Uncovering the structures of privacy research using bibliometric network analysis and topic modelling

Friso van Dijk, Joost Gadellaa and Chaïm van Toledo
Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
Marco Spruit
Department of Public Health and Primary Care, Leiden University, Leiden, The Netherlands, and
Sjaak Brinkkemper and Matthieu Brinkhuis
Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

Received 23 May 2022
Revised 19 October 2022
Accepted 18 January 2023

Abstract

Purpose – This paper aims that privacy research is divided in distinct communities and rarely considered as a singular field, harming its disciplinary identity. The authors collected 119.810 publications and over 3 million references to perform a bibliometric domain analysis as a quantitative approach to uncover the structures within the privacy research field.

Design/methodology/approach – The bibliometric domain analysis consists of a combined directed network and topic model of published privacy research. The network contains 83,159 publications and 462,633 internal references. A Latent Dirichlet allocation (LDA) topic model from the same dataset offers an additional lens on structure by classifying each publication on 36 topics with the network data. The combined outcomes of these methods are used to investigate the structural position and topical make-up of the privacy research communities.

Findings – The authors identified the research communities as well as categorised their structural positioning. Four communities form the core of privacy research: individual privacy and law, cloud computing, location data and privacy-preserving data publishing. The latter is a macro-community of data mining, anonymity metrics and differential privacy. Surrounding the core are applied communities. Further removed are communities with little influence, most notably the medical communities that make up 14.4% of the network. The topic model shows system design as a potentially latent community. Noteworthy is the absence of a centralised body of knowledge on organisational privacy management.

Originality/value – This is the first in-depth, quantitative mapping study of all privacy research.

Keywords Privacy, Bibliometric, Mapping, Network, Topic model

Paper type Research paper

1. Introduction

In line with scientific research at large, privacy research output has increased significantly over the past decade (van Dijk *et al.*, 2021). The rapid expansion of privacy research is

Organizational Cybersecurity Journal: Practice, Process and People
Vol. 3 No. 2, 2023
pp. 81-99
Emerald Publishing Limited
e-ISSN: 2635-0289
p-ISSN: 2635-0270
DOI 10.1108/OCJ-11-2021-0034

attributed to the ever-increasing collection, processing and dissemination of personal data, at larger scales and by innovative means. This leads to new and increased privacy violations, collective awareness, legislative responses and new approaches for safeguarding privacy. However, despite the increased volume of scientific publications on understanding and addressing privacy, privacy research is rarely considered as a coherent field.

Symptomatic to the lack of coherence are the influential literature surveys in privacy research. Most surveys focus on a single research community, for example consumer privacy (Lanier and Saini, 2008), privacy-preserving data mining (Verykios *et al.*, 2004) or nursing (Leino-Kilpi *et al.*, 2002). Even the well-known multi-disciplinary survey by Smith *et al.* (2011) excludes the majority of privacy research due to the limits of its qualitative methodology. As a result, privacy research appears fractured between disciplines, creating a need for mapping studies.

A quantitative analysis of the bibliometric record of privacy research can offer additional insights to the structures within privacy research. We apply network analysis and topic modelling to the bibliometric record to generate insights in the structures of the privacy research community. These techniques offer the opportunity for identifying communities and influential, "authoritative" information sources in networks that function as hubs that hold the network together. Relevant examples of its uses include scientific impact measures (De Mey, 1982), finding indicators of inter-disciplinarity in journals (Leydesdorff, 2007) and the effects of co-authorship on scholarly performance (Abbasi *et al.*, 2011).

Topic modelling further enhances the network analysis by offering an additional lens through which this data can be categorised and viewed. In the past, it has been used to investigate fields such as communication research (Maier *et al.*, 2018), fisheries science (Syed and Spruit, 2018) and journalism (Jacobi *et al.*, 2016). The combination of these techniques offers a broad mapping of the similarities and distinction between the research communities within privacy research.

The purpose of this study is to perform a multi-method bibliometric domain analysis of the structures in privacy research. Central to this investigation are two research questions (RQ):

RQ1. What communities and topics make up the privacy research field?

RQ2. How are the identified communities positioned within broader privacy research?

We expand on a previously created network of privacy research (van Dijk *et al.*, 2021), which will provide the research communities of privacy research and their positioning within privacy research. A newly created topic model will provide insights on the topics of interest to privacy researchers. Overlaid on the network, the topic model offers further opportunities to identify the topical interests of each community as well as investigate potential latent communities in privacy research.

2. Research methodology

The methodology applied in this research begins with (1) re-use of a previously created network of privacy research from 83,159 publications. (2) Additional data is collected to (3) create a topic model from 79,432 publications to enhance the bibliometric record. The results of these analyses are (4) synthesised for an in-depth overview of the structures in the privacy research field.

2.1 Network analysis

This study re-uses a directed network of the bibliometric record of modern privacy research (van Dijk *et al.*, 2021). The network consists of 83,159 publications from 1965 until 2022 as

nodes and the 462,633 references between them as edges. Edges are unweighted, as each reference is considered equal. The network is scale-free and has a clustering coefficient of 0.076. This indicates that clusters and communities exist around much-cited publications. However, together with its average path length (5.6) and maximum path length (21), fragmented information exists within the network as well.

van Dijk *et al.* (2021) provided a visualisation of the network displaying the communities in the network and the primary topic(s) of each significant cluster (Figure 1). Communities in the network were identified through the Louvain method (Blondel *et al.*, 2008). Each community was already named after its significant clusters, which were visually identified in the network and labelled by random sampling and agreement between two researchers. To further analyse each identified community, we use the following averaged structural measures for each community in the network.

(1) In-degree, representing the citations received by the community from other publications in the network.

(2) Out-degree, for the references made by the community within the network.

(3) Eigenvector centrality, representing the quality of connections within the community. Communities with the most influential publications in the entire network will score higher.

(4) Betweenness centrality, for the structural position of a community within the network. Communities with a high betweenness centrality are considered to be in a bridge position in the network.

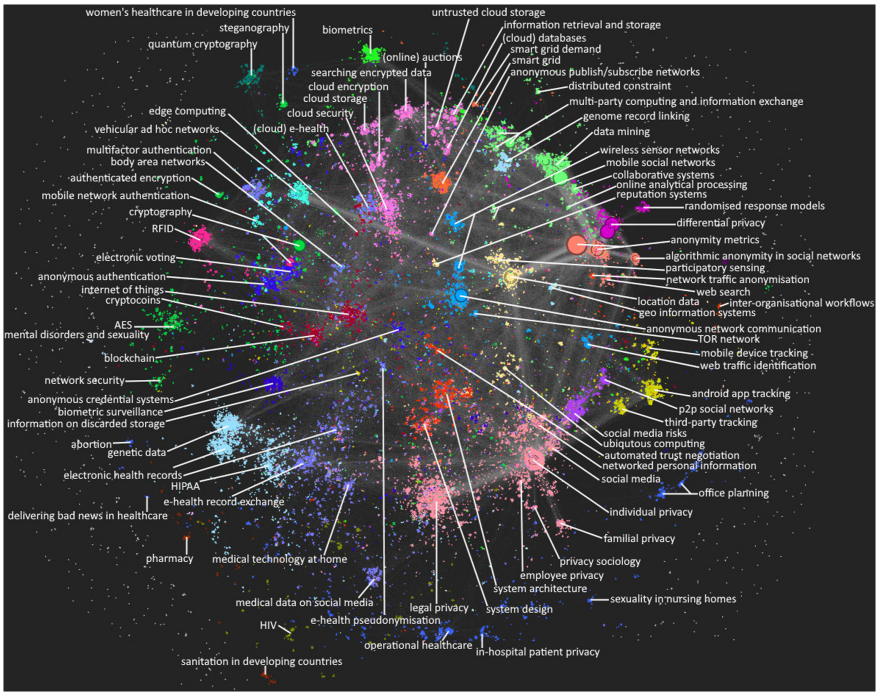

Figure 1.
Visual network representation of privacy literature with labeled clusters. Node size represents citation count. Colours represent research communities (Table 1)

Together, these centrality measures will provide an overview of the structural positioning of each community within the network.

2.2 Additional data collection

For the topic analysis, the network dataset was extended with the corresponding keywords and abstracts using Pybliometrics (Rose and Kitchin, 2019). Only keywords added by the author(s) were included in the data set to prevent the topic model from learning from previous categorisations, be it automatic or human. Only abstracts are considered, as the benefit from the inclusion of the full text is negligible and significantly increases the computational resources required (Syed and Spruit, 2017).

Since the initial query, 126 works had been retracted or otherwise removed. Of the 83,033 works left, 24,982 are without keywords, 3,534 of which without abstract. Sample inspection showed that these are primarily complete books and conference introductions. 67 works are missing just their abstract but have keywords. All documents without abstract were removed. No new publications were added to the network dataset. The final dataset used for the topic analysis consists of 79,432 publications.

2.3 Topic model

A Latent Dirichlet allocation (LDA) topic model was created from the collected data (Gadellaa, 2021). This generative probabilistic model uses the frequencies of words in the documents to uncover latent topics and assign every document a distribution over these topics (Blei *et al.*, 2003). LDA is a mixed membership model, which means that documents are seen as part of multiple topics in different proportions. This makes it particularly suitable for exploratory and descriptive analyses (Elgesem *et al.*, 2015). By capturing the heterogeneity of research topics that a specific paper can belong to, LDA overcomes the limitations of other techniques (Syed and Spruit, 2017). It is important to note that LDA is a bag-of-words (BOW) model, meaning it does not consider the order of words, just their presence and count within a document. Although this is an oversimplification from how humans understand text, it is well-founded for the purpose of uncovering latent structures (Blei and Lafferty, 2006).

2.3.1 Data preparation. To create a topic model in a reproducible and objective manner, two of the main methodological steps that need to be taken are appropriate pre-processing and adequate selection of model parameters (Maier *et al.*, 2018). A downside of LDA is that these steps can influence results in a significant way and that their order matters, as consecutive steps are dependent on previous results (Denny and Spirling, 2018). The following data preparation was applied, closely following recommendations by Maier *et al.* (2018).

(1) Copyright notices are removed from the abstracts using a regular expression.

(2) Tokenization, converting to lowercase and punctuation removal using the Gensim package (Rehurek and Sojka, 2011) for Python, and can be considered the removal of uninformative data in the BOW approach used here (Scott and Matwin, 1999).

(3) Stop-word removal, using the Natural Language Toolkit's (NLTK) English stop-word list (Bird *et al.*, 2009).

(4) Lemmatization and stemming, using WordNet (Miller, 1995) to collapse inflicted forms to a single term. Lemmatization was chosen over the more aggressive stemming procedure to maintain more interpretability. It is an effective feature-reduction procedure for English texts (Haselmayer and Jenny, 2014).

(5) Relative pruning, also referred to as removing 'corpus specific stop words'. All words that appear in less than 0.5% or more than 99% of the texts are removed. No

systematic studies have been conducted on which percentages of documents should be considered lower and upper-bound, but <1% and >95% seem to be the de facto standard (Denny and Spirling, 2018; Grimmer and Stewart, 2013) and recently, Maier *et al.* (2020) showed that such percentages do not decrease model quality.

The first two steps can be considered the removal of uninformative data in the BOW approach used here (Scott and Matwin, 1999). All further steps are feature selection, with the overarching goal to only keep those words with distribution patterns that contribute to distinction between topics. This reduction is also essential for the computational efficiency of the later hyper-parameter tuning.

The relative pruning reduced the number of terms from 75,148 to 2,421; 97% of words occurred too little or too often to contribute to the discovery of latent topics. It is noteworthy that the term privacy was not removed, while it is the keyword upon which the papers were selected. This was because more than 1% of documents were added to the database because privacy was in the keywords added by the publisher, which were not included in this analysis.

2.3.2 Parameter tuning. After preparing the documents, several of LDA's model parameters have to be defined. Much of our method for finding the optimal model is based on Syed *et al.* (2018), which used a similar approach on fisheries science literature. We optimise our topic model for C_v coherence, a measure of model quality introduced by Röder *et al.* (2015). Both their systematic exploration and a later comparison to alternatives by Syed (2019) show that this measure correlates highly with human understanding on a large variety of test data. The following considerations were made in defining the LDA's model parameters:

(1) *The optimal number of topics (K).* Setting this value too low will result in topics that are too broad to be meaningful, but setting it too high can create semantically meaningless topics that should have been combined (Syed and Spruit, 2017). In some topic model applications, there are reasons to infer the "right" number of topics from theory or the analysis' goals. In most cases, the optimal number is discovered by generating different models and measure their quality according to a defined metric. There is no right answer, just a number that produces fitting results according to the chosen evaluation metric.

(2) *The Dirichlet prior distribution on the topic probabilities within documents (α) and on the word probabilities within topics (η).* In text corpora like these, where some topics are expected to be less common than others, an asymmetrical prior is preferred (Syed and Spruit, 2018; Wallach *et al.*, 2009). Gensim's "online" LDA algorithm (Hoffman *et al.*, 2010) can use the Newton–Raphson method to learn both priors from the data (Huang, 2005). The technical details of this algorithm are beyond the scope of this analysis, but the relevant effect is that finding the most appropriate priors is not done through an additional dimension in our grid search, saving computation time.

(3) *The random state, the number of passes through the corpus during training, and the number of iterations for inferring the topic distribution.* This is not necessary for each level of K, but the convergence profile might differ between versions with different parameters. To our knowledge, no more sophisticated methods than a grid search exist for these parameters. The random state has to be varied to prevent falling into a local minimum (Boyd-Graber *et al.*, 2014).

In this analysis an initial grid search was performed on models with a number of topics (K) between 1 and 361. The steps between the different values were based on the power function x^2 (1, 4, 9, 16, etc.). This sequence was chosen ad hoc after some exploratory model generation showed coherence measurements to be less sensitive to changes in topic numbers in the

higher regions. The first search across two random states showed that coherence was maximal between 25 and 49. The search was then deepened to include four random states and additional values for K in between the values in the original sequence. Further exploration of interactions between variables was not possible because of the computational cost of evaluating models. Nevertheless, the most important variables could be varied in a structured manner, generating 56 models to evaluate (Gadellaa, 2021). The best fitting model provided 36 topics.

2.3.3 Topic labelling. The final part of topic model creation consists of labelling the topics uncovered by the algorithm. In our context of creating a second classification of the same dataset, human labelling is useful and pragmatic. The labelling was be done by two annotators, who separately reviewed each topic's top words and a visualisation using LDA vis (Sievert and Shirley, 2014). There was agreement on 31 of 36 (86%) topic labels between the annotators. Differences between annotators were settled by reviewing the options for fit and sampling a random sub-set of titles from each topic.

2.4 Synthesis of network and topic model

The topic distribution of each publication and its network outcomes were combined to create an enhanced bibliometric record of the privacy field. This resulted in a dataset consisting of 83,159 publications with, for each publication, its centrality measures, community and main topic. From this dataset, a heat map visualisation was created to identify the subjects of interest in each research community. It displays the proportion of topic occurrence within each network communities. Furthermore, we extracted a sub-set of the 65 most cited publications within the network (in-degree) by combining the 50 most cited publications of the entire timespan with the 50 most cited publications of the last five years. Citations remain the baseline influence measure in the network analysis (Newman, 2010). The period of five years was applied to address the age bias inherent to bibliometric data and is a modal value of citation analysis research (Moody *et al.*, 2010).

We use this combined dataset of the network analysis and topic model to answer the research questions as follows:

To answer *RQ1. What communities and topics make up the privacy research field?* We consider the outcomes of the community analysis and investigate patterns of interest, such as potentially latent communities.

RQ2. How are the identified communities positioned within broader privacy research? is answered by analysing the structural characteristics of the research communities within the network. We identify the central communities in privacy research, which form the disciplinary core of the research field and consider the findings against literature.

Section 3.1 presents the network, and its communities and their centrality measures. In 3.2, we show the labelled topic model. Section 3.3 brings together these results by displaying the relationship between communities and topics, as well as the topic distribution of the most influential publications. The research questions are then answered by considering the combined analysis in the discussion.

3. Results

3.1 Network communities

A graphical rendering of the network of privacy research (Figure 1) provides insight into the structures of the field. It is a layered visualisation, made up of.

(1) 83,159 nodes: dots representing each of the publications in the network. The size and colour of each node is determined by the number of references received and its community respectively.

(2) 462,633 edges: references between publications and are depicted as a line between two nodes.

(3) 94 clusters: a clustering of nodes of significant size within the network, named through 90 labels.

(4) Communities: algorithmically identified groupings of one or more clusters. The accompanying Table 1 shows the 20 largest communities from a total of 42, with their corresponding colours. These 20 communities account for 93% of the network and are colour-coded in Figure 1.

While a visualisation is only one representation of a more complex dataset (Grandjean, 2016), it provides a high-level overview of the privacy research field.

Table 1 displays the average centrality measures for each of the 20 largest communities. The meaning of the network centrality measures in this context are explained in section 2.1. Like individual centrality measures, the average in-degree remains the primary indicator of influence as the number of citations received. The average out-degree shows a close relation with in-degree, suggesting an exchange between referencing in the network and getting referenced from the network. Whether that dynamic exists within or between communities cannot be said from this data.

Colour	Community	Size (%)	Avg. In-degree	Avg. Out-degree	Eigenvector*	Betweenness*
	Individual privacy and law	16.2	6.31	6.31	0.46	0.63
	Cloud computing	7.9	5.43	5.37	0.29	0.35
	E-health and medical data	7.6	3.54	3.77	0.08	0.45
	Genetic data	6.8	3.83	3.79	0.13	1
	Data mining	5.5	7.58	7.32	0.58	0.5
	Location data	5.2	8.86	9.23	0.67	0.85
	Anonymity metrics	4.4	11.08	10.46	1	0.83
	Internet of things	4.3	3.66	4.58	0.08	0.52
	Differential privacy	3.8	10.64	10.83	0.96	0.94
	Electronic voting	3.6	4.28	4.06	0.17	0.18
	Networking	3.6	5.84	5.71	0.33	0.2
	Cybersecurity	3.2	2.49	2.13	0	0
	Mobile devices and apps	3.0	4.31	4.25	0.17	0.31
	System architecture and design	3.0	4.4	4.53	0.21	0.33
	Vehicular ad hoc networks	3.0	6.59	7.11	0.42	0.36
	RFID	2.9	6.56	6.72	0.42	0.24
	Social media	2.5	5.8	5.84	0.33	0.41
	Physical privacy	2.4	2.41	2.42	0	0.06
	Biometrics	2.1	5.06	5.13	0.25	0.27
	Smart grid	1.8	6.09	6.49	0.33	0.41

Note(s): * Normalised average values

Table 1. The 20 largest communities in the network and their averaged centrality measures. Colour coding was included to ease identification of the communities in the Figure 1 network

Eigenvector centrality appears strongly correlated with in-degree, which can be explained by its focus on influential neighbours. The higher the average in-degree of a community, the more likely a publication in that community has influential neighbours.

Betweenness centrality does not follow in-degree. As an indicator for information flow through the network, it increases for the most central communities in the network – for example *individual privacy and law*, *anonymity metrics* and *differential privacy* – and those connecting a large, otherwise less connected community (with a lower average in-degree) to the larger body of knowledge – such as *e-health and medical data* and *genetic data*. Publications bridging the gap between these communities and the rest of the network have an excessively high betweenness centrality, as they are most often on the shortest path between nodes.

On the lower end of the spectrum, a combination of a lower average in-degree and eigenvector centrality appears in communities that form common exit points in the network. The two clearest examples are *cybersecurity* and *physical privacy*. Furthermore, there appears to be a sharp drop off in average betweenness centrality for each community with an average in-degree below 3 with more than 10 publications in them. Publications in these communities are more likely to consume information from the rest of the network than to connect otherwise separate disciplines.

3.2 Topic model

The 56 generated topic models were evaluated for their quality, C_v. Most of the models held a value between 0.32 and 0.49, showing a clear pattern where C_v scores increased sharply for the initial 'increase' in the number of topics K. After the peak, C_v scores slowly decreased for models with more topics. The highest scoring model had a C_v score of 0.517 and a K of 36. This number of topics was not always the maximum across the different random states, suggesting an interaction between the random state and the effect of K, indicating that multiple coherent topic models exist for this dataset. Although the highest scoring model is chosen for interpretation, we have to keep in mind that we are looking at one of many possible representations of this data. A textual representation of the final topic model can be found in Table 2.

Notably, topic 19 seems to include both some recommender systems and healthcare. Inspecting the papers within this category does not yield a satisfying explanation of this grouping of two seemingly unrelated topics. The subjects seldom co-occur in one document, and none of the other most-frequent terms seem to connect the two. It could very well be a topic that would have been split up when 37 was chosen as value for K.

3.3 Combined models analysis

3.3.1 Topics in communities. Overlaying the topic model with the network communities shows a large coherence between the two approaches (Figure 2). Two methods using different document properties show that for most of the larger communities in the network, one or multiple counterparts can be clearly distinguished in the topics.

We identify three categories of communities within the network when enriched by the topic model. First are those with distinctive subjects; communities with a clear one-to-one relation to one of the topics. Examples are radio-frequency identification (*RFID*), *smart grid*, *social media/networks*, *vehicular ad hoc networks* and *location data*, for which all publications contained in the community are between 29 and 47% assigned to one specific topic.

Second are topic-sharing communities. These patterns require more interpretation, only to make sense after returning to the citation network and its underlying data. In the network, the communities of *data mining* and *anonymity metrics* are heavily interconnected.

id	Label	Most frequent words
1	Consumer privacy	privacy, trust, consumer, online, study, service, customer, factor, model, concern
2	Smartphones	mobile, device, application, user, apps, android, phone, privacy, app, security
3	Quantum encryption	feature, method, quantum, based, face, privacy, using, homomorphic, vector, encryption
4	Networking	network, wireless, node, sensor, privacy, data, security, routing, communication, aggregation
5	Smart grid	smart, grid, energy, power, privacy, consumption, meter, demand, game, system
6	Image processing	image, digital, data, content, video, hiding, method, proposed, privacy, technique
7	Data mining	data, privacy, mining, preserving, method, sensitive, algorithm, information, technique, anonymization
8	Legal	privacy, right, public, law, surveillance, government, state, article, legal, policy
9	Security operations	attack, security, network, detection, traffic, system, analysis, threat, based, privacy
10	Big data	data, learning, big, machine, analytics, model, library, deep, privacy, training
11	Internet of things	iot, internet, thing, device, security, smart, application, architecture, data, system
12	Multi-party computing	protocol, party, secure, computation, privacy, two, preserving, private, third, multi
13	Privacy-preserving algorithms	privacy, algorithm, query, differential, preserving, data, problem, private, mechanism, result
14	Social media	social, privacy, medium, online, student, self, facebook, study, use, disclosure
15	Digital services	user, privacy, information, service, personal, provider, system, identity, paper, management
16	RFID	protocol, rfid, authentication, security, tag, privacy, system, attack, identification, proposed
17	Biometrics	identity, biometric, authentication, system, biometrics, identification, fingerprint, template, user, security
18	Healthcare (patient privacy)	patient, care, study, privacy, hospital, family, home, nurse, staff, satisfaction
19	Recommender systems	system, healthcare, recommendation, collaborative, privacy, based, filtering, recommender, user, prediction
20	Physical monitoring	monitoring, system, sensor, activity, human, privacy, home, camera, wearable, based
21	Location data	location, privacy, user, based, service, lb, trajectory, query, spatial, mobile
22	Access control	access, control, policy, privacy, based, model, system, agent, data, attribute
23	Blockchain	blockchain, electronic, voting, system, transaction, payment, contract, mail, voter, decentralized
24	Cloud computing	cloud, computing, service, security, privacy, data, environment, resource, issue, model
25	Encryption	encryption, key, algorithm, aes, security, standard, implementation, using, based, cryptography
26	Data privacy	data, privacy, protection, information, personal, regulation, legal, right, law, individual
27	Systems design	security, privacy, system, technology, design, paper, management, research, challenge, issue
28	Ethics	research, ethical, issue, genetic, ethic, consent, data, review, human, researcher
29	Vehicular ad hoc networks	vehicle, network, communication, ad, privacy, message, vehicular, anonymity, hoc, anonymous

(*continued*)

Table 2.
Labelled topics with most frequent words per topic

id	Label	Most frequent words
30	Social networks	social, network, user, privacy, online, sharing, information, profile, content, networking
31	Patient data	health, patient, medical, record, care, information, healthcare, data, electronic, clinical
32	Statistics	privacy, private, information, bound, case, show, function, problem, optimal, model
33	Emperical studies	study, risk, privacy, survey, participant, result, testing, test, attitude, woman
34	Encryption schemes	scheme, key, signature, based, security, authentication, proposed, group, privacy, secure
35	Web applications	web, user, peer, privacy, search, site, website, tool, log, process
36	Cloud storage	data, cloud, encryption, storage, scheme, encrypted, search, user, privacy, server

Table 2.

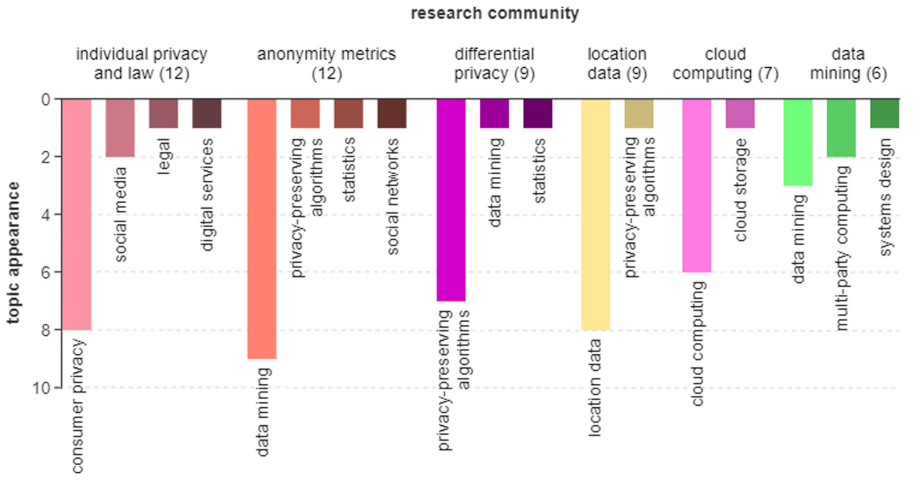

Figure 2.
Average proportion of topic occurrence within network communities

Since *anonymity metrics* was not identified as a separate topic by the LDA algorithm, it is logical to find *anonymization* as one of the terms within the topic of *data mining*.

Third, there are multi-topic communities. Since the communities are not uniform in size, it is to be expected that the larger communities cover multiple subjects. The community of *individual privacy and law* contains literature on *consumer privacy*, *legal* and *social media* according to the topic model. The community of *cloud computing* contains literature on *cloud storage* and *cloud computing*. Often, the topics that are part of larger communities are primarily found in those communities. Since each topic is the same size, it shows its limits of the topic model in community identification when dealing with larger communities. Finally, several topics are a proportion of many of the communities. This is notably the case for the topics of *systems design*, *digital services* and *encryption schemes*. There are multiple possible explanations: publications on this subject could be very distinctive but widely used in different research areas, or the words related to these subjects could be a smaller part of many papers about different topics. It could also indicate latent research communities that may benefit from forming further centralisation.

3.3.2 Most influential publications. The research community and primary topic of the 65 most cited publications in privacy research were extracted. These publications received the most references from within the network over two time periods: all time and the last 5 years in the dataset. Together, they account for 40,340 references (8.7%) in the network. This sub-set contains 15 communities and 19 topics in varying pairs. One publication did not have a topic classification, as its abstract could not be retrieved. 8 out of 15 communities appear only once among the most influential publications. 1 community, *RFID*, appears twice. The final 6 communities stand at a distance, having between 6 and 12 publications. Figure 3 displays the 6 most occurring communities and the topic distributions of the most cited publications within them.

Most notable from the communities in Figure 3 is that their occurrence does not relate to the size of each community alone. Rather, the 6 communities that emerge are the 2 largest communities – *individual privacy and law* and *cloud computing* – and the 4 communities with the highest centrality measures (Table 1). The third and fourth largest communities are notably absent from the most cited publications, with *e-health and medical data* appearing only once and *genetic data* not at all while respectively making up 7.6 and 6.8% of all publications under consideration.

4. Discussion

In this section, we discuss the results to answer the research questions posed in the introduction with the combined results of the network analysis and the topic model. First, we consider the subjects in privacy research and each community for insights on the community structure of the field. Here it shows that the combination of methods offers a cross-section of the structures in the privacy research field, with which we can identify macro communities and potentially latent communities to highlight areas of interest.

Second, the structural position of each community within the network is considered together with the field's most cited publications to identify the central communities in privacy research. In its core we find individual and legal privacy, as well as privacy-enhancing technologies (PETs) in data science, location data and general computing. However, significant gaps appear this central body of knowledge. The translation of individual privacy

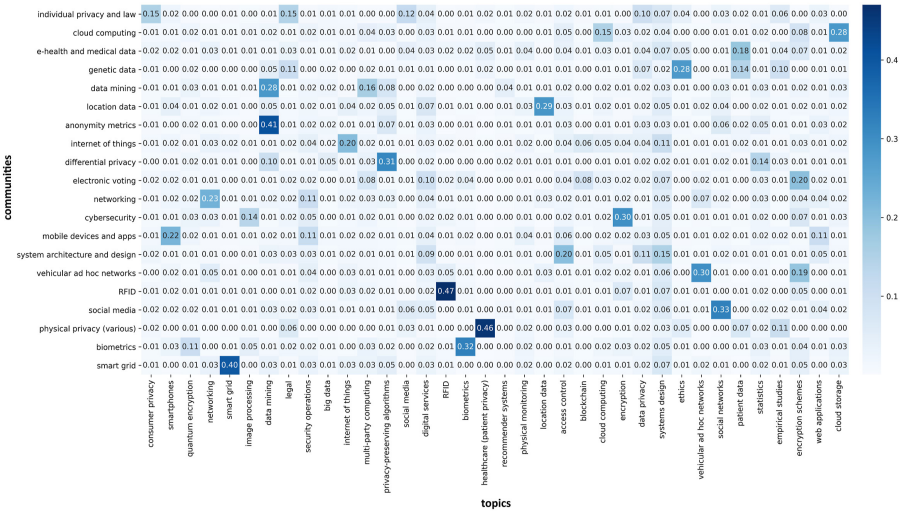

Figure 3.
Topic distribution per community of the 65 internally most cited publications in privacy research ($n > 2$)

concerns to the use of PETs in information systems is not reflected in the core of the field. Organisations are primarily studied from the outside rather than from within (Belanger and Crossler, 2011; Smith *et al.*, 2011), at the risk of treating the behavioural processes within organisations as a black box.

Overall, our analysis of the privacy research field offers new insights that provide more questions than answers. If we are to translate privacy concerns and values into practical solutions, then these gaps need to be addressed. Future studies should take note of the current strengths and shortcomings of the privacy research field in its current state and attempt to bridge the gaps. Finally, we discuss opportunities for future research and limitations in our approach.

4.1 Subjects in privacy research

Anyone who has investigated privacy can attest to it being a multi-disciplinary field. The resolution limit inherent to manual review offers a challenge in obtaining a broad perspective. In this current analysis we sought for an ordering of privacy research through a combination of network and topic analysis. This section discusses *RQ1* . What communities and topics make up the privacy research field?

What stands out immediately is the breadth and connectedness of privacy research. The 83,159 publications formed 94 notable clusters, with subjects ranging from operational healthcare and office planning to smart grids and quantum cryptography. The network analysis and topic model prove to be complementary methods in investigating the makeup of the research field with a strong coherence between their outcomes. Considered as a whole, the results show a diverse research field with communities that have distinct identities but share topics of interest with others.

Three community-topic sets stand out from this perspective. The first regards the community of *cybersecurity*. The community's primary topics are *encryption*, *encryption schemes* and *image processing* in order of proportion. However, topics such as *security operations* and *access control* primarily appear outside of this community. The topic model thus indicates a greater topical breadth of cybersecurity research within privacy than the network analysis alone suggests. As adequate cybersecurity is a pre-requisite for maintaining information privacy (Spiekermann and Cranor, 2009) and with secrecy, security and confidentiality considered to be central aspects of privacy (Smith *et al.*, 2011), this distribution of topics over various communities shows a healthy relationship between privacy and security research.

Then, there are three interconnected communities that share distinct but related research interests, indicating a broader research community than the individual network communities suggest. The communities of *data mining*, *anonymity metrics* and *differential privacy* all have the highest topical proportions of *data mining* and *privacy-preserving algorithms*, albeit in different distributions. Especially noteworthy is that the *data mining* topic is proportionally largest within the community of *anonymity metrics*. While they are classified as distinct communities, their shared topical interests and interconnectedness suggest a larger macro-community in this corner of the network. This perspective is supported by its most cited literature survey, which draws heavily from each of these communities under the name of privacy-preserving data publishing (PPDP) (Fung *et al.*, 2010).

The third result of interest concerns the topic of *system design*, which we would expect to strongly relate to the community of *system architecture and design*. While the community itself is relatively small (3% of the network) and spread out, the topic model suggests that much of the research relevant to this community happens all over the network. Illustrative is the only publication in the most cited publications of the field classified as a *system design*

paper, as it appears in the community of *data science*. One use of the topic model is to discover potentially latent communities in the investigated area (Syed *et al.*, 2019). While it is possible that system design is a distinct topic applied in all of privacy research, the existence of an overlapping community suggests a more structural issue. These results indicate that system design likely is a latent research community in the privacy field, which has yet to form its core theories or centralise around them. The absence of a sizeable and central system design research community is problematic when we consider the challenges faced by organisations globally in implementing Privacy-by-Design to comply with recent and upcoming privacy legislation. The field should take note of this theoretical gap of how privacy is achieved in information systems.

A final observation is that the community of *physical privacy* exists at the fringes of the network in a variety of small clusters, such as operational healthcare, sexuality in nursing homes and office planning. Its primary significance is that, however loosely connected this community may be, the distinction between physical and information privacy is mentioned in many of the field's influential theories (Belanger and Crossler, 2011; Phelps *et al.*, 2000; Smith *et al.*, 2011; Solove, 2006). While some theories consider aspects of physical privacy, most use it to distinguish information privacy as a separate concept. However, aspects of both information and physical privacy do appear in many communities, among them *location data*, *Internet of things*, *vehicular ad hoc networks*, *RFID*, *biometrics* and *smart grid*. For example, a smart speaker can be an intrusion in the physical space when it collects audio data during a fight. The physical nature of these information technologies is absent from the field's influential theories and physical privacy is often not considered at all. With further developments in information technology, this line between physical and information privacy will continue to blur. A monitoring device in the operation room can nowadays be connected to an Artificial Intelligence system that offers the operating team recommendations on physical procedures. A smartphone can continuously monitor someone's location. The disappearing boundaries between physical and information privacy offer a new challenge for privacy researchers that have yet to be represented in its theories.

4.2 Central communities in privacy research

While our approach produced new insights in the makeup of privacy research, a major concern with this dataset is that privacy is rarely considered a field of its own. When comparing the network against the full bibliometric record, it only curtails 14.7% of all references made by the collected publications. This is a significant portion, but also includes research communities with different views towards privacy. Due to the broad inclusion criteria in establishing the bibliometric record, only a sub-set of the identified communities forms the core of the privacy field. Others are more likely to apply the field's theories while still identifying as privacy researchers. Others again will view privacy as a reference discipline. In this section, we interpret the results through answering *RQ2*. How are the identified communities positioned within broader privacy research?

Four communities stand out by their positioning in the network as well as frequent occurrence in the most cited publications. These four communities can be considered the *central communities in privacy research* around which other research organises itself.

(1) *Individual privacy and law.* This is the largest community in the network with 16.2% of all publications. It primarily concerns itself with consumer privacy, the legal aspects of privacy, social media and data privacy. This view is strengthened further by its theoretical contributions, as the most influential theories in privacy research primarily revolve around individual privacy (van Dijk *et al.*, 2021; Smith *et al.*, 2011).

(2) *Cloud computing*. The second-largest community in the network nominally concerns the current computing paradigm in providing Information Technology capabilities to users (Wang *et al.*, 2010). Researchers in this community, however, focus on a more fundamental privacy issue highly relevant to cloud computing: remote and centralised data processing. As such, many of its disciplines are not specific to cloud computing: untrusted storage, databases, encryption and searching encrypted data. Rather, cloud is the paradigm in which these fundamental issues converge.

(3) PPDP is a macro-community consisting of the *data mining*, *anonymity metrics* and *differential privacy*. PPDP is defined as the methods and tools designed to protect the disclosure of any individual's identity while sharing and transporting data (Fung *et al.*, 2010). All three communities share a topical interest in *data mining* and *privacy-preserving algorithms*. As evidenced by the high centrality measures of its individual communities, the macro-community of PPDP holds an essential position within privacy research.

(4) *Location data*. The location data community is primarily concerned with its topical namesake. Its most influential publications relate to k-anonymity (Gedik and Liu, 2008) and privacy guarantees in location-based queries (Ghinita *et al.*, 2008). What differentiates it from the other central communities is its combination of physical and information privacy, as many techniques revolve around the obfuscation of a physical location while still retaining usefulness of information (Beresford and Stajano, 2003). It appears a key subject in privacy research, evidenced by its central position in the network and its usefulness in a wide variety of application domains.

Surrounding these central communities we find applied privacy research. Communities in this category are highly connected to the central communities, such as *cloud computing*, *networking*, *vehicular ad-hoc networks* and *mobile devices and apps*. Each of these communities has mid-range centrality values and strong connections to one or more central communities. While privacy is at the heart of these research communities, they are defined by their application domain.

Then there are communities with little influence in the network. Scoring relatively low on most or all centrality measures, they are less connected to the wider network and thus of less structural importance. Most notable among them are the communities of *e-health and medical data* and *genetic data*. Despite their cumulative 14.4% of publications, they hold little authority within the network. A pattern that was also identified in privacy's most influential publications, where a small number of little referenced publications were found to monopolise the connection of medical research to the broader network (van Dijk *et al.*, 2021). Exploring this divide between medical and broader privacy research may be a fruitful direction for future research.

A notable observation is the absence of a central community regarding organisational privacy knowledge, such as privacy management, privacy governance or a similar term. A wide variety of influences have been prescribed to the organisational context, including but not limited to legislation, organisational policies, (organisational) culture, privacy expertise and the business model (Acquisti *et al.*, 2015; Belanger and Crossler, 2011; Culnan and Armstrong, 1999; Smith *et al.*, 2011; Spiekermann and Cranor, 2009; Warkentin *et al.*, 2011). Privacy professionals, besides operating in this complex environment, are asked to weigh organisational interests, privacy concerns, legal requirements, organisational and technical capabilities, social norms and ethics (Bieker *et al.*, 2016; Clarke, 2009; Raab, 2020; Wright, 2013). However, despite various and frequent calls for more organisational research within the field's most influential publications (Belanger and Crossler, 2011; Smith *et al.*, 2011) and an abundance of research on organisational outcomes (Lanier and Saini, 2008; Smith *et al.*, 2011),

no influential research community on how to manage privacy at an organisational level has risen to the surface. This absence is also apparent in the field's core theories, which concern themselves with privacy concerns, the individual level of analysis and, in one case, system design (van Dijk *et al.*, 2021). There appears to be a scientific knowledge gap in organisational privacy management in a complex organisational context, while grey literature provides a variety of models and tooling (Swartz *et al.*, 2019). New privacy regulations and increased societal privacy awareness have only increased the need for organisational privacy knowledge, making the absence of such a community even more apparent.

The positioning of legal privacy research next to, and in the same community as, individual privacy research, furthers concerns on the absence of a central community for organisational privacy management. While individual concerns undoubtedly need to be understood and addressed for privacy regulations to be effective, the close relationship between research on the individual level of privacy and legal privacy appears to colour the perspective of legal researchers. Its most influential publications concern themselves with individual privacy harms (Solove, 2006), privacy as contextual integrity (Nissenbaum, 2004) and the necessity of privacy to individual development (Cohen, 2012). However, understanding organisational privacy management is required for the design of effective interventions in the application of privacy legislation. One example of conflict between organisational reality and legislation is the European Union's General Data Protection Legislation (GDPR), which is intentionally technology-agnostic. A summary of unforeseen complications in the implementation of the GDPR finds conflicts in the areas of backups, non-volatile storage, auditability, biometrics, data integrity and increased costs for certain processing activities (Politou *et al.*, 2018). While this cannot be classified as an oversight of existing research, legal privacy researchers should be aware that we currently lack a shared understanding of the domain of organisational privacy management, factors influencing privacy decision-making and what effective treatments can be applied.

4.3 Future research and limitations

This research contributes to our overall understanding of the privacy research field. Its primary data source, Scopus, is considered a reliable source for bibliometric data (Baas *et al.*, 2020), though it provides several limitations. First, various publications have no indexed references or lack an abstract, limiting their usefulness in or excluding them altogether from the analyses performed. Second, the earliest paper was published in 1965 and thus does not include influential publications as least as far back as 1890 (Warren and Brandeis, 1890). The impact of this missing data on the network and topic model has not been quantified. Keeping these limitations in mind, the outcomes of this research provide an overview of current privacy research. Future research could expand the data sources used to gain a more complete picture of privacy research over time and further investigate the temporal developments of the field.

As can be expected of a mapping study, its outcomes also raise new questions. In considering privacy as a singular research field, a number of findings ask for further exploration. First is the presence of system design research as a potentially latent community. It requires in-depth investigation to assess whether this is indeed a latent community within privacy research or rather a distinct and influential reference discipline. Then, the lack of influence and connection of medical research to the broader research field. There is a disconnection between the medical sciences and other privacy research that cannot be explained within the scope of this research. Finally, the absence of organisational privacy management in the network, the topic model and the most cited publications provides what may be the currently most significant knowledge gap in privacy research.

After intermediate presentation of the results we were approached by several scholars interesting in repeating our approach to bibliometric analysis in their own domains. We invite interested researchers in re-using and expanding on this methodology and software used (van Dijk, 2021; Gadellaa, 2021).

5. Conclusions

The goal of this research was to obtain insights in the structure of the privacy research field. With a combination of network analysis and topic modelling, we analysed both the structural and topical makeup of the field. The network analysis shows privacy as a heterogeneous field, divided into 20 significant research communities consisting of one or more disciplines. The topic model divided the field into 36 distinct topics of equal size. Overlaying these outcomes provided a fruitful ground for further analysis. The combination of techniques allowed us to identify three central communities of the field: individual privacy and law, cloud computing, PPDP and location data. PPDP is a macro-community identified through a topical overlap as well as the community's literature surveys, and is made up of the communities of data mining, anonymity metrics and differential privacy. The topic model further shows a healthy presence of cybersecurity research throughout the network, not only in communities with a direct relation to the topic.

An assessment of the combined results also provided insight in what is missing from this picture. First, the topic model identified systems design research as a topic of interest in almost all research communities, rather than centralised in the community of system architecture and design. This may indicate a latent community that could benefit from further centralisation. Second is the apparent disconnect between medical privacy research and the rest of the network. Despite the two largest communities containing 14.4% of publications in the dataset, they do not possess a significant position in the network. Finally, we considered the absence of organisational privacy management research in the network, topic model and the most cited publications. While it is frequently mentioned, there appears to be no influential research on this subject. This hampers the design of effective interventions at the place personal data is processed.

Altogether, analysing the bibliometric record by combining network analysis with topic modelling has proven to be a fruitful ground for investigating the privacy research field. The quantitative nature of these methods has allowed us to analyse the overarching structures of the field from more than 80,000 publications, furthering our understanding of the multi-disciplinarity, makeup and deficiencies of the privacy research field.

References

Abbasi, A., Altmann, J. and Hossain, L. (2011), "Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures", *Journal of Informetrics*, Vol. 5 No. 4, pp. 594-607.

Acquisti, A., Brandimarte, L. and Loewenstein, G. (2015), "Privacy and human behavior in the age of information", *Science*, Vol. 347 No. 6221, pp. 509-514.

Baas, J., Schotten, M., Plume, A., Côté, G. and Karimi, R. (2020), "Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies", *Quantitative Science Studies*, Vol. 1 No. 1, pp. 377-386.

Belanger, F. and Crossler, R. (2011), "Privacy in the digital age: a review of information privacy research in information systems", *MIS Quarterly*, Vol. 35, pp. 1017-1041.

Beresford, A.R. and Stajano, F. (2003), "Location privacy in pervasive computing", *IEEE Pervasive Computing*, Vol. 2 No. 1, pp. 46-55.

Bieker, F., Friedewald, M., Hansen, M., Obersteller, H. and Rost, M. (2016), "A process for data protection impact assessment under the European general data protection regulation", *Privacy Technologies and Policy*, Springer International Publishing, Cham, Vol. 9857, pp. 21-37.

Bird, S., Klein, E. and Loper, E. (2009), *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, Sebastopol, CA.

Blei, D.M. and Lafferty, J.D. (2006), "Dynamic topic models", *Proceedings of the 23rd International Conference on Machine Learning*, Association for Computing Machinery, New York, NY, pp. 113-120.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *The Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008), "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008 No. 10, P10008.

Boyd-Graber, J., Mimno, D. and Newman, D. (2014), "Care and feeding of topic models: problems, diagnostics, and improvements", *Handbook of Mixed Membership Models and Their Applications*, Vol. 225255.

Clarke, R. (2009), "Privacy impact assessment: its origins and development", *Computer Law and Security Review*, Vol. 25 No. 2, pp. 123-135.

Cohen, J.E. (2012), "What privacy is for", SSRN Scholarly Paper No. ID 2175406, Social Science Research Network, Rochester, NY, available at: https://papers.ssrn.com/abstract=2175406 (accessed 13 July 2018).

Culnan, M.J. and Armstrong, P.K. (1999), "Information privacy concerns, procedural fairness, and impersonal trust: an empirical investigation", *Organization Science*, Vol. 10 No. 1, pp. 104-115.

De Mey, M. (1982), *The Cognitive Paradigm*, D. Reidel Publishing Company, Dordrecht, Holland.

Denny, M.J. and Spirling, A. (2018), "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it", *Political Analysis*, Vol. 26 No. 2, pp. 168-189.

Elgesem, D., Steskal, L. and Diakopoulos, N. (2015), "Structure and content of the discourse on climate change in the blogosphere: the big picture", *Environmental Communication*, Vol. 9 No. 2, pp. 169-188.

Fung, B.C.M., Wang, K., Chen, R. and Yu, P.S. (2010), "Privacy-preserving data publishing: a survey of recent developments", *ACM Computing Surveys*, Vol. 42 No. 4, pp. 1-53.

Gadellaa, J.F. (2021), "Topical analysis of privacy literature, with an application on citation network interpretation", available at: https://github.com/JoostGadellaa/capita-selecta

Gedik, B. and Liu, L. (2008), "Protecting location privacy with personalized k-anonymity: architecture and algorithms", *IEEE Transactions on Mobile Computing*, Vol. 7 No. 1, pp. 1-18.

Ghinita, G., Kalnis, P., Khoshgozaran, A., Shahabi, C. and Tan, K.-L. (2008), "Private queries in location based services: anonymizers are not necessary", *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 121-132.

Grandjean, M. (2016), "A social network analysis of Twitter: mapping the digital humanities community", Edited by Mauro, A., *Cogent Arts and Humanities*, Vol. 3 No. 1, doi: 10.1080/23311983.2016.1171458.

Grimmer, J. and Stewart, B.M. (2013), "Text as data: the promise and pitfalls of automatic content analysis methods for political texts", *Political Analysis*, Vol. 21 No. 3, pp. 267-297.

Haselmayer, M. and Jenny, M. (2014), "Measuring the tonality of negative campaigning: combining a dictionary approach with crowd-coding", *Political Context Matters: Content Analysis in the Social Sciences*, Mannheim.

Hoffman, M.D., Blei, D.M. and Bach, F. (2010), "Online learning for latent Dirichlet allocation", *In Advances in Neural Information Processing Systems*, NIPS '10, Vol. 23.

Huang, J. (2005), "Maximum likelihood estimation of dirichlet distribution parameters", *CMU Technique report*, Vol. 18.

Jacobi, C., van Atteveldt, W. and Welbers, K. (2016), "Quantitative analysis of large amounts of journalistic texts using topic modelling", *Digital Journalism*, Vol. 4 No. 1, pp. 89-106.

Lanier, C.D. and Saini, A. (2008), "Understanding consumer privacy: a review and future directions", *Academy of Marketing Science Review*, Vol. 12 No. 2, p. 48.

Leino-Kilpi, H., Välimäki, M., Dassen, T., Gasull, M., Lemonidou, C., Scott, P. and Arndt, M. (2002), "Privacy: a review of the literature", *International Journal of Nursing Studies*, Vol. 38, pp. 663-671, doi: 10.1016/S0020-7489(00)00111-5.

Leydesdorff, L. (2007), "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 9, pp. 1303-1319.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., *et al.* (2018), "Applying LDA topic modeling in communication research: toward a valid and reliable methodology", *Communication Methods and Measures*, Vol. 12 Nos 2-3, pp. 93-118.

Maier, D., Niekler, A., Wiedemann, G. and Stoltenberg, D. (2020), "How document sampling and vocabulary pruning affect the results of topic models", *Computational Communication Research*, Vol. 2 No. 2, pp. 139-152.

Miller, G.A. (1995), "WordNet: a lexical database for English", *Communications of the ACM*, Vol. 38 No. 11, pp. 39-41.

Moody, D., Iacob, M.-E. and Amrit, C. (2010), *Search of Paradigms: Identifying the Theoretical Foundations of the IS Field", Proceedings of the 22nd European Conference on Information Systems*, Pretoria, pp. 1-13.

Newman, M. (2010), *Networks: an Introduction*, 2nd ed., Oxford University Press, Oxford, New York.

Nissenbaum, H. (2004), "Privacy as contextual integrity", *Washington Law Review*, Vol. 79, p. 41.

Phelps, J., Nowak, G. and Ferrell, E. (2000), "Privacy concerns and consumer willingness to provide personal information", *Journal of Public Policy and Marketing*, Vol. 19 No. 1, pp. 27-41.

Politou, E., Alepis, E. and Patsakis, C. (2018), "Forgetting personal data and revoking consent under the GDPR: challenges and proposed solutions", *Journal of Cybersecurity*, Vol. 4 No. 1, doi: 10.1093/cybsec/tyy001.

Raab, C.D. (2020), "Information privacy, impact assessment, and the place of ethics", *Computer Law and Security Review*, Vol. 37, 105404.

Rehurek, R. and Sojka, P. (2011), *Gensim–Python Framework for Vector Space Modelling*, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Vol. 3 No. 2.

Röder, M., Both, A. and Hinneburg, A. (2015), "Exploring the space of topic coherence measures", *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, New York, NY, Association for Computing Machinery, pp. 399-408.

Rose, M.E. and Kitchin, J.R. (2019), "Pybliometrics: scriptable bibliometrics using a Python interface to Scopus", *SoftwareX*, Vol. 10, 100263.

Scott, S. and Matwin, S. (1999), "Feature engineering for text classification", *ICML*, Citeseer, Vol. 99, pp. 379-388.

Sievert, C. and Shirley, K. (2014), "LDAvis: a method for visualizing and interpreting topics", *Proceedings of the Workshop on Interactive Language Learning*, Visualization, and Interfaces, pp. 63-70.

Smith, H.J., Dinev, T. and Xu, H. (2011), "Information privacy research: an interdisciplinary review", *MIS Quarterly*, Vol. 35, pp. 989-1015.

Solove, D.J. (2006), "A taxonomy of privacy", *University of Pennsylvania Law Review*, Vol. 154 No. 3, p. 477.

Spiekermann, S. and Cranor, L.F. (2009), "Engineering privacy", SSRN Scholarly Paper No. ID 1085333, Social Science Research Network, Rochester, NY.

Swartz, P., Da Veiga, A. and Martins, N. (2019), "A conceptual privacy governance framework", *2019 Conference on Information Communications Technology and Society*, IEEE, pp. 1-6.

Syed, S. (2019), "Topic discovery from textual data: machine learning and natural language processing for knowledge discovery in the fisheries domain", March, available at: https://dspace.library.uu.nl/handle/1874/374917 (accessed 5 February 2021).

Syed, S. and Spruit, M. (2017), "Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation", *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 165-174.

Syed, S. and Spruit, M. (2018), "Exploring symmetrical and asymmetrical dirichlet priors for latent dirichlet allocation", *International Journal of Semantic Computing*, Vol. 12 No. 03, pp. 399-423.

Syed, S., Borit, M. and Spruit, M. (2018), "Narrow lenses for capturing the complexity of fisheries: a topic analysis of fisheries science from 1990 to 2016", *Fish and Fisheries*, Vol. 19 No. 4, pp. 643-661.

Syed, S., Aodha, L., Scougal, C. and Spruit, M. (2019), "Mapping the global network of fisheries science collaboration", *Fish and Fisheries*, Vol. 20, doi: 10.1111/faf.12379.

van Dijk, F. (2021), "Network analysis of literature", available at: https://github.com/FrisovanDijk/bibliometric-privacy-network-analysis

van Dijk, F., Spruit, M., van Toledo, C. and Brinkhuis, M. (2021), "Pillars of privacy: identifying core theory in a network analysis of privacy literature", *Proceedings of the Twenty-Ninth European Conference on Information Systems*, (ECIS 2021), available at: https://aisel.aisnet.org/ecis2021_rp/84

Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y. and Theodoridis, Y. (2004), "State-of-the-art in privacy preserving data mining", *ACM SIGMOD Record*, Vol. 33 No. 1, pp. 50-57.

Wallach, H., Mimno, D. and McCallum, A. (2009), "Rethinking LDA: why priors matter", *Advances in Neural Information Processing Systems*, pp. 1973-1981.

Wang, L., von Laszewski, G., Younge, A., He, X., Kunze, M., Tao, J. and Fu, C. (2010), "Cloud computing: a perspective study", *New Generation Computing*, Vol. 28 No. 2, pp. 137-146.

Warkentin, M., Johnston, A.C. and Shropshire, J. (2011), "The influence of the informal social learning environment on information privacy policy compliance efficacy and intention", *European Journal of Information Systems*, Vol. 20 No. 3, pp. 267-284.

Warren, S.D. and Brandeis, L.D. (1890), "Right to privacy", *Harvard Law Review*, Vol. 4, p. 193.

Wright, D. (2013), "Making privacy impact assessment more effective", *The Information Society*, Vol. 29 No. 5, pp. 307-315.

Corresponding author
Friso van Dijk can be contacted at: f.w.vandijk@uu.nl