УДК 004.912:004.942 DOI: 10.31673/2786-8362.2024.021324

Ланде Д.В., д.т.н.; **Пучков О.О.,** к.ф.н.;
Субач І.Ю., д.т.н.

МЕТОДИКА ПОБУДОВИ ТА АНАЛІЗУ МЕРЕЖІ ЗВ'ЯЗКІВ ДЖЕРЕЛ ІНФОРМАЦІЇ

Lande D.V., Puchkov O.O., Subach I.Yu. Methodology for building and analyzing the network of information sources. In today's cyberspace, monitoring and analyzing information flows are becoming critical to preventing cyber threats and detecting disinformation campaigns. Given the increasing number of cyber threats, an important task is to determine the influence of key information sources and analyze the links between sources that publish cybersecurity materials. The article proposes a methodology for building a network of links between information sources based on horizontal visibility graphs with time constraints, which allows analyzing the thematic similarity and temporal proximity of publications, which makes it possible to identify key sources of information and determine the likely initiators of information and cyberattacks. Criteria for establishing links between sources based on thematic similarity and temporal proximity have been developed, which is key to analyzing the spread of information in real time. An algorithm for constructing a horizontal visibility graph to form a network of information sources has been improved, which simultaneously takes into account the thematic similarity and proximity in time of publications. The use of network metrics, such as centrality and PageRank, is proposed, which allows assessing the credibility of information sources in the context of specific topics and time intervals, which significantly improves the understanding of the structure of information flows. The proposed approach not only helps to identify cyber threats and disinformation campaigns, but also assesses the credibility of resources, which is important for improving cybersecurity. Experimental results have shown the high accuracy of the method for identifying key sources and analyzing their role in the dissemination of information.

Keywords information flow, disinformation campaign, horizontal visibility graph, network of connections, thematic similarity, time constraints, credibility of the information source, network building algorithm, network influence metrics

Ланде Д.В., Пучков О.О., Субач І.Ю. Методика побудови та аналізу мережі зв'язків джерел інформації. У сучасному кіберпросторі моніторинг та аналіз інформаційних потоків набувають критичного значення для запобігання кіберзагрозам та виявлення дезінформаційних кампаній. В умовах збільшення кількості кіберзагроз важливою задачею є визначення впливу ключових інформаційних джерел та аналіз зв'язків між джерелами, які публікують матеріали з кібербезпеки. У статті пропонується методика побудови мережі зв'язків між джерелами інформації, що ґрунтується на графах горизонтальної видимості з часовими обмеженнями, яка дозволяє аналізувати тематичну схожість та часову близькість публікацій, що дає змогу виявляти ключові джерела інформації та визначати ймовірних ініціаторів інформаційних та кібератак. Розроблено критерії для встановлення зв'язків між джерелами на основі тематичної схожості та часової близькості, що є ключовим для аналізу поширення інформації у реальному часі. Удосконалено алгоритм побудови графу горизонтальної видимості для формування мережі інформаційних джерел, який одночасно враховує тематичну схожість і близькість у часі публікацій. Запропоновано використання мережевих метрик, таких як центральність та PageRank, що дозволяє оцінювати авторитетність інформаційних джерел у контексті конкретних тем і часових інтервалів, що значно покращує розуміння структури інформаційних потоків. Запропонований підхід не лише сприяє виявленню кіберзагроз та дезінформаційних кампаній, але й оцінює авторитетність ресурсів, що є важливим для підвищення рівня кібербезпеки. Результати експериментів показали високу точність методу виявлення ключових джерел та аналізу їхньої ролі у поширенні інформації.

Ключові слова: інформаційний потік, дезінформаційна кампанія, граф горизонтальної видимості, мережа зв'язків, тематична схожість, часові обмеження, авторитетність інформаційного джерела, алгоритм побудови мережі, метрики мережевої впливовості

Вступ

У сучасному кіберпросторі моніторинг та аналіз інформаційних потоків мають критичне значення для запобігання кіберзагрозам, виявлення дезінформаційних кампаній та визначення впливу ключових джерел.

У зв'язку зі зростанням кількості кіберзагроз і інформаційних атак, ефективне виявлення джерел дезінформації та розуміння поширення інформації стали пріоритетними завданнями для кібербезпеки. Однією з ключових задач є аналіз зв'язків між джерелами, які публікують матеріали з кібербезпеки, оскільки це дозволяє оцінювати авторитетність ресурсів, розуміти ризики поширення дезінформації та відстежувати її канали й аналізувати кіберзагрози.

Особливо важливим аспектом при аналізі інформаційних потоків є побудова мережі зв'язків між джерелами. Підставою для встановлення зв'язків між двома джерелами може бути схожість або збіг опублікованих ними документів за тематикою або близькістю у часі публікації. Така мережа надає можливість визначити, які інформаційні джерела є основними, найбільш впливовими або навіть ініціаторами певних інформаційних хвиль.

Аналіз подібних мереж допомагає виявити потенційні джерела кіберзагроз або дезінформаційних кампаній, вказуючи на ті джерела, які постійно випереджають інших у публікаціях на певні теми. Якщо джерело першим розповсюдило інформацію, воно може бути ймовірним джерелом цієї інформації, якщо інші чинники не свідчать про протилежне. Тому важливо враховувати не лише тематичну схожість публікацій, а й часові параметри публікацій.

Постановка завдання. У даній статті пропонується методика, що базується на застосуванні концепції графів горизонтальної видимості з обмеженням часу для побудови мереж зв'язків між інформаційними джерелами в кіберпросторі. Передбачається, що такий підхід дозволить визначати ключові та першоджерела інформації, аналізуючи як тематичну подібність, так і часові параметри публікацій.

Аналіз останніх досліджень і публікацій. Огляд основних концепцій та наукових напрямків, що мають відношення до цієї роботи охоплює концепцію графів горизонтальної видимості, їх обмежень за часом видимості, аналіз інформаційних мереж, підходів до моніторингу контенту, дослідження часових рядів, що відповідають динаміці тематичних публікацій.

Графи видимості (Visibility Graphs) — це метод перетворення часового ряду в мережу, де кожна точка часового ряду відповідає вершині графа [1]. Граф горизонтальної видимості (Horizontal Visibility Graph, HVG) — це одна з модифікацій графів видимості, коли вершини двох точок виявляються з'єднаними ребром, якщо між ними немає перешкод за висотою, тобто немає інших точок з більшою або рівною висотою між ними [2]. Цей підхід активно застосовується для аналізу часових рядів і дослідження складних систем, оскільки дозволяє виявити приховані структури і закономірності у даних, у тому числі текстових [3].

Робота [1] стала основою для концепції графа видимості. З того часу графи видимості використовуються у багатьох задачах, включаючи аналіз фінансових ринків, метеорологічні дані, а також кібербезпеку [4]. У роботі [2] представлено алгоритм побудови графа горизонтальної видимості, який є спрощеною геометричною та аналітично розв'язуваною версією алгоритму графа видимості. Показано, що цей алгоритм видимості є дієвим методом для відрізнення випадковості в часових рядах. У [3] запропоновано компактифікований граф горизонтальної видимості для мовної мережі та ідентифікації слів, які визначають інформаційну структуру тексту.

Слід зауважити, що аналіз інформаційних мереж передбачає вивчення зв'язків між джерелами інформації. Важливим є підхід до побудови мереж на основі контенту та часових характеристик. Дослідження [5] показує, як зміст соціальних зв'язків між журналістами та їхніми джерелами, а особливо множинність цих зв'язків, відображається в практиках виявлення новин у політичній журналістиці.

Класична робота [6] описує методи визначення авторитетних джерел на основі гіперпосилань, що є аналогічним до нашого підходу в кіберпросторі для виявлення основних джерел інформації.

Велика кількість досліджень зосереджена на розробці алгоритмів для автоматичного визначення ваги вузлів і ребр у графах. Наприклад, у [7] аналізується алгоритм PageRank, який оцінює важливість вебсторінок, використовуючи принципи теорії графів і теорії матриць. Подібні алгоритми можуть бути адаптовані для аналізу динамічних інформаційних мереж [8, 9].

Метою роботи є розробка нової методики побудови мережі зв'язків між інформаційними джерелами на основі схожості контенту та близькості у часі публікацій, яка дозволяє не лише виявити ключові та впливові джерела, але й встановити ймовірність того, що певне джерело є первинним розповсюджувачем інформації.

Для досягнення мети необхідно вирішити наступні часткові взаємопов'язані задачі:

1. Проаналізувати та обґрунтувати критерії зв'язку між джерелами на основі спільності або схожості контенту та близькості часу публікації.
2. Розробити метод та алгоритм побудови мережі, який будує мережу інформаційних джерел на основі заданих критеріїв, з урахуванням як тематичних, так і часових параметрів.
3. Визначити впливовість джерел у мережі, визначаючи найбільш авторитетні або центральні джерела у контексті тематики зі сфери кібербезпеки.
4. Встановити ймовірного первинного джерела інформації на основі часової пріоритетності публікацій та аналізу схожості контенту.

Вирішення цих задач дозволить на практиці створити ефективний інструмент для моніторингу та аналізу інформаційних потоків у сфері кібербезпеки, який може бути застосований для виявлення дезінформаційних кампаній та запобігання кіберзагрозам.

Виклад основного матеріалу дослідження.

Запропонована методика побудови мережі джерел інформації, ґрунтується на застосуванні графа горизонтальної видимості (Horizontal Visibility Graph, HVG), який дозволяє виявляти впливові джерела на основі аналізу тематичних інформаційних потоків та припущенні, що джерела інформації є заздалегідь ранжованими за обсягами публікацій з досвіду спостереження їх протягом тривалого часу. З формальної точки зору, їм уже приписані деякі вагові значення.

Задача, що підлягає рішенню може бути сформульованою наступним чином:

Дано: D – множина документів $\{D_1, D_2, \dots D_n\}$, що надходять у тематичний інформаційний потік;

S – множина джерел $\{S_1, S_2, \dots S_m\}$, що публікують документи $\{D_1, D_2, \dots D_n\}$;

t_i – час публікації документа D_i;

w_i – вага джерела (попередньо визначена на основі обсягу публікацій), яке опублікувало документ D_i;

τ – максимальний допустимий проміжок часу для встановлення зв'язку між джерелами.

Необхідно: побудувати мережу $G' = (V', E')$ для відображення зв'язків джерел інформації на задану тему зі сфери кібербезпеки та оцінки їхньої ваги.

Методика включає наступну послідовність кроків:

Крок 1. Побудова графу горизонтальної видимості. Здійснюється побудова часового ряду, елементи якого відповідатимуть документам з інформаційного потоку в порядку надходження їх у тематичний інформаційний потік. Значення ряду відповідатимуть ваговому значенню інформаційного джерела, яке опублікувало відповідний документ:

часовий ряд $S=\{(t_1,w_1),(t_2,w_2),\ldots,(t_n,w_n)\}$, де кожен елемент (t_i,w_i) відповідає документу D_i, опублікованому в момент часу t_i джерелом S_j, до якого відноситься документ D_i із вагою $w_i=w(S_j), D_i \sim S_j$

Зауважимо, що спочатку в якості ваги джерела S_j, можна використовувати, наприклад, його продуктивність – питому кількість повідомлень в інформаційному потоці, що відповідають цьому джерелу.

По цьому ряду будується граф горизонтальної видимості із заданим “напрямом погляду” у минуле (дивись рисунок 1):

для кожного елемента (t_i,w_i) у часовому ряді визначається з якими попередніми елементами (t_i,w_j) (де $i<j$) може бути встановлений зв’язок згідно з умовами горизонтальної видимості:

зв’язок між (t_i,w_i) та (t_i,w_j) існує, якщо: $w_k<\min(w_i,w_j)$, (1)

$$t_i-t_j\le\tau\,. \quad (2)$$

При виконанні умов (1), (2) формується граф $G=(V,E)$, де вузли V відповідають документам D_i, а ребра E – відповідають встановленим зв’язкам між документами.

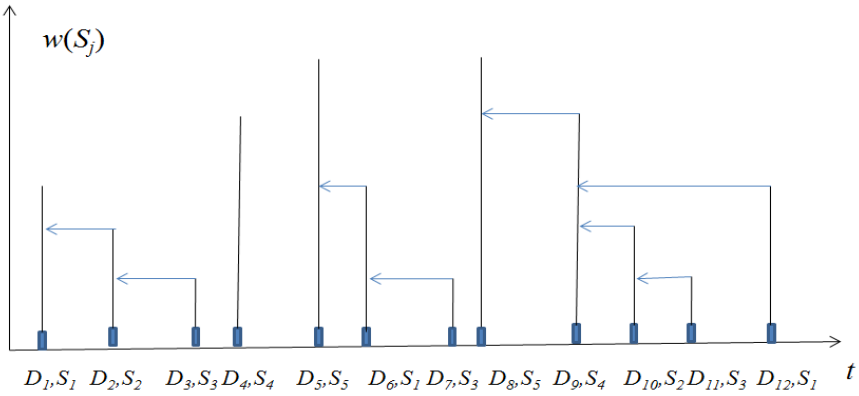

Рис. 1. Горизонтальна видимість документів D_i тематичного інформаційного потоку з “напрямом погляду” у минуле . Горизонтальна вісь – вісь часу з відкладеними точками – документами, вертикальна вісь – вага джерела, що відповідає документу

Крок 2. Представлення графу горизонтальної видимості у вигляді мережі для відображення зв’язків джерел інформації на задану тему зі сфери кібербезпеки.

Встановлюється зв’язок джерела інформації з іншим, більш рейтинговим, якщо він існує і опублікував інформацію раніше. Як можна побачити на рис. 1, зв’язок встановлюється, якщо “горизонтальний погляд” у минуле з вершини стовпчика (вага джерела, що відповідає поточному документу) впирається у стовпчик значення ваги джерела іншого документа, яке, вочевидь, має більше значення. При чому, зв’язок між джерелами встановлюється, при умові, що період часу не перевищив деякого порогу τ, наприклад, 24 години. Це припущення відповідає звичайній динаміці поширення новин, крім окремих випадків інформаційних операцій [10].

Таким чином, на кроці 2 створюється мережа $G'=(V',E')$, де вузли V' відповідають джерелам інформації, а ребра E' відображають зв’язки між джерелами, якщо між відповідними документами є зв’язок у графі горизонтальної видимості G.

Отже, вузол $v'\in V$ є пов’язаним із вузлом $v''\in V''$, якщо у графі G існує зв’язок між документом, опублікованим джерелом v', та документом, опублікованим джерелом v''.

Для здійснення оцінки ваги зв’язків, кожному ребру $e'\in E'$ мережі G' може бути присвоєна вага, що відображає частоту встановлення зв’язку між двома джерелами.

Крок 3. Оцінка ваги джерел, як першоджерел інформації. Ця оцінка здійснюється на основі алгоритму типу PageRank, який надає ймовірність того, що джерело інформації є першоджерелом або найбільш впливовим джерелом у потоці інформації. Зміст цього кроку, полягає у визначенні ваги вузлів у побудованій мережі горизонтальної видимості HVG джерел інформації.

Не зважаючи на те, що алгоритм PageRank розроблено для аналізу зв'язків між вебсторінками, його можна адаптувати для аналізу мереж зв'язків між джерелами інформації.

Ґрунтуючись на тому, що кожен вузол (джерело) буде отримувати вагу на основі кількості і якості зв'язків, що входять до нього від інших джерел припустимо, що у нашому випадку вузли – це джерела, а зв'язки – це публікації, які вказують на взаємодію між джерелами через подібність контенту та близькість публікацій у часі.

Основним принципом за яким працює алгоритм PageRank є, так зване, "голосування": кожен вузол передає свою вагу тим вузлам, на які посилається. У результаті отримуємо ситуацію: чим більше на джерело вказують інші вагомі джерела, тим більша ймовірність того, що це джерело є важливим і може бути першоджерелом інформації.

З формальної точки зору це можна представити наступним чином.

Будується граф горизонтальної видимості, де вузли відповідають джерелам інформації, а ребра між вузлами відображають зв'язки, що виникають через публікації, які є схожими за темою і близькі за часом.

Для кожного вузла i (джерела інформації) його PageRank $PR(i)$ можна визначити за наступною формулою:

$$PR(i) = (1-d)/N + d\sum_{j\in M(i)} \frac{PR(j)}{L(j)}, \tag{3}$$

де: $PR(i)$ – PageRank вузла i;

d – коефіцієнт демпінгу (зазвичай дорівнює 0.85);

N – загальна кількість вузлів (джерел інформації);

$M(i)$ – набір вузлів, які мають посилання на вузол i;

$L(j)$ – кількість вихідних зв'язків вузла j.

Результуючі значення $PR(i)$ для кожного вузла будуть відображати ймовірність того, що це джерело є першоджерелом або ключовим джерелом інформаційного потоку. Таким чином, більше значення $PR(i)$ вказує на більшу важливість джерела в мережі.

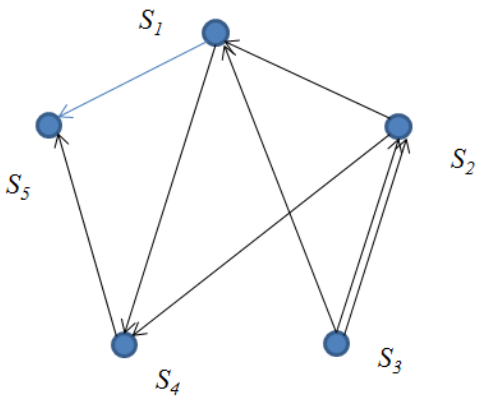

Рис. 2. Мережа джерел, що відповідає тематичному інформаційному потоку, зображеному на рис. 1.

Наведений підхід дозволяє формалізувати процес ідентифікації першоджерел інформації на основі аналізу зв'язків між публікаціями та їх авторами. У результаті ми отримуємо нову

мережу з оновленими вагами вузлів, яка відображає ймовірність того, що певне джерело стало початковим у ланцюгу публікацій або поширення інформації.

Побудована таким чином мережа джерел (див. рисунок 2) відображає зв'язки джерел за заданою тематикою зі сфери кібербезпеки та дозволяє визначати лідерів серед них не просто по кількості повідомлень, а й по їх тематиці, а також робити припущення щодо ймовірних першоджерел інформації

Таким чином, запропонована методика дозволяє визначити основні та найвпливовіші джерела на основі обробки заздалегідь ранжируваних за обсягами публікацій джерел та формуванні графу горизонтальної видимості на базі часового ряду, елементи якого відповідають документам з інформаційного потоку в порядку надходження їх у тематичний інформаційний потік.

Застосування сформованої мережі можливо для визначенні впливових джерел і визначення першоджерел. При цьому, джерела з більшою кількістю звя'зків або з вищою сумарною вагою зв'язків можна вважати більш впливовими у тематичному інформаційному потоці. Аналіз мережі дозволяє ідентифікувати джерела, які найчастіше виступають початковими вузлами для поширення інформації.

Висновки

Запропонована методика побудови мережі зв'язків між джерелами інформації, що базується на графах горизонтальної видимості з часовими обмеженнями, дозволяє ефективно аналізувати інформаційні потоки в кіберпросторі. Цей підхід забезпечує можливість не лише виявляти ключові джерела інформації, але й оцінювати їхню роль у розповсюдженні інформації, а також виявляти потенційні дезінформаційні кампанії та кіберзагрози. Завдяки одночасному використанню тематичної схожості контенту і близькості публікацій у часі, методика надає можливість краще зрозуміти процеси взаємодії між інформаційними джерелами та виявити ймовірні первинні джерела.

Наукова новизна отриманого в результаті проведеного дослідження результату полягає у наступному:

- удосконалено алгоритм побудови графу горизонтальної видимості для формування мережі інформаційних джерел, який на відміну від існуючих, одночасно враховує тематичну схожість і близькість у часі публікацій. Це дозволяє гнучко відстежувати взаємодію між джерелами та аналізувати процес поширення інформації про події у кіберпросторі;
- запропоновано новий метод, який дозволяє на основі часових пріоритетів та змістовної схожості визначати, які джерела можуть бути першими у розповсюдженні інформації, що є важливим для ідентифікації можливих ініціаторів дезінформаційних атак та кіберзагроз;
- запропоновано використання мережевих метрик, таких як центральність та PageRank, що дозволяє оцінювати авторитетність інформаційних джерел у контексті конкретних тем і часових інтервалів, що значно покращує розуміння структури інформаційних потоків.

Таким чином, вирішено низку часткових задач, а саме:

- розроблено критерії для встановлення зв'язків між джерелами на основі тематичної схожості та часової близькості, що є ключовим для аналізу поширення інформації у реальному часі;
- розроблено алгоритм побудови мережі зв'язків між інформаційними джерелами, який дозволяє автоматизовано генерувати мережеву структуру, яка відображає взаємодію джерел у кіберпросторі;
- на основі застосування мережевих метрик виконано оцінку впливовості різних джерел, що дає змогу виявити найбільш важливі вузли у мережі інформаційних потоків.

У свою чергу, використання часових параметрів і контентного аналізу дозволило розробити методику, що визначає найбільш ймовірне першоджерело інформації у мережі.

На практиці, розроблена методика може бути застосована для виявлення дезінформаційних кампаній, аналізу інформаційних та кіберзагроз, а також підвищення рівня кібербезпеки в умовах стрімкого зростання інформаційних та кібератак.

Список використаної літератури:

1. Lacasa, L., Luque, B., Ballesteros, F., Luque, J., Nuno, J. C. From time series to complex networks: The visibility graph // *Proceedings of the National Academy of Sciences*. 2008. vol. 105, no 13. P. 4972-4975. DOI: 10.1073/pnas.0709247105.

2. Luque, B., Lacasa, L., Ballesteros, F., Luque, J. Horizontal visibility graphs: Exact results for random time series // *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*. 2009. vol. 80, no 4. DOI: 10.1103/PhysRevE.80.046103.

3. Lande, D. V., Snarskii, A. A., Yagunova, E. V., Pronoza, E. V. The use of horizontal visibility graphs to identify the words that define the informational structure of a text // *2013 12th Mexican International Conference on Artificial Intelligence*. IEEE, 2013. P. 209-215. DOI: 10.1109/MICAI.2013.33.

4. Donner, R. V., Heitzig, J., Donges, J. F., Zou, Y., Marwan, N., Kurths, J. The geometry of chaotic dynamics – A complex network perspective. The European Physical Journal B. 2010. vol. 84, no. 4, P. 653–672. DOI: 10.1140/epjb/e2010-00108-2.

5. Malling, M. Sources that trigger the news: Multiplexity of social ties in news discovery. *Journalism Studies*. 2021. vol. 22, no 10. P. 1298-1316. DOI: 10.1080/1461670X.2021.1951331.

6. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*. 1999. vol. 46, no 5. P. 604-632. DOI: 10.1145/324133.324140.

7. Langville, A. N., Meyer, C. D. Google's PageRank and beyond: The science of search engine rankings. Princeton university press, 2006. DOI: 10.1515/9781400830329.

8. Sallinen, S., Luo, J., & Ripeanu, M. Real-time pagerank on dynamic graphs // *Proceedings of the 32nd International Symposium on High-Performance Parallel and Distributed Computing*. 2023. С. 239-251. DOI: 10.1145/3588195.3593004.

9. Lande, D., Snarskii, A., Dmytrenko, O., & Subach, I. Relaxation time in complex network // *Proceedings of the 15th International Conference on Availability, Reliability and Security, ARES '20*, 2020. pp. 1–6. DOI: 10.1145/3407023.3409231.

10. Dodonov, A., Lande, D., Tsyganok, V., Andriichuk, O., Kadenko, S., Graivoronskaya, A. Information Operations Recognition. From Nonlinear Analysis to Decision-Making. LAP Lambert Academic Publishing, 2019. 292 p. ISBN-13: 978-620-0-27697-1.

Автори статті

Ланде Дмитро – доктор технічних наук, професор, Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.

ORCID: 0000-0003-3945-1178

Пучков Олександр – кандидат філософських наук, професор, Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.

ORCID: 0000-0002-8585-1044

Субач Ігор – доктор технічних наук, професор, Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна.

ORCID: 0000-0002-9344-713X

Authors of the article

Lande Dmytro –Doctor of Science **(**technic), Professor**,** National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

ORCID: 0000-0003-3945-1178

Puchkov Oleksandr – Candidate of Science (philosophy), Professor, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

ORCID: 0000-0002-8585-1044

Subach Ihor – Doctor of Science **(**technic), Professor**,** National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

ORCID: 0000-0002-9344-713X