A Data-Driven Reflection on 36 Years of Security and Privacy Research

Aniqua Baset and Tamara Denning
School of Computing, University of Utah
{aniqua, tdenning}@cs.utah.edu

Abstract

Meta-research—research about research—allows us, as a community, to examine trends in our research and make informed decisions regarding the course of our future research activities. Additionally, overviews of past research are particularly useful for researchers or conferences new to the field. In this work we use topic modeling to identify topics within the field of security and privacy research using the publications of the IEEE Symposium on Security & Privacy (1980-2015), the ACM Conference on Computer and Communications Security (1993-2015), the USENIX Security Symposium (1993-2015), and the Network and Distributed System Security Symposium (1997-2015). We analyze and present data via the perspective of topics trends and authorship. We believe our work serves to contextualize the academic field of computer security and privacy research via one of the first data-driven analyses. An interactive visualization of the topics and corresponding publications is available at https://secprivmeta.net.

1 Introduction

As computing systems are pervading every aspect of our life, from the mundane (e.g., electronic payments) to the futuristic (e.g., IoT, virtual reality), security and privacy concerns surrounding these technologies are growing in number and seriousness. As a result, we are seeing security and privacy turning into an essential component in the design of any computer system or application. With this paradigm shift, the research field of security and privacy is experiencing an upsurge of interest from a broader spectrum of researchers and professionals. However, the lack of a structured, data-driven analysis of the current and past research trends of this field can obstruct the big picture view of the security and privacy landscape for the researchers or professionals outside or new to the field. To that end, in this paper, we present our meta-research—research about research—and discuss the trends in past 36 years of security and privacy research (from 1980–2015). We base our analysis on the publications from four of the top-tier conferences in the field—the IEEE Symposium on Security & Privacy (S&P), the ACM Conference on Computer and Communications Security (CCS), the USENIX Security Symposium (USENIX), and the Network and Distributed System Security Symposium (NDSS). We apply the technique of topic modeling to infer and categorize the publications into different security and privacy topics (e.g., browser security, secure computation, anonymity). We analyze trends over time: recently prominent topics (e.g., mobile apps and browser security), diminishing topics, and topics (e.g., access control) that have been steady over time. We next analyze the trends in authorship such as author retention, author accumulation over time, and inter-topic author movement. We also investigate intra-topic entropy to determine which topics have a particularly small (e.g., Java security) or large (e.g., browser security) authorship pool. We also analyze trends in topics with respect to academics versus non-academics authors.

While some of the results in this paper may be folk knowledge in the community, we pursue a structured method for generating a list of research topics and analyzing those. We believe that this is a first step towards building a richer body of introspective scholarship that will aid in understanding the landscape of security and privacy research for a broader spectrum of audience.

2 Identifying research topics

The first and foremost task needed for our analysis is to identify the research topics within the security and privacy field since not all conferences that we study categorize their publications: for example, USENIX does not. Additionally, CCS uses different categorizations than S&P, which uses keywords. Finally, we want to explore categorizations driven directly from the text of the publications. Therefore, to get a data-driven, cohesive categorization across different conference publications, we use generative topic modeling to extract a list of topics from a textual corpus—the full publication contents of S&P, CCS, USENIX, and NDSS. Specifically, we choose to use Latent Dirichlet allocation (LDA) [8], which is a well-

accepted and widely-used technique in the machine learning community. To provide context for our work, we first provide a brief overview of LDA in Section 2.1. Next, in Section 2.2, we describe our methodologies of all of the steps involved in inferring the security and privacy research topics—from collecting data to getting a refined, labeled list of topics.

2.1 Background: LDA

Conceptually, LDA considers documents to be equivalent to bags-of-words. It then assumes that these bags were (hypothetically) generated by one or more topics that exist latent within the corpus. From the input documents (bags) and a topic number, LDA tries to find out the latent distribution of topics over words—and hence, the distribution of a document over topics. As output, LDA produces:

- a list of word-sets that represent different topics, along with the weight of each word in the set; and
- topic distribution of each document in the corpus (i.e., a list of all topics with their associated topic portion—e.g., 60% in Topic A, 15% in Topic B, etc.).

While LDA—and topic modeling in general—is a useful technique to infer topics from documents, it is hard to evaluate whether or not a model is high-quality. There is no clear-cut way to examine whether the generated topics are meaningful and coherent, or to determine whether the assignments of topics to documents are appropriate. To overcome some of these problems, several metrics have been proposed to measure the quality of a topic or an overall model [30]. However, a *high-scoring* topic is not always a *high-quality* topic for people [10]. For the latter it is necessary to incorporate human perception in the process of topic modeling [9]. Some recent works on topic modeling have been focused on designing interactive topic modeling that can adapt to human input; however, this line of work is still in progress. In this work, we focus on using regular LDA and add post-processing steps to sanitize the resulting topic model. It should be noted that the manual tweaking of topic models is not uncommon in topic modeling works. For example, Hall et al. used hand-seeded words to improve topic coverage when topics were otherwise missing from the model [13].

2.2 Our methodologies

2.2.1 Data collection

We performed our study on publications from S&P (1980-2015), CCS (1993-2015), USENIX (1993-2015), and NDSS(1997-2015). The distribution of the publications is presented in Table 1. For each venue and conference year, we collected: (a) the full content of all publications; (b) all publication titles, authors, and affiliations; and (c) corresponding session names for all publications*.

Table 1: Publication distribution at different venues

Venue	Years	# of papers
CCS	1993–1994, 1996–2015	1066
NDSS	1997–2015	456
S&P	1980–2015	932
USENIX	1993, 1995-96, 1998–2015	608
	Total:	3062

2.2.2 Pre-Processing

Pre-processing the input documents is a crucial step in the modeling process since the basis of topic modeling is the words within a document and their frequency. We therefore take some time to indicate the pre-processing that we performed on the publications.

Step 1. Most of the publications were available in PDF format, although a few were only available in PostScript, HTML, or text formats. To get the corresponding text version of the PDF, PostScript, and HTML publications we used the PDFBox [1], ps2ascii [4], and html2text [3] tools, respectively. From full text of the publications we then extracted the main body of each publication, leaving out the title, author information, acknowledgement, and references.

Step 2. For some PDF formats, the text conversions are very noisy. We noticed the following problems in the the converted text of some of the PDF formats:

1. ligatures like “fi”, “fl”, “ffi”, or “ffl” had been converted to ‘.’, ‘-’, or a space;
2. homo-glyphs had been converted to their similar looking counterpart, e.g. “rn” → “m”, “cl” → “d”
3. words had been fragmented: e.g., “determination” → “deter mination”
4. words had been coupled together: e.g., “this publication presents” → “thispublicationpresents”

We tackled these problems as follows. For all non-dictionary words, we went through the possible homoglyphs and checked whether replacing a glyph with its counterpart made the word a dictionary word (2). We also ran a spell checker (1, 2, 3, 4). We manually verified any changes before committing them. This is important in cases such as system names, which are sometimes treated as misspelled words by the spell checker. For words that were fragmented (3), we fixed them by looking for sequences of non-dictionary words such that merging them results in a dictionary word.

Step 3. In order to match different forms of words we converted them to a more standardized form using lemmatization.

*Ethics. For all publishers—Internet Society, USENIX, ACM, and IEEE—we examined and complied with both the Terms of Service and the robots.txt regarding rules for access, sharing, and crawling. We contacted IEEE and ACM in order to gain permission to conduct this study using the publications available to our institution; Internet Society and USENIX did not require such special permissions.

For example, “attacks", “attacked", and “attacking" all correspond to a single form: “attack". Lemmatization is a standard process in linguistics for converting different inflected forms of a word to a single version [19]. We used the Stanford CoreNLP tool for the lemmatization [2].

Topic modeling does not take into account multi-word phrases. So, we decided to convert technical phrases (e.g., ‘man in the middle’) to single hyphenated words (e.g., ‘man-in-the-middle’). The problem was identifying such phrases. We noticed that most common phrases like this were available at least once in their hyphenated forms in the corpus. Therefore, we examined all hyphenated words for technically meaningful word groups and converted their non-hyphenated versions to a canonical hyphenated form[†]. We also converted phrases that ended with ‘-based’, ‘-system’, ‘-related’ or ‘-dependent’ to the base phrase[‡].

To handle the acronyms, we created a list from the publication corpus by searching for phrases that matched forms similar to ‘Axxx Bxxx Cxxx (ABC)’. After manually verifying the acronym-full-form pairs, we converted the acronyms to their full forms and represented them as hyphenated phrases (see above). To handle acronyms with multiple (overloaded) meanings, we only converted an acronym to its full form if the full form appeared in the same publication. For some common acronyms like SSL and HTTP, we did not necessarily expect that people would include the full form somewhere in the publication; so, for common acronyms, we chose to use the short form as default and converted the full forms to acronyms. Interested readers can check the acronym list online at `http://secprivmeta.net`.

Finally, we removed all numbers[§] and symbols from words with the exception of hyphens (‘-’), dots (‘.’), and slashes (‘/’). We also converted all non-acronym text to lower case.

Step 4. We created a stopword list: the list of words to be ignored during topic modeling. Using a stopword list reduces the chance of getting a “junk” topic. We included words in the list that had low IDF (Inverse Document Frequency) values, the most common English words [5], and words that appeared in only one document (in order to avoid algorithm/system names appearing as topic words).

2.2.3 Topic Modeling

We performed LDA topic modeling on the text from a total of 3019 publications. Among the 3062 publications we collected, we discarded 39 of them for pdf-to-text conversion problems and 3 others because their full texts were not available online.

We used the MALLET tool to perform the topic modeling [21]. MALLET optimizes the LDA hyperparameters over multiple iterations. We generated different models by varying the number of topics, the number of iterations, and the starting hyperparameter values. To evaluate the quality of the generated models, we used averaged topic Pointwise Mutual Information (PMI). As defined by Newman et al. [26], PMI of a topic t is as follows:

$$PMI(t, w_{1..N}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} log(\frac{D(w_i, w_j)}{D(w_i) + D(w_j)})$$

where $w_{1..N}$ are the top N words of a topic t, $D(w_i)$ and $D(w_j)$ are the numbers of documents where word w_i and w_j appeared respectively, and $D(w_i, w_j)$ is the number of documents where both words w_i and w_j appeared [¶].

As described in Section 2.1, there is no commonly accepted way to evaluate a model and topic quality metrics do not always match with human judgment. Therefore, we manually examined the top 5 highest-scoring models according to the average PMI scores. There is no one perfect model: a human might find multiple topics in different high-scoring models to be good, but those topics might not all appear in a single model. We manually selected one of the 5 top-scoring models and performed post-processing in order to refine its quality. The model we finally chose had a total of 118 topics; after post-processing and excluding some method-based topics, we had a total of 95 topics.

2.2.4 Post-Processing

In this section we explain how we refined the topics in our model. There were some “mixed” topics where two completely different concepts mixed together because of some common words they share (which is not uncommon to experience in topic modeling [9]). For example, we had a mixed topic that had both garbled circuit and integrated circuit publications. These two different kinds of publications were (logically) in the same topic because they share words like ‘circuit,’ ‘gate,’ and ‘bit.’ We used the following process to fix such mixed topics:

1. For each pair of publications in a topic, we found the correlation between the publications by computing the Kullback-Leibler divergence of their topic distributions [14].
2. We created a graph where vertices represent publications in a topic and edge weight between two vertices is the correlation value between the corresponding publications.
3. We ran a graph modularity algorithm to identify different sub-communities in the graph.
4. We analyzed the publications in different sub-communities to check if they are valid sub-community.

[†]For example, “webapplication” and “web application” were both converted to “web-application”

[‡]For example, “context-based” would become “context.”

[§]One exception is numbers starting with “802."

[¶]The proponents of PMI used external datasets to compute co-occurrences whreas we chose to use our publication corpus.

The intuition behind this process is that if a topic has publications from different domains, the topic distribution of a publication from one domain will be different from the topic distribution of a publication from a different domain, while the distributions of publications from the same domain will be close to each other. Therefore, we would be able to detect different sub-domains via sub-communities in the graph.

When we found sub-communities in a topic that represented sub-domains, we either created a new topic to move the publications from one sub-domain or moved them to an already existing topic, if appropriate. For example, for the garbled circuit and integrated circuit topic we moved the publications in the integrated circuits publications to an already existing hardware topic.

We also merged some topics. For example, we had two topics that were related to mobile apps. For our trend analysis and other computations, we consider these two topics as one.

2.2.5 Topic Labeling

We labelled a topic according to the top words in that topic, the top publications that appear in the topic, their keywords or CCS index (if available), and session name (if available).

2.2.6 Assigning Documents

As mentioned in Section 2.1, each document has a percentage for every topic in the model. We applied thresholds in order to have "yes/no" assignments of documents to topics for our trend analysis. For each topic, we computed the $mean + 2 * standard\ deviation$ value of the topic percentage of all documents (for that topic) and used that as the threshold.

3 Topics in security and privacy research

We now describe the topics themselves from our refined topic model. All of the topics are listed in Table 2[‖].

In the table, timeline for each topic appears to the right of the topic title which shows the publication count for the topic for each year (1980–2015). All timelines use the same scale and the highest bar corresponds to 20 publications (though Mobile Apps has more than 20 publications in its peak year, we use a general threshold of maximum 20 publications in order to keep the bars of small number of publications legible).

In the table, the topics are shown grouped by categories. Throughout this paper categories are referenced in small capitals. The categories (and their orderings) are not meant to be definitive; rather, they were formed via researcher inspection and are meant to aid in viewing and discussing the data.

Several topics do not fit well into the defined categories (e.g., Electronic Voting, E-commerce) and are thus placed in a MISCELLANEOUS category. The topics under METHODS are a slightly different kind of topic: they represent the techniques that were used in publications rather than the research topics or domains of the publications. For example, a publication might belong to User Study topic if it reports on a user study or to Machine Learning if machine learning was used in the course of the study. Four method topics not listed in the table are Probabilistic Methods, Performance Evaluation, Graphs, Hash Trees, Vectors and Matrices, and Logic. We excluded these because they are more general ideas that do not represent the main subject of the publication and because they have remained in publications at a fairly constant rate throughout the conferences' history (in contrast to Machine Learning or User Study, which are topics that show definite changes over time). We included System Calls and String Matching and Regular Expressions because these topics have publications that strongly identify with the topics and the publications do not belong to other topics.

As topics from similar domains share common words, some publications from one topic might instead be found in a highly related topic. For example, some public-key cryptography publications can be found under the Cryptographic Protocols or Encryption topics instead of the Public-Key Cryptography topic. Topics under FORMALISM category are another such example.

There are some research themes that did not appear as one single topic in our topic model, despite the fact that researchers might think to create such a topic. Instead, the publications related to those themes ended up in different topics. We posit that this is possibly because the words used in the publications from these domains are not cohesive enough in our corpus to form their own topics. For example, there is no one "Privacy" topic that includes all sorts of privacy publications. However, there is a Data Privacy topic. This is quite an old topic that helped to have enough publications to form a cohesive group of words in the publications to constitute a topic. Newer privacy publications, that are different than data privacy, ended up in different but relevant topics. For example, the publication [24] that describes privacy preserving payment protocol is under the E-commerce topic.

4 Investigating the inferred topics

We now delve into analyzing the inferred topics and corresponding publications to uncover trends over time, trends in authorship, trends in topics with respect to academics versus non-academics authors, etc.

How has the appearance of topics changed over time? Are there some categories that are/were more prominent than others?

Fig. 1 gives an overall idea of trends in different topic categories. We provide the timeline with respect to categories instead of the individual topics in order to see what happened to a group of topics rather than trends in individual ones.

[‖] We encourage readers to check our interactive visualization at http://secprivmeta.net for better understanding of the topics.

Table 2: 95 topics from the topic model, grouped into related categories. Categories are approximately ordered by age. Timelines for each topic show the paper count for each year (1980-–2015). All timelines use the same scale. The highest bar correspond to 20 papers. (Mobile Apps actually peaked at 28 papers.)

Crypto

Cryptographic protocols 

Key distribution/management

Group communication 

Public-key cryptography 

Digital signature 

Network authentication 

Encryption 

Crypto & number theory 

Random numbers

Trust

Trusted computing/system

Trust management

Software & trust 

Formalism

Formal methods 

Formal methods & verification 

Formal specification & verification 

Security labeling

Security policies 

System

File & file system security 

Kernels 

Compartmentalization

Storage security 

Hardware

RFIDs & ICs 

Low level

Physical properties

Malicious hardware

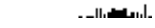
Embedded & hardware security

Networks

Attacks, defenses & detections 

Network design

Perimeter controls 

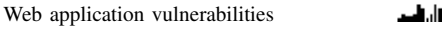
Traffic analysis: attacks & defences 

Routing

Web

JavaScript security 

Browser security 

Web application vulnerabilities 

DOM & documents 

Online services 

Auth

Access control 

Passwords 

CAPTCHA 

Computation

Secure (multiparty) computation 

Verifiable computation & zero knowledge proofs

Data

Data privacy

Databases

Genomic privacy

Malware

Malware

Viruses & worms: propogation & scanning

Bots & Botnet

Intrusion/anomaly detection

Programs

Static & dynamic analysis

Binary code analysis

Program exploitations: attacks & defenses

Vulnerabilities: exploits, disclosure & patches

(De)obfuscation & decompilation

Memory exploits & defenses

Control flow

Information flow

Java security

Information leakage

Covert channel

Side-channel attack

Memory disclosure attacks & defenses

Internet

Domain Name System (DNS)

SSL/TLS

TCP/IP

Mobile

Mobile app

Mobile devices

Mobile network

Location privacy/tracking

Crime & fraud

Online crime

Dark web

Spam, scam & fraud

Online advertising

Anonimity & Censorship

Censorship

Tor

Anonymity

Peer-to-peer communications

Social network & (de)anonymization

Virtual

Virtual machines & virtualization

Cloud

Client-server accountability

Miscellaneous

Fingerprints & fingerprinting

Real-world sensing

Electronic voting

Cards & tokens

(User) interfaces

Encoding/decoding

Automated analysis: protocols & files

Game & game theory

Wireless security

Institutional security

E-commerce

Bitcoin & crypto-currency

Methods

Machine learning

User study

String matching & regular expressions

System calls

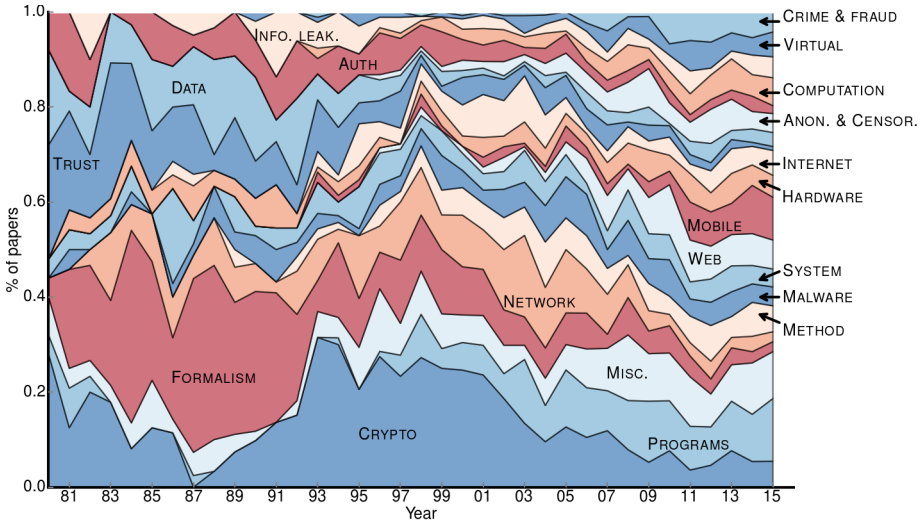


Figure 1: The prevalence of publications in a given category over time.

For example, it is more interesting to see what happened to the CRYPTO category as a whole than the different trends in Crypto & Number Theory and Encryption. For each category and year we present the number of publications from that category over the total publication counts for that year. It can be seen from the timeline that FORMALISM topics were the most prominent ones during 1980s (proportionately). Around 1993 their proportion started to decrease, while CRYPTO topics started to become prominent. From the early 2000s the field became more diverse; publications are distributed over a larger number of categories rather than one or two having dominance as in the 1980s and 1990s.

What are the topics that are strongly emerging or that have experienced a recent surge of publications?

We next identify the topics in the field that have experienced a recent upsurge in publications. We calculated 2-year period moving averages for each of the topic timelines. We then considered topics with strong upward trends in their moving averages to be prominent. These topics are (in the order of their total publication count in last 3 years): Mobile Apps, Verifiable Computation & ZK Proofs, Machine Learning, Program Exploitations, Side-Channel Attacks, SSL/TLS, Binary Analysis, Control Flow, Memory exploits & defenses, Bitcoin/Crypto-currency, and VMs & virtualization. The JavaScript Security and Browser Security topics did not make into this list. However, if we consider them together, they qualify as prominent. Another interesting topic is Bitcoin/Crypto-Currency which is showing a steep trend in the last two years (2014-2015). It is much newer than the other topics and therefore has fewer publications, so did not make in the list.

Are there any topics that are dying out, or do all topics continue to have publications?

While new topics regularly arose—adding to the list of active topics—no topics particularly died out and ceased to appear. Two exceptions to this are Databases and Java Security, which have not appeared in any publication since 2010. However, this does not necessarily mean that databases ceased to be a subject for research. Older publications on this theme appeared in Databases, while newer publications on this theme most likely appear in Data Privacy, in which common themes are re-identification, preserving privacy during database queries, and similar.

Some other topics that appeared in only 2 publications since 2010 are Group Communication, Institutional Security, and Software & Trust. Software & Trust is a topic that predominantly appears in the earlier years of the research and is one that primarily seems based on characteristic language usage of the time rather than a strongly cohesive topic theme.

Are there any topics that have been "slow and steady" over the years, or do all topics experience fluctuation?

Most of the topics have some inactive years. We present the exceptions, which have continuously appeared each year since their start in Table 3. There are also a handful of steady topics; that is, topics that tended to have 1–2 publications each year (with a maximum of 1–2 year gaps) since the 1980s. These topics are: Access Control, Intrusion/Anomaly Detection, Files and File Systems, Cards and Tokens, Security Policies, Security Labeling, and Covert Channels. Another topic that has kept showing up since the 80s is Network Design; however, the context of the publications of this topic has changed over time. While late 90s publications discussed ATM networks, the recent ones are mostly focused on Software Defined Networks.

Do authors tend to publish continuously, or do they come and go?

Table 3: Steady topics since start

Topic	Start
Network attacks, defenses, & detection	1997
(De)obfuscation & decompilation	2003
Social networks & (de)anonymization	2005
Web application vulnerabilities	2005
Bots and Botnets	2006
Mobile apps	2009

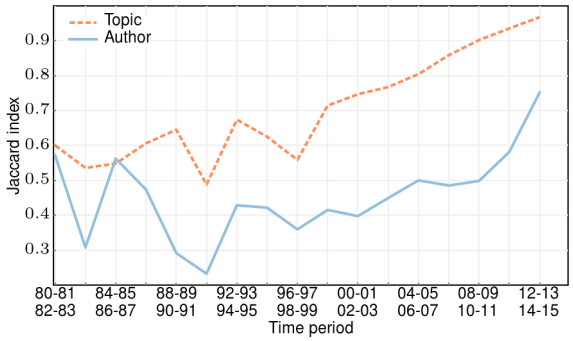


Figure 2: Flux in the appearance of particular authors and topics over time, as given by the Jaccard index. A higher value means more overlap between consecutive time periods.

We analyze how the sets of authors change over time. For this, we calculate the Jaccard index of author sets for contiguous periods, $J(A_1,A_2) = (|A_1 \cap A_2|)/(|A_1| + |A_2| - |A_1 \cap A_2|)$; where, A_1 and A_2 are the sets of authors who published in period p_1 and p_2, respectively. Intuitively, this gives us the proportion of common authors in contiguous periods. We use a two year time period; that is, we calculate the Jaccard index using the set of authors who published during 1980–1981 and the set of authors who published during 1982–1983 (next for 1982–1983 and 1984–1985, and so on). When generating these sets we only considered authors who have published at least 5 publications. We use this same concept—a retention metric via the Jaccard index—to calculate the proportion of similar topics in contiguous periods. Both the author and topic trends are presented in Fig. 2, which shows the flux of the authorship and topic pools from year to year.

From approximately 1984–1991 topics showed an upward trend, indicating an increased overlap of topics in consecutive periods. In contrast, during the same time period there was a steep decline in authorship overlap, suggesting large churn in the authorship pool. This suggests that new authors joined the field while authors in general began publish more on established topics.

Author retention was the lowest for 1990–1991, which also aligns with the lowest topic overlap. This might be taken to suggest that new authors joined the field and brought in new topics; however, as we note in previous discussion, this was primarily not the case.

After approximately 1991, both authors and topics show overall upward trends, indicating decreasing flux. It is apparent that in recent years both topics and authors tend to appear more continuously.

Table 4: Topics introduced by new authors

Topic	Start	Topic	Start
Real-world sensing	1985	Java security	1996
Viruses & worms	1988	Vulnerabilities	1998
Machine learning	1992	JavaScript security	1998
Verifiable comp. & ZK proofs	1993	Electronic voting	1999
Random numbers	1995	P2P comm.	2002
Passwords	1995	Memory disclosure	2003
Anonymity	1996	Games & game theory	2003

Table 5: Diverse and tight topics in terms of authorship

Diverse	Tight
HW: low level	Institutional security
Security policies	Malicious HWs
Browser	Game & game theory
Verifiable comp. & ZK proofs	Java
Machine learning	Genomics
TCP/IP	CAPTCHA
Static & dynamic analysis	Online crime
Data privacy	Cloud
Crypto protocols	Memory disclosure
Mobile app	Databases
Mobile devices	Tor
File & file systems	P2P communication
Side-channels	Bitcoin/crypto-currency
(User) interfaces	Automated analysis
Client-server accountability	HW: RFIDs & ICs

Do new authors introduce new topics to the conferences or are new topics begun by existing authors?

We next examine whether new topics are often introduced by new authors or whether authors publishing from existing topics begin the new topics. We find that most of the topics were started by authors from existing topics, *except* for the topics listed in Table 4, which were introduced by authors who had not previously published in the conferences we studied.

Which topics have a small core of authors ("tight")? Which have a broad variety of authors ("diverse")?

We measure the diversity of authorship in a topic using Shannon entropy, $E_t = -\sum_{i=1}^{A} p_i * log(p_i)$; where p_i is the probability of author i to publish in topic t and A is the total number of authors. Higher entropy indicates more diversity in the authorship pool. The 15 diverse topics (in decreasing entropy order) and the 15 tight topics (in increasing entropy order) are presented in Table 5. Hardware: Low-Level has the most diverse authorship. We posit that this is because this topic is applied to different domains, thereby diversifying authorship. The next most diverse topic is Security Policies, which has appeared almost every year since the beginning of the corpus; we posit that this has helped the topic to increase its authorship pool. In fact, most of the topics in the diverse end have a large number of publications. Notably, some of the topics with low entropy also have high publication counts: Databases and FORMALISM topics. While Games & Game Theory and Malicious Hardware are close-knit in terms of

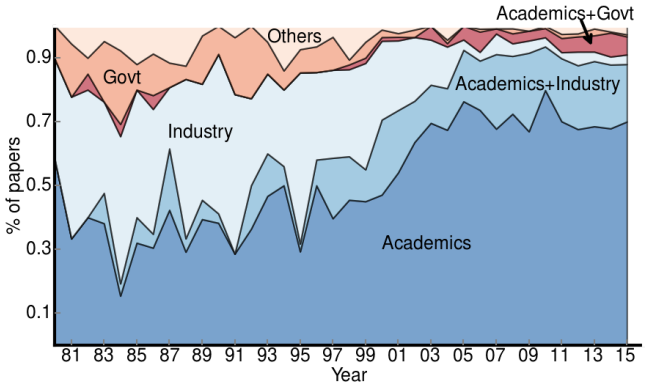


Figure 3: Author affiliations over the years, given as the proportion of that affiliation type for the given year.

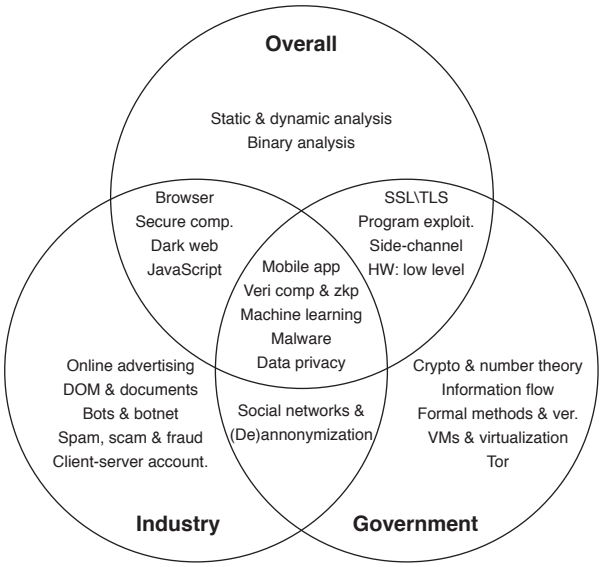


Figure 4: Comparison of overall top 15 topics in past 5 years of Industry (and their collaborations) with that of Government (and their collaborations)

authorship, these topics have fewer publications than many of the other topics, so it is not surprising that the authorship is not diverse.

How has industry and government participation in research changed over time?

We now examine the trends in authorship affiliations. We divide authors' affiliations into 5 categories: Academics, Industry, Government, Non-Profit/Research Institutes, and Independent Researchers.

Fig. 3 shows the proportion of each category of affiliation in publications for each year. In the figure, 'Academics', 'Industry', 'Academic + Industry', 'Govt', and 'Academics + Govt' represent the publications from the groups as their names suggest. However, these categories in the figure also include all collaborations with Non-Profit/Research Institutes or Independent Researchers. The 'Other' category in the figure contains publications from all other combinations: for example, publications with authors from both Industry and Government. It is evident from the figure that a smaller percentage of publications are being published from just Industry or just Government in recent years. Instead, they are collaborating more with Academics than they did in the 1980s and early 1990s. Around 2001 the Academics-only publications has started to dominate and the proportion has been steadily rising since then.

Do industry or government entities publish on the same things that academics do or do they focus on different topics? Has this changed over time or held steady?

We next analyze the similarity between the overall topic distributions versus the topic distributions of Industry (and their collaborations) and Government (and their collaborations) using Jensen–Shannon divergence [11]. We find that Industry publication topics are closely aligned to the overall trend of topic publications than that of Government publications.

To provide more insight on the different interests expressed by Industry (and their collaborations) and Government (and their collaborations), we contrast their recent top 15 topics with the overall top topics (2010–2015) in Fig. 4.

Are there any topics for which the non-academics are most interested in than the academics?

For most of the topics, the number of publications from Industry and Government (including collaborations with Academics) are fewer than the Academics (including their non-Industry, non-Government collaborations). However, the exceptional topics where Industry and Government publications outnumber the Academics publications are: Formal specification & verification, Network authentication, Formal methods, Software & trust, Formal methods & verification, Databases, Security labeling, and Crypto & number theory.

We find that among these topics Software & Trust, Formal Specification & Verification, Formal Methods, and Databases in fact have more publications from *just* non-Academics than from Academics or Academic collaborations. All the topics in the list above are some of the oldest topics. Given the dominance of non-Academic publications during the 1980s and 1990s, it is not too surprising to see that these are the topics with a prominence of non-Academics authors.

5 Related work

While the security and privacy research field lacks much formalized meta-research, several past invited papers and panel talks presented expert insights.

On its 20th anniversary, IEEE S&P had a number of panel talks reflecting on the past and considering the future. These panels were on: operating systems [12], covert channels [23], formal methods [22], networking [16, 27], evaluation criteria and commercial technology [18], ubiquitous computing [31], hardware [25], software technology [28], and assurance [29].

IEEE S&P also had three invited meta-papers on its 30th anniversary. Bishop et al. discussed 30 years of IEEE S&P by, for example, mentioning particularly prominent papers from a given year [7]. Landwehr presented the history of government funding in security research [17]. Last, Maughan et al. discussed the transitioning of security research to practice [20].

NDSS at its 20th anniversary in 2013 had a keynote talk on 20 years of security research in network and distributed systems [15]. The content of the talk was both expert insight and some analysis of the prevalence of certain words (e.g., 'PGP') in the NDSS paper abstracts.

These papers, talks, and panels were very informative and incorporated expert knowledge. However, they were not formalized studies of the corpus of published papers, and were largely restricted to S&P and NDSS topics. In contrast, we seek to perform a data-driven study of publications from S&P, CCS, USENIX, and NDSS.

Balzarotti has a data-driven analysis of publications and authorship on his website [6] . He supplies data such as average author count, acceptance rates, submitted papers, author nationalities, top authors, and collaboration networks. These results deal with paper metadata and do not include analysis on paper text or paper topic like ours.

6 Conclusion

Security and privacy meta-research is an under-explored area of research. We have taken the first step towards a formalized, introspective study of the field of security and privacy research. We believe our work will provide a comprehensive view of the overall security and privacy field to new security researchers as well as researchers from different fields.

Acknowledgments

We would like to thank Dr. Vivek Srikumar for his guidance on topic modeling and Sahar Mehrpour for her help in the visualizations presented at https://secprivmeta.net.

References

[1] Apache PDFBox. https://pdfbox.apache.org.

[2] CoreNLP. https://stanfordnlp.github.io/CoreNLP.

[3] html2text. https://pypi.org/project/html2text.

[4] ps2ascii. https://linux.die.net/man/1/ps2ascii.

[5] Stopwords. https://www.ranks.nl/stopwords.

[6] System security circus 2018. http://s3.eurecom.fr/~balzarot/notes/top4_2018/.

[7] Martin Bishop et al. Reflections on the 30th Anniversary of the IEEE Symposium on Security and Privacy (Invited paper). IEEE S&P, 2010.

[8] David M Blei et al. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[9] Jordan Boyd-Graber et al. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of mixed membership models and their applications*, 2014.

[10] Jonathan Chang et al. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

[11] Bent Fuglede et al. Jensen-shannon divergence and hilbert space embedding. In *ISIT*, 2004.

[12] Virgil Gligor. 20 Years of Operating Systems Security (Panel talk). IEEE S&P, 1999.

[13] David Hall et al. Studying the history of ideas using topic models. In *EMNLP*, 2008.

[14] James M Joyce. Kullback-leibler divergence. In *International Encyclopedia of Statistical Science*. 2011.

[15] Richard Kemmerer. 20 Years of Network and Distributed Systems Security: The Good, the Bad, and the Ugly (Keynote talk). NDSS, 2013.

[16] Stephen Kent. Network Security: Then and Now, or, 20 Years in 10 Minutes (Panel talk). IEEE S&P, 1999.

[17] Carl E Landwehr. History of US Government Investments in Cybersecurity Research: A Personal Perspective (Invited paper). IEEE S&P, 2010.

[18] S. Lipner. Twenty Years of Evaluation Criteria and Commercial Technology (Panel talk). IEEE S&P, 1999.

[19] Christopher Manning et al. Introduction to information retrieval. *Natural Language Engineering*, 2010.

[20] Douglas Maughan et al. Crossing the" Valley of Death": Transitioning Cybersecurity Research into Practice (Invited paper). 2013.

[21] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit, 2002.

[22] J. McLean. Twenty Years of Formal Methods (Panel talk). IEEE S&P, 1999.

[23] J. Millen. 20 Years of Covert Channel Modeling and Analysis (Panel talk). IEEE S&P, 1999.

[24] Pedro Moreno-Sanchez et al. Privacy preserving payments in credit networks. In *NDSS*, 2015.

[25] R. Needham. The Hardware Environment (Panel talk). IEEE S&P, 1999.

[26] David Newman et al. Automatic evaluation of topic coherence. In *ACL HLT*, 2010.

[27] H. Orman. Twenty Year Time Capsule Panel: The Future of Networking (Panel talk). IEEE S&P, 1999.

[28] H. Shrobe. Software Technology of the Future (Panel talk). IEEE S&P, 1999.

[29] B. Snow. The Future Is Not Assured. But It Should Be (Panel talk). IEEE S&P, 1999.

[30] Hanna M Wallach et al. Evaluation methods for topic models. In *ICML*, 2009.

[31] M. Weiser. How Computers Will Be Used Differently in the Next Twenty Years (Panel talk). IEEE S&P, 1999.