SAND2018-2822C

Collaborative Analytics for Biological Facility Characterization

Jacob P. Caswell[a], Kelsey L. Cairns[a], Christina L. Ting[a], Mark W. Hansberger[b], Matthew A. Stoebner[b], Thomas R. Brounstein[a], Christopher R. Cueller[a], and Elizabeth R. Jurrus[b]

[a]Sandia National Laboratories, Albuquerque, NM
[b]Defense Threat Reduction Agency, Ft. Belvoir, VA

ABSTRACT

Thousands of facilities worldwide are engaged in biological research activities. One of DTRA's missions is to fully understand the types of facilities involved in collecting, investigating, and storing biological materials. This characterization enables DTRA to increase situational awareness and identify potential partners focused on biodefense and biosecurity. As a result of this mission, DTRA created a database to identify biological facilities from publicly available, open-source information. This paper describes an on-going effort to automate data collection and entry of facilities into this database. To frame our analysis more concretely, we consider the following motivating question: *How would a decision maker respond to a pathogen outbreak during the 2018 Winter Olympics in South Korea?* To address this question, we aim to further characterize the existing South Korean facilities in DTRA's database, and to identify new candidate facilities for entry, so that decision makers can identify local facilities properly equipped to assist and respond to an event. We employ text and social analytics on bibliometric data from South Korean facilities and a list of select pathogen agents to identify patterns and relationships within scientific publication graphs.

Keywords: bibliometrics, graph analytics, collaboration analytics, open source analytics, biodefense, biosecurity, community detection

1. INTRODUCTION

DTRA's database of biological facilities is designed to provide information related to potential pathogen repositories worldwide to support capability assessments of a country's bio-infrastructure and other bio-related activities. This database is composed of consolidated open-source information and provides a novel user interface prototype for displaying all pathogen repository data points, which allows decision makers to better understand the location of research facilities, hospitals, diagnostic laboratories, etc., that collect, investigate, and/or store biological materials. Due to the open-source nature of its information, this database can be used to facilitate cross-agency sharing of information regarding vital biological and technical content to aid in the development of i) responses to global bio-related humanitarian crises and ii) countermeasures to prevent pathogens from impacting local communities.

Since 2012, the Defense Threat Reduction Agency (DTRA) has identified over 1750 facilities worldwide, based on limited open-source research. Of these facilities, roughly half are considered fully characterized within the database. However, work is still required to ingest data to improve both the database's characterization of existing facilities and to discover new facilities that, if included, could help DTRA and its global partners react to threats to public health. This effort is enabled by the proliferation of freely available information, including public literature repositories such as PubMed, which provide ready access to references, abstracts, and links to full-text articles on life sciences and biomedical topics. However, the sheer volume of information necessitates the development of analytics for a semi-automated, analyst-driven, pipeline for data ingestion and entry into DTRA's database. DTRA has partnered with Sandia National Laboratories to leverage the laboratories' capabilities in data analytics, with a focus here on bibliometrics.

E-mail: ERJ: elizabeth.r.jurrus.civ@mail.mil

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Pathogen Name	Current Identifier	Other Identifiers
Avian Influenza virus (highly pathogenic)	influenza a	hpai, h5n1
Bacillus anthracis	anthracis	anthrax
Clostridium botulinum (pathogen and toxin)		botulinum
Burkholderia mallei	mallei	
Burkholderia pseudomallei	pseudomallei	
Ebola virus		ebola, ebolavirus
Foot-and-mouth disease virus		fmdv
Francisella tularensis	tularensis	
Marburg virus		marburg, marburgvirus
Reconstituted 1918 Influenza	influenza a	h1n1, spanish flu
Rinderpest virus		rinderpest, morbillivirus
Variola major		variola
Variola minor		variola
Yersinia pestis	pestis	

Table 1. List of select pathogen agents used in this study.

Analysis of bibliometric data may utilize text and social analytics to identify patterns and relationships within scientific publication graphs.[1–3] These analytics may also be utilized to interpret the flow of information through communities of agents,[4] as well as identify sub-communities given complex rules governing inter-agent relationships.[5] Additionally, graph analytics have been shown to provide weak predictors of collaborative behavior by encoding similarities among agents such as authors or institutions.[6] Combining institution-level text and social analytics to construct various community graphs, this paper will document the effort to enhance DTRA's ability to characterize known facility information and to identify new potential data points for inclusion in DTRA's database, all while reducing the manual load imposed on analysts.

This paper is structured as follows. Section 2 discusses the data, including a description of the steps necessary for collection and pre-processing. Section 3 contains a description of our graph-fusion pipeline and graph analytics for community detection and characterization. Section 4 provides results and analyses of graphs constructed from existing data and newly-ingested PubMed data. Finally, Section 5 concludes with a summary and brief discussion of future work.

2. DATA

2.1 Data Collection

This section describes the two datasets used in this work: the existing DTRA dataset and a PubMed dataset collected from PubMed's publicly available metadata. This work is framed in the context of a motivating question: *What resources are available to respond to a pathogen outbreak during the 2018 Winter Olympics in South Korea?* By focusing on this question, we are able to focus our dataset to South Korean facilities.

2.1.1 DTRA Dataset

DTRA's database is a small, hand-curated database of over 1750 facilities worldwide. Each entry in the dataset represents an institution and contains data related to the activities, pathogens and equipment present at the facility. Each data field has been parsed and added to our dataset. The following is a sample of the potentially interesting and relevant fields: a) *facility name*, b) *country*, c) specific location, d) *facility description*, e) *pathogens stored at the facility*, and f) known equipment. The results and analyses described in this work will make use of the italicized fields. The database, when filtered to South Korea, contains 72 facilities. Of these facilities, 15 contain a facility description and 14 contain pathogens. For the remainder of the paper, the *DTRA dataset* refers to the reduced dataset of 72 South Korean facilities with a description and/or known pathogens.

2.1.2 PubMed Dataset

PubMed openly publishes metadata corresponding to its vast collection of papers from medical and biological fields. All data is free to download through PubMed's web interface.[7] We started by collecting all records containing 'Korea' in the affiliation field that also mentioned at least one of the pathogens from Table 1 anywhere in the record. This approach returned 311 papers. We parsed these records for the following fields: a) PubMed identifier (PMID, a unique value assigned to each PubMed record), b) article title, c) author list, d) list of institutions affiliated with each paper, e) abstract (if available), and f) keywords (if available).

For this study, we focus on the institutions present in the PubMed data. Once parsed, our dataset contains over 4,600 unique institution strings. Each string contains the name of an institution, as well as location information including the institution's country. Filtering to remove institutions that did not include "Korea" in the name yielded 566 unique institution strings.

Although the PubMed dataset contains many more institutions than are present in DTRA's dataset, the formatting of institution strings is extremely inconsistent and sometimes contains typographic errors. Furthermore, many of these strings actually refer to the same physical institution, necessitating a cleaning and disambiguation phase before applying further analysis. The disambiguation process (described in Section 2.2) yields 128 South Korean facilities. For the remainder of the paper, the *PubMed dataset* refers to the reduced dataset of 128 disambiguated South Korean facilities and their associated papers, on which we perform the analyses described in this work.

2.2 Facility Name Disambiguation

The lack of standardization in institution names necessitates the need to disambiguate these names, which is the process of identifying groups of distinctly formatted strings that refer to the same physical institution. Disambiguation proved necessary for two reasons: 1) to determine when two facilities listed in the affiliation list on different PubMed documents are in actuality the same facility with a different format, and 2) to map facilities found in the *DTRA dataset* to facilities appearing in the *PubMed dataset*.

The disambiguation process comprises two phases, each with several steps. The first phase consists of cleaning the data, and removing inconsistencies that interfere with the second phase, which involves grouping institution strings that correspond to the same physical institution. In the following sections, we describe each phase separately.

2.2.1 Cleaning

Several of the affiliation strings contain multiple institutions; others include additional information about the authors such as email addresses and telephone numbers. In order for the second phase of the disambiguation process to work well, it is necessary to split affiliation lists into as many distinct institution name strings as possible and to remove as much of the unrelated information as possible. This cleanup process involves a series of semi-supervised steps involving visual inspection of the data to determine how to split strings and what patterns to remove.

To split affiliation strings into multiple institutions, common delimiters are identified and used to separate institutions.[*] To remove unrelated information, patterns that indicate email addresses, parenthesized phrases, and common name suffixes (e.g., "PhD") are used to trigger removing characters. The strings are also standardized (e.g. ampersands are converted to "and") and run through a spell checker.

Despite splitting many affiliation lists into multiple new institution name strings, this cleaning process ultimately reduced the number of unique strings. In many instances, removing unwanted characters resulted in exact matches with other strings. After cleanup, 517 out of 566 unique strings remained to be disambiguated.

[*] Separators used in this work: ";", "Korea and", "Korea ", "Korea, and", "Korea.", " USA.", " USA,", "*", "‡", "§", "||", "†", "[2]", "[3]", "[4]", "[5]", "[6]"

Step	Counts
Unique strings parsed from raw data	4606
Unique strings parsed from raw data containing “Korea”	566
Unique strings after cleaning	517
Institutions found by matching algorithm	129
Institutions after manual correction	128

Table 2. Reduction in the number of strings through the disambiguation process, which involves a cleaning and a matching phase.

2.2.2 Matching

After the institution name strings are considered clean, they are run through a matching algorithm. The matching algorithm considers strings in a pairwise fashion and determines whether or not two strings should be merged, meaning they represent the same institution. Merged strings are included together in a set, along with any other strings they had previously been merged with. Each set denotes a unique institution. Thus, the output of this algorithm is a grouping of institution name strings, where each group represents a unique physical institution.

At a high level, the process involves searching for a match between *important* substrings from two different institution name strings. If a match is found, the institution name strings are considered the same physical institution. This process is outlined below. For each pair of facilities:

1. Split each name string into substrings based on commas.
2. Identify substrings containing at least one keyword from a prioritized list of keywords.[†]
3. Identify *important* substrings as those containing the n highest-priority keywords found within the substrings.
4. Compare the sets of *important* substrings for two facilities. If a substring in one set contains a substring in the second set, a match is found indicating that the two institution name strings from which the substring sets were derived are the same physical institution.

For example, by setting “university” as a higher priority keyword than “institute”, and by excluding “department” from our list of keywords, this process attempts to match facilities at the university level, followed by the institute level, all while ignoring the department level. The disambiguation algorithm reduced the 517 unique name strings to 129 unique institutions. Manual verification of this list found one pair of strings that failed to be merged. The final list, after manually merging the final pair, contained 128 institutions. Reduction of the number of strings through the cleaning and disambiguation process is summarized in Table 2.

3. METHODS

Graph analytics have been shown to provide weak predictors of collaborative behavior by encoding similarities between agents such as authors or institutions.[6] Combining institution-level text and social analytics to construct various community graphs, this section will describe our graph-fusion pipeline for generating fused-similarity graphs and discuss the discovery analytics we use to identify or characterize potentially interesting research communities for analyst-driven entry into DTRA’s database.

[†] The prioritized list of keywords used in this work: “samsung”, “university”, “universe”, “administration”, “hospital”, “ministry”, “agency”, “brewery”, “institute”, “institut”, “academy”, “arboretum”, “foundation”, “laboratory”, “labs”, “center”, “facility”, “school”, “service”, “bank”, “products”

```
/* ---- Fusion Pipeline Example---- */

// Adjacency Matrix Creation
graph = new graphMatrix(numNodes)

// Pipeline processors
// Draws all edges for shared pathogens.
g1=boolSharedPathogenProcessor.compute()
// Selects only the edges in which the institution have collaborated.
g2=boolCollabProcessor.compute()
// Boolean AND corresponding entries in g1,g2
graph = andProcessor.compute(g1, g2)

// Gephi format writer
gexfGraphWriter.write(graph, filename)
```

Figure 1. An example fusion pipeline that will create a graph of institutions who share pathogens and collaborate, then write the graph to a Gephi-compatible format.

3.1 Graph Fusion Pipeline

To aid rapid and modular analytic development, we developed a graph fusion pipeline that is capable of fusing basic similarity processors to produce graphs which encode composite measures of similarity. To develop such a pipeline, we represent graphs as adjacency matrices $\mathbf{G} \in \mathbb{R}^{N \times N}$ where a given entry, $g_{ij} \in [0, 1]$, encodes a similarity score used to draw edges between institutions i and j, and N represents the number of nodes in $\mathbf{G}$. A series of user-defined similarity processors and fundamental operator (e.g., AND, XOR, NOT, ADD, MULT) processors then compute various similarity scores between nodes i and j, and combines the scores. For the PubMed data, we consider similarity scores based off collaboration or shared pathogens. Edges in the graph may or may not be weighted. A weighted edge directly corresponds to the calculated similarity score. However, in an unweighted graph, edges are present if and only if some similarity criteria is met. We call these boolean similarity metrics. Finally, a writer translates the adjacency matrix into a graph format interpretable by existing graph analytics software such as Gephi.[8]

An example pipeline is provided in Fig. 1 as a Groovy script that calls on Citrus,[9] a text analysis library developed at Sandia National Laboratories. The user-defined similarity processors, `boolSharedPathogenProcessor` and `boolCollabProcessor`, in Fig. 1 are processors that will construct an edge between two institutions i and j if the two institutions share at least one pathogen or have collaborated on at least one publication, respectively. The fundamental `andProcessor` will select only the edges for which the two institutions have collaborated AND share pathogens. By combining fundamental operation processors and user-defined similarity processors, we allow for the generation of arbitrary graph fusions.

Weighted similarities based on shared pathogens or collaborations make use of the Jaccard similarity coefficient to determine a similarity score. For pathogen similarity,

$$J^{p}_{ij} = |P_i \cap P_j| / |P_j \cup P_j|, \tag{1}$$

where P_i is the set of pathogens for institution i. Similarly for collaboration similarity,

$$J^{c}_{ij} = |C_i \cap C_j| / |C_j \cup C_j|, \tag{2}$$

where C_i is the set of publications for institution i. The score of the fused graph using the AND operator is then the product of the two scores for pathogen and collaboration:

$$g_{ij} \equiv J^{p}_{ij} \times J^{c}_{ij}. \tag{3}$$

Once these similarity scores have been fused, we may then view the resulting adjacency matrix as a graph and utilize graph analytics algorithms for the discovery and characterization of research communities.

An additional similarity metric, which we applied to the *DTRA dataset*, is the similarity between facility descriptions. Our similarity metric for text descriptions uses a metric called *term frequency-inverse document frequency* ($\mathrm{TFIDF}_{t,d}$).[10] $\mathrm{TFIDF}_{t,d}$ is commonly used in information retrieval and text mining as a reflection of how important a term t is to a document d, relative to a corpus of N documents. It is important to note that here "document" really refers to "facility description", since we are comparing facilities. Several versions of $\mathrm{TFIDF}_{t,d}$ exist. We use the default version implemented in Citrus, which is defined by, $\mathrm{TFIDF}_{t,d} = \mathrm{tf}_{t,d} \times \log(N/\mathrm{df}_t)$, where the first term $\mathrm{tf}_{t,d}$ is the term frequency and the second term idf_t is the definition for the inverse document frequency. Importantly, this expression is highest when a term t shows up many times only within a small number of documents.

Description similarity scores for between facilities are based off of the cosine distance between $\mathrm{TFIDF}_{t,d}$ vectors. For each facility, we can construct a document vector, $\mathbf{V}(d)$. Each vector has n components, corresponding to the $\mathrm{TFIDF}_{t,d}$ weight of the n terms in the dictionary. The similarity of two documents d_1 and d_2 is then defined by,

$$\mathrm{sim}(d_1, d_2) = \mathbf{v}(d_1) \cdot \mathbf{v}(d_2), \tag{4}$$

where $\mathbf{v}(d)$ is the document vector normalized by the Euclidean norm: $\mathbf{v}(d) = \mathbf{V}(d)/||\mathbf{V}(d)||$.

3.2 Discovery Analytics

The graphs that are written out at the end of the fusion pipeline are ingested into Gephi[8] to facilitate visualization and analysis of the graphs. In particular, Gephi provides a modularity optimization algorithm known as the Louvain community detection algorithm.[11] Modularity measures the density of edges inside verses the density of edges between assigned communities. Optimizing the modularity theoretically results in an optimal grouping of the nodes of a given graph.

Once we have applied Louvain for community detection, a major issue remains with characterizing the results of these communities. One option is to compute a degree of *focus* of the pathogen profile for each community, as defined by the entropy:

$$H_j = -\sum_i p_{ij} \log p_{ij}, \tag{5}$$

where p_{ij} is the probability of pathogen i in community j. H_j is calculated by taking the sum over each of the pathogens in community j. A second option is to compute the *spread* of the top pathogen $\mathcal{P}$ in community j, where $\mathcal{P}$ is defined as the pathogen present at the highest number of institutions in community j. We can then define α_j, which is the proportion of facilities in community j that have published on $\mathcal{P}$:

$$\alpha_j \equiv \frac{n_{\mathcal{P}}}{n_j}. \tag{6}$$

Here $n_{\mathcal{P}}$ is the number of institutions in j associated with $\mathcal{P}$ and n_j is the number of institutions in community j. A third option is to compute the relative number of known facilities in community j:

$$\beta_j \equiv \frac{n_D}{n_j}, \tag{7}$$

where n_D is the number of institutions in j that also appear in the DTRA dataset. This metric can be used to link discovered communities to established ground truth of known facilities.

Each of these characterizations may be used to help an analyst determine if a community and its associated facilities warrant further attention. Additional metrics used in this work are described as they are introduced.

4. RESULTS

In what follows, we discuss results obtained from analysis of the *DTRA dataset* and the *PubMed dataset*.

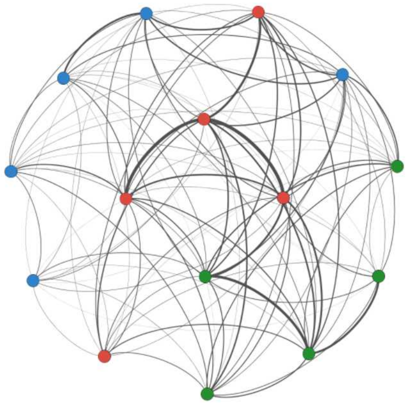

Figure 2. Graph constructed from the description similarity between known institutions. Node colors corresponds to community and edge weights represent strength of the similarity score.

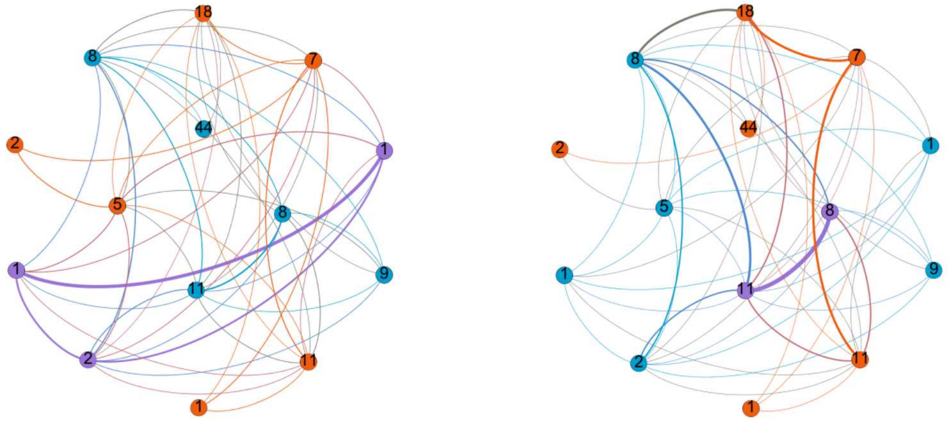

Figure 3. Pathogen similarity between institutions. Edge weights on the left graph were computed using the Jaccard similarity coefficient. Edge weights in the graph on the right reflect total number of pathogens common between facilities. The nodes are labeled by the size of the pathogen set.

4.1 DTRA Database

To set up a baseline for comparison with results from our *PubMed dataset*, we construct graphs based on the *DTRA dataset*, where nodes represent known facilities and edge weights represent a similarity score between two facilities. In this work, we consider similarities based on 1) the cosine distance of $\text{tfidf}_{t,d}$ vectors constructed from the facility description; 2) the Jaccard distance of the pathogens stored at the facilities; and 3) the total number of pathogens shared between two facilities, essentially the un-normalized Jaccard distance. Of the 72 South Korean facilities in the DTRA database, 15 contain a facility description and 14 contain pathogen attributions, where we have not limited ourselves to the pathogens in Table. 1. The resulting graphs are quite small, further motivating the use of PubMed data to enhance the DTRA database.

Following construction of the graphs, we apply the Louvain algorithm[11] in Gephi to identify communities of institutions. In the following figures, node color represents community, and edge thickness reflects similarity scores between institutions. In Fig. 2 we show the facility description similarity graph, with similarities computed as defined in Eq. 4. We can see that all facilities are related to some extent, i.e. the graph is fully connected. Although the community detection algorithm identifies three communities, each with several of the most strongly connected nodes in the graph, manual inspection of the facility descriptions of the most strongly connected nodes does not reveal any notable similarity. This is not surprising, however, as the cosine similarities range from

$0.01 - 0.28$, indicating that even the most similar descriptions are not very similar.

Next, in Fig. 3, we show the similarity graphs constructed from the overlap in pathogen sets between DTRA's known facilities. The left graph shows edges weighted by the Jaccard similarity, J^p_{ij}; the right graph shows edges weighted by the number of overlapping pathogens, essentially the un-normalized Jaccard similarity, $\bar{J}^p_{ij} = |P_i \cap P_j|$. The nodes are labeled by the number of pathogens listed in the *DTRA dataset*. By comparing these two graphs, we can see that the graph based on J^p_{ij} (left) tends to emphasize focused similarity of pathogens, where three purple nodes with $1 - 2$ pathogens each are connected by the strongest edge weights and are therefore detected as their own community. In contrast, the graph based on $\bar{J}^p_{ij}$ (right) identifies overall similarity of pathogens, where now the nodes connected by the strongest edge weights contain $7 - 18$ pathogens. The two purple nodes represent the most similar facilities, with 5 shared pathogens. Interestingly, the node with 44 pathogens is not strongly connected in either shared pathogen graph.

4.2 PubMed

In addition to characterizing known institutions from DTRA's database, we seek to develop a flexible framework to guide the analyst-driven discovery of new institutions and collaborative research communities using openly available data retrieved from PubMed. As discussed previously, this work is framed in the context of a pathogen outbreak during the 2018 Winter Olympics in South Korea so that the specific goals of our PubMed analysis are to identify and characterize the research communities of South Korean institutions that have published on the set of pathogens previously defined in Table 1. Our approach is to apply social and text analytics to construct fused-similarity graphs that encode 1) communities with a *shared* specific research topic and 2) communities with a *focused* research interest.

We examine the graphs based on four metrics, which are described in more detail in Section 3.2:

1. $\bar{n}$, the average community size.
2. $\bar{\beta}$, the average percentage of institutions from DTRA's database in discovered research communities.
3. $\bar{\alpha}$, the average coverage coefficient, which describes the proportion of institutions in a community that share the community's most prominent pathogen $\mathcal{P}$.
4. $\bar{H}$, the average entropy in the distribution of a community's focus over the set of pathogens.

Table 3 shows a summary of these statistics for the different graphs described in the following sections.

Graph	$\overline{n}$	$\overline{\beta}$	$\overline{\alpha}$	$\overline{H}$
Bool. Shared Pathogen	26	27.1%	0.831	2.315
Bool. Collaboration	11	15.1%	0.689	1.972
Bool. Composite	8.4	21.4%	0.801	2.236
Weighted Shared Pathogen	20.4	26.6%	0.928	1.167
Weighted Collaboration	6.2	19.6%	0.786	1.564
Weighted Composite	5.8	23.6%	0.873	1.343

Table 3. Comparison of graph level average statistics.

4.2.1 Boolean Graph Analysis

To obtain a community structure of research institutions from PubMed that can compare with the baseline community structure identified in DTRA's database, we first construct a boolean shared pathogen (BSP) graph. That is, a network of nodes (institutions) is connected by edges with weight $g_{ij} = 1$ if institutions i and j share at least one pathogen. Otherwise, $g_{ij} = 0$ (i.e. nodes i and j are unconnected). From this graph, we apply the Louvain community detection algorithm to identify research communities based on shared pathogens within the graph. From the BSP graph depicted in Fig. 4, we see two large well-connected communities dominating

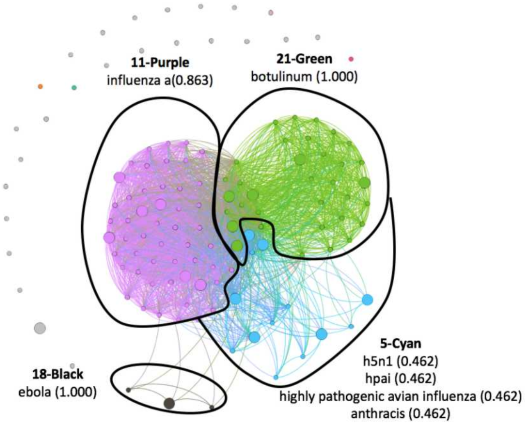

Figure 4. The boolean shared pathogen (BSP) graph visualized using Gephi's Fruchterman Reingold layout. Community labels denote top pathogen(s), $\mathcal{P}$, with the coverage, α, in parentheses.

ID	Color	Size	β	α	H
11	Purple	51	11.8%	0.863	2.520
21	Green	25	17.1%	1.000	2.603
5	Cyan	13	46.2%	0.462	4.137
18	Black	3	33.3%	1.000	0.000
Global Averages		**26**	**27.1%**	**0.831**	**2.315**

Table 4. Description of the community and graph level statistics for the BSP graph.

the graph; statistics of the communities are provided in Table. 4. Characterization of the pathogens in the communities reveals that the average coverage coefficient $\bar{\alpha} = 0.831$, suggesting that communities in the BSP graph contain a high coverage of institutions that publish on the top pathogen $\mathcal{P}$ of the community. This result makes intuitive sense, since the BSP graph is identifying similarity based a shared pathogen. Recall that the pathogen similarity graph constructed from DTRA's database consisted of well-connected nodes without any notable community structure. Similarly, the BSP graph constructed from the PubMed data contains a high-level community structure of well-connected institutions that share a predominant pathogen: 11-Purple has a coverage of $\alpha = 0.863$ (influenza a) and 21-Green has a coverage of $\alpha = 1.000$ (botulinium), providing a relatively distinct separation of institutions that favor influenza research and those that favor botulinum research. However, while this graph illustrates which institutions publish on similar pathogens and what those pathogens are, it does not reveal research community structure based on social collaboration.

To identify collaborative research communities in our PubMed dataset, we create a boolean collaboration (BC) graph. That is, a network of nodes (institutions) connected by edges with weight $g_{ij} = 1$ if institutions i and j have co-published at least one paper and $g_{ij} = 0$ otherwise; see Fig. 5. From this graph, we apply the Louvain community detection algorithm to identify highly collaborative research communities within the graph.[‡] The statistics of the communities from the boolean collaboration network are provided in Table 5.

Table 5 shows that, on average, the communities cannot be well characterized by a single defining pathogen, as indicated by an average coverage coefficient, $\bar{\alpha} = 0.689$. This trend is especially true of the three largest communities, 29-Purple, 17-Green, and 22-Cyan, whose coverage coefficient α falls below the graph average. Furthermore, for 17-Green and 22-Cyan, there exists no pathogen shared by a majority of the community's institutions. However, for the remaining smaller communities, at least half of their institutions share the most prevalent pathogen $\mathcal{P}$ in that community. This is not surprising, however, as the remaining communities consist mainly of either isolated or largely isolated research pairs and small communities. Within these smaller communities, it is easier to create a clear characterization of the pathogens present.

The decrease in $\bar{\alpha}$ between the BSP and BC graphs makes intuitive sense since the BC graph is not directly identifying similarity based a shared pathogen. However, the fact that the average entropy over the communities' pathogen distribution also decreases from $\bar{H} = 2.315$ in the BSP graph to $\bar{H} = 1.972$ in the BC graph is perhaps more surprising. The decreased entropy indicates an *increase* in research focus even though we are not directly using any information on pathogen similarity for the BC graph. This result can be understood by noting that

[‡] It should be noted that this dataset selected only South Korean institutions that had published on a specific list of pathogens. As such, while there are 29 disconnected institutions, they may have published with institutions outside of South Korea. Alternatively, they may have published with institutions within our dataset, but on topics not covered.

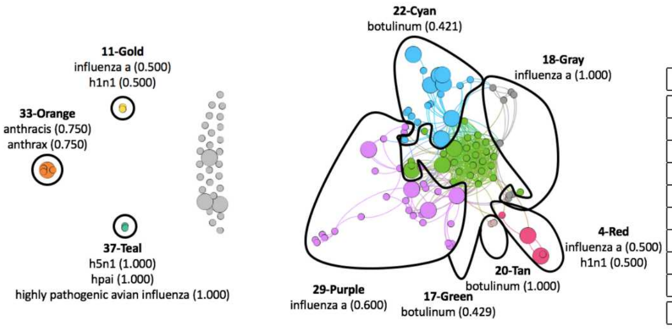

Figure 5. Boolean collaboration (BC) graph over detected institutions visualized using Gephi's Force Atlas layout. Community labels denote top pathogen(s), $\mathcal{P}$, with the coverage, α, in parentheses.

ID	Color	Size	β	α	H
29	Purple	30	13.3%	0.600	3.064
17	Green	28	10.7%	0.429	3.324
22	Cyan	19	36.8%	0.421	2.306
18	Gray	8	0.0%	1.000	1.264
33	Orange	4	50.0%	0.750	2.626
4	Red	4	25.0%	0.500	1.192
37	Teal	2	0.0%	1.000	2.250
20	Tan	2	0.0%	1.000	0.000
11	Gold	2	0.0%	0.500	1.000
Global Averages		**11**	**15.1%**	**0.689**	**1.972**

Table 5. Description of the community and graph level statistics for the BC graph.

t

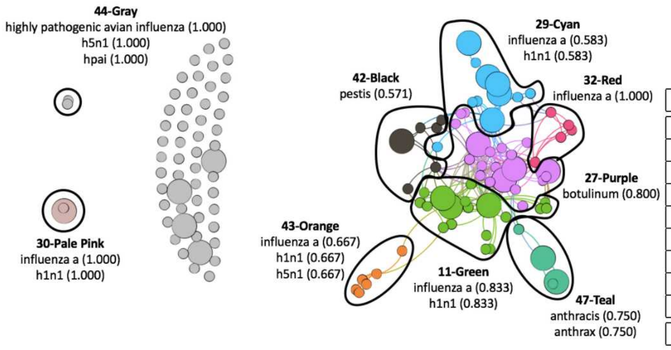

Figure 6. Composition of boolean collaboration (BC) applied to boolean shared pathogens (BSP) visualized using Gephi's Force Atlas layout. Community labels denote top pathogen(s), $\mathcal{P}$, with the coverage, α, in parentheses.

ID	Color	Size	β	α	H
27	Purple	20	20.0%	0.800	3.268
11	Green	18	16.7%	0.833	2.967
29	Cyan	12	41.7%	0.583	1.968
42	Black	7	14.3%	0.571	1.959
43	Orange	6	0.0%	0.667	2.308
32	Red	5	0.0%	1.000	0.863
47	Teal	4	50.0%	0.750	2.626
30	Pale Pink	2	50.0%	1.000	1.918
44	Gray	2	0.0%	1.000	2.250
Global Averages		**8.4**	**21.4%**	**0.801**	**2.236**

Table 6. Description of the community and graph level statistics for the BC-BSP graph.

in the BSP graph, an edge is drawn between any two facilities that share at least one pathogen, independent of the number of pathogens that are not shared. These pathogens that are not accounted for when identifying communities in the BSP graph decrease the focus of a community's pathogen distribution. Thus, when not accounting for the BSP similarity, the entropy H of the graph decreases and the research focus increases.

In order to combine the similarities in the BC graph with the similarities in the BSP graph, we fuse the two graphs such that a network of nodes (institutions) is now connected by edges with weight $g_{ij} = 1$ if institutions i and j have co-published at least one paper *and* share at least one pathogen. Otherwise, $g_{ij} = 0$. This graph is shown in Fig. 6. The resulting composite boolean collaboration and boolean shared pathogen (BC-BSP) graph structurally resembles the BC graph. However, a direct comparison of the two graphs reveals a more refined community structure in the BC-BSP graph. In particular, the main community of green nodes in Fig. 5 is no longer present in the composite fused graph accounting for BSP. This is because the abstract, title or keywords of the publication that the nodes published on did not share a pathogen, and thus were removed by the shared pathogen processor filter. The result is nodes being clustered into a greater number of smaller communities, as can be seen by the reduction in the average community size from $\bar{n} = 11$ to $\bar{n} = 8.4$.

At the community level, we see that fusing the BSP graph into the BC graph increases the coverage coefficient from $\alpha = 0.689$ for the BC graph to $\alpha = 0.801$ for the BC-BSP graph, implying that certain nodes which did not share the predominant pathogen $\mathcal{P}$ migrated into new communities based on a common shared pathogen. However, we also see a general increase in entropy from $\bar{H} = 1.972$ to $\bar{H} = 2.236$ in these emerging communities. That is to say, by incorporating information about shared pathogens into the collaboration graph, we have

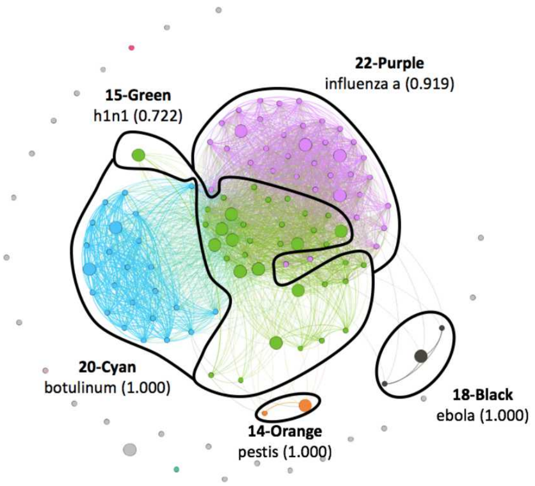

Figure 7. Weighted shared pathogen (WSP) graph over detected institutions visualized using Gephi's Force Atlas layout. Community labels denote top pathogen(s), $\mathcal{P}$, with the coverage, α, in parentheses.

ID	Color	n	β	α	H
22	Purple	37	10.8%	0.919	1.535
15	Green	36	30.6%	0.722	3.423
20	Cyan	24	8.3%	1.000	0.877
18	Black	3	33.3%	1.000	0.000
14	Orange	2	50.0%	1.000	0.000
Global Averages		**20.4**	**26.6%**	**0.928**	**1.167**

Table 7. Description of the community and graph level statistics for the WSP graph.

produced a collaboration network that has an increased ability to characterize communities based on a commonly shared pathogen, however at the expense of decreasing the focus of the research facilities in each community.

One factor that may be complicating our analysis could be the unintended consequences of employing boolean graphs in application. In particular, boolean edges will weigh all collaborations equally, and likewise all numbers of shared pathogens equally. In the next subsection, we will describe our efforts to further refine our analysis by integrating a sense of weighted similarity into our networks.

4.2.2 Weighted Graph Analysis

As with the boolean graph analysis, we first create a weighted shared pathogen (WSP) graph, where edges between institutions i and j are now weighted according to Eq. 1. The resulting graph is shown in Fig. 7. In general, edge weights reflect the number of pathogens shared between two institutions, relative to the combined pathogens present at *either* institution. Thus institutions that share a large number of pathogens relative to their total combined pathogens will have a higher edge weight, and thus will be *more* likely to be clustered when using community detection; conversely, edges between two communities that share few pathogens, but have many unshared pathogens, will have weaker edge weights, causing these institutions to be *less* likely to be clustered.

Similarities between the BSP graph depicted in Fig. 5 and the WSP graph in Fig. 7 are recognizable, although a few differences are apparent. Two dominant, tightly connected communities are visible in both graphs. However, the large anomalous cluster connecting the two dominant communities now shows much more overlap (compare 5-Cyan in the BSP graph with 15-Green in the WSP graph). Additionally, a fifth community (14-Orange) has emerged on the periphery of the WSP graph. Table 7 provides community and graph-level statistics for WSP graph. A summary comparison with the BSP graph in Table 3 reveals an increase in the average coverage coefficient ($\bar{\alpha} = 0.928$ compared to 0.831) and a decrease in the average entropy of the communities ($\bar{H} = 1.167$ compared to 2.315). These results suggest that the WSP graph increases the focus and coverage of a research community.

Next we construct a weighted collaboration (WC) graph, with edge weights computed according to Eq. 2. The graph is shown in Fig. 8 and statistically summarized in Table 8. Analogous to the WSP graph, higher edge weights reflect a larger number of shared publications, relative to the total size of their shared corpus. The WC graph demonstrates the benefit of using weighted edges for collaboration networks. The 31-Cyan community, for instance, consists of eleven institutions whose only collaboration is on a single conference proceeding. This tightly connected component is immediately isolated from its neighbors and is sequestered into its own community that

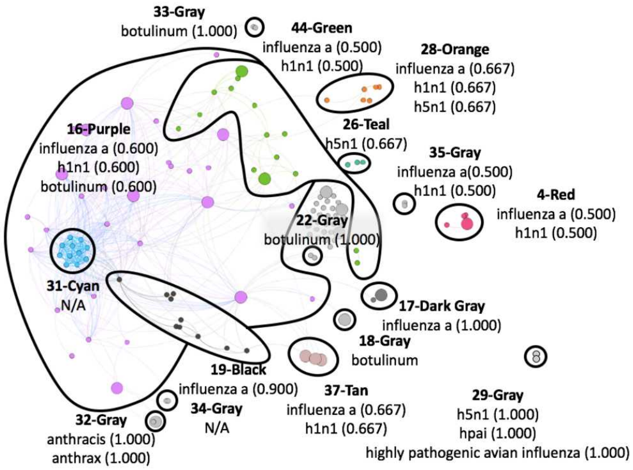

Figure 8. Weighted collaboration (WC) graph over detected institutions visualized using Gephi's Force Atlas layout. Community labels denote top pathogen(s), $\mathcal{P}$, with the coverage, α, in parentheses.

ID	Color	n	β	α	H
16	Purple	30	26.7%	.600	3.372
44	Green	16	12.5%	0.500	2.822
31	Cyan	11	0.0%	N/A	N/A
19	Black	10	0.0%	0.900	1.253
28	Orange	6	0.0%	0.667	2.308
4	Red	4	25.0%	0.500	1.918
26	Teal	3	0.0%	0.667	2.522
37	Tan	3	100.0%	0.667	1.522
32	Gray	2	50.0%	1.000	1.000
33	Gray	2	0.0%	1.000	0.000
35	Gray	2	0.0%	0.500	1.000
18	Gray	2	50.0%	1.000	0.000
17	Dark Gray	2	50.0%	1.000	1.922
22	Gray	2	0.0%	1.000	0.000
34	Gray	2	0.0%	N/A	N/A
29	Gray	2	0.0%	1.000	2.250
Global Averages		**6.2**	**19.6%**	**0.786**	**1.564**

Table 8. Description of the community and graph level statistics for the WC graph.

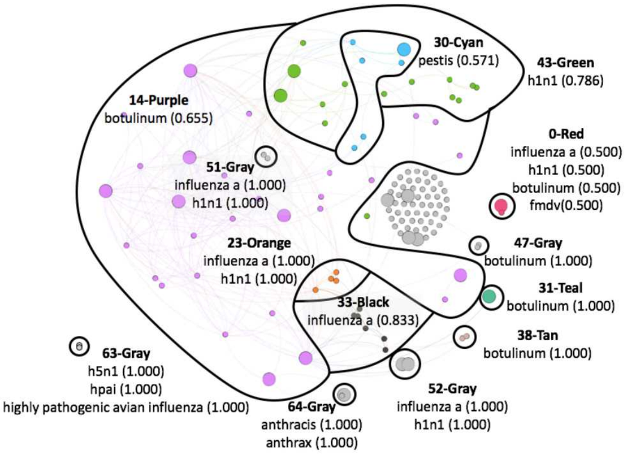

Figure 9. Composition of weighted collaboration (WC) applied to weighted shared pathogens (WSP) visualized using Gephi's Force Atlas layout. Community labels denote top pathogen(s), $\mathcal{P}$, with the coverage, α, in parentheses.

ID	Color	n	β	α	H
14	Purple	29	27.6%	0.655	3.404
43	Green	14	14.3%	0.786	2.576
30	Cyan	7	14.3%	0.571	1.959
33	Black	6	0.0%	0.833	0.954
23	Orange	4	0.0%	1.000	1.392
0	Red	2	50.0%	0.500	1.918
31	Teal	2	50.0%	1.000	0.000
38	Pale Pink	2	0.0%	1.000	0.000
47	Gray	2	0.0%	1.000	0.000
64	Gray	2	50.0%	1.000	1.000
51	Gray	2	0.0%	1.000	1.000
52	Gray	2	100.0%	1.000	1.000
63	Gray	2	0.0%	1.000	2.250
Global Averages		**5.8**	**23.6%**	**0.873**	**1.343**

Table 9. Description of the community and graph level statistics for the WC-WSP graph.

could then be largely ignored by an analyst. Table 3 provides a summary comparison of the WC with the BC graph, where we can see that the WC graph provides several preferable community characteristics, including: a higher average percentage of previously identified institutions per community ($\bar{\beta} = 19.6\%$ compared to 15.1%), a higher coverage coefficient ($\alpha = 0.786$ compared to 0.689), and a lower entropy/higher focus ($H = 1.564$ compared to 1.972).

Now that we have produced and analyzed both the WC and WSP graphs, we show the results of fusing the two using Eq. 3, resulting in a WC-WSP graph with a more fine-grained characterization of research communities. The graph generated by this fusion can be seen in figure 9. This graph largely maintains the structure from the original weighted collaboration graph with a few key differences. Namely, the tightly clustered and potentially problematic 31-Cyan cluster has been removed as it did not share pathogens with its peer institutions. Additionally, the graph has created more, smaller, communities that can be found scattered throughout the graph. The combination of these two have a noticeable impact on the computed graph level statistics in Table 9.

In particular, a comparison of the WC-WSP graph with the WC graph reveals that the composite WC-WSP graph provides a higher average coverage coefficient ($\overline{\alpha} = 0.873$ compared to 0.786), as well as a smaller average community size ($\overline{n} = 5.8$ compared to 6.2) and a lower average entropy ($\overline{H} = 1.343$ compared to 1.564).This result implies that the fusion of the WSP graph with the WC graph has boosted the average coverage of the communities' most prevalent pathogen, and increased the focus of each institution when compared to the WC graph. Therefore, the application of graph fusion can provide a finer grained analysis over a collaboration network without perturbing the underlying network structure.

4.2.3 Composite Graph Analysis

Graph	$\overline{n}$	$\overline{\beta}$	$\overline{\alpha}$	$\overline{H}$
Bool. Composite	8.4	21.4%	0.801	2.236
Weighted Composite	5.8	23.6%	0.873	1.343
% Change	-30.1%	+10.3%	+8.9%	-39.9%

Table 10. Comparison of composite boolean and weighted graphs.

Finally, we compare the composite boolean and weighted graphs, summarized in Table 10. It is apparent that the weighted composite graph decreases both the average community size $\bar{n}$ and the average intra-community entropy $\bar{H}$, leading to smaller, more focused institutions than the boolean collaboration graph. Furthermore, it provides increases to both the relative population of previously identified institutions within a community and the coverage coefficient $\bar{\alpha}$, meaning a better characterization of the communities with a higher presence of known institutions.

5. CONCLUSIONS AND FUTURE WORK

In conclusion, we applied text, graph, and social analytics on bibliometric data to discover and characterize research communities in South Korea. This work is framed in the context of a hypothetical pathogen outbreak during the 2018 Winter Olympics in South Korea, during which decision makers need to identify local facilities that are properly equipped to aid and support the crisis response. DTRA and Sandia National Laboratories have partnered to develop analytics for semi-automated, analyst-driven characterization of research facilities, hospitals, and diagnostic laboratories based on patterns and relationships identified within scientific communities. More specifically, we developed a graph fusion pipeline for generating, combining, and analyzing different similarity graphs based on open-source PubMed data. Our results show that an approach that fuses a collaboration graph with a shared research focus allows for a finer grained analysis of communities in a collaboration network, while still preserving the underlying network structure. Furthermore, more sophisticated similarity measures, such as weighted edges based on Jaccard similarities, reduce the noise introduced by simple boolean similarity measures. Future work will include exploring other combinations of fundamental boolean operators, such as shared pathogens but NOT shared collaboration, and other user-defined similarity measures, such as similarities of facility descriptions based on topical modeling algorithms such as latent dirichlet allocation (LDA).[12,13]

While the use of PubMed and ability to perform facility disambiguation to identify all of the possible biological facilities is valuable by itself, the graph approach has the potential to provide important context for identifying which facilities have the most impact for responders and planners. We plan to continue to evaluate this new method and use this information to summarize facility holdings around regions of concern.

ACKNOWLEDGMENTS

The mission of the Defense Threat Reduction Agency is to safeguard the United States and its allies from global WMD threats by integrating, synchronizing, and providing expertise, technologies, and capabilities. The Defense Threat Reduction Agency reports to United States Special Operations Command (USSOCOM) and Office of the Under Secretary of Defense for Acquisition and Sustainment (A&S). This work was conducted under contract between the Defense Threat Reduction Agency and Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND-2018XXXX.

REFERENCES

[1] Newman, M. E. J., "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences of the United States of America* **101**, 5200–5205 (2004).

[2] Bender, M. E., Edwards, S., von Philipsborn, P., Steinbeis, F., Keil, T., and Tinnemann, P., "Using co-authorship networks to map and analyse global neglected tropical disease research with an affiliation to Germany," *PLoS Neglected Tropical Diseases* **9**(12), e0004182 (2015).

[3] de Paula Fonseca e Fonseca, B., Sampaio, R. B., de Arajo Fonseca, M. V., and Zicker, F., "Co-authorship network analysis in health research: method and potential use," *Health Research Policy and Systems* **9**, 14–34 (2016).

[4] Kuhn, T., Perc, M., and Helbing, D., "Inheritance patterns in citation networks reveal scientific memes," *Physical Review X* **4**, 041036 (2014).

[5] Fortunato, S., "Community detection in graphs," *Physics Reports* **486**, 75–174 (2010).

[6] Liben-Nowell, D. and Kleinberg, J., "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology* **58**(7), 1019–1031 (2007).

[7] "Pubmed advanced search." https://www.ncbi.nlm.nih.gov/pubmed/advanced. Accessed: 22 February, 2018.

[8] Bastian, M., Heymann, S., and Jacomy, M., "Gephi: An open source software for exploring and manipulating networks," (2009).

[9] Bauer, T. and Garcia, D., "Accessibility, adaptability, and extendability: Dealing with the small data problem," *Advances in Intelligent Systems and Computing* **497** (2016).

[10] Manning, C. D., Schütze, H., and Raghavan, P., [*Introduction to Information Retreival*], Cambridge University Press, Cambridge, England (2009).

[11] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E., "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment* , P1008 (2008).

[12] Pritchard, J. K., Stephens, M., and Donnelly, P., "Inference of population structure using multilocus genotype data," *Genetics* **155**, 945–959 (2000).

[13] Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent dirichlet allocation," *Journal of Machine Learning Research* **3**, 993–1022 (2003).