УДК 004.855:519.216

*О.С. Балабанов*

# ПРИНЦИПИ ТА АНАЛІТИЧНІ ЗАСОБИ РЕКОНСТРУКЦІЇ СТРУКТУР ЙМОВІРНІСНИХ ЗАЛЕЖНОСТЕЙ У СПЕЦІАЛЬНОМУ КЛАСІ

Запропоновано та обґрунтовано набір емпіричних резолюції, які спираються виключно на безумовні залежності двох змінних та забезпечують ідентифікацію безпосередніх зв'язків (ребер) у структурах залежностей в класі монопотокових графів. Цей клас структур є підкласом ациклонних орграфів та суперкласом для полі-лісів. Охарактеризовано властивості монопотокових моделей. Коректність розроблених емпіричних резолюцій ґрунтується на емпірично надійному припущенні безумовної (маргінальної) реберної неоманливості.

Ключові слова: відтворення структури залежностей, монопотоковий граф, безумовна залежність, емпіричне проявлення залежностей, емпіричні резолюції ідентифікації ребер.

## Вступ

Відтворення структур зв'язків та впливів між характеристиками середовища, які неявно відбиті в багатовимірних масивах статистичних даних – одна з центральних задач глибокого аналізу даних та відкриття знань у базах даних. Хоча за останню чверть століття отримано багато результатів для розв'язання цієї важкої задачі, інтенсивні дослідження й розробки тривають [1–5]. Найбільш важка постановка задачі постає в ситуації, коли на структуру ймовірнісних залежностей не накладено жодних обмежень і зовсім нічого не відомо про цю структуру. Зазвичай передбачається єдине обмеження – в структурі немає орієнтованих циклів (циклонів). Відсутність циклонів можна вважати вимогою коректного збору даних. Тобто в процесі генерації одного запису даних кожна змінна $X$ вимірюється досить швидко, так що «сигнал» від $X$ до інших змінних не може встигнути «оббігти коло» й вплинути на $X$ у цьому записі. Структури ймовірнісних залежностей, де заборонено циклони, об'єднуються в клас ациклонних орієнтованих графів (АОГ). Задача відтворення автентичної (генеративної) АОГ-структури за відсутності апріорних знань у найгіршому випадку є експоненційно важкою [4, 5]. Натомість якщо відомо, що структура генеративної моделі задовольняє певним «топологічним» обмеженням, то відтворення такої структури значно спрощується. Для різних спеціальних класів структур й моделей розроблено спеціалізовані методи відтворення моделі. Коли генеративна модель дійсно належить обраному класу, то така задача характеризується як реконструкція (відтворення) моделі. Коли генеративна модель виходить за межі обраного класу, то задачу називають «редукція» (структурна апроксимація, «покриття») моделі. В статті розглянуто постановку, коли відомо, що структура моделі належить до класу монопотокових графів залежностей. Строго обґрунтовано нові засоби вдосконалення методів реконструкції моделей у цьому класі.

## Характеристика монопотокових моделей залежностей та огляд методів їх відтворення

*Ребро* відображає безпосередній зв'язок між двома змінними (вершинами). В більшості практичних задач працюють з моделями в підкласі ациклонних орієнтованих графів, де всі ребра є одноорієнтовані (тобто мають вигляд $X \rightarrow Y$). На основі таких графів (задавши параметри локальних залежностей) утворюється клас ординарних АОГ-моделей (оАОГ-моделей), у тому числі байєсові та гауссові мережі [1, 3, 6, 7]. Спеціальні підкласи оАОГ-моделей визначаються певними «топологічними» обмеженнями на структуру (вводиться заборона циклів або шляхів певних типів). Нагадаємо необхідні графові поняття.

Ребро зображується як неорієнтоване $X - Y$, коли орієнтація цього ребра є невідома або несуттєва в даному контексті. *Цикл* у графі – це шлях з ребер $X - Y - \cdots - X$, де всі проміжні (некінцеві) вершини – різні, а кінцеві вершини – тотожні. *Колізор* – це шлях вигляду $X \to Y \leftarrow Z$. *Ланцюг* – це шлях, на якому немає жодного колізора. Оршлях – це ланцюг вигляду $X \to \cdots \to Z \to Y$. Циклон – це оршлях, де остання вершина збігається з першою.

Мабуть найбільш відомим підкласом оАОГ є ліси залежностей. В лісі (дереві) залежностей немає ані циклів, ані колізорів. Його розширенням є полі-ліси; в них заборонені цикли (а колізори – дозволені). Ліси й полі-ліси підтримують прості паттерни залежностей, недостатні для моделювання багатьох об'єктів. Проміжне місце між полі-лісами та загальним випадком оАОГ посідає клас так званих монопотокових графів залежностей (МПГЗ). Моделі на базі МПГЗ можуть мати цикли і є значно експресивнішими за полі-ліси. (В літературі вони називалися простими або спрощеними оАОГ). Аксіоми та властивості МПГЗ-моделей описані в [6, 7]. Можна дати кілька еквівалентних варіантів **визначення** МПГЗ:

1) якщо в МПГЗ існує ланцюг $X \to Q \to Y$ або ребро $X - Y$, то між вершинами $X$ та $Y$ не існує жодного іншого ланцюга;

2) між батьками однієї й тієї ж вершини в МПГЗ немає жодного ланцюга;

3) в МПГЗ неможливий цикл з одним колізором.

Наслідками визначення є властивості МПГЗ.

*Властивість* 1: якщо в МПГЗ між вершинами $X$ та $Y$ існує більше одного ланцюга, то серед тих ланцюгів немає жодного оршляху, тобто всі ті ланцюги мають вигляд $X \leftarrow \cdots \to Y$. Візьмемо довільно два таких ланцюга $\lambda$ та $\mu$. На ланцюгах $\lambda$ та $\mu$ є змінні відповідно $Q$ та $R$, такі, що між $Q$ та $R$ не існує жодного ланцюга.

*Властивість* 2: якщо в МПГЗ між вершинами $X$ та $Y$ існують різні ланцюги $\lambda$ та $\mu$, з тим, що вони проходять через спільну вершину $Q$, яка не тотожна ані $X$, ані $Y$, то тоді ланцюги $\lambda$ та $\mu$ закінчуються спільною дугою $X \leftarrow L$ або (та) спільною дугою $R \to Y$. (Можливо, що $Q \equiv L$ або $Q \equiv R$).

Сенс цієї властивості: коли в МПГЗ кілька ланцюгів між $X$ та $Y$ «зійшлися», вони вже не можуть розійтися (розділитися), бо це породило б одноколізорний цикл.

Залежність між змінними забезпечується неблокованим шляхом між змінними. Змінні $X$ та $Y$ є безумовно залежні тоді й тільки тоді, якщо між $X$ та $Y$ існує хоча б один ланцюг. Марковські властивості класу оАОГ-моделей визначаються критерієм d-сепарації [1, 3, 6, 8]. Кожному факту d-сепарації відповідає умовна незалежність. Процедурі перевірки факту d-сепарації можна зіставити статистичний тест незалежності. Ранг тесту незалежності визначається кількістю змінних в умові тесту. (Тест безумовної незалежності має нульовий ранг).

Відомим показником тісноти (сили) залежності між дискретними змінними $X$ та $Y$ є взаємна інформація (за Шенноном). Безумовну взаємну інформацію позначимо $\mathrm{Info}(X,Y)$. Елементарний наслідок марковської властивості формулюється наступним чином. Якщо дана умовна незалежність $X$ та $Y$ за умови на $Q$, то буде

$$\mathrm{Info}(X,Y) \le \mathrm{Info}(X,Q)$$

та

$$\mathrm{Info}(X,Y) \le \mathrm{Info}(Y,Q).$$

Безумовну залежність також називають асоціацією. Наступне твердження випливає з властивості 2 та критерія d-сепарації.

*Властивість* 3. Якщо в МПГЗ-моделі всі ланцюги між змінними $X$ та $Y$ взаємно перетинаються на одній змінній

$Q$, то буде $\text{Info}(X,Y) \le \text{Info}(X,Q)$ й $\text{Info}(X,Y) \le \text{Info}(Q,Y)$.

З властивостей 2 та 3 випливає

*Властивість* 4. Якщо в МПГЗ-моделі для заданої пари асоційованих змінних ($X$, $Y$) немає жодної змінної $Q$ такої, що

$$\text{Info}(X,Q) \ge \text{Info}(X,Y)$$

та

$$\text{Info}(Q,Y) \ge \text{Info}(X,Y),$$

то не існує такої змінної, через яку проходили б усі ланцюги між $X$ та $Y$.

*Властивість* 5. Якщо в МПГЗ-моделі між змінними $X$ та $Y$ існує єдиний ланцюг, то для кожного ребра $Q — R$ на цьому ланцюзі буде

$$\text{Info}(X,Y) \le \text{Info}(Q,R).$$

**Визначення 1.** В МПГЗ-моделі асоціація між змінними $X$ та $Y$ називається *обманною нереберною*, якщо немає ребра $X — Y$ і якщо на кожному ланцюзі між $X$ та $Y$ існує щонайменше одне ребро $R — Z$ таке, що

$$\text{Info}(R,Z) < \text{Info}(X,Y).$$

**Визначення 2.** В МПГЗ-моделі асоціація між змінними $X$ та $Y$ називається *двійниковою*, якщо вона обманна нереберна і якщо не існує змінної $Q$, яка сепарує $X$ та $Y$.

Кожна обманна нереберна асоціація між $X$ та $Y$ або є двійниковою, або утворена конкатенацією двійникової асоціації з оршляхом. (В другому випадку всі ланцюги між $X$ та $Y$ проходять через певну спільну вершину (вершини)).

Зрозуміло, що задача відтворення моделі з даних стає найпростішою, коли відомо, що генеративна модель належить класу лісів або полі-лісів. Для лісів є відомий алгоритм квадратичної складності, запропонований C. Chow та C. Liu [9, 10]. Принцип алгоритму Chow&Liu – встановлення ребер у порядку зменшення міцності залежностей. Збагативши цей алгоритм засобами орієнтації ребер, можна отримати завершений алгоритм відтворення полі-лісів залежностей з даних. (Варіант такого алгоритму – 'SpaPolyTree' [6]). Моделі в класі лісів або полі-лісів можна відтворювати також методами, основаними на тестах незалежності. Для цього потрібні тести тільки нульового та першого рангів. (Прототипи таких алгоритмів – процедури 'Branch' («Гілка») та 'PolyForesyn' [6]).

Отже, є принциповий розрив між двома групами методів відтворення структури моделі з даних. Маємо елементарні спеціалізовані методи для однозв'язаних структур (якими є полі-ліси); і маємо універсальні комбінаторні методи для всього класу оАОГ-моделей. Втім, як показано в [11], можна надати універсальному методу здатності автоматично розпізнавати той факт, що генеративна модель є полілісом, і в такому разі наближатися до спеціалізованих методів за обчислювальними витратами.

Задача реконструкції моделі в класі монопотокових графів залежностей потребує нестандартних рішень. Відтворювати МПГЗ за допомогою універсальних методів – вкрай нерозумно. Водночас елементарні методи не працюють, бо в МПГЗ існують цикли, а сепаратори можуть бути великорозмірними, як і в довільних оАОГ.

Запропоновано кілька високоспеціалізованих алгоритмів реконструкції МПГЗ, які базуються на особливостях монопотокових графів. Перший алгоритм реконструкції МПГЗ, описаний в [12], логічно простий і потребує малої кількості тестів. Питання про присутність будь-якого ребра $X — Y$ вирішується за допомогою двох тестів. 1) тест безумовної незалежності; 2) тест умовної незалежності $X$ та $Y$, причому в умову включено всі змінні, окрім $X$ та $Y$. Це призводить до тестів умовної незалежності високого рангу, які є ненадійними. Отже, і сам алгоритм має низьку надійність. Уявимо, що змінних багато і що в генеративній моделі ребра $X — Y$ немає, але взаємозалежність змінних $X$ та $Y$ – сильна. Тоді (навіть за використання вибірки даних реалістичного обсягу) тест умовної незалежності може помилково не виявити незалежність (з огляду на складний формат). Тож буде поставлене помилкове ребро. Зазначимо, що сильна безумовна залежність може виникати, коли змінні пов'язані ланцюгом з двох ребер, або коли між змінними є двійникова асоціація.

Згодом з'ясовано, що для відтворення монопотокових моделей достатньо спиратися на тести нульового та першого рангу [13–15]. Методи серії «Генеалогія», зокрема, алгоритм «Генеалог-2», описаний в [16], та більш досконалий алгоритм «Генеалог-С» [6] – непридатні для стандартних обставин, бо потребують розпізнавання слабких безумовних залежностей, у тому числі тих, що утворені довгими ланцюгами [6, 17]. Метод CH1, описаний в [13], для кожної пари залежних змінних виконує всі можливі тести першого рангу. Це призводить до надто великої кількості тестів. Можна уникнути цього недоліку, якщо змінити головний принцип методу, а саме, перейти від пошуку сепаратора до використання провокації залежності [14, 15, 17]. Замість інтенсивного пошуку спростування ребра можна знаходити підтвердження присутності ребра.

Феномен провокованої залежності полягає у наступному [18]. Якщо змінні $X$ та $Z$ – взаємонезалежні, і вони разом впливають на змінну $Y$, то кондиціонування змінної $Y$ зазвичай породжує залежність між $X$ та $Z$. Тобто в результаті введення умови на змінну $Y$ виникає умовна залежність між $X$ та $Z$. Згідно результатів стохастичної симуляції дискретних моделей, сила (рівень) провокованої залежності в середньостатистичному сенсі приблизно дорівнює силі маргінальної (реберної) залежності між $Y$ та змінною $X$ або $Z$. На провоковану залежність припадає біля $1/3$ повної «родинної» інформації колізора [18]. Провокована залежність може колапсувати (не виникати) тільки в особливих випадках у дискретних моделях.

Позначимо як $\mathrm{Ind}(X,Y\,|\,\mathbf{S})$ факт умовної незалежності змінних $X$ та $Y$ за умови $\mathbf{S}$. Факт умовної залежності виражається як $\sim\mathrm{Ind}(X,Y\,|\,\mathbf{S})$. Безумовна незалежність та залежність позначаються відповідно $\mathrm{Ind}(X,Y)$ та $\sim\mathrm{Ind}(X,Y)$. Теоретично провокована залежність визначається як сполучення (паттерн) $\mathrm{Ind}(X,Z)\,\&$ $\&\sim\mathrm{Ind}(X,Z\,|\,Y)$. Оскільки в МПГЗ, згідно їх визначення, в кожному колізорі $X\rightarrow Y\leftarrow Z$ змінні $X$ та $Z$ – взаємонезалежні, то на базі кожного колізора має породжуватися провокована залежність.

В групі методів, що спираються на інструмент провокованих залежностей, найбільш ефективним можна вважати алгоритм «Proliferator-D» [6, 17]. За задумом, цей алгоритм оминає тести сепарації, але виконує тести провокації. Виведення структури МПГЗ-моделі виконується в два основні етапи. На першому етапі обчислюються всі парні асоціації та ідентифікуються всі провоковані залежності. Для уникнення зайвої роботи провокація залежності виконується тільки за наявності передумов – коли трійка змінних утворює квазі-колізорний паттерн. На другому етапі алгоритм конструює структуру моделі, спираючись, зокрема, на принцип Chow&Liu. При цьому (за достатніх підстав) відразу встановлюються орієнтовані ребра. Хоча алгоритм «Proliferator-D» переважає своїх попередників, ретельний аналіз показує, що він теж недосконалий й часто виконує необов'язкові обчислення. «Proliferator-D» є алгоритмом субкубічної складності. Але в багатьох нетривіальних випадках МПГЗ-модель можна відтворити, виконавши роботу лінійної складності. Потужним резервом підсилення методів відтворення структури моделі є правила мінімальної сепарації [6, 8]. Більшість цих правил виведено для всього класу оАОГ-моделей, але декілька створено спеціально для МПГЗ-моделей [6]. Правила мінімальної сепарації обґрунтовані в апараті графів, а для обґрунтування емпіричних еквівалентів («зліпків») цих правил необхідні відповідні форми припущення каузальної неоманливості [6]. Ті форми припущення неоманливості можна оцінити як недостатньо надійні, тож й аргументація ефективності правил мінімальної сепарації може сприйматися як непереконлива. В даній роботі пропонується більш переконливий шлях гарантування надійності відтворення моделі з даних. Спочатку формулюються найбільш надійні припущення про емпіричні (статистично-ймовірнісні) прояви залежностей, а потім на базі цих припущень будуються емпіричні резолюції та засоби для методів відтворення МПГЗ.

## Базові припущення та інструментарій

Статистична залежність може бути слабкою. У ході аналізу реалістичних даних статистичний тест може відрізнити залежність від незалежності тільки якщо емпірична залежність досить міцна (велика). Метод відтворення моделі з даних може спиратися тільки на статистично значущі залежності. Позначатимемо предикатом $\mathrm{Dep}(X,Y)$ факт *значущої безумовної залежності* між $X$ та $Y$. (Домовимося, що предикат $\mathrm{Dep}(X,X)$ завжди істинний).

*Квазі-колізорна трійка* змінних – це паттерн $\mathrm{Dep}(X,Y)\,\&\,\mathrm{Dep}(Z,Y)\,\&\,\mathrm{Ind}(X,Z)$.

*Емпірична провокована залежність* – це сполучення $\sim\mathrm{Dep}(X,Z)$ & $\mathrm{Dep}(X,Z\,|\,Y)$.

Обчислювальна складність алгоритму ‘Proliferator-D’ в основному визначається кількістю тестів провокації залежності, а значить, кількістю квазі-колізорних трійок. На рис. 1 показано приклади структур МПГЗ, які породжують велику кількість квазі-колізорних трійок. В першій структурі на рис. 1, *а* теоретично маємо $(n-1)(n-2)/2$ квазі-колізорних трійок. В другій структурі на рис. 1, *б* ця кількість менша. В третій структурі на рис. 1, *в* кількість квазі-колізорних трійок дорівнює $(n-2)(n-3)$. І хоча теоретично кількість квазі-колізорних трійок не може бути вище за квадратичну від кількості змінних, у багатьох випадках ‘Proliferator-D’ буде виконувати зайву роботу.

Дійсно, модель, показана на рис. 1, *а*, – дуже простий випадок полі-лісу й може бути тривіально відтворена на основі тестів нульового рангу та логіки. Інші моделі, показані на цьому рисунку теж реконструюються просто.

Бажано базувати метод відтворення моделі з даних на надійних припущеннях про емпіричні прояви залежностей [1, 3, 6]. Мабуть, найбільш надійним припущенням, яке можна сформулювати, є припущення *безумовної (маргінальної) реберної неоманливості* (0-РНО): будь-які дві змінні, поєднані ребром, є безумовно залежні. Формально: якщо існує $(X — Y)$, то буде $\mathrm{Dep}(X,Y)$.

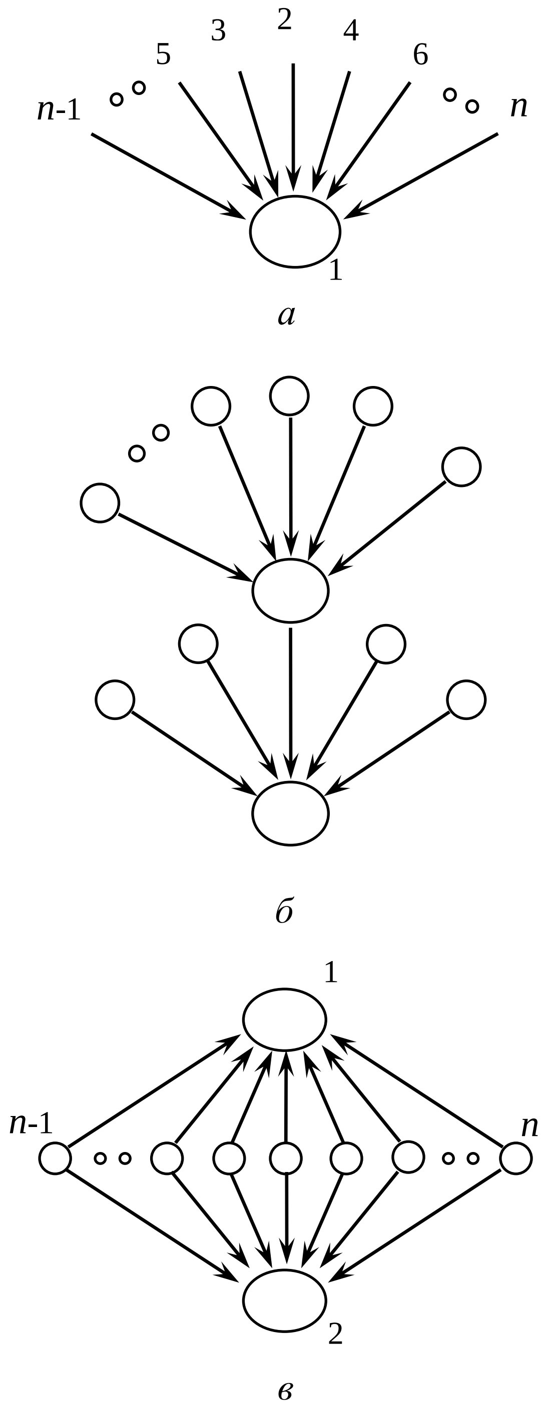


Рис. 1. Приклади МПГЗ:
*а, б* – полі-ліси; *в* – МПГЗ з циклами

Для МПГЗ-моделей це припущення – більш надійне, ніж для загального випадку оАОГ-моделей, тому що в оАОГ можуть бути одночасно присутні ребро $X - Y$ і ще якісь ланцюги між вершинами $X$ та $Y$. (Тоді може статися взаємна компенсація залежностей). В МПГЗ таке виключено.

Припущення 0-РНО є необхідним та може виявитися недостатнім для ідентифікації всіх ребер. Тоді знадобиться залучити дещо складніший варіант. Припущення *реберної неоманливості першого рангу* (1-РНО): будь-які дві змінні, поєднані ребром, є умовно залежні, з умовою на довільну третю змінну.

Можна передбачити, що надійність припущення 1-РНО також буде вище для МПГЗ-моделей, ніж для загального випадку оАОГ-моделей, бо в оАОГ існує більше можливостей для контр-прикладів. Ризик порушення 1-РНО є досить високий, зокрема, у трикутнику $X \to Y \to$ $\to Z \leftarrow X$, де кондиціонування змінної $Z$ створює провоковану залежність, яка може анігілювати з залежністю, створеною ребром $X - Y$.

Для обґрунтування здатності ідентифікувати спрямування (орієнтацію) ребер потрібні додаткові припущення. Для класу МПГЗ-моделей таким може бути одне з двох наступних припущень.

Припущення *безумовної каузальної неоманливості двох-реберних ланцюгів* (0-Л2НО): якщо в моделі між змінними $X$ та $Z$ присутній якийсь ланцюг довжиною у два ребра, то буде безумовна залежність $\mathrm{Dep}(X,Z)$. Іншими словами, якщо існує ланцюг $X \to Y \to Z$ або $X \leftarrow Y \to Z$, то він забезпечує значущу транзитну залежність між $X$ та $Z$.

Припущення *елементарної провокації залежності на колізорі* (ЕПЗК): якщо в моделі присутній колізор $X \to Y \leftarrow Z$, то буде умовна залежність $\mathrm{Dep}(X,Z \mid Y)$. Для МПГЗ-моделей припущення ЕПЗК – більш надійне, ніж для загального випадку оАОГ-моделей, тому що для колізора $X \to Y \leftarrow Z$ в складі МПГЗ-моделі змінні $X$ та $Z$ обов'язково взаємонезалежні (а в оАОГ – не обов'язково).

(Зауважимо, що для орієнтації ребер в оАОГ-моделях цих припущень недостатньо).

Доцільно також зафіксувати варіант припущення, спрощений відносно 0-Л2НО. Припущення *безумовної каузальної неоманливості елементарних двохреберних ланцюгів* (0-ЕЛ2НО): якщо в моделі між змінними $X$ та $Z$ присутній один й тільки один ланцюг довжиною у два ребра, то буде безумовна залежність $\mathrm{Dep}(X,Z)$. Можна очікувати, що це припущення – більш надійне, ніж 0-Л2НО, оскільки воно уникає ситуацій з певними ускладненнями.

Варто зазначити, що усі методи відтворення моделі з даних постулюють (за замовчуванням) наступне припущення: будь-які дві змінні, не поєднані жодним ланцюгом (ребром), тестуються як безумовно незалежні. (Звісно, статистичні тести мають ризики помилок, оцінювані значенням p-value).

**Визначення 3.** *Набір близьких до змінної* $X$ визначається як $\mathrm{Clo}(X) =$ $= \{Y \mid \mathrm{Dep}(X,Y)\}$, тобто як множина всіх тих змінних, які безумовно залежать від $X$ (включаючи саму $X$).

Тривіально $\mathrm{Clo}(X) \supseteq \{X\}$. Згідно припущення 0-РНО, якщо існує $X - Y$, то $Y \in \mathrm{Clo}(X)$. Якщо вершина $X$ дотична до $k$ ребер, то $|\mathrm{Clo}(X)| \geq (k+1)$.

Висока ефективність згаданих вище алгоритмів відтворення лісів та полілісів завдячує принципу Chow&Liu (який повторює відомий алгоритм Крускала). Коректність принципу Chow&Liu для полі-лісів забезпечується відсутністю циклів у структурі. На жаль, цей принцип не є коректним для МПГЗ-моделей в загальному випадку (бо в МПГЗ може бути кілька паралельних ланцюгів). Встановлення ребер в МПГЗ-моделі за принципом Chow&Liu призводить до помилок у певних ситуаціях [6, 7]. Виконання алгоритму Chow&Liu призведе до встановлення ребра на місці двійникової чи обманної нереберної асоціації (а певне автентичне ребро буде втрачене) [6, 7, 17]. Тому для використання принципу Chow&Liu необхідно озброїти метод відтворення МПГЗ-моделей засобами розпізнавання двійникових асоціацій. Далі буде показано, як можна ідентифікувати багато ребер у МПГЗ-моделях на основі систематичного аналізу сукупності безумовних залежностей. Коректність пропонованих засобів ґрунтується на найбільш надійному припущенні – безумовної (маргінальної) реберної неоманливості (0-РНО). Для коректності резолюцій, які вдаються до принципу Chow&Liu, потрібне також припущення *монотонності залежності на марковському ланцюзі*. Формулювання: якщо маємо $\mathrm{Dep}(X,Y)$, $\mathrm{Dep}(Y,Z)$ та $\mathrm{Ind}(X,Z\,|\,Y)$, то чинне

$$\mathrm{Info}(X,Z)\le \min\{\mathrm{Info}(X,Y),\mathrm{Info}(Y,Z)\}.$$

Останнє припущення та припущення 0-РНО не поглинають один одного. (Властивість 3 треба розуміти в асимптотичному сенсі).

## Емпіричні резолюції ідентифікації ребер в МПГЗ-моделях

З визначення двійникової асоціації та властивості 4 випливає наступний факт.

*Факт* 1. Якщо для МПГЗ-моделі асоціація між змінними $X$ та $Y$ – «двійникова», то $|\mathrm{Clo}(X)|\ge 4$ та $|\mathrm{Clo}(Y)|\ge 4$.

З властивостей 1 та 5 випливає наступний факт.

*Факт* 2. Якщо в МПГЗ-моделі асоціація між змінними $X$ та $Y$ – двійникова, то між вершинами $X$ та $Y$ існує не менше двох ланцюгів $\lambda$ та $\mu$ таких, що кожна некінцева змінні на ланцюзі $\lambda$ незалежна від кожної некінцевої змінної на ланцюзі $\mu$.

*Наслідок*. Якщо для МПГЗ-моделі маємо $Y\in\mathrm{Clo}(X)$ і серед змінних, що входять до $\mathrm{Clo}(X)$, або серед змінних, що входять до $\mathrm{Clo}(Y)$, немає принаймні двох взаємонезалежних, то асоціація між $X$ та $Y$ не може бути двійниковою.

Якщо спиратися виключно на припущення 0-РНО, то серед усіх виведених [6, 8] правил мінімальної сепарації можна транслювати в емпіричну форму тільки одне. Це правило єдиного родича, або безальтернативного ребра [6] (воно чинне для всього класу оАОГ-моделей). В емпіричній формі воно формулюється наступним чином.

*Резолюція єдиного близького*. Якщо в оАОГ-моделі маємо $\mathrm{Clo}(X)=\{X,Y\}$, то існує ребро $X — Y$.

Для класу МПГЗ-моделей є чинним більш сильний результат.

*Резолюція перемички* (відсутність спільної суміжної). Нехай для МПГЗ-моделі маємо $Y\in\mathrm{Clo}(X)$ та $\left(\mathrm{Clo}(X)\setminus\{X\}\right)\cap\left(\mathrm{Clo}(Y)\setminus\{Y\}\right)=\varnothing$. Тоді, якщо $|\mathrm{Clo}(X)|\le 3$ або $|\mathrm{Clo}(Y)|\le 3$, то в моделі присутнє ребро $X — Y$. Інакше (тобто якщо $|\mathrm{Clo}(X)|>3$ та $|\mathrm{Clo}(Y)|>3$) є дві можливі альтернативи:

а) в моделі присутнє ребро $X — Y$, або

б) всі шляхи між $X$ та $Y$ є ланцюгами, з тим, що кожний ланцюг між $X$ та $Y$ має не менше трьох ребер, і існує не менше двох таких ланцюгів.

Ясно, що альтернатива «б» імплікує існування двійникової асоціації.

З умови «$Y\in\mathrm{Clo}(X)$ та $\left(\mathrm{Clo}(X)\setminus\{X\}\right)\cap\left(\mathrm{Clo}(Y)\setminus\{Y\}\right)=\varnothing$» випливає (властивість 3 та 4), що не існує такої змінної $Z$, що всі ланцюги між $X$ та $Y$ перетинаються на $Z$. Якщо немає ребра $X — Y$, то між $X$ та $Y$ існує не менше двох ланцюгів (властивість 5). Якби якийсь ланцюг між $X$ та $Y$ мав два ребра $X-Q-Y$, то згідно припущення 0-РНО було б $Q\in\mathrm{Clo}(X)$ та $Q\in\mathrm{Clo}(Y)$.

*Резолюція кінцевого двох-реберного ланцюга*. Нехай для МПГЗ-моделі маємо $\mathrm{Clo}(X)=\mathrm{Clo}(Y)=\{X,Y,Z\}$. Тоді змінні $X,Y,Z$ поєднані в генеративній моделі двома ребрами, які утворюють ланцюг й ідентифікуються згідно принципу Chow&Liu.

*Доведення*. Якби якісь дві змінні з трьох $X,Y,Z$ були поєднані через ланцюг, який проходить через якусь четверту змінну $W$, то було б $W \in \mathrm{Clo}(X)$ або $W \in \mathrm{Clo}(Y)$, що суперечить умовам. Якби змінні $X, Y, Z$ були поєднані як колізор, то дві з них були б взаємонезалежні, що суперечить умовам. Оскільки трикутник ребер в МПГЗ неможливий, отримуємо бажаний висновок. (Вказані два ребра ідентифікуються згідно властивості 5).

Зауважимо, що в такій конфігурації одна з трьох змінних (в даному разі $Z$) може мати «додаткові» близькі змінні, тобто можливе $|\mathrm{Clo}(Z)| > 3$.

Об'єднуючи властивість 3 та наслідок з факту 2, отримуємо результат.

*Резолюція виключення нереберної асоціації*. Нехай для МПГЗ-моделі маємо $Y \in \mathrm{Clo}(X)$, і серед змінних, що входять до $\mathrm{Clo}(X) \setminus \{X, Y\}$, або серед змінних, що входять до $\mathrm{Clo}(Y) \setminus \{X, Y\}$, немає жодної пари взаємонезалежних. Тоді, якщо серед змінних, що входять до $\mathrm{Clo}(Y) \setminus \{X, Y\}$, немає хоча б одної змінної $Q$, такої, що $\mathrm{Info}(X;Q) \geq \mathrm{Info}(X;Y)$ та $\mathrm{Info}(Q,Y) \geq \geq \mathrm{Info}(X,Y)$, то присутнє ребро $X — Y$.

Об'єднавши резолюцію перемички та резолюцію виключення обманної нереберної асоціації, отримаємо більш сильний результат.

*Резолюція перемички* (об'єднана). Нехай для МПГЗ-моделі маємо $Y \in \mathrm{Clo}(X)$ і виконується $(\mathrm{Clo}(X) \setminus \{X\}) \cap \cap (\mathrm{Clo}(Y) \setminus \{Y\}) = \varnothing$. Для ідентифікації присутності ребра $X — Y$ в моделі достатнім є виконання принаймні однієї з наступних умов:

а) маємо $|\mathrm{Clo}(Y)| \leq 3$ або $|\mathrm{Clo}(Y)| \leq 3$;

б) серед змінних, що входять до $\mathrm{Clo}(X)$ немає жодних двох взаємонезалежних;

в) серед змінних, що входять до $\mathrm{Clo}(Y)$, немає жодних двох взаємонезалежних.

*Ідея доведення*. Якби асоціація між $X$ та $Y$ забезпечувалася єдиним ланцюгом або була обманною нереберною (але не двійниковою) асоціацією, то існував би 1-сепаратор для пари ( $X$ , $Y$ ), а тоді, не виконувалася б передумова правила. Додаткові умови («а», «б, «в») резолюції виключають існування двійникової асоціації.

*Резолюція слабкої пари ребер*. Якщо для МПГЗ-моделі маємо $\mathrm{Clo}(Y) = \{X, Y, Z\}$ й $Z \notin \mathrm{Clo}(X)$, то в моделі присутня пара ребер $X — Y — Z$.

*Доведення*. Припустимо від протилежного, що ребра $X — Y$ немає. Тоді, з огляду на умову $\mathrm{Clo}(Y) = \{X, Y, Z\}$, суміжною до змінної $Y$ буде тільки одна змінна $Z$. Отже, змінні $X$ та $Y$ мають поєднуватися ланцюгами (ланцюгом), які проходять через ребро $Z — Y$. Тоді, згідно властивості 3, буде $\mathrm{Info}(X, Y) \leq \mathrm{Info}(X, Z)$. А це разом з фактом $Y \in \mathrm{Clo}(X)$ тягне $Z \in \mathrm{Clo}(X)$, що суперечить умовам. З огляду на симетрію, доведення для другого ребра аналогічне.

Резолюцію слабкої пари ребер можна узагальнити.

*Резолюція центральної вершини* (*єдиного вузла*). Нехай для МПГЗ-моделі всі змінні, що входять до складу множини $\mathrm{Clo}(Q) \setminus \{Q\}$, є взаємонезалежними. Тоді, якщо $|\mathrm{Clo}(Q)| = 3$ або серед змінних множини $\mathrm{Clo}(Q) \setminus \{Q\}$ немає жодної такої $Z$, що $|\mathrm{Clo}(Z)| \geq 4$, то в моделі присутні ребра $Q — X$ для всіх змінних $X \in \mathrm{Clo}(Q) \setminus \{Q\}$.

*Доведення*. У випадку $|\mathrm{Clo}(Q)| = 3$ отримуємо резолюцію слабкої пари ребер. Для іншого випадку припустимо від протилежного, що є змінна $W$, $W \in \mathrm{Clo}(Q) \setminus \{Q\}$ несуміжна до $Q$. Тоді асоціація між $Q$ та $W$ була б двійникова (інші варіанти нереберної асоціації відпадають, бо потребують існування деякої спільної близької до $Q$ та $W$ змінної, що суперечить умовам згідно властивості 3). Але для можливості вказаної двійникової асоціації (згідно факту 1) необхідно $|\mathrm{Clo}(W)| \geq 4$. Суперечить умовам.

*Резолюція крайнього трикутника асоціацій*. Нехай для МПГЗ-моделі маємо

$\text{Clo}(X)=\{X,Y,Z\}$, $\text{Clo}(Y)\supset\{X,Y,Z\}$ та $\text{Clo}(Z)\supset\{X,Y,Z\}$ (а асоціація між $Y$ та $Z$ може задовольняти необхідним вимогам обманної нереберної асоціації). Тоді чинне:

а) якщо $\text{Info}(X;Y)>\text{Info}(X;Z)>$ $>\text{Info}(Y;Z)$, то в моделі присутній ланцюг $Y — X — Z$;

б) якщо $\text{Info}(X;Y)>\text{Info}(Y;Z)>$ $>\text{Info}(X;Z)$ або $\text{Info}(Y;Z)>\text{Info}(X;Y)>$ $>\text{Info}(X;Z)$, то в моделі присутнє ребро $X — Y$, а питання щодо другого ребра ланцюга залишається відкритим.

Нагадаємо, що *кліка* асоційованих змінних – це така множина змінних $\mathbf{C}$, що для кожних двох її елементів $X,Y\in\mathbf{C}$ чинне $Y\in\text{Clo}(X)$. *Ізольована кліка* – це така множина змінних $\mathbf{C}$, що для кожної $X\in\mathbf{C}$ чинне $\text{Clo}(X)=\mathbf{C}$.

*Резолюція ізольованої кліки*. Якщо для МПГЗ-моделі деяка підмножина змінних $\mathbf{C}$ утворює ізольовану кліку, то в генеративній моделі підмножина змінних $\mathbf{C}$ поєднана деревом ребер, і це дерево повністю відтворюється за принципом Chow&Liu.

Дійсно, якби якісь три змінні утворювали б колізор, то була б пара взаємно незалежних змінних. МПГЗ без колізорів – це дерево (ліс).

Можна розширити резолюцію ізольованої кліки (і охопити резолюцію кінцевого двох-реберного ланцюга як спеціальний випадок). Введемо допоміжні поняття.

*Граф парних асоціацій* (ГПА) для даних (генерованих з моделі з набором змінних $\mathbf{V}$) – це неорієнтований граф з вершинами $\mathbf{V}$, в якому встановлене ребро $X — Y$ для кожної пари змінних ($X$, $Y$), для якої чинне $\text{Dep}(X,Y)$, й немає ребра $Q — Z$ для кожної пари змінних ($Q$, $Z$), якщо чинне $\sim\text{Dep}(X,Y)$.

*Факт 3*. Якщо в МПГЗ-моделі асоціація між змінними $X$ та $Y$ – обманна нереберна, то в ГПА існує цикл, який проходить через $X$ та $Y$ і який не входить у склад жодної кліки.

Нехай дві змінні $X$ та $Y$ поєднані в МПГЗ двома чи більше ланцюгами і входять до деякої компоненти ГПА $\mathbf{K}$. Тоді ясно, що за припущення 0-РНО компонента $\mathbf{K}$ не може бути клікою.

*Максимальна кліка змінних* – це така кліка $\mathbf{C}$, що не існує жодної змінної $X\notin\mathbf{C}$ такої, що $\mathbf{C}\cup\{X\}$ теж є клікою змінних.

*Резолюція кліки асоціацій у деревовидному оточенні*. Нехай для МПГЗ-моделі маємо в ГПА максимальну кліку $\mathbf{C}$. Далі, нехай для кожної пари змінних $X,Y$ з кліки $\mathbf{C}$ чинне наступне: між змінними $X$ та $Y$ немає жодного шляху в ГПА поза клікою $\mathbf{C}$. Тоді в генеративній моделі всі змінні кліки $\mathbf{C}$ з'єднані деревом, яке можна ідентифікувати за принципом Chow&Liu.

Умови резолюції формалізуються так: $\mathbf{C}=\{X,Y,Z,..\}$; для всіх $X\in\mathbf{C}$ маємо $\text{Clo}(X)\supseteq\mathbf{C}$; для будь-якої пари $X,Y\in\mathbf{C}$ між $X$ та $Y$ *не існує* жодної послідовності $Z_1\in\text{Clo}(X)$, $Z_2\in\text{Clo}(Z_1)$, ... ... $,Z_k\in\text{Clo}(Z_{k-1})$, …, $Y\in\text{Clo}(Z_n)$, де $Z_k\notin\mathbf{C}$.

Дійсно, якби між якимись змінними $X,Y\in\mathbf{C}$ в генеративній моделі існувало б два чи більше ланцюгів, то, згідно властивості 1, ці ланцюги не входили б в жодну кліку. Тоді був би шлях в ГПА між $X$ та $Y$ поза клікою $\mathbf{C}$. Отже, в генеративній моделі в рамках множини $\mathbf{C}$ немає циклів, тобто всі змінні кліки $\mathbf{C}$ поєднані в генеративній моделі деревом. (Вимога, щоб кліка $\mathbf{C}$ була максимальною, є технічною. Якщо кліка не максимальна, неможливо виконати решту вимог резолюції).

Можна зауважити, що більшість наведених резолюцій автоматично виконується алгоритмом Chow&Liu. Проте ці резолюції гарантують коректну реконструкцію ребер (а алгоритм Chow&Liu – ні).

Пропоновані резолюції значно прискорюють реконструкцію моделей, структура яких показана на рис. 1. Для випадку рис. 1, а всі ребра ідентифікуються резолюцією єдиного близького. Ця сама резолюція дозволяє відтворити ребра «нижнього ярусу» в структурі рис. 1, б (крім одного ребра). Всі ребра «верхнього ярусу» цієї структури відтворюються за допомогою резолюції крайнього трикутника асоціацій. Припустимо, що в моделі із структурою рис. 1, в асоціація між змінними *X*1 та *X*2 – значуща. Досить, щоб для однієї трійки змінних виконалися умови п. «а» резолюції крайнього трикутника асоціацій, і тоді всі ребра швидко ідентифікуються.

## Експериментальне випробування

Продемонструємо ефективність роботи запропонованих резолюцій на прикладі. Структура генеративної моделі (в класі МПГЗ) для експерименту показана на рис. 2. Модель має 32 змінні та 37 ребер. Всі змінні – тризначні. Для цієї структури було генеровано (стохастичним механізмом) три варіанти параметризації, і для кожної параметризації генеровано кілька вибірок даних обсягом 500, 1000, 2000, 3000 та 5000 записів. Тестування безумовної незалежності виконувалось за спрощеною схемою – як порівняння взаємної інформації з пороговим значенням. Попередній експеримент проведено для перевірки надійності тестів та припущення 0-РНО. Для цього фіксувалися помилкові рішення двох типів:

1) ідентифікація залежності змінних, які згідно генеративної моделі є взаємонезалежними;

2) ідентифікація незалежності змінних, поєднаних ребром в генеративній моделі. Результати цього експерименту наведено в табл. 1.

Підсумуємо результати експерименту. В середньому для вибірки розміром 500 записів і величини порогу 0,01 ризик втрати ребра досяг 13 %, а кількість помилкових залежностей дорівнює 11. Аналогічні показники для величини порогу 0,02 становлять 28 % та 0,17 частки залежності. Коли вибірка має 1000 записів або більше, не з'явилося жодної помилкової залежності. Для вибірки розміром

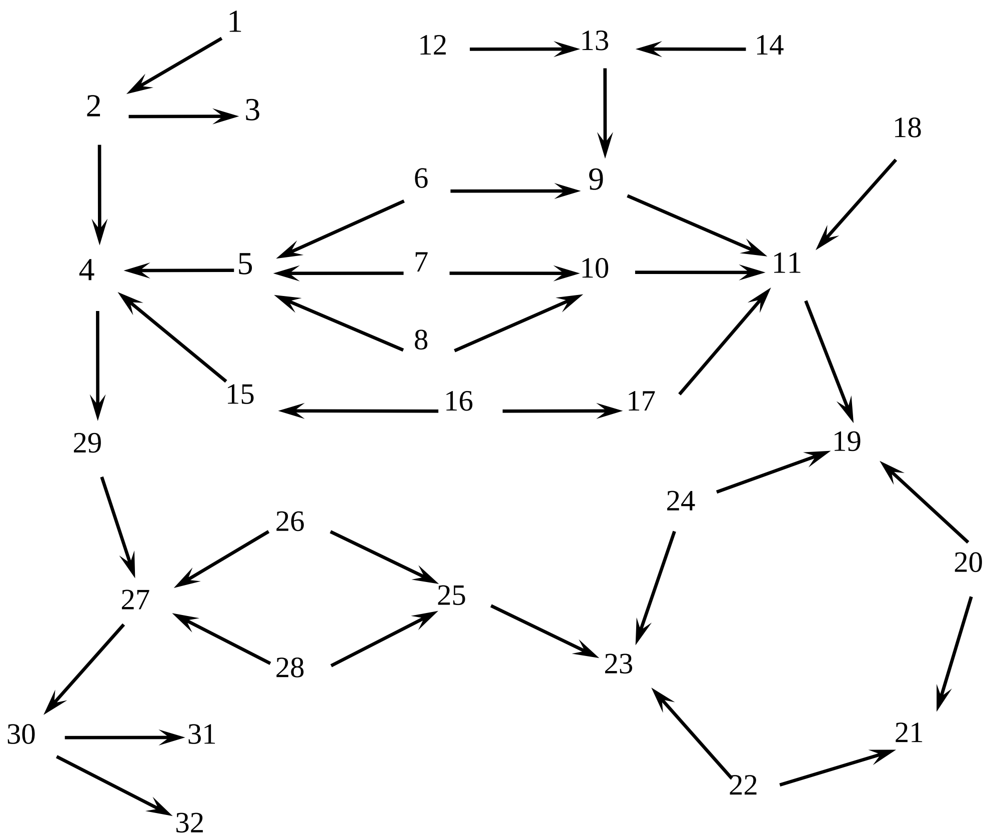


Рис. 2. Структура МПГЗ ('MF32-37') для експерименту

Таблиця 1

Емпіричні залежності для ребер моделі

| Варіант парамет-ризації | Розмір вибірки | Поріг значущості = 0,01 | | Поріг значущості = 0,02 | |
|---|---|---|---|---|---|
| | | Кількість втрачених ребер | Кількість помилкових залежностей | Кількість втрачених ребер | Кількість помилкових залежностей |
| 1-й | 500 | 4 | 12 | 8 | 0 |
| | 500 | 5 | 13 | 8 | 0 |
| | 500 | 3 | 10 | 8 | 0 |
| | 500 | 4 | 9 | 10 | 0 |
| | 1000 | 4 | 0 | 7 | 0 |
| | 1000 | 3 | 0 | 9 | 0 |
| | 2000 | 3 | 0 | 8 | 0 |
| | 3000 | 1 | 0 | 9 | 0 |
| | 5000 | 1 | 0 | 7 | 0 |
| 2-й | 500 | 3 | 8 | 6 | 0 |
| | 500 | 4 | 2 | 9 | 0 |
| | 500 | 5 | 14 | 10 | 0 |
| | 500 | 5 | 10 | 9 | 0 |
| | 1000 | 4 | 0 | 10 | 0 |
| | 1000 | 5 | 0 | 11 | 0 |
| | 2000 | 4 | 0 | 10 | 0 |
| | 3000 | 3 | 0 | 6 | 0 |
| | 5000 | 2 | 0 | 6 | 0 |
| 3-й | 500 | 6 | 11 | 13 | 0 |
| | 500 | 7 | 18 | 15 | 0 |
| | 500 | 6 | 24 | 15 | 1 |
| | 500 | 8 | 6 | 15 | 1 |
| | 1000 | 7 | 0 | 13 | 0 |
| | 1000 | 7 | 0 | 13 | 0 |
| | 2000 | 7 | 0 | 15 | 0 |
| | 3000 | 7 | 0 | 14 | 0 |
| | 5000 | 6 | 0 | 13 | 0 |

1000 записів ризик втрати ребра становить 13 % (для порогу 0,01) та 28 % (для порогу 0,02).

Для вибірок розміром від 2000 до 5000 записів середній ризик втрати ребра склав 10 % (для порогу 0,01) та 26 % (для порогу 0,02). Той факт, що при зростанні розміру вибірки показники втрат ребер зменшуються слабо, вказує на значну кількість майже маскованих та деградованих ребер.

Відомо, що завдяки співвідношенням параметрів моделі певне ребро генеративної моделі може перетворитися на масковане [18] або навіть деградувати (тобто абсолютно втратити вплив). На користь таких ситуацій у нашому експерименті свідчить той факт, що для різних вибірок даних повторюються одні й ті самі пропуски ребер. Зокрема, для 2-го варіанта параметризації в результатах виведення з усіх дев’яти вибірок (з обома значеннями порогу значущості) пропущено ребра 7—5 та 24—19. А для третього варіанту параметризації в результатах виведення з усіх дев’яти вибірок даних (коли використано поріг 0,02) пропущено вісім ребер. Не всі зафіксовані пропуски ребра можна вважати справжніми помилками. Пропуск деградованого ребра не є помилкою. А що стосується синдрому маскованих ребер, то відомі методи відтворення моделей (основані на незалежності) також не захищені від нього.

Головний експеримент випробував роботу запропонованих резолюцій. Для цього експерименту використано вибірки розміром 2000 та поріг значущості 0,01. Оскільки деякі вищенаведені резолюції взаємно поглинаються, було реалізовано тільки чотири резолюції. Кожна з обраних резолюцій – єдиного близького, перемички (об’єднана), центральної вершини (єдиного вузла) та кліки асоціацій у деревовидному оточенні, – застосовувалась автономно, як на початку процесу реконструкції. Тривалість застосування чотирьох резолюцій до вибірки даних складала біля третини секунди в середовищі MATLAB. (Потенційно складною задачею міг стати пошук максимальних клік, але вдалося уникнути важкості відомої абстрактної задачі через те, що процедура відразу шукає кліки, які задовольняють всім вимогам резолюції).

Результати головного експерименту наведено в табл. 2. Зафіксовані в таблиці помилки чотирьох резолюцій насправді відображають одну помилку, яка сталася внаслідок того, що транзитна залежність двох-реберного ланцюга перевищила залежність одного з ребер того ланцюга. Якщо розглянути 1-й варіант параметризації, то чотири резолюції у підсумку відтворили 24 ребра (без помилок), що становить дві третини ребер моделі. Для 2-го варіанта параметризації ці резолюції відтворили 18 ребер без помилок. Для 3-го варіанта параметризації правильно відтворено 26 ребер та вставлено одне помилкове ребро.

Таблиця 2

Ефективність застосування емпіричних резолюцій відтворення ребер

| Резолюція | Варіант параметризації | Правильно відтворених ребер | Помилкових ребер |
|---|---|---|---|
| Р. єдиного близького | 1 | 16 | 0 |
| | 2 | 16 | 0 |
| | 3 | 19 | 1 |
| Р. перемички (об’єднана) | 1 | 19 | 0 |
| | 2 | 10 | 0 |
| | 3 | 19 | 1 |
| Р. центральної вершини (єдиного вузла) | 1 | 16 | 0 |
| | 2 | 7 | 0 |
| | 3 | 18 | 1 |
| Р. кліки асоціацій у деревовидному оточенні | 1 | 6 | 0 |
| | 2 | 5 | 0 |
| | 3 | 16 | 1 |

## Висновки

Показано, яким чином (апріорі знаючи тільки приналежність моделі до класу монопотокових структур) можна ідентифікувати більшість ймовірнісних зв'язків (безпосередніх залежностей), спираючись виключно на безумовні залежності двох змінних. Ідентифікація ребер (зв'язків) виконується емпіричними резолюціями, які строго обґрунтовані на базі припущення безумовної (маргінальної) реберної не-оманливості. Вже за своєю конструкцією це припущення є одним з найнадійнішим серед простих версій припущень каузальної неоманливості, відомих у галузі каузального виведення. Висока емпірична надійність цього припущення та побудованих на ньому резолюцій підтверджується результатами експерименту. Плануємо розвинути і перенести ідеї та принципи на загальний клас моделей із структурою ациклонних орграфів (байєсівських мереж).


1. *Spirtes P*. Introduction to causal inference. Journal of Machine Learning Research. 2010. Vol. 11, P. 1643−1662.
2. *Kalisch M*., *Bühlmann P*. Causal structure learning and inference: a selective review. Quality Technology & Quantitative Management. 2014. Vol. 11, N 1. P. 3–21.
3. *The TETRAD* Project: Constraint based aids to causal model specification / R. Scheines, P. Spirtes, C. Glymour et al. Multivariate Behavioral Research. 1998. Vol. 33, N 1. P. 65–118.
4. *Chickering D., Heckerman D., Meek C.* Large-sample learning of Bayesian networks is NP-hard. Journal of Machine Learning Research. 2004. Vol. 5. – P. 1287–1330.
5. *Балабанов О.С.* Відтворення каузальних мереж на основі аналізу марковських властивостей. Математичні машини та системи. 2016. № 1. С. 16–26.
6. *Балабанов О.С.* Каузальні мережі: аналіз, синтез та виведення з статистичних даних: дис. … доктора фіз.-мат. наук (01.05.01–теоретичні основи інформатики та кібернетики) / Балабанов Олександр Степанович. – К.: Ін-т кібернетики ім. В.М. Глушкова НАНУ, 2014. 305 с.
7. *Балабанов О.С.* Системи ймовірнісних залежностей: графові та статистичні властивості. Математичні машини та системи. 2009. № 3. С. 80–97.
8. *Балабанов О.С.* Правила підбору сепараторів в баєсівських мережах. Проблеми програмування. 2007. № 4. С. 33–43.
9. *Chow C.K., Liu C.N.* Approximating discrete probability distributions with dependence trees. IEEE trans. on Information Theory. 1968. Vol. 14, N 3. P.462–467.
10. *Балабанов О.С.* Індуктивне відтворення деревовидних структур систем залежностей. Проблемы программирования. 2001. № 1–2. С. 95–108.
11. *Балабанов О.С.* Прискорення алгоритмів відтворення байєсівських мереж. Адаптація до структур без циклів. Проблеми програмування. 2011. № 1. С. 63–69.
12. *Geiger D., Paz A., Pearl J.* Learning simple causal structures. Internat. Journal of Intelligent Systems. 1993. Vol. 8, N 2. P. 231–247.
13. *de Campos L.M., Huete J.F.* On the use of independence relationships for learning simplified belief networks. Intern. Journal of Intelligent Systems. 1997. Vol. 12. Issue 7. P. 495–522.
14. *Балабанов О.С.* Ефективний метод виявлення структур залежностей в статистичних даних. Проблемы программирования. 2004. № 2–3. С. 312–319.
15. *Балабанов А.С.* К выводу структур моделей вероятностных зависимостей из статистических данных. Кибернетика и системный анализ. 2005. № 5. С. 19–31.
16. *Балабанов А.С.* Индуктивный метод восстановления монопотоковых вероятностных графовых моделей зависимостей. Проблемы управления и информатики. 2003. № 5. С.75–84.
17. *Балабанов А.С.* Реконструкция модели вероятностных зависимостей по статистическим данным. Инструментарий и алгоритм. Проблемы управления и информатики. 2009. № 6. С. 90–103.
18. *Балабанов А.С.* Индуцированная зависимость, взаимодействие факторов и дискриминация каузальных структур. Кибернетика и системный анализ. 2016. № 1. С. 10–22.

## References


1. Spirtes P. (2010). Introduction to causal inference. Journal of Machine Learning Research. 11, 1643−1662.
2. Kalisch M., Bühlmann P. (2014). Causal structure learning and inference: a selective review. Quality Technology & Quantitative Management. 11 (1), 3–21.
3. Scheines R., Spirtes P., Glymour C. et al. (1998). The TETRAD Project: Constraint based aids to causal model specification. Multivariate Behavioral Research. 33 (1), 65–118.
4. Chickering D., Heckerman D., Meek C. (2004). Large-sample learning of Bayesian networks is NP-hard. Journal of Machine Learning Research. 5, 1287–1330.
5. Balabanov O.S. (2016). Vidtvorennya kauzalnych merezh na osnovi analizu markovskich vlastyvostej [Reconstruction of causal networks via analysis of Markov properties]. Mathematical Machines and Systems. (1), 16–26. [In Ukrainian].
6. Balabanov O.S. (2014). ‘Causal nets: analysis, synthesis and inference from statistical data’, Doctor of math. sciences thesis, V.M. Glushkov Institute of Cybernetics, Kyiv, Ukraine. [In Ukrainian].
7. Balabanov O.S. (2009) Probabilistic dependency models: graphical and statistical properties. Mathematical Machines and Systems. (3), 80–97. [In Ukrainian].
8. Balabanov O.S. (2007). Rules for picking up separators in Bayesian networks. Problems in programming. (4), 33–43. [In Ukrainian].
9. Chow C.K., Liu C.N. (1968). Approximating discrete probability distributions with dependence trees. IEEE trans. on Information Theory. 14 (3), 462–467.
10. Balabanov O.S. (2001). Inductive recovery of structures of dependency trees. Problems in programming. (1–2), 95–108. [In Ukrainian].
11. Balabanov O.S. (2011). Accelerating algorithms for Bayesian networks recovery. Adaptation to structures without cycles. Problems in programming. (1), 63–69. [In Ukrainian].
12. Geiger D., Paz A., Pearl J. (1993). Learning simple causal structures. Internat. Journal of Intelligent Systems. 8 (2), 231–247.
13. de Campos L.M., Huete J.F. (1997). On the use of independence relationships for learning simplified belief networks. Intern. Journal of Intelligent Systems. 12 (7), 495–522.
14. Balabanov O.S. (2004). Efficient method for discovery of dependency structures in statistical data. Problems in programming. (2–3), 312–319. [In Russian].
15. Balabanov A.S. (2005). Inference of structures of models of probabilistic dependences from statistical data. Cybernetics and Systems Analysis. 41 (6), 808–817. – Springer, New York.
16. Balabanov A.S. (2003). Inductive reconstruction method for “mono-flow” probabilistic graphical models of Dependencies. Journal of Automation and Information Sciences. 35 (10), 1–8. – Begell House Publishers, Danbury.
17. Balabanov A.S. (2009). Reconstruction of the model of probabilistic dependences by statistical data. Tools and algorithm. J. of Automation and Information Sciences. 41 (12), 32–46.
18. Balabanov O.S. (2016). Induced dependence, factor interaction, and discriminating between causal structures. Cybernetics and Systems Analysis. 52 (1), 8–19.


Одержано 20.12.2016

***Про автора:***

*Балабанов Олександр Степанович*,
доктор фізико-математичних наук,
провідний науковий співробітник.
Кількість наукових публікацій в українських виданнях – 50.
Кількість наукових публікацій в зарубіжних виданнях – 9.
http://orcid.org/0000-0001-9141-9074.

***Місце роботи автора*:**

Інститут програмних систем
НАН України,
03187, м. Київ-187,
проспект Академіка Глушкова, 40.
Тел.: (044) 5263420.
E-mail: bas@isofts.kiev.ua