Analytics on Google Scholar Profiles: Fake Citations in Scholar Profiles

Muhammad Hamza Madni
Muhammad Hassan Iqbal

Supervisor: Dr.Fareed Zaffar

Lahore University of Management Sciences

CONTENTS

1 ABSTRACT

Google Scholar is one of the most eminent web search engines that allow users to search for scholarly literature across all publishing formats and disciplines. The Google Scholar searches include academic journals, conference papers, books, and any other form of scholarly literature. Among other features, Google Scholar also allows authors to create their profiles and link articles and journals to their profile based on which the profile is given a total number of citations and various other indexed rankings. However, it was observed that some of the profiles included articles that were not written by the author, but the Google Scholar still included the citations for those articles in calculating the total number of articles for the scholar profile. This claim is validated in this paper by analyzing various data sets and deploying different methods to support the hypothesis and provide evidence of the flaws within the search engine. Previously, research conducted on Google Scholar shows that Google has always modified its website to cater to those faults, but the research presented in this report is novel and does not derive from any previous research. Initially, a randomly selected sample of 736 profiles consisting of well-renowned scholars around the United States of America, was chosen for analysis. Out of those 736 profiles, 690 profiles showed that at least one of the articles did not belong to them and out of those 690 profiles, 19 profiles had more than 100 such articles which did not belong to the author. To further investigate, a profile was created from scratch to see what conditions Google Scholar does not cater to while populating a new profile with its associated scholarly literature. Finally, another dataset was collected which grouped the authors based on regions to see whether this issue is region-specific or not.

2 LITERATURE REVIEW

This section summarizes the previous work done on Google Scholar and any related work which helped reach any final conclusions for defining the methodology applied in section 3.

2.1 GOOGLE SCHOLAR AS A NEW SOURCE FOR CITATION ANALYSIS

https://www.int-res.com/articles/esep2008/8/e008p061.pdf

The paper compares Google Scholar to other databases like Thomson ISI Web of Knowledge. It highlights some of Google Scholar's disadvantages like its uneven coverage between different fields of studies and its poor performance for older publications compared to other databases. The paper also mentions Google Scholar's limitations when dealing with authors with special characters in their names. Furthermore, at times the search engine's automatic processing, produces nonsensical results like wrong titles or wrong authors names. Another problem highlighted by this paper is such that the database update frequency of this search engine is low in comparison to other such platforms. Despite these drawbacks the paper concludes that Google Scholar's H-index (a parameter which highlights authors that have a continuous contribution towards the field) is a much better and robust metric to use compared to average citations. As this parameter gives more recognitions to authors that have contributed more papers with an average number of citations than an author that might only have a handful number of papers with many citations.

2.2 Advisor Reviews—Standard Review Google Scholar

https://www.researchgate.net/publication/263572400_Google_Scholar

The literature talks about some of flaws in the design and implementation of the search engine in Google Scholar and provides solutions for correcting them. One minor flaw is that search results are ordered by relevance rather than the date of publish, by default, unlike other academic databases. The paper also explains that since Google Scholar did not use publisher supplied metadata there are several errors in the database which cause inflated citation counts due to the inclusion of both master records and citation records for individual articles. There is also confusion about what sources google crawls to build and update its database. Another error that occurs due to not using publisher metadata is the appearance of ghost authors. These "ghost authors" often take their names from other fields in the document, resulting in clearly erroneous author names such as P Login (for Please Login) or A Registered (for Already Registered). The paper also highlights the publication date errors with Google Scholar. The author states that conducting an Advanced Scholar Search and limiting the date range to articles published between 2012 and 2025, for example, returns more than 1,700 articles, all with wrong dates of publication.

2.3 Using Google Scholar to track the scholarly output of research groups

https://link.springer.com/content/pdf/10.1007/s40037-019-0515-4.pdf

This piece of literature discusses how Google Scholar profiles for research groups can be used to track their progress and productivity, using metrics like the h-index, several papers published per year and citation counts. This can be used by organizations while giving research grants and enables them to easily assess the productivity of the group or individuals. The paper also emphasizes the importance of manually approving updates to profiles as Google Scholar might add articles automatically if manual approval of updates is not turned on, overestimating the productivity of the individual or group. However, Google Scholar profiles can, nevertheless, be a useful and powerful tool for tracking an individual's or research group's progress compared to other databases as the other databases are not well structured or require premium access. This trait of the search engine allows it to be used by people when looking for competent research groups or individuals in a certain field butt at the same time it can be easily fashioned to make a person look more productive than they are by tampering with their profile and adding articles that don't belong to them.

2.4 Google Scholar: An Internet-Based Program for Finding and Tracking Citations

https://archives.ourheritagejournal.com/index.php/oh/article/view/4273

The research article explains the key features of Google Scholar and how to use the service to get better search results. The author gives tips on how to improve your search queries and get relevant results as searching specific content could be hard for certain people. This research article also mentions the option to automatically add articles to your profile that google assumes belong to you and advises to manually check or confirm them. The article further explains all the features and limitations of Google Scholar in simple. It is a great resource for anyone new to Google Scholar regardless if they are making a profile or just searching for something as it gives tips for both.

3 METHODOLOGY

Google Scholar has shown itself to be a credible source for journals, articles and scholarly literature. The number of citations on a paper or the number of citations on a profile has been a measure for the general population to gauge the credibility and popularity of the specific paper. The initial hypothesis was that several profiles contained articles which were not written by the author of that profile. To look deeply into this and to see the granularity at which this problem could be found, some forms of datasets were collected, and analysis was performed on them. Other than collecting data for analysis, a scholar profile was created on Google Scholar from scratch for an unknown professor which was later populated with articles to see the checks that Google Scholar applies before making the profile public.

3.1 INITIAL COLLECTION OF DATASET FOR PROBLEM IDENTIFICATION

Names of 1000 computer science professors and researchers from top-ranked universities were scraped from a website csrankings.org. CSRankings is a metrics-based ranking of top computer science institutions around the world. The website was used solely to get the names and institutions of the professors because it is assumed that a professor from one of the top-ranked institute is likely to have more articles published under his name, therefore, the profiles scraped would have had enough data for analysis. These professors belonged to institutions such as Carnegie Mellon University, University of Illinois and University of Michigan. A complete list of these professors and their institutions can be found in the dataset provided.

A scraper was used to scrape the data from Google Scholar. The scraper was coded in Python3 and it used selenium to perform its tasks. The scraper opens every public profile on Google Scholar from the list of professors provided to it from CSRanking. It then pulls all the data for the Scholar's profile into a text file. Out of the 1000 names from the CSRanking, 736 text files were created as these professors had public profiles and the other empty profiles were discarded. For each of those 736 profiles, a separate text file was created with the following information scraped:

- Link of the scholar profile
- Total citations of the profile
- Names of the first 500 articles on the profile with its citations, authors and publication name.

From these text files, a cumulative data was further extracted in an excel file which was then used for the analysis of these profiles. This final dataset in the excel file contained the following attributes:

File Name: Text file name
Author Name: Author Name (As on Google Scholar page)
Link: Link to the page of the Google Scholar profiles
Total Articles Count: Count of total articles on that page
Total Citations Count (Copied): Total citations for the Author profile as on Google Scholar
Total Citations Count (Calculated): Total citations for the Author profile by adding the citations for each article on that page.
… Count: Count of the article on a page with a lesser number of authors shown ***
… Citation Count: Total count for the citations for the articles for lesser number of authors shown.

False Count: Discrepancies count on the page
False Citation Count: Total count for the citations of the articles with discrepancies found.

*** The articles for lesser number of authors shown refer to the articles on a page where not all the authors were shown on the main google scholar page. In such cases, the google scholar page shows a "..." symbol after some author names and the rest of the author names are mentioned inside the document of the paper. The mentioned author names for the research paper/article do not contain the page Author name in it before the "…" hence these articles are not considered as discrepant yet and might need manual verification.

The dataset was thoroughly analyzed, and the findings are shared in the Results section 4.1.

3.2 Creating a profile from scratch on Google Scholar

A scholar profile was created from scratch on Google Scholar to test for the conditions that the webpage accounts and check for the loopholes which allow an author's profile to contain articles which do not belong to it.

A professor at Lahore University of Management Sciences was chosen at random to create a duplicate of his profile by filling in the same details, as on his Google Scholar profile, in a newly generated file. The website asked for the name of the professor, Affiliation, areas of interest and a verification email for the instructor. Figure 3.2.a shows the information required by the Google Scholar. Later the profile was populated using the different options provided by Google Scholar automatically which included all the articles containing the author's name in its list of authors and other ways of adding an article manually. Finally, the profile update was set to automatic and the profile was set to the public after verification.

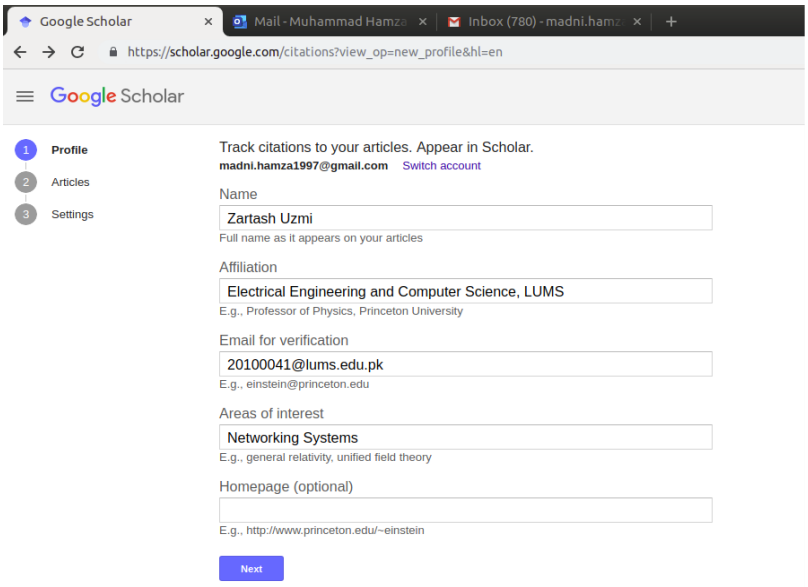

Figure 3.2.a

3.3 Scraping Scholar Profiles by region in the World

For a more in-depth clarity into the problem, we analyzed our hypothesis over another dimension: Region. This was necessary to see if the problem of fake citations and articles was generic to all the profiles around the world or found in certain regional profiles more than the others.

To achieve this, the countries in the world were roughly divided into 12 regions and then 10 universities were selected at random from every region. Among each of those universities, 2 instructors were chosen at random with public Google Scholar profiles adding up to a total of 240 public profiles. A dataset was formed with the Professor Name, Link to the Scholar profile, University and Region. The list of these regions and the universities selected in each region can be seen in Appendix.

The scraper used in section 3.1 was modified for pulling some additional information from the profiles this time and then it was given the list of the 240 professors along with their links to the scholar profiles. Due to the limited computation power only first 300 articles were scraped for each of the authors, given the author had more than 300 articles on his profile. A separate text file was generated for each of the scholar profiles with the following attributes:

Author Name: Author Name (As on Google Scholar page)
Link: Link to the page of the Google Scholar profiles
University: University or Institute associated with the author
Region: Region to which the University belongs.
Total Citations: Total citations for the Author profile as on Google Scholar
For Each Article:
Name of Article: Name of the Article.
Citations of the Article: Total citations for the article
Authors for the article: Complete list of Authors for the article on Google Scholar
Publication Date: Date of Publishing
Publication: The Publication House or Institute by which the article was published
Crossed: Whether the article citations are marked with an asterisk ****
Starred: Whether the article citations are crossed over *****
Year: Year in which the article was released

****An asterisk on citation means that the citations on this article might be different than what has been shown on the profile.

*****A cross over the citations means that the citations for this specific article have been merged into another article and therefore not counted towards the total citations for the profile. The crossed and starred citations are not used in any analysis in the results section.

From these 240 text files a dataset was further derived consisting of the following attributes:

Name: Author Name (As on Google Scholar page)
Region: Region to which the University belongs.
University: University or Institute associated with the author
Link: Link to the page of the Google Scholar profiles
Total Citations: Total citations for the Author profile as on Google Scholar

Total Citations since 2015: Increase in Total citations since 2015
H Index: The max value of h such that an author has written h number of papers with each of those papers cited at least h times.
H Index since 2015: The max value of h since 2015.
I10 Index: Number of publications with at least 10 citations.
I10 Index since 2015: The value for I10 index since 2015.
Total Articles: Total number of articles scraped from the profile. (The total number will be 300 if of the profile has more than 300 articles).
Calculated Citations: Provides us with the total of citations for the articles scraped from the profile.
Fake Articles: The number of articles in a profile that do not enlist the author of the profile as one of the authors on Google Scholar.
Fake Citations: The number of total citations for the fake articles in a profile.
Star Articles: The number of articles with citations marked with an asterisk.
Star Citations: The total number of citations for the Star Articles.
Dash Articles: The number of articles with citations crossed over with a line.
Dash Citations: The number of citations for Dash Articles.
Per cent of Fake Articles: Fake Articles divided by the Total Articles.
Per cent of Fake Citations: Fake Citations divided by Calculated Citations.

A thorough analysis for this dataset is shared in the Results section 4.3.

4 RESULTS AND FINDINGS

This section summarizes the results for all the 3 modes of research shared in the previous sections and shares insights about the Google Scholar Profiles.

4.1 INITIAL DATASET ANALYSIS

- ➔ Out of the 736 profiles scraped and tested for discrepancies, 690 profiles contained at least one article which did not belong to the author of the profile.
- ➔ The distribution for the number of discrepancies in the 690 profiles is shown in Figure 4.1.a. It shows that 186 profiles had exactly 1 discrepancy whereas 19 profiles had more than 100 discrepancies.

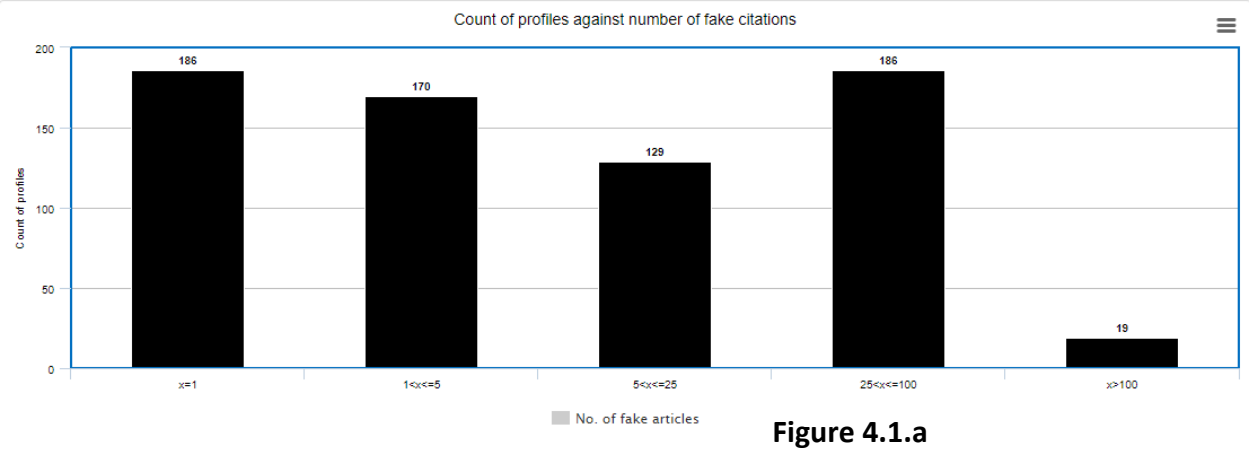

Figure 4.1.a

- ➔ Out of the 690 pages with discrepancies, the highest number of discrepancy (false count) was seen in the profile for Gang Wang with a total false count of 377 articles from the total of 500 articles present on his Scholar profile with a hit rate of about 75.40%. Figure 4.1.b shows the author names with the highest number of False Counts, Hit ratio and other relevant data.

	A	B	C	D	E	F
1	Author Name	False Count	False Citation Count	Total Articles Count	Total Citations (Copied)	Hit rate (False Count)
2	Gang Wang	377	13928	500	25943	75.40
3	Taesoo Kim	341	4450	500	9865	68.20
4	Xiaojin Zhu	317	3649	500	20399	63.40
5	Ming Lin	167	1464	500	29377	33.40
6	Graham Neubig	131	126	484	3777	27.07
7	Mark Hill	99	665	447	27449	22.15
8	Krste Asanovic	86	3513	340	16159	25.29
9	Borivoje Nikolic	81	3298	318	17998	25.47
10	Insup Lee	79	736	500	16231	15.80

Figure 4.1.b

➔ The dataset was sorted by “False citation count” and another field was added to the dataset with attribute name False citation ratio where:

$$False\ Citation\ Ratio = \frac{False\ Citation\ Count}{Total\ Citations\ (Copied)}$$

The new dataset showed, for each data entry, the percentage of citations that did not belong to the Scholar on his Google Scholar profile.

Figure 4.1.c shows the profiles with the highest counts for False citations and their respective False Citation ratios.

Author Name	False Citation Count	Total Citations (Copied)	Total Articles Count	False Citation Ratio
Serafim Batzoglou	41176	61279	151	67.19
Takeo Kanade	32653	145586	500	22.43
Stephen Wright	26248	51937	325	50.54
Niklas Elmqvist	16141	21057	239	76.65
Edmund Clarke	13961	71810	500	19.44
Gang Wang	13928	25943	500	53.69
Jennifer Widom	11896	62714	349	18.97
Hector Garcia-Molina	11234	88953	500	12.63
Pavel Pevzner	11047	48071	398	22.98

Figure 4.1.c

- The figure above shows that “Serafim Batzoglou” has the highest False Citation Count and a False Citation Ratio of 67.19%. The very high False Citation Count is due to the first few articles posted on the profile which do not list the Author as one of the authors for the articles. The profile for Serafim Batzoglou can be found in the dataset provided.
- Another high False Citation Count case can be seen in the profile of Takeo Kanade who is an instructor at Carnegie Mellon University. The False Citation Ratio is 22.43 per cent and it is majorly due to the first article in his profile. The Google scholar profile for Takeo Kanade can be found in the dataset provided.

4.2 Creating a profile from scratch on Google Scholar

A profile was created on Google Scholar for an instructor from Lahore University of Management Sciences. The following section describes the findings in this process.

➔ The profile was created using any Google Mail ID therefore the name used in the Gmail ID and the Scholar profile could be different.

- The name for the Gmail ID was Muhammad Hamza Madni and the Scholar name was set to Zartash Uzmi. Due to this any person will be able to create a profile for any professor in the world.

➔ Since Dr. Zartash was affiliated to LUMS in this profile therefore the profile could be verified using any email address associated to the given institute.

- The Email used for the verification was 20100041@lums.edu.pk as seen in Figure 3.2.a.
- A non LUMS email ID could not verify the profile.
- Due to this any person belonging to the same institute, instructor, staff or student, would be able to create a profile for the instructor and be able to verify it using his own assigned university Email ID.

➔ Selecting Articles to add to the profile:

- The profile shows by default grouped articles. These articles were grouped by Google Scholar itself based on naming conventions used generally. A total of 77 articles were grouped into a single check box and therefore can be selected altogether. All the articles in this case had Dr. Zartash as a listed author.

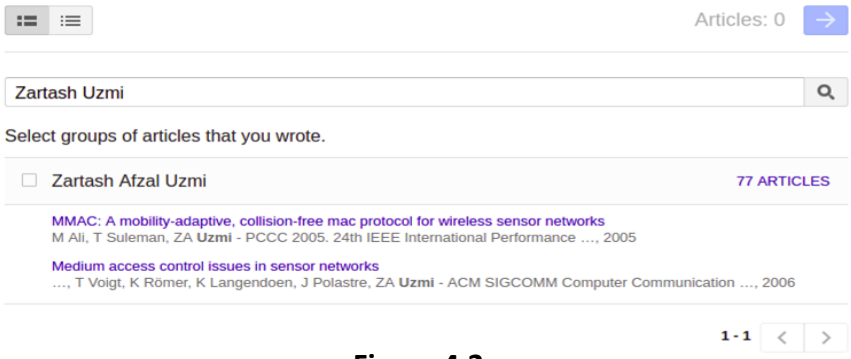

Figure 4.2.a

- Selecting the List view for these articles, gives an option to select each article manually from the given list of articles. The non-exhaustive list has some initial articles with Dr. Zartash listed as an author but as we go deeper into this list, we start to find articles from different authors. These articles can still be selected for being a part of the profile and their citations will be counted towards the total citations of the profile.

➔ Profile Update Settings:

- Google Scholar provides two options while creating a profile regarding the profile updates; “Apply updates automatically” or “Email me updates for review”; Figure 4.2.b. With time Google Scholar finds more relevant articles for the created profile and either automatically updates them without notifications or sends them in email for review and then updates accordingly.

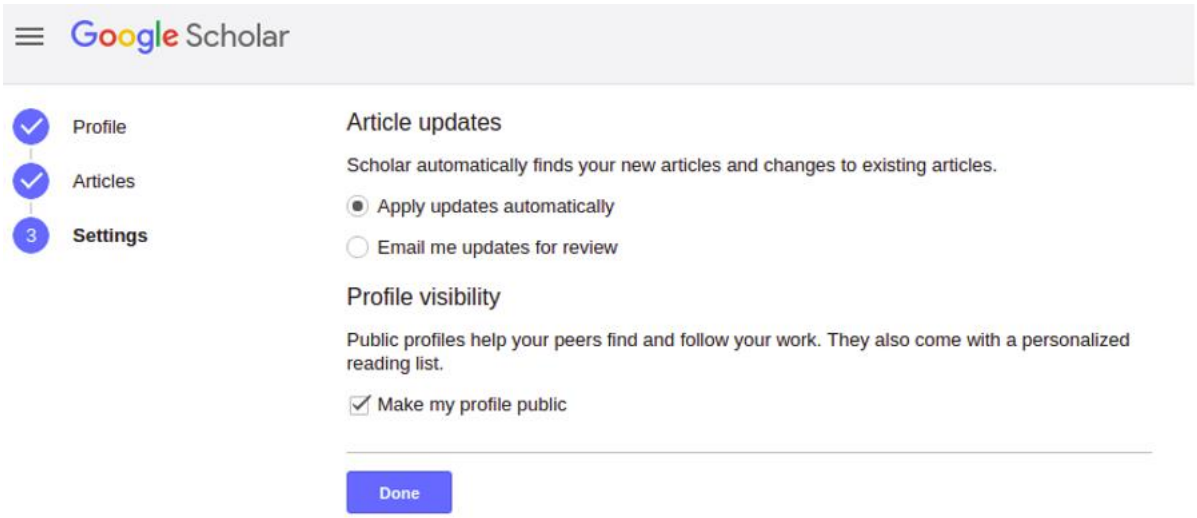

Figure 4.2.b

4.3 Scraping Scholar Profiles by Region in the World

A new dataset was collected for this study and the profiles were chosen randomly among regions. A total of 12 regions were selected and 20 profiles for each region were shortlisted for the study. The section refers to the findings for the section 3.3.

- ➔ The Percentage of Fake Articles for each Region are shown in form of a Boxplot in Figure 4.3.a.
  - It shows that the mean percentage fake articles for each region is less than 10 percent except for Eastern Europe where it is approximately 17 percent. The western and southern Asia show the least mean percentage of fake articles.
  - Regions of Eastern Europe, Eastern Asia, North America, and South Africa show very huge 3rd quartiles going up to 65 percent in case of Eastern Europe.

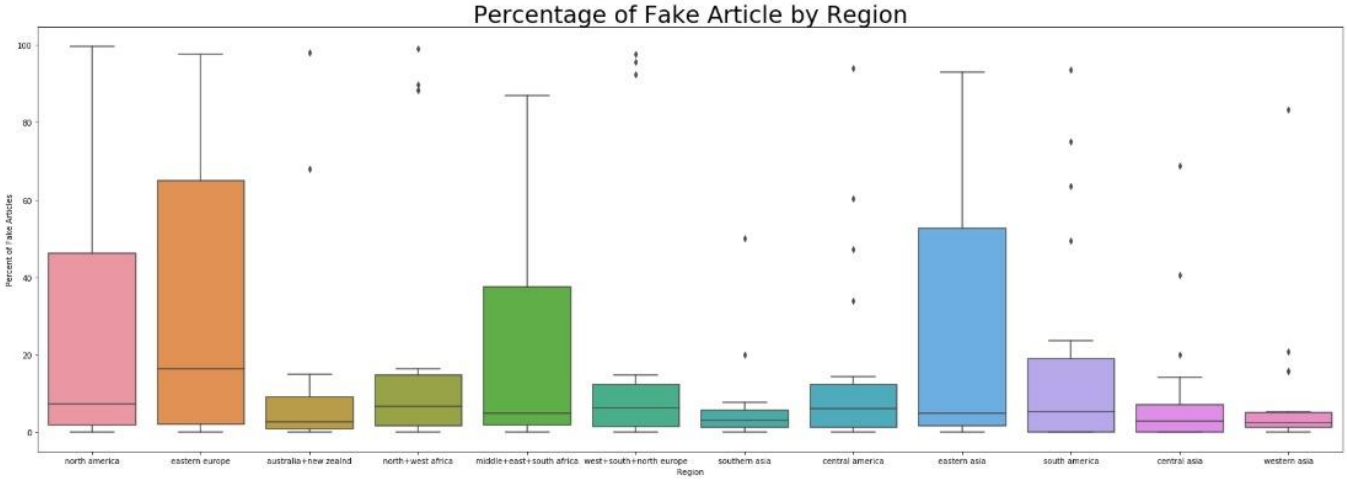

Figure 4.3.a

➔ The Percentage of Fake Citations for each Region are shown in form of a Boxplot in Figure 4.3.b.

- The Regions for North America, South Africa, East Asia and South America rose by a great margin as this shows that the number of fake articles in profiles in these regions are lesser but the citations in those articles are very high.
- Asian Regions and Australia seem to be least involved with their lowest means and quartiles.

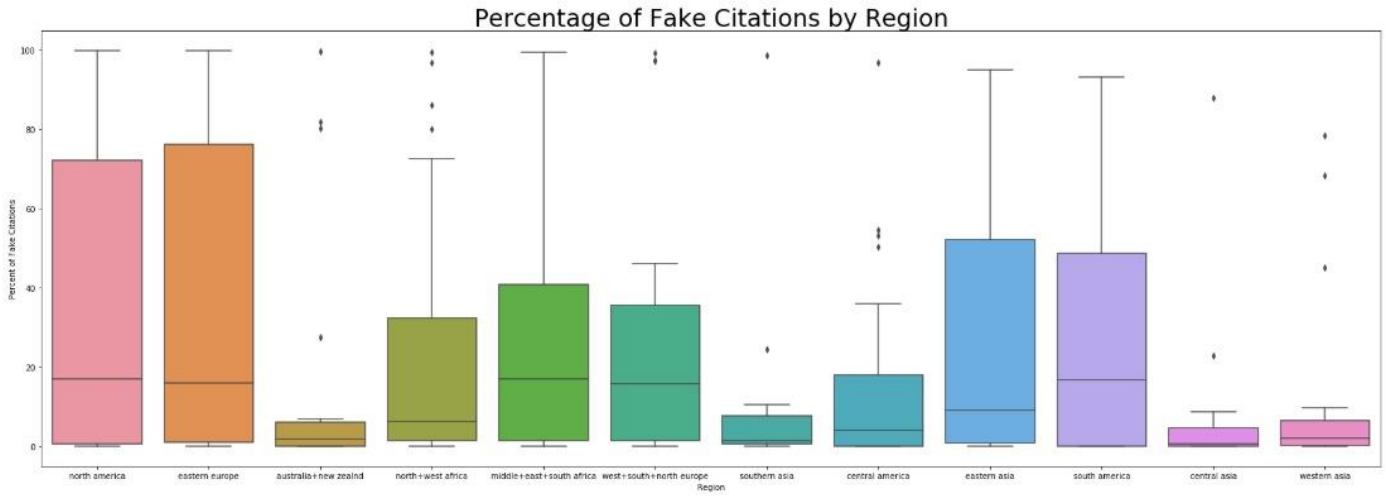

Figure 4.3.b

5 CONCLUSION:

The findings in the previous section summarize that a lot of profiles that can be publicly accessed on Google Scholar contain articles that do not belong to the author of the profile and the citations of these articles are then accounted towards the total citations of the profile and also for calculating the H-index and I-10 index of the profile. With a greater number of citations and highly cited articles, the profile gets a higher index value due to which the profile gets more popularity than the other profiles. With the current research conducted it is unsure that whether it is all Google Scholar's fault that there are lesser checks while creating and updating profiles due to which the profiles are allowed to keep these articles or should we blame the profile owner for this dishonesty but both the parties do play a role. It can also be seen that this issue is more seen in profiles belonging to the regions of Eastern Europe, North and South America and Eastern Asia as they show the highest numbers for percentage of fake articles and citations. For a further study and analysis, more data should be collected to make a stronger claim and some duplicate profiles should be created to be analyzed over time.

6 APPENDIX:

LIST OF REGIONS AND UNIVERSITIES USED IN METHODOLOGY 3.3

Middle, east and south africa	North and west africa	central asia	southern asia	western asia	eastern asia	south america	north america	central america	eastern europe	West, south and north europe	Australia and new zealnd
University of Cape Town	Mohammed V University of Rabat	Tajik National University	University of Colombo	Alfaisal University	University of Hong Kong	University of Campinas	University of toronto	University of Guanajuato	Novosibirsk State University	National and Kapodistrian University of Athens	University of Cantebury
University of Johannesburg	Cairo University	Kyrgyz State Technical University	University of Peradeniya	United Arab Emirates University	Hanyang University	University of São paulo	Georgia institute of technology	University of Sonora	Saint Petersburg State University	Ruprecht Karls Universität Heidelberg	University of sydney
University of the Witwatersrand	Mansoura University	American University of Central Asia	University of Delhi	Jordan University of Science and Technology	Tsinghua University	Universidad Peruana Cayetano Heredia	california Institute of technology	The Unbiversity of the West Indies	Lomonosov Moscow State University	University of helsinki	University of Melbourne
University of Pretoria	University of Ibadan	Turkmen State Power Engineering Institute	Quaid-i-azam University	Al-Balqa Applied University	Korean Advanced Institute of Science and Technology	pontifical Javeriana University	McGill University	University of Havana	National Research University Higher School of Economics	University of Cambridge	University of Queensland
North-West University	Aswan University	International Ataturk Alatoo University	University of Dhaka	University of Science and Technology Sana'a	National University of Mongolia	University of Desarrollo	Duke University	National polytechnic Institute,mexico	National Taras Shevchenko University of Kyiv	University of Copenhagen	University of New South Wales
Stellenbosch University	Sidi Mohamed Ben Abdellah University	National University of Uzbekistan	Alzahra University	Sultan Qaboos University	Mongolian University of Science and Technology	University of Antioquia	University of British Columbia	University of Costa Rica	Vilnius University	Trinity College Dublin	University of Auckland
University of Nairobi	Universit of Ghana	Tashkent University of Information Technologies	Aligarh Muslim University	King Andulaziz University	University of Tokyo	University of the Andes	university of Michigan-Ann Arbor	Autonomous University of the State of Mexico	University of Tartu	University of Oxford	Victoria University of Wellington
Makerere University	Covenant University	Al-Farabi Kazakh National University	Amity University	Khalifa University	Peking University	Del Rosario University	McMaster University	Autonomous University of Sinaloa	Sumy State University	Université de Strasbourg	University of otago
University of Souh Africa	University of Biskra	University of Central Asia	Amirkabir University of Technology	Cankaya University	Kyoto University	National University of Cuyo	University of Montreal	Metropolitian Autonomous University	Tallinn University of Technology	Catholic University of Leuven	Divine Word University
University Of Dar es Salaam	Ferhat Abbas Setif University 1	L.N.Gumilyov Eurasian National University	COMSATS University	Qatar University	Seoul National University	National University of Corboba	University of texas at Austin	Monterrey Instute of Techonology	University of Latvia	University of Greenland	University of Papua New Guinea

DRIVE LINK TO THE COMPLETE DATASET FOR THE PROJECT.

https://drive.google.com/open?id=1cHLH1eLX-6T8RyvtkARo3nbpjw2XBYFJ