Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2017 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 2017

Clustering and Topic Modelling: A New Approach for Analysis of National Cyber security Strategies

Farzan Kolini
The University of Auckland, f.kolini@auckland.ac.nz

Lech Janczewski
Department of Information Systems and Operation Management The Business School University of Auckland, lech@auckland.ac.nz

Clustering and Topic Modelling: A New Approach for Analysis of National Cybersecurity Strategies

Completed Research Paper

Farzan Kolini
Department of Information Systems
and Operation Management
The Business School
University of Auckland, New Zealand
f.kolini@auckland.ac.nz

Lech Janczewski
Department of Information Systems
and Operation Management
The Business School
University of Auckland, New Zealand
l.janczewski@auckland.ac.nz

Abstract

The consequences of cybersecurity attacks can be severe for nation states and their people. Recently many nations have revisited their national cybersecurity strategies (NCSs) to ensure that their cybersecurity capabilities is sufficient to protect their citizens and cyberspace. This study is an initial attempt to compare NCSs by using clustering and topic modelling methods to investigate the similarity and differences between them. We also aimed to identify underlying topics that are appeared in NCSs. We have collected and examined 60 NCSs that have been developed during 2003-2016. By relying on institutional theories, we found that memberships in the international intuitions could be a determinant factor for harmonization and integration between NCSs. By applying hierarchical clustering method, we noticed a stronger similarities between NCSs that are developed by the EU or NATO members. We also found that public-private partnerships, protection of critical infrastructure, and defending citizen and public IT systems are among those topics that have been received considerable attention in the majority of NCSs. We also argue that topic modeling method, LDA, can be used as an automated technique for analysis and understanding of textual documents by policy makers and governments during the development and reviewing of national strategies and policies.

Keywords: National cybersecurity strategy, NCS, similarity, topic modelling, hierarchical cluster, LDA, latent dirichlet allocation

Introduction

Cybersecurity remains relatively important on the top of the harsh and turbulent business environment. The internet has been recently changed into a minefield for digital crime, information leakage, cyber harassment, and cyber-attack on a large scale. In the aftermath of cyber-attacks on nation-states such as Estonia (2007), Georgia (2008) Kyrgyzstan (2009), South Korean's banks (2010), Stuxnet malware (2010), Cyber espionage against US (2012), New York dam's SCADA systems (2016) or mass data breaches on LinkedIn (2013), Yahoo (2014), Dropbox(2014), and Telegram(2016) have enforced many national governments to reconsider their perception of cybersecurity risks, and potential impacts on society, economy, and critical infrastructures. In this regards the key dilemma for national government is to protect nation-state against crippling cyber threats and respond effectively to minimize the adverse impact of cyber related incidents (kolini & Janczewski, 2015). Therefore, by developing a national cybersecurity strategy (NCS), governments aim to bolster the security of internet, which is pivotal for driving the modern economy and technology-reliant societies and to reinforce the resiliency of critical national infrastructures. Thus, Understanding the multifaceted nature of cybersecurity is a key enabler for national governments and to adjust themselves to rapidly changing nature and complexity of cybersecurity ecosystems (Stevens, 2012).However, In spite of government's best endeavor to address all the facets of cybersecurity, it is still not clear which aspects of NCS are more prevalent and has been received more attention during the development of cybersecurity strategies. Besides, many governments are failed to understand how to measure the effectiveness of their NCS in compare to other NCSs (Klimburg, 2012). For instance, Lehto (2013) compared and classified eight NCSs to understand similarities and differences between NCSs based on three comparison metrics. ITU (2005) investigated 14 NCSs to identify common themes and topics for global culture of cybersecurity based on five measurement metrics. Luiijf (2013) has analyzed 19 NCSs to address whether they are harmonized according to 9 different common themes and topics.

To our best knowledge, a common set of performance indicators or metrics that can be used as a gold-standard for systematically evaluating the effectiveness of NCSs is yet to be developed. Therefore, the researchers have identified a set of comparison or measurement metrics based on their experience and knowledge, which is followed by a manual approach for classification and interpretation of cyber strategies. In general, all these studies are resource-intensive and difficult to apply when dealing with large documents and the study needs to be refreshed whenever NCSs has been changed or revised by national governments. Hence, with the abundant research on the systematic analysis and classification of NCSs, in this research we aim to investigate the following research objectives: (1) to understand the relationships between NCSs by using clustering and topic modelling techniques (2) to identify and investigate the most frequent topics in the collection of NCSs.

Our dataset for this study includes 60 NCSs that have been published in the English language from 2003 to 2016 from various online databases. The rest of this paper is organized as follows. Following the brief introduction, we present a review of the literature and theory that forms the basis of this research. Next, we present the research methodology, followed by data collection techniques and analysis. Finally, we discuss the findings and elaborate its implication on theory and practices as long as research constraints and future studies.

Literature Review

National Cybersecurity Comparison

ENISA (2012) described NCS as a tool to improve the overall security and resilience of national information system networks, internet, infrastructure and IT-based services. It is designed to be a top-down approach to cybersecurity that elucidated a set of high-level national initiatives and objectives that should be achieved in a specific timeframe. Drawing on prior literature, NCSs are usually analysed through comparative analysis approach, which has been used previously in cross-culture analysis studies. The comparative analysis demonstrates the similarity and differences between NCSs by identifying crucial differences or similarities between them. This method can provide enough foundation to create an explanatory model for findings and analysis. We noted that most of the previous studies have followed aforementioned approach for comparison and analysis of NCSs. Table 1 showed list of these studies with a summary of methodology and their findings.

Author	NCSs & Distributions	Methodology	Description	Theory
Dunn (ITU, 2005)	14 NCSs” ▪ 10 from EU ▪ 2 form Australasia ▪ 2 from America	Used 5 comparison metric from a cross comparison surveys	▪ Identified similar topics in NCSs such as cyber warnings, legal aspects, and public-private-partnership. ▪ Identified significant differences in technical and national security approaches.	Political science theory
Luiijf et al. (2011)	10 NCSs: ▪ 5 form EU ▪ 3 form Australasia ▪ 2 form America	Comparison based on 9 topics identified by researchers	▪ Compared similarities and differences between 10 NCSs.	N/A
Wamala (ITU,2012)	N/A	Used 8 elements as main features for cybersecurity programme	▪ Proposed a theoretical reference model for cybersecurity strategy and a guidance for implementation of NCS	N/A
Kilimburg (NATO,2012)	N/A	Three workshops with expert in cybersecurity.	Proposed a theoretical framework for better understanding of different facets of NCS and practical manual for development of NCS	N/A
OECD (2012)	10 NCSs” ▪ 6 from EU ▪ 2 form Australasia ▪ 2 from America	Circulated an open-ended questionnaire survey from governments and	highlighted that some concepts such as government co-ordination responsibility, public-private co-operation, international co-	N/A

		non-government stake holders.	operation, and social value are shared common subjects among NCSs.	
Luiijf et al. (2013)	19 NCSs: ▪ 11 from EU ▪ 4 form Australasia ▪ 2 from America ▪ 2 from Africa	Comparison based on 11 categories	Compared 19 NCSs to propose a formal structure for NCS development.	N/A
Letho (2013)	8 NCSs: ▪ 5 from EU ▪ 1 form Australasia ▪ 2 from America	Used 3 comparison elements	Found a significant variance in scope and depth of cybersecurity strategies.	N/A
Shafqat et al. (2016)	20 NCSs: ▪ 11 from EU ▪ 7 form Australasia ▪ 2 from America	Used 10 comparison metrics	Found number of differences in the scope and approached in NCSs and provider some recommendation for development of NCSs.	N/A

Table 1. Summary of Previous Literatures on NCS's Comparison

As can be seen, the majority of research analysing NCSs has been defined comparison metrics, elements or topics to identify similarities and differences between them. On the other hand, a number of studies have been conducted to design and develop a new set of measures for strategy development and implementation. By analysing studies listed in Table 1, it is evident that there is no consensus among researchers on the identification of metrics or topics for comparison analysis. Besides, the analysis of NCSs are resource-intensive and to extensively rely on the experience and knowledge of the researchers, which lead to a limited number of NCSs have been chosen for analysis and comparison. To our best knowledge, we could not find any study that follows a more holistic approach to consider all the available NCSs for comparison and analysis.

Institutions for Cybersecurity

Unlike other conventional theories, a unified body of thought does not form institutional theories. Instead, it can be invoked to investigate a variety of disciplines like political science, economics, organizational behaviour, information technology, and international relations (Steinmo, 1992; Moe, 1984; Hall, 1996; Backhouse & Silva, 2006). Powell & DiMaggio (2012) suggested that institutional isomorphism could be occurred because of coercive, normative and mimetic forces, which together promote the success and legitimacy of actors in their institutional environment. Since institutional forces, mimetic or coercive often surfaced at the time of uncertainty and cyberspace has been suffered from significant uncertainties, nation states aim to follow other nations that are perceived to be more legitimate or superior in cybersecurity. Uncertainty in cyberspace is caused by volatile nature of cyber-attacks and their impacts on the states, people and economy. Relying on institutional theory, in this study we aim to understand whether the emergence of institutions of cybersecurity can be investigated by semantic analysis of NCSs. Further, within the institutional field, cybersecurity strategies can be correlated to each other in spite of inevitable conflicts in national interests, legislations, culture, and diplomacy. The characteristic of cybersecurity landscape including international institutions has been studied by kolini & Janczewski (2015).Nation states are enacted in various international institutions like EU, ACAN, OECD, NATO, ITU, CERT and UN. Whilst some of these institutions have catered guideline or manuals for cybersecurity implementation (ITU and OECD), the others suggested that cybersecurity is a key initiative for national security and to address the role of national security in the development of NCSs. Table 2 has briefly addressed several international institutions for cybersecurity.

Institution for Cybersecurity	Liability	Cybersecurity Operations	Members
United Nations-UN	International Forum	Policy maker	193 countries
International Telecommunication Union-ITU	International Forum	Guidelines, technical standards and education	193 countries
European Union-EU	Regional Forum	Legislations and policy maker	28 countries
North Atlantic Treaty Organization-NATO	Regional Forum	Guidelines, education, cyber/military Prevention and Response cooperation	33 countries

Organization for Economic Co-operation and Development-OECD	International Forum	Publications and reports	34 countries
Asia-Pacific Economic Cooperation-APEC	International Forum	Publications and reports	21 countries
Organization for Security Co-operation in Europe -OSCE	Regional Forum	Training , workshops and reports	57 countries 11 partners
Five Eyes-FVEY	International Alliance	Security and Intelligence sharing	5 countries

Table 2. Institutions for Cybersecurity

Since collective actions of institutions can considerably reduce the transactional costs of cybersecurity and improve the overall effectiveness of incident response programs, we perceive that membership in these institutions can potentially influence actors to align their cybersecurity strategies with institutional requirements to gain a higher level of legitimacy and efficacy. Hence, we argue that these mutual interactions can ultimately result in integration and harmonisation between NCS. Although institutional frameworks can be an appropriate tool for examining the cybersecurity initiatives, Reich (2000) and Choucri et al. (2014) suggested that intuitional theory should be revisited with more pragmatic approaches to consider empirical studies in the forefront of theoretical propositions. Therefore, in this study, we are examining 60 NCSs to understand the similarities and integration between these strategies. To our best knowledge, while several studies have explored the coercive and normative forces of institutions at organisational and government level, more empirical studies are necessary to investigate NCSs from the lens of intuitional theories

Data Collection, Methodology, and Analysis

National Cybersecurity Strategies (NCSs) Data Collection

The presence of NCSs is a relatively recent and ongoing project. For instance, outside of the US, the formulation of NCS has been started in late 2010. Many nation states have been enforced to introduce a NCS to treat the emergence of cyber threats as a new countrywide challenge to their national security and to align their cybersecurity capabilities with cyber threat ecosystem. Security of cyberspace is a cross-sector challenge that could adversely threaten critical infrastructures such as energy, healthcare, transportation and financial institutions. To get a broader picture of cybersecurity strategies, this study decided to collect as many NCSs from different available databases such as NATO[1], ENISA[2], and OSCE[3]. The data were collected from the official website of each database. The first NCS was drafted by the U.S in 2003 while the UK has developed the latest refreshed strategy during November 2016. A total of 60 NCS documents in a period of 13 years is obtained and included in this study. Table 3 provides a brief summary of this study dataset. We have excluded non-English NCS for this study. Besides, all NCS's pdf files are converted to text format by a postgraduate student who has annotated all tables and figures into text. This approach is appropriate for text mining and clustering analysis. The principle researcher has also manually examined the accuracy and constituency of this transformation.

Dataset	Number of Strategies	Published Year
Europe	34	2011-2016
Asia, Middle East and Africa	19	2011-2016
America (North, Central, and South)	5	2003-2015
Australasia	2	2015-2016
Total	60	2003-2016

Table 3. Collected National Cybersecurity Strategies (NCSs)

[1] https://ccdcoe.org/cyber-security-strategy-documents.html
[2] https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map
[3] http://www.osce.org/

Hierarchical Clustering of NCSs

With the emergence of Big Data, in digital libraries, internet and online repositories, clustering approaches can be used increasingly to analyse the relations among terms in documents. The cluster analysis attempts to classify similar textual documents in-group clusters to distinguish them from each other groups. A variety of methods has been introduced for representing data and measuring proximity of text documents (Jain et al. 1999).Hence we are trying to find the answer for the following question:

Research Question 1: What are the relationships between NCSs and how they can be explained?

The similarity measures which describe the closeness or separation of text documents can be determined by various hierarchical or partial clustering algorithms including Euclidean Distance, Jaccard Coefficient, Pearson Correlation Coefficient, Cosine Similarity, and K-means (Sneath et al. 1973, Willett 1998, Salton 1998, and Jain et al. 1999). Jain (1999) pointed out that there is no universally acceptable scheme for computing similarity between patterns represented and all the above mentioned approaches have been widely applied for clustering texts and documents. In particular, Leydesdorff (1998), Strehl et al. (2000), and Huang (2008) suggested that Pearson's Coefficient, Jaccard's Correlation, and Cosine's Similarity produce often a better and more accurate results for similarity measures and text document clustering. Besides, other unsupervised clustering approaches like K-means can be more suitable and noise tolerance while processing a larger number of not textual datasets (Punj, 1983). One immediate advantage of the hierarchical cluster over partial algorithm such as K-means is that we can successfully merge all NCSs during clustering procedures a nested series of clusters.

Since this study considered relevantly smaller datasets and aimed to merge all the NCSs during clustering procedures, we followed extant literature by applying Pearson's Correlation Coefficient distance (Leydesdorff, 1998 and Lin et al., 2016), and drawing the linkage between clusters by Ward's linkage. The Ward's Linkage is based on the linear model criterion of least square to identify the pair of clusters in such a way that the within-group sum of squared errors can be minimized (Ward, 1963).The hierarchical cluster is often illustrated as a dendrogram as shown in Figure 1. . To compare or interpret the result of hierarchical cluster, researchers specify a number of clusters by visual examination of the dendrogram (Borcard et al., 2011). In this study, we produced graphs of fusion level, Mantel Correlation and Heat map to identify the optimal number of clusters, for analysis and comparison (Borcard et al., 2011).

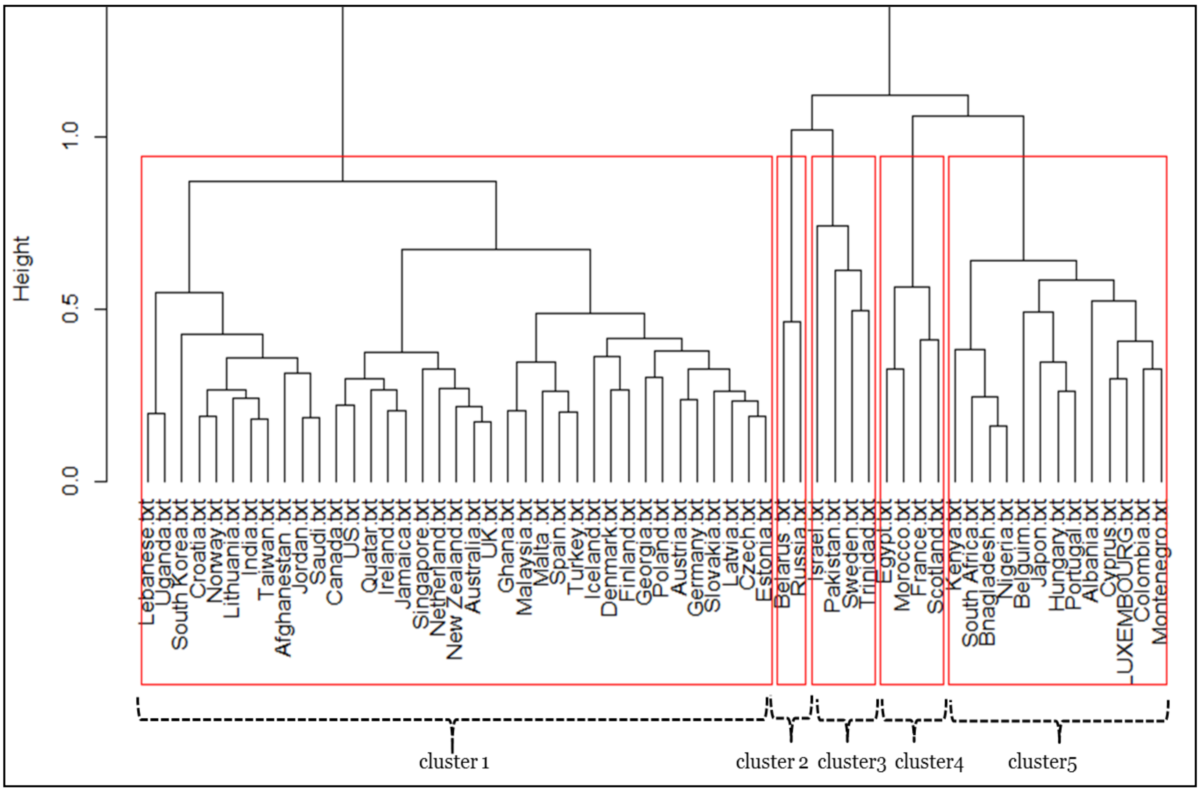

Figure 1. Hierarchical Cluster Dendrogram

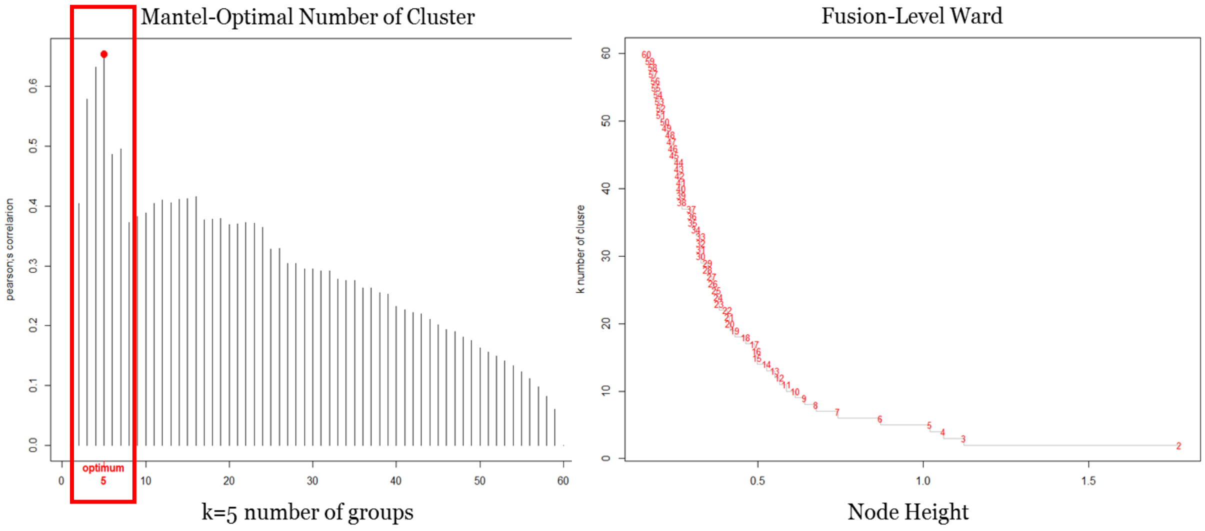

Figure 2. Optimal Number of Cluster

Analysis of NCS's Hierarchical Clusters

We broadly examined nested clusters in dendrogram and Mantel Correlation to determine an optimal number of clusters that can appropriately represent our collected NCSs (Figure 2.). Mantel Correlation address the original distance matrix and binary matrices calculated from dendrogram cut at various level (Borcard et al., 2011). The five high-level clusters can be modestly categorised NCSs across different data sources (Figure 1.). This hierarchical clustering[4] computes a similarity matrix to determine the linkage between NCSs to classify similar strategies into groups, and to distinguish them from other groups. A summary of cluster's characteristics has been outlined in table 4.

Clusters	Members	Hierarchical Cluster Characteristic
Cluster1	37	▪ 22 NATO members ▪ 5 FVEY members ▪ 2 members (Georgia and Estonia) suffered from destructive cyber-attacks ▪ 7 members has refreshed their NCSs during the last two years
Cluster2	2	▪ 2 members have a very close political and cultural ties ▪ Not member of regional institutions
Cluster3	4	▪ Not member of regional institutions ▪ 2 members has been involved in regional turmoil for many years
Cluster4	4	▪ 2 NATO members ▪ 2 Arab League members
Cluster5	13	▪ 5 NATO members ▪ 3 African union members

Table 4. Cluster's Characteristics

The result of our analysis suggested potential influences from participation in international institutions in the development of NCSs. For instance, we noted a stronger similarity between NCS, which created by EU and NATO members where 27 members of clusters 1 and 4 with the highest value of Pearson's Correlations are the NATO and EU members. We also found that the majority of NCSs have been created or updated after the introduction of NATO's National Cybersecurity Framework by NATO Cooperative Cyber Defence Centre (CCDCOE) in 2012 (Klimburg, 2012).Our results indicated a close harmonisation between NCSs that have suffered from a series of motivated cyber-attacks. For instance, Both Estonia and Georgia experienced an unprecedented cyber-attack between 2008 and 2009, have developed associated NCSs. We also found that there is a tendency among nation states to demonstrate a closer alignment with their neighbours or politically-aligned partners. For instance, like-minded nations with similar political views such as (Russia and Belarus) or FEY's members, (UK,

[4] Due to space limit, we cannot include the Pearson correlation matrix that has been used for this analysis. However, we are able to provide the calculation upon any request to the authors.

Canada, US, and New Zealand), have developed proximate NCSs. Similar to this, geographically and culturally closed countries such as (Egypt and Morocco), (Denmark, Iceland, and Finland), (Jordan and Saudi Arabia), (Nigeria, South, Africa, and Kenia), (Poland, Austria, Germany, and Slovakia), and (Estonia and Latvia) have also shared some sort of similarities and integration in their cyber strategies.

Latent Dirichlet Allocation (LDA) Topic Modeling

Most of the today's information are generated and stored in the form of texts like news, blogs, web pages, scientific articles, books, policies, and social media. As a result, organizations, researchers, politicians, and decision makers are actively seeking for new approaches that could help them to search, organise, synthesise and understand this volume of information. Historically scholars have applied several interpretive and qualitative approaches to discover and understand underlying themes that ultimately suggest a meaningful picture of the phenomenon under investigation.

David Blei (2012) highlighted that topic modelling algorithms are statistical methods to understand underlying latent topics that are inherent in text documents and help researchers toward better summarization and interpretation of collected information with topic labels. Several variations of topic modelling and document clustering techniques have been introduced by various researchers to enhance data representation, information retrieval and machine learning. Field began an earnest with Latent Semantic Indexing (LSI) (Deerwester, 1990), Mixture of Unigrams model (McCallum, 1999), Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999), Latent Dirichlet Allocation (LDA) (Blei, Ng & Jordan, 2003), and Correlated topic models (CTM) (Blei & Lafferty, 2007). We resorted on LDA model to identify the most pertinent topics in NCSs. The LDA approach has been widely applied by various IS scholars to analyse IS academic research (Chen & Zhao, 2015), online App Stores (Vakulenko et al., 2014), social media (Shahbaznezhad & Tripathi, 2015), phishing and spam discovery (Ramanthan & Wechler, 2013), and Organisations (Syed & Dhillon, 2015).

LDA is a generative probabilistic algorithm, which has been frequently utilized in text mining and topic extraction (Blei et al., 2003). LDA treats each document as a bag of words and aims to discover latent topics from a distribution over words without considering the order of occurrence (Blei et al., 2003).Since LDA is an unsupervised learning algorithm there is no need for manual labelling of each document. Topic modelling package in R has introduced several libraries such as text mining "tm", topic modelling "topic models", and "pdftotext" for identifying latent topics. The R package also offered two sampling algorithms for fitting topic models namely VEM (Blei et al., 2003) and Gibbs (Chang, 2009). Although both approaches have been used in different studies, we relied on Gibbs sampling for inference and analysis. For this purpose, we downloaded NCSs from various online databases and converted them to text format for computation of a single corpus document. The corpus has been pre-processed by removing non-influential words such as punctuations, stop and common words, numbers, and capitalization. We used Term Frequency (TF) to remove terms with lower frequency, irrelevant terms and those terms occurred in many documents (Hornik & Grun, 2011; Nolasco & Oliveira, 2016).

We computed Document Term Matrix (DTM) and used LDA algorithm to perform textual analysis of corpus document to determine the topic matrix and its associated words. Each identified topic can be linked to relevant cybersecurity strategies based on the occurrence of words in each strategy document. Whilst determination of number of topics remains an open research problem (Vakulenko et al., 2014; Syed & Dhillon, 2015; Shi et al., 2015), only of few studies have suggested methods for estimating the optimal number of topics (Greene et al., 2014; Grun & Hornik, 2011). Previous studies with very large datasets may suggest up to 300 topics to fit and train topic models (Wei & Croft, 2006); however other researchers have suggested that the standard measures for estimating the number of topics can be misleading while the semantic analysis of topics needs to be performed through human intervention(Chang et al., 2009; Vakulenko et al., 2014).Similarly, Blei (2012) signified that accuracy in a number of topics should not be considered as the only measure for the model selection; Instead, interpretability of results and evaluation of topics is the salient fact in topic modelling approach. In this study, we re-ran LDA algorithm for 5, 10, 15, 20, 50 and 100 topics, since our dataset is relatively small, after careful interpretation and analysis we noted that 10 topics could represent a reasonable result and be adequate for interpretation and analysis of the latent topics.

Topic Modelling Analysis

Since NCSs do not include topic labels identifying their content, we aim to extract the underlying topics and corresponding words to answer the following question:

Research Question 2: Which topics can be identified in the collection of NCSs?

Topic words can be grouped by common higher-level themes to compile topic labels. Previous studies have suggested that topic labelling can be performed by human judgement (Sidorova et al. 2008; Chang et al., 2009; Shi et al., 2015; Syed & Dhillon, 2015), semi-supervised approach (Vakulenko et al., 2014; Patton et al., 2011), or automatic labelling (Nolasco & Oliveira, 2016). Since automatic labelling for semantic analysis of topics is still in its early stage and it is not practical to find a gold-standard list of topics to compare against all datasets, we decided to follow a manual method for identifying topic labels. Therefore, labelling of NCS topics were performed by two independent researchers with prior knowledge and extent experiences in cybersecurity and public policy studies. Next, the corresponding words under each topic are systematically explored and topic labels have been identified. Our topic model discriminates among 10 different topics Table 5 presents 10 identified topics and 15 corresponding words that used to described each topic in our dataset.

Topic #	Topic Words	Topic Labels
Topic 1	digital, defense, strategy, service, data, international, measures, domain, systems, citizens, level, public, approach, vital, use	Defending citizens and public IT systems
Topic 2	development, service, sector, digital, program, technology, initiative, crime, promoting, content, strategic, developing internet, support	Organization/Sector for cybersecurity
Topic 3	Cyberspace, systems, attack, critical, private, vulnerabilities, networks, response, infrastructure, effort, sector, agencies, strategy, internet, defend	Cyberspace resiliency against attacks for critical sectors and infrastructure
Topic 4	Risk, objectives, management, area, technology, standards, assessment, infrastructure, establish, environment, policies, implementation, research, develop, initiative	Develop policy and standard for technology and infrastructure
Topic 5	State, law, ministry, compute, international, legal, republic, council, personal, criminal, internet, access, protection, bodies, convention	Legislation and laws for cybercrime
Topic 6	Public, cooperation, private, International, necessary, implementation, electronic, level, protection, communication, system, systems infrastructure, administration, space	Public-Private and International cooperation
Topic 7	Cybersecurity, measures, international, cyberspace, including, development, activities, business, capabilities, attacks, necessary, promote, management, service, systems	Cybersecurity measure for cyber capabilities
Topic 8	Businesses, online, sector, threat, strategy, internet, governments, awareness, threats, business, organizations, skills, private, protect, public	Training and awareness for public, private sector, and online businesses
Topic 9	Access, data, management, business, use, policy, ensure, organizations, risks, requirement, network, physical, procedures, controls, users	Risk Management Procedures
Topic 10	Cybersecurity, strategy, critical, infrastructure, cybercrime, policy, framework, threats, sector , awareness, protection, private, cyberspace, need, public	Critical Infrastructure (CI) protection

Table 5. NCS's Identified Topics by LDA Algorithm

Topics Probability and Distribution

We now look at some illustrative instances to answer:

Research Question 3: how frequently do topics appear in NCS and hierarchical clusters?

The first Topic, addresses requirements for safeguarding citizens and public IT systems against crippling cyber-attacks. The second topic emphasises on organization or sector for cybersecurity. The third topic outlines resiliency of critical sectors against cyber threats. The forth topic articulates the need for cybersecurity framework, policy or standards. The fifth theme, Legislation and laws for cybercrime, disputes the need to develop and integrate national and international legislation (e.g. law, legal, council) for battle against cybercriminals. The sixth category includes topics that seeks attention for improving public-private cooperation in cyber domain. The next topic focuses on measurement of cybersecurity capabilities. Topic 8, Training and Awareness Programme, promotes the requirement for improving the overall cybersecurity awareness and skills among individual, online businesses and public sector. The ninth topic focuses on a cybersecurity risk management to identify cyber risks at national and international level. Topic 10, Critical Infrastructure (CI) Protection, depicts the importance of national critical infrastructures protection against catastrophic events (e.g. public-private-partnerships). We also calculated the probability of topic words outlined in Table 5 for each NCS to understand how well each topic is represented in the document corpus. Figure 3 illustrates an extract of the topic probabilities, as calculated by the LDA topic modelling, for each NCS and the normalised percentage of each topic is computed accordingly. By looking at the probability of topics in each column and their distribution among NCSs, we can extrapolate some initial findings from Figure 3. First, some topics show lower distribution among NCSs, such as topic 2 (7.4%), topic 7 (7.1%) and topic 9 (7.2%). On the other hand, there are other topics, Topic 1(12.2%), Topic 6 (16.3%), and Topic 10(12.1%), which appears almost across all NCs. These correspond to the topic of "defending citizens and public IT systems", "public-private and international cooperation", and "critical infrastructure protection (CIP)".When a topic appears more frequently, it means that that topic has gained enough attention and considered as a key initiatives from several governments. For instance, many national governments have bolstered the security of their critical infrastructures in the aftermath of the Stuxnet malware. Thus, it is not surprising that CIP has been considered as a key concern in most of NCSs.

In order to illuminate an explanation for this topic distribution, we are looking at extant literature and noted that all these three topics have been extensively scrutinised and recommended by international institutions such as NATO, EU or ITU. Besides, in line with our hierarchical clustering findings, we noted that nation states such as Belarus or Russia which are not members of any regional institutions (Refer to Table 4) are among those who have the lowest distribution for these topics. As another example, we noted that countries such as Australia, Canada, New Zealand, and UK, with the highest probability for topic 8, "training and awareness for public and private sector, and online businesses" are members of the same International Cybersecurity Institution (FVEY) and three of them except Canada have been recently updated their NCSs by emphasizing on new initiatives for uplifting cybersecurity awareness and skills among their citizen and businesses.

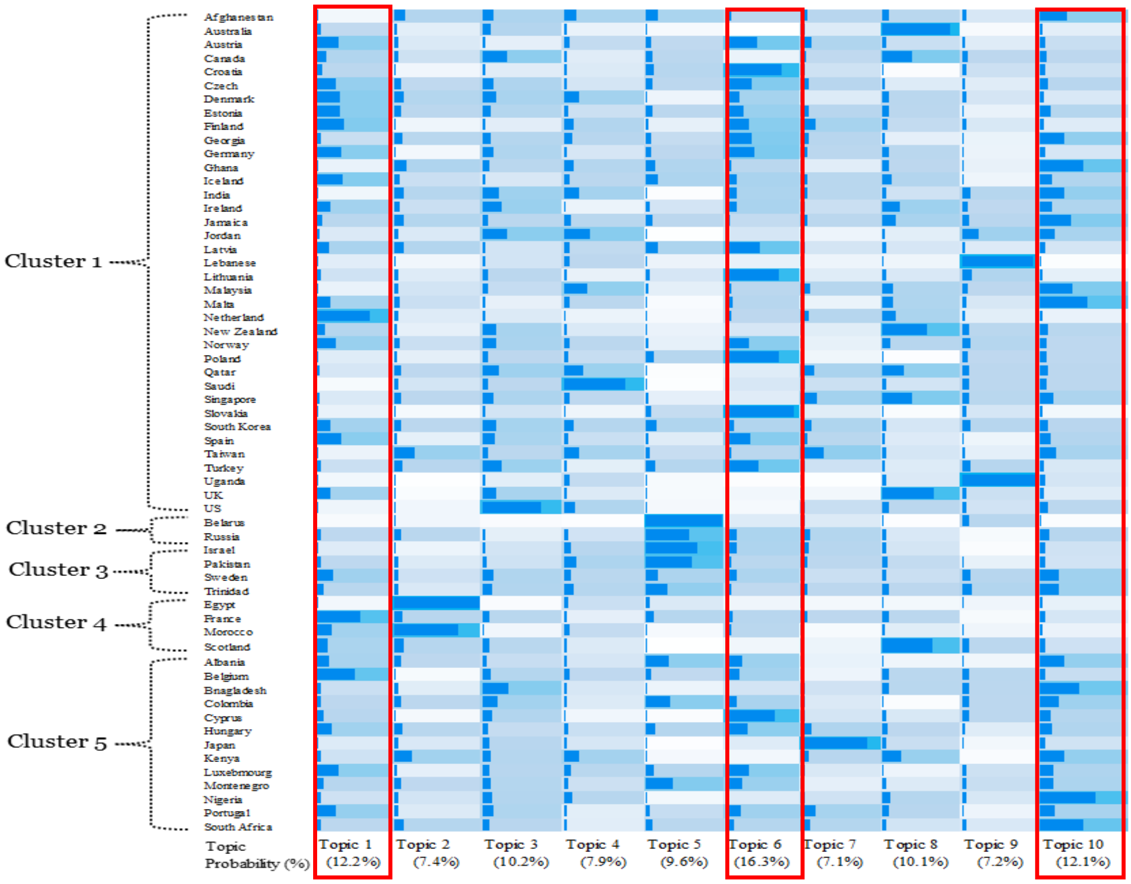

Figure 3. Topic's Distribution over NCSs

For the final step of this study we aimed to calculate per-cluster topic's distribution to understand which of the latent topics are appeared more frequently in the hierarchical clusters computed in Figure 1. Overall, Figure 4 denotes topic's distribution over the hierarchical clusters at a high level that can be used as an input for further statistical analysis. Our results indicated that most of topics identified in Table 5 have been appeared in both clusters 1 and 5 except topic 5, which is related to cyber legislation and laws. These findings are consistent with extant literatures as members of these clusters are mostly among EU countries with a considerable number of cybercrime laws and legislations governed by EU (Schmitt, 2013). As a result, topic 5 doesn't seem to be a great concern for the EU members. In contrast, we found that topic 5 is closely associated with cluster 2 members Russia and Belarus. Since both countries are not EU members, development of cybercrime laws and legislation can be a considerable stride in combat against cyber criminals and organized cyber cartels. Similarly, Topic 2, *cybersecurity organization,* is the most probable topic in cluster 4. In this cluster, France (Liveri, 2014) expresses its desire to become world power in the area of cybersecurity or Egypt invests on other organizations or sectors (e.g. telecommunication and postal services) as one of the main pillars for achieving its cyber strategy objectives.

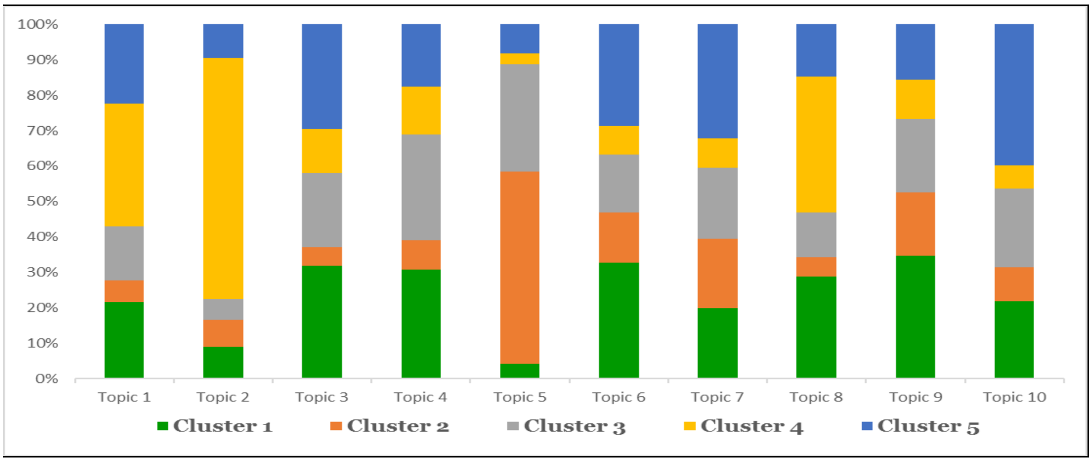

Figure 4. Topic's Distribution over the hierarchical clusters

Discussion and Conclusion

Development and implementation of effective NCS is an ambitious project for many nation states. Since the formulation of NCSs has been started recently, a comparison analysis of NCSs is common among national government and particularly policy makers. Thus a systematic assessment of NCS considering its consistency and harmonization with other strategies can be an ongoing challenge. In this study, we examined NCSs from institutional theories to determine whether membership in regional or international cybersecurity institutions can be determinant for development of proximate NCSs. In this study we examined 60 NCSs by applying machine learning approaches such as hierarchical clustering and topic modelling techniques. Previous studies have been used a limited number of metrics or indicators for such analysis. Our results pinpointed that quantitative analytical method such as LDA and clustering can be called during the analysis of qualitative data such as textual policies, strategies and legislations to get a bigger picture and insights during the formulation of NCS. We also noted that approach could be regarded as a complimentary approach to assist policy makers for better identification of topics that are neglected or not covered appropriately.

The result of our clustering method helped us for a better understanding of the overall similarities between NCSs. Our results suggested that members of an institution like NATO or like-minded allies have developed more integrated and harmonized NCS. We also noted that coercive forces of international cybersecurity institutions are appeared during the development of NCSs, other factors such as political and geographical imperatives can be determinant for convergence between NCSs. One of the limitations of this study is that not all the governments have published their NCSs to the public websites and some others are not translated into English. Besides, we couldn't find a globally accepted best practices that can be compared with the result of our topic modelling and topic labels. Among many possibilities for future studies, other theoretical lenses in politics, power, and economics can be considered for examining of NCSs.

References

Backhouse, J., Hsu, C. W., & Silva, L. (2006). “Circuits of power in creating de jure standards: shaping an international information systems security standard,” *MIS QUARTERLY*, 413-438.

Blei, D. M. (2012). “Probabilistic topic models,” *Communications of the ACM, 55*(4), 77-84.

Blei, D. M., & Lafferty, J. D. (2007). “A correlated topic model of science,” *The Annals of Applied Statistics*, 17-35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). “Latent dirichlet allocation,” *Journal of machine Learning research, 3*(Jan), 993-1022.

Borcard, D., Gillet, F., & Legendre, P. (2011). “Introduction Numerical Ecology with R,” (pp. 1-7): Springer.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). “Reading tea leaves: How humans interpret topic models,” Paper presented at the Advances in neural information processing systems.

Chen, H., & Zhao, J. L. (2015). ISTopic: “Understanding Information Systems Research through Topic Models,” *Available at SSRN 2601719*

Choucri, N., Madnick, S., & Ferwerda, J. (2014). “Institutions for Cyber Security: International Responses and Global Imperatives,” *Information Technology for Development, 20*(2), 96-121.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). “Indexing by latent semantic analysis,” *Journal of the American society for information science, 41*(6), 391.

Dunn, M. (2005). “*A comparative analysis of cybersecurity initiatives worldwide*,” Paper presented at the WSIS Thematic meeting on Cybersecurity, Geneva.

Greene, D., O’Callaghan, D., & Cunningham, P. (2014). “How many topics? stability analysis for topic models,” Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Hall, P. A., & Taylor, R. C. (1996). “Political science and the three new institutionalisms,” *Political studies, 44*(5), 936-957.

Hofmann, T. (1999). “Probabilistic latent semantic indexing,” Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Huang, A. (2008).“Similarity measures for text document clustering,” Paper presented at the Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). “Data clustering: a review,” *ACM Computing Surveys (CSUR), 31*(3), 264-323.

Klimburg, A. (2012). “National cyber security framework manual,” NATO Cooperative Cyber Defense Center of Excellence.

Kolini, F., & Janczewski, L. (2015). “Cyber Defense Capability Model: A Foundation Taxonomy,” Paper presented at the CONF-IRM 2015 Proceedings paper 32, Canada

Lehto, M. (2013). “The ways, means and ends in cyber security strategies,” Paper presented at the Proceedings of the 12th european conference on information warfare and security (Jyväskylä, 2013), Academic Publishing, Reading.

Leydesdorff, L., & Zaal, R. (1988). “Co-words and citations relations between document sets and environments,”

Lin, F.-r., Hao, D., & Liao, D. (2016). *Automatic Content Analysis of Media Framing by Text Mining Techniques.* Paper presented at the 2016 49th Hawaii International Conference on System Sciences (HICSS).

Liveri, D., Sarri, A. (2014). “An evaluation framework for national cybersecurity strategy,” *ENISA*

Luiijf, E., Besseling, K., & De Graaf, P. (2013). “Nineteen national cyber security strategies,” *International Journal of Critical Infrastructures 6, 9*(1-2), 3-31.

Luiijf, H., Besseling, K., Spoelstra, M., & De Graaf, P. (2011). “Ten national cyber security strategies: A comparison,” Paper presented at the International Workshop on Critical Information Infrastructures Security.

McCallum, A. (1999). “Multi-label text classification with a mixture model trained by EM,” Paper presented at the AAAI’99 workshop on text learning.

Moe, T. (1984). The New Economics of Organization American Journal of Political Science. 28: 739—77.. 1987. “An Assessment of the Positive Theory of Congressional Dominance,” *Legislative Studies Quarterly, 12*, 475-500.

Nolasco, D., & Oliveira, J. (2016). "Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data*,"* Paper presented at the 2016 49th Hawaii International Conference on System Sciences (HICSS).

Patton, R. M., Beaver, J. M., & Potok, T. E. (2011). "Classification of Distributed Data Using Topic Modeling and Maximum Variation Sampling," Paper presented at the System Sciences (HICSS), 2011 44th Hawaii International Conference on.

Powell, W. W., & DiMaggio, P. J. (2012). "The new institutionalism in organizational analysis,": University of Chicago Press.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 134-148

Ramanathan, V., & Wechsler, H. (2013). "Phishing detection and impersonated entity discovery using Conditional Random Field and Latent Dirichlet Allocation," *Computers & Security, 34*, 123-139.

Reich, S. (2000). "The four faces of institutionalism: Public policy and a pluralistic perspective," *Governance, 13*(4), 501-522.

Schmitt, M. N. 2013. "Tallinn manual on the international law applicable to cyber warfare," Cambridge University Press

Shafqat, N., & Masood, A. (2016). "Comparative Analysis of Various National Cyber Security Strategies," *International Journal of Computer Science and Information Security, 14*(1), 129.

Shahbaznezhad, H., & Tripathi, A. (2016). "The Art of Listening on Social Media Platforms: How Firms Follow Users on Social Media Fan Pages," Paper presented at the Twenty-Fourth European Conference on Information Systems (ECIS), İstanbul, Turkey.

Shi, Z., Lee, G. M., & Whinston, A. B. (2015). "Towards a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence," *Management Information Systems Quarterly (MISQ), Forthcoming*

Small, H., Sweeney, E., & Greenlee, E. (1985). "Clustering the Science Citation Index using co-citations," II. Mapping science. *Scientometrics, 8*(5-6), 321-340.

Sneath, P. H., & Sokal, R. R. (1973). "Numerical taxonomy," *The principles and practice of numerical classification*.

Steinmo, S., & Thelen, K. (1992). "Structuring politics: historical institutionalism in comparative analysis," Cambridge University Press.

Stevens, T. (2012). "A cyberwar of ideas? Deterrence and norms in cyberspace," *Contemporary Security Policy, 33*(1), 148-170.

Strehl, A., Ghosh, J., & Mooney, R. (2000). "Impact of similarity measures on web-page clustering," Paper presented at the Workshop on Artificial Intelligence for Web Search (AAAI 2000).

Syed, R., & Dhillon, G. (2015. "Dynamics of Data Breaches in Online Social Networks: Understanding Threats to Organizational Information Security Reputation," Paper presented at the ICIS USA

Vakulenko, S., Müller, O., & Brocke, J. v. (2014). "Enriching iTunes App Store categories via topic modeling," Paper presented at the ICIS, Auckland, New Zealand.

Wamala, F. (2011). "ITU National Cyber Security Strategy Guide," International Telecommunication Union, Sep 2011.

Ward Jr, J. H. (1963). "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association, 58*(301), 236-244.

Wei, X., & Croft, W. B. (2006). "LDA-based document models for ad-hoc retrieval," Paper presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.

Willett, P. (1988). "Recent trends in hierarchic document clustering: a critical review," *Information Processing & Management, 24*(5), 577-597.