Mining and Utilizing Patent Related Data in Quantifying the Societal Impact of Security Technologies

ZHONGYU SHI, University of Twente, The Netherlands

Different from academic impact evaluation, the quantitative societal impact evaluation tool is still limited, many of them are based on altmetrics, including social media data in measuring societal impact. During practical usage, it shows many limitations, especially in terms of lacking data in old papers, can be gamified, and undervaluing basic studies.

In this paper, patent-related data is used as the indicator. Three methodologies are designed beyond the traditional direct patent citations, to mine more data and improve the evaluation of "basic study" compared with Altmetric. They are applied in all IEEE S&P papers from 2000 to 2005 to test the performance. Methodology 1 can extract more related patents, methodology 2 extends the length limit between papers and patents, and methodology 3 proposed a new topic-based algorithm. Based on methodology 3, we did several case studies to verify the relationship between papers and patents on the same topics. The result shows a strong relationship between them.

Additional Key Words and Phrases: Societal Impact Evaluation, Patent, Altmetrics, Security, Natural Language Processing

1 INTRODUCTION

There are many reasons exist for research impact evaluation, research organizations need to monitor and manage their performance and disseminate contribution, government; stakeholders, and the wider public want to know the value of research; and researchers want to ensure future funding, as well as improving the understanding to develop better impact delivering methodologies. On the other hand, it is important to notice that the research impact evaluation can lead to a devaluation of 'blue skies' research[17]. Therefore, it's meaningful to look back and analyze what we can learn from these cases for better research impact evaluation.

Research impact consists of academic impact and societal impact. Academic impact means the significance of research for stakeholders within the academic world, there are already many well-known and widely used tools to measure the academic impact, like H-index, citations, field normalized citation impact, etc[20]. Societal impact evaluation means the assessment of social, cultural, environmental, and economic returns (impact and effects) from results (research output) or products (research outcome) of publicly funded research. However, more so than with academic impact measurement, the assessment of societal impact research is still badly needed, societal impact is much harder to access than is academic impact and is still in the early stages[8].

The societal impact research can be divided into two types, general methodologies like altmetrics which focuses on social media citations and other indicators like policy documents, news, patents, etc., it can be used widely in all fields, also there are methodologies specifically in fields like Health Technology Assessment (HTA) that make use of medical data[16, 18]. Similar to the health field, the importance of security technologies assessment is noteworthy, both ethically and economically, as a lack of security input could lead to money and information loss, while proper use of the assessment can inform decision making. Take security breaches as an example, today's security breaches are severe that the number and the cost of damages are increasing year by year, one of every eight websites has at least one critical vulnerability[11]. Nevertheless, relevant studies regarding the impact valuation of security technologies are limited.

As highlighted by L. Bornmann's survey, the approach of societal impact valuation should be as broadly based as possible, identify appropriate indicators for different disciplines, and also develop mechanisms to collect accurate and comparable data[8]. Therefore, to make the assessment methodology as extensive as possible so it can fit various fields, it is vital to first analyze and evaluate the indicators of the current benchmark Altmetric tool in quantitative societal impact valuation. After that, develop methodologies beyond that and then evaluate their performance using data from the security field.

By combining the aspects mentioned above, the main requirements of our methodologies are:

- **R1**: The methodologies should be general so they can be used in different disciplines.
- **R2**: The methodologies should try to mitigate the problem of undervaluing "blue skies" research (basic research).
- **R3**:The methodologies should improve the performance in certain aspects based on the Altmetric tool.
- **R4**:The methodologies should be evaluated using the data from the security field.

To achieve these requirements, we need to answer these research questions:

- **RQ1**: How to evaluate the Altmetric tool and identify appropriate indicator(s) to quantify the societal impact?
- **RQ2**: How to create methodologies based on the indicator(s) to fulfill **R2** and **R3**?
- **RQ3**: In which way(s) the performance of the methodologies can be evaluated?

2 RELATED WORK

The current existing social impact evaluation methodologies can be generally divided into two types, **exclusively for one field** or **fit all fields**.

2.1 Social Impact Evaluation for one field

Health Technology Assessment (HTA) is a well-known example focusing on informing decision-making to promote better health systems, which needs to determine the value of health technology

TScIT 37, July 8, 2022, Enschede, The Netherlands

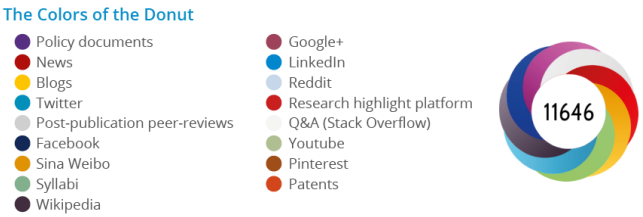

Fig. 1. Altmetric Score Components[1]

at different points in their lifecycles[16]. It has shown remarkable growth over the last decades, using machine-based technology, and widened to include smaller technologies [6]. In recent years, artificial intelligence techniques are proposed to use in this field[3].

In addition, some attempts are made in the water technology area to create a framework for quantifying the real-world impact of water technologies, using the total number of companies using the technology, number of countries using the technology, annual market value, etc.[14]

In the security field, the study done by Grag, A introduced the way of assessing the financial impact of cyberattacks by calculating the stock fluctuation after security attacks in big companies[10].

2.2 Altmetrics – Social Impact Evaluation for all fields

Different from the assessment specifically on fields, the altmetrics introduced by J. Priem can be used to assess all fields. It has garnered increasing attention, analyzing the social media citations of research papers to measure their societal impact [18, 19]. Since the idea has come out, some altmetrics tools have emerged, like Altmetric on https://www.altmetric.com, plumx used by Elsevier, etc., with the development of such tools, they have added more and more sources to improve the societal impact evaluation. In this research, we use Altmetric as the benchmark, as its wide use of research stakeholders, including some authoritative journals like Nature in its article metrics, although Nature names the section clearly as online attention, it's indeed an indicator of impact as mentioned on altmetrics website[4].

Despite the novelty and utility of altmetrics, there exist many difficulties as well. As the indicators of altmetrics were initially based on social media citations, it has some drawbacks mentioned by J. Mingers's paper, altmetrics can be gamed by "buying" likes or tweets, high altmetrics scores may come from controversial topics, and it can under-represent older papers[13]. Another noteworthy concern is that the Altmetric score is largely based on the number of citations to the identifiers (DOI, ISBN, etc.) of the scientific papers. It seems plausible from the article in NISO written by Lin and Fenner that citations might be the most important measure of impact, but as they also mentioned, citations only represent a small fraction of the user engagement with a paper, there are also viewed, saved, discussed, recommended types that can be used to measure the societal impact[12]. Hence, the purpose of this study is to provide methodologies that can alleviate these difficulties.

News	8
Blog	5
Policy document (per source)	3
Patent	3
Wikipedia	3
Peer review (Publons, Pubpeer)	1
Weibo (not trackable since 2015, but historical data kept)	1
Google+ (not trackable since 2019, but historical data kept)	1
F1000	1
Syllabi (Open Syllabus)	1
LinkedIn (not trackable since 2014, but historical data kept)	0.5
Twitter (tweets and retweets)	0.25
Facebook (only a curated list of public Pages)	0.25
Reddit	0.25
Pinterest (not trackable since 2013, but historical data kept)	0.25
Q&A (Stack Exchan)	0.25
Youtube	0.25
Number of Mendeley readers	0
Number of Dimensions and Web of Science citations	0

Fig. 2. Altmetric Score Weight[5]

3 METHODOLOGY

3.1 Choose Indicators

3.1.1 Societal Impact Indicator. Since the Altmetric tool has already collected quite comprehensive sorts of indicators as shown in figure 1, we can pick reliable indicators from them to improve.

Considering the reliability of the indicators, it's necessary to prevent various difficulties described above. First, the score can be deceived, the easily accessible social media data should be removed, including blogs, Twitter, Reddit, etc. Moreover, to evaluate earlier research outputs, indicators that were not widely used before should be considered carefully, like post-publication peer reviews. There are still some existing indicators, including policy documents, patents, and news. As policy documents are limited in quantity, the focus should be either patents or news. Among news and patents, news can still be generated because of controversial topics, however the patents need more serious application and review procedures.

Another crucial factor is the link between research papers and indicator(s). A paper written by Mohammad Ahmadpoor and Benjamin F. Jones. in Science claimed their findings are consistent with theories that emphasize substantial and fruitful connections between patenting and prior scientific inquiry. Among all these fields, computer science and nanotechnology have the shortest distances between published research and patents. [2].

Therefore, the choice is to exclusively use patent related data as the societal impact evaluation indicator in this research paper.

3.1.2 Performance Evaluation Indicator. Based on the indicator of patent we have selected, what can be further improved is the quantity of patents, as well as the performance of evaluating "blue skies" studies. The former can be directly calculated by the amount and the percentage of improvement, but the latter can be quite tricky to define.

Inspired by the directed citation graph established by Mohammad Ahmadpoor et al., a possible solution is to calculate the average distance from a paper to its patent citations[2]. This approach has

to start from D=1, find direct patent citations. After that look for D=2 patents, go through all scholar citations of this paper, and then find all patent citations of these citing papers. By sustaining this process with larger D, it can ultimately halts since only later papers can cite preceding papers. If the average distance is high, then it's considered as a basic research. However, such a strategy involves a lot of data collecting, because of the time constraint and unavailability of API, we want to find an easier approach. Therefore, we turn to a simpler solution of using scholarly citations, to measure whether the scholarly contribution is undervalued.

3.2 Data Source Selection

3.2.1 Research Data. As one of the requirements mentioned above, although the methodologies need to be designed to fit all fields, the data used for evaluation should focus on the security field. Hence, the decision is to use the papers from IEEE Security & Privacy, one of the most influential research journals in the security field. Considering both the problem that Altmetric under-represents earlier papers and the delay of more than 10 years between publications and patents with large distances, the papers will be collected from 2000 to 2005 in IEEE Security& Privacy[2, 13].

3.2.2 Patent Data. Since the main process of this research requires investigating the relationship between research papers and patents, it's ideal to choose a platform that consists of both these two types of data and well link them together. To ensure the performance of evaluation, the database should be as comprehensive as possible, ideally easier for data analysis. The WIPO Manual written by D. Oldham has provided an overview of the available open-source patent databases. He outlined the strength of The Lens database on https://www.lens.org that strongly links the scientific and patent literature through citations, provides rapid access for patent analytics purposes [15]. Another point is that it covers many different sorts of patents beyond the generally used granted patents, like patent application, limited patent, etc. These types of patents are valuable, as they are also recorded and cited by other patents as well, so we also include them in the evaluation.

3.3 Evaluate Altmetric

Although many research papers have already pointed out various limitations of altmetrics, it's still worthwhile to collect the quantitative data from Altmetric before developing our methodologies, focus on the score and patent citations, and set it as a benchmark in evaluating the performance of our methodologies. The data is collected through the official API provided by Altmetric.

In figure 3, the data collected demonstrates how many percentages of Security papers we selected can have a score and patent citations by Altmetric, it turns out that more than 60% of them have a score of 0, and more than 70% of them don't have any direct patent citations collected by the Altmetric.

Moreover, to test if Altmetric score undervalues the scholarly contribution of papers, the Pearson correlation between Altmetric scores and scholar citations is calculated. The result is 0.42, which seems not very strong and has space to improve, the data is shown in figure 4.

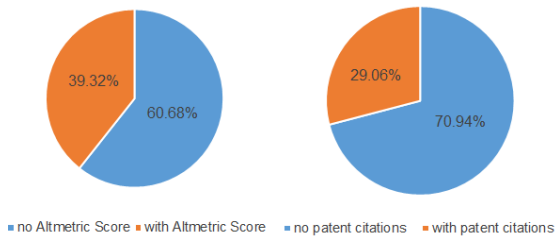

Fig. 3. Altmetric Score and Patents (IEEE S&P 2000-2005)

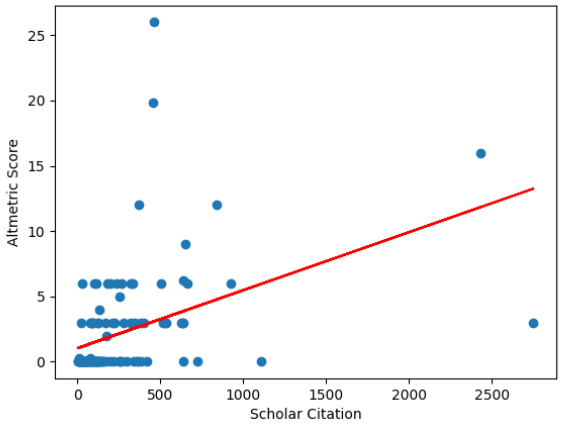

Fig. 4. The Altmetric Score and Scholar Citations

As IEEE S&P is one of the top journals in the Security field, which should have received more attention than the other papers. The result demonstrates the fact that the Altmetric tool seems doesn't perform well in older papers without much social media data, since more than half of them have zero scores. The patent related data collected by Altmetric seems not adequate for evaluation.

3.4 Methodology 1: Including "Cited with full Title"

Because of the fact illustrated by Lin "citations only represent a small fraction of the user engagement with a paper", the goal of this methodology is to somehow capture the "viewed, saved, discussed, recommended" data related to patents, so more patent related data can be collected for evaluation. Among them, the "discussed" type is defined as "science blogs" and "journal comments" in scholars, and "social media" in the public. Inspired by this, the possibility of discussing or citing the exact title of the papers, but not citing their DOIs deserves some notice. The most straightforward type is "discussed or cited the full title".

On such a basis, several small experiments are conducted. Taking the paper "Practical techniques for searches on encrypted data" as an example, to find "cited" data, we need unique identifiers like DOI, we found 49 recorded citations. To calculate the "discussed with full title data", the way is to use the full title as the identifier to conduct "exact search", which looks for the data that contains exactly the

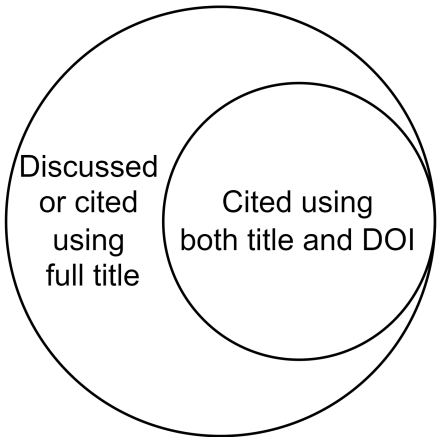

Fig. 5. Hypothetical Venn Diagram in Methodology 1

full text of the query. By doing this, there exist 193 patent records, which is more than three times of the records.

Therefore, a hypothetical Venn diagram is made as figure 5. The strategy is then applied in all selected research papers by getting the union of "cited with DOI" and "cited with full title" data using the unique identifier "the lens id" in the lens database.

3.5 Methodology 2: Extending Distance Limit between papers and patents

According to the research in evaluating the relationship between papers and patents done by M. Ahmadpoor, he defined a distance metric between patent inventions and prior papers as shown in figure 6. By analyzing a huge amount of data in various fields, the discovery is that Computer Science is one of the fields with the shortest average distance between papers and patents, which is only around 2. Moreover, about 42% of papers in computer science has direct citations, and more than 30% of papers have distance over 2, figure 7 includes such distribution in many fields with also longer distance[2].

Such discovery is also quite instructional to societal impact evaluation to gain more related patent data. As existing approaches like Altmetric only concentrate on the citations with distance 1, which results to only fewer than 30% of paper having patent citations. Why not include the scenarios with distance 2, or even longer when the evaluated paper doesn't perform well with distance 1? In addition to the performance benefits, another crucial point is that such methodology is also beneficial to linking basic research to practical research and then to patents. The point was also highlighted by M. Ahmadpoor et al., such distance metric could potentially be useful in "quantifying and tightening traditional but loose descriptors around 'basic' and 'applied' scientific research"[2].

Therefore, such a metric can be used in numerous ways to fulfill diverse societal impact evaluation tasks. For instance, if someone is interested in the societal impact of basic search, he can extend the distance range to be much larger, to get a more comprehensive insight. If the goal is merely to achieve the direct practicality, using a shorter distance is more reasonable.

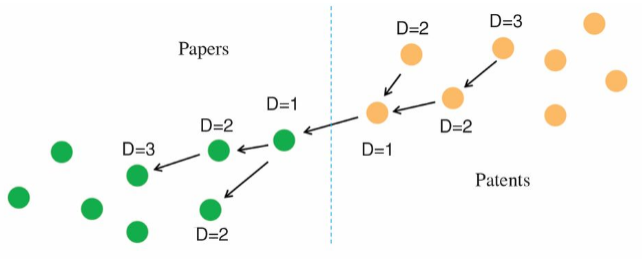

Fig. 6. Directed Citation Graph from Patents to Papers, D refers to the vertice's distance to the closest vertice on the other side, namely the number of edges in a shortest path connecting them[2]

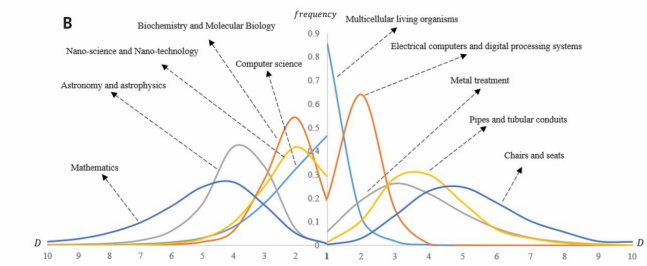

Fig. 7. Distance Distribution of fields[2]

Because of the diversity of tasks, as well as the difficulty in collecting data with longer distances, the demonstrated methodology 2 only adds distance 2 data to fit general usage. According to figure 7, in theory, this methodology can already on average reach more than 70% of all connections instead of 42%, by including the data from D=2 papers in the papers side to D=1 patents on the patents side, and also from D=1 papers on the papers side to D=2 patents in the patents side in figure 6.

3.6 Methodology 3: Calculating a Score using Patent Citations and Paper Related Topics

In methodology 1, the "discussed" data is collected solely based on using "exact search" with the full research paper title. However, it can happen in some cases that only some keywords or topics related to the research papers are "discussed", but not all of the papers in such topics are "cited". Therefore, it would be interesting to try if related keywords or topics can be extracted from papers, so more relevant patents can be found in the evaluation.

3.6.1 Extracting Keywords from Paper Full Text. The initial study aimed at extracting keywords from the full text of research papers, then using the keywords to do "exact search" in the lens database to evaluate their impact. It sounds plausible if such keyword extraction can maintain certain consistent accuracy in different types of research papers. However, even if the information extracted already has little noise, the key obstacle is the generality of extracted words, and the unavoidable huge impact of little noise.

In the paper "Fang: a firewall analysis engine", by applying keyword extraction from this paper, "Fang" and "firewall analysis engine" are extracted, which seems specific enough to be used in

filtering related patents. However, from papers like "Privacy technology lessons from healthcare", the extracted words can be much more general like "healthcare", " privacy technology lessons", which can result in much more results. The generality can highly affect the number of patents, which make the result less reliable. The other essential point is the huge impact of even little noise. Because of the different writing styles of papers and the accuracy barrier of keyword extraction tools, it seems impossible to steadily mitigate the noise. Since the objective is to do "exact search", even little noise can easily make the results deviate a lot.

3.6.2 Extracting Topics from Paper Citations. To ensure the reliability of this methodology, the way of extracting keywords from their full text is aborted. However, its idea can still work if there exists another stable way to extract topics from papers. Along with the full text of papers, the papers or patents cite them can convey a lot of information, showing the topics they have impact on and even the weight by quantity. Another crucial benefit is that with the increment of extracted data, the noise can be filtered by the occurrence time.

Since the information from citing papers are richer than citing patents, the decision is to extract topics from citing papers. Nevertheless, the topics collected in this way can no longer be the identifiers to filter patents, instead they represent the related topics that papers have impact on, and searching them results in all the related patents of these topics. Among all collected topics, the high-frequency topics can be selected, and the rest are considered as noises, the occurrence frequency of the selected topics can be used in calculating the weights of the relationship to these topics.

Having the related topics and the weight of the relationships to them, a societal impact evaluation score can be designed, using the weighted sum of the product of research papers' impact on each topic and the societal impact of each topic, represented by this formula:

$$\text{Score} = \sum \text{Weight} \times \text{Paper Impact in Topic} \times \text{Topic Societal Impact}$$

- Weight = the strength of Relationship to a Related Topic $= \frac{\text{Single Keyword Occurrence Time}}{\text{Sum of all Selected Keywords Occurrence Time}}$
- Paper Impact in Topic $= \frac{\text{Paper Scholar Citations}}{\text{Sum of Scholar Citations in Topic}}$
- Topic Societal Impact = the Amount of Patent Records

In the lens database, titles and abstracts of most scholarly works are directly exportable, they contain core information and makes the process much more efficient rather than full text. Thus, the idea is to use the combination of titles and abstracts, doing text cleaning (Normalization, Remove Unicode Characters, Remove Stopwords, and Lemmatization) and then applying keyword extraction from these papers. Since using a single word in keyword extraction can introduce lots of noise, the keyword length is limited from 2 to 3, unsupervised keyword extraction tool keyBERT provided on "https://github.com/MaartenGr/KeyBERT" is chosen, as the extracted length can be customized. The top 20 keywords are extracted from each paper, and only several top keywords should be selected, to reduce noise and also make the methodology more practical for personal users because the lens doesn't provide free API. The number of top keywords extracted should be tested and then determined by the result.

	Altmetric	DOI based	Full Title based	Union of DOI based and Full Title basd
Average Increased Percentage	0%	7.85%	15.09%	15.40%
Median Increased Percentage	0.00%	3.42%	9.31%	9.56%
Average Increased Amount	0.00	10.09	27.27	27.72
Median Increased Amount	0	2	6	7

Fig. 8. Methodology 1 compared with Altmetric

To make the result more precise, some attempts were made in using zero-shot text classification to classify all papers by more keywords, however, the result turned out that more general words like "network security" are closer to more papers. Stop words can be accumulated to avoid this problem, but as the scope of data being used in this research is not comprehensive, this step is suspended.

4 RESULTS

4.1 Methodology 1

When applying methodology 1, four types of data are collected for analysis. To omit the difference between patent datavases, DOI-based methodology is also used in the lens database, to simulate Altmetric in collecting "cited" data.

According to figures 8, the "full title" based methodology can retrieve more data. The reason behind that is not all patents use scholar DOIs when citing papers, sometimes they just use their full title, so there is a big gap between DOI based and the union. On the other hand, different from the hypothesis Venn diagram of figure 5, the smaller gap between "full title based" and the union demonstrates that in rare cases, patents only include DOI in the citations.

However, there is one exceptional case happens in the data that the title of the paper "Authentication tests" cannot be used as an identifier, since it represents a large area.

In general, this methodology works well by using the union of "full text" and DOI instead of DOI only, especially when the performance of using DOI is poor. However, handling exceptional cases is also necessary. A potential solution is not to use "full title search" with titles smaller or equal to three words (based on the experience gained from 3.6.1), because they are likely to refer to bigger topics which lead to the titles not being unique. Hence, the main problem of this methodology is that papers with shorter topics can cause some noise.

4.2 Methodology 2

When applying methodology 2, four types of data are collected, including the patent citations from the Altmetric, D=1 in the lens, and two types of D=2 data. The accuracy can be ensured since all citations are collected based on unique identifiers used by the Lens database.

From the result table in figure 9, far more related patents can be discovered than methodology 1, which is very useful when directly citing patents are lacking.

	Altmetric	D=1 the Lens	D=2 Paper <- Citing Papers <- Citing Patents	D=2 Paper <- Citing Patents <- Citing Patents
Average Increased Percentage	0%	7.85%	185.44%	171.08%
Median Increased Percentage	0.00%	3.42%	114.75%	48.75%
Average Increased Amount	0.00	10.09	205.34	452.65
Median Increased Amount	0	2	108	21
Correlation with scholarly citations	0.42	0.28	0.73	0.13

Fig. 9. Methodology 2 compared with Altmetric

Regarding the goal of considering "basic research", the Pearson correlation between their citations and scholar citations is calculated, the result in figure 12 shows that such consideration can be customized according to the requirement. When the direct societal impact needs to be considered, more distance can be put on the patents side of the metric in figure 6; if "basic research" needs to be considered, then more distance can be put on the papers side of the metric in figure 6. The distance can be adjusted for various usage.

Hence, this methodology seems appropriate for a variety of societal impact evaluation purposed, and the reliability can be ensured by using "cited" data.

4.3 Methodology 3

Based on the formula in section 3.6.2, the scores calculated using top five related topics are shown in figure 10. To evaluate the effectiveness of this methodology, the correlation with patent citations and scholar citations both need to be considered. The goal is to see whether such a topic based method can well reflect single paper's data.

Figure 11 demonstrates the trend, when extracting more than 6 top topics for each paper, the sum of correlation starts dropping, when using 5 or 6 top topics for each paper, the sum of these two correlations reaches the maximum. When top two related topics of a paper are extracted, the correlation between a paper's real patent citations and its score calculated by the algorithm reaches 0.9. This means by combining a paper's scholarly impact in its top two related topics and the patent citations of these topics, its patent citations can be roughly predicted.

Hence, when using the top 6 topics for each paper, the correlation with scholar citations reaches 0.52, which is higher than the Altmetric score's correlation with scholar citations 0.42. This method raises a bit of performance in reflecting scholarly contribution. At the same time, the correlation with patent citations is 0.77, which means this score can still well reflect single paper's patent citations data.

Another key improvement compared to ALtmetric is that it can calculate a score for all papers as long as they have some scholar citations or patent citations to extract topics from, which is easier than only using patent citations, since for even D=1, according to M.Ahmadpoor's study, it has a mean delay of 6.66 years[2]. By using this method, 99.15% of papers have a score high than 0, while only 39.32% of papers have a Altmetric score.

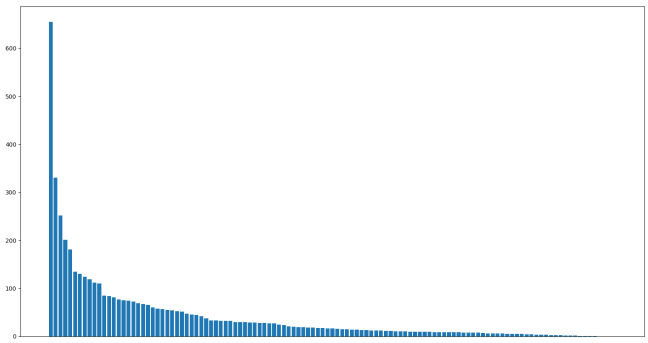

Fig. 10. Methodology 3 Score Distribution, when top five related topics are extracted for each paper

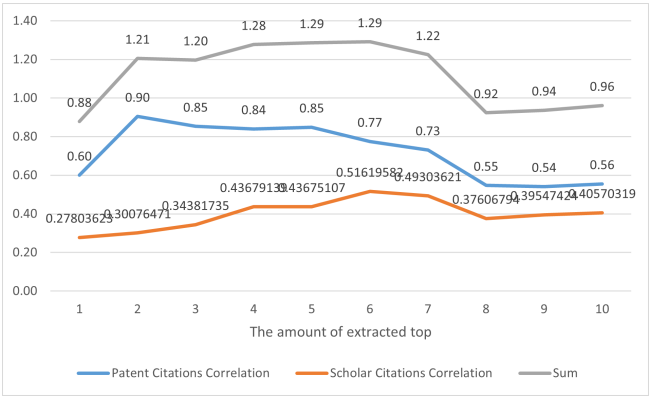

Fig. 11. Methodology 3 Correlation Trend, the x axis represents how many extracted top topics are used for each paper

The main drawbacks are that this methodology can be quite time-consuming for papers with a lot of citations, and it's not that accurate for papers with very little citations.

For the former case, three papers with over 1000 citations are studied in this research, comparing the score between using its original scholar citations or only 1000 citations, their scores don't fluctuate much. To mitigate the latter difficulty, zero-shot classification combined with keyword extraction from its full text can be potentially used, which requires further study on it.

4.4 Overall Comparison

A overall comparison table is made in figure 12 to clearly demonstrate the performance of the methodologies, as well as their advantages and drawbacks.

5 CASE STUDY

As a strong relationship was discovered by methodology 3, between patent citations of a paper and its related topics' patent citations combined with scholarly citations. It's necessary to do some case studies to show the effect of this methodology in practice and why such a topic-based method can reflect papers' societal impact. We first extract the related topics and rank them by frequency in 5.1, then we study the relationship between patents and publications

	Altmetric Score	Methodology 1	Methodology 2 D=2 Paper <- Citing Papers <- Citing Patents	Methodology 2 D=2 Paper <- Citing Patents <- Citing Patents	Methodology 3 (top 6 extracted topics)
Percentage of Papers with Patent Citations > 0	29.06%	82.05%	94.87%	66.67% (available patent records in the Lens)	n.a.
Correlation with Scholarly Citations	0.42	0.16	0.73	0.13	0.52
Correlation with Direct Patent Citations	0.52 (with Altmetric patent citations)	1	0.26	0.23	0.77
Percentage of Papers with Score > 0	39.32%	n.a.	n.a.	n.a.	99.15%
Benefits	• Good at catching latest societal impact including social media data	• Able to find more relevant patents	• It's able to be customized to fit different needs		• The coverage is good • Improved the relationship to both patent and scholarly citations based on Altmetric
Drawback	• Poor performance in old papers • Not good enough in reflecting scholarly contribution	• Not stable, shorter titles sometimes are not unique identifiers	• The data collection for longer distance can be hard		• The running speed is slow for highly cited papers (although in limited examples, the scores fluctuate very little) • The extracted topics can be inaccurate when papers have little citations

Fig. 12. Overall Comparison between all Methodologies

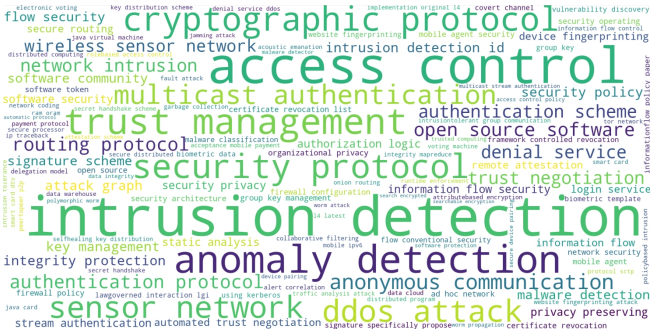

Fig. 13. Word Cloud of Related Topics

on the same topics in 5.2, since the prior study from Mohammad Ahmadpoor et al. was only based on DOI[2].

5.1 Extract related Topics and Rank them by frequency

In 3.6, the correlation reaches 0.9 when the top two related topics are extracted. Thus, all papers' top two related topics are collected and displayed as a wordcloud in figure 15.

5.2 Study the relationship between papers and patents in the context of topics

Different from the study of Mohammad Ahmadpoor et al. which built a citation graph, our study focuses more on the relationship between scholarly publications and patents in the context of topics. There are mainly two questions that need to be answered:

- **Q1**: Is there a time gap between scholarly publications and patents? Which one usually comes first?
- **Q2**: How strong is the correlation between the number of scholarly publications and patents in annual data?

Among these related topics, the most popular ones like "intrusion detection", "access control", "anomaly detection", etc. are topics with a long history, which can be traced back to the last century, and they grows steadily till now. Some smaller and more specific topics that were proposed around 2000, like "multicast authentication", "java card", etc. Some of them developed well, whereas some others stagnated at some stage. To ensure the diversity of data, the topics are divided into two types to analyze, the popular topics are analyzed in chapter 5.2.1 and the smaller ones are discussed in chapter 5.2.2

5.2.1 Popular Topics. In this section, the most popular topics "intrusion detection", "access control", "anomaly detection", and "trust management" are studied. Since a lot of scholarly and patent records exist on these big topics, the main focus of the study is to check the correlation between the number of papers and the number of patents filed based on yearly data. One important thing is that patents often have a filing date and a publication date, the time gap between them varies from case to case. Thus, we only use the filing date to reduce the noise.

Trust Management. In the computer science field, trust management is a concept introduced by Matt Blaze in the paper "Decentralized trust management" in 1996[7]. However, as displayed in figure 16, such a keyword is also used in some other fields like business, this can introduce some noise. As "trust management" is a general phrase, applicants or owners of patents can also include this keyword, which also brings some noise.

Thanks to the lens database, such noise of scholarly work can be directly filtered by limiting the field to computer science only. However, there is no field information on patents in the current mainstream patent databases. Although this problem can be potentially solved by zero-shot classification, using patents' topic and abstract as text input and scholarly fields from the lens as types, it needs much further work to validate the accuracy.

Therefore, in these case studies, the scholarly and patent data from topics is used without filtering the field. To reduce the effect of noise, several cases with diverse types are studied.

Back to the topic itself, the time gap between the paper "Decentralized trust management" and patents is very short. In the same year this paper was published, a patent "Determination of software functionality" was filed, testing the ability of a trust management system, such a patent can't be captured by the traditional DOI-based method.

In terms of the relationship between the number of scholarly publications and patents, by reducing the noises before 1996, the correlation reaches 0.86, which means they are highly correlated.

Other Popular Topics. Different from "Trust Management", "Intrusion Detection" has fewer noises from quite different fields. It's too general that hardly a paper initially introduced this field can be found, thus only the correlation is calculated. The correlation data from 1969 to 2021 is 0.9.

Similar to "Trust Management", "Access control" has a correlation of 0.96 from 1966 to 2021, and "Anomaly Detection" has a correlation of 0.97 from 1974 to 2021.

5.2.2 Specific and Smaller Topics. In this section, more specific and smaller topics are studied, including "multicast authentication" and "java card". These smaller topics can gelp identify a single paper's impact on patents.

"Multicast Authentication" is a small topic that only has 152 scholarly works and 121 patent records using the exact search in the lens database. It doesn't start from a famous scholarly work, and the correlation is only 0.08. The reason could be this is a practical topic, early patents mainly cited other patents instead of scholarly works. In terms of the time gap, the first paper was published in 1996, and there are patents on this topic from 2002.

Fig. 14. Trust Management Wordcloud

Similar to "Multicast Authentication", "java card" has a correlation of -0.03. Although there is a highly cited book from Zhiqun Chen in 2000, this topic was already mentioned in four prior scholarly works starting from 1997 [9]. One year later, there are three patents that mentioned this topic.

5.3 Conclusion of Case Study

From the diverse topics studied, papers are usually written before patents, which has a similar result to the DOI-based method from Mohammad Ahmadpoor's work. In large and general topics, the correlation between annual scholarly publications and patents is high, but in small and specific topics, the correlation is weak. However, the correlation between the total number of papers and patents on the above six topics is 0.84, so the annual correlation seems to have some noise when the topic is small. Therefore, the conclusion is that there is a strong relationship between papers and patents on the same topic.

6 CONCLUSION

In this research, many extra patent related data is mined by these three methodologies beyond direct patent citations. Methodology 1 shows the fact that many patents cite papers using either DOI or full title. By using the union of them, the number of related patents can be raised, Methodology 2 is built inspired by Mohammad Ahmadpoor et al.'s idea, which can help in catching more patent related data over longer distances. Based on different needs of societal impact evaluation, it can be customized. Methodology 3 is a new societal impact evaluation metric based on papers' related topics. The result shows that when the top two related topics of a paper are extracted, the patent citations of a paper are highly correlated with the combination of the paper's scholarly impact on these topics and the topics' patent citations.

Through further case studies, we a found strong relationship between papers and patents on the same topic. Patents are usually created based on prior papers, and the number of papers has a high correlation with the number of patents.

This research paper also has some limitations. The scope of data being tested only includes old papers in the security field, to mitigate the main difficulties of Altmetric, due to the time limitation and unavailability of API in the Lens database. In addition, these methodologies all have some limitations that can be further improved.

There is much future work that can be done based on the discoveries of this research. The distance metric can be attempted to identify basic studies. The noise of methodology 1 and methodology 3 can be potentially mitigated by text classification. The efficiency problem of methodology 3 may be reduced by testing the effect of setting a threshold. Moreover, the idea of using topics instead of direct citations may be applied to other indicators like news.

7 ACKNOWLEDGEMENT

I would like to sincerely appreciate my supervisors Ralph Holz and Abhishta Abhishta for their constant support and inspiration.

REFERENCES

[1] 2022. *The donut and Altmetric Attention Score – Altmetric.* https://www.altmetric.com/about-our-data/the-donut-and-score/
[2] Mohammad Ahmadpoor and Benjamin F. Jones. 2017. The dual frontier: Patented inventions and prior scientific advance. *Science* 357, 6351 (aug 2017), 583–587. https://doi.org/10.1126/SCIENCE.AAM9527/SUPPL_FILE/AAM9527_AHMADPOOR_SM.PDF
[3] Hassane Alami, Pascale Lehoux, Yannick Auclair, Michèle de Guise, Marie Pierre Gagnon, James Shaw, Denis Roy, Richard Fleet, Mohamed Ali Ag Ahmed, and Jean Paul Fortin. 2020. Artificial intelligence and health technology assessment: Anticipating a new level of complexity. https://doi.org/10.2196/17707
[4] Altmetric. [n.d.]. What are altmetrics? – Altmetric. https://www.altmetric.com/about-altmetrics/what-are-altmetrics/
[5] Altmetric. 2022. Altmetric Attention Score : Altmetric. https://help.altmetric.com/support/solutions/articles/6000233311-how-is-the-altmetric-attention-score-calculated-
[6] David Banta. 2003. The development of health technology assessment. *Health Policy* 63, 2 (feb 2003), 121–132. https://doi.org/10.1016/S0168-8510(02)00059-3
[7] Matt Blaze, Joan Feigenbaum, and Jack Lacy. 1996. Decentralized trust management. In *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*. IEEE Comput. Soc. Press, 164–173. https://doi.org/10.1145/3362168
[8] Lutz Bornmann. 2013. What is societal impact of research and how can it be assessed? a literature survey. , 217–233 pages. https://doi.org/10.1002/asi.22803
[9] Zhiqun; Chen. 2000. Java Card Technology for Smart Cards. *Taiwan health smart card project pp 7* (2000). https://books.google.com/books/about/Java_Card_Technology_for_Smart_Cards.html?id=H8VQAAAAMAAJ
[10] Ashish Garg, Jeffrey Curtis, and Hilary Halper. 2003. Quantifying the financial impact of IT security breaches. *Information Management and Computer Security* 11, 2-3 (2003), 74–83. https://doi.org/10.1108/09685220310468646
[11] Naveen Kumar and A Capstone. 2014. TODAY'S IMPORTANCE OF CYBERSECURITY. (2014).
[12] Jennifer Lin and Martin Fenner. 2013. Altmetrics in Evolution: Defining and Redefining the Ontology of Article-Level Metrics. *Information Standards Quarterly* 25, 2 (2013), 20. https://doi.org/10.3789/ISQV25NO2.2013.04
[13] John Mingers and Loet Leydesdorff. 2015. A review of theory and practice in scientometrics. , 19 pages. https://doi.org/10.1016/j.ejor.2015.04.002 arXiv:1501.05462
[14] Paul O'Callaghan, Lakshmi Manjoosha Adapa, and Cees Buisman. 2021. Assessing and anticipating the real world impact of innovative water technologies. *Journal of Cleaner Production* 315 (sep 2021), 128056. https://doi.org/10.1016/j.jclepro.2021.128056
[15] Dr. Paul Oldham. 2022. Chapter 7 Databases | The WIPO Manual on Open Source Patent Analytics (2nd edition). https://wipo-analytics.github.io/manual/databases.html#introduction-1
[16] Brian O'Rourke, Wija Oortwijn, and Tara Schuller. 2020. The new definition of health technology assessment: A milestone in international collaboration. , 187–190 pages. https://doi.org/10.1017/S0266462320000215
[17] Teresa Penfield, Matthew J. Baker, Rosa Scoble, and Michael C. Wykes. 2014. Assessment, evaluations, and definitions of research impact: A review. , 21–32 pages. https://doi.org/10.1093/reseval/rvt021

[18] Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger. 2012. Altmetrics in the wild: Using social media to explore scholarly impact. (mar 2012). https://doi.org/10.48550/arxiv.1203.4745 arXiv:1203.4745

[19] Mike Thelwall and Tamara Nevill. 2018. Could scientists use Altmetric.com scores to predict longer term citation counts? *Journal of Informetrics* 12, 1 (feb 2018), 237–248. https://doi.org/10.1016/j.joi.2018.01.008

[20] University of Twente. 2022. academic-societal-impact @ www.utwente.nl. https://www.utwente.nl/en/service-portal/university-library/research-impact/academic-societal-impact