Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Розробка та дослідження методу виявлення спільнот у мережах на основі модулярності
(тема)

Виконав:
здобувач другого року навчання,
групи ДСм-23-1

Мусієнко М.Ю.
(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 проф. Кіріченко Л.О.
(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 Мусієнку Михайлу Юрійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Розробка та дослідження методу виявлення спільнот у мережах на основі модулярності

затверджена наказом університету від 22 листопада 20 24 р. № 1238Ст
2. Термін подання студентом роботи до екзаменаційної комісії 16 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
2) Виявлення мережевих спільнот на основі модулярності
3) Чисельні експерименти на тестових мережах

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	09.12.2024	виконано
	спільноти		
5	Дослідження методів розбиття мереж,	16.12.2024	виконано
	заснованих на максимізації модулярності		
6	Алгоритмічна реалізація виявлення мережевих	23.12.2024	виконано
	спільнот на основі максимізації модулярності		
7	Програмна реалізація алгоритмів розбиття мереж	30.12.2024	виконано
8	Проведення експериментальних досліджень	06.01.2025	виконано
9	Написання пояснювальної записки	13.01.2025	виконано
10	Захист перед ЕК	18.01.2025	

Дата видачі завдання 25 листопада 20 24 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Кіріченко Л.О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 59 с., 15 рис., 2 дод., 15 джерел.

ЖАДІБНИЙ АЛГОРИТМ, МЕРЕЖІ, МОДИФІКОВАНА МОДУЛЯРНІСТЬ, МОДУЛЯРНІСТЬ, СПІЛЬНОТИ.

Об’єктом досліджень є мережі.

Предметом досліджень кваліфікаційної роботи є методи розбиття мереж на спільноти.

Метою роботи є дослідження методів та алгоритмів розбиття мереж на спільноти на основі максимізації критерію модулярності мережі.

ABSTRACT

Master’s thesis contains: 59 p., 15 fig., 2 ann., 15 references.

COMMUNITIES, GREEDY ALGORITHM, MODIFIED MODULARITY, MODULARITY, NETWORKS.

The object of research are 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 communities based on maximizing the network modularity criterion.

ЗМІСТ

ВСТУП

Однією з актуальних проблем аналізу структури мереж є проблема виявлення спільнот. Наприклад, люди, як користувачі соціальних мереж, можуть групуватись у професійні, мовні, географічні спільноти [1], [2]. Групування вузлів мереж біологічної, біохімічної природи відображає спеціалізацію відповідних організмів, клітин, речовин тощо. При цьому ознаки, за якими відбувається групування, їхня кількість, ступінь впливу, зазвичай є невідомими. Так само, як невідомою є й кількість спільнот, існуючих у мережі. З огляду на зазначені властивості проблеми, яка розглядається, цілком природнім є те, що не існує (і не може існувати [3], [4]) загальної постановки задачі виявлення спільнот, загальних критеріїв якості розбиття тощо. Через це, незважаючи на велику кількість публікацій із даної тематики, кількість та розмаїття запропонованих методів та алгоритмів, проблема розбиття мережі на спільноти залишається актуальною в науковому сенсі та одночасно практично значущою.

Всі існуючі методи розбиття мереж на спільноти можна умовно розділити на модель-орієнтовані, критерій-орієнтовані та алгоритм-орієнтовані. В основі модель-орієнтованих методів є ймовірнісно-статистична генеративна модель мережі [2], [5]. Тоді розбиття мережі еквівалентно пошуку параметрів цієї моделі. Наявність обгрунтованої моделі є перевагою методів даного класу.

В основі критерій-орієнтованих методів є деякий критерій – модулярність, який описує якість розбиття. Відповідно, алгоритм розбиття за своєю сутністю є пошуком локального максимуму цього критерію [6]. Основною перевагою цього підходу є гнучкість у визначенні критерію та відносно велика швидкодія відповідних алгоритмів.

В основі алгоритм-орієнтованих методів лежать чітко визначені алгоритми розбиття мереж на спільноти. Наприклад, виділення клік, або покрокове видалення вузлів, які є найпотужнішими посередниками [7].

Перевагами цього підходу є «прозорість» процедури розбиття, легка інтерпретованість результатів.

Основну увагу в кваліфікаційній роботі було зосереджено на методах розбиття, заснованих на максимізації модулярності мережі. В основі критеріїв модулярності лежить така, цілком природня, вимога, що щільність зв’язків всередині спільнот має бути вищою, ніж між вузлами з різних спільнот.

Зазначено, що розбиття мереж за традиційним критерієм модулярності часто призводить до поглинання малих спільнот більшими. Тому велику увагу було приділено застосуванню модифікованої модулярності [4], яка безпосередньо залежить від кількості вузлів у спільнотах.

Здійснюється алгоритмічна реалізація методів розбиття мереж на основі максимізації модифікованої модулярності. Запропоновані алгоритми відносяться до класу жадібних та є доволі швидкодіючими.

У практичній частині роботи виконується програмна реалізація обраних методів та алгоритмів, здійснюється тестування їхньої працездатності та ефективності. Тестування виконується на типових мережевих датасетах.

1 ОГЛЯД ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Огляд сучасного стану проблеми виявлення спільнот у мережах

Багато різних мереж реального світу, включаючи біологічні та соціальні мережі, природно діляться на модулі або спільноти, тобто групи вершин із відносно щільними зв’язками всередині груп та з розрідженими зв’язками між вузлами різних груп [1], [3], [7]. Важливою науковою проблемою
є виявлення таких спільнот в мережах. Існує 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 багато прикладів мереж, як природних, так і синтетичних, для яких структура спільнот є відомою, і успішний метод виявлення спільнот повинен знайти їх. З іншого боку, якщо «справжня» структура спільнот мережі заздалегідь невідома, то емпірично знайдена структура повинна бути інтерпретованою.

По-друге, бажано, щоб метод мав у своїй основі науково обгрунтовані теоретичні підстави. Проте, наявність зазначеної бази не гарантує практичну значущість отриманого розбиття.

По-третє, для того щоб метод міг успішно застосовуватись, обчислювальна складність його алгоритмічної реалізації повинна бути відносно невеликою. Інакше кажучи, алгоритми повинні бути швидкими. Так, розмір досліджуваних мереж може сягати багатьох мільйонів вузлів та мільярдів зв’язків між ними. Тому алгоритм виявлення спільнот, час роботи якого лінійно залежить від розміру мережі, має надзвичайну перевагу, над алгоритмом з квадратичною чи вищою складністю.

Одним з найважливіших класифікуючих факторів щодо методів виявлення спільнот є припустимість, або неприпустимість перетинань між спільнотами. При цьому будь-яку задачу чіткого розбиття (тобто в умовах неприпустимості перетинань) можна розглядати як частковий випадок задачі нечіткого розбиття, тобто розбиття з перетинаннями. Так, у практичних ситуаціях спільноти часто можуть частково перекриватись. Моделі спільнот посилань відкривають можливість виявлення динамічних або накладених спільнот, де одна вершина або ребро може належати до кількох груп одночасно. Наприклад, у соціальних мережах люди часто належать до кількох кіл знайомих – родина, друзі, колеги тощо.

Тоді ці кола слід розглядати як такі, що перекриваються, 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.

Одним з підходів до розв’язання задачі виявлення спільнот є підхід, заснований на використанні стохастичної генеративної моделі мережі. Цей підхід, який застосовує методи статистичного виводу до мереж, досліджувався для випадку неперекриття [2]. Однак поширення цього підходу на випадок припустимості перекриттів є нетривіальним [5]. 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» [5], 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.

Найбільш продуктивними серед ймовірнісно-статистичних методів виявлення спільнот, що перекриваються, є такі, що грунтуються на моделюванні та дослідженні спільнот за зв’язками, а не за вершинами. Природньо, що будь-яке розбиття зв’язків мережі на спільноти автоматично породжує й розбиття вершин. І якщо мережа є зв’язною, то спільноти вершин будуть перетинатись. Ця ідея була незалежно запропонована рядом авторів, зокрема й у машинному навчанні [8]. 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.

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.

Іншим класом методів розбиття мереж на спільноти є такі, що засновані на евристичних функціях якості, оптимізованих щодо можливих розділень вузлів або зв’язків мережі [3]. Такі функції якості, зокрема функція модулярності [6], 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 мати багато локальних максимумів. Крім того, з суто формально-теоретичної точки зору, модулярність (у традиційній формі) є не єдино можливою евристичною функцією якості.

Таким чином, першим етапом роботи є аналіз поняття складних мереж, властивостей мереж, моделей мереж, понять центральності, асортативності та модулярності мереж, методів обчислень цих параметрів.

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.

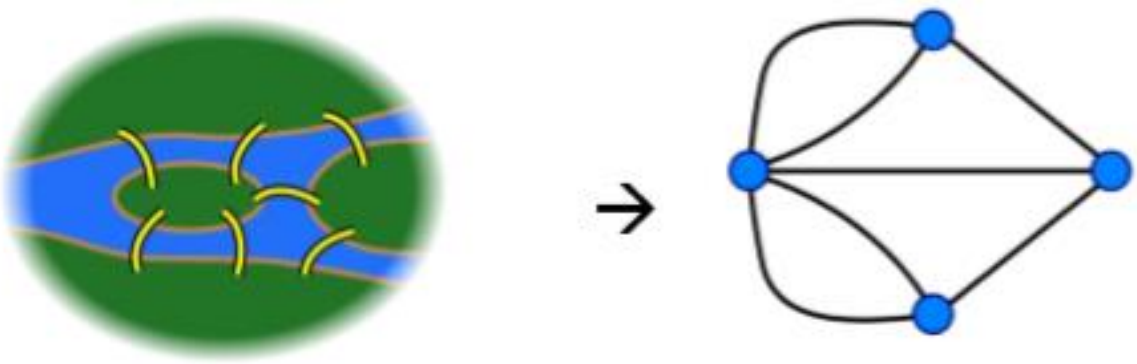

Рисунок 1.1 – Схема мостів Кенигсберга

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 – $O(|V|+|\mathrm{E}|)$. 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 –

$$\rho_{\log} = \lambda - 1, \qquad (1.2)$$

де параметр λ є показником ступеню, який пов'язує кількість зв'язків та кількість вузлів мережі: $E(n) \underset{n\to\infty}{\to} Cn^{\lambda}$. З цього випливає, що

$$\lambda = \lim_{n\to\infty} \frac{\log(E/2)}{\log(n/2)}. \tag{1.3}$$

Формула (1.3) є асимтотичною, а більш точно значення λ (яке називається показником еластичності мережі) можна знайти як розв'язок рівняння

$$E = \frac{\Gamma(n+\lambda-1)}{\Gamma(n-1)\Gamma(\lambda+1)}. \tag{1.4}$$

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

Для будь-якої деревовидної мережі (зокрема, тієї, що на рисунок 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 компоненти 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 повного графу довільного розміру дорівнює двом. Відповідно, ступенева щільність повного графу є $\rho_{log} = 1$, тобто співпадає із звичайною щільністю (1.1).

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 $\rho_{\log} = (\sqrt{33}-5)/2 \approx 0.3723$. 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:

$$l_s = \frac{2}{n(n-1)} \sum_{i \geq j} l_{ij}\,, \tag{1.5}$$

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:

$$l_{\max} = \max_{i,j}\left(l_{ij}\right). \tag{1.6}$$

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 (1.5) та (1.6) також дорівнюватимуть нескінченості, тобто не будуть мати сенсу. Тому більш універсальною мірою ніж середня найкоротша відстань (1.5) є довжина середнього інверсного шляху, яка визначається наступним чином:

$$l_{inv} = \frac{n(n-1)}{2}\left(\sum_{i\geq j} l_{ij}^{-1}\right)^{-1}. \tag{1.7}$$

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:

$$C_i = \frac{\Delta_i}{V_i}. \qquad (1.8)$$

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: $C_4 = 0/0 = 0$.

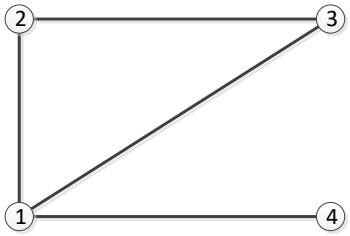

Рисунок 1.3 – Приклад мережі з чотирма вузлами та чотирма зв’язками

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

$$C_i = \frac{\sum_{j,m} A_{ij} A_{jm} A_{mi}}{k_i (k_i - 1)}, \qquad k_i = \sum_j A_{ij} = \deg(i). \qquad (1.9)$$

де підсумовування ведеться по всіх вузлах j, m.

Якщо коефіцієнти кластеризації окремих вузлів ((1.8), або (1.9)) відомі, то можна обчислити коефіцієнт кластеризації всієї мережі шляхом усереднення коефіцієнтів кластеризації вузлів:

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i \,. \qquad (1.10)$$

Для мережі, яка розглядалась (рисунок 1.3) $C = (1/3 + 1 + 1 + 0)/4 = 7/12$.

Дещо схожою з (1.10), але трохи іншою мірою кластеризації вузлів мереж, є показник транзитивності:

$$T = \frac{tr(A^3)}{\sum_i k_i(k_i - 1)} = \frac{\sum_i (A^3)_{ii}}{\sum_i k_i(k_i - 1)}, \qquad (1.11)$$

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

$$T = \frac{\sum \Delta_i}{\sum V_i}. \qquad (1.12)$$

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

.

1.3 Типові моделі 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.

Наразі існує багато різних моделей складних мереж. Найбільш

поширеними да дослідженими з них є модель випадкового графу (модель Ердеша-Рен`ї), модель безмасштабної мережі (модель Барабаші-Альберт) та модель «тісного світу».

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

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:

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}. \qquad (1.13)$$

Довжина середнього інверсного шляху між вузлами (1.7) становить

$$l_{inv} \propto \log(N). \qquad (1.14)$$

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):

$$p_k = C \cdot k^{-\alpha}. \qquad (1.15)$$

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 (1.15) 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:

$$p_i = \frac{k_i}{2(n-1)}, \tag{1.16}$$

де n – поточний розмір мережі.

Для мережі БА середня довжина найкоротшого шляху між вузлами набагато менше, ніж для ER-моделі (1.14) та дорівнює

$$l_s \propto \log(\log(n)). \tag{1.17}$$

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

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}, \quad (k \geq m). \tag{1.18}$$

Варто зауважити, що розподіл (1.18) є дискретним варіантом ступеневого розподілу (1.15) 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 рисунок 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.

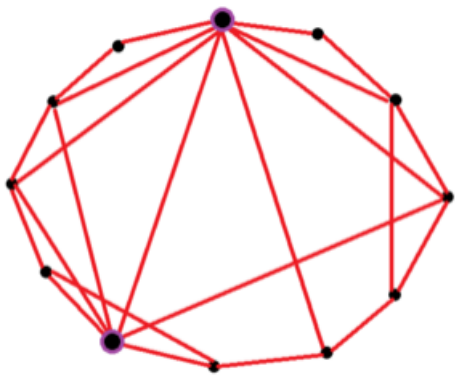

Рисунок 1.5 – Модель WS мережі «тісного світу»

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 $1 < \ln n < k \ll n$.

1.4 Постановка задачі дослідження

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

– дослідження критерію модулярності, властивостей розбиттів, заснованих на максимізації модулярності;

– дослідження модифікованого критерію модулярності;

– алгоритмічна реалізація розбиття вузлів мережі на спільноти на основі максимізації модифікованого критерію модулярності;

– оцінка обчислювальної складності алгоритмів розбиття;

– програмна реалізація алгоритмів виявлення мережевих спільнот на основі максимізації модулярності, експериментальна перевірка працездатності та ефективності алгоритмів розбиття за допомогою відомих тестових датасетах мереж.

2 ВИЯВЛЕННЯ МЕРЕЖЕВИХ СПІЛЬНОТ НА ОСНОВІ МОДУЛЯРНОСТІ

2.1 Поняття модулярності мереж

Загальновідомо, що мережі реального світу не є однорідними. Виявлення угруповань є одним із найважливіших завдань аналізу складних мереж. Так, розбиття мережі на спільноти дозволяє масштабувати мережу, зіставивши спільнотам вихідної мережі вузли нової та перейти таким чином до розгляду мережі у більшому масштабі. При цьому уточнюється та прояснюється структура зв'язків між елементами мережі та виявляється масштаб спільнот. Крім того, розбиття мережі на спільноти дозволяє виявити нетипові вузли та зв'язки, що є критично важливими з погляду цілісності мережі, поширення інформації в ній та т. і. Таким чином, розробка методів виявлення спільнот у мережах є актуальним практичним завданням.

При цьому, незважаючи на високу практичну значущість проблеми, яка розглядається, на сьогоднішній день не існує універсальних, науково обґрунтованих методів її вирішення. Насамперед це зумовлено прикладним, інженерним характером самої проблеми, що зазвичай не передбачає єдиної формалізованої постановки задачі, а також відсутністю обґрунтованих критеріїв якості розбиття мережі на спільноти. Другою причиною є велика різноманітність самих мереж. Так, використання деякого формалізованого критерію та розробка відповідних методів та алгоритмів розбиття мережі можуть бути науково обґрунтованими, визнаватися і використовуватися фахівцями-прикладниками, але з часом ці методи та алгоритми «спростовуються» контрприкладом, тобто наявністю такої мережі, для якої запропоноване розбиття буде далеким від оптимального. Крім того, важливою властивістю розглянутої задачі є її висока обчислювальна складність, що унеможливлює розв’язання методом повного перебору навіть для мереж невисокої розмірності. З цієї причини деякі методи,

незважаючи на вагоме математичне обґрунтування, не застосовуються на практиці. І навпаки, існують та успішно застосовуються напівемпіричні методи та підходи, що не мають строгого обґрунтування, але показують ефективність при вирішенні практичних задач. Можна сказати, що нестрогість постановки та нерозв'язність загальної проблеми оптимального розбиття вузлів мережі на спільноти є причиною та обґрунтуванням актуальності наукових досліджень, спрямованих на вирішення задач виявлення спільнот у складних мережах.

Як було зазначено вище, завдання поділу мережі на спільноти є дуже загальним, відповідно, існує безліч різних відтінків у постановці цього завдання, що відображають очікувані властивості об'єкта (тобто особливості зв'язків у мережі та вимоги до якості її розбиття). Так, зв'язки у мережі можуть бути спрямованими чи неспрямованими; мати довільну дійсну вагу або, наприклад, лише невід'ємну чи лише бінарну (0 / 1); множинні зв'язки та петлі можуть бути припустимими або ні.

Характерною класифікуючою вимогою до властивостей розбиття є припустимість або неприпустимість перетинів виділених спільнот. Найбільш простим для аналізу (але і найбільш практично важливим) є випадок розбиття вузлів мережі на множини, що попарно не перетинаються, який і є основним предметом подальшого аналізу.

Таким чином, об'єктом дослідження є складні мережі, математичною моделлю яких є неорієнтований зважений граф $G(V,E)$.

Для розглянутої задачі існує три основні сімейства методів розв'язання: на основі методу максимальної правдоподібності [2], [5], на основі центральності [7] та на основі максимізації модулярності [4], [6]. Методи, засновані на максимальній правдоподібності, мають суворе статистичне обґрунтування: вони зводяться до максимізації міри Кульбака-Лейблера (розбіжності по розподілу між розбиттям вузлів по спільнотах, що формуються) і деякою нуль-моделлю, що описує випадкове розбиття. Важливою перевагою таких методів є суворе обгрунтування як самих

методів, і властивостей нуль-моделі, у протистояння з якою й здійснюється розбиття. Найважливішим недоліком методів, заснованих на максимальній правдоподібності, є досить велика обчислювальна складність відповідних алгоритмів, що обмежує їхню практичну застосовність. Важливо, що у всіх випадках задача виявлення угруповань вузлів мережі має високу обчислювальну складність, тому оптимальне її рішення для мереж великого розміру неможливо. Це породжує необхідність застосування різних субоптимальних, зазвичай жадібних, алгоритмів. Крім того, варіативність математичного опису нуль-моделей є невисокою, а вплив моделі, що використовується, на властивості одержуваних варіантів поділу мережі не завжди очевидний.

Метод Гірвана-Ньюмана [7] реалізує поділ мережі шляхом покрокового видалення ребер, що мають найбільший вплив на центральність (у сенсі посередництва). Цей метод має високу обчислювальну складність ($O(m^2 n)$, де m – число ребер, а n – число вузлів мережі) і тому не застосовується для більшості мереж реального світу.

У прикладному аспекті найбільш поширеними методами розбиття вузлів мережі є методи, що ґрунтуються на максимізації модулярності [4], [6]. Головною їх перевагою є відносна простота реалізації і, відповідно, висока продуктивність.

Проблема полягає в розбитті множини вузлів мережі на спільноти. Припускається, що ці спільноти не перетинаються попарно, тому розбиття на спільноти $C_u, u = 1,...,K$ називається припустимим, якщо ці спільноти утворюють повну групу: $Nodes = \bigcup_{u=1}^{K} C_u, \quad \forall u \neq v: \; C_u \bigcap C_v = \varnothing$. Кількість шуканих спільнот (K) у загальному випадку вважається невідомою.

Розв'язання поставленої задачі засноване на максимізації цільової функції $Q(G, C_1,...,C_K)$, яка називається модулярністю.

Модулярність мережі є показником схильності вузлів групуватися в

кластери, які також називаються спільнотами. Якщо вузли в одній спільноті тісно пов’язані один з одним, а зв’язки між вузлами з різних спільнот рідкісні, то показник модулярності є високим, інакше – низьким.

Визначення модулярності мережі засноване на використанні в якості нуль-моделі випадкового графу із збереженням ступенів вузлів (тобто ваг зв'язків) аналізованої мережі. У випадковому неорієнтованому графі $G(V, E)$ з N вузлами з вагою вузлів k_i, $i = 1,...,N$ очікувана вага ребра між вузлами i, j дорівнює $k_i k_j / (2m)$, де $m = \frac{1}{2}\sum_{i=1}^{N} k_i$ це сумарна вага ребер мережі.

Можна зауважити, що для незважених мереж ваги вершин дорівнюють їх ступеням ($k_i = \deg_i$), вага всієї мережі дорівнює кількості ребер $m = size(G)$, а очікувана вага ребра між вузлами i, j дорівнює ймовірності наявності цього ребра.

Для кожної пари вузлів i, j можна розрахувати відхилення фактичної ваги зв'язку між ними (тобто елемента матриці суміжності $A_{i,j}$ зваженого графа) від очікуваної ваги зв'язку:

$$\Delta_{i,j} = A_{i,j} - \frac{k_i k_j}{2m}. \qquad (2.1)$$

Модулярність спільноти (C_u) дорівнює сумі значень (2.1), взятої по всіх вузлах, що входять до цієї спільноти:

$$q_u = \frac{1}{2m}\sum_{i,j \in C_u}\left(A_{i,j} - \frac{k_i k_j}{2m}\right) = \frac{1}{2m}\left(L_u^{in} - \frac{(L_u^{tot})^2}{2m}\right), \qquad (2.2)$$

де L_u^{in} – вага всіх посилань у спільноті u (для незваженого графа), що дорівнює подвоєній кількості ребер, обидва кінці яких належать спільноті; $L_u^{tot} = \sum_{i \in C_u} k_i$ – загальна вага всіх вузлів, включених до спільноти C_u.

Модулярність мережі $Q(G)$ визначається як сума модулярностей всіх спільнот цієї мережі. Таким чином, показник модулярності мережі є адитивним щодо спільнот та досить просто обчислюється. Це зумовлює широке застосування показника модулярності для оцінки якості розбиття вузлів мережі спільноти.

У той же час, вирішення зворотньої задачі (виявлення спільнот та розбиття вузлів на спільноти) шляхом максимізації модулярності мережі пов'язане з відомими проблемами. Поперше, кількість розбиттів множини з n елементів (відома як число Белла) зростає дуже швидко – субфакторіально:

$$B_n = O\left((n / \log n)^n\right). \tag{2.3}$$

Тому максимізація модулярності мережі $Q(G)$ відносно всіх можливих розбиттів шляхом прямого перебору є неможливою.

Зазвичай для розбиття вузлів на спільноти шляхом максимізації модулярності використовують субоптимальні жадібні алгоритми, найвідомішим з яких є алгоритм Louvain [6].

Другою із проблем є обмежена роздільна здатність [3] алгоритмів пошуку угруповань, заснованих на модулярності. Це означає, що вони прагнуть формувати близькі за розміром спільноти і тому погано виділяють спільноти, що складаються з небагатьох вузлів, тоді як у реальних мережах такі можуть існувати. Для подолання цього недоліку у формулу визначення модулярності спільноти (2.2) вводиться додатковий параметр – коефіцієнт роздільної здатності $\gamma > 0$ (за замовчуванням $\gamma = 1$):

$$q_u = \frac{1}{2m} \sum_{i,\, j \in C_u} \left(A_{i,j} - \gamma \frac{k_i k_j}{2m} \right) = \frac{1}{2m} \left(L_u^{in} - \gamma \frac{(L_u^{tot})^2}{2m} \right). \tag{2.4}$$

Якщо $\gamma < 1$, то алгоритм надає перевагу більшим спільнотам, інакше – меншим.

З урахуванням (2.4) формула розрахунку модулярності мережі набуває наступного вигляду:

$$Q(G) = \sum_{u=1}^{K} q_u = \frac{1}{2m} \sum_{u=1}^{K} \left(L_u^{in} - \gamma \frac{(L_u^{tot})^2}{2m} \right). \quad (2.5)$$

Важливо відзначити, що значення $\gamma \neq 1$ порушує статистичний зміст модулярності як відхилення фактичної кількості зв'язків усередині спільноти від очікуваного. Більш того, варіювання значення коефіцієнту роздільної здатності не завжди дозволяє виділити невеликі спільноти.

Згідно з [4], проблема полягає в тому, що нуль-модель (випадковий граф), що використовується в класичному підході, не передбачає наявності спільнот. Більше того, модулярність спільнот (2.2), (2.4) не залежить безпосередньо від кількості вузлів у цих спільнотах.
У зв'язку з цим було запропоновано модифікувати визначення модулярності окремих спільнот мережі (2.4) та мережі в цілому (2.5). Пропоновані зміни відображатимуть вплив чисельності вузлів у спільнотах, відповідно максимізація модифікованої модулярності дасть можливість виявляти дрібні спільноти, якщо вони присутні у мережі.

Крім того, у традиційних методах розбиття вузлів шляхом максимізації критерію (2.5) (наприклад, алгоритм Louvain [3], [6]) не приділяється належної уваги обґрунтованості вибору значення кількості спільнот (K): реалізується найпростіший підхід, згідно з яким злиття спільнот припиняється, коли модулярність мережі (2.5) перестає зростати. Такий підхід до вибору кількості спільнот не обґрунтований статистично.

Таким чином, методи виявлення спільнот у мережі, засновані на максимізації модулярності, хоч і є найбільш перспективними, але мають серйозні недоліки: наявність статистично необґрунтованих настроювальних параметрів (γ), необґрунтованість кількості спільнот (K), несхильність до виділення спільнот малого розміру.

2.2 Модифікований критерій модулярності

Припустимо, що мережа розділена на K спільнот C_u, $u = 1, \ldots, K$, які попарно не перетинаються. Позначимо кількість вузлів у спільноті з номером u як n_u. Числові значення K та n_u вважаються невідомими. Позначимо загальну кількість вузлів у мережі як $N = \sum_{u=1}^{K} n_u$. У випадковому графі будь-який вузол підключається до інших вузлів мережі з однаковою ймовірністю. Тоді ймовірність p_u зв'язування вузла $i \in C_u$ з довільним вузлом j із тієї ж самої спільноти C_u дорівнює

$$p_u = \Pr(j \in C_u \mid i \in C_u) = \frac{n_u - 1}{N - 1}. \tag{2.6}$$

Отже, очікувана вага зв'язків між вузлом i та іншими вузлами спільноти C_u становить $\mathrm{E}\left\{L_u^{in}(i)\right\} = p_u k_i$, де p_u визначається згідно з (2.6).

За цих умов цілком природною мірою модулярності вузла може служити різниця між фактичною вагою зв'язків поточного вузла i з всіма вузлами класу C_u та очікуваним значенням цієї ваги:

$$\Delta L_u^{in}(i) = L_u^{in}(i) - \mathrm{E}\left\{L_u^{in}(i)\right\} = L_u^{in}(i) - p_u k_i. \tag{2.7}$$

Варто зазначити, що подібна до (2.7) модель використовувалася в [10] для оцінки асортативності мережі. Крім того, проблема вимірювання асортативності [11] тісно пов'язана з проблемою виявлення спільнот.

Як і у випадку з традиційним визначенням, модулярність всієї спільноти дорівнює нормалізованій сумі модулярностей вузлів у цій спільноті:

$$\mu_u = \frac{1}{2m}\sum_{i\in C_u}\left(L_u^{in}(i) - \mathrm{E}\left\{L_u^{in}(i)\right\}\right) = \frac{1}{2m}\left(L_u^{in} - p_u L_u^{tot}\right). \qquad (2.8)$$

Легко бачити, що отриманий коефіцієнт модулярності (2.8) визначений для будь-яких непорожніх спільнот (тобто таких, що $n_u > 0$).

Щоб оцінити модулярність усієї мережі, підсумуємо локальні модулярності (2.8) за всіма спільнотами мережі:

$$\mu(G) = \sum_{u=1}^{K}\mu_u = \frac{1}{2m}\sum_{u=1}^{K}\left(L_u^{in} - p_u L_u^{tot}\right). \qquad (2.9)$$

Нижня межа модулярності мережі (2.9) відповідає випадку $\forall u:\ L_u^{in} = 0$ і не може бути меншою за –1, тоді як верхня межа, досяжна за ідеального розбиття ($\forall u:\ L_u^{in} = L_u^{tot}$), не перевищує +1. Отже, $\mu(G) \in (-1; 1)$.

Порівнюючи запропоноване визначення модулярності (2.8) – (2.9) із традиційним (2.4) – (2.5), можна зробити висновок, що єдина (але дуже суттєва) відмінність між ними зумовлена методом оцінки очікуваного значення ваги усіх посилань у спільноті u, тобто $\mathrm{E}\{L_u^{in}\}$. За традиційним підходом вона дорівнює $\gamma(L_u^{tot})^2 / (2m)$, тоді як у модифікованому методі ця оцінка дорівнює $p_u L_u^{tot}$. Враховуючи визначення p_u (2.6), можна зробити висновок, що ці оцінки збігаються за умови, що параметр роздільної здатності γ дорівнює

$$\gamma_u = p_u\frac{2m}{L_u^{tot}} = \frac{2m}{N-1} / \frac{L_u^{tot}}{n_u - 1} \approx \frac{\bar{k}}{\bar{k}_u}, \qquad (2.10)$$

де $\bar{k}, \bar{k}_u$ – середні ваги вузлів у всій мережі та у спільноті u відповідно.

Таким чином, модифікований метод розрахунку модулярності мережі, з одного боку, є статистично обґрунтованим через імовірнісну модель (2.6),

а з іншого боку, може розглядатися як різновид традиційного методу з використанням індивідуальних налаштувань параметра роздільної здатності (2.10).

В обох варіантах, традиційному (2.5) і модифікованому (2.9), модулярність мережі є сумою відхилень фактичної кількості зв'язків у спільнотах L_u^{in} від очікуваної. В обох випадках значення L_u^{in} можна розглядати як випадкові величини з очікуваними значеннями $\gamma(L_u^{tot})^2 / (2m)$ і $p_u L_u^{tot}$ відповідно та, що важливо, скінченими дисперсіями.

Звідси випливає, що дисперсія суми цих відхилень (тобто дисперсія значення модулярності всієї мережі) також є скінченою і зростає пропорційно кількості термів, тобто спільнот (K). Тому для забезпечення можливості порівнянь розбиттів на різну кількість спільнот у [4] було запропоновано нормалізувати модулярність мережі (2.9), поділивши її значення на $\sqrt{K}$:

$$\bar{\mu}(G) = \frac{1}{\sqrt{K}} \sum_{u=1}^{K} \mu_u = \frac{1}{2m \cdot \sqrt{K}} \sum_{u=1}^{K} \left(L_u^{in} - p_u L_u^{tot} \right). \qquad (2.11)$$

Запропонована нормалізація дозволяє порівнювати модулярність мережевих поділів для різної кількості спільнот.

2.3 Алгоритмічна реалізація виявлення мережевих спільнот на основі максимізації модифікованого критерію модулярності

Як відзначалося вище, виявлення спільнот шляхом максимізації модулярності (2.9) або (2.11) для великих значень N та K за допомогою повного перебору всіх можливих розбиттів (2.3) є неможливим через високу складність потрібних обчислень. Тому пропонується використати два варіанти жадібних алгоритмів, які можна назвати «помірно жадібним» та «дуже жадібним». Їхні псевдокоди показані на рисунок 2.1 і рисунок 2.2

відповідно.

В обох алгоритмах спочатку кожен вузол розглядається як окрема спільнота ($C_u = \{u\}, \quad u = 1, \dots, N$) і тоді початкові значення модулярностей (2.9) та (2.11) дорівнюють нулю. Обидва варіанти використовують однакову процедуру `findbestcomm4U(u,Comm)` для вибору спільноти C_v, до якої додається поточний вузол u. В цій процедурі перебираються всі вузли v, суміжні з u, та обирається така спільнота $C_v \ni v$, приєднання u до якої максимізує загальну модулярність (2.9) мережі. Поточне значення загальної модулярності мережі позначене як `gainU`.

```
algorithm medium_greedy(G):
iterate for u in Nodes:
     comm[u] = C[u] = set(u)
end_of_iterate
mod_prev = -inf
mod_curr = 0
while mod_curr > mod_prev:
     random_permutation(comm)
     iterate for u in comm:
          gainU, v = findbestcomm4U(u, C)
          Exclude(u, C[u])
          C[v] = union(C[v], C[u])
     end_of_iterate
     mod_prev = mod_curr
     mod_curr = modularity(C)
     comm = C
end_of_while
return C
```

Рисунок 2.1 – Псевдокод «помірно жадібного» алгоритму

У помірно-жадібному алгоритмі у циклі по первинних спільнотах (вузлах) $u = 1, \dots, k_{iter}$ безпосередньо після того як для поточної спільноти $u \in C_u$ буде знайдена (через виклик `findbestcomm4U(u,Comm)`) така спільнота C_v, приєднання u до якої максимізує `gainU`, виконуються операції приєднання u до C_v ($C_v = C_v \cup \{u\}$) та виключення u з C_u ($C_u = C_u \backslash \{u\}$). При цьому спільнота C_u може

стати порожньою, а може й ні. Після завершення циклу список спільнот оновлюється та виконується наступна ітерація зовнішнього циклу. Отже, цей варіант по суті збігається з відомим алгоритмом Louvain [6] із заміною традиційного критерію модулярності (2.5) на модифікований (2.9).

Згідно з дуже жадібним алгоритмом (рисунок 2.2) у циклі по спільнотах обирається така пара спільнот $C_u,\ C_v$, щоб їх злиття призвело до максимально можливого підвищення критерію (2.9) або (2.11). Після завершення циклу C_v оновлюється як $C_v = C_v \cup C_u$, а C_u виключається зі списку спільнот. Таким чином, після кожної ітерації кількість спільнот зменшується на одну. Цей процес (для зовнішнього циклу) триває до тих пір, поки є можливі злиття (тобто поки збільшується модулярність мережі).

```
algorithm very_greedy(G)
iterate for u in Nodes:
     comm[u] = C[u] = set(u)
end_of_iterate
mod_prev = -inf
mod_curr = 0
while mod_curr > mod_prev:
     ubest = u[1]
     gainbest = -inf
     iterate for u in comm:
          gainU, v = findbestcomm4U(u, C)
          if gainU > gainbest:
               gainbest = gainU
               ubest = u
               vbest = v
     end_of_iterate
     Exclude(ubest, C[ubest])
     C[vbest] = union(C[vbest], C[ubest])
     mod_prev = mod_curr
     mod_curr = modularity(C)
     comm = C
end_of_while
return C
```

Рисунок 2.2 – Псевдокод «дуже жадібного» алгоритму

Як видно з псевдокодів (рисунок 2.1, рисунок 2.2), складнішою та найглибше вкладеною частиною обох алгоритмів є виклик функції

`findbestcomm4U`. Ця функція виконує ітерацію з перебору спільнот, які мають спільні ребра з вузлами поточної спільноти, тому її обчислювальна складність дорівнює $O(k)$, де $k \leq d_{max},\ k \leq k_{iter}$, k_{iter} – поточна кількість спільнот d_{max} – максимальна ступінь первинної спільноти (якщо розглядати її як вузол). Таким чином, складність помірно жадібного алгоритму дорівнює

$$T_1 = O(N \cdot k), \quad k = \max\{d_{\max}, K\}. \tag{2.12}$$

Обчислювальна складність однієї ітерації дуже жадібного алгоритму дорівнює $O(nk)$, де n – поточна кількість спільнот, $n = N, N-1, \ldots, K$. Таким чином, складність дуже жадібного алгоритму дорівнює

$$T_2 = O(N^2 k), \quad k = \max\{d_{\max}, K\}. \tag{2.13}$$

Очевидно, що помірно-жадібний алгоритм значно швидший за дуже жадібний, але його результат (як значення модулярності, так і конфігурація спільнот) залежить від порядку проходження спільнот у циклі $u = 1, \ldots, k_{iter}$. З іншого боку, дуже жадібний алгоритм, є незалежним від випадкового порядку спільнот, а також дозволяє відстежувати зміну модулярності мережі в покроковому режимі.

3 ЧИСЕЛЬНІ ЕКСПЕРИМЕНТИ НА ТЕСТОВИХ МЕРЕЖАХ

Програмна реалізація алгоритму розбиття мереж на спільноти виконувалась на мові Python із використанням бібліотек, таких як NetworkX для роботи з графами та SciPy для виконання ієрархічної кластеризації. Для візуалізації розбиттів використовувався пакету Gephi [12], який забезпечує інтерактивний інтерфейс для аналізу та представлення структурних властивостей мереж, таких як модульність, центральність вузлів та візуальний розподіл спільнот.

Для перевірки працездатності та ефективності досліджуваних методів та алгоритмів використовувались широко відомі тестові датасети реальних мереж.

Першим з них було обрано датасет карате-клуба, дослідженого Закарі [13]. Це неорієнтована зважена мережа, яка містить $N = 34$ вузлів, які мають загальну вагу $2m = 462$. Цей набір даних з відомим розбиттям на дві групи широко використовується при вивченні спільнотної структури мереж. Вузли груп позначені як «Mr. Hi» і «Officer», а кожна група складається з 17 вузлів (рисунок 3.1).

Рисунок 3.1 – Мережа karate_club

Максимізація неусередненого критерію модулярності (2.9) за допомогою дуже жадібного алгоритму призводить до поділу цієї мережі на три великі спільноти (з 14, 10 і 5 вузлів) і 5 одиничних вузлів (‘J’, ‘S’, ‘c’, ‘L’, та ‘R’). Досягнуте значення критерію дорівнює $\mu(G) = 0.4773$. Дендрограма розбиття представлена на рисунок 3.2. Для наочності візуалізації дендрограми цифрові позначки вузлів (1,...,34) замінено літерами *A,...,Z,a,...,h*. Значення осі висот на дендрограмі відповідають значенням критерію модулярності (2.9).

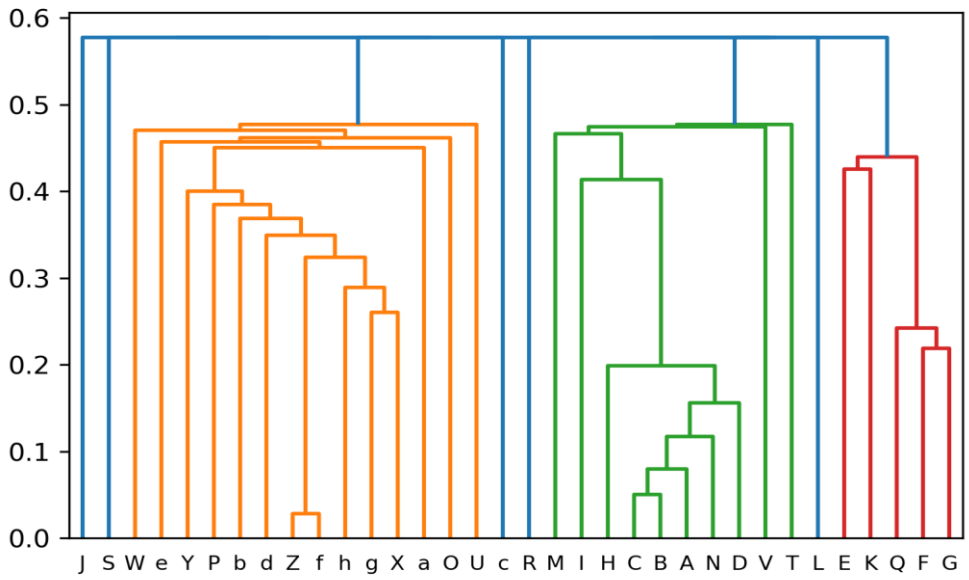

Рисунок 3.2 – Дендрограмма розбиття мережі karate_club за критерієм неусередненої модулярності (2.9) за дуже жадібним алгоритмом розбиття

Важливо відзначити, що залежність неусередненої модулярності (2.9) мережі від кількості спільнот (рисунок 3.3) є плоскою, тобто її максимум виражений слабо. Таким чином, отримане значення для кількості спільнот ($K = 8$) значною мірою зумовлено властивостями цього конкретного алгоритму. При використанні іншого алгоритму поділу можна очікувати значення K в діапазоні від 3 до 12.

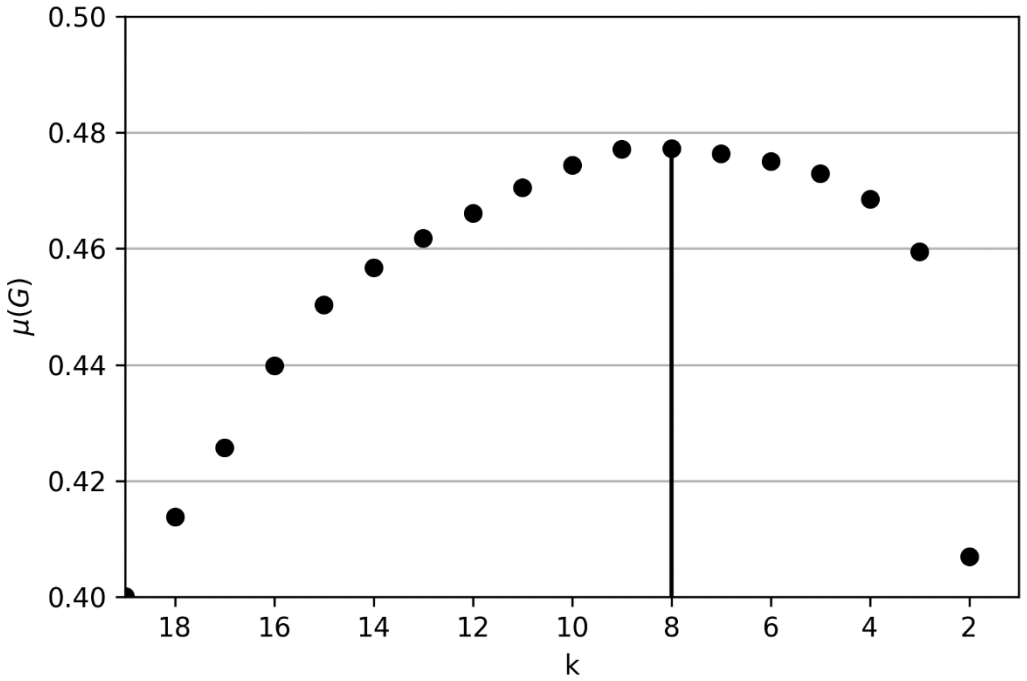

Рисунок 3.3 – Залежність неусередненої модулярності мережі karate_club від кількості виявлених спільнот за дуже жадібним алгоритмом розбиття

Відповідно до чисельного експерименту, максимізація того ж самого неусередненого критерію модулярності (2.9) за допомогою помірно-жадібного алгоритму призводить до поділу цієї мережі на дві спільноти – з 16 та 18 вузлів. У порівнянні з розбиттям за дуже жадібним алгоритмом (рисунок 3.2), вузли «J», «S», «c» об'єднані в спільноту з 14 вузлів («Officer»), вузли «L» та «R» приєдналися до спільноти з 10 та 5 вузлами («Mr.Hi»), але вузол «I» перейшов із спільноти «Officer» до «Mr. Hi», що є неправильно. Відомо, що цей вузол «I» (№9) є проблематичним для багатьох процедур класифікації, перевірених на цьому наборі даних. Таким чином, помірно-жадібний алгоритм створює розбиття, близьке до ідеального, досягаючи високого значення критерію $\mu(G) = 0.4185$ і є істотно швидшим, ніж дуже жадібний варіант запропонованого алгоритму.

Максимізація нормованої на $\sqrt{K}$ модулярності (2.11) за дуже жадібним алгоритмом призводить до поділу цієї мережі на дві спільноти (по 17 вузлів кожна). Варто відзначити, що цей поділ повністю відповідає вихідним міткам вузлів («Mr. Hi» та «Officer», рисунок 3.1), тобто отримане

розбиття, яке наведено на рисунок 3.4, є ідеальним. Досягнуте значення критерію (2.11) становить $\bar{\mu}(G) = 0.2877$, що дорівнює неусередненому значенню $\mu(G) = 0.4069$, тобто є дещо менше ніж для неідеального розподілу 16/18.

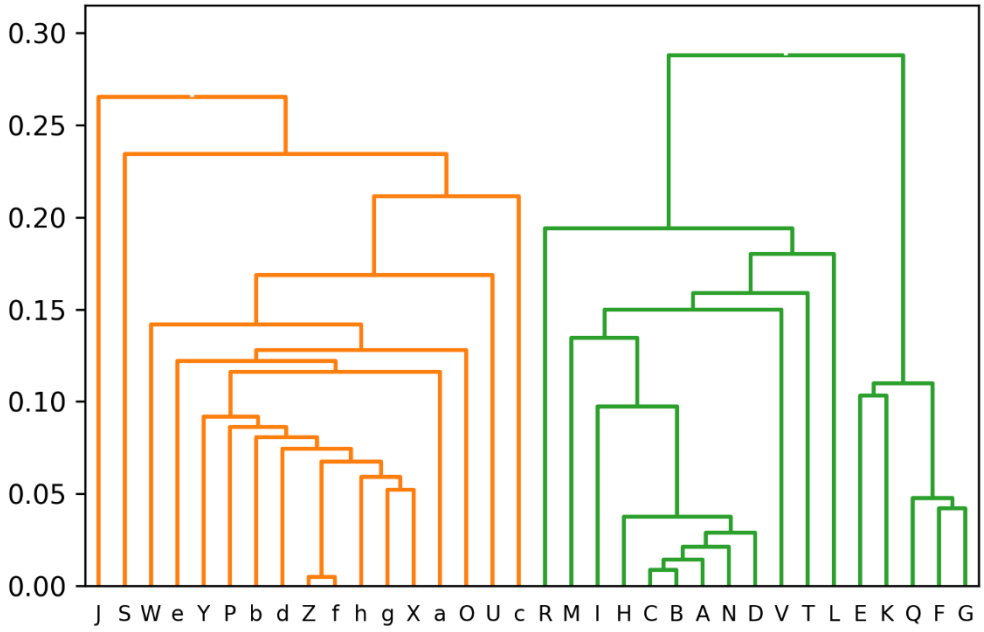

Рисунок 3.4 – Дендрограмма розбиття мережі karate_club за критерієм усередненої модулярності (2.11) з використанням дуже жадібного алгоритму розбиття

Другим тестовим набором даним, який використовувався для дослідження пропонованого методу розділення мереж, є популярний набір Newman's Polbooks [14]. Цей набір часто використовується для демонстрації алгоритмів розбиття мереж на спільноти. Вузлами цього набору є книги про політику США, які продає онлайн-продавець книг Amazon.com. Вузли (книги) з'єднані ребром, якщо ці книги були замовлені одним й тим же самим покупцем. Ребра між вузлами позначають той факт, що ці книги часто купувалися разом. Мережа містить 105 вузлів та 441 ребр, що з'єднують ці вузли. Вузли позначені літерами «c», «l», «n» (49, 43 і 13 вузлів відповідно), щоб вказати, чи є вони консервативними, ліберальними чи

нейтральними за дотриманням відповідних політичних поглядів (рисунок 3.5). Мережа має природний поділ на кластери, які відповідають політичним ідеологіям. Марк Ньюман (Mark Newman) присвоїв ці мітки окремо на основі читання описів та оглядів книг, розміщених на Amazon. Мережа містить 105 вузлів та 441 ребро, а максимальний ступінь вузла становить 25.

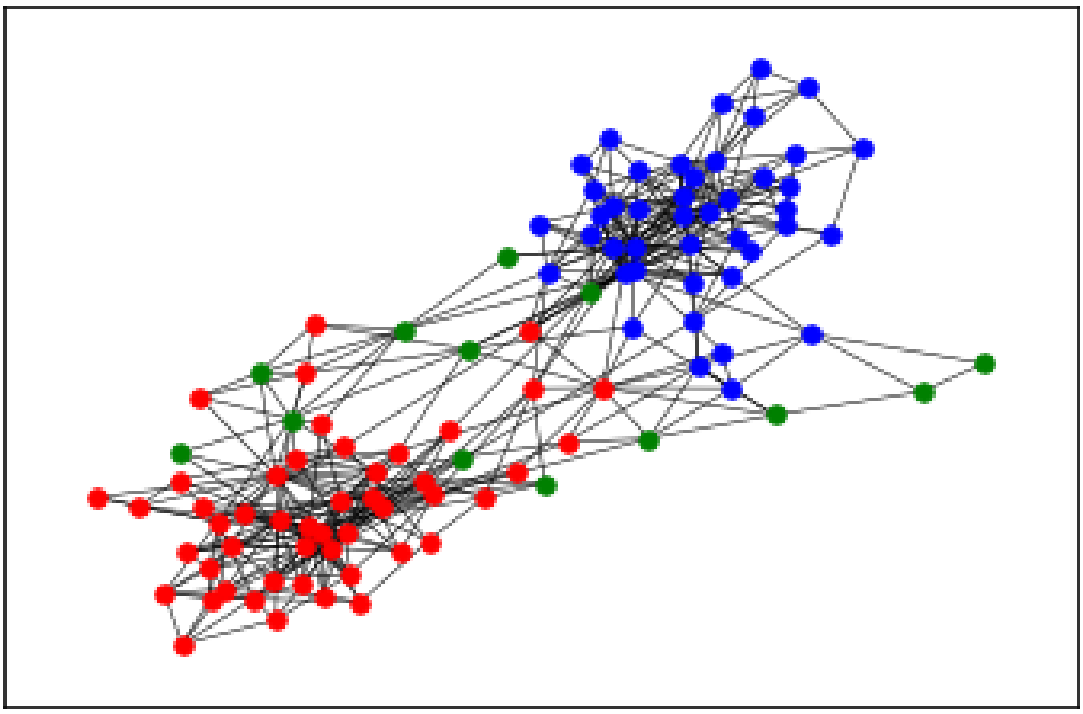

Рисунок 3.5 – Мережа polbooks; червоні вузли – консервативні книги («c»), сині – ліберальні («l»), зелені – нейтральні («n»)

Максимізація неусередненого критерію модулярності (2.9) дуже жадібним алгоритмом призводить до поділу цієї мережі на шість спільнот (з 4, 38, 37, 10, 4 і 12 вузлів). Досягнуте значення критерію $\mu(G) = 0.5628$. Дендрограма розбиття мережі наведена на рисунок 3.6, а залежність використовуваного критерію від кількості виявлених спільнот (рисунок 3.7) аналогічна відповідній залежності для клубу карате (рисунок 3.3).

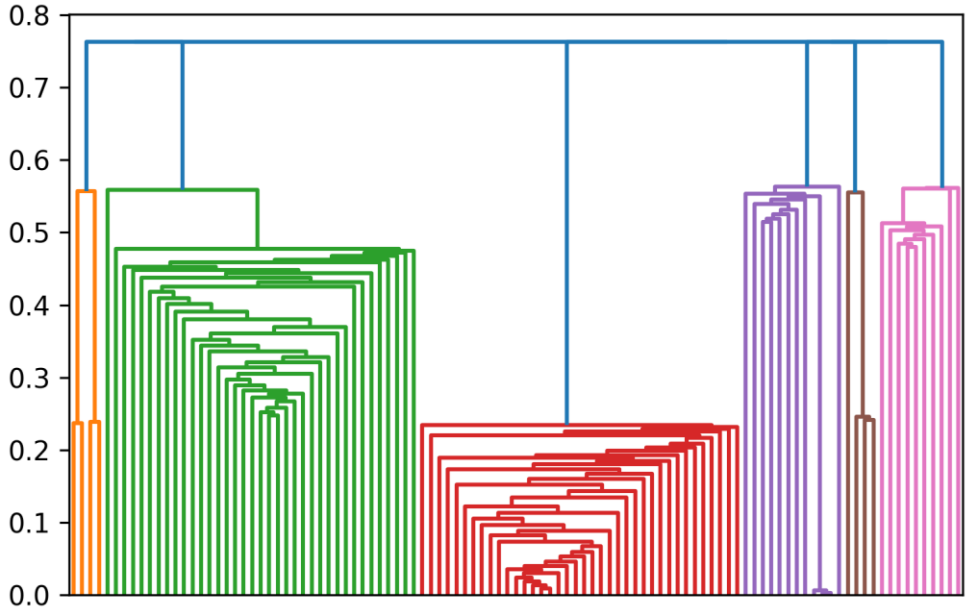

Рисунок 3.6 – Дендрограма розбиття мережі polbooks за критерієм неусередненої модулярності (2.9) з використанням дуже жадібного алгоритму розбиття

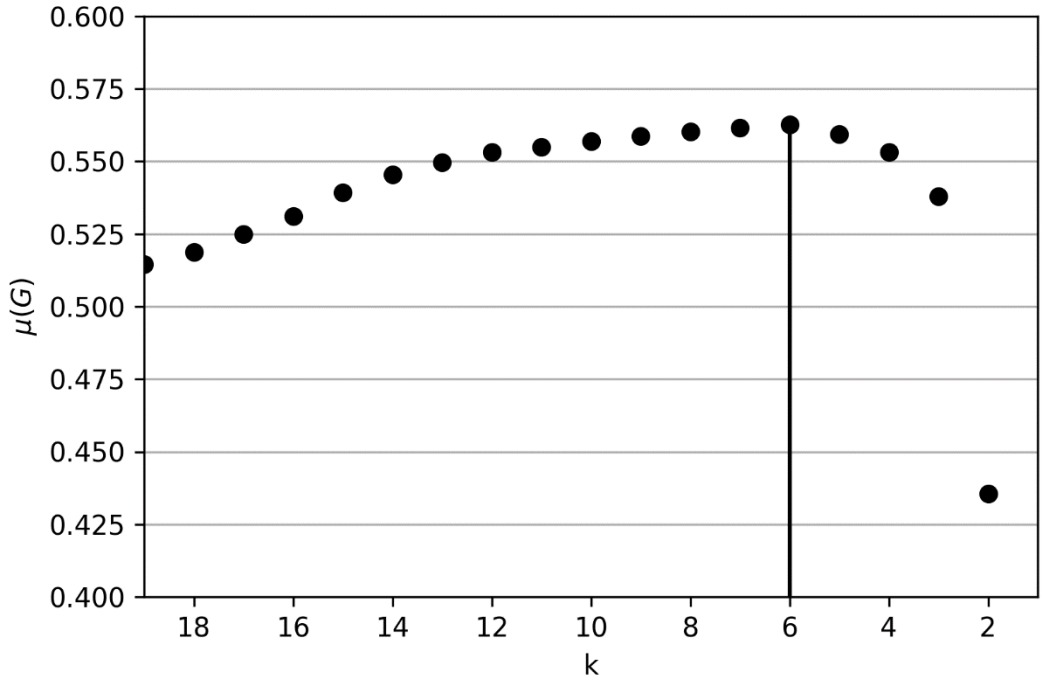

Рисунок 3.7 – Залежність неусередненої модулярності мережі polbooks від кількості виявлених спільнот за дуже жадібним алгоритмом розбиття

У той же час розбиття за нормалізованим критерієм модулярності (2.11) призводить до поділу мережі на три спільноти (рисунок 3.8) з 38, 41 і 26 вузлів відповідно.

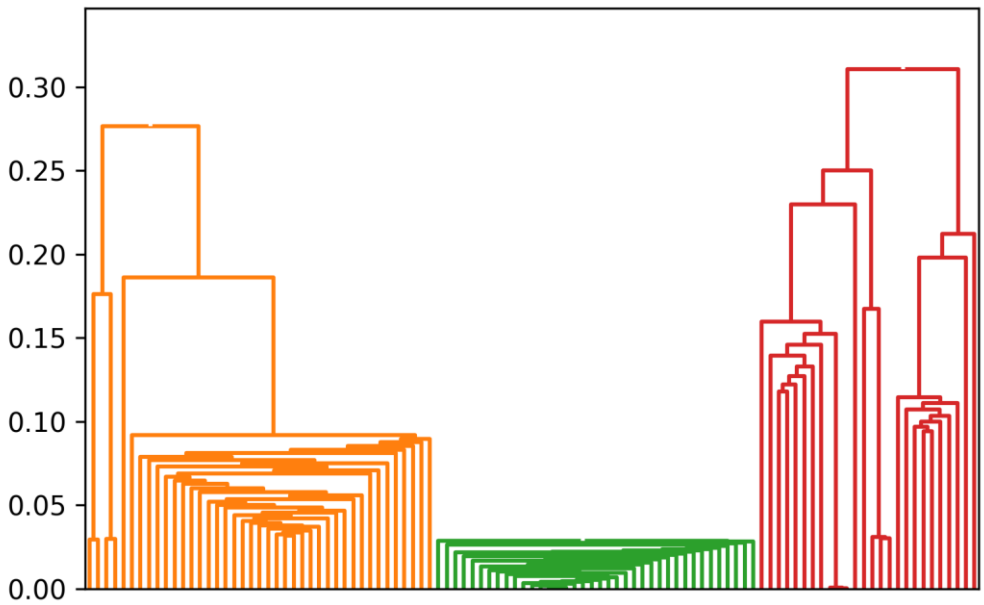

Рисунок 3.8 – Дендрограма розбиття мережі polbooks за критерієм нормалізованої модулярності (2.11) з використанням дуже жадібного алгоритму розбиття

Отримане розбиття не можна назвати точним: половина вузлів, віднесених до спільноти «n», насправді належать до класів «c» та «l». При цьому кількість спільнот ($K = 3$) відповідає дійсному значенню. Варто зазначити, що первісна («справжня») розмітка вузлів є суб'єктивною. Отже, якість розбиття запропонованим методом мережі polbooks можна вважати помірною.

ВИСНОВКИ

Предметом досліджень кваліфікаційної роботи є задача розбиття мережі на спільноти. Було проведено дослідження варіацій та особливостей постановки задачі виявлення мережевих спільнот, проаналізовано основні підходи щодо їх розв’язання.

Визначено, що однією з найважливіших вимог щодо методів виявлення спільнот є їхня швидкодія. Крім того, важливим чинником є інтерпретованість результатів розбиття. Через ці причини дослідження були зосереджені на класі критерій-орієнтованих методів, тобто на виявленні спільнот на основі максимізації критерію модулярності.

Було проаналізовано визначення модулярності, методів її розрахунку та алгоритмів її максимізації. Зазначено, що традиційний критерій модулярності має схильність до поглинання малих спільнот. Тому було взято до уваги модифікований критерій модулярності, який явним чином залежить від кількості вузлів у кожній спільноті. Ця властивість підвищує схильність до виявлення невеликих спільнот. Проведено нормалізацію модифікованої функції модулярності щодо кількості спільнот, що робить її більш актуальною у випадку, коли кількість спільнот апріорі невідома.

Розроблено два алгоритми поділу вузлів мережі на спільноти на базі модифікованої модулярності. «Дуже жадібний» алгоритм має квадратичну складність відносно кількості вузлів. «Помірно-жадібний» алгоритм виявлення спільнот має лінійну складність, тобто є значно швидше, тому саме він є основним варіантом алгоритмічної реалізації досліджуваного методу розбиття мереж на спільноти.

Досліджувані алгоритми були реалізовані програмно та були протестовані на відомих мережевих датасетах. Результати експерименту в цілому підтверджують відповідність властивостей досліджуваного методу розбиття мереж вихідним вимогам.

Результати роботи було представлено на конференції ICT-2024 [15].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. 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: 02.01.2025).

2. 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: 02.01.2025).

3. Cultural scene detection using reverse Louvain optimization / M. Hamdaqa et al. *Science of Computer Programming*. 2014. Vol. 95. P. 44–72. URL: https://doi.org/10.1016/j.scico.2014.01.006 (date of access: 02.01.2025).

4. Network community detection using modified modularity criterion / V. Shergin et al. Eejet. 2024. Vol. 6, no. 4 (132). P. 6-13. URL: https://doi.org/10.15587/1729-4061.2024.318452 (date of access: 03.01.2025).

5. 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: 02.01.2025).

6. Fast unfolding of communities in large networks / V. D. Blondel et al. *Journal of statistical mechanics: theory and experiment*. 2008. Vol. 2008, no. 10. P. P10008. URL: https://doi.org/10.1088/1742-5468/2008/10/p10008 (date of access: 02.01.2025).

7. 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: 02.01.2025).

8. Parkkinen, Juuso, et al. A block model suitable for sparse graphs. *Proceedings of the 7th International Workshop on Mining and Learning with Graphs* (MLG 2009), Leuven. 2009. Vol. 5.

9. Assortativity of an elastic network with implicit use of information about

nodes degree / V. Shergin et al. *International scientific symposium "intelligent solutions"*. 2021. URL: https://api.semanticscholar.org/CorpusID:245386355 (date of access: 02.01.2025)

10. Noncorrelational assortativity coefficient for networks with nominative attributes / V. L. Shergin et al. 10-th IEEE international scientific-practical conference problems of infocommunications : Science and Technology, Kharkiv.

11. 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: 02.01.2025).

13. 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: 02.01.2025).

14. Newman M. Network data. URL: https://public.websites.umich.edu /~mejn/ netdata/ (date of access: 02.01.2025).

15. Мірошниченко Т., Мусієнко М., Пономаренко А. Виявлення мережевих спільнот на основі критерію модулярності. 13-та Міжнародна науково-технічна конференція : Інформ. системи та технології ІСТ-2024, м. Харків, 26–28 листоп. 2024 р.