НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»
Навчально-науковий Фізико-технічний інститут
Кафедра математичного моделювання та аналізу даних

До захисту допущено:
Виконувач обов'язків завідувача кафедри
_____ Іван ТЕРЕЩЕНКО
«___»_____2023 р.

Дипломна робота
на здобуття ступеня бакалавра
за освітньо-професійною програмою «Математичні методи моделювання, розпізнавання образів та безпеки даних»
спеціальності 113 «Прикладна математика»
на тему: «Дослідження українського сегменту Вікіпедії засобами аналізу складних мережевих структур»

Виконав:
студент IV курсу, групи ФІ-91
Дербаль Богдан Васильович _____
Керівник:
к.ф.-м.н. Орєхов Олександр Арсенійович _____
Рецензент:
_____ _____

Засвідчую, що у цій дипломній роботі немає
запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ – 2023 року

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Навчально-науковий Фізико-технічний інститут

Кафедра математичного моделювання та аналізу даних

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 113 «Прикладна математика»

Освітня програма «Математичні методи моделювання, розпізнавання образів та безпеки даних»

ЗАТВЕРДЖУЮ

Виконувач обов'язків завідувача кафедри

_____ Іван ТЕРЕЩЕНКО

«__» _____ 2023 р.

ЗАВДАННЯ

на дипломну роботу студенту

Дербаль Богдан Васильович

1. Тема роботи «Дослідження українського сегменту Вікіпедії засобами аналізу складних мережевих структур», керівник роботи к.ф.-м.н. Орєхов Олександр Арсенійович, затверджені наказом по університету від «__» ____ 2023р. №_____

2. Термін подання студентом роботи _____.

3. Вхідні дані до роботи – розвинені методи аналізу складних мережевих структур.

4. Зміст роботи – 1. Опис предметної області; 2. Теоретичні основи роботи; 3. Реалізація поставленої задачі і аналіз результатів.

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

6. Консультанти розділів роботи

7. Дата видачі завдання 16.02.2023

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Огляд опублікованих джерел за тематикою дослідження	Січень-лютий 2023 р.	Виконано
2	Підготовка першого розділу	Лютий-березень 2023 р.	Виконано
3	Підготовка другого розділу	Березень 2023 р.	Виконано
4	Підготовка даних до роботи	Квітень 2023 р.	Виконано
5	Аналіз даних	Квітень-травень 2023 р.	Виконано
6	Підготовка третього розділу	Травень 2023 р.	Виконано
7	Оформлення дипломної роботи	Травень 2023 р.	Виконано
8	Підготовка презентації доповіді	Травень 2023 р.	Виконано

Студент _____ Богдан ДЕРБАЛЬ

Керівник _____ Олександр ОРЄХОВ

РЕФЕРАТ

Дипломна робота присвячена аналізу складних мережевих структур українського сегменту Вікіпедії. Робота складається з 47 сторінок, включає 16 ілюстрації та посилається на 8 джерел літератури.

Об'єктом дослідження є український сегмент Вікіпедії, який є складною мережевою структурою, що включає безліч статей, гіперпосилань та взаємозв'язків між ними.

Мета дослідження полягає в аналізі та розумінні структури, взаємозв'язків та характеристик українського сегменту Вікіпедії. Метою є отримання інсайтів щодо взаємодії та динаміки інформації, виявлення потенційних дірок у покритті українського сегменту Вікіпедії.

Для досягнення мети було використано:

1. Огляд існуючих досліджень що до аналізу мережевих структур.
2. Використання API Вікіпедії для збору даних необхідних для подальшого дослідження.
3. Використання програмних засобів для роботи з графовими структурами та аналізу мереж, таких як Gephi та бібліотека Python NetworkX.
4. Використовуючи зібрані метрики було виконано аналіз мережі.

Ключові слова: АНАЛІЗ МЕРЕЖЕВИХ СТРУКТУР, УКРАЇНСЬКИЙ СЕГМЕНТ ВІКІПЕДІЇ, ГРАФОВА СТРУКТУРА, КЛАСИФІКАЦІЯ МЕРЕЖ, RANDOM NETWORK, SCALE-FREE NETWORK, SMALL WORLD NETWORK.

ABSTRACT

The diploma thesis is dedicated to the analysis of complex network structures in the Ukrainian segment of Wikipedia. The work consists of 47 pages, includes 16 illustrations, and references 8 sources.

The object of the research is the Ukrainian segment of Wikipedia, which is a complex network structure comprising numerous articles, hyperlinks, and interconnections between them.

The purpose of the research is to analyze and understand the structure, interconnections, and characteristics of the Ukrainian segment of Wikipedia. The goal is to gain insights into information interaction and dynamics, identify potential gaps in coverage, and forecast the development of the Ukrainian segment of Wikipedia.

To achieve this goal, the following steps were taken:

1. Review of existing research on network structure analysis.
2. Utilization of the Wikipedia API to gather the necessary data for further analysis.
3. Employment of software tools for working with graph structures and network analysis, such as Gephi and the Python library NetworkX.
4. Classification of networks using the collected metrics.

Keywords: NETWORK STRUCTURE ANALYSIS, UKRAINIAN SEGMENT OF WIKIPEDIA, GRAPH STRUCTURE, NETWORK CLASSIFICATION, RANDOM NETWORK, SMALL WORLD NETWORK.

ЗМІСТ

РЕФЕРАТ **4**
ABSTRACT **4**
ЗМІСТ **5**
ВСТУП **7**
РОЗДІЛ 1. ОГЛЯД ПРЕДМЕТНОЇ ОБЛАСТІ **9**
1.1 Постановка задачі 9
1.2 Опис даних 10
1.3 Актуальність роботи 10
РОЗДІЛ 2. ТЕОРЕТИЧНІ ОСНОВИ РОБОТИ **11**
2.1 Загальні поняття теорії графів 11
2.1.1 Мережі та графи 11
2.1.2 Орієнтовані та неорієнтовані графи 13
2.1.3 Ступінь, середній ступінь та розподіл ступенів 13
2.1.4 Матриця суміжності 15
2.1.5 Зважені графи 17
2.1.6 Шлях, відстань 17
2.1.7 Зв'язність 18
2.1.8 Коефіцієнт кластеризації 19
2.2 Випадкові мережі 19
2.2.1 Модель випадкової мережі 19
2.2.2 Варіативність випадкової мережі 20
2.2.3 Розподіл ступенів випадкової мережі 22
2.2.4 Розвиток випадкової мережі 23
2.2.5 Властивість “світ малий” 26
2.3 Властивість безмасштабності 28
2.3.1 Степеневий закон та безмасштабні мережі 28
2.3.2 Хаби 31
2.3.3 Властивість малого світу для безмасштабних мереж 32
2.3.4 Універсальність безмасштабних мереж 32
РОЗДІЛ 3. РЕАЛІЗАЦІЯ ПОСТАВЛЕНОЇ ЗАДАЧІ І АНАЛІЗ РЕЗУЛЬТАТІВ **35**
3.1 Опис інструментів для реалізації задачі 35
3.2 Збір даних та формування графу 35
3.3 Аналіз та візуалізація мережі 36
3.3.1 Визначення розподілу вершин 36

3.3.2 Наявність хабів..... 39
3.3.3 Властивість “світ малий”..... 40
3.3.4 Аналіз результатів..... 41
ВИСНОВКИ..... 42
СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ..... 43
ДОДАТКИ..... 44
Додаток А Код програми для побудови мережі..... 44
Додаток В Код для побудови графіків розподілу вершин в логарифмічному представленні..... 47

ВСТУП

Українська Вікіпедія є одним з найбільших і найпопулярніших онлайн-джерел інформації, яке має значний вплив на формування знань та розуміння світу. Ця енциклопедія, яка постійно розширюється та оновлюється користувачами з різних куточків України та світу, утворює складну мережеву структуру, яка відображає взаємозв'язки між статтями, авторами та їхнім контентом. Проте, через неймовірно велику кількість статей та зв'язків між ними, досить складно уявити, що насправді являє собою ця структура. Саме тому ця робота буде присвячена її дослідженню.

Мета дослідження. Метою даного дипломного проекту є проведення детального аналізу українського сегменту Вікіпедії з використанням методів аналізу складних мережевих структур. Головною метою є виявлення ключових факторів, що впливають на структуру інформації в українському сегменті Вікіпедії. Результати цього дослідження можуть бути використані для покращення якості та доступності інформації в Вікіпедії та надати уявлення щодо подальшого розвитку цієї мережі.

Об'єкт дослідження. Об'єктом дослідження є український сегмент Вікіпедії - онлайн-енциклопедія, яка працює на основі принципу колективного редагування та забезпечує доступ до інформації українською мовою.

Предмет дослідження. Предметом дослідження є аналіз складних мережевих структур в українському сегменті Вікіпедії.

Новизна дослідження. Основною новизною даного дослідження є застосування методів аналізу складних мережевих структур для вивчення структури Вікіпедії. Цей підхід дозволяє отримати нові інсайти та розуміння процесів, що відбуваються в мережі Вікіпедії, з'ясувати ключові чинники, що впливають на розвиток та якість інформації. Дослідження специфічно фокусується на українській версії Вікіпедії, враховуючи особливості української мови, культурного контексту та користувацьких звичаїв. Це дозволяє отримати унікальні висновки, специфічно адаптовані до українського сегменту. Результати дослідження можуть мати значний вплив на подальший розвиток Вікіпедії та поліпшення її функціональності.

РОЗДІЛ 1. ОГЛЯД ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Постановка задачі

Постановка задачі включатиме наступні етапи:

- Збір та підготовка даних: за допомогою API Вікіпедії будуть зібрані дані про український сегмент, такі як статті, першоджерела та їхні взаємозв'язки. Дані будуть оброблені та підготовлені для подальшого аналізу.

- Побудова мережевої структури: з використанням бібліотеки networkx будуть побудовані мережі, де статті представлятимуть вузли, а зв'язки між ними відображатимуться ребрами.

- Візуалізація мережі: з використанням програмного забезпечення Gephi буде проведена візуалізація отриманих мережевих структур. Це дозволить наглядно представити зв'язки між статтями, виявити структурні особливості та зрозуміти глобальну організацію мережі.

- Розподіл вершин та класифікація мереж: буде побудовано графік розподілу вершин за їх ступенями для визначення ролі та значущості вузлів у мережі. Далі проведемо класифікацію мереж за типами графів (наприклад, безмасштабна мережа, випадкова мережа тощо) для отримання додаткової інформації про структуру мережі Вікіпедії.

Отримані результати дослідження допоможуть виявити особливості та характеристики українського сегменту Вікіпедії, розкрити взаємодії між статтями, а також зрозуміти мережеву структуру та розвиток цієї онлайн-енциклопедії.

1.2 Опис даних

Дані є основою для проведення аналізу та отримання результатів. Джерело даних: Український сегмент Вікіпедії. Для збору даних було використано API Вікіпедії, яке надає доступ до різних аспектів платформи, включаючи статті, першоджерела та їхні взаємозв'язки.

Статті: Зібрані дані містять інформацію про українські статті відповідно до їхніх заголовків, категорій та метаданих.

Мережева структура: Дані про зв'язки між статтями утворюють мережеву структуру українського сегменту Вікіпедії. Ця структура включає граф, де вершини представляють статті, а ребра відображають зв'язки між ними.

Отримані дані є основою для подальшого аналізу та дослідження складних мережевих структур. Використання цих даних дозволить отримати нові уявлення про взаємозв'язки, властивості та динаміку цієї мережі, що сприятиме подальшому розвитку та покращенню української Вікіпедії.

1.3 Актуальність роботи

Вікіпедія забезпечує доступ до широкого спектру інформації для україномовних користувачів. Її важливість полягає в поширенні знань, збереженні культурної спадщини та сприянні навчанню. Проте, український сегмент Вікіпедії представляє собою складну та малодосліджену структуру. Тому дослідження українського сегменту Вікіпедії є актуальною задачею, що дозволяє краще розуміти цю мережу та сприяти її подальшому розвитку.

РОЗДІЛ 2. ТЕОРЕТИЧНІ ОСНОВИ РОБОТИ

2.1 Загальні поняття теорії графів

Теорія графів є потужним інструментом для вивчення структури та взаємозв'язків в складних системах. Вона розглядає об'єкти, які складаються з вузлів (або вершин) і зв'язків між ними. Ця теорія дозволяє абстрагуватися від конкретного вмісту системи і зосередитися на її структурних властивостях.

Теорія графів надає нам інструменти для аналізу різноманітних проблем, таких як маршрутизація в мережах, соціальні взаємозв'язки, логістика, та багато інших. Вона допомагає розкрити складні залежності між об'єктами та виявити основні властивості системи, які впливають на її функціонування. Тож давайте ознайомимось з базовими поняттями.

2.1.1 Мережі та графи

Мережа - це сукупність компонентів системи, які називають **вузлами** або **вершинами**, та прямих взаємодій між ними, які називають **зв'язками** або **ребрами**.

Граф являє собою графічне представлення мережі. Рисунок 2.1 чудово демонструє приклади мереж та два базові поняття:

Кількість вузлів (N), представляє собою кількість компонентів у системі. Величину N часто називають розміром мережі. Щоб відрізнити вузли, їх позначають індексами i = 1, 2, ..., N.

Кількість ребер (L) представляє загальну кількість взаємодій між вузлами. Зазвичай потреби в окремих позначках для ребер немає, оскільки їх можна легко ідентифікувати за допомогою вузлів, які вони з'єднують. Наприклад, ребро (2, 4) з'єднує вузли 2 та 4.

Рисунок 2.1 показує підмножину (a) Інтернету, де маршрутизатори підключені один до одного; (b) мережу акторів Голлівуду, де два актори з'єднані, якщо вони зіграли в одному фільмі; (c) мережу взаємодії білків, де два білки пов'язані, якщо є експериментальні докази того, що вони можуть зв'язуватися один з одним у клітині.

Хоча природа вузлів та ребер відрізняється, ці мережі мають одне й те ж графічне представлення, складаючись з N = 4 вузлів та L = 4 зв'язків, показаних на (d).

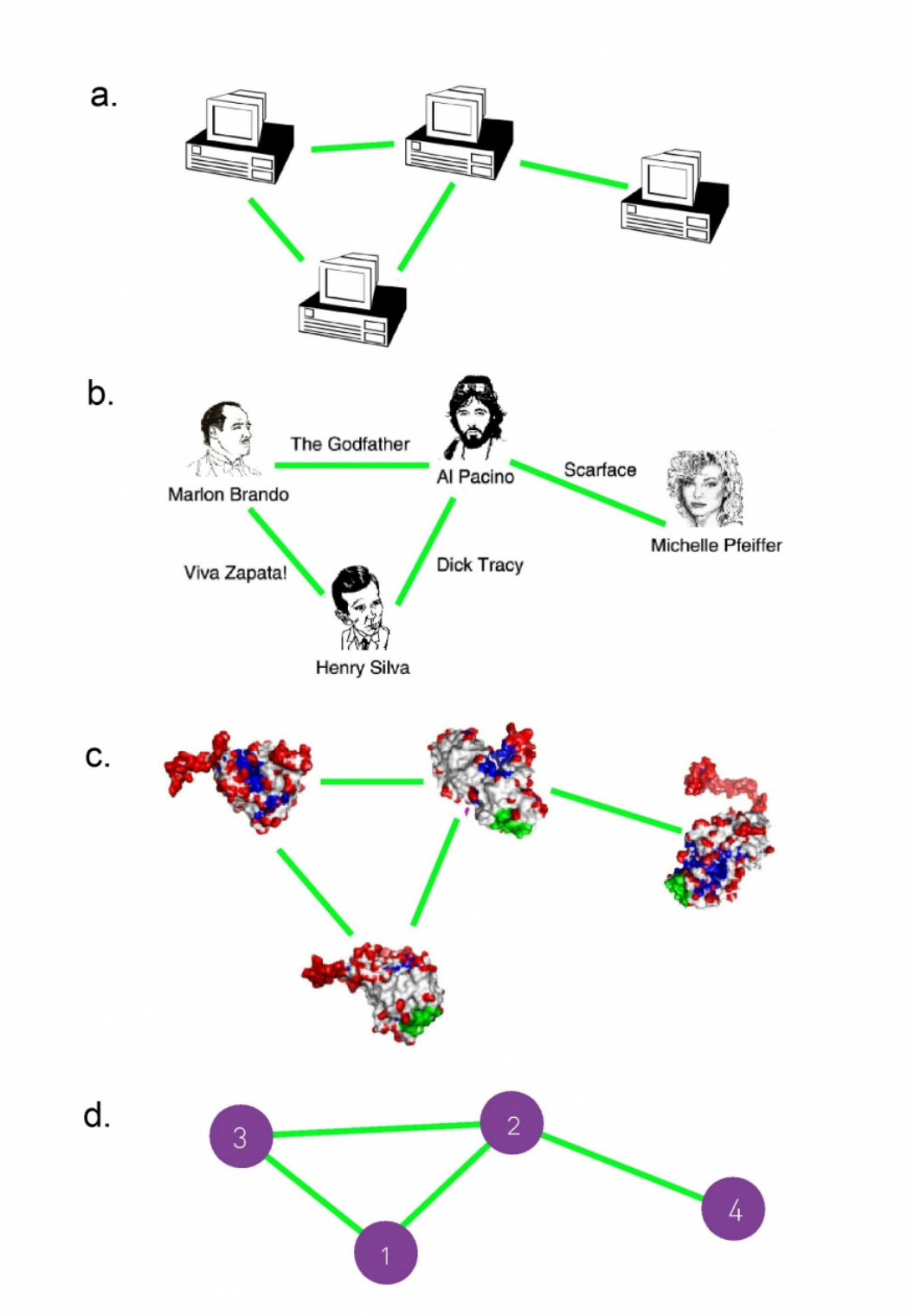

Рис. 2.1 Приклади мереж.

2.1.2 Орієнтовані та неорієнтовані графи

В теорії графів ребра мережі можуть бути **орієнтованими** або **неорієнтованими**, в залежності від того, чи вказується напрямок зв'язку між вузлами.

Граф називається **орієнтованим**, якщо всі його ребра орієнтовані. Це вказує, що комунікація або взаємодія між вузлами відбувається в одному напрямку. Наприклад, якщо граф представляє мережу залізниць, орієнтоване ребро може вказувати напрямок руху потягу від однієї станції до іншої.

Граф називається **неорієнтованим**, якщо всі його ребра неорієнтовані. Це означає, що комунікація або взаємодія між вузлами відбувається в обидва напрямки. Наприклад, якщо граф представляє дорожню мережу, неорієнтоване ребро може позначати двосторонній рух між двома перехрестями.

2.1.3 Ступінь, середній ступінь та розподіл ступенів

Однією з ключових характеристик кожного вузла є його **ступінь**, який представляє кількість зв'язків, які він має з іншими вузлами. Ступінь і-того вузла позначається k_i. На рисунку 2.1 маємо $k_1 = 2$, $k_2 = 3$. У неорієнтованій мережі загальна кількість ребер, L, може бути виражена як сума ступенів вузлів:

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i, \tag{2.1}$$

Важливою властивістю мережі є її **середня ступінь**, яка для неорієнтованої мережі обчислюється як:

$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}, \tag{2.2}$$

У направлених мережах розрізняють вхідний ступінь, k_i^{in}, що представляє кількість зв'язків, які вказують на вузол i, та вихідний ступінь, k_i^{out}, що представляє кількість

зв'язків, які виходять з вузла і в інші вузли. Нарешті, загальний ступінь вузла, k_i, задається формулою:

$$k_i = k_i^{in} + k_i^{out} , \tag{2.3}$$

Загальну кількість посилань у спрямованій мережі також можна отримати як:

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} , \tag{2.4}$$

Ступінь орієнтованої мережі можна виразити:

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \frac{L}{N} , \tag{2.5}$$

Розподіл ступенів p_k виражає ймовірність того, що випадково обраний вузол у мережі має ступінь k. Оскільки p_k це ймовірність, то ця величина має бути нормалізована, тобто:

$$\sum_{k=1}^{\infty} p_k = 1 , \tag{2.6}$$

Для мережі з N вузлами розподіл ступенів, який візуалізований в гістограмах (рисунок 2.2) обчислено за формулою:

$$p_k = \frac{N_k}{N} , \tag{2.7}$$

Де N_k кількість вершин з ступенем k, а N - загальна кількість вершин

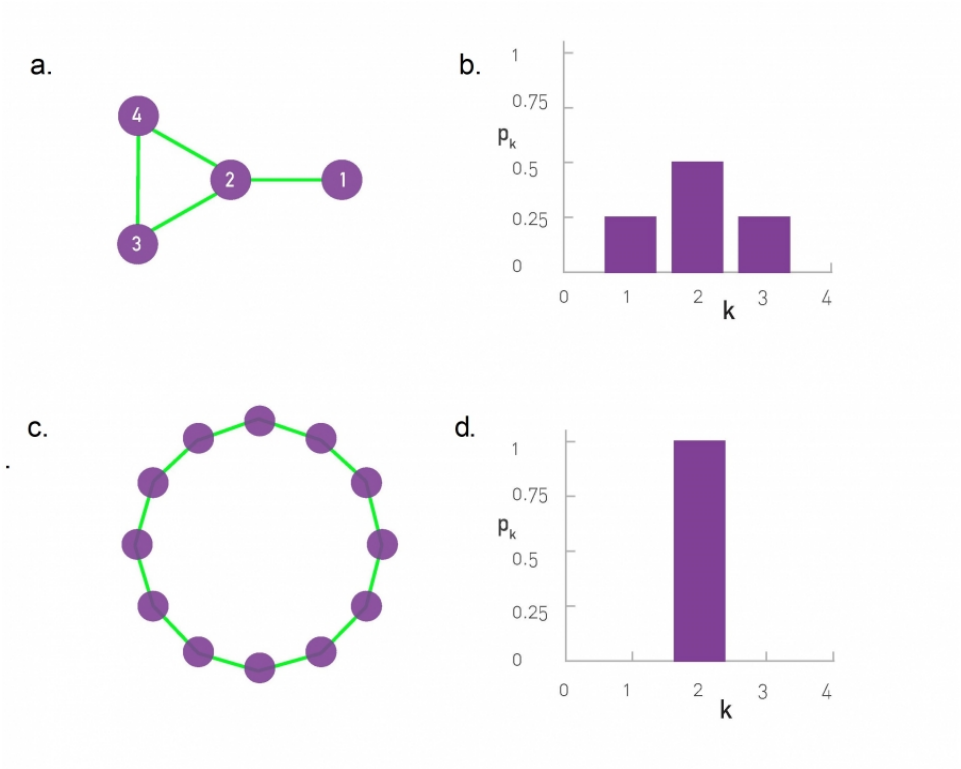

Рис. 2.2 Візуалізація розподілу ступенів.

2.1.4 Матриця суміжності

Існує кілька різних способів математичного представлення мережі. Розглянемо неорієнтовану мережу на рисунку 2.1. Вона однозначно описується переліком ребер $\{(1,\ 2),\ (1,\ 3),\ (2,\ 3),\ (2,\ 4)\}$. Один з найпопулярніших способів представлення мережі це представлення за допомогою **матриці суміжності**. Матриця суміжності має вигляд матриці $(N\ \times\ N)$. Для орієнтованого графу $A_{ij}\ =\ 1$, якщо існує зв'язок, що вказує від вузла j до вузла i, в іншому випадку $A_{ij}\ =\ 0$. Для неорієнтованого графу $A_{ij}\ =\ 1$, якщо існує зв'язок між вузлами j та i, а в іншому випадку $A_{ij}\ =\ 0$. Матриця суміжності неорієнтованої мережі є симетричною, тобто $A_{ij}\ =\ A_{ji}$.

Ступінь k_i вузла може бути отриманий з елементів матриці суміжностей. Для неорієнтованих мереж ступінь вузла обчислюється як сума по рядках або стовпцях матриці:

$$k_i = \sum_{j=1}^{H} A_{ji} = \sum_{i=1}^{H} A_{ji}, \tag{2.8}$$

Для спрямованих мереж розрізняємо вхідні та вихідні ступені відповідно:

$$k_i^{in} = \sum_{j=1}^{H} A_{ij} \qquad k_i^{out} = \sum_{j=1}^{H} A_{ji}, \tag{2.9}$$

Оскільки в неорієнтованій мережі ми маємо кількість вихідних посилань дорівнює кількості вхідних посилань, ми маємо наступне рівняння:

$$2L = \sum_{j=1}^{H} k_i^{in} = \sum_{j=1}^{H} k_i^{out} = \sum_{ij} A_{ij}, \tag{2.10}$$

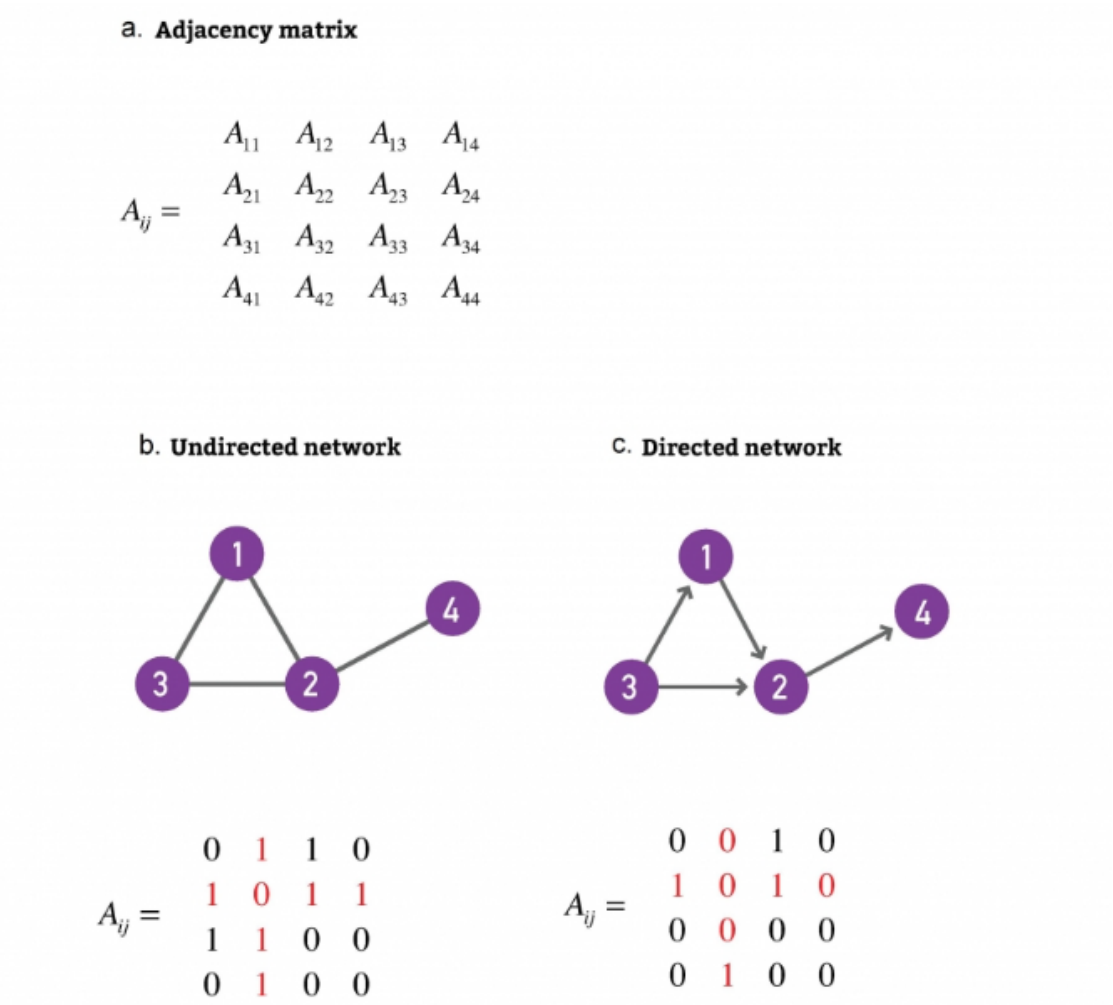

Рис. 2.3 Приклади матриці суміжності

2.1.5 Зважені графи

В деяких випадках ребра мережі корисно представляти з певною силою, вагою або значенням, як правило дійсним числом. Таким чином, у Інтернеті ребра можуть мати ваги, що представляють обсяг даних, які по ним передаються, або пропускну здатність. У соціальній мережі зв'язки можуть мати ваги, що представляють частоту контактів між учасниками. Такі мережі називаються **зваженими** і можуть бути представлені, задаючи елементам матриці суміжності значення, що відповідають вагам відповідних зв'язків $A_{ij} = \omega_{ij}$.

2.1.6 Шлях, відстань

Шлях у мережі - це будь-яка послідовність вершин, в якій кожна наступна пара вершин в послідовності з'єднана ребром в мережі. Інакше кажучи, шлях - це маршрут через мережу, який проходить від вершини до вершини вздовж ребер мережі. Шляхи можуть бути визначені як для орієнтованих, так і для неорієнтованих мереж. У орієнтованій мережі кожне ребро, яким проходить шлях, повинно бути пройденим у правильному напрямку. У неорієнтованій мережі ребра можуть бути пройденими в будь-якому напрямку.

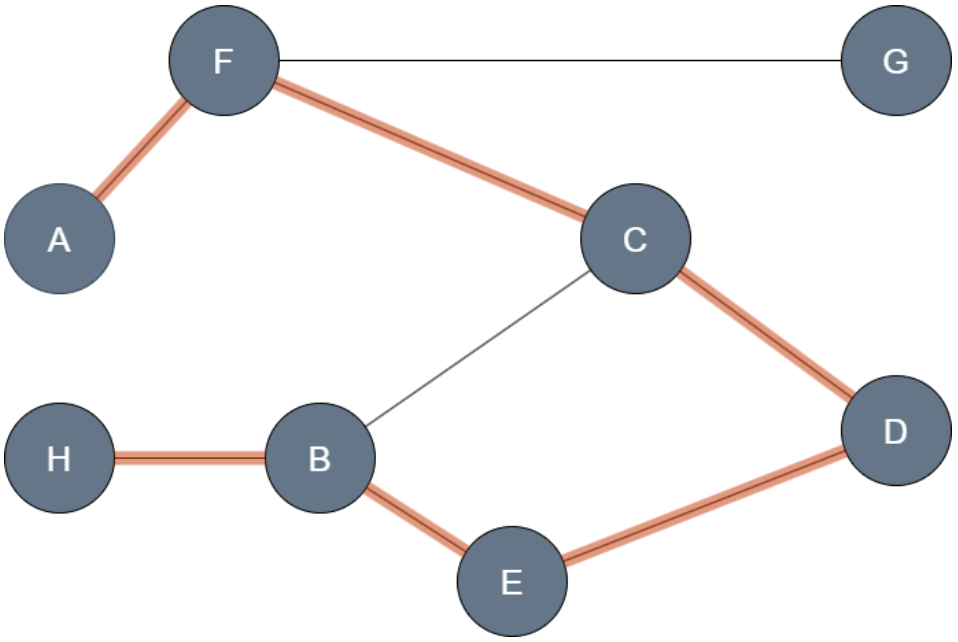

Рис. 2.4 Шлях з вершини *A* до вершини *H* в неорієнтованій мережі

Довжина шляху відображає кількість посилань, які входять в шлях. На рисунку 2.4 значення довжини шляху дорівнює 6.

Найкоротший шлях між вузлами i та j - це шлях з найменшою кількістю ребер. Найкоротший шлях також називають відстанню між вузлами i та j позначають як d_{ij} або просто d. Між парою вузлів може існувати кілька найкоротших шляхів однакової довжини d. Найкоротший шлях ніколи не містить петель або перетинає себе. У реальних мережах часто потрібно знайти найкоротший шлях і якщо у випадку мережі з рисунку 2.4 це проста задача, то для мережі з мільйонами зв'язків це може зайняти багато часу. На практиці для цього використовується алгоритм пошуку в ширину.

Середня довжина шляху - це середня відстань між усіма парами вершин в мережі.

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j=1,N;\, i \neq j} d_{i,j}, \tag{2.11}$$

2.1.7 Зв'язність

У неорієнтованій мережі вузли i та j з'єднані, якщо між ними існує шлях. Вони роз'єднані, якщо такий шлях не існує, у такому випадку ми маємо $d_{ij} = \infty$. Мережа є **зв'язною**, якщо всі пари вузлів у мережі з'єднані. Мережа є **роз'єднаною**, якщо є принаймні одна пара з $d_{ij} = \infty$. Очевидно, що мережа, показана на рисунку 2.5, є роз'єднаною, і ми називаємо її дві підмережі компонентами або кластерами.

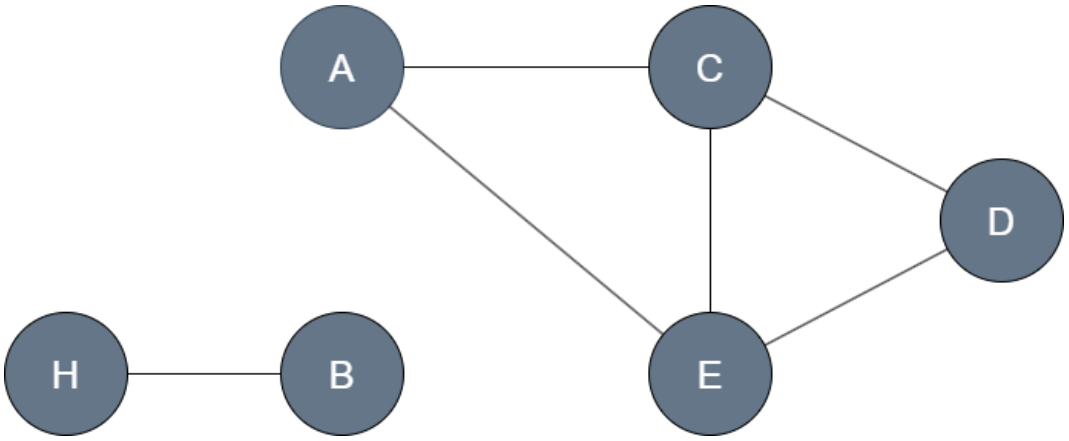

Рис. 2.5 Мережа з двома кластерами

Кластер - це підмножина вузлів в мережі, така, що існує шлях між будь-якими двома вузлами, які належать до кластеру, але неможливо додати до неї будь-які інші вузли з такою самою властивістю. Якщо мережа складається з двох кластерів, можна підключити одне правильно розташоване ребро, що з'єднує їх, і зроб

Мостом є будь-яке ребро таке, що якщо його перервати, то мережа ити мережу зв'язною. Таке посилання називається мостом. роз'єднається. Наприклад на рисунку 2.5 уявне ребро (0, 5) було б мостом.

2.1.8 Коефіцієнт кластеризації

Коефіцієнт кластеризації є мірою локальної зв'язності вузла в мережі. Він вказує на те, наскільки ймовірно, що сусідні вузли підключені один до одного. Коефіцієнт кластеризації вимірюється від 0 до 1 і для вузла з ступенеа k_i може бути розрахований за формулою:

$$C_i = \frac{2L_i}{k_i(k_i-1)} \quad , \tag{2.12}$$

L_i - кількість зв’язків між k_i сусідніми вузлами вузла i

Коефіцієнт кластеризації всієї мережі визначається як середнє значення коефіцієнту кластеризації в цій мережі:

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i \, , \tag{2.13}$$

2.2 Випадкові мережі

2.2.1 Модель випадкової мережі

Випадкова мережа - це модель мережі, де вузли вже задані, а зв'язки (ребра) між ними формуються з певною ймовірністю. Випадкові графи використовуються як інструмент для дослідження і розуміння властивостей реальних мереж.

Цінність випадкових графів полягає в їхній важливості як точки відліку для порівняння з реальними мережами. Вони надають нам можливість створити

контрольне середовище, де ми можемо вивчати різні аспекти мережевої структури та поведінки без урахування конкретних деталей реальної системи.

Досліджуючи випадкові графи, ми можемо розуміти такі важливі поняття, як розподіл степенів вузлів, кластеризація, відстані між вузлами, шляхи в мережі та багато інших. Це дозволяє нам отримати узагальнені знання про те, як працюють різні типи мереж і як вони впливають на поширення інформації, передачу сигналів, епідемії та інші процеси.

Випадкова мережа складається з N вузлів, де кожна пара вузлів з'єднана з ймовірністю p.

Загальний алгоритм побудови випадкової мережі виглядає так:

1. Починаємо з N ізольованих вузлів.
2. Обираємо пару вузлів і генеруємо випадкове число від 0 до 1. Якщо число перевищує значення p, з’єднуємо вибрану пару вузлів; в іншому випадку залишаємо їх не з'єднаними.
3. Повторюємо крок (2) для кожної з N(N-1)/2 пар вузлів.

2.2.2 Варіативність випадкової мережі

Випадкові мережі з однаковими параметрами N та p можуть виглядати по різному (рисунок 2.6). Як бачимо змінюється не лише схема підключення а й загальна кількість зв'язків L.

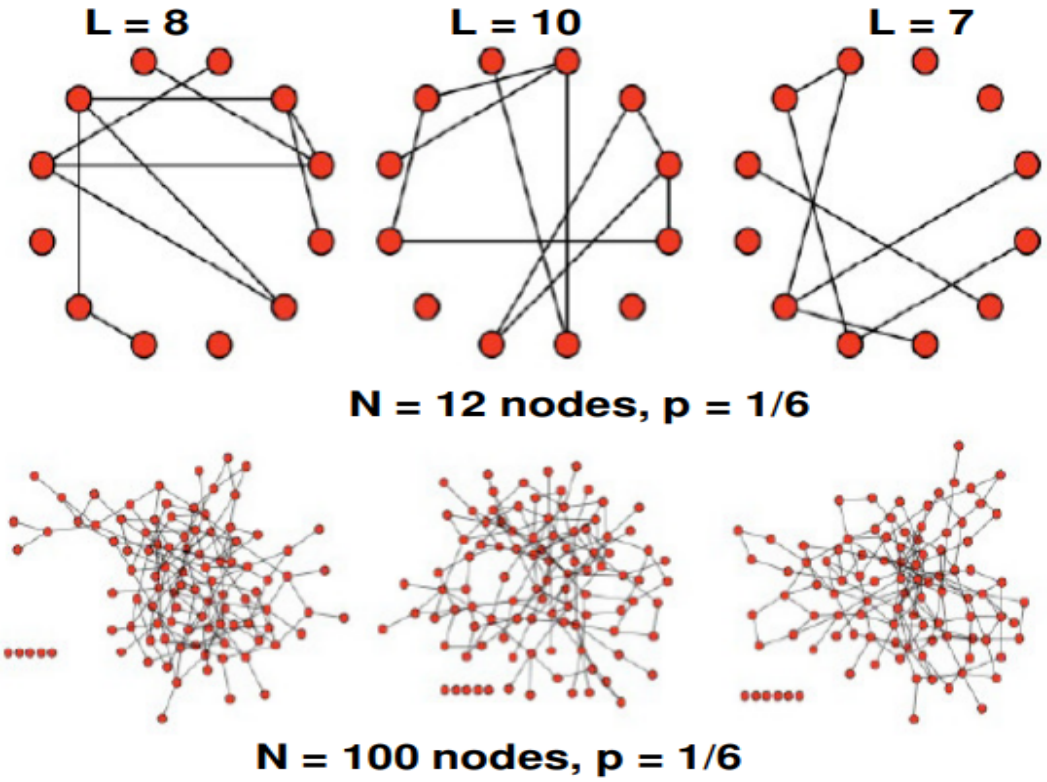

Рис. 2.6 Приклади випадкових мереж з однаковими параметрами

Спробуємо обчислити очікуване значення L з параметрами N і p. Ймовірність того, що випадкова мережа матиме саме L зв'язків може бути обчислена як добуток трьох факторів:

- Ймовірність того, що L спроб з'єднати $\frac{N(N-1)}{2}$ пар вузлів привели до зв'язку. Ця ймовірність дорівнює p^L.
- Ймовірність того, що $\frac{N(N-1)}{2} - L$ не привели до зв'язку. Ця ймовірність дорівнює $(1 - p)^{\frac{N(N-1)}{2} - L}$
- Також слід врахувати всі можливі комбінації того як ми можемо розмістити L зв'язків серед $\frac{N(N-1)}{2}$ пар вузлів. Тобто $C^L_{\frac{N(N-1)}{2}}$

Отже маємо ймовірність того, що конкретна реалізація випадкового графу має саме L зв'язків можна обчислити за формулою:

$$p_L = C^L_{\frac{N(N-1)}{2}} \, p^L \, (1 - p)^{\frac{N(N-1)}{2} - L}, \tag{2.14}$$

Оскільки це форма **біноміального розподілу**, то очікувана кількість зв'язків обчислюється за формулою:

$$\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} L p_L = p\frac{N(N-1)}{2}, \tag{2.15}$$

Тобто середня кількість зв'язків $\langle L \rangle$ у мережі задається добутком ймовірності що два вузли будуть пов'язані і максимальної кількості можливих зв'язків. Аналогічно ми можемо визначити середню ступінь вершини у випадковому графі як добуток ймовірності p, що дві вершини будуть з'єднані і максимально можливого ступеню для вершини $p(N - 1)$:

$$\langle k \rangle = p(N - 1), \tag{2.16}$$

2.2.3 Розподіл ступенів випадкової мережі

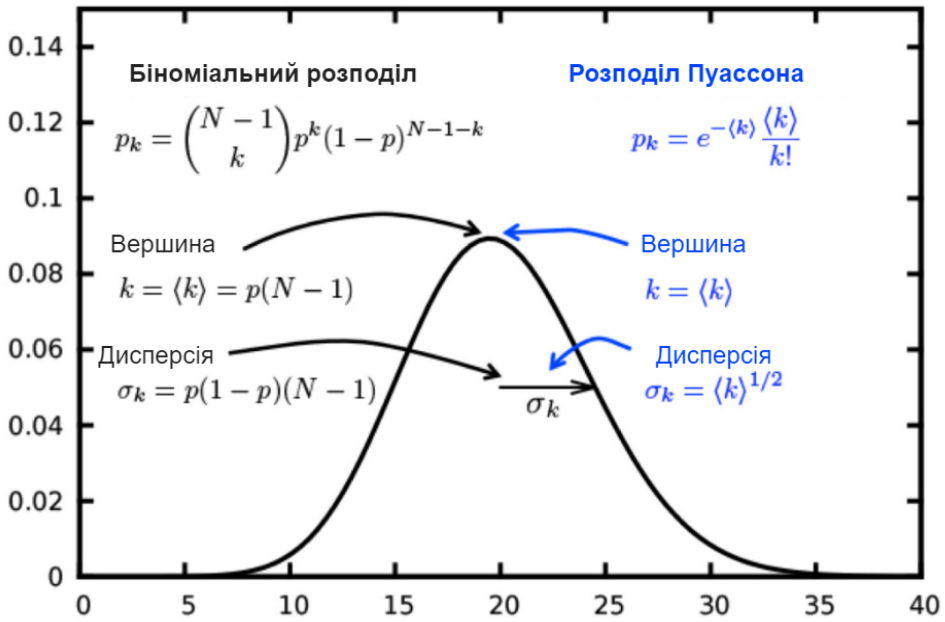

Рис. 2.7 Порівняння представлення розподілу випадкової мережі в біноміальній і пуассонівській формі.

Як ми вже з'ясували вигляд розподілу ступенів у випадковій мережі є біноміальним розподілом (ліва половина рисунку 2.7). При $N \gg \langle k \rangle$ біноміальний розподіл наближається до розподілу Пуассона (права половина рисунку 2.7). Оскільки обидві формули можуть описувати той самий розподіл, вони мають ідентичні властивості, але вони виражені в термінах різних параметрів: біноміальний розподіл залежить від p і N, тоді як у розподілі Пуассона є лише один

параметр, *‹k›*. Саме ця простота робить форму Пуассона більш зручною для обчислень.

Якщо розглядати біноміальний розподіл, то ймовірність того, що довільний вузол матиме *k* зв'язків дорівнює:

$$p_k = C^k_{N-1} p^k (1 - p)^{N-1-k}, \tag{2.17}$$

Проте більшість реальних мереж задовольняють умову $N \gg \langle k \rangle$. В такому випадку розподіл може бути апроксимовано до розподілу Пуассона і в такому випадку отримаємо:

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}, \tag{2.18}$$

Біноміальний і пуассонівський розподіли описують одну й ту саму величину, тому вони мають схожі властивості. Проте точний результат для розподілу ступенів є біноміальною формою, а форма з розподілу Пуассона є лише наближенням, яке є дійсним в умовах обмеження $N \gg \langle k \rangle$. Оскільки більшість реальних мереж є розрідженими, то ця умова зазвичай виконується.

Перевага пуассонівської форми полягає в тому, що ключові характеристики мережі, мають набагато простішу форму (Рис. 2.7) і залежать від одного параметра - ‹k›.

Також слід пам'ятати, що пуассонівський розподіл у не залежить від числа вузлів N, тому передбачає, що розподіл ступенів мереж різних розмірів, але з однаковим середнім ступенем ‹k›, неможливо відрізнити один від одного.

2.2.4 Розвиток випадкової мережі

Уявимо мережу що складається з *N* ізольованих вузлів в якій поступово з'являються нові зв'язки. Це відповідає поступовому збільшенню *p*. Щоб кількісно оцінити цей процес, ми спочатку перевіряємо, як розмір найбільшого підключеного кластера в мережі, N_G , змінюється в залежності від *‹k›*. Крайні випадки є очевидними:

Для $p = 0$ ми маємо $\langle k \rangle = 0$, тому всі вузли ізольовані. Тому найбільший компонент має розмір $N_G = 1$ і $\frac{N_G}{N} \to 0$ для великих N .

Для $p = 1$ ми маємо $\langle k \rangle = N - 1$, отже, мережа є звязним графом і всі вузли належать одному компоненту. Тому $N_G = N$ і $\frac{N_G}{N} \to 1$.

Очікувалося, що найбільший компонент поступово зростатиме від $N_G = 1$ до $N_G = N$ при зростанні $\langle k \rangle$ від 0 до N-1. Проте, як показано на рисунку 2.8, це не так: співвідношення $\frac{N_G}{N}$ залишається рівним нулю для невеликих значень $\langle k \rangle$, що свідчить про відсутність великого кластера.

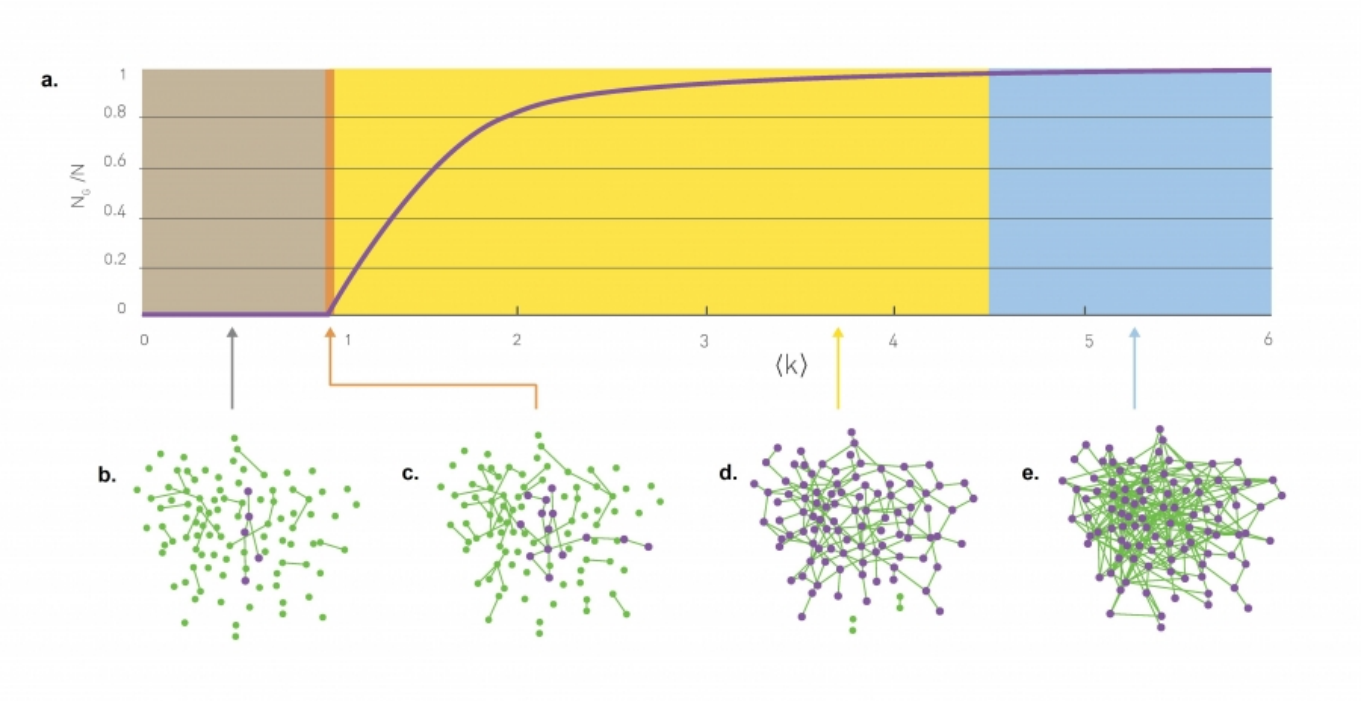

Рис.2.8 Розвиток випадкової мережі

Як тільки $\langle k \rangle$ перевищує критичне значення, $\frac{N_G}{N}$ зростає, що сигналізує швидку появу великого кластера, який ми називаємо гігантським компонентом. Ердьош і Реньї передбачили, що умова появи виконується тоді і тільки тоді, коли кожен вузол в середньому має більше одного зв'язку:

$$\langle k \rangle = 1, \tag{2.19}$$

Використовуючи формулу 2.16 можемо виразити 2.19 як:

$$p_c = \frac{1}{N-1} \approx \frac{1}{N}, \tag{2.20}$$

Отже, чим більша мережа, тим менший p буде достатнім для утворення гігантського компонента.

Дивлячись на рисунок 2.8 ми можемо розрізнити чотири режими які топологічно розрізняються і кожен з яких має свої унікальні характеристики:

1. **Субкритичний режим**: $0 < \langle k \rangle < 1$ ($p < \frac{1}{N}$, рисунок 2.8b)

   При *‹k›* = 0 мережа складається з *N* ізольованих вузлів. Збільшення *‹k›* означає, що ми додаємо до мережі *N‹k› = pN(N*-1*)/*2 зв'язків. Однак, оскільки *‹k›* ‹ 1, в цьому режимі у нас є лише невелика кількість зв'язків, тому ми спостерігаємо переважно маленькі кластери.

   Ми можемо визначити найбільший кластер у будь-який момент як гігантський компонент. Проте в цьому режимі відносний розмір найбільшого кластера, $\frac{N_G}{N}$, залишається рівним нулю. Причина полягає в тому, що для *‹k›* < 1 найбільший кластер є деревом розміром $N_G \sim lnN$, тому його розмір зростає набагато повільніше, ніж розмір мережі. Тому $\frac{N_G}{N} \simeq \frac{lnN}{N} \to 0$, при $N \to \infty$.

2. **Критична точка:** $\langle k \rangle = 1$ ($p = \frac{1}{N}$, рисунок 2.8c)

   Критична точка відділяє режим, коли ще немає гігантського компонента ($\langle k \rangle < 1$) від режиму, коли він вже є ($\langle k \rangle > 1$). У цій точці відносний розмір найбільшого компонента все ще дорівнює нулю. Насправді розмір найбільшого компонента складає $N_G \sim N^{2/3}$. Відповідно N_G зростає набагато повільніше, ніж розмір мережі, а отже його відносний розмір зменшується при $N \to \infty$

Проте варто зауважити, що в цій точці відбувається значний стрибок у розмірі найбільшого компонента при ‹*k*› = 1. Наприклад, для випадкової мережі з $N = 7 \times 10^9$ вузлами, при$\langle k \rangle < 1$ найбільший кластер становить $N_G \simeq lnN = ln(7 \times 10^9) \simeq 22.7$. У порівнянні, при $\langle k \rangle = 1$ очікується $N_G \sim N^{2/3} \simeq 3.6 \times 10^6$. Проте, як у субкритичному режимі, так і в критичній точці найбільший компонент містить лише мізерну частку від загальної кількості вузлів в мережі.

3. **Надкритичний режим:** $\langle k \rangle > 1$ ($p > \frac{1}{N}$, рисунок 2.8d)

Цей режим найбільш актуальний для реальних систем, оскільки тут ми вже маємо гігантський компонент, який виглядає як мережа. В околі критичної точки розмір гігантського компонента змінюється як:

$$\frac{N_G}{N} \sim \langle k \rangle - 1, \tag{2.21}$$

Гігантський компонент містить скінченну частину вузлів. Чим далі ми відходимо від критичної точки, тим більша частина вузлів належатиме до нього.Варто зауважити, що (2.21) є дійсним лише в околі $\langle k \rangle = 1$. Для великих $\langle k \rangle$ залежність між N_G та $\langle k \rangle$ є нелінійною.

4. **Зв'язний режим:** $\langle k \rangle > lnN$ ($p > \frac{lnN}{N}$, рисунок 2.8e)

Для достатньо великих значень *p* гігантський компонент поглинає всі вузли та компоненти, тому $N_G \simeq N$. У відсутності ізольованих вузлів мережа стає зв'язною. Середня ступінь, при якій це відбувається, залежить від *N* і дорівнює

$$\langle k \rangle = lnN, \tag{2.22}$$

Варто зауважити, що коли ми переходимо до зв'язного режиму, мережа все ще є досить розрідженою, оскільки $\frac{lnN}{N} \to 0$ для великих N . Мережа перетворюється на повний граф лише при $\langle k \rangle = N - 1$.

2.2.5 Властивість “світ малий”

Концепція "світ малий" (англ. small-world) є одним із ключових понять в теорії комплексних мереж. Вона відноситься до властивості мережі, де вершини мають високу кластеризацію та короткі шляхи між собою.

Один з найвідоміших прикладів "світу малого" є "експеримент шести рукостискань". Цей експеримент передбачає, що кожна людина на планеті може знайти шлях до будь-якої іншої людини через серію ланцюжків зі знайомих людей, в яких кожна людина знає не більше шести інших людей. Це свідчить про те, що навіть у величезних мережах, таких як соціальні мережі, існує дуже короткий шлях між будь-якими двома людьми.

Основна ідея "малого світу" полягає в тому, що в більшості реальних мереж, включаючи соціальні мережі, мережі співробітництва, мережі дружби, мережі залежностей між веб-сторінками та багато інших, можна знайти дуже короткий шлях між будь-якими двома вершинами через декілька посередників (проміжних вершин).

Це означає, що в мережі існує висока кластеризація, що означає, що вершини тенденцію формувати локальні групи, де багато з'єднань між сусідніми вершинами. У той же час, незважаючи на цю локальну структуру, відстань між будь-якими двома вершинами в мережі є відносно невеликою. Це можна інтерпретувати як те, що мережа має "короткі шляхи" між вершинами, де відстань вимірюється кількістю ребер, які потрібно пройти, щоб дістатися від однієї вершини до іншої. Це твердження викликає запитання: що означає “короткий”, тобто короткий порівняно з чим?

Розглянемо випадкову мережу із середнім ступенем *‹k›*. Вузол у цій мережі має в середньому ‹k› вузлів на відстані $d = 1$, ‹k›^{2} вузлів на відстані $d = 2$, … , ‹k›^{d} вузлів на відстані *d*.

Наприклад, якщо ‹k› $\approx$ 1000, що є оціночною кількістю знайомих, які має людина, ми очікуємо 10^6 осіб на відстані два, тобто майже все населення Землі, на відстані три від нас.

Отже очікувана кількість вузлів на відстані d від нашого початкового вузла дорівнює:

$$N(d) \approx 1 + \langle k \rangle + \langle k \rangle^2 + \ldots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1}, \tag{2.23}$$

$N(d)$ не має перевищувати загальну кількість вузлів. Ми можемо ідентифікувати діаметр мережі d_{max} встановивши $N(d_{max}) \approx N$. Припускаючи, що ‹k› $\gg 1$ ми можемо знехтувати (-1) в чисельнику і знаменнику (2.23). Таким чином маємо:

$$\langle k \rangle^{d_{max}} \approx N, \tag{2.24}$$

Тому можемо отримати діаметр випадкової мережі як:

$$d_{max} \approx \frac{\ln N}{\ln \langle k \rangle}, \tag{2.25}$$

Формула (2.25) показує математичне формулювання феномену малого світу. Проте для більшості мереж (2.25) надає кращу апроксимацію середньої відстані між двома випадково обраними вузлами *‹d›*, ніж d_{max}. Тому зазвичай властивість малого світу визначається умовою:

$$\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}, \tag{2.26}$$

Ця умова описує залежність середньої відстані в мережі від *N* та *‹k›*.

Загалом $\ln N \ll N$, тому залежність *‹d›* від *ln N* означає, що відстані в випадкових мережах набагато менші за розмір мережі. Отже, під "малим" у "феномені малого світу" ми розуміємо те, що середня довжина шляху або діаметр пропорційні логарифму від розміру системи.

2.3 Властивість безмасштабності

2.3.1 Степеневий закон та безмасштабні мережі

Якби Інтернет був випадковою мережею, то ступені веб сторінок мали б відповідати розподілу Пуассона. Проте насправді форма Пуассона погано підходить для розподілу ступенів (на рисунку 2.9 ми бачимо розподіли степенів для мережі Інтернет у логарифмічному представленні). Натомість отриманий графік добре апроксимується з:

$$p_k \sim k^{-\gamma}, \tag{2.27}$$

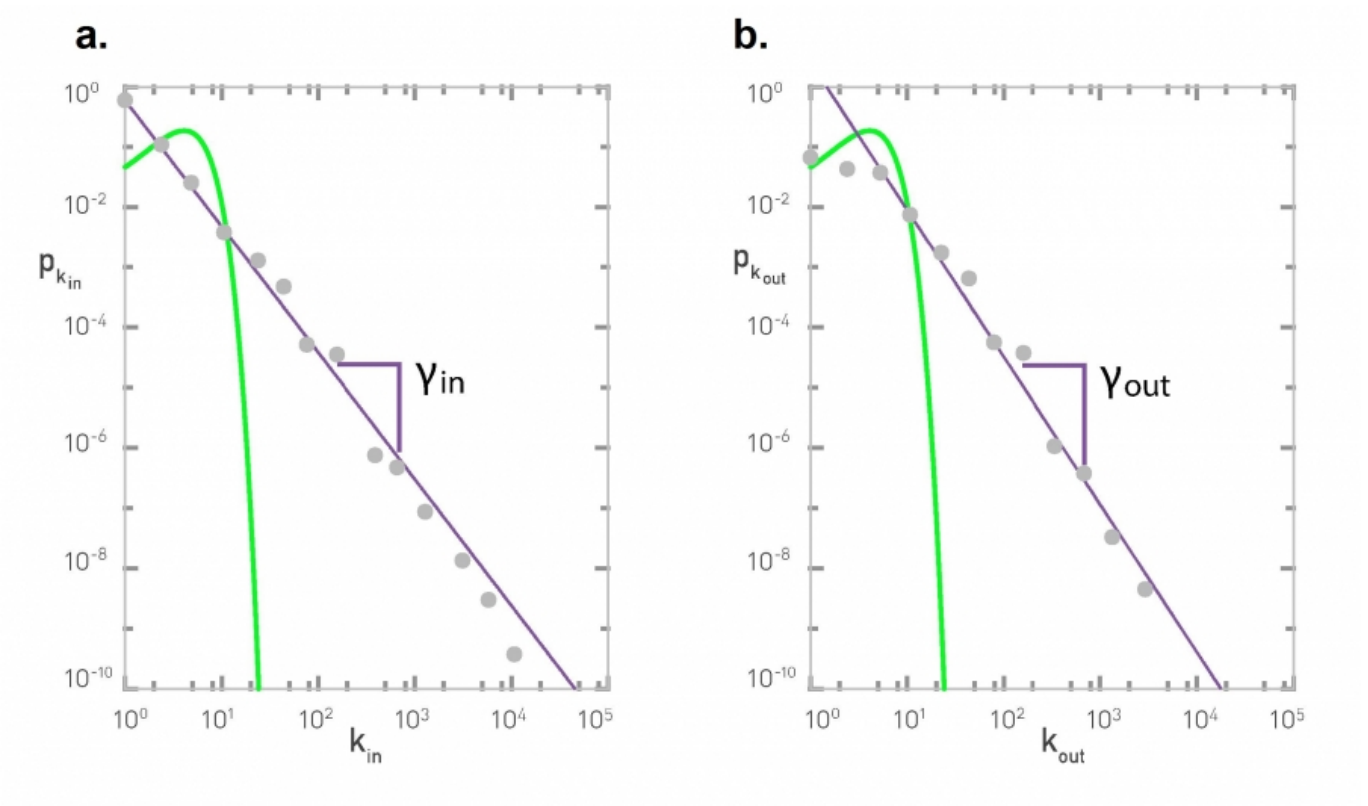

Рис 2.9 Розподіл ступенів для мережі Інтернет (зеленим кольором зображено розподіл Пуассона)

Рівняння (2.23) називається **розподілом ступеня степеневого закону**, а γ є його ступеневим показником. Взявши логарифм справа і зліва отримаємо:

$$log\, p_k \sim -\ \gamma\, log\, k, \tag{2.28}$$

Отже, якщо (2.23) справедливе, тоді $log\ p_k$ залежатиме лінійно від $log\ k$, кут нахилу цієї лінії буде експонентою ступеня γ.

Емпіричні результати, показані на рисунку 2.9, показують існування мережі, чий розподіл ступенів досить відмінний від розподілу Пуассона, що характеризує випадкові мережі. Ми називаємо такі мережі безмасштабними. Отже **безмасштабна мережа** - це мережа, розподіл ступенів якої підпорядковується розподілу степеневого закону.

Щоб краще зрозуміти властивості масштабної вільності, нам потрібно визначити розподіл степеневого закону більш чіткими термінами:

Дискретне формулювання

Оскільки ступені вузлів є додатніми цілими числами, $k = 0, 1, 2, ...$, дискретний формалізм надає ймовірність p_k того, що вузол має саме k зв'язків як:

$$p_k = Ck^{-\gamma}, \tag{2.29}$$

Тут C представляє собою нормалізуючий коефіцієнт, оскільки $\sum_{k=1}^{\infty} p_k = 1$. Отже отримуємо:

$$C \sum_{k=1}^{\infty} k^{-\gamma} = 1, \tag{2.30}$$

Тобто:

$$C = \frac{1}{\sum_{k=1}^{\infty} k^{-\gamma}} = \frac{1}{\zeta(\gamma)}, \tag{2.31}$$

Де $\zeta(\gamma)$ - дзета-функція Римана. Таким чином для $k > 0$ дискретний розподіл степеневого закону має вигляд:

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}, \tag{2.32}$$

Неперервне формулювання

У аналітичних розрахунках часто зручно припускати, що ступені можуть мати будь-яке додатне дійсне значення. У цьому випадку ми записуємо розподіл ступенів за законом степенів як:

$$p(k) = Ck^{-\gamma}, \tag{2.33}$$

Оскільки $\int_{k_{min}}^{\infty} p(k)dk = 1$, то ми отримуємо:

$$C = \frac{1}{\int_{k_{min}}^{\infty} k^{-\gamma}dk} = (\gamma - 1)k_{min}^{\gamma-1}, \tag{2.34}$$

Тому неперервне формулювання розподілу ступенів має форму:

$$p(k) = (\gamma - 1)k_{min}^{\gamma-1}k^{-\gamma}, \tag{2.35}$$

2.3.2 Хаби

Основна різниця між випадковою та масштабно-вільною мережами полягає в розподілі ступенів. Щоб це проілюструвати, на рисунку 2.9 порівнюється розподіл степеневого закону з розподілом Пуассона. Видно певні властивості:

- Для невеликих k степеневий закон перевищує функцію Пуассона, що свідчить про те, що безмасштабна мережа має велику кількість вузлів з невеликим ступенем, більшість з яких відсутні у випадковій мережі.
- Для k в околі $\langle k \rangle$ розподіл Пуассона перевищує степеневий закон, що свідчить про те, що в випадковій мережі існує надлишок вузлів зі ступенем $k \approx \langle k \rangle$.
- Для великих k степеневий закон знову перевищує криву Пуассона. Це свідчить про те, що ймовірність спостереження вузла з високим ступенем, тобто хабу, на декілька порядків вища в безмасштабній мережі, ніж у випадковій.

З’ясуємо як розмір мережі впливає на розмір її хабів. Для цього обчислимо максимальний ступінь вершини в мережі k_{max} який називають природним обмеженням розподілу ступеня p_k. Величина k_{max} і буде відображати розмір

найбільшого хабу. З метою наочності проведемо обчислення для експоненційного розподілу:

$$p(k) = Ce^{-\lambda k}, \tag{2.36}$$

Для мережі з мінімальним ступенем k_{min} маємо таку умову нормалізації:

$$\int_{k_{min}}^{\infty} p(k)dk = 1, \tag{2.37}$$

Звідси маємо $C = \lambda e^{\lambda k_{min}}$. Щоб обчислити k_{max} ми припускаємо, що в мережі з N вузлами ми очікуємо не більше одного вузла в межах (k_{max}, ∞). Тобто ймовірність того, що навмання обраний вузол ступінь якого рівна k_{max} дорівнює $\frac{1}{N}$:

$$\int_{k_{max}}^{\infty} p(k)dk = \frac{1}{N}, \tag{2.38}$$

З рівняння (2.34) маємо:

$$k_{max} = k_{min} + \frac{lnN}{\lambda}, \tag{2.39}$$

Оскільки функція lnN зростає повільно, то це означає, що найбільший ступінь не буде сильно відрізнятися від найменшого.

Для розподілу ступенів Пуассона розрахунок трохи більш складний, але отримана залежність k_{max} від N є ще повільнішою, ніж логарифмічна залежність.

Для безмасштабної мережі залежність k_{max} від N буде наступною:

$$k_{max} = k_{min} N^{\frac{1}{\gamma-1}}, \tag{2.40}$$

Отже, чим більша мережа, тим більший ступінь її найбільшого вузла. Поліноміальна залежність k_{max} max від N означає, що в великій безмасштабні мережі може бути значна різниця між найменшим вузлом, k_{min}, і найбільшим вузлом, k_{max}.

2.3.3 Властивість малого світу для безмасштабних мереж

Наявність хабів у випадковій мережі не може не відобразитись на властивості “малого світу”. Більшість безмасштабних мереж, що представляють практичний інтерес, мають цю властивість, яка власне і є наслідком хабів, які діють як мости між багатьма малими вузлами.

2.3.4 Універсальність безмасштабних мереж

За останнє десятиліття виявлено, що багато реальних мереж мають властивість, яку називають безмасштабністю, і це має велике значення у науці, технології та суспільстві. На рисунку 2.10 показані розподіли ступенів вузлів для різних мереж: інфраструктурної (Інтернет, рис. 2.10 (a)), біологічної (взаємодії білків, рис. 2.10 (b)), комунікаційної (електронні листи, рис. 2.10 (c)) та мережі наукових комунікацій (цитування, рис. 2.10 (d)). Для кожної мережі розподіл ступенів відрізняється від пуассонівського розподілу і краще піддається апроксимації степеневим законом.

Наприклад, Інтернет - це штучна мережа з історією, що нараховує всього три десятиліття, тоді як мережа взаємодій білків є результатом чотирьох мільярдів років еволюції. У деяких мережах вузлами є молекули, в інших - комп'ютери. Така різноманітність змушує нас називати безмасштабність універсальною характеристикою мереж.

Для того щоб класифікувати мережу, ми можемо скористатися побудовою розподілу ступенів в ній. Таким чином ми зразу бачимо, чи є мережа безмасштабною: у таких мережах ступені найменших і найбільших вузлів значно відрізняються. В звичайній мережі ці вузли мають схожі ступені.

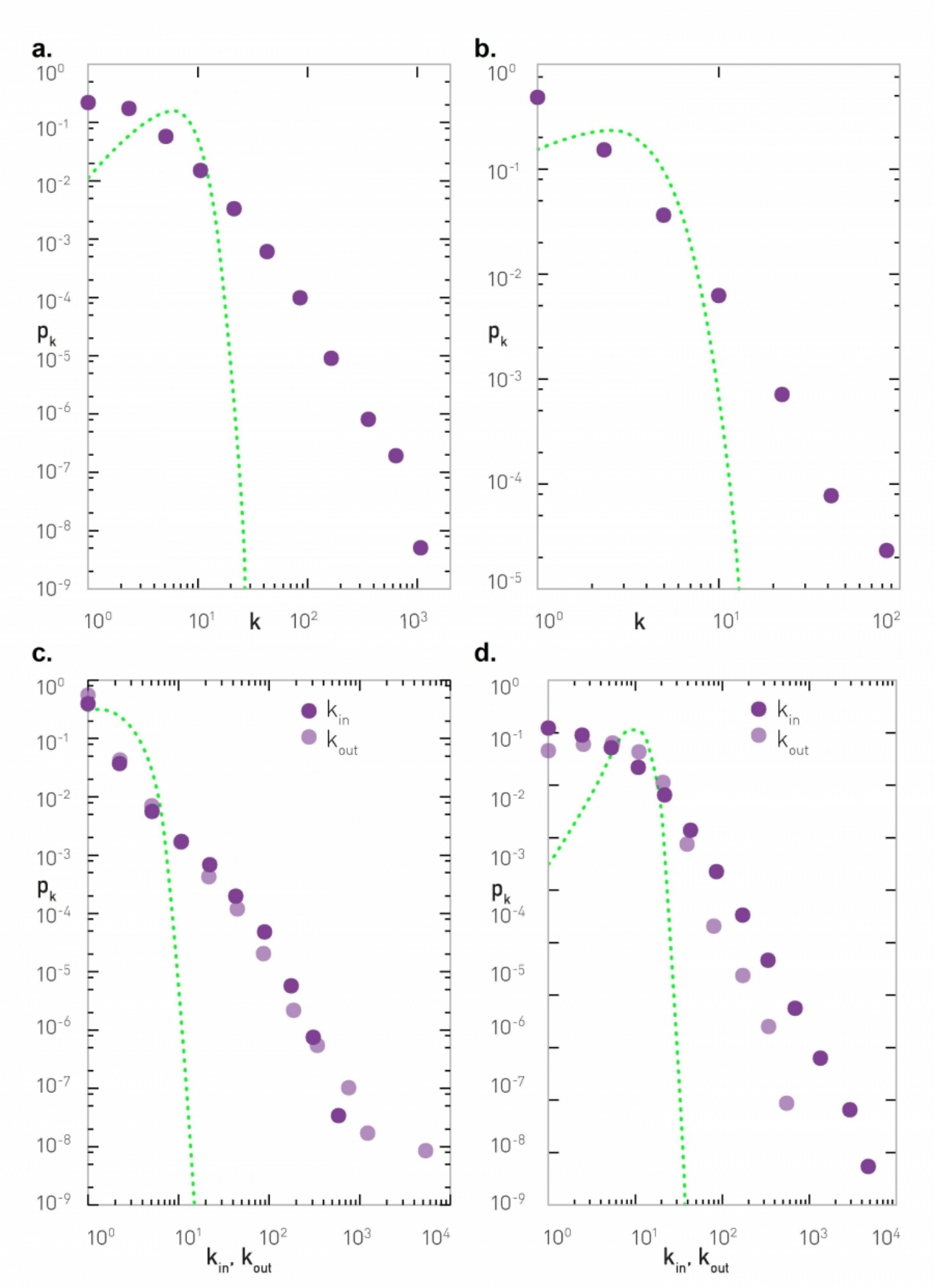

Рис. 2.10 Приклади розподілів ступенів для різних реальних мереж.

РОЗДІЛ 3. РЕАЛІЗАЦІЯ ПОСТАВЛЕНОЇ ЗАДАЧІ І АНАЛІЗ РЕЗУЛЬТАТІВ

3.1 Опис інструментів для реалізації задачі

Бібліотека *wikipediaapi* - це Python-інтерфейс до веб-сайту Вікіпедії, який дозволяє нам отримувати доступ до статей, категорій, посилань і багато іншого. Вона надає зручні методи для пошуку статей, отримання текстового змісту статей із Вікіпедії, а також інформації про їхню структуру. Ця бібліотека підтримує мовну локалізацію і дозволяє вибрати різні мови Вікіпедії для отримання даних.

Бібліотека *NetworkX* - це потужний інструмент для аналізу складних мережевих структур. Вона дозволяє створювати, маніпулювати та досліджувати різноманітні типи графів і мереж, такі як орієнтовані і неорієнтовані графи, мультиграфи, діграфи тощо. Бібліотека має широкий набір функцій для аналізу графів, розрахунку метрик, виявлення підграфів, генерації графів та багато іншого.

Gephi - це потужний візуалізаційний інструмент для аналізу та візуалізації складних мережевих структур. Він дозволяє імпортувати дані мережі з різних джерел, виконувати різні аналізи графів, відображати мережі у вигляді візуальних графів, застосовувати різні алгоритми розташування вузлів для покращення читабельності та дослідження структури мережі. Gephi також має розширення для виконання аналізу та візуалізації соціальних мереж, мереж транспорту, біологічних мереж і багато іншого.

3.2 Збір даних та формування графу

Для цього етапу була реалізована процедура збору даних з Вікіпедії та формування графу, яка дозволить вивчати взаємозв'язки між сторінками. Основна мета цього процесу - отримати інформацію з Вікіпедії та представити її у вигляді графічної структури.

Спочатку ми ініціалізуємо об'єкт *wikipediaapi*, що встановлює з'єднання з Вікіпедією. Потім вибирається категорія сторінок, яка буде взята як початкова точка дослідження, та перевіряється, чи існує ця категорія.

Наступний крок полягає у рекурсивному проходженні через елементи категорії. Було реалізовано функцію *log_categorymembers* яка зберігає назви сторінок у список *category_members.* Отриманий список зберігається у файлі *category_members.json*.

Далі зчитуємо дані з файлу *category_members.json*. Після чого проводиться збір посилань з усіх сторінок Вікіпедії з заданим рівнем глибини пошуку. Отримані посилання зберігаються у вигляді графу, де кожна сторінка є вузлом, а посилання між ними - ребром.

На основі зібраних даних будується граф який містить всі посилання з усіх знайдених сторінок. Граф зберігається у форматі GEXF для подальшої роботи з ним.

шлях до У кінці виводиться збережених графів, щоб користувач міг легко знайти їх у вказаному місці. Загалом, код забезпечує збір даних з Вікіпедії та побудову графу, що дає можливість подальшого аналізу та візуалізації інформації з Вікіпедії.

3.3 Аналіз та візуалізація мережі

Для аналізу і візуалізації мережі будемо використовувати інструмент Gephi. В результаті роботи попереднього алгоритму, обравши в якості початкової точки категорію “Ссавці” і виконавши збір статей з глибиною 3 ми отримали орієнтовану мережу що складається з 8052-х вершин та 25116 граней.

3.3.1 Визначення розподілу вершин

Як нам відомо з теоретичного розділу - більшість реальних є безмасштабними. Тож оскільки мережа статей вікіпедії є спрямованою мережею, то для перевірки цього припущення побудуємо розподіли ступенів входу і виходу для вершин отриманої мережі. Для початку для цього було застосовано Gephi:

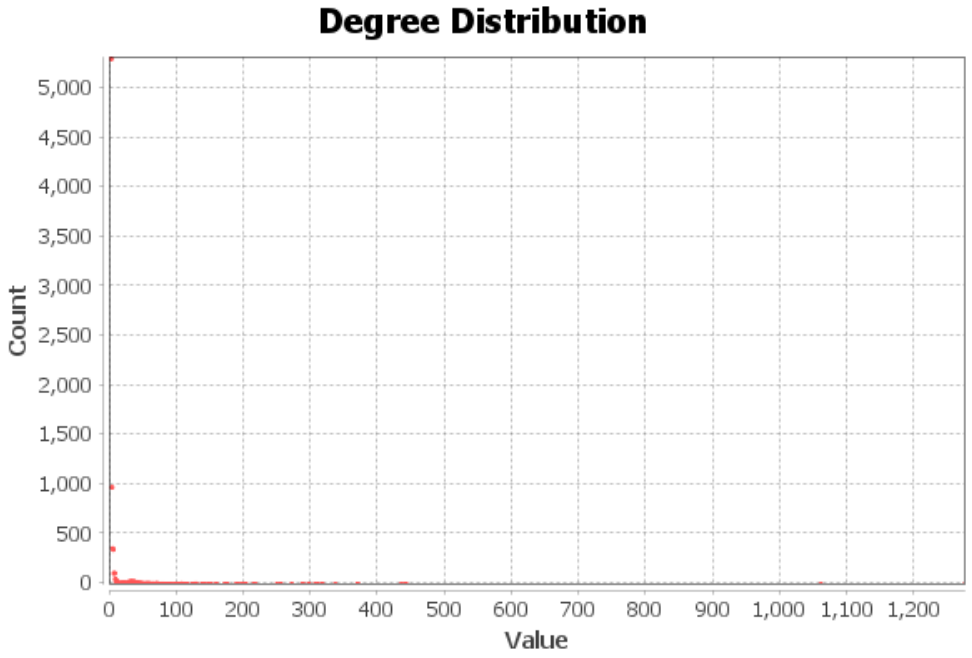

Рис. 3.1 Загальний розподіл ступенів.

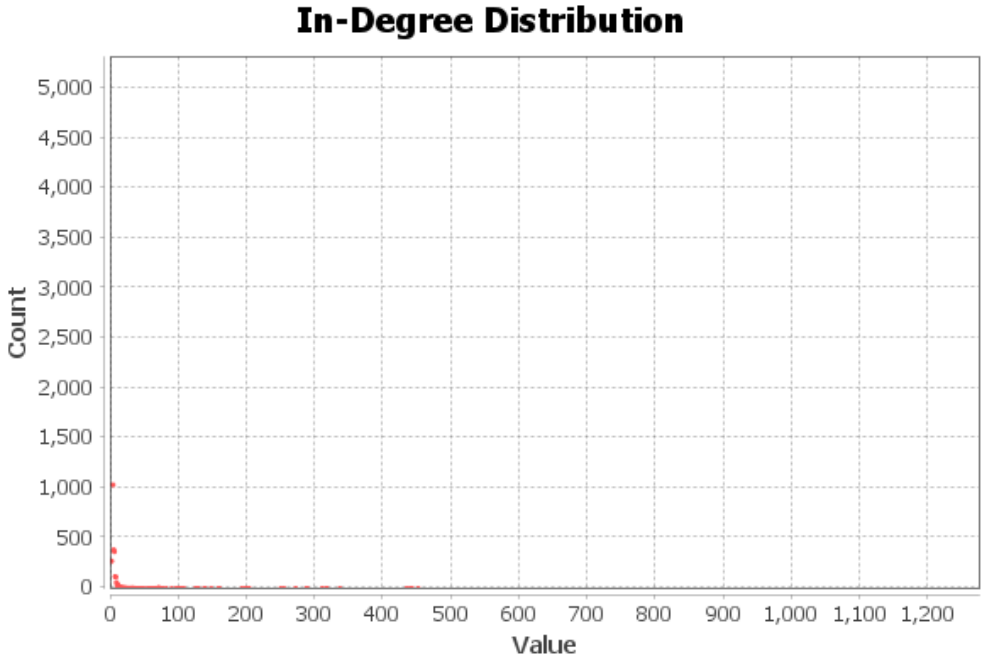

Рис. 3.2 Розподіл ступенів входу.

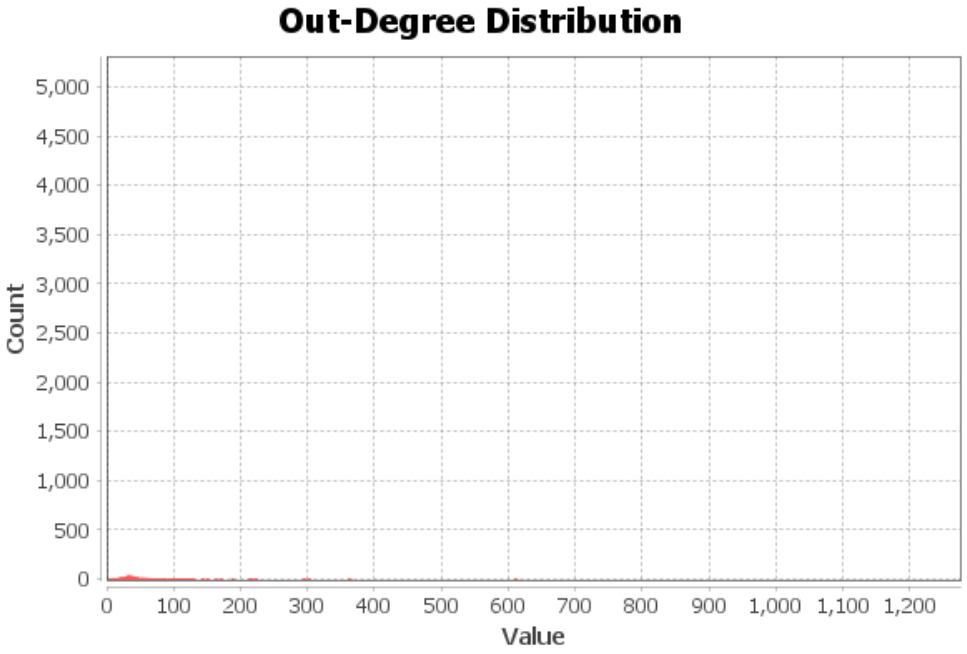

Рис. 3.3 Розподіл ступенів виходу.

Проте, як бачимо, коли ми маємо вузли з дуже різним ступенем, лінійний графік не може відобразити їх усі. Щоб отримати зрозумілі розподіли ступенів, було прийнято рішення використати логарифмічне представлення, гарантуючи таким чином, що кожна точка даних чітко відображатиметься. Ось що отримали в результаті:

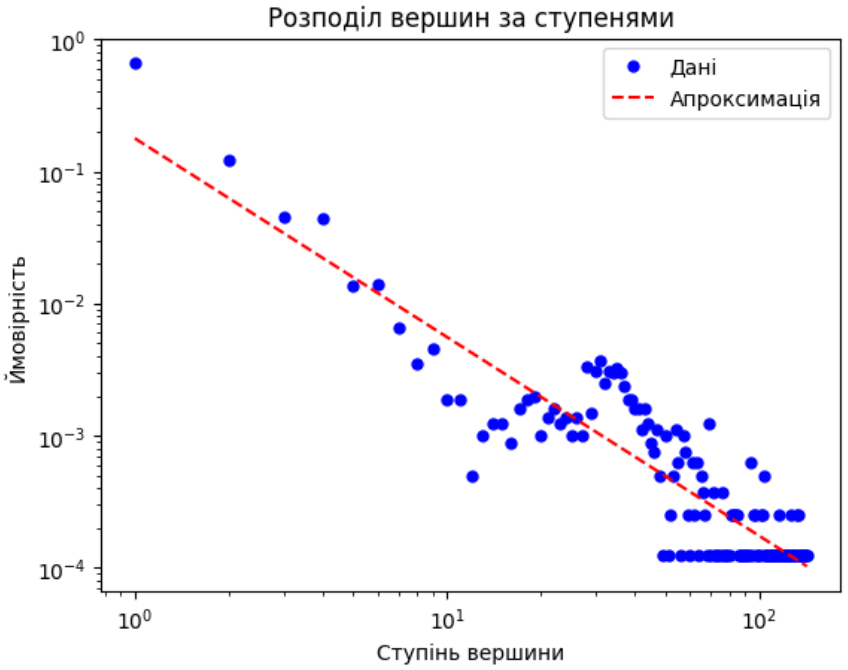

Рис. 3.4 Загальний розподіл ступенів в логарифмічному представленні.

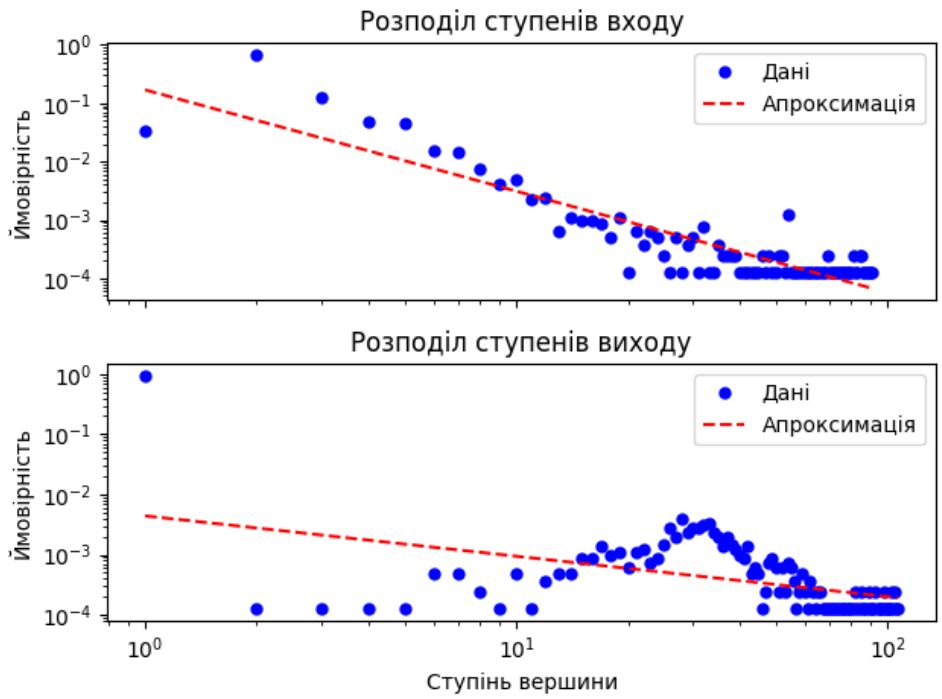

Рис. 3.5 Розподіл ступенів входу і виходу в логарифмічному представленні.

Як бачимо розподіл ступенів входу відповідає степеневому закону, але розподіл ступенів виходу має певні відхилення. Тому потрібно детальніше перевірити властивості отриманої мережі.

3.3.2 Наявність хабів

Однією з властивостей безмасштабних мереж є наявність хабів. Це легко помітити візуалізувавши мережу.

Для візуалізації було використано Gephi, розмір та колір вершин відображає їх ступінь. Відповідно чим більша і темніша вершина, тим більшим є її ступінь. Дивлячись на рисунок 3.6 легко помітити наявність двох великих хабів. Отже, ще одна властивість безмасштабної мережі присутня.

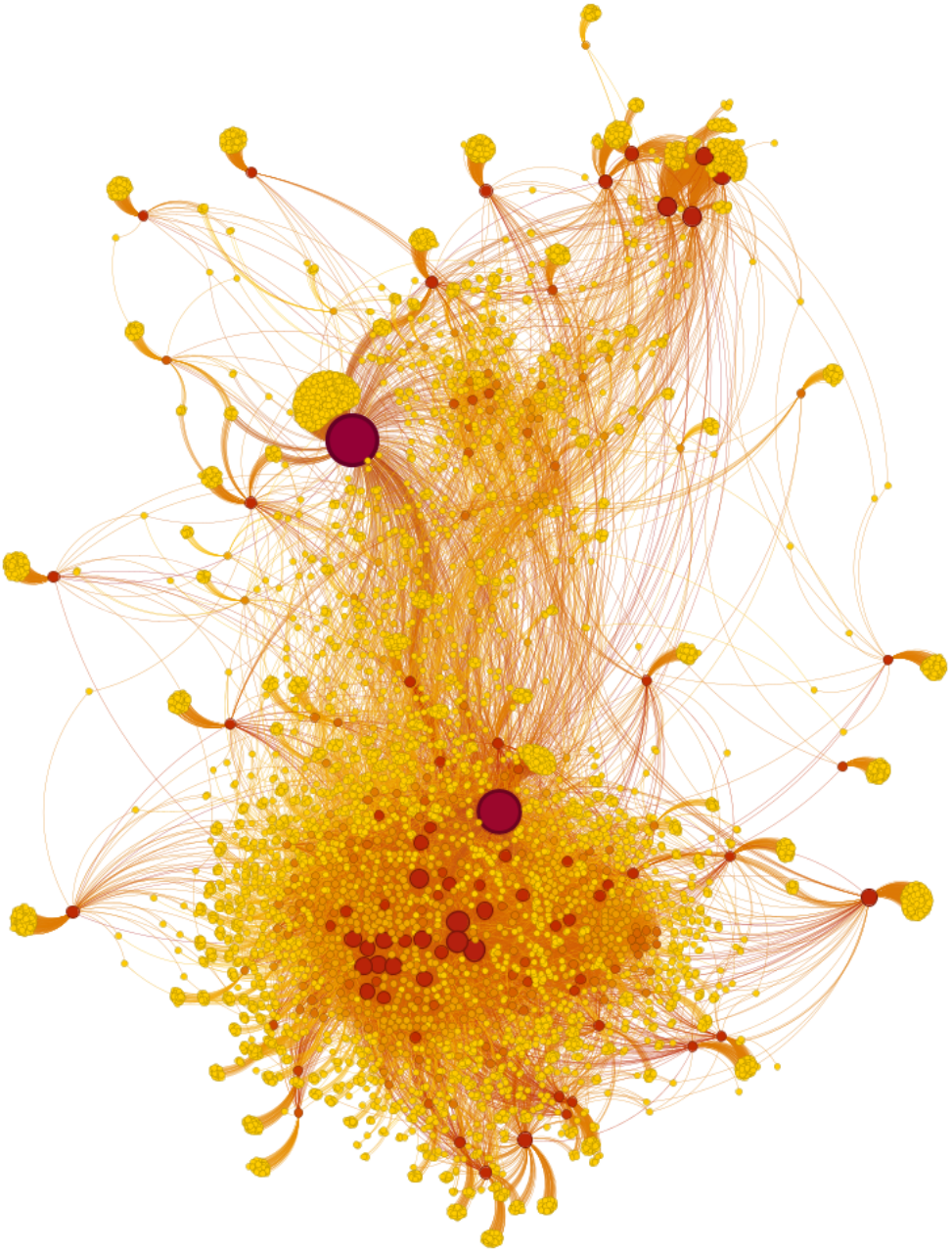

Рис. 3.6 Мережа статей вікіпедії отримана з категорії ссавці з глибиною пошуку - 3.

3.3.3 Властивість “світ малий”

Перевіримо чи виконується умова для властивості світ (2.26). Використовуючи Gephi отримали, що середній ступінь цієї мережі $\langle k \rangle = 3.119$, а середня довжина $\langle d \rangle = 4.006$. Отже враховуючи що кількість вершин 8052

підставивши ці значення в (2.26) маємо: $\frac{ln\,N}{ln\,\langle k\rangle} = 7,906$ а отже $\langle d\rangle < \frac{ln\,N}{ln\,\langle k\rangle}$. Умова властивості виконується.

3.3.4 Аналіз результатів

Отже цих даних достатньо для того аби зробити висновок що перед нами саме безмасштабна мережа.

Оскільки безмасштабні мережі характеризуються кількома хабами з високим ступенем, можна розглянути залучення таких головних статей наприклад в рекламній діяльності. Ці статті можуть бути важливими та популярними темами, які привертають багато уваги користувачів.

Безмасштабні мережі можуть бути використані для покращення алгоритмів рекомендацій. З урахуванням структури мережі та зв'язків між статтями, можна розглянути розробку алгоритмів, які забезпечують кращу рекомендацію статей користувачам, враховуючи їхні інтереси та поведінку у мережі.

Безмасштабні мережі мають ще одну дуже корисну властивість: вони часто проявляють явище "прив'язаності до нових вершин", тобто нові статті мають більшу ймовірність отримати зв'язки зі статтями високого ступеня.

ВИСНОВКИ

Дослідження українського сегменту Вікіпедії засобами аналізу складних мережевих структур виявило,що ц ей сегмент є складною мережею з багатою взаємодією між статтями. Використання теорії графів дозволило отримати глибоке розуміння структури та характеристик українського сегменту Вікіпедії.

У першому розділі було сформульовано задачу дослідження українського сегменту Вікіпедії та описано доступні дані. Також була обґрунтована актуальність роботи, вказуючи на важливість розуміння структури та характеристик цієї мережі.

У другому розділі роботи було розглянуто загальні поняття теорії графів, включаючи мережі та графи, орієнтовані та неорієнтовані графи, ступінь та середній ступінь вершин, розподіл ступенів, матрицю суміжності, зважені графи, шляхи та відстані, зв'язність та коефіцієнт кластеризації, тощо. Також було досліджено властивості випадкових мереж та безмасштабних мереж, зокрема степеневий закон розподілу ступенів, наявність хабів, властивість "світу малого" та універсальність безмасштабних мереж.

У третьому розділі було реалізовано поставлену задачу та проведено аналіз отриманих результатів. Було описано інструменти для реалізації задачі, збір даних та формування графу. Також було проведено аналіз та візуалізацію мережі, включаючи визначення розподілу вершин, наявність хабів та властивість "світу малого". Результати аналізу свідчать про відповідність українського сегменту Вікіпедії основним принципам безмасштабних мереж.

Це дослідження є важливим з практичної та наукової точок зору. Воно допомагає зрозуміти структуру та динаміку цієї мережі, що має значення для розробки ефективних алгоритмів рекомендацій статей та покращення якості інформації. Таким чином, висновки роботи підтверджують важливість та актуальність проведеного дослідження українського сегменту Вікіпедії.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Barabasi, A.L., Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.
2. Random graphs - Erdos and Renyi (1959)
3. Albert, A and Barabási, AL (2002) ‘Statistical mechanics of complex networks’, Reviews of Modern Physics, vol.74, pp. 47–97.
4. Barabási, A.-L. (2016). Network Science. Cambridge University Press
5. Newman, M. E. J. (2010). Networks: An Introduction. Oxford University Press
6. NetworkX Documentation: https://networkx.org/documentation/stable/
7. Gephi Documentation: https://gephi.org/users/documentation/
8. Wikipedia API Documentation: https://www.mediawiki.org/wiki/API:Main_page

ДОДАТКИ

Додаток А. Код програми для побудови мережі

```
# Імпорт необхідних модулів
import wikipediaapi
import networkx as nx
import json

# Перевірка чи існує обрана категорія
wiki_wiki = wikipediaapi.Wikipedia('uk')

category_name = 'Ссавці'
page_py = wiki_wiki.page('Category:' + category_name)

if (page_py.exists()):
    print("Обрана категорія існує")
else:
    print("Обраної категорії не існує")

def print_links(page):
    links = page.links
    for title in sorted(links.keys()):
        print("%s: %s" % (title, links[title]))

print_links(page_py)

# Збір всіх статтей з первним рівнем глибини
search_depth = 3
category_members = []

def log_categorymembers(categorymembers, level=0):
    for c in categorymembers.values():
        if c.ns == 0:
            entry = {'title':c.title,'level':level}
            category_members.append(entry)
        if c.ns == 14 and level < search_depth:
            log_categorymembers (c.categorymembers, level + 1)

cat = wiki_wiki.page("Category:"+category_name)
log_categorymembers(cat.categorymembers)

json_path = 'category_members.json'
with open(json_path, 'w', encoding='utf-8') as outfile:
    json.dump(category_members, outfile, ensure_ascii=False)
```

```
# Збір посилань з кожної збереженої сторінки
with open(json_path, 'r', encoding='utf-8') as jsonfile:
    category_members = json.load(jsonfile)
    jsonfile.close()

seen = []
network = {}

print("Збираємо всі посилання з "+str(len(category_members))+" сторінок вікіпедії. Це
може зайняти певний час...")
print("")

count=1

for each in category_members:
    title=each["title"]
    if count % 50 == 0:
        print("Посилання зібрано з "+str(count)+" сторінок. Загальна кількість
"+str(len(category_members))+". Продовжуємо збирати...")
    if not title in seen:
        seen.append(title)
        try:
            page=wiki_wiki.page(title)
            text_links = []
            links = page.links
            for link_title in sorted(links.keys()):
                text_links.append(link_title)
            network.update({title:text_links})

        except:
            print('Пропущено: '+title)
            print("")
    count=count+1

print("Посилання з всіх сторінок успішно зібрано")

# Формування мережі
all_edges = []
members = network.keys()

print("Обчислення зв'язків...")
print("")
for source in network:
    for target in network[source]:
        edge = (source,target)
```

```
        all_edges.append(edge)

print("Збереження мереж...")
print("")

new_cat_members={}
for each in category_members:
    new_cat_members[each["title"]]={"level":each["level"]}

G = nx.DiGraph()
G.add_edges_from(all_edges)
for each in G:
    if each in members:
        G.nodes[each]['member_level'] = 'Level
'+str(new_cat_members[each]["level"])
    else:
        G.nodes[each]['member_level'] = 'Not a member'
nx.write_gexf(G, 'network_allpages.gexf')

print("Мережі успішно збережено. Ви зможете знайти їх за наступними шляхами: ")
print("")

locale=!cd
print(locale[0]+"\\"+'network_allpages.gexf')
```

Додаток Б. Код для побудови графіків розподілу вершин в логарифмічному представленні

```
import matplotlib.pyplot as plt
import numpy as np

G = nx.read_gexf('./network_allpages.gexf')

# Отримуємо ступені входу та виходу вершин
in_degrees = [degree for (_, degree) in G.in_degree()]
out_degrees = [degree for (_, degree) in G.out_degree()]

# Розраховуємо ймовірність для ступенів входу
in_degree_prob = np.bincount(in_degrees) / len(in_degrees)
in_degree_prob = in_degree_prob[in_degree_prob != 0]
log_in_degrees = np.log(range(1, len(in_degree_prob) + 1))
log_in_probs = np.log(in_degree_prob)

# Розраховуємо ймовірність для ступенів виходу
```

```
out_degree_prob = np.bincount(out_degrees) / len(out_degrees)
out_degree_prob = out_degree_prob[out_degree_prob != 0]
log_out_degrees = np.log(range(1, len(out_degree_prob) + 1))
log_out_probs = np.log(out_degree_prob)

# Виконуємо апроксимацію поліномом степеня 1 для ступенів входу
in_coefficients = np.polyfit(log_in_degrees, log_in_probs, deg=1)
in_poly = np.poly1d(in_coefficients)

# Виконуємо апроксимацію поліномом степеня 1 для ступенів виходу
out_coefficients = np.polyfit(log_out_degrees, log_out_probs, deg=1)
out_poly = np.poly1d(out_coefficients)

# Створюємо графіки
fig, (ax1, ax2) = plt.subplots(2, 1, sharex=True)

# Графік для ступенів входу
ax1.loglog(range(1, len(in_degree_prob) + 1), in_degree_prob, 'bo', markersize=5,
label='Дані')
ax1.loglog(np.exp(log_in_degrees), np.exp(in_poly(log_in_degrees)), 'r--',
label='Апроксимація')
ax1.set_title('Розподіл ступенів входу')
ax1.set_ylabel('Ймовірність')
ax1.legend()

# Графік для ступенів виходу
ax2.loglog(range(1, len(out_degree_prob) + 1), out_degree_prob, 'bo',
markersize=5, label='Дані')
ax2.loglog(np.exp(log_out_degrees), np.exp(out_poly(log_out_degrees)), 'r--',
label='Апроксимація')
ax2.set_title('Розподіл ступенів виходу')
ax2.set_xlabel('Ступінь вершини')
ax2.set_ylabel('Ймовірність')
ax2.legend()

plt.tight_layout()
plt.show()
```