УДК 004.891:004.414.2

***Олексій КОВЕНЬКО***
аспірант кафедри інтелектуальних кібернетичних систем, «Київський авіаційний інститут»,
3414934@stud.kai.edu.ua
***ORCID:*** 0009-0007-2418-7761

***Наталія АПЕНЬКО***
кандидат технічних наук, доцент,
доцент, кафедри інтелектуальних кібернетичних систем, «Київський авіаційний інститут»,
nataliia.apenko@npp.kai.edu.ua
***ORCID:*** 0000-0001-6891-0869

# МЕТОД СЕМАНТИЧНОЇ ПРЕФІЛЬТРАЦІЇ ПРИЧИННО-НАСЛІДКОВИХ ЗВ'ЯЗКІВ У ВИСОКОВИМІРНИХ МЕРЕЖАХ

**Анотація.** У статті розв'язується фундаментальна проблема експоненційної обчислювальної складності алгоритмів виявлення причинно-наслідкових зв'язків у мережах надвисокої розмірності. Запропоновано новий метод семантичної префільтрації, що ґрунтується на гіпотезі семантичної розрідженості та використовує векторні представлення вузлів. Експериментально доведено, що метод дозволяє звузити простір пошуку (до 88.3% відкинутих пар для великих графів) за умови збереження понад 90% істинних ребер. Розроблений алгоритм успішно долає проблему «холодного старту» в умовах повної відсутності історичних спостережень та природним шляхом масштабується зі збільшенням складності системи.

**Мета.** Розробка та емпірична валідація методу семантичної префільтрації, здатного звузити простір пошуку причинно-наслідкових зв'язків у високовимірних графах.

**Методологія.** Методологічною основою є гіпотеза семантичної розрідженості та включає чотири етапи: формування текстових тлумачень для кожного вузла, їх векторизацію у латентному просторі, розрахунок матриці косинусної подібності та застосування стратегії адаптивного відсікання.

**Наукова новизна.** Вперше сформульовано та емпірично підтверджено гіпотезу семантичної розрідженості для причинно-наслідкових мереж. На її основі розроблено новий метод до префільтрації каузальних графів надвисокої розмірності на основі семантики вузлів та без залучення статистичних рядів спостережень.

**Висновки.** Запропонований метод семантичної префільтрації ефективно долає експоненційну обчислювальну складність пошуку причинно-наслідкових зв’язків та дозволяє відсіяти до 88.3% нерелевантних пар вузлів при збереженні цільової повноти ≥ 0.90. Емпірично підтверджено, що алгоритм масштабується пропорційно до розміру мережі та розв'язує проблему недостатності історичних даних (HDLSS).

**Ключові слова:** причинно-наслідкові графи, семантична префільтрація, прокляття розмірності, великі мовні моделі, векторні представлення, адаптивне відсікання, причинно-наслідкове відкриття.

**Oleksii KOVENKO, Natalia APENKO. METHOD OF SEMANTIC PREFILTERING OF CAUSAL RELATIONSHIPS IN HIGH-DIMENSIONAL NETWORKS**

**Abstract.** The article addresses the fundamental problem of the exponential computational complexity of causal discovery algorithms in high-dimensional networks. A novel semantic pre-filtering method is proposed, based on the semantic sparsity hypothesis and utilizing vector representations of nodes. It has been empirically demonstrated that the method allows narrowing the search space (with up to 88.3% of pairs rejected for large graphs) while preserving over 90% of true edges.

The developed algorithm successfully overcomes the "cold start" problem in the complete absence of historical observations and scales naturally with the increasing complexity of the system.

**Objective.** Development and empirical validation of a semantic pre-filtering method capable of narrowing the search space for causal relationships in high-dimensional graphs.

**Methodology.** The methodology is based on the semantic sparsity hypothesis and comprises four stages: generating textual interpretations for each node, their vectorization in a latent space, calculating a cosine similarity matrix, and applying an adaptive pruning strategy.

**Scientific Novelty.** For the first time, the semantic sparsity hypothesis for causal networks has been formulated and empirically confirmed. Based on this, a novel method for the pre-filtering of ultra-high-dimensional causal graphs has been developed, relying solely on node semantics and without involving statistical observation series.

**Conclusions.** The proposed semantic pre-filtering method effectively overcomes the exponential computational complexity of causal discovery and allows filtering out up to 88.3% of irrelevant node pairs while maintaining a target recall $\geq 0.90$. It has been empirically confirmed that the algorithm scales proportionally to the network size and resolves the problem of insufficient historical data (HDLSS).

**Keywords:** causal graphs, semantic pre-filtering, curse of dimensionality, LLM, embeddings, adaptive pruning, causal discovery.

**Постановка проблеми.** Побудова точних причинно-наслідкових моделей (Causal Discovery) є критичним етапом у створенні надійних систем підтримки прийняття рішень (DSS) та пояснюваного штучного інтелекту (XAI). У сучасних високотехнологічних доменах, таких як системна біологія, нейронаука та моніторинг розподілених IT-систем, спостерігається тенденція до надвисокої розмірності даних. Наприклад, аналіз функціональної МРТ (fMRI) вимагає обробки графів із понад 50 000 вокселів [2], геномні дослідження оперують

мережами діапазону 20 000 – 1 000 000 вузлів [2, 3], а сучасні системи телеметрії генерують тисячі метрик у реальному часі.

Ключовою проблемою в таких умовах стає «прокляття розмірності» (Curse of Dimensionality). При лінійному збільшенні кількості вузлів N кількість можливих спрямованих ациклічних графів (DAGs) зростає суперекспоненційно, створюючи комбінаторно складний простір пошуку [11, 19]. Класичні алгоритми пошуку причинності, такі як PC або GES, стикаються з фундаментальними обмеженнями: їхня часова складність на щільних графах зростає експоненційно [7, 19]. Емпіричні дослідження демонструють, що навіть оптимізовані версії цих алгоритмів стають обчислювально нездійсненними або не досягають збіжності при обробці щільних графів розмірністю понад 1000 вузлів через вичерпання ресурсів пам'яті та часу [1, 2].

Ситуація ускладнюється в умовах недостатності історичних даних, що призводить до сценарію «висока розмірність, малий обсяг вибірки» (HDLSS). Традиційні методи вимагають, щоб обсяг вибірки зростав пропорційно до розмірності мережі; порушення цієї умови спричиняє різке падіння статистичної потужності. Наприклад, для надійного відновлення структури мережі алгоритмом FullCI або Ancestral Causal Inference точність виявлення зв'язків знижується до критичного рівня вже при 20 змінних, якщо обсяг вибірки є недостатнім відносно розмірності простору ознак [3].

Окремим аспектом проблематики є фундаментальний компроміс між повнотою виявлення зв'язків та часовою складністю алгоритму. Оскільки задача точного відновлення структури баєсової мережі належить до класу NP-складних [7], досягнення 100% повноти на мережах високої розмірності є практично недосяжним [11]. Сучасні дослідження вказують на доцільність переходу до наближених рішень: використання «грубозернистих» представлень або евристичних обмежень дозволяє скоротити час обчислень на порядки [20]. У практичних задачах (наприклад, Root Cause Analysis) здатність алгоритму швидко локалізувати більшість ключових причин є пріоритетнішою за пошук

глобального оптимуму, який є недосяжним через обмеження обчислювальних ресурсів [8, 20].

Таким чином, актуальним науковим завданням є розробка методів префільтрації, які б звужували простір пошуку ще до етапу статистичної перевірки. Відсутність ефективних методів, які б не залежали від обсягу історичних даних (на відміну від SIS або Lasso) [10], а спиралися б на семантичну природу змінних, гальмує впровадження каузального ШІ у великомасштабних системах.

**Аналіз останніх досліджень і публікацій.** Проблема масштабованості алгоритмів виявлення причинно-наслідкових зв'язків у високовимірних просторах вирішується за трьома основними напрямами: алгоритмічною оптимізацією, статистичною фільтрацією та інтеграцією знань LLM.

Фундаментальні дослідження у цій сфері спираються на constraint-based (PC, FCI, ACI) [18, 19] та score-based (GES) [7] підходи, розвинені P. Spirtes, C. Glymour [19] та J. Ramsey [2]. Попри теоретичну обґрунтованість, ці методи характеризуються суперекспоненційним зростанням простору пошуку [11]. Емпіричні дані свідчать, що перехід від 50 000 до 1 000 000 вузлів призводить до зростання часу виконання у 83 рази навіть за умови високої розрідженості графа [2], а на щільних графах розмірністю $N > 1000$ застосування цих алгоритмів стає обчислювально непрактичним. Апаратне прискорення та розпаралелювання (Parallel-PC, GPU-PC) лише частково нівелюють цю проблему, не змінюючи алгоритмічної складності перевірок умовної незалежності [1, 12, 13].

Окремою проблемою є залежність класичних методів від обсягу вибірки. Дослідження показують, що алгоритми втрачають статистичну потужність детекції в умовах обмежених даних (HDLSS) [3]. Методи статистичного скрінінгу (SIS, Lasso) ефективні для зменшення розмірності [10], проте вони повністю залежать від історичних даних і є незастосовними у сценаріях «холодного старту», коли доступні лише метадані змінних.

Інтеграція великих мовних моделей (LLM) дозволила використовувати семантичні зв'язки як джерело апріорних знань [6, 15, 16]. Попри високу точність

у задачах причинно-наслідкового виявлення, існуючі методи стикаються з критичними проблемами масштабованості [17] та фундаментальними сумнівами щодо здатності моделей до істинного каузального мислення («Causal Parrots») [5]. Підходи на основі попарних запитів (Pairwise Querying) вимагають квадратичної кількості звернень до моделі $O(N^2)$, що є економічно та часово неприйнятним для мереж із тисячами вузлів [9]. Навіть оптимізовані стратегії, такі як пошук у ширину (BFS) [9] або ітеративна генерація підграфів, залишаються обмеженими через високу латентність генеративних моделей. Значним кроком став фреймворк IRIS (Feng et al., 2024), який уможливлює Causal Discovery за відсутності табличних даних через ітеративний аналіз текстових корпусів [14]. Однак цей підхід орієнтований на глибокий видобуток нових сутностей (Knowledge Extraction) та є обчислювально надлишковим для задач швидкої структурної префільтрації фіксованих просторів ознак великої розмірності. Водночас існуючі векторні підходи фокусуються переважно на передбаченні зв’язків, а не на побудові скелета графа [4].

Таким чином, у літературі відсутній метод, який би поєднував незалежність від історичних даних, лінійну обчислювальну складність та здатність ефективно розріджувати простір пошуку для графів надвисокої розмірності. Саме розробці такого методу семантичної префільтрації присвячена ця стаття.

**Виклад основного матеріалу.** Для подолання експоненційної складності алгоритмів пошуку причинності у високовимірних просторах розроблено метод семантичної префільтрації. Запропонований метод ґрунтується на гіпотезі семантичної розрідженості (Semantic Sparsity Hypothesis): у великих системах причинно-наслідкові зв’язки існують переважно між семантично спорідненими об'єктами, тоді як семантично віддалені змінні є ймовірно незалежними. Розроблений метод складається з чотирьох послідовних етапів (рис. 1). На першому етапі формується текстовий опис кожної змінної графа, який може бути складений доменними експертами або згенерований автоматично на основі доступних метаданих. На другому етапі отримані тексти перетворюються у числові вектори (embeddings) за допомогою моделі векторизації, що дозволяє

кількісно оцінювати семантичну близькість через геометричну відстань у векторному просторі. На третьому етапі обчислюється матриця косинусної подібності між усіма парами змінних графа. На четвертому етапі застосовується стратегія адаптивного відсікання (Adaptive Top-k Pruning), за якої для кожного вузла зберігаються лише k сусідів з найвищим показником подібності, а решта пар відкидаються. Отриманий розріджений граф слугує вхідними даними для класичних алгоритмів (наприклад, PC або GES), радикально звужуючи простір їхнього пошуку.

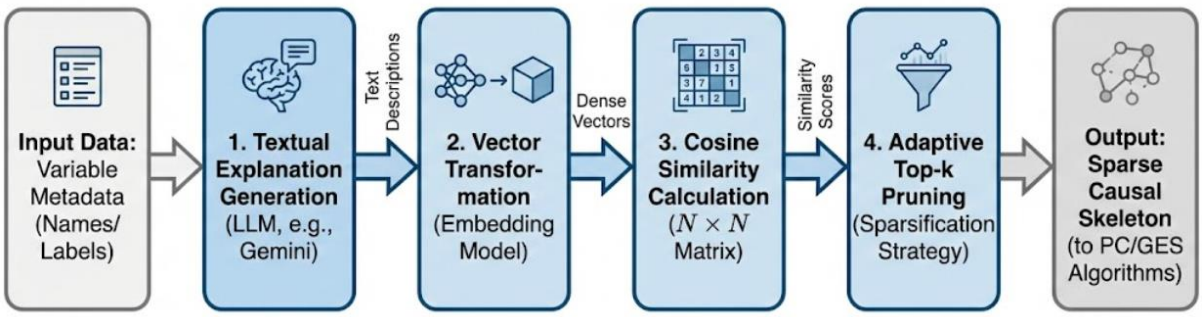


**Рис. 1 Блок-схема методу семантичної префільтрації.**

Для емпіричної валідації методу було використано набір стандартних бенчмарк-мереж із репозиторію bnlearn, які охоплюють різні предметні області (медицина, агрономія, генетика) та варіюються за розміром від малих (35 вузлів) до масивних (1041 вузол). До вибірки увійшли мережі: mildew (35), alarm (37), barley (48), hepar2 (70), win95pts (76), diabetes (413), link (724) та munin (1041). Для першого етапу (генерації тлумачень) було використано велику мовну модель gemini-2.5-flash, яка формувала розгорнуті описи виключно на основі назв змінних. Для виконання другого етапу (векторизації) застосовано модель gemini-embedding-001.

Оцінка ефективності префільтрації проводилася за чотирма математичними метриками. Нехай $E_{true}$ – множина ребер істинного графа, $E_{filtered}$ – множина ребер, збережених у скелеті після фільтрації, а $E_{full}$ – множина всіх можливих

пар вузлів графа. Повнота (Recall) визначалася як частка збережених істинних зв'язків:

$$Recall = \frac{|E_{true} \cap E_{filtered}|}{|E_{true}|}$$

Ступінь стиснення простору пошуку (Reduction Rate, RR) обчислювався як частка відкинутих ребер відносно повного графа:

$$RR = 1 - \frac{|E_{filtered}|}{|E_{full}|}$$

Точність (Precision) визначалася за формулою:

$$Precision = \frac{|E_{true} \cap E_{filtered}|}{|E_{filtered}|}$$

Коефіцієнт переваги над випадковим відбором (Lift) обчислювався як відношення точності розробленого методу до загальної густини істинного графа (Density):

$$Density = \frac{|E_{true}|}{|E_{full}|}$$

$$Lift = \frac{Precision}{Density}$$

Критерієм успішності було визначено збереження не менше 90% істинних причинних зв'язків (Recall ≥ 0.90), оскільки їхня втрата на етапі префільтрації є незворотною. Отримані результати зведено у таблицю 1.

Аналіз даних виявив зростання ефективності методу зі збільшенням розмірності графа. На масивному графі munin (1041 вузол) алгоритм досяг найвищих показників: для збереження 90.1% істинних зв'язків знадобилося залишити лише 91 найближчого сусіда для кожного вузла. Це дозволило відсіяти 88.3% усіх можливих пар змінних. Крім того, найвище значення метрики Lift зафіксовано саме для графа munin (7.67). Це доводить здатність методу ідентифікувати потенційні причинно-наслідкові зв'язки майже у 8 разів ефективніше за випадкове вгадування. Високу загальну ефективність також зафіксовано для мережі diabetes (413 вузлів), де Reduction Rate склав 0.773, а

показник Lift — 3.97. Це підтверджує гіпотезу про те, що у великих системах змінні формують виражені семантичні кластери, які успішно виявляються мовними моделями.

Таблиця 1

**Показники ефективності семантичної префільтрації при цільовому рівні Recall ≥ 0.90**

| Dataset | Вузли (N) | Обране k | Recall | Reduction Rate | Lift |
|---|---|---|---|---|---|
| mildew | 35 | 19 | 0.978 | 0.343 | 1.49 |
| alarm | 37 | 13 | 0.935 | 0.557 | 2.11 |
| barley | 48 | 37 | 0.905 | 0.097 | 1.00 |
| hepar2 | 70 | 43 | 0.927 | 0.240 | 1.22 |
| win95pts | 76 | 49 | 0.920 | 0.239 | 1.21 |
| diabetes | 413 | 73 | 0.900 | 0.773 | 3.97 |
| link | 724 | 217 | 0.902 | 0.613 | 2.33 |
| munin | 1041 | 91 | 0.901 | **0.883** | **7.67** |

Натомість на мережах меншої розмірності результати виявилися неоднорідними. Наприклад, для мережі barley (48 вузлів) метод показав найнижчу ефективність (Reduction Rate = 0.097, Lift = 1.00), що свідчить про слабку кореляцію між структурою зв'язків та семантичною близькістю назв змінних у цьому датасеті або ж про надмірну щільність семантичного простору даної предметної області.

Візуалізацію переваги розробленого методу над випадковим відбором (за метрикою Lift) для всіх наборів даних наведено на рис. 2.

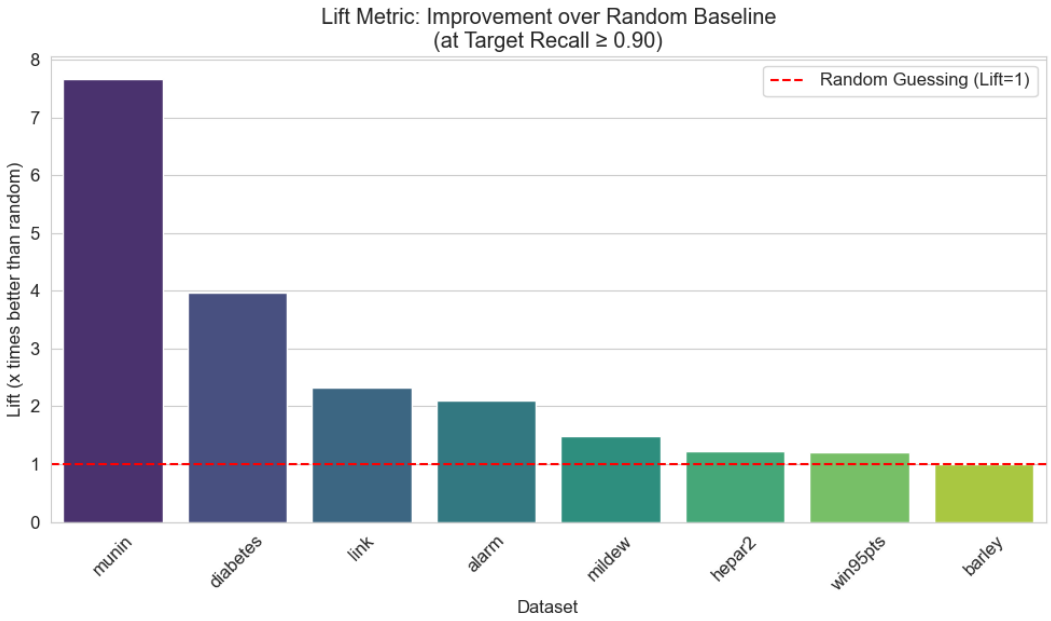


**Рис. 2. Залежність метрики Lift від набору даних.**

Наступним важливим аспектом дослідження є аналіз компромісу між повнотою виявлення зв'язків (Recall) та ефективністю стиснення простору пошуку (Reduction Rate). Побудована крива ефективності (рис. 3) показує, що для масивних графів точки оптимального співвідношення зміщені у цільову зону (верхній правий кут графіка).

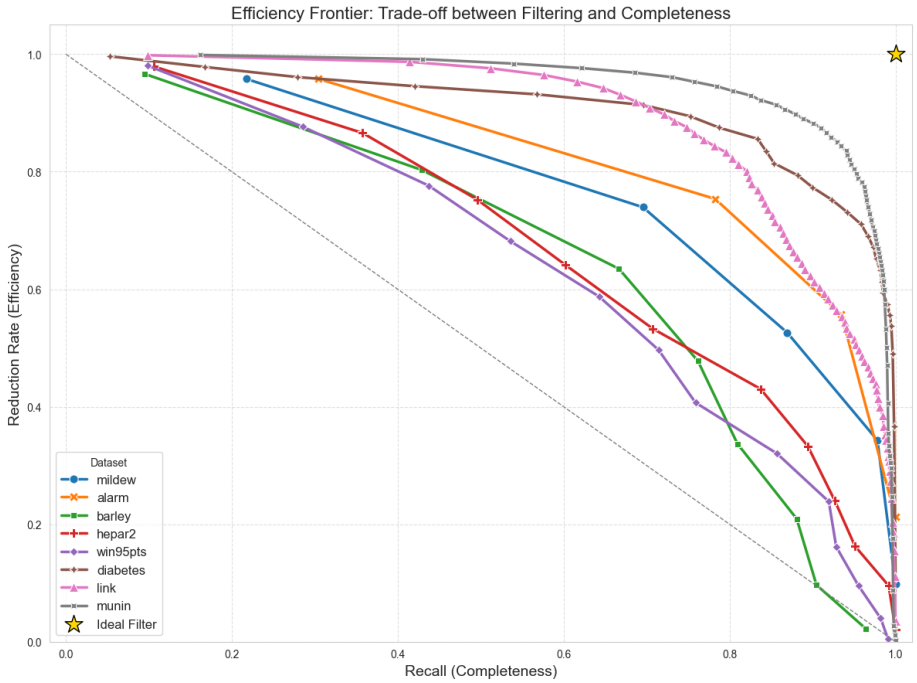


**Рис. 3. Залежність ступеня стиснення від повноти.**

Для остаточного підтвердження масштабованості методу було проаналізовано залежність відносного показника збережених сусідів від загальної розмірності графа ($k/N$) за умови фіксованого рівня повноти Recall ≥ 0.90.

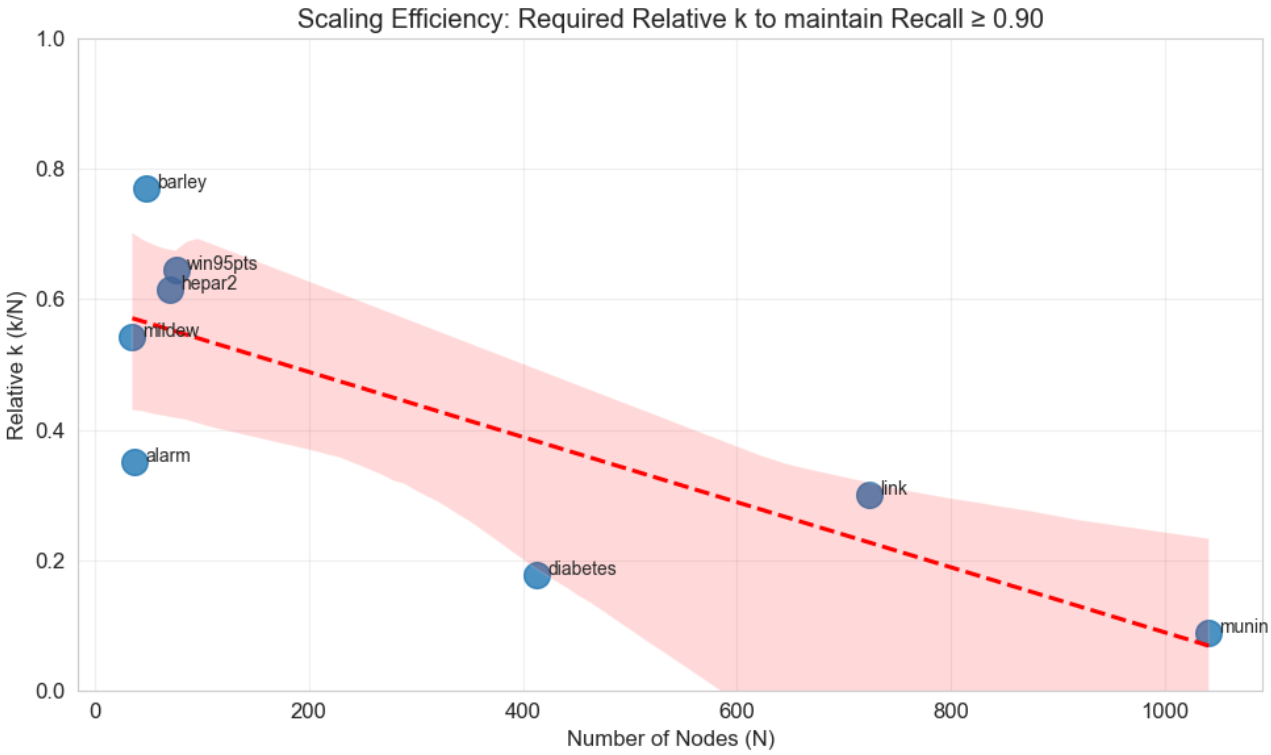


**Рис. 4. Залежність відносного показника k від розмірності мережі.**

Аналіз лінії тренду демонструє чітку зворотну залежність: якщо для малих мереж (наприклад, barley або win95pts) алгоритм зберігає від 50% до 80% вузлів як потенційних кандидатів на утворення ребер, то для масивного графа munin цей показник падає нижче 10%. Цей ефект пояснюється топологічною природою розростання мереж: зі збільшенням загальної кількості вузлів середня кількість зв'язків для кожного окремого елемента не зростає пропорційно. Як наслідок, у високовимірних просторах граф стає екстремально розрідженим у відносних показниках, а локальні кластери стають значно віддаленішими один від одного. Завдяки цьому моделі векторизації отримують змогу легко ідентифікувати та відсікати переважну більшість очевидно непов'язаних сутностей. Таким чином, ефективність семантичної префільтрації масштабується природним шляхом: що

більшою є загальна розмірність системи, то вищу відносну частку нерелевантних зв'язків метод здатний превентивно видалити з простору пошуку.

**Висновки.** У дослідженні розроблено метод семантичної префільтрації, який ефективно розв'язує фундаментальну проблему «прокляття розмірності» та експоненційної обчислювальної складності в задачах пошуку причинно-наслідкових зв’язків. Експериментально підтверджено гіпотезу семантичної розрідженості: використання семантичних векторних представлень вузлів дозволяє радикально звузити простір пошуку ще до етапу застосування класичних алгоритмів структурного навчання. Крім того, метод продемонстрував високу результативність в умовах повної відсутності історичних спостережень, успішно розв'язуючи проблему «холодного старту».

Перспективи подальших досліджень фокусуються над вдосконаленням логіки фільтрації шляхом упровадження гібридного відсікання (поєднання стратегії Top-k із жорстким порогом семантичної подібності), динамічного розширення параметра k для вузлів-концентраторів (hubs) та підвищення якості семантичного сигналу за допомогою застосування ансамблевих векторних представлень (агрегація кількох тлумачень).

### *Список використаних джерел:*

1. A Fast PC Algorithm for High Dimensional Causal Discovery with Multi-Core PCs / T. D. Le et al. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019. Vol. 16, no. 5. P. 1483–1495. DOI: 10.1109/TCBB.2016.2591526.

2. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images / J. Ramsey et al. International Journal of Data Science and Analytics. 2017. Vol. 3, no. 2. P. 121–129. DOI: 10.1007/s41060-016-0032-z.

3. Ancestral causal learning in high dimensions with a human genome-wide application / U. Noè et al. arXiv preprint arXiv:1910.05166. 2019. DOI: 10.48550/arXiv.1910.05166.

4. Balashankar A., Subramanian L. Learning Faithful Representations of Causal Graphs. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL). 2021. P. 1002–1011. DOI: 10.18653/v1/2021.acl-long.81.

5. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal / M. Zečević et al. Transactions on Machine Learning Research. 2023. DOI: 10.48550/arXiv.2308.13067.

6. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality / E. Kıcıman et al. Transactions on Machine Learning Research. 2024. DOI: 10.48550/arXiv.2305.00050.

7. Chickering D. M. Optimal structure identification with greedy search. Journal of Machine Learning Research. 2002. Vol. 3. P. 507–554. DOI: 10.1162/153244303321897717.

8. Detecting and quantifying causal associations in large nonlinear time series datasets / J. Runge et al. Science Advances. 2019. Vol. 5, no. 11. DOI: 10.1126/sciadv.aau4996.

9. Efficient Causal Graph Discovery Using Large Language Models / T. Jiralerspong et al. arXiv preprint arXiv:2402.01207. 2024. DOI: 10.48550/arXiv.2402.01207.

10. Fan J., Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008. Vol. 70, no. 5. P. 849–911. DOI: 10.1111/j.1467-9868.2008.00674.x.

11. Feigenbaum I., Khanna S., Vempala S. S. On the Unlikelihood of D-Separation. Proceedings of Machine Learning Research. 2024. Vol. 246. P. 1–17. DOI: 10.48550/arXiv.2303.05628.

12. Guo C., Luk W. Accelerating Constraint-Based Causal Discovery by Shifting Speed Bottleneck. Proceedings of the 2022 ACM/SIGDA International Symposium on

Field-Programmable Gate Arrays (FPGA '22). 2022. P. 123–134. DOI: 10.1145/3490422.3502363.

13. Hagedorn C., Huegle J. GPU-Accelerated Constraint-Based Causal Structure Learning for Discrete Data. SIAM International Conference on Data Mining (SDM). 2021. P. 37–45. DOI: 10.1137/1.9781611976700.5.

14. IRIS: An Iterative and Integrated Framework for Verifiable Causal Discovery in the Absence of Tabular Data / T. Feng et al. arXiv preprint arXiv:2406.10526. 2024. DOI: 10.48550/arXiv.2406.10526.

15. Darvariu V.-A., Hailes S., Musolesi M. Large Language Models are Effective Priors for Causal Graph Discovery. arXiv preprint arXiv:2401.12838. 2024. DOI: 10.48550/arXiv.2401.12838.

16. Large Language Models for Causal Discovery: Current Landscape and Future Directions / G. Wan et al. International Joint Conference on Artificial Intelligence (IJCAI). 2024. DOI: 10.24963/ijcai.2024/889.

17. LLM-Driven Causal Discovery via Harmonized Prior / T. Ban et al. IEEE Transactions on Knowledge and Data Engineering. 2024. DOI: 10.1109/TKDE.2025.3353067.

18. Magliacane S., Claassen T., Mooij J. M. Ancestral Causal Inference. Advances in Neural Information Processing Systems. 2016. Vol. 29. P. 4473–4481. URL: https://papers.nips.cc/paper/6266-ancestral-causal-inference.

19. Spirtes P., Glymour C., Scheines R. Causation, Prediction, and Search. 2nd ed. Cambridge : MIT Press, 2000. 543 p. DOI: 10.7551/mitpress/1754.001.0001.

20. Ultra-Scalable and Efficient Methods for Hybrid Observational and Experimental Local Causal Pathway Discovery / A. Statnikov et al. Journal of Machine Learning Research. 2015. Vol. 16. P. 3219–3267. URL: https://jmlr.org/papers/v16/statnikov15a.html.