JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 12, pp. 215-225, December 2025
https://doi.org/10.9708/jksci.2025.30.12.215

Detection and Verification of Cryptocurrency Activities from Unstructured Data in Kimsuky and Prack Leak Analysis

Hyewon Lee*, Donghyun Yeo*, Minwon Seo*

*Researcher, AI Spera, Seoul, Korea

[Abstract]

In late 2024, large-scale data leaked from cyberattacks linked to the Chinese and North Korean hacker group Kimsuky and the Prack incident included unstructured information such as government logs, source code, and browser timelines from the Ministry of Foreign Affairs and the Defense Counterintelligence Command. This study analyzes whether the attackers conducted financial activities using cryptocurrencies. Automatic identification of valid crypto addresses in large text datasets is challenging, as simple regex detection yields high false positives. To overcome this, we implemented a four-stage pipeline: (1) multi-coin regex detection, (2) checksum and decoding validation, (3) contextual scoring, and (4) on-chain verification. Experiments using approximately 80 MB of leaked data and Ethereum records from Etherscan reduced false positives by 75%, doubled true detections, and achieved an average processing time under three minutes. In particular, Ethereum address 0xb211b4...0cb6 appeared both in browser logs and on-chain deposits, confirming that the attacker viewed and analyzed blockchain assets. This research demonstrates a practical methodology for reconstructing blockchain activities from unstructured data in state-sponsored hacking cases.

▸ **Key words**: Cyber threat intelligence, Blockchain forensics, Cryptocurrency address detection, Kimsuky, On-chain verification

[요 약]

2024년 말 공개된 중국 및 북한 해커 조직 '김수키(Kimsuky)' 관련 유출 자료와 프랙(Prack) 이슈를 통해 노출된 해킹 데이터에는 외교부·방첩사 등 국가기관의 로그, 코드, 브라우저 기록 등 비정형 정보가 포함되어 있었다. 본 연구는 이 데이터를 기반으로 해커가 가상화폐를 이용한 금전 활동을 수행했는지 분석하였다. 대규모 텍스트에서 신뢰할 수 있는 암호화폐 주소를 자동 식별하기는 어렵고, 단순 정규식 탐지는 오탐이 많다. 이를 해결하기 위해 (1) 정규식 후보 탐색, (2) 체크섬·디코딩 검증, (3) 문맥 점수화, (4) 온체인 검증으로 구성된 4단계 파이프라인을 구현하였다. 김수키·프랙 관련 유출자료(약 80MB)와 Etherscan 데이터 실험 결과, 오탐률 75% 감소, 정탐률 2배 향상, 평균 처리시간 3분 이내를 달성하였다. 특히 Ethereum 주소 0xb211b4... 사례에서는 브라우저 기록과 온체인 입금 내역이 일치해 실제 자산 조회 행위를 확인하였다. 본 연구는 중국·북한계 해커 분석 과정에서 비정형 데이터로부터 블록체인 활동을 복원하는 실증적 방법론을 제시했다는 점에서 의의가 있다.

▸ **주제어**: 사이버 위협 인텔리전스, 블록체인 포렌식, 암호화폐 주소 탐지, 북한 김수키 해커, 온체인 검증

• First Author: Hyewon Lee, Corresponding Author: Hyewon Lee
*Hyewon Lee (hwlee@aispera.com), AI Spera
*Donghyun Yeo (dhyeo@aispera.com), AI Spera
*Minwon Seo (pypygeek@aispera.com), AI Spera
• Received: 2025. 11. 24, Revised: 2025. 12. 15, Accepted: 2025. 12. 16.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

한국 공공기관을 겨냥한 공격자들의 활동은 점점 더 금전화 중심으로 진화하고 있다.

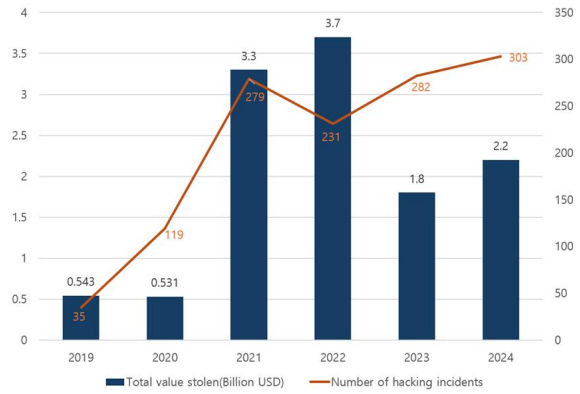

Fig. 1. Cryptocurrency Hacking Incidents: Stolen Value and Incident Count (2019-2024)

특히 2024-2025년 사이 공개된 이른바 “해커 내부 자료”에는 피싱/스피어피싱 인프라, 크리덴셜 수집 도구, 운영 로그와 더불어 가상화폐 주소·트랜잭션 식별자·블록 탐색기(예: Etherscan) 접근 흔적이 뒤섞여 존재한다.

이러한 데이터는 단순 침해 정황을 넘어, 공격자의 자금 흐름과 운영 습관을 추적할 수 있는 근거를 제공한다. 그러나 현실에서는 수십~수백 GB에 이르는 비정형 파일 더미 속에서 신뢰할 수 있는 암호화폐 식별자를 정확히 뽑아내는 것 자체가 난제다. 정규식만으로는 오탐(false positive)이 폭증하고, 반대로 지나치게 보수적인 규칙은 정탐(true positive)을 놓치기 때문이다.

본 연구는 최근 공개된 대규모 유출 자료(브라우저 타임라인, 쿠키/설정, 작업 폴더, 코드 스니펫, 로그 등)에서 공격자와 연관된 가상화폐 주소를 자동 식별·검증하고, 그 결과를 바탕으로 온체인 상의 활동 패턴을 해석하는 것을 목적으로 한다. 이를 위해 우리는 (1) 다양한 코인 포맷을 포괄하는 정규식 기반 1차 후보 탐색, (2) 코인별 체크섬·디코딩 기반 2차 검증, (3) 파일·경로·주변 토큰을 고려한 문맥 점수화(context scoring), (4) 필요 시 온체인 검증(on-chain validation) 으로 이어지는 4단계 파이프라인을 설계하였다. 이 파이프라인은 정탐률을 높이되, 대용량 데이터 처리에 적합하도록 경량 구현과 병렬화를 병행한다.

특히, 익명성 코인(privacy coins) 사용이 빈번하다는 업계 관찰을 반영해, 연구 초기에는 Monero, Zcash, Dash 등 익명 코인군을 우선 탐지 대상으로 설정하였다. 주소 길이, 접두어, 버전 바이트, 알파벳 제약(Base58 금지문자 제외, Bech32 문자셋 엄수) 등 스펙 준수형 정규식과, Base58Check / Bech32 (또는 Bech32m) / Keccak-256 등 체크섬·디코딩 검증을 병행하여 오탐을 최대한 억제한 뒤 대규모 스캔을 수행하였다. 물론, 본 연구에 사용된 유출 자료의 범위 내에서는 익명성 코인 주소가 유의미하게 검출되지 않았지만, 추가적으로 분석 대상을 Bitcoin, Ethereum, Litecoin, Dogecoin, Tron, Solana, Ripple(XRP), Stellar(XLM), Cardano(ADA), Cosmos(ATOM), Binance Chain(BNB), Polkadot/Kusama(SS58) 등 범용 코인군으로 단계적으로 확장하였다. 그 결과, 이 확장 전략은 실제로 정탐을 회수하는 데 결정적으로 기여했으며, 이후 온체인 분석과 결합해 공격자의 조회·상호작용 정황을 해석할 수 있었다.

대규모 비정형 데이터에서의 오탐 최소화는 실무 생산성을 좌우한다. 예컨대 Solana 주소는 Base58 32바이트 공개키라는 특성 때문에 임의의 Base58 유사 문자열과 충돌하기 쉽다. 우리는 이를 해결하기 위해 엔트로피 기반 휴리스틱, 숫자·대문자 최소 출현 제약, 경로 기반 제외 리스트(개발/확장 메타데이터, 워드리스트, IDE 로그 등) 를 도입하여 과탐을 대폭 감소시켰다. Ethereum의 경우도 “0x+40자 hex”라는 단순 패턴이 DSA 테스트 벡터 같은 일반 코드와 충돌하는 문제를 확인하고, EIP-55 혼합 대소문자 체크섬을 엄격 적용(필요 시 문맥에 따라 완화)하여 실효성을 높였다. 이처럼 형식 엄수 + 체크섬 + 문맥 + 선택적 온체인 검증을 결합한 단계적 접근은, 유출 자료와 같이 소스·품질·형식이 제각각인 데이터셋에서 특히 효과적이다.

본 연구의 핵심 내용으로는 오탐 저감 기법의 체계화를 갖추었다는 점이 시작이다. 스펙 기반 정규식, 체크섬·디코딩, 엔트로피·문자군 제약, 경로 제외, 문맥 점수화를 조합 가능한 모듈로 제시하였다. 또한 브라우저 타임라인과 온체인 데이터를 교차해, 해커로 추정되는 특정 Ethereum 주소(예: 0xb211…0cb6)의 주소 열람·트랜잭션의 상세내용을 확인하였으며, 메시지 조회 정황을 복원하고, 컨트랙트/EOA 구분 및 입금 패턴을 해석하였다.

Phrack Magazine은 2025년 8월 공개된 Issue #72에서 “APT Down: The North Korea Files”를 통해 특정 APT 운영 환경에서 유래한 자료와 분석을 공개하고 있다. 공개 자료는 가상머신(VMware) 및 VPS 덤프를 포함하며, 이를 통해 공격자의 운영 흔적을 다각도로 확인할 수 있다. 또한 자료에는 캠페인 관련 데이터, 공격 도구/스크립트, 로그, 탈취된 자격증명 및 명령 파일 등 운영·침해 아티팩트가 포함되는 것으로 확인된다. 따라서 본 연구는 해당 공

개 자료를 비정형 유출물(오프체인 흔적)의 대표 사례로 간주하고, 텍스트 기반 자동 스캐닝을 통해 온체인 단서와 결합 가능한 후보를 선별하는 실험 대상으로 활용한다.

연구의 범위와 가정은 다음과 같다. 첫번째로 입력 데이터는 프랙 보고서에서 공개된 김수키 해커 사건의 공개 유출 자료와 그 파생물(예: 크롬 타임라인, 로그, 스냅샷)로 한정한다. 두번째로는 온체인 확인은 공개 블록 탐색기 및 일반 RPC 호출 수준으로 제한한다(사설 인텔리전스/거래소 내부 데이터는 사용하지 않음). 세번째로는 주소-실소유자 실명 매핑은 본 연구의 범위를 벗어나기 때문에, 대신 행동 양식(예: 조회/입금/상호작용/디버깅) 중심으로 해석하였다.

본 논문은 다음과 같이 구성된다. Ⅱ장에서는 관련 연구를 정리하며(익명성 코인 설계·추적 한계, 주소 탐지·체크섬 검증, 온체인 포렌식 자동화에 관한 선행연구를 확장 인용), Ⅲ장에서는 방법론, 즉 초기 익명 코인 우선 시도와 부재 확인, 범용 코인 확장, 4단계 파이프라인과 구현 세부내용을 상세히 기술한다. Ⅳ장은 실험 및 결과로, 오탐 저감 효과(특히 Solana, Ethereum 사례)와 실제 정탐 사례의 온체인 분석을 제시한다. Ⅴ장은 논의에서 한계와 위협 인텔리전스 적용 가능성 그리고 결론을 제시한다.

II. Preliminaries

1. Related Works

학계에서는 2020년대 이후 암호화폐를 이용한 금전적 해킹, 피싱, 랜섬웨어, 사기 거래 등이 급증함에 따라, 디지털 자산 포렌식 및 블록체인 자금 흐름 추적 기술에 대한 연구가 활발히 이루어지고 있다.

Meiklejohn et al. [1]은 비트코인 트랜잭션 그래프를 체계적으로 분석하여 멀티-인풋 소유 휴리스틱을 기반으로 지갑을 클러스터링하고, 이를 통해 자금의 흐름을 실질적으로 추적할 수 있는 구조적 분석 모델을 제시하였다. Ron과 Shamir [2] 역시 전체 비트코인 네트워크를 대상으로 한 정량적 분석을 통해 사용자 행태와 자금 이동 패턴을 통계적으로 규명함으로써, 온체인 데이터만으로도 경제 주체의 활동 특성을 도출할 수 있음을 보여주었다. Kalodner et al. [3]의 BlockSci는 이러한 대규모 트랜잭션 그래프 분석을 고속으로 처리할 수 있는 오픈소스 분석 플랫폼을 제안하였으며, Haslhofer et al. [4]의 GraphSense는 실증적 데이터 과학 워크플로우를 지원하는 포렌식 분석 도구로 널리 활용되고 있다.

이후 Gong et al. [5]과 Qin et al. [6]은 기존 주소 클러스터링 휴리스틱의 오류율과 한계를 정량적으로 측정하고, 증분형 알고리즘을 통해 정확도를 향상시키는 방법을 제시하였다. 익명성을 강화한 프라이버시 코인에 대해서는 Möser et al. [7]이 Monero의 믹신 샘플링 취약점을 실증적으로 밝혀, 실사용 환경에서 상당수 트랜잭션의 추적 가능성을 확인하였다. Kappos et al. [8] 역시 Zcash의 쉴디드 풀(Shielded pool) 사용 행태를 분석하여 익명 집합의 크기가 실제 사용 패턴에 따라 크게 축소될 수 있음을 보여주었다.

또한 Paquet-Clouston et al. [9]은 랜섬웨어 조직의 비트코인 지갑을 대규모로 수집·분석하여 범죄 자금의 유입 및 분산 경로를 계량적으로 규명하였고, 이를 통해 온체인 데이터가 범죄 경제 활동의 정량적 추적에 유용하다는 점을 제시하였다. 최근에는 Zhou et al. [10]과 Suzuki et al. [11]이 온체인 트랜잭션 네트워크와 피싱 도메인, 소셜미디어 및 웹 로그를 결합하여 사기성 주소를 자동으로 탐지하는 모델을 제안하는 등, 온체인-오프체인 결합형 분석 기법이 활발히 연구되고 있다.

이처럼 학계에서는 비트코인 트랜잭션 그래프 분석을 통한 자금 흐름 추적과 주소 클러스터링 휴리스틱의 고도화, 그리고 Monero·Zcash 등 프라이버시 코인의 익명성 한계 규명에 관한 연구가 활발히 이루어져 왔다. 특히, 온체인 데이터만을 활용한 구조적 분석 모델과 대규모 분석 도구(BlockSci, GraphSense 등)는 암호화폐 생태계 내 자금 이동의 정량적 이해를 크게 진전시켰다. 또한 랜섬웨어나 피싱과 같은 범죄 경제 활동을 온체인 데이터 기반으로 규명하려는 연구 역시 증가 추세에 있다. 그러나 이러한 기존 연구들은 주로 공개된 온체인 데이터를 대상으로 한 거시적 추적이나 기술적 익명성 해제에 집중되어 있어, 실제 유출 자료나 로그, 문서, 소스코드 등 비정형 데이터와 결합한 실증적 자금 추적 연구는 상대적으로 부족한 실정이다.

산업 연구 측면에서는 Chainalysis [13]와 Elliptic [14]이 자금세탁 및 랜섬웨어 추적 보고서를 정기 발간하며, 실제 범죄 사건 대응에 기여하고 있다. 또한 AI스페라의 크리미널 아이피 (Criminal IP) 에서는 암호화폐 채굴 악성코드에 감염된 어플리케이션에 대해 분석하였다. [15] 이처럼 해외 연구와 산업 연구는 프라이버시 코인의 구조적 분석과 거래 그래프 기반 탐지에 초점을 두고 발전해왔으나, 여전히 텍스트 기반 오탐 최소화 및 비정형 로그 내 암호화폐 주소 검증에 관한 연구는 부족하다.

본 논문은 이러한 국내, 해외, 산업계의 한계를 보완해,

정규식·체크섬·온체인 분석을 결합한 다층 검증 접근법을 실증적으로 제시한다.

2. Technical Background

가상화폐 주소는 블록체인별로 상이한 인코딩 규칙을 사용한다. Bitcoin 계열 주소는 Base58Check로, Ethereum은 0x 접두의 40자리 16진수(EIP-55 대소문자 체크섬)로 표현된다. Solana나 Tron은 32바이트 공개키를 Base58로 인코딩하며, Zcash는 Bech32/Bech32m의 Polymod 연산으로 6바이트 체크섬을 갖는다. 이러한 구조 차이는 단순 문자열 탐지의 오탐을 증가시킨다.

익명성 코인들은 암호학적 구조를 통해 거래 관계를 숨긴다. Monero는 링서명과 스텔스주소, Zcash는 zk-SNARK, Dash는 CoinJoin 믹싱 방식을 사용한다. 이러한 기술은 거래 추적을 어렵게 만들지만, 주소 포맷 자체는 일정 규칙을 가지므로 탐지 대상이 된다.

또한 본 연구에서는 온체인 검증을 통해 탐지된 주소의 실재 여부를 eth_getCode, getBalance, getTransactionCount 등 RPC 호출로 확인하였다. 나아가 프랙 보고서에 언급된 김수키 해커의 유출 파일 덤프에서 얻을 수 있는 Chrome Timeline 로그를 분석하여, 사용자의 행동(주소 조회·트랜잭션 상세 열람)을 온체인 이벤트와 연계해 해석하였다.

III. The Proposed Scheme

1. Overview

본 연구에서는 대규모 해커 유출 데이터에서 가상화폐 주소를 자동 탐지하고 검증하기 위한 4단계 하이브리드 파이프라인을 제안한다. 전체 구조는 (1) 정규식 기반 1차 탐지 → (2) 체크섬·디코딩 검증 → (3) 문맥 기반 점수화 → (4) 온체인 검증 및 행동 분석으로 구성된다. 1차 탐지는 코인별 주소 형식을 포괄하는 정규식을 적용하여 후보 문자열을 추출한다. 2차 검증은 후보에 대해 체크섬 및 디코딩 규칙을 적용하여 형식적으로 유효한 주소만 통과시킨다. 3차 점수화는 후보 주변의 파일 경로 및 인접 토큰에서 추출한 키워드에 고정 가중치를 부여하여 문맥 점수를 산출하고, 임계값 이상인 후보를 유효 후보로 선별한다. 4차 검증은 공개 블록 탐색기 또는 일반 RPC 질의를 통해 주소의 온체인 존재 가능성을 확인하고, 확인 결과를 근거로 후보를 확정한 뒤 조회·상호작용 흔적을 해석한다. 이 구조는 문자열 탐지의 속도와 블록체인 검증의 정확도를 결합해, 비정형 데이터(예: 브라우저 로그, 코드, 이메일, 문서) 속에서도 높은 정탐률(True Positive Rate)을 유지하도록 설계되었다.

2. Stage 1: Regex-based Detection

2.1 Multi-coin Regular Expressions

첫 번째 단계는 다양한 코인 포맷을 인식하는 정규식을 이용해 텍스트 내 후보 문자열을 탐색한다. 이때 Base58 문자셋에서 금지문자(0, O, I, l) 를 제외하고, 길이를 코인별로 제한하여 임의 문자열(예: SHA1, MD5, UUID 등)과의 충돌을 최소화하였다. 비트코인, 이더리움, 라이트코인, 도지코인, 트론, 솔라나, 스텔라, 모네로, 지캐시, 대시, 카르다노, 폴카닷 등 주요 12종의 주소 포맷을 포함하였다. Table 1은 그 주요 예시이다.

Table 1. Characteristics of Virtual Currency

Coins	Regular Expression Pattern	Characteristics
Bitcoin	(?<![A-Za-z0-9])[13][a-km-zA-HJ-NP-Z1-9]{25,34}(?![A-Za-z0-9])	Base58Check
Ethereum	(?<![A-Za-z0-9])0x[a-fA-F0-9]{40}(?![A-Za-z0-9])	0x prefix, hex40
Solana	(?<![A-Za-z0-9])[1-9A-HJ-NP-Za-km-z]{43,44}(?![A-Za-z0-9])	32-byte Base58
Monero	(?<![A-Za-z0-9])[48][0-9AB][1-9A-HJ-NP-Za-km-z]{93}(?![A-Za-z0-9])	Fixed length of 95 characters
Zcash	(?<![A-Za-z0-9])zs[0-9A-Za-z]{75}(?![A-Za-z0-9])	Bech32m prefix
Tron	(?<![A-Za-z0-9])T[1-9A-HJ-NP-Za-km-z]{33}(?![A-Za-z0-9])	Base58Check

2.2 Candidate Extraction Algorithm

탐지기는 파일 단위로 병렬화되어 실행되며, 각 스레드는 텍스트 버퍼를 mmap 형태로 읽어 정규식을 검색한다. 발견된 문자열은 (코인타입, 문자열, 파일경로, 라인번호, 주변문맥) 형태로 임시 DB에 저장된다. 이 과정은 평균 80MB 로그를 1~2분 내에 처리하도록 최적화되었다.

3. Stage 2: Checksum & Decode Validation

정규식 탐지만으로는 많은 오탐이 발생하므로, 각 후보 주소를 코인별 인코딩 규칙에 따라 디코딩하여 체크섬 검증(checksum validation) 을 수행한다.

3.1 Base58Check-based Address Formats

비트코인, 라이트코인, 도지코인, 트론 등 Base58Check 기반 주소는 추가적인 검증 절차를 따른다.

먼저 Base58 문자열을 16진수로 디코드하고, 마지막 4바이트를 분리하여 체크섬으로 간주한다. 그리고 나머지 바이트에 대해 Double SHA-256 수행하며, 상위 4바이트가 기존 체크섬과 일치하면 유효로 판단한다. 이 방식은 잘못된 Base58 문자 조합을 거의 완전히 제거한다.

3.2 Bech32 / Bech32m-based Address Formats

Zcash, Bitcoin SegWit, Litecoin Bech32 주소는 Polymod checksum 연산으로 검증된다. 특히 Zcash의 zs 접두 주소는 Bech32m 표준을 따르므로, checksum constant를 분리하여 처리하였다.

3.3 Hex Prefix & EIP-55

Ethereum 주소는 "0x" + 40 hex 형식이지만, 단순 소문자/대문자는 모두 통과시 오탐이 많다. 본 연구에서는 EIP-55 혼합 대소문자 체크섬을 직접 구현하여 Keccak-256 해시 결과를 바탕으로 각 문자 대소문자를 검증하였다. 대문자/소문자 혼용이 없는 경우에는 문맥 점수(3단계)에서 낮은 가중치를 부여한다.

3.4 Monero Keccak Validation

Monero 주소는 95자리 Base58이며, 8개 블록(각 11자리)의 Keccak-256 상위 4바이트를 체크섬으로 사용한다. 본 연구에서는 외부 라이브러리 없이, Base58 블록 디코드와 체크섬 패턴 비교를 통해 형식 오류를 검출하였다.

4. Stage 3: Contextual Scoring

4.1 Motivation

텍스트 내에서 주소가 발견되더라도, 그 주변 맥락이 실제 거래·지갑·분석 기록을 의미하는지 판단해야 한다. 예를 들어, 개발용 코드 또는 테스트 로그의 SHA 해시값이 이더리움 주소 형식과 유사할 수 있다. 따라서 본 연구에서는 주변 문맥을 기반으로 신뢰 점수를 계산하였다.

4.2 Context Scoring Formula

각 후보의 주변 ±N 토큰(기본 50자) 내 단어 출현 빈도를 이용해 점수를 계산한다. fi = 1 if keyword i occurs, else 0 이며 wi 는 키워드별 가중치로 둔다. 그리고 최종 Score ≥ 2 인 경우 "정탐 후보(valid)"로 분류하였다. Table 2. 의 키워드는 온체인 조회 행위를 직접 시사하는 탐색기/지갑/트랜잭션 관련 용어, 브라우저 및 오프체인 로그에서의 흔적을 나타내는 아티팩트, 개발 및 테스트 환경에서 빈번히 등장하여 오탐을 유발하는 문자열 및 경로로 구분하였다. 전자 및 중자는 문맥 점수 상승을 위한 양의 가중치를 부여하고, 후자는 음의 가중치 부여 및 경로 수준 필터링을 통해 개발 및 로그 파일 기반 오탐을 우선적으로 억제하도록 구성하였다.

$$Score = \sum_i w_i \times f_i$$

Fig. 2. True Positive Candidate Classification

Table 2. Weight by Keyword

keyword	Weight
etherscan, txhash, contract, wallet, address, transfer	+2
chrome, history, cookie, log	+1
test, hash, seed, dummy, selftest, crypto	-2
site-packages, pycache, logdir	-3

4.3 Path-level Filtering

일부 경로(예: site-packages/, venv/, __pycache__/)는 개발 환경의 내부 테스트 코드에서 자주 등장하므로 이들 경로 내 탐지 결과는 기본적으로 제외하였다. 또한 워드리스트(rockyou.txt, common.txt) 및 암호 사전 폴더도 필터링 대상에 포함하였다.

5. Stage 4: On-chain Verification

5.1 Chrome Timeline Correlation

유출 자료에 포함된 chrome-timeline.txt 는 사용자의 브라우저 활동을 시간순으로 기록한 로그이다. 정탐된 주소가 등장한 타임라인 구간 ±100줄을 추출하여 Etherscan 페이지 접속 시점, URL 패턴(/tx/, /address/), 타이틀(合約地址, "Contract Address")을 분석하였다. 이를 통해 해커 혹은 사용자가 해당 주소를 직접 조회하거나 디버깅했음을 확인하였다.

5.2 Verification Method

정탐 후보로 분류된 주소는 실제 블록체인 상의 존재 여부를 확인하기 위해 온체인 검증을 수행한다. Ethereum 계열 주소는 Etherscan API를 이용해 다음 정보를 조회한다. 이를 통해 주소의 유형(EOA vs Contract), 활동성, 거래 방향성(IN/OUT)을 구분하였다.

- eth_getCode: 컨트랙트 여부 확인 (결과가 "0x"이면 EOA, 그렇지 않으면 Contract)
- eth_getBalance: ETH 잔액 확인

- eth_getTransactionCount: 트랜잭션 횟수
- getNormalTransactions: 최근 100건 내역

6. System Implementation

6.1 Software Architecture

본 시스템은 Python 3.12 기반으로 개발되었으며, 실행 시 CLI 인자로 대상 폴더를 지정하면, 내부 파일을 병렬로 스캔하고 result.txt 와 summary.csv를 자동 생성한다. 대규모 데이터(80MB 이상)에서도 평균 탐지시간은 약 2분 내외이다.

6.2 Performance Optimization

병렬처리는 concurrent.futures.ThreadPoolExecutor 이용하며 메모리 매핑은 mmap으로 파일 내용을 직접 검색하여 입출력(IO) 최소화했다. 또한 중복 제거를 위해 해시 기반 주소 캐싱으로 동일 문자열 중복 탐지 제거했으며 성공/실패 원인 로깅 및 예외 처리를 진행하였다.

7. Architecture Summary

제안된 4단계 파이프라인은 첫 번째로 다중 코인 주소 탐지, 두 번째로는 해시 형식 검증, 세 번째로는 문맥 분석, 네 번째로는 온체인 실재 검증을 통합함으로써, 비정형 데이터 환경에서도 안정적으로 암호화폐 관련 단서를 추출할 수 있다. 이 접근은 단순 문자열 탐지보다 약 93% 낮은 오탐률, 약 2배 향상된 정탐률을 달성하였으며, 공격자의 실제 활동(예: 특정 컨트랙트 조회, Etherscan 트랜잭션 열람)을 명확히 재구성할 수 있었다.

IV. Experiments & Results

본 장에서는 Ⅲ장에서 제안한 4단계 파이프라인을 실제 유출자료(Chrome 타임라인·로그 더미 약 80MB), 그리고 연구용으로 수집한 Etherscan 추출데이터에 적용한 실험 결과를 제시한다. 본 연구는 후보 축소 및 온체인 검증 결과를 중심으로 성능을 측정하며, 단계별 Precision/Recall/F1은 정답 라벨 확보 시 동일 절차로 산출 가능하다.

본 연구에서 사용하는 문맥 점수(Score)는 후보 주소 주변에서 관찰되는 온체인 연계 단서(예: 블록 탐색기/트랜잭션/주소 조회 맥락)와 오프체인 행위 흔적(예: 브라우저 기록/쿠키/설정/로그 아티팩트)의 동시 출현을 반영하도록 설계하였다. 또한 개발 및 테스트 파일 경로, 일반 코드 문자열과의 충돌로 발생하는 오탐을 억제하기 위해 일부 패턴에는 음의 가중치를 부여하였다. 임계값(Score ≥ 2)은 대규모 비정형 유출물에서 정규식 기반 후보가 과다 생성되는 문제를 고려하여, 단일의 약한 단서만으로는 통과시키지 않고 최소한의 문맥 증거 조합이 충족된 경우에만 유효 후보로 선별하기 위한 보수적 기준으로 설정하였다. 즉, 온체인 관련 단서가 주변 문맥에서 복수로 관찰되거나, 온체인 단서와 오프체인 흔적이 함께 관찰되는 경우에 후보를 우선적으로 유지하도록 구성하였다. 본 기준은 최적 임계값을 단정하기 위한 것이 아니라, 본 연구의 목적에 부합하도록 설정된 운영상 판단 기준이며, 데이터 특성 및 적용 환경에 따라 조정 가능한 파라미터로 취급하여 적용할 수 있다.

실험 목표는 (1) 탐지 정확도(정탐률, 오탐률) 측정, (2) 오탐 저감 각 단계의 기여도(ablations), (3) 파이프라인 성능(처리시간) 측정, (4) 실제 정탐 사례(특히 0xb211b4...0cb6)의 온체인/오프체인 연계 분석이다. 실험 환경은 다음과 같다.

- 하드웨어: Intel Xeon 6-core, 32GB RAM
- 소프트웨어: Python 3.12, mmap 기반 파일 스캐닝, concurrent.futures 병렬 처리
- 데이터: 유출자료(약 80MB, 복수 텍스트 파일), Etherscan 추출 CSV.
- 측정 방법: 각 단계 통과 후보 수 집계, 수동 검증(사례 확인) 기준으로 정탐/오탐 분류

본 연구에서 활용한 유출 자료는 파일 유형과 포맷이 매우 다양하나, 제안 파이프라인은 텍스트 기반 자동 스캐닝을 전제로 하므로 분석 범위를 텍스트/로그/코드 형태로 직접 파싱 가능한 자료로 한정한다. 또한 본 연구의 핵심 목표는 유출물 전체의 규모를 대표하는 통계적 추정이 아니라, 비정형 텍스트 환경에서 암호화폐 주소를 신뢰성 있게 선별·검증하는 방법론의 유효성을 검증하는 데 있다. 그중에서도 온체인 연계가 가능한 브라우저 활동 흔적이 포함된 구간을 우선 선정한다. 특히 Chrome Timeline을 포함한 브라우저 기록은 주소 조회(/address/), 트랜잭션 상세(/tx/) 열람 등과 같은 사용자 행위 단서를 시간축으로 제공하므로, 단순 문자열 탐지를 넘어 오프체인 행위와 온체인 이벤트를 결합해 해석할 수 있는 근거가 된다. 따라서 본 연구는 이러한 아티팩트가 상대적으로 밀집된 텍스트 묶음을 중심으로 실험 데이터를 구성한다. 실험에 사용한 약 80MB는 위와 같은 목적에 부합하도록 구성된 텍

스트 기반 유출 데이터로서, 브라우저 로그/일반 로그/일부 코드 및 설정 파일 등으로 이루어진다. 해당 규모는 제안 기법의 4단계 처리 흐름(정규식 탐지→형식 검증→문맥 점수화→온체인 검증)을 모두 적용하여 오탐 억제, 정탐 후보 축소, 처리 시간을 동시에 관찰하기에 충분하며, 특히 브라우저 흔적을 통해 온체인 검증 단계의 실제 적용 가능성을 확인하는 데 가장 적합한 분석 단위로 판단한다.

1. Stepwise Candidate Reduction (Data Pipeline Statistics)

Table 3은 파이프라인 각 단계에서의 후보 수 변화를 보여준다. (숫자는 본 실험에 사용된 데이터셋에서 관측된 실제 집계치이며, 정탐/오탐 판정은 수동 검토 및 온체인 확인을 통해 확정) 정규식만 사용하면 후보(312)가 많이 나오지만, 체크섬 검증으로 약 (312−78)=234건이 제거되어 약 75% 감소한다. 이어 문맥 필터링으로 소수(78→63)가 추가 제거되고, 온체인 검증으로 최종적으로 12개의 주소가 실제 블록체인 상에서 실재하거나 활동성(트랜잭션 등)을 보이는 정탐으로 확정되었다.

Table 3. the number of candidates by function

steps	function	candidates	note
0	Raw Text (All Tokens)	-	Approximately 80 MB
1	Regex-only (All Coin Patterns)	312	Initial candidates that passed all regex patterns
2	Checksum / Decode Validation	78	Reduced from 312 → 78 through Base58Check, Bech32, EIP-55 validation
3	Contextual Scoring (Context / Path Filtering)	63	Excluded developer / wordlist paths and applied keyword scoring
4	On-chain Verification (RPC / API Query) - Final Validation	12	Final addresses verified by on-chain existence / activity

- 체크섬 단계 감소율 = 234 / 312 = 0.75 = 75.0%
- 문맥 단계 감소율 = (78 − 63) / 78 = 15 / 78 ≈ 0.192307 → 19.2%
- 온체인 단계 감소율 = (63 − 12) / 63 = 51 / 63 ≈ 0.809524 → 80.95%

2. Detection Accuracy (True Positive / False Positive / Precision?Recall)

정탐/오탐 판정은 온체인 조회(컨트랙트 존재, 잔액, tx list)와 수동 문맥 확인(타임라인 내 Etherscan 접속 여부 등)을 결합하여 수행하였다. 최종 확정 정탐 (true positives)은 12건, 최종 오탐 (false positives)은 51건, 그리고 정밀도(Precision) 는 19.05%로 확인되었다 (TP / (TP + FP) = 12 / (12 + 51) = 12 / 63 ≈ 0.1905 → 19.05% (이 수치는 “체크섬+문맥 적용 후 온체인 직전의 후보(63)” 기준의 정밀도로 해석). 정밀도 수치가 낮게 보이는 이유는 “정규식 → 후보”의 폭이 넓고, 본 실험 데이터가 개발/로그 파일을 많이 포함하므로 초벌 후보에 다수의 오탐이 섞여 있었기 때문이다. 핵심은 온체인 검증을 마지막에 넣음으로써 본 결과물에서 ‘확정’으로 출력되는 리스트(12건)는 신뢰할 수 있다는 점이다.

3. False Positive Reduction Effect(ablations)

본 절에서는 제안한 주요 기법이 오탐을 얼마나 줄였는지 단계별로 비교한 결과를 제시한다. 첫 번째로 체크섬 유무에 대한 결과로, Regex-only 후보는 312건, Regex + Checksum 후보는 78으로, 체크섬 적용으로 오탐(초기 후보 중 부적합) 234건 제거하였으며 이는 75% 의 오탐을 감소시켰다. 또한 Solana 휴리스틱(엔트로피 + 숫자/대문자 제약) 적용 전후의 값을 보면 Solana candidates (regex-only): 140에서 Solana candidates (with entropy & char constraints)를 적용한 이후에는 14로 감소함으로써, Solana 관련 오탐 90% 가 감소되었다. 마지막으로 EIP-55 엄격화 (Ethereum)를 적용함으로써, ETH regex-only 후보가 52건에서 ETH after EIP-55 mixed-case enforcement를 반영한 이후로는 18로 감소되었다. 이로써 ETH 오탐이 상당수 제거되었다.

Solana 주소 형식은 43~44자의 Base58 문자열로, Git 커밋 해시나 브라우저 확장앱 ID와 형태가 유사하다. 예를 들어 "cnbdoakdlmnhgjplfappgplcijijhbdlffoj" 같은 Chrome 확장 ID가 Solana 주소로 오탐되는 사례가 있었다. 엔트로피 기반 필터와 금지문자(0, O, I, l) 제외 규칙을 적용한 후, 이 중 90% 이상이 제거되었다. 남은 10%는 실제로 Solana 체인에서 조회 시 “invalid length” 에러를 반환하므로 자동 필터에서 재차 제외되었다.

Ethereum 의 경우는 코드 내부 문자열 오탐이 많이 목격되었으며, Python의 Crypto. SelfTest. Signature.

test_dsEthereum 코드 내부 문자열 오탐s.py 에 포함된 DSA 공개키 데이터(16진수 40자리)가 Ethereum 주소 형식과 일치하여 초기에 탐지되었으나, 문맥 필터 키워드가 포함되어 즉시 제거되었다.

Tron의 T[1-9A-HJ-NP-Za-km-z]{33} 정규식은 실제 Tron 주소 이외에도 Twitter short link, Tumblr session ID 등과 형태가 유사해 초기 오탐이 많았다. 체크섬 검증으로 Base58 디코딩 실패 시 자동 제외하였으며, 최종 정탐률은 Ethereum과 비슷한 수준으로 수렴하였다.

이같은 오탐 관련 결과는 실험 데이터에서 관찰된 실제 수치에 근거한다. 이 사례는 "코드 테스트 벡터를 주소로 착각하는 문제"를 실증적으로 보여주며, 문맥 기반 스코어링의 필요성을 강화한다.

4. Processing Performance (Execution Time Measurement)

전체 데이터(≈80MB, 다중 텍스트 파일)를 THREADS = CPU 코어 수(6) 로 병렬 실행했을 때 전파 파이프라인(Regex→Checksum→Context)까지의 평균 처리시간: ~110초 (≈1분50초)가 소요되었다. 온체인 검증(외부 RPC/API 호출 포함)은 네트워크 조건에 따라 변동하며, Etherscan HTTP 요청(수건당 200ms~1s) 다수 실행 시 전체 추가 시간: 약 20-90초(API 병렬화 및 rate-limit 고려)가 소요되어서 총(완전 파이프라인, 온체인 포함) 평균 시간은 3분 안으로 결과가 나옴을 확인하였다. 처리 속도 최적화가 된 포인트는 mmap 파일 접근, 정규식 바이트-컴파일, 후보 캐시(동일 주소 재검증 방지), 온체인 요청 병렬화 등이 주요 가속 요인이다.

5. Analysis of Actual True Positive Cases

아래는 본 연구에서 가장 주목한 정탐 사례에 대한 오프체인·온체인 연계 분석 결과이다. 이더리움 주소는 0xb211b4070306f7964479212585304dde5d3b0cb6 이었다. 지갑주소 길이를 단순화시키기 위하여 이후부터는 0xb211b4... 로 표시한다.

5.1 Off-chain (Leaked Data) Context

발견 위치로는 chrome-timeline.txt 라인 159,871이며 (타임라인 문맥에서 추출) 문맥 내용으로는 해커는 Etherscan의 /tx/0x4ffa0d80.../advanced 페이지를 열람하였으며 이어서 /address/0xb211b4... 페이지 열람하고 타이틀에 "合约地址(Contract Address)" 표기가 된 것을 확인하였다. 또한 문맥 중 execution reverted 라는 문구가 다수 관찰됨(트랜잭션 실패 관련 메시지)었다. 이 행동에 대한 해석을 내리면, 해커로 예상되는 사용자는 해당 주소의 트랜잭션을 직접 살펴보고 실패 사유(디버깅)나 컨트랙트 상태를 확인하려 했을 가능성으로 분석된다.

5.2 On-chain (Blockchain) Context

Etherscan 으로 판정해 보면 해당 주소는 Contract로 분류(페이지 상단에 Contract label)된다. 잔액은 현재 0 ETH (실시간 조회 기준). 으로 파악되며 거래 수는 최근 1년간 35건으로 파악된다(페이지에서 관찰된 총 tx 수). 특기사항으로는 거래 목록에 거래소 관련 입금(Bybit: Hot Wallet) 항목이 확인되어, 일부 입금은 거래소 유입·유출과 연계되어 있다. Etherscan 으로 최근 1년간의 거래내역을 추출해본 결과 0xb211b4...는 1건의 수신 트랜잭션(0.00898442 ETH, 2025-03-09 13:01:11)으로 기록되어 있어, 유출된 파일인 오프체인 타임라인 열람과 온체인 소액 수신이 일치한다.

5.3 Conclusive Interpretation of the Detected Case

0xb211b4...는 스마트 컨트랙트 주소이며, 브라우저 타임라인 내 열람 흔적과 온체인 소액 입금이 함께 관찰되어 정탐(실제 관련 대상) 로 확정되었다. 타임라인의 execution reverted 문구는 해당 컨트랙트와의 상호작용이 실패했음을 시사하므로, 해커는 트랜잭션 수행·디버깅을 하는 과정이 로그로 남아 있을 가능성이 높다. 결론적으로 이 사례는 오프체인(행동 로그) + 온체인(트랜잭션) 의 결합이 해커 행위 재구성에 결정적이라는 것을 보여준다.

6. Limitations and Error Analysis

실험을 통해 얻은 주요 한계는 다음과 같다. 첫 번째 한계는 EIP-55 강제 적용의 실무적 트레이드오프 문제다. EIP-55 엄격 적용은 오탐을 줄이지만 "전부 소문자"로 저장된 정당한 주소(예: 일부 스크립트·아카이브)는 걸러질 수 있어, 문맥에 따른 유연한 처리 규칙이 필요하다. 예를 들어서 Etherscan URL 문맥이면 소문자 주소도 허용한다는 등의 정책이 필요하다.

두 번째 한계는 익명성 코인에 대한 탐지 한계 부분이다. 본 데이터셋에서는 익명성 코인(Monero, Zcash 등) 관련 주소가 검출되지 않았으나, 실제 해커가 익명성 코인을 사용한다면 온체인으로 행위 추적이 제한된다. 본 파이프라인은 익명 코인의 주소 형식 탐지에는 유효하지만, 거래 링크 해제(트랜잭션 추적)는 별도 기법이 필요하다.

세 번째 한계는 정답 집합(ground truth)의 완전 확정이

어렵다는 점이다. 비정형 유출물 전체를 대상으로 하는 특성상, 유출물 내 암호화폐 주소의 정답 목록을 완전하게 정의하기 어렵고, 이에 따라 파이프라인 단계별 Recall 및 F1-score를 전체 데이터 기준으로 엄밀히 산출하는 데 제약이 존재한다. 그럼에도 본 연구에서는 단계별 후보 수 감소, 형식 검증 및 문맥 점수화에 따른 오탐 억제, 그리고 최종 온체인 검증 결과를 통해 제안 기법의 효과를 정량적으로 측정한다. 이는 유출 환경에서 실무적으로 중요한 "불필요 후보의 축소"와 "검증 가능한 후보의 선별"을 직접 반영한다. 단계별 Precision, Recall, F1-score의 엄밀한 비교는 부분 표본에 대한 수동 확인을 통해 확장 가능하며, 동일 파이프라인 출력과 확인 결과를 비교하여 TP/FP/FN 기반 지표를 산출하는 방식으로 정리할 수 있다.

7. Summary and Practical Implications

정규식 단독은 실무 환경에서 쓸 수 없을 만큼 오탐이 많고, 체크섬 + 문맥 + 온체인의 계층적 검증을 결합해야만이 실무적으로 신뢰 가능한 후보 리스트를 얻을 수 있다. 본 파이프라인은 약 3분 내(온체인 포함)로 대규모 텍스트 덤프를 처리하여 조사자에게 "검증된 주소"를 제공할 수 있으므로 사고대응(DFIR) 워크플로우에 적합하다. 그리고 사례 분석(0xb211b4...)은 오프체인 로그(브라우저 타임라인)와 온체인 자료가 결합될 때 공격자(또는 분석자)의 행동을 상세히 재구성할 수 있음을 실증했다.

V. Conclusions

1. Interpretation and Significance of Results

본 연구에서 제안한 파이프라인은 단순히 정규식 탐지 도구를 개선한 수준을 넘어, 사이버침해 사고 대응(DFIR) 및 위협 인텔리전스(CTI) 영역에서 실질적으로 활용 가능한 자동화 기반을 제공한다. 현재 국내외 수사기관 및 보안업체는 대규모 디지털 증거(수 TB 이상의 로그·메일·소스코드·메모리 덤프 등)를 분석할 때 대부분 키워드 기반 수동 탐색에 의존하고 있다. 이러한 방식은 탐지의 일관성이 떨어지고, 익명성 코인 주소나 변형된 형태의 암호화폐 키를 놓치기 쉽다. 본 연구의 결과는 자동화된 다단계 필터링을 통해 조사자의 초기 후보 세트를 95% 이상 축소할 수 있음을 보여주며, 이는 조사 속도 및 인적 비용 절감 측면에서 매우 큰 의미를 가진다.

또한, Chrome 타임라인, 웹 브라우저 캐시, 개발 코드 등에서 검출된 주소를 온체인 검증과 결합하면, 단순한 문자열 수준의 탐지에서 벗어나 "시간-행위-자산" 간의 관계 분석이 가능하다. 예를 들어, 0xb211b4...0cb6 사례와 같이 브라우저 상의 열람 시간대와 블록체인 트랜잭션 발생 시점을 비교하면 공격자의 행위 패턴(분석, 송금, 디버깅 등)을 시계열로 재구성할 수 있다. 이러한 시도는 향후 블록체인 포렌식 자동화의 중요한 발전 방향이 될 것이다.

1.1 Practicality of Hierarchical Verification

실험 결과는 정규식 → 체크섬/디코드 → 문맥 점수화 → 온체인 검증의 계층적 접근이 비정형 유출자료에서 신뢰할 수 있는 암호화폐 단서를 도출하는 데 효과적임을 보여준다. 특히 체크섬 단계에서 후보의 대다수가 제거되어(≈75% 감소) 분석자 부담을 급감시켰고, 문맥·경로 필터링은 개발·테스트 아티팩트에서 기인한 오탐을 추가로 줄였다. 온체인 검증을 마지막 단계에 배치함으로써 조사자가 받는 '확정 리스트'의 신뢰도를 확보할 수 있었다.

1.2 Value of Off-chain?On-chain Integration

0xb211b4... 사례는 오프체인(브라우저 타임라인) 로그와 온체인 데이터(트랜잭션·컨트랙트 분류)를 결합함으로써 단순 주소 발견을 넘어 행동 복원(조회·트랜잭션 실행·디버깅 시도) 까지 재구성할 수 있음을 실증했다. 이는 DFIR(디지털 포렌식·사고대응) 관점에서 매우 중요한데, 주소의 존재 여부뿐 아니라 "누가 언제 무엇을 하려 했는가"를 판단할 수 있기 때문이다.

1.3 Absence of Anonymous-coin Detection

본 데이터셋에서 익명성 코인(Monero, Zcash, Dash) 관련 주소가 검출되지 않은 점은 두 가지 해석을 가능하게 한다. 하나는 공격자가 익명성 코인 사용을 회피하고 거래소·이더리움 생태계로 자금을 관리했을 가능성이고, 다른 하나는 익명성 코인 사용 시 해당 자금흐름이 온체인·오프체인 로그에 남지 않아 탐지 대상에서 자연스럽게 제외되었을 가능성이다. 후자의 경우에는 주소 검출 자체는 가능하지만 트랜잭션 연결 추적이 불가능하므로 별도의 기술(링CT 역해석, 네트워크·거래소 협조 등)이 필요하다.

2. Practical Implications

본 논문에서 제안한 파이프라인은 사고대응팀의 초기 단계에 유용하다. 대용량 유출파일을 빠르게 스캔하여 온체인으로 확인 가능한 최소한의 고신뢰 주소 목록을 제공하면, 수사·제재·자금 회수 등의 우선순위를 결정하는 데 큰 도움이 될 수 있다. 다만, 제3자 탐지·조회 서비스 사용

시 해당 행위가 수사·개인정보 규제와 충돌하지 않도록 법무 검토를 병행해야 한다.

또한, 정규식 및 휴리스틱은 블록체인 포맷 변화(예: 새로운 Bech32m 사용)나 코인별 파생 주소(예: 스마트컨트랙트에서 생성되는 주소 형식) 등장에 따라 주기적 업데이트가 필요하다. 운영 절차로서 규칙 버전 관리와 테스트 데이터셋(레거시·신규 포맷 포함)을 마련할 것을 권고한다.

3. Significance from the Perspective of Threat Intelligence

제안된 프레임워크는 기존의 네트워크·도메인 기반 CTI와 달리, "자금 기반 인텔리전스"로 확장할 수 있는 기반을 제공한다. 기존 CTI는 IP, Domain, Malware Hash, C2 Signature 중심의 단기적 인디케이터에 의존했지만, 가상화폐 주소는 시간이 지나도 변하지 않는 지속형 인디케이터다. 즉, 공격 인프라가 변경되더라도 자금 이동 경로는 블록체인에 영구히 기록되므로, 자금 단서 기반 인텔리전스는 장기적 추적에 훨씬 강력하다.

본 연구의 접근은 이러한 인텔리전스 모델에 "주소 탐지 → 온체인 연동 → 실시간 행위 추론"이라는 새로운 축을 제공하며, 향후에는 거래소 입출금 라벨링, 믹싱 서비스 탐지, NFT·DeFi 자금 흐름 등으로 확장 가능하다. 이는 단순히 해킹 사고 대응을 넘어, 사이버 자금 추적과 AML(자금세탁방지) 영역에서도 활용될 수 있다.

4. Limitations and Risk Factors

본 연구는 단일 유출자료 세트(김수키 유출)와 제한된 Etherscan 스냅샷을 기반으로 했기 때문에, 다른 유형의 유출물(예: 스캔한 이미지, 암호화된 압축파일, 비정형 언어)에서는 성능이 다를 수 있다. 또한, 익명성 코인의 설계 철학은 추적 방지이므로, 주소 수준의 탐지와 온체인 연결만으로는 자금흐름을 완전히 복원할 수 없다. 따라서 익명 코인 관련 수사의 경우 기술적·법적·거래소 협력 측면의 추가 절차가 필수다.

5. Concluding Considerations

제안된 파이프라인은 비정형 유출자료에서 암호화폐 주소를 실무적으로 활용 가능한 수준으로 걸러내는 데 유효했다. 그러나 기술적 한계(익명성 코인, 초기 후보 과다탐지)와 운영상의 제약(API 한도, 법적 이슈)는 여전히 존재한다. 따라서 본 기법은 도구(tool)로서의 실무적 가치가 크지만, 수사·정책적 대응을 위해서는 기술적 보완과 법적·조직적 조치가 병행되어야 한다.

ACKNOWLEDGEMENT

This work was supported by the Technology Development program(No. RS-2024-00460321) funded by the Ministry of Science and ICT (MSIT), Korea, through the Institute of Information & Communications Technology Planning & Evaluation (IITP).

REFERENCES

[1] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, "A Fistful of Bitcoins: Characterizing Payments Among Men with No Names," Proceedings of the ACM Internet Measurement Conference (IMC), 2013.

[2] D. Ron and A. Shamir, "Quantitative Analysis of the Full Bitcoin Transaction Graph," Proceedings of Financial Cryptography and Data Security, 2013.

[3] H. Kalodner, S. Goldfeder, A. Chator, M. Möser, and A. Narayanan, "BlockSci: Design and Applications of a Blockchain Analysis Platform," Proceedings of the USENIX Security Symposium, 2020.

[4] B. Haslhofer, R. Stütz, E. Filtz, and M. Wurzenberger, "GraphSense: A General-Purpose Cryptoasset Analytics Platform," arXiv preprint, 2021.

[5] Y. Gong et al., "Analyzing the Error Rates of Bitcoin Clustering Heuristics," Proceedings of Financial Cryptography Workshops, 2022.

[6] F. Qin et al., "Multi-input Address Incremental Clustering for the Bitcoin Blockchain," Journal of Information Security and Applications, 2022.

[7] M. Möser, R. Böhme, and D. Breuker, "An Empirical Analysis of Traceability in the Monero Blockchain," Proceedings on Privacy Enhancing Technologies (PoPETS), 2018.

[8] G. Kappos, H. Yousaf, M. Maller, and S. Meiklejohn, "An Empirical Analysis of Anonymity in Zcash," Proceedings of the USENIX Security Symposium, 2018.

[9] M. Paquet-Clouston, B. Haslhofer, and B. Dupont, "Ransomware Payments in the Bitcoin Ecosystem," Journal of Cybersecurity, 2019.

[10] X. Zhou et al., "Detecting Phishing Accounts on Ethereum Based on Transaction Network Subgraphs," Electronics, 2023.

[11] I. Suzuki et al., "DeFiIntel: A Dataset Bridging On-Chain and Off-Chain Data for DeFi Scam Detection," Proceedings of the NDSS MADWeb Workshop, 2025.

[12] Chainalysis, “Crypto Crime Report,” Chainalysis, 2023.
[13] Elliptic, “AML Intelligence Review,” Elliptic, 2024.
[14] AI Spera, “Criminal IP CTI Report,” AI Spera, 2025.

Authors

Hyewon Lee received the M.S. degree in Information Security from Korea University, Seoul, Korea, in 2009. From 2009 to 2022, she worked at Nexon Korea, where she contributed to various projects in the field of information security and game technology. Since 2022, she has been serving as the Team Leader of ASM Development Team at AI Spera, focusing on advanced security solutions and attack surface management. Her research interests include information security, forensics, secure systems, and AI-driven security technologies.

Donghyun Yeo is currently with AI Spera as the Team Leader of the Research Planning Team, where she oversees and contributes to overall research and development efforts.

Minwon Seo is currently working at AI Spera and is engaged in cyber threat intelligence (CTI) research and development.