Building Taxonomies based on Human-Machine Teaming: Cyber Security as an Example

Mohamad Imad Mahaini
School of Computing
University of Kent
Canterbury, UK
mim@kent.ac.uk

Shujun Li
School of Computing
University of Kent
Canterbury, UK
S.J.Li@kent.ac.uk

Rahime Belen Sağlam
Computer Engineering Department
Ankara Yıldırım Beyazıt University
Ankara, Turkey
rbsaglam@ybu.edu.tr

ABSTRACT

Taxonomies and ontologies are handy tools in many application domains such as knowledge systematization and automatic reasoning. In the cyber security field, many researchers have proposed such taxonomies and ontologies, most of which were built based on manual work. Some researchers proposed the use of computing tools to automate the building process, but mainly on very narrow sub-areas of cyber security. Thus, there is a lack of *general* cyber security taxonomies and ontologies, possibly due to the difficulties of manually curating keywords and concepts for such a diverse, inter-disciplinary and dynamically evolving field.

This paper presents a new human-machine teaming based process to build taxonomies, which allows human experts to work with automated natural language processing (NLP) and information retrieval (IR) tools to co-develop a taxonomy from a set of relevant textual documents. The proposed process could be generalized to support non-textual documents and to build (more complicated) ontologies as well. Using the cyber security as an example, we demonstrate how the proposed taxonomy building process has allowed us to build a general cyber security taxonomy covering a wide range of data-driven keywords (topics) with a reasonable amount of human effort.

KEYWORDS

cyber security, taxonomy, ontology, knowledge representation, natural language processing (NLP), information retrieval (IR), online social network (OSN), Twitter, visualization

ACM Reference Format:
Mohamad Imad Mahaini, Shujun Li, and Rahime Belen Sağlam. 2019. Building Taxonomies based on Human-Machine Teaming: Cyber Security as an Example. In *Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES '19), August 26–29, 2019, Canterbury, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3339252.3339282

1 INTRODUCTION

Taxonomies and ontologies are both useful knowledge representation tools for systematically and structurally conceptualizing human knowledge about objects (or things) and concepts in many domains, especially in sciences, engineering, business and education [1]. The two words "taxonomy" and "ontology" have very similar meanings, but the latter has a more theoretical flavor and normally requires more advanced components such as relations between concepts, therefore allowing formal reasoning about the meanings of sentences [18].

Cyber security is a highly inter-disciplinary and dynamically evolving subject. It is not surprising to see that many researchers and practitioners have attempted to build and use taxonomies and ontologies to better organize our knowledge on different sub-areas of the broad subject. See Section 2 for a brief overview of some related work on cyber security taxonomies and ontologies. Although there has been a lot of work on taxonomy and ontology building, there is a general lack of more automated processes for building taxonomies and ontologies. In addition, more general taxonomies and ontologies covering the whole subject are rare, possibly due to the more demanding human effort required, the complexity of putting everything together and the constant effort to keep such taxonomies and ontologies up to date.

In this paper, we propose to apply the new concept of human-machine teaming [8] for taxonomy building, in order to reduce the human effort involved in the building process and to make it easier to create and maintain the built taxonomy. The proposed process includes three stages: A) the data collection phase for preparing a large set of textual documents of interests and also documents in other areas, B) the text analysis phase for processing the textual documents to produce a list of relevant terms (keywords); C) the taxonomy building phase for creating and refining the taxonomy based on a defined structure and assignment of all terms from Stage B to the structure. Stage A involves manual selection of textual documents done by the human analyst and automatic processing of collected data to form a properly formatted dataset ready for Stage B. Stage B is heavily automated using natural language processing (NLP) and information retrieval (IR) tools, but the final selection of terms is controlled by the human analyst. Stage C is mostly done manually based on the human analyst's expert knowledge, but can be facilitated by an automated tool for visualizing the taxonomy. Taking cyber security as an example domain, we demonstrate how the proposed process has been used to build a general cyber security taxonomy with a reasonable amount of human effort and a large set of textual documents processed by automated NLP and IR tools.

The rest of the paper is organized as follows. Related work is given in Section 2. The proposed taxonomy building process is described in Section 3. Then, Section 4 explains the data collection stage for building the example cyber security taxonomy. The second

ARES '19, August 26–29, 2019, Canterbury, United Kingdom

ACM ISBN 978-1-4503-7164-3/19/08.
https://doi.org/10.1145/3339252.3339282

stage on text processing is covered with greater details in Section 5. The last stage on building the actual cyber security taxonomy is presented in Section 6. Possible future work is discussed in Section 7. Finally, Section 8 concludes the paper. Some supplementary material is provided in the Appendix, covering more details of the constructed cyber security taxonomy and its possible applications.

2 RELATED WORK

2.1 Automatic and Semi-Automatic Taxonomy and Ontology Building

Automatic and semi-automatic processing of information for building taxonomies and ontologies has been an active topic in different research fields. For instance, one popular technique, Formal Conceptual Analysis (FCA) [21], has been widely used to automatically construct a formal ontology from a given set of objects and their properties[5]. NLP and machine learning techniques have also been widely used to automate taxonomy and ontology building, especially based on natural language texts [26]. Such techniques are less used in building cyber security taxonomies and ontologies, as reviewed in the next two subsections.

2.2 Selected Taxonomies in Cyber Security

Critical infrastructure (CI) protection is one of the cyber security sub-domains where researchers have proposed frameworks for building taxonomies. In [16], Luiijf and Nieuwenhuijs proposed a generic threat taxonomy for CI that is made up of 325 nodes. They built an extensible taxonomy to support adding more elements to their taxonomy as they did not develop their taxonomy from scratch but relied on existing threats databases instead. Moreover, in [14] Jiang et al. proposed a domain-specific language for security in CI. They created a simple taxonomy for CI and cyber components.

There are also studies that focused on building taxonomies for cyber-physical threats and attacks. In [12], Heartfield et al. built a taxonomy for the cyber security threats that affect smart homes. In their taxonomy, they considered the impact on both the system and users. On the other hand, Loukas t al. [15] worked on vehicle security attacks and they created a taxonomy for the characteristics of Intrusion Detection Systems (IDS) for different types of vehicles. In [24] Sedjelmaci and Senouci also worked on vehicles but aerial ones. They examined the current detection schemes for aerial vehicle security and then classified them into a small taxonomy.

In their study [22], Radmans et al. focused on creating a taxonomy for the cyber security attacks on wireless sensor networks. They studied many possible attacks on such networks. Their final taxonomy was limited to classes about attack categories (internal or external) and whether they are active or passive.

Cyber threat intelligence (CTI) is a sub-domain in which a lot of ontologies have been developed to facilitate information sharing among different organizations and computer systems. Such ontologies are mostly represented as a common data format such as IODEF (Incident Object Description and Exchange Forma) [7], STIX (Structured Threat Information eXpression) [19] and OpenIOC (Open Indicators Of Compromise) [11]. In [3], Burger et al. created a taxonomy model to analyze and classify existing CTI ontologies. On the other hand, in [17] Mavroeidis and Bromander surveyed existing CTI taxonomies and ontologies at the time, and they concluded that none of them are readily available to be used within CTI due to lack of expressiveness. They also suggested some actions in order to address this problem.

In [10], Elnagdy et al. built a knowledge structure (i.e., a mini taxonomy) of cyber insurance for practitioners in this specific industry.

One of the most comprehensive taxonomies in the cyber security domain was proposed by Canbek et al. in [4], where they focused on the mobile security domain and built a large taxonomy covering different concepts. Supporting their taxonomy with two sub-taxonomies for mobile malware and mobile malware analysis, they proposed an overall hierarchy with over 1,300 nodes.

2.3 Selected Ontologies in Cyber Security

Quite a number of studies about building – or using – ontologies for the cyber security domain exist in the literature. However, most of them were created for a particular application such as detecting vulnerabilities excluding several important concepts in cyber security. Here, we review some typical work on this topic.

In [9] Elahi et al. proposed a design for an ontology about the security concepts related to vulnerabilities in software. Their main goal was to integrate the captured knowledge in the security system requirements. In [23], Razzaq et al. proposed a method based on semantic web techniques to detect attacks on web applications by analyzing users' requests since those requests cover rich attack information. They also created ontology models for attacks and communication protocols. In [25] Wang and Guo proposed an ontology for vulnerabilities. They populated their ontology using the description of some common vulnerabilities taken from the NVD (National Vulnerability Database). In [27], Zamfira and Ciocarlie proposed a meth for creating an ontology that can be used for detecting cyber security attacks. The ontology they built focuses on cyber operation and conceptualizes different data needed in the processes. They also tested the use of the ontology with a prototype web firewall, and showed the ontology did help.

Maybe the most similar work to ours reported in this paper is [6], in which Costa et al. set an ontology for modelling insider threat attacks. The most interesting part – for this paper – is their semi-automated approach to developing their ontology. They collected sources related to insider threats cases and used NLP tools to automatically parse the extracted text sentences, then they used human analysts to determine the meaning of each sentence for building the ontology. Their work was only on cyber indicators related to this type of threats. This differs from our work on several aspects, (1) their output is an ontology while ours is a taxonomy. (2) Our focus is the cyber security domain as a whole, not just insider threats. (3) We use NLP and *n*-gram ranking and the human expert is only involved in taxonomy creation stage.

3 PROPOSED TAXONOMY BUILDING METHOD

The high-level overview of our proposed taxonomy building process is illustrated in Figure 1. This process consists of three main stages as outlined below.

Stage A is for **data collection**. The main goal of this stage is to prepare a properly formatted set of textual documents for analysis

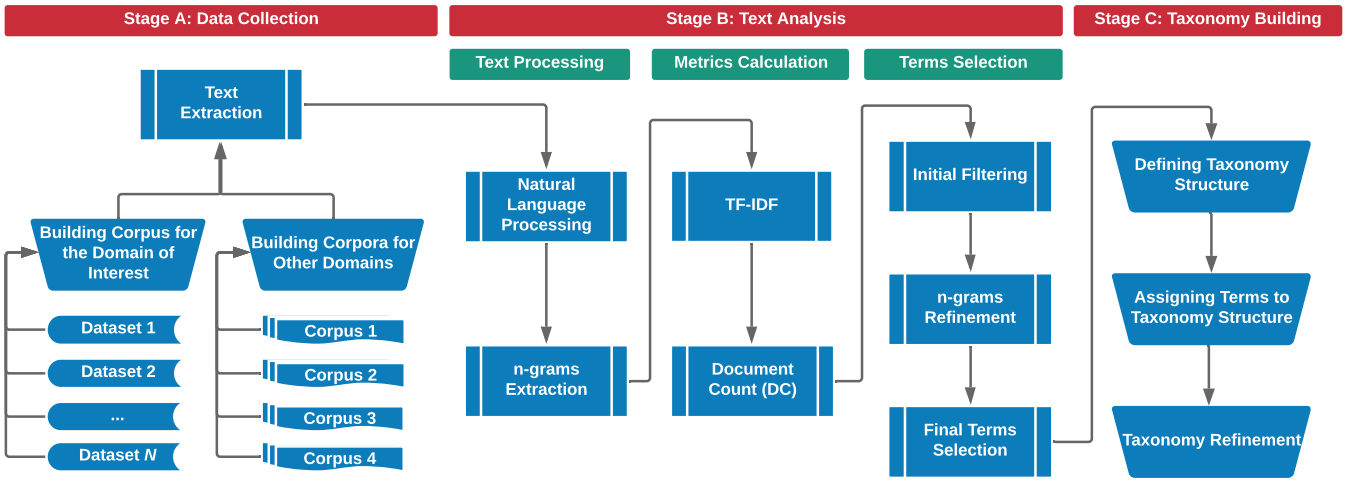

Figure 1: The proposed human-machine teaming based process for building taxonomies

in Stage B. Different datasets are needed here, so that the built taxonomy can cover more useful terms that can help separate the domain of interest (e.g., cyber security) from other domains. This stage will involve human effort to identify relevant documents from different domains and to aggregate all documents together, using automated tools, into a format ready for Stage B.

Stage B is for **text analysis**. The purpose of this stage is to produce a number of relevant terms to be assigned to a taxonomy structure defined in Stage C. Stage B consists of three steps. The first one is about processing the prepared corpora for generating n-grams[1]. The second step is about calculating some useful metrics including TF-IDF and the Document Count (DC) values, which will be explained later in the next three sections using cyber security as an example domain. The last step is about the selection of relevant terms from all n-grams based on the calculated metrics, which include three sub-steps: initial filtering, n-gram refinement, and final term selection based on TF-IDF ranking. The stage is largely automated by using NLP and IR tools, but the last step needs to involve the human analyst to determine some parameters empirically and to exclude irrelevant terms that should be excluded based on the human analyst's domain expert knowledge.

Stage C is for **building the taxonomy** from the relevant terms produced in Stage B. Stage C consists of the following steps. The first is defining the basic structure of the taxonomy especially the top-level and the second level classes and other components that are known before inspecting all the relevant terms. The definition of the basic structure can be done based on the human analyst's domain knowledge (independently of the relevant terms), but can be informed by what key concepts can be used to cover all those terms. The second step is assigning the relevant terms produced in Stage B to the defined taxonomy structure. In the final step, the whole taxonomy is refined based on any issues identified from the term assignment step, including necessary adjustments of the basic structure of re-assignments of some terms. This stage is mostly done by the human analyst, although some software tools could be developed to facilitate the term assignment (e.g., an automated recommendation system for suggesting where a term should be mapped to) and to create a visualization of the taxonomy.

The whole process can be repeated in full, or partially, to keep the created taxonomy up to date. For example, after the taxonomy is created, the text processing step in Stage B can be run again to produce more candidate n-grams to enrich the taxonomy, or some new documents can be added and be processed to update the taxonomy for reflecting new changes from relevant topics. Only new terms or out-dated terms need processing for such updates of the taxonomy, so its maintenance can be relatively light.

As a whole, the process combines the work of humans and machines well to co-create the taxonomy. A vital feature of the process is that the human analyst does not need to arbitrarily define a list of keywords, which is often the hardest, the most time-consuming, error-prone and "random" part of any taxonomy building process, but is able to work with a list of automatically produced terms. Such a human-machine teaming process does not only help reduce human effort but also increase the accuracy of the built taxonomy.

In the following three sections, we will use cyber security as an example domain to demonstrate how the proposed process was used by us to build a general cyber security taxonomy. Each section focuses on one stage of the process.

Note that the proposed methodology can be used for any domain or purpose, not just for the cyber security domain. This is because the approach is data-driven, i.e., the candidate terms are automatically harvested from a given dataset. Domain knowledge experts are still needed to select relevant documents and process the automatically produced candidate terms, but the most time-consuming and arbitrary part of the work – defining what terms to use – is largely automated. Therefore, by using a different dataset containing documents from a different domain, one can build the taxonomy for that domain.

[1] An n-gram is a sequence of n adjacent words. When n is 1, then it is called a uni-gram; when n is 2, it is called a bi-gram; and so on.

4 DATA COLLECTION

For building the cyber security taxonomy, we collected five datasets (corpora) of textual documents. The first dataset consists of cyber security related sources, while the others cover documents from four selected non-cyber security domains. The reason for collecting non-cyber security documents is to eliminate terms that are common in all domains, therefore not indicative for the cyber security domain. This will be accomplished mainly by evaluating the TF-IDF (Term Frequency-Inverse Document Frequency, explained later in this paper) scores of *n*-grams, in order to identify good candidate terms for the cyber security domain.

4.1 Building Cyber Security Dataset

Since our main objective is to build a general taxonomy for cyber security, we needed first to create a representative corpus for cyber security using human-written textual documents. For this purpose, different sources were collected in order to cover more diverse *n*-grams that are commonly used by different cyber security related people such as professionals, academics and hackers. The created textual corpus consisted of documents selected from the following four representative types of data sources.

1) **Professional reports** focusing on cyber security, issued by well-known organizations such as the ENISA (European Network and Information Security Agency, https://www.enisa.europa.eu/) and the UK's NCSC (National Cyber Security Centre, https://www.ncsc.gov.uk/). These reports correspond to cyber security professionals who are associated mostly with government and industry.

2) **Academic papers** written by cyber security researchers. Some papers in this category were collected by searching Google Scholar with some broad keywords such as "cyber security" and "information security". The others were cyber security related papers already known to the authors who are all cyber security researchers. This type of data corresponds to people mostly from academia.

3) **OSN data** (Twitter timelines) of cyber security related Twitter accounts, produced by a trained machine learning based classifier reported in [2]. Each timeline was a text file that resulted from merging all the tweets that were retrieved for a given Twitter account. The maximum number of tweets that could be collected per account was 3,200 as this was a limitation set by the Twitter API.

4) **Underground forum posts** that contain discussions taking place in some underground forums used by hackers and cyber criminals. For this data source, we used a database from the Cambridge Cybercrime Centre at the University of Cambridge [20]. This data source was used to gain insights into the terms that are usually used among cyber criminals.

The number of documents and the formats that were used in the cyber security corpus are listed in Table 1 including the numbers of tokens and words counts.

4.2 Building Non-Cyber Security Datasets

For non-cyber security documents, we used four other corpora in the following domains: news, law, (general) science and English literature. Each corpus is big enough to be considered as representative for its domain. See Table 2 for statistics about these textual corpora, from which we can see the different corpora have 2-10m words and their sizes are comparably rich.

Table 1: Statistics of the cyber security corpus

Data Source	Documents	Tokens	Words
Type 1	117	1,850,804	736,280
Type 2	385	5,465,389	2,001,726
Type 3	219	10,635,807	3,532,320
Type 4	69	10,554,781	3,448,526

Table 2: Statistics of all five corpora used

Corpus	Documents	Tokens	Words
Cyber Security	790	28,506,781	9,718,852
English Language	3321	24,581,859	7,544,295
Law	635	19,474,191	6,422,816
News	4561	8,536,780	3,346,196
Science	5149	5,553,724	2,282,068

4.3 Text Extraction

The processing pipeline started with reading the sources after which the texts were extracted. For each file type, we used a different parser. For example, the web pages parser removes all HTML tags (i.e., characters between < and >) and extract the remaining text. Moreover, the parser of PDF files removes the meta-data fields and links and extract the plain text. For the Twitter timeline of an account, the parser converted it into one text file by concatenating only the plain text of the tweets and removing all other data fields.

5 TEXT ANALYSIS

As mentioned previously, the text analysis stage consisted of the following steps, which are explained below.

5.1 Text Processing

In this step, and before applying the NLP tools, several preprocessing steps were applied to reduce the number of tokens for the later steps. We removed URLs, emails, independent numeric strings and punctuation symbols. Other strings related to Twitter data such as (retweet indicator "RT", @usernames and hashtag symbol "#") were removed as well.

5.1.1 Natural language processing. An NLP tool takes the raw text as an input and returns the annotated text as an output. We used several annotators from the Stanford CoreNLP library (https://stanfordnlp.github.io/CoreNLP/) to first tokenize the text, split tokens into sentences, assign part of speech (POS) tags to each token and then to apply lemmatization for each token to obtain the original root without any suffixes or prefixes. After that, we removed the stop words and applied a number rules to eliminate words that are less useful for our purpose, e.g., words that were too short, too long or non-English.

5.1.2 n-grams extraction. After the aforementioned NLP processing, we extracted *n*-grams where ($n = 1 \rightarrow 5$). Initially, only unigrams and bigrams were considered. However, we noticed that for this domain many valid terms are long *n*-grams (e.g., "Access

Control Policy", "Cyber Threat Information Sharing", "National Cyber Security Awareness Month"). Thus, trigrams, four-grams and five-grams were considered as well. This reflects how the proposed process can easily adapt to help refine the built taxonomy.

Statistics about all the corpora are presented in Table 3, which shows the number of n-grams of each size (1 to 5) for each corpus. The table clearly shows that longer terms are used more often in the cyber security domain than in others.

5.2 n-Grams Metrics Calculation

At the end of the last step, we ended up with over 6.7 million n-grams including 2.2 million for the cyber security corpus. They are too many to work with for the manual process in Stage C of the taxonomy building process, so we need to filter them down to a more manageable size. To this end, we calculated a number of metrics for each n-gram, which are then used to filter and select a smaller number of n-grams as valid terms.

5.2.1 Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF has been widely used for information retrieval tasks. It is defined based on a number of "documents", which correspond to "corpora" in our case. For a given word w and a particular "document" d of interest (the cyber security corpus in our case), it is defined as the product of the term frequency (TF), $\mathrm{TF}_{w,d}$ is defined as the count of w in the "document" d, and the inverse document frequency (IDF) defined as follows:

$$\mathrm{IDF}_w = \log \frac{N}{1 + N_w} \tag{1}$$

where N is the total number of "documents" (5 in our case), and N_w is the number of "documents" that cover the word w.

We used the TF-IDF scores to rank all the n-grams. The ranked list is used in the next step to filter and select top n-grams with higher TF-IDF values as candidates for building the taxonomy.

5.2.2 Document Count (DC). Some terms are assigned a relatively high TF-IDF value although they appear just in a very small number of documents in the cyber security corpus. Those highly document-specific terms are more likely unrelated to the cyber security domain, otherwise they should have been more widely used. To exclude such terms, we also calculated each n-gram's Document Count (DC), the number of documents in the corpus of interest (cyber security for our case) it appears at least once.

5.3 Terms Selection

In order to select a more manageable size of relevant terms for the taxonomy building stage, we followed the following three steps.

5.3.1 Initial filtering. First, we set three simple rules to reduce the number of candidate n-grams to around 62k. To be selected as a candidate, an n-gram must satisfy the following two conditions: 1) $\mathrm{IDF}_w \geq \log(5/3)$, meaning that the n-gram should have appeared in no more than 2 corpora (otherwise it is not unique enough for the cyber security domain); 2) $\mathrm{DC}_w \geq 5$, meaning that the n-gram appears in at least 5 different cyber security documents.

5.3.2 n-grams refinement. To help reduce irrelevant terms further, two automatic sub-processes were applied. Note that this refinement step can actually appear anywhere after n-gram extraction.

One reason is that part of the used cyber security corpus contains underground forum posts, which contain a lot of spam texts, links, advertisements for selling different products (especially medicines and drugs) and other less useful texts. To eliminate those "spam" n-grams, We followed a simple approach that proved very effective: we identified the most used names of products that usually appear after the word "buy" and created a blacklist for those names. Then we eliminated any n-gram containing at least one word from the blacklist. This removed more than 95% of those terms. We did not need an extensive advertisement terms removal mechanism as the remaining 5% had a lower TF-IDF value and did not appear among the top extracted n-grams.

Moreover, in order to eliminate redundant n-grams that are *completely* covered by other ones, we applied what we named as n-gram "coverage rules". By "covered", we mean that an n-gram t of size S is a sub "word sequence" of another n-gram t' of size $> S$ and the former appears only as part of the latter, e.g., if a unigram "formalizing" appears only when "formalizing security". Since we extracted n-grams from sizes one to five, there are four different types of coverage rules that if combined can cover all coverage cases: (1) a unigram covered by a bigram, (2) a bigram covered by a trigram, (3) a trigram covered by a four-gram, and (4) a four-gram covered by a five-gram. See Table 4 for some examples. When applied correctly, coverage rules can help reduce the number of candidate n-grams for future processing. However, in some cases an n-gram completely covered by another one may not be redundant as it can bear a broader semantic meaning than the latter, e.g., in a corpus "cyber" may accidentally appear only with 'cyber security", but "cyber" clearly should be kept as a standalone n-gram since it has a richer semantic meaning. To this end, the coverage rules should not be used alone, but its results can be always manually checked to avoid mistakes ("good" n-grams got wrongly eliminated).

5.3.3 Final terms selection. Third, for each n-gram size (1 to 5), we set an empirically determined threshold for TF and a size-specific threshold for DC to further eliminate some n-grams that do not appear frequently enough. Then, we ranked all remaining n-grams by their TF-IDF values, and then selected the top 1,000 unigrams, the top 1,500 bigrams, the top 1,000 trigams, the top 500 fourgrams and the top 500 fivegrams, which led to a set of 4,000 n-grams as candidate terms for further processing. The 4,000 candidate terms were then examined manually to remove irrelevant terms, correct wrongly extracted terms, and merge some terms.

6 TAXONOMY BUILDING

This is the third stage of the proposed taxonomy building process, which consists of the following steps.

6.1 Defining Taxonomy Structure

We needed to define a basic (not necessarily complete) structure for the cyber security taxonomy with a sufficient level of details to facilitate the term assignment in the next step. To this end, we studied existing cyber security taxonomies in order to get some insights into how we could design our taxonomy's initial structure.

We decided to choose ten top-level classes defined as below as a starting point. More details about those classes and associated subclasses can be seen in Figure 2 of the Appendix.

Table 3: Statistics of extracted *n*-grams for each corpus

Corpus	Unigrams	Bigrams	Trigrams	Fourgrams	Fivegrams	Total
Cyber Security	260,737	978,997	547,868	250,389	150,080	2,188,071
English Literature	280,380	1,119,025	470,754	143,686	46,662	2,060,507
Law	60,477	414,052	208,730	61,662	16,371	761,292
News	121,513	511,277	226,911	75,255	24,871	959,827
Science	95,618	350,101	180,699	56,999	16,950	700,367

Table 4: Examples of *n*-grams eliminated by the coverage rules

Case	Unigram	Bigram	Thrigram	Fourgram	Fivegram
A bigram covered by a trigram	default	windows			
A trigram covered by a fourgram	default	windows	kernel		
A fourgram covered by a fivegram	default	windows	kernel	debugging	
Accepted *n*-gram (fivegram)	default	windows	kernel	debugging	setting

"**Individual**": Human users are the main actors in the cyber security domain. Therefore, a more detailed view should be provided about the different roles that cyber security related people can have. Some of the subclasses under this class are: "End User", "Expert", "Academic", "Hacker", "Cybercriminal", "Activist" and "Journalist".

"**Party**": This is a main class added to represent different types of human gatherings and organisations that play a role in the cyber security domain. Some of the subclasses are: "Research Centre", "Educational Institute", "Government", "Critical Infrastructure", "NGO", "Business", and "Group". The "Business" subclass has more subclasses about different types of business entities, and the "Group" subclass was added to groups of people or organizations.

"**Event**": This is a main class covering all the activities and things that take place in the cyber security domain. An event can be attended by an "Individual" or any of its subclasses. Some of the subclasses under "Event" are: "Conference", "Expo", "Workshop", "Awareness Event" and "Training Event".

"**Vulnerability**": This main class covers vulnerabilities that can be exploited by attackers [13]. Three subclasses were created under this class: "Dataset", which refers to existing vulnerabilities databases such as CVE (Common Vulnerabilities and Exposures, https://cve.mitre.org/) and CWE (Common Weakness Enumeration, https://cwe.mitre.org/); "Software", which covers different categories of software vulnerabilities, e.g., "OS (Operating System)", "Application" or "Web Server", and finally "Hardware", which covers hardware related vulnerabilities.

"**Threat**": This main class is about "potential causes of an unwanted incident, which may result in harm to a system or organization" [13]. According to the nature of a "Threat", these subclasses were created: "Criminal", "Technical", "Business", "Legal", and "Other".

"**Attack**": This main class is about "attempts to destroy, expose, alter, disable, steal or gain unauthorized access to or make unauthorized use of an asset" [13]. This main class covers a wide range of cyber security attacks mapped to the following sub-classes: "Physical Attack", "Software Attack", "Network Attack", "Social Engineering", "Data Breach", "Unauthorized Access".

"**Technical**": This is a main class that covers technical concepts such as "Cryptography", "Protocol" and "Standard". Under "Cryptography" there are "Encryption" and "Hashing" subclasses. "Encryption" contains well-known encryption algorithms (e.g., "DES", "AES", "Diffie-Hellman" and "RSA"). The same applies to "Hashing", where several subclasses were added beneath it to represent hashing algorithms (e.g., "BSD", "MD5", "SHA-1", "SHA-256") and related concepts such as "salting" and "rainbow table".

Security "**Control**": This main class refers to any measure or course of actions that can be taken in order to reduce a risk. Control mechanisms contain processes, policies, devices, practices, and other actions that can alter risk [13]. Some of the "Control" subclasses are: "Firewall", "Access Control", "Standard", "Policy", "Regulation", "Training", "Detection", and "Sandbox". The "Policy" subclass contains more than 15 subclasses which represent the different kinds of policies such as data protection policies, access control policies and other general policies.

"**Risk**": This is a main class that covers concepts related to cyber risks. We defined a subclass named "Type" to reflect the nature of a risk, e.g. "Application", "Insider", "Internet", "IoT", "Privacy", "Technical", "Third-party". Usually, a risk has a numeric value to quantify it. Thus, we added a subclass named "Score". Additionally, we added a subclass "Operation" to cover all the common operations that usually associated with cyber risks, e.g., "Aggregation", "Assessment", "Identification", "Management", "Mitigation" and "Modeling".

"**Sub-domain**": This main class is for covering cyber security topics that can each have a sub-taxonomy such as "Cloud Security", "Mobile Security", "IoT Security", "Automotive Security", "Smart Grid Security", "Information Security Management", "Trust", "Cyber Insurance" and "Cyber Threat Intelligence".

6.2 Assigning Terms to Taxonomy Structure

Each term produced in the text analysis stage should be examined manually in order to assign it to the right class or subclass. In some cases, new subclass will be added to accommodate a term, which will involve refinement of the taxonomy structure (explained in the next subsection).

During this step, any different spellings and synonyms for the same term should be considered. Each term has a list of words that represent different spellings that a term can have. For example, the term "cyber security" has a list of the following words with the same meaning: "cyber security", "cybersecurity", "cyber-security". Another example is the "zero-day" attack. This term has the following words with the same meaning:"zero_day", "zeroday", "zero-day", "0day", "zer0day".

6.3 Taxonomy Refinement

While assigning terms to their potential classes and subclasses, changes of the taxonomy's basic structure may be necessary. A class or a subclass may need renaming or moving from one place to another to improve the semantic hierarchy of the proposed taxonomy. In addition, mergers of two or more subclasses or splitting a class or subclass could also happen. Such changes are normally done in an embedded manner as part of the term assignment step, so these two steps are often working in parallel.

The refinement process can also apply to the terms themselves because mapping terms to the taxonomy structure can change how the human analyst understand and organize all the terms. For instance, some terms may be discarded, and some new terms can be added to classes and subclasses with fewer children.

After following all the steps, we managed to build a general cyber security taxonomy. Since the cyber security taxonomy is based on a limited set of textual documents, it is actually not complete and more like a base-line subset of the full taxonomy. For instance, the automatically produced terms contain names of some cyber security experts, but many others are not included. Therefore, the built taxonomy should be further refined by using more textual documents, other existing taxonomies, and manually added nodes. The evolution of the domain also requires the taxonomy to be dynamically updated by re-running the process with new textual documents from time to time.

To some extent, the built (incomplete) taxonomy can be considered a good guideline to allow the human analyst to find ways to enrich the taxonomy further, e.g., after seeing a small number of terms for a specific concept (e.g., encryption algorithm, cyber security company, and cyber security expert, just to name a few), the human analyst can systematically look for more relevant terms and consider how to refine the taxonomy's structure and content.

7 FUTURE WORK

The work presented in this paper can be extended in a number of ways. The tree-based taxonomy is actually not enough to capture complicated concepts and relations between them. Thus, extending the taxonomy to a more complicated cyber security ontology will be needed to support more advanced analysis, such as automatic reasoning based on input data (e.g., when a particular cyber security event happens, what consequences it will generate and what defenses can be taken).

Second, we will work on enhancing the level of automation of the proposed process, to reduce the required human effort further. For instance, more advanced NLP and IR tools can be used to reduce the number of irrelevant and wrongly extracted terms, and an automated recommendation system can map relevant terms to the defined taxonomy structure. It is also possible to automate the collection of input textual documents at a larger scale.

Third, we want to consider the use of structural features to improve the *n*-grams ranking and terms selection. We will investigate the use of words' styles and positions to set weights for the extracted *n*-grams. For example, an *n*-gram appears in a document title, abstract section or section header should receive more importance than other *n*-grams in other parts of the document. Also, a word highlighted by a **bold** or *italic* style may be more important than other un-styled words.

Finally, with the aim of validating generalizability and consistency of the proposed taxonomy building method, the process can be conducted by independent data sources and experts, and the results are cross-validated and enhanced. We can also compare the taxonomy built using the proposed method with others built manually by experts without using an semi-automated approach. We however would like to point out that a real ground truth can hardly be established for quality checking since the taxonomy is *qualitative* and *subjective* by nature.

8 CONCLUSION

This paper presents a human-machine teaming based process to build taxonomies, starting from a given set of textual documents, followed by mostly automated processing done by NLP and IR tools, which produce a list of relevant terms to be assigned to a defined taxonomy structure. The key feature of the proposed process is a higher level of automation, which helps reduce human effort and make the selection of relevant terms more data driven (less subjective). The process also allows built taxonomies to be maintained more easily. An example is given to show how a general taxonomy was constructed using this process for the cyber security domain. The example taxonomy was built with a reasonable amount of human effort and a large number of candidate terms automatically collected from multiple data sources, which can be extremely time consuming and more error-prone if done by humans alone.

ACKNOWLEDGMENTS

Mohamad Imad Mahaini was supported by European Union's Horizon 2020 project NeCS (http://www.necs-project.eu/), under the Marie Skłodowska-Curie Grant Agreement No 675320.

Shujun Li's work was partly supported by the research projects ACCEPT (https://accept.cyber.kent.ac.uk/) and PriVELT (https://privelt.ac.uk/), funded by the EPSRC (Engineering and Physical Sciences Research Council) in the UK, under grant numbers EP/R033749/1 and EP/P011896/2, respectively.

The authors thank the Cambridge Cybercrime Centre (https://www.cambridgecybercrime.uk/) of the University of Cambridge, for granting us the access to their dataset on underground forums.

REFERENCES

[1] June Abbas. 2010. *Structures for Organizing Knowledge: Exploring Taxonomies, Ontologies, and Other Schema.* Neal-Schuman Publishers, Inc.

[2] Çağrı B. Aslan, Rahime Belen Sağlam, and Shujun Li. 2018. Automatic Detection of Cyber Security Related Accounts on Online Social Networks: Twitter As an Example. In *Proceedings of the 9th International Conference on Social Media and Society*. ACM, 236–240. https://doi.org/10.1145/3217804.3217919

[3] Eric W. Burger, Michael D. Goodman, Panos Kampanakis, and Kevin A. Zhu. 2014. Taxonomy Model for Cyber Threat Intelligence Information Exchange

Technologies. In *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security*. ACM, 51–60. https://doi.org/10.1145/2663876.2663883
[4] Gürol Canbek, Seref Sagiroglu, and Nazife Baykal. 2016. New Comprehensive Taxonomies on Mobile Security and Malware Analysis. *International Journal of Information Security* 5, 4 (2016), 106–138.
[5] Philipp Cimiano, Andreas Hotho, and Andreas Hotho. 2004. Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text. In *Proceedings of the 16th European Conference on Artificial Intelligence*. IOS Press, 435–439.
[6] Daniel Costa, Michael Albrethsen, Matthew Collins, Samuel Perl, George Silowash, and Derrick Spooner. 2016. *An Insider Threat Indicator Ontology*. Technical Report CMU/SEI-2016-TR-007. Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, USA. http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=454613
[7] R. Danyliw. 2016. The Incident Object Description Exchange Format Version 2. IETF RFC 7970. https://tools.ietf.org/html/rfc7970
[8] Development, Concepts and Doctrine Centre, Ministry of Defence, UK. 2018. Human-machine teaming. Joint Concept Note 1/18. https://www.gov.uk/government/publications/human-machine-teaming-jcn-118
[9] Golnaz Elahi, Eric Yu, and Nicola Zannone. 2009. A Modeling Ontology for Integrating Vulnerabilities into Security Requirements Conceptual Foundations. In *Conceptual Modeling - ER 2009: 28th International Conference on Conceptual Modeling, Gramado, Brazil, November 9-12, 2009. Proceedings*. Springer, 99–114. https://doi.org/10.1007/978-3-642-04840-1_10
[10] Sam Adam Elnagdy, Meikang Qiu, and Keke Gai. 2016. Understanding Taxonomy of Cyber Risks for Cybersecurity Insurance of Financial Industry in Cloud Computing. In *Proceedings of 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing*. IEEE, 295–300. https://doi.org/10.1109/CSCloud.2016.46
[11] Will Gibb. 2013. Back to Basics Series: OpenIOC. https://www.fireeye.com/blog/threat-research/2013/09/basics-series-openioc.html
[12] Ryan Heartfield, George Loukas, Sanja Budimir, Anatolij Bezemskij, Johnny R.J. Fontaine, Avgoustinos Filippoupolitis, and Etienne Roesch. 2018. A taxonomy of cyber-physical threats and impact in the smart home. *Computers & Security* 78 (2018), 398–428. https://doi.org/10.1016/j.cose.2018.07.011
[13] ISO/IEC. 2018. Information technology – Security techniques – Information security management systems – Overview and vocabulary. ISO/IEC 27000:2018. https://www.iso.org/standard/73906.html
[14] Yuning Jiang, Manfred Jeusfeld, Yacine Atif, Jianguo Ding, Christoffer Brax, and Eva Nero. 2018. A language and repository for cyber security of smart grids. In *Proceedings of 2018 IEEE 22nd International Enterprise Distributed Object Computing Conference*. IEEE, 164–170. https://doi.org/10.1109/EDOC.2018.00029
[15] George Loukas, Eirini Karapistoli, Emmanouil Panaousis, Panagiotis Sarigiannidis, Anatolij Bezemskij, and Tuan Vuong. 2019. A taxonomy and survey of cyber-physical intrusion detection approaches for vehicles. *Ad Hoc Networks* 84 (2019), 124–147. https://doi.org/10.1016/j.adhoc.2018.10.002
[16] H.A.M. Luiijf and A.H. Nieuwenhuijs. 2008. Extensible threat taxonomy for critical infrastructures. *International Journal of Critical Infrastructures* 4, 4 (2008), 409–417. https://doi.org/10.1504/IJCIS.2008.020159
[17] Vasileios Mavroeidis and Siri Bromander. 2017. Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In *Proceedings of 2017 European Intelligence and Security Informatics Conference*. IEEE, 91–98. https://doi.org/10.1109/EISIC.2017.20
[18] New Idea Engineering, Inc. 2018. What's the difference between Taxonomies and Ontologies? - Ask Dr. Search. http://www.ideaeng.com/taxonomies-ontologies-0602
[19] OASIS Open. 2019. Introduction to STIX. https://oasis-open.github.io/cti-documentation/stix/intro
[20] Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. 2018. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1845–1854. https://doi.org/10.1145/3178876.3186178
[21] Uta Priss. 2006. Formal Concept Analysis in Information Science. *Annual Review of Information Science and Technology* 40, 1 (2006), 521–543. https://doi.org/10.1002/aris.1440400120
[22] Pedram Radmand, Alex Talevski, Stig Petersen, and Simon Carlsen. 2010. Taxonomy of Wireless Sensor Network Cyber Security Attacks in the Oil and Gas Industries. In *Proceedings of 2010 24th IEEE International Conference on Advanced Information Networking and Applications*. IEEE, 949–957. https://doi.org/10.1109/AINA.2010.175
[23] Abdul Razzaq, Khalid Latif, H. Farooq Ahmad, Ali Hur, Zahid Anwar, and Peter Charles Bloodsworth. 2014. Semantic security against web application attacks. *Information Sciences* 254 (2014), 19–38. https://doi.org/10.1016/j.ins.2013.08.007
[24] Hichem Sedjelmaci and Sidi Mohamed Senouci. 2018. Cyber security methods for aerial vehicle networks: taxonomy, challenges and solution. *Journal of Supercomputing* 74, 10 (2018), 4928–4944. https://doi.org/10.1007/s11227-018-2287-8
[25] Ju An Wang and Minzhe Guo. 2009. OVM: An ontology for vulnerability management. In *Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research*. ACM, Article 34.
[26] Hui Yang and Jamie Callan. 2009. A Metric-based Framework for Automatic Taxonomy Induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*. Association for Computational Linguistics, 271–279.
[27] Andrei C. Zamfira and Horia Ciocarlie. 2018. Developing An Ontology Of Cyber-Operations In Networks Of Computers. In *Proceedings of 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing*. IEEE, 395–400. https://doi.org/10.1109/ICCP.2018.8516644

A CYBER SECURITY TAXONOMY: VISUALIZATION

An overview of the high-level structure of the initial cyber security taxonomy is shown in Figure 2. The diagram contains over 170 objects showing the main classes and subclasses of the taxonomy, before all terms are allocated. The high-level structure was refined while terms are allocated, and future adaptation is expected in future. This visualization provides a simple way to quickly understand the hierarchy of the proposed taxonomy.

The root node of the taxonomy is "Cyber Security", which is connected to all top-level main classes. We can distinguish a main class by node size and the bold style of its internal text. Each main class is connected to its subclasses and each subclass might be connected to other subclasses beneath it and so on.

We will maintain the cyber security taxonomy at a web page (https://cyber.kent.ac.uk/research/cyber_taxonomy/). Our aim is to provide both machine readable files and an interactive visualization of the most stable version of the complete taxonomy so that people can use it right away. We welcome other cyber security researchers and experts to help co-develop the taxonomy further.

B CYBER SECURITY TAXONOMY: POSSIBLE APPLICATIONS

For the cyber security taxonomy presented in this paper, we explain some applications where such a taxonomy can be used.

This taxonomy can be used to capture cyber security related discussions on OSNs. This can be achieved by analyzing OSN feeds like tweets to determine if those tweets are related to the cyber security domain and then identify the topics and concepts that are discussed. This also can help in building monitoring applications for OSNs with security purposes e.g. monitoring the spread of a malware on the Internet and its impact on people and organizations.

We can analyze the timeline of a Twitter account to determine if the author is related to the cyber security domain. Also, we can determine which cyber security related concepts (s)he is interested in. Such analysis can help understand human behavior on OSNs for security purposes, e.g., cyber security awareness campaigns.

This taxonomy can be used to select a set of keywords as features for a machine learning classifier to automatically classify cyber security related people into different classes, which can help provide useful information about cyber security activities, such as impeding or fresh attacks and first responses.

The taxonomy can further be used to analyze cyber security related textual sources in a semantic way, leading to better systematic analysis for such sources. One interesting application would be connecting such semantic analysis with eye-tracking data to understand how human users understand cyber security related documents such as privacy policies and security warnings.

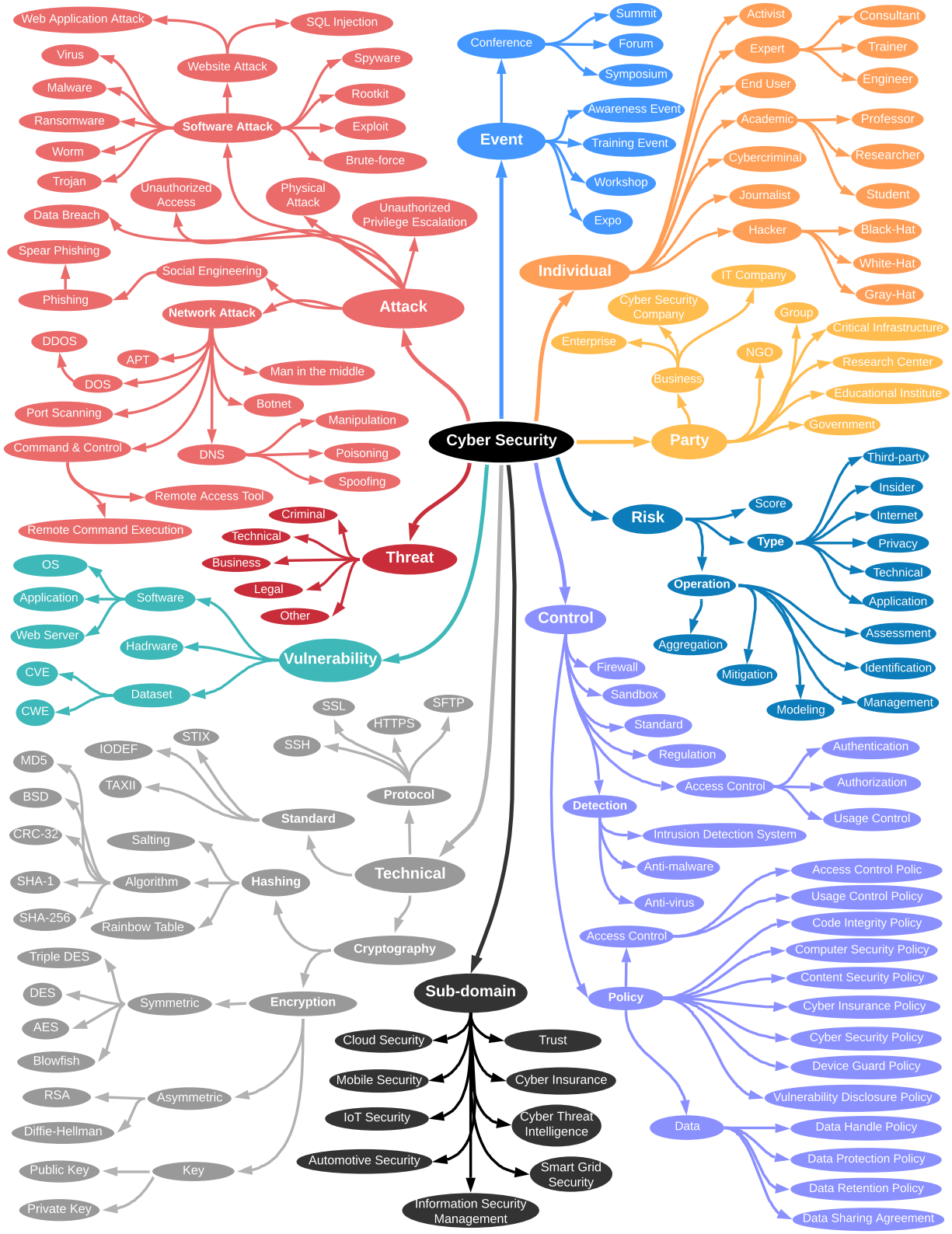

Figure 2: A visualization of the high-level structure of the initial cyber security taxonomy, showing the main classes and subclasses created before the term allocation phase. Note that node colors were used for illustration purposes only.