

Citation Graph Analysis of Publications on “Disease Control through Social Network Surveillance”

Moses Boudourides^{#*1,2}, Andrew Stevens^{*3}, Giannis Tsakonas^{##4}, Sergios Lenis^{**5}

[#]*Department of Computer Science, Haverford College, USA*

¹Moses.Boudourides@cs.haverford.edu

^{*}*SPS, Northwestern University, USA*

²Moses.Boudourides@northwestern.edu

³andrewstevens24@u.northwestern.edu

^{##}*Library & Information Center, University of Patras, Greece*

⁴gtsak@upatras.gr

^{**}*Citrix, Patras, Greece*

⁵sergioslenis@gmail.com

Abstract— This paper presents a study in Data-Network Science on the dataset of publications archived in The Semantic Scholar Open Research Corpus under the area of "Disease Control through Social Network Surveillance." In particular, the dataset consists of 10866 documents (i.e., articles and reviews), retrieved through boolean search, published in the period from 1983, which was the first year of cataloguing such publications, to 2020. Considering the corpus of abstracts (titles included) of these documents and applying the standard LDA Topic Modeling technique, one may find an optimal number of six topics. Unpacking the lists of references of publications in this dataset, the network of our study is the citation directed acyclic graph of publications with nodes/publications labelled according to types, themes, and (dominant) topics of publications. Our aim is to study local network properties in the adjacency sets of nodes with regards to assortativity, triadic relationships amongst connected nodes, and the attributes related to content. In this way, we are analyzing the interplay between semantics and structure in the area of publications on "Disease Control through Social Network Surveillance," by examining how the considered attributes of publications at the semantic level convey the meaning of their affiliations into identifiable local adjacency patterns at the structural level.

Keywords— Scientometric Networks, Citation Graphs, DAG, Nodal Degrees, Triadic Closure and Openness, Assortativity, Topic Modeling.

I. INTRODUCTION

A graph or network is a pattern of pairwise interactions between nodes or vertices, where such interactions are formally represented by links or edges, which, in general (among other typologies), might be directed or undirected, simple (binary) or multiple (weighted), while both nodes and attributes can be labelled by various attributes that may

characterize them. The relatively old mathematical field of Graph Theory is the discipline studying graphs and the relatively new interdisciplinary fields of Social Network Analysis and Network Science are the typical domains for the study of networks.

In Scientometrics, the sub-field of Bibliometrics that measures and analyzes scientific literature (Garfield, 2009), one of the most important purposes is the study of the distribution of citations in and among documents or publications. A citation is a reference, embedded in the body of a document (usually summarized in the section of the bibliography or list of references) and referencing (attributing to) another document, the relevance of which is acknowledged and discussed in the former.

II. CITATION GRAPHS AND CITATION ADJACENCIES

Starting from a *bibliographic dataset* (i.e., a dataset of publications together with their reference lists), a *citation graph* can be composed, nodes of which are the dataset publications and links are the attributions of citations among publications. The resulting citation graph is a *directed acyclic graph* (DAG), because a publication can only cite chronologically earlier publications. Notice that

every node/publication in a citation graph can be either citing or cited by other publications according to the values of the corresponding total degree (i.e., the sum of in-degree and out-degree) of this node/publication. Often in Bibliometrics, Library and Information Science, documents in a bibliographic dataset are generically referred to as sources and, although the common use of the term source is for citations (as references to sources), here we will be using the term source to mean (only) “citing document,” holding the term of citation to “cited documents.” In this terminology and after removing isolated publications (with zero total degree) from a citation graph, every node/publication in this graph can be one of the following three types (adopting the convention that in a citation graph the direction of edges/links goes from citations to sources):

- (A) source, but not citation (zero out-degree, positive in-degree);
- (B) citation, but not source (zero in-degree, positive out-degree);
- (C) source and citation (positive in- and out-degree).

Although this classification (according to the distribution of in- and out-degrees) might appear to be complete, there are certain circumstances eluding its purview. For example, assuming that on the average most publications in a bibliographic dataset might have the length of their reference lists being of the same order, the resulting citation graphs may consist of relatively more nodes/publications of type (A) compared to type (B) and more nodes/publications of type (B) compared to type (C). In such cases, either the citation graph is highly disconnected, when the average in-degree of citations is sufficiently low, or, otherwise, the induced co-citation graph is highly

disconnected (although, in the latter case, the corresponding graph of bibliographic coupling might happen to be well connected). In the extreme case that the number of nodes of type (C) is zero, then the citation graph becomes a bipartite graph with the bipartition of types (A) and (B). Thus, the problem in these circumstances comes from the fact that the linkages among nodes of the same type is very low (actually, only nodes of type (C) can be linked with their own type). Consequently, the citation graph becomes insignificantly assortative (or “homophilic”) with regards to the citation type attribute of nodes.

To what can such high type-assortativity (or low type-mixing) be attributed? To understand this, one might argue both in terms of the inherent endogenous structural patterns in the citation graph and the observed exogenous attributes with which nodes/publications of this graph might be labelled. First, let us consider the case of non-structural nodal labels. These attributions may originate either from the data collection protocols or from the content of documents in the collected corpora. In the former case, typically a bibliographic data set is harvested from a big bibliographic database by querying the occurrence of certain search keywords. Apparently, the latter can be reduced to a number of elementary *themes* such that any node/publication of the resulting citation graph might display. For example, if the search query is a composite statement involving certain basic terms (variables) assembled with the help of certain Boolean connectives, then each of these basic search terms might be a theme-attribute to nodes/publications of the citation graph. Of course, this presupposes that the database from which a bibliographic dataset is extracted by keyword search querying is already categorized in certain fields to which every publication in the database might be attributing.

Normally, such fields on the elements of a database are derived from taxonomies already embedded in the available information about publications (for instance, title or abstract words or keywords given by authors or by any other classification scheme used in the database etc.). Hence, in some way, nodal theme-attributes are always related to the content, i.e., to the semantics of nodes/publications of a citation graph extracted from a bibliographic database. However, there exist other nodal attributes which hinge directly on the contents of such publications or, partially, on the text of their abstracts. For example, *Topic Modeling* is a popular unsupervised machine-learning classification technique which categorizes a corpora of documents (now, the content of all the publications or the content of their abstracts) to certain *topics*, in such a way that each document/publication is associated to a *dominant topic*. So, in short, theme-labels and topic-labels are two attributions of certain non-structural (exogenous) characteristics on nodes of a citation graph.

Now, let us examine possible structural reasons being responsible for the occurrence of high or low type-assortativity of citation graph nodes/publications. A first reason might be sought in the possible clustering of these nodes. After the seminal work of Ronald Burt on structural holes (Burt, 1995), it is known that the mechanism of triadic closure (or closure-producing transitivity) might increase the clustering patterns in a graph. However, in the context of citation graphs having the above defined three types of nodes, the only possible transitivity completion that can be attained is by the brokerage of nodes of type (C) bridging linkages among nodes of type (A) and (B). As we have already seen, complete absence of nodes of type (C) creates a bipartition among nodes of type (A) and type (B), implying zero clustering inside

each of these partitions. Nevertheless, when nodes of type (C) come to play a role by being attached to nodes of either type (A) or type (B), the attained mixing is not always the desired one. There are two extreme ends in the way that nodes of type (C) are articulating linkages with nodes of type (A) or (B). These ends are operating through the following two mechanisms of type assembling:

1. At the one end, all nodes of type (C) might be placed exclusively as adjacent nodes to nodes of type (A) or (B) in such a way they are completely subordinated by the latter creating a configuration of segmented ego-nets (with egos being nodes of type (A) or (B)), which (ego-nets) are not linked to each other. In this case, what increases is the mixing of nodes of type (A) with nodes of type (C) or of nodes of type (B) with nodes of type (C), but not the mixing among nodes of type (A) and (B).
2. At the opposite end, each node of type (C) might be bridging a node of type (A) with a node of type (B). The more often this end occurs, the higher the indirect (as mediated by nodes of type (C)) clustering among nodes of type (A) and (B) can be attained and, at the same time, the higher the overall mixing (or disassortativity) of nodal types might be achieved.

What happens in mechanism (I) is that triadic closure occurs exclusively around nodes of either type (A) or type (B), while in mechanism (II) nodes of (C) might create structural holes (triadic incompleteness) among nodes of types (A) and (B). Motivated by these two extreme mechanisms, we are introducing the following four new nodal degrees. For this purpose, let us consider a node u of the citation graph (DAG) $G = (V, E)$. By

convention, the direction in the edge-pairing $(u, v) \in E$ in a citation graph G is interpreted as node/publication v is citing node/publication u (i.e. the direction of links in G goes from citation to source). Moreover, the *in/out-neighbors* of u are denoted as follows:

$$N_{\text{in}}(u) = \{v \in V: (v, u) \in E\},$$

$$N_{\text{out}}(u) = \{v \in V: (u, v) \in E\},$$

and, thus, the *in/out-degrees* of u are:

$$\text{in-degree}(u) = |N_{\text{in}}(u)|,$$

$$\text{out-degree}(u) = |N_{\text{out}}(u)|,$$

where, for a set X , $|X|$ denotes the *cardinality* of X , i.e., the number of elements of X . Furthermore, the symbol “TC” stands for **triadically closed** and “TO” for **triadically open**.

- The **TC-in-adjacency set** of u , denoted as $\text{TC}_{\text{in}}(u)$, consists of all the in-neighbors v of u having all of their in-neighbors to also be in-neighbors of u (which is a case of triadic in-adjacency completion). Symbolically,

$$\text{TC}_{\text{in}}(u) = \{v \in N_{\text{in}}(u): N_{\text{in}}(v) \subseteq N_{\text{in}}(u)\}.$$

Moreover, the **TC-in-degree** of u is defined as

$$\text{TC-in-degree}(u) = |\text{TC}_{\text{in}}(u)|,$$

i.e., the $\text{TC-in-degree}(u)$ is equal to the number of those in-neighbors of u , which inherit the same in-adjacency property to their own in-neighbors. Thus,
 $\text{TC-in-degree}(u) \leq \text{in-deg}(u)$.

- The **TC-out-adjacency set** of u , denoted as $\text{TC}_{\text{out}}(u)$, consists of all the out-neighbors v

of u having all of their out-neighbors to also be out-neighbors of u (which is a case of triadic out-adjacency completion). Symbolically,

$$\text{TC}_{\text{out}}(u) = \{v \in N_{\text{out}}(u): N_{\text{out}}(v) \subseteq N_{\text{out}}(u)\}.$$

Moreover, the **TC-out-degree** of u is defined as

$$\text{TC-out-degree}(u) = |\text{TC}_{\text{out}}(u)|,$$

I.e., the $\text{TC-in-degree}(u)$ is equal to the number of those out-neighbors of u , which inherit the same out-adjacency property to their own out-neighbors. Thus,
 $\text{TC-out-degree}(u) \leq \text{out-deg}(u)$.

- The **TO-in-adjacency set** of u , denoted as $\text{TO}_{\text{in}}(u)$, consists of those in-neighbors v of u that they might possess at least one in-neighbor which is not u 's in-neighbor (which is a case of triadic in-adjacency incompleteness). Symbolically,

$$\text{TO}_{\text{in}}(u) = \{v \in N_{\text{in}}(u): N_{\text{in}}(v) - N_{\text{in}}(u) \neq \emptyset\}$$

Moreover, the **TO-in-degree** of u is defined as

$$\text{TO-in-degree}(u) = |\text{TO}_{\text{in}}(u)|,$$

i.e., the $\text{TO-in-degree}(u)$ is equal to the number of those in-neighbors of u , which break the transitivity of the in-adjacency property that they are holding with u . Thus,
 $\text{TO-in-degree}(u) \leq \text{in-deg}(u)$.

- The **TO-out-adjacency set** of u , denoted as $\text{TO}_{\text{out}}(u)$, consists of those out-neighbors v of u that they might possess at least one out-neighbor which is not u 's out-neighbor

(which is a case of triadic out-adjacency incompleteness). Symbolically,

$$\text{TO}_{\text{out}}(u) = \{v \in N_{\text{out}}(u): N_{\text{out}}(v) - N_{\text{out}}(u) \neq \emptyset\}$$

Moreover, the **TO-out-degree** of u is defined as

$$\text{TO-out-degree}(u) = |\text{TO}_{\text{out}}(u)|,$$

i.e., the $\text{TO-out-degree}(u)$ is equal to the number of those out-neighbors of u , which break the transitivity of the out-adjacency property that they are holding with u . Thus, $\text{TO-out-degree}(u) \leq \text{in-degree}(u)$.

Proposition: For any node $u \in V$ in a DAG $G = (V, E)$,

$$\begin{aligned} \text{TC-in-degree}(u) + \text{TO-in-degree}(u) &= \text{in-degree}(u), \\ \text{TC-out-degree}(u) + \text{TO-out-degree}(u) &= \text{out-degree}(u). \end{aligned}$$

The Handshaking Lemma: For any node $u \in V$ in a DAG $G = (V, E)$,

$$\begin{aligned} \text{TC-in-degree}(u) + \text{TC-out-degree}(u) + \\ \text{TO-in-degree}(u) + \text{TO-out-degree}(u) &= 2|E|. \end{aligned}$$

In other words, the higher/lower is the TC-in/out-degree of a node/publication u , the stronger/weaker is the dependence or influence of u from its in/out-adjacency set. Similarly, the higher/lower is the TO-in/out-degree of u , the stronger/weaker is the mixing or the association of u with non-in/out-adjacent nodes in the citation graph, with which u is connected through directed 2-paths. In particular, the extreme case of zero TC-in/out-degree of a non-isolated node u means

that u depends solely on non-in/out-adjacent nodes, which are accessible through directed 2-paths, while zero TO-in/out-degree means that u is sustained exclusively by its in/out-adjacent nodes, which are again accessible through directed 2-paths.

As an example, in order to demonstrate the above introduced degrees, let us display the radius 2 egonet citation subgraph of the paper with id 203619203. Incidentally, this is a paper entitled “The centers for disease control and prevention strive initiative: construction of a national program to reduce health care associated infections at the local level,” which was published in the *Annals of Internal Medicine* in 2019.

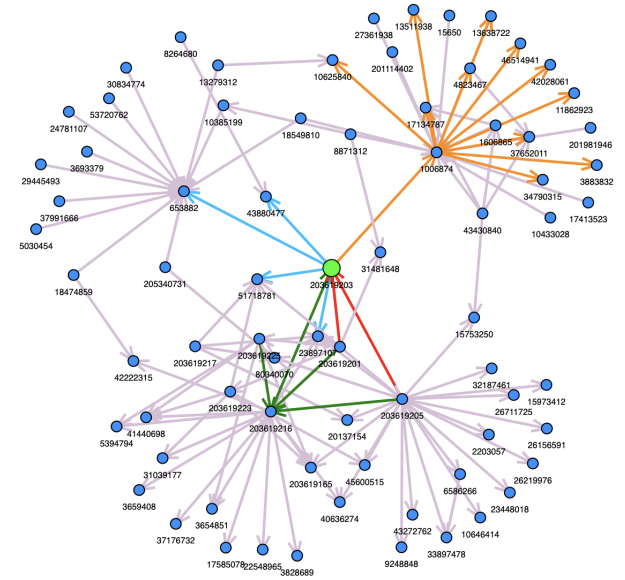


Figure 1: The radius 2 egonet around paper 203619203.

Paper 203619203 (the ego) has in-degree 3 (citing the three papers 203619205, 203619201 and 203619216) and out-degree 5 (being cited by the five papers 1006874, 653882, 23897107, 43880477 and 51718781). Moreover, there exist two predecessors of the ego which are TC-in-adjacent to it (papers 203619201 and 203619205 with their citation links colored red) and one predecessor which is TO-in-adjacent to the ego (paper

203619216 with green colored citation link). Furthermore, four successors of the ego are TC-out-adjacent to it (papers 23897107, 43880477, 51718781 and 653882 with cyan colored inverse citation links) and one successor of the ego is TO-out adjacent (paper 1006874 with orange colored inverse citation links). Notice that the six triadically closed neighbors of the ego either are cited/citing other papers cited/citing by the ego or they do not possess any such corresponding citations. On the other hand, the two triadically open neighbors of the ego are cited/citing papers which are not cited/citing the ego.

III. DATA

The Semantic Scholar Open Source Research Corpus "S2ORC", which includes 81.1 million publications, served as the source for retrieving a subcorpus related to the area of interest (Lo, et.al, 2020). Similar usage of this resource includes A Text Mining Approach to Discovering COVID-19 Relevant Factors (Sastre, et.al. 2020 - DOI:10.1109/BIBM49941.2020.9313149), published in IEEE BIBM. All processing has been performed in python (Rossum, 1995). The rake_nltk (Rose, 2010) and supplementary yake (Campos, 2020) packages were used to extract keywords as features on graph nodes, and each citation is represented as an edge in a networkX construct (Hadberg, 2008).

Attempts to obtain data to be used for analysis included Web of Science (WoS), Scopus, Elsevier, their ICSR Lab, Microsoft Academic Graph (MAG), and finally S2ORC. Efforts in using sources before S2ORC were impeded by various degrees of bureaucracy, paywalls, lack of support, or difficulty in implementation. S2ORC is the most substantial source available while also being the easiest to obtain and mine. It is available as uniformly shuffled shards of compressed JavaScript Object Notation (JSON) line , `jsonl`, files which were downloaded onto the Northwestern Quest Analytics computing platform. Of the 25 fields

available in the data set, `paper_id` was utilized as the unique identifier for a record, and the following fields were associated: `title`, `abstract`, `year`, `outbound_citations`.

Titles and abstracts of the full S2ORC corpus were searched using the boolean phrase "('network' or 'surveillance') and ('disease control' or 'disease network')". The search returned 3,566 publications which we will use as entries of a bibliographic dataset in the area of Disease Control through Social Network Surveillance (from now on referred to by the acronym DCSNS). For each of the source nodes, the `outbound_citations` field (list) was traversed to lookup and obtain outbound citations, totalling 67,572. The corresponding citation graph was constructed, the nodes of which are the paper IDs with edges created on each `paper_id` to `outbound_citation` link. The graph was then pruned of nodes with a degree of one and then zero. This reduced the "source set", publications matching the search phrase, to 3,044 publications. The reference set has 8,467 documents with 645 items existing in both sets (sources referenced by other sources) resulting in a final corpus size of 10,866.

Publication titles and abstracts were then combined from the S2ORC dataset into documents representing each text. The four component search terms/phrases were split into binary fields named `themes`, as indicators for whether they are found in each document. The `LatentDirichletAllocation` (LDA) module of scikit-learn (Pedregosa et al., 2011) generated topic modeling classification. Fields representing the dominant topic from each classification added 6 more fields for each document. These processes result in a single topic and between zero and four themes attributed to each document. All data processing code and resulting data is available in the github repository <https://github.com/zaphodnothingth/dcsns>.

IV. THE DCSNS CITATION GRAPH

The DCNS citation graph extracted from the above described S2ORC dataset is a directed acyclic graph (DAG) composed of 10,852 nodes/publications and 23,173 links/edges, which are either citations or

inverse citations (but never both, since the graph is a DAG). As a DAG, it is not strongly connected (it has as many strongly connected components as the number of its nodes, i.e., 10852). Neither it is weakly connected (it has 25 weakly connected components with the largest weakly connected component composed of 10749 nodes and 23065 edges). The density of this graph is 0.0002 and its transitivity is 0.009. Since plotting a graph of this size would result in a hardly intelligible visualization, what we are displaying below is the 8-core of the citation graph, which is the maximal subgraph that contains nodes of (total) degree 8 or more (Batagelj and Zaversnik, 2003).

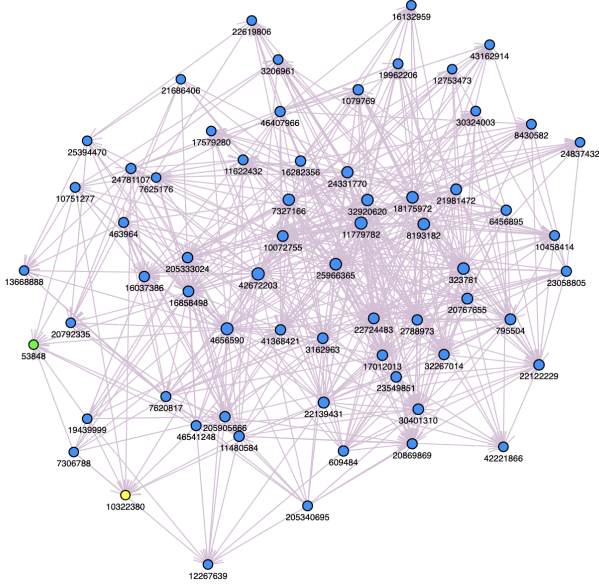


Figure 2: The 8-core of the citation graph

(with nodes colored in 3 Girvan-Newman communities).

V. NODAL ATTRIBUTES OF THE DCSNS CITATION GRAPH

A. Degrees

In the following two diagrams, the boxplots and the correlation matrix among all the degrees of the DCSNS citation graph are displayed:

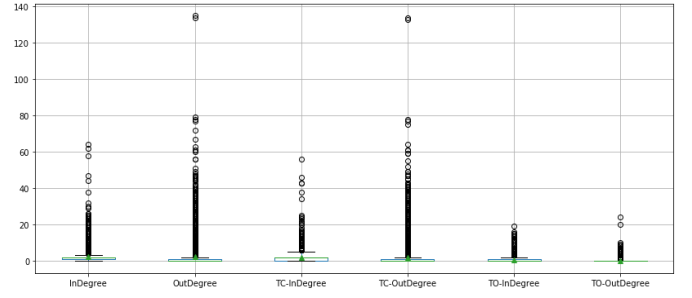


Figure 3: Boxplots of various degrees of nodes of the DCSNS citation graph.

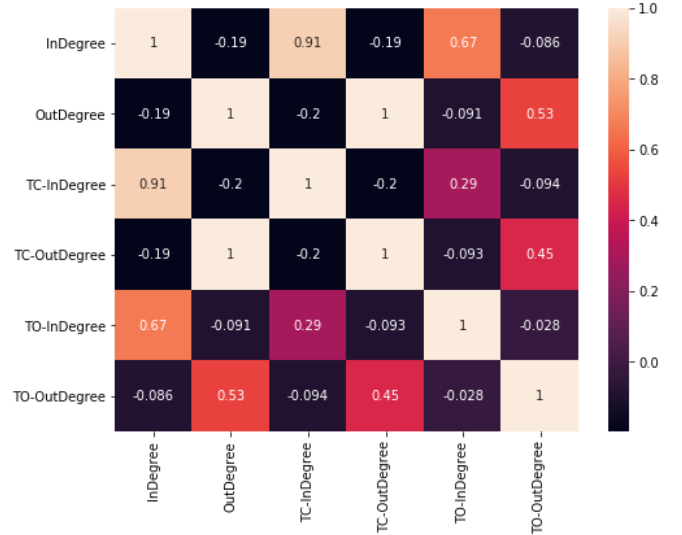


Figure 4: Correlation matrix of various degrees of nodes of the DCSNS citation graph.

B. Types

As we have already specified, the citation graph nodes/publications are grouped with regards to their type. Here are the three types of nodes (where in parenthesis the number of corresponding nodes/publications belonging to each type is given):

- Type (A): Source, not Citation (7,840)
- Type (B): Citation, not Source (2,450)
- Type (C): Source and Citation (562)

This is the first of totally three attributes labeling the nodes of the citation graph. Before examining the other two attributes, let us display the reduced graph of types, when nodes of the citation graph are

aggregated according to the type to which they belong, and correspondingly edges are aggregated among types. Apparently, the reduced graph which is aggregated in this way is a digraph of three nodes (the three types of nodes) and six weighted edges. These edges are the only possible, because the node “Citation - not Source” cannot be citing, because otherwise its in-degree would become positive and this would make it to be a source (but it is not). Similarly, the node “Source - not Citation” cannot be cited, because otherwise its out-degree would become positive and this would make it to be a citation (but it is not).

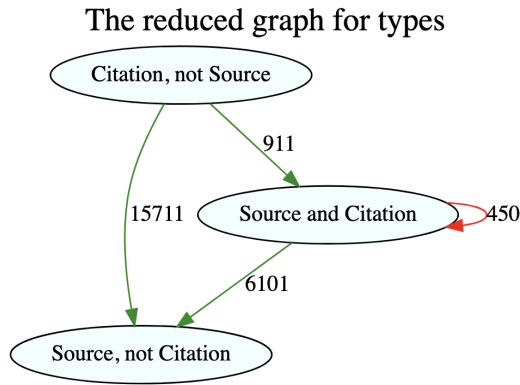


Figure 5: The reduced citation graph for types of publications.

C. Themes

As we have already discussed, the nodes/publications of the DCSNS citation graph were classified into three categories, called **themes**, which were related to terms appearing in the keyword searches employed for the collection of the DCSNS bibliographic dataset from the S2ORC database. Four themes were, thus, identified (in parenthesis, the number of nodes/publications characterized by the corresponding theme):

- Disease Control Theme (3,089)
- Disease Network Theme (354)
- Network Theme (2,243)
- Surveillance Theme (3,663)

Moreover, among the 10,852 nodes/publications of the citation graph, 5,143 of them did not possess any theme categorization, because they appeared as

extra publications inside the reference lists of the primary dataset collected by the aforementioned keyword-searches in the S2ORC database. In addition, among the remaining 5,709 publications classified to these themes, several of them were assigned to more than one theme. In other words, the theme attribute on nodes was non-exclusionary (overlapping). Thus, in order to partition the citation graph nodes into distinct thematically determined groups, we had to consider the set of **combined themes** (according to the existing combinations of occurrences of the above four themes as categories of the citation graph nodes). In this way, the following eleven combinations of themes were identified (in parenthesis, the numbers of corresponding nodes):

- Disease Control Theme (247)
- Network Theme (905)
- Surveillance Theme (1,207)
- Network & Disease Control Theme (548)
- Network & Disease Network Theme (344)
- Network & Surveillance Theme (159)
- Network, Disease Control & Disease Network Theme (2)
- Network, Surveillance & Disease Control Theme (277)
- Network, Surveillance & Disease Network Theme (5)
- Network, Surveillance, Disease Control & Disease Network Theme (3)
- Surveillance & Disease Control Theme (2,012)

Of course, there still exist the 5,143 nodes/publications, not categorized by any of these combined themes. For the sake of completeness, we may categorize them in a twelfth category of combined themes designated as “Reference without Theme.” In this way, the reduced DCNS citation graph of combined themes is the following weighted digraph composed of 12 nodes (combined themes) and 41 aggregated edges:

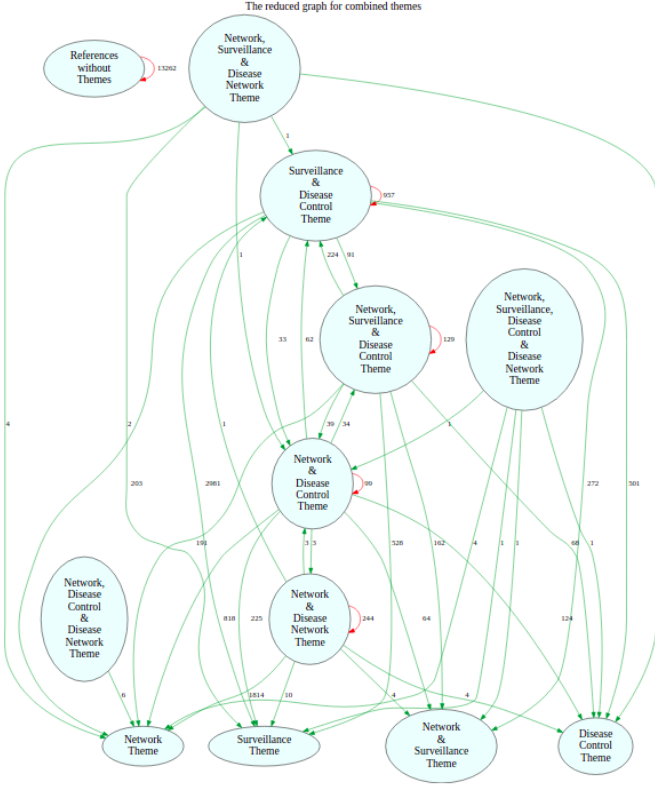


Figure 6: The reduced citation graph for combined themes.

D. Topics

The 10,852 nodes of the DCSNS citation graph are publications which form a corpora of documents. In this corpora, the document of a publication consisted of the abstract of the publication extracted from the S2ORC database, when the publication was in the primary dataset collected by the aforementioned keyword-searches, or simply the existing title words, when the publication was inside the lists of references of the former. This textual corpora was processed through the unsupervised machine learning technique of **Topic Modeling** (using the LDA model) in order to classify the content of the corpora into six topics and to associate to each document (i.e, to each publication) a dominant topic. In this way, one finds an optimal number of six topics producing the maximum topic coherence score in the corresponding topic models with varying numbers of topics. What follows is the list of the six resulting topics (which were interpreted with the given names according to the

top terms in each topic). Notice that in parenthesis is the number of publications, for which the corresponding topic is dominant.

- Disease Networks Topic (719)
- Infectious Diseases Topic (1,245)
- Disease Control Topic (2,636)
- Health-Related Data Topic (2,053)
- Surveillance Topic (2,925)
- Network Models Topic (1,274)

Now, the categorization of the citation graph nodes/papers in one of these six topics becomes a partition in the set of nodes, leaving no node uncategorized and without any overlapping of topics among the nodes. Thus, the reduced DCNS citation graph of (dominant) topics is the following weighted digraph composed of 6 nodes (topics) and 36 aggregated edges:

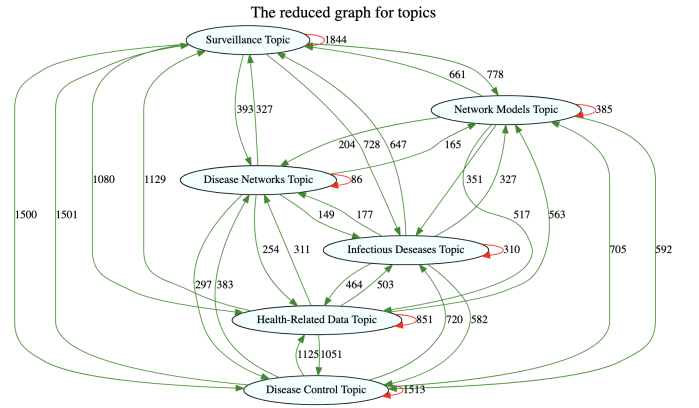


Figure 7: The reduced citation graph for dominant topics.

V. DEGREE AND ATTRIBUTE ASSORTATIVITIES

Here, we are examining all the degree and all attribute assortativity coefficients for the DCSNS citation graphs (Newman, 2003). The results are summarized in the following table:

TABLE I
ASSORTATIVITY COEFFICIENTS OF THE CITATION GRAPH

	Degree assortativity coefficient	Attribute assortativity coefficient
InDegree	0.094	-
OutDegree	0.028	-
TC-InDegree	-0.007	-
TC-OutDegree	-0.001	-
TO-InDegree	0.348	-
TO-OutDegree	0.217	-
Type	-	0.003
Combined Theme	-	0.029
Topic	-	0.019

W. CONCLUSIONS

The DCSNS dataset (from S2ORC) exhibits a very low ratio (5.18%) of citation-sources nodes over the total number of nodes of the citation graph and an even lower ratio (1.94%) of links among citation-sources nodes over the total number of links of the citation graph. As a result, the citation type assortativity coefficient is negligible (0.003).

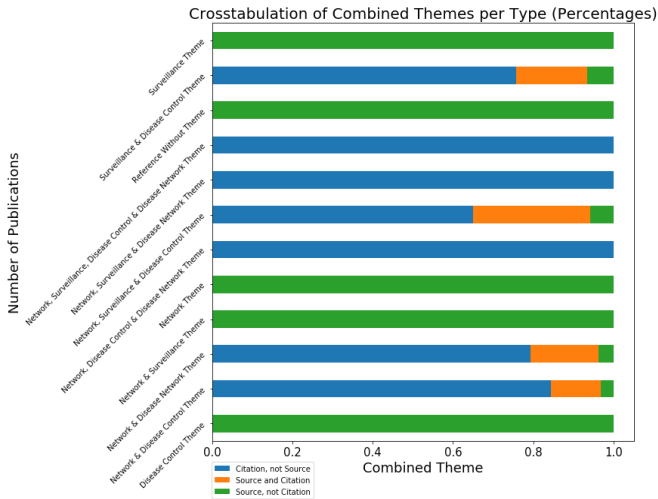


Figure 8: Cross-tabulation of Combined Themes per Type.

On the other hand, among publications of Source and Citation type, there is a relatively high number

(355 out 562, i.e., 63.1%) of publications, which are cross-tabulated within the Surveillance & Disease Control Theme (see Figure 8).

This suggests that by restricting the DCSNS citation graph as a subgraph induced by nodes/publications of the Source & Citation Type (C), one would expect that the resulting assortativity (or homophily) vs. mixing (or disassortativity) of the filtered graph on Sources & Citations would be relatively more significant than what was displayed in the previous table. As a matter of fact, the various assortativity coefficients of the Sources & Citations subgraph are computed to take the following values on the aforementioned subgraph:

TABLE 2
ASSORTATIVITY COEFFICIENTS OF THE SOURCES & CITATIONS SUBGRAPH

	Degree assortativity coefficient	Attribute assortativity coefficient
InDegree	0.158	-
OutDegree	0.336	-
TC-InDegree	0.046	-
TC-OutDegree	0.316	-
TO-InDegree	0.513	-
TO-OutDegree	0.276	-
Combined Theme	-	0.519
Topic	-	0.437

Thus, we conclude that, in the DCSNS citation graph, the attribute of (dominant) topics, at the semantic level, is the one among all attributes of nodes/publications which tends to be predominantly aligned with the citation patterns, which are exhibited at the structural level. Such an alignment between meaning and structure on the DCNS citation graph signifies the fact that the bibliographies of publications in the original dataset of our study were quite focused on the topics, in which Topic Modeling classified the content of publications (in the way that it was represented by their abstracts and titles). In particular, we are observing that it is “Surveillance and Disease

Control” the topic that best captures the structural features of the DCSNS citation graph.

As future directions of our study of the DCSNS bibliography, we intend to consider the temporal citation graph and determine which topics (or possibly themes) might best reflect the citation structures during each time period. Furthermore, a very important issue that we would like to tackle in our future investigations on the DCSNS bibliography is whether the (computational) problem of link prediction might be still contingent on the attribute of the topics or whether link prediction would be controlled exclusively by structural characteristics of the underlying citation graph.

ACKNOWLEDGMENT

This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology.

REFERENCES

- [1] G. van Rossum, Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.
- [2] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008
- [3] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
- [4] Garfield, Eugene (2009). "From the science of science to Scientometrics visualizing the history of science with HistCite software". *Journal of Informetrics*. 3 (3): 173–179.
- [5] Burt, Ronald S. (1995). *Structural Holes: The Social Structure of Competition*. Cambridge: Harvard University Press.
- [6] M. E. J. Newman, Mixing patterns in networks, *Physical Review E*, 67 026126, 2003.
- [7] Rose, Stuart & Engel, Dave & Cramer, Nick & Cowley, Wendy. (2010). Automatic Keyword Extraction from Individual Documents. 10.1002/9780470689646.ch1.
- [8] Sharma, Vishwas B. Rapid Automatic Keyword Extraction algorithm. Retrieved from <https://github.com/csurfer/rake-nltk>
- [9] Campos, R., Mangaravite, V., Pasquali, A., Jatowt, A., Jorge, A., Nunes, C. and Jatowt, A. (2020). YAKE! Keyword Extraction from Single Documents using Multiple Local Features. In *Information Sciences Journal*. Elsevier, Vol 509, pp 257-289
- [10] Python Package Index - PyPI. (n.d.). Python Software Foundation. Retrieved from <https://pypi.org/>