분할된 Shannon 엔트로피 값을 이용한 파일 암호화 판별 정확성 향상에 대한 연구

고주성*, 곽진**
*아주대학교 컴퓨터공학과 정보보호응용및보증연구실
**아주대학교 사이버보안학과
e-mail:*jsko.isaa@gmail.com, **security@ajou.ac.kr

Accuracy Enhancement of Determining File Encryption Status through Divided Shannon Entropy

Ju-Seong Ko*, Jin Kwak**
*ISAA Lab., Department of Computer Engineering, Ajou University.
**Department of Cyber Security, Ajou University.

요 약

랜섬웨어는 사용자의 중요 파일을 암호화한 후 금전을 요구하는 형태의 악성코드로, 전 세계적으로 큰 피해를 발생시켰다. 안드로이드 환경에서의 랜섬웨어는 앱을 통해 동작하기 때문에, 앱의 악의적인 암호화 기능 수행을 실시간으로 탐지할 수 있는 방안에 대한 연구들이 진행되고 있다. 자원 제한적인 안드로이드 환경에서 중요한 파일들에 대한 암호화 수행 여부를 실시간으로 탐지하기 위한 방안으로 Shannon 엔트로피 값 비교가 있다. 하지만 파일의 종류에 따라 Shannon 엔트로피 값이 크게 달라질 수 있으며, 암호화 기능 수행에 대한 오탐이 발생할 수 있다. 따라서 본 논문에서는 파일에 대한 분할된 Shannon 엔트로피 값을 측정하여 암호화 기능 수행 탐지의 정확성을 높이고자 한다.

1. 서론

2017년 매 분기에서 가장 많이 발생한 악성코드는 랜섬웨어이며, 안드로이드를 포함한 다양한 환경에서 발생하였다[1]. 랜섬웨어는 사용자의 중요 데이터를 암호화한 후 복호화에 대한 대가로 금전을 요구하는데, 금전을 지불했을 때 파일을 복구할 수 있는지에 대한 여부는 불확실하다. 랜섬웨어로 인해 파일이 암호화된 경우 이를 복구하는 것은 매우 어렵기 때문에 사전에 예방하는 것이 가장 중요하다[2].

안드로이드 환경에서 랜섬웨어 앱의 동작을 실시간으로 탐지하기 위한 연구들이 진행되고 있으며, 그 예로는 Jing Chen 등의 연구[3] 등이 있다. 해당 연구의 첫 단계는 모니터링 중인 중요 파일들이 암호화되었는지 실시간으로 검사하며, 이를 위해 Shannon 엔트로피를 이용하고 있다.

하지만 Shannon 엔트로피를 이용할 경우 파일에 따라 엔트로피 측정값이 일정하지 않다. 평문/암호문을 구분하는 엔트로피 기준값을 정하더라도 평문의 엔트로피값이 기준보다 높거나 암호문의 엔트로피 값이 기준보다 낮은 예외적인 상황이 발생할 수 있다. 따라서 엔트로피 값 비교를 통한 암호화 판별의 정확성을 향상시킬 방안이 필요하다.

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2017R1E1A1A01075110).

본 논문의 2장에서는 Shannon 엔트로피와 랜섬웨어의 암호 알고리즘 및 접근 파일 종류에 대한 관련 연구를 다룬다. 3장에서는 분할된 Shannon 엔트로피 값을 측정하여 파일 암호화 판별의 정확성을 향상시킬 수 있는 방안을 제시하며, 4장에서는 향후 연구에 대해 언급하며 결론을 맺는다.

2. 관련 연구

2.1 Shannon 엔트로피

Shannon 엔트로피는 파일 내의 값 분포를 이용해 암호화 여부를 판단하는 데에 사용될 수 있다. 암호 알고리즘의 성질 중 혼돈에 의해 암호문의 값 분포는 평문에 비해 균일해진다[4]. 따라서 특정 파일의 Shannon 엔트로피 값이 높은 경우, 해당 파일은 암호화된 것으로 판단할 수 있다. 엔트로피 값은 최대 8을 가지며, 다음의 연산을 통해 엔트로피 값을 측정할 수 있다.

$$e = \sum_{i=0}^{255} P_{B_i}(-\log_2 P_{B_i}),\ P_{B_i} = \frac{F_i}{filesize}$$

(F_i는 파일 내 i의 개수이며, filesize는 파일의 크기(바이트 단위)이다.)

2.2 랜섬웨어에 이용되는 암호 알고리즘 및 파일 종류

암호 알고리즘과 암호화 타겟은 주요 랜섬웨어들이 이용했던 것들을 사용한다. 킬디스크, 워너크립터 등은 암호 알고리즘으로 Triple-DES CBC, AES-128 CBC, AES-256 CBC를 이용했으며[5] 주로 문서 파일들에 접근

하였다[4]. <표 1>은 주요 랜섬웨어들이 이용한 암호 알고리즘과 접근 파일 종류들을 정리한 것이다.

<표 1> 주요 랜섬웨어의 암호 알고리즘 및 접근 대상

분 류	세부 항목
암호 알고리즘	Triple-DES, AES-128, AES-256
접근 파일	pptx, docx, pdf, xlsx

3. 제안 사항

기존 Shannon 엔트로피를 이용해 파일의 암호화 여부를 판별하는 방안은 파일 전체에 대한 엔트로피를 계산한다. 하지만 기존 방안은 판별 결과가 부정확할 수 있는 문제가 있기 때문에 실제 적용에는 어려움이 있다. 따라서 기존 방안의 문제점을 분석하고, 정확성을 향상시킬 수 있는 방안을 제안한다.

3.1 기존 방안의 문제점 분석

평문의 엔트로피가 암호문의 엔트로피보다 높을 수 있다. 한 파일에 대해 암호화를 적용했을 때, 암호문의 엔트로피 값이 평문의 값보다 낮아질 수 있다. 이는 평문을 암호문으로 또는 암호문을 평문으로 오탐하는 상황을 초래할 수 있다.

이를 해결하기 위해 암호문의 엔트로피 값은 높게, 평문의 엔트로피 값은 낮게 측정할 수 있는 방안이 필요하다.

3.2 분할된 엔트로피 측정을 통한 정확성 향상 방안

평문의 경우 파일 시그니처 정보, 파일 내용 등이 차지하는 부분이 나누어져 있다. 이를 이용해 파일을 여러 구역으로 분할한 후 여러 개의 엔트로피 값의 평균을 내면 엔트로피 값이 감소하는 결과를 얻을 수 있다. 하지만 암호문은 파일 전체에 균일한 분포를 가지기 때문에 평문에 비해 엔트로피 값이 감소하는 폭이 작다.

따라서 평문의 엔트로피 값 감소 폭은 크게, 암호문의 엔트로피 값 감소 폭은 작게 하여 기존 방안보다 평문과 암호문 간 엔트로피 값 차이를 크게 만들 수 있다. 이를 위해 엔트로피 값 측정 연산을 수정한 것은 다음과 같다.

$$e = \sum_{k=1}^{Areanum} \left(\sum_{i=0}^{255} P_{B_i}(-\log_2 P_{B_i}) \right) / Areanum,\ P_{B_i} = \frac{F_i}{Areasize}$$

(Areasize는 분할된 구역의 크기이며, Areanum은 분할된 구역의 수이다.)

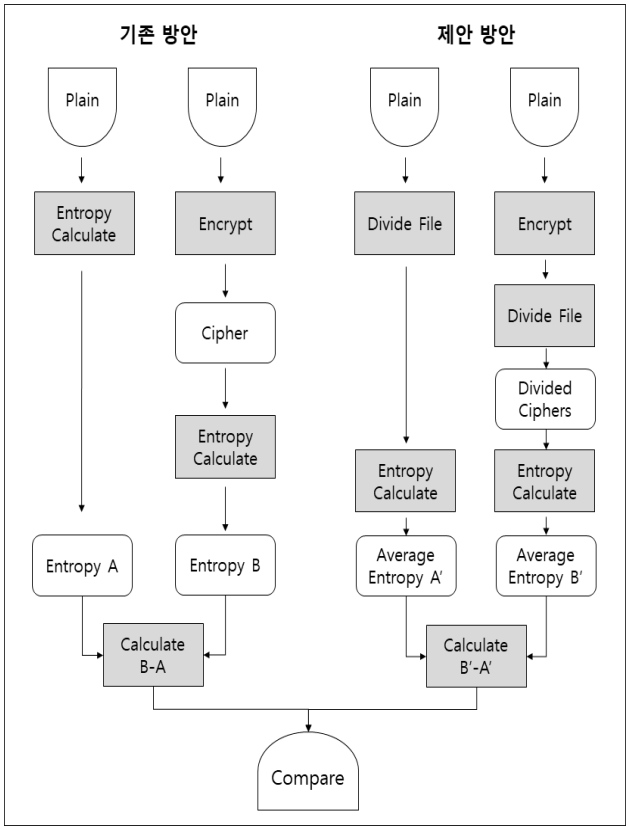
<표 2>는 기존 방안의 문제점과 정확성 향상 방안에 대해 정리한 것이다.

<표 2> 기존 문제점 및 정확성 향상을 위한 개선 방안

분 류	내용
문제점	엔트로피 값 불확실성으로 인한 파일 암호화 여부 판단 부정확
개선 방안	평문과 암호문 간 엔트로피 값 차이 발생으로 정확성 향상

3.3 구현 및 측정 결과

구현을 위해 접근 파일 네 종류, 총 80개의 파일들을 각각 Triple-DES, AES-128, AES-256(CBC 모드)로 암호화하였다. 파일 분할 크기는 암호 알고리즘들의 2048바이트에 16바이트를 더해주며 정확성 높은 분할 크기를 찾을 수 있도록 반복하였다. (그림 1)은 기존 방안과 제안 방안의 비교 과정을 나타낸 것으로, 해당 과정을 80개의 파일에 대해 분할 크기를 변경하며 반복한다.



(그림 1) 구현 및 비교 방안

다양한 분할 크기들 중 4,096바이트로 분할 했을 때 가장 암호문과 평문의 엔트로피 값 차이가 큰 결과를 얻을 수 있었다. 세 가지 암호 알고리즘과 80개의 파일을 대상으로 분할된 엔트로피 값을 측정한 결과, 모든 파일에 대해 엔트로피 값 차이가 커진 결과를 얻을 수 있다.

랜섬웨어의 특성 상 암호 알고리즘과 접근 파일 종류의 구분 없이 엔트로피 값 차이를 측정하였으며, <표 3>은 기존 방안과 제안 방안으로 측정한 엔트로피 차이 결과의 일부이다.

<표 3> 기존 및 제안 방안의 엔트로피 차이 측정 일부

기존 방안의 엔트로피 차이	제안 방안의 엔트로피 차이
0.179398	1.773794
0.271839	1.286309
0.052972	1.586222
0.002484	1.428367
0.632714	1.765784
0.009872	1.735918
0.010744	1.720880
0.074877	1.710801
0.033368	1.720209
0.261086	1.751970
0.079661	1.717322
0.045219	1.739751
0.079325	0.228221
0.078539	0.418082
0.017506	1.283649
…	…

4. 결론

본 논문에서는 파일 암호화 여부 판별의 정확성을 높이기 위해 분할된 Shannon 엔트로피 측정 방안을 제안하였다. 여러 종류의 암호 알고리즘과 랜섬웨어 타겟 파일들을 대상으로 분할된 엔트로피 값 측정을 시도하였으며, 그 중 4096바이트 단위로 분할한 경우 엔트로피 값 차이가 커지는 것을 확인할 수 있었다.

하지만 제한된 환경에서 이용되는 파일들의 크기는 매우 작을 수 있다. 파일 크기가 4096바이트 이하인 경우 기존 방식과 동일하게 적용되어 정확성 향상을 기대하기 어렵다.

따라서 향후 연구에서는 안드로이드의 db파일과 같은 크기가 작은 파일에 대해서도 분할된 Shannon 엔트로피 값을 이용해 파일 암호화 판별의 정확성을 높일 수 있도록 해야 한다. 또한, Shannon 엔트로피를 이용한 실시간 파일 암호화 여부 판단의 실현 가능성을 위해 연산 효율에 대한 추가적인 연구가 진행되어야 한다.

참고문헌

[1] 한국인터넷진흥원, “2017년 4분기 사이버 위협 동향 보고서”, 2018.01.

[2] 한국인터넷진흥원, “‘16년 랜섬웨어 동향 및 ‘17년 전망”, 2017.01.

[3] Jing Chen, Chiheng Wang, Ziming Zhao, Kai Chen, Ruiying Du, Gail-Joon Ahn, “Uncovering the Face of Android Ransomware: Characterization and Real-Time Detection”, IEEE Transaction on Information Forensics and Security, Vol.13 No.5, 2017.12.

[4] Nolen Scaife, Henry Carter, Patrick Traynor, Kevin R.B. Butler, “CryptoLock (and Drop It): Stopping Ransomware Attacks on User Data”, 2016 IEEE 36th International Conference on Distributed computing Systems, 2016.06.

[5] 안랩 시큐리티대응센터(ASEC) 분석팀, “‘리눅스 랜섬웨어’ 본격화되나”, AhnLab 보안이슈 게시판, 2017.07.