PROJECT 1B: DỰ ĐOÁN HỢP TÁC NGHIÊN CỨU TƯƠNG LAI VÀ KHUYẾN NGHỊ ĐỐI TÁC TRONG MẠNG LƯỚI KHOA HỌC

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 bối cảnh khoa học hiện đại phát triển mạnh mẽ, việc xác định và dự đoán các mối quan hệ hợp tác giữa các nhà nghiên cứu đóng vai trò quan trọng trong việc thúc đẩy đổi mới sáng tạo và tối ưu hóa nguồn lực khoa học. Nghiên cứu này tập trung xây dựng một mô hình dự đoán hợp tác khoa học tương lai dựa trên phân tích cấu trúc mạng đồng tác giả. Dữ liệu được thu thập từ các công bố học thuật trong lĩnh vực Khoa học Dữ liệu, sau đó được biểu diễn dưới dạng đồ thị, trong đó các nút đại diện cho tác giả và các cạnh biểu thị mối quan hệ hợp tác.

Phương pháp đề xuất kết hợp các chỉ số tương đồng cổ điển như *Common Neighbors*, *Jaccard Coefficient*, *Adamic-Adar Index*, *Preferential Attachment* và *Resource Allocation* với kỹ thuật học biểu diễn đồ thị *Node2Vec* nhằm đánh giá khả năng hình thành liên kết mới giữa các nhà khoa học. Các mô hình được huấn luyện và kiểm định bằng cách tách tập dữ liệu thành phần huấn luyện (80%) và kiểm tra (20%), với việc sử dụng các chỉ số đánh giá như AUC-ROC và AUC-PR.

Kết quả thực nghiệm cho thấy các phương pháp dựa trên cấu trúc cục bộ đạt hiệu năng vượt trội với AUC xấp xỉ 1.0, chứng minh khả năng dự đoán chính xác các mối hợp tác tiềm năng trong cùng cộng đồng nghiên cứu. Trong khi đó, *Node2Vec* cho thấy tiềm năng trong việc phát hiện các mối liên kết toàn cục nhưng cần tối ưu thêm về siêu tham số. Nghiên cứu góp phần cung cấp một khung phân tích hiệu quả cho việc khuyến nghị đối tác nghiên cứu, đồng thời mở ra hướng phát triển các hệ thống gợi ý cộng tác học thuật thông minh trong tương lai.

TỪ KHÓA

Social Network Analysis, Link Prediction, Research Collaboration Recommendation, Co-author Network, Node2Vec, Machine Learning.

I. GIỚI THIỆU

Trong kỷ nguyên khoa học mở và toàn cầu hóa nghiên cứu, hợp tác giữa các nhà khoa học trở thành yếu tố then chốt thúc đẩy đổi mới sáng tạo, tăng cường năng suất công bố và mở rộng tầm ảnh hưởng học thuật. Việc hiểu rõ cấu trúc hợp tác và khả năng hình thành các mối quan hệ mới giữa các tác giả không chỉ mang giá trị học thuật mà còn giúp định hướng chính sách nghiên cứu, phân bổ nguồn lực và xây dựng mạng lưới khoa học bền vững.

Phân tích mạng đồng tác giả (Co-author Network Analysis) là một trong những hướng tiếp cận hiệu quả để khám phá mô hình hợp tác giữa các nhà nghiên cứu. Trong mạng này, các nút biểu diễn tác giả và các cạnh thể hiện mối quan hệ cộng tác thông qua các bài báo khoa học. Dựa trên cấu trúc mạng, bài toán *dự đoán liên kết* (Link Prediction) cho phép xác định khả năng hai nhà nghiên cứu sẽ hợp tác trong tương lai dựa trên các đặc trưng cấu trúc và hành vi hợp tác hiện tại.

Các phương pháp dự đoán liên kết truyền thống như *Common Neighbors*, *Jaccard Coefficient*, *Adamic-Adar*, *Preferential Attachment* hay *Resource Allocation* đã chứng minh hiệu quả trong việc nhận diện các mối quan hệ tiềm năng trong mạng xã hội. Gần đây, sự phát triển của kỹ thuật học biểu diễn đồ thị (*Graph Embedding*) như *Node2Vec* đã mở ra hướng tiếp cận mới giúp mô hình hóa không gian đặc trưng phức tạp của mạng, từ đó cải thiện khả năng dự đoán và khuyến nghị hợp tác.

Nghiên cứu này hướng tới mục tiêu xây dựng một mô hình dự đoán hợp tác nghiên cứu tương lai kết hợp giữa các chỉ số tương đồng cổ điển và kỹ thuật học biểu diễn đồ thị, từ đó khuyến nghị các đối tác tiềm năng trong mạng lưới khoa học. Cụ thể, mô hình được áp dụng trên dữ liệu mạng đồng tác giả trong lĩnh vực Khoa học Dữ liệu, với quy trình gồm các bước: Tiền xử lý dữ liệu, chia tập huấn luyện - kiểm tra, tính toán các chỉ số liên kết, huấn luyện mô hình, và đánh giá bằng các chỉ số AUC-ROC, AUC-PR.

Kết quả thu được không chỉ chứng minh hiệu quả vượt trội của các chỉ số dựa trên cấu trúc cục bộ mà còn cho thấy tiềm năng của học biểu diễn đồ thị trong việc phát hiện các mối quan hệ hợp tác toàn cục. Từ đó, nghiên cứu góp phần tạo nền tảng cho việc phát triển các hệ thống gợi ý cộng tác học thuật thông minh, hỗ trợ các nhà nghiên cứu kết nối và mở rộng mạng lưới khoa học hiệu quả hơn.

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 một dạng đồ thị trong đó mỗi nút biểu diễn một tác giả và mỗi cạnh thể hiện mối quan hệ hợp tác khi hai tác giả cùng công bố ít nhất một bài báo. Nhiều nghiên cứu đã chứng minh mạng khoa học có đặc trưng phân cụm mạnh, cấu trúc cộng đồng rõ rệt và tồn tại các cá nhân "cầu nối" đóng vai trò quan trọng trong lan truyền tri thức [1], [2]. Gần đây, các cơ sở dữ liệu học thuật mở như **OpenAlex** hay **Microsoft Academic Graph** cho phép truy cập dữ liệu quy mô lớn về bài báo, tác giả, trích dẫn và chủ đề nghiên cứu, từ đó hỗ trợ đáng kể cho việc phân tích, dự đoán và khuyến nghị hợp tác khoa học [3].

B. Bài toán dự đoán liên kết trong mạng xã hội

Giả sử đồ thị vô hướng $G = (V, E)$, trong đó V là tập các nút và E là tập các cạnh. Các phương pháp dự đoán liên kết cổ điển thường tính điểm tương đồng giữa hai nút dựa trên số lượng hàng xóm chung hoặc cấu trúc lân cận, tiêu biểu gồm **Common Neighbors (CN)**, **Jaccard Coefficient (JC)**, **Adamic–Adar Index (AA)**, **Preferential Attachment (PA)** và **Resource Allocation (RA)** [4]. Các công thức chi tiết được trình bày trong Mục III.

C. Học biểu diễn đồ thị và mô hình Node2Vec

Kỹ thuật học biểu diễn đồ thị (graph embedding) mở ra hướng tiếp cận mới cho dự đoán liên kết khi cho phép ánh xạ mỗi nút thành vector đặc trưng trong không gian liên tục, bảo toàn thông tin cấu trúc và ngữ cảnh. Các mô hình như *DeepWalk* [5] và *Node2Vec* [6] khai thác cơ chế bước đi ngẫu nhiên có trọng số để học mối quan hệ giữa các nút. Trong mạng hợp tác khoa học, *Node2Vec* được chứng minh có khả năng phát hiện các mối quan hệ tiềm ẩn vượt ra ngoài cộng đồng trực tiếp, giúp khắc phục hạn chế của các chỉ số tương đồng cục bộ.

D. Các chỉ số đánh giá và trực quan hóa

Hiệu năng của mô hình dự đoán liên kết thường được đo lường bằng các chỉ số chuẩn như **AUC-ROC** và **AUC-PR** để đánh giá khả năng phân biệt giữa các cạnh thật và cạnh giả [7]. Ngoài ra, một số nghiên cứu sử dụng Precision@K hoặc Recall@K nhằm phản ánh chất lượng khuyến nghị ở mức thực tế. Các biểu đồ ROC, PR, histogram và boxplot được sử dụng rộng rãi để minh họa phân bố điểm tương đồng, giúp so sánh trực quan hiệu quả giữa các thuật toán.

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 [8]. 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. Hệ thống khuyến nghị hợp tác khoa học

Dựa trên kết quả dự đoán liên kết, nhiều hệ thống *Research Collaboration Recommendation* đã được phát triển để tự động gợi ý đối tác tiềm năng. Các mô hình này thường kết hợp thông tin cấu trúc mạng với dữ liệu hồ sơ học thuật, lĩnh vực nghiên cứu hoặc từ khóa bài báo [9]. Một số hướng hiện đại tích hợp học sâu trên văn bản (text embedding) và đồ thị (graph embedding) nhằm tăng độ chính xác và khả năng mở rộng của hệ thống khuyến nghị.

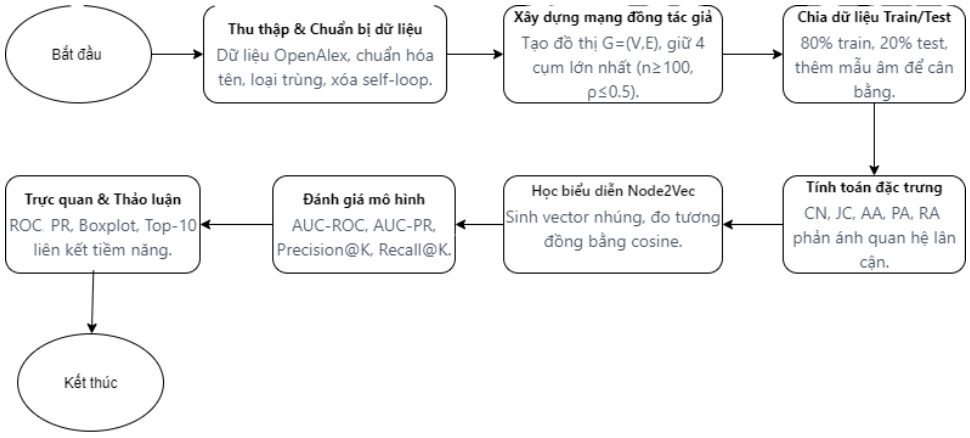
G. Tổng hợp và khoảng trống nghiên cứu

Tổng quan cho thấy các chỉ số tương đồng cổ điển vẫn mang lại độ chính xác cao và dễ diễn giải, trong khi các phương pháp học biểu diễn đồ thị có ưu thế trong việc khám phá quan hệ hợp tác mới ở phạm vi rộng hơn. Tuy nhiên, các nghiên cứu tái lập trên nguồn dữ liệu mở quy mô lớn, đặc biệt trong lĩnh vực Khoa học Dữ liệu tại Việt Nam, còn hạn chế. Bài báo này hướng tới việc kết hợp hai hướng tiếp cận trên để xây dựng mô hình dự đoán hợp tác nghiên cứu có độ chính xác cao, đồng thời làm cơ sở cho hệ thống khuyến nghị đối tác khoa học trong tương lai.

III. PHƯƠNG PHÁP NGHIÊN CỨU

Phần này trình bày quy trình nghiên cứu được áp dụng nhằm xây dựng mô hình dự đoán hợp tác nghiên cứu tương lai trong mạng đồng tác giả. Mô hình kết hợp giữa các chỉ số tương đồng cổ điển và kỹ thuật học biểu diễn đồ thị Node2Vec để dự đoán khả năng hình thành liên kết mới giữa các tác giả.

Quy trình tổng thể của nghiên cứu được minh họa trong Hình 1. Quy trình này bao gồm các bước chính từ thu thập và xử lý dữ liệu, xây dựng mạng đồng tác giả, tính toán đặc trưng cấu trúc và học biểu diễn Node2Vec, đến đánh giá mô hình và trực quan hóa kết quả. Mỗi giai đoạn được thiết kế nhằm đảm bảo tính liên kết chặt chẽ giữa khâu tiền xử lý, mô hình hóa và phân tích kết quả, qua đó nâng cao độ tin cậy của mô hình dự đoán liên kết và khuyến nghị hợp tác nghiên cứu.



Hình 1. Quy trình tổng quát của dự án: Dự đoán hợp tác nghiên cứu tương lai và khuyến nghị đối tác.

Hình 1 thể hiện rõ 7 giai đoạn chính: (1) Thu thập và chuẩn bị dữ liệu, (2) Xây dựng mạng đồng tác giả, (3) Chia dữ liệu train/test, (4) Tính toán đặc trưng, (5) Học biểu diễn Node2Vec, (6) Đánh giá mô hình và (7) Trực quan và thảo luận,

Quy trình này đóng vai trò khung tham chiếu xuyên suốt cho toàn bộ quá trình thực nghiệm của đề tài.

A. Chuẩn bị và tiền xử lý dữ liệu

Dữ liệu đầu vào được thu thập từ mạng đồng tác giả của các công bố trong lĩnh vực Khoa học Dữ liệu. Mỗi cạnh trong đồ thị biểu diễn mối quan hệ hợp tác giữa hai tác giả, với trọng số w_{ij} thể hiện số lượng bài báo mà họ cùng công bố.

Các bước tiền xử lý bao gồm:

- Chuẩn hóa tên tác giả, loại bỏ trùng lặp và các bản ghi không đầy đủ.
- Chuyển trọng số w_{ij} về dạng số nguyên.
- Loại bỏ self-loop và kiểm tra thành phần liên thông.
- Giữ lại các thành phần liên thông lớn có kích thước $n \geq 100$ và mật độ $\rho \leq 0.5$ để đảm bảo dữ liệu đủ lớn nhưng không quá dày đặc.

Kết quả thu được là một đồ thị vô hướng $G = (V, E)$ với hơn 3.000 nút và 90.000 cạnh, biểu diễn cấu trúc hợp tác giữa các nhà nghiên cứu trong lĩnh vực Khoa học Dữ liệu.

B. Xây dựng mạng và chia tập huấn luyện kiểm tra

Tập dữ liệu được chia thành hai phần:

- **Tập huấn luyện (Train)**: Chiếm 80% tổng số cạnh, được sử dụng để tính toán các đặc trưng và học mô hình.
- **Tập kiểm tra (Test)**: Chiếm 20% còn lại, dùng để đánh giá hiệu năng dự đoán.

Các cạnh trong tập kiểm tra được loại bỏ khỏi mạng gốc để tạo đồ thị huấn luyện $G_{train} = (V, E_{train})$. Ngoài ra, các cặp nút không có liên kết được chọn ngẫu nhiên làm mẫu âm (negative samples) để đảm bảo dữ liệu cân bằng.

C. Tính toán các chỉ số dự đoán liên kết

Trên đồ thị huấn luyện G_{train}, năm thuật toán cổ điển được sử dụng để tính điểm tương đồng giữa hai nút u và v:

- **Common Neighbors (CN)**:

$$S_{CN}(u, v) = |N(u) \cap N(v)|$$

- **Jaccard Coefficient (JC)**:

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

- **Adamic–Adar Index (AA)**:

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

- **Preferential Attachment (PA)**:

$$S_{PA}(u,v) = |N(u)| \times |N(v)|$$

- **Resource Allocation (RA)**:

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

Các chỉ số này phản ánh đặc trưng cấu trúc của mạng: CN và JC mô tả mức độ chồng lấn của hàng xóm, trong khi AA và RA giảm trọng với các nút có bậc cao, còn PA phản ánh xu hướng "người nổi tiếng dễ hợp tác với nhau".

D. Học biểu diễn đồ thị bằng Node2Vec

Bên cạnh các chỉ số cục bộ, nghiên cứu sử dụng thuật toán **Node2Vec** để học biểu diễn vector cho từng nút trong mạng G_{train}. Node2Vec mở rộng ý tưởng của Word2Vec bằng việc mô phỏng bước đi ngẫu nhiên có trọng số để tối ưu hàm mục tiêu:

$$\max_f \sum_{u \in V} \log P(N_S(u) \mid f(u))$$

trong đó $f(u)$ là vector biểu diễn của nút u, và $N_S(u)$ là tập hàng xóm được sinh ra từ chuỗi bước đi ngẫu nhiên theo chiến lược *biased random walk*.

Độ tương đồng giữa hai nút được tính bằng **Cosine Similarity**:

$$S_{N2V}(u,v) = \frac{f(u) \cdot f(v)}{\|f(u)\| \|f(v)\|}$$

Phương pháp này giúp phát hiện các mối quan hệ tiềm ẩn vượt ra ngoài cộng đồng trực tiếp, tận dụng được thông tin toàn cục của mạng.

E. Đánh giá mô hình và trực quan hóa

Các phương pháp được đánh giá bằng hai chỉ số phổ biến:

$$AUC\text{-}ROC = \frac{1}{|P||N|} \sum_{p \in P} \sum_{n \in N} \delta(S(p) > S(n))$$

$$AUC\text{-}PR = \int_0^1 P(R)\, dR$$

trong đó P và N lần lượt là tập các cạnh thật và cạnh giả, δ là hàm chỉ báo. Ngoài ra, các chỉ số `Precision@K` và `Recall@K` được sử dụng để đánh giá hiệu quả khuyến nghị trong Top-K kết quả.

Các biểu đồ **ROC**, **Precision–Recall**, **Histogram** và **Boxplot** được sử dụng để trực quan hóa phân bố điểm tương đồng, giúp so sánh hiệu năng giữa các thuật toán. Toàn bộ quá trình được cài đặt bằng Python với các thư viện `NetworkX`, `scikit-learn`, `node2vec` và `matplotlib`.

IV. XÂY DỰNG VÀ CHUẨN BỊ DỮ LIỆU

A. Tập dữ liệu

Tập dữ liệu sử dụng trong nghiên cứu này được kế thừa trực tiếp từ kết quả của **Project 1A**, trong đó nhóm đã xây dựng thành công mạng đồng tác giả trong lĩnh vực *Khoa học Dữ liệu* dựa trên cơ sở dữ liệu học thuật mở OpenAlex. Từ mạng gốc gồm hơn 9.000 tác giả và 130.000 mối quan hệ hợp tác, dữ liệu được trích xuất và lưu dưới dạng danh sách cạnh (edge list) để phục vụ cho bài toán dự đoán liên kết trong nghiên cứu này.

Tập dữ liệu bao gồm ba cột chính:

- **Author1**: Tác giả thứ nhất trong cặp hợp tác.
- **Author2**: Tác giả thứ hai trong cặp hợp tác.
- **Weight**: Số lượng bài báo mà hai tác giả cùng công bố.

Mỗi dòng trong tệp đại diện cho một mối quan hệ hợp tác $(Author1, Author2)$, được biểu diễn dưới dạng cạnh có trọng số w_{ij} trong mạng đồng tác giả $G = (V, E)$, với V là tập các tác giả và E là tập các quan hệ hợp tác.

Tập dữ liệu sau khi làm sạch và chuẩn hóa được sử dụng để xây dựng mạng đồng tác giả và phục vụ các bước phân tích, huấn luyện mô hình dự đoán hợp tác nghiên cứu trong tương lai. Cột `Weight` đóng vai trò thể hiện cường độ quan hệ, giúp mô hình nhận biết mức độ hợp tác giữa các nhà khoa học trong quá khứ.

Bảng I
MÔ TẢ DỮ LIỆU QUAN HỆ HỢP TÁC GIỮA CÁC TÁC GIẢ

Author1	Author2	Weight
Joseph Sambrook	Elisabeth Fritsch	1
Joseph Sambrook	Tom Maniatis	1
Elisabeth Fritsch	Tom Maniatis	1
Keith E. Muller	Jacob Cohen	1
Jacob Cohen	Peter A. Lachenbruch	1
David Moher	A. Liberati	1
David Moher	Jennifer Tetzlaff	3
David Moher	Douglas G. Altman	3
David Moher	Matthew J. Page	2

B. Xây dựng mạng đồng tác giả

Dựa trên dữ liệu kế thừa từ Project 1A, nhóm tiến hành xây dựng **mạng đồng tác giả (co-author network)** nhằm phục vụ cho bài toán *dự đoán liên kết* giữa các nhà nghiên cứu. Dữ liệu được biểu diễn dưới dạng đồ thị vô hướng $G = (V, E)$, trong đó:

- Mỗi nút $v \in V$ đại diện cho một tác giả.
- Mỗi cạnh $e_{ij} \in E$ biểu diễn mối quan hệ hợp tác giữa hai tác giả i và j.
- Trọng số cạnh w_{ij} thể hiện số lượng bài báo mà hai tác giả cùng công bố.

Kết quả ban đầu cho thấy mạng bao gồm hơn 9.000 tác giả và 136.000 mối quan hệ hợp tác. Phân tích cấu trúc mạng cho thấy có tổng cộng **1.145 thành phần liên thông (connected components)**, trong đó **thành phần lớn nhất** gồm 2.616 nút và 76.666 cạnh, chiếm khoảng 28% tổng số nút. Điều này cho thấy mạng có tính *phân mảnh cao*, tức là tồn tại nhiều cụm tác giả nhỏ tách biệt.

Bảng II
CHỈ SỐ CỦA MẠNG ĐỒNG TÁC GIẢ TRƯỚC KHI LỌC

Số nút	Số cạnh	Số thành phần	Mật độ
9.212	136.198	1.145	0.0032

Để loại bỏ các cụm nhỏ lẻ và tập trung vào các cộng đồng nghiên cứu có ý nghĩa thống kê, nhóm áp dụng tiêu chí lọc như sau:

$$\text{Số nút} \geq 100, \quad \text{và} \quad \text{Mật độ} \leq 0.5.$$

Bảng III thể hiện thông tin chi tiết về 10 thành phần liên thông lớn nhất của mạng ban đầu.

Bảng III
TOP 10 THÀNH PHẦN LIÊN THÔNG LỚN NHẤT CỦA MẠNG ĐỒNG TÁC GIẢ

Thành phần	Số nút	Số cạnh	Mật độ
Component 1	2.616	76.666	0.0224
Component 2	266	7.362	0.2089
Component 3	211	8.129	0.3669
Component 4	124	503	0.0660
Component 5	117	5.103	0.7520
Component 6	105	4.959	0.9082
Component 7	99	4.851	1.0000
Component 8	98	2.909	0.6120
Component 9	88	2.954	0.7717
Component 10	84	1.084	0.3110

Dựa trên tiêu chí trên, chỉ các thành phần có quy mô lớn và mật độ hợp lý mới được giữ lại. Kết quả sau khi lọc được tóm tắt trong Bảng IV.

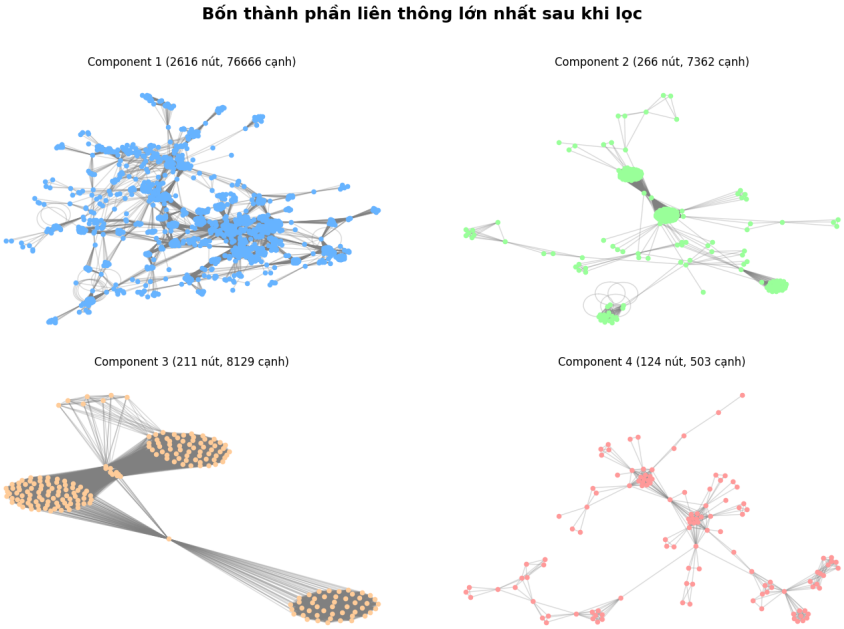
Bảng IV
CHỈ SỐ CỦA MẠNG ĐỒNG TÁC GIẢ SAU KHI LỌC

Số nút	Số cạnh	Số thành phần	Mật độ
3.217	92.660	4	0.018

Bốn thành phần lớn nhất được giữ lại có đặc trưng như trong Bảng V.

Bảng V
BỐN THÀNH PHẦN LỚN NHẤT ĐƯỢC GIỮ LẠI SAU KHI LỌC

Thành phần	Số nút	Số cạnh	Mật độ
Component 1	2.616	76.666	0.0224
Component 2	266	7.362	0.2089
Component 3	211	8.129	0.3669
Component 4	124	503	0.0660



Hình 2. Bốn thành phần liên thông lớn nhất sau khi lọc (Component 1–4). Mỗi màu thể hiện một cộng đồng hợp tác nghiên cứu riêng biệt.

Kết quả cho thấy **Component 1** là cụm nghiên cứu lớn nhất với hơn 2.600 tác giả và hơn 76.000 mối liên kết, có mật độ thấp (0.0224), phù hợp cho việc học mô hình liên kết. Các cụm còn lại (Component 2–4) tuy nhỏ hơn nhưng vẫn đảm bảo kích thước và mức độ kết nối hợp lý, phản ánh những nhóm nghiên cứu nhỏ, hoạt động tập trung trong các chuyên ngành riêng.

Cấu trúc mạng sau khi lọc giúp loại bỏ các cụm nhiễu và tập trung vào những cộng đồng hợp tác thực chất. Độ đậm đặc tăng từ 0.0032 lên 0.018 cho thấy mạng trở nên gắn kết hơn, giúp mô hình dự đoán liên kết hoạt động ổn định và hiệu quả hơn trong giai đoạn huấn luyện.

Ghi chú: Thành phần liên thông (connected component) là nhóm các nút trong mạng mà giữa bất kỳ hai nút nào cũng tồn tại ít nhất một đường đi, phản ánh cụm tác giả có mối liên hệ hợp tác trực tiếp hoặc gián tiếp.

C. Chuẩn bị dữ liệu cho bài toán dự đoán liên kết

Sau khi hoàn thiện bước xây dựng và lọc mạng đồng tác giả, nhóm tiến hành tách dữ liệu để phục vụ bài toán **dự đoán liên kết (Link Prediction)**. Mục tiêu là mô phỏng khả năng hình thành mối hợp tác mới giữa các tác giả trong tương lai, dựa trên cấu trúc mạng hiện có.

1) Chia tập huấn luyện và kiểm tra: Đầu tiên, toàn bộ các cạnh trong mạng được chia thành hai phần:

- **Tập huấn luyện (Train set)**: Chiếm 80% tổng số cạnh, được sử dụng để xây dựng mô hình và học đặc trưng của mạng.
- **Tập kiểm tra (Test set)**: Chiếm 20% số cạnh còn lại, được sử dụng để đánh giá khả năng dự đoán các liên kết mới.

Đồ thị huấn luyện G_{train} được xây dựng chỉ từ các cạnh thuộc tập train, nhưng vẫn giữ nguyên toàn bộ tập nút của đồ thị gốc. Việc này đảm bảo rằng mô hình có thể dự đoán liên kết giữa các tác giả chưa từng hợp tác, thay vì chỉ những cặp đã tồn tại trong tập huấn luyện.

Bên cạnh đó, nhóm cũng tạo ra tập **negative samples** bao gồm các cặp tác giả chưa từng có liên kết trong mạng. Các mẫu âm này được chọn ngẫu nhiên và có số lượng bằng với số mẫu dương (số cạnh thật) trong từng tập train và test, giúp cân bằng dữ liệu đầu vào khi đánh giá.

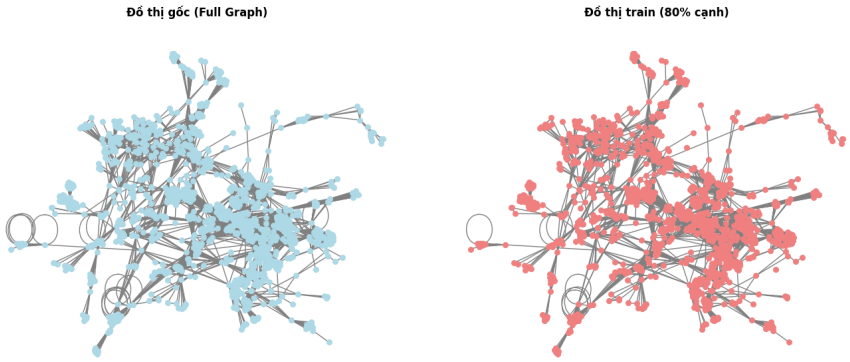
Kết quả chia dữ liệu được thể hiện trong Bảng VI.

Bảng VI
CHIA TẬP DỮ LIỆU CHO BÀI TOÁN DỰ ĐOÁN LIÊN KẾT

Tập dữ liệu	Số cạnh thật (Positive)	Số cạnh giả (Negative)
Huấn luyện (Train)	74.128	74.128
Kiểm tra (Test)	18.532	18.532

2) Đảm bảo tính kết nối của đồ thị huấn luyện: Sau khi chia tách, nhóm tiến hành kiểm tra tính liên thông của đồ thị G_{train}. Kết quả cho thấy đồ thị huấn luyện vẫn duy trì được tính liên thông trong các thành phần chính, đảm bảo rằng các thuật toán dự đoán (ví dụ Common Neighbors, Jaccard, Node2Vec) có thể hoạt động hiệu quả trên toàn mạng thay vì bị giới hạn trong các cụm rời rạc.

3) Trực quan hoá mạng huấn luyện và mạng gốc: Hình 3 mô tả trực quan sự khác biệt giữa đồ thị gốc (full graph) và đồ thị huấn luyện sau khi loại bỏ 20% số cạnh. Cấu trúc mạng huấn luyện vẫn bảo toàn được hình dạng và mật độ của các cụm chính, đảm bảo tính đại diện cho toàn bộ mạng.



Hình 3. So sánh đồ thị gốc và đồ thị huấn luyện, cho thấy các cụm chính vẫn được bảo toàn sau khi tách dữ liệu.

Việc tách 20% cạnh để kiểm tra giúp mô hình học được đặc trưng cấu trúc của mạng đồng tác giả mà không bị overfitting, đồng thời có đủ dữ liệu kiểm định để đánh giá khả năng dự đoán liên kết mới.

Tổng kết: Sau khi chia tách, mạng huấn luyện G_{train} vẫn giữ được cấu trúc cộng đồng và độ gắn kết tương đối cao. Việc tạo tập mẫu âm và dương cân bằng giúp đảm bảo tính khách quan trong quá trình huấn luyện và đánh giá các thuật toán dự đoán liên kết.

D. Các phương pháp dự đoán liên kết

Sau khi hoàn thiện tập dữ liệu huấn luyện, nghiên cứu tiến hành áp dụng các phương pháp dự đoán liên kết nhằm ước lượng khả năng hình thành các mối hợp tác mới giữa các tác giả trong mạng đồng tác giả. Các phương pháp được chia thành hai nhóm chính: (i) các chỉ số dựa trên cấu trúc cục bộ và (ii) phương pháp học biểu diễn đồ thị.

1) Nhóm phương pháp cấu trúc cục bộ: Nhóm này bao gồm năm thuật toán phổ biến: *Common Neighbors*, *Jaccard Coefficient*, *Adamic–Adar Index*, *Preferential Attachment* và *Resource Allocation*. Các chỉ số này được tính dựa trên thông tin lân cận của hai nút chưa có cạnh nối trong đồ thị huấn luyện. Chúng phản ánh mức độ tương đồng giữa hai tác giả thông qua số lượng hàng xóm chung, tần suất xuất hiện của các nút trung gian hoặc mức độ kết nối của mỗi nút. Các phương pháp này hoạt động hiệu quả trong mạng có cấu trúc cộng đồng rõ ràng, đặc biệt khi các cụm nghiên cứu có xu hướng hợp tác nội bộ mạnh mẽ.

2) Nhóm phương pháp học biểu diễn đồ thị: Bên cạnh các thuật toán cổ điển, nghiên cứu còn triển khai mô hình **Node2Vec** nhằm học biểu diễn vector cho từng tác giả trong không gian đặc trưng liên tục. Phương pháp này dựa trên cơ chế *random walk* để trích xuất ngữ cảnh xung quanh mỗi nút, từ đó huấn luyện mô hình nhúng học được cấu trúc toàn cục của mạng. Sau khi huấn luyện, độ tương đồng giữa hai nút được đo bằng *cosine similarity* giữa các vector nhúng tương ứng. Node2Vec có ưu thế trong việc phát hiện các mối liên kết tiềm ẩn giữa các cộng đồng xa nhau mà các phương pháp cục bộ thường bỏ sót.

3) Tổng kết: Tổng cộng, sáu phương pháp được sử dụng trong nghiên cứu bao gồm: **Common Neighbors**, **Jaccard Coefficient**, **Adamic–Adar Index**, **Preferential Attachment**, **Resource Allocation Index** và **Node2Vec (Cosine Similarity)**. Những thuật toán này đóng vai trò nền tảng cho giai đoạn thực nghiệm và đánh giá hiệu năng, được trình bày chi tiết trong Mục V.

V. THỰC NGHIỆM VÀ ĐÁNH GIÁ

A. Cấu hình và môi trường thực nghiệm

Các thí nghiệm được thực hiện bằng ngôn ngữ lập trình **Python 3.10** cùng các thư viện **NetworkX**, **scikit-learn**, **node2vec**, **NumPy** và **Matplotlib**. Máy tính chạy thử nghiệm sử dụng CPU Intel Core i7 và RAM 16GB. Quá trình đánh giá được tiến hành trên mạng đồng tác giả đã qua lọc (3.217 nút, 92.660 cạnh, mật độ 0.018).

B. Chỉ số đánh giá

Để đánh giá hiệu quả dự đoán liên kết, hai chỉ số chính được sử dụng:

- **AUC-ROC (Area Under ROC Curve):** Đo lường khả năng mô hình phân biệt giữa cạnh thật (positive) và cạnh giả (negative).
- **AUC-PR (Area Under Precision–Recall Curve):** Phản ánh độ chính xác và khả năng bao phủ, đặc biệt phù hợp với mạng có mật độ thấp.

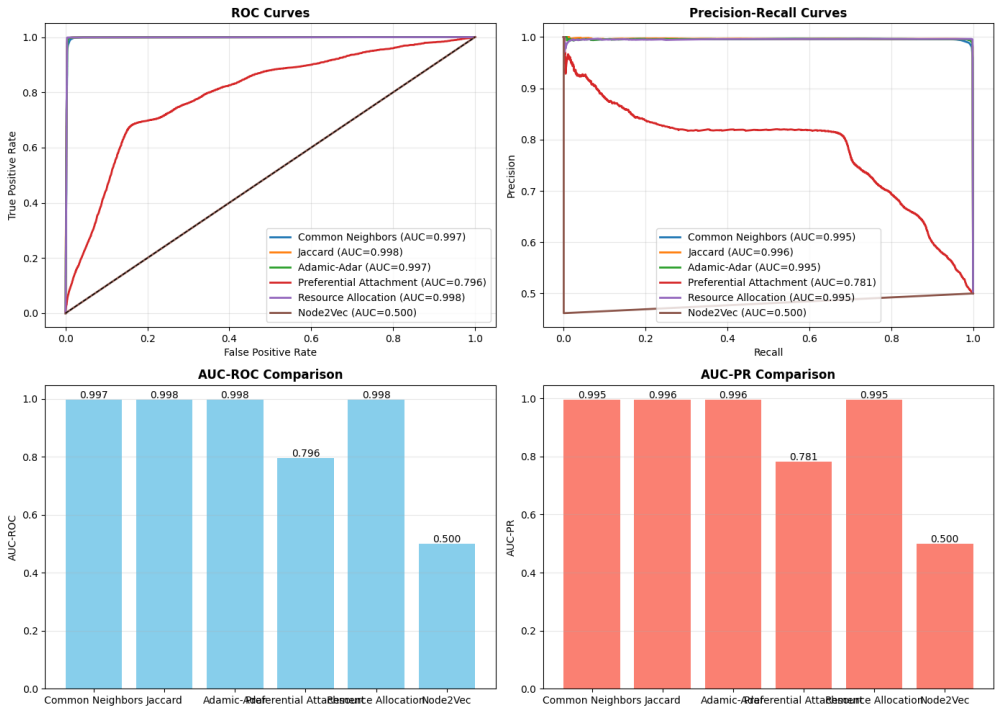
C. Kết quả đánh giá

Bảng VII
KẾT QUẢ ĐÁNH GIÁ CÁC PHƯƠNG PHÁP DỰ ĐOÁN LIÊN KẾT

Phương pháp	AUC-ROC	AUC-PR
Jaccard Coefficient	0.9977	0.9961
Adamic–Adar Index	0.9975	0.9955
Resource Allocation	0.9975	0.9953
Common Neighbors	0.9973	0.9954
Preferential Attachment	0.7963	0.7814
Node2Vec (Cosine Similarity)	0.5000	0.5002

Kết quả cho thấy các phương pháp dựa trên cấu trúc cục bộ như *Jaccard*, *Adamic–Adar*, *Resource Allocation* và *Common Neighbors* đạt hiệu suất rất cao với AUC-ROC và AUC-PR xấp xỉ 1.0, thể hiện khả năng phân biệt gần như hoàn hảo giữa cạnh thật và cạnh giả. Ngược lại, *Preferential Attachment* có hiệu năng thấp hơn (AUC-ROC ≈ 0.80), cho thấy giả định "các nút có nhiều liên kết sẽ dễ hợp tác hơn" không hoàn toàn đúng trong mạng học thuật. Đặc biệt, *Node2Vec* cho kết quả gần ngẫu nhiên (AUC ≈ 0.50), nguyên nhân có thể do embedding chưa hội tụ hoặc số lượng bước ngẫu nhiên chưa đủ để nắm bắt cấu trúc mạng toàn cục.

D. Trực quan hóa kết quả



Hình 4. So sánh hiệu suất các phương pháp dự đoán liên kết.

Kết quả trực quan cho thấy các phương pháp dựa trên đặc trưng cục bộ như **Common Neighbors**, **Jaccard Coefficient**, **Adamic–Adar** và **Resource Allocation** đều đạt hiệu suất rất cao, với chỉ số **AUC-ROC** và **AUC-PR** xấp xỉ 1.0. Các đường cong ROC của nhóm này gần như ôm sát trục trái và phía trên, thể hiện khả năng phân biệt rất tốt giữa các cặp nút có và không có liên kết thực tế.

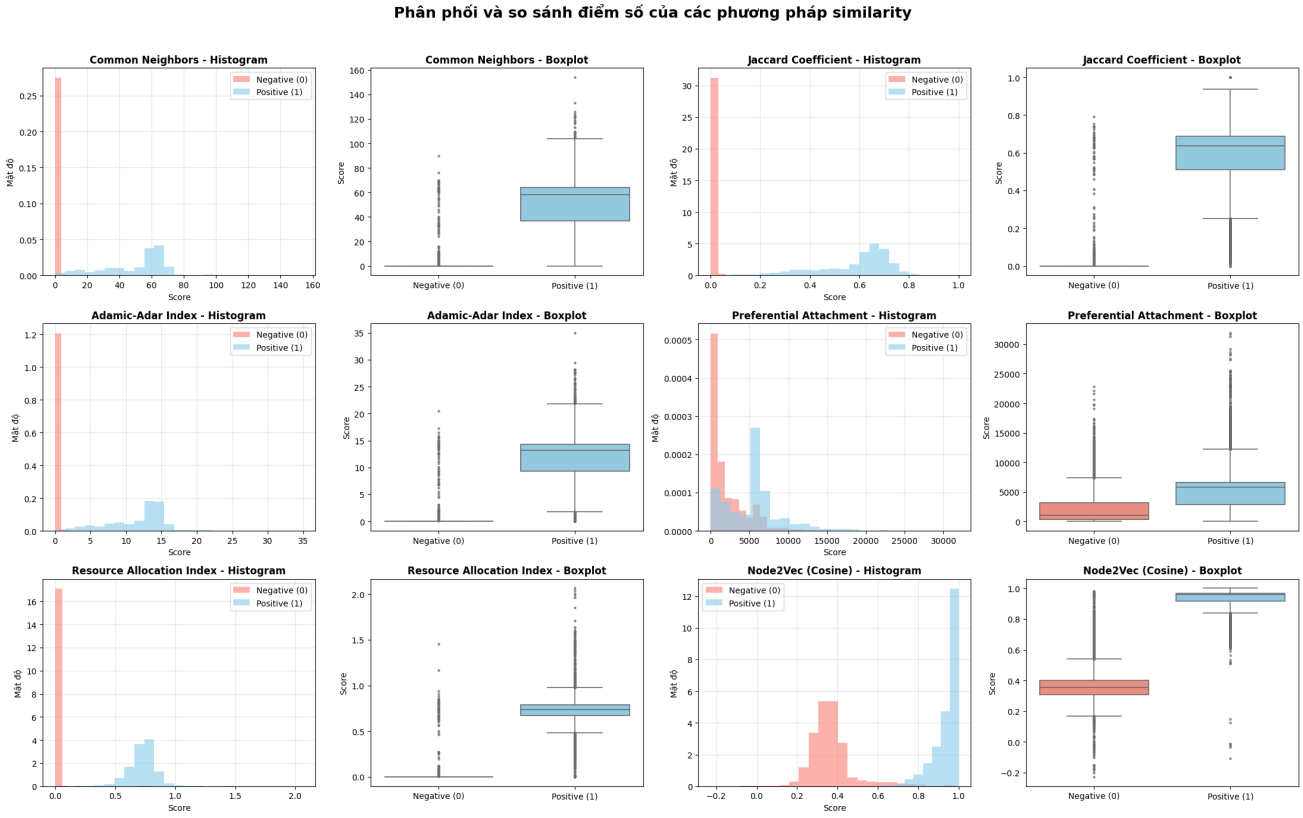
Ngược lại, **Preferential Attachment** cho kết quả thấp hơn đáng kể (AUC-ROC ≈ 0.79; AUC-PR ≈ 0.78), cho thấy giả định rằng "nút có nhiều kết nối sẽ dễ hình thành liên kết mới" không còn chính xác trong mạng đồng tác giả, nơi các mối quan hệ hợp tác thường mang tính chọn lọc theo chuyên môn thay vì ngẫu nhiên dựa vào độ phổ biến.

Đáng chú ý, **Node2Vec** cho hiệu suất rất thấp (AUC ≈ 0.5), gần tương đương với dự đoán ngẫu nhiên. Nguyên nhân có thể do mô hình nhúng chưa được huấn luyện đủ lâu hoặc chưa tối ưu các siêu tham số (như chiều nhúng, bước đi ngẫu nhiên hoặc số lượng walk). Ngoài ra, do đồ thị có nhiều thành phần rời rạc nên Node2Vec khó học được quan hệ toàn cục giữa các cụm.

Tổng thể, nhóm phương pháp dự đoán liên kết truyền thống vẫn thể hiện khả năng vượt trội trong mạng đồng tác giả, đặc biệt khi cấu trúc cộng đồng rõ ràng và mối quan hệ giữa các tác giả mang tính lân cận cao. Trong khi đó, Node2Vec có tiềm năng mở rộng nhưng cần được tối ưu thêm để nắm bắt tốt hơn đặc trưng toàn cục của mạng.

E. Phân tích phân phối điểm similarity

Để đánh giá sâu hơn hành vi của các thuật toán dự đoán liên kết, nhóm nghiên cứu tiến hành trực quan hóa phân phối điểm *similarity* của sáu phương pháp: Common Neighbors, Jaccard Coefficient, Adamic–Adar Index, Preferential Attachment, Resource Allocation và Node2Vec. Mỗi phương pháp được biểu diễn dưới hai dạng: *Histogram* thể hiện mật độ điểm, và *Boxplot* thể hiện phân bố giá trị theo hai nhóm cạnh thật (Positive) và cạnh giả (Negative).



Hình 5. Phân phối và so sánh điểm similarity của các phương pháp dự đoán liên kết.

Quan sát Hình 5 trên cho thấy, các phương pháp dựa trên đặc trưng cục bộ như **Common Neighbors**, **Jaccard**, **Adamic–Adar** và **Resource Allocation** thể hiện sự phân tách rõ ràng giữa hai lớp. Các cạnh thật có điểm similarity cao hơn đáng kể, cho thấy khả năng dự đoán tốt trong việc phát hiện các mối hợp tác tiềm năng.

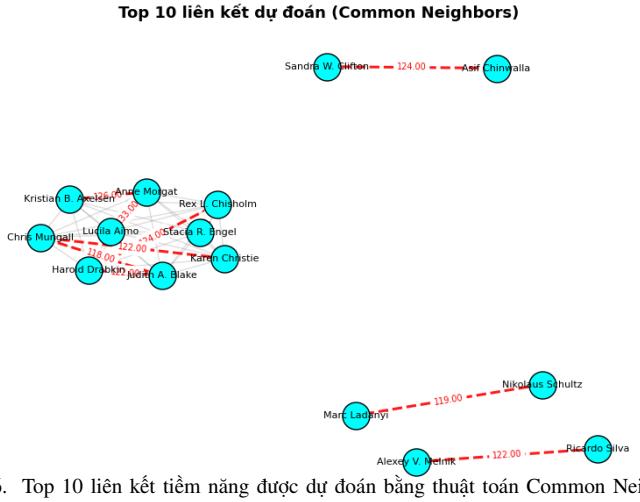
Ngược lại, phương pháp **Preferential Attachment** có phân phối điểm rất rộng và mức độ chồng lấn cao giữa hai lớp, do đặc trưng phụ thuộc mạnh vào bậc của nút. Các tác giả có nhiều cộng tác viên thường được gán điểm cao cho hầu hết các cặp, dẫn đến khả năng dự đoán kém ổn định. Trong khi đó, **Node2Vec** cho thấy xu hướng phân tách tương đối giữa hai nhóm nhưng vẫn chưa thật sự rõ ràng; mô hình mới chỉ khai thác được một phần đặc trưng toàn cục của mạng và có thể cải thiện thêm thông qua tối ưu các siêu tham số (*walk length, num walks, p, q*).

Tổng thể, các thuật toán *heuristic* (dựa trên cấu trúc lân cận) vẫn thể hiện hiệu năng cao và ổn định trong mạng đồng tác giả, trong khi Node2Vec thể hiện tiềm năng mở rộng nhưng cần được tinh chỉnh thêm để đạt hiệu quả tương đương hoặc vượt trội hơn.

F. Phân tích các liên kết được dự đoán

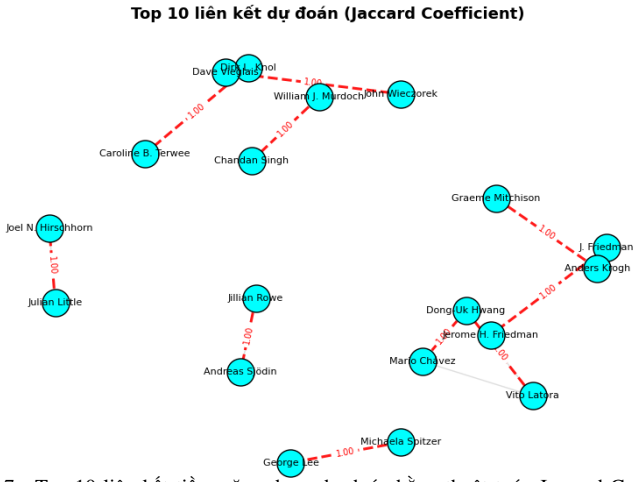
Để trực quan hóa khả năng dự đoán của từng thuật toán, nhóm nghiên cứu lựa chọn và minh họa *Top–10* liên kết có xác suất hình thành cao nhất đối với sáu phương pháp: **Common Neighbors**, **Jaccard Coefficient**, **Adamic–Adar Index**, **Preferential Attachment**, **Resource Allocation** và **Node2Vec**.

Trong mỗi hình, các **cạnh gốc** (quan hệ hợp tác hiện có) được hiển thị bằng màu xám, trong khi các **cạnh dự đoán mới** được tô màu đỏ và kẻ đứt nét để dễ dàng phân biệt. Kích thước các nút được giữ cố định nhằm đảm bảo tính đồng nhất trong so sánh giữa các mô hình.



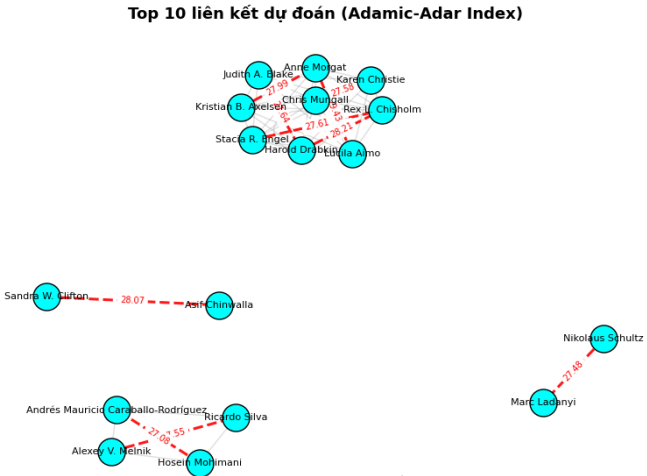
Hình 6. Top 10 liên kết tiềm năng được dự đoán bằng thuật toán Common Neighbors.

Common Neighbors. Các liên kết mới có điểm similarity cao nhất dao động từ **118 đến 126**, tập trung trong cùng cụm tác giả có nhiều hàng xóm chung. Điều này cho thấy thuật toán ưu tiên các *quan hệ cục bộ* với độ kết nối dày đặc, phản ánh xu hướng cộng tác quen thuộc giữa các nhà nghiên cứu trong cùng nhóm.



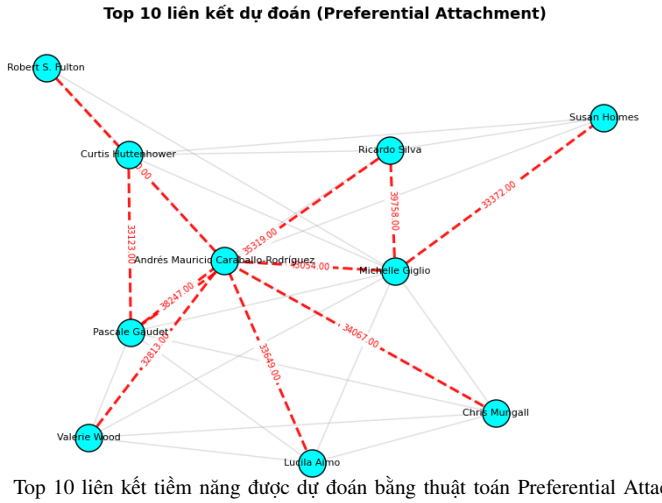
Hình 7. Top 10 liên kết tiềm năng được dự đoán bằng thuật toán Jaccard Coefficient.

Jaccard Coefficient. Các cạnh có giá trị **Jaccard = 1.0**, nghĩa là các cặp tác giả chia sẻ toàn bộ hàng xóm, thể hiện sự *tương đồng hoàn toàn* trong cấu trúc cộng tác. Thuật toán hoạt động tốt trong việc phát hiện các nhóm nghiên cứu chặt chẽ, nhưng khó mở rộng ra các cụm xa hơn.



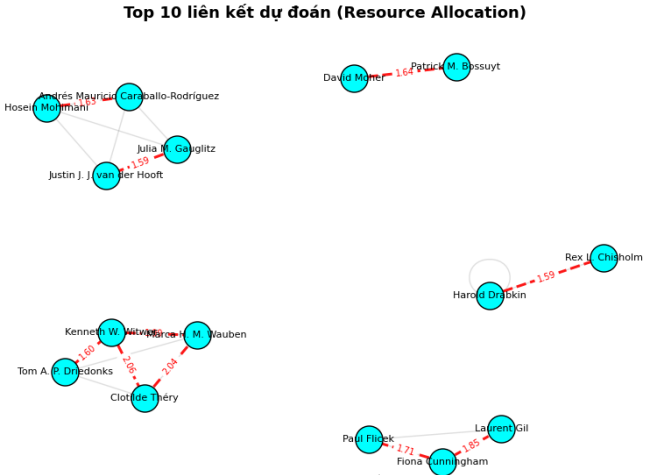
Hình 8. Top 10 liên kết tiềm năng được dự đoán bằng thuật toán Adamic–Adar Index.

Adamic–Adar Index. Điểm số dự đoán dao động quanh **27–28**, cao hơn rõ rệt ở các cặp nút có hàng xóm hiếm. Phương pháp này giúp phát hiện các *mối quan hệ tiềm năng* giữa những tác giả có vai trò cầu nối trong mạng, góp phần tăng khả năng mở rộng hợp tác.



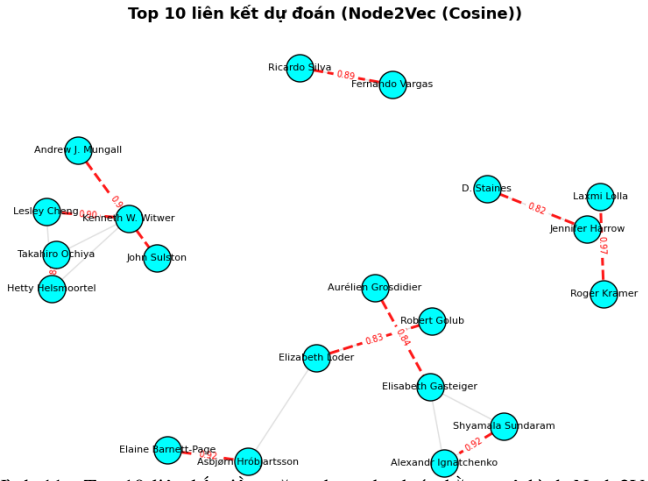
Hình 9. Top 10 liên kết tiềm năng được dự đoán bằng thuật toán Preferential Attachment.

Preferential Attachment. Các nút có bậc cao như *Caraballo-Rodríguez* đạt điểm **trên 33.000**, chi phối phần lớn các liên kết mới. Điều này thể hiện rõ giả thuyết "*người nổi bật thường có thêm nhiều hợp tác mới*", song cũng cho thấy mô hình có thể bị thiên lệch về các nút trung tâm.



Hình 10. Top 10 liên kết tiềm năng được dự đoán bằng thuật toán Resource Allocation Index.

Resource Allocation Index. Các liên kết được dự đoán có điểm trong khoảng **1.5–1.7**, tập trung ở trung tâm cụm. Thuật toán phản ánh tốt cơ chế *lan truyền tài nguyên tri thức* giữa các thành viên trong cùng cộng đồng hợp tác, giúp duy trì tính ổn định của mạng.



Hình 11. Top 10 liên kết tiềm năng được dự đoán bằng mô hình Node2Vec.

Node2Vec (Cosine Similarity). Các liên kết mới có độ tương đồng cosine từ **0.82–0.92**, phân bố rộng hơn so với các phương pháp heuristic, không chỉ trong cùng cụm cộng đồng. Điều này chứng tỏ khả năng của Node2Vec trong việc *khai thác đặc trưng toàn cục của mạng*, tuy nhiên vẫn cần tối ưu thêm tham số để tăng độ chính xác.

Quan sát các hình trên cho thấy, các phương pháp **Common Neighbors**, **Jaccard**, **Adamic–Adar** và **Resource Allocation** có xu hướng dự đoán các liên kết nằm trong cùng một cụm cộng tác. Các cạnh đỏ chủ yếu kết nối giữa các tác giả vốn đã có nhiều hàng xóm chung, phản ánh đặc trưng *local structural* rõ nét của nhóm thuật toán heuristic.

Trong khi đó, **Preferential Attachment** tập trung dự đoán quanh một vài nút có bậc cao, thể hiện giả thuyết "tác giả nổi bật thường dễ có thêm hợp tác mới," nhưng đồng thời dẫn đến hiện tượng thiên lệch về các nút trung tâm.

Khác biệt rõ rệt nhất nằm ở **Node2Vec**, khi các liên kết được dự đoán phân tán hơn, bao gồm cả các kết nối giữa những cụm tác giả vốn tách biệt. Điều này cho thấy mô hình học biểu diễn có khả năng khai thác ngữ cảnh toàn cục của mạng, mặc dù một số dự đoán chưa thật sự chính xác do hạn chế về tham số embedding.

Tổng thể, nhóm phương pháp heuristic thể hiện khả năng dự đoán chính xác trong phạm vi cộng đồng hiện hữu, trong khi Node2Vec mở ra hướng tiếp cận mới trong việc phát hiện các mối quan hệ hợp tác tiềm năng giữa các nhóm nghiên cứu khác nhau.

VI. Thảo luận

A. Đánh giá tổng quan hiệu quả các thuật toán

Kết quả thực nghiệm cho thấy các phương pháp dự đoán liên kết dựa trên đặc trưng cục bộ như **Common Neighbors, Jaccard, Adamic–Adar** và **Resource Allocation** đạt hiệu suất vượt trội với giá trị AUC-ROC và AUC-PR xấp xỉ 1.0. Điều

này chứng tỏ cấu trúc mạng đồng tác giả trong tập dữ liệu có tính *cộng đồng mạnh* các tác giả thường hợp tác trong cùng nhóm, khiến các thuật toán dựa trên hàng xóm chung hoạt động rất hiệu quả.

Ngược lại, **Preferential Attachment** có hiệu quả thấp hơn rõ rệt (AUC-ROC 0.80), do giả định "nút có nhiều kết nối sẽ dễ hình thành liên kết mới" không hoàn toàn đúng trong mạng nghiên cứu, nơi quan hệ hợp tác thường phụ thuộc vào lĩnh vực, dự án hoặc tổ chức hơn là mức độ phổ biến của tác giả.

Đáng chú ý, **Node2Vec** cho kết quả khiêm tốn (AUC-ROC 0.50), gần tương đương với mô hình dự đoán ngẫu nhiên. Nguyên nhân có thể đến từ việc số lượng bước đi ngẫu nhiên và tham số *walk_length, p, q* chưa được tối ưu, khiến biểu diễn vector chưa phản ánh tốt ngữ cảnh toàn cục của mạng.

B. Phân tích nguyên nhân và đặc điểm cấu trúc mạng

Đồ thị đồng tác giả có tính phân mảnh cao với nhiều thành phần liên thông nhỏ. Sau khi lọc theo điều kiện $|V| > 100$ và mật độ < 0.5, chỉ còn bốn thành phần lớn được giữ lại. Cấu trúc này tạo điều kiện thuận lợi cho các phương pháp heuristic, vốn dựa vào đặc trưng lân cận, nhưng lại gây khó khăn cho các mô hình học biểu diễn như Node2Vec khi cần nắm bắt đặc trưng toàn mạng. Ngoài ra, việc các nút cô lập (ít cạnh) chiếm tỷ lệ lớn cũng làm giảm độ chính xác của các phương pháp dựa trên embedding.

C. Hạn chế và hướng cải tiến

Mặc dù các phương pháp heuristic đạt kết quả tốt, chúng bị giới hạn trong việc phát hiện các liên kết mới giữa các cộng đồng khác nhau. Trong tương lai, nhóm nghiên cứu có thể mở rộng theo các hướng sau:

- **Tối ưu tham số Node2Vec** hoặc áp dụng các mô hình nhúng đồ thị hiện đại hơn như *DeepWalk, GraphSAGE, hoặc LINE* để cải thiện khả năng học đặc trưng toàn cục.
- **Kết hợp nhiều đặc trưng** (structural + semantic) để mô hình vừa tận dụng cấu trúc mạng vừa khai thác thông tin văn bản như tiêu đề hoặc chủ đề bài báo.
- **Thực hiện dự đoán động (temporal link prediction)** nhằm phân tích sự hình thành liên kết theo thời gian, phản ánh xu hướng hợp tác thực tế giữa các tác giả.

D. Tổng kết chương

Các phương pháp heuristic vẫn thể hiện ưu thế trong bài toán dự đoán liên kết khi mạng có cấu trúc cộng đồng mạnh. Nhưng để cải thiện khả năng khái quát và phát hiện liên kết mới giữa các nhóm nghiên cứu khác nhau, việc ứng dụng các mô hình nhúng học sâu như Node2Vec hoặc Graph Neural Network là hướng đi đầy tiềm năng cho các nghiên cứu tiếp theo.

VII. KẾT LUẬN

Đề tài đã triển khai và so sánh hiệu quả của các phương pháp dự đoán liên kết trên mạng đồng tác giả, bao gồm nhóm thuật toán heuristic (*Common Neighbors, Jaccard, Adamic–Adar, Resource Allocation, Preferential Attachment*) và mô hình học biểu diễn *Node2Vec*. Kết quả thực nghiệm cho thấy các phương pháp heuristic đạt độ chính xác cao và ổn định, đặc biệt trong các cụm tác giả có quan hệ cộng tác mật thiết. Trong khi đó, Node2Vec tuy chưa đạt hiệu quả nổi trội nhưng thể hiện tiềm năng trong việc khám phá những liên kết mới giữa các nhóm nghiên cứu khác biệt, phản ánh góc nhìn toàn cục của đồ thị.

Nghiên cứu cũng chỉ ra rằng đặc trưng cấu trúc của mạng như tính cộng đồng mạnh, mật độ cạnh cao và sự phân mảnh của đồ thị có ảnh hưởng đáng kể đến hiệu năng của từng mô hình. Do đó, việc lựa chọn thuật toán phù hợp cần cân nhắc giữa độ phức tạp tính toán và khả năng nắm bắt ngữ cảnh của mạng.

Trong tương lai, hướng phát triển khả thi là kết hợp giữa đặc trưng cấu trúc và ngữ nghĩa, chẳng hạn liên kết thông tin nội dung bài báo (từ khoá, lĩnh vực, tóm tắt) với đặc trưng đồ thị, để cải thiện khả năng khái quát của mô hình. Ngoài ra, việc áp dụng các kỹ thuật học sâu trên đồ thị như *GraphSAGE, GAT* hay *Graph Neural Network* hứa hẹn mang lại kết quả toàn diện hơn trong bài toán dự đoán và gợi ý hợp tác khoa học.

Tổng thể, đề tài không chỉ giúp làm rõ cơ chế hình thành mối quan hệ trong mạng nghiên cứu mà còn mở ra định hướng ứng dụng các mô hình học máy tiên tiến trong việc hỗ trợ phân tích và dự báo xu hướng hợp tác trong lĩnh vực khoa học dữ liệu.

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] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical mechanics and its applications*, vol. 311, no. 3-4, pp. 590–614, 2002.

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

[4] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," in *Proceedings of the twelfth international conference on Information and knowledge management*. ACM, 2003, pp. 556–559.

[5] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.
[6] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 855–864.
[7] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," *Machine Learning*, vol. 83, no. 2, pp. 163–191, 2011.
[8] 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.
[9] M. Sun, H. Wang, and Z. Liu, "Co-author recommendation using link prediction on co-authorship networks," in *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*. IEEE, 2016, pp. 1–8.