Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.26, NO.2, Apr. 2016 ISSN 2288-2715(Online)
http://dx.doi.org/10.13089/JKIISC.2016.26.2.405

선제 대응을 위한 의심 도메인 추론 방안

강 병 호,† 양 지 수, 소 재 현, 김 창 엽‡
안랩

A Proactive Inference Method of Suspicious Domains

Byeongho Kang,† JISU YANG, Jaehyun So, Czang Yeob Kim‡
AhnLab

요 약

본 논문에서는 선제 대응을 위한 의심 도메인 추론 방안을 제시한다. TLD Zone 파일과 WHOIS 정보를 이용하여 의심 도메인을 추론하며, 후보 도메인 탐색, 기계 학습, 의심 도메인 집단 추론의 세 과정으로 구성되어 있다. 첫 번째 과정에서는 씨앗 도메인과 동일한 네임 서버와 업데이트 시간을 가진 다른 도메인을 TLD Zone 파일로부터 추출하여 후보 도메인을 형성하며, 두 번째 과정에서는 후보 도메인의 WHOIS 정보를 정량화하여 유사한 집단끼리 군집화 한다. 마지막 과정에서는 씨앗 도메인을 포함하는 클러스터에 속한 도메인을 의심 도메인 집단으로 추론한다. 실험에서는 .COM과 .NET의 TLD Zone 파일을 사용하였으며, 10개의 알려진 악성 도메인을 씨앗 도메인으로 이용하였다. 실험 결과, 제안하는 방안은 55개의 도메인을 의심 도메인으로 추론하였으며, 그 중 52개는 적중하였다. F1은 0.91을 기록하였으며, 정밀도는 0.95을 보였다. 본 논문에서 제안하는 방안을 통해 악성 도메인을 추론하여 사전에 차단할 수 있을 것으로 기대한다.

ABSTRACT

In this paper, we propose a proactive inference method of finding suspicious domains. Our method detects potential malicious domains from the seed domain information extracted from the TLD Zone files and WHOIS information. The inference process follows the three steps: searching the candidate domains, machine learning, and generating a suspicious domain pool. In the first step, we search the TLD Zone files and build a candidate domain set which has the same name server information with the seed domain. The next step clusters the candidate domains by the similarity of the WHOIS information. The final step in the inference process finds the seed domain's cluster, and make the cluster as a suspicious domain set. In experiments, we used .COM and .NET TLD Zone files, and tested 10 seed domains selected by our analysts. The experimental results show that our proposed method finds 55 suspicious domains and 52 true positives. F1 scores 0.91, and precision is 0.95 We hope our proposal will contribute to the further proactive malicious domain blacklisting research.

Keywords: Suspicious Domain Inference, Proactive Detection, DNS Zone File, WHOIS Information, Machine Learning

I. 서 론

블랙리스트 기법은 악성 요소의 주요 대응 방안 중 하나로, 알려진 악성 요소가 시스템에서 발견되는지 조사하여 차단한다[19, 20]. 블랙리스트 기법은 알려진 악성 요소에 대해서는 높은 탐지율과 낮은 오진율을 보이며 빠른 검사 속도를 가진다는 강점이 있으나, 경험적 판단에 근거하므로 새로운 유형의 악성 요소 대응에는 약점이 있다[2, 11]. 가령, 침해 사고가 발생했을 때 동원되는 악성 요소는 사후에 블랙리스트에 등록되어 차단되며, 재발 시에는 과거의 이력을 토대로 대응한다. 이에 따라 블랙리스트 기법은

Received(07. 13. 2015), Modified(1st: 10. 05. 2015, 2nd: 01. 28. 2016), Accepted(04. 08. 2016)
† 주저자, byeongho.kang@ahnlab.com
‡ 교신저자, matthew.kim@ahnlab.com(Corresponding author)

변형된 공격에 대해서는 대응 시간이 소모되어 빠르게 확산되는 공격을 효과적으로 차단하기에는 어려움이 있다. 이러한 약점을 보완하기 위해 의심되는 악성 요소를 공격 이전에 블랙리스트에 등록해 두는 선제 방어 기법이 연구되고 있으며, 대표적으로 DNS 기반의 도메인 검열[4, 6, 7], 행위 판단[5, 9], 악성 URL 분석[3, 17, 18] 등의 분야가 있다.

본 논문에서는 선제 대응을 위한 의심 도메인 추론 방안을 제안한다. 알려진 악성 도메인과 유사한 등록 정보를 가진 도메인을 의심 도메인으로 가정하였으며, 동일한 네임 서버에 등록되어 있고 유사한 WHOIS 정보를 가진 도메인을 탐색하여 추론한다. 실험에서는 분석가들이 임의로 선택한 10개의 악성 도메인을 씨앗 도메인으로 사용하여 의심 도메인을 추론하였다. 실험 결과, 제안하는 방안은 총 55개의 도메인을 추론하였고 52개의 도메인은 적중하였다.

본 논문의 구성은 다음과 같다. 2장에서는 의심 도메인의 블랙리스트 기법에 관한 선행 연구를 다루며 제안하는 방안과 비교한다. 3장에서는 제안하는 방안에 대해 상세히 설명하며, 4장에서는 실험 환경 및 평가 기준을 정의하고 실험 결과를 보인다. 5장에서는 제안하는 방안의 성능과 강점 및 약점을 실험 결과에 근거하여 논의하며, 6장에서 결론을 내린다.

II. 관련연구

블랙리스트 기법은 과거의 공격 요소를 토대로 악성 요소를 탐지하므로, 악성 요소의 변종 탐지에 약점이 있다[14,15]. 이러한 약점을 보완하기 위한 여러 연구가 진행되고 있으며, 그 중 Jian Zhang 등[1]은 HPB (Highly Predictive Blacklisting)를 제안하였다. HPB는 로그 기반의 악성코드 탐지 기법으로, 알려진 악성 코드의 행위 로그와 현재 시스템의 행위 로그를 대조하여 악성코드를 탐지한다. 행위 분석 과정에서 페이지랭크(PageRank)와 유사한 연결망 분석을 통해 경험 기반 탐지의 약점을 줄이기 위해 노력하였다. 이 방안은 과거의 공격으로부터 완전히 변경된 공격은 대응하지 못하지만 대부분의 변종 공격에는 효과적으로 대응할 수 있다는 장점이 있다. Justin Ma 등[3]은 악성 URL 분석을 통해 웹페이지의 악성 주소를 탐지하는 연구를 진행하였다. 알려진 악성코드가 참조하는 URL의 문자열을 통계 분석하여 악성 URL을 추론하였으며, 호스트 정보를 추가하여 진단율을 향상시켰다. 이 과정을 통해 형성된 추론 모델은 95-99%의 정확도를 보였다. 하지만 Justin Ma 등이 제안한 방안은 알려진 악성코드가 참조하는 악성 URL을 추론하므로 새로운 유형의 악성 요소 탐지에는 약점이 있다. 즉 Jian Zhang 등의 연구와 Justin Ma 등의 연구 모두 과거의 공격에 기반을 둔다는 블랙리스트 기법의 단점을 완전히 해소했다고 보기는 어렵다. 따라서 새로운 유형의 공격이 시작되면 악성 요소를 분석하고 블랙리스트에 등록시키는 대응 시간이 소모된다. Fig. 1은 블랙리스트의 경험적 판단으로 인한 대응 시간의 소모를 보인다.

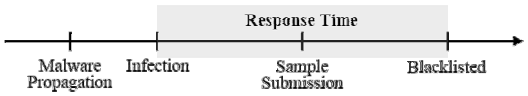

Fig. 1. Reactivity of Blacklisting

Mark Felegyhazi 등[2]은 악성 요소의 선제 대응을 위해 의심 도메인을 추론하였으며, 본 논문에 많은 귀감이 되었다. 이들은 알려진 악성 도메인과 유사한 도메인 등록 정보를 가진 다른 도메인을 탐색하였으며 할당된 네임서버의 나이와 변경 이력을 토대로 의심 도메인을 추론하였다. 실험 결과 약 3,600여개의 알려진 악성 도메인을 씨앗으로 사용하여 약 12,800여개의 도메인을 추론하였으며, 정밀도는 76%을 보였다.

III. 의심 도메인 추론 방안

공격자는 미리 생성된 도메인 집단으로부터 공격에 동원할 도메인을 선택하여 사용한다[8,11]. 따라서 자동화 도구로 생성된 도메인 집단은 유사한 WHOIS 정보와 네임서버를 공유한다는 특징을 가진다[2]. 공격의 초기에는 도메인 집단 중 일부만이 공격에 동원되며, 이 도메인이 차단된 이후 공격자는 새로운 도메인을 공격에 동원한다. 즉 블랙리스트 기법은 대응 시간을 필요로 하므로 공격자는 새로운 도메인을 계속적으로 동원함으로써 공격을 지속할 수 있다는 문제점이 있다. 이러한 약점을 보완하기 위해 본 논문에서는 공격의 초기 시점에 발견된 도메인으로부터 그 도메인이 속한 집단을 추론하여 향후 예상되는 추가 공격을 사전에 차단하는 선제 대응 방안을 제안한다.

제안하는 방안은 후보 도메인 탐색, 기계 학습, 의심 도메인 집단 추론의 세 과정으로 구성되어 있

다. 첫 번째 과정에서는 씨앗 도메인과 동일한 네임서버와 업데이트 시간을 가진 다른 도메인을 탐색한다. 분석 경험에 따르면 동일한 공격에 동원되는 도메인은 동일한 네임서버와 네임서버 정보 수정 시간을 가지고 있었기 때문이다. 두 번째 과정에서는 후보 도메인의 WHOIS 정보를 정량화하여 유사한 집단끼리 군집화한다. 마지막 과정에서는 씨앗 도메인을 포함한 클러스터를 탐색하며, 그 클러스터에 속한 도메인을 의심 도메인 집단으로 추론한다.

3.1 후보 도메인 탐색

DNS Zone 파일은 도메인에 할당된 IP, 네임서버, TTL 정보 등을 포함하고 있다. 본 연구에서는 후보 도메인 탐색을 위해 네임서버 정보에 집중하였으며, 씨앗 도메인과 동일한 네임서버와 업데이트 시간을 가진 도메인을 TLD Zone 파일로부터 탐색하여 후보 도메인으로 선정하였다. 후보 도메인은 악성 도메인뿐만 아니라 정상 도메인 또한 탐색되는 것이 일반적이나, 네임서버 자체가 악성인 경우에는 후보 도메인 전체가 악성 도메인인 사례도 찾을 수 있었다.

3.2 기계 학습

3.2.1 특징 선택

WHOIS 정보 정량화를 위해 WHOIS 정보의 일부를 선별 및 통합하여 특징으로 사용하였다. Id와 Handle 정보는 임의의 값이 부여되는 관리용 값으로, WHOIS 등록 정보 유사성과의 상호 관계를 찾기 어려워 특징으로 사용하지 않았다. 또한 WHOIS에서 조회되는 네임서버 정보는 신뢰도가 낮아서 사용하지 않았다. WHOIS 정보는 일정 주기로 갱신되기 때문에 갱신 주기 내에 네임서버 정보가 변경되면 과거의 정보를 출력할 수 있다. 따라서 도메인에 등록된 실제 네임서버와 WHOIS에서 조회되는 네임서버가 다른 경우가 있을 수 있으며, 공격자는 도메인의 네임서버를 변경한 직후 공격을 시작할 수 있다. WHOIS의 지연된 업데이트로 인한 과거의 네임서버 정보는 잘못된 추론으로 이어질 수 있기 때문에 DNS의 TLD Zone 파일에서 네임서버 정보를 취득하여 사용하였다. Table 1은 공격에 동원된 악성 도메인의 네임서버 정보가 지연된 업데이트로 인해 상이하게 조회되는 예시를 보인다.

Table 1. Different Name Servers by Data Origins

Data Origin	Namep Servers
TLD Zone	NS1.xxx.xx NS2.xxx.xx
WHOIS Info.	NS1.yyyyyyyyyy.yyy NS2.yyyyyyyyy.yy

Street, PostalCode, City, State, Country 정보는 동일한 맥락을 가진다고 판단하여 하나의 특징인 Address로 통합하여 사용하였다. 같은 방법으로 Email0, Email1, Email2 정보를 *Emails* 특징으로 통합하여 사용하였다. Table 2는 선택된 12개의 특징과 그에 상응하는 자료 구조를 나타낸다.

Table 2. Features, Data Types, and WHOIS Records

Feature	Data Type	WHOIS Record
DomainName	*String*	DomainName
Status	*String*	Status
Registrar	*String*	Registrar
CreationDate	*DateTime*	CreationDate
ExpirationDate	*DateTime*	ExpirationDate
UpdatedDate	*DateTime*	UpdatedDate
Emails	*String*	Email0
		Email1
		Email2
Name	*String*	Name
Organization	*String*	Organization
Address	*String*	Street
		PostalCode
		City
		State
		Country
Phone	*String*	Phone
Fax	*String*	Fax

3.2.2 정량화

특징 값을 정량화하여 모든 후보 도메인을 12차원의 벡터로 표현하였으며, 각 특징 값은 *String*과 *DateTime*의 두 가지 자료 구조를 갖는다. String 자료 구조의 특징 값은 자연어 문자열 발생 빈도[16]와의 코사인 유사도를 연산하여 정량화하였으며, DateTime 자료 구조의 특징 값은 추론 시점으로부터의 경과일 또는 도래일을 연산하여 정량화하였다. DateTime 자료 구조의 특징 값 정량화 수식은

$$\text{Score}_{CreationDate} = \left(1 - \frac{1}{AnalysisDate - CreationDate}\right)$$

$$\text{Score}_{UpdatedDate} = \left(1 - \frac{1}{AnalysisDate - UpdatedDate}\right)$$

$$\text{Score}_{ExpirationDate} = \left(1 - \frac{1}{ExpirationDate - AnalysisDate}\right)$$

Fig. 2. Score Evaluation for the DateTime Type Features

Fig. 2에서 나타낸다.

3.2.3 군집화

후보 도메인은 다수의 클러스터를 포함할 수 있으며 군집화될 클러스터의 총 수는 미리 파악하기 어렵다. 씨앗 도메인과 동일한 네임서버 정보를 가진다 하더라도 반드시 씨앗 도메인과 연관되었다고 보기는 어렵기 때문이다. 따라서 결과로 도출되는 클러스터의 수는 어떠한 후보 도메인을 군집화 하는지에 따라 결정되어야 한다. 또한 후보 도메인 중에서는 다른 어떤 후보 도메인과도 유사하지 않은 도메인이 있을 수 있다. 본 연구는 자동화된 도구로 생성된 도메인 집단을 탐색하는 것을 목적으로 하므로, 단독 도메인은 노이즈로 간주하여 클러스터링 되지 않도록 처리하였다. 본 연구에서는 이러한 제약 조건을 만족시키는 DBSCAN 알고리즘을 사용하여 후보 도메인을 군집화하였다. DBSCAN은 클러스터를 형성하는 최소 오브젝트 수 min_pts와 클러스터를 형성하는 최대 탐색 반경 eps를 인자 값으로 받는다. 단독 도메인을 노이즈로 간주하므로 min_pts 값은 2를 사용하며, eps 값은 실험을 통해 탐색한 최적의 값을 사용한다.

3.3 의심 도메인 집단 추론

기계 학습 과정을 통해 군집화 된 도메인 클러스터 중 씨앗 도메인을 포함하는 클러스터를 의심 도메인 집단으로 추론한다. 이 집단에 속한 도메인은 악성 행위를 나타내는 씨앗 도메인과 WHOIS 정보와 네임서버 정보가 유사하므로 자동화된 도구를 통해 대량으로 생성된 도메인일 가능성이 높다고 가정한다. 따라서 이 집단에 속한 도메인을 블랙리스트에 미리 등록함으로써 씨앗 도메인과 연관된 도메인을 사전에 차단할 수 있다.

IV. 실 험

4.1 실험 환경 및 평가 기준

실험은 피싱 사이트, 봇넷 서버, 악성코드 유포지에 사용된 도메인 중 10개를 씨앗 도메인으로 사용하여 의심 도메인을 추론하였다. 추론된 도메인은 분석을 통해 악성 여부를 직접 판별하였으며, 씨앗 도메인과의 연관성을 고려하여 추론의 정당성을 검증한다. 또한 한 개의 도메인으로부터 평균 몇 개의 도메인을 추론하는지 평가하기 위한 추론 인수(Inference Factor)를 정의하여 사용하며, F1과 정밀도(precision), 재현율(recall) 또한 함께 평가한다.

실험에서는 .COM과 .NET의 TLD Zone 파일을 이용하였다. 그 이유로는 첫째, APWG의 2014년 하반기 Global Phishing Survey에 의하면 악성 도메인의 65.1%가 .COM 또는 .NET TLD를 사용하기 때문이며[13], 둘째, .COM과 .NET의 TLD Zone 파일은 접근이 용이하고 VeriSign[12]에 의해 잘 관리되고 있어 신뢰할 수 있기 때문이다.

4.2 성능 조율

DBSCAN 알고리즘의 eps 인자에 대해 성능 조율을 수행하여 최적의 값을 탐색한다. 추론 인수와 정밀도, 오경보(false positive) 값을 평가하였으며, 정밀도를 최대로 유지하는 선에서 높은 추론 인수와 낮은 오경보 값을 보이는 수치를 선택하였다. 우리는 단독 도메인을 노이즈로 간주하여 min_pts 값은 2로 고정하여 사용하였다. Fig. 3과 Table 3은 성능 조율 실험 결과를 나타낸다.

실험 결과, eps 값이 0.2와 0.21일 때 주어진 평가 기준에 부합하는 최적의 성능을 보이는 것을 확인하였다. 실험 과정에서는 eps 값으로 0.21을 사용하였으며, 이 경우 평균 추론 인수는 5.2, 평균 정밀도는 0.95, 평균 오경보 값은 0.33을 나타내었다. 10개의 씨앗 도메인에 대한 성능 조율 실험의 전체 결과는 Appendix A에 수록하였다.

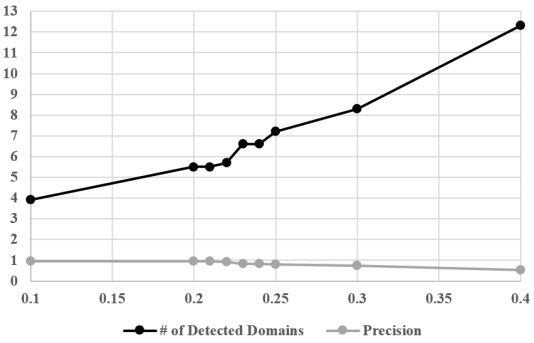

Fig. 3. Performance Graph with eps (min_pts=2)

Table 3. DBSCAN Parameter Evaluation with eps (min_pts=2)

eps	Average Inference Factor	Average Precision	Average False Positive
0.1	3.9	0.95	0.20
0.2	5.5	0.95	0.33
0.21*	5.5	0.95	0.33
0.22	5.7	0.91	0.46
0.23	6.6	0.84	1.04
0.24	6.6	0.84	1.04
0.25	7.2	0.79	1.49
0.3	8.3	0.75	2.10
0.4	12.3	0.54	5.60

4.3 실험 결과

제안하는 방안은 10개의 씨앗 도메인으로부터 55개의 의심 도메인을 추론하였다. 의심 도메인을 포함한 모든 후보 도메인을 실험 결과와 실제 사이트를 비교 분석하였다. 그 결과, 52개의 도메인은 적중하였고 3개 도메인은 오경보하였으며, 7개의 도메인은 악성임에도 추론에 실패하였다고 판단하였다. F1은 0.91을 기록하였으며, 정밀도와 재현율은 각각 0.95와 0.88을 나타내었다. 실험 결과는 Table 4에 정리하여 수록하였다.

연구의 약점을 파악하기 위해 3개의 오경보된 도메인에 대해 조사하였다. 그 중 2개는 인터넷 쇼핑몰로, 모바일 폰 주변기기를 판매하는 곳이다. 이 두 사이트는 서로 다른 곳임에도 씨앗 도메인과 유사한 도메인 등록 정보를 가지고 있었다. 이 두 사이트는 쇼핑몰 제작 도구로 만들어졌으며 특정 업체에 의해 공동으로 관리되는 사이트라고 추정하였다. 오경보

Table 4. Experimental Results (eps=0.21, min_pts=2)

Suspicious Domains	55	Inference Factor	5.2
True Positive	52	F1	0.91
False Positive	3	Precision	0.95
False Negative	7	Recall	0.88

된 나머지 1개의 사이트는 소규모 호텔 사이트로, 이 사이트 또한 쇼핑몰 제작 도구로 만들어지고 공동으로 관리되는 사이트로 추정하였다.

추론에 적중한 사이트 중에서 여전히 악성 행위에 동원되지 않은 2개의 도메인을 발견하였다. 이 도메인은 동일한 집단에 속한 다른 악성 도메인과 동일한 등록정보와 응답 특성을 가지지만, 악성 행위에 동원된 증거를 찾을 수 없었다. 이 두 도메인은 잠재적으로 악성 행위에 동원될 가능성이 높다고 판단하여 휴면 악성 도메인으로 분류하였다.

4.4 Mark Felegyhazi 등의 연구와의 비교

본 논문에서 제안하는 방안의 실험 결과를 Mark Felegyhazi 등의 연구[2]와 비교하였다. 평가 항목으로는 씨앗 도메인 수, 추론 도메인 수, 추론 인수, 정밀도를 선정하였다. Mark Felegyhazi 등의 연구에서는 적중률만을 제공했기 때문에 F1 값과 재현율은 비교할 수 없었다.

Mark Felegyhazi 등의 연구는 3,653개의 씨앗 도메인으로부터 총 12,799의 도메인을 추론하여 3.4의 추론 인수를 기록하였다. 반면 본 논문에서 제안하는 방안의 추론 인수는 5.2로, Mark Felegyhazi 등의 연구에 비해 약 1.5배가량 높은

Table 5. Comparison with Mark Felegyhazi's Approach

	Our Approach	Mark Felegyhazi's Approach
# of Seed Domains	10	3,653
# of Inferred Domains	52	12,799
Inference Factor	5.2	3.4
Precision	0.95	0.76

수치를 보인다. 또한 정밀도에서도 강점을 보였다. Mark 등의 연구는 0.76을 기록한 반면, 제안하는 방안은 0.95를 기록하였다. 하지만 본 논문에서 제안하는 방안은 실험 규모가 작기 때문에 Mark Felegyhazi 등의 연구와 유사한 수준의 데이터에서도 동일하게 우수한 성능을 나타낼지에 대해서는 추가 실험이 필요하다.

V. 논 의

본 논문에서 제안하는 방안은 상용 서비스로 구축하기 위해 높은 정밀도와 낮은 오경보에 집중하여 성능 조율을 진행하였다. 그 결과, 기존의 연구 대비 높은 정밀도를 확보할 수 있었다. 하지만 반대급부로 진단에 실패한 도메인 또한 관찰할 수 있었다. Appendix A의 전체 실험 결과에서, 제안하는 방안은 2번째, 4번째, 7번째, 10번째 씨앗 도메인으로부터 최소 7개의 도메인이 실제 악성임에도 의심 도메인으로 추론해 내지 못하였다. 이 7개의 도메인은 eps 값을 증가시킴으로써 추론할 수 있으나, 이 경우에는 정밀도가 낮아지고 오경보가 증가하는 문제를 수반하였다. 미진단된 7개의 도메인을 분석하였고, 이 도메인들의 등록일, 만료일, 갱신일이 씨앗 도메인과 차이가 크며, 그 외의 정보는 동일하다는 점을 찾을 수 있었다. 분석을 통해 이 도메인이 각각 다른 시점에 등록되었으나, 현재는 동일한 공격자가 자동화 도구로 관리하는 것임을 추정할 수 있었다.

VI. 결 론

블랙리스트 기법은 악성 요소의 주요한 대응 방안 중 하나로, 알려진 악성 요소에 대해서는 우수한 성능을 보인다. 하지만 이 방법은 과거의 공격 이력을 토대로 진단하기 때문에 새로운 유형의 악성 요소 대응에는 한계가 있다. 본 논문에서는 이러한 약점을 보완하기 위한 의심 도메인 추론 방안을 제시하였다. 분석 경험에 따르면, 악성코드 제작자는 자동화된 도메인 관리 도구를 이용하여 많은 수의 도메인을 일괄 관리하며 동일한 네임서버와 업데이트 시간을 가지고 있고 WHOIS 정보 또한 같거나 유사하다. 본 연구에서는 이 점에 착안하여 TLD Zone 파일을 통해 네임서버와 그 업데이트 시간이 동일한 도메인을 탐색하였으며, WHOIS 정보를 토대로 군집화 하였다. 그리고 클러스터 중 씨앗 도메인을 포함하는 클러스터를 탐색하여 의심 도메인을 추론하였다. 실험은 10개의 씨앗 도메인을 이용하여 진행하였으며, 그 결과 총 55개의 의심 도메인을 추론하였다. 추론된 도메인 중 52개는 악성 도메인으로 판별되었고, 3개의 도메인은 오경보로 밝혀졌다. 오경보된 도메인은 인터넷 쇼핑몰과 지역 호텔 사이트이며, 특정한 솔루션을 이용하여 공동 관리 중인 도메인으로 추정하였다. F1은 0.91을 기록하였고 정밀도는 0.95, 재현율은 0.88을 보였다. 본 논문에서 제안하는 방안을 통해 악성 요소의 선제 대응에 기여할 것으로 기대한다.

Appendix A. Full Experimental Results

Table 6. DBSCAN Parameter Evaluation with eps (min_pts=2)

Seed No. (Category)	eps	# of Suspicious Domains	FP	Precision
1 (Botnet Server)	0.1	8	0	1.00
	0.2	8	0	1.00
	0.21	**8**	**0**	**1.00**
	0.22	8	0	1.00
	0.23	11	3	0.73
	0.24	11	3	0.73
	0.25	14	6	0.57
	0.3	21	13	0.38
	0.4	21	13	0.38
2 (Botnet Server)	0.1	3	0	1.00
	0.2	6	0	1.00
	0.21	**6**	**0**	**1.00**
	0.22	6	0	1.00
	0.23	6	0	1.00
	0.24	6	0	1.00
	0.25	6	0	1.00
	0.3	7	0	1.00
	0.4	23	14	0.39
3 (Botnet Server)	0.1	3	0	1.00
	0.2	11	0	1.00
	0.21	**11**	**0**	**1.00**
	0.22	11	0	1.00
	0.23	14	3	0.79
	0.24	14	3	0.79
	0.25	14	3	0.79
	0.3	14	3	0.79
	0.4	22	11	0.50
4 (Botnet Server)	0.1	3	0	1.00
	0.2	3	0	1.00
	0.21	**3**	**0**	**1.00**
	0.22	3	0	1.00
	0.23	3	0	1.00
	0.24	3	0	1.00

Seed No. (Category)	eps	# of Suspicious Domains	FP	Precision
	0.25	6	2	0.67
	0.3	7	3	0.57
	0.4	11	7	0.36
5 (Phishing Site)	0.1	3	0	1.00
	0.2	3	0	1.00
	0.21	**3**	**0**	**1.00**
	0.22	4	1	0.75
	0.23	6	3	0.50
	0.24	6	3	0.50
	0.25	6	3	0.50
	0.3	6	3	0.50
	0.4	6	3	0.50
6 (Phishing Site)	0.1	4	1	0.75
	0.2	5	2	0.60
	0.21	**5**	**2**	**0.60**
	0.22	5	2	0.60
	0.23	5	2	0.60
	0.24	5	2	0.60
	0.25	5	2	0.60
	0.3	7	4	0.43
	0.4	9	4	0.56
7 (Phishing Site)	0.1	4	0	1.00
	0.2	6	0	1.00
	0.21	**6**	**0**	**1.00**
	0.22	7	0	1.00
	0.23	7	0	1.00
	0.24	7	0	1.00
	0.25	7	0	1.00
	0.3	7	0	1.00
	0.4	7	0	1.00
8 (Phishing Site)	0.1	3	0	1.00
	0.2	4	0	1.00
	0.21	**4**	**0**	**1.00**
	0.22	4	0	1.00
	0.23	4	0	1.00
	0.24	4	0	1.00
	0.25	4	0	1.00
	0.3	4	0	1.00
	0.4	4	0	1.00
9 (Malware Repository)	0.1	4	1	0.75
	0.2	5	1	0.80
	0.21	**5**	**1**	**0.80**
	0.22	5	1	0.80
	0.23	5	1	0.80
	0.24	5	1	0.80
	0.25	5	1	0.80
	0.3	5	1	0.80
	0.4	9	5	0.44
10 (Malware Repository)	0.1	4	0	1.00
	0.2	4	0	1.00
	0.21	**4**	**0**	**1.00**
	0.22	4	0	1.00
	0.23	5	0	1.00
	0.24	5	0	1.00
	0.25	5	0	1.00
	0.3	5	0	1.00
	0.4	11	6	0.45

References

[1] Jian Zhan, Phillip Porras, and Johannes Ullrich, "Highly Predictive Blacklisting," Proceedings of the 17th USENIX Security Symposium, pp. 107-122, Jul. 2008.

[2] Mark Felegyhazi, Christian Kreibich, and Vern Paxson, "On the Potential of Proactive Domain Blacklisting," Proceedings of the 3rd USENIX Conference on Large-scale Exploits and Emergent Threats (LEET '10), pp. 6-13, Apr. 2010.

[3] Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs," Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1245-1254, Jun. 2009.

[4] Leyla Bilge, Engin Kirda, Christopher Kruegel, and Marco Balduzzi, "EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis," Proceedings of 18th Annual Network and Distributed System Security Symposium (NDSS), Feb. 2011.

[5] M. Patrick Collins, Timothy J. Shimeall, Sidney Faber, Jeff Janies, Rhiannon Weaver, Markus De Shon, and Joseph B. Kadane, "Using Uncleanliness to Predict Future Botnet Addresses," Proceedings of the 7th SIGCOMM Conference on Internet Measurement, pp. 93-104, Aug. 2007.

[6] Yuanchen He, Zhenyu Zhong, Sven

Krasser, and Yuchun Tang, "Mining DNS for Malicious Domain Registrations," Proceedings of the 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), pp. 1-6, Oct. 2010.

〔7〕 Manos Antonakakis, Roberto Perdisci, Wenke Lee, Nikolaos Vasiloglou II, and David Dagon, "Detecting Malware Domains at the Upper DNS Hierarchy," Proceedings of the 20th USENIX Security Symposium, pp. 16-30, Aug. 2011.

〔8〕 Aditya Kapoor, and Rachit Mathur, "Predicting the Future of Stealth Attacks," Proceedings of the 21st Virus Bulletin International Conference (VB2011), pp. 5-7, Oct. 2011.

〔9〕 Byeongho Kang, TaeGuen Kim, BooJoong Kang, Eul Gyu Im, and Minsoo Ryu, "TASEL: Dynamic Taint Analysis with Selective Control Dependency," Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems (RACS), pp. 272-277, Oct. 2014.

〔10〕 Byeongho Kang, and Eul Gyu Im, "Analysis of Binary Code Topology for Dynamic Analysis," Proceedings of the 29th ACM Symposium on Applied Computing (SAC), pp. 1731-1732, Mar. 2014.

〔11〕 Byeongho Kang, JISU YANG, Jaehyun So, and Czang Yeob Kim, "Detecting Trigger-based Behaviors in Botnet Malware," Proceedings of the 2015 Research in Adaptive and Convergent Systems (RACS), pp. 274-279, Oct. 2015.

〔12〕 VeriSign, Zone Files for Top Level Domains (TLDs), http://www.verisigninc.com/en_US/channel-resources/domain-registry-products/zone-file/index.xhtmlVeriSign

〔13〕 APWG, Global Phishing Survey 2H2014, http://apwg.org/download/document/245/APWG_Global_Phishing_Report_2H_2014.pdf

〔14〕 Abhijit Bose, Xin Hu, Kang G. Shin, and Taejoon Park, "Behavioral Detection of Malware on Mobile Handsets," Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys '08), pp. 225-238, Jun. 2008.

〔15〕 Adrian Tang, Simha Sethumadhavan, and Salvatore J. Stolfo, "Unsupervised Anomaly-Based Malware Detection using Hardware Features," Research in Attacks, Intrusions, and Defenses, pp. 109-129, Sep. 2014.

〔16〕 Robert Edward Lewand, Cryptological Mathematics, The Mathematical Association of America, Dec. 2000.

〔17〕 Hyunsang Choi, Bin B. Zhu, and Heejo Lee, "Detecting Malicious Web Links and Identifying Their Attack Types," Proceedings of the 2nd USENIX Conference on Web Application Development (WebApps '11), pp. 125-136, Jun. 2011.

〔18〕 Kyle Zeeuwen, Matei Ripeanu, and Konstantin Beznosov, "Improving Malicious URL Re-Evaluation Scheduling through an Empirical Study of Malware Download Centers," Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality 2011), pp. 42-49, Mar. 2011.

〔19〕 Yu Feng, Saswat Anand, Isil Dillig, and Alex Aiken, "Apposcopy: Semantics-based Detection of Android Malware through Static Analysis," Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2014), pp. 576-587, Nov. 2014.

〔20〕 Bhimshankar Mantur, Abhijeet Desai, K.S. Nagegowda, "Centralized Control

Signature-Based Firewall and Statistical-Based Network Intrusion Detection System (NIDS) in Software Defined Networks (SDN)," Emerging Research in Computing, Information, Communication and Applications, pp. 497-506, Jul. 2015.

〈저 자 소 개〉

강 병 호 (Byeongho Kang) 정회원
2013년: 한양대학교 공과대학 컴퓨터공학부 학사 졸업
2015년: 한양대학교 컴퓨터·소프트웨어학과 석사 졸업
2015년 1월~현재: 안랩 ASEC 위협분석팀 연구원

〈관심분야〉 머신러닝

양 지 수 (JISU YANG) 정회원
2013년: 한양대학교 공과대학 컴퓨터공학부 학사 졸업
2012년: 한양대학교 전자컴퓨터통신공학과 석사 졸업
2012년 1월~현재: 안랩 ASEC 위협분석팀 주임연구원
〈관심분야〉 네트워크보안

소 재 현 (Jaehyun So) 정회원
2009년 7월~현재: 안랩 ASEC 위협분석팀 선임연구원
〈관심분야〉 디지털포렌식

김 창 엽 (Czang Yeob Kim) 정회원
2007년 1월~현재: 안랩 ASEC 위협분석팀 선임연구원
〈관심분야〉 데이터마이닝, 머신러닝