Unveiling Scientific Articles from Paper Mills with Provenance Analysis

João Phillipe Cardenuto (✉ phillipe.cardenuto@ic.unicamp.br)
State University of Campinas

Daniel Moreira
Loyola University Chicago

Anderson Rocha
State University of Campinas

Article

Keywords:

Posted Date: May 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2791141/v1

License: ☛ ☞ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Unveiling Scientific Articles from Paper Mills with Provenance Analysis

João Phillipe Cardenuto1,*, Daniel Moreira2, and Anderson Rocha1

1University of Campinas, Institute of Computing, Artificial Intelligence Lab., Recod.ai, Campinas, SP, Brazil
2Loyola University Chicago, Department of Computer Science, Chicago, IL, USA
*phillipe.cardenuto@ic.unicamp.br

ABSTRACT

The increasing prevalence of fake publications created by paper mills poses a significant challenge to maintaining scientific integrity. While integrity analysts typically rely on textual and visual clues to identify fake articles, determining which papers merit further investigation can be akin to searching for a needle in a haystack, as these fake publications have non-related authors and are published on non-related venues. To address this challenge, we developed a new methodology for provenance analysis, which automatically tracks and groups suspicious figures and documents. Our approach groups manuscripts from the same paper mill by analyzing their figures and identifying duplicated and manipulated regions. These regions are linked and organized in a provenance graph, providing evidence of systematic production. We tested our solution on a paper mill dataset of hundreds of documents and also on a larger version of the dataset that deliberately included thousands of documents intentionally selected to distract our method. Our approach successfully identified and linked systematically produced articles on both datasets by pinpointing the figures they reused and manipulated from one another. The technique herein proposed offers a promising solution to identify fraudulent manuscripts, and it could be a valuable tool for supporting scientific integrity. All datasets, annotations, trained models, and implementations from this research are freely available at https://github.com/phillipecardenuto/upm.

Introduction

In 2018, Jana Christopher has raised concerns about the systematic and large-scale fabrication of image results in biomedical manuscripts1. While working for the Federation of European Biochemical Societies’ Press (FEBS Press), she reported sets of papers formally submitted to peer review containing suspicious Western blots. The unveiled problem resided mostly on the recurrence of unrelated experimental outcomes presenting identical substrate backgrounds and too similar individual bands. This suggested the probable composition of images by splicing together the same set of Western blots onto the same empty background to support ungrounded results. The involved publications – all rejected by FEBS Press – did not present an obvious relation regarding authorship or authors’ affiliation.

Elizabeth Bik and other investigators later confirmed Christopher’s concerns, who reported an extensive collection of manuscripts allegedly belonging to paper mills2 (i.e., undisclosed actors who seem to systematically fabricate scientific articles and forge images to support the articles’ claims, regardless of the absence of scientific ground). As of March 2021, Bik et al. have listed more than 1,300 documents suspected of coming from paper mills, of which 370 were retracted to date3.

Paper mills pose a new challenge to the community of scientific integrity verification. Typical cases of scientific misconduct comprise the inappropriate reuse, duplication, or manipulation of images executed by the same group of researchers, who usually split paper authorship and belong to the same laboratory4. Paper mills, on the contrary, are the source of several suspicious manuscripts that, despite sharing fabricated content, do not present a relation in authorship or authors’ affiliation1. Consequently, fake articles from paper mills may appear in diverse venues with non-related authors and topics, thus generating a needle-in-the-haystack problem. This new challenge is changing the scientific integrity landscape forcing the community to rethink its guidelines and detection tools5.

To understand the relevance of paper mills in the context of scientific integrity and problems related to images, we inspected the reasons for article retraction involving controversial images in the past decade (from 2010 to 2021), according to the Retraction Watch database6. As one might observe in Figure 1, the systematic fabrication of suspect papers has drastically increased in the last two years. We group this same data by scientific area in Table 1. Biological and Medical articles concentrate most of the retractions (89%), encompassing all the cases of paper mills. Based on this scenario, we decided to focus on biomedical manuscripts, planning to expand to other areas in the future.

As the outcome of researching and developing a computer-aided solution to automatically detect scientific manuscripts from paper mills, we propose a new forensic method extending provenance analysis7 to the case of biomedical scientific
Figure 1. Retracted papers due to image problems from 2010 to 2021. The number of retracted articles due to paper mill production has drastically increased from 2020 to 2021, asking for immediate action. Data source: Retraction Watch database as of September 2021.

<table>
<thead>
<tr>
<th>Retraction Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper Mill</td>
</tr>
<tr>
<td>Image Duplication</td>
</tr>
<tr>
<td>Image Falsification/Fabrication</td>
</tr>
<tr>
<td>Others</td>
</tr>
<tr>
<td>Image Error</td>
</tr>
</tbody>
</table>

Table 1. Distribution of retracted papers due to problematic images by scientific area from 2010 to 2021.

<table>
<thead>
<tr>
<th>Scientific Area</th>
<th>Biological and Medical</th>
<th>Physics</th>
<th>Business</th>
<th>Social Sciences</th>
<th>Environmental Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Retracted Papers</td>
<td>2697</td>
<td>303</td>
<td>10</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>#Paper Mill Retractions</td>
<td>375</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

articles. This is needed because these documents depict imagery that does not follow the usual distribution and nature of figures previously treated by provenance analysis (e.g., microscopic material and Western blots, instead of natural scenes and everyday objects). The proposed method starts with a collection of thousands of articles (i.e., the haystack) in Portable Document Format (PDF) and ends with pinpointing the most problematic figures within their suspicious documents (i.e., the needles). We rely on forensic image artifacts to link the figures – and ultimately the documents – that share content.

We test our technique on a dataset of papers suspected of coming from mills, as reported by Bik. Dr. Bik named this dataset the Stock Photos Paper mill (SPP). In addition, we also test the solution in an extended version of SPP, to which we have added thousands of distractor documents (i.e., papers without known issues). The extended SPP helps us understand the solution’s performance in a more realistic needle-in-the-haystack scenario, where there is a predominance of authentic articles and a smaller group of problematic ones. Our experiments demonstrate that the proposed method can efficiently and effectively unveil suspicious relationships among documents in large-scale scenarios of thousands of articles, grouping them according to the category of the reused images (e.g., microscopy imaging, blots, graphs).

Disclaimer. Although we rely on the impressive work of Bik and collaborators, who disclosed and collected the potential problems of the papers belonging to the SPP dataset, we must highlight that the proposed solution can neither establish the intentions of the authors of these papers nor it is our purpose to judge or denounce their actions. From our standpoint, the mere presence of an article within SPP (and its extended version) does not mean the presence or absence of misconduct. We are simply finding potential sharing of content between articles. Our ultimate goal is to make this tool available to the community and officials from institutions and integrity offices, who will have the final word about the cases.

In summary, the contributions of this work are:

• A novel automated solution to the problem of unveiling scientific articles suspected of coming from paper mills by relying on the articles’ figures. The proposed method starts with automatically extracting the figures from thousands of PDF articles and ends with a rich visualization of the shared content at both figure and document levels.

• A new provenance analysis method tailored for biomedical images that can track reused images throughout multiple publications.

• An annotated dataset composed of scientific articles suspected of coming from paper mills added to distractor publications (i.e., regular papers without known problems). This dataset is an extension of Bik’s work, containing 4,869 scientific
Figure 2. Extension of the SPP dataset annotations. The original annotations (left) are based on spreadsheets and do not provide details on the shared visual content. The proposed new annotations (right) rely on documents in JSON format to track and register all the figures within a document and all the panels within a figure that suspiciously share regions. The panel in this figure is present in the dataset and extracted from https://doi.org/10.1042/BSR20191453 under a creative commons license.

documents, of which 121 (~2.5%) were documented as suspects of belonging to mills. We report quantitative and qualitative results of the proposed technique and two other baselines from the literature over this dataset.

• A machine-learning-based content extractor of biomedical scientific figures, which segments the compound figures into multiple independent panels for further analyses. The extractor also filters the panels according to their type (e.g., microscopy imaging, blots), prioritizing the types with a high prevalence of misuse, according to scientific integrity analysts.

We are open-sourcing and releasing all the datasets, annotations, trained models, and implementations used in this research. The content is available at https://github.com/phillipecardenuto/upm.

Results

We organized the outcomes from this work in (i) Paper Mill Datasets, a collection of datasets for testing paper mill detection techniques; (ii) Quantitative evaluation including the metrics and effectiveness analysis for image reuse and systematic production detection; (iii) Qualitative analysis and output visualization of the proposed method herein; and (IV) Performance Analysis of the proposed automated panel extraction solution.

Paper Mill Datasets

Bik and collaborators have first reported a set of scientific articles suspected of belonging to paper mills. We started with their work and increased the amount and detail of the annotated problems. We also included distractor documents (i.e., documents without known issues) to challenge any proposed solutions to work with a more realistic needle-in-the-haystack scenario.

Stock Photos Paper Mill (SPP) Dataset. The SPP dataset contains 121 biomedical articles describing cancer types and cell tissue samples. Bik and collaborators annotated the suspicious occurrences of similar images throughout these papers. Such annotations were made publicly available on Bik’s website via spreadsheets. As one might observe on the leftmost panel within Fig. 2 (see “Original Annotation”), each row of a spreadsheet represents a particular article; it contains a column to identify the publication through its Digital Object Identifier (DOI), and a column with its label.

As a limitation of this annotation, one cannot pinpoint the similarities and shared visual content between two papers with the same label. To provide more complete annotations with a focus on media forensics, we are now detailing the content-sharing relationships between pairs of papers at the level of documents (121 samples), figures (498 samples), and panels (2,581 items). The latter level (panels) is grouped into categories (e.g., panels that depict microscopy imagery, western blots, graphs). Fig. 2 puts in perspective the original SPP dataset spreadsheet-based annotations (on the left side) with the proposed ones (on the right side). The new annotations are stored in JavaScript Object Notation (JSON) to cope with the more complex information.

Extended SPP Dataset Versions. Aiming to challenge the proposed solutions and understand their performance in terms of false alarms (when programs wrongly accuse issues over non-problematic data), we extended the SPP dataset by adding documents without known problems (distractors). To make more realistic scenarios — where the majority of the analyzed
papers do not have problems — and to understand the progression of the challenge over larger and larger sized corpora of articles, we created two versions of the SPP extension. In the first one (namely “v1”), 969 papers containing biomedical figures were added to the SPP dataset. Each paper was found through its figures, which were queried using the Open Access Biological Image Search Engine. We explicitly selected interest figures through the engine’s categories, such as microscopy images, flow cytometry, and Western blots. In the second version (namely “v2”), 3,635 papers were similarly added to the first version. As of the date of this writing, all added articles were not associated with any suspected image-related misconduct.

Table 2 summarizes the respective numbers of the annotated documents, figures, and panels that constitute the SPP, extended SPP (v1), and extended SPP (v2) datasets.

### Table 2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Documents</th>
<th>#Figures</th>
<th>#Panels</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPP</td>
<td>121</td>
<td>498</td>
<td>2581</td>
</tr>
<tr>
<td>Extended SPP (v1)</td>
<td>1090</td>
<td>1467</td>
<td>10143</td>
</tr>
<tr>
<td>Extended SPP (v2)</td>
<td>4725</td>
<td>5303</td>
<td>47540</td>
</tr>
</tbody>
</table>

**Quantitative Results**

To verify the effectiveness of the proposed solutions and baselines from the literature, we assess their performance in the face of three tasks:

- Finding images and documents that share content one with another (Content Pairing).
- Grouping and determining the categories of reused elements within a collection of scientific images and documents (Content Grouping).
- Classifying images and documents as either produced by a paper mill or not (Content Classification).

In the following, we report the results of the baselines and the proposed solutions applied over the three versions of the SPP dataset (regular and extended “v1” and “v2”). Results are organized into two levels: image versus document. The first one is dedicated to finding forensic traces that link the scientific images (i.e., analysis at image level), while the second one aims at finding shared content between the documents (i.e., analysis at document level).

**Baselines.** We selected two state-of-the-art methods previously proposed in the literature of digital forensics and scientific integrity as baselines to put the herein-proposed solution into perspective:

- **Bioscience-scale automated detection of figure element reuse.** Acuna et al. introduced a human-assisted methodology to detect image reuse in biomedical databases. The proposed pipeline contains one module for copy-move detection and another for biological image classification. The copy-move stage is an interest-point-based extraction and matching algorithm similar to the one developed by Amerini et al., except for an adaptation to the case of scientific images (e.g., microscopy imagery) instead of natural scenes (e.g., outdoor landscapes). Due to their time-consuming algorithm, their analysis was contained within only hundreds of articles with the same last or/and first authors. However, articles from the same paper mill — with few exceptions — do not share any authors. Because of that, many articles should be considered to unveil paper mills. In contrast to Acuna et al.’s solution, the proposed method herein filters the most suspect figures before running any time-consuming analysis by grouping the most similar items. This filtering process enables our solution to process more data in a more feasible time. We were unable to locate an open-source implementation of the copy-move detection method developed by Acuna et al., and therefore, we developed our own version based on their article, which we employed in our experiments. Our in-house implementation is available on the GitHub repository of this article. Throughout the article, we refer to our implementation of Acuna et al.’s method as BSRD, which stands for Bioscience-Scale Reuse Detector.

- **SILA.** A System for Scientific Image Analysis (SILA), which contains several modules (including a provenance analysis one) from the literature of digital forensics tuned to scientific images. SILA’s provenance analysis is based on the work of Moreira et al. and Bharati et al., with components added to cope with scientific images (e.g., a text detection component that avoids matching sub-panel legend letters within the scientific figures). In contrast to the proposed solution herein, SILA’s method was designed to find the relationships of a single query image with a scientific image corpus. On the other hand, our solution is designed to find all relationships within a collection of either document (document level) or their inner figures (image level), with no need for an explicit query image of interest.
Fig. 3. Content pairing results per image type for each SPP dataset version. Blots are the most challenging type of image for all solutions. Contrary to the other versions and regardless of the version of the SPP dataset, the proposed method does not suffer significant drops of CP in the presence of distractors. BSRD refers to our implementation of Acuna et al.’s method.

Table 3. Results of content pairing at the image level. The best results are in bold. The proposed solution is more robust to the addition of distractor documents (i.e., documents that present no known issues) and their respective images, being more indicated in more realistic scenarios of a needle in the haystack (a few problematic documents and the majority of non-problematic elements). BSRD refers to our implementation of Acuna et al.’s method.

<table>
<thead>
<tr>
<th>Dataset \ Method</th>
<th>BSRD</th>
<th>SILA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPP</td>
<td>0.54</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>Extended SPP (v1)</td>
<td>0.25</td>
<td>0.66</td>
<td>0.74</td>
</tr>
<tr>
<td>Extended SPP (v2)</td>
<td>0.04</td>
<td>0.36</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 4 reports the CP results at the document level. The proposed method and SILA are both stable in finding reused
Proposed reports the CG results at the document level. As expected, considering only the articles from SPP (no distractors),

Table 4. Results of content pairing at the document level. The best results are in bold. The proposed solution works on par with SILA, presenting a better CP in the extended SPP (v2) version of the dataset, which presents the most distractors. Blots (i.e., gel electrophoresis images) are the most challenging type for all methods.

<table>
<thead>
<tr>
<th>Dataset \ Method</th>
<th>BSRD</th>
<th>SILA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPP</td>
<td>0.72</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Extended SPP (v1)</td>
<td>0.03</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>Extended SPP (v2)</td>
<td>0.01</td>
<td>0.73</td>
<td>0.76</td>
</tr>
</tbody>
</table>

images across different documents, even with distractors (i.e., documents without known image reuse issues). On the other hand, BSRD (our implementation of Acuna et al.’s method) significantly drops its performance in the scenario with distractors.

**Content Grouping**

The task of content grouping (CG) measures the effectiveness of a method in identifying and grouping images that share content (image-level) or publications that reuse the same images (document level). This task was inspired by the reused image categories identified by Dr. Bik and other investigators (e.g., colony formation assay photo category one, CF1), in which multiple images with small variation among themselves are reused throughout multiple articles. Within the SPP dataset annotation, each reused image from the same group appears linked to another (a.k.a., connected component in computer science).

To assess the effectiveness of the solutions for this task, we rely on node precision (NP) and node recall (NR). NP is the number of identified images or documents (the nodes) correctly included in the same predicted group divided by the number of elements annotated for this group. It answers: “how many items from a predicted group correctly belong to the same category?”. On the other hand, NR is the number of elements from the same category correctly grouped and divided by the annotated number of elements of their category. It answers the following question: “from an annotated category, how many elements of this category was a solution able to identify and insert into the same group?”. As close NR and NP are to one, the better the solution for this task. Similarly to the case of CP, CG uses the harmonic mean of NP and NR (see Eq. 2). CG is inspired by the metric Node Overlap (NO) concept proposed by NIST to evaluate the nodes of a provenance graph. CG value is within the interval [0, 1]. The close its value to one, the better the solution.

$$CG = \frac{2 \times NP \times NR}{NP + NR}$$  \(2\)

Table 5 expresses the CG results at the image level for each SPP dataset version. Consistent with the results obtained in CP, both BSRD and SILA solutions demonstrate a marked decline in performance as the number of distractors increases from SPP to Extended SPP (v2). In contrast, the proposed solution exhibits only a slight deviation across all dataset versions, with a difference of only four points in the dataset version with the most distractors.

Fig. 4 depicts the CG performance of each evaluated method among all dataset versions according to the type of problematic image. As observed in the case of CP results, blots also pose the most challenging type of image.

At the document level, CG considers the group of documents that belongs to the same paper mill. Given that we only included one group of suspect papers, the SPP, we expected all 121 publications from this group to be included in the same group. Table 6 reports the CG results at the document level. As expected, considering only the articles from SPP (no distractors), all solutions scored close to one, the maximum score. However, all solutions lose effectiveness when including more and more distractors from Extended SPP (v1) to Extended SPP (v2). BSRD solution is the most affected by the distractors, dropping the CG value from 1.00 to 0.01 with the addition of the distractors. The proposed solution has the lowest impact when facing distractors, dropping only eight percentage points in the worst scenario.
Figure 4. Content grouping results per image type for each SPP dataset version. Blots are the most challenging type of image for all solutions. Like the case of CP, the proposed method does not suffer significant drops of CG in the presence of distractors.

<table>
<thead>
<tr>
<th>Dataset \ Method</th>
<th>BSRD</th>
<th>SILA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPP</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Extended SPP (v1)</td>
<td>0.04</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Extended SPP (v2)</td>
<td>0.01</td>
<td>0.86</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 6. Results of content grouping at the document level. The best results are in bold. As the number of distractors increases from SPP to Extended SPP (v2), the performance of all methods decreases, however, the proposed exhibited the most robust performance across different numbers of distractors.

Content Classification

Content classification (CC) involves determining whether a picture or publication is systematically produced (i.e., created by a paper mill). This task tries to answer: “how effective is a solution in identifying an item from a paper mill production?”. Within the SPP dataset, each reused image (image-level) and article (document-level) reported by Dr. Bik and the other investigators were marked as systematically produced – for this task. All other items were considered as not coming from paper mills.

To assess CC, we rely on precision and recall metrics. The precision metric is defined as the ratio of the number of items correctly identified by a solution as products of paper mills to the total number of items identified by the solution. Recall is the ratio of elements correctly predicted as systematically produced to the total number of elements annotated as paper mill results. To deal with the haystack of non-problematic articles in the SPP extended datasets (i.e., imbalanced datasets), CC uses the harmonic mean of precision and recall (Eq. 3). CC can take any value between zero and one, with a value closer to one indicating better performance.

\[
CC = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

(3)

Table 7 indicates the CC results at the image level for each SPP dataset version. As expected, when applied to the isolated images of SPP, all solutions have a high performance, indicating that they are precise and could identify most of the reused images. However, with the addition of image distractors to the Extended SPP v1 and v2, BSRD and SILA failed to identify the reused images. In turn, the proposed method kept its high performance in all scenarios, indicating robustness to distractors.

Fig. 5 depicts the CC performance of each evaluated method’s overall dataset version and image types. The results show that identifying Blots poses the biggest challenge for CC compared to other image types. BSRD has a great CC value when applied to the SPP dataset without distractor – the best performance for the Flow Cytometry type. However, its performance significantly drops when applied to the needle in a haystack scenario (the extended SPP datasets). SILA exhibits a similar but less pronounced drop in performance. In turn, the proposed solution has stable and higher CC even when applied to the worst scenario.

Table 8 reports the CC results at the document level. As expected, all solutions achieve CC values close to one when applied to the dataset without distractors. However, when the number of distractors is increased, the performance of the solutions diminishes. Specifically, when non-problematic items are added to the dataset, BSRD solution is the most affected, with a drop of ninety and six percentage points in CC. In contrast, the proposed solution is robust when the haystack size increases, maintaining stable and higher CC values across all dataset versions.
Table 7. Results of content classification at the image level. The best results are in bold. The proposed solution outperforms the other evaluated techniques on all datasets for the task of content classification. Especially on the Extended SPP (v2) dataset, our method achieves nearly double the performance of SILA, the second-best performing method.

<table>
<thead>
<tr>
<th>Dataset \ Method</th>
<th>BSRD</th>
<th>SILA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPP</td>
<td>0.87</td>
<td>0.84</td>
<td>0.91</td>
</tr>
<tr>
<td>Extended SPP (v1)</td>
<td>0.44</td>
<td>0.77</td>
<td>0.91</td>
</tr>
<tr>
<td>Extended SPP (v2)</td>
<td>0.08</td>
<td>0.44</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Figure 5. Content classification results per image type for each SPP dataset version. Differently from the other solutions, the proposed method does not suffer significant drops of CC when more and more distractors are added from the SPP dataset to its extended versions.

Table 8. Results of content classification at the document level. The best results are in bold.

<table>
<thead>
<tr>
<th>Dataset \ Method</th>
<th>BSRD</th>
<th>SILA</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPP</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Extended SPP (v1)</td>
<td>0.26</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Extended SPP (v2)</td>
<td>0.06</td>
<td>0.86</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Qualitative Results

In this section, we present the provenance graphs generated by our solution when applied to the SPP dataset (v2), providing an explainable and user-friendly visualization of the images and documents identified as being produced by paper mills. Furthermore, the qualitative results are presented separately for image level and document level, providing insights into the performance of our solution in each context.

Image-level Results

Our solution provides a provenance graph visualization in which the graph nodes represent images and the links represent their relationships of reuse. For example, a link from a node A to another node B indicates that part of A is reused in B.

Figure 6 depicts the provenance graph generated by our solution when applied to the SPP Extended (v2) dataset. Each node in the graph represents an image labeled with its digital object identification (DOI) and a reference to the reused figure. This graph has the largest number of reused images among all the output graphs and consists entirely of microscopy images. To avoid cluttering the visualization, we only show the links that keep the graph connected as a single group, i.e., the connected component. Consequently, the absence of a link between two nodes does not necessarily indicate that they do not share content. After comparing Fig. 6 graph with Bik’s annotations, we have found that the automatically-computed outcome matches the category labeled as “TW14” within their annotation schema. Despite the large presence of distractors, all annotated images were correctly found and matched by our method, not including any false positive link.

Fig. 7 depicts the output of our method when applied to Western blots. Blue nodes represent correctly found figures from the image group labeled as “SWB1” within Bik’s annotations. In turn, red nodes comprise the images our method failed to detect (a.k.a. misses or false negatives). As presented throughout the quantitative results for all the tested methods, Western blots are a challenging type of image to find regions of content sharing. The graph presented in Fig. 7 shows that only six out of 18 images within the group “SWB1” were correctly predicted as sharing content. This result indicates both a limitation of our method and a research opportunity for future work.

Document level. Figure 8 illustrates the largest provenance graph that the proposed solution generates when applied to the extended SPP dataset (v2) at the document level. Each node in the graph represents a scientific document that shares content
Figure 7. Provenance graph of western blots related to the group labeled as “SWB1” within Bik’s annotations. Blue nodes refer to correctly predicted figures that share content. Red nodes indicate missed figures not found by the proposed method but that share visual content. Due to copyright issues, we did not include the real figures in the graph. Below each node, we provide the DOI of the document that is the source of the involved figure, as well as the reference used in the document to such figure.
Figure 8. Provenance graph generated by the proposed solution, representing the document-level relationships between articles sharing content in the extended SPP dataset (v2). All documents within this graph were reported as problematic by Dr. Bik’s investigation, without any false positives. Each node in the graph corresponds to a publication, with its DOI indicated below. The most densely connected region of the graph is magnified, revealing a document that shares its content with many others.

with another. Similar to the image-level graph, the absence of a link in this graph does not necessarily mean that the nodes do not share content. To reduce clutter, we only draw the links that keep the graph connected. Upon comparing this graph with the paper mill documents identified in Dr. Bik’s investigation, we found that all suspect publications were correctly included without adding distractor articles.

Fig. 9 contains additional provenance graphs predicted by the proposed method over the extended SPP dataset (v2). Below each graph node (which represents a publication), there is a PubMed Central identifier (PMC); it refers to one of the distractors we have added to extend the SPP dataset and, therefore, comprise false alarms (only 21 out of 4604). Differently from the previous graph containing all reported articles, these additional seven graphs have smaller sizes, with at most five nodes. To understand the reason for such cases, we analyzed the images identified as false alarms by the method. We found that there is a high visual similarity among these images. Notably, most detected images are creative commons material commonly used as illustrations in the biomedical field. Additionally, these images were properly cited during their reuse and should not raise any concerns for human reviewers.

Figure 9. Seven provenance graphs at document level that were found by the proposed solution herein but do not represent paper mills (i.e., false alarms) over the extended SPP dataset (v2). Each document is identified through its PMC identification ID, provided below the publication’s graph node. Out of 4096 PMC distractors, only 21 triggered findings. This was caused by documents presenting very similar figures.
Automated Panel Extraction

Panel extraction is essential to focus on the image regions of interest to the scientific integrity problem, and filter out those that might raise false alarms due to their intrinsic similarity (e.g., diagrams, drawings, and legend indicative letters). For this purpose, we collected and annotated 3,836 biomedical scientific figures under creative commons license from different journals, creating a dataset of 3,236 figures (32,507 panels) for training the detector of panels, and 600 figures (4,888 panels) for testing it. We used the annotation tool Label Studio during the labeling process to draw the location of each panel within the figures. For example, Figure 10 depicts a sample from our dataset. Each colored rectangle depicts a different type of panel. Each figure’s URL and annotations are available in the GitHub repository of this work.

![Figure 10. Compound Figure Annotation for Panel Extraction. Each colored rectangle corresponds to a panel that should be extracted as part of the panel extraction task. The categories of each rectangle are indicated by their color. For instance, the green panels are annotated as microscopy imaging. To generate this example, we used the figure distributed under a creative commons license found at https://doi.org/10.1371/journal.pone.0152712.g002.](image)

The proposed panel extractor is based on Yolo-v5, an object detection model for digital images. Given our biomedical figures scope, we trained the solution to pinpoint the location of the most problematic biomedical types of panels:

- **Microscopy Imaging**: Photos taken by a microscope that can be fluorescent labeling, histology staining, or other types of tagging of cells or tissues, or components within, captured by light, electron, fluorescent, or other microscope types.
- **Blots**: The resulting image from techniques to detect specific proteins in a tissue sample or extract them using electrophoresis. This category includes Western, Northern, and Southern blots.
- **Body Imaging**: Image of the whole organ – from any living being – including images from Magnetic Resonance Imaging (MRI), Computerized Tomography/Computerized Axial Tomography (CT/CAT) scans, Positron Emission Tomography (PET), X-ray, and ultrasounds.
- **Graphs and Plots**: Experimental data charts including bar, line, scatter, pie, area, and histogram plots.
- **Flow Cytometry**: The resulting plot from a technique used to detect and measure physical and chemical characteristics of a population of cells or particles.

We evaluate the proposed technique using the average precision (AP) – a popular metric in computer vision for object detection. The higher the AP, the more accurate the detection. Typically, AP is used with a parameter (i.e., a threshold) that will only consider a panel as successfully detected if its overlap (i.e., intersection over union, namely IoU) with the annotated area is higher than the threshold. In this study, we considered an IoU threshold of 0.5 (AP @ 0.5). The model achieved an AP score of 93.4% on the test set. Table 9 shows the model’s performance on the test set per type of panel. The herein-proposed panel extractor achieved high scores for all evaluated classes.
Table 9. Performance of the panel extraction solution by image panel type. We report the average precision with an intersection over union of at least 0.5 (AP @ 0.5). A perfect solution would lead to an AP value of 1.0.

<table>
<thead>
<tr>
<th>Panel Type</th>
<th>Blots</th>
<th>Graphs</th>
<th>Microscopy</th>
<th>Body Imaging</th>
<th>Flow Cytometry</th>
<th>Mean AP @ 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP @ 0.5</td>
<td>0.996</td>
<td>0.958</td>
<td>0.922</td>
<td>0.825</td>
<td>0.969</td>
<td><strong>0.934</strong></td>
</tr>
</tbody>
</table>

Discussion

Digital forensics is crucial for maintaining scientific image integrity in the digital age. Early forensic tools have focused on detecting scientific image tampering, using image segmentation techniques as proposed by Farid and pixel-wise comparison as proposed by Koppers et al. However, with the increase in image reuse and manipulation across different scientific articles, new automated methods have emerged to address these issues, e.g., Bucci. Furthermore, even sophisticated forensics techniques that leverage artificial intelligence have been proposed to detect scientific image tampering, such as BioFors. While these methods have shown some effectiveness, further research and development are still needed to improve their performance and make them more robust for scientific integrity assessment.

To our knowledge, no digital forensic technique has been proposed that uses provenance analysis for detecting paper mills. Additionally, research is lacking in analyzing large datasets for scientific image integrity. One of the most well-known methods is Acuna et al., which uses a human-assisted methodology to identify image reuse in the bioscience scale. Acuna et al.’s solution starts with a copy-move detector across figures, followed by an automated model that classifies the matched image pair as biological or not. Finally, a human revises the biological matched pairs looking for problematic reuse. Although they tested their method on 2.7 million scientific figures, their analysis was limited to figures from the same first and last authors in the collection. Consequently, an image would only be compared to a few hundred others, i.e., the other figures from the same authors. While their method represents an advance for scientific image integrity, the authors acknowledge that their approach is computationally complex and not scalable for comparing thousands of images.

A more complex tool for scientific image analysis was proposed by Moreira et al. using digital forensic-based state-of-the-art techniques. Their system can deal with real-world cases, starting from PDF documents and ending with copy-move detection and provenance analysis of their figures. Their work presents the first solution that leverages the provenance analysis pipeline for scientific image integrity to the best of our knowledge.

In contrast to prior work, our proposed technique is specifically designed to identify articles produced systematically in a large-scale scenario, leveraging the content of their figures. Our method starts by extracting figures from a collection of suspicious documents, then isolating and filtering the most concerning panels within those figures, which is essential for dealing with the large scale. Finally, it conducts a provenance analysis on the extracted panel images to track potential manipulations and reuse.

To assess the efficacy of our automated extraction and filtering solution, we tested it on a dataset generated as part of this research work. The results demonstrated strong performance (Mean AP = 93.4%) in accurately extracting and filtering the relevant panels of interest. In addition, to put the performance of our proposed forensic solution in perspective with existing ones, we compared it to Acuna et al, and Moreira et al. works. We use three metrics: Content Pairing, which evaluates if the method can find pairs of images and documents sharing content; Content Grouping, which assess the groups of images or documents conveying the same type of content type; and Content Classification, which classifies images or documents as being systematically reused.

Our experiments showed that all three solutions were effective when applied to small document sets containing only paper mill-related articles, as indicated by all evaluated metrics. However, when applied to larger datasets with multiple distracting articles (i.e., non-problematic papers), only the proposed solution remained stable with a high effectiveness score, varying only a few percentage points for all metrics. Thus, in contrast to the baselines, the proposed method proved more robust in a realistic scenario where most analyzed papers do not have problems.

The most challenging task we evaluated was Content Pairing. Our proposed solution scored 71% (the score value is between 0 and 100%) when applied to the Extended SPP (v2) dataset. Our method scored higher than the others, as Moreira et al. scored 36% and Acuna et al. scored 4%, but there is still room for improvement. These improvements could be mainly concentrated on Western blot images, which were particularly challenging and obtained the lowest scores for all evaluated solutions. Western blot images often have low entropy (i.e., their foreground objects do not have a clear boundary), making it difficult for SIFT-based methods – used in all evaluated solutions – to find the best points for matching. Future research could focus on developing more advanced techniques to improve matching these types of images.

Overall, our results indicate that this work provides a viable solution using provenance analysis for the recently reported paper mill cases. However, the future of paper mills remains uncertain, and further research and discussions are necessary to address such a challenging problem. Eventually, paper mills will adopt artificial intelligence algorithms to generate realistic solutions.
scientific figures. Although this may seem like a futuristic threat, recent studies have demonstrated the ability of Generative Adversarial Networks (GANs), a type of Deep Learning method also used for creating Deepfakes, to create convincing Western Blots even on low computational power devices such as domestic notebooks. Investigating the potential of such a Deep Learning algorithm, Qi et al.²⁶ were able to artificially create realistic Western Blots that even biomedical specialists were unable to distinguish from pristine ones.

Because of these new threats and the ones that may appear, our endeavor to develop integrity tools alone may not be sufficient to assure scientific integrity. We believe such a complex problem should be addressed by a multi-pronged approach that leverages technology, policy, and education to create a culture of integrity and accountability within the scientific community to prevent scientific misconduct, such as paper mills.

Method

The methodology is structured into two sections: 1. Filtering & Evidence Collection and 2. Scientific-Based Provenance Analysis. The first describes the proposed filtering process that allows our method to be applied to large-sized datasets. The second presents the proposed provenance analysis method tailored for scientific images that can be used to identify articles produced by paper mills.

Filtering & Evidence Collection

The Filtering & Evidence Collection stages aim to parse, organize, and gather evidence for further analysis, as Figure 11 shows. The process begins with collecting suspicious PDF documents, from which all figures are extracted using an automated method (Document Parsing). The solution identifies and extracts each figure’s panel, filtering them into the five types of interest (Filtering). Finally, an artificial intelligence model is used to identify each image’s unique features, and their descriptions are saved in a database (Evidence Storage), which will be used in the Provenance Analysis phase to identify similar panels.

Document Parsing: Given the Portable Document Format (PDF) as the standard for scientific documents, we included an automated PDF figure extraction in the pipeline. Of course, this step is not required if it is possible to collect the scientific figures directly from another source. However, many online publications do not release their figures separately from their PDF document, making the scientific image analysis depend on a PDF parser. For example, we experience this behavior in most articles from the SPP dataset, whose images were not available for download, only the PDF.

To parse a PDF document and extract its figures, we use Moreira et al.’s PDF parser.¹³ Their method identifies the PDF instruction regarding figure rendering and does not apply any image processing operation. By not applying such operations, they avoid including noise or additional information in the final image that could confuse further analysis.

Figure 11. Filtering & Evidence Collection Workflow. A suspect collection of documents undergoes parsing and figure extraction, resulting in a set of figures. These figures are then processed by a filtering stage that identifies and extracts the panels of interest. Later, a machine-learning model creates a robust evidence representation that can withstand transformations commonly applied to panels, such as resizing, compression, and color changes. Finally, each panel representation is stored in a database for further analysis. The figure used to depict the workflow is available under the creative commons license at https://doi.org/10.1371/journal.pone.0152712.g002.

Filtering: Biomedical articles often contain complex figures with multiple panels that depict various analyses of an experiment, as Figure 10 shows. However, certain regions within these figures may not provide informative content and could hinder image analysis due to their inherent similarities, such as letter-based labels or schematic diagrams. Filtering and retaining only the informative regions is crucial to ensure scientific image integrity. To identify these regions, we consulted recognized guidelines and studies and focused on the most problematic panel types: Microscopy Images, Blots, Body Imaging, Graphs and Plots, and Flow Cytometry, as described in the Automated Panel Extraction section of the Results.
The filtering stage employs a YoloV5\textsuperscript{18} deep-learning model-based object detection method to identify and locate informative panels in an input figure. While other detection models could be used, we opted for a lightweight model like YoloV5 to ensure that our filtering stage remains computationally efficient and accessible to integrity analysts with limited computing resources.

The proposed filtering model was pre-trained on natural images – e.g., social-media photos from the COCO\textsuperscript{28} dataset – and later fine-tuned it for scientific figures. For the pre-trained model, we used the weights from YOLOv5x6 provided by the YoloV5 authors\textsuperscript{18}. To fine-tune this model on scientific data, we collected 3,836 biomedical article figures under a creative commons license, considering the ones with at least one informative panel type (i.e., Microscopy or the other four image types).

We annotated the articles’ figures using Label Studio\textsuperscript{17} to locate informative panels within them. Subsequently, we split the figures into 3,236 for training (comprising 32,507 informative panels) and 600 figures for testing (totaling 4,888 informative panels) using the annotated images. To obtain the figures, we used the Open Access Biomedical Image Search Engine’s API\textsuperscript{10} and queried for the informative panel types.

**Evidence Storage:** While preparing figures for scientific articles, researchers often apply various post-processing operations, such as resizing, color adjustments, and compression, to enhance the visibility of their findings. However, these operations can inadvertently introduce artifacts or remove crucial information from the image. For example, even a simple photograph resizing to fit it into a panel placeholder can result in distortions or loss of important details, affecting the integrity analysis.

To avoid such artifacts distracting provenance analysis, we create evidence representations of each panel image content using a deep-learning model robust to several types of image processing (e.g., color-changing, noise, rescale, mirroring, crop, among others). This model’s robustness ensures that the original image and its processed version (e.g., a rescaled image) are similar, allowing the provenance analysis to retrieve the original image by giving its processed version, and vice versa, all based on their visual similarity.

For creating these representations, we utilize the MobileNetV3\textsuperscript{27} model, which is pre-trained on ImageNet\textsuperscript{29}, a large dataset for image classification. MobileNetV3 is a lightweight and powerful model for image description that identifies robust features from an image using low computational power. For extracting the evidence descriptions of each image, we get the feature vector from the last layer of MobileNetV3 after removing its classification portion.

**Scientific-Based Provenance Analysis**

Provenance analysis is a powerful technique for identifying and understanding object relationships within a collection, especially when suspicious links exist. One common application of provenance analysis is understanding the relationships between a single media item and a data collection\textsuperscript{14}. We are interested in identifying potentially suspicious links between multiple scientific documents and their figures, which may indicate papers produced by a paper mill. To achieve this, we have divided the provenance analysis into two levels: image and document. Both levels provide invaluable insights for analysts conducting investigations. The image level identifies the reuse and manipulation of images across a set of articles, while the document level visualizes the most suspicious systematically produced documents.

**Image Provenance Analysis Workflow**

Image provenance analysis tracks forensic clues that indicate image manipulation and duplication. Leveraging the collected figure panels extracted through the Filtering & Evidence Collection process, our method tracks reused panels and identifies possible manipulated versions.

The provenance workflow steps are divided into 1. Content Retrieval; 2. Consistency Check and Matching; 3. Content Sharing Score Calculation; 4. Content Shared Table Building; 5. Identification of Suspicious Images; and 6. Provenance Graph Output. Figure 12 depicts the complete analysis. For didactic sake, in the following, we describe the provenance workflow using a generic image panel $P$ from the database returned by the Filtering & Evidence Collection. All other panels from the database will also be processed using the same workflow.

**Content Retrieval:** Provenance analysis starts by performing a content-based image retrieval using $P$ as a query. This step is depicted in Fig. 12 (step 1). To perform content retrieval, the method compares the features of $P$ (extracted during the Filtering & Evidence Collection stage) with all other features from the database using cosine similarity. Then, the top-$K$ similar panels to $P$ ($K = 100$) are inserted into a processing queue. Panels extracted from the same source document as $P$ are not included in the processing queue, as our goal is to identify reused images across different documents.

**Consistency Check and Matching:** This step compares each element from the queue with $P$ using a local-description strategy, which involves comparing the regions of $P$ to the regions of the items from the queue at the pixel level. During this analysis, the method searches for reused regions and possible manipulation traces. Fig. 12 (step 2) depicts the process. The local-description analysis is computationally expensive, and comparing all elements in the database with $P$ would take a significant amount of time (weeks or months for large databases). Content Retrieval plays a crucial role in the Provenance
Analysis by reducing the local analysis of \( P \) to the top \( K \) most similar panels in the database. In addition, this approach enables matching to be quickly accomplished (in minutes or seconds).

Let the next item from the processing queue be denoted as \( R_I \) for the sake of illustration. The consistency process starts by selecting interest points from the image of \( P \) and the image of \( R_I \) to verify if their contents are consistent. The method relies on scale-invariant features transform (SIFT)\textsuperscript{30}. SIFT identifies interest points from the images and generates local descriptions from them (feature vectors) that are robust to scale, rotation, and color. The method stores its coordinates (x and y position within the image) and SIFT descriptions for each selected point. Using the interest points descriptions, the method matches the most similar interest points of \( P \) and \( R_I \) using a brute-force strategy (i.e., checking all possible pairs of descriptions).

As expected, some matches from \( P \) to \( R_I \) are false. To consider only the consistent ones, the method finds a homography transformation \( H \) that aligns the largest possible number of matched points from \( P \) to \( R_I \). This transformation is depicted in the multiple green parallel lines of Fig. 12 (step 2), which map \( P \) points into \( R_I \) points. Homography transformations will possibly re-scale, rotate, and translate the interest points of \( P \) to match its content to \( R_I \)\textsuperscript{31}. To find the homography transformation and eliminate incorrectly matched interest points, we use the MAGSAC method\textsuperscript{32}.

As SIFT features are not robust to mirroring, we also perform the same interest point matching and image alignment in the mirrored version of \( P \), named \( P' \). If neither \( P \) nor \( P' \) panels have at least 20 consistent matching interest points with \( R_I \), we consider the pair \((P, R_I)\) inconsistent and discard it.
**Content Sharing Score Calculation:** This step verifies if the content between $P$ and $R_I$ is relevant to an integrity analysis, taking into account the region of $P$ that was matched with $R_I$, as illustrated in Fig. 12 (step 3). To delimit this region, the method finds the smallest convex polygon (a.k.a. convex hull) that involves all matched interest points from $P$ to $R_I$ (Fig. 12, step 3). Then, using the area of this polygon divided by the total area of $P$, we calculate the Content Sharing Score of $P$. Mathematically, the content sharing score is given by the convex hull area built using the matched interest points from $P$ to $R_I$ divided by the total area of $P$. In the following equation, $P \cap R_I$ represents the Content Sharing Score of $P$ with $R_I$, while $A(X)$ represents the area of $X$.

$$P \cap R_I = \frac{A(P_{\text{Convex Hull}})}{A(P)}$$

By comparing the content sharing score between $P$ and $R_I$ with a threshold, we can determine if the region of $P$ matched to $R_I$ is relevant for an integrity analysis. Conversely, if the content-sharing score is below the threshold, we consider the pair $(P, R_I)$ inconsistent and discard it.

The content sharing score is a value between 0 and 1, indicating the degree of similarity between the content of $P$ and $R_I$. A score close to 0 indicates that $P$ shares few portions with $R_I$, while a score close to 1.0 indicates that $P$ shares most of its content with $R_I$. It is important to note that the content sharing score of $P \cap R_I$ differs from that of $R_I \cap P$, as these values are relative to each image. For instance, if $P$ is a cropped version of $R_I$ that covers 50% of the area of $R_I$, then the content sharing score $P \cap R_I$ would be equal to 1.0, while the content sharing score $R_I \cap P$ would be 0.5.

**Content Shared Table Building** In this stage, a table is generated to indicate the Content Sharing Score between $P$ and all items in the processing queue (as shown in Fig. 12, step 4). Using this table, the method determines the most relevant images for integrity analysis. An image is considered relevant for integrity analysis (i.e., a possible duplication) if its content sharing score in the Content Shared Table exceeds a threshold. Empirically, a threshold of 10% is effective, meaning that the $(P, R_I)$ pair is deemed suspicious if $P \cap R_I \geq 0.1$ or $R_I \cap P \geq 0.1$.

Note that if $R_I$ is relevant to $P$ (i.e., they share content), other similar images to $R_I$ may also be relevant to compare with $P$. Thus, the method expands the processing queue of $P$ by adding the top-$L$ most similar images of $R_I$, similarly as performed in the Content Retrieval step, but using $R_I$ as a query. We use $L = 40$ in our experiments. To avoid redundancy, images that have already been processed will not be included twice in the queue.

**Processing Queue Loop:** Once panel $R_I$ has been analyzed, it is removed from the processing queue, and the next panel is selected for examination. This process is repeated until either 400 panels have been analyzed or the queue becomes empty. Based on our experiments, panels with a queue position of 400 or higher generally have significantly different content from panel $P$, and therefore do not provide meaningful evidence for the provenance analysis. Additional panels beyond the 400th iteration are not considered for analysis.

**Provenance Graph Construction:** This step aims to group all items that share content and creates a provenance graph visualization (as illustrated in Fig. 12, steps 5 and 6). This visualization presents a set of images that relate one to another, tracking the image with the highest content sharing score.

Before starting the Graph Construction step, all panels within the database should be processed through the previous steps, and their Content Shared Table must be filled. This allows the method to determine how much content an image shares with others. Subsequently, the method identifies and links all pairs of images with content-sharing scores greater than 10%, isolating them into groups. Within a group, each element is related to at least one other element in the same group. These groups are formally known as connected components in computer science graph theory.

Within a connected component, an image may be linked to multiple others, resulting in a dense graph that can be difficult to understand and visualize. Figure 13 illustrates an example of a connected component before and after improving its visualization. To enhance the visualization and identify duplicated images, the method removes all cycles within the connected components and preserves the links that maximize the sum of all content-sharing scores. Specifically, the method computes the maximum spanning tree, an undirected graph with the maximum weights on its links, for each connected component. Figure 13 shows the output before and after pruning the links of a connected component. The pruned version of the connected component is the provenance graph, which identifies images connected to numerous others, providing evidence of systematic production. For instance, the central blue node in Figure 13, which has the most connections to other nodes in the graph, may indicate the source of a systematic production and could be the starting point for a scientific integrity investigation.

**Provenance Analysis at Document-Level**

The goal of document provenance analysis is to indicate the group of documents that reuse elements of each other, providing clues of systematically produced articles. This analysis uses the content shared table from the image level to link the image
Figure 13. Visualization of a graph before and after computing its maximum spanning tree, referred to as the provenance graph. Each blue node represents a different image, and the links between nodes indicate that their corresponding images share content. On the left is the connected component graph, which shows a connected group of images identified by the proposed method. On the right, the corresponding provenance graph obtained by pruning the links of the connected components graph by computing the maximum spanning tree (MST). MST removes all cycles within the graph while keeping the edges that maximize the sum of all content-sharing scores between each linked node.

Figure 14. Provenance Analysis at Document Level. The process of document-level provenance analysis begins with Filtering & Evidence Collection (1), followed by image-level analysis to identify relationships and graphs of the collected figures (2). Finally, it tracks related documents through their linked figures and creates a provenance visualization of them. The images used in this figure are public domain and were used only for illustration’s sake.

source documents. Figure 14 depicts the complete investigation from 1. Filtering & Evidence Collection; 2. Provenance Analysis at Image Level; and 3. Document Provenance Analysis.

The document analysis starts by creating a squared table $M$ with $n$ rows and $n$ columns, where $n$ is the size of the document dataset. Each row and column in this table represents a document from the dataset, and the value of a cell $d_{i,j}$ (row $i$, column $j$) represents the number of duplicated elements between document $i$ and $j$. The content-shared tables from the image level are used to populate $M$. For instance, if document $D_i$ shares $k$ elements with document $D_j$, the cell in row $i$ (representing $D_i$) and column $j$ (representing $D_j$) of $M$ is assigned the value $k$. This table can be interpreted as the adjacency matrix of the documents. Next, the document analysis identifies the connected components of $M$ similar to the image level. To improve the visualization of these components, a document provenance graph is created (as illustrated in Fig. 14, step 3). This graph represents the maximum number of shared elements between documents, similar to the graph constructed at the image level.

Through this analysis, we observed that articles produced by paper mills are connected in the same provenance graph, despite having unrelated authors or subjects, i.e., we could find the needles from the haystack.

References


**Data availability statement**

All datasets, annotations, trained models, and implementations from this research are freely available at https://github.com/phillipecardenuto/upm.

**Acknowledgements**

We thank São Paulo Research Foundation for supporting this research (FAPESP) under the thematic project DéjàVu, grants 2017/12646-3 and 2020/02211-2. Additionally, we extend our gratefulness to Dr. Elizabeth Bik for generously allowing us to use the Stock Photo Paper Mill while developing our dataset. Lastly, we would like to acknowledge the Retraction Watch team for providing and sharing their valuable database, which greatly contributed to our understanding of the current state of scientific integrity.

**Author contributions statement**

JP.C. collected and annotated the scientific figure panel dataset and developed the automated panel extraction model. He led the development of paper mill detection solution and experiments, the organization of the paper mill dataset, the analysis of the results, and the writing of the paper. D.M. assisted in designing all the experiments and executed some of the paper-mill detection experiments, and the analysis of the results. He also helped with writing the paper. A.R. advised on all tasks and helped with writing the text. All authors reviewed the manuscript.

**Additional information**

**Competing interests:** The authors declare no competing interests.