Emerging Threats in Internet of Things Voice Services

Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey | University of Illinois at Urbana–Champaign

In this study, we conduct an empirical analysis of interpretation errors made by Amazon Alexa, the speech-recognition engine that powers the Amazon Echo family of devices. We show how common misinterpretations made by Alexa can be used to build a new class of attacks, called *skill squatting attacks*, and discuss its security implications.

Smart speakers, such as the Amazon Echo and Google Home, have become staple Internet of Things devices in the modern home. In fact, analysts estimated that more than 75 million smart speakers will have been sold by the end of 2018. These devices eschew traditional computing inputs, such as a keyboard, mouse, or touch—instead, they rely entirely on the human voice and speech-recognition systems as their primary control interface.

Unfortunately, the voice services powering these devices are not perfect, and they often make mistakes when interpreting speech. Errors and misuse of smart speakers are already being reported in the wild. For families tuning into a news broadcast in San Diego last year, coverage of a girl using an Amazon Echo to buy a dollhouse caused their own homes' smart speakers to buy dollhouses. Elsewhere, in Oregon, someone's Amazon Echo surreptitiously recorded his or her conversation and then sent it to a random contact as a voice message. These incidents join a growing body of anecdotal evidence that users are subject to frequent misinterpretation of their voice in everyday use. In spite of this, we are unaware of any independent, public effort to study these speech recognition errors in more depth.

Our Experiments: An Overview

In this study,[5] we investigate speech recognition misinterpretations in Amazon Alexa—quantifying how often and why they occur. We chose Alexa because it currently holds the largest share of the smart speaker market.[4] We then leveraged our understanding of misinterpretations to build a new class of attacks on speech recognition systems, called *skill squatting attacks*. Alexa skills are analogous to any other application, except that they run on the Amazon Alexa platform. Skill squatting attacks leverage common misinterpretations to surreptitiously route users to malicious Alexa skills without their knowledge. Finally, we demonstrate how skill squatting attacks can be used to launch complex phishing attacks on users.

What Is an Alexa Skill?

Amazon Alexa is the voice service system that powers the Amazon Echo family of devices. To add

Digital Object Identifier 10.1109/MSEC.2019.2910013
Date of publication: 21 May 2019

extensibility to the Alexa platform, Amazon allows the development of third-party applications, called *skills*, that leverage Alexa voice services. Many companies are actively developing Alexa skills to provide easy access to their services through voice. For example, users can now request rides through the Lyft skill and conduct everyday banking tasks with the American Express skill.

Users interact with skills directly through their voice. Figure 1 illustrates a typical interaction. The user first invokes the skill by saying the skill name or its associated invocation phrase ①. The user's request is then routed through Alexa cloud servers ②, which determine where to forward it based on the user input ③. The invoked skill then replies with the desired output ④, which is finally routed from Alexa back to the user ⑤. Up until April 2017, Alexa required users to enable a skill to their account, in a manner similar to downloading a mobile application onto a personal device. However, Alexa now offers the ability to interact with third-party skills without first installing them.[6]

Measuring Misinterpretations

To measure misinterpretations in speech recognition, we first need a way to identify when Alexa misinterprets a speech sample. To this end, we built a test harness that sends audio through to Amazon Alexa and receives a transcription of the audio content. The test harness takes in audio files as the input, sends them through the Alexa cloud, receives transcriptions of the audio files, and stores them for further analysis. Queries to Amazon Alexa are limited to 400/min to avoid overloading Amazon's production servers.

To study specific misinterpretations and their causes, we rely on an externally collected speech corpus called the *Nationwide Speech Project* (*NSP*). The NSP is an effort led by The Ohio State University, Columbus, to provide structured speech data from a range of speakers across the United States.[3] The NSP corpus provides speech from a total of 60 speakers from six geographical dialect regions.

In particular, five male and five female speakers from each region provided a set of 188 single-word recordings, 76 of which were single-syllable words (e.g., "mice," "dome," and "bait") and 112 of which were multisyllable words (e.g., "alfalfa" and "nectarine"). These single-word files provided a total of 11,460 speech samples for further analysis and served as our primary source of speech data. We queried each audio sample 50 times to Alexa to ensure that our results were consistent. For all queries, Alexa did not return a response on 681 (0.1%), which we excluded from our analysis. We collected this data set of 572,319 Alexa transcriptions on 14 January 2018 over a period of 24 h.

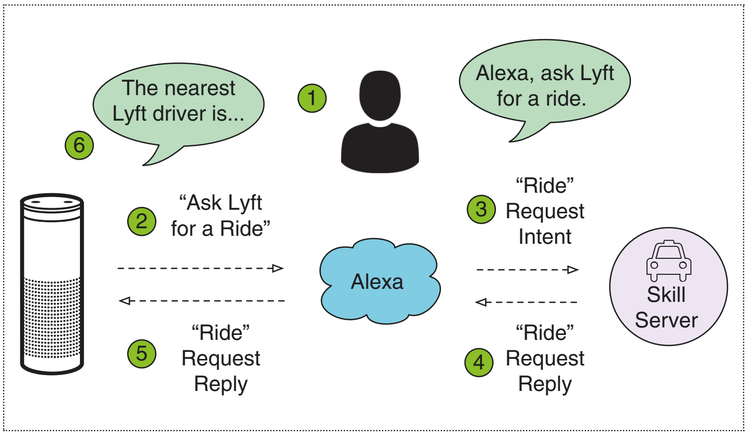


Figure 1. The user–skill interaction in Alexa. A typical user interaction with an Alexa skill, using an Echo device. In this example, a user interacts with the Lyft skill to request a ride.

Ethical Considerations

Although we used speech samples collected from human subjects, we never interacted with subjects during the course of this research. We used public data sets and ensured that our usage was in line with their provider's terms of service. All requests to Alexa were throttled so to not affect the availability of production services. For all attacks presented in this article, we tested them only in a controlled, developer environment. Finally, we did not attempt to publish a malicious skill to the public skill store. We disclosed these attacks to Amazon and worked with them through the standard disclosure process to ensure that these problems are known.

How Accurate Is Alexa?

We started our analysis by investigating how well Alexa transcribed words in our data set. Consistent with anecdotal evidence, Alexa only correctly interpreted 394,715 (68.9%) of the 572,319 audio samples.

In investigating where Alexa makes interpretation errors, we found that accuracy rates varied across words. Figure 2 shows the interpretation accuracy by individual words in our data set. Only three words (2%) were always interpreted correctly. In contrast, 9% of words were always interpreted incorrectly, indicating that Alexa is poor at correctly interpreting some classes of words. Words with the lowest accuracy tended to be small, single-syllable words, such as "bean," "calm," and "coal." Words with the highest accuracy were mixed. Many of the top words contained two or three syllables, such as "forecast" and "robin." In one counterexample, the word "good" was interpreted correctly 99.9% of the time.

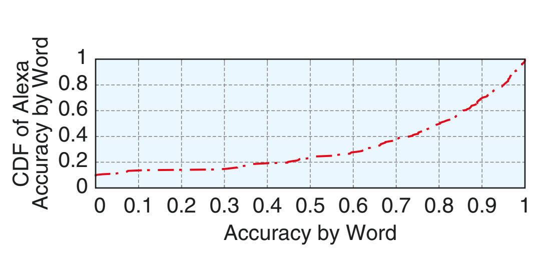


Figure 2. The accuracy of Alexa interpretations by word is shown as a cumulative distribution function (CDF); 9% of the words in our data set were never interpreted correctly, and 2% were always interpreted correctly. This shows substantial variance in misinterpretation rate among words.

Classifying Voice Service Errors

Alexa's speech recognition service is nondeterministic—even when playing back the exact same audio file, the distribution ways in which a word is misinterpreted vary greatly. In investigating the distributions of misinterpretations per word, we observed that, for each of the 188 words, there were one or two interpretations that Alexa outputted more frequently than the others. We called this interpretation the *most common error* (*MCE*) for a given word.

For example, take the word "boil," which was misinterpreted by Alexa 100% of the time. The MCE of "boil" was the word "boyle," which accounted for 94.3% (the MCE rate) of the errors. In this sense, the rate at which the MCE occurred served as a measure of how random the distribution of misinterpretations was. Because "boyle" accounted for the majority of its interpretation errors, we can argue that "boil" has a predictable misinterpretation distribution.

To visualize the rate and randomness of interpretation errors per word, we plotted the error rate for each word along with its MCE rate (Figure 3). This graphical representation provided us with a clearer picture of interpretation errors in Alexa. We then split this plot into three sections: sections 1 (upper left), 2 (upper right), and 3 (bottom half).

The majority (77.7%) of words in our data set fell into section 3. These words were interpreted correctly by Alexa most of the time. There were 9.6% of the words in our data set that appeared in section 2, meaning they were misinterpreted often but did not feature a prevalent MCE. These were likely to be words that Alexa was poor at understanding altogether. As an example, the word "unadvised," which has 147 unique misinterpretations, appeared in this section. The final class of words, in section 1, were those that were misinterpreted more than 50% of the time and had an MCE that appeared in more than 50% of the errors. These were words that Alexa misunderstood both frequently and in a consistent manner. There were 24 (12.8%) such words in our data set.

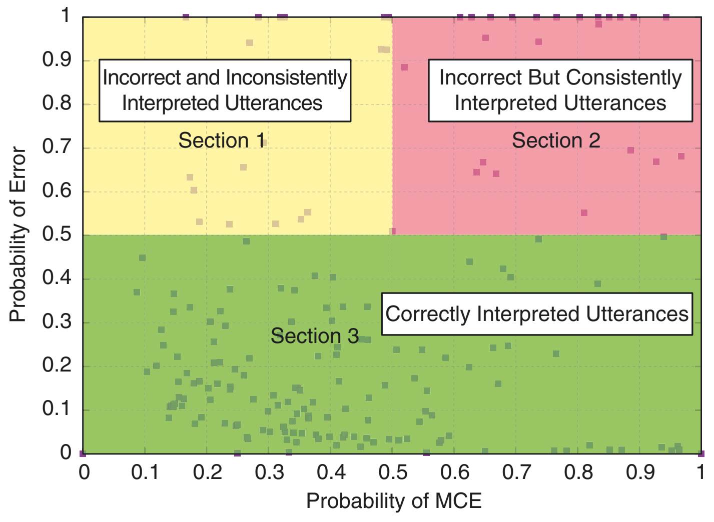


Figure 3. The error rate versus MCE. We plot the error rate by the rate of the MCE for all of the words of our data set. Points in section 1 (upper left) represent words that are misinterpreted both frequently and consistently. In our data set of 188 words, 24 (12.8%) fall in section 1.

Explaining Voice Service Errors

We now have a classification for interpretation errors from our data set. Moreover, we identified 24 words for which Alexa consistently outputs one wrong interpretation. We next investigate why these systematic errors occur.

Homophones

Unsurprisingly, eight (33.3%) of these errors, including "sail" to "sale," "calm" to "com," and "sell" to "cell" were attributable to the fact that these words are homophones, meaning that they have the same pronunciation but different spellings. Of these, five are cases where Alexa returns a proper noun (of a person, state, band, or company) that is a homophone with the spoken word, for example, "main" to "Maine," "boil" to "Boyle," and "outshine" to "Outshyne."

Compound Words

Two (8.3%) other systematic errors occurred due to compound words. Alexa appeared to break these into their constituent words, rather than return the continuous compound word. For example, "superhighway" was split into "super highway" and "outdoors" was split into "out doors."

Phonetic Confusion

Ten (41.7%) of the systematic errors could be explained by examining the underlying phonetic structures of the input words and their errors: in each case, the MCE differed from the spoken word by just a single phoneme. For example, the MCE for the word "wet" was the word "what." The phonetic spelling of "wet" is W EH T, whereas the phonetic spelling of "what" is W AH T. The errors show that Alexa often misunderstands certain specific phonemes within words while correctly interpreting the rest of them. A full list of the phonetic structures for these cases is shown in Table 1.

Other Errors

We could not easily explain three (12.5%) of the errors: "mill" to "no," "full" to "four," and "earthy" to "Fi." Even in listening to each speech sample individually, we found no auditory reason why this interpretation error occurred. One surprising error ("preferably" to "preferrably") occurred because Alexa returned a common misspelling of the intended word. This may be caused by a bug in the Alexa system itself.

Table 1. Phonetic structure of systematic errors.

Word	MCE	Word phonemes	MCE phonemes
rip	rap	R IH P	R AE P
lung	lang	L AH NG	L AE NG
wet	what	W EH T	W AH T
dime	time	D AY M	T AY M
bean	been	B IY N	B IH N
dull	doll	D AH L	D AA L
coal	call	K OW L	K AO L
luck	lock	L AH K	L AA K
loud	louder	L AW D	L AW D ER
sweeten	Sweden	S W IY T AH N	S W IY D AH N

NOTE: We show the underlying phonetic structure of the 10 systematic errors that seem to appear due to Alexa confusing certain phonemes with others. In each case, the resultant MCE is at an edit distance of just one phoneme from the intended word.

Squatting on Alexa Skill Names

After finding that, for some words, Alexa made frequently occurring predictable errors, we leveraged these errors to build a new class of attacks, called *skill squatting*, which exploited predictable errors to surreptitiously route users to a malicious Alexa skill.

The core idea is simple—given a systematic error from one word to another, an adversary constructs a malicious skill that has a high likelihood of confusion with a target skill on the Alexa skills store. When a user tries to access a desired skill using his or her voice, he or she is instead routed to the malicious skill because of a systematic error in the interpretation of the input.

As an example, consider the real Alexa skill "cat facts." The skill is very simple to use—a user invokes the skill by saying "Alexa, tell me some cat facts" to his or her Echo device, and the skill returns some facts about cats. Unfortunately, in this example, Alexa commonly misinterprets "facts" as the word "fax." If attackers know this, they can publish a new skill onto the Alexa store called "cat fax," instead of "cat facts." Now, when the user says "Alexa, tell me some cat facts," he or she is instead routed to the malicious "cat fax" skill instead of his or her desired skill.

It may not surprise you that such an attack vector exists. After all, there are plenty of other attacks that are similar in flavor—for example, domain name typosquatting. In typosquatting, an attacker predicts a common typo in a domain name and abuses it to hijack a request.[8–11] However, typosquatting relies on the user to make a mistake while typing a domain; in contrast, skill squatting is intrinsic to the speech-recognition service itself.

Targeting Existing Alexa Skills

We next investigate how an adversary can craft maliciously named skills targeting existing skills in the Alexa skills store. To start, we collected all of the skills available on the Alexa store as of 27 December 2017, amounting to 23,238 unique skill names. Then, we split each skill name into its individual words. If a word in a skill existed in our spoken data set of 188 words, we checked whether that word was squattable. If so, we exchanged that word with its MCE to create a new skill name. As an example, the word "calm" was systematically misinterpreted as "com" in our data set. Therefore, a skill with the word "calm" can be squatted by using the word "com" in its place (e.g., "quick com" squats the existing Alexa skill "quick calm").

Using the 24 squattable words that we identified previously, we found that we could target 31 skill names that currently exist on the Alexa Store. Only 11 (45.8%) of the squattable words appear in Alexa skill names. Table 2 shows one example of a squattable skill for each of these 11 words. We note that the number of squattable skills we identify is primarily limited by the size of our data set and that it is not a ceiling for the pervasiveness of this vulnerability in the Amazon market.

Predicting Squattable Words

An adversary who attempts this attack using the techniques described thus far would be severely restricted

Table 2. Squattable skills in the Alexa skills store.

Skill	Squatted skill
Boil an egg	Boyle an egg
Main site workout	Maine site workout
Quick calm	Quick com
Bean stock	Been stock
Test your luck	Test your lock
Comic Con dates	Comic khan dates
Mill Valley guide	No valley guide
Full moon	Four moon
Way loud	Way louder
Upstate outdoors	Upstate out
Rip Ride Rockit	Rap ride rocket

NOTE: We show 11 examples of squattable skills publicly available in the Alexa skill store as well as squatted skill names an attacker could use to squat them.

Table 3. Squatted skills in the Alexa skills store.

Skill A	Skill B
Cat fats	Cat facts
Pie number facts	Pi number facts
Cat facts	Cat fax
Magic hate ball	Magic eight ball
Flite facts	Flight facts
Smart homy	Smart home
Phish geek	Fish geek
Snek helper	Snake helper

NOTE: We show examples of squatted skills in the Alexa skills store that drew our attention during manual analysis. Notably, a customer review of the phish geek skill noted they were unable to use the application due to common confusion with the fish geek skill.

by the size and diversity of his or her speech corpus. Without many recordings of a target word from a variety of speakers, he or she would be unable to reliably identify systematic misinterpretations of that word. Considering that many popular skill names make use of novel words (e.g., WeMo) or words that appear less frequently in discourse (e.g., Uber), acquiring such a speech corpus may prove prohibitively costly and, in some cases, infeasible.

We previously observed that, in some cases, phoneme errors could explain why Alexa made a systematic misinterpretation. Motivated by this observation, we now consider how an attacker can amplify the value of his or her speech corpus by reasoning about Alexa misinterpretations at the phonetic level. To better understand this, consider the misinterpretation of the word "luck." "Luck" (L AH K) is frequently misinterpreted as "lock" (L AA K), suggesting that Alexa experiences confusion specifically between the phonemes AH and AA. As such, an attacker might predict confusion in other words with the AH phoneme (e.g., "duck" to "dock," "cluck" to "clock") without having directly observed those words in their speech corpus.

Given this intuition, we used our seed set of word misinterpretations to build a phoneme model of misinterpretations. The output of this model was a mapping from input phoneme to potential output phonemes. We then applied this model to identify already existing instances of confused skills in the Alexa skills store. In total, we found 381 unique skill pairs that exhibited phoneme confusion. The largest single contributor was the word "fact," which was commonly misinterpreted as "facts" and "fax." Given the large number of fact-related skills available on the skill store, it is unsurprising that many of these exist in the wild.

To determine whether these similarities were due to chance, we investigated each pair individually on the skill store. We found eight examples of squatted skills that we marked as worth investigating more closely (Table 3). We cannot speak to the intention of the skill creators. However, we found it interesting that such examples currently exist on the store. For example, "cat facts" has a corresponding squatted skill, "cat fax," which seemingly performs the same function, though published by a different developer. In another example, "Phish Geek,"[2] which purports to give facts about the American rock band Phish, is squatted by "Fish Geek,"[1] which gives facts about fish. Anecdotally, one user of "Phish Geek" appeared to have experienced squatting, writing in a review: "I would love it if this actually gave facts about the band. But instead, it tells you things like 'Some fish have fangs!' "

Ultimately, we have no clear evidence that any of these skills of interest were squatted intentionally. However, this does provide interesting insight into some examples of what an attacker may do and further validates our assertion that our phoneme-based approach can prove useful in finding such examples in use today. Beyond this, we found that of the 23,238 unique skills in the Alexa skill store, 16,836 (72.5%) could potentially be squatted using our phoneme model. Unfortunately, without additional speech samples, there is no way for us to validate the potential attacks.

Case Study: Skill-Based Phishing Attack

In some cases, the skill squatting attack can be used to create complex phishing attacks. In this section, we show an example phishing attack on the American Express skill. The American Express skill allows Amex users to use their Echo device to perform standard banking tasks, like checking account balances and initiating money transfers.

Recall that the typical workflow for interacting with an Alexa skill is for a user to send his or her Echo device some command, for example, "Alexa, ask Amex to make a bank transfer of 30 dollars." However, if a user is unauthenticated with the Amex service at the time of the request, the Echo device will prompt the user to sign in to his or her account on their mobile device. An example prompt is shown in Figure 4(a).

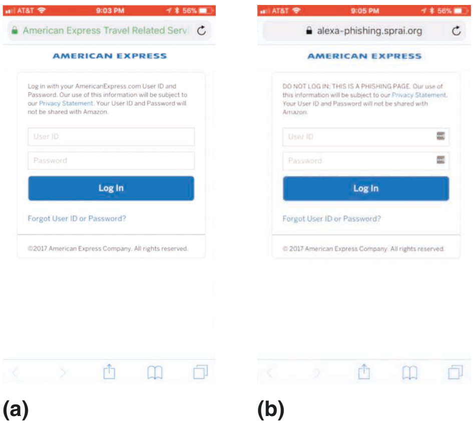
In our testing, we observed that Alexa makes specific systematic errors with phonemes that sound like letters—for example, in our data set, the word "accelerate" was often interpreted by Alexa as "x.celerate." Motivated by this observation, we uncovered that if a skill existed on the skill store called "Am X," Alexa would prefer it over the valid "Amex" application in every case.

As a result, an attacker could publish a skill called "Am X" and have it replicate the flow of authentication for the real American Express skill, displaying a prompt to the user that looks nearly identical to that of the correct Amex login page [Figure 4(b)]. If a user were to enter his or her credentials into this login page, he or she would be unwittingly giving away his or her login credentials to an adversary. As such, skill squatting attacks have a more dangerous potential beyond simply confusing "phish" and "fish"—they could provide a new avenue by which adversaries can phish users.

Limitations

A core limitation of our analysis is the scope and scale of the data set we used in our analysis. The NSP data set only provided 188 words from 60 speakers, which was inadequate for measuring the full scale of systematic misinterpretations of Amazon Alexa. Although our phoneme model extends our observed misinterpretation results to new words, it was also confined by just the errors that appeared from querying the NSP data set.

Another limitation of our work is that we relied on the key assumption that triggering skills in a development environment worked similarly to triggering publicly available skills. However, we did not attempt to publish skills or attack existing skills on the Alexa skills store due to ethical concerns. A comprehensive validation of our attack would require that we work with Amazon to test the skill squatting technique safely in their public production environment.



Log in with your AmericanExpress.com User ID and Password. Our use of this information will be subject to our Privacy Statement. Your User ID and Password will not be shared with Amazon.

©2017 American Express Company. All rights reserved.

DO NOT LOG IN; THIS IS A PHISHING PAGE. Our use of this information will be subject to our Privacy Statement. Your User ID and Password will not be shared with Amazon.

© 2017 American Express Company. All rights reserved.

Figure 4. The American Express login prompts. After launching a skill squatting attack, an adversary can launch a phishing attack by constructing a card that looks nearly identical to the valid login page. (a) The American Express standard prompt and (b) the American Express phishing prompt.

Countermeasures

The skill squatting attack relies on an attacker registering squatted skills. All skills must go through a certification process before they are published. To prevent skill squatting, Amazon could add to the certification process both a word-based and a phoneme-based analysis of a new skill's invocation name to determine whether it may be confused with skills that are already registered. This idea is further explored in an article by Zhang et al. appearing at the IEEE Security and Privacy Symposium in 2019.[12]

As a similar example, domain name registrars commonly restrict the registration of homographs—domains that look very similar visually—of well-known domains.[7] These checks seem not to be currently in place on Alexa because we found 381 pairs of skills with different names, but likely to be squatted on the store.

Short of pronunciation-based attacks, there already exist public skills with identical invocation names on the Alexa skills store. For example, there are currently more than 30 unique skills called "cat facts," and the way in which Amazon routes requests in these cases is unclear. Although this is a benign example, it demonstrates that some best practices from other third-party app store environments have not made their way to Alexa yet.

Our study suggests that systematic errors in Amazon Alexa give rise to a new class of attack with potentially dangerous security implications.

We showed how an attacker can leverage systematic errors to surreptitiously trigger malicious applications for users in the Alexa ecosystem. Further, we demonstrated how this attack could be extended to launch complex phishing attacks. We hope our results inform the security community about the implications of interpretation errors in speech-recognition systems and ultimately provide the groundwork for future work in this area. ■

Acknowledgments

This work was supported in part by the National Science Foundation under contracts CNS 1750024, CNS 1657534, and CNS 1518741. This work was additionally supported by the U.S. Department of Homeland Security contract HSHQDC-17-J-00170. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or the sponsors.

References

1. Amazon, "Fish geek," Accessed on: 2018. [Online]. Available: https://www.amazon.com/Matt-Mitchell-Fish-Geek/dp/B01LMN5RGU/
2. Amazon, "Phish geek," Accessed on: 2018. [Online]. Available: https://www.amazon.com/EP-Phish-Geek/dp/B01DQG4F0A
3. C. G. Clopper, "Linguistic experience and the perceptual classification of dialect variation," Ph.D. dissertation, Dept. Linguistics, Indiana Univ., Bloomington, 2004.
4. B. Kinsella, "56 million smart speaker sales in 2018 says Canalys," voicebot.at, Accessed on: 2018. [Online]. Available: https://www.voicebot.ai/2018/01/07/56-million-smart-speaker-sales-2018-says-canalys/
5. D. Kumar et al., "Skill squatting attacks on Amazon Alexa," in *Proc. 27th USENIX Security Symp.*, 2018 , pp. 33–47.
6. T. Martin, "You can now use any Alexa skill without enabling it first," cnet.com, Accessed on: 2017. [Online]. Available: https://www.cnet.com/how-to/amazon-echo-you-can-now-use-any-alexa-skill-without-enabling-it-first/
7. Namecheap, "Do you support IDN domains and emoticons?" Accessed on: 2018. [Online]. Available: https://www.namecheap.com/support/knowledgebase/article.aspx/238/35/do-you-support-idn-domains-and-emoticons
8. N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen, "Soundsquatting: Uncovering the use of homophones in domain squatting," in *Proc. Int. Conf. Information Security*, 2014, pp. 291–308.
9. J. Spaulding, S. Upadhyaya, and A. Mohaisen, "The landscape of domain name typosquatting: Techniques and countermeasures," in *Proc. IEEE 11th Int. Conf. Availability, Reliability and Security (ARES)*, 2016, pp. 284–289.
10. J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich. "The long 'taile' of typosquatting domain names," in *Proc. 23rd USENIX Security Symp. (USENIX)*, 2014, pp. 191–206.
11. R. Tahir et al., "It's all in the name: Why some URLs are more vulnerable to typosquatting," in *Proc. IEEE Int. Conf. Computer Communications*, 2018, pp. 2618–2626.
12. N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian. "Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems," in *Proc. 40th IEEE Symp. Security and Privacy*, 2019.

Deepak Kumar is a Ph.D. student in computer science at the University of Illinois at Urbana–Champaign. His research interests include computer security and Internet measurement. Contact him at dkumar11@illinois.edu.

Riccardo Paccagnella is an M.S. student in computer science at the University of Illinois at Urbana–Champaign. His research interests include systems security research. Contact him at rp8@illinois.edu.

Paul Murley is a Ph.D. student in computer science at the University of Illinois at Urbana–Champaign. His research interests include web security and measurement. Contact him at pmurley2@illinois.edu.

Eric Hennenfent is a software engineer at Trail of Bits. Hennenfent received a B.S. in electrical and computer engineering from the University of Illinois at Urbana–Champaign. His research focuses on low-level systems security. Contact him at hennenf2@illinois.edu.

Joshua Mason is a research scientist at the University of Illinois at Urbana–Champaign. He received a Ph.D. in 2009 from Johns Hopkins University. His research interests include low-level systems security and reverse engineering. Contact him at joshm@illinois.edu.

Adam Bates is an assistant professor at the University of Illinois at Urbana–Champaign. He received a Ph.D. in 2016 from the University of Florida. His research interests include systems security and data provenance. Contact him at batesa@illinois.edu.

Michael Bailey is an associate professor at the University of Illinois at Urbana–Champaign. He received a Ph.D in 2006 from the University of Michigan. His research interests include the security, performance, and availability properties of computing systems. Contact him at mdbailey@illinois.edu.