Document Co-citation Analysis using the Concept Lattice

Anamika Gupta
S.S. College of Business Studies, University of Delhi, India
anamikargupta@sscbsdu.ac.in

Shikha Gupta
S.S. College of Business Studies, University of Delhi, India
shikhagupta@sscbsdu.ac.in (corresponding author)

Mukul Bisht
S.S. College of Business Studies, University of Delhi, India
mukul.20529@sscbs.du.ac.in

Prestha Hooda
S.S. College of Business Studies, University of Delhi, India
prestha.20533@sscbs.du.ac.in

Md Salik
S.S. College of Business Studies, University of Delhi, India
salik.20528@sscbs.du.ac.in

Received: 16 July 2023 | Revised: 23 August 2023 | Accepted: 27 August 2023

 DOI: https://doi.org/10.48084/etasr.6201

ABSTRACT

Document Co-citation Analysis (DCA) is a method to identify and analyze the relationships between co-cited documents. In this paper, we attempt to use concept lattice for DCA. Concept lattice is a graph structure given in Formal Concept Analysis (FCA), a branch of mathematics based on the concept and its hierarchy. The experiments are conducted on an extensive repository of citations extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources, having a total of 5,354,309 papers and 48,227,950 citation relationships. In this paper, it is established that the concept lattice supports DCA and helps to identify a set of co-cited documents and their co-citation strength. It also provides navigation to reflect the subset-superset relationship of the co-citations. Further, the concept lattice helps identify the hierarchy among the documents and answers the most relevant queries related to DCA.

Keywords-Document Co-citation Analysis (DCA); concept lattice; co-citation strength; citation dataset

I. INTRODUCTION

A citation is a reference to a document, a book, a publication, or an article. The analysis of such citations to understand the relationships and patterns between the documents, authors, keywords, or journals is termed as citation analysis [1-4]. It helps studying the importance and relevance of the work done by others, studying the trends of research, knowing about evolving areas of research, evolving journals in a specific area, etc. [5-10]. The cited work is the document mentioned in the reference and the citing work is the one which is referencing the cited work. DCA is a method to study the relationship among the documents based on the citations [1-4]. Two or more documents are said to be co-cited when they are referenced by the same document. Co-citation strength measures the strength of the relationship between two documents based on the co-citation frequency [4]. For example, if two documents D and E, are cited by documents A, B, and C, then documents D, and E, are co-cited and have a co-citation strength of 3 (Table I). DCA is a process to discover co-cited documents along with their co-citation strength. The discovered co-citations are examined to find emerging patterns. The process helps to understand the connections and the similarities between different research areas [5-10]. It helps to identify the related research directions, the most influential work done by others, the most cited sub-themes within a particular theme, the evolution of new academic areas, etc. [6-8].

A concept lattice is a graph structure used in FCA to represent the relationships between concepts in a given context [11-13]. FCA is a mathematical field based on the graph structure which has been explored in the area of data analysis, knowledge representation, and information management [11-15]. In FCA, the elements of the graph are called Concepts, the lattice of the Concepts is a hierarchical structure of the Concepts which represents the hierarchy based on shared properties or attributes. A Concept has two components (Extent, Intent), where Extent is the set of objects and Intent is the set of attributes. For example, the concept lattice for the dataset given in Table I is shown in Figure 1. Seven Concepts have been identified, namely (ABC, DE), (ABCD, E), (CE, F), (DEF, G), (D,EG), (E,FG), (C, DEF).

TABLE I. SAMPLE CITATION DATASET

Document	References
A	D, E
B	D, E
C	D, E, F
D	E, G
E	F, G
F	G

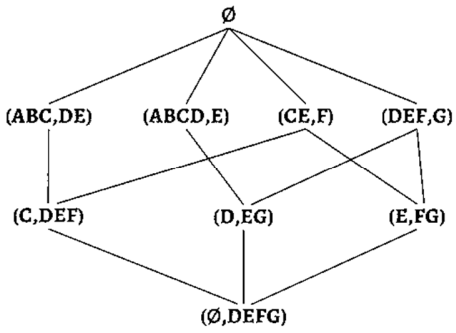

Fig. 1. The concept lattice of the dataset given in Table I.

A. Our Contribution

DCA is an established powerful means for analyzing the relationships between a set of documents. Researchers working in the area of FCA have established the concept lattice structure as a means of analyzing the hierarchy of Concepts based on shared properties. Our contribution is an effort to establish a connection between FCA and DCA. For example, let us consider the dataset given in Table I and the concept lattice generated from this dataset shown in Figure 1. We can observe that a node in the concept lattice has (Extent, Intent) where Extent is showing the citing papers and Intent is showing the cited papers. The documents of the Intent are co-cited documents with the co-citation strength equal to the number of documents in the Extent. For example, the Intent having documents D and E has the Extent having documents A, B, C. Thus, documents D, and E, are co-cited with a co-citation strength equal to 3. This establishes the relationship between concept lattice and document co-citation. To the best of our knowledge, not much work has been done in this area.

B. Why FCA?

The concept lattice structure of FCA is an efficient tool for knowledge and knowledge discovery. The advantage of using concept lattice for DCA is that the concept lattice provides a hierarchical view of the co-citations. It provides navigation to reflect the subset-superset relationship of the co-citations. The rest of the paper is based on an analysis of concept lattice to answer the following queries related to document co-citation on the large citation network dataset downloaded from [16] and identify trends and useful patterns, such as:

- Most cited publications
- List of all co-cited publications.
- Co-citation strength of all co-cited publications.
- Evolution of new academic fields.
- The pattern of co-cited publications year-wise.

II. METHODOLOGY

A. Data Collection and Software Used

There are various software applications available for generating the concept lattice [17]. The citation dataset was extracted from [16] and Charm software [18] was chosen to discover the concept lattice.

B. Pre-processing of the Available Citation Data

Pre-processing of the available citation data was done to provide the input to the concept lattice software, and the output of the software was interpreted for citation analysis. The data were in json format which was converted to text format. Two columns, namely id and references, were extracted. The id represents the publication/paper id and the reference has the list of cited publications. The extracted columns were converted to a file in the format desired by the Charm software. The data were huge to be kept in memory, so they were converted to smaller files and all the operations were performed on individual files and the results were combined. So, Charm software was run on the two columns of the citation dataset and the concept lattice was generated from that. A sample format of those two columns is shown in Table II.

TABLE II. SAMPLE FORMAT OBTAINED FROM THE CITATION DATASET

Paper Id: References
id1: id4, id5, id7
id2: id1, id4, id5
id3: id2
id4: id76, id5
id5: id1, id4, id5

Here, the publication with id1 cites three papers (id4, id5, id7). id1 and id4 are co-cited by id2 and id5. Hence the set (id1, id4) are co-cited documents with strength equal to two. Charm software is used to create a concept lattice from the publicly available dataset in [16] which has more than 6 million papers and citations. The features are extracted from the mentioned dataset and then Charm is run on those two features. No new dataset was created.

C. Generation of the Concept Lattice

When the concept lattice software Charm was run, an appropriate minimum support threshold was chosen after experimentation. The dataset has 3926533 rows (transactions) with 5347758 columns (items) after data cleaning. Discovering concepts from such a large dataset is a difficult task. So, we chose a minimum support threshold to find the frequent concepts. We experimented with different values. Finally, we selected 0.0001%, i.e. 393 transactions, as the minimum support threshold.

D. Exploration of the Concept Lattice

The generated concept lattice was analyzed to understand the thematic structure and the relationships between concepts. The most significant or influential concepts within the lattice were identified. This can help uncover key themes or research areas within the document collection.

E. Result Interpretation

The output is the lattice of concepts (Extent, Intent), where Intent is the list of co-cited publications, and Extent is the list of citing papers. Original Python code was written to extract the relevant information from the lattice, and present the same in informative manner. The Concepts and their relationships within the lattice were interpreted to gain insight into the intellectual structure of the field. The connections between different concepts were explored along with their associated documents to identify emerging research trends, influential works, and potential research gaps.

III. RESULTS

A. Dataset

The data were extracted from [16]. Different versions of this dataset are available in different formats [19-21]. For this project, version 13, which is stored in json format, was chosen. Version 13 of the data has 5,354,309 papers, and 48,227,950 citation relationships. The downloaded file has a size of 17.35 GB.

B. Data Description

The downloaded dataset has the characteristics shown in Table III. After processing the provided dataset, a large number of Concepts were generated. We summarized the results based on the number of publications in Intent (co-cited publications). The Extent of the Concept stores the publication ids of the citing papers. The frequency of the Extent is the co-citation strength. The summary of the results provides interesting insight about that data. The maximum number of publications in a set of co-cited publications is 6, and there is only one such set whose co-citation strength is 431. Similar to this, there are 13 sets of 5 co-cited publications, 67 sets of 4, 251 set of 3, and 997 sets of 2. Each of such set has different co-citation strength. The results are summarized in Table IV.

C. Most Cited Publications

As the number of publications in the set of co-cited publications increases, the co-citation strength decreases. Hence, the first level nodes of the lattice with one publication are the most cited publications. There 7955 such nodes. We can sort them according to their frequency, and get the top cited references. Table V shows one such list of the top 10 most cited papers.

TABLE III. CITATION DATASET DESCRIPTION

Name	Type	Description
id	string	Paper id
title	string	Paper title
authors.name	string	Author name
author.org	string	Author affiliation
author.id	string	Author ID
venue.id	string	Paper venue ID
venue.raw	string	Paper venue name
year	int	Published year
keyword	list of strings	Keywords
fos.name	string	Fields of study
fos.w	float	Fields of study weight
references	list of strings	Paper references
ncitation	int	Citation number
pagestart	string	Page start
pageend	string	Page end
doctype	string	Paper type: journal, book title
lang	string	Detected language
publisher	string	Publisher
volume	string	Volume
issue	string	Issue
issn	string	Issn
isbn	string	Isbn
doi	string	doi
pdf	string	pdf URL
url	list	pdf URL
abstract	string	Abstract
indexedabstract	dict	Indexed abstract

TABLE IV. CO-CITED SETS IN THE CITATION DATASET

No. of publications in a co-cited set	Number of co-cited sets
1	7955
2	997
3	251
4	67
5	13
6	1

TABLE V. TOP 10 CITED PUBLICATIONS

Freq	Title	Year	References
26,873	Deep Residual Learning for Image Recognition	2016	36
24,814	ImageNet Classification with Deep Convolutional Neural Networks	2012	20
24,040	Distinctive Image Features from Scale-Invariant Keypoints	2004	33
20,464	LIBSVM: A library for support vector machines	2011	36
16,999	Long short-term memory	1997	27
14,684	Histograms of Oriented Gradients for Human Detection	2005	15
14,450	Latent dirichlet allocation	2001	15
12,998	Image quality assessment: from error visibility to structural similarity.	2004	29
12,220	Support-Vector Networks	1995	3
12,067	MapReduce: simplified data processing on large clusters	2008	18

D. List of All Co-Cited Publications

Table IV lists down the number of co-cited publications having 2, 3, 4, 5, and 6 publications. These data were derived from the generated lattice. So there are 1329 sets of co-cited publications.

E. Co-Citation Strength

The nodes of the generated lattice have the Extent and Intent fields, as mentioned above. Finding the cardinality of the Extent gives the co-citation strength of the related set. Further, we can find the top *n* number of co-citations having 2, 3, 4, 5, or 6 publications, and their corresponding co-citation strength.

F. Evolution of New Academic Fields

The lattice of publications yielded 9284 nodes, each representing one Concept. There were six levels of nodes where each level has that many numbers of publications in the Intent of the node representing the Concept. The set of 6 co-cited publications had small co-citation strength, so for further analysis, we chose a set of 5 co-cited publications. Out of 13 such sets, we chose the set having maximum co-citation strength. Let us denote that set as *S*. The paper ids of the publications of set *S* are shown in Table VI.

TABLE VI. PAPER IDS OF THE PUBLICATIONS IN THE SET *S*

573696056e3b12023e5186fe
556f622a2401b4b38c23635c
573696f46e3b12023e5f1198
573696026e3b12023e516718
5736986b6e3b12023e730129

Since the paper id is a long string, we are using the first 8 characters of those ids for any further reference. We chose a subset of the complete lattice. Figure 2 shows the subset having the paper ids given in Table VI. Each node of the lattice represents a Concept (Extent, Intent). For the co-cited publications in *S*, the number of publications in the Extent are 748. Since it is not possible to show all the publications of the Extent, a node of the lattice in Figure 2 represents (Intent, *f*) where Intent stores the list of co-cited publications, and *f* is the cardinality of the Extent representing the co-citation strength.

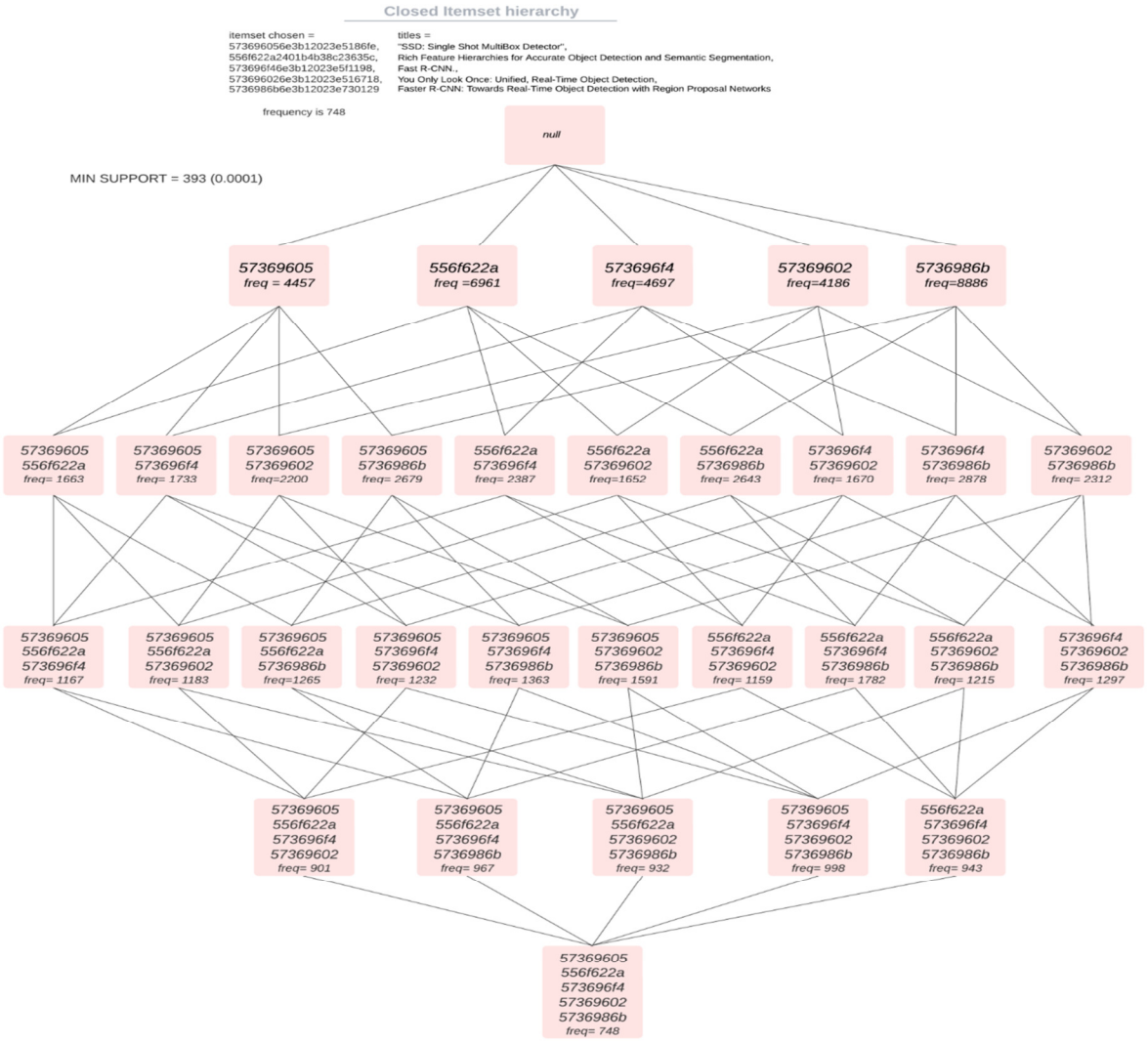

Fig. 2. Lattice of 5 co-cited publications in set *S*.

The lattice in Figure 2 shows the hierarchy of all the co-cited publications derived from the set *S*. The lattice has null root. We can derive the following observations:

- The first level nodes of the lattice of publications in set *S* have one publication, so they will not be categorized in the co-cited publication list. But they provide important information on the most cited publications. The publication with id 5736986*b* is cited by 8886 papers, and is the highest among the given set *S*.
- There are 10 second level nodes which have two co-cited publications. Each of these 10 nodes has different cardinality. The co-cited publications 556*f*622*a* and 573696*f*4 have the highest co-citation strength.
- There are 20 possible sets of 3 co-cited publications which can be derived from the second level nodes. Out of the possible 20 sets, only 10 have the minimum support threshold of 0.0001 %, hence there are 10 third level nodes.
- Publications 556*f*622*a*, 573696*f*4, and 5736986*b* have the highest co-citation strength of 1782.
- There are 5 sets of 4 co-cited publications. The co-cited publications 57369605, 573696*f*4, 57369602, and 5736986*b* have the highest co-citation strength of 998.
- There is 1 set of 5 co-cited publications. The co-cited publications 57369605, 556*f*622*a*, 573696*f*4, 57369602, and 5736986*b* have the highest co-citation strength of 748.
- Analyzing the fields keyword and fos.name, we obtained the word cloud given in Figure 3. Leaving the common words, we can note that words like object, detection, machine, image are prominent. The keywords give an idea of the discipline to which these publications belong.

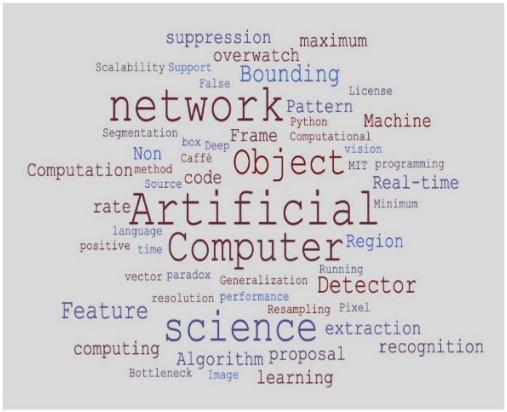

Fig. 3. Word cloud of 5 co-cited publications in *S*.

G. Yearwise Pattern of the Co-Cited Publications

The year-wise analysis of the publications of the set *S* is given in Figure 4. It was found that co-citation strength is increasing every year. This gives an idea of the growing popularity of the research area indicating its continued relevance and influence.

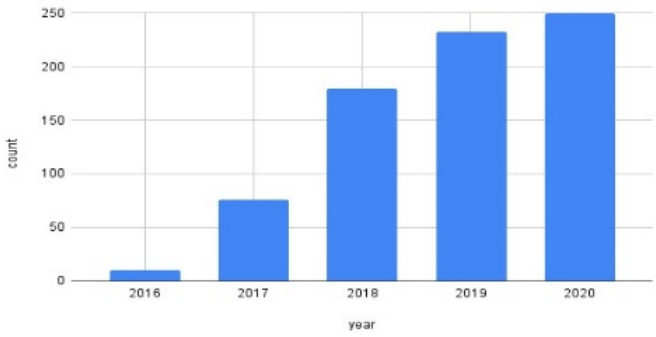

Fig. 4. Year-wise co-citation frequency.

IV. CONCLUSIONS AND FUTURE SCOPE

In this paper, an attempt has been made to conduct Document Co-citation Analysis (DCA) by using the concept lattice structure of the Formal Concept Analysis. It has been observed that concept lattice structure reveals a lot of insight about DCA. This work can be extended to analyze author co-citation and journal co-citation by having the concept lattice of authors and journals, respectively. Further, the concept lattice of keywords can be discovered to find the most relevant research areas.

REFERENCES

[1] Y. Ding, G. Chowdhury, and S. Foo, "Mapping the intellectual structure of information retrieval studies: an author co-citation analysis, 1987-1997," *Journal of Information Science*, vol. 25, no. 1, pp. 67–78, Feb. 1999, https://doi.org/10.1177/016555159902500107.

[2] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973, https://doi.org/10.1002/asi.4630240406.

[3] H. Small, "The synthesis of specialty narratives from co-citation clusters," *Journal of the American Society for Information Science*, vol. 37, no. 3, pp. 97–110, 1986, https://doi.org/10.1002/(SICI)1097-4571(198605)37:3<97::AID-ASI1>3.0.CO;2-K.

[4] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for Information Science*, vol. 24, no. 4, pp. 265–269, 1973, https://doi.org/10.1002/asi.4630240406.

[5] S. Khalid and S. Wu, "Supporting Scholarly Search by Query Expansion and Citation Analysis," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 6102–6108, Aug. 2020, https://doi.org/10.48084/etasr.3655.

[6] S. Khalid, S. Khusro, I. Ullah, and G. Dawson-Amoah, "On The Current State of Scholarly Retrieval Systems," *Engineering, Technology & Applied Science Research*, vol. 9, no. 1, pp. 3863–3870, Feb. 2019, https://doi.org/10.48084/etasr.2448.

[7] A. J. Singh and S. Ravikumar, "Newspaper Citation in Scholarly Publications: A Study on Financial Times Newspaper during 2001-2010 as reflected in Web of Science," *Library Philosophy and Practice (e-journal)*, Feb. 2018, Art. no. 1630.

[8] S. Khalid, S. Wu, and F. Zhang, "A multi-objective approach to determining the usefulness of papers in academic search," *Data Technologies and Applications*, vol. 55, no. 5, pp. 734–748, Jan. 2021, https://doi.org/10.1108/DTA-05-2020-0104.

[9] X.-Y. Liu and B.-C. Chien, "Applying Citation Network Analysis on Recommendation of Research Paper Collection," in *Proceedings of the 4th Multidisciplinary International Social Networks Conference*, New York, NY, USA, Apr. 2017, pp. 1–6, https://doi.org/10.1145/3092090.3092138.

[10] Z. A. Shaikh, "Keyword Detection Techniques: A Comprehensive Study," *Engineering, Technology & Applied Science Research*, vol. 8, no. 1, pp. 2590–2594, Feb. 2018, https://doi.org/10.48084/etasr.1813.

[11] M. Kaytoue, S. O. Kuznetsov, and A. Napoli, "Revisiting Numerical Pattern Mining with Formal Concept Analysis," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011, pp. 1342–1347, http://doi.org/10.5591/978-1-57735-516-8/IJCAI11-227.

[12] U. Priss, S. Polovina, and R. Hill, Eds., *Conceptual Structures: Knowledge Architectures for Smart Applications: 15th International Conference on Conceptual Structures, ICCS 2007, Sheffield, UK, July 22-27, 2007. Proceedings*, vol. 4604. Berlin, Heidelberg, Germany: Springer, 2007.

[13] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*. Berlin, Heidelberg, Germany: Springer Science & Business Media, 2012.

[14] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, "Mining Minimal Non-redundant Association Rules Using Frequent Closed Itemsets," in *Computational Logic — CL 2000*, London, UK, 2000, pp. 972–986, https://doi.org/10.1007/3-540-44957-4_65.

[15] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Pruning closed itemset lattices for association rules," in *BDA'1998 international conference on Advanced Databases*, Hammamet, Tunisia, Oct. 1998, pp. 177–196.

[16] "Citation Network Dataset: DBLP+Citation, ACM Citation network." https://www.aminer.org/citation.

[17] "FCA Software." https://upriss.github.io/fca/fcasoftware.html.

[18] M. J. Zaki, "CHARM Algorithm." Jul. 14, 2022, [Online]. Available: https://github.com/zakimjz/CHARM.

[19] J. Tang, D. Zhang, and L. Yao, "Social Network Extraction of Academic Researchers," in *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, USA, Jul. 2007, pp. 292–301, https://doi.org/10.1109/ICDM.2007.30.

[20] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, Mar. 2009, pp. 807–816, https://doi.org/10.1145/1557019.1557108.

[21] A. Sinha *et al.*, "An Overview of Microsoft Academic Service (MAS) and Applications," in *Proceedings of the 24th International Conference on World Wide Web*, New York, NY, USA, Feb. 2015, pp. 243–246, https://doi.org/10.1145/2740908.2742839.