Mapping PhD Theses of UK Universities to CyBOK

Virginia N.L. Franqueira (PI); Jason R.C. Nurse and Shujun Li (Co-Is);
and Rahime Belen Saglam (RA)

4 October, 2021

1 Introduction and Project Summary

The goal of this project was to analyse the areas of cyber security research that have been investigated in PhD theses published in the UK in the past four years (2017-2020), and to reflect on the evolution of such research over that period of time. More specifically, we collected a dataset of 315 unique cyber security-related PhD theses awarded by universities in the UK via searches into the ProQuest database[1]. Those theses were then mapped manually to one or more CyBOK Knowledge Areas (KAs) based on the topics covered. This mapping was designed to allow us to categorise the research topics of the collected PhD theses and to observe major trends.

However, during the project we encountered a number of major challenges, which resulted in inconclusive findings from the analysis conducted. Therefore, instead of reporting on the findings from the mapping study, we consider the main results from this project to be more about the challenges we faced (Section 2) and recommendations for future work focusing on this domain (Section 3).

2 Challenges Identified

The project sought to apply a series of automated and manual approaches to achieve its aims. The challenges we identified during the process are elaborated below.

- In terms of mapping of PhD theses to CyBOK KAs, a number of automated approaches were attempted, i.e., topic modelling, unsupervised clustering techniques, and cosine similarity based correlation analysis. However, they had to be abandoned because they returned unreliable results on topic coverage. As a consequence, in order to guarantee the mapping accuracy, this task was completed manually and independently by several members of the project team to allow cross-validation. A side-effect of the manual process, however, was that limitations in the ProQuest database search observed at a later stage of the project, e.g., that ProQuest had a software bug that led to missing records of PhD theses published in 2016 (a year we originally planned to cover), could not be fixed because it was not possible to re-do the labour-intensive manual process within the available time frame.

- We undertook a set of manual analyses to validate our work. The manual validation of our dataset of theses revealed that the PhD theses covered in our work were an incomplete subset of all cyber security related theses published in the UK during the period of 2017-2020. This reality was a consequence of two main reasons.

  Firstly, we found that there are many relevant PhD theses missing from ProQuest, the database on which we based our search. The approaches taken by different UK universities and PhD sponsoring bodies in the UK seem to be inconsistent in regards to submission of PhD theses to ProQuest. It is unclear what factors led to such a phenomenon, e.g., whether the embargo period or reservations against ProQuest are the main driving reason of not submitting to ProQuest. The main alternative

[1] `https://www.proquest.com/`

to ProQuest for UK PhD theses, EThOS[2], did not allow us to set the minimum set of attributes required for the search query, i.e., time period, and therefore turned out to be unusable.

Secondly, we had to rely on keyword-based searches due to the lack of reliable ground truth labels accompanying relevant PhD theses. We first used a generic query, identified by inspecting a number of survey papers, to capture an initial set of PhD theses that are likely to be related to cyber security. Then, we defined a search query for each CyBOK KA, obtained in one of the following two approaches: derived from the KA titles, and via manual evaluation of the corresponding Knowledge Tress. The latter approach required experts with substantial knowledge on the corresponding KA, so we could not use it for all KAs, but just four for which the first (simpler) approaches did not work well. An intrinsic problem here—that we noticed during this study—is that no keyword-based research queries can guarantee the return of all relevant PhD theses. In addition, we also faced the dilemma of using broader keywords that result in a large number of false positives and using more restrictive keywords that lead to false negatives. We made the latter choice to be able to comply with the tight time frame of the project.

- The majority of PhD theses we collected were mapped to more than one CyBOK KA; this therefore represents a wide coverage of topics per thesis. We decided to limit the maximum number of KAs assigned to a thesis to four, treating all KAs equally. This mapping led to over-counting the number of theses mapped to some KAs. Mapping more than four KAs or using precise weights (i.e., defining the degree to which a thesis covered a particular KA) would require a substantial amount of efforts of experts who have a very good knowledge of all relevant KAs, which was simply unfeasible to do given the short nature of the project.

3 Recommendations

Recognising the strategically important insights that a reliable mapping of PhD theses published in the UK could bring to inform future UK cyber security capacity building activities, we propose the following recommendations based on the challenges we described above.

First, we recommend developing mechanisms to establish **"ground truth" mappings of cyber security related PhD theses to CyBOK KAs**. For instance, this can be achieved via a collaborative effort of all UK universities by asking future PhD students and their supervisors to add the relevant CyBOK KAs as new metadata of their theses. In principle, this can be done for many already published PhD theses as long as the PhD student or the supervisor are still contactable. During our mapping exercise we noticed that some PhD theses are only loosely related to cyber security, which we decided to exclude from our collection. This suggests that developing some guidelines for mapping PhD theses to CyBOK KAs would be very useful.

Second, we recommend further studies be conducted to better **understand how PhD theses are indexed by relevant databases**, particularly ProQuest and EThOS. This should cover studies on the process followed by UK universities and PhD sponsoring bodies in terms of submission to such databases, in order to determine the root cause of any incomplete coverage. Such studies will help identify the best databases to use and help inform UK universities and PhD sponsoring bodies in the UK about future best practices. If the best database turns out to be EThOS, it would be worth working with the database's owner – the British Library – to develop a more powerful API that can support more sophisticated search queries. An alternative to retrieve PhD theses would be to obtain this information directly from universities since PhD theses are normally made public on their library website or an open research portal. While this approach can eliminate uncertainties in relation to incomplete coverage of PhD theses in databases such as ProQuest and EThOS, the different systems used by UK universities can be a major challenge in automating the process. If future studies sought to collect such information via academic departments directly, special care has to be paid to contact *all* departments (e.g., Computing, Engineering, Politics, Psychology, etc.) and other relevant bodies such as research centres and institutes, UKRI-funded Centres for Doctoral Training (CDTs), European Commission funded Incoming Training

[2] `https://ethos.bl.uk/`

Networks (ITNs), where cyber security related PhD theses, especially interdisciplinary ones, may be produced.

Third, we recommend further research be conducted to investigate **how to automate the mapping process between PhD theses and CyBOK KAs**, considering each thesis can be mapped to multiple KAs. This will allow cross-KA analysis to identify how different CyBOK KAs interact with each other. We envisage that this line of research will benefit from the ground truth mappings of PhD theses and the standard search-oriented keywords we recommended above. We envisage a key challenge in this area is how to automatically determine the weight of each CyBOK KA for a given thesis. Another challenge would be to map different chapters and section of a PhD thesis to CyBOK KAs so that we can produce a hierarchical KA map for the thesis to facilitate more complicated topical analysis.

Last but not the least, we also recommend **studying the mappings between PhD theses and CyBOK KAs in the context of different disciplines**, in order to study the interactions between CyBOK KAs and relevant disciplines. While each KA itself can also be mapped to relevant disciplines, looking at the actual discipline(s) a PhD thesis belongs to can reveal more insights about how interdisciplinary cyber security research is conducted within different disciplines. Such studies can also help provide useful input for future revisions of CyBOK KAs, by identifying researchers whose work has not been but should be covered in one or more KAs.