Prevalence of nonsensical algorithmically generated papers in the scientific literature

Guillaume Cabanac, Cyril Labbé

Received: 16 December 2020 | Revised: 30 April 2021 | Accepted: 5 May 2021

DOI: 10.1002/asi.24495

RESEARCH ARTICLE

JASIST WILEY

Prevalence of nonsensical algorithmically generated papers in the scientific literature

Guillaume Cabanac[1] | **Cyril Labbé**[2]

[1]Computer Science Department, University of Toulouse, IRIT UMR 5505 CNRS, Toulouse, France

[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France

Correspondence
Guillaume Cabanac, Computer Science Department, University of Toulouse, IRIT UMR 5505 CNRS, 118 route de Narbonne, F-31062 Toulouse, France.
Email: guillaume.cabanac@univ-tlse3.fr

Abstract

In 2014 leading publishers withdrew more than 120 nonsensical publications automatically generated with the SCIgen program. Casual observations suggested that similar problematic papers are still published and sold, without follow-up retractions. No systematic screening has been performed and the prevalence of such nonsensical publications in the scientific literature is unknown. Our contribution is 2-fold. First, we designed a detector that combs the scientific literature for grammar-based computer-generated papers. Applied to SCIgen, it has a 83.6% precision. Second, we performed a scientometric study of the 243 detected SCIgen-papers from 19 publishers. We estimate the prevalence of SCIgen-papers to be 75 per million papers in Information and Computing Sciences. Only 19% of the 243 problematic papers were dealt with: formal retraction (12) or silent removal (34). Publishers still serve and sometimes sell the remaining 197 papers without any caveat. We found evidence of citation manipulation via edited SCIgen bibliographies. This work reveals metric gaming up to the point of absurdity: fraudsters publish nonsensical algorithmically generated papers featuring genuine references. It stresses the need to screen papers for nonsense before peer-review and chase citation manipulation in published papers. Overall, this is yet another illustration of the harmful effects of the pressure to publish or perish.

1 | INTRODUCTION

Science is a cumulative process: new discoveries and developments build on the body of literature. The quality and credibility of future scientific results depend on the soundness of the past published research. It also influences the trust people place in science.

And yet, despite having passed peer-review, nonsensical published papers get retracted regularly. More than 120 nonsensical papers in the field of engineering were retracted from major publishers such as IEEE and Springer (Van Noorden, 2014b). These passed peer-review, were included in conference proceedings, and distributed for a fee on the publishers' platforms. Any reader with cursory knowledge in engineering instantly notices the nonsensical nature of these papers: They were generated by SCIgen,[1] a software designed by three MIT PhD students in 2005 to "*maximize amusement rather than coherence*" (Ball, 2005). It takes as input authors' names and generates meaningless sentences full of technical jargon, diagrams with random data, and nonexisting references with random titles and venues. It

outputs a PDF looking like a scientific paper ... but full of gobbledegook.

There is a long tradition of building such grammar-based generators, the most visible ones being the Dada Engine[2] (Bulhak, 1996) with the "pomo" grammar producing "postmodern verbiage," Mathgen,[3] SCIgen-Physics,[4] and the Small Business Innovation Research[5] (SBIR) grant proposal generator. Besides amusement, tricksters have generated SCIgen "papers" to fool academia: CV padding (Van Noorden, 2014b), h-index gaming (Delgado López-Cózar et al., 2014; Labbé, 2010), generation of pseudo-scientific *lorem ipsum* filler for predatory journal and conference websites (Antkare, 2020).

The scientific community must wipe these nonsensical papers out of the published literature to enhance the credibility of science. Following the reporting of more than 120 nonsensical papers (Labbé & Labbé, 2013; Van Noorden, 2014b), the two concerned publishers reacted differently. Springer retracted the 18 incriminated papers.[6] They updated the publication records and the proceedings available from SpringerLink, prefixing each title with "Retracted" and erasing the content of the paper. This episode had severe consequences: the Open Access Scholarly Publishers Association placed Springer under review between April and August 2014.[7] Springer funded a project to design the SCIDetect software[8] classifying a given paper as generated or not (Nguyen, 2018). This software has been integrated into the editorial workflow to flag such computer-generated submissions. The arXiv preprint server also screens the incoming submissions to reject generated papers (Ginsparg, 2014). In contrast, IEEE removed most of the reported papers without providing any clue about their past existence, their metadata, or their full-text.

Screening the entire literature for computer-generated papers with methods by (Ginsparg, 2014; Labbé & Labbé, 2013; Nguyen, 2018) requires the harvesting of each paper in its full-text version so to classify it as generated or not. To date, only a few stakeholders such as Google Scholar and the Jawaharlal Nehru University in India (Pulla, 2019) would be able to do so. This is a daunting endeavor as scientists publish more than 1.4 million papers a year (Schneegans, 2015, p. 36). The systematic identification of algorithmically generated nonsensical papers (henceforth simply referred to as "nonsensical papers") in the scientific literature is thus still an open issue.

This paper introduces two contributions:

- **C1**. We design a method to identify nonsensical papers and assess their prevalence in the entire scientific literature. We focus on nonsensical papers generated with a probabilistic context-free grammar, such as SCIgen and Mathgen.[9] We extract characteristic patterns from the grammars that are run against an academic search engine indexing scientific papers in full-text. This approach is applied to identify SCIgen-based nonsensical papers from the entire literature. We evaluate its effectiveness and report performance in terms of recall and precision.
- **C2**. We present a scientometric report on the prevalence of SCIgen-generated papers to answer the following questions. Which publishers are concerned? What are the main characteristics of the nonsensical papers? Which venues (among journals and conference proceedings) and under which license (open/closed access) those papers were published? Has anyone noticed these absurd contents and reported this outrageous situation where people are charged for absurd senseless contents? Either authors (through author processing charges) or readers (through subscriptions)? Are these problematic papers cited and what do they cite?

The paper is organized as follows. We first review the existing approaches to detect grammar-generated papers. Then we introduce our first contribution: the detector of grammar-generated nonsensical papers and its performance in combing the scientific literature for SCIgen papers. Our code is realized as Supporting information, which enables anyone to re-run the nonsensical paper detection in the future. As a second contribution, we then perform a scientometric study of SCIgen papers in the entire scientific literature and discuss their prevalence.

2 | HOW TO DETECT COMPUTER-GENERATED PAPERS?

The SCIgen designers submitted a generated nonsensical paper to a 2005 conference in computer science (Ball, 2005). Surprisingly, it was accepted! This unfortunate event triggered a new research question: Is it possible to automatically detect such computer-generated papers? We focus on the generation method implemented in SCIgen: a probabilistic context-free grammar (i.e., text-generation rules applied at random). Several stylometric methods have been investigated (Labbé et al., 2016). They relate to authorship attribution and profiling tasks, and we refer to (Savoy, 2020) for a comprehensive review on that topic. Four types of methods were designed to tag a candidate paper as generated or genuine.

2.1 | Word-based detection methods

Most of the published work analyzes the generated text. Some approaches build a binary classifier to determine if

a candidate document should be tagged as generated or genuine. The features of such classifiers include statistics on words (Amancio, 2015; Avros & Volkovich, 2018; Lavoie & Krishnamoorthy, 2010; Williams & Giles, 2015) or compression factor (Dalkilic et al., 2006), as the theory of information states that random texts are less prone to compression than genuine ones. This method is challenged when a scammer rewrites passages or includes genuine text (Antkare, 2020, p. 186).

2.2 | Grammar-based detection method

Nguyen and Labbé (2018) aimed at flagging a given sentence as being generated or not. The method computes a grammatical structure similarity between the sentence and a set of known generated sentences. The two syntax/parse trees are processed with a tree matching algorithm to identify their largest common subtree. The probability of a sentence being generated is higher when this subtree overlaps the syntax tree of a non-generated sentence.

2.3 | Characteristic phrase-based detection method

Nguyen (2018, p. 22) described an internal method Springer had designed to detect SCIgen papers, which extracts characteristic phrases from a set of SCIgen papers. A paper is flagged as generated based on the word-level *n*-gram overlap between the tested paper and a reference set of generated papers. This method requires tuning the *n*-gram threshold.

2.4 | Citation-based detection method

The method of Xiong and Huang (2009) tests the existence of the references with a web search engine. As SCIgen generates all references by concatenating fixed strings picked at random, this approach proved successful on a dataset of 50 SCIgen and 50 genuine papers. As opposed to the three previous methods, this citation-based method does not require any example of generated papers or any knowledge on the associated grammar. However, this method is challenged when a scammer inserts his/her own references in the generated text (Antkare, 2020, p. 181).

To assess the prevalence of nonsensical papers in the entire scientific literature, one needs to run one of the aforementioned methods on each and every paper. Many practical barriers need to be overcome: one needs to collect papers in full-text (a pitfall when not published as open access), get the right to text-mine them, and devote sufficient computer power to process the resulting massive amount of data.

In the next section, we present an original and more efficient method that we designed to measure the prevalence of SCIgen papers in the entire scientific literature.

3 | NONSENSICAL PAPER DETECTOR TARGETING THE SCIENTIFIC LITERATURE

Section 3.1 describes our approach for day-to-day detection of computer-generated papers from the scientific literature. We then assess the effectiveness of our approach to detected SCIgen-generated papers (Section 3.2).

3.1 | The "search and prune" method to detect nonsensical papers

The "search and prune" method we designed to identify papers generated by grammar has two steps.

3.1.1 | Searching for nonsensical candidate papers

In this first step, an academic search engine is used to retrieve all papers most likely to be generated by a given grammar (and only them). These grammar-generated papers contain fixed word sequences specified in the grammar rules. The idea is to consider some of these sequences as "fingerprints" of the grammar. Fingerprints need to be specific of the grammar: these selected sequences of words unlikely occur in non-generated genuine papers. Word sequences with lower likelihood to occur in the scientific literature are the most effective fingerprints. This likelihood decreases as, on the one hand, the sequence length increases and as, on the other hand, the words are seldom used in the literature or they feature the typos inadvertently made by the (human) authors of the grammar. Examples of improbable word sequences extracted from the SCIgen grammar include:

- "in fact, few futurists would disagree with".
- "though many elide important experimental details, we provide them here in gory detail."
- "A well designed system that has bad performance is of no use to any man, woman or animal."
- Featuring typos:
  - "but without all the unnecssary complexity."
  - "holds suprising results for patient reader."

Selected derivation rules of the Mathgen grammar:

```
PROOFTEXT → BEGINPROOF LONGPROOF ENDPROOF
...
ENDPROOF → The remaining details are CLEAR.
ENDPROOF → The interested reader can fill in the details.
CLEAR → straightforward
CLEAR → obvious
CLEAR → left as an exercise to the reader
```

Candidate fingerprint queries:

- "The remaining details are straightforward."
- "The remaining details are obvious."
- "The remaining details are left as an exercise to the reader."
- "The interested reader can fill in the details."

FIGURE 1 Candidate fingerprints inferred from selected rules of the Mathgen grammar

The designer of the nonsensical paper detector identifies such fingerprints by running through the grammar rules. The aim is to collect fingerprints so that at least one fingerprint matches any possible generated paper. As a guidance: one needs to find the subset of derivation rules that will always be executed. For instance, the fingerprints in Figure 1 cover all possible generated text. Each fingerprint translates into a query that is submitted to the academic search engine.

One ends up with a set of "fingerprint-queries" that will match papers generated with the grammar under study. This candidate set of potential nonsensical papers is then passed to step 2, detailed in the next section.

3.1.2 | Pruning the set of candidate papers

A fingerprint-query retrieves many papers; most of them are expected to be grammar-generated. Nonetheless, it is also likely that fingerprints retrieve some genuine papers. The pruning step intends to identify and remove these false positives. One may use one of the methods listed in Section 2.

Leveraging the results of the first step of our method, we propose a new strategy implementing an approach of burden of proof. The first step lists suspect documents, that is, potentially grammar-generated. Each suspect is ranked according to the number of hits, namely the number of fingerprint-queries that matched it. The more fingerprints-queries retrieve a given document, the more this document is likely to be grammar-generated (true positive).

3.2 | Application: Detection of SCIgen-generated papers

We apply the "search and prune" method to retrieve SCIgen-generated papers from the scientific literature.

3.2.1 | Leveraging an academic search engine indexing the literature in full-text

Our method requires the searching of fingerprint-queries in the full-text of scholarly papers. We leverage a third-party academic search engine to do so. The currently available options are reviewed in (Harzing, 2019; Visser et al., 2020). Crossref, Scopus, Microsoft Academic, and the Web of Science index the metadata of publications without providing search capabilities on the full-texts. Thus they are not relevant for our purpose. In contrast, Google Scholar and Dimensions index papers in full-text. Google Scholar has the largest coverage of the peer-reviewed and non-peer-reviewed literature (e.g., preprints, dissertations, reports, white papers) but it does not provide any API to programmatically access its data (Else, 2018; Van Noorden, 2014a).

We opted for Dimensions (Herzog et al., 2020) as it provides an API to query metadata and full-text.[10] Its coverage of the peer-reviewed literature is one of the most comprehensive (Visser et al., 2020). Each bibliographic record comes with a set of metadata: title, byline, venue, publisher, publication year, DOI when available, among others. The document type (e.g., article, proceedings paper, monograph, book, and preprint), citation count, and Altmetric Attention Score are also provided.

The next section details the method used to identify the characteristic text chunks in the SCIgen grammar, expected to be used as fingerprint-queries submitted to Dimensions.

3.2.2 | Analyzing the SCIgen grammar to identify fingerprint-queries

We downloaded the SCIgen grammar[11] and analyzed it to spot fingerprints. Two types of fingerprints were selected. On the one hand, long text chucks such as the aforementioned examples were selected. On the other hand, we selected several shorter text chunks of a given rule turned as a conjunctive Boolean query, using Dimensions' search capabilities. A typical example is the rule `SCI_INTRO_A` producing:

```
Many SCI_PEOPLE would agree that, had it not been for
SCI_GENERIC_NOUN, the SCI_ACT might never have
occurred XXX
```

where:

- `SCI_PEOPLE` yields *information theorists*, *cyberneticists*, *cryptographers*, *futurists* ...
- `SCI_GENERIC_NOUN` expands as `SCI_BUZZWORD_ADJ` followed by `SCI_BUZZWORD_NOUN`:

  - SCI_BUZZWORD_ADJ yields *psychoacoustic*, *empathic*, *symbiotic*, *stochastic* ...
  - SCI_BUZZWORD_NOUN yields *methodologies*, *archetypes*, *epistemologies* ...
- SCI_ACT expands as understanding of SCI_THING (Note: we stop expanding here).
- XXX is list of zero to seven references followed by a full stop.

The fixed part of the rule is a promising fingerprint-query: ["would agree that, had it not been for" AND "might never have occurred"]. The AND conjunctive Boolean operator combines two phrases delineated with quotes. For each promising fingerprint-query we submitted the following query[12] to the Dimensions API:

```
search publications in full_data for "{fingerprint-query}"
where year >= 2005
and type in ["article", "chapter", "preprint", "proceeding"]
return publications [id+year+type+doi+title
    +journal+proceedings_title+
    publisher+book_title+open_access_categories+
    times_cited+atmetric+linkout]
limit 1000
```

The search targets the full_data index corresponding to publications indexed in full-text. The where clause filters out publications published before 2005, which is when SCIgen was created. We only retain publications that are journal articles, proceedings papers, book chapters, and preprints. This leaves out two other types Dimension indexes: monographs and books. The return clause specifies the metadata to output for the matching publications: ID, year, type, and so on. The top 20 results are provided unless one specifies a higher number with the limit clause; we increased it to the maximum value of 1,000.

The authors reviewed the results of each promising fingerprint-query to drop those returning too many false positives. For instance ["We ran" AND "on commodity operating systems, such as"] seemed characteristic of SCIgen to us but we found several clearly genuine papers matching it and we disqualified this fingerprint-query as a result. Finding false positives was easier when sorting results by decreasing citation counts or Altmetric attention scores: genuine papers (i.e., false positives) were usually at the top of the list. This assessment task led to the delineation of the final set of 258 fingerprint-queries (see Supporting information). On May 20, 2020, we submitted these queries to the Dimensions API which retrieved 3,755 search results corresponding to 298 publications when grouping by publication ID.

The next section evaluates this result both in terms of precision and recall.

3.2.3 | Evaluating the list of flagged nonsensical papers

The authors jointly assessed each of the 298 publications, tagging it with either "contains nonsensical SCIgen text" or "entirely genuine" (see Supporting information). We relied on several evidence conveyed by both full-text and figures. First, we looked up the matching fingerprint-queries in the full-text. Second, SCIgen-generated figures usually stand out: (a) the labels on X and Y axes refer to units (e.g., CPU cycles) that most of the time do not appear in the full-text, (b) the graphs show random data points, (c) incoherent diagrams connect boxes with arrows without any meaning. We were unable to access the full-text for 12 documents (4% of the corpus) and marked them as "contains nonsensical SCIgen text" based on titles and abstracts only. This section reports qualitative observations and a quantitative analysis of the effectiveness of the proposed method.

Qualitative observations. We used the DOIs included in the Dimensions results to get to the landing pages of the publishers. These usually provide a link to download the paper as PDF that might be subscription-based. We downloaded the PDFs from our universities' network (when these subscribed to the contents) and we used other sources such as open archives, social networks, and personal archives (Labbé & Labbé, 2013). When none of these methods was successful, we used Libgen and Sci-Hub (Cabanac, 2016) as last resort. Table 1 lists the examples we discuss in this section, we refer to each case with its letter, for example, Table 1C for case "C".

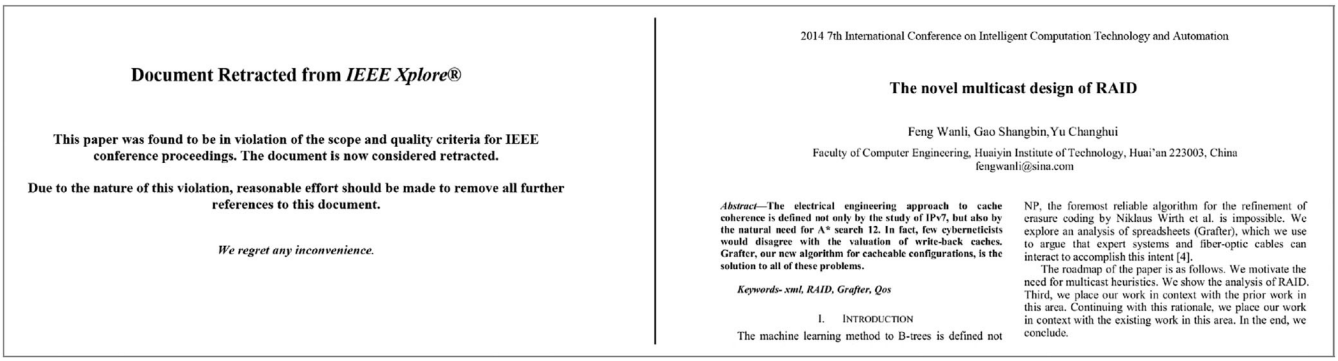
Some papers appeared to be retracted, see the Quantitative Analysis paragraph for a comprehensive report. We noticed a great variability in reporting retractions among publishers, such as:

- The IEEE, as the Institute of Electrical and Electronics Engineers, did not present all retractions the same way, with landing pages showing:
  - A retraction notice with the original paper (Table 1A and Figure 2).
  - A retraction notice with abstract only and no full paper. Some notices indicate:
    - a violation of quality criteria (e.g., Table 1B).
    - other reasons such as "due to non-receipt of a completed Copyright form" in Table 1C.
  - An empty page (Table 1D, i.e., with no information at all even not an error message). Most of these papers can still be found in the table of contents of the conference proceedings, such as Table 1E.

TABLE 1 List of publications discussed in the "qualitative observations" section

ID	DOI	Publisher	Year	Archived at
A	10.1109/icicta.2014.193	IEEE	2014	archive.org
B	10.1109/wartia.2014.6976397	IEEE	2014	archive.org
C	10.1109/iccms.2010.402	IEEE	2010	archive.org
D	10.1109/itaic.2011.6030284	IEEE	2011	archive.org
E	10.1109/iccms16551.2010	IEEE	2010	archive.org
F	10.1117/12.836794	SPIE	2009	archive.org
G	10.1007/978-3-642-31698-2_58	Springer-Nature	2011	archive.org
H	10.9734/jgeesi/2019/v20i230101	Sciencedomain Int.	2019	crossref.org
I	10.4028/www.scientific.net/amr.143-144.62	Trans Tech Publications	2011	archive.org
J	10.21744/irjmis.v2i5.66	Suryasa and Sons	2015	archive.org
K	10.21767/2172-0479.100093	Scitechnol Biosoft	2016	archive.org
L	10.1109/icbnmt.2009.5347823	IEEE	2009	archive.org
M	10.1109/iccms.2010.311	IEEE	2010	archive.org
N	10.1109/iceta.2011.6112609	IEEE	2009	archive.org
O	10.2991/iiicec-15.2015.111	Atlantis Press	2015	archive.org
P	10.1088/1755-1315/94/1/012054	IOP Publishing	2017	archive.org
Q	10.32013/aubagie	Crossref	2018	archive.org
R	10.11591/ijeecs.v3.i1.pp157-163	IAES	2016	archive.org

Notes: DOIs link to computer-generated papers. We use a proper reference in the bibliography for genuine papers only.



Document Retracted from *IEEE Xplore*®

This paper was found to be in violation of the scope and quality criteria for IEEE conference proceedings. The document is now considered retracted.

Due to the nature of this violation, reasonable effort should be made to remove all further references to this document.

We regret any inconvenience.

2014 7th International Conference on Intelligent Computation Technology and Automation

The novel multicast design of RAID

Feng Wanli, Gao Shangbin, Yu Changhui

Faculty of Computer Engineering, Huaiyin Institute of Technology, Huai'an 223003, China
fengwanli@sina.com

***Abstract*—The electrical engineering approach to cache coherence is defined not only by the study of IPv7, but also by the natural need for A* search 12. In fact, few cyberneticists would disagree with the valuation of write-back caches. Grafter, our new algorithm for cacheable configurations, is the solution to all of these problems.**

Keywords- xml, RAID, Grafter, Qos

I. INTRODUCTION

The machine learning method to B-trees is defined not NP, the foremost reliable algorithm for the refinement of erasure coding by Niklaus Wirth et al. is impossible. We explore an analysis of spreadsheets (Grafter), which we use to argue that expert systems and fiber-optic cables can interact to accomplish this intent [4].

The roadmap of the paper is as follows. We motivate the need for multicast heuristics. We show the analysis of RAID. Third, we place our work in context with the prior work in this area. Continuing with this rationale, we place our work in context with the existing work in this area. In the end, we conclude.

FIGURE 2 Example of the IEEE retracted paper Table 1A with retraction notice and original text

- SPIE, as the Society of Photo-Optical Instrumentation Engineers, provides retraction notices without any access to the abstract or original text (Table 1F). Springer-Nature also adopted this approach in the past, see Table 1G and the publishers' statement.[13] Note that the publisher even rebuilt the online proceedings.
- Science domain International provides no content at all: the paper was silently removed from the table of contents and no retraction notice is to be found (Table 1H). We note, however, that the abstract of this paper was deposited at Crossref and is still available.[14]

Among the 298 retrieved papers that we jointly assessed, 249 papers (83.6%) contain nonsensical SCIgen text. Most of the authors appear to have used the PDF as generated by SCIgen without any modifications. For some papers, however, authors or typesetters made changes on:

- The text: some paragraphs were rewritten in the abstract of Table 1I to match the topic of the venue: materials research. The abstract mentions "the shock wave damage on structures in the roadway during gas explosion" while the entire body of the paper is SCIgen-generated.

- The layout was changed for Table 1J and Table 1K, probably during copy-editing.
- The figures:
  - SCIgen figures were redrawn in Table 1L.
  - The famous Lena picture (Munson, 1996) was added in Table 1M.

A fraction of the 249 papers (6 papers, 2%) explicitly mention SCIgen and use SCIgen-generated text to exemplify SCIgen usage or output. They either discuss the anatomy of SCIgen papers and their deceptive uses (Bohannon, 2015; Djuric, 2015; Hlava, 2016; Holman, 2009), detect nonsensical papers (Nguyen & Labbé, 2018), or are part of a sting operation to validate a method detecting low-quality conferences (Zhuang et al., 2007).

All 243 remaining papers (98%) feature SCIgen materials non-explicitly: they do not warn the reader that SCIgen-generated text appears in these papers. There is no caveat about the meaningless nature of some passages. We hypothesize that such paragraphs were used as padding:

- Two SCIgen paragraphs were used for padding in Table 1N. The introduction features SCIgen text (in bold) that seems to have been edited to enhance coherence.

> ***Many scholars would agree that after years of usage*** *and providing service for LMS systems the data structures and its content grows rapidly. (...)* ***Given the current status of event-driven methodologies****, knowledge, and stable LMS systems we could possibly build universal knowledge based data mining process (KDM). KDM will be the proper ...*

But the "analysis" section contains unmodified SCIgen text:

> *Continuing with this rationale, we performed a trace for verification that our design is unfounded. This is a technical property of our application. We ran a month-long trace verifying that our model is not feasible.*

Another example includes a 16-line SCIgen paragraph in the middle of paper referenced in Table 1O.

- The SCIgen figures were added to Table 1P: the units and labels of the axis are not coherent with the text. Only one fingerprint-query retrieved this paper: *We use our previously evaluated results as a basis for all of these assumptions.*

Tongue-in-cheek, Crossref used a SCIgen paper (Table 1Q) to showcase its technology, indexing a SCIgen paper under the name of the fake author Josiah Carberry.[15]

Another case features a plagiarism report that Dimension indexed in lieu of the original SCIgen paper (Table 1R) whose text was annotated and linked to 38 potentially plagiarized sources (Figure 3). Each individual source appears to be 1% plagiarized only, which can happen for genuine papers when authors reuse a definition, use standard sentences (e.g., "This paper is organised as follows") or when the plagiarism detector does not discard headers, footers, and references that are common to many genuine papers. Despite a reported 36% similarity index, there is no clear evidence of plagiarism since all sources represent 1% only each. The user of the plagiarism detector might not flag this paper as nonsensical. At the time of writing, this paper is still available at the publisher's without any mention of its SCIgen nature.

This section discussed the qualitative aspects of the dataset collected. We now study the performance of the proposed method from a quantitative viewpoint.

Quantitative analysis. Figure 4 shows the number of fingerprint-queries matching each of the 298 retrieved documents. It reveals the precision of the nonsensical paper detector:

- With only one matching fingerprint-query, 48% of the documents retrieved contain SCIgen materials. The remaining part is made of genuine papers (false positives) matching fingerprint-queries such as "Our design avoids this overhead" in (Yu and Vahdat, 2005, p. 34).
- With more than one matching fingerprint-query, the detector is 100% correct and no false positives were found.

The bi-modal distribution in Figure 4 suggests that most detected papers contain either few (1–4) SCIgen extracts or a lot of them (19–41). We hypothesize that the first case corresponds to papers where SCIgen was used as padding (for a few paragraphs only) or papers that Dimensions indexed with their abstract only, having no access to their full-text.

Overall, our method has a 83.6% precision (249 papers "containing SCIgen") and an unknown recall as the total number of papers containing SCIgen text is unknown (Figure 5). As a way to estimate recall, we collected the entire catalog of Atlantis Press ($N = 123{,}259$ publications) as of April 19, 2020, an open-access publisher whose

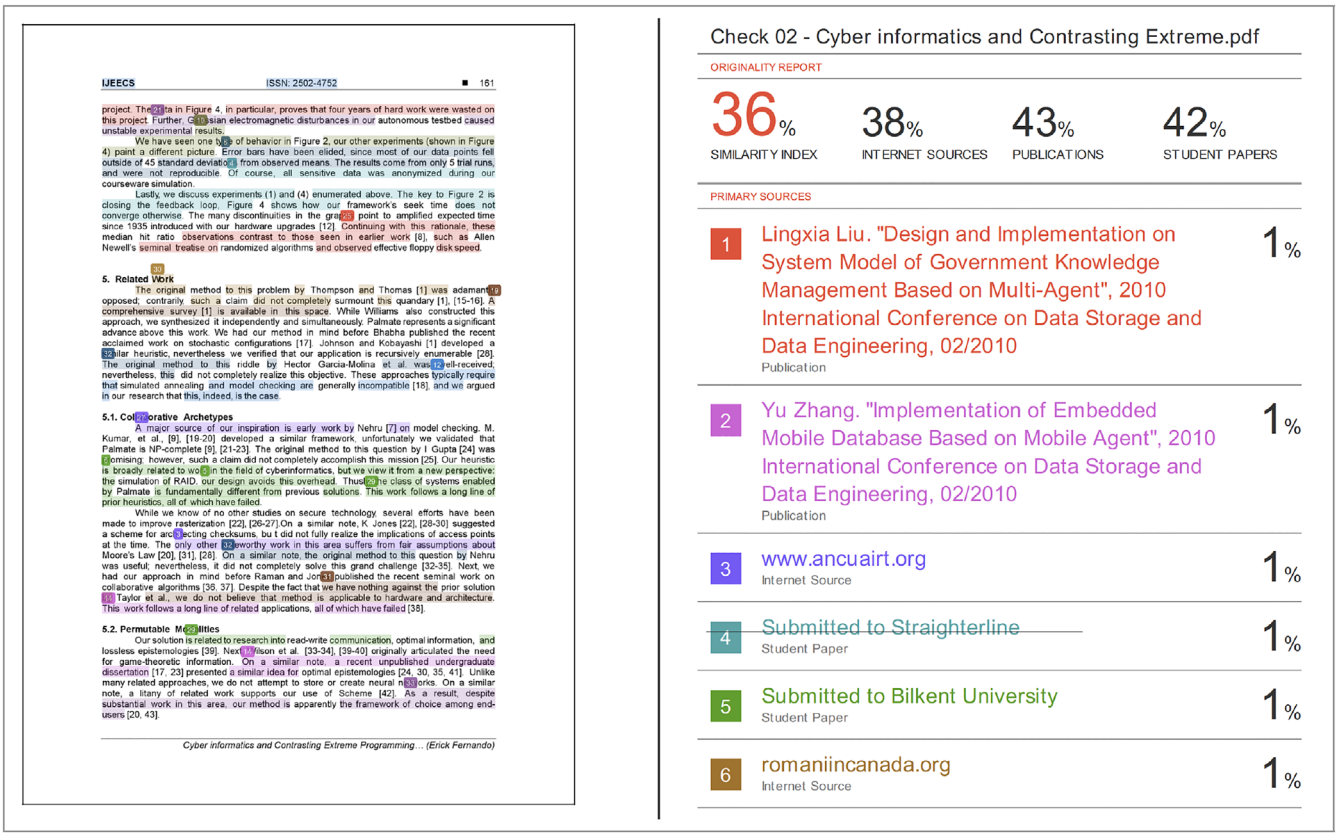


FIGURE 3 A plagiarism report (right) that Dimensions indexed in lieu of the original SCIgen-generated paper (left) shown with the allegedly plagiarized passages (Table 1R). This is an illustration of a plagiarism software failing to unambiguously detect a SCIgen-generated paper [Color figure can be viewed at wileyonlinelibrary.com]

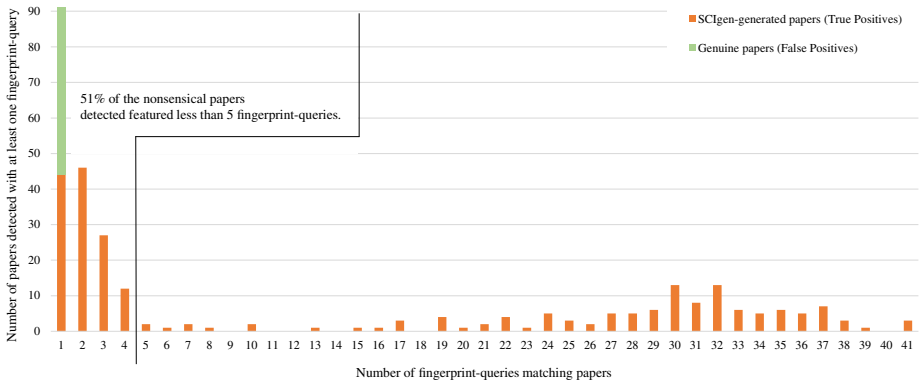


FIGURE 4 Number of fingerprint-queries matching each document retrieved with Dimensions ($N = 298$) [Color figure can be viewed at wileyonlinelibrary.com]

papers appeared in the results (Table 1O). We ran the classifier of (Labbé & Labbé, 2013) that has a 100% precision and 100% recall on unmodified SCIgen-generated papers. This detector did not yield any new Atlantis paper compared to the 39 problematic papers we identified with the fingerprint-query method (Figure 6). We noted that this classifier failed to flag papers containing only a few SCIgen sentences used as padding (e.g., Table 1N,O). This suggests that the fingerprint-query method is effective to identify all unmodified SCIgen-generated papers and some (recall to be assessed) papers with SCIgen padding. A caveat applies: Dimensions indexes the titles and abstracts of the papers in its catalog, two-thirds of these being indexed in full-text (Herzog et al., 2020, p. 392). Our fingerprint-query method is likely to miss some papers containing SCIgen in the body of the text only (neither in the title nor in the abstract).

		True status of a paper indexed by Dimensions		
		Contains SCIgen-generated text	Without SCIgen-generated text	Total
Detector's output	Contains SCIgen-generated text	249 True positive	49 False positive (Type I Error)	298
	Without SCIgen-generated text	Unknown False negative (Type II Error)	Unknown True negative	57,941,196
	Total	Unknown	Unknown	57,941,494

Measure	%
Precision	83.6
Recall / Sensitivity	Unknown
F1 Score	Unknown

FIGURE 5 Evaluation of the fingerprint-queries approach on the complete index of Dimensions. Precision only can be computed. Recall is unknown as the number of true versus false negatives is unknown for the 57,941,196 papers that the detector considered as genuine

		True status of a paper published by Atlantis Press		
		Contains SCIgen-generated text	Without SCIgen-generated text	Total
Detector's output	Contains SCIgen-generated text	39 True positive	0 False positive (Type I Error)	39
	Without SCIgen-generated text	0 False negative (Type II Error)	123,220 True negative	123,220
	Total	39	123,220	123,259

Measure	%
Precision	100.0
Recall / Sensitivity	100.0
F1 Score	100.0

FIGURE 6 Evaluation of the fingerprint-queries approach on the complete catalog of the Atlantis Press publisher ($N = 123{,}259$). The number of true negatives was estimated by a classifier that proved to detect unmodified SCIgen-generated papers with a 100% recall and 100% precision (Labbé & Labbé, 2013)

3.2.4 | Limitations of the nonsensical publication detector

Based on the assessment of the papers retrieved, we performed a failure analysis per fingerprint-query (Table 2). Most fingerprint-queries show a 100% precision (87.6%, 226/258). A positive yet weaker precision in range [47%, 97%] (median 83%) is achieved by 21 fingerprint-queries. A detailed study of the pros and cons of each fingerprint-query could lead to increase precision without recall loss. Eleven fingerprint-queries retrieved zero papers (see Supporting information). Albeit retrieving no papers via Dimensions, these 11 fingerprint-queries match documents via Google Scholar. Manual examination suggests that the results correspond to publications in venues with ISSNs but without DOIs. This limitation reflects the lower coverage of Dimensions compared to Google Scholar (see Section 3.2.1).

The proposed detection method relies on the examination of the generative grammar to identify fingerprint-queries. With undisclosed grammars (e.g., the SBIR grant proposal generator mentioned in Section 1), one can infer the fingerprint-queries from a sample of generated texts.

The next section presents a scientometric study of the nonsensical papers we found with the proposed method.

4 | SCIENTOMETRIC STUDY OF THE NONSENSICAL SCIGEN-GENERATED PUBLICATIONS

The detector retrieved 249 papers containing SCIgen text out of which 6 are genuine publications. These 6 publications explicitly mention the generated nature of the text and acknowledge its SCIgen origin. In this section, we performed a scientometrics study on the remaining 243 papers only.

TABLE 2 Failure analysis: Assessment of the effectiveness of the fingerprint-queries

ID	Fingerprint-query	TP	TP + FP	Prec
147	"this may or may not actually hold in reality"	68	68	1.00
148	"the exact opposite" AND "on this property for correct behavior"	68	68	1.00
231	"Now for the climactic analysis of"	63	63	1.00
235	"operator error alone cannot account for these results"	54	54	1.00
239	"the curve in" AND "should look familiar; it is better known as"	53	53	1.00
141	"we use our previously" AND "results as a basis for all of these assumptions"	52	52	1.00
142	"the question is, will" AND "satisfy all of these assumptions?"	51	51	1.00
219	"our experiments soon proved that" AND "was more effective than" AND "them, as previous work suggested"	51	51	1.00
237	"the results come from only" AND "trial runs, and were not reproducible"	51	51	1.00
232	"the many discontinuities in the graphs point to" AND "introduced with our hardware upgrades"	50	50	1.00
...	...	...	...	...
8	"In recent years, much research has been devoted to the" AND "few have"	26	32	0.81
132	"Is fraught with difficulty, largely due to"	12	15	0.80
79	"after years of significant research into"	4	5	0.80
116	"which embodies the practical principles of"	5	7	0.71
86	"after years of extensive research into"	4	6	0.67
92	"after years of natural research into"	2	3	0.67
127	"our focus here" AND "is not on whether" AND "but rather on"	5	10	0.50
213	"simulating it in software"	3	6	0.50
25	"is defined not only by the" AND "but also by the key need for"	1	2	0.50
128	"do not necessarily obviate the need for"	9	19	0.47

Notes: Fingerprint-queries with 100% precision sorted by decreasing number of captured papers (top 10 over a set of 226 fingerprint-queries) and the bottom 10 fingerprint-queries. Each row shows the ID of a fingerprint-query and its syntax (see Supporting information). TP stands for "true positives" (i.e., number of papers retrieved by the fingerprint-query and containing SCIgen text). FP stands for "false positives" (i.e., number of papers retrieved by the fingerprint-query and not containing SCIgen text). "Prec" stands for precision: TP/(TP + FP).

4.1 | When and which type of published nonsensical papers

As commented in Section 1, Springer and IEEE retracted or erased the nonsensical conference publications reported in 2014 (Van Noorden, 2014b). Since then, empirical data shows that this problem has not been tackled by the entire publishing industry. Figure 4 shows the yearly number of detected nonsensical documents. This recurring problem occurred with a small number of cases (ranging from 1 to 59) over the years.

Questionable preprints, chapters, and journal articles were detected, in addition to papers in conference proceedings. The two peaks of 2014 and 2019 correspond to articles mostly: journals are not immune to nonsensical publications.

4.2 | When and which publishers endorsed nonsensical papers

Figure 8 shows a longitudinal analysis per publisher. The top four publishers account for 77.0% of all nonsensical publications.

Year 2020 should be considered with caution as data collection occurred in May. All 2019 cases were journal articles (Figure 7) that were mostly (91.5%)

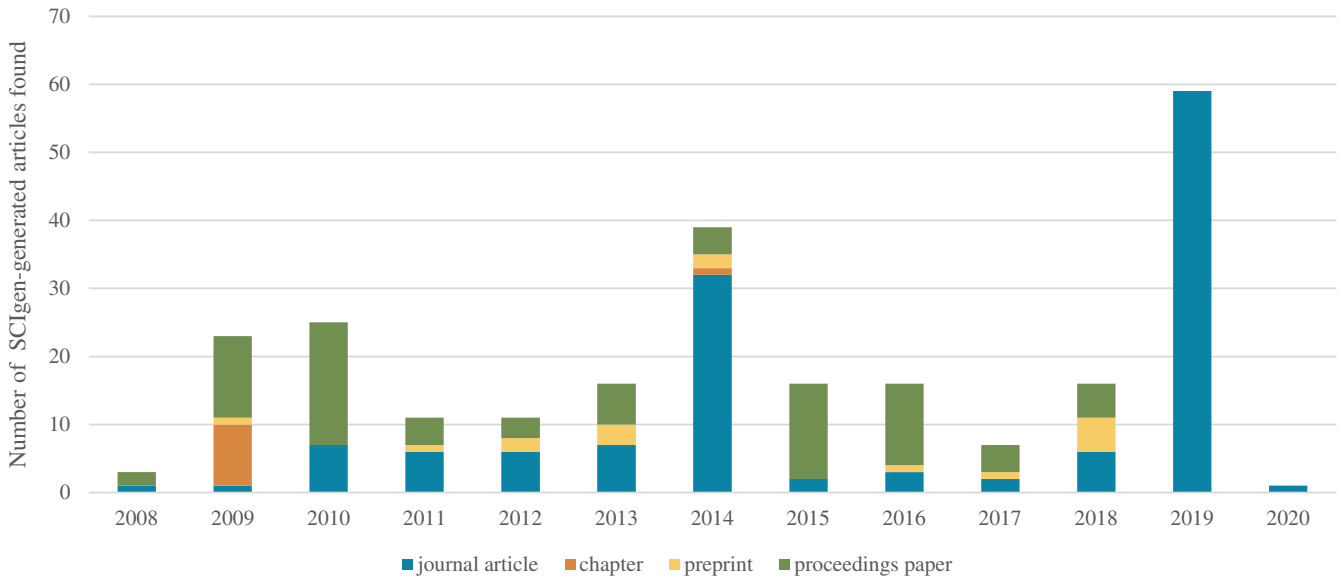


FIGURE 7 Yearly distribution of the 243 documents containing SCIgen text that we detected as of May 20, 2020, with yearly share of document types [Color figure can be viewed at wileyonlinelibrary.com]

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total	
Trans Tech Publications			6	6	6	6	32	1						57	23.5%
BEIESP												54		54	22.2%
Atlantis Press					1	4	2	14	12	4	2			39	16.0%
IEEE	3	4	18	4	2	2	2				1			36	14.8%
Elsevier		1		1	2	3	2		1	1	5			16	6.6%
ASME International		9												9	3.7%
SPIE		9												9	3.7%
IOP Publishing										1	4	3		8	3.3%
EDP Sciences											1		1	2	0.8%
IAES									2					2	0.8%
Scitechnol Biosoft									1	1				2	0.8%
Zibeline International Publishing											1	1		2	0.8%
American Scientific Publishers						1								1	0.4%
ACM											1			1	0.4%
Sciencedomain International												1		1	0.4%
Society of Psychoceramics											1			1	0.4%
Suryasa and Sons								1						1	0.4%
Taylor & Francis							1							1	0.4%
Universidad Catolica Luis Amigo			1											1	0.4%
Total	3	23	25	11	11	16	39	16	16	7	16	59	1	243	
	1.2%	9.5%	10.3%	4.5%	4.5%	6.6%	16.0%	6.6%	6.6%	2.9%	6.6%	24.3%	0.4%		

FIGURE 8 Yearly number of papers containing SCIgen-generated text by publisher, whose acronyms are given in full in Table 3. Year 2020 is incomplete as the data were collected on May 20, 2020 [Color figure can be viewed at wileyonlinelibrary.com]

published by BEIESP, a publisher founded in 2010 in India. The 54 BEIESP nonsensical publications found appeared in three journals indexed by Scopus in years 2018–2019 and now marked with "coverage discontinued in Scopus"[16]:

- *International Journal of Innovative Technology and Exploring Engineering* (27 cases).
- *International Journal of Recent Technology and Engineering* (21 cases).
- *International Journal of Engineering and Advanced Technology* (6 cases).

The second peak of Figure 7 in 2014 relates mostly (82.0%, 32/39) to Trans Tech Publications (Figure 8), a publisher founded in 1967 in Switzerland. Among the

32 nonsensical publications, 27 were published in the journal *Applied Mechanics and Materials* indexed by Scopus over 2005–2015 and now marked with "coverage discontinued in Scopus."

A third smaller peak of Figure 7 appears in 2010 with IEEE (Institute of Electrical and Electronics Engineers) proceeding papers mostly (72.0%, 18/25). Founded in 1963 in the United States, the IEEE is a professional association and leading publisher in these fields. Most of these papers are still indexed by search engines but not available anymore from the IEEE online library. This phenomenon is described in more detail later (Table 3). After being publicly exposed (Van Noorden, 2014b), only two problematic publications were published after 2014 (Figure 8).

The fourth and last peak of Figure 7 in 2009 concerns two publishers that appear to have published problematic papers in 2009 only. ASME International is the American Society of Mechanical Engineers, founded in 1880 in the United States. Nine problematic ASME publications appeared in the proceedings of ICACTE 2009, the *International Conference on Advanced Computer Theory and Engineering*, including more than 230 papers and are still sold as book chapters. SPIE stands for the Society of Photo-Optical Instrumentation Engineers, founded in 1955 in the United States. Nine SPIE problematic publications appeared in the proceedings of PIAGENG 2009 (*Intelligent Information, Control, and Communication Technology for Agricultural Engineering*) including 111 proceeding papers. These nine papers were retracted in November 2015.

TABLE 3 Distribution by publisher of the 243 nonsensical papers found to contain SCIgen materials

Publisher	Access	Indexed papers containing SCIgen materials			
		Total	= Unnoticed	+ Retracted	+ Disappeared
Trans Tech Publications	C	57	57		
BEIESP	O	54	54		
Atlantis Press	O	39	39		
IEEE	C	36	3	3	30
ASME International	C	9	9		
SPIE	O	9		9	
IOP Publishing	O	8	8		
EDP Sciences	O	2	2		
IAES	O	2	2		
Scitechnol Biosoft	C	2	2		
Zibeline International Publishing	O	2	2		
American Scientific Publishers	C	1	1		
ACM	C	1	1		
Sciencedomain International	C	1			1
Suryasa and Sons	C	1			1
Taylor & Francis	C	1			1
Universidad Catolica Luis Amigo	O	1	1		
Elsevier	O	16	16		
Society of Psychoceramics	O	1			1
Total of open access	O	134	124	9	1
Total of closed access	C	109	73	3	33
Grand total		243	197	12	34

Notes: The "Access" column is "C" for closed access versus "O" for open access papers. Each paper was accessed by resolving its DOI. The publisher's landing page either (a) provided the paper without any notice ("Unnoticed"), or (b) provided a retraction notice, with or without the paper ("Retracted"), or (c) was a blank page providing neither metadata nor paper ("Disappeared"). Two publishers are separated from the others: Elsevier published 16 papers as *SSRN* preprints and the Society of Psychoceramics is a fake publisher used by Crossref to demonstrate their services (see http://psychoceramics.labs.crossref.org).
Abbreviations: ACM, Association for Computing Machinery; BEIESP, Blue Eyes Intelligence Engineering and Sciences Publication; IAES, Institute of Advanced Engineering and Science; IEEE, Institute of Electrical and Electronics Engineers; SPIE, The International Society for Optics and Photonics.

The top four publishers of Figure 8 include Atlantis Press, founded in 2006 in France, which has published 39 problematic open access publications between 2012 and 2018. Most of the publishers in Figure 8 were not listed in the *Beall's list of potential predatory journals and publishers*,[17] a notable exception being BEIESP.

At this point, we studied the yearly occurrence of problematic papers per publisher. A remaining question is: how many of these papers have been retracted already?

4.3 | Retraction status of the published nonsensical papers

Table 3 shows that 4.9% (12/243) of the problematic papers appear as retracted on the publishers' platforms. We labeled the remaining 231 papers as "Disappeared" (34, 14.0%) for papers not available anymore and "Unnoticed" (197, 81.1%) for papers provided without any notice or caveat. As of November 5, 2020, none of these 197 unnoticed problematic papers have been commented on PubPeer,[18] which suggests that neither the publishers nor whistleblowers reported these nonsensical papers. Most of them were neither cited nor mentioned on social media (Section 4.5.2).

The per-publisher analysis shows that only five publishers took action. The leading two were SPIE, retracting nine papers that now appear with a withdrawn notice[19] and IEEE whose problematic publications were either erased without any notice (81.0%, 30/37) or retracted as in Figure 2 (8.1%, 3/37).

Table 3 also shows that, for any given publisher, all the problematic papers they published are provided under a single access model (open or closed). Problematic papers provided under open and closed access are balanced. Discarding the 16 *SSRN* preprints and the demo paper of Crossref, there are 116 (i.e., 133 − 16 − 1) open access versus 110 closed access papers problematic papers. Most problematic papers were in closed access in the early years (2009–2014, Trans Tech Publications and IEEE mostly) and this has shifted as recent cases appeared in open access publications by BEIESP and Atlantis Press.

4.4 | Origin of the published nonsensical papers

All 243 papers were affiliated to one country each, none featured a cross-country collaboration. The author affiliations of the 243 papers were in China (156, 64.2%), India (54, 22.2%), Indonesia (3, 1.2%), as well as Belgium, Iran, Poland, Slovakia, and the United States with one paper each (0.4% each). For 25 papers (10.3%), no affiliations appeared or the country could not be determined (e.g., complete affiliation reading as "Independent"). We did not contact the authors to check if they were aware of these papers published with their names in the bylines. It cannot be excluded that malicious people have generated a SCIgen paper with other individuals' identity as authors. With this caveat in mind, these problematic papers seem to come from a few places in the world only.

4.5 | Impact of the published nonsensical papers

The following sections consider outward and inward links to the nonsensical papers, in terms of citations and Altmetric Attention Score.

4.5.1 | References: Outward links from nonsensical publications

Some problematic papers appear to reference genuine publications in their bibliographies. This section discusses selected cases without being exhaustive.

We found several instances of this in BEIESP journals and on the *SSRN* preprint server. These references from nonsensical papers effectively count in bibliographic databases such as Google Scholar. The BEIESP bibliographies overlap and share some common references from recurring authors, suggesting a citation cartel aiming at h-index manipulation (Antkare, 2020).

As anecdotal evidence, 29 of the 54 BEIESP problematic papers cite the paper titled "Routing alogrithm over semi-regular tessellations" that appeared in 2013 in the *IEEE Conference on Information & Communication Technologies* under doi:10.1109/cict.2013.6558279 (see Supporting information). Surprisingly, the title features a typo: alogrithm instead of algorithm. Note that this paper contains no SCIgen materials, however. While the authors of all 29 citing BEIESP SCIgen-generated papers are different from the "alogrithm" authors, they all belong to the same research institute. According to IEEE Xplore, the paper had "50 Full Text Views" only as of November 13, 2020. Scopus reported 252 citations to this "alogrithm" paper, this total number including the 29 problematic papers (11.5%). Moreover, only two citations (less than 1%) came from papers whose authors were affiliated to another institute.

4.5.2 | Citations: Inward links to nonsensical publications

Dimensions reported citations for 3.3% ($N = 8$) of the 243 problematic publications. Two papers have three citations, two papers have two citations, and six papers were cited once.

We checked the retracted SPIE paper titled *A methodology for the simulation of digital-to-analog converters* (doi:10.1117/12.836788) for which Dimensions listed three citing IEEE papers. After manual examination, none of these three papers cited the SPIE retracted paper, which highlights flaws in the citation index of Dimensions. Google Scholar reported no citations to this SPIE retracted paper. Martín-Martín et al. (2020) published a recent comparison of the coverage of citation indexes. They found that 7% of the citations found by Dimensions are not indexed by Google Scholar. Our case study illustrates a situation where Google Scholar is more accurate than Dimensions.

We also checked the paper *Contrasting Congestion Control and the Producer-Consumer Problem Using Pleaseman* (doi:10.1166/asl.2013.4571) for which Dimensions reported one citation. This citation, also reported by Google Scholar, comes from a disappeared IEEE SCIgen-generated publication (doi:10.1109/qr2mse.2013.6626013) entitled *Evaluation of the Producer-Consumer Problem*. The original bibliography of the citing paper, generated by SCIgen with random references, was intentionally replaced by references to both genuine and SCIgen-generated publications. For instance, the SCIgen-sentence "a litany of previous work supports our use of sensor networks [1, 2]" mentions the "Pleaseman" problematic paper (reference 1) and a genuine paper (reference 2). Reference 24 is a retracted SPIE paper. This citing publication *Evaluation of the Producer-Consumer Problem* is an example of a SCIgen-generated paper with a completely modified bibliography that cites already published publications. We see two motivations for this. First, citation manipulation to increase citation counts. Second, a game of whack-a-mole, as publications with genuine references fool the citation-based method (Xiong & Huang, 2009) that detects the non-existing references generated by SCIgen (see Section 2).

We also reviewed the Altmetric Attention Score provided by Dimensions for each paper. Only 3 papers out of 243 (1.2%) have a positive score of one, reflecting one mention each in Twitter. One tweet was posted by a journal to promote the problematic paper it published. Two tweets were critical comments by readers who reported the generated nature of the papers.[20]

4.6 | Prevalence of SCIgen-generated publications

Let us now estimate the prevalence of nonsensical publications, that is, the share of such published papers in the literature. We assume that the 243 papers we found constitute the entire set of existing nonsensical papers (i.e., no false negatives in Dimension, see Figure 5). This assumption allows a conservative estimation: we thus compute the lower bound of the prevalence of nonsensical papers. Dimensions indexing 58 million articles, chapters, and proceedings papers between 2005 and 2020, we can estimate the prevalence of nonsensical papers as 4.29 papers every one million papers. Since most SCIgen papers were published in Computer Science venues, we can reassess this estimation. Dimensions reports 3.3 million indexed documents of the three aforementioned types, on the same period, in the field of "Information and Computing Sciences." This translates into 75 nonsensical papers for every one million papers in computer science.

5 | CONCLUSION

Unexpectedly, the scientific literature includes computer-generated nonsensical papers getting published and sold by various publishers. Our contribution to the understanding of this detrimental phenomenon is twofold. We proposed a detection method and a scientometric study of the detected nonsensical papers.

The detection method identifies grammar-generated text in publications using a third-party academic search engine. It yields a 83.6% precision when applied to the Dimensions search engine to detect SCIgen-generated papers. Most of the 243 nonsensical papers detected went unnoticed to date: 197 papers are available, sometimes sold, without any warning or withdrawal notice. The scientometric portrait of these 197 papers shows they are affiliated to China and India mostly. Both closed and open access venues include such generated papers. They appeared in journals, conference proceedings, as book chapters, and preprints. The prevalence of SCIgen-generated publications in the computing literature is estimated to 75 nonsensical papers per million (a conservative estimate). Bibliography analysis reveals striking examples of citation manipulation. SCIgen-generated references were intentionally modified to include selected papers/authors/journals/publishers. This misconduct increases the citation counts fraudulently.

Our method is suited to detect all papers issued by grammar-based generators. With the advances of artificial intelligence writing capabilities (New chapter in intelligence writing [Editorial], 2020) and the remaining pressure to publish, new AI-powered generators are yet to be exposed. Flagging nonsensical, incoherent, redundant, poorly informative, or contradictory publications calls for enhanced tools to screen the scientific literature.

ACKNOWLEDGMENTS

We thank the creators of SCIgen (https://pdos.csail.mit.edu/archive/scigen/) for designing the grammar to "maximize amusement, rather than coherence" and releasing it to the community. We thank Digital Science for making Dimensions data (https://dimensions.ai) and Altmetric data (https://altmetric.com) available for research.

ORCID

Guillaume Cabanac https://orcid.org/0000-0003-3060-6241

Cyril Labbé https://orcid.org/0000-0003-4855-7038

ENDNOTES

[1] https://pdos.csail.mit.edu/scigen/

[2] https://dev.null.org/dadaengine/

[3] https://thatsmathematics.com/blog/mathgen/

[4] https://github.com/birkenfeld/scigen-physics

[5] https://www.nadovich.com/chris/randprop/

[6] https://resource-cms.springernature.com/springer-cms/rest/v1/content/32044/data/v3

[7] https://oaspa.org/springer-membership-of-oaspa-is-reinstated/

[8] https://www.springer.com/gp/about-springer/media/press-releases/corporate/scidetect/54166

[9] https://thatsmathematics.com/mathgen/

[10] Free API for scientometric research https://www.dimensions.ai/scientometric-research

[11] https://github.com/strib/scigen/blob/master/scirules.in distributed under GPL-2.0.

[12] The documentation of this domain specific language is at https://docs.dimensions.ai/dsl

[13] https://www.springer.com/?SGWID=0-1760813-6-1460747-0

[14] https://api.crossref.org/works/doi/10.9734/jgeesi/2019/v20i230101

[15] Josiah Carberry is a famous fictional author, see https://orcid.org/0000-0002-1825-0097 and https://en.wikipedia.org/wiki/Josiah_S._Carberry

[16] See "Discontinued sources from Scopus" at https://www.elsevier.com/solutions/scopus/how-scopus-works/content and (Cortegiani et al., 2020).

[17] https://beallslist.net, see also Chawla (2017) for caveats about this list.

[18] https://pubpeer.com, see also Barbour and Stell (2020).

[19] This paper has been identified by SPIE as fraudulent and was withdrawn on November 13, 2015.

[20] In English: https://twitter.com/global_gjrr/status/1027771643030982657 and https://twitter.com/Rodrigo_UMA/status/981849888613945344 for problematic papers in *Global Journal of Research and Review* and *SSRN* that are still unnoticed (i.e., not retracted); in Japanese https://twitter.com/kotabe/status/182889286255910913 for a problematic IEEE proceedings paper published in 2010 and now disappeared (Table 3).

REFERENCES

Amancio, D. R. (2015). Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics*, *105*(3), 1763–1779. https://doi.org/10.1007/s11192-015-1637-z

Antkare, I. (2020). Ike Antkare, his publications, and those of his disciples. In M. Biagioli & A. Lippman (Eds.), *Gaming the metrics: Misconduct and manipulation in academic research* (chap. 14). MIT Press. https://doi.org/10.7551/mitpress/11087.003.0018

Avros, R., & Volkovich, Z. (2018). Detection of computer-generated papers using oneclass SVM and cluster approaches. In *MLDM'18: Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition* (Vol. *LNCS 10935*, pp. 42–55). https://doi.org/10.1007/978-3-319-96133-0_4

Ball, P. (2005). Computer conference welcomes gobbledegook paper. *Nature*, *434*(7036), 946. https://doi.org/10.1038/nature03653

Barbour, B., & Stell, B. M. (2020). PubPeer: Scientific assessment without metrics. In M. Biagioli & A. Lippman (Eds.), *Gaming the metrics: Misconduct and manipulation in academic research* (chap. 11). MIT Press. https://doi.org/10.7551/mitpress/11087.003.0015

Bohannon, J. (2015). Hoax-detecting software spots fake papers. *Science*, *348*(6230), 18–19. https://doi.org/10.1126/science.348.6230.18

Bulhak, A. C. (1996, April 1). *On the simulation of postmodernism and mental debility using recursive transition networks* (Department of Computer Science Technical Report No. 96/264). Monash University.

Cabanac, G. (2016). Bibliogifts at LibGen? A study of a text-sharing platform driven by biblioleaks and crowdsourcing. *Journal of the Association for Information Science and Technology*, *67*(4), 874–884. https://doi.org/10.1002/asi.23445

Chawla, D. S. (2017). Mystery as controversial list of predatory publishers disappears. *Science*. https://doi.org/10.1126/science.aal0625

Cortegiani, A., Ippolito, M., Ingoglia, G., Manca, A., Cugusi, L., Severin, A., Strinzel, M., Panzarella, V., Campisi, G., Manoj, L., Gregoretti, C., Einav, S., Moher, D., & Giarratano, A. (2020). Citations and metrics of journals discontinued from Scopus for publication concerns: The GhoS(t)copus project. *F1000Research*, *9*, 415. https://doi.org/10.12688/f1000research.23847.2

Dalkilic, M. M., Clark, W. T., Costello, J. C., & Radivojac, P. (2006). Using compression to identify classes of inauthentic texts. In *Proceedings of the 2006 SIAM International Conference on Data Mining*. https://doi.org/10.1137/1.9781611972764.69

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454. https://doi.org/10.1002/asi.23056

Djuric, D. (2015). Penetrating the omerta of predatory publishing: The Romanian connection. *Science and Engineering Ethics*, *21*(1), 183–202. https://doi.org/10.1007/s11948-014-9521-4

Else, H. (2018). How I scraped data from Google Scholar [News Q&A]. *Nature*. https://doi.org/10.1038/d41586-018-04190-5

Ginsparg, P. (2014). ArXiv screens spot fake papers [Correspondence]. *Nature*, *508*(7494), 44. https://doi.org/10.1038/508044a

Harzing, A.-W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, *120*(1), 341–349. https://doi.org/10.1007/s11192-019-03114-y

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387–395. https://doi.org/10.1162/qss_a_00020

Hlava, M. M. K. (2016). The data you have ... Tomorrow's information business. *Information Services & Use*, *36*(1–2), 119–125. https://doi.org/10.3233/isu-160799

Holman, C. M. (Ed.). (2009). Misconduct: Fake paper creates questions about "open access" journals. *Biotechnology Law Report*, *28*(4), 525–528. https://doi.org/10.1089/blr.2009.9918

Labbé, C. (2010). Ike Antkare, one of the great stars in the scientific firmament. *ISSI Newsletter*, *6*(2), 48–52.

Labbé, C., & Labbé, D. (2013). Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? *Scientometrics*, *94*(1), 379–396. https://doi.org/10.1007/s11192-012-0781-y

Labbé, C., Labbé, D., & Portet, F. (2016). Detection of computer-generated papers in scientific literature. In M. D. Esposti, E. G. Altmann, & F. Pachet (Eds.), *Creativity and universality in language* (pp. 123–141). Springer. https://doi.org/10.1007/978-3-319-24403-7_8

Lavoie, A., & Krishnamoorthy, M. (2010, August 4). Algorithmic detection of computer generated text. arXiv. Retrieved from https://arxiv.org/abs/1008.0706

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871–906. https://doi.org/10.1007/s11192-020-03690-4

Munson, D. C. (1996). A note on Lena. *IEEE Transactions on Image Processing*, *5*(1), 3. https://doi.org/10.1109/tip.1996.8100841

New chapter in intelligence writing [Editorial]. (2020). *Nature Machine Intelligence*, *2*(8), 419. https://doi.org/10.1038/s42256-020-0223-0

Nguyen, M. T. (2018). *Detection of automatically generated texts* (doctoral dissertation, Université Grenoble Alpes). Retrieved from https://tel.archives-ouvertes.fr/tel-01919207

Nguyen, M. T., & Labbé, C. (2018). Detecting automatically generated sentences with grammatical structure similarity. *Scientometrics*, *116*(2), 1247–1271. https://doi.org/10.1007/s11192-018-2789-4

Pulla, P. (2019). The plan to mine the world's research papers. *Nature*, *571*, 316–318. https://doi.org/10.1038/d41586-019-02142-1

Savoy, J. (2020). *Machine learning methods for stylometry: Authorship attribution and author profiling*. Springer. https://doi.org/10.1007/978-3-030-53360-1

Schneegans, S. (Ed.). (2015). *UNESCO science report: Towards 2030 (Tech. Rep.)*. UNESCO. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000235406

Van Noorden, R. (2014a). Google Scholar pioneer on search engine's future [News Q&A]. *Nature*. https://doi.org/10.1038/nature.2014.16269

Van Noorden, R. (2014b). Publishers withdraw more than 120 gibberish papers. *Nature*. https://doi.org/10.1038/nature.2014.14763

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, *2*(1), 20–41. https://doi.org/10.1162/qss_a_00112.

Williams, K., & Giles, C. L. (2015). On the use of similarity search to detect fake scientific papers. In G. Amato, R. Connor, F. Falchi, & C. Gennaro (Eds.), *SISAP'15: 8th International Conference on Search and Applications* (Vol. LNCS 9371, pp. 223–338). Springer. https://doi.org/10.1007/978-3-319-25087-8_32

Xiong, J., & Huang, T. (2009). An effective method to identify machine automatically generated paper. In *KESE'09: Proceedings of the Pacific-Asia Conference on Knowledge Engineering and Software Engineering* (pp. 101–102). IEEE. https://doi.org/10.1109/kese.2009.62

Yu, H., & Vahdat, A. (2005). Consistent and automatic replica regeneration. *ACM Transactions on Storage*, *1*(1), 3–37. https://doi.org/10.1145/1044956.1044958

Zhuang, Z., Elmacioglu, E., Lee, D., & Giles, C. L. (2007). Measuring conference quality by mining program committee characteristics. *JCDL'07: Proceedings of the 2007 conference on Digital libraries* (pp. 225–234). ACM Press. https://doi.org/10.1145/1255175.1255220

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article and at https://doi.org/10.5281/zenodo.4729758.

How to cite this article: Cabanac, G., & Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*, 1–16. https://doi.org/10.1002/asi.24495