



Explorations in Cyber International Relations

Massachusetts Institute of Technology Harvard University

What is Cybersecurity? Explorations in Automated Knowledge Generation

Nazli Choucri

Political Science Department
Massachusetts Institute of
Technology

Gihan Daw Elbait

Masdar Institute of Science
and
Technology, Abu Dhabi,
UAE

Stuart E. Madnick

Sloan School of Management
Massachusetts Institute of
Technology

November 6, 2012

This material is based on work supported by the U.S. Office of Naval Research, Grant No. N00014-09-1-0597. Any opinions, findings, conclusions or recommendations therein are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.



Citation: Choucri, N., Elbait, G. D., Madnick, S. E. (2012). *What is cybersecurity? Explorations in automated knowledge generation* (ECIR Working Paper No. 2012-4). MIT Political Science Department.

Unique Resource Identifier: ECIR Working Paper No. 2012-4.

Publisher/Copyright Owner: © 2012 Massachusetts Institute of Technology.

Version: Author's final manuscript.



Explorations in Cyber International Relations

Massachusetts Institute of Technology Harvard University

What is Cybersecurity? Explorations in Automated Knowledge Generation

Nazli Choucri

Department of Political Science
Massachusetts Institute of Technology

Gihan Daw Elbait

Masdar Institute of Science and Technology and
Massachusetts Institute of Technology

Stuart Madnick

Sloan School of Management and Engineering Systems Division
Massachusetts Institute of Technology

This work is funded by the Office of Naval Research under award number N00014-09-1-0597. Any opinion or findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Office of Naval Research

ECIR Conference on
Who Controls Cyberspace?

November 6-7, 2012



What is Cybersecurity?

Explorations in Automated Knowledge Generation

Nazli Choucri

Department of Political Science
Massachusetts Institute of Technology

Gihan Daw Elbait

Masdar Institute of Science and Technology and
Massachusetts Institute of Technology

Stuart Madnick

Sloan School of Management and Engineering Systems Division
Massachusetts Institute of Technology

Abstract

This paper addresses a serious impediment to theory and policy for cybersecurity: Trivial as it might appear on the surface, there is no agreed upon understanding of the issue, no formal definition, and not even a consensus on the mere spelling of the terms — so that efforts to develop policies and postures, or capture relevant knowledge are seriously hampered. In this context, we present a “proof of concept” for a new research strategy based on a close examination of a large corpus of scholarly knowledge, and the extent to which it enables us to generate new knowledge about cybersecurity of relevance to international relations and to national security relevant to the nation’s security and to international relations. Given the new cyber realities, this paper is also a “proof” of how to create new knowledge through automated investigations of the record to date.

1. Introduction

The construction of cyberspace and its worldwide use is a fact of daily life, for almost everyone and almost everywhere. At the same time, it has created a new set of technical, social and policy challenges of daunting proportions. The policy community in the United States has become increasingly concerned about threats to national security enabled by hostile uses of cyber access. While there may be a general understanding of such threats, the domain of cybersecurity remains to be better understood, if not fully mapped, than is currently the case. The U.S. government is fully cognizant of this situation. Observers and analysts alike have noted the importance of generating a holistic view of cybersecurity.

1.1 National Security and Cybersecurity

This paper addresses a serious impediment to theory and policy for cybersecurity: Trivial as it might appear on the surface, there is no agreed upon understanding of the issue, no formal definition, and not

even a consensus on the mere spelling of the terms — so that efforts to develop policies and postures, or capture relevant knowledge are seriously hampered.

In both national and international policy and scholarly contexts, the terms “cybersecurity” and “cyber security” – and sometimes even “cyber-security” – are used by different sources, in different countries, and for different purposes. This simple difference compounds difficulties in capturing all of the relevant knowledge in any investigation. Later on, we define this as the nomenclature problem.

In the international arena, such spelling and meaning variations are particularly problematic. Not only do they create difficulties in electronic and other communication modes, they also call attention to differences in policy postures – even among allies. Given that the international community already exhibits different understandings as well as policy postures on matters of cybersecurity, the inability to settle on a common spelling and meaning of the term highlights divergences and does little to facilitate convergence.

Inevitably, threats to security transmitted through cyber venues are contingent on structures and processes that constitute cyberspace, the new environment of human interaction. It is fair to say that we do not, as yet, have a holistic perspective on cyberspace or a common understanding of its various facets and manifestations. We know the Internet, not well, but enough to track policy relevance. Clearly the Internet is the core of cyberspace, the essential enabler, but not the entirety of the cyber domain. There is still much that we need to know.

This paper reports on a research initiative undertaken as a proof of concept for the value of automated generation of knowledge structure including identification of key elements, relationships among elements, and changes over time. We proceed from the assumption that national security is best served by improved knowledge of cyberspace and more detailed understanding of cybersecurity. It will also be better served by a holistic rather than a partial view of cyberspace and a comprehensive understanding of cybersecurity and its evolution.

1.2 International Relations

By now, it is evident that the construction of cyberspace has generated significant impacts on international relations. There is every indication that cyberspace has become an arena of considerable political importance, nationally and internationally. If there is clear evidence of the salience of cybersecurity – and the internationalization and politicization of cyberspace – it is the range of contentions around the current renegotiation of the International Telecommunication Regulations (ITR’s) – established in 1988.

Much has taken place in world politics between then and now – including such diverse events such as Stuxnet, Wikileaks, and the role of the Internet in the Arab spring, among many others. It is therefore surprising to find that a recent survey of eighteen major international relations journals as well as six influential policy journals for all years between 2000 and 2012 finds little attention to cybersecurity or to cyberspace. (See Appendix 1 for the list of Journals)

By contrast, the international political community has recognized the salience of cyberspace and the potentials for cyberthreats that undermine the security of states and create instability in the overall modes of international relations. For example, in 2010, the Secretary General of the United Nations transmitted to the General Assembly the *Report of the Group of Governmental Experts in the Field of Information and Telecommunications in the Context of International Security* [UN 2012]. The Report states bluntly that “there is increased reporting that States are developing ICTs as instruments of warfare and intelligence, and for political purpose.” The Report proceeds with an explicit enumeration of the Experts’ assessments and recommendations. Interestingly, the words “cybersecurity” and “cyberspace” – spelled

in any form, with or without hyphens or spaces – *do not appear anywhere in the text*. The words that are conventionally used are “information and communications technology”.

In the formal diplomatic context, there are considerable differences in the ways in which individual states refer to cyber issues, as is often the case in politics among nations. China and Russia do not refer to cybersecurity – hyphenated or otherwise – but use only the term “information security”, whereas the U.S. recognizes explicit threats to “cybersecurity.” The choice of words reflects national priorities and policy implications. Unlike the United States, Russia and China are concerned mainly with control over, and security of, the content that is available to their own citizens. Their focus is almost entirely on their own internal context. By contrast, security matters are about the sanctity and safety of networks, connectivity, and various facets of communication systems, not about controlling content available for its citizens. None of this addresses “what is cybersecurity” or yield any insights into the constituent elements thereof. It shows only that words matter – and matter a lot.

All of this creates something of a paradox: on the one hand, cyberspace is now firmly set as a new domain of high politics, and almost all states express concern for insecurities in the cyber domain. But, we have as yet no full mapping of the building blocks of this new domain nor its full implications for national security. Few would argue that our understanding of cybersecurity requires no further improvement. If, as noted above, the fields of international relations and security studies have given limited attention to cybersecurity, it is an exception in the scientific community as a whole.

The epistemic community, which is the broad range of knowledge-generators, has created a large and complex body of knowledge on security issues in cyberspace, coupled with all the ambiguities and uncertainties associated with the new reality. Delving deeply into such body of knowledge– and understanding the lessons to date – is essential if we are to develop viable policies for managing the domain as a whole and for reducing vulnerabilities to sources or consequences of cyber threats to security.

Automated knowledge aggregation systems consist of massive amounts of published materials, usually supported by search, retrieval, and select classification functions. Such knowledge-aggregation systems can now be explored for improving knowledge of the cyber domain. They can be mined to extract a full set of conceptual and technical facets of cybersecurity (or of the broader context of cyberspace). And they can be used to generate empirically based structures of knowledge for any issue of interest.

1.3 Proof of Concept

In this paper the “proof of concept” is about the extent to which our research strategy, based on a close examination of a large corpus of scholarly knowledge, enables us to generate new knowledge about cybersecurity that is relevant to the nation’s security and to international relations. Given the new cyber realities, this paper is also a “proof” of how to create new knowledge through automated investigations of the record to date.

Throughout we make only one critical assumption, namely that the corpus of scientific knowledge is an unexplored asset of significant proportions. Accordingly, we seek to:

- (1) Construct a holistic empirical view of cybersecurity built on the knowledge created by the scientific community;
- (2) Identify the constituent concepts, the individual building blocks of cybersecurity as well as the broader domain of cyberspace;
- (3) Highlight differences, if any, between science and engineering versus the social sciences perspectives, and capture the policy relevance of the difference; and
- (4) Determine notable changes over time for any of the above.

2. Research Strategy and Methodology

The quest for a holistic and empirically based view of cybersecurity and interactions in the cyber context begins with the choice of research approach for the investigation. We have selected to concentrate on customizing and implementing automated knowledge exploration with targeted applications for proof of concept purposes in the four issues above. The overarching task is to generate an empirically based hierarchical structure of knowledge of cybersecurity, a taxonomy, and its constituent components. In so doing, it is also useful to explore the structure of cyberspace.

2.1. Hierarchical Organization

The concept of **taxonomy** refers to a hierarchical organization of terms relevant to a particular domain of research. In general, taxonomies usually expand from a more general term at the top (root) to more specific terms at the bottom (leaves). The ordering of the terms in the taxonomy should reflect the inter-relationships between the words (and concepts) of relevance.

Of the many uses and functions of taxonomy generation, the following are among the most common:

- For decision makers and researchers, a taxonomy is used for better understanding the state or characteristic features of the area of interest.
- For investigators of a complex field consisting of several sub-fields, the structural organization of taxonomy will yield information about the ways in which sub-fields are connected within a given system.
- For tracking changes in over time, the taxonomy structure shows transformation of significance.

Accordingly, taxonomies are especially useful where it is beneficial to abstract from plain words, text, or descriptions to ordered, hierarchical, concepts and relationships.

The common purpose across these motivations and uses is to enhance internally consistent, ordered, and reliable information pertaining to a domain of interest from a credible academic online publication database. Taxonomy generation algorithms are based on the analysis of a data set of bibliometric information obtained using automated taxonomy generation mechanisms help advance the extraction of order and meaning. These provide a degree of built in quality assurance for the results. Automated systems also generate ways of *visually representing* the data in a manner that is easily usable and understandable for end users.

2.2 Automated Methodology

The automated taxonomy generation methodology *per se* consists of sequential steps, each contingent on the previous one. Figure 1 below shows the overall sequence of steps. With varying functions and complexities, each step is necessary to generate the product at the end of this process: An empirically-based structure of cybersecurity and its constituent elements presented in an ordered hierarchical order – and visual representation – derived from the corpus of knowledge and data base utilized. The result of the automated taxonomy generation method *is in the form of new knowledge derived from meta systems of aggregated knowledge*.

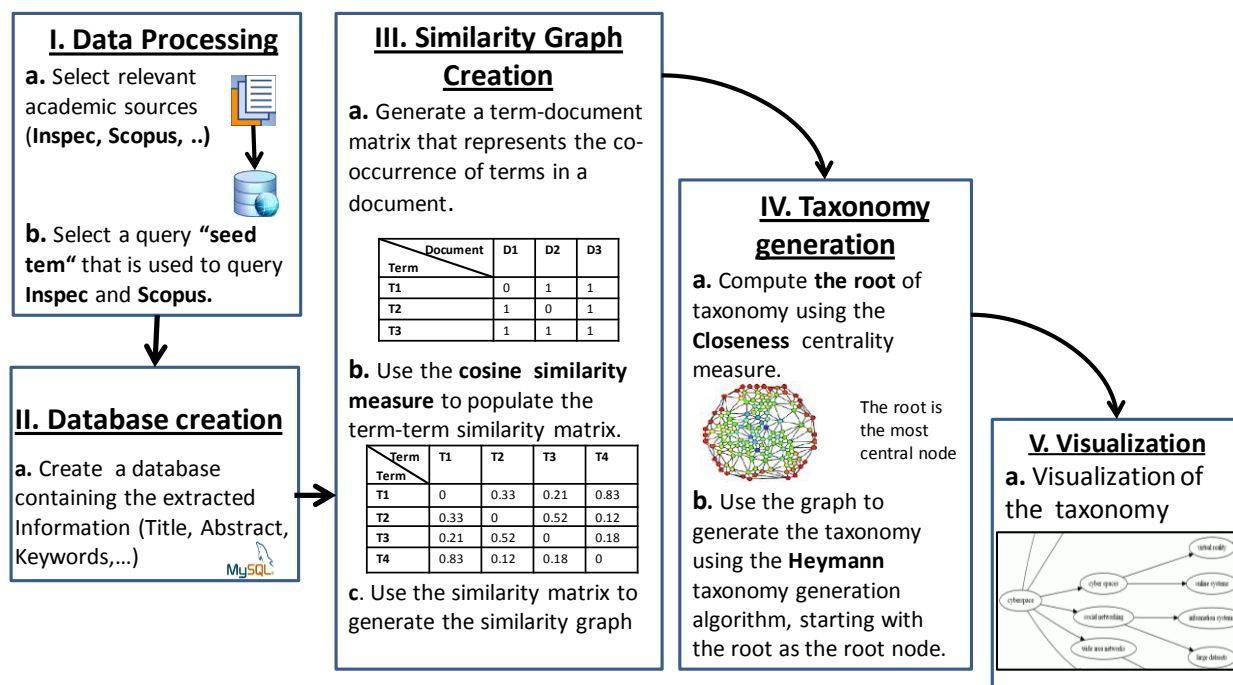


Figure 1: Overview of Method

The overview presented in Figure 1 provides only a broad schematic view of the method. We now turn to each step, noting the decisions, logic and specific operation. This automated taxonomy generation mechanism is characterized by high coupling, whereby each step consists of *actions*, *outcomes* or *results*.

2.3 Data Processing: Databases and Seed Term

First step, to select the databases and to choose seed terms, is foundational. For proof of concept purposes, we have identified two major data sets: one from *science and technology* that focuses on the technical construction and operation features (the supply side), and the other from the *social sciences* where attention to cyberspace and cybersecurity are driven more by end users and uses (the demand side), rather than technological features.

Following as a comparative analysis of different databases to that aggregate scientific and technical perspective, we selected *Inspec*, which contains over 13 million bibliographic records of engineering and computer science from more than 5,000 journals and conferences in the specific fields of Physics, Electronics and Electrical, Computing and Control, Information Technology and Manufacturing, Mechanical and Production engineering.

For the social sciences perspective, we chose *Scopus*, a large database containing more than 47 million records of citations and abstracts for articles published in more than 19,500 peer-reviewed journals from more than 5,000 international publishers and conferences world-wide in diverse fields, such as Arts and Humanities, Business, Management and Accounting, Decision Sciences, Economics, Econometrics and Finance, Environmental Science, Psychology, and Social Sciences.

These aggregated knowledge systems provide the content, the “raw” data, for the subsequent steps. The choice of a “seed term” – such as “cyber” or “cyberspace” – is used to query these databases via their online interface. With the appropriate software we then extract (“scrape”) each document’s bibliometric

information from the website. The result is a body of data containing bibliometric matter for all of the articles produced by the scraping process. The data produced is data is stored in a *local database file*.

2.4 Database Creation: Construct and Store

With the local database file in hand, the next step is to construct the data structure and store the data (that has just been extracted from the selected knowledge corpuses). The *stored database structure* includes the documents' title, abstract, and keywords.

2.5 Graph Construction: The Similarity Graph

The third step is to create a *similarity graph* that reflects the congruence between terms. It is the process by which the set of keywords in the stored database structure that will be used as the terms for the final taxonomies. This step is the most methodologically complex as well as analytically and computationally rich.

We start with building a term *co-occurrence matrix*. This is done by:

- (i) creating a term-document occurrence matrix that represents whether a certain term occurs (or not) in a document (see Figure 2a),
- (ii) converting this initial matrix to a term-term co-occurrence matrix (see Figure 2b), and

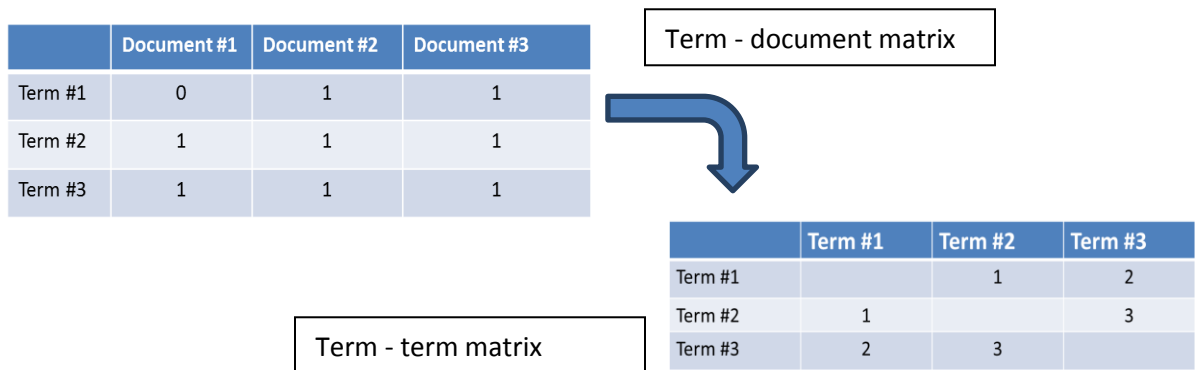


Figure 2 (a) Term-Document and (b) Term-Term matrices

- (iii) using the term-term co-occurrence matrix in conjunction with a similarity measure to convert the set of keywords to a term similarity graph, where the nodes are the terms and the edges represent the strength of their relationship.

The *similarity measures* utilized for this proof of concept investigation are:

- i. Cosine similarity (for undirected graph),
- ii. Symmetric normalized Google distance similarity (sNGD) (for directed graphs), and
- iii. Asymmetric normalized Google distance similarity (aNGD) (for directed graphs).

Each has specific benefits.

The *cosine similarity* is a similarity measure between two vectors based on the cosine of the angle between them. Given two vectors (in our case we generate these vectors e.g., ({"cyberspace": Set (12, 33, 41..., "doc-id"), }, ({"cyber security": Set (12, 33, 55..., 'doc-id>'), }

$$\text{Cosine similarity} = \frac{n_{x,y}}{\sqrt{n_x} \cdot \sqrt{n_y}}$$

Where n_x and n_y are the number of articles that contain terms x and y respectively, and $n_{x,y}$ is the number of articles that contain both x and y . The resulting similarity ranges from 0 and 1, where 0 means independent, and 1 means exactly similar.

Using the similarity measure, we then *populate a similarity matrix* (see Figure 3.) A matrix generated using cosine similarity or sNGD similarity will be symmetric across the diagonal, whereas one generated using asymmetric NGD similarity will not.

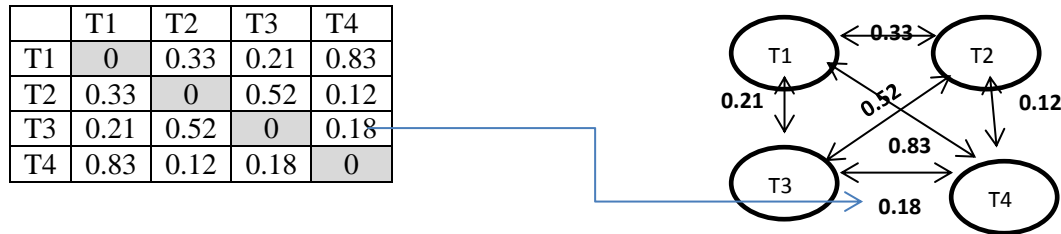


Figure 3: Populating a similarity matrix

2.5. Taxonomy Construction

The final steps in automated taxonomy generation is shown in Figure 4. The terms similarity (relationships) can be visualized as a graph shown as (a) below, and a spanning tree is then selected from among the edges in the graph. The spanning tree is transformed into taxonomy by instantiating one term, in this case term 5, as the root node. The rest of the taxonomy is formed by remaining consistent with the connections in the spanning tree, and thus creating a final taxonomy as shown in (b) below.

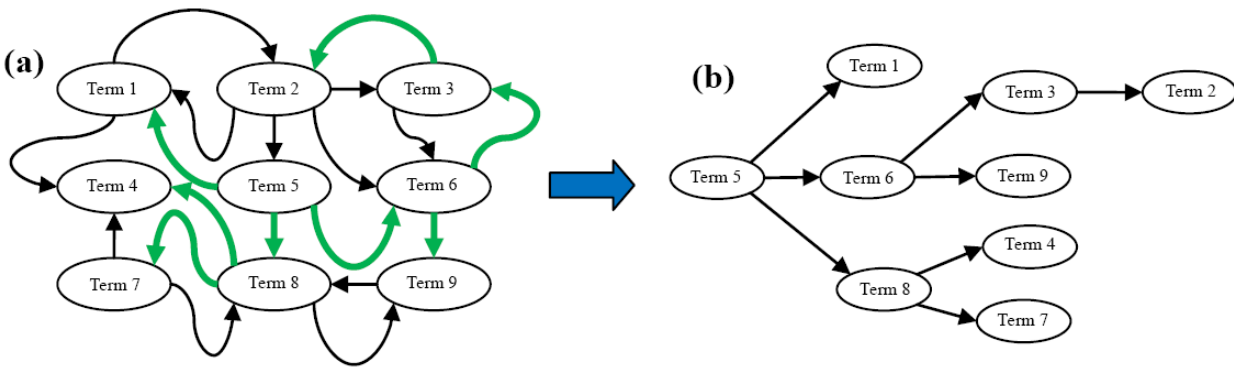


Figure: From (a) similarity graph to (b) a taxonomy

The notion of the root *node* of a taxonomy is shown in the figure below. The root node is identified by *the closeness centrality measure*; where closeness refers to the mean of shortest paths to other nodes starting from it, and those that have smaller mean shortest path lengths will have a higher closeness centrality metric.

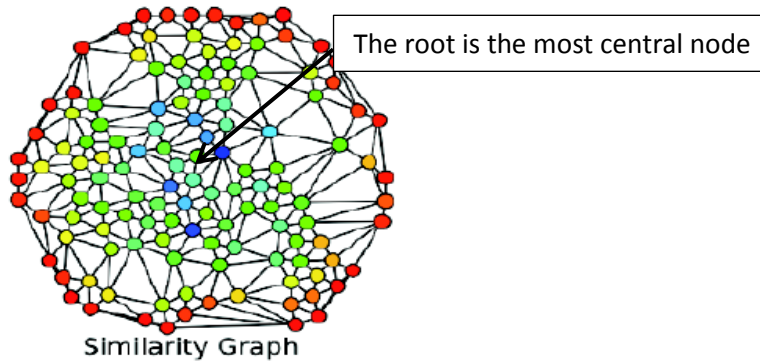


Figure 5: Determining the Root

The centrality measure is as follows:

$$C_{\text{closeness}}(v) = \frac{1}{\sum_{\substack{t \in V \\ t \neq v}} d_{v,t}}$$

Where $d_{v,t}$ is the length of the shortest distance in the graph between vertices v and t .

The next step is to select a *taxonomy algorithm* that takes a term-term similarity matrix and a root term to generate the taxonomy. In the course of this investigation, we implemented four taxonomy-generating algorithms, presented [Camina 2010, Ziegler 2009]. These are:

- i. Dijkstra-Jarnik-Prim's (DJP) Algorithm
- ii. Kruskal's Algorithm
- iii. Edmond's Algorithm
- iv. Heymann Algorithm

For the proof of concept we chose to work with the Heymann Algorithm [Heymann 2006], a method originally developed and intended for analysis of social networks and briefly explained below:

- The user tags documents or images with keywords.
- Each term (or tag) is associated with a vector that contains the annotation frequencies for all documents,
- which can then be compared to the vectors of the other terms using a variety of similarity measures thus
- produce a taxonomy where tags that are very similar to each other are linked together.
- The algorithm requires a list of tags in descending order of their generality.
- It then obeys the order starting with the most general tag and iteratively inserts each tag into a growing taxonomy by attaching them to either the most similar tag or the taxonomy's root.
- Only the tag threshold is used to represent the value of the similarity measure above which a link is permitted to be a child of a tag other than the root.

The algorithm process is depicted in Figure 6 below.

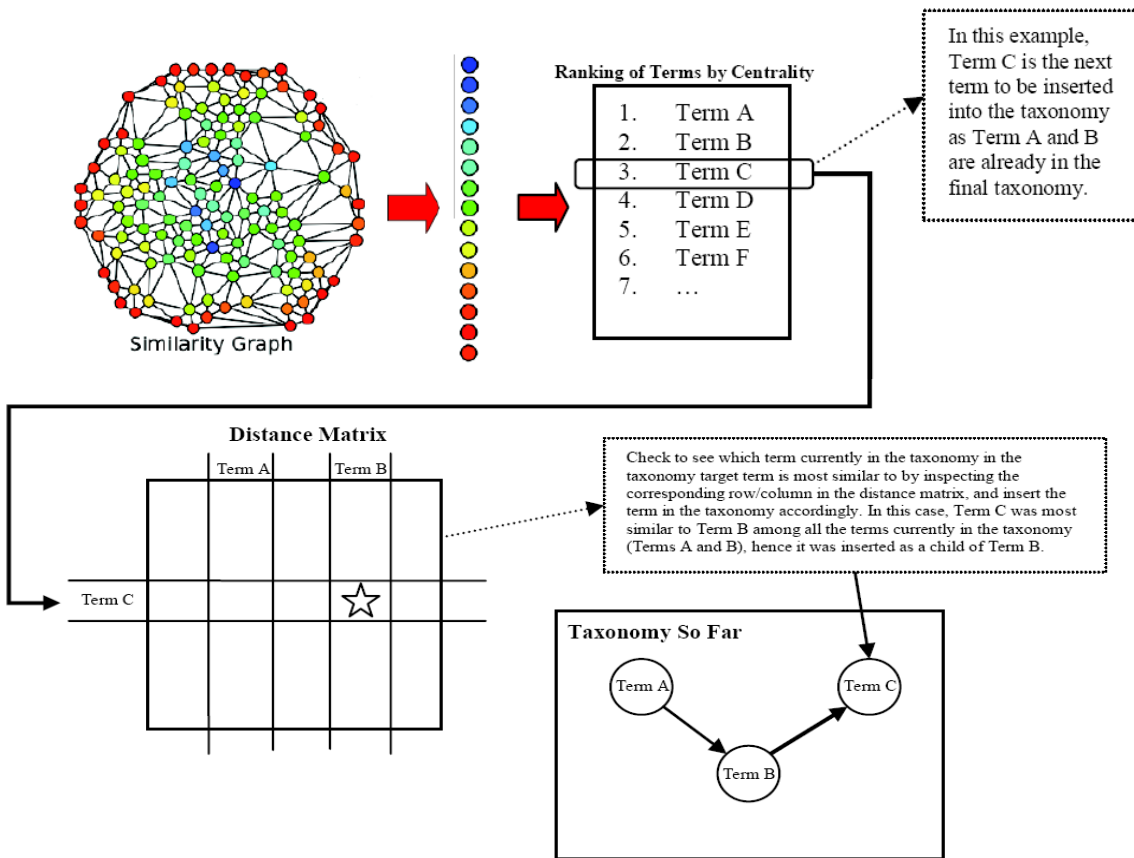


Figure 6: A sketch summarizing the process of the Heymann Algorithm

To simplify: the Heymann Algorithm iteration inserts new terms into the taxonomy by first sorting the terms not yet in the taxonomy by centrality, then taking the most central term not yet in the taxonomy and comparing it to each of the terms in the taxonomy to determine the one to which it is most similar. We have adapted the standard Heymann algorithm to work with the term similarity matrix described earlier. For more details on the specifics of these algorithms, refer to [Camina 2010; Henschel et al. 2009; Woon et al. 2009; and Ziegler 2009.]

The taxonomy generation mechanism is now complete and the results are available for visualization, the fifth step in this process.

2.6. Taxonomy Visualization

Using an interactive visualizer, we can now observe a full or partial representation of the knowledge structure, in Figure 7.

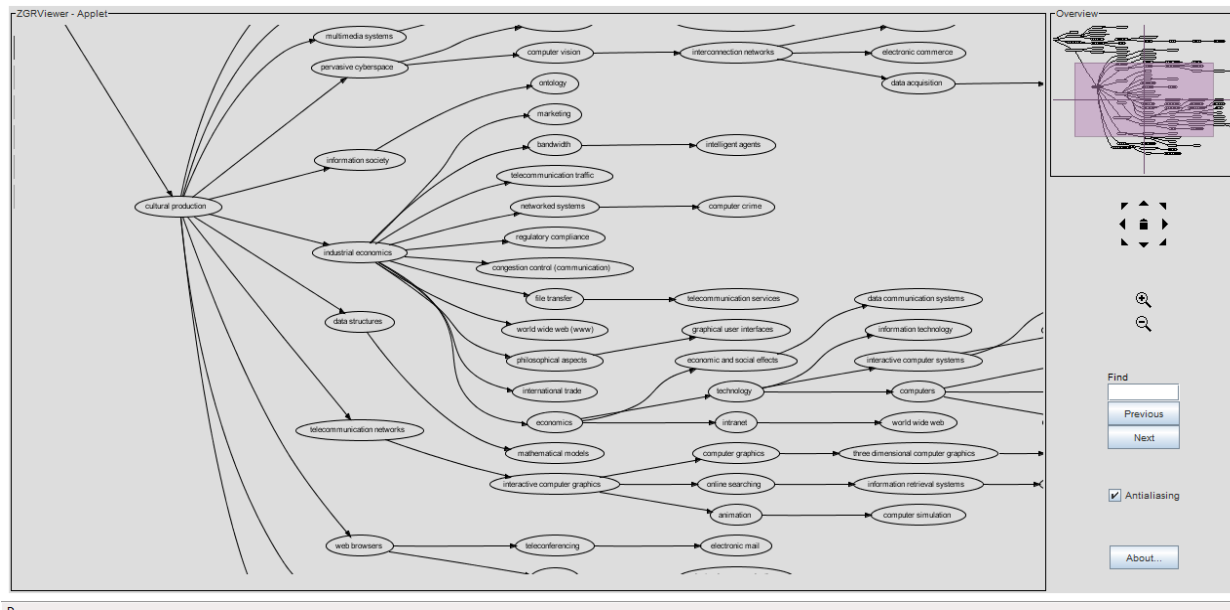


Figure 7: Visualization

This visual product presents the “map” of the domain generated from the body of knowledge considered that serves as the empirical database. The overall structure of the entire taxonomy is shown in the upper right corner and the highlighted section indicates the part of the taxonomy currently being viewed in the large window. This facilitates browsing and “zooming” around the taxonomy without losing track of where you are in the taxonomy.

3. Cyberspace as a Holistic Domain

For *contextual* purposes to help situate cybersecurity in its broader domain, we report first on the results of mapping cyberspace and the holistic view constructed by the taxonomy generation mechanism.

3.1 Mapping Cyberspace

We recognize that the construction of cyberspace was largely a science and engineering initiative. Nonetheless, the consequences and concerns associated with uses and misuses of cyber access are likely to encompass the socio-economic and political domain and addressed by social scientists.

As a first step, we realized that the concept of cyberspace must be framed both as “cyber space” and as “cyberspace” to capture the full references in the knowledge corpus examined. Table 1 shows the differences in uses of seed terms. Our purpose here is only to highlight the absence of unanimity or common norm of the core term in both the engineering and the social science knowledge corpuses.

Table 1 Number of results for the different databases queries using different seed terms

Seed terms	Inspec_ technical	Scopus_ social science
Cyberspace	2597	1703
“Cyber space”	628	68

The above table carries an important message, which, at first glance, may appear trivial. In the absence of agreed upon nomenclature in the scientific community, analysts cannot assume that the words they chose over all relevant options used. Given that cyber space and cyberspace are both used in the corpus of knowledge with no underlying “rule” to distinguish between the two, both must be utilized. The selection of one or the other will introduce unwarranted bias by excluding the knowledge to be derived from the other computationally. Any a priori assumption or choice of term will inevitably result in missing a body of data, that is, the data that is represented by alternative spelling, hyphens or other sources of variations. This is a simple case. For cybersecurity, and other related concepts, such as cyberthreat and cybercrime, the variations are more extensive.

3.2 Visual Representation: Comparative Results

We now turn to the visual representation of empirically derived taxonomy for cyberspace (using either spelling) in engineering (Figure 8a) and in the social sciences (Figure 8b) for the entire databases, that is, since earliest uses of the term. We applied the Heymann algorithm restricting the results to 500, 200, and 100 keywords on the assumption that we would derive added insights from this more extended investigation.

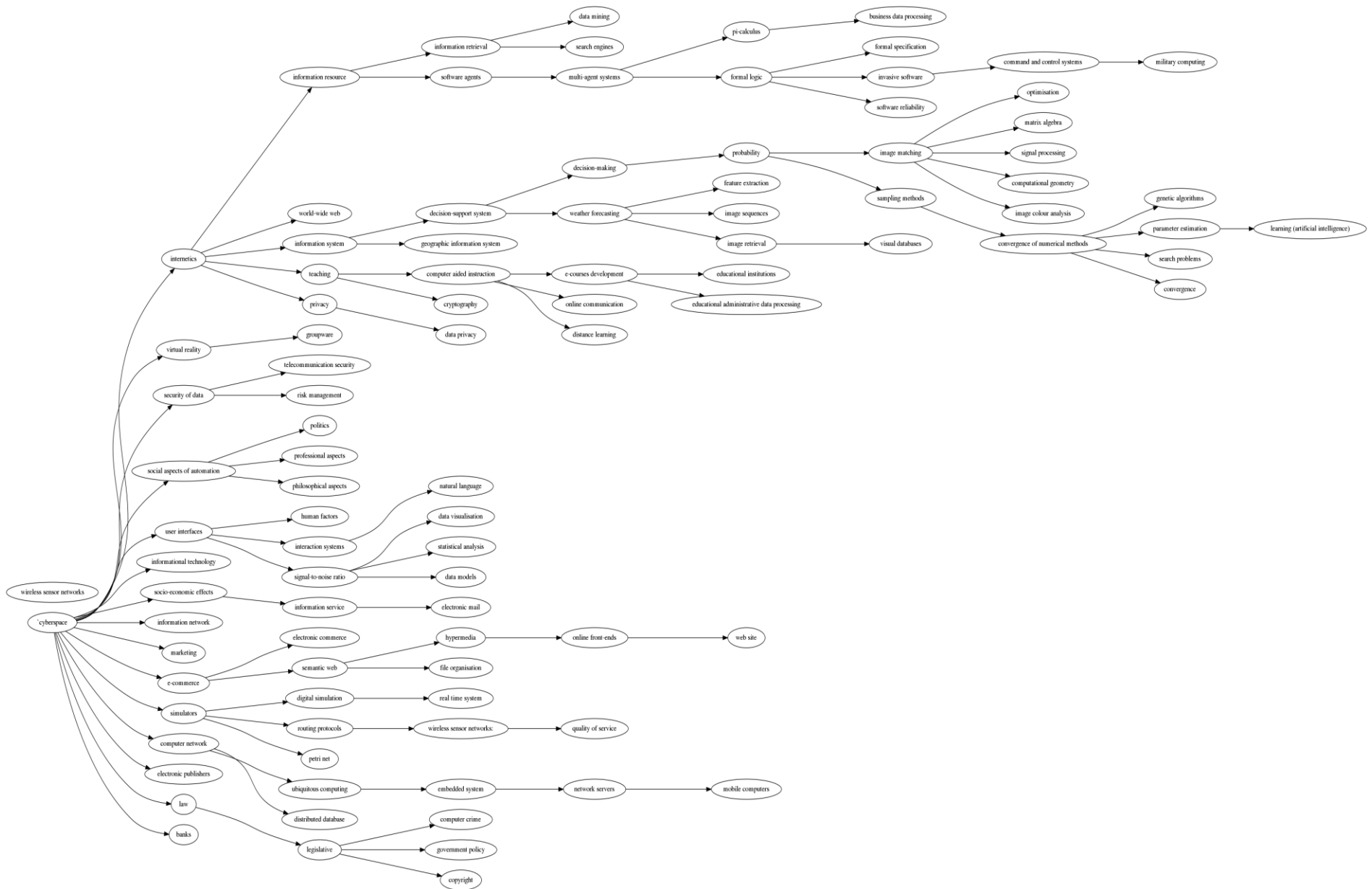


Figure 8a: Cyberspace Taxonomy using Inspec (Social Science perspective) with 100 terms

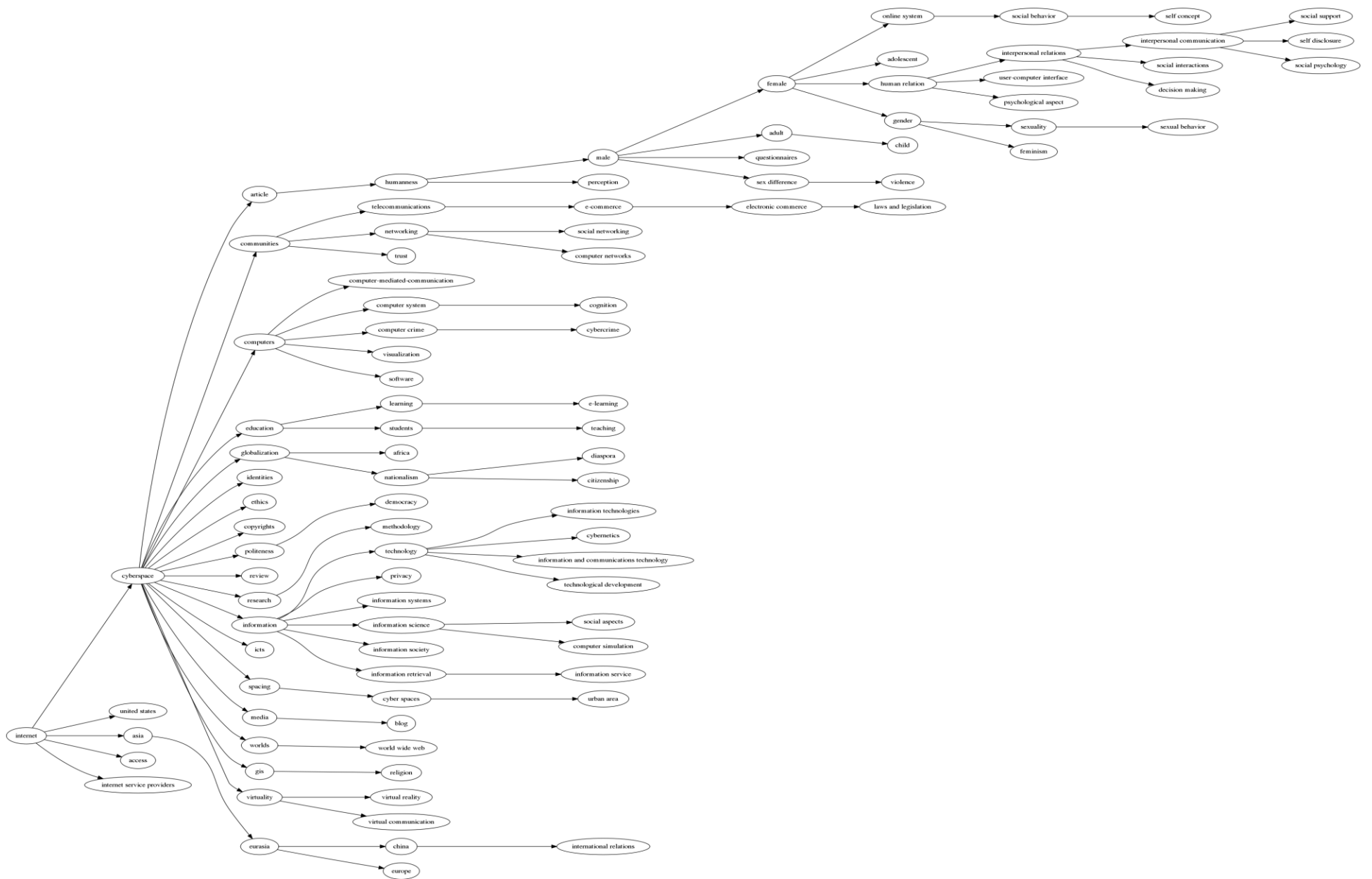


Figure 8b: Cyberspace Taxonomy using Scopus (engineering and technology perspective) with 100 terms

For this proof of concept study, we note that out of the 500 terms used to generate the taxonomies, only 35 terms are found to be in common in the taxonomy results when we use the seed term “cyberspace” in the social and technical databases. This signals a rather limited degree of overlap at the higher level of granularity.

When we focus on the 100 terms taxonomies, shown in Figure 8, we find that both the social sciences (Figure 8a) and the engineering and technology (Figure 8b) literatures produce a view of cyberspace that is highly *structured but the structures are different*.

In the social sciences, the root is the Internet which then leads to cyberspace, the US, Asia, access, and service providers. International politics (and embedded power relations) is salient and immediate in its association with cyberspace. The term “communities”, and its conceptual referents, is the most extensive outshoot of cyberspace and all of the derivatives are varieties of social features. A quick look at the first level off cyberspace provides a set of terms that represent the most salient issues for the social science community.

In the engineering literature, we see “internetics” as a level one offshoot of cyberspace and its further extension provides a set of seemingly unordered terms now associated with features and uses of the Internet. Interestingly, the only socio-economic effects (offshoot of cyberspace) are information service and electronic mail. Other socio impacts are notable off socio-aspects of automation. These patterns represent the literatures, not the underlying empirical “realities”.

These differences suggest that jointly the two views are likely to provide a more comprehensive perspective – especially for policy purposes – than either one individually. Again, this might seem trivial, but its implications for purposes of covering the entire domain as currently understood – as well as for policy purpose – are not.

4. Changes in Structure over Time: Breakpoint Analysis

4.1 Results from Science and Technology

Selecting the year 2000, the turn of the century, we focused on the cyber domain, with cyberspace as the root term. (It could be argued that the breakpoint choice is arbitrary, a most reasonable observation for a final investigation, rather than for a proof of concept.) Considering all of the empirically identifiable change in both the engineering and the social science scholarly literatures, we highlight five notable features for each perspective.

For science, engineering and technical representation of cyberspace based on the 100 key word inquiry, the notable features are as follows:

1. ***Consolidation of understanding and representation of cyberspace***
 - Decrease of terms derived from cyberspace from 23 to 9
 - Internet more specified, with almost half of the terms coming off the Internet
2. ***Emergence of a focus on banking***
 - Differentiation of banking from general business
 - Disappearance of specific business terms earlier noted
3. ***Organized representation of the Internet***
 - Internet is more specifically represented and organized

- Almost half of the terms come off the Internet
4. ***Changes in legal and legislative priorities***
 - Reduced attention to intellectual property and copyright
 - Increased focus on data privacy, computer crime, and government policy
 5. ***Growth in multimedia issues***
 - Decline in attention to communication and computing per se
 - Development from video data and academic computing

4.2 Results from the Social Sciences

Despite the existence of some common features, the social science literature carries a different representation of cyberspace and its evolution. The characteristic features for the social sciences are the following:

1. ***Consolidated technological understanding***
 - Cyberspace becomes a central term (not a branch), derived from the Internet (not the other way around)
 - Internet differentiated by countries and regions (rather than off communities)
2. ***Recognized complexity of cyberspace***
 - The linear “thin” representation is replaced by greater intricacy
 - Extraneous terms are no longer evident, they “drop off”
3. ***Redefined attention to people***
 - Earlier literatures treat humans as impersonal users, focusing on technical features as processing and development
 - The shift is to greater focus on self, gender, social interaction, and interpersonal relations
4. ***Emergence of new issue-concerns***
 - Technology and technological development
 - Global Information Systems (GIS) related to marketing and globalization (as well as religion, possibly for identification purposes?)
5. ***Shifts in attention to matters of ethics***
 - Earlier ethics is defined largely as research ethics, thus derivative of research
 - Ethics are repositioned as a function of war and terror; but also related to software (Possibly reflecting concern for underlying norms?)

4.3 Relevance to International Relations

Of the many features relevant to international relations, we note here only the most obvious. The social science literature appears far more sensitive to geographic (and therefore strategic) factors than engineering and sciences. This is hardly surprising. Nonetheless, the salience of U.S. and Asia in the first level of cyberspace is important. This, in itself, reflects the consolidation of concerns over the politics of cyberspace, with the U.S. and Asia potentially positioned in a seemingly “bipolar” system. (Elsewhere is an empirically based confirmation of this emergent bipolarity with the U.S. and China as the basic poles in a system of cyber access.)

Later on, with greater technical understanding, the domain of cyberspace is no longer represented in strictly bipolar terms. For the engineering knowledge base, the notable consolidation around the Internet after 2000 is significant, reflecting possibly, what is missing in the international relations literature noted in the introduction, namely attention to emergent structures in world politics.

Another distinctive feature is about education and learning, “process” matters. These factors are developed off information/technology, and extend to teaching, searches, e-learning and the like. By contrast, the science and engineering database shows a strong “product” focus on issues related to education (i.e., libraries, automated resources, etc.) and no attention to “processes”.

In many ways, this reinforces the results we find in the social science knowledge corpus of an increasing attention to individuals and human being in cyber domain. That the Human Rights Council referred specifically to the Internet reflects a departure from the uses of information and communication technologies (ICTs) noted earlier. It also indicates a potential diversification of references to cyber-related matters.

Finally, we note with interest, that the term “cybersecurity” – in any form of spelling – does not appear in the 100 word taxonomies derived from cyberspace as the seed term – either for the social sciences or for science and engineering.

5. Cybersecurity: Complexity or Coherence?

We turn next to the analysis and comparison of “cybersecurity” coverage in the two databases we have used (see Table 2.). In each case, we show the number of results the whole period for the 100-word specification. The important point here is to note the variation in the results by different treatment uses of cyber security as the seed term. At a minimum, we can infer that the nomenclature problem noted earlier remains unresolved.

Table 2 Number of results for the different databases queries using different seed terms

Seed terms	Inspec_ technical	Scopus_social science
cybersecurity	321	94
“cyber security”	702	140

Further, the taxonomy generation mechanism shows a far greater attention to cybersecurity by scientists and engineers than by social scientists. Again, this is entirely consistent with the critical gap we noted at the onset.

5.1 Keyword Explorations

As is often said: “What you see depends on how you look at it”. The same may be said, with appropriate caveats, for automated taxonomy generation. In this case, however, the empirical approach provides important anchors for the inquiry. We show in Table 3 results for automated investigations: with 100 keywords and with 500 keywords.

Table 3 Comparative Results

Taxonomy size and levels	Cybersecurity_scopus_social	Cybersecurity_inspec_technical
100 level 1	cybersecurity	Cybersecurity
100 level 2	Security systems, network security, security, engineering	Security of data, computer crime, internet, information security, investments, contracts
Interesting nodes	Security, law	Computer crime—cyber attacks—cybercrime--legislation
500 level 1	Cyber-security	US
500 level 2	Security systems, network security, security, engineering, computers, war, modeling , design , attacks	Data categories, automation, standardization, security
500 level 3	Security, attacks, law	Cybersecurity--risks, USA, SCADA, cyber security threats

5.2 Inferences and New Knowledge

The 100-word taxonomy generated is shown in Figure 9a and Figure 9b. First, there are notable differences in representations of the taxonomy structures, but also in their coverage when we look at the entire proof of concept period. For example, we note that when comparing the first and second level terms from cybersecurity for both the social sciences and the engineering literatures, we find that the engineering knowledge base is more differentiated in its reference to the offshoots of cybersecurity, distinguishing between data, computer crime, information security, investments and contract. Again, this may be a functional of operational experience.

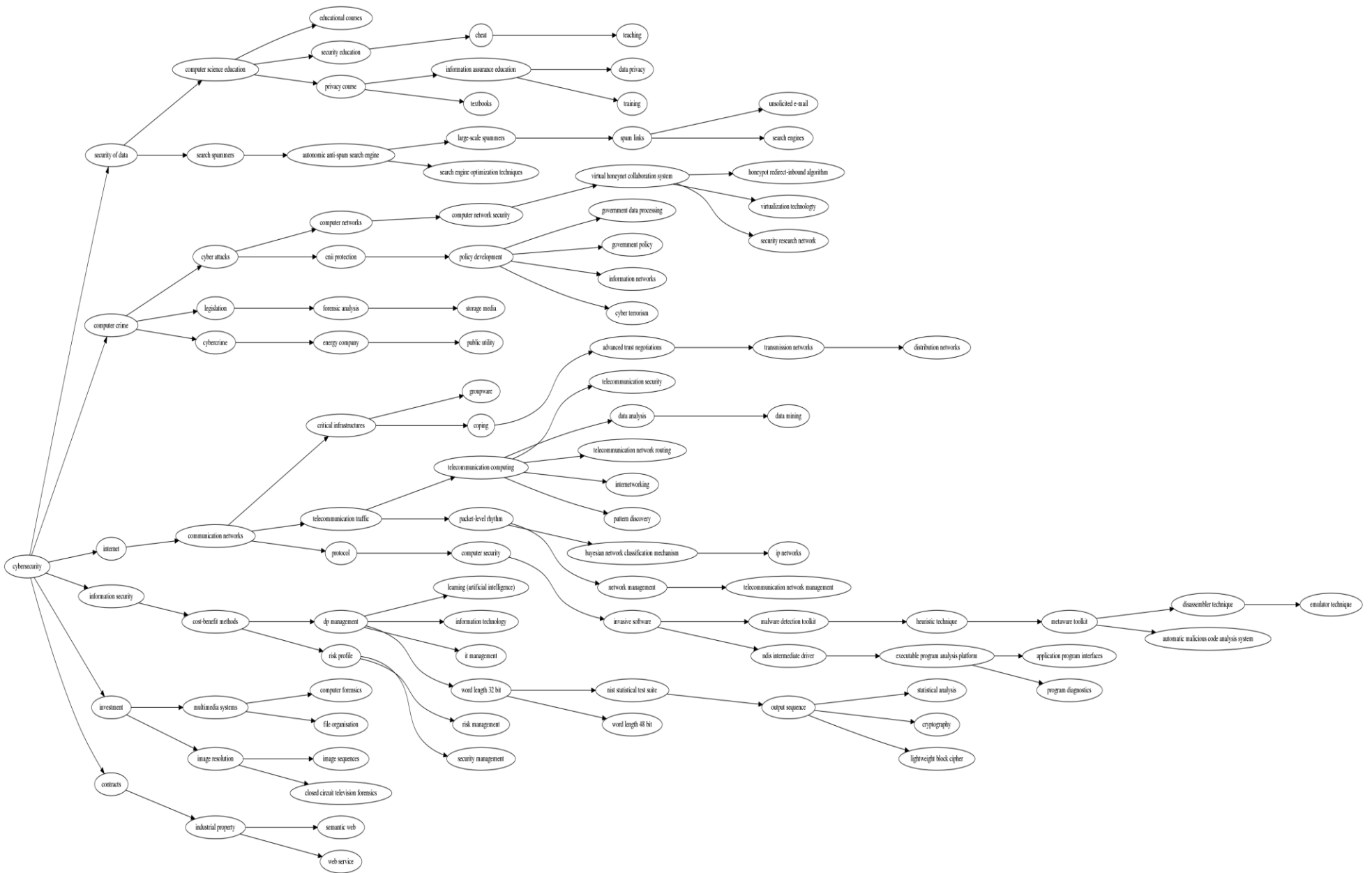


Figure 9a: Cybersecurity_Inspec_100 Words

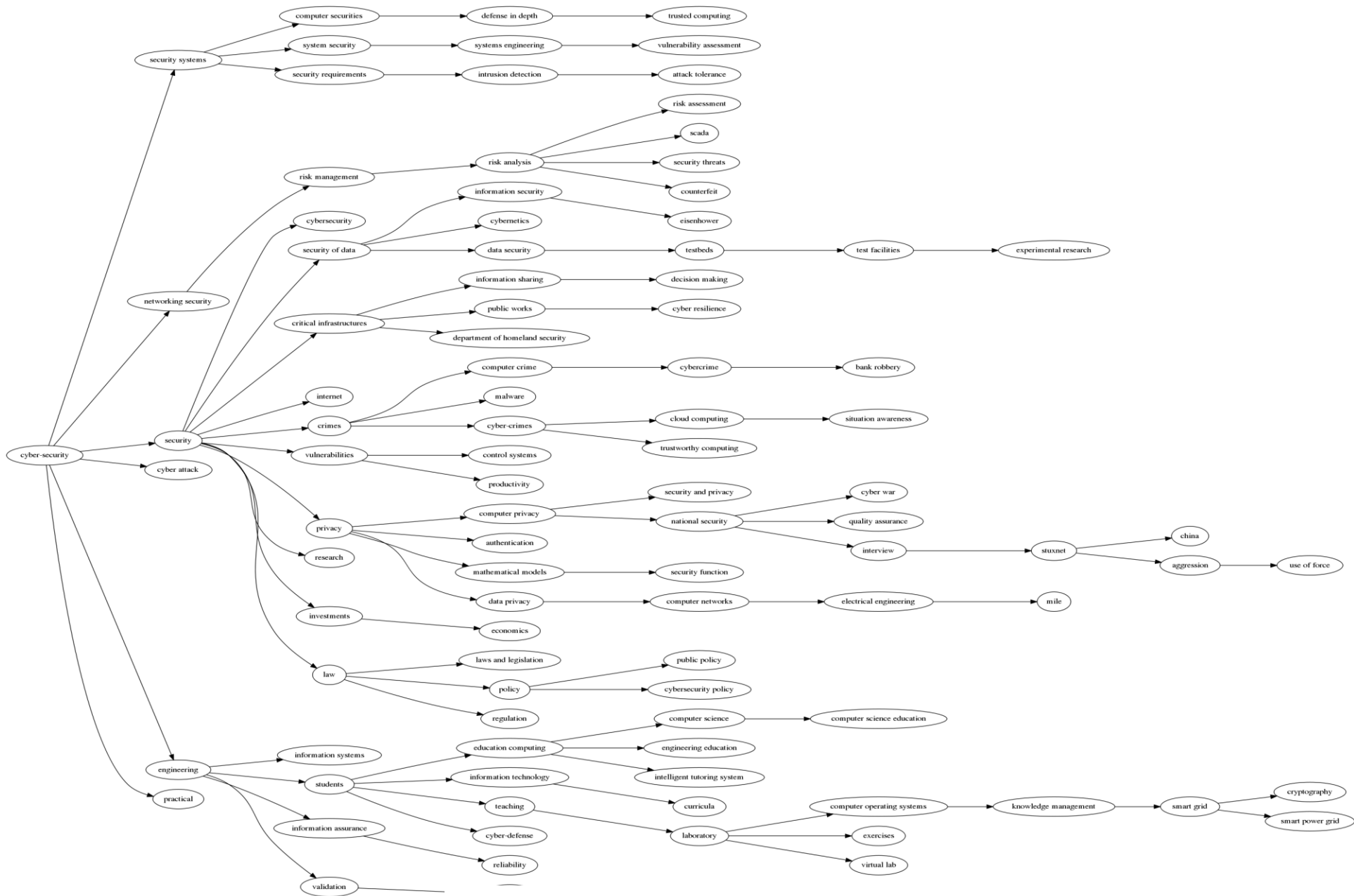


Figure 9b: Cybersecurity_Scopus_100 Words

Second, security, per se, as an offshoot of cybersecurity is highly developed in the social sciences; whereas as a clear indication of vulnerability literally has no derivatives. By contrast, the engineering cyber attack literature is highly developed along the trajectory of cyber attacks (off computer crime) and developed into four additional levels.

Turning to the taxonomy of 500 terms, some additional patterns emerge.

1. **Greater differentiations and a larger number of terms and relationships.** This is to be expected, but only to some extent, since the differentiation in level 2 is greater for the social sciences.
2. **There is no consistency on nomenclature,** cybersecurity and cyber-security are both utilized.
3. **The social science literature connects cybersecurity to war in** level 2 of the taxonomy, but conflict war and other related terms do not appear in the engineering mapping.
4. **Professional divisions consolidate** in level 3 of the taxonomy, in that the terms of security, attacks and law emerge in the social sciences. In the science literatures, we see USA, SCADA, and cyber security threat.
5. **Politicization of cybersecurity** is more apparent in the engineering knowledge bases.
6. **Relatively limited convergence,** only 88 terms are in common out of the 500 terms used to generate the taxonomy among the terms derived using the seed term “cybersecurity” in both the social and the technical literatures.

The low degree of convergence is significant: Neither of the bodies of knowledge alone can provide a holistic view of cybersecurity.

5.3 Changes in “Cybersecurity” and Shifts in Understanding

Consistent with the earlier reporting of results for 100-term taxonomy representing cyberspace both before and after the year 2000, we now focus on the 100 terms for cybersecurity, in the social science and in the engineering literatures.

Again, we see notable changes overall, which reflect, most likely, greater investments in research with targeted responses to national needs. The key features are noted below. These highlight about increased differentiation and understandings as well as notable “dead-ends” of arrested developments.

1. **Remarkable expansion and complexity of the cybersecurity domain,** reflecting greater granularity:
 - Early on, two major branches stem from security as the root, (i.e., security and cybersecurity) with all major extensions of terms stemming from cybersecurity.
 - Subsequently, cybersecurity is the root term with 13 offshoot branches.
2. **Greater differentiation in the characteristic features of cybersecurity,** consistent with the above:
 - Greater focus on data security issues is significant: security of data becomes its own branch directly off cybersecurity.
 - Similar shifts take place for information technology. Early on information technology stemmed from security of data, but later it appears at a direct branch.
3. **Bounded-system view of cybersecurity early on with no dead-ends off the root term.**
 - The two autonomous branches off the root expand and develop into further differentiations of concepts.
 - The terms off security and cybersecurity, together, provide an initial mapping of considerable coherence.

4. ***Open-system view of cybersecurity later on , with 6 direct dead-end offshoots of cybersecurity,*** that is, stems that do not lead to further key word (or knowledge) beyond their initial recognition.
 - **Five dead ends terms** are risk management, information secrecy, power engineering computing, information technology and information systems.
 - **National security is the sixth dead end.** This finding signals an underdevelopment of knowledge and research of the national security implications of cybersecurity and vice versa.
 - **The dead end terms** may well define priorities for the next round of investments in *cybersecurity research*.
5. ***Limited attention to homeland security,*** and only in social science. This is surprising but it may reflect the normal lag in the framing of research and reaching results.

6. Cybersecurity in International Relations

To date the scholarly field of international relations has shown relatively little attention of cyberspace and cybersecurity, as noted earlier. For this reason, the above results – in both the social sciences and science and engineering – will continue to provide the basis for knowledge on cybersecurity. As noted earlier in this paper, the paucity of articles in 18 major journals since 2000 indicates either a lack of interest by scholars in the field or, alternatively, a significant lag in the expression of interest.

At the same time, if we consider the proposed revisions of the International Telecommunication Regulations (ITR's) of the ITU, scheduled for a December 2012 conference, it is clear that the international community is still not fully converging on its images of the future. Implicit in this process is the increasing salience of security related issues, which are associated with cybersecurity – irrespective of the specific spelling or meaning involved.

We have found that both the social science and the science and engineering literatures represent overlapping but different understandings of the phenomenon. Given the relatively low overlap between the two sets of bodies of knowledge, it is fair to say that individually they cannot be relied upon to provide a comprehensive span of the constituent elements of cybersecurity, nor of the different ways in which the knowledge of cyberspace is structured.

Further, we cannot rely on knowledge about cyberspace to provide a sense of the knowledge structure of cybersecurity by social scientists or by scientists, technologists and engineers. Based on the results presented earlier in this paper, we can now highlight some overarching features of knowledge of cybersecurity – over and above the detailed features presented earlier.

- (1) While considerably more attention has been given to the “supply” side of cyberspace (that is, the technical and operational features) than to the “demand” side (patterns of access and use) this attention can also provide important information about potential points of vulnerability in cyber systems.
- (2) The levels of the hierarchical structure that represents cybersecurity shows the relative salience of its different features as reflected by the occurrence of various terms as well as the relationship among the terms. This helps the understanding of a much broader range of relevant elements.
- (3) There is a consolidated understanding within each of the knowledge data bases about the characteristic features of cybersecurity. We see the emergence of a more differentiated representation of the characteristic features of cybersecurity embedded in each of the knowledge corpuses.

- (4) While engineers and scientists do not agree with the social scientists regarding the critical features of cybersecurity, they each have distinctive understandings.
- (5) The consolidation identified in this paper provides “critical mass” to the coherence of concept and salience of cybersecurity in the international arena, thus assisting in its internationalization. This is an important, but not sufficient prerequisite for an international recognition of the issues area and its legitimacy in international discourse.
- (6) At the same time, the divergence of views is useful for generating a more holistic view of the issue-area compared to a view that is either solely social science orientated or dominated by technology and engineering.
- (7) For policy purposes, in international relations a more holistic perspective provides something of an “insurance policy” as we begin to develop more coherent strategies to manage this threat.
- (8) While we expect that the central tendencies in the two knowledge systems used to construct and analyze the database are shaped professional considerations, cyber crime currently appears in the engineering knowledge base but not in the social science corpus.
- (9) Unsurprisingly, networks appear in the social sciences, but not in engineering, science and technology. Further, it is the social scientists that stress security and law.
- (10) This “proof of concept” paper shows that automated examination of the corpus of scholarly knowledge enables us to generate *new knowledge* about cybersecurity relevant international relations and national security. It also highlighted features of cybersecurity beyond the more commonly cited.
- (11) This research strategy generated a more holistic empirical perspective – including constituent concepts, individual building blocks, and derivative – based on science and engineering as well as the social sciences. It yielded not only knowledge structure and coherence, but also identified features that are usually not considered relevant to cybersecurity in common discourse.
- (12) Despite an overlap in coverage between the knowledge bases of the social sciences and of engineering materials noted earlier in this paper, the two data sources generate distinct and different patterns of knowledge about the critical features of cyberspace and cybersecurity. While caution is always necessary, we have no basis for suggesting that the overlap creates distortions tending toward convergence. In other words, despite the overlap, divergence dominates.

References

- [Blaschke 2002] Blaschke, C., Valencia, A. *Automatic Ontology Construction from the Literature*. Genome Informatics, Volume 13, 2002, pp. 201-213.
- [Camina 2010] Camina, Steven. *A Comparison of Taxonomy Generation Techniques Using Bibliometric Methods: Applied to Research Strategy Formulation*. EECS Thesis, Massachusetts Institute of Technology, 2010.
- [Chuang et al. 2002] Chuang, S., Chien, L., *Towards Automatic Generation of Query Taxonomy: A Hierarchical Query Clustering Approach*. Academia Sinica, Taipei, 2002.
- [Henschel et al. 2009] Henschel, A., Woon, W., Wachter, T., Madnick, S. *Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2009.
- [Heymann 2006] Heymann, P., Garcia-Molina, H., *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. InfoLab Technical Report, Stanford University, 2006.

- [Krishnapuram 2003] Krishnapuram, R., Kimmamuru K., *Automatic Taxonomy Generation: Issues and Possibilities*. Lecture Notes in Computer Science, Springer, Berlin, 2003.
- [Reardon and Choucri 2012] Reardon, R. and Choucri, N. *The Role of Cyberspace in International Relations: A Review of the Literature*, paper presented at the 2012 Annual Meeting of the International Studies Association, San Diego, CA.
- [Sánchez 2004] Sánchez, D., Moreno, A., “Automatic Generation of Taxonomies from the WWW.” In *Proceedings of the Practical Aspects of Knowledge Management, (PAKM 2004)*, Vienna, Austria. Volume 3336, pp. 208-219.
- [UN 2012] United Nations General Assembly, Human Rights Council, Twentieth Session Agenda item 3 A/HRC/20/L.13, June 2012, p. 2.
- [Woon et al. 2009] Woon, W., Henschel, A., Madnick, S. *A Framework for Technology Forecasting and Visualization*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2009.
- [Ziegler 2009] Ziegler, B. *Methods for Bibliometric Analysis of Research: Renewable Energy Case Study*. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology, 2009.

Appendix 1. Journals Reviewed

Scholarly IR Journals (18):

- American Journal of Political Science
- American Political Science Review
- Annals of the American Academy of Political and Social Science
- British Journal of Political Science
- European Journal of International Relations
- International Interactions
- International Organization
- International Political Science Review
- International Security
- International Studies Quarterly
- Journal of Conflict Resolution
- Journal of Peace Research
- Journal of Strategic Studies
- Millennium
- Perspectives on Politics
- Political Science Quarterly
- Security Studies
- World Politics

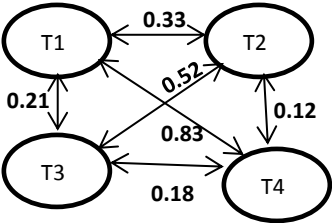
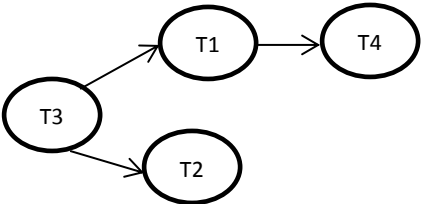
Policy Journals (8):

- Current History
- Foreign Affairs
- Foreign Policy
- International Affairs
- Journal of International Affairs
- National Interest
- Survival
- Washington Quarterly

Source: Robert Reardon and Nazli Choucri *The Role of Cyberspace in International Relations: A Review of the Literature*, paper presented at the 2012 Annual Meeting of the International Studies Association, San Diego.

Appendix 2

Summary of the Operational steps of the taxonomy generation process: Actions and Outcomes

	Actions	Outcomes																																									
Step 1: Data processing	Review the set of relevant online academic sources(Inspec, Compendex, Scopus) a. Choose a query “seed term” and use it to query online databases	a. Verify selected data bases from a list of data bases most relevant sets b. Select “cyberspace” and “cybersecurity” for different queries of the online databases and extract relevant fields.																																									
Step 2: Database creation	a. Construct the database containing Title, Authors, Abstract and Keywords	a. An SQL database that includes each document’s title, authors, abstract, and keywords																																									
Step 3 Graph creation	<p>a. Generate a term-document matrix that represents the co-occurrence of terms</p> <p>b. Convert the keywords set to term-term similarity matrix by using the cosine similarity measure</p> <p>c. Use the term similarity matrix to generate a graph</p>	<p>a. A term-term matrix</p> <table border="1" data-bbox="1000 747 1349 884" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th></th> <th>D1</th> <th>D2</th> <th>D3</th> </tr> </thead> <tbody> <tr> <th>T1</th> <td>0</td> <td>2</td> <td>2</td> </tr> <tr> <th>T2</th> <td>3</td> <td>0</td> <td>1</td> </tr> <tr> <th>T3</th> <td>4</td> <td>1</td> <td>2</td> </tr> </tbody> </table> <p>b. A term similarity matrix</p> <table border="1" data-bbox="940 984 1409 1159" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th></th> <th>T1</th> <th>T2</th> <th>T3</th> <th>T4</th> </tr> </thead> <tbody> <tr> <th>T1</th> <td>0</td> <td>0.33</td> <td>0.21</td> <td>0.83</td> </tr> <tr> <th>T2</th> <td>0.33</td> <td>0</td> <td>0.52</td> <td>0.12</td> </tr> <tr> <th>T3</th> <td>0.21</td> <td>0.52</td> <td>0</td> <td>0.18</td> </tr> <tr> <th>T4</th> <td>0.83</td> <td>0.12</td> <td>0.18</td> <td>0</td> </tr> </tbody> </table> <p>c.</p> 		D1	D2	D3	T1	0	2	2	T2	3	0	1	T3	4	1	2		T1	T2	T3	T4	T1	0	0.33	0.21	0.83	T2	0.33	0	0.52	0.12	T3	0.21	0.52	0	0.18	T4	0.83	0.12	0.18	0
	D1	D2	D3																																								
T1	0	2	2																																								
T2	3	0	1																																								
T3	4	1	2																																								
	T1	T2	T3	T4																																							
T1	0	0.33	0.21	0.83																																							
T2	0.33	0	0.52	0.12																																							
T3	0.21	0.52	0	0.18																																							
T4	0.83	0.12	0.18	0																																							
Step 4 Taxonomy generation	<p>a. Compute the root of the taxonomy using the “Closeness” centrality measure.</p> <p>b. Starting with the root as the root node, use the similarity matrix to generate the taxonomy by applying the Heymann algorithm</p>	<p>a. The root of the taxonomy</p> <p>b. A taxonomy</p> 																																									

<p>Step 5 Visualization</p>	<p>a. Use an interactive visualizer (ZGRViewer) to represent the taxonomy</p>	<pre>graph LR; cyberspace --> cyber_spaces[cyber spaces]; cyberspace --> social_networking[social networking]; cyberspace --> wide_area_networks[wide area networks]; cyber_spaces --> virtual_reality[virtual reality]; cyber_spaces --> online_systems[online systems]; social_networking --> information_systems[information systems]; wide_area_networks --> large_datasets[large datasets];</pre>
---------------------------------	---	---