Appendix A. Data Extraction and Processing Procedure

We processed the data from three forums involving two Chinese forums and one English forum. Although the languages are different, we followed the same data processing procedures for all the forums. The main objectives of the data processing were as follows:

- Identify cybersecurity-relevant posts; knowledge provision posts; knowledge acquisition posts; money-involved posts.
- Count the occurrence of cybersecurity keywords in each post
- Calculate similarity with cybersecurity professional reports

The theoretical underpinning for our classification scheme can be summarized as follows:

(1) Cybersecurity-relevant posts: A hack forum is a technical discussion board for users interested in hacking techniques. However, there still exist plenty of posts in hack forums unrelated to cybersecurity techniques with no risk of violations of CMA. In addition to cybersecurity, they also discuss other things such as general computer skills, career, hobbies, and so on. According to our classification, the number of cybersecurity-related posts account for 26.1%, 40.0%, and 60.3% for the two Chinese forums and English forum, respectively. Therefore, although cybersecurity-relevant leading posts capture a unique aspect of a hack forum, they are not the general theme of all the leading posts in these top-ranked hack forums. Therefore, the classification of cybersecurity-relevant posts plays a vital role in examining the chilling effect of CMA enforcement.

(2) Knowledge acquisition and provision posts: Our classification of knowledge acquisition and knowledge provision is consistent with the literature that recognizes knowledge acquisition and provision as distinct activities with different motivations (Welser et al. 2007, Zhang et al. 2007, Hwang et al. 2015). In addition, prior studies (Rogers 2006, Zhang et al. 2015, Park et al. 2018) have specifically focused on knowledge exchange behaviors among hackers. For example, messages posted in a hack forum were extracted and classified into knowledge acquisition (e.g., questions, requests), knowledge provision (e.g., answers, tutorials), or neither based on a support vector machine (SVM) algorithm (Rogers 2006, Zhang et al. 2015, Park et al. 2018), which is similar to our approach. There is a growing body of studies on hacker motivations and incentives in which knowledge acquisition is identified as one of important motivations among hackers (Thomas 2005, Rogers 2006, Zhang et al. 2015, Park et al. 2018). Thus, we believe that our classification of knowledge acquisition and provision is well-founded in the literature.

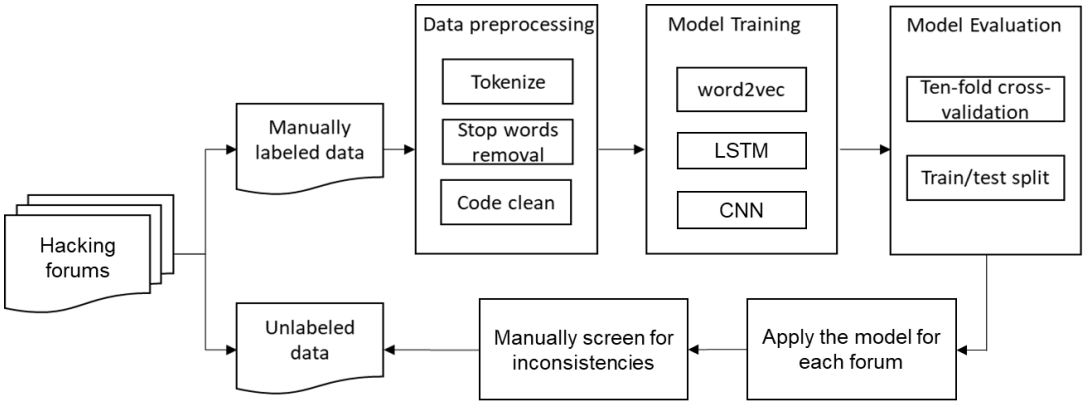
(3) Money-involved posts: The distinct role of intrinsic and extrinsic motivation has been broadly recognized in the literature. Today more and more hackers are motivated by financial incentives (Kshetri 2010), which represent a very strong form of extrinsic motivation. Therefore, it is worthwhile to study the money-involved posts as a separate category of posts in a hack forum. For a similar reason, prior studies have also tried to identify money-involved posts in a hack forum (Mikhaylov and Frank 2016, Siu et al. 2021). For example, Siu et al. (2021) focused on currency exchange behaviors in an underground hack forum and identified the types of activities discussed by those involved in currency exchange (illicit or not criminal). In our context, with money-involved posts, contributors' intent is easier to judge

because these posts involve financial incentives and motivations, leading to relatively lower uncertainty of false penalization. Money-involved posts are particularly targeted by CMA enforcement if hacking tools or activities are also involved in the transaction. Therefore, given one of the key mechanisms of the chilling effect is increased uncertainty of false prosecution under CMA enforcement, money-involved lawful posts are less likely to be chilled by CMA enforcement.

A1. Procedures for identifying cybersecurity-relevant posts

We developed a three-step process to identify cybersecurity-relevant posts, knowledge provision posts, knowledge acquisition posts, and money-involved posts. First, we constructed two representative sample sets for the manual labeling of posts in the Chinese and English hack forums. The manually labeled data became the training and testing data sets of the LSTM and CNN model. Second, we trained a Long Short-Term Memory (LSTM) model (Rao and Spasojevic 2016) and a Convolutional Neural Network (CNN) model (Kim 2014) for classification of posts based on word embeddings. According to Kim (2014), the CNNs are trained on top of pretrained word vectors for sentence-level classification tasks with little hyperparameter tuning and static vectors but achieve excellent results compared with frequently used recurrent neural network (RNN) architecture such as LSTM. We subsequently deployed the trained models onto the full text sample to classify posts based on the four dimensions described in the paper (Table A1). Third, we aggregated the predictive results of the LSTM and CNN models and resolved the inconsistencies with manual screening, which resulted in final labels for all posts in the hack forums. Details of the text classification process are presented in Figure A1.

Figure A1. Hack forum text classification process



Step1. Manual Labeling

Because the manually labeled data would become the training dataset of the LSTM and CNN models, constructing a representative training set would determine the overall representation power of our trained model and directly influence its predictive power. To ensure the quality of the manual labels while balancing the cost, we limited the initial scope of the sample selection process to all leading posts coming from the three hack forums. From the establishment of the forums until March 2011, Chinese hack forum users contributed 188,025 (161,052 from Hackbase and 26,973 from 2cto) leading posts, and English hack forum users contributed 101,660 leading posts. We randomly selected 25% of all leading posts (based on unique post ID) in each particular year, which yielded 47,006 unique Chinese leading posts and 25,415 unique English leading posts. We randomly divided all the

selected text records into blocks. Each block contained 2,000 text records. In total, we constructed a sample set of 47,006 text records from the Chinese hack forums and a sample set of 25,415 text records from the English hack forum. The block segmentation was designed to facilitate the subsequent manual labeling process.

The manual labeling process involved two Chinese coders independently labeling 40,263 leading posts out of the 161,052 leading posts in Hackbase and 6,743 leading posts out of the 26,973 leading posts in 2cto and two native English-speaking coders independently labeling 25,415 leading posts out of the 101,660 leading posts in Hackforums.net. All coders — three postgraduate and one senior undergraduate — were majors in information systems and received more than six months of training, during which they gained domain knowledge in information security and hacker communities through course and projects before working on labeling. To ensure independence between coders, we used the block segmentation of the sample data according to the method of Yue et al. (2019). Each coder only received one block of the data at any given time, and we made sure the blocks given to them in the same time period were different across coders. Each coder would need to label through all the text records. After all coders in the Chinese or English groups had completed the labeling of all text records, we compared their results and identified inconsistent posts. We further asked the coders to resolve the initial inconsistent labels independently to see if it was because of unintended mistakes. Their interrater agreement, using kappa statistics, was 0.85 for Hackbase, 0.91 for 2cto, and 0.83 for English forum, which suggests sufficient interrater reliability (Hallgren 2012). After consulting with cybersecurity professionals, we obtained the finalized screening results of 47,006 Chinese posts and 25,415 English posts for machine learning.

Table A1. Content Classification Schema

Cybersecurity-relevant Posts
The cybersecurity relevance of a post is decided based on explicit or implicit references or discussions about cybersecurity techniques and/or any related subtopics mentioned in the title, questions, and answers of a post. For example, a post on how to use/develop cybersecurity tools/functions/frameworks (e.g., login/authentication) is considered cybersecurity-relevant. However, discussing a non-security task (e.g., integrating third-party APIs) is not considered cybersecurity-relevant.
Knowledge Acquisition Posts
Posts that include questions, doubts, requests, advice seeking, and anything else reflecting a user's request for information.
Knowledge Provision Posts
Posts that include answers, tutorials, teaching, troubleshooting, guidelines, demonstrations, and anything reflecting the provision of information.
Money-Involved Posts
Posts concerning buying or selling a product or a service, exchanging currency, or any other behavior related to transactions.

Table A2. Classification Examples of Cybersecurity-relevant, Knowledge Acquisition, Knowledge Provision, and Money-involved Posts

Cybersecurity-relevant Posts
Positive Examples (posts labeled as cybersecurity-relevant posts)

English Post (id= 493511), “I would like to know how to protect myself from attacks from other people, including DDoS attacks. Do I need to get on a VPN, etc? I need to know the necessary steps to make sure i'm not able to be attacked). This may be due to the difference in nature between attack and protection. While attacks are normally associated with specific ports, cybersecurity protection is system-wide”

Chinese Post (id=36288), “I would like to ask what kind of SQL to download can invade a database of the network 7.0 and LeadBBS v3.14 version! If I am wrong, please forgive me, please explain it to me.”

Negative Examples (posts NOT labeled as cybersecurity-relevant posts)

English Post (id= 9447), “I have Firefox 2 0 0 11 can someone find out how to put a proxy please”

Chinese Post (id=252819), “The solution to the problem that Win 7 desktop background cannot be changed”

Knowledge Acquisition Posts

Positive Examples (posts labeled as knowledge acquisition posts)

English Post (id= 2480791), “Hello everyone, Just wondering, is there any RAT, logger, etc... that starts up as a service, or are they all just using the more commonly used HKLM/HKCU startup method (that I see on every single HF program)? Just curios, thanks.
IMG[images/smilies/smile.gif]***IMG***”

Chinese Post (id=38665), “Help me, please. Why my Rising Antivirus can’t be upgraded. Once I click on upgrade. it says I have upgraded too many times”

Negative Examples (posts NOT labeled as knowledge acquisition posts)

English Post (id= 5059576), “If you need fud i will crypt for you ! autoit so non dependent ! :)”

Chinese Post (id=36502), “Microsoft's most valuable expert Nuo Yan's blog was hacked!!
It is the so-called hacker organization called BlackPeon. The hacker organization said that it found that there is a bug in the blog hall that can directly enter the admin management.”

Knowledge Provision Posts

Positive Examples (posts labeled as knowledge provision posts)

English Post (id= 3702352), “anyone need help setting up? darkcomet or anything just post your skype and ill get back to you asap to help you”

Chinese Post (id=3579873), “A solution to the problem of not being able to open remote desktop (share)”

Negative Examples (posts NOT labeled as knowledge provision posts)

English Post (id= 2494716**)**, “hey guys i just want to have a keylogger on my laptop so everytime someone logs on any site which requires to enter their user and pass it autosaves on my computer without knowing by the user. If u know any please reply to this post. Thanks in Advance! :) more power to ya'll”

Chinese Post (id=36363), “Please help me find a software named as what kind of eyes, what kind of demon knife...something like that. I remember hackbase has...but I can't find it. I have seen it posted once in the forum. But I can’t find the post. I hope everyone can help me.
Very urgent! Waiting online!!!”

Money-Involved Posts

Positive Examples (posts labeled as money-involved posts)

English Post (id= 4875310), “I need to get 1,000 likes on a page. does anyone know a way to help? willing to pay for 1,000 likes”

Chinese Post (id=23514), “Anyone help me find my email password and I will give him 20 QQ coins. How about this! My email address. This email address I need urgently. I'm online waiting for the ah. QQ: 165328019”

Negative Examples (posts NOT labeled as money-involved posts)

English Post (id= 1326592), “Hello all, I want to test my blackshades rat. I know i can test it at my own but i already tryd some things, I want to try more on victims so im asking could somebody give me 1 or 2 victims for free. Should be very great.”

Chinese Post (id=2812163), “Good news! Tencent is celebrating 6th anniversary now, 6 digit QQ number can be applied for free. Apply for it soon!”

Step 2. Machine Classification

We first conducted a standard preprocessing step of tokenizing raw text before applying the deep-learning algorithms for text classification. For English, the tokenization process involves punctuation splitting and separation of some affixes like possessives. Because of the major difference between Chinese and Western languages is at the lexical level, Chinese language requires more extensive token pre-processing, commonly referred to as Chinese word segmentation (Wong et al. 2009). We split Chinese text into a sequence of words, which is defined according to the Chinese word segmenter provided in the FudanNLP toolkit (Qiu et al. 2013). Before word segmentation, stop words and useless punctuation in this classification task were removed. As of now, numerous stop word lists have been developed for English language. These stop word lists are traditionally derived through frequency analysis of the entire lexicon within a large corpus (Yang 1995). Unlike in the English language, the Chinese language lacks a universally accepted stop word list. Although some studies on Chinese information retrieval utilize manually constructed stop word lists (Du et al. 2000, Chen and Chen 2001, Nakagawa et al. 2004), others may automatically generate stop word lists. We adopted the most comprehensive collection of stop words for the Chinese language[1] and evaluated the effectiveness of the lists according to Zou et al. (2006). In addition, special format texts (e.g., scripts, website, time stamps, numeric strings, etc.) were replaced by uniform codes. Prior experiments (Zou et al. 2006) have demonstrated that segmentation based on an effective stop word list significantly outperforms segmentation without a stop word list. When constructing word embedding vectors, we used the publicly available word2vec vectors that have been trained on 100 billion words from Google News for English words and ngram2vec[2] word representations that were learned upon multiple corpora of Chinese Wikipedia, Baidubaike, People’s Daily News, Sogou News, Zhihu QA, etc. for Chinese words (Zhao et al. 2017). These vectors, with a dimensionality of 300, were trained by using the continuous bag-of-words architecture (Mikolov et al. 2013). Words absent from the set of

[1] https://github.com/stopwords-iso/stopwords-zh

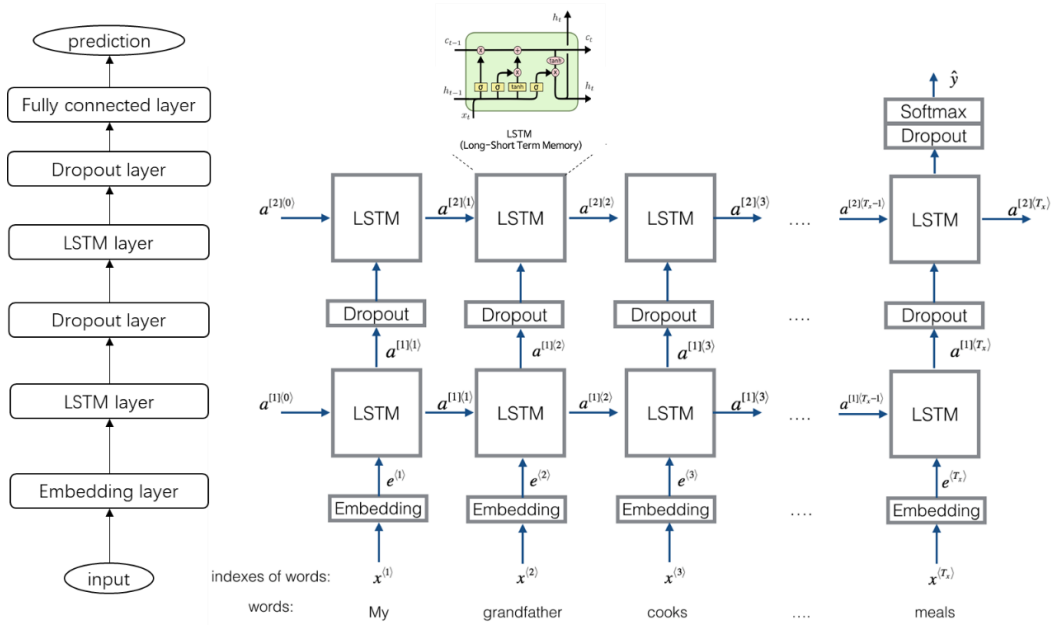
[2] Ngram2vec toolkit is a superset of word2vec and fasttext toolkit in which arbitrary context features and models are supported. Please refer to https://github.com/Embedding/Chinese-Word-Vectors

pretrained words were initialized randomly. To accommodate the convolution operation in the subsequent CNN model, we fixed the length[3] of the model input sequences for each post at 200.[4] According to the statistics, the average length of a Chinese leading post in our sample is around 67 words, and the maximum length is 915 words, while the corresponding numbers for an English leading post are 45 and 1,088. In some extreme cases, authors never used any punctuation in one post because it includes long codes and scripts. We preprocessed the text with codes and scripts by replacing them with formatted labels "CODES" to reduce the length of posts. Similarity, we also replaced the site URL with formatted labels "LINKS," time stamps with "DATETIME," etc. We balanced the potential loss of information against computing complexity and chose a fixed post length of 200 words. Hence, if the length of a post exceeds 200 words, the overly long section will be cut off. If the length of a post is less than 200 words, the missing section will be padded with 0. Python was used as programming language. The models of LSTM and CNN were built by a TensorFlow deep learning framework.

The LSTM model architecture shown in Figure A2 uses word embeddings coupled with LSTM units for the purpose of contextual text classification. The first layer, the embedding layer, transforms positive integer indices from the input into dense real-valued vectors of fixed sizes. The second layer in the network is composed of multiple LSTM units that include four primary components: input gate, self-recurrent connection, forget gate, and output gate. Dropout is a regularization technique to avoid overfitting, which is accomplished by randomly omitting a fraction of the units from coadaptation. The fully connected layer, as the name suggests, is characterized by comprehensive connections to all activations from the preceding layer. Based on this framework, we adopted the binary cross entropy function as a cost function. The number of hidden nodes in the LSTM is 128 and Dropout value set to 0.5. The learning rate is set to 0.001. The model was trained with sigmoid activation and the Adam optimizer. We picked a batch size of 64, which gives relative high accuracy with a relatively fast training speed.

Figure A2. Structure of LSTM (Rao and Spasojevic 2016)

[3] The length of a leading post in our context is not a simple count of Chinese characters, but the number of words after text segmentation. For example, in Chinese, a word could consist of two or more characters.



The CNN model architecture shown in Figure A3 follows the design of Kim (2014). The first layer, known as the embedding layer, executes the function of mapping vocabulary word indices into low-dimensional vector representations. The next layer uses multiple filter sizes to perform convolutions over the embedded word vectors. Three convolution operations, each with a filter size of 3×200, 4×200, and 5×200, were applied to each post representation separately to generate three different sets of features. The number of filters per filter size is 128. Next, we applied max-pooling to convert the convolutional layer's output into a long feature vector, added dropout regularization, and classified the result by using a softmax layer. The training of the model was conducted via stochastic gradient descent over shuffled minibatches, utilizing the Adadelta update rule. For model hyperparameters, rectified linear units were used, filter windows (h) of 3, 4, 5 with 128 feature maps each, a dropout rate of 0.5, an L2 constraint of 3, and a minibatch size of 64. These specific values were determined via a grid search on the SST-2 development set.

Ten-fold cross validation was applied during the training process for both the LSTM and CNN models. In addition, we conducted a hold-out evaluation method to evaluate the performance of our model. Before training, we split the manually labeled 40,360 Chinese posts and 21,000 English posts randomly into training (80%) and test sets (20%). We used accuracy, precision, recall, and F1 scores to evaluate the classification performance on the test set. The overall classification accuracy on cybersecurity-relevant posts in the English forum reached 89.85% for the LSTM model and 93.29% for the CNN model. The figures for Chinese hack forum are 91.54% for the LSTM model and 94.16% for the CNN model. Table A3 reports the text classification performance for the previously mentioned four classification dimensions of the English and Chinese posts. Finally, for each hack forum, the trained LSTM and CNN models were applied to labeling all the remaining posts.

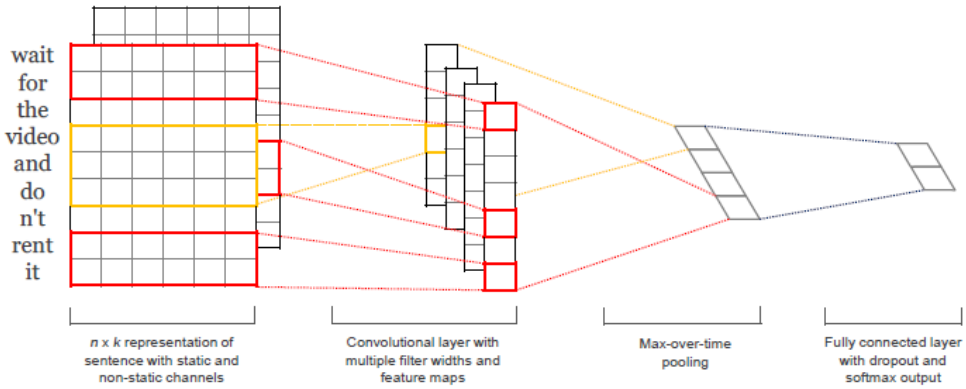


Figure A3. Structure of CNN (Kim 2014)

Table A3. Classification Performance

			Cybersecurity -relevant or not	Knowledge provision or not	Knowledge acquisition or not	Money-involved or not
LSTM	**English posts**	Precision	0.8875	0.9142	0.7731	0.5770
		Recall	0.9526	0.9254	0.8541	0.9198
		Accuracy	0.8985	0.8902	0.8945	0.9284
	Chinese posts	Precision	0.8473	0.9278	0.8668	0.7520
		Recall	0.9315	0.9203	0.9361	0.9323
		Accuracy	0.9154	0.9078	0.9122	0.9081
CNN	**English posts**	Precision	0.9207	0.9621	0.7542	0.5863
		Recall	0.9724	0.8894	0.9230	0.9875
		Accuracy	0.9329	0.9010	0.8996	0.9327
	Chinese posts	Precision	0.9049	0.9500	0.9019	0.8815
		Recall	0.9353	0.9013	0.8697	0.9287
		Accuracy	0.9416	0.9110	0.9048	0.9519

Step 3. Manually Screen for Inconsistencies

We compared the predictive results of the LSTM and CNN models and identified inconsistent classification labels. Consequently, we got 10,564 inconsistently classified Chinese posts and 4,778 inconsistently classified English posts. We further asked the Chinese and English coders to independently screen all the inconsistent classifications produced by the machines following the same procedure as described in Step 1. The final inconsistency rates for the English and Chinese forums after manual screening were all less than 1%. The posts that could not be consistently classified after manual screening were treated as negative outcomes (i.e., cybersecurity-irrelevant). Ultimately, we obtained finalized classification results for all posts.

A2. Procedures for counting the occurrence of cybersecurity keywords

In a bid to augment the specificity of our analysis, we further performed keyword extraction and measured the extent of intensive use of cybersecurity keywords in each post based on multiple glossaries of cybersecurity terminology. For all the posts in the full data set, all the stop words, underlinings, punctuation, numbers, and spaces in each post were removed first. Then the N-gram

stemming procedure was implemented on the content of each post, where N={1, 2, 3}. All the stemmed words were then recompiled into a space-delimited text, serving the purpose of keyword extraction (Porter 1980). We adopted a knowledge-based approach to extract security keywords in each post based on the following three collections of keywords commonly used in cybersecurity industry. The first source (noted as "src1") is the glossary of cybersecurity terms designed by the cybersecurity training institute SANS[5] and the cybersecurity agency NICCS.[6] The second source (noted as "src2") is a dictionary of information security terms, abbreviations and acronyms by Calder and Watkins (Calder and Watkins 2007). The third source (noted as "src3") is a catalog of cybersecurity vulnerabilities publicly disclosed since 1999 by the Common Vulnerabilities and Exposures (CVE) program.[7] For each keyword identified from the CVE data, we conducted multiple rounds of manual screening to pick up only security-related keywords. The vulnerability keywords extracted from CVE records by year were matched with the posts generated in the same year. All the terminologies from the three sources were translated into Chinese by Chinese cybersecurity professionals to accommodate for keyword extraction from Chinese hack forums. The three terminology dictionaries captured different aspects of the emergence of new cybersecurity techniques and increases in cybersecurity threats. Each word in the posts was compared with the three dictionary files, respectively, and the number of keywords in each post corresponding to each terminology source was calculated as measures of the extent of relevance to cybersecurity in hack forum discussions.

A3. Procedures for calculating similarity to cybersecurity professionals' reports

We measured the similarity of each leading post compared with contemporaneous content generated by cybersecurity professionals. We collected two sources of cybersecurity archives, the Diary Archive provided by the SANS Internet Storm Center (Angrist and Pischke 2008)[8] and the Internet Security Threat Reports published by Symantec.[9] We extracted *the daily handler diaries* from the SANS ISC according to a data structure defined as [date, author, thread_id, title, comment_num, content] and aggregated them on a monthly basis. Symantec annually published the internet threat reports. So, we just downloaded all the reports published during our study period from the Broadcom security center's archived publications. For each leading post, we calculated the document-level semantic

[5] https://www.sans.org/security-resources/glossary-of-terms/

[6] https://niccs.us-cert.gov/about-niccs/glossary

[7] https://cve.mitre.org/data/downloads/index.html

[8] https://isc.sans.edu/diaryarchive.html

[9] https://www.broadcom.com/support/security-center/publications/archive

similarity with the diaries generated in the same month or the threat reports generated in the same year based on the term frequency–inverse document frequency (TF-IDF) method (Salton et al. 1975).

To be specific, we quantified a word in a piece of a document by attaching a weight to each word that signifies the importance of the word in the document and corpus. We vectorized each piece of document on a vocab composed of a list of all possible words in the cybersecurity professional corpus (i.e., the two sources of cybersecurity archives), which generated a fixed-length feature representation of the document. The words in the vocab were also translated into Chinese by Chinese cybersecurity professionals to accommodate calculation of similarity with posts in Chinese hack forums. By vectorizing each document, we could compute the cosine similarity between each hack forum post and the contemporaneous reports created by cybersecurity professionals. The formula for cosine similarity is shown below; it considers vector orientation, independent of the vector magnitude. If the cosine similarity score equals 0, it indicates the content from different sources does not share any common word. In another extreme case, the cosine similarity score equals 1 if all the words from content generated by cybersecurity professionals can be retrieved in a hack forum post. The normalized cosine similarity value ranges from 0 to 1, both inclusive.

$$cossim(u, v) = \frac{u \cdot v}{\|u\|\|v\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \quad \text{(A-1)}$$

Appendix B. Supplementary Tables and Figures

Table B1. Legislation against Production/Distribution/Possession of Computer Misuse Tools[10]

Country	Law	Enforcement date	Amendment
Australia	Criminal Code Act 1995 (Cth)	01/03/2013	ss 478.3 and 478.4
Croatia	The New Criminal Law	01/01/2013	Article 272
Canada	Protecting Canadians from Online Crime Act	09/12/2014	Section 342.2
China	Criminal Code	28/02/2009	Article 285
Colombia	Penal Code Act 1273 of 2009	29/01/2009	Article 269A-J
Ethiopia	Telecom Fraud Offence Proclamation	04/09/2012	Article 3
Fiji	Crimes Decree 2009	01/02/2010	Article 346
France	Monetary and Financial Code	15/07/2009	Article L163-4
Germany	German Criminal Code	05/11/2007	Acts 202c
Italy	Penal Code	18/03/2008	Art 615
Netherlands	Computer Crime II Act	01/09/2006	Article 139d
New Zealand	Crimes Amendment Act 2013 (2013 No 27)	01/07/2013	subsection 1 of 251
Qatar	Cybercrime Law (No. 14 of 2014)	02/10/2014	Article 66
Russia	Criminal Code	30/12/2008	Act 273 and 138.1
Serbia	Criminal Code	31/08/2009[11]	Article 304a
Singapore	Computer Misuse and Cybersecurity Act	13/03/2013	Article 10(1)
Sweden	Criminal Code	01/07/2013	Article 9b
Switzerland	Swiss Criminal Code	01/01/2012	Article 144bis

[10] This table was compiled by the authors based on extensive collection and research from the online cyberlaw resources of each country.

[11] http://ilo.org/dyn/natlex/natlex4.detail?p_lang=en&p_isn=84244

United Kingdom	Computer Misuse Act	22/11/2006	s3A

Table B2. Summary Statistics of Content Features at Leading Post-level

Hackbase (Chinese Forum)					
Variable	**Obs.**	**Mean**	**Std. Dev.**	**Min**	**Max**
num_replies	140689	8.4843	60.6802	0	7065
Is_cybersecurity_relevant	140689	0.2607	0.4390	0	1
Is_knowledge_acquisition	140689	0.6868	0.4638	0	1
Is_knowledge_provision	140689	0.2588	0.4380	0	1
Is_money_involved	140689	0.0048	0.0694	0	1
src1_keyword_counts_per_post	140689	1.1221	1.9874	0	44
src2_keyword_counts_per_post	140689	2.1393	3.3969	0	56
src3_keyword_counts_per_post	140689	1.5159	2.9795	0	42
2cto (Chinese Forum)					
Variable	**Obs.**	**Mean**	**Std. Dev.**	**Min**	**Max**
num_replies	10393	6.0290	52.8821	0	4763
Is_cybersecurity_relevant	10393	0.3997	0.4899	0	1
Is_knowledge_acquisition	10393	0.8633	0.3436	0	1
Is_knowledge_provision	10393	0.1334	0.3400	0	1
Is_money_involved	10393	0.0008	0.0277	0	1
src1_keyword_counts_per_post	10393	7.7606	7.9760	0	56
src2_keyword_counts_per_post	10393	13.9486	12.6708	0	75
src3_keyword_counts_per_post	10393	11.0580	9.8735	0	70
Hackforums.net (English Forum)					
Variable	**Obs.**	**Mean**	**Std. Dev.**	**Min**	**Max**
num_replies	101660	12.2551	40.6075	0	2571
Is_cybersecurity_relevant	101660	0.6033	0.4892	0	1
Is_knowledge_acquisition	101660	0.6796	0.4666	0	1
Is_knowledge_provision	101660	0.2659	0.4418	0	1
Is_money_involved	101660	0.0873	0.2823	0	1
src1_keyword_counts_per_post	101660	3.1458	4.8208	0	107
src2_keyword_counts_per_post	101660	7.0416	8.1340	0	157
src3_keyword_counts_per_post	101660	2.9917	4.5179	0	72

Table B3. Changes in Outcomes after CMA

Variables (per user per week)	Period	English Forum Mean	Chinese Forums Mean	English Forum % drop after CMA	Chinese Forums % drop after CMA
num_security_posts	Before CMA	0.0369	0.0012	55%	83%
	After CMA	0.0166	0.0002		
perc secposts	Before CMA	0.0317	0.0012	56%	83%
	After CMA	0.0141	0.0002		
src1_keyword_counts	Before CMA	0.0545	0.0033	51%	85%
	After CMA	0.0265	0.0005		
src2_keyword_counts	Before CMA	0.09133	0.0049	53%	84%
	After CMA	0.04336	0.0008		
src3_keyword_counts	Before CMA	0.05694	0.0038	54%	84%
	After CMA	0.02642	0.0006		

Table B4. Correlation Matrix of Outcome Variables

Treatment = 0 (obs=377,939)					**Treatment = 1 (obs=8,229,412)**				
Variables (per user per week)	(1)	(2)	(3)	(4)	Variables (per user per week)	(1)	(2)	(3)	(4)
(1) num_security_posts					(1) num_security_posts				
(2) perc_secposts	0.8647				(2) perc_secposts	0.8522			
(3) src1_keyword_counts	0.7031	0.6847			(3) src1_keyword_counts	0.4858	0.4152		
(4) src2_keyword_counts	0.7685	0.7802	0.9293		(4) src2_keyword_counts	0.4657	0.3962	0.8618	
(5) src3_keyword_counts	0.7112	0.7072	0.8917	0.9312	(5) src3_keyword_counts	0.4437	0.3571	0.8107	0.8634

Table B5a. Descriptions of Variables in Weekly Panel Data

Variable	Description
Dependent variables (per user in week *t* in logarithmic form)	
num_security_posts (log)	The number of cybersecurity-relevant leading posts
perc secpost	The percentage of cybersecurity-relevant leading posts
src1_keyword_counts (log)	The average number of source 1 cybersecurity keywords in leading posts
src2_keyword_counts (log)	The average number of source 2 cybersecurity keywords in leading posts
src3_keyword_counts (log)	The average number of source 3 cybersecurity keywords in leading posts
ISC_similarity	Average cosine similarity scores of leading posts with ISC diaries
Symantec_similarity	Average cosine similarity scores of leading posts with Symantec threat reports
num_nonsecurity_posts (log)	The number of cybersecurity-irrelevant leading posts
num_money involved_posts (log)	The number of money-involved leading posts
Independent variables (per user in week *t*-1 in logarithmic form)	
auth_life (log)	The time interval since the first posting time or the registration time (if exist)
cum_kp_rpl_to_sec (log)	Cumulative number of knowledge provision replies to cybersecurity-relevant leading posts
cum_kp_posts (log)	Cumulative number of knowledge provision leading posts
cum_leading_posts (log)	Cumulative number of leading posts
cum_mon_rpl_to_sec (log)	Cumulative number of money-involved replies to cybersecurity-relevant leading posts
cum_mon_posts (log)	Cumulative number of money-involved leading posts
cum_replies (log)	Cumulative number of replies
cum_replies_to_sec (log)	Cumulative number of replies to cybersecurity-relevant leading posts
cum_ka_rpl_to_sec (log)	Cumulative number of knowledge acquisition replies to cybersecurity-relevant leading posts
cum_ka_posts (log)	Cumulative number of knowledge acquisition leading posts
cum_sec_rpl_to_sec (log)	Cumulative number of cybersecurity-relevant replies to cybersecurity-relevant leading posts
cum_security_posts (log)	Cumulative number of cybersecurity-relevant leading posts
cum_ISCsimilarity (log)	Cumulative average cosine similarity scores of cybersecurity-relevant leading posts with ISC diaries
cum_ISCsimi_rpl_to_sec (log)	Cumulative average cosine similarity scores of replies to cybersecurity-relevant leading posts with ISC diaries
cum_Symantecsimilarity (log)	Cumulative average cosine similarity scores of cybersecurity-relevant leading posts with Symantec threat reports
cum_Synsimi_rpl_to_sec (log)	Cumulative average cosine similarity scores of replies to cybersecurity-relevant leading posts with Symantec threat reports
cum_src1_keyword_counts (log)	Cumulative average number of source 1 cybersecurity keywords in leading posts
cum_src1_rpl_to_sec (log)	Cumulative average number of source 1 cybersecurity keywords in replies to cybersecurity-relevant leading posts
cum_src2_keyword_counts (log)	Cumulative average number of source 2 cybersecurity keywords in leading posts
cum_src2_rpl_to_sec (log)	Cumulative average number of source 2 cybersecurity keywords in replies to cybersecurity-relevant leading posts
cum_src3_keyword_counts (log)	Cumulative average number of source 3 cybersecurity keywords in leading posts

cum_src3_rpl_to_sec (log)	Cumulative average number of source 3 cybersecurity keywords in replies to cybersecurity-relevant leading posts
cum_nonsec_posts (log)	Cumulative number of cybersecurity-irrelevant leading posts

Independent variables (at forum level in week *t-1*)

num_authors (log)	Total number of active authors in the forum
total_kp_leading_posts (log)	Total number of knowledge provision leading posts in the forum
total_kp_replies (log)	Total number of knowledge provision replies in the forum
total_leading_posts (log)	Total number of leading posts in the forum
total_mon_leading_posts (log)	Total number of money-involved leading posts in the forum
total_mon_replies (log)	Total number of money-involved replies in the forum
total_replies (log)	Total number of replies in the forum
total_replies_to_sec (log)	Total number of replies to cybersecurity-relevant leading posts in the forum
total_ka_leading_posts (log)	Total number of knowledge acquisition leading posts in the forum
total_ka_replies (log)	Total number of knowledge acquisition replies in the forum
total_sec_leading_posts (log)	Total number of cybersecurity-relevant leading posts in the forum
total_sec_replies (log)	Total number of cybersecurity-relevant replies in the forum
total_kp_rpl_to_sec (log)	Total number of knowledge provision replies to cybersecurity-relevant leading posts in the forum
total_mon_rpl_to_sec (log)	Total number of money-involved replies to cybersecurity-relevant leading posts in the forum
total_ka_rpl_to_sec (log)	Total number of knowledge acquisition replies to cybersecurity-relevant leading posts in the forum
total_sec_rpl_to_sec (log)	Total number of cybersecurity-relevant replies to cybersecurity-relevant leading posts in the forum
total_ISCsimi_rpl_to_sec (log)	Total average cosine similarity scores of replies to cybersecurity-relevant leading posts in the forum with ISC diaries
total_Synsimi_rpl_to_sec (log)	Total average cosine similarity scores of replies to cybersecurity-relevant leading posts in the forum with Symantec threat reports
total_src1_rpl_to_sec (log)	Total average number of source 1 cybersecurity keywords in replies to cybersecurity-relevant leading posts in the forum
total_src2_rpl_to_sec (log)	Total average number of source 2 cybersecurity keywords in replies to cybersecurity-relevant leading posts in the forum
total_src3_rpl_to_sec (log)	Total average number of source 3 cybersecurity keywords in replies to cybersecurity-relevant leading posts in the forum
total_lead_ISCsimilarity (log)	Total average cosine similarity scores of leading posts in the forum with ISC diaries
total_lead_Synsimilarity (log)	Total average cosine similarity scores of leading posts in the forum with Symantec threat reports
total_lead_src1_keyword (log)	Total average number of source 1 cybersecurity keywords in leading posts in the forum
total_lead_src2_keyword (log)	Total average number of source 2 cybersecurity keywords in leading posts in the forum
total_lead_src3_keyword (log)	Total average number of source 3 cybersecurity keywords in leading posts in the forum
total_rpl_ISCsimilarity (log)	Total average cosine similarity scores of replies in the forum with ISC diaries

total_rpl_Synsimilarity (log)	Total average cosine similarity scores of replies in the forum with Symantec threat reports
total_rpl_src1_keyword (log)	Total average number of source 1 cybersecurity keywords in replies in the forum
total_rpl_src2_keyword (log)	Total average number of source 2 cybersecurity keywords in replies in the forum
total_rpl_src3_keyword (log)	Total average number of source 3 cybersecurity keywords in replies in the forum
total_nonsec_leading_posts (log)	Total number of cybersecurity-irrelevant leading posts in the forum

Table B5b. Summary Statistics of Variables in Weekly Panel Data (Including 50,050 users and 205 weeks)

Variable	Obs.	Mean	Std. Dev.	Min	Max
Dependent variables (per user in week *t*)					
num_security_posts (log)	8607351	0.0018	0.0437	0	4.8903
perc secpost	8607351	0.0017	0.0385	0	1
src1_keyword_counts (log)	8607351	0.0036	0.0716	0	4.1431
src2_keyword_counts (log)	8607351	0.0056	0.1015	0	4.7536
src3_keyword_counts (log)	8607351	0.0039	0.0769	0	3.9703
ISC_similarity	8607351	0.0002	0.0052	0	.4018
Symantec_similarity	8607351	0.0001	0.0026	0	.3322
num_nonsecurity_posts (log)	8607351	0.0028	0.0542	0	4.5326
num_money involved_posts (log)	8607351	0.0002	0.0114	0	2.0794
Independent variables (per user in week *t-1*)					
auth_life (log)	8607351	4.8476	0.7865	0	5.9296
cum_kp_rpl_to_sec (log)	8607351	0.4068	0.7978	0	7.8272
cum_kp_posts (log)	8607351	0.2515	0.5233	0	6.4329
cum_leading_posts (log)	8607351	1.0661	0.6174	0	6.5511
cum_mon_rpl_to_sec (log)	8607351	0.3172	0.6801	0	7.1785
cum_mon_posts (log)	8607351	0.0132	0.1092	0	3.9318
cum_replies (log)	8607351	1.4851	1.3808	0	9.0157
cum_replies_to_sec (log)	8607351	0.6936	1.0905	0	8.3898
cum_ka_rpl_to_sec (log)	8607351	0.5790	0.9700	0	7.9666
cum_ka_posts (log)	8607351	0.8818	0.6202	0	5.4848
cum_sec_rpl_to_sec (log)	8607351	0.4986	0.8975	0	8.3598
cum_security_posts (log)	8607351	0.4393	0.5439	0	5.7869
cum_ISCsimilarity (log)	8607351	0.0204	0.1148	0	3.8127
cum_ISCsimi_rpl_to_sec (log)	8607351	0.0282	0.1960	0	5.7553
cum_Symantecsimilarity (log)	8607351	0.0118	0.0746	0	3.2643
cum_Synsimi_rpl_to_sec (log)	8607351	0.0166	0.1282	0	4.7122
cum_src1_keyword_counts (log)	8607351	0.8574	0.9356	0	8.5553
cum_src1_rpl_to_sec (log)	8607351	0.3634	0.7867	0	8.6154
cum_src2_keyword_counts (log)	8607351	1.2639	1.0786	0	9.1321
cum_src2_rpl_to_sec (log)	8607351	0.4671	0.9457	0	9.3388
cum_src3_keyword_counts (log)	8607351	0.9316	1.0043	0	8.8410
cum_src3_rpl_to_sec (log)	8607351	0.3410	0.7641	0	8.3443

cum_nonsec_posts (log)	8607351	0.7769	0.6884	0	6.4754
Independent variables (at forum level in week *t-1*)					
num_authors (log)	8607351	0.1075	0.7667	0	8.3224
total_kp_leading_posts (log)	8607351	0.0988	0.7049	0	7.9797
total_kp_replies (log)	8607351	0.1469	1.0457	0	10.4798
total_leading_posts (log)	8607351	0.1187	0.8462	0	9.3392
total_mon_leading_posts (log)	8607351	0.0310	0.2899	0	7.2923
total_mon_replies (log)	8607351	0.1334	0.9529	0	9.4414
total_replies (log)	8607351	0.1656	1.1783	0	11.5100
total_replies_to_sec (log)	8607351	0.1355	0.9688	0	10.8218
total_ka_leading_posts (log)	8607351	0.1060	0.7637	0	8.9706
total_ka_replies (log)	8607351	0.1557	1.1085	0	10.9904
total_sec_leading_posts (log)	8607351	0.0923	0.6665	0	8.8094
total_sec_replies (log)	8607351	0.1472	1.0489	0	10.9972
total_kp_rpl_to_sec (log)	8607351	0.1170	0.8375	0	9.7510
total_mon_rpl_to_sec (log)	8607351	0.1032	0.7382	0	8.5393
total_ka_rpl_to_sec (log)	8607351	0.1262	0.9037	0	10.3348
total_sec_rpl_to_sec (log)	8607351	0.1210	0.8734	0	10.8218
total_ISCsimi_rpl_to_sec (log)	8607351	0.0257	0.3416	0	6.2491
total_Synsimi_rpl_to_sec (log)	8607351	0.0197	0.2843	0	5.2952
total_src1_rpl_to_sec (log)	8607351	0.1090	0.7955	0	8.9402
total_src2_rpl_to_sec (log)	8607351	0.1162	0.8532	0	9.7717
total_src3_rpl_to_sec (log)	8607351	0.1068	0.7811	0	8.9564
total_lead_ISCsimilarity (log)	8607351	0.0252	0.2948	0	7.2009
total_lead_Synsimilarity (log)	8607351	0.0185	0.2453	0	6.3012
total_lead_src1_keyword (log)	8607351	0.1281	0.9155	0	10.2135
total_lead_src2_keyword (log)	8607351	0.1406	1.0047	0	11.0605
total_lead_src3_keyword (log)	8607351	0.1322	0.9414	0	8.7486
total_rpl_ISCsimilarity (log)	8607351	0.0378	0.3938	0	6.804
total_rpl_Synsimilarity (log)	8607351	0.0286	0.3270	0	5.8394
total_rpl_src1_keyword (log)	8607351	0.1350	0.9682	0	9.5026
total_rpl_src2_keyword (log)	8607351	0.1450	1.0414	0	10.3387
total_rpl_src3_keyword (log)	8607351	0.1353	0.9695	0	9.5286
total_nonsec_leading_posts (log)	8607351	0.1111	0.7917	0	8.4508

Note: variables cum_* represent cumulative value at the individual level until the previous week (t-1); total_* represents total amount at forum level in week (t-1); “lead” is abbreviation for leading posts; “rpl” is abbreviation for replies; “sec” is abbreviation for cybersecurity-relevant posts; “kp” is abbreviation for knowledge provision posts; “ka” is abbreviation for knowledge acquisition posts; “mon” is abbreviation for money-involved posts; “rpl_to_sec” indicates replies received in (or as of) week (t-1) responding to any of the cybersecurity-relevant leading posts contributed by a user as of week (t-1).

Table B5c. Summary Statistics of Variables for Treatment Group in Weekly Panel Data (Including 43,322 users and 205 weeks)

Variable	Obs.	Mean	Std. Dev.	Min	Max
Dependent variables (per user in week *t*)					
num_security_posts (log)	8229412	0.0007	0.0244	0	4.8903
perc_secpost	8229412	0.0007	0.0246	0	1
src1_keyword_counts (log)	8229412	0.0018	0.0484	0	3.5553
src2_keyword_counts (log)	8229412	0.0027	0.0651	0	4.0775
src3_keyword_counts (log)	8229412	0.0021	0.0559	0	3.9512
ISC_similarity	8229412	6.40E-06	0.0003	0	0.1076
Symantec_similarity	8229412	2.65E-06	0.0002	0	0.0646
num_nonsecurity_posts (log)	8229412	0.0022	0.0471	0	4.5326
num_money involved_posts (log)	8229412	2.49E-05	0.0043	0	1.7918
Independent variables (per user in week *t-1*)					
auth_life (log)	8229412	4.9187	0.6859	0	5.9296
cum_kp_rpl_to_sec (log)	8229412	0.4026	0.7774	0	7.5637
cum_kp_posts (log)	8229412	0.2481	0.5166	0	6.4329
cum_leading_posts (log)	8229412	1.0586	0.6060	0	6.5511
cum_mon_rpl_to_sec (log)	8229412	0.3240	0.6812	0	7.1785
cum_mon_posts (log)	8229412	0.0081	0.0789	0	1.9459
cum_replies (log)	8229412	1.4690	1.3658	0	8.5069
cum_replies_to_sec (log)	8229412	0.6444	1.0408	0	8.3898
cum_ka_rpl_to_sec (log)	8229412	0.5342	0.9169	0	7.9666
cum_ka_posts (log)	8229412	0.8729	0.6107	0	5.1299
cum_sec_rpl_to_sec (log)	8229412	0.4403	0.8123	0	7.5000
cum_security_posts (log)	8229412	0.4157	0.5175	0	5.7869
cum_ISCsimilarity (log)	8229412	0.0055	0.0166	0	1.0377
cum_ISCsimi_rpl_to_sec (log)	8229412	0.0029	0.0159	0	0.9590
cum_ Symantecsimilarity (log)	8229412	0.0036	0.0131	0	0.7260
cum_Synsimi_rpl_to_sec (log)	8229412	0.0019	0.0125	0	0.8156
cum_src1_keyword_counts (log)	8229412	0.8236	0.8997	0	8.5553
cum_src1_rpl_to_sec (log)	8229412	0.3036	0.6685	0	6.9735
cum_src2_keyword_counts (log)	8229412	1.2077	1.0340	0	9.1321
cum_src2_rpl_to_sec (log)	8229412	0.3817	0.7759	0	7.1476
cum_src3_keyword_counts (log)	8229412	0.8941	0.9748	0	8.8410
cum_src3_rpl_to_sec (log)	8229412	0.2749	0.6260	0	6.8926
cum_nonsec_posts (log)	8229412	0.7866	0.6839	0	6.4754
Independent variables (at forum level in week *t-1*)					
num_authors (log)	8229412	0.0892	0.6917	0	5.9814
total_kp_leading_posts (log)	8229412	0.0830	0.6448	0	6.2344
total_kp_replies (log)	8229412	0.1239	0.9602	0	8.8448
total_leading_posts (log)	8229412	0.0978	0.7583	0	6.6708
total_mon_leading_posts (log)	8229412	0.0175	0.1662	0	3.3673
total_mon_replies (log)	8229412	0.1162	0.9009	0	8.3390
total_replies (log)	8229412	0.1385	1.0731	0	9.6172
total_replies_to_sec (log)	8229412	0.1094	0.8488	0	8.0637

total_ka_leading_posts (log)	8229412	0.0858	0.6719	0	6.0521
total_ka_replies (log)	8229412	0.1298	1.0059	0	9.0926
total_sec_leading_posts (log)	8229412	0.0719	0.5582	0	5.3033
total_sec_replies (log)	8229412	0.1210	0.9382	0	8.5979
total_kp_rpl_to_sec (log)	8229412	0.0952	0.7394	0	7.2449
total_mon_rpl_to_sec (log)	8229412	0.0877	0.6814	0	6.7488
total_ka_rpl_to_sec (log)	8229412	0.1011	0.7851	0	7.5781
total_sec_rpl_to_sec (log)	8229412	0.0943	0.7321	0	7.1058
total_ISCsimi_rpl_to_sec (log)	8229412	0.0052	0.0447	0	0.8825
total_Synsimi_rpl_to_sec (log)	8229412	0.0026	0.0231	0	0.4174
total_src1_rpl_to_sec (log)	8229412	0.0821	0.6384	0	6.3767
total_src2_rpl_to_sec (log)	8229412	0.0862	0.6695	0	6.5525
total_src3_rpl_to_sec (log)	8229412	0.0800	0.6225	0	6.0521
total_lead_ISCsimilarity (log)	8229412	0.0081	0.0653	0	1.1337
total_lead_Synsimilarity (log)	8229412	0.0040	0.0326	0	0.6108
total_lead_src1_keyword (log)	8229412	0.1030	0.7981	0	8.3615
total_lead_src2_keyword (log)	8229412	0.1128	0.8741	0	8.9263
total_lead_src3_keyword (log)	8229412	0.1086	0.8416	0	8.7486
total_rpl_ISCsimilarity (log)	8229412	0.0156	0.1262	0	1.8034
total_rpl_Synsimilarity (log)	8229412	0.0097	0.0804	0	1.5294
total_rpl_src1_keyword (log)	8229412	0.1069	0.8294	0	8.4233
total_rpl_src2_keyword (log)	8229412	0.1139	0.8827	0	8.8446
total_rpl_src3_keyword (log)	8229412	0.1074	0.8326	0	8.6344
total_nonsec_leading_posts (log)	8229412	0.0939	0.7288	0	6.4394

Table B5d. Summary Statistics of Variables for Control Group in Weekly Panel Data (Including 6,728 users and 205 weeks)

Variable	Obs.	Mean	Std. Dev.	Min	Max
Dependent variables (per user in week *t*)					
num_security_posts (log)	377939	0.0270	0.1731	0	3.8067
perc secpost	377939	0.0230	0.1417	0	1
src1_keyword_counts (log)	377939	0.0408	0.2532	0	4.1431
src2_keyword_counts (log)	377939	0.0678	0.3721	0	4.7536
src3_keyword_counts (log)	377939	0.0419	0.2553	0	3.9703
ISC_similarity	377939	0.0043	0.0242	0	0.4018
Symantec_similarity	377939	0.0020	0.0124	0	0.3322
num_nonsecurity_posts (log)	377939	0.0178	0.1353	0	3.2958
num_money involved_posts (log)	377939	0.0032	0.0506	0	2.0794
Independent variables (per user in week *t-1*)					
auth_life (log)	377939	3.3001	1.1580	0	5.3845
cum_kp_rpl_to_sec (log)	377939	0.4974	1.1524	0	7.8272

cum_kp_posts (log)	377939	0.3252	0.6493	0	5.0752
cum_leading_posts (log)	377939	1.2297	0.8106	0	5.7807
cum_mon_rpl_to_sec (log)	377939	0.1683	0.6369	0	5.7652
cum_mon_posts (log)	377939	0.1255	0.3503	0	3.9318
cum_replies (log)	377939	1.8365	1.6340	0	9.0157
cum_replies_to_sec (log)	377939	1.7660	1.5146	0	8.3598
cum_ka_rpl_to_sec (log)	377939	1.5556	1.4578	0	7.6069
cum_ka_posts (log)	377939	1.0742	0.7747	0	5.4848
cum_sec_rpl_to_sec (log)	377939	1.7660	1.5146	0	8.3598
cum_security_posts (log)	377939	0.9545	0.7930	0	5.4161
cum_ISCsimilarity (log)	377939	0.3462	0.4282	0	3.8127
cum_ISCsimi_rpl_to_sec (log)	377939	0.5779	0.7438	0	5.7553
cum_ Symantecsimilarity (log)	377939	0.1914	0.2988	0	3.2643
cum_Synsimi_rpl_to_sec (log)	377939	0.3367	0.5134	0	4.7122
cum_src1_keyword_counts (log)	377939	1.5930	1.3198	0	8.0288
cum_src1_rpl_to_sec (log)	377939	1.6645	1.6105	0	8.6154
cum_src2_keyword_counts (log)	377939	2.4880	1.2823	0	8.5960
cum_src2_rpl_to_sec (log)	377939	2.3274	1.9086	0	9.3388
cum_src3_keyword_counts (log)	377939	1.7475	1.2593	0	7.8721
cum_src3_rpl_to_sec (log)	377939	1.7819	1.6103	0	8.3443
cum_nonsec_posts (log)	377939	0.5665	0.7495	0	4.7958
Independent variables (at forum level in week *t-1*)					
num_authors (log)	377939	0.5076	1.6738	0	8.3224
total_ kp_leading_posts (log)	377939	0.4412	1.4631	0	7.9797
total_kp_replies (log)	377939	0.6477	2.1361	0	10.4798
total_leading_posts (log)	377939	0.5739	1.8891	0	9.3392
total_mon_leading_posts (log)	377939	0.3253	1.1054	0	7.2923
total_mon_replies (log)	377939	0.5079	1.6916	0	9.4414
total_replies (log)	377939	0.7568	2.4863	0	11.5100
total_replies_to_sec (log)	377939	0.7034	2.3126	0	10.8218
total_ka_leading_posts (log)	377939	0.5470	1.8029	0	8.9706
total_ka_replies (log)	377939	0.7213	2.3698	0	10.9904
total_sec_leading_posts (log)	377939	0.5367	1.7682	0	8.8094
total_sec_replies (log)	377939	0.7168	2.3562	0	10.9972
total_kp_rpl_to_sec (log)	377939	0.5925	1.9582	0	9.7510
total_mon_rpl_to_sec (log)	377939	0.4408	1.4765	0	8.5393
total_ka_rpl_to_sec (log)	377939	0.6709	2.2061	0	10.3348
total_sec_rpl_to_sec (log)	377939	0.7034	2.3126	0	10.8218
total_ISCsimi_rpl_to_sec (log)	377939	0.4710	1.5512	0	6.2491
total_Synsimi_rpl_to_sec (log)	377939	0.3928	1.2973	0	5.2952
total_src1_rpl_to_sec (log)	377939	0.6934	2.2764	0	8.9402
total_src2_rpl_to_sec (log)	377939	0.7700	2.5242	0	9.7717
total_src3_rpl_to_sec (log)	377939	0.6886	2.2587	0	8.9564
total_lead_ISCsimilarity (log)	377939	0.3982	1.3194	0	7.2009

total_lead_Synsimilarity (log)	377939	0.3345	1.1146	0	6.3012
total_lead_src1_keyword (log)	377939	0.6746	2.2149	0	10.2135
total_lead_src2_keyword (log)	377939	0.7449	2.4433	0	11.0605
total_lead_src3_keyword (log)	377939	0.6466	2.1175	0	8.3067
total_rpl_ISCsimilarity (log)	377939	0.5215	1.7151	0	6.8040
total_rpl_Synsimilarity (log)	377939	0.4413	1.4545	0	5.8394
total_rpl_src1_keyword (log)	377939	0.7454	2.4455	0	9.5026
total_rpl_src2_keyword (log)	377939	0.8221	2.6934	0	10.3387
total_rpl_src3_keyword (log)	377939	0.7424	2.4341	0	9.5286
total_nonsec_leading_posts (log)	377939	0.4844	1.6009	0	8.4508

Table B6. Summary Statistics of Google Trends Index of Cybersecurity Keywords

ID	English	Chinese	Obs.	CN_index		WD_index	
				Mean	Std. Dev.	Mean	Std. Dev.
1	antivirus	杀毒	133	3.1091	1.0102	3.8263	0.3782
2	attack	攻击	133	3.3627	0.5105	3.716	0.2986
3	back door	后门	133	2.1334	1.0721	2.2973	1.0315
4	botnet	僵尸网络	133	2.7743	1.1601	3.621	0.3294
5	buffer overflow	缓存区溢出	133	2.8914	0.973	3.0939	0.7087
6	cracking	破解	133	3.2729	0.7237	3.1259	0.7636
7	DDOS	DDOS	133	3.216	0.6296	3.7528	0.285
8	decryption	解密	133	3.358	0.5046	3.6241	0.3884
9	DOS	DOS	133	2.9313	1.1335	3.1797	0.6308
10	encryption	加密	133	3.7254	0.4562	3.5856	0.4352
11	firewall	防火墙	133	2.8248	1.0486	2.9433	0.8722
12	hacker	黑客	133	2.9448	0.7957	3.8785	0.3247
13	infection	感染	133	2.4572	1.0093	1.6099	0.7126
14	injection	注入	133	2.9948	0.5269	4.0581	0.2752
15	keylogger	键盘记录器	133	2.1869	1.6576	3.817	0.4821
16	malware	恶意软件	133	2.4707	0.9257	3.704	0.442
17	patch	补丁	133	2.6697	1.0564	1.7095	0.7551
18	penetration	渗透	133	1.9219	0.8319	3.9772	0.2279
19	phishing	钓鱼	133	2.4144	0.9984	3.9187	0.2721

20	port	端口	133	3.102	0.7702	3.3334	0.6185
21	proxy server	代理	133	3.1565	0.8277	3.6502	0.3958
22	rootkit	根程序	133	2.295	0.9964	2.5667	0.5634
23	security	安全	133	3.4328	0.5397	4.1799	0.1712
24	spam	垃圾邮件	133	1.8388	1.1566	3.2749	0.6112
25	spoofing	欺骗	133	2.7836	0.9274	3.1888	0.597
26	spyware	间谍软件	133	1.8167	1.3106	2.6914	1.2051
27	trojan	木马	133	2.3788	1.3321	2.951	1.0111
28	virus	病毒	133	2.5976	1.2061	2.8176	0.6829
29	vulnerability	漏洞	133	3.3509	0.6802	3.3799	0.427
30	worm	蠕虫	133	2.2883	1.1564	1.5246	0.8674

Table B7. Robustness Check Using Google Trends as Controls

VARIABLES	Quantity		Relevance		
	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0023***	-0.0012***	-0.0043***	-0.0054***	-0.0057***
	(0.0008)	(0.0003)	(0.0009)	(0.0010)	(0.0009)
Observations	8,543,852	8,543,852	8,543,852	8,543,852	8,543,852
Adjusted R^2	0.2513	0.1651	0.2018	0.2310	0.1881
Google trends and user fixed effects are included					

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B8. PSM Performance for Matching Covariates

Variable	Unmatched / Matched	Mean: Treated	Mean: Control	%bias	%reduct |bias|	t-test: t	t-test: p>|t|	V(T)/V(C)
cum_leading_posts	U	3.0923	4.9501	-8.9		-10.80	0.000	0.09*
	M	1.1776	1.0870	0.4	95.1	3.92	0.000	1.01
cum_replies	U	15.512	37.3500	-17.3		-19.87	0.000	0.13*
	M	6.2666	5.1200	0.9	94.7	1.33	0.183	0.74*
cum_knowledge_	U	2.0953	3.4589	-22		-22.55	0.000	0.24*
acquisition	M	0.9036	0.8469	0.9	95.8	1.95	0.051	1.43*
cum_knowledge_	U	0.8327	1.2545	-2.3		-3.02	0.003	0.06*
provision	M	0.1769	0.1366	0.2	90.4	1.93	0.054	1.37*

Table B9. PSM Performance for Outcome Variables

Variable	Sample	English forum	Chinese forums	Difference	S.E.	T-stat
cum_security_posts	Unmatched	3.2552	0.8456	2.4096	0.1044	23.0689
	ATT	0.7192	0.4185	0.3007	0.0386	7.5429
cum_src1_keyword_counts	Unmatched	18.2527	4.3622	13.8905	1.2223	11.3618
	ATT	1.6894	1.2143	0.4751	0.1484	3.2047
cum_src2_keyword_counts	Unmatched	40.7357	8.2100	32.5256	2.5886	12.5713
	ATT	2.3447	3.8502	1.5055	0.2619	5.7491
cum_src3_keyword_counts	Unmatched	18.3496	5.9781	12.3715	1.1633	10.6402
	ATT	1.5903	1.3042	0.2861	0.1810	1.5810

Table B10. Robustness Check Using PSM Matched Users

VARIABLES	Quantity		Relevance		
	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0044***	-0.0037***	-0.0061***	-0.0107***	-0.0083***
	(0.0012)	(0.0006)	(0.0010)	(0.0016)	(0.0011)
Observations	4,447,337	4,447,337	4,447,337	4,447,337	4,447,337
Adjusted R^2	0.1714	0.0742	0.0869	0.1041	0.0773
Weekly fixed effects and user fixed effects are included					

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B11. DID Analysis, Including an Arrest Event as an Additional Treatment Variable

VARIABLES	Quantity		Relevance		
	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0023***	-0.0012***	-0.0043***	-0.0054***	-0.0057***
	(0.0007)	(0.0003)	(0.0009)	(0.0009)	(0.0010)
Treat×Arrest	-0.0016	-0.0016**	-0.0011	-0.0020	-0.0009
	(0.0011)	(0.0008)	(0.0009)	(0.0010)	(0.0011)
Observations	8,543,852	8,543,852	8,543,852	8,543,852	8,543,852
Adjusted R^2	0.2513	0.1650	0.2017	0.2314	0.1882
Weekly fixed effects and user fixed effects are included					

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B12. DID Analysis Cutting the Timeline after Oct 2010

VARIABLES	Quantity		Relevance		
	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0023***	-0.0012***	-0.0045***	-0.0056***	-0.0059***
	(0.0007)	(0.0004)	(0.0008)	(0.009)	(0.0010)
Observations	7,253,434	7,253,434	7,253,434	7,253,434	7,253,434
Adjusted R^2	0.2523	0.1657	0.2031	0.2330	0.1889
Weekly fixed effects and user fixed effects are included					

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B13. Internet Access Traffic by Country (Adapted from Yue et al. (2019))

	Hackforums.net		Hackbase		2cto	
	5/2015	1/2017	5/2015	1/2017	5/2015	1/2017
Algeria		0.6%				
Australia	2.7%	1.5%				
Azerbaijan						
Bangladesh	0.7%					
Belarus						
Belgium		0.7%				
Brazil	0.8%					
Canada	3.3%	4.8%				
China		1.1%	92.6%	69.2%	88.8%	96.8%
Croatia	0.7%	0.9%				
Czech Republic						
Denmark	1.6%	1.8%				
Egypt	2.2%	1.1%				
Finland						
France	2.2%	1.4%				
Germany	1.4%	5.5%				
Greece	1.2%	1.2%				
Hong Kong		0.6%		1.2%	0.6%	
Korea			4.5%	8.2%		
India	22.6%	5.1%				
Indonesia	1.2%					
Iran		1.0%				
Israel						
Italy	1.3%	1.4%				
Japan			0.9%	19.9%		1.3%
Kazakhstan						
Kuwait		1.5%				
Latvia						
Mexico		0.8%				
Morocco		0.5%				
Netherlands	5.0%	3.8%				
Nigeria		0.8%				
Norway	3.3%	3.7%				
Pakistan	0.9%					
Philippines	1.1%					
Poland	1.0%	0.5%				
Portugal	1.7%					

Romania	0.6%	1.4%				
Russia	0.9%					
Saudi Arabia	2.1%	0.6%				
Singapore		0.6%				
Slovenia	0.9%					
Spain	0.6%	2.0%				
Sweden	5.5%	1.1%				
Taiwan			1.1%	0.9%		
Turkey	1.8%	0.6%				
Ukraine						
United Kingdom	7.8%	10.6%				
United States	14.7%	28.9%	0.5%		8.8%	1.0%
Uzbekistan						

Table B14. Robustness Check by Adding Controls about Enforcement in Other Countries

VARIABLES	**Quantity**		**Relevance**		
	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0053***	-0.0042***	-0.0063***	-0.0101***	-0.0074***
	(0.0010)	(0.0008)	(0.0018)	(0.0026)	(0.0023)
Eng×Effort	-0.0614	-0.0259	-0.0313	-0.0671	0.0069
	(0.0413)	(0.0321)	(0.0482)	(0.0788)	(0.0524)
Observations	7,599,800	7,599,800	7,599,800	7,599,800	7,599,800
Adjusted R^2	0.2520	0.1649	0.2028	0.2322	0.1891
Weekly fixed effects and user fixed effects are included					

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B15. Falsification Test Using Cybersecurity-irrelevant Leading Posts as Alternative Dependent Variables

VARIABLES	Full sample	PSM sample
	(1)	(2)
	num_nonsecurity_posts	num_nonsecurity_posts
Treat×CMA	-0.0009	-0.0010
	(0.0009)	(0.0008)
Week FE	Yes	Yes
User FE	Yes	Yes
Observations	8,557,301	4,447,337
Adjusted R^2	0.1692	0.0919

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B16. Falsification Test Using Money-involved Leading Posts as Alternative Dependent Variables

VARIABLES	(1)	(2)
	num_moneyinvolved_posts	perc_monposts
Treat×CMA	0.0003	0.0001
	(0.0002)	(0.0001)
Week FE	Yes	Yes
User FE	Yes	Yes
Observations	8,543,852	8,543,852
Adjusted R^2	0.0993	0.0604

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B17. Falsification Test with Random Implementation of Pseudo Treatment

	Quantity		Relevance		
	(1)	(2)	(3)	(4)	(5)
VARIABLES	num_security_posts	perc_secposts	src1_keyword_counts	src2_keyword_counts	src3_keyword_counts
Mean of random β	-2.26E-05	2.83E-06	-8.50E-06	-2.91E-06	-6.80E-06
S.D. of random β	0.0001	0.0001	0.0002	0.0003	0.0002
z-score	-0.1592	0.0275	-0.0392	-0.0094	-0.0289
p-value	0.8735	0.9781	0.9687	0.9925	0.9769
replications	1000	1000	1000	1000	1000

Table B18. Descriptive Statistics of Moderating Variables and Alternative Dependent Variables

Variable	Description	Mean	Std. Dev.	Min	Max
Moderating variables of prior experience (observations 8,543,852) measured at individual level, as of week *t-1* and in logarithmic form					
cum_sec_posts (log)	The cumulative number of cybersecurity-relevant leading posts	0.4393	0.5439	0	5.7869
cum_src1_keyword_counts (log)	The cumulative number of source 1 cybersecurity keywords in leading posts	0.8574	0.9356	0	8.5553
cum_src2_keyword_counts (log)	The cumulative number of source 2 cybersecurity keywords in leading posts	1.2639	1.0786	0	9.1321
cum_src3_keyword_counts (log)	The cumulative number of source 3 cybersecurity keywords in leading posts	0.9316	1.0043	0	8.8410
Moderating variables for peer feedback (observations 8,543,852) including replies that responded to any of a user's cybersecurity-relevant leading posts as of week *t-1*, in logarithmic form, and proportional to the cumulative number of cybersecurity-relevant leading posts as of week *t-1*					
avg_sec_rpl_to_sec (log)	The average number of cybersecurity-relevant replies	0.0031	0.0632	0	5.3613
avg_src1_rpl_to_sec (log)	The average number of source 1 cybersecurity keywords in replies	0.0024	0.0602	0	5.2933
avg_src2_rpl_to_sec (log)	The average number of source 2 cybersecurity keywords in replies	0.0039	0.0868	0	5.8693
avg_src3_rpl_to_sec (log)	The average number of source 3 cybersecurity keywords in replies	0.0026	0.0620	0	5.0938
Moderating variables for peer participation (observations 8,543,852) at forum level in week *t-1* and in logarithmic form					
total_sec_leading_posts (log)	Total number of cybersecurity-relevant leading posts	0.0923	0.6663	0	8.8094
total_lead_src1_keyword (log)	Total number of source 1 cybersecurity keywords in leading posts	0.1280	0.9153	0	10.2135
total_lead_src2_keyword (log)	Total number of source 2 cybersecurity keywords in leading posts	0.1405	1.0045	0	11.0605
total_lead_src3_keyword (log)	Total number of source 3 cybersecurity keywords in leading posts	0.1322	0.9411	0	8.7486
Moderating variables for peer feedback (observations 8,543,852) simultaneously considering the replies to cybersecurity-relevant leading posts by a focal user (original measure in Table 3) as well as other leading posts responded to by a focal user					
avg_sec_ rpl_and_leads (log)	The average number of cybersecurity-relevant replies and leading posts	0.0113	0.1281	0	5.3613
avg_src1_ rpl_and_lead (log)	The average number of source 1 cybersecurity keywords in replies and leading posts	0.0175	0.1532	0	5.2983
avg_src2_ rpl_and_lead (log)	The average number of source 2 cybersecurity keywords in replies and leading posts	0.0240	0.2082	0	5.8861
avg_src3_ rpl_and_lead (log)	The average number of source 3 cybersecurity keywords in replies and leading posts	0.0153	0.1439	0	5.0940

Alternative dependent variables for knowledge acquisition or knowledge provision (observations 8,523,616) at individual level in week *t* and in logarithmic form					
num_sec_ka_leading_posts (log)	The number of cybersecurity-relevant and knowledge acquisition leading posts	0.0013	0.0355	0	3.0910
perc_sec_ka_leading_posts	Percentage of knowledge acquisition leading posts in cybersecurity-relevant leading posts	0.0005	0.0334	0	1
num_sec_kp_leading_posts (log)	The number of cybersecurity-relevant and knowledge provision leading posts	0.0006	0.0241	0	4.8752
perc_sec_kp_leading_posts	Percentage of knowledge provision leading posts in cybersecurity-relevant leading posts	0.0002	0.0263	0	1
num_sec_ka_replies (log)	The number of cybersecurity-relevant and knowledge acquisition replies	0.0102	0.1247	0	6.3026
num_sec_kp_replies (log)	The number of cybersecurity-relevant and knowledge provision replies	0.0082	0.1034	0	5.0499
num_replies_to_sec (log)	The number of replies to cybersecurity-relevant leading posts	0.0064	0.1200	0	6.9147
Moderating variables for prior experience of knowledge acquisition and knowledge provision (observations 8,523,616) at individual level in week *t* and in logarithmic form					
cum_ka_leading_posts (log)	The cumulative number of knowledge acquisition leading posts	0.0845	0.3072	0	5.1240
cum_kp_leading_posts (log)	The cumulative number of knowledge provision leading posts	0.0314	0.1974	0	5.7526

Table B19. Moderating Effect of Alternative Measures of Social Learning on the Chilling Effect

VARIABLES	Quantity		Relevance		
	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
(a) Test the moderating effect of sum of average number of cybersecurity-relevant replies to cybersecurity-relevant leading posts generated by focal user and number of cybersecurity-relevant leading posts replied by focal user until *t-1*					
Treat×CMA	-0.0022**	-0.0025***	-0.0072***	-0.0096***	-0.0084***
	(0.0008)	(0.0009)	(0.0013)	(0.0014)	(0.0012)
Treat×CMA×avg_sec_	0.0019**	0.0024***	0.0053***	0.0081***	0.0048***
rpl_and_leads	(0.0010)	(0.0010)	(0.0011)	(0.0020)	(0.0011)
(b) Test the moderating effect of sum of average number of Source 1 cybersecurity keywords in replies to cybersecurity-relevant leading posts generated by focal user and average number of Source 1 cybersecurity keywords in leading posts replied by focal user until *t-1*					
Treat×CMA	-0.0063***	-0.0045***	-0.0101***	-0.0147***	-0.0111***
	(0.0001)	(0.0003)	(0.0011)	(0.0012)	(0.0010)
Treat×CMA×avg_src1_	0.0020**	0.0030***	0.0054***	0.0089***	0.0049***
rpl_and_lead	(0.0010)	(0.0004)	(0.0008)	(0.0014)	(0.0010)
(c) Test the moderating effect of sum of average number of Source 2 cybersecurity keywords in replies to cybersecurity-relevant leading posts generated by focal user and average number of Source 2 cybersecurity keywords in leading posts replied by focal user until *t-1*					
Treat×CMA	-0.0074***	-0.0059***	-0.0123***	-0.0187***	-0.0134***
	(0.0001)	(0.0004)	(0.0010)	(0.0009)	(0.0011)
Treat×CMA×avg_src2_	0.0016**	0.0023***	0.0038***	0.0064***	0.0035***
rpl_and_lead	(0.0008)	(0.0003)	(0.0011)	(0.0012)	(0.0010)
(d) Test the moderating effect of sum of average number of Source 3 cybersecurity keywords in replies to cybersecurity-relevant leading posts generated by focal user and average number of Source 3 cybersecurity keywords in leading posts replied by focal user until *t-1*					
Treat×CMA	-0.0060***	-0.0046***	-0.0105***	-0.0158***	-0.0121***
	(0.0003)	(0.0003)	(0.0013)	(0.0011)	(0.0009)
Treat×CMA×avg_src3_	0.0013**	0.0026***	0.0038***	0.0069***	0.0039***
rpl_and_lead	(0.0006)	(0.0007)	(0.0011)	(0.0010)	(0.0010)
Observations	8,543,852	8,543,852	8,543,852	8,543,852	8,543,852
Weekly fixed effects and user fixed effects are included					

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B20. Moderating Effect of Forum-level Peer Participation on the Chilling Effect

VARIABLES	Quantity		Relevance		
	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Test the moderating effect of forum-level total number of cybersecurity-relevant posts in week *t-1*					
treat×CMA	-0.0029***	-0.0016***	-0.0046***	-0.0058***	-0.0058***
	(0.0007)	(0.000)	(0.0010)	(0.0013)	(0.0012)
Treat×CMA×total_sec_p	0.0052**	0.0032**	-0.0004	0.001319	-0.0020
osts	(0.0025)	(0.0016)	(0.0031)	(0.0054)	(0.0033)
Test the moderating effect of forum-level total number of Source 1 cybersecurity keywords in week *t-1*					
Treat×CMA	-0.0029***	-0.0016***	-0.0045***	-0.0058***	-0.0057***
	(0.0010)	(0.0004)	(0.0009)	(0.0011)	(0.0010)
Treat×CMA×total_src1_	0.0042**	0.0026**	0.0001	0.0015	-0.0014
keyword_counts	(0.0021)	(0.0013)	(0.0030)	(0.0043)	(0.0032)
Test the moderating effect of forum-level total number of Source 2 cybersecurity keywords in week *t-1*					
Treat×CMA	-0.0029***	-0.0016***	-0.0045***	-0.0058***	-0.0057***
	(0.0009)	(0.0003)	(0.0011)	(0.0009)	(0.0010)
Treat×CMA×total_src2_	0.0039**	0.0024**	0.0002	0.0014	-0.0011
keyword_counts	(0.0020)	(0.0012)	(0.0021)	(0.0034)	(0.0023)
Test the moderating effect of forum-level total number of Source 3 cybersecurity keywords in week *t-1*					
Treat×CMA	-0.0029***	-0.0016***	-0.0046***	-0.0058***	-0.0058***
	(0.0008)	(0.0004)	(0.0010)	(0.0013)	(0.0011)
Treat×CMA×total_src3_	0.0051**	0.0030**	0.0009	0.0026	-0.0005
keyword_counts	(0.0026)	(0.0015)	(0.0032)	(0.0041)	(0.0031)
Observations	8,543,852	8,543,852	8,543,852	8,543,852	8,543,852
Weekly fixed effects and user fixed effects are included					

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Table B21. The Complete DID Estimation Results for Full Sample

VARIABLES	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0024***	-0.0012***	-0.0044***	-0.0054***	-0.0057***
	(0.0008)	(0.0004)	(0.0009)	(0.0012)	(0.0009)
ln_auth_life	-0.0031***	-0.0028***	-0.0067***	-0.0098***	-0.0075***
	(0.0002)	(0.0002)	(0.0004)	(0.0005)	(0.0004)
cum_kp_rpl_to_sec (log)	-0.0042	-0.0035	-0.0076	-0.0098	-0.0071
	(0.0040)	(0.0027)	(0.0050)	(0.0072)	(0.0050)
cum_kp_posts (log)	0.0037	0.0104***	0.0004	-0.0025	-0.0013
	(0.0024)	(0.0016)	(0.0038)	(0.0052)	(0.0043)
cum_leading_posts (log)	0.0049	0.0050**	0.0249***	0.0352***	0.0234***
	(0.0043)	(0.0021)	(0.0072)	(0.0093)	(0.0069)
cum_mon_rpl_to_sec (log)	-0.0018	0.0026	-0.0006	-0.0014	-0.0027
	(0.0058)	(0.0034)	(0.0071)	(0.0098)	(0.0071)
cum_mon_posts (log)	0.0306***	0.0124**	0.0283**	0.0325**	0.0254**
	(0.0109)	(0.0056)	(0.0115)	(0.0163)	(0.0118)
cum_replies (log)	0.0042***	0.0038***	0.0102***	0.0149***	0.0115***
	(0.0005)	(0.0004)	(0.0008)	(0.0011)	(0.0008)
cum_replies_to_sec (log)	0.0161***	0.0143***	0.0062	0.0118	0.0079
	(0.0042)	(0.0029)	(0.0059)	(0.0079)	(0.0063)
cum_ka_rpl_to_sec (log)	0.0027	0.0011	0.0107*	0.0118	0.0058
	(0.0048)	(0.0034)	(0.0065)	(0.0088)	(0.0064)
cum_ka_posts (log)	-0.0012	0.0012	-0.0119*	-0.0263***	-0.0141**
	(0.0046)	(0.0022)	(0.0070)	(0.0091)	(0.0077)
cum_sec_rpl_to_sec (log)	-0.0100**	-0.0094***	-0.0082	-0.0073	-0.0021
	(0.0047)	(0.0031)	(0.0062)	(0.0086)	(0.0066)
cum_security_posts (log)	-0.0673***	-0.0754***	-0.0079	-0.0126	-0.0060
	(0.0041)	(0.0032)	(0.0062)	(0.0078)	(0.0072)
cum_ISCsimilarity (log)	0.15567**	0.0739*	0.1854**	0.0297	0.0750
	(0.0711)	(0.0392)	(0.0731)	(0.1087)	(0.0741)
cum_ISCsimi_rpl_to_sec (log)	0.1455***	0.0494***	0.1339***	0.2082***	0.1359***
	(0.0325)	(0.0198)	(0.0415)	(0.0575)	(0.0432)
cum_ Symantecsimilarity (log)	-0.1842*	-0.0140	-0.2936***	-0.1426	-0.1965*
	(0.1024)	(0.0543)	(0.1040)	(0.1513)	(0.1059)
cum_Synsimi_rpl_to_sec (log)	-0.2787***	-0.1306***	-0.2126***	-0.3394***	-0.2091***
	(0.0447)	(0.0251)	(0.0560)	(0.0773)	(0.0591)
cum_src1_keyword_counts (log)	0.0014	0.0009	-0.0510***	-0.0053*	0.0003
	(0.0015)	(0.0012)	(0.0022)	(0.0030)	(0.0023)
cum_src1_rpl_to_sec (log)	-0.0016	-0.0007	-0.0062**	-0.0036	-0.0037
	(0.0021)	(0.0014)	(0.0030)	(0.0041)	(0.0033)
cum_src2_keyword_counts	-0.0016	-0.0020**	-0.0015	-0.0464***	0.0026

(log)	(0.0012)	(0.0010)	(0.0018)	(0.0021)	(0.0020)
cum_src2_rpl_to_sec (log)	-0.0034*	-0.0029*	-0.0020	-0.0097**	-0.0039
	(0.0020)	(0.0016)	(0.0031)	(0.0043)	(0.0034)
cum_src3_keyword_counts	0.0019*	0.0005	0.0019	-0.0029	-0.0509***
(log)	(0.0012)	(0.0010)	(0.0019)	(0.0024)	(0.0022)
cum_src3_rpl_to_sec (log)	0.0006	0.0010	-0.0022	-0.0028	-0.0048
	(0.0018)	(0.0013)	(0.0029)	(0.0039)	(0.0032)
num_authors (log)	-0.0529***	-0.0187**	-0.0576***	-0.0885***	-0.0547***
	(0.0115)	(0.0072)	(0.0186)	(0.0249)	(0.0214)
total_ kp_leading_posts	-0.0266***	-0.0161***	-0.0351***	-0.0549***	-0.0460***
(log)	(0.0060)	(0.0044)	(0.0098)	(0.0141)	(0.0108)
total_kp_replies (log)	-0.0069	-0.0075	-0.0153	-0.0384	-0.0261
	(0.0222)	(0.0150)	(0.0301)	(0.0421)	(0.0312)
total_leading_posts (log)	0.0588***	0.0227**	0.1094***	0.1530***	0.1189***
	(0.0159)	(0.0103)	(0.0268)	(0.0350)	(0.0295)
total_mon_leading_posts	0.0026***	0.0013*	0.0041***	0.0048**	0.0022
(log)	(0.0009)	(0.0007)	(0.0015)	(0.0020)	(0.0017)
total_mon_replies (log)	0.0010	-0.0030	0.0019	0.0073	-0.0026
	(0.0137)	(0.0098)	(0.0198)	(0.0272)	(0.0197)
total_replies (log)	0.0378*	0.0287*	0.0645*	0.0798*	0.0520
	(0.0214)	(0.0154)	(0.0352)	(0.0466)	(0.0371)
total_replies_to_sec (log)	-0.0041	-0.0051	0.0015	0.0240	0.0258
	(0.0145)	(0.0105)	(0.0245)	(0.0320)	(0.0271)
total_ka_leading_posts	-0.0037	-0.0030	-0.0164*	-0.0165	-0.0200**
(log)	(0.0053)	(0.0025)	(0.0088)	(0.0112)	(0.0097)
total_ka_replies (log)	-0.0448*	-0.0314	-0.0480	-0.0869	-0.0381
	(0.0272)	(0.0213)	(0.0427)	(0.0580)	(0.0451)
total_sec_leading_posts	-0.0007	-0.0005	-0.0157**	-0.0121	-0.0139*
(log)	(0.0036)	(0.0032)	(0.0065)	(0.0101)	(0.0084)
total_sec_replies (log)	0.0324	0.0234	-0.0003	0.0302	0.0026
	(0.0228)	(0.0172)	(0.0347)	(0.0484)	(0.0356)
total_kp_rpl_to_sec (log)	-0.0023	0.0005	-0.0059	0.0027	0.0013
	(0.0163)	(0.0124)	(0.0251)	(0.0339)	(0.0263)
total_mon_rpl_to_sec (log)	0.0051	0.0013	0.0091	0.0034	0.0099
	(0.0098)	(0.0077)	(0.0154)	(0.0212)	(0.0147)
total_ka_rpl_to_sec (log)	-0.0011	0.0013	-0.0041	-0.0356	-0.0275
	(0.0176)	(0.0143)	(0.0334)	(0.0430)	(0.0352)
total_sec_rpl_to_sec (log)	0.0011	0.0002	0.0013	0.0136	-0.0033
	(0.0167)	(0.0122)	(0.0256)	(0.0338)	(0.0278)
total_ISCsimi_rpl_to_sec	0.0147**	0.0024	0.0076	0.0034	0.0039
(log)	(0.0065)	(0.0054)	(0.0115)	(0.0154)	(0.0130)
total_Synsimi_rpl_to_sec	0.0023	0.0038	0.0277*	0.0432**	0.0189
(log)	(0.0093)	(0.0069)	(0.0146)	(0.0205)	(0.0168)
total_src1_rpl_to_sec (log)	0.0016	0.0056	-0.0023	0.0021	-0.0059

	(0.0054)	(0.0044)	(0.0109)	(0.0142)	(0.0125)
total_src2_rpl_to_sec (log)	-0.0053	-0.0088**	-0.0123	-0.0177	-0.0043
	(0.0054)	(0.0045)	(0.0102)	(0.0133)	(0.0118)
total_src3_rpl_to_sec (log)	-0.0019	-0.0001	0.0003	-0.0013	-0.0054
	(0.0045)	(0.0031)	(0.0077)	(0.0099)	(0.0087)
total_lead_ISCsimilarity	0.0152	0.0152*	0.0023	0.0040	0.0022
(log)	(0.0117)	(0.0076)	(0.0198)	(0.0256)	(0.0205)
total_lead_Synsimilarity	-0.0142	-0.0062	-0.0126	-0.0225	-0.0158
(log)	(0.0164)	(0.0098)	(0.0243)	(0.0328)	(0.0257)
total_lead_src3_keyword	-0.0173*	-0.0068	-0.0102	-0.0331*	-0.0176
(log)	(0.0096)	(0.0064)	(0.0137)	(0.0198)	(0.0151)
total_lead_src1_keyword	-0.0155	-0.0125*	0.0023	-0.0230	-0.0077
(log)	(0.0097)	(0.0065)	(0.0155)	(0.0212)	(0.0182)
total_lead_src2_keyword	0.0466***	0.0277***	0.0366*	0.0901***	0.0614***
(log)	(0.0126)	(0.0081)	(0.0197)	(0.0261)	(0.0205)
total_rpl_ISCsimilarity	-0.0001	0.0016	0.0168*	0.0205*	0.0228**
(log)	(0.0048)	(0.0039)	(0.0091)	(0.0118)	(0.0103)
total_rpl_Synsimilarity	0.0073	0.0039	0.0059	0.0061	0.0061
(log)	(0.0045)	(0.0038)	(0.0086)	(0.0112)	(0.0096)
total_rpl_src3_keyword	0.0072	0.0059	0.0138	0.0222	0.0171
(log)	(0.0091)	(0.0073)	(0.0152)	(0.0197)	(0.0170)
total_rpl_src1_keyword	-0.0318***	-0.0223***	-0.0627***	-0.0774***	-0.0496**
(log)	(0.0112)	(0.0078)	(0.0199)	(0.0256)	(0.0216)
total_rpl_src2_keyword	0.0037	0.0070	0.0188	0.0258	0.0072
(log)	(0.0131)	(0.0087)	(0.0192)	(0.0259)	(0.0213)
Constant	0.0303***	0.0301***	0.0526***	0.0893***	0.0598***
	(0.0014)	(0.0012)	(0.0022)	(0.0031)	(0.0024)
Observations	8,543,852	8,543,852	8,543,852	8,543,852	8,543,852
Adjusted R-squared	0.2513	0.1649	0.2020	0.2314	0.1879

*** p<0.01, ** p<0.05, * p<0.1 Robust standard errors in parentheses

Table B22. The Complete DID Estimation Results for PSM Sample

VARIABLES	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0044***	-0.0037***	-0.0061***	-0.0107***	-0.0083***
	(0.0012)	(0.0006)	(0.0010)	(0.0016)	(0.0011)
ln_auth_life	-0.0013***	-0.0017***	-0.0035***	-0.0054***	-0.0040***
	(0.0001)	(0.0001)	(0.0002)	(0.0003)	(0.0003)
cum_kp_rpl_to_sec (log)	-0.0025	0.0021	0.0030	0.0006	0.0019
	(0.0051)	(0.0039)	(0.0068)	(0.0094)	(0.0066)
cum_kp_posts (log)	0.0038*	0.0044*	0.0049	0.0088*	0.0068*
	(0.0021)	(0.0023)	(0.0035)	(0.0052)	(0.0041)
cum_leading_posts (log)	-0.0260***	-0.0429***	-0.0341***	-0.0229**	-0.0185**
	(0.0050)	(0.0052)	(0.0081)	(0.0110)	(0.0088)
cum_mon_rpl_to_sec	0.0098	0.0027	-0.0004	-0.0004	-0.0049
(log)	(0.0136)	(0.0051)	(0.0084)	(0.0122)	(0.0083)
cum_mon_posts (log)	0.0299**	0.0138*	0.0147	0.0285	0.0180
	(0.0120)	(0.0079)	(0.0133)	(0.0185)	(0.0133)
cum_replies (log)	0.0023***	0.0029***	0.0055***	0.0086***	0.0065***
	(0.0003)	(0.0003)	(0.0005)	(0.0007)	(0.0005)
cum_replies_to_sec (log)	0.0084*	0.0032	-0.0024	0.0014	0.0000
	(0.0047)	(0.0037)	(0.0053)	(0.0077)	(0.0059)
cum_ka_rpl_to_sec (log)	0.0039	0.0072*	0.0164***	0.0200**	0.0127**
	(0.0050)	(0.0040)	(0.0054)	(0.0078)	(0.0054)
cum_ka_posts (log)	0.0066***	0.0042*	0.0138***	0.0137**	0.0137***
	(0.0025)	(0.0024)	(0.0041)	(0.0057)	(0.0045)
cum_sec_rpl_to_sec (log)	-0.0141*	-0.0117***	-0.0139**	-0.0183**	-0.0074
	(0.0085)	(0.0042)	(0.0059)	(0.0084)	(0.0059)
cum_security_posts (log)	-0.0455***	-0.0445***	0.0158*	0.0053	0.0073
	(0.0051)	(0.0047)	(0.0087)	(0.0119)	(0.0097)
cum_ISCsimilarity (log)	0.2196**	0.2113***	0.1544*	0.3065**	0.1900**
	(0.1100)	(0.0630)	(0.0830)	(0.1306)	(0.0787)
cum_ISCsimi_rpl_to_sec	-0.0066	0.0385	0.1102	0.1320	0.0706
(log)	(0.0684)	(0.0416)	(0.0810)	(0.1200)	(0.0807)
cum_ Symantecsimilarity	-0.0901	-0.1606*	-0.1141	-0.2105	-0.1331
(log)	(0.1229)	(0.0865)	(0.1065)	(0.1847)	(0.1094)
cum_Synsimi_rpl_to_sec	-0.0029	-0.0662	-0.1120	-0.1831	-0.0991
(log)	(0.1190)	(0.0652)	(0.1243)	(0.1935)	(0.1250)
cum_src1_keyword_coun	-0.0029**	-0.0033***	-0.0516***	-0.0066**	-0.0009
ts (log)	(0.0013)	(0.0012)	(0.0022)	(0.0029)	(0.0022)
cum_src1_rpl_to_sec	-0.0058**	-0.0039	-0.0030	-0.0034	0.0001
(log)	(0.0028)	(0.0024)	(0.0033)	(0.0044)	(0.0033)
cum_src2_keyword_coun	-0.0007	-0.0007	-0.0029*	-0.0528***	-0.0027

ts (log)	(0.0010)	(0.0011)	(0.0016)	(0.0025)	(0.0018)
cum_src2_rpl_to_sec	0.0019	0.0019	-0.0002	-0.0006	-0.0011
(log)	(0.0020)	(0.0022)	(0.0031)	(0.0042)	(0.0034)
cum_src3_keyword_coun	0.0034***	0.0039***	0.0020	0.0009	-0.0498***
ts (log)	(0.0011)	(0.0012)	(0.0019)	(0.0028)	(0.0024)
cum_src3_rpl_to_sec	-0.0002	0.0005	-0.0022	-0.0031	-0.0048
(log)	(0.0024)	(0.0023)	(0.0037)	(0.0051)	(0.0042)
cum_nonsec_posts (log)	0.0181***	0.0366***	0.0238***	0.0106	0.0085
	(0.0049)	(0.0049)	(0.0076)	(0.0104)	(0.0084)
num_authors (log)	0.0168**	0.0175**	0.0373***	0.0555***	0.0489***
	(0.0069)	(0.0073)	(0.0144)	(0.0189)	(0.0156)
total_ kp_leading_posts	0.0115**	0.0099**	0.0258***	0.0372***	0.0305***
(log)	(0.0046)	(0.0045)	(0.0090)	(0.0117)	(0.0099)
total_kp_replies (log)	0.0412**	0.0515***	0.0672***	0.1041***	0.0719**
	(0.0163)	(0.0150)	(0.0255)	(0.0350)	(0.0280)
total_leading_posts (log)	0.0228	0.0294	0.0025	-0.0344	-0.0058
	(0.0317)	(0.0215)	(0.0440)	(0.0541)	(0.0449)
total_mon_leading_posts	-0.0026**	-0.0022**	-0.0025*	-0.0053**	-0.0042***
(log)	(0.0012)	(0.0010)	(0.0015)	(0.0022)	(0.0016)
total_mon_replies (log)	-0.0094	-0.0156	-0.0152	-0.0247	-0.0154
	(0.0269)	(0.0131)	(0.0230)	(0.0339)	(0.0237)
total_replies (log)	-0.0215	-0.0048	-0.0586*	-0.0619	-0.0676**
	(0.0149)	(0.0167)	(0.0313)	(0.0411)	(0.0340)
total_replies_to_sec (log)	0.0164	0.0090	0.0246	0.0232	0.0301
	(0.0123)	(0.0119)	(0.0235)	(0.0305)	(0.0265)
total_ka_leading_posts	-0.0042***	-0.0036**	-0.0096*	-0.0131*	-0.0133**
(log)	(0.0016)	(0.0017)	(0.0054)	(0.0067)	(0.0057)
total_ka_replies (log)	-0.0102	-0.0266	0.0255	-0.0174	0.0333
	(0.0237)	(0.0212)	(0.0394)	(0.0544)	(0.0420)
total_sec_leading_posts	-0.0089	-0.0103*	-0.0136	-0.0112	-0.0156
(log)	(0.0083)	(0.0059)	(0.0114)	(0.0148)	(0.0123)
total_sec_replies (log)	-0.0079	-0.0093	-0.0326	-0.0187	-0.0413
	(0.0151)	(0.0145)	(0.0280)	(0.0401)	(0.0311)
total_kp_rpl_to_sec (log)	-0.0168	-0.0132	-0.0439**	-0.0501*	-0.0404*
	(0.0115)	(0.0129)	(0.0212)	(0.0303)	(0.0241)
total_mon_rpl_to_sec	0.0261**	0.0191	0.0364**	0.0460	0.0349*
(log)	(0.0118)	(0.0124)	(0.0172)	(0.0284)	(0.0189)
total_ka_rpl_to_sec (log)	-0.0162	-0.0166	-0.0220	-0.0017	-0.0227
	(0.0150)	(0.0144)	(0.0273)	(0.0355)	(0.0304)
total_sec_rpl_to_sec	-0.0150	-0.0018	0.0015	-0.0211	-0.0009
(log)	(0.0131)	(0.0134)	(0.0227)	(0.0317)	(0.0245)
total_ISCsimi_rpl_to_sec	-0.0028	-0.0083	-0.0162	-0.0179	-0.0106
(log)	(0.0064)	(0.0066)	(0.0103)	(0.0150)	(0.0117)
total_Synsimi_rpl_to_sec	0.0256***	0.0190**	0.0382***	0.0633***	0.0337**

(log)	(0.0095)	(0.0081)	(0.0145)	(0.0209)	(0.0159)
total_src1_rpl_to_sec	0.0028	0.0028	-0.0013	0.0039	-0.0007
(log)	(0.0040)	(0.0051)	(0.0090)	(0.0121)	(0.0099)
total_src2_rpl_to_sec	-0.0021	-0.0042	-0.0209**	-0.0256*	-0.0141
(log)	(0.0042)	(0.0053)	(0.0104)	(0.0133)	(0.0114)
total_src3_rpl_to_sec	-0.0025	-0.0007	0.0107	0.0070	0.0001
(log)	(0.0036)	(0.0040)	(0.0074)	(0.0095)	(0.0084)
total_lead_ISCsimilarity	-0.0113	0.0064	-0.0079	-0.0262	-0.0201
(log)	(0.0121)	(0.0087)	(0.0191)	(0.0252)	(0.0210)
total_lead_Synsimilarity	0.0227**	0.0038	0.0237	0.0464*	0.0369
(log)	(0.0111)	(0.0100)	(0.0210)	(0.0282)	(0.0236)
total_lead_src3_keyword	-0.0279***	-0.0170**	-0.0488***	-0.0655***	-0.0325***
(log)	(0.0105)	(0.0071)	(0.0132)	(0.0177)	(0.0121)
total_lead_src1_keyword	0.0039	-0.0037	0.0061	-0.0000	-0.0022
(log)	(0.0061)	(0.0062)	(0.0136)	(0.0175)	(0.0149)
total_lead_src2_keyword	0.0223**	0.0191**	0.0562***	0.0852***	0.0501***
(log)	(0.0097)	(0.0083)	(0.0151)	(0.0207)	(0.0159)
total_rpl_ISCsimilarity	0.0028	0.0052	0.0238***	0.0298***	0.0259***
(log)	(0.0041)	(0.0048)	(0.0079)	(0.0106)	(0.0087)
total_rpl_Synsimilarity	-0.0067*	-0.0043	-0.0263***	-0.0344***	-0.0249***
(log)	(0.0038)	(0.0044)	(0.0071)	(0.0098)	(0.0080)
total_rpl_src3_keyword	0.0023	0.0019	0.0213	0.0330*	0.0245
(log)	(0.0068)	(0.0073)	(0.0138)	(0.0180)	(0.0154)
total_rpl_src1_keyword	0.0067	-0.0032	0.0061	-0.0089	-0.0007
(log)	(0.0101)	(0.0089)	(0.0186)	(0.0236)	(0.0194)
total_rpl_src2_keyword	-0.0060	-0.0020	-0.0204	-0.0206	-0.0166
(log)	(0.0074)	(0.0086)	(0.0196)	(0.0243)	(0.0209)
total_nonsec_leading_pos	-0.0189	-0.0256	-0.0264	-0.0046	-0.0255
ts (log)	(0.0239)	(0.0160)	(0.0321)	(0.0398)	(0.0335)
Constant	0.0234***	0.0298***	0.0436***	0.0721***	0.0443***
	(0.0013)	(0.0012)	(0.0016)	(0.0023)	(0.0017)
Observations	4,447,337	4,447,337	4,447,337	4,447,337	4,447,337
Adjusted R-squared	0.1714	0.0742	0.0869	0.1041	0.0773

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses

Table B23. The Impacts of CMA Enforcement on Users with Non-Zero Cybersecurity-Relevant Leading Posts before CMA

	Quantity		Relevance		
VARIABLES	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0111***	-0.0081***	-0.0140***	-0.0191***	-0.0158***

	(0.0011)	(0.0007)	(0.0013)	(0.0018)	(0.0014)
Observations	946,655	946,655	946,655	946,655	946,655
Adjusted R^2	0.2563	0.1741	0.2210	0.2574	0.2119
Weekly fixed effects and user fixed effects are included					

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

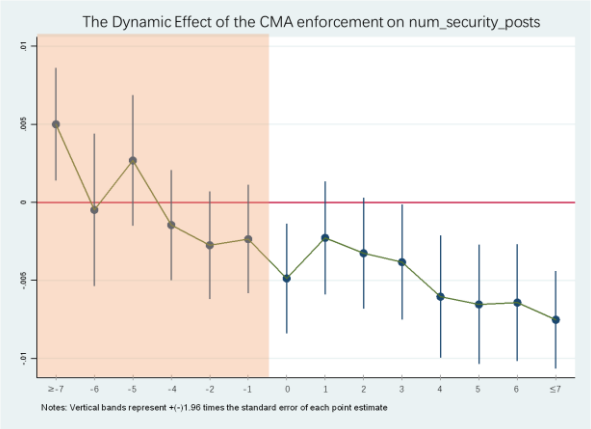
Table B24. Robustness Check Using PSM Matched Users with Non-Zero Cybersecurity-Relevant Leading Posts before CMA

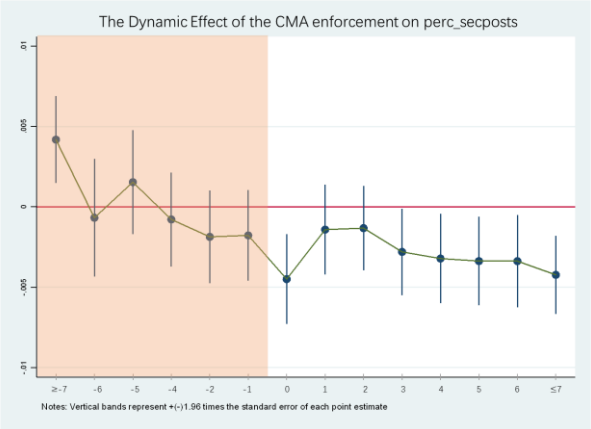
VARIABLES	**Quantity**		**Relevance**		
	(1) num_security_posts	(2) perc_secposts	(3) src1_keyword_counts	(4) src2_keyword_counts	(5) src3_keyword_counts
Treat×CMA	-0.0112***	-0.0085***	-0.0160***	-0.0272***	-0.0208***
	(0.0026)	(0.0016)	(0.0035)	(0.0049)	(0.0036)
Observations	826,335	826,335	826,335	826,335	826,335
Adjusted R^2	0.2662	0.2061	0.2327	0.2734	0.2233
Weekly fixed effects and user fixed effects are included					

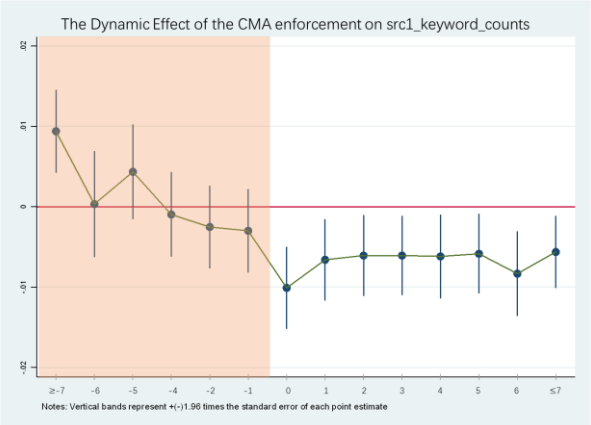
*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

Figure B1. Parallel Trend Tests

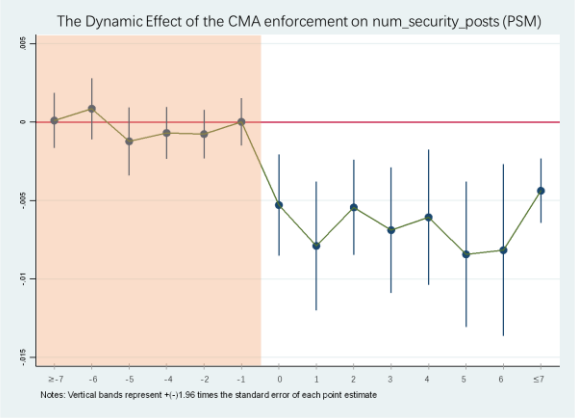
(1) Full Sample

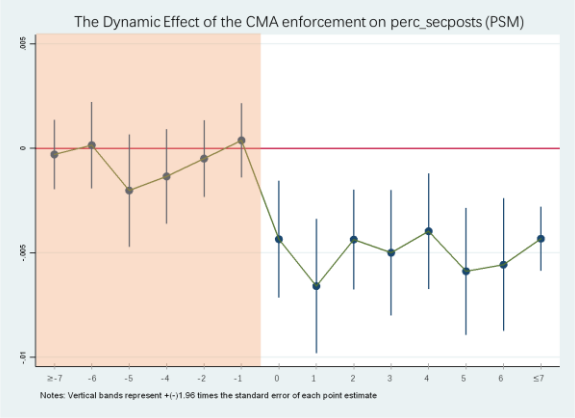


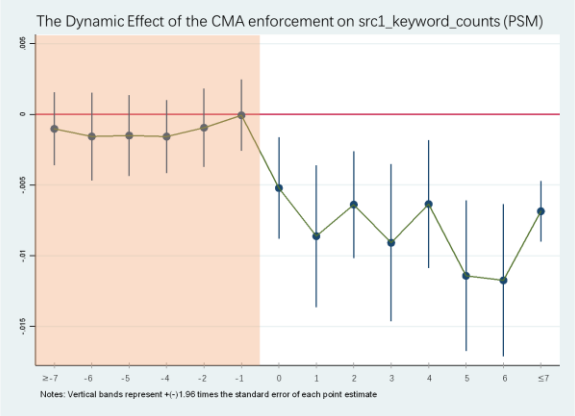


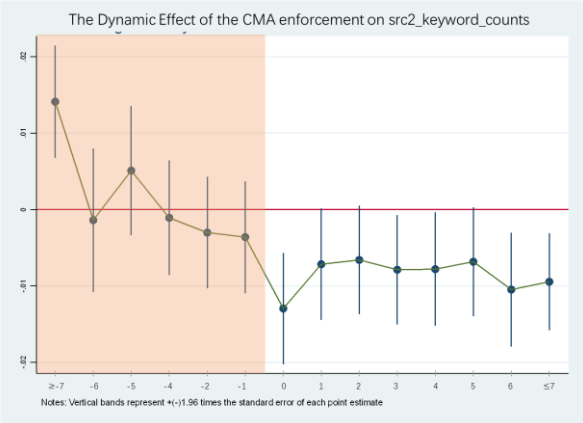


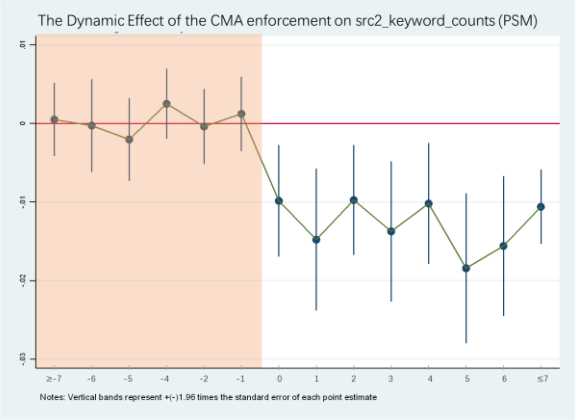
(2) Matched Sample

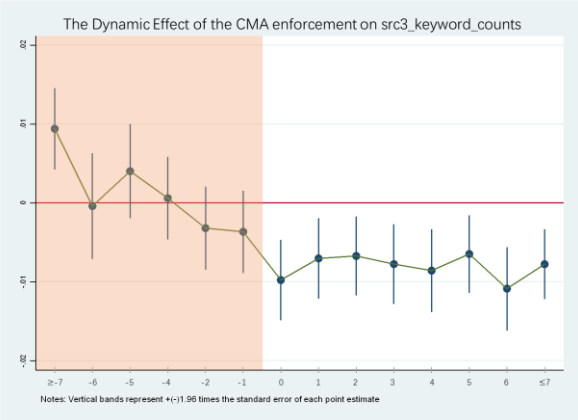


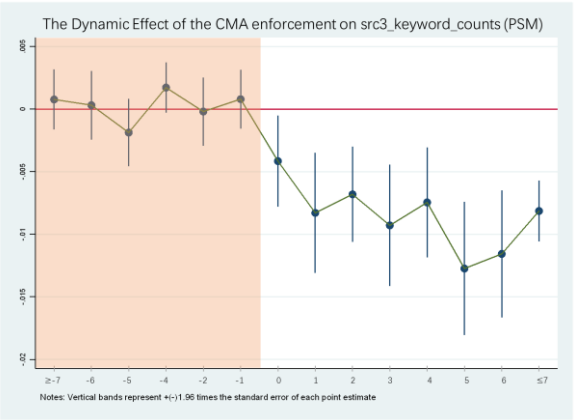








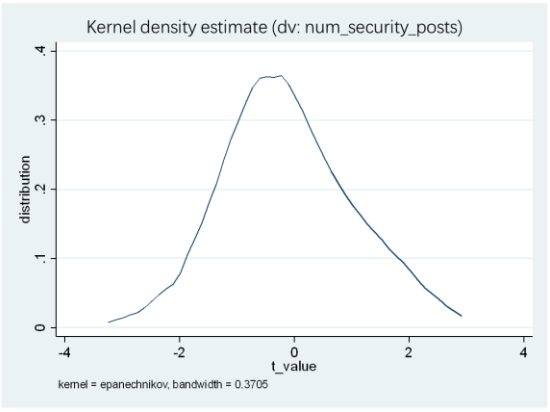




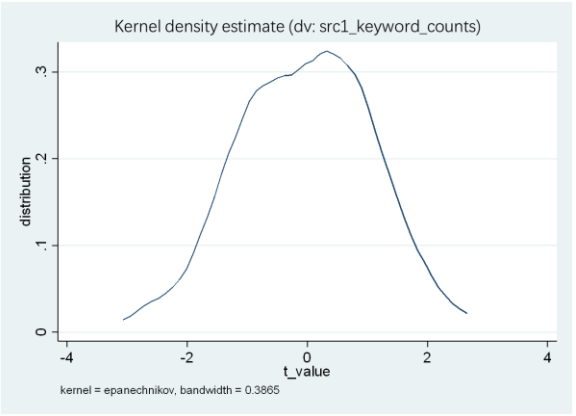
Note: Vertical bands represent +(-)1.96 times the standard error of each point estimation.

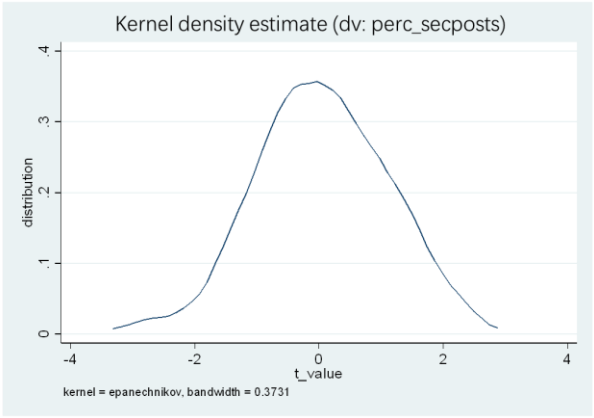
Figure B2. Kernel Density Estimation of Regression t-value of Pseudo Treatment

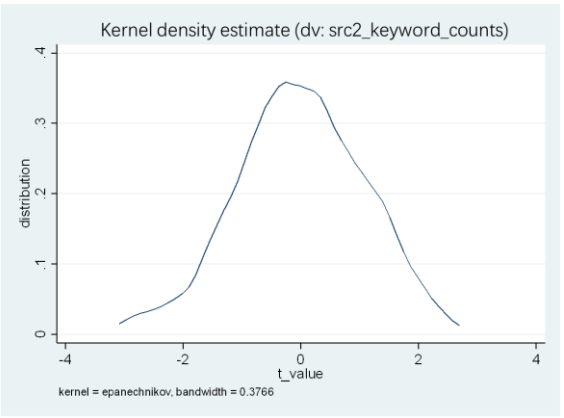
Quantity of cybersecurity-relevant discussions

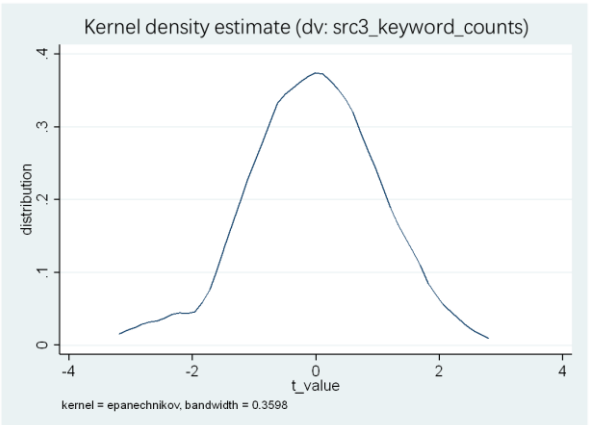


Extent of relevance to cybersecurity in discussions





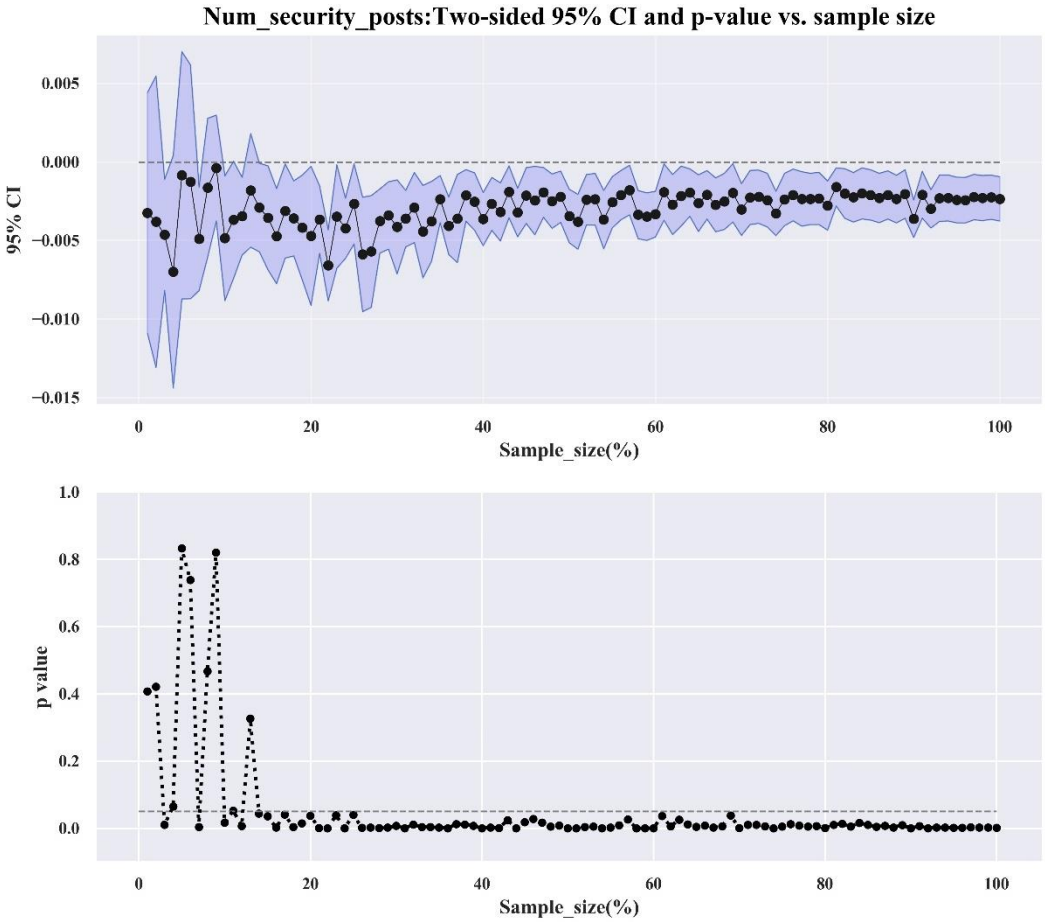




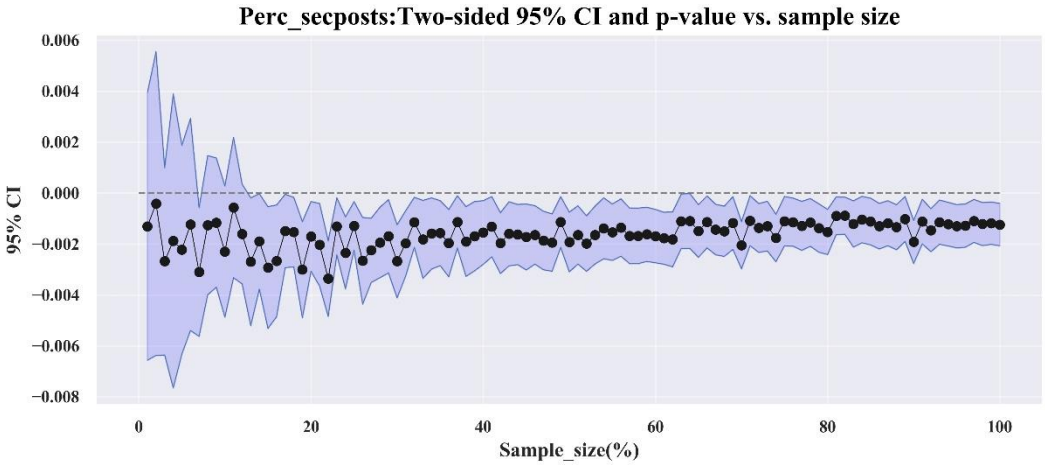
The following charts are created by recurrently sampling larger sizes from the full sample, rerunning the statistical model with each sample, determining the coefficients and p-values of interest, and graphically representing them. It helps in checking the robustness of large sample estimation (Lin et al. 2013).

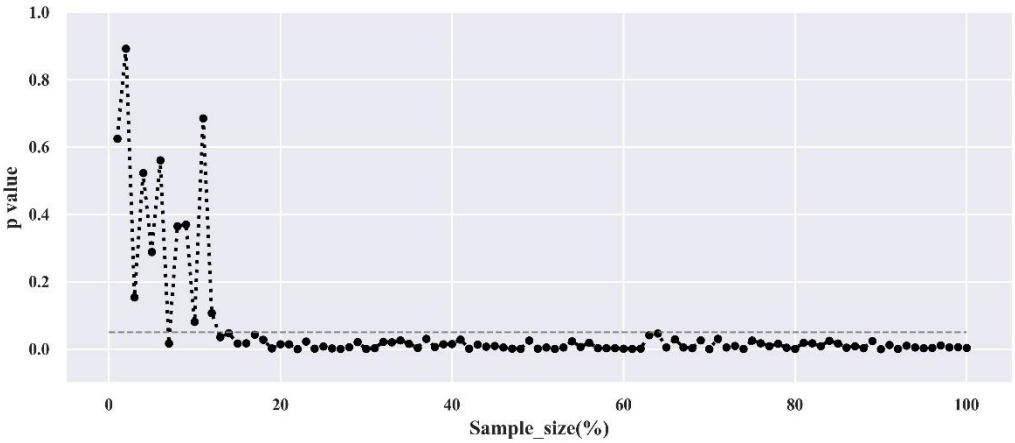
Figure B3. Two-Sided 95% Confidence Interval (Top) and p-Value (Bottom) for Chilling Effect vs. Sample Size

(a) Chilling Effect on Number of Cybersecurity-relevant Leading Posts

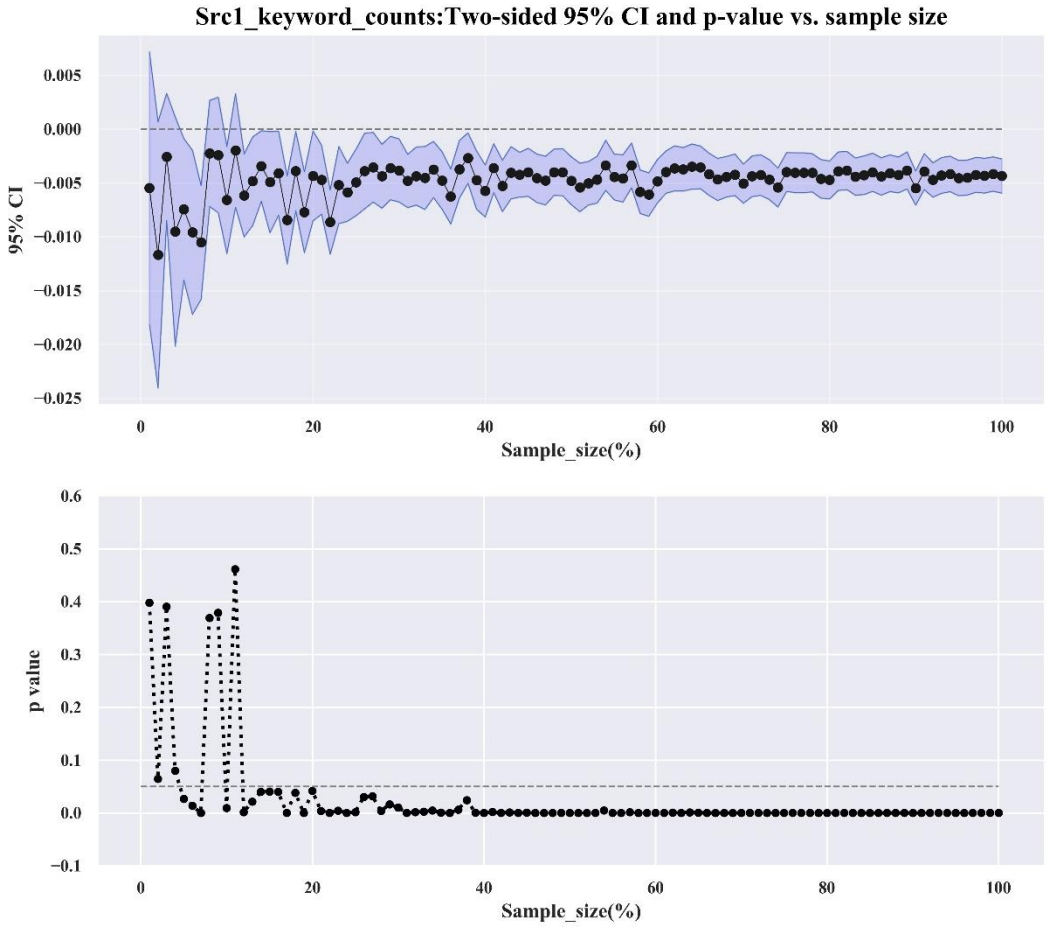


(b) Chilling Effect on Proportion of Cybersecurity-Relevant Leading Posts

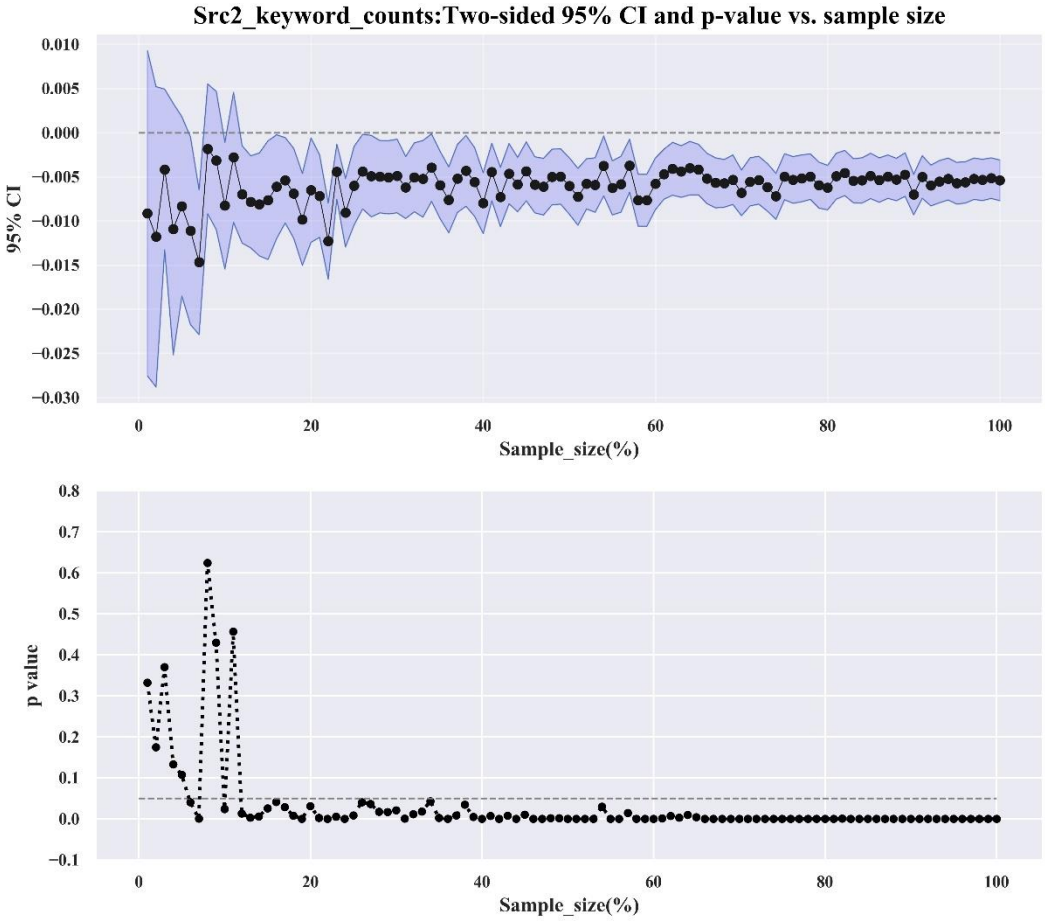




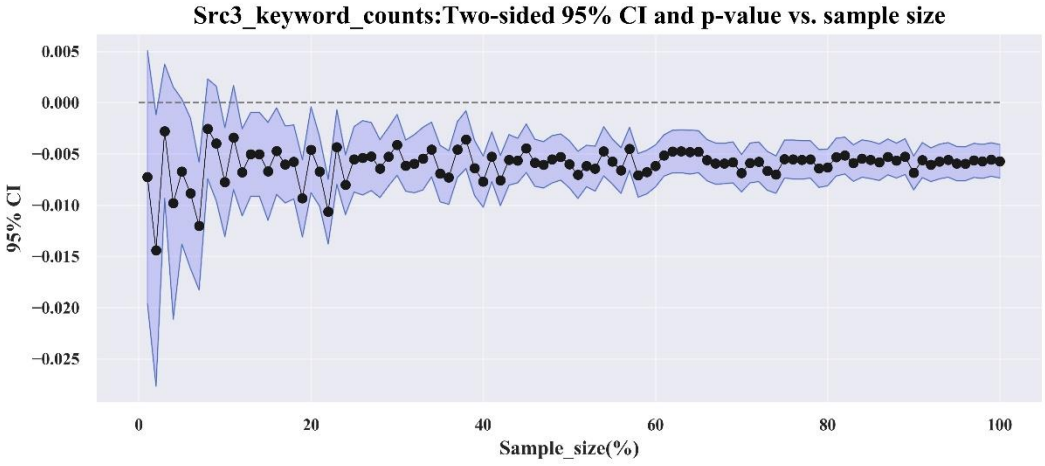
(c) Chilling Effect on Number of Cybersecurity Keywords (Referring to Source 1)



(d) Chilling Effect on Number of Cybersecurity Keywords (Referring to Source 2)



(e) Chilling Effect on Number of Cybersecurity Keywords (Referring to Source 3)



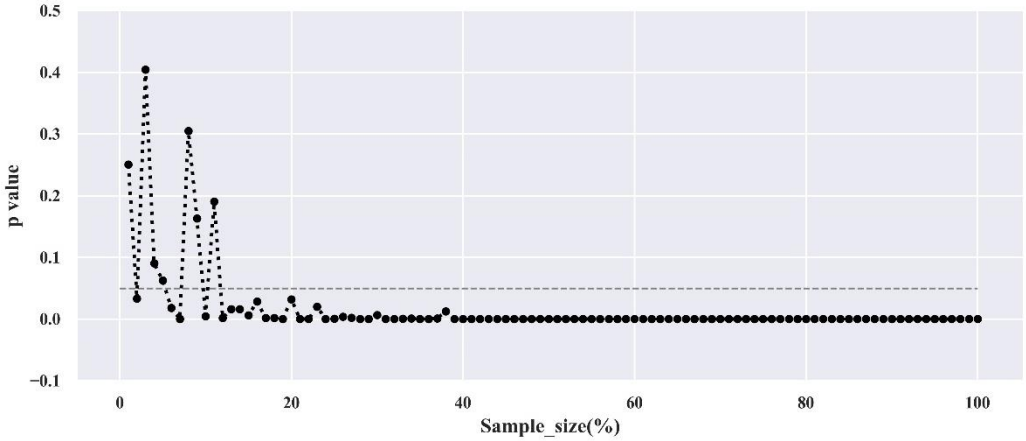
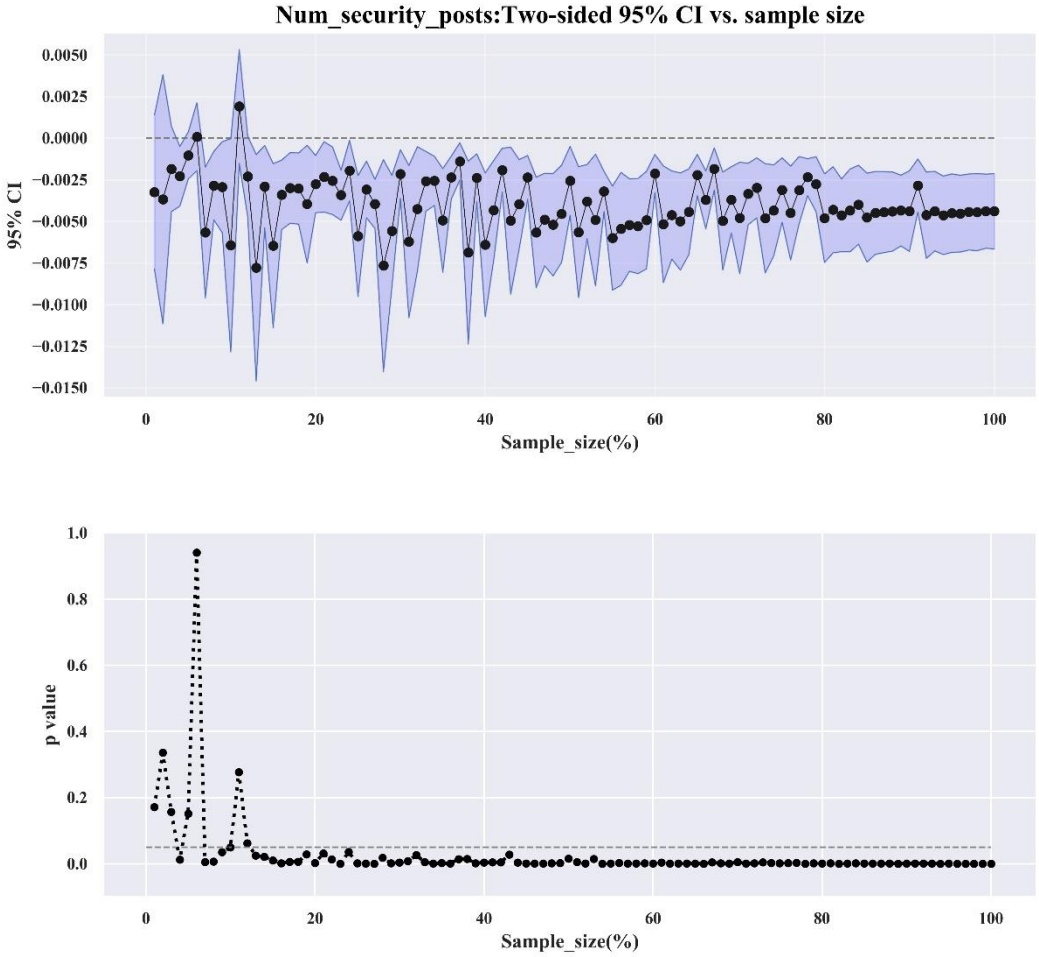
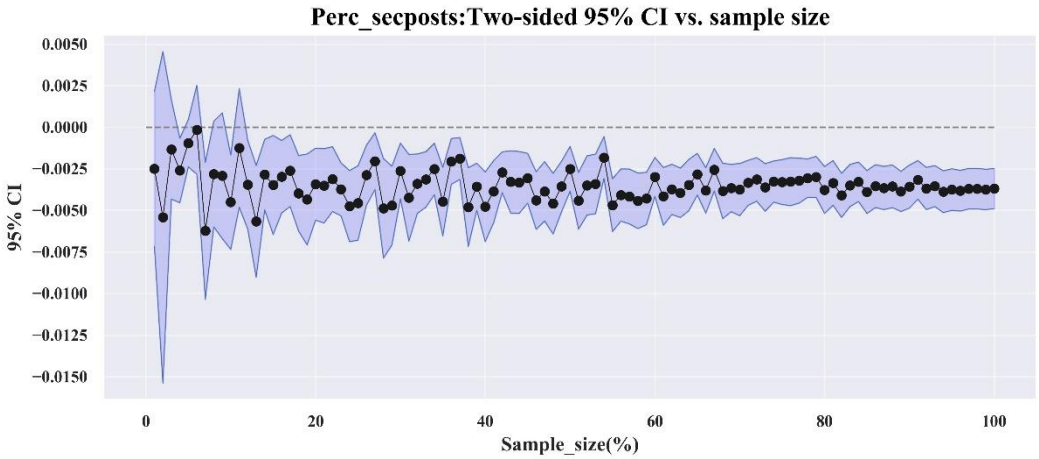

0.5
0.4
0.3
0.2
0.1
0.0
−0.1
p value
0
20
40
60
80
100
Sample_size(%)

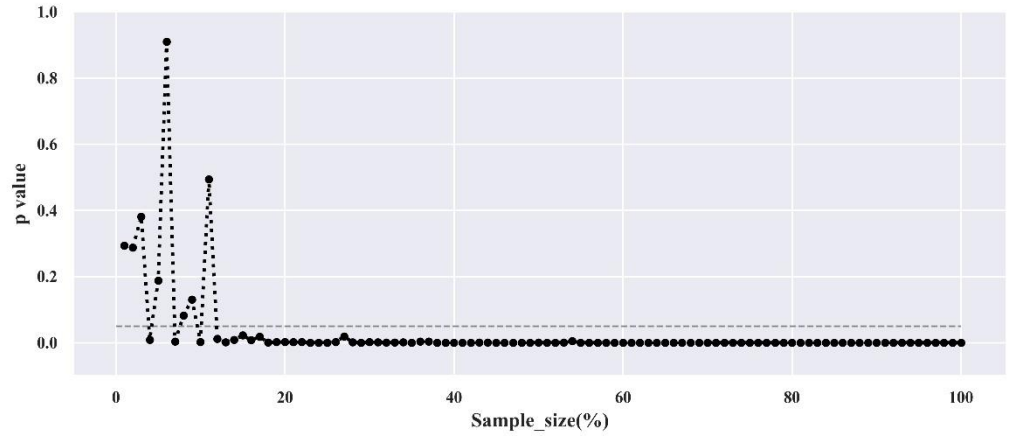
Figure B4. Two-Sided 95% Confidence Interval (Top) and p-Value (Bottom) for Chilling Effect vs. Sample Size based on PSM sample

(a) Chilling Effect on Number of Cybersecurity-relevant Leading Posts

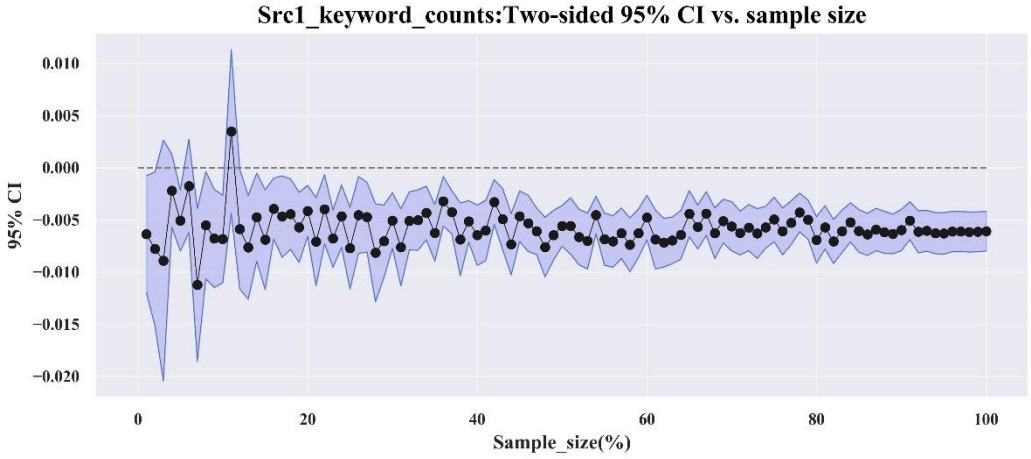


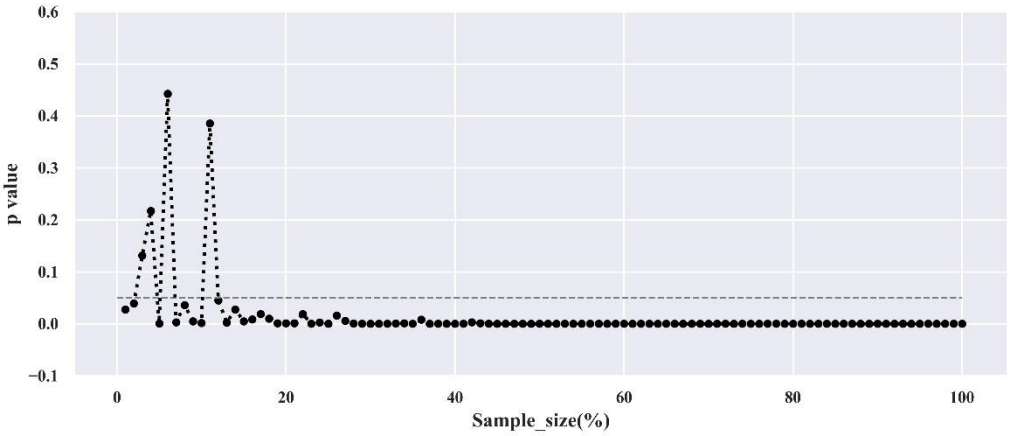
(b) Chilling Effect on Proportion of Cybersecurity-Relevant Leading Posts



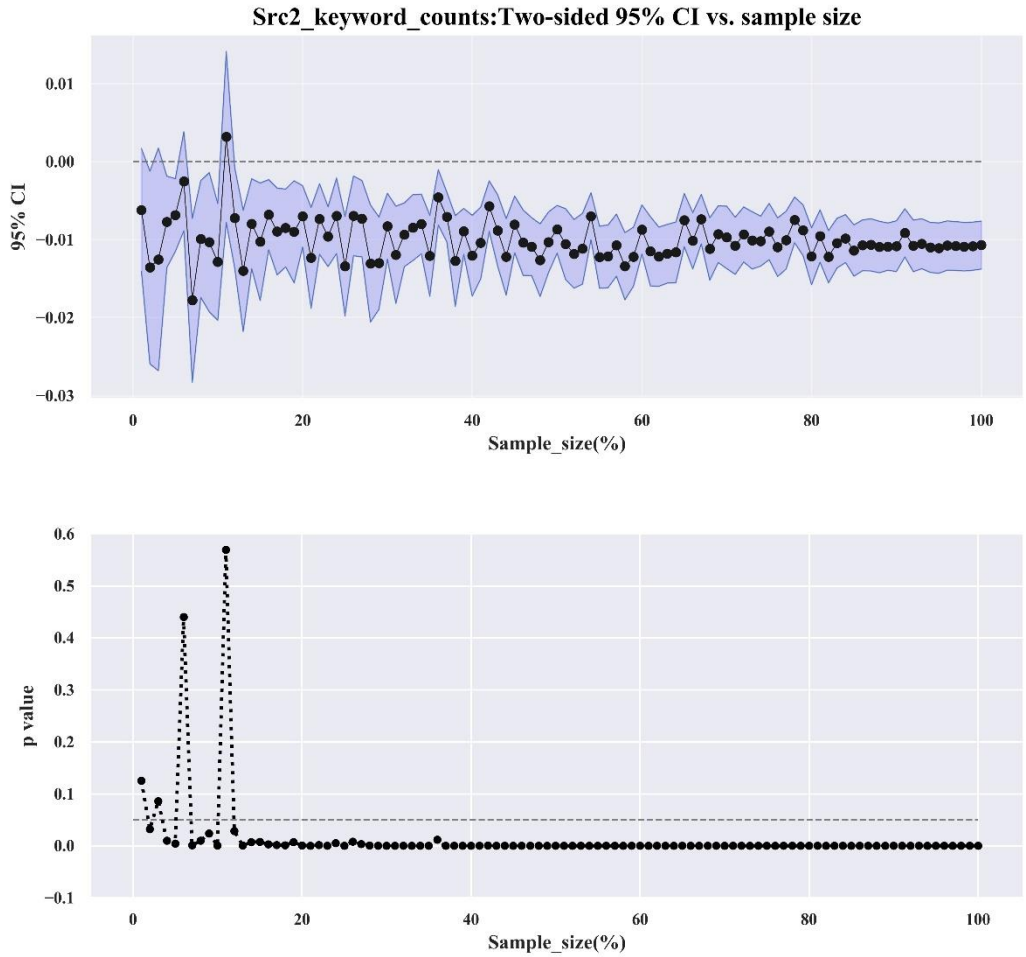


(c) Chilling Effect on Number of Cybersecurity Keywords (Referring to Source 1)

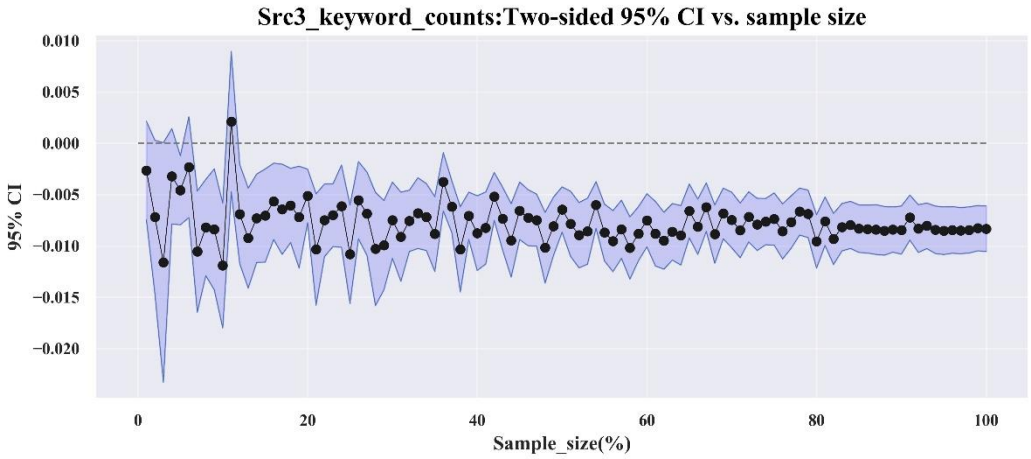


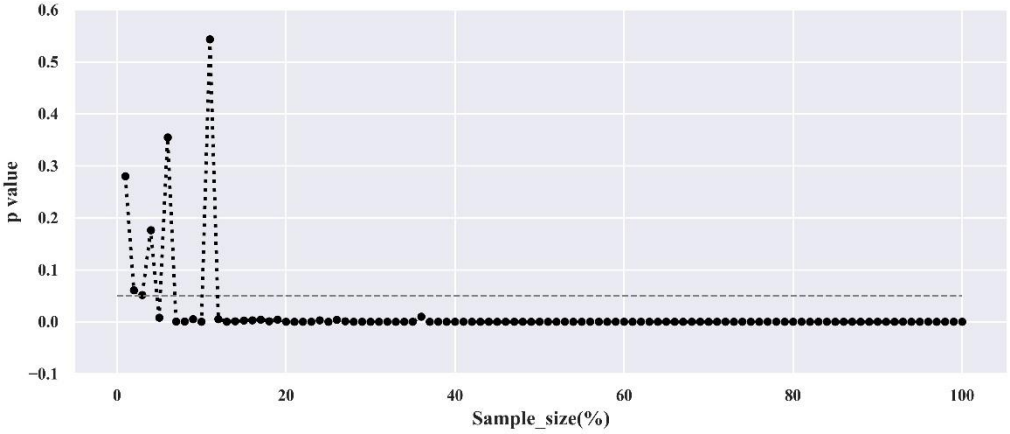


(d) Chilling Effect on Number of Cybersecurity Keywords (Referring to Source 2)



(e) Chilling Effect on Number of Cybersecurity Keywords (Referring to Source 3)




0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
p value
0
20
40
60
80
100
Sample_size(%)

Appendix C. Procedures for Translation between Chinese and English

According to our research design, we have implemented two types of translation between Chinese and English. We describe the procedures that we took to ensure the accuracy of translation below.

C. 1. Cybersecurity Keywords Translation from English to Chinese

We collected the glossary of cybersecurity keywords from three sources in English. We first searched the Chinese expressions of English cybersecurity terms using Google Translate and obtained synonyms and similar expressions from Xinhua Dictionary API using Python Package lightDict. Google Translate has been known to be one of the most popular machine translation tools and has also been widely used in multilingual translation in darknet research (Chen 2011). Xinhua Dictionary is the best-selling Chinese language dictionary with more than 260,000 entries in Chinese characters and words. To complement the Xinhua Dictionary with more online buzzwords, we also searched on Jikipedia.com, the largest Chinese online buzzwords dictionary with around 30,000 entries. A search returns the Top 5 expressions with meaning closest to the focal cybersecurity term. For example, when we searched "破解" (cracking), it returned the following synonyms: "BT 版," "pj," "pojie," "盗版," "白嫖." Drawing on this process, we could detect synonyms for the cybersecurity keywords in Chinese and also the online slang with the meanings to the focal cybersecurity keywords to complement the query from translators. All Chinese synonyms of the English cybersecurity keywords were double-checked by cybersecurity experts from Singapore who can speak both languages quite fluently and communicate cybersecurity knowledge in both languages. We recruited these experts from Singapore because the main language in education in Singapore is English but many Singaporeans also learn and speak Chinese, which ensures these experts are fluent in both languages.

C. 2. Archive Translation from English to Chinese

We referred to reputational cybersecurity archives generated by cybersecurity professionals from two sources, the Diary Archive provided by SANS Internet Storm Center (ISC) and the Internet Security Threat Reports provided by Symantec. The translation of archives written in English into Chinese is done directly by Google Translate and also edited by cybersecurity experts from Singapore to ensure the meaning of the exchange from Chinese to English is not lost or changed in translation. The cosine similarity scores (called ISC_similarity and Symantec_similarity) were calculated, respectively, by comparing the content of each leading post in Chinese forums with the pool of translated ISC diaries published within the same month and with the translated Symantec threat reports published in the same year. Despite some language gaps, these archives are expected to serve as a good benchmark for cybersecurity professionals' content for Chinese users for the following three reasons.

First, no obvious gap in cybersecurity knowledge exists between users in Chinese hack forums and users in English forums. Cybersecurity education in China is highly synchronized with the global cybersecurity knowledge community. In the computer science and security curriculums of universities in China, most of their cybersecurity textbooks are authorized translations of well-established English

textbooks. Moreover, much cybersecurity terminology is taught directly in English without ever being first translated into Chinese terms. Additional evidence can be found in the online information sharing and vulnerability disclosure activities conducted by CNCERT/CC, the National Computer Network Emergency Response Technical Team/Coordination Center of China. As stated in their official website, "CNCERT/CC has its presence in 31 provinces, autonomous regions, and municipalities across mainland China. CNCERT/CC coordinates with key network operators, domain name registrars, cybersecurity vendors, academia, civil society, research institutes and other CERTs to jointly handle significant cybersecurity incidents in a systematic way." As one of the earliest members in FIRST (the Forum of Incident Response and Security Teams, the most influential international cybersecurity information sharing community) since August 7, 2002, CNCERT/CC has for decades synchronized its vulnerability disclosure and updates with the most well-known vulnerability databases in the English cybersecurity community, for example, https://vuldb.com/ and https://www.cvedetails.com/. Therefore, any language barrier that may exist between Chinese hack forum users and English hack forum users does not affect their access to the same knowledge of cybersecurity.

Second, because cybersecurity is a global phenomenon and is based on common computer architectures and software platforms that have been adopted globally, our translation in this domain could be done with a reasonable level of accuracy. The main purpose of using these sources of cybersecurity glossaries and archives is to have an objective benchmark for both English and Chinese hack forums to examine whether the discussions in hack forums are more relevant to cybersecurity terminology and more similar to the archives generated by cybersecurity professionals. The language barriers between English cybersecurity achieves (or keywords) and Chinese hack forum posts can be solved by executing translation tasks, such as a machine translation-based approach, a corpus-based approach, and a dictionary-based approach (Zhou et al. 2005). Among them, Google Translation is known to be one of the most popular machine translation tools (http://code.google.com/apis/ajaxlanguage/documentation/#Translation). We referred to a study by Tsai (2022), which investigates the effectiveness of using Google Translate as a translingual CALL tool for English as a Foreign Language (EFL) writing and is keyed to the perceptions of both more highly proficient Chinese students majoring in English at Chinese universities and less proficient students who were not majoring in English. This study suggests students using Google Translate as a revision tool displayed better L2 performance in written language and content than in their self-writing; non-English major students showed significantly more positive attitudes toward the use of Google Translate than English major students. Thus, we have more confidence in the translation quality of Google Translate, complemented with the editing by cybersecurity experts who are proficient with both the English and Chinese languages.

Lastly, during our study period, internet users in China could always access the up-to-date and translated version of the annual reports of internet security threats published by Symantec, which, chosen as one of the two benchmarks for calculating similarity scores in our study, is considered one of the most important information security reports for both professionals and university students.

Appendix D. Possible Censorship in the Chinese and English Forums

In this section, we describe our robustness checks to examine whether any removed posts by forums could have biased our main results on the chilling effect. In both Chinese and English forums, we noted that the content in some leading posts were emptied by forum administrators because of violation of the forums' moderation rules. Although we cannot ensure all the "censored" leading posts are moderated because of concerns for CMA prosecution, we used these empty posts as a proxy for removed posts. Specifically, based on a DID approach, we checked (1) whether the censorship on Chinese forums has significantly shifted before and after CMA, compared with the English forum at both the forum and user levels, and (2) whether our results might change if we included these censored posts into our main dependent variable, cybersecurity-relate leading posts. We confirmed that no bias exists as a result of possible censorship by administrators after CMA, and the chilling effect remains significant when the censored posts are included. For this analysis, we need to highlight two points about the possible removal of leading posts in the Chinese and English forums.

1) Both self-moderation by hack forum users and moderation or censorship by community administrators reflect the chilling effect and result in a decrease in the volume of cybersecurity-relevant content, which is one of our observable outcome variables. Admittedly, our estimate of the chilling effect might be biased if any possible censorship by community administrators treated posts differently before and after CMA enforcement. In other words, we need to examine whether administrators only deleted suspicious posts generated after CMA or cleared all suspicious content from their databases whenever it was released. According to the articles of CMA, any malicious content still existing in hack forums after CMA could be considered as sharing or disseminating computer misuse tools and the forum would be subject to prosecution. Thus, the hack forums, if they had such material, would censor all the content existing in the database, whereas after CMA, hack forum users need only to do self-moderation.

2) Administrators in both Chinese and English forums had the right to transfer, remove, or delete the offending content. We found the moderation rules announced in one of the Chinese forums state, "The means of dealing with violators include: locking or removing the offending content; admonishing or warning the violator; deducting the violator's prestige; retrieving the medal; implementing a temporary ban or a permanent ban on the violator's account (blocking ID); implementing an IP ban (blocking IP)." Apart from deleting the entire thread, administrators have alternatives of just removing part of the offending content or closing the thread. We have found some leading posts marked as "Your post violates the moderation rules and has been locked! Please read the moderation rules before posting ....." in the other Chinese forum. Some leading posts, including labels saying [closed by admin] or [removed by admin] or [deleted by admin], were found in English hack forum. Given that those censored posts still exist, we can generate a proxy of the moderation behavior of administrators in both Chinese and English hack forums.

   We have applied a text analysis method to recognize all the "censored" leading posts moderated by forum admins from both English and Chinese forums. For the users included in our study, there were 5,669 leading posts that were moderated and remained in these forums during our

study period. Although we cannot be sure all the "censored" leading posts are moderated because of concerns for CMA prosecution, at least we can take advantage of the data to examine whether the censorship on Chinese forums has significantly shifted before and after CMA compared with the English forum. If not, we will largely alleviate any concerns about any removed posts.

Given the two points above, we first compared the total number of censored posts and the percentage of censored posts in both Chinese and English forums (Figure D1) and conducted a DID analysis at the forum level as well as at the individual level (Table D1). The percentage of censored posts is calculated as num_censored_posts/(num_censored_posts + num_security_posts), which aims to examine whether the standard of censorship by admins has significantly changed before and after CMA.

Figure D1. Monthly Trends of the Number of Censored Leading Posts in Hack Forums

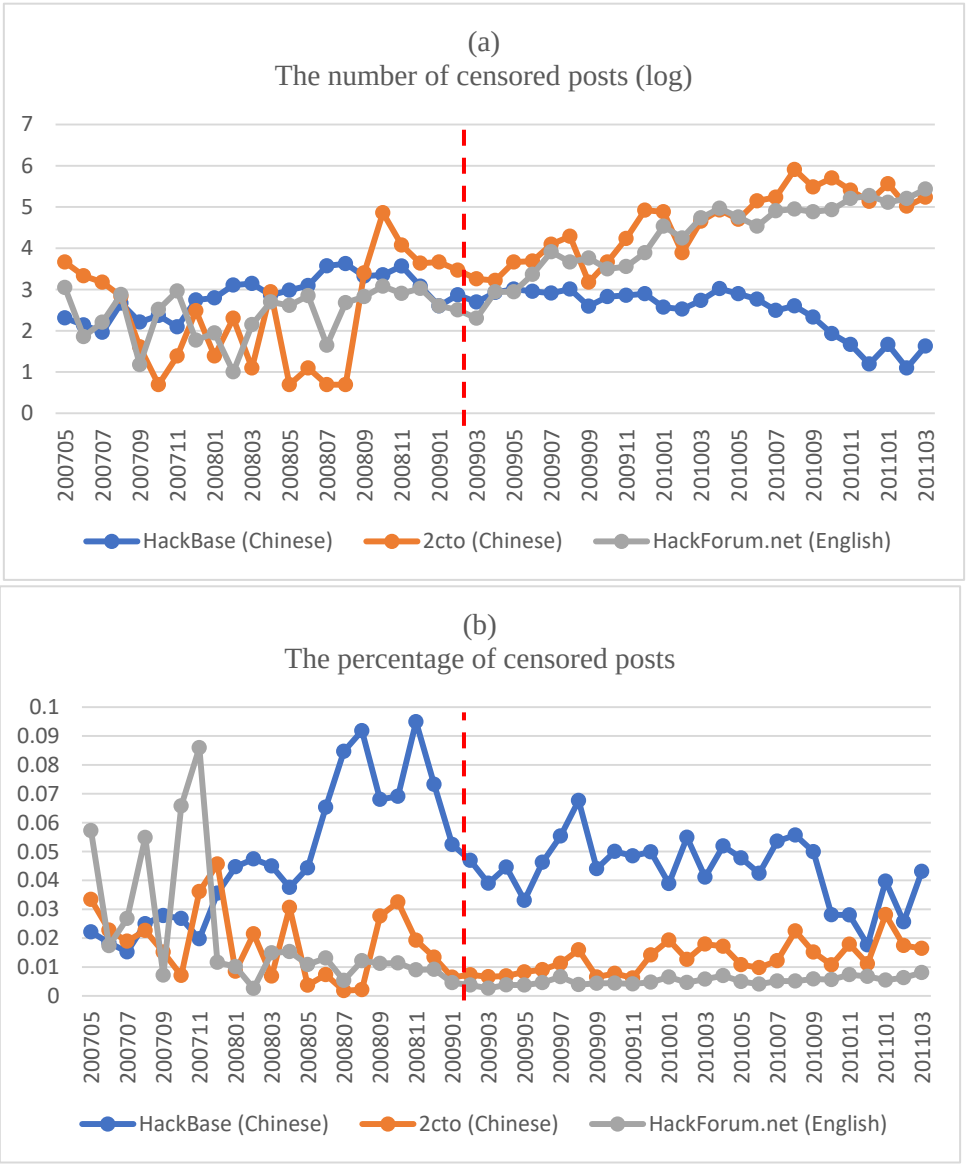


Table D1. DID Analysis on the Number and the Percentage of Censored Posts

VARIABLES	(1) num_censored_posts (forum level)	(3) perc_censored_posts (forum level)	(3) perc_censored_posts (individual level)	(4) perc_censored_posts (PSM individual level)
Treat×CMA	-0.9534	-0.0565	-0.0000	-0.0001
	(0.6543)	(0.0891)	(0.0000)	(0.0000)
Observations	612	612	8,543,852	4,447,337
Adjusted R-squared	0.6931	0.3533	0.9120	0.9069
Weekly fixed effects and user fixed effects are included				

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

The results in Figure D1 together with Table D1 suggest that neither the count of censored posts nor the percentage of censored posts exhibits any significant difference before and after the CMA enforcement. To summarize, there exists no bias because of possible moderation and censorship by administrators after the CMA enforcement.

Additionally, we also conducted robustness checks by including those censored posts as cybersecurity-relevant ones. In Table D2, we found that the chilling effect remains significant when those censored posts moderated by forum administrators are considered. Nevertheless, we would like to highlight that excluding those censored posts helped us to tease from the deterrence effect the chilling effect on desirable (or legal) behaviors.

Table D2. Robustness Checks by Including Censored Posts as Cybersecurity-Relevant Posts

VARIABLES	(1) num_security_ posts	(2) perc_secposts	(3) num_security_ posts (PSM)	(4) perc_secposts (PSM)
Treat×CMA	-0.0056***	-0.0032***	-0.0059***	-0.0032***
	(0.0010)	(0.0003)	(0.0011)	(0.0009)
Observations	8,543,852	8,543,852	4,447,337	4,447,337
Adjusted R-squared	0.2493	0.1682	0.2984	0.2347
Weekly fixed effects and user fixed effects are included				

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. The estimates for controls were omitted for brevity.

References

Angrist JD, Pischke J-S (2008) Parallel worlds: Fixed effects, differences-in-differences, and panel data. *Mostly harmless econometrics* (Princeton University Press), 221-248.

Calder A, Watkins SG (2007) *A dictionary of information security terms, abbreviations and acronyms* (IT Governance Ltd, Cambridgeshire, UK).

Chen H (2011) *Dark Web: Exploring and Data Mining the Dark Side of the Web* (Springer, New York, NY).

Chen K-H, Chen H-H (2001) Cross-language Chinese text retrieval in NTCIR workshop: towards cross-language multilingual text retrieval. *ACM SIGIR Forum* (ACM New York, NY, USA), 12-19.

Du L, Zhang Y, Sun L, Sun Y, Han J (2000) PM-based indexing for Chinese text retrieval. *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, 55-59.

Hallgren KA (2012) Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology* 8(1):23.

Hwang EH, Singh PV, Argote L (2015) Knowledge Sharing in Online Communities: Learning to Cross Geographic and Hierarchical Boundaries. *Organization Science* 26(6):1593-1611.

Kim Y (2014) Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics), 1746-1751.

Kshetri N (2010) *The global cybercrime industry: economic, institutional and strategic perspectives* (Springer).

Lin M, Lucas Jr HC, Shmueli G (2013) Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research* 24(4): 906-917.

Mikhaylov A, Frank R (2016) Cards, Money and Two Hacking Forums: An Analysis of Online Money Laundering Schemes. *2016 European Intelligence and Security Informatics Conference (EISIC)*.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111-3119.

Nakagawa H, Kojima H, Maeda A (2004) Chinese term extraction from Web pages based on compound term productivity. *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing*, 79-85.

Park AJ, Frank R, Mikhaylov A, Thomson M (2018) Hackers hedging bets: a cross-community analysis of three online hacking forums. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 798-805.

Porter MF (1980) An algorithm for suffix stripping. *Program*.

Qiu X, Zhang Q, Huang X-J (2013) Fudannlp: A toolkit for Chinese natural language processing. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 49-54.

Rao A, Spasojevic N (2016) Actionable and political text classification using word embeddings and LSTM. *arXiv preprint arXiv:.02501*.

Rogers MK (2006) A two-dimensional circumplex approach to the development of a hacker taxonomy. *Digital Investigation* 3(2):97-102.

Salton G, Wong A, Yang C-S (1975) A vector space model for automatic indexing. *Communications of the ACM* 18(11):613-620.

Siu GA, Collier B, Hutchings A (2021) Follow the money: The relationship between currency exchange and illicit behaviour in an underground forum. *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*.

Thomas J (2005) The moral ambiguity of social control in cyberspace: a retro-assessment of the 'golden age' of hacking. *New Media & Society* 7(5):599-624.

Tsai S-C (2022) Chinese students' perceptions of using Google Translate as a translingual CALL tool in EFL writing. *Computer Assisted Language Learning* 35(5-6):1250-1272.

Welser HT, Gleave E, Gleave E, Fisher D, Smith MA (2007) Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure* 8(2):1-32.

Wong K-F, Li W, Xu R, Zhang Z-s (2009) Introduction to Chinese natural language processing. *Synthesis Lectures on Human Language Technologies* 2(1):1-148.

Yang Y (1995) Noise reduction in a statistical approach to text categorization. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 256-263.

Yue WT, Wang Q-H, Hui KL (2019) See no evil, hear no evil? Dissecting the impact of online hacker forums. *MIS Quarterly* 43(1):73-95.

Zhang J, Ackerman MS, Adamic L (2007) Expertise networks in online communities: structure and algorithms. *Proceedings of the 16th international conference on World Wide Web*.

Zhang X, Tsang A, Yue WT, Chau M (2015) The classification of hackers by knowledge exchange behaviors. *Information Systems Frontiers* 17(6):1239-1251.

Zhao Z, Liu T, Li S, Li B, Du X (2017) Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 244-253.

Zhou Y, Reid E, Qin J, Chen H, Lai G (2005) US domestic extremist groups on the Web: link and content analysis. *IEEE Intelligent Systems* 20(5):44-51.

Zou F, Wang FL, Deng X, Han S (2006) Evaluation of Stop Word Lists in Chinese Language. *LREC*, 2497-2500.