Appendix A. Data Extraction and Processing of Hacker-Forum Posts

We processed the data from six forums involving three languages:

- English: Hackforums.net (Hackforums) and Hellboundhackers.org (HBH)
- Chinese: hackbase.com (Hackbase) and 2cto.com (HHLM)
- Russian: Antichat.ru (Antichat) and Xaker.name (Xaker)

All six of these forums focus on hacking discussion and are publicly accessible. They are regarded as important forums in their respective languages. In particular, under the computer/hacking category in Alexa, Hackforums ranked second (ranking was accessed on July 3, 2015) and HBH ranked 18th (ranking was accessed on August 2, 2016). The two forums are the highest ranked online forums on the list.[1]

As for the Chinese forums, Hackbase was ranked second in Alexa under the hacker category in Chinese (ranking was accessed on June 29, 2015) and HHLM was ranked fifth in the information-security category by ChinaZ.com (ranking was accessed on June 29, 2015).

Antichat is considered one of the oldest and most influential hacker forums in Russia. Inexplicably, it is not categorized by Alexa, but its domestic ranking in Alexa is higher than other Russian categorized hacker forums. Xaker was ranked fourth by Alexa under the hacker category in Russian (ranking was accessed on June 29, 2015). Table A0.1 reports the latest rankings by Alexa.

Although the languages are different, we followed the same data extraction and processing procedures for all the forums. The main objectives of the data processing are:

[1] There are other newer forums, such as EvilZone.org. However, we cannot use them because they do not match the DDOS-attack data in 2007–2011.

- Extract posts mentioning a port.
- Extract posts related to DDOS attacks.
- Extract posts related to specific DDOS-attack techniques.
- Identify orientation of the discussion content.

Table A0.1 Hacker Forum Traffic Ranking (Source: Alexa, accessed at January 22, 2017)

	Ranking in country	Rank in category	Notes	Category
Hackforums	5,727 (US)	3	1st in the subcategory of Chats and Forums	Computers > Hacking
HBH	179,550 (US)	17	7th in the subcategory of Exploits	Computers > Hacking
Hackbase	5,353 (CN)	5		World > Chinese Simplified CN > Computers > Security > Hacker
HHLM	814 (CN)	1		World > Chinese Simplified CN > Computers > Security > Hacker
Antichat	2,826 (RU)	—		This site has not been categorized
Xaker	44,071 (RU)	8		World > Russian > Computer > Hacking

The following sections explain the detailed procedures.

A1. Procedures for Port-Related Post Extraction

We developed a three-step process to identify port-related posts and extract the port numbers mentioned in the posts. In Step 1, we remove posts that are obviously not related to a port. Because this step involves removing posts classified as port unrelated, it is critical to ensure that very few port-related posts are wrongly removed. To check this and ensure a high recall rate, we apply the procedure to all test sets (see below).

In Step 2, we divide the potential posts into candidate port-related posts and irrelevant posts. We treat posts from the irrelevant set as port unrelated. We check whether this procedure gives a high recall rate with all the test sets.

In Step 3, we manually screen ALL posts within the candidate set to verify and extract the port numbers from the posts.

The details of these three steps are as follows.

1. Replace irrelevant numbers and remove irrelevant posts.
    1a. Replace known port-unrelated numbers, e.g., date, IP address, specialized terms containing numbers such as Win32, system32, sp2, etc., by some arbitrary symbols. For example, the IP address 127.0.0.1 is replaced by text *urlreplace*;
    1b. Remove posts that do not contain numbers.
2. Extract candidate posts.
    2a. Identify posts for the candidate set based on three rules.
        i. The post contains the term *port* or its variants (e.g., *ports, destination-port*);
        ii. The post contains keywords related to protocols (e.g., *TCP, UDP, Telnet,* etc.) and the corresponding port numbers (0–65535). We obtain the protocol list from the Internet Assigned Numbers Authority (IANA) – https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.txt [accessed February 16, 2017].
        iii. The post contains an IP address or URL that ends with numbers.
    2b. Classify the remaining posts as irrelevant.
3. Screen all candidate posts to identify and extract the port numbers.

Two research assistants (RA) with computer-security knowledge and extensive training manually read all candidate posts to extract the port numbers. They then compared their results and identified inconsistent posts (such as whether a post contains port numbers or inconsistent

recording of port numbers). We further asked the RAs to resolve the initial inconsistent extractions independently. They then compared their results again to resolve the inconsistencies. After the discussion, we obtained the finalized screening results for all posts.

A1.2 Test-Set Generation

To assess the accuracy of the port-extraction procedure developed above, we randomly generated three test sets. Two RAs with computer-security knowledge and extensive training read all test-set posts. They first determined whether a post mentions a port number and then discussed and resolved any inconsistencies in the extraction. In general, the inconsistent labeling mostly arose because of careless mistakes or typos. We report the details, including the total number of posts and actual number of port-related posts in each sample in Table A1.1, Columns (1) and (2).

A1.3 Evaluation of Port-Related Post Extraction

We applied the port-extraction procedure to each test set to evaluate its performance. An important indicator of performance is recall rate. A high recall rate means that our procedure can successfully identify port-related posts (equivalently, that it does not miss many port-related posts). We checked the recall rates after completing Steps 1 and 2 (Step 3 does not apply as the RAs read all posts in the test samples). In Step 2, we assumed that the posts that fall into the irrelevant group are port unrelated and checked what errors could cause a port-related post to be assigned into this group. The number of such misclassifications is reported in Table A1.1, Column (8). The recall rates in Column (9) reflect the percentage of the correctly classified posts.

Because we engaged two RAs to screen all posts in the candidate set, the precision rate was not an issue in our procedure.

Table A1.1 Recall Rates for Port-Related Post Extraction on Test Sets

	Posts	Port-related posts	Remaining posts after Step 1	Recall rate after Step 1	Posts classified as candidate posts after Step 2	Port-unrelated posts classified as irrelevant after Step 2	Port-related posts classified as candidate group after Step 2	Port-related posts classified as irrelevant group after Step 2	Recall rate after Step 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Hackforums									
Sample 1	8,415	62	2,426	100%	255	198	57	5	91.9%
Sample 2	7,860	36	2,376	100%	199	164	35	1	97.2%
Sample 3	8,685	56	2,623	100%	297	244	53	3	94.6%
HBH									
Sample 1	3,846	39	1,622	100%	128	93	35	4	90%
Sample 2	3,635	16	1,479	100%	101	89	12	4	75%
Sample 3	4,292	20	1,771	100%	128	110	18	2	90%
Hackbase									
Sample 1	4,431	93	2,804	100%	139	60	79	14	92%
Sample 2	3,810	87	2,750	100%	132	53	79	8	95%
Sample 3	4,716	54	3,381	100%	103	57	46	8	96%
HHLM									
Sample 1	1,479	85	1,336	100%	249	167	82	3	100%
Sample 2	1,456	74	1,318	100%	215	147	68	6	100%
Sample 3	1,437	80	1,318	100%	201	128	73	7	96%
Antichat									
Sample 1	5,822	41	5,494	100%	463	426	37	4	90%
Sample 2	3,597	30	3,307	100%	411	383	28	2	93%
Sample 3	6,983	29	6,468	100%	225	197	28	1	97%
Xaker									
Sample 1	2,373	66	2,375	100%	366	305	61	5	92%
Sample 2	3,285	68	3,287	100%	415	349	66	2	97%
Sample 3	2,229	62	2,229	100%	291	229	62	0	100%

Our extraction achieves recall rates that mostly exceed 90%. We then applied the procedure to all six forums. We engaged two RAs to screen all the extracted candidate posts for each of the six forums.

Table A1.2 reports the final extraction results. For hackforums.net, the 24,610 port-related posts are distributed among 15,202 threads which contain a total of 246,104 posts. We use these 246,104 posts as the forum dataset for subsequent data analysis.

Table A1.2: Final Dataset

	Posts	Candidate posts after the first two steps	Extracted port-related posts after the third step
Hackforums	2,960,893	91,481	24,610
HBH	63,300	1,924	302
Hackbase	1,733,924	12,507	5,884
HHLM	388,938	10,953	4,194
Antichat	1,211,268	43,096	9,588
Xaker	45,297	3,729	744

A2. Procedures for DDOS-Attack Post Extraction

We next extract posts related to DDOS attacks. Unlike with port-related posts where we can extract posts based on numbers and some specific keywords such as *port,* the challenge here is that we have to generate some keywords to associate the posts with DDOS attacks. We followed a four-step process.

In Step 1, we identify a number of articles related to DDOS attacks, including general discussions, techniques, and tools, from the Internet. The collection of these articles forms the DDOS-content database.

In Step 2, we extract and rank the keywords from these documents based on keyword frequency and number of documents containing the keywords after removing some common stop words. The highly ranked DDOS-specific keywords then form the basis for us to classify posts related to DDOS attacks.

In Step 3, we extract the posts using the generated keywords and assign a score to each forum post. We determine the score based on the aggregate count of DDOS-specific keywords in the posts. We then separate the posts into candidate and irrelevant sets. The candidate set contains posts that have a high score (with many DDOS-specific keywords). The irrelevant set contains posts that have a low score.

In Step 4, the two RAs manually screen all the posts in the candidate set to verify that they are indeed related to DDOS attacks.

We conducted five or six rounds (depending on languages) of Steps 1 to 3 to fine-tune our keyword lists. These trials allow us to experiment with the cutoff values that separate the DDOS-unrelated and potential DDOS-attack posts (Step 3). They also allow us to inspect why some posts are misclassified, which helps us adjust the weights assigned to the keywords. From these trials, we identify variants of the keywords (e.g., *ddosing* and *ddosed*) that are typically associated with informal use of languages in the forums. We then add these variants to the keyword lists. We also use some DDOS-unrelated keywords to better separate the non-DDOS posts from the DDOS-related posts.

We noticed a trade-off between using more or fewer keywords generated from Step 2. More keywords ensured a higher recall rate. However, they would classify a larger number of posts as potentially DDOS-related and subsequently require more intense manual processing effort in Step 4. Because the English forum, hackforums.net, has a large number of posts, we first selected the 450 highest-ranked keywords from the initial set of 1,267 keywords extracted from 284 articles. We then manually separated these keywords into general and specific DDOS-related terms. Eventually, we used 112 keywords that were more specific to DDOS attacks. As for the Chinese and Russian forums, we used a much larger number of keywords. The initial numbers of keywords are 2,234 and 1,209 from 362 and 138 articles, respectively. The final numbers of keywords for the English, Chinese and Russian forums are 112, 2,241, and 1,109. The articles used to form our DDOS keyword database and keyword lists are available at https://drive.google.com/drive/folders/0BwjvgWAzyIyweXJpMHNZMEJXWFU.

The details of the DDOS post-extraction procedures are as follows.

1. We searched the Internet with keywords such as *ddos* and other DDOS-attack methods or tools. We obtained these keywords from reading some general introductory articles on DDOS, such as Wikipedia. From this Internet search, we collected an initial set of DDOS-attack articles. Upon reading these articles, we added some keywords and performed more searches to obtain more articles.
2. We fed these articles into text-mining software packages (SAS Enterprise Miners for English and Chinese and SynchroTerm for Russian). The software generated a list of keywords based on frequency of appearance in the documents and the number of documents containing the keywords. It automatically grouped keywords with similar meaning together under a unified parent keyword. For example, *attack, attacks* and *attacking* were grouped under *attack.* We manually grouped some other keywords. For example, *arp poisoning* was grouped together with *arp cache poisoning* and *arp poison routing.* We then generated the keyword lists.
3. Based on term frequencies and numbers of documents containing the keywords, we ranked and selected the top-ranked keywords as the basis for extracting DDOS-attack posts. We assigned the weights based on whether the keywords are specific to DDOS attacks. For example, *DDOS* is considered a DDOS-specific term, whereas *server* is considered a common term and hence receives a lower weight although it is also highly ranked by the software. From the trial runs, we identified additional keywords from the forum posts and added those keywords to the list. Some examples of

common keywords used across forums of different languages are: *ddos, flood, teardrop, bot, botnet, spoof, slowloris* and *smurf*.

4. The posts with higher numbers of matched DDOS-specific keywords receive higher scores. Specifically, DDOS-specific keywords are assigned more weight. Hence, each time a DDOS-specific keyword appears in a post, it would contribute to a higher score (compared to the non-DDOS-specific keywords). We then aggregated the scores for each post. Posts receiving low scores are considered to be DDOS unrelated. The remaining DDOS posts are considered to be potentially DDOS-related.
5. The two RAs then screened all candidate DDOS-attack posts. A post has to explicitly mention the DDOS-related terms in order for the RAs to consider it to be DDOS-related. For posts that had inconsistent classification results, we asked the RAs to evaluate each post individually. They then compared the results again and discussed to resolve any remaining inconsistencies.

A2.1 Test-Set Generation

To assess the accuracy of the port-extraction procedure developed above, we randomly generated three test sets. Two RAs with computer-security knowledge and extensive training read all posts in each test set. They first determined whether the post was DDOS related. They then reexamined the posts that were inconsistently classified. Posts that could not be consistently classified after the discussion were treated as DDOS unrelated. The final inconsistency rates for the English, Chinese and Russian forums are all less than 1%.

A2.2 Evaluation of DDOS-Related Post Extraction

We applied the DDOS-extraction procedure to the test sets to evaluate its performance (Table A2.2). The results show that our extraction consistently achieved recall rates exceeding 80%. In most cases, the recall rates exceed 90%.

Table A2.2: Recall Rates for DDOS-Related Posts Extraction on Test Sets

	Posts	Posts classified as candidate DDOS-related posts	DDOS-related posts	DDOS-related posts extracted	Recall rate
Hackforums					
Sample 1	8,415	672	288	256	89%
Sample 2	7,860	673	278	255	92%
Sample 3	8,685	654	206	195	95%
HBH					
Sample 1	3,846	491	10	7	92%
Sample 2	3,635	276	25	19	83%
Sample 3	4,292	293	55	47	85%
Hackbase					
Sample 1	4,431	1,567	23	21	91%
Sample 2	3,810	1,602	8	7	88%
Sample 3	4,716	1,674	5	4	80%
HHLM					
Sample 1	1,479	1,114	51	50	98%
Sample 2	1,456	1,147	68	68	100%
Sample 3	1,437	1,126	52	51	98%
Antichat					
Sample 1	5,822	1,832	17	15	94%
Sample 2	3,597	1,603	62	55	98%
Sample 3	6,983	2,177	13	12	100%
Xaker					
Sample 1	2,738	1,494	76	74	97%
Sample 2	3,710	1,938	78	77	99%
Sample 3	2,565	1,372	63	62	98%

Because our analysis requires the DDOS-related posts to be related to a port, we apply the above procedure to all threads that contain at least one port-related post. Two RAs then independently screened the candidate DDOS-related posts. Table A2.3 presents the composition of the final dataset.

Table A2.3: Final Dataset Composition

	Posts in threads that contain port-related posts	Potential DDOS-related posts	Extracted DDOS-related posts	Threads that contain DDOS-related posts
Hackforums	246,104	23,090	13,410	2,781
HBH	2,230	404	69	26
Hackbase	46,211	13,652	430	339
HHLM	21,255	7,948	1,284	831
Antichat	198,094	31,347	626	198
Xaker	8,854	1,939	124	62

A3. Automatically Classify Content Orientation of Hacker-Forum Posts

We applied Latent Dirichlet Allocation (LDA) analysis on the port-related posts and their follow-ups. As shown in Figure A3, with respect to *Stop-Word Elimination, Stemming* and Lemmatization, we employed Porter Stemmer and WordNetLemmatizer in the NTLK package, a leading platform for building Python programs to work with human-language data. In the LDA-analysis step, we employed Gensim, one of the most commonly used open-source Python topic-modeling packages, used and cited in over 500 commercial and academic applications.

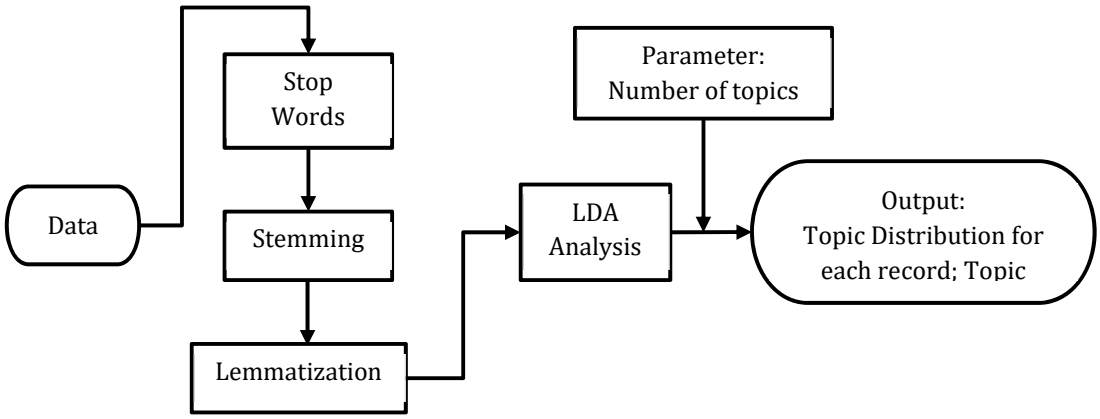

Figure A3. LDA analysis procedures

To generate topic keywords, we applied the approach introduced by Quinn et al. (2010), which takes into account how frequently a word is used in a topic and how distinctive the usage of a word is on a particular topic. The score of a keyword is calculated by

$$r_{kw} = \frac{\beta_{kw} - \underset{j \neq k}{median}(\beta_{jw})}{\underset{l \neq k}{MAD}(\beta_{lw})},$$

where β is the conventional keyword score generated by LDA and *MAD* is the median absolute deviation. Compared with the topic keywords generated directly by LDA, the topic keywords generated by this approach could extract the primary meaning of a post without being overwhelmed by secondary mention of extraneous topics. For example, in our case, the keyword *hack* can appear in any topics, but it is unable to distinguish between different topics. Table A3 reports the top 20 keywords in each topic. The intensity of the shade reflects the score of each keyword. Darker shade represents a higher score.

Table A3. LDA Results (3–10 Topics)

Topic ID	Posts	3-Topic Model									
0	3,777	http	error	file	short	includ	compil	win	messag	warn	platform
		program	found	miss	version	print	open	msg	result	invalid	data
1	48,385	work	nice	great	link	post	tutori	plea	download	tut	thank
		one	man	thread	hack	know	now	use	ing	much	want
2	30,672	port	use	server	ip	open	bot	work	connect	file	run
		know	make	password	host	comput	download	one	set	want	don
Topic ID	**Posts**	**4-Topic Model**									
0	6,861	link	download	plea	file	send	password	version	viru	sourc	upload
		updat	code	plz	bot	backdoor	messag	remov	detect	add	compil
1	40,247	nice	work	great	post	tutori	tut	man		thread	one
		hack	now	know	use	much	share	ing	keep	well	plea
2	2,679	http	error	file	includ	foundhttp	sql	result	warn	program	vulner
		found	inject	admin	open	invalid	php	commandsadd	platform	miss	print
3	33,047	use	port	server	bot	work	ip	open	connect	know	run
		one	want	make	host	comput	program	don	find	set	see
Topic ID	**Posts**	**5-Topic Model**									
0	43,500	nice	work	great	tutori	link	post	plea	tut	one	man
		thank	download	thread	know	ing	hack	now	use	much	want
1	1,836	error	compil	file	irc	channel	join	bot	includ	mod	program
		sourc	public	miss	warn	sticki	code	win	wast	lmao	reali
2	18,622	port	work	open	file	server	bot	download	connect	run	use
		set	forward	problem	make	rat	test	plea	still	metu	router
3	15,542	use	ip	password	server	ddo	site	hack	websit	port	connect
		know	host	want	find	comput	command	proxi	exploit	attack	way
4	3,334	http	click	short	view	spoiler	viewspoil	bomber	foundhttp	cheer	plz
		result	found	link	print	lmfao	admin	color	bitch	rapidshar	msg

Topic ID	Posts	6-Topic Model									
0	2,034	error	bump	short	file	includ	open	program	compil	fred	blackshad
		cheer	messag	warn	awsom	miss	invalid	platform	own	wifi	syntax
1	4,541	link	download	plea	viru	detect	file	send	plz	upload	updat
		backdoor	antiviru	trojan	remov	clean	dead	advanc	hey	infect	give
2	45,291	work	nice	great	tutori	post	one	tut	man	thank	thread
		use	know	now	rat	ing	much	plea	hack	still	bot
3	21,461	port	use	ip	server	connect	open	run	comput	work	know
		hack	site	want	host	websit	ddo	don	program	one	bot
4	7,262	file	click	password	bot	irc	server	view	email	name	log
		spoiler	command	download	send	code	add	chang	run	make	account
5	2,245	http	result	sticki	advanc	forc	tool	vulner	sock	site	read
		web	jar	found	page	print	net	sner	brute	handi	find
Topic ID	**Posts**	**7-Topic Model**									
0	19,854	nice	great	tutori	link	work	tut	man	thank	plea	post
		download	share	bro	thread	much	awesom	job	keep	mate	dude
1	2,460	http	plz	reupload	fred	result	bomark	perl	plea	die	found
		php	bren	read	jar	print	script	site	input	return	steam
2	4,576	file	download	bot	viru	short	crypt	fud	detect	crypter	link
		server	updat	exe	trojan	channel	folder	messag	infect	remov	bind
3	9,525	port	open	connect	ip	server	click	forward	use	router	run
		work	set	problem	test	noip	view	host	spoiler	make	show
4	36,133	work	know	one	use	don	bot	want	someon	make	now
		ing	rat	post	hack	still	anyon	plea	site	thread	even
5	9,045	use	password	hack	ip	file	find	proxi	exploit	window	email
		command	site	server	run	comput	send	access	account	network	web
6	1,241	error	file	compil	program	includ	open	miss	warn	platform	invalid
		commandsadd	iso	fatal	winrar	slowlori	use	windn	syntax	tuto	document

Topic ID	Posts	8-Topic Model									
0	1,949	error	short	program	messag	compil	reupload	miss	fred	warn	prob
		invalid	platform	youseemeorg	own	msg	asap	omni	char	syntax	fix
1	7,798	link	download	plea	file	viru	updat	bot	send	bump	fud
		sourc	upload	version	crypter	infect	detect	plz	backdoor	hey	crypt
2	23,568	thread	hack	know	one	post	want	site	don	make	ing
		someon	use	fuck	take	read	now	click	even	best	say
3	22,986	port	server	work	ip	use	bot	open	connect	run	set
		problem	comput	know	make	host	forward	test	sure	one	want
4	4,872	password	email	exploit	use	command	send	name	log	inform	file
		code	web	system	account	list	attack	tool	provid	servic	secur
5	18,562	nice	work	great	tutori	tut	man	thank	post	share	much
		bro	awesom	use	job	keep	now	mate	dude	one	keylogg
6	2,062	http	packet	php	foundhttp	flood	result	cheer	admin	script	sql
		vulner	site	tcp	inject	print	syn	page	host	read	websit
7	1,037	file	includ	open	error	window	path	perl	access	execut	permiss
		commandsadd	thnk	ettercap	batch	backtrack	item	kiddi	jar	modifi	fatal
Topic ID	**Posts**	**9-Topic Model**									
0	8,102	link	download	plea	send	file	password	account	email	work	bump
		plz	log	messag	someon	add	upload	hey	give	keylogg	advanc
1	23,638	know	one	don	make	post	want	now	work	click	ing
		use	rat	download	see	thread	say	someon	even	take	updat
2	8,774	bot	ip	server	host	use	booter	ddo	connect	irc	metu
		botnet	make	command	ddoser	port	hit	set	want	address	put
3	2,658	http	short	detect	fred	found	trojan	wast	tnx	avg	result
		antiviru	report	umm	bitch	clean	omfg	ftw	avast	color	sned
4	13,339	work	port	open	problem	connect	server	use	run	test	forward
		set	still	router	rat	everi	know	anyon	comput	check	sure
5	17,522	nice	great	tutori	tut	man	thank	work	post	share	thread
		bro	much	dude	awesom	job	keep	wow	mate	love	guid
6	971	foundhttp	win	version	print	die	string	name	public	white	read
		commandsadd	valu	int	msg	TRUE	nick	return	char	config	file
7	6,879	hack	use	site	tool	websit	find	exploit	comput	ip	attack
		learn	http	way	web	network	code	know	want	system	internet
8	951	file	error	open	program	compil	includ	viewspoil	warn	miss	permiss
		invalid	click	platform	path	hahaha	access	syntax	modifi	skiddi	fatal

Topic ID	Posts	10-Topic Model									
0	18,042	nice	great	tutori	tut	man	work	thank	share	post	bro
		awesom	job	much	mate	click	thread	keep	dude	love	wow
1	956	file	includ	public	error	open	reali	win	version	string	commandsadd
		fatal	sourc	conf	class	return	debug	static	final	omni	int
2	26,570	know	one	post	don	work	use	hack	want	thread	ing
		best	make	now	well	site	take	rat	read	even	learn
3	10,836	port	ip	connect	use	open	forward	work	server	router	address
		comput	network	ddo	attack	firewal	noip	internet	problem	block	know
4	12,623	bot	server	file	run	make	download	use	host	work	irc
		set	open	metu	instal	password	command	see	one	now	click
5	2,261	hack	exploit	site	inform	vulner	http	find	web	websit	admin
		sql	script	hacker	inject	thnx	foundhttp	cheer	attack	sticki	perl
6	7,151	link	plea	download	work	send	email	someon	msn	password	keylogg
		hey	test	anyon	updat	plz	give	messag	crypt	add	crypter
7	977	error	program	window	compil	miss	vista	warn	forc	pack	sandbox
		use	brute	umm	fix	invalid	platform	devic	bruteforc	winrar	syntax
8	1,705	http	short	result	php	read	tor	request	found	print	rock
		jar	data	packet	elvi	thnk	die	where	function	coie	url
9	1,713	tool	viru	detect	trojan	clean	net	download	file	report	bomber
		antiviru	sner	infect	advanc	hex	bomark	backdoor	avg	handi	softwar

A4. Extract Posts Related to Specific DDOS-Attack Techniques

We identified forum posts related to two types of hacking techniques. The first type of posts is related to botnets. To carry out large-scale and cost-effective DDOS attacks, attackers often use botnets to remotely control hundreds or thousands of compromised computers around the globe. The second type of posts are related to new hacking techniques that facilitate DDOS attacks. We describe the detailed steps of extracting posts related to these two types of DDOS-attack techniques in sections A4.1 and A4.2.

A4.1 Extract Posts Related to Botnets

We identified posts related to botnets based on the frequency of botnet keywords as measured by term-frequency–inverse-document-frequency (tf-idf). tf-idf is a numerical statistic reflecting the importance of a word in a given document in a collection of documents or corpus. It is often used as a weighting factor in information retrieval, text mining, and user modeling.

The botnet keywords should help distinguish botnet-related and not-botnet-related posts. The data processing proceeds in four steps:

1. Extract seed posts using keywords *bot, zombi* and *botnet.* From the 246,104 posts in the port-related DDOS-attack threads, we obtained 15,718 botnet-related seed posts.
2. Apply tf-idf on every seed post against 10 randomly chosen botnet-unrelated posts from the forum. Doing so generates a list of keywords for each seed post. However, not every keyword is relevant to botnet, as the content of each seed post may be related to other contexts.[2]

[2] For example, the idf of "legit" is quite significant and "legit" is a keyword in the post. "Burimidk is legit I know him to can I trade 700 bots for 700 bots with you."

3. Combine all keywords generated for each seed post and calculate the average tf-idf score for each candidate keyword. Although each seed post might contain keywords irrelevant to botnets, only botnet keywords that are common across all the seed posts would result in high average tf-idf scores. Generally, a keyword is considered a botnet-related keyword if its tf-idf score is above 0.2 and its occurrence in the corpus is greater than 10. The distribution of the tf-idf scores is shown in Figure A4.1.[3]

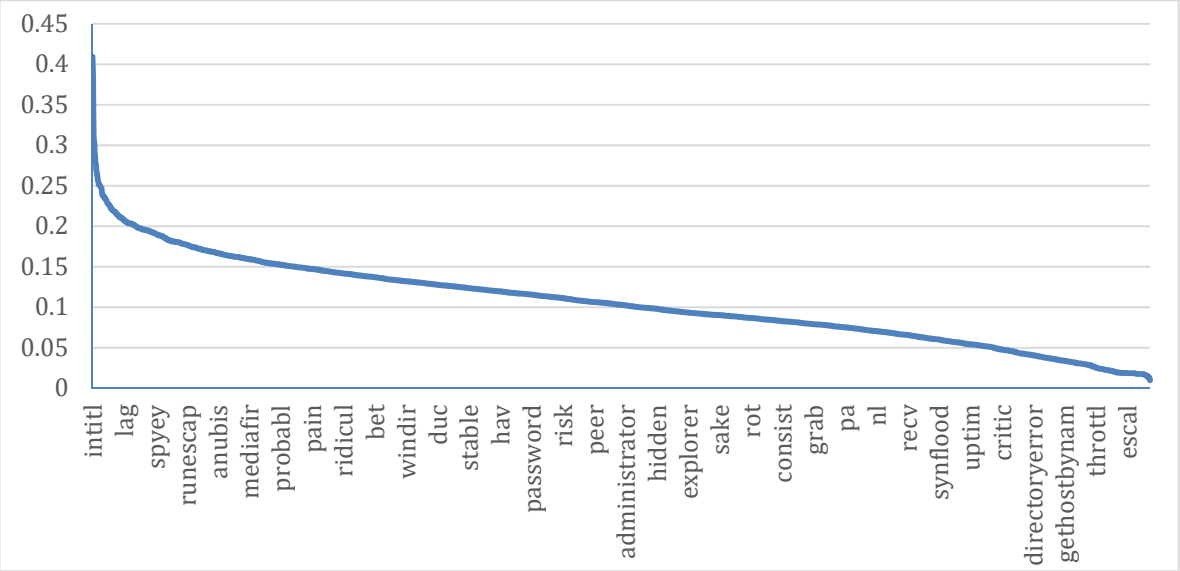

Figure A4.1 tf-idf Score Distribution

4. Label a post as botnet related if it contains one or more keywords from the list of botnet keywords obtained in Step 3. From the 246,104 port-related DDOS-attack posts, we extracted 34,039 botnet-related posts. Among these bot-related posts, 24,065 posts are inside the threads containing DDOS-related posts, and 12,249 are also DDOS-related posts.

[3] In Figure A4.1, Intitl/inurl are keywords that restrict search results from search engines. Hackers often use them to locate the information of interest; verbinden is a commonly used variable name in botnet scripts; Kaiten is a Linux bot; zrxbtcore refers to a remote administrator tool, z3r0xb0t.

A4.2 Extract Posts related to Emergent Hacking Techniques

To identify emerging discussion on hacking techniques, we identified the attack techniques that freshly appeared during our study period. We searched for the corresponding alert information from the National Cyber Awareness System, [4] which documents the incidents of new vulnerabilities and exploits. During our study window (2007–2011), Zeus botnet was first discovered in July 2007 and became more prevalent in March 2009.[5] Mariposa botnet was first discovered in December 2008.[6] We used the related keywords (*Mariposa* and *Butterfly* for Mariposa botnet, and *Zbot* and *Zeus* for Zeus botnet) to extract relevant posts from Hackforums.

From the 246,104 port-related DDOS-attack posts, we extracted 27 posts containing the keywords *mariposa* or *butterfly* and 130 posts containing the keywords *zbot* or *zeus.* Among the 27 posts related to Mariposa botnet, 20 are inside the threads containing DDOS-related posts and 12 are also DDOS-related posts. Among the 130 posts related to Zeus botnet, 86 are within the threads containing DDOS-related posts and 37 are also DDOS-related.

[4] https://www.us-cert.gov/ncas.

[5] https://www.us-cert.gov/ncas.

[6] https://en.wikipedia.org/wiki/Mariposa_botnet.

A5. Keyword Distribution and Manual Classification for Six Hacker Forums

Table A5.1 Manual Classification of the Port- and DDOS-Related Threads Based on Leading Posts

Category	Protection		Malicious attacks		Protection & malicious attacks		Irrelevant		Total leading post	Total reply post
Hackforums	104	(30)	2,458	(42)	105	(52)	114	(35)	2,781	107,902
HBH	7	(13)	10	(20)	3	(8)	6	(21)	26	327
Hackbase	39	(7)	260	(47)	9	(8)	31	(151)	339	10,238
HHLM	575	(9)	239	(19)	145	(16)	32	(17)	991	5,618
Antichat	35	(32)	100	(202)	8	(1,816)	55	(1,100)	198	79,597
Xaker	4	(66)	48	(47)	7	(14)	3	(40)	62	1,888

Note: The table entries are the number of leading posts classified in the respective categories. The parentheses contain the average numbers of replies per leading post during 2007–2011.

Table A5.2 Top 10 Keywords in the Titles of Port- and DDOS-Related Threads

Keywords	Hackforums	HBH	Hackbase	HHLM	Antichat	Xaker
help	1	3	4		38	
bot/botnet/botnets	3	2	24		4	1
hack/hacked/hacker/hacking	5	1	30	4		2
security		15		1		
icq					1	11
3389			1	39		
ddos/ddosed/ddoser	2			17	2	3
question	8		2		8	36
attack	24	5	36	2	5	
set/setting	50			25	3	
protect/protection				3	19	9
how			3	11		
port	4	9	5	9	30	32
network/networking		4	32	10	29	12
zombie						4
computer		8	15	5		
breaking/breaks					12	5
need/needed	6	34				
server	9	20	19	6	9	24
net	31		6			
open	37	6	8			
answer					6	
flooder					16	6
tut	7					
irc	14	7			28	
php		29			7	7
enter			7			
method			48	7		
version						8
system				8		
command			9	15		
ip	10	40	27	45	18	
darkcomet	40					10
school		10				
script					10	
mapping			10			

Table A5.3 Top 10 Keywords in the Leading Posts of Port- and DDOS-Related Threads

Keywords	Hackforums	HBH	Hackbase	HHLM	Antichat	Xaker
port	1	3	3	6	7	6
server	3	13	2	1	2	2
open	8	12	1	42		
file	9		31	15	1	1
connect/connected/connection		1	9	16	42	
bot/botnet/bots	2	36			36	4
style		2				
method			33	2		
run/running	29		7	3		
last					3	
system			18	7	8	3
ip	4	27	5	10	14	14
net	35		4			17
computer	50	6	6	4		
src		4				
edit/edited		23			4	30
http	5	24			6	8
start	33			5		
password	38		23	49	10	5
code	41	48			5	12
class		5				
work/working/works	6				44	9
time	23				11	7
img		7				
find/found	16	8				45
how			8			
use			21	8		
hack/hacked/hacking/hacker	30	9		43		
site	40				9	11
network			27	9	41	28
try/tried/trying	10					
div		10				
pigeon			10			

Table A5.4 Top 10 Keywords in the Reply Posts of Port- and DDOS-Related Threads

Keywords	Hackforums	HBH	Hackbase	HHLM	Antichat	Xaker
awesome/thanks/nice/good/great/better/best/fine	1		1	9		
class		1				
message	115				2	1
quote		18			1	2
use/used/using	10	41		1		
Look at it			2	4		
style		2				
system		86	54	2		
write/wrote	2	17				
file/files	21	58	76	14	7	3
html/http/https	14	3		70	3	14
Learn			4	3		
post owner			3			
work/worked/working/works	3				10	13
alt		4				
code	64				4	23
last						4
tried/try/trying	4					
forum host			5	36		
img		5				
method			44	5		
php	62	150			5	55
server	11	54		115		5
tut/tutorial	5					
computer	44		20	6		18
edited					6	6
help/helped	6					
how/how is it			6	11		
src		6				
align		7				
port/ports	7	32				
rms/rms.exe						7
support			7			
user		33	55	7		45
bot/botnet/bots	8					
com	16	21		72	8	28
div		8				
enter			68	8		
install/installation						8
Selfless			8			
add/added						9
contribution			9			
gif		9				
icq					9	
know	9	29	11	22		
assembling/assembly						10
set up			53	10		
share	113		10	15		
strong		10				

References

Lee, D., Carpenter, B. E., and Brownlee, N. (2010), "Observations of UDP to TCP Ratio and Port Numbers," Fifth International Conference on Internet Monitoring and Protection, Barcelona, pp. 99–104.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010), "How to Analyze Political Attention with Minimal Assumptions and Costs," *American Journal of Political Science* (54:1), pp. 209–228.

Appendix B. Supplementary Robustness Tests

Table B1. Robustness Tests Using Other Measures of Forum Discussion

Variables	(1) Port mention DDOS thread	(2) Port thread DDOS thread	(3) Port mention DDOS effective	(4) Port effective DDOS effective	(5) Port thread DDOS effective	(6) Port mention DDOS mention	(7) Port effective DDOS mention	(8) Port thread DDOS mention
Lagged number of victim IPs	0.579***	0.579***	0.579***	0.579***	0.579***	0.579***	0.579***	0.579***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Number of threats	0.342***	0.334***	0.343***	0.331***	0.334***	0.344***	0.343***	0.343***
	(0.108)	(0.107)	(0.109)	(0.107)	(0.107)	(0.109)	(0.109)	(0.109)
Number of vulnerabilities	0.108***	0.105***	0.108***	0.105***	0.105***	0.108***	0.108***	0.108***
	(0.028)	(0.027)	(0.028)	(0.027)	(0.027)	(0.028)	(0.028)	(0.028)
Forum post measures	−0.043***	−0.023***	−0.036***	−0.033***	−0.025***	−0.038**	−0.021**	−0.019**
	(0.012)	(0.005)	(0.013)	(0.006)	(0.005)	(0.016)	(0.008)	(0.008)
Port fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250
Adjusted R^2	0.979	0.979	0.979	0.979	0.979	0.979	0.979	0.979
Number of port	35,450	35,450	35,450	35,450	35,450	35,450	35,450	35,450

Notes: Column (1): Measure port discussion by counting the specific posts mentioning a port number, and DDOS-attack discussion by counting all posts in a thread containing at least one DDOS-attack post. Column (2): Measure port discussion by counting all posts in a thread containing at least one port post, and DDOS-attack discussion by counting all posts in a thread containing at least one DDOS-attack post. Column (3): Measure port discussion by counting the specific post mentioning a port number, and DDOS-attack discussion by counting all posts mentioning or replying to an earlier post mentioning DDOS attacks. Column (4): Measure port discussion by counting all posts mentioning or replying to an earlier post mentioning a port number, and DDOS-attack discussion by counting all posts mentioning or replying to an earlier post mentioning DDOS attacks. Column (5): Measure port discussion by counting all posts in a thread containing at least one port post, and DDOS-attack discussion by counting all posts mentioning or replying to an earlier post mentioning DDOS attacks. Column (6): Measure port discussion by counting the specific posts mentioning a port number, and DDOS-attack discussion by counting the specific posts mentioning DDOS attacks. Column (7): Measure port discussion by counting all posts mentioning or replying to an earlier post mentioning a port number, and DDOS-attack discussion by counting the specific posts mentioning DDOS attacks. Column (8): Measure port discussion by counting all posts in a thread containing at least one port post, and DDOS-attack discussion by counting the specific posts mentioning DDOS attacks. Robust standard errors clustered by port in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table B2. Robustness Tests with Other Lag Orders

Variables	(1) L1	(2) L2	(3) L7	(4) L15	(5) L30
Lagged number of victim IPs	0.579***	0.579***	0.578***	0.578***	0.575***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Number of threats	0.327***	0.325***	0.322***	0.296***	0.204***
	(0.107)	(0.107)	(0.106)	(0.099)	(0.077)
Number of vulnerabilities	0.104***	0.104***	0.102***	0.101***	0.102***
	(0.027)	(0.027)	(0.026)	(0.026)	(0.026)
Contemporaneous posts	−0.020***	−0.016***	−0.009***	−0.006***	−0.004***
	(0.004)	(0.003)	(0.002)	(0.002)	(0.001)
Posts lagged by 1 day	−0.020***	−0.014***	−0.007***	−0.004***	−0.003**
	(0.004)	(0.003)	(0.002)	(0.001)	(0.001)
Posts lagged by 2 days		−0.014***	−0.004**	−0.002	−0.000
		(0.003)	(0.002)	(0.002)	(0.002)
Posts lagged by 3 days			−0.008***	−0.005***	−0.004***
			(0.002)	(0.002)	(0.002)
Posts lagged by 4 days			−0.004***	−0.002	−0.000
			(0.002)	(0.001)	(0.001)
Posts lagged by 5 days			−0.008***	−0.005***	−0.004***
			(0.002)	(0.002)	(0.001)
Posts lagged by 6 days			−0.004**	−0.000	0.001
			(0.002)	(0.002)	(0.002)
Posts lagged by 7 days			−0.010***	−0.004**	−0.003*
			(0.002)	(0.002)	(0.002)
Posts lagged by 8 days				−0.002	0.000
				(0.002)	(0.001)
Posts lagged by 9 days				−0.002	−0.001
				(0.002)	(0.002)
Posts lagged by 10 days				−0.007***	−0.006***
				(0.002)	(0.001)
Posts lagged by 11 days				−0.002	−0.000
				(0.002)	(0.001)
Posts lagged by 12 days				−0.005***	−0.003**
				(0.001)	(0.001)
Posts lagged by 13 days				−0.003*	−0.000
				(0.002)	(0.001)
Posts lagged by 14 days				−0.007***	−0.004***
				(0.002)	(0.002)
Posts lagged by 15 days				−0.006***	−0.001
				(0.002)	(0.002)
Posts lagged by 16 days					0.001
					(0.002)
Posts lagged by 17 days					−0.005***
					(0.001)
Posts lagged by 18 days					−0.001
					(0.001)
Posts lagged by 19 days					−0.003**
					(0.001)
Posts lagged by 20 days					−0.000
					(0.001)

Variables	(1) L1	(2) L2	(3) L7	(4) L15	(5) L30
Posts lagged by 21 days					−0.002
					(0.002)
Posts lagged by 22 days					−0.003**
					(0.001)
Posts lagged by 23 days					0.001
					(0.002)
Posts lagged by 24 days					−0.005***
					(0.001)
Posts lagged by 25 days					−0.004**
					(0.002)
Posts lagged by 26 days					−0.000
					(0.001)
Posts lagged by 27 days					0.000
					(0.001)
Posts lagged by 28 days					−0.006***
					(0.002)
Posts lagged by 29 days					−0.003*
					(0.002)
Posts lagged by 30 days					−0.005***
					(0.002)
Port fixed effects	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	64,696,250	64,660,800	64,483,550	64,199,950	63,668,200
Adjusted R^2	0.979	0.979	0.979	0.979	0.979
Number of port	35,450	35,450	35,450	35,450	35,450

Notes: Column (1): Include contemporaneous discussion and discussion lagged by one day. Column (2): Include contemporaneous discussion and discussion lagged by two days. Column (3): Include contemporaneous discussion and discussion lagged by seven days. Column (4): Include contemporaneous discussion and discussion lagged by fifteen days. Column (5): Include contemporaneous discussion and discussion lagged by thirty days. Robust standard errors clustered by port in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table B3. Other Robustness Tests

Variables	(1) Omit threat and vulnerability	(2) No lagged attacks	(3) Add port–month effects	(4) Omit imputed data	(5) Omit port 0	(6) Exclude most-attacked ports	(7) After Hackforums was launched	(8) Only Hackforums ports
Lagged number of victim IPs	0.579***		0.557***	0.517***	0.579***	0.579***	0.567***	0.531***
	(0.002)		(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)
Number of threats		0.568***	0.316***	0.328***	0.333***	0.302***	0.159**	0.364***
		(0.130)	(0.108)	(0.107)	(0.107)	(0.114)	(0.077)	(0.122)
Number of vulnerabilities		0.216***	0.100***	0.086***	0.103***	0.096***	0.109***	0.113***
		(0.046)	(0.028)	(0.028)	(0.027)	(0.023)	(0.028)	(0.030)
Number of DDOS-attack posts	−0.032***	−0.077***	−0.035***	−0.028***	−0.029***	−0.031***	−0.024***	−0.031***
	(0.006)	(0.014)	(0.007)	(0.007)	(0.005)	(0.006)	(0.006)	(0.007)
Port fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Port–month fixed effects	No	No	Yes	No	No	No	No	No
Observations	64,696,250	64,696,250	64,696,250	54,286,535	64,694,425	64,687,125	60,442,250	52,669,500
Adjusted R^2	0.979	0.968	0.979	0.950	0.979	0.979	0.979	0.952
Number of port	35,450	35,450	35,450	30,082	35,449	35,445	35,450	28,860

Notes: Column (1): Omit the number of threats and number of vulnerabilities. Column (2): Omit the lagged number of victim IPs. Column (3): Add port–month fixed effects. Column (4): Omit all observations with zero DDOS attacks. Column (5): Omit port 0. Column (6): Omit the five most-attacked ports, 6881, 80, 53, 4672, and 137. Column (7): Trim all data before May 2, 2007, the day when Hackforums was officially launched. Column (8): Omit all ports not mentioned in Hackforums. Robust standard errors clustered by port in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table B4. Regression Results with LDA Topics

Variables	(1) 3-topic model	(2) 4-topic model	(3) 5-topic model	(4) 6-topic model	(5) 7-topic model	(6) 8-topic model	(7) 9-topic model	(8) 10-topic model
Lagged number of victim IPs	0.579***	0.579***	0.579***	0.579***	0.579***	0.579***	0.579***	0.579***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Number of threats	0.333***	0.332***	0.333***	0.332***	0.328***	0.329***	0.329***	0.329***
	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)
Number of vulnerabilities	0.106***	0.105***	0.106***	0.106***	0.103***	0.104***	0.103***	0.103***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Topic 1	−0.013***	0.009***	−0.006***	−0.008	−0.017***	−0.010**	0.000	−0.017***
	(0.003)	(0.003)	(0.001)	(0.005)	(0.003)	(0.005)	(0.002)	(0.003)
Topic 2	−0.005***	−0.008***	−0.000	0.008***	−0.011***	0.008***	0.003***	−0.016**
	(0.001)	(0.002)	(0.005)	(0.003)	(0.003)	(0.002)	(0.001)	(0.006)
Topic 3	−0.000	−0.013***	−0.000	−0.008***	0.004	−0.000	0.005**	0.000
	(0.001)	(0.003)	(0.001)	(0.002)	(0.003)	(0.001)	(0.002)	(0.001)
Topic 4		0.000	0.002	−0.000	−0.002*	0.000	−0.010***	−0.003**
		(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)
Topic 5			−0.013***	0.009**	0.003***	−0.001	−0.002	0.006***
			(0.003)	(0.004)	(0.001)	(0.003)	(0.001)	(0.002)
Topic 6				−0.014***	−0.005**	−0.018***	−0.018***	−0.020***
				(0.003)	(0.002)	(0.003)	(0.003)	(0.004)
Topic 7					−0.010	−0.013***	−0.015**	−0.000
					(0.006)	(0.003)	(0.006)	(0.003)
Topic 8						−0.001	−0.008***	−0.014**
						(0.005)	(0.002)	(0.006)
Topic 9							−0.011	−0.013***
							(0.008)	(0.004)
Topic 10								0.007*
								(0.004)
Port fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250
Adjusted R^2	0.979	0.979	0.979	0.979	0.979	0.979	0.979	0.979
Number of port	35,450	35,450	35,450	35,450	35,450	35,450	35,450	35,450

Notes: Column (1): Three LDA topics. Column (2): Four LDA topics. Column (3): Five LDA topics. Column (4): Six LDA topics. Column (5): Seven LDA topics. Column (6): Eight LDA topics. Column (7): Nine LDA topics. Column (8): Ten LDA topics. Robust standard errors clustered by port in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Table B5. Regression Results with Discussion from Other Forums

Variables	(1) Hackforums	(2) HBH	(3) Hackbase	(4) HHLM	(5) Antichat	(6) Xaker
Lagged number of victim IPs	0.579***	0.579***	0.579***	0.579***	0.579***	0.579***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Number of threats	0.331***	0.343***	0.345***	0.345***	0.343***	0.345***
	(0.107)	(0.108)	(0.109)	(0.109)	(0.109)	(0.109)
Number of vulnerabilities	0.104***	0.108***	0.108***	0.108***	0.108***	0.108***
	(0.027)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
Hackforums posts	−0.032***					
	(0.006)					
HBH posts		0.041				
		(0.047)				
Hackbase posts			0.006			
			(0.011)			
HHLM posts				0.002		
				(0.003)		
Antichat posts					−0.006***	
					(0.001)	
Xaker posts						−0.030
						(0.025)
Port fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250	64,696,250
Adjusted R^2	0.979	0.979	0.979	0.979	0.979	0.979
Number of port	35,450	35,450	35,450	35,450	35,450	35,450

Notes: Column (1): Include only port-related DDOS-attack posts from Hackforums. Column (2): Include only port-related DDOS-attack posts from HBH. Column (3): Include only port-related DDOS-attack posts from Hackbase. Column (4): Include only port-related DDOS-attack posts from HHLM. Column (5): Include only port-related DDOS-attack posts from Antichat. Column (6): Include only port-related DDOS-attack posts from Xaker. Robust standard errors clustered by port in parentheses. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.