THE GENERATION OF LARGE NETWORKS FROM *WEB OF SCIENCE* DATA

Loet Leydesdorff, Gohar-Feroz Khan and Lutz Bornmann

Loet Leydesdorff (Ph.D. Sociology, M.A. Philosophy, and M.Sc. Biochemistry) is professor at the *Amsterdam School of Communications Research (ASCoR)* of the *University of Amsterdam*. He is Honorary Professor of the *Science and Technology Policy Research Unit* (*SPRU*) of the *University of Sussex*, Visiting Professor of the *Institute of Scientific and Technical Information of China* (*ISTIC*) in Beijing, Guest Professor at *Zhejiang University in Hangzhou*, and Visiting Professor at the *School of Management*, Birkbeck, *University of London*. He has published extensively in systems theory, social network analysis, scientometrics, and the sociology of innovation. With Henry Etzkowitz, he initiated a series of workshops, conferences, and special issues about the *Triple Helix of University-Industry-Government Relations*. He received the *Derek de Solla Price Award* for Scientometrics and Informetrics in 2003 and held "The City of Lausanne" Honor Chair at the *School of Economics*, *Université de Lausanne*, in 2005. In 2007, he was Vice-President of the *8th International Conference on Computing Anticipatory Systems* (*Casys'07*, Liège). In 2014, *Thomson Reuters* listed him as a highly-cited author.
http://www.leydesdorff.net/list.htm
http://www.leydesdorff.net/th2/index.htm
http://highlycited.com
http://orcid.org/0000-0002-7835-3098

University of Amsterdam. Amsterdam School of Communication Research (ASCoR)
PO Box 15793, 1001 NG Amsterdam, The Netherlands
loet@leydesdorff.net

Gohar-Feroz Khan is an Assistant Professor at the *Korea University of Technology and Education*. Since his PhD from *Kaist* in 2011, Khan has published over 30 articles in refereed journals, conference proceedings, and book chapters. His research has been published in *Online information review*, *Social science computer review*, *Government information quarterly*, *Journal of the American Society for Information Science and Technology*, *Scientometrics*, *Information development*, *Internet research*, *Asian journal of communication*, and *Asia Pacific journal of information system*, among others. His research interests include, 1) social information systems & social media, 2) e-government, government 2.0, and 3) networked science. Prior to his doctoral studies, Dr. Khan held a senior management position with the *Afghan Ministry of Communications and Information Technology*. He is also an Associate Editor of *Journal of contemporary Eastern Asia* and founding Director of *Center for Social Technologies*.
http://gfkhan.wordpress.com/publications
http://orcid.org/0000-0003-2784-0918

Korea University of Technology & Education (*KoreaTECH*)
1600 Chungjol-ro Byungcheon-myun
Cheonan city, 330-708, South Korea
gohar.feroz@kut.ac.kr

Lutz Bornmann works as a sociologist of science at the *Division for Science and Innovation Studies* in the administrative headquarters of the *Max Planck Society* in Munich (Germany). Since the late 1990s, he has been working on issues in the promotion of young academics and scientists in the sciences and on quality assurance in higher education. His current research interests include research evaluation, peer review and bibliometrics, and altmetrics. *Thomson Reuters* lists him among the most-highly cited researchers worldwide over the last ten years.
http://highlycited.com
http://orcid.org/0000-0003-0810-7091

Division for Science and Innovation Studies.
Administrative Headquarters of the Max Planck Society
Hofgartenstr., 8. 80539 Munich, Germany
bornmann@gv.mpg.de

Article received on 09-29-2014
Final acceptance: 10-10-2014

Abstract

During the 1990s, one of us developed a series of freeware routines (*http://www.leydesdorff.net/indicators*) that enable the user to organize downloads from the *Web of Science* (*Thomson Reuters*) into a relational database, and then to export matrices for further analysis in various formats (for example, for co-author analysis). The basic format of the matrices displays each document as a case in a row that can be attributed different variables in the columns. One limitation to this approach was hitherto that relational databases typically have an upper limit for the number of variables, such as 256 or 1024. In this brief communication we report on a way to circumvent this limitation by using *txt2Pajek.exe*, available as freeware from *http://www.pfeffer.at/txt2pajek*

Keywords

Web of Science, Bibliometric network, *Pajek*, *txt2Pajek*.

Título: Generación de grandes redes a partir de datos de la *Web of Science*

Resumen

Durante la década de 1990, uno de nosotros desarrolló una serie de rutinas de software gratuito (*http://www.leydesdorff.net/indicators*) que permiten organizar las descargas desde la *Web of Science* (*Thomson Reuters*) en una base de datos relacional, y luego exportar matrices para su posterior análisis en varios formatos (por ejemplo, para el análisis de co-autores). El formato básico de las matrices muestra cada documento en una fila al que se le pueden atribuir diferentes variables en las columnas. Una limitación que entonces tenía este enfoque era que las bases de datos relacionales suelen tener un límite superior en el número de variables, por ejemplo, 256 o 1.024 En esta breve comunicación se presenta una forma de eludir esta limitación utilizando txt2Pajek.exe, disponible como freeware en el url *http://www.pfeffer.at/txt2pajek*

Keywords

Web of Science, Redes bibliométricas, *Pajek*, *txt2Pajek*.

Leydesdorff, Loet; **Feroz-Khan, Gohar**; **Bornmann, Lutz** (2014). "The generation of large networks from *Web of Science* data". *El profesional de la información*, v. 23, n. 6, November-December, pp. 589-593.

http://dx.doi.org/10.3145/epi.2014.nov.05

Introduction

In recent decades, one of us has developed a series of software routines that enables the user to organize downloads from the *Web of Science* (*Thomson Reuters*) into a relational database, and then to export matrices for further analysis in various formats; for example, for co-author analysis, co-citation analysis, bibliographic coupling, etc. (**Cobo** *et al*., 2011). The basic format of each matrix shows each document as a case in a row that can be attributed with different variables in the columns. Variables can be author names, institutional addresses, cited references, etc. One can also combine types of variables such as authors, title words, and institutional addresses (**Leydesdorff**, 2014; **Vlieger**; **Leydesdorff**, 2011). Multiplication of the asymmetrical word/document matrix with its transposed leads to a co-word matrix; and this operation can be done *mutatis mutandis* for other (sets of) variables attributable to documents.
http://www.leydesdorff.net/indicators

One limitation to this approach was hitherto that relational databases typically have an upper limit for the number of variables, such as 256 or 1024, whereas the number of cases (documents) is limited only by considerations of disk space[1]. In this brief communication, we report on a way to circumvent this limitation easily by using *txt2Pajek.exe* (**Pfeffer**; **Mrvar**; **Bagagelj**, 2013). *Txt2Pajek* enables the user to generate a 2-mode (asymmetrical) matrix of cases (documents) and variables in the *Pajek* format for an unlimited number of variables from a text file. Within *Pajek* (**De-Nooy**; **Mrvar**; **Batagelj**, 2011) the newly generated 2-mode file can be further transformed into a 1-mode network file that can also be used in other software programs for network analysis and visualization such as *Gephi*, *UCINet*, or *VOSViewer*.
http://www.pfeffer.at/txt2pajek

Data

One of us (GFK) encountered the systems limitation of 1024 variables when generating a co-author network at the level of institutional addresses using *instcoll.exe* for analysis and visualization. Using the eight journals listed in the so-called *Senior Scholars' Basket* of the *Association for Information Systems* (*AIS*) that were used for the ranking, 3,587 documents were downloaded for the period 1995-2014 (table 1). The set contains 7,397 institutional addresses, of which 4,617 are unique (**Khan**, in preparation). The author wished to pursue a network analysis using these names of institutions as nodes and had already organized the data download in a relational database using *isi.exe*[2].
http://www.leydesdorff.net/software/instcoll/index.htm
http://www.leydesdorff.net/software/isi

Analysis

The institutional names are organized by *isi.exe* in a separate table (named *cs.dbf*) that contains the document numbers for relational database management and the address information. Using *Excel* or a similar program, one can open

Table 1: The data of 3,587 documents in the eight journals (1995-2014) in the basket used by the *Association for Information Systems* (*AIS*) for ranking.

Journals	N
European journal of information systems	613
Information systems journal	341
Information systems research	549
Journal of information technology	447
Journal of management information systems	534
Journal of strategic information systems	331
Journal of the Association for Information Systems	239
MIS quarterly	533
Total	3,587

this table and save it as a comma-separated-variables (.csv) file or as tab-delimited. A program entitled *dbf2csv.exe* has additionally been made available at this url for a direct transformation:
http://www.leydesdorff.net/software/dbf2csv/dbf2csv.exe

The comma-separated files can be read as text files into *txt2Pajek.exe* and are transformed in 2-mode *Pajek* files.

One can further refine the address information by using functions of *Excel*. For example, the first address in the file was "Unist, Sch Technol Management, Ulsan, South Korea" in cell B2. Using the function "=left(B2, find(",",B2)-1)", one obtains the institutional name "Unist" in another cell (e.g., C2). Since institutional names are now considerably standardized in *WoS*, one can drag the function along the column in Excel and thus obtain a field with only institutional names. There are 1,364 unique institutional names in the set based on 3,564 (of the 3,578) documents. Similarly, one can extract country names on the right side of the string using more composed functions or by writing a routine[3].

The .csv file should be renamed with the extension ".txt" and one should take care that the content is either lower or upper case (or capitalized case) because the default cases were changed in *WoS* during the 1990s. The transformation by *txt2Pajek* is straightforward and provides a file with the same name, but with the extension ".net" in the *Pajek* format. This file can be read into *Pajek* or another network-analysis program that is able to read this format. The *Pajek* format is nowadays increasingly the standard currency for exchanges among network analysis and visualization programs.

Network analysis in *Pajek*

When the 2-mode network generated by *txt2pajek.exe* is read into *Pajek* (v.3), it can be transformed into a 1-mode network (in this case of institutes) under *Network > 2-Mode Network > 2-Mode to 1-Mode > Columns*. The option "multiple lines" should be set ON. Thereafter the multiple lines have to be summed under *Network > Create New Network > Transform > Remove > Multiple Lines > Sum Values.* The lines of the network (edges) can now be visualized with different widths. Similarly, one can size the nodes using "weighted degrees" for the number of occurrences under *Network > Create Vector > Centrality > Weighted Degree > All.* Using the Draw-menu now visualizes the network (*Draw > Network + First Partition + First Vector*; **Bruun**, 2009).

As would be expected, institutional collaboration networks contain lots of isolates, dyads, triads, etc. These small networks are not necessarily connected among themselves. The network under study thus contains 153 components with a largest component of 1,171 (85.9% of the 1,364) nodes. Figure 1 shows this largest component as a heat map after exporting to *VOSViewer* (**Van-Eck**; **Waltman**, 2010).

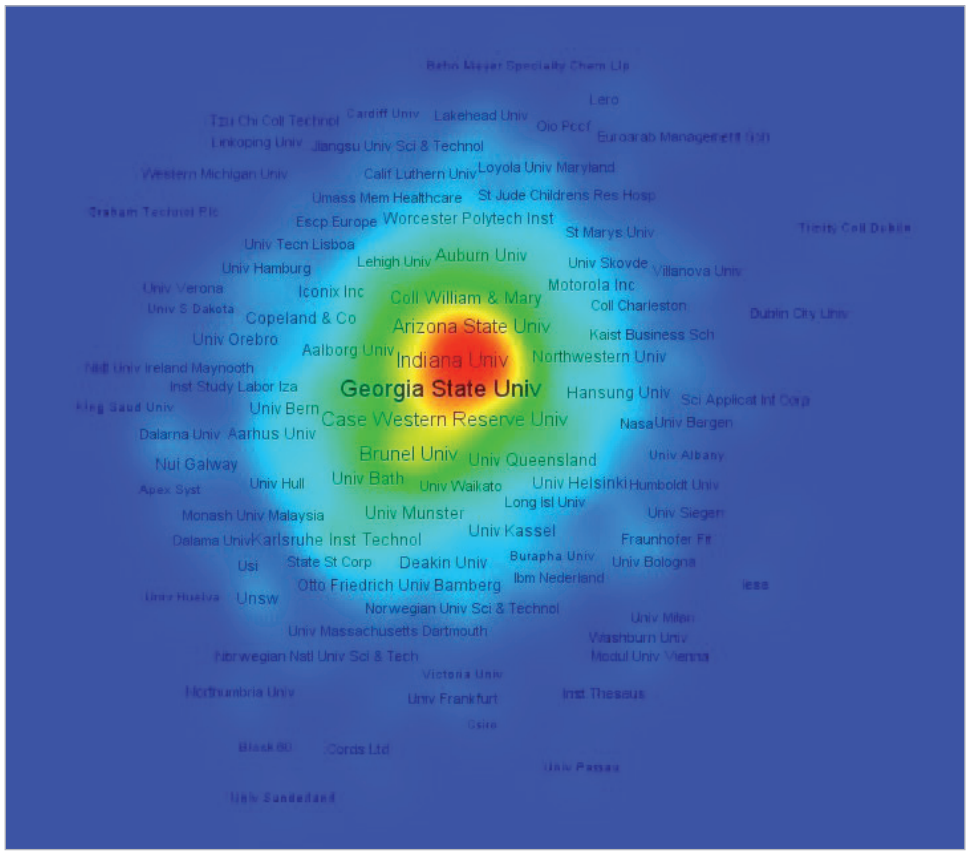

Figure 1. Heat map of the largest component (N = 1,171) of the network of institutional collaborations in the *AIS*-basket of 8 journals in "information systems".

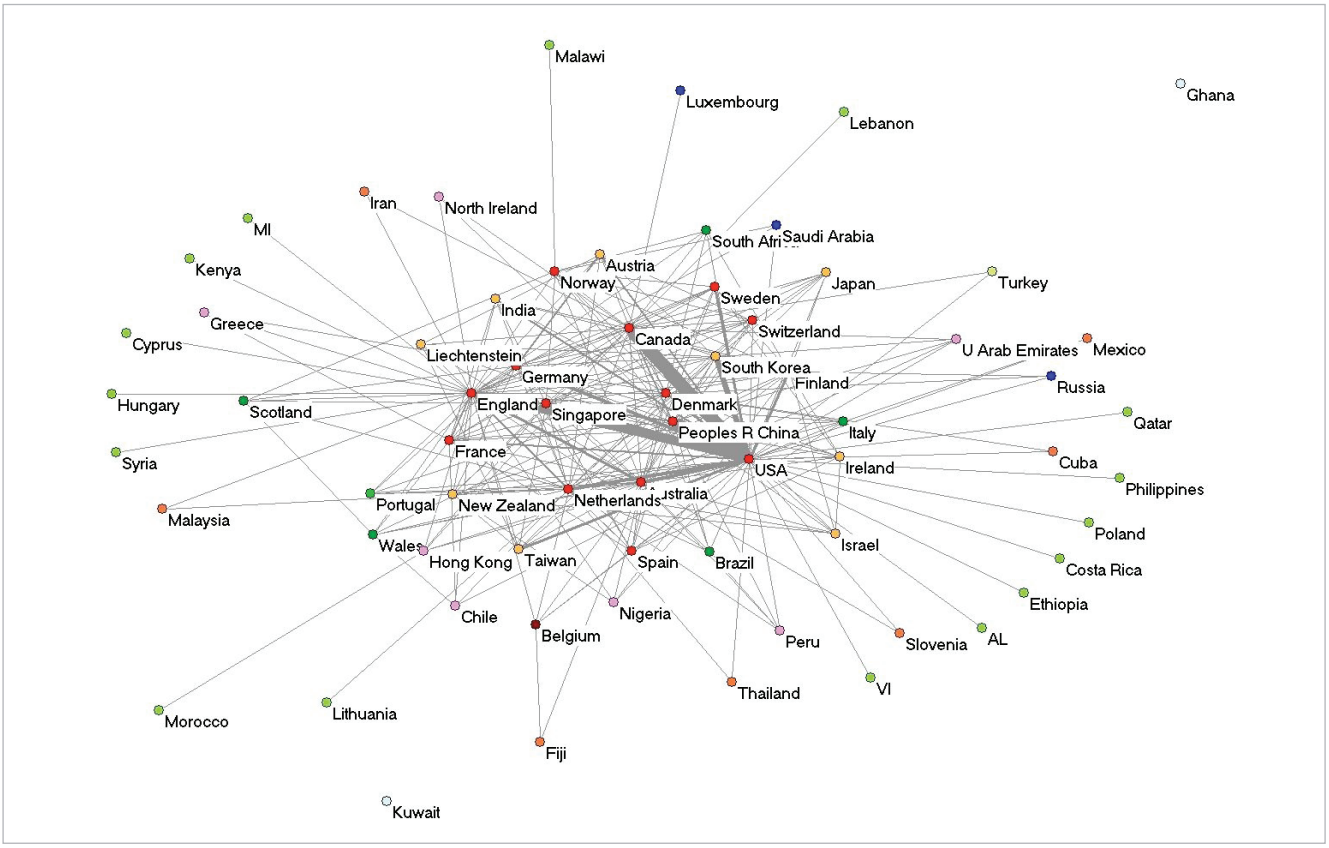

Figure 2. Network of international collaborations in the *AIS* basket; 67 countries; 3,563 documents. **Kamada** & **Kawai** (1989) used for the mapping.

The second largest component contains only nine institutes.

Figure 2 shows the network of 67 countries named in the institutional addresses of the 3,563 (99.6%) documents that provide such information. Note that in *WoS*, "England" is counted separately from the other countries of the UK.

Conclusions and summary

Using this pathway, one can visualize both smaller and very large networks, for example, of authors in large consortia (such as at *CERN*; **Milojević**, 2010). The routines *isi.exe* and *txt2pajek.exe* have no systems limitations except disk sizes. Bringing the files into network analysis and visualization programs, one can study degree distributions, clustering coefficients, modularity, etc., and visualize subsets accordingly. An alternative route for achieving this is provided by *Wos2Pajek*, but in this case the data is not organized relationally into databases. We have demonstrated the possibilities for analysis and visualization of collaborations in the specialty of "information systems" both at the institutional and international levels.
http://pajek.imfm.si/doku.php?id=wos2pajek

Notes

1. In a 32-bit operating environment, file sizes are limited to 2 GB, but this limitation is removed in the environment of a 64-bit operating system.

2. One can use *scopus.exe* at *http://www.leydesdorff.net/scopus* for transforming data from *Scopus* into this format.

3. The table cs.dbf already contains country names as a separate (third) field.

References

Bruun, Jesper (2009). *Physics and didactics: maps of text on scientific literacy*.
http://absalon.itslearning.com/jbruun/blog

Cobo, Manuel-Jesús; **López-Herrera, Antonio G.**; **Herrera-Viedma, Enrique**; **Herrera, Francisco** (2011). "Science mapping software tools: Review, analysis, and cooperative study among tools". *Journal of the American Society for Information Science and Technology*, v. 62, n. 7, pp. 1382-1402.
http://dx.doi.org/10.1002/asi.21525

De-Nooy, Wouter; **Mrvar, Andrej**; **Batagelj, Vladimir** (2011). *Exploratory social network analysis with Pajek* (2nd edition). New York, NY: Cambridge University Press. ISBN: 978 0521174800

Khan, Gohar-Feroz (in preparation). *Roots and fruits of information system domain: a network perspective*.

Kamada, Tomihisa; **Kawai, Satoru** (1989). "An algorithm for drawing general undirected graphs". *Information processing letters*, v. 31, n. 1, pp. 7-15.
http://dx.doi.org/10.1016/0020-0190(89)90102-6

Leydesdorff, Loet (2014). "Science visualization and discursive knowledge". In: Cronin, Blaise; Sugimoto, Cassidy (Eds.). *Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact*, pp. 167-185. Cambridge MA: MIT Press. ISBN: 978 0262525510

Milojević, Stasa (2010). "Modes of collaboration in modern science: beyond power laws and preferential attachment".

Journal of the American Society for Information Science and Technology, v. 67, n. 7, pp. 1410-1423.
http://arxiv.org/pdf/1004.5176.pdf
http://dx.doi.org/10.1002/asi.21331

Pfeffer, Jürgen; **Mrvar, Andrej**; **Batagelj, Vladimir** (2013). *txt2pajek: Creating Pajek files from text files technical report*, CMU-ISR-13-110. Carnegie Mellon University: School of Computer Science, Institute for Software Research.
http://www.pfeffer.at/txt2pajek/txt2pajek.pdf

Van-Eck, Nees-Jan; **Waltman, Ludo** (2010). "Software survey: VOSviewer, a computer program for bibliometric mapping". *Scientometrics*, v. 84, n. 2, pp. 523-538.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2883932
http://dx.doi.org/10.1007/s11192-009-0146-3

Vlieger, Esther; **Leydesdorff, Loet** (2011). "Content analysis and the measurement of meaning: the visualization of frames in collections of messages". *The public journal of semiotics*, v. 3, n. 1, pp. 28.
http://www.leydesdorff.net/semiotics/semiotics.pdf