

2 Plagiarism Detection

This chapter⁶ provides a background on academic plagiarism. The rapid advancement of information technology and especially the dissemination of the Internet have drastically increased the availability of information – not only for legitimate purposes. Academic plagiarism is one form of undue information use simplified by the abundance of information and ease of information access [161].

In academia, plagiarism, i.e. using the words or ideas of another person and passing them off as one's own, has been described by some as a “cardinal sin” ([249], p. 1), maybe even the “ultimate sin” ([21], p. 57). Plagiarism deprives the original authors of the benefits of their work, including gaining academic reputation or acquiring research funding. Plagiarism may even shift these benefits to the plagiarist. Furthermore, plagiarism distorts the traceability of ideas, arguments and results within academic literature, and withholds valuable resources for discovering related material from the reader [306].

Given the volume of available information, detecting plagiarism through manual inspection is time-consuming and hardly feasible ([71], p. 9). Therefore, software capable of partially automating plagiarism detection has become increasingly popular. This section reviews the extensive and rapidly growing literature on research in academic plagiarism detection. Section 2.1 provides a definition, explains the forms of plagiarism, and discusses the prevalence of academic plagiarism. Section 2.2 gives a detailed description of plagiarism detection (PD) approaches currently in use, and an overview of the most effective PDS including performance evaluations follows in Section 2.3.

2.1 Academic Plagiarism

This section introduces the problem of academic plagiarism. Section 2.1.1 provides a definition, Section 2.1.2 characterizes the forms of academic

⁶ An abridged version of the literature review in this chapter has been published with Norman Meuschke [228].

plagiarism, and Section 2.1.3 concludes with a summary of the severity of the problem.

2.1.1 Definition

Inspired by the five key characteristics of plagiarism according to Fishman⁷ ([113], p. 5), we define plagiarism to encompass:

***The use of ideas, concepts, words, or structures without
appropriately acknowledging the source to benefit in a setting
where originality is expected.***

Other researchers commonly define academic plagiarism as literary theft, i.e. stealing words or ideas from other authors [102, 250]. Theft describes the deliberate appropriation of foreign property without the consent of the rightful owner ([120], p. 125). The definition used in this thesis does not necessarily characterize academic plagiarism as theft for the following reasons.

First, academic plagiarism need not be deliberate. Authors may inadvertently fail to properly acknowledge a source, e.g., by forgetting to insert a citation, or citing a wrong source; thereby committing plagiarism unintentionally [36, 219]. Additionally, a psychological memory bias called cryptomnesia can cause humans to unconsciously attribute foreign ideas to themselves [268].

Second, academic plagiarists may act in consent with another author, but still commit plagiarism by not properly acknowledging the original source. The term collusion describes the behavior of authors, who write collaboratively, or copy from one another, although they are required to work independently [71]. We include collusion in the definition of academic plagiarism.

⁷ Note, the five characteristics of plagiarism as defined by Fishman are: (1) the use of words, ideas, or work products (2) attributable to another identifiable person or source, (3) without attributing the work to the source (4) in a situation where there is a legitimate expectation of original authorship (5) in order to obtain some benefit, credit, or gain which need not be monetary ([113], p. 5).

2.1.2 Forms of Academic Plagiarism

Real-world observations of academic plagiarism reveal a variety of commonly found forms.

Literal plagiarism describes the undue copying of text with very little or no disguise.

- ***Copy & paste (c&p)*** is the most common form of literal plagiarism and is characterized by adopting text verbatim from another source [219, 358].
- ***Shake & paste (s&p)*** refers to the copying and merging of text segments with slight adjustments to form a coherent text, e.g., by changing word order, by substituting words with synonyms, or by adding or deleting “filler” words [357].

Disguised plagiarism subsumes practices to conceal unduly copied text [185]. We identified five forms of disguised plagiarism in the literature on plagiarism.

- ***Paraphrasing*** is the intentional rewriting of foreign thoughts in the vocabulary and style of the plagiarist without acknowledging the source [71, 185].
- ***Technical disguise*** refers to techniques that exploit weaknesses of current detection approaches to make plagiarized content non-machine-detectable. Examples include using homoglyphs, symbols that visually appear similar or identical, or inserting random letters in white font [151, 170].
- ***Translated plagiarism*** is the manual or automated conversion of text from one language to another with the intention of hiding its origin [357].

- *Structural and idea plagiarism*⁸ encompasses the use of compositional elements or a broader concept without due acknowledgement of the source. Even if the text is in the author's own words, structural elements, such as outlines or the presentation of ideas or content, such as the chosen research approach, the experimental setup, the lines of argument or the background sources used, may be similar on a level that would have warranted acknowledgement [116, 219]. Inherent in its definition, structural and idea plagiarism is not "obvious" and thus it is not necessarily an indicator that a work is unoriginal or must be retracted. Thus, the term "plagiarism" for structural and idea similarity is justified often only for extreme cases. The presence of structural or idea similarity can rather be a potential quality indicator, e.g., to determine if a work qualifies to be published in a top-journal or a mediocre journal, or if a dissertation meets the highest demands or only satisfies the necessary requirement. We combine structural and idea plagiarism into a single plagiarism form, since it is extremely difficult for human examiners to judge if potential structural plagiarism also copied ideas. Structural and idea plagiarism represent one of the most controversial forms of plagiarism to verify [362], because the decision on whether structural or topical similarities exceed a legitimate level is highly subjective.

⁸ There is no consensus on whether structural and idea plagiarism should be categorized as a form of disguised plagiarism. However, for the definition of disguised plagiarism in this thesis, i.e. forms of plagiarism containing little or no verbatim text overlap and thus not being reliably detectable by PDS, structural and idea plagiarism can reasonably be included in this category. Note that exceptional cases in which structural plagiarism or idea plagiarism also contains paragraphs or sentences copied in their entirety exist; however, this holds true for all plagiarism forms, they do not have to be exclusive.

Self-plagiarism is the partial or complete reuse of one's own writings without such reuse being justified. Presenting updates or providing access to a larger community may justify re-publishing one's own work, but still requires appropriate acknowledgement of the previously published work [40]. Unjustified reasons include trying to artificially increase one's citation count [77].

2.1.3 Prevalence of Plagiarism in the Academic Environment

Academic plagiarism is not a new phenomenon. Since the 1920s, researchers have analyzed the problem, focusing mainly on North American colleges. The following studies give empirical evidence of the problem by providing reviews on academic dishonesty in general [44, 74], collegiate cheating behavior [82, 364] and plagiarism in particular [102, 250].

The majority of studies use self-report surveys to evaluate plagiarism behavior. The most extensive study on U.S. and Canadian campuses questioned around 80,000 students over three years from 2002 to 2005 [220]. McCabe reports 38 % of undergraduates and 25 % of graduate students self-reporting to have paraphrased or copied at least a few sentences without indicating the written source in the 12-month period prior to being questioned [220]. McCabe assumes the true numbers to be higher, because students were more concerned about their anonymity in this web-based assessment compared to earlier paper-based surveys [221, 222]. We agree with this assumption, since self-reports show a tendency to understate misbehavior [284].

The self-report studies often did not distinguish between the different forms of concealed plagiarism or the degree of plagiarism obfuscation. However, for studies indicating the prevalence of specific plagiarism forms, copy & paste and shake & paste plagiarism, a few sentences in length, dominates [176, 220, 222, 223, 273]. Around 20 % of participants admitted to having plagiarized large parts of a document or having obtained texts from fellow students or Internet essay banks [176, 220, 273].

Other studies completed outside of North America that employed plagiarism detection systems consistently found 20 % or more of the inspected documents to contain suspicious content [23, 83, 329]. However, the fraction out of total

plagiarism represented by the detected plagiarism remains unknown. The presented studies only serve as "spotlights" on student plagiarism in different countries. Yet, by reviewing these studies, as well as other extensive research and particular cases observed in the literature [74, 82, 102, 250], we conclude that plagiarism among students is a serious problem.

Assessments of academic dishonesty among post-graduate researchers are rare. One large-scale survey of 2,000 doctoral students and their 4,000 associated faculty members reported that 28 % of faculty members witnessed doctoral students committing plagiarism. Seven percent of doctoral students and 8 % of faculty members reported they had experienced plagiarism by faculty members [324]. Another survey of approximately 3,250 scientists asking about personal misbehavior yielded lower admitted incident rates. Only about 1 % of the respondents self-reported having committed plagiarism. Martinson and Anderson assess these results as “[...] *conservative estimates* [...]” of the true frequency ([215], p. 738). They assume understatements and a response bias from plagiarists who refused to participate.

Fröhlich, Martin and Williams, experts in the field of academic plagiarism, agreed that persons and institutions that discover academic misbehavior often treat such incidences in a clandestine manor. Therefore, only a small fraction of incidences becomes public [116, 214, 366]. The aforementioned experts deduct reasons that substantiate this assumption from known cases of misconduct. *Personal dependence and the fear of retaliation* by the accused, or peers related to the accused, may keep researchers from reporting or publicizing academic misbehavior. *Aversion of engagement* in the laborious and time-consuming inquiry needed for verifying misconduct is another obstacle to reporting. *Fear of losing credibility and scientific reputation* often keeps institutions, including universities, research centers or conferences, from publicizing cases of misconduct or handling them as rigorously as they should.

Despite these obstacles, numerous cases of plagiarism in academia have become public. Price reviews 19 cases of plagiarism, which the U.S. Office of Research Integrity publicized as a result of evaluating medical research projects between 1992 and 2005 [269]. Gutbrodt reports that the IEEE INFOCOM 2006

conference, rejected 12 out of about 1,000 submitted papers after a scan using a PDS revealed suspicious similarities [145].

Sorokina et al. used a self-developed PDS to scan approximately 285,000 texts in the scientific document database arXiv.org [307]. They found more than 500 documents to contain likely cases of plagiarism and approximately 30,000 documents (20 % of the collection) to likely be duplicates or to contain “[...] *excessive self-plagiarism* [...]” ([307], p. 12). Sorokina et al. categorized documents in the excessive self-plagiarism class if their largest contiguous amount of copy-free text was less than 20 % of total document length. As the consequence of a different investigation, arXiv.org deleted 65 articles from 14 different authors for containing substantial plagiarism [15].

The project Déjà Vu [92, 104, 105, 114, 202, 321] used a text similarity scanner [191, 254] to analyze abstracts of bioscience articles in MEDLINE[®] and their full-texts in PubMed Central[®] (PMC) if available. MEDLINE is a bibliographic index and PMC a digital full-text archive [335, 338]. The Déjà Vu project identified 79,383 articles with highly similar abstracts. Manual checks of 4,515 full-texts identified 252 cases of likely plagiarism and 89 likely cases of self-plagiarism [92]. Many reviews presented further plagiarism cases committed in part by renowned senior scholars [69, 116, 214, 313, 361, 366].

Recently, the investigations of two crowd-sourcing projects, the GuttenPlag Wiki and the VroniPlag Wiki exposed plagiarism in the doctoral thesis of former German Federal Minister of Defense and documented 48 cases of plagiarism, respectively⁹ [147, 350]. Some cases in the VroniPlag Wiki involve high-ranking politicians, including the dissertations of members of the German Federal Parliament [348], the European Parliament [64], and the former Vice President of the European Parliament [226]. To date, the responsible universities have verified and retracted the doctorates of nine offenders¹⁰ [350].

⁹ As of 2013-07-04. The VroniPlag Wiki investigations began in March 2011 and are ongoing.

¹⁰ As of 2013-07-04. For a complete and up-to-date listing of retractions visit: <http://de.vroniplag.wikia.com/wiki/Übersicht>

In a similar case, a Hungarian magazine accused Hungary's president, Pál Schmitt, of having committed substantial plagiarism in his doctoral thesis. The responsible university investigated the allegations, confirmed plagiarism on 197 of the 215 pages in the dissertation, and rescinded Schmitt's doctorate [292].

Ironically, even two European ministers of education, were recently found to have plagiarized. The Romanian Minister of Education, Ecaterina Andronescu, was accused of plagiarism and falsification of data in 2012 [163]. The same year in Germany, Annette Schavan, the German Federal Minister of Education and Research was accused of plagiarism in her doctoral thesis. The accusations of Schavan's dissertation sparked a lengthy and heated political debate. The final decision on the presence of plagiarism was made almost a year later, in February 2013, when the Heinrich-Heine University of Düsseldorf rescinded the doctorate by a nearly unanimous vote on the grounds of "willful deceit" [153]. A. Schavan stepped down from her political position but vowed to take the decision to court [309].

We conclude that academic plagiarism is a pressing unsolved problem, also among graduate and post-graduate researchers, although plagiarism research has focused mainly on undergraduate students. Applying automatic detection systems to student assignments is already common practice at many institutions [18]. Scholarly publications, however, are checked far less routinely. By applying string matching to the MEDLINE® database, the Déjà Vu project identified numerous likely cases of plagiarism [104, 114]. Investigations like these can only lead to speculations on the quantity of well-disguised plagiarism in research that goes undetected. Empirical studies on plagiarism frequencies are listed in Appendix H.

The following section describes current plagiarism detection approaches. By pointing out the strengths and weaknesses of existing systems, we find that a substantial number of plagiarism incidences are likely to remain undetected.

2.2 Plagiarism Detection Approaches

This section first gives an overview of the generic mode of operation for all plagiarism detection systems (PDS) and second presents technical descriptions of the detection approaches employed by PDS.

2.2.1 Generic Detection Approach

Plagiarism detection is a hypernym for computer-based approaches, which support the identification of plagiarism [318]. PD is an information retrieval (IR) task supported by specialized IR systems, called plagiarism detection systems (PDS). PDS implement one of two generic detection approaches: *external* or *intrinsic*.

External PDS compare a suspicious document with a reference collection, which is a set of genuine documents [318]. The comparison requires a document model with defined similarity criteria. The task is to retrieve all documents that contain passages that are similar, beyond a chosen threshold, to segments in the suspicious document [319].

Intrinsic PDS statistically examine linguistic features of a text, a process known as *stylometry*, without performing comparisons to other documents. Intrinsic PDS report changes in writing styles as indicators for potential plagiarism [97].

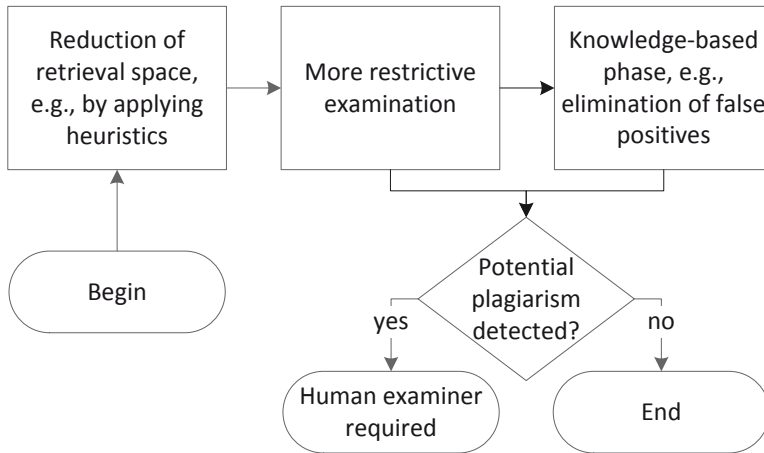


Figure 1: Generic Plagiarism Detection Process

Most external PDS follow a three-stage retrieval process as illustrated in Figure 1. In the first stage, PDS commonly apply computationally inexpensive heuristic document models to reduce the retrieval space. The goal of this stage is to identify a small fraction of the reference collection as candidate documents from which the suspicious text could originate. Coarser fingerprinting (see Section 2.2.3), string matching (see *String Matching*, page 26) or vector space models (see *Vector Space Models*, page 28) are common detection approaches used by PDS for this purpose.

In the second stage, candidate documents retrieved in the first stage undergo a computationally more expensive detailed comparison. PDS usually apply finer-grained variants of the detection approaches we will explain in Sections 2.2.3–2.2.4. PDS can either rely on a single detection approach, or implement a combination of approaches. For example, a PDS may use a coarser fingerprinting method or a vector space model for the initial retrieval stage and a more fine-grained implementation of the same detection approach for the detailed comparison stage. Likewise, a PDS may employ fingerprinting or vector space model-based retrieval for the initial retrieval stage and an elaborate string-matching procedure for the detailed comparison stage.

Citation-based Plagiarism Detection
Detecting Disguised and Cross-language Plagiarism
using Citation Pattern Analysis

Gipp, B.

2014, XXVI, 350 p. 70 illus., Softcover

ISBN: 978-3-658-06393-1