УДК 004.7:004.056

ВИЯВЛЕННЯ АНОМАЛІЙ МЕРЕЖНОГО ТРАФІКУ В ЗАДАЧАХ ЗАХИСТУ КОРПОРАТИВНИХ МЕРЕЖ

О. Д. Демченко[1,а], В. В. Демчинський[1,б]

[1] Навчально-науковий Фізико-технічний інститут

Анотація

У роботі досліджено підхід до виявлення аномалій трафіку в корпоративних мережах на основі аналізу потокових даних NetFlow та формування статистичного профілю мережної активності (baseline). Реалізовано повний цикл обробки мережного трафіку: збір, агрегація у потоки, класифікація сервісів та формування правил нормальної активності. Методика ґрунтується на підходах Network Traffic Analysis (NTA), де для виявлення підозрілої активності використовуються метадані потоків і поведінкові ознаки мережі [1, 2]. Проведено експериментальну перевірку методу на основі зібраного набору NetFlow-даних із подальшим виділенням потоків і часових вікон, що мають ознаки відхилення від сформованого baseline як опорного статистичного профілю.

Ключові слова: корпоративна мережа, NetFlow, аномалії, статистичний профіль, baseline, мережний трафік, anomaly detection, NTA, Python.

Вступ

Сучасні корпоративні мережі характеризуються високою інтенсивністю обміну даними між користувачами, серверами та зовнішніми ресурсами. Традиційні методи захисту, засновані на сигнатурному аналізі, мають обмежену ефективність, оскільки здатні виявляти переважно відомі загрози та потребують попередньо визначених шаблонів атак [3]. З поширенням протоколів шифрування, зокрема TLS/SSL, HTTPS, QUIC, IPSec знижується ефективність глибокого аналізу пакетів (Deep Packet Inspection, DPI), оскільки вміст пакетів стає недоступним для прямої перевірки засобами моніторингу [4].

У зв'язку з цим особливого значення набуває підхід Network Traffic Analysis (NTA), який базується на аналізі метаданих мережного трафіку, зокрема потокових записів NetFlow та IPFIX. За такого підходу досліджуються не інкапсульовані дані пакетів, а характеристики взаємодії: IP-адреси, порти, протоколи, напрямки передавання, кількість пакетів, обсяг даних і часові інтервали потоків [1, 2].

Одним із ключових елементів поведінкового аналізу є формування статистичного профілю нормальної активності мережі (baseline) — опорного статистичного рівня, відносно якого оцінюються подальші відхилення. Такий профіль дозволяє виявляти аномалії як відхилення від типових шаблонів мережної взаємодії [5, 6].

Метою роботи є побудова статистичного профілю (baseline) нормальної мережної активності на основі аналізу NetFlow-потоків і виявлення аномалій як відхилень від типової поведінки мережі.

Для демонстрації методики: від збору даних до автоматичного формування правил та виділення підозрілих потоків було реалізовано комплексний підхід, що охоплює збір мережного трафіку, його перетворення у потокові записи, побудову статистичного профілю та подальше виявлення аномалій на основі аналізу відхилень.

1. Постановка задачі

У задачах захисту корпоративних мереж важливо забезпечити постійний контроль мережної активності, своєчасне виявлення нетипової поведінки вузлів і сервісів, а також оперативне реагування на потенційні інциденти безпеки. Такі задачі відповідають загальній логіці мережного моніторингу безпеки, де потоки трафіку використовуються для виявлення підозрілих шаблонів і змін у поведінці мережі [1, 5].

Для досягнення поставленої мети необхідно забезпечити збір мережного трафіку, формування NetFlow-потоків, побудову статистичного профілю та класифікацію мережних сервісів. На основі отриманих характеристик формується профіль нормальної поведінки мережі (baseline), який надалі використовується для виявлення відхилень від типових шаблонів взаємодії [2, 6].

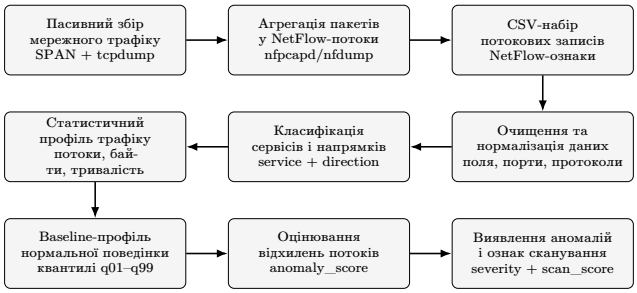
2. Методика дослідження

Методика дослідження базується на підході Network Traffic Analysis (NTA) та передбачає аналіз мережного трафіку на рівні потокових записів NetFlow з метою формування статистичного профілю мережної активності (baseline) та подальшого виявлення аномалій.

Запропонований підхід реалізує flow-based аналіз, у якому досліджуються не інкапсульовані дані пакетів, а їхні метадані: IP-адреси, порти, протоколи,

[а] olekdemch-ipt27@lll.kpi.ua
[б] vvd@bigmir.net

напрямки передавання, кількість пакетів, обсяг переданих даних і тривалість з'єднання.



Цикл дослідження: від збору трафіку до виявлення аномалій.

Збір даних виконано у локальній мережі із використанням механізму port mirroring (SPAN) і tcpdump, після чого трафік агреговано у потокові записи та експортовано у CSV. Такий підхід відповідає пасивному моніторингу, оскільки колектор отримує копію пакетів і не втручається у процес передавання даних [7]. Подальша обробка виконувалася із використанням інструментів tcpdump, nfdump та nfpcapd, призначених для роботи з пакетами та потоковими записами [8].

Для задач класифікації сервісів і напрямків потоку, побудови статистичного профілю нормальної поведінки (baseline) та розрахунку зваженого показника аномальності (`anomaly_score`) було розроблено Python-модуль.

Особливістю підходу є те, що baseline-правила нормальної активності формуються автоматично на основі фактично зібраних потоків і містять статистичні межі за кількістю пакетів, байтів і тривалістю.

2.1. Опис роботи програмного модуля

Зібраний трафік було перетворено у потоки за допомогою інструментів nfdump та nfpcapd. NetFlow є технологією збирання метаданих про IP-трафік і дає змогу аналізувати мережну активність без збереження вмісту кожного пакета [2, 8]. У межах роботи потік (flow) розглядається як сукупність пакетів з однаковими параметрами: IP-адресою джерела, IP-адресою призначення, портом джерела, портом призначення та транспортним протоколом [2].

Для аналізу мережного трафіку використовувалися такі показники, як кількість потоків, кількість пакетів, обсяг переданих даних, тривалість потоків, а також розподіл трафіку за сервісами та напрямками передавання. Такий набір ознак є типовим для flow-based аналізу, оскільки дозволяє узагальнювати велику кількість пакетів до компактних статистичних характеристик [3, 5].

Відповідно попередня обробка даних передбачала підготовку NetFlow-записів до подальшого статистичного аналізу. На вхід програма отримує CSV-файл із такими основними полями: `ts`, `duration`, `protocol`, `src_ip`, `dst_ip`, `src_port`, `dst_port`, `packets`, `bytes`. На цьому етапі виконується перевірка структури вхідного файлу, нормалізація текстових і числових полів, уніфікація значень протоколів, обробка портів та перетворення тривалості потоків у числовий формат.

Додатково формуються похідні ознаки, що характеризують інтенсивність і розмір потоків: `bytes_per_packet`, `packets_per_second`, `bytes_per_second`. Вони дозволяють оцінити середній обсяг даних на один пакет, кількість пакетів за секунду та швидкість передавання даних у межах окремого потоку.

Результатом цього етапу є очищений набір потокових записів, який містить не лише початкові NetFlow-поля, а й службові `direction`, `service`, `service_port`, `src_is_lan`, `dst_is_lan`. Ці ознаки використовуються для класифікації трафіку за напрямками та сервісами, а також для подальшого формування baseline-профілю нормальної поведінки мережі.

Напрямок потоку визначається на основі приналежності IP-адрес до локальної підмережі `192.168.177.0/24`, адреси маршрутизатора `192.168.177.1` та multicast-діапазону. Якщо джерело належить локальній мережі (LAN), а призначення є зовнішнім вузлом, потік класифікується як `LAN_TO_WAN`; у протилежному випадку — як `WAN_TO_LAN`. Взаємодія з маршрутизатором виділяється окремо як `LAN_TO_ROUTER` або `ROUTER_TO_LAN`. Такий підхід дозволяє відокремити користувацьку активність, відповіді зовнішніх сервісів, локальну взаємодію між вузлами та службовий трафік.

Класифікація сервісів виконується на основі поєднання транспортного протоколу та номера порту. Зокрема, трафік TCP/443 інтерпретується як HTTPS/TLS, UDP/443 — як QUIC/WebRTC, порт 53 — як DNS, UDP/123 — як NTP, UDP/5353 — як mDNS, TCP/445 — як WINDOWS_FILE_SHARING. Якщо комбінація протоколу та порту не відповідає жодному з визначених шаблонів, сервіс позначається як OTHER або некласифікований сервіс.

Зазначений підхід є достатнім для побудови поведінкового профілю мережі на рівні потокових характеристик та подальшого виявлення відхилень від типової активності.

2.2. Формування статистичного профілю (baseline)

Основою baseline-профілю є групування потоків за ключем:

```
protocol + direction + service_port + service
```

Тобто всі потоки з однаковим протоколом, напрямком, сервісним портом і класом сервісу об'єднуються в один шаблон нормальної поведінки. Для кожного шаблону програма обчислює кількість потоків, суму пакетів і байтів, кількість унікальних IP-адрес, середні значення, стандартні відхилення та квантильні межі для `packets`, `bytes` і `duration`.

Статистичне правило описує не лише дозволений тип взаємодії, а й типовий діапазон її параметрів. Для меж використовуються квантилі `q01` та `q99`, тобто 1-й і 99-й процентилі. Це доцільно, оскільки мережний трафік має нерівномірний розподіл: більшість потоків є короткими, але періодично виникають великі потоки. Квантилі зменшують вплив одиничних

викидів і краще описують нормальний діапазон поведінки.

Фрагмент baseline-правил статистичного профілю

```
ALLOW TCP LAN_TO_WAN service_port=443 service=HTTPS/TLS flows
    =9104; packets_q01=1; packets_q99=431.85; bytes_q01=40;
    bytes_q99=65117.65; duration_q01=0; duration_q99=300.98
```

```
ALLOW TCP LAN_TO_WAN service_port=443 service=HTTPS/TLS
```

Наведене правило означає, що HTTPS/TLS-трафік до зовнішніх ресурсів є типовим для мережі. Проте якщо конкретний HTTPS-потік має кількість пакетів, обсяг байтів або тривалість за межами статистичного діапазону, він може бути позначений як потенційна аномалія.

Кожне правило є узагальненням групи однотипних потоків, які регулярно повторювалися під час збору трафіку. Наприклад, значна кількість потоків TCP у напрямку `LAN_TO_WAN` із портом 443 узагальнюється як правило нормальної HTTPS/TLS-взаємодії. Таким чином, сформований baseline-профіль не описує кожне з'єднання окремо, а відображає типові шаблони поведінки мережі [3, 5].

Надалі ці правила використовуються як еталон: потоки, що відповідають сформованому baseline-профілю, вважаються типовими, а потоки, які не відповідають жодному правилу або суттєво відрізняються за кількістю пакетів, обсягом даних чи напрямком взаємодії, можуть розглядатися як потенційні аномалії [1, 6].

2.3. Розрахунок показника відхилення від baseline-профілю та виявлення аномалій

Після побудови baseline-профілю на навчальному наборі нормального трафіку кожен потік тестового набору порівнюється з відповідним правилом нормальної поведінки. Якщо для потоку не знайдено відповідного шаблону, фіксується ознака `UNKNOWN_BASELINE_PATTERN`. Якщо сервіс не вдалося класифікувати за визначеними правилами, додається ознака `UNCLASSIFIED_SERVICE`. У разі, коли значення `packets`, `bytes` або `duration` виходять за межі квантильного діапазону `q01--q99`, для потоку обчислюється відносна величина відхилення, яка враховується під час формування показника відхилення від baseline-профілю (`anomaly_score`).

Для узагальненої кількісної оцінки підозрілості використовується показник `anomaly_score`, який формується як зважена сума виявлених ознак відхилення від baseline-профілю. У таблиці 1 наведено вагові коефіцієнти, що використовуються під час розрахунку цього показника. Чим більше таких ознак має потік і чим суттєвішими є статистичні відхилення, тим вищим є значення цього показника.

Таблиця 1. Вагові коефіцієнти ознак відхилення для розрахунку `anomaly_score`.

Ознака відхилення	Ваговий коефіцієнт
Невідомий baseline-шаблон	+5,0
Некласифікований сервіс	+2,0
Відхилення за кількістю пакетів	+1,0
Відхилення за обсягом байтів	+1,5
Відхилення за тривалістю	+1,0

Після цього потоку призначається рівень критичності: NORMAL для 0 балів, LOW для ненульового значення, MEDIUM від 4 балів і HIGH від 7 балів. Запропонований підхід є гнучкішим за бінарну логіку «норма/аномалія», оскільки дозволяє ранжувати події за рівнем підозрілості.

Приклад інтерпретації аномального потоку

```
TCP LAN_TO_WAN HTTPS/TLS
packets=3742
bytes=841003
duration=264.231 sec
anomaly_score=7.5
severity=HIGH
reason=PACKETS_OUT_OF_BASELINE;BYTES_OUT_OF_BASELINE
```

Такий потік формально належить до дозволеного HTTPS/TLS-трафіку, однак його кількість пакетів і обсяг даних суттєво перевищують статистичні межі baseline-профілю. Саме тому він позначається як аномальний не через порт або протокол, а через нетипову поведінку всередині дозволеного шаблону.

Виявлення аномалій базується на принципі порівняння нових NetFlow-потоків тестового періоду з побудованим baseline-профілем нормальної поведінки: якщо параметри потоку не відповідають визначеним правилам, такий потік розглядається як потенційна аномалія.

2.4. Поведінковий аналіз хостів і ознаки сканування

Окремо для нових потоків, що перевіряються відносно сформованого baseline-профілю, виконується поведінковий аналіз локальних вузлів у часових вікнах тривалістю 1 хвилина. Для кожного джерела трафіку обчислюються кількість потоків, кількість унікальних IP-адрес призначення, кількість зовнішніх адрес, унікальних портів, сервісів, а також кількість коротких потоків тривалістю до 1 секунди.

Такий підхід дозволяє виявляти аномалії, які не завжди помітні на рівні одного окремого потоку. Зокрема, сканування портів або мережна розвідка зазвичай проявляються як сукупність великої кількості коротких з'єднань до різних портів або IP-адрес у межах короткого проміжку часу.

Для кількісної оцінки такої активності використовується показник підозрілої поведінки вузла у часі (`scan_score`), який формується на основі сукупності поведінкових ознак локального вузла в межах хвилинного часового вікна. До таких ознак належать підвищена кількість потоків, значна кількість унікальних IP-адрес призначення, велика кількість унікальних портів, різноманітність сервісів, а також переважання короткотривалих з'єднань. Чим більше таких ознак спостерігається одночасно, тим вищим є рівень підозрілої активності вузла.

Якщо значення `scan_score` перевищує встановлений поріг, відповідне часове вікно позначається як підозріле: наприклад, велика кількість короткотривалих потоків до різних IP-адрес і портів протягом однієї хвилини може свідчити про мережну розвідку або іншу нетипову активність.

3. Результати роботи

Найбільшу частку за кількістю потоків становили DNS — 22 472 потоки, HTTPS/TLS — 14 036, ICMP — 7 782, QUIC/WebRTC — 6 905 та NTP — 1 207.

За напрямками переважав `LAN_TO_WAN` — 22 421 потік, далі `LAN_TO_ROUTER` — 14 330, `WAN_TO_LAN` — 10 374 та `ROUTER_TO_LAN` — 7 510. Це підтверджує, що основна активність мережі пов'язана з вихідними зверненнями локальних пристроїв до зовнішніх ресурсів, DNS-обміном і відповідями зовнішніх серверів.

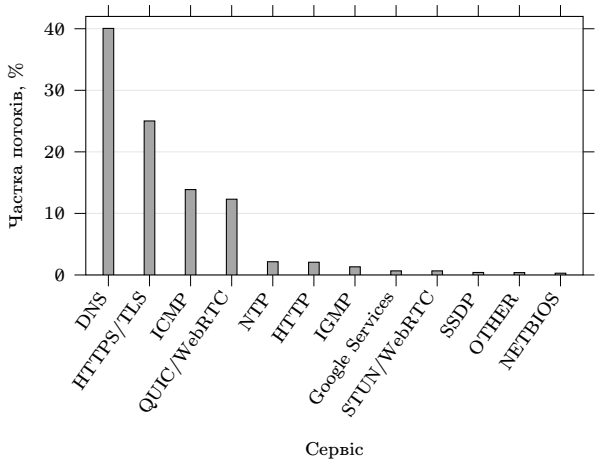


Рис. 1. Розподіл потоків за сервісами.

На рис. 1 наведено розподіл потоків за сервісами. Отримані результати свідчать про те, що мережний трафік у досліджуваній підмережі має достатньо повторювану структуру. Основна частина потоків формується типовими сервісами та напрямками взаємодії, зокрема DNS, HTTPS/TLS, QUIC/WebRTC, HTTP, NTP та службовими протоколами локальної мережі. Це дозволяє узагальнити велику кількість окремих потоків у компактний набір правил нормальної поведінки [1, 2].

Таблиця 2. Основні характеристики навчального набору та сформованого baseline-профілю.

Показник	Значення
Кількість потоків	56 102
Кількість пакетів	6 191 045
Обсяг трафіку	424,359 MB
Тривалість збору	147,218 хв
Середній розмір потоку	7 931,486 байт
Медіанний розмір потоку	96 байт
Кількість baseline-правил	40

У таблиці 2 подано основні характеристики навчального набору, на основі якого було сформовано baseline-профіль нормальної поведінки мережі. Сформований профіль складається з обмеженої кількості правил, однак кожне з них описує не одиничне з'єднання, а типовий шаблон мережної взаємодії. Наприклад, багато однотипних HTTPS-з'єднань у напрямку `LAN_TO_WAN` узагальнюються одним правилом, що описує нормальний веб-трафік користувачів. Такий підхід дозволяє зменшити обсяг даних для аналізу та виділити найбільш характерні ознаки нормального функціонування мережі.

Відхилення від сформованого baseline можуть розглядатися як ознаки потенційно аномальної активності. До таких відхилень належать поява нетипових портів, невідомих сервісів, різке збільшення кількості потоків, зростання кількості унікальних адрес або зміна звичного розподілу трафіку. Саме тому статистичний профіль може використовуватись як основа для подальшого виявлення сканування, інтенсивних запитів або інших нетипових дій у мережі [9, 6].

Висновки

У роботі розглянуто методику виявлення аномалій мережного трафіку у корпоративних мережах на основі аналізу NetFlow-даних. Реалізовано збір і аналіз потокової інформації, класифікацію сервісів та формування статистичного профілю (baseline), який описує типові шаблони мережної взаємодії.

Подальший розвиток методики може передбачати побудову статистичного профілю з урахуванням часу доби, дня тижня та сезонних змін мережної активності, а також порівняння отриманих результатів з методами машинного навчання.

Отримані результати підтверджують доцільність використання NetFlow-аналізу для задач моніторингу корпоративних мереж, оскільки він дає змогу виявляти відхилення за метаданими потоків без аналізу вмісту пакетів [1, 2]. Такий підхід є особливо актуальним в умовах широкого використання зашифрованого трафіку та може бути основою для побудови систем виявлення аномальної активності [4].

Перелік використаних джерел

1. *Cisco.* What Is Network Traffic Analysis? — URL: https://www.cisco.com/site/us/en/learn/topics/security/what-is-network-traffic-analysis.html.
2. *IBM.* What is NetFlow? — URL: https://www.ibm.com/think/topics/netflow.
3. *Collins M.* Network Security Through Data Analysis: Building Situational Awareness. — Sebastopol : O'Reilly Media, 2017.
4. *ENISA.* Encrypted Traffic Analysis / European Union Agency for Cybersecurity. — 2020. — URL: https://www.enisa.europa.eu/publications/encrypted-traffic-analysis.
5. *Davidoff S., Ham J.* Network Forensics: Tracking Hackers Through Cyberspace. — Upper Saddle River : Prentice Hall, 2012.
6. Anomaly Detection / Wikipedia. — URL: https://en.wikipedia.org/wiki/Anomaly_detection.
7. *TP-Link.* How to configure Port Mirroring on TP-Link Easy Smart/Smart Switch. — URL: https://www.tp-link.com/support/faq/527/.
8. *nfdump project.* Netflow processing tools. — URL: https://github.com/phaag/nfdump.
9. *Nmap Project.* Nmap Reference Guide. — URL: https://nmap.org/book/man.html.