Temporal Understanding of Cybersecurity Threats

Jennifer Sleeman
Computer Science and Electrical Eng.
Univ. of Maryland, Baltimore County
Baltimore. MD 21250 USA
jsleem1@umbc.edu

Tim Finin
Computer Science and Electrical Eng.
Univ. of Maryland, Baltimore County
Baltimore. MD 21250 USA
finin@umbc.edu

Milton Halem
Computer Science and Electrical Eng.
Univ. of Maryland, Baltimore County
Baltimore. MD 21250 USA
halem@umbc.edu

***Abstract*—As cybersecurity-related threats continue to increase, understanding how the field is changing over time can give insight into combating new threats and understanding historical events. We show how to apply dynamic topic models to a set of cybersecurity documents to understand how the concepts found in them are changing over time. We correlate two different data sets, the first relates to specific exploits and the second relates to cybersecurity research. We use Wikipedia concepts to provide a basis for performing concept phrase extraction and show how using concepts to provide context improves the quality of the topic model. We represent the results of the dynamic topic model as a knowledge graph that could be used for inference or information discovery.**

***Index Terms*—Cybersecurity, Knowledge Graph, Topic Modeling, Dynamic Topic Modeling**

I. Introduction

Cybersecurity is an important computing area that is vital to our society due to the rise in cyber attacks and the damage that they can do [1]. Computer exploits such as malware are often hard to identify before they have been executed. Most work in this area that uses language understanding technology has focused on extracting entities and relations without temporal analysis. As computer exploits grow in sophistication, understanding past exploits and how they are evolving could provide insight into potentially new exploits before they occur.

Work by Mittal et al. [2] described the importance of the temporal dimension in understanding cybersecurity exploits, such as a 2015 attack against GitHub that lasted 72 hours [3]. Often, it is not until the attack occurs that the information that led to it can be analyzed. However, in many cases, the attacker makes multiple attempts to execute the attack or spends time on a victim's system before detection [4].

We address the issue of temporal analysis by applying Dynamic Topic Models (DTM) to two document collections: 17 years of malware reports from Symantec and 20 years of Cybersecurity research papers from the ArXiv repository. We identify changes in trends over time for each and correlate the two models to show how the research changes in the second collection are influenced by new malware described in the first.

Among specialized fields such as cybersecurity, natural language understanding is challenged by the dependence upon specific terminology and jargon that is significantly present in text documents used for knowledge extraction. This terminology is hard to manage due to the heavy use of both acronyms and multi-word phrases, whose meanings give significant context as phrases rather than a bag of words. To address this issue, we extract common cybersecurity concepts from Wikipedia data.

For the cybersecurity domain, learning DTMs is still an unexplored area of research and performing cross-domain analysis between multiple data sets is also unexplored. We put forth this work to show how temporal analysis by means of cross-domain understanding can be applied to cybersecurity.

II. Background

As the Internet and its use in everyday tasks has become ubiquitous, so has cybersecurity-related crimes. Cybersecurity attacks can be divided into those involving software, hardware, and networks [5] and can attempt to exploit any combination of confidentiality, integrity, and availability [6]. Common attacks include the following [7].

- Man in the middle attacks that intercept communication;
- Brute force attacks that obtain protected information;
- Denial of service attacks that flood networks preventing access;
- Phishing attacks that steal information using deception;
- Social engineering attacks that manipulate users to obtain access to information; and
- Malware attacks that compromise data or resource integrity, confidentiality and/or availability.

A. Dynamic Topic Models

Dynamic Topic Modeling (DTM) [8] provides a means for performing topic modeling over time. Internally using Latent Dirichlet Allocation (LDA) [9], it creates a topic per time slice. By using a state space model, DTM links topic and topic proportions across models to 'evolve' the models over time. The early work of Blei et al. [8] modeled the evolution of a large collection of Science articles and showed topic evolution for specific scientific concepts of interest consistent with historical understanding. DTM has been used in a number of applications, including science research [8], software [10], finance [11], music [12], and climate change [13], [14] to understand how particular domains have changed over time.

III. Related Work

Work by Joshi et al. [15] used information from cyber-based attacks and exploits to process unstructured text and to generate RDF linked data that could then be used for identifying vulnerabilities. Though there are similarities to our approach, i.e., converting unstructured text to a graph-based representation, our method treats documents as mixture models enabling improved similarity detection among documents. Also, our method includes temporal analysis of documents. Matthews et al. [16] built on this work to develop a full intrusion detection system that used machine learning but addresses different issues.

More recent research by Prakash et al. [17], [18] has an approach similar to our approach in that their method is generative. They use a method called propagation-based models to represent malware trends by using phrases that provide contextual constraints to help identify malware attacks. Their goal differs from ours, which uses models as a way of gathering information about historical events and current research, that could then be used to support systems that do such predictions.

In the work by Kolini et al. [19], topic modeling was used for processing national cybersecurity strategies (NCS) documents in addition to hierarchical clustering. They used the topics as a means for finding the themes among the NCS documents. Their topic analysis includes using human annotators. Our work differs in that we are evaluating how concepts are changing over time by means of a dynamic topic model.

IV. Approach

We use Dynamic Topic Models (DTM) to evolve topics over time in a data collection. A key innovation to our method is the use of Wikipedia concepts to provide domain context for the preprocessing of documents. Typically bag-of-words are used for methods such as topic modeling, but we have found that using domain concepts that include domain-relevant phrases and specifically looking for those concepts during document preprocessing, improves topic modeling results [20].

Automatically extracting domain concepts from a text collection is a challenging problem. An important contribution of this work is our approach to automatic domain concept extraction using Wikipedia concepts. We exploit Wikipedia concepts related to cybersecurity as a context model for training the DTM. Since Wikipedia concepts can be mapped to concepts in DBpedia [21], Wikidata [22] and other background knowledge resources, we use knowledge graphs during the modeling process, enabling the results of this work to be used for additional query and inference.

With this context in place, we preemptively search for mentions of cybersecurity concepts in the text of each document before standard text processing methods are used, such as stop word removal, low-frequency removal and lexical-based processing. In addition to the Wikipedia-based concept search, standard text processing is applied to find other words of interest subject to stop word removal, low-frequency removal and lexical-based processing.

We generate one large vocabulary file across all time slices, a file which defines the words for each document, and a file which indicates to DTM how many files exist per time slice. We then use DTM to generate the topic model. The whole process is governed by a knowledge graph which is created and updated as documents in the repository are processed. The knowledge graph is populated as the DTM learns latent topics over time as shown in Figure 1.

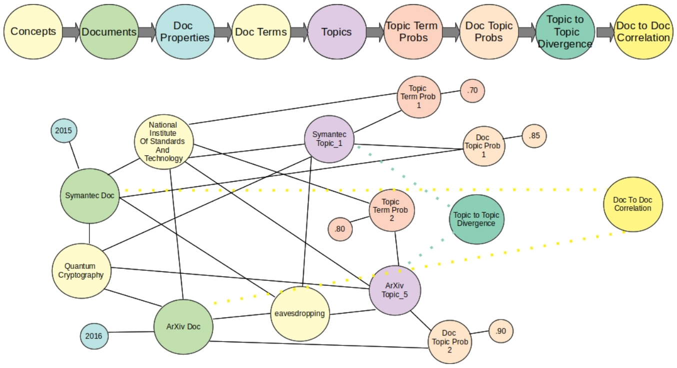

Fig. 1. Knowledge Graph Construction of Dynamic Topic Modeling.

TABLE I
Example Wikipedia Concept Terms Used

Example Concept
cryptanalysis
cryptographic protocol
cryptographic software
cryptography
cryptosystem & cryptovirology
cyber-insurance
cyber-security regulation
cyber security standards
cyber self-defense
cyberattack & cybercrime
cyberspace
cyberterrorism
cyberwarfare

Initially the knowledge graph encapsulates the concepts from Wikipedia for the specific domain. As more documents are processed, it includes graphs of the documents in the collection(s) and their properties, including their discovered concepts. During the topic modeling process, graphs of the topics and topic probabilities are added to the knowledge graph. For cybersecurity this is particular useful when looking for documents with common exploit properties.

A. Extracting Knowledge from Unstructured Text

We captured 3,836 total concepts from Wikipedia which were used to establish the context for the topic modeling portion. We started with the concept phrases ‘cybersecurity’, ‘computer security’, ‘cyber security’, and ‘cyber’. For each phrase, we retrieved Wikipedia pages, then for each page retrieved the outgoing links found on that page. We perform a one-level traversal to formulate the concept list. The longer this list of concepts, the longer the preprocessing time. When we increased the traversal to three levels, processing time doubled. Example concepts are shown in Table I.

TABLE II
EXAMPLE WIKIPEDIA CONCEPT TERMS USED

Example Concept	*Acronym*
australian information security association	aisa
advanced encryption standard	aes
denial of service	dos
department of homeland security	dhs

When we perform the preprocessing step, which is the step that finds the word and word phrases in the text and accumulate their frequencies, we then treat the acronym for a word phrase as if that word phrase was found and increase the frequency of the word phrase. Automatic acronym generation is challenged by the fact that acronyms can be confused with stop words, and acronyms can also be polysemous. For example, a simple Wikidata query shows that 32 entities have an English name or alias matching ”CPA”. We use a general heuristic for extracting acronyms from phrases. This heuristic takes the first letter of each word in the phrase (excluding pronouns) and generates the acronym from the combination of these letters. We captured 245 cybersecurity acronym lookups based on the cybersecurity Wikipedia concepts. Example acronyms based on concepts are shown in Table II. From this list one can see that our heuristic for choosing acronyms works well for finding common acronyms, however more complex acronyms are harder to identify with a heuristic.

B. Topic Models Over Time

In this work we used Dynamic Topic Models to model multiple document collections over time. The output from preprocessing is a set of files that are required by DTM including the vocabulary file representing the vocabulary for the document collection, a file which indicates the number of time slices, and a file that defines the words and frequencies for each document. We then use DTM to generate the dynamic topic model for each document collection.

C. Knowledge Graph Creation and Use

The goal of this work is to provide a tool for an end user to use for obtaining information related to large data collections. By generating knowledge graphs that represent the documents, their words and frequencies, the topics and topic pairs, and by grounding all of this by the Wikipedia concepts, the knowledge acquired from this processing can be used for other applications such as search or by providing a temporal cognitive query component [23].

V. CYBERSECURITY DATA SETS

We evaluated two different data sets over time for this work. The first is a collection of research papers from ArXiv that are categorized by their authors as relevant to Cryptography and Security. This document set tends on average to be larger in size. The second is a collection of Symantec Malware reports that are uncategorized. This document set tends on average to be smaller in size. We describe each of these data sets in more detail below.

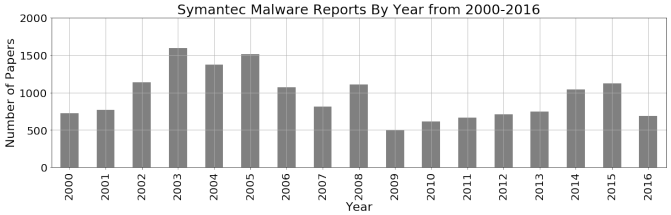

Fig. 2. Symantec Malware Reports from 2000-2017 Data Distribution by Year.

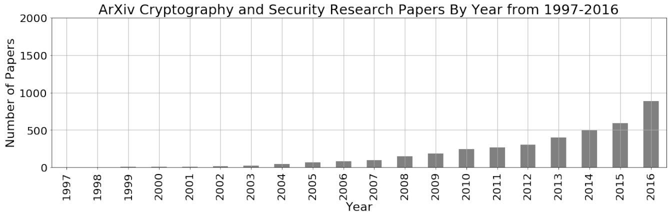

Fig. 3. Cryptography and Security ArXiv Papers From 1997-2016 Data Distribution by Year.

A. Symantec Malware Data Set

The Symantec Malware reports data set is a set of reports which describe malware incidents and actions to take to combat the malware. They are typically relatively short, with an average of about nine KB size or approximately 4,500 words. There are a total of 16,167 files with a total size of approximate 149.6 MB. This dataset spans 17 years from the year 2000 to the year 2016. This data is unlabeled, however in each report there is text that indicates the *type* of report with the types ’trojan’, ’worm’ and ’virus’ as the highest types present across reports. We show the distribution by year in Figure 2.

B. ArXiv Data Set

The ArXiv Cryptography and Security data set is a set of research papers related to cryptography and security. They are typically longer in size, with an average of about 50 KB size or approximately 20,000 words. There are a total of 3,913 files with a total size of approximate 215.5 MB. All of the documents are categorized as ‘Cryptography and Security’ but in addition to this base category, many are further tagged by the other categories, with about 94 categories in total. This dataset spans 20 years from the year 1997 to the year 2016 as shown in Figure 3. The number of ArXiv papers are increasing steadily over time and the Symantec Malware Reports are strongly correlated to exploit events.

VI. EXPERIMENTS AND ANALYSIS

To assess the value of our approach, we performed two different experiments. One experiment assesses the improvement in the quality of the topics using concept phrases which are obtained from Wikipedia. A second experiment measures how the dynamic topics can be used in understanding how the cybersecurity domain is changing over time.

A. Concept Context Experiment

In this experiment we tested how the topic model performs given the Wikipedia provided concepts as the context for the models in comparison with a standard bag of words model without a predefined context. We used the ArXiv dataset, labeled by category. We split the dataset into a train and test set (60/40 split). We then built two separate topic models, one which has the concept-provided context and the second which does not use a context but rather a standard bag of words. We performed the similarity portion of the experiment without the temporal component in order to isolate strictly the concept context portion of this work. We evaluate document similarity by grouping documents by subcategory type.

We generated topic models for the ArXiv dataset with and without the concept context using the training set and then used the test set for evaluation. We measured similarity defined by the average probabilities across documents for each topic. Figure 4 shows two heat maps comparing the ArXiv topic models with and without concepts. For visualization, we show a subset of documents from the train and test sets, organized by their paper subcategories. The heat map shows subcategory paper similarity between the train and test documents with darker cells indicating stronger similarity.

We can observe two things from the heat maps. For certain subcategories of papers, we see stronger similarity between train and test in the concept model, such as the Mathematics Graph Theory train test section in the heat maps. Sometimes this difference in similarity is subtle, but still evident, as in the Mathematics Number Theory train test section in the heat maps. Small differences can indicate significant improvement in topic word probabilities that are correlated to a given set of documents. As we will show in Tables III and IV, the top ten topic word probabilities tend to be more human-understandable with concept context models. Secondly, the heat maps show that with the concept model, there were fewer incidents of test documents with similarities that were the same across all subcategories, as indicated by a horizontal or vertical band. For those documents, the concept model was better at distinguishing between subcategories.

When comparing the DTM models using 100 topics and the full datasets, the first difference we observed is the vocabulary size. With the concept model, the size of the vocabulary is 13,016, and the vocabulary size for the non-concept model is 11,824. Between the two models, as presented in Table III, there are two real differences. In the concept model, topics tend to contain word phrases based on the Wikipedia-built concept model. This affects the topics because when a phrase is broken into a sequence of words, other interesting words become less probable. For example, *quantum cryptography* was a concept defined by our knowledge graph. The word *photon* is seen in the top ten most probable words, but it is not in the top ten most likely words for the topic without context concepts.

The second difference is since the model is guided by the domain concepts, words that are harder to interpret because they are more generic move farther towards the tails of the probability distribution. For example in Table IV, *intrusion detection* is a concept specific to the domain and found in the text. In the topic model without concepts, *intrusion detection* is never found. The only topic that mentions *intrusion* and *detection* contains words that are more generic and less specific to the domain.

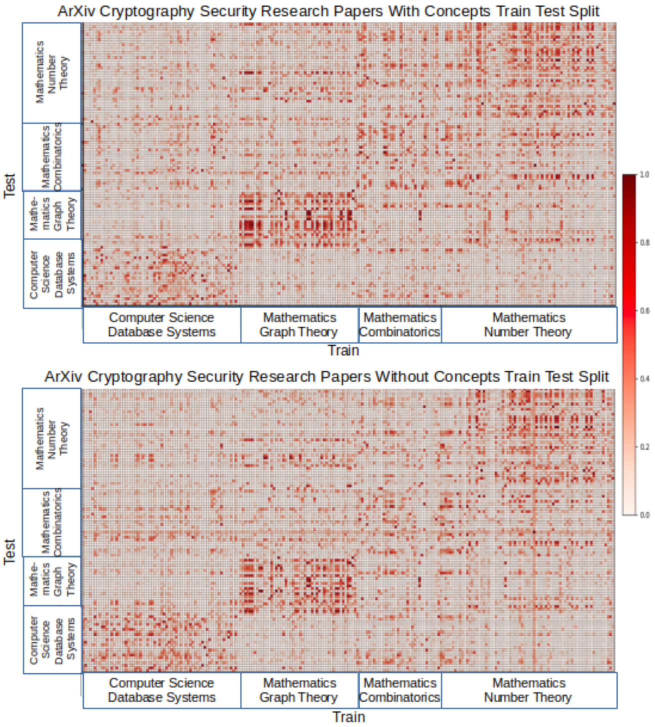

Fig. 4. ArXiv Train/Test Split Heat Maps.

TABLE III
THE ARXIV TOPIC MODEL WITH AND WITHOUT CONTEXT CONCEPTS

Year	*Topic With Concept*	*Topic Without Concept*
2000	quantum cryptography, phase, photon, cryptography, measurement, channel, system, eavesdropping, stage, polarization	quantum, state, communication, phase, cryptography, channel, eavesdropping, protocol, error, polarization

B. Dynamic Model Experiment

In this experiment we measured how reliable the model is for understanding specific concept evolution. We build the

TABLE IV
THE ARXIV TOPIC MODEL WITH AND WITHOUT CONTEXT CONCEPTS

Year	*Topic With Concept*	*Topic Without Concept*
2000	intrusion detection, universal, taxonomy, intrusion detection system, based, payload, classification, input, attack, alert	cell, network, intrusion, parameter, system, information, detection, method, space, approach

dynamic topic model for the ArXiv dataset and the Symantec Malware dataset. We compare topic evolution for a range of 20 to 100 topics using intervals of 20. We also experimented with the variance hyperparameter in DTM which controls how much time variance is allowed among topics, however, we used a low variance of .05.

When we build topic models over time, topics evolve over time based on the documents in the collection at that time point. Our observation is as we increased these number of topics we saw more granularity among topics. Topics represented more narrow mixtures. We also observed concepts that drop off of one topic and fall into another at various points in time. Deciding on the number of topics can be based on a measure called perplexity, which approximately gives a good estimate. Another approach is to observe how the topics change visually as the number of topics increases and choosing the total number of topics based on the requirements of the problem being solved. For cybersecurity we suggest using a larger number of topics for more granularity. Since discovery is a big part of the cybersecurity process, i.e. looking for information based on past exploits, the more granular the model, the more information can be inferred. However, we have found since concepts can be represented in different topics simultaneously over time due to the co-occurrence of the concept with other words, plotting the concept over time across all topics provides a visual trend of how the concept is changing over time.

Using the concept *malware* as a use case and observing all topics for a given model, Figures 5 show how the probability for the word malware changes over time. As the number of topics is increased the significant spike at year 2009 remains prominent across models for a given topic. In addition, there is a second trend that is increasing as it approaches 2016 which is more prominent as the number of topics is increased.

According to Wikipedia, the first-ever *malware* was detected on February 16, 2006 [24]. The models reflect a rise in probabilities from approximately 2005 to the spike in 2009, however as we obtain more granular details using 100 topics, it is observed that the secondary trend dominates around 2011.

In Tables V and VI, we show topics that correlate to the two trends that spike early then drop off in Figure 5 100 topics. The model begins to capture the reference of *malware* in the top 10 most probable terms for these two trends. The first trend which has a smaller spike starting around 2006 shows association of *malware* with how to protect computing devices using *antivirus software*. The second trend which has a more pronounced spike relates to mobile computing and *malware*. The *malware* presence in the top 10 most probable terms is also observed in the models created with 60 topics and 80 topics.

In the 100-topic model, the third trend observed which has a moderate spike around 2008 but then steeply increases from 2011 onward is shown in Table VII related to new types of attacks.

Though the first official *malware* detection was in 2006, in the 100-topic model, as shown in Table VIII, *malware* was in

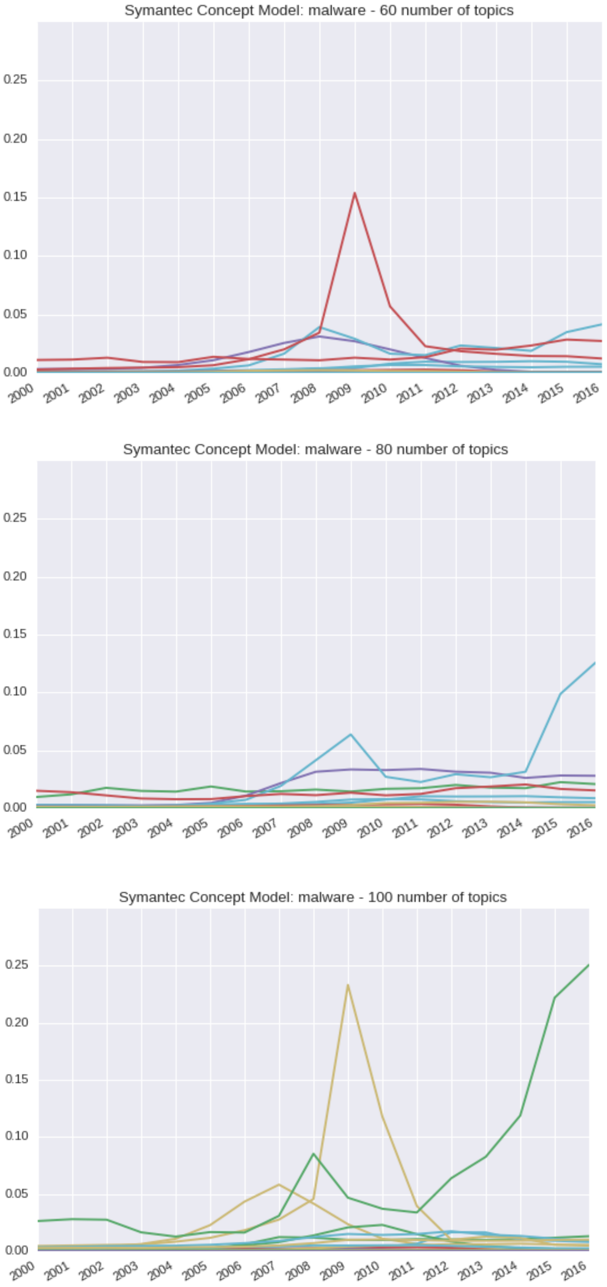

Fig. 5. 60-, 80-, and 100-Topic Symantec Reports DTM Concept: Malware

TABLE V
100-Topic Symantec Malware Report Dynamic Topic Model
Topic 5

Year	*Top 10 Most Relevant Term*
2006	remove, protection, threat, antivirus software, packed, file, malware, symantec, security, window
2007	remove, protection, threat, antivirus software, packed, file, malware, symantec, security, window
2008	protection, remove, threat, malware, packed, file, symantec, antivirus software, security, window, protection, packed, threat, file, symantec, remove, malware, window, antivirus software, security

TABLE VI
100-TOPIC SYMANTEC MALWARE REPORT DYNAMIC TOPIC MODEL TOPIC 41

2008	privacy, commander, doctor, malware, picture, movie, action, video, multi, surveillance
2009	malware, doctor, privacy, commander, action, android, intent, picture, movie, video
2010	malware, action, doctor, privacy, android, intent, commander, bluetooth, picture, provider
2011	action, android, intent, privacy malware, doctor, bluetooth, commander, picture, wifi

TABLE VII
100-TOPIC SYMANTEC MALWARE REPORT DYNAMIC TOPIC MODEL TOPIC 86

2007	malicious, component, info, scanner, attacker, rootkit, door, remote, malware, computer
2008	info, malicious, malware, scanner, attacker, rootkit, component, door, computer, remote
2009	malicious, info, attacker, malware, door, scanner, rootkit, remote, computer, component
2010	malicious, info, attacker, malware, door, computer, remote, scanner, rootkit, based
2011	info, malicious, attacker, malware, computer, door, remote, reputation, dropped, based
2012	malicious, info, attacker, malware, computer, door, remote, reputation, threat, dropper
2013	malicious, info, malware, attacker, computer, remote, door, dropper, reputation, back
2014	malicious, malware, info, attacker, computer, dropper, remote, door, reputation, back
2015	malware, malicious, info, dropper, attacker, computer, remote, payload, dropped, door
2016	malware, malicious, info, dropper, attacker, computer, remote, payload, dropped, door

the top 10 most probable words for a particular topic in the year 2000. Indeed, a Symantec Malware report did specifically reference *malware* with a timestamp of 2000. In fact, this report details an exploit titled *Infostealer* which was found on *December 8, 1997*.

When observing the same trend information for the ArXiv Cryptography and Security research papers, as shown in Figure 6, there is a spike among one topic in particular at year 2009 for 60 topics. This spike is more prominent at 2008 given 80 topics and prominent at 2009 given 100 topics. In each there appears to be a dip and then another rise as it reaches 2016.

For the ArXiv Cryptography and Security research papers, *malware* is found in the top 10 probable words by year 2007 as shown in Table IX.

TABLE VIII
100-TOPIC SYMANTEC MALWARE REPORT DYNAMIC TOPIC MODEL TOPIC 86

Year	*Top 10 Most Relevant Terms*
2000	component, computer, attacker, malicious, dropper, kernel, remote, malware, door, author
2001	component, attacker, computer, malicious, remote, dropper, door, malware, kernel, security
2002	component, attacker, malicious, door, dropper, computer, remote, kernel, malware, rootkit

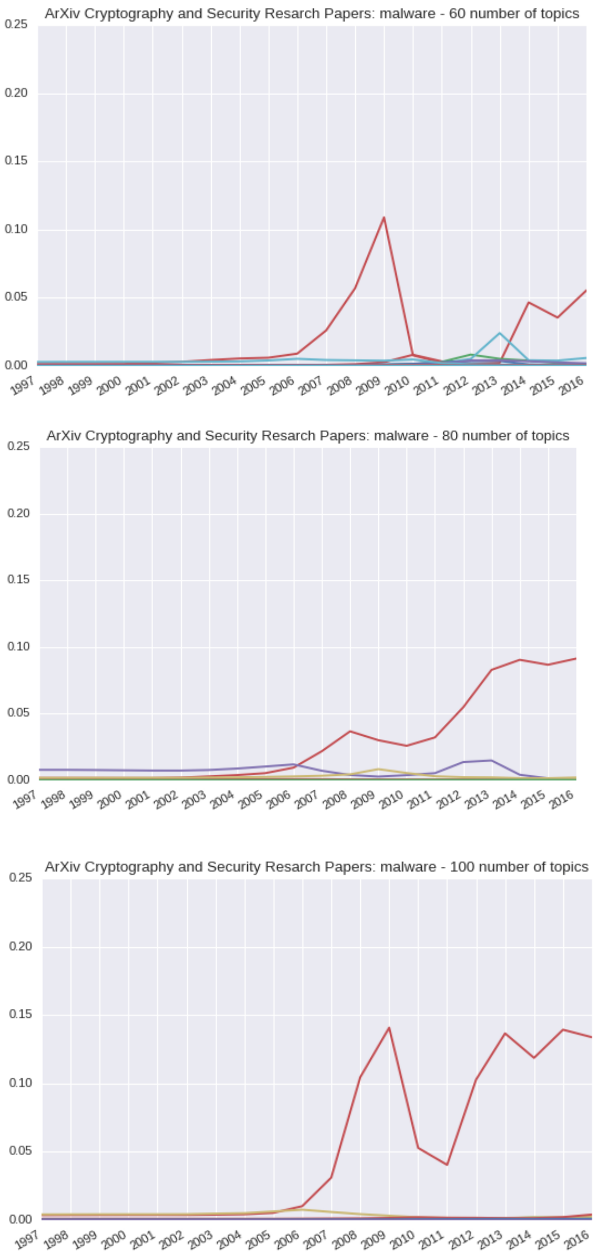

Fig. 6. 60-, 80-, and 100-Topic ArXiv Papers DTM Concept: Malware

TABLE IX
100-TOPIC ARXIV CRYPTOGRAPHY AND SECURITY RESEARCH PAPERS DYNAMIC TOPIC MODEL TOPIC 3

Year	*Top 10 Most Relevant Terms*
2007	infected, epidemic, virus, malware, infection, worm, internet, spreading
2008	wireless, spread, malware, worm, infected, infection, spreading, virus
2009	spread, internet, propagation, epidemic, malware, worm, infected, infection, virus, propagation, host, internet, simulation, spread

TABLE X
100-TOPIC ARXIV CRYPTOGRAPHY AND SECURITY RESEARCH PAPERS DYNAMIC TOPIC MODEL TOPIC 3 YEARS 2012-2016

Year	*Top 10 Most Relevant Terms*
2012	malware, infection, infected, virus, worm, malicious, host, behavior, spread, parameter
2013	malware, malicious, behavior, worm, infected, sample, call, email, file, family
2014	malware, similarity, sample, infected, behavior, infection, malicious, family, based, type
2015	malware, sample, behavior, spreading, family, infected, based, malicious, virus, benign
2016	malware, sample, virus, family, malicious, benign, infected, infection, based, anti

By the year 2012, for Topic 3 malware was the highest probable word as shown in Table X.

However, since the ArXiv dataset is a collection of research papers rather than a set of detailed exploit reports, the effects are less pronounced. Also, there tends to be a delay in concepts reflected in research papers due to the time to perform research and write the paper. As opposed to the reports which are quick, on point, and of a more urgent nature.

Given this type of analysis, one could use these models to understand how long it takes to build momentum among published work for a particular attack type. This could potentially be used to identify years that may have the most relevant work for a given attack type.

VII. CONCLUSION

Cybersecurity threats are increasing. As systems evolve, new vulnerabilities are discovered because threats are also evolving and developing new attack strategies, leading to the creation of new categories of attacks. Tools to help model and understand such trends in cybersecurity threats and attacks are useful in helping combat them. In this work we have provided insight into how to use dynamic topic models for cybersecurity documents to understand how the concepts found among documents are changing over time. We have demonstrated an approach that uses an ontology of cybersecurity concepts extracted from Wikipedia to extract phrases that can improve the readability of the topics and provide better human understanding of the topics. We represent the results of the dynamic topic model as a knowledge graph that could be used for inference or information discovery.

REFERENCES

[1] Symantec, “Internet security threats report,” http://www.symantec.com-/threatreport/.

[2] S. Mittal, P. K. Das, V. Mulwad, A. Joshi, and T. Finin, “Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities,” in *IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining*. IEEE Press, 2016, pp. 860–867.

[3] M. Nestor, “GitHub has been under a continuous DDOS attack in the last 72 hours,” http://goo.gl/KBucR0.

[4] S. N. Narayanan, A. Ganesan, K. Joshi, T. Oates, A. Joshi, and T. Finin, “Early detection of cybersecurity threats using collaborative cognition,” in *4th Int. Conf. on Collaboration and Internet Computing*. IEEE, 2018, pp. 354–363.

[5] J. Jang-Jaccard and S. Nepal, “A survey of emerging threats in cybersecurity,” *Journal of Computer and System Sciences*, vol. 80, no. 5, pp. 973–993, 2014.

[6] S. E. Goodman and H. S. Lin, Eds., *Toward a safer and more secure cyberspace*. National Academies Press, 2007.

[7] A. Bendovschi, “Cyber-attacks–trends, patterns and security countermeasures,” *Procedia Economics and Finance*, vol. 28, pp. 24–31, 2015.

[8] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *23rd Int. Conf. on Machine learning*. ACM, 2006, pp. 113–120.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[10] J. Hu, X. Sun, D. Lo, and B. Li, “Modeling the evolution of development topics using dynamic topic models,” in *22nd Int. Conf. on Software Analysis, Evolution, and Reengineering*. IEEE, 2015, pp. 3–12.

[11] T. Morimoto and Y. Kawasaki, “Forecasting financial market volatility using a dynamic topic model,” *Asia-Pacific Financial Markets*, vol. 24, no. 3, pp. 149–167, 2017.

[12] U. Shalit, D. Weinshall, and G. Chechik, “Modeling musical influence with topic models,” in *Int. Conf. on Machine Learning*, 2013.

[13] J. Sleeman, M. Halem, T. Finin, and M. Cane, “Dynamic topic modeling to infer the influence of research citations on IPCC assessment reports,” in *Big Data Challenges, Research, and Technologies in the Earth and Planetary Sciences Workshop*. IEEE, 2016.

[14] ——, “Modeling the evolution of climate change assessment research using dynamic topic models and cross-domain divergence maps,” in *AAAI Spring Symposium on AI for Social Good*. AAAI Press, 2017.

[15] A. Joshi, R. Lal, T. Finin, and A. Joshi, “Extracting cybersecurity related linked data from text,” in *2013 IEEE Seventh International Conference on Semantic Computing*. IEEE, 2013, pp. 252–259.

[16] M. L. Mathews, A. Joshi, and T. Finin, “Detecting botnets using a collaborative situational-aware IDPS,” in *Int. Conf. on Information Systems Security and Privacy*. SciTePress, 2016.

[17] C. Kang, N. Park, B. A. Prakash, E. Serra, and V. Subrahmanian, “Ensemble models for data-driven prediction of malware infections,” in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 2016, pp. 583–592.

[18] B. A. Prakash, “Prediction using propagation: From flu trends to cybersecurity,” *IEEE Intelligent Systems*, vol. 31, no. 1, pp. 84–88, 2016.

[19] F. Kolini and L. J. Janczewski, “Clustering and topic modelling: A new approach for analysis of national cyber security strategies.” in *Pacific Asia Conference on Information Systems*, 2017.

[20] J. Sleeman, T. Finin, M. Halem *et al.*, “Ontology-grounded topic modeling for climate science research,” *Emerging Topics in Semantic Technologies. ISWC 2018 Satellite Events*, 2018.

[21] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, “Dbpedia: A nucleus for a web of open data,” in *The semantic web*. Springer, 2007, pp. 722–735.

[22] C. H. Song, D. Lawrie, T. Finin, and J. Mayfield, “Gazetteer generation for neural named entity recognition,” in *Florida Artificial Intelligence Research Symposium*, May 2020.

[23] R. Bordawekar and O. Shmueli, “Using word embedding to enable semantic queries in relational databases,” in *1st Workshop on Data Management for End-to-End Machine Learning*. ACM, 2017.

[24] Wikipedia contributors, “Timeline of computer viruses and worms — Wikipedia, the free encyclopedia,” 2019, [Online; accessed 10-April-2019]. [Online]. Available: https://en.wikipedia.org/?curid=174761