Association for Information Systems
AIS Electronic Library (AISeL)

WHICEB 2014 Proceedings | Wuhan International Conference on e-Business

Summer 6-1-2014

A Modeling and Analysis Framework for Knowledge System Based on Meta-Network Approach

Xiao Liu
School of Management, Jinan University (Guangzhou), China

Jun Wang
School of Economics, Jinan University (Guangzhou), China

A Modeling and Analysis Framework for Knowledge System Based on Meta-Network Approach

Xiao Liu[1]； *Jun Wang*[2]
[1]School of Management, Jinan University (Guangzhou), China
[2]School of Economics, Jinan University (Guangzhou), China

Abstract: Nowadays some online research platforms (e.g., Web of Science or IEEE Xplore) provide bibliographic content and tools to access, analyze, and manage the world's leading journals and conference proceedings in sciences, social science, arts, and humanities. However, when facing increasingly mass literature, it's very difficult for researchers to effectively and systematically acquire the knowledge structure about a particular topic by using traditional literature reviewing method. Therefore we need explore new knowledge discovery tools for knowledge representation in an effective and efficient way. This paper proposes a knowledge system meta-network model by identifying the concepts representing entities and relationships from bibliometric data, and a methodology framework for meta-network modeling and analysis by using integrated techniques, including text mining, network text analysis, social network analysis, longitudinal network analysis and visualization. Case study using the Web of Science database as data source, explores the knowledge structure and interdisciplinary cooperation mode, as well as hot topics evolution in the field of World Trade Web.

Keywords: knowledge system, meta-network, text mining; network text analysis; world trade web

1. INTRODUCTION

The online platforms, such as Web of Science (literature citation index on the science, social science, art and humanities) and IEEE Xplore (literature citation index on engineering and technology), are one of the main source of bibliometric data. These database containing information on the academic literature are the basic unit of society and knowledge systems analysis.

Although researchers can access publication data in electronic form, including the title of article, authors, publisher, publish date, address of authors, keywords, abstract, references and citation information, etc., they can't immediately get a complete overview in certain filed because reviewing literatures is a time-consuming job. For keeping catch with the emergence of new articles, researchers have to read large relevant literature information. These situations motivate the investigation and development of automated techniques for extracting the underlying structure of knowledge from publication data.

Bibliometrics is a type of research method used in library and information science. It utilizes quantitative analysis and statistics to describe patterns of publication within a given field. Many research fields use bibliometric methods to explore the impact of their field, the impact of a set of researchers, or the impact of a particular paper [1-2].

Except for body of works in bibliometrics, the interdisciplinary applications of network-based approach have attracted a great deal of attentions during the past two decades, concerning social, economic, technological and biological systems. A use of networks as model of complex systems and processes brings a number of advantages. Firstly, networks have a convenient graphical representation: network nodes and links can be visualized as graph vertices and edges respectively. Visualization of networks as graphs has value in itself for understanding intricate relationships within complex systems. Secondly, researcher can employ graph-theoretical measures to describe the network topology structure and characteristics.

Some studies use scientific literature to build interlinked co-authorship networks [3-5], co-citation networks

[6-7], co-publication networks [8-9] and co-words networks [10-11], and then to compare scientific activities or investigate how a specific research area changes over time. Some studies focus on network dynamics by using statistical physics or simulation methods, in order to reveal the existence of specific network topologies and the structuring mechanism. These studies provide quantitative analytical results and interesting empirical findings, but merely disclose one side of knowledge rather than the whole picture. There are still lack of research framework from the view of systems thinking.

In fact, a given certain knowledge subject can be considered as a system which is made of components, relationships, and attributes. Components are various types of entities such as individual author, paper, journal and research institute (e.g., university or lab). Relationships are the links between the components. Attributes are the properties of the components and the relationships.

However new challenges have arisen because the real environment involves large amounts of entities and relational data that is dynamic. In other words, the nodes and edges (relations) may appear, disappear, or change in strength over time as new information arrives. On the other hand, large-scale data can be gathered from online database. These large data streams require more powerful tools to digest and manipulate, and extract network topology with minimal computational overhead.

Therefore the aim of this paper is to construct a modeling and analysis framework so that we can employee innovative techniques or tools for representing the whole structure of knowledge system. New analytical techniques, such as text mining, network text analysis, dynamic meta-network modeling and analysis, should be explored. Especially in the scenario of big data, the knowledge system is a dynamic and evolving complex system.

This paper is organized as follows: Section 2 introduces a meta-network model for representing the structure of knowledge system based on scientific literature dataset. Section 3 proposes a methodology framework integrated multiple kinds of techniques, such as data collection, data cleaning, network text analysis, social network and dynamic network analysis, for meeting to the needs of data pre-processing, modeling and analysis. Section 4 explores the knowledge structure of World Trade Web (WTW), by using dataset connected form the Web of Science. The final section concludes this study and suggests future work.

2. KNOWLEDGE SYSTEM META-NETWORK MODEL

The approach of meta-matrix (i.e., meta-network), proposed in Ref [12], can be thought of as a conceptual description of the organization and as an ontology for characterizing key organizational entities (e.g., people, knowledge, resource, task) and relations among them(e.g., social network, knowledge network, resource network, assignment network). There are two key points in defining a meta-matrix. Firstly a set of entities are identified. Secondly, the relations among entities are identified.

Entity-Relationship-Attribute ideas are commonly used to specify and design information system. An entity is a class of objects which have certain attributes or properties. A relationship among entities is a relation. In bibliometric dataset, there are several important kinds of entities that can be connected, modeled and analyzed. (1) An author is narrowly defined an originator of any written work. Attributes of author include personal information, e.g., name, email address, job title, nationality, etc.. (2) A paper is an academic work that is usually published in an academic journal. Attributes of paper include article title, publishing date, journal name, times cited, pages, etc.. (3) An organization/institute generally refers to university or research lab which authors belong to. Attributes of organization/institute include information on name and address. (4) Keywords are the words/phrases that academics use to indicate the topic area of the paper. Most of journals include keywords, some are not. Careful selection of keywords for a paper will increase the chance of someone retrieving, reading and citing it.

In addition, the abstract of a paper reveals the internal structure of an author's reasoning. In generally, the

abstract should include the following information: (1) purpose of the paper: what are the reasons for writing the paper or the aims of the research; (2) design/methodology/approach: how are the objectives achieved; what are the main methods used for the research; (3) findings: what was found in the course of the work? So, some of concepts (i.e., words/phrase) can be extracted from the abstract for representing the mental map of an author.

A modeling framework (see Table 1) can be characterized as a set of interlocked networks connecting entities such as author/people, paper/article, organization/institute, keywords and concepts.

Table 1. Meta-network of knowledge system

	Authors	Papers	Organizations	Keywords	Concepts
Authors	Co-authorship network Who cooperate with whom	Author-paper network Who publish which paper	Membership network Who is in which organization	Author-Keywords network Who knows what, and in which field	Author-Mental Map Who contribute what, who propose a new idea and method
Papers		Citation network Which paper cites which paper	Affiliation network Which paper is belong which institutes	Research Field What is topic in a paper, what kind of knowledge involved in	Research Outline What is research problem, motivation, method and contribution
Organizations			Co-publishing network Which institute collaborates which institute	Knowledge Distribution Which institute focuses which research filed	Knowledge outcomes/products Which institute contribute what
Keywords				Co-words network Co-occurrence of keywords in a number of papers	Knowledge Map Concepts maps to integrate or interpret knowledge
Concepts					Semitic network Word associations, word network

There are several networks commonly studied as follows: (1) Co-authorship network. Co-authorship network links cooperation relationship among authors. Links are reciprocal (symmetric), and the link weights (times of co-author publish) can increase over time. (2)Author-paper network. Author-paper network links authors to papers (the outcomes of academic collaboration). This type of network is also known as two-mode or bipartite network because the ties are only established between nodes belonging to different sets. (3) Membership network. The membership network links authors to institutes/organizations. (4) Author-keywords network. Author-keywords network links authors to keywords, representing who know what and in which research field. (5) Author-mental map. The concept of mental map was introduced by Edward Tolman.[13] Mental map (i.e., cognitive map) is a type of mental representation which serves an individual (e.g., author) to code, store, recall, and decode information. Mental map can be extracted from the abstract of paper, representing as a network of relations between author and their idea (i.e., words/phrases). (6) Co-citation network. Citation networks are directed because the links go from one paper to the other, and are acyclic because a paper can cite only existing papers. All edges in the citation networks point backwards in time. Paper citations are helpful for tracing topics or subfields that are related to a specific research interests. (7) Co-words network. The network constituted with author-keywords matrix is called author co-words network, which is a weighted network without directions. The weight is the co-occurrence frequency of two author-keywords. It reflects the relation among multiple vocabulary conceptions and also represents the structure of scientific research. (8) Semantic network. A semantic network which links are directed and labeled represents semantic relations between concepts (i.e., words/phrases). Semantic network often is used as a form of knowledge representation and supports artificial intelligence system for reasoning about the knowledge [14-16].

The advantages of using meta-network to model bibliometric data are clearly identifiable. Firstly, each cell in the meta-matrix described in Table 1 can be instantiated not only in multiple ways, but flexibility. The relation can have attributes indicating the strength, frequency or existence of the associated connection among the entities, and the relation cam be dynamic, changing with changes in author, paper and topic. Secondly, ‘hidden’

relationships between entities can be identified through analyzing their connections to nodes of different kinds. For example, co-authorship is a network expressing existence of co-authorship relation between authors of scientific papers, so it can be derived from the author-paper network by matrix computation. Thirdly, changes in one network cascade through the entire meta-network. The dynamic mechanism can be identified and effectively tested using the meta-network concept.

Even more significantly, we can answer the following questions based on meta-network comprehensive analysis: who are the key authors or scientists in this field of research? How often do authors publish? Which organizations or countries are involved in this research? How often do certain author collaborate? What do the topological features of collaboration networks tell us? How can we use text contents to detect scientific breakthroughs? How does a research field evolve over time along with the changes in the semantic network, changes of citation and collaboration networks? The answers of these questions help us understand the overall structure of knowledge area.

3. A METHODOLOGY FRAMEWORK

The scientific literature data connected from online platform by using browsing or exporting tool, is a kind of semi-structural data, different with structured data (in a fixed field within a record or file, e.g. relational databases and spreadsheets). Therefore, the challenges are how to identify entities and relations from such textual data, and how to build meta-matrix model based on various entities and relations.

In this section, we propose an interdisciplinary approach which integrates various techniques or tools for supporting the process of knowledge discovery and representation, including data collection and pre-processing, text mining for identifying concepts from textual data, network text analysis for ontology definition and meta-matrix modeling, social network analysis, statistical analysis of topology characteristics of network, longitudinal network analysis (i.e., dynamic network analysis) and community detection, etc.. A methodology framework is showed in Figure 1.

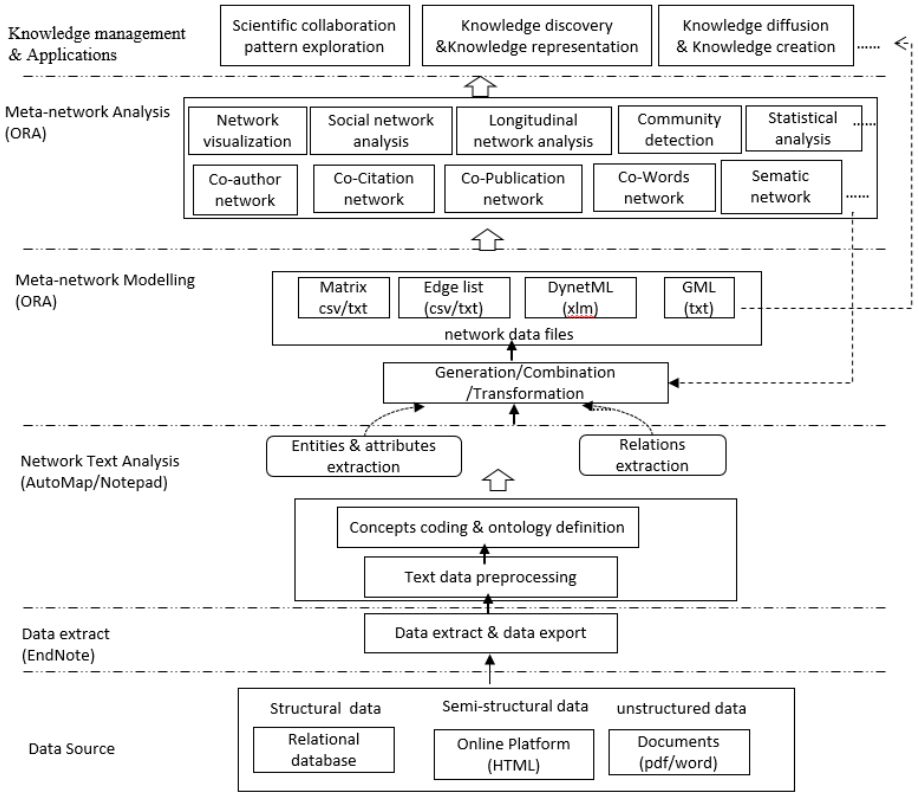

Figure 1. A methodology framework for knowledge meta-network modeling and analysis

3.1 Data extract and export

For the semi-structured data, tags or other types of markers (i.e., metadata) are used to identify certain elements. Some convenient software tools, such as EndNote (bibliographic management software), can be used as a personal database to gather and store citation records from different information sources. Most importantly, these tools also can extract metadata from huge PDF article collection, and then export to the text file. If data source stores in a relational database, then SQL (Structural Query Language) is convenient tool for data querying and export.

● Text data pre-processing. This step mainly includes data cleaning and concepts identification. The complexity of data pre-processing depends on the data sources used. The common pre-processing includes transforming typos into correct forms, remove plurals, deletion of irrelevant or meaningless terms, locate known and common n-grams, and employ thesauri listing relevant concepts set of interest. The thesauri is referred to as generalization thesauri that is organized a two-columned collection associating a concept with a corresponding higher-level concept.[20] An example would be associating "world trade network" or "international trade network" with "world trade web". Thesauri can be used to handle aliases (e.g., abbreviations are attached to the full concept), reduce specific concepts to more general concepts, and combine similar concepts. Data pre-processing need run multiple iterations for constructing the generalization thesauri and identifying the sets of relevant concepts, so this step is the most time-consuming phase in the whole process of knowledge discovery. But fortunately, some advance tools (e.g., AutoMap[21]) are developed to run preliminary processing steps automatically under support of researchers.

● Entity identification and ontology definition. The main work of this step is to create the ontology thesaurus or meta-network thesaurus, which are used to assign concepts to their corresponding entity class. The word ontology has been taken from Philosophy, now became a fashionable word in the knowledge engineering. An ontology provides the means for describing explicitly the conceptualization behind the knowledge represented in a knowledge base. The general or common ontologies include vocabulary related to things, events, time, space, behaviour, etc. According to ORA user's guild [22], there are several ontology classes including agent, event, knowledge, resource, task, organization, location, role, action, belief, and others. Entity identification and classification depends on the research questions of interests and expert's judgment. In our study, there are five entity classes including author (agent in ORA), paper (event in ORA), institute (organization in ORA), keywords (knowledge in ORA) and knowledge concepts (concept in ORA). For a more complex ontological scheme with multiple entity classes, using meta-network attribute thesauri that have more than two columns is good choice.

3.3 Meta-network extraction and modeling

The complexity of meta-network extraction depends the number of entity classes, because if the number of entity classes is N, the number of network that can be extracted is N-choose-2. This step can use AutoMap to process the set of texts, and generate out a meta-network.

Because AutoMap generates a meta-network for each text, there are thousands of network if there are thousands of texts. So analysts need to consider how to combine these meta-network in a meaningful way according to the need of research questions. ORA provides several ways to fuse network data [22].

3.4 Meta-network analysis and visualization

The ORA is a statistical analysis package for operating meta-networks. It was originally conceived as a risk assessment tool for identifying individuals of potential risks to groups or organization. It combines techniques, methods and algorithms from graph theory, machine learning, operations research, social psychology, probability theory, and matrix algebra. ORA reports provide quantitative information about traditional social network measures and extended graph measure. From the website of CASOS users can

download the latest version software.

4. CASE STUDY

4.1 Data and data preprocessing

To construct the database for this study, we choose a set of articles in a particular field as the analytical sample from Web of Science. Publications are extracted using the string "world trade web" in the titles or keywords or abstracts and restricting the research to publications in English. The period limits from January 1, 2003 to December 30, 2012. For simplify data, the citation data between papers is not be connected. 98 related articles are obtained. After the elimination of repetitions, conference papers and the ones that have nothing to do with the major topic, the actual number of availability is 58, the number of authors is 101.

We export meta-data of 58 articles by using EndNote software, and use Notepad tool to edit text file (Figure 2). The first paragraph lists coauthor information for identifying the co-authorship relationship. The second paragraph shows information on the first author, title, journal and published date for extracting the actor-event network. The third paragraph lists the abstract for identifying the semantic network. The rest of paragraphs list author's address information for extracting the membership network.

4.2 Coding and modeling

Before using AutoMap to extract meta-networks, we create three text files for supporting meta-network generation, including delete-list, generalization-thesaurus and metanetwork-thesaurus. The most technical job is the definition of concepts based on expert's knowledge. Figure 3 shows an example.

We firstly import all of the text file using AutoMap, then employ "delete list" command to deal with the meaningless words. Secondly we apply the generalization thesaurus to identify concepts. Thirdly we extract the meta-networks from text file by using metanetwork thesaurus. Fourthly, we import all of meta-networks in ORA, then manually operate meta-network data, including deletion of unnecessary nodes or links, combination of meta-networks, and matrix computation for generating new relation/network.

M. Angeles Serrano; Marian Boguna.
M. Angeles Serrano; Topology of the World Trade Web; Physical Review E; 2003.
Economy, and consequently trade, is a fundamental part of human social organization which, until now, has not been studied within the network modelling framework. Networks are mathematical tools used in the modelling of a wide variety of systems in social and natural science…….
M. Angeles Serrano; University of Barcelona, Spain.
Marian Boguna; University of Barcelona, Spain.

Figure 2. The example of text file format (each paper as a separate document)

delete-list.txt:
a
an
of
is
in
are
not
has
that
then
have
What
……

generalization-thesaurus.txt:
M.Angeles Serrano,Serrano_M_A
Topology World Trade Web,Topology_World_Trade_Web
Physical Review E,Physical_Review_E
2003,2003
trade,trade
social organization,Social_Organization
Network,network
networks,network
modelling framework,modelling_framework
mathematical tools,mathematical_tools
systems,system
……

metanetwork-thesaurus.txt:
Serrano_M_A/agent
Topology_World_Trade_Web/event
Physical_Review_E/attribute
2003/time
Social_Organization/concept
network/concept
modelling_framework/concept
mathematical_tools/concept
scale_free/concept
degree_distribution/concept
Barcelona_University/organization
Spain/location
……..

Figure 3. The data files for data preprocessing, concepts identification and meta-network extraction

4.3 Analysis and visualization

The timely evolution of publications for each year shows in Table 2. The first paper (based on our dataset) was published in 2003. The number of publications had a huge growth in 2007 and 2008.

Table 2. Articles on world trade network

Publishing Year	2003	2004	2005	2007	2008	2009	2010	2011	2012
Number of authors	5	2	2	24	25	13	17	25	12
Number of papers	2	1	1	11	17	6	7	9	4

The Top 5 journals publishing article in the field of WTW include Physica A (7 articles), Physical Review E (6 articles), European Physical Journal B (6 articles), Advances in Complex Systems (2 articles), Ecological economics (2 articles). The rest journals publishing one article include Journal of Systems Science & Complexity, Journal of International Business Studies, World Development, etc.

There are three obvious clusters in the author-paper network. The largest group (Cluster1) gathers 15 authors and 22 papers (Figure 4). There are 8 authors and 3 papers in the Cluster2. The Cluster3 include 5 authors and 2 articles. The largest strong component in the co-authorship network is showed in Figure 5.

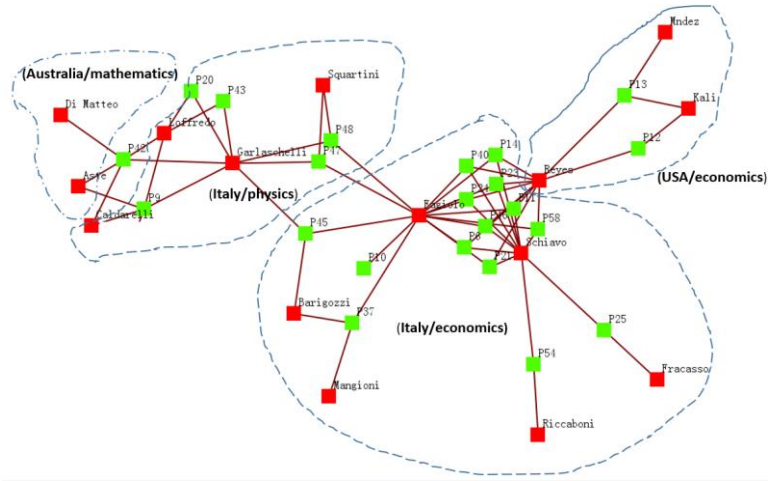

Figure 4. The largest cluster in the author-paper network (p# representing paper ID)

We use 'Key Entity' analysis tool in ORA to evaluate each author's position and role in the co-authorship network. According to the centrality characteristics (e.g., degree, betweeness, closeness and eigenvector), the Top 5 scientists are Garlaschelli, Fagiolo, Barigozzi, Schiavo and Reyes. Some of them are high productivity who also have many cooperators. Some of them are the bridge persons among different organization, subject or countries, like Garlaschelli , Fagiolo and Reyes (see Figure 5).

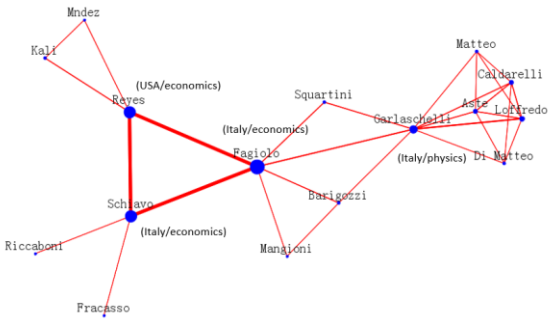

Figure 5. The largest strong component in the co-authorship network (with 2 or more papers published)

In Figure 6, the research topics involve in globalization, trade flow pattern, trade policy, trade barriers, economic crisis, economic growth and development. The mainstream network analysis methods focus on the

community structure detection and the centrality analysis. These keywords indicate that WTW has been studied from two perspectives: one originated in research on complex system in physics; the other view comes from economic and social science. Econophysics, which an interdisciplinary research field applying theories and methods originally developed by physicists in order to solve problems in economics, emerges from these studies.

Figure 6. The co-words network of World Trade Web (2003-2012)

5. CONCLUSION

The meta-network approach is useful method to model knowledge system from the systems thinking. Network text analysis technique enable the analyst to extract meta-network by integrating ontology definition with text mining. AutoMap software as a tool for NTA, the efficiency of meta-networks modeling depends on the quality and speed of text preprocessing and encoding. Once the analyst determines coding rules, the text can be quickly converted into network data file.

However our goal is to develop a more automated system which integrates data-to-model processing serving as an analytical tool in knowledge representation and human knowledge evolution. Therefore we need to further examine our methodology through more applications or bigger dataset. On the other hand, along with the development of computing power and text mining technologies, the further work is to extend our methodology framework integrating with advance techniques in data mining and data visualization**.**

REFERENCES

[1] Pritchard, A. (1969). Statistical bibliography or bibliometrics. Journal of documentation, 25(4), 348-349.

[2] Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences?. Scientometrics, 62(1), 117-131.

[3] Newman, M. E. (2001). Scientific collaboration networks. I. Network construction and fundamental results. Physical review E, 64(1), 016131.

[4] Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical review E, 64(1), 016132.

[5] Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. Proceedings of the National academy of Sciences of the United States of America, 101(Suppl 1), 5200-5205.

[6] Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. Information processing & management, 35(3), 401-420.

[7] Egghe, L., & Rousseau, R. (2002). Co-citation, bibliographic coupling and a characterization of lattice citation networks. Scientometrics, 55(3), 349-361.

[8] Mairesse, J., & Turner, L. (2005). Measurement and explanation of the intensity of co-publication in scientific research: An analysis at the laboratory level (No. w11172). National Bureau of Economic Research.

[9] Mattsson, P., Laget, P., Nilsson, A., & Sundberg, C. J. (2008). Intra-EU vs. extra-EU scientific co-publication patterns in EU. Scientometrics, 75(3), 555-574.

[10] Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. Scientometrics, 22(1), 155-205.

[11] Leydesdorff, L., & Welbers, K. (2011). The semantic mapping of words and co-words in contexts. Journal of Informetrics, 5(3), 469-475.

[12] Krackhardt, D., & Carley, K. M. (1998). PCANS model of structure in organizations (pp. 113-119). Carnegie Mellon University, Institute for Complex Engineered Systems.

[13] Tolman E.C. (July 1948). "Cognitive maps in rats and men". Psychological Review 55 (4): 189–208.

[14] Steyvers, M., & Tenenbaum, J. B. (2005). The Large‐Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Cognitive science, 29(1), 41-78.

[15] van Atteveldt, W. (2008). Semantic network analysis. Techniques for Extracting, Representing, and Querying Media Content. Charleston: BookSurge.

[16] Leydesdorff, L., & Welbers, K. (2011). The semantic mapping of words and co-words in contexts. Journal of Informetrics, 5(3), 469-475.

[17] Popping, R. (2003). Knowledge graphs and network text analysis. Social Science Information, 42(1), 91-106.

[18] Diesner, J., & Carley, K. M. (2004). Using network text analysis to detect the organizational structure of covert networks. In Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference.

[19] Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. Causal mapping for information systems and technology research: Approaches, advances, and illustrations, 81-108.

[20] Tambayong, L., & Carley, K. M. (2012). Network Text Analysis in Computer-Intensive Rapid Ethnography Retrieval: An Example from Political Networks of Sudan. Journal of Social Structure, 13.

[21] Carley, K. M., Columbus, D., & Azoulay, A. (2013). AutoMap User's Guide 2013. http://www.casos.cs.cmu. edu/ publications/papers/CMU-ISR-13-105.pdf

[22] Carley, K. M., Pfeffer, J., Reminga, J., Storrick, J., & Columbus, D. (2013). ORA User's Guide 2013. http://www.casos.cs.cmu.edu/publications/papers/CMU-ISR-13-108.pdf