Towards a More Representative Definition of Cyber Security

Daniel Schatz
University of East London, u0829943@uel.ac.uk

Rabih Bashroush
University of East London, r.bashroush@uel.ac.uk

Julie Wall
University of East London, j.wall@uel.ac.uk

Follow this and additional works at: https://commons.erau.edu/jdfs
Part of the Computer Law Commons, and the Information Security Commons

Recommended Citation
DOI: https://doi.org/10.15394/jdfs.2017.1476
Available at: https://commons.erau.edu/jdfs/vol12/iss2/8

This Article is brought to you for free and open access by the Journals at Scholarly Commons. It has been accepted for inclusion in Journal of Digital Forensics, Security and Law by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.
TOWARDS A MORE REPRESENTATIVE DEFINITION OF CYBER SECURITY

Daniel Schatz
University of East London
Architecture Computing and Engineering
u0829943@uel.ac.uk

Rabih Bashroush
University of East London
Architecture Computing and Engineering
r.bashroush@uel.ac.uk

Julie Wall
University of East London
Architecture Computing and Engineering
j.wall@uel.ac.uk

ABSTRACT

In recent years, ‘Cyber Security’ has emerged as a widely-used term with increased adoption by practitioners and politicians alike. However, as with many fashionable jargon, there seems to be very little understanding of what the term really entails. Although this is may not be an issue when the term is used in an informal context, it can potentially cause considerable problems in context of organizational strategy, business objectives, or international agreements. In this work, we study the existing literature to identify the main definitions provided for the term ‘Cyber Security’ by authoritative sources. We then conduct various lexical and semantic analysis techniques in an attempt to better understand the scope and context of these definitions, along with their relevance. Finally, based on the analysis conducted, we propose a new improved definition that we then demonstrate to be a more representative definition using the same lexical and semantic analysis techniques.

Keywords: cyber security; information security; national cyber policy; systematic review

1. INTRODUCTION

The terminology used to discuss security aspects of digital devices and information changed considerably in recent years. At the beginning of the century, terms regularly used in this context would be “Computer Security,” “IT Security,” or “Information Security.” Whilst these terms have nuanced differences understood by professionals working in this space, they were tangible enough to be meaningful to the wider populace. General conversations could be had and plans could be made based on a common understanding of what these terms imply. However, towards the end of the first decade, new terminology
started to become increasingly popular with(185,222),(868,928)
the use of the term “Cyber Security.” It had
been in use during previous years but its
popularity gained considerably when U.S.
President Barack Obama in 2009 proclaimed “I
call upon the people of the United States to
recognize the importance of cybersecurity and
to observe this month with appropriate
activities, events, and trainings to enhance our
national security and resilience” (The White
House, 2009). The immediate impact of this
press release on terminology can be illustrated
with the help of Google’s search trends which
shows a noticeable spike in this period (Figure
1). The trend lines on the chart show total
searches for a term relative to the total
number of searches done on Google over time.
We can see a steady decline of the search
terms “Computer Security” and “Information
Security” with variants of “Cyber Security”
converging and surpassing them. This finding
is only indicative but as seen in previous
research (Choi & Varian, 2012), search engine
based information is useful and of value to
identify trends.

Figure 1. Google search trends for security 2004 – 2015

This development in use of terminology is
causing some issues as the term “Cyber
Security” lacks the defining clarity of, for
example, “Computer Security.” This can lead
to confusion and misunderstanding if parties
have different assumptions of what the term
represents. Quoting Sowell (2014) on the
importance of clarity;

“What may seem like small steps in
logic, after the fact, can be a long, time-
consuming process of trial and error
groping, while creating and refining
concepts and definitions to express ideas
in clear and unmistakable terms which
allow substantive issues to be debated in
terms that opposing parties can agree on,
so that they can at least disagree on
substance, rather than be frustrated by
semantics.”

Whilst it is unlikely to be a problem in
private conversations between interested
citizens it becomes, at the very least, a
nuisance at an organizational level and is a
widely recognized issue among professionals in
the field. These problems amplify if ambiguity
continues in courts of justice, national cyber
security strategies or international treaties. Eig (2011) discusses such issues in context of statutory interpretation in greater detail.

An additional, although less impacting, issue is the inconsistent use of syntax for cyber security. Across the literature both versions, cybersecurity and cyber security, are used. Observing the search trends as illustrated we see that both terms are upward trending; however, the disjoined version (cyber security) shows prevalence in absolute numbers which is the spelling that shall be used going forward unless referring to primary source material.

Recognizing the lack of a consistent meaning of the term cyber security as a considerable issue (Baylon, 2014; Congressional Research Service, 2014; Creasey, 2013; Internet Society, 2012), we are first reviewing the current definition landscape across professional, academic and governmental literature with a goal to identify the most prevalent definitions, key components in definitions, and take a view on contentious points between proposed definitions where such exist. As second and third steps, we will identify the best match definition and contribute a new improved one.

The remainder of the paper is structured as follows; in the next section, we will take a look at existing research in this field and discuss challenges of the current definition landscape. Section 3 describes the approach and methodology followed for our systematic literature review on the topic. We continue to analyze the definition set from a semantic perspective in sections 4 and 5 with a proposal for an improved definition outlined in section 6. In sections 7 and 8, we review limitations of our approach and provide conclusive thoughts.

2. LITERATURE REVIEW

The lack of a uniformly accepted definition of cyber security as described in the previous section has been recognized across professional (Barzilay, 2013; Stubley, 2013; Walls, Perkins, & Weiss, 2013), governmental (Falessi, Gavrila, Klejnstrup Ritter, & Moulinos, 2012; Government of Montenegro, 2013; Wamala, 2011) and academic (Baylon, 2014; Giles & Hagestad, 2013) work.

2.1 Industry definitions

Walls et al. (2013) approach the topic from the perspective of a professional services provider (Gartner Inc.) and is thus focusing on tangible guidance for strategic decision makers. A key challenge highlighted is the ambiguity introduced by the thoughtless use of the term ‘cyber security,’ where nuanced definitions (Information Security or IT Security) are more appropriate and descriptive. They suggest that the term cyber security is only used in context of security practices related to the combination of offensive and defensive actions involving or relying upon information technology and/or operational technology environments and systems. The authors state that it marks a superset of security practices such as information security, IT security and other related practices. Stubley (2013) takes a different view to this and simplifies cyber security to information security based on a short analysis of the ‘cyber’ component which he defines to describe the use of information technology and computers. Barzilay (2013) again takes a different view and argues that cyber security must be defined through cyber risk which leads to his conclusion that cyber security is a sub discipline of information security which is in contrast to Walls et al. (2013). In official guidance, ISACA (2014) takes yet another position stating that cyber security is emerging within the fields of information security and traditional security. Enterprises should distinguish between standard (lower-level) information security and cyber security; the difference is in the scope, motive, opportunity and method of the attack.
2.2 **Government and nation state definitions**

In their analysis of national cyber security strategies of European Union member states, Falessi et al. (2012) provide terminology guidance in the annex explaining that there is no universally accepted nor straightforward definition of ‘cyber security.’ They write that some people regard cyber security as overlapping with information security but no definitive conclusion is provided. This view is shared by Wamala (2011) claiming that cyber security is a branch of information security. The paper highlights the risk of uncertain terminology and aims to provide clarification on the relative positions of cyber security and information security. It draws a link between cyber security and the global characteristic of the internet, as such distinguishing it from information security which, according to the author, rarely traverses jurisdictions. Wamala goes further in this definition claiming that cyber security focuses more on integrity and availability whereas information security is mainly concerned with confidentiality. He concludes that cyber security is information security with jurisdictional uncertainty and attribution issues. The Government of Montenegro (2013) agrees with the notion of a lack of clear definitions in this area and dedicates a full section in its cyber security strategy to this topic. Whilst the paper states that it presents definitions which are compliant with the basic meanings as understood in EU countries, it unfortunately does not actually provide a conclusion on the term cyber security but rather quotes various definitions from other sources. Baylon (2014) discusses the topic from a multinational cooperation perspective highlighting that the lack of or insufficiently agreed on definition of key terminology in the cyber and space security domains poses a major challenge to international treaties and arms control agreements. In particular, the considerably different interpretation of cyber security between western countries and both Russia and China causes complications in this context. Baylon states that the term ‘cyber security’ as such does not exist in Russian legislation or official doctrines. Instead, the concept of information security is prevalent. However, in this context, “information” represents a meaning extending outside the digital space which widens conversations into the information space in general. The author categorizes this into the Eastern approach, looking at cyber security emphasizing ‘social cohesion,’ and the Western approach, perceiving cyber security through a ‘national security prism.’ Godwin III, Kulpin, Rauscher, and Yaschenko (2014) concur with this challenge and provide bi-nationally (USA/Russia) agreed terminology for key phrases pertaining to the cyber space. Amongst these, the term ‘cyber security’ is defined as well; notably with a considerably different interpretation than found in official national cyber security strategies of most western countries. Giles and Hagestad (2013) extend on this by contrasting key terms and principles in this space as understood in each of their focus countries (USA, China, Russia). They find that there is a notably different understanding and approach between these countries. They conclude that in absence of a mutually agreed terminology, any potential for finding a real commonality of views on the nature and governance of cyberspace remains distant.

2.3 **Academic definitions**

Academic research has not been oblivious to the obvious challenges in this developing problem space, of course. Luiijf, Besseling, and de Graaf (2013) conducted an exhaustive study of national cyber security strategies (NCSS) for 19 countries which also discusses differences in terminology in some detail. They find that only eight nations define the term ‘cyber
security’ in their NCSS, whereas six nations do not provide any such definition. The authors note that for the ten NCSS which have the term cyber security defined either through implication, description or definition, the understanding of what it means varies greatly. This view is shared by Craigen, Diakun-Thibault, and Purse (2014) who looked at a wider range of sources attempting to define the term. They find that the term is used broadly and its definitions are highly variable, context-bound, often subjective, and, at times, uninformative. Based on a shortlist of nine definitions, the authors work towards a unified definition identifying five dominant themes of cyber security. Through consensus in a multidisciplinary group, the authors arrive at an additional definition for cyber security. Many of the definitions mentioned in this section will be the focus of the remainder of this paper.

3. SYSTEMATIC REVIEW APPROACH

To better understand the variety of relevant definitions in use for cyber security, we followed a semi systematic literature review approach (Mäntylä, Adams, Khomh, Engström, & Petersen, 2014) as further described below. Following the collection of definitions, we applied text analysis methods on the resulting dataset focusing on semantic similarity analysis with the goal to identify harmonizing definitions. This approach resulted in a ranking of definition similarity across the dataset from a text analytics perspective; i.e. we established which definitions represent most accurately the definition of ‘cyber security’ across the whole dataset. Based on further analysis of the highest scoring definitions, we created a new definition comprising the key terms identified. The new definition was then compared against the original dataset to verify its best match status across the whole dataset.

3.1 Research question

We started by defining our research questions at a high level.

Table 1

<table>
<thead>
<tr>
<th>Research questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ 1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>RQ 2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>RQ 3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>RQ 4</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

In order to answer our research questions, we first needed to identify the relevant definitions. For this, we applied a set of inclusion and exclusion criteria to our literature search as follows.

Inclusion criteria:
• **IC1:** Sources with clear intention of providing an explicit definition of cyber security
• **IC2:** Sources available in English or translation readily available

**Exclusion criteria:**
• **EC1:** Sources which provide no clear or only implicit definitions of cyber security
• **EC2:** Sources that lack rigor (peer review) or authority (governmental or professional bodies) for defining cyber security

These criteria have been applied throughout the search process, in particular EC2 (Cornell University, 2016). In the first instance, Thomson Reuters’ Web of Science database was used to identify relevant academic sources. The search scope covered a time span of ‘All years’ with a search construct of TOPIC: ("cyber security" OR Cybersecurity) NEAR definition). This produced limited results of merely 13 hits of which only one source met our criteria. Modifying the search query to include variations of the term ‘definition’ (meaning, interpretation) did not produce any additional relevant results. Our search efforts in other databases such as Science Direct (25 results) were met with similar challenges. To capture a wider range of sources we extended our search efforts to the general purpose search engine Google.com, limiting search parameters as follows ([cybersecurity AROUND(3) definition] OR ["cyber security" AROUND(3) definition]). Manual review of the top search results returned by Google was then conducted to capture the most relevant sources. Based on the sources identified, further backward and forward reference crawling was conducted (using Google Scholar) to capture additional material relevant to our research question. In addition, source lists provided by ENISAi and NATO ii were reviewed manually. Our literature review identified 28 sources which met our inclusion and exclusion criteria as shown in Appendix A in no particular order. Out of the 28 identified sources, one definition source is considered academic, five industries contributed and 22 definitions were by government or government aligned bodies. As expected there is considerable overlap in term use between definitions of which some include parts of definitions stated by another source (e.g. #3 and #18). The definition text was extracted from the source material in the context it was written.

### 4. BASIC DEFINITION ANALYSIS

To get a better understanding of the dataset, an initial exploratory text analysis (Hearst, 1999) was conducted to try and discover information inherent to the definitions. We started by applying basic information extraction procedures (Weiss, Indurkhya, Zhang, & Damerau, 2004) utilizing the Text Mining framework *tm_map* (Meyer, Hornik, & Feinerer, 2008) for the software environment for statistical computing “R.”

Before the definition data were loaded into R, minimal manual normalization was applied to standardize character encoding and remove unnecessary line breaks. The definition corpus was then prepared with common pre-processing functions as provided by *tm_map* to convert content to lower case, strip whitespaces, remove punctuation and remove stop words (English). In addition, stemming was applied (Porter, 1997) to reduce the number of distinct word types in the text corpus and to increase the frequency of occurrence of some individual types (Weiss et al., 2004).

With the corpus prepared, we created a simple document-term matrix (Salton, 1963) that allowed us to gain basic insights on how our sources define ‘cyber security.’ As
illustrated in Figure 2, the root form of ‘security,’ ‘cyber security,’ ‘cyber,’ and ‘cyberspace’ is prevalent in the corpus which was expected. However, we also get an indication of related words fundamental to the definition pool.

The basic term frequency analysis provided an intuition on term priority across the definition dataset and an indication of the importance (by way of word count) of certain words in the set, most notably ‘risk,’ ‘protect,’ ‘use,’ ‘process,’ and ‘system.’ With this information, we conducted an analysis on the definition sets.

Lexical Overlap analysis, the process of identifying how many words texts have in common, is one of the simplest methods to assess the similarity between texts (Rus, 2014). We used this to conduct a basic lexical token review on our definition set with just the most frequent ten unigrams in their stemmed form as shown in Figure 3. The heat map shows the ten most frequent terms across all 28 definitions in the dataset with an individual and total term count per document. Based on this simple analysis, we glean that some definitions incorporate a wider spectrum of key terms than others (e.g. #3 or #11), and may provide a better representation of what the entire definition pool defines as ‘cyber security.’ We also note that such simple analysis is skewed by term repetition e.g. as

![Figure 2. Word frequency in definition corpus for top terms](image-url)
observed for definition #12 where ‘cybersecur’ and ‘cyberspac’ are used frequently.

Continuing with our basic analysis, we created a correlation matrix for a sparse document-term matrix (sparsity at 0.85) to gain additional information on strongly correlated terms across the definition set. Figure 4 confirms our assumption that frequent terms such as ‘cybersecur,’ ‘cyberspac,’ or ‘secur’ are not highly correlated with other terms in this context; however, the correlation matrix shows that we have some correlated terms that are worth further exploration.

We found high correlation between the terms of the CIA triad (confidentiality, integrity, availability) which makes intuitive sense as they tend to be used together when writing about topics like cyber security. We further see noteworthy correlation of ‘inform’ and ‘integ’ along with the CIA triad which we’ll see confirmed in a later section of this paper. We also note correlation between ‘secur,’ ‘asset,’ and ‘environ’ which points towards a general agreement that those terms standing together are important to a harmonized definition of cyber security. This basic approach shows further interesting positive and negative correlations (e.g. ‘include’ and ‘infrastructur’ or ‘realibl’ and ‘protect’) that helps to better understand the definition space. But we still lack a way to identify what the most representative definition of ‘cyber security’ is. Following the maxim “a person without data is just another person with an opinion”iii, we designed an approach that would allow us to identify the most representative definition within our pool. The assumption is that our dataset includes the majority of authoritative definitions for ‘cyber security’ and as such covers all relevant aspects of the concept as proposed by the sources (Ryan & Bernard, 2003). This means we can identify the definition encompassing the majority of relevant components through lexical and semantic similarity analysis; that is the definition which is most alike to every other definition in the dataset. We made use of a wide range of advanced similarity measures as described in the next section to achieve this.
5. DEFINITION
SIMILARITY
ANALYSIS

Semantic similarity is a well-established area of research with a wide range of practical applications (Androutsopoulos & Malakasiotis, 2010; Couto, Silva, & Coutinho, 2007; Graesser, Olney, Haynes, & Chipman, 2005; Yuhua, Bandar, & McLean, 2003). For the purpose of this research, we investigated current work on short text and sentence based similarity measures. We initially planned to use the best method for sentence based similarity measures as proposed by subject matter experts on this topic, but found that this is a developing area with various competing methods proposed. Instead of picking one specific method to calculate similarity, we decided to calculate similarity with a variety of methods to balance advantages and disadvantages of individual methods. The result is an average similarity score as described in this section. We found the SEMILAR toolkit (Rus, Lintean, Banjade, Niraula, & Stefanescu, 2013) to be ideal for this as it vastly simplified the task of calculating similarity using multiple algorithms.

Figure 4. Sparse DTM correlation matrix
and options. The authors describe the toolkit as “a one-stop-shop for investigating, annotating, and authoring methods for the semantic similarity of texts of any level of granularity.” We used the toolkit to conduct both the pre-processing phase and the similarity computing phase for our dataset.

As with our basic analysis we conducted common pre-processing tasks on our dataset but with some notable differences. Again, the first step is tokenization of the text to obtain the ordered set of lexical tokens. Based on our configuration, SEMILAR calculates the initial lexical form of the token, lemma form of the word, part-of-speech (POS), weighted specificity of the word, semantic representation (WordNet (Miller, 1995) or LSA (Martin & Berry, 2007)) and a list of syntactic dependencies with the other words in the same sentence (Lintean, 2011). To capture as much context as possible, we chose Stanford CoreNLP (De Marneffe, MacCartney, & Manning, 2006) as the configuration option for tokenization, POS tagging, lemmatizer as well as syntactic parsing. Figure 5 provides a visual example of how this task processed one of the definitions in the set. The effect of lemmatization (as compared to stemming) and part of speech tagging is apparent. The function identified sentence tokens and categorized them accurately for further processing. In the sample chosen we see that the tagger associated words with their respective part of speech (“The”/Determiner, “ability”/Noun singular, “protect”/verb base, “or”/coordinating conjunction, etc.).

![Figure 5. Sample pre-processing with StanfordNLP](image)

With the definition set prepared this way, we calculated similarity between all definitions using nine methods resulting in 7056 similarity scores. The selection of the nine methods and
their configuration options we used to calculate the similarity scores was based on recommendations and insights in relevant literature (Corley & Mihalcea, 2005; Gomaa & Fahmy, 2013; Lee, 2011; Lintean, 2011; Nakov, Popova, & Mateev, 2001; Rus, 2014; Rus & Lintean, 2012; Yuhua, McLean, Bandar, O'Shea, & Crockett, 2006). The methods chosen are categorized and listed within the SEMILAR toolkit as lexical methods (five), Corley and Mihalcea (2005) (three) as well as plain LSA vector similarity.

For lexical similarity methods, we did not remove stop words, non-function words or punctuation adopting findings by Lintean (2011, p. 60) and Yuhua et al. (2006) showing the importance of these tokens for similarity calculations due to their structural information value. We did, however, convert all tokens to lower case. For lexical matching we selected optimal pairing (Rus et al., 2012) without enforcing part of speech matching. Token weights are based on entropy (Martin & Berry, 2007) rather than inverted document frequency (IDF) (Sparck Jones, 1972) following guidance by Lintean (2011), finding that entropy-based weighting leads to better results than IDF-based weighting. With this configuration set as baseline, we selected five token similarity metric methods; Jiang and Conrath (Jiang & Conrath, 1997), Leacock and Chodorow (Leacock, Miller, & Chodorow, 1998), Lin (Lin, 1998), LSA as well as Wu and Palmer (Wu & Palmer, 1994). For similarity calculations based on the ‘Class of Method’ (Corley & Mihalcea), we again chose Jiang and Conrath, Leacock and Chodorow and Lin, each with bidirectional scoring type and Touchstone Applied Science Associates corpus (TASA) derived IDF as model. Finally, for plain LSA similarity scoring, we selected a frequency-based local weight, as well as an entropy-based global weight. Further rationale and detail on each of the methods is beyond the scope of this paper and can be found in the references listed in this section.

With the scores calculated, we transformed them into a matrix format as pictured in Figure 6. This allowed us to calculate final averages for each definition.
As the similarity score is asymmetric (Lintean, 2011) for some of the methods (Text A → Text B ≠ Text B → Text A) illustrated by the different values in the upper triangle compared to the lower triangle, we calculated all row (r) and column (c) means. The combined mean provided the overall similarity score per definition. Figure 7 shows how each definition measures up in similarity against all other definitions.

![Figure 7. Average similarity score per definition in dataset](image_url)

With this information, we produced a ranked order of the most representative definitions in the dataset. Table 2 shows an excerpt of the final list with the 5 most representative definitions of the definition pool ranked by similarity score across all nine methods and definitions.

Per our semantic similarity approach, the most representative definition in our dataset of authoritative definitions is part of the South Africa NCSS;

“Cybersecurity is the collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organization and assets.”

It is worth noting that definition #18 is part of an more exhaustive definition text by the International Telecommunication Union (2008), but comes out top due to its relative conciseness. On the flip side, we note that brevity is not key to a representative definition (in context of the pool of our authoritative definitions) as illustrated by the trailing
definitions #16, #17 and #28. These are very concise but do not have sufficient descriptive depth to capture the meaning of cybersecurity; both objectively, as shown in the comparison, as well as subjectively (although this leaves plenty of room for argument).

Table 2.
Top 5 most representative definitions

<table>
<thead>
<tr>
<th>ID</th>
<th>source</th>
<th>title</th>
<th>SimScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Republic of South Africa</td>
<td>Cybersecurity Policy of South Africa</td>
<td>0.434</td>
</tr>
<tr>
<td>11</td>
<td>French Network and Information Security Agency</td>
<td>Information systems defence and security France’s strategy</td>
<td>0.426</td>
</tr>
<tr>
<td>21</td>
<td>Spanish Cyber Security Institute</td>
<td>National Cyber Security, a commitment for everybody</td>
<td>0.409</td>
</tr>
<tr>
<td>3</td>
<td>International Telecommunication Union</td>
<td>Series X: Data Networks, Open System Communications and Security</td>
<td>0.407</td>
</tr>
<tr>
<td>15</td>
<td>New Zealand Government</td>
<td>New Zealand’s Cyber Security Strategy</td>
<td>0.405</td>
</tr>
</tbody>
</table>

It is important to point out we didn’t identify this to be the most relevant definition through expert opinion but through unbiased similarity analysis based on an authoritative set of definitions. Definition #18 best captures the essence of all authoritative definitions combined.

6. TOWARDS AN IMPROVED DEFINITION

After identifying the most representative definitions for ‘cyber security’ as described in the previous section, the next step was to try and construct an improved definition. The new definition would then be measured under the same conditions to compare similarity scores. Using KH Coder (Higuchi, 2015) we investigated the previously mentioned top five definitions (18, 11, 21, 3, 15) with the assumption that they contain the most relevant attributes in the overall definition pool. To establish the key underlying concepts needed to craft an improved definition, we used co-occurrence network analysis (Rice & Danowski, 1993). In textual analysis, co-occurrence networks show words with similar appearance patterns and as such with high degrees of co-occurrence. The approach is based on the idea that a word’s meaning is related to the concepts to which it is connected. It also has the benefit that no coder bias is introduced other than to determine which words are examined (Ryan & Bernard, 2003). However, applying the function on our definition set, even though already limited to five paragraphs and with minimum spanning tree applied, proved to produce a very crowded output difficult to navigate. By filtering for term frequency (TF ≥ 2) when producing the co-occurrence network graph, we were able to reduce the information presented to a (human) manageable level while preserving important context.

Figure 8 shows the minimum spanning tree (MST) network graph model with 32 nodes and 25 edges extracted. The graph presents an at a glance a view of the underlying concepts inherent to the words used in the definition set. In addition to the minimum spanning tree, we have added community detection to further emphasize connected components. The node size illustrates the term frequency, and detected communities are highlighted in different colors. Based on the dataset, we found that the ‘random walk’ or ‘walktrap’
algorithm (Pons & Latapy, 2005) provided the subjectively best community detection approach. Combined with MST, it aids in understanding not only the key concepts but also how words group into communities and which communities are closer to each other (signified by dotted lines).

With the key components extracted, we were in a position to create our own proposal for an improved definition. Through several iterations of manual sentence construction using words and communities, we arrived at a definition that captures key components and respects community adhesion:

"The approach and actions associated with security risk management processes followed by organizations and states to protect confidentiality, integrity and availability of data and assets used in cyber space. The concept includes guidelines, policies and collections of safeguards, technologies, tools and training to provide the best protection for the state of the cyber environment and its users."

To verify that this definition is not only representative from a human reader perspective but also in terms of semantic
similarity, we repeated our semantic analysis benchmarking work (Section 5), this time including our new definition (#29).

As expected, the overall results are nearly the same as previously since the methodology and configuration of the benchmark has not changed. Individual scores have changed slightly due to the new addition (29) to the corpus. The overall ranking did not change except for our proposed definition being included at the top as seen in Table 3.

Table 3
Top results for improved definition dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>source</th>
<th>title</th>
<th>SimScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>New Definition</td>
<td>n/a</td>
<td>0.465</td>
</tr>
<tr>
<td>18</td>
<td>Republic of South Africa</td>
<td>Cybersecurity Policy of South Africa</td>
<td>0.440</td>
</tr>
<tr>
<td>11</td>
<td>French Network and Information Security Agency</td>
<td>Information systems defence and security France’s strategy</td>
<td>0.434</td>
</tr>
<tr>
<td>21</td>
<td>Spanish Cyber Security Institute</td>
<td>National Cyber Security, a commitment for everybody</td>
<td>0.416</td>
</tr>
<tr>
<td>3</td>
<td>International Telecommunication Union</td>
<td>Series X: Data networks, open system communications and security</td>
<td>0.412</td>
</tr>
<tr>
<td>15</td>
<td>New Zealand Government</td>
<td>New Zealand’s Cyber Security Strategy</td>
<td>0.409</td>
</tr>
</tbody>
</table>

7. STUDY LIMITATIONS AND CHALLENGES TO VALIDITY

In the previous section, we proposed a new definition for ‘cyber security’ which tops the ranking of most relevant definitions among authoritative sources. However, as with many similar research exercises, there is no claim to completeness or infallibility of our work. Our study is affected by limitations inherent to literature reviews as described by Kitchenham and Charters (2007) which includes limitations on search comprehensiveness and material selection. To mitigate this weakness, forward and backward reference checking was conducted on key publications to discover potentially relevant sources. Regardless, it is possible that our efforts missed sources which we would have otherwise considered authoritative and relevant (although the number of definitions covered in this study should ensure relevance irrespective). Another inherent limitation to literature reviews is the language barrier, as this work only covered definitions provided in English.

Although the study has achieved its objective of creating a representative definition for ‘cyber security,’ our approach for creating the definition is limited by manual sentence generation constrains. It is possible that an automated approach, iterating all possible combinations of our nodes and communities leveraging natural language generation (Sauper & Barzilay, 2009), would have produced another, perhaps more relevant definition. This was beyond the scope of this paper but will be considered for future work.

Lastly, considering the pace at which social communities create, adopt and modify their understanding of developing areas such as ‘cyber’ and ‘cyberspace,’ our definition is representative at the time of the research. It is expected that this definition will become less
fitting or relevant as social, political, and technological developments in this space progress. Nonetheless, our proposed model for evaluating definitions will prove useful and remain relevant in the future.

8. CONCLUSION AND RESEARCH QUESTIONS

For this research, we set out to analyze the landscape of authoritative sources defining the term ‘cyber security.’ As part of this work, we conducted a semi-systematic literature review identifying relevant sources. Through our efforts as outlined in section 3, we found 28 authoritative sources fulfilling our inclusion criteria and were included for further analysis in context of our research questions. This not only provided the fundament to answer our research questions, but also contributed the most exhaustive set of authoritative sources for further research in this field. We found the majority of definition sources to be related to governmental institutions with several additional relevant sources from industry and the academic sector (RQ1). Our review of primary sources unveiled a clear lack of congruence across the sources as to the meaning and scope of the term. Even contradictory claims in regard to scope were identified for several primary studies (RQ2). To better understand the differences in the definition set (RQ2) and to identify the most relevant definition (RQ3), we applied basic (section 4) and advanced (section 5) semantic similarity analysis methods to the data set. To our knowledge, this is the first endeavor to make use of this novel and non-biased approach to identify the most representative definition in a set of definitions (for ‘cyber security’). We were able to show that the definition contributed by the Republic of South Africa (2010) achieved the highest similarity score and as such was the most representative definition of ‘cyber security’ under the conditions of this work. To answer our final research question (RQ4), we conducted further analysis on the data set making use of co-occurrence, semantic networks, and community detection methods. By isolating key components and communities in the definition set, we produced an improved definition for ‘cyber security’ (section 6). Our new definition was shown to be the new most representative definition following the same methodology discussed in the paper. While we recognize the potential for further improvement of this approach (section 7), we believe that the methodology and the improved definition is a noteworthy contribution to the field.
REFERENCES


doi:http://dx.doi.org/10.1016/j.eswa.2010.04.043


doi:10.1504/IJCIS.2013.051608


### APPENDIX A

**Table 4**

*Complete overview of definition sources*

<table>
<thead>
<tr>
<th>ID</th>
<th>Source</th>
<th>Title</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Committee on National Security Systems</td>
<td>National Information Assurance (IA) Glossary</td>
<td>2009</td>
</tr>
<tr>
<td>2</td>
<td>National Initiative for Cybersecurity Careers and Studies,</td>
<td>Explore Terms: A Glossary of Common Cybersecurity Terminology</td>
<td>n/a</td>
</tr>
<tr>
<td>3</td>
<td>International Telecommunication Union</td>
<td>Series X: Data Networks, Open System Communications and Security</td>
<td>2008</td>
</tr>
<tr>
<td>4</td>
<td>Gartner</td>
<td>Definition: Cybersecurity</td>
<td>2013</td>
</tr>
<tr>
<td>5</td>
<td>The Institution of Engineering and Technology</td>
<td>Resilience and Cyber Security of Technology in the Built Environment</td>
<td>2013</td>
</tr>
<tr>
<td>6</td>
<td>British Standards Institute</td>
<td>Guidelines for cybersecurity</td>
<td>2012</td>
</tr>
<tr>
<td>7</td>
<td>Australian Government</td>
<td>Cyber Security Strategy</td>
<td>2009</td>
</tr>
<tr>
<td>9</td>
<td>Government of Belgium</td>
<td>Cyber Security Strategy</td>
<td>2012</td>
</tr>
<tr>
<td>10</td>
<td>Government of Finland</td>
<td>Finland's Cyber Security Strategy</td>
<td>2013</td>
</tr>
<tr>
<td>11</td>
<td>French Network and Information Security Agency</td>
<td>Information systems defence and security France’s strategy</td>
<td>2011</td>
</tr>
<tr>
<td>12</td>
<td>Federal Ministry of the Interior</td>
<td>Cyber Security Strategy for Germany</td>
<td>2011</td>
</tr>
<tr>
<td>14</td>
<td>The Netherlands, Ministry of Security and Justice</td>
<td>The National Cyber Security Strategy (NCSS) 2</td>
<td>2013</td>
</tr>
<tr>
<td>15</td>
<td>New Zealand Government</td>
<td>New Zealand's Cyber Security Strategy</td>
<td>2011</td>
</tr>
<tr>
<td>17</td>
<td>Kingdom of Saudi Arabia</td>
<td>Developing National Information Security Strategy for the Kingdom of Saudi Arabia</td>
<td>2011</td>
</tr>
<tr>
<td>18</td>
<td>Republic of South Africa</td>
<td>Cybersecurity Policy of South Africa</td>
<td>2010</td>
</tr>
<tr>
<td>19</td>
<td>Republic of Turkey</td>
<td>National Cyber Security Strategy and 2013-2014 Action Plan</td>
<td>n/a</td>
</tr>
<tr>
<td>20</td>
<td>National Institute of Standards and Technology</td>
<td>Framework for Improving Critical Infrastructure Cybersecurity</td>
<td>2014</td>
</tr>
<tr>
<td>21</td>
<td>Spanish Cyber Security Institute</td>
<td>National Cyber Security, a commitment for everybody</td>
<td>2012</td>
</tr>
<tr>
<td>ID</td>
<td>Source</td>
<td>Title</td>
<td>Year</td>
</tr>
<tr>
<td>-----</td>
<td>-------------------------------</td>
<td>----------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>22</td>
<td>Republic of Poland</td>
<td>Cyberspace Protection Policy of The Republic of Poland</td>
<td>2013</td>
</tr>
<tr>
<td>24</td>
<td>Craigen, Dan</td>
<td>Defining Cybersecurity</td>
<td>2014</td>
</tr>
<tr>
<td></td>
<td>Diakun-Thibault, Nadia Purse,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Randy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Merriam-Webster</td>
<td>Definition of Cybersecurity</td>
<td>2015</td>
</tr>
<tr>
<td>26</td>
<td>Oxford Dictionary</td>
<td>Definition of Cybersecurity</td>
<td>2015</td>
</tr>
<tr>
<td>27</td>
<td>Amoroso, Edward</td>
<td>Cyber Security</td>
<td>2007</td>
</tr>
<tr>
<td>28</td>
<td>EastWest Institute</td>
<td>Critical Terminology Foundations 2</td>
<td>2014</td>
</tr>
<tr>
<td>29</td>
<td>New Definition</td>
<td></td>
<td>2016</td>
</tr>
</tbody>
</table>


ii https://ccdcoe.org/strategies-policies.html

iii Attributed to Edward Deming