Identifying Recent Research Topics in Post-Quantum Cryptography via Topic Modelling

Boyeon Song
Division of Science and Technology Digital Convergence
Korea Institute of Science and Technology Information
Daejeon, Korea
bysong@kisti.re.kr

Tae Jong Kim
Division of Science and Technology Digital Convergence
Korea Institute of Science and Technology Information
Daejeon, Korea
k2boy3@naver.com

***Abstract*—We collect the papers of Post-Quantum Cryptography (PQC) published from August 2022 to July 2023, which are able to be accessed on Web of Science. To get an idea what topics have mainly been researched about PQC for the recent one year, we employ topic modelling using Latent Dirichlet Allocation (LDA) approach. It is a popular method to extract topics from a collection of documents and divide them into the topics. Our LDA analysis classifies the selected PQC papers into six topics based on their similarities and extracts main keywords of each topic. The result supports us to be able to understand the recent research of PQC and find the future research directions.**

***Index Terms*—post-quantum cryptography, PQC, topic model, Latent Dirichlet Allocation, LDA, quantum-proof, quantum-safe, quantum-resistant.**

I. Introduction

Post-Quantum Cryptography (PQC) (also known as quantum-proof, quantum-safe, or quantum-resistant cryptography) is a term to describe cryptographic algorithms that are secure against cryptanalytic attacks by a quantum computer [1]. National Institute of Standards and Technology (NIST) initiated a standardization process of PQC in December 2016 by calling for quantum-resistant public-key cryptographic algorithms for new public-key cryptography standards: digital signatures and encryption/key-establishment. After its six-year competition process, NIST announced the first four PQC candidates to be standardized and fourth round submission in July 2022. [5], [6]

In this paper, we collected the papers of PQC published from August 2022 to July 2023 by using Web of Science [4], that is, which have been published since the announcement of the first group of winners of NIST PQC standards and Round 4 candidates.

We then applied a topic model to understand what topics of PQC have mainly been researched for the recent one year. Topic model is a machine learning and natural language processing technique for discovering the abstract topics that occur in a collection of documents [3].

As far as we know, our work is the first analysis in PQC using topic modelling to identify recent research topics.

This research was supported by Korea Institute of Science and Technology Information (KISTI). (No. K-23-L04)

II. Recent Research Analysis of PQC

In this section, we introduce our analysis method being used to investigate recent research trends in PQC. Then, we show the analysis result with extracted topics and their main keywords.

A. Research methodology

We first collected the papers related to PQC accessed on Web of Science, which were published between August 2022 and July 2023, i.e., for the last one year. The initial papers searched with the words, *Post-Quantum* & *Cryptography*, was 198. After examining them, we eliminated 9 papers because of duplication or irrelevance to PQC. Thus, 189 papers were selected as the final analysis samples.

We then analyzed the selected papers using Latent Dirichlet Allocation (LDA), a type of topic modelling to determine topics present in a document. LDA is a generative probabilistic model for collections of discrete data, such as text corpora, and is commonly applied to discover topics in a collection of documents [2], [7]. It recognizes topics in the documents through several steps [8].

Next, we extracted and refined keywords by frequency analysis and preliminary topic modelling results, and created three dictionaries: (a) definition dictionary including definition terms such as key encapsulation and digital signature, (b) synonym dictionary containing same or similar terms, e.g., Internet of Things, IoT, and IOT, and (c) exclusion dictionary for excluding common or irrelevant words such as article, example, etc, and al.

For topic model optimization, we measured coherence score (u_mass) [9] that indicates the level of semantic similarity between words on a topic. Coherence score (u_mass) plays a pivotal role in finding the optimal parameters for topic modelling, especially in the context of LDA analysis. It is because determining the most appropriate number of topics and setting the right values for hyper parameters such as alpha and beta are challenging [9], [10].

Topics were set from 5 to 14, with alpha values ranging from 0.01 to 0.1, and beta values from 0.01 to 0.02. This resulted in 200 different coherence scores. The highest coherence score was -4.181, when the number of topics is six with an alpha value, 0.07, and a beta value, 0.02, as shown in Fig. 1. The

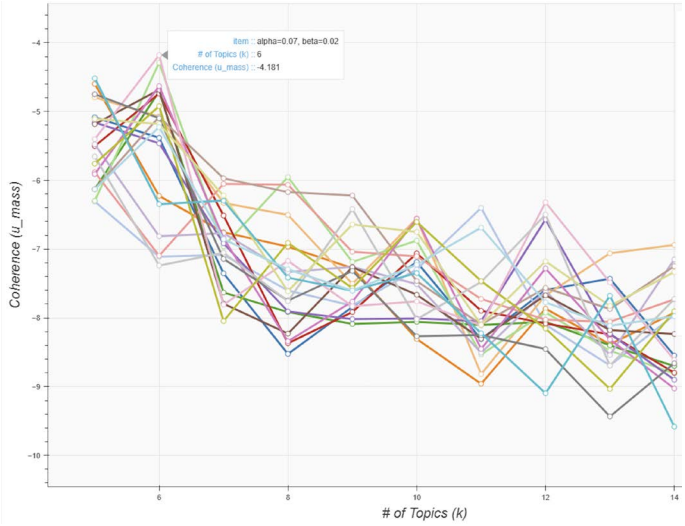

Fig. 1. Coherence score and the number of topics

parameters for the highest coherence score were used for the final LDA topic modelling analysis.

B. Research topics of PQC by LDA analysis

The LDA analysis mentioned above classified the 189 papers of PQC into 6 topics. Topic 1 includes 30 papers, Topic 2 includes 24 papers, Topic 3 includes 20 papers, Topic 4 includes 40 papers, Topic 5 includes 35 papers, and Topic 6 includes 40 papers. Fig. 2 shows the number and percent of the PQC papers allotted per topic.

Topic 4 and Topic 6 both accounted for the highest percentage of the papers, Topic 5 took 2nd place, Topic 1 took 3rd place, Topic 2 took 4th place, and Topic 3 is the last one. But, there are not much big difference between the number of papers per topic.

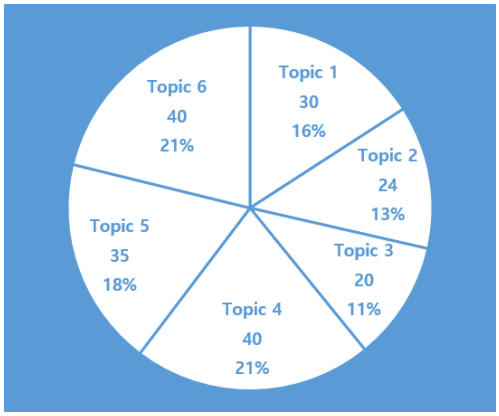

Fig. 2. the number and percent of the selected PQC papers of 6 topics

The analysis of Topic 1 generated ten main keywords: *signature, secret key, polynomial, generation, isogeny, public key, vector, Commutative Supersingular Isogeny Diffie–Hellman (CSIDH), ciphertext, and digital signature*. Fig. 3 shows the ten keywords and their probability of Topic 1.

By examining the papers belonging to Topic 1, we found that they are about post-quantum signature algorithms or key exchange algorithms especially considering isogenies, supersingular isogeny Diffie–Hellman, and CSIDH. We titled Topic 1 'Post-Quantum Signature and Key Exchange'.

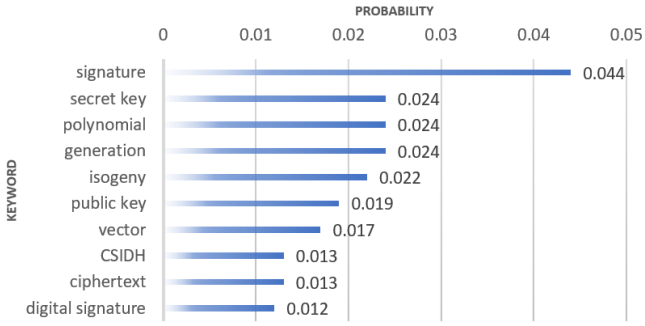

Fig. 3. Ten keywords and their probability of Topic 1

The analysis of Topic 2 displayed ten major keywords: *proof, cryptosystem, complexity, matrix, code, ring signature, zero-knowledge, lattice-based, knowledge, and error*. Fig. 4 shows the ten keywords and their probability of Topic 2.

We defined the topic as 'Post-Quantum Zero-Knowledge Proofs (ZKPs)'. The papers in the topic mainly dealt with post-quantum ZKPs and their applications, including analysis of ZKPs for a NP-complete problem, code-based ZKPs of knowledge in the rank setting, zero-knowledge argument protocols, and challenges in designing lattice-based post-quantum ZKPs.

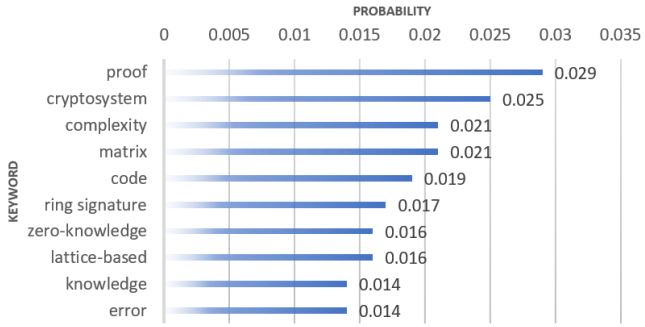

Fig. 4. Ten keywords and their probability of Topic 2

The analysis of Topic 3 outputted ten chief keywords: *blockchain, quantum, multiplication, cybersecurity, evaluation, sample, threshold, standard, Private Stream Aggregation (PSA), and Ring Learning with Error (RLWE)*. Fig. 5 shows the ten keywords and their probability of Topic 3.

They showed clearly that the topic is about 'Quantum-Secure Cryptographic Approaches in Blockchains'. The papers in the topic discussed the following issues: cryptographic primitives in blockchains, post-quantum PSA from RLWE, cybersecurity of the current blockchain technologies, and cryp-

tographic schemes for quantum-resilient security protocols in blockchain-based applications.

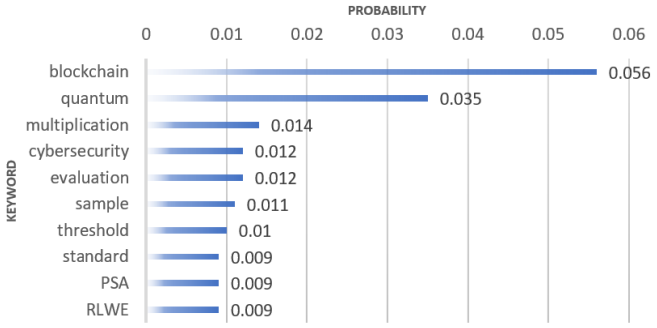

Fig. 5. Ten keywords and their probability of Topic 3

The analysis of Topic 4 revealed a distinct emphasis on hardware-related themes of quantum-resistant cryptographic algorithms. The key terms that define the topic are *hardware, NIST, Number Theoretic Transform (NTT), Saber, Field Programmable Gate Array (FPGA), Key Encapsulation Mechanism (KEM), accelerator, polynomial multiplication, key encapsulation, and memory*. Fig. 6 shows the ten keywords and their probability of Topic 4.

The topic is titled 'Hardware Acceleration of NIST PQC Algorithms'. The papers in the topic addressed various aspects of PQC hardware, for example, pipelined NTT architecture, Saber polynomial multiplication using efficient FPGA architecture, and hardware-implemented lightweight accelerators. They mainly focused on hardware implementation to increase the efficiency and speed of post-quantum cryptographic processes.

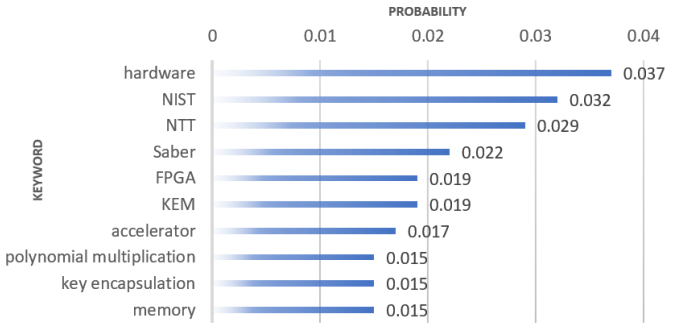

Fig. 6. Ten keywords and their probability of Topic 4

The analysis of Topic 5 extracted ten top keywords: *encryption, NIST, cryptosystem, public-key, decryption, quantum, analysis, error, cost, and Learning With Errors (LWE)*. Fig. 7 shows the ten keywords and their probability of Topic 5.

The topic indicates a specific focus on the analysis and implementation of public-key encryption algorithms being considered as potential NIST PQC standards. The related papers contained the following subjects: improved low-depth SHA3 quantum circuit, polar coding for RLWE based public key encryption, and analysis of RLWE channel. The topic is titled 'Analysis of Public-Key Encryptions of NIST PQC Candidates'.

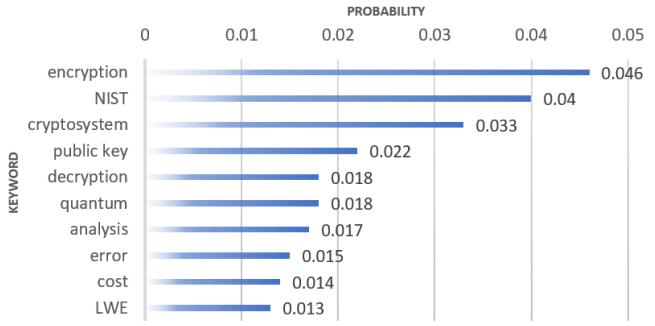

Fig. 7. Ten keywords and their probability of Topic 5

The analysis of Topic 6 obtained ten top keywords: *signature, Internet of Things (IoT), lattice-based, lattice, public-key, encryption, authentication, privacy, cloud, and quantum attacks*. Fig. 8 shows the ten keywords and their probability of Topic 6.

It was easy to find that the topic is mainly about 'Post-Quantum Digital Signature'. The works published within this topic explored various approaches of quantum-safe digital signatures, which are based on different algorithm types, that is, lattice-based, code-based, or multivariate-based cryptographic schemes, in different usages, for example, proxy signature, group signature, or multi-signature. Signature schemes were also considered to being applied to IoT devices or wireless networks, to be able to resist quantum attacks.

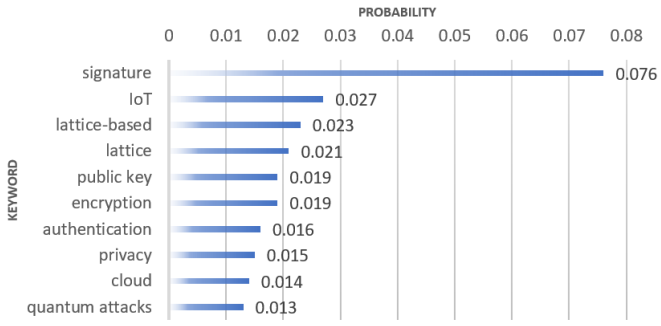

Fig. 8. Ten keywords and their probability of Topic 6

We summarized the topic analysis result of the selected PQC papers in Table I. The table briefly shows the main six topics of PQC research for the recent one year, with the number and percent of the papers belonging to each topic.

III. Conclusion

We have surveyed the 189 papers associated with PQC that were published for the recent one year (from August

TABLE I
THE RECENT RESEARCH TOPICS OF PQC

Topic #	Topics in PQC	# of papers (%)
1	Signature and key exchange	30 (16%)
2	Zero-knowledge proofs	24 (13%)
3	Cryptographic approaches in blockchains	20 (11%)
4	Hardware acceleration of NIST PQC	40 (21%)
5	Analysis of NIST PQC	35 (18%)
6	Digital signature	40 (21%)
	Total	189

2022 to July 2023), available on Web of Science. We then have analyzed them using LDA approach, which is serving a powerful tool to extract topics from a bunch of documents and divide them into the topics. For optimizing our topic modelling, we have used coherence score (u_mass) to find optimal parameters. It ensures that the topics derived are not only statistically significant, but also contentually relevant.

By systematically collecting, refining, and validating the data, we have obtained the analysis result, in which the selected PQC papers are classified into six topics with the top 10 frequent keywords, respectively. Topic 1 is about post-quantum signature and key exchange, Topic 2 is about post-quantum zero-knowledge proofs, Topic 3 is about quantum-secure cryptographic approaches in blockchains, Topic 4 is about hardware acceleration of NIST PQC algorithms, Topic 5 is about analysis of public-key encryptions of NIST PQC candidates, and Topic 6 is about post-quantum digital signature (see Table I).

Both Topic 4 and Topic 6 took up the highest percentage of the collected papers, as shown in Table I. The major research topics among six topics were hardware implementation of NIST PQC algorithms and post-quantum digital signature. But the percentage per topics are evenly distributed without big difference between them. It means that PQC has widely studied throughout topics. As we could guess, NIST PQC candidates have received considerable attention, and Topic 4 & 5 are related to them, which accounted for 39 percent of the papers. The study of post-quantum ZKPs of Topic 2 has recently emerged and is expected to be more active.

The LDA analysis helps us to understand comprehensively what fields of PQC have been researched for the last one year. Furthermore, we will continue reviewing noteworthy papers related to the topics thoroughly for further research.

REFERENCES

[1] Wikipedia, https://en.wikipedia.org/wiki/Post-quantum_cryptography
[2] Wikipedia, https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
[3] Wikipedia, https://en.wikipedia.org/wiki/Topic_model
[4] Web of Science, Clarivate™, https://www.webofscience.com/wos/woscc/basic-search
[5] Computer Security Resource Center, National Institute of Standards and Technology, https://csrc.nist.gov
[6] D. Moody, The beginning of the end: the first NIST PQC standards, March 2022, https://csrc.nist.gov/csrc/media/Presentations/2022/the-beginning-of-the-end-the-first-nist-pqc-standa/images-media/pkc2022-march2022-moody.pdf
[7] Blei, D.M., Ng, A.Y. and Jordan, M.I., Latent dirichlet allocation. Journal of machine Learning research, pp.993–1022, Jan. 2003.
[8] Enes Zvornicanin, When Coherence Score Is Good or Bad in Topic modelling?, https://www.baeldung.com/cs/topic-modelling-coherence-score, updated in May 31, 2023.
[9] Röder, M., Both, A., and 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, 2015.
[10] Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. , Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108, 2010.