Dynamic Network Analysis of Nuclear Science Literature for Research Influence Assessment

Samrat Chatterjee, Dennis Thomas, Daniel Fortin, Karl Pazdernik, Benjamin Wilson, and Lisa Newburn

Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99354 USA,
E-mail: samrat.chatterjee@pnnl.gov, dennis.thomas@pnnl.gov

Abstract:

Analyzing nuclear science literature via data-driven methods is key for assessing research influence and technology advancements. Indicators of scholarly activities may be buried in publications and collaboration networks over time. Mining for relevant scholarly influence trends in large volumes of text can be computationally challenging; however, open-source information over time can offer opportunities to extract meaningful insights. While network centrality analysis of scholarly research provides topology-based insights, additional emphasis on dynamics associated with information diffusion through these networks is important. This paper represents a step in that direction through the development of a novel dynamic network analysis framework and computational engine to identify key entities and capabilities over time within global scholarly nuclear science collaboration networks. Network theoretic, stochastic simulation, and optimization methods are used to analyze variability in scholarly interactions, influence propagation, and collaboration patterns. A topic-aware influence maximization algorithm is developed to identify key influential authors over time, along with an efficient parallelized implementation to reduce computational costs. A case study using open-source Scopus data with 33,517 published nuclear research papers from 2000-2019 is presented and representative analytic insights are generated. Broad implications of these insights are discussed, and future research directions are also identified.

Keywords: nuclear; collaboration; network; topic; influence

1. Introduction

Activities such as publishing nuclear research and expanding scholarly networks include indicators of research influence – defined as the ability to have the greatest reach (or spread) in a scholarly network – and technology advancements. These types of indicators, often buried in large volumes of technical publication data, may reveal spatiotemporal patterns associated with a dynamic research collaboration landscape. Recent advances in data science methods and computational platforms, combined with nuclear domain knowledge, might provide the appropriate mix of analytic tools that can generate key research influence insights. While data-driven learning technologies are promising for analyzing patterns in large volumes of text, applying these methods to a research influence assessment problem over time can be computationally challenging. Readily available open-source information on research collaborations, such as journal papers and technical reports, can offer insights into an evolving nuclear research and technology domain. However, there is a need for network-theoretic techniques to better exploit time-varying metadata from publications, including authorship, collaboration, and topics of interest.

In this paper, a dynamic network analysis framework is presented for addressing the challenge of identifying key entities and capabilities in nuclear research networks. An entity may be broadly defined as an author, organization, or state. The focus here is on author-level collaboration networks based on open-source publication metadata over a 20-year time period from 2000-2019. Research goals comprise of: (1) identifying key network influencers based on nuclear research topics, (2) comparing network topology-based measures with information diffusion-based outcomes over time, and (3) characterizing influential author collaboration dynamics, including persistence and emergence of connections. Data-driven approaches implemented to meet research goals include network construction, topic modeling, centrality analysis, information diffusion, influence maximization, and temporal dynamics analysis. A case study application is also presented.

The remainder of this paper is organized as follows. We first present a brief background on network analysis of nuclear science literature. Thereafter, we describe the

influential entity identification problem. Next, a modular data-driven dynamic network analysis framework is presented including network construction, topic modeling, topic-aware influence maximization, and temporal dynamics analysis. This is followed by a case study application. Finally, concluding remarks and steps for future research are included.

2. Network analysis of nuclear science literature

Nuclear technology capability development and transfer can include physical items of trade as well as knowledge shared in scientific networks by researchers (Molas-Gallart, 1997). Detecting early signs of proliferation activities, such as based on analysis of text-based data from scientific publications, may provide additional intervention options further "left of boom" (Sheffield, 2020). Analyzing scientific literature for evaluating a State's nuclear activities also leads to diversification of information sources for comparison to generate safeguards conclusions (Feldman et al., 2013). Further, the International Atomic Energy Agency (IAEA) Physical Model represents a consolidated framework for data fusion and analysis that also includes areas of nuclear research and development (Liu & Morsy, 2007). As a result, analysis of research networks may be used to discover collaboration communities and influencers engaged in nuclear fuel cycle related research and development (Iancu, Wilson, Calle, & Gagne, 2018).

Network analysis and text mining techniques have been applied before to analyze scientific networks for nuclear capabilities assessment (Kas et al., 2012; Stewart et al., 2018; Diab et al., 2018; Iancu et al., 2018; Goldblum et al., 2019). Kas et al. (2012) developed a text-mining tool to construct a terminology thesaurus based on nuclear physics research with over 20,000 articles. This tool uses citation networks to identify key entities – individuals, organizations, and nation states – engaged in nuclear research and topics of interest by mapping key terms to capabilities. Weighted citation networks have also been studied in the biomedical domain under the assumption that all cited papers may not have equal influence on the publication of interest (Kim et al., 2018). Weights are typically based on topic similarity and relative importance of a paper in terms of content.

Stewart et al. (2018) describe a data collection, fusion, storage, analysis, and visualization architecture using open-source information for nonproliferation applications. Algorithmic methods that were developed include natural language processing, knowledge and ontology-graph based approaches, link analysis, and geoparsing. Diab et al. (2018) describe a natural language processing approach for identifying key terms that distinguish uranium from other mining processes. Iancu et al. (2018) model publication co-author relations, patent ownership, and organizational affiliation with a directed multigraph. More recently, Goldblum et al. (2019) describe a multiplex network science framework for modeling state proliferation decisions using trade, conflict, alliance, and cooperative agreement networks. A policy-oriented aggregated proliferation metric over time was defined based on measures of centrality and correlation across network layers.

Network analysis methods above typically rely on topology-based measures of centrality over time to identify anomalies and key researchers. However, research influence assessment requires additional emphasis on the role of dynamics associated with these networks. Specifically, research network dynamics in terms of context or topic-based information diffusion and evolving collaboration behavior. This paper represents a step in that direction and presents network-based algorithms and insights from a case study application with focus on dynamic research network analysis.

3. Influential entity identification problem

Identifying influential entities of interest (i.e., author, organization, or state) as well as the evolution of their capabilities, in a computationally efficient manner, within a dynamic scientific research network setting is a challenging problem. Temporal patterns in collaboration networks can contain significant information on sequencing of events and evolution of technology advances. Figure 1 presents a conceptual illustration of research network dynamics and the evolution of influential authors. At a time *t,* nuclear technology capability of an entity of interest may be reflected via the prominence of authors within a nuclear research collaboration network. On the other hand, the interest in pursuing research and development on a particular topic most likely is motivated by the desire for advancing technology capability. As illustrated here at time *t +1*, with an evolving research landscape, the technology capability of an entity might continue to grow as might be evident from increased number of prominent authors within a research network.

The focus of this study is on developing data-driven algorithms and computational pipelines that may be useful for identifying potential influential entities and their capabilities over time. Using a global nuclear research collaboration network over a 20-year time period, three research questions were defined: (1) how to identify key network influencers based on evolving research topics over time? (2) how does network topology-based measures of centrality compare against information diffusion dynamics-based outcomes? and (3) how to characterize influential author collaboration dynamics (in terms of persistence and emergence of connections)? A modular dynamic network analysis framework, developed to address these research questions, is described next.

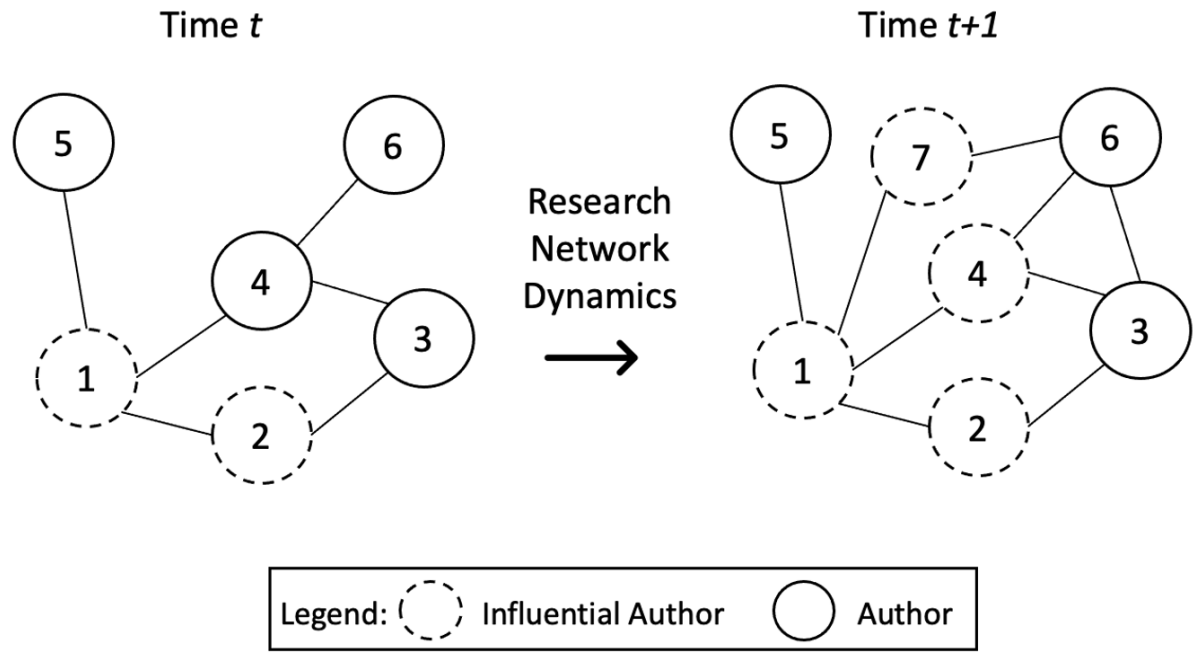

Figure 1: Conceptual Research Network Dynamics Illustration. Research network elements may evolve – grow, shrink, or remain consistent – over time. Nodes in a research network represent authors and edges represent collaboration among authors on a research paper. A node with dotted lines here represents an influential author. In this illustration, the number of influential authors in the network increases from 2 at time *t* to 4 at time *t+1*.

4. Dynamic network analysis framework

Dynamic network analysis is a scientific area of study that fuses concepts from network science, graph theory, network optimization, and stochastic simulation to characterize topology and dynamics associated with networked systems (Newman, 2018; Barabási, 2016; Carley, 2003). Mathematically, a network or graph may be defined as $G = (V,E)$, where V is a set of vertices or nodes and E is a set of edges or links. A network with order n (i.e., number of nodes) is specified by the adjacency matrix, A, an $n \square n$ square matrix where A_{ij} indicates a link connecting node i and node j. Dimensions of temporal effects on networks may be broadly categorized through: (1) node addition/removal, (2) link addition/removal, (3) dynamic flows across networks, and (4) node/link state transitions. Typically, graph analytic methods for influence and capability assessment model individuals, organizations, or events as nodes, and the relations between nodes as different types of links. For example, collaboration networks represent authors as nodes and joint authorship on a manuscript as a link between two nodes. In this study, analytics from collaboration networks over time were generated to address the research questions above associated with identifying key network influencers, comparing topology-based measures of centrality with diffusion dynamics-based outcomes, and influential author collaboration dynamics. Pairing of methods from network science, simulation, and optimization within a flexible computational environment was accomplished to generate insights about key entities and capabilities over time. Computational costs associated with various algorithms were also taken into account and parallel computing paradigms were implemented to accelerate simulation runs.

Figure 2 presents our modular dynamic network analysis framework and computational engine. On the left, a stepwise workflow begins from the top with data collection from Scopus scholarly publications on nuclear research followed by author collaboration network construction. This is followed by topic analysis using metadata information such as title and abstract. The resulting network structures were thereafter subject to dynamic network analysis algorithms that led to characterization of key entities and capabilities over time. On the right, an overview of the computational engine for dynamic network analysis is described in greater detail with three connected modules: (1) topic-aware influence maximization, (2) information diffusion cascade, and (3) temporal dynamics analysis.

Methodological details under each of the framework elements in figure 2 including compute modules within dynamic network analysis are described in sections 4.1 to 4.4. Section 4.1 describes network construction steps including data collection and author collaboration network representation. Section 4.2 focuses on topic modeling based on the Non-negative Matrix Factorization (NMF) algorithm. Next, in section 4.3, topic-aware influence maximization algorithm is discussed which is based on submodular optimization and information diffusion cascade simulation. Finally, section 4.4 includes temporal dynamic

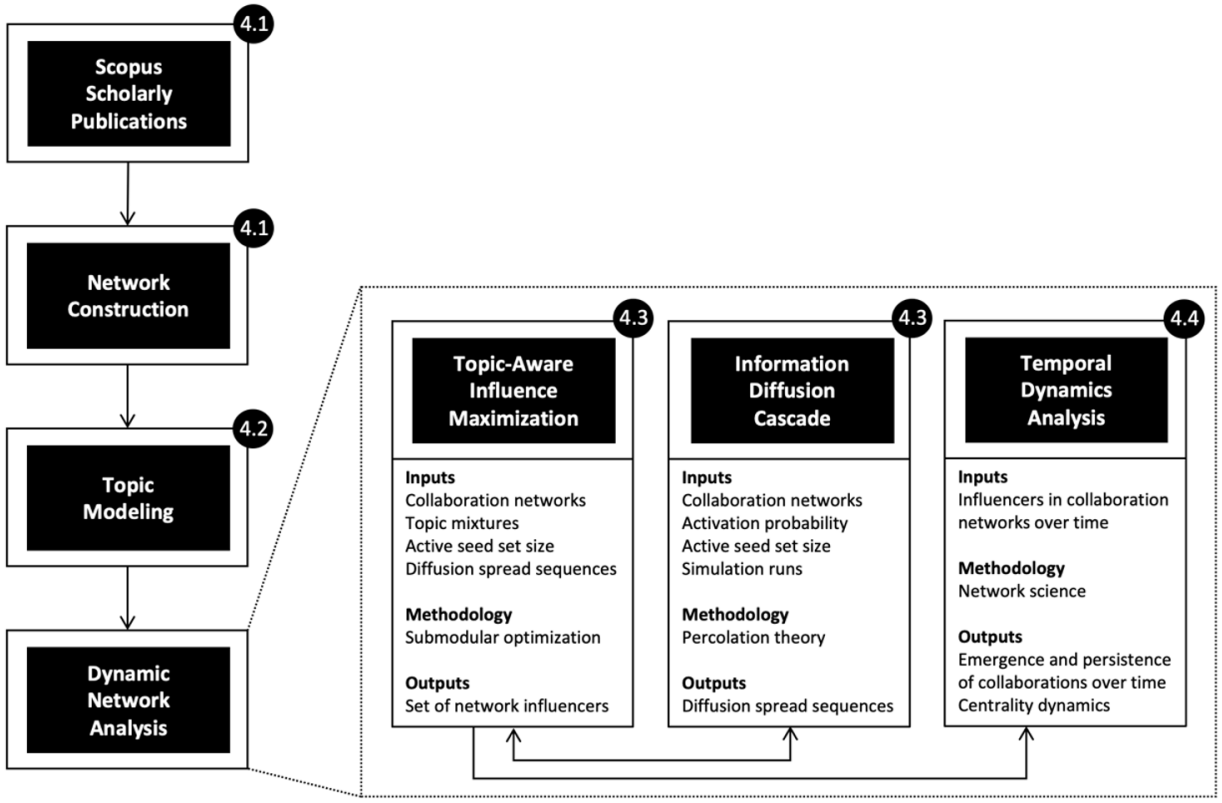

Figure 2: Modular dynamic network analysis framework. Inputs, methodology, and outputs are linked through analytic information flow pipelines within and across computational modules described in sections 4.1 to 4.4.

analysis based on network centrality measures. The overarching goal of this dynamic network analysis framework and computational engine was to generate insights on key entities and capabilities over time using network-theoretic, stochastic simulation, and optimization methods that address variability in scholarly interactions, influence propagation, and author collaboration patterns.

4.1 Network construction

Steps involved in constructing global collaboration research networks are briefly described below.

4.1.1 Data collection

The data for constructing the global collaboration research networks was obtained from Scopus (Elsevier, 2021). We used the *PyScopus* Python package (Zuo, 2023) to search and obtain the Scopus identifiers (IDs) of nuclear related articles, published in the years from 2000 to 2019, and then used the Scopus Application Programming Inferface (API) to directly search Scopus by the Scopus IDs to obtain full metadata information, particularly, author affiliations and abstracts. The list from pyscopus returned the Scopus ID, title, publication name, ISSN, volume, page range, date, doi, citation counts, publication type, author affiliations (name, city, country), author IDs, and full-text links (incomplete) of each article. The articles were identified by searching for terms such as, *nuclear fuel, nuclear energy,* and *nuclear reactor* in the title, abstract and keyword list of each article record in Scopus. For example, the query used for finding the articles published in the year 2000 was specified as: "TITLE-ABS-KEY('nuclear PRE/0 fuel') OR TITLE-ABS-KEY('nuclear PRE/0 energy') OR TITLE-ABS-KEY('nuclear PRE/0 reactor') AND PUBYEAR IS 2000. The same query format is used for other years. The list, however, did not show one-to-one correspondence between each author and their affiliation; this information was obtained from the full metadata information obtained using the Scopus API. Records from journals that published less than 10 nuclear-related articles in a year were not included in the final dataset. The final dataset contained a total of 33,517 records (articles) in .json format.

Within Scopus, each article has a unique numeric ID for each author. For the publications from the years 2000-2019 used in this study, there were 64,312 authors from around the world. We observed that authors over time may change organizational affiliations or use different names (e.g., initials or full name) in their publications. As a result, an author may get assigned multiple numeric IDs. This may lead to additional nodes (representing author IDs) in the collaboration networks over time. We wrote Python scripts to check for cases where an author had multiple IDs but maintained a single affiliation--there were 393 (or 0.61% of 64,312) such instances among all authors over the 20-year time period in

this study. In order to merge or split nodes, additional information is needed that indicates whether an author used different forms of their name or whether two authors may have the exact same name along with affiliation. While additional information related to possible author name resolution can refine our analysis further, it is a non-trivial problem to address for a global-scale network and was outside the scope of this study. Moreover, given the relatively small proportion of instances (0.61%) in this study where an author had multiple IDs but maintained a single affiliation, we did not artificially merge, split, or discard any author information.

4.1.2 Author collaboration network

The author collaboration network is a co-authorship based influence network, where a node represents an author (identified by author id) and an edge represents co-authorship between two authors if they had co-authored at least one article. Figure 3 presents author collaboration network elements and an illustration of network construction when multiple authors collaborate on the same article. The network was constructed using the *NetworkX* Python package (Schult and Swart, 2008). From each record (.json file), we extracted the author names and ids; as well as their institution name, city, and country from their respective affiliations. To quantify the strength or closeness of each pair of co-authors, we calculated Newman-Fowler (NF) weights (Fowler, 2006; Perianes-Rodriguez et l., 2016) using the formula:

$$f_{ij} = f_{ji} = \sum_{p} \frac{a_q}{n_q - 1},$$

where $a_q = 1$, if i and j are co-authors of the same publication p, and 0 otherwise; n_q is the number of authors of publication q. The –1 in the denominator $n_q - 1$ is used to ignore self-links. The NF weights were inverted and assigned as edge weights in the graph, so that edge weights can be interpreted as cost or distance between two co-authors for centrality analysis.

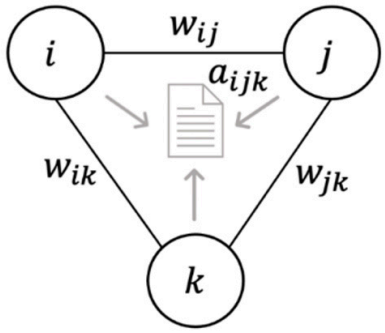

Figure 3: Author collaboration network elements. In this illustration, all authors i,j, and k (nodes) collaborate on the same article a_{ijk}, and inverted Newman-Fowler weights, w_* represent collaboration strength along edges.

Since we are interested in finding authors with maximum influence, we only consider the Giant Connected Component (GCC) of the network for the influence analysis. The GCC is defined as a sub-network that is the largest connected collection of nodes from the original network (Kitsak et al., 2018), and was determined using the *connected_components* function from *NetworkX*. Although it may be possible that influential authors are present outside of a network's GCC, that possibility was not considered for this analysis. Before creating the GCC, we removed densely and weakly connected components of the network by removing edges with weights above a threshold value of 20, which corresponds to an edge representing one paper co-authored by 21 authors. Papers co-authored by many authors will appear as densely connected networks in the author collaboration graph. These networks may represent relatively weak collaborations and can end up being selected in the GCC. Papers with more than 100 co-authors are also present in our dataset. Therefore, we applied the criterion to remove such densely and weakly connected components by removing edges with weights above a threshold value.

4.2 Topic modeling

Topic modeling was performed using state-of-the-art NMF algorithm (Kuang et al., 2015). NMF is a linear algebra based dimension reduction algorithm where the input is a normalized term frequency-inverse document frequency (TF-IDF) matrix and the outputs are two non-negative matrices representing words by topics and topics by documents. We used the NMF outcomes for subsequent influence maximization analysis.

We implemented the NMF algorithm using the *Scikit-learn* Python package (Pedregosa et al., 2011). The title and abstract of each record were combined and used as the text for the topic modeling. The NMF model was fitted with a maximum of 1,000 features extracted from the text of 33,517 records. Common English stop words and corpus-specific words occurring in the text from only one or two records or from at least 95% of the 33,517 records were removed during feature extraction. The features for the NMF model were extracted using *TfidfVectorizer* function in the Scikit-learn package. A total of ten topic categories (or topics) were obtained. The number of topics (categories) was set to ten, which we found to meaningfully classify the records with minimal overlap. The weights of the ten topics discovered by the NMF model were fitted using *sklearn.decomposition.* The NMF module with *Frobenius* norm minimization and regularization, where the L1 to L2 ratio was set at 0.5, and *alpha*, the constant multiplying the regularization terms, was set at 0.1, to avoid overfitting.

Each topic is a group of keywords (features), where each keyword contributes a certain weight to the topic. The top 10 words in each topic were used to analyze the meaning

of the topic. The model outputs were the weight distributions of the ten topics for each record. These weights were normalized using the formula, $v_i/\sum_{i=1}^{10} v_i$, where v_i is weight of topic i for a record. The topic with the highest normalized weight was treated as the dominant topic for that record.

4.3 Topic-aware influence maximization

The topic-aware influence maximization method (Chen et al., 2015) was applied to identify the top 5 authors who can influence the information diffusion of a topic mixture in a network. The topic mixture is defined as a vector, $\lambda = \{\lambda_i | \, i \in [1,10]\}$, where λ_i indicates whether topic i is to be included $(\lambda_i = 1)$ or excluded $(\lambda_i = 0)$ in the activation probability calculations. For example, $\lambda = \langle 1,0,0,0,1,0,1,0,1,0 \rangle$ represents a mixture of topics 1, 5, 7, and 9. Figure 5 illustrates the steps in our research topic-based estimation of activation probabilities for the influence analysis. The activation probability along each edge in the network was computed by taking the dot product of the topic mixture vector and NMF topic weight vector obtained by averaging the topic weights of the co-authored papers along the edge followed by normalization.

The influence maximization method (Kempe et al., 2003) uses an information diffusion model to identify the key authors that have the potential to cause maximal spread of information on a network based on activation probability at an edge. We consider the Independent Cascade (IC) approach (Kempe et al., 2003) as the information diffusion model in this study. Figure 4 illustrates this process of information diffusion cascade and the feedback with influence maximization. Mathematically, the influence maximization problem can be defined as follows. Consider a graph $G = (V, E)$ that abstracts a complex network, where V is the set of nodes V and E is the set of edges $\{(u, v) \mid u, v \in V\}$. There are three types of nodes: (1) active - refers to an author who is influenced in the current step in an iteration path, (2) inactive - refers to an author who was active before and cannot influence others in subsequent time steps in an iteration path, and (3) available - refers to an author who can be influenced in the next step in an iteration path. The edge (u, v) implies that u can influence v. Additional simulation conditions are posed by the choice of diffusion model. For instance, in the IC-based diffusion model, an activated node u has a single chance to activate its available neighboring node v with an activation probability of p_{uv}. Given the possibility of initially activating k nodes, the

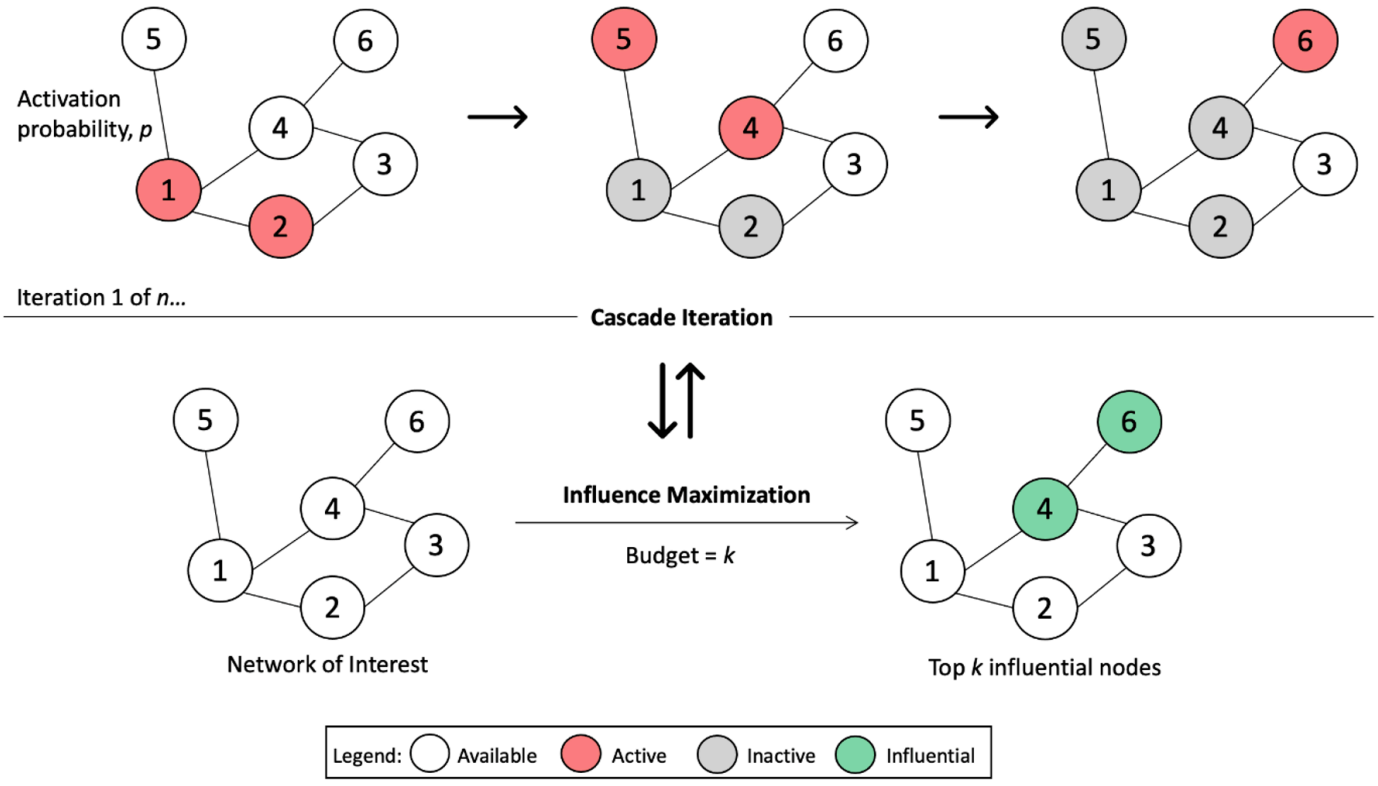

Figure 4: Information diffusion cascade and influence maximization process. In this illustration, a cascade iteration represents a simulation instantiation where an initial set of active nodes result in network impact. Simulation outcomes inform the influence maximization leading to identification of influential nodes, and candidate seed set samples are sent back to the simulation engine via a feedback loop.

```
Input: G, λ                ▷ initial author collaboration network,
                             topic mixture vector
Output: H                  ▷ network with activation probabilities
procedure ACTIVATIONPROBABILITY(G)
    H ← CreateCopy(G)
    for each edge e ∈ Edges(G) do
        a_u, a_v ← GetCoAuthorIDs(e)
        scopusIDList ← GetCoAuthoredPaperScopusIDs(a_u, a_v)
        W ← GetTopicWeightMatrix(scopusIDList)
        m ← RowSize(W)             ▷ number of papers
        n ← ColumnSize(W)          ▷ number of topics
        t ← CreateArray(size = n)
        for j = 1 : n do           ▷ average weights of all papers for each topic
            t_j ← (Σ_{i=1}^{m} W_ij) /m
        end for
        for j = 1 : n do           ▷ normalize the averaged topic weights
            t_j ← t_j / Σ_{j=1}^{n} t_j
        end for
        p ← DotProduct(λ, t)
        if p = 0 then
            RemoveEdgeFromGraph(H, e)
        else
            H ← AddEdgeAttribute(e, p)
        end if
    end for
    return(H)
end procedure
```

Figure 5: Algorithm for activation probability calculation in author collaboration network.

influence maximization problem aims to find a set of ***k*** seed nodes called the seed set ***S***, that when activated result in maximal activations on the network among all possible such sets of ***k*** nodes. The seed size ***(k)*** for all the simulation runs was set at 5 (i.e., to identify top 5 influencers), and the number of iterations ***(n)*** for each IC simulation was set at 10,000. The IC model is a stochastic simulation where the node activation is a random process and the expected cascade size is a random variable. Multiple iterations of the IC are required to obtain the largest expected cascade size. In our simulations, we determine the authors with the largest expected cascade sizes as the top influencers. The IC simulation-based optimization converges resulting in a stable list of top ***k*** authors.

The influence maximization computation was performed using a Message Passing Interface (MPI) Python implementation with parallel computing of the ordinary greedy algorithm (Kempe et al., 2003) using the *mpi4py* Python package (Dalcin et al., 2005). A Monte Carlo loop was implemented for 10,000 iterations to compute the expected spread in activations. A key input in the information cascade and influence maximization process described above is the characterization of activation probability (p) that models the likelihood of information flow across networked entities. Figure 5 illustrates the steps in our research topic-based estimation of activation probabilities for use in influence analysis. The activation probability for a given topic mixture of an author to activate another author was computed using the weight distribution of the NMF topics obtained from the topic modeling results. Starting with collection of co-authored or cited papers corresponding to edges in the network, topic modeling led to a distribution of topic weights per paper. These topic weights were averaged and normalized over the collection, and aggregated topic mixture weights were computed to serve as activation probabilities in the scientific network.

4.4 Temporal dynamics analysis

Analyzing temporal dynamics over scientific networks involves exploring multiple dimensions of information exchange. Below are brief descriptions of the methods implemented to characterize such temporal behaviors.

4.4.1 Collaboration dynamics

The collaboration dynamics of the influencers can be characterized by analyzing their ability to form new collaborations (i.e., emergence) and to maintain old collaborations (i.e., persistence) in time. Mathematically, the number of new collaborations in a given year was estimated by computing the difference in the set of collaborators in that year and the aggregated set of collaborators in prior years. The number of old collaborations in a given year was estimated by computing the intersection of the set of collaborators in that year and the aggregated set of collaborators in prior years.

4.4.2 Centrality dynamics

Centrality dynamics of the influencers was characterized by first computing the topology-based degree and betweenness measures of centrality of influencers in the collaboration network of each year (Barabási, 2012). The change in the centrality measures over time indicates the evolution of key influencers based on topology. The centrality measures (*degree and betweenness*) were calculated using the Python library *NetworkX* (Schult and Swart, 2008). Degree centrality indicates who is well-connected (popular) based on the number of connecting edges. *Betweenness* centrality indicates who controls information flow (or acts as a bridge) between two authors based on how often a node appears on the shortest paths between all other nodes in the network. The edge weights (as defined in the network construction section above) were used in the betweenness centrality calculations.

5. Case study

The case study application addresses the overarching research goals of identifying key influencers, comparing network measures of centrality with information diffusion-based outcomes, and characterizing collaboration dynamics. The results and discussion support the dynamic network analysis framework and computational engine described in Figure 2.

	Whole network				GCC network	
Year	Nodes	Edges	Edges (NF wt ≥ 0.05)	Number of components	Nodes	Edges
2000	3022	67361	5303	1136	42	116
2001	3462	26920	8443	1247	139	606
2002	3522	17219	7553	1091	110	438
2003	3505	27948	8636	1159	56	167
2004	4654	14195	12152	1209	159	573
2005	4866	24636	12921	1383	282	1457
2006	6487	431823	12668	3002	99	504
2007	5731	34945	15789	1644	617	3488
2008	6021	24681	14394	1566	131	525
2009	5890	25055	15423	1417	86	291
2010	6415	31206	16608	1666	111	538
2011	5540	32185	14635	1429	150	836
2012	5171	14128	13597	1105	223	914
2013	6296	28237	17586	1441	177	824
2014	6727	36303	19687	1426	952	4753
2015	7472	62235	20690	1789	276	1151
2016	8285	93836	22426	1998	419	1728
2017	8032	74851	23255	2047	351	1294
2018	10353	784580	26165	2747	934	3526
2019	8289	30427	23218	1523	891	3629

Table 1: Number of nodes and edges in the author collaboration networks.

5.1 Author collaboration data

The author collaboration networks were created based on 33,517 records. Table 1 presents the number of nodes and edges from the overall author collaboration network over time before and after removing edges based on the NF weight criterion. The year 2018 had the largest network, with 10,353 nodes and 784,580 edges. After removing edges with NF weights below 0.05, the number of edges reduced from 784,580 to 26,165. By counting the number of authors affiliated with each country in a year, and averaging the count over the 20-year period, we can find that the highest average number of authors were from China, followed by United States, Japan, and Russia. Figure 6 depicts how the number of authors changed over a period of 20 years for the top ten countries that had the highest average number of authors. The figure indicates that there has been a drastic increase in the number of authors from China after 2011.

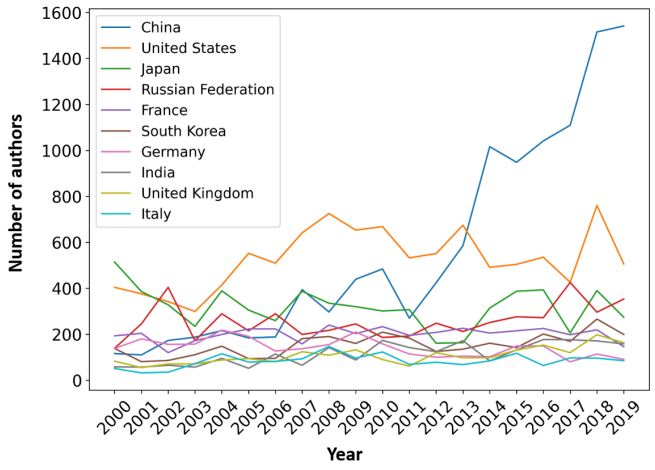

Figure 6: Top ten countries with the highest number of authors from 2000 to 2019.

The GCC networks from the author collaboration networks were computed after removing edges with zero activation probabilities. We observe that in several cases the size of GCC author collaboration networks is roughly an order of magnitude less than the original networks, thereby contributing to higher computational efficiency while capturing significant network connectivity.

5.2 Topic modeling

Figure 7 presents ten topics identified by the NMF model with distinct keywords for each topic along with normalized weight contributions of the top ten keywords in each topic. Each topic is labeled by a topic index number from 1 to 10. In each topic, the top two to three keywords are highly weighted compared to the other keywords; which, suggests that the model is able to identify distinct keywords that capture the overall meaning of each topic. Specifically, the authors' interpretations of the topics based on the keywords were:

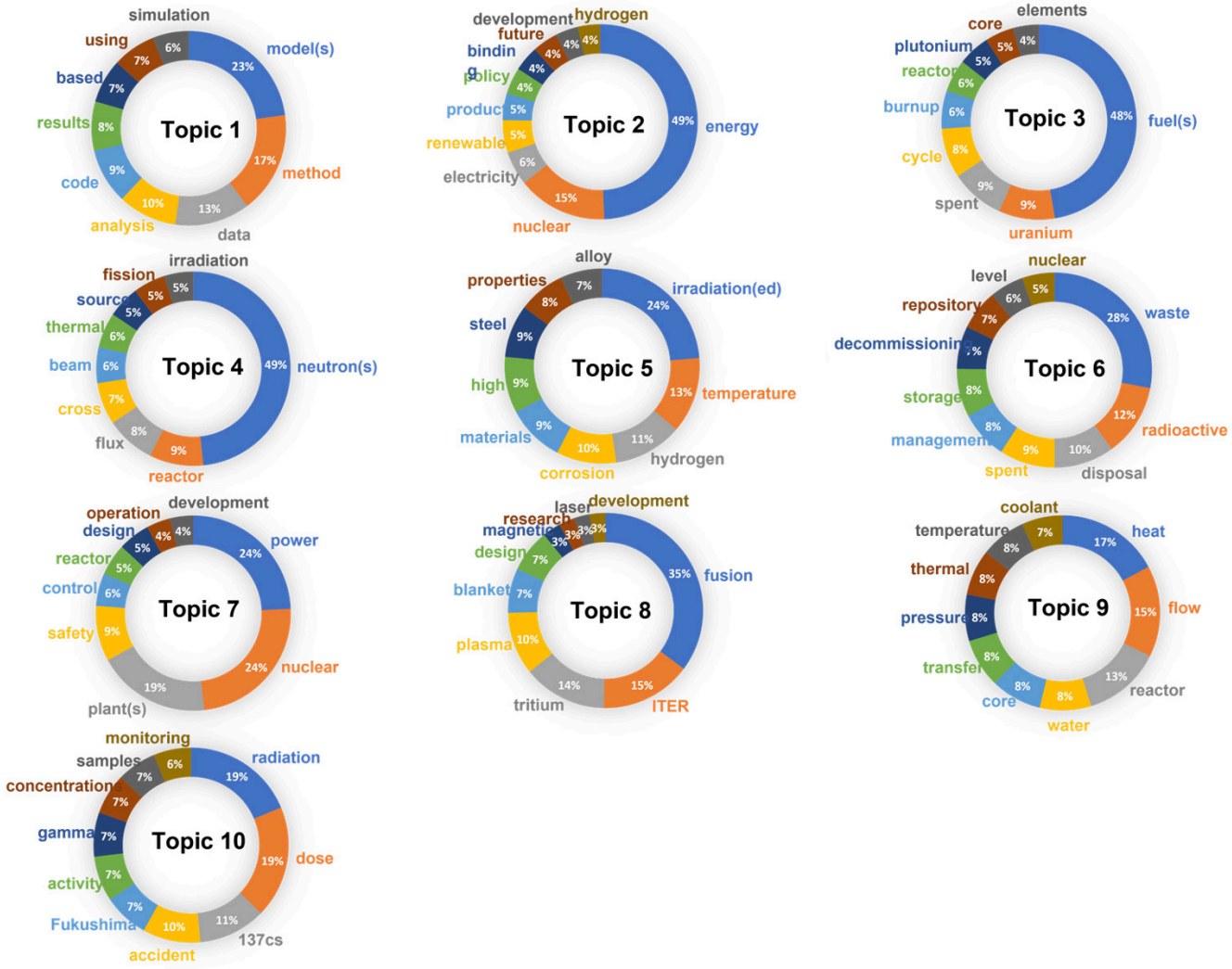

Figure 7: Normalized weight distribution of the top 10 keywords of each NMF topic.

Topic 1: *Modeling and data analysis;*

Topic 2: *Nuclear energy production;*

Topic 3: *Spent uranium in the nuclear fuel cycle;*

Topic 4: *Neutron flux in nuclear reactors;*

Topic 5: *M aterial irradiation at high temperatures;*

Topic 6: *Radioactive waste management;*

Topic 7: *Reactor control in nuclear power plants,*

Topic 8: *Nuclear fusion and ITER (International Thermonuclear Experimental Reactor);*

Topic 9: *Heat flow in nuclear reactors; and*

Topic 10: *Radiation dose from nuclear accidents.*

The word “nuclear” appears in topics 2, 6, and 7, indicating that it was not ignored during the feature extraction and more than 5% of the records do not have the word “nuclear” in their titles and abstracts. These are the records where one or more of the query phrases (*nuclear fuel, nuclear energy, or nuclear reactor*) occurred only in the keyword list and not in the title and the abstract used for the topic analysis. Although the NMF topics have unique meanings, it is typical for a record to belong to multiple topics with varying weights. Figure 8 shows an example where all the topics contribute a non-zero weight to a record, with the highest-weighted topic being topic 6. In this study, the topic with the highest weight is considered the dominant topic for a particular record. Per figure 9, topics 1, 5, 7, and 9 were the top four *dominant topics* among all records in the most recent years from 2016-2019. In addition, these were also the top four dominant topics based on cumulative number of records from 2000-2019 (see Table 2). In this study, these top four prevalent dominant topics were selected as the topic mixture for information diffusion as part of topic-aware influence maximization analysis.

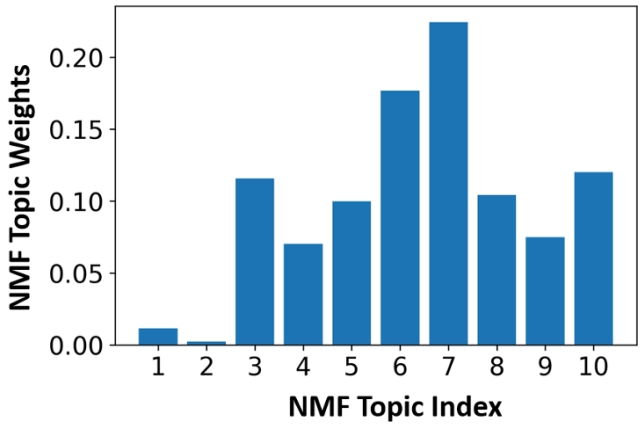

Figure 8: An example showing the non-zero normalized weights of each NMF topic of a record.

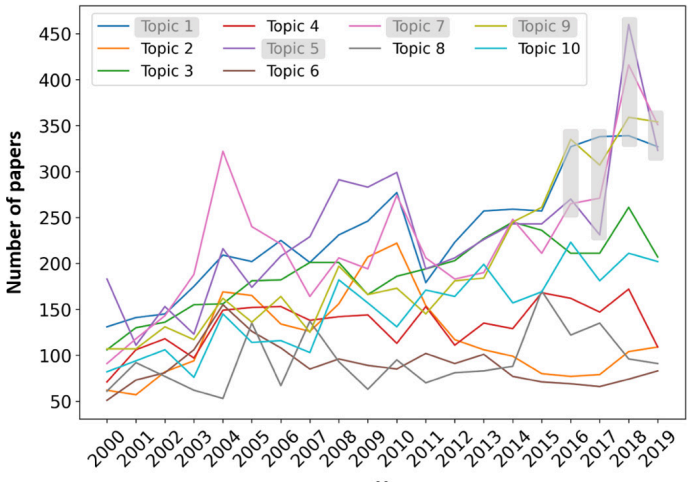

Figure 9: Number of records published every year on each dominant NMF topic. Highlighted topics were the top four dominant topics in the most recent years.

Dominant NMF Topic	Cumulative Number of Records from 2000-2019
Topic 1	4689
Topic 2	2399
Topic 3	3795
Topic 4	2669
Topic 5	4666
Topic 6	1789
Topic 7	4502
Topic 8	1871
Topic 9	3956

Table 2: Cumulative number of records for each dominant NMF topic from 2000-2019. Top four dominant topics and corresponding cumulative number of records are highlighted.

5.3 Influence analysis with collaboration network

As described in the topic-aware influence maximization section, activation probabilities for influence analysis may be computed using a mixture of NMF topic weights yielding different results from the topic-aware influence maximization (TAIM). In this case study, we present TAIM results to identify key influencers over time using an NMF topic mixture. Specifically, we use TAIM to identify the top 5 authors who can influence the diffusion of information pertaining to a mixture of the top four dominant NMF topics (see figure 9 and table 2). These topics are modeling and data analysis (topic 1), material irradiation at high temperatures (topic 5), reactor control in nuclear power plants (topic 7), and heat flow in nuclear reactors (topic 9).

5.3.1 TAIM analysis on author collaboration networks

The TAIM analysis was applied to identify the top 5 influencers in the collaboration network from each year. The computations were performed using a parallelized (MPI) version of the TAIM algorithm. For example, a single TAIM simulation run with GCC of a network with 952 nodes and 4,753 edges using our optimized algorithm converged in about 6 hours using 32 processors; without MPI, the processing time for this run was about 30 times slower or about 7 days. The computational acceleration allowed us to run the TAIM algorithm on large graphs (e.g., 5,000+ nodes and 28,000+ edges) and complete all simulation runs in the order of a few days. Figure 10 shows the top 5 influencers for the years 2000 and 2019. The top influencer in 2000 was from Netherlands, followed by Japan, France, Russia Federation, and Japan again. The top influencer in 2019 was from United States, followed by two from China, and one each from Poland and Italy. Similar analysis was applied on collaboration graphs for years 2001 to 2018 (results not shown here). The change in the top influencers from year to year is indicative of the network dynamics.

The primary advantage of the TAIM analysis is that it helps to identify the top influencers who can diffuse information about the selected topic mixture through their high-spread influence network (collaborators), more efficiently than others. Whether or not they leveraged their positions as top influencers is subject for further investigation. Particularly, it is beneficial to discover if any of the influencers have used their high influence spread in a network to gain new collaborations and prominence in a research area. If they have, then it would be important to analyze the collaboration and publication dynamics to determine whether their positions as the top influencers in a particular year affect or was affected by their collaborations, publication track records, and research impact in past and future years. At the same time, we observe that the influencers may not spread their influence if they or members of their high-spread influence

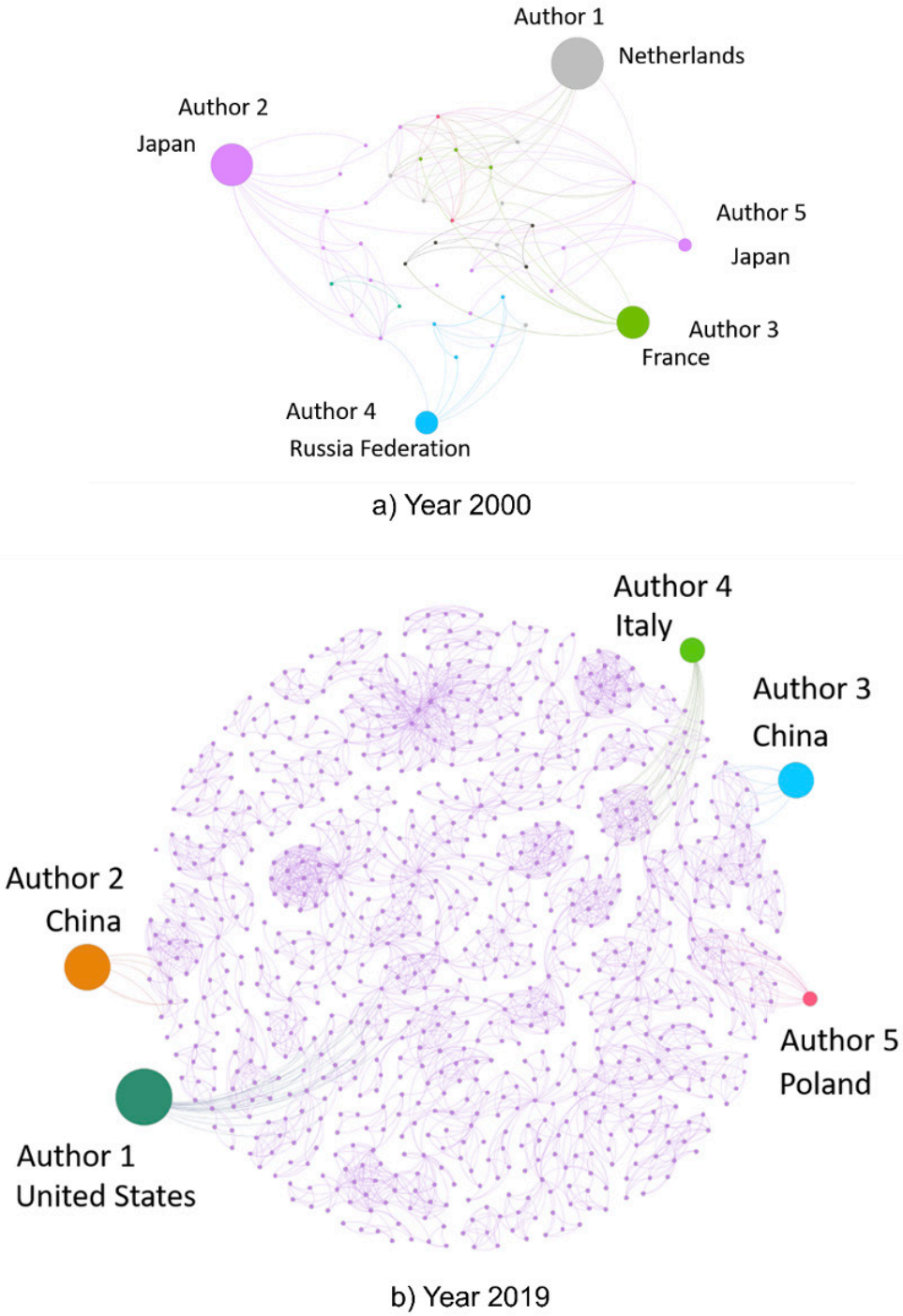

Figure 10: GCC of the author collaboration network, showing the top 5 influencers from years 2000 and 2019 for a mixture of dominant NMF topics 1, 5, 7, and 9. Author IDs have been masked for privacy, node size indicates influence, and node color indicates authors.

network are not actively collaborating or publishing for a long time period.

Among the 99 unique influencers that were identified from 2000 to 2019 (one author was a two-time top 5 influencer), 54 actively published papers for only less than 5 years, 26 were active for 5 to 9 years, 15 were active for 10 to 14 years, and the remaining 4 were active for at least 15 years. Thus, not all influencers from a given year actively published in subsequent years. This may be due to possible factors such as nature and duration of the research, transition of collaborators, or evolving research interests. Influence analysis provides a way to track the collaboration and publication dynamics of influencers pertaining to any topic mixture over a period of time. Furthermore, the ability to generate and monitor scholary influence dynamics may possibly contribute to identifying technology advances and readiness.

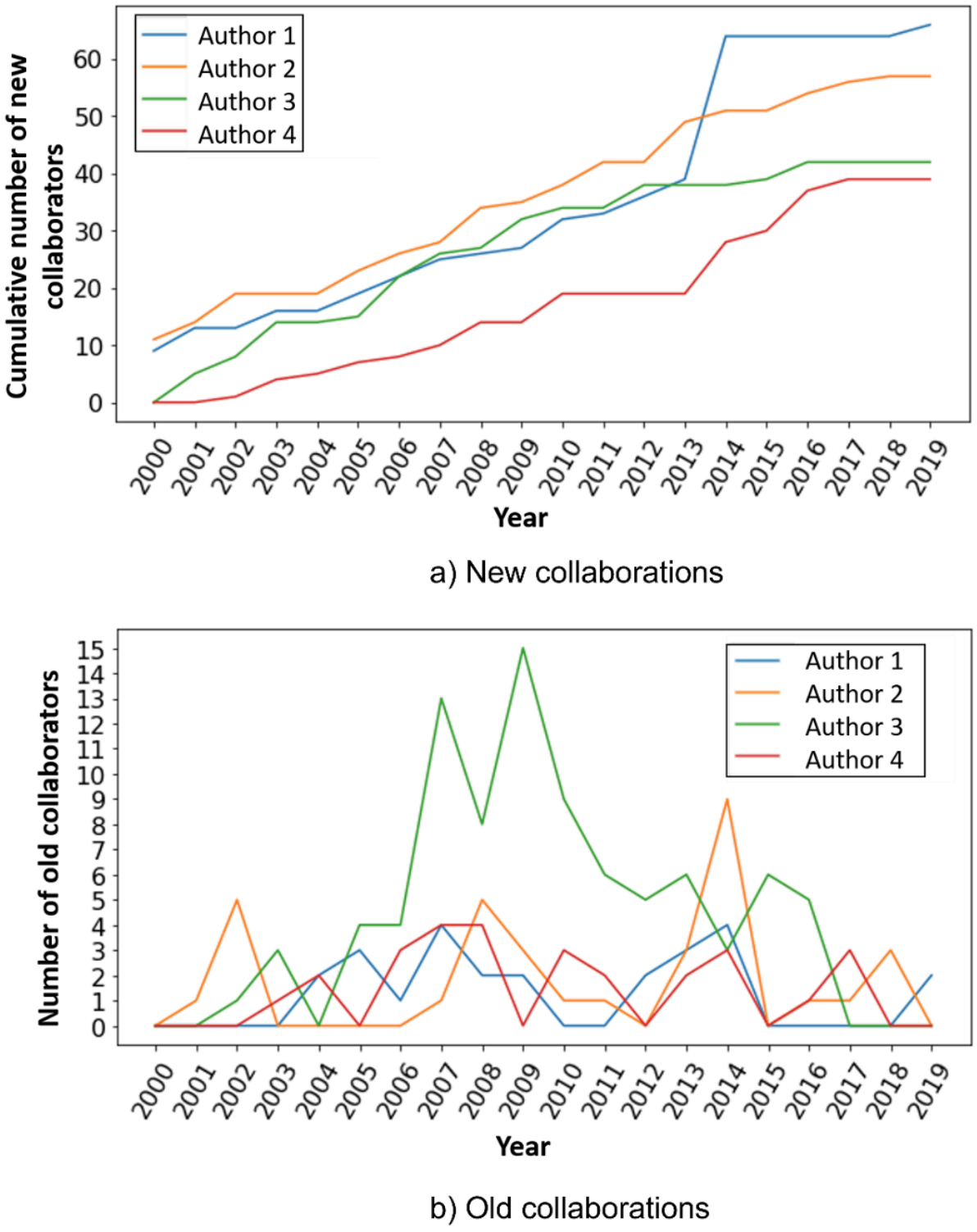

Figure 11: New and old collaborations of influential authors who published papers in at least 15 years between 2000 and 2019. Author IDs are masked for privacy.

5.3.2 Collaboration dynamics of influential authors

The collaboration dynamics of the influencers from 2000 to 2019 were analyzed based on their persistence to maintain old collaborations and the emergence of connections through their ability to form new collaborations. The new collaborators in a given year were those the influencer did not co-author a paper within the years prior to that year.

Figure 11 shows the collaboration dynamics of the four influencers who published papers for at least 15 years. Two of them were among the top 5 influencers in year 2000, one was from 2007, and the other from 2016. Figure 11a presents the rate at which each influencer formed new collaborations over time. While all the influencers have been actively making new collaborations over time, we observe that they exhibit different collaboration signatures over time.

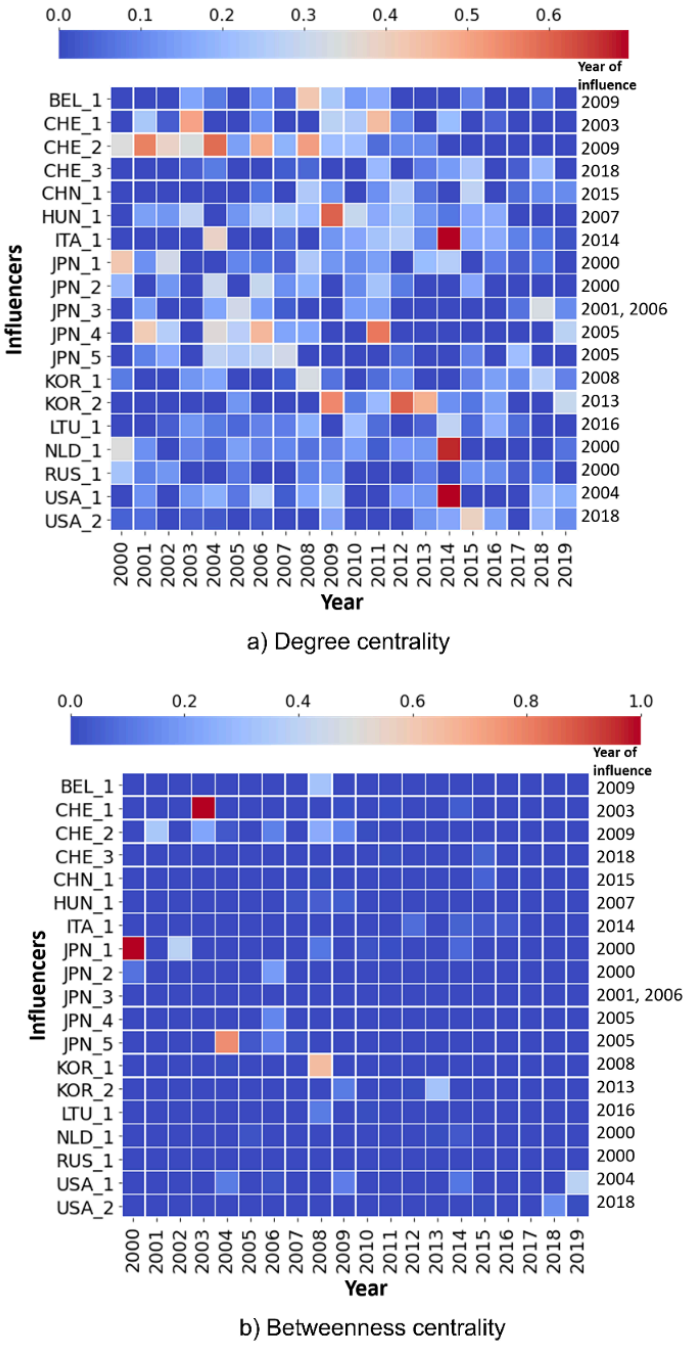

Figure 12: Heat map showing the centrality values of influencers who published at least for 10 years from 2000 to 2019. Centrality values are normalized by the maximum value in a given year. Author IDs are masked by their 3-letter country codes and count index.

From the results presented, it can be understood that it is important for an influencer to work with new collaborators to maintain or increase research productivity and influence. For example, the top influencer from 2000 (author 1 in figure 11) gained 25 new collaborations in the year 2014. Such an influencer could have suddenly gained new collaborations due to contributions on a project with large number of new team members or based on new and multiple projects, or possibly even new leadership positions over multiple projects. Regardless of the factors that might have contributed to the sudden gain (or even decline) in the number of new collaborations, the analytic insights can be used to characterize the growth rate of new collaborations for each author. The old collaboration dynamics in figure 11b indicate that the same influencers continued to maintain about two to five of their old collaborations. The influencer from 2007 (author 3 in figure 11b) in fact shows a strong collaborative behavior with their old collaborators. This might suggest continuing collaborations with known researchers from established areas of expertise and prior history of teaming together.

5.3.3 Centrality analysis of author collaboration networks

Network centrality measures (e.g., degree and betweenness) are commonly used topology-based outcomes to identify authors who are well connected and may control the flow of information in a network. It is interesting to analyze if the top influencers from TAIM analysis (including information diffusion-based outcomes) also have high degree or betweenness centrality values over time. Figure 12 presents the centrality values over time for the 19 influencers who published for at least 10 years. Three influencers, each from 2000, 2004, and 2014, have high degree centralities in the year 2014 (see figure 12a), which are indicative of their high number of collaborations. Two influencers, one from 2000 and another from 2003, also ranked among the top for the highest betweenness centrality (see figure 12b). Since high centrality values over time for influencers are not consistently observed from the heatmap, authors with high-influence spread are not necessarily also the ones with high centrality values. These results signify that a direct correspondence between rankings based on centrality measures and influence spread may not occur since the centrality analysis does not take into account topic awareness and information diffusion. Thus, the TAIM analysis provides novel information that centrality analyses do not provide, which can be valuable for understanding technology advances.

5.4 Discussion

The range of representative analytic outcomes presented in the case study provide insights from global nuclear research collaboration networks that could support influence and capability assessment. These insights are based on modeling and simulation assumptions as well as conditions described under section on dynamic network analysis framework. In light of the overarching challenge problem of identifying key entities and their capabilities over time for influence and capability assessment, four main takeaways from the case study are summarized below:

- Author collaboration networks represent different forms of influence that may lead to varied scholarly research publication patterns in support of technology advancements over time. Collaboration is more direct and indicate evolution of connections in the form of disaggregat-

ed networks over time. The GCC of these networks may serve as a computationally efficient reduced order representation while preserving significant network connectivity properties.

- Within the dynamic network analysis framework and computational engine, additional mixed-topic research areas can also be defined by subject matter experts. NMF based topic modeling algorithm may still serve as useful prior information.
- TAIM accounts for complex topologies and information cascade dynamics simultaneously to identify and rank key influencers by country over time, using topic mixture weights as inputs for computing activation probabilities of influence among authors. This represents an advance in the state-of-the-art of network analysis with nuclear science literature. Parallelized implementation of the TAIM algorithm led to up to 30 times faster compute times in some cases. These key influencers, identified via stochastic simulation-based network optimization, may possess the ability to diffuse information over a network more efficiently than others, and might include authors who are not necessarily those with high values of topological measures of centrality. This is important because using just topology-based measures may miss other influential authors and their collaborators.
- Even among influential authors there is variability in the way they form collaborations over time. Some influencers may choose to continue partnerships with their old collaborators exhibiting persistence of connections, while others may choose to continually seek new collaborators to pursue research goals exhibiting emergence. These collaboration signatures may reveal patterns of scholarly behavior that might help in a more robust assessment of technology advancements and capabilities.

6. Conclusion

The novel data-driven dynamic network analysis framework and computational engine developed in this paper is comprised of three connected computational modules: (1) topic-aware influence maximization, (2) information diffusion cascade, and (3) temporal dynamics analysis. Network theoretic, stochastic simulation, and optimization methods were leveraged to identify key entities and capabilities over time within global scholarly nuclear science research collaboration networks. The analytic insights associated with variability in scholarly interactions, influence propagation, and collaboration patterns over time via network connections can be useful for assessing technology advancements and capabilities. The main element of the dynamic network analysis engine is a topic-aware influence maximization algorithm that enables identification and ranking of key authors who have the potential to influence the spread of information in networks over time. A critical insight from our case study is that influential authors may have unique collaboration behaviors and may or may not exhibit high values of topological measures of centrality. As a result, using just topology-based measures may not lead to a comprehensive assessment of the nuclear research landscape and technology advancements.

The results described using author collaboration networks represent analytic examples to illustrate the value of our dynamic network analysis engine for assessing research influence and technology advancements. Further work may include analysis of authors who collaborate with influencers and their evolution as potential influencers in the future. Future research may further involve expansion of information sources to include other data types such as corporate, trade, patent, and professional affiliation network activities over time to yield even more comprehensive understanding of key entities and capabilities over time. Further research may also include multi-layer network representations with corresponding topology and dynamics to capture importance and influence across network layers; as well as transformer-based topic analysis along with the use of graph representation learning for characterizing uncertainty (due to missing or unobserved information) in research connections. In summary, identifying influential authors can enable estimation of research trajectories in a country, and possibly in collaborating countries over time. Such information can be vital for detecting early signs of proliferation activities and generating safeguards conclusions.

7. Acknowledgments

This study was supported by the U.S. Department of Energy's National Nuclear Security Administration (NNSA). Pacific Northwest National Laboratory (PNNL) is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05–76RL01830.

8. References

1. Barabási A-L; Network Science; Cambridge University Press; 2016
2. Carley K M; Dynamic Network Analysis; 2003
3. Chen S, Fan J, Li G, Feng J, Tan K-L, Tang J; Online Topic-Aware Influence Maximization; Proceedings of the VLDB Endowment; 8; 2015; p 666-677
4. Dalcin L, Paz R, and Storti M; MPI for Python; Journal of Parallel and Distributed Computing; 65(9): 2005; p 1108-1115
5. Diab J, Burr P, and Stohr R; Using Machine Learning and Natural Language Processing to Enhance Uranium Mining and Milling Safeguards; IAEA Symposium on International Safeguards; 51; IAEA; 2018; p 1-7
6. Elsevier; Scopus; https://www.elsevier.com/solutions/scopus; Online; accessed 30 March 2021

7. Feldman Y, Barletta M, Ferguson M, and Norman C; Scientific and Technical Information as a Source for State Evaluation; INMM 54th Annual Meeting; 2013; Atlanta, GA

8. Fowler J H; Legislative Cosponsorship Networks in the US House and Senate; Social Networks; 28; 2006; p 454-465

9. Goldblum B L, Reddie A W, Hickey T C, Bevins J E, Laderman S, Mahowald N, Wright A P, Katzenson E, and Mubarak, Y; The Nuclear Network: Multiplex Network Analysis for Interconnected Systems; Applied Network Science; 4; 2019; p 1–17

10. Iancu I, Wilson B, Calle D, and Gagne D; Detection of Undeclared Nuclear Material and Activities using the Collaborative Analysis Platform; Proceedings of the {IAEA} Symposium on International Safeguards; 2018

11. Kas M, Khadka A G, Frankenstein W, Abdulla A Y, Kunkel F, Carley L R, and Carley K M; Analyzing Scientific Networks for Nuclear Capabilities Assessment; Journal of the American Society for Information Science and Technology; 63; 2012; p 1294-1312

12. Kempe D, Kleinberg J, and Tardos É; Maximizing the Spread of Influence Through a Social Network; Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2003; p 137-146

13. Kim M, Baek I, and Song M; Topic Diffusion Analysis of a Weighted Citation Network in Biomedical Literature; Journal of the Association for Information Science and Technology; 69; 2018; p 329-342

14. Kitsak M, Ganin A A, Eisenberg D A, Krapivsky P L, Krioukov D, Alderson D L, and Linkov I; Stability of a Giant Connected Component in a Complex Network; Physical Review E; 97; 2018; p 012309

15. Kuang D, Choo J, and Park H; Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering; In Partitional Clustering Algorithms; Springer; 2015; p 215-243

16. Liu Z, and Morsy S; Development of the Physical Model; IAEA Safeguards Symposium; 29; 2007; p 1-7

17. Molas-Gallart J; Which Way To Go? Defence Technology and the Diversity of 'Dual-Use' Technology Transfer; Research Policy; 26 (3); 1997; p 367-385

18. Newman M; Networks; Oxford University Press; 2018

19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, and Duchesnay E; Scikit-learn: Machine Learning in Python; Journal of Machine Learning Research; 12; 2011; p 2825-2830

20. Perianes-Rodriguez A, Waltman L, and Van Eck, NJ; Constructing bibliometric networks: A comparison between full and fractional counting; Journal of Informetrics; 10.4; 2016; p 1178-1195

21. Schult D A and Swart P; Exploring Network Structure, Dynamics, and Function using NetworkX; Proceedings of the 7th Python in Science Conferences (SciPy 2008); 2008.

22. Sheffield A; Developing the Next-Generation of AI Systems to Push the Detection of Foreign Nuclear Proliferation Further "Left of Boom"; Countering WMD Journal; 21; 2020; 100-102

23. Stewart I J, Lee A, and ElGebaly A; Automated Processing of Open Source Information for Nonproliferation Purposes; Journal of Nuclear Materials; 46; 2018; p 21-36

24. Zuo S; pyscopus 1.0.3; https://pypi.org/project/pyscopus; Online; accessed 9 May 2023