Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет ННЦЗФН
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Розробка та дослідження методу кластерізації мереж на спільноти, які перекриваються
(тема)

Виконав:
здобувач другого року навчання,
групи ДСзм-23-1
Крамаренко Д.П.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Науки про дані (Data Science)
(повна назва спеціалізації)
Керівник доц. Шергін В.Л.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ О.В. Золотухін
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Науки про дані (Data Science)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Крамаренку Дмитру Павловичу
(прізвище, ім'я, по батькові)

1. Тема роботи Розробка та дослідження методу кластерізації мереж на спільноти, які перекриваються

затверджена наказом університету від 29 листопада 20 24 р. № 204Стз

2. Термін подання студентом роботи до екзаменаційної комісії 21 січня 20 25 р.

3. Вихідні дані до роботи Науково-технічні публікації та дані інтернет-джерел щодо моделей мереж та методів розбиття мереж на спільноти

4. Перелік питань, що потрібно опрацювати в роботі

1) Огляд предметної галузі та постановка задачі дослідження

2) Генеративна модель для спільнот зв'язків мережі

3) Програмна реалізація та тестування алгоритму розбиття мереж на спільноти

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	25.11.2024	виконано
2	Огляд предметної галузі	29.11.2024	виконано
3	Постановка завдання та узгодження з керівником	02.12.2024	виконано
4	Дослідження методів розбиття мереж на спільноти	09.12.2024	виконано
5	Дослідження та аналіз генеративних моделей мереж, які	16.12.2024	виконано
	придатні для моделювання формування спільнот		
6	Аналіз, розробка та вдосконалення алгоритмів оцінювання	23.12.2024	виконано
	параметрів генеративної моделі		
7	Програмна реалізація моделі алгоритмів генерації	30.12.2024	виконано
	мереж та оцінювання параметрів моделі		
8	Проведення експериментальних досліджень	06.01.2025	виконано
9	Написання пояснювальної записки	13.01.2025	виконано
10	Захист перед ЕК	21.01.2025	

Дата видачі завдання 25 листопада 20 24 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Шергін В.Л.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 67 с., 10 рис., 1 табл., 2 дод., 13 джерел.

МЕРЕЖІ, ОЧІКУВАННЯ-МАКСИМІЗАЦІЯ, ПРАВДОПОДІБНІСТЬ, СПІЛЬНОТИ ВУЗЛІВ, СПІЛЬНОТИ ЗВ’ЯЗКІВ.

Об’єктом досліджень є мережі.

Предметом досліджень кваліфікаційної роботи є методи розбиття мереж на спільноти.

Мета роботи – дослідження методів та алгоритмів розбиття мереж на спільноти, які перекриваються.

Методи дослідження – математичне моделювання та статистичний аналіз використовувались для розробки генеративної моделі мережевих спільнот на основі розподілу Пуассона та методу максимальної правдоподібності. Алгоритмічний аналіз дозволив дослідити та оптимізувати обчислювальну складність EM-алгоритму. Експериментальні методи включали тестування на синтетичних та реальних мережах з використанням комп'ютерного моделювання в Python із застосуванням бібліотек NetworkX та Gephi. Для валідації результатів застосовувались статистичні методи оцінки точності виявлення спільнот.

ABSTRACT

Master’s thesis contains: 67 p., 10 fig., 1 tabl., 2 ann., 13 sources.

EXPECTATION-MAXIMIZATION, LIKELIHOOD, LINK COMMUNITIES, NETWORKS, NODE COMMUNITIES.

The object of research is networks.

The subject of research are methods of splitting networks into communities

The purpose of the work is to research methods and algorithms for splitting networks into overlapping communities.

Research methods –mathematical modeling and statistical analysis were used to develop a generative model of network communities based on the Poisson distribution and maximum likelihood method. Algorithmic analysis enabled investigation and optimization of the computational complexity of the EM algorithm. Experimental methods included testing on both synthetic and real networks using computer modeling in Python with NetworkX and Gephi libraries. Statistical methods were applied to validate the accuracy of community detection results.

ЗМІСТ

ВСТУП

Предметом досліджень кваліфікаційної роботи є задача розбиття мережі на групи (кластери, ком’юніті, спільноти). У багатьох випадках складні мережі цілком природньо діляться на такі спільноти. Властивістю будь-якого розумного поділу є те, що щільність зв’язків всередині спільнот є суттєво вищою, ніж щільність зв’язків між вершинами з різних спільнот. Відповідно, пошук науково обгрунтованих методів виявлення спільнот у мережах на основі структури зв’язків в них є водночас актуальною науковою проблемою теорії графів та мереж, а з іншого боку, ця задача є вкрай актуальною у практичному сенсі.

Задача розбиття мережі на спільноти припускає дві досить суттєві відмінності у формулюванні: вершини мереж можуть входити лише до однієї з можливих спільнот, або одночасно до різних. Інакше кажучи, функція приналежності вершини до спільноти може бути звичайною, або нечіткою. Наприклад, людина, як вузол соціальної мережі, може входити одночасно до різних спільнот: коло особистого спілкування, спільнота за професійною ознакою, за мовою, релігією, географічним розташуванням тощо. Виходячи із зовнішніх вимог інколи перекриття між групами цілком припустиме, а інколи вимагається класифікувати вершину однозначно.

В роботі розглядаються обидві формулювання, причому чітке розбиття (розбиття без перекриття спільнот) розглядається як частковий випадок задачі розбиття мереж на спільноти з припустимістю перекриттів.

Проводиться дослідження методів розв’язання зазначеної проблеми, аналіз існуючих методів та алгоритмів, вимоги до них. Робиться обгрунтований вибір метода. Здійснюється налаштування та модифікація обраного метода, його алгоритмічна реалізація.

У практичній частині роботи виконується тестування працездатності та ефективності обраного підходу. Тестування виконується як на штучно згенерованих мережах, так і на типових датасетах мереж.

1 ОГЛЯД ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Огляд сучасного стану проблеми розбиття мереж на спільноти

Багато мережевих систем, включаючи біологічні та соціальні мережі, природно діляться на модулі або спільноти, тобто групи вершин із відносно щільними зв’язками всередині груп та з малою кількістю зв’язків між вузлами, які належать різним групам [1], [2]. Залежно від контексту групи можуть бути такими що перетинаються, або такими що не перетинаються. Важливою науковою проблемою в теорії мереж, яка викликала значний інтерес серед дослідників в цьому сторіччі, є те, як виявити такі спільноти в емпіричних мережевих даних [2], [3]. Є кілька бажаних властивостей, які повинен мати хороший метод виявлення спільнот.

По-перше, він повинен бути ефективним, тобто він повинен мати можливість точно визначити структуру спільнот, якщо вона дійсно присутня. Існує, наприклад, багато прикладів мереж, як природних, так і синтетичних, для яких структура спільнот є відомою, і успішний метод виявлення спільнот повинен знайти їх.

По-друге, методи, засновані на обгрунтованих теоретичних принципах, є кращими, ніж ті, які такими не є. Метод, заснований на простому передчутті, що щось може спрацювати, за своєю суттю заслуговує меншої довіри, ніж метод, заснований на підтвердженому результаті або фундаментальному математичному розумінні.

По-третє, якщо метод реалізований як комп’ютерний алгоритм, то бажано, щоб він був бути швидким і міг добре масштабуватися відповідно до розміру аналізованої мережі. Багато мереж, які вивчає сучасна наука, є великими, з мільйонами чи навіть мільярдами вершин, тому алгоритм виявлення спільнот, час роботи якого, скажімо, лінійно залежить від розміру мережі, має надзвичайну перевагу, над таким, який має складність квадрату

чи кубу від розміру мережі.

В роботі розглядається проблема виявлення спільнот у тому випадку, коли вони можуть перетинатись. Якщо ж перетинання не припустиме, то таку задачу можна розглядати як частковий випадок задачі розбиття з перетинаннями. Ранні спроби виявлення спільнот припускали, що спільноти не перетинаються, але у практичних ситуаціях спільноти часто збігаються.

У соціальних мережах, наприклад, люди часто належать до кількох кіл знайомих – родина, друзі, колеги тощо. Тому ці кола слід вважати такими, що збігаються, оскільки вони мають принаймні одного спільного члена. У біологічних мережах вершини також можуть належати до кількох груп. Метаболіти в метаболічній мережі можуть відігравати роль більш ніж в одному метаболічному процесі або циклі; види в харчовій мережі можуть потрапляти на кордон між двома субспільнотами, які інакше не взаємодіють, і відігравати певну роль в обох. Таким чином, найзагальніше формулювання проблеми виявлення спільноти має передбачати можливість перетинання.

Розв’язання задачі виявлення спільнот грунтується на використанні стохастичної генеративної моделі мережевої структури спостережуваних мережевих даних. Цей підхід, який застосовує методи статистичного виводу до мереж, досліджувався низкою авторів для випадку неперекриття [4], [5]. Однак поширення цього підходу на випадок припустимості перекриттів є нетривіальним. Вирішальним кроком є саме розробка генеративної моделі, яка відображає мережі з спільнотами, які перекриваються, подібною до тієї, що спостерігається в реальних мережах.

Моделі, які використовуються в більшості робіт, є моделями «змішаного членства» [4], в яких, як правило, вершини можуть належати до кількох груп, і дві вершини, швидше за все, будуть з’єднані, якщо вони мають більше ніж одну спільну групу. Це, однак, означає, що область перекриття між двома спільнотами повинна мати вищу середню щільність зв’язків, ніж область, яка потрапляє лише в одну спільноту. Незрозуміло, чи

точно це відображає поведінку мереж реального світу, але, безперечно, можна побудувати мережі, які не мають такого типу структури. Доцільно б було віддати перевагу менш обтяжливій моделі, тобто такої, яка робить менше припущень щодо структури спільнот.

Ще один набір підходів до виявлення спільнот, які перекриваються, являють собою такі, що базуються на структурі локальної спільноти. Замість того, щоб розділяти всю мережу на спільноти за один крок, ці методи натомість шукають локальні групи в мережі на основі аналізу шаблонів локальних з'єднань та ігнорування глобальної структури мережі.

Методи такого роду природно створюють спільноти, які перекриваються, бо створюється велика кількість незалежних локальних спільнот по всій мережі. Виявлені таким чином спільноти мають тенденцію бути компактними та зв'язними підграфами, вимога, яка не завжди відповідає іншим методам. З іншого боку, глобальні методи виявлення можуть краще охопити структуру великомасштабної мережі та є більш доцільними, коли повинні бути задоволені певні обмеження, такі як щодо кількості спільнот.

Останні роки велику увагу було приділено розробці глобального статистичного методу виявлення спільнот, що перекриваються, на основі ідеї спільнот за зв'язками. Ця ідея була незалежно запропонована рядом авторів, зокрема й у машинному навчанні [6]. Ідея полягає в тому, що спільноти виникають у тому разі, коли в мережі існують різні типи зв'язків. У соціальній мережі, наприклад, є посилання, що представляють родинні зв'язки, дружбу, професійні стосунки тощо. Якщо ми зможемо ідентифікувати типи ребер, тобто кластеризувати не вершини, а саме ребра, тоді ми після цього зможемо також згрупувати у спільноти й вузли, спираючись на класифікацію інцидентних ним ребер.

Цей підхід має приємну особливість – узгодженість з нашим інтуїтивним уявленням про походження та природу структури спільнот, водночас створюючи спільноти, що перекриваються природним чином:

вершина належить більш ніж одній спільноті, якщо вона має ребра більше ніж одного типу.

Існуючі підходи до виявлення спільнот посилань використовували евристичні функції якості, оптимізовані щодо можливих розділень зв'язків мережі [7], [8]. Такі функції якості, зокрема функція модулярності [7], використовуються для спільнот, які не перетинаються. Але хоча на практиці ці функції часто дають прийнятні результати, вони також мають деякі недоліки: модулярність, наприклад, не може бути використана для пошуку дуже малих спільнот, вона може не мати єдиного максимуму, крім того, модулярність є дещо незадовільною з формальної точки зору. Результати роботи [7] свідчать про те, що ці недоліки можна виправити, відмовившись від підходу функції якості, та натомість підігнавши генеративну модель до даних. Цей підхід є доцільним, але визначення моделі для спільнот посилань вимагає певної тонкощі.

У генеративних моделях для спільнот вершин можна спочатку розкласти вершини по групам, а потім розмістити ребра на основі цього призначення. Але для моделі спільнот посилань неможливо призначити ребра групам, доки вони не існують, тому ребра та їх групування мають бути створені одночасно. Після побудови моделі, подальший крок полягатиме в тому, щоб визначити значення параметрів моделі, які найкраще відповідають спостережуваній мережі, а потім на основі цієї моделі визначити спільноти вершин.

При цьому доцільно спочатку розробити рішення загальної проблеми розбиття (тобто з припустимістю перетинань), а вже потім показати, як варіант того самого підходу можна також застосувати до розбиття на спільноти, які не перетинаються.

Таким чином, першим етапом роботи є аналіз поняття складних мереж, властивостей мереж, моделей мереж, зокрема генеративних моделей, а також методу максимальної правдоподібності.

1.2 Основні поняття теорії графів

Математичною моделлю будь-якої мережі є граф. Графом називають такий об'єкт $G(V,E)$, який складається з множини вершин (V) та множини зв'язків (з'єднань, E) між парами вершин. Ці зв'язки також називають дугами, або ребрами. За необхідністю графи можуть містити додаткові дані, які можуть асоціюватись з вершинами або з ребрами.

Засновником теорії графів є Леонард Ейлер. Він в 1736 р. розв'язав відому задачу про Кенігсбергські мости (рисунок 1.1): довів, що не існує такого замкненого маршруту, який би проходив по всім мостам строго по одному разу по кожному з них.

Очевидно, що такий маршрут (якщо б він існував) можна б було накреслити не відриваючи олівця від бумаги та не дублюючи вже існуючи лінії.

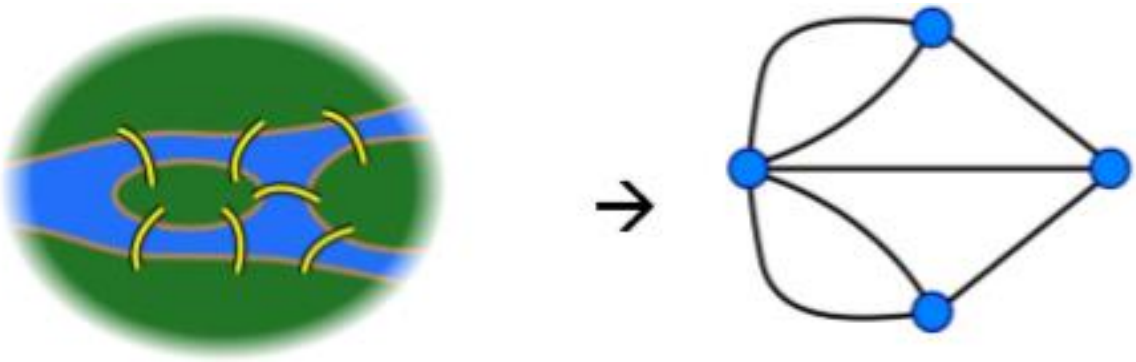


Рисунок 1.1 – Схема мостів Кенигсберга

Ейлер сформулював та довів теорему: для того, щоб граф можна було накреслити, не відриваючи олівця від паперу, необхідно, щоб число вершин з непарним числом інцідентних ребер дорівнювало нулю, або двом. Якщо воно дорівнює нулю, то цикл можна починати в довільній вершині, а якщо двом – то в одній з непарних вершин. Такий цикл має назву цикл Ейлера.

У графі Кенігсберзьких мостів (рисунок 1.1) всі вершини мають непарну кількість інцидентних їм ребер (три по три та одна – п'ять), тому

циклу Ейлера не існує. Але якщо вилучити будь-яке ребро, чи додати, то буде існувати Ейлерів ланцюг. Ця публікація Ейлера лягла в основу теорії графів.

1.2.1 Різновиди графів, способи представлення графів

Існують різні види графів.

Просто граф, або неорієнтований граф (неорграф) – впорядкована пара $G(V,E)$, де V – це непорожня множина вершин, а $E \subseteq V^2$ – множина пар вершин, званих ребрами. Множина V (а отже й E) зазвичай вважається скінченою. Варто зазначити, що деякі твердження, вірні для скінчених графів, є невірними у випадку, якщо граф нескінчений.

Вершини u та v називаються кінцями (або кінцевими вершинами) ребра (u,v). Якщо довільні вершини u та v поєднані якимось ребром (хоча б одним), то вони називаються сусідніми.

Ребра називаються суміжними, якщо вони мають спільну кінцеву вершину.

Ребра називаються кратними, якщо їхні кінці співпадають (тобто у разі якщо множини кінцевих вершин тотожні між собою).

Ребро називається петлею, якщо його кінці збігаються між собою, тобто $e=(v,v)$.

Простим називають граф, який не має ані петель, ані кратних ребер.

Ступенем вершини v (зазвичай позначається як $\deg(v)$) називають кількість ребер, інцидентних цій вершині. Якщо серед цих ребер є петлі, то вони рахуються двічі.

Вершина називається ізольованою, якщо вона не має інцидентних їй ребер. Висяча вершина (або лист) – це така вершина, що є кінцем рівно одного ребра.

Орієнтований граф (скорочено орграф) відрізняється тим, що зв’язки між вершинами мають напрямок (тобто орієнтацію). Ці зв’язки називають

дугами, а їх напрямок на рисунках позначають стрілками.

Таким чином, орграф – це впорядкована пара $G = (V, A)$, де V – непорожня множина вершин, а A – множина впорядкованих пар цих вершин, тобто дуг.

Дуга – це впорядкована пара вершин (u, v), де вершина u є початком, а вершина v – кінцем дуги. Таким чином, дуга $u \to v$ веде від вершини u до вершини v. При цьому в графі може також існувати зворотня дуга $v \to u$, а може й не існувати. У такому випадку, зазвичай, ці дуги $u \to v$ та $v \to u$ поєднують в неорієнтоване ребро $u - v$.

Якщо граф містить як дуги (орієнтовані зв'язки), так і ребра (неорієнтовані зв'язки), то він називається змішаним. Таким чином, орграфи та неорграфи (тобто звичайні графи) є частковими випадками змішаного графа.

Узагальненням звичайних графів є зважені графи, тобто такі, в яких кожному ребру поставлено у відповідність деяке число (вага ребра чи дуги). Можна сказати, що звичайні графи є частковими випадками зважених, у разі, якщо всі ваги одиничні.

Мультиграф – це граф, який містить кратні ребра.

Псевдограф – це граф, який містить петлі.

Граф, який не містить ані петель, ані кратних ребер, називається простим графом. Зазвичай, термін «граф» без уточнень («мульти» чи «псевдо») позначає саме простий граф.

Гіперграф – це граф, у якому ребро може з'єднувати більше двох вузлів.

Існує декілька способів представлення графів з метою їх подальшої комп'ютерної обробки та зберігання. Вибір способу представлення залежить від розміру графа, його щільності та задач по його обробці.

Основним з таких способів є матриця суміжності. Рядки та стовпці матриці відповідають вершинам графа. У кожному осередку цієї матриці міститься число, що визначає зв'язок між вершиною-рядка та вершиною-

стовпця. Для простих (незважених та неорієнтованих) графів таким числом елементами матриці суміжності є одиниці та нулі (відповідно у разі наявності та відсутності ребра, яке поєднує вершину-рядок з вершиною-стовпцем). Така матриця є симетричною. Матриця суміжності орграфу складається з 0, 1 та -1 (тобто розрізняють вихід чи вхід дуги у вершину). Така матриця антисиметрична. Матриця суміжності зважених графів містить ваги ребер (дуг).

Цей спосіб представлення графів є найбільш зручним та поширеним. Його ефективність відносно інших способів (щодо обсягу пам'яті) збільшується з ростом щільності графа.

У матриці інцидентності стовпці відповідають ребрам (дугам) графу, а рядки – вершинам. Елемент матриці (на перетині рядка i зі стовпцем j) дорівнює +1, якщо дуга під номером j виходить з вершини i, та -1, якщо вона входить в вершину i. Усі інші елементи матриці інцидентності дорівнюють нулю. Даний спосіб представлення графа є найбільш витратним по пам'яті (розмір пропорційний $|V|\cdot|E|$), але може бути ефективним для деяких задач з обробки графів, наприклад, в задачі знаходження циклів.

У списку суміжності графа кожній його вершині відповідає рядок, в якому зберігається список суміжних вершин. Така структура даних являє собою список списків, тобто не є квадратною чи прямокутною матрицею. Кількість пам'яті – $O(|V|+|E|)$. Цей спосіб є найбільш ефективним у разі, якщо щільність графу мала (такі графи називають розрідженими). Списки суміжності використовуються в алгоритмах швидкого обходу графа в ширину або глибину.

У списку ребер графа кожному його ребру графа відповідає рядок, який містить номери вершин-кінців цього ребра. Обсяг потрібної пам'яті складає $O(|E|)$. Цей спосіб зберігання графів – найекономніший за памяттю, тому саме він часто застосовується для зовнішнього зберігання або обміну даними.

1.2.2 Числові характеристики графів та мереж

Головною індивідуальною характеристикою вузла мережі (тобто вершини графу) є його ступінь – кількість зв’язків (тобто безпосередніх сусідів даного вузла). У разі, якщо зв’язкі є орієнтованими (тобто якщо граф є орграфом), то рахують окремо вхідну та вихідну ступені, тобто зв'язки (дуги), які входять в вузол, та ті, які з нього виходять.

За необхідністю вузли можуть мати додаткові параметри, наприклад, вік (час створення), вагу тощо. Кількість цих параметрів є довільною та залежить від предметної галузі.

Мережі (графи) в цілому характеризуються такими параметрами, як:

– розмір (кількість вузлів, $n = |V|$);

– кількість зв'язків ($|E|$);

– щільність мережі;

– кількість компонентів зв'язності;

– середня відстань від одного вузла до інших;

– діаметр мережі – найбільша відстань між вузлами в мережі;

– асортативність (схильність вузлів бути поєднаними з подібними, або протилежними собі за деякою властивістю);

– коефіцієнт кластеризації

– та інші.

Щільність мережі є числовою характеристикою кількості наявних ($|E|$) зв'язків у співставленні з максимально можливою їх кількістю для мережі даного (фіксованого) розміру. Найпростішою числовою мірою щільності є відношення між $|E|$ та кількістю максимально можливих зв'язків для мережі даного розміру $E_{\max}(n)$, яка дорівнює $E_{\max}(n) = n(n-1)/2$:

$$\rho = \frac{2|E|}{n(n-1)}. \qquad (1.1)$$

Така міра є найпростішою для розрахунку, але не дуже зручною: відношення (1.1) не дає змоги адекватно співставляти такі мережі, що суттєво розрізняються за розміром.

Наприклад, мережа (рисунок 1.2) має три зв'язки з шести можливих. Згідно з (1.1), її щільність дорівнює 50%. При цьому з практичної точки зору ця мережа є деревом, тобто не щільною, а навпаки розрідженою.

Рисунок 1.2 – Приклад простої мережі (дерева)

Якщо розглянути мережу розміром $n = 25$ (наприклад, граф товариських зв'язків між студентами групи), то максимальна можлива кількість зв’язків сягає 300, а щільності (1.1) у 50% відповідає наявність 150 зв'язків. Середня кількість вершин, суміжних з даною (тобто середня кількість товарищів) складе $2|E|/n = 12$. Якщо таку мережу зобразити графічно, то вона буде виглядати досить щільною.

Якщо ж уявити собі світову мережу аеропортів ($n > 50\,000$), то навіть за дуже низької щільності (1.1) в 1%, між ними було б більше 12 мільйонів пар зв'язків (тобто унікальних за маршрутом рейсів).

Набагато зручнішою мірою щільності мереж, є ступенева щільність – показник ступеню λ, який пов’язує кількість зв’язків та кількість вузлів $E(n) \underset{n\to\infty}{\to} Cn^{\lambda}$. З цього випливає, що

$$\lambda = \lim_{n\to\infty} \frac{\log E(n)}{\log(n)}. \qquad (1.2)$$

Цей показник збігається з показником еластичності мережі та розраховується як розв'язок рівняння

$$E = \frac{\Gamma(n+\lambda-1)}{\Gamma(n-1)\Gamma(\lambda+1)}, \tag{1.3}$$

відносно $1 \le \lambda \le 2$, де $\Gamma(x)$ – гамма-функція Ейлера.

Для будь-якої деревовидної мережі (зокрема, тієї, що на рисунку 1.2) показник еластичності дорівнює одиниці. Якщо розмір (тобто кількість вузлів) мережі (n) є фіксованим, то дерево має найменшу можливу кількість ребер серед усіх зв'язних мереж; будь-яке ребро дерева є мостом, тобто його видалення призведе до розпаду мережі на дві компонентни зв'язності. Таким чином, $\lambda = 1$ є найменшим значенням показника ступеневої щільності на множині зв'язних мереж.

Іншим крайнім випадком мережі є повна мережа, яка має максимально можливу кількість зв'язків $E_{\max}(n) = n(n-1)/2$, тобто кожна пара вузлів зв'язана ребром. Показник еластичності (тобто міра ступеневої щільності) повної мережі довільного розміру дорівнює двом.

Наприклад, зв'язна мережа з $n = 4$ вузлами може мати від трьох (дерево) до шести (повний граф) ребер. У випадку, якщо вона має $|E| = 4$ ребра її ступенева щільність складе $\lambda = (\sqrt{33}-3)/2 \approx 1.3723$. Якщо розмір мережі дорівнюватиме $n = 25$, то за зазначеній ступеневій щільності ввона матиме 65 зв'язків (тобто кожен вузол матиме в середньому 2.6 сусідніх). Мережа розміром $n = 50\,000$ матиме в середньому 46 зв'язків на вузол, а взагалі – біля 2.3 млн. зв'язків.

Як було зазначено, довжиною шляху між вузлами є кількість ребер, вздовж яких необхідно крокувати аби дістатися від одного вузла до іншого. Доцільно розглядати лише елементарні шляхи, тобто такі, в яких ребра та вузли не повторюються. Серед всіх елементарних шляхів, що з'єднують дані

вузли, виділяють найкоротший шлях (SP, shortest path). Його довжина розглядається як відстань між цими вузлами. Можна зазначити, що найкоротший шлях між заданими вузлами може бути неунікальним, але в цьому випадку всі такі найкоротші шляхи матимуть однакову довжину.

Якщо мережа є зв'язною, то між кожною парою вузлів (i, j) існує хоча б один шлях, яким можна дістатися від одного з них до іншого. У такому разі всі найкоротші відстані у мережі l_{ij} є скінченими.

Протилежним крайнім випадком є мережа, яка складається виключно з ізольованих вузлів, тобто не має жодного зв’язку. В такій мережі кожний вузол є окремою компонентою зв'язності.

Важливою характеристикою мережі як цілісної структури є середня найкоротша відстань. Усереднення робиться по всім $n(n-1)/2$ парам вузлів:

$$l_s = \frac{2}{n(n-1)} \sum_{i \geq j} l_{ij}\,, \tag{1.4}$$

де n – кількість вузлів;

l_{ij} – відстань (довжина найкоротшого шляху) між вузлами i та j.

Іншою важливою з практичної точки зору метричною характеристикою мережі є її діаметр, який визначається як максимум серед попарних відстаней:

$$l_{\max} = \max_{i,j} \left(l_{ij} \right). \tag{1.5}$$

У разі, якщо мережа має більш однієї компоненти зв’язності (тобто є незв'язною), то деякі вузли є недосяжними один з одного, тобто відстань між ними буде нескінченою. Зрозуміло, що в такому випадку міри (1.4) та (1.5) також дорівнюватимуть нескінченості, тобто не будуть мати сенсу.

Тому більш універсальною мірою ніж середня найкоротша відстань (1.4) є довжина середнього інверсного шляху, яка визначається наступним чином:

$$l_{inv} = \frac{n(n-1)}{2}\left(\sum_{i \geq j} l_{ij}^{-1}\right)^{-1}. \qquad (1.6)$$

Наприклад, для деревовидної мережі (рисунок 1.2) середня найкоротша відстань дорівнює $l_s = \frac{1+2+3+1+2+1}{6} = \frac{5}{3}$, діаметр складає $l_{max} = 3$, а довжина середнього інверсного шляху – $l_{inv} = 6\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{1}+\frac{1}{2}+\frac{1}{1}\right)^{-1} = \frac{18}{13}$.

Коефіцієнт кластеризації мережі був запропонований в 1998 році Уаттсом (D. Watts) та Строгатцем (S. Strogatz). Цей коефіцієнт характеризує схильність, чи несхильність вузлів до утворення груп, вузли яких суттєво тісніше пов’язані між собою, аніж з іншими вузлами мережі. Такі групи називаються кластерами.

Коефіцієнт кластеризації C_i окремого вузла i показує, скільки вузлів-сусідів цього вузла безпосередньо пов’язані між собою, тобто є сусідами один для одного. Цей коефіцієнт дорівнює відношенню числа «трикутників» (Δ_i) з вершиною i до числа «виделок» (V_i) (два зв'язки, що виходять з вузла) з основою в цьому ж вузлі i:

$$C_i = \frac{\Delta_i}{V_i}. \qquad (1.7)$$

Наприклад, в мережі (рисунок 1.3) існують три «виделки» з вершиною в вузлі 1: 213, 214 і 314, а «трикутник» тільки один (213). Тому коефіцієнт кластеризації першого вузла дорівнює $C_1 = 1/3$. Аналогічним чином легко

бачити, що $C_2 = C_3 = 1$. Щодо четвертого вузла, то він є «висячим», тобто не має ані трикутників, ані виделок. В такому випадку показник кластеризації невизначений, проте часто його умовно вважають нульовим: $C_4 = 0/0 = 0$.

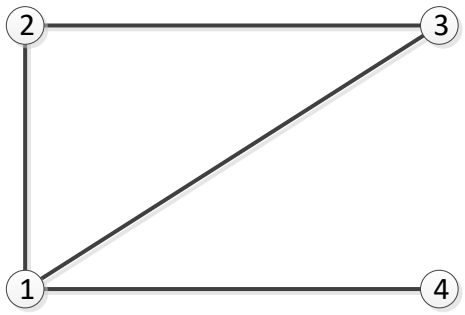


Рисунок 1.3 – Приклад мережі з чотирма вузлами та чотирма зв'язками

Зазвичай, структура мережі (графа) задається матрицею суміжності A_{ij}. В такому випадку показники кластеризації вузлів обчислюються за формулою

$$C_i = \frac{\sum_{j,m} A_{ij} A_{jm} A_{mi}}{k_i (k_i - 1)}, \qquad k_i = \sum_j A_{ij} = \deg(i), \tag{1.8}$$

де підсумовування ведеться по всіх вузлах j, m.

Якщо коефіцієнти кластеризації окремих вузлів ((1.7), або (1.8)) відомі, то можна обчислити коефіцієнт кластеризації всієї мережі шляхом усереднення коефіцієнтів кластеризації вузлів:

$$C = \frac{1}{n} \sum_{i=1}^{n} C_i . \tag{1.9}$$

Для мережі, яка розглядалась (рисунок 1.3)

$C = (1/3 + 1 + 1 + 0)/4 = 7/12$.

Дещо схожою з (1.9), але трохи іншою мірою кластеризації вузлів мереж, є показник транзитивності:

$$T = \frac{tr(A^3)}{\sum_i k_i(k_i - 1)} = \frac{\sum_i (A^3)_{ii}}{\sum_i k_i(k_i - 1)}. \qquad (1.10)$$

Транзитивність мережі пов'язана з кількістю «виделок» та «трикутників» в цій мережі простим співвідношенням

$$T = \frac{\sum \Delta_i}{\sum V_i}. \qquad (1.11)$$

Для мережі з прикладу рисунок 1.3, який розглядався, $T = \frac{1+1+1+0}{3+1+1+0} = \frac{3}{5}$

1.3 Поняття «складної мережі», моделі складних мереж

Терміни «мережа» і «граф» багато в чому схожі. Під час дослідження структури зв’язків у мережі в якості її математичної моделі застосовують саме граф, тому терміни «мережа» і «граф» часто вживаються як синоніми. Різниця між цими термінами з’являється у разі, коли до уваги береться метрика простору (навіть не сама по собі метрика, а наявність самого цього простору).

Будь-яка існуюча в реальному світі мережа (комп’ютерна, електрична, транспортна) розгортається в деякому метричному просторі, наприклад, на площині (рисунок 1.4), на поверхні сфери, в тривимірному просторі тощо. В той же час граф, є суто математичним об’єктом, а тому й позапросторовим.

Це означає, що досліджуючи мережу методами та в межах теорії графів абстрагуються як від метрики простору, в якому мережа знаходиться, так і від існування самого цього «зовнішнього» простору.

Проте, з точки зору переважаючий більшості практичних застосувань така відмінність не має значення, тому, як правило, ігнорується і термін «мережа» вживається як більш «прикладний» та сучасний синонім «теоретичного» терміну «граф».

Рисунок 1.4 – Мережа, утворена накладанням графа (а) на дискретний двовимірний простір (б)

Теорія складних мереж є відносно новим науковим напрямом; вона сформувалася наприкінці XX століття. Незважаючи на те, що обєктом досліджень цієї теорії є мережі різної природи (мережі зв’язку, електричні, транспортні, інформаційні, біологічні та ін.), найбільший внесок в розвиток теорії складних мереж внесли дослідження соціальних мереж. Вузлами соціальних мереж є соціальні об’єкти (окремі особистості, або спільноти), а зв'язками – соціальні відносини між ними. Виявилось, що хоча структуру соціальних мереж (як і будь-яких інших) можна математично описати засобами теорії графів, проте ці мережі мають такі властивості, які або не розглядаються у теорії графів, або вважаються у межах цієї теорії несуттєвими.

В даний час термін «складні мережі» охоплює мережі, які мають такі

властивості:

– великі розміри: ці розміри можуть сягати мільонів та мільярдів вузлів, що унеможливлює як візуальне зображення структури, так і індивідуальне дослідження властивостей окремих вузлів;

– зростання (зміна) в часі: теорія графів зазвичай розглядає графи як сталі об’єкти, в той же час більшість мереж реального світу є динамічними, в них безперевно виникають та зникають вузли та зв’язки між ними;

– елементи випадковості при формуванні: формування та розвиток складної мережі в цілому та окремих її вузлів підпорядковується певним закономірностям, проте ці закономірності мають статистичний характер, тобто формування та еволюція мережі є випадковим процесом.

Типовими задачами дослідження складних мереж є:

– дослідження класичних «графових» характеристик складних мереж різної природи (наприклад, щільності, кластеризації, діаметру та ін.);

– дослідження статистичних властивостей процесу еволюції мереж;

– моделювання складних мереж, їх візуалізація;

– моделювання та дослідження відповідних «фізичних» процесів на складних мережах (розповсюдження інформації, епідемії, процеси дифузії, перколяція тощо);

– передбачення поведінки мереж при частковій зміні структури зв’язків та зворотня задача – пошук змін зв’язків, необхідних для досягнення потрібних характеристик мережі.

Під час цих досліджень, так само як і в теорії графів, досліджуються як параметри окремих вузлів та зв'язків, так і статистичні характеристики мережі в цілому.

1.3.1 Типові генеративні моделі складних мереж

Наразі існує багато різних моделей складних мереж. Найбільш поширеними да дослідженими з них є модель випадкового графу (модель

Ердеша-Ренˋї), модель безмасштабної мережі (модель Барабаші-Альберт) та модель «тісного світу».

Першою з цих моделей виникла модель випадкового графу Ердеша-Ренˋї (ER) [9]. Вона була запропонована в середині XX століття відомими угорськими математиками Ердешем та Ренˋї.

В мережі ER кожна пара вузлів з'єднана ребром (неорієнтованим) з ймовірністю λ. З ростом λ від 0 до 1 граф стає все більш щільним. При $\lambda = 0$ граф є порожнім (всі вершини ізольовані, ребер немає), при $\lambda = 1$ граф є повним. Таким чином, єдиним керуючим параметром моделі ER є ймовірність зв'язку (λ).

Простота цієї моделі дозволяє досить легко отримувати аналітичні оцінки параметрів мережі. Так, розподіл ступенів вузлів p_k (тобто ймовірність того, що довільний вузол матиме ступінь k) відповідає широковідомому закону Пуассона:

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{1.12}$$

Довжина середнього інверсного шляху між вузлами (1.6) становить

$$l_{inv} \propto \log(N). \tag{1.13}$$

Всі можливі графи, які мають фіксовану кількість вузлів (n) та ребер ($E = \lambda n$) мають однакову статистичну вагу (однакову ймовірність реалізації).

Модель ER придатна для моделювання мереж, в яких множина вузлів є статичною, а динамічні властивості обумовлені виникненням, або зниканням зв'язків. До таких мереж відносяться транспортні або енергетичні мережі на початковій стадії їх розвитку. У той же час, розвиток, наприклад, транспортної мережі характеризується не тільки будівництвом

доріг (тобто зв'язків), а й супроводжується виникненням нових вузлів (станцій, населених пунктів) і за таких умов ER-модель стає непридатною. Тим більше ця модель непридатна для моделювання таких мереж (соціальних, комп'ютерних, інформаційних тощо), множина вузлів яких не є статичною, а ймовірності з'єднання довільно обраних пар вузлів не можна вважати рівними між собою.

У 1999 році американо-угорські фізики А.-Л. Барабаші (Albert-Laszlo Barabasi) та Р. Альберт (Reka Albert) вивчали розподіл вузлів деяких мереж за ступенями цих вузлів. Досліджувались, наприклад, мережі білкових взаємодій в клітинах, структура інтернету, структура авіаційних повідомлень в США тощо [9]. Результат виявився несподіваним. Замість розподілу Пуассона (який притаманний ER моделі), чи принаймні близького до нього, виявилося, що для більшості досліджуваних мереж середнього значення, навколо якого групуються значення ступенів вузлів не існує. Розподіл вузлів за ступенями підпорядковується ступеневому закону роподілу (або його дискретним аналогам):

$$p_k = C \cdot k^{-\alpha}. \qquad (1.14)$$

Інакше кажучи, в цих мережах невелике число вузлів (хабів) мають дуже багато зв'язків, а переважна більшість вузлів мають лише по декілька зв'язків. Більше того: якщо не враховувати ці хаби (тобто видалити з мережі ці вузли та інцидентні їм зв’язки), то розподіл збереже свою форму; він також буде ступеневим, матиме той самий показник розподілу. Тобто невелика частина вузлів будуть концентрувати у себе більшість тих зв’язків, що зосталися. Якщо побудувати графік розподілу ступеневого закону (1.14) та не підписати масштаб координатних вісей, то відновити його (масштаб) буде неможливо, бо «хвіст» розподілу візуально подібен до всього розподілу. Такі мережі отримали назву безмасштабних мереж (scale free networks).

Барабаші і Альберт [9] також запропонували просту модель виникнення та еволюції безмасштабних мереж (БА-модель). Вони показали, еволюція безмасштабних мереж визначається двома чинниками:

Зростання. Мережа є динамічною; на кожному кроці (n) алгоритма додається один новий вузол та з'єднується m зв'язками з вже існуючими вузлами мережі. Початково мережа складається з $m_0 \geq m$ вузлів.

Переважне приєднання (Preferencial attachment). Ймовірніть p_i того, що новий вузол приєднається до існуючого (з номером i) пропорційна ступеню (тобто кількості зв'язків) цього вузла k_i:

$$p_i = \frac{k_i}{2(n-1)}, \tag{1.15}$$

де n – поточний розмір мережі.

Для мережі БА середня довжина найкоротшого шляху між вузлами набагато менше, ніж для ER-моделі (1.13) та дорівнює

$$l_s \propto \log(\log(n)). \tag{1.16}$$

Зі зростанням мережі БА ($n \to \infty$) її числові характеристики сходяться до усталених (асимптотичних) значень. Зокрема, закон розподілу вузлів за ступенем (p_k) сходиться до

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}, \quad (k \geq m). \tag{1.17}$$

Варто зауважити, що розподіл (1.17) є дискретним варіантом ступеневого розподілу (1.14) з показником $\alpha = -3$, а точніше – розподілом Юла-Саймона (Yule-Simon).

Наразі існує дуже багато варіацій моделі БА, проте всі вони

зберігають два основних принципи: зростання та переважне приєднання.

Модель мережі «малого світу» (small world – SW) [9] імітує мережі надмалого (у порівнянні з розміром) діаметром.

Згідно з емпіричною гіпотезою Мілграма, будь-яких двох людей на Землі можна з'єднати ланцюжком з шести знайомих. Ця гіпотеза має назву «теорія шести рукостискань». Числова характеристика (тобто 6) вважається дискусійною (та занизькою), проте гіпотеза Мілграма є загальноприйнятою на якісному рівні: деякі мережі (зокрема, деякі соціальні мережі) характеризуються дуже короткими відстанями між більшістю з пар вузлів. Тому про наш світ кажуть, як про тісний світ (малий світ, small world).

Алгоритм побудови моделі тісного світу був вперше запропонований Уоттсом та Строгатцем та полягає у наступному: генерується регулярний граф розміром n, в якому кожен вузол має ступінь k. Потім деякі з ребер (їхня доля складає $0 < \beta < 1$ від загальної кількості) випадковим чином перенаправлюються, змінюючі інцидентні їм вузли. Оскільки в результаті перенаправлення граф може опинитись незв’язним, то частіше замість перенаправлення ребер просто додають нові.

На рисунку 1.5 показано приклад графа тісного світу з 12 вузлами. Розміром виділено вузли з великою кількістю зв’язків – хаби. Середня ступінь вузла складає 3.833, середня довжина найкоротшого шляху – 1.803, коефіцієнт кластеризації – 0.522.

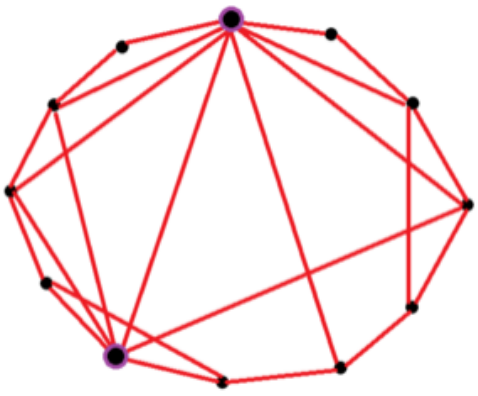


Рисунок 1.5 – Модель WS мережі «тісного світу»

Уоттс та Строгатц показали, що мережі тісного світу мають високий ступінь кластеризації та надмалу середню довжину шляху між вузлами. Такі властивості притаманні, зокрема, мережі нейронних зв'язків хробака нематода, мережі акторів Голівуду, мережі електростанцій та інші.

Згідно з моделлю Уоттса-Строгатца, властивості мереж тісного світу є суперпозицію властивостей регулярних мереж (решіток) та пуассонівських випадкових мереж на кшалт мережі ER. Ці моделі є статичними за розміром. Для генерації графа використовуються три параметри: розмір графа (тобто кількість вершин, n), ступінь вершин до перекидання чи додавання зв’язків (k) та доля зв’язків, які додаються чи перекидаються ($0 < \beta < 1$), причому $1 = \ln n = k = n$.

1.3.2 Метод максимальної правдоподібності

Метод максимальної правдоподібності (також метод найбільшої вірогідності) у математичній статистиці – це метод оцінювання невідомого параметра шляхом максимізації функції правдоподібності. Він ґрунтується на припущенні про те, що вся інформація про статистичну вибірку міститься у цій функції. Метод максимальної правдоподібності був проаналізований, рекомендований і значно популяризуваний Р. Фішером між 1912 і 1922 роками (хоча набагато раніше він використовувався Гаусом, Лапласом і іншими). Оцінка максимальної правдоподібності є популярним статистичним методом, який використовується для створення статистичної моделі на основі даних, і забезпечення оцінки параметрів моделі.

Метод максимальної правдоподібності застосовується для широкого кола статистичних моделей, зокрема:

– лінійні моделі і узагальнені лінійні моделі;

– факторний аналіз;

– моделювання структурних рівнянь;

– задачі перевірки гіпотез і формування довірчого інтервалу;

– дискретні моделі вибору.

Функція правдоподібності, часто звана просто правдоподібністю (likelihood) вимірює відповідність статистичної моделі до вибірки даних для заданих значень невідомих параметрів. Її утворюють зі спільного розподілу ймовірності реалізації отриманої вибірки, але розглядають та використовують як функцію лише від шуканих невідомих параметрів, відтак розглядаючи випадкові змінні як зафіксовані в спостережуваних значеннях.

Функція правдоподібності описує гіперповерхню, чий пік, якщо він існує, представляє поєднання значень параметрів моделі, які максимізують імовірність фактичної реалізації отриманої вибірки. Процедура визначення цих аргументів через максимізацію функції правдоподібності відома як оцінювання максимальною правдоподібністю, яке, заради обчислювальної зручності, зазвичай застосовують з використанням натурального логарифма правдоподібності, відомого як функція логарифмічної правдоподібності (log-likelihood). Крім того, форма та кривина поверхні правдоподібності несуть інформацію про стійкість цих оцінок, через що інколи здійснюють побудову графіку функції правдоподібності. Варіант використання правдоподібності першим зробив Рональд Фішер, який мав переконання, що він є самодостатньою системою для статистичного моделювання та висновування. Згодом Барнард та Бірнбаум очолили наукову школу, яка виступила за принцип правдоподібності, постулюючи, що вся доречна інформація для висновування міститься у функції правдоподібності. Але навіть і в частотницькій та баєсовій статистиці функція правдоподібності відіграє фундаментальну роль.

Функцію правдоподібності зазвичай означують по-різному для дискретних та неперервних розподілів імовірності. Загальне означення також є можливим.

Нехай X є дискретною випадковою змінною з функцією маси ймовірності (pmf) $p_\theta(x)$, яка залежить від параметра θ. Тоді функція

$$\mathrm{L}(\theta \mid X) = p_\theta(x) = \Pr(X = x \mid \theta), \tag{1.18}$$

яку розглядають як функцію від θ, є функцією правдоподібності для заданого результату x випадкової змінної X. Правдоподібність $\mathcal{L}(\theta|X)$ (1.18) не слід плутати з умовною ймовірністю $\Pr(\theta|x)$: правдоподібність дорівнює ймовірності спостерігання певного результату x за умови, що справжнім значенням параметра є θ, і відтак дорівнює густині ймовірності над результатом x, а не над параметром θ.

Нехай маємо вибірку $X_1, X_2, \dots, X_n$ з розподілу P_θ, де $\theta \in \Theta$ – невідомий параметр. Нехай $\mathcal{L}(x|\theta)$ є функцією правдоподібності, яка реалізує відображення $\Theta \to \mathbb{R}$. Тоді точкова оцінка

$$\theta(X_1,\dots,X_n) = \arg\max_{\theta\in\Theta} \mathrm{L}(X_1,\dots,X_n \mid \theta) \tag{1.19}$$

називається оцінкою максимальної правдоподібності параметра θ.

Таким чином, оцінка максимальної правдоподібності – це така оцінка, яка максимізує функцію правдоподібності при фіксованій реалізації вибірки.

Оскільки функція $\ln x$ монотонно зростає на всій області визначення $x > 0$, то максимум будь-якої функції $f(\theta)$ досягається в тій же точці θ, що й максимум функції $\ln f(\theta)$ і навпаки. У переважній більшості випадків логарифм від функції правдоподібності (1.18) диференціюється набагато простіше, ніж сама ця функція. Тому на практиці широко використовується саме логарифмічна правдоподібність (log-likelihood), тобто шукається точкова оцінка

$$\theta(X_1,\dots,X_n) = \arg\max_{\theta\in\Theta} \left(\ln \mathrm{L}(X_1,\dots,X_n \mid \theta)\right). \tag{1.20}$$

Головной рисою оцінок, отримуваних методом максимальної

правдоподібності, є їхня строга статистична обгрунтованість.

Основними недоліками цього методу є складність розв’язання рівняннь (1.19)–(1.20), можливість наявності багатьох локальних мінімумів функції правдоподібності (1.18). Крім того, оцінка максимальної правдоподібності, загалом, може бути зміщеною.

Основним методом розв’язання рівнянь (1.19)–(1.20) є метод очікування-максимізації (EM-алгоритм, Expectation-Maximization).

1.4 Постановка задачі дослідження

Виходячі з проведеного аналізу стану предметної області, в кваліфікаційній роботі ставляться наступні задачі:

– побудова генеративної моделі мережі, яка б описувала розбиття вузлів та ребер на спільноти, які можуть перекриватися;

– застосування методу максимальної правдоподібності (EM-алгоритму) для обчислення параметрів зазначеної моделі;

– аналіз обчислювальної складності EM-алгоритму, його модифікація з метою пришвидшення та зменшення використання пам’яті;

– обґрунтування можливості застосування цього алгоритму для розв’язання задачі розбиття мережі на спільноти, які не перетинаються;

– програмна реалізація алгоритму розбиття мережі на спільноти;

– тестування працездатності та ефективності алгоритму розбиття мережі на спільноти з використанням як штучно згенерованих мереж, так і датасетів реальних мереж;

– висновки по роботі.

2 ГЕНЕРАТИВНА МОДЕЛЬ ДЛЯ СПІЛЬНОТ ЗВ'ЯЗКІВ МЕРЕЖІ

2.1 Створення генеративної моделі для спільнот зв'язків

Першим етапом роботи є створення генеративної моделі мережі, тобто такої моделі, яка б описувала створення мереж із заданими властивостями.

Ця модель генерує мережі $G(n,m)$ із заданою кількістю вершин (n) та з m неорієнтованими ребрами, розділеними між заданою кількістю спільнот (K). Зручно вважати, що ребра пофарбовані K різними кольорами, що представляють спільноти, до яких вони належать. Нехай модель параметризується набором параметрів θ_{iz}, які представляють собою схильність вершини i мати ребра кольору z. Тоді добуток $\theta_{iz}\theta_{jz}$ є очікуваною кількістю ребер кольору z, які лежать між вершинами i та j.

Варто зазначити, що переважна більшість мереж реального світу є дуже розрідженими: фактична кількість ребер m набагато менша, ніж максимально можлива $m_{full} = n(n-1)/2$. Тому утворення ребра між вершинами i та j можна розглядати як рідкісне явище: ймовірність такої події є дуже малою. Якщо припустити, що такі рідкісні явища можуть відбуватись декілька разів та ймовірність кожної події не залежить від того, скільки разів ця подія вже відбулася (чи взагалі не відбулася), то розподіл кількості подій описується добре відомим у теорії ймовірностей законом Пуассона. За цим законом ймовірність того, що випадкова величина X (з невід'ємним цілим носієм) прийме значення k дорівнює

$$\Pr(X = k) = \frac{\lambda^k}{k!}\exp(-\lambda) \to Pois(X;\lambda), \qquad (2.1)$$

де λ є очікуваним значенням цієї випадкової величини ($E[X] = \lambda$).

Таким чином, будемо вважати, що точне число ребер кольору z, які лежать між вершинами i та j, розподілено за законом Пуассона (2.1) навколо

середнього значення $\lambda_{ijz} = \theta_{iz}\theta_{jz} = 1$.

Слід зауважити, що технічно це означає, що мережа $G(n,m)$ є мультиграфом та псевдографом, тобто вона може мати більше одного ребра між парою вершин, а також може мати петлі (тобто ребра які з'єднують вершину з собою). Більшість реальних мереж мають лише поодинокі ребра між вершинами, тому може здаватися, що така модель нереалістична. Проте, дозволення мультиребер та петель робить модель набагато простішою. З іншого боку, для переважної більшості графів та розбиттів їх на спільноти, значення θ_{iz} будуть дуже малими, отже ймовірність виникнення дублюючих ребер буде замалою, а отже, і внесена помилка, також мала. Можна зазначити, що аналогічне наближення зроблено в більшості інших випадкових графових моделей мереж, включаючи, наприклад, широко досліджену конфігураційну модель [4].

У запропонованій моделі, яка буде використовуватись у роботі, спільноти посилань виникають неявно під час створення мережі, а не прописані явно. Дві вершини i,j, які мають великі значення θ_{iz} і θ_{jz} для деякого значення z, мають високу ймовірність бути з'єднаними ребром кольору z, і, отже, групи таких вершин будуть мати тенденцію з'єднуватися відносно щільними мережами ребер кольору z. Це є саме та структура, яка найкраще відповідає структурі мережі з спільнотами посилань.

2.2 Виявлення спільнот, які перекриваються

Враховуючи модель, визначену вище, можна легко записати ймовірність, з якою генерується будь-яка конкретна мережа. Відомо, що сума з z незалежних випадкових величин з розподілом Пуассона (2.1) $Pois(X;\lambda_z)$ з параметрами λ_z також є випадковою величиною з розподілом Пуассона з параметром $\sum_z \lambda_z$. Тому очікувана загальна кількість ребер усіх кольорів між будь-якими двома вершинами i та j дорівнює просто

$\lambda_{ij} = \sum_z \theta_{iz}\theta_{jz}$ (або $\lambda_{ii} = \frac{1}{2}\sum_z \theta_{iz}^2$ для петель), а фактичне число ребер A_{ij} є розподіленим за законом Пуассона з цим середнім. Таким чином, ймовірність утворення графа $G(n,m)$ з елементами матриці суміжності A_{ij} дорівнює

$$P(G\,|\,\theta) = \prod_{i<j} \frac{(\lambda_{ij})^{A_{ij}}}{A_{ij}!}\exp\left(-\lambda_{ij}\right) \times \prod_i \frac{(\lambda_{ii})^{A_{ii}/2}}{(A_{ii}/2)!}\exp\left(-\lambda_{ii}\right), \qquad (2.2)$$

де $\lambda_{ij} = \sum_z \theta_{iz}\theta_{jz}$ при $i \neq j$, та $\lambda_{ii} = \frac{1}{2}\sum_z \theta_{iz}^2$.

При цьому враховано, що елемент матриці суміжності A_{ij} приймає значення $A_{ij} = 1$, якщо між вершинами i та j є ребро, але $A_{ii} = 2$, якщо вершина i має петлю. Саме тому з'явились множники 1/2 у другому добутку.

Згідно з методом максимальною правдоподібності модель є найкращим чином узгодженою з спостережуваною мережею, якщо значення ймовірності (2.2) є максимальним. Зазвичай максимізують не саму цю ймовірність (відносно параметрів θ_{iz}), а її логарифм, що є суттєво зручнішим. Взявши логарифм виразу (2.2) та відкинувши адитивні та мультиплікативні константи (які не впливають на положення максимуму), ми отримуємо логарифмічну функцію правдоподібності:

$$\log P(G\,|\,\theta) = \sum_{i,j}\left(A_{ij}\log(\lambda_{ij}) - \lambda_{ij}\right) = \sum_{i,j}\left(A_{ij}\log\left(\sum_z \theta_{iz}\theta_{jz}\right) - \sum_z \theta_{iz}\theta_{jz}\right). \qquad (2.3)$$

Пряма максимізація цього виразу шляхом диференціювання призводить до набору нелінійних неявних рівнянь відносно шуканих значень θ_{iz}, які важко розв'язати навіть чисельно. Простіший підхід полягає в наступному: спочатку застосуємо нерівність Йенсена:

$$\log\left(\sum_z x_z\right) \geq \sum_z \left(q_z \log \frac{x_z}{q_z}\right), \tag{2.4}$$

де x_z – довільний набір позитивних чисел, а q_z – довільні ймовірності, тобто невід'ємні числа, які задовольняють умові $\sum_z q_z = 1$.

Можна зазначити, що завжди можна досягти точної рівності у (2.4), якщо призначити $q_z = x_z / \sum_z x_z$. Застосовуючи (2.4) до рівняння (2.3), можна отримати, що

$$\log P(G \mid \theta) \geq \sum_{i,j,z} \left(A_{ij} q_{ijz} \log \frac{\theta_{iz}\theta_{jz}}{q_{ij}(z)} - \theta_{iz}\theta_{jz}\right), \tag{2.5}$$

де ймовірності q_{ijz} можна вибрати будь-яким способом, за умови, що вони задовольняють нормуванню $\sum_z q_{ijz} = 1$.

Важливо зазначити, що q_{ijz} визначені лише для таких пар вершин i, j, які фактично з'єднані ребром у мережі (таких що $A_{ij} = 1$), а отже, для кожного кольору z цих коефіцієнтів є стільки, скільки спостережуваних ребер, тобто m (а не $n^2 \gg m$).

Оскільки, як зазначалося, точна рівність у цьому виразі завжди може бути досягнута відповідним вибором q_{ijz}, то звідси випливає, що подвійна максимізація правої частини (2.5) відносно обох змінних q_{ijz} і θ_{iz} еквівалентна максимізації первісної логарифмічної правдоподібності (2.3) відносно θ_{iz}. Може здатися, що це не спрощує нашу задачу оптимізації: нам вдалося лише перетворити одну оптимізацію на подвійну, що можна було б уявити як ускладнення проблеми, проте це не так: подвійна оптимізація насправді дуже проста.

Для заданих значень θ_{iz}, оптимальні значення q_{ijz} дорівнюють

$$q_{ijz} = \frac{\theta_{iz}\theta_{jz}}{\sum_z \theta_{iz}\theta_{jz}}, \tag{2.6}$$

оскільки саме такі значення роблять нерівність (2.5) точною рівністю.

Враховуючи (2.6), оптимальні значення θ_{iz} можна знайти шляхом диференціювання (2.5), що дає

$$\theta_{iz} = \frac{\sum_j A_{ij} q_{ijz}}{\sum_i \theta_{iz}}. \tag{2.7}$$

Підсумовуючи цей вираз по i, отримуємо:

$$\left(\sum_i \theta_{iz} \right)^2 = \sum_{i,j} A_{ij} q_{ijz} \,. \tag{2.8}$$

Підставивши (2.8) у (2.7), матимемо:

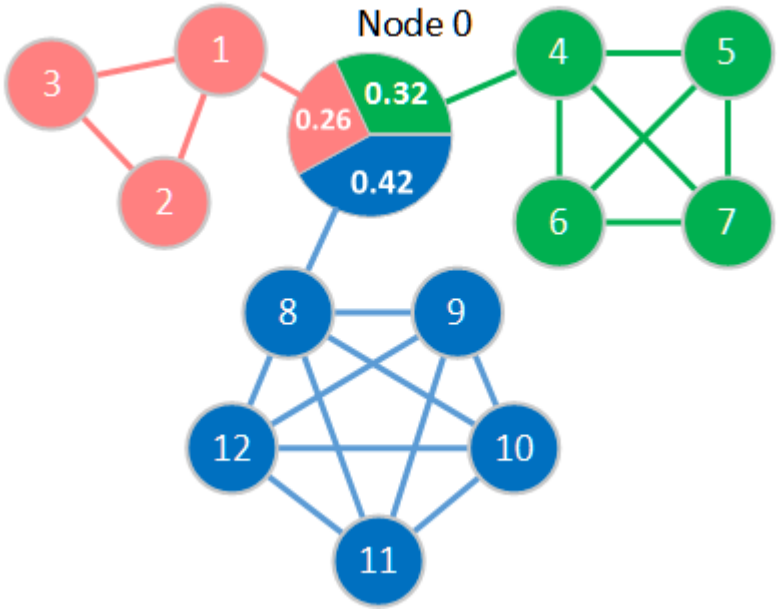
$$\theta_{iz} = \frac{\sum_j A_{ij} q_{ijz}}{\sqrt{\sum_{i,j} A_{ij} q_{ijz}}}. \tag{2.9}$$

Таким чином, максимізація логарифмічної правдоподібності (2.3) зводиться до одночасного вирішення рівнянь (2.6) і (2.9), яке можна виконати ітераційно, вибравши випадковий набір початкових значень і чередуючи обчислення між θ_{iz} та q_{ijz} за формулами (2.6) і (2.9). Цей підхід відомий як EM-алгоритм (Expectation-Maximization algorithm). Цей алгоритм широко використовується в математичній статистиці для знаходження оцінок максимальної схожості параметрів ймовірних моделей, у випадку, коли модель залежить від прихованих змінних. Кожна ітерація

алгоритму складається з двох кроків. На E-кроці (expectation) вираховується очікуване значення параметрів моделі θ_{iz}, при цьому приховані змінні q_{ijz} розглядаються як спостережувані. На M-кроці (maximization) вираховується оцінка прихованих параметрів, яка максимізує логарифмічну правдоподібність. Потім це значення використовується для E-кроку на наступній ітерації. І так далі. Алгоритм виконується до збіжності обох параметрів, і доведено, що логарифмічна правдоподібність монотонно зростає за результатами кожної ітерації, хоча збігається не обов'язково до глобального максимуму, тобто часто – саме до одного з багатьох локальних.

Для того, щоб запобігти застряганню в локальному максимумі, можна продублювати весь розрахунок кілька разів із випадковими початковими умовами та обрати той результат, який дає найвищу остаточну логарифмічну правдоподібність.

На рисунку 2.1 показано приклад застосування EM-алгоритму (2.6), (2.9) для простої мережі, яка, очевидно, містить три спільноти. Після 20-ї ітерації значення θ_{iz} та q_{ijz} стабілізувались.



Theta values			
	Community		
Node	**1**	**2**	**3**
0	0.21	0.27	0.35
1	0	0	1.06
2	0	0	0.71
3	0	0	0.71
4	0	1.07	0
5	0	0.80	0
6	0	0.80	0
7	0	0.80	0
8	1.07	0	0
9	0.85	0	0
10	0.85	0	0
11	0.85	0	0
12	0.85	0	0

Рисунок 2.1 – Приклад розбиття мережі на спільноти

При цьому кольори вузлів $1 \div 12$ визначились чітко, а вузол 0 відноситься одночасно до всіх трьох спільнот у пропорціях 0.26, 0.32 та 0.42. Можна зазначити, що ці пропорції відповідають кількості вузлів у «чистих» спільнотах (3, 4 та 5 відповідно).

Значення q_{ijz} у рівнянні (2.6) мають просту інтерпретацію: це ймовірності того, що ребро між i та j має колір z, тобто q_{ijz} є саме тією величиною, яка потрібна, щоб зробити висновок про спільноти посилань у мережі. При цьому можна зазначити, що q_{ijz} є симетричною матрицею відносно i, j, як це і має бути для неорієнтованої мережі.

2.3 Зв’язок моделей виявлення спільнот у мережі та моделі PLSA щодо статистичного аналізу тексту

Наведений метод математично тісно пов’язаний з методами, розробленими в спільноті машинного навчання для аналізу текстових документів. Зокрема, генеративну модель, яка застосовується, можна розглядати як варіант моделі, що використовується в ймовірнісному прихованому семантичному аналізі (PLSA) – техніці для автоматичного виявлення тем у корпусі тексту – адаптований до поточного контексту спільнот посилань. Особливий інтерес становить робота [10], яка, хоч і не зосереджується на спільнотах зв’язків, але використовує інший варіант моделі PLSA, поєднуючи її з ітеративним алгоритмом підгонки, що називається факторизацією невід’ємної матриці, щоб знайти спільноти, що перекриваються, у спрямованих мережах. Також варто відзначити роботу [6], де розглядається модель спільноти посилань, але при цьому застосовується алгоритмічний підхід, заснований на байєсівській генеративній моделі.

Використовувана генеративна модель є дещо модифікованим еквівалентом моделі, яка застосовується в аналізі тексту і називається ймовірнісним латентним семантичним аналізом (PLSA) [10].

Класична проблема аналізу тексту, яка вирішується за допомогою методу PLSA, полягає в аналізі «корпусу» текстових документів для пошуку наборів слів, які всі (або більшість) зустрічаються в тих самих документах. Припускається, що ці набори слів відповідають темам, які можна використовувати для групування документів за змістом.

Підхід PLSA розглядає документи як так званий «мішок слів», тобто враховується лише кількість разів, коли кожне слово зустрічається в документі, а не порядок, у якому вони зустрічаються. Крім того, часто розглядається лише підмножина цікавих слів, а не всі слова, які містяться у корпусі. Математично корпус із D документів і W цікавих слів представлений матрицею входжень A з елементами A_{wd}, які дорівнюють кількості входжень слова w в документ d. Цю модель можна розглядати як зважену дводольну мережу, яка має один набір вершин для документів, інший для слів, а вага ребер, які з'єднують слова з документами, дорівнює кількості входжень слова w в документ d.

У PLSA кожна пара слово-документ асоціюється з неспостережуваною змінною z, яка позначає одну з K тематичних груп. Припускається, що кожне ребро розміщується незалежно випадковим чином у дводольному графі, причому ймовірність того, що ребро потрапляє між словом w і документом d, представляється у формі

$$P(w,d) = \sum_z P(w|z)P(d|z)P(z), \qquad (2.10)$$

де $P(z)$ – це ймовірність того, що довільне ребро належить до теми z;

$P(w|z)$ – ймовірність того, що ребро з темою z з'єднується зі словом w;

$P(d|z)$ – ймовірність того, що ребро із темою z з'єднується з документом d.

Враховуючи існування тем, кінці кожного ребра документа та слова розміщуються незалежно. Автор PLSA Хофманн [10] називає цю

параметризацію «симетричною», маючи на увазі, що слово та документ відіграють математично еквівалентні ролі. Проте, ця «симетричність» має не той сенс, який є звичним: мережа дводольна, а матриця входжень A_{wd} не тільки не є симетричною, а й навіть не обов'язково квадратна.

Альтернативний опис моделі, який є корисним для створення матриці входжень, та який відповідає формулюванню проблеми еквівалентної мережі, полягає в тому, що кожен матричний елемент A_{wd} приймає випадкове значення, отримане незалежно від інших за розподілом Пуассона із середнім значенням

$$\sum_z P(w|z)P(d|z)w_z . \qquad (2.11)$$

Таким чином, кожне ребро розміщується незалежно від інших із ймовірністю (2.10), де $P(z) = w_z / \sum_z w_z$.

В моделі спільнот мереж матриця суміжності (на відміну від матриці входжень) завжди симетрична, тому параметр w_z є зайвим, і у генеративній моделі мереж зі спільнотами він був опущений.

PLSA передбачає використання реберної ймовірності (2.10) для обчислення правдоподібності за всім набором слів та документів, а потім її максимізацію відносно невідомих ймовірностей $P(w|z)$, $P(d|z)$ та $P(z)$. Отримані ймовірності дають змогу визначити, наскільки сильно кожне слово або документ пов'язано з певною темою z, але оскільки теми є довільними, це фактично те саме, що просто групувати слова та документи в «спільноти». Крім того, можна використовувати ймовірності, щоб розділити ребра дводольного графа між тематичними групами, даючи текстовий еквівалент «спільнот посилань».

Було досліджено ряд методів для максимізації правдоподібності в задачі статистичного аналізу текстів. У первісній роботі з PLSA [10] використовувався саме Expectation-Maximization алгоритм, однак його

відповідність версії (2.6), (2.9) не є точною, оскільки матриця входжень слів у документи є несиметричною. Можна узагальнити цей метод на орієнтовані одномодові мережі, розглядаючи неорієнтовану мережу як орієнтовану із спрямованими ребрами (дугами), які проходять в обох напрямках між кожною зв'язаною парою вершин. Проте, на жаль, цей підхід не працює, оскільки загалом він призводить до апостеріорної ймовірності q_{ijz}, яка є несиметричною щодо i та j. Ця асиметрія викликає проблеми у ассоціюванні спільнот з неорієнтованими ребрами. Якщо $q_{ij} \neq q_{ji}$, то ребро може бути в одній спільноті, якщо розглядати дугу від i до j, і в іншій спільноті, якщо це ребро розглядати як дугу з j до i.

Таким чином, використувану у [10] версію EM-методу неможливо безпосередньо застосувати для стандартних ненаправлених одномодових мереж. Замість застосування стандартної моделі PLSA до моделей мереж, кращим підходом є використання моделі, яка є внутрішньо симетричною з самого початку. Саме такою є модель, сформульована у розділі 2.1.

Їснує декілька інших пов'язаних підходів, у тому числі заснованих на методах, відомих як невід'ємна матрична факторизація (NMF) та латентне виділення Діріхле (LDA). Ці моделі також є асиметричні (і, отже, непридатні для неорієнтованих мереж) і використовують різні алгоритмічні підходи для максимізації ймовірності. Модель NMF схожа за стилем на алгоритм EM, в ній використовується ітераційна схема максимізації, але конкретні ітераційні рівняння відрізняються.

2.4 Зменшення обчислювальної складності EM-алгоритма розбиття мереж на спільноти

Описаний вище метод може бути реалізований безпосередньо як комп'ютерний алгоритм для пошуку спільнот, які перекриваються, та добре працює для мереж помірного розміру, до десятків тисяч вершин. Для великих мереж як обсяг використаної пам'яті, так і час виконання стають

значними та обмежують застосуванню методу до великих мереж. Проте, обидві характеристики складності можуть бути покращені за допомогою більш тонкої реалізації, яка має зробити можливим обробку мереж з мільйонами вершин.

Використання пам'яті алгоритму визначається простором, необхідним для зберігання параметрів: θ_{iz} вимагає $O(nK)$ пам'яті, тоді як q_{ijz} потребує $O(mK)$ одиниць, де n і m це кількість вершин і ребер у мережі. Оскільки m значно більше за n, це означає, що q_{ijz} є домінуючою складовою щодо використання пам'яті. Проте, є можливим реорганізувати використувану версію EM-алгоритму (2.6), (2.9) таким чином, щоб q_{ijz} не потрібно було зберігати. Введемо нову змінну k_{iz} – середню кількість кінців ребер кольору z, з'єднаних з вершиною i:

$$k_{iz} = \sum_j A_{ij} q_{ijz} . \tag{2.12}$$

Тоді алгоритм обчислення величин θ_{iz}, k_{iz} можна реалізувати наступним чином: спочатку визначається новий набір величин k_{iz}^{new}, в якому будуть зберігатись нові значення k_{iz}. На початку ітерації всі вони встановлюються на нуль. Потім обчислюється середнє число d_z ребер кольору z, підсумованих за всіма вершинами:

$$d_z = \sum_i k_{iz} . \tag{2.13}$$

Тоді первісні параметри θ_{iz} (2.9) дорівнюватимуть

$$\theta_{iz} = \frac{k_{iz}}{\sqrt{d_z}} . \tag{2.14}$$

Проводиться ітерація по всім ребрам (i,j) мережі. Спочатку для кожного ребра мережі обчислюється знаменник рівняння (2.6), який з урахуванням (2.14) матиме вигляд

$$D = \sum_z \theta_{iz}\theta_{jz} = \sum_z \frac{k_{iz}k_{jz}}{d_z}. \tag{2.15}$$

Маючи значення D (2.15) та d_z (2.13), можна обчислити q_{ijz} з (2.6):

$$q_{ijz} = \frac{\theta_{iz}\theta_{jz}}{\sum_z \theta_{iz}\theta_{jz}} = \frac{k_{iz}k_{jz}}{D \cdot d_z}. \tag{2.16}$$

Тепер можна додати ці значення до величин $k_{iz}^{new}, k_{jz}^{new}$ згідно з (2.12) та повторити розрахунки (2.15)–(2.16) для наступного ребра мережі.

Після завершення цього циклу величини k_{iz}^{new} дорівнюватимуть сумі (2.12), а отже будуть правильними новими значеннями k_{iz}.

Запропонована варіація методу вимагає зберігати лише старі та нові значення k_{iz} із загальною кількістю $2nK$, а не значення q_{ijz} кількістю mK, що призводить до значної економії пам’яті для великих мереж.

Що стосується часу роботи, то EM-алгоритм (2.12)–(2.16) має обчислювальну складність $O(mK)$ операцій на ітерацію, де m є кількістю ребер у мережі. Основним визначником загального часу виконання є кількість ітерацій, які необхідно виконати, перш ніж значення k_{iz} зійдуться. Можна зазначити, що в багатьох випадках потрібна досить велика кількість ітерацій, що сповільнює продуктивність методу, проте швидкість алгоритму можна покращити.

У типовому випадку розбиття мереж на спільноти кінцевим результатом є те, що кожна вершина належить не всім можливим K спільнотам (кольорам), а лише невеликій підмножини з K. Іншими словами,

велика частка параметрів k_{iz} прагне до нуля під час ітерації ЕМ.

З рівнянь (2.12), (2.16) можна побачити, що якщо деякий з коефіцієнтів k_{iz} коли-небудь стане нульовим, то він залишиться таким для всіх майбутніх ітерацій. Це означає, що його більше не потрібно оновлювати, і таким чином можна заощадити час, виключивши його з обчислень. Можна застосувати дві корисні стратегії для скорочення набору цих параметрів. Згідно з першій з них, якщо деякий k_{iz} падає нижче попередньо визначеного порогу δ, то він встановлюється на нуль. Після того, як k_{iz} було встановлено на нуль, відповідні значення q_{ijz} на всіх суміжних ребрах також дорівнюють нулю, тому їх не потрібно обчислювати. Таким чином, для кожного ребра потрібно обчислити значення q_{ijz} лише для тих кольорів z, для яких і k_{iz}, і k_{jz} відмінні від нуля, тобто лише для перетину наборів кольорів у вершинах i та j.

Ця стратегія призводить до значного збільшення швидкості в тих випадках, коли кількість спільнот K велика. Для невеликих значень K економія швидкості переважується додатковими обчислювальними витратами, і ефективніше було б просто обчислити всі q_{ijz}, але все ж таки обрізку k_{iz} на нуль нижче порогового значення δ робити доцільно, оскільки це відкриває шлях для використання другої стратегії прискорення.

Друга стратегія пришвидшення, яку можна використовувати як окремо, так і разом із першою, мотивується ти фактом, що якщо для деякої вершини i значення k_{iz} для всіх кольорів z, крім одного, встановлені на нуль, то колір цієї вершини фіксується на цьому одному значенні і більше не змінюватиметься. Більше того, якщо обидві вершини на кінцях ребра (i, j) мають цю властивість, тобто зійшлися до одного кольору і більше не змінюються, то ребро (i, j), яке їх з'єднує, більше не впливає на обчислення і може бути повністю видалено з мережі.

Завдяки використанню цих двох стратегій швидкість обчислень помітно покращується. Кількість параметрів k_{iz} і кількість невидалених ребер швидко і суттєво зменшується з ходом обчислень, так що більшість

ітерацій включає лише підмножину, пов’язану з вершинами, ідентифікація спільноти яких є найбільш неоднозначною. Якщо значення порогу δ встановлено рівним нулю, то модифікований (одній чи обома стратегіями) алгоритм точно еквівалентний первісному EM-алгоритму (2.12)–(2.16), але навіть за такого вибору покращення швидкості є суттєвим. Якщо ж δ мале, але ненульове (наприклад, $\delta = 0.001$), то розрахунок стає наближеним. Проте, ця похибка невелика, в той час як покращення швидкості є значним.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ТЕСТУВАННЯ АЛГОРИТМУ РОЗБИТТЯ МЕРЕЖ НА СПІЛЬНОТИ

Для перевірки працездатності та ефективності досліджуваних алгоритмів використовувались як синтетичні (згенеровані комп'ютером) мережі, так і датасети реальних мереж. Синтетичні мережі дозволяють перевірити здатність алгоритму виявляти заздалегідь відому (створену під час генерації мережі) структуру спільнот в контрольованих умовах, тоді як датасети реальних мереж дозволяють дослідити швидкодію алгоритма в практичних умовах.

Програмна реалізація алгоритму розбиття мереж на спільноти виконувалась з використанням мови Python у середовищі Spyder 5.5.5, бібліотеки NetworkX, пакету Gephi.

3.1 Тестування алгоритму розбиття на синтетичних мережах

Тестування алгоритму виявлення спільнот на синтетичних мережах має форму класичного тесту узгодженості: мережі створюються на підставі тієї самої стохастичної моделі, на якій базується сам алгоритм. Потім оцінюється здатність алгоритму відновлювати відомі поділи мереж на спільноти для різних значень параметрів цих мереж та спільнот. Можна варіювати параметри для створення мереж із чіткою структурою спільноти (що має полегшити виявлення) або взагалі без структури спільноти (що робить це неможливим), або проміжні варіанти. Таким чином можна змінювати складність завдання, яке ставиться перед алгоритмом.

Синтезовані мережі мають $n = 10000$ вершин, розділених на дві спільноти, що перекриваються. x вершин розміщуються в першій спільноті, тобто вони мають зв'язки лише в межах цієї спільноти, y вершини розташовуються у другій спільноті, а решта $z = n - x - y$ вершин – в обох спільнотах з рівною середньою кількістю зв'язків з вершинами кожної із

спільнот. Середнє значення ступіню усіх вершин є однаковим, дорівнює k.

Проведено три серії тестів. У першій розмір перекриття між спільнотами встановлений на $z = 500$, а решта вершин розподіляється між спільнотами порівну: $x = y = 4750$. Параметр k варіюється. Якщо $k \to 0$, то у мережі майже немає ребер і, отже, немає структури спільноти, і будь-який алгоритм розпізнавання спільнот не спрацює. З іншого боку, якщо k велике, то визначити структуру спільнот, має бути просто.

У другій серії тестів перекриття також становить $z = 500$, але цього разу значення $k = 10$ зафіксоване та варіюється баланс вершин між x та y.

Нарешті, у третій серії тестів $k = 10$, $x = y$, але варіюється розмір перекриття z.

На рисунку 3.1 показано виміряну частку правильно класифікованих вершин (чорна крива) у кожному з цих трьох наборів тестів (трьох окремих панелей). Вершина вважається правильно класифікованою, якщо в середньому вона має принаймні одне ребро відповідного кольору за результатом максимізації правдоподібності. Точніше кажучи, вершина належить до спільноти z, якщо її очікуваний ступінь відносно кольору z, визначений як k_{iz} (9), більший за одиницю.

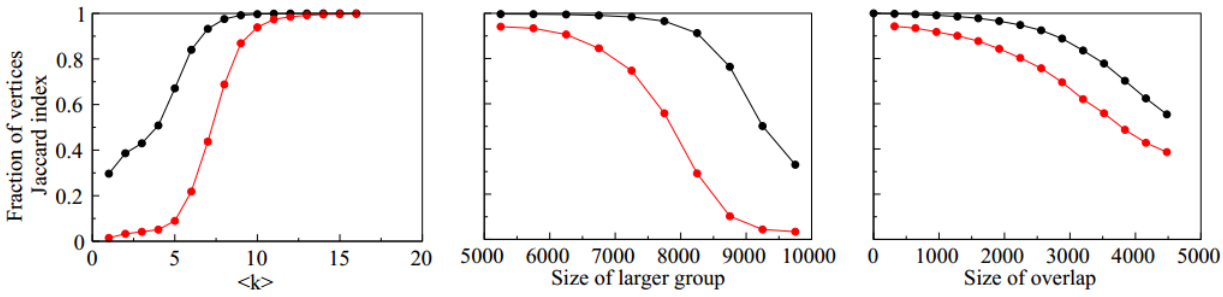


Рисунок 3.1 – Результати трьох серії тестів на синтезованих мережах

Кількість синтезованих мереж становить $N = 100$. Кожна точка даних усереднена для 100 мереж. Для кожної мережі було використано 20

випадкових ініціалізацій змінних (тобто 20 прогонів). Прогон, який дав найвище значення логарифмічної правдоподібності (2.3), був прийнятий як кінцевий результат. На кожній панелі чорна крива показує частку вершин, призначених правильним спільнотам за допомогою алгоритму, тоді як червона крива є індексом Жаккара для вершин у перекритті.

Як показано на рисунку 3.1, існують значні діапазони параметрів для всіх трьох тестів, для яких алгоритм працює добре, правильно класифікуючи більшість вершин у мережі. Як і очікувалося, точність у першому тесті зростає зі збільшенням k, і для значень $k > 10$ (значення, яке легко досягається багатьма мережами реального світу) алгоритм ідеально ідентифікує відому структуру спільноти. У двох інших тестах точність знижується через збільшення асиметрії двох груп або розміру перекриття, але наближається до 100%, коли ці фактори не є великими.

Щоб більш детально дослідити здатність алгоритму ідентифікувати спільноти, які перекриваються, для тих самих тестових мереж вимірювався індекс Жаккара: якщо S є множиною вершин у справжньому перекритті, а V – множина вершин, які алгоритм ідентифікує як такі, що знаходяться в зоні перекриття, то індекс Жаккара дорівнює

$$J = \frac{|S \cap V|}{|S \cup V|}. \tag{3.1}$$

Цей індекс є стандартним показником подібності між наборами, який «карає» як за помилкові позитивні, так і за помилкові негативні результати щодо включення (в даному випадку) вершин у перекриття. Значення індексу Жаккара (3.1) показані червоними кривими на рисунку 3.1. Можна бачити, що загальна форма кривих схожа на загальну частку правильно визначених вершин. Зокрема, для мереж з середнім ступенем $k > 13$ значення J прагне до 1, тобто перекриття ідентифікуються ідеально.

3.2 Тестування алгоритму розбиття на реальних мережах

Досліджуваний метод був протестований на реальних мережах. Першою тестовою мережею була та, яка стала фактично обов'язковою для тестів виявлення спільнот: це мережа клубу карате, вперше досліджена Захарі (Zachary's karate club, [11], [12]). Вона описує дружні стосунки між членами університетського спортивного клубу. Ця мережа цікава тим, що під час дослідження клуб розділився на дві частини в результаті внутрішньої суперечки, таким чином, кількість спільнот та розбиття на спільноти є відомими. Проте, неодноразово виявлялося, що майбутню «лінію розколу» цього клуба можна було спрогнозувати заздалегідь спираючись саме на структуру зв'язків між вершинами – членами клубу.

На рисунку 3.2 показано розбиття мережі членів клубу карате на дві групи, що перекриваються, яке знайдено запропонованим алгоритмом. Кольори на малюнку показують як поділ вершин, так і поділ ребер. Розкол між двома групами в клубі чітко проявляється в результатах і добре відповідає фактичній лінії розколу. Проте, EM-алгоритм призначає кілька вершин обом групам. Індивіди, представлені цими вершинами, є тими, хто має друзів в обох таборах. Можна припустити, що вони мали певні труднощі, вирішуючи, на чий бік суперечки виступити, і дійсно, оригінальне дослідження Захарі містить деякі вказівки на те, що це був саме той випадок.

Варто зазначити, що запропонований EM-алгоритм не тільки ідентифікує вершини, які перекриваються, але й визначає ступінь цього перекриття, тобто ступінь приналежності кожної вершини до тієї чи іншої спільноти. Частка обчислюється як очікувана частка ребер кожного кольору, що падає на вершину: k_{iz}/d_i. Це відображено на рисунку 3.2 шляхом фарбування вершин у перекритті круговими діаграмами. Кольори ребер відповідають найвищому значенню q_{ijz} для даного ребра, тоді як кольори вершин вказують частку інцидентних ребер, які потрапляють у

кожну спільноту. Для вершин у більш ніж одній спільноті вершини намальовані більшими для ясності та поділені на кругові діаграми, що представляють їх розподіл між спільнотами.

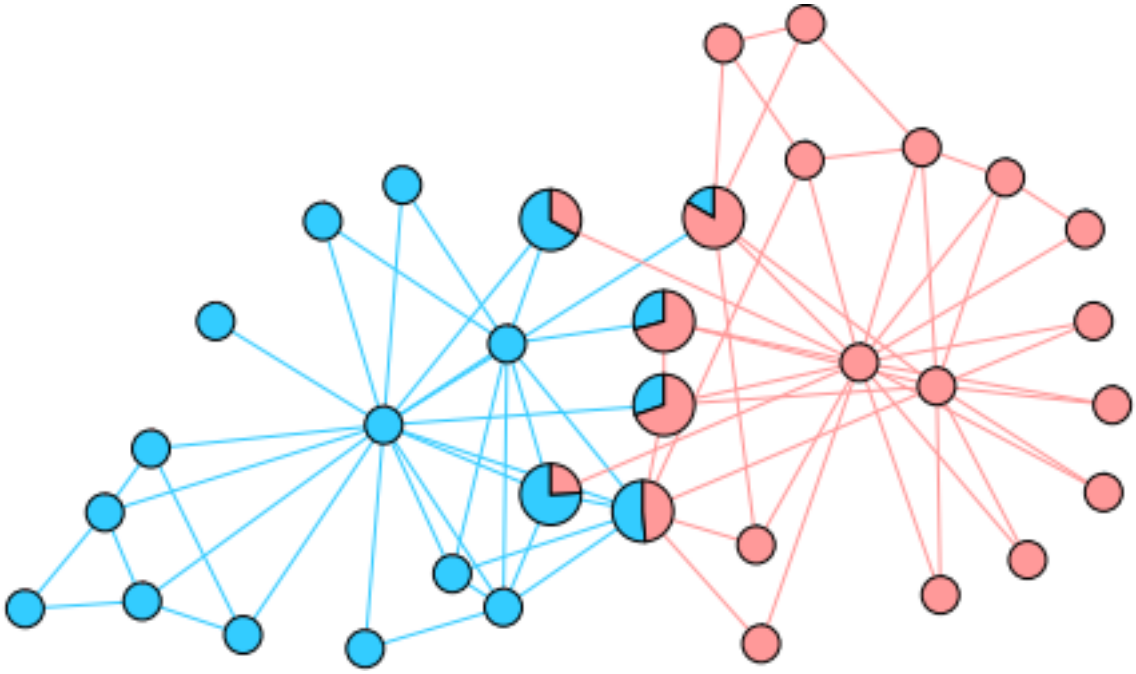


Рисунок 3.2 – Спільноти мережі клубу карате

Другим прикладом реальної мережі була також соціальна мережа, проте іншої природи: це мережа взаємодії між персонажами роману Віктора Гюго «Знедолені» («Les Misérables»), складена Кнутом [13]. У цій мережі два персонажі з’єднані ребром, якщо вони з’являються в одному розділі книги.

На рисунку 3.3 показано розділення мережі за допомогою запропонованого алгоритму на шість спільнот, що перекриваються. Цей розподіл приблизно відповідає соціальним розбіжностям і сюжетам у сюжетній лінії роману. Але що особливо цікаво в даному випадку, так це роль, яку відіграють хаби в мережі головні персонажі, які представлені вершинами особливо високого ступеня. У мережах багатьох типів часто зустрічаються вузли високого ступеня, які зв’язані з кожною частиною

мережі. Присутність таких хабів створює проблеми для традиційних схем виявлення таких спільнот, які не перекриваються, оскільки ці хаби незручно вписуються в будь-яку з чітких спільнот: незалежно від того, в яку спільноту цей хаб буде віднесено, він матиме багато зв’язків з вершинами в інших спільнотах, що погіршить чіткість виділення спільнот. Спільноти, що перекриваються, забезпечують елегантне вирішення цієї проблеми, оскільки подібні хаби можна розмістити в перекриттях.

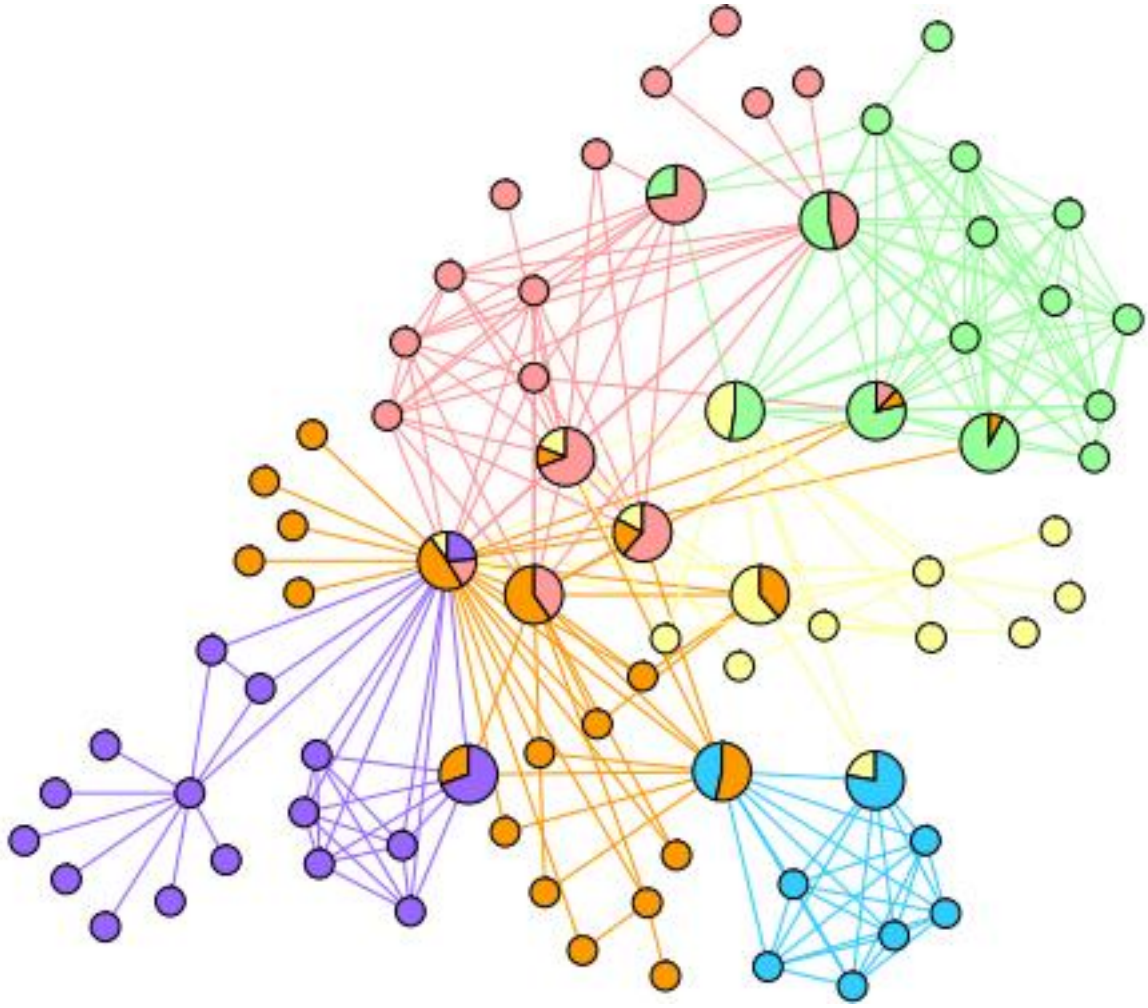


Рисунок 3.3 – Спільноти мережі персонажів з «Les Misérables»

Як показано на рисунку 3.3, EM-алгоритм робить саме це, розміщуючи хаби в двох або більше спільнотах. Таке розподілення виглядає цілком природним і для досліджуваної мережі: головні персонажі,

представлені хабами, є саме тими, хто з'являється більш ніж в одному з сюжетів (розділів) книги. Водночас, другорядні персонажі мають відносно «жорстку» спеціалізацію: вони висвітлюють головних персонажів (кожен свого), тому й з'являються у зв'язці саме з своїм головним персонажем та з його «світою другорядних».

3.3 Тестування алгоритму розбиття на мережах з спільнотами, які не перетинаються

Досліджуваний ЕМ-алгоритм пошуку спільнот призначений для пошуку у мережах таких спільнот, що перекриваються, проте його можна використовувати і для пошуку таких спільнот, що не перетинаються. Як відомо [4], будь-який алгоритм, який обчислює пропорційну приналежність вершин у спільнотах, може бути адаптований до випадку неперекриття шляхом віднесення кожної вершини до однієї спільноти – до тієї, до якої вона належить найбільше. У нашому випадку це означає віднесення вершини i до спільноти z, для якої значення θ_{iz} найбільше відносно z. Більше того, ця процедура може бути строго виправдана, якщо розглядати модель спільноти посилань як релаксацію стохастичної блочної моделі, що не перекривається, скоригованої за ступенем. При цьому виявляється критично важливим, щоб використовувана блокова модель включала знання про ступені вершин мережі. Це спонукає до розгляду так званої блокової моделі з корекцією ступенів: розглядається мережа з n вершин, кожна з яких належить рівно одній спільноті. Присвоєння спільноти представлене змінною-індикатором S_{ir}, яка приймає значення 1, якщо вершина i належить спільноті r, і нуль в іншому випадку. Число ребер між кожною парою вершин i, j вважається розподіленим за законом Пуассона (2.1) з очікуваним значенням $\theta_i \omega_{rs} \theta_j$, якщо вершина i належить групі r, а вершина j належить групі s, де θ_i та ω_{rs} є параметрами моделі. Таким чином, середнє значення елемента матриці суміжності A_{ij}дорівнює $\theta_i \left(\sum_{rs} S_{ir} \omega_{rs} S_{js}\right) \theta_j$ для кожної

пари вершин. Нормалізація параметрів може бути довільною, оскільки завжди можна масштабувати всі θ_i на будь-яку константу, якщо одночасно масштабувати всі ω_{rs}. Проте для спрощення розрахунків доцільно нормувати параметри θ_i таким чином, щоб сума їхня сума дорівнювала одиниці в межах кожної спільноти:

$$\forall r: \quad \sum_i \theta_i S_{ir} = 1. \tag{3.2}$$

Тепер можна адаптувати цю модель до спостережуваної мережі, записавши ймовірність генерації мережі як добуток ймовірностей Пуассона для кожного (мульти)ребра, а потім максимізуючи її щодо параметрів θ_i та ω_{rs} та призначень спільноти S_{ir}. Але, у той час як максимізація ймовірності (чи логарифмічної правдоподібності) відносно безперервних параметрів θ_i та ω_{rs} виконується відносно просто, через диференціювання, то максимізація відносно дискретних змінних S_{ir} набагато складніша.

Поширеним способом вирішення таких задач є релаксація (послаблення) дискретних змінних, тобто дозвіл їм приймати безперервні дійсні значення. Тоді оптимізацію можна було б виконати шляхом диференціювання. У даному випадку можна б було дозволити S_{ir} приймати довільні невід'ємні значення з урахуванням нормалізації (3.2). За таким підходом S_{ir} представлятиме частку, на яку вершина i належить групі r. Після такої релаксації можна поглинути параметри θ_i в S_{ir}, визначивши $\theta_{ir} = \theta_i S_{ir}$ з нормуванням $\sum_i \theta_{\mathrm{ir}} = 1$. Тоді середня кількість ребер між вершинами i та j стає $\sum_{rs} \theta_{ir} \omega_{rs} \theta_{js}$. Таким чином ми отримали розширену форму моделі спільнот, які перекриваються, що досліджувалась в даній кваліфікаційній роботі, узагальненої для включення додаткової $K \times K$ матриці ω_{rs}. Це узагальнення дає модель, у якій два кінці ребра можуть належати до різних спільнот. Можна вважати, що кожен кінець ребра забарвлений власним кольором, замість того, щоб усе ребро набувало лише

одного кольору. Ребра будуть одноколірними тільки якщо матриця ω_{rs} буде діагональною.

Проведемо підгонку загальної (недіагональної) моделі до спостережуваної мережі за допомогою вже використаного раніше алгоритму ЕМ. Позначаючи ймовірність q_{ijrs} того, що ребро між i та j має кольори r та s, рівняння ЕМ (2.6), (2.9) матимуть наступний вигляд:

$$q_{ijrs} = \frac{\theta_{ir}\omega_{rs}\theta_{js}}{\sum_{rs}\theta_{ir}\omega_{rs}\theta_{js}}, \tag{3.3}$$

$$\theta_{ir} = \frac{\sum_{js} A_{ij}q_{ijrs}}{\sum_{ijs} A_{ij}q_{ijrs}}, \qquad \omega_{rs} = \sum_{ij} A_{ij}q_{ijrs}. \tag{3.4}$$

Ітеруючи ці рівняння, можна знайти рішення для параметрів θ_{ir}. Оскільки $\theta_{ir} = \theta_i S_{ir}$, то підсумовуючи обидві частини цієї рівності по r, та враховуючи, що $\sum_r S_{ir} = 1$, ми отримуємо, що $\sum_r \theta_{ir} = 1$. Отже,

$$S_{ir} = \frac{\theta_{ir}}{\theta_i} = \frac{\theta_{ir}}{\sum_r \theta_{ir}}. \tag{3.5}$$

Обчисливши S_{ir} згідно з (3.5), можна змінити релаксацію моделі, округливши значення до нуля або одиниці, що еквівалентно присвоєнню кожної вершини i тієї спільноти r, для якої S_{ir} є найбільшим, або, що еквівалентно, тієї спільноти r, для якої θ_{ir} є найбільшим.

Таким чином, алгоритм для поділу мережі у разі неперекриття спільнот полягає в тому, щоб просто послабити задачу, припустивши перекриття, розв’язати задачу поділу мережі на спільноти, які перетинаються, застосувавши для цього ЕМ-алгоритм у вигляді (2.6), (2.9), або (2.12)–(2.16), а після цього призначити кожну вершину такій спільноті,

для якої θ_{ir} є найбільшою.

У звичайній версії ЕМ-алгоритму матриця ω_{rs} була відсутня, що (з урахуванням можливості її довільної нормалізації) еквівалентно припущенню, що вона діагональна. Однак було виявлено, що навіть якщо дозволити ω_{rs} бути матрицею загального виду, то алгоритм зазвичай зберігає ненульовими лише діагональні значення. Це означає, що результат первісного ЕМ-алгоритму (для розбиттів з перекриттями) та узагальненого алгоритму (для розбиттів без перекриттів) має бути однаковим. В той же час, важливим є те, що діагональна версія алгоритму працює значно швидше, ніж узагальнена.

Проте, цілком можливо, що можуть існувати мережі з цікавою недіагональною груповою структурою, які можна виявити за допомогою більш загальної моделі. Модель, що включає матрицю ω_{rs}, може в принципі знайти дизасортативну структуру спільноти, тобто структуру, в якій зв'язки менш поширені всередині спільнот, ніж між ними, а також краще вивчену асортативну структуру. Наприклад, модель може виявити двосторонню структуру в мережах, тоді як невідкоригована модель не може.

Для тестування працездатності досліджуваного алгоритму на мережах із спільнотами, які не перетинаються, використовувалось тестування як на штучних (синтезованих) мережах, так і на реальних.

На рисунку 3.4 наведено приклад тестування досліджуваного алгоритму розбиття на мережі ігор з американського футболу [11]. У цій мережі вершини представляють університетські команди, а ребра представляють розклад ігор на сезон 2000 року. Дві команди з'єднуються, якщо вони грали між собою. Відомо, що університетські команди поділені на конференції. У 2000 році було 11 конференцій серед команд Дивізіону I-A, які складають мережу, а також 8 команд, які не входять до жодної з конференцій.

Кластери вершин представляють спільноти, знайдені за допомогою модифікованого алгоритму, тоді як кольори вершин представляють

конференції, на які офіційно розділені коледжі. Темно-фіолетові вершини представляють незалежні коледжі, які не належать до жодної конференції. Можна бачити, що всі команди, які належать до якоїсь конференції, класифікуються вірно.

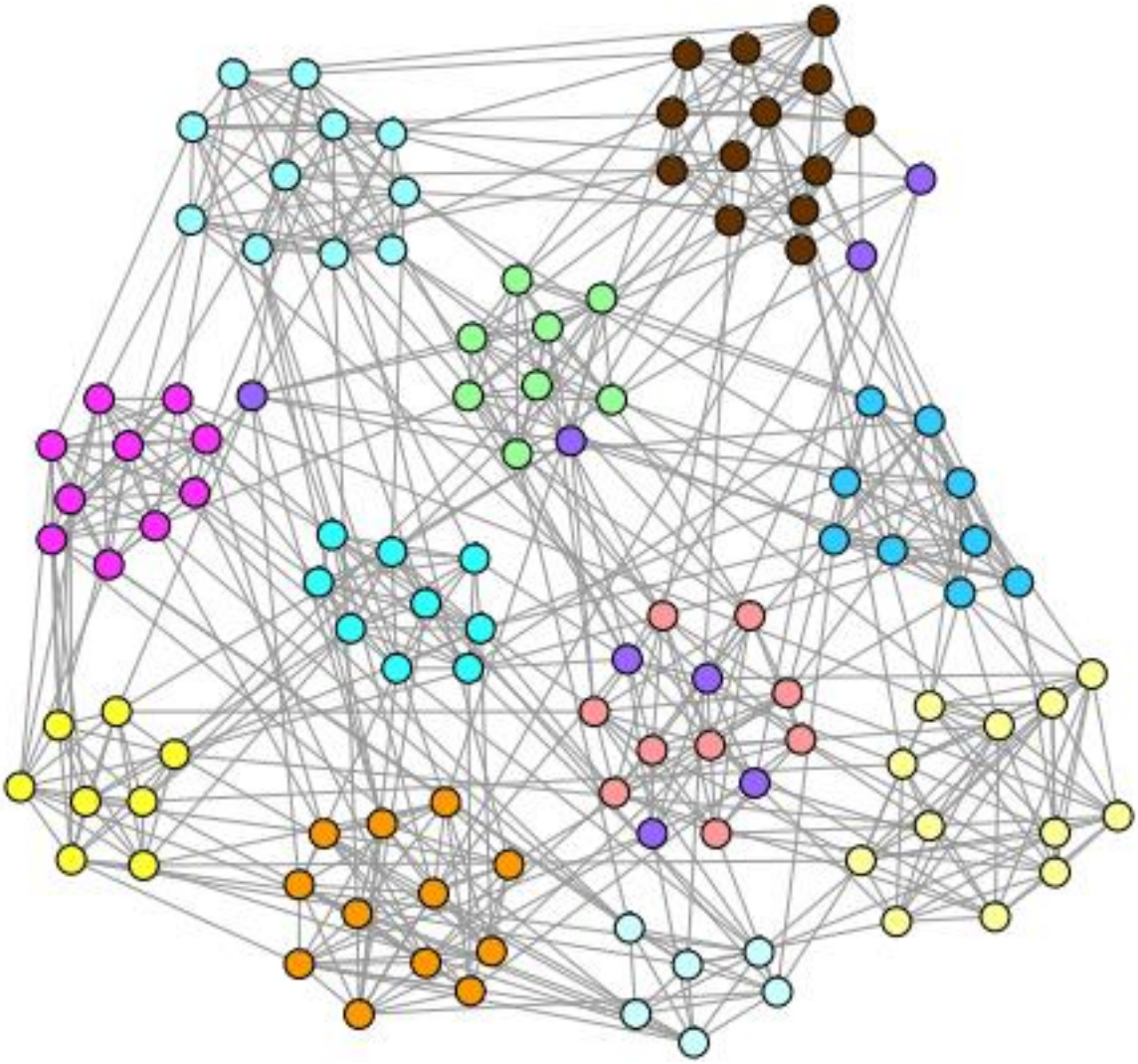


Рисунок 3.4 – Спільноти, які не перекриваються, знайдені в мережі футбольних змагань коледжів США

3.4 Дослідження швидкодії застосовуваної версії ЕМ-алгоритму

Як було зазначено у розділі 2, проста реалізація рівнянь ЕМ у формі (2.6), (2.9) дає алгоритм, який є недостатньо швидким для великих

мереж. Тому було запропоновано більш досконалу реалізацію цього методу у вигляді (2.12)–(2.16), під час якої непотрібні змінні поступово видаляються з ітерацій, що має призвести до значно більшої швидкості. Було проведено порівняння цих двох версій алгоритму за швидкодією на тестових мережах.

Результати підсумовані в таблиці 3.1, де наведено час центрального процесора (у секундах), витрачений на виявлення спільнот, що перекриваються для кожної з тестових мереж.

Таблиця 3.1 – Результати порівняння алгоритмів розбиття мереж на спільноти за швидкодією

Алгоритм	Час роботи (с)	Кількість ітерацій	Логарифмічна правдоподібність
Штучна мережа з 13 вузлів (рис. 2.1), n = 13, m = 22, K = 3			
Простий	0.127	2317	-53.33
Швидкий, $\delta = 0$	0.094	2781	-53.33
Швидкий, $\delta = 0.001$	0.109	2594	-52.37
Мережа karate club, n = 34, m = 78, K = 2			
Простий	0.640	3378	−235.11
Швидкий, $\delta = 0$	0.390	6160	−235.11
Швидкий, $\delta = 0.001$	0.484	6325	−233.56
Мережа персонажів «Знедолені», n = 77, m = 254, K = 6			
Простий	5.79	7874	−660.03
Швидкий, $\delta = 0$	3.50	14563	−660.03
Швидкий, $\delta = 0.001$	3.85	15266	−685.07
Мережа політичних блогів, n = 1224, m = 16 715, K = 2			
Простий	892	3583	−65360
Швидкий, $\delta = 0$	171	15797	−65350
Швидкий, $\delta = 0.001$	369	25347	−66100

У цих тестах було використувано 100 випадкових ініціалізацій змінних. Під час тестів орієнтовані мережі були штучно симетризовані. З мережі polblogs було прибрано висячі вершини. В якості кінцевого

результату береться прогін, який дає найвище значення логарифмічної правдоподібності (2.3).

Для кожної мережі представлено результати трьох різних розрахунків:

– розрахунок, виконаний за допомогою простого алгоритму EM (2.6), (2.9);

– обчислення з використанням скороченого алгоритму (2.12)–(2.16) з пороговим параметром δ, встановленим на нуль; такий алгоритм дає результати, ідентичні простому алгоритму;

– обчислення, виконане з використанням скороченого алгоритму з $\delta = 0.001$, який вводить додаткове наближення, яке зазвичай призводить до дещо нижчого кінцевого значення логарифмічної правдоподібності, але дає значне додаткове підвищення швидкості.

Таким чином, результати експерименту підтверджують працездатність та ефективність реалізованого алгоритму розбиття спільнот. Можна зазначити, що під час розбиття мереж на невелику кількість спільнот (значення K є маленьким) найшвидшим є швидкий алгоритм без додаткового вилучення ребер (тобто з $\delta = 0$). Але чим більшою є кількість спільнот, тим ефективнішим стає додаткове вилучення ребер ($\delta = 0.001$).

ВИСНОВКИ

Предметом досліджень кваліфікаційної роботи є задача розбиття мережі на групи (кластери, ком’юніті, спільноти). Відомо, що пошук науково обгрунтованих методів виявлення спільнот у мережах на основі структури зв’язків в них є водночас актуальною науковою проблемою теорії графів та мереж, а з іншого боку, ця задача є важливою у практичному сенсі.

В роботі розглядається задача розбиття мережі на спільноти з припустимістю перекриттів між ними на основі побудови генеративної моделі мережі. У такому разі розбиття мережі визначається параметрами цієї моделі, які в свою чергу можуть бути знайдені методом максимальної правдоподібності. Зазначений підхід є обгрунтованим з ймовірнісно-статистичної точки зору.

Проведено дослідження генеративних моделей мереж. Зазначено, що задача, яка розглядається, має тісний зв’язок з задачею статистичного аналізу текстів, проте відповідний алгоритм PLSA, відомий у машинному навчанні, не може бути безпосередньо перенесений на задачу розбиття мереж. Визначено, що ефективним методом пошуку параметрів моделі є широко відомий метод очікування-максимізації (EM-алгоритм).

В роботі проведено аналіз властивостей EM-алгоритма, зокрема його обчислювальної складності. Виконано модифікацію цього алгоритма, яка дозволяє суттєво прискорити його та одночасно зменшити витрати пам’яті.

Здійснено програмну реалізацію дослідженого методу та відповідних алгоритмів. Проведено тестування працездатності та ефективності обраного підходу з використанням як штучно згенерованих мереж, так і типових датасетів мереж. Результати тестування підтвердили високу швидкість та одночасно прийнятну ефективність досліджуваного методу розбиття мереж на спільноти та відповідних алгоритмів, які реалізують цей метод.

Результати роботи були представлені на 13-й Міжнародній науково-технічній конференції «Інформаційні системи та технології ІСТ-2024» [14].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Girvan M., Newman M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*. 2002. Vol. 99, no. 12. P. 7821–7826. URL: https://doi.org/10.1073/pnas.122653799 (date of access: 07.12.2024).

2. Fortunato S. Community detection in graphs. *Physics Reports*. 2010. Vol. 486, no. 3–5. P. 75–174. URL: https://doi.org/10.1016/j.physrep.2009.11.002 (date of access: 07.12.2024).

3. Newman M. E. J. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter*. 2004. Vol. 38, no. 2. P. 321–330. URL: https://doi.org/10.1140/epjb/e2004-00124-y (date of access: 07.12.2024).

4. Ball B., Karrer B., Newman M. E. J. Efficient and principled method for detecting communities in networks. *Physical Review E*. 2011. Vol. 84, no. 3. URL: https://doi.org/10.1103/physreve.84.036103 (date of access: 07.12.2024).

5. Karrer B., Newman M. E. J. Stochastic blockmodels and community structure in networks. *Physical Review E*. 2011. Vol. 83, no. 1. URL: https://doi.org/10.1103/physreve.83.016107 (date of access: 07.12.2024).

6. Parkinnen J., Gyenge A., Sinkkoken J., Kaski S. A block model suitable for sparse graphs. In Proceedings of the 7th International Workshop on Mining and Learning with Graphs, Association of Computing Machinery, New York. 2009.

7. Network community detection using modified modularity criterion / V. Shergin et al. *Eastern-European Journal of Enterprise Technologies*. 2024. Vol. 6, no. 4 (132). P. 6–13. URL: https://doi.org/10.15587/1729-4061.2024.318452 (date of access: 07.12.2024).

8. Shergin V.L., Grinyov S.A., Chala L.E., Udovenko S.G. Noncorrelational assortativity coefficient for networks with nominative attributes. 10-th IEEE International Scientific-Practical Conference Problems of

Infocommunications, Science and Technology (PIC S&T), Kharkiv, Ukraine. 2024.

9. Noldus R., Van Mieghem P. Assortativity in complex networks. *Journal of Complex Networks*. 2015. Vol. 3, no. 4. P. 507–542. URL: https://doi.org/10.1093/comnet/cnv005 (date of access: 07.12.2024).

10. Hofmann T. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*. 2004. Vol. 22, no. 1. P. 89–115. URL: https://doi.org/10.1145/963770.963774 (date of access: 07.12.2024).

11. Network data. *University of Michigan*. URL: https://public.websites.umich.edu/~mejn/netdata/ (date of access: 07.12.2024).

12. Zachary W. W. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*. 1977. Vol. 33, no. 4. P. 452–473. URL: https://doi.org/10.1086/jar.33.4.3629752 (date of access: 07.12.2024).

13. Knuth D. E. The Stanford GraphBase: A platform for combinatorial computing. New York, N.Y : ACM Press, 1993. 576 p.

14. Шергін В., Крамаренко Д., Гриньов С. Дослідження складності алгоритмів розбиття мереж на спільноти, які перекриваються. 13-та Міжнародна науково-технічна конференція «Інформаційні системи та технології ІСТ-2024», 26–28 листопада, 2024, Харків. 2024