9 Big Data Analytics and Text Mining in Internet Governance Research: Computational Analysis of Transcripts from 12 Years of the Internet Governance Forum

Derrick L. Cogburn

Since the late 1990s, Internet governance has been a critical topic for multidisciplinary academic research (Bygrave and Bing 2009; Cogburn 2003, 2005; Cogburn et al. 2005; DeNardis 2009; Goldsmith 2007; Mueller 2009, 2010; Paré 2003; Thierer and Crews 2003). One simple measure of the increasing interest in Internet governance research is Google Trends statistics. In figure 9.1 we see November 2005 as the height of popularity for web searches of "Internet governance." This was the final month leading up to the second phase of the World Summit on the Information Society (WSIS), held November 16–18, 2005, in Tunis, after which the United Nations Internet Governance Forum (IGF) was launched (Cogburn 2017). Internet governance was certainly important before WSIS, but this global meeting helped accelerate a broad, multistakeholder focus on the issues beyond the narrow technical and academic focus that had dominated the study of Internet governance since the mid-1990s (Cogburn 2017). In chapter 3, Mueller and Badiei highlight these same trends.

In some ways research on the narrower Internet governance domain of cybersecurity has eclipsed the broader study of Internet governance. During the same time period, 2004–2017, there was a steady rise in searches for the term "cybersecurity," with a relatively sharp increase starting in 2013 (figure 9.2). Many researchers turned to a focus on better understanding the political and strategic implications of decisions made by infrastructure providers (Musiani et al. 2016). This rise is correlated with a steady increase in actual cybersecurity attacks, including the March 2007 hack of TJX, parent company of T. J. Maxx (Pepitone 2014), the June 2010 Stuxnet attack against Iran's nuclear centrifuges (Kushner 2013), the November 2013 Target (Kassner 2015), September 2014 Home Depot (Vinton 2014), and the April and June

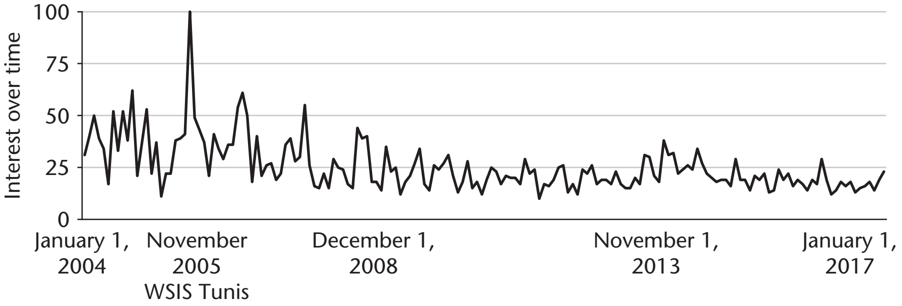


Figure 9.1
Google Trends searches for "Internet governance," 2004–2017.

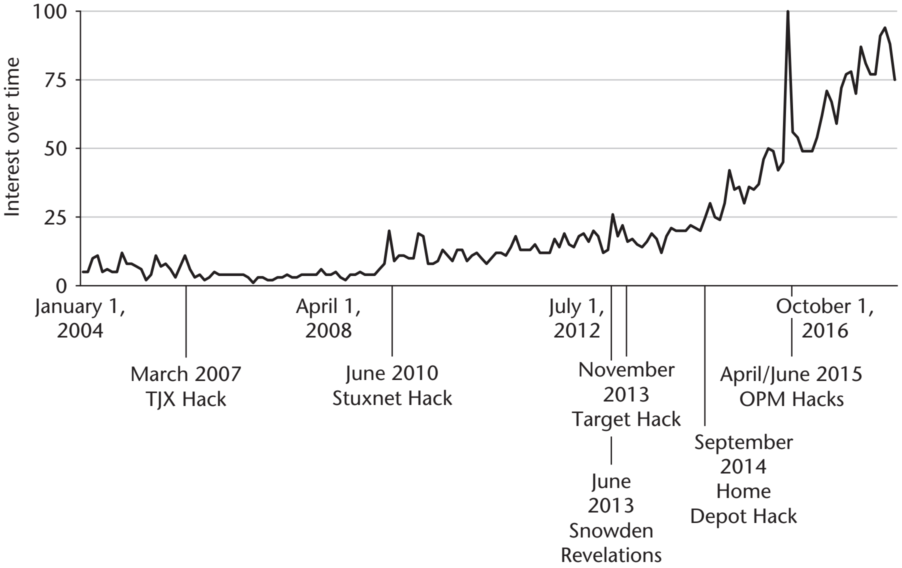


Figure 9.2
Google Trends searches for "cybersecurity," 2004–2017.

2015 Office of Personnel Management data breaches (Koerner 2016). This rise is also correlated with the revelations by Edward Snowden of US government widespread surveillance of the Internet, which began in June 2013 (Greenwald 2013). It also corresponded with the launch of the 2013 Barack Obama administration executive order on cybersecurity (Executive Order No. 13,636, 2013) and the subsequent National Institute of Standards and Technology (NIST 2014) Framework for Improving Critical Infrastructure Cybersecurity.

During the same period, the terms "big data" and "analytics" also became much more widespread. Between 2004 and 2012, Google Trends indicates search for the term "big data analytics" was relatively flat, with an average popularity score of 5. But then it exploded, jumping to 15 in January 2012 and to a high of 100 in March 2017. Figure 9.3 illustrates this trend. Schneider (2016) highlights some of the potential reasons for these trends, such as the increasing use of technologies that produce digital data, including unstructured text, and the corresponding increase in computational power available to analyze them.

Some describe big data, the Gartner Group being the first, by its 3Vs, or the volume, velocity, and variety of data that are available today (Laney 2001). Some scholars add veracity, variability, and value to those for understanding the concept of big data.

A particularly interesting type of data is underutilized: unstructured textual data. By unstructured, we mean text that has not yet been tagged, coded, or organized in some predetermined data model. Schneider (2016) estimates that up to 80 percent of the world's available data is unstructured text. The growing digital production and digitization of text adds significant amounts of textual data for analysis. This includes text on websites and in blog posts, speeches, meeting transcripts, email archives, reports, published articles, and especially social media (e.g., Twitter, Facebook, RSS feeds). Twitter is a particularly interesting source of unstructured textual data, making "nearly all of its data available via APIs [application programming interfaces] that enables [*sic*] realtime programmatic access to its massive seven-year archive" (Leetaru et al. 2013), and "Twitter users produce

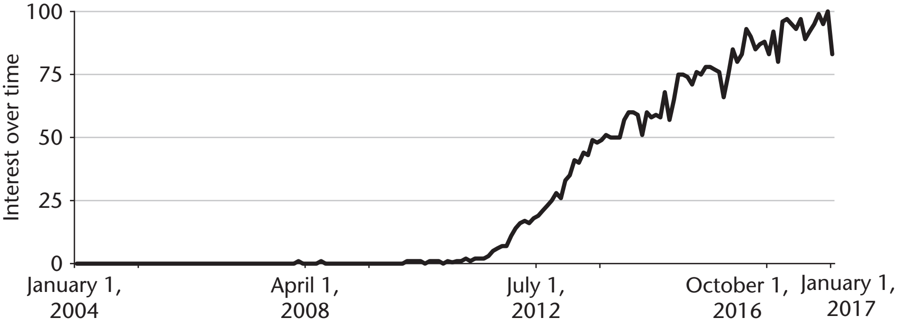


Figure 9.3
Google Trends searches for "big data analytics," 2004–2017.

8 billion words—every single day" (Kaisler et al. 2014). Of course, each of these data sources has "structure," but it is considered unstructured data because it has not yet been organized into a predetermined data model (e.g., into a spreadsheet with columns and rows).

In addition, each of these genres of data has its own characteristics that can be harnessed to augment analysis (Lee and Myaeng 2002). However, these many text-based datasets are not very large in terms of file size. One gigabyte of storage, an extremely small and common amount carried around by most students and faculty these days, can contain over 894,784 pages of plain text. A terabyte of data can contain 916,259,689 pages of plain text. It would take most humans an inordinately long time to read that many pages, but the data could be easily carried around on the average external drive. The big data of most social scientists or economists, however, especially those exploiting text as their data source, are not at the scale of the big data of earthquake engineering and upper atmospheric research. Thus, the "big" in "big data" is a relative term, with its meaning changing according to domain and the infrastructure available to the researcher. There is no singularly accepted definition of big data.

This increasing availability of data is coupled with a corresponding increase in computational tools, storage, and big data analytics and text mining software, including both commercial and open source options. The combination of this infrastructure and the available data allows us to combine insights from both quantitative and qualitative big data, but especially from unstructured textual data sources.

Purpose

This chapter, first, examines the substantive issues related to one of the major global institutional venues for debating issues related to Internet governance, specifically the annual UN IGF. Thus, this chapter focuses on identifying core themes and key issues discussed over the 12-year history of the IGF, and understanding which issues have remained constant, which have changed, and when they emerged or changed. Given the increasing importance of cybersecurity research in Internet governance, this chapter also explores the extent to which cybersecurity issues were debated at the IGF and the relationship between the NIST cybersecurity framework introduced by the Obama administration and related IGF debates. Second, the chapter demonstrates the potential of big data analytics and text mining techniques in Internet governance research.

These inductive and deductive text mining techniques are powerful tools to exploit the voluminous textual data available to Internet governance researchers and are also extremely useful for current research interests related to computational propaganda, or "fake news," and state- and nonstate-actor influences on national electoral systems through social media and other communication platforms. In chapter 7, Jardine highlights the continued growth of the Internet, including new users and content they create. He also argues for the suitability of quantitative methods for subsections of Internet governance research and particularly for cybersecurity research. The text mining approaches discussed in this chapter provide a nice complement to Jardine's arguments, by taking a quantitative analytical approach to analyzing text-based data, which are inherently qualitative. My use of a categorization model (dictionary) to measure the extent to which the NIST cybersecurity framework was present in the IGF debates reinforces his claims. In addition, Mueller and Badiei in chapter 3 discuss the widespread availability of materials on specific Internet governance institutions, such as the IGF, Internet Engineering Task Force, Internet Corporation for Assigned Names and Numbers, and others. There are many, many more, including the email archives, websites, and policy papers of transnational nongovernmental organization networks and of groups such as the Internet Assigned Numbers Authority Transition Committee. This chapter illuminates the reasons for analyzing those materials and making much better use of the voluminous resources available from these institutions for Internet governance research.

Conceptual Framework: Approaches to Text Mining

Text can contain substantial meaning and value to researchers. There are two important dimensions to text: semantics and syntax. *Semantics* refers to the meaning of words within their surrounding framework. *Syntax* is the structure of language, how individual words are arranged to make well-formed sentences and paragraphs. For decades, qualitative researchers have analyzed texts, doing deep and careful reading of relevant documents. As these qualitative research projects grew in size and complexity, computer-assisted qualitative data analysis software (CAQDAS) was developed to help facilitate this process. While extremely helpful, these CAQDAS tools still require researchers to closely read all documents and add codes to the text, developed a priori or in vivo while reading the documents.

The field of text mining is highly interdisciplinary and encompasses multiple theoretical approaches and methods with one common element—unstructured text as input information. Text mining has been aided by the widespread availability of machine-readable text. However, advances in the field of text mining, aided by concurrent increases in computational power and storage, have now accelerated the potential to use these techniques across a range of fields (Schneider 2016). With these tools, researchers can take unstructured text and transform it into a structured numerical format, based on term frequencies, and subsequently apply standard data mining techniques, finally unlocking the vast amount of valuable information in texts.

Many techniques are available to exploit the power and potential of big data analytics and text mining in specific research projects, including text classification, text clustering, ontology and taxonomy creation, document summarization, and latent corpus analysis. In general, there are two philosophical approaches to text mining, statistical and natural language processing. The statistical approach to text mining takes the "bag of words" route. It assumes there is value in the words themselves and does not require the analysts to understand the syntax of the words. In contrast, the natural language processing approach first tags parts of speech and then considers word and sentence structure. This study takes a statistical text mining approach, while recognizing the value of the natural language process approach.

Statistical text mining has two broad divisions—inductive and deductive, each with its own methodologies and techniques. The inductive approach asks broad exploratory questions about a large-scale text-based dataset, without specific a priori goals. For example, we can ask what key words and phrases characterize a dataset and determine what topics, themes, and trends exist. We can identify named entities within the dataset, including countries, people, organizations, and acronyms. Cross-tabulation determines how each element changes in relation to other key variables, such as date, region, and organizational type.

The deductive approach, in contrast, is confirmatory, allowing us to ask specific research questions of the data and to even test hypotheses. We can build, adopt, or adapt dictionaries (categorization models) to help us explore specific concepts in the dataset and to determine the degree of their presence or absence (Bengston and Xu 1995; Deng, Hine, and Sur 2017; Rousu et al. 2005). Variants of these models allow sentiment analysis, to characterize positive and negative sentiment, or polarity, within the dataset (Liu and Zhang 2012). Further, we can use supervised machine learning to develop classification models that allow us to predict text with a high degree of accuracy (Rousu et al. 2005). Through the use of these inductive and deductive techniques, we can begin to illustrate the tremendous potential of computational text mining for Internet governance and cybersecurity research.

Case Study: UN IGF

The WSIS action lines adopted at the end of the 2003 WSIS included promoting the continued development of the Internet with its potential impact on all aspects of the world (Cogburn 2017; International Telecommunication Union, n.d.). The action lines included four key references to Internet and Internet governance, and the 2005 WSIS Tunis Agenda mentions the Internet 80 times and Internet governance 30 times. At the conclusion of WSIS Tunis, participants adopted the Tunis Agenda, which in addition to coordinated implementation of the WSIS action lines, included a commitment to establish and support the UN IGF. The IGF was given an initial 5-year mandate and was subsequently approved for another 10 years.

The first IGF was held in Athens, Greece, in 2006, immediately after the conclusion of WSIS 2005 in Tunisia (Bygrave and Bing 2009). As of December 2017, there have been 12 IGFs, the last in Geneva. Going back to the first IGF in 2006, transcripts of sessions have been made available to the public on the IGF website (http://intgovforum.org), and over time, this process has increased and become more comprehensive. For example, in 2006, only 11 transcripts were made available, while in 2017, as many as 215 transcripts were made available (table 9.1).

Table 9.1
Number of IGF transcripts by year.

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
11	14	14	15	114	61	8	63	138	162	205	215	1,020

Research Questions

To demonstrate the computational text mining techniques described above, this case study of IGF transcripts asks four broad research questions, two inductive and two deductive.

RQ1. What are the key themes, topics, and entities discussed at IGF over its lifetime?

RQ2. Which key issues have remained consistent at IGF, and which ones have changed?

RQ3. In what ways is the Internet of Things (IoT) represented at IGF?

RQ4. To what extent is the 2014 NIST cybersecurity framework represented at IGF?

Methodology

This study is organized using the cross-industry standard process for data mining (CRISP-DM) for text mining.[1] Since text mining is still a relatively new and somewhat unstandardized field, the CRISP-DM approach can provide a well-understood, documented, and somewhat standardized process for executing and managing complex text mining projects. The CRISP-DM for text mining has six stages through which each text mining project must proceed (figure 9.4). In Stage 1, the researcher determines the purpose of the text mining study: what the researcher wants to accomplish and the problem or opportunity identified by the researcher. In Stage 2 the researcher explores the availability and nature of the unstructured textual data to exploit. The researcher has to determine if the data are available, in what format they are stored, and in what quantity. Stage 3 focuses on preparing data, which could include data cleaning, preprocessing, applying stopwords (exclusion lists that remove common, insignificant words), and further data reduction techniques of stemming and lemmatization. In Stage 4, the researcher develops the models and specific techniques to analyze the data. In Stage 5 the researcher evaluates the results of the analysis. In Stage 6 the researcher deploys the results, in the form of recommendations and presentations. At any point along the way, the researcher may decide to go back to a previous stage or all the way to the beginning.

Using an automated commercial tool called SiteSucker,[2] the researcher collected all the publicly available IGF transcripts. This data collection

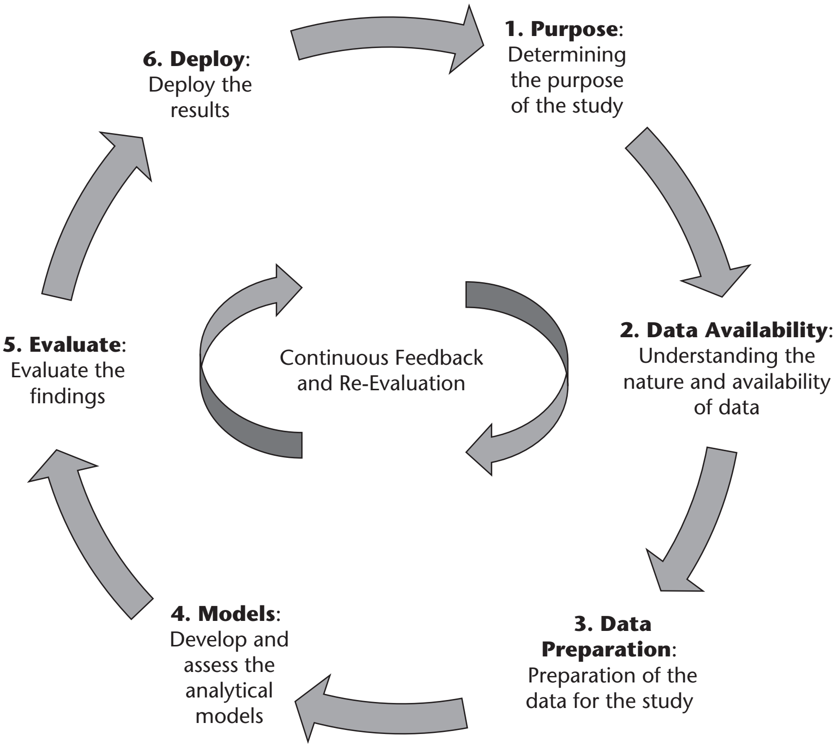

1. Purpose: Determining the purpose of the study
2. Data Availability: Understanding the nature and availability of data
3. Data Preparation: Preparation of the data for the study
4. Models: Develop and assess the analytical models
5. Evaluate: Evaluate the findings
6. Deploy: Deploy the results
Continuous Feedback and Re-Evaluation

Figure 9.4
The CRISP-DM approach.

yielded 1,020 documents (made up of .txt, .html, .doc, and .pdf formats), in a file of 109.9 megabytes. The data collection represents the available transcripts of main sessions and workshops where available.

Once the data were collected, the researcher used a commercial software tool called Provalis ProSuite to organize the project and conduct the text mining.[3] There are open source software tools that can also perform this type of analysis, such as the R programming language and related packages. Most R programmers use RStudio, an open source integrated development environment (IDE), which makes it easy to install and use free and open source text mining packages such as tm, Rvest (for data collection), and tidytext. The first step is to build the corpus, which includes converting the textual data into numerical data, based on the word frequencies across documents. This corpus, containing all the available IGF transcripts, is constructed primarily as a document-term-matrix (DTM), which means each document is represented as a row in the matrix, while each term (word) in each document is represented in the columns of the matrix, with a numerical value for how many times that term occurs in the document. Upon import, I used the file structure (organized by date of the IGF) to automatically create a Date variable for filtering the dataset by date and conducting a longitudinal analysis. I also used this Date variable for cross-tabulation analysis. Then, I preprocessed the data, applying a typical English-language stopword dictionary (or exclusion list) to remove frequent words that add little value to the analysis (such as "a," "and," "the"). The exclusion list may be modified for a specific dataset (e.g., words deemed important to include or remove in the analysis). I did not apply stemming or lemmatization, which preprocesses a textual dataset to reduce its overall size.

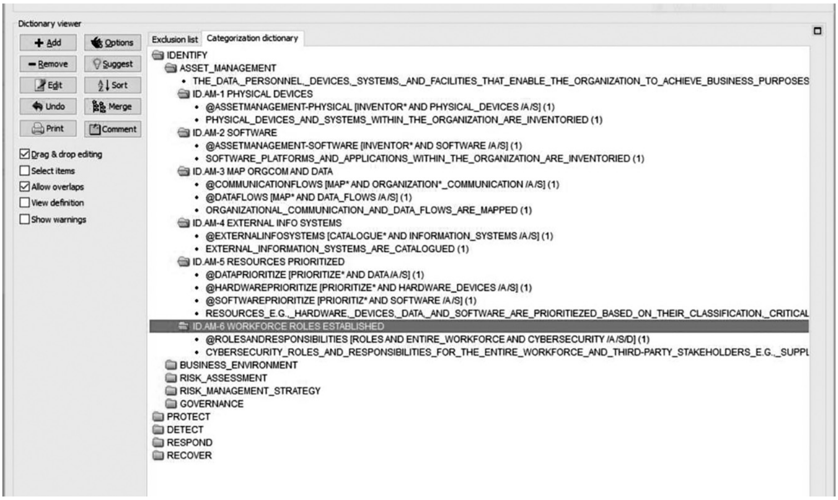
To answer the first two of the four research questions, I began with an inductive approach, a count-based evaluation, that focuses on term and document frequency, followed by phrase frequency. This is a common approach, and is one of the simplest techniques for text mining, similar to basic descriptive analysis of project variables in a statistical study. A word or phrase, an n-gram, occurring frequently in a dataset, with some important limitations discussed later, is considered important. In this analysis, I used the term frequency by inverse document frequency (TFxIDF) technique: if a word appears frequently in a document, it is important; but if it appears in many documents, it is less important. This is a common text mining heuristic to identify important words and phrases in a corpus.

Next, I used an inductive technique called topic modeling, which essentially does exploratory factor analysis on the underlying numerical representation of the IGF transcripts to identify factors, which are interpreted as topics. However, unlike factor analysis, since the dataset is based on text, the software provides a textual suggestion of what the topic seems to represent. I applied topic modeling on the entire dataset and separately for each of the 12 years. In addition, I identified key organizations, countries, acronyms, and people across the entire dataset, and again for each year, using a named-entity extraction tool.

To answer the third and fourth research questions, I took a deductive approach: hierarchical cluster analysis and categorization modeling (also known as dictionary development) to answer the third question on the IoT, and categorization modeling to determine the extent to which the 2014 NIST cybersecurity framework is included in IGF discussions for the fourth.

Hierarchical cluster analysis allows examination of the entire dataset for co-occurrences and to assess the themes or topics represented by specific clusters, such as a cluster that appears to represent the IoT.

Categorization modeling is an explicitly deductive technique. Essentially, it requires the researcher to develop a semantic representation of the concept of interest, and then apply that dictionary to the corpus to determine the extent to which that concept is present or absent in the text. Dictionary development generally starts with the broadest categories within the concept (for example, in sentiment analysis, these top categories tend to be the binary categories of positive and negative). Then, those broadest categories can be further divided into broad subcategories. Once the lowest levels of categories and subcategories have been determined, specific words, phrases, and rules (which allow you to formulate criteria for text that includes negations and specifications for proximity of words and phrases) can be developed. When any of these elements occur, they accrue to the subcategory, which in turn aggregates up to its category. This is a very powerful technique to identify the extent to which a specific concept the researcher is interested in exploring is either present or absent in the dataset. In this study, I developed a categorization model (figure 9.5) starting from the 2014 NIST cybersecurity framework core spreadsheet.[4] This framework has five primary categories (identify, protect, detect, respond,



☑ Drag & drop editing
□ Select items
☑ Allow overlaps
□ View definition
□ Show warnings

```
IDENTIFY
  ASSET_MANAGEMENT
    • THE_DATA_PERSONNEL_DEVICES_SYSTEMS_AND_FACILITIES_THAT_ENABLE_THE_ORGANIZATION_TO_ACHIEVE_BUSINESS_PURPOSES
    ID.AM-1 PHYSICAL DEVICES
      • @ASSETMANAGEMENT-PHYSICAL [INVENTOR* AND PHYSICAL_DEVICES /A /S] (1)
      • PHYSICAL_DEVICES_AND_SYSTEMS_WITHIN_THE_ORGANIZATION_ARE_INVENTORIED (1)
    ID.AM-2 SOFTWARE
      • @ASSETMANAGEMENT-SOFTWARE [INVENTOR* AND SOFTWARE /A /S] (1)
      • SOFTWARE_PLATFORMS_AND_APPLICATIONS_WITHIN_THE_ORGANIZATION_ARE_INVENTORIED (1)
    ID.AM-3 MAP ORGCOM AND DATA
      • @COMMUNICATIONFLOWS [MAP* AND ORGANIZATION*_COMMUNICATION /A /S] (1)
      • @DATAFLOWS [MAP* AND DATA_FLOWS /A /S] (1)
      • ORGANIZATIONAL_COMMUNICATION_AND_DATA_FLOWS_ARE_MAPPED (1)
    ID.AM-4 EXTERNAL INFO SYSTEMS
      • @EXTERNALINFOSYSTEMS [CATALOGUE* AND INFORMATION_SYSTEMS /A /S] (1)
      • EXTERNAL_INFORMATION_SYSTEMS_ARE_CATALOGUED (1)
    ID.AM-5 RESOURCES PRIORITIZED
      • @DATAPRIORITIZE [PRIORITIZE* AND DATA /A /S] (1)
      • @HARDWAREPRIORITIZE [PRIORITIZE* AND HARDWARE_DEVICES /A /S] (1)
      • @SOFTWAREPRIORITIZE [PRIORITIZ* AND SOFTWARE /A /S] (1)
      • RESOURCES_E.G._HARDWARE_DEVICES_DATA_AND_SOFTWARE_ARE_PRIORITIEZED_BASED_ON_THEIR_CLASSIFICATION_CRITICAL
    ID.AM-6 WORKFORCE ROLES ESTABLISHED
      • @ROLESANDRESPONSIBILITIES [ROLES AND ENTIRE_WORKFORCE AND CYBERSECURITY /A /S/D] (1)
      • CYBERSECURITY_ROLES_AND_RESPONSIBILITIES_FOR_THE_ENTIRE_WORKFORCE_AND_THIRD-PARTY_STAKEHOLDERS_E.G._SUPPL
  BUSINESS_ENVIRONMENT
  RISK_ASSESSMENT
  RISK_MANAGEMENT_STRATEGY
  GOVERNANCE
PROTECT
DETECT
RESPOND
RECOVER
```

Figure 9.5
Overview of 2014 NIST cybersecurity categorization model.

and recover) and within each category are multiple subcategories and sub-subcategories. All these elements are captured in my categorization model. I deployed these categorization models across the entire 12-year period.

Similarly, I could have built another categorization model representing the EU Cybersecurity Framework and compared the degree to which each framework was represented in the dataset. Or I could have explored the dataset to assess the degree to which the priorities of one stakeholder—say, the private sector, represented by the group Business Action in Support of the Information Society (BASIS) and supported by the International Chamber of Commerce (ICC)—were represented in the dataset relative to, say, the statements of the civil society Internet Governance Caucus (IGC). I also could have used supervised machine learning to build a classifier to distinguish between the content of each stakeholder group and then deployed that classifier to assess which stakeholder group had the most influence in the IGF processes. It would be a little tricky to do this in the IGF context, because there are no concrete outcome documents of an IGF, but I used this technique to great effect in an analysis of stakeholder contributions to the NetMundial conference (Cogburn 2014).

Research Limitations and Challenges in Text Mining Internet Governance

As far as I know, the corpus of 1,020 IGF transcripts makes this the largest study to date of these important data, and these techniques have proved to be extremely valuable in studying Internet governance. However, this only scratches the surface, and this approach has important limitations. First, although the data for this study—transcripts from 12 years of the IGF—are tremendously revelatory, they do not cover all the workshops and side events associated with an IGF meeting, and they capture only formal statements from IGF sessions. This focus on what Goffman (1959) called “front stage” behavior ignores his argument of the importance “back stage” behavior can have on policy debates. Of course, much of the work of the IGF is accomplished outside the formal conference structure. Backstage behavior occurs during the coffee breaks, lunches, dinners, and the many receptions and parties associated with an IGF. Analyzing only the formal language is a major limitation of this approach. Also, there is the limitation of this particular dataset of who said what. It would be tremendously valuable to have each statement attributed to a particular actor, who could then be coded in a variety of ways (e.g., by multistakeholder sector or by regional, ethnic, or gender demographic). However, I would have had to be present for each session and have conducted more traditional participant-observation research or interviews or focus groups. Finally, most text analytics techniques wrestle with understanding syntactic meaning. Sarcasm, euphemism, and double entendre, all common in human language, continue to elude many of these computational approaches. Nonetheless, while limitations of these text mining techniques have no silver bullet, this study demonstrates some of their tremendous value in Internet governance research.

Findings

To answer the first research question, What are the key themes, topics, and entities discussed at IGF over its lifetime?, I applied TFxIDF to explore, first, keywords and then phrases across all 12 years. Figure 9.6 shows the top ten themes at IGF represented by keyword frequency, and figure 9.7 shows the

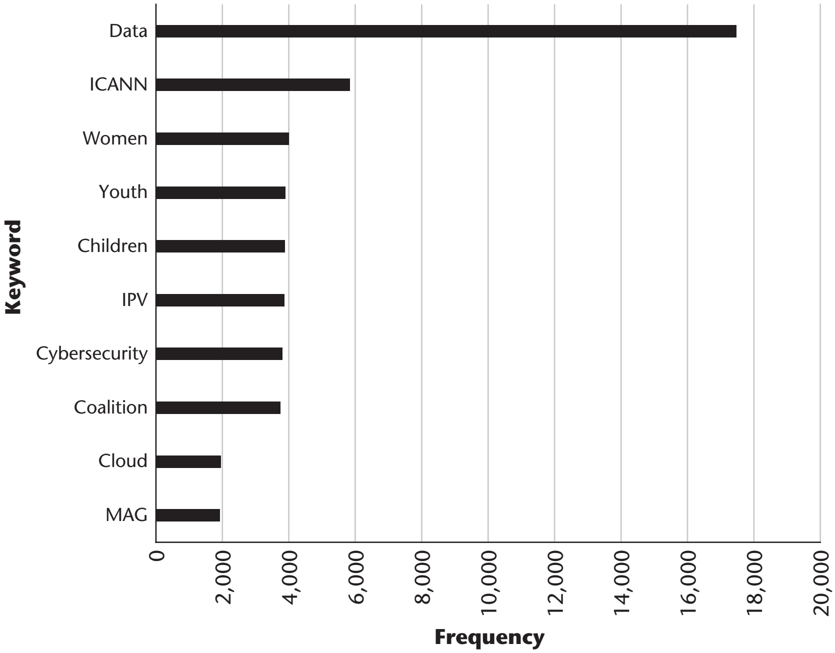


Figure 9.6
Key themes across 12 years of IGF represented by keywords.

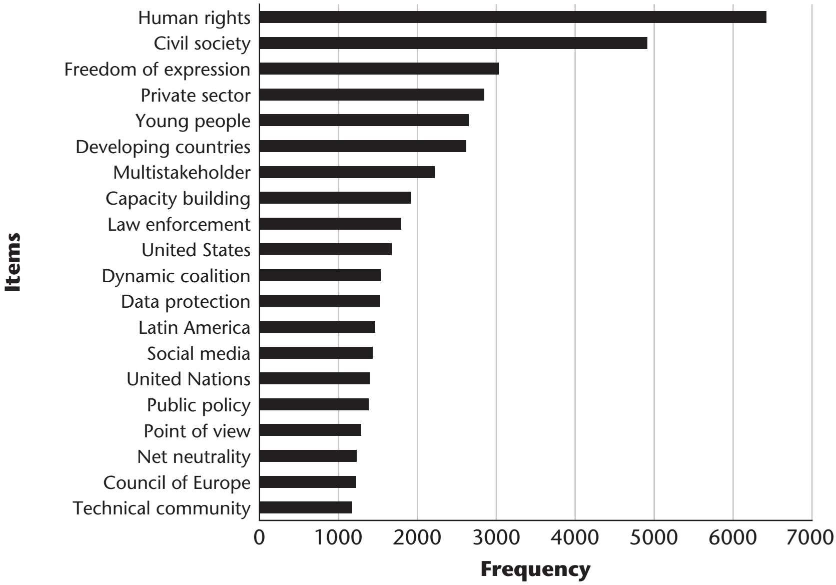


Figure 9.7

Key themes across 12 years of IGF represented by phrases.

top 20 themes of 12 years of IGF represented by phrase frequency. Then I used named-entity extraction techniques to identify all the people referenced in the dataset across 12 years. Figure 9.8 shows the top twenty names in the dataset.

To answer the second research question, Which key issues have remained consistent at IGF, and which ones have changed?, I explored the changes in key themes over the 12 years of IGF by identifying the top 20 themes at the beginning (2006; figure 9.9), middle (2011; figure 9.10), and most recently (2017; figure 9.11).

Also, when using the entity extraction tools, we identify the most frequently listed organizations, acronyms, countries, and people across all 12 years of the IGF. Figure 9.12 illustrates the top 25 organizations, acronyms, and countries across 12 years of the IGF. Figure 9.8 represents the top twenty names appearing in the IGF transcripts over the 12 years represented in this dataset.

To answer the third research question, In what ways is the Internet of Things (IoT) represented at IGF?, I conducted a cluster analysis across all 12 years. There were initially 60 clusters identified, representing significant

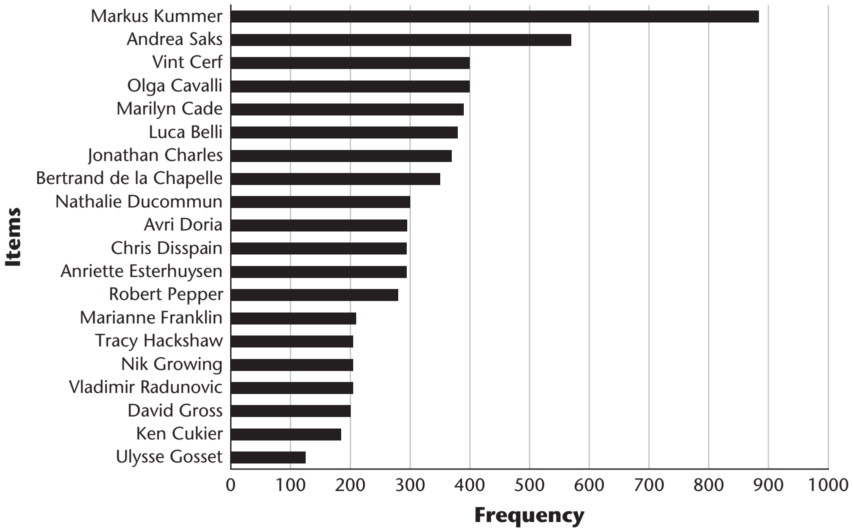


Figure 9.8
Top twenty person names at IGF across 12 years.

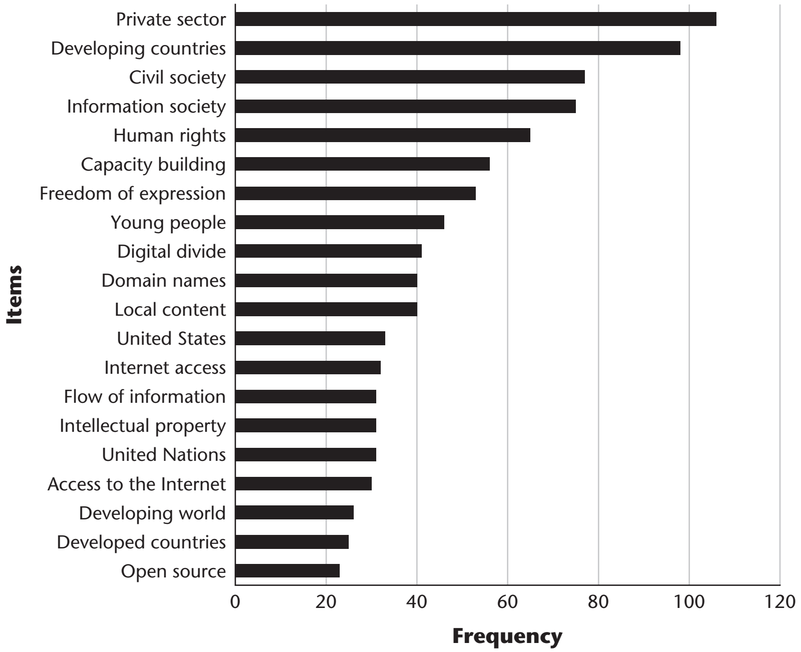


Figure 9.9
IGF top phrases 2006.

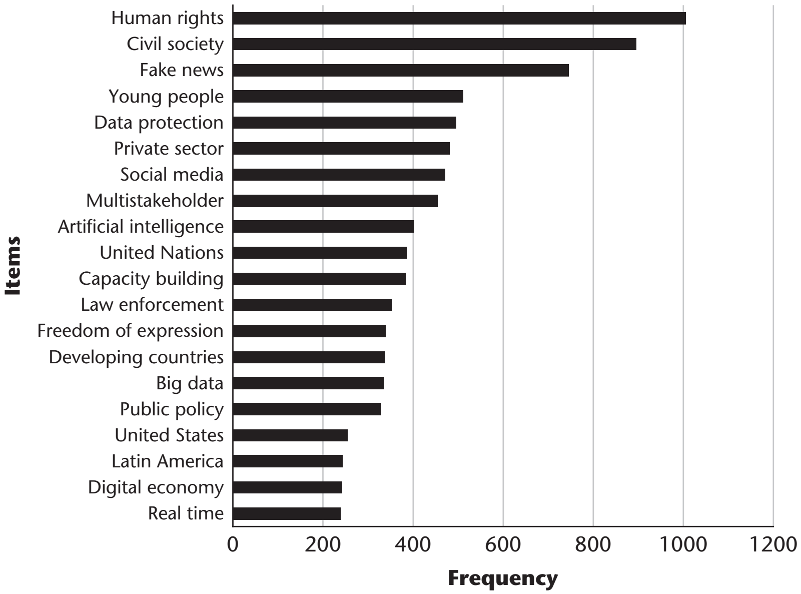


Figure 9.10
IGF top phrases 2011.

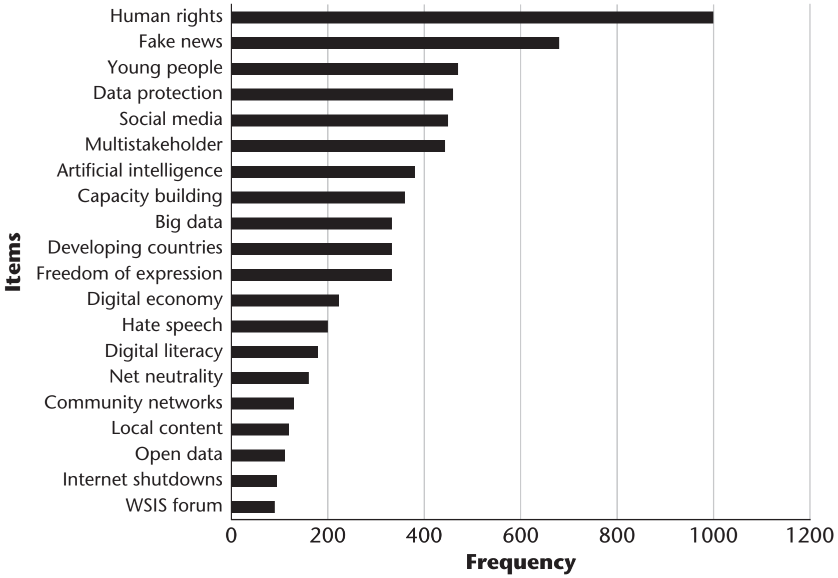


Figure 9.11
IGF top phrases 2017.

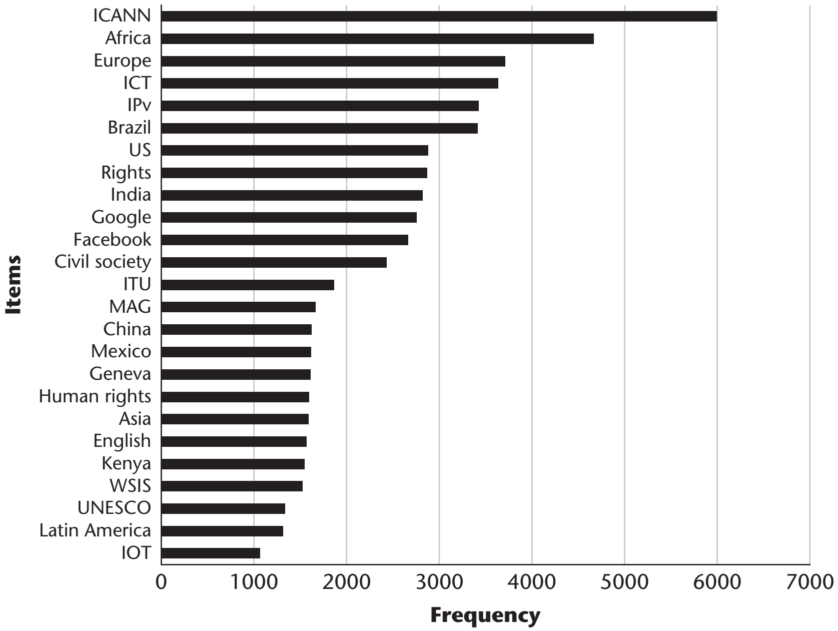


Figure 9.12
IGF entity extraction over 12 years.

thematic groupings. Figure 9.13 illustrates four of those clusters around child protection, capacity building, innovation in infrastructure (including broadband, mobile, net neutrality, and cloud computing), and smart cities and IoT.

Using the inductive technique of topic modeling to look across all 12 years of IGF and the middle and most recent IGFs in the dataset, I found that freedom of expression and human rights are the most durable and consistent topics. Internationalized domain names and mobile phones were taken over in the most recent IGF by fake news and media freedom and multistakeholder discussions. Table 9.2 highlights this topic modeling across the IGFs.

Finally, to answer the fourth research question, To what extent is the 2014 NIST cybersecurity framework represented at IGF?, I deployed the categorization model that captured all the primary categories, subcategories, and sub-subcategories of the framework.

These figures show the frequency of keywords later codified in the 2014 NIST cybersecurity framework when looking at the entire 12-year IGF

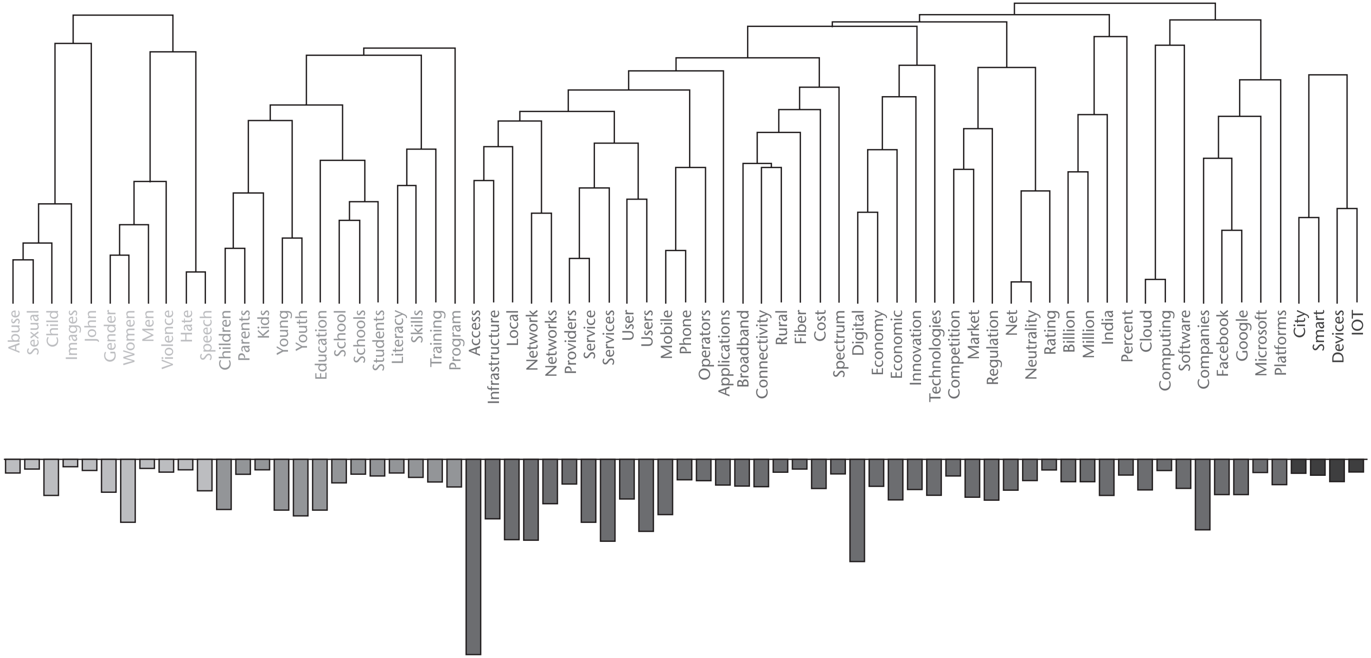


Figure 9.13

Partial illustration of the cluster analysis highlighting four of the 60 clusters.

Table 9.2
Topic modeling over 12 years of the IGF.

2008	2009	2010	2011	2012
IPV4-6 transition	DNS and IANA	Budapest convention	Human rights	Cybercrime
Child pornography	Accessibility	Mobile devices/ Wi-Fi	IDNs	IDNs
Enhanced cooperation		Human rights	Mobile phones	Disaster risk
Freedom of expression			Accessibility	
2013	2014	2015	2016	2017
Journalist/ bloggers	Human rights	Human rights	Human rights	IANA transition
Budapest convention	IANA transition	Net neutrality	Wi-Fi/fiber	Cybersecurity
Mobile devices	DNS	Child abuse	Children online	Fake news
Human rights	CERTs/CSIRTs	IANA transition	IXPs	Human rights
Intellectual property	Accessibility		SDGs	
IDNs			IANA transition	

dataset. Specifically, figures 9.15 and 9.16 allow us to drill down into the categorization model to the sub and sub-subcategories. For example, the most frequently occurring sub-category is PR.PT-3 Least Functionality. This subcategory corresponds with the overall main category labeled “protect,” identified as “PR” in the categorization model. The protect “function” of the NIST cybersecurity framework has six categories, once of which is called “protective technology” (labeled PR-PT in the categorization model). The term “protective technology” refers to how an organization manages protective technology. As defined by NIST, in this subcategory “technical security solutions are managed to ensure the security and resilience of systems and assets, consistent with related policies, procedures, and agreements.” Within protective technology, the third sub-subcategory (labeled PR-PT-3 in the categorization model) refers to the principle of least functionality. As defined by NIST, “the principle of least functionality is incorporated by configuring systems to provide only essential capabilities.” In order to attempt to detect

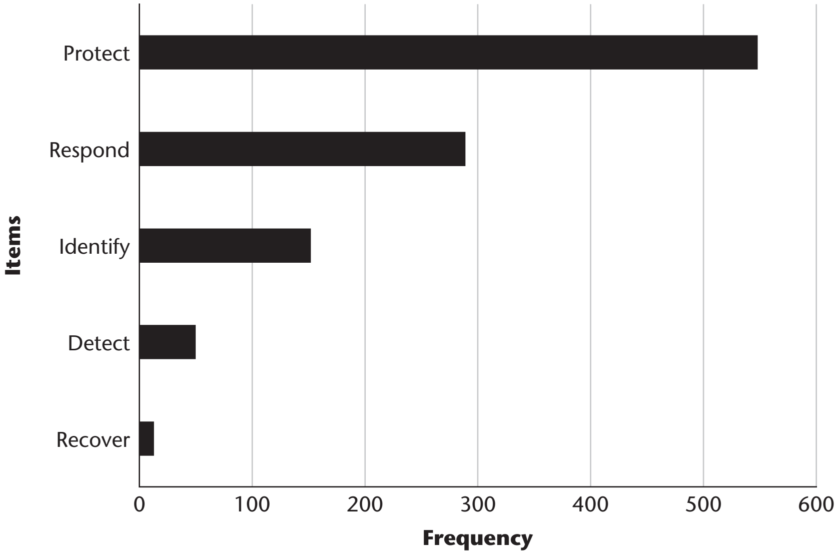


Figure 9.14

Distribution of NIST cybersecurity framework primary categories.

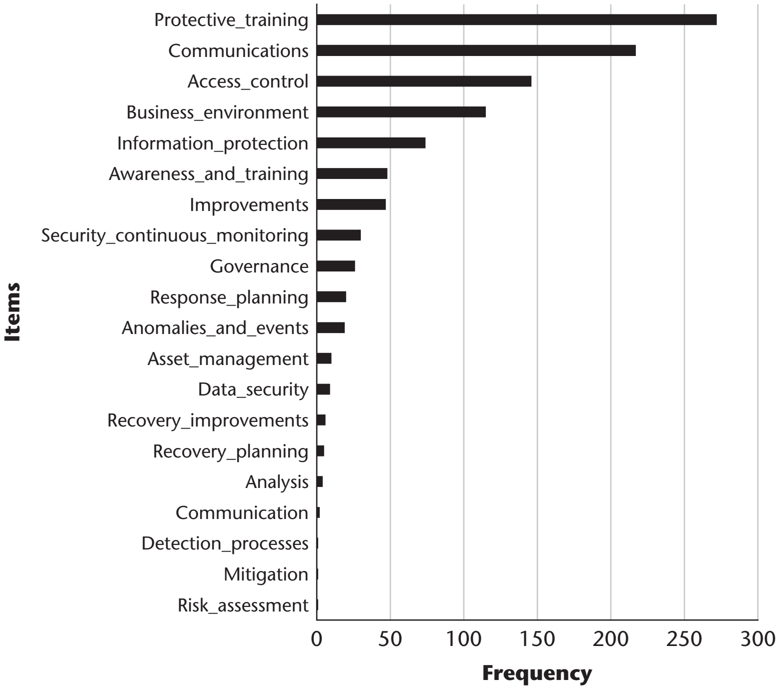


Figure 9.15

Distribution of NIST cybersecurity framework subcategories.

this level of the NIST cybersecurity framework, this component of the cybersecurity model included words and phrases related to the principle of least functionality. Figure 9.16 illustrates the finding that within the 12 years of IGF transcripts, the most frequently occurring component of the NIST cybersecurity framework (as represented by this categorization model) is the principle of least functionality. Each component and subcomponent of the NIST model is represented in this way, with RS.CO-3, the next most frequently occurring concept, representing the response function/category of the framework, and the communication subcategory. The response-communication subcategory focuses on response activities that are coordinated with internal and external stakeholders (e.g., external support from law enforcement agencies). The third element of the response-communication subcategory, labeled RS.CO-3, is designed to assess the degree to which "information is shared consistent with response plans." In this way, we can take this complex and highly detailed government framework and assess the degree to which each of its detailed parts is represented in a database of unstructured text.[5]

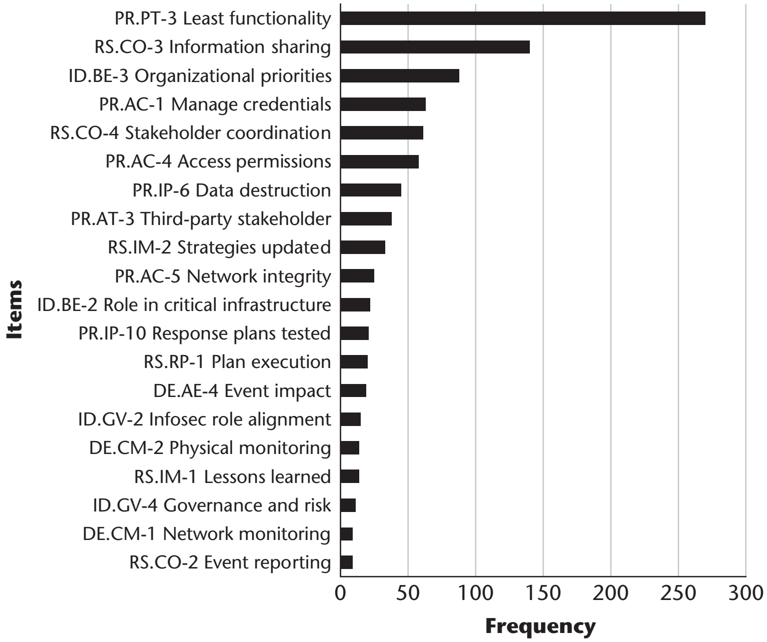


Figure 9.16

Distribution of NIST cybersecurity framework sub-subcategories.

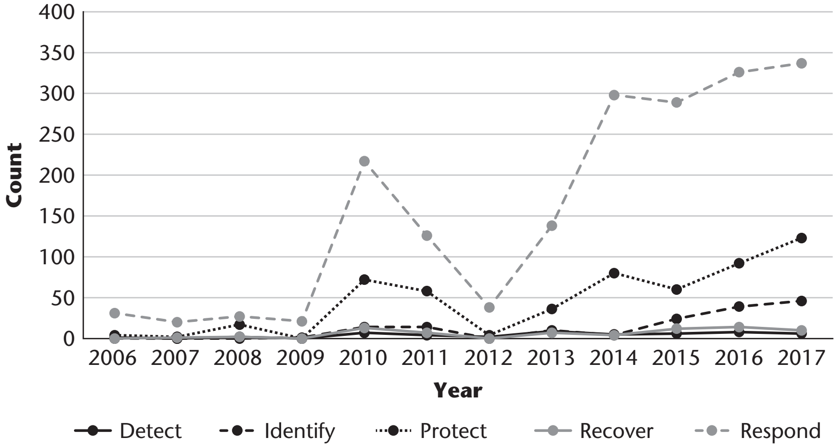


Figure 9.17
Year-by-year line chart distribution of NIST cybersecurity framework keywords.

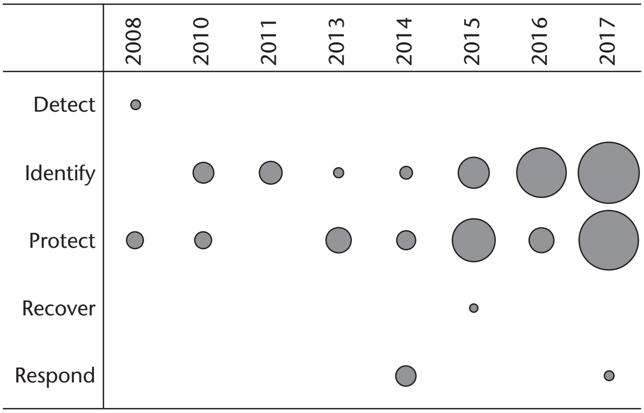


Figure 9.18
Year-by-year bubble chart distribution of NIST cybersecurity framework keywords.

Through the use of this detailed categorization model, we are able to estimate the degree to which concepts in the framework are present (or absent) in the 12 years of IGF transcripts. As such, we see that in 2014, the year the framework was introduced, there was a substantial increase in two of the five components: identify and protect. See figure 9.17.

Another way of viewing these data is via a bubble chart. In figure 9.18, we see that for these same two categories, identify and protect, of the NIST cybersecurity framework there was also an increase after 2014, when the first version of the framework was introduced.

Discussion

With this relatively brief analysis, I have identified the key thematic focus areas of the IGF over its 12-year lifespan. Even without the expensive and time-consuming participant observation most researchers studying the IGF would want to use, I identified key trends, patterns, changes in foci, and important actors—people—operating in this emblematic Internet governance institution. For example, one of the most surprising findings was how early and prominently disability and accessibility issues were included in the IGF. On the basis of topic modeling, we see accessibility appearing as a topic as early as 2009; it appears again in 2011 and 2014. Given the sensitivity of this analysis, and the difficulty of any term making it into the small list of core topics for any given year, this result is striking (only human rights [n=7] and the IANA transition [n=4] appeared more frequently). This finding is most likely a result of the work of the Dynamic Coalition on Accessibility and Disability and its long-term coordinator, Andrea Saks. In addition, while our named-entity extraction analysis identified the name of Markus Kummer appearing most frequently in the dataset, nearly twice as frequently as any other name (which makes sense, given that Markus was the head of the IGF secretariat, and a key leader of the movement within the UN), the second most frequently occurring name is Andrea Saks. Most of the remaining names on the list would be immediately recognizable to anyone studying or participating actively in the IGF over its lifetime.

I have also shown how, in line with Google Trends, cybersecurity topics have become increasingly important over the lifespan of the IGF, with a fairly clear correlation between the introduction of the NIST cybersecurity framework in the United States and these discussions globally at IGF. This finding helps to illustrate how a well-defined major power policy framework—in this case the US under President Obama—can potentially influence global discussions on that issue. My goal in this chapter was not to exhaust an analysis of the policy issues at IGF but to demonstrate the important role these big data analytics and text mining techniques can play in Internet governance research.

Conclusion and Future Research

This study identified some interesting substantive components of the IGF, including the key thematic focus areas over its 12-year lifespan and in a year-by-year comparison. In addition, although I have only scratched the surface, I believe I have demonstrated the power of big data analytics and text mining in Internet governance and cybersecurity research. In this and in other work I have tried to highlight the importance and potential impact of these techniques in monitoring and evaluating the UN's Sustainable Development Goals and implementation of the WSIS action lines.

Regarding future research, I have already highlighted some possibilities to pursue in the near term. Some of these will require adding more variables to the dataset, including type of gathering (e.g., main session, workshop), identifying which dynamic coalition organized the event, and finally, identifying the speaker by name or stakeholder grouping. However, nearer-term studies will focus on building other categorization models: first, to represent different approaches to cybersecurity in order to compare the degree to which each framework is represented in the dataset, and then to identify, represent, and compare other concepts, such as net neutrality and Internet freedom. I also believe exploring the dataset to assess the degree to which the priorities of various stakeholders are represented will be fruitful.

Notes

1. More information on the CRISP-DM process model (1999) is available at http://www.crisp-dm.org/.

2. See the SiteSucker website at https://ricks-apps.com/osx/sitesucker/index.htm.

3. Provalis ProSuite is available at http://provalisresearch.com/.

4. Framework V1.1 Core (Excel). NIST Framework for Improving Critical Infrastructure Cybersecurity. Available at https://www.nist.gov/cyberframework/framework.

5. For more detail, see the NIST cybersecurity framework website: https://www.nist.gov/cyberframework.

References

Bengston, D. N., & Xu, Z. (1995). Changing national forest values: A content analysis (Research paper NC-323). St. Paul, MN: US Department of Agriculture, Forest

Service, North Central Forest Experiment Station. Retrieved from http://www.nrs.fs.fed.us/pubs/rp/rp_nc323.pdf

Bygrave, L. A., & Bing, J. (Eds.). (2009). *Internet governance: Infrastructure and institutions*. Oxford, UK: Oxford University Press on Demand.

Cogburn, D. L. (2003). Governing global information and communications policy: Emergent regime formation and the impact on Africa. *Telecommunications Policy, 27*(1–2), 135–153.

Cogburn, D. L. (2005). Partners or pawns? The impact of elite decision-making and epistemic communities in global information policy on developing countries and transnational civil society. *Knowledge, Technology & Policy, 18*(2), 52–82.

Cogburn, D. L. (2014, September 1). *Uncovering the conceptual antecedents of the NET-Mundial Outcome Document on the future of global Internet governance*. Paper presented at the Annual Symposium of the Global Internet Governance Academic Network (GigaNet), Istanbul, Turkey.

Cogburn, D. L. (2017). *Transnational advocacy networks in the information society: Partners or pawns?* New York, NY: Palgrave Macmillan Springer.

Cogburn, D. L., Mueller, M., McKnight, L., Klein, H., & Mathiason, J. (2005). The US role in global Internet governance. *IEEE Communications Magazine, 43*(12), 12–14.

DeNardis, L. (2009). *Protocol politics: The globalization of Internet governance*. Cambridge, MA: MIT Press.

Deng, Q., Hine, M., Ji, S., & Sur, S. (2017, January). Building an environmental sustainability dictionary for the IT industry. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. Retrieved from http://hdl.handle.net/10125/41264

Exec. Order No. 13,636. (2013, February 12). *Improving critical infrastructure cybersecurity*. Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/executive-order-improving-critical-infrastructure-cybersecurity

Goffman, E. (1959). *The presentation of self in everyday life*. New York, NY: Doubleday.

Goldsmith, J. (2007). Who controls the Internet? Illusions of a borderless world. *Strategic Direction.*

Greenwald, G. (2013, June 6). NSA collecting phone records of millions of Verizon customers daily. *The Guardian*. Retrieved from https://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order

International Telecommunication Union. (n.d.). *WSIS Action Lines*. Available at https://www.itu.int/net/wsis/stocktaking/help-action-lines.html

Kaisler, S. H., Espinosa, J. A., Armour, F., & Money, W. H. (2014, January). Advanced analytics: Issues and challenges in a global environment. In *2014 47th Hawaii International Conference on System Sciences* (pp. 729–738). Waikoloa, HI: IEEE Computer Society.

Kassner, M. (2015, February 2). Anatomy of the Target data breach: Missed opportunities and lessons learned. *ZDnet*. Retrieved from https://www.zdnet.com/article/anatomy-of-the-target-data-breach-missed-opportunities-and-lessons-learned/

Koerner, B. I. (2016, October 23). Inside the cyberattack that shocked the US government. *Wired*. Retrieved from https://www.wired.com/2016/10/inside-cyberattack-shocked-us-government/

Kushner, D. (2013, February 26). The real story of Stuxnet. *IEEE Spectrum*. Retrieved from https://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*(70), 1. Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lee, Y. B., & Myaeng, S. H. (2002, August). Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 145–150). New York, NY: ACM.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013, May). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday, 18*(5–6). Retrieved from https://firstmonday.org/article/view/4366/3654

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 415–463). Boston, MA: Springer.

Mueller, M. L. (2009). *Ruling the root: Internet governance and the taming of cyberspace*. Cambridge, MA: MIT Press.

Mueller, M. L. (2010). *Networks and states: The global politics of Internet governance*. Cambridge, MA: MIT Press.

Musiani, F., Cogburn, D. L., DeNardis, L., & Levinson, N. S. (Eds.). (2016). *The turn to infrastructure in Internet governance*. New York, NY: Palgrave Macmillan.

National Institute of Standards and Technologies (NIST). (2014, February 12). Framework for improving critical infrastructure cybersecurity. Retrieved from https://www.nist.gov/system/files/documents/cyberframework/cybersecurity-framework-021214.pdf

Paré, D. J. (2003). *Internet governance in transition: Who is the master of this domain?* Lanham, MD: Rowman & Littlefield.

Pepitone, J. (2014, January 12). 5 of the biggest-ever credit card hacks. TJX: 94 million. *CNN Business*. Retrieved from https://money.cnn.com/gallery/technology/security/2013/12/19/biggest-credit-card-hacks/3.html

Rousu, J., Saunders, C., Szedmak, S., & Shawe-Taylor, J. (2005, August). Learning hierarchical multi-category text classification models. In *Proceedings of the 22nd international conference on Machine learning* (pp. 744–751). New York, NY: ACM.

Schneider, C. (2016, May 25). The biggest data challenges that you might not even know you have. *IBM Watson*. Retrieved from https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

Thierer, A. D., & Crews, C. W. (Eds.). (2003). *Who rules the net? Internet governance and jurisdiction*. Washington, DC: Cato Institute.

Vinton, K. (2014, September 18). With 56 million cards compromised, Home Depot's breach is bigger than Target's. *Forbes*. Retrieved from https://www.forbes.com/sites/katevinton/2014/09/18/with-56-million-cards-compromised-home-depots-breach-is-bigger-than-targets