Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра прикладної математики
(повна назва)

# АТЕСТАЦІЙНА РОБОТА
# Пояснювальна записка

рівень вищої освіти другий (магістерський)

Методи аналізу та пошуку лідерів
в соціальних мережах
(тема)

Виконав:
студент _2_ курсу, групи _САУм-18-1
Закутній С.В.
(прізвище, ініціали)
Спеціальність
124 Системний аналіз
(код і повна назва спеціальності)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма
Системний аналіз і управління
(повна назва освітньої програми)
Керівник проф. Кіріченко Л.О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ПМ _____ Тевяшев А.Д.
(підпис) (прізвище, ініціали)

2019 р.

Харківський національний університет радіоелектроніки

Факультет _інформаційно-аналітичних технологій та менеджменту
Кафедра _прикладної математики
Рівень вищої освіти _другий (магістерський)
Спеціальність _124 Системний аналіз
(код і повна назва)
Тип програми _освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма _Системний аналіз і управління
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри ПМ _____
(підпис)
"____" _____ 2019 р.

# ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові _Закутньому Сергію Валерійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи _Методи аналізу та пошуку лідерів в соціальних мережах

затверджена наказом по університету від 31 жовтня 2019 р. № 1601 Ст

2. Термін подання студентом роботи до екзаменаційної комісії _9_ грудня 2019 р.

3. Вихідні дані до роботи _дані соціальної мережі

4. Перелік питань, що потрібно опрацювати в роботі
1. Системний аналіз проблеми пошуку лідерів у соціальних мережах
2. Вибір і обґрунтування методу розв'язання
3. Програмна реалізація
4. Результати обчислювального експерименту
5. Аналіз можливих застосувань

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій

1. Актуальність теми роботи
2. Постановка задачі
3. Системний аналіз проблеми
4. Метод чисельного аналізу
5. Результати обчислювального експерименту

## КАЛЕНДАРНИЙ ПЛАН

| № | Назва етапів роботи | Терміни виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Підбір та вивчення технічної літератури за темою роботи | вересень 2019 р. | виконано |
| 2 | Вибір та обґрунтування методу | жовтень – листопад 2019 р. | виконано |
| 3 | Розробка алгоритму і програми | листопад – грудень 2019 р. | виконано |
| 4 | Проведення аналітичних досліджень та розрахунків | листопад – грудень 2019 р. | виконано |
| 5 | Робота над текстом пояснювальної записки | грудень 2019 р. | виконано |
| 6 | Представлення роботи на рецензію в ЕК | грудень 2019 р. | виконано |

Дата видачі завдання _2_ вересня 2019 р.

Студент _____
(підпис)

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

# РЕФЕРАТ

Пояснювальна записка: 93 с., 22 рис., 15 табл., 1 додаток, 17 джерел.

АНАЛІЗ СОЦІАЛЬНИХ МЕРЕЖ, ТЕОРІЯ ГРАФІВ, МЕТРИКА ЦЕНТРАЛЬНОСТІ, ПОШУК ЛІДЕРІВ, NETWORKX, PYTHON, КОЕФІЦІЄНТ ЦЕНТРАЛЬНОСТІ, МАТРИЦЯ СУМІЖНОСТІ.

Об’єкт дослідження – соціальні мережі.

Мета роботи – пошук лідерів соціальної мережі за допомогою методів теорії графів.

Метод дослідження – визначення коефіцієнтів значимості вершин за допомогою метрик центральності.

У роботі проведений системний аналіз проблеми пошуку лідерів у соціальних мережах. За допомогою мови програмування Python та бібліотеки NetworkX на основі визначення коефіцієнтів метрик центральності був створений програмний продукт та проведений порівняльний аналіз досліджуваних метрик.

## ABSTRACT

Introductory note: 93 pages, 15 tables, 22 figures, 1 appendix, 17 sources.

SOCIAL NETWORK ANALYSIS, GRAPH THEORY, CENTRALITY MEASURE, SEARCH OF LEADERS, NETWORKX, PYTHON, IMPORTANCE SCORE, ADJACENCY MATRIX.

Object of research – the social networks.

Purpose of work – search for social network leaders by methods of graph theory.

Methods of research – defining importance score of edges by applying centrality measures.

Work contains system analysis of the problem of searching leaders in social networks. Program for computing importance score for each measure and analysis for comparing of measures that were researched was created using Python programming language and NetworkX framework.

# ЗМІСТ

# ВСТУП

Початок XXI століття можна сміливо назвати часом інформаційних технологій, тому що важко згадати галузь, яка так стрімко набрала обороти у відносно короткі проміжки часу. Завдяки цьому розвитку кожна людина тепер має змогу дізнатися майже про все, що її цікавить, а також володіє можливістю зв'язатися з будь-ким на цій планеті.

Людина завжди прагне розширити свій кругозір, знайти нових знайомих та друзів, з якими можна було б поділитися своїми враженнями чи спогадами, висловити свою точку зору щодо різних проблем та вислухати коментарі інших. Саме тому соціальні мережі набрали шалену популярність у всьому світі та зараз важко найти людину, яка б взагалі нічого не чула про них.

Якщо ми хочемо отримати максимальну користь від наданої мережами нам інформації, потрібно вміти її аналізувати та застосовувати у конкретних задачах. Такими задачами можуть бути різноманітні соціальні дослідження, статистика по конкретним групам людей, пошук неявних зв'язків чи інтересів між особами на різних континентах і т.п.

При дослідженні соціального графу або мережі одними з основних питань для дослідника є питання про те, наскільки важлива кожна вершина для графу, як вона взаємодіє з іншими, як швидко проходять новини чи повідомлення через мережу, яке значення вона відіграє для мережі в цілому та наскільки зміниться її роль при видаленні або навпаки, при додаванні до неї інших зв'язків.

У даній роботі була вирішена проблема пошуку лідерів у соціальних мережах з використанням теорії графів та метрик центральності. Соціальна мережа представляється у вигляді графа з вершин (користувачів) та ребер (зв'язків) між ними.

Розглянута проблема є актуальною через те, що аналіз соціальних мереж зараз став ключовим методом досліджень у сучасній соціології, а також за допомогою нього можна вирішити безліч проблем у різних галузях. Наприклад, задача пошуку лідерів розглядається в маркетингових задачах, коли необхідно в

цілях просування продукту чи інновації розповсюдити інформацію серед користувачів, або в спортивній діяльності, коли необхідно визначити взаємодію партнерів у команді. Навіть у воєнній галузі за допомогою аналізу соціальних мереж може бути виявлена потенційно небезпечна група людей, що може займатися протиправною діяльністю.

Метою даної роботи буде розробка та програмна реалізація пошуку лідерів у соціальних мережах на основі методів теорії графів.

Задача пошуку лідерів є однією з основних у дослідженні соціальних мереж і є тісно пов’язаною з задачею про пошук кластерів у графі, тобто якихось спільнот, які схожі за загальними ознаками.

Основною задачею дослідження буде розпізнавання лідерів у доступних соціальних графах, введення нової метрики для поєднання різних характеристик та врахування значимості кожної метрики для розглядуваної задачі та перевірка її роботи на достовірність результатів у реальних мережах.

# 1 СИСТЕМНИЙ АНАЛІЗ ПРОБЛЕМИ ПОШУКУ ЛІДЕРІВ У СОЦІАЛЬНІЙ МЕРЕЖІ

## 1.1 Системний аналіз проблеми пошуку лідерів соціальної мережі та постановка задачі дослідження

### 1.1.1 Вербальна модель системи

Об’єктом аналізу є соціальна мережа, яку можна представити у вигляді графу. За допомогою аналізу зв’язків у такій мережі та метрик центральності можна визначити основних користувачів або лідерів, тобто тих людей, які користуються особливою популярністю серед інших.

Задача дослідження соціальних мереж останнім часом викликала великий інтерес серед науковців, тому що на основі таких досліджень можна значно полегшити собі життя, коли ми хочемо, наприклад, розповсюдити інформацію, або навпаки, зібрати інформацію про комунікабельність тієї чи іншої групи осіб.

Дані можуть бути отримані декількома способами: з якого-небудь електронного ресурсу за допомогою спеціальних програмних інструментів або з використанням даних дослідницьких фірм у відкритому доступі. У даній роботі дані будуть отримані саме першим способом.

Дослідження в цій області вимагають уважності та правильного аналізу отриманих результатів.

Соціальна мережа представлена у вигляді графа з вершин (користувачів) та ребер (зв’язків) між ними. В залежності від вибору метрики центральності можна отримати різні результати про властивості цієї мережі.

Аналіз соціальних мереж широко використовується у ряді додатків та дисциплін. Деякі розповсюджені додатки аналізу мереж включають в себе збір та накопичення даних, моделювання розповсюдження мережі, моделювання мережі та вибірок, аналіз характерних ознак та поведінки користувача. Галузей, де

може застосовуватися обрана проблема безліч. Так, наприклад, у спортивній діяльності можна відстежувати наскільки гравець добре взаємодіє з іншими партнерами по команді, у воєнній галузі в процесі аналізу мереж виконується пошук у глибину на три вузли для пошуку лідера мережі для нанесення ударів по захвату або знищенню найбільш значимих цілей, що приводить до порушення функціонування мережі. Це можна також легко перевірити, досліджуючи гігантську компоненту графа. Коли видаляються перші три вершини, що мають найбільше зв'язків з іншими, то гігантська компонента ще зберігає свою структуру, але якщо видалити четверту чи п'яту, вона почне розпадатися на зовсім невеликі групи, що звичайно, порушить її значимість для мережі в цілому.

### 1.1.2 Морфологічний опис системи

Модель «чорний ящик» – модель досліджуваної системи, що зосереджена на дослідженні реакції системи, як цілого, на зміни зовнішнього середовища (рис. 1.1). Система є максимально простою і відображає входи та виходи досліджуваного явища. Цей метод дослідження системи найбільш підходить при виявленні реакції системи на її входи. І хоча цей «ящик» є відокремленим, але він не є повністю ізольованим. Система пов'язана з навколишнім середовищем та за допомогою цих зв'язків впливає на середовище. Зв'язки, що направлені від системи до середовища називають виходами системи, а навпаки, із середовища до системи – входами.

Поняття «чорний ящик» було запропоновано У. Р. Ешбі. У кібернетиці воно дозволяє вивчати поведінку систем, тобто їх реакцій на різноманітні зовнішні впливи, і в той же час абстрагуватися від їх внутрішнього устрою. Таким чином, система вивчається не як сукупність взаємопов'язаних елементів, а як ціле, яке взаємодіє з середовищем на своїх входах і виходах [1].

Дослідження за допомогою метода «чорного ящика» полягає в тому, що відбувається попереднє дослідження за взаємодією системи з навколишнім се-

редовищем та встановлюються усі вхідні та вихідні параметри, серед яких виділяють найбільш суттєві джерела впливу. Потім відбувається вибір входів та виходів для дослідження з врахуванням необхідних засобів впливу на систему та засобів контролю та нагляду за поведінкою системи.

На наступному етапі проводяться впливи на вхід системи та реєстрація її виходів. У процесі вивчення дослідник та «чорний ящик» утворюють систему зі зворотним зв’язком, а первинні результати дослідження – множина пар станів входу й виходу, аналіз яких дозволяє встановити між ними причинно-наслідковий зв'язок.


Рисунок 1.1 – Модель «чорний ящик»

### 1.1.3 Функціональна модель системи

Функціональна модель системи може бути представлена графічно за допомогою контекстної діаграми IDEF0. Спочатку проводиться загальний опис системи, після чого відбувається функціональна декомпозиція. В результаті застосування IDEF0 до поточної системи отримаємо модель цієї системи, що складається з ієрархічно впорядкованої множини діаграм. На рисунку 1.2 зображена загальна функціональна модель системи із зовнішніми зв’язками.

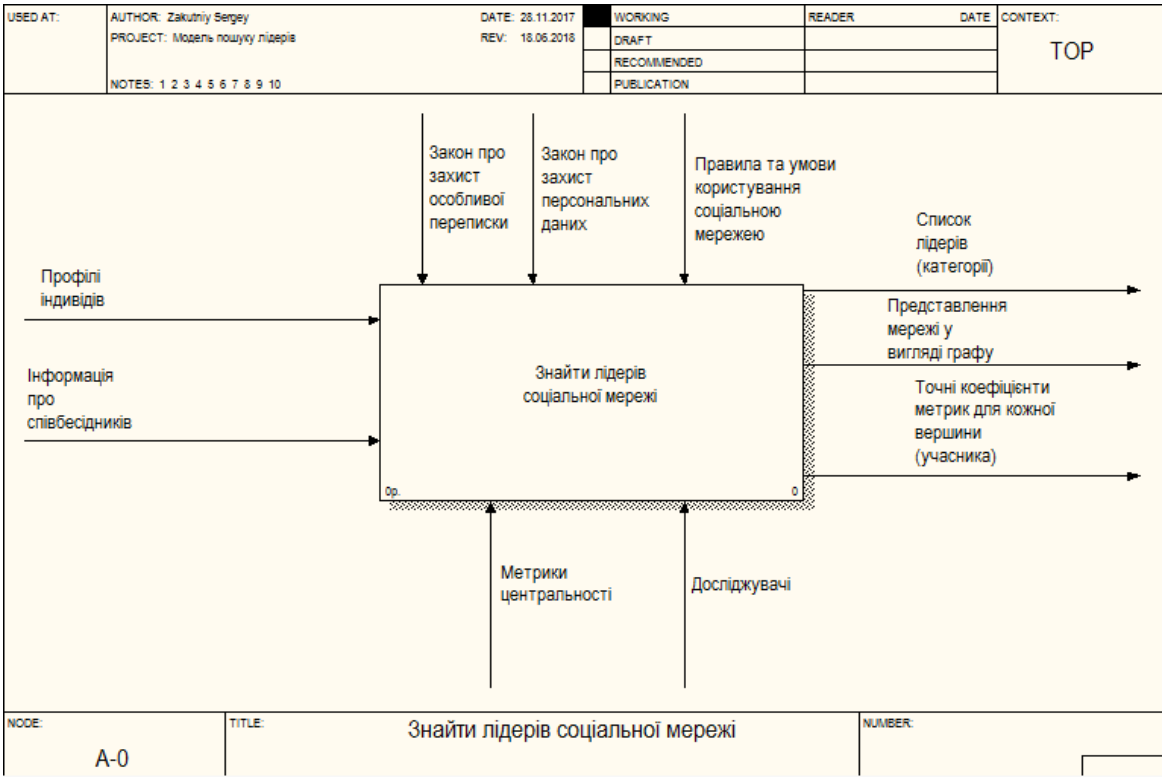


Рисунок 1.2 – Контекстна IDEF0 діаграма

Рисунок 1.3 являє собою декомпозицію минулої діаграми на 3 функціональні блоки, між якими встановлені послідовні прямі зв’язки.

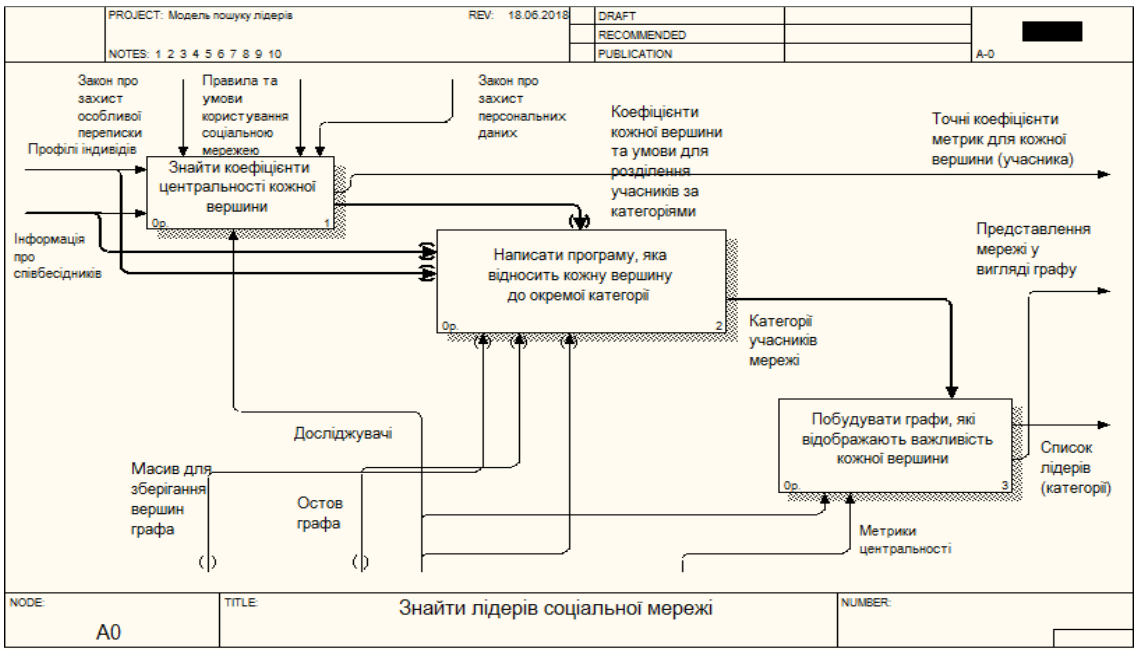


Рисунок 1.3 – Декомпозиція першого рівня функціонального блоку «Знаходження лідерів соціальної мережі»

### 1.1.4 Інформаційна модель системи

Діаграми потоків даних (Data Flow Diagrams – DFD) – методологія графічного структурного аналізу, що описує зовнішні по відношенню до системи джерела і адресати даних, логічні функції, потоки даних і сховища даних, до яких здійснюється доступ.

Необхідність використання DFD-діаграм полягає в потребі описати існуючі в структурі організації потоків даних. Діаграми потоків даних містять елементи двох видів: чотирикутники – описують функції (роботи, процеси) і стрілки які описують інформаційні потоки між цими функціями [2].

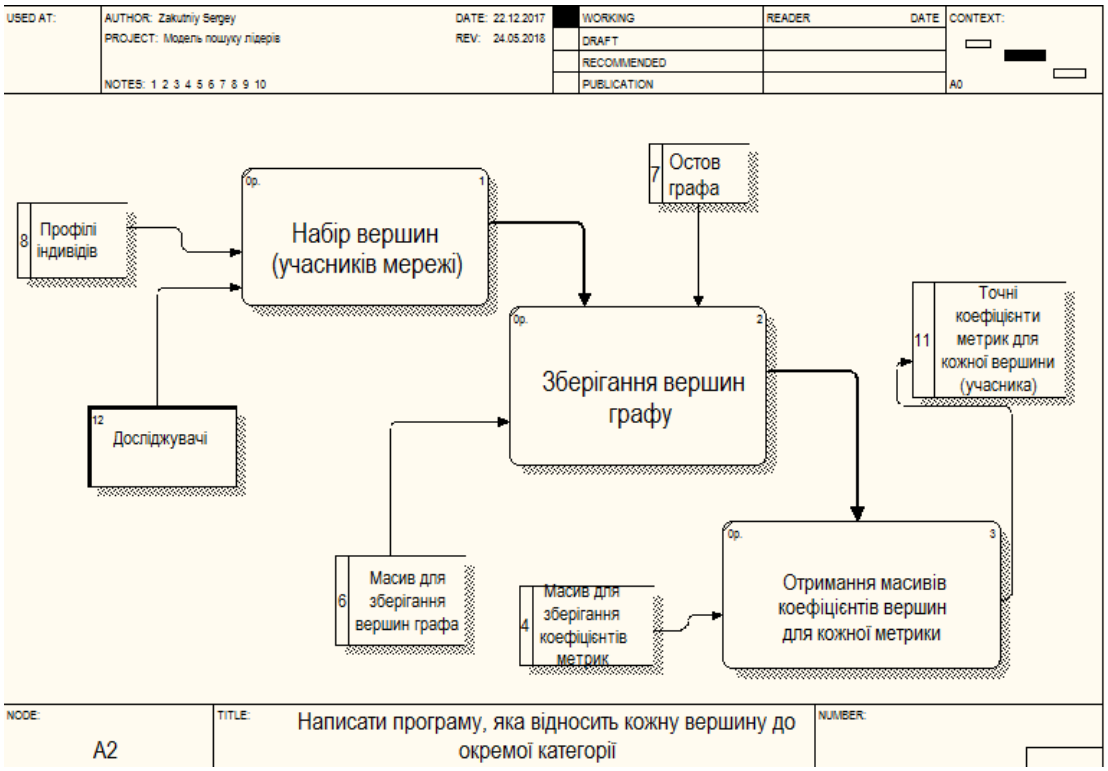


Рисунок 1.4 – Контекстна DFD діаграма

Наявність в діаграмах DFD елементів для опису джерел, приймачів і сховищ даних дозволяє більш ефективно і наглядно описати процес документообігу. Однак для опису логіки взаємодії інформаційних потоків більш підходить стандарт опису процесів IDEF3, який називається також workflow diagramming – методологією моделювання, що використовує графічний опис інформаційних

потоків, взаємин між процесами обробки інформації та об'єктів, що є частиною цих процесів. Діаграми workflow можуть бути використані в моделюванні бізнес-процесів для аналізу завершеності процедур обробки інформації. За допомогою них можна описати сценарії дій співробітників організації. IDEF3 – це метод, основна мета якого описати ситуацію, коли процеси виконуються в певній послідовності, а також описати об'єкти, які беруть участь спільно в одному процесі.

Для більш детального моделювання опишемо послідовність операцій у невизначених процесах, за допомогою діаграми IDEF3 (рис. 1.5).

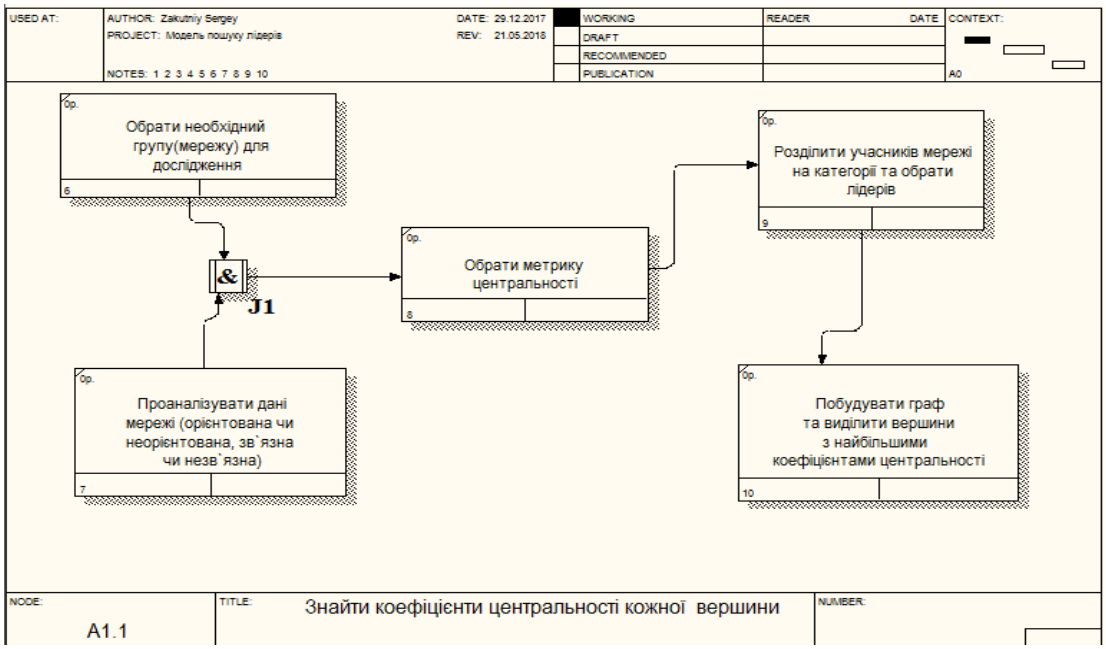


Рисунок 1.5 – IDEF3 діаграма

## 1.2 Аналіз сценаріїв вирішення проблеми пошуку лідерів у соціальних мережах

### 1.2.1 Модель аналізу проблеми

Об’єкт дослідження – модель порівняльного аналізу метрик для визначення лідерів мережі.

Мета дослідження – дослідити запропоновані метрики пошуку лідера та визначити оптимальну.

Для проведення дослідження необхідно обрати критерії, які впливають на результат найбільше. Такими критеріями є:

– коректність (К1);

– швидкодія (К2);

– складність реалізації (К3);

– сучасність (К4).

Необхідно розглянути всі ці критерії та більш детально їх описати. У якості альтернатив будуть обрані метрики, за якими розраховується коефіцієнт центральності кожної вершини.

Порівнюючи коректність, ми маємо на увазі наскільки отримані коефіцієнти кожної з метрик будуть відображати реальне положення вершини у графі, бо наприклад, кількість соціальних контактів не може відображати їх якість, а просто показувати ступінь активності індивіда.

Критерій швидкодії буде порівнювати проміжки часу для розрахунку програмою коефіцієнтів центральності вершин кожної з метрик.

Порівнюючи складність реалізації визначається можливість виконання розрахунків коефіцієнтів метрики для будь-якого користувача самостійно з базовими знаннями теорії графів.

Сучасність показує наскільки дана метрика актуальна в даний момент часу та її застосовність при дослідженні соціальних мереж.

Будуть розглянуті такі альтернативи:

– степенева метрика (degree centrality) (А1);

– метрика близькості (closeness centrality) (А2);

– проміжна метрика центральності (betweenness centrality) (А3)

– центральність за власним вектором (eigenvector centrality) (А4).

Проведемо аналіз альтернатив та виділимо основні переваги та недоліки метрик центральності за кожним критерієм.

Степенева або показникова метрика є історично першою та вважається

найпростішою у реалізації. Вона відображає відношення кількості зв’язків однієї вершини до загальної кількості зв’язків між вершинами. Значним недоліком цієї метрики є те, що кількість соціальних контактів, часто відображує не їхню якість, а ступінь комунікабельності індивіда.

Метрика близькості показує відношення числа вузлів графа до суми відстаней між конкретним вузлом та усіма іншими. Основним недоліком є неможливість розрахунку такого коефіцієнту для незв`язних графів.

Проміжна метрика центральності аналізує не тільки зв’язки однієї вершини, а й сусідніх до неї, а отже буде давати більш правдоподібні результати. Вона показує скільки найкоротших шляхів між всіма вузлами мережі проходить через один конкретний вузол.

І нарешті Eigenvector centrality демонструє залежність між центральностями вузла та центральностями її сусідів. Якщо учасник має багато зв’язків з іншими, у яких також багато зв’язків, то це свідчить про те, що він має високу центральність за власним вектором. На рисунку 1.6 зображена ієрархічна модель процесу аналізу метрик.

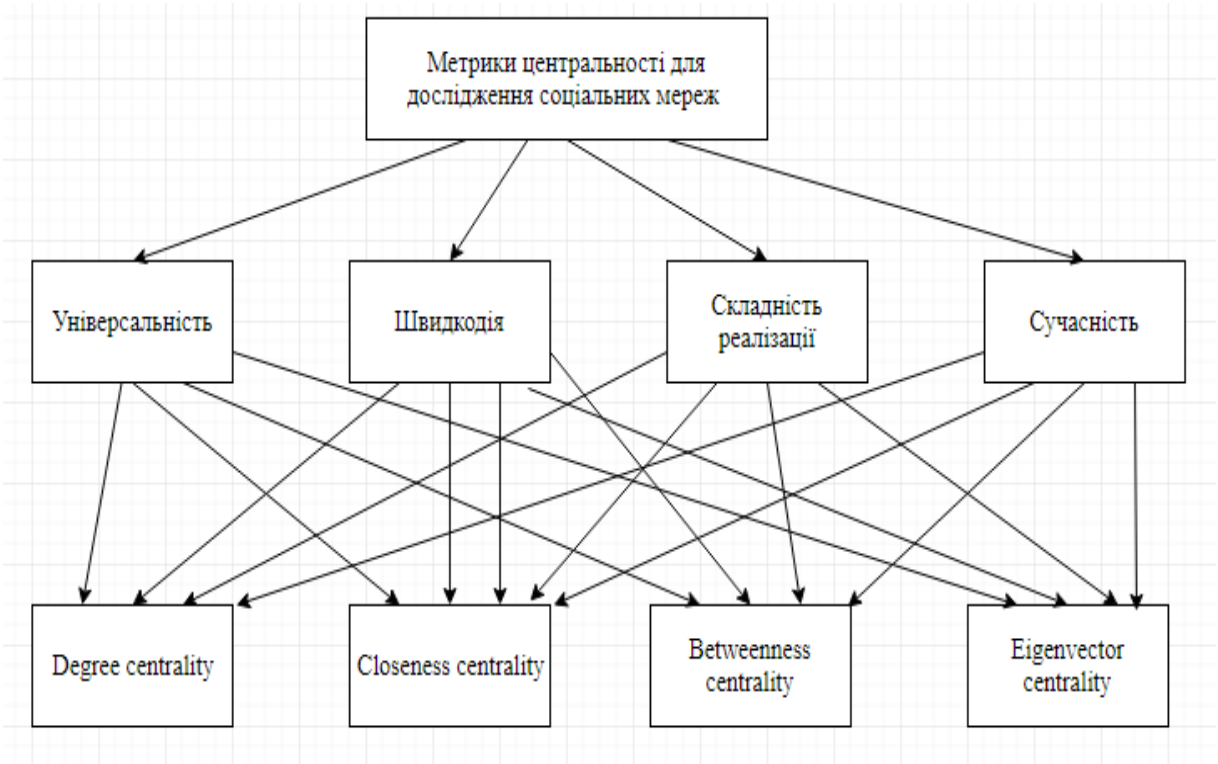


Рисунок 1.6 – Ієрархічна модель процесу аналізу метрик

### 1.2.2 Оцінювання вектора пріоритетів незадоволеностей методом аналізу ієрархій

За допомогою методу попарних порівнянь побудуємо модель процесу аналізу метрик. Аналіз метрик включає в себе такі рівні:

– нульовий рівень – компоненти проблеми;

– перший рівень – класифікація метрик;

– другий рівень – характеристики компонентів, які впливають на результат поставленої задачі.

Нульовий рівень проблеми представлений на рисунку 1.6.

На першому рівні аналізу проблеми побудуємо матрицю попарних порівнянь критеріїв. Результати наведені в таблиці 1.1. Для встановлення відносної важливості критеріїв використана шкала відношень Сааті.

Таблиця 1.1 – Матриця попарних порівнянь критеріїв

| | К1 | К2 | К3 | К4 | Вектор пріоритетів |
|---|---|---|---|---|---|
| К1 | 1 | 4 | 1/3 | 5 | 0,27981 |
| К2 | 1/4 | 1 | 1/5 | 3 | 0,10837 |
| К3 | 3 | 5 | 1 | 7 | 0,55742 |
| К4 | 1/5 | 1/3 | 1/7 | 1 | 0,05439 |

Для знаходження індексу узгодженості знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1 + 1/4 + 3 + 1/5 = 4{,}45,$$

$$y_2 = 4 + 1 + 5 + 1/3 = 10{,}333,$$

$$y_3 = 1/3 + 1/5 + 1 + 1/7 = 1{,}67619,$$

$$y_4 = 5 + 3 + 7 + 1 = 16.$$

Тоді

$$\lambda_{\max} \approx 4{,}1697 .$$

Індекс узгодженості дорівнює

$$CI^k = \frac{4{,}1697 - 4}{4 - 1} = 0{,}056.$$

Оскільки матриця попарних порівнянь критеріїв – це матриця четвертого порядку, то відношення узгодженості:

$$CR^k = \frac{CI^k}{0{,}9} = 0{,}062.$$

Оскільки відношення узгодженості є близьким до 0,1, то вважатимемо, що матриця попарних порівнянь критеріїв побудована правильно.

Далі формуємо матриці попарних альтернатив за кожним критерієм з метою порівняння методів між собою за кожним критерієм окремо.

Таблиця 1.2 – Матриця порівнянь за критерієм К1

| К1 | А1 | А2 | А3 | А4 | Вектор пріоритетів |
|---|---|---|---|---|---|
| А1 | 1 | 1/4 | 1/5 | 1/6 | 0,05648 |
| А2 | 4 | 1 | 1/3 | 1/3 | 0,14205 |
| А3 | 5 | 3 | 1 | 1/2 | 0,30938 |
| А4 | 6 | 4 | 2 | 1 | 0,49208 |

Для знаходження індексу узгодженості знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1+4+5+6 = 16,$$

$$y_2 = 1/4+1+3+4 = 8,25,$$

$$y_3 = 1/5+1/3+1+2 = 3,5333$$

$$y_4 = 1/6+1/4+1/2+1 = 1,91667.$$

Тоді

$$\lambda_{\max} \approx 4,11199,$$

а індекс узгодженості буде дорівнювати

$$CI_{K1}^{A} = 0,03733.$$

Оскільки матриця попарних порівнянь альтернатив – це матриця четвертого порядку, то відношення узгодженості:

$$CR_{K1}^{A} = \frac{CI^{k}}{0,9} = 0,041478.$$

Таблиця 1.3 – Матриця порівнянь за критерієм К2

| К2 | А1 | А2 | А3 | А4 | Вектор пріоритетів |
|---|---|---|---|---|---|
| А1 | 1 | 2 | 3 | 4 | 0,45305 |
| А2 | 1/2 | 1 | 3 | 4 | 0,32035 |
| А3 | 1/3 | 1/3 | 1 | 2 | 0,14053 |
| А4 | 1/4 | 1/4 | 1/2 | 1 | 0,08606 |

Для знаходження індексу узгодженості знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1 + 1/2 + 1/3 + 1/4 = 2{,}0833,$$

$$y_2 = 2 + 1 + 1/3 + 1/4 = 3{,}5833,$$

$$y_3 = 3 + 3 + 1 + 1/2 = 7{,}5$$

$$y_4 = 4 + 4 + 2 + 1 = 11.$$

Тоді

$$\lambda_{\max} \approx 4{,}09248,$$

а індекс узгодженості в цьому випадку буде рівний

$$CI_{K2}^{A} = 0{,}308.$$

Відношення узгодженості:

$$CR_{K2}^{A} = \frac{CI^{k}}{0{,}58} = 0{,}0342.$$

Таблиця 1.4 – Матриця порівнянь за критерієм К3

| К3 | А1 | А2 | А3 | А4 | Вектор пріоритетів |
|---|---|---|---|---|---|
| А1 | 1 | 3 | 6 | 8 | 0,57017 |
| А2 | 1/3 | 1 | 4 | 6 | 0,27681 |
| А3 | 1/6 | 1/4 | 1 | 5 | 0,1112 |
| А4 | 1/8 | 1/6 | 1/5 | 1 | 0,04182 |

Для знаходження індексу узгодженості знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1 + 1/3 + 1/6 + 1/8 = 1,625,$$

$$y_2 = 3 + 1 + 1/4 + 1/6 = 4,41667,$$

$$y_3 = 6 + 4 + 1 + 1/5 = 11,2$$

$$y_4 = 8 + 6 + 5 + 1 = 20.$$

Тоді

$$\lambda_{\max} \approx 4,23091,$$

а індекс узгодженості для критерію К3:

$$CI_{K3}^{A} = 0,76969.$$

Відношення узгодженості:

$$CR_{K3}^{A} = \frac{CI^{k}}{0,58} = 0,085.$$

Таблиця 1.5 – Матриця порівнянь за критерієм К4

| К4 | А1 | А2 | А3 | А4 | Вектор пріоритетів |
|---|---|---|---|---|---|
| А1 | 1 | 1/3 | 1/4 | 1/8 | 0,05140 |
| А2 | 3 | 1 | 1/2 | 1/7 | 0,10947 |
| А3 | 4 | 2 | 1 | 1/5 | 0,18096 |
| А4 | 8 | 7 | 5 | 1 | 0,65817 |

Для знаходження індексу узгодженості знаходимо суми елементів матриці за стовбцями:

$$y_1 = 1+3+4+8 = 16,$$

$$y_2 = 1/3+1+2+7 = 10{,}333,$$

$$y_3 = 1/4+1/2+1+5 = 6{,}75$$

$$y_4 = 1/8+1/7+1/5+1 = 1{,}468.$$

Тоді

$$\lambda_{\max} \approx 4{,}14119,$$

а індекс узгодженості критерію К4 матиме вигляд:

$$CI_{K4}^{A} = 0{,}047.$$

Відношення узгодженості:

$$CR_{K4}^{A} = \frac{CI^{k}}{0{,}58} = 0{,}0522.$$

Розрахуємо вектор глобальних пріоритетів альтернатив. Для цього знаходимо добуток

$$\vec{p} = \begin{bmatrix} 0{,}38549 \\ 0{,}23469 \\ 0{,}17361 \\ 0{,}20611 \end{bmatrix}.$$

Розрахуємо індекс узгодженості та відношення узгодженості для всієї ієрархії:

$$CI = 0,531969,$$

$$RI = 0,90 + 0,90 = 1,8,$$

$$CR = \frac{CI}{RI} = 0,295538,$$

що теж можна вважати доброю узгодженістю.

Найбільша компонента вектора локальних пріоритетів критеріїв відповідає першому критерію. Отже, маємо наступні пріоритети за критеріями порівняння: складність реалізації, коректність, швидкодія, сучасність.

Порівнюючи альтернативи за обраними критеріями, отримали вектор глобальних пріоритетів, найбільша компонента якого відповідає першій альтернативі, тобто степеневій метриці.

## 1.3 Змістовна та формальна постановка задачі

### 1.3.1 Змістовна постановка задачі

При дослідженні соціального графу або мережі одними з основних питань для дослідника є питання про те, наскільки важлива кожна вершина для графу, як вона взаємодіє з іншими, як швидко проходять новини чи повідомлення через мережу, яке значення вона відіграє для мережі в цілому та наскільки зміниться роль при її видаленні або навпаки, при додаванні до неї інших зв`язків.

У випадку, коли досліджується достатньо великий граф зі значною кількістю вершин, то ми можемо ввести таке поняття, як гігантська компонента графа – це компонента, що складається з $\gamma \times n$ вершин, де $n$ – загальна кількість вершин у графі, а $\gamma$ – коефіцієнт від 0 до 1, що показує, наскільки компонента графа, якщо він незв`язний, велика чи мала у розмірах по відношенню до загальної кількості вершин. У випадку повністю зв’язного графа цей коефіцієнт дорівнює 1. При видаленні вершин, що мають багато сполучень (ребер) з інши-

ми, значимість цієї компоненти втрачається і мережа стає більш розрідженою та розірваною, що призводить та втрати її стійкості. Саме тому дуже важливо знати, які вершини можна видаляти,а які не потрібно.

Пошук лідерів є однією з найважливіших задач у процесі дослідження соціальних мереж, тому що за їх допомогою ми можемо вирішити проблеми розповсюдження інформації, аналізу зв'язків у структурі організації чи команди, можна досліджувати дані різного рівня: поведінку окремих учасників мережі, різних підгруп, особливості позицій в суспільстві та властивості мережі в цілому. Також, знаючи лідерів у соціальній мережі, можна з'ясувати найвпливовіших особистостей та знаменитостей, що допоможе в рекламному бізнесі, коли необхідно знати, що розміщені оголошення буде бачити велика кількість людей.

Задача пошуку лідерів зводиться до задачі розрахунку коефіцієнтів центральності графа, а так як існує не одна можливість визначення цих коефіцієнтів, то необхідно обрати оптимальну метрику та проаналізувати, що вона дає точні результати.

### 1.3.2 Формальна постановка задачі

Задачу пошуку лідерів соціальної мережі будемо розглядати за допомогою теорії графів. Соціальний граф – це граф, вузли якого представлені у вигляді соціальних об'єктів, такими як профілі користувачів з різноманітними атрибутами (ім'я, день народження, рідне місто), а ребра – соціальними зв'язками між ними.

Тоді нехай $G = \langle V, E \rangle$ – це досліджуваний граф, де $V$ – множина вузлів (vertices), а $E$ – множина ребер (edges).

Визначимо деякі основні поняття для дослідження зв'язків та відстаней у графі.

Одним з найважливіших понять теорії графів є матриця суміжності. Вона є одним із способів представлення графу у вигляді матриці.

Матриця суміжності графа $G$ зі скінченним числом вершин $n$ – це квадратна матриця $A$ розміру $n$, у якій значення елементу $a_{ij}$ дорівнює числу ребер з $i$ вершини до $j$ вершини. У графі без петель та кратних ребер матриця суміжності містить нулі на головній діагоналі.

Матриця суміжності є основною структурою даних, що використовується для представлення графів у комп'ютерних програмах.

Відстанню між двома вершинами в графі називають число ребер у найкоротшому шляху між цими вершинами. Ця відстань також називається геодезичною.

Ексцентриситетом $\varepsilon(v)$ вершини $v$ називається найбільша геодезична відстань між вершиною $v$ та будь-якою іншою вершиною графа. Іншими словами – це відстань до самої віддаленої вершини від $v$:

$$\varepsilon(v) = \max_{u \in V} d(v,u),$$

де $d(v,u)$ – відстань між двома вершинами $v$ та $u$.

Радіусом графа називається мінімальний ексцентриситет серед усіх вершин графа.

$$r = \min_{v \in V} \varepsilon(v).$$

Діаметром графа називається максимальний ексцентриситет серед усіх вершин графа.

$$d = \max_{v \in V} \varepsilon(v).$$

Центральною вершиною графа називається вершина, ексцентриситет якої рівний $r$:

$$r = \varepsilon(v).$$

Для визначення коефіцієнтів центральності у степеневій метриці використовуємо формулу:

$$C_D(i) = k(i) = \sum_j A_{ij} = \sum_j A_{ji}, \tag{1.1}$$

де $A_{ij}$ – елементи матриці суміжності.

Для розрахунку коефіцієнтів за метрикою близькості використовується формула:

$$C_C(i) = \frac{1}{\sum_j d(i,j)}, \tag{1.2}$$

де $d(i,j)$ – відстань між вершинами $i$ та $j$.

Коефіцієнти для розрахунку центральності вершин у проміжній метриці виконуються за формулою:

$$C_B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}, \tag{1.3}$$

де $\sigma_{st}(i)$ – кількість усіх найкоротших шляхів від вершини $s$ до вершини $t$, що проходять через вершину $i$;

$\sigma_{st}$ – загальна кількість найкоротших шляхів від $s$ до $t$.

Визначення коефіцієнту центральності вершини за допомогою Eigenvector метрики відбувається за формулою:

$$Ax = \lambda x, \tag{1.4}$$

де $A$ – це матриця суміжності графа $G$ з власним значенням $\lambda$.

Також одним з частинних випадків метрики за власним вектором є метрика, запропонована Лео Катцем у 1953 році і носить його ім'я – Katz centrality. Розрахунок коефіцієнтів значимості кожної вершини у графі відбувається за формулою:

$$x_i = \alpha \sum_j A_{ij} x_j + \beta, \tag{1.5}$$

де $A_{ij}$ – елемент матриці суміжності графу $G$ з власним значенням $\lambda$;

$\alpha, \beta$ – додатні константи.

## 1.4 Постановка задачі дослідження

Виходячи з проведеного системного аналізу системи «Аналіз пошуку лідерів у соціальних мережах» сформулюємо наступні задачі для дослідження в рамках даної атестаційної роботи:

– дослідити метрики центральності, що використовуються для розрахунку коефіцієнтів значимості вершин графа;

– обрати інструменти програмної розробки та розробити архітектуру програми, в якій будуть реалізовані обрані метрики;

– на основі отриманих результатів про центральність кожної вершини визначити, для яких цілей підходить даний граф та як можна застосувати отриману інформацію у реальних задачах;

– на основі отриманих даних зробити висновки про проведену роботу.

# 2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ

## 2.1 Огляд метрик центральності для розрахунку коефіцієнтів значимості вершин мережі

Перший дослідницький додаток про метрики центральності було зроблено під керівництвом Бавеласа в Лабораторії групових мереж в Массачусетського університету наприкінці 1940-х років.

Ці дослідження були проведені Гарольдом Лівіттом і Сідні Смітом в 1949 році. Про них повідомили Бавелас і Барретт і вперше були детально описані Льюіттом у 1951 році. Їх звіти привели до такого висновку, що центральність була пов'язана з груповою ефективністю у вирішенні проблем, пов'язаних зі сприйняттям лідерства та особистого задоволення учасників соціальної групи.

Ці доповіді послужили поштовхом для багатьох експериментів в 1950-х і 1960-х роках. Узагальнюючи експериментальну літературу в 1968 році, Берджесс дійшов до висновку, що «Дослідження не дало послідовних і кумулятивних результатів». Проте, результати показують, що центральність має відношення до того, як групи стають більш організованими для вирішення деяких видів проблем.

При виборі метрики центральності з числа багатьох доступних, концептуальна або теоретична мета є способом, завдяки якому конкретна міра центральності відповідає уявленню про, наприклад, владу, статус або вплив. Одним з важливих аспектів цього є основна відмінність між показниками здатності вершини отримувати доступ до інших і середнім показником її потенціалу як передавач [3].

Коефіцієнти центральності дають відповідь на запитання «Що характеризує значиму або важливу для графа вершину?». Відповідь дається у термінах дійсних функцій на вершинах графу, де отримані значення розраховуються таким чином, що можливо провести їхнє ранжування з подальшим визначенням найбільш важливих вершин [4].

Слово «значимість» може інтерпретуватися у широкому спектрі визначень, в залежності від того яку метрику центральності ми обираємо. Запропонуємо дві схеми категоризації. За першою «значимість» може мати на увазі важливість передачі даних через мережу. За другою схемою можемо визначити її як рівень впливу конкретної вершини на мережу в цілому. Обидва підходи розділяють центральності у чіткі категорії. При подальшому аналізі результатів легко зробити висновок про те, що застосування центральності, яка підходить для однієї категорії може давати невірні результати при застосуванні її до другої категорії [5].

Центральність вузлів, або визначення, які ж вершини є більш «центральними» ніж інші, є ключовою проблемою аналізу мереж. Лінтон Фріман стверджував, що центральні вузли – це ті вузли, що знаходяться в гущі подій та являються координаційними центрами мережі. Для прикладу він використав мережу, що складалася з п’яти вершин (рис. 2.1). Середній вузол має аж три переваги над іншими: він має більше зв’язків, він може найшвидше дістатись до будь-якої іншої вершини, а також він контролює весь потік даних, що проходить через мережу. Базуючись на цих принципах він зміг формалізувати три різноманітні метрики – показникові, проміжну та метрику близькості [6].

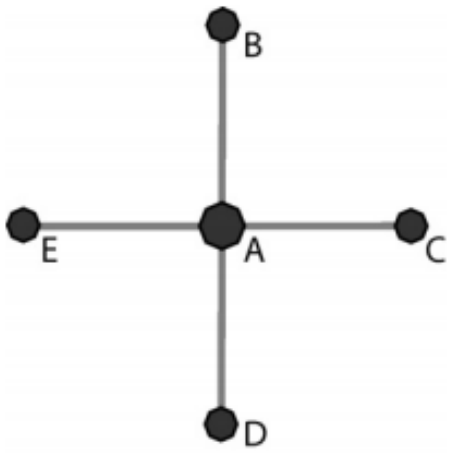


Рисунок 2.1 – Граф «зірка» з 5 вузлами та 4 ребрами

Існують чотири загальні метрики центральності – це показникова, проміж-

жна, метрика близькості та метрика за власним значенням. Інші метрики в основному є похідними від цих «базових» метрик. Спробуємо розглянути їх більш детально.

## 2.2 Показникова (степенева) метрика (degree centrality measure)

Показникова центральність є найпершою та найпростішою метрикою. Показник або степінь є кількістю ребер, що інцидентні розглядуваній вершині та є мірою участі вершини в мережі. Він є базовим індикатором та найчастіше використовуваним показником при найпершому аналізі мереж.

Основна ідея метрики полягає в тому, що вузли з великою кількістю з'єднань є більш впливовими в мережі. Іншими словами, людина з більшою кількістю друзів у соціальній мережі, більш цитована стаття в мережі наукового цитування є більш центральною і важливішою. для мережі відповідно до цієї метрики.

Ступінь центральності – це розрахунок загального числа з'єднань, пов'язаних з вершиною. Його можна розглядати як свого роду міру популярності, але грубу, яка не розпізнає різницю між кількістю і якістю. Ступінь централізації не робить різниці між посиланням на президента Сполучених Штатів і посиланням на відсів із середньої школи.

Висока вихідна центральність за ступенем вказує на те, що вузол є «владним»; це такий тип людини або сайту, який може швидко поширити інформацію серед інших людей. Висока вхідна центральність за ступенем вказує, що вузол – «знаменитість»; це означає, що за таким типом людини або сайту буде стежити багато людей. Google.com має мільярди зовнішніх посилань на інші сайти, що є показником влади у мережі. YouTube.com має відносно небагато посилань на інші сайти, однак, багато людей розміщують посилання на YouTube або вбудовують його контент на власні сторінки. Ця властивість показує популярність даного відеохостингу [7].

Для визначення коефіцієнтів центральності у степеневій метриці викори-

стовуємо формулу:

$$C_D(i) = k(i) = \sum_j A_{ij} = \sum_j A_{ji}, \tag{2.1}$$

де $A_{ij}$ – елементи матриці суміжності.

Для того, щоб звести отримані дані до приблизно одного вигляду та розглядати коефіцієнти в зрозумілому діапазоні, коефіцієнти необхідно нормалізувати, для цього розраховують нормалізований ступінь центральності $C_D^*(i)$ за наступною формулою:

$$C_D^*(i) = \frac{1}{(n-1)} * C_D(i), \tag{2.2}$$

де $n$ – загальна кількість вершин у графі.

Простота цієї метрики є її найбільшою перевагою, тому що потрібно знати тільки локальне середовище розглядуваного вузла. Проте, вона є також і обмеженою в можливостях через те, що звертаючи увагу лише на локальну структуру, метрика не бере участі у розгляді усієї мережі в цілому і це є проблемою в ситуаціях,коли потрібно дістатись до деякого іншого вузла. Так, конкретний актор може мати багато сусідів навколо себе, але якщо його сусіди будуть мати замало зв’язків з іншими, то у випадку коли буде необхідно доставити якусь інформацію на інший кінець графу, ця задача не зможе бути виконана настільки швидко, як хотілося б.

Ще однією перевагою такої метрики є те, що вона може бути застосовна як до орієнтованих, так і до неорієнтованих мереж. В такому випадку можна ввести два поняття: поняття «престижу» та поняття «впливу». Чим більший «престиж» вершини, тим більше вона має вхідних до себе ребер (in-degree), а чим більший «вплив», тим більша кількість вихідних ребер з вершини (out-degree). Таким чином, використовуючи in-degree метрику можна сказати, наскі-

льки популярним серед найближчих сусідів є актор, out-degree метрика про те, як швидко розповсюджується інформація від даної вершини.

Також її можна використовувати для дослідження мереж, в якій кожне ребро є зваженим (тобто йому відповідає деякий коефіцієнт). В такому випадку вона визначається наступним чином:

$$s_i = C_D^w(i) = \sum_j^N w_{ij},$$

де $w$ – зважена матриця суміжності, в якій $w_{ij}$ більший за нуль у випадку, якщо вершина $i$ з'єднана з $j$ та значення дорівнює вазі ребра.

## 2.3 Метрика близькості (closeness centrality measure)

Ця метрика використовується для того, щоб відстежити взаємодії акторів (учасників) між собою в мережі. Вона показує, наскільки близько знаходиться актор по відношенню до усіх інших вузлів мережі. Вузол, який є найближчим до усіх інших вузлів графа, найбільш підвладний до сприйняття нової інформації або віруса. Формально виражається як відношення числа інших вузлів до суми відстаней між конкретним вузлом та усіма іншими. Для того, щоб мати високу ступінь даного виду центральності, необхідно не тільки самому мати достатню кількість зв'язків, а і щоб у сусідів чи друзів цих зв'язків було достатньо.

Якщо мова йде про поширення даних і виявлення інформаційних потоків в організації, а дослідник зацікавлений в пошуку акторів, які можуть найбільш ефективно приймати і передавати їх, то найбільше підходить метрика центральності по близькості, оскільки для отримання інформації потрібно бути поруч з іншими. В цьому випадку актори, які мають в середньому більш коротку дистанцію до інших учасників мережі, можуть найбільш ефективно передавати і отримувати інформацію, а отже, мати більш високий коефіцієнт центральності.

Для розрахунку коефіцієнтів за метрикою близькості використовується формула:

$$C_C(i) = \frac{1}{\sum_j d(i,j)}, \tag{2.3}$$

де $d(i,j)$ – відстань між вершинами $i$ та $j$.

Нормалізоване значення розраховується за формулою:

$$C_C^*(i) = (n-1) * C_C(i). \tag{2.4}$$

Актор, що має високий нормалізований коефіцієнт центральності за метрикою близькості має найкоротші шляхи для комунікації з іншими акторами, а отже, за допомогою нього можна найшвидше дістатись до будь-якої вершини у графі. Це значить, що ця метрика може продемонструвати, наскільки конкретний учасник мережі підходить у цілях швидкого розповсюдження інформації. Також за допомогою цієї метрики можна дізнатись, наскільки компактною є дана мережа, наскільки добре вона відображає її загальну структуру.

Недоліком є те, що вона застосовна лише для зв’язних графів, тому що неможливо визначити відстань від однієї вершини до іншої, якщо компоненти не з’єднані між собою. У такому випадку їхня відстань рівна нескінченності. Для вирішення цієї проблеми використовують Harmonic centrality (гармонічну центральність), при отриманні результатів за цією метрикою такі зв’язки не враховуються і не додаються до загальної суми.

Розрахунок коефіцієнтів для гармонічної центральності здійснюється за формулою:

$$h_i = \sum_{i \neq j} \frac{1}{d(i,j)}. \tag{2.5}$$

Також може бути нормалізована за допомогою множення $h_i$ на $n-1$.

## 2.4 Проміжна метрика (betweenness centrality measure)

Проміжна метрика центральності базується на понятті відстаней між двома вершинами та на пошуку найкоротших шляхів від однієї вершини до іншої. Ця ідея належить Бавеласу в його першій роботі над даним предметом. Він припустив, що коли конкретна персона в групі стратегічно розташована на найкоротшому між двома іншими вершинами шляху, то ця персона знаходиться в центральній позиції. Цю саму думку висловив і Шимбел, лише іншими словами: «Припустимо, що для з'єднання вершин $i$ та $j$, необхідна проміжна вершина $k$. Вершина $k$ в такій мережі має певну відповідальність між $i$ та $j$. Тоді якщо розрахувати кількість усіх найкоротших шляхів, що проходять через вершину $k$, тоді матимемо міру напруги, якій піддається вершина $k$ на протязі всієї активності мережі.» Отже, центральна вершина контролює весь потік інформації, що проходить між двома іншими учасниками [8].

Для кожної пари вершин у зв'язному графі існує як мінімум один найкоротший шлях між ними. У випадку незв'язного графу ця відстань буде дорівнювати нулю. Проміжна метрика для кожного вузла враховує кількість найкоротших шляхів, що проходять через цей вузол.

Коефіцієнти для розрахунку центральності вершин у проміжній метриці виконуються за формулою:

$$C_B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}. \qquad (2.6)$$

де $\sigma_{st}(i)$ – кількість усіх найкоротших шляхів від вершини $s$ до вершини $t$, що проходять через вершину $i$;

$\sigma_{st}$ – загальна кількість найкоротших шляхів від $s$ до $t$.

Нормалізований коефіцієнт обчислюється за наступною формулою:

$$C_B^*(i) = \frac{2}{(n-1)*(n-2)} * C_B(i). \quad (2.7)$$

Ця метрика знайшла дуже широке застосування у теорії мереж. Наприклад, в телекомунікаційних мережах, актор з найбільшим коефіцієнт значимості за цією метрикою буде мати найбільший контроль над мережею, тому що через нього буде проходити найбільша кількість інформації. Крім аналізу соціальних мереж, також може застосовуватись у біології, дослідженні транспортних шляхів, соціології.

При аналізі футбольного матчу проміжна центральність дозволяє судити, наскільки робота з м'ячем між двома гравцями залежить від третього гравця. Гравці з високим рівнем центральності з посередництва грають ключові ролі в підтримці темпу гри.

Як правило, прикладні завдання виявлення ключових гравців пов'язані з їх подальшим використанням або нейтралізацією. Одним із завдань соціально-освітніх проектів є збільшення кількості зв'язків між суб'єктами та об'єктами суспільної діяльності і, як наслідок, збільшення числа ключових гравців, які володіють значним соціальним капіталом.

Чим вище показник мережевої проміжної центральності, тим вище ймовірність того, що даний учасник володіє значним соціальним капіталом і системними компетенціями, що дозволяють йому контролювати інформаційні потоки всередині системи спільної мережевої діяльності. Саме тому центральність з посередництва є найбільш точною мірою для визначення ступеня здатності індивіда контролювати взаємодію людей в своєму соціальному оточенні.

Основним недоліком цієї метрики є те, що вона є однією з найбільш затратних у часі та має порядок $O\left(|N|^3\right)$. Це пояснюється тим, що метрика відсте-

жує саме зв’язки між вершинами, а не кількість самих вершин як, наприклад, показникова метрика, і навіть у порівняно маленькій мережі з 2200 вузлів (мережа сімейних зв’язків, друзів, колег та знайомих) має аж 4,8 мільйони пар зв’язків.

## 2.5 Метрика за власним вектором (eigenvector centrality measure)

При розгляданні показникової метрики як недолік було зауважено те, що не враховується значимість сусідів досліджуваної вершини і метрика за власним вектором вирішує цю проблему. Наприклад, одна вершина може мати однакову кількість сусідів з іншою, але ця вершина може бути важливою для поєднання компонентів графу, а інша знаходитись на периферії та не буде відігравати великої ролі для мережі в цілому. Саме тому ідея Eigenvector метрики в тому, щоб врахувати кількість зв’язків не тільки конкретної вершини, а ще і її сусідів. Центральність показує ступінь впливу конкретної вершини на мережу в цілому.

Центральність за власним вектором для вершини $i$ розраховується за формулою:

$$Ax = \lambda x, \tag{2.8}$$

де $A$ – це матриця суміжності графа $G$ з власним значенням $\lambda$.

Також можна використовувати формулу:

$$v_i = \frac{1}{\lambda}\sum_j A_{ij} v_j . \tag{2.9}$$

Загалом, існує багато різних власних значень $\lambda$, для яких розв’язок власного вектора існує. Згідно з теоремою Перрона-Фробеніуса, існує єдиний дода-

тній розв’язок цієї рівності, якщо $\lambda$ – найбільше власне значення власного вектору матриці суміжності $A$.

Центральність за власним вектором застосовується у багатьох областях. Першою з них, звичайно є соціологія. Соціологи використовують її для визначення зв’язків

## 2.6 PageRank центральність

PageRank (PR) – це алгоритм, який використовується пошуковим механізмом Google для сортування веб-сторінок в результатах пошуку. PageRank був названий на честь Ларрі Пейджа, одного із засновників Google. PageRank – це спосіб вимірювання важливості сторінок сайту. За даними Google PageRank працює шляхом визначення кількості та якості посилань на сторінку, щоб визначити приблизну оцінку важливості веб-сайту. Основна ідея алгоритму полягає в тому, що більш важливі веб-сайти можуть отримувати більше посилань з інших веб-сайтів.

В даний час PageRank – це не єдиний алгоритм, який використовується Google для упорядкування результатів пошуку, але це перший алгоритм, який був використаний компанією, і він є найвідомішим.

Цей алгоритм будується на понятті важливості вузла для мережі, при цьому під важливістю розуміється кількість посилань на конкретний вузол з інших вузлів. Наприклад, така сторінка як http://www.yahoo.com/, буде містити десятки тисяч зворотних посилань, вказуючи на цей сайт.

Той факт, що домашня сторінка Yahoo має так багато зворотних посилань зазвичай має на увазі, що це досить важливе джерело інформації. Дійсно, багато пошукових систем використовують розрахунок зворотних посилань, щоб спробувати зробити їх механізм пошуку на користь більш якісних або більш важливих сторінок. Проте, просте зворотне посилання має безліч проблем в мережі. Деякі з цих проблем пов`язані з характеристиками мережу, якій немає в зви-

чайних базах даних академічного цитування.

Причина того, що PageRank цікавий, полягає в тому, що в багатьох випадках кількість зворотних посилань не відповідає нашому розумінню важливості. Наприклад, якщо веб-сторінка має посилання на домашню сторінку Yahoo, це може бути тільки одне посилання, але вона дуже важлива саме через те, що Yahoo важливий для мережі.

Саме тому ця сторінка повинна бути вище, ніж багато сторінок з великою кількістю посилань з незрозумілих місць.

PageRank – це спроба побачити, наскільки добре можна отримати наближення до «важливості» зі структури посилань.

Ґрунтуючись на обговоренні вище, ми даємо наступний інтуїтивний опис PageRank – сторінка має високий ранг, якщо сума рангів його зворотних посилань висока. Це стосується як випадку, коли сторінка має багато зворотних посилань і коли сторінка має декілька високо оцінюваних зворотних посилань.

Нехай $u$ – це веб-сторінка. Тоді нехай $F_u$ буде набором посилань сторінок, що виходять з $u$, а $B_u$ буде набором сторінок, які входять в $u$. Нехай $N_u = |F_u|$ буде числом посилань, що виходять з $u$, а $c$ – нормалізуючий коефіцієнт.

Починаємо визначати просте ранжування $R$, яке є спрощеною версією PageRank. Воно визначається за формулою:

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}. \qquad (2.10)$$

Це рівняння формалізує висновки, наведені вище. У ньому враховується, що ранг сторінки розподілений рівномірно, щоб внести вклад в ранг сторінок, на які вони вказують або посилаються. При цьому $c$ повинен бути менше одиниці, тому що є сторінки без прямих посилань і їх вага в системі втрачається. Рівняння є рекурсивним і може бути розраховано, починаючи з будь-якого на-

бору рейтингів вершин до тих пір, доки воно не зійдеться.

Слід зауважити, що з цією спрощеною функцією є невелика проблема. Припустимо, що дві веб-сторінки вказують одне на одного і більше ні на кого іншого, але є сторінка, яка вказує на одну з них. Тоді при проході циклу він буде накопичувати ранг і не буде його розподіляти через те, що не буде вихідних ребер. Цей цикл утворює деяку пастку, яка називається ранговим провалом.

Щоб уникнути цієї проблеми було введено таке визначення.

Нехай $E(u)$ – це деякий вектор на веб-сторінках, який відповідає джерелу рангу. Коефіцієнт PageRank буде обчислюватися таким чином:

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u). \qquad (2.11)$$

При цьому коефіцієнт $c$ максимізується та $\|R'\|_1 = 1$ у просторі $L_1$.

Єдиним недоліком цієї моделі є «звисаючі» посилання, тобто ті, які вказують на сторінки без вихідних посилань. Вони впливають на модель через те, що невідомо де їхня вага повинна бути розподілена і таких посилань велика кількість.

Через те, що «звисаючі» посилання не впливають на ранжування будь-якої іншої сторінки безпосередньо, ми просто видаляємо їх з системи, поки всі коефіцієнти PageRank не будуть розраховані. Після цього вони можуть бути повернені назад без істотних змін.

Головна область використання PageRank – це пошук. Його основна перевага полягає в тому, що він здатний виділяти найважливіше для запитів, які не визначені однозначно. Наприклад, запит «Стенфордський університет» може повернути величезну кількість веб-сторінок, в яких згадується Стенфорд, але завдяки PageRank, головна сторінка сайту університету буде показана першою.

Основною метою PageRank, як було згадано вище, є пошук, а точніше, сортування прихованих посилань таким чином, щоб «кращі» посилання з'явля-

лися в списку результатів першими. Також PageRank може бути використаний при виборі цікавих статей або новин для користувача. Наприклад, людина яка використовує сайт новин, завжди хоче відстежувати будь-які важливі посилання, тобто новини, які користуються особливою популярністю і на які веде безліч посилань з інших джерел. Крім того, PageRank може допомогти користувачеві вирішити, чи заслуговує цей сайт довіри або ж ні. Наприклад, користувач схильний довіряти новинам, які показуються на головній сторінці сайту Стенфордського університету і тому в списку інших новин вони будуть показуватися першими [9].

## 2.7 Центральність за завантаженням (Load centrality)

Метрики центральності відіграють ключову роль у розгляді структури і аналізу графів і грають ключову роль в проблемах, пов'язаних з мережами, такими як розміщення сервісів, аналіз стабільності мережі і їх оптимізація. Проміжна центральність є однією з найпопулярніших, але алгоритм для її розрахунку є досить складним через те, що для цього необхідно знати усі найкоротші шляхи між будь-якими вершинами графа.

У комп’ютерних мережах проміжна центральність може бути використана для розміщення сервісів, для поліпшення маршрутизації, для управління і контролю над топологіями, для безпеки і т.д.

Центральність за завантаженістю є однією з варіацій проміжної центральності з алгоритмом, який є більш точним і розподіленим. Вона була введена для того, щоб зменшити час збіжності алгоритму визначення найкоротших шляхів між вершинами і може бути застосований для побудови бездротових розподілених мереж. Одним з додатків цієї метрики центральності може бути поліпшення розширюваності і стійкості протоколів для маршрутизації.

Ця центральність дозволяє використовувати ефективний розподілений алгоритм на основі алгоритму Беллмана-Форда для застосування до будь-якого

протоколу маршрутизації. Алгоритм розраховує точне число центральності, яка передбачає очікуване навантаження на вершину, розглядаючи маршрутизацію по шляхах з рівними вагами. У найпростішому випадку, коли існує тільки один шлях з мінімальною вагою між двома вершинами, результати будуть збігатись з алгоритмом проміжної центральності.

Алгоритм може бути застосований до будь-яких зважених і не зважених графів. Його час збіжності зростає лінійно з ростом діаметра мережі $D$, при цьому кожен вузол буде знати власну центральність і центральність інших вузлів.

Нехай $G(V,E)$ – це граф, у якому $V$ – множина усіх вершин графа, а $E$ – множина усіх ребер графа. Тоді центральність за завантаженістю або load centrality буде визначатися наступним чином.

Розглянемо граф $G(V,E)$ і алгоритм для визначення шляхів з мінімальною вагою між любими двома парами вершин $(s,d)$. Тоді нехай $\theta_{s,d}$ – це якість посилань, що надходять від вершини $s$ до вершини $d$. Припускаємо, що наступний перехід завжди здійснюється по шляху з мінімальною вагою, а в разі наступного переходу трафік між ними ділиться порівну. Тоді число $\theta_{s,d}(v)$ (commodity) буде показувати загальну кількість переданого трафіку через вершину $v$ і завантаженість цієї вершини буде визначатися рівнянням:

$$LC(v) = \sum_{s,d \in V} \theta_{s,d}(v). \qquad (2.12)$$

Зазвичай припускається, що $s \neq d, s \neq v, d \neq v$ і в загальному випадку $\theta_{s,d} = 1$. У випадку, коли граф неорієнтований, існує $\frac{N \cdot (N-1)}{2}$ пар $(s,d)$ і load centrality коефіцієнти можуть бути нормалізовані як

$$\overline{LC}(v) = \frac{N \cdot (N-1)}{2} \sum_{s,d \in V} \theta_{s,d}(v). \qquad (2.13)$$

Центральність за завантаженістю і проміжна центральність є дуже схожими між собою, але вони не збігаються. Це видно з прикладу, описаному нижче. На малюнку однієї і тієї ж мережі представлені коефіцієнти проміжної центральності та центральності по завантаженості, припускаючи, що кожне ребро має однакову вагу. Якщо число $\theta_{s,t}$ ділиться навпіл між двома наступними кроками, що лежать на двох маршрутах з еквівалентною загальною вагою, то вузли $v$ и $w$ будуть ділити коефіцієнт порівну, як бачимо далі з вершини $v$ знову виходять два шляхи і вони знову ділять завантаженість з вузла $v$ навпіл. У той же час, коефіцієнт для розрахунку проміжної центральності показує відношення найкоротших шляхів, що проходять через вершину $v$ з $s$ до вершини $t$ відносно загального числа усіх шляхів з $s$ в $t$. В цьому і є основна відмінність цих двох метрик. У самому простому випадку, коли шлях тільки один, очевидно, що результати будуть збігатися [10].

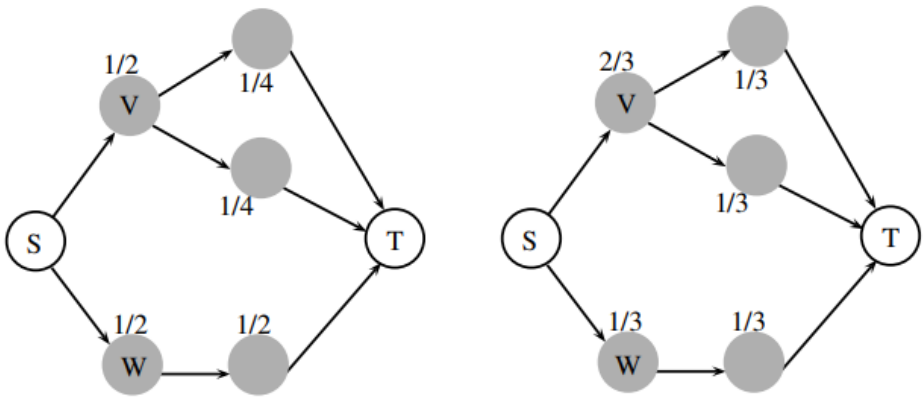


Рисунок 2.2 – Різниця в коефіцієнтах між load та betweenness centrality

## 2.8 Сумісна метрика

Кожна з розглядуваних вище метрик підходить для вирішення якогось конкретного типу задач, але якщо необхідно врахувати відразу декілька факторів, врегулювавши їхню значимість, тоді для цього потрібен інший підхід.

Задля вирішення цієї проблеми на основі ідеї правила Борда визначення місць для кожної вершини після отримання коефіцієнтів значимості було вирішено побудувати сумісну метрику, яка б враховувала важливість кожної з метрик в залежності від її характеристик, простоти реалізації, точності та швидкості алгоритму.

Ідея полягає в тому, щоб після визначення коефіцієнтів значимості присвоювати їй відповідний бал (або місце), відштовхуючись від отриманого результату. Наприклад, якщо ми маємо граф з 10 вершин і вершина має найнижчий коефіцієнт, то їй присвоюється один бал, а якщо найвищий – то десять балів, а потім це значення помножити на коефіцієнт важливості самої метрики, який може бути обраний в залежності від типу розглядуваної задачі або за бажанням самого дослідника. Отримане число буде коефіцієнтом значимості сумісної метрики і буде розраховуватись за формулою

$$k_i = \sum_{j=0}^{n} c_j v_j \,, \tag{2.14}$$

де $i = \overline{1,m}$ – кількість вузлів у розглядуваному графі;

$c_j$ – коефіцієнт важливості розглядуваної метрики для дослідника, $\sum c_j = 1$;

$v_j$ – отримане місце вершини для розглядуваної метрики після розрахунку за правилом Борда;

$j = \overline{1,n}$ – номер розглядуваної метрики;

$n$ – кількість метрик, що були обрані для розв’язання поставленої задачі.

Для подальших розрахунків коефіцієнтів сумісної метрики у розділі 4 необхідно також визначити коефіцієнти $c_j$, які відображають важливість кожної метрики, а також повинні враховувати точність алгоритму для визначення коефіцієнтів центральності та призначення цієї метрики.

Усі характеристики та коефіцієнти були зведені до таблиці 2.1.

Таблиця 2.1 – Коефіцієнти важливості метрик та їхні характеристики

| Назва метрики | Характеристика | Коефіцієнт важливості |
|---|---|---|
| Degree | Є найпростішою метрикою, не вказує на «якість» зв’язків, а тільки визначає їхню кількість. Може бути застосовна для визначення «престижу» вершин. | 0,05 |
| Closeness | Вершини, що володіють високим коефіцієнтом центральності за цією метрикою мають здатність до швидкого доступу до усіх інших вершин і добре підходять для передачі інформації. | 0,15 |
| Betweenness | Вершини з високим коефіцієнтом центральності виконують роль «моста» між вершинами та є ланцюгом для компонентів мережі. | 0,1 |
| Eigenvector | Метрика враховує не тільки кількість зв’язків розглядуваної вершини, але і зв’язки її сусідів, щоб визначити «якість» цих зв’язків. Вона показує ступінь впливу вершини на мережу. | 0,2 |
| PageRank | Більш вдосконалена версія Eigenvector метрики, вершина буде мати високий коефіцієнт центральності, якщо на неї є багато посилань з інших вершин | 0,3 |
| Load | Володіє більш точним та швидким алгоритмом визначення коефіцієнтів центральності, ніж у проміжної метрики і використовується для побудови бездротових мереж. | 0,2 |

# 3 ПРОГРАМНА РЕАЛІЗАЦІЯ

## 3.1 Вибір мови програмування та необхідних бібліотек

У якості мови програмування була обрана високорівнева мова загального призначення – Python. Python підтримує декілька парадигм програмування, в тому числі структурну, об’єктно-орієнтовану, функціональну, імперативну та аспектно-орієнтовану. Основні архітектурні риси – динамічна типізація, автоматичне керування пам’яттю, механізм обробки винятків та зручні високорівневі структури даних. Python організується у функції та класи, які можуть об’єднуватися в модулі.

Розробка мови була розпочата на початку 1980-х років співробітником нідерландського інституту CWI Гвідо ван Россумом. Для розподіленої операційної системи Amoeba була потрібна скриптова мова програмування, і Гвідо почав займатися її розробкою. Важливою метою розробників мови було зробити її кумедною для користування. Це відображено в грайливому підході до створення документації та навчальних програм [11].

Python стане в нагоді у випадках, коли виникають завдання, пов’язані з аналізом даних, з роботою веб-додатків, або якщо статистичний код потрібно інкорпорувати в робочу базу даних. Python, будучи повнофункціональною мовою програмування, відмінно підходить для реалізації алгоритмів з їх подальшим практичним використанням. Ще нещодавно пакети для аналізу даних на Python перебували в зародковому стані, що представляло певну проблему, але в останні роки ситуація значно покращилася.

У цієї мови є дуже багато переваг перед іншими, якщо постає задача аналізу соціальних мереж через те, що зараз дуже багато інструментів для обробки Big Data створено саме для Python. Крім цього, мова є дуже простою, тому що код є зрозумілим навіть для людини, яка до цього ще не була знайома з цією мовою, саме на простоті та читабельності розробники мови і робили акцент при її написанні. Своєрідною «фішкою» мови є її доступність. Весь код, написаний

на Python є відкритим.

Ще однією перевагою цієї мови є і те, що її підтримують такі гіганти IT-сфери як: Google, Dropbox, Mozilla, Facebook та ін. Це говорить про те, що великі компанії не бояться робити ставку на Python, тому що впевнені у її надійності та знають, що розробників для цієї мови вдосталь і технології, що створені за допомогою Python будуть жити.

На мою думку, в рамках цієї роботи та й у цілому для математичних розрахунків Python підходить найбільше, тому що в основному загальні та найбільш потужні для обчислень бібліотеки створені саме під Python. До них можна віднести SciPy, NumPy, Matplotlib для побудови графіків та візуалізації, sklearn для машинного навчання а також безліч інших.

Для програмної реалізації у даній роботі була використана бібліотека NetworkX.

Бібліотека NetworkX призначена для роботи з графами та іншими мережевими структурами. Основними можливостями бібліотеки є:

а) велика кількість класів для роботи з простими, зваженими та орієнтованими графами;

б) збереження/завантаження графів в/з найбільш розповсюджених форматів для зберігання файлів з графами;

в) вбудовані процедури для створення графів базових типів;

г) візуалізація мереж у вигляді 2D і 3D графіків.

І хоча самі розробники кажуть, що візуалізація не є пріоритетною метою даної бібліотеки, але вкупі зі зручними інструментами для маніпуляції, її застосування доволі перспективне. Також у неї досить непогана документація,тому ця бібліотека є популярною для аналізу соціальних мереж та графів.

## 3.2 Вибір засобу візуалізації

У якості засобу візуалізації отриманих результатів був використаний

Gephi – це пакет програмного забезпечення з відкритим кодом для мережевого аналізу. Він дозволяє доволі чітко розрізняти усі можливі зв’язки між вершинами та добре взаємодіє з бібліотекою NetworkX, завдяки цьому можна задавати колір та розмір вершин, а сам Gephi надає можливість задавати необхідні примітки для кожної вершини, регулювати розміри міток для вузлів графа, а також має функцію затінення інших вузлів графа при наведенні на конкретну вершину.

Gephi є дуже потужним засобом візуалізації різних даних, який підійде як новачкові, так і досвідченому користувачеві. Використання баз даних дозволяє швидко і легко отримати необхідні вхідні дані для візуалізації, але крім цього Gephi підтримує багато інших форматів таких як .gexf, .gephi, .gml, .graphml та ін.

Був розроблений студентами французького Технологічного університету Комп'єня (UTC) у 2009 році. Gephi обирався для Google Summer of Code (ініціативана програма для проектів з відкритим вихідним кодом) п’ять разів поспіль: у 2009, 2010, 2011, 2012 та 2013.

Gephi активно використовується в цілому ряді академічних дослідницьких проектів, зокрема соціологічних; також швидко здобув популярність серед журналістів. Зараз його користувацьке середовище значно розширилося — за допомогою цього пакета можна займатися будь-якою темою мережевого аналізу. Gephi використовувався, серед іншого, для візуалізації глобальної зв'язності контенту New York Timesta вивчення мережевого трафіку Twitter під час соціальних заворушень: Gephi надихав створення LinkedIn InMaps і був використаний для візуалізації цілої мережі Truthy.

Загалом Gephi широко використовується в так званих «цифрових гуманітарних науках» (Digital humanities): в історії, літературі, політології тощо. Спеціалісти з такого середовища активно приймають участь у подальшій розробці та популяризації продукту [12].

Gephi містить в собі набір основних укладок, а також безліч інших інструментів для аналізу графів. Від програмістського community для Gephi написано багато плагінів для експорту укладання в інтерактивну веб-сторінку. Та-

кож оригінальна імплементація OpenOrd міститься саме в Gephi. У Gephi є інструменти розмальовки вершин і ребер за їх властивостями, настройка підписів, розмірів і інших параметрів відтворення. Є експорт в основні формати зображень, включаючи векторні.

Дуже прикрий факт в тому, що Gephi вже кілька років не підтримується. Два основних розробника не володіють ресурсами щоб передати свої знання, необхідні для подальшої розробки комусь ще, а також заявили, що вже не можуть більше активно підтримувати Gephi. З останніх новин, в блозі проекту з'явилася заява про те, що потужності сучасного WebGL вже обганяють старий Gephi і є шанси побачити його відродження у вигляді веб-додатку.

## 3.3 Опис програми

Програма виконана в середовищі Jupyter Notebook на повнофункціональній мові програмування Python.

Основні елементи програми складаються з трьох файлів.

Перший файл – Neural_network_graph.ipynb. У цьому модулі виконується завантаження графа орієнтованої нейронної мережі нейронів і синапсів черв'яка C.elegans за допомогою вбудованих функцій бібліотеки NetworkX, визначення коефіцієнтів центральності за усіма розглянутими метриками та встановлення необхідних параметрів для візуалізації за допомогою додатку Gephi. Першою розглядається показникова метрика центральності, потім метрика близькості, проміжна, метрика за власним вектором, PageRank метрика та метрика за навантаженістю. Після розрахунку усіх місць для кожної вершини визначаються значення вершин сумісної метрики.

Для візуалізації необхідно задати кожній вершині розмір та колір. Чим більший коефіцієнт центральності, тим яскравішим кольором володіє вершина і має більший розмір.

Після аналізу отриманих результатів для кожної метрики були визначені

максимальний та мінімальний коефіцієнти значимості вершин.

Другий файл – Email_graph.ipynb. У цьому файлі розглянута мережа електронних листів Європейського інституту досліджень, що складається з 986 вершин та 24929 ребер.

Розрахунки аналогічні тим, що проводились для нейронної мережі, відмінність лише у діапазоні коефіцієнтів для кожної з груп.

У третьому файлі під назвою Gnutella_graph.ipynb розглядається орієнтований граф нейронної мережі з 8717 вершинами та 31525 ребрами.

Відмінністю від попередніх файлів є те, що вершинам не присвоювались колір і розмір, а тільки були виконані необхідні розрахунки для визначення коефіцієнтів сумісної метрики.

# 4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ

## 4.1 Результати досліджень для орієнтованої нейронної мережі

Так як основною задачею дослідження є розпізнавання лідерів у доступних соціальних графах та порівняння отриманих коефіцієнтів значимості за кожною з метрик центральності, то метою створення програмного продукту було належне порівняння графів різної розмірності та різних типів, обчислення цих коефіцієнтів та візуалізація у зручному для користувача вигляді.

Для проведення досліджень було обрано три соціальні мережі:

– орієнтована мережа нейронної системи з усіма нейронами та синапсами черв’яка C.elegans з 296 вершин та 2344 ребер;

– орієнтована мережа електронних листів Європейського інституту досліджень, що представляє зв’язки між користувачами. Складається з 986 вершин та 24929 ребер;

– орієнтована мережа системи для передачі файлів Gnutella, де вузли представляють собою хости (сервери), а ребра – взаємодію між серверами. Мережа має 8717 вершин та 31525 ребер.

Першою для дослідження була використана мережа нейронної системи черв’яка C.elegans, яка ідеально підходить для візуалізації, тому що має відносно невелику кількість вершин та ребер. Загальний вигляд усієї мережі зображений на рисунку 4.1.

Також для аналізу обраної мережі спочатку було обрано показникову метрику, що обчислює кількість інцидентних кожній вершині ребер, а потім нормалізує отримане значення для кожної вершини.

Для розпізнавання лідерів у мережі вузли з найбільшим коефіцієнтом значимості за показниковою метрикою мають більший розмір та яскравіший колір. Так, перша група (з найвищими коефіцієнтами – лідери) має червоний колір, друга – жовтий, третя – зелений, четверта – синій та п’ята – без кольору (з міткою).

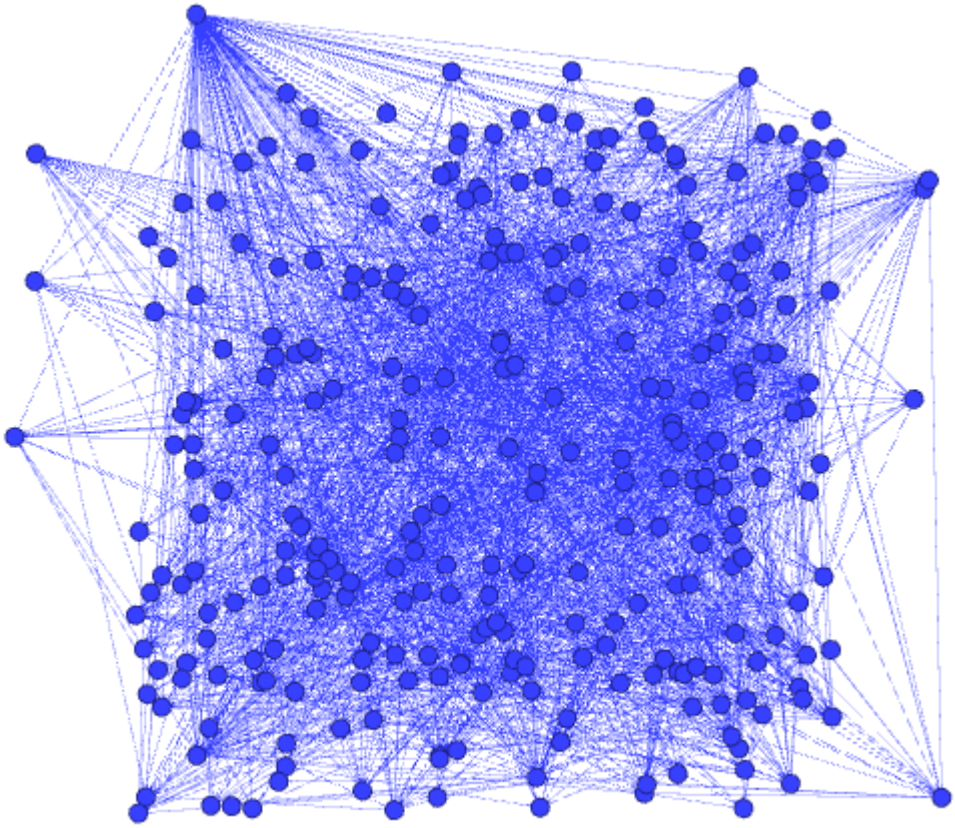

Рисунок 4.1 – Мережа нейронів черв’яка C.elegans

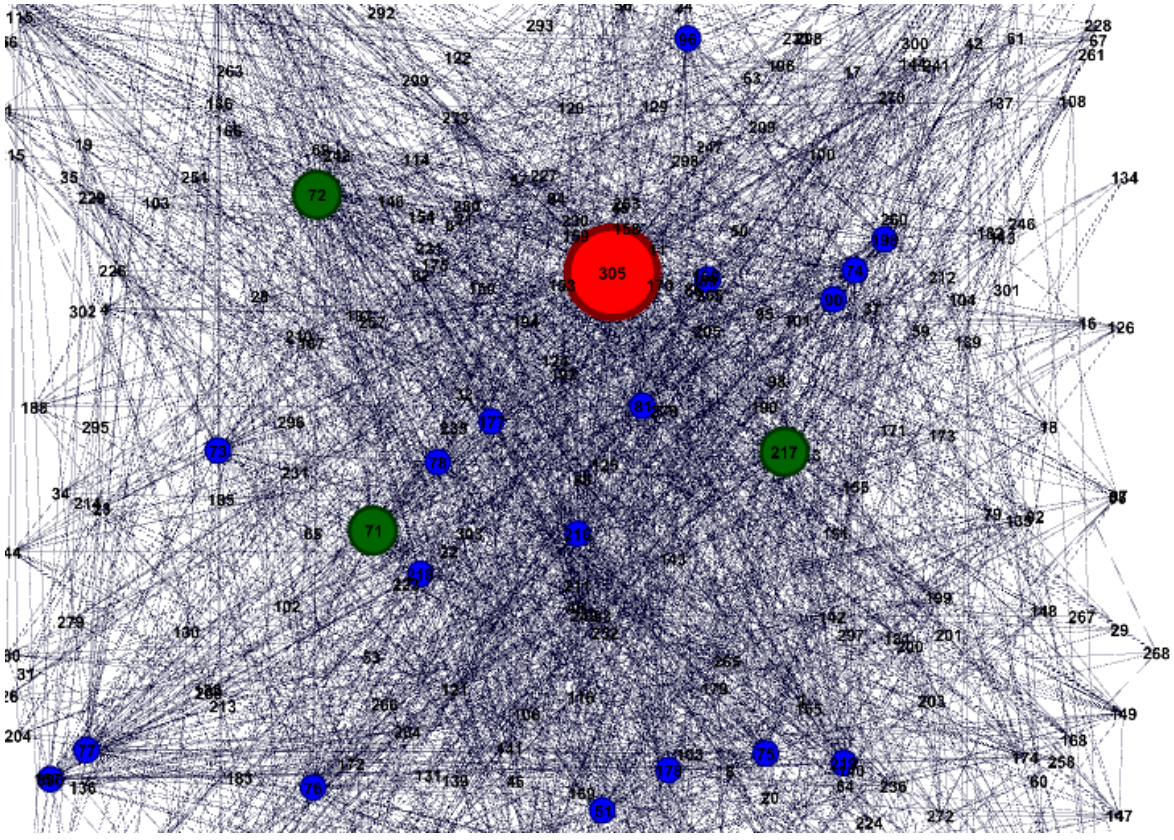


Рисунок 4.2 – Лідери нейронної мережі за показниковою центральністю

До першої групи потрапили вершини з коефіцієнтами від 0,4 до 0,5, до другої – від 0,3 до 0,4, до третьої – від 0,2 до 0,3, до четвертої – від 0,1 до 0,2 і до п’ятої усі інші вершини, коефіцієнт яких менше ніж 0,1.

Як видно з наведеного рисунку лідером з коефіцієнтом значимості рівним – 0,45085 стала вершина з номером 305. До другої групи не попала жодна з вершин, а ось до третьої групи одразу три – 71 з коефіцієнтом 0,28136 та 72-га з коефіцієнтом 0,27119, а також 217-та з коефіцієнтом 0,20339. Отже, можна зробити висновок, що найбільше зв’язків з іншими вершинами мають саме ці вище наведені вузли. Але, знову ж таки, зазначимо, що ці зв’язки не відображають їхню якість та важливість для графа, а лише кількість.

Наступною для розгляду є метрика близькості, у ній також усі вершини були розділені на групи, кольори для груп були обрані тим же чином, як і для показникової центральності. Результати відображені на рисунку 4.3.

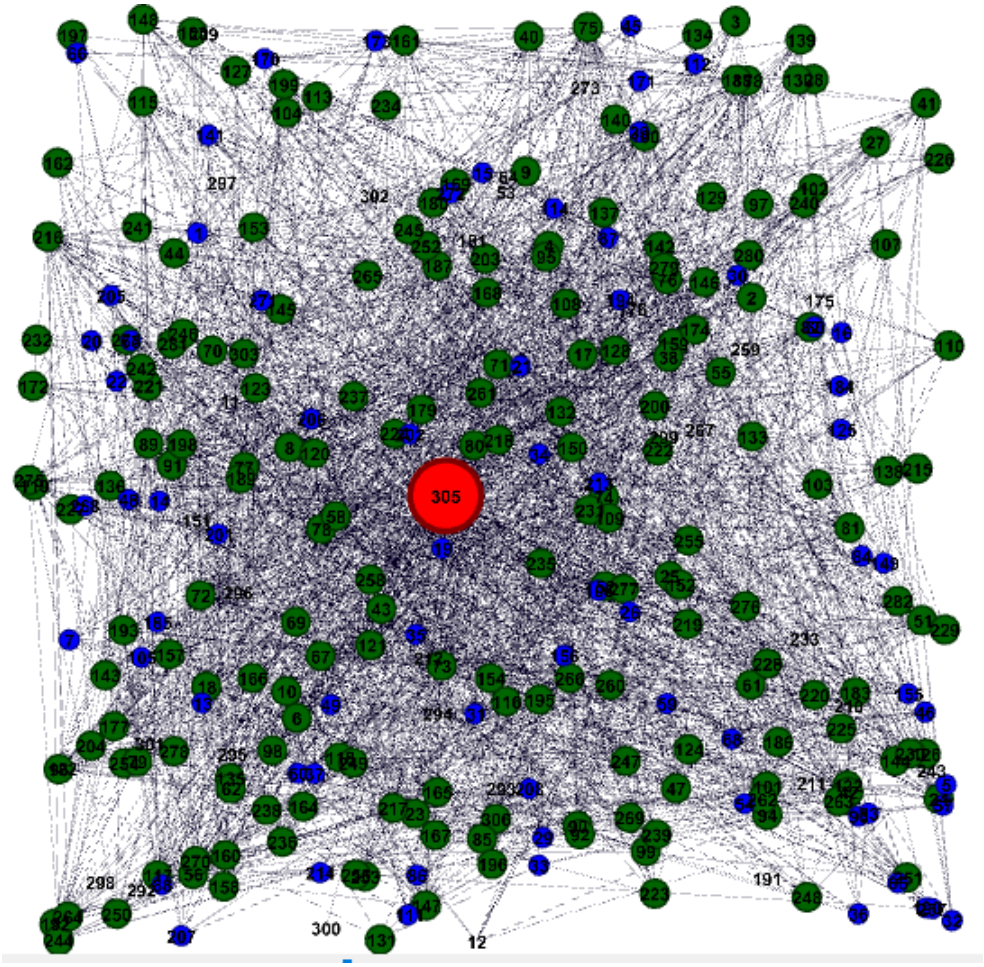


Рисунок 4.3 – Лідери нейронної мережі за метрикою близькості

Єдиною відмінністю є те, що відсутні група з коефіцієнтами у проміжку між 0,3 до 0,5, до цих груп не потрапила жодна з вершин. Максимальний коефіцієнт був знову у вершини 305 і дорівнював він 0,57578. Вершини синього кольору мають коефіцієнт від 0,1 до 0,2, а від 0,2 до 0,3 – зеленого. Як бачимо, вершина 305 має не тільки багато зв'язків з іншими, але й добре підходить для передачі даних.

Наступною для розгляду взято проміжну метрику. В неї майже всі коефіцієнти є значно меншими, ніж у попередніх груп, саме тому прийняте рішення зменшити діапазони для груп з різними кольорами. Так група з червоним кольором має коефіцієнти від 0,09 до 0,11, до другої жовтого кольору – від 0,03 до 0,09, до третьої зеленого кольору – від 0,02 до 0,03, до четвертої синього кольору – від 0,01 до 0,02, та до п'ятої усі інші, що менші, ніж 0,01. Результати приведені на рисунку 4.4.

Максимальним коефіцієнтом 0,10597 володіє вершина з міткою 195. Вона підходить для зв'язування компонентів графа та при її видаленні можна легко розірвати структуру мережі, що показує її справжню значимість.

У метрики за власним вектором також майже усі вершини, крім 305-ї (з максимальним коефіцієнтом 0,58048) мали низькі (менші за 0,1) коефіцієнти значимості. Тому до першої групи червоного кольору потрапила тільки 305-та вершина, до другої жовтого кольору – вершини з коефіцієнтами від 0,1 до 0,5, до третьої зеленого кольору – від 0,07 до 0,1, до четвертої синього кольору – від 0,04 до 0,07, до п'ятої сірого кольору – від 0,01 до 0,04.

Тепер можна стверджувати, що вершина 305 має дійсно «якісні» зв'язки тому, що її сусіди також мають високі у рамках розглядуваної мережі коефіцієнти значимості для метрики за власним вектором. Слід зазначити і те, що багато вершин потрапили до групи жовтого кольору, що свідчить про наявність деякого кластеру вершин, що мають серйозний вплив на мережу.

Отримані результати для метрики за власним вектором наведені на рисунку 4.5

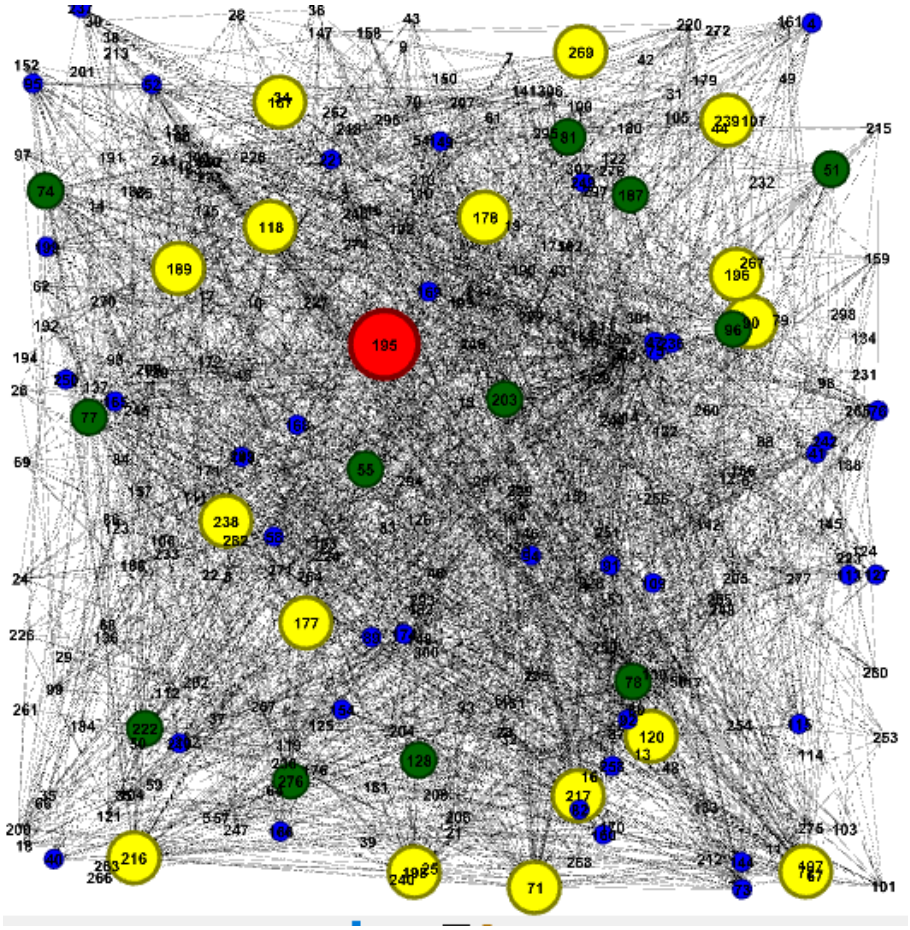

Рисунок 4.4 – Лідери нейронної мережі для проміжної метрики

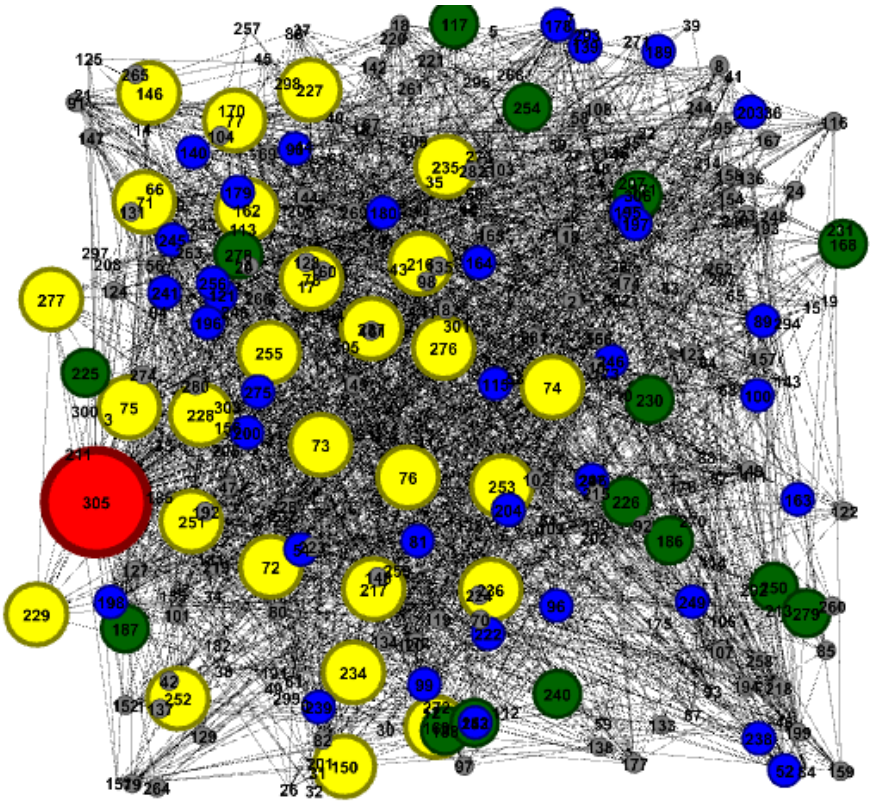

Рисунок 4.5 – Лідери нейронної мережі метрики за власним вектором

Для PageRank метрики групи були визначені наступним чином: перша група червоного кольору мала коефіцієнти від 0,1 до 0,13 та у неї потрапила лише одна вершина з максимальним коефіцієнтом, до другої жовтого кольору – від 0,004 до 0,008, до третьої зеленого кольору – від 0,003 до 0,004, до четвертої синього кольору – від 0,002 до 0,003, до п’ятої сірого кольору – від 0,001 до 0,002. Усі інші вершини кольору не мали, а мали лише мітку, як і в метриках, що наведені вище.

Вершиною з максимальним коефіцієнтом стала знову 305-та (коефіцієнт значимості – 0,12464). Високий коефіцієнт значимості за цією метрикою свідчить про те, що вершина має високий вплив і популярність на мережу, тому що на неї вказують багато інших «якісних» вершин.

Граф з розподілом на групи приведений на рисунку 4.6.

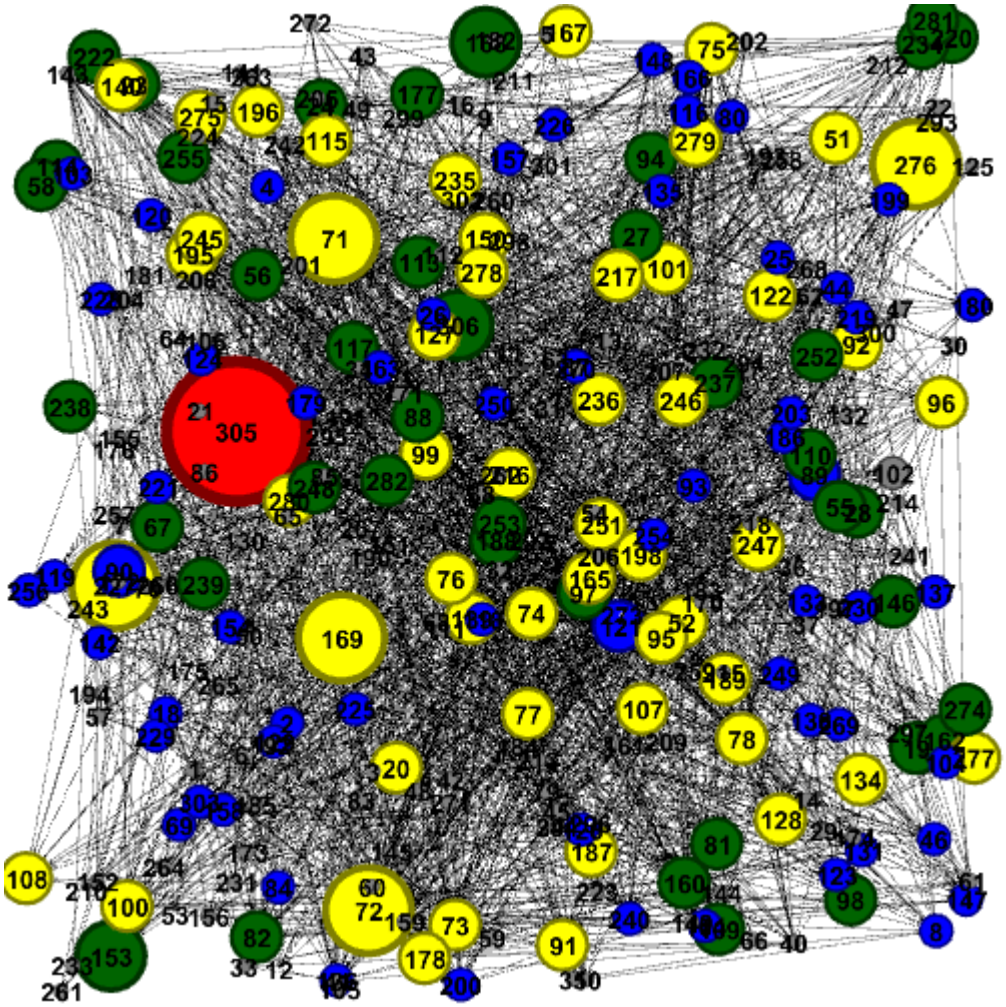


Рисунок 4.6 – Лідери нейронної мережі за PageRank метрикою

І нарешті для останньої метрики, що буде проілюстрована у рамках розгляду цієї орієнтованої нейронної мережі є метрика за навантаженістю, яка дуже схожа на проміжну метрику, але володіє більш точним і модифікованим алгоритмом на основі алгоритму Беллмана-Форда і використовується також для контролю потоку інформації.

Групи були розподілені наступним чином: до першої групи червоного кольору потрапили вершини з коефіцієнтами від 0,09 до 0,11, до другої жовтого кольору – від 0,04 до 0,09, до третьої зеленого кольору – від 0,03 до 0,04, до четвертої синього кольору – від 0,02 до 0,03, до п’ятої сірого кольору – від 0,01 до 0,02, а також усі інші вершини з коефіцієнтом, меншим за 0,01 потрапили до групи з відсутністю кольорового забарвлення. Як видно з отриманих результатів, вони дуже схожі на результати за betweenness метрикою.

Загальний граф з усіма групами приведений на рисунку 4.7.

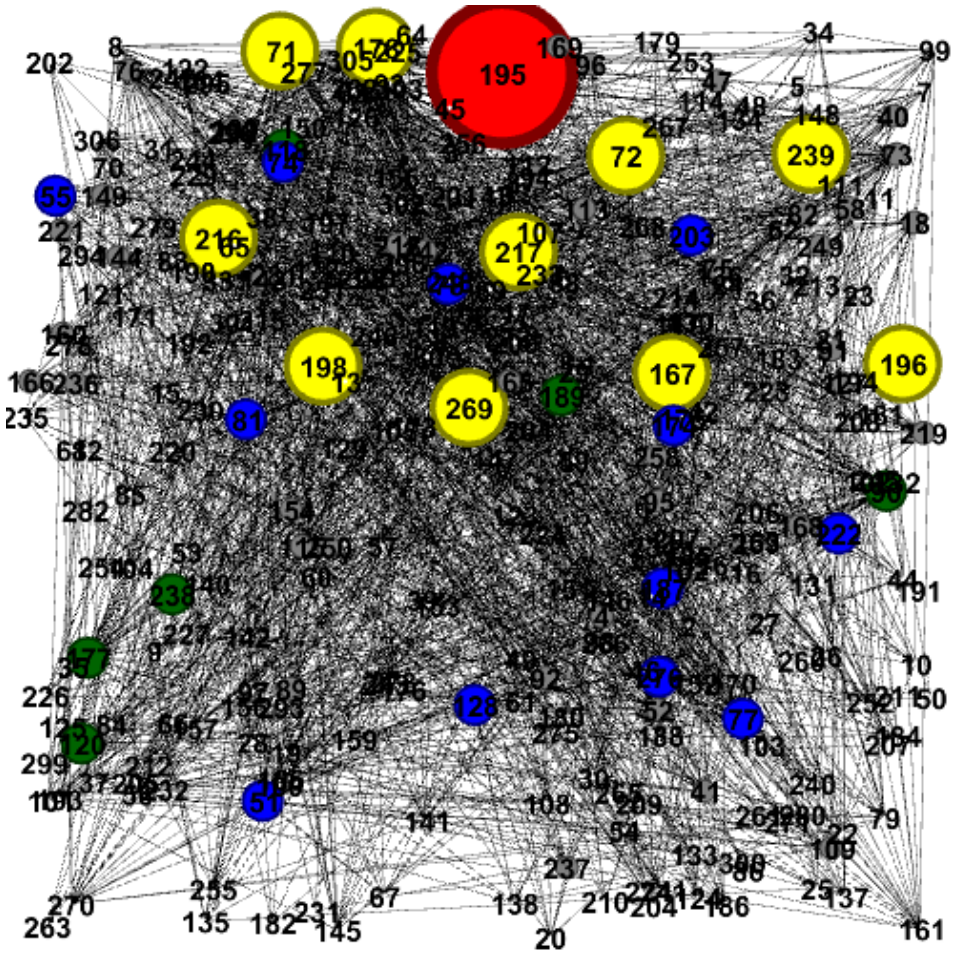


Рисунок 4.7 – Лідери нейронної мережі метрики за навантаженістю

Згідно з таблицею у пункті 2.8 для сумісної метрики були обрані наступні коефіцієнти для кожної з метрик: для показникової – 0,05, для метрики близькості – 0,15, для проміжної метрики – 0,1, для метрики за власним вектором – 0,2, для Pagerank – 0,3, для метрики за навантаженістю – 0,2. Після розрахунку отриманих місць вершин за кожною з метрик та домноженні цього місця на відповідний коефіцієнт було отримано наступні результати.

Таблиця 4.1 – Перша п’ятірка вершин з коефіцієнтами за сумісною метрикою

| Номер вершини | Коефіцієнт значимості |
|---|---|
| 71 | 292,6 |
| 72 | 290,7 |
| 276 | 280,75 |
| 74 | 280,05 |
| 77 | 279,95 |

Як видно з таблиці 4.1 за сумісною метрикою лідером стала вершина 71 з коефіцієнтом 292,6, а ось вершина 305, яка була лідером майже для усіх метрик не потрапила до неї через те, що мала вкрай низькі коефіцієнти для метрики за навантаженістю і проміжної.

## 4.2 Результати досліджень для орієнтованої мережі електронних листів

Наступною мережею для аналізу є також орієнтована мережа, що представляє собою зв’язки між членами усіх департаментів Європейського інституту досліджень.

Вона складається з 986 вершин та 24929 ребер. Повна структура графа зображена на рисунку 4.8.

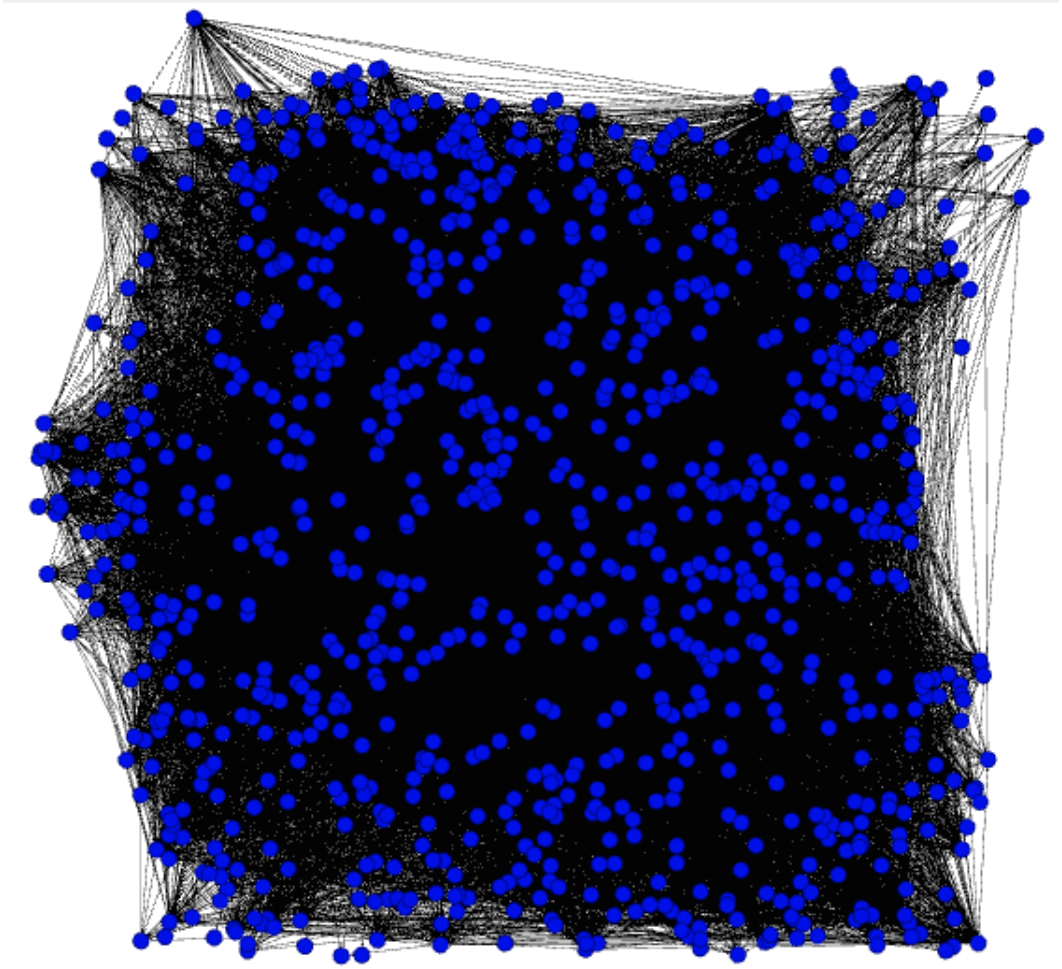

Рисунок 4.8 – Повна структура мережі електронних листів

Таким же чином, як і для попередньої мережі усі вершини були розбиті на групи в залежності від коефіцієнтів значимості. Для показникової метрики було виділено шість груп: до першої червоного кольору потрапили вершини з коефіцієнтами від 0,5 до 0,6, до другої жовтого кольору – від 0,4 до 0,5, до третьої зеленого кольору – від 0,3 до 0,4, до четвертої синього кольору – від 0,2 до 0,3, до п’ятої сірого кольору – від 0,2 до 0,1. Лідером за цією метрикою стала вершина з міткою 90 і коефіцієнтом значимості, що дорівнює 0,55228.

Для метрики близькості коефіцієнти були розподілені за тими ж підгрупами, але до першої групи не потрапила жодна з вершин, тому що найвищий коефіцієнт дорівнює 0,45834. Лідером за цією метрикою стала також вершина 90.

Розподіл груп зображений на рисунках 4.9 та 4.10 відповідно.

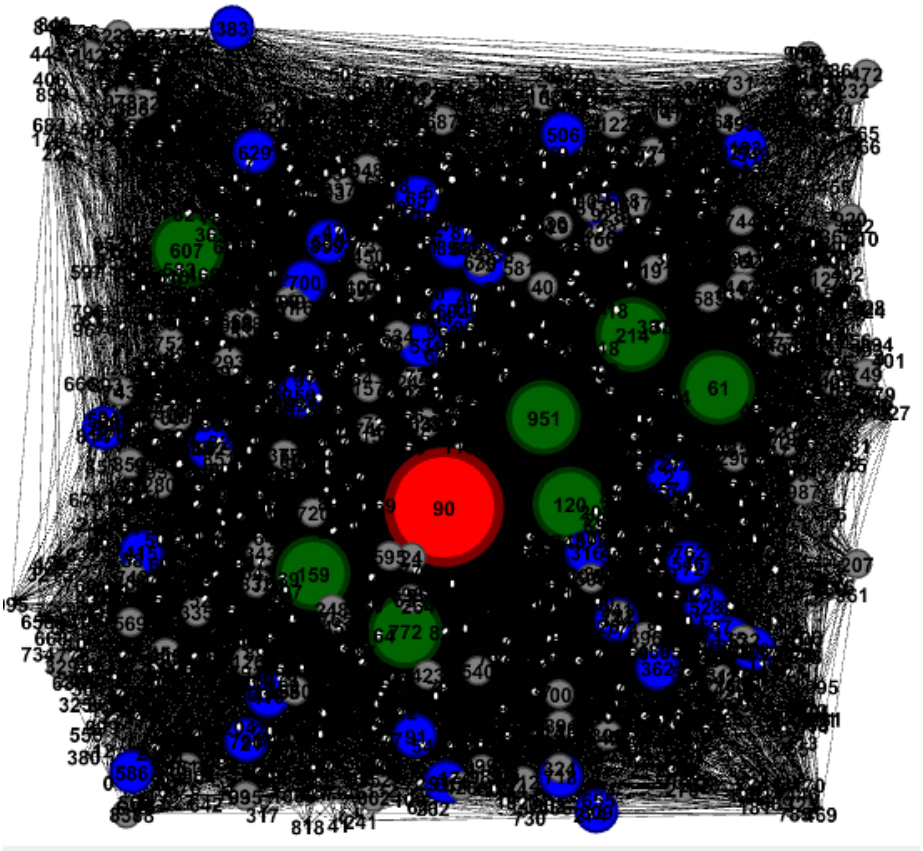


Рисунок 4.9 – Лідери мережі електронних листів показникової метрики

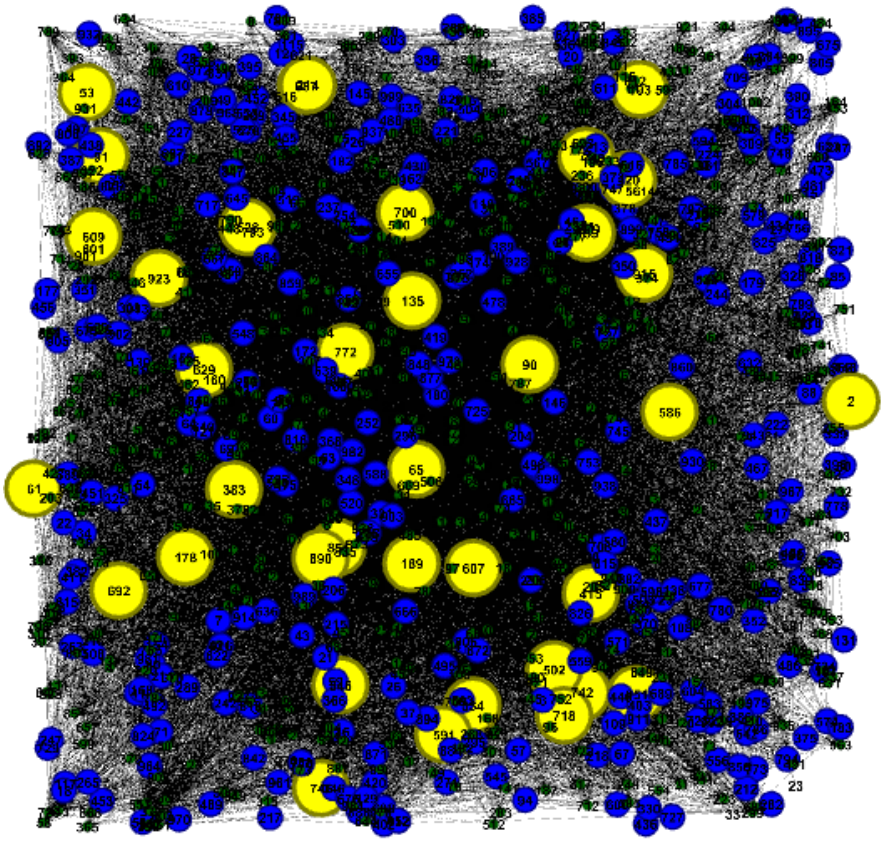


Рисунок 4.10 – Лідери мережі електронних листів за метрикою близькості

Коефіцієнти проміжної метрики було значно нижчими, ніж у показникової та метрики близькості і були розподілені на п’ять груп: перша жовтого кольору мала коефіцієнти від 0,07 до 0,1, друга зеленого кольору – від 0,03 до 0,05, третя синього кольору – від 0,01 до 0,03, четверта сірого кольору – від 0,001 до 0,01 і до п’ятої групи без забарвлення потрапили усі інші вершини з коефіцієнтами, нижчими за 0,001.

Лідером цієї метрики стала вершина з міткою 90 і коефіцієнтом 0,07493.

У метрики за власним вектором коефіцієнти були вже вищими і розподілились вже на шість груп: перша червоного кольору мала вершини з коефіцієнтами від 0,1 до 0,2, друга жовтого кольору – від 0,07 до 0,1, третя зеленого кольору – від 0,05 до 0,07, четверта синього кольору – від 0,03 до 0,05, п’ята сірого кольору – від 0,01 до 0,03 та усі інші потрапили до групи без забарвлення.

Лідером метрики знову стала вершина 90 з коефіцієнтом 0,14898.

Для PageRank метрики коефіцієнти розподілились наступним чином: перша група червоного кольору мала діапазон від 0,007 до 0,01, друга жовтого кольору – від 0,005 до 0,007, третя зеленого кольору – від 0,003 до 0,005, четверта синього кольору – від 0,002 до 0,003, п’ята сірого кольору – від 0,001 до 0,002.

Лідером, як і у попередніх метриках стала вершина 90 з коефіцієнтом 0,00752.

І нарешті для метрики за навантаженістю коефіцієнти були розподілені наступним чином: до першої групи червоного кольору потрапили вершини з коефіцієнтами від 0,05 до 0,071, до другої жовтого кольору – від 0,03 до 0,05, до третьої зеленого кольору – від 0,007 до 0,03, до четвертої синього кольору – від 0,004 до 0,007, до п’ятої сірого кольору – від 0,001 до 0,004.

Лідером знову стала вершина 90 з коефіцієнтом 0,07052, тому розподіл груп засновувався саме відштовхуючись від максимального коефіцієнту.

На рисунках 4.11 – 4.14 зображений розподіл усіх груп відповідно до вказування їх у тексті.

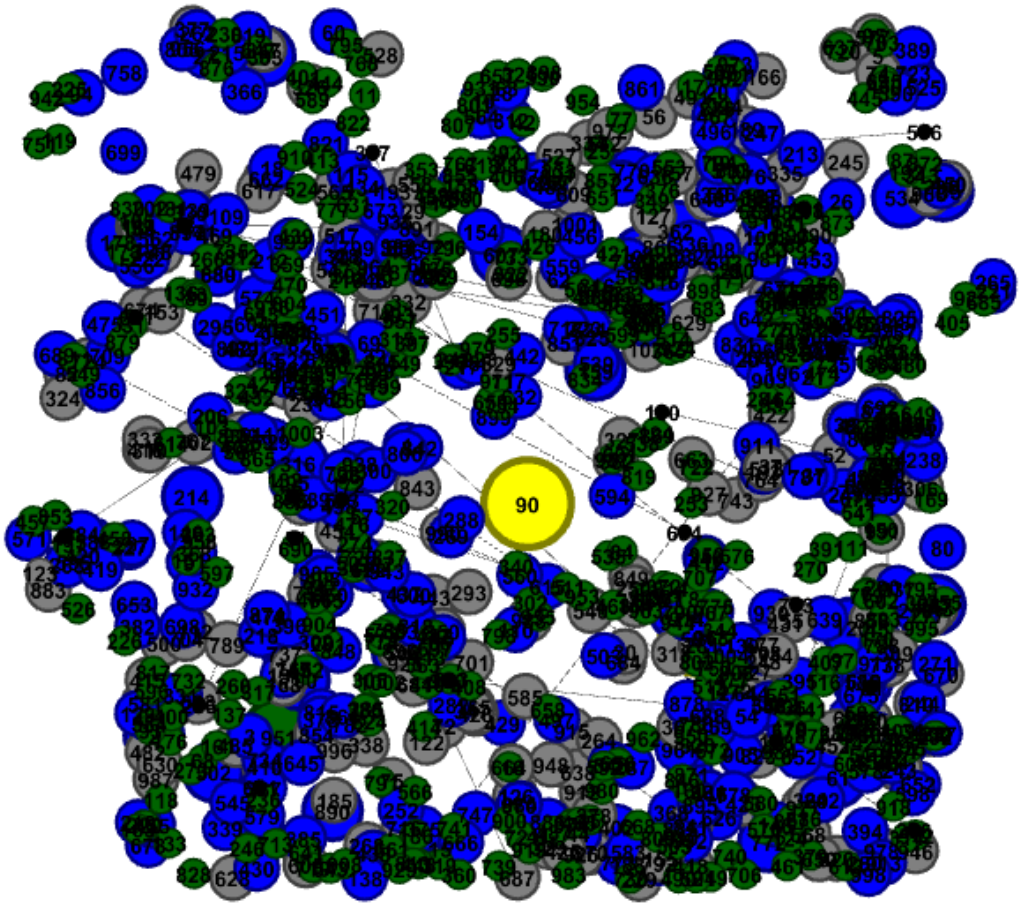


Рисунок 4.11 – Лідери мережі електронних листів проміжної метрики

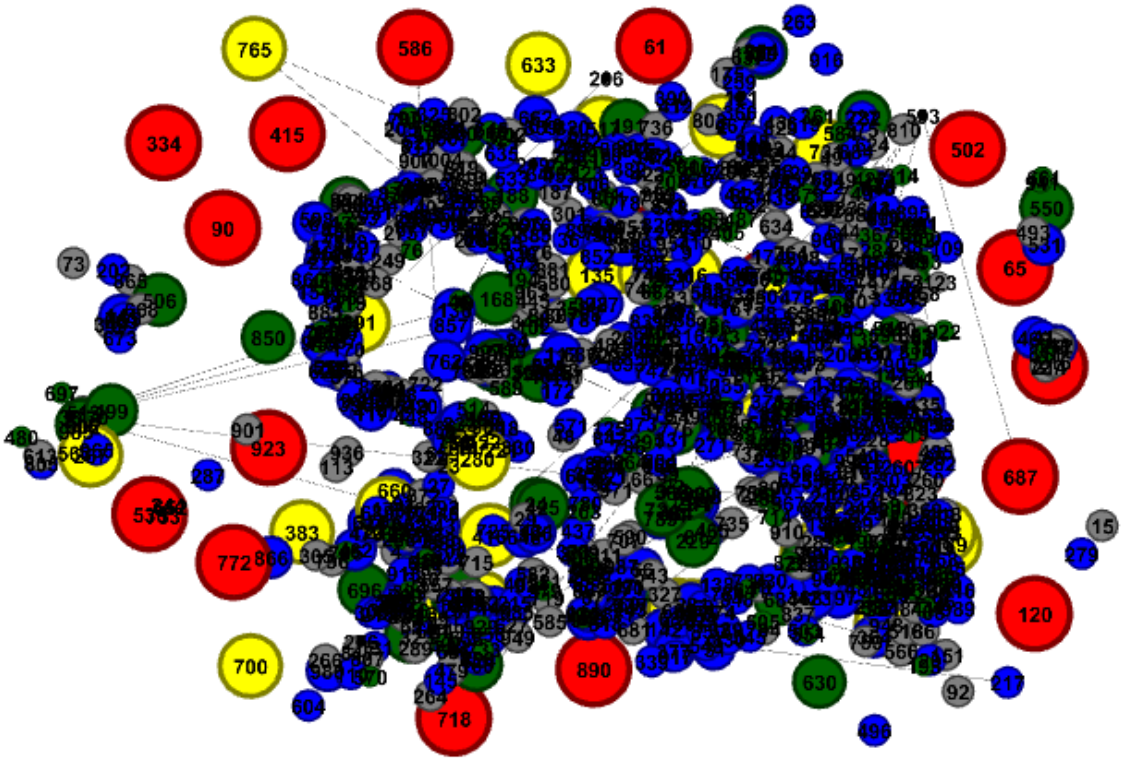


Рисунок 4.12 – Лідери мережі електронних листів метрики за власним вектором

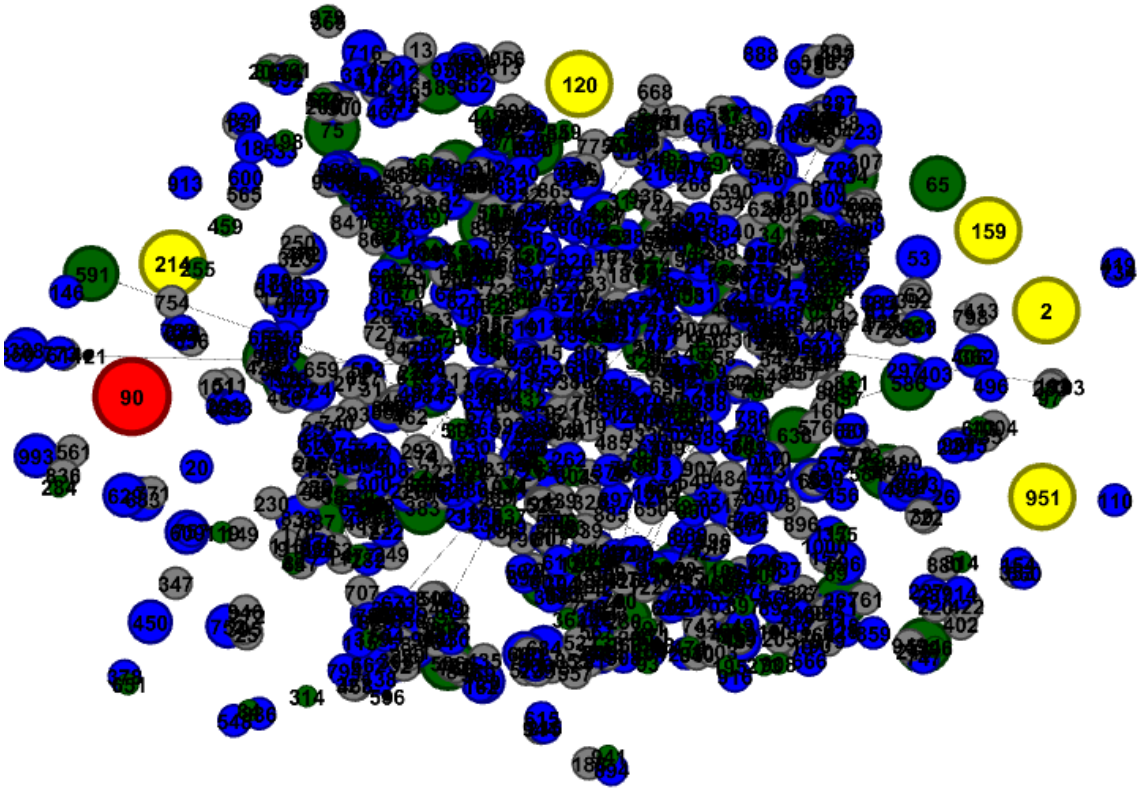


Рисунок 4.13 – Лідери мережі електронних листів за PageRank метрикою

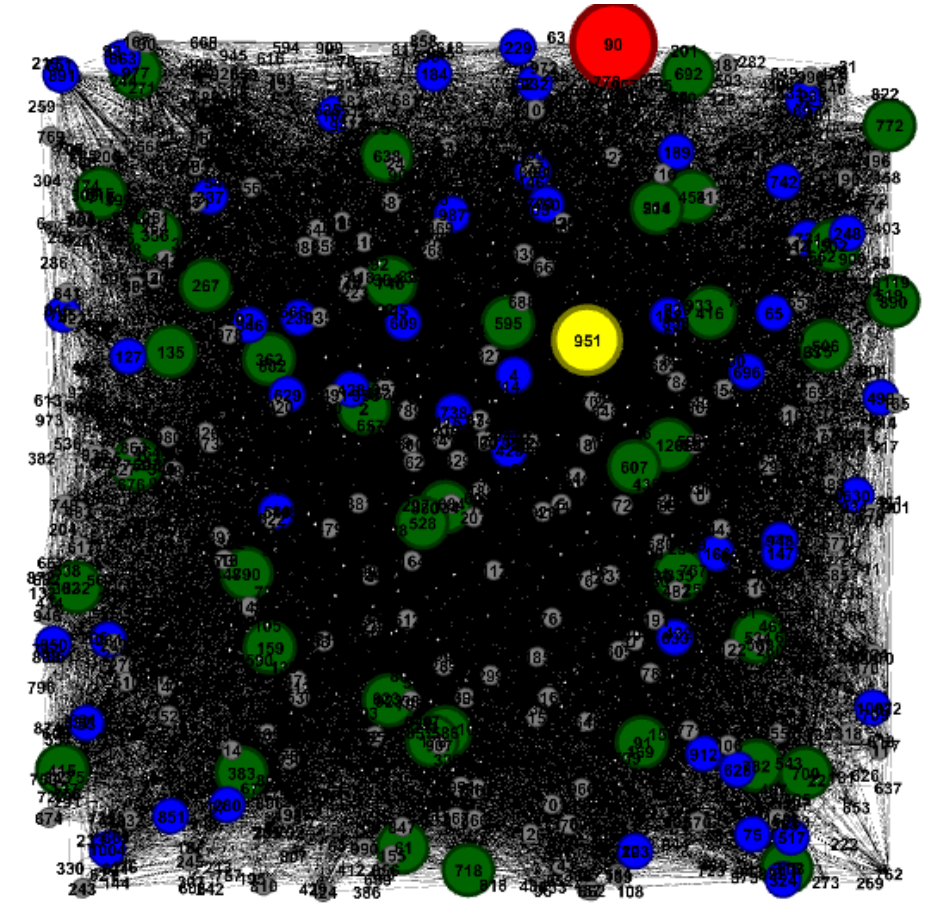


Рисунок 4.14 – Лідери мережі електронних листів метрики за навантаженістю

Після розрахунку усіх коефіцієнтів значимості за кожною з метрик та розподілення місць серед усіх вершин були виконані розрахунки і для сумісної метрики з такими ж коефіцієнтами, як і для попередньої мережі.

Результати зведені до таблиці 4.2.

Таблиця 4.2 – Перша п’ятірка вершин з коефіцієнтами за сумісною метрикою мережі електронних листів

| Номер вершини | Коефіцієнт значимості |
|---|---|
| 90 | 986 |
| 159 | 983,8 |
| 214 | 983 |
| 120 | 982,5 |
| 951 | 979,95 |

## 4.3 Результати досліджень для орієнтованої мережі Gnutella

У третьому модулі програми розглядалася орієнтована мережа файлообмінної системи Gnutella між хостами (серверами). Вона складається з 8717 вузлів та 31525 ребер.

Графічної ілюстрації у додатку Gephi приводити немає особливого сенсу через те, що вершин дуже велика кількість і неможливо досконало розібрати мітки на вершинах та чітко зобразити їх на рисунках.

Саме тому усі результати для кожної з розглянутих метрик було зведено до таблиць 4.3–4.8 з вказуванням перших п’яти вершин-лідерів за своїми метриками.

До таблиці 4.9 були занесені дані для сумісної метрики, що також включають у себе п’ятірку лідерів та коефіцієнти важливості кожної з метрик, що дорівнюють тим же значенням, що вказані для розрахунків у минулих пунктах.

Таблиця 4.3 – П’ятірка лідерів за показниковою метрикою для мережі Gnutella

| | |
|---|---|
| 6494 | 0,01319 |
| 356 | 0,00837 |
| 176 | 0,00826 |
| 31 | 0,00803 |
| 556 | 0,00803 |

Таблиця 4.4 – П’ятірка лідерів за метрикою близькості для мережі Gnutella

| | |
|---|---|
| 174 | 0,10622 |
| 556 | 0,10557 |
| 31 | 0,10535 |
| 99 | 0,10525 |
| 558 | 0,10502 |

Таблиця 4.5 – П’ятірка лідерів за проміжною метрикою для мережі Gnutella

| | |
|---|---|
| 878 | 0,00863 |
| 556 | 0,00848 |
| 357 | 0,00798 |
| 173 | 0,00787 |
| 567 | 0,00710 |

Таблиця 4.6 – П’ятірка лідерів метрики за власним вектором мережі Gnutella

| | |
|---|---|
| 296 | 0,14044 |
| 636 | 0,14019 |
| 99 | 0,13984 |
| 300 | 0,13277 |
| 299 | 0,13200 |

Таблиця 4.7 – П’ятірка лідерів метрики PageRank для мережі Gnutella

| | |
|---|---|
| 556 | 0,00092 |
| 841 | 0,00086 |
| 300 | 0,00086 |
| 558 | 0,00086 |
| 176 | 0,00086 |

Таблиця 4.8 – П’ятірка лідерів метрики за навантаженістю для мережі Gnutella

| | |
|---|---|
| 878 | 0,00830 |
| 556 | 0,00804 |
| 357 | 0,00762 |
| 173 | 0,00737 |
| 292 | 0,00687 |

Таблиця 4.9 – П’ятірка лідерів сумісної метрики для мережі Gnutella

| | |
|---|---|
| 556 | 8715,15 |
| 558 | 8709,55 |
| 174 | 8709,2 |
| 176 | 8707,9 |
| 353 | 8706,45 |

Як видно з отриманих результатів, сумісна метрика добре підходить для врахування усіх коефіцієнтів за кожною з метрик, тому що 556 вершина потрапила майже до усіх п’ятірок-лідерів.

# 5 АНАЛІЗ МОЖЛИВИХ ЗАСТОСУВАНЬ

Будь-яка організація, товариство чи підприємство є соціальною мережею. Саме тому аналіз таких соціальних мереж та пошук лідерів може застосовуватися у будь-яких галузях науки і техніки.

Соціологи використовують поняття «соціальна мережа» з початку XX століття для визначення наборів відношень між членами соціальних систем на усіх рівнях, від міжособових до міжнародних. Дослідження соціальних мереж вперше були представлені у роботах соціологів, саме тому в цій області вони мають найпоширеніше застосування.

Крім соціології пошук лідерів може застосовуватись, наприклад, у спортивній діяльності. У футболі, що є командною грою, однією з найголовніших задач тренерського персоналу є створення максимальної зіграності команди. Необхідно знайти варіанти найкращого стартового складу, а для цього необхідно знати як взаємодіють гравці одне з одним. Цю задачу допоможе виконати проміжна метрика центральності, яка відстежує зв’язки між вершинами (у даному випадку гравцями). За допомогою degree метрики, можна визначити гравця, до якого надходить найбільша кількість передач, або від якого йде найбільша небезпека для суперника. Всі ці дані можуть зіграти роль у визначенні складу команди на гру та відповісти на запитання про комунікабельність гравців між собою.

Неабияку роль аналіз соціальних мереж та пошук лідерів відіграє у розвідувальних та правоохоронних заходах. Ця техніка дозволяє аналітикам відобразити на карті нелегальну або приховану організацію, таку як шпіонське коло або вуличну банду. Агентство національної безпеки (NSA) використовує програми таємних масових систем електронного нагляду для генерації даних, необхідних для представлення цього типу аналізу у терористичних мережах, що мають відношення до національної безпеки. У процесі мережевого аналізу Агентство національної безпеки проводить пошук у глибину на три вузли. Після того як закінчилось початкове відображення соціальної мережі, виконується

аналіз для визначення структури та лідера мережі. Це дозволяє військовим або правоохоронним органам завдати ударів для захвату або знищенню найбільш значимих цілей, що призводить до порушення функціонування мережі.

У приватному секторі фірми використовують аналіз соціальних мереж для підтримки таких діяльностей як взаємодія та аналіз клієнтів, маркетинг та бізнес-аналітика. Використання аналізу соціальних мереж державним сектором включає в себе розвиток стратегій участі керівництва, аналіз індивідуальної та групової участі, використання засобів масової інформації та розв’язок проблем, що засновані на дослідженні спільнот [13].

Метрики центральності також можуть використовуватися у медицині як спосіб пошуку найбільш активних клітин, що можуть впливати на орган людини. Таким чином можна розпізнавати захворювання за допомогою штучного інтелекту.

У роботі Network Centrality Measures and Systemic Risk: An Application to the Turkish Financial Crisis турецькими дослідниками банківської сфери метрики центральності використовувались для розпізнавання системно важливих фінансових інституцій задля дослідження фінансової кризи 2000 року. Вони вивчали роль «Демірбанку» у падінні банківської системи за допомогою теорії графів та аналізу соціальних мереж.

Відома в усьому світі корпорація Google використовує метрики центральності для пошуку найбільш відвідуваних Web-сторінок. Саме ними була розроблена більш вдосконалена PageRank метрика. Google PageRank визначає індекс авторитетності веб-сторінок, що зв’язані між собою через посилання.

# ВИСНОВКИ

В рамках даної атестаційної роботи була розглянута проблема пошуку лідерів у соціальних мережах та визначені основні методи її розв’язання за допомогою теорії графів. Для визначення коефіцієнтів центральності були використані різноманітні метрики, такі як степенева (або показникова) метрика, метрика близькості, проміжна метрика, метрика за власним вектором, PageRank метрика та метрика за навантаженістю. Також було введено поняття сумісної метрики для визначення найважливіших вершин мережі, враховуючи вагу та значимість кожної з вище наведених метрик. Кожна з метрик була детально проаналізована та для неї визначені типи задач, до яких вона може бути застосовна. Був проведений порівняльний аналіз усіх метрик за різними критеріями та визначені їхні переваги та недоліки.

Для візуалізації та обчислення результатів за допомогою бібліотеки NetworkX та мови програмування Python були створені три програмні модулі, що відображують застосовність метрик центральності до графів різної розмірності та різних типів. Для кращої візуалізації було використано графічний додаток Gephi.

Отже, після проведення усіх досліджень над обраними мережами можна дійти висновку, що для кожного типу вирішуваної задачі підходить своя метрика, для визначення впливу – PageRank метрика, для розповсюдження інформації – метрика близькості, для контролю потоку інформації – проміжна або метрика за навантаженістю, але якщо необхідно врахувати декілька факторів одразу, то введена сумісна метрика якнайкраще підходить для визначення значимості тієї чи іншої вершини для мережі.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Лапыгин Ю.Н. Теория организации. Москва : “Инфра-М”, 2007. 311 с.

2. Киселев Д.Ю., Киселев Ю.В., Макарьев В.Д. Структурный анализ потоков данных (Data Flow Diagrams – DFD) : метод. указания. Самара : Изд-во СГАУ, 2014. 12 с.

3. Freeman L. C. Centrality in Social Networks Conceptual Clarification // Social Networks. 1978/79. № 1. P. 215–239.

4. Bonacich P. F. Power and Centrality: A Family of Measures // American Journal of Sociology. University of Chicago Press. 1987. Vol. 92. № 5. P. 1170–1182.

5. Borgatti P. S. Centrality and Network Flow // Social Networks. 2005. V. 27, № 1. P. 55–71.

6. Opsahl T., Skvoretz J. Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths // Social Networks. 2010. V. 32, № 3. P. 245–251.

7. Басов Н. В. Создание знания в сетевых коммуникативных структурах // Социологический журнал. 2014. № 1. С. 106–123.

8. Freeman L. C. A Set of Measures of Centrality Based on Betweenness // Sociometry. 1977. V. 40, № 1. P. 35–41.

9. Langville A., Meyer C. A survey of eigenvector methods of web information retrieval // SIAM Review. 2005. V. 47, № 1. P. 135–161.

10. On the Distributed Computation of Load Centrality and its Application to DV Routing / L. Maccari, L. Ghiro, A. Guerrieri, A. Montresor, R. Cigno // IEEE INFOCOM 2018 – IEEE Conference on Computer Communications. 2018. №1. P. 2582–2590.

11. Основна інформація про аналіз соціальних мереж. URL : https://ru.wikipedia.org/wiki/Анализ_социальных_сетей. (дата звернення: 09.10.2019).

12. Gephi. URL : https://uk.wikipedia.org/wiki/Gephi (дата звернення: 06.11.2019).

13. Відеокурс аналізу соціальних мереж. URL : https://www.youtube.

com/channel/UCqwvCUUnfyL_MdA_uQPZF_A (дата звернення:11.09.2019).

14. Kirichenko L., Bulakh V., Radivilova T. Fractal time series analysis of social network activities // 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), Kharkov, Ukraine. 2017. № 17. P. 456–459.

15. Анализ взаимозависимости временных рядов биткоина и активности сообществ в социальных сетях / Л. Кириченко, Т. Радивилова, В. Булах, В. Чакрян // International Journal Information Technologies & Knowledge. 2018. V. 12, № 1. С. 43–55.

16. Соціальні мережі для проведення досліджень. URL : http://www-personal.umich.edu/~mejn/netdata/ (дата звернення: 01.11.2019).

17. Основна інформація про метрики центральності. URL : http://letopisi.org/index.php/Анализ_социальных_сетей (дата звернення: 24.10.2019).