

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

імені ІГОРЯ СІКОРСЬКОГО»

НАВЧАЛЬНО-НАУКОВИЙ

ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

Кафедра математичних методів аналізу даних

«До захисту допущено»

завідувача кафедри

_____ Наталія КУССУЛЬ

«__» _____ 2022 р.

Дипломна робота

на здобуття ступеня бакалавра

зі спеціальності: 113 Прикладна математика

на тему: «Соціальні мережі. Коефіцієнти мережевого впливу»

Виконала: студентка 4 курсу, групи ФІ-81.
(шифр групи)

Чугай Дар'я Володимирівна
(прізвище, ім'я, по батькові) _____ (підпис)

Керівник: старший викладач кафедри ММАД Тітков Д.В.
(посада, науковий ступінь, вчене звання, прізвище та ініціали) _____ (підпис)

Рецензент _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) _____ (підпис)

Засвідчую, що в цій дипломній роботі
немає запозичень із праць інших
авторів без відповідних посилань.

Студент _____
(підпис)

Київ — 2022 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

Кафедра математичних методів аналізу даних

Рівень вищої освіти — перший (бакалаврський)

Спеціальність (освітня програма) — 113 Прикладна математика

(«Математичні методи комп'ютерного моделювання, розпізнавання образів та безпеки даних»)

ЗАТВЕРДЖУЮ

В. о. завідувача кафедри

_____ Наталія

КУССУЛЬ

(підпис)

«___» _____ 2022 р.

ЗАВДАННЯ
на дипломну роботу студенту

Чугай Дар'ї Володимирівні

(прізвище, ім'я, по батькові)

1. Тема роботи: Соціальні мережі. Коефіцієнти мережевого впливу.

Науковий керівник роботи: старший викладач кафедри ММАД Тітков Д.В.,
консультант Смирнов С.А.

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «___» _____ 2022 р. №

2. Термін подання студентом роботи ___ червня 2022 р.

3. Вихідні дані до роботи

4.Зміст роботи:

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо)

Соціальні мережі. Коефіцієнти мережевого впливу — презентація.

6. Дата видачі завдання 25.03.2022

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів дипломної роботи	Примітка
1	Отримання завдання	25.03.2022	
2	Розробка плану роботи	30.03.2022	
3	Дослідження предметної області та існуючих рішень	10.04.2022	
4	Збір інформації	20.04.2022	
5	Огляд основних положень моделювання соціальних мереж	05.05.2022	
6	Пошук методів опису та прогнозування поширення впливу в соціальних мережах	15.05.2022	
7	Пошук методів вирішення недоліків алгоритмів та моделювання соціальних мереж	25.05.2022	
8	Огляд результатів	05.06.2022	
9	Оформлення дипломної роботи	14.06.2022	
10	Отримання допуску до захисту	15.06.2022	

Студент _____ Д. В. Чугай
(підпис) (ініціали, прізвище)

Керівник роботи _____ Д. В. Тітков
(підпис) (ініціали, прізвище)

РЕФЕРАТ

Ця робота містить 35 сторінок, 6 ілюстрацій, 15 джерел.

У цій роботі йдеться про математичне представлення поширення впливу в соціальних мережах. Змістовно її можна поділити на 3 частини. У першій буде введено поняття індексів впливу та представлено їхні види. У другій частині статті йдеться про важливі алгоритми дослідження соціальних мереж: лувенський метод та PageRank. Третій розділ буде заключним і розкриє типи соціальних мереж, їхні особливості й математичний підхід для їхнього представлення.

Ключові слова: коаліція, коаліція, що виграє, ключовий учасник, Алгоритм PageRank, Лувенський метод, SNA (social networks analysis), силовий алгоритм, індекс впливу Банцафа, індекс впливу Шеплі-Шубіка, індекс впливу Джонстона, індекс впливу Дигена-Пакела, індекс впливу Холлера-Пакела

ABSTRACT

This work contains 35 pages, 6 illustrations, 15 sources.

This paper deals with the mathematical representation of the spread of influence in social networks. Its content can be divided into 3 parts. The first will introduce the concept of impact indices and present their types. The second part of the article deals with important algorithms for researching social networks: the Louvain method and PageRank. The third section will be final and will reveal the types of social networks, their features and mathematical approach to their presentation.

Keywords: PageRank algorithm, Louvain Method, SNA, force-directed algorithm, Deegan-Packel power index, Johnston power index, Shapley-Shubik power index, Banzhaf, power index

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів	5
Вступ	6
Основна частина	7
Коефіцієнти мережевого впливу	7
1.1. Коефіцієнт Банцафа (Banzhaf power index)	7
1.2. Коефіцієнт Шаплі-Шубіка (Shapley-Shubik power index)	8
1.3. Коефіцієнт Джонстона (Johnston power index)	9
1.4. Коефіцієнт впливу Дигена-Пакела (Deegan-Packel)	10
1.5. Коефіцієнт Холлера-Пакела (Holler-Packel power index)	10
Висновки до розділу	11
2. Алгоритми дослідження соціальних мереж	12
2.1 SNA	12
2.2 Лувенський метод	12
2.3 PageRank	18
3. Моделі соціальних мереж	22
3.1. Класифікація моделей впливу в соціальних мережах.	22
3.2. Стохастичні моделі в соціальних мережах.	25
3.3. Моделі інформаційного керування й інформаційного протистояння в соціальних мережах.	26
Висновки	28
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ	29

Перелік умовних позначень, символів, одиниць, скорочень і термінів

Коаліція S - підмножина множини учасників, що ухвалюють рішення.

Квота - кількість голосів, необхідна для прийняття рішення

Коаліція, що перемагає - коаліція, якої вистачає для перевищення квоти

Ключовий учасник - особа, що бере участь у голосуванні, і якщо вона вийде з коаліції, що перемагає, то та перетворюється на коаліцію, що програє.

Агенти — суб'єкти соціальних мереж, що можуть бути окремими індивідами, або групами індивідів, що об'єднані за певною ознакою.

“Лідери думок” - агенти соціальних мереж, що мають найбільший вплив.

Вступ

Сучасні соціальні мережі — складна структура, що містить у собі учасників та зв'язки між ними. Окремою ланкою є онлайніві соціальні мережі, які за останні два десятиліття стали потужним джерелом поширення впливу.

Одним із красномовних підтверджень цього факту є кількість людей, що використовують онлайніві соціальні мережі для пошуку інформації в умовах повномасштабної війни. Так, згідно з нещодавним соціологічним опитуванням [13], проведеним Київським міжнародним інститутом соціології (КМІС) понад 3/4 українців використовують саме їх як джерело новин. І така тенденція спостерігається в багатьох інших країнах Заходу. Саме тому вивчення поширення впливу в соціальних мережах є досить важливим завданням, яке є й буде актуальним у найближчому майбутньому.

Соціальні мережі можна лаконічно описати двома елементами: учасниками та зв'язками між ними. Така характеристика дозволяє нам провести аналогію між соціальними мережами та графами, користувачів представити у вигляді вершин, а зв'язки між ними за допомогою ребер. Це дозволяє перенести деякі властивості графів на соціальні мережі й застосовувати до них відповідні методи.

Окремої уваги заслуговує дослідження зв'язків між членами мережі, їхній вплив одне на одного. Це знаходить відгук зокрема й у теорії ігор, тож часто в джерелах можна натрапити на терміни, що стосуються саме цієї області математики.

Актуальність роботи

Соціальні мережі - явище, яке невинно вдосконалюється та набуває нових форм. Зокрема за останні кілька десятиліть було створено чималу кількість онлайнівих соціальних мереж, що змінили спосіб комунікації людей і поширення впливу в них. Відповідно, оновлення потребують і методи їх дослідження.

Саме тому під час ознайомлення з алгоритмами, що стосуються соціальних мереж, важливо було знайти їхні найсучасніші концепції та

версії. Тому ця робота присвячена знаходженню недоліків у існуючих алгоритмах та запропонуванню методів для їх усунення.

Мета і завдання

Метою дослідження є моделювання соціальних мереж з оглядом на особливості їхньої структури і поширення впливу.

Завдання дослідження полягає у адаптації старих методів до сучасних соціальних мереж. Для цього робота поділена на 2 етапи:

1. Пошук недоліків у старих алгоритмах, пошук їх невідповідності сучасним соціальним мережам.
2. Знаходження покращень у методах, що дозволяють усунути ці проблеми.

Об'єкт дослідження - соціальні мережі

Предмет дослідження - поширення впливу у соціальних мережах

Методи дослідження:

- Знаходження коефіцієнтів мережевого впливу: Банцафа, Шеплі-Шубіка, Джонстона, Дигена-Пакела, Холлера-Пакела. Введення
- Метод аналізу соціальних мереж SNA.
- Алгоритм кластеризації соціальних мереж (Лувенський метод).
- Метод представлення графа, з огляду на ваги його ребер (силовий алгоритм)
- Алгоритм оцінки впливовості вебсторінок PageRank.
- Моделі дифузії інновацій: порогові та каскадні моделі.
- Стохастичні моделі.
- Моделі інформаційного керування і протиборства

Апробація результатів роботи

Результати роботи були представлені на Всеукраїнській науково-практичній конференції студентів, аспірантів та молодих вчених 2022 р.

Публікації

Стислий огляд результатів цієї роботи було оприлюднено у статті
“Моделі впливу у соціальних мережах” [15]

Моделі впливу у соціальних мережах

1. Коефіцієнти мережевого впливу

1.1. Коефіцієнт Банцафа (Banzhaf power index)

Перш ніж ввести коефіцієнт Банцафа, зауважимо, що термін «коаліція» S позначає підмножину множини учасників, що ухвалюють рішення. Нехай ми маємо певне число — квоту q , що якщо кількість голосів k_i (i відповідає номеру учасника голосування) їй дорівнює або її перевищує, то рішення вважатиметься ухваленим, а коаліцію, кількість голосів якої виявилася достатньою для ухвалення рішення називатимемо переможною.

$$\sum_{i \in W} k_i \geq q,$$

де W — множина осіб, що беруть участь у голосуванні

$$\sum_{i \in S} k_i \geq q,$$

де S — множина осіб, що беруть участь у голосуванні

Введемо також індикатор того, що коаліція переможна.

$$\nu(S) = \begin{cases} 1, & \text{якщо } S - \text{коаліція, що виграє} \\ 0 & \text{в іншому випадку} \end{cases}$$

Нехай учасник переможної коаліції буде ключовим, якщо при його виході з неї, вона перестає бути переможною.

Введемо кілька означень [2]:

1. Диктатором називається учасник голосування, якщо він один уже складає переможну коаліцію.
2. Блокуючим є учасник, який входить у всі переможні коаліції.

Тоді індекс Банцафа знаходимо як відношення кількості коаліцій, у яких i учасник є ключовим, до суми кількості коаліцій з усіма ключовими учасниками.

$$\beta(i) = \frac{b_i}{\sum_j b_j}$$

1.2. Коефіцієнт Шаплі-Шубіка (Shapley-Shubik power index)

Цей коефіцієнт дуже схожий на індекс Банцафа, втім його відмінність полягає в тому, що тепер у голосуванні беруть участь послідовні коаліції (тобто тепер нам важливо знати у якому порядку учасники входять до коаліцій). До того ж слід зауважити, що до послідовних коаліцій входять усі учасники голосування (але в різному порядку) [9]. Сенс цього коефіцієнта впливу полягає в тому, щоб оцінити наскільки кожен учасник може бути ключовим.

Знаходимо коефіцієнт Шаплі-Шубіка в кілька кроків.

1. У кожній послідовній коаліції знаходимо ключову особу
2. Порахувати, скільки разів кожна особа буває ключовою.
3. Знаходимо відношення кількості разів того, що учасник є ключовим до кількості всіх послідовних коаліцій.

Нехай n — кількість учасників голосування. Оскільки, усі послідовні коаліції мають містити всіх учасників голосування, які, до того ж, не можуть поворюватися, то кількість усіх послідовних коаліцій знайдемо як перестановку з n учасників :

$$P_n = n \cdot (n - 1) \cdot \dots \cdot 2 \cdot 1 = n!$$

Тоді коефіцієнт Шаплі-Шубіка набуває такого вигляду:

$$\sigma(i) = \frac{b_i}{n!},$$

де b_i — кількість разів того, що учасник є ключовим

1.3. Коефіцієнт Джонстона (Johnston power index)

Відмінність коефіцієнту Джонстона від коефіцієнту Банцафа полягає в тому, що він оцінює впливовість учасника голосування з огляду на всі ключові партії, що присутні в коаліції.

Для обчислення коефіцієнту Джонстона нам необхідно визначити:

1. Для кожного учасника p_i усі коаліції $C_1, C_2, \dots, C_k$, у яких p_i є ключовим.
2. Кількість ключових учасників $n_1, n_2, \dots, n_k$ у $C_1, C_2, \dots, C_k$.

Тоді можемо знайти загальний коефіцієнт Джонстона для учасника p_i за такою формулою:

$$TJI(p_i) = \frac{1}{n_1} + \frac{1}{n_2} + \dots + \frac{1}{n_k}$$

Потім ми знаходимо суму загальних коефіцієнтів Джонстона для всіх ключових учасників p_i

$$\sum_j TJI(p_j)$$

Зрештою коефіцієнт впливу Джонстона для учасника p_i ми знаходимо як відношення загального коефіцієнту Джонстона для учасника p_i до суми загальних коефіцієнтів Джонстона для всіх ключових учасників p_i

$$JI(p_i) = \frac{TJI(p_i)}{\sum_j TJI(p_j)}^1$$

Зауваження: якщо учасник не є ключовим у жодній із коаліцій, то його коефіцієнт Джонстона дорівнює 0.

1.4. Коефіцієнт впливу Дигена-Пакела (Deegan-Packel)

Для визначення коефіцієнту впливу Дигена-Пакела потрібно ввести поняття мінімальної коаліції, що виграє.

¹
[1, 10]

Коаліція називається такою, що мінімально виграє, якщо вона стає такою, що програє після усунення з неї будь-якого учасника.

Для знаходження коефіцієнту впливу Дигана-Пакела нам потрібно знайти:

1. Усі коаліції, що мінімально виграють $C_1, C_2, \dots, C_k$, у яких перебуває учасник p_i
2. Кількість учасників $m_1, m_2, \dots, m_k$ у коаліціях $C_1, C_2, \dots, C_k$.

Тоді можемо знайти загальний коефіцієнт Дигана-Пакела для учасника p_i

$$TDPI(p_i) = \frac{1}{m_1} + \frac{1}{m_2} + \dots + \frac{1}{m_k}$$

Аналогічно до знаходження коефіцієнта впливу Джонстона, наступним кроком стане обчислення суми загальних коефіцієнтів Дигана-Пакела

$$\sum_j TDPI(p_j)$$

і, власне, знаходження самого коефіцієнта впливу Дигана-Пакела як відношення загального коефіцієнту Дигана-Пакела для учасника p_i до суми загальних коефіцієнтів Дигана-Пакела для всіх ключових учасників p_i

$$DPI(p_i) = \frac{TDPI(p_i)}{\sum_j TDPI(p_j)}$$

1.5. Коефіцієнт Холлера-Пакела (Holler-Packel power index)

Коефіцієнт впливу Холлера-Пакела має в собі ознаки коефіцієнтів впливу Банцафа (знаходимо коаліції, у яких учасник є ключовим) і Дигана-Пакела (берем до уваги тільки коаліції, що мінімально виграють).

$$HPI(p_i) = \frac{h_i}{\sum_j h_j}, \text{ де}$$

h_i — кількість коаліцій, що мінімально виграють, і до яких входить учасник i

Висновки до розділу

Були описані основні коефіцієнти мережевого впливу. Головним недоліком усіх описаних коефіцієнтів є те, що вони ігнорують відмінність різних учасників у інтенсивності надання переваги щодо створення коаліцій. Цю хибу можна усунути, якщо ввести певні статистичні характеристики голосування. [1]

Одним із прикладів може послугувати введення функції інтенсивності $f(i, S)$ зв'язку учасника i з коаліцією S .

$$\chi_i = \sum f(i, S)$$

Тоді коефіцієнт впливу α описуємо за допомогою такої формули

$$\alpha(i) = \frac{\chi_i}{\sum_j \chi_j}$$

2. Алгоритми дослідження соціальних мереж

Дослідження соціальних мереж тісно пов'язане з теорією графів, адже саме графове представлення об'єктів та зв'язків між ними щонайкраще підходить для представлення таких моделей. Зокрема це дає змогу застосовувати деякі алгоритми притаманні графам.

2.1 SNA

SNA (social network analysis) — це метод, що допомагає наочно зобразити об'єкти та співвідношення з урахуванням сили (важливості цих зв'язків) за допомогою графової інтерпретації моделей. Зазвичай, вершини таких графів позначають людей, а ребра — стосунки між ними.

Для опису зв'язків у цьому методі існує 4 метрики (характеристики) [4]:

1. мультиплексність (multiplexity) — кількість окремих зв'язків між об'єктами
2. гомофілія (homophily) — ступінь здатності утворювати зв'язки з іншими об'єктами
3. взаємодія (reciprocity) — ступінь співдії об'єктів

Силовий алгоритм (force-directed algorithm) допомагає відобразити силу зв'язків між об'єктами в зручнішому форматі. Його ідея полягає в тому, що об'єкти без зв'язків відштовхуються, а зі зв'язками — притягуються на відстань, що відповідає вазі ребра між ними [5].

2.2 Лувенський метод

Для кластеризації зв'язків використовують лувенський метод (Louvain Method). Це жадібний алгоритм навчання без учителя. Мета Лувенського методу — розділити мережу на кластери (групи) за допомогою збільшення модулярності кожного кластеру. Це означає, що він максимізує (жадібно) кількість та силу зв'язків всередині кластеру й мінімізує зв'язки з іншими кластерами. Слід зауважити, що цей метод дозволяє розпізнати ієрархічні групи (тобто групи з груп). Алгоритм

закінчує свою роботу, коли модулярність максимізована, а нових змін у мережі не спостерігається.

Розглянемо детальніше кроки алгоритму. Спершу визначимо за якою формулою ми обчислюємо модулярність. Для її знаходження ми маємо побудувати випадковий граф для того, щоби знайти очікувану кількість ребер.

Q — це значення, яке ми прагнемо максимізувати. Лувенський метод робить це в кілька етапів. На етапі ініціалізації ми маємо мережу, представлену графом із послідовністю вершин. Вважаємо, що кожен вузол належить до групи, що представлена цим же вузлом. Далі на кожній ітерації алгоритму ми виконуємо послідовно такі дії:

1. Модулярність оптимізується завдяки можливості локальної зміни належності до груп.
2. Після того, як усі вершини отримали належність до своїх груп, вони об'єднуються у вузли, що формують нову мережу.

Нехай маємо неорієнтований граф $G = (V, E)$ і його матрицю суміжності² $A(G)$. Нехай кількість ребер цієї мережі становить m . Оскільки граф неорієнтований, а кожне ребро виходить із якоїсь вершини та входить в іншу, то кількість таких «півребер» буде парною й дорівнюватиме $2m$. Тепер нехай ми можемо з'єднувати ці півребра випадковим чином, щоб утворити випадковий граф. Нехай із вузла i виходить k_i півребер, а з j — k_j . Порахуємо ймовірність того, що з'єднані два вузли i та j . Отже, ми перебираємо всі можливі ребра k_i для i та k_j для j , але оскільки i -тий вузол ми можемо обрати серед усіх вузлів цього графу (їхня кількість дорівнює $2m$), то для j -го вузла в нас залишається тільки $2m-1$ варіантів. Відповідно так виглядає ймовірність того, що між i та j існує ребро

$$\frac{k_i k_j}{2m - 1}$$

Якщо число m дуже велике, то можемо зробити спрощення в знаменнику

$$p_{ij} = \frac{k_i k_j}{2m}$$

Це і є очікуваною кількістю ребер між вузлами i та j .

² Матриця суміжності - квадратна матриця розміру $n \times n$, де n - кількість вершин графа, елементами якої є такі числа a_{ij} , що відповідають кількості ребер, що виходять з i -ої вершини до j -ої.

Слід зауважити, що ми запобігаємо утворенню кратних ребер.

Далі знаходимо різницю між реальною кількістю ребер та очікуваною

$$A_{ij} - p_{ij}$$

Тепер просумуємо по всіх парах i та j і поділимо цю суму на $2m$.

Тоді модулярність знаходимо як

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - p_{ij}] \delta(c_i, c_j)$$

де c_i та c_j позначають групи (community),

$\delta(c_i, c_j)$ — дельта Кронекера, тобто функція, що дорівнює 1, коли $c_i = c_j$ і 0, коли $c_i \neq c_j$.

Тепер внесемо знак суми в дужки

$$\begin{aligned} Q(C) &= \frac{1}{2m} \left[\sum_{i,j \in C} A_{ij} - \sum_{i,j \in C} p_{ij} \right] \\ Q(C) &= \frac{1}{2m} \left[\sum_{i,j \in C} A_{ij} - \frac{\sum_{i \in C} k_i \sum_{j \in C} k_j}{2m} \right] \\ Q(C) &= \frac{\sum_{i,j \in C} A_{ij}}{2m} - \frac{\sum_{i \in C} k_i \sum_{j \in C} k_j}{(2m)^2} \end{aligned}$$

Введемо позначення:

$$\sum_{in} \equiv \sum_{i,j \in C} A_{ij} \quad \text{— сума вагів усіх зв'язків між вузлами всередині групи}$$

C

$$\sum_{tot} \equiv \sum_{i \in C} k_i \sum_{j \in C} k_j \quad \text{— сума вагів усіх зв'язків між вузлами групи C}$$

Отже, отримаємо вираз

$$Q(C) = \frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2$$

Отже, з цієї формули бачимо, що максимальної модулярності можна досягти, якщо відповідна різниця буде найбільшою. Оскільки в першому доданку всі зв'язки всередині групи C , а в другому ще й зовнішні, тож для максимізації цього виразу, потрібно максимізувати зв'язки всередині групи й мінімізувати їх зовні.

Введемо подальші позначення:

$k_{i,in} \equiv \sum_{i \in C} A_{ij} + \sum_{j \in C} A_{ij}$ — сума вагів усіх зв'язків, що сполучають вузол i з групою C ;

k_i — сума вагів усіх зв'язків вузла i .

Залишилося ввести ще одну величину ΔM . Вона позначає зміну модулярності через перехід вузла від однієї групи до іншої.

Нехай маємо вузол i , що належить групі D і хочемо, щоби він належав групі C . Тоді зміну модулярності для вузла i можемо записати так:

$$\Delta Q(D \rightarrow i \rightarrow C) = \Delta Q(D \rightarrow i) + \Delta Q(i \rightarrow C)$$

Цю формулу можна інтерпретувати так: як зміниться модулярність, якщо ізолювати вузол i від групи D , і як зміниться модулярність, якщо приєднати вузол i до групи C .

Порахуємо модулярність вузла i (який поки не належить ні до групи C , ні до групи D до з'єднання з групою C).

$$Q_{before} = Q(C) + Q(\{i\})$$

Підставимо знайдене значення $Q(C)$ у вираз:

$$Q_{before} = \frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 + Q(\{i\})$$

Тепер знайдемо $Q(\{i\})$

Оскільки вузол i не має зв'язків сам із собою, то $k_{i,in} = 0$. Відповідно, перший доданок теж $= 0$. Тоді залишається значення k_i , що і є сумою всіх зв'язків, що виходять із вузла i . Тоді

$$Q(\{i\}) = 0 - \left(\frac{k_i}{2m}\right)^2$$

Підставимо знайдені вирази у Q_{before}

$$Q_{before} = \frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 + \left(\frac{k_i}{2m}\right)^2$$

Поглянемо на зміну модулярності через приєднання вузла i до групи C

$$Q_{after} = Q(C + \{i\})$$

$$Q_{after} = \frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2$$

Зауваження! Тут $k_{i,in}$ не дорівнює 0, оскільки є сумою зв'язків із групою C .

Отже, зміна модулярності від переходу i від ізоляції до групи C набуває такого вигляду:

$$\Delta Q = Q_{after} - Q_{before}$$

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 + \left(\frac{k_i}{2m}\right)^2 \right]$$

Це і є значення, яке ми будемо перевіряти для того, щоби максимізувати модулярність. Жадібність Лувенського методу полягає в тому, що на кожному кроці він рахує модулярність для груп і визначає серед них найбільшу.

Проілюструємо роботу алгоритму на прикладі такого графу:

1. Ініціалізація (рис. 2.1). Маємо граф із послідовністю вершин. Кожен вузол належить групі, що складається з цієї ж вершини.

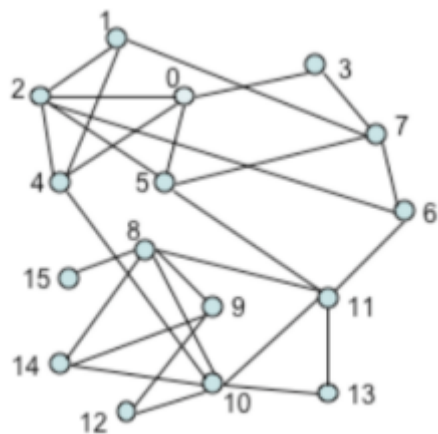


Рис 2.1 Ініціалізація [6]

2. 1 крок (рис. 2.2). Знаходимо й надаємо вузлам кластери на підставі оптимальної модулярності.

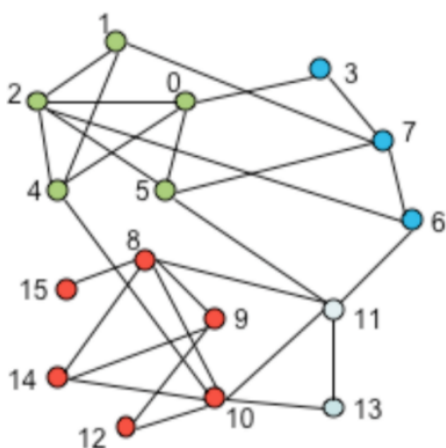


Рис 2.2 Крок 1: поділ на кластери [6]

3. Вершини, що перебували в одному кластері, об'єднуємо в один вузол. Утворюється нова мережа (рис. 2.3.).

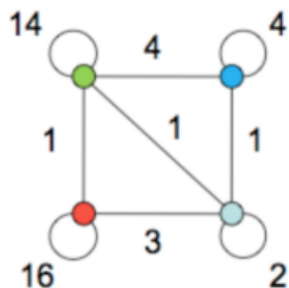


Рис 2.3 Крок 2: об'єднання елементів кластерів [6]

4. Повторюємо попередні пункти й отримаємо мережу з максимізованою модулярністю (рис. 2.4)



Рис 2.4 Результат роботи алгоритму [6]

Переваги алгоритму:

- Підходить для застосування на великих мережах оскільки забезпечує швидке виконання
- На виході отримуємо високий ступінь модулярності
- Забезпечує швидку збіжність
- Підходить для зважених графів
- Розпізнає ієрархію груп

Недоліки алгоритму:

- У результаті роботи цього методу часом залишаються найбільші кластери, а важливі, проте маленькі кластери можуть бути поглинені іншими в процесі роботи алгоритму. Їх можна знайти на проміжних ітераціях.
- Для ієрархічних кластерів важко визначити оптимальну кількість груп.

2.3 PageRank

Цей алгоритм особливий тим, що його використовує один із найпопулярніших сучасних пошукових сервісів Google search.

Для опису алгоритму введемо кілька понять. Нехай сторінки мережі будуть представлені у вигляді вузлів графа, а гіперпосилання між ними будуть ребрами між відповідними вузлами. Вважатимемо, що більше посилань має сторінка, то впливовою вона є. Також посилання з важливіших сайтів мають більшу вагу. Тоді ідея алгоритму полягає в тому, щоби визначити найвпливовіші вебсторінки з огляду на кількість та якість гіперпосилань на них. Інакше кажучи, на виході алгоритму ми отримаємо вектор сайтів зі значеннями їхньої впливовості.

Нехай i -тий сайт має впливовість, яку ми можемо описати числом r_i . Уявимо i -тий сайт має d_i зовнішніх посилань (тобто таких, що виходять із цього сайту на інші).

Припустимо, що ми хочемо знайти впливовість сторінки j . Для цього ми маємо знайти суму всіх вагів ребер, що входять до цього вузла

Вагу одного такого ребра ми знаходимо як відношення значення впливовості сайту, з якого посилання нас відправляє на сайт j до кількості зовнішніх посилань.

$$r_i = \frac{r_i}{d_i}$$

Тоді впливовість сторінки j дорівнює

$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$

Побудуємо для нашої мережі стохастичну матрицю суміжності M . Її елементи знаходимо за формулою (), якщо вузли i та j суміжні:

$$M_{ij} = \frac{1}{d_j}$$

В іншому випадку

$$M_{ij} = 0$$

Тобто для кожного сайту j у стовпчику прописані коефіцієнти ребер, що виходять з j і входять у i .

Зауважимо, що сума всіх коефіцієнтів у кожному стовпчику має дорівнювати 1, бо ми d_j разів додаємо $\frac{1}{d_j}$. Тож цю матрицю можна інтерпретувати як матрицю розподілу ймовірностей суміжних вершин.

Також введемо вектор рангу r , компонентами якого є r_i . Тут також

$$\sum_i r_i = 1$$

Тоді маємо, що

$$\mathbf{r} = M \cdot \mathbf{r}$$

Отже, тепер можемо виписати кроки алгоритму:

1. Ініціалізація. Нехай маємо мережу з N сайтами. Прописуємо вектор $\mathbf{r}$, заповнюючи його елементи випадковими значеннями ймовірностей, наприклад

$$\mathbf{r}^0 = \left(\frac{1}{N}, \frac{1}{N}\right)^T$$

2. Далі ітеративно рахуємо впливовість сайтів за формулою

$$r_j^{t+1} = \sum_{i \rightarrow j} \frac{r_i^t}{d_i}$$

або

$$r^{t+1} = M \cdot r^t$$

3. Задаємо точність ϵ для зупинки алгоритму
Критерій зупинки:

$$\sum_i |r_i^{t+1} - r_i^t| < \epsilon$$

Переваги алгоритму:

- простота використання

Недоліки алгоритму:

- Якщо сторінка з гарним контентом була створена нещодавно й не встигла набрати велику аудиторію, то PageRank не вкаже її серед впливових сайтів. Цей недолік можна прибрати, якщо постійно оновлювати значення PageRank.
- Сторінка посилається сама на себе (spider trap). Тобто виникають ситуації, коли алгоритм може зациклюватися.

Рішення проблеми полягає в тому, що протягом кількох кроків алгоритм заходить на один і той самий сайт, а потім його викидає за

межі цього зациклення на будь-який інший сайт. Цей спосіб можна реалізувати, ввівши додатковий параметр β , що означає ймовірність залишитися на цьому посиланні, а $1 - \beta$ означає ймовірність перейти за іншим посиланням. Зазвичай значення параметру β обирають у межах від 0.8 до 0.9.

Тоді формула для розв'язання цієї задачі набуде такого вигляду:

$$r_j = \sum_{i \rightarrow j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N}$$

або введемо нову матрицю

$$\mathbf{G} = \beta \mathbf{M} + (1 - \beta) \left[\frac{1}{N} \right]_{N \times N}$$

і тоді формула набуде такого вигляду:

$$\mathbf{r} = \mathbf{G} \cdot \mathbf{r}$$

- Є сторінки, з яких не виходять інші посилання (dead ends). Тобто сума коефіцієнтів зовнішніх посилань у стовпчику з таким сайтом у стохастичній матриці суміжності буде дорівнювати 0, що, власне, порушує її стохастичність.

Для цього існує такий підхід: коли ми натрапляємо на сайт, у якого відсутні посилання на інші ресурси, то будемо вважати, що існує однакова ймовірність того, щоби перейти з цього сайту на будь-який інший, що є об'єктом цієї мережі. Відповідно, ми заповнюємо цей стовпчик у матриці однаковими значеннями для відновлення її стохастичності.

Слід зауважити, що введення параметру β також вирішує цей недолік.

Висновки до розділу

У цьому розділі описано метод SNA, що забезпечує представлення соціальних мереж у зручному для математичного дослідження вигляді.

Надалі був розглянутий спосіб кластеризації соціальних мереж - лувенський метод. Проблеми, що можуть виникати під час роботи алгоритму безпосередньо пов'язані з максимізацією модулярності, адже через присутність вкладених кластерів важко визначити їхню оптимальну кількість. Тому цей алгоритм краще застосовувати для мереж з дискретними кластерами. Ще одним суттєвим недоліком цього методу є те, що важливі, але малі кластери можуть поглинутися алгоритмом, що суперечитиме відображенню дійсної соціальної мережі. Тож доведеться контролювати процес і призначати таким кластерам ідентифікатори, які сповіщатимуть алгоритм про те, що вони не можуть входити до інших.

3. Моделі соціальних мереж

Цей розділ присвячений моделям соціальних мереж, які дають змогу прогнозувати та структурувати поведінку груп.

Для розуміння моделей впливу, введемо кілька понять.

- 1) агенти — суб'єкти соціальних мереж, що можуть бути окремими індивідами, або групами індивідів, що об'єднані за певною ознакою;
- 2) відношення — зв'язки між агентами
- 3) «Лідери думок» — агенти, з найбільшим впливом серед інших.
- 4) Зовнішні агенти впливу — зазвичай медіа, бізнес, що створюють зовнішній вплив (рекламу, маркетингові акції тощо).
- 5) Соціальна кореляція — фактори, що можуть бути спільними для груп агентів.
- 6) Локалізація груп — поділ груп за інтересами, думками тощо.
- 7) Стадії — етапи поширення якоїсь думки, ідеї, нововведення тощо

3.1. Класифікація моделей впливу в соціальних мережах.

Цей підрозділ присвячений моделям соціальних мереж, у яких поведінка агентів є визначеною

Ось деякі класи моделей у соціальних мережах.

1) Порогові моделі

Ці моделі застосовують, коли необхідно спрогнозувати ухвалення рішення людиною з огляду на кількість інших осіб, що вже ухвалили певне рішення. Цю кількість ми можемо позначити певним порогом. Тобто якщо кількість осіб що ухвалили певне рішення більша за цей поріг, то й особа, для якої ми досліджуємо цей показник, ухвалить те саме рішення. Відповідно, якщо кількість осіб менша від порогу, то таке рішення певною особою ухвалене не буде.

Тепер розглянемо математичне формулювання цих моделей.

Вважатимемо агентів, що ухвалили певне рішення (тобто на них поширився вплив) активними, а тих, хто не ухвалив (вплив не поширився) — відповідно, неактивними.

На кроці ініціалізації деяких агентів мережі робимо активними. Нехай маємо поріг для i агента з проміжку $\varphi_i \in [0, 1]$. Він може бути однаковим для всіх агентів мережі або бути власним для кожного агента. Якщо кількість активних агентів-сусідів перевищує цей поріг, то наш агент

теж стає активним. В іншому випадку він залишається неактивним. Варто зауважити, що на ребрах графа, що представляє цю мережу відкладенні значення впливу, і для кожного агента сума всіх впливів, що надходять від активних агентів-сусідів, не може перевищувати 1:

$$\sum_{j-i} a_{ij} \leq 1$$

Відповідно, щоби поточний агент перейшов із неактивного стану до активного необхідно, щоби виконалася така умова:

$$\sum_{j-i} a_{ij} \geq \varphi_i$$

Переваги моделі з порогами:

- Враховує фактори «соціальної кореляції»
- Відображає фактори зовнішнього впливу
- Показує «лідерів думок»
- Враховує фактори «соціальної кореляції»
- Відображає різні ступені агентів схильності до впливу
- Існує поріг, що визначає чи зміниться рішення агента від впливу інших думок
- Присутність власних міркувань агентів

Недоліки:

- Не бере до уваги непрямого впливу

2) Модель незалежних каскадів

У цій моделі вплив поширюється каскадом через мережу. Вона є однією з найпростіших серед дифузійних моделей. Вона дещо нагадує моделі з порогами. Ми так само починаємо з набору активних вершин, але додатково ще зазначаємо ймовірність поширення впливу p_{ij} від i до j агента. Фактично ця ймовірність означає можливість для активного i вузла активізувати кожного його неактивного сусіда. Зупиняється алгоритм, коли припиняється активізація вузлів.

Відмінність між моделями з порогами та каскадами полягає в тому, що в перших активація одного вузла залежить від кількості активних вузлів-сусідів. Натомість у других кожен активний агент має лише одну можливість змінити стан неактивного агента.

Переваги моделі з порогами:

- Враховує існування «лідерів думок»
- Присутність власних міркувань агентів

Недоліки:

- Не бере до уваги непрямого впливу
- Не враховує фактори «соціальної кореляції»
- Не відображає фактори зовнішнього впливу
- Не враховує фактори «соціальної кореляції»
- Не відображає різні ступені схильності агентів до впливу

Каскадні та порогові моделі є базовими моделями дифузії інновацій.

3) Моделі просочування та зараження

Моделі просочування дозволяють описати поведінку кластеризованих компонент у випадкових графах³. У таких моделях соціальна мережа асоціюється з губкою, а вплив - з рідиною. Кожен вузол є аналогом клітини, яку має пройти рідина, щоб дістатися інших.[16]

Для моделі зараження можлива різна поведінка залежно від природи впливу. Наприклад, якщо розглядати явище блогерства, то така модель повинна мати певний поріг популярності, перевищивши який певний набуває значної популярності та поширюється всією мережею. Втім, якщо йдеться про розповсюдження комп'ютерного вірусу, то такий поріг відсутній: якщо ним інфікується бодай один агент, то вся мережа буде заражена.

4) Модель Ізинга

Ідея цієї моделі була запозичена з магнетизму. Нехай перед кожним агентом соціальної мережі постає завдання: ухвалити чи не ухвалити певне рішення. Тоді агенти можуть перебувати у двох станах (спінах): “+1”, якщо агент це рішення ухвалює і “-1”, якщо не ухвалює. Тоді агенти, що мають однакову думку, мають знакосталу поведінку, а якщо мають місце протилежні міркування, то позначаємо їх різними знаками.

5) Моделі на базі клітинних автоматів

6) Моделі на базі ланцюгів Маркова

7) Теоретико-ігрові моделі

³

Випадковий граф - це граф, на якому визначений розподіл ймовірності чи випадковий процес, що задає цей граф.

- Моделі взаємної інформативності
- Моделі узгоджених колективних дій
- Моделі комунікацій
- Моделі стабільності мережі
- Моделі інформаційного впливу й керування
- Моделі інформаційного протиборства

Моделі «дифузії інновацій» описують поширення нововведень у суспільстві з плином часу. Під терміном «нововведення» ми матимемо на увазі будь-який новий для агентів продукт, що може набувати матеріальної чи нематеріальної форми (наприклад, думка, ідея, технологія)

Визначимо групи людей, що досліджуватимуть ці винаходи:

- новатори (ті, хто першими починає вивчати й користуватися винаходами)
- ранні послідовники (ті, хто починає досліджувати нововведення після новаторів)
- рання більшість (основна частина людей, яка на підставі попередніх досліджень раніше від решти починає активно використовувати винаходи)
- пізня більшість (велика частина людей, яка пізніше використовує ці новаторства)
- пізні послідовники

Динаміку цього поширення (тобто залежність кількості послідовників від часу) можна графічно представити за допомогою S-подібної функції.

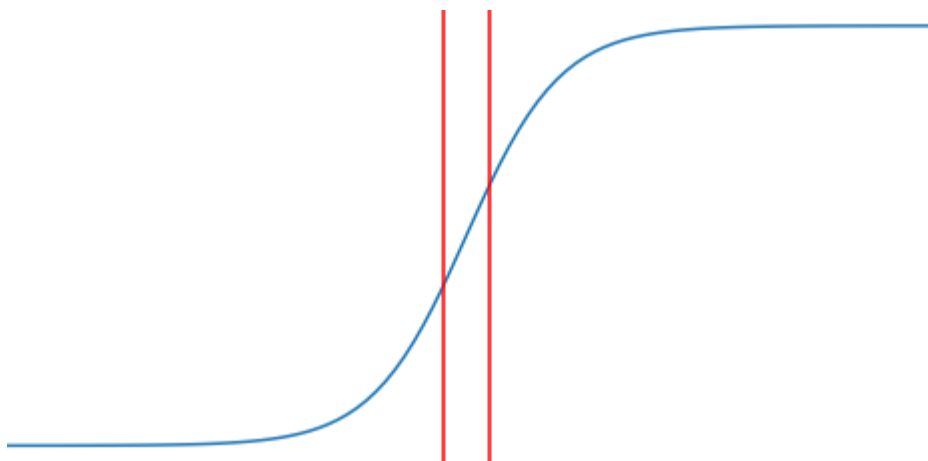


Рис. 3.1 Залежність кількості послідовників від часу

Проекція частини графіка S-подібної функції, відрізана червоними лініями від решти графіка — називається періодом швидкого зростання. Саме в цьому часовому проміжку різко зростає попит на винаходи.

А тепер зобразимо розподіл груп цих людей або, інакше кажучи, криву стадій.

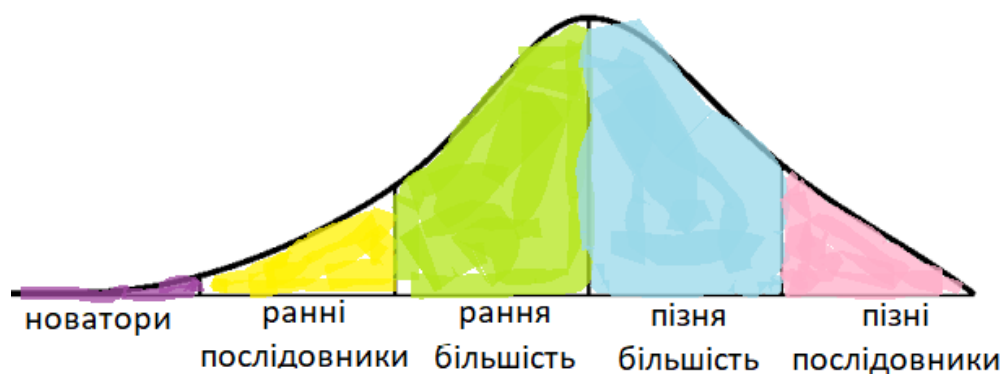


Рис. 3.2. Розподіл груп послідовників

3.2. Стохастичні моделі в соціальних мережах.

Ці моделі описують схильність агентів ухвалювати рішення за допомогою ентропії розподілу ймовірності їхніх дій. На графах це означає, що ваги ребер матимуть стохастичний характер. Через це поведінка агентів є непередбачуваною.

Нехай маємо стохастичний граф $G = \langle V, E, W \rangle$, де

V позначає множину агентів

E — множину ребер

W — матрицю, елементами якої є випадкові величини w_{ij} множину можливих дій учасників соціальної мережі

Така концепція моделі дозволяє ввести нові лаконічні характеристики і застосувати їх до великої кількості агентів. Такими можуть бути математичне очікування та дисперсія. Це дозволяє нам визначити схильність агентів утворювати нові зв'язки.

Недоліком такої концепції є невизначеність, впроваджена у поведінку агентів.

3.3. Моделі інформаційного керування й інформаційного протиборства в соціальних мережах.

3.3.1 Модель інформаційного керування

Введемо поняття довіри. Нехай j агент впливає на i . Тоді i агент, відповідно довіряє j .

Групою вважатимемо мінімальну сукупність агентів, які прямо або непрямо впливають один на одного.

Супутником є такий агент, що довіряє агентам з деяких груп, водночас впливу на жодну з цих груп не має.

Задаємо множину агентів N , кількість яких становить n . У кожного агента є своя початкова думка, відображена у векторі-стовпці $\vec{x}^0$. Ступінь довіри i агента j -ому позначатимемо $a_{ij} \geq 0$. Ступінь довіри всіх агентів мережі один одному будемо позначати матрицею прямого впливу $A = \|a_{ij}\|$. Зауважимо, що вона є стохастичною, тобто розподіл ступенів довіри i агента решті агентів відбувається так, що їхня сума дорівнює 1, тобто

$$\forall i \in N \sum_j^n a_{ij} = 1.$$

Вплив одного агента на його «сусіда» вважатимемо прямим. Якщо i агент впливає на j агента, а j на k , відповідно, i на k , то це — непрямий вплив.

Тоді зміну довіри агентів із часом можна представити як

$$x_i^t = \sum_j a_{ij} x_j^{t-1}, i \in N$$

Тобто на кожному кроці, ми множитимемо ту саму матрицю A на вектор, що отримали на попередній ітерації. Втім тоді виникає питання про критерій закінчення роботи такого алгоритму. За постійного обміну впливом думки агентів з часом стабілізуються. Це означає, що значення матриці A прямують до певного вигляду. Математично можна представити це за допомогою границь:

$$A^\infty = \lim_{k \rightarrow \infty} (A)^k,$$

де k - кількість ітерацій

Для існування цієї границі має виконуватися умова, що у кожній групі існує принаймні один агент i , що довіряє сам собі зі значенням $a_{ii} > 0$. Матриця A^∞ називається матрицею результуючого впливу.

Тоді кінцеву думку агентів можна знайти за такою формулою:

$$\vec{X} = A^\infty \vec{x}^0,$$

де $\vec{X}$ - вектор кінцевих думок

Оскільки ми встановили, як буде змінюватися думка агентів з часом, то можемо додати змінну, що встановлюватиме вплив на думку. Для цього створимо вектор керування $\vec{u} \in \mathbb{R}^n$.

Тоді на кожному кроці будемо додавати цей вектор до поточного вектору x_i

$$\vec{X} = A^\infty (\vec{x}^0 + \vec{u})$$

або у координатному представленні

$$X_{ui} = \sum_{j \in N} A_{ij}^\infty (x_j^0 + u_j) = \sum_{j \in N} A_{ij}^\infty x_j^0 + \sum_{j \in N} A_{ij}^\infty u_j, \quad i \in N$$

Введемо критерій ефективності керування $\Phi(\vec{X}, u)$.

Тоді задача поширення впливу на агентів полягає у тому, щоб максимізувати функцію

$$\Phi(A^\infty (\vec{x}^0 + \vec{u}), \vec{u}) \longrightarrow \max_{u \in U}$$

Отже, задача поширення впливу в соціальній мережі звелася до задачі керування.

Переваги моделі інформаційного керування:

- дозволяє оптимізацію інформаційної взаємодії
- можливість об'єднання у коаліції

Існує певна схожість між цим моделями та ланцюгами Маркова.

Ланцюг Маркова - послідовність випадкових елементів $\{x_n, n \geq 0\}$ у скінченній або зліченній множині E з початковим розподілом $\Pi_i, i \in E$ і перехідною матрицею $P = (p_{ij}), i, j \in E$

Відповідність між марківськими моделями та ланцюгами Маркова можемо встановити, якщо введемо такі перепозначення:

- станом ланцюга маркова вважатимемо агента соціальної мережі;
- довіра агента i агентові j означатиме ймовірність переходу між станами i та j ;
- відповідно, матриця прямого впливу A стане перехідною матрицею P .
- непрямий вплив - досяжність
- група - клас існуючих станів, який не можна розкласти
- супутник - неістотний стан

Висновки до розділу

Цей розділ є заключним, адже демонструє як адаптувати різні соціальні мережі для математичного дослідження.

Першими були досліджені такі основні моделі дифузії інновацій: порогові та каскадні. З огляду на принцип їхньої дії, їхні властивості дещо відрізняються, проте обидві враховують присутність власних міркувань агентів, зміну думок внаслідок впливу інших членів тієї ж соціальної мережі та присутність “лідерів думок”.

Стохастичні моделі завдяки невизначеній поведінці агентів дозволяють описувати соціальні мережі з великою кількістю агентів, але з малою кількістю кількісних характеристик (наприклад, математичним очікуванням чи дисперсією).

Також було представлено модель інформаційного впливу, у якій не тільки зображено як змінюється думка агентів з часом, а і як поширити вплив на них.

Висновки

У цій роботі було розкрито тему дослідження поширення впливу між різноманітними об'єднаннями учасників соціальних мереж: від двох осіб до цілої мережі. Методи, що були наведені для математичного опису таких структур, дозволяють піддавати їх науковому дослідженню навіть у їх сьогоденному вигляді. Втім, сучасні мережі невпинно модернізуються, що вимагає одночасного оновлення методів їхнього дослідження. З огляду на це в роботі були наведені приклади проблем, які виникають через оновлення таких структур, а також запропоновані шляхи їх усунення за допомогою корекції алгоритмів та впровадження більш гнучких величин, що точніше можуть описати зв'язки між учасниками мережі.

Однак, слід зауважити, що хоча ці методи можуть гарно впоратися з моделюванням впливів у мережах, але все ще залишається місце для їх покращення та модернізації [15].

Однією з таких може бути пошук рішення для знаходження оптимальної кількості кластерів у Лувенському методі. Відомо, що він гарно працює у соціальних мережах з дискретними кластерами, втім існує потреба у гарних результатах і щодо соціальних мереж з ієрархічними кластерами.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Ф. Т. Алескеров, Е.Л. Хабіна, Д.А. Шварц «Бінарні відношення, графи й колективні рішення»
2. Pradeep Dubey, Lloyd S. Shapley «Mathematical Properties of the Banzhaf Power Index»
3. Svitlana Popereshnyak, Iryna Yurchuk «Social Networks: Analysis, Algorithms and Their Implementation»
4. James Powell, Matthew Hopkins «A Librarian's Guide to Graphs, Data and the Semantic Web», 2015
5. Annalyn Ng, Kenneth Soo «Теоретический минимум по Big Data. Всё, что нужно знать о больших данных»
6. Prof. Jure Leskovec «CS224W: Machine Learning with Graphs | 2021»
7. Д.А. Губанов, Д. А. Новіков, А.Г. Чхартішвілі «Соціальні мережі: моделі інформаційного впливу, керування та протиборства»
8. Mathispower4u «Weighted Voting: The Shapley-Shubik Power Index» (Youtube)
9. Jeremy L. Martin «Sequential Coalitions»
10. Nathan Garbig «The Johnston Power Index»
11. Mark Granovetter «Threshold Models of Collective Behavior»
12. Alireza Rezvanian, Mohammad Reza Meybodi «Stochastic graph as a model for social networks»
13. https://opora.ua/report/polit_ad/24068-mediaspoznavannia-ukrayintsiv-v-umovakh-povnomasshtabnoyi-viini-opituvannia-opori
14. David Kempe, Jon Kleinberg, Eva Tardos «Maximizing the Spread of Influence through a Social Network»
15. Чугай Д.В., Смирнов С.А. «Моделі впливу у соціальних мережах» — 2022 р.
16. Henry Hexmoor “Diffusion and Contagion” — 2015