

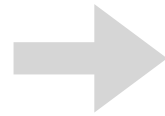
A Data-Driven Reflection on 36 Years of Security and Privacy Research

Aniqua Baset, Tamara Denning
School of Computing, University of Utah

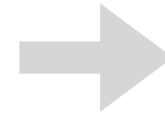
What we did



3062 publications
1980-2015



Topic modeling

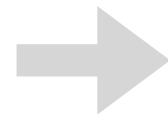


Trends in authorship
and contents

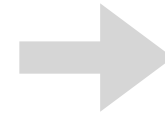
What we did



3062 publications
1980-2015

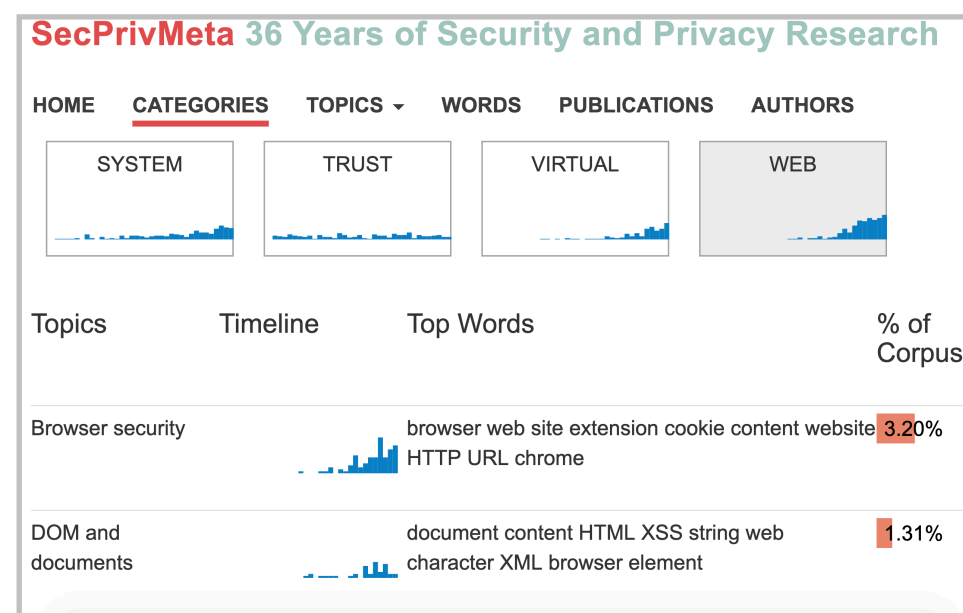


Topic modeling



Trends in authorship
and contents

Online visualizations + Data





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

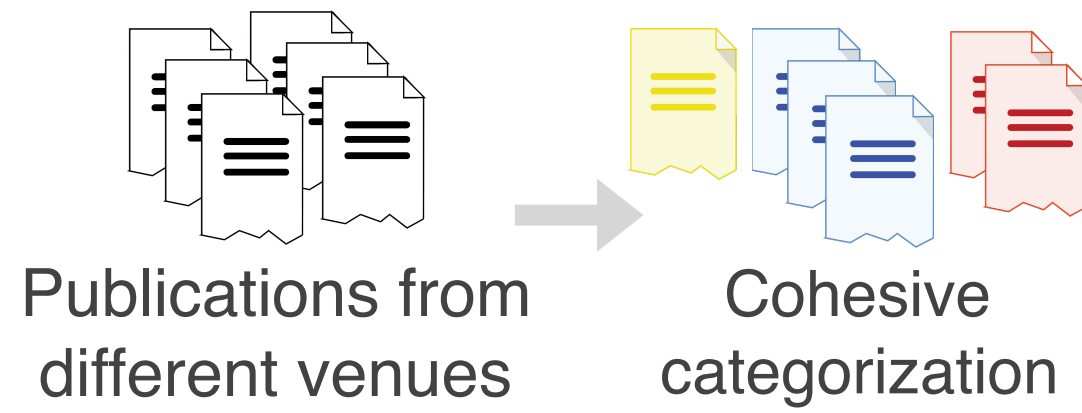
- Panel talks, keynote, invited papers with valuable expert insights

... But, we lack structured, data-driven efforts

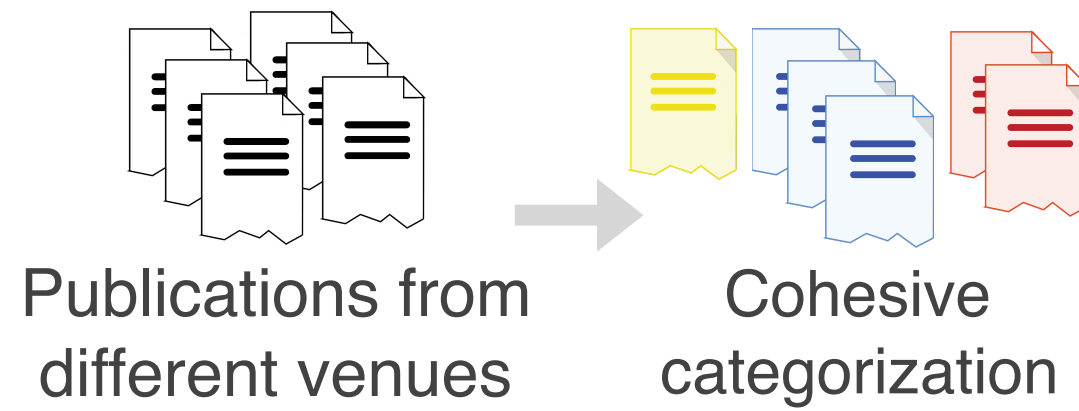


How should we do data-driven introspection?

We need



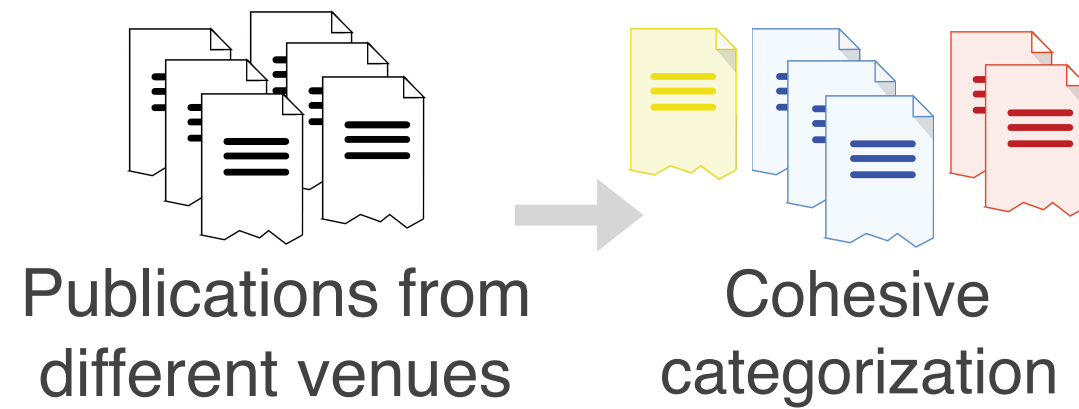
We need



We want

Data-driven approach from publications

We need



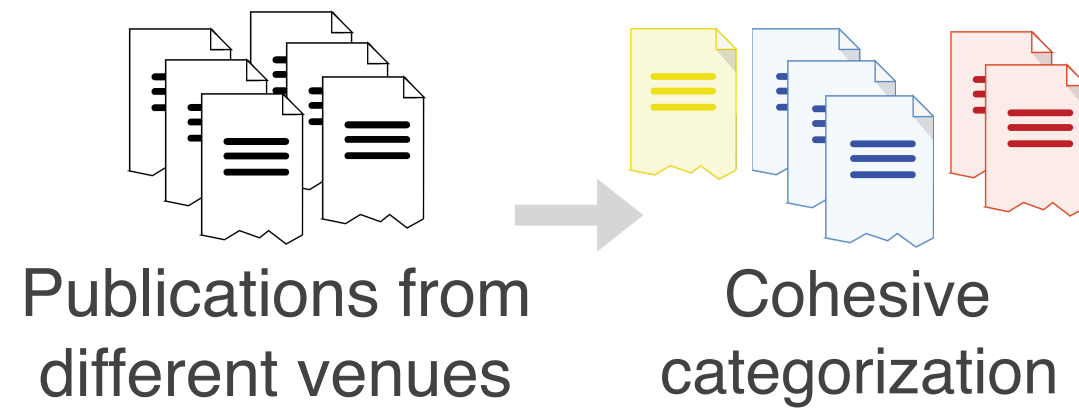
We want

Data-driven approach from publications

Problem

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

We need



We want

Data-driven approach from publications

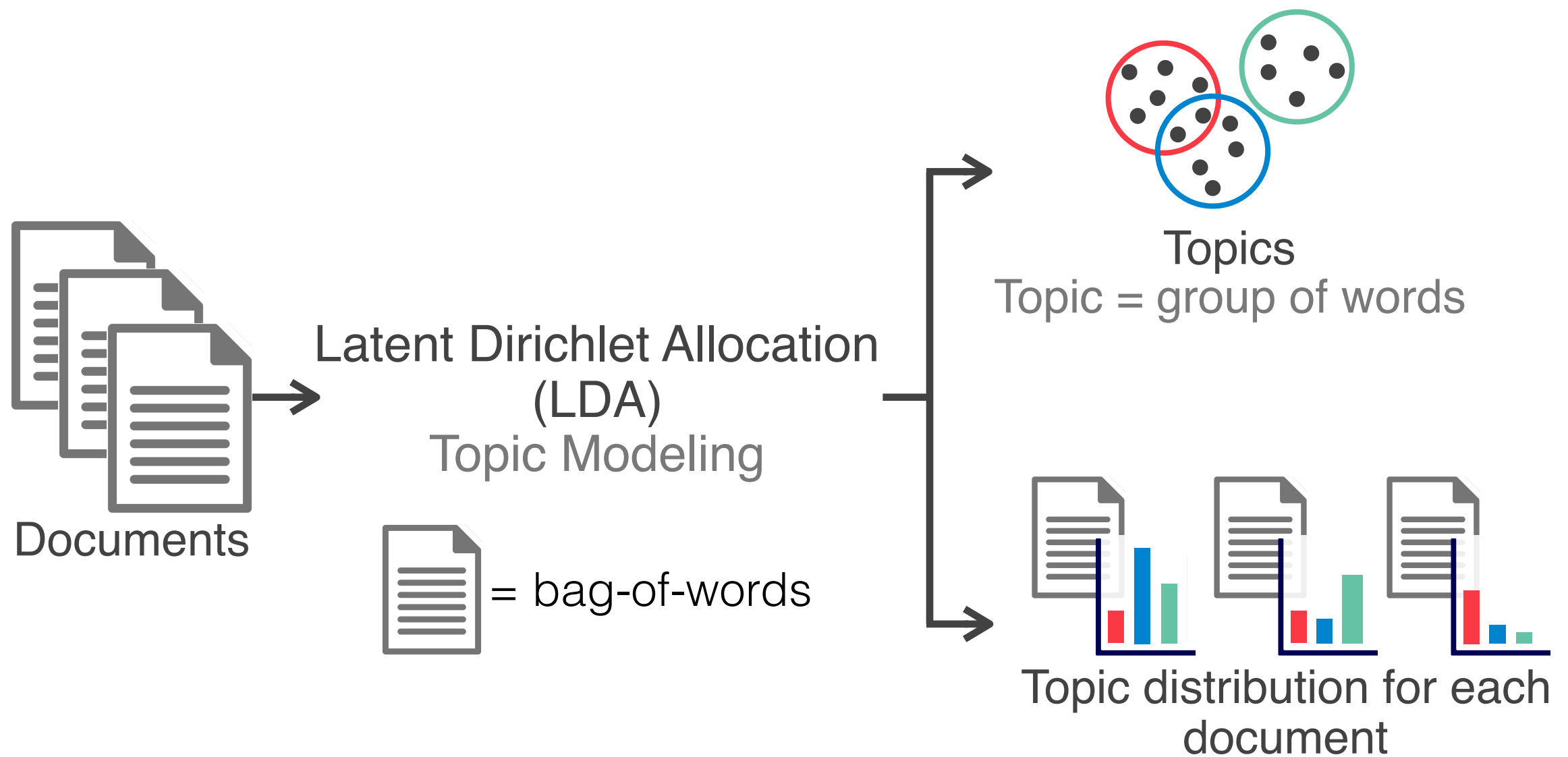
Problem

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

Our approach

Topic modeling on full
contents of the publications

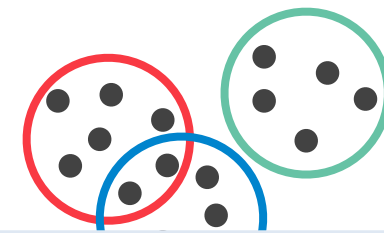
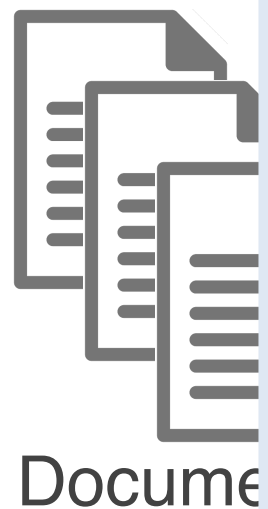
Overview of topic modeling



Overview of topic modeling

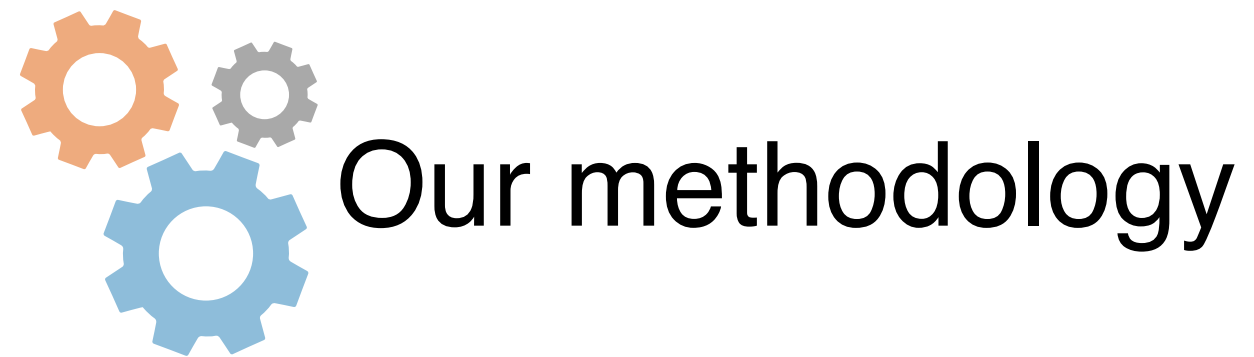
Challenges

- Measuring quality of a topic model is hard
- High-scoring topic \neq High-quality for people
- Pre-processing texts is crucial



Topic distribution for each document





Data collection: 3062 publications, 1980-2015

S&P

From publishers

IEEE Symposium on
Security & Privacy



1980-2015



456

CCS

From publishers

ACM Computer and
Communications Security



1993-1994
1996-2015



1066

USENIX

From website

USENIX Security
Symposium



1993
1995-1996
1998-2015



608

NDSS

From website

Network and Distributed
System Security Symposium



1997-2015

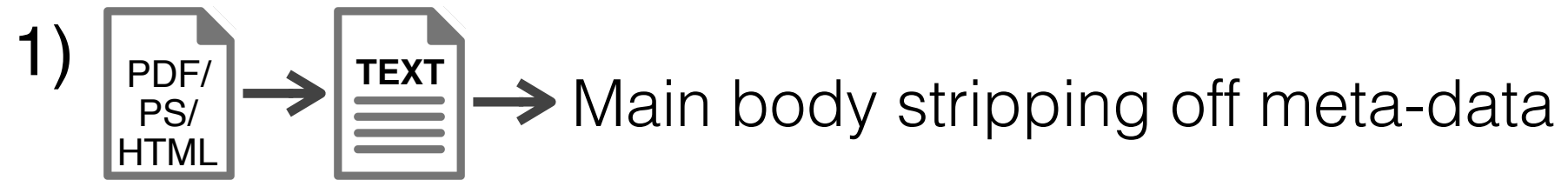


932








Full content + Title + Authors and their affiliations + Session name

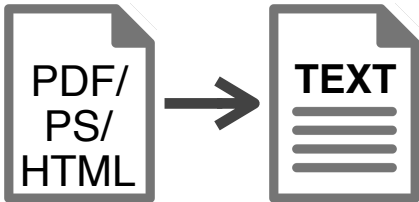
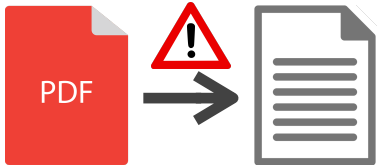
Pre-processing input for topic modeling



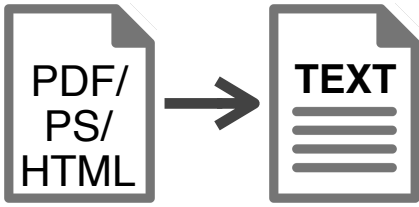
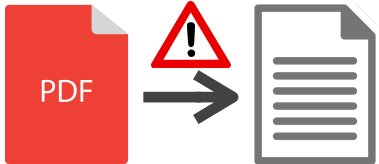
Pre-processing input for topic modeling

- 1)  →  → Main body stripping off meta-data
- 2)   →  Fixing conversion errors [e.g., fi/fl/ffl ligatures, homoglyphs]

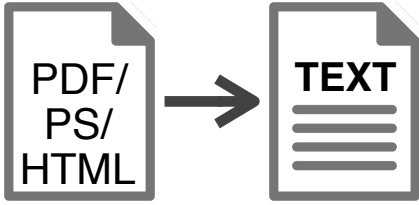
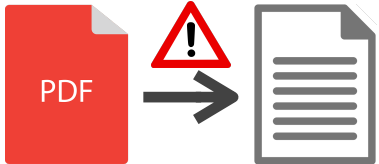
Pre-processing input for topic modeling

- 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]

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]

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)]

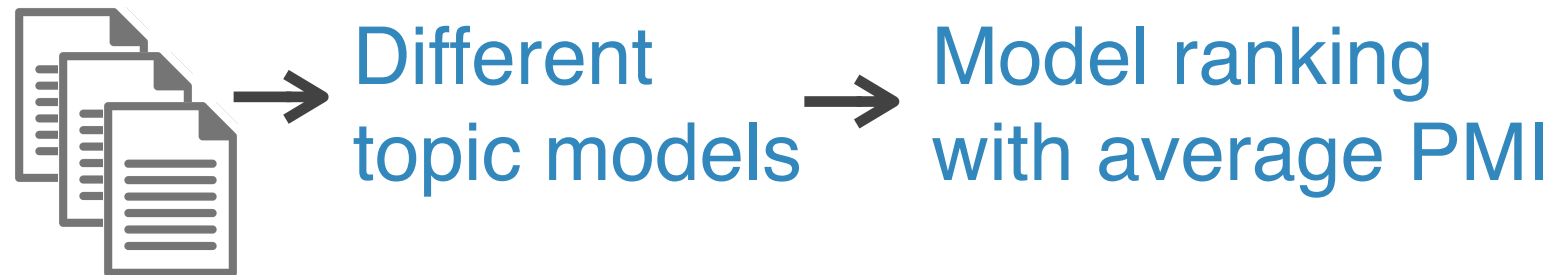
Generating and selecting a topic model



By varying

- # of topics (60 to 120)
- hyperparameters

Generating and selecting a topic model

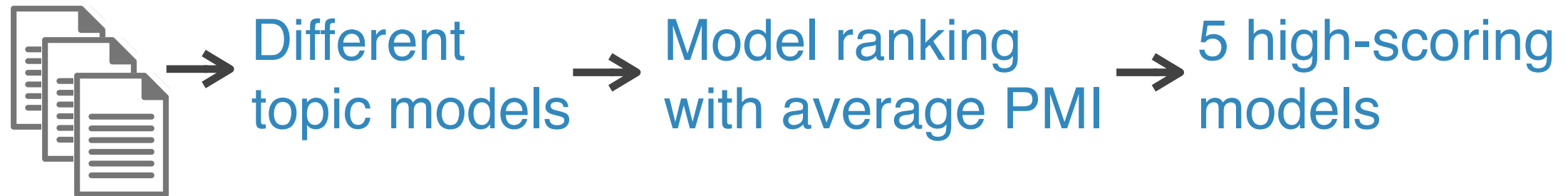


Pointwise Mutual Information (PMI)

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

PMI↑ Topic quality↑

Generating and selecting a topic model



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]

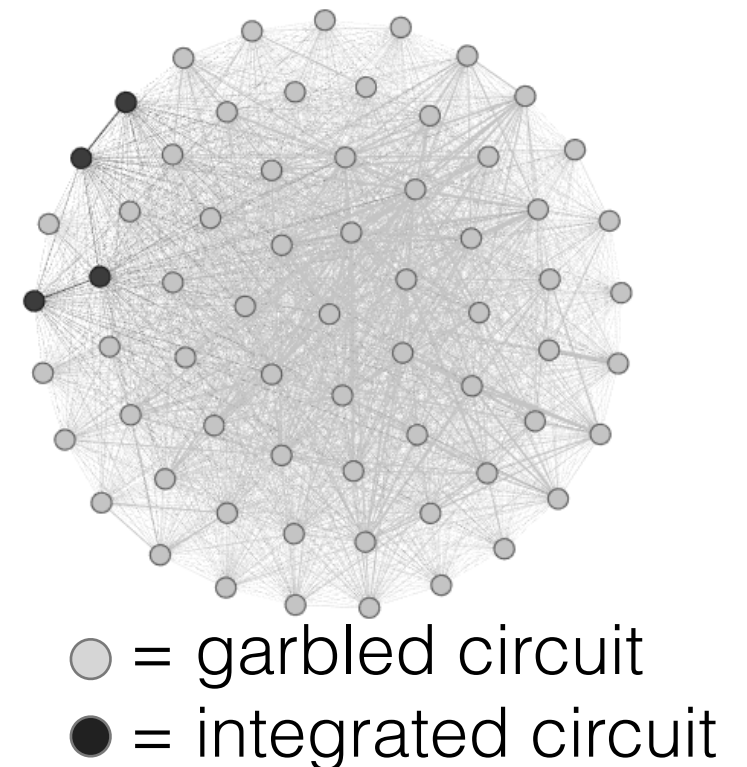
Refining selected topic model



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

(De)obfuscation and decompilation	Domain Name System (DNS)	Location privacy/tracking	Real-world sensing
(User) interfaces	E-commerce	Malicious hardware	Routing
Access control	Electronic voting	Malware	SSL/TLS
Automated analysis: protocols and files	Embedded and hardware security	Memory disclosure attacks and defenses	Secure (multiparty) computation
Android
Bitcoin
BitTorrent
Blockchain
Browser
C/A
Ce
Ce
Cli
Clo
Co
Co
Co
Cr
Cr
DOM and documents	Intrusion/anomaly detection	Passwords	Phishing and worm propagation and scanning
Dark web	Java security	Peer-to-peer communications	Vulnerabilities: exploits disclosure and patches
Data privacy	JavaScript security	Program exploitations: attacks and defenses	Web application vulnerabilities
Databases	Kernels	Public-key cryptography	Wireless signal
Digital signature	Key distribution/management	Random numbers	

Grouped into 20 categories

CRYPTO	TRUST	FORMALISM	SYSTEM
HARDWARE	NETWORKS	WEB	AUTH
COMPUTATION	DATA	MALWARE	PROGRAMS
INFORMATION LEAKAGE	INTERNET	MOBILE	CRIME & FRAUD
ANONYMITY & CENSORSHIP	VIRTUAL	METHOD	MISCELLANEOUS

Example: CRYPTO category

Topic label

Top 5 words from LDA

Cryptographic protocols

protocol session party session-key secret

Encryption

encryption ciphertext encrypted decryption
decrypt

Network authentication

authentication authenticate kerberos secret
service

Crypto and number theory

mod bit prime rsa random

Digital signature

signature sign public signer verification

Public-key cryptography

certificate CA trust revocation sign

Key distribution/management

round broadcast secret threshold secret-sharing

Group communication

group member multicast join communication

Random numbers

entropy output random pool randomness

Example: CRYPTO category

Topic label

Top 5 words from LDA

Cryptographic protocols

protocol session party session-key secret

Encryption

encryption ciphertext encrypted decryption
decrypt

Network authentication

authentication authenticate kerberos secret
service

Crypto and number theory

mod bit prime rsa random

Digital signature

signature sign public signer verification

Public-key cryptography

certificate CA trust revocation sign

Key distribution/management

round broadcast secret threshold secret-sharing

Group communication

group member multicast join communication

Random numbers

entropy output random pool randomness

Example: CRYPTO category

Topic label

Top 5 words from LDA

Cryptographic protocols

protocol session party session-key secret

Encryption

encryption ciphertext encrypted decryption
decrypt

Network authentication

authentication authenticate kerberos secret
service

Crypto and number theory

mod bit prime rsa random

Digital signature

signature sign public signer verification

Public-key cryptography

certificate CA trust revocation sign

Key distribution/management

round broadcast secret threshold secret-sharing

Group communication

group member multicast join communication

Random numbers

entropy output random pool randomness

Example: CRIME & FRAUD category

Topic label

Top 5 words from LDA

Dark web

site URL search web website

Spam, scam, and fraud

spam email account mail post

Online advertising

ad ads publisher click target

Online crime

account market service customer country

Example: CRIME & FRAUD category

Topic label

Dark web

Top 5 words from LDA

site URL search web website

Spam, scam, and fraud

spam email account mail post

Online advertising

ad ads publisher click target

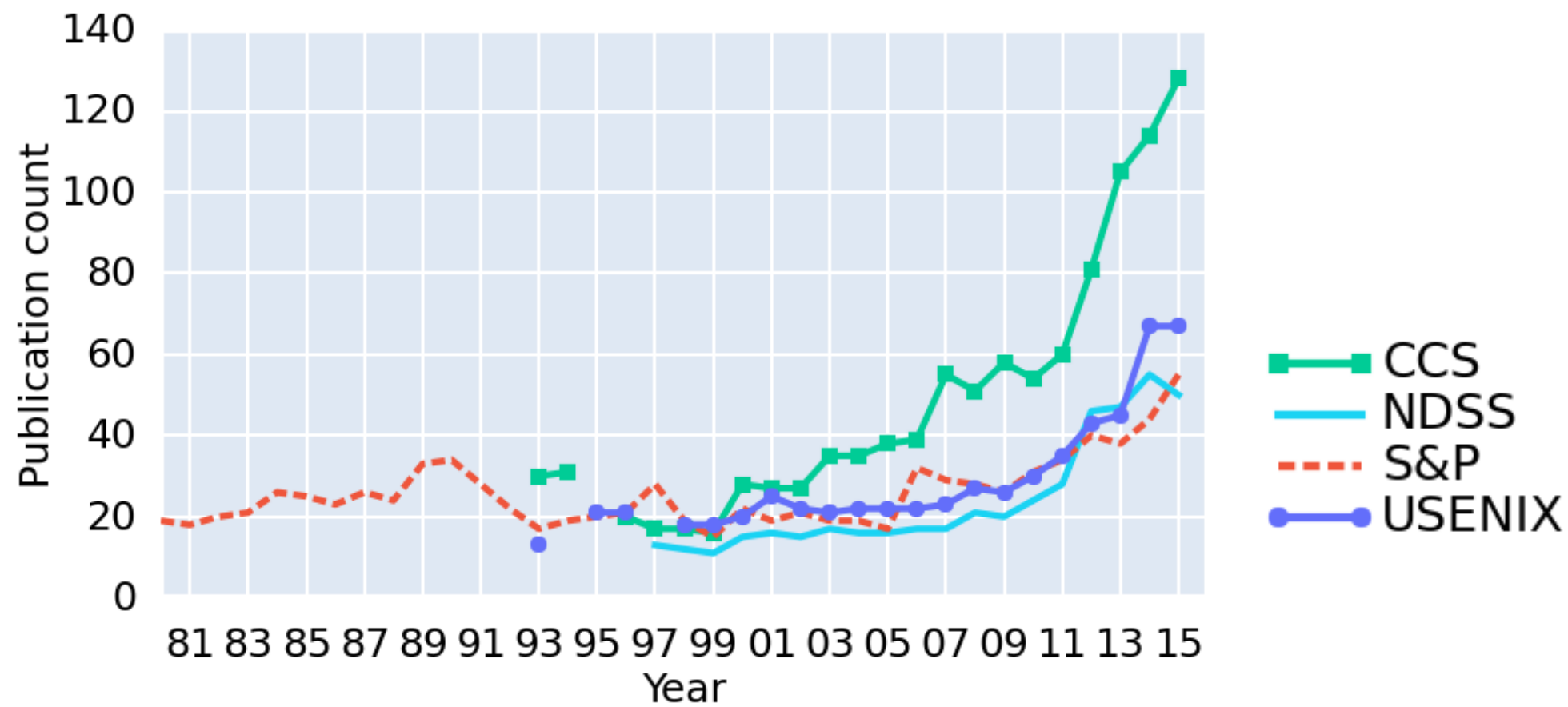
Online crime

account market service customer country

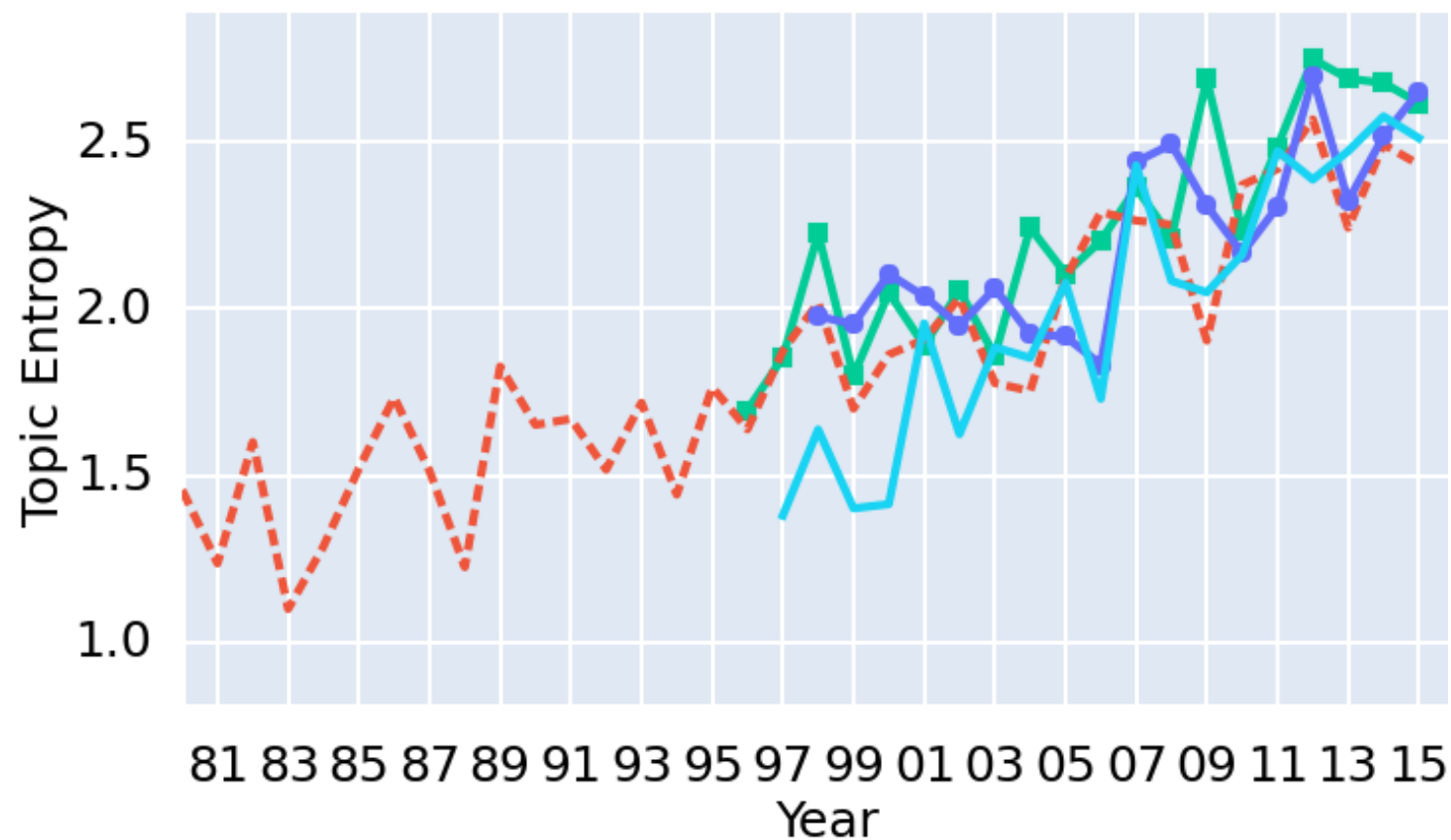
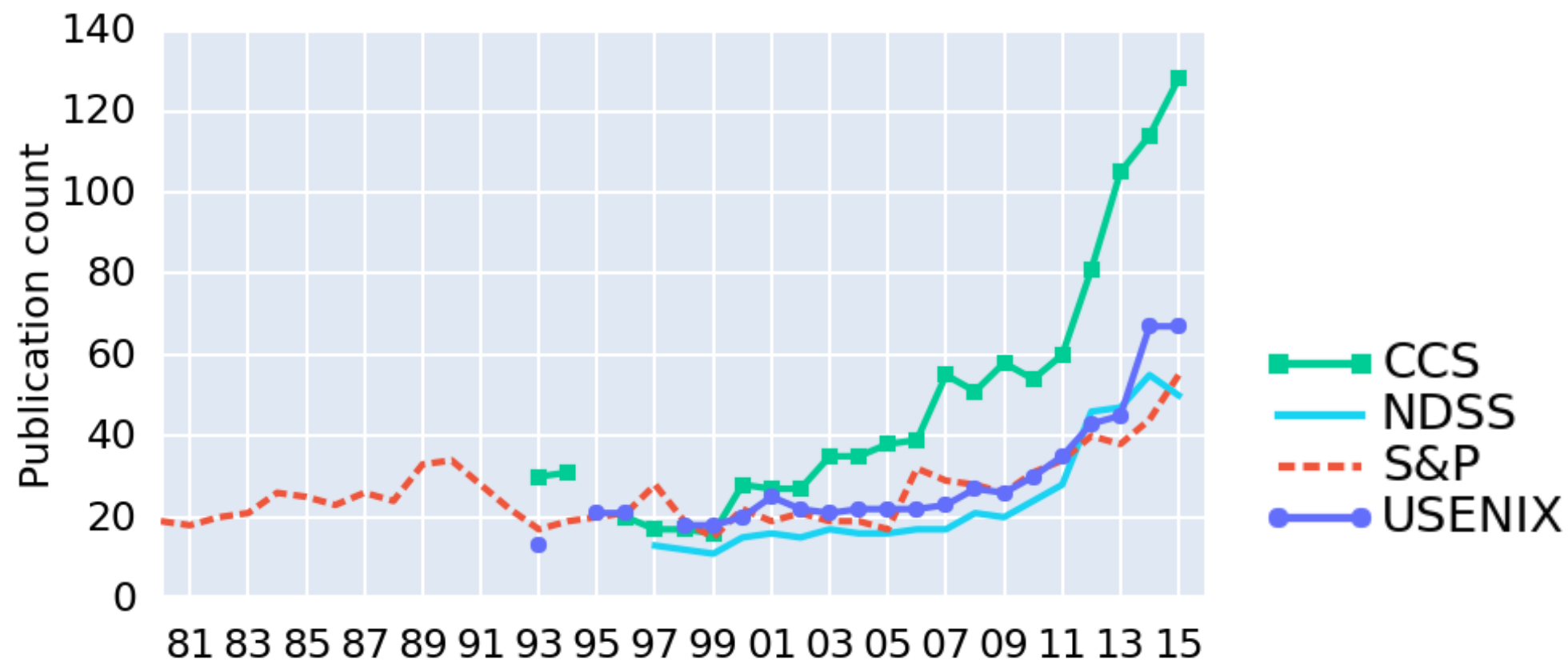


Let's look at some trends

How have the venues changed over time?

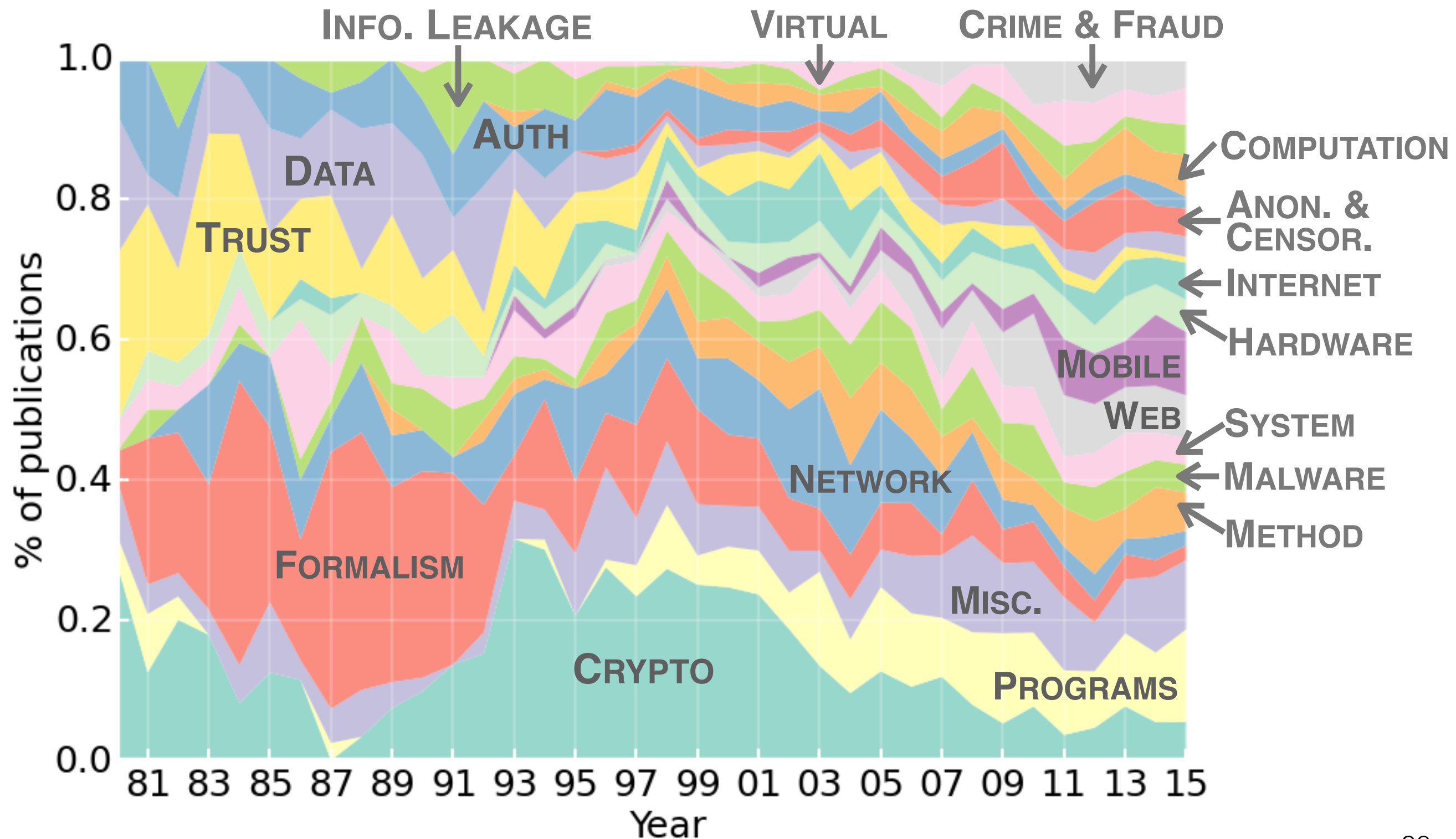


How have the venues changed over time?

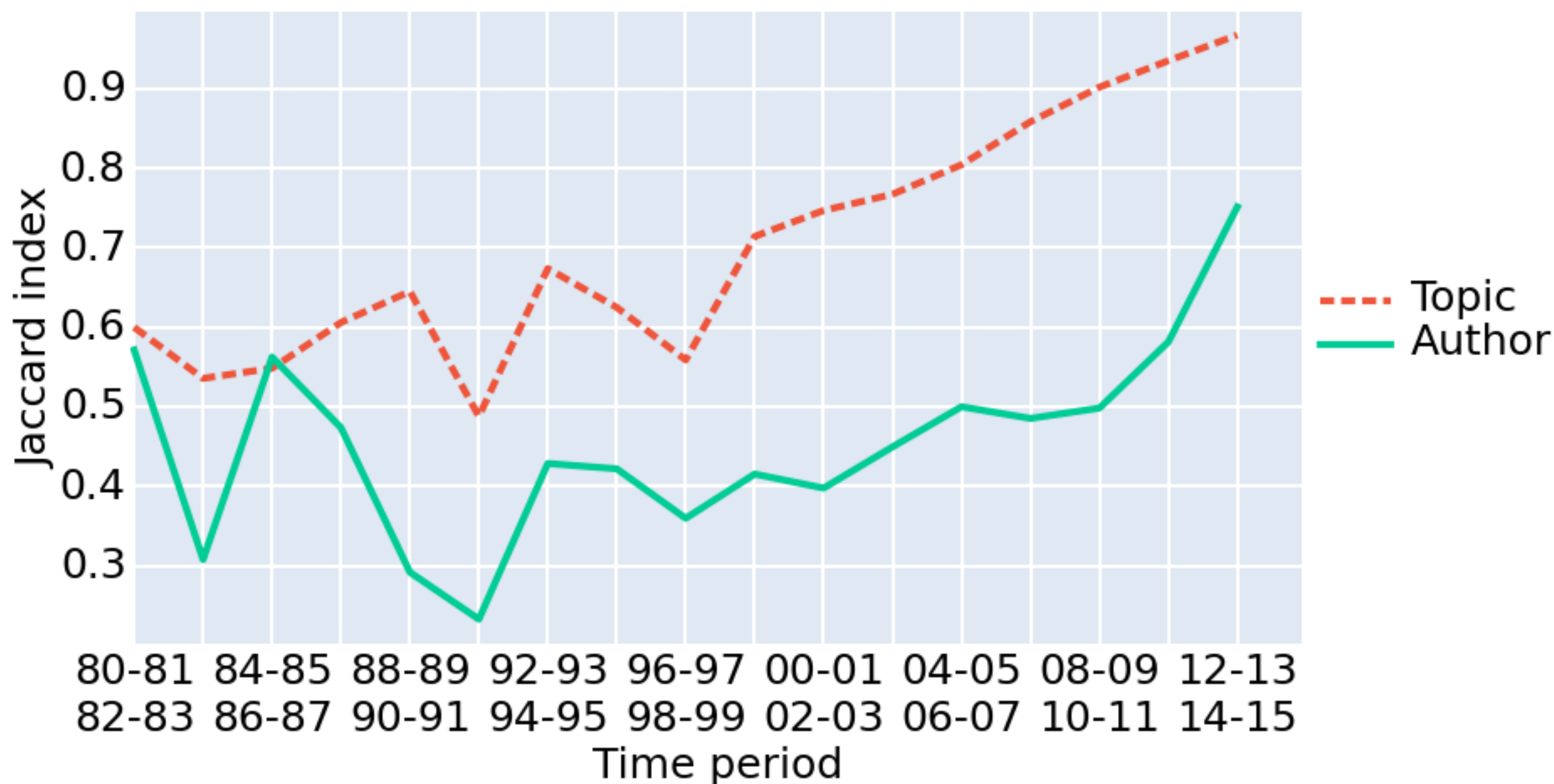


Entropy $\uparrow \approx$ Diversity \uparrow

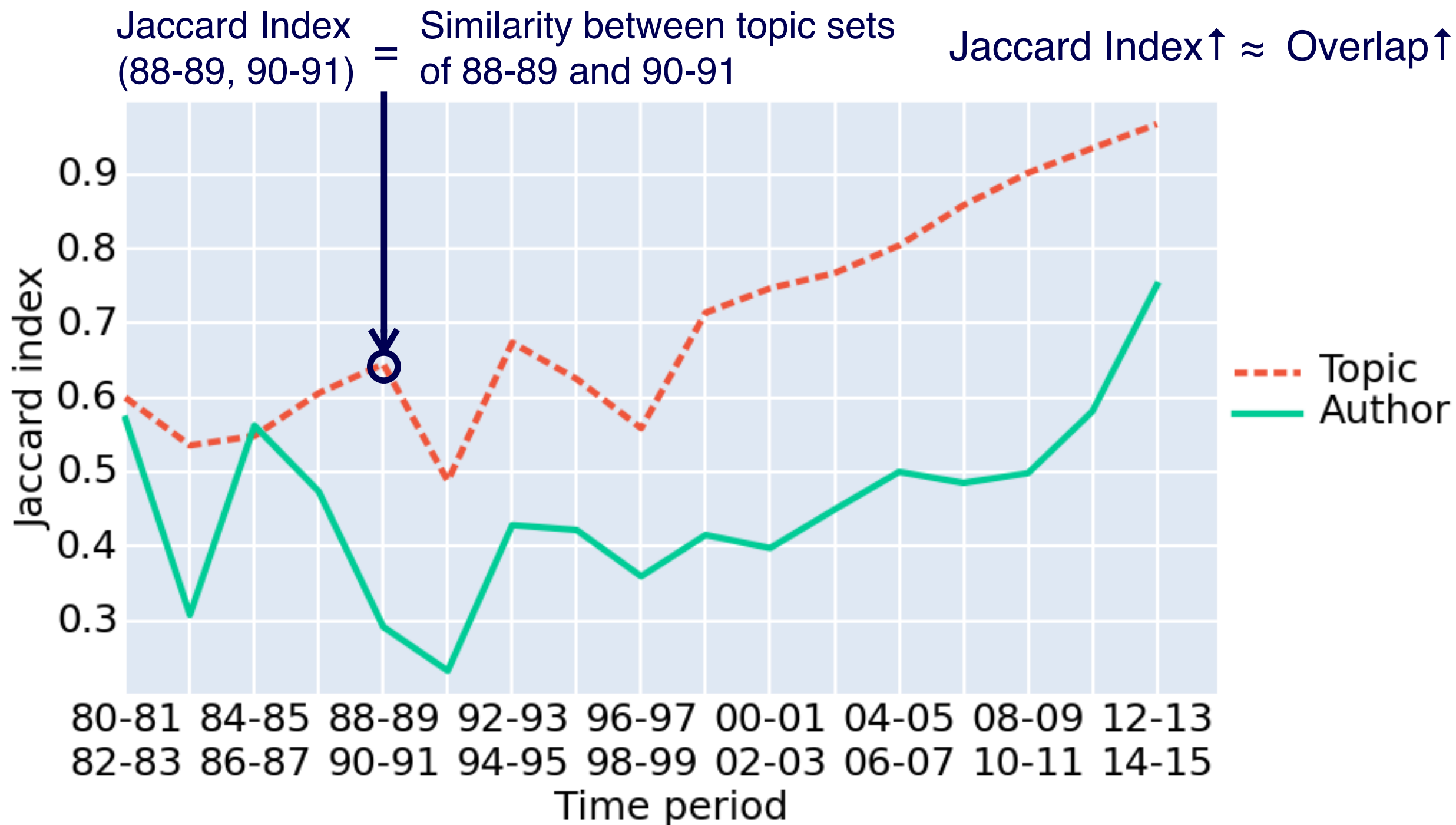
How has the category distribution changed over time?



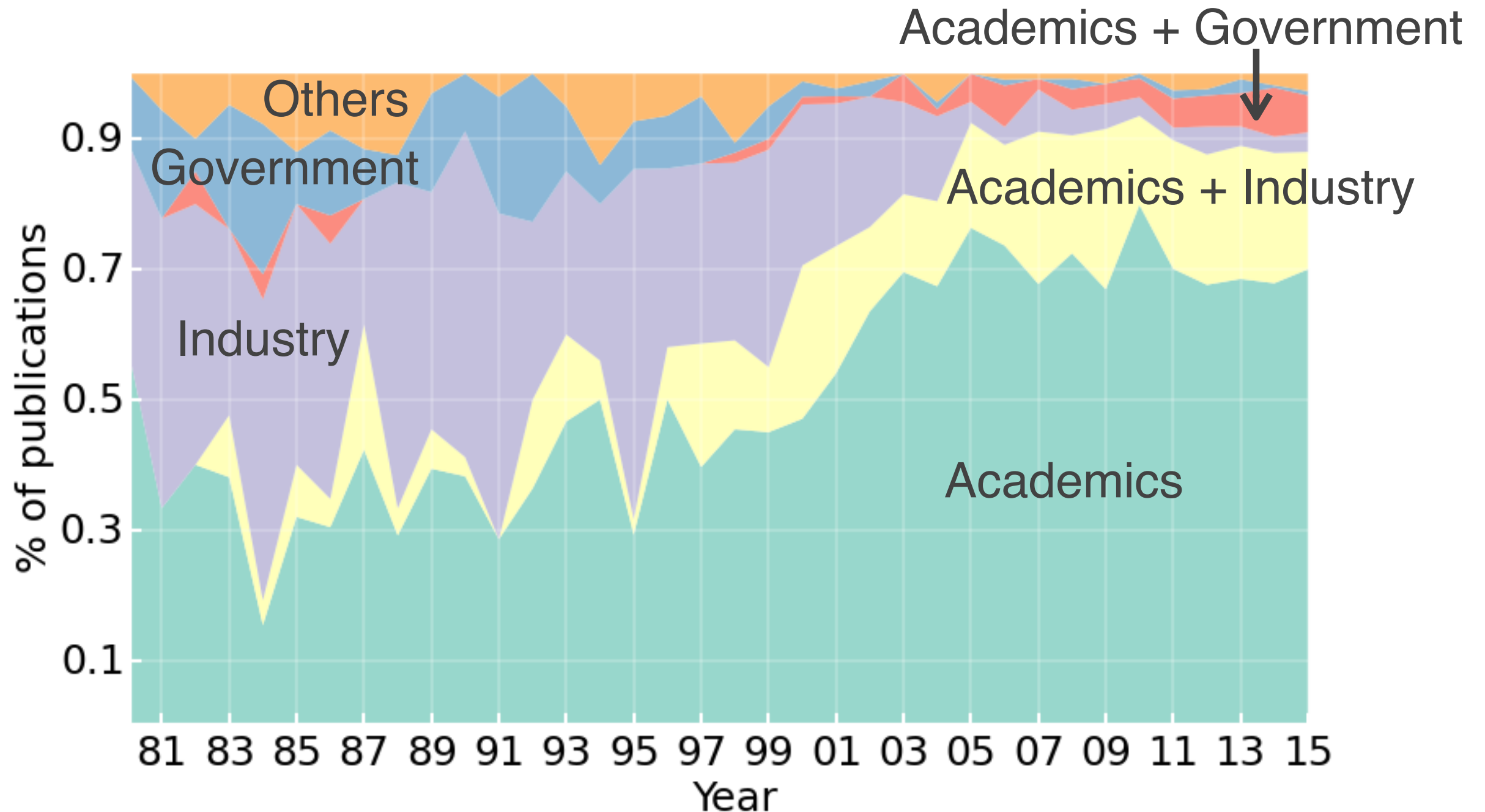
How consistent are authors and topics year-to-year?



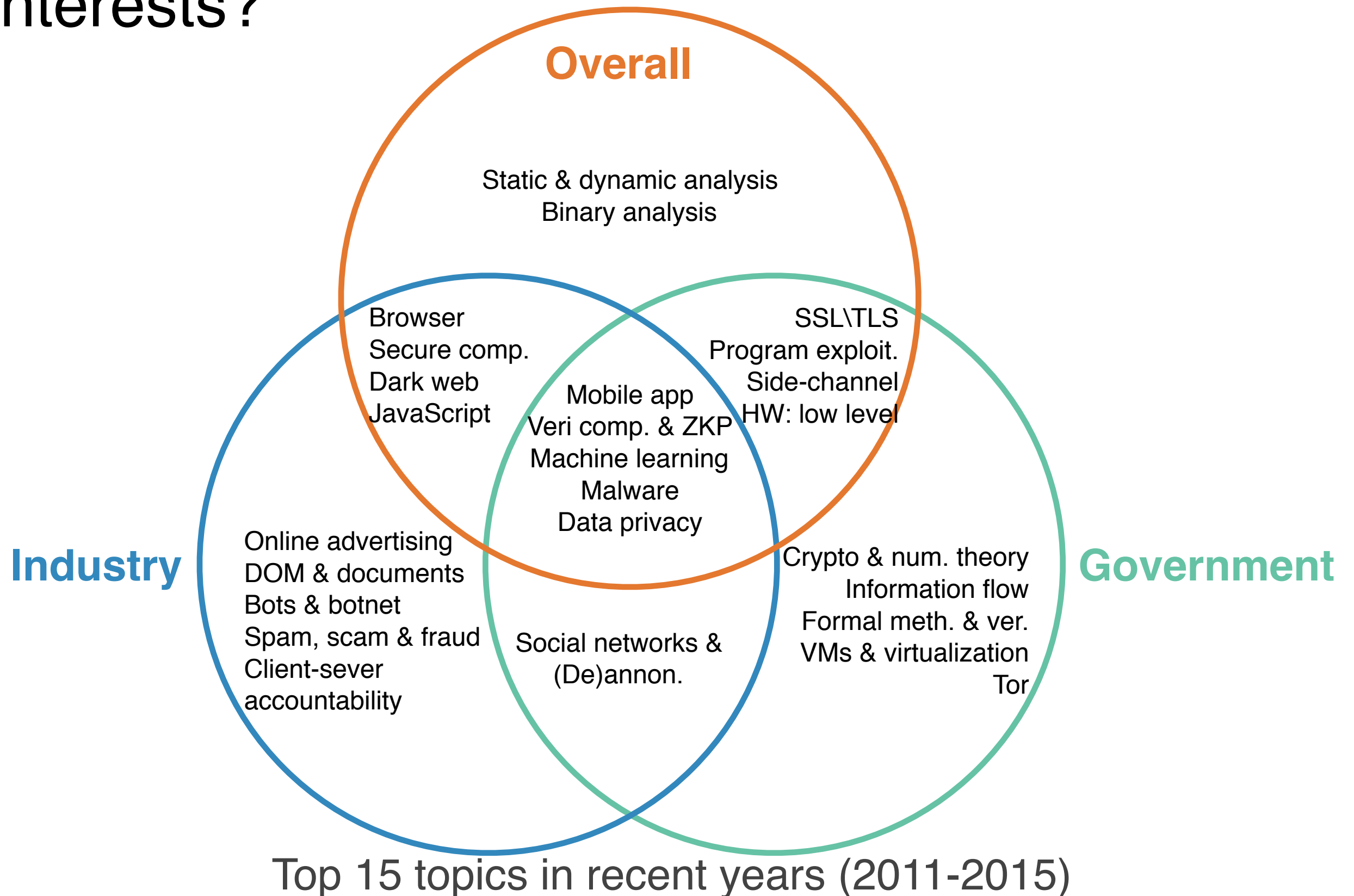
How consistent are authors and topics year-to-year?



How has industry and government participation changed over time?



Do non-academic collaborators have same interests?





Tool and data availability

Our site secprivmeta.net

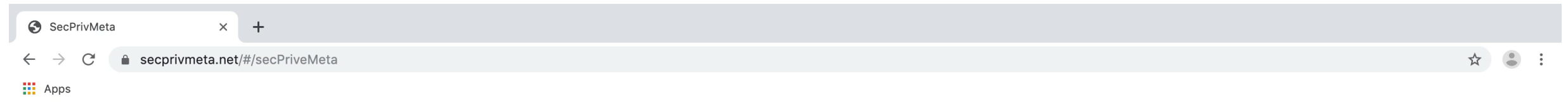
Interactive visualizations

Topics | Topic-timelines | Topic-words | Publications |
Authors | ...

Available data

Meta-data with categorized affiliations | Acronym list |
Stop-word list | Original topic model | ...

secprivmeta.net demo



SecPrivMeta 36 Years of Security and Privacy Research

[HOME](#) [CATEGORIES](#) [TOPICS](#) [WORDS](#) [PUBLICATIONS](#) [AUTHORS](#)


About

We present a topic modeling on 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). Check out our [CSET'19 paper](#) for more details!

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

Team

Aniqua 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 95 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 Sahar Mehrpour for helping us in the visualizations presented at this site.

The visualizations of topics and papers are inspired by the [dfr-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

