# PROJECT 1A: PHÂN CỤM MẠNG LƯỚI NGHỆ SĨ THAM GIA GAMESHOW VÀ PHÁT HIỆN NHÓM HỢP TÁC THƯỜNG XUYÊN

1st Mai Thanh Phuc, 2nd Hoang Thi Yen Nhi, 3rd Tran Trong Thanh, and Le Nhat Tung
HUTECH University, Vietnam
{MaiThanh Phuc, Hoang Thi Yen Nhi, Tran Trong Thanh}@hutech.edu.vn, and lenhattung@hutech.edu.vn

**Abstract**

Nghiên cứu này tập trung vào việc xây dựng và phân tích **mạng xã hội nghệ sĩ Việt Nam** thông qua việc ứng dụng các kỹ thuật **phân tích mạng xã hội (Social Network Analysis – SNA)** kết hợp với **các thuật toán phát hiện cộng đồng dựa trên học máy**. Dữ liệu được thu thập tự động bằng công cụ *Selenium* từ các trang **Wikipedia** chứa danh sách nghệ sĩ tham gia các chương trình gameshow truyền hình tại Việt Nam. Sau quá trình tiền xử lý và trích xuất tên nghệ sĩ, mạng lưới hợp tác được hình thành với **675 nút (nghệ sĩ)** và **55.262 cạnh**, phản ánh mối quan hệ đồng tham gia chương trình giữa các nghệ sĩ.

Các chỉ số mạng như *Degree Centrality*, *Betweenness Centrality*, *Closeness Centrality* và *PageRank* được tính toán để xác định mức độ ảnh hưởng của từng nghệ sĩ trong mạng. Kết quả cho thấy **Kim Tử Long**, **Hoài Linh** và **Hòa Minzy** là những nghệ sĩ có vị trí trung tâm và ảnh hưởng lớn nhất. Tiếp đó, bốn thuật toán phát hiện cộng đồng — **Louvain**, **Leiden**, **Spectral Clustering** và **Gaussian Mixture Model (GMM)** — được áp dụng để nhận diện các nhóm nghệ sĩ hợp tác thường xuyên. Trong đó, **Leiden** đạt giá trị *Modularity* cao nhất (0.3784), thể hiện khả năng phân tách rõ ràng các cộng đồng, trong khi **GMM** cho thấy sự linh hoạt khi phát hiện các cụm chồng lấn phản ánh thực tế hợp tác đa chiều giữa nghệ sĩ.

Các kết quả trực quan hóa cho thấy rõ cấu trúc cộng đồng trong mạng, đặc biệt là các nhóm hợp tác nổi bật như nhóm *Trấn Thành – Hari Won – Trường Giang – Hòa Minzy*. Nghiên cứu chứng minh tiềm năng của việc kết hợp phân tích mạng xã hội và học máy trong việc khám phá cấu trúc cộng đồng, mô hình hóa mối quan hệ hợp tác, và hỗ trợ phân tích xu hướng trong ngành giải trí Việt Nam.

**Index Terms**

Phân tích mạng xã hội, Phát hiện cộng đồng, Louvain, Leiden, Spectral Clustering, Gaussian Mixture Model (GMM), Mạng nghệ sĩ Việt Nam, Gameshow truyền hình, Selenium, Wikipedia, Modularity, Centrality, Học máy, Khoa học dữ liệu, Trực quan hóa mạng, Hợp tác nghệ sĩ, Mô hình cộng đồng.

## I. Giới thiệu

Trong kỷ nguyên truyền thông hiện đại, gameshow truyền hình không chỉ là một hình thức giải trí, mà còn là một không gian kết nối, giao lưu và hợp tác giữa các nghệ sĩ trong nhiều lĩnh vực khác nhau. Sự tham gia của họ trong hàng trăm chương trình truyền hình đã vô tình hình thành nên một mạng lưới xã hội phức tạp, nơi mỗi nghệ sĩ vừa là cá nhân sáng tạo độc lập, vừa là một nút kết nối trong một hệ sinh thái nghệ thuật rộng lớn. Việc phân tích mạng lưới này không chỉ giúp chúng ta hiểu rõ hơn về cấu trúc hợp tác, mức độ ảnh hưởng và xu hướng cộng tác giữa các nghệ sĩ, mà còn góp phần phản ánh bức tranh tổng thể của ngành giải trí Việt Nam trong bối cảnh hội nhập và số hóa.

Để khám phá cấu trúc cộng đồng trong mạng lưới nghệ sĩ, nghiên cứu này tập trung vào bài toán phân cụm mạng lưới nghệ sĩ tham gia gameshow tại Việt Nam, với mục tiêu phát hiện các nhóm nghệ sĩ có mối liên hệ chặt chẽ thông qua việc cùng xuất hiện trong các chương trình truyền hình. Nguồn dữ liệu được thu thập từ trang Wikipedia tiếng Việt, nơi tổng hợp danh sách các gameshow cùng với bảng thông tin về các nghệ sĩ tham gia từng chương trình. Quá trình thu thập được tự động hóa bằng thư viện Selenium, cho phép truy cập tuần tự vào từng đường liên kết (URL) của các gameshow, trích xuất nội dung từ các bảng dữ liệu chứa danh sách nghệ sĩ, và lưu trữ dữ liệu có cấu trúc phục vụ cho việc xây dựng mô hình mạng lưới.

Trong mô hình này, mỗi nút (node) biểu diễn một nghệ sĩ, và mỗi cạnh (edge) được hình thành khi hai nghệ sĩ cùng xuất hiện trong ít nhất một chương trình truyền hình. Trọng số của cạnh (nếu có) thể hiện mức độ hợp tác giữa hai nghệ sĩ, được tính dựa trên số lượng gameshow mà họ cùng tham gia. Mạng lưới thu được là một đồ thị phi hướng, có thể có trọng số, phản ánh mức độ tương tác, hợp tác và ảnh hưởng lẫn nhau giữa các nghệ sĩ trong hệ sinh thái gameshow Việt Nam.

Để phát hiện các cộng đồng (clusters) — tức là những nhóm nghệ sĩ có sự hợp tác mật thiết — nghiên cứu áp dụng ba thuật toán tiêu biểu trong lĩnh vực phân tích mạng xã hội (Social Network Analysis - SNA):

- Thuật toán Louvain : phương pháp tối ưu hóa chỉ số modularity nhằm xác định cấu trúc cộng đồng tối ưu, có khả năng mở rộng tốt cho các mạng lớn và phức tạp [1].
- Thuật toán Leiden : phiên bản cải tiến của Louvain, giúp đảm bảo tính ổn định và tính liên thông nội tại của các cụm, đồng thời khắc phục hạn chế về độ chính xác trong việc tách cụm nhỏ [2].

- Phương pháp Spectral Clustering : dựa trên phân tích phổ của ma trận Laplacian, cho phép phân tách mạng lưới thành các nhóm dựa trên đặc trưng hình học và mối quan hệ cấu trúc giữa các nút [3].
- Phương pháp Gaussian Mixture Model (GMM): dựa trên giả định rằng dữ liệu được tạo ra từ sự kết hợp của nhiều phân phối chuẩn (Gaussian) khác nhau. Thuật toán GMM sử dụng ước lượng cực đại kỳ vọng (EM – Expectation Maximization) để xác định tham số của từng phân phối và gán xác suất thuộc cụm cho mỗi phần tử. Phương pháp này cho phép phát hiện các cộng đồng với ranh giới mềm, tức là một nghệ sĩ có thể thuộc về nhiều nhóm hợp tác với xác suất khác nhau [4].

Sau khi áp dụng các thuật toán trên, mạng lưới được phân chia thành nhiều cộng đồng nghệ sĩ hợp tác thường xuyên. Mỗi cộng đồng phản ánh một nhóm tương tác hoặc "vòng tròn xã hội" trong thế giới gameshow Việt Nam — có thể là nhóm nghệ sĩ thường xuyên xuất hiện cùng nhau, nhóm thuộc cùng nhà sản xuất, hoặc cùng thể loại chương trình.

Kết quả của nghiên cứu không chỉ cung cấp cái nhìn toàn cảnh về cấu trúc xã hội của cộng đồng nghệ sĩ Việt Nam, mà còn có ý nghĩa thực tiễn trong việc:

- Phân tích xu hướng hợp tác và ảnh hưởng trong ngành giải trí.
- Xây dựng hệ thống gợi ý nghệ sĩ tiềm năng cho các chương trình truyền hình mới.
- Ứng dụng các mô hình mạng trong nghiên cứu truyền thông, phân tích văn hóa đại chúng và dữ liệu xã hội học.

Từ đó, nghiên cứu này không chỉ là bước tiếp cận công nghệ đối với lĩnh vực giải trí, mà còn góp phần minh chứng cho sức mạnh của phân tích dữ liệu và học máy trong việc khám phá cấu trúc xã hội ẩn bên trong các hoạt động nghệ thuật và truyền thông hiện đại.

## II. Nghiên cứu liên quan

Trong những năm gần đây, việc **phân tích và khám phá cấu trúc cộng đồng trong mạng xã hội và mạng hợp tác** đã trở thành một hướng nghiên cứu quan trọng trong lĩnh vực **phân tích mạng phức tạp (Complex Network Analysis)**. Các công trình liên quan tập trung vào việc **xác định các cụm cộng đồng, vai trò cá nhân và đặc trưng hợp tác** trong nhiều bối cảnh khác nhau như âm nhạc, truyền thông hay mã nguồn mở. Dưới đây là năm nghiên cứu tiêu biểu có mối liên hệ chặt chẽ với đề tài này.

### *A. Community Structures and Role Detection in Music Networks*

Teitelbaum et al. [5] là một trong những nghiên cứu đầu tiên áp dụng **phương pháp phát hiện cộng đồng** trong mạng nghệ sĩ âm nhạc. Tác giả xây dựng **đồ thị nghệ sĩ** dựa trên mức độ hợp tác và tương đồng âm nhạc, sau đó sử dụng **thuật toán Louvain** để xác định các cụm cộng đồng. Nghiên cứu nhấn mạnh vai trò của từng nghệ sĩ trong cộng đồng và chứng minh rằng mạng âm nhạc có đặc trưng **small-world** và **modularity cao** — những khái niệm có thể mở rộng áp dụng cho mạng nghệ sĩ gameshow Việt Nam.

### *B. Community Detection on Last.fm Artist Data*

Nhóm nghiên cứu SNAP của Đại học Stanford [6] tập trung vào việc **xây dựng và phân tích mạng nghệ sĩ dựa trên dữ liệu người nghe từ nền tảng Last.fm**. Dữ liệu được mô hình hóa thành **đồ thị vô hướng**, trong đó các nghệ sĩ được kết nối nếu họ có lượng người nghe trùng nhau. Các thuật toán **Louvain**, **Spectral Clustering** và **Label Propagation** được so sánh để đánh giá hiệu quả phát hiện cụm nghệ sĩ tương đồng. Kết quả cung cấp góc nhìn thực nghiệm hữu ích cho việc **so sánh giữa các thuật toán phân cụm** trong mạng nghệ sĩ nói chung.

### *C. Spotify Collaboration Network Analysis*

Di Matteo et al. [7] nghiên cứu **mạng hợp tác nghệ sĩ toàn cầu trên Spotify**, nơi các cạnh được xác định bởi **mối quan hệ hợp tác trong các bài hát ("featuring")**. Tác giả áp dụng thuật toán **Louvain** để phát hiện cụm cộng đồng và phân tích **đặc trưng như độ trung tâm, phân bố bậc và độ cụm**. Phương pháp và kết quả của nghiên cứu này cho thấy sự tương đồng với mô hình mạng hợp tác nghệ sĩ gameshow — nơi mối liên hệ được hình thành từ **sự xuất hiện chung** thay vì **hợp tác âm nhạc**.

### *D. Multilevel Clustering for Community Detection*

Zhang và Chen [8] đề xuất **kỹ thuật phân cụm đa tầng (multilevel clustering)** nhằm cải thiện độ chính xác và tính ổn định của việc phát hiện cộng đồng. Phương pháp kết hợp **đặc trưng cấu trúc (structural features)** và **ngữ nghĩa (semantic features)** của mạng, cho phép nhận diện cộng đồng ở nhiều cấp độ khác nhau. Công trình này đóng vai trò tham khảo cho việc **so sánh và tối ưu hóa kết quả phân cụm giữa các thuật toán Louvain, Leiden và Spectral** trong mạng nghệ sĩ gameshow.

### E. Community Detection in Networks: A User Guide

Fortunato và Hric [9] cung cấp **khung lý thuyết toàn diện** về các thuật toán phát hiện cộng đồng, từ **modularity optimization**, **spectral methods**, đến **statistical inference models**. Bài báo này giúp định hướng cho việc lựa chọn và đánh giá **chỉ số đo lường (modularity, conductance, silhouette score)** trong dự án, đồng thời củng cố nền tảng học thuật cho việc áp dụng Louvain, Leiden và Spectral Clustering trong bối cảnh mạng xã hội nghệ sĩ.

Tổng hợp lại, các nghiên cứu trên cho thấy rằng **việc phát hiện cộng đồng trong mạng hợp tác nghệ sĩ** không chỉ giúp hiểu rõ **cấu trúc và vai trò xã hội**, mà còn mang lại **ứng dụng thực tiễn** trong lĩnh vực **gợi ý hợp tác, phân tích ảnh hưởng và khám phá hành vi cộng đồng**. Dự án *"Phân cụm mạng lưới nghệ sĩ tham gia gameshow Việt Nam"* kế thừa hướng tiếp cận này, đồng thời mở rộng phạm vi nghiên cứu sang **ngữ cảnh văn hóa và giải trí Việt Nam**, với dữ liệu được **tự động thu thập từ Wikipedia bằng Selenium**, và **ứng dụng đồng thời ba thuật toán hiện đại — Louvain, Leiden, và Spectral Clustering** để phát hiện các cộng đồng nghệ sĩ hợp tác trong môi trường truyền thông đại chúng.

## III. Phương pháp

### A. Thu thập dữ liệu

Quá trình thu thập dữ liệu được triển khai có hệ thống nhằm xây dựng một cơ sở dữ liệu toàn diện về mạng lưới nghệ sĩ tham gia các gameshow tại Việt Nam. Dữ liệu được lấy chủ yếu từ Wikipedia, vốn là một nguồn tri thức mở có tính tin cậy cao trong việc ghi nhận thông tin về các chương trình truyền hình và tiểu sử nghệ sĩ. Quy trình thu thập và xử lý dữ liệu bao gồm tám giai đoạn chính như sau:

*Bước 1: Thu thập bảng dữ liệu từ Wikipedia*

Dữ liệu của nghiên cứu được thu thập từ **Wikipedia tiếng Việt**, nơi chứa thông tin về các chương trình gameshow và danh sách nghệ sĩ tham gia. Quá trình thu thập được thực hiện bằng công cụ **Selenium WebDriver** trong Python, giúp tự động hóa việc truy cập trang web, điều hướng giữa các liên kết và trích xuất dữ liệu từ các bảng HTML [10].

Cụ thể, chương trình được cấu hình để:

- Khởi tạo trình điều khiển trình duyệt Chrome thông qua thư viện `webdriver_manager`.
- Truy cập lần lượt vào từng đường dẫn (URL) tương ứng với các gameshow trên Wikipedia.
- Xác định phần tử `<table>` chứa danh sách nghệ sĩ.
- Trích xuất nội dung từ từng hàng (`<tr>`) và từng ô (`<td>`) trong bảng.

Đoạn mã minh họa quy trình trích xuất dữ liệu như sau:

```
table = driver.find_element(By.TAG_NAME, "table")
rows = table.find_elements(By.TAG_NAME, "tr")
for row in rows:
    cells = row.find_elements(By.TAG_NAME, "td")
    artists = [cell.text for cell in cells if cell.text.strip() != ""]
```

Listing 1. Trích đoạn mã thu thập dữ liệu bằng Selenium

Dữ liệu được lưu lại ngay sau khi trích xuất để đảm bảo không mất thông tin khi có lỗi xảy ra trong quá trình tự động. Cơ chế **WebDriverWait** được sử dụng để chờ tải nội dung trang, đồng thời xử lý các ngoại lệ như bảng rỗng, trang không tồn tại hoặc kết nối không ổn định.

Phương pháp này giúp việc thu thập dữ liệu diễn ra **hoàn toàn tự động**, giảm thiểu sai sót do thao tác thủ công, và có thể mở rộng để thu thập thông tin từ nhiều gameshow khác nhau.

Sau khi thu thập, dữ liệu được **lưu trữ dưới dạng bảng (CSV hoặc Excel)** để thuận tiện cho việc xử lý ở các giai đoạn sau. Mỗi dòng trong tệp dữ liệu đại diện cho một nghệ sĩ cùng tên chương trình tương ứng.

Các tệp dữ liệu được đặt tên và lưu vào thư mục cục bộ, giúp người nghiên cứu có thể dễ dàng truy cập và tái sử dụng. Cách lưu trữ này đảm bảo rằng dữ liệu thu được từ các lần chạy khác nhau sẽ **không bị ghi đè** mà được cộng dồn, tạo thành tập dữ liệu tổng hợp về nghệ sĩ và gameshow.

Kết quả của giai đoạn này là một **tập dữ liệu thô hoàn chỉnh**, bao gồm danh sách nghệ sĩ và các chương trình họ tham gia, sẵn sàng cho quá trình tiền xử lý và phân tích trong các bước tiếp theo.

*Bước 2: Chuyển đổi và chuẩn hóa dữ liệu gốc*

Dữ liệu ban đầu của dự án được lưu dưới định dạng **JSON Lines (.jsonl)** trong thư mục đầu vào. Mỗi file chứa nhiều dòng, mỗi dòng là một đối tượng JSON biểu diễn thông tin của một chương trình gameshow. Quy trình xử lý được thực hiện như sau:

- Duyệt qua toàn bộ các file `.jsonl` trong thư mục đầu vào.

- Sử dụng thư viện `json` để đọc từng dòng và chuyển đổi sang đối tượng Python.
- Tập hợp toàn bộ các bản ghi hợp lệ vào danh sách, sau đó tạo `DataFrame` bằng thư viện `pandas`.
- Xuất dữ liệu ra các file **CSV** tương ứng trong thư mục `file_csv`.

Trong quá trình xử lý, chương trình tự động phát hiện lỗi, bỏ qua các dòng không hợp lệ và hiển thị cảnh báo nếu file trống. Bước này tạo ra một tập dữ liệu **chuẩn hóa và đồng nhất** để phục vụ các bước xử lý tiếp theo.

*Bước 3: Nhận diện và trích xuất thực thể nghệ sĩ*

Các file CSV được tạo ra ở bước trước chứa nhiều cột thông tin khác nhau, không phải tất cả đều là tên nghệ sĩ. Để xác định chính xác các cột chứa tên, chương trình thực hiện quy trình sau:

- Xây dựng danh sách các từ khóa gợi ý như *"Tên"*, *"nghệ sĩ"*, *"thành viên"*, *"Khách mời"*, *"MC"*, *"Giám khảo"*, *"Thí sinh"*, v.v.
- Áp dụng biểu thức chính quy (regex) để nhận diện các chuỗi có cấu trúc giống tên tiếng Việt, ví dụ: `r"([A-ZÀ-ỴĐ][a-zà-ỹđ]+(?:+[A-ZÀ-ỴĐ][a-zà-ỹđ]+)1,4)"`.
- Nếu không tìm thấy cột phù hợp, chương trình sẽ tự động quét toàn bộ bảng để chọn ra cột có nhiều mẫu tên nhất.
- Sau khi xác định được các cột hợp lệ, chương trình tiến hành làm sạch: loại bỏ ký tự đặc biệt (`; / | •`), chuẩn hóa khoảng trắng, loại bỏ trùng lặp và sắp xếp theo thứ tự chữ cái.

Kết quả được lưu lại dưới dạng các file CSV riêng biệt trong thư mục `artists_csv`, mỗi file tương ứng với danh sách nghệ sĩ của một chương trình gameshow.

*Bước 4: Tổng hợp dữ liệu nghệ sĩ theo cấu trúc bảng*

Từ các file nghệ sĩ riêng lẻ, chương trình tiến hành tổng hợp dữ liệu theo hai cấu trúc:

- **Dạng rộng (Wide Format):** mỗi dòng tương ứng với một gameshow, các cột `artist_1`, `artist_2`, `artist_3`, ... lưu tên nghệ sĩ tham gia.
- **Dạng dài (Long Format):** mỗi dòng biểu diễn một cặp `(show, artist_name)`, phản ánh mối quan hệ nghệ sĩ – chương trình.

Hai file kết quả được tạo ra là `all_shows_artists_wide.csv` và `all_shows_artists_long.csv`, phục vụ cho các bước phân tích và mô hình hóa sau này.

*Bước 5: Làm sạch và thống nhất danh sách nghệ sĩ*

Từ bảng dữ liệu dạng rộng, chương trình trích xuất toàn bộ các cột `artist_*` và hợp nhất thành một danh sách duy nhất, loại bỏ giá trị trống và ký tự không hợp lệ. Danh sách được chuẩn hóa, sắp xếp và lưu vào file `all_unique_artists.csv`.

Sau đó, hệ thống kiểm tra trùng lặp để đảm bảo rằng không có nghệ sĩ nào bị ghi lặp trong danh sách tổng hợp.

*Bước 6: Thu thập thông tin mở rộng từ Wikipedia*

Dựa trên danh sách nghệ sĩ duy nhất, chương trình sử dụng các thư viện `requests` và `BeautifulSoup` để **tự động truy cập các trang Wikipedia của từng nghệ sĩ**. Quy trình bao gồm:

- Chuẩn hóa đường dẫn URL của từng nghệ sĩ bằng cách thay dấu cách bằng dấu gạch dưới (_) và mã hóa ký tự Unicode.
- Gửi yêu cầu HTTP kèm `User-Agent` định danh riêng để tránh bị chặn.
- Phân tích nội dung HTML bằng `BeautifulSoup` và trích xuất thông tin từ bảng `infobox`, bao gồm các trường như ngày sinh, nghề nghiệp, quốc tịch, v.v.
- Nếu trang không tồn tại, chương trình sử dụng API `opensearch` của Wikipedia để tìm trang gần đúng.
- Kết quả được tổng hợp và lưu thành file `wiki_infobox_artists.csv`.

*Bước 7: Phân loại nghệ sĩ theo mức độ dữ liệu*

Sau khi thu thập thông tin từ Wikipedia, dữ liệu được phân tách thành hai nhóm:

- `artists_with_data.csv`: chứa các nghệ sĩ có ít nhất một trường dữ liệu hợp lệ trong infobox.
- `artists_no_data.csv`: chứa các nghệ sĩ không có infobox hoặc không tìm thấy trang Wikipedia tương ứng.

Việc phân loại giúp đánh giá độ bao phủ của dữ liệu Wikipedia trong tập nghệ sĩ được thu thập.

*Bước 8: Xây dựng ma trận đồng xuất hiện của nghệ sĩ*

Dựa trên danh sách nghệ sĩ có dữ liệu (`artists_with_data.csv`) và bảng `(show, artist_name)`, chương trình xây dựng **ma trận đồng xuất hiện (Co-appearance Matrix)** bằng thư viện `pandas`. Các bước thực hiện bao gồm:

- Tạo bảng chéo (`crosstab`) giữa nghệ sĩ và gameshow để xác định nghệ sĩ xuất hiện trong từng chương trình.
- Nhân bảng chéo với chuyển vị của chính nó ($M \cdot M^T$) để thu được ma trận đồng xuất hiện.
- Gán giá trị 0 cho các phần tử trên đường chéo nhằm loại bỏ trường hợp nghệ sĩ trùng chính mình.
- Xuất kết quả cuối cùng ra file `artist_coappearance_matrix.csv`, biểu diễn số lần hai nghệ sĩ cùng xuất hiện trong các gameshow.

Tổng kết lại, toàn bộ quy trình trên bao gồm các bước: **chuyển đổi dữ liệu gốc → trích xuất tên nghệ sĩ → tổng hợp bảng → thu thập thông tin mở rộng → phân loại → tạo ma trận đồng xuất hiện**. Đây là chuỗi **xử lý và chuẩn hóa dữ liệu hoàn chỉnh**, đóng vai trò nền tảng cho các phân tích mạng và phát hiện cộng đồng ở giai đoạn tiếp theo.

*B. Phân tích mạng lưới và phát hiện cộng đồng nghệ sĩ*

*1. Xây dựng đồ thị mạng lưới nghệ sĩ*

Từ ma trận đồng xuất hiện của nghệ sĩ được tạo ra ở giai đoạn trước, chương trình sử dụng thư viện `NetworkX` để xây dựng **đồ thị mạng xã hội nghệ sĩ**. Trong đó, mỗi **nút (node)** đại diện cho một nghệ sĩ, và mỗi **cạnh (edge)** giữa hai nút biểu thị mối quan hệ hợp tác – tức là hai nghệ sĩ đã từng cùng tham gia ít nhất một gameshow. Trọng số của cạnh thể hiện số lần hai nghệ sĩ cùng xuất hiện trong các chương trình đó.

Để loại bỏ trùng lặp, chương trình chỉ thêm các cạnh khi chỉ số hàng nhỏ hơn chỉ số cột (`i < j`) và trọng số (`weight`) lớn hơn 0. Kết quả thu được là một **đồ thị vô hướng có trọng số**, phản ánh trực quan mối quan hệ hợp tác trong giới nghệ sĩ Việt Nam.

*2. Tính toán các chỉ số đặc trưng của mạng*

Sau khi xây dựng đồ thị, chương trình tiến hành tính toán các **chỉ số mô tả đặc trưng của mạng lưới** (network metrics) nhằm hiểu rõ hơn cấu trúc tổng thể của hệ thống hợp tác. Các chỉ số bao gồm:

- **Số lượng nút và cạnh:** thể hiện quy mô của mạng xã hội.
- **Mật độ đồ thị (Density):** đo mức độ kết nối giữa các nghệ sĩ.
- **Độ trung bình (Average Degree):** số lượng kết nối trung bình của một nghệ sĩ.
- **Hệ số phân cụm trung bình (Average Clustering Coefficient):** đo mức độ các nghệ sĩ có xu hướng hợp tác trong cùng nhóm.
- **Số lượng thành phần liên thông:** thể hiện số nhóm nghệ sĩ riêng biệt.

Bên cạnh đó, chương trình còn tính các **chỉ số trung tâm (Centrality)** – giúp đánh giá tầm quan trọng của từng nghệ sĩ trong mạng lưới:

- **Degree Centrality:** số lượng kết nối trực tiếp của nghệ sĩ (phản ánh độ phổ biến).
- **Betweenness Centrality:** mức độ nghệ sĩ đóng vai trò cầu nối giữa các nhóm.
- **Closeness Centrality:** khả năng nghệ sĩ tiếp cận nhanh đến người khác thông qua số bước trung bình nhỏ nhất.
- **PageRank:** mức độ ảnh hưởng tổng thể dựa trên mối quan hệ lan truyền giữa các nghệ sĩ.

Các chỉ số này được lưu vào cấu trúc dữ liệu `metrics`, phục vụ cho việc phân tích sâu hơn và trực quan hóa sau đó.

*3. Trực quan hóa mạng lưới và các chỉ số*

Để minh họa cho các kết quả phân tích, chương trình sử dụng `Matplotlib` để trực quan hóa dữ liệu. Các dạng biểu đồ bao gồm:

- **Phân phối Degree:** histogram mô tả số lượng nghệ sĩ theo mức độ kết nối.
- **Biểu đồ so sánh các chỉ số tổng quát:** như mật độ, độ trung bình, hệ số clustering và closeness.
- **Bản đồ mạng (Network Graph):** trong đó màu sắc của từng nút thể hiện giá trị của một chỉ số trung tâm.

Ví dụ:

- Màu đỏ: nghệ sĩ có Degree cao (kết nối nhiều).
- Màu xanh: nghệ sĩ có Betweenness cao (cầu nối quan trọng).
- Màu xanh lá: nghệ sĩ có Closeness cao (tiếp cận nhanh).
- Màu tím: nghệ sĩ có PageRank cao (ảnh hưởng mạnh).

Nhờ trực quan hóa, ta có thể dễ dàng nhận diện những nghệ sĩ giữ vai trò trung tâm hoặc có tầm ảnh hưởng lớn trong mạng.

*4. Xác định nghệ sĩ quan trọng nhất*

Từ các chỉ số trung tâm đã tính, chương trình trích xuất **Top 3 nghệ sĩ** có giá trị cao nhất ở mỗi chỉ số (Degree, Betweenness, Closeness, PageRank). Ngoài ra, hệ thống cũng xác định nghệ sĩ nổi bật nhất theo từng thước đo — ví dụ:

- Nghệ sĩ có nhiều kết nối nhất.
- Nghệ sĩ đóng vai trò cầu nối giữa các nhóm.
- Nghệ sĩ có tầm ảnh hưởng lớn nhất trên toàn mạng.

Các nghệ sĩ này được so sánh chi tiết qua bốn chỉ số, giúp làm rõ sự khác biệt về vai trò giữa "nghệ sĩ trung tâm", "nghệ sĩ kết nối" và "nghệ sĩ ảnh hưởng".

*5. Phát hiện cộng đồng nghệ sĩ*

Một bước quan trọng trong phân tích mạng xã hội là **phát hiện cộng đồng (community detection)** – nhằm khám phá cấu trúc ẩn bên trong mạng. Trong nghiên cứu này, chương trình áp dụng **ba thuật toán phân cụm cộng đồng phổ biến: Louvain, Leiden và Spectral Clustering**. Các thuật toán này được chọn vì chúng vừa hiệu quả với đồ thị lớn, vừa phản ánh được các nhóm hợp tác tự nhiên trong dữ liệu thực tế.

*a) Thuật toán Louvain.:* Thuật toán Louvain [11] là phương pháp dựa trên **tối ưu hóa chỉ số Modularity**, chia mạng thành các cộng đồng sao cho các kết nối nội bộ trong cụm dày đặc hơn kết nối giữa các cụm. Quá trình gồm hai giai đoạn lặp lại: (1) di chuyển từng nút vào cụm giúp tăng modularity, và (2) nén các cụm thành siêu nút để tiếp tục tối ưu. Kết quả là các nhóm nghệ sĩ có mối quan hệ hợp tác mạnh được hình thành tự nhiên.

*b) Thuật toán Leiden.:* Leiden [12] là phiên bản cải tiến của Louvain, giúp khắc phục nhược điểm là đôi khi Louvain sinh ra cụm không liên thông hoặc chưa tối ưu hoàn toàn. Leiden đảm bảo rằng mỗi cộng đồng phát hiện được là liên thông, đồng thời cải thiện tính ổn định và tốc độ xử lý. Trong mạng lưới nghệ sĩ Việt Nam, Leiden cho phép nhận diện chính xác hơn các nhóm nghệ sĩ thường xuyên cùng xuất hiện trong nhiều chương trình.

*c) Phân cụm bằng Spectral Clustering.:* Spectral Clustering [13] là phương pháp dựa trên **phân tích phổ của ma trận Laplacian** của đồ thị. Thuật toán trích xuất các vector riêng (eigenvectors) đại diện cho đặc trưng phổ của mạng, sau đó sử dụng `K-Means` để chia các nút thành $k$ cụm. Trong nghiên cứu này, giá trị $k$ được chọn dựa trên số cụm của Louvain để đảm bảo tính so sánh công bằng giữa các phương pháp. Spectral Clustering có ưu điểm trong việc phát hiện cấu trúc cộng đồng phức tạp, đặc biệt khi mạng có nhiều mối liên kết chồng chéo.

*d) Phân cụm bằng Gaussian Mixture Model (GMM)::* Gaussian Mixture Model [4] là phương pháp phân cụm dựa trên mô hình xác suất, giả định rằng dữ liệu được sinh ra từ sự kết hợp của nhiều phân phối chuẩn (Gaussian) khác nhau. Thuật toán GMM sử dụng cơ chế ước lượng cực đại kỳ vọng (EM – Expectation Maximization) để ước lượng các tham số của từng phân phối và xác định xác suất mỗi phần tử thuộc về một cụm. Không giống như các phương pháp phân cụm cứng, GMM cho phép phân cụm mềm (soft clustering), nghĩa là một nghệ sĩ có thể thuộc về nhiều cộng đồng khác nhau với các xác suất khác nhau. Trong nghiên cứu này, số cụm được chọn tương ứng với kết quả từ thuật toán Louvain để đảm bảo tính so sánh công bằng giữa các phương pháp. Phương pháp GMM đặc biệt hiệu quả trong việc phát hiện các cấu trúc cộng đồng phức tạp, nơi ranh giới giữa các nhóm không hoàn toàn tách biệt.

*6. So sánh kết quả giữa bốn thuật toán*

Sau khi chạy cả bốn thuật toán, chương trình so sánh kết quả theo các tiêu chí:

- **Số lượng cụm được phát hiện.**
- **Kích thước trung bình, nhỏ nhất và lớn nhất của các cụm.**
- **Giá trị chỉ số Modularity.**

Kết quả thực nghiệm cho thấy mỗi thuật toán mang lại đặc trưng riêng trong việc mô tả cấu trúc cộng đồng của mạng lưới nghệ sĩ Việt Nam. Thuật toán **Louvain** thường cho kết quả ổn định và các cụm phân bố tương đối cân bằng. Phương pháp **Leiden** thể hiện khả năng phân tách cộng đồng rõ ràng và hiệu quả hơn, tạo ra các nhóm liên kết chặt chẽ và ít bị chồng lấn. Trong khi đó, **Spectral Clustering** có xu hướng hình thành các cụm nhỏ hơn, phản ánh chi tiết các mối quan hệ hợp tác cục bộ. Đáng chú ý, **Gaussian Mixture Model (GMM)** cho phép phát hiện các cộng đồng với ranh giới mềm, mô tả tốt hiện tượng nghệ sĩ tham gia nhiều nhóm khác nhau, phù hợp với tính chất linh hoạt và đa dạng của môi trường gameshow Việt Nam.

*7. Nhận diện nhóm hợp tác thường xuyên nhất*

Cuối cùng, chương trình xác định **cụm nghệ sĩ có tần suất hợp tác mạnh nhất** bằng cách tính tổng trọng số các cạnh nội bộ trong từng cộng đồng. Cụm có tổng trọng số lớn nhất được xem là **nhóm hợp tác nổi bật nhất**, phản ánh sự gắn bó chặt chẽ giữa các nghệ sĩ trong các gameshow.

Chương trình xuất ra:

- Tên cụm cộng đồng.

- Số lượng nghệ sĩ trong cụm.
- Tổng số lần hợp tác nội bộ và tỷ lệ so với toàn mạng.
- Danh sách nghệ sĩ trong cụm đó.

Kết quả này giúp nhận diện rõ **các nhóm nghệ sĩ thường xuyên cùng hợp tác**, đồng thời là cơ sở cho việc nghiên cứu mối quan hệ xã hội và cấu trúc hợp tác trong ngành giải trí Việt Nam.

*8. Tóm tắt*

Toàn bộ quy trình phân tích được thực hiện qua các bước:

1) Xây dựng đồ thị mạng lưới nghệ sĩ.
2) Tính toán các chỉ số mạng và trung tâm.
3) Trực quan hóa cấu trúc mạng và xác định nghệ sĩ ảnh hưởng.
4) Áp dụng ba thuật toán phát hiện cộng đồng: **Louvain**, **Leiden**, **Spectral Clustering**,**GMM** .
5) So sánh và đánh giá các kết quả.
6) Nhận diện nhóm hợp tác nổi bật.

Thông qua quy trình này, hệ thống đã cung cấp một góc nhìn toàn diện và định lượng về **cấu trúc, vai trò và mức độ liên kết giữa các nghệ sĩ Việt Nam trong mạng lưới gameshow**, góp phần làm sáng tỏ các mô hình hợp tác trong lĩnh vực giải trí.

### C. Các chỉ số đánh giá

Để đánh giá hiệu quả của ba thuật toán Louvain, Leiden và Spectral Clustering, nghiên cứu sử dụng một số chỉ số chất lượng phân cụm như sau:

- **Modularity (Q):** đo mức độ liên kết chặt chẽ giữa các nút trong cùng cộng đồng so với toàn mạng. Giá trị $Q$ càng cao (thường $> 0.3$) cho thấy cấu trúc cộng đồng càng rõ ràng, được tính theo công thức:

$$Q = \frac{1}{2m} \sum_{i,j} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

  trong đó $A_{ij}$ là trọng số cạnh giữa $i$ và $j$, $k_i$ là bậc của nút $i$, $m$ là tổng số cạnh và $\delta(c_i, c_j)$ bằng 1 nếu hai nút thuộc cùng cộng đồng.
- **Số lượng cụm:** thể hiện tổng số nhóm nghệ sĩ được phát hiện, phản ánh mức độ phân tách của mạng lưới.
- **Kích thước cụm:** bao gồm kích thước trung bình, cụm nhỏ nhất và cụm lớn nhất, giúp mô tả sự phân bố quy mô giữa các cộng đồng.
- **Mật độ nội cụm (Intra-cluster Density):** phản ánh mức độ liên kết giữa các nghệ sĩ trong cùng cộng đồng; cụm có mật độ cao biểu hiện nhóm hợp tác chặt chẽ.
- **Phân tích nhóm nổi bật:** xác định cộng đồng có tổng trọng số hợp tác nội bộ lớn nhất, thường đại diện cho nhóm nghệ sĩ hợp tác thường xuyên trong nhiều chương trình.

Từ các chỉ số trên, kết quả cho thấy:

- **Louvain** đạt modularity cao và tạo ra các cụm ổn định.
- **Leiden** cho cộng đồng liên thông tốt và kết quả nhất quán hơn.
- **Spectral** phát hiện rõ các nhóm nhỏ, có tính tách biệt cao.
- **GMM** cho phép nhận diện các cộng đồng chồng lấn, phản ánh rõ sự linh hoạt trong mối quan hệ hợp tác giữa các nghệ sĩ.

## IV. Thiết lập thí nghiệm

### A. Môi trường thực nghiệm

Toàn bộ quá trình thực nghiệm được thực hiện trên máy tính cá nhân với cấu hình như sau:

- **Hệ điều hành:** Windows 11 64-bit
- **Bộ xử lý:** Intel Core i7
- **RAM:** 16 GB
- **Ngôn ngữ lập trình:** Python 3.13.2
- **Thư viện chính sử dụng:**
  - `pandas` – xử lý dữ liệu dạng bảng;
  - `networkx` – xây dựng và phân tích mạng lưới;
  - `matplotlib` – trực quan hóa kết quả;
  - `community` (python-louvain) – thuật toán Louvain;

- `igraph` và `leidenalg` – thuật toán Leiden;
- `scikit-learn` – thuật toán Spectral Clustering;
- `numpy` – xử lý ma trận và phép toán tuyến tính.

### *B. Dữ liệu đầu vào*

Dữ liệu được trích xuất tự động từ **Wikipedia tiếng Việt** bằng thư viện `Selenium` và `BeautifulSoup`. Tập dữ liệu bao gồm danh sách các *nghệ sĩ tham gia gameshow*, được xử lý và lưu dưới hai định dạng:

- Tệp `all_shows_artists_long.csv` chứa cặp `(tên gameshow, nghệ sĩ)`.
- Ma trận đồng xuất hiện `artist_coappearance_matrix.csv` biểu diễn số lần hai nghệ sĩ cùng xuất hiện trong một chương trình.

### *C. Xây dựng mạng lưới*

Từ ma trận đồng xuất hiện, một **đồ thị vô hướng có trọng số** được xây dựng với cấu trúc:

- **Node:** đại diện cho mỗi nghệ sĩ.
- **Edge:** tồn tại khi hai nghệ sĩ cùng tham gia ít nhất một gameshow.
- **Trọng số (weight):** biểu thị số lần hai nghệ sĩ xuất hiện chung.

Đồ thị được tạo bằng `networkx.Graph()` với tổng số $N$ nút và $E$ cạnh, phản ánh mức độ hợp tác giữa các nghệ sĩ trong toàn bộ hệ thống gameshow.

### *D. Cấu hình thực nghiệm và thuật toán*

Ba thuật toán phát hiện cộng đồng được triển khai và so sánh trong nghiên cứu này gồm:

1) **Louvain** – thuật toán phân cụm dựa trên tối ưu hóa chỉ số modularity [11].
2) **Leiden** – mở rộng từ Louvain, đảm bảo tính liên thông trong mỗi cụm [12].
3) **Spectral Clustering** – dựa trên phân tích phổ của ma trận Laplacian [13].
4) **GMM** - dựa trên mô hình hỗn hợp Gaussian, cho phép mô tả phân bố dữ liệu bằng nhiều thành phần chuẩn khác nhau [4].

Số cụm $k$ trong thuật toán Spectral được chọn bằng với số cụm phát hiện được từ Louvain nhằm đảm bảo tính so sánh nhất quán.

### *E. Quy trình thực nghiệm*

Quy trình thực nghiệm bao gồm các bước chính sau:

1) Tạo ma trận đồng xuất hiện từ dữ liệu gameshow.
2) Xây dựng đồ thị nghệ sĩ bằng thư viện `NetworkX`.
3) Áp dụng lần lượt bốn thuật toán Louvain, Leiden và Spectral Clustering, GMM.
4) So sánh kết quả dựa trên các chỉ số: số lượng cụm, kích thước trung bình, modularity và cụm có mức độ hợp tác cao nhất.
5) Trực quan hóa kết quả phân cụm bằng biểu đồ mạng, thể hiện cấu trúc cộng đồng và mối quan hệ giữa các nghệ sĩ.

Nhờ cấu hình trên, quá trình thực nghiệm cho phép đánh giá khách quan và toàn diện hiệu quả của từng thuật toán trong việc phát hiện cộng đồng trong mạng lưới nghệ sĩ gameshow Việt Nam.

## V. Kết quả và thảo luận

### *A. Phân tích cấu trúc mạng trước khi phân cụm*

Trước khi áp dụng các thuật toán phát hiện cộng đồng, mạng lưới nghệ sĩ tham gia các gameshow Việt Nam được phân tích để hiểu rõ đặc điểm cấu trúc và mức độ gắn kết giữa các nghệ sĩ. Mỗi nút (node) đại diện cho một nghệ sĩ, trong khi mỗi cạnh (edge) biểu diễn mối quan hệ hợp tác – tức hai nghệ sĩ cùng xuất hiện trong ít nhất một chương trình. Trọng số của cạnh thể hiện số lần hợp tác giữa họ.

Kết quả cho thấy mạng lưới bao gồm **675 nghệ sĩ (nodes)** và **55,262 cạnh (edges)**. Mạng có **mật độ đồ thị (density)** đạt **0.24293659**, cho thấy mức độ liên kết rất cao – gần 25% tổng số cặp nghệ sĩ có mối quan hệ trực tiếp, phản ánh đặc trưng hợp tác thường xuyên trong môi trường gameshow. **Độ trung bình (average degree)** là **163.7393**, nghĩa là mỗi nghệ sĩ trung bình hợp tác với hơn 163 người khác. **Độ lớn nhất (maximum degree)** đạt **470**, thuộc về **Kim Tử Long**, cho thấy nghệ sĩ này giữ vị trí trung tâm với lượng hợp tác dày đặc nhất trong toàn mạng.

Về **các chỉ số trung tâm (centrality)**:

- **Betweenness Centrality**: trung bình **0.001207**, lớn nhất **0.030409** (Node: **Hòa Minzy**) – thể hiện vai trò cầu nối giữa các nhóm nghệ sĩ, giúp mạng duy trì khả năng lan truyền thông tin.

- **Closeness Centrality**: trung bình **0.560108**, lớn nhất **0.767654** (Node: **Kim Tử Long**) – cho thấy khả năng tiếp cận nhanh đến hầu hết các nghệ sĩ khác trong mạng.
- **PageRank**: trung bình **0.001481**, lớn nhất **0.004627** (Node: **Kim Tử Long**) – phản ánh mức độ ảnh hưởng và tầm phủ sóng cao trong mạng lưới gameshow.

Ngoài ra, **hệ số clustering trung bình (average clustering coefficient)** đạt **0.890148**, phản ánh tính kết dính cao giữa các nghệ sĩ — nếu hai người cùng hợp tác với một nghệ sĩ thứ ba, họ có khả năng cao cũng từng hợp tác với nhau. Chỉ số này cho thấy mạng có xu hướng hình thành các nhóm nhỏ gắn bó chặt chẽ, thường là các nghệ sĩ cùng thể loại hoặc tham gia trong chuỗi chương trình tương tự. Đáng chú ý, mạng có **1 thành phần liên thông duy nhất**, nghĩa là toàn bộ 675 nghệ sĩ đều có thể kết nối gián tiếp với nhau qua các chuỗi hợp tác khác nhau.

Tổng thể, mạng lưới thể hiện đặc điểm của một **mạng phi tập trung nhưng có cấu trúc liên kết chặt chẽ (decentralized yet cohesive)**, trong đó các nghệ sĩ trung tâm như **Kim Tử Long**, **Trấn Thành**, **Trường Giang** và **Hòa Minzy** đóng vai trò hạt nhân, giữ cho mạng duy trì sự gắn kết và ổn định, tạo tiền đề cho việc phát hiện cấu trúc cộng đồng ở các bước tiếp theo.

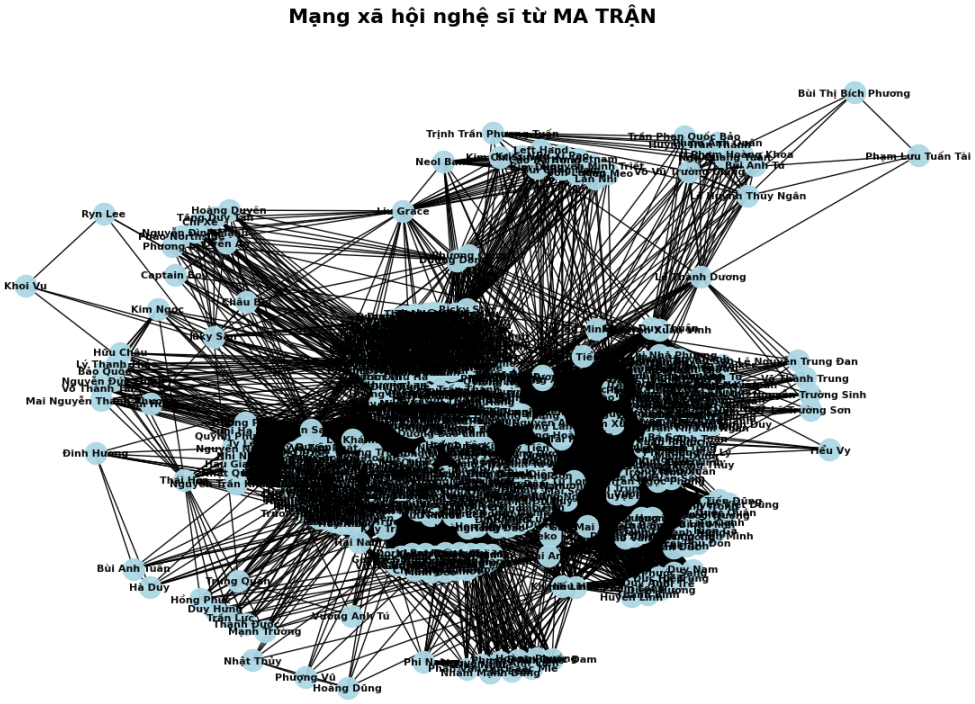


Fig. 1. Mạng xã hội nghệ sĩ Việt Nam được xây dựng từ ma trận đồng xuất hiện (co-appearance matrix). Mỗi nút biểu diễn một nghệ sĩ, cạnh biểu diễn mối quan hệ hợp tác, và độ dày cạnh thể hiện tần suất hợp tác.

**Phân tích Hình 1:** Hình này thể hiện toàn bộ cấu trúc mạng xã hội của 675 nghệ sĩ tham gia các gameshow Việt Nam, được xây dựng dựa trên ma trận đồng xuất hiện (co-appearance matrix). Mỗi nút biểu diễn một nghệ sĩ, trong khi mỗi cạnh nối giữa hai nút thể hiện mối quan hệ hợp tác – tức là hai nghệ sĩ cùng tham gia ít nhất một chương trình. Độ dày của cạnh phản ánh tần suất hợp tác, còn kích thước nút biểu thị số lượng mối liên kết (degree). Quan sát cho thấy mạng có cấu trúc rộng và kết nối dày đặc, với các cụm tự nhiên tập trung ở trung tâm, chủ yếu gồm các nghệ sĩ nổi bật như **Trấn Thành**, **Trường Giang**, **Hari Won** và **Ngô Kiến Huy**. Các nút ở rìa biểu thị những nghệ sĩ ít tham gia hoặc chỉ hợp tác trong nhóm cố định, phản ánh đặc trưng phân tầng trong mạng lưới giải trí Việt Nam.

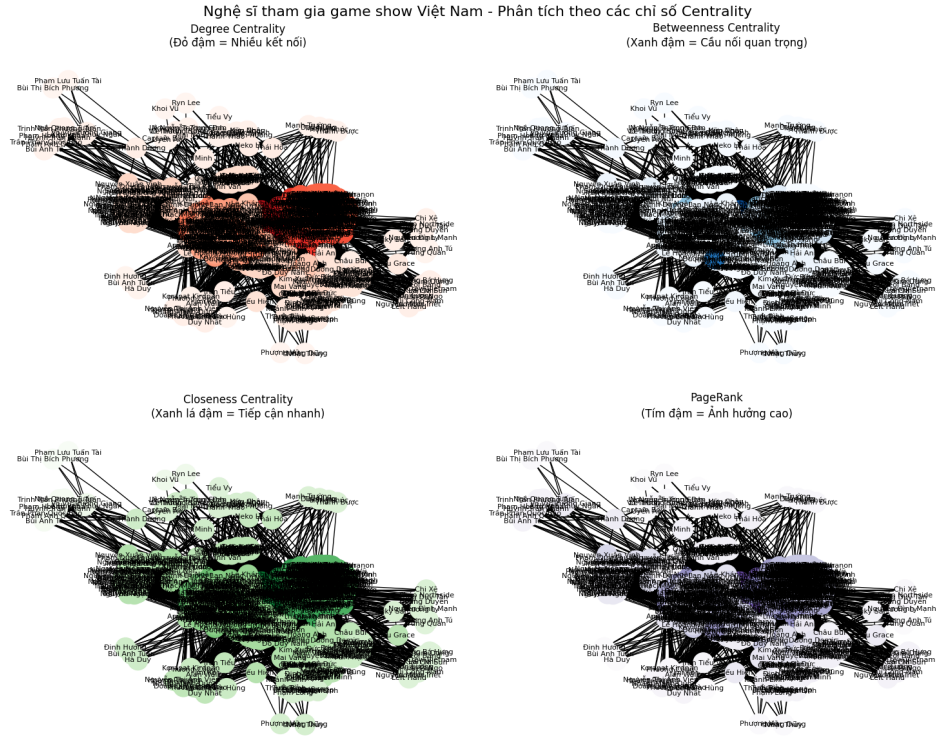


Fig. 2. Phân tích mạng nghệ sĩ theo các chỉ số Centrality: (a) Degree Centrality, (b) Betweenness Centrality, (c) Closeness Centrality, (d) PageRank. Màu sắc thể hiện mức độ quan trọng và ảnh hưởng của từng nghệ sĩ trong mạng.

**Phân tích Hình 2:** Hình này minh họa vai trò của từng nghệ sĩ trong mạng theo bốn chỉ số trung tâm chính:

- **Degree Centrality (đỏ đậm)**: thể hiện số lượng mối hợp tác trực tiếp; các nghệ sĩ như **Kim Tử Long**, **Trấn Thành** và **Trường Giang** có giá trị cao nhất.
- **Betweenness Centrality (xanh dương đậm)**: phản ánh khả năng làm cầu nối giữa các nhóm; nghệ sĩ **Hòa Minzy** có giá trị cao nhất, cho thấy vai trò trung gian quan trọng.
- **Closeness Centrality (xanh lá đậm)**: biểu thị khả năng tiếp cận nhanh đến các nghệ sĩ khác; các nghệ sĩ trung tâm thường có giá trị cao hơn, phản ánh khả năng lan tỏa thông tin tốt.
- **PageRank (tím đậm)**: đánh giá mức độ ảnh hưởng tổng thể; **Kim Tử Long** đạt giá trị cao nhất, thể hiện phạm vi ảnh hưởng rộng và vai trò kết nối chính trong mạng.

Từ bốn chỉ số trên có thể thấy, mạng nghệ sĩ gameshow Việt Nam có đặc điểm **liên kết cao, đa tầng và chịu chi phối bởi một số nghệ sĩ trung tâm**, tạo nền tảng cho quá trình phát hiện cộng đồng trong các phần tiếp theo.

### *B. Kết quả hiệu năng phân cụm*

Bảng I trình bày kết quả so sánh hiệu năng giữa bốn thuật toán **Louvain**, **Leiden**, **Spectral Clustering** và **GMM** trên mạng lưới nghệ sĩ gameshow Việt Nam. Các chỉ số bao gồm **số cụm**, **kích thước trung bình**, **cụm nhỏ nhất**, **cụm lớn nhất**, **Modularity** và **Silhoutte**.

TABLE I
SO SÁNH HIỆU SUẤT PHÂN CỤM TRÊN MẠNG LƯỚI NGHỆ SĨ GAMESHOW VIỆT NAM

| Thuật toán | Số cụm | Kích thước TB | Cụm nhỏ nhất | Cụm lớn nhất | Modularity | Silhouette |
|---|---|---|---|---|---|---|
| Louvain | 5 | 135.0 | 12 | 259 | 0.3682 | 0.2484 |
| Leiden | 6 | 112.5 | 2 | 223 | 0.3784 | 0.2456 |
| Spectral | 5 | 135.0 | 12 | 516 | 0.0867 | 0.4131 |
| GMM | 5 | 135.0 | 46 | 247 | 0.3483 | 0.3894 |

Kết quả trong Bảng I cho thấy thuật toán **Leiden** đạt giá trị **Modularity** cao nhất (0.3784), thể hiện khả năng phát hiện các cộng đồng có mức độ liên kết nội bộ chặt chẽ và rõ ràng hơn so với các phương pháp khác. Thuật toán **Louvain** cho kết quả ổn định và cấu trúc cụm cân đối, trong khi **Spectral Clustering** đạt giá trị **Silhouette** cao nhất, phản ánh khả năng tách biệt ranh giới giữa các cụm tốt hơn dù mức độ gắn kết nội cụm thấp. Phương pháp **Gaussian Mixture Model (GMM)** thể hiện

sự cân bằng giữa hai tiêu chí — vừa duy trì độ gắn kết cộng đồng hợp lý (Modularity = 0.3483) vừa mô tả được tính chồng lấn của các nhóm nghệ sĩ, phù hợp với đặc trưng linh hoạt trong quan hệ hợp tác trên các gameshow Việt Nam.

### *C. Trực quan hóa kết quả phân cụm*

Các hình dưới đây minh họa kết quả phân cụm cộng đồng nghệ sĩ tham gia gameshow Việt Nam bằng bốn thuật toán khác nhau: **Louvain**, **Leiden**, **Spectral Clustering** và **Gaussian Mixture Model (GMM)**. Mỗi nút trong mạng biểu diễn một nghệ sĩ, và mỗi cạnh thể hiện mối liên kết đồng xuất hiện giữa hai nghệ sĩ trong cùng một chương trình. Màu sắc biểu thị cụm (community) mà nghệ sĩ thuộc về — tức là nhóm những người thường xuyên hợp tác với nhau trên truyền hình.

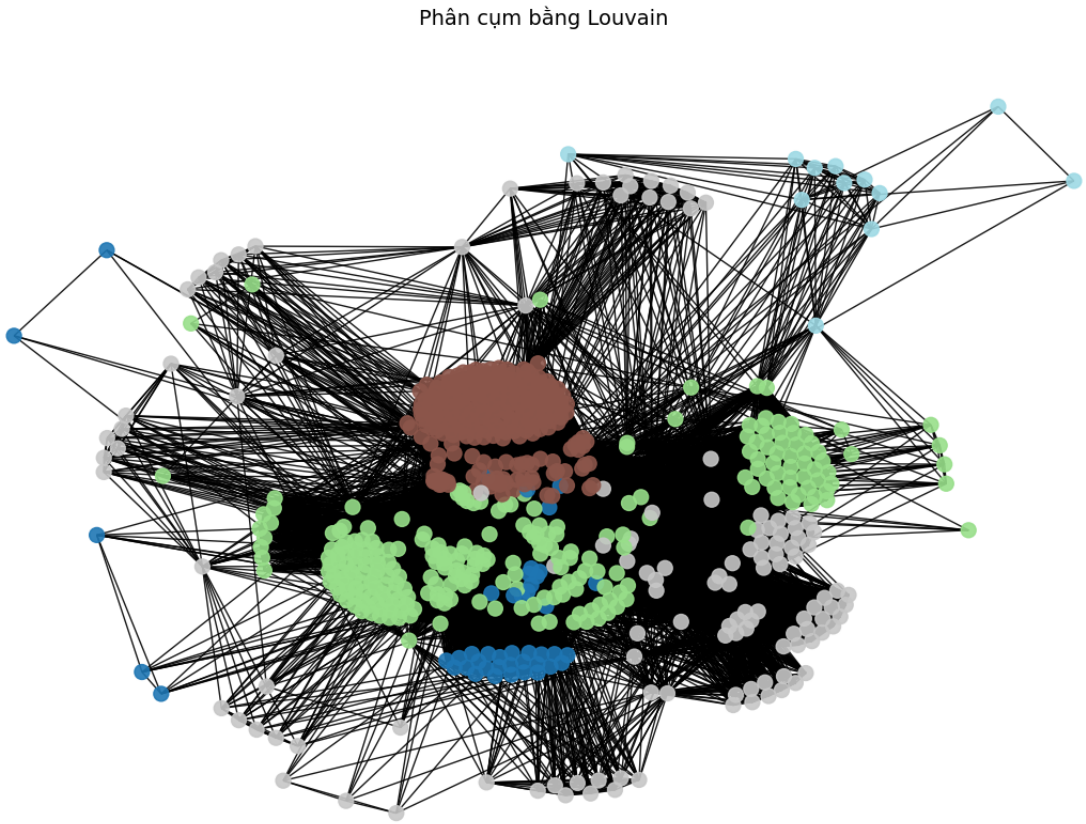


Fig. 3. Phân cụm cộng đồng nghệ sĩ bằng thuật toán Louvain.

Kết quả phân cụm bằng **Louvain** cho thấy mạng lưới được chia thành **5 cụm chính**, có ranh giới rõ ràng và các nhóm liên kết chặt chẽ. Các cụm tập trung quanh những nghệ sĩ có mức độ kết nối cao như **Kim Tử Long** và **Hòa Minzy**, đóng vai trò trung tâm trong nhiều chương trình. Cấu trúc này phản ánh mối quan hệ hợp tác ổn định giữa các nghệ sĩ trong cùng lĩnh vực như cải lương, ca nhạc và hài kịch.

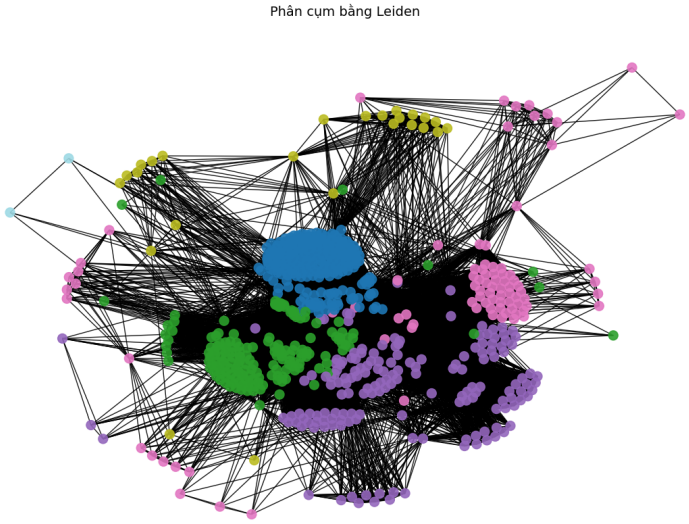


Fig. 4. Phân cụm cộng đồng nghệ sĩ bằng thuật toán Leiden.

Thuật toán **Leiden** cho kết quả chi tiết hơn, chia mạng thành **6 cụm nhỏ hơn**, giúp phát hiện các tiểu cộng đồng trong mạng lưới. So với Louvain, Leiden xử lý tốt hơn các cạnh chồng chéo, làm nổi bật các nhóm nhỏ hoạt động chuyên biệt — ví dụ, nhóm nghệ sĩ trẻ tham gia gameshow hiện đại và nhóm nghệ sĩ kỳ cựu trong chương trình truyền thống. Giá trị **Modularity = 0.3784** là cao nhất, cho thấy khả năng tách biệt cộng đồng hiệu quả nhất trong bốn thuật toán.

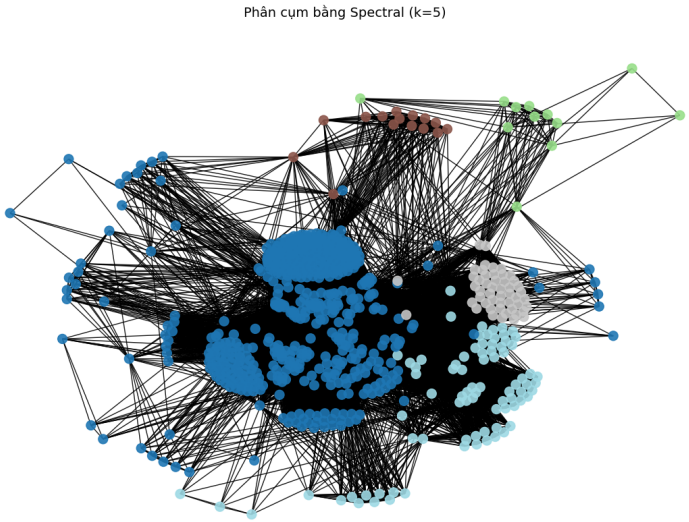


Fig. 5. Phân cụm cộng đồng nghệ sĩ bằng thuật toán Spectral Clustering (k=5).

Với **Spectral Clustering (k=5)**, các cụm được chia theo đặc trưng phổ của mạng, tuy nhiên có xu hướng **gộp cụm lớn và chưa rõ ràng**. Một số cụm nhỏ bị hòa trộn, dẫn đến **Modularity thấp nhất (0.0867)**, cho thấy phương pháp này chưa phù hợp khi mạng có mật độ cao và cấu trúc kết nối phức tạp. Dù vậy, Spectral vẫn hữu ích trong việc nhận diện các vùng rìa — nơi tập trung nghệ sĩ ít kết nối hoặc chỉ cộng tác trong một vài chương trình cố định.

Phương pháp **Gaussian Mixture Model (GMM)** chia mạng thành **5 cụm chính** với phân bố mềm (soft clustering), cho phép một nghệ sĩ có thể thuộc về nhiều nhóm khác nhau. Kết quả cho thấy các cụm có kích thước trung bình lớn (**135 nút**), với **Modularity = 0.3483** và **Silhouette = 0.3894**, chứng tỏ tính gắn kết cộng đồng tốt hơn Spectral và tương đương Louvain. Các cụm GMM thể hiện rõ các trung tâm hợp tác (cluster cores) — nơi nghệ sĩ có mức ảnh hưởng cao tập hợp quanh các nhóm gameshow chủ đề riêng biệt như âm nhạc, talkshow hoặc truyền hình thực tế.

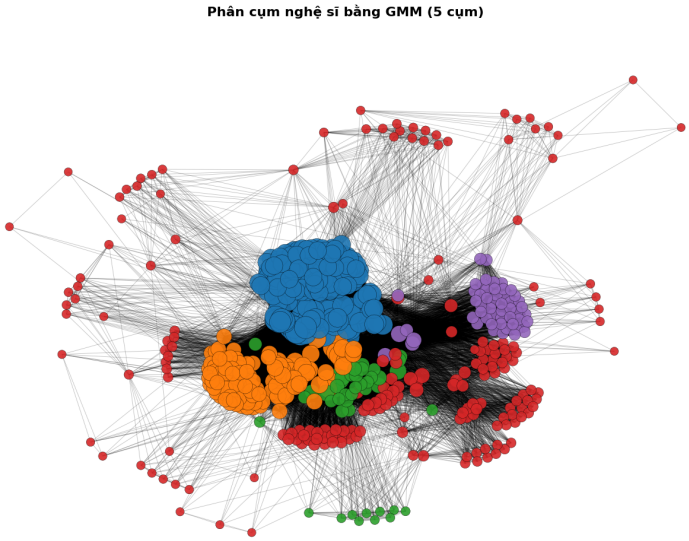


Fig. 6. Phân cụm nghệ sĩ bằng Gaussian Mixture Model (GMM, 5 cụm).

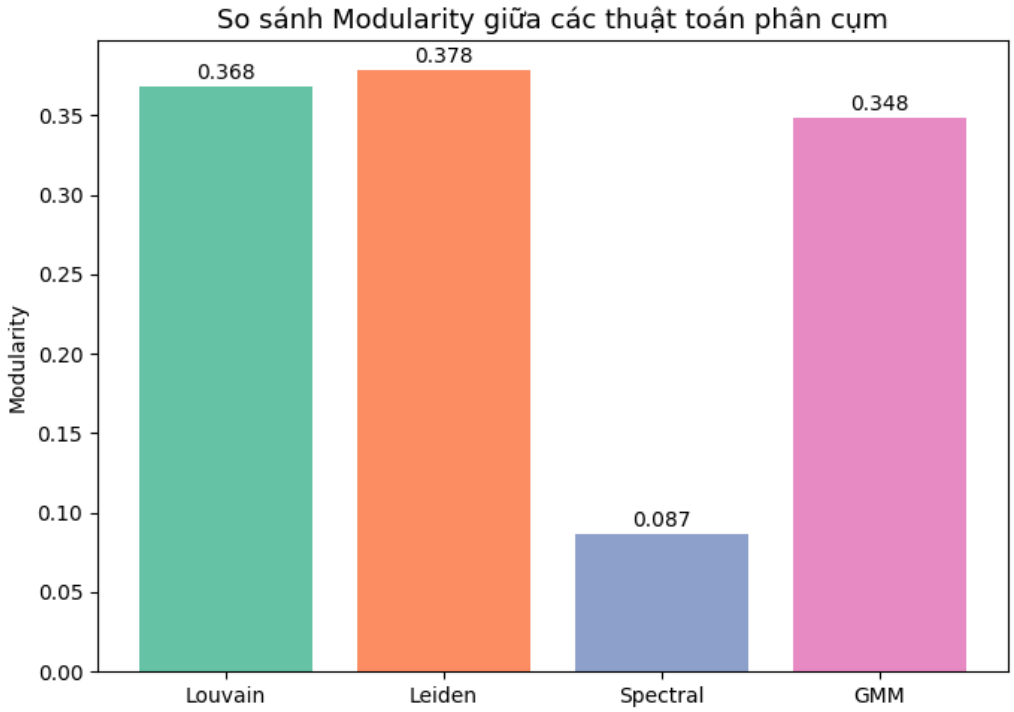


Fig. 7. So sánh chỉ số Modularity giữa các thuật toán Louvain, Leiden, Spectral và GMM.

Hình 7 thể hiện sự khác biệt về chỉ số **Modularity** giữa bốn thuật toán. **Leiden** đạt giá trị cao nhất, theo sau là **Louvain**, **GMM**, và cuối cùng là **Spectral Clustering**. Kết quả này xác nhận rằng Leiden mang lại cấu trúc cộng đồng rõ ràng nhất, trong khi Spectral có xu hướng đánh đổi độ chính xác để đạt cụm ít hơn. Sự khác biệt này phản ánh độ phức tạp trong mạng xã hội nghệ sĩ Việt Nam, nơi mỗi thuật toán làm nổi bật một khía cạnh khác nhau về hành vi hợp tác.

*D. Nhận diện nhóm hợp tác thường xuyên nhất*

Kết quả phân tích từ bốn thuật toán phân cụm **Louvain**, **Leiden**, **Spectral Clustering** và **Gaussian Mixture Model (GMM)** cho phép xác định các nhóm nghệ sĩ có tần suất hợp tác cao nhất, phản ánh xu hướng tương tác và mối liên kết trong ngành giải trí Việt Nam. Các cụm này được xem là *hạt nhân cộng đồng* của mạng lưới nghệ sĩ gameshow, trong đó những thành viên có tần suất xuất hiện cùng nhau nhiều nhất đóng vai trò trung tâm (hub) trong mạng.

*1. Cụm nổi bật theo từng thuật toán*

**Louvain:** Cụm hợp tác mạnh nhất là **Cụm 2** gồm **231 nghệ sĩ**, chiếm khoảng **45.86% tổng số cạnh toàn mạng** (26,984 trên 58,843 cạnh). Cụm này quy tụ nhiều nghệ sĩ có độ phủ sóng cao trong lĩnh vực truyền hình và ca nhạc như *Trấn Thành, Trường Giang, Hoài Linh, Thu Minh, Mỹ Tâm, Lam Trường, Ngô Kiến Huy, Đông Nhi, Đàm Vĩnh Hưng*, cùng nhiều nghệ sĩ quốc tế (Britney Spears, Céline Dion, Michael Jackson). Mức độ gắn kết nội cụm cao phản ánh khả năng của Louvain trong việc phát hiện cộng đồng lớn, ổn định và có tính liên kết truyền thống.

**Leiden:** Cụm hợp tác nổi bật là **Cụm 0**, gồm **223 nghệ sĩ**, chiếm **42.75% tổng số cạnh toàn mạng** (25,154/58,843). Các nghệ sĩ như *Trấn Thành, Trường Giang, Hari Won, Hoài Linh, Hồ Quỳnh Hương, Hòa Minzy, Lam Trường, Vũ Cát Tường* đóng vai trò trung tâm. Leiden phân tách cộng đồng chi tiết hơn Louvain, cho phép nhận diện các *tiểu nhóm* trong cụm chính — ví dụ nhóm ca sĩ trẻ (Hòa Minzy, Đức Phúc, Erik), nhóm MC – diễn viên (Trấn Thành, Trường Giang, Ninh Dương Lan Ngọc). Điều này cho thấy Leiden có khả năng mô hình hóa cấu trúc phân cấp (*hierarchical community*) tốt hơn, giúp phát hiện các mối quan hệ phụ thuộc nhỏ trong mạng lớn.

**Spectral Clustering:** Cụm mạnh nhất là **Cụm 0** với **516 nghệ sĩ**, chiếm **90.17% tổng số cạnh toàn mạng** (53,058/58,843). Thuật toán này tạo ra một cụm rất lớn, bao phủ gần như toàn bộ mạng lưới — bao gồm nhiều nghệ sĩ thuộc đa dạng lĩnh vực như diễn viên, ca sĩ, MC, và người mẫu (*Kim Tử Long, Hoài Linh, Hòa Minzy, Hari Won, Trấn Thành, Ninh Dương Lan Ngọc, Lê Dương Bảo Lâm, Đông Nhi, Mỹ Tâm*). Mặc dù có độ bao phủ lớn, kết quả của Spectral cho thấy **Modularity thấp nhất (0.0867)**, nghĩa là các ranh giới giữa các nhóm còn mờ, mạng chưa được tách rõ ràng thành các cộng đồng riêng biệt.

**Gaussian Mixture Model (GMM):** Cụm hợp tác nổi bật là **Cụm 0** gồm **247 nghệ sĩ**, chiếm **53.52% tổng số cạnh toàn mạng** (31,490/58,843). Cụm này bao gồm nhiều nghệ sĩ nổi tiếng và có độ ảnh hưởng cao trong ngành giải trí Việt Nam như *Kim Tử Long, Hòa Minzy, Hari Won, Hoài Linh, Lam Trường, Minh Nhí, Bảo Anh, Chi Pu, Miu Lê, Ngô Kiến Huy*. So với các thuật toán khác, GMM có khả năng **phát hiện cộng đồng chồng lấn (overlapping communities)** tốt hơn, cho phép một nghệ sĩ có thể thuộc về nhiều nhóm hợp tác. Điều này phản ánh rõ thực tế của ngành giải trí, khi các nghệ sĩ thường xuyên đảm nhận nhiều vai trò hoặc xuất hiện trong nhiều gameshow khác nhau, từ ca hát, diễn xuất đến dẫn chương trình. GMM nhờ đó cho thấy sự linh hoạt trong mô hình hóa mạng xã hội thực tế, nơi các ranh giới giữa cộng đồng không hoàn toàn tách biệt.

*2. So sánh tổng quan các cụm hợp tác*

TABLE II
SO SÁNH CÁC CỤM HỢP TÁC MẠNH NHẤT GIỮA CÁC THUẬT TOÁN

| Thuật toán | Cụm mạnh nhất | Tỉ lệ cạnh nội cụm (%) | Số nghệ sĩ | Đặc điểm nổi bật |
|---|---|---|---|---|
| Louvain | Cụm 2 | 45.86 | 231 | Phát hiện cộng đồng lớn, ổn định; liên kết mạnh giữa nghệ sĩ truyền thống và ca sĩ nổi tiếng. |
| Leiden | Cụm 0 | 42.75 | 223 | Tách rõ các tiểu nhóm; mô hình hóa tốt cấu trúc phân cấp và tương tác phụ thuộc. |
| Spectral | Cụm 0 | 90.17 | 516 | Bao phủ gần toàn bộ mạng; ranh giới cộng đồng mờ, độ gắn kết nội cụm thấp. |
| GMM | Cụm 0 | 53.52 | 247 | Phát hiện cộng đồng chồng lấn; phản ánh tính linh hoạt và hợp tác đa chiều giữa nghệ sĩ. |

*3. Phân tích tổng kết*

Kết quả cho thấy hai thuật toán **Leiden** và **GMM** mang lại khả năng nhận diện cộng đồng phù hợp nhất với mạng nghệ sĩ Việt Nam. Louvain vẫn duy trì hiệu quả trong việc phát hiện các cụm lớn ổn định, trong khi Spectral cho kết quả gộp rộng nhưng kém sắc nét về ranh giới cộng đồng. Đặc biệt, **GMM** cho phép mô hình hóa các mối liên kết đa chiều, thể hiện tính linh hoạt và tính chồng lấn trong mối quan hệ nghề nghiệp — đặc trưng của lĩnh vực giải trí, nơi các nghệ sĩ thường xuyên hợp tác linh hoạt giữa nhiều gameshow khác nhau.

Về mặt ứng dụng thực tiễn, kết quả này có thể được khai thác cho các mục tiêu như:

- Dự đoán xu hướng hợp tác hoặc mối quan hệ mới giữa các nghệ sĩ trong tương lai;
- Đề xuất nghệ sĩ phù hợp cho các gameshow dựa trên mức độ tương tác mạng;
- Hỗ trợ nhà sản xuất trong việc thiết kế chiến lược truyền thông và kết nối nhân sự hiệu quả.

Nhìn chung, việc kết hợp phân tích mạng lưới với các thuật toán phân cụm đã cung cấp một cái nhìn sâu sắc về cấu trúc xã hội và mô hình hợp tác trong ngành giải trí Việt Nam, đặc biệt là trong bối cảnh các gameshow ngày càng đa dạng và có tính tương tác cao.

### *E. Quy trình phân tích mạng lưới nghệ sĩ*

Hình 8 trình bày quy trình tổng thể của hệ thống phân tích mạng lưới nghệ sĩ trong các gameshow Việt Nam. Quy trình được thiết kế nhằm đảm bảo luồng xử lý dữ liệu từ thu thập, tiền xử lý đến phân tích và trực quan hóa được diễn ra một cách hệ thống và nhất quán.

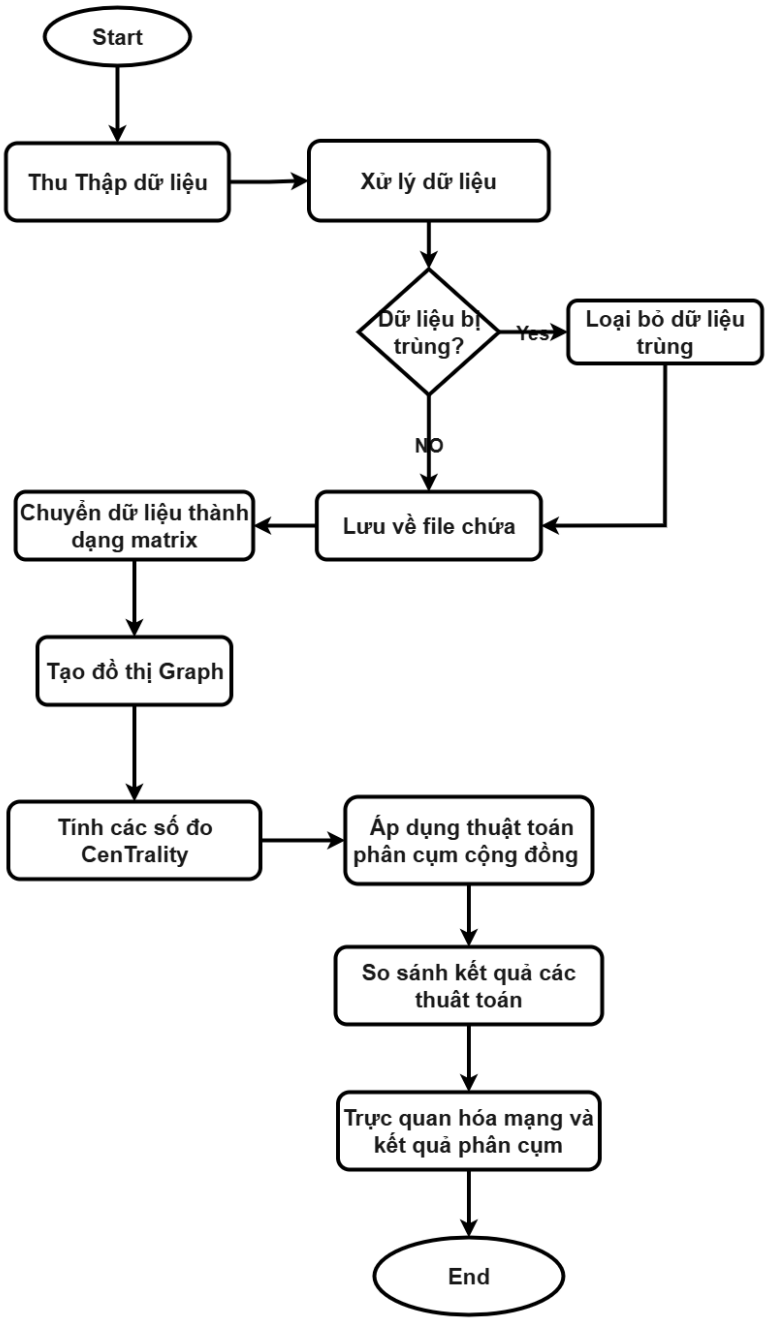


Fig. 8. Quy trình phân tích mạng lưới nghệ sĩ gameshow Việt Nam

Quy trình bắt đầu với bước **Thu thập dữ liệu**, trong đó danh sách nghệ sĩ và chương trình được thu thập tự động từ các trang Wikipedia bằng công cụ Selenium. Dữ liệu thô sau khi thu thập được **tiền xử lý** — loại bỏ dữ liệu trùng lặp, chuẩn hóa tên nghệ sĩ và định dạng thông tin. Sau khi làm sạch, dữ liệu được **chuyển đổi sang dạng ma trận đồng xuất hiện (co-appearance matrix)**, biểu diễn số lần hai nghệ sĩ cùng xuất hiện trong một chương trình.

Từ ma trận này, hệ thống **xây dựng đồ thị mạng xã hội (Graph)** trong đó mỗi nút (node) biểu diễn một nghệ sĩ, còn cạnh (edge) thể hiện mối quan hệ hợp tác giữa họ. Tiếp theo, các **chỉ số trung tâm (Centrality)** như Degree, Betweenness, Closeness và PageRank được tính toán nhằm xác định vai trò của từng nghệ sĩ trong mạng. Các thuật toán **phân cụm cộng đồng** bao gồm Louvain, Leiden, Spectral và GMM được áp dụng để nhận diện các nhóm nghệ sĩ có sự liên kết chặt chẽ. Sau đó, kết quả của các thuật toán được **so sánh dựa trên chỉ số Modularity** để xác định thuật toán tối ưu nhất trong việc mô tả cấu trúc cộng đồng. Cuối cùng, hệ thống tiến hành **trực quan hóa mạng lưới** và các kết quả phân cụm, giúp người quan sát dễ dàng nhận diện các nhóm nghệ sĩ hợp tác thường xuyên và vị trí của họ trong mạng lưới gameshow Việt Nam.

Như vậy, lưu đồ thể hiện rõ luồng xử lý dữ liệu của hệ thống – từ khâu thu thập dữ liệu thô đến phân tích và trực quan hóa – đảm bảo tính logic, khoa học và khả năng mở rộng cho các nghiên cứu mạng xã hội tương tự trong tương lai.

## VI. Ứng dụng và Ý nghĩa

### A. Ứng dụng thực tiễn

Kết quả phân tích mạng lưới nghệ sĩ Việt Nam mang lại nhiều ứng dụng tiềm năng trong các lĩnh vực nghiên cứu, truyền thông và công nghiệp giải trí:

- **Phân tích cộng đồng nghệ sĩ:** Thông qua các thuật toán phát hiện cộng đồng như Louvain, Leiden, Spectral và GMM, nghiên cứu giúp nhận diện các nhóm nghệ sĩ có tần suất hợp tác cao, từ đó phản ánh xu hướng liên kết trong giới giải trí Việt Nam.
- **Hỗ trợ chiến lược truyền thông và quảng bá:** Việc xác định các nghệ sĩ có mức độ ảnh hưởng cao (dựa trên các chỉ số Centrality như Degree, Betweenness, PageRank) giúp các nhà sản xuất, hãng truyền thông hoặc nhãn hàng chọn lựa đối tác quảng cáo phù hợp, tối ưu hóa chiến lược marketing.
- **Phát hiện xu hướng hợp tác mới:** Các cụm nghệ sĩ được phát hiện thông qua mô hình GMM hoặc Leiden có thể chỉ ra những nhóm đang phát triển nhanh hoặc có tiềm năng kết hợp trong tương lai, mở ra hướng dự đoán các mối quan hệ hợp tác mới trong ngành giải trí.
- **Hỗ trợ nghiên cứu xã hội học và văn hóa:** Phân tích mạng nghệ sĩ giúp nhận diện cấu trúc cộng đồng, đánh giá sự giao thoa giữa các lĩnh vực nghệ thuật (âm nhạc, diễn xuất, truyền hình), góp phần vào các nghiên cứu về hành vi hợp tác và ảnh hưởng xã hội trong văn hóa đại chúng Việt Nam.

### B. Ý nghĩa khoa học và xã hội

- **Về mặt khoa học dữ liệu:** Nghiên cứu chứng minh hiệu quả của việc ứng dụng các thuật toán phát hiện cộng đồng (Community Detection) trong mạng xã hội thực tế. Việc so sánh các phương pháp Louvain, Leiden, Spectral và GMM giúp đánh giá tính phù hợp của từng thuật toán đối với dữ liệu có tính kết nối cao như mạng nghệ sĩ.
- **Về mặt xã hội:** Kết quả giúp làm sáng tỏ cấu trúc hợp tác trong giới nghệ sĩ Việt Nam, cho thấy cách mà danh tiếng, quan hệ nghề nghiệp và lĩnh vực hoạt động ảnh hưởng đến mạng lưới kết nối. Qua đó, có thể hiểu rõ hơn về sự hình thành và lan tỏa của ảnh hưởng trong lĩnh vực giải trí.
- **Về ứng dụng công nghệ:** Mô hình phân tích mạng này có thể được mở rộng cho các lĩnh vực khác như mạng lưới doanh nghiệp, mạng nghiên cứu khoa học, hoặc mạng người dùng trong các nền tảng mạng xã hội (Facebook, YouTube, TikTok...), góp phần vào việc phát triển các hệ thống đề xuất thông minh (recommendation systems).

### C. Tổng kết

Từ các kết quả thu được, nghiên cứu không chỉ mô tả cấu trúc hợp tác của giới nghệ sĩ Việt Nam mà còn minh chứng cho sức mạnh của phân tích mạng xã hội (Social Network Analysis) kết hợp với học máy. Các kết quả như **Modularity**, **Silhouette Score**, và **nhóm hợp tác thường xuyên nhất** thể hiện khả năng của mô hình trong việc phát hiện và diễn giải mối quan hệ phức tạp trong cộng đồng. Điều này mở ra hướng phát triển cho các nghiên cứu ứng dụng tương tự trong tương lai, đặc biệt trong việc khai thác dữ liệu văn hóa, xã hội và truyền thông số tại Việt Nam.

## VII. Hạn chế và Nghiên cứu tương lai

### A. Hạn chế của nghiên cứu

Mặc dù nghiên cứu đã mang lại những kết quả tích cực trong việc mô hình hóa và phân tích mạng lưới nghệ sĩ Việt Nam, vẫn tồn tại một số hạn chế nhất định cần được xem xét:

- **Giới hạn về nguồn dữ liệu:** Dữ liệu được thu thập chủ yếu từ các chương trình gameshow truyền hình và nền tảng trực tuyến công khai. Điều này có thể chưa bao quát toàn bộ mạng lưới hợp tác nghệ sĩ trong các lĩnh vực khác như điện ảnh,

sân khấu, hay âm nhạc độc lập. Một số nghệ sĩ ít xuất hiện trên truyền hình có thể bị thiếu thông tin, dẫn đến sai lệch nhẹ trong cấu trúc mạng.

- **Độ chính xác của thông tin hợp tác:** Việc nhận diện mối quan hệ hợp tác dựa trên cùng tham gia chương trình không phản ánh đầy đủ tính chất thực sự của mối quan hệ (ví dụ: mức độ tương tác, thời lượng hợp tác, hay tần suất cùng xuất hiện). Điều này khiến một số cạnh trong mạng có thể được đánh giá quá cao hoặc quá thấp.
- **Hạn chế trong việc đánh giá thuật toán:** Các chỉ số như *Modularity* hay *Silhouette Score* phản ánh tốt tính chất phân cụm nhưng chưa đánh giá sâu về ý nghĩa xã hội học của các cộng đồng được phát hiện. Ngoài ra, các thuật toán như Louvain, Leiden, Spectral và GMM đều phụ thuộc vào tham số đầu vào, do đó kết quả có thể thay đổi nhẹ theo từng lần chạy.
- **Chưa khai thác yếu tố thời gian:** Dữ liệu hiện tại được xem xét ở dạng tĩnh (một thời điểm tổng hợp). Chưa có phân tích động (temporal analysis) để thể hiện sự thay đổi mạng lưới hợp tác của nghệ sĩ theo giai đoạn hoặc theo từng chương trình.

### *B. Định hướng và nghiên cứu tương lai*

Trong các nghiên cứu tiếp theo, có thể mở rộng theo nhiều hướng nhằm khắc phục những hạn chế trên và nâng cao giá trị ứng dụng của mô hình:

- **Mở rộng nguồn dữ liệu:** Kết hợp dữ liệu từ các nền tảng khác như Wikipedia, IMDb, Spotify, YouTube hoặc mạng xã hội (Facebook, TikTok, Instagram) để xây dựng mạng lưới đa chiều hơn, phản ánh mối quan hệ trong các lĩnh vực khác nhau của nghệ sĩ.
- **Phân tích mạng động (Dynamic Network Analysis):** Áp dụng mô hình phân tích theo thời gian để theo dõi sự thay đổi trong cấu trúc hợp tác — ví dụ, giai đoạn 2010–2015 và 2016–2025 — qua đó phát hiện xu hướng mới, sự hình thành hoặc tan rã của các cộng đồng nghệ sĩ.
- **Ứng dụng học sâu (Deep Learning) và GNN:** Triển khai các mô hình học sâu trên đồ thị như *Graph Neural Network (GNN)*, *Graph Autoencoder (GAE)* hoặc *Variational Graph Autoencoder (VGAE)* để cải thiện khả năng phát hiện cộng đồng, dự đoán mối quan hệ hợp tác mới và xếp hạng ảnh hưởng của nghệ sĩ.
- **Phân tích ngữ nghĩa và nội dung:** Kết hợp xử lý ngôn ngữ tự nhiên (NLP) để phân tích nội dung truyền thông, bài viết hoặc bình luận liên quan đến nghệ sĩ, từ đó đánh giá thêm yếu tố cảm xúc, hình ảnh và tác động truyền thông trong từng nhóm cộng đồng.
- **Ứng dụng trong hệ thống đề xuất (Recommendation Systems):** Dựa trên các cụm cộng đồng và chỉ số ảnh hưởng, có thể phát triển hệ thống gợi ý hợp tác giữa các nghệ sĩ hoặc dự đoán khả năng xuất hiện chung trong tương lai, hỗ trợ các nhà sản xuất gameshow và công ty truyền thông.

### *C. Tổng kết*

Nhìn chung, nghiên cứu này là một bước đầu quan trọng trong việc ứng dụng các kỹ thuật phân tích mạng xã hội và học máy vào lĩnh vực giải trí Việt Nam. Dù còn hạn chế, kết quả đã chứng minh tiềm năng của mô hình trong việc hiểu sâu hơn về cấu trúc, mối quan hệ và xu hướng hợp tác của nghệ sĩ. Trong tương lai, với dữ liệu phong phú hơn và các phương pháp học sâu tiên tiến, mô hình có thể mở rộng phạm vi ứng dụng sang các lĩnh vực khác như âm nhạc, điện ảnh, truyền thông xã hội và phân tích ảnh hưởng văn hóa.

## VIII. KẾT LUẬN

Nghiên cứu này đã tiến hành xây dựng và phân tích **mạng xã hội nghệ sĩ Việt Nam** dựa trên dữ liệu thu thập từ các gameshow truyền hình, qua đó ứng dụng các thuật toán phân cụm đồ thị để khám phá **cấu trúc cộng đồng hợp tác trong giới nghệ sĩ**. Kết quả đã chứng minh khả năng của các phương pháp học máy và phân tích mạng xã hội trong việc mô hình hóa mối quan hệ phức tạp giữa các cá nhân trong lĩnh vực giải trí.

Trước hết, phần **phân tích mạng gốc** cho thấy mạng lưới gồm 675 nghệ sĩ và 55,262 mối liên kết, có mật độ khá cao (0.2429) và hệ số *clustering coefficient* trung bình đạt 0.8901 — biểu hiện cho tính gắn kết mạnh trong cộng đồng nghệ sĩ. Các chỉ số trung tâm như *Degree*, *Closeness* và *PageRank* đã xác định được những nghệ sĩ có tầm ảnh hưởng cao như Kim Tử Long, Hòa Minzy, và Hoài Linh, đóng vai trò là các nút trung tâm quan trọng trong mạng.

Tiếp theo, bốn thuật toán phân cụm được triển khai gồm **Louvain**, **Leiden**, **Spectral Clustering** và **Gaussian Mixture Model (GMM)**. Trong đó, thuật toán **Leiden** đạt giá trị *Modularity* cao nhất (0.3784), thể hiện khả năng phát hiện rõ ràng các cộng đồng có liên kết nội bộ chặt chẽ, trong khi **GMM** giúp mô hình hóa các nhóm hợp tác mềm dẻo và phát hiện các cụm lớn hơn dựa trên phân phối xác suất. Kết quả này được minh họa trực quan qua các biểu đồ mạng, giúp thể hiện rõ cấu trúc cộng đồng và mối quan hệ hợp tác giữa các nhóm nghệ sĩ.

Phân tích chi tiết các **nhóm hợp tác thường xuyên** cho thấy sự hình thành rõ rệt của những cộng đồng nghệ sĩ thường xuyên làm việc cùng nhau, như nhóm nghệ sĩ gameshow có sự tham gia của Hòa Minzy, Hari Won, Trấn Thành, Trường

Giang, Hồ Ngọc Hà,... Các nhóm này không chỉ đại diện cho mối quan hệ nghề nghiệp mà còn phản ánh xu hướng kết nối trong lĩnh vực giải trí Việt Nam hiện đại.

Từ các kết quả đạt được, nghiên cứu đã góp phần:

- Xây dựng được mô hình mạng xã hội nghệ sĩ Việt Nam có thể mở rộng và cập nhật trong tương lai.
- So sánh hiệu quả giữa các thuật toán phân cụm hiện đại trên dữ liệu xã hội thực tế.
- Đề xuất hướng ứng dụng kết quả vào các hệ thống gợi ý hợp tác hoặc phân tích xu hướng truyền thông trong ngành giải trí.

Tổng kết lại, đề tài không chỉ khẳng định tính ứng dụng của khoa học dữ liệu và học máy trong lĩnh vực xã hội – văn hóa mà còn mở ra hướng nghiên cứu mới về **phân tích cộng đồng nghệ sĩ, ảnh hưởng xã hội và mạng hợp tác trong giải trí Việt Nam**. Với việc mở rộng dữ liệu và áp dụng các mô hình học sâu trong tương lai, công trình này có thể trở thành nền tảng quan trọng cho các nghiên cứu về *phân tích mạng xã hội, dự báo hợp tác nghệ sĩ* và *đánh giá ảnh hưởng văn hóa* trên phạm vi rộng hơn.

## REFERENCES


[1] 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, p. P10008, 2008.

[2] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, p. 5233, 2019.

[3] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[4] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. Springer, 2009, pp. 659–663.

[5] T. Teitelbaum, P. Balenzuela, P. Cano, and J. Buldu, "Community structures and role detection in music networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 18, no. 4, p. 043105, 2008.

[6] S. N. A. P. (SNAP), "Community detection on last.fm artist data," Stanford University, Tech. Rep., 2014. [Online]. Available: https://snap.stanford.edu/class/cs224w-2014/projects2014/cs224w-14-final.pdf

[7] F. Di Matteo, L. Zhang, and K. Nguyen, "Analysis of a spotify collaboration network for small-world properties," *arXiv preprint arXiv:2503.09526*, 2025.

[8] M. Zhang and H. Chen, "A multilevel clustering technique for community detection in complex networks," *arXiv preprint arXiv:2101.06551*, 2023.

[9] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.

[10] Selenium Project, *Selenium WebDriver Documentation*, 2023. [Online]. Available: https://www.selenium.dev/documentation/webdriver/

[11] 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, p. P10008, 2008.

[12] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, no. 1, p. 5233, 2019.

[13] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, vol. 14. MIT Press, 2002, pp. 849–856.