



---

## CRÍTICA DE LIBROS / BOOK REVIEWS

---

### Analysis and Visualization of Citation Networks

Dangzhi Zhao y Andreas Strotmann

Morgan & Claypool Publishers, 2015. 208 pag. ISBN: 978-1-60845-938-4

La investigación estructural, que examina las interacciones entre los objetos que analiza, ha trascendido su ámbito originario en la Sociología y se ha introducido en la práctica totalidad de las disciplinas. Como la ciencia es un sistema complejo construido por las interacciones fuertes entre muchos componentes diversos (Shi, Foster & Evans, 2015) no resulta extraño que la Documentación Científica (DC) haya adoptado de forma natural a lo largo de los últimos 50 años el concepto de red y luego la metodología de su análisis.

A lo largo de su desarrollo, el análisis de redes sociales (ARS) ha ido incorporando cuatro enfoques que en la actualidad se complementan para definir el campo y las técnicas de investigación que utiliza (Freeman, 2004):

1. El ARS está motivado por una intuición estructural basada en los lazos que conectan actores sociales,
2. Está basado en datos empíricos sistemáticos,
3. Se apoya en gran medida en las representaciones gráficas, y
4. Se basa en el empleo de modelos matemáticos o computacionales.

En el ámbito de la DC, los trabajos sobre las relaciones de citas de los años 60 y 70 corresponden a la primera de esas fases y su impulso inicial, el de la "intuición estructural". En segundo lugar, los índices de citas (los tradicionales WoS, Scopus o CiteSeer) y los índices con citas (PsyncINFO) sirven como recopilaciones sistemáticas de datos empíricos. En tercer lugar, la nómina de instrumentos automáticos para el análisis, la visualización y la exploración de patrones en la literatura científica

es cada vez más amplia, ya se trate de programas dedicados (CiteSpace, Sci2) o de propósito general (UCInet, Gephi). Al cuarto enfoque me referiré más adelante.

Desde su encuentro en Alberta, a principios de los 2000, Dangzhi Zhao y Andreas Strotmann han sido coautores de casi 20 publicaciones, todas ellas relacionadas directa o indirectamente con las redes de citas. El libro que aquí se comenta es una recopilación de esos trabajos y, como en otros casos (Vargas-Quesada & Moya-Anegón, 2007) su estructura expositiva responde al esquema tradicional del artículo original de investigación, aunque aquí bastante desequilibrada. Me propongo seguir esa misma estructura para examinar su contenido. Después, situaré la obra en la tradición de los análisis de redes de citas y en el contexto de las obras contemporáneas sobre el mismo asunto.

#### El contenido del libro

El primer capítulo presenta los fundamentos. Para definir el análisis de citas, recurren a la habitual convención: una obra citada ejerce una influencia cognitiva sobre las que la citan, que se traduce tanto en un flujo de conocimiento como en una relación intelectual. El paso lógico siguiente hubiera sido exponer los tres tipos de relaciones que se derivan de las citas entre trabajos científicos. Sin embargo, hay que esperar otras cuatro páginas para encontrar las definiciones de cita directa (aquí llamada inter-citation) co-cita y enlace bibliográfico. Y esto porque el grueso del capítulo se dedica a las aplicaciones del análisis de citas. En conjunto, el capítulo recuerda mucho a la estructura del clásico de Eugene Garfield (1979). Así, el epígrafe "What is citation analysis" equivaldría al

capítulo inaugural del libro de Garfield "A Conceptual View of Citation Indexing"; los titulados "Tracking knowledge flows and the diffusion of ideas" y "Mapping Research Fields" eran en el clásico "Citation Analysis as a Method of Historical research into Science" y "Mapping the Structure of Science"; "Assessing information resources and evaluating scholarly contributions" correspondería a "A Science Management Tool".

De los cinco campos de aplicación que se comentan, sólo el dedicado a la representación gráfica de campos y especialidades y el que se centra en estudio de las trayectorias cognitivas están relacionados directamente con las redes de citas. Este comentario abunda en una de las principales limitaciones del libro: de los tres tipos de redes que las relaciones de citas permiten trazar, se ignoran las redes de citas (basadas en la relación de cita directa) y, sorprendentemente, las redes de enlace bibliográfico sólo se tratan en el ejemplo del capítulo dos. El libro se centra en las redes de co-citación, la técnica más común entre las tres según los autores (p. 5) porque en realidad trata de visualización de dominios. Además de esta limitación, el enfoque de la obra es ambiguo. Ya desde el resumen, los autores declaran en dos ocasiones que el libro se centra en el análisis estructural y descarta la evaluación. Sin embargo, ya he mencionado que la primera serie de aplicaciones que se comenta en este capítulo inicial es "Assessing information resources and evaluating scholarly contributions".

El resto del primer capítulo se dedica a justificar la validez y fiabilidad de las relaciones y los análisis de citas con una muy amplia revisión de los trabajos críticos y favorables al uso de los análisis de citas.

### Tres capítulos dedicados a los métodos

La parte metodológica del libro abarca los tres capítulos siguientes. Más de 90 páginas de las 147 que ocupa el texto. El capítulo 2, "Conducting Citation Network Analysis", detalla en su subtítulo título que contiene los pasos, los conceptos, las técnicas y las herramientas para realizarlo. Pero no es así. Arranca con una enumeración de los pasos habituales en la visualización de dominios. Los epígrafes que extienden cada uno de ellos, sin embargo, contienen un conjunto de consideraciones, comentarios y criterios, pero nada especialmente instrumental. Al menos en los apartados dedicados a los dos primeros pasos, la delimitación del campo y la selección de los objetos a analizar. El tercer paso del análisis, literalmente la medición de la conexión entre los

objetos, introduce las medidas de similitud derivadas de la cita directa, en enlace bibliográfico y la co-citación. La tabla con supuestos artículos y listas de referencia que utiliza como ejemplo (Tabla 2.1 en p. 32) es muy confusa y las llamadas a esa misma tabla a lo largo del texto, más aún. Un diagrama similar al que empleó Vargas-Quesada en su tesis (y reproduce en p. 35 o en la 65 de la obra antes citada) hubiera resultado mucho más claro. Por lo demás, esta parte del capítulo renuncia a tratar las medidas derivadas de las redes de cita directa y de enlace bibliográfico. Esto último sorprende porque los propios autores del libro introdujeron en 2008 el concepto de enlace bibliográfico entre autores, al que además han dedicado un par de trabajos empíricos. Como ya he dicho, es la co-citación la relación que centra no sólo el contenido de este apartado, sino del resto del libro.

El siguiente epígrafe de este mismo capítulo, titulado "Statistical Analysis of Citation Networks" trata de la aplicación de técnicas de análisis multivariado a la reducción de dimensiones y el descubrimiento de estructuras subyacentes en la red. Mientras que Vargas-Quesada y Moya-Anegón presentan una excelente discusión de los métodos de reducción y sus tipos (p 37-50 de su libro) Zhao y Stromann sólo repasan las técnicas estadísticas que han utilizado en sus trabajos empíricos, las mismas que se empleaban a principio de los años 80. Por lo demás, se embarcan en una serie de consideraciones sobre las operaciones a realizar sobre las matrices de citas que, a la vista de las prestaciones de cualquier aplicación actual de análisis, complican más que aclaran el procedimiento.

A pesar de su carácter metodológico, el capítulo se cierra con la exposición de un caso; en realidad la reproducción de uno de los últimos originales de los autores. Este apéndice presenta (esta vez sí) el procedimiento completo desde la extracción de datos al trazado de la red (la primera figura del libro, que se ha hecho esperar 58 páginas) y, aunque de forma algo desordenada, con remisiones a secciones de capítulos futuros, justifica el título del capítulo.

El tercer capítulo se titula "Field Delineation and Data Sources for Citation Analysis" y abunda en las fase de extracción de datos, examinando los requisitos que cualquier fuente bibliográfica debería cumplir. A continuación, repasa los pros y cons del uso de WoS y Scopus y el proceso de descarga de registros desde cada uno de estos sistemas, con una sucesión de volcados de pantalla que un diagrama de flujo podría haber resumido. Finalmen-

te, el capítulo recurre a otro de los originales de los autores para ilustrar el proceso de combinar un índice de citas (Scopus) con una base de datos temática, que es Index Medicus/PubMed. Al reproducir un original de 2011, los autores no cuentan que los códigos de identificación de PubMed (PMID) se incorporan en la actualidad a muchos registros de WoS. Resulta también extraño que no dediquen un comentario al uso del DOI en el "record linkage".

El capítulo 4 sigue siendo metodológico y está dedicado al problema de la desambiguación de los nombres de autores. Este problema se ve agravado por el hecho de que son los autores (y no los artículos individuales) las entidades que centran los análisis de Zhao y Strotmann.

El capítulo final se aparta del esquema convencional que los anteriores seguían y se estructura como si de un artículo original se tratase. Su título es "Visualization of Citation Networks" y se inicia revisando las redes que otros grupos han ido trazando sobre el campo de la Información y Documentación. Sigue una sección metodológica al uso, donde se describe la procesión de datos por Python, SPSS, Pajek e Inkscape y las manipulaciones a realizar hasta obtener la imagen de la red y decorarla. Vuelven a echarse en falta los esquemas o diagramas de flujo para representar los pasos, cuyo seguimiento no es lineal ni sencillo y sorprende sobremedida que los autores se aparten de la práctica habitual del análisis de redes al considerar la red que conecta autores y áreas como una red simple (figura 5.12 en p. 134) en lugar de bipartita. Las conclusiones del capítulo son también las del libro, porque vienen seguidas sólo por un par de apéndices y una extensa bibliografía.

### **Desequilibrios**

La primera impresión que la lectura del libro transmite es que se trata de una recopilación de trabajos originales apresurada y desequilibrada. El apresuramiento ha originado las redundancias en el texto y en la estructura del libro, cuyo segundo capítulo está de más. En cuanto al desequilibrio, lo hay tanto por exceso como por defecto. El exceso ha conducido a la inflación de epígrafes, como el que trata del análisis de las apariciones de las obras citadas en los textos (in-text citations) que nada aportan a la delimitación del análisis. Los autores podrían haberse ahorrado el gazapo (p. 93) que revela el recorte y pegue de esta sección. No tiene tampoco mucho sentido enfatizar las ventajas de Scopus como fuente si va ser WoS la finalmente elegida y de la combinación de índices de citas con

índices temáticos tampoco se ofrecen ejemplos en lo que queda de libro. En cuanto a sus carencias, la obra resulta incompleta en su conjunto y en los detalles de su contenido. Las redes de citas no se limitan a las de co-citación entre autores. El libro de Zhao Strotmann es una obra sobre visualización de dominios a través del análisis de citas, no sobre redes de citas. Y la obra de Vargas-Quesada y Moya-Anegón lo supera por claridad, orden expositivo y rigor, a pesar de ser ocho años anterior, diez en el caso de la tesis de Vargas. Un trabajo mucho menos ambicioso, como el de Miray Kas y sus colaboradores (2012) abarca un espectro más amplio de relaciones y métodos de análisis.

El descuido de los autores ha causado, por ejemplo, que el script en Python que les sirve para generar la matriz de adyacencia inicial entre los autores a partir de los registros descargados de WoS no aparezca en apéndice. Pero a eso me voy a referir a continuación.

### **El contexto actual del análisis de redes y su visualización**

Al repasar la evolución del análisis de redes sociales, Linton Freeman se ha referido en varias ocasiones a la ignorancia de algunos investigadores procedentes de la física estadística, que redescubrieron ideas y procedimientos instaurados con anterioridad entre los investigadores sociales sin reconocerlos (2004, p. 164-167). Zhao y Strotmann, a pesar de los ejemplos de visualización que ofrecen en su capítulo correspondiente, parecen ignorar el contexto actual del análisis de redes aplicado a datos bibliométricos e insisten en una metodología propia que han ido refinando con nuevas medidas de similitud para explorar repetidos conjuntos de datos pero que, en esencia, es la heredada de los 80.

Tras la descarga y preparación de los datos, el analista puede elegir por extraer una matriz de adyacencia o bien una sencilla lista de vértices (autores, revistas, trabajos) y las aristas que expresan sus relaciones (cita directa, co-autoría, co-cita, enlace bibliográfico). Si se opta por la primera opción, es preciso seguir el camino artesanal que pasa por introducir la matriz de adyacencia en SPSS y generar el análisis factorial. Después hay que convertir las matrices resultantes en archivos planos con vértices y aristas que se procesan con cualquier aplicación. Zhao y Strotmann optan por Pajek argumentando que quienes eligen otra aplicación tienden a perder el control si no realizan desambiguación o limpieza de datos.

Las alternativas a esta trabajosa rutina son numerosas. Entre las aplicaciones orientadas a la bibliometría, que toman directamente los registros descargados, los procesan y visualizan, se cuentan CiteSpace (<http://cluster.cis.drexel.edu/~cchen/citespace/>) Bibliotools ([http://www.sebastian-grauwin.com/?page\\_id=427](http://www.sebastian-grauwin.com/?page_id=427)) Sci2 (<https://sci2.cns.iu.edu/user/index.php>) o VOSviewer (<http://www.vosviewer.com/Home>). Todos los programas mencionados pueden representar redes de enlace bibliográfico y co-citación. Por otra parte, existen aplicaciones generales orientadas a la visualización, por ejemplo Gephi (<https://gephi.github.io/>).

Las redes representan sistemas y contienen valiosos datos sobre la organización de tales sistemas. Las herramientas que se usan para identificar, valorar y visualizar la estructura subyacente a una red se denominan métodos de detección de comunidades. Zhao y Strotmann recurren a técnicas estadísticas y se refieren al agrupamiento o clustering como un procedimiento que "clasifica un conjunto de objetos en dos o más grupos exclusivos [disjuntos] de forma que la asociación entre los miembros de un mismo grupo es fuerte y la relación entre miembros de clusters diferentes es débil" (p. 40). También observan que el análisis factorial se emplea en el análisis de co-citas entre autores para "revelar las especialidades de los campos de investigación y la pertenencia de los autores a una o más de esas especialidades" (p. 41). Estas frases resultan familiares porque coinciden a grandes rasgos con los conceptos de comunidades y su detección en el análisis de redes. Compárese si no con: "A common approach is to define a community as a group of vertices such as the density of edges between the vertices of the group is higher than the average edge density in the graph (Fortunato y Castellano, 2012). Todos los programas que he mencionado realizan de forma automática y controlada la partición y detección de comunidades.

## REFERENCIAS

- Batagelj, V., Doreian, P., Ferligoj, A., & Kejzar, N. (2014). Scientific Citation and Other Bibliographic Networks en *Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution*. John Wiley & Sons, pp. 105-154.
- Fortunato, S., Castellano, C. (2012). Community structure in graphs En R. A. Meyers (ed.)

## Los modelos del análisis de redes de citas

El análisis de redes científicas, apoyado en la naturaleza colectiva del esfuerzo investigador, está empezando a rendir una amplia gama de resultados (Yan & Ding, 2014). La delineación de dominios mediante análisis de citas, que ya se puede realizar casi de forma rutinaria (Grauwin & Jensen, 2011) es una de las más importantes aplicaciones del estudio de estas redes, pero sólo una de ellas. Y, por descontado, la visualización de redes es un ejercicio de gran valor exploratorio, pero no el objetivo único de los análisis. Vladimir Batagelj y sus colaboradores acaban de dedicar tres capítulos de su libro más reciente (2014) a las redes de citas, el primero de ellos a las citas científicas. Combinan órdenes en Pajek y [R] para realizar un análisis exhaustivo de las redes de citas directas, extendiendo así el contenido del capítulo 11 del Manual de Pajek. Chaomei Chen (2014) ha puesto en relación la visualización de dominios con la dinámica científica a través de un apasionante estudio de casos. El grupo de la Escuela de Gestión de la Innovación del Instituto Tecnológico de Tokio ([http://www.k-lab.mot.titech.ac.jp/index\\_eng.html](http://www.k-lab.mot.titech.ac.jp/index_eng.html)) está aportando interesantísimos resultados en sus numerosos análisis de redes de citas de artículos y de patentes. Los capítulos sobre citas y co-citas del libro colectivo sobre dinámica de la ciencia ofrecen enfoques evolutivos sobre las redes (Scharnhorst et al., 2012).

Estas y otras muchas actividades desarrollan modelos cuantitativos propios o los incorporan desde otros ámbitos para comprender los procesos de innovación, colaboración y desarrollo del conocimiento, el cuarto enfoque de Freeman. El libro de Zhao y Strotmann se atiene a modelos analíticos anteriores, tiene un valor informativo limitado y una utilidad formativa aún menor. Carece del orden del libro de Vargas y Moya, tampoco tiene el alcance del libro de Chen y su recopilación antes cierra una época que abre nuevas perspectivas.

**Carlos Benito Amat**

Ingenio (CSIC-UPV)

- Computational Complexity. Springer New York. pp. 490-512. Accesible también en <http://arxiv.org/pdf/0712.2716.pdf>
- Chen, C. (2014). *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. 2nd ed. New York: Springer.
- Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of*

- Science. Vancouver, BC : North Charleston, S.C: Empirical Press.
- Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. New York: Wiley.
- Grauwin, S., & Jensen, P. (2011). Mapping scientific institutions. *Scientometrics*, 89(3), 943-954. DOI: <http://dx.doi.org/10.1007/s11192-011-0482-y>
- Kas, M., Carley, K., & Carley, L. (2012). Trends in science networks: understanding structures and statistics of scientific networks. *Social Network Analysis and Mining*, 2(2), 169-187. DOI: <http://doi.org/10.1007/s13278-011-0044-6>
- Scharnhorst, A., Börner, K., & van den Besselaar, P. (2012). *Models of science dynamics: encounters between complexity theory and information sciences*. Springer Science & Business Media.
- Shi, F., Foster, J. G., & Evans, J. A. (2015). Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43, 73-85. DOI: <http://dx.doi.org/10.1016/j.socnet.2015.02.006>
- Vargas-Quesada, B., & de Moya-Anegón, F. (2007). *Visualizing the structure of science*. Springer Science & Business Media. DOI: <http://dx.doi.org/10.1007/3-540-69728-4>
- Vargas-Quesada, B. (2005). *Visualización y análisis de grandes dominios científicos mediante redes pathfinder (PFNET)*. Tesis Univ Granada. Disponible en <http://digibug.ugr.es/bitstream/10481/590/1/15407974.pdf>
- Yan, E., & Ding, Y. (2014). *Scholarly Networks Analysis*. En Alhajj R y Rokne j: *Encyclopedia of Social Network Analysis and Mining*. Springer New York. pp. 1643-1651.