JOURNAL OF CYBERSECURITY

*Journal of Cybersecurity*, 0(0), 2015, 1–24
doi: 10.1093/cybsec/tyv008
Research Article

Research Article

# Identifying patterns in informal sources of security information

**Emilee Rader[1] and Rick Wash[2,*]**

[1]Department of Media and Information, Michigan State University, East Lansing, MI, USA and [2]School of Journalism and Department of Media and Information, Michigan State University, East Lansing, MI, USA

*Corresponding author: 404 Wilson Rd #305, East Lansing, MI 48824, USA. Tel: 5173552381; E-mail: wash@msu.edu

Received 31 May 2015; revised 18 September 2015; accepted 29 September 2015

## Abstract

Computer users have access to computer security information from many different sources, but few people receive explicit computer security training. Despite this lack of formal education, users regularly make many important security decisions, such as "Should I click on this potentially shady link?" or "Should I enter my password into this form?" For these decisions, much knowledge comes from incidental and informal learning. To better understand differences in the security-related information available to users for such learning, we compared three informal sources of computer security information: news articles, web pages containing computer security advice, and stories about the experiences of friends and family. Using a Latent Dirichlet Allocation topic model, we found that security information from peers usually focuses on who conducts attacks, information containing expertise focuses instead on how attacks are conducted, and information from the news focuses on the consequences of attacks. These differences may prevent users from understanding the persistence and frequency of seemingly mundane threats (viruses, phishing), or from associating protective measures with the generalized threats the users are concerned about (hackers). Our findings highlight the potential for sources of informal security education to create patterns in user knowledge that affect their ability to make good security decisions.

**Key words**: news; informal learning; security; users.

## Introduction

Cybersecurity has a people problem. A large number of the exploited vulnerabilities in computing systems involve users of those systems making bad choices. For example, Anderson [1] found that the majority of security issues with automated banking machines are due to users making incorrect or inappropriate decisions. A large proportion of attacks on the Internet targets vulnerabilities in end users rather than vulnerabilities in technology [2]. End users are vulnerable because they often have a relatively poor understanding of computer security issues [3], yet they still make many security-relevant decisions every day.

Few people are innately talented in security; most need to learn about cybersecurity threats and how to protect themselves and the technologies they use. Cybersecurity is not easy to learn, though; feedback is rare and often difficult to associate with specific decisions [4]. Instead of direct learning, people rely on others [5–7] to help them learn indirectly what cannot be directly experienced. This *social learning* is common in many places in life [5], and often occurs when people tell stories or provide advice to each other [6].

We identified three important sources from which nonexpert computer users can learn about cybersecurity: *articles* in traditional news outlets such as newspapers, *web pages* from third parties intended to educate end users about security, and personal *stories* told, much like gossip, between people. All three sources represent different ways that security knowledge is communicated to end users. Web pages are generally the most authoritative; people often turn to these when seeking computer security expertise online. They also communicate the concerns that important organizations like the government think nonexperts should be aware of. Personal

stories reveal both the knowledge of nonexperts and what nonexperts are concerned about. And news articles tend to focus on issues relevant to a larger society rather than mundane, everyday issues.

These communications are the raw material that end users have to learn from. However, most studies that address what nonexpert end users know about security do not analyze potential sources of their knowledge. To better understand similarities between potential sources in what they communicate about security, we collected a dataset of security communications from each source: 301 personal stories, 1072 news articles, and 509 web pages. Using a Latent Dirichlet Allocation based topic model, we identified 10 major topics that were covered by these communications, which we describe in detail.

Most of the communications were about Phishing and Spam, Data Breaches, Viruses and Malware, and Hackers and Being Hacked, while fewer communications cover Mobile Privacy and Security, or Criminal Hacking. We found that hackers are a major concern in personal stories, but rarely appear in expert advice web pages. Both phishing/spam and viruses/malware commonly appear in web pages and personal stories, but have largely disappeared from news articles. Personal stories often draw connections between who is attacking (hackers) and how they are attacking (viruses), whereas the web pages usually draw connections between attacks (phishing, viruses) and protective measures (passwords, encryption). Our results suggest that communications between end users focus more on *who* conducts attacks, communications representing expert advice focus on *how* attacks are conducted, and communications from the news focus on the *consequences* of attacks. No single source is sufficient for an end user to learn from. However, there were some topics that were addressed by all three sources. In particular, Credit Card and Identity Theft was of relatively equal interest to all three.

## Related work

### Making security decisions

For most everyday computer users, protecting one's computer from security-related problems is difficult. Threats and attacks [2] are constant and pervasive, and as a result end users must make a wide variety of computer security decisions, despite not having the training or expertise for such decisions [8].

Many experts consider end users to be inherently insecure [9, 10]. Because of this, designers of computer security systems have advocated removing users' decision making from security systems as much as possible [11, 12]. However, there are some tasks that humans are simply better at than computers [13], and there are some activities (like rebooting) over which users should be able to exercise some discretion [14]. Therefore, system designers often involve users in everyday security decision making.

Most people find such decisions difficult to make. People generally think about security only when something goes wrong [15], and do not have a good understanding of what a security risk looks like in practice [16]. Most people use simplified mental models of attackers [3] to help make decisions. These simplified models do not capture the complexity of many real-world situations, but instead mostly use metaphors to describe and reason about security problems [17]. The mental models of novice users are often very different than those of security experts [18]. Another common strategy among end users is to delegate security decisions to a trusted other, such as a security expert or an organization [19].

All of these strategies, however, require some amount of knowledge about computer security. Awareness of risks, threats, and remedies is important for being able to cope effectively with problems and resolve them [20]. That awareness and knowledge may be incomplete or inaccurate [3, 21]. Even when people recognize threats related to security, like Viruses and Malware, this recognition is only of broad categories rather than specific details and actionable knowledge they would need to adequately protect themselves [22]. Nevertheless, people need to learn about security, because they must make security-related decisions as they use their computers on a daily basis [4]. While experts and novices sometimes follow the same advice, experts are much more likely to follow security advice that defends against larger classes of attacks, such as using different passwords across different websites [23]. And Kang *et al*. [24] found that awareness of threats among both experts and novices is related to taking protective action, but people who learned about threats through past negative security experiences were more motivated to protect themselves than those with no such experiences.

One avenue for computer users to learn about computer security is from one's employer in the workplace, through security education, training, and awareness (SETA) programs [25]. However, computer security training programs tend to be motivational and persuasive, rather than factual—more about encouraging compliance with policy than communicating knowledge and skills [26]. This training also tends to be decontextualized (best practices rather than situation-specific responses) or focuses on routine activities, which makes it hard to apply to real problems or situations which are more complex [16]. Users who have not received formal training report that they are mostly self-taught, or have learned from experience, or from people they know like coworkers, friends, and family members [27].

### Informal learning

Learning is not limited to formal educational settings like classrooms. Most people continue to learn as adults in less formal ways. Marsick and Watkins [28] make a distinction between *informal* learning and *incidental* learning. The primary difference is intentionality: informal learning happens when people intentionally choose to seek out new ideas, while incidental learning happens "en passant," as a by-product of other daily activities [28].

When people learn informally, they intentionally choose to seek out and learn about new ideas. However, this learning is usually much less structured than in a classroom setting and is usually self-directed [28]. Informal learning is integrated with existing daily routines, though it is usually triggered by some internal or external "jolt." Despite being intentional, it usually is not a highly conscious or structured activity, and is often haphazardly conducted and influenced by random chance. It is also often linked to the learning of other people and done as part of a group [29].

In contrast, incidental or implicit learning happens "independently of conscious attempts to learn and in the absence of explicit knowledge about what was learned" [30]. Often this learning happens during everyday activities. Eraut [31] separates this type of learning by how much cognition is happening when the learning takes place. He talks about near-spontaneous or *reactive* learning that happens in the middle of some other action when there is little time to think. This is distinct from *deliberative* learning where a person takes the time to deliberate and think through some situation and engage in deliberative activities such as planning and problem solving. Deliberative learning occurs when there is a clear work-based goal, and learning happens as a by-product [31].

Almost all of these theories of learning posit some form of feedback loop [28, 31]: people make decisions, act on those decisions, observe the consequences of those actions, and then update their knowledge. However, for many cybersecurity decisions, this feedback loop is broken; people often cannot observe the consequences

of their actions. This means that people often do not have enough information from past experience to estimate the likelihood that they might experience a computer security issue in the future, or what the consequences of that issue might be [32]. For example, if a person's credit card information is stolen and used, it is very difficult to trace the breach back to the decision that enabled it, and therefore to learn a lesson that would help them avoid the problem next time.

This broken feedback loop can inhibit learning about security. In particular, it makes it prevent incidental, implicit, or deliberative learning from occurring [31]. If it is not possible to observe outcomes, then people cannot connect the consequences of decisions with the initial choices, and therefore they cannot update their knowledge.

Broken feedback loops are not unique to cybersecurity. It can be difficult to connect actions to consequences in many domains, including health, business, and politics. To cope, humans have developed sophisticated methods for *social learning*: learning how to behave from other people rather than from direct experience [33]. Social learning can occur simply by observation, watching other people take actions and incur consequences, and by observing when others receive rewards or suffer punishment [33]. Modeling one's behavior after watching what others do is especially common in unfamiliar situations [34]; this can even happen unconsciously when people follow descriptive social norms [35].

Not all social learning comes from direct observation, though. Much of what people learn comes from exchanging knowledge and experiences through interacting with others. Most formal schooling, for example, includes direct instruction that teaches people how to behave. In addition, informal stories told about things that have happened to other people can serve as implicit instruction and indirect observation [6].

### Learning about security

Many studies of computer users' security-related intentions and behaviors focus on awareness and knowledge as a necessary but not sufficient condition for people to make appropriate security decisions to protect themselves and their computers [36–39]. In other words, people need to know something about computer security threats and how to mitigate them in order to make good security-related decisions and behave in a secure manner. However, these studies typically do not address where that awareness and knowledge might come from in the first place.

Several researchers have hypothesized that there are many possible sources of security-related information available for computer users, such as retailers and vendors of software and professional IT services, websites of varying provenance and credibility, friends and family, corporations and governments, and the media [22, 40]. Furnell *et al*. [36] asked computer users who they would turn to for help if they had a computer security-related problem, and around 40% said friends or relatives, public information or websites, and IT professionals. However, very little is known about whether and how much computer users rely on these ways of informal learning about computer security-related topics and behaviors [41]. We examine three different sources of information that people can use to help them indirectly observe the behaviors and outcomes of cybersecurity decisions in others, and receive information and instruction about how they should behave.

Professionally produced *web pages* are a method of semi-formal instruction that organizations and governments are currently using to help people learn more about cybersecurity. Organizations already do this for internal purposes, hosting web pages that employees use for mandatory security training [16]. These web pages often include lists of best practices, definitions, and "dos and don'ts." Companies, organizations, and governments have an interest in improving computer security on the Internet, and as a result they make information like this available to the public as well.

Interpersonal *stories*—basically, cybersecurity gossip—allow people to hear about the decisions and consequences of others indirectly, and often also include lessons about how to behave [7]. Social information sharing is an emerging area of research into how knowledge about computer security might be obtained by computer users. In the workplace, employees say they rely on coworkers for information about what to do in a security-related situation when coping with a problem, and also learn from coworkers' mistakes [38]; the same is true of home computer users and their family and friends [41].

*News articles* often include noteworthy descriptions of cybersecurity incidents and security advice important to society. Exposure to information via mass media can cause examples of potential threats and harms to be more accessible in a person's memory, and therefore people may come to believe occurrences are more likely than they actually are. An example of this is fear of violent crime initiated by exposure to accounts via television news [42]. Cultivation effects like this also exist for print news [43]. There is some evidence that people pay attention to news articles about security threats and breaches, and share them via social media and other mechanisms with people they care about to warn them about potential problems [41].

## Method

Our goal in this study is to assess and describe the content of communications from each of three sources and compare the topics covered by each one. We focused on topics because we are interested in what a user might learn about security from each of these types of documents. All three sources are primary methods where users can informally learn about security. In contrast to formal education that one might encounter in an organized high school or college curriculum, these are sources that a user might encounter as they socialize, read the news, and surf or search the web.

### Data collection

We collected three separate datasets as the part of an ongoing research project. Our final corpus for analysis consists of 301 interpersonal stories about security, 1072 news articles, and 509 web pages, for a total of 1882 items.

#### Interpersonal stories

We began by collecting examples of stories that people tell each other about computer security [7]. We conducted a survey in December 2011 and January 2012, and based our questions on a similar survey that collected examples of interpersonal gossip for analysis [6]. We recruited subjects from undergraduate courses at a large Midwestern university. Subjects received course credit for participating. The survey announcement went out to 728 students across five different course sections. This number is based on total enrollment and does not control for students who may have received the announcement from multiple courses. We received 301 valid and complete responses, for a 41% response rate.

Eliciting computer security stories is difficult to do without biasing subjects. In pilot tests, we found that providing a definition of computer security biases subjects to focus on examples in the definition rather than tell stories from their own experiences. In the final

survey, we asked subjects to follow four steps, each of which resulted in unstructured text responses being recorded by the survey:

1. List "as many computer security problems as you can think of."
2. List ways to "protect yourself and your computer from computer security problems or threats."
3. List "times in the past when you remember being told or reading about a story related to computer security."
4. Choose one story "for which you can most easily recall details" and "write the story as if you were telling it to a friend" using as much detail as possible.

Responses to this last prompt are the stories that we analyze here. Appendix 4 includes some example stories that subjects told.

Rader *et al.* [7] presented additional analyses of these stories; however, that paper focused on self-reported responses to survey questions asking about features of the stories, rather than the actual content of the stories. Also, it only briefly presented a high-level content analysis of the stories that was conducted by human coders, unlike the computational topic model including three types of documents that we present here.

### News

To identify what everyday computer users might learn about computer security from journalists and the news media, we collected a dataset of newspaper articles. We selected 16 large newspapers and collected all computer security-related news articles that appeared in those newspapers during 2011. We collected news stories from newspapers that target regional, national, and international audiences.

When choosing regional publications, we identified nine newspapers that represent all US regional areas including the Northeast, the West, the Midwest, and the South. Each regional newspaper had a Monday through Friday daily circulation average of more than 20 000. We chose three newspapers which focused on the USA at a national level, each with a Monday through Friday circulation average of more than one million. We also chose four English-language non-US newspapers, including one from Australia, one from Great Britain, one from Canada, and one from India. All were published daily at the time the study was conducted and were printed in the traditional newspaper format, and all had a "technology" or similarly themed section. Appendix 2 lists the newspapers and their circulation at the time of data collection.

To identify news articles about computer security, we created a list of 25 phrases commonly used when discussing computer security issues and used those phrases to search the newspaper archives for articles. Most newspapers could be searched via LexisNexis (http://www.lexisnexis.com/, 9 November 2015, date last accessed), but for those with restricted availability, we were able to search for articles via ProQuest, Google News, and in the case of The Chicago Tribune, a subscription service run by the newspaper itself. The search phrases used are in Appendix 2, and include phrases such as "computer hacker" and "online firewall." We avoided using words such as "virus" and "infected" that might be ambiguous and used in other fields such as medicine, although in spot checking the data we found that our searches did return articles about computer viruses. Based on manual spot checking of newspaper contents during the study timeframe, these 25 phrases identified a large proportion of relevant articles for 2011.

As a result of the searches, approximately 1100 articles were initially retained for our sample. Blog posts were not considered "articles" and were not selected for analysis. Editorials were considered "articles" and were included due to the amount of attention readers typically devote to editorial pages of reputable newspapers nationwide. After removing duplicates, we were left with 1072 news articles.

### Web pages

In the summer of 2012, we collected a dataset consisting of web pages related to informal (i.e., nonclassroom) education about computer security. We defined "security education" as any information produced by an organization that would benefit in some way if users behaved more securely, that can also be said to be an authority in the field for the purposes of instructing a consumer or user on the topic of computer or network security. This definition includes web pages and other online documents that, although targeted toward different audiences and varying in terms of technical complexity, are united by their intent to inform members of the general public about computer security-related topics.

We focused on state and federal government agencies, university IT departments, and corporations such as Internet service providers, social media companies, and banks, as sources for computer security information for the general public. These sources all have a vested interest in a well-informed and secure public. While the authors of the web pages are not necessarily security experts, these web pages come from organizations that can be reasonably believed to employ experts, and the advice contained in the web pages is likely to carry the credibility of expertise.

We began by collecting materials from the websites for US government agencies (chosen from the official list at http://www.usa.gov/directory/federal/index.shtml, 9 November 2015, date last accessed). We also randomly selected five US states and collected materials from the websites from those state governments (all 50 states provide computer security materials on their websites). We collected materials from IT department subdomains of a random sample of institutions on the Carnegie Foundation's list of US universities and colleges. And finally, we brainstormed a list of types of corporations that would be motivated to inform their customers and users about computer security, including software producers that release updates frequently, like operating systems, web browsers, and PDF readers; social network sites; ISPs; antivirus companies; and banks. Within each type of company, we selected the top two by market share in the summer of 2012 and collected materials from their websites. Appendix 3 contains the final list of organizations.

For each organization, we conducted a series of Google searches restricted to that organization's domain. We identified 45 keyword pairs that commonly return computer security education documents (listed in Appendix 3), and conducted a separate search for each keyword pair, after disabling Google's personalized search feature. A member of the research team downloaded each of the top 50 results for each search, identified only the documents that concerned computer security, and then removed duplicates. In total, we identified 916 web pages. Two other members of the research team reviewed each of these pages and removed those from the dataset which were not about informing users about computer security, which were targeted at computer security experts, or which were primarily multimedia (images, video) and not text. The final dataset includes 509 web pages.

## Documents and context

Stories, with a mean word count of 95, were much shorter than both news articles ($M = 617$) and web pages ($M = 971$). Both the news article and web page datasets had a number of outliers that were significantly longer than other documents. In the news dataset, 12 items (1%) were longer than 2000 words ($M = 3152$,

SD = 1709). Thirty-one items (6%) in the web pages dataset were longer than 2000 words ($M = 3763$, SD = 1972). Table 1 has additional descriptives.

For the stories, survey respondents were instructed to write the stories in first person, and most of them did. Most stories consisted of a short description of a computer security-related incident that had happened to a family member or friend of the respondent. The stories were written as narratives that included features like symptoms that allowed the people referred to in the story to recognize that there was a problem that might be related to computer security, and whether the problem was resolved or not. Some stories contained explicit advice in addition to the narrative elements (e.g., "Do NOT respond to it [shady email] or click on the link," STORY344), but most did not. Examples stories can be found in Appendix 4, and more details about their content can be found in Rader *et al.* [7].

The news articles were quite diverse in both format and style. They range from hard news, covering events of local or national economic and political importance, to softer stories about celebrities that had been victims of data breaches and ways that hackers are portrayed in popular culture. Some news articles took an approach that was more educational, like an article that contained a Q&A with a computer security expert about security issues related to using public wifi networks (R395, "Free Wi-Fi Can Cost You," Chicago Tribune). Others were narrative descriptions of efforts organizations are undertaking to recover from security-related incidents (N236, "RSA Faces Angry Users After Breach," New York Times). Example news articles can be found in Appendix 5.

The web pages we collected describe security threats and concerns, and provide advice, tips, or instructions to readers about how to deal with these issues or incidents. Many take the form of definitions of computer security-related terms, or checklists of things users should and should not do to keep their computing equipment safe online, or recover from a breach or identity theft situation. Some consist of software feature descriptions and tutorials meant to educate users about how to use tools that can protect them. Many contain references to additional content users can read if they want more information. Example web pages can be found in Appendix 6.

### Current events in 2011–12

Naturally, many of the documents we collected refer to events that were recently occurring around the time we collected the data. This focus on current events can shape which topics were included in the stories, news articles, or web pages.

One of the major events that occurred was a breach of the Sony Playstation network that exposed many users' personal and financial information (http://www.wired.com/2011/04/playstation-network-hacked/, 9 November 2015, date last accessed). A similar attack occurred against RSA data security (http://www.nytimes.com/2011/03/18/technology/18secure.html, 9 November 2015, date last accessed). Two major hacker groups, LulzSec (http://knowyourmeme.com/memes/events/lulzsec-hacks, 9 November 2015, date last accessed) and Anonymous (http://www.forbes.com/sites/andygreenberg/2011/04/04/anonymous-hackers-bring-down-sony-websites/, 9 November 2015, date last accessed), entered the public spotlight when they conducted hacks of a number of highly visible websites. Also, the movie *The Girl with the Dragon Tattoo* was released (http://www.imdb.com/title/tt1568346/, 9 November 2015, date last accessed); the title character is a hacker in the movie, which caused much discussion about hackers in popular culture. Multiple documents in our corpus mention each of these events.

**Table 1.** Number of words per document for each dataset: web pages, news articles, and personal stories

| Type | Mean | Median | SD |
|---|---|---|---|
| Web pages | 795 | 566 | 972 |
| News articles | 617 | 532 | 458 |
| Personal stories | 95 | 83 | 50 |

## Analysis

To understand what topics were discussed in these documents, we used a computational topic modeling algorithm to identify multiple distinct topics. Since "human communication is complex and multi-layered and therefore interpretation is rarely simple or straightforward" [44], we felt that human coding of the documents could be biased by properties of the documents such as form, organization, and style. We used Latent Dirichlet Allocation (LDA) to analyze the words used in the documents and identify a set of topics for further investigation.

LDA is a mature technique (introduced in 2003 [45]) that has been used for topic analysis by researchers in history [46], literature [47], sociology [44], political science [48], public policy [49, 50], and science and technology studies [51], among others [52].

LDA is a type of probabilistic topic modeling. It is a "bag of words" technique. This means that it looks at frequencies and co-occurrences of words within documents, and in common across documents. The order of the words and documents does not matter for the way it detects topics.

Topic modeling with LDA makes some assumptions. At a conceptual level, it reverse-engineers the hidden structure of underlying topics from which the observed documents were assumed to be generated as they were created, based on the words used in the documents. LDA assumes that documents in a corpus are composed of a known, fixed number of topics or themes, and that the words in each document are all related to the underlying topics within that document. The words in one document are evaluated in the context of the words in all the other documents in the corpus being analyzed [53]. LDA also assumes that all documents in the corpus share the same set of topics, just to varying degrees. This means that we cannot claim that we have found ALL the relevant topics to computer security informal learning; the topics described in this article are entirely based on the words used in our corpus, which were determined by our sampling frame.

Topic modeling using LDA produces a set of themes present to varying degrees in the documents. These themes or topics are corpus-wide patterns in the way words are used. LDA does not produce a definitive categorization for what each document is "about," or a representation of what any given person would take away from reading each of the documents, or a quality assessment of the information within each document.

We used a topic modeling toolkit called MALLET [54] for our analysis. MALLET is commonly used by digital humanities researchers for text analysis projects [55]. We combined the three datasets (news articles, web pages, and stories) into one corpus for analysis, and identified topics without regard for source dataset. We used a standard list of stopwords (words that the topic modeling software ignores), augmented with words common to specific datasets but unrelated to computers or security.

LDA requires us to prespecify the number of topics to identify. We generated topic models for 8–20 topics, and also 25, 30, 40, 50, and 100 topics. After careful examination, we determined that a model with 10 topics produced conceptually distinct topics without identifying individual newsworthy events or creating topics including only very small numbers of documents.

## Computer security topics

We identified 10 computer security topics across the three datasets. Figure 1 illustrates their prominence in the entire corpus. In all of the security education materials we gathered, the most commonly discussed topic is *Phishing and Spam*. The second most common topics, with roughly the same prevalence in the entire corpus, are *Data Breaches* and *Viruses and Malware*. The least common topic is *Mobile Privacy and Security*.

LDA assumes that each document in a corpus is composed of *all* topics. However, some topics are more prevalent in any particular document than others. This allows us to identify which topics are the most commonly discussed. The weight of each topic in the full corpus is listed in Table 2. Nearly all documents consist of at least two or three topics with a weight greater than 0.10. For each topic, we counted the number of documents that had that topic listed as the primary topic (largest weight for that document) and the number of documents that listed the topic as the secondary topic. On average, the primary topic had a weight of 0.56 (SD = 0.17), and the secondary topic had a weight of 0.21 (SD = 0.09).

The topic modeling algorithm assumes that topics are made up of words; it produces a set of words for each topic that have a high probability of being associated with that topic. For each topic below, we present this list of *high probability words*, and describe the relevance of each topic for computer security. We do not provide an example document for each topic, because documents consist of multiple topics. Instead, we describe common patterns in how these documents communicate about these topics to end users.

### Phishing and Spam (PhaS)

*email information account phishing mail message spam personal Internet site website address messages click password web facebook links link*

Phishing is a common form of online scam where criminals attempt to trick users into revealing sensitive personal information via emails that upon first glance can appear genuine, but in reality are not [56]. The information users reveal is then typically used for financial or Internet fraud. Phishing and Spam are a large problem with email in society right now. Approximately 1 in 900 emails was a phishing scam in 2014 [2]. Every day, about 28 billion spam emails are sent around the globe [2]. Dhamija *et al.* [57, 58] found that these types of attacks work because most users are either not aware of indicators of scams or do not pay attention to such indicators. Since these types of scams directly target and exploit end users, end users need education to protect themselves from such attacks.

Out of the 10 topics we identified, *Phishing and Spam* was the most prevalent in the corpus, with overall weight of 0.27. Most of

**Table 2.** List of topics identified

| *Topic name* | *Corpus Weight* | *Main topic* | | *Second topic* | |
|---|---|---|---|---|---|
| | | # | % | # | % |
| PhaS | 0.27 | 266 | 14 | 286 | 15 |
| DtBr | 0.23 | 220 | 11 | 241 | 12 |
| VraM | 0.23 | 243 | 13 | 220 | 11 |
| HaBH | 0.23 | 139 | 7 | 282 | 15 |
| PsaE | 0.20 | 139 | 7 | 170 | 9 |
| NtnC | 0.19 | 245 | 13 | 181 | 9 |
| CCaIT | 0.19 | 166 | 8 | 177 | 9 |
| PaOS | 0.17 | 124 | 6 | 143 | 7 |
| CrmH | 0.14 | 239 | 12 | 107 | 5 |
| MPaS | 0.10 | 101 | 5 | 75 | 4 |

"Corpus weight" is the weight of each topic by the LDA algorithm across the entire corpus. "Main topic" and "Second topic" show the number and percent of documents in the entire corpus with each topic as the most prevalent topic in the document, and as the second most prevalent topic in the document.

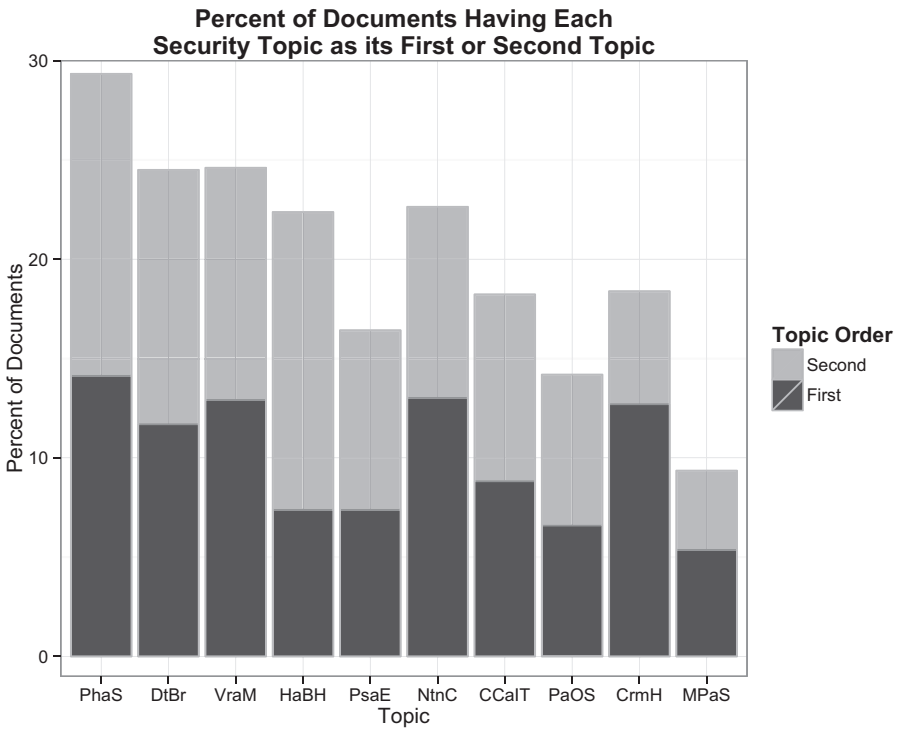


**Figure 1.** For each topic, the percent of documents that has that topic as its main or secondary topic.

our documents about phishing focus on its delivery method, including words such as "email, account, mail, spam." Many documents that have high weights for this topic include definitions and examples of phishing (including specific forms like "spear phishing" and SMiShing), advice for how to identify phishing both before and after one has become a victim, what to do if you become a victim, and tools to help users avoid being exposed to phishing scams. Some documents try to help users identify what phishing attacks look like, and give examples of tactics scammers use to prevent the messages they create from being blocked by spam filters. Many include reminders that companies like banks and employers will not send requests via email asking for login credentials or other personal or account information. Finally, a few documents describe tools such as browser plugins and spam filters that help users to not become victims.

### Data Breaches (DtBr)

*data sony customers information hackers breach online network attack users services accounts playstation attacks personal service systems customer card*

This topic focuses on instances where user information (account details or other personal data) were exposed by "hackers" or "attacks," or by users inadvertently revealing information publicly that should have been kept secure. Data breaches are a growing problem. In 2014, 312 companies publicly reported a breach that exposed data from approximately 350 million users, such as real names, government ID numbers (e.g., US Social Security Numbers (SSN)), home addresses, and financial information [2]. Fifty-seven percent of Europeans reported having their information exposed at some point in the past via a data protection failure or data breach [59]. Data breaches also affect corporations; firms notice a drop in their stock prices after announcing a data breach that involves confidential information [60].

There were many specific examples of data breaches in the corpus, focusing mostly on highly visible organizations such as Sony, RSA Security, Citigroup, Nintendo, Dell, Best Buy, and Walgreen Co. How a breach occurred is usually unknown or goes unreported; instead, the documents focus on the aftermath in terms of costs to both organizations and users. A number of other attacks were included as part of this topic as well. For example, distributed denial-of-service attacks and other security-related events that caused systems to become unavailable used words like "online, services, systems," which are part of this topic. In addition, there were several real-world examples in which files containing personal data (SSNs, health records, drug, and alcohol test results) were exposed publicly on the Internet by mistake when they should not have been, causing embarrassment and potential liability for the organizations at fault.

### Viruses and Malware (VraM)

*computer software anti virus malware windows internet microsoft spyware program viruses firewall malicious programs file files computers system install*

This topic focuses on educating users about "viruses." It contains definitions of viruses, malware, spyware, adware, and worms that are aimed at informing users about the nature of threats from malicious software that self-propagates or spreads after the user has taken some action. These definitions sometimes included detailed descriptions of particular malware (e.g., DNSChanger, Koobface, Mac Defender), or a history of the evolution of computer viruses. Symantec reports that 317 million new pieces of malware were created in 2014 [2]. This represents almost 1 million new pieces of malware every day. Approximately 1 in 244 emails included a malware attachment or a link to malware [2]. End users frequently think about viruses, and use the term "virus" to represent all malicious software [3]. There are many different kinds of malware, and users have difficulty understanding the threats and taking action to protect their computers [3]. Instead, they often delegate that responsibility to software tools like antivirus [19].

Much of the content in this topic is focused on how to avoid being compromised or infected. Tools like antivirus are mentioned frequently, as well as antispyware and firewalls. However, there is also a lot of behavioral advice, such as admonitions not to use p2p file sharing software, and to download only trusted software. This topic also includes advice to install software updates regularly. Finally, this topic includes descriptions of the kinds of symptoms users experience when using a computer that may have been infected. These symptoms are often nonspecific, like slow performance, pop-up windows in a web browser, or settings that have been changed—things that are very difficult for users to attribute directly to malware. This topic contains very little about what users should do to cope if they experience symptoms like this, or who to turn to for help.

### Hackers and Being Hacked (HaBH)

*hacker computer money asked wrote hacked wanted hard eventually hacking game worked left twitter idea night gave half reason*

There are a variety of different contexts and interpretations in which the word "hacker" is used. It usually means someone who is technically skilled that breaks into computers to gain unauthorized access, but it can also mean an especially talented or skilled programmer. Because "hacker" is an overloaded term, the mentions of hackers in the corpus range widely and also do not overlap very much with each other. For example, documents that depict pop culture impressions of hackers include discussions about the movie "Girl with the Dragon Tattoo," which was released in 2011 in the USA and featured a hacker (someone who breaks into computers) as one of the main characters. There were also documents reviewing books that had been published about famous or well-known hackers, or written by hackers about hacking.

This topic also includes descriptions of things "good" hackers do, like attend hacker conferences and work for the government or companies to try to identify vulnerabilities. It also includes the idea of "hacking" as demonstrating one's skill as a programmer, and using those skills to generate new ideas and invent new things. There were also mentions of the Silicon Valley "hacker ethos" as a way of solving problems. Finally, this topic includes depictions of hacking as criminal activity, although there are few specifics about exactly how that activity is undertaken. Instead, the documents included examples of compromised computers or systems. The Sony Playstation hack appears in this topic as well, but depicted as a "hack" rather than a data breach. There were many mentions of high-profile celebrity account compromises, also referred to as "hacks." There were also examples of problems with one's computer, like porn popups or other symptoms similar to those in the Viruses and Malware topic, but in this topic the source of the problem was attributed to an attack by a "hacker"—a person—rather than malicious software.

### Passwords and Encryption (PsaE)

*information data password network access passwords secure wireless computer system encryption public networks devices sensitive personal computers protect [wi]fi*

Many users are concerned about protecting their computers and safeguarding their digital information. This topic includes two main ways to do this: use encryption, and have good password habits and practices. In general, there is a tradeoff between security and usability. Highly secure systems such as email encryption are often difficult for people to use [61]. However, people do perceive that using stronger passwords makes them more secure [62]. This topic does not address the tradeoff; instead, it focuses on the behaviors and practices users can adopt to take full advantage of the benefits of these technologies. It also includes some information about physical security, such as watching out for shoulder surfing, and controlling physical access to one's devices, especially while traveling.

This topic includes advice about creating passwords, though always from a security standpoint rather than a usability standpoint. This includes descriptions of what a strong password looks like, some of which is contradictory: long, mixed case with numbers and symbols, avoid dictionary words, changed frequently—and yet easy to remember. It also addresses encryption in the context of wireless network security, including advice not to use open wireless networks, to check websites to make sure they use SSL, and how to configure a home wireless network to make it more secure.

### National Cybersecurity (NtnC)

*government cyber internet attacks computer china officials state military iran attack systems united national states department nuclear chinese networks*

Documents in this topic cover computer security in relation to national security concerns. In recent years, cyber attacks either against or allegedly perpetrated by governments have gained widespread coverage and attention, and have also been increasing in frequency. There is much concern about the future of cyber warfare, and the role the security of global networks and infrastructure such as water supplies and the electric grid may play [63].

In our corpus, this topic included specific examples of cyber attacks such as Stuxnet; instances of online espionage; attacks against the US State Department, White House, and Chamber of Commerce; and discussions of whether or not such attacks should be classified as acts of war. There was also coverage of what should be done to protect critical infrastructure from attack, and the marshaling of national security resources such as recruiting "white hat" hackers, and training for the military in cyber warfare. In addition, this topic included discussions of repressive regimes and authoritarian governments using tactics to restrict access to the Internet. There were stories about when Egypt shut down access to the Internet in January 2011, mentions of Internet censorship by Iran and China, and Russia jamming smartphones as a protest in 2011.

### Credit Card and Identity Theft (CCaIT)

*credit information identity theft card report bank number fraud personal account money social online accounts consumer file contact victim*

Identity theft and financial fraud are topics of considerable concern. Identity theft is a growing problem, and is associated with computer security because often the information necessary to steal someone's identity is obtained through compromising enterprise or business systems, or through email or other online scams that trick people into compromising their accounts. A stolen credit card can be sold in the black market for anywhere between $0.50 and $20.00; a scan of a real passport is worth about $1–$2; and a stolen gaming account can be sold for as high as $15 [2].

This topic contains definitions of identity theft, primarily related to criminal efforts to commit financial fraud by obtaining or using credit in someone else's name. It includes definitions of what identity theft is, depictions of the emotional cost and stress of dealing with identity theft, and how to cope with the consequences and aftermath of becoming a victim to identity theft. In addition, this topic includes more detailed and specific advice and instructions for how to prevent identity theft. For example, many documents describe what kind of information a criminal would need to steal someone's identity, and how they might obtain that information. Some documents recommend using strong passwords for financial accounts as a way to prevent criminals from accessing them, and even using cash instead of credit cards to pay for things. Finally, this topic covers how to recognize signs that one has become a victim of identity theft, including strategies such as regularly monitoring accounts and obtaining one's free yearly credit report.

### Privacy and Online Safety (PaOS)

*online facebook social information privacy internet sites kids users children personal web child networking share post content safety protect*

This topic contains information about staying safe online. Much has been written about interpersonal risks associated with Internet use. These risks include unwanted disclosures, interactions with bullies and others out to do harm, and hostile online situations that can transition to real-world dangers. Many people believe that privacy and online safety are personal issues and that we should place personal responsibility on end users for their online safety [39, 64].

Present in documents associated with this topic are discussions of privacy issues related to the use of online social networks, and effectively managing one's digital footprint. In particular, many documents focus specifically on Facebook and using location-based services as activities that involve particularly strong risks. Online bullies, harassment, and sexual predators are among the negative safety outcomes associated with Internet use that we found in the corpus. For example, there are descriptions of the behavior of online predators, and advice for parents on how to identify when children might be involved with one. Cyberbullying also appeared in the documents as part of this topic, as well as exhortations not to become someone who bullies or intimidates others online. Finally, many documents contained online safety tips for parents and children to help them stay safe online. These tips included age-based guidelines for appropriate Internet use, information for parents about age-appropriate limits, and other advice not to trust everything people say online or meet up alone with someone from an online forum or chat room.

### Criminal Hacking (CrmH)

*police anonymous computer hacking lulzsec wikileaks law court crime twitter hackers manning website arrested hacker cyber posted investigation members*

This topic is made up of examples and instances of cyber crime. It is distinct from "Hackers and Being Hacked" in that it is entirely focused on the criminal acts that may be perpetrated by "hackers," and any legal consequences that may occur. Cybercrime can include traditional crimes that are now conducted online (such as harassment or stalking), crimes that have substantially changed as they have moved online (such as credit card fraud), and new crimes that are solely online (such as creating botnets) [65]. While most of the costs of cybercrime to victims are based in traditional crimes moving online, most security expenditures go toward the new crimes [65].

This topic contains general descriptions of criminal activity involving digital technologies, as well as reports of the prevalence of said activity. For example, some of the documents in this topic contain descriptions of the crimes and consequences in the legal system of activities like harassing, stalking, and spying on others using computers. This includes things like hacking webcams to access naked pictures and video streams of women, spouses spying on each other, etc. In addition, this topic includes instances where the criminal activity resulted in some public display or evidence that a hack had taken place, like taking over and defacing an organizations website or posting offensive things on its social media account, and posting information like passwords or confidential documents that were obtained through the criminal activity on some public forum or other website. Finally, this topic includes documents talking about Anonymous, WikiLeaks, and Lulz Security that some might classify as "hacktivism." The activities of these entities are treated in most of the documents that mention them as instances of cybercrime.

**Mobile Privacy and Security (MPaS)**

*mobile phone apps device google app devices apple data android users cloud phones location smartphones store market malware software*

This topic contains information about privacy and security related to mobile devices. This is its own topic, rather than falling under other topics related to privacy and security, because the discussion of mobile security is different from other kinds of computer security advice. Because mobiles are easier to lose and therefore fall into others' hands more often, physical device security is a concern addressed in this topic. Also, approximately 17% of apps on the Android apps store were malware in 2014 [2]; therefore, the app download and software update model are aspects of mobile privacy and security that do not exist in the same way for other kinds of computing devices. As a result, users tend not to think of their mobiles in the same way they do their personal computers, for security and privacy purposes. Few people use antivirus for their mobiles, and few understand that smartphones and tables can be vulnerable in the same ways computers are. These beliefs were reflected in this topic.

Many of the documents focused on trying to educate and encourage users to adopt better mobile security practices, by communicating things like how mobile apps can be shady from a security and privacy perspective, and that users should be very careful when downloading and installing apps. Mobile app permissions and the risk of spyware and tracking technologies in particular, were discussed. Finally, the documents made a platform-related distinction between Apple and Google, and the review policies of the different app stores for mobile apps. In particular, Apple makes more of an effort to review submissions to its app store than Google does. This ostensibly means more malware is available for Android, and Android users must therefore be more careful than iOS users. This was illustrated in our corpus by more documents about security tips for the Android platform than the iOS platform.

## Results

The 10 topics described in the previous section comprise most of the topics that everyday computer users are likely to hear about concerning computer security, and they correspond with existing, known security issues and concerns. Next, we examine patterns in how these topics are presented to users, and what users can learn about them.

### Methods of communication: understanding sources

LDA topic models assume that all topics are present in all documents, though each topic may be present to a varying degree. Some documents feature a particular topic more prominently than other topics. In Fig. 1, we showed which topic was the most prominent topic in each document in the entire corpus, and also which topic was the second most prominent. Fig. 2 breaks each topic down further, by source: interpersonal stories, news articles, or web pages. An overall chi-square test of equality of proportions for the prevalence of each topic within each document source was statistically significant [$\chi^2(18, N = 1882) = 1558$, $P < 0.000$]. We used the Holm–Bonferroni correction for post-hoc chi-squared tests for each topic, and these were all statistically significant at the $P < 0.01$ level. See Appendix 1 for the contingency table and details of each post-hoc test.

By far the most prevalent topic in the stories dataset is *Hackers and Being Hacked*, with 58% of stories discussing hackers as their primary or secondary topic. This topic is also sometimes discussed in the News dataset (22% of documents). However, *Hackers and Being Hacked* is only rarely mentioned in the Web Pages dataset, with only 2% of web pages covering this topic. Interpersonal stories primarily focus on the aspects of computer security that everyday users are most concerned about. The prevalence of *Hackers and Being Hacked* in the stories suggests that this is one of the biggest concerns articulated by end users. Other research has also found this to be a major concern [3]. However, even though this is a concern, our results show that the only place that everyday computer users can really learn about this topic is from each other. Advice from experts communicated via web pages very rarely discusses this topic.

The most prevalent topic covered by the web pages is *Phishing and Spam* at 55% of documents, followed closely by *Viruses and Malware*. These two topics garner the most attention from experts trying to educate end users. However, both of these topics are rarely mentioned in the news articles, only being discussed in approximately 15% and 7%, respectively. This suggests that while advice from experts focuses on these topics, they are likely mundane and not of sufficient interest to warrant news articles being written about them. Both topics also have a strong presence in the interpersonal stories dataset.

The most prevalent topic in the news dataset is *Data Breaches* at 37% of documents, followed closely by *National Cybersecurity* at 36%. These topics are newsworthy and of broad interest to society, but largely do not help everyday users make security decisions to protect themselves. As a result, these topics are rarely discussed in interpersonal stories (12% and 4%) or in web pages (5% and 4%). Similarly, *Criminal Hacking* also follows this pattern. Thirty-one percent of news articles discuss this topic, but only about 5% of stories mention this topic and virtually none of the web pages discuss this. *Criminal Hacking* focuses mostly on the investigation and description of computer-based crimes and criminal groups such as Lulzsec and Anonymous. It is unclear why people do not tell many stories about these incidents and why web pages do not use these real-world incidents when providing expert security advice. However, the depictions of these incidents in news articles focus mostly in investigations and legal ramifications, which are also unlikely to be helpful to end users in thinking about how to protect themselves from attacks.

*Passwords and Encryption* is much more prevalent in the web pages dataset (33%) than in the news articles (11%) or interpersonal stories (9%), though it is present to a degree in all three datasets. The presence of this topic in all three of our samples indicates that organizations, journalists, and end users agree that Passwords and Encryption are relevant for computer security. However, they differ in terms of the emphasis or importance of the topic. In contrast, *Credit Card and Identity Theft* (stories = 22%, news = 14%, web

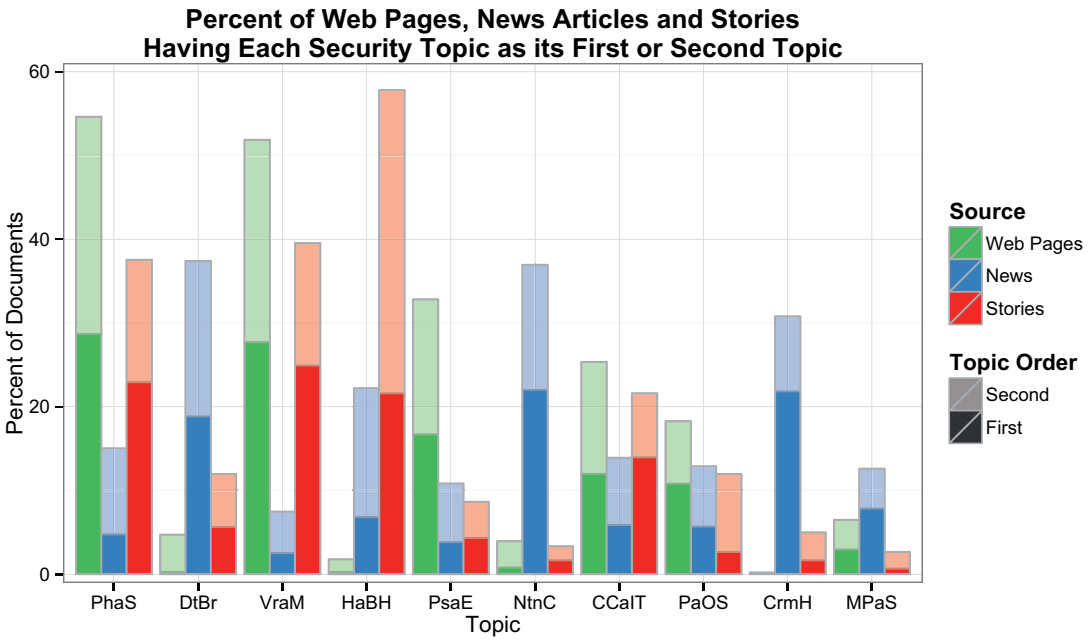


**Figure 2.** Bar chart showing how many documents have each topic as the first or second most prevalent topic, broken down by source type.

pages = 25%) and *privacy and online safety* (stories = 12%, news = 13%, web pages = 11%) feature somewhat more equal prevalence in the three datasets, when compared to the differences in the other topics across datasets. This means that in all three datasets, these are secondary but still important topics. Finally, the topic *Mobile Privacy and Security* is more prevalent in the news and web pages than it is in the stories, indicating that threats and remedies that fall under this topic are not something end users have personal experience with or have much to say about.

The emphasis placed on different topics across the three document sources can help us learn more about what aspects of computer security the producers of these documents are attempting to communicate with their audiences about. These findings illustrate that there are many large differences between the document sources in terms of the topics they cover.

## Content of communication: topic focus

Each individual document in our corpus can substantially include one topic, or possibly many different topics. In the previous section, we identified the top two most prevalent topics in each document, and then used that to characterize patterns across datasets. However, some documents are more focused on a single topic than others. A web page could be solely about passwords, and a news article could easily discuss four or five different topics in a single article. For each topic in a document, LDA produces a weight of that topic in the document, which approximately corresponds to the percentage of the document about that topic. To analyze the topical focus of each document, we decided that a document can be said to be "about" a topic if it has a weight greater than 0.10 for that topic. We chose this cutoff by manually examining a random subset of documents and identifying a cutoff that approximately matched our judgment about when a topic would be recognizably present to a casual reader of the document.

Figure 3 shows the overall distribution of topic focus for the entire corpus. An overall chi-square test of equality of proportions for the prevalence of the levels of topic focus within document source was statistically significant [$\chi^2$ (8, $N = 1882$) = 97.57, $P < 0.000$]. Only 13% of documents in the corpus are focused on a single topic. Most documents cover either two (37%) or three topics (34%). While a document focused on a single topic might provide greater information about that topic, documents that cover multiple topics allow users to discover information about topics other than the one they are searching for. Multi-topic documents, then, have the potential for being better for learning about security because they have the ability to spread information about additional topics.

Not all sources of information have the same degree of focus. Figure 4 shows how focused documents in each of the three sources of information are. In general, web pages are the most focused, as indicated by the greater skew of the distribution to the left side of the graph. Approximately 18% of the web pages only include a single topic, and less than 10% discuss four or more topics. This is potentially a missed opportunity; when novice users go to web pages looking for expert advice, they might learn about other important aspects of computer security if the web pages included other, related topics. This finding could also be an artifact of filtering topics at a weight of 0.10; if a web page included a list of very short statements about each topic such as bullet-pointed advice, it is possible that those topics would fall below the threshold.

In contrast, news articles frequently include multiple topics, with over 20% of documents including information about four or more topics. The news media seems to be doing a good job drawing connections between multiple computer security topics, and when people learn from news stories they are likely to learn about a variety of security issues. Finally, interpersonal stories have a tighter distribution; most stories include exactly two or three topics. There are very few stories focused on a single topic (6%) and also few stories that cover four or more topics (11%). When people talk about security with each other, they tend to talk about exactly two or three topics.

## Content of communication: topic co-occurrence

Because a large percentage of the documents in our corpus discuss more than one topic, we took a closer look at which topics co-occur

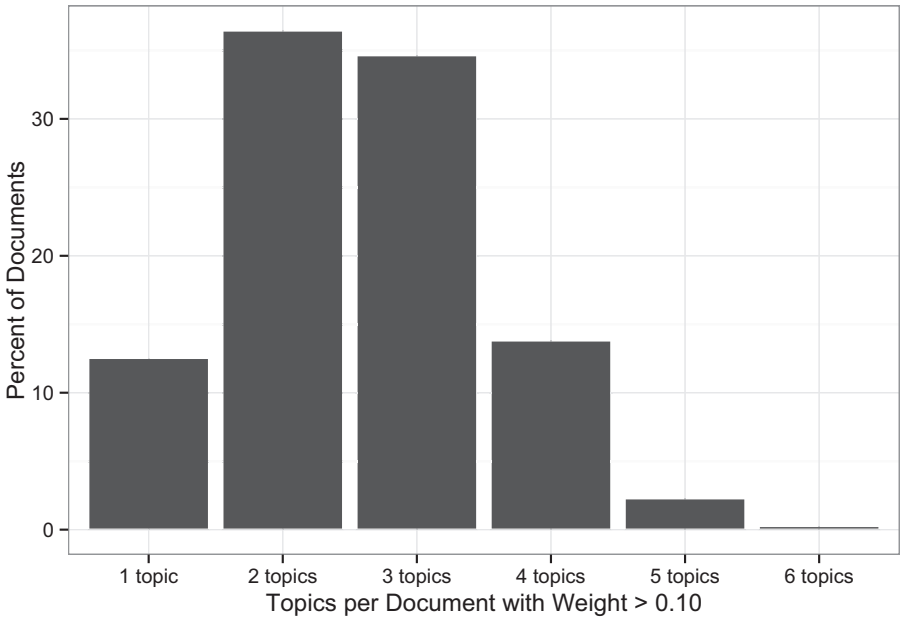


**Figure 3.** Bar chart showing the distribution of the number of topics in each document.

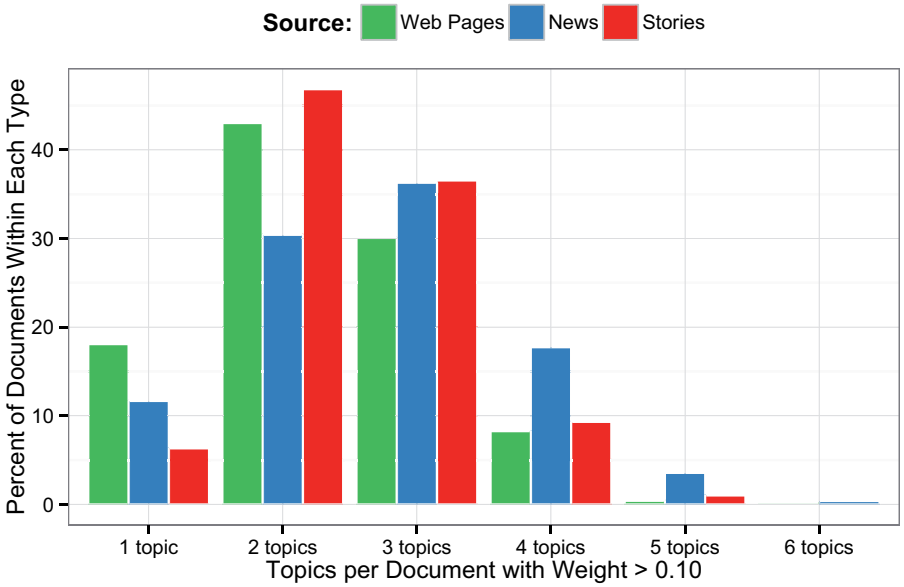


**Figure 4.** Bar chart showing the distribution of the number of topics in each document, by source.

in the same document. As in the previous section, we consider two topics to be present in the same document if their weights of both topics for that document generated by the topic model are greater than 0.10. For each source, we identified which topics commonly co-occur and have graphically displayed this information with a network diagram. Figures 5–7 depict topic co-occurrence relationships between all 10 topics for each source. A thicker line connecting two topics means that the two topics co-occur more frequently in documents from that source than topics connected by a thin line. Only topics that co-occur in at least 1% of documents have lines between them. Node size in the network diagrams represents what proportion of documents from that source have each topic as their first or second most prevalent topic.

### Topic co-occurrence within each source

**Interpersonal stories.** Despite being the shortest documents, most interpersonal stories discuss more than one topic. Figure 5 contains a network representation of topic co-occurrence in interpersonal stories. The most frequent topics to co-occur in the stories are *Viruses and Malware* and *Hackers and Being Hacked*, with 33% of documents including both these topics. *Phishing and Spam* is also strongly connected to *Hackers and Being Hacked*, with 28% of documents including both these terms. (*Phishing and Spam* and *Viruses and Malware* appear in 16% of documents together.) *Hackers and Being Hacked* also appears with *Credit Card and Identity Theft* in approximately 18% of documents. While it is not definitive, this evidence suggests that many of the stories are about various types of attacks (viruses,

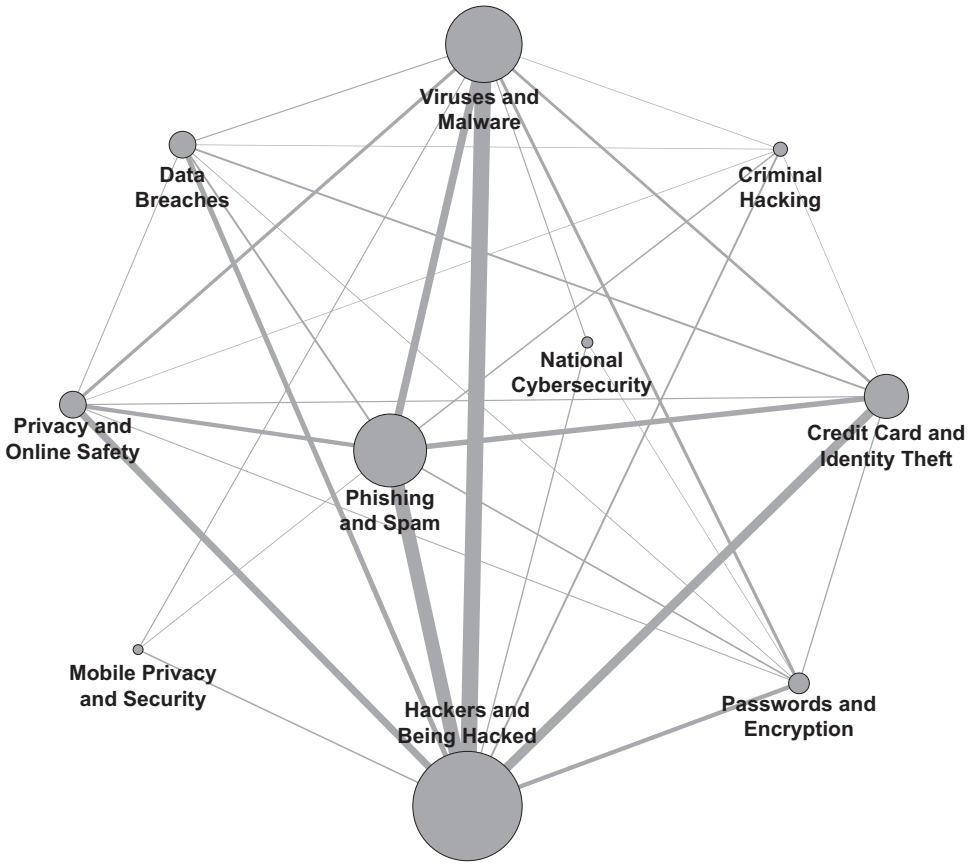


**Figure 5.** Topic co-occurrence in interpersonal stories.

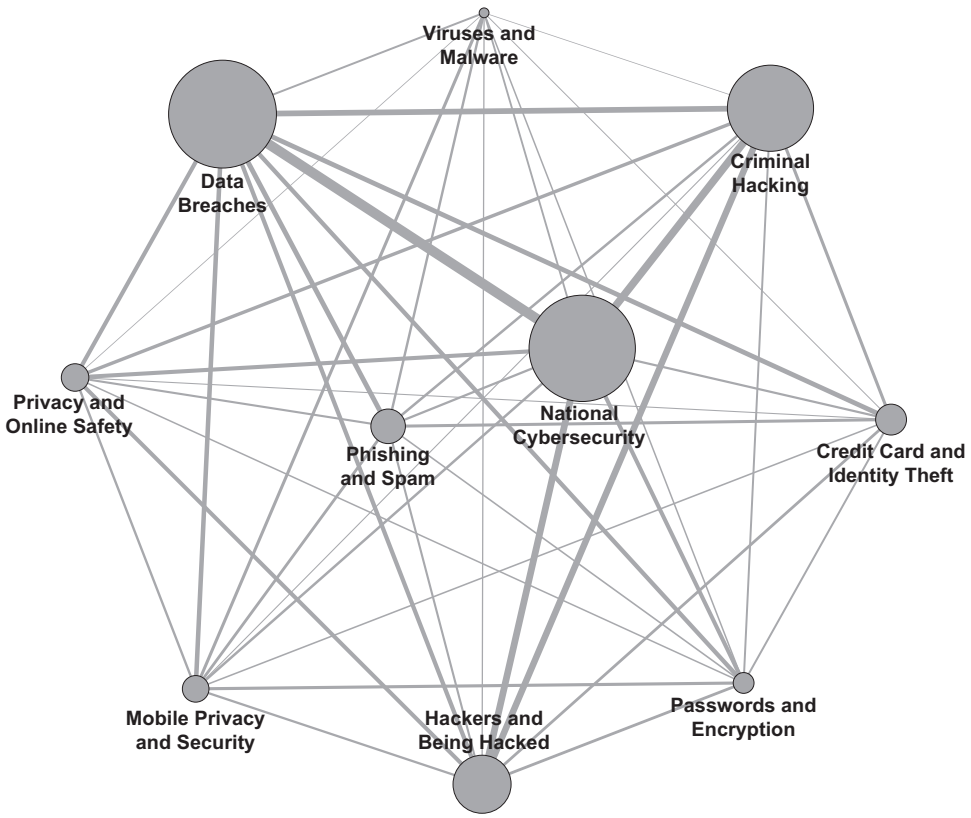


**Figure 6.** Topic co-occurrence in news articles.

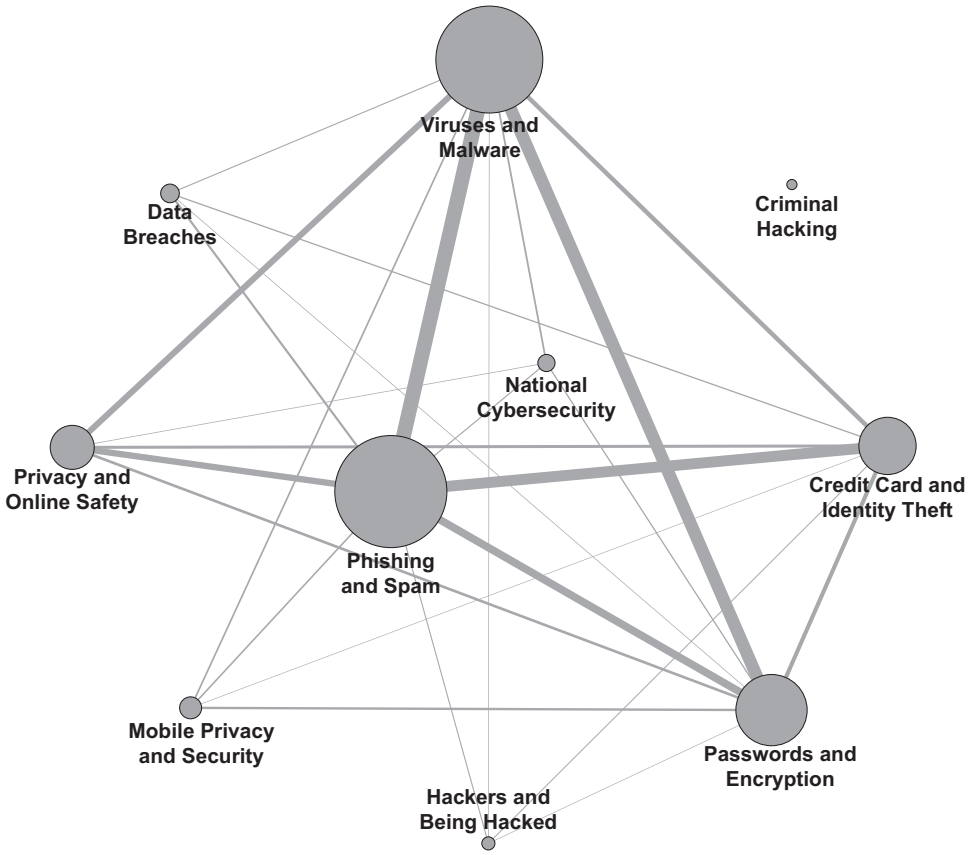


**Figure 7.** Topic co-occurrence in education web pages.

phishing, or stolen credit card information) and also include speculation about who might be behind these attacks (i.e., hackers). It is also possible that users are having difficulty disambiguating the sources or threats that cause the outcomes they experience. Finally, since interpersonal stories rarely discuss issues like *Data Breaches*, *Criminal Hacking*, or *National Security*, these topics rarely co-occur.

**News articles**. News articles discuss multiple topics at approximately average rates. However, there is no pair of topics that frequently co-occurs in the news articles; all 10 topics co-occur with all the other topics. Only four pairs co-occur in more than 10% of news articles, with the most common connection between *Data Breaches* and *National Security* (20% of news articles). This suggests that newspapers are doing a good job drawing lots of different connections across topics related to computer security.

**Web pages**. Expert-produced web pages are generally the most focused documents. When they do connect multiple topics, they frequently connect multiple types of attacks, such as *Phishing and Spam* and *Viruses and Malware* (29% of web pages), or *Phishing and Spam* and *Credit Card and Identity Theft* (21% of web pages). Manually looking through these documents, many of them included lists of potential attacks and the advice for how to protect against them. However, as shown in Fig. 7, this graph is more sparse than the other two graphs, which means that there are co-occurrences between fewer pairs of topics. Interestingly, *Viruses and Malware* is connected to *Passwords and Encryption* in 26% of web pages. This likely occurs because "use anti-virus" and "use strong passwords" are the most commonly repeated security advice from experts.

**Comparing topic co-occurrence across sources**

We can compare patterns in topic co-occurrence across the three document sources to identify ways in which the different producers of computer security documents draw connections between the same topics. For example, the *Hackers and Being Hacked* topic is strongly connected to both *Viruses and Malware* and *Phishing and Spam* among the interpersonal stories. However, in both the web pages and the news articles, *Hackers* rarely co-occurs with either *Viruses* [$\chi^2$ (2, $N = 134$) = 360.62, $P < 0.000$] (all chi-square tests in this section use a null hypothesis of equal proportions within topics and across document sources, and use the Holm–Bonferroni correction) or *Phishing* [$\chi^2$ (2, $N = 141$) = 217.72, $P < 0.000$]. We suspect that users either want to identify who to blame or are looking for a cause for the problems they are experiencing, and this tends to be whoever the person might be that is behind the attack. Similarly, *Credit Card and Identity Theft* is connected to *Hackers and Being Hacked* in the stories but not very strongly in the other two datasets [$\chi^2$ (2, $N = 126$) = 84.55, $P < 0.000$]. Very rarely does expert advice attribute attacks to the people who caused them.

*Viruses and Malware* and *Phishing and Spam* are very strongly connected with each other in the web pages (29%), but less so in interpersonal stories (16%), and barely at all in the news articles [6%, $\chi^2$ (2, $N = 248$) = 177.54, $P < 0.000$]. Web pages tend to provide advice about multiple kinds of threats and protective actions all together in the same document, whereas stories, both interpersonal and news were usually about a single occurrence or event.

*Viruses and Malware* is also strongly connected to *Passwords and Encryption* in the web pages dataset (26%), but barely at all in the other two datasets [stories = 6%, news = 3%, $\chi^2$ (2, $N = 177$) = 190.91,

$P < 0.000$]. Advice in web pages often includes multiple ways to protect oneself, like using antivirus and having stronger passwords, all in the same document. However, end users focus more on cause and effect, and tell stories in narrative order. Neither strong passwords nor encryption fit neatly into a narrative order, and were not something that came up very much in the interpersonal stories (only 8.6% of stories had *Passwords and Encryption* as one of the top two topics).

Other interesting differences in co-occurrence patterns include *Phishing and Spam* and *Credit Card and Identity Theft*, which co-occur in 21% of web pages. This reflects that experts know of the common relationship between attack (phishing) and consequence (identity theft) when providing advice. However, only 6% of news articles and 11% of interpersonal stories draw this connection [$\chi^2$ $(2, N = 208) = 81.73$, $P < 0.000$]. Finally, the *Data Breaches* topic is connected with most other topics in news articles; however, it is not strongly connected to any topics in interpersonal stories except *Hackers and Being Hacked* (10%). This may reflect a belief by end users that hackers are the source of Data Breaches; however, in reality Data Breaches are more often a result of phishing attacks, malware, and human error. *Data Breaches* is not connected at all to *Hackers and Being Hacked* in expert-produced web pages [$\chi^2$ $(2, N = 125) = 47.74$, $P < 0.000$].

## Document composition: similarities and differences

In the previous section, we described the differences we found regarding how information about computer security is scoped and discussed from the three different sources based on our analysis of how topics co-occur within documents from each source. Focusing on the relationship between topics and sources allows us to consider differences in how the documents are created or produced. In other words, when organizations, end users, and the news media communicate about computer security, how do they organize what they say into topics and what topics do they cover? We found that the three sources place a different amount of emphasis on each topic, and that topics which are likely to co-occur from one source are unlikely to appear together when discussed by a different source. This gives us an interesting view into what these documents are communicating about regarding computer security.

We can also examine the data from the perspective of the consumer of the information, such as a hypothetical end user who is seeking information about computer security. This allows us to consider how a consumer might search for information, and what they might find if they were to encounter documents from these different sources. For example, if a user were to go looking for information about, say, a shady looking email they received from a friend, where might that person find information about this? Would an end user searching Google for information using their own vocabulary be likely to come across information that would be helpful to them? In other words, how are the documents from each source different from each other, and what might this mean for end users who are in need of help or who want to learn more?

To answer these questions, we created a network graph to help us visualize the similarity between all of the documents in our dataset, based on the topic composition of each document (Fig. 8). The edges in the graph each represent how similar a pair of documents is to each other, weighted by the Pearson correlation between the topic vectors for both documents. (A topic vector is the list of all 10 topic weights for a given document.) We started with a fully connected graph and then filtered out edges with weight less than 0.80, which resulted in 84 345 edges (connections between documents). The size of each node represents how many other documents that node is connected to, and the nodes in the graph are colored based on which source each document came from: red for stories, green for web pages, and blue for news articles. The edges are colored based on the types of the nodes they connect. For example, if two stories are connected, the edge is colored red. But, if a story and a news article are connected the edge is either blue or red, and the color selection is effectively random in these cases.

We used the Fruchterman Reingold layout algorithm as implemented in the Gephi software [66], which is a well-known graph layout algorithm that produces clusters of tightly connected nodes, to lay out the graph for the visualization. The clusters that the algorithm identified correspond to the topics in the topic model, such that each node within a cluster has the same topic as its most highly weighted topic.

This graph does not provide new insights above what we presented above; however, it provides a different way of visualizing the above results, all in a single image rather than split across many. It is based on the same topic model, though it uses a more detailed visualization that provides some additional evidence that our findings are present in the data.

Our interpretation of the graph focuses on the patterns in how the documents from each source do or do not cluster tightly together into groups. Similarity between interpersonal stories (red) and other kinds of documents are an indication of areas where the way end users talk about security overlaps with the way organizations seeking to educate and news media seeking to inform talk about the same issues. The clusters in the graph where red nodes are closely linked to nodes of other colors are particularly interesting, as well as clusters where red nodes are all but absent.

For example, there are three clusters in the graph which are mostly news (blue), like "Criminal Hacking" in the top right of Fig. 8. This illustrates that documents that are primarily about newsworthy aspects of computer security, like legal consequences of hacking activities, do not overlap much with other computer security-related topics discussed in documents from other sources. Users would therefore be unlikely to encounter information in the news that appears similar to the issues they are facing and hear others like them talking about.

Alternatively, a cluster like *Credit Card and Identity Theft* in the lower left of the figure has very similar proportions of documents from all three sources tightly clustered together. Because the clusters are formed based on similarities between the proportions of topics in each document, this means that the words used in all three sources to talk about causes, consequences, and coping related to identity theft is similar. It means the overall topic composition of documents that are primarily about this topic are similar as well. Organizations create web pages to educate people about it, it is newsworthy, and everyday computer users also experience it and are worried about it. Users concerned about identity theft would therefore be able to find information they can recognize as related to their experiences from any of the three sources, because the words they themselves used to talk about identity theft are similar to the words used in the other types of documents in our corpus.

The cluster for *Passwords and Encryption*, a little above and to the left of center in the graph, is mostly web pages (green) with a few red and blue nodes. This indicates that it is a topic organizations are trying to educate end users about, but that users themselves did not bring up very often in the stories they told about computer security. Since everyday computer users are the target audience for educational web pages created by organizations, this indicates a mismatch between what end users talk about as related to computer security and what organizations want them to know. This disconnect is also reflected in the behaviors of end users, like writing down passwords, which is something that experts advise against as a bad security practice but end users do it anyway [9], and in policies of organizations that consist of "do's and don'ts" rather than cause and effect [16]. If end users do not consider passwords to be something they think is related to computer security, our analysis

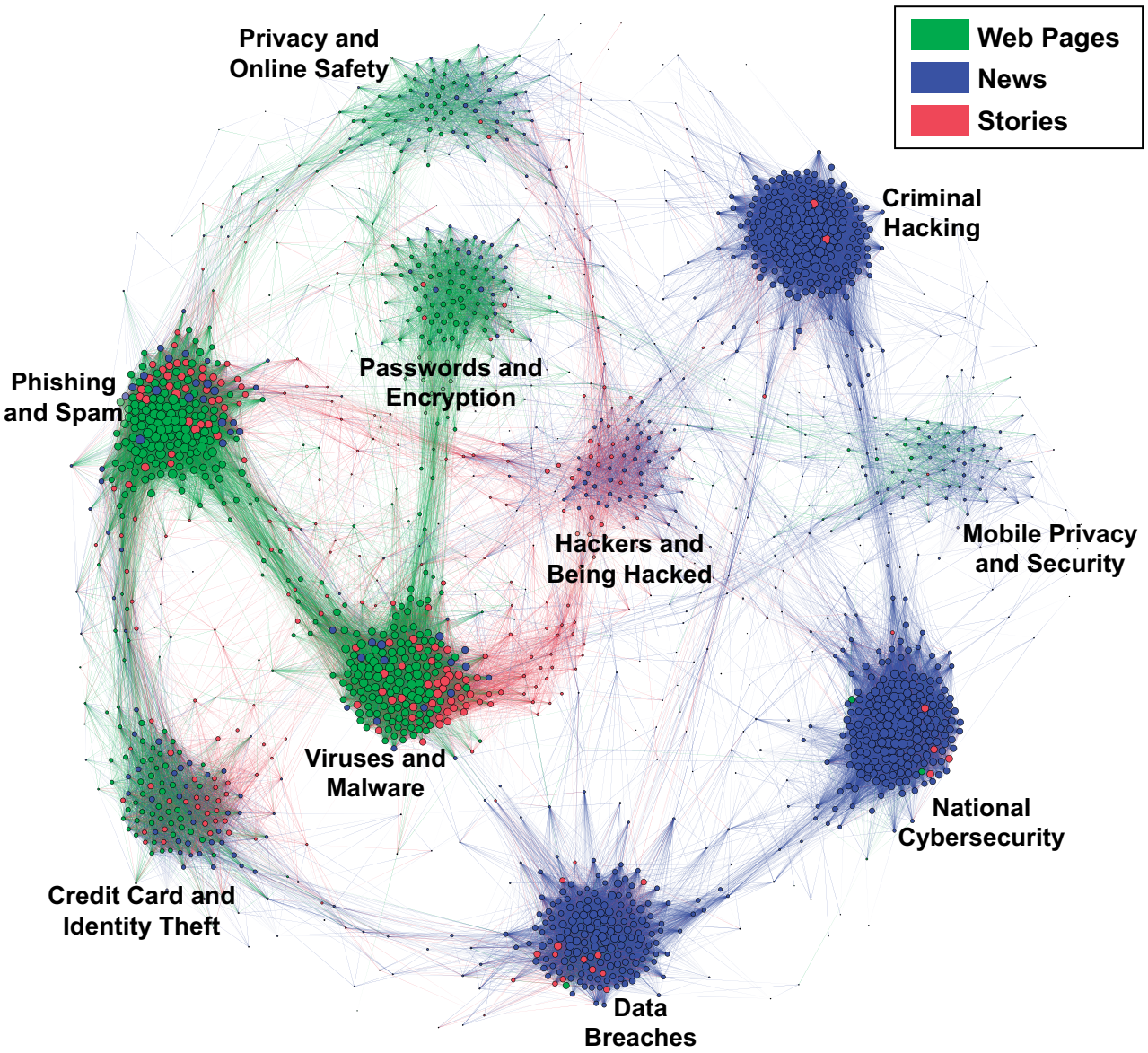


**Figure 8.** The document similarity graph, with clusters for each topic. There is one node for each document in the dataset. The red nodes are stories, green are web pages, and blue are news articles. Larger nodes are connected to more other documents. Edges represent the Pearson correlation between the topic vectors for a pair of documents.

reveals that when they need information about what they consider to be computer security related they are unlikely to encounter advice about protective measures like passwords and encryption online or in the news, because they do not think and talk about it in the same way.

The *Phishing and Spam* and *Viruses and Malware* clusters, center-left in the graph, both contain predominantly web pages but also have some stories and news articles mixed in, indicating that these are topics that are both related to users' experiences, and also discussed in web pages intended to educate them. This is encouraging because this means that some education web pages are using similar language and terminology as end users when addressing pervasive problems such as phishing and viruses. However, from our analysis we cannot tell if it is because the web pages are tailored for the users, or because everyday computer users are using similar language as the education web pages without knowing what they mean. Either way, these clusters indicate that users experiencing problems who turn to the Internet for help have at least some chance of encountering information related to the problems they are having. As we have mentioned before, however, these topics are not very common in the news articles.

Finally, a little above and to the right of center in the graph is the cluster for *Hackers and Being Hacked*. It is mostly blue (news) with some red (stories). This means that stories and news articles resemble each other in the way they talk about hackers and hacking, and the web pages do not talk about hackers much or in the same way as end users and news articles do. This is interesting because one can imagine that end users who see people—hackers—as the source of the threats they face could completely miss information online about protective measures like how to use encryption. Also, reading about legal proceedings faced by those caught hacking or about cyber warfare, two topics that co-occur with *Hackers and Being Hacked* in the news stories, is unlikely to provide useful information to everyday computer users about computer security threats.

## Discussion

### Communication between experts versus everyday computer users

Topic models, including the one we use above, focus on word use; a topic is a group of words that consistently appears within individual documents, and is found across multiple documents. The words that people use are an indication of how they think about an issue, and focusing on language and vocabulary is an approach that has been used by others to study how people think about computer security [18].

Our findings suggest that everyday computer users and experts use different words to talk about computer security concerns. Everyday computer users tend to use a lot of words related to *Hacking and Being Hacked* when discussing computer security: *hacker, hacking, hacked, money, wanted, reason*. These words communicate about who the people are that are carrying out the attacks and their underlying motivations. They also frequently communicate about multiple security topics at the same time. Web pages created by experts, however, mostly use words related to specific attacks such as *Viruses and Malware (computer, software, [anti]virus, malware)* and *Phishing and Spam (email, information, account, phishing)*. Experts focus much less on "who" is attacking and "why" they are attacking, and instead focus on "what" the attack vector is and "how" an attack might be carried out. They also focus on less diverse topics within each document, while drawing more connections between attacks and protective measures.

These findings shed more light on a disconnect that is known to exist between experts and novices in the way they communicate about computer security issues, and also present an interesting opportunity for both sides to learn from each other. By ignoring who is conducting computer attacks and why they do so, experts miss an opportunity to connect with everyday computer users who think and talk about these same kinds of attacks from the perspective of who does them and why. In other words, our findings indicate that a nonexpert user would care more about who an identity thief is and why they want the user's data, than the specifics of what phishing mails look like. Wash [3] found that most people do not necessarily want to protect themselves from every possible attack, and use mental models of "who" the hackers are and "why" they might attack to decide what protections they need to put in place. Information from experts that is intended to educate may miss its audience entirely because everyday computer users are more worried about the source of the attack than how it might be carried out. Gossip about people and their motivations is much more memorable [6]; including additional information about potential attackers and reasons for attacks might make expert advice more approachable and understandable for everyday computer users.

This approach to communicating about security may be challenging for computer security experts, who do not often focus on this aspect. Their attention is directed more toward technical rather than interpersonal issues. Also, the specific identity of an attacker is often unknown. Experts undoubtedly communicate a mental model that is more useful for security: it does not matter who is attacking; what matters is "how" they attack. The method of attacking (phishing versus malware, e.g.) is what determines which security protections are needed. However, speaking to everyday computer users about things they care about using words they are likely to use themselves might help to create a dialogue about protections that is rooted in everyday computer users' concerns, and generalities about characteristics and motivations of attackers may be enough to get users' attention.

When novices communicate with each other, they should focus on spreading information they might already be aware of concerning how attacks are carried out and draw more connections between the method of attack and techniques for protection. The *Credit Card and Identity Theft* topic, which all three of the sources talk about, presents an interesting example that may be a model for other areas of computer security education and training. It is an issue that is newsworthy and for which experts and novices use the same kinds of language. An everyday computer user who has fallen victim to identity theft might focus in conversations with her friends not only about why someone would want to do such a thing, but also any steps she has taken to prevent it from happening again. Even nonexpert users know some important pieces of security advice that can be shared [23].

### Common attacks are important but mundane

Newspaper reporters are taught to include the *who*, *what*, *how*, *when*, and *why* of whatever incident they are reporting in [67]. In this respect, newspaper articles have the potential to be a bridge between the way that novices communicate about computer security, and the way that experts provide advice. However, the news articles in our sample mostly ignore the mundane but important types of attacks that both novices and experts frequently communicate about. Both expert-written web pages and novice-told interpersonal stories frequently discuss *Phishing and Spam* and *Viruses and Malware*. These topics are important types of attacks that affect many people, and also attacks that require user attention and good decisions to protect against. However, newspapers very rarely discuss these attacks, which may mean that the attacks are sufficiently mundane that few specific attacks warrant a news article about it. As a place to learn about computer security, news articles are falling short in this regard.

Instead, news articles related to security are frequently about large-scale attacks such as *Data Breaches* and *National Cybersecurity* issues. While these attacks are clearly important in society, there is little that individuals can do about them, which is probably why few interpersonal stories are about them. As a source of practical informal learning about computer security, news articles mostly focus on larger scale issues that individuals cannot effect while ignoring the mundane but important attacks that computer users face frequently and are able to do something about.

### Informal and incidental learning about security

Informal learning is unstructured and takes place as people seek out and encounter new ideas as they go about their lives, and learn new things that they incorporate into their understanding of the world around them. It is often triggered by a "jolt" [28] that highlights something that they do not know or are wrong about. Das *et al*. [41] wrote about what jolts or "catalysts" like this look like for everyday computer users, in the context of informal social learning about computer security: observing others' novel or insecure behavior, negative experiences, starting to use new technologies and having to configure them, and conversations with experts. This aligns with previous research about formation of mental models; as people have experiences where they encounter an inconsistency between their beliefs and a situation they are experiencing or a problem to be solved, they incorporate new information into their existing mental models [68].

Incidental learning occurs when computer security issues arise as part of everyday experiences such as talking with family and friends or reading newspapers [31]. While incidental learning is not always

as deliberative and careful as informal learning, it happens much more often and can have a strong influence on people's mental models [30]. Both informal and incidental learning are important for computer security because of the broken feedback loop: it is hard for people to learn about how to effectively protect themselves and their computers via direct experience. The contribution of this study is therefore to describe what everyday computer users are likely to encounter and learn from as part of informal or incidental learning.

Users who seek out information about computer security for *informal* learning are likely to encounter mostly news articles and web pages from organizations. In these, they have the opportunity to learn about a wide variety of attacks and how to protect against such attacks. On the other hand, people whose computer security knowledge mostly comes from *incidental* sources such as stories from other people can learn ideas about the kinds of people who attack computers and connected them to broad classes of attacks. Incidental sources are currently very bad at providing information about protections or about connecting related attacks. But sources for informal learning are potentially less memorable. They do not include as much information about who is conducting attacks and why they attack, which is much easier for most people to remember [6].

Additionally, we found that web pages with computer security advice are generally more focused than other sources for informal and incidental learning. When computer users seek information about security for informal learning, they are less likely to encounter information about security topics other than the one they are seeking. Since informal learning is often haphazardly conducted, not well structured, and influenced by random chance [29], this focus limits informal learning. Because web pages intended to educate everyday computer users are more focused, people can only learn about topics that they already are aware of from them. They are less likely to be exposed to information connecting what they already know (like threats) to things they are not aware of (like protective measures or sources of attacks) because it does not co-occur in the documents they are finding.

### Limitations

For each dataset, there is no equivalent of a phone book from which we can randomly sample documents. As such, all three datasets have some amount of bias due to the sampling. For example, when examining the news dataset, we were not able to search for the word "virus" because it is also associated with a large number of medical articles. We tried to address sampling biases with spot checking: in the news dataset, we picked one week and manually looked at every article posted in the Technology, National, and International news sections of multiple newspapers. We then verified that our search terms found all of the computer security-related articles for that week (they did), including ones about topics (like computer viruses) not necessarily covered by the terms. While this does not guarantee coverage, it suggests that we did not miss that much. We spot checked both the news articles and web pages datasets.

All three datasets have biases. The interpersonal stories are all told by undergraduate students (aged 18–24) at a large Midwestern university, and as such might not represent the concerns or experiences of broader groups of people. They do have similar patterns to existing research, though, such as the focus on hackers and viruses that Wash [3] found. The news articles might not include some stories about topics not explicitly searched for. And the web pages includes biases from both the choice of organizations to sample and the use of Google's search engine to find relevant documents. We have interpreted most of our findings as differences between populations of documents, but it is possible that some of the findings are artifacts of the sampling process rather than representative of the larger population of interest.

Also, these documents represent *communications*: what everyday computer users, journalists, and web page authors have chosen to communicate with others about computer security. People have a wide variety of motivations for communication, and not all of them lead to the communications being accurate representations of what the communicator believes or knows. While each document source is aimed at the general population and not technical computer security experts, they each serve a different communication function and differences between the three sources may be caused by this difference in focus.

In addition, communications are often intended to persuade or to mislead or they simply try to make something easier to understand. We cannot know for sure what the underlying population of people believes or knows from these communications; however, we can see how they communicate about it and talk with others about computer security. All of our results should be taken in the context of opportunities for informal learning: what kinds of knowledge is it possible for end users to learn from each other, from newspaper articles, or from expert-produced communications? Additionally, we did not evaluate the effectiveness of the communications; we do not know if people were successfully able to learn anything from these documents.

Since this data was collected, Edward Snowden revealed information about the US Government's use of computer security, and a large public discussion has occurred about the role of government in computer security. This article currently focused exclusively on protection from criminal rather than governmental actions, since that is the focus of the materials we collected. However, it is possible that the dialog has changed to include governmental actors as a result of this public discussion.

## Conclusion

For most computer users, learning how to make appropriate security decisions to protect your computer is rather difficult. Few people have direct experience with the majority of computer-based attacks, and those attacks are constantly evolving. Instead, people generally get their knowledge from informal and incidental sources of social learning: interpersonal stories, news articles, and web pages with security advice.

We collected examples of all three of these sources of informal social learning about computer security, and used a computational topic model to determine which computer security topics they discussed. The interpersonal stories focus mostly on *who* attacks, and drawing connections between attacker and the broad class of attack (virus, phishing). Web pages that the users can go to for expert advice, however, focus on *how* attacks are conducted, and on drawing connections between the type of attack and protective measures. News articles cover the *consequences* of attacks, and draw a wide range of connections across computer security topics.

Users who actively but informally seek out computer security information are likely to find information about attacks and preventative measures, but are unlikely to learn who is attacking or why. Users who only come across computer security information incidentally are likely to know more about the kinds of attackers and some nonspecific types of attacks, but have little opportunity to learn more about protecting themselves. Computer users cannot simply look toward a single source to get a complete picture of computer

security protections; instead they must collect information from multiple sources in order to have the knowledge they need to make good security decisions.

## Acknowledgments

We thank Alcides Velasquez, Zack Girourd, Katie Hoban, Lauren McKown, and Nathan Zemanek for their assistance with sampling, collecting, and cleaning the data. We are also grateful to everyone associated with the BITLab at MSU for helpful discussions and feedback.

## Funding

This material is based upon work supported by the U.S. National Science Foundation under Grant No. CNS-1116544 and CNS-1115926. Funding to pay the Open Access publication charges for this article was provided by US National Science Foundation.

*Conflict of interest statement*. None declared.

## Appendix 1. Statistical details

This table reports the number of documents that include each topic as either the primary or secondary topic. It also reports results of the post-hoc $\chi^2$ test for each topic. *P*-values are corrected with the Holm–Bonferroni correction to correct the family-wise error rate, top 5% for this set of tests. The null hypothesis of each test is that the proportion of documents with the given topic as primary or secondary is the same across all three datasets. Since all tests reject at the 1% level, we can be confident that all differences we observe across datasets are not due to random chance.

| | **Web** | **News** | | | | | |
|---|---|---|---|---|---|---|---|
| *Topic* | Pages | Articles | Stories | | $\chi^2$ | df | *P* |
| PhaS | 278 | 161 | 113 | *** | 272.7 | 2 | 0.000 |
| DtBr | 24 | 401 | 36 | *** | 229.9 | 2 | 0.000 |
| VraM | 264 | 80 | 119 | *** | 409.9 | 2 | 0.000 |
| HaBH | 9 | 238 | 174 | *** | 342.1 | 2 | 0.000 |
| PsaE | 167 | 116 | 26 | *** | 137.4 | 2 | 0.000 |
| NtnC | 20 | 396 | 10 | *** | 291.1 | 2 | 0.000 |
| CCaIT | 129 | 149 | 65 | *** | 33.1 | 2 | 0.000 |
| PaOS | 93 | 138 | 36 | ** | 9.7 | 2 | 0.008 |
| CrmH | 1 | 330 | 15 | *** | 258.1 | 2 | 0.000 |
| MPaS | 33 | 135 | 8 | *** | 34.1 | 2 | 0.000 |

## Appendix 2. Newspapers and news search keywords

| Newspaper | Country | Region | Circulation |
|---|---|---|---|
| The Australian | Australia | Oceania | 135 000 |
| The Globe and Mail | Canada | North America | 306 985 |
| Daily Telegraph | Great Britain | Europe | 874 000 |
| Times of India | India | Asia | 3 146 000 |
| USA Today | USA | National | 1 784 242 |
| Wall Street Journal | USA | National | 2 096 169 |
| New York Times | USA | National | 1 150 589 |
| Philadelphia Inquirer | USA | Northeast | 331 134 |
| The Boston Globe | USA | Northeast | 205 939 |
| Washington Post | USA | South | 507 465 |
| Dallas Morning News | USA | South | 409 642 |
| Chicago Tribune | USA | Midwest | 425 370 |
| Detroit Free Press | USA | Midwest | 234 579 |
| Denver Post | USA | West | 353 115 |
| San Jose Mercury | USA | West | 527 568 |
| Los Angeles Times | USA | West | 572 998 |

| *Search terms* | *News articles* |
|---|---|
| Computer break in | 24 |
| Computer firewall | 24 |
| Computer hacker | 194 |
| Computer identity theft | 83 |
| Computer malicious | 129 |
| Computer password | 107 |
| Computer security | 484 |
| Computer spam | 46 |
| Facebook hacker | 63 |
| Facebook password | 58 |
| Internet hacker | 171 |
| Internet identity theft | 68 |
| Internet malicious | 104 |
| Internet password | 27 |
| Internet security | 415 |
| Internet spam | 56 |
| Online firewall | 24 |
| Online hacker | 168 |
| Online identity theft | 101 |
| Online malicious | 104 |
| Online password | 109 |
| Online security | 431 |
| Online spam | 56 |
| Twitter hacker | 75 |
| Twitter password | 41 |

## Appendix 3. Websites and web search keywords

**Federal Government Agencies**

- Federal Bureau of Investigation (FBI)
- National Institute of Standards and Technology (NIST)
- US Computer Emergency Readiness Team (US-CERT)
- OnGuardOnline (Stop. Think. Connect. campaign)
- Federal Communications Commission (FCC)
- Federal Trade Commission (FTC)

**State Government Agencies**

- New York
- Arkansas
- North Carolina
- Colorado
- Michigan

| Search terms | Web pages |
|---|---|
| Account malware | 138 |
| Account phishing | 167 |
| Account security | 146 |
| Computer attacks | 122 |
| Computer authentication | 35 |
| Computer encryption | 90 |
| Computer malware | 140 |
| Computer phishing | 145 |
| Computer security | 165 |
| Cyber attacks | 44 |
| Cyber dns | 12 |
| Cyber malware | 98 |
| Cyber phishing | 109 |
| Cyber security | 167 |
| Data malware | 101 |
| Data phishing | 114 |
| Email attacks | 97 |
| Email malware | 140 |
| Email phishing | 144 |
| Flash malware | 36 |
| Flash phishing | 39 |
| Flash security | 20 |
| Identity malware | 124 |
| Identity phishing | 121 |
| Internet attacks | 75 |
| Internet malware | 129 |
| Internet phishing | 151 |
| Microsoft attacks | 24 |
| Microsoft malware | 33 |
| Microsoft phishing | 51 |
| Network attacks | 68 |
| Network malware | 92 |
| Network security | 96 |
| Online attacks | 76 |
| Online malware | 151 |
| Online phishing | 148 |
| Online security | 170 |
| Site malware | 132 |
| Site phishing | 139 |
| Software malware | 134 |
| Software phishing | 138 |
| Software security | 122 |
| Web malware | 103 |
| Web phishing | 138 |
| Web security | 116 |

**University IT Departments**

- University of California-Santa Barbara
- Fairfield University
- Life University
- University of Indianapolis
- Mississippi College
- East Central College
- Saint Augustines College
- Washington State Community College
- University of Wisconsin-La Crosse
- Stratford University

Companies

- *Operating Systems* (Mkt Share, 2012)
  - Microsoft (85%)
  - Apple (11%)
- *Social Network Sites* (# users, 2012)
  - Facebook (901 million)
  - Google+ (43 million)
- *Internet Service Providers* (Mkt Share, 2012)
  - AT&T (20%)
  - Verizon (12%)
  - Comcast (5%)
- *Antivirus Companies* (Mkt Share, 2012)
  - Avast (17.4%)
  - Symantec (10.3%)
- *Third-Party Software*
  - Adobe
  - Mozilla
- *Banks*
  - JP Morgan Chase
  - Bank of America

## Appendix 4. Example stories

### STORY460:

I was on the phone with my mom the other day and asked her about a strange email that she had sent me that was talking about working online and how I should apply. I almost clicked on the link but because I don't want to work this semester I decided not to. My mom said she was so glad that I didn't open it because apparently it was spam and was being sent to all of her contacts who notified her that this was going on even before I had. Thankfully, her computer was not affected by the email.

### STORY377:

My friend decided he wanted to watch some inappropriate videos and went to a shady site. He did not have a firewall or any sort of anti-virus so his computer got infected. His computer slowly got worse and worse until he couldn't handle it and took it to his parents. His parents did not know what to do and before they could figure it out, the computer died.

### STORY344:

I heard there was an email going around that looks like it comes from your bank. They ask you for your account and credit card information. Do NOT respond to it or click on the link. It is a scam and they are only looking for access to your account to steal your

information and your money. The bank already has your information so they have no need to ask for it. They will also never terminate your account for such a reason.

## Appendix 5. Example news articles

### NEWS236:

The nation's biggest banks and large technology companies like SAP rushed Tuesday to accept RSA Security's offer to replace their ubiquitous SecurID tokens as many computer security experts voiced frustration with the company.

The company's admission of the RSA tokens' vulnerability on Monday was a shock to many customers because it came so long after a hacking attack on RSA in March and one on Lockheed Martin last month. The concern of customers and consultants over the way RSA, a unit of the tech giant EMC, communicated also raises the possibility that many customers will seek alternative solutions to safeguard remote access to their computer networks.

Bank of America, JPMorgan Chase, Wells Fargo, and Citigroup said they planned to replace the tokens as soon as possible. The banks declined to say how many customers would be affected, although SAP said that most of its 50 000 employees used RSA's tokens and that it was seeking to replace them all.

Defense industry officials said Tuesday that concerns about the tokens had prompted some of the nation's largest military contractors to accelerate their plans to shift to computer smart cards and other emerging security technology.

The RSA tokens provide security by requiring users to enter a unique number generated by the token each time they connect to their networks.

Competitors eyeing the dominant market share of RSA are offering special deals like $5 rebates per token to customers that are considering a switch.

For now, however, the biggest worry for RSA is how to appease angry customers as well as mollify computer security consultants, who have been increasingly critical of how long it took for the company to acknowledge the severity of the problem.

Industry officials said that Lockheed, the nation's largest military contractor, made the security changes suggested by RSA after its attack in March. They included increased monitoring and addition of another password to its remote log-in process. Yet the hackers still got into Lockheed's network, prompting security experts to say that the tokens themselves needed to be reprogrammed.

Arthur W. Coviello Jr, RSA's executive chairman, made the offer in a letter posted on the company's website on Monday. He said RSA was expanding the offer to companies other than military contractors, particularly those focused on protecting intellectual property and their corporate networks. He also said it was suggesting that banks use two additional RSA services to avert fraud in authenticating computer log-ins.

Mr Coviello said in the letter that characteristics of the attack on RSA "indicated that the perpetrator's most likely motive" was to steal security information that could be used to obtain military secrets and intellectual property. He said that RSA had worked with military companies to replace their tokens "on an accelerated timetable."

Michael Gallant, an EMC spokesman, said, "We have not withheld any information that would adversely affect the security of our customers' systems."

"We provided very specific recommendations, we provided details of the attack, and we worked closely with customers to strengthen their overall security," Mr Gallant said.

The company's admissions were too little, too late, industry experts said.

"They got pushed really hard by some of their customers, particularly in the financial services sector," said Gary McGraw, chief technology officer for Cigital, a computer security consulting company based in Washington. "They came around, but they came around late."

Mr McGraw said that companies would be wise to replace RSA's tokens and that some companies—banks, in particular—had done so. Like many people, he criticized RSA for failing to disclose the potential danger of the problem to its customers.

Until Monday, RSA said publicly and privately in meetings with customers that replacements were unnecessary, he said. "They shared their party line that everything is fine – pay no attention to the explosion in the corner," Mr McGraw said.

Another security consultant, Alex Stamos, chief technology officer for iSEC Partners, said that many companies that use RSA tokens were irate about the hacking and RSA's response. He claimed that RSA misled customers about the potential problems after the initial hacking came to light. "Their whole excuse doesn't hold water," he said.

By minimizing the problem for six to seven weeks, Mr Stamos said that RSA made companies more vulnerable.

"There would have been huge benefit for RSA customers to know the truth," he said.

In the short term, customers are focused on getting new tokens but the overall outlook is cloudy.

"Companies are asking for the new tokens and looking long term to switching away from RSA," Mr Stamos said. "If you have 30,000 employees, switching to a new access solution is a yearlong process."

Avivah Litan, a longtime financial technology analyst for Gartner, estimated that it would cost banks just under $1 per customer to clean up the mess, even though RSA had agreed to supply new tokens. That would amount to as much as $95 million in customer service, mailing and other costs—a tiny fraction of the roughly $29 billion in profit the banking industry earned in the first quarter of this year.

As a result, most bankers see the recent breach as an annoyance, not a major security threat. Ms. Litan said that most of the biggest banks would step up other fraud protection measures, like monitoring their websites and customer accounts for suspicious behavior.

Moving to a new token provider would be costly because it would require them to redesign their online-banking applications as well as help customers—typically high-net-worth customers they do not want to alarm—make the shift to a new system.

Still, to increase security, Ms. Litan predicted that more banks would instead turn to new fraud prevention technologies that have been gaining adoption recently.

Such technologies help banks make sure that customers' PCs are malware free, send text messages or call customers to confirm transactions, and use analytics to look for unusual behavior that might point to fraud.

But the blow to RSA's reputation could hurt the company's ability to win new business, she said. While RSA was once the safe, conservative choice, "now when people talk about them, they will always be associated with this breach," Ms. Litan said.

Experts have speculated that the hackers obtained at least part of the RSA databases holding serial numbers and other critical data for the tens of millions of tokens. But to make use of the data stolen from RSA, security experts said, the hackers of Lockheed would also

have needed the passwords of one or more users on the company's network.

RSA has said that in its own breach, the hackers did this by sending "phishing" e-mails to small groups of employees, including one worker who opened an attachment that unleashed malicious software, enabling the hacker to obtain the worker's passwords.

Lockheed has said it would keep using the SecurID tokens and would replace 45 000 of them. L-3 Communications, a military contractor in New York, is also still using the tokens.

The military industry officials said that even before the breach at RSA, Northrop Grumman, another giant military contractor, had begun shifting from SecurID tokens to smart cards. The Pentagon also uses the smart cards, and other military contractors are accelerating plans to switch to them as well, the officials said.

Indeed, analysts say rivals like Vasco Data Security, Symantec, VeriSign, and dozens of small security vendors are circling. On Tuesday, PhoneFactor, which offers a phone-based password service to hundreds of companies, offered live Webcasts and a rebate to companies that wanted to switch.

"Since the Lockheed story, it's been crazier than ever," said Steve Dispensa, the chief technology officer of PhoneFactor.

## NEWS217:

The Pentagon, trying to create a formal strategy to deter cyberattacks on the USA, plans to issue a new strategy soon declaring that a computer attack from a foreign nation can be considered an act of war that may result in a military response.

Several administration officials, in comments over the past two years, have suggested publicly that any American president could consider a variety of responses—economic sanctions, retaliatory cyberattacks, or a military strike—if critical American computer systems were ever attacked.

The new military strategy, which emerged from several years of debate modeled on the 1950s effort in Washington to come up with a plan for deterring nuclear attacks, makes explicit that a cyberattack could be considered equivalent to a more traditional act of war. The Pentagon is declaring that any computer attack that threatens widespread civilian casualties—e.g., by cutting off power supplies or bringing down hospitals and emergency-responder networks—could be treated as an act of aggression.

In response to questions about the policy, first reported Tuesday in The Wall Street Journal, administration and military officials acknowledged that the new strategy was so deliberately ambiguous that it was not clear how much deterrent effect it might have. One administration official described it as "an element of a strategy," and added, "It will only work if we have many more credible elements."

The policy also says nothing about how the USA might respond to a cyberattack from a terrorist group or other nonstate actor. Nor does it establish a threshold for what level of cyberattack merits a military response, according to a military official.

In May 2009, four months after President Obama took office, the head of the US Strategic Command, Gen. Kevin P. Chilton, told reporters that in the event of a cyberattack "the law of armed conflict will apply," and warned that "I don't think you take anything off the table" in considering a response. "Why would we constrain ourselves?" he asked, according to an article about his comments that appeared in Stars and Stripes.

During the cold war, deterrence worked because there was little doubt the Pentagon could quickly determine where an attack was coming from—and could counterattack a specific missile site or city. In the case of a cyberattack, the origin of the attack is almost always unclear, as it was in 2010 when a sophisticated attack was made on Google and its computer servers. Eventually Google concluded that the attack came from China. But American officials never publicly identified the country where it originated, much less whether it was state sanctioned or the action of a group of hackers.

"One of the questions we have to ask is, How do we know we're at war?" one former Pentagon official said. "How do we know when it's a hacker and when it's the People's Liberation Army?"

A participant in the debate over the administration's broader cyberstrategy added, "Almost everything we learned about deterrence during the nuclear standoffs with the Soviets in the '60s, '70s and '80s doesn't apply."

White House officials, responding to the article that appeared in The Journal, argued that any consideration of using the military to respond to a cyberattack would constitute a "last resort," after other efforts to deter an attack failed.

They pointed to a new international cyberstrategy, released by the White House two weeks ago, that called for international cooperation on halting potential attacks, improving computer security, and, if necessary, neutralizing cyberattacks in the making. General Chilton and the vice chairman of the Joint Chiefs of Staff, Gen. James E. Cartwright, have long urged that the USA think broadly about other forms of deterrence, including threatening a country's economic well-being, or its reputation.

The Pentagon strategy is coming out at a moment when billions of dollars are up for grabs among federal agencies working on cyber-related issues, including the National Security Agency, the Central Intelligence Agency, and the Department of Homeland Security. Each has been told by the White House to come up with approaches that fit the international cyberstrategy that the White House published in May.

## NEWS395:

After oxygen, your wallet, and cell phone, nothing is more vital to the business traveler than wireless Internet. It is our connection to work, home, fantasy sports teams, and shopping. On the hotel, café, or convention center networks, we flip through our online tasks with nary a care. But a care would be a good idea.

Jason Glassberg, co-founder of Casaba Security, a Seattle-based technology security company, said the hazards associated with public Wi-Fi networks are so numerous that he does not log on to them; he connects to the Internet through his iPhone. When he must access the Internet on a public network, he does so through a virtual private network—VPN in industry speak—that allows him to encrypt his data through a personal server back home.

"A personal level of encryption definitely makes me feel safer," he said. "But I'm probably more paranoid than most."

Though Glassberg doesn't encourage everyone to be as cautious as he, he does say the average road warrior needs to pay closer attention to Internet habits.

Q. How safe are public wireless networks?

A. There are basically two kinds: unsecured and secured. An unsecured has no log-in, no password, and nothing is encrypted. Those are the most dangerous; if they're free for you, they're free for anybody, and anybody can be on them, looking for people doing online transactions. You should never enter bank account information on that. A secured network makes it harder, but it's not the biggest deterrent. It's another step someone would have to go through, so they'll probably go for one that doesn't have a password first.

Q. Would you personally enter banking information on a secured network?

A. It's a bit safer, but if I didn't have to do it, I wouldn't do it.

Q. Is Internet information theft usually a crime of opportunity?

A. It's the car-thief analogy: if someone's targeting your car, they'll find a way to get in. Similarly, if someone is targeting you or your business, they'll probably find a way to get in. But a lot of time, people are looking for people who let their guard down. You don't want to be the guy out there laying yourself bare.

Q. How easy is it to pick off information from someone on a public network?

A. Very easy. The largest theft of credit card information was by a guy sitting in a parking lot, picking up the information through an unsecured network. He was able to pick up passwords and start his hack. People with virtually no skill can collect the data.

Q. Do you need to be more cautious of a public network at, say, a chain hotel in a major city than a rural bed-and-breakfast?

A. Cybercrime is an equal-opportunity pain. It boils down to who's doing what, when, and where. In the middle of nowhere Iowa, maybe people are bored and pass the time this way. It's easy to do with tools that are very easy to acquire.

Tips from Jason Glassberg

*Be sure any sensitive information is sent on websites beginning with https, not just http. The "s" is proof of a security certificate.

*Be aware of the kind of network you're joining. A WEP network is least secure. WPA and WPA2 networks are more secure.

*Be sure file sharing and printer sharing are turned off on your laptop.

*Run up-to-date anti-virus software and a firewall on your computer.

*Do as little banking and make as few sensitive transactions as possible on public networks; do these instead on your phone, which is safer.

## Appendix 6. Example web pages

Only the textual content of the web pages was retained for analysis.

### CM35:

Enable or disable links and functionality in phishing email messages.

Phishing is the malicious practice of using email messages to lure you into disclosing personal information, such as your bank account number and account password. Often, phishing messages use untrustworthy links to fake websites that request your personal information. This information can be used by criminals to steal your identity, your money, or both. Learn more about phishing schemes.

Because it can be difficult to distinguish a phishing email message from a legitimate email message, the Outlook Junk Email Filter evaluates each incoming message to see whether it includes suspicious characteristics common to phishing scams. Such characteristics can include untrustworthy links, or content common to phishing messages, or the message was sent from a spoofed (fake) email address. Suspicious message detection is always turned on in Microsoft Outlook 2010, even if other junk email filtering is turned off.

What happens in Outlook 2010 with suspected phishing messages?

When a suspected phishing message arrives, it is processed as follows:

*If the Junk Email Filter doesn't consider a message to be spam but does consider it to be phishing, the message is left in the Inbox, but any links in the message are disabled and you can't use the Reply and Reply All commands. In addition, any attachments in the suspicious message are blocked.

*If the Junk Email Filter considers the message to be both spam and phishing, the message is automatically sent to the Junk E-mail folder. Any message sent to the Junk E-mail folder is saved in plain text format and all links are disabled. In addition, the Reply and Reply All commands are disabled and any attachments in the message are blocked.

*If the Junk Email Filter considers the message to be both spam and phishing, and the sender (someone@example.com) or domain (@example.com) is on your Safe Senders List, the message is left in the Inbox. However, the links and attachments in the message are disabled.

The InfoBar (InfoBar: Banner near the top of an open email message, appointment, contact, or task. Tells you if a message has been replied to or forwarded, along with the online status of a contact who is using Instant Messaging, and so on.) in the message describes the action taken on the message.

Move suspicious messages from the Junk E-mail folder.

You can move a message considered suspicious back to the Inbox. In the Reading Pane (Reading Pane: A window in Outlook where you can preview an item without opening it. To display the item in the Reading Pane, click the item.) or open message, click the InfoBar, and then click Move to Inbox.

InfoBar menu

*The original message format is restored but the links the message contains remain disabled. In addition, the Reply and Reply All functionality remains disabled and any attachments in the message remain blocked.

*If the Junk Email Filter considers the message to be both spam and phishing but you don't agree, open the Junk E-mail folder, right-click the message, and then click Add Sender to Safe Senders List. The message is moved to your Inbox. Disabled links remain disabled. The original message format is restored.

Important: After you add the sender or domain to your Safe Senders List, any new messages from that sender or domain are evaluated by the filter but aren't moved to the Junk E-mail folder. We recommend that your Safe Senders List not include banks, credit card companies, or e-commerce senders or domains, because these senders' addresses are the most frequently used by phishers.

Turn on disabled links

If you want to enable the links in a message, do the following:

1. In the Reading Pane or open message, click the InfoBar text at the top of the message.
2. Click Enable links and other functionality (not recommended).

Turn off automatic disabling of links

1. On the Home tab, in the Delete group, click Junk, and then click Junk E-mail options.
2. On the Options tab, clear the Disable links and other functionality in phishing messages (recommended) check box.

Note: If you later turn on this feature, links in previous messages that were evaluated as suspicious by the Junk Email Filter are disabled.

Turn off warnings about potentially spoofed email addresses

1. On the Home tab, in the Delete group, click Junk, and then click Junk E-mail options.

2. On the Options tab, clear the Warn me about suspicious domain names in e-mail addresses (recommended) check box.

### GFUC21:

Understanding Hidden Threats: Corrupted Software Files

Malicious code is not always hidden in web page scripts or unusual file formats. Attackers may corrupt types of files that you would recognize and typically consider safe, so you should take precautions when opening files from other people.

What types of files can attackers corrupt? An attacker may be able to insert malicious code into any file, including common file types that you would normally consider safe. These files may include documents created with word processing software, spreadsheets, or image files. After corrupting the file, an attacker may distribute it through email or post it to a website. Depending on the type of malicious code, you may infect your computer by just opening the file.

When corrupting files, attackers often take advantage of vulnerabilities that they discover in the software that is used to create or open the file. These vulnerabilities may allow attackers to insert and execute malicious scripts or code, and they are not always detected. Sometimes the vulnerability involves a combination of certain files (such as a particular piece of software running on a particular operating system) or only affects certain versions of a software program.

What problems can malicious files cause? There are various types of malicious code, including viruses, worms, and Trojan horses (see Why is Cyber Security a Problem? for more information). However, the range of consequences varies even within these categories. The malicious code may be designed to perform one or more functions, including

*interfering with your computer's ability to process information by consuming memory or bandwidth (causing your computer to become significantly slower or even "freeze")

*installing, altering, or deleting files on your computer

*giving the attacker access to your computer

*using your computer to attack other computers (see Understanding Denial-of-Service Attacks for more information)

How can you protect yourself?

*Use and maintain anti-virus software—Anti-virus software can often recognize and protect your computer against most known viruses, so you may be able to detect and remove the virus before it can do any damage (see Understanding Anti-Virus Software for more information). Because attackers are continually writing new viruses, it is important to keep your definitions up to date.

*Use caution with email attachments—Do not open email attachments that you were not expecting, especially if they are from people you do not know. If you decide to open an email attachment, scan it for viruses first (see Using Caution with Email Attachments for more information). Not only is it possible for attackers to "spoof" the source of an email message, but your legitimate contacts may unknowingly send you an infected file. If your email program automatically downloads and opens attachments, check your settings to see if you can disable this feature.

*Be wary of downloadable files on websites - Avoid downloading files from sites that you do not trust. If you are getting the files from a supposedly secure site, look for a website certificate (see Understanding Web Site Certificates for more information). If you do download a file from a website, consider saving it to your computer and manually scanning it for viruses before opening it.

*Keep software up to date—Install software patches so that attackers cannot take advantage of known problems or vulnerabilities (see Understanding Patches for more information). Many operating systems offer automatic updates. If this option is available, you should enable it.

*Take advantage of security settings—Check the security settings of your email client and your web browser (see Evaluating Your Web Browser's Security Settings for more information). Apply the highest level of security available that still gives you the functionality you need.

## References


1. Anderson R. Why cryptosystems fail. In: *CCS '93: Proceedings of the 1st ACM conference on Computer and communications security*. New York: ACM, 1993, 215–27.
2. Symantec Corporation. Internet security threat report. 2015. http://www.symantec.com/security_response/publications/threatreport.jsp (9 November 2015, last accessed date).
3. Wash R. Folk models of home computer security. In: *Symposium on Usable Privacy and Security (SOUPS)*. New York, NY: ACM, 2010.
4. Wash R, Rader E. Influencing mental models of security: a research agenda. In: *NSPW '11: Proceedings of the 2011 Workshop on New security paradigms workshop*. New York, NY: ACM, 2011.
5. Bandura A. Human agency in social cognitive theory. *Am Psychol* 1989;**44**:1175–84.
6. Baumeister RF, Zhang L, Vohs KD. Gossip as cultural learning. *Rev Gen Psychol* 2004;**8**:111–21.
7. Rader E, Wash R, Brooks B. Stories as informal lessons about security. In: *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. New York, NY: ACM, 2012.
8. Besnard D, Arief B. Computer security impaired by legitimate users. *Computers & Security* 2004;**24**:253–64.
9. Adams A, Sasse M. Users are not the enemy. *Commun ACM* 1999;**42**:46.
10. Kaemer S, Carayon P. Human errors and violations in computer and information security: The viewpoint of network administrators and security specialists. *Appl Ergonom* 2007;**38**:143–54.
11. Cranor LF. A framework for reasoning about the human in the loop. In: *Proceedings of the 1st Conference on Usability, Psychology, and Security (UPSec)*. Berkeley, CA: USENIX Association, **2008**.
12. Yee K-P. User interaction design for secure systems. In: *Proceedings of the International Conference on Information and Communications Security (ICICS)*. Springer, Lecture Notes in Computer Science 2513, 2002, 278–90.
13. von Ahn L, Blum M, Hopper NJ *et al*. CAPTCHA: using hard AI problems for security. In *Proceedings of the EUROCRYPT '03*. Springer, International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT 03) Lecture Notes in Computer Science 2656, 2003, 294–311.
14. Wash R, Rader E, Vaniea K *et al*. Out of the loop: how automated software updates cause unintended security consequences. In: *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. Berkeley, CA: USENIX Association, 2014, 89–104.
15. Zurko ME. User-centered security: Stepping up to the grand challenge. In: *21st Annual Computer Security Applications Conference (ACSAC'05)*. New York, NY: IEEE, 2005, 187–202.
16. Kirlappos I, Beautement A, Sasse MA. "comply or die" is dead: long live security-aware principal agents. In: Financial Cryptography and Data Security, *number 7862 in Lecture Notes in Computer Science*. Springer, 2013, 70–82.
17. Camp LJ. Mental models of privacy and security. *IEEE Technol Soc* 2009;**28**:37–46.
18. Asgharpour F, Liu D, Camp L. Mental models of computer security risks. In: *Workshop on the Economics of Information Security (WEIS)*, 2007.
19. Dourish P, Grinter RE, De La Flor JD *et al*. Security in the wild: user strategies for managing security as an everyday, practical problem. *Pers Ubiquit Comput* 2004;**8**:391–401.
20. Anderson CL, Agarwal, R. Practicing safe computing: a multimedia empirical examination of home computer user security behavioral intentions. *MIS Quart* 2010;**34**:613–43.

21. Prettyman SS, Furman S, Theofanos M *et al*. Privacy and Security in the Brave New World: The Use of Multiple Mental Models. In: *Human Aspects of Information Security, Privacy, and Trust*. Springer International Publishing, 2015, 260–70.
22. Furnell S, Moore L. Security literacy: the missing link in today's online society? *Comput Fraud Secur Bull* 2014;**2014**:12–18.
23. Ion I, Reeder R, Consolvo S. " . . . no one can hack my mind": comparing expert and non-expert security practices. In: *Symposium on Usable Privacy and Security (SOUPS)*. Berkeley, CA: USENIX Association, 2015, 327–46.
24. Kang R, Dabbish L, Fruchter N *et al*. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In: *Symposium on Usable Privacy and Security (SOUPS)*. Berkeley, CA: USENIX Association, 2015, 39–52.
25. Whitman ME. Enemy at the gate: threats to information security. *Commun ACM* 2003;**46**:91–95.
26. Karjalainen M, Siponen M. Toward a new meta-theory for designing information systems (is) security training approaches. *J Assoc Inf Syst* 2011;**12**:518–55.
27. Furman SM, Theofanos MF, Choong Y-Y *et al*. Basing cybersecurity training on user perceptions. *IEEE Secur Priv* 2012;**10**:40–49.
28. Marsick VJ, Watkins KE. Informal and incidental learning. *New Dir Adult Contin Educ* 2001;**2001**:25–34.
29. Marsick VJ, Volpe M. The nature and need for informal learning. *Adv Develop Hum Resour* 1999;**1**:1–9.
30. Reber AS. *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. Oxford: Oxford University Press, 1993.
31. Eraut M. Informal learning in the workplace. *Stud Contin Educ* 2004;**26**:247–73.
32. Quigley K, Burns C, Stallard K. 'Cyber Gurus': a rhetorical analysis of the language of cybersecurity specialists and the implications for security policy and critical infrastructure protection. *Gov Inf Q* 2015;**32**:108–17.
33. Bandura A. *Social Learning Theory*. Upper Saddle River, NJ: Prentice Hall, 1977.
34. Cialdini R. *Influence: The Psychology of Persuasion*, revised edn. New York, NY: Harper Business, 2006.
35. Goldstein NJ, Cialdini RB, Griskevicius V. A room with a viewpoint: using social norms to motivate environmental conservation in hotels. *J Consum Res* 2008;**35**:472–82.
36. Furnell SM, Bryant P, Phippen AD. Assessing the security perceptions of personal Internet users. *Comput Secur* 2007;**26**:410–17.
37. Nicolas-Rocca TS, Schooley BL, Spears JL. Exploring the effect of knowledge transfer practices on user compliance to is security practices. *Intl J Knowl Manage* 2014;**10**:62–78.
38. Posey C, Roberts TL, Lowry PB *et al*. Bridging the divide: a qualitative comparison of information security thought patterns between information security professionals and ordinary organizational insiders. *Inf Manage* 2014;**51**:551–67.
39. LaRose R, Rifon NJ, Enbody R. Promoting personal responsibility for internet safety. *Commun ACM* 2008;**51**:71–76.
40. James T, Nottingham Q, Kim BC. Determining the antecedents of digital security practices in the general public dimension. *Inf Technol Manage* 2013;**14**:69–89.
41. Das S, Kim TH-J, Dabbish LA *et al*. The effect of social influence on security sensitivity. In: *Symposium on Usable Privacy and Security (SOUPS)*, 2014, 143–57.
42. Romer D, Jamieson KH, Aday S. Television news and the cultivation of fear of crime. *J Commun* 2003;**53**:88–104.
43. Arendt F. Cultivation effects of a newspaper on reality estimates and explicit and implicit attitudes. *J Media Psychol Theories Methods Appl* 2010;**22**:147–59.
44. Mohr JW, Wagner-Pacifici R, Breiger RL *et al*. Graphing the grammar of motives in National Security Strategies: cultural interpretation, automated text analysis and the drama of global politics. *Poetics* 2013;**41**:670–700.
45. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003;**3**:933–1022.
46. Miller IM. Rebellion, crime and violence in Qing China, 1722–1911: a topic modeling approach. *Poetics* 2013;**41**:626–49.
47. Jockers ML, Mimno D. Significant themes in 19th-century literature. *Poetics* 2013;**41**:750–69.
48. Grimmer J. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Polit Anal* 2010;**18**:1–35.
49. Quinn KM, Monroe BL, Colaresi M *et al*. How to analyze political attention with minimal assumptions and costs. *Am J Polit Sci* 2010;**54**:209–28.
50. Bonilla T, Grimmer J. Elevated threat levels and decreased expectations: how democracy handles terrorist threats. *Poetics* 2013;**41**:650–69.
51. Jurowetzki R, Hain DS. Mapping the (R-)Evolution of technological fields - a semantic network approach. In: *SocInfo*. Springer International Publishing, 2014, 359–83.
52. Mohr JW, Bogdanov P. Introduction—topic models: what they are and why they matter. *Poetics* 2013;**41**:545–69.
53. Blei DM. Probabilistic topic models. *Commun ACM* 2012;**55**:77–78.
54. McCallum AK. Mallet: a machine learning for language toolkit. 2002. http://mallet.cs.umass.edu (9 November 2015, date last accessed).
55. Graham S, Milligan I. Review of MALLET, produced by Andrew Kachites McCallum. *J Digi Human* 2012;**2**:73–76.
56. Blythe M, Petrie H, Clark JA. F for fake: four studies on how we fall for phish. In: *Proceedings of the Conference on Human Factors in Computing (CHI) '11*, New York, NY: ACM, 2011, 3469–78.
57. Dhamija R, Tygar JD, Hearst M. Why phishing works. In: *CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, 2006, 581–90.
58. Schechter SE, Dhamija R, Ozment A *et al*. The emperor's new security indicators. In: *SP '07: Proceedings of the 2007 IEEE Symposium on Security and Privacy*. New York, NY: IEEE Computer Society, 2007, 51–65.
59. Symantec Corporation. State of privacy report. 2015. http://www.symantec.com/content/en/us/about/presskits/b-state-of-privacy-report-2015.pdf (9 November 2015, date last accessed).
60. Campbell K, Gordon LA, Loeb MP *et al*. The economic cost of publicly announced information security breaches: empirical evidence from the stock market. *J Comput Secur* 2003;**11**:431–48.
61. Whitten A, Tygar JD. Why Johnny can't encrypt: a usability evaluation of pgp 5.0. In: *Proceedings of the USENIX Security Symposium*. Berkeley, CA: USENIX Association, 1999.
62. Shay R, Komanduri S, Kelley PG *et al*. Encountering stronger password requirements: user attitudes and behaviors. In: *Symposium on Usable Privacy and Security (SOUPS)*. New York, NY: ACM, 2010, 2.
63. Langner R. Stuxnet: dissecting a cyberwarfare weapon. *Secur Priv IEEE* 2011;**9**:49–51.
64. Shillair R, Cotten SR, Tsai H-YS *et al*. Online safety begins with you and me: Convincing Internet users to protect themselves. *Comput Hum Behav* 2015;**48**:199–207.
65. Anderson R, Barton C, Böhme R *et al*. Measuring the cost of cybercrime. In: *The Economics of Information Security and Privacy*. Berlin, Heidelberg: Springer, 2013, 265–300.
66. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: *International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA: AAAI, 2009.
67. Bender J, Davenport L, Drager M *et al*. *Reporting for the Media*, 10th edn. Oxford: Oxford University Press, 2011.
68. Gelman SA, Legare CH. Concepts and folk theories. *Ann Rev Anthropol* 2011;**40**:379–398.