

Іван ПАРХОМЕНКО, канд. техн. наук, доц.
ORCID ID: 0000-0001-6889-9284
e-mail: ivan.parkhomenko@knu.ua

Київський національний університет імені Тараса Шевченка, Київ, Україна

Юрій СОКИРАН, асп.
ORCID ID: 0009-0004-7041-2307
e-mail: sokyran@knu.ua

Київський національний університет імені Тараса Шевченка, Київ, Україна

МЕТОДИ ОЦІНЮВАННЯ ГЕОГРАФІЧНОЇ КОНЦЕНТРАЦІЇ ТА СТАБІЛЬНОСТІ РОЗРОБНИКІВ У ПРОЄКТАХ ІЗ ВІДКРИТИМ КОДОМ З OSINT-ПІДХОДОМ

Вступ. Представлено комплексну OSINT-орієнтовану систему для оцінювання ризиків ланцюгів постачання програмного забезпечення у критичній інфраструктурі з відкритим кодом. Після сплеску атак на ланцюги постачання, систематичне оцінювання вразливостей проєктів з відкритим кодом стало необхідністю. Запропоновано багатовимірну методологію оцінювання ризиків, яка інтегрує аналіз географічної концентрації, метрики стабільності розробників (bus factor), методи атрибуції учасників. Система поєднує методи Mining Software Repositories з OSINT-підходами, включаючи аналіз доменів електронної пошти, комітів і публічних профілів.

Методи. Досліджено модель оцінювання ризиків ланцюгів постачання програмного забезпечення з відкритим кодом. Модель географічної концентрації базується на модифікованому індексі Герфіндала – Гіршмана (HHI) для кількісного оцінювання розподілу внесків розробників за країнами. Модель стійкості супроводжувачів використовує метрику bus factor для визначення мінімальної кількості ключових учасників, втрата яких критично вплине на проєкт. Модель атрибуції учасників поєднує OSINT-методи аналізу доменів електронної пошти, комітів і публічних профілів для географічної ідентифікації розробників.

Результати. Емпірична валідація на п'яти критичних проєктах з відкритим кодом на основі аналізу 1500 комітів кожного проєкту виявила значні варіації у профілях ризиків. Індекс HHI коливається від 1000+ (Django – оптимальна диверсифікація) до 7000+ (Swiper – критична концентрація). Ідентифіковано проєкти з екстремальною географічною залежністю: Tarantool – 81 % внесків із росії, що створює ризики. Встановлено, що географічна концентрація та bus factor є незалежними вимірами ризику, що підтверджує необхідність багатовимірного оцінювання.

Висновки. Запропонована багатовимірна OSINT-методологія оцінювання ризиків ланцюгів постачання програмного забезпечення з відкритим кодом продемонструвала ефективність у виявленні вразливостей. Емпірична валідація на п'яти проєктах підтвердила, що географічна концентрація та bus factor є незалежними вимірами ризику, що вимагають окремого оцінювання. Ідентифіковано проєкти з критичними профілями: Swiper (bus factor 1, ризик 100 %) повторює передінцидентний патерн XZ Utils, а Tarantool (81 % внесків із росії) створює санкційні ризики. Використання метрик забезпечує комплексне оцінювання вразливостей, що дозволяє приймати обґрунтовані рішення щодо вибору залежностей у критичній інфраструктурі.

Ключові слова: OSINT, програмне забезпечення з відкритим кодом, географічна концентрація, критична інфраструктура, оцінювання ризиків.

Вступ

Згідно з дослідженням Linux Foundation (2020), відкрите програмне забезпечення нині відіграє ключову роль у глобальній цифровій екосистемі, забезпечуючи потреби мільярдів користувачів. Більше того, динаміка зростання залежності від таких систем, як ядро Linux, OpenSSL, curl, лише посилюється, створюючи нові виклики для національної безпеки та стабільності глобального цифрового простору.

Дослідження Freund (2024) виявило критичний інцидент із бекдором XZ Utils. Вразливість ланцюгів постачання програмного забезпечення з відкритим кодом була яскраво продемонстрована інцидентом у березні 2024 р., коли витончена багаторічна кампанія соціальної інженерії призвела до впровадження шкідливого коду в бібліотеку стиснення, яка використовується практично всіма дистрибутивами Linux. Цей інцидент виявив фундаментальні слабкості в тому, як критична інфраструктура залежить від проєктів, які супроводжуються окремими особами або невеликими командами, підкресливши, що безпека ланцюгів постачання охоплює не лише цілісність коду, а й стабільність розробників, верифікацію учасників та організаційну стійкість.

Однією з проблем сучасного етапу розвитку програмного забезпечення з відкритим кодом є багатогранний характер ризиків ланцюгів постачання. Вони включають географічну концентрацію основних розробників, залежності від одного розробника, що створюють критичні точки відмови, корпоративний контроль над важливими бібліотеками та недостатні механізми верифікації учасників. За даними GitHub Octoverse (2024), США продовжують лідирувати за кількістю розробників, тоді як Індія швидко наближається до першого місця з прогнозом досягнення лідерства до 2028 р. Така концентрація у поєднанні з проблемою "bus factor", коли проєкти залежать від дуже небагатьох активних супроводжувачів, формує багаторівневі загрози для екосистеми постачання програмних продуктів.

Враховуючи це, актуальним є створення багатовимірної моделі аналізу ризиків у процесі постачання програмного забезпечення, що враховуватиме територіальну локалізацію, надійність команд підтримки, залежність від корпорацій і верифікацію учасників. Така система дозволить виявляти потенційні вразливості ланцюгів постачання до їхньої експлуатації, надавати дієві метрики для пріоритизації ризиків і підтримувати прийняття обґрунтованих рішень щодо вибору залежностей і стратегій пом'якшення ризиків.

Метою цієї статті є розроблення комплексної OSINT-орієнтованої системи оцінювання ризиків ланцюгів постачання програмного забезпечення в критичній інфраструктурі з відкритим кодом. Ця система охоплює множинні



виміри ризиків через систематичний аналіз географічного розподілу та верифікації учасників, забезпечуючи виявлення потенційних вразливостей ланцюгів постачання та формування практичних рекомендацій для підвищення стійкості програмних залежностей.

Огляд літератури. У сфері геополітичної безпеки програмного забезпечення з відкритим кодом можна виділити кілька ключових напрямів досліджень, що розглядають різні аспекти впливу географічних і політичних факторів на розвиток критично важливих технологій.

Дослідження Atlantic Council (2024) наголошує, що код із відкритим кодом став критично важливою інфраструктурою, яка потребує державного регулювання та захисту. Статистика демонструє, що 34 % внесків в екосистему відкритого коду генерується американськими розробниками, 13 % – російськими, формуючи регіональну концентрацію загроз для ключових програмних проєктів.

Експерти французького Ifri (2024) досліджують економічний і геополітичний вплив залежності від open-source рішень, відмічаючи, що урядова залученість до розроблення відкритого коду набуває дедалі виразнішого політичного характеру й підпорядковується цілям держав у сфері кібербезпеки.

Дослідження Hunted Labs (2024) також зазначає, що російські розробники з VK (компанії під санкціями за зв'язки з російськими спецслужбами) контролюють пакет `easyjson`, який використовується в критичних компонентах, таких як Kubernetes, Helm та Istio.

Концепція "bus factor" – мінімальна кількість членів команди, відхід яких критично вплине на проєкт – стала важливою метрикою для оцінювання стійкості проєктів із відкритим кодом. Дослідження Avelino et al. (2016). (2016) демонструє, що значна частина проєктів на GitHub має bus factor один або два, що означає, що втрата одного учасника може повністю зупинити розроблення. Ця вразливість була проілюстрована інцидентом XZ Utils, де виграшного єдиного розробника було використано через соціальну інженерію для впровадження бекдору. Cosentino et al. (2015) розробили методи обчислення truck factor (синонім bus factor) у програмних проєктах, надавши алгоритмічні підходи, які можна інтегрувати в автоматизовані системи оцінювання ризиків. Ці проблеми стійкості супроводжувачів безпосередньо перетинаються з ризиками географічної та корпоративної концентрації, оскільки проєкти з низьким bus factor, сконцентровані в певних регіонах або компаніях, стикаються з комплексними вразливостями.

Аналіз від prn (2018) останніх патернів атак на ланцюги постачання виявляє витончені стратегії експлуатації, націлені на екосистеми відкритого коду. Інцидент event-stream 2018 р. продемонстрував, як злоумисники можуть отримати доступ через соціальну інженерію та активність у внесках. Ці вектори атак підкреслюють потребу в комплексних системах оцінювання ризиків ланцюгів постачання, які оцінюють не лише якість коду, а й управління проєктом, верифікацію учасників та аналіз дерева залежностей.

Згідно з дослідженням Vidoni (2022), для систематичного аналізу цих ризиків ланцюгів постачання дослідники розробили методології Mining Software Repositories (MSR), які витягують та аналізують дані з артефактів розроблення програмного забезпечення. Комплексний систематичний огляд літератури з досліджень MSR демонструє їх застосування в багатьох сферах, включаючи характеристики мереж розробників, аналіз географічного розподілу, відстеження вразливостей.

Методи

Методологічну основу роботи становить інтегрований підхід до збирання й опрацювання інформації з репозиторіїв ключових open-source проєктів, що об'єднує інструментарій Mining Software Repositories із методами геолокації контриб'юторів та аналізом залежностей у постачанні.

З метою досягнення репрезентативності вибірки відібрано п'ять значущих проєктів із відкритим вихідним кодом: Django (вебфреймворк), Tarantool (база даних), Swiper (UI-бібліотека), Node.js (середовище виконання) та Model Context Protocol (AI-фреймворк). Ці проєкти характеризуються рівнем використання в критичній інфраструктурі, значними залежностями від інших систем та активними спільнотами розробників (CISA, 2024).

Збір метаданих здійснюється за допомогою GitHub REST API та безпосереднього аналізу git-логів із клонованих репозиторіїв. GitHub API надає доступ до статистики внесків, інформації про розробників і структур залежностей.

Географічна атрибуція розробників здійснюється за допомогою трьох основних методів. Аналіз доменів електронної пошти дозволяє ідентифікувати країну та корпоративну приналежність через зіставлення доменів із базами даних геолокації DNS-записів. Аналіз часових зон комітів забезпечує приблизне визначення географічних регіонів через параметр `--date=raw` команди `git log`, який надає інформацію про зсув відносно UTC (Opensource.com, 2015). Аналіз публічних профілів GitHub доповнює географічні дані через кінцеву точку API `/users/{username}`, яка може містити інформацію про місцезнаходження розробника.

Результати

Для оцінювання стійкості проєкту до атак, спрямованих на розробників, впроваджено обчислення bus factor на основі концентрації авторства комітів. Bus factor представляє мінімальну кількість учасників, відхід яких призведе до втрати критичних знань про проєкт, і обчислюється через аналіз розподілу авторства коду по кодовій базі.

Для кожного проєкту ми аналізуємо історію комітів для ідентифікації мінімального набору учасників, відповідальних за 50 % загальної кількості комітів, відповідно до методології Avelino et al., 2016. Ця метрика визначається кількістю учасників, що покривають 50 % комітів. Проєкти з bus factor ≤ 2 класифікуються як критично вразливі до атак соціальної інженерії, спрямованих на супроводжувачів, як це було продемонстровано інцидентом XZ Utils, де єдиний перевантажений розробник був експлуатований.

Компонент ризику bus factor кількісно оцінює вразливість до атак, спрямованих на супроводжувачів:

$$R_{\text{busfactor}} = \max(0, (4 - \text{BF}) / 3) \times 100 \%, \quad (1)$$

де BF – обчислений bus factor. Проєкти з $\text{BF} \leq 3$ отримують оцінки ризику 33–100 %, що відображає критичну вразливість до сценаріїв єдиної точки відмови.



Додатково обчислюємо коефіцієнт концентрації супроводжувачів:

$$R_{\text{maintainer}} = \frac{C_{\text{top3}}}{C_{\text{total}}} \times 100 \%, \quad (2)$$

де C_{top3} – сума комітів від трьох найактивніших контриб'юторів (відсортованих за кількістю комітів), а C_{total} – загальна кількість комітів (без урахування ботів). Значення, що перевищують 80 %, вказують на високий ризик концентрації супроводжувачів.

Індекс Герфіндала – Гіршмана для географічної концентрації внесків обчислюють шляхом групування розробників за країнами та розрахунку відсоткового розподілу їхніх внесків:

$$HHI = \sum_{i=1}^n \left(\frac{P_i}{100} \right)^2 \times 10000, \quad (3)$$

де P_i – відсотковий внесок i -ї країни.

Ризик концентрації обчислюють за формулою

$$R_{\text{concentration}} = \begin{cases} \min(100, (G_{\text{top3}} - 50) \times 2, & G_{\text{top3}} > 50 \\ 0, & G_{\text{top3}} \leq 50 \end{cases}, \quad (4)$$

де G_{top3} – сума відсоткових внесків від трьох країн із найбільшими внесками. 50 % для трьох провідних країн визначено як критичний поріг.

Процес оцінювання корпоративного домінування розрізняє корпоративних та індивідуальних розробників, обчислюючи їхні відносні частки в загальному обсязі внесків:

$$S_{\text{corp}} = \frac{C_{\text{corp}}}{C_{\text{total}}} \times 100, \quad (5)$$

$$R_{\text{corp}} = \begin{cases} \min(100, (S_{\text{corp}} - 40) \times 2.5, & S_{\text{corp}} > 50 \\ 0, & S_{\text{corp}} \leq 50 \end{cases}, \quad (6)$$

де S_{corp} – корпоративні внески, C_{total} – загальні внески. Пороговий рівень корпоративного домінування встановлено на рівні 50 %.

Система також оцінює якість даних шляхом аналізу рівнів визначення походження розробника:

$$R_{\text{uncertainty}} = \min(100, U_{\text{unknown}} \times 2 + U_{\text{low}} \times 1.5), \quad (7)$$

де U_{unknown} – відсоток розробників із невизначеним місцезнаходженням і характеристиками часового поясу, U_{low} – відсоток розробників із низькою впевненістю в результаті. Впевненість залежить від методу вилучення інформації: прямі збіги назв країн і корпоративних доменів – висока впевненість; часткові збіги за місцезнаходженням і національними доменами верхнього рівня (TLD) – середня; типові припущення для загальних TLD – низька, а якщо визначення не вдалося – впевненість невідома.

Оцінювання диверсифікації зосереджено на аналізі 10 % найактивніших розробників:

$$R_{\text{diversity}} = \max \left(0, \left(1 - \frac{T_{\text{top}}}{T_{\text{total}}} \right) \times 100 \right), \quad (8)$$

де T_{top} – це країни, що входять до 10 % найактивніших розробників, T_{total} – загальна кількість країн.

Загальна оцінка ризику формується як зважена комбінація компонентів:

$$R_{\text{overall}} = \sum_{i=1}^n R_i \times w_i. \quad (9)$$

Ваги відкалібровано для оцінювання ризиків ланцюгів постачання: географічна концентрація (0,30), корпоративне домінування (0,20), невизначеність (0,20), диверсифікація (0,15), ризик bus factor (0,15).

Застосування запропонованої методології до п'яти проєктів із відкритим кодом виявило значні варіації у профілях ризиків ланцюгів постачання. Валідація охопила Django (вебфреймворк), Tarantool (база даних), Swiper (UI-бібліотека), Node.js (середовище виконання) та Model Context Protocol (AI-фреймворк), проаналізувавши по 1500 комітів кожного проєкту та загалом 560 учасників. Індекс Герфіндала – Гіршмана варіювався від 1097 (Django) до 7734 (Swiper), демонструючи суттєві варіації у ризиках концентрації в екосистемі (табл. 1).

Таблиця 1

Метрики ризику географічної концентрації

Проєкт	Учасники	Атрибут, %	HHI	$R_{\text{концентрації}}$, %	Провідні країни, %
Django	200	56,5	1097	0	AU (19,8), NL (15,5), US (10,9)
Tarantool	91	35,2	6769	84.6	RU (81,8) , CA (7,3), UK (3,1)
Swiper	17	58,8	7734	96.8	US (87,3), IL (10,6), BR (0,5)
Node.js	200	64,0	1793	19.6	US (37,6), NL (11,4), DE (10,9)
MCP	52	48,1	3887	82.6	UK (47,9), US (39,7), NL (3,8)

Tarantool виявився проєктом із найвищим геополітичним ризиком – 81,8 % внесків із росії. Така екстремальна концентрація в одній країні для системи баз даних створює критичні вразливості ланцюгів постачання з огляду на чинні санкції та геополітичну напруженість. Крім цього, проєкт Swiper також має велику концентрацію розробників у США, але це пов'язано з низькою кількістю розробників. Це натомість спричинятиме підвищення bus factor ризику.

Django продемонстрував оптимальний профіль ризику зі збалансованим географічним розподілом по 29 країнах, доводячи, що диверсифікація в критичних проєктах є досяжною. Жодна країна не перевищила 20 % внесків, а лідери (Австралія, Нідерланди, США) сумарно становлять лише 46,1 % (табл. 2).

Node.js є еталоном здорового проєкту: bus factor 10 (ризик 0 %), ризик концентрації 23,1 %, стабільність 100 % (усі 25 історичних супроводжувачів активні). Django також показує здорові метрики: bus factor 7, стабільність 72,2 %, 13 активних із 18 історичних супроводжувачів.

Django також показує здорові метрики з bus factor 7, стабільністю 72,2 % та 13 активними супроводжувачами з 18 історичних.



Tarantool та MCP демонструють задовільні показники стійкості: bus factor 5 та 4 відповідно, стабільність 73,9 % та 100 %. Примітно, що Tarantool, незважаючи на критичний географічний ризик (81 % росія), має здорові показники супроводжувачів, що підтверджує незалежність цих вимірів ризику.

Swiper демонструє найкритичніший ризик стійкості супроводжувачів із bus factor 1, що означає фактичну залежність проекту від одного розробника (nolimits4web). Цей патерн повторює стан XZ Utils до інциденту та є привабливою ціллю для атак соціальної інженерії. Ризик концентрації 88,7 % вказує на критичну вразливість проекту.

Таблиця 2

Метрики стабільності розробників

Проект	Bus factor	R _{maintainer} , %	R _{busfactor} , %	Активні розробники	Історично активні розробники	Рівень
Django	7	29,3	0,0	13	18	Низький
Tarantool	5	32,1	0,0	17	23	Низький
Swiper	1	88,7	100,0	1	3	Критичний
Node.js	10	23,1	0,0	25	25	Низький
MCP	4	42,2	0,0	17	17	Низький

Дискусія і висновки

У цій роботі представлено комплексну OSINT-орієнтовану систему оцінювання ризиків ланцюгів постачання програмного забезпечення в критичній інфраструктурі з відкритим кодом. Система охоплює множинні виміри вразливостей, включаючи географічну концентрацію, стійкість супроводжувачів (bus factor) і достовірність верифікації учасників.

Емпірична валідація на п'яти проектах демонструє значні варіації у профілях ризиків ланцюгів постачання. Критичні висновки включають: ідентифікацію проекту з одним супроводжувачем (Swiper, bus factor = 1, що повторює передінцидентний патерн XZ Utils; екстремальну географічну концентрацію (Tarantool: 81 % росія), що створює санкційні ризики; та важливість багатовимірного оцінювання – Tarantool має критичний географічний ризик, але здорові показники супроводжувачів (bus factor 5, стабільність 73,9 %), тоді як MCP демонструє протилежний патерн із високою концентрацією на союзних країнах (UK–US 87,5 %) та відмінною стабільністю команди (100 %).

Система надає дієві метрики, що дозволяють організаціям приймати обґрунтовані рішення щодо вибору залежностей, пріоритизувати аудити безпеки для високоризикових залежностей і розробляти цілеспрямовані стратегії пом'якшення ризиків. Оскільки атаки на ланцюги постачання продовжують ескалацію, таке систематичне оцінювання ризиків стає необхідним для підтримки безпечних програмних екосистем.

Перспективні напрями подальших досліджень включають розроблення автоматизованих систем моніторингу ризиків ланцюгів постачання в реальному часі, створення міжнародних стандартів оцінювання та використання поведінкової аналітики для раннього виявлення атак соціальної інженерії, спрямованих на супроводжувачів проектів.

Практичне застосування моделей у реальних мережах: запропоновані моделі мають практичне застосування у кількох ключових напрямках безпеки екосистеми постачання програмних продуктів. Під час вибору нових залежностей розробники можуть порівнювати альтернативні бібліотеки за профілями ризику, обираючи компоненти зі здоровішими показниками стійкості команди. Для існуючих систем методологія дозволяє автоматизовано проаналізувати всі залежності та пріоритизувати аудит безпеки, починаючи з найризикованіших компонентів.

Можливості для подальших досліджень такі: подальші дослідження можуть бути спрямовані на розширення можливостей моделей шляхом інтеграції додаткових автоматизованих алгоритмів зіставлення, які співвідносять імена користувачів GitHub, справжні імена та шаблони електронної пошти з профілями LinkedIn, збір перевіреної історії зайнятості та національної приналежності з професійних профілів.

Результати дослідження показують, що запропонована багатовимірна OSINT-методологія оцінювання ризиків екосистеми постачання програмного забезпечення з відкритим кодом продемонструвала високу ефективність у виявленні вразливостей. Використання метрик забезпечує комплексне оцінювання вразливостей, що не досягається традиційними одновимірними підходами, та дає змогу приймати обґрунтовані рішення щодо вибору залежностей у критичній інфраструктурі.

Внесок авторів: Іван Пархоменко – концептуалізація; методологія; аналіз джерел, підготування огляду літератури або теоретичних засад дослідження; Юрій Сокиран – збір емпіричних даних та їхня валідація; емпіричне дослідження.

Джерела фінансування. Це дослідження не отримало жодного гранту від фінансової установи в державному, комерційному або некомерційному секторах.

Список використаних джерел

- Atlantic Council. (2024). *Avoiding the success trap: Toward policy for open-source software as infrastructure*. <https://www.atlanticcouncil.org/in-depth-research-reports/report/open-source-software-as-infrastructure/>
- Avelino, G., Passos, L., Hora, A., & Valente, M. T. (2016). A novel approach for estimating truck factors. In *Proceedings of the 24th IEEE International Conference on Program Comprehension (ICPC)* (pp. 1–10). IEEE. <https://doi.org/10.1109/ICPC.2016.7503718>
- CISA. (2024). *CISA open-source software security roadmap*. <https://www.cisa.gov/resources-tools/resources/cisa-open-source-software-security-roadmap>
- Cosentino, V., Cánovas Izquierdo, J. L., & Cabot, J. (2015). Assessing the bus factor of Git repositories. In *Proceedings of the 22nd IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER)* (pp. 499–503). IEEE. <https://doi.org/10.1109/SANER.2015.7081864>
- Freund, A. (2024, March 29). Backdoor in upstream xz/liblzma leading to SSH server compromise [Mailing list post]. oss-security. <https://www.openwall.com/lists/oss-security/2024/03/29/4>
- GitHub. (2024). *Octoverse 2024: The state of open source*. <https://octoverse.github.com/>
- Hunted Labs. (2024). *The Russian open-source project we can't live without*. <https://huntedlabs.com/the-russian-open-source-project-that-we-cant-live-without/>
- Ifri. (2024). *Software power: The economic and geo-political implications of open-source software*. <https://www.ifri.org/en/publications/etudes-de-ifri/soft-power-economic-and-geopolitical-implications-open-source>
- Linux Foundation. (2020). *2020 FOSS contributor survey*. <https://www.linuxfoundation.org/resources/publications/foss-contributor-2020>



- npm, Inc. (2018, November 27). *Details about the event-stream incident*. *npm Blog*. <https://blog.npmjs.org/post/180565383195/details-about-the-event-stream-incident>
- Opensource.com. (2015, November 10). *Using Git and mailing lists time zones to find out where developers live*. <https://opensource.com/business/15/11/git-mailing-list-time-zone-analysis>
- Vidoni, M. (2022). A systematic process for mining software repositories: Results from a systematic literature review. *Information and Software Technology*, 144, Article 106791. <https://doi.org/10.1016/j.infsof.2021.106791>

References

- Atlantic Council. (2024). *Avoiding the success trap: Toward policy for open-source software as infrastructure*. <https://www.atlanticcouncil.org/in-depth-research-reports/report/open-source-software-as-infrastructure/>
- Avelino, G., Passos, L., Hora, A., & Valente, M. T. (2016). A novel approach for estimating truck factors. In *Proceedings of the 24th IEEE International Conference on Program Comprehension (ICPC)* (pp. 1–10). IEEE. <https://doi.org/10.1109/ICPC.2016.7503718>
- CISA. (2024). *CISA open-source software security roadmap*. <https://www.cisa.gov/resources-tools/resources/cisa-open-source-software-security-roadmap>
- Cosentino, V., Cánovas Izquierdo, J. L., & Cabot, J. (2015). Assessing the bus factor of Git repositories. In *Proceedings of the 22nd IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER)* (pp. 499–503). IEEE. <https://doi.org/10.1109/SANER.2015.7081864>
- Freund, A. (2024, March 29). Backdoor in upstream xz/liblzma leading to SSH server compromise [Mailing list post]. *oss-security*. <https://www.openwall.com/lists/oss-security/2024/03/29/4>
- GitHub. (2024). *Octoverse 2024: The state of open source*. <https://octoverse.github.com/>
- Hunted Labs. (2024). *The Russian open-source project we can't live without*. <https://huntedlabs.com/the-russian-open-source-project-that-we-cant-live-without/>
- Ifri. (2024). *Software power: The economic and geo-political implications of open-source software*. <https://www.ifri.org/en/publications/etudes-de-ifri/software-power-economic-and-geopolitical-implications-open-source>
- Linux Foundation. (2020). *2020 FOSS contributor survey*. <https://www.linuxfoundation.org/resources/publications/foss-contributor-2020>
- npm, Inc. (2018, November 27). *Details about the event-stream incident*. *npm Blog*. <https://blog.npmjs.org/post/180565383195/details-about-the-event-stream-incident>
- Opensource.com. (2015, November 10). *Using Git and mailing lists time zones to find out where developers live*. <https://opensource.com/business/15/11/git-mailing-list-time-zone-analysis>
- Vidoni, M. (2022). A systematic process for mining software repositories: Results from a systematic literature review. *Information and Software Technology*, 144, Article 106791. <https://doi.org/10.1016/j.infsof.2021.106791>

Отримано редакцією журналу / Received: 09.12.25
Прорецензовано / Revised: 20.12.25
Схвалено до друку / Accepted: 25.12.25

Ivan PARKHOMENKO, PhD, Assoc. Prof.
ORCID ID: 0000-0001-6889-9284
e-mail: ivan.parkhomenko@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Yurii SOKYRAN, PhD Student
ORCID ID: 0009-0004-7041-2307
e-mail: sokyrans@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

METHODS FOR ASSESSING GEOGRAPHIC CONCENTRATION AND DEVELOPER STABILITY IN OPEN-SOURCE PROJECTS USING AN OSINT APPROACH

Background. This paper presents a comprehensive OSINT-oriented system for assessing software supply chain risks in open-source critical infrastructure. Following the surge in supply chain attacks, systematic vulnerability assessment of open-source projects has become essential. A multidimensional risk assessment methodology is proposed, integrating geographic concentration analysis, developer stability metrics (bus factor), and contributor attribution methods. The system combines Mining Software Repositories methods with OSINT approaches, including analysis of email domains, commits, and public profiles.

Methods. The paper examines a risk assessment model for open-source software supply chains. The geographic concentration model is based on a modified Herfindahl-Hirschman Index (HHI) for quantifying the distribution of developer contributions by country. The maintainer resilience model employs the bus factor metric to determine the minimum number of key contributors whose loss would critically impact the project. The contributor attribution model combines OSINT methods for analyzing email domains, commits, and public profiles for geographic identification of developers.

Results. Empirical validation on five critical open-source projects based on the analysis of 1,500 commits per project revealed significant variations in risk profiles. The HHI ranged from 1,000+ (Django — optimal diversification) to 7,000+ (Swiper — critical concentration). Projects with extreme geographic dependency were identified: Tarantool — 81% of contributions from Russia, creating risks. It was established that geographic concentration and bus factor are independent risk dimensions, confirming the necessity of multidimensional assessment.

Conclusions. The proposed multidimensional OSINT methodology for assessing open-source software supply chain risks demonstrated effectiveness in identifying vulnerabilities. Empirical validation on five projects confirmed that geographic concentration and bus factor are independent risk dimensions requiring separate assessment. Projects with critical profiles were identified: Swiper (bus factor 1, risk 100%) replicates the pre-incident pattern of XZ Utils, while Tarantool (81% of contributions from Russia) creates sanction-related risks. The use of these metrics enables comprehensive vulnerability assessment, allowing informed decisions regarding dependency selection in critical infrastructure.

Keywords: OSINT, open-source software, geographic concentration, critical infrastructure, risk assessment.

Пархоменко Іван є членом редакційної колегії цього видання, тому він не брав участі в процесі рецензування та прийняття рішення щодо публікації цієї статті.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

Parkhomenko Ivan is a member of the Editorial Board of this publication; therefore, he did not participate in the peer review process or the decision-making regarding the publication of this article.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.