УДК 004.056.53:004.62

А. Р. Магденко
А. В. Грицак

МЕТОД КРОСПЛАТФОРМНОЇ КОРЕЛЯЦІЇ ЦИФРОВИХ СЛІДІВ ДЛЯ ДЕАНОНІМІЗАЦІЇ СУБ'ЄКТІВ КІБЕРЗАГРОЗ

Вінницький національний технічний університет

Анотація

У тезах розглядається інноваційний метод кросплатформної кореляції цифрових слідів, спрямований на автоматизацію процесів деанонімізації суб'єктів тіньової економіки. Обґрунтовано архітектурну модель розробленого програмного комплексу, що інтегрує мікросервісний підхід (Apache Kafka) та гібридну систему баз даних (MongoDB, Neo4j). Запропоновано алгоритм детермінованої екстракції криптографічних ідентифікаторів із неструктурованих текстів, який усуває колізії форматів між різними блокчейн-мережами. Доведено, що синтез комунікаційних графів (Telegram, DarkWeb) та фінансових транзакцій дозволяє експоненціально підвищити ефективність розслідувань у кіберпросторі в режимі реального часу.

Ключові слова: OSINT, деанонімізація, графові бази даних, кібербезпека, блокчейн-аналітика, обробка природної мови, тіньова економіка.

Abstracts

The paper presents an innovative method of cross-platform correlation of digital footprints aimed at automating the de-anonymization processes of shadow economy subjects. The architectural model of the developed software complex, which integrates a microservice approach (Apache Kafka) and a hybrid database system (MongoDB, Neo4j), is substantiated. An algorithm for the deterministic extraction of cryptographic identifiers from unstructured texts is proposed, eliminating format collisions between various blockchain networks. It is proven that synthesizing communication graphs (Telegram, DarkWeb) and financial transactions exponentially increases the efficiency of real-time cyberspace investigations.

Keywords: *OSINT, de-anonymization, graph databases, cybersecurity, blockchain analytics, natural language processing, shadow economy.*

Вступ

Сучасний етап розвитку кіберпростору характеризується фундаментальною трансформацією архітектури кіберзлочинності. Вектори координації протиправної діяльності поступово зміщуються з ізольованих сегментів оверлейних мереж (DarkWeb) у площину загальнодоступних месенджерів із наскрізним шифруванням, ключовим з яких є Telegram. Зловмисники свідомо застосовують стратегію диверсифікації ризиків: месенджери використовуються для агресивного маркетингу, швидкої комунікації та реалізації автоматизованих схем формату B2C (Business-to-Consumer), тоді як класичні onion-маркетплейси залишаються незамінними для розміщення бекенд-інфраструктури та забезпечення системи ескроу (Escrow) при багатотисячних криптовалютних транзакціях [1]. Утворюється складна гетерогенна екосистема, функціонування якої зумовлює прогресуючу технологічну фрагментацію цифрових слідів.

Результати дослідження

Аналіз існуючих рішень та постановка проблеми. Критичний аналіз існуючого програмного інструментарію демонструє суттєві обмеження у можливостях автоматизованої кросплатформної кореляції цифрових артефактів. Наявну системну проблему можна охарактеризувати як «фрагментарність екосистеми розслідувань». Класичні OSINT-фреймворки (зокрема Maltego) дозволяють агрегувати дані з різнорідних джерел, проте вимагають ручного керування графами та не здатні автоматизовано аналізувати контекст тисяч неструктурованих повідомлень [2].

З іншого боку, платформи професійної блокчейн-аналітики класу Enterprise (Chainalysis Reactor, Crystal) використовують потужні евристики кластеризації для відстеження руху коштів, проте їхнім фундаментальним недоліком є абсолютна «сліпота» до даних поза межами блокчейну (Off-chain data).

З математичної точки зору цю проблему описано як ізольованість графів: існуючі інструменти будують граф фінансових транзакцій $G_f = (V_f, E_f)$ та комунікаційний граф $G_c = (V_c, E_c)$ як абсолютно незалежні топологічні структури [3]. Відсутність автоматизованої функції відображення $f : V_c \to V_f$, яка б дозволила екстрагувати криптовалютні адреси з текстового контексту закритих чатів і прив'язувати їх до фінансових потоків, створює критичну прогалину в методології проактивної деанонімізації.

Проєктування мікросервісної архітектури та конвеєра даних. З метою подолання виявлених емпіричних обмежень розроблено архітектурну модель модульного асинхронного скрапінгу з імплементацією мікросервісного підходу. Під час експериментальної експлуатації базових монолітних прототипів збору даних було виявлено проблему асиметрії швидкості (I/O Bottlenecks): швидкість парсингу інформації з Telegram-каналів суттєво перевищувала пропускну здатність баз даних, що призводило до втрати цифрових доказів.

Принциповим удосконаленням стало впровадження проміжного шару потокової обробки подій (Event Streaming Layer) на базі розподіленого брокера повідомлень Apache Kafka [4]. Колектори (Telegram Userbot, Tor Headless Browsers) асинхронно публікують зібрані сирі JSON-об'єкти у тематичні черги (Topics) за патерном «видавець-підписник» (Publish-Subscribe). Це створює надійний буферний шар, який математично гарантує збереження всіх перехоплених комунікацій навіть у випадку пікових навантажень або тимчасової відмови обчислювальних вузлів.

Специфіка розслідувань на основі відкритих даних вимагає імплементації гібридної (поліглотської) архітектури зберігання. Для забезпечення суворих криміналістичних вимог щодо незмінності первинних цифрових доказів (Data Provenance) спроєктовано «Озеро Даних» (Data Lake) на базі документо-орієнтованої СУБД MongoDB. Кожен артефакт зберігається як атомарний BSON-документ із додаванням мітки часу (NTP) та криптографічного хешу за алгоритмом SHA-256. Це формує безперервний ланцюг збереження доказів (Chain of Custody), унеможливлюючи подальшу модифікацію даних [5]. Водночас для зберігання вилучених сутностей та швидкого обчислення транзитивних кореляцій впроваджено спеціалізовану графову базу даних Neo4j, чия архітектура первинно оптимізована для роботи з багатовимірною топологією. Інтеграція цих підсистем реалізується через програмний механізм перехресних посилань (Cross-Referencing об'єктних ідентифікаторів).

Стохастичне моделювання поведінки колекторів та обхід антифрод-систем. Критичним бар'єром під час автоматизованого збору розвідувальних даних у прихованій мережі Tor є протидія сучасним системам виявлення ботів (Bot Mitigation Systems), які імплементують багаторівневі алгоритми поведінкового та апаратного аналізу. Для забезпечення безперервного доступу до цільових тіньових форумів розроблена архітектура інтегрує модуль стохастичної імітації людської поведінки (Stochastic Human Behavior Simulation). Цей математичний апарат генерує синтетичні патерни рухів комп'ютерної миші, швидкості прокручування сторінок та динаміки натискання клавіш, спираючись на приховані марковські моделі (Hidden Markov Models, HMM) [6].

Додатковим вектором обфускації модулів збору є динамічна підміна криптографічних відбитків мережевого рівня, зокрема TLS-фінгерпринтів стандарту JA3 та JA4, а також обфускація Canvas/WebGL-відбитків. Система автоматизовано розраховує та генерує унікальні конфігурації еліптичних кривих та шифрувальних наборів (Cipher Suites), які на математичному рівні відповідають валідним сигнатурам найпопулярніших легальних веб-браузерів. Механізми обходу CAPTCHA реалізовані через інтеграцію з хмарними API комп'ютерного зору (Computer Vision) та алгоритмами навчання з підкріпленням (Reinforcement Learning) [6].

Протидія концептуальному дрейфу в семантичному аналізі. Фундаментальною проблемою довгострокового семантичного аналізу тіньових комунікацій є явище концептуального дрейфу (Concept Drift), що проявляється у безперервній еволюції кіберзлочинного сленгу та виникненні нових термінів обфускації. Статичні NLP-моделі неминуче стикаються з експоненціальною деградацією точності (Model Decay) при появі невідомих токенів. Для вирішення цієї проблеми програмне ядро імплементує парадигму безперервного машинного навчання (Continual Learning) та механізми виявлення семантичних аномалій.

Система регулярно застосовує тест Пейджа-Хінклі (Page-Hinkley Test) для статистичного аналізу розподілу нових текстових токенів у векторному просторі: якщо густина кластеризації невідомих термінів навколо визначених онтологічних класів перевищує встановлений математичний поріг, алгоритм автоматично ініціює процедуру донавчання. Завдяки використанню методів класифікації з нульовим навчанням (Zero-shot Learning) та механізмів уваги, система здатна самостійно

виокремлювати контекстуальне значення новостворених сленгових конструкцій виключно за їхнім синтаксичним оточенням, гарантуючи резистентність конвеєра до лінгвістичної обфускації [7].

Темпоральне топологічне моделювання та метрики центральності. Статичне моделювання, яке фіксує всі виявлені взаємозв'язки як одночасні події, є недостатнім для реверс-інжинірингу динамічних злочинних синдикатів. Розроблена методологія розширює апарат графів до темпоральних мультиграфів (Temporal Networks), де кожне імовірнісне ребро характеризується дискретним часовим інтервалом існування. До історичних кореляцій застосовуються експоненційні функції затухання довіри (Time-Decay Functions) за законом $w(t) = w_0 \cdot e^{-\lambda \Delta t}$, що автоматично знижує релевантність застарілих даних (наприклад, скомпрометованих акаунтів) та запобігає хибній кластеризації [7].

Для автоматизованого виявлення ізольованих злочинних угруповань (кіберсиндикатів) усередині глобальної мережі застосовується метод оптимізації модулярності Лувена (Louvain Method), який максимізує щільність внутрішніх зв'язків кластера. Ідентифікація ключових фігурантів здійснюється через метрики центральності: центральність за посередництвом (Betweenness Centrality) виявляє OTC-брокерів (тіньові обмінники), через які проходять транзитні потоки, тоді як центральність власного вектора (Eigenvector Centrality) ідентифікує прихованих лідерів угруповання на основі впливовості їхніх зв'язків [7].

Математична модель топологічного аналізу та кросплатформної кореляції. Акумульовані масиви сирого тексту проходять двоконтурну нормалізацію: детерміновану (через регулярні вирази RegEx з обов'язковою криптографічною валідацією контрольних сум адрес) та імовірнісну (із застосуванням моделей розпізнавання іменованих сутностей NER на базі архітектури Transformer) [8].

Здобуті верифіковані ідентифікатори передаються до графового рушія Neo4j. Якщо пряме дублювання артефактів відсутнє, система оперує імовірнісними зв'язками, розраховуючи композитний індекс довіри (Confidence Score). Кожному частковому збігу (стилометрична подібність текстів, синхронність циркадних ритмів, фонетична схожість нікнеймів за метрикою Левенштейна) призначається вага $w_i \in [0.0,1.0]$. Фінальна ймовірність злиття ізольованих профілів обчислюється за формулою об'єднання незалежних подій [8]:

$$P_{total} = 1 - \prod_{i=1}^{n} (1 - w_i)$$

Для виявлення транзитивних зв'язків застосовується апарат марковських ланцюгів та обходу зваженого мультиграфа. Якщо композитна ймовірність перевершує заданий статистичний поріг, база даних транзакційно об'єднує сутності (Entity Fusion), інстанціюючи новий об'єкт – Мета-Вузол (Meta-Node). Він виступає цифровим двійником кіберзлочинця, акумулюючи ієрархію соціальних зв'язків та історію фінансових операцій.

Експериментальне дослідження. Для перевірки алгоритмічної працездатності розробленого комплексу було проведено натурний експеримент. Програмний агент (Userbot), що функціонував усередині Docker-контейнера, перехопив у закритому Telegram-каналі публікацію тіньового прайс-листа. Система успішно заблокувала дублікати завдяки механізму кешування (collections.deque), здійснила детерміновану нормалізацію тексту та вилучила криптовалютну адресу стандарту TRC-20 (мережа Tron). Далі мікросервісний маршрутизатор виконав асинхронний запит до публічного API-шлюзу TronGrid, отримавши статуси останніх смарт-контрактів.

Фінальним кроком стало транзакційне формування графового зв'язку (Джерело → Повідомлення → Гаманець) у базі даних Neo4j та автоматична генерація дворівневого OSINT-звіту через ізольований шлюз Telegram Bot API на обліковий запис адміністратора.

Висновки

Підсумовуючи результати проведеного дослідження, можна констатувати успішне вирішення фундаментальної науково-прикладної проблеми технологічної сегментації цифрових слідів у сучасних гетерогенних кіберзлочинних екосистемах. Розроблена та емпірично верифікована алгоритмічна модель кросплатформної кореляції довела, що подолання інфраструктурної обфускації можливе виключно через синергетичне поєднання мікросервісної потокової обробки даних, методів нейромережевого семантичного аналізу та топологічного моделювання на базі зважених мультиграфів.

Імплементація безперервного конвеєра даних, який об'єднує криміналістично стійке документо-орієнтоване сховище з графовим аналітичним ядром, дозволила повністю автоматизувати процеси

вилучення, нормалізації та математичної оцінки транзитивних зв’язків між ізольованими комунікаційними й фінансовими профілями. Застосування розроблених стохастичних евристик та алгоритмів кластеризації успішно нівелює спроби зловмисників приховати рух капіталу за допомогою протоколів криптографічного мікшування чи децентралізованих фінансових сервісів.

У результаті роботи комплексу розрізнений інформаційний шум трансформується у єдиний концептуальний об’єкт – Мета-Вузол, що формує математично обґрунтовану та юридично значущу доказову базу для ідентифікації зловмисників.

Таким чином, створена архітектура спростовує концепцію абсолютної анонімності в мережі та забезпечує надійний науково-технічний фундамент для переходу від реактивної цифрової криміналістики до систем проактивного моніторингу та нейтралізації транснаціональних кіберзагроз у режимі реального часу.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Internet Organised Crime Threat Assessment (IOCTA) 2024 / European Cybercrime Centre. The Hague : Europol, 2024. 84 p. URL: https://www.europol.europa.eu/publication-events/main-reports/internet-organised-crime-threat-assessment-iocta-2024 (дата звернення: 12.05.2026).

2. Hassan N. A., Hijazi R. Open Source Intelligence Methods and Tools: A Practical Guide to Online Intelligence. Berkeley : Apress, 2018. 331 p. URL: https://content.e-bookshelf.de/media/reading/L-11293072-01e9ad742f.pdf (дата звернення: 12.05.2026).

3. The 2024 Crypto Crime Report / Chainalysis Team. New York : Chainalysis Inc., 2024. 142 p. URL: https://www.chainalysis.com/blog/2024-crypto-crime-report-introduction/ (дата звернення: 13.05.2026).

4. Narkhede N., Shapira G., Palino T. Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale. Sebastopol : O'Reilly Media, 2017. 322 p.

5. Rabiner L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 1989. Vol. 77, no. 2. P. 257–286. URL: https://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf (дата звернення: 14.05.2026).

6. Sutton R. S., Barto A. G. Reinforcement Learning: An Introduction. 2nd ed. Cambridge : MIT Press, 2018. 552 p. URL: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf (дата звернення: 14.05.2026).

7. Fast unfolding of communities in large networks / V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre. *Journal of Statistical Mechanics: Theory and Experiment*. 2008. Vol. 2008, no. 10. P. P10008. URL: https://www.researchgate.net/publication/1913681_Fast_Unfolding_of_Communities_in_Large_Networks (дата звернення: 15.05.2026).

8. Robinson I., Webber J., Eifrem E. Graph Databases. 2nd ed. Sebastopol : O'Reilly Media, 2015. 256 p. URL: https://web4.ensiie.fr/~stefania.dumbrava/OReilly_Graph_Databases.pdf (дата звернення: 15.05.2026).

Магденко Анастасія Романівна – студентка групи 1КІТС-22б, Факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, м. Вінниця, e-mail: anastasiiamahdenko@gmail.com

Грицак Анатолій Васильович – доцент кафедри менеджменту та безпеки інформаційних систем, Вінницький національний технічний університет, м. Вінниця, e-mail: grytsak.a.v@gmail.com

Mahdenko Anastasiia R. – student of group 1KITS-22b, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: anastasiiamahdenko@gmail.com

Hrytsak Anatoly V. – Associate Professor of the Department of Management and Security of Information Systems Vinnytsia National Technical University, Vinnytsia, e-mail: grytsak.a.v@gmail.com