Open University of the Netherlands
faculty of Management, Science & Technology
Master Computer Science of Master Software Engineering

Automating outlier detection in academic publishing

Author:
ing. Niels Tielenburg
851376878

Chair(wo)man:
prof. dr. Marko van Eekelen
Supervisor:
dr. ir. Hugo Jonker
2nd Supervisor:
prof. dr. Marko van Eekelen

Presentation date:
21-06-2017

June 10, 2017

Course code:
IM9906

Abstract

More and more, scientists are judged based on their research performance. As scientists are therefore under pressure to produce more and better research, some scientists focus on increasing on their research performance by fraudulent ways.

Current fraud detection measures are, however, insufficient. This research intends to design measures to assist and support fraud investigation, and develop tooling to automate application of these measures.

As fraudsters stand out to other scientists, they are outliers. We propose to investigate whether an outlier is a potential fraudster by comparing outliers to their scientific peers. The following research questions were formulated around this topic:

- RQ1: How to find scientific outliers?
- RQ2: How to compare the research output of scientific outliers to that of their scientific peers?

To find potential fraudsters, we propose to use a methodology using two phases, an outlier detection phase and a peer comparison phase. The outlier detection phase identifies outliers among a pool of scientists by investigating their citation and publication data. The outlier detection phase compares the outliers to their scientific peers, by investigating more detailed information about their publications and citations.

To find outliers among the publication and citation data of scientists, we propose to calculate certain measures and use an outlier detection mechanism. Outliers were found using an adjusted boxplot as the outlier detection mechanism, which takes skewness of data into account.

After outliers are found, outliers are compared to their peers in the peer comparison phase. Detailed data of the outliers and peers were acquired and used to calculate other measures. These measures consider different characteristic, targeted at indicating potential fraudulent behavior. After finding all the data necessary for the measures, the final step of the peer comparison phase is to compare the outliers to their peers.

Both phases were implemented using different publication data processors. The first phase was implemented using the publication data processors DBLP, Google Scholar and Semantic Scholar, where the peer comparison phase was implemented using Google Scholar and Elsevier as primary publication data processor.

The result of this research is a framework and different API's that can easily be changed, adapted and extended. Furthermore, we provided a set of indicative measures that could help indicate scientists who might be defrauding by increasing on their research performance in fraudulent ways.

Various experiments have been conducted to test the methodology. In an experiment to test the outlier detection phase, we were in most cases able to find significant outliers using different publication data processors. However, the combination of outlier detection mechanism, publication data processor and measure is in some cases not capable of finding significant outliers. We also performed an experiment investigating two outliers in the peer comparison phase. These outliers were compared to their peers by calculating the indicative measures and using an outlier detection mechanism. An experiment has also been performed to show outstanding scientists were indicated as outliers. Out of 23 outstanding scientists investigated, sixteen were identified as being an outlier.

Our methodology is therefore to some extent capable of finding the expected outliers. However, as we were not able to validate the methodology with actual fraudsters, we cannot conclude if the methodology is suitable of finding potential fraudsters. More research need to be done to verify if real fraudsters can be found using this methodology.

Contents

List of Figures

List of Tables

Abbreviations

AHCI Arts & Humanities Citation Index

API Application Programmable Interface

captcha completely automated public Turingtest to tell computers and humans apart

DBLP DataBase systems and Logic Programming

ESCI Emerging Sources Citation Index

HTML HyperText Markup Language

IQR Interquartile Range

IT Information Technology

LNCS Lecture Notes in Computer Science

MAD Median Absolute Deviation

PC Personal Computer

PhD Philosophiae Doctor

SCIE Science Citation Index Expanded

SSCI Social Sciences Citation Index

USB Universal Serial Bus

WoS Web of Science

XML eXtensible Markup Language

1 Introduction and Motivation

More and more, scientists are judged based on their research performance. Research performance plays a key role in making funding and hiring decisions[Whi; KJ06]. For example: the Dutch science funding agency NWO requires proposals for TOP Grants Chemical Sciences to include research performance of the last five years[1]. Thus, scientists are under pressure to produce more and 'better' research.

Research performance is often measured by looking at the quantity and quality of research. Research quantity can be measured easily. Quality of research, however, cannot. To overcome this, quality is often equated to scientific impact. To determine the impact of research in an objective way, various author level metrics have been proposed. As the pressure is high to increase research performance, scientists might be tempted to behave fraudulent. Reports have already shown scientists behave fraudulent to improve on their research performance. Scientists, for example, created false accounts and reviewed papers themselves [Hau15], and they have been stealing ideas and presented them as their own [Smi06]. Numerous examples of scientists who committed fraud that have eventually been caught exists[2]. These scientists evaded detection by existing fraud-detection mechanisms, sometimes for years[3]. Current fraud detection measures are therefore insufficient. This research serves two main purposes:

- Design measures to assist and support fraud investigation
- Develop tooling to automate application of these measures

Methodology We propose to systematically investigate output characteristics from different public sources. As the amount of scientists and the amount of data is enormous, we propose to use a two-phase approach. First, we investigate citation and publication measures, thereby identifying a pool of outliers. Next, we take a closer look at those outliers by investigating various other measures. We will develop tooling that calculates and combines these measures, thereby automating the investigation of fraudulent behavior and minimizing interaction.

As fraudsters try to imitate outstanding researchers, using this methodology might result in finding potential outstanding scientists and potential fraudulent scientists. Our methodology and tools are not able to tell the difference. The aim for this research is therefore not to provide evidence of fraud. The aim is to provide a systematic approach to find outliers that might have committed fraud. Further (manual) inspection of every outlier is still necessary to determine whether the outlier has actually engaged in fraud.

Contributions Currently, there's no systematic approach in identifying suspects of fraudulent behavior. This research aims to change that by contributing in the following ways:

- Design a systematic approach for finding potential fraudsters
- Develop automated tooling minimizing interaction
- Initial attempt at finding and using measures for the detection of potential fraudsters

[1] http://www.nwo.nl/en/funding/our-funding-instruments/nwo/top-grants/top-grants-chemical-sciences/top-grants/top-grants.html

[2] http://www.onlineuniversities.com/blog/2012/02/the-10-greatest-cases-of-fraud-in-university-research/

[3] http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html

Ethical considerations Repercussions for scientists guilty of fraudulent behavior can be severe. Scientists could, for example, be fired from the university they work for. As already mentioned, outstanding behavior cannot be distinguished from actual fraud using the systematic approach proposed in this research. Scientists found to be outliers, even when compared to their peers, can therefore not immediately be marked as fraudsters, but only as potential fraudsters. As such, no names of will be mentioned in this research, as this might have severe consequences for the scientist. Every scientist found to be a potential fraudster should be investigated by different means to assure the scientist is not falsely accused of fraud.

Thesis overview This thesis will first provide in some background information on the peer review process, available publication data processors and different kind of bibliometrics that can be used to measure quality of research. Next, we will describe the problem statement, immediately followed by the methodology we propose to assist and support fraud investigation. The next section outlines related work, where we describe, among else, some of the attacks found in practice. Section 6 explains the measures used during the outlier detection phase, followed by a section that describes the outlier detection method used. The following section explains the peer types used and their suitability to be used as peers, to be followed by a section on the comparison of outliers with their peers. The methodology has been implemented, which is described in section Implementation, and to test the methodology and implementation, experiments have been conducted which are described in the section Experiments. Finally, the thesis is finished by a conclusion and some proposals for future work.

This research was executed as the final project for the degree Master of Science in Software Engineering.

2 Publishing process and Bibliometrics

2.1 Publication types

Scientists can publish publications in a range of venues. Every venue is specialized in a certain subject, has different demands for publications that will be published, and is of different quality. Scientists therefore select a venue to which they want to submit or present their publication. In IT, important venues are journals, and workshop and conference proceedings.

Peer-review process A peer-review process is used for many kinds of venues[4,5,6,7]. This process is conducted by one or multiple editors. It has different stages and always starts with submitting a publication by a scientist. The publication will first be checked on basic criteria, such as the importance of the topic, grammar and the relevance of the publication for the venue. If the publication meets these basic criteria, the publication will be send to the chief editor. The chief editor can decide to get the publication peer-reviewed.

There are three types of peer-review possible[7]:

- Open Review. The author of the publication will know who the reviewer and the reviewer will know the author.
- Single Blind Review. The author of the publication will not know the reviewer.
- Double Blind Review. The author of the publication will not know the reviewer, and the reviewer(s) will not know the author.

As all types have up- and downsides, there may be a difference in the type of peer-review used between different venues.

After peer-reviewing a publication, commentary will be send back to the author. The author may revise the publication, or respond to the commentary. After the publication is revised, the author can send it back for a second review. The chief editor will finally decide if the publication is ready to be published. If the publication is rejected at any stage, the author can decide to revise the publication and submit it to another venue to try getting it published.

Editors are responsible for finding reviewers and delegating publications to be reviewed to the reviewers. Editors may also decide to review publications themselves.

Together, editors and reviewers form the editorial board or a program committee chair. All members are active in the field of the venue[8,9]. Members are also responsible for:

- Approaching potential contributors for the venue.
- Identifying peer reviewers.
- Identifying new topics for special issues.

As every member is an expert in some field, it is possible a specific member of the board always edits publications on a particular subject.

[4] https://icer.hosting.acm.org/icer-2015/review-process/
[5] http://www.markbernstein.org/elements/Reviewing.pdf
[6] https://www.springer.com/gp/adis/resources/for-authors/the-review-process
[7] https://www.elsevier.com/reviewers/what-is-peer-review
[8] https://www.springer.com/gp/authors-editors/editors/editorial-boards/32688
[9] https://www.elsevier.com/editors/editorial-boards

2.2 Publication data processors

After a publication has been indexed by a venue, metrics can be calculated about the publication. There are multiple publication data processors that use publications published in venues to calculate metrics. Some of these publication data processors keep track of how many publications have been published by a certain author (Publication Database), and some publication data processors also keep track of the amount of citations to and from publications (Citation Database).

2.2.1 Citation Databases

A citation database is an index of citations between publications. Citation databases can be used by researchers to extract, for example, patterns and links between publications, authors and venues. Data acquired from a citation database may include, for example, the amount of citations to a specific publication. There are at least four databases that could be used for citation analysis: Google Scholar, Web of Science, Scopus and Semantic Scholar.

Google Scholar "Google Scholar includes scholarly articles from a wide variety of sources in all fields of research, all languages, all countries, and over all time periods"[10]. It uses automated software (crawlers or robots) to search the internet for publications. To let the crawler or robot search your website, the website must be structured in a particular way. When a publication is found, it will be parsed to identify bibliographic data and references. Therefore, when you want your publication to be indexed by Google Scholar, you need to publish your publication in such a way the parser is able to identify certain properties. This includes meta information like the title, the author and the publication date.

Web of Science The Web of Science (WoS) includes four citation indexes: Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), the Arts & Humanities Citation Index (AHCI) and the Emerging Sources Citation Index (ESCI)[11]. Three of these (the SCIE, SSCI or AHCI) are considered the flagship indexes. ESCI covers all areas covered in SCIE, SSCI or AHCI, but journals covered in ESCI are not covered in SCIE, SSCI or AHCI. All four citation indexes have certain criteria to which journals are selected to be indexed.

WoS is selective about journals as only the most important papers are published in relatively few journals, and most of the citations come from relatively few journals[11]. As the indexes are not static, the composition of the indexes can change by adding or removing journals.

Scopus Scopus is an "abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings"[12]. Scopus uses a content selection and advisory board to decide which journals, books and conference proceedings will be selected. The board consists out of sixteen persons, all representing a specific field.

Scopus uses a set of benchmarks and metrics to determine the quality of the journal over time. After some time it may be that the journal is not of sufficient quality anymore. To detect the deterioration of the quality, the journal is compared to peer journals in the same field. If

[10] https://scholar.google.nl/intl/nl/scholar/about.html

[11] http://wokinfo.com/essays/journal-selection-process/

[12] https://www.elsevier.com/solutions/scopus

the journal stands out too much, it will be red flagged. The journal will be informed and might try to solve the issue. If the journal fails to do so, the journal will be removed from Scopus.

Different selection criteria exists for books and conference proceedings[13].

Semantic Scholar Semantic scholar currently indexes millions of publications, related to computer science and neuroscience[14]. It uses techniques such as machine learning and machine vision to help find publications faster than other sources[15]. Only high-quality publications are indexed in Semantic Scholar, using "carefully tuned mechanisms"[16].

2.2.2 Publication Databases

DBLP DataBase systems and Logic Programming (DBLP) is “an on-line reference for bibliographic information on major computer science publications”[17]. DBLP is free to use and it provides in access to bibliographic meta-data and links to electronic editions of publications. Currently, it indexes over 3.3 million publications and 1.7 million authors. The complete database of DBLP can be downloaded as an XML file. DBLP only covers publications from computer science. Therefore, DBLP has an advisory board consisting out of a number of computer scientists, each a specialist in a different computer science field.

Arxiv Arxiv is an “electronic archive and distribution server for research articles” [18]. To get archived on Arxiv, a scientists need to create an account and submit his or hers publications. The publications do not get peer reviewed, but may be moderated [19]. In 2014, Arxiv had one million preprints archived [20].

2.3 Bibliometrics

Bibliometrics is statistical analysis of books, articles, or other publications[21]. It is used to compute and evaluate research performance of journals, publications and authors. Bibliometrics can be computed with help of publication data processors.

Bibliometrics can be divided into multiple sub-categories[22]. Author-level metrics is the sub-category to compute research performance of authors, and is therefore of great importance for scientists. As this research investigates authors improving their research performance, this sub-category is also of importance to this research. Some of the metrics used to calculate research performance are:

[13] https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection

[14] https://www.semanticscholar.org/faq

[15] http://www.sciencemag.org/news/2016/11/computer-program-just-ranked-most-influential-brain-scientists-modern-era

[16] https://www.semanticscholar.org/faq

[17] http://dblp.uni-trier.de/

[18] https://arxiv.org/help/general

[19] https://arxiv.org/help/moderation

[20] http://www.the-scientist.com/?articles.view/articleNo/41677/title/Q-A--1-Million-Preprints-and-Counting/

[21] http://stats.oecd.org/glossary/detail.asp?ID=198

[22] https://www.elsevier.com/solutions/scopus/features/metrics

H-index The h-index is computed by determining how many papers a person has written and how many times those papers have been cited. For example: if a person has an h-index of x, this scientist has written at least x papers, each of which has been cited at least x times. The h-index, proposed by Hirsch, is an easily computable index. The h-index "gives an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions"[Hir10]. The total number of papers and citations are normally larger than the h-index. Hirsch empirically found out that the total amount of citations is even three to five times larger than the minimum amount of citations needed for a particular h-index.

Author Impact Factor The author impact factor is computed by first determining all the papers a person has written during year $t-1$ and year $t-2$. The next step is obtaining citations to those publications in year t. The author impact factor is finally computed by dividing the number of citations by the amount of papers written during year $t-1$ and year $t-2$. The author impact factor "is capable to capture trends and variations of the impact of scientific output of scholars in time"[PF13].

G-index It is computed by determining the number of publications g, that received a total number of g^2 citations. The g-index was introduced as an improvement of the h-index[Egg06]. The g-index was introduced as the h-index is insensitive to one or several outstandingly highly cited publications.

i10-index The i10-index is simply computed by determining the number of publications which received at least ten citations.

3 Problem statement

Whenever a metric is being used, behavior changes. This is also known as Goodhart's law: "When a measure becomes a target, it ceases to be a good measure." Metrics used to measure research performance play a key role in making funding and hiring decisions[Whi; KJ06; Law07]. Scientists are therefore tempted to use attacks to increase their research performance. However, using attacks to increase research performance is unethical. Scientists using these kind of attacks should therefore be found.

To investigate a scientist of fraud is a lengthy and costly process. For example: it took three persons over one year to investigate Stapel[23]. In practice, the decision to investigate a scientist is therefore based on a strong indication of fraud. Currently, no systematic approach exists to identify scientists engaging in fraud. Fraudulent scientists are caught either by luck or flagrant behavior.

Defrauding and excellent scientists have at least one thing in common: they stand out one way or another to the other scientists. Therefore, these scientists are outliers. However, not all outliers are interesting enough to be further investigated. To determine which outliers are interesting enough to be investigated, outliers could be compared to their scientific peers. If a scientist is also an outlier when compared to their scientific peers, then the scientist is interesting enough to be investigated further.

The research questions formulated around this problem statement are:

- RQ1: How to find scientific outliers?
- RQ2: How to compare the research output of scientific outliers to that of their scientific peers?

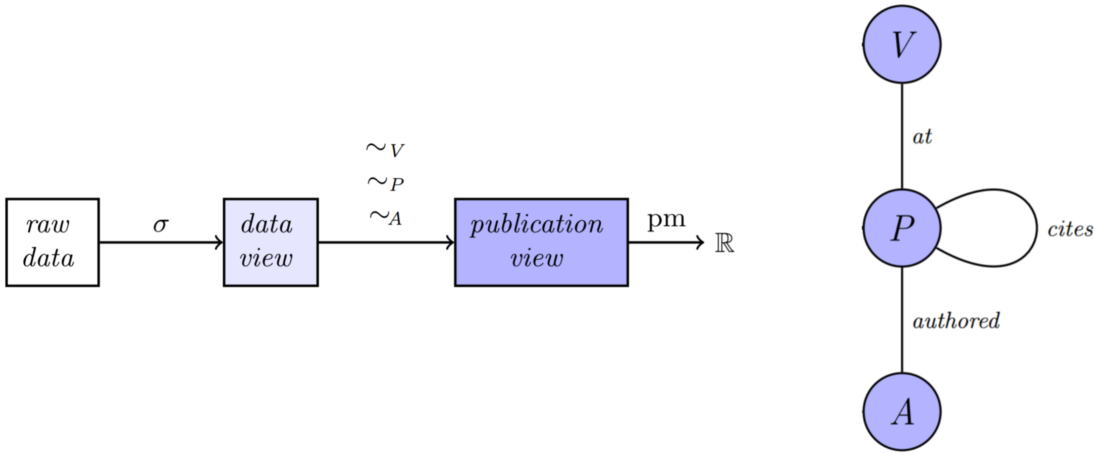
We try to answer these research questions by only looking at publicly available data from a select amount of publication data processors. Furthermore, we will only limit our scope to scientists active in the field of IT. However, defrauding scientists try to mimic outstanding scientists. This research will therefore not only focus on identifying scientists who might be engaging in fraud to increase their research performance, but also focus on identifying potential excellent scientists.

The purpose of this research is not to automate every step, but to support the process with help of data. We think it will never be possible to prove with 100% certainty it will be possible to find fraudsters by automated means. Human interaction will always be necessary to investigate potential fraudsters.

This research is a continuation on existing research [JM17]. Notations used in this research originated from that research. Figure 1a displays how a publication data processor calculates metrics. For example, Google Scholar searches the internet (raw data) to obtain a data view. Of that data view a publication view is constructed by finding out, for example, which publication cites which publication or who wrote which publication. From that publication view, metrics can finally be calculated. Figure 1b displays a publication view, where the rest of the notations used in this research can be found. A, P and V indicate, respectively, the sets of authors, papers, and publication venues. authored indicates the relation where an author authored a publication, cites indicates the relation where a publication cites a publication or the relation

[23] https://www.tilburguniversity.edu/nl/over/profiel/kwaliteit-voorop/commissie-levelt/

where a publication contains a citation to a publication, and at indicating the relation at what venue a publication is published.



(a) (b)

Figure 1: (a) From data to metric and (b) Induced publication view, taken from [JM17].

4 Methodology

We propose to use a methodology to compare scientists to their scientific peers, in order to determine whether a scientist might be using fraudulent behavior, and should therefore be further investigated. Scientists should roughly display the same characteristics in their scientific data when compared to their scientific peers. If a scientist does not display the same characteristics in their scientific data by a substantial amount, it might be the scientist uses fraudulent behavior.

To be able to make a peer comparison, we propose to calculate certain measures. These measures are expressed using the sets and relations in Figure 1b. However, the amount of data necessary for these sets and functions is prohibitive. For example, if we want to acquire all the incoming citations to a certain publication p, we would need to investigate every publication of a publication data processor whether it cites the publication p. Such a prohibitive amount of data is necessary, that this would only be feasible if we have complete access to a publication data processor's database. We can only use the (limited) publication view on the data (see Figure 1a) provided by the publication data processors. Furthermore, not every scientist need to be compared to their scientific peers. Some scientists with, for example, one publication and one citation do not need to be peer compared. As data acquisition takes time and effort, and we can only zoom in on detailed data for a limited amount of scientists, we would only want to use those resources sparingly and only do the peer comparison for the scientists that stand out.

Therefore, we propose to use a two-phase approach, an outlier detection phase and a peer comparison phase. In the outlier detection phase we will find outliers among a set of scientists, that stand out one way or another. In this phase, we only need to acquire certain publication and citation data about all the scientists we want to investigate. Using this data, we propose to calculate certain measures, specifically designed to indicate scientists using potential fraudulent behavior. In the peer comparison phase of the methodology, first a set of outliers is necessary, which can be acquired by our outlier detection phase. Next, we propose to compare these outliers to their peers. Therefore, first we need to acquire the scientific peers of an outlier. Next, we need to acquire more detailed data to calculate other measures for the outlier and the peer-group. Finally, the measures of the outlier can be compared to the measures of the peer-group, by investigating whether the outlier is also an outlier when compared to the peer-group. If an outlier is also an outlier when compared to the peer-group, this scientist is a potential fraudster, especially when the outlier is an outlier on multiple measures. This scientist could therefore be further investigated for fraud. This methodology is graphically depicted as a framework in Figure 2.

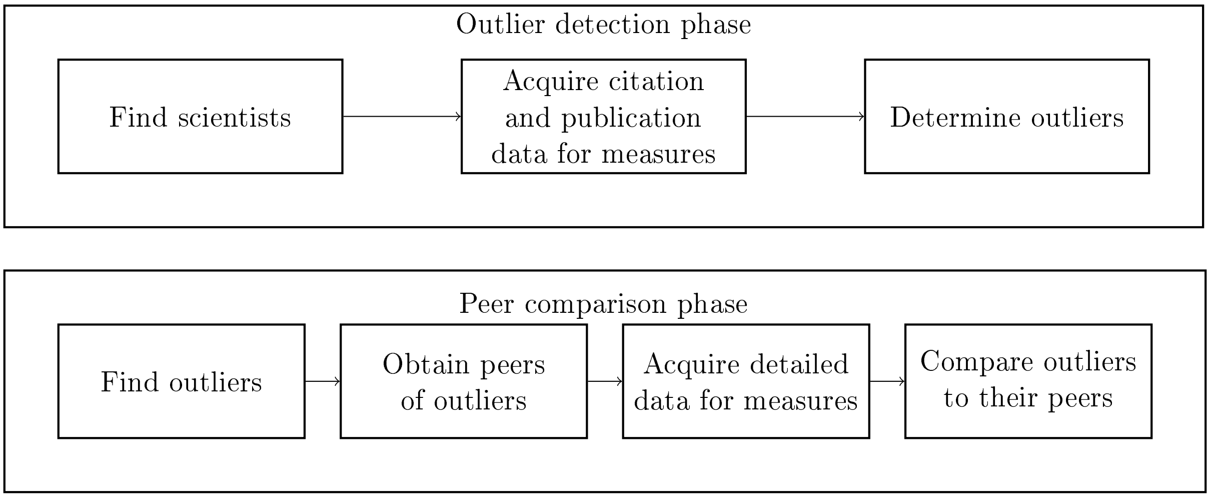

Outlier detection phase
Find scientists
Acquire citation and publication data for measures
Determine outliers
Peer comparison phase
Find outliers
Obtain peers of outliers
Acquire detailed data for measures
Compare outliers to their peers

Figure 2: Methodology as a framework.

4.1 Data used by methodology

In the two phases, different kind of data is used. In the outlier detection phase, only citation and publication data is used of all the scientists found in the scientist database. With help of the citation and publication data, measures can be calculated. The outliers between the values can be found by an outlier detection method.

In the peer comparison phase, more detailed information is necessary to be able to calculate all of the proposed measures. The proposed measures use, for example, meta-data about specific publications and specific venues. As we need to acquire peers of the outliers, a peer database is also necessary to find peers.

Figure 3 displays the different kind of data used by the two phases.

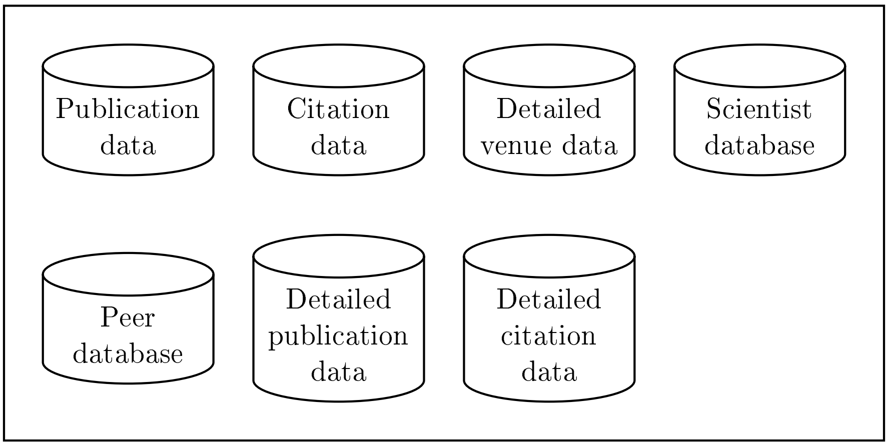


Figure 3: Data used in the phases

5 Related Work

5.1 Attacks found in practice

Different kind of attacks to increase author-level metrics have already took place. One attack is self plagiarism [CK05]. With this attack, scientists construct new publications, using fragments of old publications, and try to get these published. Another attack is to misuse your position in an editorial board. Different kinds have already taken place, such as the citation ring attack [Hau15; FMO14]. This attack uses fake accounts to peer-review publications. With this attack, an editor creates false accounts and sends publications to those accounts to be reviewed. The editor can review these publications, and approve the publications without a proper peer-review. Reviewers might also coerce citations [WF12]. Reviewers or editors may only accept a publications to be published if it contains a certain citation. It is also possible to make excessive use of self citing [LRT12]. To receive more citations, scientists can easily cite (non-relevant) publications they published themselves. Scientists might also form a cartel and work together to receive more citations[24]. Scientists will cite each other publications using this attack. Salami slicing is another attack already used [Rog99]. Using this attack, scientists divide their research in as many publications as possible, just to get as many publications as possible out of it.

5.2 Fraud detection software

Software has already been developed to detect certain kind of fraud. SciDetect is an open-source program to detect automatically generated papers created with SCIgen and similar programs [Boh15]. With SCIgen it was possible to create generated papers that looked as written by researchers, but on closer inspection looked clearly fabricated. To detect plagiarism, many programs are available [AAS11]. iThenticate[25] is an example of a service that uses plagiarism detection software. This service is being used by Elsevier, IEEE, Springer, and other scholarly journals to detect scientific fraud. A tool specifically designed to find self plagiarism is also available [Col+03]. Although not specifically developed as fraud detection software, Publish or Perish “is a software program that retrieves and analyzes academic citations”[Har07]. With help of this software, for example, citation analysis is possible that can be used to detect fraudulent behavior.

5.3 Outlier detection

Outlier detection is used in different topics of IT. It is used, for example, in the detection of outliers in network traffic [Ste12]. An outlier detection algorithm was used where full network payload data or low-level access to the hardware was not available. Software was also developed to track down software bugs, using outlier detection techniques[HL02]. While a program is running, the software observes its behavior and tries to detect errors and their causes. Sensor networks is another topic were outlier detection can be applied. As data volumes can be very large in sensor networks, outlier detection can be used to only send the relevant data to save energy[She+07]. Outlier detection has also been effectively used to detect botnet clients[BS06].

[24] `https://scholarlykitchen.sspnet.org/2012/04/10/emergence-of-a-citation-cartel/`

[25] `http://www.ithenticate.com/`

6 Outlier detection

To indicate potential fraudsters, it is necessary to understand the citation and publication behavior of scientists active in the field of IT. After a characterization of this behavior, it is possible to define what kind of data characteristics potential fraudsters display, and measures can be used to indicate these characteristics in the data.

6.1 IT Publication and citation model/observations

To find outliers, we defined assumptions that describe the behavior of publication and citation data of scientists active in the field of IT. These assumptions have been realized along with studying various sources of citation and publication data. The following assumptions have been made by observing the publication and citation model:

Assumption 1 The amount of publications a scientist can produce, without collaborating, will not likely exceed eight per year.

Assumption 2 A scientist is not likely to collaborate on more than sixteen publications per year.

Assumption 3 In general, the more publications a scientists has produced, the more citations the scientist has received.

Assumption 4 There are scenarios possible, such as while completing a PhD, where scientists can show a high increase in their amount of publications produced per year. After some time, however, it is expected the amount of publications produced per year stays roughly constant, as writing good quality publications take time.

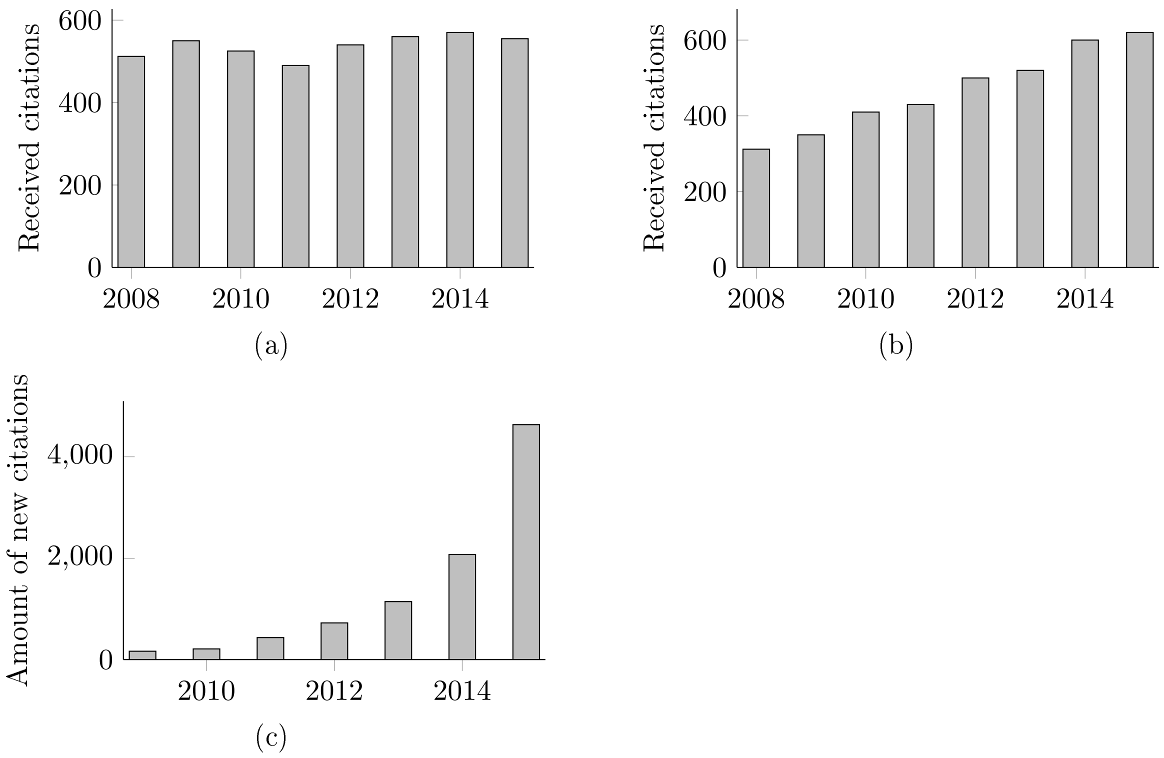
Assumption 5 After scientists reach a threshold of roughly 1,000 citations, the derivative of the amount of citations per year will not increase rapidly. Most scientists show a linear growth (see Figure 4b), or stay constant (see Figure 4a). Trends showing a quadratic or even faster growth is not sustainable in the long run, such as depicted in Figure 4c.

6.2 Potential fraud characteristics

With the observations described in Section 6.1 it is possible to characterize the data displayed by potential fraudsters. As fraudsters try to mimic outstanding researchers, these characteristics might indicate scientists using fraudulent behavior, but they also might indicate outstanding scientists. Therefore, not only potential fraudsters will be found, but also potential outstanding scientists.

To indicate the characteristics, we propose to calculate certain measures of each scientist. Outliers can be found among the values by an outlier detection mechanism. The measures are divided into the following categories, each category indicating different characteristics of potential fraudster:

- Measures using publication data
- Measures using citation data
- Measures using publication and citation data



(a)

(b)

(c)

Figure 4: (a) Stable citation count, (b) Linear growth of citation count and (c) Non-linear growth of citation count

6.2.1 Measures using publication data

According to Assumption 4, the amount of publications produced per year should stay roughly constant. To publish publications, there are certain standards that must be met. Each venue has its own standards. To meet those standards, publications must be of a certain quality, and producing publications of good quality takes time. If a scientist publishes a lot of publications per year, while the scientist did not do that in previous years, a scientist might be researching a problem that produces lots of new insights that are all fit to be published. However, it might also indicate a scientist using questionable means just to gain lots of published publications.

To find these outliers, a measure is proposed that calculates the maximum of the derivative of publications published per year. However, scientists can have multiple publications being peer-reviewed, where all of these publications could be published in the same year, which may lead to a very high value of this measure. Therefore, the possibility must be taken into account that no publications are published in a certain year, as all of them are being peer-reviewed, and all of the peer-reviewed publications get published in the next year. As it normally will not take longer than two years to get a publication published, we average the amount of publications over two years to dampen fluctuations, before taking the derivative of this set of values. For a scientist OP, the measure is calculated by Equation 3, where the set of publications of OP is given by Equation 1, and the amount of publications of a given year is given by Equation 2.

$$\text{pubs}(OP) = \{p \in P \mid \text{authored}(OP, p)\}. \tag{1}$$

$$\mathit{pubsinyear}(OP, \mathit{year}) = |\{p \mid p \in \mathrm{pubs}(OP) \land p.\mathit{year} = \mathit{year}\}|. \tag{2}$$

$$\mathit{maxdiffpubcount}(OP) = \max_{y} \left(\sum_{i=y-1}^{y} \mathit{pubsinyear}(OP, i) - \mathit{pubsinyear}(OP, i-1) \right). \tag{3}$$

Another characteristic can also be investigated when investigating the publication data of scientists. According to Assumption 1 it is possible scientists can produce up to eight publications per year. The same scientist can, in the same year according to Assumption 2, collaborate on sixteen publications. Therefore, scientists can produce a maximum of 24 publications per year. For a scientist *OP* we could simply determine if the scientist published more than 24 publications in a year. If this occurs, the scientist should be added to the set of potential fraudsters. Note that this is not an outlier detection measure. Outliers are not being determined using this measure. This measure simply investigates the characteristic whether more than 24 publications are published in a certain year. We will refer to this measure in the following sections using the name *toomanypubs*.

6.2.2 Measures using citation data

Every publication a scientist publishes can receive citations. As the amount of citations can roughly be related to the amount of publications (Assumption 3), and the amount of new publications produced will likely stay constant (Assumption 4), the amount of citations will also stay constant or increase linear. According to Assumption 5, the derivative in the amount of citations will therefore not increase rapidly. However, if the derivative is very high, some of the research, for example, might be of outstanding quality. A scientists could also have tried to increase on the amount of citations in a fraudulent manner. In both cases, the scientist will have citation data that is deviant when compared to the other scientists.

To find these outliers, a measure is proposed that calculates the maximum of the derivative of the amount of citations received per year. However, as it is possible a few publications are being peer-reviewed (see Section 6.2.1), it will also be possible an abnormal high increase in citations will follow in one or two years after the publications are published. And again, as it normally will not take longer than two years to get a publication published, the average amount of citations of two years is calculated to dampen fluctuations, before taking the derivative of this set of values. For a scientist *OP*, the measure is calculated by Equation 5, where the amount of citations of *OP* of a given year is given by Equation 4.

$$\mathit{citesinyear}(OP, \mathit{year}) = \big|\{p \mid p \in P \land p' \in \mathrm{pubs}(OP) \land \mathit{cites}(p, p') \land p.\mathit{year} = \mathit{year}\}\big|. \tag{4}$$

$$\mathit{maxdiffcitecount}(OP) = \max_{y} \left(\sum_{i=y-1}^{y} \mathit{citesinyear}(OP, i) - \mathit{citesinyear}(OP, i-1) \right). \tag{5}$$

6.2.3 Measures using publication and citation data

The more publications a scientist produces, the more citations are expected (Assumption 3). Therefore, the amount of extra citations per year can also be compared to the amount of extra publications of the previous year (assuming the citations occur after publication). If this quotient is very large, the researcher in question received a large amount of extra citations, while the amount of extra publications produced of the previous year was less.

At least two scenarios are possible where a scientists can obtain a large quotient: the researcher might have received these citations by publishing excellent research, or the researcher used fraudulent ways to improve on bibliometrics. As already mentioned before, it is necessary to take an average, as there is a possibility a few publications are being peer-reviewed and all of these get published in the next year. Again, as it normally will not take longer than two years to get a publication published, the average of two years is calculated. And, as scientists receive citations after a publication is published, we take the publication data preceding the citation data by one year. For a scientist OP, the maximum ratio of the derivative of the number of citations versus the derivative of the number of publications is calculated by Equation 6.

$$\mathit{maxratiocitsvspubs}(OP) = \max_{y} \left(\sum_{i=y-1}^{y} \frac{\mathit{citesinyear}(OP, i) - \mathit{citesinyear}(OP, i-1)}{\mathit{pubsinyear}(OP, i-1) - \mathit{pubsinyear}(OP, i-2)} \right). \quad (6)$$

A caveat of this measure is the possibility that scientists suddenly stop producing publications by, for example, retiring or passing away. In that case, the amount of publications can become zero. If the scientist keeps receiving citations, the resulting quotient can become very large.

6.3 Combining outcome of the measures

Outliers need to be found of all the measures, except for the measure where we indicate scientists producing more than 24 publications, with help of an outlier detection mechanism. After outliers are found for each individual measure, the result is eventually different sets of outliers for each measure. During the peer comparison phase we are only interested in a single set of outliers. We therefore propose to combine all of the resulting sets of outliers into one set. For example, if a publication data processor contains only publication data and therefore the measures *maxdiffpubcount* and *toomanypubs* are used, the set of scientists to investigate in the next phase could be given by Equation 7.

$$\mathit{InvestigateInNextPhase}(OP) = \mathit{maxdiffpubcount}(OP) \cup \mathit{toomanypubs}(OP). \quad (7)$$

7 Outlier detection method

After the measures are calculated for each scientist, the outliers among these values should be determined. To be able to do this, an outlier detection method need to be used. There are many outlier detection methods, each with up- and downsides. Some of these methods are [Ole11] [26,27]:

- Standard deviation method: Every sample outside the interval $[\bar{x} - a * \sigma; \bar{x} + a * \sigma]$ is an outlier (where a is $>$ 0 and σ the standard deviation).
- Z-score: the z-value is calculated by $Z_i = \frac{Y_i - \bar{Y}}{\sigma}$. If the Z_i-value of $Y_i > 3.5$, Y_i is considered to be an outlier.
- Modified z-score: the m-value is calculated by $M_i = \frac{0.6745(x_i - \bar{x})}{median(|x_i - \bar{x}|)}$. If the M_i-value of $x_i > 3.5$, x_i is considered to be an outlier.
- Tukey's method (Boxplot): every sample outside the interval $[Q_1 - 1.5IQR; Q_3 + 1.5IQR]$ is considered to be an outlier (where Q_1 is the 25th percentile value, Q_3 is the 75th percentile value and IQR the distance between Q_3 and Q_1)
- Adjusted boxplot: The same as Tukey's method, except the interval is adjusted by multiplying the value 1.5 with a skewness factor, to take the skewness of the data into account.
- MAD: every sample outside the interval $[\bar{x} - a * (b * median(|x_i - \bar{x}|)); \bar{x} + a * (b * median(|x_i - \bar{x}|))$, is considered to be an outlier (where b depends on the distribution of the data, and a is $>$0).
- Median rule: every sample outside the interval $[\bar{x} - 2.3IQR; \bar{x} + 2.3IQR]$ is considered to be an outlier.

The shape of the data is one of the most important aspects when determining which method to use to find outliers. If data is normally distributed, and therefore symmetrical, all of the listed methods are proper candidates to be used. However, the proposed measures do not result in symmetrical data, all of the measures result in skewed data. Not all of the methods listed are capable of handling skewness in the data. Therefore, some of these methods can be rejected at once. Of the remaining methods (Tukey's method, MAD, the Median rule, and the Adjusted boxplot), the adjusted boxplot especially takes into account the skewness of the data[Seo06]. This outlier detection method was therefore chosen to determine the outliers.

7.1 Adjusted boxplot

One of the most commonly used method for determining outliers is the boxplot method, also known as Tukey's method[WS11] (see Figure 5).

[26] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

[27] https://www.r-bloggers.com/absolute-deviation-around-the-median/

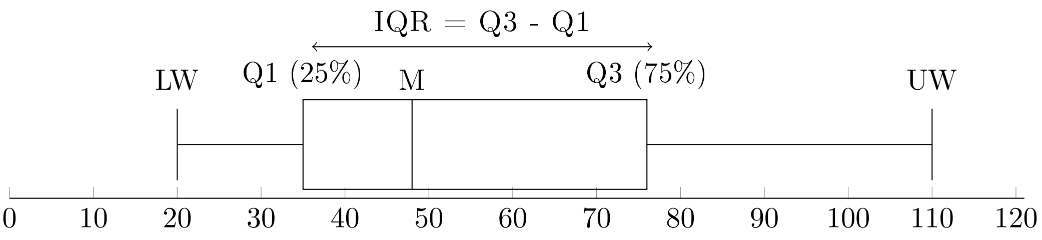


Figure 5: Boxplot example

A boxplot uses quartiles calculated from the data. The first quartile is calculated by taking the 25th percentile and the third quartile is calculated by taking the 75th percentile of the data (indicated by Q1 and Q3 in Figure 5. The median of the data is also visualized in a boxplot, indicated by M. To find outliers, two whiskers are calculated, the lower whisker and the upper whisker (indicated by LW and UW respectively in Figure 5). These whiskers are calculated by taking the inter quartile range (indicated by IQR in Figure 5), and subtracting this distance from Q1 to obtain the lower whisker, and adding it to Q3 to obtain the upper whisker. If there is no sample in the set equal or lower than the lower whisker, the lower whisker will be the lowest sample of the set. The same holds for the upper whisker, the upper whisker will be the highest value of the set in case the calculated upper whisker is higher than the highest value of the set. In all cases, an outlier is always a sample outside the interval

$$[Q_1 - 1.5IQR; Q_3 + 1.5IQR]$$

However, as stated before, the boxplot method only works well if the data is not skewed. If the data is skewed, too many points are considered to be outliers. An adjustment on the boxplot was proposed in [HV08] that takes the skewness of the data into account. A skewness measure measures the asymmetry of data with a value ranging from -1 to 1, where a negative skew indicates a right-leaning curve (Figure 6b) and a positive skew indicates a left-leaning curve (Figure 6a).

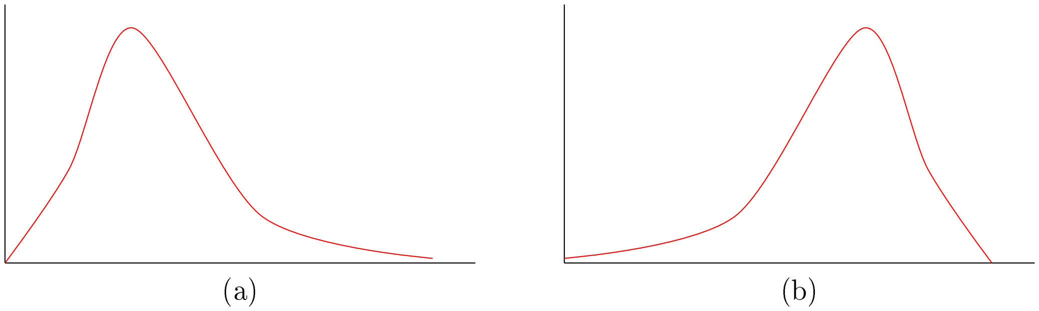


Figure 6: (a) Positive skew and (b) Negative skew

Different measures to indicate the skewness exist, where the overall best measure indicating skewness is the medcouple(MC)[BHS03; BHS04]. The proposed adjustment for the boxplot method also uses the medcouple to indicate the skewness of the data. By an empirically conducted study, the following interval was proposed for the adjusted boxplot:

$$[Q_1 - 1.5e^{-3.5MC}IQR; Q_3 + 1.5e^{4MC}IQR]$$

Three possibilities can occur using this adjusted interval:

1. If the data is not skewed, the medcouple will be 0 and the original interval of the boxplot will be used.
2. If the data is skewed to the right, the medcouple will indicate the negative skew by resulting in a value between -1 and 0 and the interval will be adjusted to include more samples of the tail on the left side.
3. If the data is skewed to the left, the medcouple will indicate the positive skew by resulting in a value between 0 and 1 and the interval will be adjusted to include more samples of the tail on the right side.

As the data is skewed and the adjusted boxplot is an excellent way of determining outliers, the adjusted boxplot was used to find the outliers. The proposed interval of the adjusted boxplot will also be used for detecting our outliers. However, as this interval is proposed using an empirically conducted study, it might not suit our own data set.

8 Peers of outliers

In the peer comparison phase of the methodology we need to compare the outliers found in the outlier detection phase to their peer-group. Many scientists or even publications can be classified as peers of a certain scientist in one way or another. We identified the following categories as peers:

- Co-authors
- Co-editors and co-program committee members
- Co-publications

8.1 Peer types

Co-Authors Co-authors are one of the obvious choices for peers. Scientists write publications with other scientists in the same field, and their data should therefore display roughly the same characteristics. The co-authors of peers can be determined by Equation 8. However, co-authors can also have a different connection than only co-authorship. For example, a PhD student can have a publication with his or hers supervisor as a co-author. In this case, the supervisor is expected to have different characteristics in citation an publication data.

$$\mathit{coauthorpeers}(OP) = \{a \in A \setminus \{OP\} \mid p \in \mathrm{pubs}(OP) \wedge \mathrm{authored}(a, p)\}. \tag{8}$$

Co-editors and co-program committee members Co-editors and co-program committee members of peers can also be used as a peer-group to compare to the outliers. An outlier which is an editor or a program committee member might have used his or hers influence to obtain more citations and/or publications, by, for example, influencing which papers are published. To help indicating outliers that might have used this kind of behavior, the outlier can be compared to their co-editors and/or co-program committee members as the peer-group. The co-program committee members can be determined by Equation 9. A similar equation can be used to determine the co-editors.

$$\mathit{comemberpeers}(OP) = \{a \in A \setminus \{OP\} \mid v \in V \wedge \{OP, a\} \subseteq v.\mathit{reviewers}\}. \tag{9}$$

Co-publications Peers do not necessarily need to be persons. Peers can also be other publications, when, for example, we want to investigate the citation origin to publications of an outlier. In that case it makes more sense to investigate the origin of citations of publications from the same venue the publications of the outlier are published in. The co-publications of an OP can be determined by Equation 10.

$$\mathit{copublicationspeers}(OP) = \{p \in P \mid v \in V \wedge \mathit{at}(p, v) \wedge p' \in \mathrm{pubs}(OP) \wedge \mathit{at}(p', v)\}. \tag{10}$$

8.2 Peer-group suitability

A peer is, according to its definition, “a person who has equal standing with another or others, as in rank, class, or age”[28]. As already mentioned, a co-author, for example, could also be a

[28] http://www.thefreedictionary.com/peers

scientist who does not resemble the outlier. It is important to find the correct members of a peer-group. Selecting the wrong members can result in the outlier also being an outlier when compared to the peer-group, or it can result in the outlier not being an outlier when compared to the peer-group. For example, a peer-group of only the students the outlier supervises is not a suitable peer-group.

Therefore, to compare outliers to their peers, the peers need to resemble the outlier. Any peer that does not resemble the outlier should be rejected. The suitability of a peer-group must therefore be taken into account. To indicate the suitability of the peer-group, we propose to use the following indicators:

- Peer group size: if the group is too small outliers cannot be found.
- Overlap in venues the peer and outlier published in: if the overlap of venues is too small, data of the peer might not be comparable to the suspect.
- Publication and citation dates of a peer: a peer might have reached his or hers performance peak a long time ago, data might therefore not be comparable to the suspect's data.

8.2.1 Peer-group suitability indicators

Peer group size Finding true outliers can only be done if the size of the data set of the peers is large enough. With only two peers, for example, it is impossible to conclude with any amount of certainty the outlier is also an outlier when compared to its peers. The size of the peer group therefore need to be taken into account. There is, however, no definition of what a small data set is. During this research we will define small as less than fifteen. If a peer group size is less than fifteen, the peer-group should be rejected, or other peers need to be added. The peer group size is calculated by equation 11.

$$\mathit{peergroupsize}(\mathit{Peers}) = |\mathit{Peers}|. \tag{11}$$

Overlap in venues the peer and outlier published in An outlier might have published in ten different venues. Comparing this outlier with a peer that has published in ten completely different venues may lead to an inaccurate comparison. For example, if we want to know the venue where citations originate from, the venue the publication was published in might be of great importance.

To indicate the similarity between an outlier and a peer, we propose to calculate and compare the overlap in venues the outlier and peer published in. As peers and outliers might occasionally publish in completely different venues, we propose to only take the venues into account that have published more than one publication. For an *outlier* and a *peer*, the overlap in venues can be calculated by Equation 13. Equation 12 determines the set of venues where an *OP* published more than once.

$$\mathit{publishedin}(OP) = \{v \in V \mid p \in \mathrm{pubs}(OP) \wedge \mathrm{at}(p, v) \wedge p' \in \mathrm{pubs}(OP) \wedge p \neq p' \wedge \mathrm{at}(p', v)\}. \tag{12}$$

$$\mathit{overlapvenues}(\mathit{outlier}, \mathit{peer}) = \frac{|\mathit{publishedin}(\mathit{outlier}) \cap \mathit{publishedin}(\mathit{peer})|}{|\mathit{publishedin}(\mathit{outlier})|}. \tag{13}$$

Example 8.1. An outlier published more than once in venues A, B, C, D and E. A peer published more than once in venues C, D, E and F. The peer published in three out of five venues where the outlier also published. The overlap of the peer is therefore $\frac{3}{5} = 0.6$.

Publication and citation dates As outliers might have recently written many publications, a peer might have written most of their publications a long time ago. Again, the same also holds for the amount of citations, an outlier might have received many citations recently, while a peer might have received most of their citations a long time ago. As some of the venues might not have existed a long time ago, data from peers can sometimes contain different characteristics than data of the outlier. Therefore, the publication and citation dates should also be used to indicate the suitability of a peer.

To indicate the suitability of the publication and citation dates between an outlier and a peer, the overlap in the amount of publications or citations per year can be compared. Naturally, there might be a difference in the amount of citations and publications between an outlier and a peer in a certain year. However, if the difference is relatively small, the peer should be indicated as a valid peer and not be rejected. If, for example, an outlier obtained ten publications and a peer obtained nine publications, there is not much of a difference. The same can be said about an outlier who obtained 200 citations, and a peer who obtained 210 citations. If an outlier obtained twice as much citations or publications in a certain year, compared to a peer, it might still be a valid peer. However, if an outlier obtained three times as much, we think the difference becomes too substantial. Therefore, to compensate for the amount of absolute publications and citations, we propose to take the log of base three of the amount of citations and publications in a certain year. The overlap between the amount of publications for an outlier *outlier* and a peer *peer* in a certain time frame can be calculated by Equation 14, and the overlap between the amount of citations in a certain time frame by Equation 15.

$$\begin{gathered} overlapPubs(startyear, endyear, outlier, peer) = \\ \frac{1}{(endyear - startyear) + 1} * \sum_{i=startyear}^{endyear} \\ \begin{cases} 1 - \frac{|(\log_3 |pubsinyear(outlier,i)| - \log_3 |pubsinyear(peer,i)|)|}{\log_3 |pubsinyear(outlier,i)|} & \text{if} |pubsinyear(outlier, i)| > 0 \\ 0 & \text{if} |pubsinyear(outlier, i)| = 0 \end{cases} . \end{gathered} \tag{14}$$

$$\begin{gathered} overlapCites(startyear, endyear, outlier, peer) = \\ \frac{1}{(endyear - startyear) + 1} * \sum_{i=startyear}^{endyear} \\ \begin{cases} 1 - \frac{|(\log_3 |citesinyear(outlier,i)| - \log_3 |citesinyear(peer,i)|)|}{\log_3 |citesinyear(outlier,i)|} & \text{if} |citesinyear(outlier, i)| > 0 \\ 0 & \text{if} |citesinyear(outlier, i)| = 0 \end{cases} . \end{gathered} \tag{15}$$

Example 8.2. An outlier obtained nine citations in 2007, 27 in 2008 and 81 in 2009. A peer obtained nine citations in 2006, 27 in 2007, 27 in 2008 and 81 in 2009. When investigating the period 2006-2009, this will result in: $\frac{1}{4} * (0 + (1 - |\frac{2-3}{2}|) + (1 - |\frac{3-3}{3}|) + (1 - |\frac{4-4}{4}|)) = \frac{1}{4} * (0 + 0.5 + 1 + 1) = 0.625$

8.2.2 Using peer-group suitability indicators

Thresholds for some of the peer-group suitability indicators should be chosen to reject certain peers. For example, when determining the overlap in venues of all peers, a threshold could be set at 0.5. A peer who scores a value lower than 0.5 should be rejected. After some of the peers are rejected, the peer-group size need to be calculated again. If the peer-group size is large enough, the outlier can be compared to the peer-group.

9 Comparing outliers to their scientific peers

After outliers are detected and their scientific peers are found, the next step is to compare the outlier to their peers. There are many ways outliers can be compared to their peers. In this research, we propose eight measures that are specifically designed to indicate scientists who might be defrauding.

9.1 Measures

We propose to use the following measures to indicate scientists who might be defrauding:

- Fraction of amount of citations originating from the most citing venue
- Fraction of amount of citations originating from the most citing scientist
- Amount of publications in most publishing venue
- Amount of early citation dates
- Maximum of fraction of early citation dates
- Max derivative of citation count
- Max derivative of publication count
- Max fraction of derivative of citation versus derivative of publication count

The following sections elaborates on each measure by giving a description of why it is used as a measure for indicating possible fraud. As fraudulent behavior cannot be distinguished from outstanding scientists behavior, all of the measures might indicate outstanding scientists and fraudulent scientists. Therefore, an intuitive benign and malicious example are also given. Finally, an example on how to calculate the measure is given.

9.1.1 Fraction of amount of citations originating from the most citing venue

Citations to a publication all originate from another publication published in the same or a different venue. An abnormal high amount of citations originating from a publication in a specific venue might indicate an outlier committing fraud. For a scientist OP, Equation 17 determines the measure how high the largest fraction of citations originating from one venue is, and the amount of citations is given by Equation 16.

$$\mathit{totcits}(OP) = \big|\{p \in P \mid p' \in \text{pubs}(OP) \wedge \text{cites}(p, p')\}\big|. \tag{16}$$

$$\mathit{fracmostcitsfromvenue}(OP) = \max_{v} \left(\frac{|\{p \in P \mid \text{at}(p, v) \wedge p' \in \text{pubs}(OP) \wedge \text{cites}(p, p')\}|}{\mathit{totcits}(OP)} \right). \tag{17}$$

Intuitive benign example An outlier might be highly specialized in a certain field. This field is relatively small and only a few venues publish publications on this topic. Therefore, the outlier might receive a relative high amount of citations originating from a specific venue.

Intuitive malicious example An outlier might be a member of an editorial board or a program committee of a certain venue. The outlier could use his/hers influence to receive more citations. For example, the outlier could only accept publications if it includes citations to his/hers own work. In this case there might be an abnormal high citation count originating from the venue the outlier edits for.

Example 9.1. An *OP* published five publications with the following amount of citations and citation origins:

- Publication p_1 with three citations originating from venue v_1 and eight citations originating from venue v_2
- Publication p_2 with four citations originating from venue v_1 and four citations originating from venue v_3
- Publication p_3 with five citations originating from venue v_1 and two citations originating from venue v_2
- Publication p_4 with eight citations originating from venue v_4
- Publication p_5 with four citations originating from venue v_2 en three citations originating from venue v_4

In this example, the *OP* received twelve citations from venue v_1, fourteen from venue v_2, four from venue v_3 and eleven from venue v_4. The result of the equation using the values in this example leads to $\frac{14}{14+12+11+4} = 0.3414$

9.1.2 Fraction of amount of citations originating from the most citing scientist

Scientists cite other related work in their publications. A scientist might cite multiple publications written by a specific author. An abnormal high amount of citations to work of a certain author might indicate an outlier committing fraud. For a scientist *OP*, Equation 18 determines the measure the largest fraction of citations originating from one author is.

$$\mathit{fracmostcitsfromscientist}(OP) = \max_{a} \left(\frac{|\{p \in P \mid p' \in \mathrm{pubs}(OP) \wedge \mathrm{cites}(p, p') \wedge \mathrm{authored}(a, p)\}|}{\mathit{totcits}(OP)} \right). \tag{18}$$

Intuitive benign example Scientists might work on the same topic. These scientists are likely to cite each other's work and therefore an abnormal amount of citations might be originating from work of a certain author.

Intuitive malicious example Scientists can make arrangements between each other to cite each other's work. If such an arrangement has been made, there might be an abnormal amount of citations originating from work of a certain author to an outlier.

Example 9.2. An *OP* wrote five publications that are cited by the following authors:

- Publication p_1 cited by a publication of author a_1 and a publication of author a_2

- Publication p_2 cited by a publication of author a_1
- Publication p_3 cited by a publication of author a_2 and author a_3
- Publication p_4 cited by a publication of author a_1 and author a_4
- Publication p_5 cited by a publication of author a_1 and a publication of author a_2 and a_4

In this example, the *OP* received four citations from author a_1, three citations from author a_2, one citation from author a_3 and two citations from author a_4. The result of the equation using the values in this example leads to $\frac{4}{10} = 0.4$

9.1.3 Amount of publications in most publishing venue

All publications are published in a certain venue. An abnormal high amount of publications published in a specific venue might indicate an outlier committing fraud. To indicate how popular a specific venue is, when compared to the other venues, we propose to calculate for an *OP* the largest fraction of publications published in one venue by Equation 20. The venue popularity of a single venue is given by Equation 19, where the fraction of publications published in a venue of an *OP* is calculated.

$$\mathit{venuepopularity}(OP) = \{(v,n) \mid v \in V \wedge n = \frac{|\{p \in P \mid \mathrm{at}(p,v) \wedge \mathrm{authored}(OP,p)\}|}{|\mathrm{pubs}(OP)|}\}. \tag{19}$$

$$\mathit{fracpubsatmostpopularvenue}(OP) = \max\left\{n \mid (v,n) \in \mathit{venuepopularity}(OP)\right\}. \tag{20}$$

Intuitive benign example An outlier might be highly specialized in a certain field. This field is relatively small and only a few journals publish publications on this topic. Therefore, the outlier might publish a relative high amount of publications in a specific venue.

Intuitive malicious example An outlier might be a member of an editorial board or a program committee of a certain venue. The outlier could use his/hers influence to publish more publications. For example: the outlier could only accept publications if he/she is mentioned as an author in the publication. In this case there might be an abnormal high number of publications published in the venue the outlier edits for.

Example 9.3. An *OP* published six publications in the following venues:

- Publication p_1 published in venue v_1
- Publication p_2 published in venue v_2
- Publication p_3 published in venue v_1
- Publication p_4 published in venue v_1
- Publication p_5 published in venue v_3
- Publication p_6 published in venue v_2

In this example, the *OP* published three publications in venue v_1, two in venue v_2 and one in venue v_3. The VenuePopularity of the *OP* is $\{(v_1, 0.5), (v_2, 0.33), (v_3, 0.17)\}$. The most popular venue is therefore venue v_1 with a value of 0.5.

9.1.4 Amount of early citation dates

Scientists need time to write good publications. After a publication is published, other scientists also need time to read that publication before they can use it in their own research. If a publication is already cited before the year of publication, this might indicate an outlier committing fraud. For an OP, the fraction of early citations versus the total amount of citations is given by Equation 21.

$$\mathit{earlycitesdates}(OP) = \frac{|\{p \in P \mid p' \in \mathrm{pubs}(OP) \wedge \mathrm{cites}(p, p') \wedge p'.\mathit{year} < p.\mathit{year}\}|}{\mathit{totcits}(OP)}. \tag{21}$$

Intuitive benign example Scientists might publish their not yet accepted publications as pre-prints on websites such as Arxiv. This way, scientists let people know they are working on a publication that might soon be published. Other scientists can already read these publications and use them in their own research. When the publication finally gets published, scientists may already have read the publication and may already have cited the publication before it was published.

Intuitive malicious example Scientists might use their position in an editorial board to advertise publications not yet published. These publications therefore might already be referenced in the year(s) preceding the official publication.

Example 9.4. An OP published three publications and received citations in the following years:

- Publication p_1 published in 2015, with four citations in 2014, ten in 2015 and twenty in 2016
- Publication p_2 published in 2016, with two citations in 2016 and 50 in 2017
- Publication p_3 published in 2016, with ten citations in 2017

In this example, the OP received a total of sixteen citations of which the year preceded the official publication year. A total amount of 96 citations was received. This measure therefore evaluates to $\frac{16}{96} = 0.16667$

9.1.5 Maximum of fraction of early citation dates

A derivative of the measure *earlycitesdates*, is the measure *fracearlycites*. Here, we do no investigate the number of early citations with respect to the total amount of citations, but we investigate the number of early citations with respect to the total amount of citations received in three years after publication. The maximum of this fraction of early citations for any paper published by OP is then used. For an OP, the maximum of the fraction of early citation dates is calculated by Equation 25. Equation 24 calculates the fraction of early citations versus the citations in the first three years since publication, where Equation 23 determines the citations to a publication p before a given year. Equation 22 determines all of the publications that cite to a publication of an OP.

$$\mathit{citingpubs}(OP) = \{p \in P \mid p' \in \mathrm{pubs}(OP) \wedge \mathrm{cites}(p, p')\}. \tag{22}$$

$$citingpubsbeforeyear(p, year_i) = \{p' \in citingpubs(p) \mid p'.year < p.year_i\}. \tag{23}$$

$$fracearlycites(p) = \begin{cases} \frac{|\{citingpubsbeforeyear(p,p.year)\}|}{|\{citingpubsbeforeyear(p,p.year+3)\}|} & \text{if}|\{citingpubsbeforeyear(p, p.year + 3)\}| > 0 \\ 0 & \text{if}|\{citingpubsbeforeyear(p, p.year + 3)\}| = 0 \end{cases}. \tag{24}$$

$$maxearlycites(OP) = \max_{p \in \mathrm{pubs}(OP)} fracearlycites(p). \tag{25}$$

Intuitive benign example The same examples given in Section 9.1.4 also apply to this measure.

Intuitive malicious example The same examples given in Section 9.1.4 also apply to this measure.

Example 9.5. An *OP* published three publications and received citations in the following years:

- Publication p_1 published in 2015, with four citations in 2014, ten in 2015, twenty in 2016 and 35 in 2017
- Publication p_2 published in 2016, with two citations in 2016 and 50 in 2017
- Publication p_3 published in 2016, with eight citations in 2015, four in 2016 and ten citations in 2017

fracearlycites evaluates to $\frac{4}{69}$ for p_1, $\frac{0}{52}$ for p_2 and $\frac{8}{22}$ for p_3. *maxearlycites* therefore evaluates to $\frac{8}{22}$ for the *OP*.

9.1.6 Max derivative of citation count

An outlier already became an outlier by looking at their citation data in the outlier detection phase of the methodology. However, this data can also be compared to the citation data of the peer-group only. The same measure can be calculated the same way as in Section 6.2.2.

Intuitive benign example Scientists can make a great discovery and publish this discovery. This publication might be cited often in the years after the publication is published.

Intuitive malicious example An outlier might be a member of an editorial board or a program committee of a certain venue. The outlier could use his/hers influence to receive more citations in the same way as described in Section 9.1.1.

Example 9.6. An *OP* received the following citations in the following years:

- 20 citations received in year 2001
- 40 citations received in year 2002
- 80 citations received in year 2003

- 70 citations received in year 2004
- 65 citations received in year 2005
- 90 citations received in year 2006

The derivative of the citations per year is given by the set {20, 40, -10, -5, 25}. Adding two consecutive values results in the following set: { 60, 30, -15, 20 }. The maximum is 60, which is the outcome of this measure.

9.1.7 Max derivative of publication count

An outlier already became an outlier by looking at their publication data in the outlier detection phase of the methodology. However, this data can also be compared to the publication data of peers. Therefore the same measure can be calculated again as in Section 6.2.1.

Intuitive benign example Scientists may become a supervisor of PhD students. These PhD students write publications with the name of the supervisor as co-author. The supervisor might publish a lot of extra publications in the years after he or she became a supervisor, compared to the years before.

Intuitive malicious example An outlier might be a member of an editorial board or a program committee of a certain venue. The outlier could use his/hers influence to publish more publications in the same way as described in Section 9.1.3.

Example 9.7. An *OP* published the following publications in the following years:

- 2 Publications published in year 2001
- 5 Publications published in year 2002
- 20 Publications published in year 2003
- 15 Publications published in year 2004
- 3 Publications published in year 2005
- 8 Publications published in year 2006

The derivative of the publications per year is given by the set {3, 15, -5, -12, 5}. Adding two consecutive values results in the following set: { 18, 10, -17, -7 }. The maximum is 18, which is the outcome of this measure.

9.1.8 Max fraction of derivative of citation versus derivative of publication count

An outlier already became an outlier by looking at their publication and citation data combined in the outlier detection phase of the methodology. However, this data can also be compared to the combination of publication and citation data of peers. Therefore, the same measure can be calculated again as in Section 6.2.3.

Intuitive benign example As already mentioned, a caveat of this measure is the possibility that scientists suddenly stop producing publications by, for example, retiring or passing away. If the scientist keeps receiving citations without producing publications, the resulting quotient can become very large.

Intuitive malicious example The same examples given in Section 9.1.6 and Section 9.1.7 also apply to this measure.

Example 9.8. An *OP* published the following publications, and received the following amount of citations in the following year:

2 Publications published obtained in year 2000 and 20 citations received in year 2001

5 Publications published obtained in year 2001 and 40 citations received in year 2002

20 Publications published obtained in year 2002 and 80 citations received in year 2003

15 Publications published obtained in year 2003 and 70 citations received in year 2004

3 Publications published obtained in year 2004 and 65 citations received in year 2005

8 Publications published obtained in year 2005 and 90 citations received in year 2006

The derivative of the citations per year is given by the set {20, 40, -10, -5, 25}. The derivative of the publications per year is given by the set {3, 15, -5, -12, 5}. Calculating the quotient of these values results in the set $\{\frac{20}{3}, \frac{40}{15}, \frac{-10}{-5}, \frac{-5}{-12}, \frac{25}{5}\}$. Adding two consecutive values result in a different set, of which the maximum is $\frac{20}{3} + \frac{40}{15} = \frac{28}{3} = 9.33$. This is the outcome of this measure.

9.2 Comparing

To investigate if the outlier is an outlier on these measures when compared to their peers, we propose to use the same outlier detection method as the method used to determine the set of outliers in the outlier detection phase of the methodology (Section 7). If the data is not skewed, a normal boxplot will be used, and when the data is skewed the boxplot will be adjusted accordingly. If the measure of the scientist is outside the maximum whisker of the boxplot, the scientist is also an outlier when compared to their peers on that specific measure.

10 Implementation

The methodology has been implemented by designing a framework and two API's. All of the software has been developed in the Python language. Three main publication data processors were used during the implementation: DBLP, Google Scholar and Semantic Scholar. DBLP was chosen as it contains major computer science journals and proceedings, and we could download the complete database. Google Scholar was used as this is one of the most known source for bibliometric data, and the data could be relatively easy extracted. Furthermore, Google Scholar contains every data necessary to calculate each measure. Semantic Scholar was eventually also used during the implementation, as Google Scholar proved to be too slow to extract data. Data could be acquired a lot faster using Semantic Scholar.

The manual of the software can be found in Appendix A.

10.1 Design

A framework has been designed implementing the methodology used for the detection of outliers and comparison of peers. The framework can be seen in Figure 2. Some design decisions had to be made during the implementation of the framework. The most import decisions were related to acquiring the necessary data for the API's, and handling different publication data processors.

10.1.1 API's

Creating and testing new measures is simplified by designing and creating two API's. The first API (`OutlierDetectAlgorithms`)can be used to calculate the different measures of the outlier detection phase. This API can be easily adjusted by adapting or adding other measures. It depends on a certain data structure. Every data acquired from a publication data processor need to be mapped to this data structure before functions of this API can be used.

The other API (`InducedPubViewRelations`) contains the functions of the induced publication view displayed in Figure 1. However, this API is dependent on data of the induced publication view. A prohibitive amount of data is therefore necessary before this API can be used in the way intended, especially the cites relation needs a huge amount of data. This API can be used in the outlier detection phase and the peer comparison phase. Again, data acquired from a publication data processor need to be mapped to a data structure that can be used by the API.

During implementation, DBLP has been successfully converted to the data structure necessary for the API `InducedPubViewRelations` API. Therefore, the outlier detection phase has been implemented in two ways for DBLP, once making use of the API `OutlierDetectAlgorithms`, and once making use of the API `InducedPubViewRelations`. However, as DBLP does not contain citation data, only the relations at and authored can be used in a meaningful way.

An attempt has also been made to convert the Google Scholar data to the data structure needed for the API `InducedPubViewRelations`. However, as already mentioned, the amount of data necessary is prohibitive. Therefore, as more detailed information were acquired of some of the outliers and their peers during the peer comparison phase, only a single measure has been implemented using this API. The partial acquired Google Scholar data has been converted to the data structure needed by the API, which was successfully used to determine the measure.

10.1.2 Data structures

The amount and kind of data necessary during the peer comparison phase is completely different compared to the data necessary during the outlier detection phase. Algorithm 1 and Algorithm 2 show the pseudo code, demonstrating how to acquire the data for both the phases. In the outlier detection phase, only the amount of publications and citations per author per year are needed. In the peer comparison phase, that more (detailed) data is needed, the amount of data explodes when compared to the outlier detection phase.

Algorithm 1 Extraction of data necessary for the outlier detection phase

```
for all $authora \in A$ do
    for all $y \in a.publications.years$ do
        Count[a][y].PubCount = count(pubs(a))
    end for
    for all $y \in a.citations.years$ do
        Count[a][y].CitCount = count(cites(a))
    end for
end for
```

Algorithm 2 Extraction of data necessary for the peer comparison phase

```
for all $authora \in A$ do
    PeerGroup = findPeers(a)
    for all $authorb \in PeerGroup \cup a$ do
        Publications = findPublications(b)
        for all $Publicationp \in Publications$ do
            DetailedInfo[b][p].PubMetaInfo = p.MetaInfo
            DetailedInfo[b][p].CitedBy= p.CitedBy
            for all $PublicationcinMetaInfo[b][p].CitedBy$ do
                DetailedInfo[b][p][c].CitMetaInfo = c.MetaInfo
            end for
        end for
    end for
end for
```

Therefore, the decision was made to design a different data structure for the data of the outlier detection phase and the peer comparison phase. The specifics of the data structures can be found in Appendix A.

10.1.3 Handling different publication data processors

Different publication data processors can be used or linked in the two different phases. For example, during the outlier detection phase, outliers can be found using publication and citation data of Google Scholar and Semantic Scholar. During the peer comparison phase, peers can be found using LNCS as the peer database and data for the measures that need publication and venue meta-data can be calculated using data from WoS. However, attention must be given to linking data from different publication data processors, as this might prove to be difficult. For example, when searching the name 'A. De Vries', using DBLP, resulted in eighteen matches. When searching the same name using Semantic Scholar, ten matches are returned, of which some are the same as the ones found with DBLP. Therefore, to successfully link accounts between the different publication data processors a reliable mechanism must be chosen that is able to link data from the different publication data processors.

Combining different data types of different publication data processors is, however, not recommended. For example, DBLP does not contain citation data. Using citation data of another source (like Google Scholar) to calculate the measures could give inaccurate results, as the publication data processors do not contain the same publications or authors.

Different kind of publication data processors were used during the implementation of the methodology. An overview of the publication data processors used can be found in Figure 7. The method used for acquiring data from a publication data processor (e.g. sending database queries or downloading an XML file), and the data format itself (e.g. HTML or XML) of the publication data processor can differ between the different data sources, while calculating the measures is independent of the publication data processors. Therefore, the framework was designed in such a way that it makes a distinction between acquiring data and calculating the measures. As the calculation of the metrics expects a certain generic data structure, every publication data processor should map the data acquired to the generic data structure.

A generic data structure for all possible publication data processors was also designed that can be used with the API that contains the functions of the induced publications view in Figure 1b (`InducedPubViewRelations`). However, as it was not possible to completely fill this generic data structure using any of the publication data processors, it was only possible to calculate a limited amount of measures in each phase.

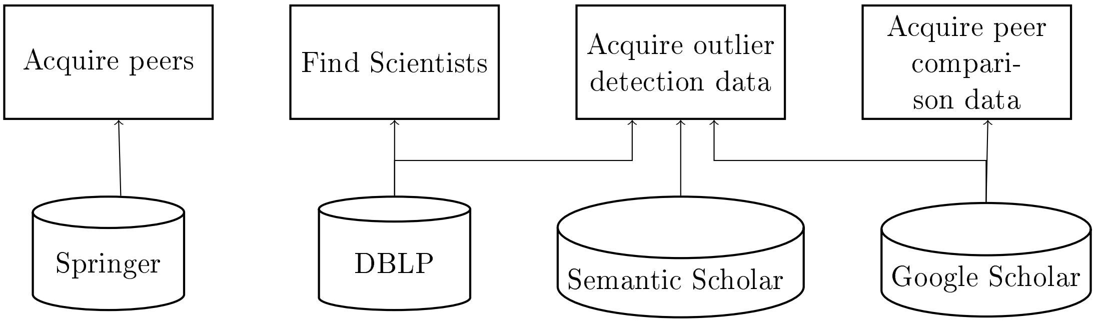


Figure 7: Publication data processors used

The measures used in the outlier detection phase can be calculated in a generic way for all publication data processors, using the generic data structure. The measures used in the peer comparison phase, however, were calculated by adapting the calculation of the measures to specific data acquired from Google Scholar. This means the peer comparison phase is not yet capable of handling other publication data processors than Google Scholar, as it uses the raw data acquired from Google Scholar directly.

10.2 Framework

This section contains the implementation details of the different steps used in the framework.

10.2.1 Finding scientists

To find the names of scientists we want to investigate for potential fraudsters, we searched the database of DBLP for authors. Every author in the database of DBLP was written to a file, in random order. The various publication data processors used in the outlier detection phase used this file as input for acquiring the data necessary in the outlier detection phase.

10.2.2 Acquiring citation and publication data for measures

The data necessary for the outlier detection phase were acquired using different publication data processors. The data processors used were DBLP for publication data, Semantic Scholar for citation data, and Google Scholar for publication and citation data. Acquiring the data from these sources was relatively easy. Google Scholar and Semantic Scholar provide the user with a limited view on the data. This limited view of both publication data processors show the amount of citations per year received, which can relatively easy be extracted. Data from DBLP could be easily downloaded and converted to the generic data structure.

Linking publication data processors As already mentioned, linking data from different sources can be difficult. However, an attempt has been made to link the names of scientists found using DBLP as input for Google Scholar and Semantic Scholar. Some names in DBLP returned multiple results when searching the same name in Google Scholar or Semantic Scholar. We chose to acquire the profiles of all the returned results. However, there were also names not found with Google Scholar or Semantic Scholar, that are in DBLP. When searching for abbreviated names, for example, some names were not found by Google Scholar. An attempt was made to still find the name, but without the abbreviation. For example, if the name Alie B. De Vries returned no result, an attempt was made to find the name without the abbreviation B.. Thus, in this case, the name Alie De Vries was searched for. This made sure we found more Google Scholar and Semantic Scholar profiles of scientists that are also in DBLP. However, this also lead to some other unwanted profiles. For example, searching for the name A. B. C. D. E. De Vries results in no profiles found using Google Scholar. Searching again for the name De Vries results in the profiles of two completely different persons. Both these profiles were then used. No further attempt has been made to link the correct names of the different publication data processors, as it was expected the amount of unwanted profiles to be small when compared to the complete data set. However, by using this method, potential outliers might be found that are active in a completely different field than IT.

10.2.3 Determine outliers

Outliers of the data were found by implementing the different measures and providing these in an API. For further details about specifics of the software and how to use the software, see Appendix A.

10.2.4 Obtain peers of outliers

Due to time constraints, only two outliers were investigated, and only two kind of peers were searched for: co-editors of editorial-boards (using Elsevier as the source of the data), and co-publications (using Google Scholar). Unfortunately, this step is not fully automated. As already mentioned, attention must be given when linking different publication data processors. In this step we tried to link Google Scholar and Elsevier to find the co-editors of editorial boards. As Elsevier uses other naming conventions than Google Scholar, searching for an outlier in the data of Elsevier proved to be difficult. Therefore, software was written that provided all of the editorial boards that might contain a part of the outlier his or hers name. For example, a Google Scholar profile with the name D. Runhart was found to be an outlier. When searching for this name in the editorial boards of Elsevier, the names Denise Runhart and Dennis Runhart occurs. As both names contains Runhart, both the editorial boards will be written to a file to be inspected manually. The user must (manually) select all the boards that apply to the outlier.

An attempt has been made, however, to acquire the profiles of the co-editors of editorial-boards automatically, using the affiliation of the scientist. The affiliation found using Elsevier was compared to the affiliation found using Google Scholar, using fuzzy string matching. If this match was above a threshold, the assumption was made the correct profile was found and this profile was used.

10.2.5 Acquiring detailed data for measures

In the peer comparison phase data was only acquired from Google Scholar. Only a partial amount of data were acquired for two outliers and their peers. This raw data was used to calculate the peer comparison measures. This data was also converted to the generic data structure needed for the `InducedPubViewRelations` API, so one measure could be calculated using this API.

10.2.6 Comparing outliers to their peers

One measure has been implemented using the `InducedPubViewRelations` API in this step. The other measures were calculated by adapting the calculations to specific data acquired of Google Scholar. After the measures are calculated for all of the peers, resulting boxplots were constructed. The measures of the outliers were then compared to the upper whisker value of the boxplot. If the outlier measure value was above the upper whisker, the outlier was also an outlier when compared to their peers.

10.3 Obstacles

Many obstacles were encountered when using Google Scholar as a publication data processor. However, this publication data processor was still used as it contained all the necessary information. For example, when searching for profiles, Google Scholar detects abnormal activity, as it uses web scraping detection techniques. After two hours of profile scraping, Google Scholar requested a captcha to be solved before continuing. Therefore we were limited to using the Google Scholar services for only two consecutive hours, before manually solving a captcha and continue the scraping. This hugely limited the amount of profiles that could be acquired during the time frame of this research. Acquiring more detailed citation data was also hugely limited. After only acquiring around 50 citations of a certain publication, we were not able to use that functionality of Google Scholar any more for at least the rest of the day. As some publications already received hundreds of citations, it was impossible in the time frame of this research to find all of the detailed citation data.

Google Scholar also limits the amount of publications returned by the profile of a scientist by 1,000 publications. Although the data is incomplete in these cases, we chose to still use these profiles. In some cases this may lead to scientists becoming an outlier while they are not an outlier when all the data is available, and in some cases it may lead to the scientists not becoming an outlier while the scientist is an outlier. Therefore, all the profiles with 1,000 publications should to be checked again if the scientists might or might not be an outlier.

A disadvantage of using Google Scholar is that scientists need to create a profile on Google Scholar. No profile is available if the scientist does not create one. As not all scientists choose to create Google Scholar profiles, it is impossible to find Google Scholar profiles of every scientist returned by the DBLP publication data processor. However, scientists willfully defrauding are likely to have a Google Scholar profile. They intend to increase on their bibliometric data and they want other scientists to notice the excellence of their research. Therefore, we assume most of the scientists we are interested in are available in Google Scholar.

Another disadvantage of Google Scholar is the way it collects the data to fill the database of Google Scholar. Google Scholar indexes publications automatically. It also tries to find the meta data, such as the publication date, using automated means. This does not always succeeds. Some publications, for example, therefore have no publication date, or an incorrect publication date.

As scraping is used by some of the publication data processors to fill their database, profiles, citations and publications might appear or disappear over time. For example, some of the publications of scientists found in the outlier detection phase with Google Scholar disappeared within a few months after acquiring the profile. When searching for the detailed information in the peer comparison phase at a later time, the detailed information sometimes did not exist anymore.

Calculating a medcouple value of thousands of values also proved to be difficult. Calculating the medcouple consumes a lot of time and memory. That much memory was sometimes needed that not enough was available, and the medcouple had to be calculated using less values.

Checking affiliations still does not guarantee 100% the correct profiles are linked. Especially with affiliations in different languages, or with multiple affiliations, not every profile was correctly linked.

11 Experiments

Four experiments were conducted to test the methodology. In the outlier detection phase experiment, we investigated if we could find significant outliers among different data sets acquired from different publication data processors. In the outlier detection method experiment, we tested whether a simpler outlier detection method could also be used that does not take skewness into account. In the peer comparison phase experiment, we tested the methodology whether it is capable of comparing outliers to their scientific peers. The final experiment we conducted was to test if we could find outstanding scientists. As outstanding scientists are also outliers, we figured these scientists should also be found by our outlier detection phase.

The experiments were conducted using data from different publication data processors. The data acquisition differs between the publication data processors, and was therefore also conducted by different means. For DBLP, for example, the complete database of August the 2nd of 2016 was downloaded. However, for Google Scholar, profiles were used that were acquired in the period November 2016 to May 2017, and for Semantic Scholar, profiles were used that were acquired in the period April 2017 to May 2017.

11.1 Outlier detection phase experiment

The measures of the outlier detection phase were applied to data acquired from different publication data processors: Semantic Scholar, Google Scholar and DBLP. For the data acquired from Semantic Scholar, only the citation measure was applied. Publication data is available in this publication data processor, but due to time constraints we were unable to include this data. Data acquired from DBLP was only used to calculate the publication measures, as DBLP does not contain any citation data. And finally, all the measures were applied to the data acquired from Google Scholar.

In the following sections, all of the distributions of the measures are displayed per publication data processor. In each of the distributions, the distribution of scientists is shown that have a certain measure value. Also, all of the corresponding adjusted boxplots of the calculated measures are shown, where applicable. The distribution images are kept small for readability. The full size images can be found in Appendix B.

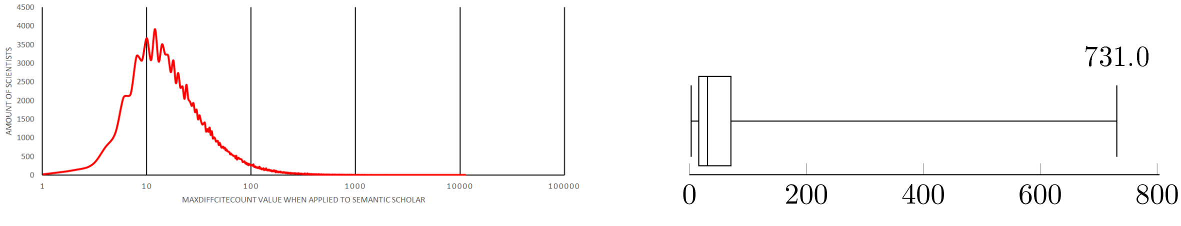
There is, however, some noise in the data. As we are trying to link data from different publication data processors, an attempt was made to use all of the names of DBLP as the resource for the names. Some data acquired using Semantic Scholar or Google Scholar therefore consists out of scientists not active in the field of IT.

11.1.1 Semantic Scholar

The citation measure *maxdiffcitecount* was applied to 139,190 profiles found with Semantic Scholar.

Measure *maxdiffcitecount* The medcouple was calculated using 40,000 profiles, due to memory requirements. The medcouple calculated of those 40,000 profiles was 0.52, indicating a strong left leaning curve. With this skewness factor, 1,267 scientists were above the upper whisker (0.91%), which had a value of 731. The distribution and the boxplot of the result of the measure is shown in Figure 8. The scale is logarithmic, which clearly shows the data is left leaning.

Conclusion 1,267 scientists were indicated as outliers, using measure *maxdiffcitecount* on the Semantic Scholar publication data processor.



(a) (b)

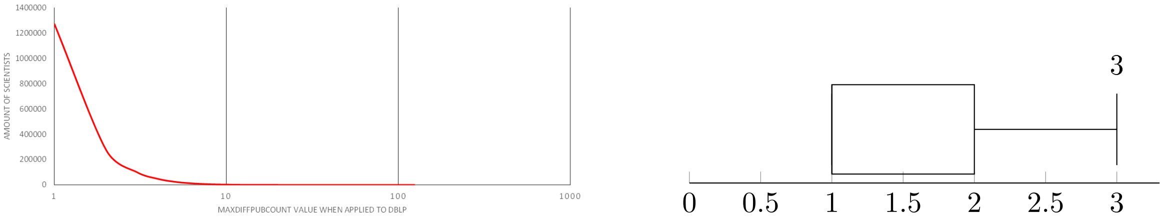
Figure 8: Distribution (a) and resulting boxplot (b) of measure *maxdiffcitecount* applied to Semantic Scholar.

11.1.2 DBLP

The publication measures *maxdiffpubcount* and *toomanypubs* were applied to 1,760,321 profiles found with DBLP.

Measure *maxdiffpubcount* The medcouple was calculated using 30,000 profiles, due to memory requirements. The medcouple calculated of those 30,000 profiles was 0.0, indicating no skewness. With this skewness factor, 110,738 scientists were above the upper whisker (6,29%). The distribution and the boxplot of the result of the measure is displayed in Figure 9. It turns out the majority of the scientists averages one or two publication in two years, resulting in a value of one or two. The upper whisker therefore has a low value. Every scientist, who on average over two years, published more then three publications is an outlier.

Conclusion 110,738 scientists were indicated as outliers, using measure *maxdiffpubcount* on the DBLP publication data processor.



(a) (b)

Figure 9: Distribution (a) and resulting boxplot of measure *maxdiffpubcount* applied to DBLP.

Measure *toomanypubs* Publication measure *toomanypubs* was also applied to the 1,760,321 profiles of DBLP. 1823 scientists were found (0.10%) to have published more than 24 publications in a certain year. Figure 10 displays the distribution of maximum amount of publications published per year.

Conclusion 1,823 scientists were indicated as having produced over 24 publications in one year, using measure *toomanypubs* on the DBLP publication data processor.

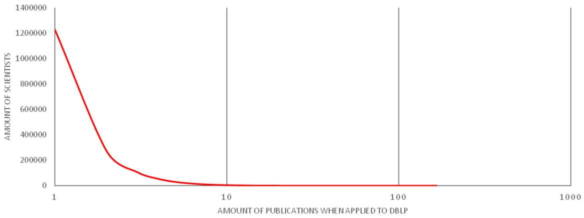


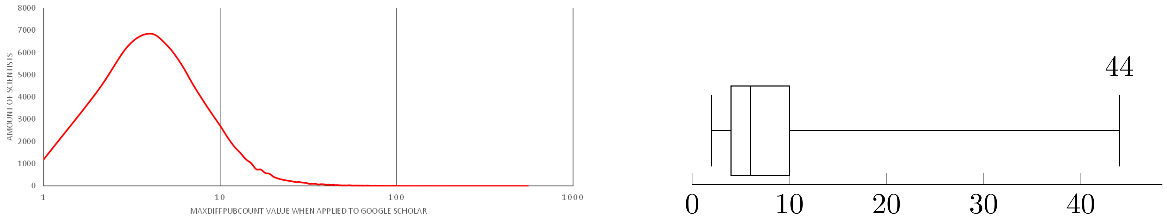
Figure 10: Distribution of maximum amount of publications published per year applied to DBLP data

11.1.3 Google Scholar

All of the measures in Sections 6.2.1, 6.2.2 and 6.2.3 were applied to 59,530 profiles acquired with Google Scholar. The medcouples were calculated using 30,000 profiles, due to memory requirements.

Measure *maxdiffpubcount* Applying the data to measure *maxdiffpubcount* resulted in 776 (1.3%) outliers, with a value of over 44. The medcouple calculated was 0.33, indicating a strong left leaning curve. The distribution and the boxplot of the result of the measure is shown in Figure 11. The scale is logarithmic, which also clearly shows the data is left leaning.

Conclusion 776 scientists were indicated as outliers, using measure *maxdiffpubcount* on the Google Scholar publication data processor.



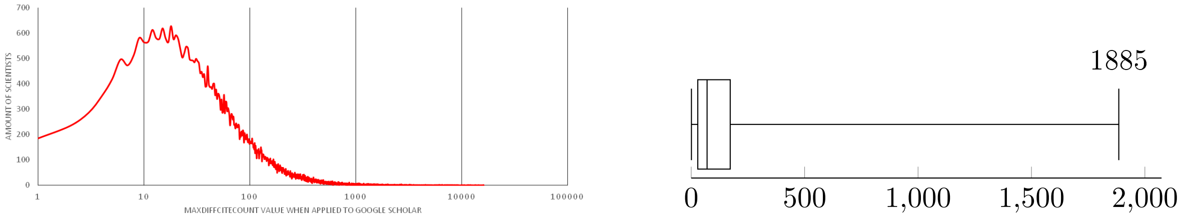
(a)

(b)

Figure 11: Distribution (a) and resulting boxplot (b) of measure *maxdiffpubcount* applied to Google Scholar.

Measure *maxdiffcitecount* Applying the data to measure *maxdiffcitecount* resulted in 726 (1.22%) outliers, with a value of over 1885. The medcouple calculated was 0.52, indicating a strong left leaning curve. The distribution and boxplot of the result of the measure is shown in Figure 12. The scale is logarithmic, which also clearly shows the data is left leaning.

Conclusion 726 scientists were indicated as outliers, using measure *maxdiffcitecount* on the Google Scholar publication data processor.

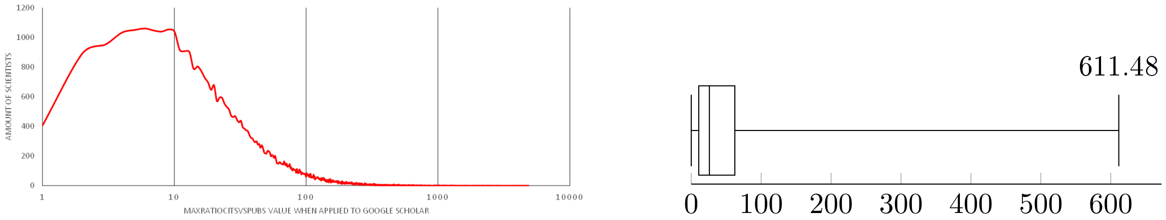


(a) (b)

Figure 12: Distribution (a) and resulting boxplot (b) of measure *maxdiffcitecount* applied to Google Scholar.

Measure *maxratiocitsvspubs* Applying the data to measure *maxratiocitsvspubs* resulted in 402 (0.67%) outliers, with a value of over 611.48. The medcouple calculated of the results of the measure was 0.49, indicating a strong left leaning curve. The distribution and boxplot of the result of the measure is shown in Figure 13. The scale is logarithmic, which also clearly shows the data is left leaning.

Conclusion 402 scientists were indicated as outliers, using measure *maxratiocitsvspubs* on the Google Scholar publication data processor.



(a) (b)

Figure 13: Distribution (a) and resulting boxplot (b) of measure *maxratiocitsvspubs* applied to Google Scholar.

Measure *toomanypubs* Applying the data to measure *toomanypubs* resulted in 8,076 scientists(13.57%). The distribution of this measure is shown in Figure 14.

Conclusion 8,076 scientists were indicated as having produced over 24 publications in one year, using measure *toomanypubs* on the Google Scholar publication data processor.

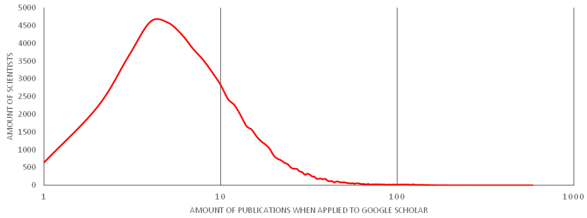


Figure 14: Distribution of maximum amount of publications published per year applied to Google Scholar data

Combining outcome of the measures Combining all Google Scholar results, according to Equation 7, resulted in a set of 8,318 individual scientists to be investigated further in the peer comparison phase.

11.1.4 Discussion

With all the publication data processors used, outliers were to some extent successfully indicated. However, some of the measures in combination with the used publication data processor and outlier detection mechanism, resulted in a huge amount of outliers. This indicates some of the outliers do, in fact, not differ significantly from the other values. Therefore, the combination of outlier detection mechanism, the publication data processor used and measures used may not be suitable to indicate significant outliers.

Within a normal distribution, results differ significant if they are outside the two-sigma interval[29]. This means a significant different result is among the 5% of the most extreme values, (2.5% of high extreme values and 2.5% of low extreme values). To indicate whether a measure is suitable to be used with the publication data processors, we will also verify if no more than 5% is indicated as an outlier, therefore indicating all the outliers are significant different than the rest. As we are only interested in the high extreme values, we therefore need to verify no more then 2.5% of the sample size is indicated as an outlier for each combination of measure and publication data processor.

DBLP With all the venues and scientists that DBLP collects, most of the scientists do not produce more than one or two publications on average in two years. This has a huge impact on detecting outliers for the measure *maxdiffpubcount*. The medcouple value resulted into a value of zero. As a consequence, all of the scientists producing three or more publications on average of two years are outliers. 6.29% of all the scientists were indicated as being an outlier. As this is almost three times more then our threshold of 2.5%, the measure is not suitable to be used with this publication data processors. However, as the medcouple was not calculated using all the values, DBLP might still be suitable to indicate outliers using measure *maxdiffpubcount*. If a medcouple larger than zero would have been used, less outliers would have been found. As DBLP does not contain citation information, this automatically means this publication data processor is not able to indicate outliers using the measures using citation data. The other measure that uses only publication data (*toomanypubs*), resulted in 0.10% of the sample data

[29] `http://www.graphpad.com/www/data-analysis-resource-center/blog/statistical-significance-defined-using-the-five-sigma-standard/`

indicated as outlier. However, as we are not looking for outliers with this measure and therefore do not use an outlier detection algorithm, verifying if no more then 2.5% of the sample size is returned does not apply to this measure. DBLP is therefore suitable to be used in combination with measure *toomanypubs*.

Semantic Scholar Semantic Scholar is able to find significant outliers among the citation data, as 0.93% of the sample data was indicated as outlier. Other measures could not be verified as the publication data was not available.

Google Scholar Using measure *toomanypubs* with Google Scholar data, resulted in lots of scientist who produced more than 24 publications in a single year. 13.57% of the sample data was indicated as having more than 24 publications produced per year. However, we are not looking for outliers with this measure. Therefore we did not use an outlier detection algorithm, and verifying if no more then 2.5% of the sample size is returned does not apply to this measure. As all of the other measures returned less than 2.5% of the sample data, Google Scholar is suitable of indicating significant outliers using all of the other measures.

However, 13.57% is still a high number. As Google Scholar indexes every publication found on the internet by automated means, it might be some publications are indexed by mistake, resulting in more publications per year for some scientists. Still, the purpose of the outcome of the outlier detection phase is to find scientists which should be investigated in the peer comparison phase. Therefore, this measure is suitable to be used with Google Scholar.

Summary Table 1 summarizes whether a publication data processors is suitable to be used to indicate outliers, using a certain measure. If a measure could not be calculated using the publication data processor, it is indicated in the Table by n/a. Our method is thus capable of finding significant outliers among bibliometric data using different publication data processors. However, our method of finding significant outliers in publicly available bibliometric data is dependent on the combination of outlier detection mechanism, the measure, and the publication data processor used.

Table 1: Publication data processors suitable to indicate outliers using a certain measure

Measure	DBLP	Semantic Scholar	Google Scholar
maxdiffpubcount	–	n/a	✓
toomanypubs	✓	n/a	✓
maxdiffcitecount	n/a	✓	✓
maxratiocitsvspubs	n/a	n/a	✓

11.2 Outlier detection method experiment

The outlier detection method has a huge impact on the amount of outliers returned. We chose the adjusted boxplot as it was capable of handling skewness of data. Other methods might, of course, provide different results. However, to verify if we really need an outlier detection method capable of handling skewness, we verified if we could have used the simpler standard boxplot method. This simpler boxplot method does not need to calculate a skewness factor, and is therefore also faster in providing results.

A quick experiment was performed by using zero as the medcouple value, thereby using the standard boxplot method, for all the measures applied to Google Scholar and Semantic Scholar data (DBLP was not used as DBLP already used a medcouple value of zero). The results showed a substantial difference in the amount of outliers found. The results of this experiment are displayed in Table 2 and Table 3.

Table 2: Amount of scientists indicated as outlier with the adjusted boxplot method and normal boxplot method, applied to Google Scholar

Measure	Adjusted boxplot	Normal boxplot
maxdiffpubcount	776 (1.3%)	4,466 (7.5%)
maxdiffcitecount	723 (1.2%)	6,431 (10.8%)
maxratiocitsvspubs	402 (0.67%)	4,147 (6.9%)

Table 3: Amount of scientists indicated as outlier with the adjusted boxplot method and normal boxplot method, applied to Semantic Scholar

Measure	Adjusted boxplot	Normal boxplot
maxdiffcitecount	1,267 (0.91%)	14,604 (10.5%)

Summary Without using a medcouple, and thereby using the standard boxplot method, none of the measures are suitable of indicating significant outliers. In combination with Google Scholar and Semantic Scholar, every measure returned more than 2.5% of the sample size. This indicates that, with the proposed measures, an outlier detection method capable of handling skewness in the data is necessary to find significant outliers, and the adjusted boxplot method is capable of doing so. Although the outlier detection method was not capable of finding significant outliers with DBLP as the publication data processor, this primarily had to do with the skewness value of zero. If the skewness value was larger than zero, less outliers would have been found.

11.3 Peer comparison phase experiment

For some of the measures used in the peer comparison phase, an enormous amount of data is required for just a single outlier to be compared to its peers. As Google Scholar was our main publication data processor for the peer comparison phase, we could not access and acquire all the information necessary to investigate all outliers in the time period of this research. Therefore, for the peer comparison phase, we only focused on two outliers, who are both outliers of the outlier detection phase, in all three publication data processors.

For some of the measures, detailed data of publications was necessary. Not all of the detailed data of publication were acquired for the outliers and their peers to calculate some of the measures. During testing of acquiring data and development of the measures, we already acquired detailed publication data of specific years for both the outliers and their peers. As the data was already available, the decision was made to use this existing data of only the specific years to calculate measures *fracpubsatmostpopularvenue* and *earlycitesdates* with. This means that for outlier O_1 and the peer-group of O_1, only publications published in the years 2014, 2015 and 2016 were investigated, and only publications published in the years 2013, 2014 and 2015 were investigated for outlier O_2 and the peer-group of O_2.

In the following sections, adjusted boxplots are shown that are calculated using the data of the peer-group of an outlier. The value on the right side of the boxplots indicates the high-whisker value. If an outlier has a measure value higher than this value, the outlier is also an outlier compared to its peer-group. The black dot in the boxplot indicates the value of the outlier.

11.3.1 Acquiring peers

This step was partially manually executed. The name of the outliers was searched for in the lists of editorial board members of Elsevier. If a part of the name of the outlier was in the name of an editorial board member, we had to check manually if this was the correct person.

The first outlier (O_1) was found to be an editor of multiple journals. We were able to find two journals where O_1 is a co-editor. Google Scholar profiles were acquired of the co-editors of those journals. Eventually, 34 Google Scholar profiles were found of the first journal, and twelve Google Scholar profiles were found of the second journal.

For the second outlier (O_2) we were able to find one journal where O_2 is a co-editor of. Again, Google Scholar profiles were acquired of all of the co-editors of this journal. This resulted in 31 Google Scholar profiles that could be investigated.

During this experiment, we did not investigate all of the peer-group suitability indicators as we could not acquire all the data necessary to investigate all the suitability indicators. The only indicator we used was *peergroupsize*. As both the peer-groups are larger than fifteen, both groups were considered to be suitable.

11.3.2 Measure *fracmostcitsfromvenue* and *fracmostcitsfromscientist*

Unfortunately, for these measures a huge amount of data was necessary, which could not be achieved in the time-frame of this research.

For the *fracmostcitsfromvenue* measure, meta-data of every citation to all publications of an outlier need to be acquired. From this meta-data the citing venue could be extracted. After the venue is known, the co-publications could be extracted from the venue. Then, every co-publication need to be investigated for their citing venue to be able to compare the publication of the outlier with. As this consumed too much time and not enough data were acquired, we can not show any results of this measure.

For the *fracmostcitsfromscientist* measure, meta-data of every citation to all publications of an outlier and all of the publications of the peer-group was necessary. As this proved to be too slow and therefore would take too much time with our data-source, we can also not show any results of this measure.

11.3.3 Outlier O_1

Table 4 displays the value of the measure calculated for the outlier and whether the outlier is also an outlier when compared to the peer group with this value. The resulting boxplots of the measures applied to the peer group can be seen in Figure 15.

Table 4: Results of measures and comparison to peer group for outlier O_1

Measure	Calculated value	Outlier to peer group
fracpubsatmostpopularvenue	0.065	–
earlycitesdates	0.03	–
fracearlycites	1.72	–
maxdiffcitecount	3,544	✓
maxdiffpubcount	46	–
maxratiocitsvspubs	315.4	–

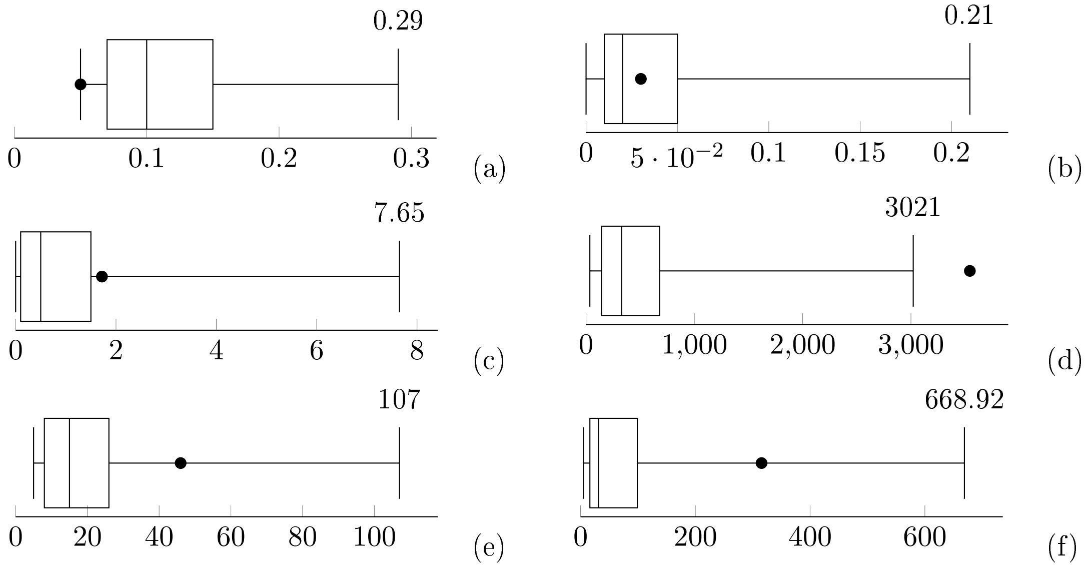


Figure 15: Boxplots of measures *fracpubsatmostpopularvenue* (a), *earlycitesdates* (b), *fracearlycites* (c), *maxdiffcitecount* (d), *maxdiffpubcount* (e) and *maxratiocitsvspubs* (f) for outlier O_1

Conclusion Outlier O_1 is only an outlier on measure *maxdiffcitecount*, when compared to the peer-group.

11.3.4 Outlier O_2

Table 5 displays the value of the measure calculated for the outlier and whether the outlier is also an outlier when compared to the peer group with this value. The resulting boxplots of the measures applied to the peer group can be seen in Figure 16.

Table 5: Results of measures and comparison to peer group for outlier O_2

Measure	Calculated value	Outlier to peer group
fracpubsatmostpopularvenue	0.1	–
earlycitesdates	0.004	–
fracearlycites	0.5	–
maxdiffcitecount	1325	✓
maxdiffpubcount	18	–
maxratiocitsvspubs	31.5	–

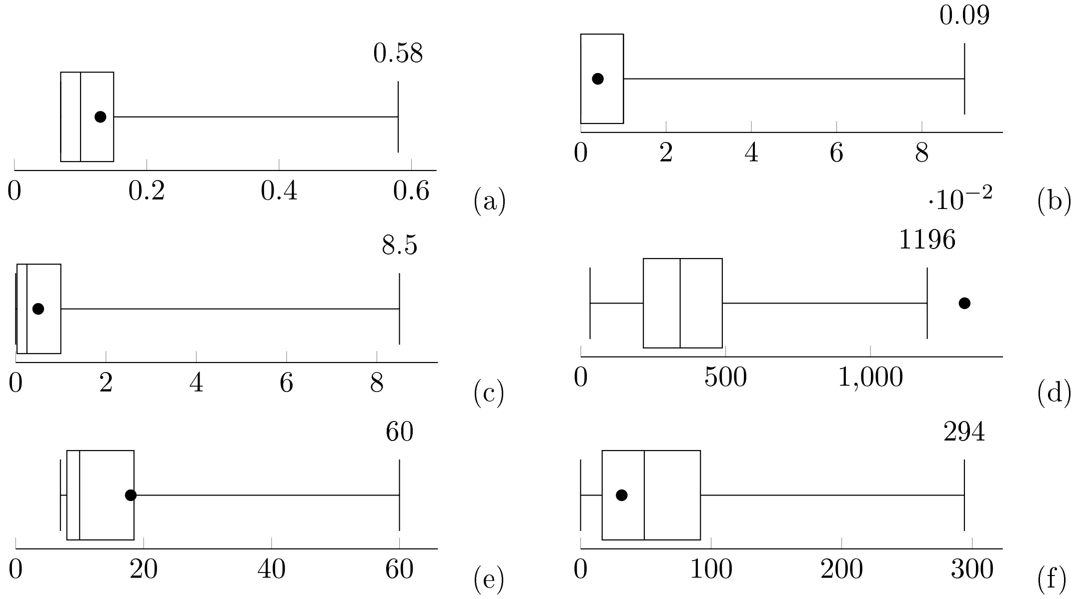


Figure 16: Boxplots of measures *fracpubsatmostpopularvenue* (a), *earlycitesdates* (b), *fracearlycites* (c), *maxdiffcitecount* (d), *maxdiffpubcount* (e) and *maxratiocitsvspubs* (f) for outlier O_2

Conclusion Outlier O_2 is only an outlier on measure *maxdiffcitecount*, when compared to the peer-group.

11.3.5 Discussion

Table 6 summarizes whether the outliers O_1 and O_2 are outliers when compared to their peers.

Table 6: Outlier is also an outlier when compared to their peer-group.

Measure	O_1	O_2
fracpubsatmostpopularvenue	–	–
earlycitesdates	–	–
fracearlycites	–	–
maxdiffcitecount	✓	✓
maxdiffpubcount	–	–
maxratiocitsvspubs	–	–
fracmostcitsfromvenue	n/a	n/a
fracmostcitsfromscientist	n/a	n/a

The more measures an outlier is an outlier on when compared to its peers, the more this might give the impression the outlier is a potential fraudster. As both outliers are only outliers when compared to their peers on only one measure, this might give the impression these outliers might not be potential fraudsters. However, as we were not capable of collecting all the data necessary to calculate all the measures, and we therefore were only able of checking two outliers partially, they might be outliers on the other measures, and therefore these outliers might be potential fraudsters. No conclusion can therefore be drawn from these results.

Summary No definitive answer can be given as to whether the method and measures used in the peer comparison phase are suitable to be used for fraud investigation. Although our method is clearly capable of comparing outliers to their scientific peers, this does not conclude it is also suitable to be used to indicate potential fraudsters. A lot depends on the measures used. One could argue therefore we did not use the correct measures. However, these measures were designed while keeping in mind how data of fraudsters is displayed. We therefore think these measures are suitable to be used as measures that help in assisting and supporting fraud investigation. To verify this, all the necessary data should be acquired and this test need to be executed again for known fraudsters. If the known fraudsters are indicated as potential fraudsters by this methodology and these measures, the method and measures used can be declared suitable to be used for fraud investigation.

11.4 Detecting outstanding scientists experiment

To validate the methodology, a set of known fraudsters should be mixed among the set of scientists already available, and verify if some or all are indicated as outliers in the outlier detection phase, and marked as outlier when compared to their peers in the peer comparison phase. As the data necessary to test this for the peer comparison phase is too large, validating this phase was not feasible. Furthermore, as there are not enough known fraudsters in our set of scientists, we cannot validate if all known fraudsters will be found.

However, it was possible to extract profiles of Google Scholar and Semantic Scholar of some outstanding scientists. With those outstanding scientists we can show scientists with remarkable publication and citation data, and thus they should become outliers in our outlier detection phase, can be found using our methodology.

For this experiment, we constructed a set of outstanding scientists by calculating the amount of publications a scientist has published in the top ranking computer security conferences[30], using

[30] `http://faculty.cse.tamu.edu/guofei/sec_conf_stat.htm`

the database of DBLP. To keep the size manageable, every scientist who published more than 40 publications in the following conferences was included in the set of outstanding scientists:

- IEEE Symposium on Security and Privacy (S&P)
- ACM Conference on Computer and Communications Security (CCS)
- International Cryptology Conference (Crypto)
- European Cryptology Conference (Eurocrypt)
- Usenix Security Symposium (Security)
- ISOC Network and Distributed System Security Symposium (NDSS)

This resulted in a set of 23 names. The next step was to acquire the Google Scholar and Semantic Scholar profiles of these scientists and verify whether these scientists are indicated as outliers by the outlier detection phase. We simply put the outstanding scientists among the set of scientists already available. Table 7 shows whether an outstanding scientist is indicated as an outlier by the outlier detection phase. If a profile was not found of a scientist, this is indicated in the Table by n/a. Fifteen scientists were indicated as outlier by the outlier detection phase.

Table 7: Outstanding IT scientists indicated as outliers in the outlier detection phase.

	maxdiffpubcount Google Scholar	*toomanypubs* Google Scholar	*maxdiffcitecount* Google Scholar	*maxratiocitsvspubs* Google Scholar	*maxdiffcitecount* Semantic Scholar
$OutstandingScientist_1$	n/a	n/a	n/a	n/a	✓
$OutstandingScientist_2$	–	✓	–	–	✓
$OutstandingScientist_3$	–	✓	✓	✓	✓
$OutstandingScientist_4$	–	✓	–	–	–
$OutstandingScientist_5$	–	–	✓	–	✓
$OutstandingScientist_6$	–	✓	–	–	–
$OutstandingScientist_7$	–	✓	–	–	✓
$OutstandingScientist_8$	–	–	–	–	✓
$OutstandingScientist_9$	–	✓	–	–	–
$OutstandingScientist_{10}$	–	–	–	–	–
$OutstandingScientist_{11}$	–	–	–	–	–
$OutstandingScientist_{12}$	–	–	–	–	–
$OutstandingScientist_{13}$	–	✓	–	–	–
$OutstandingScientist_{14}$	–	–	–	–	✓
$OutstandingScientist_{15}$	n/a	n/a	n/a	n/a	✓
$OutstandingScientist_{16}$	–	✓	–	–	–
$OutstandingScientist_{17}$	–	✓	–	–	–
$OutstandingScientist_{18}$	–	–	–	–	✓
$OutstandingScientist_{19}$	–	–	–	–	–
$OutstandingScientist_{20}$	n/a	n/a	n/a	n/a	–
$OutstandingScientist_{21}$	–	–	–	–	–
$OutstandingScientist_{22}$	–	–	–	–	n/a
$OutstandingScientist_{23}$	n/a	n/a	n/a	n/a	–

11.4.1 Discussion

Clearly, the outlier detection phase is capable of detecting some of the outstanding scientists, with fifteen out of 23 scientists (65,2%) being indicated as outliers. These scientists were only detected by applying the proposed measures in this research. To also be able to find the other scientists, other measures could be proposed, designed for finding those scientists by investigating their publication and citation data. However, care must be taken in not designing curious measures just to find these scientists.

Most of the outliers found using Google Scholar where found on the measures *toomanypubs*. As this was also the measure that returned most of the scientists to be investigated during the outlier detection phase (13.57%), this was to be expected.

There are, however, some limitations that apply to our methodology and implementation. Some of the top scientists are detected only by Google Scholar or Semantic Scholar. As both publication data processors use different data, there might be a difference in the measures calculated, and therefore a scientists may or may not be an outlier depending on the publication data processor. Another limitation is that some of the top scientists do not have a Google Scholar or Semantic Scholar profile. Without any of the profiles, a scientist cannot be investigated. A current limitation of the Semantic Scholar publication data processor means that we cannot use the number of publications per year and therefore we cannot apply the publication measures to this publication data processor.

12 Conclusion and future work

In this research we presented a systematic approach of finding potential fraudsters and outstanding researchers, by providing a framework which can be used to find the potential fraudsters and outstanding researchers among a set of scientists. This framework was developed using the two-phase methodology proposed in this research. In the outlier detection phase, outliers were indicated using certain measures and an outlier detection method. In the peer comparison phase, the outliers were compared to their peers, also using an outlier detection method and certain measures.

The following research questions were answered in this research:

RQ1: How to find scientific outliers? To find scientific outliers, we calculated measures indicating potential fraudsters and outstanding researchers of all scientists in a set. Outliers were detected using the adjusted boxplot method. However, an experiment showed not all combinations of publication data processors and measures are capable of finding significant outliers. Therefore, not all measures can be used to indicate outliers with certain publication data processors.

Another experiment showed that we were able to find outliers with the specific characteristics in their data we are interested in. We were able to indicate over 65% of the outstanding researchers as being actual outliers. There is, however, still room for improvement. Other measures could be implemented to make sure we find more or all outstanding scientists.

RQ2: How to compare the research output of scientific outliers to that of their scientific peers? To compare outliers to their peers, we proposed to calculate certain measures designed to indicate potential fraudsters of the peers and the outlier. An experiment showed we were only able to acquire a partial data set using Google Scholar as the publication data processor. With this data set we could calculate the measures, and use these for the comparison. However, as we could not acquire all the data necessary, no definitive answer can be given if the proposed method is suitable of finding potential fraudsters. Due to the data volume necessary, we were not able to calculate all of the measures properly.

The provided implementation of the framework is almost completely automated. The most difficult part to automate was to link data of different publication data processors. Therefore, finding peers of outliers is not fully automated. Manual inspection is still necessary in this step. And, depending on the publication data processor, some interaction is needed when acquiring the data necessary for the outlier detection phase. The publication data processors using web scraping detection techniques do still require manual intervention by solving captchas.

Adjusting the provided implementation can easily be achieved to suit other needs or insights. Other publication data processors, measures or outlier detection mechanism can be relatively easily implemented, especially in the outlier detection phase, thereby providing an easy way to find potential fraudsters using different means. The peer comparison phase currently only supports Google Scholar for all of the measures. However, all of the measures can be implemented by using the `InducedPubViewRelations` API, thereby making this phase also generic for other publication data processors. The downside of the API is the amount of data needed. When completely implementing the peer comparison phase using the API, a publication data processor is needed which can provide all of the data in a timely manner.

It is not possible, given the results, to conclude that we are able to find potential fraudsters using this methodology. A proper validation need to be performed before we can come to this conclusion. Before such a validation can be performed, every data necessary to calculate all the measures need to be acquired. If after the validation real fraudsters are among the indicated potential fraudsters, the conclusion can be made the methodology is successful in finding potential fraudsters. However, by implementing this framework we are now able to assist and support fraud investigation. The experiments showed promising results with the current implemented measures, as already 65% of outstanding scientists were indicated as outliers. Different measures investigating different characteristics can be easily implemented to complement the current measures. To use the full potential of our methodology, however, a huge amount of data is necessary.

Future work This research only scratched the surface of the possibilities of using a systematic approach in search for potential fraudsters. In future work, this research can be extended or other opportunities could be researched. For example: this research is highly dependable on Google Scholar in the peer comparison phase which proved to be problematic in acquiring data. Other publication data processors could therefore be implemented to overcome this issue. Other publication data processors could also be implemented to increase the certainty of the possibility an outlier is a potential fraudster. When searching for other publication data processors, we recommend to find publication data processors of which it is relative easy to acquire the data from (e.g. not limited by any amount of time or amount of queries), as our methodology and measures needs huge quantities of data. Semantic Scholar seems to be a promising publication data processor to be used for this purpose, as it currently does not use web scraping detection techniques.

To link data from different sources, a reliable way need to be found. For example, finding the same scientist using abbreviated or common names proved to be difficult when searching for only the names of these scientists in different publication data processors. Venues or publications might also have different names when using different publication data processors. Research can be conducted in finding a reliable way to link data from different sources.

The proposed suitability indicators were not implemented during this research. These could be implemented and refined in further research.

Other measures could also be considered and implemented that might use other data to investigate other properties. For example: information about who edited a publication might be of great value. The time it took for a publication from being peer-reviewed to actually being accepted by the venue, might also be of great value when developing other measures.

This research only focused on finding suspicious scientists. Other research might also be conducted at finding suspicious publications or suspicious venues. Publications and venues can all be compared to their peer publications and venues, so the same systematic approach could be used as proposed in this research.

13 Bibliography

Articles

[Rog99] Lee F Rogers. "Salami slicing, shotgunning, and the ethics of authorship." In: *AJR. American journal of roentgenology* 173.2 (1999), pp. 265–265.

[BHS04] G Brys, Mia Hubert, and Anja Struyf. "A robust measure of skewness". In: *Journal of Computational and Graphical Statistics* 13.4 (2004), pp. 996–1017.

[CK05] Christian S. Collberg and Stephen G. Kobourov. "Self-plagiarism in computer science". In: *Commun. ACM* 48.4 (2005), pp. 88–94. DOI: `10.1145/1053291.1053293`. URL: `http://doi.acm.org/10.1145/1053291.1053293`.

[BS06] James R Binkley and Suresh Singh. "An Algorithm for Anomaly-based Botnet Detection." In: *SRUTI* 6 (2006), pp. 7–7.

[Egg06] Leo Egghe. "Theory and practise of the *g*-index". In: *Scientometrics* 69.1 (2006), pp. 131–152. DOI: `10.1007/s11192-006-0144-7`. URL: `http://dx.doi.org/10.1007/s11192-006-0144-7`.

[KJ06] Clint D Kelly and Michael D Jennions. "The h index and career assessment by numbers". In: *Trends in Ecology & Evolution* 21.4 (2006), pp. 167–170.

[Smi06] Richard Smith. "Peer review: a flawed process at the heart of science and journals". In: *Journal of the royal society of medicine* 99.4 (2006), pp. 178–182.

[Law07] Peter A Lawrence. "The mismeasurement of science". In: *Current Biology* 17.15 (2007), R583–R585.

[HV08] M. Hubert and E. Vandervieren. "An adjusted boxplot for skewed distributions". In: *Computational Statistics & Data Analysis* 52.12 (2008), pp. 5186–5201. DOI: `10.1016/j.csda.2007.11.008`. URL: `http://dx.doi.org/10.1016/j.csda.2007.11.008`.

[Hir10] Jorge E. Hirsch. "An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship". In: *Scientometrics* 85.3 (2010), pp. 741–754. DOI: `10.1007/s11192-010-0193-9`. URL: `http://dx.doi.org/10.1007/s11192-010-0193-9`.

[Ole11] NP Olewuezi. "Note on the comparison of some outlier labeling techniques". In: *Journal of Mathematics and Statistics* 7.4 (2011), pp. 353–355.

[WS11] Hadley Wickham and Lisa Stryjewski. "40 years of boxplots". In: *Am. Statistician* (2011).

[LRT12] Emilio Delgado López-Cózar, Nicolas Robinson-Garcia, and Daniel Torres-Salinas. "Manipulating Google Scholar Citations and Google Scholar Metrics: simple, easy and tempting". In: *CoRR* abs/1212.0638 (2012). URL: `http://arxiv.org/abs/1212.0638`.

[WF12] Allen W. Wilhite and Eric A. Fong. "Coercive Citation in Academic Publishing". In: *Science* 335.6068 (2012), pp. 542–543. ISSN: 0036-8075. DOI: `10.1126/science.1212540`. eprint: `http://science.sciencemag.org/content/335/6068/542.full.pdf`. URL: `http://science.sciencemag.org/content/335/6068/542`.

[PF13] Raj Kumar Pan and Santo Fortunato. "Author Impact Factor: tracking the dynamics of individual scientific impact". In: *CoRR* abs/1312.2650 (2013). URL: `http://arxiv.org/abs/1312.2650`.

[FMO14] Cat Ferguson, Adam Marcus, and Ivan Oransky. "Publishing: The peer-review scam". In: *Nature* 515 (2014), pp. 480–482.

[Boh15] John Bohannon. "Hoax-detecting software spots fake papers". In: *Science* 348.6230 (2015), pp. 18–19.

[Hau15] Charlotte J Haug. "Peer-review fraud - hacking the scientific publication process". In: *New England Journal of Medicine* 373.25 (2015), pp. 2393–2395.

In Proceedings

[HL02] Sudheendra Hangal and Monica S Lam. "Tracking down software bugs using automatic anomaly detection". In: *Proceedings of the 24th international conference on Software engineering.* ACM. 2002, pp. 291–301.

[Col+03] Christian S. Collberg et al. "SPLAT: A System for Self-Plagiarism Detection". In: *Proceedings of the IADIS International Conference WWW/Internet 2003, ICWI 2003, Algarve, Portugal, November 5-8, 2003.* IADIS, 2003, pp. 508–514. ISBN: 972-98947-1-X.

[She+07] Bo Sheng et al. "Outlier detection in sensor networks". In: *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing.* ACM. 2007, pp. 219–228.

[AAS11] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and Václav Snásel. "Overview and Comparison of Plagiarism Detection Tools". In: *Proceedings of the Dateso 2011: Annual International Workshop on DAtabases, TExts, Specifications and Objects, Pisek, Czech Republic, April 20, 2011.* Ed. by Václav Snásel, Jaroslav Pokorný, and Karel Richta. Vol. 706. CEUR Workshop Proceedings. CEUR-WS.org, 2011, pp. 161–172. URL: `http://ceur-ws.org/Vol-706/poster22.pdf`.

[JM17] H. Jonker and S. Mauw. "A Much-needed Security Perspective on Publication Metrics". In: *Proc. 25th Security Protocols Workshop (SPW'17).* LNCS. To be published. Springer, 2017.

In Collection

[BHS03] Guy Brys, Mia Hubert, and Anja Struyf. "A comparison of some new measures of skewness". In: *Developments in robust statistics.* Springer, 2003, pp. 98–113.

Other

[Seo06] Songwon Seo. *A review and comparison of methods for detecting outliers in univariate data sets.* 2006.

[Har07] A.W. Harzing. *Publish or Perish.* 2007. URL: `http://www.harzing.com/pop.htm`.

[Whi] *Using Bibliometrics: A Guide to Evaluating Research Performance with Citation Data.* White Paper. Thomson Reuters, 2008.

[Ste12] Joel Stemmer. *detecting outliers in web-based network traffic*. 2012. URL: `http://essay.utwente.nl/61640/`.

A Software Manual

A.1 Software requirements

The following software was used, and might be necessary for all the scripts to work properly:

- Python V3.5.1
- Beautifulsoup V4.5.1
- Selenium V3.0
- numpy V1.12.1
- scipy V0.19.0
- lxml V3.6.1
- pandas V0.20.1
- statsmodels V0.8.0
- pybtex V0.21

A.2 Finding scientists

In folder `/DataObtainersOutlierDetectionPhase/dblp/Authors` you can find the scripts used to find all of the scientists in the DBLP database. Running the script will extract all the authors out of the `/DataObtainersOutlierDetectionPhase/dblp/database/dblp.xml` database and save all the names to `/DataObtainersOutlierDetectionPhase/dblp/Authors/authors.txt`. This is the file that is used by other publication data processors in other steps. A new XML database of DBLP can be downloaded from `http://dblp.uni-trier.de/xml/`. Make sure to put the new XML database in the correct folder (`/DataObtainersOutlierDetectionPhase/dblp/database/`), and to also include the .dtd file.

A.3 Acquiring data for the outlier detection phase

Scripts used to find outliers during the outlier detection phase use citation and publication data, saved as pickles, as input. In these pickles must be a dictionary with the name of a scientist as key. The value of the dictionary is another dictionary. This dictionary contains the year as key and the amount of citations or publications as value. For example:

{'N. Tielenburg' : {2008 : 1, 2009 : 2, 2004: 2} , 'D. Runhart' : {2009: 5, 2012: 3}}

Every publication data processor need to extract this information and export it to a different pickle for the publication data and the citation data. This way, other sources can be easily added by creating a folder in `DataObtainersOutlierDetectionPhase` and implementing the scripts to acquire data and convert this data to these pickles.

DBLP In folder `/DataObtainersOutlierDetectionPhase/dblp/PublicationsPerYear` the scripts can be found that are used to create the pickle for the publication data. This script uses the same database file as the one used for finding the scientists.

Google Scholar In folder `/DataObtainersOutlierDetectionPhase/Google_Scholar` different scripts, folders and files can be found. In `authors.txt`, the copied file containing extracted names from the DBLP database can be found. The script `profiles_finder_all_publications.py` uses this file to search for profiles of authors in Google Scholar. The file `progress.txt` keeps track of the progress. This way, the script can be stopped and started again without starting all over again. If a name has been found, the complete profile of the scientist will be saved as a .html file in the folder `google_scholar_profiles_all_pubs_x`, where x is a number. As there is a limit to how many profiles can be saved to one directory, there are multiple directories containing the profiles. When the limit is reached, a new folder need to be created and the script needs to be adjusted to save the profiles to that directory.

Make sure to provide a valid cookie of Google Scholar when acquiring profiles. With the use of this cookie it will be possible to search for Google Scholar profiles for up to two hours. The location of the cookie should be manually changed in the script `profiles_finder_all_publications`.

After enough profiles are found, the publication and citation data can be extracted using the `Extract_citations_publications_profiles.py` script. This script converts every profile into the data structure necessary for the next step. Make sure to run this script for every directory containing profiles. The resulting pickle is saved in the same directory of the profiles.

To manually add profiles, simply create a new directory, download the HTML of the profile, save this in the newly created directory, and run the extraction script. The resulting pickle can then be used in later steps.

Semantic Scholar In folder `/DataObtainersOutlierDetectionPhase/Semantic_Scholar` the scripts can be found that are used to create the pickle for the citation data. The profiles used during the research were acquired and saved using another PC. All of these profiles were compressed to one zip file, and the zip file was split into smaller files that fit on the USB drive. All the profiles were converted to a `citations.pickle` file, as only citations were acquired. The `citations.pickle` file contains all the citation data of all the profiles acquired with Semantic Scholar that are in the zip file. This `citations.pickle` file can be found in the directory `semantic_scholar_profiles_cits_1`.

To manually add profiles, simply create a new directory, create a new `author.txt` file containing the names to be searched, and set the value in `progress.txt` to zero. Then run the `profiles_finder_all_citations.py` script to collect the profiles of the names.

After enough profiles are found, the citation data can be extracted using the `Extract_citations_publications_profiles.py` script. This script converts every profile into the data structure necessary for the next step. Make sure to run this script for every directory containing profiles.

An attempt has also been made to create scripts to acquire the publications of Semantic Scholar, which can be found in the same folder. As we were not successful, these scripts only show the attempts. Further research might use these scripts as a starting point.

A.4 Finding outliers

Parsing many HTML files and extracting information is time sensitive. When testing and adjusting the different measures, this will actually take most of the time. And, to keep everything generic, extracted data saved to a generic data structure as pickles are used. All the different publication data and citation data (if available) of the different publication data processors should convert their data to this data structure.

In the root folder, the script can be found to find the outliers of the outlier detection phase. There are two scripts, one that uses the API `FindOutliers_usingAPI.py`, and one that does not (`FindOutliers.py`). Both scripts make use of another API: `OutlierDetectAlgorithms`. The measures defined in this API can be found in the folder `OutlierMeasures`. In the API, the algorithms are defined that implement the measures. In this folder, another script can be found that calculates helper functions (`helperFunctions.py`). This script can be used for calculating the boxplot values.

When running the `FindOutliers.py` script, all the measures will be executed of all the data sources. This script can easily be manually adjusted to only run the measures for a specific publication data processor, by out commenting the other publication data processors in the main function.

Other measures and extra data sets can easily be added manually. Other measures can be implemented in the API `OutlierDetectAlgorithms`. Every new measure should accept a list of publications and a list of citations, even when they are not used. The new measures need to be added to a list in the script `FindOutliers.py` that contains all the measures to run for a specific publication data processor. Extra data sets can be added by adding the extra pickle for a specific publication data processor to the set.

Of every outlier found with Google Scholar, the option exists to create an outlier peer folder structure. Simply set the option copyProfiles to True. This will not only create the outlier peer folder structure (`OutliersWithPeers`), but it will also copy all the outlier HTML files to the folder `OutliersFirstPhase`. This folder can be investigated to see the profiles of the Google Scholar outliers. The peer folder structure will be used in the next step to find the peers of the outliers.

A.5 Finding peers

A database of journals and their editors can be created by running the script `/PeersFinders/GetCoEditors/Elsevier/GetEditorialBoards.py`. This will find and svae all the editorial boards of Elsevier and their editors.

A selection of the IT boards has been made available in the folder `/PeersFinders/GetCoEditors/Elsevier/boards`. Finding peers is made available by the script `/PeersFinders/GetCoEditors/Elsevier/GetCoEditorOf.py`. This script searches whether an outlier found in the peer folder structure is an editor of a journal. As names can be written many ways, this script also searches for partial names. Every possible journal the outlier might be an editor of will be written to the `results.txt` file. Every outlier needs to be checked manually by name in this file. After a correct journal has been found, the profiles of the members of the journal should be copied manually from `PeersFinders/GetGoogleScholarProfilesCoEditors/Elsevier/profiles` to the `OutliersWithPeers` folder. The Google Scholar profiles of the editorial board members can be acquired by the script found in folder `PeersFinders/GetGoogleScholarProfilesCoEditors/Elsevier`.

A.6 Acquiring data for peer comparison phase

Acquiring the more detailed data used in the peer comparison phase is done by the scripts that can be found in the folders
`/DataObtainersPeerComparisonPhase/Google_Scholar/FindArticlesOutliersAndPeers` and
`/DataObtainersPeerComparisonPhase/Google_Scholar/FindCitationDataArticlesOutliers`.
These scripts will process every outlier and peer in the folder `OutliersWithPeers` and `OutliersFirstPhase`, and acquire the data necessary. Note that this will take a very long time, especially acquiring the citations. As Google Scholar is the only publication data processor used in this step, captchas need to be filled in once in a while to obtain a new cookie to be used to extract all the data necessary. The data acquired will be stored in the subfolder `Outliers_processed`, found in
`/DataObtainersPeerComparisonPhase/Google_Scholar/FindArticlesOutliersAndPeers` and
`/DataObtainersPeerComparisonPhase/Google_Scholar/FindCitationDataArticlesOutliers`.

The script `CreateAPIDataForTest` in the folder `API` creates the generic data structure used by the `InducedPubViewRelations` API. This script extracts data extracted from Google Scholar from all the folders where data is stored, maps it to the generic data structure and saves it as a pickle. The generic data structure used by this API must have the following structure:

{'N. Tielenburg' : [{publicationName1 : ({'property1' : 'value', 'property2' : 'value'}, ['citingArticle1', 'citingArticle2']) }, {publicationName2, : ({'property1' : 'value'}, ['citingArticle1'])}] ,
'D. Runhart' : etc. }

In short, this structures is a dictionary with the author name as key. It contains another dictionary as value. This dictionary contains all of the publications titles the author authored as key. The value is a tuple of a dictionary and a list. The dictionary contains meta data about the publication, like the year it is published. The list contains all the publications citing the publication.

A.7 Comparing outliers and peers

In the root folder the script `CompareOutliersToTheirPeers.py` can be found to compare the outliers of the peer comparison phase. This script investigates all outliers found during the outlier detection phase. Note that this script is not generic, it is not suitable to be used with other publication data processors than Google Scholar. This script uses functions defined in the scripts in the folder `PeerComparisonMeasures`. These two scripts (`ObtainMeasureDataOfCitations` and `ObtainMeasureDataOfPublications`) extract data from raw Google Scholar profiles and the more detailed data acquired of Google Scholar of the outliers and peers to calculate the measures.

The API `InducedPubViewRelations` can also be used in this phase. An example has been given in the `CompareOutliersToTheirPeers_usingAPI`. This script is generic and therefore need only slight modifications to be used with other publication data processors. Other publication data processors simply need to convert their data to the generic data structure. Note that not all measures are implemented in this script, this is just a proof of concept as not all the data was available to fill the generic data structure to be able to use the API to its full potential.

A.8 Detecting outstanding scientists experiment

The folder `/DataObtainersOutlierDetectionPhase/dblp/experimentScientists` contains the script to find the outstanding scientists. Run this script to acquire all the names of scientists with over 40 publications in the named venues. Use these names to find and acquire the profiles as explained above. Rename all the profiles so they start with valid_. Create the pickles of these profiles accordingly and add these pickles to the list of pickles used when detecting outliers. Then, run the outlier detection scripts to find out if the scientists are outliers.

A.9 Other noteworthy details

The folder `Driver` contains the chrome driver. This driver is used by selenium to acquire profiles of Google Scholar and Semantic Scholar. It is also used to acquire the more detailed information in the peer comparison phase.

The folder `API` contains the `InducedPubViewRelations` API. The functions defined in this API can only be used with the generic data structure.

A simple unittest has been created to test the calculation of the medcouple in the folder `UnitTests`. This unittest tests whether our brute force calculation of the medcouple provides the same result as the medcouple calculated by statsmodels.

B Distributions

B.1 *maxdiffcitecount* distribution of Semantic Scholar

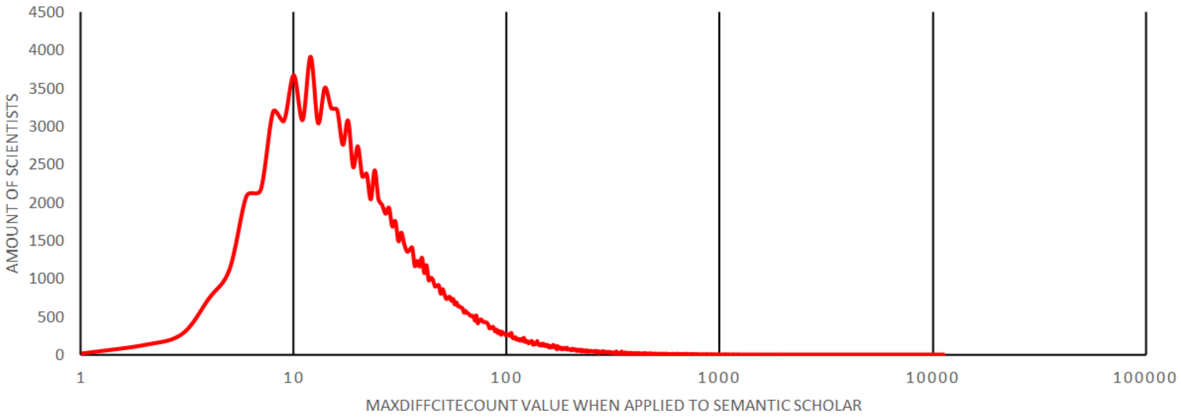


Figure 17: Distribution of the results of measures *maxdiffcitecount* applied to Semantic Scholar data

B.2 *maxdiffpubcount* distribution of DBLP

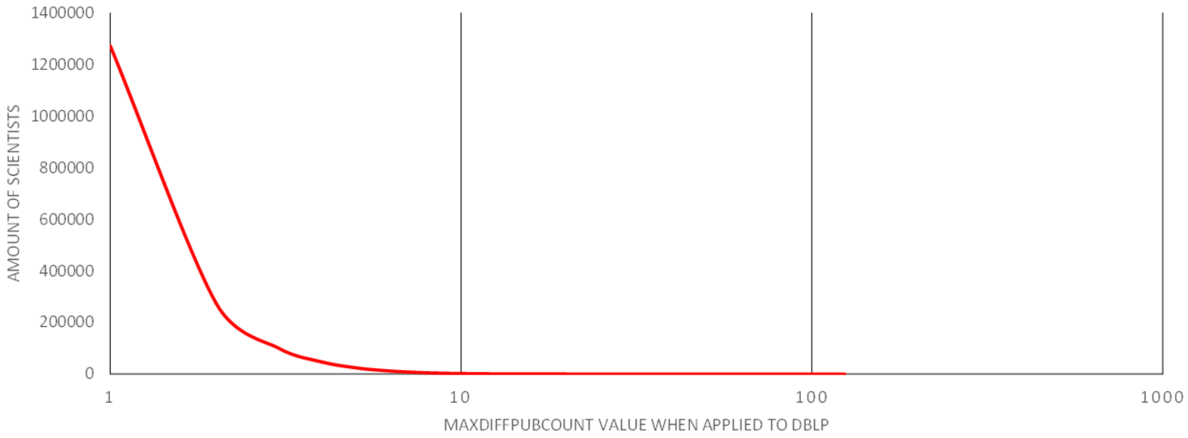


Figure 18: Distribution of the results of measures *maxdiffpubcount* applied to DBLP data

B.3 *toomanypubs* distribution of DBLP

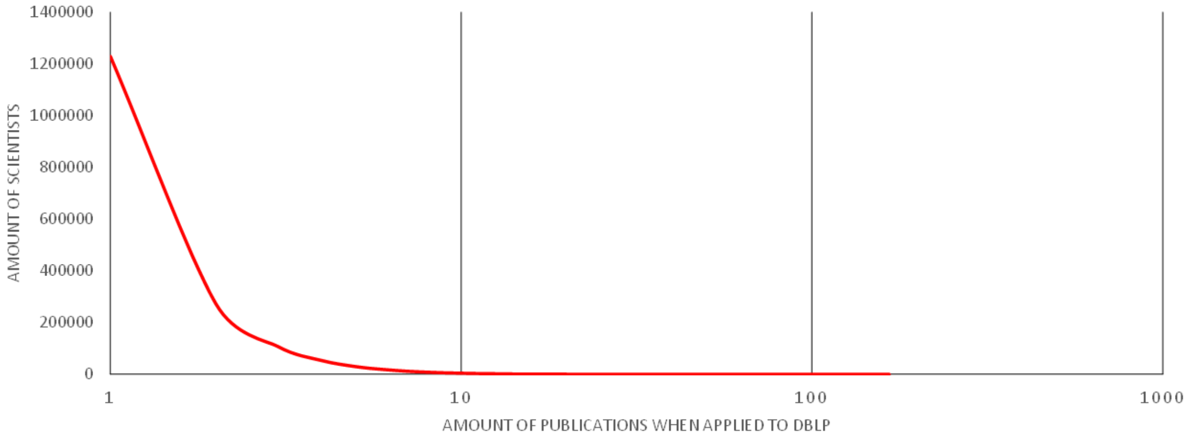


Figure 19: Distribution of the results of measure *toomanypubs* applied to DBLP data

B.4 *maxdiffpubcount* distribution of Google Scholar

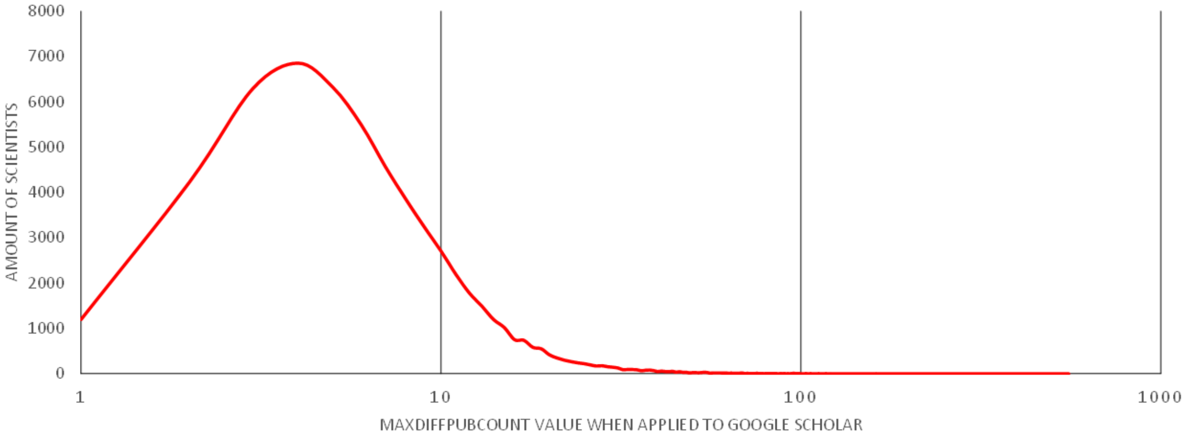


Figure 20: Distribution of the results of measures *maxdiffpubcount* applied to Google Scholar data

B.5 *maxdiffcitecount* distribution of Google Scholar

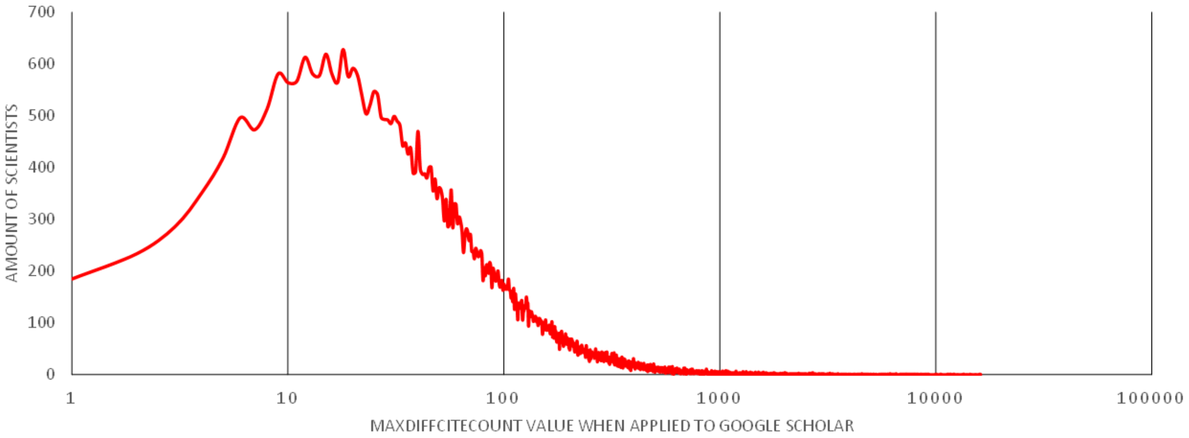


Figure 21: Distribution of the results of measure *maxdiffcitecount* applied to Google Scholar data

B.6 *maxratiocitsvspubs* distribution of Google Scholar

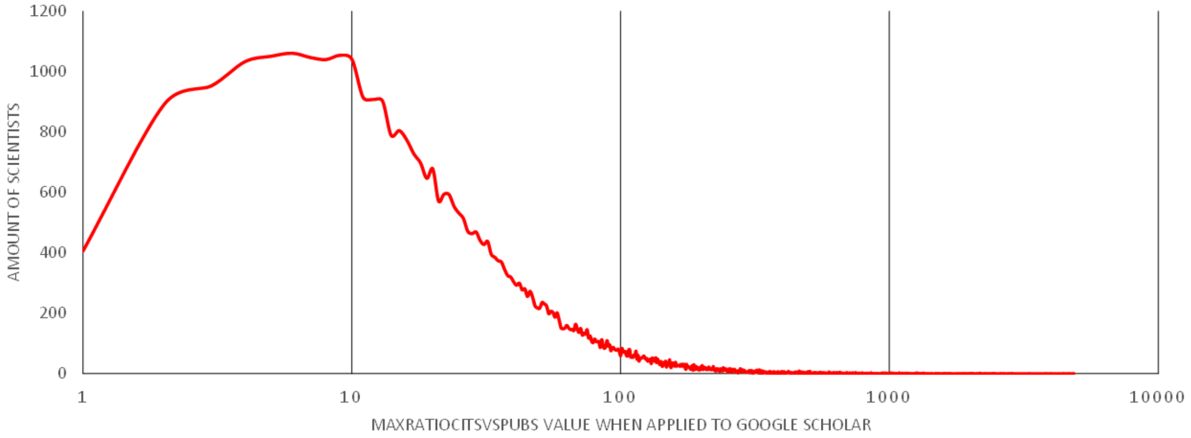


Figure 22: Distribution of the results of measures *maxratiocitsvspubs* applied to Google Scholar data

B.7 *toomanypubs* distribution of Google Scholar

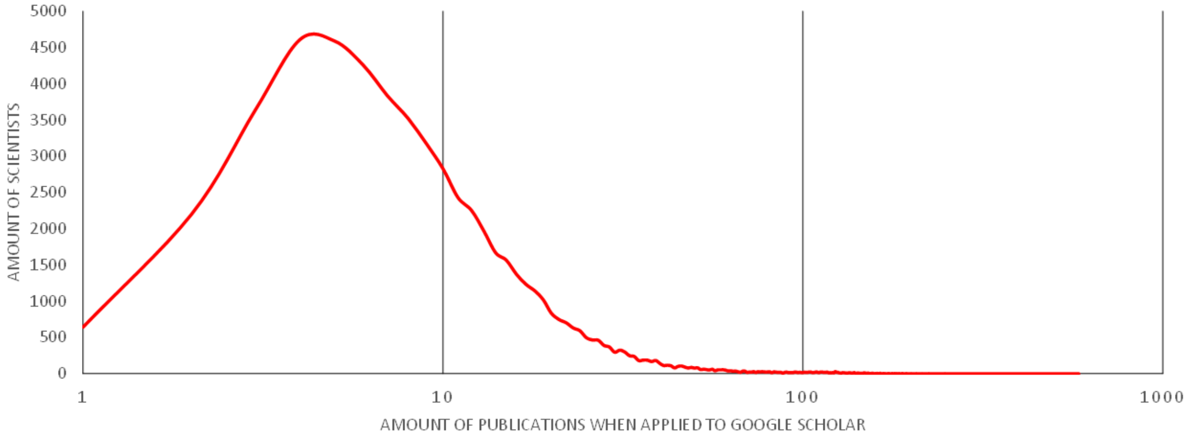


Figure 23: Distribution of the results of measure *toomanypubs* applied to Google Scholar data