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

Aniqua Baset, Tamara Denning
School of Computing, University of Utah
What we did

3062 publications 1980-2015

Topic modeling

Trends in authorship and contents
What we did

3062 publications 1980-2015

Topic modeling

Trends in authorship and contents

Online visualizations + Data

SecPrivMeta 36 Years of Security and Privacy Research

Topics: Browser security, DOM and documents

Timeline: browser web site extension cookie content website HTTP URL chrome

Top Words: document content HTML XSS string web, character XML browser element

% of Corpus: 32.0%, 31.0%
🤔 So, why even do this kind of study?
Introspection is important!

- Past evolution and future direction
- Comprehensive overview for new/external audience
- Many communities have introspective studies
  - Computational linguistics, Human computer interaction, Ubiquitous computing, Games, …
Introspection is important!

- Past evolution and future direction
- Comprehensive overview for new/external audience
- Many communities have introspective studies
  - Computational linguistics, Human computer interaction, Ubiquitous computing, Games, …

Past security & privacy introspection

- Panel talks, keynote, invited papers with valuable expert insights
Introspection is important!

- Past evolution and future direction
- Comprehensive overview for new/external audience
- Many communities have introspective studies
  - Computational linguistics, Human computer interaction, Ubiquitous computing, Games, …

Past security & privacy introspection

- Panel talks, keynote, invited papers with valuable expert insights

… But, we lack structured, data-driven efforts
How should we do data-driven introspection?
We need

Publications from different venues → Cohesive categorization
We need Publications from different venues

We want Data-driven approach from publications
We need Publications from different venues → Cohesive categorization

We want Data-driven approach from publications

Problem Not all S&P venues have keywords e.g., USENIX security
We need
Publications from different venues
Cohesive categorization

We want
Data-driven approach from publications

Problem
Not all S&P venues have keywords e.g., USENIX security

Our approach
Topic modeling on full contents of the publications
Overview of topic modeling

Documents → Latent Dirichlet Allocation (LDA) Topic Modeling

Topics
Topic = group of words

Topic distribution for each document

= bag-of-words
Overview of topic modeling

Challenges

• Measuring quality of a topic model is hard
• High-scoring topic ≠ High-quality for people
• Pre-processing texts is crucial
Our methodology

- **S&P** (IEEE Symposium on Security & Privacy)
  - From publishers
  - 1980-2015
  - 456 publications

- **CCS** (ACM Computer and Communications Security)
  - From publishers
  - 1993-1994
  - 1066 publications
  - 1996-2015

- **USENIX** (USENIX Security Symposium)
  - From website
  - 1993
  - 1995-1996
  - 1998-2015
  - 608 publications

- **NDSS** (Network and Distributed System Security Symposium)
  - From website
  - 1997-2015
  - 932 publications

**Full content** + **Title** + **Authors and their affiliations** + **Session name**
Pre-processing input for topic modeling

1) PDF/PS/HTML → TEXT → Main body stripping off meta-data
Pre-processing input for topic modeling

1) **PDF/PS/HTML** → **TEXT** → Main body stripping off meta-data

2) **PDF** → **[Red icon with exclamation mark]** Fixing conversion errors [e.g., fi/fl/ffl ligatures, homoglyphs]
Pre-processing input for topic modeling

1) PDF/PS/HTML ➔ TEXT ➔ Main body stripping off meta-data

2) PDF ➔ Fixing conversion errors [e.g., fi/fl/ffl ligatures, homoglyphs]

3) Lemmatization [e.g., attacks/attacked/attacking ➔ attack]
Pre-processing input for topic modeling

1) PDF/PS/HTML ➔ TEXT ➔ Main body stripping off meta-data

2) PDF ➔ Fixing conversion errors [e.g., fi/fl/ffl ligatures, homoglyphs]

3) Lemmatization [e.g., attacks/attacked/attacking ➔ attack]

4) Preserving technical phrases
   [e.g., man in the middle ➔ man-in-the-middle, MITM ➔ man-in-the-middle]
Pre-processing input for topic modeling

1) **Main body stripping off meta-data**

2) **Fixing conversion errors** [e.g., fi/fl/ffl ligatures, homoglyphs]

3) **Lemmatization** [e.g., attacks/attacked/attacking → attack]

4) **Preserving technical phrases**
   [e.g., man in the middle → man-in-the-middle, MITM → man-in-the-middle]

5) **Stopword list**
   - Most common English words [a, has, the]
   - Common across our corpus
     [words with low Inverse Document Frequency (IDF)]
Generating and selecting a topic model

Different topic models

By varying
• # of topics (60 to 120)
• hyperparameters
Generating and selecting a topic model

Pointwise Mutual Information (PMI)

PMI = coherence score of a topic based on topic-words

PMI↑ Topic quality↑
Generating and selecting a topic model

Different topic models → Model ranking with average PMI → 5 high-scoring models

No perfect model (not uncommon)

Human intervention (accepted)

Highest scoring model → Post-processing to refine
Refining selected topic model

⚠ Mixed topics

E.g.: Garbled circuit and integrated circuit [share words: circuit, gate, bit]
Refining selected topic model

⚠ Mixed topics

E.g.: Garbled circuit and integrated circuit [share words: circuit, gate, bit]

🔧 1. Graph with
   • vertices = publications
   • edge weights = divergences between topic distributions (using Kullback-Leibler divergence)

2. Find sub-community using graph modularity

3. Is a sub-community is a valid topic? Yes → divide
Topic labeling

• Top words
• Top publications
  - keywords or CCS index (if available)
  - session name (if available)
Let’s take a look at our final model!
Total of 95 topics

<table>
<thead>
<tr>
<th>Categories</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRYPTO</td>
<td>(De)obfuscation and decompilation, Domain Name System (DNS), Location privacy/tracking, Real-world sensing</td>
</tr>
<tr>
<td>HARDWARE</td>
<td>Access control, Automated analysis: protocols and files, E-commerce, Electronic voting, Malicious hardware, Routing</td>
</tr>
<tr>
<td>COMPUTATION</td>
<td>Malware, Memory disclosure attacks and defenses, SSL/TLS</td>
</tr>
<tr>
<td>INFORMATION LEAKAGE</td>
<td>Anonymous, Encoding/decoding, Memory errors: exploits and defenses, Secure (multiparty) computation</td>
</tr>
<tr>
<td>ANONYMITY &amp; CENSORSHIP</td>
<td>Anonymous, Encoding/decoding, Memory errors: exploits and defenses, Secure (multiparty) computation</td>
</tr>
<tr>
<td>TRUST</td>
<td>User interfaces, E-commerce, Electronic voting, Malicious hardware, Routing, SSL/TLS</td>
</tr>
<tr>
<td>NETWORKS</td>
<td>Malware, Memory disclosure attacks and defenses, SSL/TLS</td>
</tr>
<tr>
<td>DATA</td>
<td>Virtual, Method, Password, Peer-to-peer communications, Program exploitations: attacks and defenses, Public-key cryptography, Random numbers</td>
</tr>
<tr>
<td>INTERNET</td>
<td>Virtual, Method, Password, Peer-to-peer communications, Program exploitations: attacks and defenses, Public-key cryptography, Random numbers</td>
</tr>
<tr>
<td>Mobile</td>
<td>Virtual, Method, Password, Peer-to-peer communications, Program exploitations: attacks and defenses, Public-key cryptography, Random numbers</td>
</tr>
<tr>
<td>CRYPTOGRAPHIC SYSTEM</td>
<td>Virtual, Method, Password, Peer-to-peer communications, Program exploitations: attacks and defenses, Public-key cryptography, Random numbers</td>
</tr>
<tr>
<td>CRIME &amp; FRAUD</td>
<td>Virtual, Method, Password, Peer-to-peer communications, Program exploitations: attacks and defenses, Public-key cryptography, Random numbers</td>
</tr>
<tr>
<td>MISCELLANEOUS</td>
<td>Virtual, Method, Password, Peer-to-peer communications, Program exploitations: attacks and defenses, Public-key cryptography, Random numbers</td>
</tr>
</tbody>
</table>

Grouped into 20 categories

- CRYPTO
- HARDWARE
- COMPUTATION
- INFORMATION LEAKAGE
- ANONYMITY & CENSORSHIP
- TRUST
- NETWORKS
- DATA
- INTERNET
- Mobile
- CRYPTOGRAPHIC SYSTEM
- CRIME & FRAUD
- MISCELLANEOUS
## Example: CRYPTO category

<table>
<thead>
<tr>
<th>Topic label</th>
<th>Top 5 words from LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cryptographic protocols</td>
<td>protocol session party session-key secret</td>
</tr>
<tr>
<td>Encryption</td>
<td>encryption ciphertext encrypted decryption decrypt</td>
</tr>
<tr>
<td>Network authentication</td>
<td>authentication authenticate kerberos secret service</td>
</tr>
<tr>
<td>Crypto and number theory</td>
<td>mod bit prime rsa random</td>
</tr>
<tr>
<td>Digital signature</td>
<td>signature sign public signer verification</td>
</tr>
<tr>
<td>Public-key cryptography</td>
<td>certificate CA trust revocation sign</td>
</tr>
<tr>
<td>Key distribution/management</td>
<td>round broadcast secret threshold secret-sharing</td>
</tr>
<tr>
<td>Group communication</td>
<td>group member multicast join communication</td>
</tr>
<tr>
<td>Random numbers</td>
<td>entropy output random pool randomness</td>
</tr>
</tbody>
</table>
## Example: CRYPTO category

<table>
<thead>
<tr>
<th>Topic label</th>
<th>Top 5 words from LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cryptographic protocols</strong></td>
<td>protocol session party session-key secret</td>
</tr>
<tr>
<td>Encryption</td>
<td>encryption ciphertext encrypted decryption</td>
</tr>
<tr>
<td>Network authentication</td>
<td>decrypt</td>
</tr>
<tr>
<td>Crypto and number theory</td>
<td>authentication authenticate kerberos secret</td>
</tr>
<tr>
<td>Digital signature</td>
<td>service</td>
</tr>
<tr>
<td>Public-key cryptography</td>
<td>mod bit prime rsa random</td>
</tr>
<tr>
<td>Key distribution/management</td>
<td>signature sign public signer verification</td>
</tr>
<tr>
<td>Group communication</td>
<td>certificate CA trust revocation sign</td>
</tr>
<tr>
<td>Random numbers</td>
<td>round broadcast secret threshold secret-sharing</td>
</tr>
</tbody>
</table>


## Example: CRYPTO category

<table>
<thead>
<tr>
<th>Topic label</th>
<th>Top 5 words from LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cryptographic protocols</td>
<td>protocol session party session-key secret</td>
</tr>
<tr>
<td>Encryption</td>
<td>encryption ciphertext encrypted decryption decrypt</td>
</tr>
<tr>
<td>Network authentication</td>
<td>authentication authenticate kerberos secret service</td>
</tr>
<tr>
<td><strong>Crypto and number theory</strong></td>
<td><strong>mod bit prime rsa random</strong></td>
</tr>
<tr>
<td>Digital signature</td>
<td>signature sign public signer verification</td>
</tr>
<tr>
<td>Public-key cryptography</td>
<td>certificate CA trust revocation sign</td>
</tr>
<tr>
<td>Key distribution/management</td>
<td>round broadcast secret threshold secret-sharing</td>
</tr>
<tr>
<td>Group communication</td>
<td>group member multicast join communication</td>
</tr>
<tr>
<td>Random numbers</td>
<td>entropy output random pool randomness</td>
</tr>
</tbody>
</table>
## Example: Crime & Fraud category

<table>
<thead>
<tr>
<th>Topic label</th>
<th>Top 5 words from LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dark web</td>
<td>site URL search web website</td>
</tr>
<tr>
<td>Spam, scam, and fraud</td>
<td>spam email account mail post</td>
</tr>
<tr>
<td>Online advertising</td>
<td>ad ads publisher click target</td>
</tr>
<tr>
<td>Online crime</td>
<td>account market service customer country</td>
</tr>
</tbody>
</table>
## Example: Crime & Fraud category

<table>
<thead>
<tr>
<th>Topic label</th>
<th>Top 5 words from LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dark web</td>
<td>site URL search web website</td>
</tr>
<tr>
<td>Spam, scam, and fraud</td>
<td>spam email account mail post</td>
</tr>
<tr>
<td><strong>Online advertising</strong></td>
<td>ad ads publisher click target</td>
</tr>
<tr>
<td>Online crime</td>
<td>account market service customer country</td>
</tr>
</tbody>
</table>
Let’s look at some trends
How have the venues changed over time?
How have the venues changed over time?

![Graph showing changes over time with indicators for entropy and diversity increasing.](image-url)
How has the category distribution changed over time?
How consistent are authors and topics year-to-year?
How consistent are authors and topics year-to-year?

Jaccard Index (88-89, 90-91) = Similarity between topic sets of 88-89 and 90-91

Jaccard Index↑ ≈ Overlap↑
How has industry and government participation changed over time?

The graph shows the percentage of publications by different sectors over time, with categories including Academics, Academics + Industry, Academics + Government, Government, Industry, and Others. The graph indicates an increasing trend in collaboration between academia and industry, as well as a steady increase in government involvement.
Do non-academic collaborators have same interests?

Top 15 topics in recent years (2011-2015)

Industry:
- Online advertising
- DOM & documents
- Bots & botnet
- Spam, scam & fraud
- Client-sever accountability

Overall:
- Static & dynamic analysis
- Binary analysis
- Mobile app
- Veri comp. & ZKP
- Machine learning
- Malware
- Data privacy

Government:
- SSL\TLS
- Program exploit.
- Side-channel
- HW: low level
- Crypto & num. theory
- Information flow
- Formal meth. & ver.
- VMs & virtualization
- Tor

Browser
- Secure comp.
- Dark web
- JavaScript

Secure comp.
- Online advertising
- DOM & documents
- Bots & botnet
- Spam, scam & fraud
- Client-sever accountability
- Static & dynamic analysis
- Binary analysis
- Malware
- Data privacy

Mobile app
- Veri comp. & ZKP
- Machine learning
- Malware
- Data privacy

Social networks & (De)annon.

Top 15 topics in recent years (2011-2015)
Tool and data availability
Our site secprivmeta.net

Interactive visualizations
Topics | Topic-timelines | Topic-words | Publications | Authors | …

Available data
Meta-data with categorized affiliations | Acronym list | Stop-word list | Original topic model | …
SecPrivMeta 36 Years of Security and Privacy Research

About


More visualizations to come in the future. Please check again!

Team

Anjua Baset, PhD Student, School of Computing, University of Utah

Dr. Tamara Denning, Assistant Professor, School of Computing, University of Utah

Sitemap

- The CATEGORIES page shows each of the 96 topics, clustered by category.
- The Topics dropdown provides a quick way to jump to a topic without navigating to the Categories page.
- The WORDS shows all the prominent words in different topics. Selecting a word lists related topics.
- The PUBLICATIONS provides a list of all publications from the four conferences (from 1980-2015).
- The AUTHORS lists all authors who published in the four conferences (from 1980-2015).

Acknowledgment

We would like to thank Dr. Vivek Srikumar for his helpful guidance on topic modeling. We also thank Sehar Mehrpour for helping us in the visualizations presented at this site.

The visualizations of topics and papers are inspired by the efr-browser developed by Andrew Goldstone.

Downloads

- Acronym list
- Stopword list
- Phrase list
Aniqua Baset, aniqua@cs.utah.edu
Dr. Tamara Denning, tdenning@cs.utah.edu
References/Backup
References: introspective studies in other fields

Publications

- Towards a computational history of the ACL: 1980-2008
- Studying the history of ideas using topic models
- Identifying crisis of Ubicomp?: mapping 15 years of the field’s development and paradigm change
- CHI 1994-2013: mapping two decades of intellectual progress through co-word analysis
- Games research today: Analyzing the academic landscape 2000-2014

• Online tool/dataset

  - ACL Anthology Network (All About NLP)
    • Text, meta-data created using papers from ACL Anthology which hosts 51975 papers on the study of computational linguistics and natural language processing
Category trends using number of papers