Малик І.В., Кнігніцька Т.В., Горбатенко М.Ю.

# КЛАСТЕРИЗАЦІЯ: МАРКОВСЬКИЙ АЛГОРИТМ

У даній роботі розглянуто проблему кластеризації на графах на основі власних значень стохастичної матриці графа. Доведено, що власні значення стохастичної матриці для великих графів ($N > 100$) поділяються на три групи, одна із яких є визначальною для числа кластерів у графі. Використовуючи теорію випадкових матриць, вдалося показати, що асимптотичний розподіл підгрупи дійсних частин власних значень стохастичної матриці графу описується напівколовим розподілом Вігнера, причому параметр даного розподілу є $O\left(\frac{1}{\sqrt{N}}\right)$. Використання стохастичних матриць дало змогу точно локалізувати власні значення, що відповідають за кількість кластерів, чого не вдавалося зробити для матриць суміжності. Основні припущення моделі пов'язані з властивостями дискретних ланцюгів Маркова, що дозволяє розширити область застосування отриманих результатів на більш ширший клас об'єктів. Теоретичні результати перевірені на кластеризації часових рядів, що описують вартості $N = 470$ акцій S&P500 в період 2013 –2018 рр. Результати кластеризації даних часових рядів показали наявність 5 чітко виражених груп, які узгоджується із використаними даними.

*Ключові слова і фрази:* стохастичні матриці, кластеризація часових рядів, неструктуровані дані.

Чернівецький національний університет імені Юрія Федьковича, Чернівці, Україна (Малик І.В., Кнігніцька Т.В., Горбатенко М.Ю.)
e-mail: *malyk.igor.v@gmail.com (Малик І.В.), knig.tatyana.v@gmail.com (Кнігніцька Т.В.), mykola.gorbatenko@gmail.com (Горбатенко М.Ю.)*

## Вступ

Кластеризація об'єктів різної природи є на даний час однією з найбільш важливих задач машинного навчання та глибинного навчання [11]. Всі задачі кластеризації (*Unsupervised learning*) об'єднує декілька спільних проблем: визначення метрики в просторі об'єктів, для яких проводиться кластеризація, та визначення оптимальної кількості кластерів для даних об'єктів. У випадку числових даних (структурованих даних) перша проблема вирішується за допомогою розгляду відстані в $R^n$ як метрики. Проте для нечислових (неструктурованих даних) вибір потрібної метрики відіграє важливу роль при подальшій кластеризації. Неструктуровані об'єкти повністю (або частково)

УДК 004.852, 519.177, 519.217.2
2010 *Mathematics Subject Classification:* 60J10, 65C05, 68W99.

© Малик І.В., Кнігніцька Т.В., Горбатенко М.Ю., 2019

можуть визначатися часовими рядами та графами. Кластеризація на графах та кластеризація часових рядів є одними із найважливіших розділів машинного навчання, оскільки багато об'єктів задаються саме за допомогою даних неструктурованих об'єктів [3]. Тому розгляд нових алгоритмів кластеризації для даних об'єктів є одним із основних напрямків машинного навчання в даний час.

Алгоритми кластеризації для часових рядів можна умовно розділи на дві категорії:

- ***Наївні алгоритми***. Дана категорія алгоритмів базується на припущенні, що часовий ряд є вектором у Евклідовому просторі $R^n$. Недолік даних алгоритмів очевидний, оскільки не враховується структура моделей часових рядів [10].

- ***Модель – орієнтовані алгоритми.*** Дані алгоритми враховують припущення щодо моделей часових рядів, які потрібно кластеризувати.

Кожна із запропонованих категорій володіє своїми перевагами та недоліками, проте спільною проблемою для кожного алгоритму є визначення оптимальної кількості кластерів $k_{opt}$. Найбільш вживаними методами для визначення $k_{opt}$ є метод ліктя, метод силуета, $k$-Core decomposition.

В даній роботі основна увага буде приділена саме визначенню оптимальної кількість кластерів $k_{opt}$ в незалежності від категорії, до якої належить метод для побудови метрики. На відміну від багатьох аналогічних робіт, в даному дослідженні буде використано марковський алгоритм кластеризації, який грунтується на використанні стохастичної матриці графа.

Так як власні значення стохастичних матриць є визначальними при кластеризації даних, варто відзначити роботу [4]. Biely та Thurner вивчали форму спектрів власних значень корреляційних матриць часових рядів, отриманих із наборів броунівських випадкових блукань зі зсувом за часом. Дані матриці можна розглядати як випадкові, асиметричні матриці із спеціальною структурою. У роботі показано, що спектр власних значень є круговим для стохастичних матриць. Щільність власних значень отримано на основі некорельованих гаусових часових рядів. Теоретичні результати порівнюються зі щільностями власних значень, отриманими з високочастотних (5 хв) даних S&P500. У роботі також ідентифіковано різні невипадкові закономірності та знайдено асиметричні залежності, пов'язані з власними знаннями.

## Основні позначення та припущення

Як було зазначено вище, часові ряди та графи є неструктурованими об'єктами, що робить їх подібними в контексті задачі кластеризації та визначення оптимальної кількості кластерів $k_{opt}$. Тому далі розглядатимемо лише графи, розуміючи, що часові ряди мають аналогічну структуру.

Розглянемо основні позначення графів, які будуть використовуватися у даній роботі. Граф позначатимемо через $G = (V, E)$, де $V = \{1, 2, \ldots, N\}$ – множина вершин графа (окремі часові ряди), $E = \{e_1, \ldots, e_m\}$ – множина ребер графа, що з'єднують

вершини $V$ (визначаються модулями коваріацій для часових рядів). Припустимо, що граф задається матрицею суміжності $A$:

$$A = A_{N\times N},$$

де елемент $A_{ij}$ дорівнює ваговому коефіцієнту між вершинами $i$ та $j$. Припустимо, що $A_{ij} \in R_+$, причому для графів $A_{ij}$ ототожнюють з кількільстю ребер між вершинами $i$ та $j$. Для часових рядів $A_{ij}$ будемо ототожнювати із модулем коваріації між часовими рядами, що визначають $i$-й та $j$-й об'єкти. Стохастична матриця $P$, що відповідає графу $G$ та представлена матрицею суміжності $A$, задається співвідношенням

$$P_{ij} = \frac{A_{ij}}{\sum\limits_{j=1}^{N} A_{ij}}. \tag{1}$$

Дана матриця є стохастичною матрицею, яка і буде основним об'єктом дослідження в даній роботі.

Розглянемо декілька основних тверджень, які будуть далі використані при оцінці кількості кластерів $k_{opt}$. Основне твердження щодо визначення оптимальної кількості кластерів $k_{opt}$ тісно пов'язане із алгебраїчною кратністю власного значення $\lambda = 1$ для стохастичної матриці $P$ [5].

**Теорема 1.** *Алгебраїчна кратність власного значення $\lambda = 1$ для стохастичної матриці $P$ рівна кількості незвідних класів для ланцюга Маркова, що описується матрицею $P$.*

Отже, дана теорема стверджує, що кількість неперетинних класів рівна кратності власного значення $\lambda = 1$ для матриці $P$, що в свою чергу співпадає з кількістю кластерів. Даний результат є теоретичним, оскільки в реальних системах всі стани є сполучними. Використовуючи неперервну залежність між елементами матриці та її власними значеннями, буде показано, що кількість кластерів в стохастичній матриці характеризує кількість власних значень, близьких до 1 . Дану залежність можна описати наступним співвідношенням, яке буде використано при доведенні основного твердження роботи:

$$\sum_{i=1}^{N} \left(\lambda_i\left(A\right) - \lambda_i\left(B\right)\right)^2 \ \leq Tr(A-B)^2, \tag{2}$$

де $A,\ B$ – квадратні матриці розмірності $N \times N$; $\lambda_i\left(A\right),\ \lambda_i\left(B\right),\ i = 1,..,N$ – власні значення матриць $A,\ B$, відповідно, упорядковані за спаданням чи зростанням.

## Випадкові та стохастичні матриці

Теорія випадкових матриць є однією із математичних моделей автоматизованих розумних енергосистем [9], що описують функціонування великих незалежних систем. Одним із основних завдань, пов'язаних із побудовою розумної системи, є розподіл функцій системи між основними "центрами" системи. Ще одним прикладом застосування випадкових матриць є системи зв'язку та фінансова математика, розглянуті в роботі

[2]. Таким чином, випадкові матриці є однією із математичних моделей для опису Big Data. Основні результати теорії випадкових матриць по'вязані із спектральним аналізом даних матриць, тобто із властивостями власних значень. Одним із найбільш важливих результатів, отриманих в даному напрямку, є результати українських математиків Марченко та Пастура [11], якими було знайдено асимптотичний розподіл власних значень симетричної випадкової матриці

$$Z = \frac{1}{m} X X',$$

де $X$ – симетрична матриця розмірності $n \times m$, елементами якої є незалежні випадкові величини із нульовим середнім та скінченною дисперсією $\sigma^2$. Основний результат роботи [11] стверджує, що випадкова величина $\Lambda$, яка визначається наступним розподілом

$$P\left(\Lambda \in \boldsymbol{B}\right) = \frac{1}{m} \# \left\{\lambda_i\left(Z\right) \in B\right\},$$

має граничним розподілом розподіл Марченко-Пастура з параметрами $(\mu, a, b)$, який задається щільністю

$$f\left(x\right) = \frac{\sqrt{(b-a)(x-a)}}{2\pi\sigma^2\mu x}, \ x \in \left[a, b\right],$$

де

$$\mu = \lim_{m\to\infty} \frac{n(m)}{m} \ ;$$

$$a = \sigma^2(1-\sqrt{\mu})^2;$$

$$b = \sigma^2(1-\sqrt{\mu})^2.$$

Розглянемо поняття стохастичних випадкових матриць, яке буде використане при дослідженні основного результату роботи.

**Означення 1.** *Стохастичною випадковою матрицею $P$ називається квадратна матриця, яка володіє наступними властивостями:*

1. *Елементи матриці $P_{ij}$ є невід'ємними випадковими величинами.*

2. *Елементи матриці $P_{ij}$ є нормованими, тобто*

$$\sum_{j=1}^{N} P_{ij} = 1.$$

3. *Рядки матриці $P$ є незалежними випадковими векторами.*

Розглянемо приклад стохастичної випадкової матриці та розглянемо основні особливості даної матриці. Припустимо, що матриця суміжності на графі $A$ розмірності $N \times N$ є випадковою матрицею, елементи якої мають рівномірний розподіл на довільному інтервалі $[a, b]$, де

$$0 \leq a < b < \infty.$$

Тобто в даному прикладі елементи матриці мають рівномірний розподіл на $[a, b]$:

$$A_{ij} \sim U(a, b),$$

де $U(a, b)$ – рівномірний розподіл на $[a, b]$. Використовуючи матрицю $A$, утворимо стохастичну матрицю за формулою (1). Розглянемо деякі властивості власних значень матриці $P$, які безпосередньо випливають із означення стохастичної випадкової матриці.

1. З імовірністю 1, власним значенням матриці $P$ є $\lambda = 1$.

2. Власні значення матриці $P$ задовольняють співвідношення

$$0 \leq \sum_{i=1}^{N} \lambda_i(P) \leq N.$$

3. Математичне сподівання власних значень рівне 1:

$$\sum_{i=1}^{N} E(\lambda_i(P)) = 1.$$

Остання властивість є наслідком рівності суми власних значень матриці та сліду матриці:

$$\sum_{i=1}^{N} \lambda_i(P) = \sum_{i=1}^{N} P_{ii}$$

та того факту, що всі елементи матриці $P$ мають однаковий розподіл. Враховуючи дане зауваження, отримаємо наступні співвідношення для елементів матриці $P$:

$$EP_{ij} = \frac{1}{N}.$$

Таким чином для власних значень матриці $P$ справедлива наступна рівність:

$$\sum_{i=2}^{N} E(\lambda_i(P)) = 0,$$

де власні значення матриці $P$ впорядковані за спаданням дійсних частин

$$1 = Re(\lambda_1) \geq Re(\lambda_2) \geq ... \geq Re(\lambda_N).$$

Розглянемо результати симуляції власного значення $\lambda_2$, яке є найближчим до 1, тобто

$$\lambda_2 = argmin_{\lambda_i \neq 1} \{|\lambda_i - 1|\}.$$

Із наведеного нижче рис. 1 видно, що розподіл власного значення $\lambda_2(N)$ "стягується" до 0 у випадку рівномірного розподілу всіх елементів матриці суміжності $A$. Даний результат може бути інтерпретований як узагальнення класичного результату в класичному алгоритмі Маркова [7], який можна сформулювати наступним чином:

**Лема 1.** *[11] Нехай всі елементи стохастичної матриці $P$ рівні між собою, тобто*

$$P_{ij} = \frac{1}{N}.$$

*Тоді кількість кластерів, обчислена згідно із класичним алгоритмом рівна* 1 *для довільних значень параметрів $r$ та $s$.*

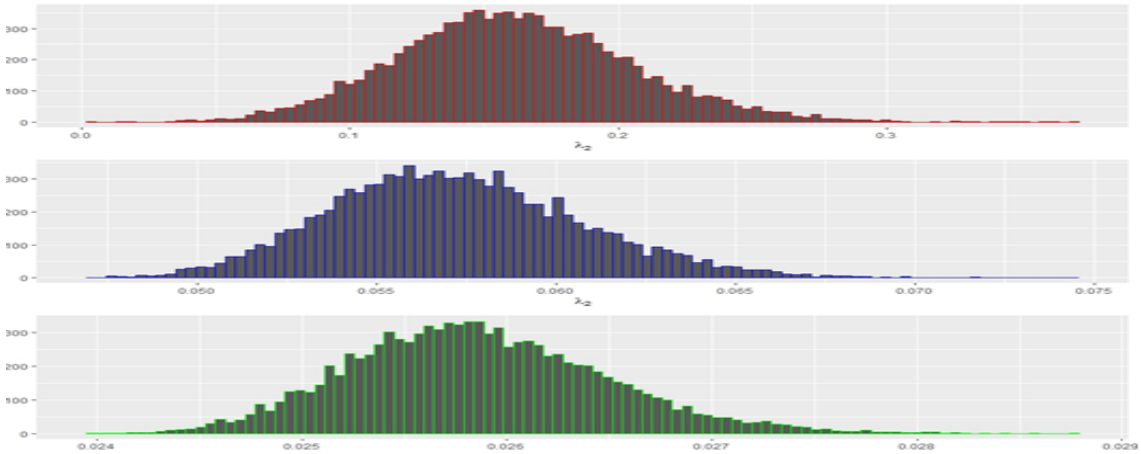

Рис. 1: Розподіл $\lambda_2$ матриці $P$ за умови $A_{ij} \sim Unif\,(0,1)$ з розмірністю $N = 10,\ 100,\ 500$ (зверху вниз)

Даний результат є тривіальним наслідком з використанням двох операцій у класичному методі:

$$P = P^r,$$

$$P = L\,(P; s)\,,$$

де оператор $L\,(P; s)$ визначає операцію інфляції для матриці $P$ з параметром $s$. Результат леми 1 є частинним випадком навіть для графів невеликої розмірності, оскільки при найменших відхиленнях значень $P$ від свого середнього значення, результати використання класичного марковського алгоритму змінюються в залежності від параметрів алгоритму $(r, s)$. Нижче буде доведено більш сильний результат для довільного розподілу на $(0, \infty)$, що узагальнює лему 1. Даний результат буде використано для відслідковування наявності одного кластера в графі, що задається стохастичною випадковою матрицею $P$.

**Теорема 2.** *Нехай виконуються наступні умови:*

1. *Всі елементи матриці $A$ є незалежними та мають одинаковий розподіл*

$$A_{ij} \sim Distr,$$

   *де розподіл $Distr$ має носієм підмножину множини $(0, \infty)$ та має скінченний момент порядку $2 + \delta, \delta > 0$.*

2. *Елементи матриці переходу за 1 крок $P_{ij}$ визначаються із співвідношення (1).*

*Тоді за ймовірністю має місце збіжність*

$$\lambda_2 (P; N) \to 0,$$

*де* $\lambda_2 (P; N) \neq 1$ *– найближче до* 1 *власне значення матриці* $P$, $N \times N$ *– розмірність матриці* $P$.

*Доведення.* Для доведення даного факту відзначимо, що значення $\lambda_2 (N)$, яке відповідає значенням матриці

$$P_{ij} = E\left(\frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}}\right)$$

рівне 0. Для доведення цього факту розглянемо власні значення матриці $P$ у наступному вигляді

$$\det (\lambda I - P) \ = \lambda^n + c_1 \lambda^{n-1} + \ldots c_{n-1}\lambda + c_n,$$

де коефіцієнти $c_i$ визначаються за формулою

$$c_i = (-1)^i \sum_k \det (M_k (i)) \ ,$$

де $M_k (i)$ – головні мінори матриці $P$ розмірності $i \times i$. Згідно побудови,

$$c_2 = ... = c_N = 0.$$

Використовуючи даний факт, знайдемо, що характеристичний поліном для матриці $E(P)$ рівний

$$E (\det (\lambda I - P) \ ) = E \left(\lambda^n + c_1 \lambda^{n-1} + \cdots + c_{n-1}\lambda + c_n\right) =$$

$$= \lambda^n + E (c_1) \lambda^{n-1} + \cdots + E (c_{n-1}) \lambda + E (c_n) =$$

$$= \lambda^n - \lambda^{n-1}.$$

Таким чином, розв'язками характеристичного рівняння є

$$\lambda_1(E(P)) = 1, \ \lambda_i(E(P)) = 0, \ i = 2, \ldots, N.$$

Отже, справедливим є той факт, що математичне сподівання власного значення $\lambda_2$(P) рівне 0 за умов теореми.

Для доведення збіжності до 0 випадкової величини $\lambda_2$ розглянемо дисперсії діагональних елементів

$$D\left(\sum_{i=1}^{N} P_{ii}\right) = \sum_{i=1}^{N} D (P_{ii}) = ND (P_{11}) ,$$

оскільки всі значення $P_{ii}$ є незалежними та однаково розподіленими. Обчислимо дисперсію випадкової величини

$$D (P_{11}) = E \left(P_{11}^2\right) - \frac{1}{N^2}.$$

Використовуючи означення матриці $P$, можна показати, що

$$E\left(P_{11}^2\right) = \frac{1}{N^2} + O\left(\frac{1}{N^2}\right).$$

Таким чином

$$D\left(P_{11}\right) = O\left(\frac{1}{N^2}\right)$$

та

$$D\left(\sum_{i=1}^{N} P_{ii}\right) = O\left(\frac{1}{N}\right).$$

Для доведення того факту, що $\lambda_2\left(P;N\right)$ прямує до 0 за ймовірністю, використаємо нерівність (2). Використовуючи дане співвідношення, отримаємо наступну оцінку для дисперсії $\lambda_2\left(P;N\right)$:

$$E(\lambda_2\left(P;N\right))^2 = D(\lambda_2\left(P;N\right))^2 \leq$$

$$\leq \sum_{i=1}^{N} D\left(\lambda_i\left(N\right)\right) \leq E\left(\sum_{i=1}^{N}\sum_{k=1}^{N}\left(P_{ik} - \frac{1}{N}\right)\left(P_{ki} - \frac{1}{N}\right)\right) =$$

$$= E\left(\sum_{i=1}^{N}\left(P_{ii} - \frac{1}{N}\right)^k\right) = \sum_{i=1}^{N} E\left(P_{ii} - \frac{1}{N}\right)^2 =$$

$$= \sum_{i=1}^{N} D\left(P_{ii}\right) = D\left(\sum_{i=1}^{N} P_{ii}\right).$$

Отже, має місце нерівність

$$D\left(\lambda_2\left(N\right)\right) \leq O\left(\frac{1}{N}\right).$$

Використовуючи дану нерівність та нерівності Чебишова, отримаємо твердження теореми. □

Результати моделювання власних значень для рівномірного розподілу $A_{ij} \sim U(a,b)$ наведені на рис. 2. З даного рисунка видно, що власні значення $\lambda_2\left(P;N\right), \ldots, \lambda_N\left(P;N\right)$ "стягуються" до 0 зі зростанням $N$. Цей факт описує поведінку власних значень стохастичної матриці графа у випадку наявності тільки одного кластера.

Загальний вигляд розподілу власних значень $\lambda_2\left(N\right), \ldots, \lambda_N\left(N\right)$ при довільній кількості кластерів можна знайти, використовуючи методику, запропоновану Марченком та Пастуром [11]. Зауважимо, що на відміну від класичного випадку випадкових матриць, стохастичні випадкові матриці володіють трьома особливостями: по-перше, середнє значення кожного елемента стохастичної випадкової матриці має математичне сподівання не менше за 0; по-друге, сума елементів в кожному рядку повинна бути рівною 1 з імовірністю 1; по-третє, елементи стохастичної випадкової матриці повинні бути невід'ємними. Дані додаткові обмеження на елементи матриці змінюють розподіл власних значень стохастичної випадкової матриці $P$. Проте загальна асимптотична поведінка власних значень має ту ж поведінку, що і для загального випадку випадкових матриць – рис. 3. Використовуючи отриману вище теорему, нам вдалося довести, що власні

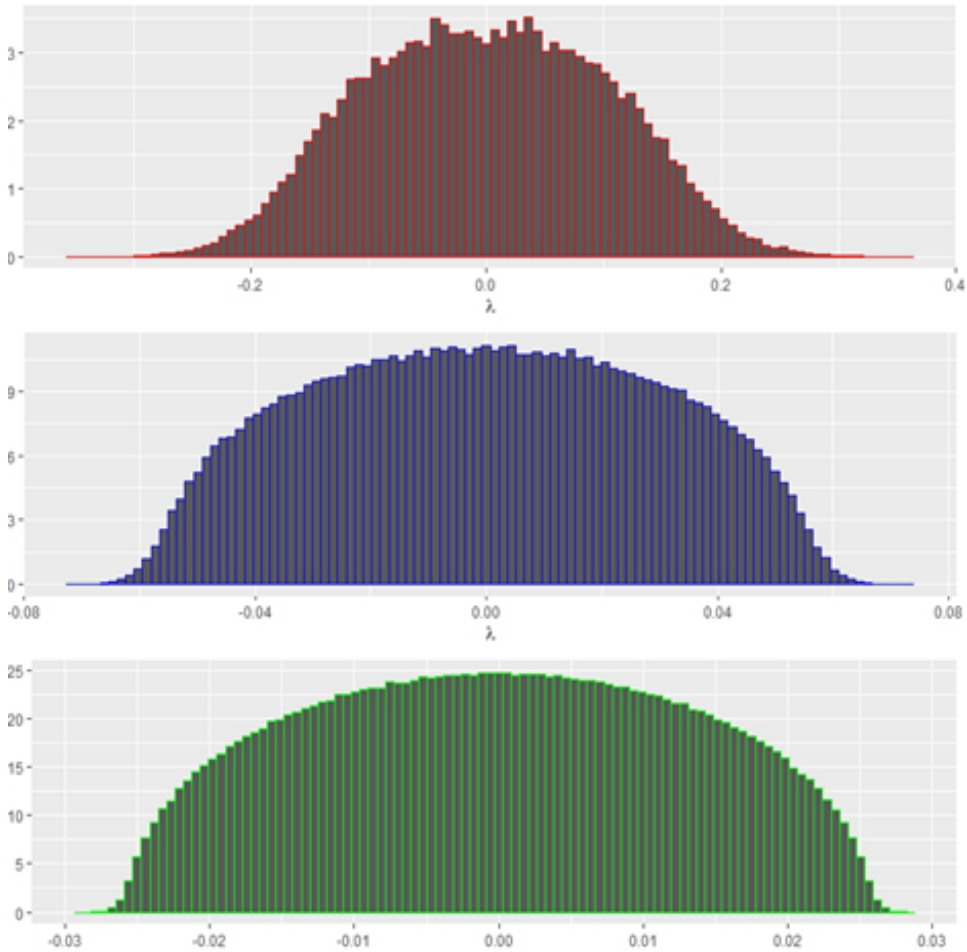


Рис. 2: Розподіл $\lambda_2$ матриці $P$ за умови $A_{ij} \sim Unif\,(0,1)$ з розмірністю $N = 10,\ 100,\ 500$ (зверху вниз)

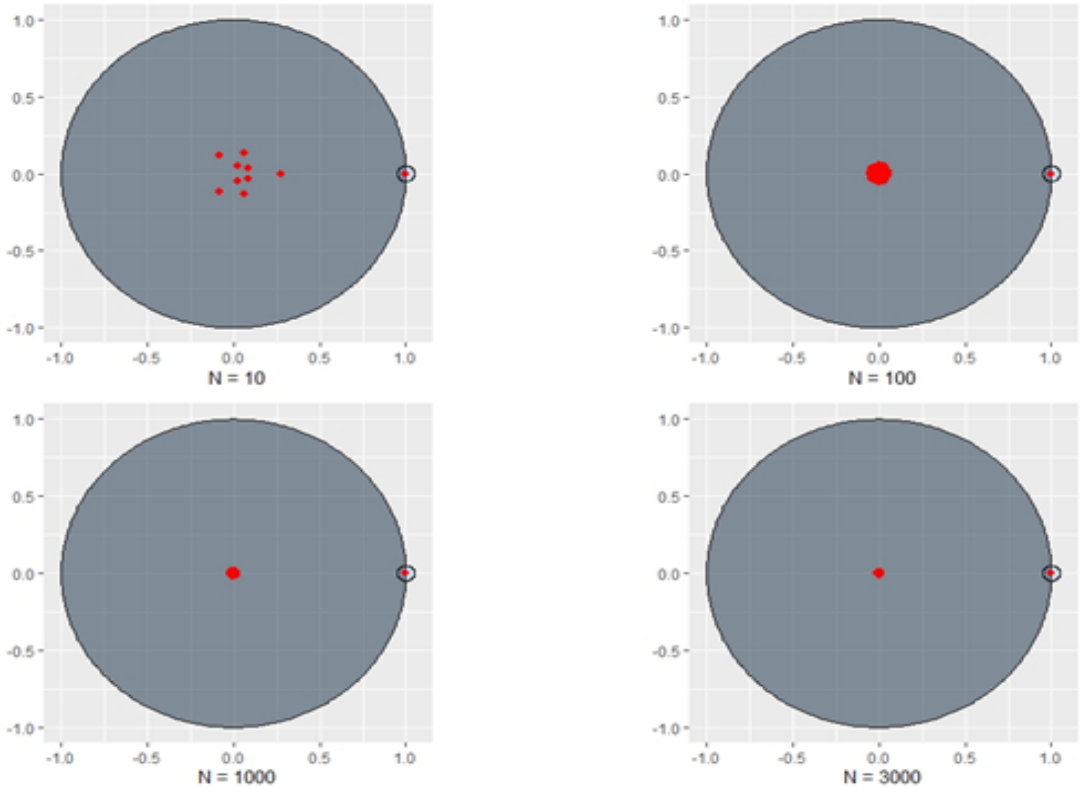


Рис. 3: Власні значення матриці $P$ за умови $A_{ij} \sim Unif\,(0,1)$ з розмірністю $N = 10,\ 100,\ 1000,\ 3000$

значення матриці $P$, які не несуть інформації про кластерність графу, визначаються співвідношенням

$$|\lambda_i(P;N)| \leq \frac{1}{\sqrt{N}}. \tag{3}$$

Використовуючи означення власних значень (3), будемо визначати оптимальну кількість кластерів в графі за допомогою співвідношення

$$k_{opt} = \#\left\{i = 1, ..., N : |\lambda_i(P) - 1| < 1 - \frac{1}{\sqrt{N}}\right\}. \tag{4}$$

Аналіз методом Монте-Карло та аналіз S&P500

## Метод Монте-Карло

Перевіримо точність визначення кластерів за допомогою класичного методу Монте-Карло [6]. Для перевірки визначення точності кластеризації, розглянемо наступну процедуру утворення матриці суміжності для графа

$$A = D + Error,$$

де $D$ – діагональна матриця

$$D = \begin{pmatrix} D_1 & 0 & 0 & \dots & 0 \\ 0 & D_2 & 0 & \dots & 0 \\ 0 & 0 & D_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & D_k \end{pmatrix}.$$

Матриці $D_i$ є матрицями розмірності $n_i \times n_i$, які визначають кластери в моделі. Матриця $Error$ – матриця розмірності $N \times N$, що задає міжкластерні зв'язки. Зрозуміло, що при умові $Error = 0$, кількість власних значень, які дорівнюють 1 з імовірністю близькою до 1, близька до $k$. Розглянемо випадок, коли матриця помилок $Error$ відмінна від 0. Для моделювання матриць $D$ та $Error$ будемо використовувати наступні параметри:

1. Кількість кластерів – $k$.

2. Рівномірний дискретний розподіл $Ud\,(n_a, n_b)$ для моделювання розмірності кластерів, а саме для моделювання $n_i$.

3. Рівномірний дискретний розподіл $Ud\,(d_a, d_b)$ для моделювання ненульових елементів матриці $D$.

4. Рівномірний дискретний розподіл $Ud\,(e_a, e_b)$ для моделювання ненульових елементів матриці $D$.

Таким чином, модель залежить від 7 гіперпараметрів

$$par = (k,\, n_a, n_b, d_a, d_b, e_a, e_b)\,.$$

Для порівняння отриманих у даній статті результатів (наш метод) використаємо марковський алгоритм, описаний в роботі [7], та метод ліктя [8]. Для моделювання методом Монте-Карло будемо використовувати $N = 10^5$ ітерацій.

Використовуючи наведений вище алгоритм для кластеризації на графах, кількість кластерів будемо визначати наступним чином

$$k_{opt} = \#\left\{ i : |\lambda_i\,(P) - 1| < 1 - \frac{2}{\sqrt{N}} \right\}.$$

Результати оцінки кластерів за трьома алгоритмами відображені в таблиці 1.

| Гіперпараметри $(k,\, n_a, n_b, d_a, d_b, e_a, e_b)$ | Марковський алгоритм [7] | | Метод ліктя | Наш метод |
|---|---|---|---|---|
| | $r = 2$ $s = 2$ | $r = 5$ $s = 2$ | | |
| $(15, 10, 20, 4, 5, 0, 1)$ | 40 | 19 | 12 | 15 |
| $(15, 10, 20, 10, 20, 2, 5)$ | 43 | 22 | 11 | 15 |
| $(25, 30, 50, 10, 20, 2, 5)$ | 67 | 34 | 24 | 25 |
| $(40, 4, 20, 10, 20, 2, 5)$ | 100 | 51 | 34 | 38 |
| $(15, 10, 20, 10, 20, 2, 5)$ | 42 | 28 | 14 | 15 |
| $(15, 10, 20, 10, 20, 8, 15)$ | 36 | 23 | 13 | 15 |
| $(50, 10, 20, 10, 20, 8, 15)$ | 115 | 54 | 53 | 46 |
| $(20, 10, 20, 10, 20, 8, 15)$ | 53 | 35 | 23 | 20 |
| $(20, 10, 20, 10, 20, 1, 15)$ | 32 | 39 | 17 | 20 |

Таб. 1. Результати визначення оптимального числа кластерів $k_{opt}$

Як ми бачимо із таблиці 1, запропонований в роботі метод дозволяє найточніше визначати оптимальну кількість кластерів, базуючись на стохастичних матрицях графів. Даний факт не є дивним, оскільки дійсні частини власних значень матриці $P$ розбиваються на три групи:

1. Власні значення, дійсні частини яких близькі до 0 та не рівні 1. Дані власні значення не є визначальними для кластеризиції, оскільки відображають лише однорідність елементів матриці. Однорідність елементів в свою чергу згідно з теоремою 2 свідчить про наявність одного кластера.

2. Власні значення, дійсні частини яких "відмінні' від 0. Дані власні значення є визначальними для кластеризації, оскільки саме на їх кількості і ґрунтується визначення оптимальної кількості кластерів $k_{opt}$.

3. Власне значення $\lambda = 1$.

На рис. 4-6 зображено групування власних значень в методі Монте – Карло для різних значень параметрів. Як можна бачити із даних рисунків, при помірній величині шуму власні значення із другої групи чітко відокремлені від власних значень із першої групи. Це й дозволяє точно визначати кількість кластерів в даних випадках. Крім того, власні значення із першої групи розподілені згідно з напівколовим розподілом Вігнера з параметром $\frac{1}{\sqrt{N}}$, що узгоджується із результатами теореми 2. Зауважимо, що матриця шумів $Error$ визначається величиною шуму, тобто дисперсією шумів. Використовуючи дану аналогію, аналогічно до теорії сигналів з гаусовими шумами [1], можна визначити показник

$$SNR = \frac{\max\{\sigma_i^2\}}{\sigma^2},$$

де $\sigma_i^2$ – дисперсія елементів в матриці $D_i$, $\sigma^2$ – дисперсія елементів матриці шумів $Error$. Назва показника $SNR$ (Signal to Noise Ratio) має аналогіну назву і при дослідженні

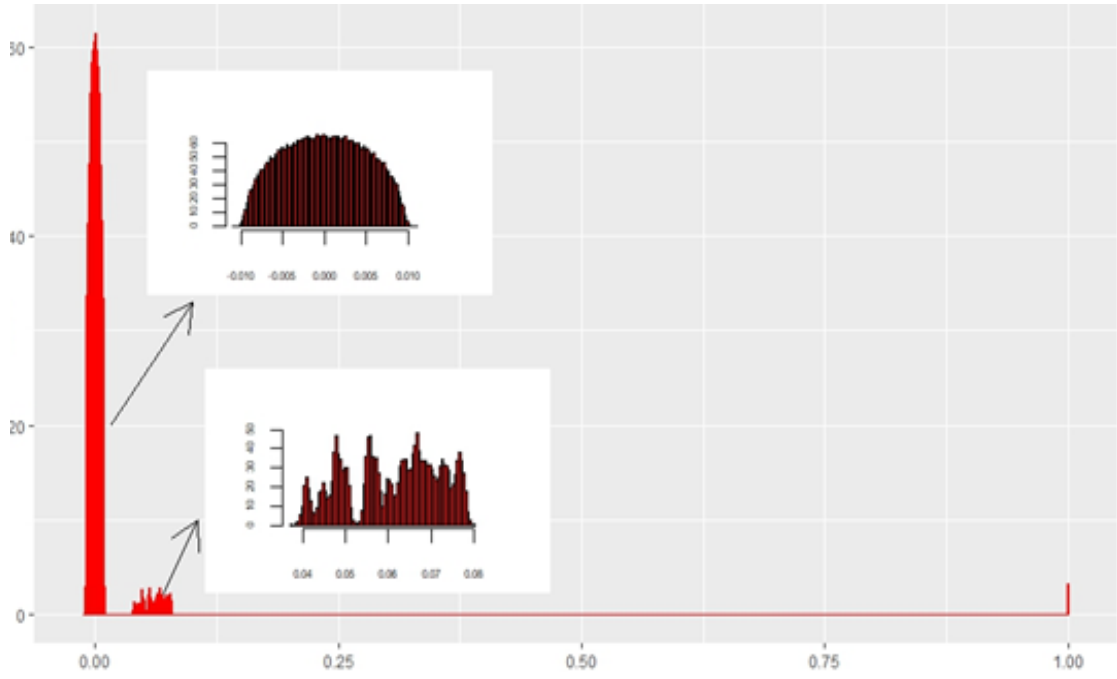


Рис. 4: Розподіл дійсних частин власних значень матриці $P$ з параметрами $par = (20, 10, 20, 10, 8, 15)$

сигналів.

У випадку сильного зашумлення даних (рис. 6) власні значення із груп 1 та 2 об'єднуються. У цьому випадку практично неможливо визначити точне значення $k_{opt}$. Проте слід відмітити, що розподіл власних значень із групи 1 у даному випадку має статистично додатні непарні моменти. У випадках, коли вплив зашумлення є меншим, асиметрія дійсних частин власних значень із групи 1 рівна 0. Даний факт може слугувати відправною точкою для подальшого аналізу значення $k_{opt}$ у випадку сильної зашумленості даних.

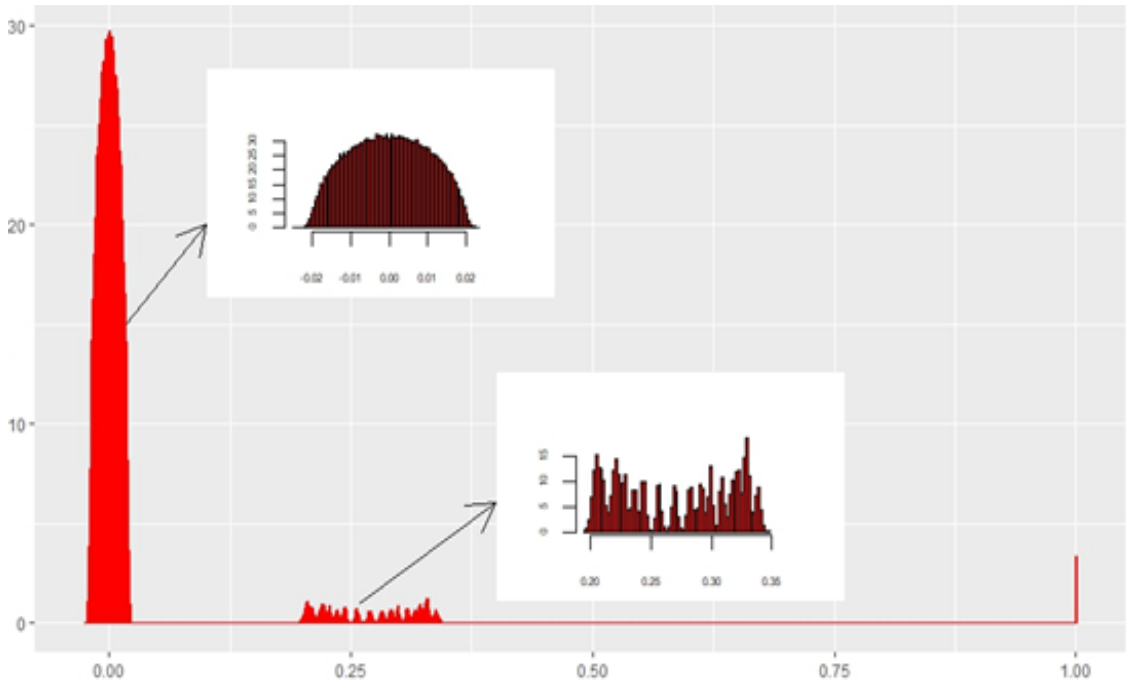


Рис. 5: Розподіл дійсних частин власних значень матриці $P$ з параметрами $par = (20, 10, 20, 10, 1, 3)$

## Аналіз акцій S&P500

Розглянемо використання методики, описаної в даній роботі, для кластеризації акцій $N = 470$ компаній S&P500 даних, фіксованих в період 2013 – 2018 років. Для аналізу розглянуто $M = 1249$ денних показників вартості акцій при закритті бірж (close price). Власні значення стохастичної матриці зображено на рис. 7. Аналогічно роботі [4], розглянемо акції із 10 секторів економіки, тобто реальна кількість кластерів рівна 10 для даного набору даних. Аналогічно більшості робіт з аналізу часових рядів вартості, використаємо відсоткові ставки акцій, які визначаються наступним чином

$$r_{i,t} = \log\left(\frac{S_{i,t}}{S_{i,t-1}}\right),$$

де $r_{i,t}$ – відсоткова ставка для $i$-ї акції в період $t$, $S_{i,t}$ – вартість $i$-ї акції в період $t$, $i = 1, ..., N$, $t = 2, ..., M$. При аналізі кількості кластерів на основі методу, запропонованого в даній роботі, було виявлено, що оптимальна кількість кластерів для даного набору даних рівна

$$k_{opt} = 5.$$

Таким чином, оптимальна кількість кластерів майже вдвічі менша ніж кількість галузей економіки. Аналогічний факт для $N = 400$ акцій S&P500 в період 2002 – 2004 років з $M = 44720$ відмічений також авторами роботи [4], де зазначено, що реальна кількість груп (кластерів) із різною поведінкою часових рядів приблизно вдвічі менша за відповідну кількість галузей економіки. Проте в роботі [4] не вказано точного критерію визначення $k_{opt}$, оскільки проводиться спектральне дослідження графа на основі матриці суміжності $A$, що не дозволяє зробити висновків щодо кількості кластерів.

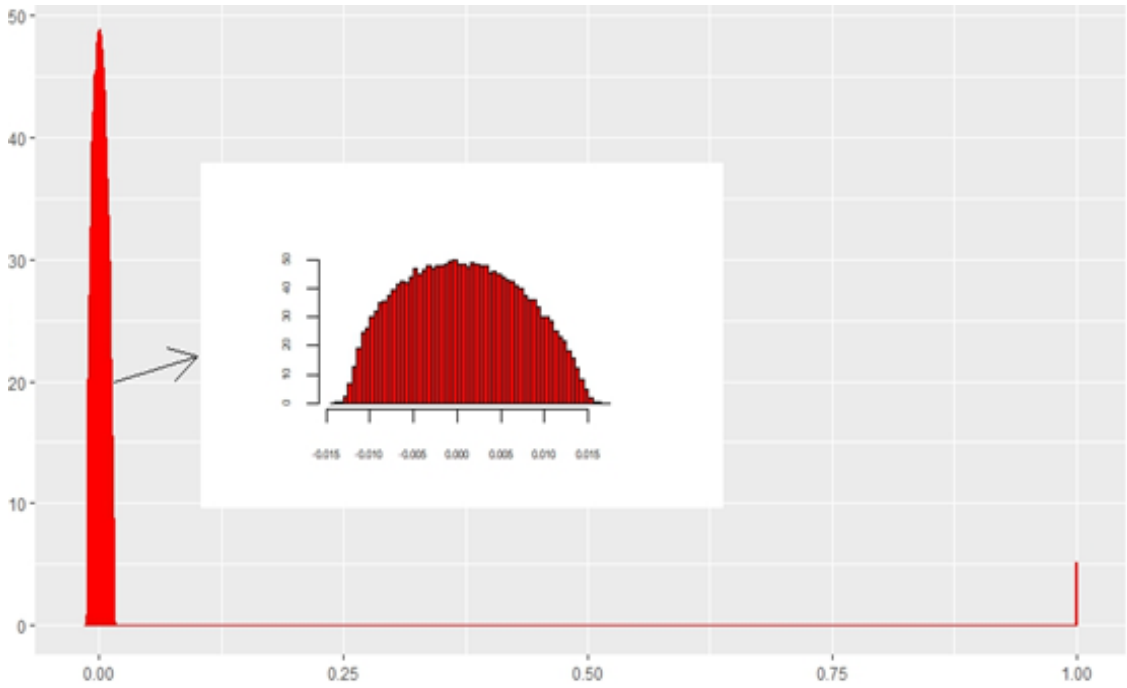


Рис. 6: Розподіл дійсних частин власних значень матриці $P$ з параметрами $par = (20, 10, 20, 10, 80, 150)$

Значення $k_{opt} = 5$ означає, що деякі сектори економіки є дуже пов'язані і не можуть бути відслідковані за допомогою вартостей акцій компаній із даних галузей. Крім того, вплив "міжгалузевих" компаній[1] також призводить до зменшення числа кластерів $k_{opt}$ для даного набору даних.

## Висновки

У роботі запропоновано новий метод визначення оптимальної кількості кластерів $k_{opt}$ при кластеризації об'єктів, що задаються неструктурованими даними (графами та часовими рядами) на основі спектрального аналізу стохастичної матриці даного графу. Використовуючи метод Монте-Карло, вдалося показати, що запропонований метод дає кращі результати для визначення оптимальної кількості кластерів $k_{opt}$ у порівнянні із деякими класичними методи. Оскільки даний алгоритм є спектральним, то його складність співпадає із складністю знаходження власних значень для матриці $P$. Ще однією перевагою даного алгоритму є те, що він не є чутливим до кластерів різного розміру, тобто співвідношення між розмірами кластерів практично не впливають на точність алгоритму.

Теоретичні результати роботи було перевірено для $N = 470$ акцій S&P500, розглянутих в період за 2013 – 2018 роки. Результати оцінки оптимального значення $k_{opt}$ співпали із відповідними оцінками для даних компаній в інший період часу.

---

Наприклад, компанія Apple Inc є міжгалузевою компанією, оскільки реалізує свої товари на ринках різних галузей.

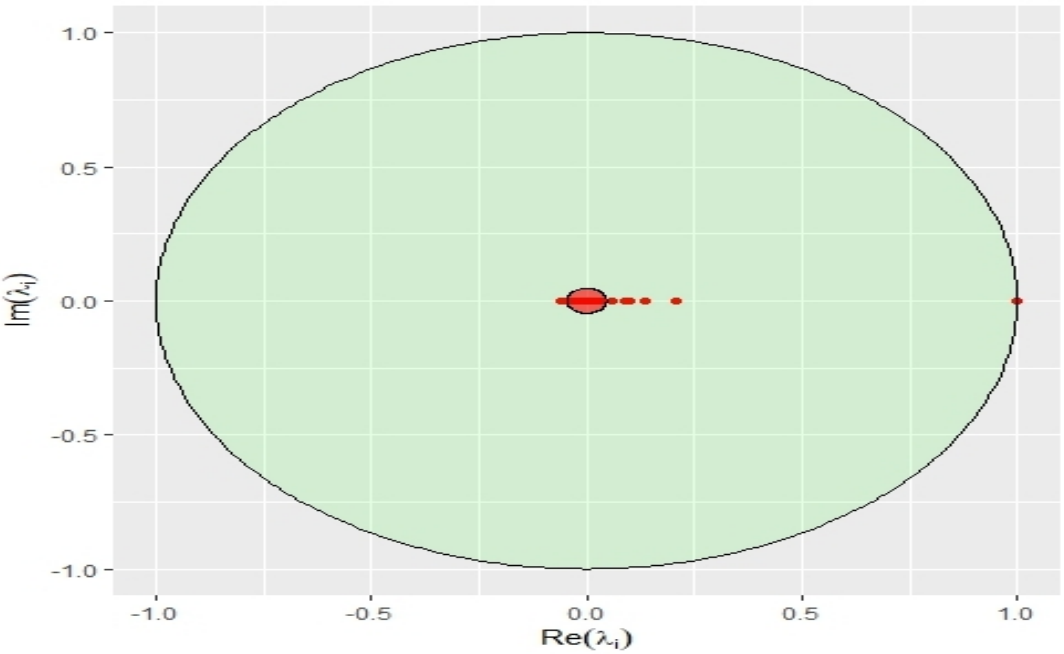


Рис. 7: Власні значення стохастичної матриці, побудованої для $N = 470$ акцій S&P500 в період 2013 – 2018 років з $M = 1249$ показниками.

## Список літератури


[1] Bagherikaram G., Motahari A., Khandani A. *The Secrecy Capacity Region of the Gaussian MIMO Broadcast Channel.* Information Theory 2013, **59**, 2673-2682. doi: 10.1109/TIT.2012.2236972.

[2] Bai Z., Fang Z., Liang Y. Spectral Theory of Large Dimensional Random Matrices and Its Applications to Wireless Communications and Finance Statistics : Random Matrix Theory and Its Applications, World Scientific Publishing Company, 2014.

[3] Bender E.A., Williamson S.G. Lists, Decisions and Graphs – With an Introduction to Probability. University of California at San Diego, San Diego, 2010.

[4] Biely C., Thurner S. *Random matrix ensembles of time-lagged correlation matrices: Derivation of eigenvalue spectra and analysis of financial time-series.* Quantitative Finance 2008. **8**, 705–722. doi: 10.1080/14697680701691477.

[5] Bolch G., Greiner S., Meer H., Trivedi K. *Queueing Networks and Markov Chains.* John Wiley, 2nd edition 2006, **40** (5), 567–568. doi:10.1080/07408170701623187.

[6] Crisan D., Del Moral P., Lyons T. *Discrete filtering using branching and interacting particle systems.* Markov Processes and Related Fields 1999, **5** (4), 293–318.

[7] Dongen S. Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, 2000.

[8] Girvan M., Newman M. *Community structure in social and biological networks.* Proc. Natl Acad. Sci. USA 2002, **99** (12), 7821–7826. doi: 10.1073/pnas.122653799.

[9] Qiu R., Antonik P. Smart Grid using Big Data Analytics. A Random Matrix Theory Approach. Wiley Online Library, 2017.

[10] Knignitska T.V. *Estimate of time series similarity based on models.* Problems of Control and Informatics 2019, **51** (4), 94–104.

[11] Marchenko V.A., Pastur L.A. *Distribution of eigenvalues for some sets of random matrices.* Mat. Sb. N.S. 1967, **72(114)** (4), 507-–536. doi:10.1070/SM1967v001n04ABEH001994

*Надійшло 04.12.2019*

---

Malyk I.V., Knignitska T.V., Gorbatenko M. Yu. *Clustering: Markov algorithm*, Bukovinian Math. Journal. **7**, 2 (2019), 59–75.

This paper deals with the problem of clustering on graphs based on the eigenvalues of the stochastic matrix of the graph. It is proved, that the eigenvalues of the stochastic matrix for large graphs ($N > 100$) are divided into three groups, one of which is decisive for the number of clusters in the graph. Using the theory of random matrices, it was possible to show, that the asymptotic distribution of the subgroup of the real parts of the eigenvalues of the stochastic matrix of the graph is described by a semicircular Winger distribution, and the parameter of this distribution is $O\left(\frac{1}{\sqrt{N}}\right)$. The use of stochastic matrices made it possible to accurately localize the eigenvalues responsible for the number of clusters, which was not possible for adjacency matrices. The basic assumptions of the model are related to the properties of Markov discrete chains, which allows to extend the scope of the obtained results to a wider class of objects. Theoretical results were verified by clustering time series describing the value of $N = 470$ shares of S&P500 data for the 6 years period (2013 – 2018). The clustering of these time series showed the results of the presence of 5 clearly defined groups, consistent with the data used. A covariance matrix with zeros on the diagonal elements was used to construct a stochastic matrix for time series. This approach made it possible to localize the main clusters more precisely. The main result of the work is devoted to asymmetric matrices with mathematical expectations not equal 0, which allowed to generalize some results of Big Data theory. Also, the dependence of clustering results on the cluster size, the number of clusters and the asymmetry of cluster size is noted in the paper. Using the Monte-Carlo method, it is shown that the proposed Markov algorithm is more stable to noise in the graph than some of classical algorithms.