Correction notice

In Section 2.5, it is indicated that the model with the highest C_v coherence score was chosen for analysis. Unfortunately, it turned out that the wrong model was manually loaded for analysis. As explained in Section 2.4 the algorithm used in this paper has the option for it's user to specify a so-called random state: a seed that the model uses for the required randomness. While processing the results from the model search phase, two random states got mislabeled, resulting in the wrong model being selected for analysis. The number of topics is equal between the two models, but the chosen seed was different and so was the coherence. The best model had a C_v of 0.517, while the wrongfully chosen model scored 0.504.

In terms of topics, the two models differ in various ways. Comparing the two, 26 of the topics overlap in a one-to-one relation. For the other 10 topics there are differences. Some examples are given, from the perspective of the 'right' model, i.e., the one with a higher C_v coherence: The two topics that were labelled as *Social networks* and *Social media use* seems to have 'merged' into one, so have the topics of *Cryptography* and *Signature*. Additionally, this model uncovered a specific topic with terms like *image* and *video* as distinguishing terms, as well as a topic with *blockchain*, *voting*, and *payment*. These terms were previously contained in different topics. On the other hand, the topics of *Big data*, *Office design* and *Application security* are not uncovered as such by the 'right' model. This once again signals the somewhat arbitrary nature of a topic model. As mentioned in Section 3.2, every model we choose to analyse is one of various possible models with a different representation of the underlying documents.

Because this type of modelling is very susceptible to changes in factors like the random state, the discussion in Section 4 already explicitly avoids drawing conclusions that would not be generalizable to different models. At a first glance, partly repeating the analysis of Section 3.5 shows that the specific relations between communities and topics change as an effect of the new topic model, but the general pattern of some 1-on-1 relations, larger communities covering multiple topics, and differences of which we cannot know the reasons hold. Because the observations in this paper are not on the level of specific patterns in the privacy literature, but mainly on this method's usage and usefulness, all findings remain valid.

Topical Analysis of Privacy Literature, with an Application on Citation Network Interpretation

Joost Gadellaa

Universiteit Utrecht, Heidelberglaan 8, 3584CS Utrecht
j.f.gadellaa@students.uu.nl

Abstract. Privacy is a versatile field of research, studied in a multitude of disciplines. A novel way to explore how research concepts like these are embedded in scientific literature is through the use of topic models. This study aims to investigate the usefulness of such modelling techniques. We use a latent Dirichlet allocation (LDA) topic model to analyse the title, abstract and keywords of 83,159 works in the research field of privacy, evaluating the model both on its own as well as in comparison to a previously made citation network analysis. We show that, although the resulting topic model can be meaningfully labeled, interpretation yields little insights. When comparing the model to the citation network analysis, remarkable similarities between the resulting classifications of these very different methods show. However, many methodological caveats pose a threat to conclusion validity, and computational costs limit possibilities for additional exploration. This method needs further research in order to generate insights, instead of just being an accurate representation of the data.

Keywords: Privacy · Topic model · LDA · Privacy · Network Analysis · Bibliometrics

1 Introduction

Privacy is a broad concept studied in various disciplines of science. The scientific discussion about the topic originates when the "Right to be let alone" was recognized "in common law, in its eternal youth, [...] to meet the demand of society" [39]. Since then, it has become an increasingly multidisciplinary concept studied by medicine, psychology, engineering, computer science, and law scholars amongst others. In a time when data-driven technological possibilities are developing fast, people are more and more concerned about their personal privacy [19]. The topic seems more pressing than ever.

The scattered nature the privacy research field creates a need for mapping studies. Understanding privacy's scientific discourse is essential to aid literature reviews, facilitate interdisciplinary sharing of knowledge, identify lacunae, and develop policy supporting the research field's development [26]. Future research paths are hard to identify and a clear direction more challenging to articulate without a coherent connection between the different theoretical understandings, practical artifacts, and empirical models.

A relatively novel approach to conduct mapping studies is through the usage of topic models. Topic modeling is an area of unsupervised machine learning in which text corpora are modeled by means of their latent topics: groups of co-occurring words associated with a single theme [10]. This method uncovers similarities between documents in the corpus and describes the documents using these relations without a need for categorization by human annotators. This method has found its way from computer science and linguistics to less technical fields like communication research [22], fisheries science [36], and journalism [18]. Topic models have also shown some promising results for the purpose of automatic discovery of sub-fields in scientific literature [25], which will be the focus of this analysis.

Taking an approach strongly focused on methods, this analysis' goal is to explore the value of topic models for the use of understanding privacy's scientific discourse. The main research question is: How can topic models contribute to our understanding of privacy's embedding in the scientific literature. Two sub-goals will be introduced below: creating and exploring a topic model, and comparing it to a citation network analysis of the same data.

The first goal is to create a coherent topic model that captures the diverse nature of privacy as a scientific research field. Using this model, we might be able to answer exploratory questions on the existence of different sub-fields and relations between them. Special attention will be on the method design. Although the existence of easy-to-use open-source libraries like Gensim [28] and Mallet [23] have enabled many scholars to employ topic models in novel ways, the method's underlying complex statistical nature may put scientific rigor at stake when not utilized properly. Researchers often treat these kinds of algorithms as 'black boxes', without thoroughly exploring and tuning underlying parameter values [7]. Therefore, the creation of a topic model for analysis must not be secondary to its interpretation.

The second goal is to compare the then discovered topic model to a network analysis of the same data by Van Dijk [9]. Their quantitative analysis and its translation to qualitative identification of the theoretical foundation of privacy as a research field is a prime example of what this kind of research can contribute. Seeing what similarities and differences between a classification based on citation communities and one on latent topics exist could tell us more about the value of these methods. The topic model might additionally be used to interpret or label citation communities more accurately or objectively. Above all, it stimulates discussion about the research field and the application of novel methods.

This paper will explore how topic models can contribute to our understanding of privacy's embedding in the scientific literature. We will elaborate on the methodology in Section 2 including a short non-technical description of LDA topic models. After a valid model has been found, it will be interpreted through its topic labels, topic correlation, and a comparison to the citation network analysis in Section 3. Section 4 will reflect on the research goals and the results' consequences for the research question and possible directions for future research.

2 Methods

2.1 Latent Dirichlet Allocation

In the context of bibliometric analyses and in general, one of the most frequently used topic models is the Latent Dirichlet Allocation (LDA) [4]. 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. 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 [12]. By capturing the heterogeneity of research topics that a specific paper can belong to, LDA overcomes the limitations of other topic models like Latent Semantic Indexing (LSI) [11] and pLSI [16]; besides the main topic it is assigned to, a paper covers multiple topics in different proportions [35]. 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 a heavy simplification from how humans understand text, it is well-founded for the purpose of uncovering latent structures [3].

2.2 Data retrieval

For this analysis, bibliometric data on the field of privacy gathered by Van Dijk [9] is used and extended. The set consists of 119,710 published works with privacy in their title, abstract, author-specified keywords, or publisher keywords. The data is further filtered on having at least one reference to or from another work in the set. This results in a subset of 83,159 works. This data was extended by querying the Scopus API for the corresponding keywords and abstracts using Pybliometrics [29]. Here, only keywords added by the author(s) were included in the data set to prevent the Topic Model from learning from previous categorizations, be it automatic or human. Only abstracts, in contrast to full texts, are considered for computational feasibility reasons. For large document collections like these, the benefit from the inclusion of more text is negligible [35]. Scopus was used because of its high quality and advanced curation of sources to include [1].

2.3 Data processing

One of this analysis' main goals is to create a topic model in a reproducible and objective way. As Maier *et al.* [22] stated, two of the main methodological steps that need to be taken to reach this goal are appropriate pre-processing and adequate selection of model parameters. It is a downside of LDA that these steps can influence results in a significant way and that their order matters [8]; consecutive steps are dependent on previous results. The follow in appraoch is chosen, closely following recommendations by Maier et al..

The first step of pre-processing consists of the cleaning of data. Copyright notices are removed from the abstracts using a regular expression to prevent the topic distribution from being influenced by the organization that published the work. Afterward, we take the following steps in this specific order: Tokenization, converting to lowercase, punctuation removal, stop-word removal, lemmatization and stemming, and relative pruning.

The sentences are split into a list of words, also known as tokens. Using the Gensim package [28] for Python, this step is easily combined with converting all letters to lowercase and removing punctuation. Converting to lowercase helps to unify the terms, and removing punctuation reduces complexity. This pre-processing step can be considered the removal of uninformative data in the bag-of-words approach used here [32].

All further steps can be considered 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 first of these is stop-word removal: Using NLTK's English stop-word list [2], we remove common stop words like *the*, *is*, and *are*.

After that, it is common practice to stem or lemmatize all words. Stemming is a procedure in which derivational affixes are cut off to reduce all words with the same stem to a common form using several set heuristics [20]. We choose lemmatization using WordNet [24] instead, which is less aggressive and enough feature-reduction for English texts [14]. This procedure only collapses inflicted forms to one term instead of many related words. Most importantly, this keeps nouns and verbs separate tokens, closer to the way a human interprets language.

The last of the feature selection steps is relative pruning. This step consists of removing words that either appear in very little or in almost all documents, sometimes referred to as 'corpus specific stop words'. These words do not contribute to the aforementioned distinctive distribution patterns, and removing them makes model computation less resource-intensive. 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 standard [13, 8] and recently, Maier *et al.* showed that such percentages do not decrease model quality [21]. We remove all words that appear in less than 0.5% or more than 99%. These percentages are very much on the safe side but still decrease the number of unique terms by around 95%, enabling the computation of more models during the model search phase.

To illustrate these steps, the beginning of an abstract by Perera *et al.* [27] is given below. Of these words, none were later removed by relative pruning.

> *"© 2019 Elsevier Inc. Internet of Things (IoT) applications typically collect and analyse personal data that can be used to derive sensitive information about individuals."*

```
['internet', 'thing', 'iot', 'application', 'typically',
'collect', 'analyse', 'personal', 'data', 'used', 'derive',
'sensitive', 'information', 'individual']
```

2.4 Parameter tuning

After cleaning 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.* [36], who performed a very similar analysis on fisheries science literature. First and foremost, the optimal number of topics (K) has to be found. 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 [35]. In some topic model applications, there are reasons to infer the 'right' number of topics from theory or the analysis' goals. Still, in most cases, the optimal number has to be discovered by generating different models and measure their quality according to a certain metric. There is no 'true' number, just a number that produces fitting results according to a chose evaluation metric. The metric of quality for which the model is optimized will be discussed in the next section.

Apart from the number of topics, the Dirichlet prior distribution on the topic probabilities within documents (α) and on the word probabilities within topics (η) have to be set. The most important choice here is if the topics are expected to have an equal probability of being assigned (a symmetrical prior) or not (an asymmetrical prior). Since in text corpora like these, it is to be expected that some topics are less common than others, an asymmetrical prior is preferred [38, 37]. Gensim's 'online' LDA algorithm [15] can use the Newton-Rapson method [17] to learn these priors from the data. 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.

If computational resources allow for it, the random state, the number of passes through the corpus during training, and the number of iterations for inferring the topic distribution should also be tuned. This is not necessary for each level of K, but the convergence profile might differ between versions with different parameters. Currently, 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 [5]. In this analysis, we will start with two random states and add more when the grid search has narrowed the possible values for other variables down to a computationally manageable level. Exact values are specified in the accompanying code repository[1] for reproducibility purposes.

2.5 Measuring of model quality

Any hyper-parameter optimization procedure needs a metric to optimize for. Perplexity [4, 31], a measurement of goodness of fit on held-out data, is often used for this purpose, but there is no evidence that a good fit aids the interpretability of a topic model by humans[5]. Worse still, perplexity might negatively correlate with human judgments of goodness of fit [6]. Röder *et al.* [30] systematically

[1] https://github.com/JoostGadellaa/Topical-Analysis-of-Privacy-Literature

explored alternatives, and found the combined measure of C_v coherence to show the highest correlation with human understanding on a large variety of test data. This measure is based on the idea that topics have a higher quality when it's top words are more related. Again building on and thorough exploration of alternatives in a similar context by Syed [34], we adopt C_v as our measurement of model quality.

2.6 Topic labeling

The final part of our topic-model pipeline consists of labeling the topics uncovered by the algorithm. Although human labeling introduces a degree of subjectivity to the process, it is necessary for understanding and reasoning about the obtained model. With automatic methods, like selecting and filtering the most frequent words with feature selection techniques, the downside occurs that the results can be too general and do not account for semantic relationships between words. For example, frequent words can be *red, white* and *blue*, but the best descriptor for this topic would be *colours*. Sometimes, external sources are used for a more semantic approach. Although different algorithms have been developed, they are seldom validated, let alone compared in a structured and rigid manner. As long as the introduced subjectivity is kept in mind during the result evaluation, manual labeling is still the most pragmatic option.

The labeling will be done by two annotators, who separately review each topic's top-words, the most representative papers per topic (i.e., documents for which the topic distribution has the highest proportion of the current topic), and a visualization using LDAvis [33]. The latter serves to judge relative differences in word frequency and add context.

3 Results

3.1 Processed data

Since Van Dijk's [9] query of literature mentioning privacy, 126 works have been retracted or otherwise removed. Of the 83,033 works left, 24,982 are without keywords, of which 3534 without abstract. Inspecting a sample shows that these are primarily complete books and conference introductions. 67 works are missing just abstract but have keywords. All documents without abstract are removed since the lack of data would not allow us to properly assign these works a distribution of topics. We don't have a reason to believe this removal causes bias. Copyright notices were removed with a custom regex available in the code. The above-described relative pruning steps reduced the number of terms from 75148 to 2421; 96.78% 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 turned out to be because more than 1% of documents were added to the database because *privacy* was in the keywords added by the publisher. These publisher keywords were not included in this analysis, as described in Section 2.2.

3.2 Model search

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 covariance measurements to be less sensitive to changes in topic numbers in the higher regions. The first search across two random states showed that covariance was maximal somewhere between 25 and 49. The search was then deepened to include a total of four random states and additional values for K in between the values in the original sequence. Further exploration of interactions between variables, as described in Section 2.4 was not possible: the current model search already took many days on a regular PC. Nevertheless, the most important variables could be varied in a structured manner, generating a total of 56 models to evaluation.

Most models in this range had a C_v value between 0.32 and 0.49, showing a clear pattern where C_v scores increased sharply for the initial 'increase' in K. After the peak, C_v scores slowly decreased for models with more topics. The highest scoring model was one uncovering 36 topics, with a C_v score of 0.517. This number of topics was not always the local maximum across the different random states, suggesting an interaction between the random state and the effect of K. The fact that the ideal number of topics is not the same for all random states can be interpreted as evidence that multiple coherent topic models exist. Although the highest scoring model is chosen for interpretation, we have to keep in mind that we are looking at one of many fitting topic models.

3.3 Topic labels

Out of 36 labels, 26 contained only syntactic differences between the two annotators. The last 10 were agreed upon without the need for a third annotator. A textual representation of the final topic model can be found in Table 1.

Table 1: Labeled topics with most frequent words per topic

	Label	**Most frequent words**
1	Healthcare	patient, care, health, study, woman, result, method, service, hiv, privacy
2	Legal	privacy, right, law, surveillance, public, protection, legal, state, article, individual
3	Privacy algrorithms	algorithm, method, problem, privacy, graph, result, data, model, proposed, based
4	Data mining	data, privacy, mining, preserving, sensitive, information, anonymity, anonymization, technique, method
5	Web services	user, service, privacy, web, system, context, agent, based, information, management
6	Software design	security, privacy, system, requirement, paper, design, issue, framework, research, threat

7	Consumer perception	privacy, consumer, online, trust, study, concern, service, customer, factor, perceived
8	Government IT	technology, government, issue, information, challenge, new, public, development, management, digital
9	Social media use	social, privacy, medium, study, student, use, self, people, research, communication
10	Cloud storage	data, cloud, storage, image, privacy, encryption, user, security, access, file
11	Cryptography	key, scheme, authentication, protocol, security, proposed, attack, secure, communication, based
12	Big data	data, big, research, protection, personal, privacy, collection, analytics, use, sharing
13	Patient data	health, patient, medical, record, healthcare, information, care, system, electronic, access
14	Encryption	encryption, search, scheme, secure, encrypted, computation, privacy, homomorphic, server, preserving
15	Privacy-preserving protocols	protocol, party, privacy, secure, voting, proof, computation, two, knowledge, transaction
16	Social networks	user, social, privacy, network, online, information, site, profile, facebook, sharing
17	Location	location, privacy, user, based, service, lb, trajectory, information, protection, spatial
18	Signature	scheme, signature, based, attribute, access, group, user, key, anonymous, control
19	Machine learning	learning, privacy, data, model, machine, based, classification, preserving, accuracy, recommendation
20	Privacy-preserving controls	privacy, data, policy, model, control, access, protection, based, preserving, paper
21	Genetics	research, genetic, ethical, consent, testing, ethic, study, test, issue, participant
22	Network security	attack, detection, traffic, system, network, channel, security, analysis, based, threat
23	Cloud computing	cloud, computing, service, security, data, resource, privacy, environment, issue, edge
24	Internet of things	internet, iot, device, thing, security, privacy, application, blockchain, based, technology
25	Vehicular ad hoc networks (and AES)	vehicle, aes, ad, vehicular, encryption, hoc, security, standard, network, implementation
26	Personal privacy risks	information, risk, privacy, sharing, child, personal, disclosure, leakage, parent, private
27	Application security	application, system, security, code, software, analysis, control, flow, based, access
28	Mobile device	mobile, user, device, application, apps, privacy, phone, app, smartphone, android

29	Networks	data, network, node, sensor, privacy, wireless, aggregation, distributed, source, scheme
30	Office design	design, space, privacy, environment, quality, study, speech, pattern, satisfaction, office
31	Monitoring devices	sensor, monitoring, system, video, activity, wearable, detection, device, camera, signal
32	Peer-to-peer networks	network, social, privacy, peer, user, trust, communication, content, networking, based
33	Biometrics	identity, biometric, authentication, system, user, privacy, biometrics, security, card, fingerprint
34	Databases	query, privacy, database, mechanism, differential, private, auction, user, preserving, game
35	Smart grid	smart, grid, energy, privacy, power, city, home, system, consumption, meter
36	RFID	rfid, tag, identification, protocol, privacy, system, security, authentication, radio, frequency

A few topics stand out. Topic 25 seems to include both the very specific topic of vehicular ad hoc networks and the topic of AES. 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. The fact that such edge-cases can always exist could be seen as a downside of this method. Topic 35 illustrates another peculiarity of LDA: since it focuses on words without context, smart grids, smart cities, smart homes, smart meters, and smart contracts seem to have ended up in the same topic. We cannot know what the topical distribution would have been if the algorithm had considered more context. Still, it illustrates that LDA can recognize only half of the difference between a smart city and a smart contract in terms of meaning.

3.4 Topic correlation

To get a sense of relations between the different topics, we can make a (Pearson) correlation matrix of the occurrence of topics within the documents, which can be found in Figure 1. The overall impression this gives is good: Topics seldom correlate much, and if they do, the coefficient is often negative. This is a sign of good separation between them. Where topics positively correlate the strongest are understandable. The high correlation between cloud computing and cloud storage, for instance, is typical for this method. Since the distribution is based on words, a single word relating to multiple real-world subjects will make topics hard to distinguish. In this case, other words can be used to differentiate between the topics, like the frequently occurring but not co-occurring *storage* and *computation*, see table 1, correlations like these are to be expected.

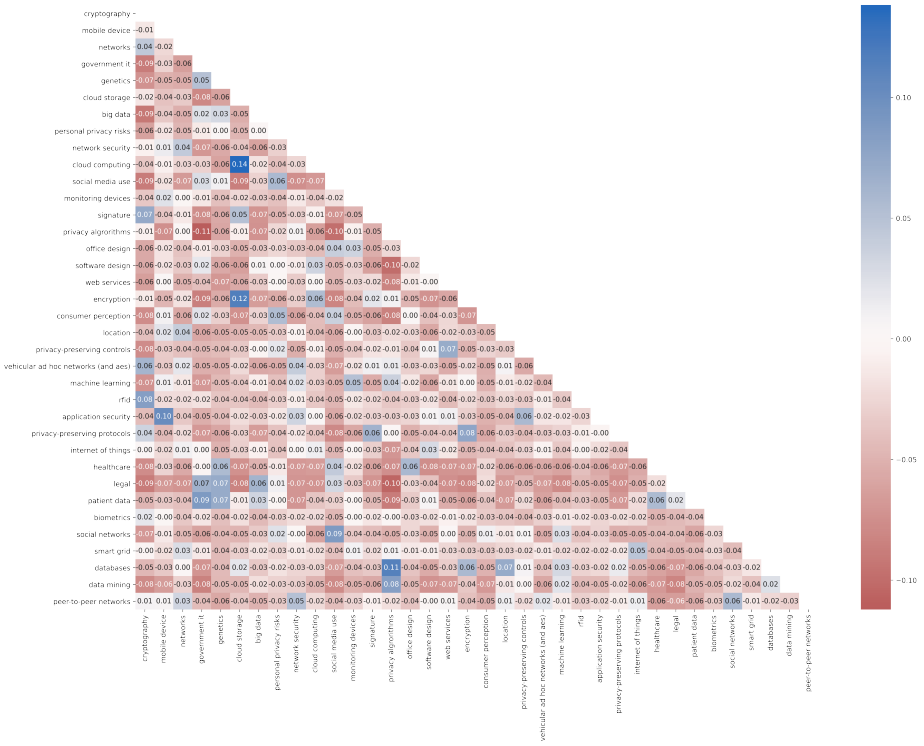

Fig. 1. Pearson correlation between topics

3.5 Comparison to citation network

We compare the model's assigned distribution of topics for each paper to the classification made by dividing the data into communities using the Louvain method, as done by Van Dijk [9]. To aid this analysis, we visualize the relative distribution of topics within each community as a heatmap. Numbers shown in Figure 2 are the average proportions for topic occurrence within each community. Communities on the y-axis are ordered from largest to smallest. From *individual privacy and law* containing about 16.2% of publications to *smart grid* containing about 1.8%. The latent topics as discovered by the topic model cannot be ordered by size in this way, since a document is not assigned to a single topic. The topics are therefore not in any specific order.

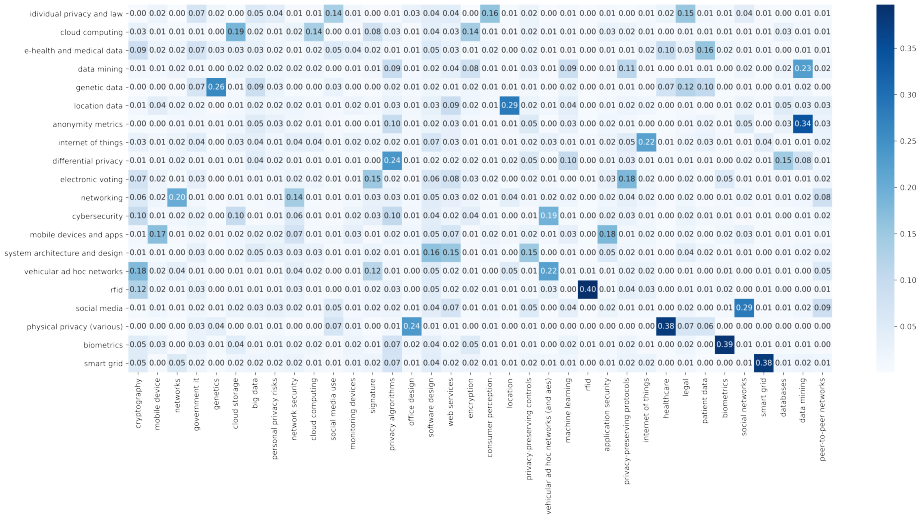

Fig. 2. Average proportion of topic occurrence within network communities

Overall, this overview of relations between the two classifications[2] shows some clear patterns. Especially because values close to 1 are impossible since all documents in a community contain multiple topics. Most importantly, it is a remarkable result that two methods that use completely different document properties end up in a relation to each other in which for most of the classifications from the network method, one or multiple counterparts can be clearly distinguished in the topics from the topic model.

For some communities, there is a clear 1-on-1 relation to one of the topics. Examples are *biometrics*, *RFID*, *smart grid*, *location data* and *genetic data*, for

[2] Technically speaking, the topic model does not classify documents into a single category, but we will speak of the 'assignment of topic distribution' as a classification for readability purposes

which all papers contained in the community are between 26% and 40% 'assigned to' one specific topic. Apparently, these areas of research were equally identifiable with both methods. Some patterns make sense after looking at the citation network. In the network, the communities of *data mining* and *anonymity metrics* are heavily interconnected. Since *anonymity metrics* was not identified as a separate topic by the LDA algorithm, it is not strange to find it to be contained in the topic that was labeled as *data mining*. The fact that *anonymity* is one of the top words in this topic seems to confirm this.

Since the communities are not uniform in size, it is to be expected that the larger communities cover multiple topics. The citation network community of *individual privacy* contains literature 'assigned to' topics on *social media use*, *consumer perception* and *legal* matters. The community of *cloud computing* contains literature on *cloud storage, cloud computing* and *encryption*. Often, the topics that are part of bigger communities are not common as a proportion of other communities; another sign of many similarities between the two classifications.

On the other hand, some topics are a proportion of many of the communities. This is notably the case for the topics of *software design*, *web services*, *cryptography* and *privacy algorithms*. There are multiple possible explanations: papers 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. We cannot know whether one or both are mislabeled, nor can we know whether the community or topic was wrongfully identified or uncovered. We will get back to this in the discussion.

4 Discussion

This paper has explored the usage of an LDA topic model to gain insights into the embedding of privacy in the scientific literature. Firstly, by carefully working out what kind of procedure is needed to create a model that is up to scientific standards and secondly, by comparing it to a different model created on other properties of the literature.

Reflecting on the creation of the topic model, we have to conclude that its full potential is currently limited by the amount of computation time needed. Because we need a grid search to find the 'best' model, it is computationally unfeasible to do the type of exploratory research that fits these kinds of creative methods. Ideally, a researcher could easily adjust model parameters and get immediate feedback to 'get a feeling' for the model. Being able to look at the multitude of possible topic models is especially important because there is no true value for N, the number of topics. Picking just one model for analysis is somewhat arbitrary, and comparing a multitude of models might result in new insights into the scientific discourse.

Observations of just the latent topics did not reveal much insight into the scientific discourse about privacy. Notable associations between topics were often on the method level, distracting from analysis on a higher level. Tweaking the

algorithm or relaxing some of the model's assumptions, like the bag-of-words approach, could increase its usefulness. However, doing exploratory research with novel algorithms applied to bibliometric analysis requires a level of mathematical understanding that might be a rare combination with domain knowledge or incentives to perform further research on applications like these.

The second part of the results, where the topic model was compared to a citation network analysis of the same data, showed how two different methods of categorizing literature could result in surprisingly similar results. The fact that so many topics could be related to communities in the citation network is remarkable and gives evidence for a certain extent of both methods' validity. For cases in which the topic-community relation was not one-to-one, this could be explained by differences in size: one community contained multiple topics.

When the two classifications do not match one-to-one, the human capacity to see patterns is lurking. However, we have to be very careful with our interpretation. Combining the two classifications stacks uncertainties of subjectivity when choosing a community detection method, choosing the number of topics, and labeling the communities and topics. When a relationship between a topic and a community is unusual, we can not know whether there is actual meaning to be found behind it. Since, for each observation, the possibility exists that the topic or community was erroneously uncovered, just the coincidental consequence of the chosen method or algorithm parameter, or mislabeled. For scientific rigor, a way to check which of the observed patterns still holds when these factors change would be needed; a sensitivity analysis of conclusions.

Furthermore, the question arises on which level of abstraction the documents are classified. It is challenging to distinguish groupings on the basis of methods as opposed to groupings by subject. Topics showing up as small proportions in several communities leave us wondering whether we uncovered subjects relevant to many disciplines or topics not corresponding to subjects but containing widely used techniques. Additionally, there is still the possibility we are looking at one of the errors mentioned above in our methods, and there is no actual pattern in the literature.

More research into these methods is essential before these novel applications can be more than something to stimulate a discussion and inspire further research.

References

1. Baas, J., Schotten, M., Plume, A., Côté, G., Karimi, R.: Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies **1**(1), 377–386 (2020)
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. "O'Reilly Media, Inc." (Jun 2009)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. pp. 113–120. ICML '06, Association for Computing Machinery, New York, NY, USA (Jun 2006). https://doi.org/10.1145/1143844.1143859

4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
5. Boyd-Graber, J., Mimno, D., Newman, D.: Care and feeding of topic models: Problems, diagnostics, and improvements. Handbook of mixed membership models and their applications **225255** (2014)
6. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.M.: Reading Tea Leaves: How Humans Interpret Topic Models p. 10
7. Chen, T.H., Thomas, S.W., Hassan, A.E.: A survey on the use of topic models when mining software repositories. Empirical Software Engineering **21**(5), 1843–1919 (Oct 2016). https://doi.org/10.1007/s10664-015-9402-8
8. Denny, M.J., Spirling, A.: Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. Political Analysis **26**(2), 168–189 (2018)
9. van Dijk, F.W.: Pillars of Privacy: Identifying Core Theory in a Network Analysis of Privacy Literature
10. DiMaggio, P., Nag, M., Blei, D.: Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics **41**(6), 570–606 (Dec 2013). https://doi.org/10.1016/j.poetic.2013.08.004
11. Dumais, S.T., et al.: Latent semantic indexing (lsi) and trec-2. Nist Special Publication Sp pp. 105–105 (1994)
12. Elgesem, D., Steskal, L., Diakopoulos, N.: Structure and content of the discourse on climate change in the blogosphere: The big picture. Environmental Communication **9**(2), 169–188 (2015)
13. Grimmer, J., Stewart, B.M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis **21**(3), 267–297 (2013)
14. Haselmayer, M., Jenny, M.: Measuring the tonality of negative campaigning: Combining a dictionary approach with crowd-coding. political context Matters: Content analysis in the social sciences, Mannheim, Germany (2014)
15. Hoffman, M.D., Blei, D.M., Bach, F.: Online learning for latent Dirichlet allocation. In: In Advances in Neural Information Processing Systems 23 (NIPS '10 (2010)
16. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50–57 (1999)
17. Huang, J.: Maximum Likelihood Estimation of Dirichlet Distribution Parameters (Jan 2005)
18. Jacobi, C., Atteveldt, W.v., Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. Digital Journalism **4**(1), 89–106 (Jan 2016). https://doi.org/10.1080/21670811.2015.1093271
19. Kokolakis, S.: Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Computers & Security **64**, 122–134 (Jan 2017). https://doi.org/10.1016/j.cose.2015.07.002
20. Lovins, J.B.: Development of a stemming algorithm p. 10
21. Maier, D., Niekler, A., Wiedemann, G., Stoltenberg, D.: How Document Sampling and Vocabulary Pruning Affect the Results of Topic Models. Computational Communication Research **2**(2), 139–152 (Nov 2020)
22. Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., Adam, S.: Applying LDA Topic Modeling in Communication Research: Toward a Valid and Re-

liable Methodology. Communication Methods and Measures **12**(2-3), 93–118 (Apr 2018). https://doi.org/10.1080/19312458.2018.1430754
23. MCCALLUM, A.: MALLET:A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (2002)
24. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM **38**(11), 39–41 (Nov 1995). https://doi.org/10.1145/219717.219748
25. Mimno, D., McCallum, A., Mann, G.S.: Bibliometric impact measures leveraging topic analysis. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06). pp. 65–74 (Jun 2006). https://doi.org/10.1145/1141753.1141765
26. Noyons, E.C.M.: Bibliometric mapping as a science policy and research management tool. Ph.D. thesis, DSWO Press (Dec 1999), iSBN: 9789090132501
27. Perera, C., Barhamgi, M., Bandara, A.K., Ajmal, M., Price, B., Nuseibeh, B.: Designing privacy-aware internet of things applications. Information Sciences **512**, 238–257 (Feb 2020). https://doi.org/10.1016/j.ins.2019.09.061
28. Rehurek, R., Sojka, P.: Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic **3**(2) (2011)
29. Rose, M.E., Kitchin, J.R.: pybliometrics: Scriptable bibliometrics using a python interface to scopus. SoftwareX **10**, 100263 (2019)
30. Röder, M., Both, A., Hinneburg, A.: Exploring the Space of Topic Coherence Measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 399–408. WSDM '15, Association for Computing Machinery, New York, NY, USA (Feb 2015). https://doi.org/10.1145/2684822.2685324
31. Scott, J., Baldridge, J.: A recursive estimate for the predictive likelihood in a topic model. In: Artificial Intelligence and Statistics. pp. 527–535. PMLR (Apr 2013), iSSN: 1938-7228
32. Scott, S., Matwin, S.: Feature engineering for text classification. In: ICML. vol. 99, pp. 379–388. Citeseer (1999)
33. Sievert, C., Shirley, K.: Ldavis: A method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. pp. 63–70 (2014)
34. Syed, S.: Topic Discovery from Textual Data: Machine Learning and Natural Language Processing for Knowledge Discovery in the Fisheries Domain (Mar 2019), https://dspace.library.uu.nl/handle/1874/374917, accepted: 2019-02-01T17:21:47Z ISBN: 9789039370865 Publisher: Utrecht University
35. Syed, S., Spruit, M.: Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 165–174 (Oct 2017). https://doi.org/10.1109/DSAA.2017.61
36. Syed, S., Borit, M., Spruit, M.: Narrow lenses for capturing the complexity of fisheries: A topic analysis of fisheries science from 1990 to 2016. Fish and Fisheries **19**(4), 643–661 (2018). https://doi.org/https://doi.org/10.1111/faf.12280
37. Syed, S., Spruit, M.: Exploring Symmetrical and Asymmetrical Dirichlet Priors for Latent Dirichlet Allocation. International Journal of Semantic Computing **12**(03), 399–423 (Sep 2018). https://doi.org/10.1142/S1793351X18400184
38. Wallach, H.M., Mimno, D.M., McCallum, A.: Rethinking lda: Why priors matter. In: Advances in neural information processing systems. pp. 1973–1981 (2009)
39. Warren, S.D., Brandeis, L.D.: Right to Privacy. Harvard Law Review **4**, 193 (1890)