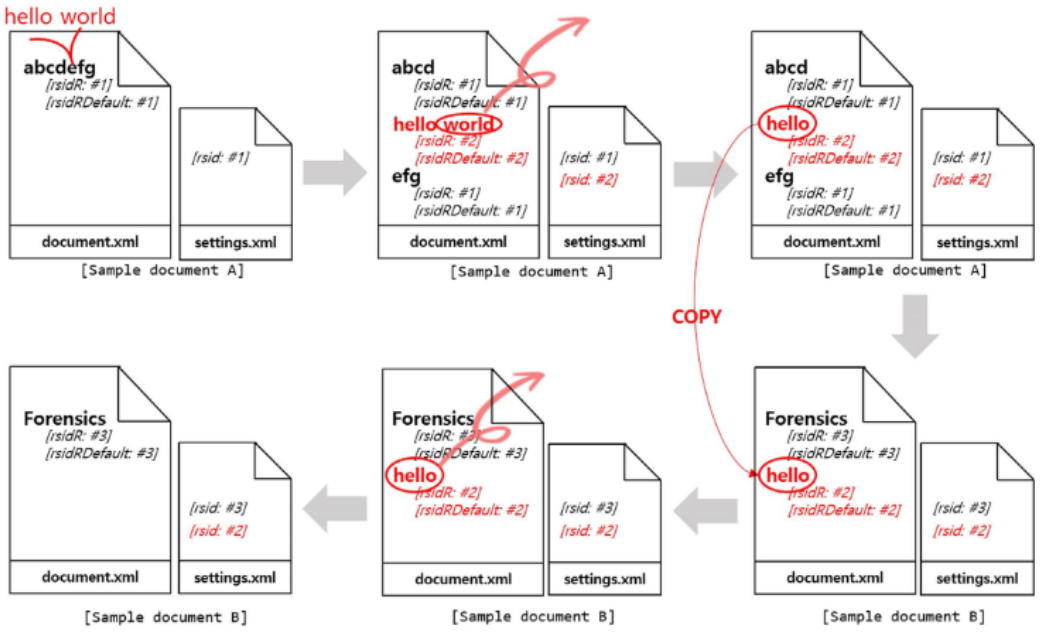

FIGURE 2 RSID rules (RSID status when a content is modified or copied) [Color figure can be viewed at wileyonlinelibrary.com]

grouping RSID, the RSID values in the default templates provided by MS Word should be excluded and analyzed.

3.2 | LibreOffice Writer

LibreOffice Writer document files are easier to investigate, because they are simpler than MS Word files, and their RSID types and rules are also straightforward. The LibreOffice Writer has a total of four types of RSID in the "content.xml" and "settings.xml" files. Paragraph-rsid and rsid are in the "content.xml" file, while RsidRoot and Rsid are in the "settings.xml" file.

The paragraph-rsid is the RSID generated when a short circuit is created and always occurs with the rsid. When a sentence in units smaller than a paragraph is created, rsid is assigned, and additionally when the font is changed. In addition, when the user copies the document, the paragraph-rsid is stored in the copy. RsidRoot and Rsid in the "settings.xml" file are created and maintained with a different rule than in MS Word, and the values change each time it is stored. Besides, unlike MS Word, the RSID for "content.xml" is not stored in the "settings.xml" file.

As with MS Word, if a file is copied, all RSIDs within two document files are the same. However, unlike MS Word, the RSID in LibreOffice remains, even when the copied content is deleted totally. RSID assigned once remains in the body of "content.xml" even if all characters are deleted. In LibreOffice Writer, since RSID is copied when the users copy content to another document even without changing the font, it is easy to trace the history of the copy.

3.3 | Other programs

In addition to MS Word and LibreOffice Writer, there are many different word processors to write and modify a document. This section describes how RSID values are created/modified while creating and modifying MS Word files on different word processors, such as MS Office Online Word, Office 365, Google Docs, and MS Office for Mobile (iOS).

When creating a new document using MS Office Online, a "document2.xml" file is created automatically instead of a "document.xml" file, and an RSID exists in it. In MS Word, the first two digits constituting 8 bits are fixed with "00", and the second six digits making up 24 bits are randomly generated, but when MS Office Online Word creates an RSID, the first two digits constituting 8 bits are assigned a value other than "00". When the user modifies the document content created with MS Office Online to MS Word, the "document2.xml" file is changed back to the "document.xml" file, and RSID is assigned as per the RSID generation rule in MS Word. Thus, if the "document2.xml" file exists inside the document, or if the first two digits of the RSID are assigned a hexadecimal number other than "00", the user can see that the file was created/modified by MS Office Online. A document created with Office365 creates two additional RSIDs in the "settings.xml" file that do not exist in

4.4 | Document tracking

If the files have the same hash value as another file, it means that they have been copied, but if anyone modifies or adds content after copying, it is difficult to determine whether the files are related. Similarity comparisons done using the author or last modified by attributes, cannot be used to specify documents as entirely copied. Thus, in this paper, using the rules of creation of RSIDs, it is possible to distinguish between file originals and copies, and to track the history of modification. If the target files share one or more of the same RSID, they are copied. Figure 5 shows the method for analyzing and tracking the history of a document when the hash values are different, and procedure is shown below:

1. Extract all RSIDs from the source file.
2. Extract all MS Word files from the PC or external storage of the analyzed target to collect the RSID values.
3. Navigate to files with the same RSID value as the original file.

There are numerous ways to track the history of a document. First, it is possible to determine the actions taken by the user to copy files or parts thereof. Methods for copying include the following: 1) copying the original file entirely and 2) using the "Save As" function in the original file. If the file is copied as in the first method, its RSIDs match the original files. However, while using the "Save As" method, one additional RSID is generated along with the RSIDs that the original file has, and its value is recorded in "settings.xml." By comparing the RSIDs of the original file with the list of the RSIDs of the copy file, we can see how they were copied. Moreover, if the other file's content is copied, RSIDs are copied and stored only in the "document.xml" file and not in the "settings.xml" file. The "settings.xml" file is assigned an RSID each time it is stored, and if there are more than three differences between the number of RSIDs in the "settings.xml" file and revision number in property information, the file is copied with the "Save As" option.

Next, if there is a file with the same RSID value, it can be seen that it was copied at least once. There are two main ways to analyze the tracking history of a document, depending on the situation. First, assume that the original file has not been modified when there are two files, A and B. If all the RSIDs in the "settings.xml" of file A exist in the "settings.xml" of file B with additional RSIDs, it means that the file A has not been modified after copy. Thus, the file containing less RSID values in the "settings.xml" is the original file. Second, when any same rsidRPr is in the A file and the B file, the sentence assigned to the rsidRPr is a copied sentence. To distinguish between the original and the copy, if the rsidRDefault and rsidR, which are the parent tags where rsidRPr is allocated, are the same as rsidRPr, it means it is the original file. But if rsidRPr is different as rsidRDefault and rsidR, then it means it was copied.

Finally, the user does not always write in the same order, so the content alone does not express the exact order in which it was created. However, RSID and metadata can be used to track the order in which the documents are created. The rsidRoot value in the "settings.xml" file indicates the first piece of content created by the user, so the first text in the "document.xml" file with the same rsidRoot RSID value is the user's first content. If there is a rsidRDefault entry in the document that has the same RSID value as rsidR, then the content with rsidR has been created or modified before rsidRDefault. If the content contains a rsidRPr entry, it means that the user modified the content to a different font type. The last sentence written and modified by the user is within the tags <bookmark Start> and <bookmark End>. In addition, if the user works with another word processor as described in the section "Revision Identifier (rsid)," the details of which word processor was used can be found.

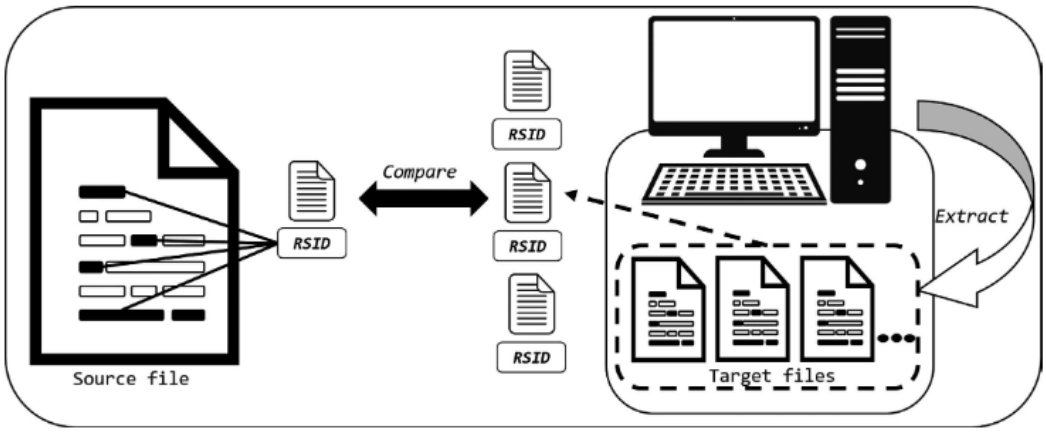

FIGURE 5 The document tracking procedure for finding the same RSID value in the storage device as target file

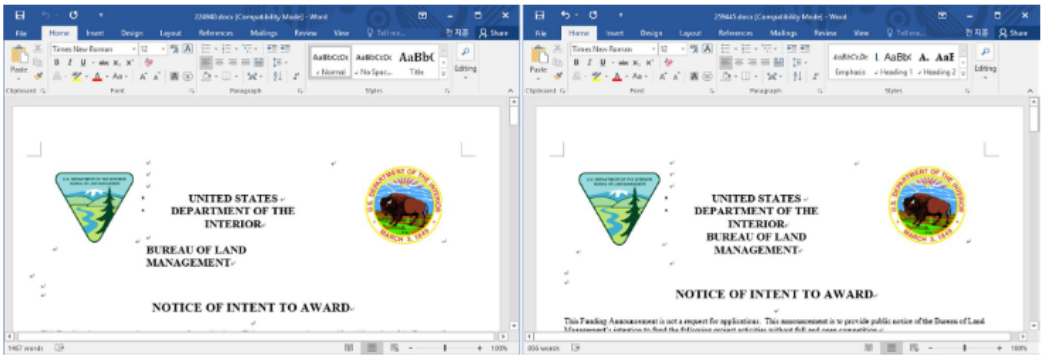

FIGURE 6 The document content of 224940.docx and 259445.docx files which are discovered an association by grouping using property (author, last saved by) [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Detail description of the example documents, 224940.docx and 259445.docx

Metadata	224940.docx	259445.docx
Company	Bureau of Land Management	U.S. Geological Survey
Content created	2008-06-10	2008-07-29
Date last saved	2008-06-10	2008-07-29
Last printed	N/A	2008-07-29
Authors	tthaler	Peter Lyttle
Last saved by	tthaler	tthaler
Title, subject, tags	N/A	N/A
Pages	4	3
Word count	1576	856
Total editing time	27 min	87 min
Size	93.0 KB	98.1 KB

found between two organizations that should not be relevant, the investigators can conclude that there is a relationship between the two.

5.3 | Tracking experiment

A document tracking experiment was conducted with minutes of meetings shared via e-mail to researchers. It describes how to find files related to minutes of meetings shared by the author on other researchers' PCs and how to track the history of the file. In this study, we proceeded by tracking minutes of the meeting file; however, actual confidential leakage events are carried out similarly, so adapting this procedure helps the investigation. The name of the file shared by the author is "[102019][DF]meeting_log.docx," and the following procedure shows how to track the files associated with it.

1. Extract RSIDs from "[102019] [DF]meeting_log.docx."
2. Extract MS Word files from PCs.
3. Identify files with the same RSID.

As a result of browsing the files in the PC, two files were discovered, and the path is as follows. After checking the hash value of the file, we found that the first file is a file that has the same hash value and is downloaded by e-mail. The hash value of the second file was not the same as that of the original minutes of the meeting. This happens when the original file is copied and modified, or only some of the content of the file is copied.

- C:/Users/naaya/downloads/[102019][DF]meeting_log.docx
- D:/DF/meeting_log/102019/work.docx

To begin with, the "work to do.docx" file, like Figure 7, is the content-copied file of the "[102019][DF]meeting_log.docx" file, not a file copy. Figure 8 shows the contents of the two documents and the RSID of the corresponding parts; although similar content exists as shown, investigators cannot be sure that they have been copied. However, since RSID "00CB0D5F" exists in both files, we can be convinced that the sentences that are in red font have been copied and modified. In the sentence, "Normal file," there are two different RSIDs, "00CB0D5F" and "005B1559," even though it is a single sentence. This means that the user wrote and saved it twice. As a result, despite the differences in content, RSID is the same, which indicates that a few modifications were made after copying the content.

If investigators search keywords, metadata, and compare similarities, they cannot find the file "work to do.docx." RSID is a unique value that is stored together when copied, using which they can find the file needed with certainty and efficiency. In cases of leakage of confidential information, suspects often delete the file, or change/add to the file name or file content before an investigation is conducted, denying the release of confidential files. Therefore, using the RSID-based document tracking method as proposed here may be helpful for the investigation.

19. da Cruz Nassif LF, Hruschka ER. Document clustering for forensic analysis: An approach for improving computer inspection. IEEE Trans Inf Forensics Secur. 2012;8(1):46–54. https://doi.org/10.1109/TIFS.2012.2223679.
20. Mundhe MRK, Maind A. Automatic labelling and document clustering for forensic analysis. Int J Recent Innovation Trends Comput Commun. 2014;2(9):2934–41.
21. Prachi KK, Phalke DA. Document clustering approach for forensic analysis: a survey. Int J Sci Res. 2014;3(12):1787–91.
22. Vaijayanthi RP. Digital forensic analysis through document clustering. Int J Recent Innovation Trends Comput Commun. 2017;5(7):245–8.
23. Dagher GG, Fung BC. Subject-based semantic document clustering for digital forensic investigations. Data Knowl Eng. 2013;86:224–41. https://doi.org/10.1016/j.datak.2013.03.005.
24. Chung H, Park J, Lee S. Forensic analysis of residual information in adobe pdf files. In: Park JJ, Yang L, Lee C, editors. Proceedings of the 6th International Conference of Future Information Technology (FutureTech 2011); 2011 June 28-30; Loutraki, Greece. New York, NY: Springer, 2011;100–9. https://doi.org/10.1007/978-3-642-22309-9_12.
25. Park J, Lee S. Forensic investigation of Microsoft PowerPoint files. Digit Investig. 2009;6(1–2):16–24. https://doi.org/10.1016/j.diin.2009.05.001.
26. Jeong D, Lee S. Study on the tracking revision history of MS word files for forensic investigation. Digit Investig. 2017;23:3–10. https://doi.org/10.1016/j.diin.2017.08.003.

How to cite this article: Joun J, Chung H, Park J, Lee S. Relevance analysis using revision identifier in MS word. *J Forensic Sci*. 2021;66:323–335. https://doi.org/10.1111/1556-4029.14584