BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP.HCM

ĐỒ ÁN MÔN HỌC
MẠNG XÃ HỘI

PHÂN TÍCH MẠNG LƯỚI TƯƠNG TÁC VÀ DỰ ĐOÁN KHẢ NĂNG TƯƠNG TÁC GIỮA CÁC THÀNH VIÊN TRONG NHÓM VỀ TRÒ CHƠI ĐIỆN TỬ TRÊN NỀN TẢNG FACEBOOK

Ngành: **KHOA HỌC DỮ LIỆU**
Chuyên ngành: **KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn: **ThS. LÊ NHẬT TÙNG**

Sinh viên thực hiện

Nguyễn Công Hùng	MSSV: 2186400267	Lớp: 21DKHA1
Phan Phú Hào	MSSV: 2186400335	Lớp: 21DKHA1
Nguyễn Hữu Huy	MSSV: 2186400230	Lớp: 21DKHA1

TP. Hồ Chí Minh, 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP.HCM

ĐỒ ÁN MÔN HỌC
MẠNG XÃ HỘI

PHÂN TÍCH MẠNG LƯỚI TƯƠNG TÁC VÀ DỰ ĐOÁN KHẢ NĂNG TƯƠNG TÁC GIỮA CÁC THÀNH VIÊN TRONG NHÓM VỀ TRÒ CHƠI ĐIỆN TỬ TRÊN NỀN TẢNG FACEBOOK

Ngành: **KHOA HỌC DỮ LIỆU**
Chuyên ngành: **KHOA HỌC DỮ LIỆU**

Giảng viên hướng dẫn: **ThS. LÊ NHẬT TÙNG**

Sinh viên thực hiện

Nguyễn Công Hùng	MSSV: 2186400267	Lớp: 21DKHA1
Phan Phú Hào	MSSV: 2186400335	Lớp: 21DKHA1
Nguyễn Hữu Huy	MSSV: 2186400230	Lớp: 21DKHA1

TP. Hồ Chí Minh, 2025

LỜI CAM ĐOAN

Mọi thông tin được trình bày trong đồ án này là trung thực và khách quan, được thu thập và phân tích một cách cẩn thận, dựa trên các nguồn chính thống và đáng tin cậy. Bất kỳ thông tin hoặc ý kiến nào được trích dẫn từ các nguồn khác đều được nêu rõ nguồn gốc và được trích dẫn theo đúng quy định.

Đồ án này là công trình nghiên cứu độc lập của chúng tôi, chưa từng được công bố ở bất kỳ nơi nào khác. Chúng tôi cam đoan rằng đã tuân thủ đầy đủ các quy tắc và quy định, bao gồm cả việc tham khảo và sử dụng dữ liệu cũng như các công cụ nghiên cứu.

Chúng tôi hy vọng rằng đồ án "PHÂN TÍCH MẠNG LƯỚI TƯƠNG TÁC VÀ DỰ ĐOÁN KHẢ NĂNG TƯƠNG TÁC GIỮA CÁC THÀNH VIÊN TRONG NHÓM VỀ TRÒ CHƠI ĐIỆN TỬ TRÊN NỀN TẢNG FACEBOOK" sẽ cung cấp góc nhìn toàn diện về cách người chơi tương tác và kết nối trong môi trường trực tuyến, đồng thời đóng góp giá trị khoa học và thực tiễn cho nghiên cứu mạng xã hội và cộng đồng trò chơi điện tử.

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

CHƯƠNG 1. TỔNG QUAN

1.1. Giới thiệu đề tài

Sự phát triển của ngành công nghiệp trò chơi điện tử trên toàn thế giới đã chứng kiến một đà tăng trưởng mạnh mẽ trong những năm gần đây. Cùng với sự phổ biến của các nền tảng mạng xã hội, cộng đồng người chơi trò chơi điện tử, hay còn được gọi là game thủ, đã tìm đến các nhóm trên nền tảng Facebook như một không gian để kết nối, chia sẻ kinh nghiệm và tương tác với nhau. Điều này tạo nên một mạng lưới xã hội phong phú, nơi mà các mối quan hệ và tương tác giữa các thành viên đóng vai trò quan trọng trong việc duy trì và phát triển cộng đồng.

Trong bối cảnh này, việc tập trung nghiên cứu về liên kết mạng xã hội trở nên cần thiết để hiểu rõ hơn về cấu trúc và động lực tương tác của cộng đồng. Nhằm đạt được mục tiêu này, nghiên cứu sẽ áp dụng các phương pháp phân tích dữ liệu tương tác và dự đoán khả năng tương tác giữa các thành viên. Qua đó, chúng tôi hy vọng có thể cung cấp những hiểu biết giá trị về cách thức các thành viên tương tác trong cộng đồng người chơi trò chơi điện tử tại Việt Nam.

1.2. Ý nghĩa khoa học và thực tiễn của đề tài

1.2.1. Ý nghĩa khoa học

Đề tài bước đầu đề xuất phương pháp phân tích tương tác trong cộng đồng về trò chơi điện tử trên mạng xã hội, góp phần phân tích và dự đoán tương tác. Nghiên cứu kỳ vọng đóng góp vào sự phát triển của khoa học dữ liệu thông qua việc xây dựng bộ dữ liệu và phương pháp luận.

1.2.2. Ý nghĩa thực tiễn

Nghiên cứu cung cấp một số thông tin hữu ích trong việc tối ưu hóa tương tác và tăng cường gắn kết cộng đồng mạng xã hội. Đồng thời, kết quả có thể hỗ trợ các nhà phát triển và doanh nghiệp trong việc thiết kế tính năng và thuật toán gợi ý nhằm nâng cao trải nghiệm người dùng.

1.3. Mục tiêu, đối tượng và phạm vi nghiên cứu

1.3.1. Mục tiêu nghiên cứu

- Xây dựng bộ dữ liệu đồ thị hoàn chỉnh từ dữ liệu tương tác trong nhóm Facebook của người chơi game, bao gồm thông tin về bài đăng, bình luận và

phản ứng của thành viên

- Phân tích các chỉ số quan trọng của mạng tương tác và ứng dụng các thuật toán để dự đoán khả năng tương tác giữa các thành viên trong cộng đồng

1.3.2. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là mạng lưới tương tác xã hội được hình thành từ các hoạt động trong một nhóm Facebook của những người chơi trò chơi điện tử. Cụ thể, nghiên cứu tập trung vào các mối quan hệ tương tác giữa thành viên thông qua hoạt động bình luận và phản ứng (reactions) đối với bài viết.

Nghiên cứu sẽ phân tích các đặc trưng cấu trúc của mạng lưới, xác định các nhóm tương tác chính và vai trò của từng thành viên trong mạng thông qua các chỉ số và thuật toán phổ biến trong lĩnh vực phân tích mạng xã hội.

1.3.3. Phạm vi nghiên cứu

- **Phạm vi không gian:** Giới hạn trong phạm vi một nhóm Facebook cụ thể "Chém Gió về Fear and Hunger" (có Facebook Group ID là `811896080494851`).
- **Phạm vi thời gian:** Dữ liệu được thu thập trong khoảng thời gian từ ngày 01/01/2025 đến ngày 03/01/2025.
- **Phạm vi nội dung:** Tập trung vào cấu trúc mạng lưới và mối quan hệ tương tác, không bao gồm nội dung cụ thể của các bài viết, bình luận và yếu tố nhân khẩu học của các thành viên.

1.4. Cấu trúc báo cáo

Chương 1. Tổng quan: Giới thiệu đề tài, trình bày mục tiêu, đối tượng và phạm vi nghiên cứu.

Chương 2. Cơ sở lý thuyết: Giới thiệu các khái niệm, chỉ số mạng xã hội và thuật toán phát hiện cộng đồng, cùng các nghiên cứu liên quan.

Chương 3. Trực quan hóa dữ liệu: Mô tả quy trình thu thập, xử lý, phân tích và trực quan hóa dữ liệu, cùng việc phát hiện cộng đồng.

Chương 4. Huấn luyện mô hình: Xây dựng, đánh giá và phân tích các mô hình dự đoán xu hướng tương tác.

Chương 5. Kết luận: Tổng kết đóng góp, hạn chế và định hướng tương lai.

Tài liệu tham khảo: Danh mục các tài liệu sử dụng.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Phân tích tương tác mạng xã hội

2.1.1. Tổng quan

Phân tích tương tác mạng xã hội là một lĩnh vực nghiên cứu tập trung vào việc hiểu và đánh giá các mối quan hệ tương tác giữa các thành viên trong mạng lưới xã hội. Trong bối cảnh các nền tảng như Facebook, việc phân tích mạng lưới tương tác giúp khám phá hành vi người dùng, cấu trúc nhóm và dự đoán khả năng tương tác giữa các thành viên[1][2].

Mạng xã hội, đặc biệt là Facebook, là nơi kết nối và hỗ trợ các cộng đồng người chơi game thông qua các nhóm chuyên biệt. Tại nhóm "Chém Gió về Fear and Hunger", các thành viên tương tác bằng cách bình luận, phản ứng, và chia sẻ nội dung, tạo nên mạng lưới kết nối phức tạp. Những tương tác này góp phần xây dựng các mối quan hệ và thúc đẩy sự gắn kết trong cộng đồng.

Mạng lưới tương tác xã hội được mô tả thông qua các thành phần sau [3]:

- **Nút (Node):** Đại diện cho các thành viên trong nhóm, như người chơi hoặc người dùng Facebook.
- **Liên kết (Edge):** Đại diện cho tương tác giữa các thành viên, như gửi tin nhắn, bình luận hoặc chia sẻ bài viết.
- **Hướng (Direction):** Phân biệt giữa mạng có hướng (ví dụ: A tương tác với B) và vô hướng (ví dụ: A và B tương tác qua lại).

2.1.2. Các chỉ số chính

Một số chỉ số quan trọng trong phân tích mạng lưới tương tác bao gồm:

- **Mật độ đồ thị (Graph Density):** Đánh giá mức độ kết nối tổng thể của mạng bằng tỷ lệ giữa số liên kết thực tế và tối đa [4].

$$D = \frac{2k}{n(n-1)}$$

Trong đó:

  - k: Tổng số liên kết thực tế trong mạng.
  - n: Số lượng nút.

- **Bậc trung bình (Average Degree):** Số liên kết trung bình của mỗi thành viên, thể hiện mức độ tương tác [5].

$$\bar{k} = \frac{2k}{n}$$

- **Số đo trung tâm (Centrality Metrics):** Đánh giá vai trò của các thành viên trong mạng [6]:
  - **Số đo bậc trung tâm (Degree Centrality):** Số lượng tương tác mà một nút tham gia.
  - **Số đo trung tâm trung gian (Betweenness Centrality):** Mức độ một nút là cầu nối giữa các nút khác.
  - **Số đo trung tâm gần gũi (Closeness Centrality):** Khoảng cách trung bình từ một nút đến các nút khác.

2.1.3. Các thuật toán phân tích mạng xã hội

- **Thuật toán Louvain**: Thuật toán hiệu quả để phát hiện cộng đồng dựa trên tối ưu hóa mô đun (modularity) [7].
- **Thuật toán Label Propagation**: Thuật toán đơn giản và nhanh chóng để phân cụm các nút trong mạng [8].
- **Thuật toán Girvan-Newman**: Dựa trên việc loại bỏ các liên kết có giá trị trung tâm trung gian (betweenness centrality) cao để phát hiện cộng đồng [9].

2.2. Dự đoán khả năng tương tác mạng xã hội

2.2.1. Tổng quan về dự đoán khả năng tương tác

Dự đoán khả năng tương tác trong mạng xã hội là một nhánh quan trọng của phân tích mạng, nhằm dự báo khả năng xuất hiện hoặc củng cố mối liên kết giữa hai nút trong mạng lưới. Điều này có ý nghĩa quan trọng trong việc hiểu rõ hành vi tương tác, phát triển cộng đồng và tối ưu hóa hoạt động của mạng xã hội [10].

2.2.2. Phương pháp dự đoán liên kết (Link Prediction)

Dự đoán khả năng tương tác giữa các thành viên trong mạng lưới được thực hiện dựa trên các thuật toán phổ biến, bao gồm:

- **Common Neighbors:** Phương pháp dựa trên số lượng láng giềng chung (neighbors) giữa hai nút u và v [11]. Công thức:

$$S(u, v) = |N(u) \cap N(v)|$$

Trong đó, $N(u)$ và $N(v)$ là tập hợp các nút láng giềng của u và v

- **Jaccard Coefficient:** Đo lường mức độ tương đồng giữa hai nút bằng tỷ lệ giữa số láng giềng chung và tổng số láng giềng của cả hai nút [12]. Công thức:

$$S(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

- **Adamic/Adar:** Đánh giá mức độ gần gũi giữa hai nút dựa trên láng giềng chung, đồng thời giảm trọng số của các láng giềng chung có nhiều kết nối (ít giá trị phân biệt) [13]. Công thức:

$$S(u, v) = \sum_{w \in N(u) \cap N(v)} \frac{1}{\log(|N(w)|)}$$

Trong đó, w là các nút láng giềng chung của u và v, và $|N(w)|$ là số lượng kết nối của w

- **Random Forest:** Là một thuật toán học máy dựa trên việc kết hợp nhiều cây quyết định (Decision Trees) để đưa ra dự đoán chính xác hơn [14]. Phương pháp này sử dụng các đặc trưng của cặp nút, như Common Neighbors, Jaccard Coefficient, Adamic/Adar.

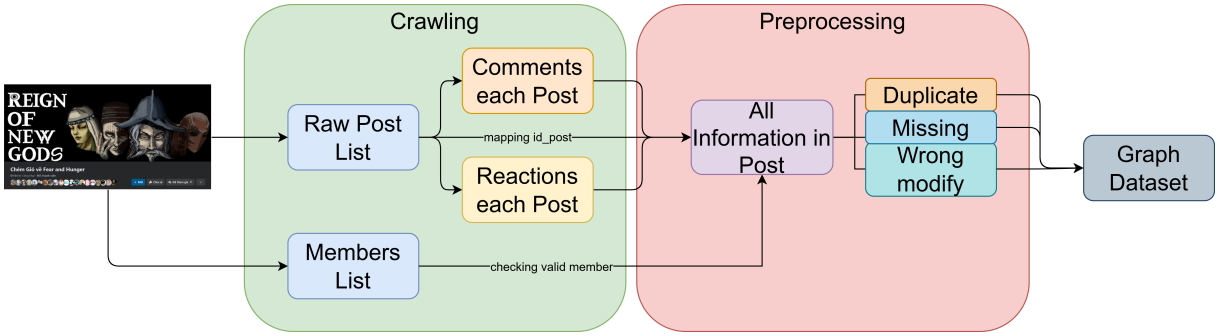
2.3. Các thư viện, công cụ được sử dụng

- **Selenium:** Tự động hoá trình duyệt để thu thập dữ liệu từ web động [15].
- **Pandas:** Xử lý và phân tích dữ liệu dạng bảng trong Python [16].
- **NetworkX:** Phân tích và trực quan hóa mạng lưới, bao gồm các thuật toán đo lường và phát hiện cộng đồng [17].
- **Scikit-learn:** Cung cấp các thuật toán học máy, công cụ huấn luyện và kiểm thử mô hình [18].
- **Gephi:** Phần mềm trực quan hóa và phân tích mạng lưới lớn [19].

CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN

3.1. Thu thập và tiền xử lý dữ liệu

Toàn bộ quy trình được triển khai bằng ngôn ngữ lập trình Python với sự hỗ trợ của thư viện Selenium, một công cụ mạnh mẽ cho phép tự động hóa việc tương tác với trình duyệt web và thư viện pandas cung cấp khả năng xử lý dữ liệu bảng.



Hình 3.1: Sơ đồ quy trình thu thập và tiền xử lý dữ liệu

Như được minh họa trong Hình 3.1, quá trình thu thập và tiền xử lý dữ liệu được thiết kế theo một quy trình hai giai đoạn có tính hệ thống và tự động hóa cao. Giai đoạn đầu tiên tập trung vào việc thu thập dữ liệu, trong khi giai đoạn thứ hai tập trung vào việc tiền xử lý:

- **Thu thập danh sách bài viết:** Quá trình thu thập danh sách bài viết được thực hiện thông qua cách cuộn trang để tải nội dung liên tục. Trong quá trình thu thập, hệ thống lưu trữ các thông tin như ID và nội dung bài viết, cùng với ID người đăng.
- **Thu thập tương tác từ bài viết:** Đối với mỗi bài viết, hệ thống thu thập phản ứng và bình luận trực thuộc bài viết.
- **Thu thập danh sách thành viên:** Song song với quá trình thu thập bài viết, hệ thống thu thập thông tin thành viên bao gồm thông tin cơ bản như ID, tên hiển thị.
- **Tổng hợp và làm sạch dữ liệu:** xử lý dữ liệu trùng lặp, xử lý dữ liệu thiếu, và sửa chữa dữ liệu không chính xác thông qua kiểm tra tính hợp lý.
- **Tạo tập dữ liệu đồ thị:** được lưu trữ dưới dạng tập tin Excel để thuận tiện cho việc phân tích và trực quan hóa trong các bước tiếp theo.

3.2. Phân tích mạng tương tác

3.2.1. Mục tiêu

Dựa trên dữ liệu đồ thị tương tác đã được thu thập và tiền xử lý, mục tiêu tiếp theo là phân tích cấu trúc và phát hiện các cộng đồng có liên kết chặt chẽ, qua đó phục vụ việc khám phá xu hướng tương tác giữa các thành viên.

Trong giai đoạn này, mạng lưới được biểu diễn dưới dạng đồ thị, trong đó mỗi thành viên được xem như một đỉnh (node) và mỗi mối tương tác được xem như một cạnh (edge). Tiếp đến, các thuật toán phát hiện cộng đồng được áp dụng nhằm xác định các nhóm có đặc trưng tương tác mật thiết.

3.2.2. Quy trình thực hiện

- **Xây dựng đồ thị:** Từ tập dữ liệu đã qua bước tiền xử lý, hệ thống trích xuất các cặp tương tác *(nguồn, đích)* làm đầu vào để tạo đồ thị. Các thuộc tính liên quan đến thành viên hoặc tần suất tương tác được đưa vào trong quá trình này nhằm giúp việc quan sát và đánh giá có chiều sâu hơn.
- **Lựa chọn và áp dụng các phương pháp phát hiện cộng đồng:**
  - *Label Propagation:* Được triển khai để xác định cộng đồng dựa trên nguyên lý các nhãn lan truyền qua lại trong đồ thị. Nhãn thống nhất sẽ phản ánh các vùng có kết nối chặt chẽ.
  - *Girvan–Newman:* Áp dụng cơ chế loại bỏ dần các liên kết "cầu nối" có tần suất xuất hiện cao trên đường đi ngắn nhất giữa các cặp đỉnh. Mỗi lần loại bỏ cạnh, đồ thị được tách dần thành các cụm nhỏ hơn.
  - *Louvain:* Triển khai phương pháp tối ưu modularity theo từng vòng lặp, gộp và rút gọn các đỉnh để hình thành cộng đồng. Đây là phương pháp tối ưu cho các mạng có quy mô lớn.
- **Trực quan hóa và đánh giá kết quả:**
  - Mỗi cộng đồng được biểu diễn bằng một màu sắc khác nhau trên đồ thị nhằm giúp quan sát sự phân bổ cộng đồng và mức độ gắn kết giữa các nhóm.
  - Bộ chỉ số *(số lượng cộng đồng, kích thước cộng đồng lớn nhất, kích thước cộng đồng nhỏ nhất)* được thống kê để so sánh tính hiệu quả, mức độ phân tách của từng phương pháp.

3.3. Dự đoán khả năng tương tác

3.3.1. Mục tiêu

Mục tiêu của giai đoạn này là dự đoán khả năng hình thành tương tác mới giữa các thành viên trong mạng (hay còn gọi là bài toán *dự đoán liên kết*). Để thực hiện, tập dữ liệu đồ thị được tách thành hai phần: một phần để huấn luyện (training) và phần còn lại để kiểm thử (testing).

3.3.2. Quy trình thực hiện

Các bước chính bao gồm:

- **Phân tích mạng:** Trước tiên, mạng lưới được phân tích sơ bộ, trích rút các chỉ số cơ bản như số lượng nút, số lượng cạnh và hệ số cụm (clustering coefficient). Qua đó, cho phép đánh giá tổng quan mức độ liên kết và mật độ kết nối trong hệ thống.
- **Chuẩn bị dữ liệu huấn luyện và kiểm thử:** Từ tập cạnh (edges) ban đầu, một tỉ lệ được chọn làm dữ liệu kiểm thử. Các cạnh này tạm thời bị loại khỏi mạng huấn luyện. Đồng thời, một số cặp đỉnh chưa có cạnh (tức là chưa tương tác) được lấy làm *ví dụ âm* để đánh giá khả năng dự đoán.
- **Rút trích đặc trưng:** Đối với mỗi cặp đỉnh (có thể là cạnh dương hoặc âm), ba loại đặc trưng phổ biến được tính toán:
  - *Số lượng láng giềng chung* (Common Neighbors)
  - *Hệ số Jaccard* (Jaccard Coefficient)
  - *Chỉ số Adamic-Adar* (Adamic-Adar)
- **Đánh giá các phương pháp truyền thống:** Các đặc trưng trên được trực tiếp sử dụng để dự đoán việc có cạnh hay không.
- **Xây dựng mô hình học máy:** Dữ liệu đặc trưng của cặp đỉnh trong mạng huấn luyện được dùng để huấn luyện một mô hình phân loại. Ở đây, mô hình được huấn luyện để phân biệt "có cạnh" (1) và "không có cạnh" (0) dựa trên các đặc trưng nêu trên. Sau khi huấn luyện, mô hình được đánh giá trên tập kiểm thử để kiểm chứng khả năng tổng quát.
- **Trình bày kết quả và so sánh:** Các kết quả được biểu diễn thông qua *bảng so sánh* và *đường cong ROC*. Điều này cho phép đánh giá trực quan độ nhạy (True Positive Rate) và độ đặc hiệu (False Positive Rate) của từng phương pháp, cũng như so sánh mô hình học máy với các phương pháp truyền thống.

3.3.3. Bộ đặc trưng dùng cho dự đoán

Trong quá trình dự đoán, ba đặc trưng sau được tính toán cho mỗi cặp đỉnh:

- **Common Neighbors:** Số lượng láng giềng chung cho biết mức độ "chồng lấn" trong mối quan hệ của hai đỉnh, giả định rằng nếu hai đỉnh đã có nhiều bạn chung thì khả năng kết nối giữa chúng cao hơn.
- **Jaccard Coefficient:** Đo lường độ tương đồng tương đối giữa tập láng giềng của hai đỉnh, bình thường hóa trên tổng số láng giềng duy nhất của chúng. Giá trị càng lớn thể hiện hai đỉnh càng có nhiều mối quan hệ chung so với quy mô tập láng giềng tổng.
- **Adamic-Adar:** Tăng mức trọng số cho các láng giềng chung "hiếm" (có bậc thấp), mang tính nhấn mạnh nếu hai đỉnh cùng kết nối đến các nút đặc biệt.

3.3.4. Các phương pháp dự đoán

Sau khi tách dữ liệu thành hai nhóm "có cạnh" (tạo mẫu dương) và "không có cạnh" (tạo mẫu âm), có hai hướng tiếp cận chính:

- **Phương pháp thống kê:** Dựa trực tiếp trên các chỉ số như Common Neighbors, Jaccard Coefficient, Adamic-Adar. Thông qua việc sử dụng một ngưỡng giá trị, có thể quyết định dự đoán liệu có hình thành cạnh tương tác hay không.
- **Phương pháp học máy:** Sử dụng mô hình phân loại (ví dụ, rừng ngẫu nhiên) với bộ đặc trưng {Common Neighbors, Jaccard, Adamic-Adar} làm đầu vào. Mô hình này được huấn luyện và kiểm thử qua quá trình chia dữ liệu (train/test) để nâng cao độ chính xác so với việc sử dụng trực tiếp từng chỉ số.

3.3.5. Đánh giá hiệu năng

Cả hai hướng tiếp cận đều được đánh giá trên cùng một tập kiểm thử với nhiều chỉ số:

- **AUC (Area Under the ROC Curve):** Đánh giá tổng quát khả năng phân tách giữa hai lớp (có cạnh hay không) của mô hình. Công thức tính AUC thường được tính toán từ đường cong ROC, không có công thức đơn giản cụ thể nhưng dựa trên tích phân của đường cong ROC:

$$AUC = \int_0^1 TPR(FPR)\, dFPR$$

- **Accuracy:** Tỷ lệ dự đoán đúng trên tổng số mẫu:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Precision:** Cho biết trong số các dự đoán là "có cạnh", có bao nhiêu dự đoán đúng:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall:** Cho biết trong số các mẫu thực sự "có cạnh", mô hình phát hiện được bao nhiêu:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1-score:** Trung bình điều hoà giữa Precision và Recall, phản ánh cân bằng chung của mô hình:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Trong đó:

- TP (True Positives): Số mẫu dự đoán đúng là có cạnh.
- FP (False Positives): Số mẫu dự đoán sai là dương (thực tế âm).
- TN (True Negatives): Số mẫu dự đoán đúng là âm.
- FN (False Negatives): Số mẫu dự đoán sai là âm (thực tế dương).

CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM

4.1. Tổng quát dữ liệu

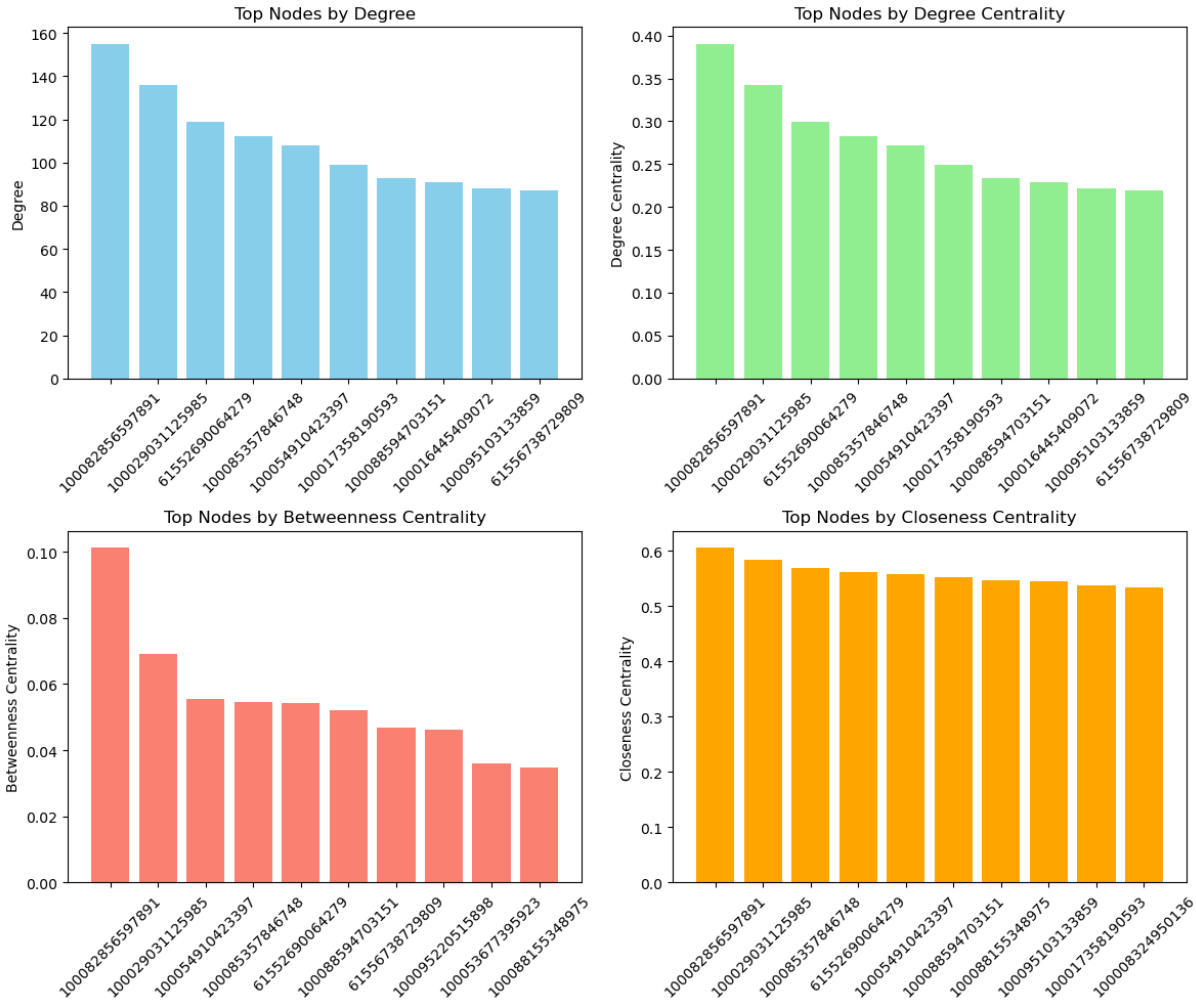
Dựa trên quá trình thu thập và xử lý dữ liệu một cách cẩn thận, đã tổng hợp được các thông tin cơ bản về mạng được trình bày trong Bảng 4.1 dưới đây:

Bảng 4.1: Thông tin cơ bản về mạng

Số nodes	398
Số cạnh	4011
Mật độ	0.0507702
Bậc trung bình	20.1558
Bậc lớn nhất	155

Mạng này bao gồm 398 nút và 4011 cạnh, tạo thành một cấu trúc phức tạp với mật độ là 0.0507702. Mặc dù mật độ này khá thấp, bậc trung bình của mỗi nút là 20.1558, cho thấy mỗi thành viên trung bình kết nối với khoảng 20 thành viên khác, phản ánh một mức độ liên kết cao. Đáng chú ý, có một thành viên trong mạng có bậc lên tới 155, làm nổi bật sự chênh lệch lớn về mức độ kết nối so với các thành viên khác. Sự chênh lệch này, cùng với số lượng cạnh lớn so với số nút, cho thấy sự tồn tại của một số nút trung tâm hoặc nổi bật, phản ánh một cấu trúc dày đặc và phức tạp trong mạng.

Phân tích sâu hơn về các số đo trung tâm của mạng này cho thấy những đặc điểm rõ rệt và tầm quan trọng của các nút trong việc duy trì và điều phối thông tin qua lại trong mạng.



Hình 4.1: Top 10 thành viên các số đo trung tâm lớn nhất

Bảng 4.2: Các số đo trung tâm của mạng

Số đo trung tâm	**Trung bình**
Degree Centrality	0.0508
Betweenness Centrality	0.0037
Closeness Centrality	0.4163

Các giá trị trình bày trong Hình 4.1 và Bảng 4.2, thể hiện rằng một số thành viên trong nhóm hoạt động rất năng nổ, thu hút được sự chú ý của nhiều thành viên khác trong nhóm. Điều này được minh chứng thông qua các chỉ số trung tâm như sau:

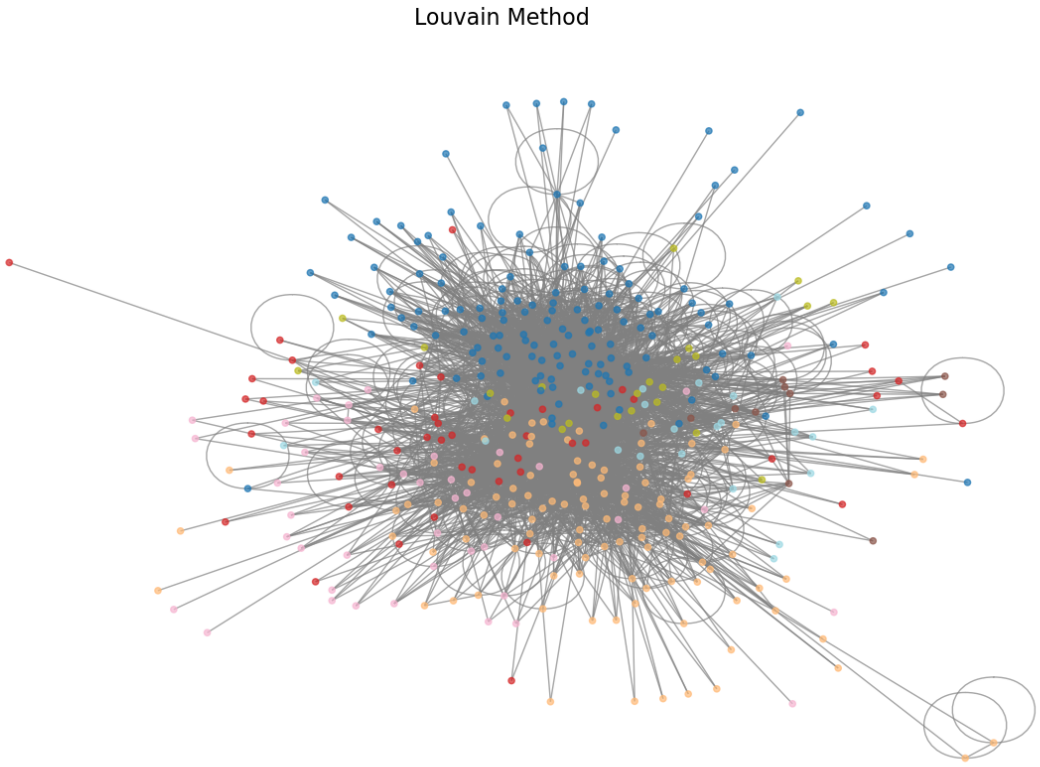
- **Degree Centrality (Trung bình 0.0508)**: Chỉ số này cho thấy một số thành viên có nhiều kết nối trực tiếp, biểu thị sự năng động trong việc tương tác với các thành viên khác. Mức độ trung bình thấp cho thấy đa số các thành viên có ít kết nối hơn, dẫn đến sự thiếu tương tác rộng rãi trong nhóm.

- **Betweenness Centrality (Trung bình 0.0037)**: Chỉ số này phản ánh vai trò của một số thành viên như là cầu nối quan trọng trong mạng, điều khiển dòng chảy thông tin. Mức trung bình thấp chỉ ra rằng vai trò này không phổ biến, hạn chế khả năng lan truyền thông tin và tương tác giữa các thành viên không trực tiếp kết nối.

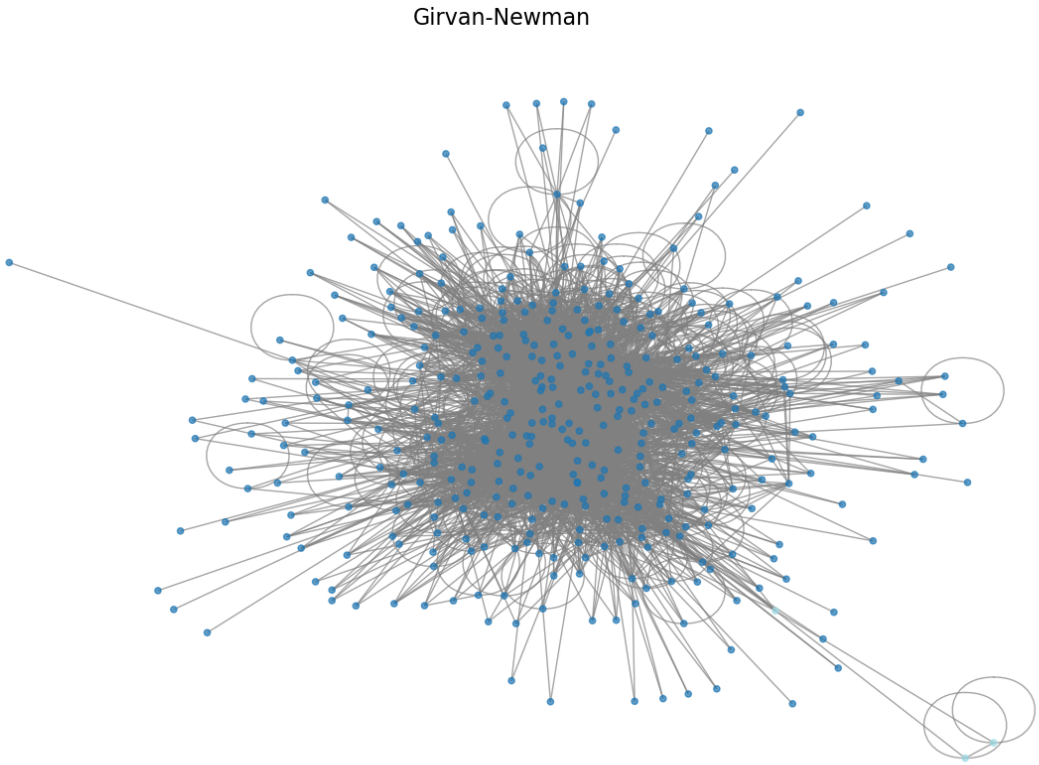
- **Closeness Centrality (Trung bình 0.4163)**: Chỉ số này đo lường khả năng một thành viên tiếp cận nhanh chóng tới các thành viên khác trong nhóm. Các giá trị cao cho thấy một số thành viên có khả năng tương tác nhanh và rộng, nhưng mức độ trung bình cho thấy không phải tất cả mọi người đều có khả năng này, làm giảm khả năng gắn kết chặt chẽ giữa các thành viên.

4.2. Phân tích mạng lưới tương tác

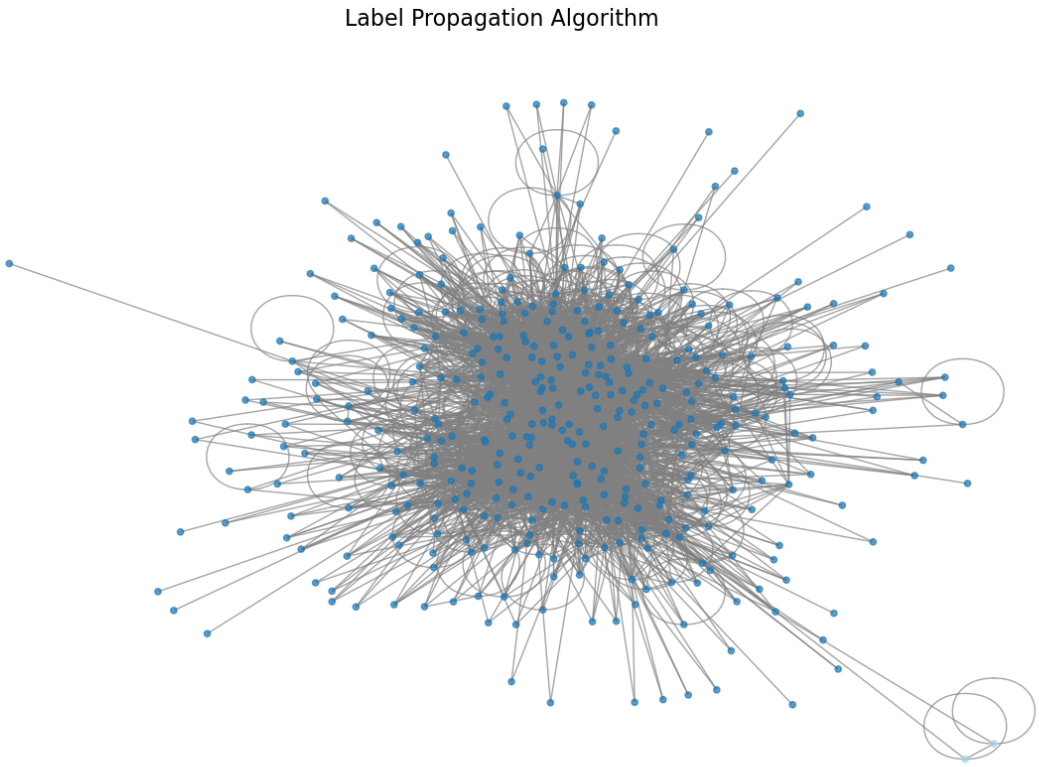
Sau khi phân tích các chỉ số trung tâm, rõ ràng là có một mức độ tương tác đáng kể giữa các thành viên trong nhóm. Dựa trên những tương tác này, chúng ta có thể nhận dạng các cộng đồng nhỏ hơn trong nhóm lớn hơn. Kết quả của quá trình phát hiện cộng đồng theo các phương pháp khác nhau được trình bày chi tiết ở Hình 4.2, 4.3, 4.4 và Bảng 4.3 như sau:



Hình 4.2: Biểu diễn kết quả của thuật toán Louvain



Hình 4.3: Biểu diễn kết quả của thuật toán Girvan-Newman



Hình 4.4: Biểu diễn kết quả của thuật toán Label Propagation

Bảng 4.3: Kết quả của các thuật toán

Phương pháp	**Số cộng đồng**	**Cộng đồng lớn nhất**	**Modularity**
Louvain	8	131	0.321978
Girvan-Newman	2	395	0.0024852
LPA	2	396	0.0014939

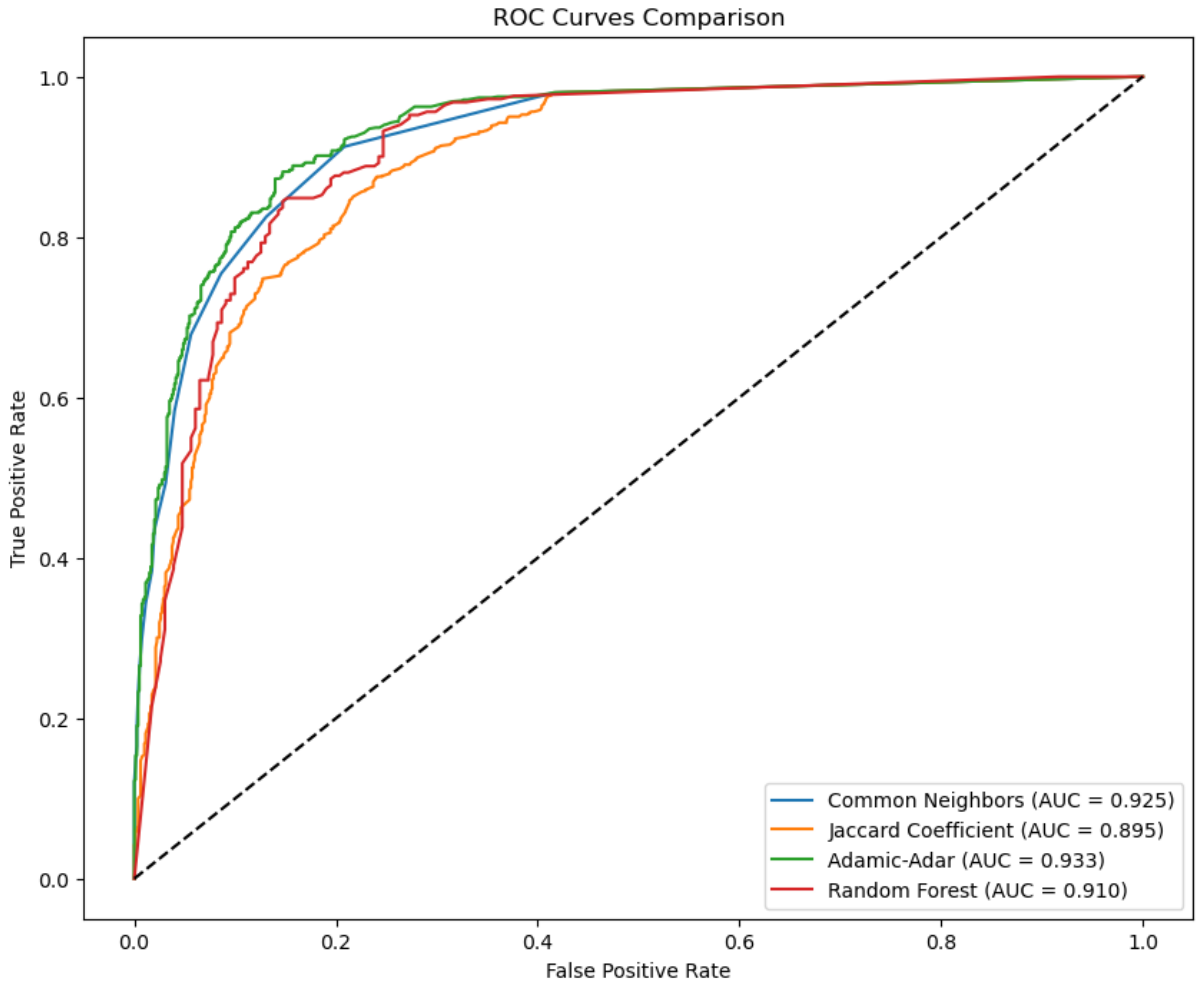
Các kết quả thu được từ các phương pháp phân cụm khác nhau cho thấy sự khác biệt đáng kể trong cách nhóm các thành viên. Trong số các phương pháp được áp dụng, phương pháp Louvain đã nổi bật lên nhờ khả năng phân định cộng đồng một cách hiệu quả hơn hẳn cụ thể là 8 cộng đồng và modularity là 0.32 so với các phương pháp còn lại. Điều này cũng cho thấy được Louvain sẽ phù hợp với các bộ dữ liệu có chủ đề tương tự cùng với việc tìm ra được các cộng đồng nhỏ hơn thể hiện được việc các thành viên trong nhóm sẽ có xu hướng tập trung về một chủ đề nhất định trong nhóm dẫn đến việc hình thành các tương tác tạo nên các cộng đồng nhỏ hơn.

4.3. Dự đoán tương tác

Từ kết quả phân cụm, chúng ta có thể dự đoán khả năng tương tác giữa các thành viên trong tương lai. Bằng cách áp dụng các mô hình dự đoán tương tác lên dữ liệu này, chúng ta có thể kiểm chứng độ chính xác của các dự đoán.

Bảng 4.4: So sánh các phương pháp học máy

Mô hình	**AUC**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Common Neighbors	0.925112	0.781796	0.701786	0.980050	0.817898
Jaccard Coefficient	0.894753	0.781796	0.701786	0.980050	0.817898
Adamic-Adar	0.932743	0.781796	0.701786	0.980050	0.817898
Random Forest	0.910462	0.844398	0.852000	0.848606	0.850299



Hình 4.5: Biểu diễn đường ROC của các mô hình

Kết quả từ mô hình dự đoán cho thấy Adamic-Adar và Common Neighbors đều có hiệu suất xuất sắc, với AUC lần lượt là 0.933 và 0.925. Điều này chứng tỏ khả năng phân biệt mạnh mẽ giữa các cặp thành viên có hoặc không có khả năng tương tác, nhấn mạnh vai trò của quá trình phân cụm trong dự đoán tương tác.

Mặc dù Random Forest có AUC thấp hơn (0.910), nhưng lại có F1-score cao nhất (0.850299), thể hiện sự cân bằng giữa độ chính xác và độ bao phủ. Điều này cho thấy Random Forest có thể là lựa chọn tối ưu trong các tình huống cần đến độ chính xác cao và khả năng phát hiện đầy đủ các tương tác.

CHƯƠNG 5. KẾT LUẬN

Trên nền tảng dữ liệu đã thu thập và phân tích, kết quả cho thấy mạng có sự khác biệt đáng kể về mức độ kết nối giữa các thành viên, với một nhóm nhỏ thể hiện tần suất tương tác vượt trội so với phần còn lại. Điều này góp phần tạo nên sự phân tầng trong cấu trúc chung, cũng như hình thành các cụm tương đối gắn kết.

Việc áp dụng các kỹ thuật dự báo cho thấy tiềm năng trong việc nhận diện xu hướng tương tác giữa các cá nhân. Những chỉ số đo lường hiệu suất của phương pháp dự báo gợi ý rằng thông qua việc kết hợp nguồn dữ liệu phong phú và thuật toán thích hợp, ta có thể đạt độ chính xác và khả năng bao quát tương đối cao.

Tuy nhiên, quy mô dữ liệu chưa lớn và thời gian thu thập tương đối ngắn cũng ảnh hưởng đến tính tổng quát của kết luận. Mặt khác, một số yếu tố khác như sở thích, nội dung thảo luận hoặc hồ sơ người dùng cũng chưa được xem xét đầy đủ. Các yếu tố này có thể đóng góp đáng kể vào việc lý giải nguyên nhân dẫn đến sự khác biệt trong mức độ kết nối, cũng như hỗ trợ đưa ra các đề xuất cụ thể để cải thiện chất lượng tương tác.

Trong tương lai, việc mở rộng phạm vi thu thập dữ liệu, kết hợp cả thông tin định lượng và định tính, hứa hẹn cung cấp bức tranh toàn diện hơn về hành vi của người tham gia. Đồng thời, hướng đi này cũng tạo cơ hội để triển khai những phương pháp khai thác dữ liệu sâu hơn, nhằm phát hiện những quy luật tiềm ẩn về cách thức thành viên tìm đến và duy trì kết nối. Chính những thông tin này sẽ trở thành nền tảng để đề ra các chính sách quản lý hoặc chiến lược xây dựng cộng đồng một cách hiệu quả và bền vững.

TÀI LIỆU THAM KHẢO

[1] D. Boyd and N. B. Ellison, "Social network sites as networked publics: Affordances, dynamics, and implications," in *A networked self: Identity, community, and culture on social network sites*, Z. Papacharissi, Ed., Routledge, 2011, pp. 39–58.

[2] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook "friends:" social capital and college students' use of online social network sites," *Journal of Computer-Mediated Communication*, vol. 12, no. 4, pp. 1143–1168, 2007.

[3] J. Scott, *Social network analysis: A handbook*. SAGE Publications, 2011.

[4] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge University Press, 1994.

[5] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[6] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[7] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, P10008, 2008.

[8] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," in *Physical Review E*, vol. 76, American Physical Society, 2007, p. 036 106.

[9] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[10] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[11] M. E. Newman, "Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality," *Physical review E*, vol. 64, no. 1, p. 016 132, 2001.

[12] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2006.

[13] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] SeleniumHQ, *Selenium webdriver*, https://www.selenium.dev, Accessed: 2023-12-01, 2023.

[16] T. P. D. Team, *Pandas: Powerful Python data analysis toolkit*. 2021, Accessed: 2023-12-01.

[17] N. Developers, *Networkx: High productivity software for complex networks*, https://networkx.org, Accessed: 2023-12-01, 2023.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] M. Bastian, S. Heymann, and M. Jacomy, *Gephi: An open graph visualization platform*, https://gephi.org, Accessed: 2023-12-01, 2009.