Analyzing Scientific Networks for Nuclear Capabilities Assessment

Miray Kas
Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA. E-mail: mkas@ece.cmu.edu

Alla G. Khadka
Graduate School of Public & International Affairs, University of Pittsburgh, Pittsburgh, PA. E-mail: asg38@pitt.edu

William Frankenstein
Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA. E-mail: frankenstein@cmu.edu

Ahmed Y. Abdulla
Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA. E-mail: aya1@cmu.edu

Frank Kunkel
Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA. E-mail: fkunkel@cs.cmu.edu

L. Richard Carley
Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA. E-mail: carley@ece.cmu.edu

Kathleen M. Carley
Engineering and Public Policy, Carnegie Mellon University and Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA. E-mail: kathleen.carley@cs.cmu.edu

The capability to build nuclear weapons is a key national security factor that has a profound influence on the balance of international relations. In addition to long-standing players, regional powers and peripheral countries have sought for ways of acquiring and/or developing them. The authors postulate that to express the capabilities, relative positions, and interrelations of the countries involved in the production of nuclear weaponization knowledge, dynamic network analysis provides valuable insight. In this article, the authors use a computational framework that combines techniques from dynamic network analysis and text mining to mine and analyze large-scale networks that are extracted from open theoretical and experimental nuclear research publications of the last two decades. More specifically, they build interlinked, dynamic networks that model relationships of nuclear researchers based on the open literature and supplement this information with text mining to classify the nuclear weaponization capabilities of each publication—of each author, organization, city, and country. Using such a comprehensive computational framework, they are able to (a) elicit the hot topics in nuclear weaponization research, (b) assess the nuclear expertise level of each country, (c) differentiate between established and emergent players, and (d) identify the key entities at various levels such as organization, city, and country.

Introduction

Weapons of mass destruction (WMDs) have been used in various forms throughout history. Although humanity did not harness the power of the atom until the middle of the 20th century, the manner in which nuclear weapons were introduced to the public—with two explosions that leveled the Japanese cities of Hiroshima and Nagasaki, killing more than 200,000 people—quickly established a public dread of these destructive weapons. Since then, the proliferation and testing of nuclear weapons have continued to attract public

Received September 22, 2011; revised February 20, 2012; accepted February 21, 2012

 • Published online 18 April 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.22678

attention. Today, with the possibility of nuclear weapons falling into the hands of terrorists, stopping the proliferation of such destructive weapons and understanding the evolution and dissemination of knowledge on the design and testing of nuclear weapons have arguably become imperative (Ferguson & Potter, 2005; Levi, 2007).

Nuclear engineers and theoretical and experimental scientists working in various fields at many international institutions are actively enhancing the state of nuclear knowledge. An increasing number of nuclear researchers have become engaged in collaborations with their counterparts at other institutions, contributing to the growth and dissemination of nuclear expertise across the world. Developments in both the breadth and depth of nuclear know-how since the inception of nuclear weapons a mere 60 years ago are worthy of attention.

Given the growth and dissemination of nuclear knowledge, collecting useful and relevant information that can serve the needs of policy makers is challenging. Data contained in open-source documents and in journal articles can be used to partially address such needs. Much of the work designed to assist policy makers has focused on information retrieval and on improving the relevance of the extracted information (Pia, 2003), developing question-answering systems (Saracevic & Kantor, 1988a, 1988b), or evaluating information retrieval systems (Tague-Sutcliffe, 1996). In contrast, we ask, "Can the structure of coauthorship/collaboration networks, when examined in conjunction with other open-source information, provide critical insight?" This study represents a first attempt at investigating whether such insights can be gleaned through a case study of nuclear weaponization that relies on a subset of the existing nuclear physics literature.

We note that numerous techniques have been used to study the proliferation of nuclear weapons (Sagan, 2011), including case studies analyzing why individual countries decide to develop nuclear weapons, statistical studies that attempt to specify timelines to weapons development, and studies that focus on specific past events and on specific technology acquisitions (Freedman, 2002; Graham, 1996; Nolan, 1989).

However, most of these studies are historical. We maintain that using a combination of coauthorship/collaboration networks and open-source information may possibly aid in the painting of a clearer picture, mainly by providing insight into the nature of a country's nuclear capability, both historical and current. We note that coauthorship/collaboration networks and bibliometrics are frequently used to assess the state of scientific disciplines, identify critical groups or authors, and assess the state of research in an area (White & Griffith, 1981; White & McCain, 1998). In fact, these techniques have been used to evaluate the diffusion of ideas related to terrorism and the terror-nuclear connection (Reid, 1993). Here we build on this legacy, but focus on the country level rather than that of the individual researcher. We also enhance the state-of-the-art on dynamic social network analysis by unifying it with the assessment and mining of publicly available, unclassified literature from the past two decades.

Over the past two decades, significant research has been conducted in the area of social network analysis that evaluates the publication patterns in various scientific disciplines (Hummon & Doreian, 1989; Newman, 2004; Small, 1999). However, this literature centers on assessing general social network analysis metrics of publications, such as frequency of publication, coauthorship, and publication trends. Most of these analyses exclude text-mining analysis of the contents of these articles, focusing on analyzing metadata exclusively (e.g., authors, publication name, publication date, etc.).

This study utilizes both the metadata and the contents of the research articles. The analysis we perform is two-fold: (a) we identify the institutions involved in nuclear knowledge production based on dynamic network analysis, and (b) we classify the knowledge produced by these institutions using text mining on the contents of these articles, identifying their relation to nuclear weapons capabilities. Entities under investigation include individuals, universities, research centers, national laboratories, international collaborations, and countries. Extraction and analysis of such vast amounts of information can be best accomplished through the use of computational solutions that are able to parse thousands of files with nuclear content and analyze networks with thousands of nodes and edges.

Towards this end, we use Cornell University's online preprint archive (Cornell University, 2011), as our primary data source for a preliminary analysis to demonstrate the utility of the method/computational framework we propose in this article. The arXiv Web site provides a comprehensive coverage of physics publications such as astrophysics, mathematical physics, fluid dynamics, plasma physics, nuclear physics, and many others. From this big digital library, we focus on 20,000+ publications uploaded to arXiv Web site from 1992 to the end of 2010, listed under the "physics–nuclear experiment" and "'physics–nuclear theory" categories. These documents provide us with an extensive research archive that encompasses a subset of the publications produced in the areas of experimental and theoretical nuclear physics over the past two decades.

Using the information available in each of these articles, we are able to determine (a) the principal researchers, (b) the affiliations of these researchers (at organization, city, state, and country levels), (c) the general topic of the article, and by extension, the general nuclear capability associated with it, and (d) the time (month and year) when this knowledge is being produced. Hence, we are able extract multiple, interlinked networks such as publication, coauthorship, and affiliation networks that are indexed by both time and location. By analyzing all of these networks, we are able to identify key players, major influencers, emerging countries, actively researched concepts, and the topics being investigated by each entity and when.

Objectives

To summarize, our objective is to demonstrate how remote assessment of the key entities and nuclear capabilities

DOI: 10.1002/asi

of different regions can be performed. Developing a comprehensive understanding of the activities performed by various entities and their level of expertise can be a key first step to detecting suspicious changes in activity that can lead to potentially threatening situations. A sudden, unexplained change in these activities, or in a country's level of expertise, may be considered a "red flag" that merits further attention.

To achieve our goals, we adopt a computational approach that combines dynamic network analysis with text mining to analyze complex scientific networks extracted from theoretical and experimental nuclear physics publications. Following this approach, we aim to find answers for various critical questions, including

- Who are the key entities engaged in nuclear research (individuals, organizations, cities, countries)?
- How often do researchers from different countries collaborate?
- In this subset of nuclear research, which countries are the major players, which ones are the emerging ones?
- What nuclear capabilities are addressed by recent publications?

Contributions and Impact on Research

Following an interdisciplinary approach that touches upon various aspects of nuclear expertise identification, we make contributions to the scientific literature on multiple fronts. Our contributions can be summarized as affecting three major research areas.

Computational framework. In this article, we craft a computational framework that consists of a unique combination of techniques from the fields of dynamic network analysis, text mining, and nuclear subject matter expertise. We contribute to the literature by providing a distinctive computational framework for the use of researchers from various fields who are willing to work on complex, dynamic networks and merge content analysis and field expertise with dynamic network data. The set of tools we use in this article are software tools and datasets that are publicly available at no cost to researchers, making our methodology and results repeatable.

Dynamic network analysis. In this article, we utilize large-scale scientific networks that are extracted from open-source research publications to understand how the nuclear research community works, yielding both social and geopolitical conclusions in the process. To the best of our knowledge, this study is the first study to apply dynamic network analysis techniques on open-source research literature to assess the nuclear weaponization capabilities of countries remotely, introducing a novel application area for the field of dynamic network analysis.

Our approach to dynamic network analysis also significantly differs from well-known dynamic network publications that cover small groups (e.g., classical examples like Sampson's Monastery; Sampson, 1969) due to the size of our dataset. It also differs from other large-scale dynamic networks studies that only focus on observing/modeling the structural properties of these networks and exclude the unique spatiotemporal and geopolitical conclusions that can be derived from these studies (e.g., Barabási & Albert, 1999; Newman, 2004).

Representing nuclear capabilities in dynamic social networks. Nuclear research, especially research that can be related to weaponization, is a dynamic field that attracts significant attention from policy makers and researchers both in the social sciences and in technical fields. However, there exists a significant gap between the social and technical researchers' perceptions of the field. In this article, we make a unique contribution that is geared towards bridging this gap. In particular, we provide a method of representing technical nuclear weaponization knowledge as standalone entities (nodes) in social networks. Therefore, our method demonstrates how technical nuclear experts can contribute to interdisciplinary research projects (e.g., dynamic network analysis on complex digital coauthorship networks) that do not fall primarily in their field. Our study provides the first example of the use of such nuclear-specific knowledge. To further illustrate this, in a subsequent section provide a more detailed description of our attempt to represent nuclear weaponization capabilities in a dynamic social network.

Paper Organization

We structure the rest of the article as follows. In the following section, we discuss how we can classify nuclear capabilities into various processes and stages and show how we can represent nuclear capabilities in dynamic networks. Then we discuss our dataset, elaborate on our computational framework, report our results, and discuss potential directions for future research. In the final section, we present our conclusions and highlight our key findings.

Representing Nuclear Weaponization Capabilities in Dynamic Networks

"Nuclear research" serves as an umbrella term that covers numerous research fields, from theoretical physics to laser isotope separation to advanced nuclear reactor technologies to nuclear safety. Any nuclear program will inevitably entail research into some of these fields (and many others). There are different ways of structuring a nuclear weaponization program, and even of developing a nuclear weaponization capability. These various, much-analyzed pathways (e.g., processes and stages listed in the Federation of American Scientists [FAS] Web site; FAS, 2011) revolve around the acquisition—either legitimately or illicitly, either through indigenous development or with help from outside actors—of certain capabilities on the road to developing an overall nuclear capability, which requires developing expertise in several different, but interrelated areas.

DOI: 10.1002/asi

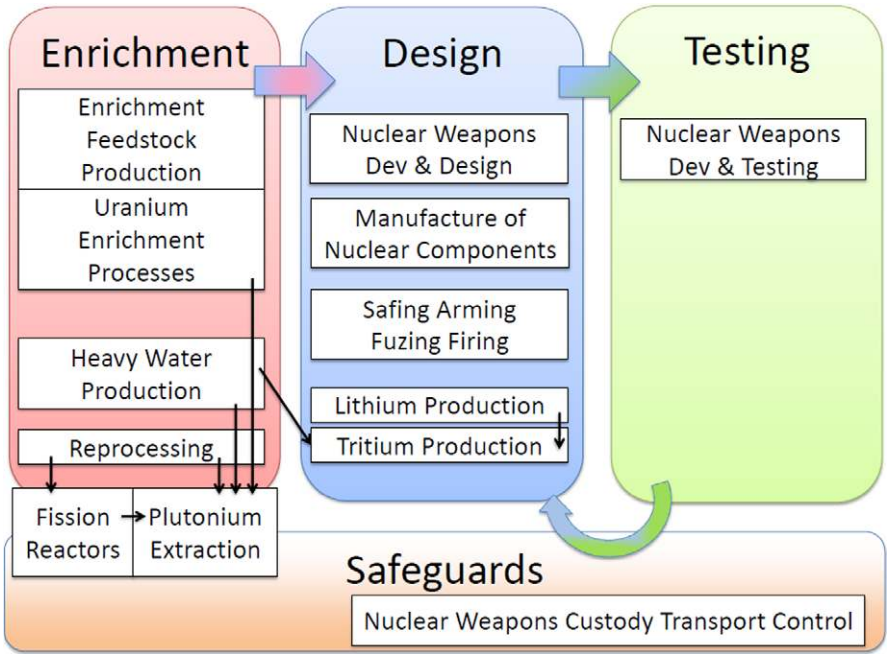

FIG. 1. Types of nuclear weaponization capabilities and associated processes outlined in the Department of Defense's Military Critical Technologies List, grouped by major stages: Enrichment, Design, Testing, and Safeguards. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The relationship that these areas of expertise have with each other can be best represented using a network due to the nonhierarchical nature of these relationships. For example, a capacity for uranium enrichment can be developed in parallel with nuclear design capabilities, which makes the existence of a strictly sequential/hierarchical workflow less likely. Furthermore, using networks for representing nuclear capabilities helps visually crystallize clusters of interrelated technical expertise types and how they relate to the overall nuclear capability development process. This can be observed in Figure 1, which groups the areas of nuclear expertise (i.e., processes) by the type of nuclear weaponization capability, and highlights the dependencies and associations between different areas of nuclear weaponization capability. Notably, the knowledge associated with nuclear fission reactors and plutonium extraction relates to nuclear weaponization capabilities associated with both enrichment and safeguards; they are drawn in Figure 1 on the boundaries of both clusters.

In this study, we loosely identify four types of nuclear weaponization capability based on the U.S. Department of Defense Militarily Critical Technologies List (MCTL) section on nuclear weapons technology (Department of Defense Security Institute, 2010). It is regarded as one of the clearest official documents detailing critical nuclear technologies that might be associated with emerging nuclear capability, and is an ideal source for identifying stages and potential technologies associated with them (Government Accountability Office, 2006).

The four types of nuclear weaponization capabilities we focus on based on the MCTL are enrichment, design, testing, and safeguards.

Enrichment capability should be interpreted as the capability to adequately enrich fuel to weapons-grade material. Enrichment capability refers to uranium and plutonium enrichment, as well as the capability to manufacture heavy water.

Design capability refers to the capability to design and manufacture nuclear weapons. This primarily refers to machining capabilities—the ability to mass-produce extremely precise tools and parts that are essential in a nuclear device. A nuclear program that is interested in developing a nuclear weapon will also be interested in controlling the weapon, which is why development of safing, arming, fuzing, and firing technology is a key process. Many weapons designs rely on tritium, which is why tritium production is a potential indicator of nuclear design capability.

Testing capability refers to the ability to adequately capture information about failed and successful nuclear weapon tests to incorporate feedback into the upcoming iterations of the design, manufacturing, and testing processes.

Finally, *safeguard capability* refers to an awareness of the proper way to handle and transport radioactive materials. These are key human worker safety issues that a nuclear program would be interested in developing, and this type of expertise is common throughout the commercial nuclear energy industry today. For instance, nuclear emergency planning, which can be rightfully considered within the scope of safeguard capabilities, is required in all phases of nuclear development, and it necessitates serious teamwork. Currently, most plants have at least 200 employees focusing on performing risk analysis, planning, and preparing for

DOI: 10.1002/asi

emergency cases that might require evacuation and sheltering of, in some cases, more than 10,000 people (Nuclear Energy Institute, 2005).

To be able to represent the nuclear-weaponization-related capabilities presented in Figure 1 in a network, we represent nuclear weaponization capabilities and the specific, weaponization-related expertise areas (i.e., processes) within each of these capabilities as two different entity classes (i.e., two different types of nodes). One entity class, *capabilities*, represents nuclear capabilities, and it has four nodes (enrichment, design, testing, and safeguards). The other entity class, *processes*, represents specific subprocesses and areas of expertise within each field such as plutonium extraction, tritium production, and enrichment feedstocks production. Each *process* is associated with a particular *capability*. The only two processes that are explicitly associated with two capabilities are the nuclear fission reactor and plutonium extraction processes. These are classified to be associated with both the enrichment and safeguard capabilities.

We extend this network representation of nuclear weaponization capabilities and related processes to research publications through the use of the named entity extraction technique borrowed from the field of text mining. This way we are able to extract the related terminology mentioned in the articles and classify them into the right area of expertise and nuclear weaponization capability. Further information on how text mining and mapping of nuclear terms to nuclear weaponization capabilities will be discussed in more detail in the Computational Framework section.

Dataset

In this article, we use a dataset we have compiled from experimental and theoretical nuclear articles that are publicly available on the Cornell University's Preprint Library, the arXiv Web site (Cornell University, 2011).

There are numerous reasons why nuclear research publications are arguably valuable for evaluating nuclear weapons capability. First, research requires funding regardless of field; however, for nuclear research, funding is even more critical, as projects for building reactors and performing nuclear experiments require multimillion dollar budgets (United States Nuclear Regulatory Commission, 2011). Channeling such significant amounts of money to research is beyond industry-only sponsorship; it usually correlates tightly with national agendas. Second, coauthorship networks are a tangible, reliable, and well-documented source through which authors disclose their collaborators. These basic science research activities may not signify progress towards weaponization, but they are concrete steps that signify national intent to utilize nuclear technology and to develop national technical know-how.

Our dataset consists of 21,080 articles' texts and the related metadata. The articles are in the field of theoretical and experimental nuclear physics, and they were added to the online library from 1992 to 2010. Each article has a unique identifier. The article IDs follow a certain convention, making it easy to create monthly, quarterly, and yearly snapshots. For instance, if the article was published online in June 1998, the first four digits of its ID are 9806. Similarly, if the article was published in December 2000, the article's ID begins with 0012.

In addition to the text, in the arXiv Web site, there is structured metadata available for bulk download in the form of XML files. Metadata XML files contain the following fields of information for each article:

Header: identifier, datestamp, setSpec

Metadata: authors, title, categories, comments, and abstract

The setSpec field shows the specific article set a metadata file is generated for. The arXiv Web site includes articles from various fields such as physics: nuclear-experimental, physics: nuclear-theory, physics: high-energy, physics: astrophysics, mathematics, computer science, and statistics. The value of the setSpec field can be any of these. However, the value of the categories field might contain multiple set names listed. For instance, an article can primarily be considered as a high-energy physics (HEP) article, and it can be cross-listed in nuclear physics, and/or nuclear theory as well. Our dataset covers the articles that are primarily listed in theoretical nuclear physics or experimental nuclear physics. Inside the authors' field, there are as many author fields as needed. Each author field has keyname (surname) and forenames as subfields. In some cases, an author field has affiliation as a subfield. However, because it is not mandatory, not all records include this information.

Using this dataset, we have extracted two interlinked networks:

Publication Network, which models information on which author wrote which article. It consists of 107,621 edges between 16,404 author nodes and 21,080 article nodes.

Coauthorship Network, which models information on who writes articles with whom. There is an edge between two authors if they have coauthored at least one article. It consists of 38,524 edges between 16,404 author nodes.

Extracting Interlinked Nuclear Physics Networks via Matrix Multiplication

To be able to construct multiple interlinked networks from our dataset, we rely heavily on matrix multiplication. For instance, to be able to build a coauthorship network, we multiply the publication network by its transpose. We also use matrix algebra to get country-to-country, city-to-city, and organization-to-organization coauthorship networks. In the computation of each of these matrices, we utilize affiliation information, as we have information on which city and country each organization is in.

Let matrix L denote the location matrix, which represents which author is affiliated with which country. The rows of matrix L are countries, and the columns are authors. To obtain matrix L, we need the information on the affiliation of

DOI: 10.1002/asi

the authors and information on the countries these organizations are based in. Matrix L' is the transpose of matrix L, whereas matrix A represents coauthorship relationships among the authors. Following this notation, the country-by-country coauthorship matrix, X, is obtained by $X = L \times A \times L'$.

Preparation and Cleaning of the Dataset

In addition to providing access to the research publications, the arXiv Web site facilitates download of metadata about the articles (e.g., information about the authors of each article and their affiliations). However, there are four major issues that need to be addressed before the data from arXiv Web site can be used for analysis.

First, when an author uploads his or her article to arXiv, entering the affiliation information is not enforced as a mandatory field; hence, not all records include this information. Therefore, we had to go through additional manual processing to capture the affiliation information of as many authors as possible.

Second, when affiliation information is entered, the same organization might be listed in different forms. For instance, Massachusetts Institute of Technology might be listed as M.I.T, MIT, or Massachusetts Institute of Technology in different articles. This requires converting all different representations of the same organization to a single representation to prevent appearance of redundant nodes in the network. We have processed the affiliation names using a thorough organization thesaurus that converts different representations of the same organization to a uniform organization name. We performed additional manual processing to fix the entries that are not captured by the thesaurus.

The third issue is the disambiguation of the author names that requires processing author entries in a way similar to processing organization names. An additional problem with author names is that there might be two different people with the same first and last name. When the author names are converted to their uniform representations, two different people might be represented by a single network node. To ensure that this does not happen, such people are further disambiguated according to their affiliations and coauthors.

The fourth and last issue is deduplication, the removal of duplicate entries. In some cases, duplication (i.e., repeated articles) might amplify the relative importance of certain terms, requiring deduplication. Because arXiv is an online library where authors upload their articles on a voluntary basis, observing duplicate entries for the same article is possible. However, we have noticed that many files have been removed by the Web site administrators upon detection of duplicate entries. Similarly, if an article is uploaded multiple times, creating different versions of the same article, we only process the latest version; earlier versions are discarded. The remaining duplicate entries, if any, are automatically removed from our database as a part of our text-mining procedure. Hence, duplicate entries are not a concern for us.

Additional Discussion: Pros and Cons of Using arXiv Data

Although there are major challenges in the preparation of the dataset that should be addressed before data from the arXiv Web site can be used for analysis, there are still reasons for preferring a database over the data that might be obtained from a selected list of journals. For instance, *Physical Review Letters*, the *Physical Review* series, *Reviews of Modern Physics*, the *Nuclear Physics* series, *Journal of Environmental Radioactivity*, *Journal of Applied Radiation and Isotopes*, *Environmental Science & Technology*, and *Nuclear Instruments and Methods in Physics Research* are among the most popular journals that publish on nuclear physics and they concentrate on producing and defusing novel knowledge and advancements in nuclear physics. Being relatively narrow in scope, it is uncommon for these venues to offer any metadata that facilitates bulk download and analysis studies that evaluate the evolution of the entire field of publications. In addition, because the goal here is to analyze nuclear physics research on the subject of where it stands in relation to developing nuclear capabilities, mining a database provides broader data than what could have been obtained if we were to mine a selected list of journals.

Another issue with the coverage of databases like arXiv is that there may be gaps for certain individuals and areas. However, analyzing only a selected list of journals can do nothing but aggravate this problem; after all, researchers publish in different journals, and it is unlikely that our data mining effort will capture every one of their publications. Therefore, in this work, we decide to use a publicly available, open-access database that has previously been used by many other researchers for various analyses (Leskovec, Kleinberg, & Faloutsos, 2005; Newman, 2004).

One major caveat to using arXiv as our primary database is that uploaded publications on arXiv are not necessarily going to be peer-reviewed articles. Therefore, the publications available through arXiv represent a stock of knowledge composed of (a) articles that have gone through review, (b) articles that are still in the early draft stage, and (c) articles that document a considerable amount of research that would not otherwise be published.

In our opinion, this is hardly a limitation; actually, the fact that arXiv is open for all kinds of research regardless of its status as peer-reviewed is a net positive. For instance, a research document that reports null results and therefore has a very slim chance of being published in any reputable journal may still be posted on arXiv, thus indicating some effort or interest that went into a particular area of investigation. This is what we are trying to capture. The main thrust of our research is remote capability detection; therefore, we seek sources that not only serve as a venue for distribution of scientific articles but also allow us to detect any ongoing nuclear activities or research efforts regardless of the quality of the research as judged by the academic world. We understand that it is difficult to incorporate the whole stock of nuclear knowledge into an analysis such as this; however,

DOI: 10.1002/asi

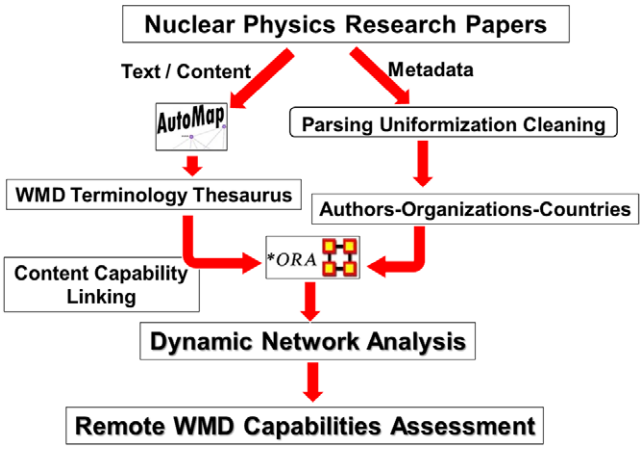

FIG. 2. A workflow of the methodology adopted in this investigation. Both the text (content) and the metadata of each article are analyzed, as the fork after "Nuclear Physics Research Papers" shows. After much cleaning, thesauri are obtained from the text, and lists of authors, organizations, cities, and countries are obtained from the metadata (sometimes supplemented from the text). ORA performs our network analysis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

we absolutely do wish to incorporate all sorts of articles hosted by arXiv, as they only add to the sum of our knowledge and help us gauge ongoing research activities at all levels and quality.

To summarize, considering that the main goal of this study is to capture research being conducted in the area of nuclear physics, we conclude that the arXiv database—although very much incomplete—does provide a valuable tool with which to commence this investigation; it captures a considerable portion of research in the area of nuclear physics, as far as open-source publications go. Finally, as far as the reliability of research submitted to the arXiv database goes, it is of a secondary concern to us because our main goal is to capture all research in the area of nuclear physics—regardless of quality—and to see whether there are some knowledge hubs that generate research that can be dangerous. The analytic inferences that we draw in the study may be cross-validated or improved by supplementing our data with the data gathered from peer-reviewed journals.

Computational Framework

In this section, we describe the computational framework we have employed in this article. Figure 2 depicts a high-level description of our overall workflow.

On one branch of our workflow (left, Figure 2), we process the contents (texts) of the articles to build a nuclear physics-specific thesaurus and link the articles' contents with the weaponization capabilities and processes. The other branch (right, Figure 2) consists of parsing, uniformization, and cleaning the dataset to extract clear author, organization, city, state, and country information associated with each article. Once both lines of work produce satisfactory, sufficiently complete data, we fuse data from both ends in order to generate a large dynamic meta-network that can be analyzed using ORA (Carnegie Mellon University, 2011a). Next, we provide detailed information on each of these subphases.

Text Mining

In this section, we discuss how we process the contents (texts) of the articles to build a nuclear physics-specific thesaurus to link each article with the weaponization capabilities and processes covered in it.

Data-to-model. We use Automap (Carnegie Mellon University, 2011b) as our main text-mining tool for constructing the terminology thesaurus for nuclear physics research and associating each article with its terminology content, and iterate over the steps of the data-to-model (D2M) process until we get our dataset in a form that is appropriate for performing network analysis. Data-to-model is a computerized data mining procedure for extracting social networks and building thesauri from text files. The details on the steps of data-to-model procedure are as follows:

Step 1: In this step, we perform detailed cleaning on the text files, such as removing extra space, blank lines, numbers, and individual letters. For nuclear physics articles, this is important because such articles use advanced mathematical formulations with many single-letter, super-/subscripted variables and numbers. This step is important for forming n-grams (i.e., a contiguous sequence of *n* words that should be interpreted altogether like an idiom) and identifying proper nouns. Those characters would otherwise appear as valid characters interfering with the named entity extraction. *Named entity extraction* typically refers to the extraction of named examples or instances of *Named Entities* (NE) that are referred to by a name. The entities might be from various classes like *agent*, *organization*, *location*, or *knowledge* (Bikel, Schwartz, & Weischedel, 1999; Diesner & Carley, 2008).

Step 2: This step involves text refinement. We create stemmed versions of the nouns and verbs (e.g., detensing/depluralization). Detensing refers to removing the tense affixes such as –ed and –ing from verbs; depluralization refers to converting a word that is in plural form to its singular form. Both techniques are used for reducing the redundancy in the set of words we analyze by converting different appearances of the same word to a uniform representation. In addition, to reduce the noise in the set of words we analyze, we also delete noise words such as prepositions, articles (e.g., a, an, the) and helping verbs (e.g., to). This step is completely automated.

Step 3: In this step, we extract entities and n-grams that are listed as named entities. The result of this extraction is a thesaurus of named entities. However, the initial thesaurus can contain invalid information that requires additional semiautomated cleanup.

Step 4: In this step, we form a thesaurus for ontological cross classification. The named entities that are identified in

DOI: 10.1002/asi

TABLE 1. Enrichment feedstock technology parameters (example for nuclear terminology-process mapping).

Technology	Sufficient technology level	Export control reference	Critical materials	Unique test, production, and inspection equipment	Unique software and parameters
Purification of yellowcake (wet process)	Knowledge of liquid-liquid extraction systems Experience in using HNO_3	NTL 8F; NRC J	Yellowcake nitric acid (HNO_3) tri-n-butyl phosphate (TBP) Refined kerosene	Filters; centrifuges; pulse columns; concentration/thermal denitration systems; tanks resistant to HNO_3	Distribution coefficients for many elements Aqueous solubility for many compounds

Step 3 are classified into the following ontological categories: Agent—Resource—Organization—Task—Location—Knowledge—Time—Belief—Event.

Ontological classification enables us to construct social networks such as (Agent × Agent), (Knowledge × Location), and (Time × Event), which allow us to focus on interaction between specific domains of information. Knowledge, resource, and task classes are the entity classes that hold nuclear-related content. They are also the classes we classify further with respect to the capabilities and processes we introduced in an earlier section. Nuclear-weaponization-related classification will be discussed further in the following sections.

Step 5: As initial input to the first iteration of the Data-To-Model procedure, we have prepared an initial nuclear terminology thesaurus file based on the information available in the International Nuclear Information System (INIS, 2011). As we mine the documents in our dataset, additional valid phrases and terms which have not been covered by the initial thesaurus file show up. In Step 5 of every iteration, we incrementally merge valid terms that are identified by the data-to-model procedure with our nuclear thesaurus from the previous iteration.

Depending on the size and coverage of the dataset, several (e.g., 3–10) iterations of these above-mentioned steps may be required. At each iteration, the text is first automatically scanned for named entities. Then, manual inspection of the list of named entities is required to remove spurious entries and tag legitimate entities by their ontological category (e.g., "Harvard" is an *organization*). The entries that are identified to have legitimate ontological tags are merged into the master thesaurus file, and therefore do not appear in the list of entities to be cleaned or tagged in the following iterations. This way, the number of entries requiring manual attention decreases iteratively. This gives a feel of the completeness of the dataset, signaling when we can finalize the text processing and cleaning.

Mapping terminology to capabilities. One output we obtain from the data-to-model procedure is a domain-specific thesaurus file that covers (a) named entities belonging to different ontologies extracted from the articles in our dataset, (b) the terms from the rather generic nuclear thesaurus built based on the information in INIS (2011), and (c) the nuclear terms from the MCTL section on nuclear weapons technology MCTL (Department of Defense Security Institute, 2010). The primary goal of using MCTL information is to link nuclear terms that have knowledge, resource, and task meta-ontology tags to specific nuclear weaponization capabilities and processes as discussed in the section, Representing Nuclear Weaponization Capabilities in Dynamic Networks.

As mentioned earlier, we treat knowledge, resource, and task as the ontology classes that might contain terms that are specifically related to nuclear weaponization research. To give examples, we ontologically classify "quantum theory" as knowledge, "uranium" as resource, and "enrichment" as task. All the knowledge, resource, and task terms in our thesaurus are mapped to individual nuclear weaponization capabilities and processes discussed in the section, Representing Nuclear Weaponization Capabilities in Dynamic Networks, if they are related to any. The nuclear terms that are rather generic, and are not specifically related to a weaponization capability/process are not tagged as weaponization related content. To illustrate this process, Table 1 lists enrichment feedstock technology parameters. The first line describes the wet process of producing yellowcake, a mixture of uranium oxide.

From this table, the following terms were extracted and mapped to the process of enrichment feedstock production: yellowcake, yellowcake purification, tri-en-butyl phosphate, centrifuge, centrifuges, denitration system, denitration systems. Many of the additional terms used in this table (wet, HNO_3, kerosene) are so common that they were not included as keywords that should be associated with enrichment production process.

Content-capability linking. The next and final step of the text-mining process is to form the networks from the text files and the stages associated with each and perform network analysis. Each text file contains numerous nuclear terms. Some of these terms are tightly coupled with one another and have a relatively frequent co-occurrence which in turn represents a link between these two terms in the semantic network. As described in the previous subsection, we tag all nuclear terms with the nuclear weaponization process and capability that they are correlated with, if they are correlated with any. For instance, an article that contains multiple testing stage related words is tagged to be a testing article due to its content. Once we have all articles linked to the weaponization stages their contents are correlated with,

DOI: 10.1002/asi

we can start drawing links between articles and weaponization capabilities, representing our findings in a network format.

Dynamic Network Analysis

As we mentioned earlier, our primary goal in this article is to perform dynamic network analysis on nuclear research networks extracted from publicly available, online publications to have a better understanding of what nuclear capabilities can be associated with major institutions in national innovation systems considering what kind of nuclear knowledge each organization/city/state/country has been producing.

Our main dynamic network analysis tool is ORA, which is built and maintained by Carnegie Mellon University's CASOS research center (Carnegie Mellon University, 2011a). ORA is an interactive dynamic, meta-network analysis tool that maintains the internal structure of an organization/social network as a set of agents, tasks, and resources. ORA provides a rich set of statistical analysis tools for comparing and contrasting networks along with graph-theoretical social centrality measures specifically designed for social networks.

TABLE 2. Key locations ranked high in terms of degree centrality.

Degree centrality		
Country	State	City
USA (1.00)	NY (1.00)	Moscow (0.48)
Germany (0.64)	CA (0.79)	Darmstadt (0.43)
France (0.46)	OH (0.44)	Tokyo (0.42)
Russia (0.35)	TX (0.37)	Orsay (0.41)
Italy (0.30)	MI (0.36)	Cedex (0.38)
Japan (0.29)	BC (0.35)	Strasbourg (0.29)

TABLE 3. Key locations ranked high in terms of closeness centrality.

Closeness centrality		
Country	State	City
Russia (0.009)	WA (0.021)	Darmstadt (0.002)
USA (0.009)	NY (0.021)	Vancouver (0.002)
Italy (0.009)	OH (0.021)	Moscow (0.002)
Germany (0.009)	QC (0.021)	Frankfurt (0.002)
UK (0.009)	MD (0.021)	Bonn (0.002)
Mexico (0.009)	MI (0.021)	Seattle (0.002)

Results

In this section, we report our findings, grouping them under separate categories. We first provide results on key entities at both state and organization levels of analysis. Then, we present another set of results that comment on the differences between established and emergent players of nuclear research and show how these differences are reflected in the networks we analyze. The last section covers text-mining techniques, showing how the content of the publications can be linked with nuclear capabilities of each country.

Social Network Analysis: Centrality Measures

In this section, we present results on key entities in the network. Decades of research on social network analysis (SNA) has led to a wealth of SNA centrality measures that focus on finding the key actors in a social network. There are four major network centrality measures that are most commonly used: degree centrality, closeness centrality (Sabidussi, 1966), betweenness centrality (Freeman, 1977; Freeman, 1979), and eigenvector centrality (Bonacich, 1987).

To briefly recap, degree centrality of a node is defined as the number of its immediate neighbors. The nodes that are ranked high on this metric have more connections to other nodes in the network. Closeness centrality of a node is calculated as the inverse of the average distance of a node to all other nodes in the network. Closeness centrality describes how close a node is to others in the network; it is interpreted as the efficiency of the node's ability to propagate information to the rest of the network. Similar to closeness centrality, betweenness centrality also depends on the shortest paths in a network. Betweenness centrality is defined as the fraction of shortest paths a node is on across the entire network, describing how critical it is for connecting different components of the network or for partitioning the network into different components when removed. Lastly, eigenvector centrality is used to characterize the influence of a node on the network; connections to well-connected nodes are weighted higher than connections to nonconnected nodes. Eigenvector centrality is especially useful for identifying the nodes that have the power to mobilize other nodes. We refer the reader to Wasserman and Faust (1994) for a more detailed introduction on the use and definition of social centrality metrics.

Key locations. In this section, we present a list of key locations in terms of cities, states, and countries ranked according to different social centrality metrics. States are only valid for the United States and Canada. In Tables 2–5, the nodes that are high in terms of a certain centrality metric are listed along with their centrality values. Tables 6 and 7 evaluate the robustness of the results, and finally Table 8 summarizes entities that are ranked highly by 150 metrics and reflect overall rankings across multiple metrics.

The results presented in Tables 2–5 reveal interesting features of the network. Because the computation of closeness values is affected by the nodes that are not within the same component as the node of interest, the closeness values are relatively at a smaller scale than those of other metrics and cannot distinguish between nodes that are highly ranked. Although betweenness is also based on the

DOI: 10.1002/asi

TABLE 4. Key locations ranked high in terms of betweenness centrality.

Betweenness centrality		
Country	State	City
Germany (0.16)	NY (0.170)	Darmstadt (0.106)
Russia (0.14)	OH (0.144)	Vancouver (0.034)
USA (0.13)	WA (0.142)	Tokyo (0.031)
Italy (0.09)	BC (0.112)	Frankfurt (0.030)
Mexico (0.79)	CA (0.101)	Seattle (0.026)
Canada (0.75)	QC (0.069)	Moscow (0.026)

TABLE 5. Key locations ranked high in terms of eigenvector centrality.

Eigenvector centrality		
Country	State	City
USA (1.000)	NY (1.000)	Moscow (1.00)
Germany (0.118)	CA (0.087)	New York (1.00)
France (0.030)	OH (0.036)	Frascati (1.00)
Canada (0.026)	MI (0.022)	Wuhan (1.00)
Russia (0.021)	TX (0.021)	Coimbra (1.00)
Japan (0.013)	WA (0.015)	Calabria (1.00)

TABLE 6. Robustness context of centrality measures.

Robustness context		
Degree centrality	Betweenness	Eigenvector
USA (–3.05)	USA (86.2)	USA (–0.491)
Germany (–3.50)	Germany (21.3)	Germany (–0.983)
France (–3.63)	Russia (12.8)	Russia (–1.20)
Russia (–3.65)	Italy (9.89)	France (–1.25)

Note. The numbers in parentheses are the number of standard deviations from the mean centrality measure for a random network of identical size and density.

TABLE 7. Comparing centrality measures of co-authorship network at the country level with a random network with identical density and size.

	Robustness context		
	Degree Centrality	Betweenness	Eigenvector
Mean	0.003	0.003	0.036
$Mean_{random}$	0.128	0.006	0.655
SD	0.008	0.024	0.095
SD_{random}	0.024	0.004	0.160

shortest paths between nodes, it is able to distinguish between highly ranked nodes and primarily provides a slightly different ordering of the nodes listed by other metrics. In general, nodes that are high on these metrics in this data are distinct from what would be seen in a random network (see Table 6).

TABLE 8. Overall rankings for key locations (averaged across 150 different social centrality metrics).

Across 150 centrality metrics		
Country	State	City
USA	NY	Moscow
Germany	CA	Darmstadt
France	OH	Tokyo
Russia	MI	Orsay
Italy	TX	Frankfurt
Canada	WA	Berkeley

The other major result is that, at the country level, there is significant overlap in the top ranked actors based on eigenvector and degree centralities. As we further discuss in the section Major Players Versus Emerging Countries, the country-level coauthorship network has a core-periphery network structure, and this result is to be expected, as eigenvector centrality is based on the relative connectiveness of node neighbors; in a core-periphery network, only core countries are well connected and have a high degree of centrality. At the city level, the locations listed by different metrics do not reflect the country-level network. There are more nodes in the city-level network, and the network is significantly sparser than the network at city or state level. Hence, the nodes that stand out are usually parts of smaller disconnected components. This is also reflected in the low closeness values in the network (e.g., too many nodes that are not reachable, and hence have a distance of infinity) as well as the eigenvector values where the nodes appear to be central to the small components they belong to.

To verify the robustness of these centrality measures, Table 6 compares highly ranked centrality actors to random networks of identical size and density, and Table 7 provides a network-level comparison of the distribution of centrality measures to a random network. More specifically, Table 7 compares the mean and standard deviation of centrality measures to random networks. The statistical context in Table 6 represents the number of standard deviations from the mean of the centrality measure of the random network to the actor's centrality measure. Both degree and eigenvector centralities are lower than would be expected in a random network; both the means and the variance are lower than a random network. This suggests that countries are collaborating in very specific clusters, and that this behavior is unique.

Finally, in Table 8 we present centrality results in the aggregate form, averaged across 150 different metrics. According to these results, the United States appears as the top producer of nuclear research. However, U.S. cities do not dominate the list of top cities. This is because, in the United States, research is distributed across the country as opposed to other countries that have significant centralization, either at their capital (e.g., Moscow-Russia) or at a city with major universities or national labs (e.g., Orsay-France, Darmstadt-Germany). In this regard, Italy is similar to the United States

DOI: 10.1002/asi

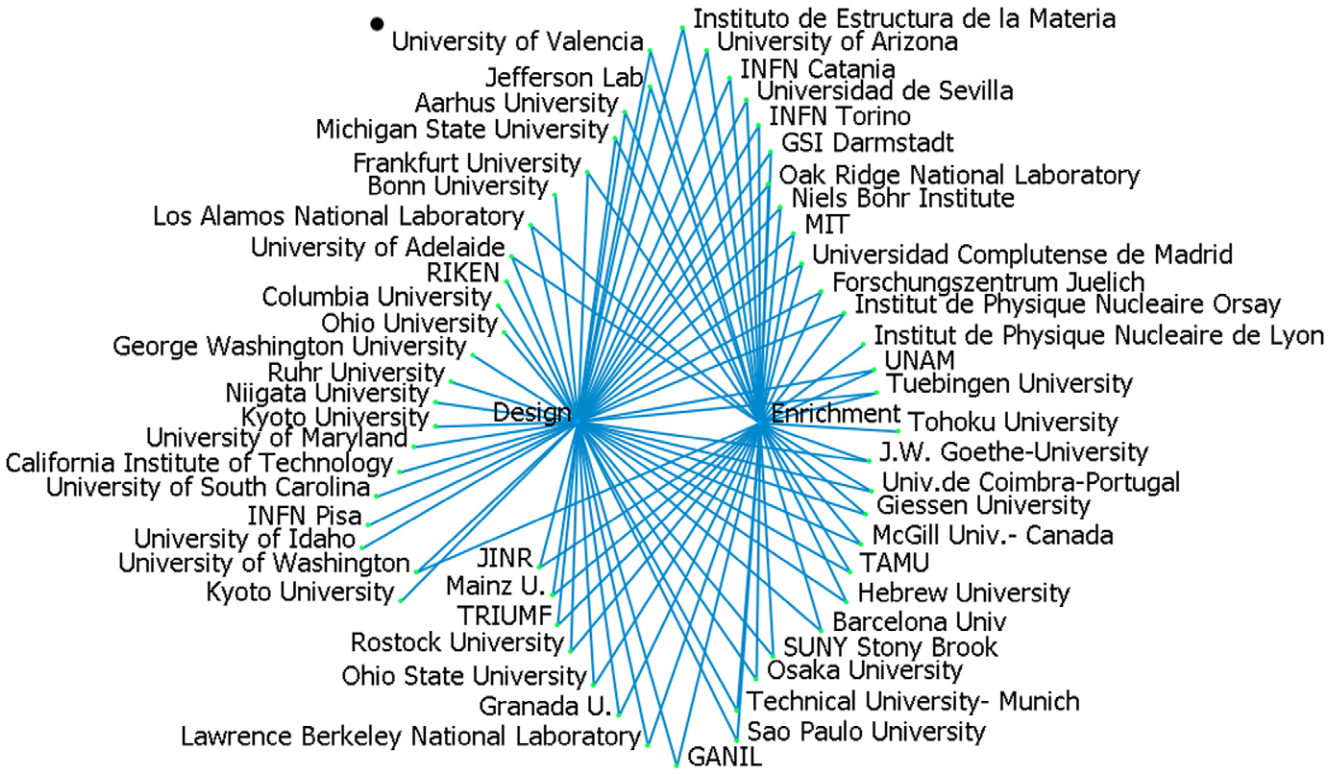

FIG. 3. Key organizations across the world associated with "Design" and "Enrichment" capabilities that have published at least 50 articles covering a certain nuclear capability. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

in terms of having less centralized research. Italy's National Institute of Nuclear Physics (INFN) is distributed across the country, with facilities in 19 university physics departments and four laboratory sites. Therefore, we again do not observe any single city from Italy ranked in the top five.

In terms of states, New York (NY) and California (CA) are the two major states both in terms of the number of publications they produce and in terms of the researchers and connections they have. New York is the highest ranked state primarily because of Brookhaven National Laboratory at Upton, New York, and State University of New York (SUNY) at Stony Brook, New York. Brookhaven National Labs is home to various research projects and the leader of a significant number of multinational, multiorganizational experiments.

For instance, the STAR collaboration appears as a very active organization in our high-energy/nuclear physics publication databases and it is composed of 55 institutions from 12 countries, with a total of 547 collaborators (Brookhaven National Laboratories, 2011). The STAR collaboration focuses on high-energy nuclear collisions that create an energy density similar to that of the Big Bang. Their first publications appeared in 2001, and the project is continuing as of 2011 under the guidance of Brookhaven National Laboratory, USA. SUNY Stony Brook benefits from its geographical proximity to Brookhaven National Laboratory, with many joint programs in place. A similar case holds for California. California is home to Lawrence Berkeley National Laboratory and UC Berkeley, which are geographically proximate. Much like New York, California hosts numerous additional universities and institutions, such as Lawrence Livermore National Laboratory (LLNL) and one of Sandia National Laboratory's two facilities, also in Livermore, which contribute significantly to its overall ranking.

Key organizations. In this part, we present the key organizations that are actively publishing articles, i.e., organizations that have written more than 50 articles in at least one of the three stages.

The number of testing articles is relatively small compared to design and enrichment articles. And, when a high threshold (threshold in terms of the number of articles published) is applied, testing disappears; we are left with organizations focused on design and enrichment. Figure 3 depicts key organizations involved in design and enrichment research. The second point is that the set of key organizations does not consist of universities or national laboratories only; it is a combination of both. However, national laboratories usually find their places as key organizations. Examples include the Lawrence Berkeley National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, and various branches of INFN in Italy.

In addition to the organizations from the United States (which constitute the largest fraction of key organizations),

DOI: 10.1002/asi

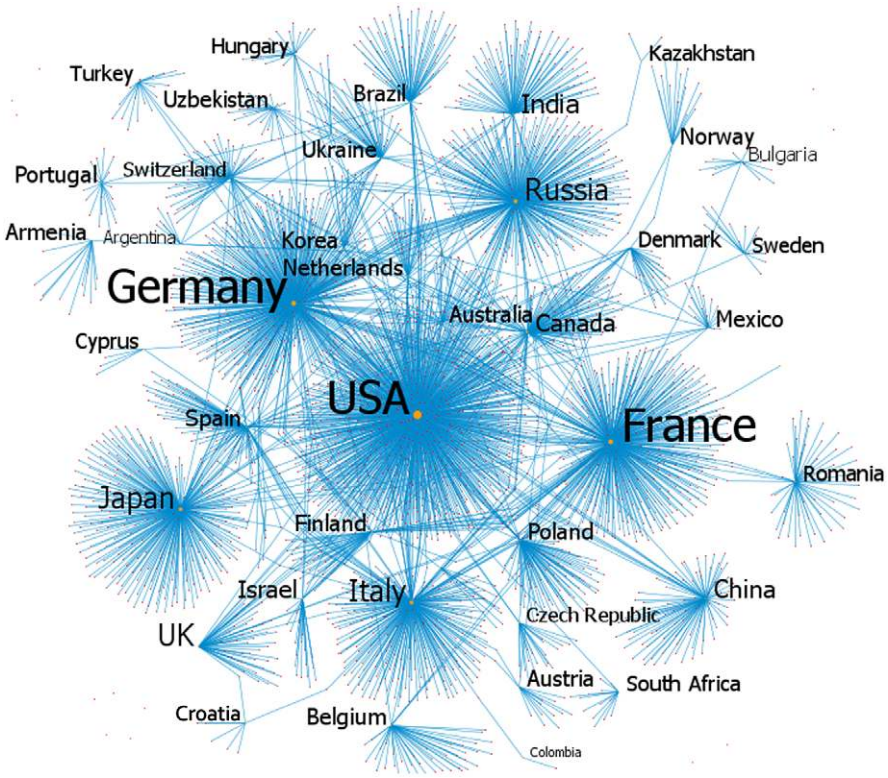

FIG. 4. Country nuclear researcher clouds. This figure is obtained on the author × country network where an author is connected to more than one country if he or she has affiliations with institutions in both. The size of the immediate sphere surrounding a country indicates how crowded its nuclear research group is. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

there is a significant number of organizations from Germany and Japan, which are the other two major countries in this line of research, as per our analysis on social centrality metrics on location entities.

Major Players Versus Emerging Countries

Emergence of interest from countries that were not previously active in nuclear research requires further attention if we are to detect threatening activities because the countries that are active in this field of research are countries that either have nuclear power or are interested in developing nuclear power. To discuss how we elicit emerging countries, we present visualizations of two different aspects of the meta-network. First, we discuss the nuclear researcher task force size of each country and the strength of collaboration bonds among various countries. Second, we discuss clustering coefficient results and explain how this metric might be useful for identifying emerging countries.

Nuclear research clouds. In Figure 4, we provide results on countries' respective researcher clouds. This graph is generated from the aggregated form of the country-by-actor network, with author nodes' name labels turned off. The labels of countries are scaled with respect to the number of nuclear researchers affiliated with the organizations in that country. Several observations are in order. In line with results presented in previous subsections, the major players are the United States, Germany, and France; these are followed by Russia, Japan, and Italy.

When Figure 4 is examined closely, there is a strong bond between Germany and France, which appears as a blurry thick line going across the United States. This is because they have many shared authors. Similar cases are observed for other pairs of countries such as Germany–Russia and USA–China.

In this network, we observe a core-periphery network kind of structure. An interesting point here is the sizes of India and China. They are relatively small, especially when their populations and gross domestic products (GDPs) are taken into account. The other point about the peripheries of the network presented in Figure 4 is that most of the peripheral countries (except China and India) are emerging ones or economically smaller countries that are not among the major players yet. We discuss emerging countries more in the next subsection, taking clustering coefficients into account.

Emerging Countries: Clustering Coefficients

In this section, we examine countries' clustering coefficients in terms of the number of shared authors. We first

DOI: 10.1002/asi

characterize the network as a core-periphery network to justify the use of clustering coefficients, and then discuss results from using the clustering coefficient measure.

Network characterization and topology. Although the clustering coefficient provides some insight into the extent of localization, it may provide a misleading amount of overconfidence due to the existence of complete subgraphs among colocated, coauthored articles. However, if we identify the underlying country coauthorship network as core-periphery, and verify that countries with high clustering coefficients are also peripheral actors, we can identify potentially suspicious behavior at the country level. We analyze the clustering coefficient at the country level as complete graphs of coauthors from the same country are collapsed into a single country node, providing more accurate results in terms of local density.

We compare some generalized network measures of centrality to random networks of identical size and density to try and characterize the country-level coauthorship network (Frantz & Carley, 2005). In this analysis, the inverse of the number of coauthored articles is used to represent the distance between two collaborating countries. To provide statistical context, we compare the means betweenness centrality in our country coauthorship network to randomly generated stylized networks using a *t* test.

A *t* test is a technique that is commonly used in social sciences to measure whether the means of two groups are statistically different from one another (Boneau, 1960). We consider a two-tailed *t* test to test the null hypothesis that the mean country betweenness centrality of the randomly generated stylized networks is identical to the mean of our observed country coauthorship network. As we do not know which mean will be higher or lower, we consider a two-tailed *t* test. Due to the low overall network density, we only use average betweenness centrality to compare networks; both closeness and inverse closeness centrality measures of randomly generated networks had distributions that failed to match our network.

We compare the country coauthorship network to three major types of networks: an Erdös-Rényi network (Rényi & Erdös, 1959), a core-periphery network (Borgatti & Everett, 2000), and a cellular network (Airoldi & Carley, 2005). Erdös-Rényi networks are random networks where the probability of an edge between two points is equal to the network density; if the country coauthorship network were of this type, it would imply that countries collaborate randomly. Core-periphery networks have a small "core" set of actors that are very well connected, but also have peripheral actors that are only connected to actors in the core. Cellular networks consist of modules or cells where actors are well connected inside the cell, but weakly connected outside the cell. A coauthorship network similar to a cellular network would imply explicit coordination among groups of countries.

We find that the (country × country) network appears most similar to a core-periphery network. The Erdös-Rényi network has average country betweenness centrality that is significantly lower than our observed network, and we reject the null hypothesis with the *t* test. Although we fail to reject the null hypothesis that the mean network betweenness centrality is identical to a randomly generated cellular network with a mean cell size of 4, we also fail to reject the null hypothesis that the mean network betweenness centrality is identical for a cellular network with a mean cell size of 10. Due to the existence of some large international collaborations, one might think that there may exist cell-like clusters of collaborating countries. However, an empirical evaluation of the network shows that this cannot be the case due to the small number of countries actually contributing to the development of nuclear knowledge. Results are summarized in Table 9.

TABLE 9. Table of *t*-test results ($\alpha = 0.1$, $df = 5$) of average betweenness for random networks of identical density and size, varied on network topology. This test provides statistical context for our characterization of the network as having a core-periphery structure.

Network topology	Average betweenness centrality	*t* Value	*p* Value (Two-tailed)
Country × country (data)	0.003		
Erdös-Rényi	–0.026	16.69	0
Core-Periphery	0.005	–0.771	0.48
Cellular (mean cell size = 4)	0.005	–0.800	0.46
Cellular (mean cell size = 10)	0.005	–0.734	0.50

After identifying that the overall country-by-country network has a core-periphery structure, we can use the clustering coefficients to identify peripheral countries that have still contributed to the general literature.

Clustering coefficient. The local clustering coefficient, initially proposed by Watts and Strogatz (1998), is a metric that clearly distinguishes social networks from random networks. In social networks, the formation of triangles is tightly correlated with the notion of transitivity. Transitivity refers to the probability of agents *i* and *k* becoming directly connected given that there exist direct links between (*i*, *j*) and (*j*, *k*). The local clustering coefficient of node *i* is defined as follows:

$$Clustering\_Coef(i) = 2\frac{|e_{j,k}|}{d_i(d_i - 1)}$$

In this formulation, nodes *j* and *k* are neighbors of node *i*. $|e_{j,k}|$ defines the number of links (edges) that exist between the neighbor of node *i*. d_i denotes the degree of node *i*, which is the number of immediate connections node *i* has. The number of possible connections that can exist among the neighbors of node *i* is $\frac{d_i(d_i-1)}{2}$ in an undirected graph, for which this formulation is provided. Therefore, the clustering coefficient of node *i* quantifies how close its neighborhood is

DOI: 10.1002/asi

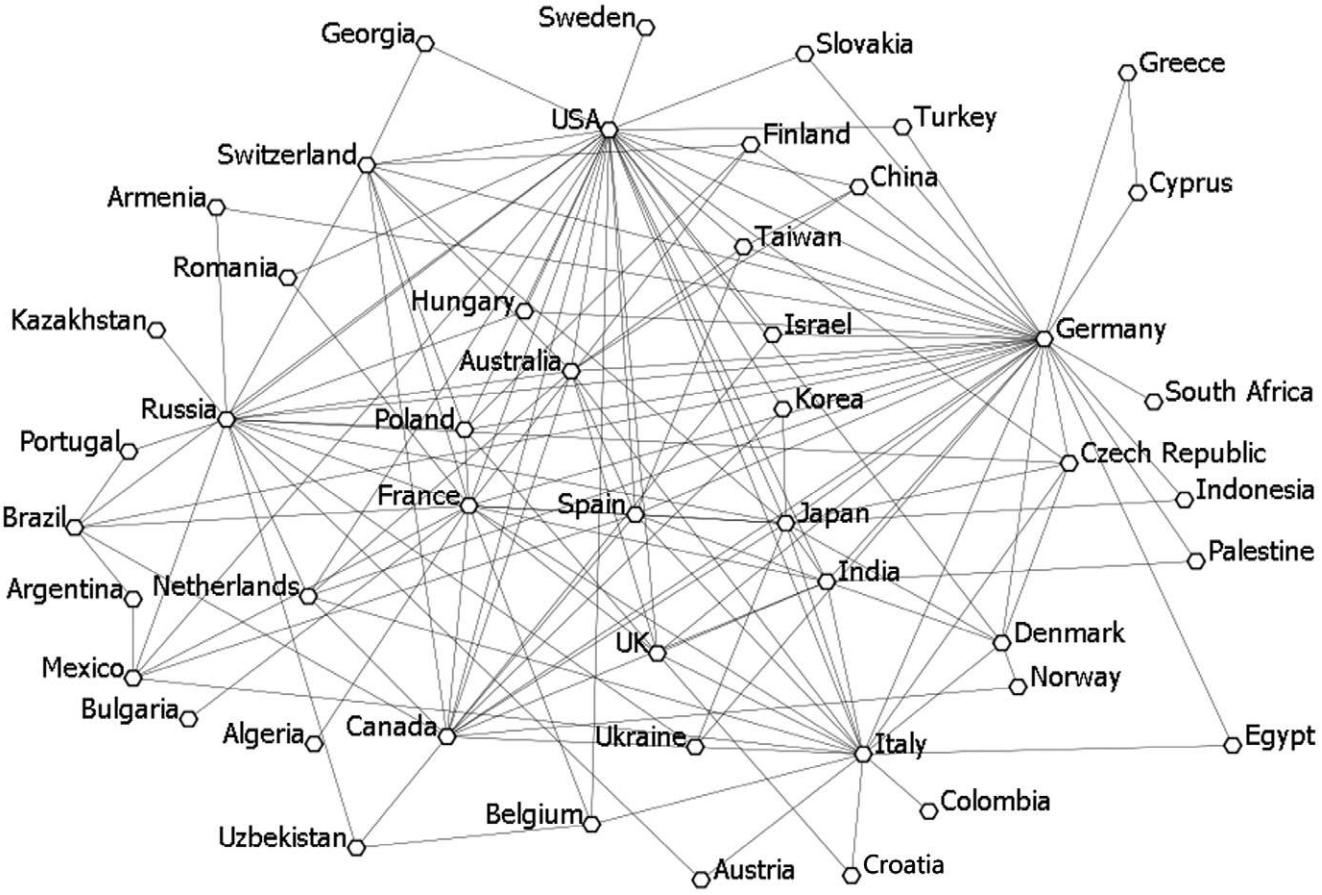

FIG. 5. Country × country network (based on shared authors). There is a line between two countries if there is at least one researcher that has/had affiliations with institutions in both countries.

to forming a clique. In other words, the clustering coefficient of a node provides insight into triangulation density and how localized its neighborhood is.

We next discuss the list of countries that have high clustering coefficient values in terms of the authors shared with other countries. The clustering coefficient results are obtained on a homogeneous (country × country) network that is generated based on affiliation information. Assume that $author_a$ had affiliations with two countries, $country_x$ and $country_y$. In this network, a link is drawn between any two countries if an author has had affiliations with organizations in those two countries. Hence, in this case, a link is inserted between $country_x$ and $country_y$ due to $author_a$ in a shared-authors network.

In a shared-authors network, major countries have many connections to other countries, which can still be considered sparse when all possible triangles joining nodes are considered. For instance, the United States, whose researchers are well distributed throughout the country and have many connections with researchers from abroad, turns out to be very low on this measure.

Emergent countries we discuss in this section are countries that are not as influential in the international arena as major players. However, they produce a substantial amount of publications on nuclear physics to be meaningful. Network analysis performed on those countries indicates that their publications are considerably localized, in the sense that authors from those states, on average, are not connected to too many authors from other countries. It is also observed that most of them tend to write articles with authors from the same institutions or cities.

Hence, countries that have high clustering coefficient values can be considered emergent countries in the country × country network presented in Figure 5. These are Kazakhstan, Turkey, Egypt, Greece, Bulgaria, Palestine, Colombia, Slovakia, Romania, and China.

Due to space constraints, we will consider the cases of four countries in detail. These four represent only a subset of the countries in our dataset; however, as case studies, they represent major actors from the emergent country list: Kazakhstan, Palestine, Turkey, and China.

Kazakhstan. In Kazakhstan, researchers primarily publish locally. There are two principal scholarly institutions engaged in nuclear research in Kazakhstan: the Institute for Nuclear Physics (National Nuclear Center of Republic of

DOI: 10.1002/asi

Kazakhstan) and Al-Farabi Kazakh State University, both located in Almaty. Considering Kazakhstan's historic relationship with Russia, there is an expected link between Kazakh and Russian researchers. The fact that there are nuclear research centers in Kazakhstan is hardly surprising. After all, following the dissolution of the Soviet Union, Kazakhstan had the potential to become one of the world's largest nuclear states. Although the Kazakh government chose to denuclearize, it is still interested in developing domestic nuclear energy for civilian use. In 2009, the Kazakh government announced its plan to build a nuclear power plant consisting of two 300 megawatt reactor units in the city of Aktau (NTI, 2010). Additionally, Kazakhstan possesses the world's second largest uranium reserves; over the past few years its uranium industry is moving towards expanding Kazakhstan's role in the international nuclear energy market (NTI, 2010).

In light of these facts, Kazakhstan's inclination towards conducting nuclear research is expected. Interestingly, Kazakhstan's industry appears to be more internationally positioned than its nuclear research centers. For example, Kazakhstan's Kazatomprom—a major Kazakh company engaged in extraction of and production of uranium—has a number of contracts with companies from countries including Russia, Canada, France, and China (Embassy of the Republic of Kazakhstan, 2011).

In our results, the two key institutions in Kazakhstan are the Institute for Nuclear Physics (INP) and Al-Farabi Kazakh State University. The Institute for Nuclear Physics is supported by the national government and does research that closely corresponds to national agendas. The aim of this institute is to conduct applied research in the field of nuclear physics and advise the government on possible uses of atomic energy in different spheres of the national economy (INP-Kazakhstan, 2011). Research performed at Al-Farabi Kazakh State University encompasses various aspects of experimental and theoretical physics and is larger in scope than what is being produced at INP. In addition, it is Kazakhstan's largest training research center preparing the next generation of Kazakh physicists (Kazakh State University, 2011).

Palestine. In contrast to Kazakhstan, Palestine has never had an association with nuclear weapons. Yet, there has been some interest on the part of the Palestinian National Authority in developing nuclear energy for peaceful purposes. Thus, in 1998, the Palestinian Energy Authority established a new unit called the Nuclear Energy and Radiation Protection Department (NERP; Palestanian Energy Department, 1998). NERP is tasked with designing policy aimed at supervising and promoting the peaceful application of ionizing radiation. It is also charged with establishing a Palestinian radiation protection and safety program (Palestanian Energy Department, 1998). One of the items on the NERP agenda is to promote and oversee the exchange of experts and information in the area of nuclear energy application and radiation protection. However, information on NERP's Web site suggests that not much has been done on this front since the agency was set up. The question, then, is why Palestine emerges as one of the entities producing nuclear knowledge.

A closer look at the data indicates that all the publications generated in Palestine are coming from a single university, Bethlehem University. Bethlehem University is a Catholic institution located at the West Bank. Due to its religious affiliation, it holds memberships in the International Federation of Catholic Universities, the Lasallian Association of Colleges and Universities, and the International Association of Universities (Bethlehem University, 2011). Of course, due to its religious affiliation, the researchers at this university do not accurately represent the demographics of the Palestinian population, i.e., most researchers are not Muslim Palestinians. In short, considering its international affiliations, there is a tendency on the part of faculty and students to reach out to Western publication venues, such as arXiv (Cornell University, 2011).

Turkey. Turkey is another emergent country in the field of nuclear research. In our dataset, a number of universities from Turkey appear: Middle East Technical University (METU), Gazi University, Izmir Institute of High Technology, Konya Selcuk University, and a few others, with METU being the most active of all. The organizations are mostly state universities, located in various cities. METU, which is the leading organization in this line of research, is a major technical university with active research programs in many different scientific and technical disciplines. METU is located in Ankara, the capital of Turkey, and due to its domestic and international prestige, is the most likely to be involved in cutting-edge technical research. Therefore, it is not surprising that the majority of Turkish publications in the field of nuclear physics come out of this university.

Another point to note is that the Turkish articles are predominantly theoretical nuclear physics articles and not experimental articles. This may reflect publishing trends at the one Turkish institution with access to a functional nuclear research reactor: Istanbul Technical University. At the national level, Turkey is currently developing its nuclear knowledge by having contracts signed for four 1,200 MWe Russian nuclear reactors at one site and is negotiating for a similar capacity at another (World Nuclear Association, 2011). The country does not currently have an operational commercial nuclear power plant.

China. Compared to its population and the number of its researchers in other fields, the number of Chinese nuclear researchers is relatively small, especially when compared to the United States and France. They are not one of the top-three influencers in the nuclear field yet; however, they are the fastest growing nuclear power user in the world. This growth is observed both in terms of the amount of research produced by Chinese researchers as well as onsite deployment and reactor construction. China is also one of the more active developers of niche reactor technologies; for instance, it is active in small modular reactor (SMR) research.

DOI: 10.1002/asi

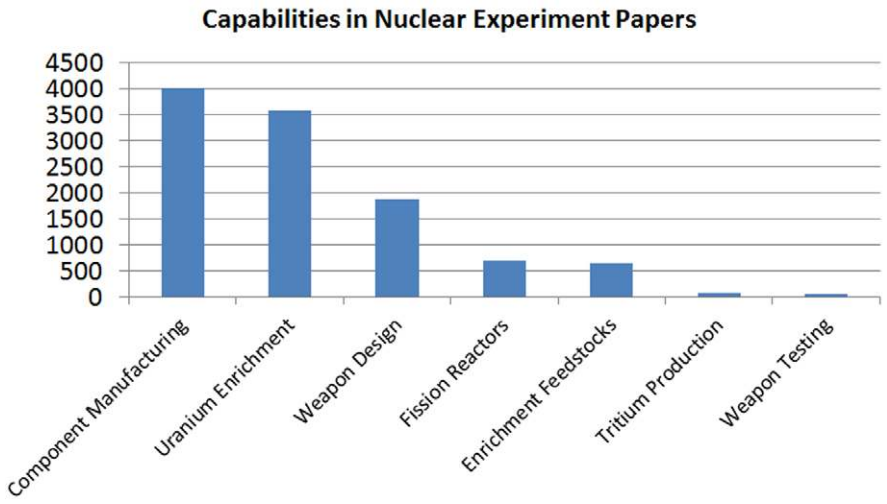

FIG. 6. Hot topics identified by text-mining techniques in nuclear physics publications (experimental articles). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

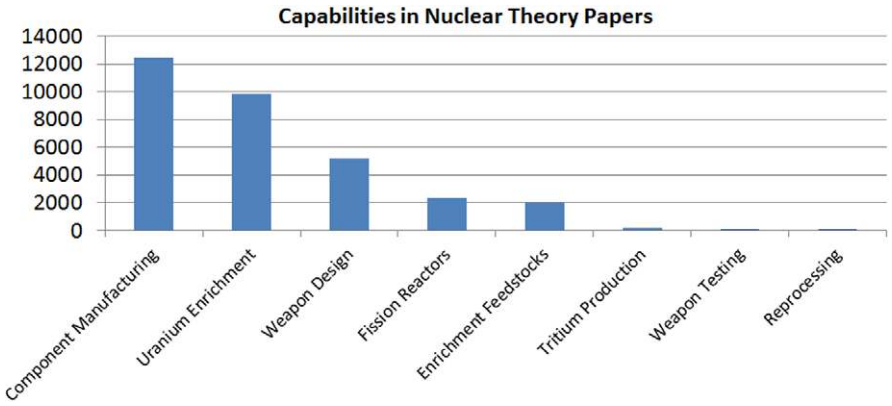

FIG. 7. Hot topics identified by text-mining techniques in nuclear physics publications (theoretical articles). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

According to public information the World Nuclear Association lists on their Web site (World Nuclear Association, 2011), with 13 operating reactors on the mainland, China is now speeding into the next phase of its nuclear power program. Currently, it has 27 reactors under construction, with an additional 28 of them being actively planned: some of these are state-of-the-art (generation III+) first-of-a-kind reactors, such as Westinghouse's AP1000. China is expected to quadruple its current nuclear capacity by the end of 2020, which is also reflected in the publications front: It is the fastest growing country in the entire world in this respect. Therefore, its status as a rapidly emerging country in this field is indicative of nuclear power's newfound prominence on the Chinese national agenda.

Nuclear Capabilities

In this section, we present our text-mining results based on the extraction of nuclear terms as named entities. We classify articles in our dataset according to nuclear weaponization processes and capabilities discussed in the section, Representing Nuclear Weaponization Capabilities in Dynamic Networks.

Hot topics in nuclear research. Out of 21,080 nuclear research articles, 4,627 articles are tagged as experimental. The remaining 16,543 are tagged as theoretical articles by their authors. Figures 6 and 7 show the coverage of nuclear weaponization processes that are explained in an earlier section in experimental and theoretical articles, respectively. In these results, an article might have content that is related to multiple nuclear weaponization capabilities or processes. Our results show that roughly 12,000 nuclear theory articles talk about component manufacturing, although they might also talk about further processes and/or capabilities.

In both categories of articles (experimental and theoretical), the same set of nuclear processes dominate the others: component manufacturing, uranium enrichment, weapons

DOI: 10.1002/asi

design, nuclear fission reactors, enrichment feedstocks production, and tritium production. The observation that reprocessing is not getting more attention, along with the dominance of uranium enrichment, may speak more to the prevalence of American once-through fuel cycles in commercial nuclear power over the French recycling/reprocessing approach to waste management.

The significance of component manufacturing speaks to just how difficult this process is, involving various research questions. A lot of attention in the media is focused on plutonium extraction and uranium enrichment rather than component manufacturing. Despite this focus, the most accurate indicator of nuclear capabilities that a country may have developed is the ability to produce the required components; it is a major competitive edge that advanced nations like the United States, Japan, and France have over developing countries.

In addition, the difference in the scale of research coverage of various processes is partially affected by the differences in the scopes of different processes. Component manufacturing, for instance, is a big research field in its own right, whereas tritium production is a far smaller field of research. One other point is that testing does not show up as a highly ranked process in nuclear physics research articles. This is because the way we mark an article to have testing content is rather conservative. We specifically focus on testing for weaponization purposes instead of any nuclear experiments the researchers might perform.

Nuclear capabilities of countries. In this section, we discuss how we can interpret the contents of nuclear research articles as nuclear weaponization capability indicators at a country level. Although both experimental and theoretical articles are dominated by discussions on a similar set of nuclear processes, we consider experimental articles to provide a better indicator of nuclear capabilities due to investments and facility construction time, and the costs involved in nuclear experiments.

Within the set of experimental articles, a significant number are written by multinational, multiorganizational collaborations. To give a few examples, the STAR Collaboration has institutions from Poland, China, United States, Czech Republic, Brazil, Germany, India, Russia, South Korea, Netherlands, Croatia, and France. The INDRA Collaboration has researchers from France, Italy, Poland, and Canada. As discussed previously, countries that are identified as emerging countries (except China) are usually not a part of these collaborations. China is a special case, as it is already known to have significant nuclear capability. Its significant growth rate is what makes it a rapidly emerging nuclear state.

As far as country positions in terms of nuclear capability, we primarily focus on two key processes. One is component manufacturing, in which the United States, France, and Japan have a significant advantage over other countries. The other is testing. Testing is perhaps the sole indicator of the ultimate level of weaponization. For instance, it is possible to do research on weapon design and write articles while holding certain assumptions about the acquisition of enriched uranium. Similarly, it is possible to conduct research on both enrichment and design capabilities in parallel. However, publishing results at the testing level is different, as it requires having the materials, the technology, and the weapon(s) ready for detonation.

Considering their prominence in these two lines of research (component manufacturing and testing), we classify the countries as having different levels of nuclear weaponization capability. We loosely cluster countries into three levels, with Level 1 countries representing the countries with the highest level of nuclear capability based on the number of publications and researchers working in the area. More specifically, we rank countries in terms of these two criteria and cluster them according to gaps between the rankings.

Level 1: United States, France, Japan, Canada, Russia, Italy

Level 2: Brazil, Denmark, Finland, Germany, Greece, Hungary, India, Japan, South Korea, Mexico, Norway, Palestine, Israel, Romania, Switzerland, China, UK

Level 3: Egypt, Morocco, Slovakia, South Africa, Argentine, Australia, Armenia, Sweden, Uzbekistan, Austria, Ukraine, Colombia, Belarus, Portugal, Netherlands, Czech Republic, Turkey, Indonesia, Belgium, Bulgaria, Croatia, Chile, Poland, Spain, Taiwan, Serbia, Kazakhstan, Cyprus (Southern, Greek part).

To investigate the question of which countries do research on which specific processes, we also list countries interested in some of the other nuclear processes. We exclude the list of countries researching components manufacturing, enrichment processes, enrichment feedstocks production, nuclear weapons design, and nuclear fission reactors, as almost all countries interested in nuclear research look into these areas. However, each of the remaining processes receives attention from different subsets of countries, which makes it interesting to list countries per nuclear process, as they are further indicators of nuclear capability.

Custody transport control: United States, Russia

Heavy water production: United States, France, Japan, Canada, Denmark, India, Israel

Plutonium extraction: Australia, France, Germany, Palestine, Russia, United States

Tritium production: Australia, France, Germany, Canada, Israel, Italy, Japan, South Korea, Poland, Russia, South Africa, Spain, United States

Weapons testing: Brazil, Canada, Denmark, Finland, Germany, Greece, Hungary, India, Italy, Japan, South Korea, Mexico, Norway, Palestine, Romania, China, Russia, Switzerland, United States

Discussion

It is important to keep in mind that inferences presented in our results are (a) based on a subset of the research on

DOI: 10.1002/asi

nuclear capability, namely that on nuclear physics, and (b) that it accounts only for unclassified literature. The fact that the data used in this study come from open-source material has two limitations. First, our data do not include research published in nuclear states of North Korea, Iran, and Pakistan, i.e., known nuclear proliferators. Although this situation might give us some hints about the general stance of the researchers from these countries on volunteering their data to publicly open, international outlets, it is still a limitation that the evolution of nuclear knowledge within these countries was not part of our discussion above. However, this does not necessarily reflect on the quality of the methodology—the computational framework—proposed in this article, as these methods would be applicable to data from those countries as well.

Another point that might be considered a limitation is that countries such as the United States, Russia, and China conduct a considerable amount of classified nuclear research (Zuberi, 2000). The results presented in this section pertain only to nuclear research that occurs in the open-source literature and provide assessments of a country's position in terms of its unclassified nuclear knowledge production.

In this article, we focus on the research capabilities of different countries as exposed by their scientific publications. The incorporation of additional data, from both physics and non-physics article databases, is an important next step. After all, it is undeniable that a lot of research in the fields of chemistry, material science, chemical engineering, and others may prove relevant to this case study. Efforts to incorporate more publications into this investigation will require, at the very least, much manpower. Moreover, not all countries are producers of research/knowledge in this area; some of them are silent consumers. Therefore, there are numerous directions open for further research that might supplement the findings we have reported.

One potential extension of this research would involve examining researcher mobility, tracking researchers from which countries visit which other countries. When the authors are tagged with certain nuclear capabilities in accordance with the content of their articles, this would enable understanding the spatiotemporal patterns of the dissemination of knowledge on various nuclear processes and capabilities.

One other possible direction is to examine the nuclear material or technology trade nuclear power countries are involved in and perform a comparison of dates of activity. For instance, one country might start buying materials for research on a certain nuclear process and start producing related publications in the following years. Performing such joint, multivariate analysis might enable early detection of emerging nuclear capabilities.

Another potential direction would be to focus on case studies that closely investigate countries with relatively closed regimes such as Iran, Pakistan, and North Korea. These countries have fewer publications that appear in the international areas, and their researchers do not necessarily volunteer their publications to the venues researchers from other countries do. Case studies that would handle these countries as special cases in nuclear capability assessment might prove fruitful.

Conclusions

This article provides a comprehensive dynamic network analysis of a subset of publications in the field of nuclear research. In particular, we build and analyze interlinked dynamic networks that are extracted from experimental and theoretical nuclear physics publications from the last two decades. Our goal is to provide a methodological, computational analysis to answer the questions of which countries are influential key players in the nuclear world, which countries have recently gained interest in this line of research, and what level of nuclear capability each country has acquired, as reflected in the stock of publications authored by its researchers.

As an intermediate step to forming this bigger picture, we present an extensive array of results that reveal the key entities involved at the organization, city, state, and country levels. In addition, our results also cover methods for distinguishing major versus emergent players using social network analysis techniques. The final part of our results demonstrates nuclear weaponization capability levels acquired by countries, along with a breakdown of publications according to the specific area of expertise they relate to.

To briefly summarize, our findings reveal that the United States, Germany, France, Russia, Japan, and Italy are leaders in the production of nuclear knowledge, while Kazakhstan, Turkey, and Egypt are among the emerging nuclear knowledge producers of the world. Along the same lines, China, the world's fastest growing economy, emerges as the country with the fastest growing nuclear infrastructure, with many reactors under construction and many more currently being planned. This growth in deployment is also well reflected in the publications emerging from China, as our network analysis has suggested.

Acknowledgments

The authors would like to thank Matthew Wachs and anonymous reviewers for insightful comments that helped us revise and improve our article. Financial support was provided by the Defense Threat Reduction Agency (DTRA) under grant number HDTRA11010102. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Threat Reduction Agency (DTRA) or the U.S. government.

References

Airoldi, E., & Carley, K.M. (2005). Sampling algorithms for pure network tologies. SIGKDD Explorations, 7(2), 13–22.

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512.
Bethlehem University. (2011). Brief history. Retrieved from http://www.bethlehem.edu/about/history.shtml
Bikel, D.M., Schwartz, R., & Weischedel, R.M. (1999). An algorithm that learns what's in a name. Machine Learning, 34(1–3), 211–231.
Bonacich, P. (1987). Power and centrality: A family of measures. American Journal of Sociology, 92(5), 1170–1182.
Boneau, A.C. (1960). The effects of violations of assumptions underlying the t test. Psychological Bulletin, 57(1), 49–64.
Borgatti, S.P., & Everett, M.G. (2000). Models of core/periphery structures. Social Networks, 21, 375–395.
Borland, P. (2003). The concept of relevance in IR. Journal of the American Society for Information Science and Technology, 54(10), 913–925.
Brookhaven National Laboratories. (2011). Star collaboration. Retrieved from http://www.star.bnl.gov/
Carnegie Mellon University. (2011a). CASOS Research Center. Retrieved from http://www.casos.cs.cmu.edu/projects/ora/
Carnegie Mellon University. (2011b). CASOS Research Center. Retrieved from www.casos.cmu.edu/projects/automap
Cornell University. (2011). Retrieved from http://arxiv.org/
Department of Defense Security Institute. (2010). The Militarily Critical Technologies List (MCTL) (Security Awareness Bulletin Number 2-95). Richmond, VA: Author.
Diesner, J., & Carley, K.M. (2008, June). Conditional random fields for entity extraction and ontological text coding. Journal of Computational and Mathematical Organization Theory, 14(3), 248–262.
Embassy of the Republic of Kazakhstan. (2011, August). Embassy of the Republic of Kazakhstan—Brief history. Retrieved from: http://www.kazakhembus.com/
Federation of American Scientists (FAS). (2011). Federation of American Scientists homepage. Retrieved from www.fas.org
Ferguson, C., & Potter, W. (2005). The four faces of nuclear terrorism. New York: Taylor & Francis Group.
Frantz, T.L., & Carley, K.M. (2005). Relating network topology to the robustness of centrality measures. CASOS. Pittsburgh, PA: Carnegie Mellon University ISR.
Freedman, L. (2002). Pugwash Meeting no. 279. London, England. Retrieved from http://www.pugwash.org/reports/nw/freedman.htm
Freeman, L.C. (1977). A set of measures of centrality based on betweenness. Sociometry, 40, 35–41.
Freeman, L.C. (1979). Centrality in social networks conceptual clarification. Social Networks, 1(3), 215–239.
Government Accountability Office. (2006). DOD's critical technologies lists rarely inform export control and other policy decisions. Washington, DC: Government Printing Office.
Graham, A. (1996). Avoiding nuclear anarchy: Containing the threat of loose Russian nuclear weapons and fissile material. Cambridge, MA: MIT Press.
Hummon, N., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. Social Networks, 11(1), 39–63.
INIS. (2011). International nuclear information system. Retrieved from www.iaea.org/inisnkm
INP-Kazakhstan. (2011). Institute of Nuclear Physics of the National Nuclear Center, Republic of Kazakstan homepage. Retrieved from http://www.inp.kz/
Kazakh State University. (2011). Al-Farabi, Physics-Technical Education in the Republic of Kazakhstan homepage. Retrieved from http://www.kaznu.kz/en/353/page
Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05) (pp. 177–187). Chicago, IL: ACM.
Levi, M. (2007). On nuclear terrorism. Cambridge, MA: Harvard University Press.
Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. Proceedings of National Academy of Sciences, 101, 5200–5205.
Nolan, J. (1989). Guardians of the arsenal: The politics of nuclear strategy. New York, NY: Basic Books.
NTI. (2010, May). Kazakhstan profile. Retrieved from http://www.nti.org/country-profiles/kazakhstan/
Nuclear Energy Institute. (2005). Nuclear power plant emergency preparedness. Retrieved from http://pbadupws.nrc.gov/docs/ML0602/ML060230348.pdf
Palestinian Energy Department. (1998). Nuclear Energy and Radiation Protection Department homepage. Retrieved from http://pea-pal.tripod.com/nuclear_energy_and_radiation_pro.htm
Reid, E. (1993). Terrorism research and the diffusion of ideas. Knowledge, Technology & Policy, 6(1), 17–37.
Rényi, A., & Erdös, P. (1959). On random graphs. Publicationes Mathematicae, 6, 290–297.
Sabidussi, G. (1966). The centrality index of a graph. Psychometrika, 31(4), 581–603.
Sagan, S.D. (2011). The causes of nuclear weapons proliferation. Annual Review of Political Science, 14, 225–244.
Sampson, S. (1969). Crisis in a cloister (Unpublished doctoral dissertation). Ithaca, NY: Cornell University.
Saracevic, T., & Kantor, P. (1988a). A study of information seeking and retrieving. II. Users, questions and effectiveness. Journal of the American Society for Information Science, 39(3), 177–196.
Saracevic, T., & Kantor, P. (1988b). A study of information seeking and retrieving. III. Searchers, searches and overlap. Journal of the American Society for Information Science, 39(3), 197–216.
Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50(9), 799–813.
Tague-Sutcliffe, J.M. (1996). Some perspectives on the evaluation of information retrieval systems. Journal of the American Society for Information Science, 47(1), 1–3.
United States Nuclear Regulatory Commission. (2011). FY 2012 budget press briefing. Retrieved from http://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr1100/v27/fy2012-press-briefing.pdf
Wasserman, S., & Faust, K. (1994). Social network analysis. Cambridge, England: Cambridge University Press.
Watts, D., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. Nature, 393(6684), 440–442.
White, H.D., & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. Journal of the American Society for Information Science, 32(3), 163–171.
White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. Journal of the American Society for Information Science, 49(4), 327–355.
World Nuclear Association. (2011). World Nuclear Association homepage. Retrieved from www.world-nuclear.org
Zuberi, M. (2000). Soviet and American technological assistance and the pace of Chinese nuclear tests. Strategic Analysis, 24(7), 1247–1266.

DOI: 10.1002/asi