a nivel de países. Situación que cobra sentido, por cuanto a medida que se van agregando capacidades de personas para representar la producción de instituciones o países, se pierde fineza en la caracterización de las capacidades productivas, lo que redunda en que a mayores niveles de agregación, la predicción de la diversificación y el desarrollo de los productores se pueda realizar con similares resultados utilizando cualquiera de los dos mapas. Vale decir, las bondades del espacio investigación se aprecian más mientras mejor es la resolución de los productores de ciencia.

Esto conlleva también a validar el hecho de que las redes que representan flujos de información, como el mapa UCSD, y que se originaron esencialmente con la idea de clasificar el conocimiento, no necesariamente representan las capacidades productivas o las habilidades de los productores, hecho que hemos comprobado al poner a prueba su capacidad para medir la diversificación y el desarrollo de productores a distintos niveles.

### 7.2.2. Ortogonalidad en relación a un mapa de citas

Al inicio de esta investigación planteábamos la necesidad de un nuevo enfoque para la construcción de mapas de la ciencia, por cuanto los mapas basados en patrones de citas, básicamente representan los flujos de información o conocimiento entre áreas, más que las habilidades de los productores de ciencia para diversificarse de un área a otra. Vale decir un biólogo bien puede citar un *paper* en estadística de algún método que utilizó para validar sus hipótesis, pero esto no implica, necesariamente, que este biólogo es capaz —tiene las capacidades o el deseo— de publicar en estadística, y viceversa, no necesariamente un estadístico tiene las capacidades para publicar en Biología.

Para develar estas diferencias entre los dos métodos para construir mapas de la ciencia, uno basado en patrones de citas, y el otro basado en capacidades productivas; hemos calculado la correlación entre los enlaces del mapa UCSD y los enlaces del espacio investigación, encontrando que efectivamente son ortogonales ($R = 0.038$), vale decir, hay enlaces que están sobrevalorados en el mapa UCSD y lo propio en el espacio investigación (ver Figura 7.1).

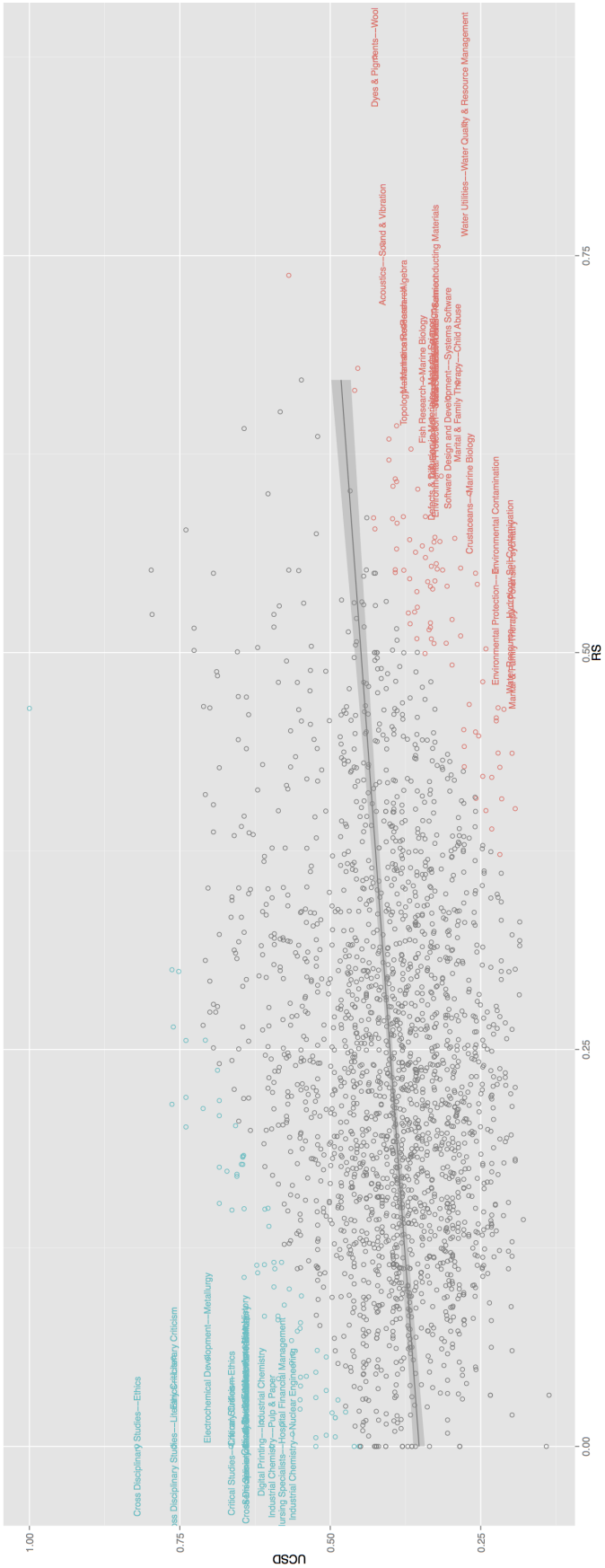



**Figura 7.1:** Análisis de correlación entre los enlaces del mapa UCSD y los enlaces del espacio investigación. En color verde se destacan enlaces sobrevalorados por mapa UCSD mientras que en color naranjo se destacan enlaces sobrevalorados por el espacio investigación. Coeficiente de correlación R=0, 038.

### 7.2.3. Robustez en cuanto a la clasificación

A la hora de cambiar de clasificación, esto es, desde UCSD a SCImago, el espacio investigación obtuvo buenos resultados en cada transición medida para cada productor, que si bien no son diferencias ampliamente significativas (valor-p <0.05) en comparación con el mapa SCImago, sí fueron mejores medidos en promedio o a través de la mediana. Además ninguna de las transiciones medidas a distintos niveles de granularidad, fue menor que 0.65 en promedio del valor bajo la curva ROC, y a nivel de individuos, este valor supera los 0.85 lo que es bastante bueno para un sistema de recomendación.

En resumen, hemos podido validar que nuestro método es robusto a la hora de cambiar de clasificación, incluso al tratarse de una clasificación de casi la mitad de categorías, en comparación con la clasificación UCSD (229 versus 554) lo que podría haberse manifestado en menores diferencias entre mapas.

### 7.2.4. Comportamiento de la diversificación

Un hallazgo necesario de destacar es que resulta más fácil predecir la diversificación de individuos que la diversificación de instituciones o países, situación que sucede en ambas clasificaciones y no solo con el espacio investigación, sino también con los mapas UCSD y SCimago. Esto básicamente porque los movimientos de un individuo entre áreas de la ciencia están limitados a un número reducido de áreas que no requieren habilidades tan distintas a las áreas en las que ya ha publicado anteriormente.

De nuestro conjunto de datos podemos aprender por ejemplo que de los 29,856 científicos que han publicado al menos dos *papers* en Biología Molecular, 45.6 % de ellos también han publicado en Bioquímica Clínica, pero solamente 1.2 % han publicado también en Economía y Econometría. Esto hace ver que los saltos (diversificación) de un área a otra, cuando se trata de individuos, son más predecibles por cuanto ocurren en un reducido conjunto de categorías cercanas.

En el caso de instituciones, cuya productividad es básicamente la agregación de la productividad de los científicos adscritos a la institución, la calidad de la predicción de la diversificación es más baja en todos los mapas, lo que empeora a nivel de países, donde en muchos casos las posibilidades de diversificación son nulas, por cuanto ya se han ocupado todas las categorías de la ciencia, salvo en países con menor desarrollo.

### 7.2.5. Comportamiento del desarrollo

En términos de desarrollo —medido con RCA— es interesante notar que a nivel de instituciones, en las dos clasificaciones, el espacio investigación entrega mejores resultados (mayor valor promedio de las áreas bajo la curva ROC) en la transición de áreas Nacientes ($RCA < 0.5$) hacia áreas desarrolladas ($RCA > 1$) que de áreas Intermedias (RCA entre 0.5 y 1) hacia áreas Desarrolladas. Vale decir, es capaz de predecir de mejor forma un salto de áreas no tan competitivas que se convierten en áreas competitivas en comparación con otras instituciones.

Esto no sucede a nivel de países donde las dos transiciones se predicen con niveles comparables. Esto es explicable en el sentido de que la dinámica de las instituciones puede tender a ser más variable en el tiempo, según la cantidad de proyectos que adjudican sus investigadores o en función de las nuevas contrataciones de científicos.

Sin embargo a nivel de países, estos cambios son menos drásticos, por cuanto los movimientos de científicos generalmente son locales y cuando son externos al país, tampoco producen un impacto considerable en la producción científica de ese país en las distintas áreas, la que sí se puede ver afectada por políticas públicas más transversales como por ejemplo el porcentaje de la inversión del PIB en Investigación y Desarrollo o la reorganización de la estructura organizativa del país, como la creación de entidades administrativas como agencias o ministerios de Ciencia y Tecnología.

### 7.2.6. Diversidad en la producción científica

Hemos realizado una importante contribución a la medición de la diversidad científica, por cuanto hemos aportado un mapa que deja entrever las similitudes y diferencias entre áreas desde un nuevo enfoque. Este aporte facilita la medición de diversidad con medidas que consideran la *disparidad* entre sus componentes. La utilización de este aporte lo hemos facilitado a través de la creación de una aplicación de software que permite calcular las medidas más utilizadas de diversidad. Los detalles de esta aplicación se encuentran en la Sección 6.1.

### 7.2.7. Conjunto de datos para analizar la producción científica

El conjunto de datos que debimos construir para verificar las hipótesis de esta tesis, es un conjunto de datos pionero por cuanto, primero, son datos a nivel de individuo; segundo, se encuentra desambiguado (respecto de los nombres) y tercero, mapea cada persona con las áreas científicas en las que ha publicado. Este conjunto de datos, puede facilitar la realización de nuevos estudios y aplicaciones, que no se habían podido conducir previamente debido a la poca disponibilidad de datos de este tipo a este nivel de resolución. Producto de esta tesis nos encontramos desarrollando una aplicación para visualizar la producción y diversificación de individuos, instituciones y países. Los detalles de esta aplicación se encuentran en la Sección 6.2.

### 7.2.8. Mapas superpuestos

Hemos avanzado la visualización de mapas superpuestos, dotando de semántica a cada nodo en el espacio investigación de los productores evaluados. También fuimos capaces de generar de manera masiva estos mapas, situación que en trabajos anteriores estaba limitada a pocos casos de estudio.

### 7.2.9. Método cuantitativo para evaluar mapas de la ciencia

Hemos propuesto un método cuantitativo para la evaluación de mapas de la ciencia como herramientas de predicción de áreas de desarrollo o nueva actividad. Este método fue propuesto originalmente para redes de complejidad económica [Hidalgo et al., 2007] y su aplicación al espacio investigación nos permitió evaluar su desempeño y también compararlo con otros mapas de la ciencia.

## 7.3. Trabajo futuro

Los hallazgos de esta tesis, facilitarán la investigación de nuevos tópicos relacionados, así como la creación de nuevas aplicaciones de uso general. En esta sección detallamos estos dos tipos de trabajo futuro que se pueden desarrollar a partir del trabajo de esta tesis.

### 7.3.1. Nuevas clasificaciones

Una evolución natural del espacio investigación es la utilización de nuevas clasificaciones, siendo que existen otras clasificaciones que son ampliamente usadas, dependiendo del país o la institución. Por ejemplo, la clasificación de campos de la ciencia de la OECD o la clasificación de las áreas de investigación FoR de la Excellence in Research for Australia (ERA), también de empresas privadas como la clasificación de categorías de la ciencia de Thomson Reuters$^{TM}$. Este trabajo futuro se puede desarrollar en la medida que las clasificaciones de *journals* en categorías se creen en algunos casos y se hagan disponibles, en otros, situación que al término de esta tesis aún no sucedía.

### 7.3.2. Mapas locales

Dotar de mayor *localidad* al espacio investigación, esto es mayor nivel de detalle en el contexto propio de una disciplina, es un camino posible con el conjunto de datos y el método utilizado. Esto es, crear mapas locales a nivel de disciplina, permitiría conducir análisis al interior de cada disciplina, descubrir sus características de diversificación y desarrollo, así como también comparar esas características con otras disciplinas. Varias áreas de la ciencia, ya cuentan con clasificaciones propias que están ampliamente difundidas, como en el caso de la clasificación DBLP en computación.

### 7.3.3. Nuevas medidas para predecir diversificación y desarrollo

En esta tesis hemos aplicado la medida de densidad para predecir la futura activación o desarrollo de áreas de la ciencia. Esta medida toma en consideración las áreas inactivas, evaluando qué tan cercanas se encuentran de áreas ya activadas. Vale decir, es un enfoque desde los nodos inactivos. Sería interesante evaluar otras posibles medidas de activación, tomando en cuenta por ejemplo, como punto de partida los nodos activos en lugar de los inactivos u otras ideas basadas por ejemplo en modelos de contagio en redes complejas.

### 7.3.4. Análisis de la dinámica del espacio investigación

No hemos abordado en esta tesis, aunque es perfectamente plausible, analizar la dinámica de cómo varía la estructura del espacio investigación en el tiempo. En esta tesis contruimos el espacio investigación con datos agregados entre los años 1971 a 2010. Un tópico posible de investigar, es analizar cómo varía la estructura del espacio investigación, año a año. Esto es por ejemplo, cómo varía el ranking de enlaces incluidos en el Árbol Recubridor Mínimo (MST) que daría cuenta de qué áreas mantenían mayor cercanía año a año, al igual que evaluar por ejemplo la centralidad de cada nodo a medida que avanza el tiempo. Este tópico de investigación permitiría responder preguntas acerca de cómo ha ido evolucionando la ciencia a lo largo de los años.

### 7.3.5. Similaridades entre productores

Si bien en esta tesis, nos hemos concentrado en las similitudes entre categorías de la ciencia, es posible evaluar la similitud entre productores, basados en las áreas en las que producen. Esto permitiría la conformación de clusters o comunidades de productores similares, entre los que sería posible realizar comparaciones o encontrar características que hacen que sean similares, o también complementarios, situación interesante de analizar, sobre todo si se busca recomendar alianzas para proyectos multidisciplinarios. Hemos realizado un estudio exploratorio de esta línea de investigación para los denominados países BRIC, Brazil, Rusia, India y China [Guevara and Mendoza, 2016].

### 7.3.6. Estudio de casos

El espacio investigación propicia el estudio de casos particulares, por ejemplo de instituciones o países, en los que se pueden evaluar sus políticas institucionales o públicas, a la luz del desarrollo de sus áreas científicas, con el objetivo de *recomendar* futuros caminos de desarrollo o inversión. Sin embargo, esto conlleva consigo también el desafío de ser capaz de cuantificar cuál es el nivel de inversión que se requiere para cada área de investigación, puesto que se conoce de antemano que algunas áreas de investigación requieren más recursos que otras para desarrollarse, como es el caso de áreas que requieren de altos niveles de inversión en equipamiento tecnológico —como el caso de la Biotecnología o la Neurociencia— versus otras que no —como el caso de la Economía o la Sociología—.

### 7.3.7. Analizar datos propios

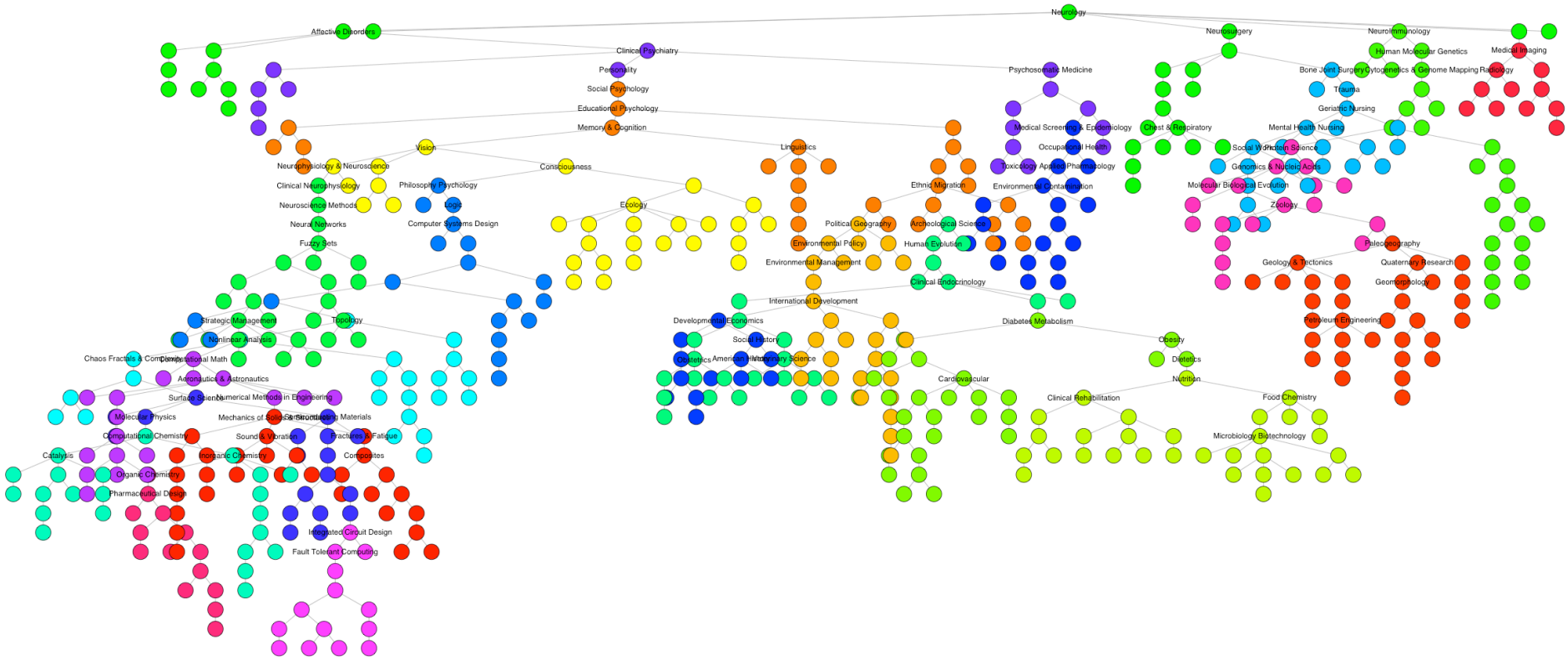
Si bien producto de esta tesis, hemos construido dos aplicaciones (ver Capítulo 6) una tendiente a facilitar la medición de la diversidad y la otra que facilita la visualización de la diversificación y el desarrollo de los productores de ciencia; se ha detectado la necesidad tanto de instituciones como de personas de analizar su propio conjunto de datos de productividad científica, la misma que no necesariamente constituye trabajos científicos sino que puede abarcar un amplio espectro de productos asociados al quehacer académico y científico, tales como: proyectos, consultorías, noticias, artículos de opinión, libros, coloquios, seminarios y charlas, software o datos de patentes.

Surge entonces la necesidad de construir una aplicación que permita al usuario ingresar datos propios y que estos datos sean mapeados y analizados en base al espacio investigación que dé cuenta de la posición actual de la institución o persona, de su productividad y diversificación en el tiempo, y que además recomiende áreas de futuro desarrollo.

Apéndice A

# Ejemplo de Arbol Recubridor Mínimo. Imagen en alta resolución

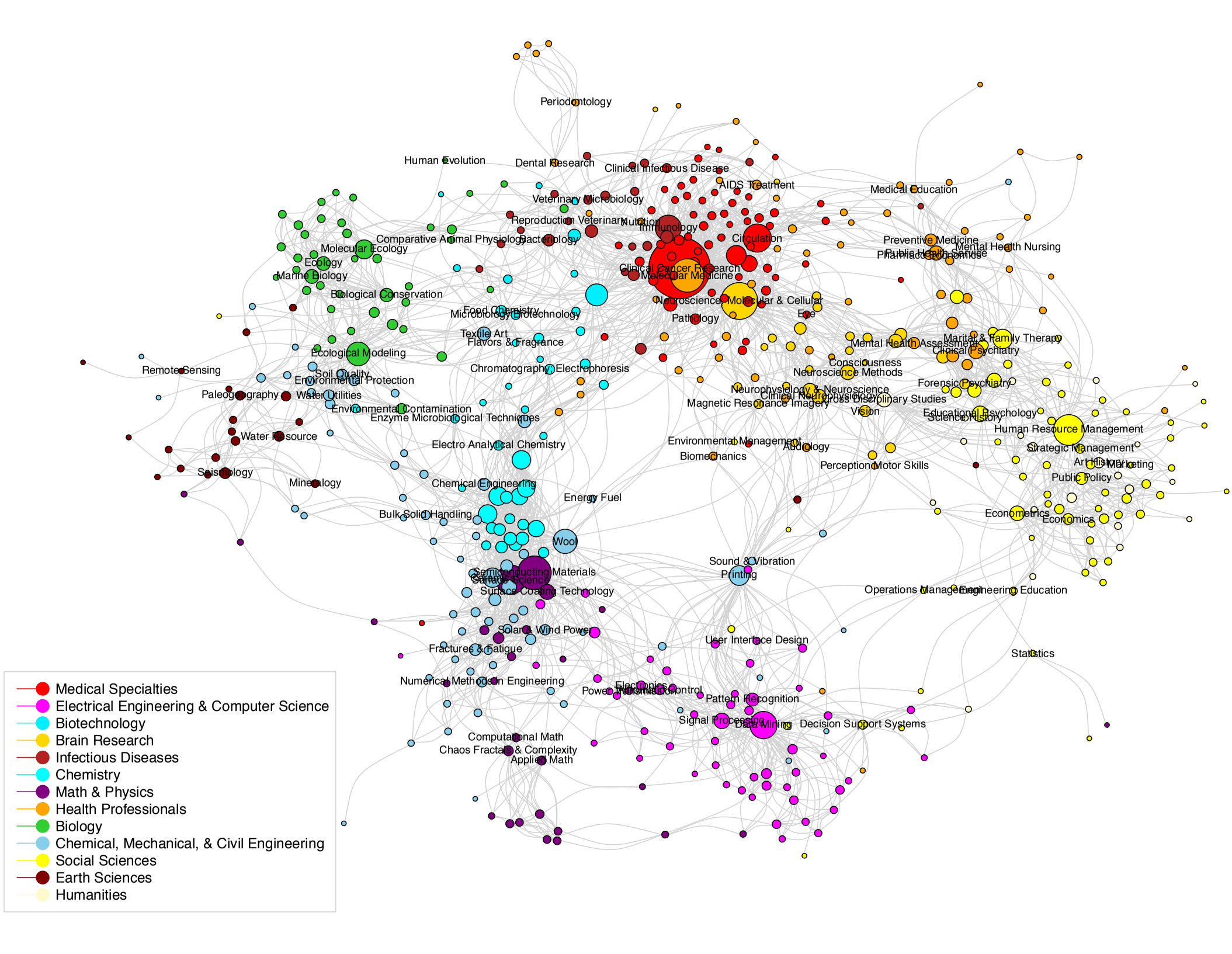
**Research Space**
**Data: GSCHOLAR | Taxonomy: UCSD | Interval: 2000 - 2009**
**Minimum Spanning Tree**




Number of nodes: 546 / 554 | Mean degree: 1 | Detected communities: 24

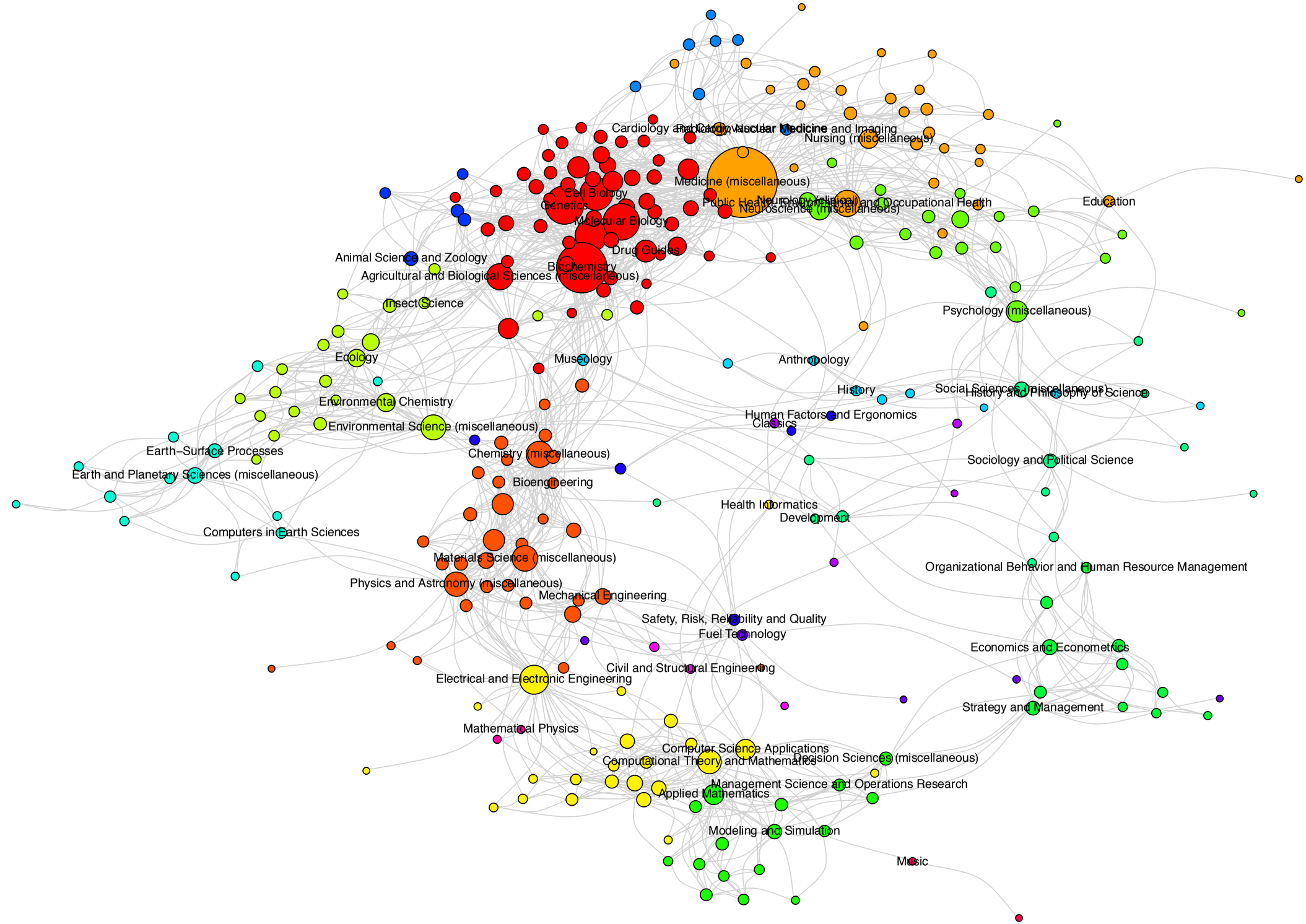
Apéndice B

# Espacio investigación en clasificación UCSD. Imagen en alta resolución


Periodontology
Human Evolution
Dental Research
Clinical Infectious Disease
AIDS Treatment
Medical Education
Veterinary Microbiology
Reproduction Veterinary
Nutrition
Immunology
Circulation
Comparative Animal Physiology
Bacteriology
Preventive Medicine
Public Health
Mental Health Nursing
Pharmacoepidemiology
Molecular Ecology
Ecology
Marine Biology
Clinical Cancer Research
Molecular Medicine
Biological Conservation
Neuroscience, Molecular & Cellular
Food Chemistry
Microbiology Biotechnology
Pathology
Eye
Textile Art
Flavors & Fragrance
Marital & Family Therapy
Mental Health Assessment
Clinical Psychiatry
Ecological Modeling
Remote Sensing
Soil Quality
Chromatography, Electrophoresis
Consciousness
Neuroscience Methods
Environmental Protection
Forensic Psychiatry
Paleoceanography
Water Utilities
Neurophysiology & Neuroscience
Clinical Neurophysiology
Cross Disciplinary Studies
Magnetic Resonance Imaging
Vision
Environmental Contamination
Enzyme Microbiological Techniques
Educational Psychology
Science History
Human Resource Management
Water Resource
Electro Analytical Chemistry
Environmental Management
Audiology
Strategic Management
Biomechanics
Art History
Marketing
Seismology
Perception
Motor Skills
Public Policy
Mineralogy
Chemical Engineering
Energy Fuel
Econometrics
Economics
Bulk Solid Handling
Wool
Sound & Vibration
Printing
Semiconducting Materials
Surface Science
Operations Management
Engineering Education
Surface Coating Technology
Solar & Wind Power
User Interface Design
Statistics
Fractures & Fatigue
Numerical Methods in Engineering
Electronics
Power Transmission Control
Pattern Recognition
Signal Processing
Data Mining
Decision Support Systems
Computational Math
Chaos Fractals & Complexity
Applied Math
Medical Specialties
Electrical Engineering & Computer Science
Biotechnology
Brain Research
Infectious Diseases
Chemistry
Math & Physics
Health Professionals
Biology
Chemical, Mechanical, & Civil Engineering
Social Sciences
Earth Sciences
Humanities

Apéndice C

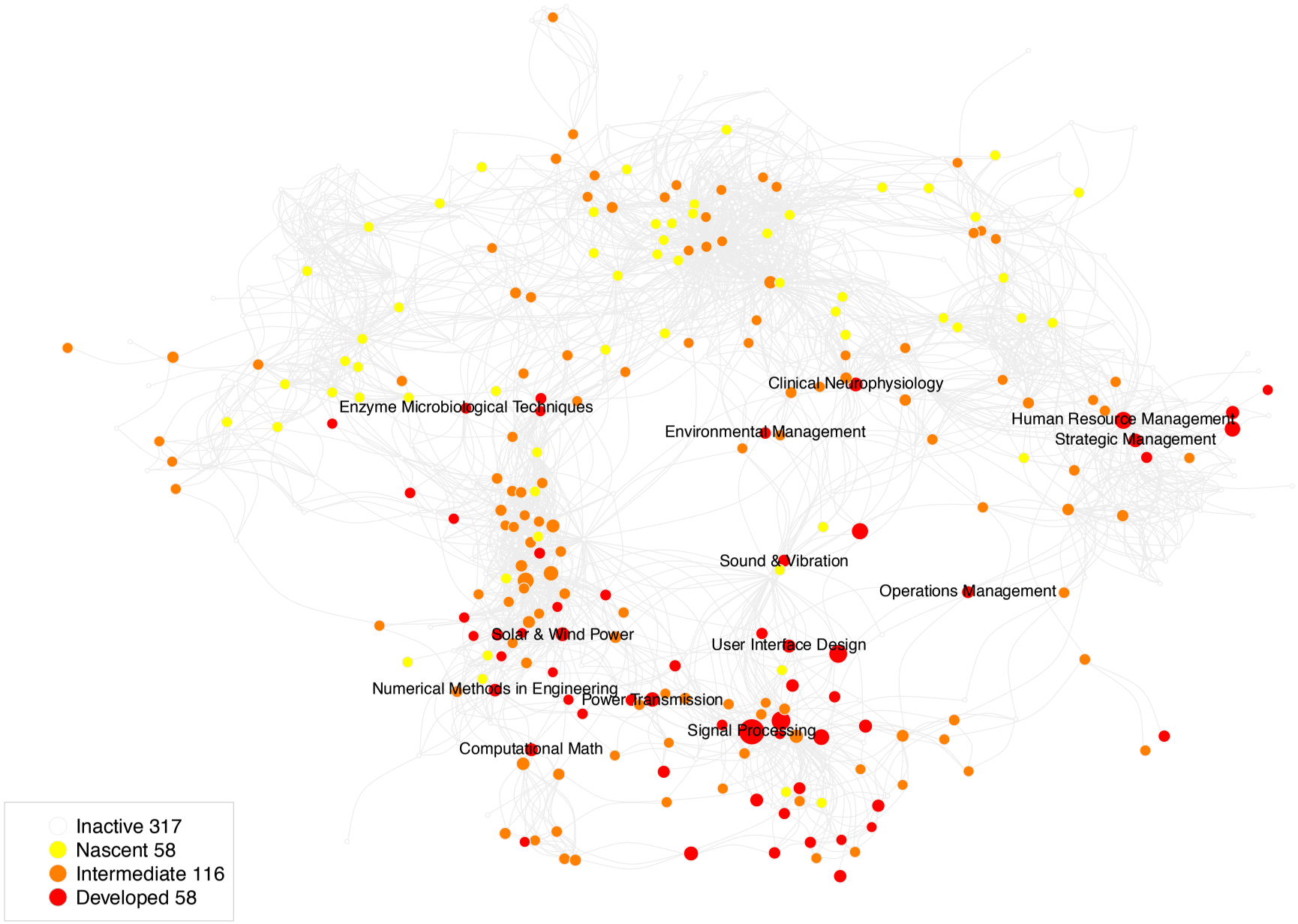
# Espacio investigación en clasificación SCImago. Imagen en alta resolución


Cardiology and Medicine
Nursing (miscellaneous)
Medicine (miscellaneous)
Cell Biology
Genetics
Molecular Biology
Public Health, Environmental and Occupational Health
Neuroscience (miscellaneous)
Education
Drug Guides
Animal Science and Zoology
Biochemistry
Agricultural and Biological Sciences (miscellaneous)
Insect Science
Psychology (miscellaneous)
Ecology
Mycology
Anthropology
History
Social Sciences (miscellaneous)
History and Philosophy of Science
Environmental Chemistry
Human Factors and Ergonomics
Classics
Environmental Science (miscellaneous)
Earth–Surface Processes
Chemistry (miscellaneous)
Sociology and Political Science
Earth and Planetary Sciences (miscellaneous)
Bioengineering
Health Informatics
Development
Computers in Earth Sciences
Materials Science (miscellaneous)
Organizational Behavior and Human Resource Management
Physics and Astronomy (miscellaneous)
Mechanical Engineering
Safety, Risk, Reliability and Quality
Fuel Technology
Economics and Econometrics
Civil and Structural Engineering
Electrical and Electronic Engineering
Strategy and Management
Mathematical Physics
Computer Science Applications
Computational Theory and Mathematics
Decision Sciences (miscellaneous)
Management Science and Operations Research
Applied Mathematics
Modeling and Simulation
Music

Apéndice D

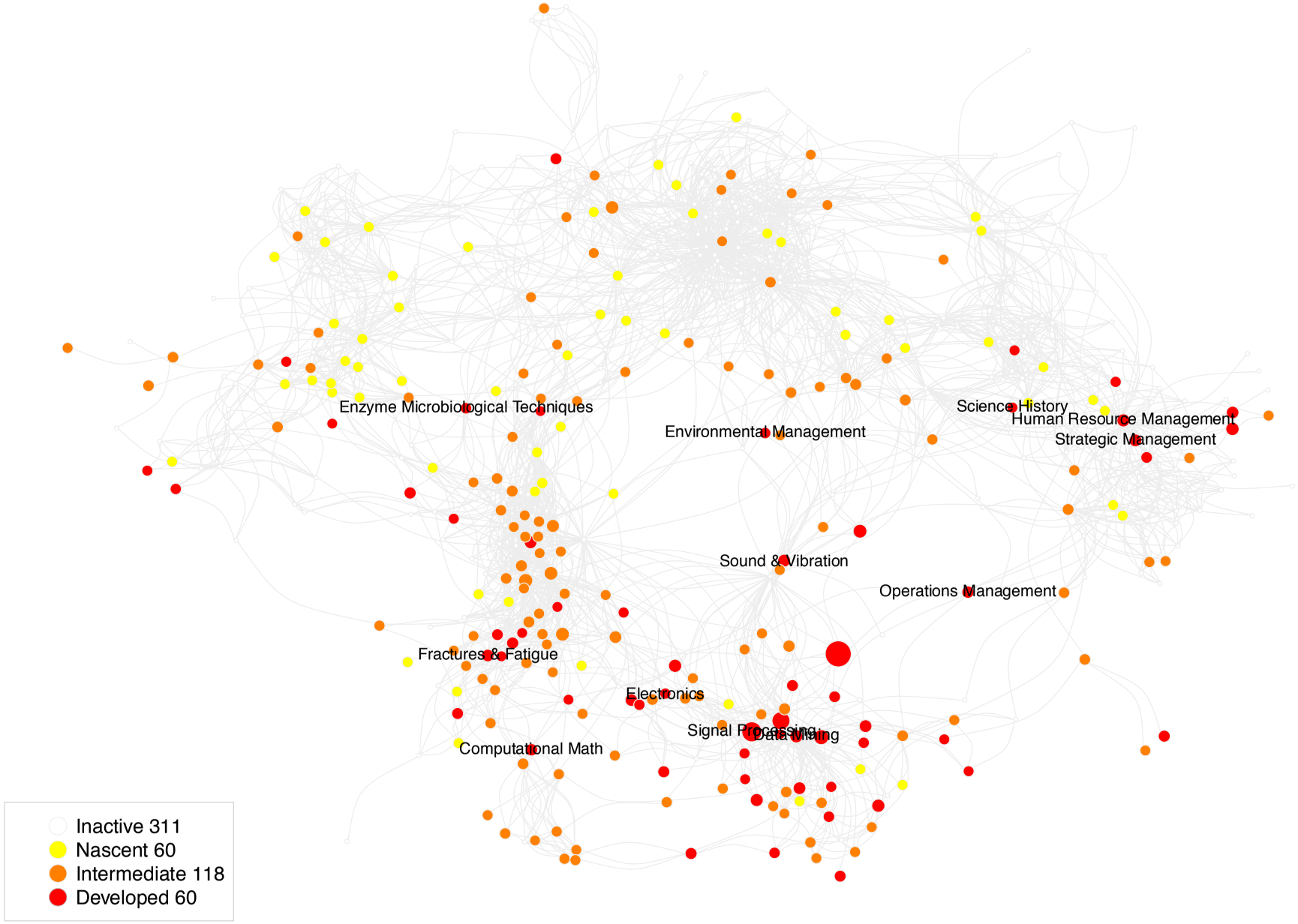
# Mapas superpuestos para instituciones

**Aalto University / Aalto–yliopisto 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



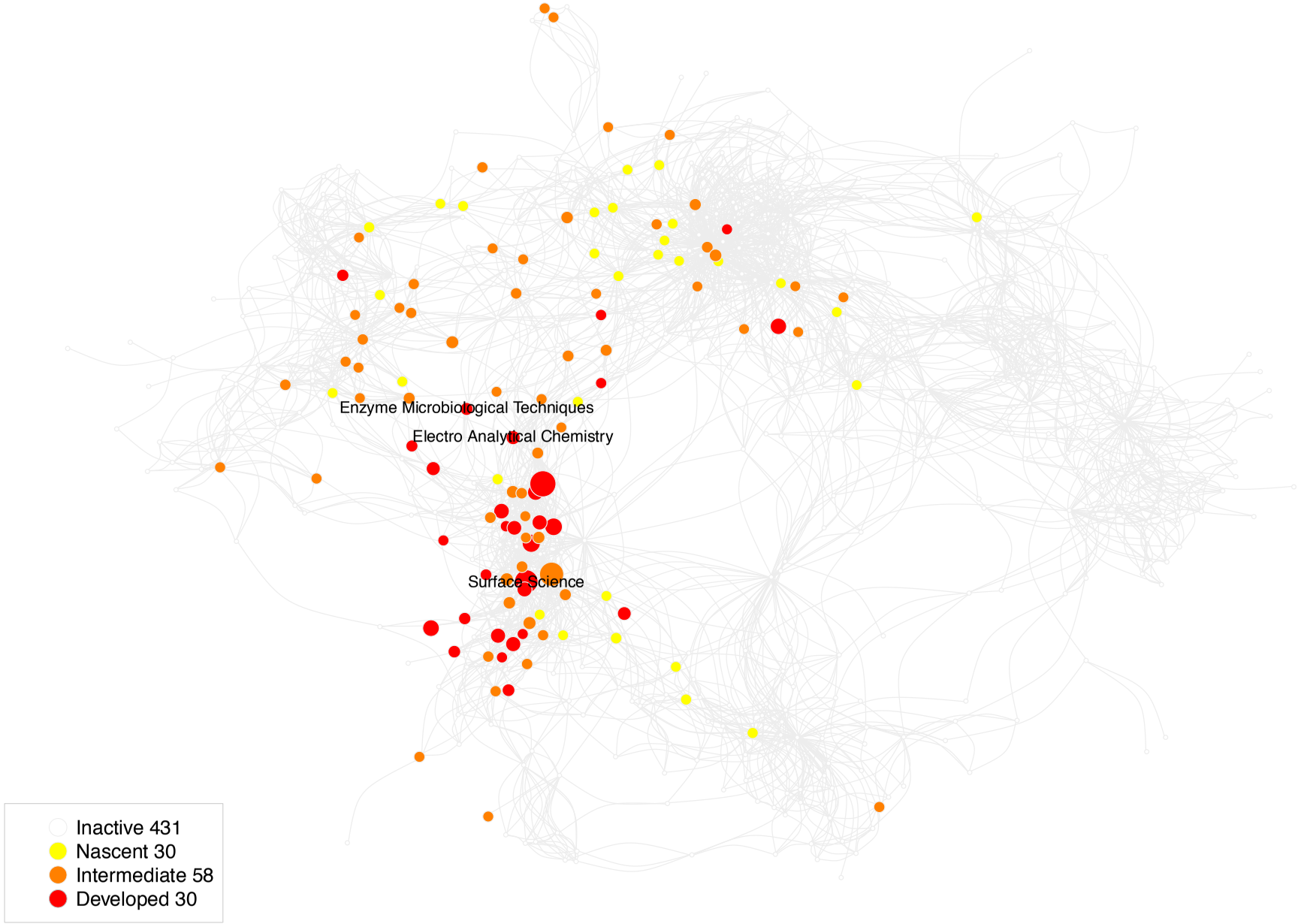

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2904 < Growing Areas < 1.7837 Developed Areas

**Aalto University / Aalto–yliopisto 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



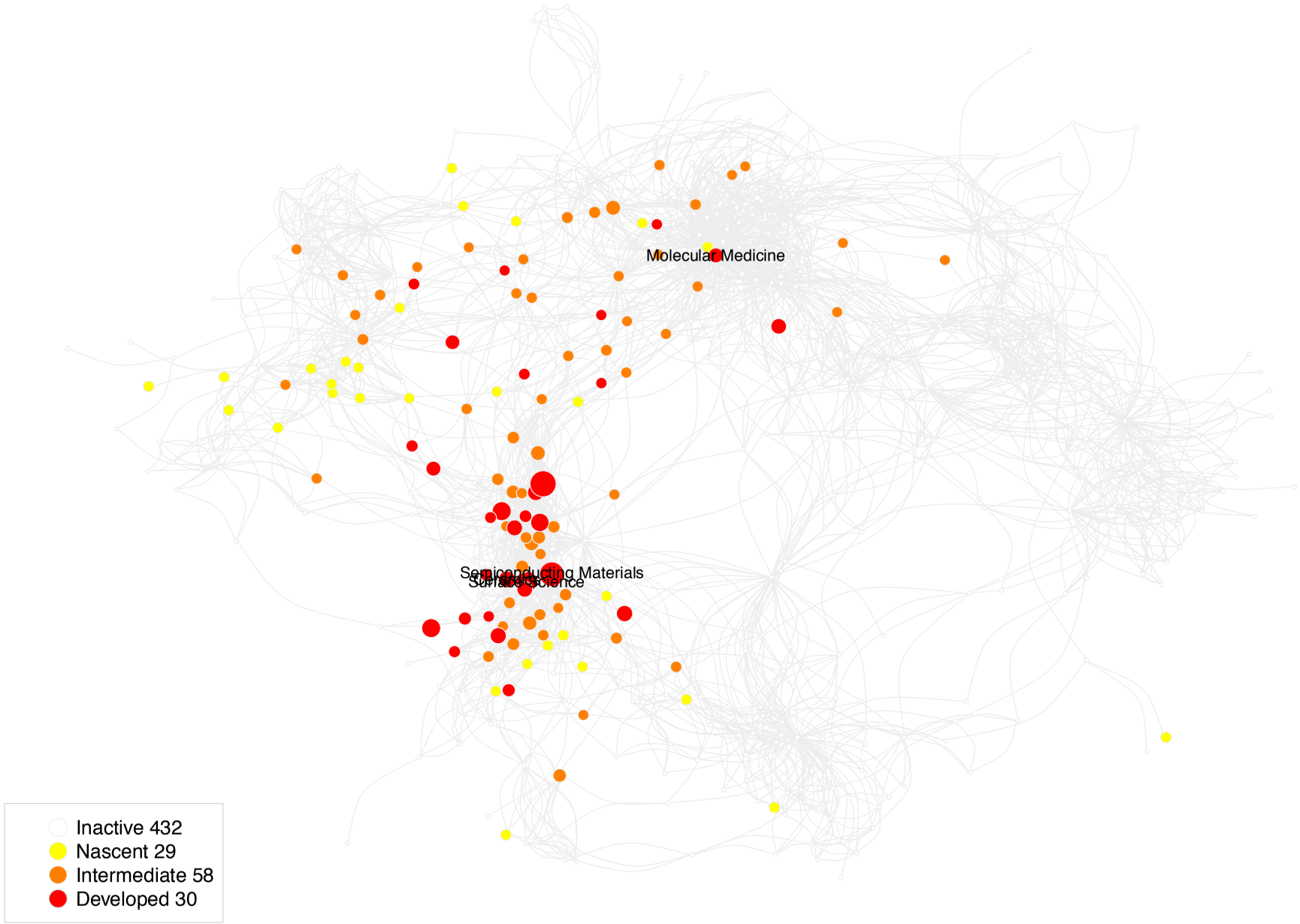

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3677 < Growing Areas < 1.7453 Developed Areas

**Bhabha Atomic Research Centre 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



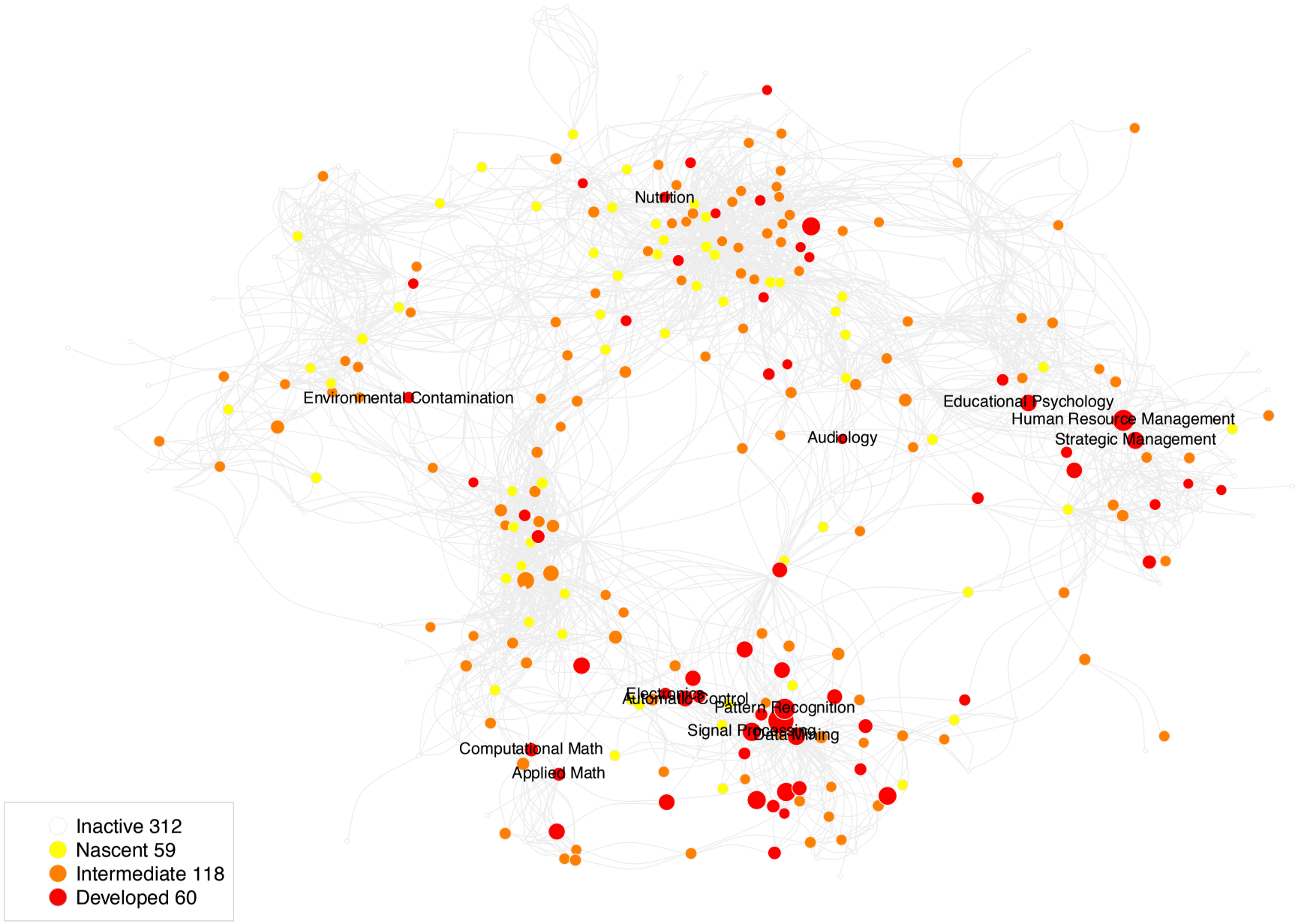

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.5808 < Growing Areas < 4.6576 Developed Areas

**Bhabha Atomic Research Centre 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



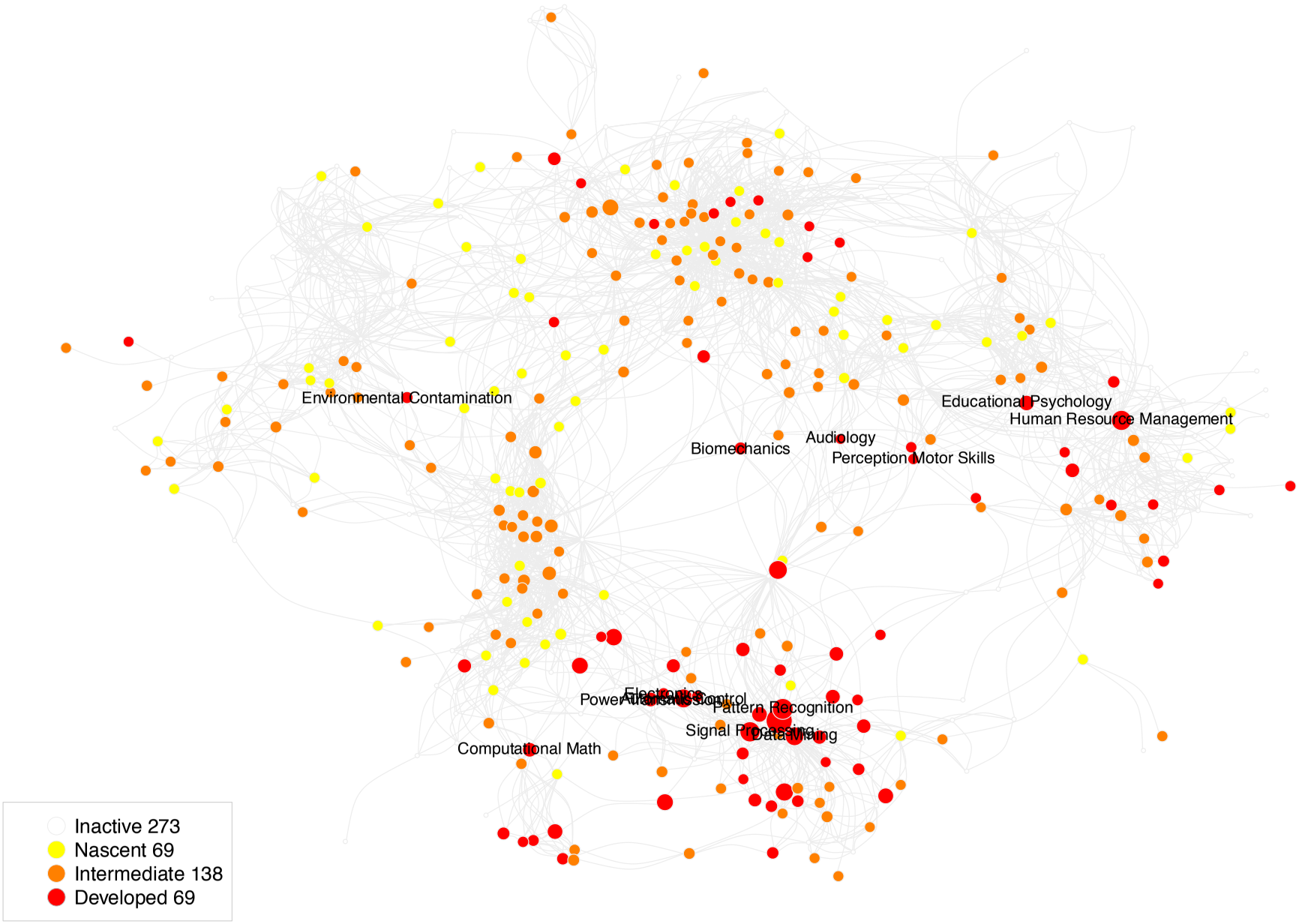

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.513 < Growing Areas < 3.9043 Developed Areas

**Chinese University of Hong Kong 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



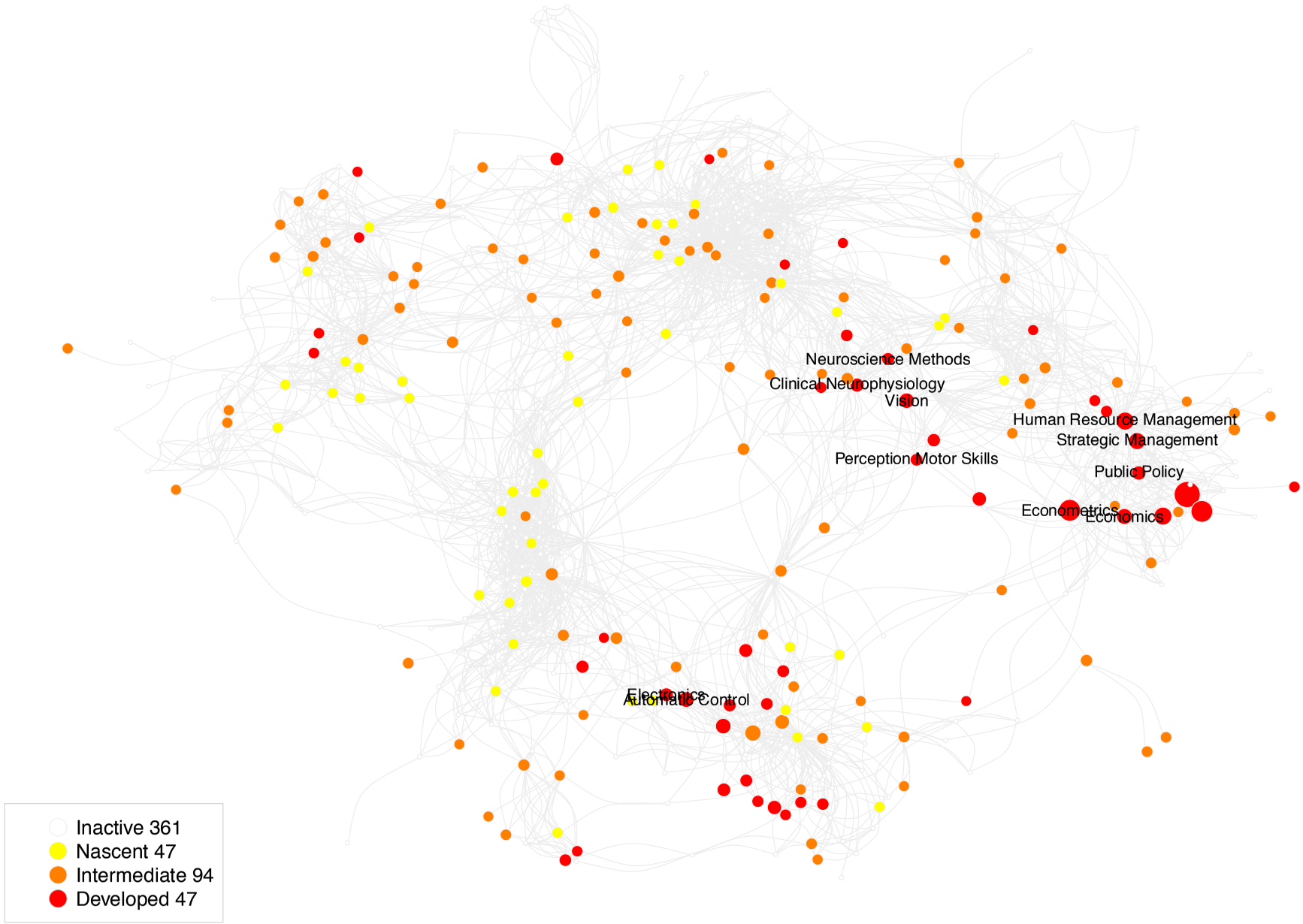

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.4879 < Growing Areas < 1.8864 Developed Areas

**Chinese University of Hong Kong 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



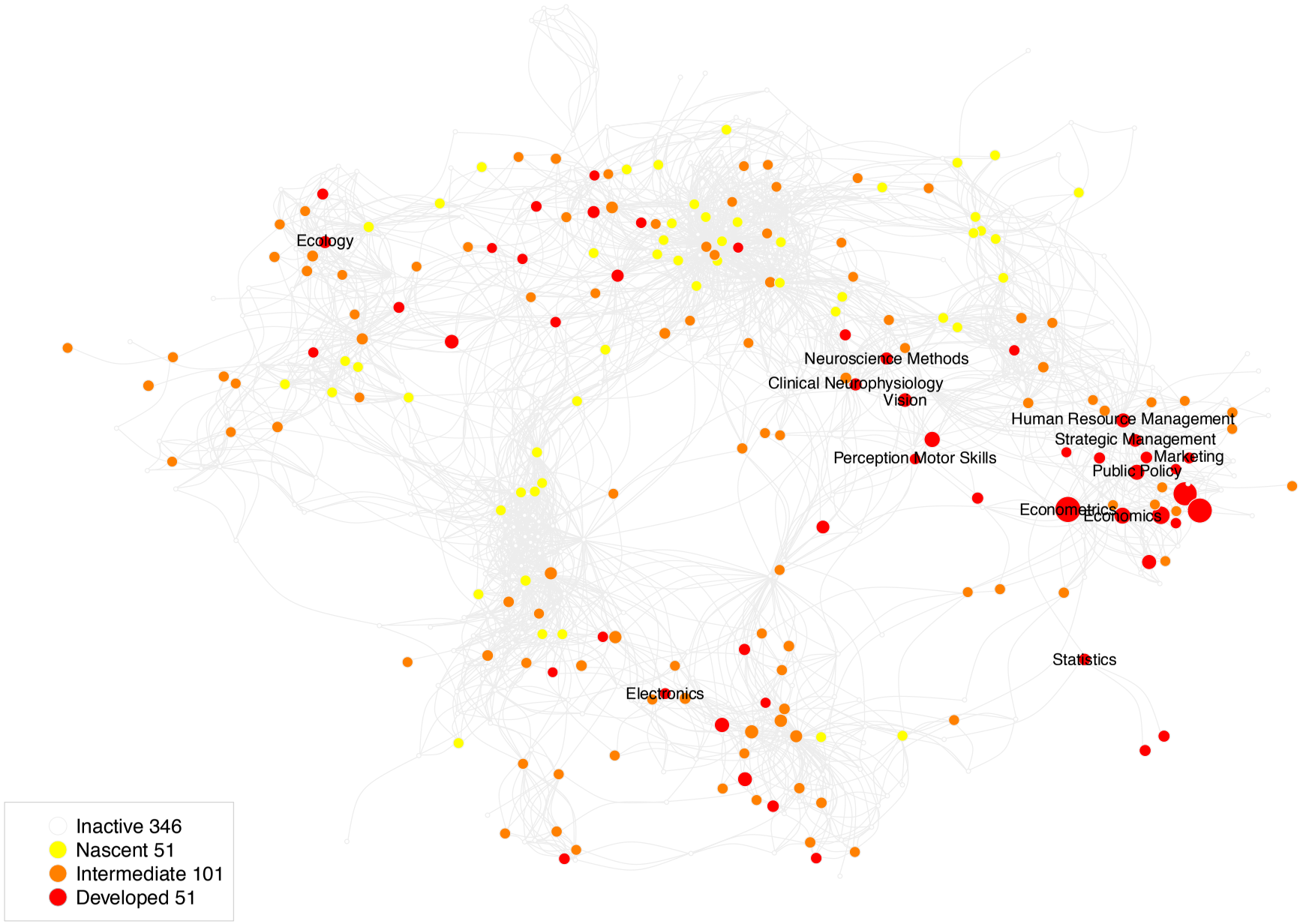

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3699 < Growing Areas < 1.6766 Developed Areas

**University of Essex 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



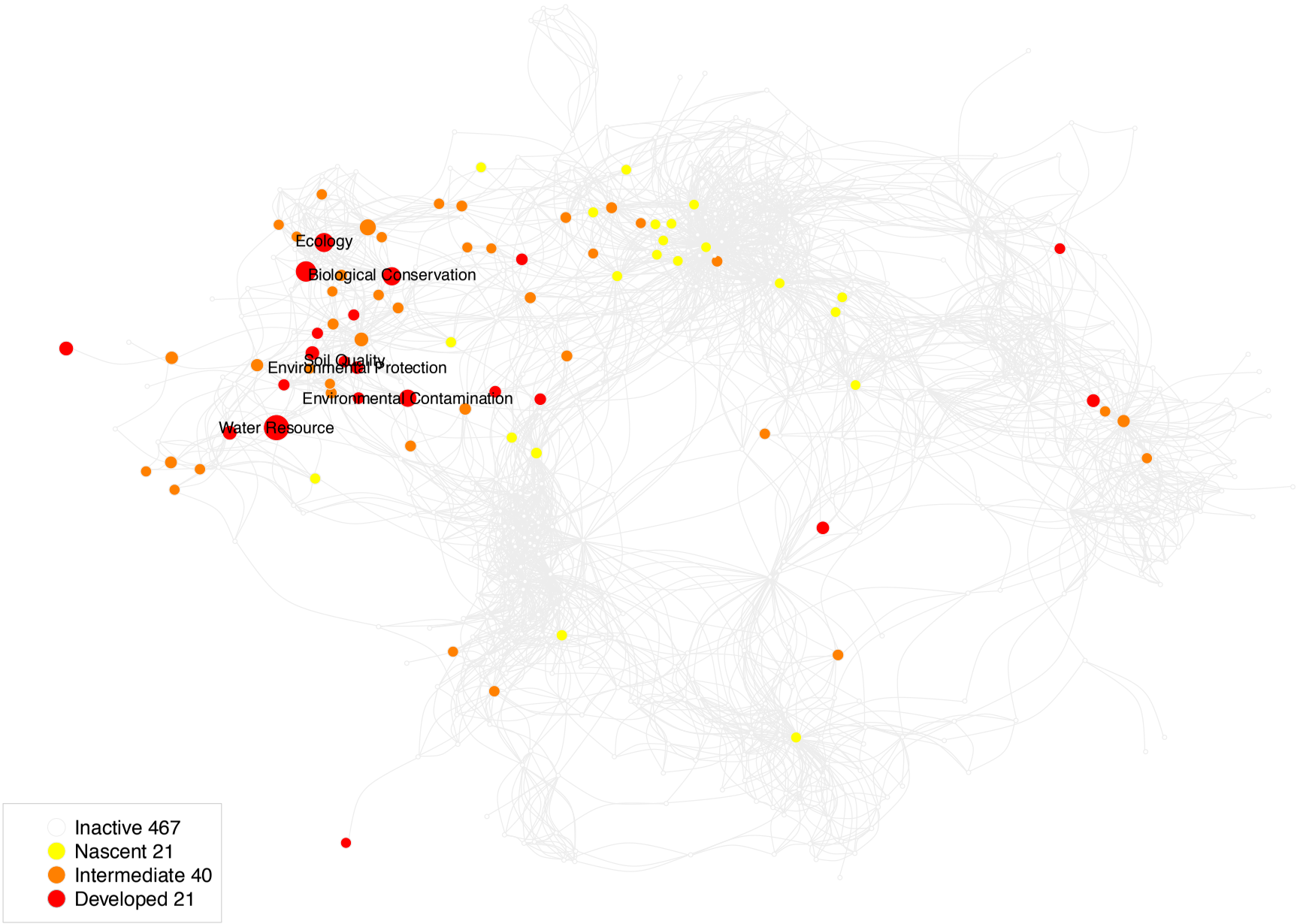

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.369 < Growing Areas < 2.2755 Developed Areas

**University of Essex 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



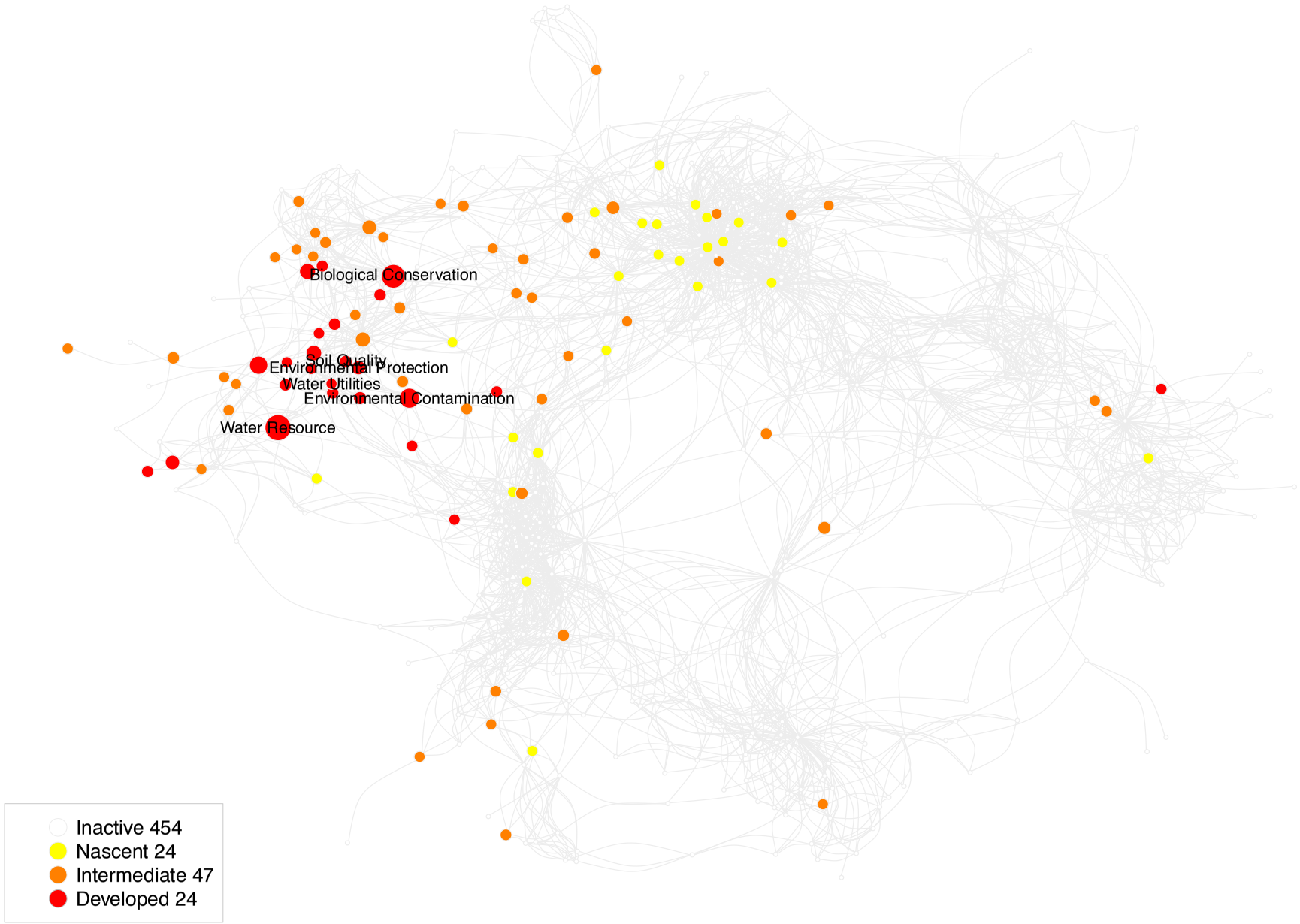

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3408 < Growing Areas < 1.7793 Developed Areas

**Helmholtz Zentrum für Umweltforschung 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



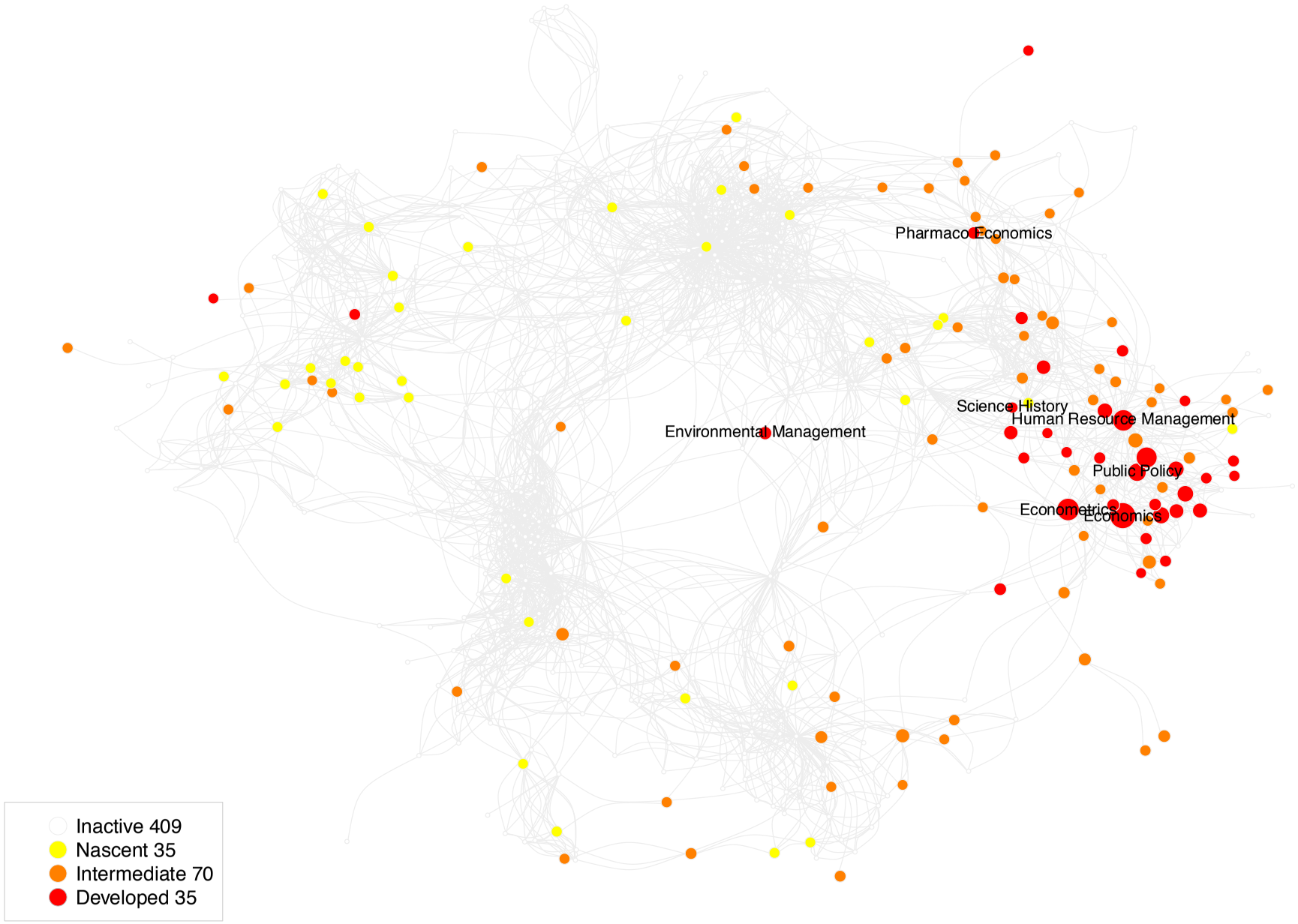

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 1.0614 < Growing Areas < 8.1195 Developed Areas

**Helmholtz Zentrum für Umweltforschung 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



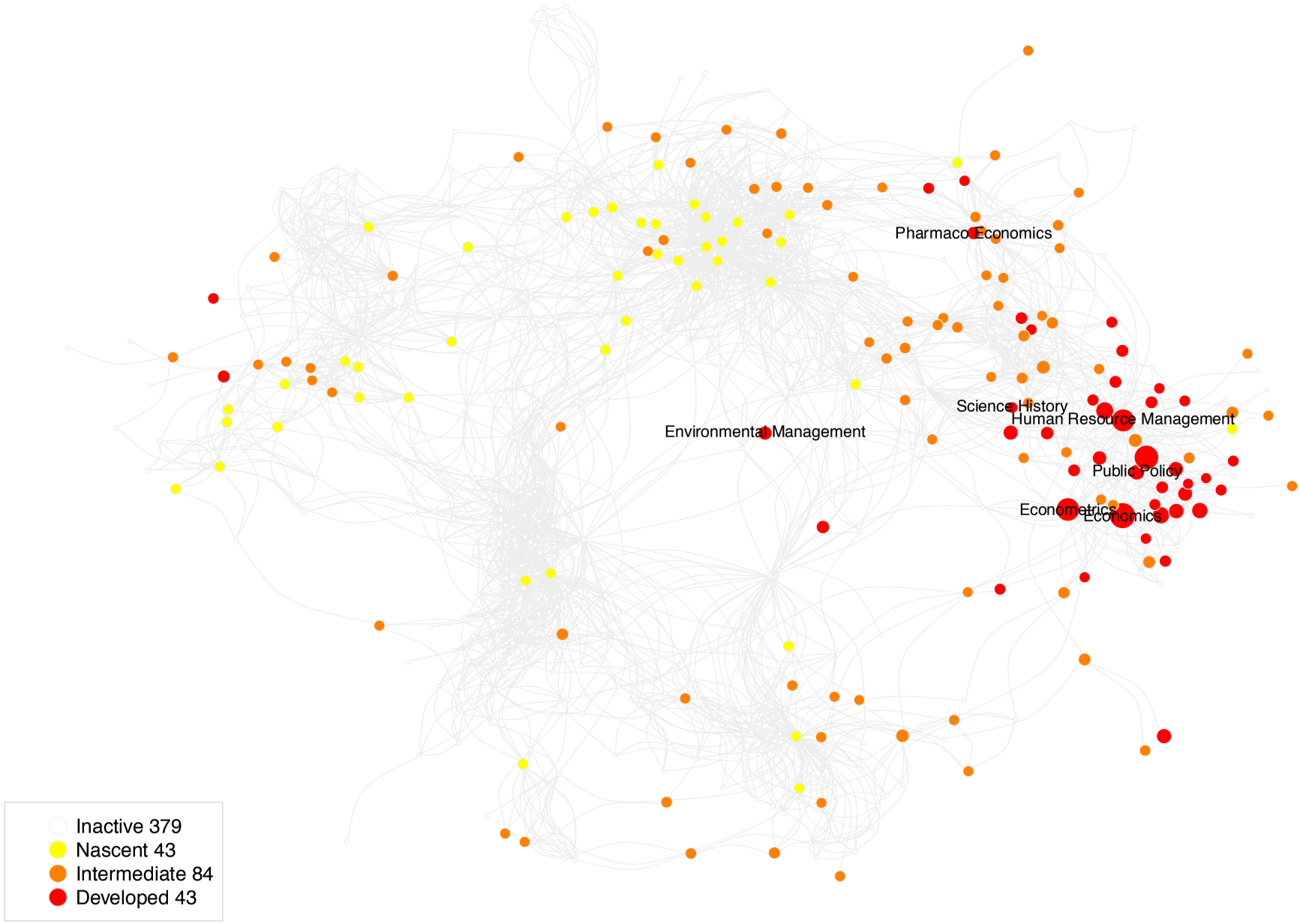

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.8086 < Growing Areas < 6.0615 Developed Areas

**London School of Economics and Political Science 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



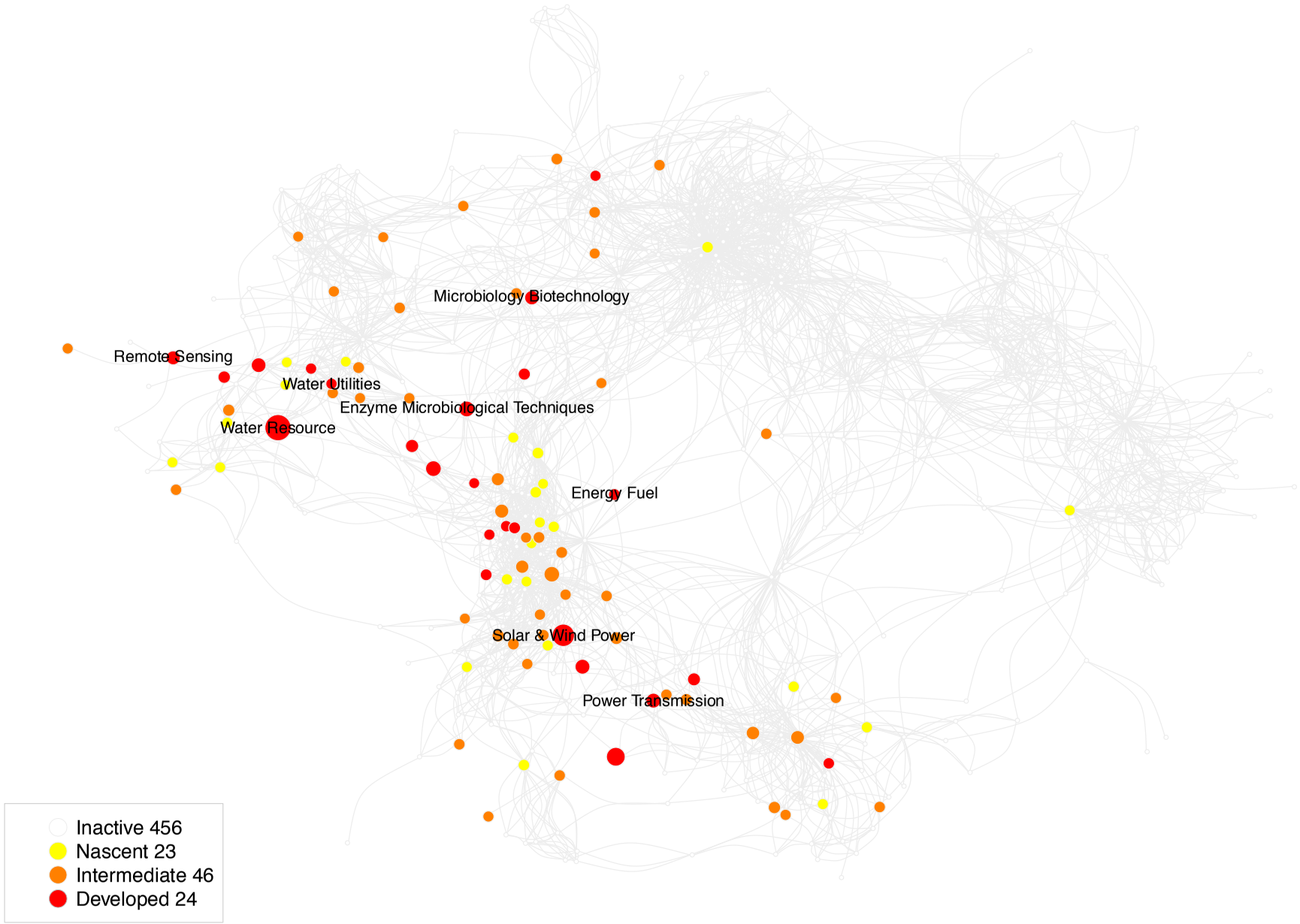

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.5498 < Growing Areas < 4.621 Developed Areas

**London School of Economics and Political Science 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



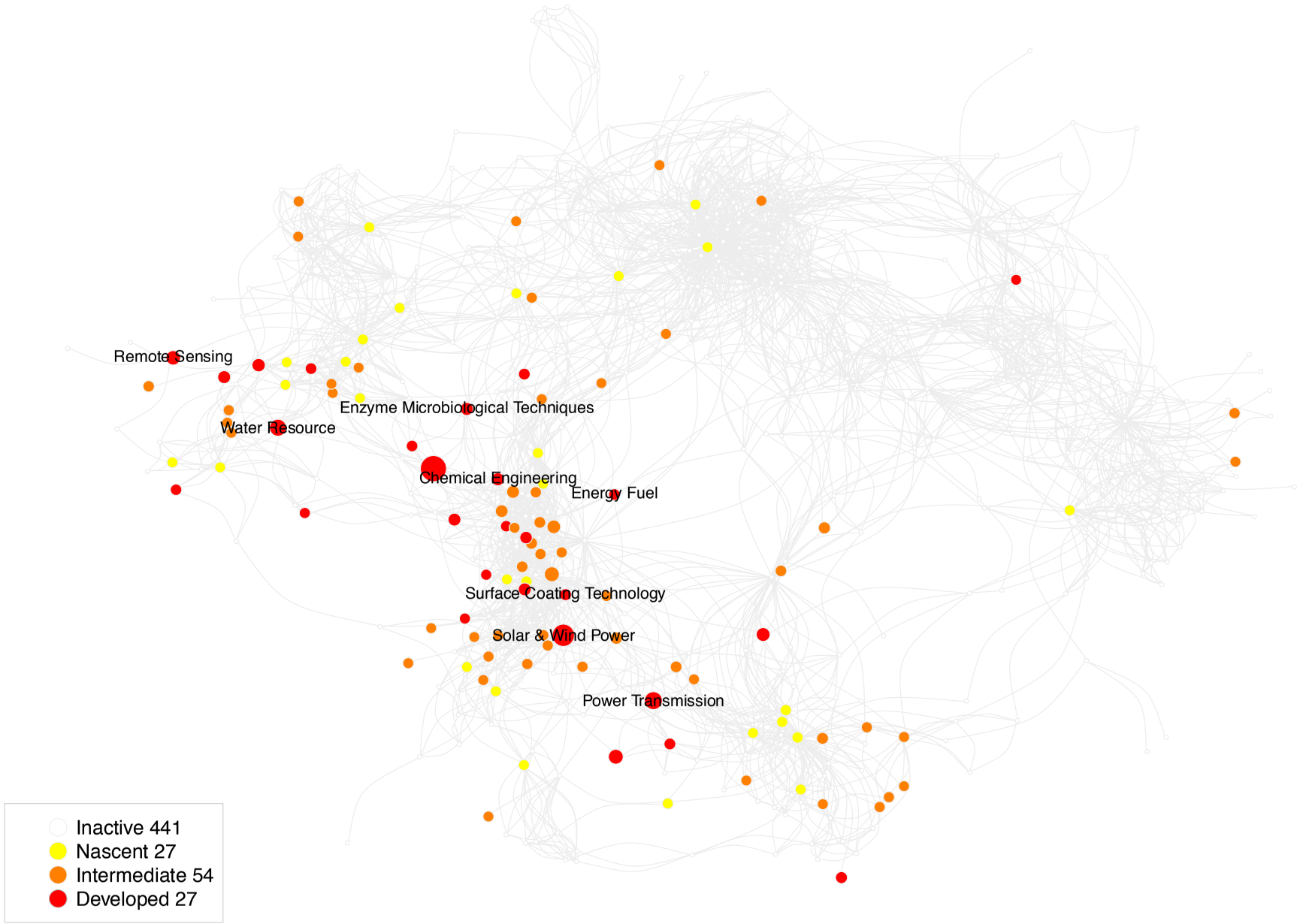

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.1794 < Growing Areas < 3.9473 Developed Areas

**Masdar Institute of Science and Technology 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



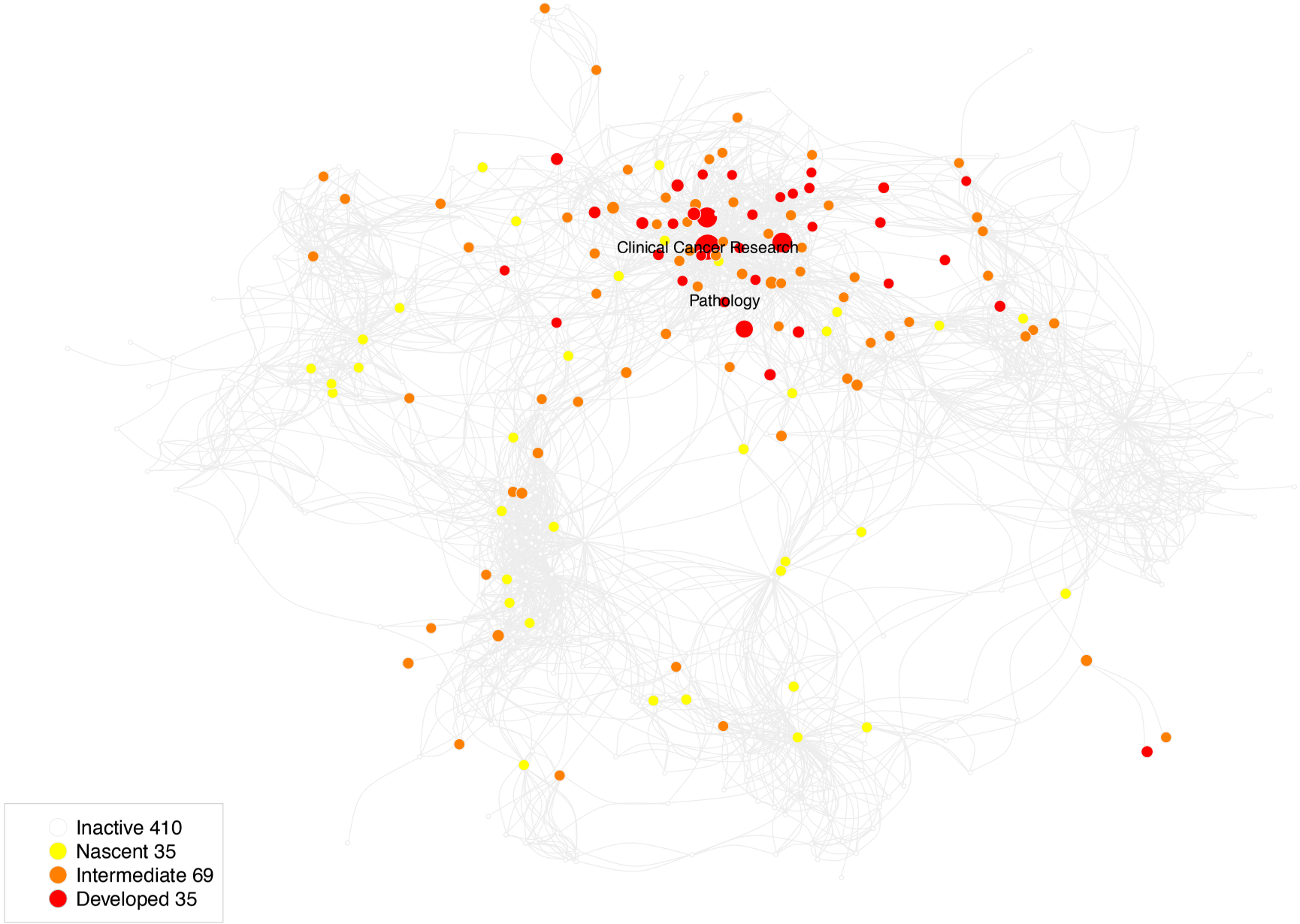

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 1.316 < Growing Areas < 5.5561 Developed Areas

**Masdar Institute of Science and Technology 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



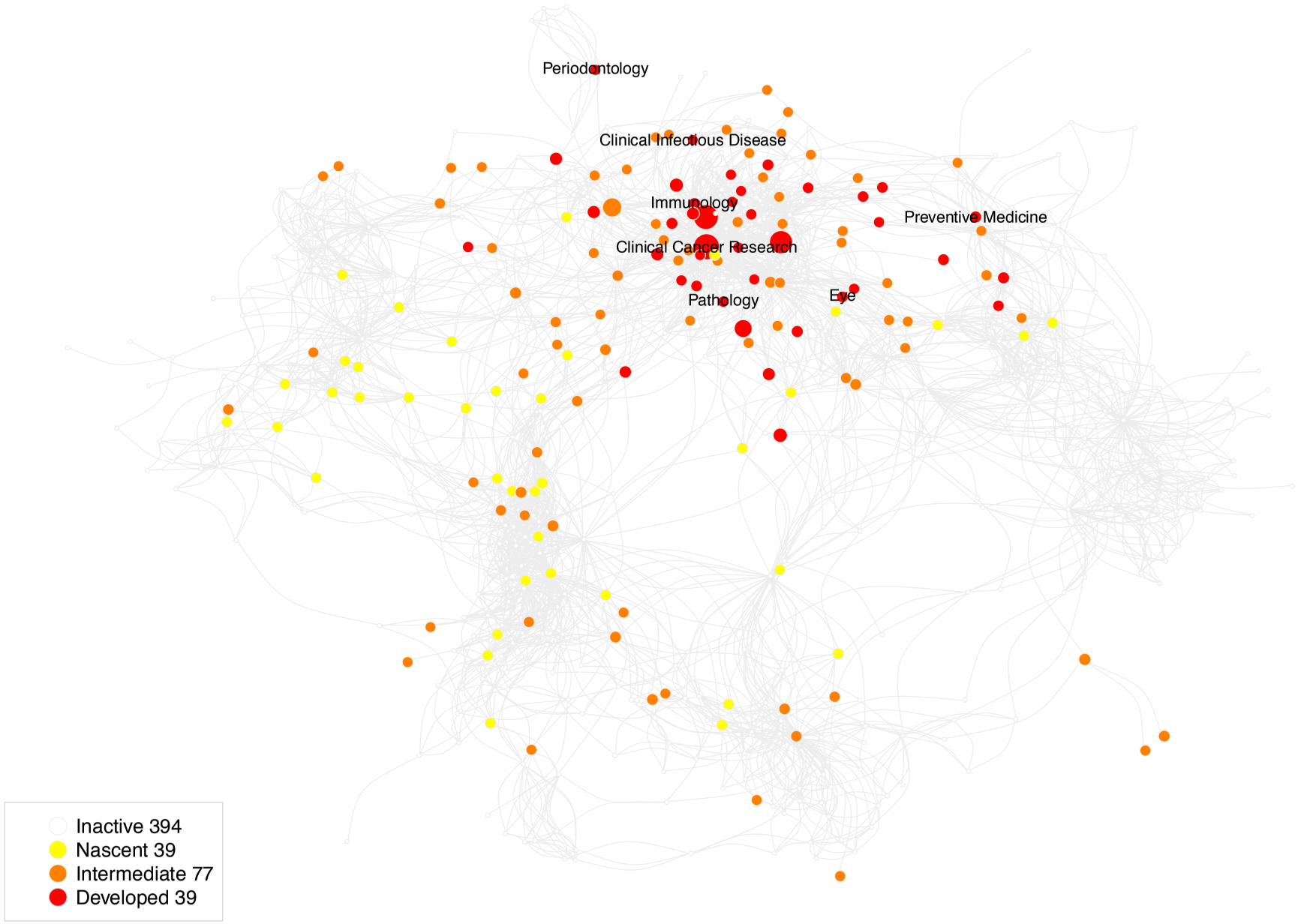

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.6525 < Growing Areas < 3.8929 Developed Areas

**M D Anderson Cancer Center University of Texas 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



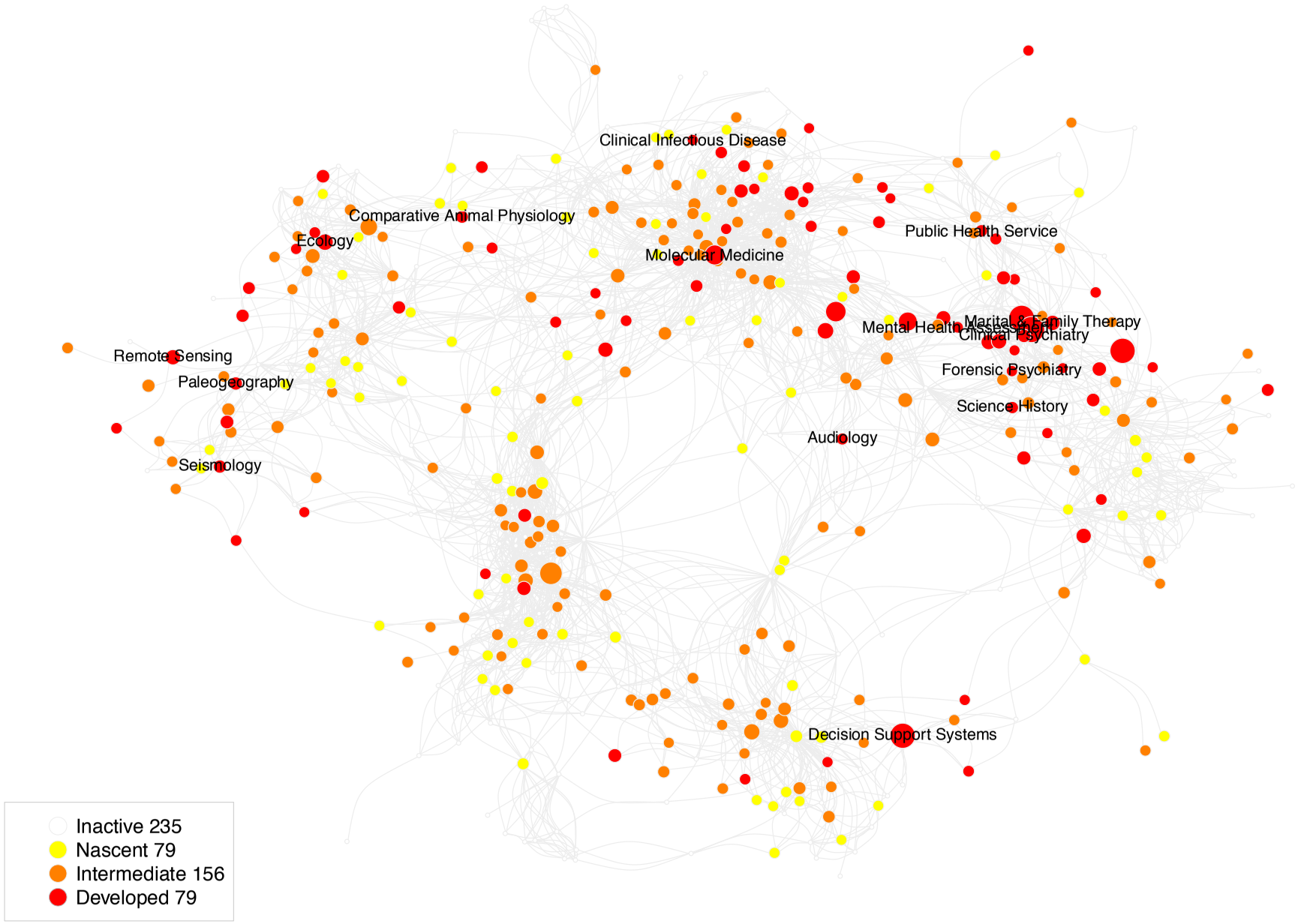

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.8435 < Growing Areas < 4.5026 Developed Areas

**M D Anderson Cancer Center University of Texas 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



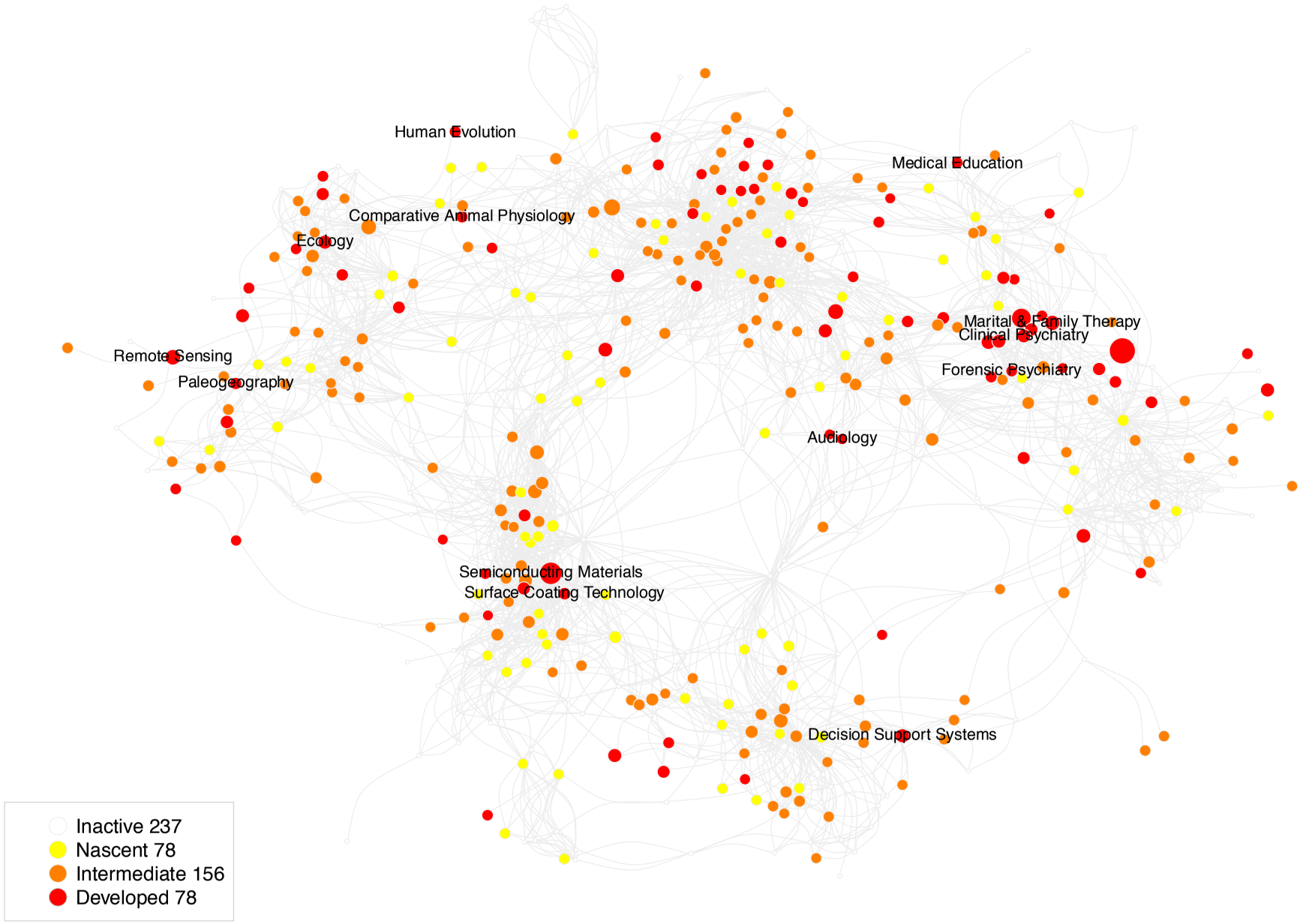

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.5491 < Growing Areas < 3.5958 Developed Areas

**University of South Florida 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**




Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.527 < Growing Areas < 1.8992 Developed Areas

**University of South Florida 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**

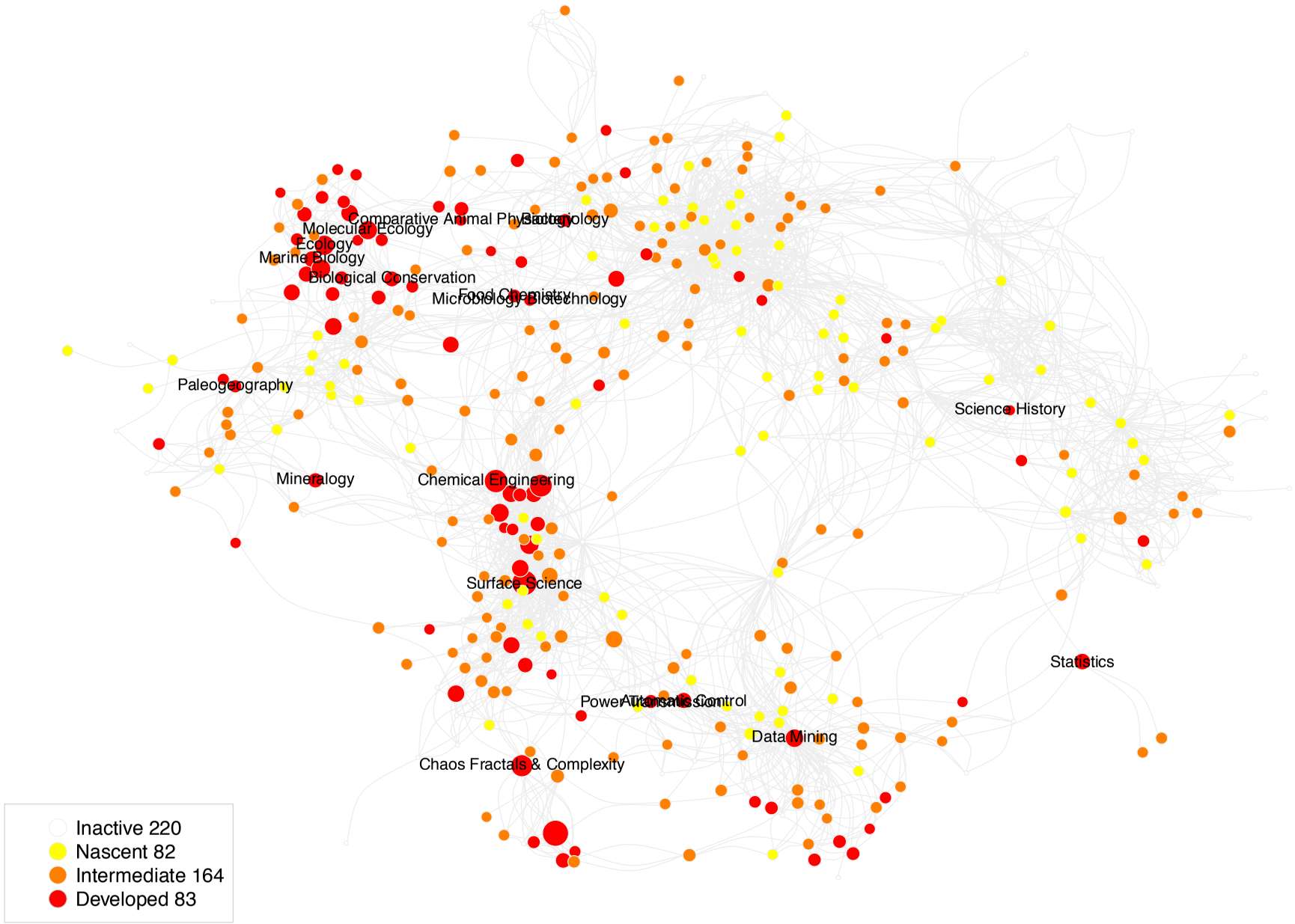



Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.4977 < Growing Areas < 2.0023 Developed Areas

Apéndice E

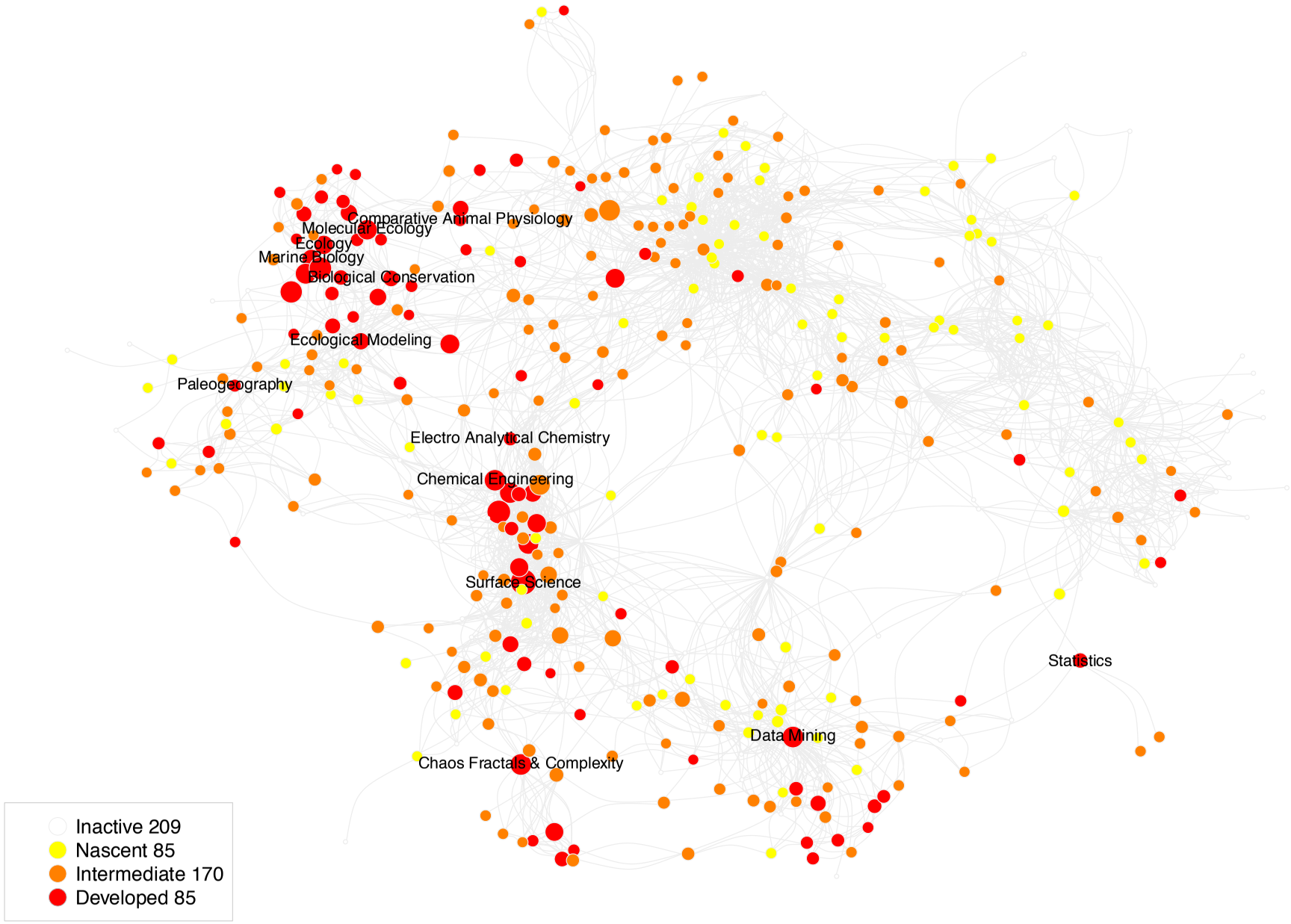
# Mapas superpuestos para países

**Argentina 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



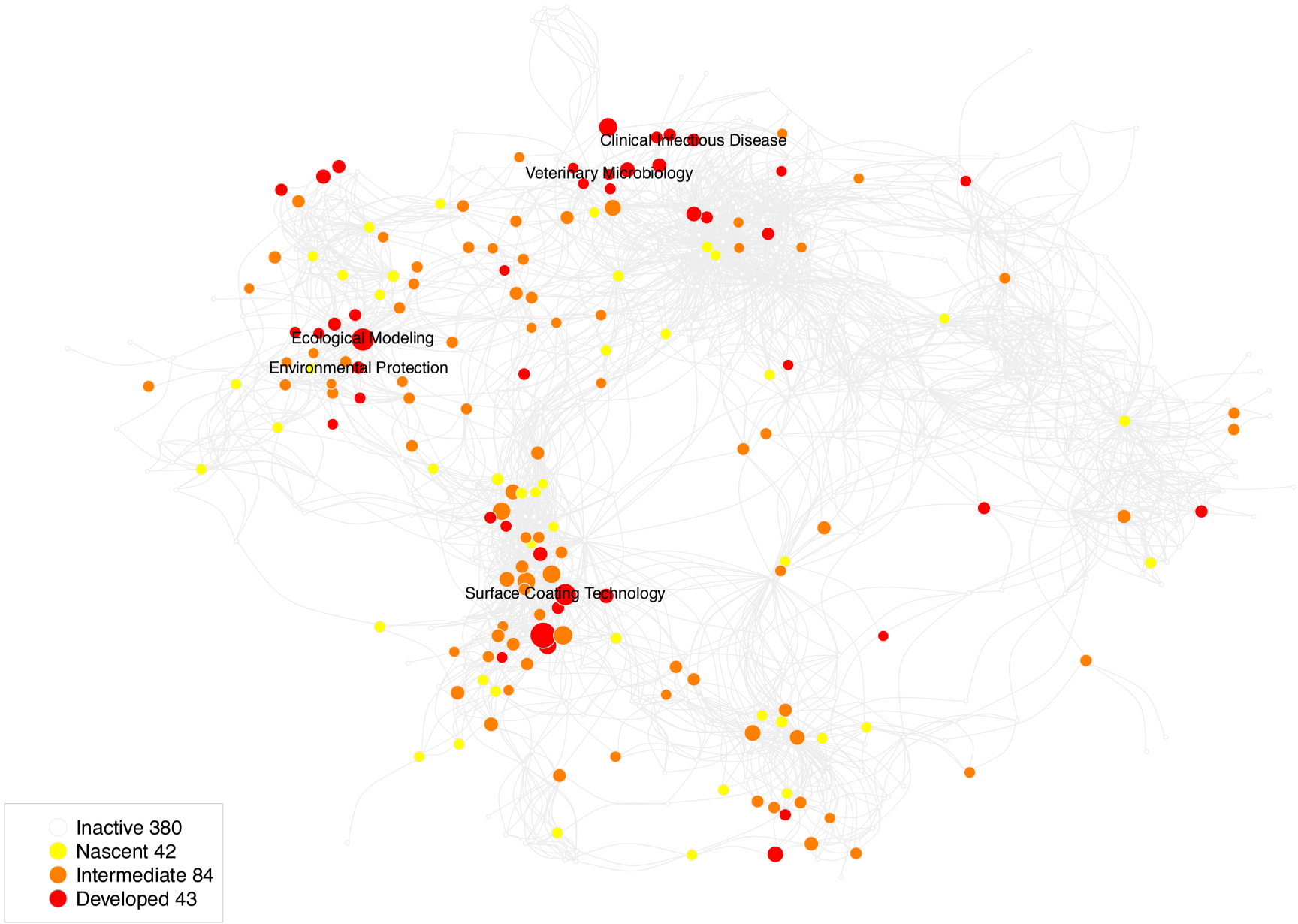

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3331 < Growing Areas < 1.4749 Developed Areas

**Argentina 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



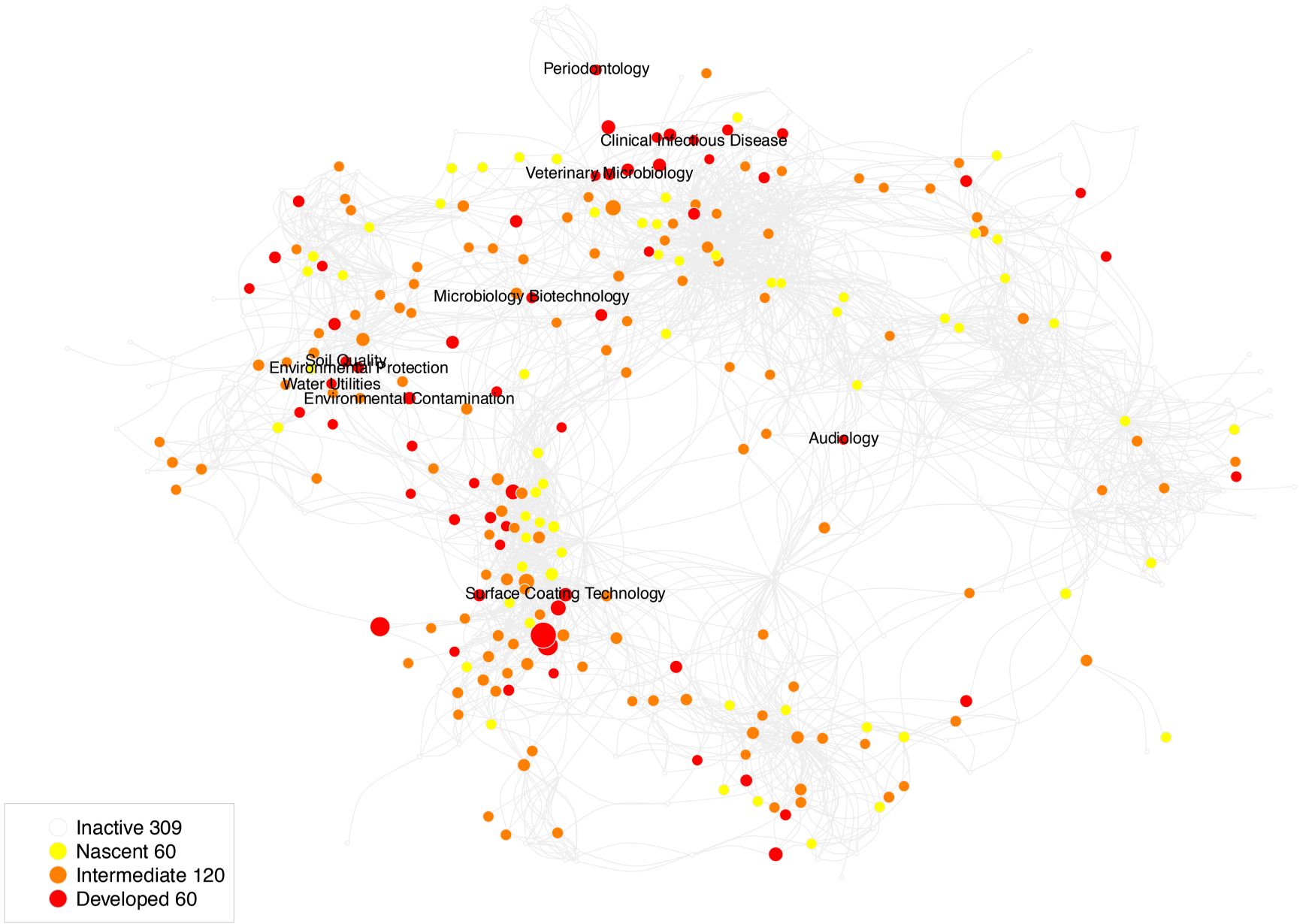

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3894 < Growing Areas < 1.5368 Developed Areas

**Bangladesh 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



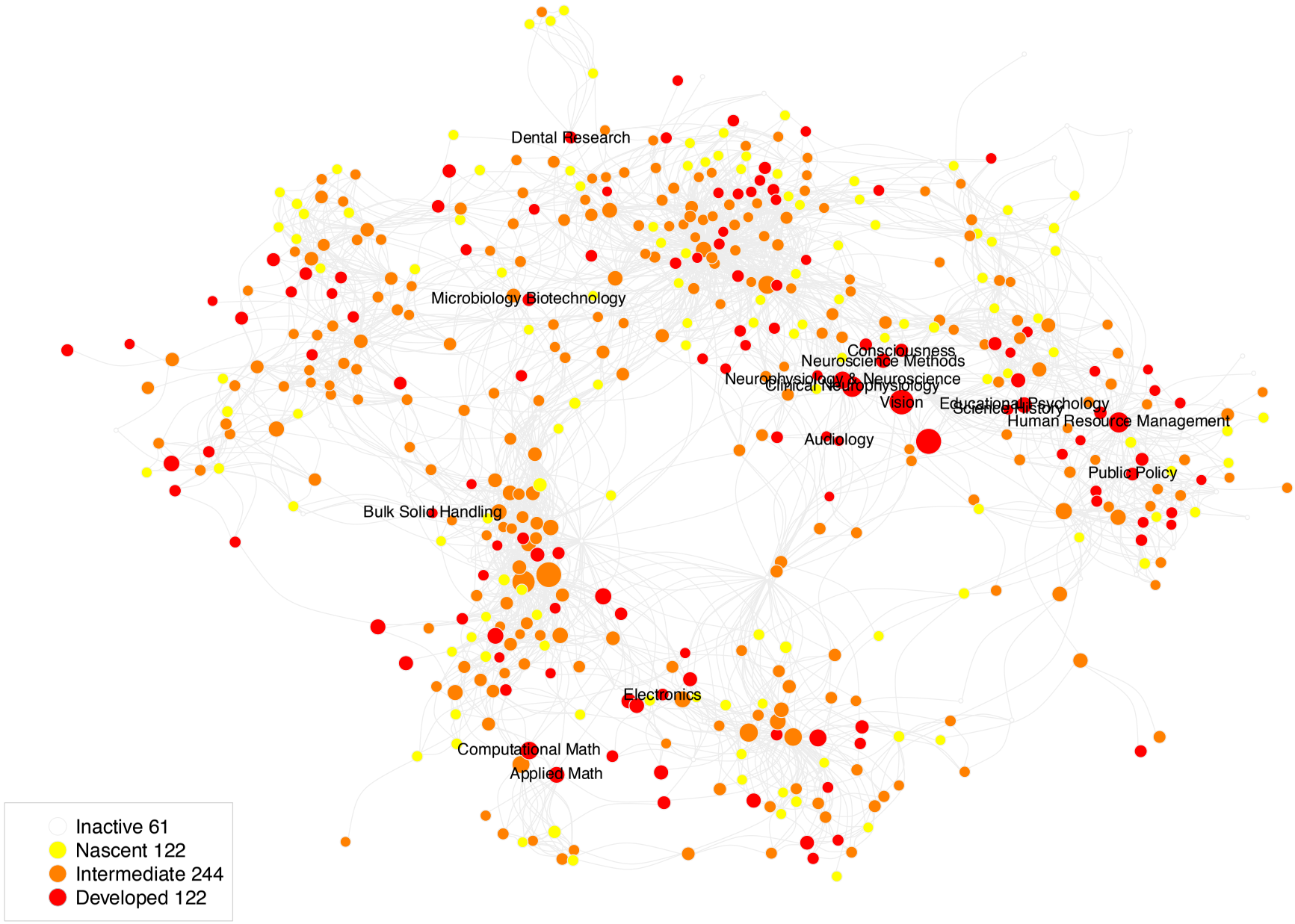

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 1.0244 < Growing Areas < 4.0662 Developed Areas

**Bangladesh 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



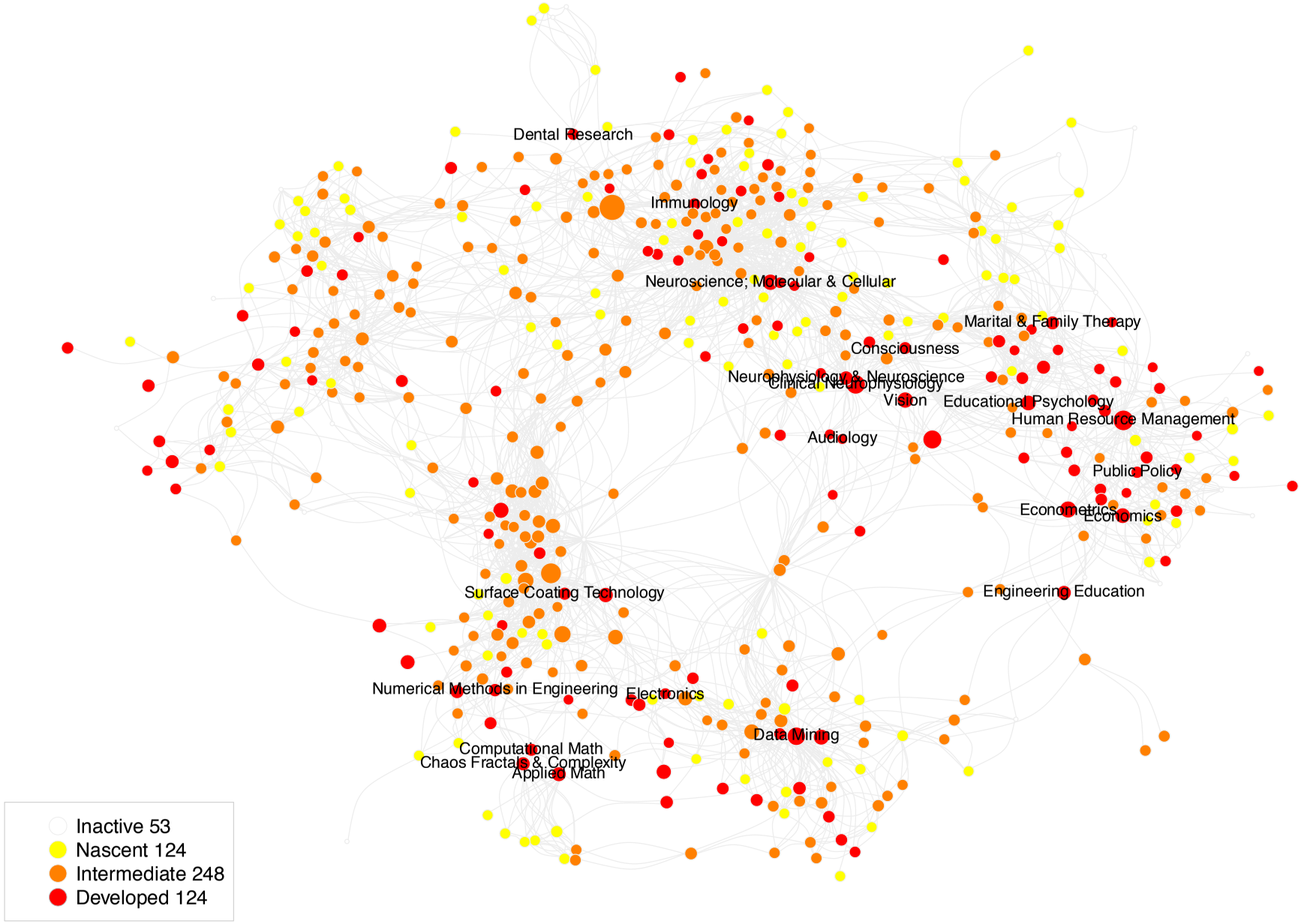

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.6229 < Growing Areas < 2.8187 Developed Areas

**Belgium 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



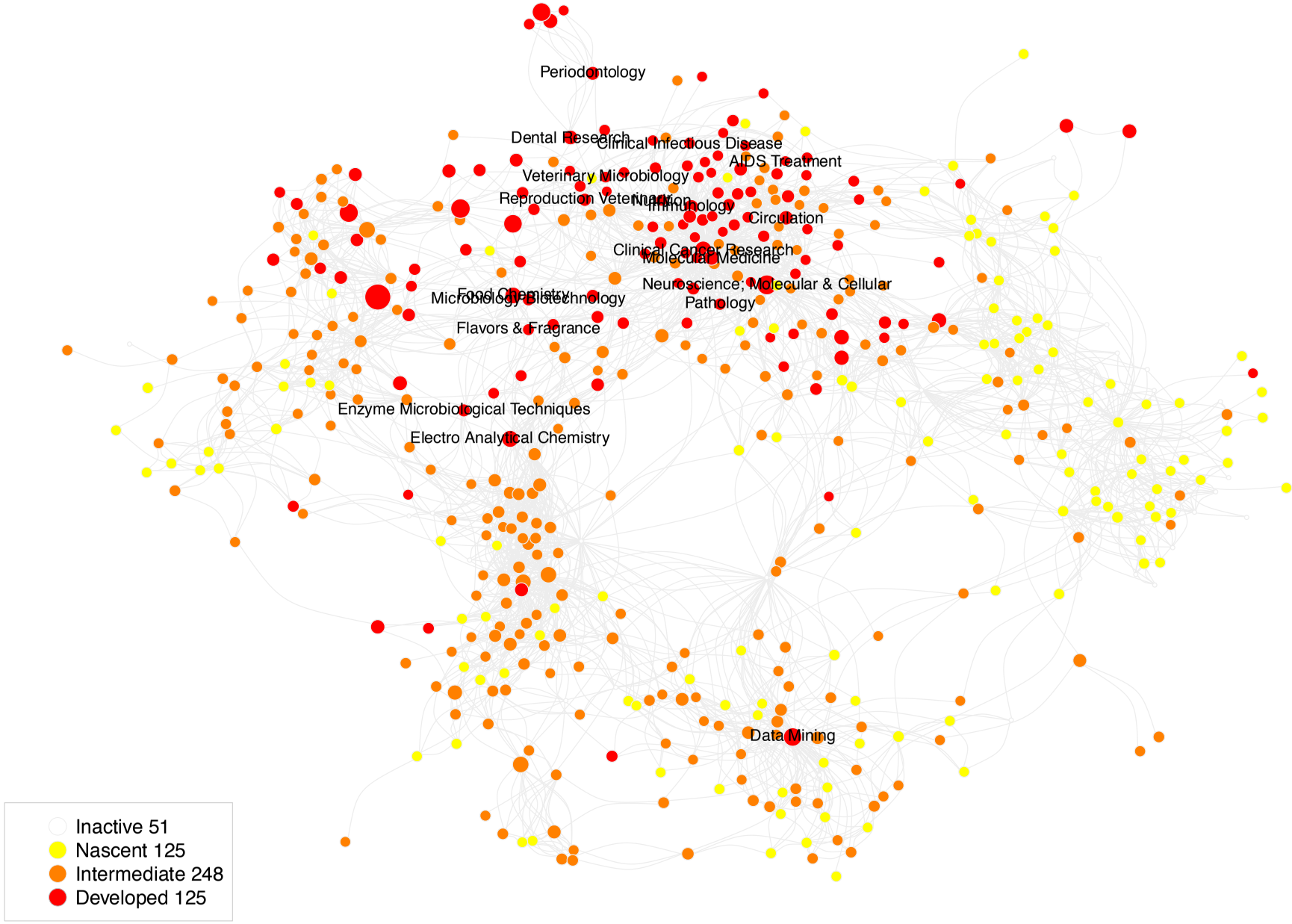

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.4891 < Growing Areas < 1.3616 Developed Areas

**Belgium 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



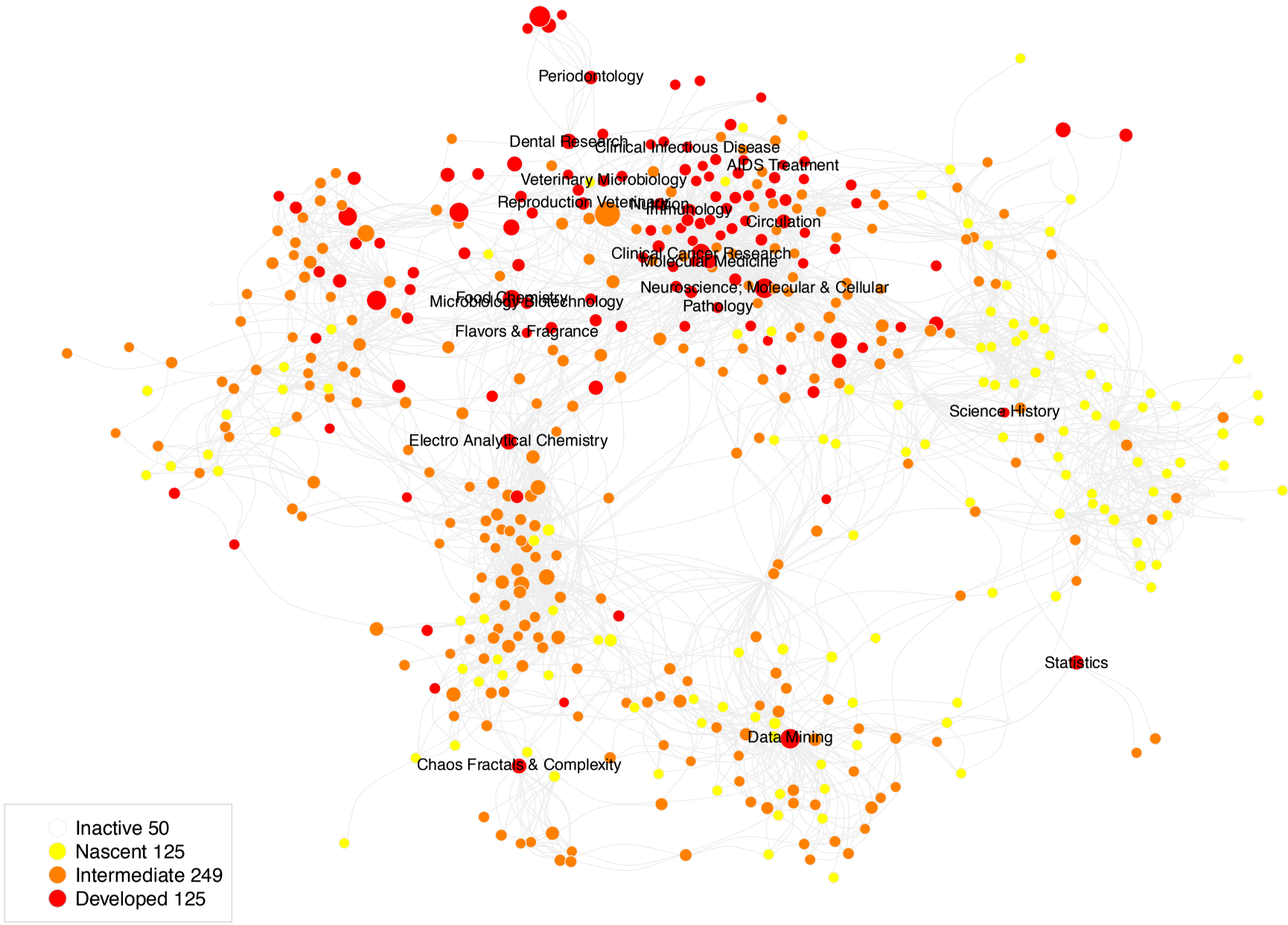

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.5178 < Growing Areas < 1.2656 Developed Areas

**Brazil 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



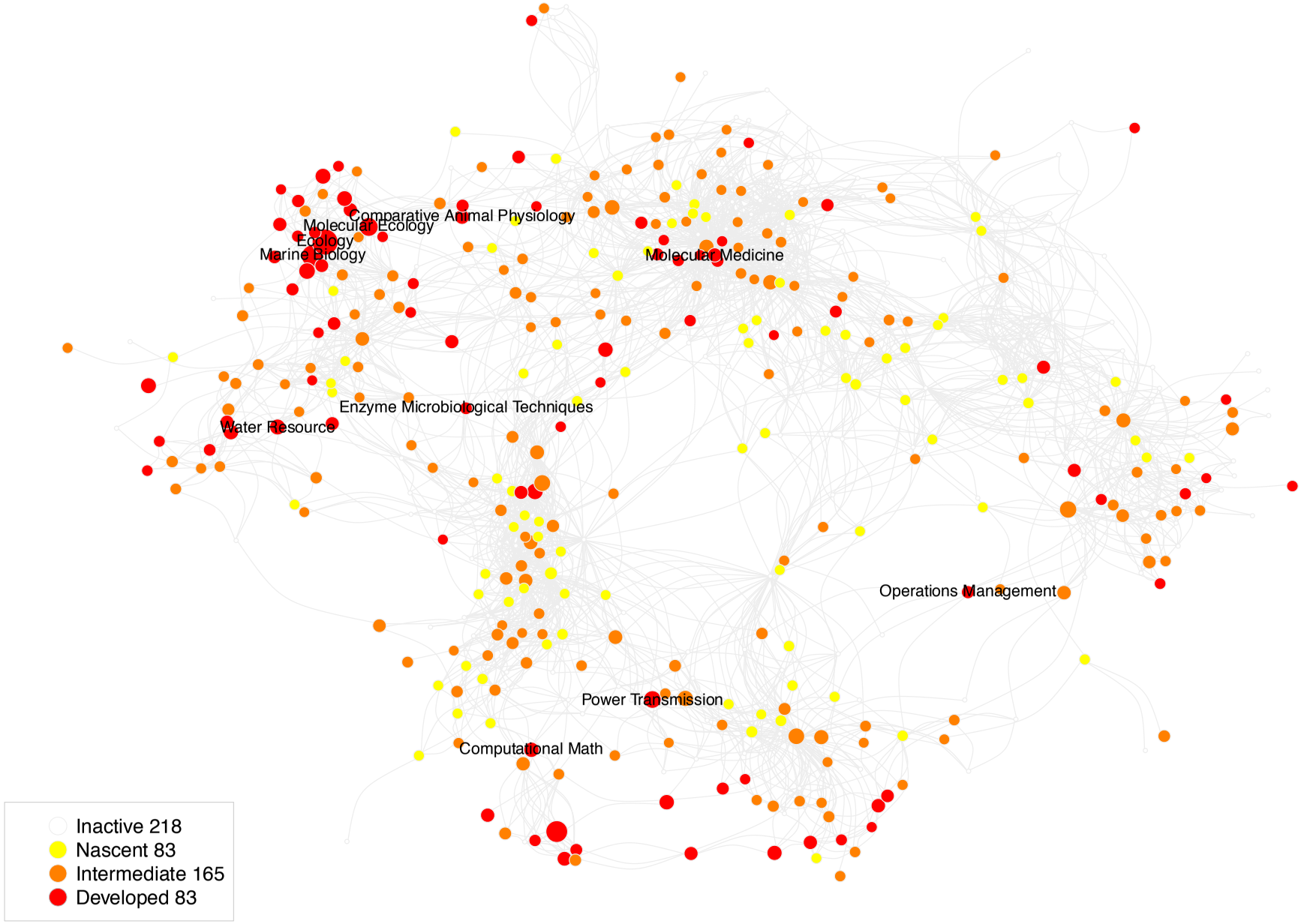

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3384 < Growing Areas < 1.6227 Developed Areas

**Brazil 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



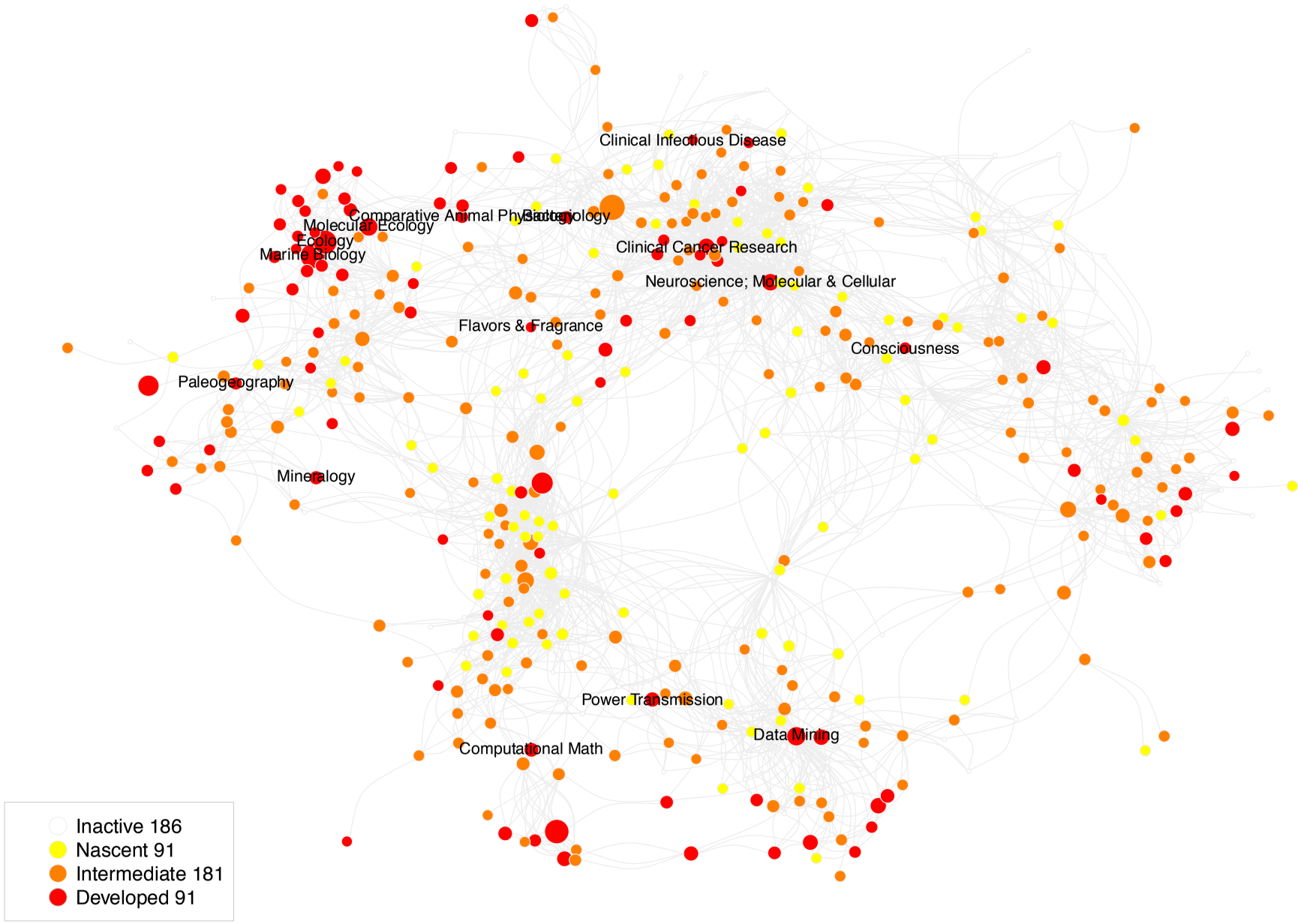

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3757 < Growing Areas < 1.6195 Developed Areas

**Chile 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



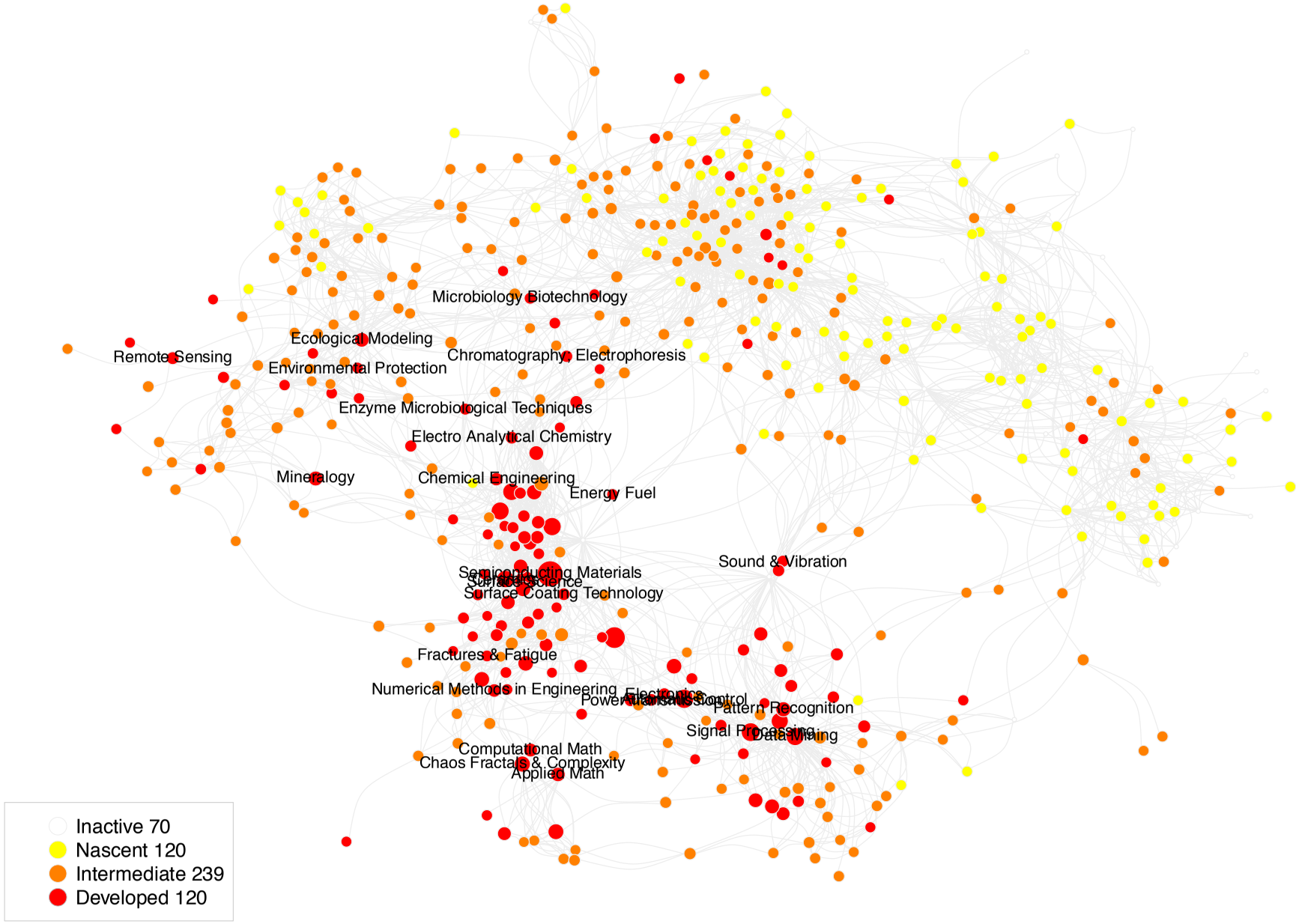

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3907 < Growing Areas < 1.6849 Developed Areas

**Chile 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



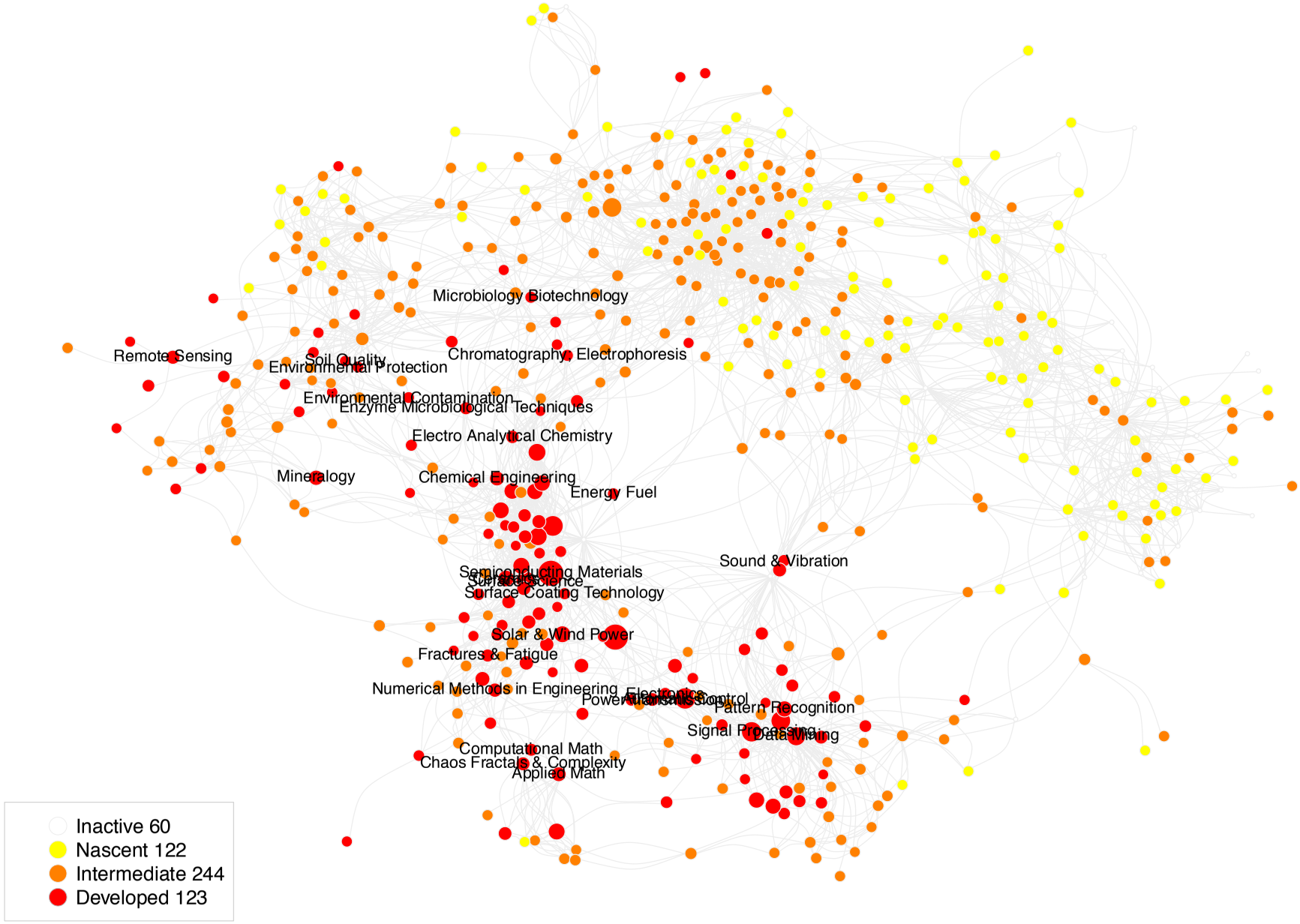

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3939 < Growing Areas < 1.6447 Developed Areas

**China 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



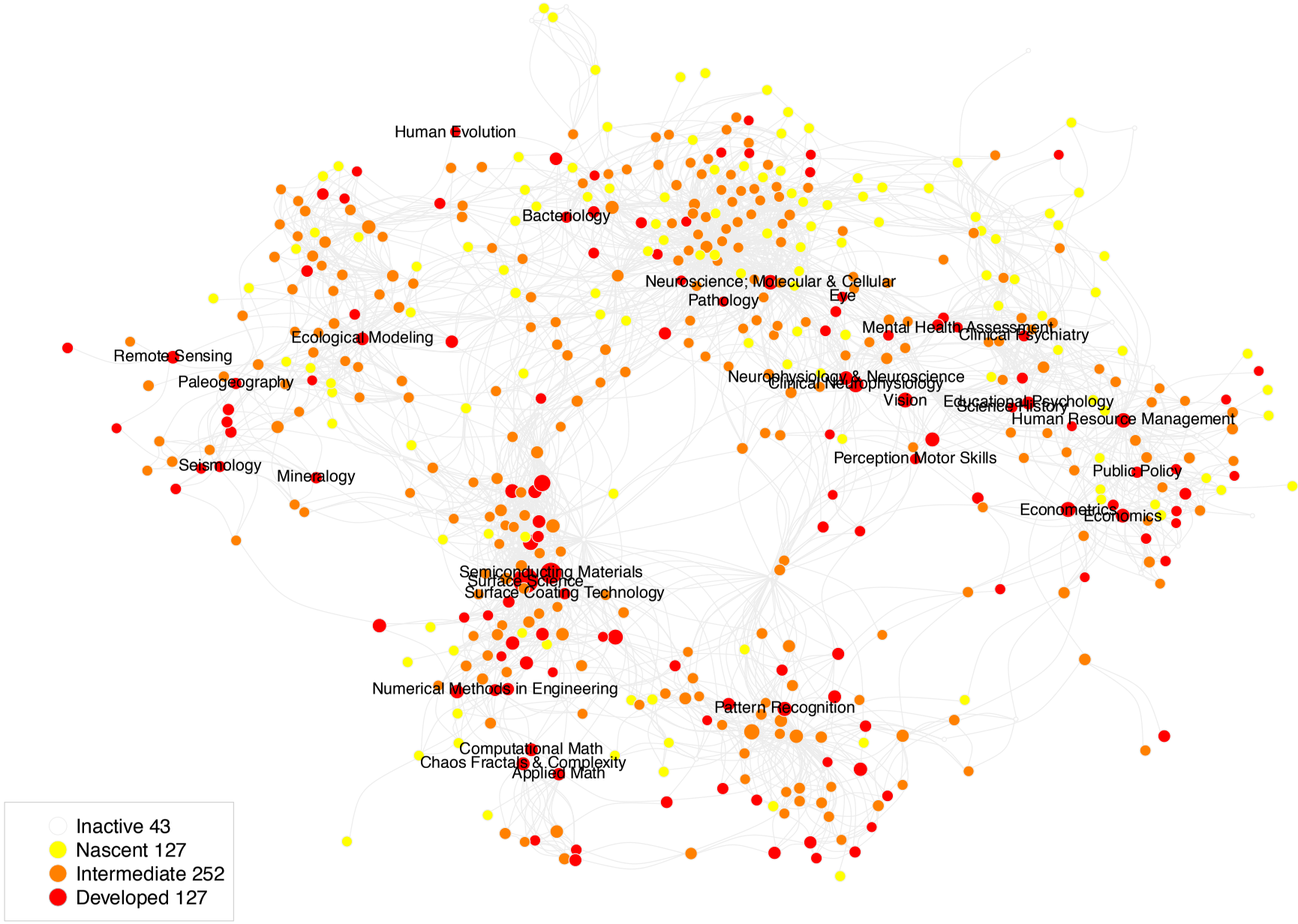

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2031 < Growing Areas < 1.1541 Developed Areas

China 2011_2013
Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA



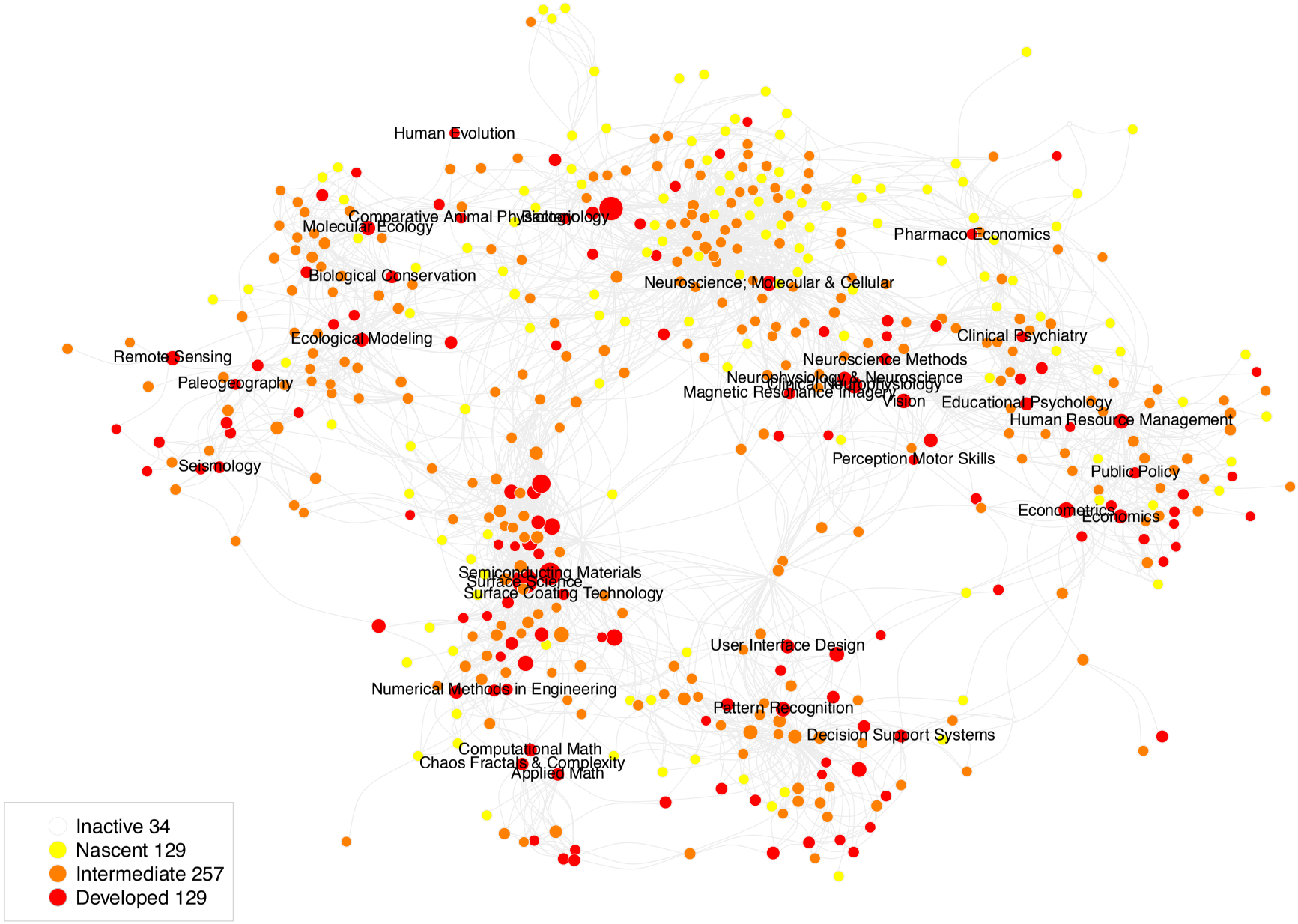

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2003 < Growing Areas < 1.1518 Developed Areas

**Germany 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



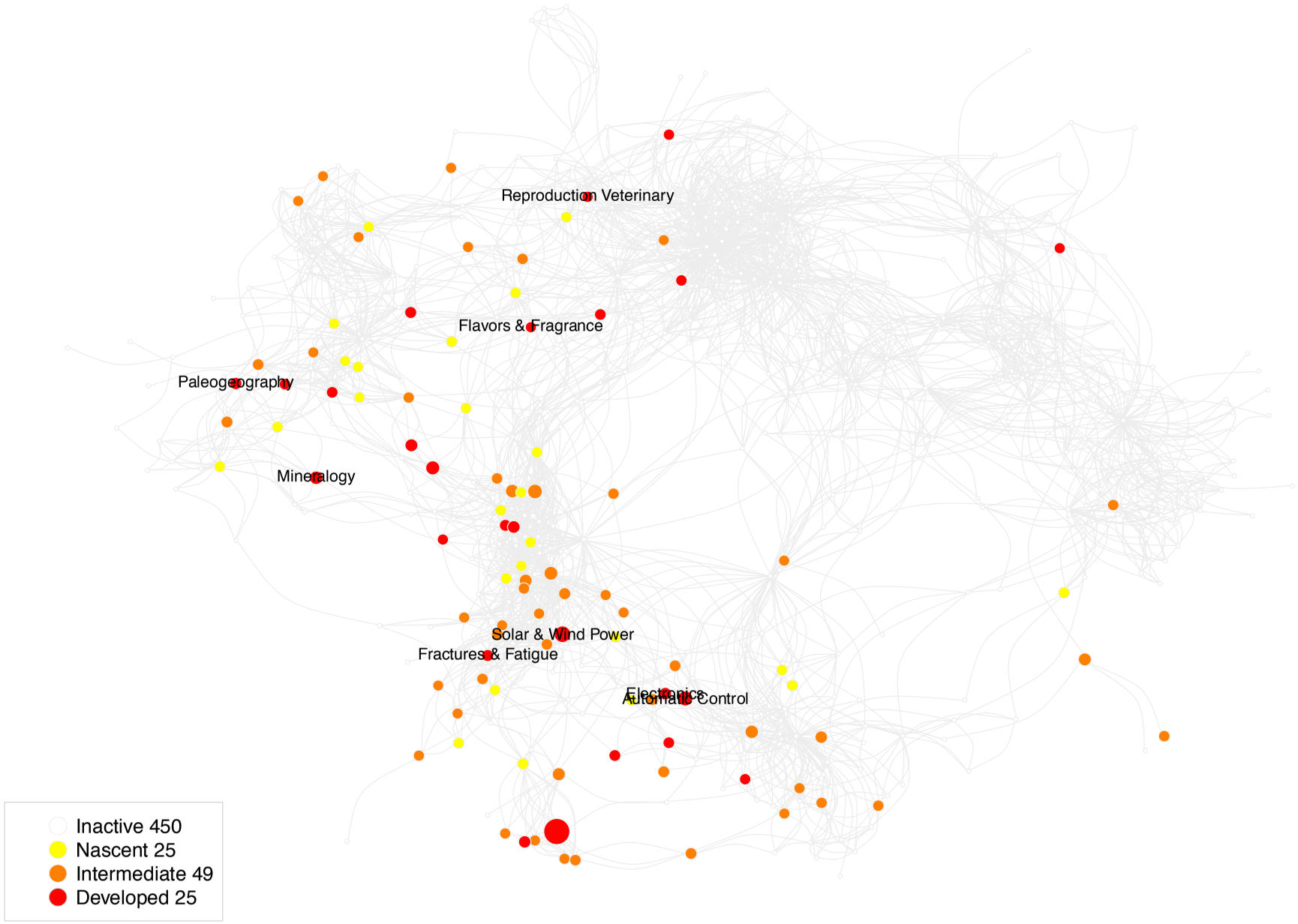

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3978 < Growing Areas < 1.1071 Developed Areas

**Germany 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



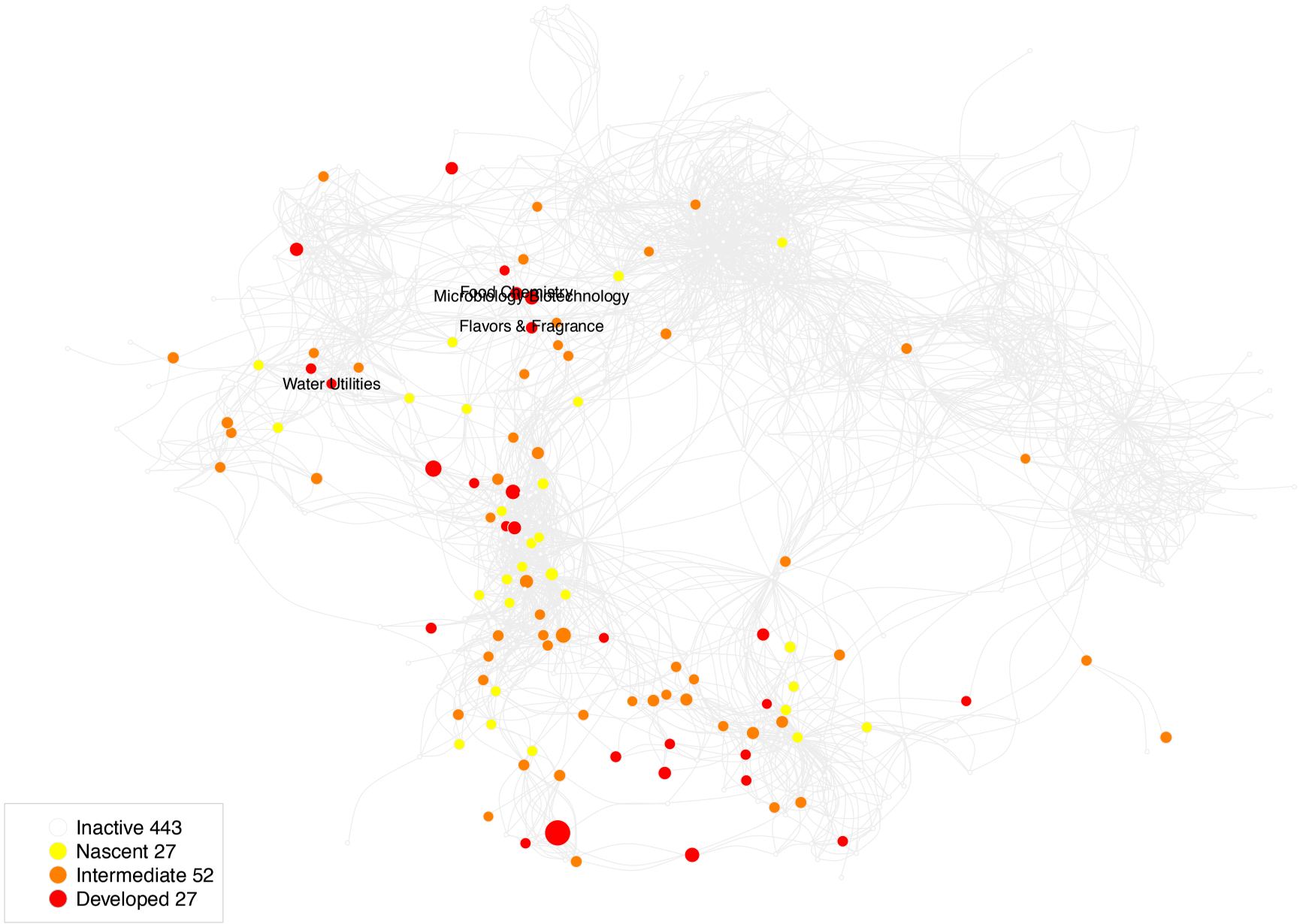

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3924 < Growing Areas < 1.1177 Developed Areas

**Algeria 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



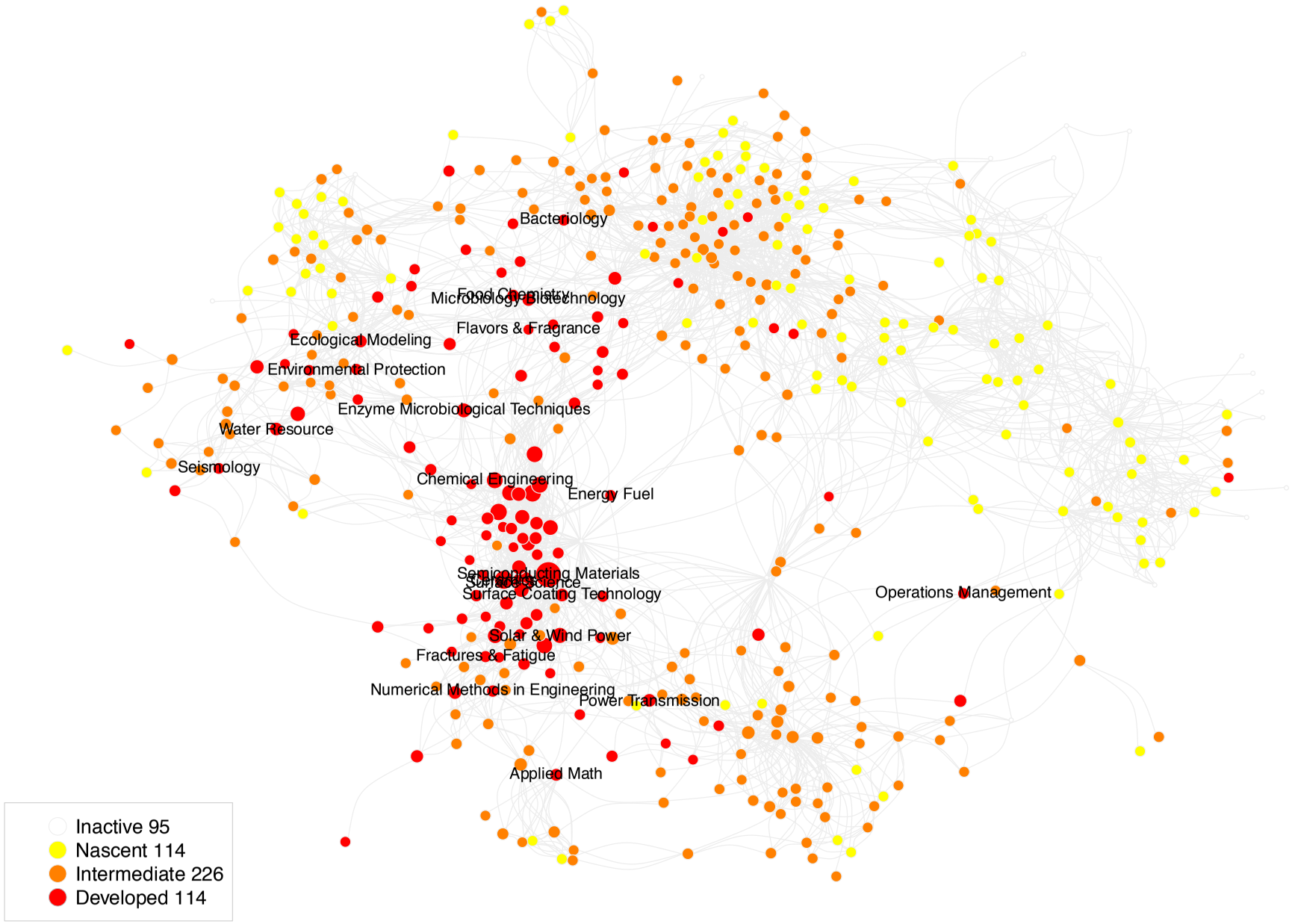

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 1.2053 < Growing Areas < 5.2321 Developed Areas

**Algeria 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



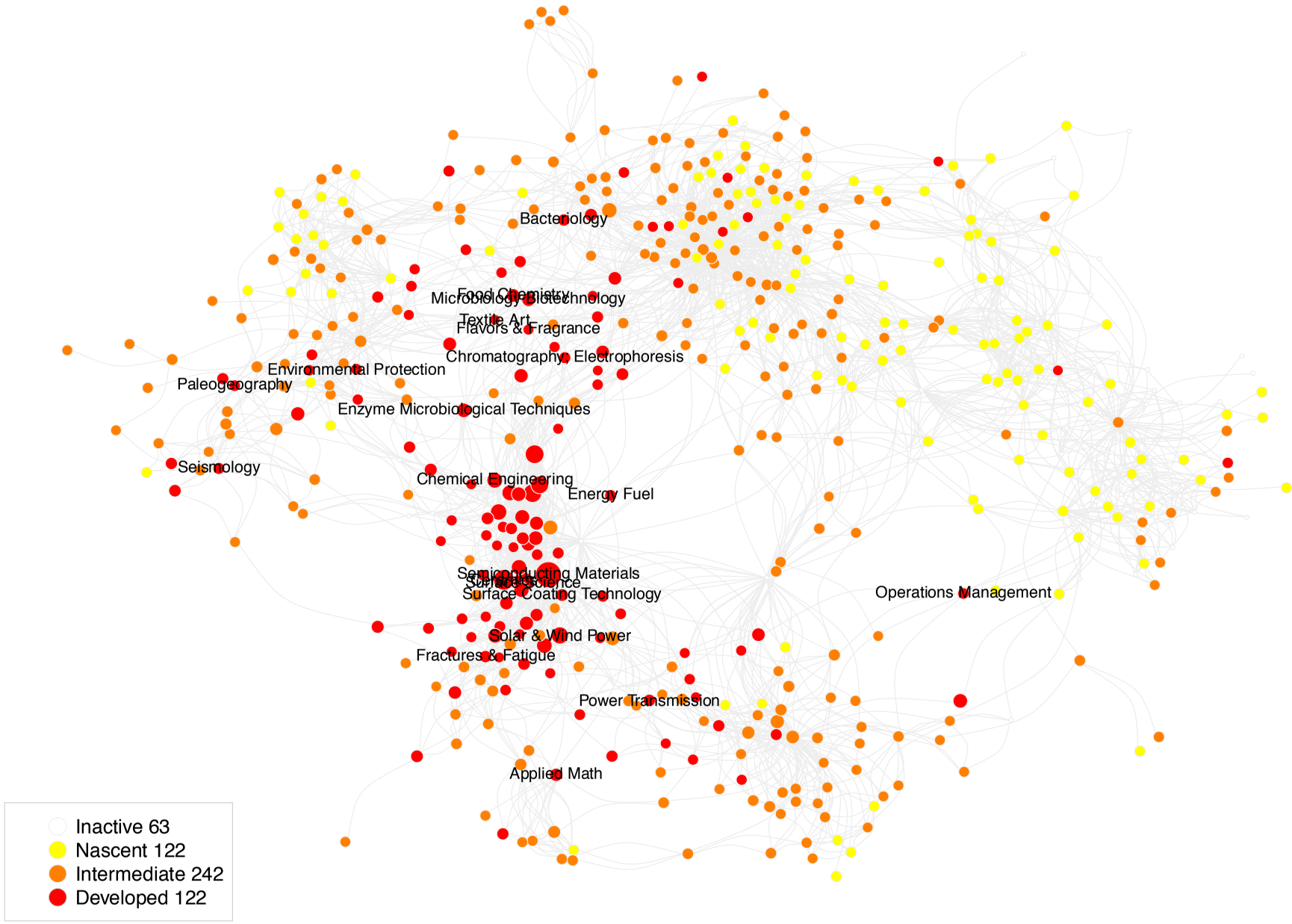

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 1.2812 < Growing Areas < 4.5406 Developed Areas

**India 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



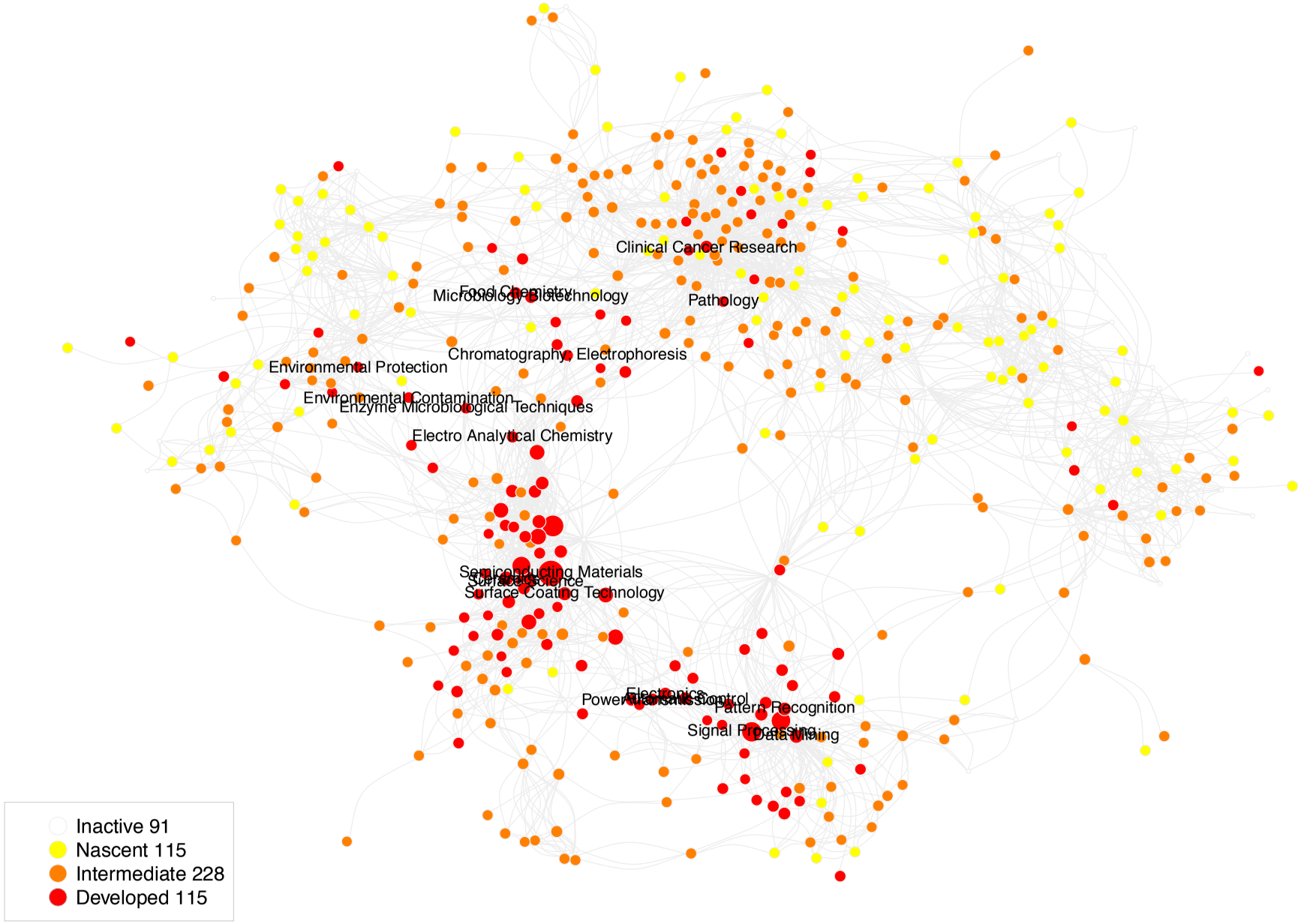

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2236 < Growing Areas < 1.2711 Developed Areas

**India 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



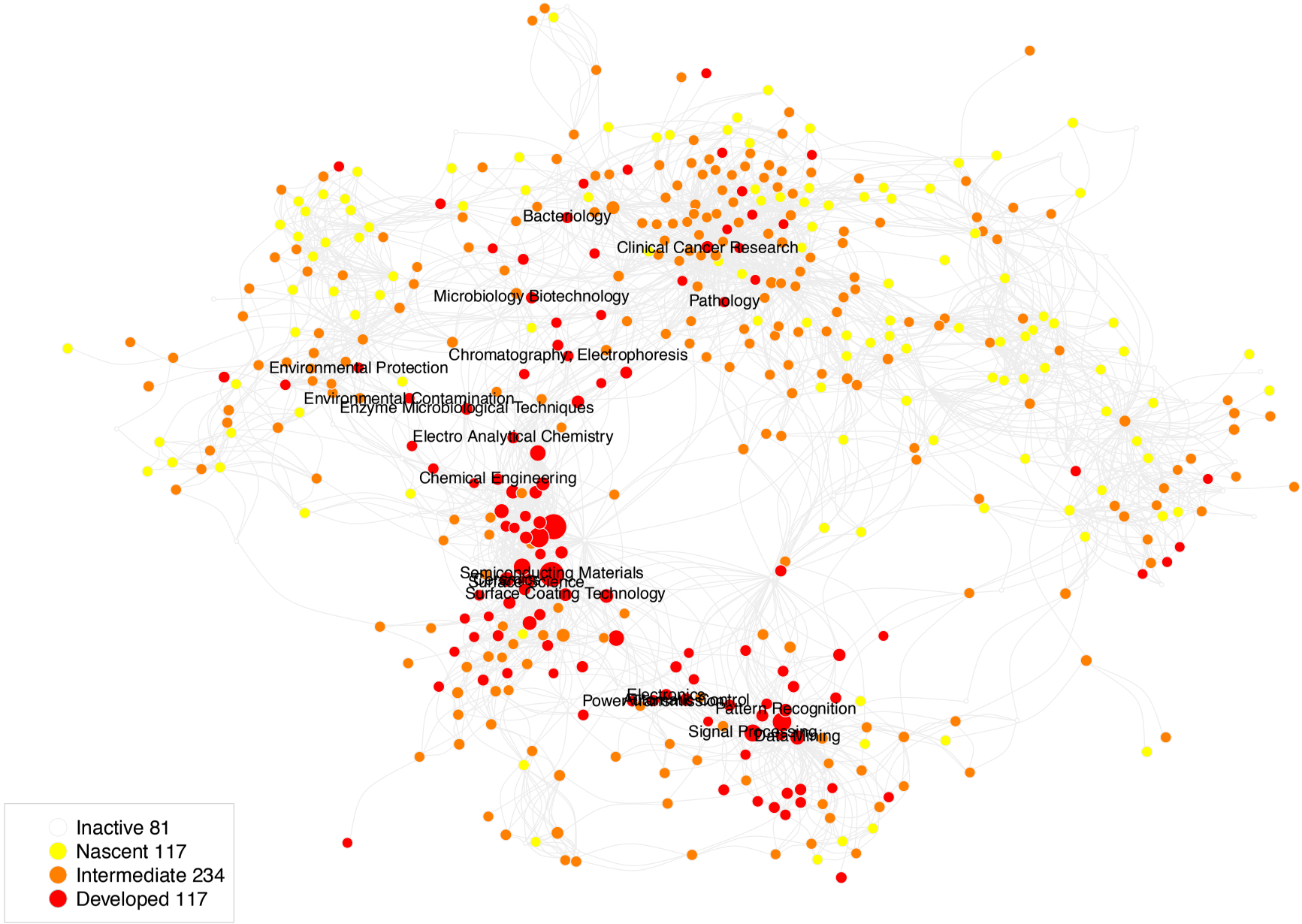

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2252 < Growing Areas < 1.2749 Developed Areas

**South Korea 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



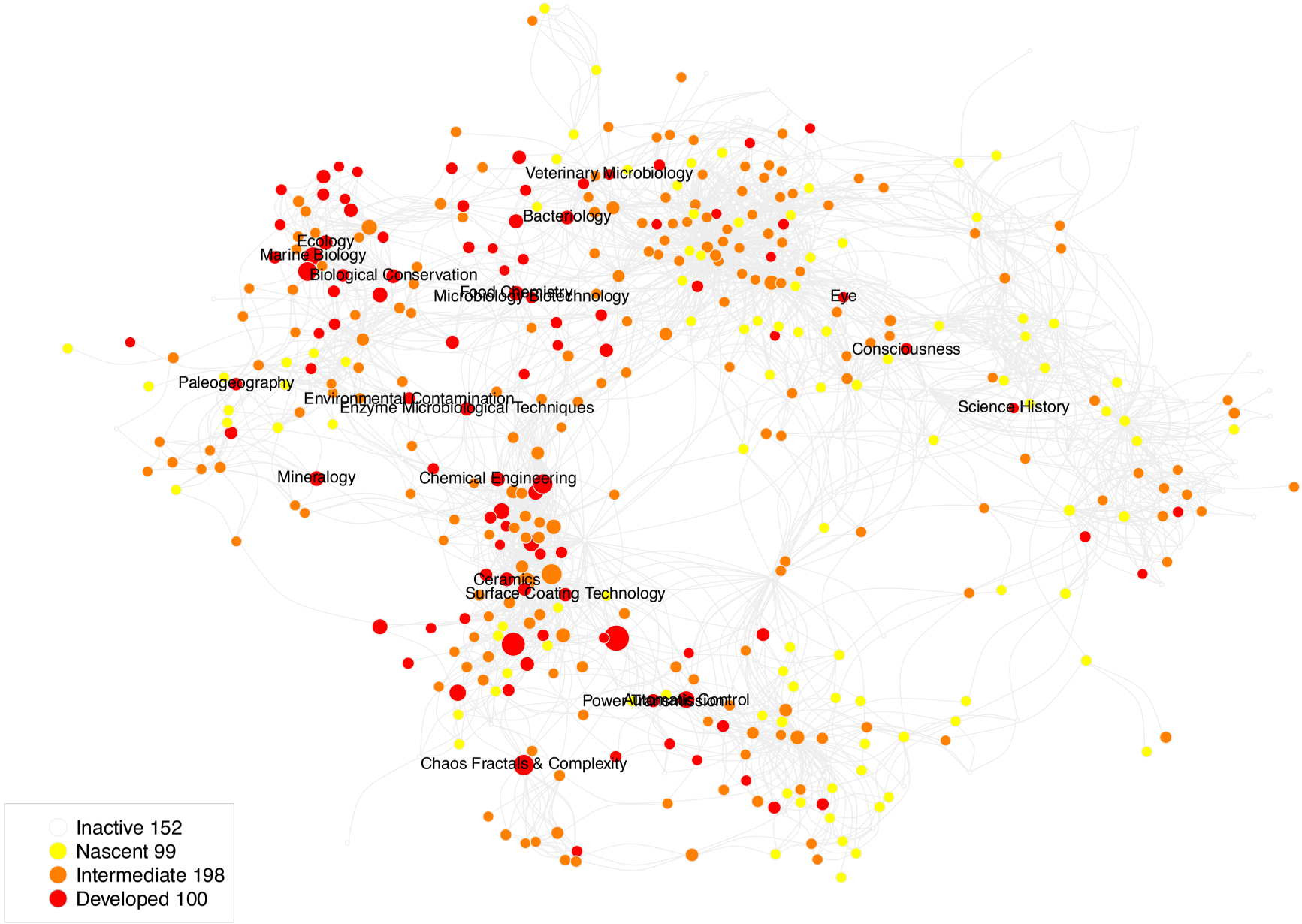

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2449 < Growing Areas < 1.0006 Developed Areas

**South Korea 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



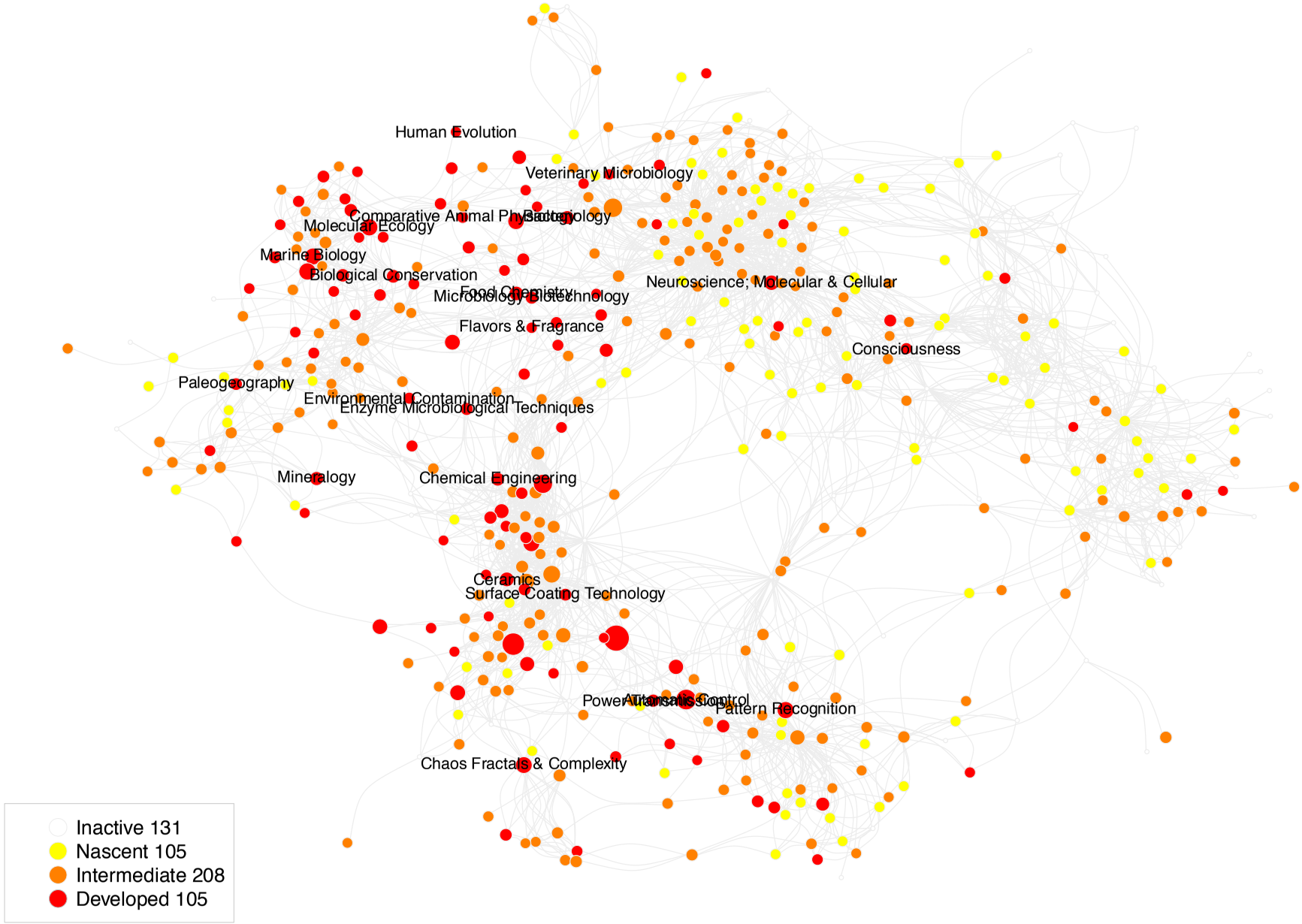

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2689 < Growing Areas < 1.029 Developed Areas

**Mexico 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**

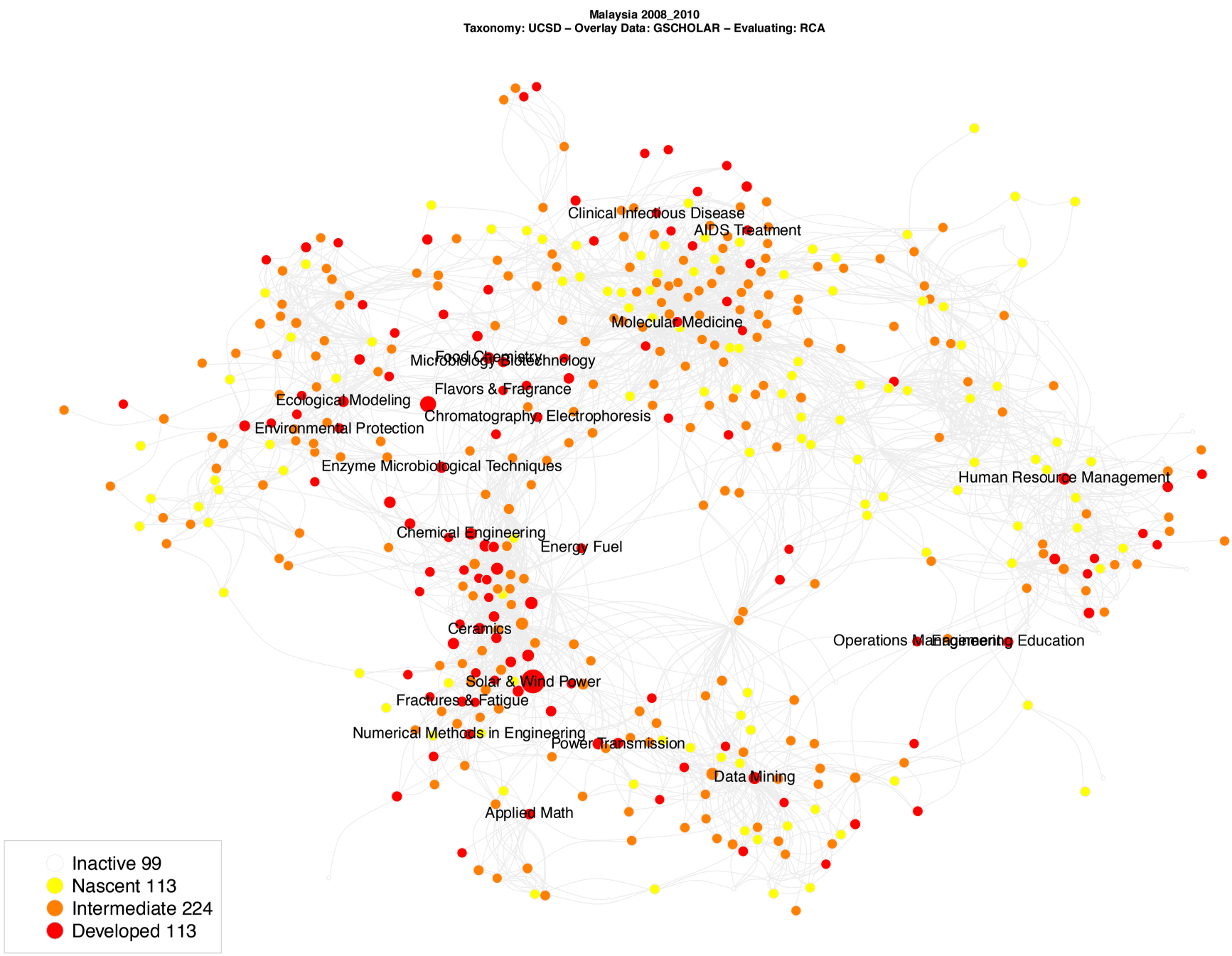



Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.3404 < Growing Areas < 1.5546 Developed Areas

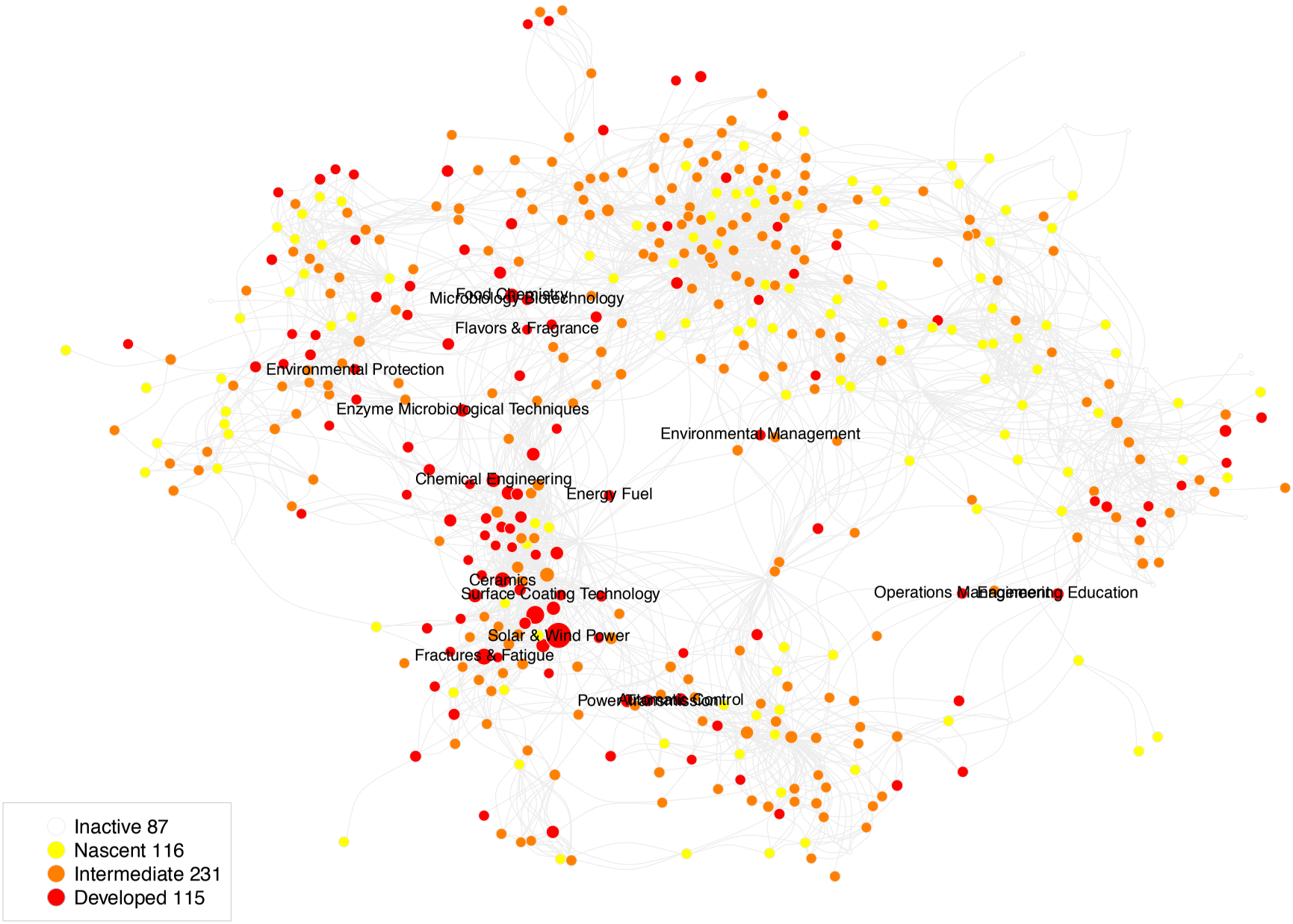
**Mexico 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**




Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.272 < Growing Areas < 1.5047 Developed Areas

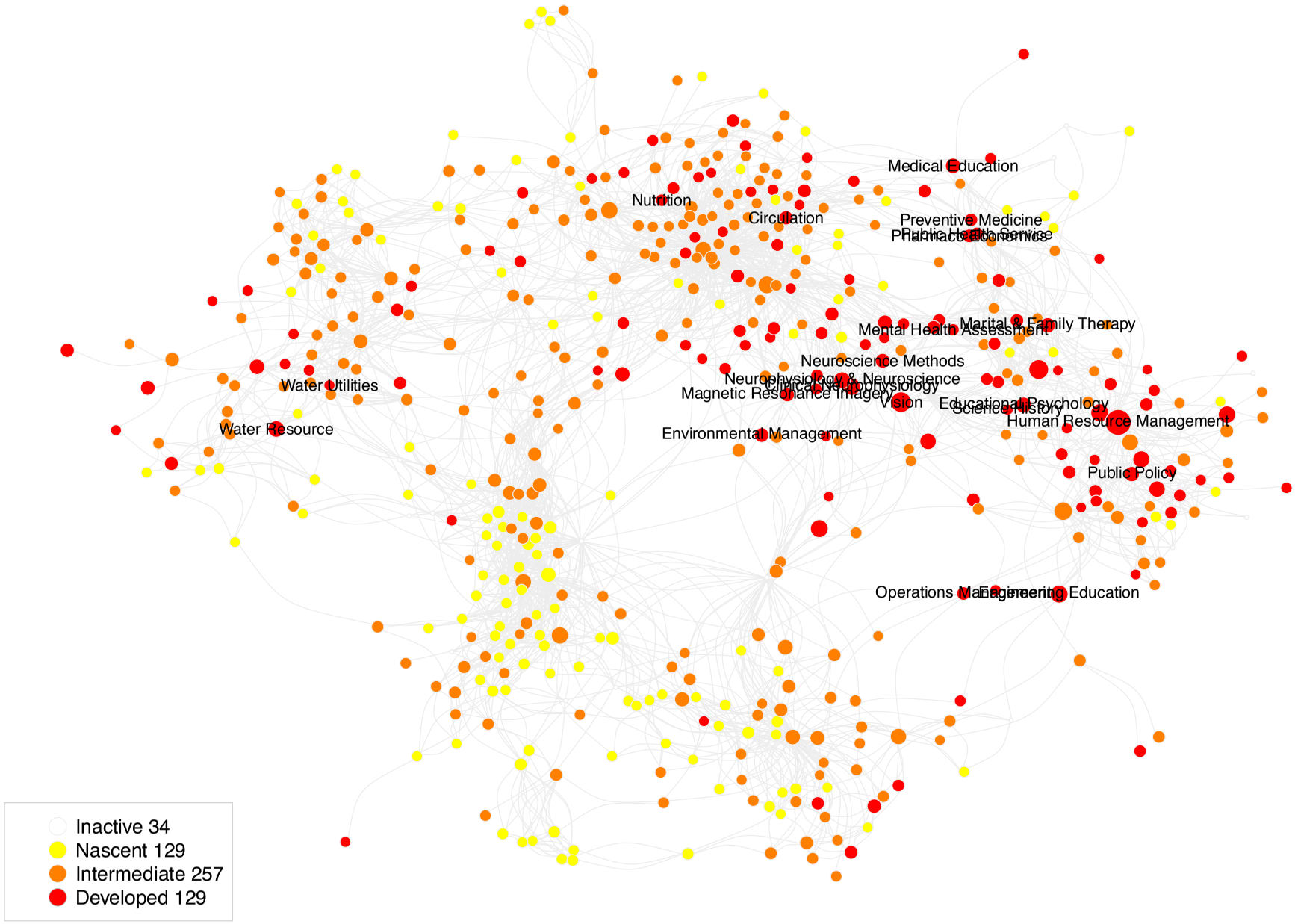

Malaysia 2008_2010
Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA
Clinical Infectious Disease
AIDS Treatment
Molecular Medicine
Food Chemistry
Microbiology Biotechnology
Flavors & Fragrance
Ecological Modeling
Chromatography, Electrophoresis
Environmental Protection
Enzyme Microbiological Techniques
Human Resource Management
Chemical Engineering
Energy Fuel
Ceramics
Operations Management
Engineering Education
Solar & Wind Power
Fractures & Fatigue
Numerical Methods in Engineering
Power Transmission
Data Mining
Applied Math
Inactive 99
Nascent 113
Intermediate 224
Developed 113
Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2749 < Growing Areas < 1.3248 Developed Areas

**Malaysia 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



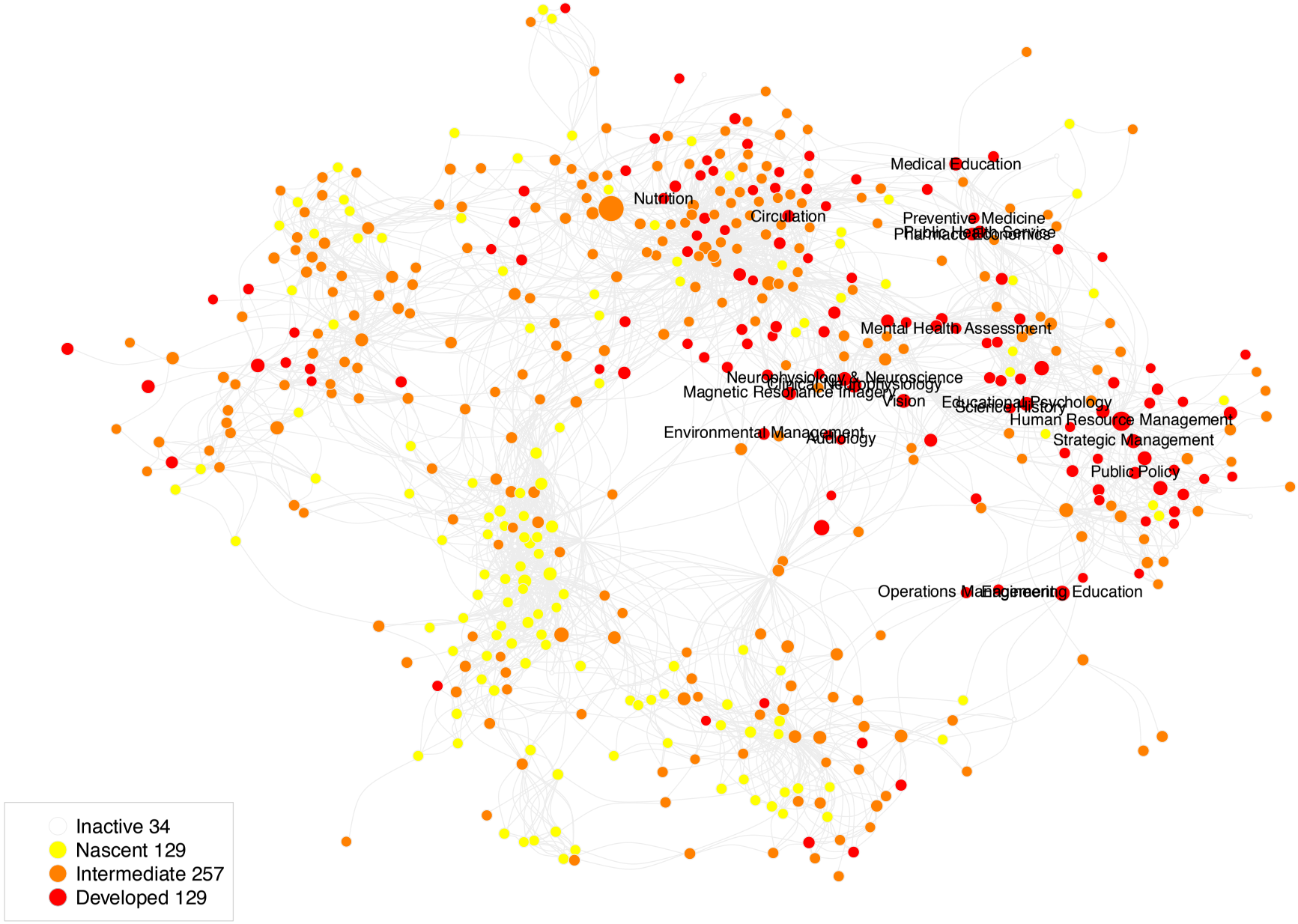

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.2622 < Growing Areas < 1.3345 Developed Areas

**Netherlands 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



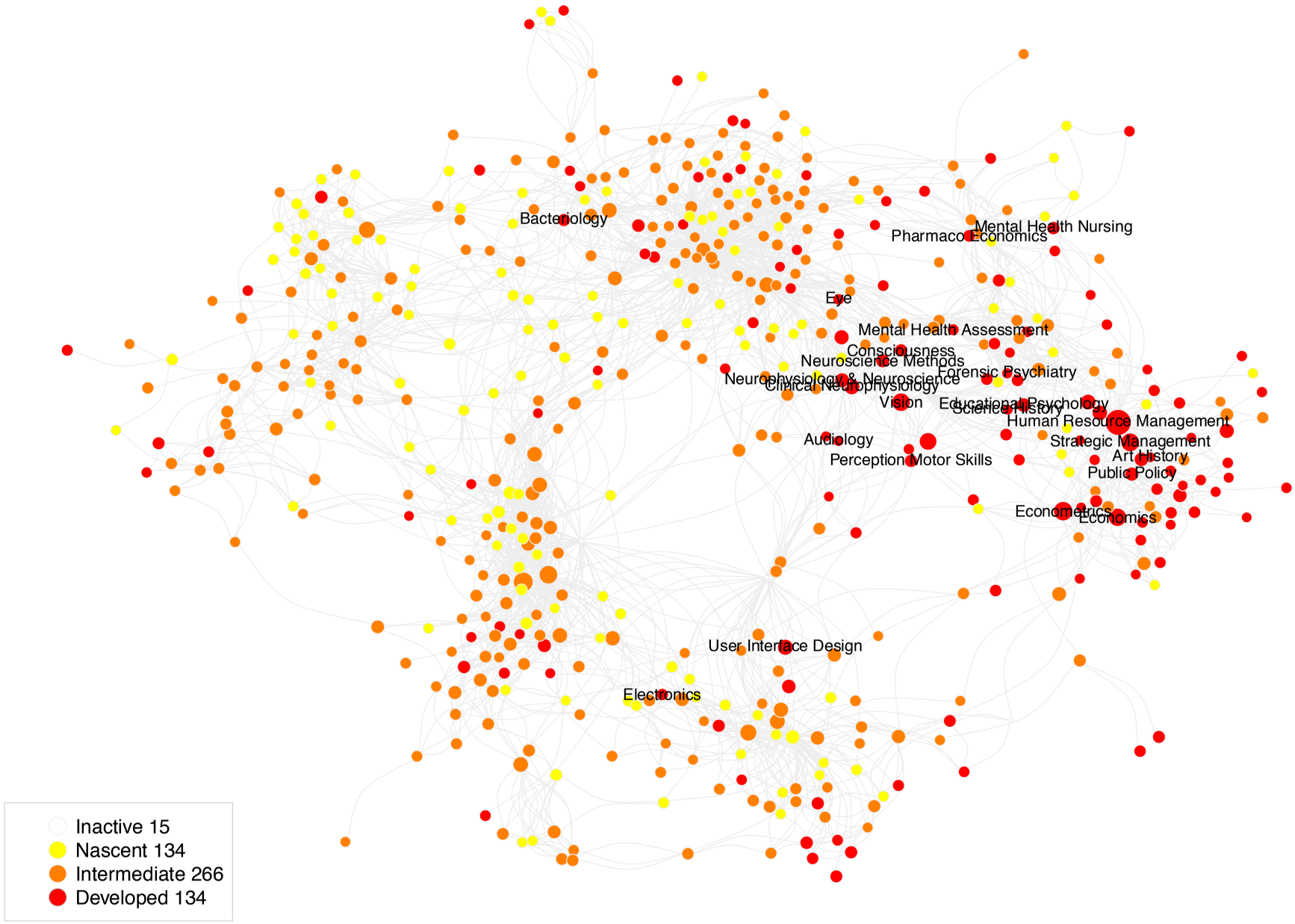

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.4881 < Growing Areas < 1.494 Developed Areas

**Netherlands 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



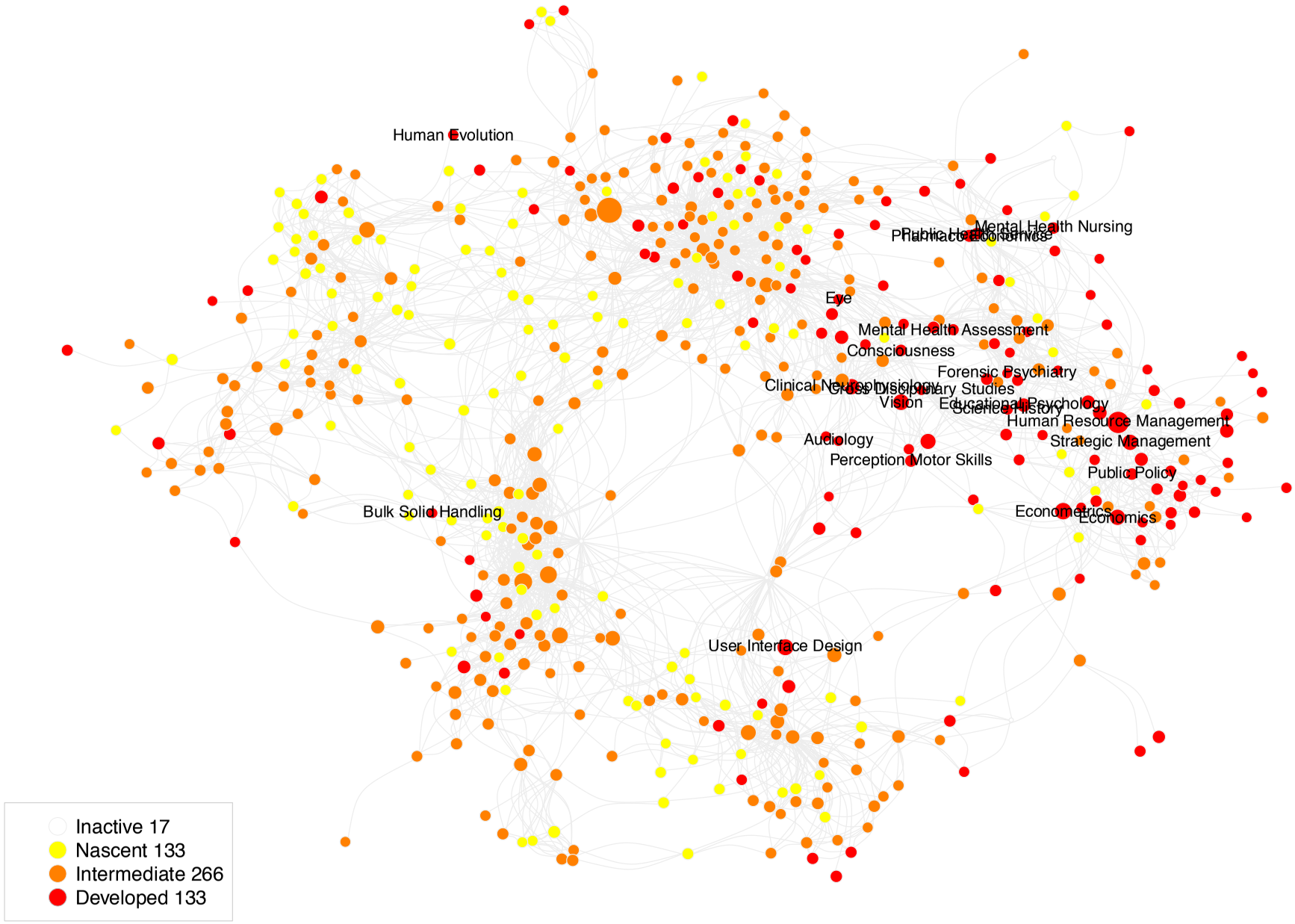

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.5214 < Growing Areas < 1.5304 Developed Areas

**United Kingdom 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



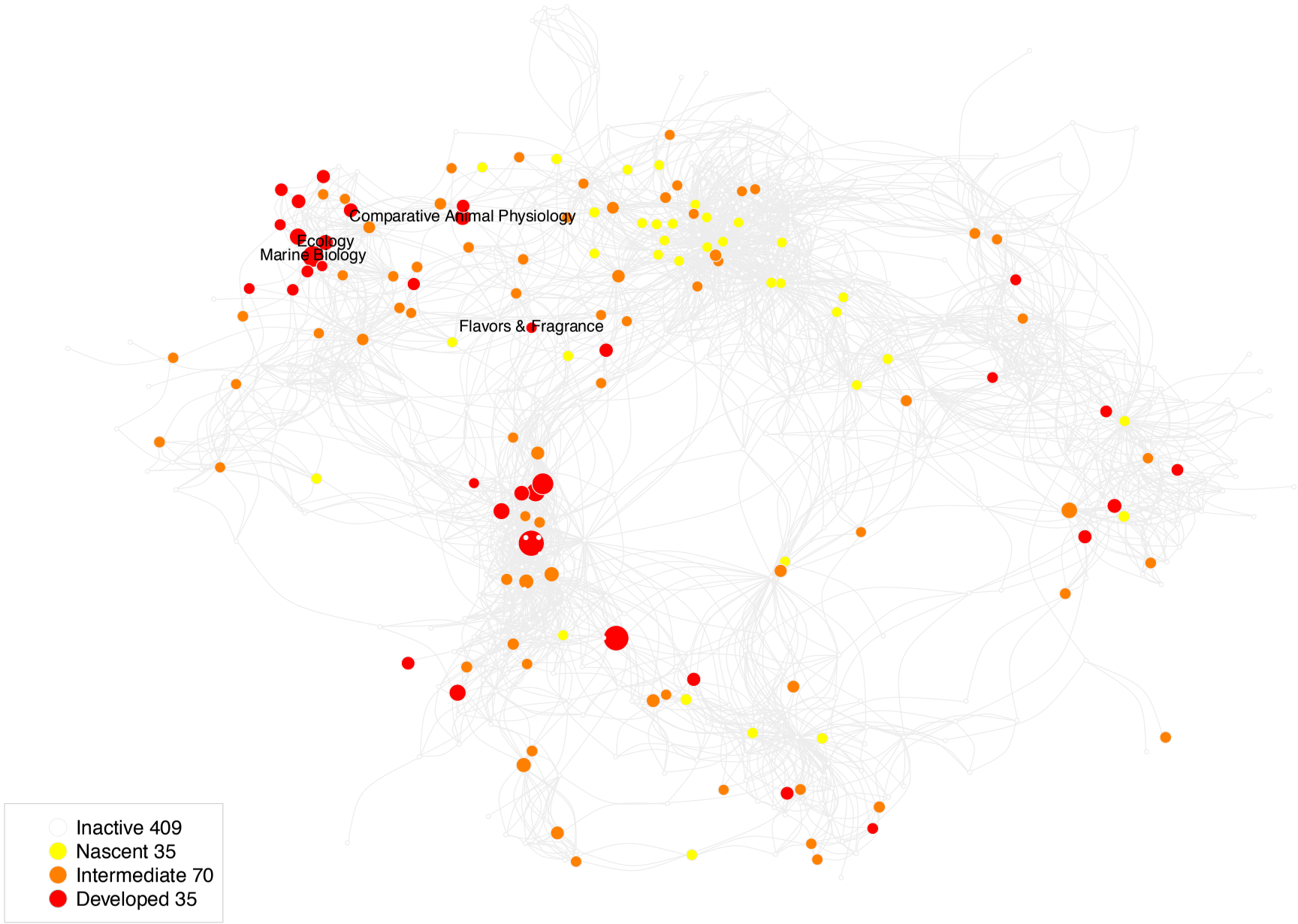

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.6085 < Growing Areas < 1.3184 Developed Areas

**United Kingdom 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



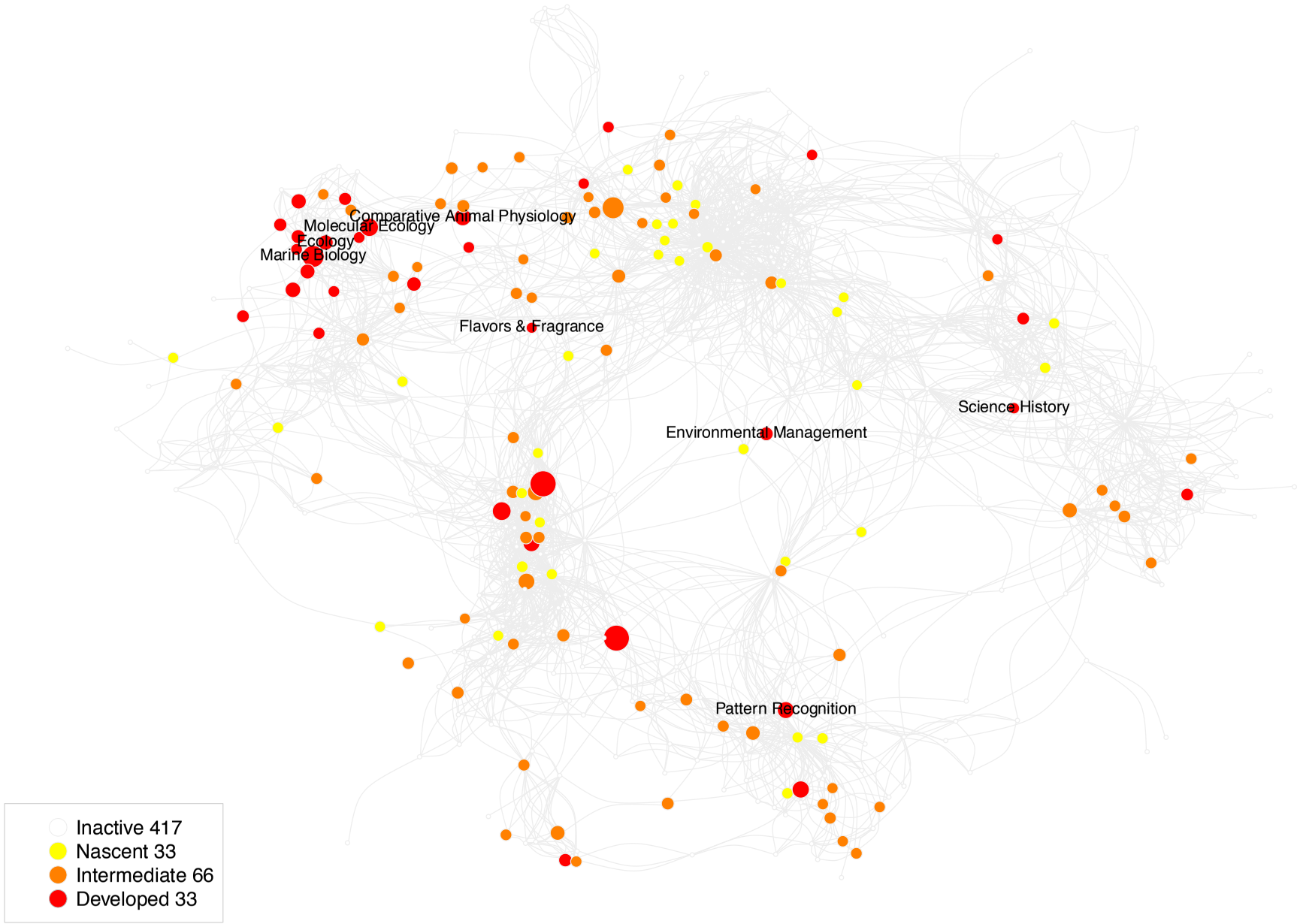

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.6296 < Growing Areas < 1.2861 Developed Areas

**Uruguay 2008_2010**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**




Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.6307 < Growing Areas < 3.3468 Developed Areas

**Uruguay 2011_2013**
**Taxonomy: UCSD – Overlay Data: GSCHOLAR – Evaluating: RCA**



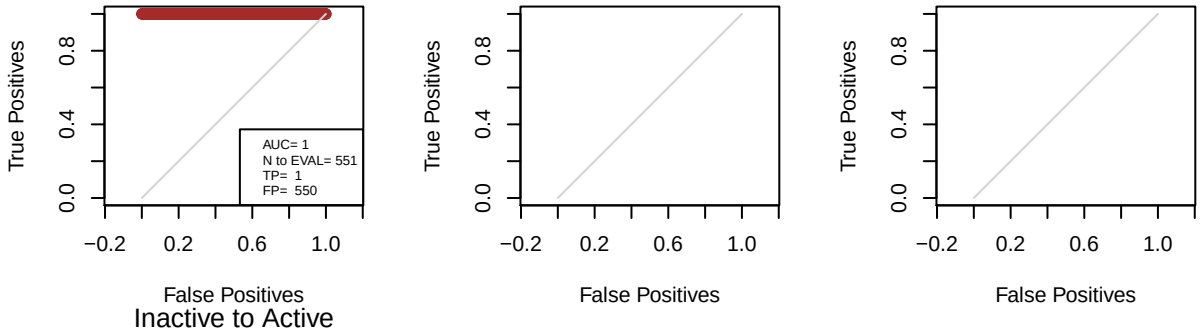

Layout: Fruchterman...Reingold I Size: Share of authorships I Color: Values of RCA
Agg. function: sum
Undeveloped < 0.871 < Growing Areas < 3.5841 Developed Areas

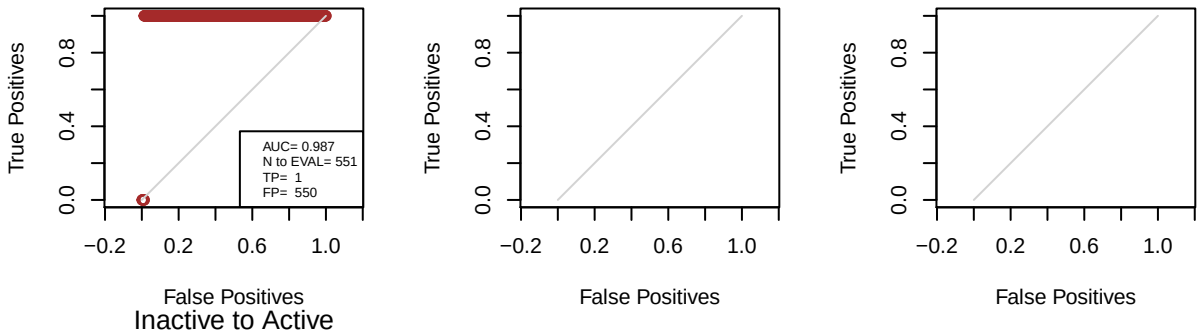
Apéndice F

# Comparación de curvas ROC para individuos entre el espacio investigación y el mapa UCSD

Jeffrey Andrews

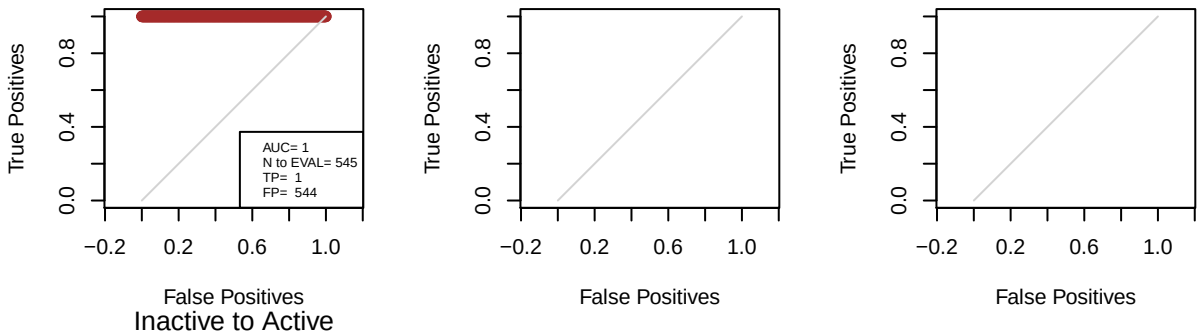
2008_2010 − 2011_2013 | RS vs UCSD

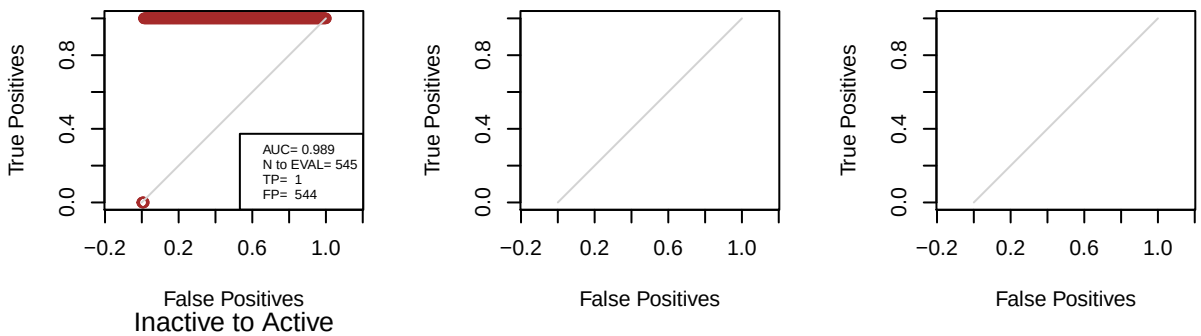



Pablo A. Denis

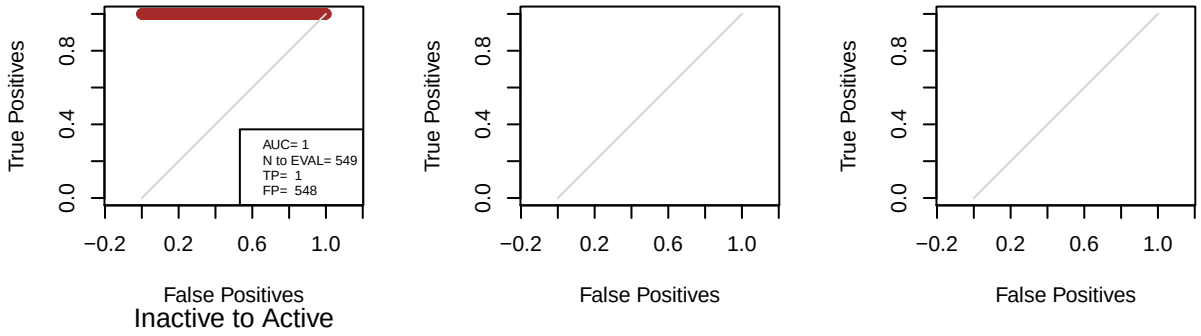
2008_2010 − 2011_2013 | RS vs UCSD

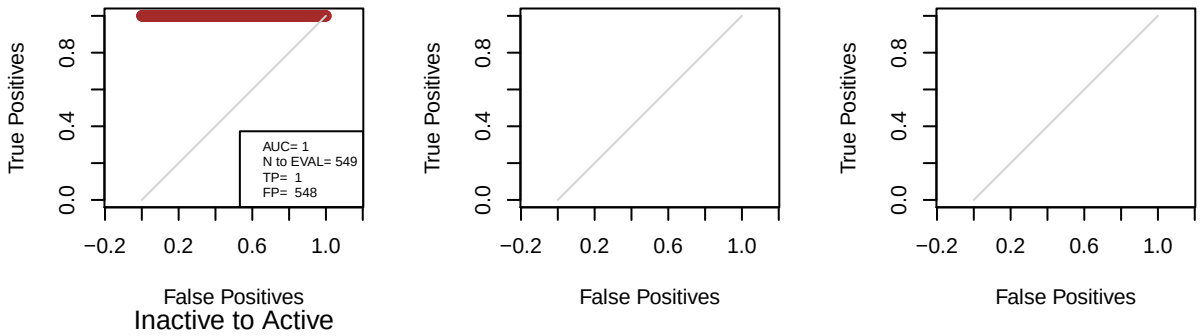


Michael A. Henning

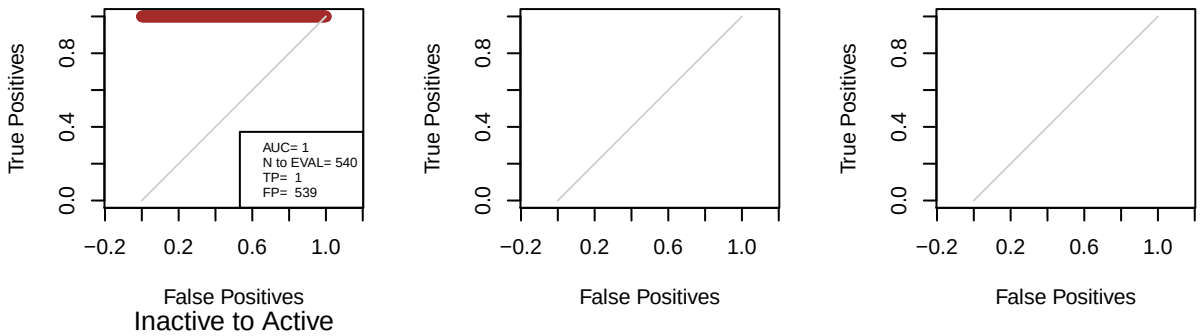
2008_2010 − 2011_2013 | RS vs UCSD

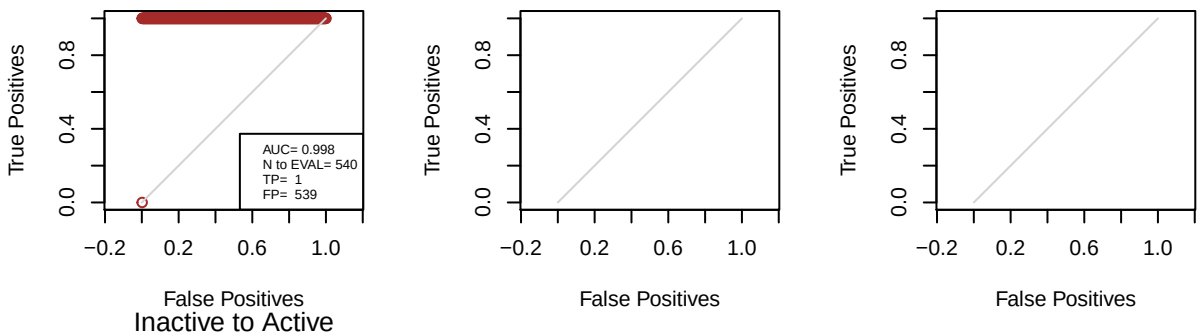


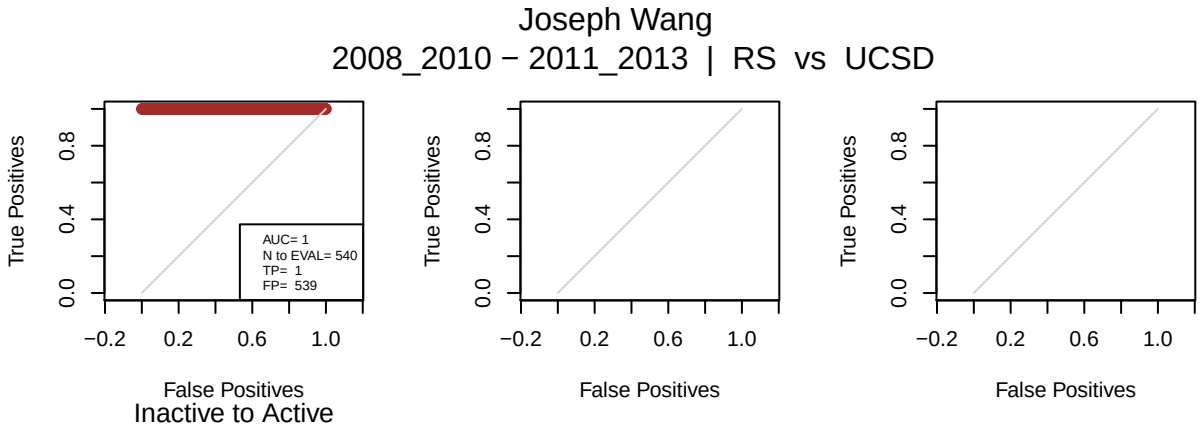



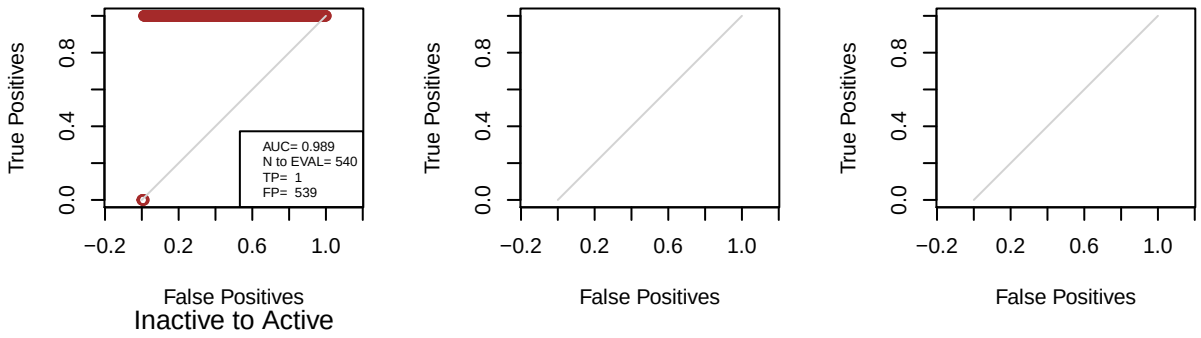

Chad Cook

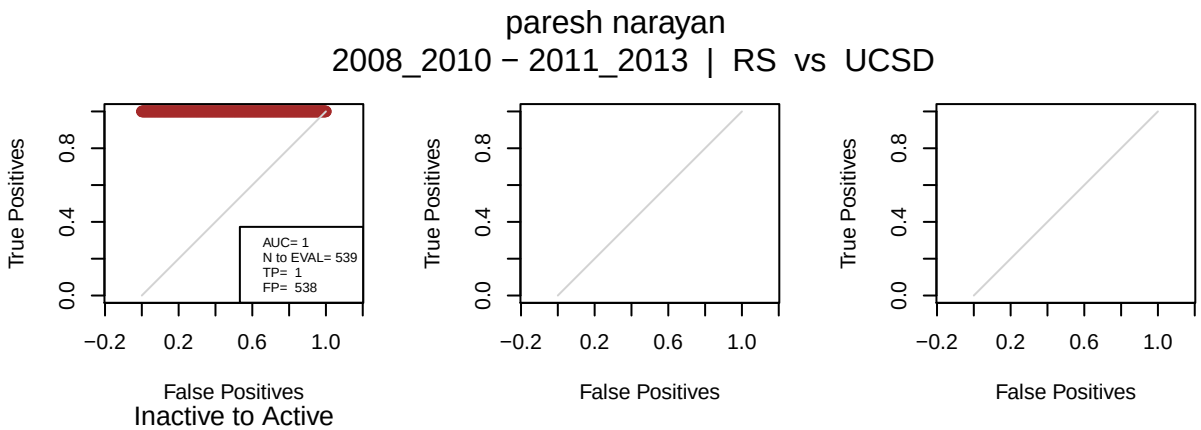
2008_2010 − 2011_2013 | RS vs UCSD

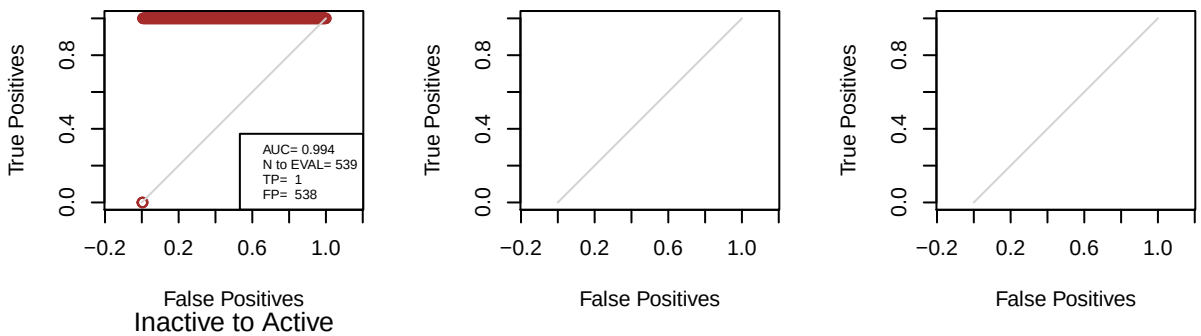


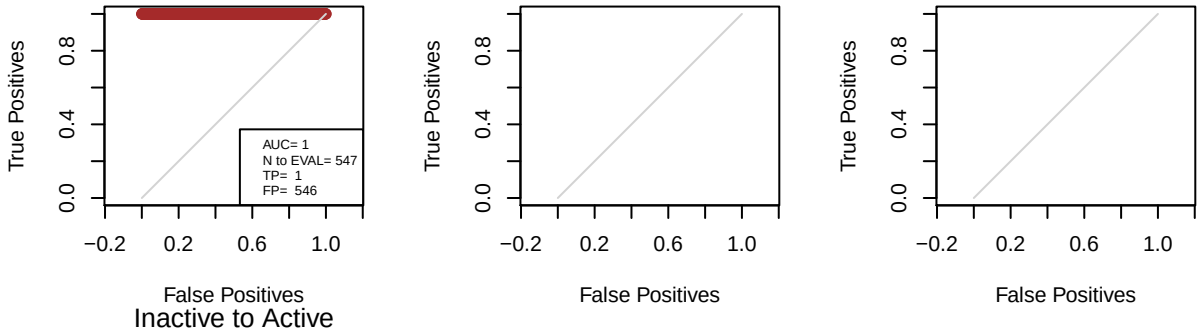

Joseph Wang
2008_2010 − 2011_2013 | RS vs UCSD
True Positives
False Positives
AUC= 1
N to EVAL= 540
TP= 1
FP= 539
Inactive to Active

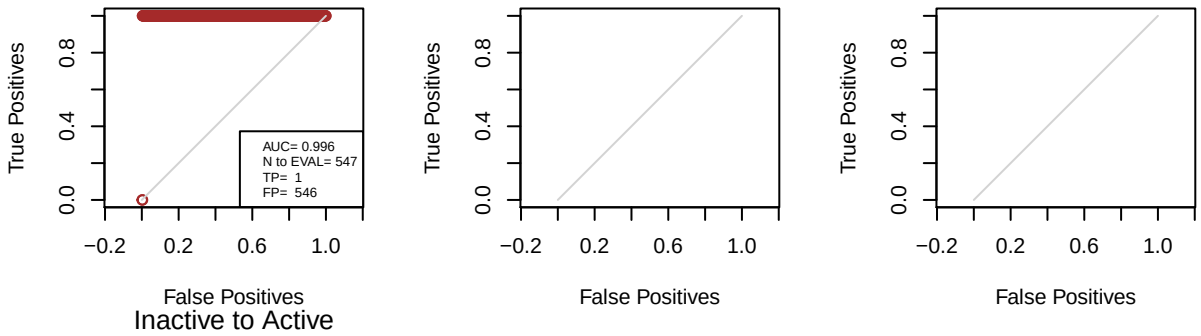

True Positives
False Positives
AUC= 0.989
N to EVAL= 540
TP= 1
FP= 539
Inactive to Active


paresh narayan
2008_2010 − 2011_2013 | RS vs UCSD
True Positives
False Positives
AUC= 1
N to EVAL= 539
TP= 1
FP= 538
Inactive to Active

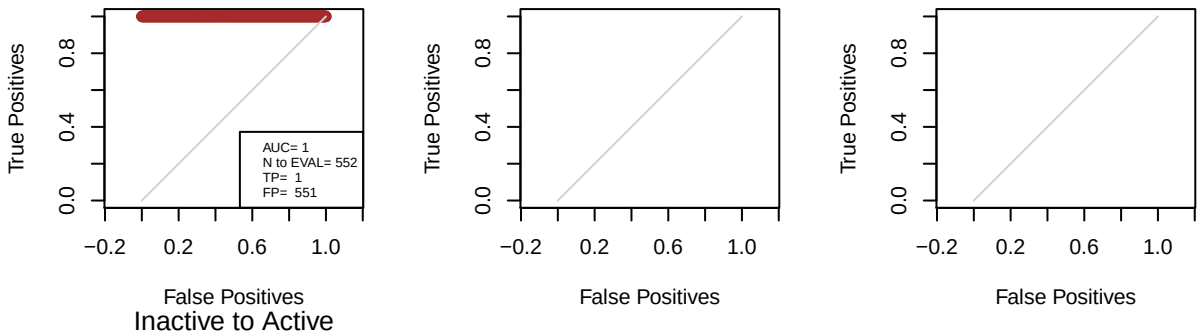

True Positives
False Positives
AUC= 0.994
N to EVAL= 539
TP= 1
FP= 538
Inactive to Active

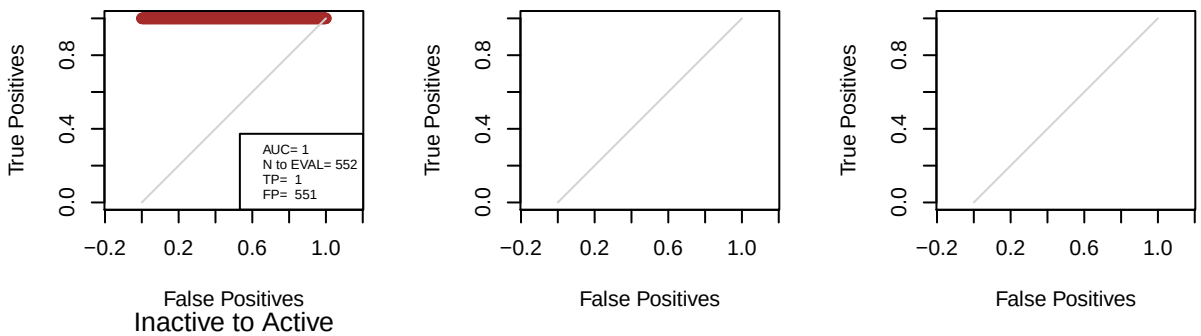
joshua aizenman
2008_2010 − 2011_2013 | RS vs UCSD




Matthijs J Warrens
2008_2010 − 2011_2013 | RS vs UCSD

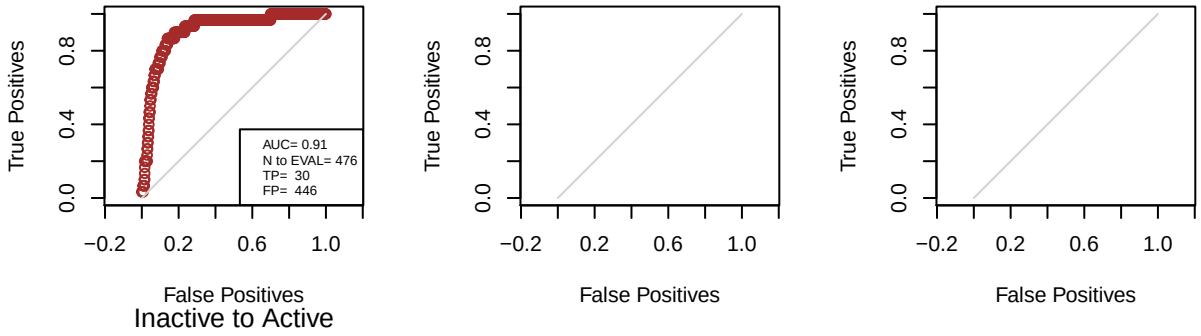


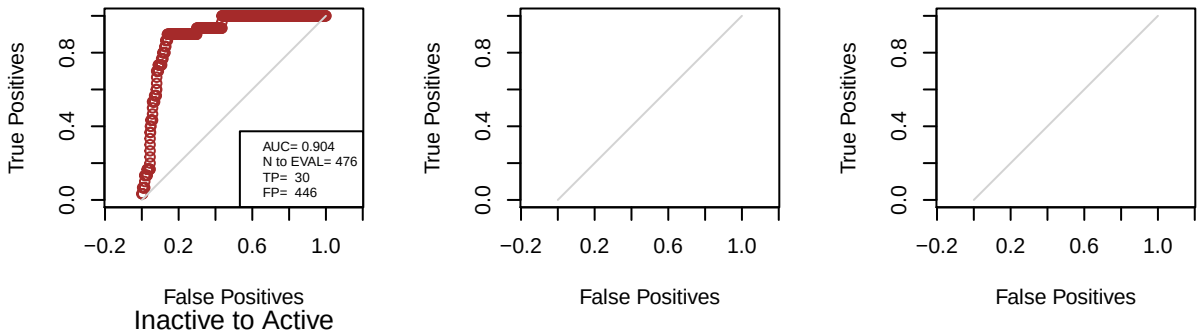
Apéndice G

# Comparación de curvas ROC para instituciones entre el espacio investigación y el mapa UCSD

## Assistance Publique Hôpitaux de Paris
## 2003_2005 − 2006_2008 | RS vs UCSD

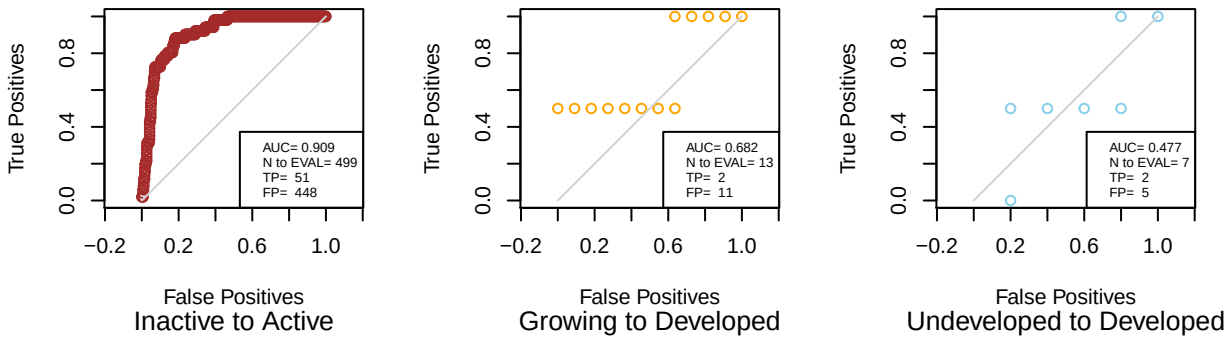



Inactive to Active

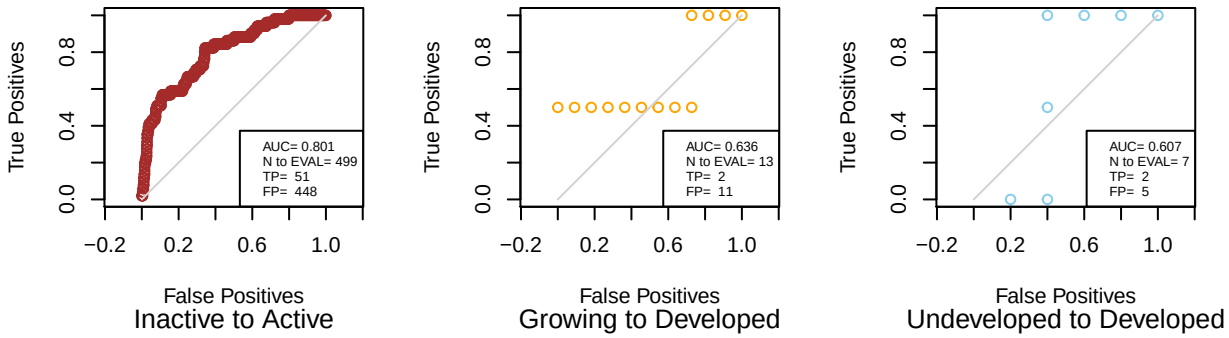



Inactive to Active

## Russian Academy of Sciences
## 2003_2005 − 2006_2008 | RS vs UCSD

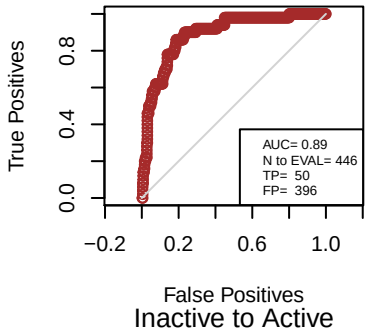



Inactive to Active | Growing to Developed | Undeveloped to Developed

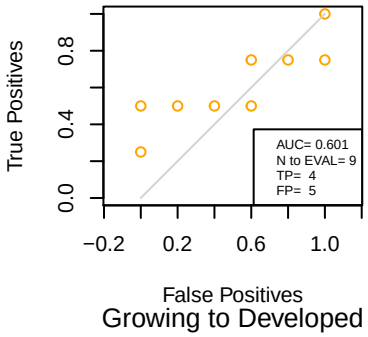



Inactive to Active | Growing to Developed | Undeveloped to Developed

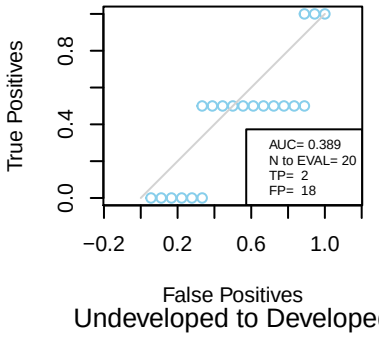
Amirkabir University of Technology
2003_2005 − 2006_2008 | RS vs UCSD



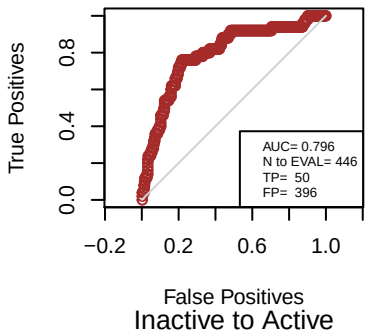
Inactive to Active



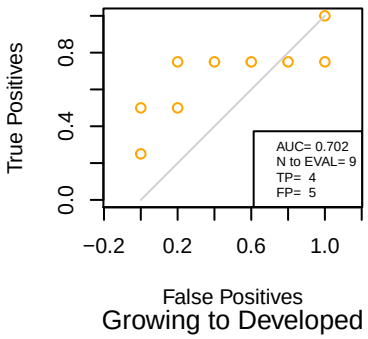
Growing to Developed



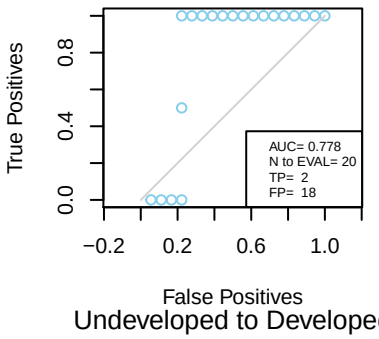
Undeveloped to Developed



Inactive to Active

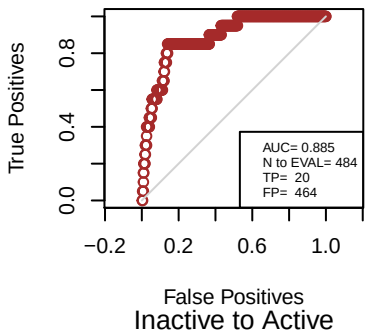


Growing to Developed

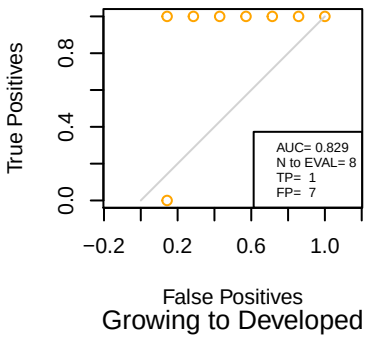


Undeveloped to Developed

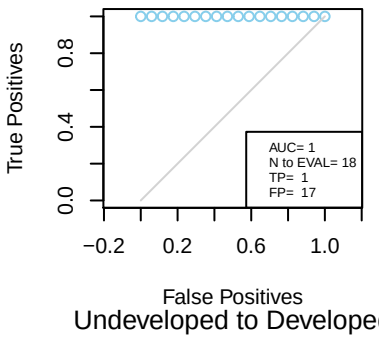
Tata Institute of Fundamental Research
2003_2005 − 2006_2008 | RS vs UCSD



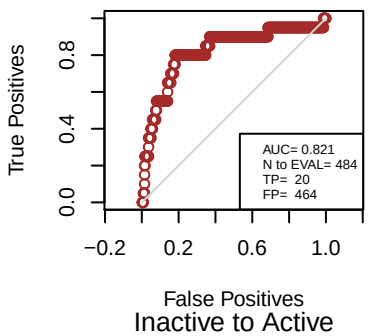
Inactive to Active



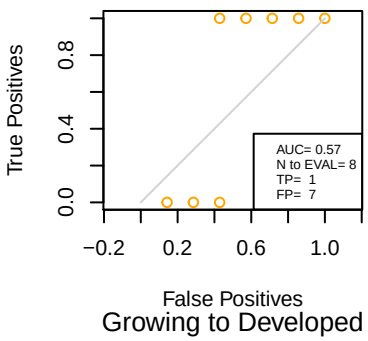
Growing to Developed



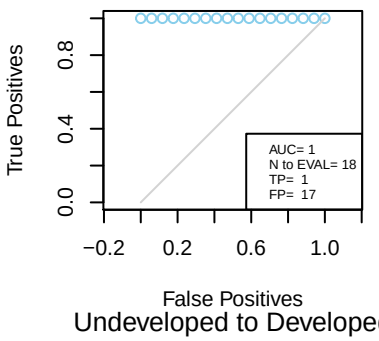
Undeveloped to Developed



Inactive to Active



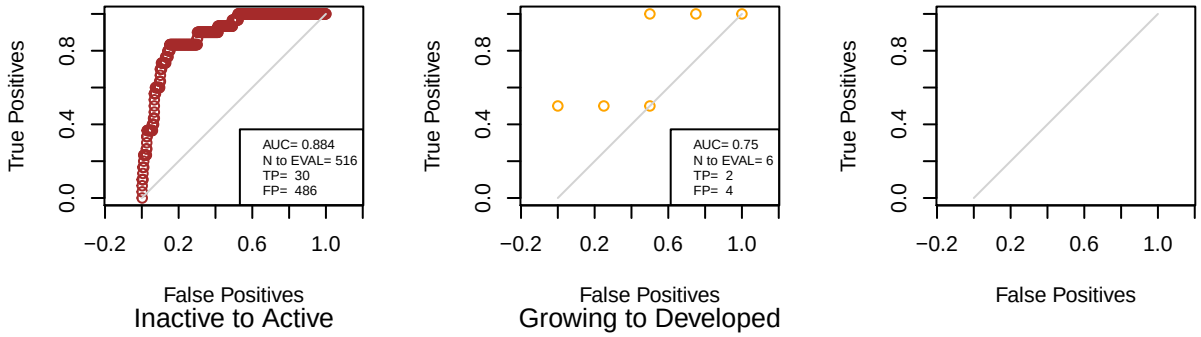
Growing to Developed



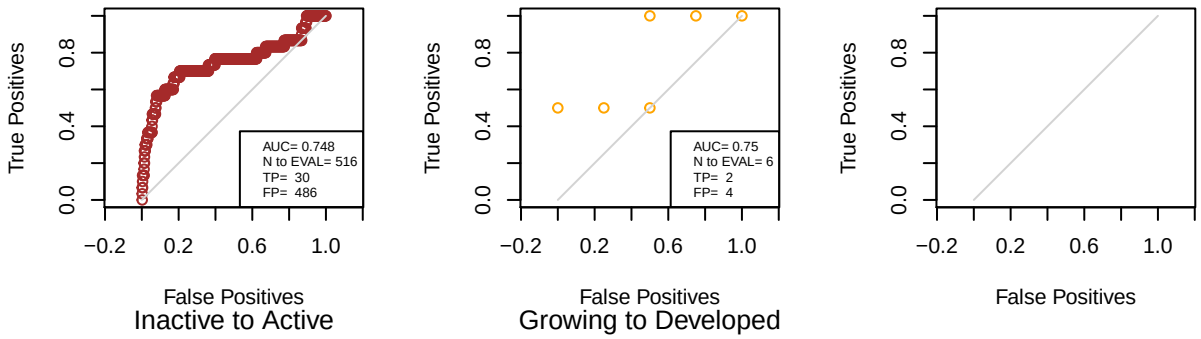
Undeveloped to Developed

Estación Biológica de Doñana

2003_2005 − 2006_2008 | RS vs UCSD



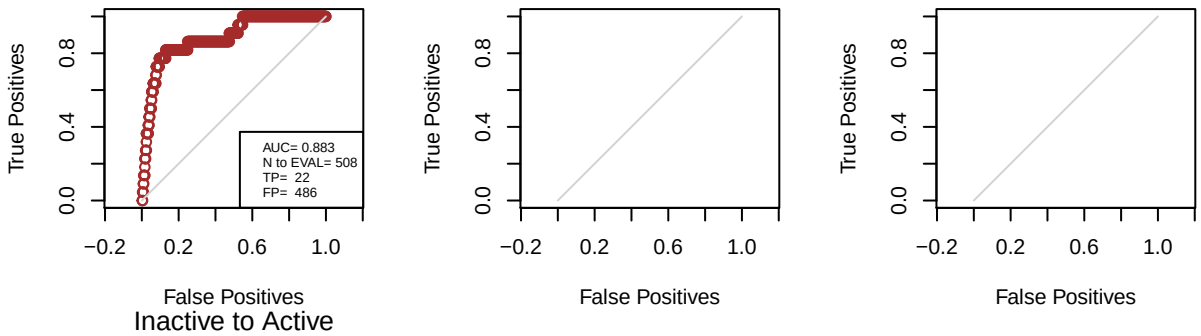

Inactive to Active | Growing to Developed



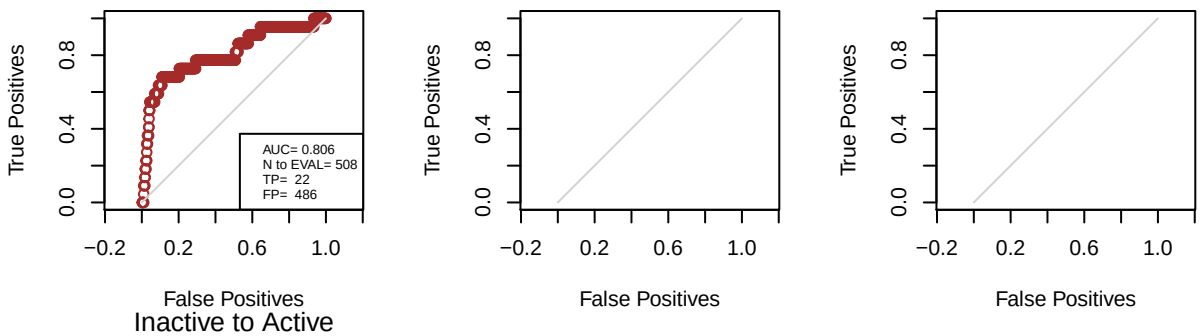

Inactive to Active | Growing to Developed

Landcare Research

2003_2005 − 2006_2008 | RS vs UCSD




Inactive to Active



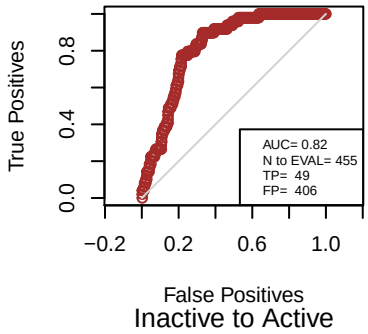

Inactive to Active

Apéndice H

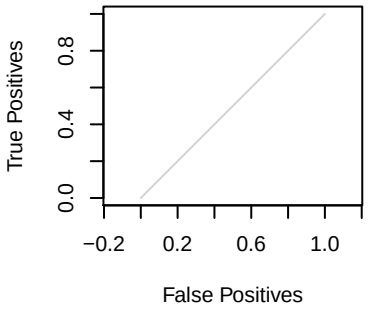
# Comparación de curvas ROC para países entre el espacio investigación y el mapa UCSD

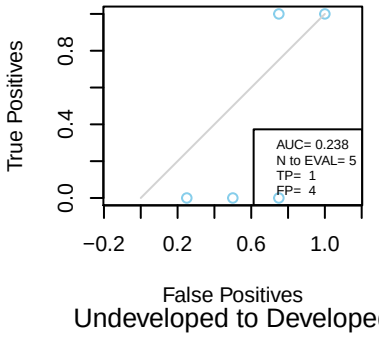
Algeria

2008_2010 − 2011_2013 | RS vs UCSD

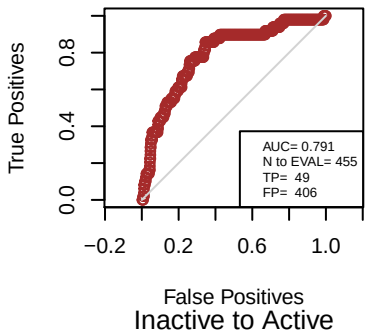



Inactive to Active

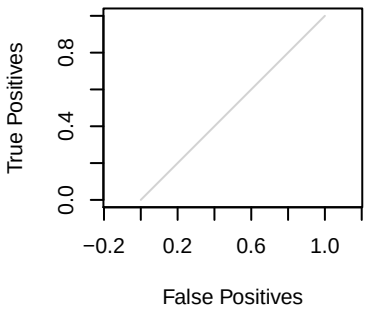


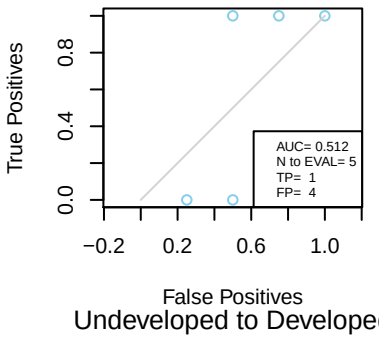




Undeveloped to Developed




Inactive to Active

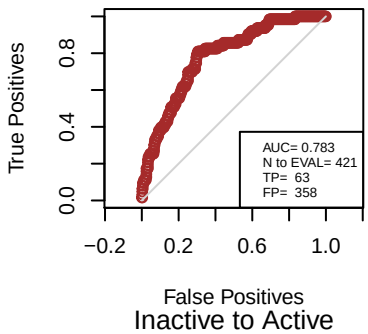


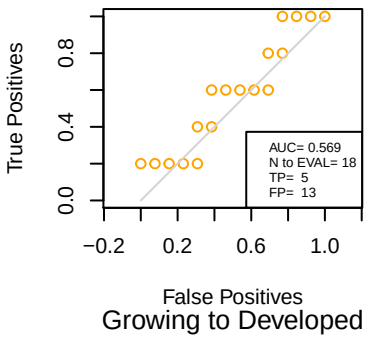

Undeveloped to Developed

Tunisia

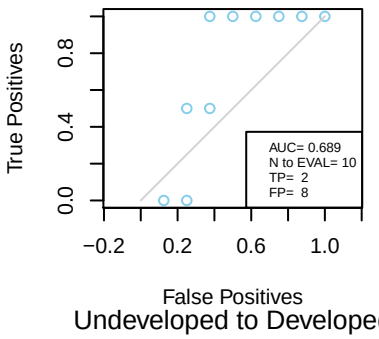
2008_2010 − 2011_2013 | RS vs UCSD



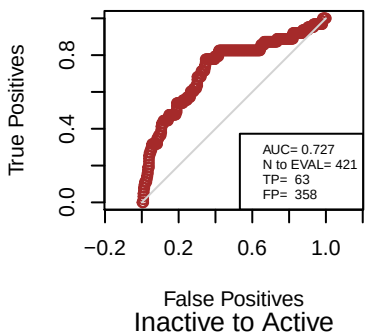

Inactive to Active



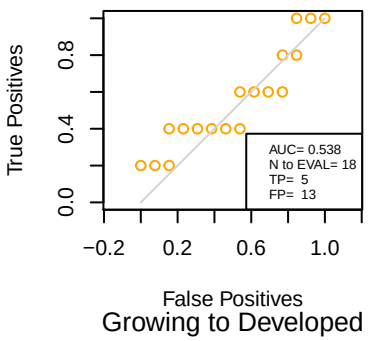

Growing to Developed



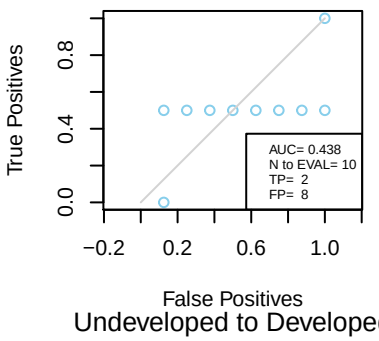

Undeveloped to Developed




Inactive to Active



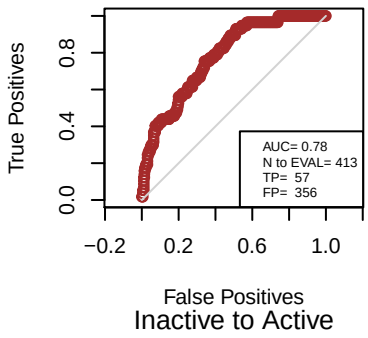

Growing to Developed



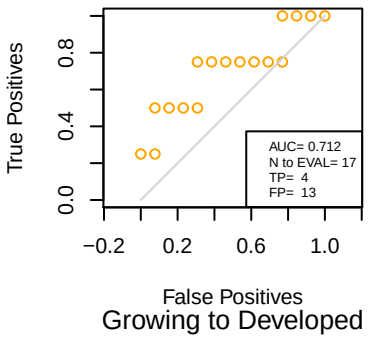

Undeveloped to Developed

Venezuela

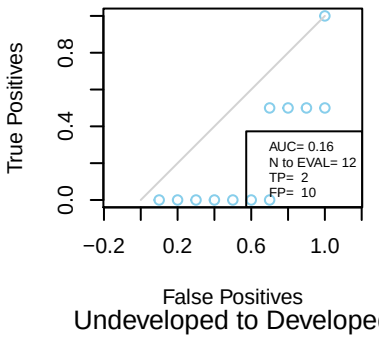
2008_2010 − 2011_2013 | RS vs UCSD



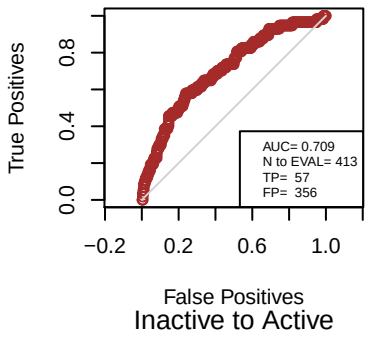

Inactive to Active



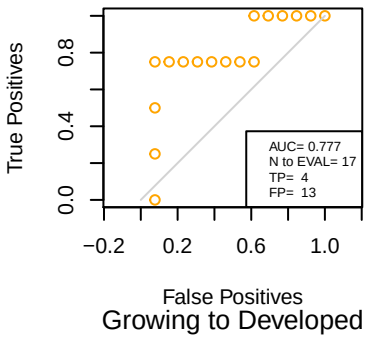

Growing to Developed



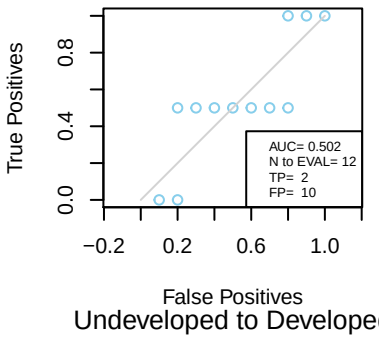

Undeveloped to Developed




Inactive to Active



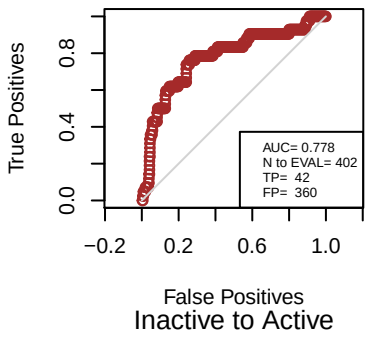

Growing to Developed



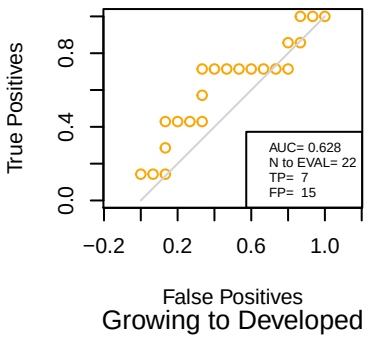

Undeveloped to Developed

Lebanon

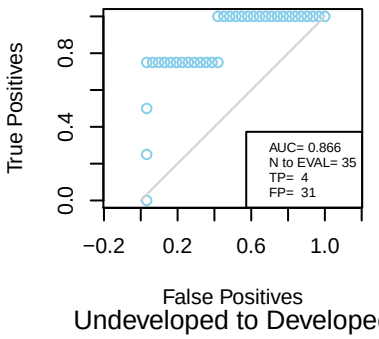
2008_2010 − 2011_2013 | RS vs UCSD



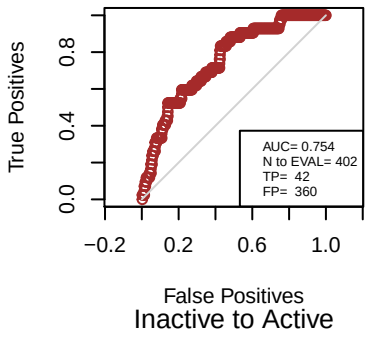

Inactive to Active



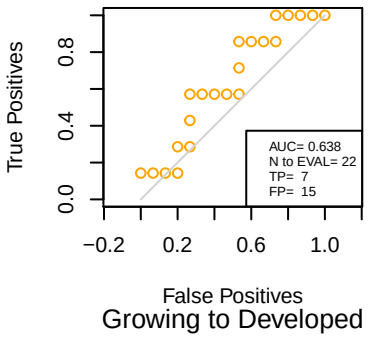

Growing to Developed



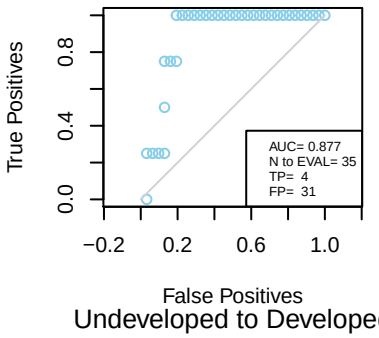

Undeveloped to Developed




Inactive to Active



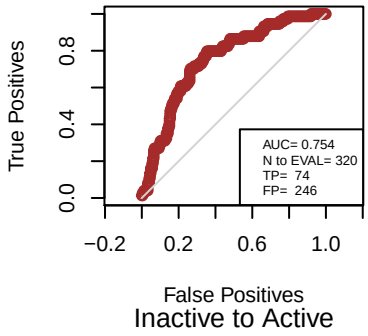

Growing to Developed



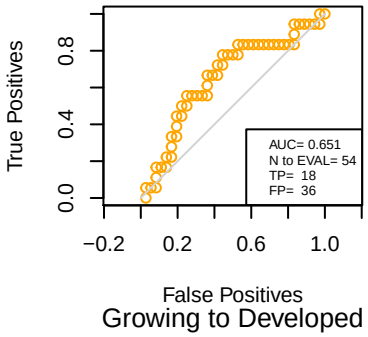

Undeveloped to Developed

Serbia

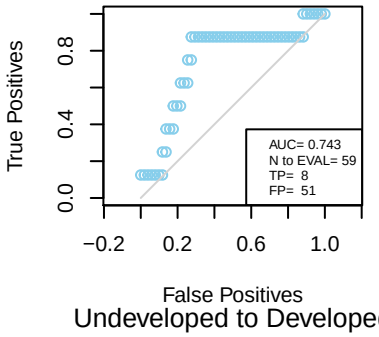
2008_2010 − 2011_2013 | RS vs UCSD



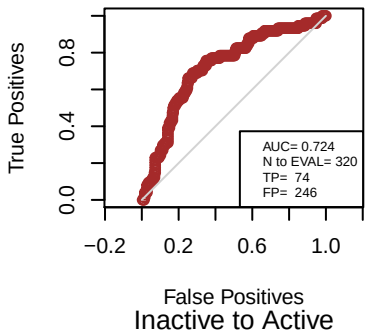

Inactive to Active



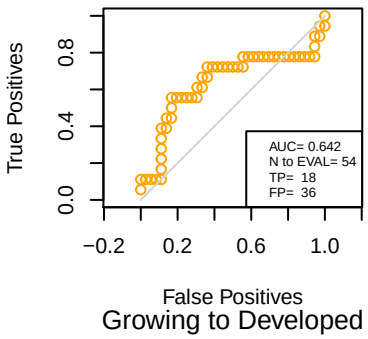

Growing to Developed



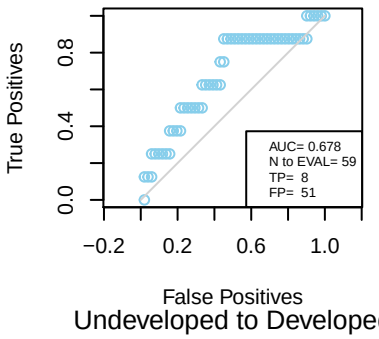

Undeveloped to Developed




Inactive to Active



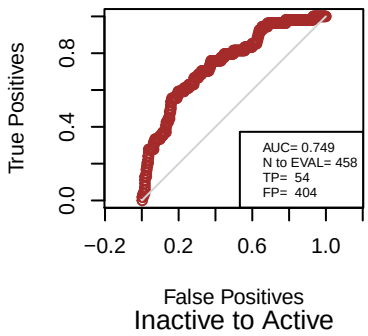

Growing to Developed



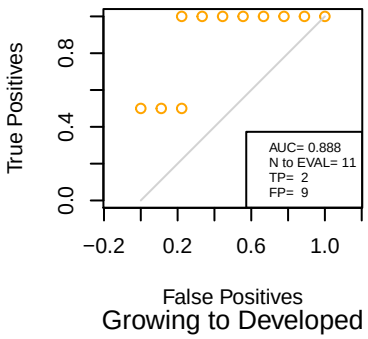

Undeveloped to Developed

Philippines

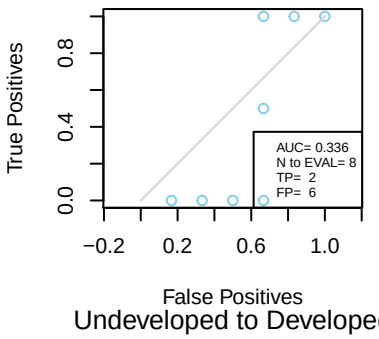
2008_2010 − 2011_2013 | RS vs UCSD



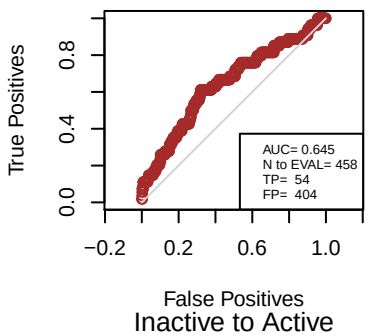

Inactive to Active



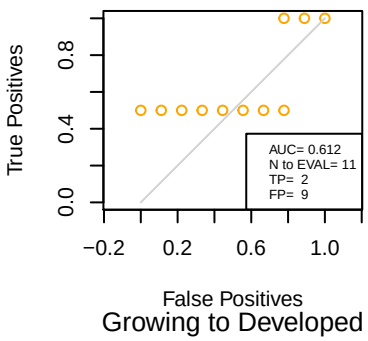

Growing to Developed



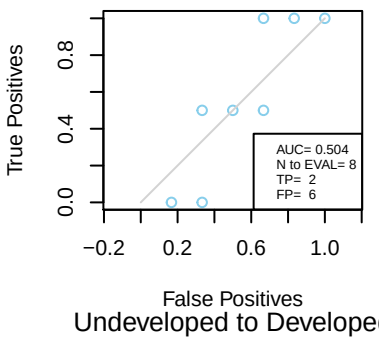

Undeveloped to Developed




Inactive to Active




Growing to Developed



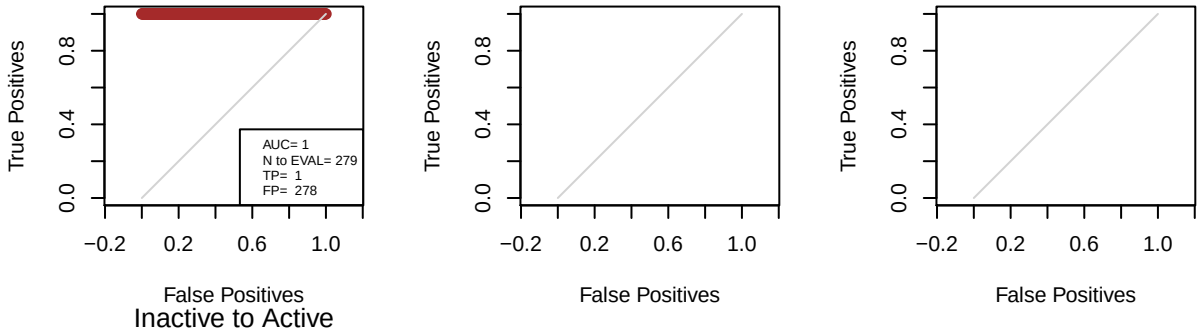

Undeveloped to Developed

Apéndice I

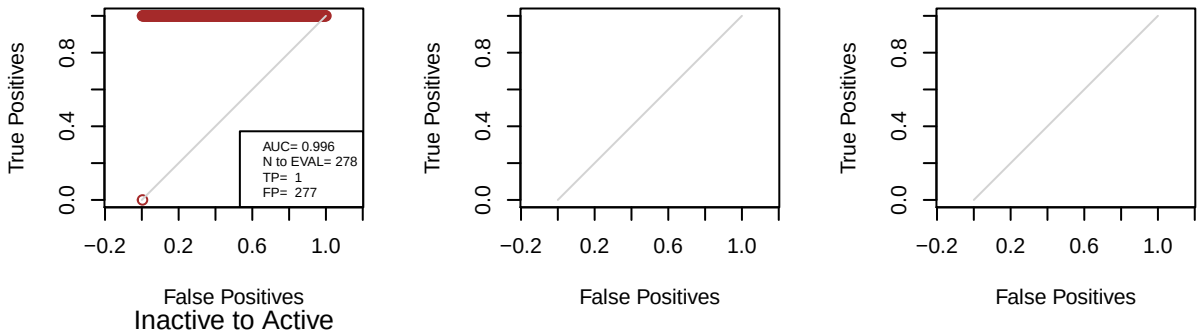
# Comparación de curvas ROC para individuos entre el espacio investigación y el mapa SCImago

Shi Jin

2008_2010 − 2011_2013 | RS vs SCIMAGO



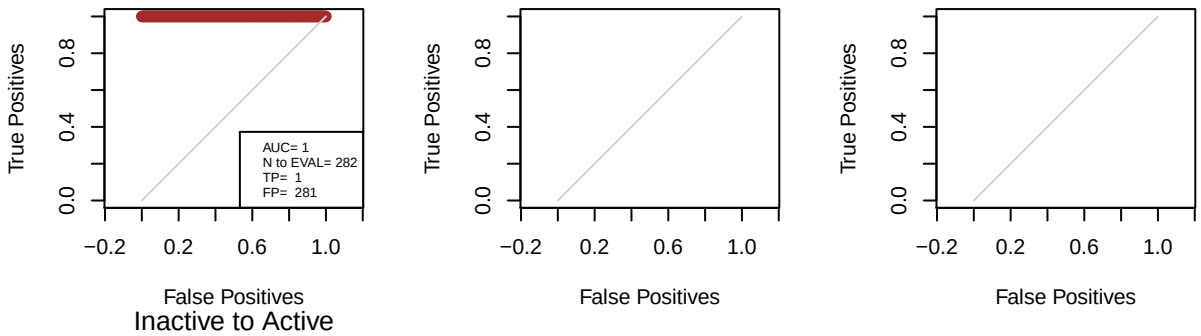

Inactive to Active



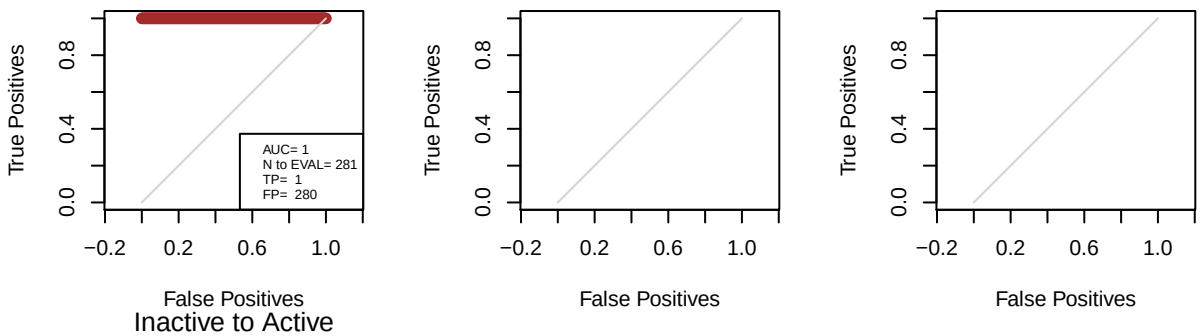

Inactive to Active

Nader Masmoudi

2008_2010 − 2011_2013 | RS vs SCIMAGO



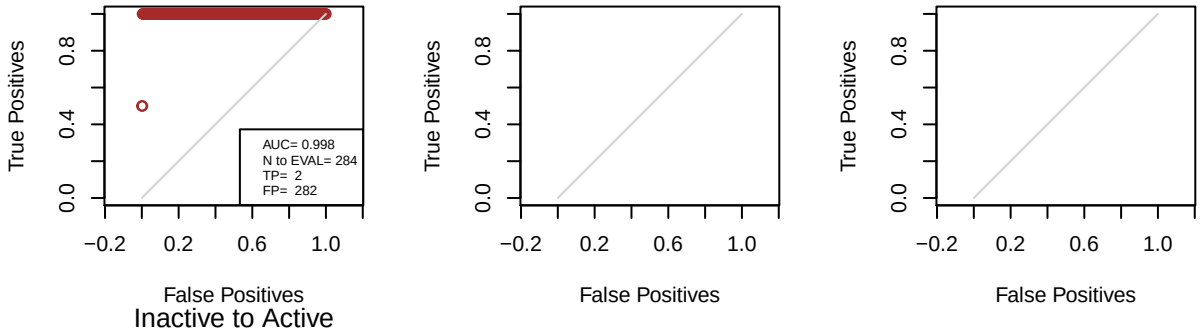

Inactive to Active



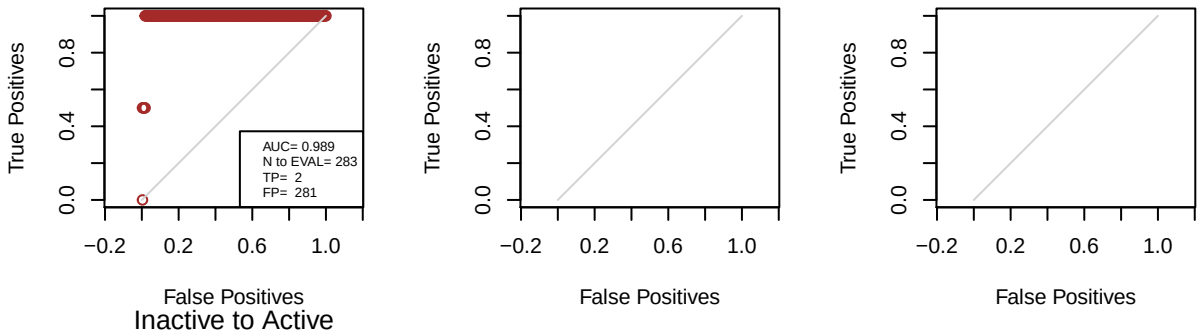

Inactive to Active

Herman Aguinis

2008_2010 − 2011_2013 | RS vs SCIMAGO



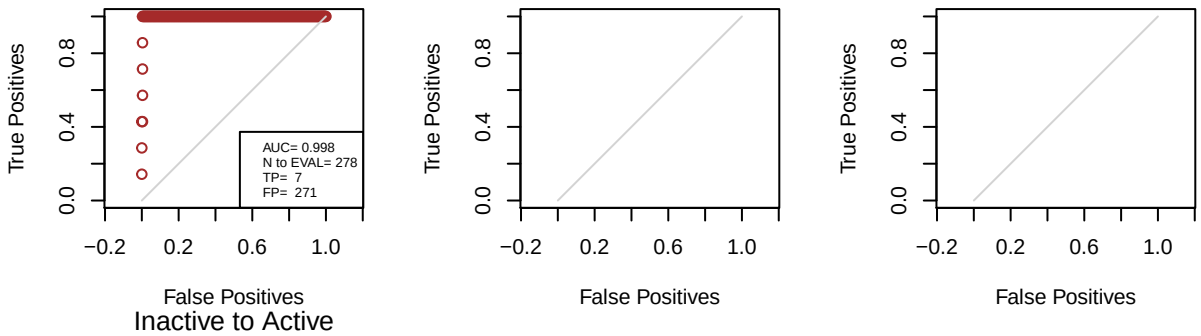

Inactive to Active



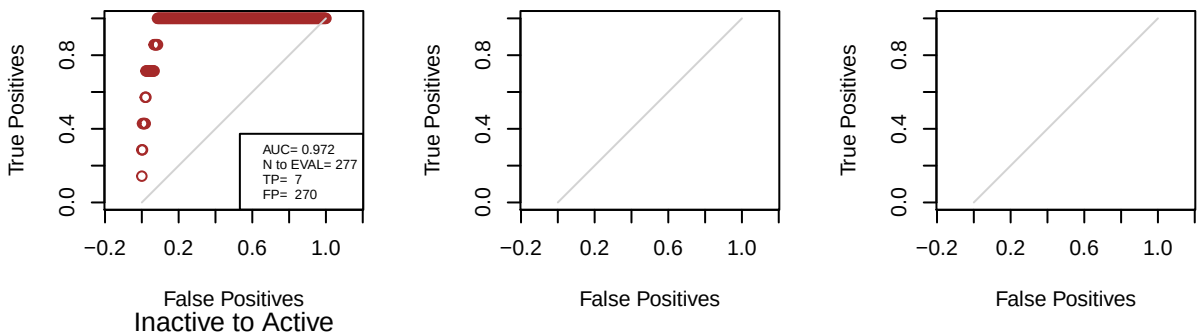

Inactive to Active

James W. Tanaka

2008_2010 − 2011_2013 | RS vs SCIMAGO



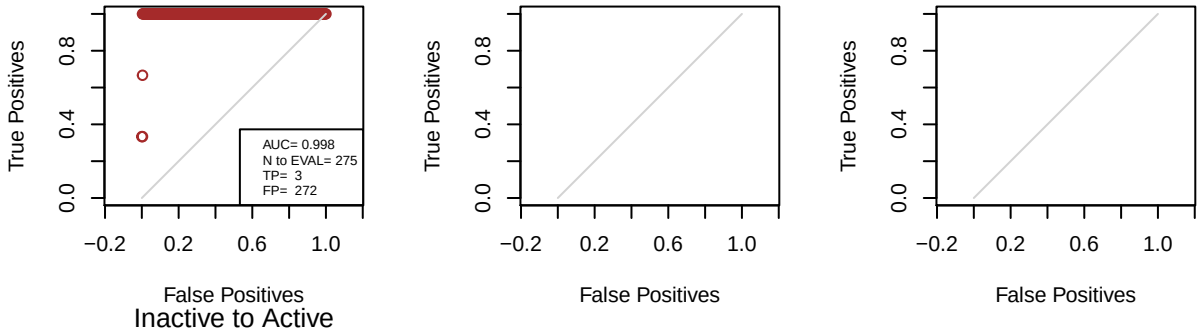

Inactive to Active



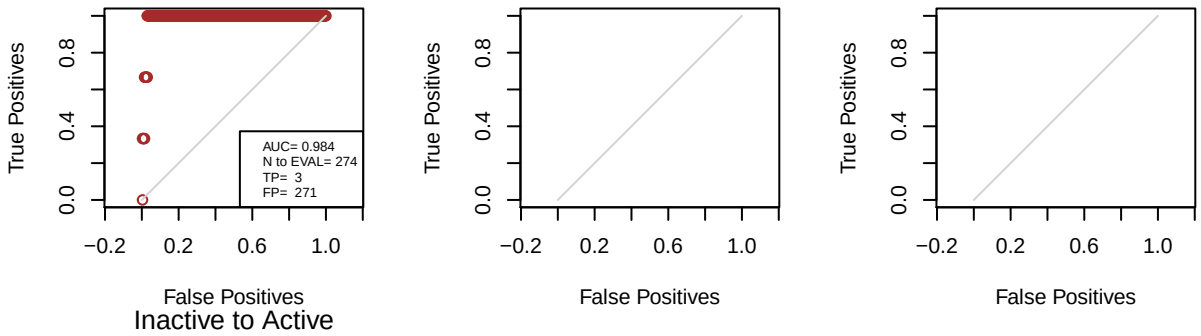

Inactive to Active

Giancarlo Salviati

2008_2010 − 2011_2013 | RS vs SCIMAGO



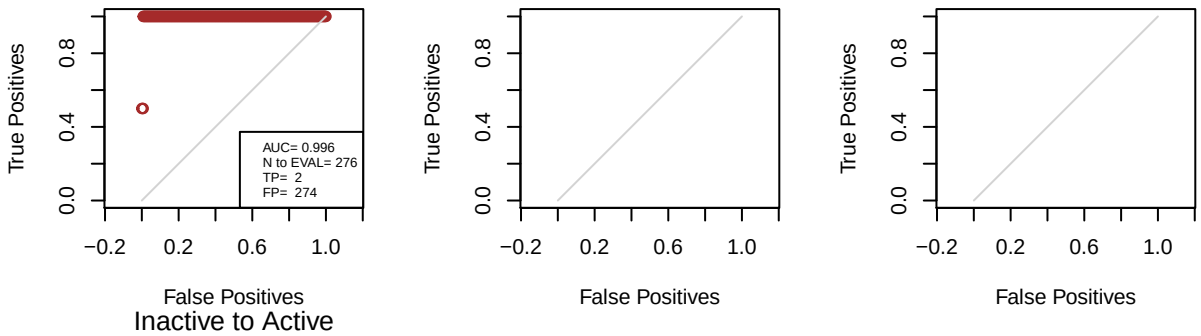

Inactive to Active



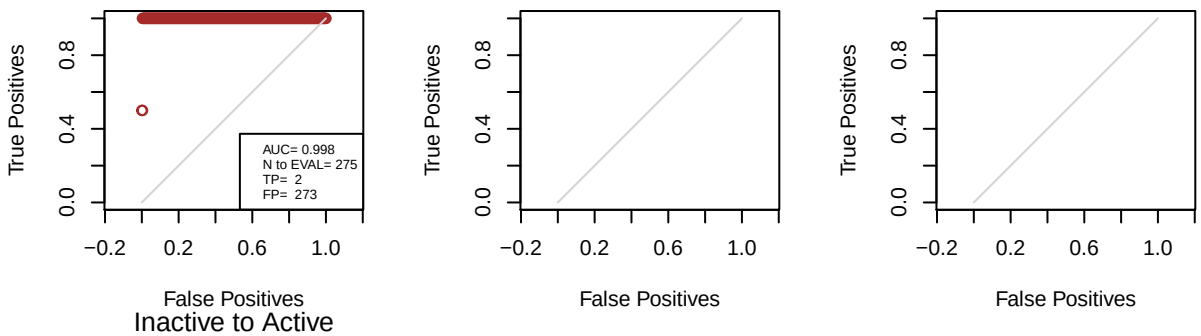

Inactive to Active

Alireza Dolatshahi−Pirouz

2008_2010 − 2011_2013 | RS vs SCIMAGO



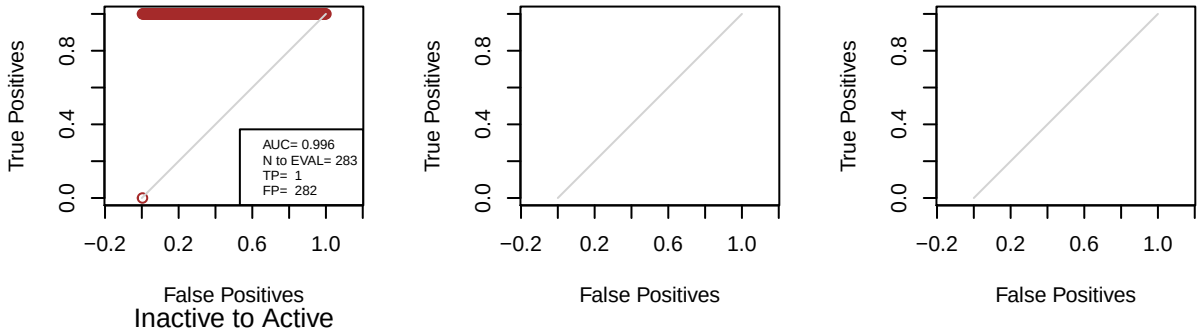

Inactive to Active

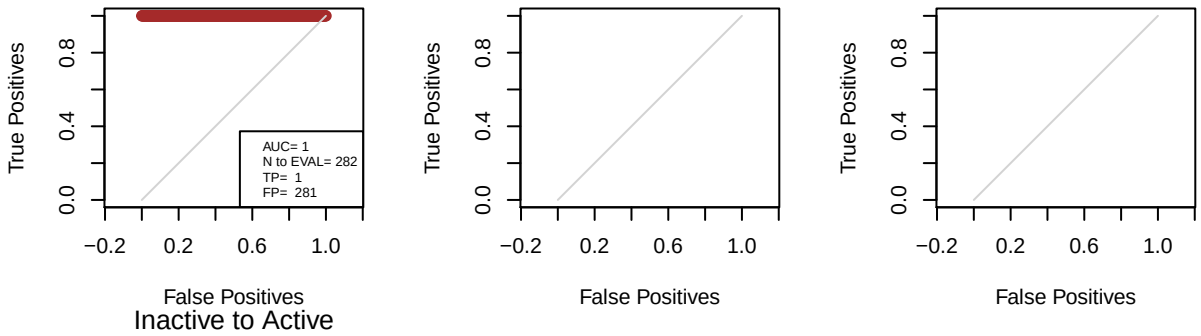



Inactive to Active

Liping Zhu

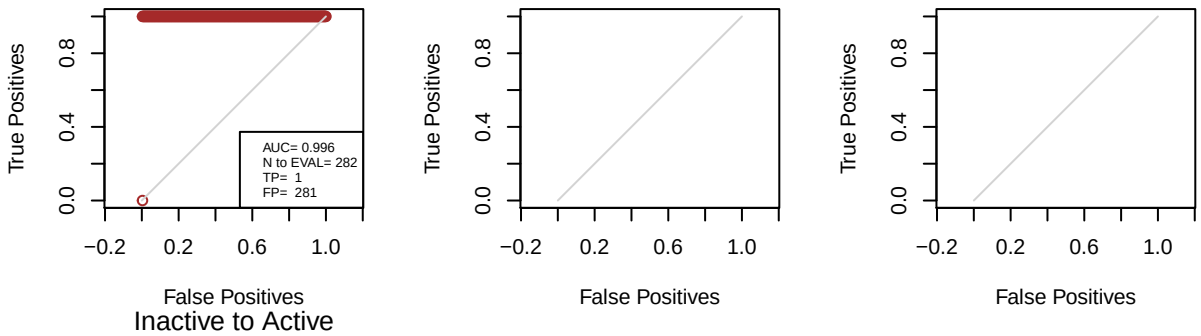
2008_2010 − 2011_2013 | RS vs SCIMAGO

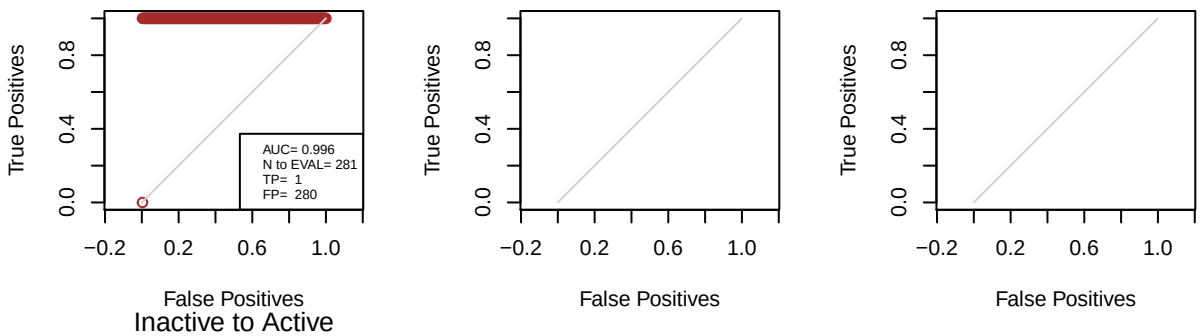



Alok Satapathy

2008_2010 − 2011_2013 | RS vs SCIMAGO

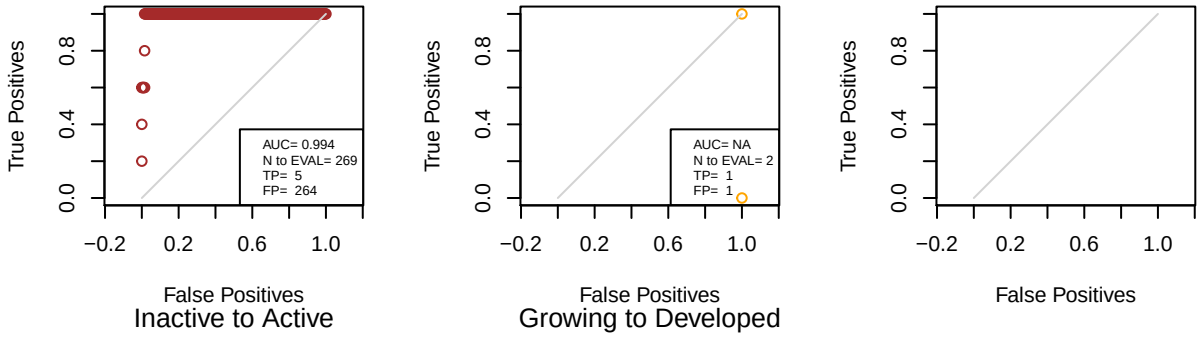


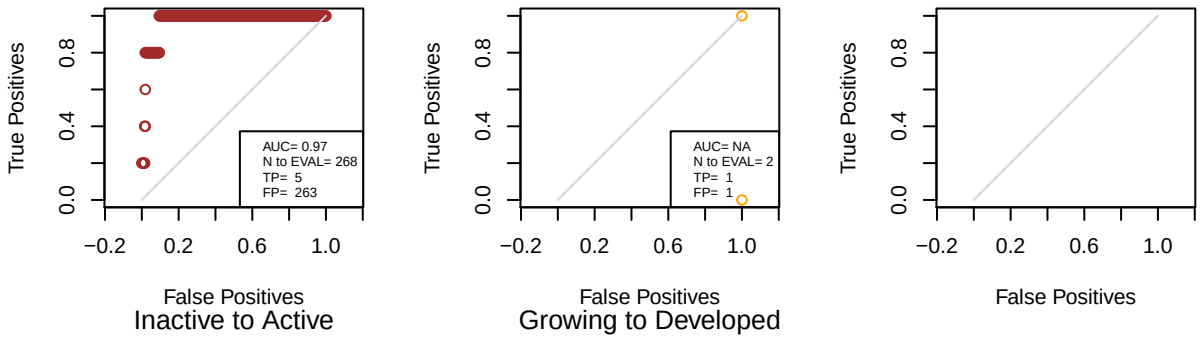



Apéndice J

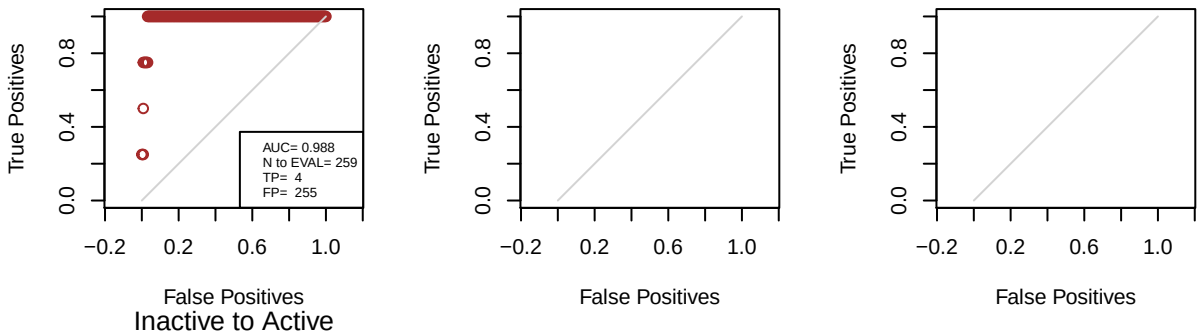
# Comparación de curvas ROC para instituciones entre el espacio investigación y el mapa SCImago

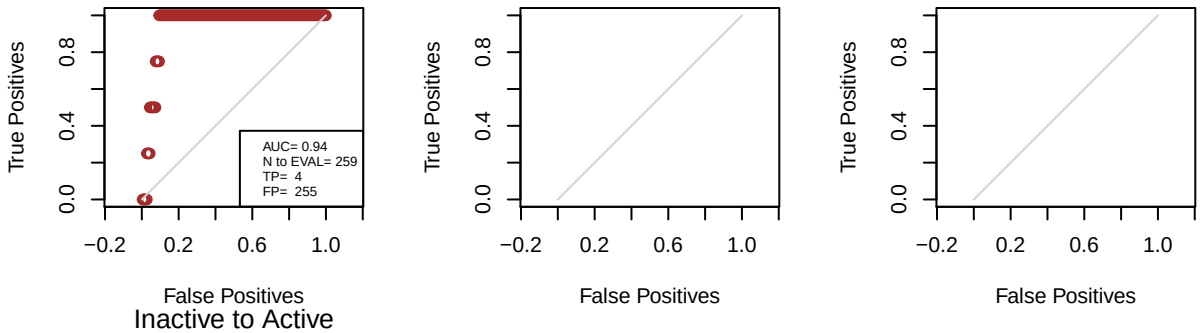
National Institute for Laser, Plasma and Radiation Physics
2008_2010 − 2011_2013 | RS vs SCIMAGO



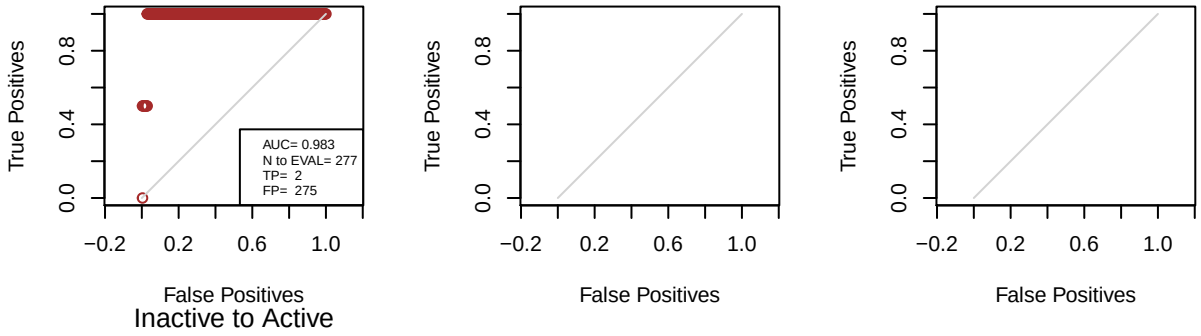



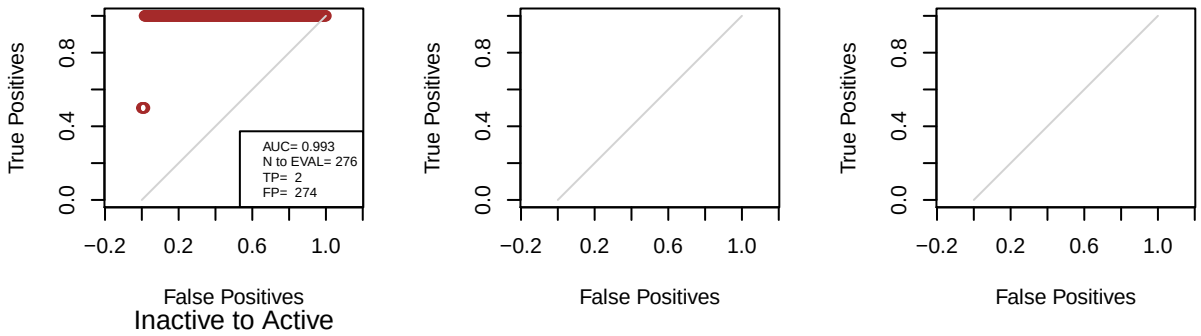

Gaziantep University
2008_2010 − 2011_2013 | RS vs SCIMAGO



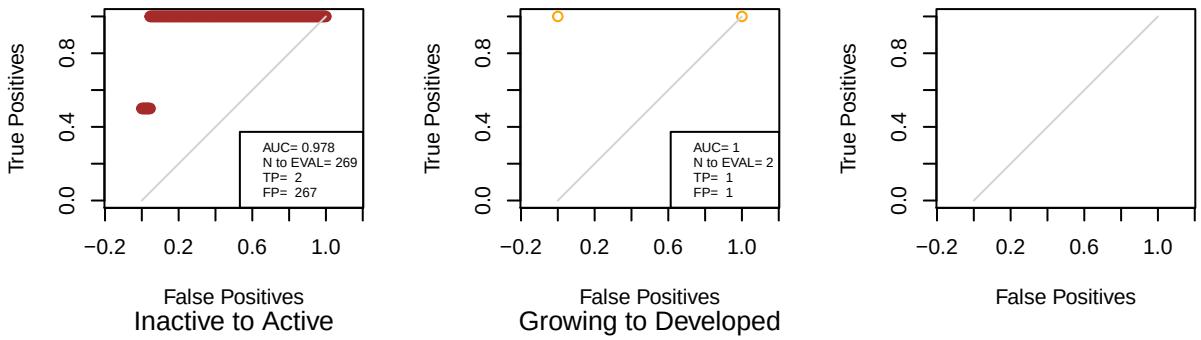



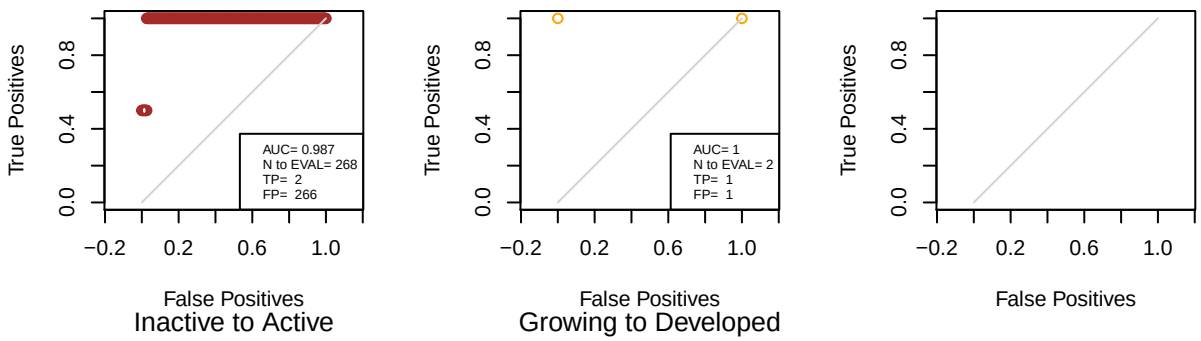
Istituto di Metodologie Inorganiche e dei Plasmi
2008_2010 − 2011_2013 | RS vs SCIMAGO



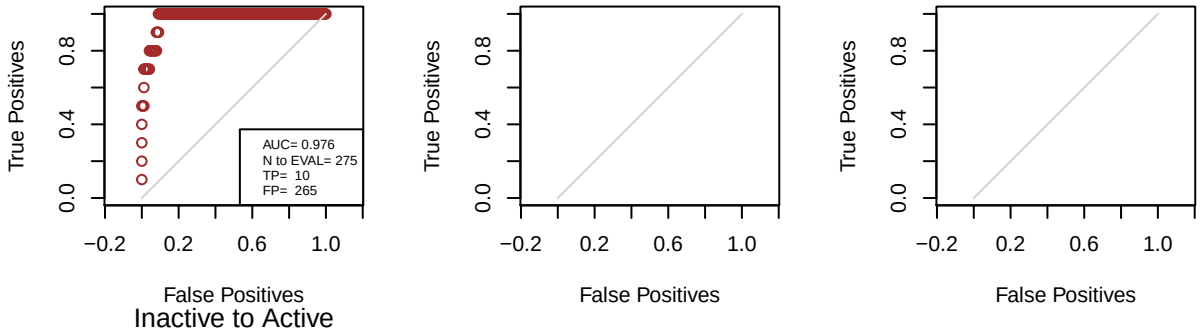

Geological Survey of Norway
2008_2010 − 2011_2013 | RS vs SCIMAGO

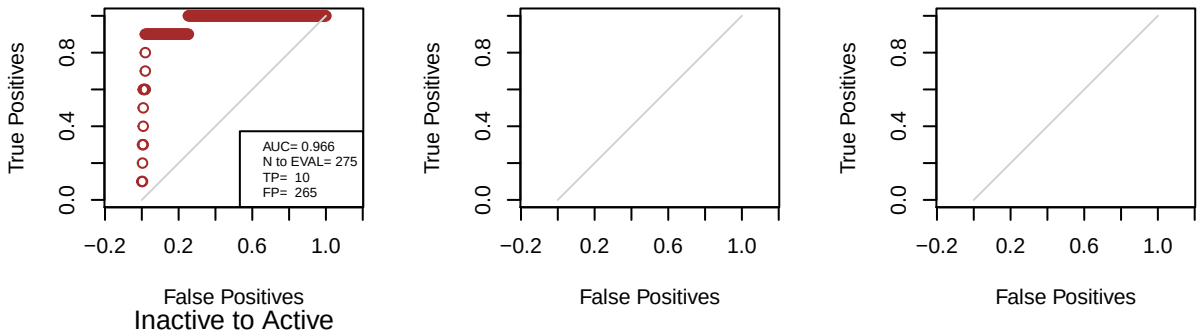


hai Institute of Microsystem and Information Technology CAS / .....

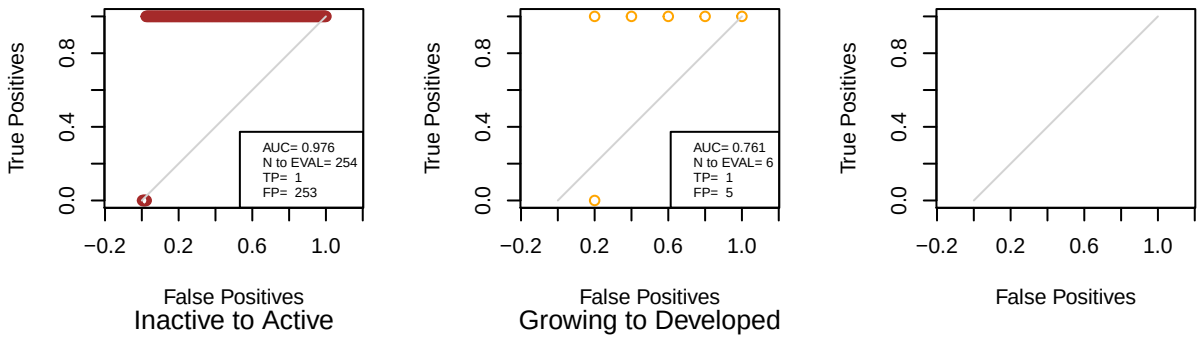
2008_2010 − 2011_2013 | RS vs SCIMAGO

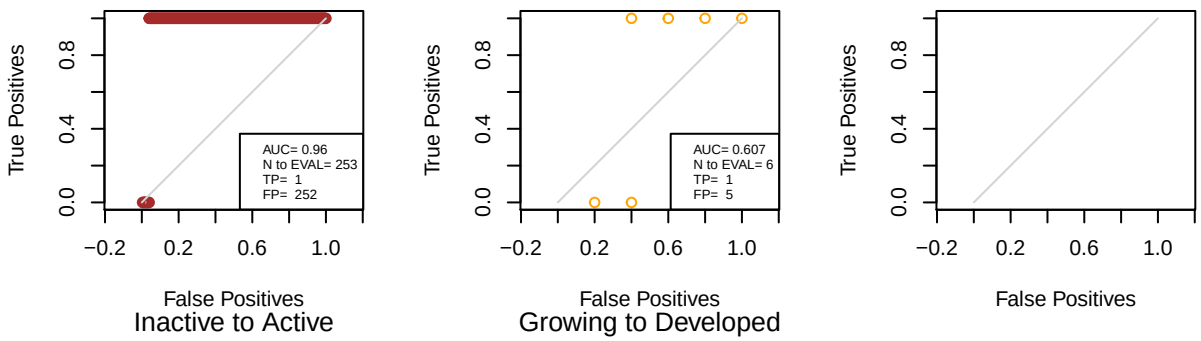



Alfa Institute of Biomedical Sciences

2008_2010 − 2011_2013 | RS vs SCIMAGO



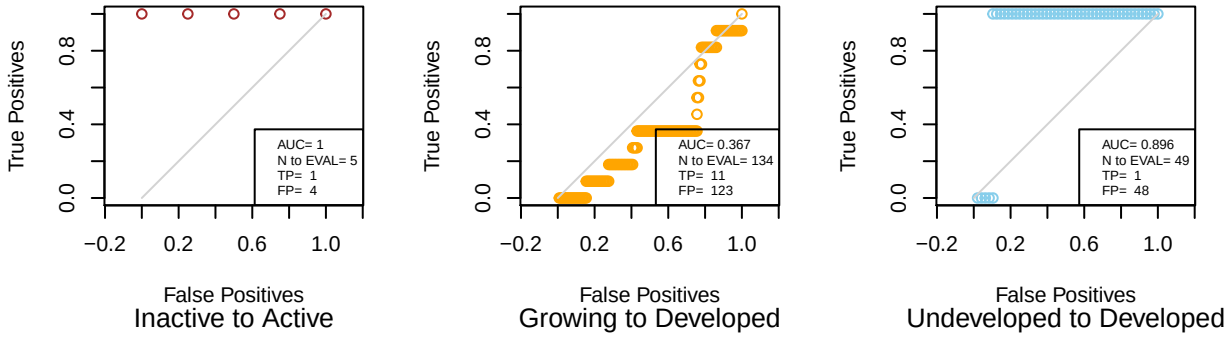



Apéndice K

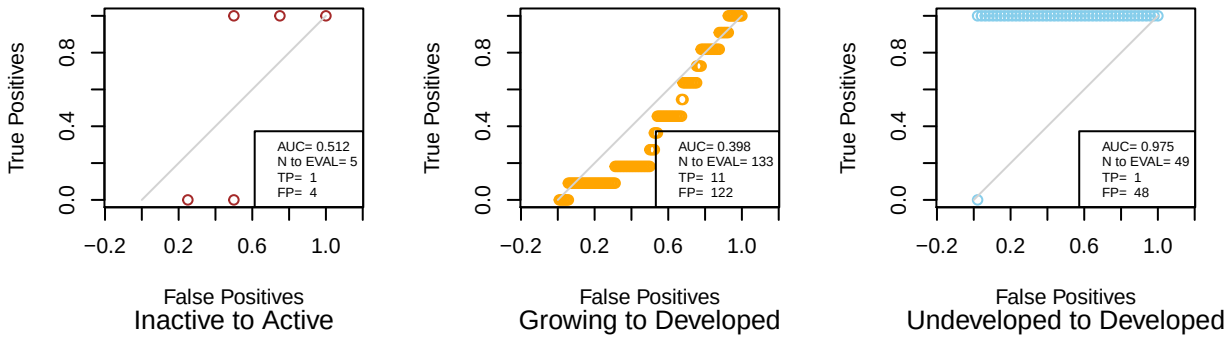
# Comparación de curvas ROC para países entre el espacio investigación y el mapa SCImago

Brazil

2008_2010 − 2011_2013 | RS vs SCIMAGO



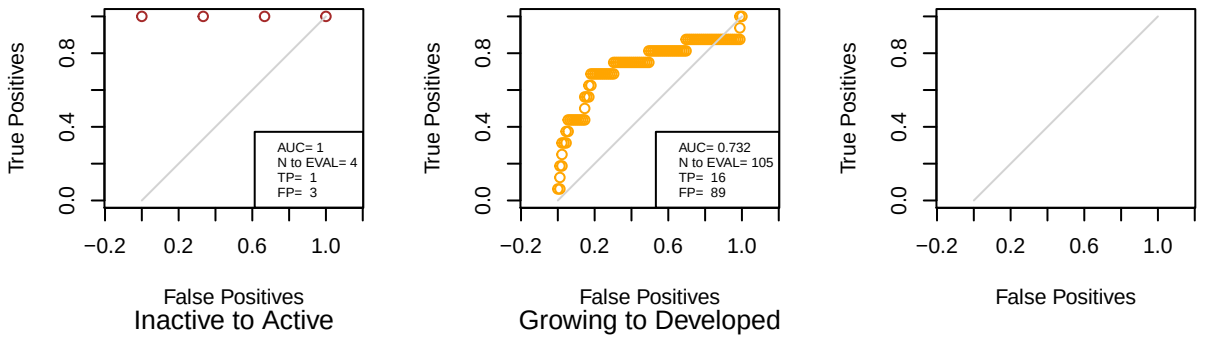

Inactive to Active | Growing to Developed | Undeveloped to Developed



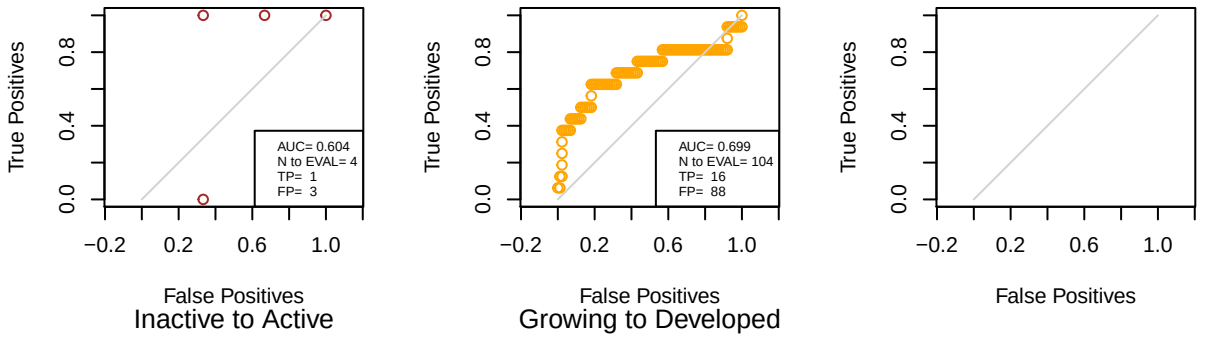

Inactive to Active | Growing to Developed | Undeveloped to Developed

Japan

2008_2010 − 2011_2013 | RS vs SCIMAGO



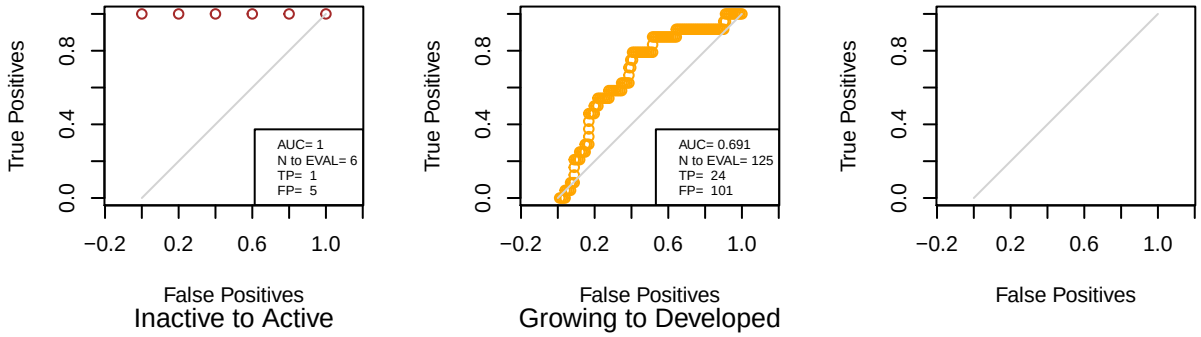

Inactive to Active | Growing to Developed



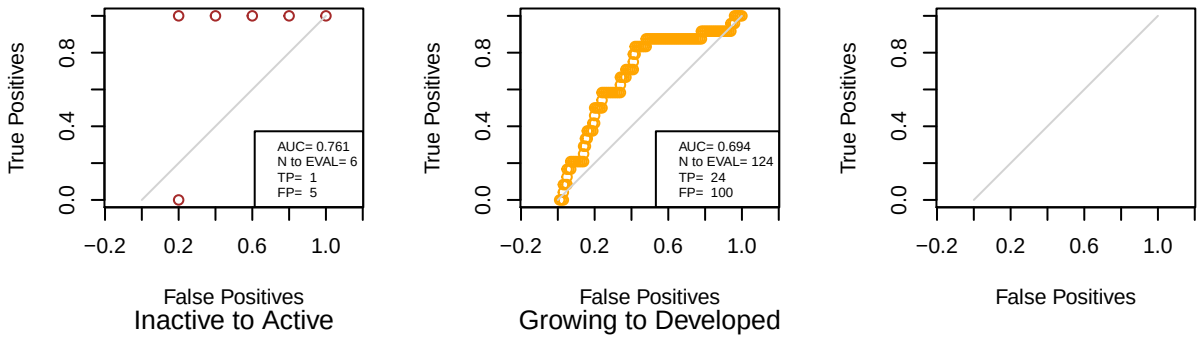

Inactive to Active | Growing to Developed

Finland

2008_2010 − 2011_2013 | RS vs SCIMAGO



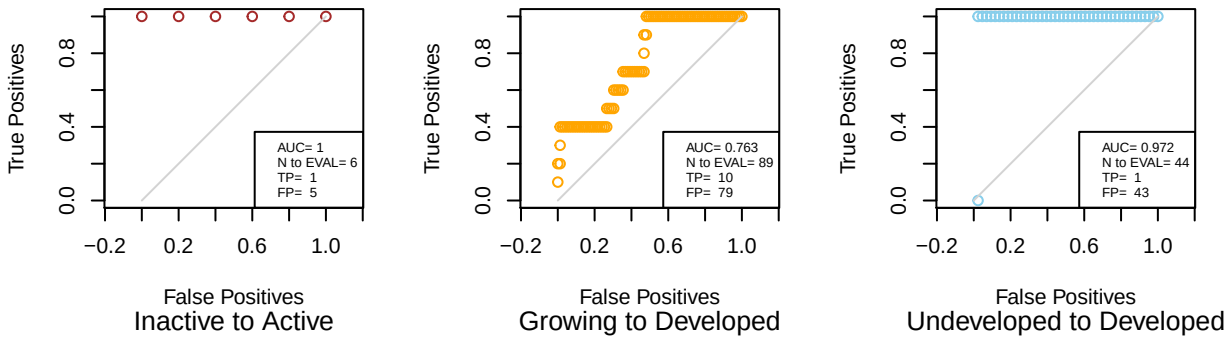

Inactive to Active | Growing to Developed



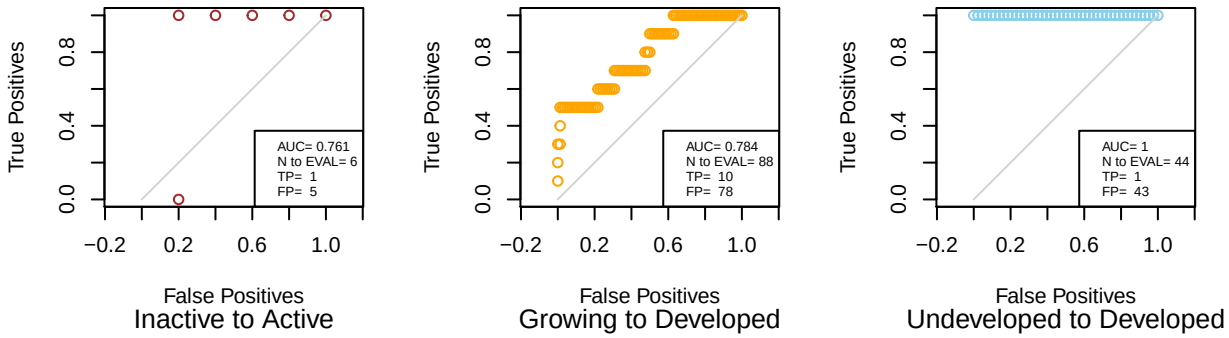

Inactive to Active | Growing to Developed

New Zealand

2008_2010 − 2011_2013 | RS vs SCIMAGO



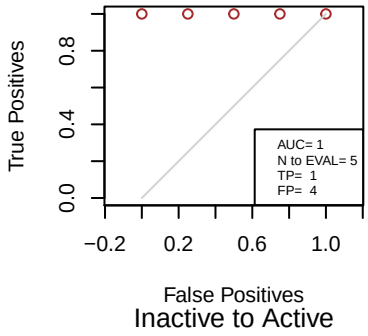

Inactive to Active | Growing to Developed | Undeveloped to Developed



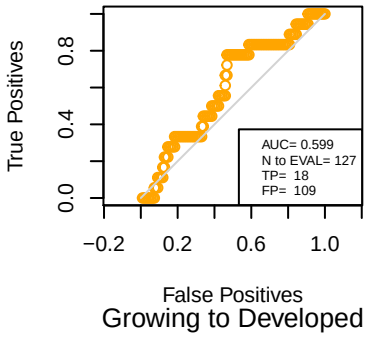

Inactive to Active | Growing to Developed | Undeveloped to Developed

France

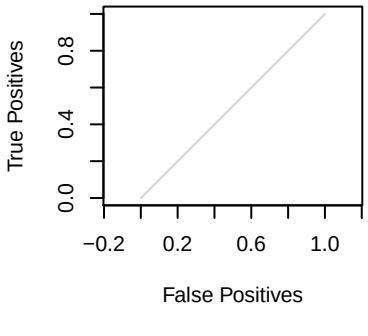
2008_2010 − 2011_2013 | RS vs SCIMAGO

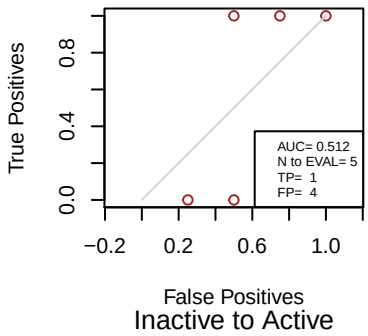



Inactive to Active

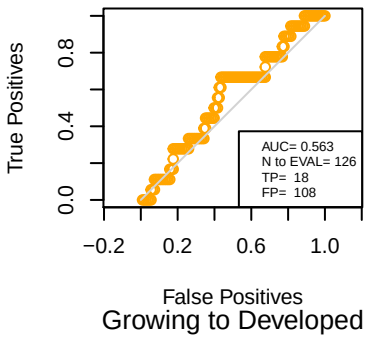



Growing to Developed

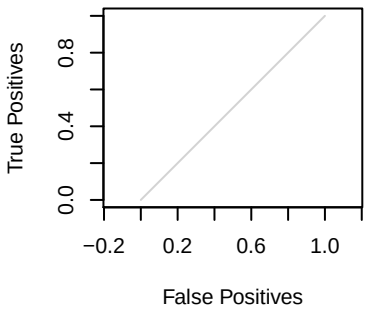



Inactive to Active

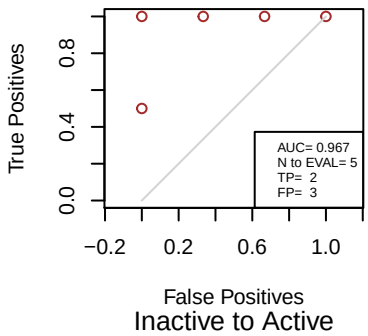



Growing to Developed

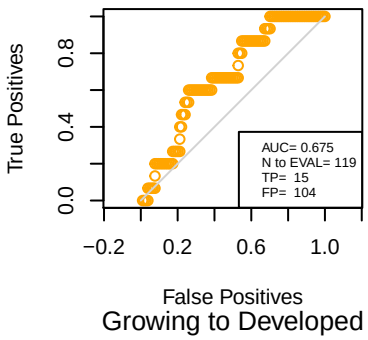



Denmark

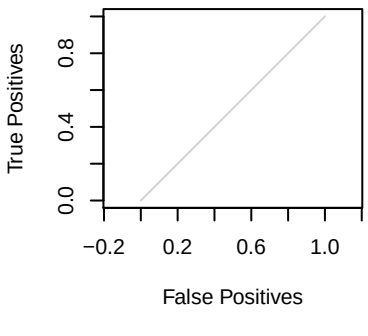
2008_2010 − 2011_2013 | RS vs SCIMAGO

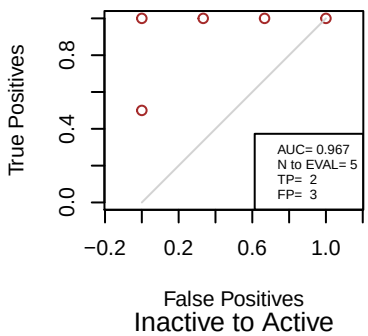



Inactive to Active

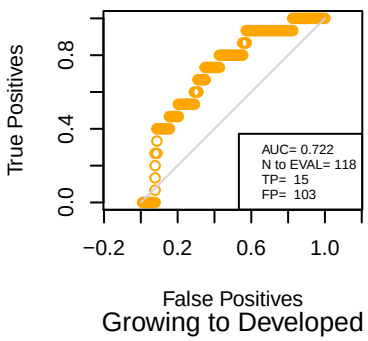



Growing to Developed

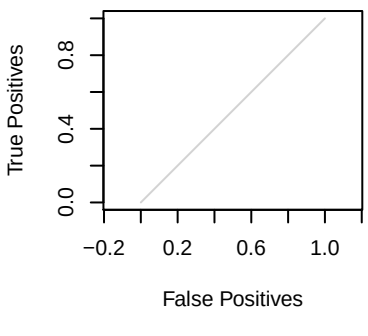



Inactive to Active




Growing to Developed



# Glosario de términos y definiciones

**Árbol Recubridor Mínimo (MST)** Árbol que recubre al grafo, en el que la suma de sus pesos (considerados como distancias) es la mínima. En inglés se denomina *Minimum Spanning Tree en inglés*. 68, 69, 116

**índice-H** Índice que propone la calidad científica de un investigador como un balance entre autoría de *papers* y las citas ganadas por esos *papers*. El algoritmo para calcularlo, es el siguiente: se toman todos los *papers* de un científico y se ordenan de forma descendiente según el número de citas. Después se asigna un número de ranking ascendente a cada paper de la lista ordenada. Luego, se desciende en la lista mientras el número del ranking del paper es menor o igual al número de citas ganadas por ese paper. Al llegar a la condición de término, se ha obtenido el índice-H que corresponde al último número de ranking que cumplía la condición evaluada. Por ejemplo un índice-H de 5 implica que un investigador a sido coautor de *journal*5 *papers* que han ganado al menos 5 citas. Este índice, ha tenido amplia difusión puesto que ayuda a reducir la hiperproducción de *papers* de baja calidad que terminan por no ser citados. 10, 13

**API** Application Program Interface. 58

**Arts & Humanities Citation Index® (AHCI)** Indice de Thomson Reuters que indexa *journals* en las áreas de las ciencias sociales. Este índice está incluido en el *core collection* de WoS. 45

**CCS** ACM Computing Classification System. 24

**CERN** European Organization for Nuclear Research. 12

**clique** Tipo especial de grafo en el que todos los nodos están conectados entre sí. 68

**CONICYT** Comisión Nacional de Investigación Científica y Tecnológica. 9, 14, 15

**CRAN** The Comprehensive R Archive Network. 95

**DBLP** Computer Science Bibliography. 8

**eigenfactor** Índice que mide la importancia de un determinado autor al interior de una comunidad científica. Esta importancia se mide en base a la centralidad del autor en la red de coautoría de la comunidad. Para esto se aplican técnicas basadas principalmente en vectores propios, similar a la idea del algoritmo de PageRank de Google. Este índice se incluye como indicador en la base de datos que indexa *papers* del área de las ciencias sociales Social Science Research Network SSRN. 13

**enlace** También denominada en inglés *link* o arista, es un elemento que pertenece al conjunto de enlaces $E$ de un grafo $G$. Los enlaces pueden tener diversas propiedades, entre ellas la más común es el *peso*. 5

**ERA** Excellence in Research for Australia. 115

**Factor de Impacto** Índice que mide la calidad de un *journal*, a través de la ponderación de las citas recibidas por sus artículos en los últimos dos años. Se calcula con el siguiente algoritmo: Para cada *journal* en la lista de un año Y, sumar las citas acumuladas por cada paper en ese *journal* en los años Y-1, Y-2. Dividir este valor por la cantidad de documentos citables publicados por ese *journal* durante Y-1, Y-2. Nótese que las citas acumuladas se consideran sobre *journals* en el mismo sistema de indexación. El resumen de factores de impacto por *journal*, se publica por Thomson Reuters en un informe conocido como *Journal Citation Report JCR*. 13, 17, 18

**FoR** Field of Research. 49, 50, 115

**FOS** Field of Science. 11

**Google Scholar** Servicio provisto por Google, que permite la búsqueda de científicos y de sus publicaciones. GS indexa información científica en todas las áreas del conocimiento y provee indicadores de citación para individuos. https://scholar.google.cl/. 58, 62, 64

**grado** Suma de los enlaces que conectan un nodo. En inglés, *degree*. El grado ponderado, es la suma de los enlaces que conectan el nodo, pero ponderados por alguna característica de los enlaces, generalmente el peso. 68

**grafo** Un grafo $G = V, E$ está compuesto por un conjunto $V$ de nodos o vértices y por un conjunto $E$ de enlaces o aristas, que relacionan los nodos. Los grafos pueden ser dirigidos, si los enlaces entregan una noción de direccionalidad o no dirigidos en caso contrario. 5

**ISI** Institute for Scientific Information. 42

**journal** Revista científica. En este documento se utiliza la palabra inglesa o la traducción al español de forma indistinta plural. 6, 8–11, 13, 17, 18, 23, 24, 28, 29, 31, 42, 44–47, 51, 52, 54, 59, 60, 62–64, 66, 70, 71, 96, 115, 203–206

**Journal Citations Report® (JCR)** Servicio provisto por la empresa Thomson Reuters™que clasifica y evalúa la calidad de *journals* según índices bibliométricos, principalmente el Factor de Impacto. 17, 18, 45, 48

**Katy Börner** Profesora de Ciencias de la Información en la Universidad de Indiana en Estados Unidos. Es reconocida por ser uno de los científicos que más ha aportado al desarrollo moderno de los mapas de la ciencia. 58, 59

**layout** Algoritmo de visualización de grafos, que determina la forma en que se visualizan los nodos y enlaces. 69

**log** Archivo de registro que se crea en base a las acciones de personas u otros programas.. 32

**mapa UCSD** Mapa de la ciencia UCSD, creado por Katy Börner y colaboradores en el año 2005 en atención a la solicitud de la Universidad de California en San Diego (UCSD) de donde proviene el nombre. El mapa fue actualizado y los datos liberados en 2012. Este mapa propone también una clasificación de la ciencia basada en clusters de *journals*. 22–24, 48, 85, 86

**Mendeley** Herramienta para buscar y organizar información científica. Orientada a estudiantes e investigadores. 32

**metadata** Se define tradicionalmente como los datos de los datos. Vale decir, datos que proveen información detallada o adicional de los datos en estudio. 11

**MIT** Massachusetts Institute of Technology. 103, 107

**nodo** También denominado vértice, (en inglés *node*), es un elemento que pertenece al conjunto de vértices $V$ de un grafo $G$. Los nodos pueden tener diversas propiedades, entre ellas la más común es el *grado*. 5

**OECD** Organization for Economic Co-operation and Development. 11, 115

**paper** Trabajo científico publicado en una revista científica o *journal*. 6, 9–17, 21, 24, 27–30, 37, 40, 42, 64–66, 70, 75, 103, 106, 110, 113, 203, 204

**peer review** Revisión por pares. Se dice del acto por el cual un trabajo científico se revisa o evalúa por parte de pares o personas de la misma área disciplinar. 6

**proceeding** Actas de un congreso o conferencia. 6

**PubMed** Public domain information on the National Library of Medicine. 8

**R** Software para computación estadística y gráficos. 28, 95

**RCA** Revealed Comparative Advantajes. 23, 24, 76, 77, 96, 97, 113

**Recuperación de Información** Area de las ciencias de la información, encargada de estudiar los métodos y algoritmos utilizados para almacenar, recuperar y rankear información, según criterios de búsqueda del usuario. Si bien sus orígenes datan de 1970, se ha popularizado debido a su aplicación a los servidores de búsqueda en línea como Yahoo! o Google. 32, 33

**ROC** Nombre con el que se conoce a las curvas ROC por sus siglas en inglés *Receiver Operating Characteristic* que es un gráfico que ilustra el rendimiento de un clasificador binario y que puede extrapolarse al rendimiento de una predicción basada en ranking. Esta curva presenta en el eje X la tasa de falsos positivos mientras que en el eje Y la tasa de verdaderos positivos. 84

**SciELO** *Scientific Electronic Library*, es una biblioteca digital que indexa *journals* principalmente de Iberoamérica. Tiene contrapartes locales en la mayoría de países. Es una corporación sin fines de lucro y se inició originalmente en Brasil en 1998. 8

**Science Index ™(SCCI)** Indice de la empresa Thomson Reuters que indexa *journals* en las áreas de las ciencias naturales y exactas. Este índice fue desactualizado y actualmente se utiliza el Science Index Expanded. 45

**SCImago** Nombre con el que usualmente se denomina al sitio web SCImagoJR, SCImago Journal & Country Ranking, que es parte de la organización SCImagoLabs liderada en sus inicios por Félix Moya de Anegón y la Universidad de Granada en España. El sitio SCImagoJR, disponibiliza de forma gratuita, rankings y visualizaciones de journals y países, con datos provistos por Scopus. 11, 23, 62, 64–66, 97, 109

**SCImago Journal & Country Ranking (SJR)** Portal web que incluye indicadores para *journals* y países con información contenida en la base de datos Scopus de Elsevier. http://www.scimagojr.com/. 8, 35

**Scopus®** Base de datos bibliográfica de acceso pagado bajo suscripción y mantenida por la editorial Elsevier. 8, 11, 35, 42, 51, 62

**scraper** Nombre con el que se suele conocer a códigos de programación que tienen como objetivo extraer y almacenar información desde la Web. Uno de los software más utilizados es Scrapy que se encuentra programado en lenguaje Python. 58, 60, 62

**SITC** Standard International Trade Classification. 11

**Social Science Index® (SSCI)** Indice de Thomson Reuters que indexa *journals* en las áreas de las ciencias sociales. Este índice está incluido en el *core collection* de WoS. 45

**SSRN** Social Science Research Network. 8

**Thomson Reuters™** Corporación cuyo foco de negocio se encuentra en la información. Adquirieron en 1992 el Instituto para la Información Científica ISI fundado en 1960 por Eugene Garfield. En el ámbito de publicaciones científicas se suele referir a Thomson

Reuters o ISI-Thomson Reuters como la base de datos que indexa publicaciones y revistas científicas. En la actualidad este servicio se provee a través del sitio Web Of Science$^{TM}$WoS, pero que también se conoce como Web of Knowledge WoK, por cuanto el URL de este servicio es webofknowledge.com. 8, 11, 15, 17, 18, 45, 48, 51, 115

**TOS** Términos de Servicio. 58

**treemap** Conocido también como *rectangular treemap* o árbol rectangular, es una forma de visualización de rectángulos anidados, que considera en cada rectángulo una rama, dentro de la cual se insertan otros cuadrados que corresponden a sus ramas de nivel jerárquico inferior, hasta llegar a sus hojas. Este tipo de visualización se popularizó por Ben Shneiderman en la década de los 90, quien introdujo la idea de recursividad. Estas visualizaciones son ampliamente utilizadas por cuanto utilizan el espacio óptimamente lo que permite representar mucha información sin perder calidad. 103, 104

**Turing Award** Premio conocido como el equivalente al Nobel para el área de computación. Se entrega desde al año 1966 a un científico destacado en el área de la computación. El premio lleva su nombre en honor a Alan Turing que es reconocido como el padre de la computación moderna. Desde el año 2014, con el auspicio de Google, el premio entregado es de un millón de dólares americanos. 12

**UCSD** University of California, San Diego. 42, 64, 66

**UPLA** Universidad de Playa Ancha. 15

**web scraping** Técnica computacional para extraer y estructurar información de sitios web, de forma automática. Generalmente utilizada cuando el sitio de interés no provee de un API para acceso a sus datos. 58, 62

**WoS** Web of Science$^{TM}$. 8, 11, 15, 42, 52, 62

**WWW** World Wide Web. 12, 13

# Bibliografía


Giovanni Abramo and Ciriaco Andrea D'Angelo. Assessing national strengths and weaknesses in research fields. *Journal of Informetrics*, 8(3):766–775, July 2014. ISSN 1751-1577. doi: 10.1016/j.joi.2014.07.002. URL http://www.sciencedirect.com/science/article/pii/S1751157714000625.

Jonathan Adams. Collaborations: The rise of research networks. *Nature*, 490(7420):335–336, October 2012. ISSN 0028-0836. doi: 10.1038/490335a. URL http://www.nature.com/nature/journal/v490/n7420/full/490335a.html.

Jonathan Adams. The brain scan of research impact | HEFCE blog, July 2015. URL http://blog.hefce.ac.uk/2015/07/20/the-brain-scan-of-research-impact/.

Altmetric. What Does Altmetric Do? - Altmetric, 2015. URL http://www.altmetric.com/whatwedo.php.

Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Improving search engines by query clustering. *Journal of the American Society for Information Science and Technology*, 58(12):1793–1804, October 2007. ISSN 1532-2890. doi: 10.1002/asi.20627. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.20627/abstract.

Bela Balassa. Trade Liberalisation and "Revealed" Comparative Advantage1. *The Manchester School*, 33(2):99–123, 1965. ISSN 1467-9957. doi: 10.1111/j.1467-9957.1965.tb00050.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9957.1965.tb00050.x/abstract.

Albert-László Barabási. *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, April 2003. ISBN 0-452-28439-2.

Albert-László Barabási. *Network Science*. Cambridge University Press, Cambridge, United Kingdom, April 2016. ISBN 978-1-107-07626-6.

Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, reprint edition edition, November 2012. ISBN 978-1-107-62625-6.

Wolfgang H. Berger and Frances L. Parker. Diversity of Planktonic Foraminifera in Deep-Sea Sediments. *Science*, 168(3937):1345–1347, December 1970. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.168.3937.1345. URL http://www.sciencemag.org/content/168/3937/1345.

John Desmond Bernal. *The social function of Science*. George Routledge, 1939.

Peter Michael Blau. *Inequality and heterogeneity: A primitive theory of social structure*, volume 7. Free Press New York, 1977.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning research*, 3(Jan):993–1022, 2003.

Johan Bollen, Herbert Van de Sompel, Aric Hagberg, Luis Bettencourt, Ryan Chute, Marko A. Rodriguez, and Lyudmila Balakireva. Clickstream Data Yields High-Resolution Maps of Science. *PLoS ONE*, 4(3):e4803, March 2009. doi: 10.1371/journal.pone.0004803. URL http://dx.doi.org/10.1371/journal.pone.0004803.

Kevin Boyack. Using detailed maps of science to identify potential collaborations. *Scientometrics*, 79(1):27–44, 2009. ISSN 0138-9130. URL http://dx.doi.org/10.1007/s11192-009-0402-6. 10.1007/s11192-009-0402-6.

Kevin W. Boyack and Richard Klavans. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12):2389–2404, December 2010. ISSN 1532-2890. doi: 10.1002/asi.21419. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.21419/abstract.

Kevin W. Boyack and Richard Klavans. Creation of a Highly Detailed, Dynamic, Global Model and Map of Science. *Journal of the Association for Information Science & Technology*, 65(4):670–685, April 2014. ISSN 23301635. doi: 10.1002/asi.22990.

Kevin W. Boyack, Richard Klavans, and Katy Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, August 2005. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-005-0255-6. URL http://link.springer.com/article/10.1007/s11192-005-0255-6.

Katy Börner. *Atlas of Science: Visualizing What We Know*. The MIT Press, September 2010. ISBN 0-262-01445-9.

Katy Börner. *Atlas of Knowledge: Anyone Can Map*. The MIT Press, March 2015. ISBN 978-0-262-02881-3.

Katy Börner, Kevin W. Boyack, Staša Milojević, and Steven Morris. An Introduction to Modeling Science: Basic Model Types, Key Definitions, and a General Framework for the Comparison of Process Models. In Andrea Scharnhorst, Katy Börner, and

Peter van den Besselaar, editors, *Models of Science Dynamics*, Understanding Complex Systems, pages 3–22. Springer Berlin Heidelberg, January 2012a. ISBN 978-3-642-23067-7 978-3-642-23068-4. URL http://link.springer.com/chapter/10.1007/978-3-642-23068-4_1.

Katy Börner, Richard Klavans, Michael Patek, Angela M. Zoss, Joseph R. Biberstine, Robert P. Light, Vincent Larivière, and Kevin W. Boyack. Design and Update of a Classification System: The UCSD Map of Science. *PLoS ONE*, 7(7):e39464, July 2012b. doi: 10.1371/journal.pone.0039464. URL http://dx.doi.org/10.1371/journal.pone.0039464.

Carlos Castillo. *Effective Web Crawling*. PhD thesis, Universidad de Chile, 2004. URL http://chato.cl/research/crawling_thesis.

Diego Chavarro, Puay Tang, and Ismael Rafols. Interdisciplinarity and research on local issues: evidence from a developing country. *Research Evaluation*, 23(3):195–209, 2014.

Chaomei Chen and Loet Leydesdorff. Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the Association for Information Science and Technology*, 65(2):334–351, February 2014. ISSN 2330-1643. doi: 10.1002/asi.22968. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.22968/abstract.

Rex H.-G. Chen and Chi-Ming Chen. Visualizing the world's scientific publications. *Journal of the Association for Information Science and Technology*, pages n/a–n/a, September 2015. ISSN 2330-1643. doi: 10.1002/asi.23591. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.23591/abstract.

Giulio Cimini, Andrea Gabrielli, and Francesco Sylos Labini. The Scientific Competitiveness of Nations. *PLoS ONE*, 9(12):e113470, December 2014. doi: 10.1371/journal.pone.0113470. URL http://dx.doi.org/10.1371/journal.pone.0113470.

John Owen Edward Clark. *100 maps: the science, art and politics of cartography throughout history*. Sterling Publishing Company, Inc., 2005.

CSAIL. Michael Stonebraker wins $1 million Turing Award, March 2015. URL http://news.mit.edu/2015/michael-stonebraker-wins-turing-award-0325.

Digital Science. Home Digital Science, 2016. URL https://www.digital-science.com/.

Digital Science, Tamar Loach, Jonathan Adams, and Martin Szomszor. The Diversity of UK Research and Knowledge - Analyses from the REF impact case studies, July 2015. URL http://figshare.com/articles/The_Diversity_of_UK_Research_and_Knowledge_Analyses_from_the_REF_impact_case_studies/1476881.

EC3Metrics. A las puertas de Scielo (Web of Science), May 2014. URL https://ec3metrics.com/las-puertas-de-scielo-web-science/.

J. Paul Elhorst and Katarina Zigova. Competition in Research Activity among Economic Departments: Evidence by Negative Spatial Autocorrelation. *Geographical Analysis*, 46(2):104–125, April 2014. ISSN 1538-4632. doi: 10.1111/gean.12031. URL http://onlinelibrary.wiley.com/doi/10.1111/gean.12031/abstract.

Elsevier. Scopus Content, 2015. URL http://www.elsevier.com/solutions/scopus/content.

Matthew E. Falagas, Eleni I. Pitsouni, George A. Malietzis, and Georgios Pappas. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal*, 22(2):338–342, January 2008. ISSN 0892-6638, 1530-6860. doi: 10.1096/fj.07-9492LSF. URL http://www.fasebj.org/content/22/2/338.

Daniel Fried and Stephen G. Kobourov. Maps of Computer Science. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific*, pages 113–120, March 2014. doi: 10.1109/PacificVis.2014.47.

Eugene Garfield. Citation indexes for Science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 1955.

Corrado Gini. *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche.* Bologna, c. cuppini edition, 1912.

Google. About Google Scholar, 2015. URL https://scholar.google.com/intl/en/scholar/about.html.

Miguel R. Guevara and Marcelo Mendoza. Publishing Patterns in BRIC Countries: A Network Analysis. *Publications*, 4(3):20, July 2016. doi: 10.3390/publications4030020. URL http://www.mdpi.com/2304-6775/4/3/20.

Miguel R. Guevara, Dominik Hartmann, and Marcelo Mendoza. *diverse: Diversity Measures for Complex Systems*. 2015. URL http://CRAN.R-project.org/package=diverse. R package version 0.1.1.

Miguel R. Guevara, Dominik Hartmann, Manuel Aristarán, Marcelo Mendoza, and César A. Hidalgo. The research space: using career paths to predict the evolution of the research output of individuals, institutions, and nations. *Scientometrics*, pages 1–15, September 2016a. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-016-2125-9. URL http://link.springer.com/article/10.1007/s11192-016-2125-9.

Miguel R. Guevara, Dominik Hartmann, and Marcelo Mendoza. diverse: an R Package to Measure Diversity in Complex Systems. *R Journal*, (en evaluacion), 2016b.

Antonio J. Gómez-Núñez, Benjamín Vargas-Quesada, Félix de Moya-Anegón, and Wolfgang Glänzel. Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, 89(3):741–758, August 2011. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-011-0485-8. URL http://link.springer.com/article/10.1007/s11192-011-0485-8.

Anne-Wil Harzing and Axèle Giroud. The competitive advantage of nations: An application to academia. *Journal of Informetrics*, 8(1):29–42, January 2014. ISSN 1751-1577. doi: 10.1016/j.joi.2013.10.007. URL http://www.sciencedirect.com/science/article/pii/S1751157713000849.

Yusef Hassan-Montero, Vicente P Guerrero-Bote, and F Moya-Anegón. Graphical interface of the SCImago Journal and Country rank: An interactive approach to accessing bibliometric information. *El profesional de la información*, 23(3):272–278, 2014.

Ricardo Hausmann and César A. Hidalgo. *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. The MIT Press, September 2013. ISBN 0-262-52542-9.

HEFCE. Research Excellence Framework 2014: The results. Technical report, Higher Education Funding Council for England (HEFCE), December 2014. URL http://www.ref.ac.uk/media/ref/content/pub/REF%2001%202014%20-%20full%20document.pdf.

César A. Hidalgo. Disconnected, fragmented, or united? a trans-disciplinary review of network science. *Applied Network Science*, 1(1), July 2016. ISSN 2364-8228. doi: 10.1007/s41109-016-0010-3. URL http://appliednetsci.springeropen.com/articles/10.1007/s41109-016-0010-3.

César A. Hidalgo, Bailey Klinger, Albert-László Barabási, and Ricardo Hausmann. The Product Space Conditions the Development of Nations. *Science*, 317(5837):482–487, July 2007. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1144581. URL http://www.sciencemag.org/content/317/5837/482.

J. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46): 16569–16572, 2005.

Charles Huamaní and Percy Mayta-Tristán. Producción científica peruana en medicina y redes de colaboración, análisis del Science Citation Index 2000-2009. *Rev Peru Med Exp Salud Publica*, pages 315–325, 2010. URL http://www.ins.gob.pe/insvirtual/images/artrevista/pdf/rpmesp2010.v27.n3.a3.pdf.

David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

Marco A. Janssen, Michael L. Schoon, Weimao Ke, and Katy Börner. Scholarly networks on resilience, vulnerability and adaptation within the human dimensions of global environmental change. *Global Environmental Change*, 16(3):240–252, August 2006. ISSN 0959-3780. doi: 10.1016/j.gloenvcha.2006.04.001. URL http://www.sciencedirect.com/science/article/pii/S0959378006000367.

Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, May 2006. ISSN 1600-0706. doi: 10.1111/j.2006.0030-1299.14714.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.2006.0030-1299.14714.x/abstract.

Jongkwang Kim and Thomas Wilhelm. Spanning tree separation reveals community structure in networks. *Physical Review E*, 87(3):032816, March 2013. doi: 10.1103/PhysRevE.87.032816. URL http://link.aps.org/doi/10.1103/PhysRevE.87.032816.

Richard Klavans and Kevin W. Boyack. Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, 60(3):455–476, 2009. ISSN 1532-2890. doi: 10.1002/asi.20991. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.20991/abstract.

Eric D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, softcover reprint of hardcover 1st ed. 2009 edition, December 2010. ISBN 1-4419-2776-X.

Cyril Labbé. Ike Antkare one of the great stars in the scientific firmament. *International Society for Scientometrics and Informetrics Newsletter*, 6(6, 2):48–52, June 2010. URL http://hal.univ-grenoble-alpes.fr/hal-00713564/. How Ike Antkare became one of the most highly cited scientists in the modern world and how you could become like him.

Heller Lambert. What will the scholarly profile page of the future look like? Provision of metadata is enabling experimentation., July 2015. URL http://blogs.lse.ac.uk/impactofsocialsciences/2015/07/16/scholarly-profile-of-the-future/.

Loet Leydesdorff and Ismael Rafols. A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60(2):348–362, 2009. ISSN 1532-2890. doi: 10.1002/asi.20967. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.20967/abstract.

Loet Leydesdorff, Stephen Carley, and Ismael Rafols. Global maps of science based on the new Web-of-Science categories. *Scientometrics*, 94(2):589–593, February 2013a. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-012-0784-8. URL http://link.springer.com/article/10.1007/s11192-012-0784-8.

Loet Leydesdorff, Caroline S Wagner, Han-Woo Park, and Jonathan Adams. Colaboración internacional en ciencia: mapa global y red. *El profesional de la información*, 22(1):87–94, 2013b.

Loet Leydesdorff, Félix Moya-Anegón, and Vicente P Guerrero-Bote. Journal maps, interactive overlays, and the measurement of interdisciplinarity on the basis of Scopus data (1996–2012). *Journal of the Association for Information Science and Technology*, 66(5):1001–1016, 2015. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.23243/pdf.

Loet Leydesdorff, Lutz Bornmann, and Ping Zhou. Construction of a Pragmatic Base Line for Journal Classifications and Maps Based on Aggregated Journal-Journal Citation Relations. *arXiv:1604.02716 [cs]*, April 2016. URL http://arxiv.org/abs/1604.02716. arXiv: 1604.02716.

Loet Leydresdof and Ismael Rafols. Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5:87–100, January 2011.

Manuel Lima and Ben Shneiderman. *The Book of Trees: Visualizing Branches of Knowledge*. Princeton Architectural Press, April 2014. ISBN 978-1-61689-218-0.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1 edition, July 2008. ISBN 0-521-86571-9.

Irena Marshakova-Shaikevich. System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya*, (6):3–8, 1973.

David Meyer and Christian Buchta. proxy: Distance and Similarity Measures, 2015. URL `http://CRAN.R-project.org/package=proxy`. R package version 0.4-14.

Jacob L. Moreno. *Who Shall Survive?* Beacon House, 1934.

Félix Moya-Anegón, Benjamín Vargas-Quesada, Victor Herrero-Solana, Zaida Chinchilla-Rodríguez, Elena Corera-Álvarez, and Francisco J. Munoz-Fernández. A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1):129–145, September 2004. ISSN 0138-9130, 1588-2861. doi: 10.1023/B:SCIE.0000037368.31217.34. URL `http://link.springer.com/article/10.1023/B%3ASCIE.0000037368.31217.34`.

Mark Newman. *Networks: An Introduction*. Oxford University Press, 1 edition edition, May 2010. ISBN 978-0-19-920665-0.

Mark Newman, Albert-László Barabási, and Duncan J. Watts. *The Structure and Dynamics of Networks:*. Princeton University Press, 1 edition edition, May 2006. ISBN 978-0-691-11357-9.

Mark E. J. Newman. Who Is the Best Connected Scientist?A Study of Scientific Coauthorship Networks. In Eli Ben-Naim, Hans Frauenfelder, and Zoltan Toroczkai, editors, *Complex Networks*, number 650 in Lecture Notes in Physics, pages 337–370. Springer Berlin Heidelberg, Berlin, Germany, January 2004. ISBN 978-3-540-22354-2 978-3-540-44485-5. URL `http://link.springer.com/chapter/10.1007/978-3-540-44485-5_16`.

OECD. Frascati Manual: Proposed Standard Practice for Surveys on Research and Experimental Development. Technical report, 2002. URL `http://dx.doi.org/10.1787/9789264199040-en`.

OECD. Revised field of science and technology (FOS) classification in the Frascati Manual., February 2007. URL `http://www.oecd.org/innovation/inno/38235147.pdf`.

ORCID. About ORCID, August 2012. URL `http://orcid.org/content/about-orcid`.

José Luis Ortega and Isidro F. Aguillo. Institutional and country collaboration in an online service of scientific profiles: Google Scholar Citations. *Journal of Informetrics*, 7(2):394–403, April 2013. ISSN 1751-1577. doi: 10.1016/j.joi.2012.12.007. URL http://www.sciencedirect.com/science/article/pii/S1751157713000023.

Abel L. Packer, Nicholas Cop, Adriana Luccisano, Amanda Ramalho, and Ernesto Spinak. *SciELO - 15 Years of Open Access: an analytic study of Open Access and scholarly communication*. UNESCO, 2014. ISBN 978-92-3-001237-3. URL http://scielo.org/php/level.php?lang=en&component=42&item=31.

Manh Cuong Pham, Ralf Klamma, and Matthias Jarke. Development of computer science disciplines: a social network analysis approach. *Social Network Analysis and Mining*, 1(4): 321–340, November 2011. ISSN 1869-5450, 1869-5469. doi: 10.1007/s13278-011-0024-x. URL http://link.springer.com/article/10.1007/s13278-011-0024-x.

Evelyn Chrystalla Pielou. *Introduction to Mathematical Ecology*. John Wiley & Sons Inc, New York, January 1970. ISBN 978-0-471-68918-8.

Brian Pink and Geoff Bascand. Australian and New Zealand Standard Research Classification (ANZSRC). Technical report, Australian Bureau of Statistics and Statistics New Zealand, 2008. URL http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1297.02008?OpenDocument#Publications.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

Ismael Rafols. Knowledge Integration and Diffusion: Measures and Mapping of Diversity and Coherence. In Ying Ding, Ronald Rousseau, and Dietmar Wolfram, editors, *Measuring Scholarly Impact*, pages 169–190. Springer International Publishing, 2014. ISBN 978-3-319-10376-1 978-3-319-10377-8. URL http://link.springer.com/chapter/10.1007/978-3-319-10377-8_8.

Ismael Rafols, Alan L. Porter, and Loet Leydesdorff. Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science & Technology*, 61(9):1871–1887, September 2010. ISSN 15322882. doi: 10.1002/asi.21368. URL http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=53286068&site=eds-live.

C. Radhakrishna Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, February 1982. ISSN 0040-5809. doi: 10.1016/0040-5809(82)90004-1. URL http://www.sciencedirect.com/science/article/pii/0040580982900041.

Research Gate. ResearchGate-About Us, 2015. URL https://www.researchgate.net/about.

Stephen A. Rhoades. Herfindahl-Hirschman Index, The. *Federal Reserve Bulletin*, 79: 188, 1993. URL http://heinonline.org/HOL/Page?handle=hein.journals/fedred79&id=376&div=&collection=.

Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, January 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0706851105. URL http://www.pnas.org/content/105/4/1118.

Alfréd Rényi. On Measures of Entropy and Information. The Regents of the University of California, 1961. URL http://projecteuclid.org/euclid.bsmsp/1200512181.

Science Metrix. Who we are, 2016. URL http://www.science-metrix.com/en/about-us/who-we-are.

SCImago. SJR-SCImago Journal and Country Rank, 2007. URL http://www.scimagojr.com/countryrank.php.

SciTech Strategies. Map of Science, 2012a. URL http://www.mapofscience.com/.

SciTech Strategies. Our Team, November 2012b. URL http://www.mapofscience.com/?page_id=102.

Claude E. Shannon. A mathematical theory of communication. *The Bell System technical Journal*, 27:379–423,623–656, October 1948.

Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, and Katsumori Matsushima. Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology*, 60(3):571–580, March 2009. ISSN 1532-2890. doi: 10.1002/asi.20994. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.20994/abstract.

Edward H Simpson. Measurement of diversity. *Nature*, 163, 1949.

Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4): 265–269, 1973.

Henry Small. Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9):799–813, 1999. ISSN 1097-4571. doi: 10.1002/(SICI)1097-4571(1999)50:9<799::AID-ASI9>3.0.CO;2-G. URL http://web.simmons.edu/~benoit/lis466/visdoc/ProQuest_43241275.pdf.

Andy Stirling. A general framework for analysing diversity in science, technology and society. *Interface The Journal of Royal Society*, 4(15):707–719, August 2007.

Arho Suominen and Hannes Toivanen. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10):2464–2476, October 2016. ISSN 2330-1643. doi: 10.1002/asi.23596. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.23596/abstract.

David Sweeney. The leading edge of impact, March 2015. URL http://www.nature.com/nature/supplements/collections/npgpublications/impact/hefce1.pdf.

The Royal Society. Knowledge, Networks and Nations: Global scientific collaboration in the 21st century. Technical report, 2011. URL https://royalsociety.org/topics-policy/projects/knowledge-networks-nations/report/.

Thomson Reuters. Web of Science Core Collection - IP & Science - Thomson Reuters, 2015. URL http://wokinfo.com/products_tools/multidisciplinary/webofscience/.

Michele Tumminello, Claudia Coronnello, Fabrizio Lillo, Salvatore Miccichè, and Rosario N. Mantegna. Spanning trees and bootstrap reliability estimation in correlation-based networks. *International Journal of Bifurcation and Chaos*, 17(07):2319–2329, July 2007. ISSN 0218-1274. doi: 10.1142/S0218127407018415. URL http://www.worldscientific.com/doi/abs/10.1142/S0218127407018415.

Nees van Eck and Ludo Waltman. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2009.

Ludo Waltman and Nees Jan van Eck. A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12):2378–2392, December 2012. ISSN 1532-2890. doi: 10.1002/asi.22748. URL http://onlinelibrary.wiley.com/doi/10.1002/asi.22748/abstract.

Jevin D. West, Michael C. Jensen, Ralph J. Dandrea, Gregory J. Gordon, and Carl T. Bergstrom. Author-Level Eigenfactor Metrics: Evaluating the Influence of Authors, Institutions and Countries Within the SSRN Community. SSRN Scholarly Paper ID 1636719, Social Science Research Network, Rochester, NY, August 2012. URL http://papers.ssrn.com/abstract=1636719.