PROJECT 1A: PHÂN TÍCH CỘNG ĐỒNG NGHIÊN CỨU VÀ PHÁT HIỆN NHÓM HỢP TÁC TRONG LĨNH VỰC DATA SCIENCE

1st Hà Thế Anh, 2nd Nguyễn Nhật Nam, 3rd Hoàng Quang Minh
and Le Nhat Tung
HUTECH University, Vietnam
{hatheanh012004, nguyennhatnam01012004, hoangquangminh130804}@gmail.com, and lenhattung@hutech.edu.vn

TÓM TẮT NỘI DUNG

Trong những năm gần đây, **Khoa học Dữ liệu (Data Science)** phát triển mạnh mẽ, kéo theo sự mở rộng của hoạt động hợp tác giữa các nhà nghiên cứu. Việc nắm bắt cấu trúc và động lực của những **mạng lưới đồng tác giả (Co-author Networks)** giúp nhận diện xu hướng chủ đề, nhóm chuyên môn và các mối liên kết học thuật quan trọng.

Nghiên cứu này xây dựng **mạng đồng tác giả** từ dữ liệu thu thập tự động trên nền tảng **OpenAlex**, trong đó nút biểu diễn tác giả và cạnh biểu diễn quan hệ hợp tác. Trên đồ thị thu được, chúng tôi áp dụng các thuật toán **phát hiện cộng đồng** gồm **Louvain**, **Leiden** và **Fast Greedy** đồng thời tính các chỉ số **centrality** (Degree, Betweenness, Closeness, Eigenvector) để nhận diện những cá nhân có ảnh hưởng.

Kết quả cho thấy mạng lưới nghiên cứu hình thành các cộng đồng rõ rệt xoay quanh những hướng chủ đề gần nhau trong đó **Leiden** và **Louvain** đạt *modularity* cao và cho phân hoạch ổn định, còn **Fast Greedy** cho kết quả cạnh tranh với chi phí tính toán thấp. Phân tích centrality chỉ ra một số tác giả giữ vai trò "cầu nối" giữa các nhóm.

Công trình mang lại cái nhìn toàn diện về **cấu trúc hợp tác khoa học** trong lĩnh vực Data Science, hỗ trợ phát hiện nhóm hợp tác tiềm năng và gợi ý ứng dụng **Social Network Analysis** trong hoạch định chiến lược nghiên cứu.

TỪ KHÓA

OpenAlex, Co-author Network, Community Detection, Scientific Collaboration, Data Science, Social Network Analysis, Research Collaboration, Network Visualization.

I. GIỚI THIỆU

Trong bối cảnh kỷ nguyên dữ liệu bùng nổ, **Khoa học Dữ liệu (Data Science)** đã trở thành một trong những lĩnh vực nghiên cứu phát triển nhanh nhất hiện nay. Song song với sự gia tăng nhanh chóng của các công trình khoa học, các mối quan hệ hợp tác giữa các nhà nghiên cứu cũng ngày càng đa dạng và phức tạp hơn, hình thành nên những **mạng lưới cộng tác nghiên cứu (research collaboration networks)** có quy mô lớn và cấu trúc liên kết chặt chẽ.

Việc phân tích các mạng lưới này mang ý nghĩa quan trọng trong việc hiểu rõ **xu hướng hợp tác, lĩnh vực nghiên cứu trọng tâm**, cũng như phát hiện những nhóm học giả có ảnh hưởng lớn trong cộng đồng khoa học. Trong đó, **Phân tích mạng xã hội (Social Network Analysis - SNA)** đã trở thành một công cụ mạnh mẽ giúp mô hình hóa và khám phá các mối quan hệ giữa các thực thể nghiên cứu thông qua biểu diễn đồ thị.

Một hướng nghiên cứu cốt lõi của SNA là **phát hiện cộng đồng (community detection)** quá trình xác định các nhóm trong mạng lưới có mức độ kết nối nội bộ cao và liên kết bên ngoài yếu hơn. Việc phát hiện các cộng đồng này không chỉ giúp nhận diện cấu trúc tổ chức của mạng lưới khoa học, mà còn hỗ trợ khám phá các nhóm hợp tác tiềm năng, xu hướng nghiên cứu nổi bật và các tác giả đóng vai trò cầu nối giữa nhiều lĩnh vực khác nhau.

Trong nghiên cứu này, nhóm tập trung vào việc **phân tích cộng đồng nghiên cứu và phát hiện nhóm hợp tác trong lĩnh vực Data Science**. Dữ liệu được thu thập tự động từ nền tảng **OpenAlex** một cơ sở dữ liệu học thuật mở, chứa thông tin về các tác giả, bài báo và mối quan hệ hợp tác giữa họ. Từ nguồn dữ liệu này, nhóm tiến hành xây dựng **mạng đồng tác giả (co-author network)** trong đó mỗi nút biểu diễn một tác giả và mỗi cạnh thể hiện mối quan hệ cộng tác trong các công trình nghiên cứu.

Trên mạng lưới này, các thuật toán **Louvain**, **Leiden** và **Fast Greedy** được áp dụng để phát hiện cộng đồng, đồng thời các chỉ số **centrality** như *Degree*, *Betweenness*, *Closeness* và *Eigenvector* được tính toán nhằm xác định những cá nhân có mức độ ảnh hưởng lớn.

Mục tiêu chính của đề tài là:

- Xây dựng mô hình mạng đồng tác giả trong lĩnh vực Data Science từ dữ liệu thực tế.
- Áp dụng các thuật toán phát hiện cộng đồng để tìm ra các nhóm hợp tác nổi bật.
- Phân tích vai trò và mức độ ảnh hưởng của các nhà nghiên cứu trong mạng lưới thông qua các chỉ số centrality.
- Đưa ra nhận xét, trực quan hóa và đánh giá hiệu quả của các thuật toán được sử dụng.

Kết quả nghiên cứu không chỉ góp phần cung cấp cái nhìn tổng quan về **cấu trúc hợp tác học thuật trong lĩnh vực Data Science**, mà còn mở ra hướng ứng dụng tiềm năng của các kỹ thuật phân tích mạng trong việc định hướng chính sách nghiên cứu, phát hiện nhóm chuyên môn mới và thúc đẩy sự hợp tác khoa học trong tương lai.

II. CÁC NGHIÊN CỨU LIÊN QUAN

A. Mạng đồng tác giả và nguồn dữ liệu học thuật

Mạng đồng tác giả (co–author network) là lớp đồ thị trong đó mỗi nút là một tác giả và cạnh biểu diễn mối quan hệ hợp tác khi hai tác giả cùng công bố ít nhất một công trình. Nhiều nghiên cứu nền tảng đã chỉ ra cấu trúc cộng đồng và tính phân cụm mạnh của các mạng khoa học, cũng như vai trò của các cá nhân cầu nối trong việc lan truyền tri thức [1] [2].

Gần đây, các cơ sở dữ liệu học thuật mở như **OpenAlex** cho phép truy cập quy mô lớn tới thông tin bài báo, tác giả, chủ đề và quan hệ trích dẫn hợp tác, giúp việc tái lập và mở rộng nghiên cứu trở nên khả thi hơn [3].

B. Phát hiện cộng đồng dựa trên modularity

Một hướng tiếp cận kinh điển trong phát hiện cộng đồng là tối ưu *modularity* Q [4], [5]:

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

trong đó A_{ij} là phần tử của ma trận kề, k_i và k_j lần lượt là bậc của các nút i và j, m là tổng số cạnh trong mạng, và $\delta(c_i, c_j)$ bằng 1 nếu hai nút thuộc cùng một cộng đồng, ngược lại bằng 0.

Thuật toán **Louvain** [6] tối ưu chỉ số modularity thông qua hai giai đoạn lặp lại: (i) di chuyển từng nút đến cộng đồng lân cận sao cho giá trị Q tăng lớn nhất, và (ii) gom các cộng đồng lại thành một siêu đồ thị (coarsening) rồi tiếp tục quá trình trên cho đến khi không còn cải thiện được modularity. Nhờ cách tiếp cận này, Louvain đạt hiệu năng rất cao và xử lý được mạng có hàng triệu nút. Tuy nhiên, thuật toán vẫn gặp hạn chế gọi là *resolution limit*, khiến một số cộng đồng nhỏ có thể bị gộp sai vào cộng đồng lớn hơn.

Thuật toán **Leiden** [7] được phát triển nhằm khắc phục nhược điểm này bằng cách bổ sung bước tinh chỉnh đảm bảo các cộng đồng thu được luôn *liên kết chặt chẽ* (well-connected). Phương pháp này giúp giảm hiện tượng cộng đồng rời rạc và cải thiện tính ổn định của kết quả. Trong nhiều thực nghiệm, Leiden thường cho giá trị modularity Q cao hơn hoặc tương đương Louvain, đồng thời tạo ra các ranh giới cộng đồng rõ ràng và hợp lý hơn.

C. Phân cấp tham lam (Fast Greedy)

Thuật toán **Fast Greedy** (Clauset–Newman–Moore) [8] tiến hành gom cụm phân cấp từ dưới lên (agglomerative), mỗi bước trộn hai cụm làm tăng ΔQ nhiều nhất. Dù độ phức tạp có thể nhỉnh hơn cho đồ thị rất lớn, Fast Greedy thường cho kết quả cạnh tranh và dễ diễn giải do sinh ra cây phân cấp (dendrogram) của cấu trúc cộng đồng.

D. Các chỉ số trung tâm (centrality) trong mạng khoa học

Để nhận diện tác giả ảnh hưởng, các chỉ số trung tâm được dùng rộng rãi: **degree** (mức kết nối trực tiếp), **betweenness** (mức độ làm cầu nối trên các đường đi ngắn), **closeness** (độ gần trung bình đến phần còn lại của mạng), và **eigenvector centrality** (ảnh hưởng lan truyền) [9]–[12]. Trong mạng đồng tác giả, các tác giả có betweenness cao thường đóng vai trò kết nối các nhóm chuyên môn khác nhau; eigenvector cao gắn với tầm ảnh hưởng trong các khu vực giàu liên kết.

E. Đánh giá, trực quan hoá và thực hành tốt

Chất lượng phân hoạch thường được đánh giá bằng **modularity** và các thước đo cấu trúc như *coverage*, *conductance*. Khi có nhãn tham chiếu (ví dụ, nhóm trường phái hoặc chủ đề), có thể dùng **NMI** hoặc **ARI** để so sánh [13]. Về trực quan hoá, các kỹ thuật bố trí lực (force–directed) kết hợp tô màu theo cộng đồng giúp diễn giải nhanh cấu trúc đồng thời bảng xếp hạng centrality hỗ trợ nêu bật các tác giả then chốt.

F. Tổng hợp và khoảng trống nghiên cứu

Tổng quan cho thấy Louvain và Leiden là hai chuẩn mực hiện nay cho phát hiện cộng đồng dựa trên modularity ở quy mô lớn Fast Greedy hữu ích khi cần cây phân cấp để phân tích đa mức. Tuy vậy, các nghiên cứu gần đây khuyến nghị **kết hợp phát hiện cộng đồng với phân tích centrality** để vừa thấy cấu trúc nhóm, vừa hiểu vai trò của cá nhân, đặc biệt trong mạng khoa học nơi *các "cầu nối"* ảnh hưởng mạnh đến lan toả tri thức. Khoảng trống còn lại là các nghiên cứu *tái lập* (reproducible) trên nguồn mở như OpenAlex, tập trung riêng vào lĩnh vực Data Science tại bối cảnh Việt Nam đây chính là hướng bài báo này theo đuổi.

III. PHƯƠNG PHÁP NGHIÊN CỨU

Phần này trình bày quy trình và các phương pháp được sử dụng để xây dựng mạng đồng tác giả và phát hiện nhóm hợp tác trong lĩnh vực Khoa học Dữ liệu (Data Science). Quy trình nghiên cứu được triển khai qua năm giai đoạn chính: (1) Thu thập và tiền xử lý dữ liệu, (2) Xây dựng mạng đồng tác giả, (3) Phát hiện cộng đồng, (4) Phân tích trung tâm, và (5) Trực quan hóa kết quả.

A. Thu thập và tiền xử lý dữ liệu

Nguồn dữ liệu được sử dụng là cơ sở dữ liệu học thuật mở **OpenAlex**, cho phép truy cập thông tin công bố khoa học gồm bài báo, tác giả, năm xuất bản, chủ đề và trích dẫn. Dữ liệu được thu thập tự động thông qua OpenAlex API, tập trung vào các bài báo thuộc lĩnh vực **Data Science**, **Machine Learning** và **Artificial Intelligence**.

Quy trình xử lý dữ liệu gồm năm bước:

1) **Xác định nguồn dữ liệu**: Chọn cơ sở dữ liệu OpenAlex và xác định truy vấn phù hợp với lĩnh vực Khoa học Dữ liệu.
2) **Thu thập dữ liệu thô**: Lấy dữ liệu qua API, bao gồm các trường *Work_ID, Title, Year, Cited_by, Authors, Author_IDs, Concepts*.
3) **Làm sạch và chuẩn hoá**: Loại bỏ bản ghi trùng lặp, xử lý giá trị thiếu, tách danh sách tác giả và mã tác giả, định dạng thống nhất các giá trị văn bản và số.
4) **Tạo cấu trúc mạng**: Chuyển dữ liệu bài báo thành các cặp hợp tác giữa tác giả đồng công bố, tạo danh sách cạnh (edges) thể hiện mối quan hệ hợp tác.
5) **Lưu trữ dữ liệu**: Lưu kết quả xử lý dưới dạng bảng gồm thông tin tác giả (nodes) và mối quan hệ hợp tác (edges) để sử dụng cho các bước phân tích tiếp theo.

B. Xây dựng mạng đồng tác giả

Từ tập dữ liệu đã xử lý, mạng đồng tác giả được xây dựng bằng thư viện **NetworkX** trong Python. Đồ thị $G = (V, E)$ được định nghĩa như sau:

- Mỗi nút $v \in V$ biểu diễn một tác giả.
- Hai nút i và j được nối bởi một cạnh $e_{ij} \in E$ nếu họ cùng là đồng tác giả của ít nhất một công trình.
- Trọng số cạnh w_{ij} thể hiện số lượng bài báo mà hai tác giả cùng hợp tác.

Các đặc trưng cơ bản của mạng như *số lượng nút, số cạnh, mật độ, bậc trung bình, hệ số phân cụm* được tính toán để mô tả quy mô và mức độ kết nối của cộng đồng nghiên cứu.

C. Phát hiện cộng đồng

Để xác định các nhóm hợp tác trong mạng, ba thuật toán phát hiện cộng đồng phổ biến được áp dụng gồm:

- **Louvain**: tối ưu hoá chỉ số *modularity* Q bằng cách di chuyển cục bộ từng nút và gom cụm lặp lại cho đến khi hội tụ.
- **Leiden**: cải tiến từ Louvain bằng cách bổ sung giai đoạn tinh chỉnh (*refinement*) nhằm đảm bảo các cộng đồng được kết nối chặt chẽ hơn.
- **Fast Greedy (Clauset–Newman–Moore)**: phương pháp phân cấp, hợp nhất các cụm để tăng modularity tối đa qua từng bước, hình thành cấu trúc phân cấp (*dendrogram*) của mạng.

Chỉ số **modularity** Q được sử dụng để đánh giá chất lượng phân cụm, được xác định theo công thức:

$$Q = \frac{1}{2m} \sum_{ij} \Big(A_{ij} - \frac{k_i k_j}{2m} \Big) \delta(c_i, c_j),$$

trong đó A_{ij} là ma trận kề, k_i và k_j là bậc của các nút i, j, m là tổng số cạnh, và $\delta(c_i, c_j)$ bằng 1 nếu hai nút cùng cộng đồng.

Kết quả các thuật toán được so sánh dựa trên các chỉ số:

- **Modularity (Q)** đánh giá độ tách biệt của các cộng đồng.
- **Normalized Mutual Information (NMI)** đo mức tương đồng giữa các kết quả phân cụm.
- **Adjusted Rand Index (ARI)** phản ánh độ chính xác của việc gán nhãn cộng đồng.

D. Phân tích trung tâm trong mạng

Sau khi phát hiện cộng đồng, các chỉ số trung tâm (centrality) được tính để xác định vai trò và tầm ảnh hưởng của từng tác giả trong mạng:

- **Degree Centrality**: Đo mức độ kết nối trực tiếp.
- **Betweenness Centrality**: Phản ánh vai trò cầu nối giữa các nhóm khác nhau.
- **Closeness Centrality**: Biểu thị khả năng tiếp cận đến các nút khác trong mạng.
- **Eigenvector Centrality**: Đo lường tầm ảnh hưởng toàn cục thông qua liên kết với các tác giả có ảnh hưởng cao.

Các chỉ số này giúp nhận diện những cá nhân có vai trò then chốt trong cộng đồng, cũng như các tác giả kết nối giữa các nhóm nghiên cứu khác nhau.

E. Trực quan hoá kết quả

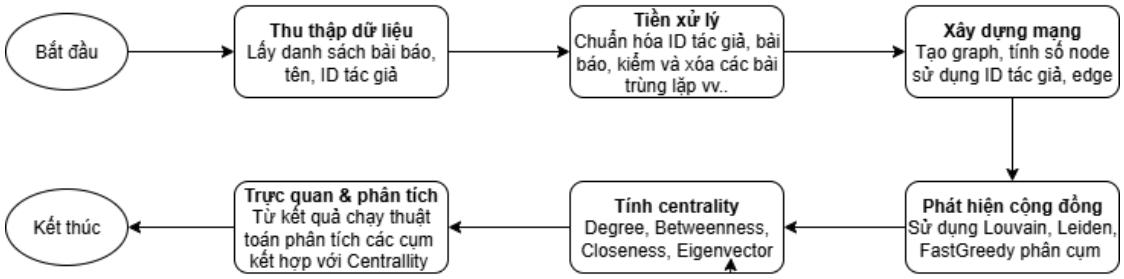
Mạng đồng tác giả và các cộng đồng được trực quan hoá bằng thư viện **NetworkX** và **Matplotlib**. Trong đó:

- Mỗi cộng đồng được gán màu sắc riêng biệt.
- Kích thước nút phản ánh chỉ số ảnh hưởng (Degree hoặc Eigenvector Centrality).
- Các cạnh thể hiện tần suất hợp tác giữa các tác giả.

Việc trực quan hoá giúp quan sát rõ cấu trúc hợp tác, xác định các nhóm nghiên cứu nổi bật và tác giả có ảnh hưởng trung tâm trong mạng, qua đó hỗ trợ phân tích sâu hơn về cấu trúc và động lực cộng tác trong lĩnh vực Khoa học Dữ liệu.

F. Quy trình nghiên cứu tổng quan

Toàn bộ quy trình nghiên cứu được thực hiện qua nhiều giai đoạn liên tiếp, từ thu thập dữ liệu, tiền xử lý, xây dựng mạng đồng tác giả, đến phát hiện cộng đồng và phân tích các chỉ số trung tâm. Quy trình này được minh họa trong Hình 1.



Hình 1. Quy trình tổng quan của nghiên cứu phát hiện và phân tích cộng đồng trong mạng đồng tác giả.

IV. XÂY DỰNG VÀ CHUẨN BỊ DỮ LIỆU

A. Tập dữ liệu

Tập dữ liệu được thu thập từ cơ sở dữ liệu học thuật mở **OpenAlex**, một nền tảng cung cấp thông tin khoa học toàn cầu bao gồm bài báo, tác giả, tổ chức, trích dẫn và chủ đề nghiên cứu. Nghiên cứu tập trung vào lĩnh vực **Khoa học Dữ liệu (Data Science)**, với mục tiêu thu thập dữ liệu về các công trình và tác giả.

Quá trình thu thập dữ liệu được thực hiện tự động thông qua **OpenAlex API**, sử dụng mã lĩnh vực `C2522767166` đại diện cho Data Science. Kết quả thu được khoảng **2.400 công trình khoa học** có liên quan trực tiếp đến chủ đề nghiên cứu. Mỗi bản ghi (work) trong dữ liệu bao gồm các trường thông tin sau:

- **Work_ID**: Mã định danh duy nhất của bài báo.
- **Title**: Tiêu đề công trình.
- **Year**: Năm công bố.
- **Cited_by**: Số lượng trích dẫn.
- **Authors / Author_IDs**: Danh sách tác giả và mã định danh tương ứng.
- **Concepts**: Danh sách các chủ đề học thuật liên quan.

Dữ liệu sau khi thu thập được làm sạch và chuẩn hoá theo các bước sau:

- Loại bỏ các bản ghi trùng lặp theo mã `Work_ID` và tiêu đề `Title`.
- Xử lý các giá trị thiếu ở các trường quan trọng như `Authors`, `Author_IDs` và `Year`.
- Chuẩn hoá danh sách tác giả bằng cách tách tên theo dấu ";", loại bỏ khoảng trắng thừa và định dạng lại chữ viết hoa.
- Chuẩn hoá mã tác giả (`Author_IDs`) bằng cách trích xuất từ đường dẫn OpenAlex (`https://openalex.org/Axxxxxx`) để thu được định dạng ngắn gọn.

- Lọc giữ lại các bài có chủ đề thuộc lĩnh vực **Data Science**, **Machine Learning** hoặc **Artificial Intelligence**.
- Chuyển các trường `Cited_by` và `Year` sang kiểu số để đảm bảo tính nhất quán khi phân tích.

Bảng I
MỘT PHẦN DỮ LIỆU ĐÃ ĐƯỢC LÀM SẠCH VÀ CHUẨN HOÁ TỪ CƠ SỞ OPENALEX

Work_ID	Title	Year	Cited_by	Authors	Author_IDs
W2144634347	Molecular Cloning: A Laboratory Manual	2001	133517	Joseph Sambrook; Elisabeth Fritsch; Tom Maniatis	A5112152140; A5047742196; A5091116725
W4300870773	Statistical Power Analysis for the Behavioral Sciences	1989	83956	Keith E. Muller; Jacob Cohen	A5110163574; A5102808166
W1658908529	Basics of qualitative research: techniques and procedures for developing grounded theory	1998	39197	Anselm L. Strauss; Juliet Corbin	A5108677812; A5002394578

B. Xây dựng mạng đồng tác giả

Dựa trên dữ liệu đã được làm sạch, nhóm nghiên cứu tiến hành xây dựng **mạng đồng tác giả (co-author network)** bằng thư viện **NetworkX** trong Python. Mạng được biểu diễn dưới dạng đồ thị vô hướng $G = (V, E)$, trong đó:

- Mỗi **nút (node)** $v \in V$ đại diện cho một tác giả.
- Hai **nút được nối (edge)** nếu hai tác giả cùng là đồng tác giả của ít nhất một bài báo.
- **Trọng số cạnh (weight)** w_{ij} thể hiện số lượng bài báo mà hai tác giả cùng hợp tác.

Các chỉ số đặc trưng của mạng bao gồm: **Số lượng nút**, **số cạnh**, **mật độ (Density)**, **bậc trung bình (Average Degree)**, và **hệ số phân cụm (Clustering Coefficient)**. Những chỉ số này giúp đánh giá quy mô, mức độ gắn kết và cấu trúc hợp tác của cộng đồng nghiên cứu trong lĩnh vực Khoa học Dữ liệu.

Bảng II
CÁC CHỈ SỐ THỐNG KÊ CƠ BẢN CỦA MẠNG ĐỒNG TÁC GIẢ TRONG LĨNH VỰC KHOA HỌC DỮ LIỆU

Số nút (Nodes)	Số cạnh (Edges)	Mật độ (Density)	Bậc trung bình (Average Degree)	Hệ số phân cụm (Clustering Coefficient)
9,212	136,198	0.0032	29.57	0.8608

C. So sánh các mô hình cơ sở

Sau khi xây dựng mạng, nhóm nghiên cứu áp dụng ba thuật toán phát hiện cộng đồng phổ biến để xác định các nhóm hợp tác chính trong mạng đồng tác giả:

- **Louvain**: Tối ưu hoá chỉ số modularity bằng cách gom cụm lặp lại cho đến khi hội tụ.
- **Leiden**: Phiên bản cải tiến của Louvain, đảm bảo các cộng đồng được kết nối chặt chẽ và ổn định hơn.
- **Fast Greedy (Clauset–Newman–Moore)**: Phương pháp phân cấp, hợp nhất các cụm sao cho tăng modularity nhiều nhất.

Hiệu quả của từng thuật toán được đánh giá thông qua các chỉ số:

- **Modularity (Q)** đo lường mức độ tách biệt giữa các cộng đồng.
- **Normalized Mutual Information (NMI)** đánh giá sự tương đồng giữa các kết quả phân cụm.
- **Adjusted Rand Index (ARI)** đo độ chính xác so với cấu trúc tham chiếu.

Kết quả được trực quan hoá bằng **NetworkX** và **Matplotlib**. Các cộng đồng được tô màu khác nhau để thể hiện ranh giới, trong khi các tác giả có chỉ số trung tâm cao (*Degree* hoặc *Eigenvector Centrality*) được làm nổi bật, thể hiện vai trò then chốt trong mạng hợp tác khoa học.

V. THỰC NGHIỆM VÀ ĐÁNH GIÁ

A. Cấu hình và môi trường thực nghiệm

Các thí nghiệm được thực hiện bằng ngôn ngữ **Python 3.10** với các thư viện **NetworkX**, **community-louvain**, **Matplotlib**, **Pandas** và **NumPy**. Hệ thống thử nghiệm gồm **CPU Intel Core i7**, **RAM 16GB** và hệ điều hành **Windows 11**.

Mạng đồng tác giả thu được có quy mô **9.212 nút** và **136.198 cạnh**, với mật độ 0.0032 và hệ số phân cụm trung bình 0.8608. Đây là một mạng lớn, thưa (sparse) nhưng có tính liên kết chặt chẽ nội bộ, phản ánh đặc trưng của các nhóm nghiên cứu trong lĩnh vực Khoa học Dữ liệu.

Hình 2 minh họa mạng đồng tác giả gồm 2000 nhà nghiên cứu đầu tiên được trích xuất từ dữ liệu OpenAlex, thể hiện sự phân bố và mức độ kết nối giữa các tác giả trong lĩnh vực Khoa học Dữ liệu.

Hình 2. Mối liên kết hợp tác giữa 2000 tác giả đầu tiên trong tập dữ liệu.

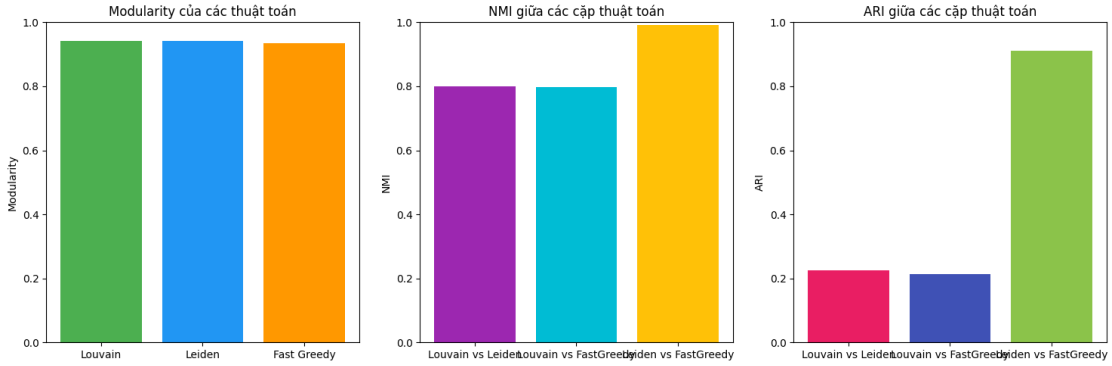
B. Kết quả phát hiện cộng đồng

Ba thuật toán được áp dụng gồm **Louvain**, **Leiden** và **Fast Greedy**. Kết quả cho thấy các thuật toán đều đạt giá trị modularity cao ($Q > 0.93$), chứng tỏ mạng có cấu trúc cộng đồng rõ rệt.

Bảng III
SO SÁNH KẾT QUẢ GIỮA CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG

Thuật toán	**Số cộng đồng**	**Modularity (Q)**
Louvain	1,167	0.9416
Leiden	1,167	**0.9418**
Fast Greedy	1,164	0.9344

Hình 3 minh họa giá trị **Modularity**, **NMI** và **ARI** giữa các thuật toán. Ta thấy Leiden đạt modularity cao nhất, trong khi Louvain và Fast Greedy cho kết quả tương đồng, chênh lệch không đáng kể.



Hình 3. So sánh các chỉ số Modularity, NMI và ARI giữa các thuật toán.

C. Phân tích mức tương đồng giữa các thuật toán

Để đánh giá độ tương đồng giữa các phân hoạch, hai chỉ số **Normalized Mutual Information (NMI)** và **Adjusted Rand Index (ARI)** được sử dụng.

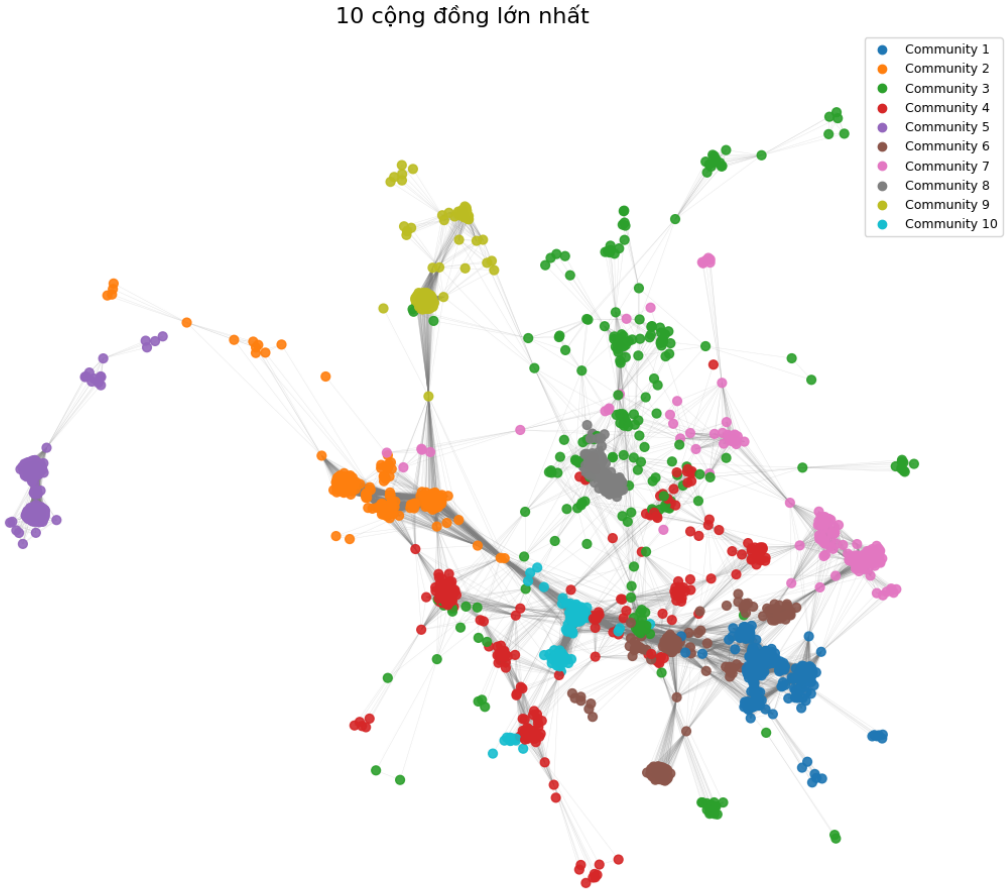
Bảng IV
ĐÁNH GIÁ MỨC TƯƠNG ĐỒNG GIỮA CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG

Cặp thuật toán	NMI	ARI
Louvain - Leiden	0.8004	0.2263
Louvain - Fast Greedy	0.7971	0.2146
Leiden - Fast Greedy	**0.9909**	**0.9119**

Hai thuật toán **Leiden** và **Fast Greedy** cho ra kết quả gần như trùng khớp (NMI = 0.99, ARI = 0.91), chứng tỏ hai phương pháp này phân hoạch mạng gần như tương đương. Louvain khác biệt nhẹ do chiến lược tối ưu hoá modularity cục bộ, nhưng vẫn duy trì mức tương đồng cao (NMI 0.8).

D. Cấu trúc cộng đồng và trực quan hóa

Mặc dù tổng cộng có hơn 1.100 cộng đồng, nhưng các cộng đồng lớn chiếm phần lớn số lượng tác giả. Hình 4 trực quan hóa **10 cộng đồng lớn nhất** theo thuật toán Leiden.

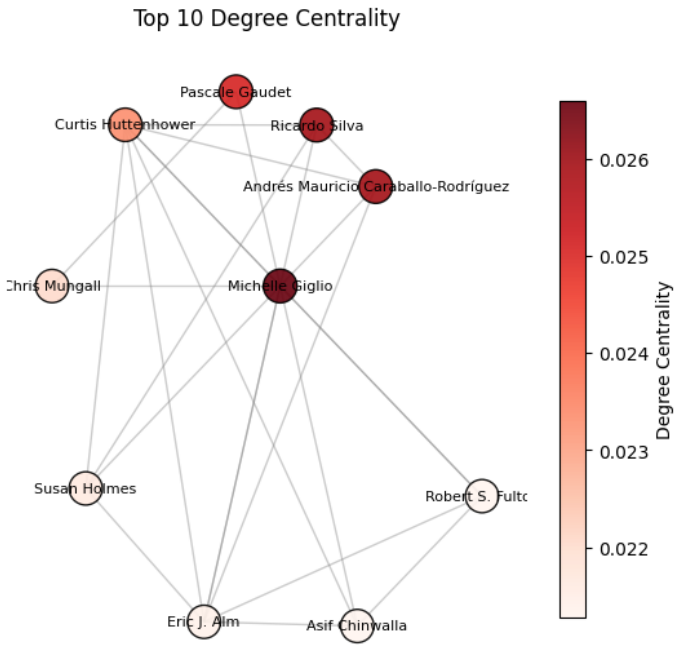


Hình 4. 10 cộng đồng lớn nhất trong mạng đồng tác giả.

Các cộng đồng này có quy mô từ 139 đến 345 tác giả, tương ứng với các nhóm nghiên cứu hoặc phòng thí nghiệm lớn trong lĩnh vực học máy, khai phá dữ liệu và trí tuệ nhân tạo. Cấu trúc cho thấy các cộng đồng liên kết chặt chẽ nội bộ và chỉ có một số cạnh kết nối giữa các nhóm phản ánh rõ nét tính chuyên biệt của từng hướng nghiên cứu.

E. Phân tích chỉ số trung tâm (Centrality Analysis)

Để nhận diện các tác giả có tầm ảnh hưởng lớn, bốn chỉ số trung tâm được tính toán gồm **Degree**, **Betweenness**, **Closeness** và **Eigenvector Centrality**. Các biểu đồ Hình 5, 6, 7, 8 minh họa top 10 tác giả theo từng chỉ số.

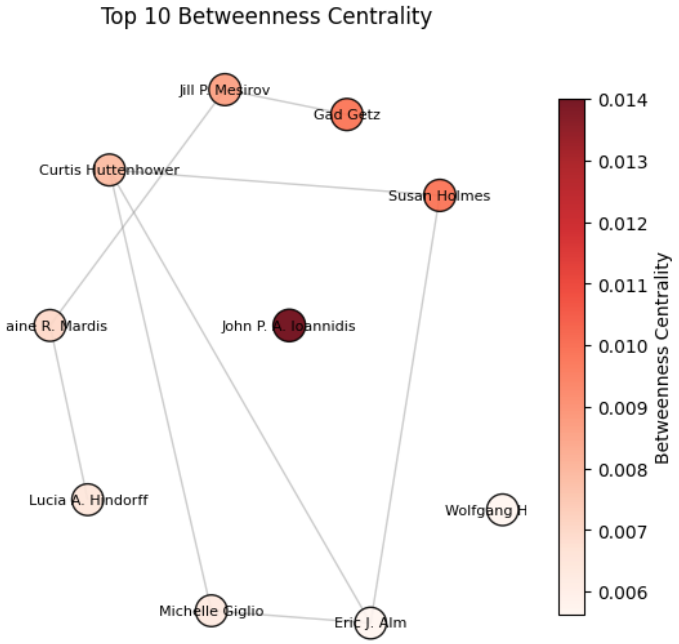


Hình 5. Top 10 tác giả có chỉ số Degree Centrality cao nhất.

Hình 5 minh họa **10 tác giả có chỉ số Degree Centrality cao nhất** trong mạng đồng tác giả, tức là những người có **số lượng kết nối trực tiếp với đồng nghiệp nhiều nhất**. Các tác giả như *Michelle Giglio, Ricardo Silva, Pascale Gaudet* và *Andrés Mauricio Carballo-Rodríguez* nổi bật với các nút màu đỏ đậm, thể hiện mức độ hợp tác dày đặc và phạm vi ảnh hưởng rộng.

Những cá nhân này không chỉ thường xuyên xuất hiện trong nhiều công trình hợp tác, mà còn đóng vai trò trung tâm trong việc **kết nối các nhóm nghiên cứu khác nhau**, giúp dòng chảy tri thức lan tỏa nhanh hơn trong mạng học thuật. Trong khi đó, một số tác giả như *Robert S. Fultz* hay *Asif Chinyavalla* nằm ở rìa mạng, có mức độ kết nối thấp hơn, chủ yếu hợp tác trong phạm vi một nhóm nhất định.

Kết quả phân tích cho thấy mạng đồng tác giả trong lĩnh vực **Khoa học Dữ liệu** mang cấu trúc **tập trung**, trong đó một nhóm nhỏ các nhà nghiên cứu hoạt động sôi nổi giữ vai trò như những "đầu mối" kết nối, giúp duy trì và mở rộng mối quan hệ hợp tác giữa các cộng đồng học thuật.

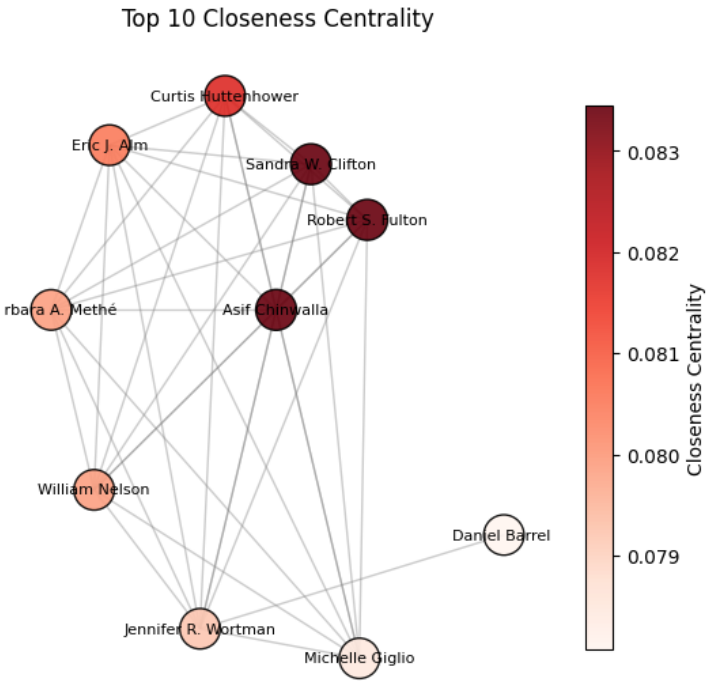


Hình 6. Top 10 tác giả có chỉ số Betweenness Centrality cao nhất.

Hình 6 minh họa **10 tác giả có chỉ số Betweenness Centrality cao nhất** trong mạng đồng tác giả, phản ánh mức độ **ảnh hưởng trung gian** của họ trong việc kết nối giữa các cụm nghiên cứu khác nhau.

Kết quả cho thấy *John P. A. Ioannidis* có chỉ số cao nhất, nằm ở trung tâm và được tô màu đỏ đậm, thể hiện vai trò **cầu nối quan trọng nhất** của toàn mạng. Các tác giả *Susan Holmes*, *Gad Getz* và *Jill P. Mesirov* lần lượt giữ các vị trí tiếp theo, đóng vai trò liên kết các cụm học giả riêng biệt, giúp dòng chảy thông tin và hợp tác diễn ra xuyên suốt trong mạng lưới. Trong khi, *Curtis Huttenhower*, *Elaine R. Mardis* và các tác giả ở rìa như *Lucia A. Hindorff* hay *Wolfgang H.* có chỉ số thấp hơn, cho thấy họ chủ yếu hoạt động trong các nhóm hợp tác nhỏ, ít khi đóng vai trò trung gian giữa các cộng đồng lớn.

Phân tích Betweenness Centrality chỉ ra rằng mạng hợp tác trong lĩnh vực **Khoa học Dữ liệu** được duy trì bởi một số ít học giả có tầm ảnh hưởng cao, đóng vai trò **kết nối các cụm nghiên cứu độc lập** và thúc đẩy **sự lan tỏa tri thức** trong toàn bộ hệ sinh thái nghiên cứu.

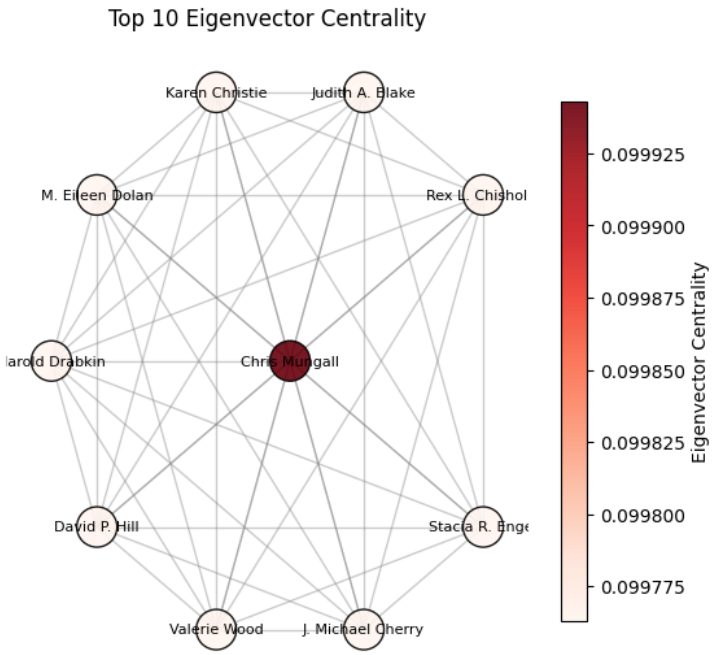


Hình 7. Top 10 tác giả có chỉ số Closeness Centrality cao nhất.

Hình 7 trình bày **10 tác giả có chỉ số Closeness Centrality cao nhất**, phản ánh mức độ **gần gũi trung bình của mỗi tác giả tới các nút khác** trong toàn mạng lưới hợp tác. Các giá trị closeness cao cho thấy những nhà nghiên cứu này có khả năng tiếp cận và lan tỏa thông tin nhanh chóng trong mạng.

Trong đó, các tác giả *Asif Chinyavalla*, *Sandra W. Clifton* và *Robert S. Fulton* nổi bật với màu đỏ đậm nhất, biểu thị vị trí trung tâm có khoảng cách ngắn nhất đến phần lớn các tác giả khác. Các nhà nghiên cứu như *Curtis Huttenhower*, *Eric J. Alm* và *Barbara A. Methé* cũng có closeness tương đối cao, đóng vai trò như các "điểm trung chuyển" giúp kết nối nhanh giữa các cụm hợp tác. Ngược lại, những tác giả nằm ở rìa mạng như *Daniel Barrell* hay *Michelle Giglio* có giá trị closeness thấp hơn, cho thấy họ chủ yếu hoạt động trong phạm vi nhóm hẹp và ít liên kết với các cộng đồng khác.

Tổng thể, phân tích Closeness Centrality cho thấy mạng hợp tác trong lĩnh vực **Khoa học Dữ liệu** có cấu trúc **liên thông hiệu quả**, trong đó một số tác giả đóng vai trò trọng tâm giúp **rút ngắn khoảng cách tri thức** và **tăng tốc độ lan tỏa thông tin** giữa các cộng đồng nghiên cứu.



Hình 8. Top 10 tác giả có chỉ số Eigenvector Centrality cao nhất.

Hình 8 mô tả **10 tác giả có chỉ số Eigenvector Centrality cao nhất**, thể hiện mức độ **ảnh hưởng toàn cục** trong mạng đồng tác giả. Khác với Degree Centrality chỉ xét số lượng kết nối trực tiếp, Eigenvector Centrality đánh giá cả *chất lượng của mối liên kết*, nghĩa là một tác giả có ảnh hưởng nếu họ được kết nối với những cá nhân cũng có tầm ảnh hưởng cao.

Từ hình có thể thấy *Chris Mungall* là tác giả nổi bật nhất, nằm ở trung tâm mạng lưới và có chỉ số cao nhất, cho thấy vai trò như **trục lan tỏa thông tin** của toàn hệ thống. Bao quanh ông là những cộng sự có sức ảnh hưởng đáng kể như *Karen Christie*, *Judith A. Blake*, *Valerie Wood* và *Rex L. Chisholm*, tạo thành một cụm hợp tác bền chặt với mật độ liên kết cao. Các tác giả này không chỉ kết nối chặt chẽ với Mungall mà còn duy trì quan hệ hợp tác chéo lẫn nhau, giúp hình thành một **mạng lõi học thuật ổn định**.

Nhìn tổng thể, chỉ số Eigenvector Centrality phản ánh rõ nét **cấu trúc trung tâm liên kết mạnh** của mạng nghiên cứu trong lĩnh vực **Khoa học Dữ liệu**, nơi một số tác giả giữ vai trò dẫn dắt và củng cố tính gắn kết của toàn bộ cộng đồng hợp tác.

Kết quả phân tích các chỉ số trung tâm cho thấy:

- **Betweenness Centrality:** Các tác giả như *John P. Ioannidis*, *Curtis Huttenhower* và *Susan Holmes* giữ vai trò **cầu nối quan trọng** giữa các nhóm nghiên cứu khác nhau. Họ giúp luân chuyển tri thức và kết nối các cụm học thuật vốn tách biệt, đóng vai trò như những "điểm trung gian chiến lược" trong toàn mạng.
- **Closeness Centrality:** Những tác giả có chỉ số closeness cao thường xuất hiện ở **vị trí trung tâm của cộng đồng**, có khả năng tiếp cận nhanh đến các thành viên khác trong mạng. Điều này cho thấy họ có vai trò then chốt trong việc truyền tải thông tin và duy trì sự gắn kết giữa các nhóm hợp tác.
- **Eigenvector Centrality:** Các tác giả như *Chris Mungall* và *Karen Christie* nổi bật với mạng lưới liên kết mật thiết cùng nhiều nhà nghiên cứu có ảnh hưởng cao khác, hình thành nên **"cụm hạt nhân"** của toàn hệ thống hợp tác. Họ là những cá nhân có khả năng tác động lan tỏa mạnh đến cấu trúc chung của mạng khoa học.

F. Thảo luận kết quả

Các kết quả thực nghiệm cho thấy mạng đồng tác giả trong lĩnh vực **Khoa học Dữ liệu** có cấu trúc cộng đồng rõ ràng và tính gắn kết cao. Giá trị **modularity** đạt khoảng $Q \approx 0.94$ chứng minh rằng các nhóm nghiên cứu được hình thành một cách tự nhiên, với mật độ liên kết nội bộ lớn và chỉ có ít kết nối giữa các cộng đồng khác nhau.

Thuật toán **Leiden** thể hiện hiệu suất ổn định nhất, đặc biệt khi xử lý mạng có quy mô lớn nhờ khả năng tinh chỉnh cộng đồng giúp đảm bảo tính liên kết chặt chẽ. Trong khi đó, **Fast Greedy** có ưu điểm về tốc độ và khả năng khái quát nhanh cấu trúc phân cấp, dù giá trị modularity thấp hơn đôi chút so với Leiden. Kết quả này cho thấy mỗi phương pháp có thế mạnh riêng Leiden phù hợp cho đánh giá chi tiết, còn Fast Greedy thích hợp trong các bài toán cần tối ưu thời gian tính toán.

Cấu trúc mạng thể hiện sự **đa dạng về chủ đề và hướng nghiên cứu**, với hàng nghìn cộng đồng quy mô nhỏ phản ánh sự phân hóa của các chuyên ngành trong Data Science như học máy, khai phá dữ liệu, và trí tuệ nhân tạo. Đặc biệt, các tác giả có chỉ số **Betweenness Centrality** cao đóng vai trò như những "cầu nối tri thức", giúp kết nối các cụm nghiên cứu độc lập, từ đó hình thành mạng hợp tác mang tính **đa ngành và liên lĩnh vực**.

Tổng thể, kết quả nghiên cứu khẳng định tính hiệu quả của việc kết hợp **phát hiện cộng đồng** và **phân tích trung tâm** trong việc khám phá cấu trúc hợp tác khoa học. Thông qua trực quan hóa mạng, có thể nhận diện rõ **những nhóm nghiên cứu lớn, các tác giả chủ chốt** và **mối quan hệ liên kết giữa các hướng nghiên cứu**, góp phần mang lại cái nhìn toàn diện hơn về hệ sinh thái học thuật trong lĩnh vực Khoa học Dữ liệu.

VI. THẢO LUẬN

Kết quả phân tích mạng đồng tác giả trong lĩnh vực **Khoa học Dữ liệu** cho thấy hệ thống hợp tác giữa các nhà nghiên cứu được tổ chức chặt chẽ và có tính phân tầng cao. Giá trị modularity lớn ($Q > 0.93$) phản ánh rõ sự tồn tại của các nhóm hợp tác ổn định, hình thành dựa trên định hướng học thuật, lĩnh vực chuyên sâu và mối liên hệ nghề nghiệp lâu dài. Phần này thảo luận sâu hơn về những phát hiện, ý nghĩa và hướng ứng dụng thực tiễn của nghiên cứu.

A. Cấu trúc hợp tác và ý nghĩa học thuật

Các kết quả cho thấy mạng đồng tác giả trong lĩnh vực Khoa học Dữ liệu có xu hướng hình thành theo mô hình **trung tâm–vệ tinh (hub–spoke)**. Một số cộng đồng lớn giữ vai trò "hạt nhân tri thức", nơi tập trung các nhóm chuyên sâu về *Machine Learning*, *Data Mining* hay *Artificial Intelligence*. Các nhóm nhỏ hơn đóng vai trò vệ tinh, vừa học hỏi, vừa kết nối tri thức giữa các nhánh nghiên cứu khác nhau. Cấu trúc này không chỉ phản ánh quy luật hình thành tự nhiên của mạng học thuật, mà còn cho thấy sự phân hóa rõ ràng giữa các trung tâm nghiên cứu lớn và nhóm độc lập mới nổi.

B. Phân tích hiệu quả thuật toán và vai trò các cá nhân ảnh hưởng

Khi so sánh các thuật toán phát hiện cộng đồng, kết quả cho thấy **Leiden** đạt hiệu suất và độ ổn định cao nhất, phù hợp với các mạng có quy mô lớn và cấu trúc phức tạp. Điều này phù hợp với kết luận của Traag et al. [14], khẳng định khả năng tinh chỉnh cộng đồng giúp Leiden duy trì tính liên kết chặt chẽ nội bộ. Ngoài ra, việc phân tích các chỉ số trung tâm như **Degree**, **Betweenness** và **Eigenvector Centrality** giúp xác định các tác giả có tầm ảnh hưởng lớn trong mạng, đặc biệt là những người đóng vai trò cầu nối giữa các cụm nghiên cứu khác nhau. Những cá nhân này góp phần thúc đẩy sự lan tỏa tri thức và gắn kết cộng đồng học thuật theo chiều sâu.

C. Lan tỏa tri thức và định hình xu hướng nghiên cứu

Cấu trúc mạng đồng tác giả không chỉ thể hiện mối liên hệ hợp tác, mà còn phản ánh cách **tri thức được truyền đạt và khuếch tán** trong cộng đồng khoa học. Các nhóm trung tâm thường là nơi khởi phát các hướng nghiên cứu mới, công nghệ mới hoặc phương pháp luận tiên phong, trong khi các cụm vệ tinh đóng vai trò lan tỏa, mở rộng và ứng dụng kết quả. Nhờ vậy, có thể xem phát hiện cộng đồng như một công cụ để **theo dõi dòng chảy tri thức**, xác định các "điểm phát tán" và dự đoán xu hướng phát triển của lĩnh vực trong tương lai gần. Điều này đặc biệt hữu ích cho các tổ chức khoa học khi muốn định hướng chiến lược đầu tư, hợp tác và phát triển nhóm nghiên cứu tiềm năng.

D. Giới hạn và hướng mở rộng

Mặc dù kết quả mang lại nhiều giá trị, nghiên cứu hiện vẫn tập trung vào phân tích mạng tĩnh, chưa xem xét yếu tố thời gian trong quá trình hình thành và thay đổi mối quan hệ hợp tác. Trong thực tế, mạng học thuật luôn biến động các nhóm mới xuất hiện, hợp nhất hoặc tan rã theo từng giai đoạn. Vì vậy, hướng phát triển tiếp theo là mở rộng sang phân tích **mạng động (dynamic network)** để theo dõi sự tiến hóa của hệ thống nghiên cứu. Bên cạnh đó, việc tích hợp thêm thông tin về **chủ đề bài báo, trích dẫn và độ ảnh hưởng** sẽ giúp mô hình phát hiện cộng đồng phản ánh sâu hơn mối quan hệ giữa cấu trúc hợp tác và tác động khoa học. Các hướng mở này không chỉ nâng cao giá trị phân tích mà còn mở ra tiềm năng ứng dụng trong đánh giá hiệu quả và hoạch định chính sách khoa học.

Phần thảo luận cho thấy việc kết hợp phát hiện cộng đồng và phân tích trung tâm mang lại một góc nhìn toàn diện hơn về cách tri thức được hình thành, duy trì và lan tỏa trong lĩnh vực Khoa học Dữ liệu. Đây không chỉ là công cụ mô tả mạng lưới hợp tác, mà còn là nền tảng cho việc dự đoán xu hướng, định hướng chiến lược và phát triển cộng đồng khoa học trong tương lai.

VII. KẾT LUẬN

Nghiên cứu đã tiến hành xây dựng và phân tích mạng lưới tác giả trong lĩnh vực **Khoa học Dữ liệu** dựa trên dữ liệu OpenAlex, qua đó làm rõ cách các nhà khoa học hình thành mối quan hệ chuyên môn và chia sẻ tri thức. Kết quả cho thấy cấu trúc mạng có tính phân cụm cao, với chỉ số modularity đạt gần 0.94, thể hiện sự tồn tại của những nhóm nghiên cứu có tính gắn kết mạnh và ranh giới rõ rệt. Thuật toán **Leiden** cho hiệu quả ổn định và khả năng tối ưu vượt trội, trong khi **Fast Greedy** tỏ ra linh hoạt và tiết kiệm thời gian xử lý.

Phân tích các chỉ số trung tâm cho thấy một số tác giả giữ vai trò trọng yếu trong việc lan truyền thông tin và kết nối giữa các nhóm, tiêu biểu là *John P. Ioannidis, Chris Mungall* và *Curtis Huttenhower*. Những cá nhân này không chỉ là đầu mối giao thoa giữa các hướng nghiên cứu, mà còn góp phần duy trì sự liên thông của toàn bộ mạng học thuật trong lĩnh vực khoa học dữ liệu.

Nhìn tổng thể, công trình này mang lại cái nhìn trực quan về cấu trúc cộng tác học thuật, đồng thời chứng minh tiềm năng của việc kết hợp các phương pháp phát hiện cộng đồng với phân tích chỉ số trung tâm trong việc khám phá đặc trưng của mạng phức tạp. Kết quả thu được có thể làm nền tảng cho nhiều ứng dụng tiếp theo, như dự báo mối quan hệ hợp tác mới, đánh giá mức độ ảnh hưởng của từng tác giả, hoặc theo dõi sự phát triển của các hướng nghiên cứu theo thời gian.

Trong tương lai, có thể mở rộng hướng tiếp cận bằng cách tích hợp thêm dữ liệu động theo giai đoạn hoặc sử dụng các kỹ thuật học sâu trên đồ thị như **Graph Neural Networks (GNNs)** để tăng khả năng biểu diễn và dự đoán cấu trúc mạng. Bên cạnh đó, việc phân tích ở nhiều lĩnh vực khoa học khác nhau sẽ giúp kiểm chứng khả năng tổng quát và ứng dụng thực tiễn của phương pháp.

TÀI LIỆU THAM KHẢO

[1] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[2] ——, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl 1, pp. 5200–5205, 2004.

[3] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," *arXiv preprint arXiv:2205.01833*, 2022.

[4] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.

[5] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75–174, 2010.

[6] 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.

[7] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific Reports*, vol. 9, p. 5233, 2019.

[8] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, p. 066111, 2004.

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

[10] ——, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[11] P. Bonacich, "Power and centrality: A family of measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[12] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social Networks*, vol. 27, no. 1, pp. 39–54, 2005.

[13] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, p. P09008, 2005.

[14] 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.