УДК 681.3

О. Г. Додонов, Д. В. Ланде
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна
тел. (044) 454-21-63

Імовірнісна модель виявлення латентних зв'язків у мережах понять

Запропоновано модель виявлення неявних (латентних) зв'язків у мережах понять у рамках концепції складних мереж. Як приклад розглянуто мережу понять (компаній), які зв'язуються одне з одним через сумісні згадки у веб-публікаціях. Розглянутий напрям аналізу складних мереж актуальний сьогодні при прийняті рішень у таких галузях як маркетинг, соціальні дослідження, конкурентна розвідка.

Ключові слова: *складні мережі, латентні зв'язки, відновлення структури мережі, моделювання, імовірнісна модель.*

На сьогоднішній день розвиток комп'ютерних технологій дозволяє не тільки аналізувати існуючі складні мережі, але й здійснювати моделювання, спираючись, зокрема, на принципи емерджентності [1]. Разом з тим, при моделюванні у багатьох галузях не завжди можна діяти методом проб і помилок, тому необхідно розвивати методи, що дозволяють узагальнювати дані, і на їхній основі перевіряти адекватність моделей [2].

На цей час уже доведено, що реальні складні мережі найчастіше мають властивості так званої безмасштабності [3], що можна розглядати як базу для моделювання, аналізу та прогнозування сценаріїв їхньої еволюції. Крім того, багатьом складним мережам притаманна властивість реорганізації, відродження після руйнувань структури. Відомо, наприклад, що безмасштабні мережі досить толерантні щодо випадкових помилок [4]. У мережі з рівномірним розподілом ступенів вузлів невелика кількість випадкових помилок може її зруйнувати. Безмасштабна мережа може поглинати випадкові помилки, що охоплюють до 80 % її вузлів, і лише потім розпадається. Причина такої стійкості полягає у тому, що помилки більш імовірні у відносно невеликих вузлах. Разом з тим, безмасштабні мережі дуже вразливі з точки зору навмисних атак на їхні концентратори. Атаки, які одночасно знищують всього лише 5–15 % вузлів-концентраторів безмасштабних мереж, можуть зруйнувати всю мережу.

Висока ймовірність виникнення безмасштабних мереж, на противагу рівномірно розподіленим випадковим мережам, виникає завдяки тому, що швидкий ріст створює переваги для перших вузлів мережі у процесі її еволюції. Чим довше діє вузол, тим більша кількість його зв'язків, тому важливість найбільших вузлів дуже велика. Р. Ротенберг [5] відзначив, що властивість безмасштабності реальних терористичних мереж вступає в протиріччя із вказівками щодо комунікаційної інфраструктури, наведеними, наприклад, у навчальному посібнику Аль-Каїди [6]. Можна стверджувати, що безмасштабна природа таких мереж не є предметом цілеспрямованого планування, а результатом природного приведення мережі до певного порядку.

Складні, зокрема, терористичні мережі характеризуються наявністю так званої «структури співтовариства», коли існують групи вузлів, які мають високу щільність ребер між собою, при тому, що щільність ребер між окремими групами низька. Традиційний метод для виявлення структури співтовариств — кластерний аналіз. Існують десятки прийнятних для цього методів, які базуються на різних мірах відстаней між вузлами. Деякі мережі часто характеризуються як клітинні — створені з майже незалежних кліток (рис. 1). З урахуванням цієї особливості конфігурації створюється спеціальний клас моделей терористичних мереж [7].

Формальне визначення клітинних мереж було дане у [8] в термінах мережних компонентів і властивостей. Клітинні мережі мають такі властивості, як надмірність, наявність тісно зв'язаних кліток (4–6 вузлів), відсутність управління вертикальним способом (нечіткі директиви), відсутність планування (формування за рахунок локальних обмежень), можливість еволюціонування у відповідь на деструктивну діяльність [9].

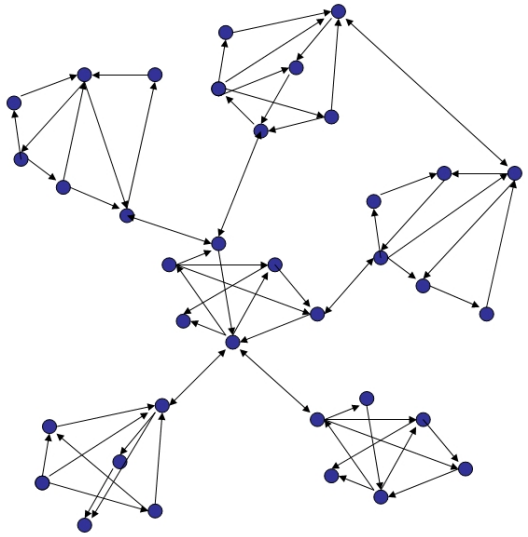

Рис. 1. Фрагмент «клітинної мережі»

Виходячи з результатів наведеного вище моделювання, здається, що руйнування будь-якої мережі є відносно нескладною задачею — досить вилучити ключові елементи — вузли та відповідні зв'язки. Разом з тим, у реальному житті здій-

снюються дещо інші процеси. Соціальні (зокрема, терористичні) мережі мають властивість відновлення після деструктивних впливів, залучення невідомих раніше прихованих (латентних) зв'язків [10]. На рис. 2 наведено схему відновлення зв'язків у мережі після вилучення вузла-концентратора [11]. Після того як мережа розділяється на ізольовані осередки, вона продовжує використовувати свої латентні ресурси та швидко відновлює втрати.

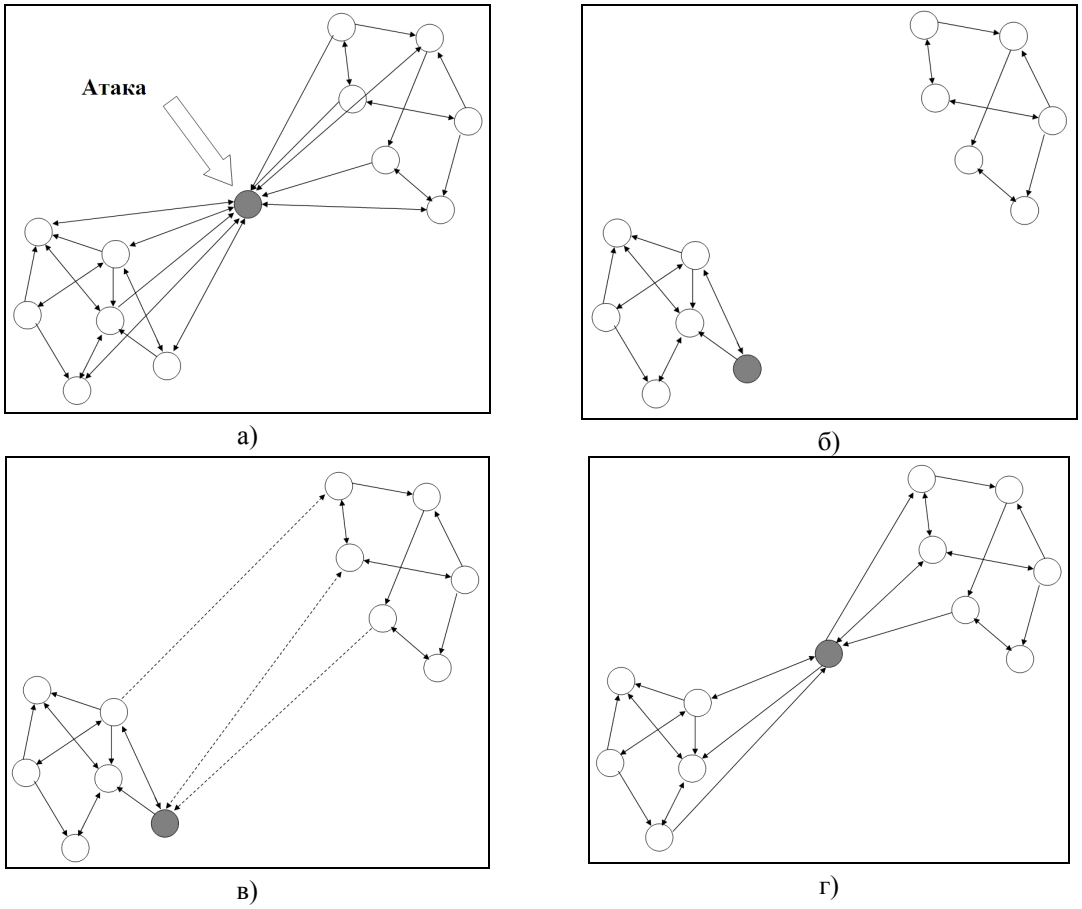

Рис. 2. Відновлення структури мережі шляхом вибору нового посередника (комутатора):
а) атака на мережу; б) незв'язна мережа з вилученим посередником;
в) відродження прихованих (латентних) зв'язків; г) зв'язність мережі відновлено

Процес відновлення заснований на використанні прихованих контактів між членами різних осередків без комутатора. Возз'єднання частин мережі не відбудеться, якщо жодна з пар вузлів не зможе знайти інформацію щодо взаємних посилань. У цьому випадку вплив роз'єднання на показники діяльності мережі залежить від того, чи зможуть знову роз'єднані частини мережі одержати взаємні зв'язки, недолік яких спостерігається у цій частині мережі. Якщо частина мережі була близькою до самодостатності, то вона продовжує функціонувати самостійно. У противному випадку частина мережі припиняє функціонування доти, поки не сформується новий зв'язок.

Якщо одне із возз'єднань виявляється успішним, то його ініціатор стає новим комутатором, який поєднує дві частини мережі. Саме завдяки наведеним механізмам складним динамічним мережам властива можливість самостійного відновлення. Тому, зокрема, пріоритет у дослідженні дестабілізації соціальних мереж віддається пошуку ключових осіб, вилучення яких розділить мережу на окремі фрагменти. Проте експерименти показують, що після того, як така мережа розділяється на ізольовані осередки, вона продовжує використовувати свої приховані ресурси та швидко відновлює втрати.

Аналізуючи зв'язки в мережі, можна виявити багато важливих властивостей, наприклад, визначити наявність кластерів, їхній склад, розходження у зв'язності всередині та між кластерами, ідентифікувати ключові елементи, які зв'язують кластери між собою, тощо. Разом з тим, серйозною перешкодою під час аналізу мереж є неповна інформація щодо зв'язків між її окремими вузлами. Група дослідників з Інституту Санта Фе представила алгоритм [10], за допомогою якого, знаючи, наприклад, лише частину зв'язків ієрархічної мережі, можна з високою ймовірністю відновити відомості щодо відсутніх ланок. Навіть не маючи повного опису системи, можна одержувати репрезентативну вибірку зв'язків і по ній намагатися добудовувати всю мережу. З погляду на природу латентних зв'язків [12, 13] сьогодні досліджуються чисельні мережі, що утворюються також багатьма об'єктами (партіями, компаніями, персонами). Це дозволяє аналітикам робити висновки щодо загальних інтересів окремих груп об'єктів у часі, виявляти ключові елементи мереж, нехтувати неістотними тощо.

Нижче наведено метод дослідження мережі понять, що характеризується великою кількістю вузлів, ребер (зв'язків) з різними ваговими значеннями, високою динамікою появи нових вузлів і зв'язків. Відомо, що матриці взаємозв'язків понять є однією із форм представлення мережених структур, що еквівалентні їхньому графовому представленню. На практиці ці матриці найчастіше відображають близькість окремих понять (сумісну появу в документах або близькість за супутнім контекстом). При різних підходах до їхньої побудови — це, як правило, симетричні матриці, елементи яких — коефіцієнти взаємозв'язків. Ребрам цих графів приписуються вагові коефіцієнти, які пропорційні кількості документів з деякого масиву, одночасно відповідні обом вузлам (об'єктам), що сполучаються цими ребрами. Існують також інші численні підходи до визначення близькості понять у масивах неструктурованих текстів, серед таких можна назвати контекстні, імовірнісні та ентропійні (Mutual Information), але всі вони є лише передумовами для побудови матриць взаємозв'язків, їхнього перегрупування і візуалізації [14, 15].

Розглянемо одне з формальних визначень матриці взаємозв'язків понять M, що відповідне приведеним у роботах [16, 17].

Позначимо: p_i $(i=1,...,K)$ — поняття; $d^{(j)}$ $(j=1,...,N)$ — документ; $d^{(j)} \in D$ — масив документів; $e_i^{(j)}$ — ознака відповідності поняття p_i документа $d^{(j)}$:

$$e_i^{(j)} = \begin{cases} 1, & p_i \in d^{(j)}, \\ 0, & p_i \notin d^{(j)}. \end{cases}$$

Можна визначити рівень зв'язку понять p_i і p_k:

$$M_{ik} = \sum_{j=1}^{N} e_i^j e_k^j.$$

Увівши позначення $E = \left\| e_i^{(j)} \right\|_{i=1,\ldots,K}^{j=1,\ldots,N}$, отримуємо матрицю взаємозв'язку понять

$$M = E^T E = \left\| M_{i,k} \right\|_{i,k=1,\ldots,K}.$$

Недіагональний елемент $M_{i,k}\left(i \neq k\right)$ цієї матриці дорівнює кількості одночасних згадок вузлів *i* та *k* у всіх записах бази даних. Діагональний елемент матриці $M_{i,i}$ — це кількість згадок *i*-го вузла у всіх записах. У реальних ситуаціях вагу зв'язку між особами *i* та *j* можна визначити також, наприклад, за кількістю телефонних переговорів між ними за певний проміжок часу, за кількістю спільно надрукованих наукових робіт тощо.

Далі розглянемо мережу понять (компаній, суб'єктів господарювання), які зв'язуються одне з одним через сумісні згадки у веб-публікаціях (вага зв'язку — кількість появ у тих самих публікаціях). На рис. 3 наведено приклад тривимірного зображення матриці взаємозв'язку понять, що складається з 30 вузлів. Дана матриця була отримана на підставі аналізу масиву веб-публікацій, що отримані системою контент-мониторингу InfoStream протягом березня 2011 року з тематики діяльності Антимонопольного комітету України. Об'єм початкових даних склав понад 1 млн. документів.

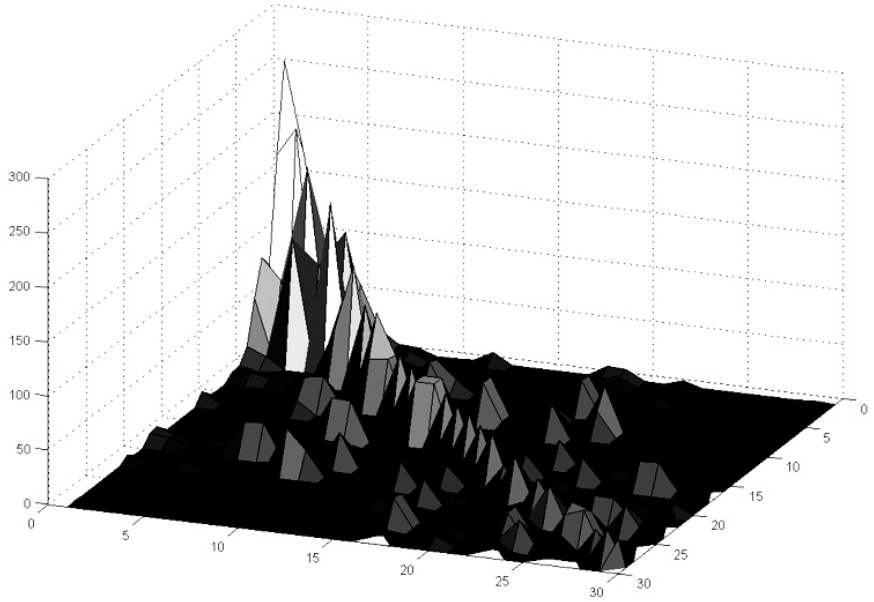

Рис. 3. Матриця зв'язків *M*. По горизонтальних осях відкладені номери об'єктів (компаній), по вертикальній — вагові значення зв'язків (відносні)

У даній матриці інцидентності оцінкою зв'язків об'єктів є значення її відповідних елементів. Якщо перейти до аналізу реальної ситуації, то такий зв'язок можна розглядати як імовірнісний. Відповідним чином нормалізувавши значення елементів матриці, можна перейти до так званої «матриці нечітких зв'язків» елементами яких є ймовірність зв'язку між об'єктами.

Тобто, передбачається, що $p_{i,j}$ — оцінка ймовірності зв'язку об'єктів *i* та *j*. У загальному випадку, передбачається, що ця оцінка експертна, не залежна від інших вузлів мережі. Ці оцінки можна було б уточнити, враховуючи не тільки їхні прямі зв'язки, але й їхні зв'язки через треті, четверті і т.д. вузли. Припустимо, що вузли 1 і 2 зв'язані безпосередньо один з одним, а також через вузол 3 (рис. 4). Відповідні оцінки ймовірності зв'язків складають $p_{1,2}p_{1,3}$.

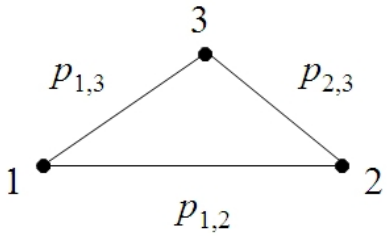

Рис. 4. Початкові оцінки ймовірності зв'язків

Тоді можна зробити наступну оцінку «нечіткої» ймовірності того, що зв'язку між вузлами 1 і 2 не існує: $\overline{p}_{1,2}^{(1)} = (1-p_{1,2})(1-p_{1,3}p_{3,2})$.

Тобто нова оцінка ймовірності зв'язку між вузлами 1 і 2 складає:

$$p_{1,2}^{(1)} = 1-(1-p_{1,2})(1-p_{1,3}p_{3,2}).$$

Відповідно формула обліку всіх зв'язків через «треті вузли» має вигляд:

$$p_{i,j}^{(1)} = 1-(1-p_{i,j})\prod_{k\neq i,j}(1-p_{i,k}p_{k,j}).$$

Очевидно, при $p_{i,j}\in[0,1]$ для будь-яких *i* та *j* буде мати місце те, що й $p_{i,j}^{(1)}\in[0,1]$ для будь-яких *i* та *j*. Дійсно, це витікає з того, що величина $(1-p_{i,j})\prod_{k\neq i,j}(1-p_{i,k}p_{k,j})$ не більше одиниці і не менше нуля як добуток ненегативних співмножників, кожен з яких менше одиниці.

При цьому завжди $p_{i,j}^{(1)} \geq p_{i,j}$. Доведемо це, для чого введемо позначення $\alpha = \prod_{k\neq i,j}(1-p_{i,k}p_{k,j})$. Очевидно $\alpha\in[0,1]$. Необхідно довести, що $p_{i,j}^{(1)} - p_{i,j} \geq 0$. Це витікає з викладення: $p_{i,j}^{(1)} - p_{i,j} = 1-\alpha(1-p_{i,j}) - p_{i,j} = (1-p_{i,j})(1-\alpha)$. Кожен з отриманих співмножників не негативний, отже, їхній добуток також не негативний.

Можна також оцінити ймовірність з урахуванням зв'язків через 4-й, 5-й і т.д. вузли, модифікувавши функцію розрахунку $p_{i,j}^{(1)}$ таким чином:

$$p_{i,j}^{(1)} = 1-(1-p_{i,j})\prod_{k\neq i,j}(1-p_{i,k}p_{k,j})\prod_{k\neq i\neq l\neq j}(1-p_{i,k}p_{k,l}p_{l,j})\prod_{k\neq i\neq l\neq m\neq j}(1-p_{i,k}p_{k,l}p_{l,m}p_{m,j})\ldots$$

Отримана матриця відображатиме не тільки явні зв'язки, що виражені оцінками ймовірності, але і зв'язки 2-го, 3-го і т.д. рівнів. На практиці співмножники, починаючи вже з $\prod_{k\neq i\neq l\neq j}(1-p_{i,k}p_{k,l}p_{l,j})$, виявляються достатньо близькими до одиниці, щоб їх зазвичай можна було не враховувати в практичних розрахунках.

До отриманої у результаті матриці, елементами якої є $p_{i,j}^{(1)}$, також можна застосувати приведений вище алгоритм, розглядаючи його результати, всього лише як першу ітерацію. Тобто до отриманої на першому кроці ітерації матриці можна застосувати описаний алгоритм:

$$p_{i,j}^{(m+1)} = 1-(1-p_{i,j}^{(m)})\prod_{k\neq i,j}(1-p_{i,k}^{(m)}p_{k,j}^{(m)}).$$

Тут $p_{i,j}^{(m)}$ для будь яких *i* та *j* є монотоно неубутною функцією від *m*, крім того, при достатньо великих *n* всі елементи матриці $\left\|p_{i,j}^{(m)}\right\|$, окрім діагональних, виявляються близькими до одиниці. За необхідності проведення декількох ітерацій здійснюється нормування величин $p_{i,j}^{(m)}$ шляхом піднесення їх у степінь α.

Для оцінки процесу збіжності елементів випадкової матриці до одиниці навіть при $\alpha = 1{,}5$ було застосовано чисельне моделювання. Виявилося, що для випадкової симетричної матриці розміром 100×100 з елементами з діапазону (0,1] для досягнення мінімальними значеннями одиниці при α = 1,5 кількість ітерацій складає всього 4–5, тобто процес сходиться достатньо швидко.

Таким чином, враховуються зв'язки, що визначені в початковій матриці не тільки між третім, але й між четвертим і п'ятим і т.д. рівнями. Разом з тим, саме внаслідок того, що кубічним і більш порядком величин можна нехтувати на практиці, в реальній роботі використовується не більше 3-х ітерацій.

Для оцінки дієвості запропонованого підходу будувалася матриця, що відповідна мережі із степеневим розподілом ваги вузлів, у якої віддалялися (обнулялися вагові значення) ребра із значеннями в середньому діапазоні. При одиничному видаленні ребер вони практично завжди відновлювалися за допомогою приведеного методу за 1 крок ітерації. При видаленні 20 % ребер вони відновлювалися приблизно в 75 % випадків.

Застосовуючи даний підхід (одну ітерацію) до приведеної вище матриці персон, що відібрані з тематики діяльності Антимонопольного комітету України (розглядаючи частоти появ як основу переходу до «нечіткої» ймовірності), були отримані результати, представлені на рис. 5.

Аналізуючи набутих значень, можна, зокрема, відмітити, що між вузлами 1 і 7 (за даними початкової матриці) немає прямого зв'язку (рис. 6). У той самий час значення нечіткої ймовірності складає 0,381, що понад як у три рази вище середнього коефіцієнта близькості за всією матрицею.

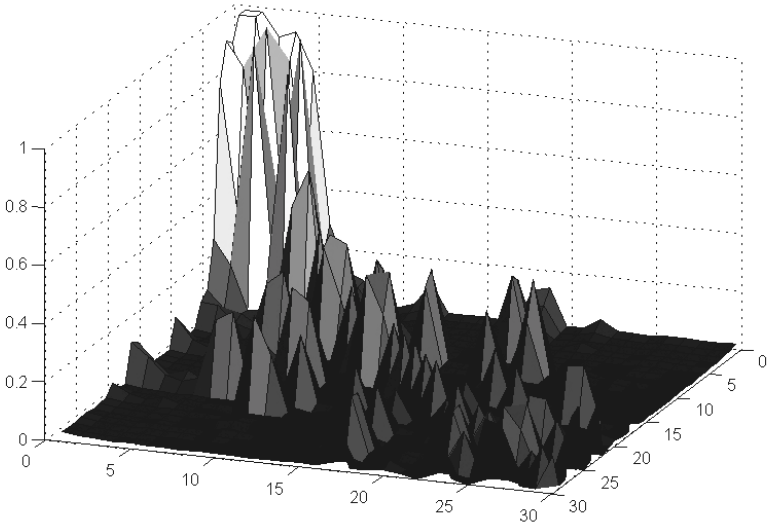

Рис. 5. Матриця «нечітких» імовірнісних зв'язків. По горизонтальній осі відкладені номери вузлів (об'єктів), по вертикальній — значення матриці

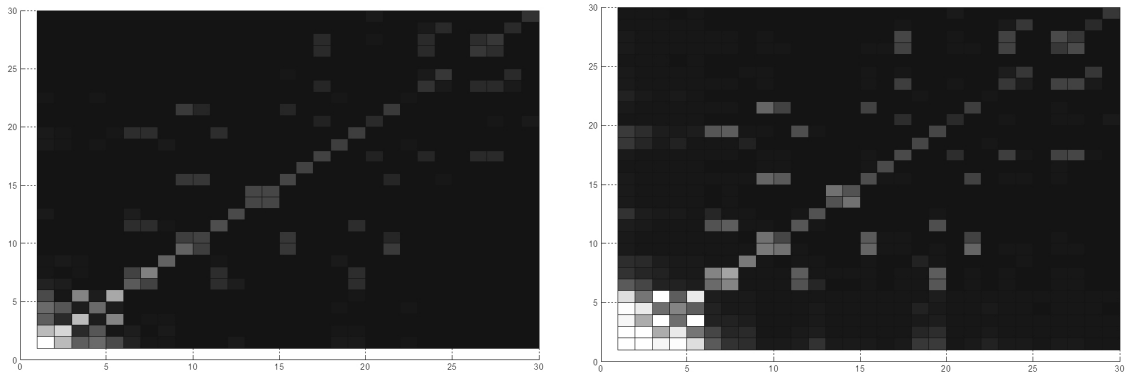

Рис. 6. Порівняння проекцій матриць: а) явних імовірнісних зв'язків; б) «нечітких» імовірнісних зв'язків

Наведений метод багато в чому нагадує підходи, що базуються на кластерному аналізі, проте його принципова відмінність у тому, що він ґрунтується на теоріях імовірності та надійності. На відміну від підходів, що існують на даний час, до виявлення взаємозв'язків понять, запропонований метод дозволяє виявляти, визначати відносну вагу і відобразити неявні зв'язки будь-яких рівнів. Слід зазначити, що аналоги подібних методів з теорії надійності до цих пір не знаходили широкого застосування в практиці аналітичної обробки інформації. Разом з тим, розглянутий напрям аналізу складних мереж сьогодні актуальний при прийняті

рішень у таких галузях як маркетинг, соціальні дослідження, конкурентна розвідка, в завданнях виявлення і візуалізації різних співтовариств.

1. *The* Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion: ed. by P. Clayton, P. Davies. — Oxford University Press, 2006. — 346 p.

2. *Ланде Д.В.* Новітні підходи і технології інформаційно-аналітичної підтримки прийняття рішень // Національна безпека: український вимір: щокв. наук. зб. / Рада нац. безпеки і оборони України, Ін-т пробл. нац. безпеки. — К., 2008. — Вип. 1–2 (20–21). — С. 87–105.

3. *Robb J.* Scale-Free Terrorist Networks / J. Robb. — 2004. — Jef Allbrights Web Files. — Режим доступу: URL:www.jefallbright.net/node/view/2632

4. *Newman M.E.J.* The Structure and Function of Complex Networks // SIAM Review. — 2003. — **45**. — P. 167–256.

5. *Rothenberg R.* From Whole Cloth: Making up the Terrorist Network // Connections. — 2002. — **24**, N 3. — P. 36–42.

6. *Al Quaeda* Training Manual: Declaration of Jihad Against Unholy Tyrants // Al-Quaeda. — 2001. — Режим доступу: URL: http://www.usdoj.gov/ag/trainingmanual.htm

7. *Carley K.* Destabilizing Networks / K. Carley, J. Lee , D. Krackhardt // Connections. — 2002. — **24**, N 3. — P. 79–92.

8. *Frantz T.* A Formal Characterization of Cellular Networks / T. Frantz, K.M. Carley // Carnegie Mellon University School of Computer Science Institute for Software Research International. — Tech. Rep. CMU-ISRI-05-109, 2005.

9. *Sageman M.* Understanding Terror Networks / M. Sageman. — University of Pennsylvania Press, 2004.

10. *Clauset A.* Hierarchical Structure and the Prediction of Missing Links in Networks / A. Clauset, C. Moore, M.E.G. Newman // Nature. — 2000. — **453**. — P. 98–101.

11. *Цветоват М.* Симуляция человеческих обществ с искусственным интеллектом. Случай террористических сетей / М. Цветоват // Ежеквартальный Интернет-журнал «Искусственные общества». — 2007. — Т. 2, № 2. — С. 5–29.

12. *Lande D.V.* Dynamic Network of Concepts from Web Publications / D.V. Lande, A.A. Snarskii // ePrint Arxiv (0806.1439).

13. *Снарский А.А.* Метод выявления неявных связей объектов / А.А. Снарский, Д.В. Ландэ, М.И. Женировский // Труды 11-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2009, Петрозаводск (Россия), 2009. — С. 46–49.

14. *Danon L.* Comparing Community Structure Identification / L. Danon, A. Dнaz-Guilera, J. Duch, A. Arenas // J. Stat. Mech. (2005) P09008. doi: 10.1088/1742-5468/2005/09/P09008 PII: S1742-5468 (05) 07477-7.

15.*Knepper M.M.* Information Retrieval and Visualization Using SENTINEL / M.M. Knepper, R. Killam, K.L. Fox, O. Frieder // TREC. — 1998. — P. 336–340.

16. *Додонов А.Г.* Выявление понятий и их взаимосвязей в рамках технологии контент-мониторинга / А.Г. Додонов, Д.В. Ландэ // Реєстрація, зберігання і оброб. даних. — 2006. — Т. 8, № 4. — С. 45–52.

17. *Калиткин Н.Н.* Математические модели природы и общества / Н.Н. Калиткин, Н.В. Карпенко, А.П. Михайлов и др. — М.: Физматлит, 2005. — 360 с.

Надійшла до редакції 06.06.2011