Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.27, NO.3, Jun. 2017 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2017.27.3.489

OOXML 문서에 대한 향상된 데이터 은닉 및 탐지 방법*

홍 기 원,[1†] 조 재 형,[1] 김 소 람,[1] 김 종 성[1,2‡]
[1]국민대학교 금융정보보안학과, [2]국민대학교 정보보안암호수학과

Improved Data Concealing and Detecting Methods for OOXML Document*

Kiwon Hong,[1†] Jaehyung Cho,[1] Soram Kim,[1] Jongsung Kim[1,2‡]
[1]Dept. of Financial Information Security, Kookmin University,
[2]Dept. of Information Security, Cryptology and mathematics, Kookmin University

요 약

MS 오피스는 국내뿐만 아니라 세계적으로 널리 사용되는 오피스 소프트웨어이다. 여러 버전 중 MS 오피스 2007부터 최신 버전인 MS 오피스 2016까지 문서 구조에 OOXML 형식이 사용되고 있다. 이와 관련해 대표적인 안티-포렌식 행위인 데이터 은닉에 대한 방법이 연구, 개발되어 은닉된 데이터에 대한 탐지 방법은 디지털 포렌식 수사 관점에서 매우 중요하다. 본 논문에서는 기존에 발표된 OOXML 형식의 MS 오피스 문서에 데이터 은닉 및 탐지에 관한 두 가지 연구를 소개한 뒤 두 연구의 탐지 방법을 우회하는 데이터 삽입 방법과 MS 오피스 엑셀, 파워포인트의 데이터인 시트, 슬라이드 등을 은닉하는 방법을 제시한다. 이와 같은 방법으로 은닉된 데이터를 탐지할 수 있는 향상된 탐지 알고리즘 또한 제시한다.

ABSTRACT

MS office is a office software which is widely used in the world. The OOXML format has been applied to the document structure from MS office 2007 to the newest version. In this regard, the method of data concealing, which is a representative anti-forensic act has been researched and developed, so the method of detecting concealed data is very important to the digital forensic investigation. In this paper, we present an improved data concealing method bypassing the previewers detecting methods for OOXML formatted MS office documents. In addition, we show concealment of the internal data like sheets and slides for MS office 2013 Excel and PowerPoint, and suggest an improved detecting algorithm against this data concealing.

Keywords: Digital forensics, OOXML, MS office, Data concealing, Detection of concealed data

I. 서 론

최근 디지털 포렌식이 알려짐에 따라, 관련된 기술 정보들을 전문가가 아닌 사람도 쉽게 접할 수 있게 되었다. 이에 따라 안티-포렌식 행위도 늘어나 디지털 포렌식 수사에 어려움을 주고 있다. 데이터 은닉은 안티-포렌식 행위 중 하나로 디스크의 빈 공간, 파일 슬랙 영역, 파일 포맷의 특징 등을 이용하는 방법이 있다. 본 논문에서는 MS 오피스의 파일 포맷을 이용한 데이터 은닉 방법과 은닉된 데이터를 탐지하는 알고리즘을 제시한다.

MS 오피스 2007 이전 버전에서는 CFBF (Compound File Binary Format)을 사용한다. 그러나 문서의 가용성 등 여러 개선 사항을 이유로 MS 오피스 2007부터 OOXML(Office Open XML)을 사용하고 있으며, 포맷이 변경됨에 따라 오피스 파일의 구성이 완전히 바뀌었다. 기존

Received(02. 17. 2017), Modified(1st: 04 26. 2017, 2nd: 06.08. 2017) Accepted(06. 08. 2017)
* 이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(2017-0-00344, 최신 모바일 기기에 대한 암호해독 및 포렌식 분석).
† 주저자, utopia3@kookmin.ac.kr
‡ 교신저자, jskim@kookmin.ac.kr(Corresponding author)

Table 1. Difference between MS Office versions

	Previous MS Office 2007	MS Office 2007 ~ 2016
File format	Compound File Binary	Office Open XML
Document extension	xls, ppt, doc	xlsx, pptx, docx

CFBF 문서에서의 데이터 조작 및 은닉 방법이 OOXML에는 적용되지 않아 이와 관련된 새로운 방향의 연구가 필요하다.

2015년 기준으로 MS 오피스는 국내 설치형 오피스 시장 점유율의 약 71%를 차지하고 있으며, 가장 널리 사용되는 오피스 애플리케이션이다[1]. MS 오피스 문서를 네트워크상에서 전송, 공유하는 것은 비밀 정보를 공유하는 행위로 의심하기 힘들고 해당 문서에 데이터가 은닉되었다 하더라도 일일이 확인하기란 쉽지 않다. 따라서 데이터를 은닉하려는 사람의 입장에서 MS 오피스 문서를 이용해 데이터를 은닉하는 것은 타당한 접근이라 볼 수 있다. 만약 MS 오피스 문서에 데이터(파일)를 은닉하였거나 시트, 슬라이드와 같은 MS 오피스 문서 일부를 은닉하였다면 통합 포렌식 도구, 장비로는 이를 탐지할 수 없다. 또한 MS 오피스 문서의 '통합 문서 검사' 기능으로도 데이터 은닉과 관련한 특이점을 발견할 수 없다. 본 논문에서 제시하는 향상된 탐지 알고리즘은 기존의 탐지 방법을 우회하여 은닉한 데이터를 찾아낼 수 있고 탐지에 필요한 정보를 이용해 분석 시 시간을 단축시킬 수 있는 알고리즘이다. 이와 같은 은닉된 데이터를 탐지하는 알고리즘 개발과 자동화 도구 개발은 디지털 포렌식 관점에서 매우 의미가 있는 연구이다.

논문의 구성으로, 2장은 OOXML에 대한 간략한 소개를 다루며 3장은 MS 오피스 문서에서의 데이터 은닉 및 탐지에 대한 관련 연구를 소개한다. 4장에서는 관련 연구의 한계점과 이를 이용한 데이터 은닉 방법, 그리고 향상된 데이터 탐지 방법을 제시한다. 마지막 5장을 결론으로 본 논문을 맺는다.

II. OOXML 소개 및 관련 연구

2.1 OOXML(Office Open XML)

OOXML은 마이크로소프트(MS)가 개발하였으며, 국제 표준 ECMA-376과 ISO/IEC 29500으로 등록되어있다. OOXML 형식은 과거 MS 오피

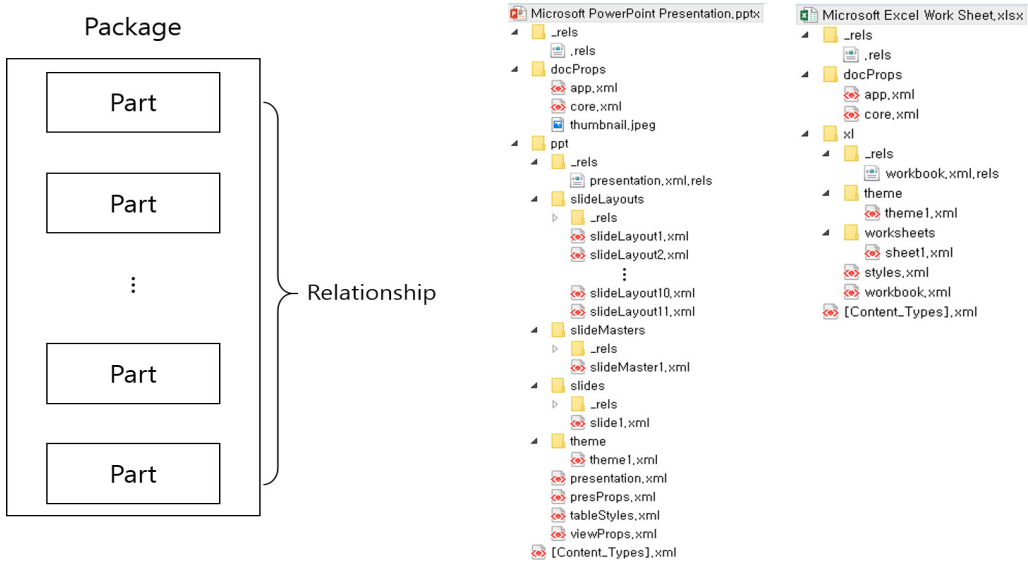

Fig. 1. Logical package and its structure model(left) and tree-structure of Excel and PowerPoint documents(right)

Table 2. Explanation of parts covered in this paper

Document type	Main part	Description
Excel(xlsx)	workbook.xml	An instance of this part type contains workbook data and references to all of its worksheets.
	workbook.xml.rels	The part containing relationship information for the worksheet part.
PowerPoint (pptx)	presentation.xml	An instance of this part type contains the definition for a slide presentation.
	presentation.xml.rels	The part containing relationship information for the presentation part.
Common	[Content_Types].xml	The part specifying information to identify parts in the package.
	app.xml	The part for document-specific unique properties at the application level.
	.rels	The relationship part that has basic relationship information for document execution. In this paper, we use it to insert a relationship of concealed data.

스 형식과 호환 가능하며, 다양한 도구와 플랫폼에서 개발할 수 있다. 또한 기업용 애플리케이션들 간에 운용성을 높이는 것이 목적이다. 표준에는 워드 프로세서(docx), 엑셀(xlsx), 파워포인트(pptx) 문서 형식이 정의되어 있고, MS 오피스 2007 버전부터 OOXML 형식이 적용되었다[2]. OOXML 형식의 문서들은 공통적으로 논리적 개체인 패키지(package)이며 ZIP 아카이브 형식을 갖는다[3]. 패키지는 여러 종류의 파트(part)와 파트 간의 관계를 정의하는 관계 파트(relationship part)로 구성되고 이와 같은 요소들을 통합해 하나의 오브젝트로 표현해 준다. 예를 들어, 어떤 문서의 한 그림은 패키지 내에 XML(eXtensible Markup Language) 문서 파트와 이미지 파트로 나누어져 있다. Table 2은 본 논문에서 제시하는 기술과 관련된 파트에 대한 설명이다.

파트는 바이트 스트림으로, 문서에 따라 정의된 콘텐츠 타입이 존재한다. MS 오피스 2007 버전 이상 문서는 대부분의 파트가 XML 형식으로 존재하고 콘텐츠 타입은 표준에 정의된 XML 스키마를 따르고 있다. 오피스 문서의 확장자를 .zip으로 바꾸어 압축을 해제하면 이를 확인할 수 있다.

패키지 내의 파트는 다른 한 파트 혹은 여러 파트들을 참조하거나 패키지 외부의 리소스를 참조할 수 있으며, 파트간의 참조 관계 역시 콘텐츠 타입으로 정의할 수 있다. 이 같은 타입의 파트를 관계 파트라 하며, 관계 파트는 패키지 내에 존재하는 '_rels' 폴더들에 있다.

2.2 관련 연구

본 절에서는 OOXML 형식이 적용된 MS 오피스 문서에 대한 연구 중 파트와 관계 파트에 관련된 데이터 은닉과 탐지에 대한 두 연구 결과를 소개한다.

2.2.1 Microsoft Office 2007 파일에 정보 은닉 및 탐지[4]

MS 오피스 문서는 응용프로그램으로 실행시켰을 때 OOXML에 정의된 형식이 맞는지 확인한다. 이때, 정의와 다른 부분이 있다면 오류가 발생해 파일이 열리지 않거나, 문서 내에 정상적이지 않은 내용을 삭제하여 해당 형식에 맞게끔 복구해 문서를 보여준다. 만약 파트 조작 없이 오피스 파일 내에 데이터(혹은 파일)를 삽입하였다면, 오류 혹은 복원 창이 팝업 되어 해당 문서는 정상적인 문서가 아님을 의심해볼 수 있다. B. Park 등은 MS 오피스 2007 버전에서 OOXML에 정의되어 있지 않은 데이터를 오류가 발생하지 않도록 은닉하고 이를 탐지하는 방법을 소개하였다[4].

Fig. 2의 위와 같이 오피스 문서에 내장되어 있

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Types xmlns="http://schemas.openxmlformats.org/package/2006/content-types">
    <Default Extension="bin"
        ContentType="application/vnd.openxmlformats-officedocument.spreadsheetml.printerSettings"/>
    <Default Extension="png" ContentType="image/png"/>
    <Default Extension="rels" ContentType="application/vnd.openxmlformats-package.relationships+xml"/>
    <Default Extension="xml" ContentType="application/xml"/>

    <Default Extension="zip" ContentType="application/zip"/>

    <Override PartName="/xl/workbook.xml"
```

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<Relationships xmlns="http://schemas.openxmlformats.org/package/2006/relationships">
    <Relationship Id="rId3" Type="http://schemas.openxmlformats.org/officeDocument/2006/relationships/extended-properties" Target="docProps/app.xml"/>
    <Relationship Id="rId2" Type="http://schemas.openxmlformats.org/package/2006/relationships/metadata/core-properties" Target="docProps/core.xml"/>
    <Relationship Id="rId1" Type="http://schemas.openxmlformats.org/officeDocument/2006/relationships/officeDocument" Target="xl/workbook.xml"/>
    <Relationship Id="rId4" Type="http://schemas.openxmlformats.org/officeDocument/2006/relationships/custom-properties" Target="docProps/custom.xml"/>

    <Relationship Id="rId100" Type="http://schemas.openxmlformats.org/officeDocument/2006/relationships/custom-properties/a" Target="xl/media/hidden.zip"/>
```

Fig. 2. Modified '[Content_Type].xml' part(upper) and modified '.rels' part(lower)

는 [Content_Type].xml 파트에는 패키지 내에 존재하는 파트들의 확장자가 명시되어있다. 은닉하고자 하는 파일이 zip 확장자를 갖는다면, Fig. 2에 표시된 부분과 같이 확장자를 넣어 준다. 그리고 Fig. 2의 아래와 같이 관계 파트에 표시된 내용을 추가, 즉 은닉된 파일에 대해 관계 파트에 관계 Id를 할당하고 타깃으로 지정한다. 그러면 MS 오피스 응용프로그램은 은닉된 데이터가 있는 문서를 실행해도 정상적인 파일로 인식하게 된다.

B. Park 등은 이와 같이 MS 오피스 문서와 관련 없는 파일 탐지뿐만 아니라 '.xml' 확장자를 갖는 파트에서 주석을 탐지하는 알고리즘과 의사 코드를 제시하였다. 먼저 탐지할 문서를 불러와 압축 해제하여 각 파트와 관계 파트에 대한 정보를 얻는다. 은닉된 파일에 대해서는 각 파트들이 관계 파트 내에서 타깃으로 존재하는지 확인한다. 만약 파트가 타깃으로 존재하지 않는다면, 은닉된 데이터로 판단한다. 그 다음으로, 타깃의 타입이 OOXML 표준에 정의되어 있지 않다면, 이를 은닉된 데이터로 판단한다. 만약 은닉된 데이터가 관계 파트이면, 해당 파트 내에 명시되어 있는 모든 타깃이 은닉된 데이터가 된다. 그리고 '.xml'을 확장자로 갖는 파트는 주석을 이용하여 데이터를 은닉할 수 있다. 하지만 기본적으로 MS 오피스 문서의 파트는 주석을 포함하지 않기 때문에, 주석이 있는 경우 사용자가 고의적으로 데이터를 삽입한 것으로 판단한다.

2.2.2 OOXML 형식의 문서에 대한 포렌식 조사[5]

Z. Fu 등은 2011년에 OOXML 형식이 적용된 문서에 대한 포렌식 조사 방법을 소개하였으며, 조작하지 않은 패키지 내부의 파트는 데이터 은닉 혹은 그 외의 목적으로 조작된 파트와 차이가 있음을 보였다. 먼저 패키지 내부 파트를 수정하지 않은 상태는 Fig. 3과 같다. Fig. 3의 위는 대표적인 디지털 포렌식 통합 분석도구인 EnCase 7.10으로 본 결과이고, 아래는 윈도우 탐색기에서 확인할 수 있는 부분이다. EnCase로 본 파트의 시간 정보는 '1980년 1월 1일 오전 12시'가 기본 값, 윈도우 탐색기 상에서 확인할 수 있는 시간 정보는 비어있다. 하지만 조작된 파트는 이와 다른 정보가 남게 된다. Fig. 4는 문서 내부의 파트를 조작한 결과이다. 직관적으로 시간 정보의 변화를 알 수 있다. 일반적인 사용자는 문서 편집 시 문서의 확장자를 바꾸어 압축을 해제해 특정 파트를 조작하지 않고, 관련된 오피스 응용프로

	Name	File Ext	Logical Size	Category	Last Written
1	printerSettings		288	Folder	01/17/17 10:14:15 오후
2	theme		152	Folder	01/17/17 10:14:15 오후
3	worksheets		248	Folder	01/17/17 10:14:15 오후
4	_rels		280	Folder	01/17/17 10:14:15 오후
5	sharedStrings.xml	xml	238	Document	01/01/80 12:00:00 오전
6	styles.xml	xml	1,574	Document	01/01/80 12:00:00 오전
7	workbook.xml	xml	717	Document	01/01/80 12:00:00 오전

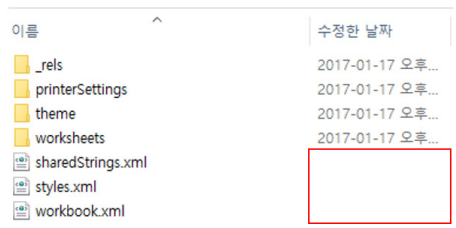

Fig. 3. Time information of part in package

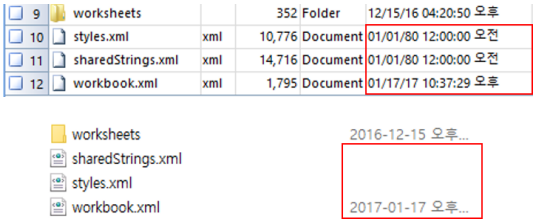

Fig. 4. Time information of modified part in package

그램을 통해 수정한다. 즉, 해당 시간 정보는 조작 행위와 직접적으로 연관된 것으로 디지털 포렌식 증거로 큰 의미가 될 수 있다.

2.3 기존 탐지 알고리즘의 한계점

앞서 소개한 관련 연구의 탐지 알고리즘에는 한계점이 존재한다. 먼저 Table 3은 3.2에서 제시할 향상된 탐지 알고리즘의 기능과 기존의 탐지 알고리즘의 기능을 비교한 것이다.

탐지 알고리즘의 한계점을 이해를 위해 가상의 두 가지 시나리오를 서술한다. 시나리오는 공통적으로 윈도우 기반의 PC를 사용, OOXML 포맷이 적용된 MS 오피스(2007 이상)를 사용한 것으로 간주한다.

시나리오 1 - MS 오피스 문서에 데이터 은닉

한 기관에서 기밀문서가 유출되어 해당 소속의 PC와 개인 저장장치에 대한 분석을 의뢰하였다. 그 중에서 유력 용의자로 짐작되는 PC에서 기밀문서(hwpx 포맷의 한글 파일)에 대한 접근 흔적과 USB 장치 연결에 대한 흔적이 발견되었다. 하지만 USB에는 한글 문서에 대한 흔적이 없고 수많은 MS 오피스 워드 문서들만 존재하였다. 이중에서 몇몇 문서들이 용의자의 PC에서 공유 드라이브에 업로드 되었지만 해당 문서들의 내용 또한 기밀문서와 관련이 없었다.

업로드 된 문서들은 공통적으로 용량이 크고 최근 수정된 날짜가 유출 시기와 비슷하였지만 기밀문서와의 관련성을 찾지 못하고 있는 상황이다.

시나리오 2 - MS 오피스 문서의 데이터 은닉

한 병원에서 MS 오피스 엑셀을 이용하여 물품, 약품 관리 대장을 작성해왔다. 문서들 중에는 코데인, 모르핀, 옥시코돈 등 마약성 진통제 입출에 대한 내용이 들어있는데, 기입된 내용과 다르게 수개월 전부터 외부인에게 넘겨졌다는 의혹이 있어 조사를 받게 되었다. 피의자 측에서 제출한 마약성 물질 입출에 대한 증빙 자료와 다른 장부가 존재할 수 있다. 한편, 내부 근무자의 인터뷰를 통해 확인한 결과 관리 대장 문서를 백업하여 관리한다는 것을 알아내었다. 확보한 사본 문서들 중 유독 몇 개의 파일의 용량이 큰 점을 파악했고 다른 문서들과 다르게 최근 수정 시간이 백업 시기가 아니었다.

해당 문서들에 대해 기존에 공개되어 있는 MS 오피스 문서에서 은닉된 데이터를 탐지하는 방법을 적용하여 분석하였지만 은닉된 데이터에 대한 흔적을 찾을 수 없었다.

두 시나리오를 통해 알 수 있는 관련 연구의 한계점은 다음과 같다. 먼저 2.2.1의 탐지 알고리즘은 파트와 관련해 두 가지, 관계파트에 타깃으로 지정되었는지 그리고 [Content_Type].xml에 정의된 확장자를 확인하여 은닉된 파일을 탐지하고 있다. 따라서 시나리오 1의 용의자가 관련 연구의 데이터 은닉 방법대로 기밀문서(hwpx 파일)를 MS 오피스 문서에 은닉하였다면 탐지 가능하다. 하지만 기밀문서의 확장자를 '.xml'로 변경하여 MS 오피스 문서 내에 삽입하고, 관계 파트에 이를 타깃으로 하는 〈Relationship〉 엘리먼트를 추가하였다면 제시한 탐지 알고리즘을 회피할 수 있다. 또한 시나리오 2와 같이 MS 오피스 문서의 데이터 은닉 역시 해당 탐지 알고리즘을 회피할 수 있다. 내부에 이미 존재하는 (4.3절에서 소개할 시트 및 슬라이드 관련 파트 등 은닉 대상)파트는 이미 '.xml' 확장자를 가지

Table 3. A comparison of detecting algorithms

	Time information	Internal data	External data	Comment string
Detecting algorithm of [4]	×	×	△	△
Detecting algorithm of [5]	○	×	×	×
Our detecting algorithm	○	○	○	○

며 몇몇 파트의 내용 수정으로 MS 오피스 문서의 데이터 은닉이 가능하다.

2.2.2의 탐지 알고리즘은 MS 오피스 문서의 파트를 조작 시 남게 되는 정보를 이용하여 데이터 은닉과 같은 정황을 탐지한다. 따라서 시나리오 1과 시나리오 2의 은닉된 데이터 모두 탐지 가능하다. 하지만 ZIP 아카이브에 대한 구조는 공개되어 있고, HxD, 010 Editor와 같은 헥사 에디터의 ZIP 아카이브의 템플릿을 활용한다면 시간 정보를 초기화할 수 있다. 시간 정보 초기화에 대한 설명은 010 Editor를 사용하여 실험, 확인하였기 때문에 이를 기준으로 설명한다.

먼저 데이터 은닉에 사용된 MS 오피스 문서에 ZIP 아카이브 템플릿을 적용하여 그 구조를 보면, 패키지 내 파트들의 시간 정보를 쉽게 구분하여 볼 수 있고 수정할 수 있다. Fig. 5와 같이 파트의 DOSTIME과 DOSDATE 각각 두 바이트가 시간 정보를 갖는 영역이다. DOSTIME은 초기값인 '00 00'으로, DOSDATE는 초기값인 '21 00'으로 바꾸어 저장하면 파트의 시간 정보만 초기화시킬 수 있다. 즉, Fig. 4와 같이 데이터를 조작한 시점의 시간 정보를 Fig. 3의 상태로 초기화한 것이다. 이와 같은 작업을 진행하면 2.2.2의 관련 연구에서 제시한 시간 정보를 활용한 은닉된 데이터 탐지를 회피할 수 있다. 즉, 파트의 시간 정보만으로 은닉된 데이터를 탐지하는 것에는 한계가 있어 이와 더불어 각 파트의 데이터를 분석해야 한다.

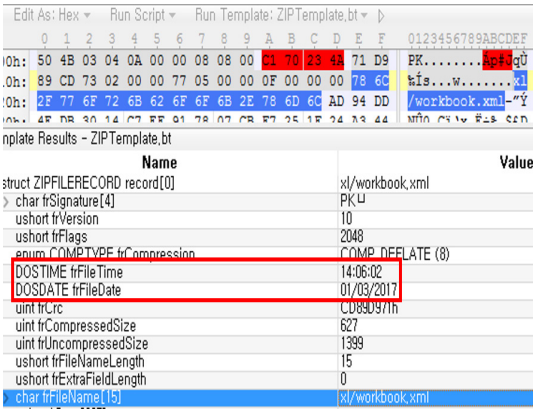

Fig. 5. The time value of modified part (e. g. workbook.xml)

III. 새로운 데이터 은닉 방법과 향상된 탐지 알고리즘

본 장에서는 Table 3의 'Internal data'와 관련된 새로운 데이터 은닉 방법을 소개하고, 이와 같이 은닉된 데이터 탐지뿐만 아니라 2.3에서 언급한 한계점을 보완한 향상된 탐지 알고리즘을 소개한다.

3.1 새로운 데이터 은닉 방법

2.2.1 관련 연구의 한계점에서 언급했던 외부 파일의 확장자를 '.xml'로 변경하여 MS 오피스 문서에 은닉하는 것은 OOXML 포맷이 적용된 MS 오피스에서 공통적으로 적용 가능하다. 즉, 이 방법은 엑셀(xlsx), 파워포인트(pptx)뿐만 아니라 워드(docx), 비지오(vsdx) 등과 같이 OOXML이 적용된 문서들에서 모두 가능하다. 관련 연구에서 제시된 방법과 다르게 [Content_Type].xml 파트를 수정하지 않고 확장자만 변경하여 파일을 삽입하더라도 문서를 여는 데에 아무런 지장이 없다. 기존의 탐지 방법을 회피하도록 관계 파트에 타깃으로 지정하여도 역시 문제가 발생하지 않는다.

시나리오 2에서 적용한 MS 오피스 문서의 데이터 은닉은 엑셀과 파워포인트에서 가능함을 확인하였다. 먼저 엑셀의 시트를 기준으로 설명하면 데이터를 은닉할 엑셀 문서의 확장자를 .zip으로 바꾸어 압축 풀기를 한다. 압축 풀기 후 생성된 폴더 내부에 'xl' 폴더를 확인할 수 있다. 그중에서 은닉할 시트에 대한 정보가 있는 \xl\workbook.xml 파트를 수정하여 저장한다. 예를 들어 엑셀 문서의 두 번째 시트(기본 시트 이름 - sheet2)를 은닉한다면, Fig. 6과 같이 workbook.xml 파트에서 시트 정보를 삭제한다. 수정한 파트를 저장한 후 다시 압축하여 확장자를 .xlsx로 바꾸어 주면 시트가 은닉된 것을 확인할 수 있다. 은닉된 시트를 다시 보이도록하기 위

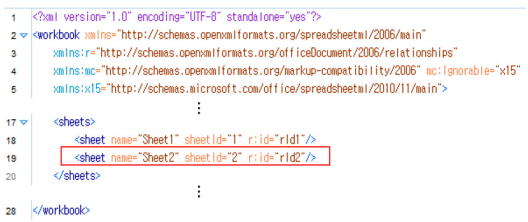

Fig. 6. 'workbook.xml' part in MS office Excel

Table 4. The file path associated with data concealing and necessary to modifying

Document	Concealing data	Related file path	Note
Excel(xlsx)	sheet	\xl_rels\workbook.xml.rels	Option
		\xl\workbook.xml	Necessary
Power point(pptx)	slide	\ppt_rels\presentation.xml.rels	Option
		\ppt\presentation.xml	Necessary
	slide master	\ppt_rels\presentation.xml.rels	Option
		\ppt\presentation.xml	Necessary
	slide note	\ppt\slide_rels\'Target part'	Necessary
Common		\[Content_Types].xml	Option
		\docProps\app.xml	Option
		_rels\.rels	Necessary

해서는 파트에서 삭제된 부분을 원래대로 돌려놓으면 된다.

파워포인트의 경우도 엑셀과 동일하게 확장자를 .zip으로 바꾸어 압축을 푼다. 생성된 폴더를 열어보면 'ppt' 폴더가 있고 \ppt\presentation.xml 파트에서 은닉할 데이터와 관련된 파트를 수정하여 저장한다. 예를 들어, 첫 번째 슬라이드를 은닉한다고 가정한다. presentation.xml 파트의 엘리먼트 〈p:sldId〉 중에서 가장 위에 있는 〈p:sldId〉가 첫 번째 슬라이드이다. 참고로, 해당 엘리먼트의 어트리뷰트 r:id 값은 보통 "rid2"이지만, 마스터 슬라이드 r:id를 부여한 뒤에 슬라이드에 r:id 값을 부여하기 때문에 고정되어 있는 값은 아니다. 이를 바탕으로 은닉할 슬라이드에 해당하는 엘리먼트를 삭제한다. 수정한 내용을 저장한 뒤 다시 압축하여 확장자를 .pptx로 변경하면, 슬라이드가 은닉된 것을 확인할 수 있다. 게다가 슬라이드뿐만 아니라 마스터 슬라이드, 슬라이드 노트 또한 은닉 가능하다. 은닉된 데이터를 다시 보이도록하기 위해서는 파트에서 삭제된 부분을 원래대로 돌려놓으면 된다.

MS 오피스에 외부 파일을 은닉하는 방법과 MS 오피스 문서의 내부 데이터인 시트, 슬라이드, 마스터 슬라이드, 슬라이드 노트가 은닉 가능한 것을 확인하였다. 그중 일부 데이터에는 텍스트뿐만 아니라 표, 사진, 동영상이 포함될 수 있고 이 역시 실험과정에서 은닉되는 것을 확인하였다. 기존의 탐지 알고리즘을 회피하면서 외부 파일 혹은 MS 오피스의 데이터 은닉을 위해 수정하는 파트를 Table 4에 정리하였다. 네 번째 열은 파트의 수정이 필수인지 구분한 것이다. Option인 항목은 은닉한 파트와 관련된 정보가 있지만, 이를 수정하지 않아도 오피스 응용프로그램으로 열었을 때와 데이터 은닉에 문제가 없다.

3.2 향상된 은닉 데이터 탐지 알고리즘

본 절에서는 4.3절에서 제시한 데이터 은닉에 대한 탐지와 기존 탐지 알고리즘의 한계점을 보완한 탐지 알고리즘을 제시한다.

향상된 탐지 알고리즘은 Fig. 7과 같다. 이 과정을 단계별로 살펴보면, 우선 문서를 바이너리 분석으로 각 파트의 시간 정보가 초기 값 '00 00 21 00'과 같은지 확인한다. 시간 값이 일치하지 않으면 은닉된 데이터가 있는 것으로 판단한다. 그 다음 단계로는 자세한 분석을 위해 문서를 압축 해제하여 모든 파트에서 각 단계별 분석에 필요한 정보를 얻는다. 분석은 문서 외부 파일을 은닉했는지(1^{st} step)와 문서 내부의 데이터를 은닉했는지(2^{nd} step)에 대한 탐지로 구분된다.

확장자를 '.xml'로 변경해 외부 파일을 은닉한 경우, XML 파일의 시작 문자열과 비교해 구분할 수 있다. '〈?xml version="1.0" encoding="UTF-8" standalone="yes"?〉'는 XML 파일의 시작 문자열로, 보통 파일의 헤더와 구분되는 점을 활용하여 외부 파일의 확장자를 수정하여 삽입하였는지 판단하게 된다. 다음으로 관계 파트의 타깃들이 OOXML 표준에 명시된 타입인지 확인한다. 만약 표준에 없는

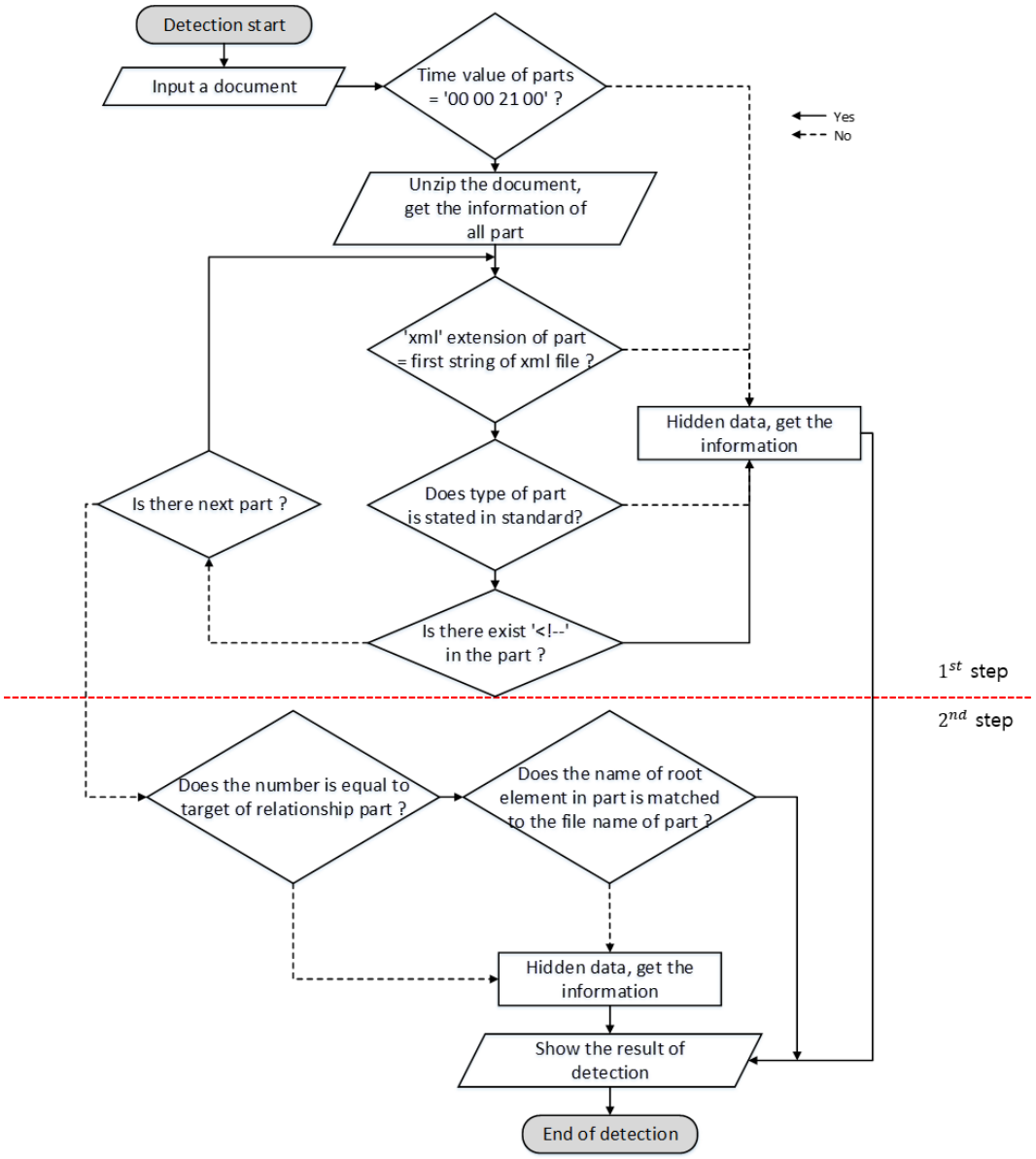

Fig. 7. Our improved detection algorithm

타입이면 이는 사용자가 삽입한 데이터로 판단하게 된다. XML 문서의 경우 주석을 이용하여 문자열 은닉을 고려할 수 있는데, MS 오피스 문서 내에 기본적으로 생성되는 XML 문서에는 주석이 없다. 따라서 '〈!--주석 내용--〉'이 있다면 은닉 데이터로 판단한다.

엑셀과 파워포인트에서 시트, 슬라이드 등 데이터 은닉에 대한 탐지는 각 관계 파트에서 타깃으로 참조되는 수와 실제 파트에 대한 파일의 수가 일치하는지 확인한다. 즉, 엑셀과 파워포인트 각 관계 파트 'workbook.xml.rels', 'presentation.xml.rels'에서 〈Relationship〉 엘리먼트의 'target' 어트리뷰트를 확인한다. 그리고 실제 파트의 파일이 타깃으로 정해져 있는 만큼 존재하는지 비교한다. 타깃 파트의 수와 실제 파트인 파일의 수가 일치하지 않으면 데이터를 은닉하였다고 판단할 수 있다. 하지만 수가

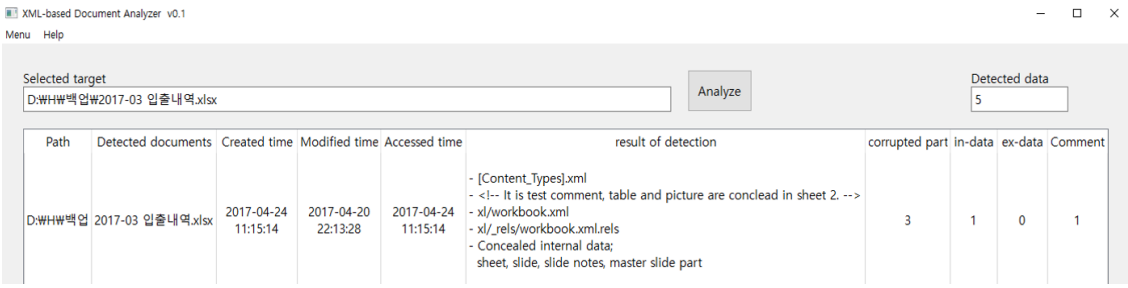

Fig. 8. Screen shot of our data detection tool implementing the algorithm described in Fig. 7

Table 5. part and part's element information

Concealed data type	File name of part	Root element name of the file
sheet	sheet#.xml	〈worksheet〉
slide	slide#.xml	〈p:sld〉
slide master	slideMaster#.xml	〈p:sldMaster〉
slide note	notesSlide#.xml	〈p:notes〉

일치하더라도 사용자가 은닉한 파트를 다른 폴더로 이동시키거나 파일 이름을 변경할 수 있다. 따라서 Table 5의 정보와 같이 파트의 파일 이름과 파일의 루트 엘리먼트 이름이 맞지 않으면 사용자가 파트를 다른 폴더로 옮겼거나 파트의 이름을 변경한 것으로 판단할 수 있다.

Fig. 8은 검증을 위해 만든 엑셀 문서로, MS 오피스 프로그램 상으로 시트가 하나 존재한다. 하지만 내부 구조를 보면 'sheet1.xml'과 'sheet2.xml' 두 개의 파트가 존재함을 확인할 수 있다. 이를 앞서 제시한 알고리즘을 기반으로 제작한 은닉 데이터 탐지 도구로 분석한 결과 Fig. 9와 같다. 도구는 단일 파일과 폴더 내 OOXML 문서를 분석할 수 있으며,

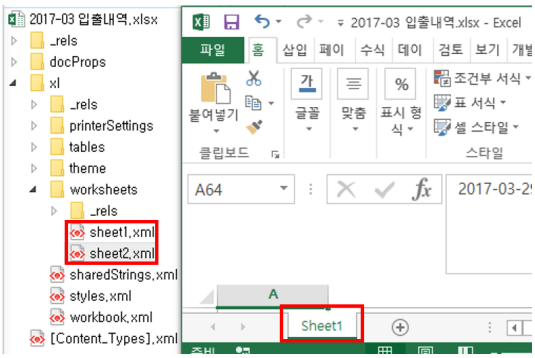

Fig. 9. Concealed sheet of Excel document and the sheet part in the document

분석 대상의 경로, 이름, MAC(Modified, Accessed, Created) 시간 정보, 탐지 결과를 확인할 수 있다. 분석에 대한 설명으로, 'corrupted'는 파트를 수정하는 과정에서 남는 정보를 발견한 것이다. 'in-data'는 시나리오 2에 해당되는 내부 데이터를 탐지한 것이고 'ex-data'는 외부 데이터를 문서 내에 삽입한 것을 탐지한 것이다. 마지막으로 'comment'는 XML 파일에 주석을 삽입한 경우 이를 탐지한 것이다. 보다 더 자세한 결과는 'result of detection'에 표현된다. 해당 그림은 엑셀 파일을 분석한 결과 시트를 숨기고 그 과정에서 파트를 수정한 것을 확인할 수 있다. 또한 주석이 하나 발견되었다는 것을 알 수 있으며, 해당 주석 문자열을 보여준다.

IV. 결 론

MS 오피스 중 엑셀과 파워포인트는 가장 널리 사용되는 문서 형식이다. 이런 문서를 이용해 비밀 정보를 은닉하고 공유하는 것은 메일, 메신저, 클라우드 서비스 등 네트워크를 활용하면 쉽게 가능하다. 그렇기 때문에 비밀 정보를 오피스 문서에 은닉하는 안티-포렌식 행위에 대응하기 위한 연구는 디지털 포렌식 관점에서 매우 중요하다.

본 논문에서는 MS 오피스 문서에 대한 데이터

은닉과 향상된 탐지 알고리즘을 제시하였다. 먼저 이와 관련하여 OOXML 포맷의 특징과 기존 연구의 데이터 은닉 및 탐지 방법을 소개하였다. 하지만 기존 알고리즘의 한계점이 존재하며, 이를 우회할 수 있는 새로운 데이터 은닉 방법이 있음을 확인하였다. 해당 내용으로는 은닉할 문서 외부 파일을 OOXML 표준의 확장자인 '.xml'로 수정하여 삽입하는 것이었다. 그리고 MS 오피스 엑셀과 파워포인트의 문서 내부 데이터인 시트와 슬라이드 등을 은닉하는 것이었다. 또한 이 과정에서 수정한 파트의 흔적을 010 Editor를 이용해 지울 수 있음을 보였다. 이와 같이 기존 탐지 알고리즘의 한계점을 이용하여 데이터를 은닉하더라도 본 논문에서 제시한 탐지 알고리즘은 은닉된 데이터를 탐지할 수 있었고, 이를 도구로 제작하여 검증하였다.

향후 연구로 MS 오피스의 다른 문서와 더불어 XML을 활용하는 문서인 HWPML, ODF 표준 등에 본 연구의 결과가 적용되는지 확인할 필요가 있다. 또한 MS 오피스는 윈도우 OS 호환 제품만이 아니라 맥 OS용 제품도 판매되고 있는데, 맥 OS는 시스템 운영 환경이 다르기 때문에 해당 OS에서의 연구도 필요하다.

References

[1] Wujin Jang, Yujung Han, "Company Report - Hangul and Computer (030520)", KB security, Jul. 2016

[2] ECMA-376-1 :2016, Edition Office Open XML File Formats - Fundamentals and Markup Language Reference, ECMA International Publication, Oct. 2015

[3] ECMA-376, 4th Edition Office Open XML File Formats - Open Packaging Conventional, ECMA International Publication, Dec. 2012

[4] Bora Park, Jungheum Park, Sangjin Lee, "Data concealment and detection in Microsoft Office 2007 files", Digital Investigation, vol. 5, pp. 104-114, Dec. 2008

[5] Zhangjie Fu, Xingming Sun, Yuling Liu, Bo Li, "Forensic Investigation of OOXML format documents", Digital Investigation, vol. 8, pp. 48-55, Apr. 2011

〈저자소개〉

홍 기 원 (Kiwon Hong) 학생회원
2016년 2월: 국민대학교 수학과 졸업
2016년 3월~현재: 국민대학교 금융정보보안학과 석사과정
〈관심분야〉 디지털 포렌식, 정보보호

조 재 형 (Jaehyung Cho) 학생회원
2015년 8월: 국민대학교 수학과 졸업
2015년 9월~현재: 국민대학교 금융정보보안학과 석사과정
〈관심분야〉 정보보호, 암호 알고리즘, 디지털 포렌식

김 소 람 (Soram Kim) 학생회원
2016년 2월: 국민대학교 수학과 졸업
2016년 3월~현재: 국민대학교 금융정보보안학과 석사과정
〈관심분야〉 디지털 포렌식, 정보보호

김 종 성 (Jongsung Kim) 종신회원
2000년 8월/2002년 8월: 고려대학교 수학 전공 학사/이학석사
2006년 11월: K.U.Leuven, ESAT/SCD-COSIC 정보보호 전공 공학박사
2007년 2월: 고려대학교 정보보호대학원 공학박사
2007년 3월~2009년 8월: 고려대학교 정보보호기술연구센터 연구교수
2009년 9월~2013년 2월: 경남대학교 e-비즈니스학과 조교수
2013년 3월~2017년 2월: 국민대학교 수학과 부교수
2014년 3월~현재: 국민대학교 일반대학원 금융정보보안학과 부교수
2017년 3월~현재: 국민대학교 정보보안암호수학과 부교수
〈관심분야〉 정보보호, 암호 알고리즘, 디지털 포렌식