BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG THÀNH PHỐ HỒ CHÍ MINH

------- 0O0 ------

BÁO CÁO TỔNG KẾT

ĐỀ TÀI THAM GIA CUỘC THI "SINH VIÊN NGHIÊN CỨU KHOA HỌC"
NĂM 2024

NGHIÊN CỨU THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG TRONG DỮ LIỆU MẠNG XÃ HỘI – ỨNG DỤNG XÂY DỰNG DỮ LIỆU DẠNG ĐỒ THỊ LƯỠNG CỰC

Thuộc lĩnh vực: Công nghệ thông tin
(thuộc lĩnh vực Khoa học Kỹ thuật)

Năm 2024

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG THÀNH PHỐ HỒ CHÍ MINH

—--- -0O0 ---- --

BÁO CÁO TỔNG KẾT

ĐỀ TÀI THAM GIA CUỘC THI "SINH VIÊN NGHIÊN CỨU KHOA HỌC"
TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG THÀNH PHỐ HỒ CHÍ MINH
NĂM 2024

NGHIÊN CỨU THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG TRONG DỮ LIỆU MẠNG XÃ HỘI – ỨNG DỤNG XÂY DỰNG DỮ LIỆU LƯỠNG CỰC

Sinh viên thực hiện: **Đỗ Thế Sang** Nam, Nữ: Nam
Dân tộc: Kinh
Lớp, khoa: 11DHTH10 – Công nghệ Thông tin Năm thứ: 4 /số năm đào tạo: 4
Ngành học: Khoa học Phân tích Dữ liệu

Sinh viên thực hiện: **Nguyễn Trường Phát** Nam, Nữ: Nam
Dân tộc: Kinh
Lớp, khoa: 11DHTH10 – Công nghệ Thông tin Năm thứ: 4 /số năm đào tạo: 4
Ngành học: Khoa học Phân tích Dữ liệu

Người hướng dẫn chính: **TS. Nguyễn Thị Bích Ngân**

Năm 2024

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Đồ án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Đồ án này đã được cảm ơn và các thông tin trích dẫn trong Đồ án đã được chỉ rõ nguồn gốc.

TP.HCM, ngày…tháng…năm…..

Sinh viên thực hiện Đồ án

Sinh viên 1	**Sinh viên 2**
(Ký và ghi rõ họ tên)	(Ký và ghi rõ họ tên)
Đỗ Thế Sang	Nguyễn Trường Phát

LỜI CẢM ƠN

Trước tiên, chúng tôi xin bày tỏ lòng biết ơn chân thành và sâu sắc nhất tới giáo viên hướng dẫn cô Nguyễn Thị Bích Ngân đã tận tình hướng dẫn, động viên và giúp đỡ chúng tôi trong suốt quá trình thực hiện đề tài nghiên cứu này.

Chúng tôi xin bày tỏ lời cảm ơn sâu sắc đến các thầy cô giáo đã giảng dạy chúng tôi trong suốt bốn năm học qua, đã cho chúng tôi những kiến thức quý báu để chúng tôi có thể vững bước trên con đường của chính mình.

Và lời cuối cùng, chúng tôi xin gửi lời cảm ơn đến trường Đại học Công Thương TP.HCM và đặc biệt là các thầy cô giáo khoa Công Nghệ Thông Tin đã luôn cung cấp các kiến thức, cũng như là tạo điều kiện và môi trường học tập tốt nhất cho sinh viên để chúng tôi có được như ngày hôm nay.

Vì kiến thức bản thân còn hạn chế, trong quá trình thực hiện đề tài nghiên cứu không thể tránh khỏi những sai sót, rất mong nhận được lời góp ý và nhận xét từ các thầy cô.

TP.HCM, ngày…tháng…năm…..

Sinh viên thực hiện Đồ án

Sinh viên 1	**Sinh viên 2**
(Ký và ghi rõ họ tên)	(Ký và ghi rõ họ tên)
Đỗ Thế Sang	Nguyễn Trường Phát

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Tiếng Anh	Tiếng Việt
E	Edge	Cạnh
G	Graph	Đồ thị
GIM	Group Influence Maximizationo	Tối đa hóa ảnh hưởng trên nhóm
IM	Influence Maximization	Tối đa hóa ảnh hưởng
MXH	Social Network	Mạng xã hội
SEO	Search Engine Optimization	Tập hợp các chiến lược và kỹ thuật được sử dụng để cải thiện và tối ưu hóa hiển thị của một trang web trên các công cụ tìm kiếm
V	Vertex	Đỉnh/Nút

DANH MỤC CÁC BIỂU ĐỒ, ĐỒ THỊ, SƠ ĐỒ

DANH MỤC CÁC HÌNH ẢNH

MỞ ĐẦU

Những thập kỷ gần đây đã chứng kiến sự bùng nổ của môi trường trực tuyến, nơi hàng trăm triệu người tương tác với nhau và tạo ra lượng thông tin chưa từng có. Mạng xã hội (MXH) đóng vài trò quan trọng trong nhiều hiện tượng của xã hội, chẳng hạn: một thông tin có thể nhanh chóng lan truyền nhanh chóng thông qua quá trình chia sẻ giữa bạn bè; một ví dụ gần đây là chiến dịch tổng thống của Donald Trump vào năm 2016, trong đó Twitter gần như hàng ngày được sử dụng như một công cụ chiến dịch [1]; hoặc thông qua MXH, với sự lan truyền thông tin nhanh chóng và tác động mạnh mẽ, tiếp thị lan truyền trên MXH là một trong những chiến lược quan trọng mà các nhà quảng cáo muốn quảng bá sản phẩm hoặc dịch vụ của họ luôn quan tâm hàng đầu [2]. Sự phổ biến và vai trò quan trọng của mạng xã hội đã thu hút nhiều sự chú ý của giới nghiên cứu các bài toán liên quan việc lan truyền thông tin.

Trong số những bài toán liên quan lĩnh vực phân tích MXH, nổi bật và thu hút nhiều nhóm nghiên cứu đó là tối thiểu hóa chi phí quảng cáo nhưng vẫn đạt hiệu quả tiếp cận nhiều khách hàng tiềm năng nhất trên MXH. Trong bài toán này, cốt lõi của là tìm tập những người có tầm ảnh hưởng rộng, có khả năng lan truyền thông tin, tác động ảnh hưởng đến nhiều người nhất, và những người bị ảnh hưởng phải thuộc cộng đồng các khách hàng tiềm năng của sản phẩm hoặc dịch vụ cần quảng bá. Để giải quyết vấn đề, người ta sẽ tìm những cộng đồng chứa khách hàng tiềm năng, "gieo tin" cho những người có tầm ảnh hưởng trong các cộng đồng đó để tối thiểu chi phí và tối đa hiệu quả của quảng cáo [3].

Để giải bài toán này chúng ta cần biểu diễn dữ liệu và cộng đồng của MXH dưới dạng đồ thị lưỡng cực (bipartite graph[1]). Tuy nhiên, hiện nay, qua quá trình khảo sát, dữ liệu của MXH thường chỉ có dạng đồ thị thông thường (graph[2]). Xuất phát từ nhu cầu trên, trong đề tài này chúng tôi nghiên cứu các thuật toán phát hiện cộng đồng trong MXH, từ đó xây dựng đồ thị dạng lưỡng cực cho MXH.

---

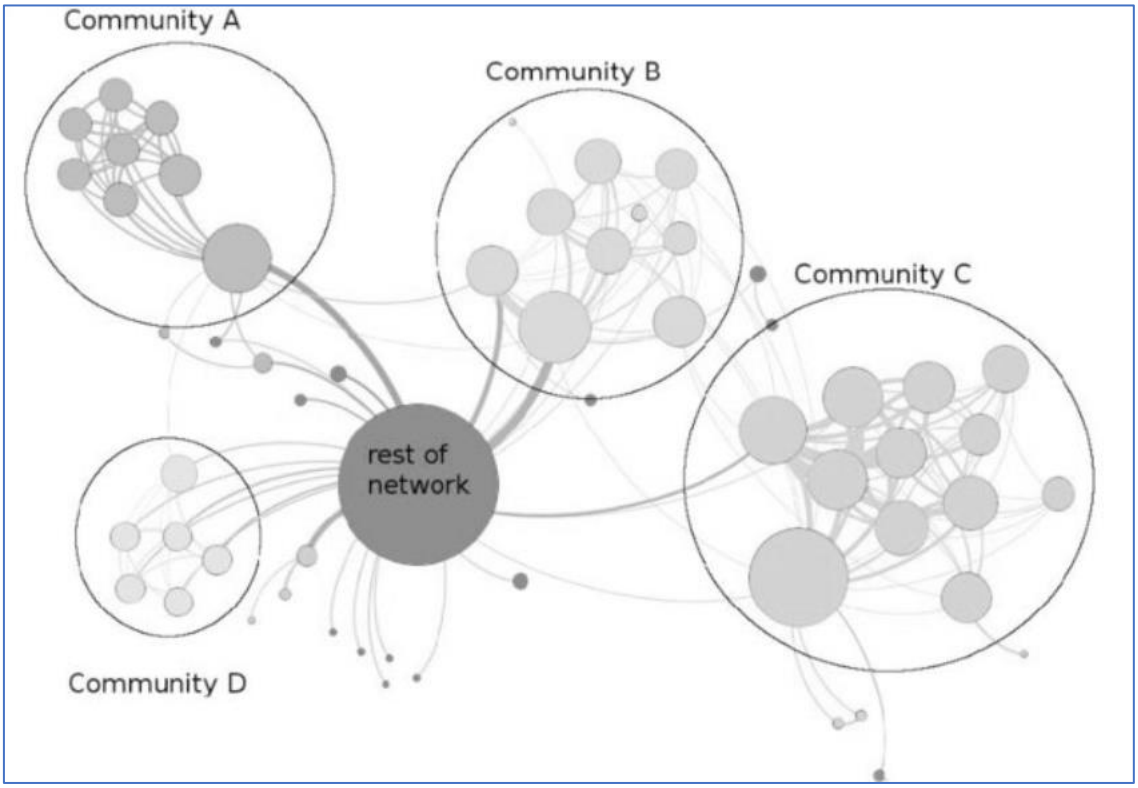
[1] https://mathworld.wolfram.com/BipartiteGraph.html

[2] https://mathworld.wolfram.com/Graph.html

TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU THUỘC ĐỀ TÀI

1. Khái niệm cộng đồng trong MXH

Mạng xã hội có thể được hiển thị dưới dạng đồ thị. Trong đó người dùng được biểu thị bằng các nút/đỉnh và kết nối (tình bạn, lượt theo dõi) được biểu thị bằng các cạnh. Cộng đồng (community) là các nhóm đỉnh có mật độ kết nối cao trong nhóm, nghĩa là các thành viên được kết nối tốt với nhau. Ngược lại, mối liên hệ giữa các nhóm này (các cộng đồng khác nhau) lại thưa thớt hơn [4]. Hình 1 minh họa các cộng đồng trong dữ liệu mạng xã hội dạng dưới dạng đồ thị.



Hình 1. Minh họa các cộng đồng trong dữ liệu MXH dưới dạng đồ thị[3]

2. Tổng quan các thuật toán phát hiện cộng đồng

Trong phần này, chúng tôi giới thiệu tổng quan về các nhóm thuật toán phát hiện cộng đồng phổ biến hiện nay. Qua quá trình khảo sát các nghiên cứu liên quan, nhìn chung có một số kỹ thuật như: xác định cộng đồng tĩnh, động, không giao nhau và chồng chéo [4].

---

[3] https://www.researchgate.net/publication/332207478_The_Distance_Between_Us/figures?lo=1

2.1. Các kỹ thuật phát hiện cộng đồng truyền thống

2.1.1. Phân chia đồ thị

Phân chia đồ thị thành g cụm có kích thước được xác định trước, sao cho số liên kết trong một cụm có mật độ hơn số cạnh giữa các cụm [5]. Các ví dụ nổi tiếng về kỹ thuật phân chia đồ thị là phương pháp phân chia phổ Laplace [6] và thuật toán Kernighan-Lin [7].

2.1.2. Phân cụm phân cấp

Đồ thị có thể chứa cấu trúc phân cấp, nghĩa là mỗi cộng đồng có thể là một tập hợp các cụm nhỏ ở các cấp độ khác nhau [5]. Trong trường hợp như vậy, các kỹ thuật phân cụm phân cấp có thể được sử dụng để xác định cấu trúc cộng đồng đa cấp của đồ thị. Các kỹ thuật phân cụm phân cấp dựa trên đo lường độ tương đồng giữa các đỉnh. Chúng không cần kích thước và số lượng cộng đồng được xác định trước. Chúng có thể được biểu diễn tốt hơn bằng các cây dendrogram[4] (hay còn gọi là sơ đồ cây) là một dạng biểu đồ được sử dụng để minh họa cho sự sắp xếp các cụm (cluster) đã được phân cụm theo tầng. Các kỹ thuật phân cụm phân cấp có thể được phân loại thành hai loại [5]:

- Các thuật toán kết hợp

  Đây là một kỹ thuật từ dưới lên vì ban đầu nó coi mỗi nút là một cụm riêng biệt và liên tục hợp nhất chúng dựa trên sự tương đồng cao và kết thúc với cộng đồng duy nhất.

- Các thuật toán chia

  Đây là một kỹ thuật từ trên xuống vì ban đầu nó coi toàn bộ mạng lưới là một cụm duy nhất và liên tục chia nó bằng cách loại bỏ các liên kết nối các nút với sự tương đồng thấp và kết thúc với các cộng đồng duy nhất.

2.1.3. Phân cụm theo phần

Phân chia tập dữ liệu thành *k* cụm không giao nhau với kích thước được xác định trước. Mục tiêu là phân chia các điểm dữ liệu thành *k* cụm để tối thiểu hoặc tối đa hóa hàm chi phí dựa trên đo lường độ không giống nhau giữa các nút. Một số hàm chi phí

---

[4] https://en.wikipedia.org/wiki/Dendrogram

thường được sử dụng bao gồm tối thiểu *k*-trung bình, tổng *k*-phân cụm, *k*-phân cụm và *k*-trung tâm [8]. Một số nghiên cứu về các các kỹ thuật phân cụm phần bao gồm phân cụm *k*-trung bình (*k*-means) [9] và phân cụm *c*-trung bình mờ (fuzzy c-means) [10], đặc biệt trong phân cụm *c*-trung bình mờ, một nút có thể thuộc về nhiều cụm.

2.1.4. Phân cụm phổ

Phân cụm phổ bao gồm tất cả các kỹ thuật sử dụng các vectơ riêng của các ma trận để chia tập hợp các điểm dữ liệu dựa trên sự tương đồng giữa chúng [5]. Các nghiên cứu liên quan phương pháp phân tách phổ Laplacian [11] và Donath [12].

2.1.5. Các thuật toán chia

Ý tưởng chính của các thuật toán này là loại bỏ các cạnh giữa các cụm trong một mạng dựa trên sự tương đồng thấp để tách các cộng đồng ra khỏi nhau [13]. Một số nghiên cứu về kỹ thuật này như: thuật toán Girvan-Newman [14] lặp lại loại bỏ các cạnh dựa trên giá trị edge-betweenness (Tính trung tâm của cạnh giữa được định nghĩa là số đường đi ngắn nhất đi qua một cạnh trong đồ thị hoặc mạng); kỹ thuật của Radicchi [15] lặp lại việc loại bỏ cạnh dựa trên hệ số phân cụm của cạnh.

2.2. Các kỹ thuật phát hiện cộng đồng dựa trên tối ưu hóa tính mô – đun (modularity)

Các kỹ thuật phát hiện cộng đồng dựa trên tối ưu hóa *modularity* được xem là cách tiếp cận hiện đại và được sử dụng phổ biến hiện nay. Trong các thuật toán phát hiện cộng đồng, modularity là độ đo được sử dụng để đánh giá chất lượng của các cộng đồng được xác định trong mạng. Về cơ bản, nó đo lường mức độ thuật toán đã chia mạng thành các nhóm có kết nối nội bộ mạnh và kết nối yếu giữa các cộng đồng [16].

2.2.1. Các kỹ thuật tham lam (greedy)

❖ **Phương pháp tham lam của Clauset-Newman-Moore [17]**

Thuật toán tìm kiếm tham lam này là thuật toán đầu tiên được đề xuất cho tối ưu hóa modularity. Đây là một kỹ thuật hợp nhất. Bắt đầu, mỗi nút thuộc về một mô-đun riêng biệt, sau đó chúng được lặp lại việc hợp nhất các mô-đun dựa trên giá trị của modularity. Thuật toán này có độ phức tạp thời gian là $O(n^3)$ trên các mạng có đồ thị thưa [17]

❖ **Thuật toán Louvain [18]**

Louvain là một thuật toán tham lam heuristic để khám phá các cộng đồng trong các biểu đồ có trọng số phức tạp. Thuật toán này cũng dựa trên việc tối ưu hóa mô-đun. Nó chỉ định các cộng đồng khác nhau cho mỗi đỉnh; một trên mỗi đỉnh. Nó liên tục hợp nhất các nút dựa trên mức tăng của tính mô - đun. Nếu không đạt được thì nút vẫn ở trong cộng đồng của chính nó. Quy trình được lặp lại cho đến khi không thể cải thiện được nữa. Sau đó, nó sẽ xây dựng lại mạng theo cách các cộng đồng được xác định trong giai đoạn đầu tiên được thay thế bằng các siêu nút. Độ phức tạp về thời gian của nó là $O(n\log^2 n)$.

2.2.2. Thuật toán tiến hóa

Thuật toán tiến hóa là một lớp các thuật toán tối ưu hóa *metaheuristic* dựa trên trí tuệ nhân tạo. Dạng thuật toán này có khả năng học cục bộ và tìm kiếm toàn cầu hiệu quả. Các phương pháp này được chia thành hai lớp dựa trên tối ưu hóa đơn mục tiêu và đa mục tiêu.

Các thuật toán dựa trên tối ưu hóa đơn mục tiêu như: MAGA-Net [19], MLAMA-Net [20].

Các thuật toán dựa trên tối ưu hóa đa mục tiêu như: MOEA/D [21], I-NSGAII [22].

2.3. Các kỹ thuật phát hiện cộng đồng chồng chéo

Trong các mạng thực tế, hầu hết các nút có thể đồng thời thuộc về nhiều cộng đồng. Các kỹ thuật phát hiện cộng đồng truyền thống thất bại trong việc xác định các cộng đồng chồng chéo.

Clique percolation là kỹ thuật được biết đến nhất được sử dụng để xác định các cộng đồng chồng chéo trong các mạng. Kỹ thuật dựa trên ý tưởng các nhóm nút có khả năng được hình thành từ các cạnh nội bộ được kết nối chặt chẽ hơn so với các cạnh bên ngoài kết nối thưa thớt. Các cộng đồng được tạo thành từ các *k*-điểm có thể đề cập đến các đồ thị con hoàn chỉnh có *k* đỉnh. Hai nhóm được biết đến là kề nhau nếu chúng chia sẻ $k - 1$ nút. Cộng đồng *k*-điểm là thành phần khổng lồ được tạo thành từ tất cả các *k*-đồ thị kề nhau được kết nối như một chuỗi *k*-điểm [23].

Một số nghiên cứu liên quan kỹ thuật này gồm các cụm đồ thị hàng đầu [24], thuật toán truyền nhãn [25].

3. Tổng quan các bài toán về đồ thị lưỡng cực

Đồ thị lưỡng cực là đồ thị có các cạnh kết nối các nút từ hai tập hợp riêng biệt không có kết nối nào trong các tập hợp đó, đặt ra những thách thức và cơ hội riêng cho các nhà nghiên cứu. Một số nghiên cứu tiêu biểu về đồ thị lưỡng cực:

3. 1. Phát hiện cộng đồng

Tang và các cộng sự [26] đề xuất phương pháp để phát hiện cộng đồng trong việc phát triển các biểu đồ lưỡng cực nhằm tận dụng thông tin tạm thời để cải thiện độ chính xác trong việc theo dõi sự phát triển của cộng đồng. Cách tiếp cận này đặc biệt phù hợp với các ứng dụng như mạng xã hội có sự thay đổi liên tục.

Li và cộng sự [27] đã giải quyết bài toán phát hiện cộng đồng bằng cách kết hợp các thuộc tính nút vào thuật toán truyền nhãn để cải thiện khả năng phát hiện cộng đồng trong biểu đồ lưỡng cực.

3. 2. Dự đoán liên kết mạng

Wang và cộng sự [28] dùng mạng thần kinh đồ thị (GNNs - Graph Neural Networks) tận dụng các tính năng của nút và cấu trúc lưỡng cực để dự đoán liên kết trong mạng.

3. 3. Kết hợp và tối ưu hóa

Đồ thị lưỡng cực được ứng dụng nhiều trong các bài toán tối ưu hóa hàm DR-submodular (diminishing return submodular) trong các ngữ cảnh dưới nhiều ràng buộc khác nhau.

Liu và cộng sự [29] giải quyết bài toán tối đa hóa hàm DR-submodular dưới ràng buộc “bài toán ba-lô” bằng kỹ thuật luồng phát trực tiếp trên dữ liệu lưỡng cực. Zhang và các cộng [30] sự đề xuất thuật toán luồng phát trực tuyến để giải bài toán tối đa hóa hàm DR-submodular tăng đơn điệu trên lưới nguyên với ràng buộc số lượng cho tập chọn phần tử. Gong và cộng sự [31] nghiên cứu bài toán tối đa hóa hàm DR-submodular trên lưới nguyên dưới ràng buộc “bài toán ba-lô”, và đã đề xuất thuật toán dùng kỹ thuật tham lam có ngưỡng khi xét duyệt các phần tử.

Ngoài ra, đồ thị lưỡng cực còn được sử dụng rộng rãi trong nhiều lĩnh vực khác như:

- **Lý thuyết trò chơi:** Phân tích chiến lược tối ưu trong các trò chơi cạnh tranh.
- **Học máy:** Phát triển các mô hình học tập hiệu quả cho các bài toán phân loại, dự đoán và nhóm dữ liệu.
- **Mạng xã hội:** Phân tích cấu trúc mạng và xác định cộng đồng.
- **Logistics:** Lập kế hoạch tuyến đường, tối ưu hóa giao hàng.

LÝ DO LỰA CHỌN ĐỀ TÀI

Xã hội phát triển, con người phải đối mặt với nhiều bài toán tối ưu có hàm đa mục tiêu dưới nhiều ràng buộc và ngữ cảnh khác nhau. Trong số đó, chúng tôi quan tâm đến bài toán tối ưu hóa trong quảng cáo trực tuyến qua các nền tảng MXH. Vấn đề đặt ra cho các nhà quảng cáo là làm sao đưa thông tin sản phẩm hoặc dịch vụ tiếp cận đến càng nhiều khách hàng tiềm năng càng tốt, nhưng phải tiết kiệm tối thiểu chi phí. Bởi vì MXH ngày càng phổ biến và phát triển mạnh mẽ, với lượng người dùng bao phủ hầu khắp toàn cầu. Một cách ngắn gọn, chúng tôi gọi bài toán này là *tối đa hóa tầm ảnh hưởng của việc lan truyền tiếp thị trên các cộng đồng của mạng xã hội.*

Trong bài toán này, các nghiên cứu cho thấy cần giải quyết 2 điều cốt lõi:

(1) làm sao tìm được những người có tầm ảnh hưởng rộng, có khả năng tương tác để tác động mạnh vào người khác thông qua các chia sẻ thông tin trên MXH của mình.

(2) Những người ảnh hưởng này phải thuộc 1 cộng đồng gồm những khách hàng tiềm năng mà nhà quảng cáo muốn tiếp cận, phù hợp cho sản phẩm và dịch vụ được quảng cáo.

Như vậy để giải quyết 2 điều trên, chúng ta cần tìm các cộng động thuộc nhóm khách hàng tiềm năng và "gieo tin" cho những người có khả năng lan truyền, ảnh hưởng rộng trong các cộng đồng đó với chi phí hợp lý để đảm bảo tính tối đa về lợi nhuận.

Để giải bài toán này, chúng ta thực nghiệm trên dữ liệu của MXH ở dạng đồ thị lưỡng cực. Nhưng qua khảo sát, hiện nay các bộ dữ liệu lưỡng cực hiện có không thuộc MXH. Vì vậy, chúng tôi chọn đề tài "*Nghiên cứu thuật toán phát hiện cộng đồng trong dữ liệu MXH, ứng dụng xây dựng dữ liệu dạng đồ thị lưỡng cực*" để thực hiện. Đây được xem như giai đoạn đầu trong quá trình nghiên cứu của nhóm về các bài toán tối ưu trong phân tích MXH.

MỤC TIÊU, NỘI DUNG, PHƯƠNG PHÁP NGHIÊN CỨU

1. Mục tiêu

Vì hướng nghiên cứu của đề tài thuộc lĩnh vực giải quyết các bài toán tối ưu, đặc biệt trong MXH. Đây là một hướng nghiên cứu rất thú vị nhưng cũng nhiều thách thức vì nó có nhiều ứng dụng trong thực tế. Hầu hết các khái nhiệm và kiến thức về nhóm bài toán này đều hoàn toàn mới với chúng tôi, vì vậy, mục tiêu của chúng tôi cần thực hiện trong đề tài này là:

(1) Tìm hiểu các khái niệm liên quan bài toán tối ưu trong phân tích dữ liệu MXH và tổ chức cấu trúc dữ liệu MXH.

(2) Nghiên cứu một số thuật toán phát hiện cộng đồng trong dữ liệu MXH.

(3) Xây dựng mô hình để chuyển đổi dữ liệu MXH dạng đồ thị thông thường sang dạng dữ liệu đồ thị lưỡng cực sau khi phát hiện các cộng đồng.

(4) Xây dựng được chương trình thực nghiệm để thực hiện chuyển đổi dữ liệu MXH từ đồ thị thông thường sang đồ thị lưỡng cực.

2. Nội dung

Chúng tôi trình bày tóm tắt các nghiên cứu liên quan đề tài của nhóm; song song đó là các định nghĩa, khái nhiệm liên quan bài toán phát hiện cộng đồng; 2 thuật toán phát hiện cộng đồng mà chúng tôi nghiên cứu và thực nghiệm; mô hình và quy trình thực hiện chuyển đổi từ đồ thị thông thường sang dữ liệu đồ thị lưỡng cực; cuối cùng là phân tích đánh gia kết quả thực nghiệm.

Nội dung chính của báo cáo được tổ chức thành các chương như sau:

- Chương 1. Cơ sở lý thuyết
- Chương 2. Thuật toán phát hiện cộng đồng trong MXH
- Chương 3. Mô hình chuyển đổi dữ liệu đồ thị về dạng lưỡng cực.
- Chương 4. Thực nghiệm và đánh giá kết quả

3. Phương pháp nghiên cứu

Trong đề tài này, nhóm thực hiện kết hợp các phương pháp:

- tìm hiểu bài toán phát hiện cộng đồng trong MXH và các kiến thức cơ sở lý thuyết của bài toán;
- nghiên cứu các công trình, công bố liên quan đến bài toán và dữ liệu đồ thị lưỡng cực.
- nghiên cứu 2 thuật toán phát hiện cộng đồng nổi tiếng và dùng phổ biến hiện này, gồm: thuật toán tham lam Clauset-Newman-Moore (gọi tắt là Greedy Modularity) [17] và thuật toán Directed Louvain [18].
- đề xuất quy trình chuyển đổi dữ liệu từ dạng đồ thị thông thường sang đồ thị dữ liệu lưỡng cực.
- tiến hành thực nghiệm 2 thuật toán trên các bộ dữ liệu chuẩn, được lấy từ hệ thống SNAP[5] (Stanford Network Analysis Project – thư viện khai thác đồ thị và phân tích mạng của Jure Leskovec và cộng sự, thuộc trường đại học Stanford, Hoa Kỳ [32]. Đánh giá kết quả.

[5] https://snap.stanford.edu/index.html

ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

1. Đối tượng

- Bài toán phát hiện cộng đồng trên MXH và các thuật toán giải nó;
- Đồ thị lưỡng cực.

2. Phạm vi nghiên cứu

- Phạm vi nghiên cứu: đề tài tập trung nghiên cứu và thực nghiệm 2 thuật toán phát hiện cộng đồng phổ biến hiện này là Greedy Modularity và thuật toán Directed Louvain; dữ liệu thực nghiệm là các bộ dữ liệu chuẩn đã được kiểm chứng và công bố, thuộc hệ thống SNAP.
- Phạm vi thời gian: Đề tài này được thực hiện trong khoảng 4 tháng (kể từ tháng 02/2024 đến tháng 5/2024).

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1. Dữ liệu đồ thị MXH

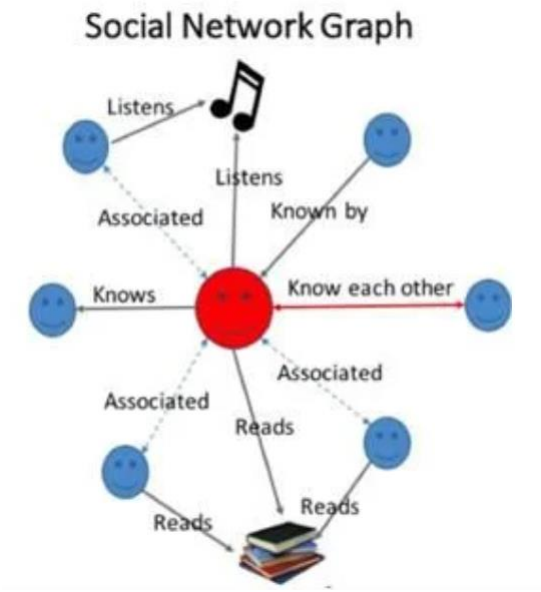
Mạng xã hội – là đồ thị về các mối quan hệ và tương tác giữa một nhóm người dùng – đóng một vai trò cơ bản như một phương tiện để lan truyền thông tin, ý tưởng và ảnh hưởng giữa các thành viên. Một ý tưởng hoặc đổi mới có thể xuất hiện – ví dụ như việc sử dụng điện thoại di động trong số sinh viên, việc áp dụng một loại thuốc mới trong ngành y, hoặc sự nổi lên của một phong trào chính trị trong một xã hội không ổn định – và nó có thể nhanh chóng chết đi hoặc xâm nhập đáng kể vào người dân.

Người dùng trên mạng (gọi tắt là người dùng) có thể giao tiếp, trao đổi thông tin với nhau bất chấp khoảng cách địa lý. Họ có thể chia sẻ thông tin, ý kiến, quan điểm, hoặc chia sẻ các bài viết của người khác,… Có thể nói, mỗi người dùng là một "phóng viên" trên MXHTT. Đặc tính này giúp cho các thông tin không chỉ phát tán nhanh chóng trên MXHTT mà nội dung còn rất đa dạng và phong phú. Ngoài ra, các MXHTT còn là nền tảng cho việc phát triển các ứng dụng, nên người dùng còn có thể tiến hành nhiều hoạt động khác do MXHTT cung cấp. Với sự phát triển của MXHTT hiện nay, ngày càng nhiều MXHTT mới được lập ra để khai thác nhiều khía cạnh khác nhau để đáp ứng nhu cầu của người dùng.

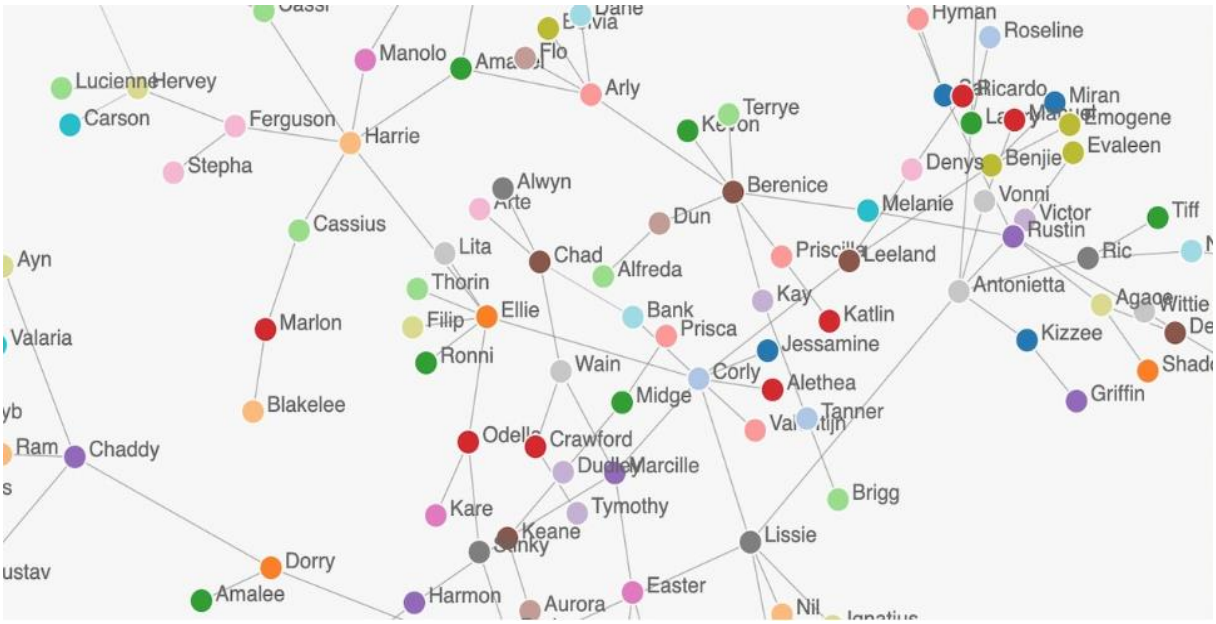
Theo như thống kê các mạng xã hội phổ biến hiện nay, vào năm 2023 có 4,9 tỷ người dùng mạng xã hội trên toàn thế giới. Theo nghiên cứu người dụng hoạt động hàng tháng theo nền tảng truyền thông xã hội của Statista, có 2,9 tỷ người sử dụng Facebook và coi đó nguồn thông tin chính thức. Những số liệu này cho thấy ngày càng có nhiều người dùng sử dụng MXHTT và chúng đóng vai trò quan trọng trong nhu cầu giao tiếp, giải trí, thu nhận thông tin của con người trong thời đại hiện nay.

Đồ thị mạng xã hội (SNG – Social Network Graph) thể hiện mối quan hệ phức tạp giữa các cá nhân hoặc thực thể trong MXH bằng các cấu trúc toán học. Các đồ thị này bao gồm các nút/đỉnh (đại diện cho các cá nhân hoặc thực thể) và các cạnh (đại diện cho các kết nối hoặc tương tác giữa chúng). Lý thuyết dữ liệu SNG cung cấp một

cấu trúc để phân tích và hiểu các mô hình phức tạp của mạng xã hội, giúp cho các nhà nghiên cứu tìm hiểu sâu vào sự phức tạp trong tương tác của con người và khám phá những hiểu biết có giá trị bên trong MXH. Hình 2, 3 minh họa ví dụ đồ thị MXH.

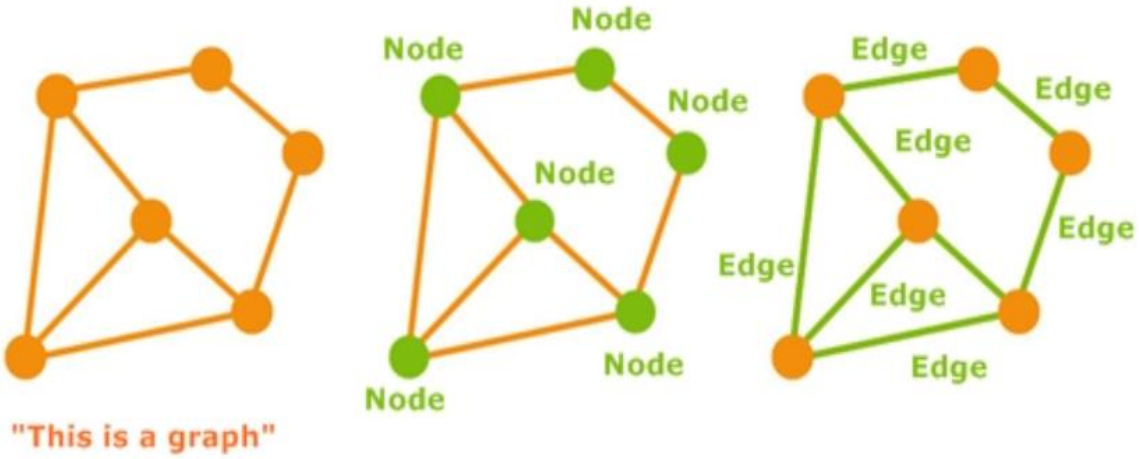


Hình 2. Ảnh minh họa đồ thị MXH



Hình 3. Ảnh minh họa các kết nối bàn bè của người dùng trên MXH

Một mạng xã hội phức tạp có thể được ánh xạ về dạng đồ thị $G(V,E)$, với V là tập các nút (đỉnh) và E là tập các cạnh. Số lượng của V và E lần lượt là n và m. Mạng $C(v,e)$ được gọi là mạng con nếu v là tập con của V và e là tập con của E.



Hình 4. Dữ liệu MXH có thể ánh xạ về dạng đồ thị thông thường

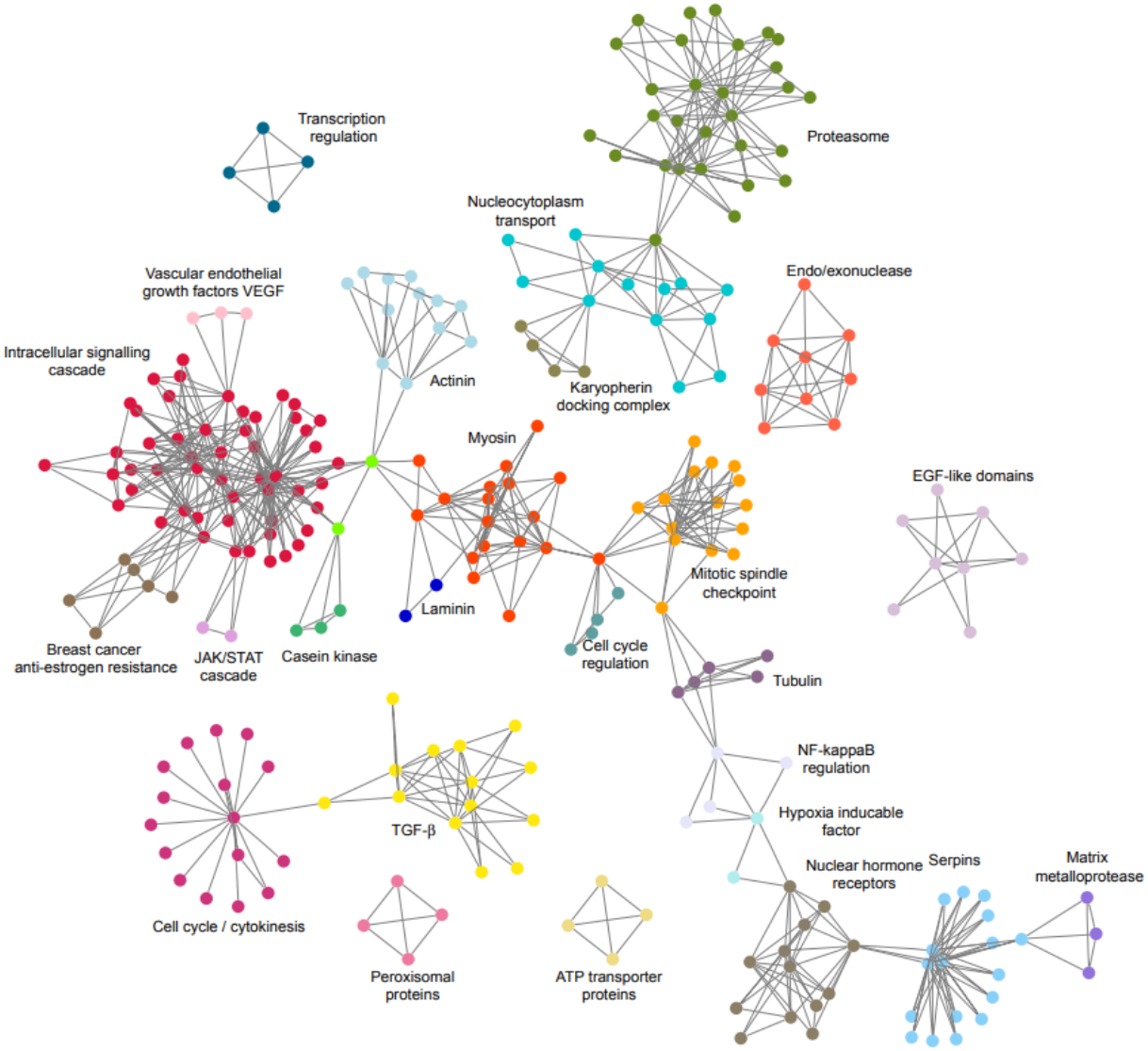
Gọi A là ma trận kề; hai nút là kề nhau nếu có một liên kết giữa chúng. Nếu tồn tại một liên kết giữa đỉnh i và đỉnh j thì $A_{ij} = 1$, ngược lại $A_{ij} = 0$. Mạng có trọng số thì w được gán vào các cạnh, trong đó w là số thực.

Gọi $K = \sum_{j=1}^{N} A_{ij}$ biểu thị bậc của đỉnh i là tổng số liên kết của đỉnh i. Mạng có hướng sẽ có hai loại bật: bậc vào và bậc ra. Số lượng cung (cạnh) mà một nút nhận được là bậc vào, sao cho $k_i^{in} = \sum_{j=1} A_{ij}$. Ngược lại, số cung mà một nút gửi ra gọi là bậc ra, sao cho $k_i^{out} = \sum_{j=1} A_{ij}$. Tổng bậc trong các mạng có hướng là $K_i = k_i^{in} + k_i^{out}$, trong khi tổng bậc của mạng vô hướng là $K_i = k_i^{in} = k_i^{out}$. Trong một mạng vô hướng thì tổng bậc là gấp đôi liên kết trong mạng.

Bậc nội của k_i^{int} của đỉnh i thuộc C là số liên kết của đỉnh i tới các nút của C sao cho $k_i^{int} = \sum_{i \in C} A_{ij}$. Bậc ngoại của đỉnh i là số cạnh kết nối với đỉnh i với các nút còn lại của G ngoài C sao cho $k_i^{ext} = \sum_{j \notin C} A_{ij}$.

2. Định nghĩa về cộng đồng

Cộng đồng trong mạng đơn, *là một tập hợp các nút được kết nối nhiều hơn trong cùng một tập so với phần còn lại của mạng. Chúng ta nói rằng cấu trúc cộng đồng trên một mạng đơn G là một phần của* $V_1 = \cup_{i-1}^{l} C_i$ *và* $V_2 = \cup_{j=1}^{k} D_j$*, trong đó* C_1 *là tập tách rời của từng cặp* V_1 *và* D_j *là từng cặp tách rời của* V_2*, sao cho tất cả các nút trong một* C_i *cụ thể được kết nối nhiều hơn với một tập con cụ thể của* V_2 *so với các nút còn lại trong* V_1*, và tương tự như vậy với phân vùng của* V_2*.*



Hình 5. Cấu trúc cộng đồng trong mạng lưới tương tác protein-protein. Biểu đồ mô tả sự tương tác giữa các protein trong tế bào ung thư của chuột. Các cộng đồng, được gắn nhãn bằng màu sắc, được phát hiện bằng Phương pháp thẩm thấu.

Nói cách khác, đồ thị G có k tập cộng đồng $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_3\}$, với $\mathcal{C}_i$ là đồ thị con của G và các đỉnh trong cộng đồng $\mathcal{C}_i$ là các đỉnh liên thông mạnh trong $\mathcal{C}_i$ và kết nối thưa thớt đến các cộng đồng hàng xóm $\mathcal{C}_j$. Ta nói các cộng đồng là không chồng chéo khi $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset \; \forall \; i, j$.

3. Tối đa hóa tính mô - đun (Modularity optimization)

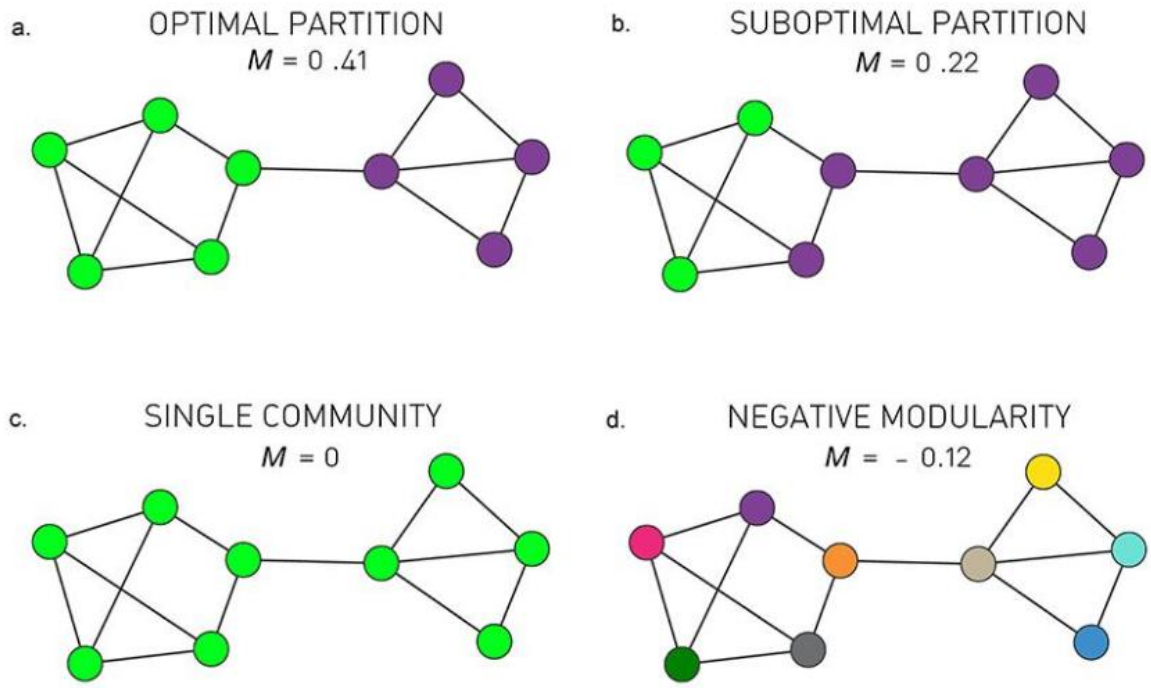
Tối đa hóa mô - đun hoạt động bằng cách xác định hàm lợi ích, được gọi là tính mô-đun, để đo lường chất lượng của việc phân chia mạng thành các cộng đồng. Người ta tối ưu hóa hàm lợi ích này qua các phân vùng có thể có của mạng quan tâm để tìm ra hàm mang lại điểm cao nhất, coi đây là sự phân chia chính xác của mạng. Vì số lượng phân chia của mạng là lớn theo cấp số nhân, nên chúng ta thường không thể thực hiện tối ưu hóa một cách triệt để, vì vậy thay vào đó chúng ta thực hiện tối ưu hóa gần đúng, trong đó nhiều phương pháp đã được thử bao gồm tham lam [14], tối ưu hóa cực trị [34], thuật toán di truyền [35].

Thuật toán Directed Louvain thì phổ biến để phát hiện cộng đồng [18] đã được áp dụng trong một số gói phần mền phân tích mạng, sử dụng sơ đồ tối ưu hóa mô - đun đa cấp và là một trong số những phương pháp phát hiện cộng đồng nhanh nhất trong thực tế.

Định nghĩa của hàm mô - đun là một hàm lợi ích của mạng. Cho một mạng và các phân vùng của mạng được phát hiện. Nó sẽ trả về điểm cao hơn nếu phân vùng đó là “tốt” và nhỏ hơn nếu nó “xấu”.

Mạng được biểu diễn bằng ma trận kề. Đối với mạng không có trọng số vô hướng gồm n nút, được đánh số từ 1 đến n, ma trận kề A là ma trận $n \times n$ đối xứng với các phần tử $A_{ij} = 1$ nếu có cạnh giữa các nút i và j và 0 nếu không có cạnh tương ứng. Chúng ta xem việc phân chia mạng thành q nhóm không chồng chéo, được đánh số từ 1 đến q và biểu diễn bằng g_isố nhóm mà nút i đươc gán. Vector $\boldsymbol{g}$ thể hiện tất cả các phân vùng của mạng. Khi đó số cạnh nằm trong các nhóm cho các phân vùng này

bằng $\frac{1}{2}\sum_{ij} A_{ij}\delta_{g_i g_j}$, với δ_{ij} là delta Kronecker và hệ số đầu công thức để tránh việc tính hai lần các cạnh [18].



Hình 6. Tối ưu hóa tính mô-đun để tìm ra phân vùng tốt nhất. Với M=0 là một cộng đồng đơn, M<0 ta có n cộng đồng (mỗi đỉnh là một cộng đồng).

Số lượng cạnh trong các nhóm không đánh giá được chất lượng phân vùng vì có thể dễ dàng tối ưu hóa bằng cách đặt tất cả các nút vào một nhóm và không có nút nào ở nhóm khác. Thay vào đó, sự tương tự đo lường sự khác biệt giữa số lượng cạnh thực tế trong các nhóm và số lượng cạnh dự kiến nếu cạnh được đặt ngẫu nhiên trong mạng. Trong trường hợp này, tác giả mô tả quá trình ngẫu nhiên hóa vị trí của các cạnh trong một mạng lưới trong khi giữ nguyên tổng số cạnh. Sau quá trình ngẫu nhiên hóa này, xác suất mà các nút i và j được kết nối bởi một cạnh được biểu diễn là P_{ij}. Sau đó, số lượng kỳ vọng của các cạnh trong các nhóm sau quá trình ngẫu nhiên hóa được tính bằng công thức $\frac{1}{2}\sum_{ij} P_{ij}\delta_{g_i g_j}$, và tính đa dạng cấu trúc được tỉ lệ thuận với số cạnh thực tế trừ đi số lượng kỳ vọng như vậy [18]:

$$Q = \frac{1}{m}\left(\frac{1}{2}\sum_{ij} A_{ij}\delta_{g_i g_j} - \frac{1}{2}\sum_{ij} P_{ij}\delta_{g_i g_j}\right) = \frac{1}{2m}\sum_{ij}\left(A_{ij} - P_{ij}\right)\delta_{g_i g_j} \quad (1)$$

trong đó m là số cạnh trong mạng và được đưa vào đây theo quy ước nó sẽ làm cho Q bằng một phần nhỏ của các cạnh chứ không phải là một số tuyệt đối làm ta dễ so sánh giữa các mạng khác nhau. Với mục đích là tối ưu hóa tính mô-đun nên m không tạo ra sự khác biệt nào cả.

Ghi chú rằng nếu ta bây giờ cho tất cả các nút trong cùng một nhóm thì $\delta_{g_i g_j} = 1$ với mọi i, j và

$$Q = \frac{1}{2m}\sum_{ijj}\left(A_{ij} - P_{ij}\right) = 0 \ (2)$$

vì như ta nói số cạnh được giữ trong mạng không đổi trong quá trình ngẫu nhiên hóa và do đó $\sum_{ij} P_{ij} = \sum_{ij} A_{ij} = 2m$. Do vậy thì ta không đạt được diểm mô-đun cao khi đặt tất cả các nút trong một nhóm lại với nhau. Tính mô-đun xảy ra với một số phân vùng không tầm thường (không tầm thường là đồ thị có nhiều hơn một đỉnh và có thể có các cạnh nối các đỉnh đó) khác mà ta xem là tốt nhất trong mạng. Đây là phương pháp tối đa hóa mô-đun.

Newman trong [4] định nghĩa xác xuất để kết nối giữa hai nút bằng

$$P_{ij} = \frac{k_i k_j}{2m} \qquad (3)$$

với $k_i = \sum_j A_{ij}$ là bật của nút i. Đây là sự lựa chọn phổ biến cho định nghĩa về tính mô-đun. Với sự lựa chọn này, tính mô - đun được đưa ra như sau

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta_{g_i g_j} \qquad (4)$$

Tuy nhiên còn một sự thật nũa là ngay cả định nghĩa này không phải lúc nào cũng đúng. Việc phát hiện cộng đồng bằng cách tối đa hóa mô-đun bằng định nghĩa (4) hoạt động trong nhiều tình huống vẫn có một thiếu sót cụ thể: nó không thể tìm thấy cấu trúc cộng đồng trong các mạng có nhiều cộng đồng nhỏ. Đặc biệt, nếu số lượng cộng đồng trong một mạng lớn hơn khoảng $\sqrt{2m}$, thì tính mô-đun tối đa sẽ không tương ứng với sự phân chia đúng. Thay vào đó, tính mô-đun tối đa sẽ có xu

hướng kết hợp các cộng đồng thành các nhóm lớn hơn và không thể giải quyết các phân chia nhỏ nhất trong mạng.

Để giải quyết vấn đề này, Arenas và các cộng sự [36] đề xuất một hàm mô-đun tổng quát được viết dưới dạng

$$Q(\gamma) = \frac{1}{2m}\sum_{in}\left(A_{ij} - \gamma\frac{k_i k_j}{2m}\right)\delta_{g_i g_j} \qquad (5)$$

Khi tham số γ = 1, Phương trình (5) giống với tính mô-đun truyền thống của Phương trình (4), nhưng các lựa chọn khác cho phép chúng ta biến đổi trọng số tương đối được gán cho các thuật ngữ cạnh quan sát và cạnh ngẫu nhiên. Nếu ta đặt nhiều trọng số hơn vào thuật ngữ cạnh quan sát (bằng cách làm γ nhỏ hơn), phân chia tính mô-đun tối đa ưa thích và phương pháp do đó có xu hướng tìm ra các cộng đồng lớn hơn. Nếu ta đặt nhiều trọng số hơn vào thuật ngữ cạnh ngẫu nhiên (γ lớn hơn), phương pháp sẽ tìm ra các cộng đồng nhỏ hơn.

Dựa theo công thức (5) và một số chứng minh đại số (xem ở phần Phụ lục), ta có thể tối giản công thức về dạng:

$$Q = \sum_{c=1}^{n}\left[\frac{L_c}{m} - \gamma\left(\frac{k_c}{2m}\right)^2\right] \qquad (6)$$

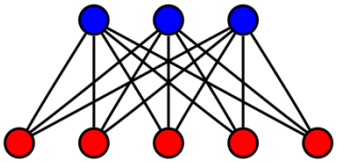
với tổng lặp trên tất cả các cộng đồng c, m là số cạnh của đồ thị, L_c là số lượng liên kết nội cho cộng đồng c, k_c là tổng bật của các đỉnh trong cộng đồng c, và γ là tham số cho độ đậm đặc của phân vùng.

Tham số γ thiết lập sự cân bằng tùy ý giữa các cạnh ngoài nhóm và các cạnh ngoài nhóm. Các mẫu nhóm phức tạp hơn có thể được phát hiện bằng cách phân tích cùng một mạng với nhiều giá trị gamma và sau đó kết hợp các kết quả lại [9]. Việc cho gamma bằng 1 là phổ biến nhất. Thông tin thêm về việc chọn lựa gamma có trong [10].

Đối với các đồ thị có hướng, ta chỉ cần thay thế k_c trong (6) bằng $k_c^{in}k_c^{out}$.

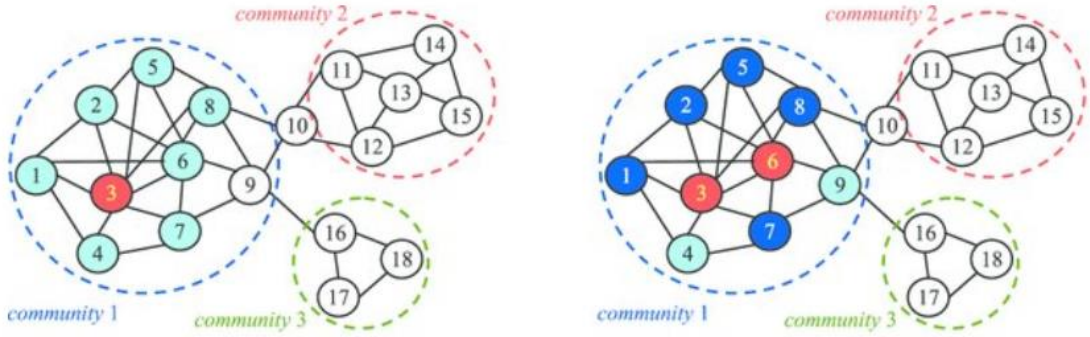
4. Dữ liệu đồ thị lưỡng cực

Đồ thị lưỡng cực (đồ thị hai phía bipartite graph[6] – hình 7): *là đồ thị trong đó các đỉnh có thể được chia thành hai tập hợp rời nhau sao cho tất cả các cạnh đều nối một đỉnh trong tập hợp này với một đỉnh trong tập hợp khác, không có cạnh nào nối giữa các đỉnh trong các tập rời rạc.*[33]



Hình 7. Đồ thị lưỡng cực với 2 tập nút V_1 có 3 đỉnh màu xanh và V_2 có 5 đỉnh màu đỏ, tập cạnh là các cạnh nối giữa các đỉnh thuộc V_1 và V_2

Diễn giải dữ liệu MXH biểu diễn trong đồ thị lưỡng cực để ứng dụng cho bài toán tối đa hóa tầm ảnh hưởng của việc lan truyền tiếp thị trên các cộng đồng của mạng xã hội như sau:



Hình 8. Các cộng đồng và kết quả lan truyền theo đỉnh số 3

Trên các MXH, mỗi tài khoản người dùng có thể tham gia nhiều nhóm (còn được gọi là cộng đồng). Mỗi cộng đồng chứa các tài khoản có mật độ liên kết "dày đặc" với nhau. Các nhà sản xuất muốn quảng bá sản phẩm thông qua nền tảng MXH, sẽ có nhiều chiến lược khác nhau. Một trong những chiến lược là *chọn và phân bổ "chi phí" tối thiểu cho các cộng đồng sao cho tác động lan truyền ảnh hưởng đến nhiều tài khoản người dùng nhất*. Vấn đề để giải quyết bài toán không chỉ dừng ở việc chỉ cần chọn các cộng đồng có nhiều thành viên là đủ, mà còn có nhiều yếu tố khác ràng buộc như: chi phí, thời gian chọn lựa cộng đồng phù hợp sản phẩm quản bá, sự tương tác và

[6] https://mathworld.wolfram.com/BipartiteGraph.html

ảnh hưởng của người dùng trong cộng đồng,… Chúng ta gọi các ràng buộc trên là "chi phí ảnh hưởng" của cộng đồng mà nhà sản xuất cần bỏ ra để quảng bá sản phẩm tới người dùng, nhưng tổng chi phí không thể vượt ngưỡng *k*. Nói cách khác, bài toán này là tìm các cộng động mà có phân bổ "chi phí ảnh hưởng" nhỏ nhất $(\leq k)$ sao cho tác động lan truyền tới nhiều người dùng nhất, nghĩa là hàm mục tiêu "lợi nhuận" đạt tối đa. Hình 8 minh họa việc gieo tin cho người dùng đỉnh số 3 trong cộng đồng để đạt hiệu quả lan truyền tốt nhất.

Diễn giải bài toán ở dạng đồ thị lưỡng cực:

Cho đồ thị *G* dạng lưỡng cực thể hiện dữ liệu của một MXH, $G(V;E)$ với *V* là tập các đỉnh được chia thành 2 phần (V_1; V_2), V_1 được định nghĩa là tập các cộng đồng của MXH, V_2 là tập các người dùng trên MXH; $E \subseteq V_1 \times V_2$ là tập các cạnh. Mỗi nút $v_1 \in V_1$ có một giá trị $\tau_{v_1} \in \mathbb{Z}_+$ thể hiện "chi phí tối đa" có thể cấp cho cộng đồng v_1. Mỗi cạnh $v_1v_2 \in E$ được liên kết có kèm trọng số $p(v_1v_2) \in [0;1]$, có nghĩa là khi chọn cộng đồng v_1 sẽ có xác suất $p(v_1v_2)$ lan truyền ảnh hưởng đến người dùng v_2. Hàm mục tiêu của bài toán là tìm tập chứa các cộng đồng v_1 sao cho tác động lan truyền đến số người dùng v_2 là tối đa.

CHƯƠNG 2: CÁC THUẬT TOÁN PHÁT HIỆN CỘNG ĐỒNG

Trong chương này, chúng tôi trình bày 2 thuật toán phát hiện cộng đồng mà nhóm chúng tôi nghiên cứu và thực nghiệm. Đó là thuật toán tham lam thuật toán tham lam Clauset-Newman-Moore (gọi tắt là Greedy Modularity) [17] và thuật toán Directed Louvain [18]. Đây là 2 thuật toán được ứng dụng phổ biến hiện nay, mỗi thuật toán sẽ có ưu và nhược điểm riêng tùy vào đặc trưng cấu trúc của bộ dữ liệu. Chúng tôi sẽ phân tích điểm này trong phần nhận xét đánh giá kết quả thực nghiệm.

1. Thuật toán tham lam Greedy Modularity

Đây là thuật toán đầu tiên được đề xuất để tối ưu hóa mô – đun (modularity). Đây là một kỹ thuật kết tụ, trong đó ban đầu, mỗi nút thuộc về một mô-đun riêng biệt, sau đó chúng được hợp nhất lặp đi lặp lại dựa trên mức tăng mô-đun. Nghĩa là nó lặp đi lặp lại việc liên kết các nút sao cho tăng được tính mô-đun của phân vùng mới. Các bước thực hiện chi tiết như sau:

- **Bước 1**: Gán mỗi nút là một cộng đồng riêng. Do vậy ta sẽ bắt đầu với n cộng đồng.
- **Bước 2**: Ta kiểm tra từng cặp cộng đồng được kết nối bằng ít nhất một liên kết và tính toán các biến thể tính mô-đun thu được khi ta hợp nhất 2 cộng đồng này.
- **Bước 3**: Xác định các cặp cộng đồng mà ΔM lớn nhất và hợp nhất chúng lại. Lưu ý rằng tính mô-đun của một phân vùng cụ thể luôn được tính toán từ cấu trúc liên kết đầy đủ của mạng. Sau đó ta ghi lại M.
- **Bước 4**: Lặp lại bước 2 cho đến khi tất cả các nút được hợp nhất thành một cộng đồng duy nhất.
- **Bước 5**: Trả về các phân vùng có tính mô-đun M cao nhất.

Ví dụ. Tìm cộng đồng sử dụng phương pháp Greedy Modularity cho đồ thị 3 nút ở hình 9:

Dựa theo công thức (6) ta có tính mô-đun cho một cộng đồng được tính như sau.

$$Q_c = \frac{L_c}{m} - \gamma \left(\frac{k_c}{2m}\right)^2 \quad (7)$$

Khi đó:

Với tất cả nút: $L_c = 0$

$$Q_A = \frac{0}{3} - \left(\frac{2}{2 \times 3}\right)^2 = -\frac{1}{9};$$

$$Q = -\frac{1}{3} = 0.33$$

Lần lặp 1 (kết hợp A và B):

$$Q_{AB} = \frac{1}{3} - \left(\frac{4}{2 \times 3}\right)^2 = -\frac{1}{9}$$

$$Q_C = \frac{0}{3} - \left(\frac{2}{2 \times 3}\right)^2 = -\frac{1}{9}$$

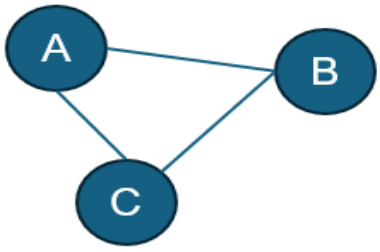
$$Q = -2/9$$

Lần lặp 2:

$$Q_{ABC} = \frac{3}{3} - \left(\frac{6}{2 \times 3}\right)^2 = 0$$

$$Q = 0$$

Kết thúc.



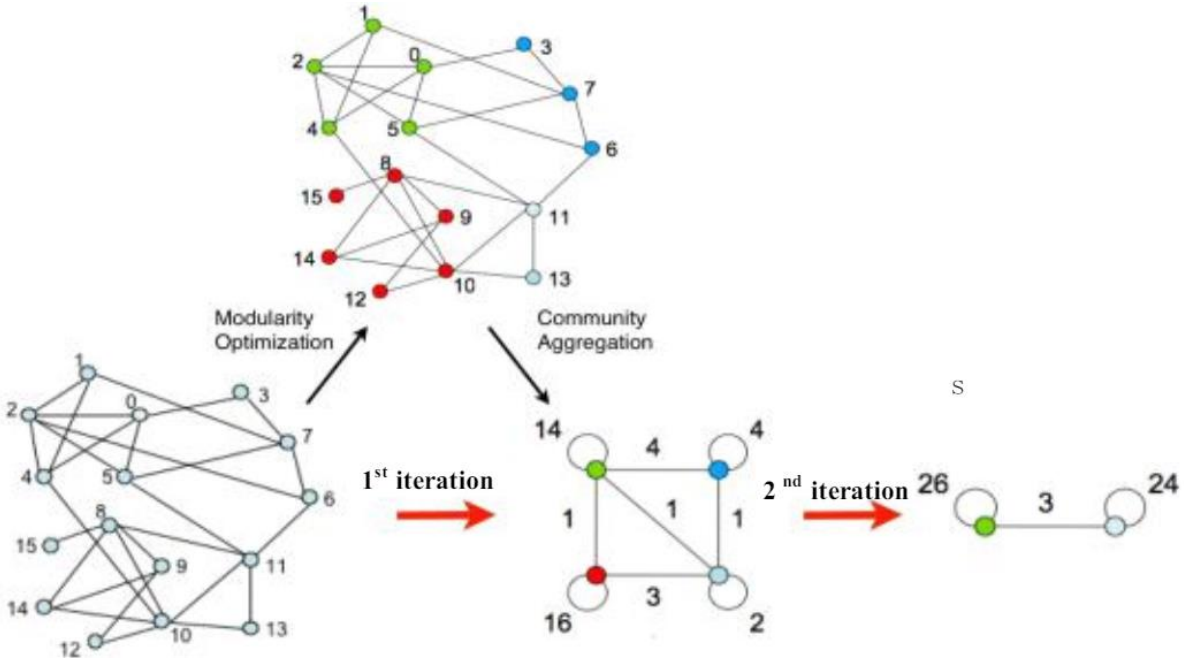
Hình 9. Tim cộng đồng chot đồ thị nhỏ gồm 3 đỉnh

Phân tích về thời gian chạy của thuật toán này, việc tính ΔM là $O(1)$ ở bước đầu, các bước tiếp theo là $O(m)$. Sau khi kết hợp lại các cộng đòng ở bước 3, việc cập nhật cho các cạnh là $O(n)$. Bước 2 và 3 lặp $n - 1$ lần suy ra độ phức tạp tổng thể cho thuật toán là $O((n + m)n)$.

2. Thuật toán Directed Louvain

Directed Louvain là một thuật toán tham lam heuristic để khám phá các cộng đồng trong các biểu đồ có trọng số phức tạp. Nó cũng dựa trên việc tối ưu hóa mô-đun. Nó chỉ định các cộng đồng khác nhau cho mỗi đỉnh; một trên mỗi đỉnh. Nó liên tục hợp nhất các nút dựa trên mức tăng của tính mô - đun. Nếu không đạt được thì nút vẫn ở trong cộng đồng của chính nó. Quy trình được lặp lại cho đến khi không thể cải thiện được nữa. Sau đó, nó sẽ xây dựng lại mạng theo cách các cộng đồng được xác định trong giai đoạn đầu tiên được thay thế bằng các siêu nút. Độ phức tạp thời gian của nó là $O(n \log n)$.

Thuật toán phát hiện cộng đồng Louvain là một phương pháp đơn giản để trích xuất cấu trúc cộng đồng của mạng. Thuật toán hoạt động theo 2 giai đoạn. Hình 10 minh họa thuật toán Louvain.



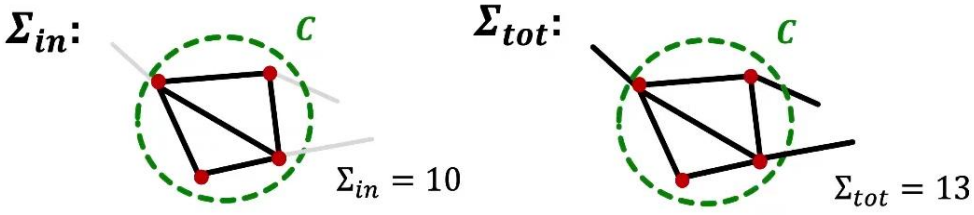
Hình 10. Minh họa thuật toán Louvain

- **Giai đoạn 1:** thuật toán gán mỗi nút nằm trong cộng đồng của chính nó. Như vậy ta có n cộng đồng. Sau đó, nó di chuyển từng nút sang tất cả các cộng đồng láng giềng của nó để tìm mức tăng mô - đun dương tối đa. Nếu không đạt được lợi ích dương thì nút vẫn ở trong cộng đồng ban đầu của nó.

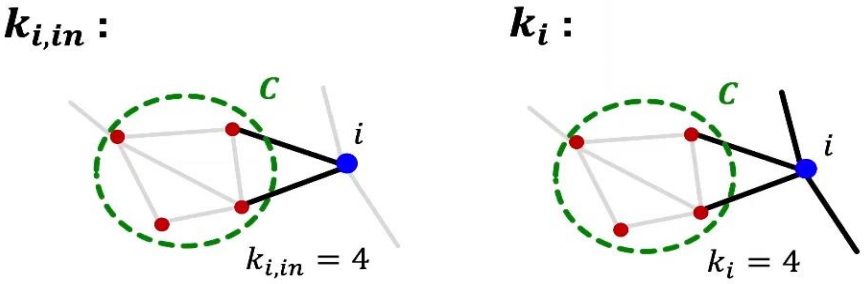
Mức tăng (lợi ích) mô-đun thu được bằng cách di chuyển một đỉnh bị cô lập i vào một cộng đồng C có thể dễ dàng tính được bằng công thức sau:

$$\Delta Q = \frac{k_{i,in}}{2m} - \gamma \frac{\sum_{tot} C \cdot k_i}{2m^2} \quad (8)$$

với m là kích thước của đồ thị (số cạnh của đồ thị), $k_{i,in}$ là tổng trọng số của các liên kết từ i đến các nút trong C, k_i là tổng trọng số của các liên kết tới nút i, $\sum_{tot} C$ là tổng trọng số của các liên kết tới các nút trong C và γ là tham số cho độ đậm đặc của phân vùng.



Hình 11. $\sum_{tot} C$ và $\sum_{in} C$.



Hình 12. $k_{i,in}$và k_i.

Trong trường hợp có hướng mức tăng mô-đun có thể được tính bằng công thức sau:

$$\Delta Q = \frac{k_{i,in}}{m} - \gamma \frac{k_i^{out} \cdot \sum_{tot}^{in} C + k_i^{in} \cdot \sum_{tot}^{out} C}{m^2} \quad (9)$$

với k_i^{out}, k_i^{in} là bậc có trọng số bên ngoài và bên trong của nút i và $\sum_{tot}^{in} C$, $\sum_{tot}^{out} C$ là tổng các liên kết vào và liên kết ra của các nút trong C.

Giai đoạn đầu kết thúc khi mà không có sự di chuyển nút riêng lẻ nào cải thiện được tính mô-đun nửa.

- **Giai đoạn 2:** gồm việc xây dựng một mạng mới có các nút là các cộng đồng được tìm thấy trong giai đoạn đầu tiên, được gọi là các siêu nút. Để làm như vậy trọng số của các liên kết giữa các nút mới được tính bằng tổng trọng số của các liên kết giữa các nút trong hai cộng đồng tương ứng. Khi giai đoạn này hoàn tất, có thể áp

dụng lại giai đoạn đầu tiên để tạo ra các cộng đồng lớn hơn với tính mô - đun tăng lên.

Thuật toán Louvain có thể tóm tắt ngắn gọn như sau:

- **Giai đoạn 1**
  - **Bước 1:** Gán mỗi nút là một cộng đồng riêng. → Bắt đầu với n cộng đồng.
  - **Bước 2**: Với mỗi nút i, lần lượt xóa i khỏi cộng đồng chứa nó, ghép i vào cộng đồng các nút j láng giềng. Sau đó tính
  - **Bước 3**: i sẽ được ghép vào cộng đồng nào có delQ lớn nhất, ngược lại nếu delQ<=0 thì i vẫn ở cộng đồng ban đầu của nó.
  - **Bước 4**: ghép thử các cặp cộng đồng có liên kết với nhau thành 1 cộng đồng mới, và tính modularity.
  - **Bước 5**: Lập lại quá trình bước 2->4 cho tới khi delQ của tất cả các nút là không tăng.
- **Giai đoạn 2**
  - Ánh xạ mỗi cộng đồng tìm được ở GĐ1 là 1 nút trong đồ thị mới, và lặp lại quá trình GĐ1 với đồ thị mới này.
  - Quá trình lặp lại cho đến khi modularity của các cộng đồng đạt được ngưỡng quy định

CHƯƠNG 3: QUY TRÌNH CHUYỂN DỮ LIỆU ĐỒ THỊ THÔNG THƯỜNG SANG ĐỒ THỊ LƯỠNG CỰC

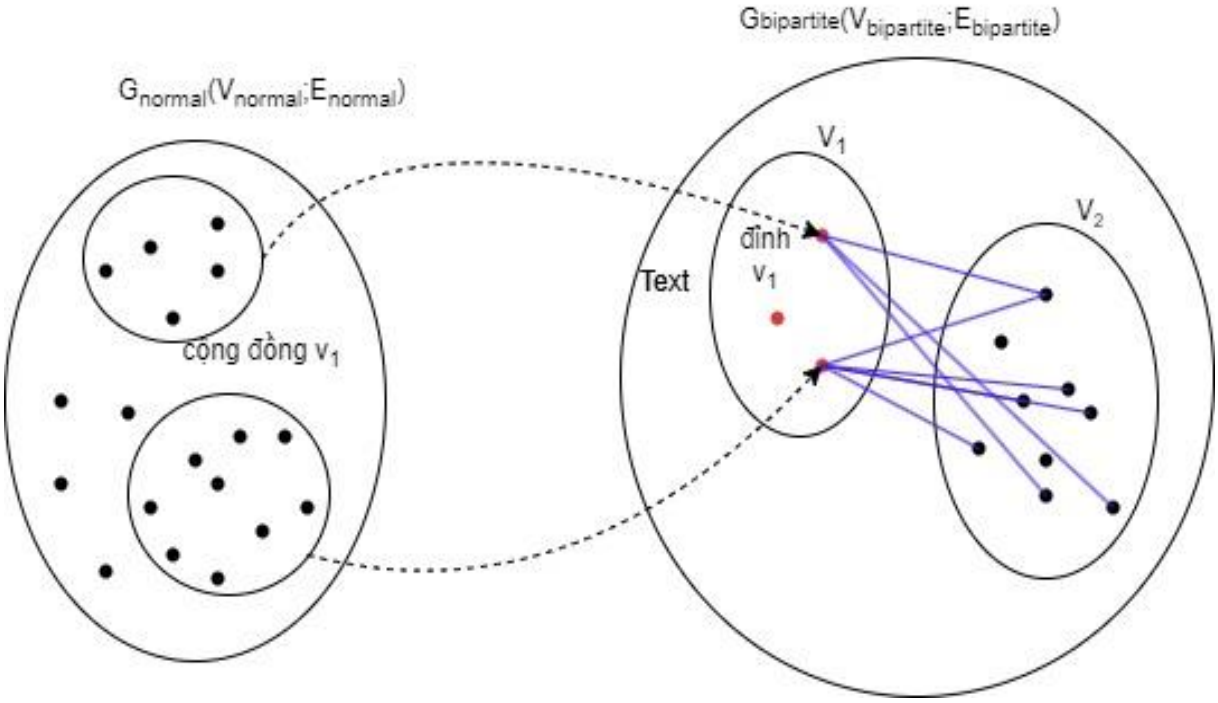
1. Chuyển đồ thị thông thường sang dạng đồ thị lưỡng cực

Quá trình thực hiện chuyển đổi đồ thị thông thường $G_{normal}(V_{normal}, E_{normal})$ sang đồ thị lưỡng cực $G_{bipartite}(V_{bipartite}, E_{bipartite})$ theo các bước sau:

- Mỗi cộng đồng tìm được trong G_{normal} chuyển thành đỉnh v_1 trong chập V_1
- Các đỉnh $v \in V_{normal}$ chuyển sang $v \in V_2$
- Cạnh $v_1, v_2 \in E_{birpartite}$ nếu v_2 thuộc về cộng đồng v_1
- Với mỗi cạnh $v_1v_2 \in E$ của $G_{bipartite}$ có một trọng số (là xác xuất ảnh hưởng của cộng đồng v_1 đến thành viên v_2) $p(v_1v_2) \in [0; 1]$. Được định nghĩa như sau:

$$p(v_1v_2) = \frac{\deg(v_2 \in \mathcal{G}_{v_1})}{\max(\deg(u)|u \in \mathcal{G}_{v_1})} \quad (10)$$

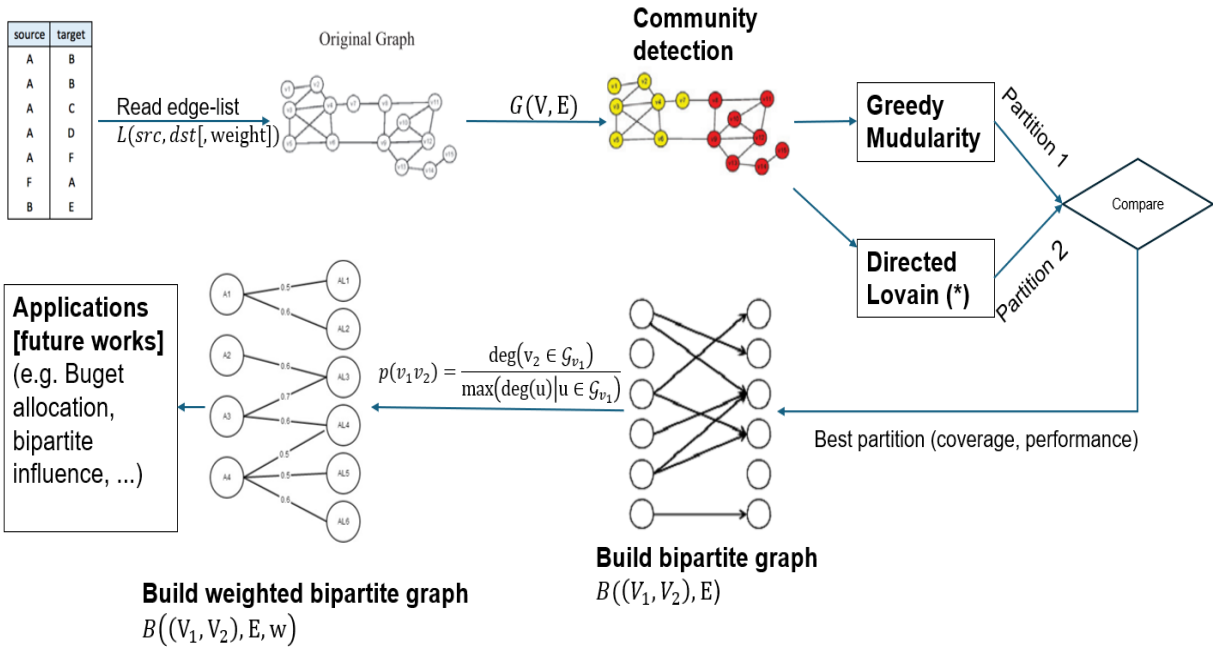
- , trong đó $\mathcal{G}_{v_1}$ đồ thị con của G chứa tất cả hàng xóm của v_1 trong $G_{bipartite}$.



Hình 13. Ánh xạ các cộng đồng tìm được sang các tập cạnh của đồ thị lưỡng cực.

2. Mô hình đề xuất

Từ cơ sở lý thuyết trên, chúng tôi để xuất mô hình để chuyển từ đồ thị mạng xã hội sang dạng đồ thị lưỡng cực có trọng số với trọng số là xác xuất ảnh hưởng của cộng đồng (tập V_1) lên các các nhân của nó (tập V_2). Hình 14 thể hiện mô hình này.



Hình 14. Mô hình đề xuất chuyển đổi dữ liệu đồ thị thông thường sang độ thì lưỡng cực, và hướng nghiên cứu tương lại là ứng dụng đồ thị lưỡng cực vào bài toán tối ưu hóa chi phí và lợi nhuận trong quảng cáo trực tuyến trên MXH.

Các bước thực hiện theo trình tự như sau:

- **Bước 1:** đọc tập cạnh $L(src, dst[, weight])$ để xây dựng một đồ thị MXH $G(V, E)$ với src là tập nguồn, dst là tập đích và $weight$ là tập trọng số tương ứng lần lượt với hai đỉnh trong tập cạnh trên, $weight$ có thể có hoặc không tùy vào nguồn của dữ liệu. Lưu ý, ta cần xác định là đồ thị có hướng hoặc vô hướng trước khi đọc danh sách cạnh này.
- **Bước 2:** tìm danh sách các tập hợp của các phân vùng (cộng đồng) của G bằng các thuật toán tìm kiếm cộng đồng như Directed Louvain hay Greedy Modularity. Khi chạy các thuật toán trên chúng trả về các danh sách chứa các phân vùng khác nhau. Ta tiến hành so sánh các thông số như độ bao phủ, hiệu suất của các danh sách này và chọn ra danh sách chứa các phân vùng phù hợp.

❖ **Bước 3:** Khi ta thu được danh sách các cộng đồng $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_3\}$ ta sẽ chuyển đổi sang đồ thị lưỡng cực $B(V, E)$, trong đó V được chia thành cặp của hai tập đỉnh (V_1, V_2), với V_1 biểu thị tập hợp của cách đỉnh nguồn trong trường hợp này là các cộng đồng, V_2 thể hiện các đỉnh đích hay mục tiêu (cá nhân, khách hàng). Do chính chất của đồ thị lưỡng cực các đỉnh trong tập V_1 không có kết nối với nhau, tương tự cho V_2, mà chỉ có các kết nối giữa các đỉnh trong V_1 và V_2.

Với mỗi cạnh $v_1v_2 \in E$ của B có một trọng số (là xác xuất ảnh hưởng của cộng đồng v_1 đến thành viên v_2) $p(v_1v_2) \in [0;\ 1]$. Được định nghĩa như công thức (10) là tỉ lệ của số bậc của $v2$ trong $\mathcal{G}_{v_1}$trên số bậc cao nhất của $\mathcal{G}_{v_1}$. Trong đó $\mathcal{G}_{v_1}$ đồ thị con của G chứa tất cả hàng xóm của v_1 trong B. Như đã nêu các đỉnh trong V_2 không có liên kết với nhau nên để tìm bật của các đỉnh trong V_2 của B ta phải dựa vào đồ thị G đã khởi tạo ban đầu và tạo ra một đồ thị con $\mathcal{G}_{v_1}$ để tìm được số bậc của nút đó và số bậc cao nhất của của nút đó trong chính cộng đồng của nó.

❖ **Bước 4.** Từ đồ thị lưỡng cực có trọng số vừa tìm được ta trả về một danh sách cạnh của $B((V_1, V_2), E, w)$ là $L_B(cmty, mem, weight)$ trong đó $cmty$ là danh sách các cộng đồng có thể lặp lại tượng ứng với danh sách các thành viên mem của nó và $weight$ là trọng số w cho từng kết nối của cạnh tương ứng. Từ tập dữ liệu đồ thị lưỡng cực đã xây dựng được, chúng tôi áp có thể áp dụng cho các bài toán phân bổ ngân sách, lan truyền trên mạng lưỡng cức, ... một trong số các bài toán đó là *phân bổ ngân sách trong mô hình ảnh hưởng của mạng lưỡng cực.*

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ CÁC KẾT QUẢ

1. Dữ liệu thực nghiệm

Chúng tôi thực nghiệm với 3 bộ dữ liệu mạng xã hội trên hệ thống SNAP[7] (Stanford Network Analysis Project – thư viện khai thác đồ thị và phân tích mạng của Jure Leskovec và cộng sự, thuộc trường đại học Stanford, Hoa Kỳ [32]) có kích thước số nút, số cạnh và cấu trúc phân bố khác nhau, gồm:

- **email-Eu-core network**[8]: Mạng được tạo bằng dữ liệu email từ một tổ chức nghiên cứu lớn ở châu Âu, Có một cạnh (*u, v*) trong mạng nếu người *u* gửi cho người *v* ít nhất một email. Các e-mail chỉ thể hiện sự liên lạc giữa các thành viên của tổ chức và tập dữ liệu không chứa các tin nhắn đến hoặc đi đến phần các người không thuộc mạng.
- **Social circles: Facebook**[9]: Tập dữ liệu này bao gồm 'danh sách bạn bè' từ Facebook. Dữ liệu Facebook được thu thập từ những người tham gia khảo sát bằng ứng dụng Facebook. Tập dữ liệu bao gồm các tính năng nút (hồ sơ), vòng kết nối trong danh sách bạn bè.
- **Social circles: Twitter**[10]: Tập dữ liệu này bao gồm 'danh sách bạn bè' từ Twitter. Dữ liệu Twitter được thu thập từ các nguồn công cộng. Tập dữ liệu bao gồm các tính năng nút (hồ sơ) và vòng kết nối trong bạn bè.

2. Kết quả thực nghiệm

Chương trình thực nghiệm với hai thuật toán phát hiện cộng đồng là Greedy Modularity (Clauset, A., Newman, M. E., & Moore) và Directed Louvain. Điểm chung của 2 thuật toán này là đều dựa trên kỹ thuật tối đa hóa tính mô-đun (modularity), hay nói cách khác là tính độ kết nối "dày đặc" của mỗi cộng đồng.

Kết quả chuyển đổi dữ liệu mạng xã hội từ dạng đồ thị thông thường thành đồ thị lưỡng cực. Kết quả thực nghiệm về phát hiện cộng đồng mô tả dưới bảng.1. Sau

---

[7] https://snap.stanford.edu/index.html

[8] https://snap.stanford.edu/data/email-Eu-core.html

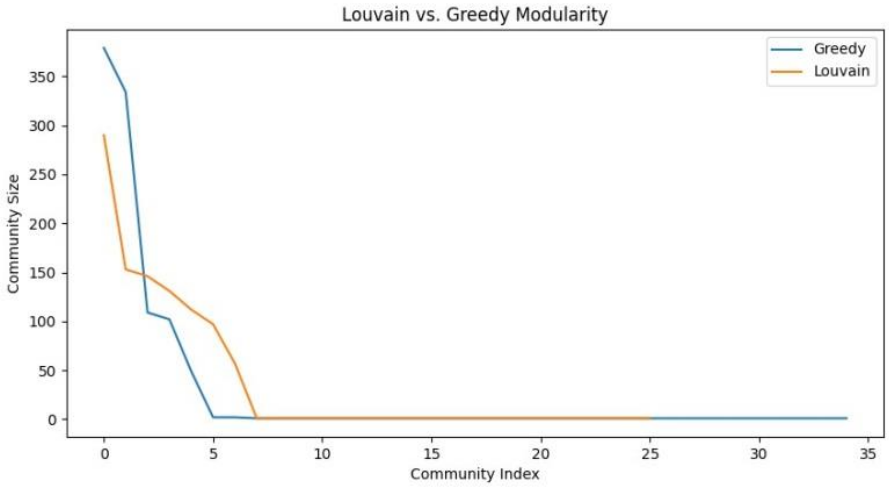
[9] https://snap.stanford.edu/data/ego-Facebook.html

[10] https://snap.stanford.edu/data/ego-Twitter.html

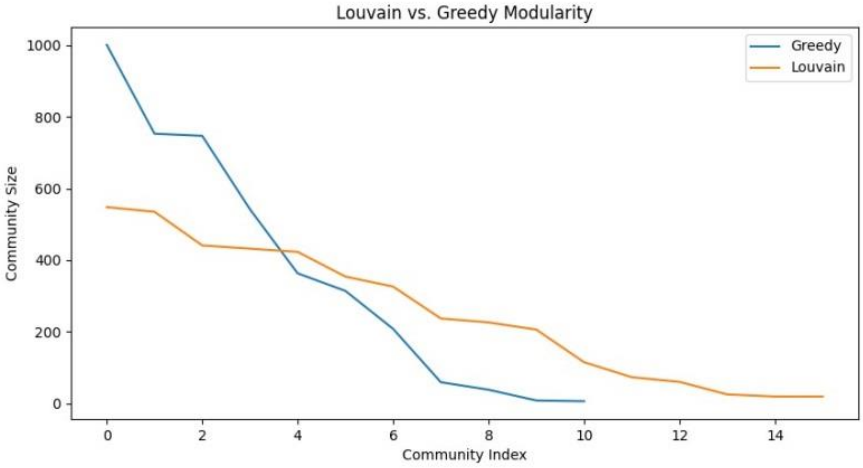
quá trình thực nghiệm, chuyển từ đồ thị thông thường sang đồ thị lưỡng cực bằng 2 kỹ thuật Greedy và Directed Louvain, kết quả thu được như ở các hình 15,16 và 17.

Bảng 1. Mô tả dữ liệu và kết quả thực nghiệm

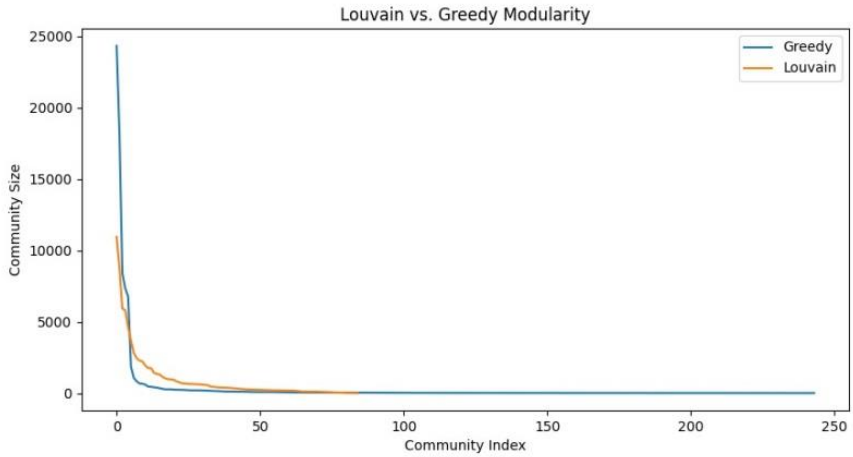
	Email Eu – Core		**Ego - Facebook**		**Ego - Twitter**	
Số đỉnh	1,005		8,039		81,306	
Số cạnh	25,571		88,234		1,768,149	
Thuật toán	Greedy	Louvain	Greedy	Louvain	Greedy	Louvain
Số cộng đồng được phát hiện	35	26	11	16	244	85
Thời gian chạy (s)	3.89	1.14	27.2	26.7	13803	12702
Thời gian trung bình phát hiện 1 cộng đồng (s)	0.11	0.04	2.47	1.67	56.57	149.44
So sánh Louvain nhanh hơn Greedy (lần)	**2.75**		**1.48**		**1.09**	



*Hình 15. Kết quả số cộng động và mật độ kích thước cộng đồng phát hiện được của bộ dữ liệu **email-Eu-core network***



Hình 16. Kết quả số cộng động và mật độ kích thước cộng đồng phát hiện được của bộ dữ liệu ***Social circles: Facebook***



Hình 17. Kết quả số cộng động và mật độ kích thước cộng đồng phát hiện được của bộ dữ liệu ***Social circles: Twitter.***

3. Phân tích và đánh giá kết quả thực nghiệm

Thông qua thực nghiệm chuyển đổi dạng đồ thị (ở bảng 1 và hình 15,16 và 17), kết quả có thể thấy:

❖ Tùy theo đặc trưng của mỗi dữ liệu mà số cộng đồng được phát hiện của mỗi thuật toán sẽ khác nhau. Nhìn chung, Greedy thường phát hiện số cộng đồng nhiều hơn và mật độ "dày đặc" của các cộng đồng cao hơn thuật toán Directed Louvain. Điều này là hiển nhiên vì Greedy có tính "tham lam" sẽ phát hiện cộng đồng, số thành viên cùng cộng đồng sao cho nhiều nhất có thể, trong khi Directed Louvain chỉ cần phát hiện cộng đồng có độ "gắn kết" đạt ngưỡng yêu cầu là nó dừng, để chuyển tìm cộng đồng tiếp theo. Đánh đổi cho kết quả đó, Greedy thường có thời gian thực thi lâu

hơn Directed Louvain. Nhưng với bộ dữ liệu Twitter, Louvain lại chạy chậm hơn và số cộng đồng thì ít hơn. Vì vậy tùy vào đặc trưng của dữ liệu đầu vào và yêu cầu của dữ liệu đầu ra mà người dùng lựa chọn thuật toán cho phù hợp.

- Khi ta cần dữ liệu với cộng đồng lớn có độ "dày đặc" cao, chúng ta chọn Greedy Modularity.
- Khi ta cần dữ liệu có nhiều cộng đồng có thể độ "dày đặc" không cao, chúng ta chọn Louvain

❖ Ưu điểm của Greedy Modularity là tìm được những cộng đồng có tính đậm đặc cao và tính mô-đun cao. Tuy vậy, kết quả mà Louvain mang đến sự phân phối đồng đều ở các cộng đồng tìm được cùng với đó là thời gian chạy tối ưu hơn với Greedy Modularity. Với các công thức đã nêu của thuật toán Louvain, ra thấy rằng thuật toán này không chỉ dựa trên bậc của các nút để phân vùng các cộng đồng mà còn sử dụng cả trọng số của đồ thị G. Do vậy nếu đồ thị MXH của ta có trọng số thì nên sử dụng Lovain để phát hiện cộng đồng tốt hơn vì trọng số có thể thể hiện dược sở thích, sự quan tâm của các cá nhân (các đỉnh) trong mạng từ đó cộng đồng mà chúng ta phát hiện được sẽ tốt hơn về mặc logic.

❖ Việc phát hiện cộng đồng bằng cách tối đa hóa mô-đun hoạt động trong nhiều tình huống vẫn có một thiếu sót cụ thể: nó không thể tìm thấy cấu trúc cộng đồng trong các mạng có nhiều cộng đồng nhỏ. Đặc biệt, nếu số lượng cộng đồng trong một mạng lớn hơn khoảng $\sqrt{2\mathrm{m}}$, thì tính mô-đun tối đa sẽ không tương ứng với sự phân vùng đúng. Thay vào đó, tính mô-đun tối đa sẽ có xu hướng kết hợp các cộng đồng thành các nhóm lớn hơn và không thể giải quyết các phân vùng nhỏ nhất trong mạng. Ví dụ như kết quả bộ Twitter, các cộng đồng phát hiện được của thuật toán Greedy Modularity có sự đậm đặc cao ở một số cộng đồng đầu và phần lớn có nhiều cộng đồng bị cô lập gây ra hiệu ứng Long- Tail. Điều này cho thấy thuật toán Louvain có nhiều ưu điểm hơn.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết quả của đề tài

❖ Nhóm thực hiện được đủ các mục tiêu đã đề ra của đề tài, gồm:

(1) Tìm hiểu các khái niệm liên quan bài toán tối ưu trong phân tích dữ liệu MXH và tổ chức cấu trúc dữ liệu MXH.

(2) Nghiên cứu một số thuật toán phát hiện cộng đồng trong dữ liệu MXH.

(3) Xây dựng mô hình để chuyển đổi dữ liệu MXH dạng đồ thị thông thường sang dạng dữ liệu đồ thị lưỡng cực sau khi phát hiện các cộng đồng.

(4) Xây dựng được chương trình thực nghiệm để thực hiện chuyển đổi dữ liệu MXH từ đồ thị thông thường sang đồ thị lưỡng cực.

(5) Nhóm có tham gia viết 1 bài báo khoa học gởi tạp chí Khoa học Đại học Công Thương TP.HCM

Tạp chí Khoa học Đại học Công Thương

NGHIÊN CỨU BÀI TOÁN TỐI ƯU HÓA HÀM DR-SUBMODULAR TRÊN MẠNG LƯỚI NGUYÊN DƯƠNG ỨNG DỤNG CHO TỐI ĐA HÓA ẢNH HƯỞNG CỦA LAN TRUYỀN TIẾP THỊ TRÊN CÁC CỘNG ĐỒNG MẠNG XÃ HỘI

Nguyễn Thị Bích Ngân[1], Nguyễn Trường Phát[1], Đỗ Thế Sang[1], Phạm Nguyễn Huy Phương[1,*]

[1]Khoa Công nghệ thông tin, Trường Đại học Công Thương Thành phố Hồ Chí Minh

Email: ngamtb@huit.edu.vn, 2001207090@hufi.edu.vn, 2001203004@hufi.edu.vn, phuongpnh@huit.edu.vn

Ngày nhận bài..... ; Ngày chấp nhận đăng: xxx, 2024

❖ Qua quá trình thực hiện đề tài, nhóm được tiếp cận những kiến thức mới liên quan các bài toán tối ưu hóa và phát hiện cộng đồng trong phân tích dữ liệu MXH. Đây là một hướng nghiên cứu có rất nhiều ứng dụng trong thực tế, khi mà internet và MXH có ảnh hưởng sâu rộng đến đời sống con người.

❖ Kết quả nghiên cứu của nhóm góp một vài trò xử lý dữ liệu trong quá trình giải bài toán *tối đa hóa tầm ảnh hưởng của việc lan truyền tiếp thị trên các cộng đồng của mạng xã hội.*

Ngoài ra, phát hiện cộng đồng xây dựng dữ liệu lưỡng cực còn được sử dụng để xác định phân khúc thị trường, phát hiện tội phạm, hệ thống khuyến nghị và nhiều lĩnh vực khác nữa.

2. **Tính mới và sáng tạo của đề tài:**
   - **Tính mới:** Tạo ra mô hình chuyển đổi dữ liệu mạng xã hội từ dạng đồ thị thông thường sang dạng lưỡng cực. Vì sau khi khảo sát các nghiên cứu liên quan, hiện chưa có công bố nào liên quan các bộ dữ liệu dạng lưỡng cực đối với mạng xã hội. Trong khi dữ liệu lưỡng cực thì có nhiều ứng dụng trong các bài toán tối ưu hóa đa mục tiêu.
   - **Tính sáng tạo:** chuyển đổi và tạo cấu trúc của dữ liệu lưỡng cực đáp ứng cho nhu cầu bài toán tối đa hóa tầm ảnh hưởng trong mạng xã hội.
3. **Phần kiến nghị**

Một trong những khó khăn nhất khi chúng tôi thực hiện đề tài này là hệ thống máy có cấu hình mạnh để chạy thực nghiệm trên các bộ dữ liệu lớn và tiếp cận các công bố khoa học mới nhất.

Vì hầu hết các bộ dữ liệu của MXH là rất lớn (dạng big data), và các nghiên cứu cũng như thực nghiệm muốn mang lại giá trị cao thì cần thực hiện trên những bộ dữ liệu lớn này. Bên cạnh đó, để tiếp cận các kiến thức mới thì phải đọc được những bài báo khoa học liên quan nhưng vì chúng tôi không có điều kiện để mua các tài khoản mà truy cập dữ liệu báo khoa học ở nước ngoài.

Do đó chúng tôi kiến nghị Khoa và Trường tạo điều kiện không chỉ cho nhóm chúng tôi mà nhiều nhóm nghiên cứu khác thuộc ngành Công nghệ thông tin nói riêng và các ngành khác nói chung:

- Xây dựng các phòng lab có máy cấu hình mạnh để các nhóm nghiên cứu thực nghiệm.
- Mua các tài khoản truy cập dữ liệu báo khoa học nước ngoài để các nhóm nghiên cứu tiếp cận dễ dàng các dữ liệu này.

4. **Hướng phát triển của đề tài**

Hướng phát triển của đề tài là:

- Tối ưu hóa các thuật toán phát hiện cộng đồng này về độ phức tạp về thời gian, hiệu suất và phạm vi bao phủ của các phân vùng được phát hiện.
- Sử dụng kết quả đầu ra của biểu đồ lưỡng cực có trọng số làm bộ dữ liệu để tối đa hóa hàm mô đun phụ lợi nhuận giảm dần (DR-submodular) trên mạng số nguyên.

PHỤ LỤC

1. Chứng minh công thức (6).

Công thức ban đầu là:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - P_{ij}\right)\delta(g_i, g_j)$$

Trong đó:

-A_{ij} là phần tử tương ứng trong ma trận kề, chỉ ra có cạnh nối giữa nút i và j.

- P_{ij} là xác suất kết nối giữa nút i và j trong mạng ngẫu nhiên.

- g_i và g_j là nhãn của các nhóm mà nút i và j thuộc về.

- $\delta\left(g_i, g_j\right)$ là delta Kronecker, bằng 1 nếu $g_i = g_j$ và bằng 0 nếu $g_i \neq g_j$.

- m là tổng số cạnh trong mạng.

Đầu tiên, chúng ta thay thế P_{ij} bằng công thức (3):

$$P_{ij} = \frac{k_i k_j}{2m}$$

trong đó k_i là tổng số cạnh kết nối với nút i.

Tiếp theo, chúng ta thay thế P_{ij} vào công thức (1):

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

Chúng ta có thể tách biệt phần tử $k_i\ k_j$ ra khỏi tổng:

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

Chúng ta biểu diễn vế đầu đầu tiên dưới dạng tổng của số cạnh nằm trong các nhóm:

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

$$= Q = \frac{1}{2m}\sum_{ij} A_{ij}\delta(g_i, g_j) - \frac{1}{4m^2}\sum_{ij} k_i k_j\,\delta\left(g_i, g_j\right)$$

$$= Q = \frac{1}{2m}\sum_{ij} A_{ij}\delta(g_i, g_j) - \frac{1}{4m^2}\sum_{c}\sum_{ij\in c} k_i k_j$$

Trong đó, $\sum_{ij\in c} k_i k_j$ là tổng của các cặp nút i và j thuộc cùng một nhóm c.

Cuối cùng, chúng ta thấy $\sum_{ij\in c} k_i k_j$ chính là k_c^2, tổng bậc của tất cả các nút trong nhóm c. Do đó:

$$Q = \frac{1}{2m}\sum_{ij} A_{ij}\delta(g_i, g_j) - \frac{1}{4m^2}\sum_{c} k_c^2$$

Kết hợp cả hai vế, chúng ta có công thức cuối cùng:

$$Q = \sum_{c=1}^{n}(\frac{L_c}{m} - \frac{\gamma k_c^2}{2m^2})$$

$$= \sum_{c=1}^{n}\left[\frac{L_c}{m} - \gamma\left(\frac{k_c}{2m}\right)^2\right]$$

với $L_c = \sum_{i,j\in c} A_{ij}$ là số lượng cạnh nằm trong nhóm c, k_c là tổng số cạnh kết nối với tất cả các nút trong nhóm c, và γ là tham số độ đậm đặc của phân vùng. ■

TÀI LIỆU THAM KHẢO

[1]. Bovet, A. and Makse, H.A., 2019. Influence of fake news in Twitter during the 2016 US presidential election. Nature communications, 10(1), p.7.

[2]. Banerjee, S., Jenamani, M. and Pratihar, D.K., 2020. A survey on influence maximization in a social network. Knowledge and Information Systems, 62, pp.3417-3455.

[3]. Amiri, B., Fathian, M. and Asaadi, E., 2021. Influence maximization in complex social networks based on community structure. Journal of Industrial and Systems Engineering, 13(3), pp.16-40.

[4]. Javed, M. A., Younis, M. S., Latif, S., Qadir, J., & Baig, A. (2018). Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, *108*, 87-111.

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

[6]. Barnes, H. L., & Olson, D. H. (1982). Parent-adolescent communication scale. In D. H. Olson (Ed.), Family inventories: Inventories used in a national survey of families across the family life cycle (pp. 33-48). Family Social Science, University of Minnesota.

[7]. Kernighan, B.W. and Lin, S., 1970. An efficient heuristic procedure for partitioning graphs. The Bell system technical journal, 49(2), pp.291-307.

[8]. Malliaros, F.D. and Vazirgiannis, M., 2013. Clustering and community detection in directed networks: A survey. Physics reports, 533(4), pp.95-142.

[9]. Year. J. MacQueen, "Some methods for classification and analysis of multivariate observations", Proc. 5th Berkeley Symp. Math. Statist Prob.1, pp. 281-297, 1967.

[10]. N. R. Pal and J. C. Bezdek, "On Cluster Validity for Fuzzy c-Means Model," IEEE Trans. Fuzzy Syst., Vol.1, pp. 370-379, 1995.

[11]. Fiedler, M., 1989. Laplacian of graphs and algebraic connectivity. Banach Center Publications, 1(25), pp.57-70.

[12]. Donath, W.E. and Hoffman, A.J., 1973. Lower bounds for the partitioning of graphs. IBM Journal of Research and Development, 17(5), pp.420-425.

[13]. Liu, X. and Murata, T., 2010. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. Physica A: Statistical Mechanics and its Applications, 389(7), pp.1493-1500.

[14]. Newman, M.E. and Girvan, M., 2003. Mixing patterns and community structure in networks. In Statistical mechanics of complex networks (pp. 66-87). Berlin, Heidelberg: Springer Berlin Heidelberg.

[15]. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D., 2004. Defining and identifying communities in networks. Proceedings of the national academy of sciences, 101(9), pp.2658-2663.

[16]. Zhang, X.S., Wang, R.S., Wang, Y., Wang, J., Qiu, Y., Wang, L. and Chen, L., 2009. Modularity optimization in community detection of complex networks. Europhysics Letters, 87(3), p.38002.

[17]. Clauset, A., Newman, M.E. and Moore, C., 2004. Finding community structure in very large networks. *Physical review E*, *70*(6), p.066111.

[18]. Blondel, V.D. et al. Fast unfolding of communities in large networks. J. Stat. Mech 10008, 1-12(2008). https://doi.org/10.1088/1742-5468/2008/10/P10008

[19. Li, Z. and Liu, J., 2016. A multi-agent genetic algorithm for community detection in complex networks. Physica A: Statistical Mechanics and its Applications, 449, pp.336-347.

[20]. Mirsaleh, M.R. and Meybodi, M.R., 2016. A Michigan memetic algorithm for solving the community detection problem in complex network. Neurocomputing, 214, pp.535-545.

[21]. Zhang, Q. and Li, H., 2007. MOEA/D: A multiobjective evolutionary algorithm based on decomposition. IEEE Transactions on evolutionary computation, 11(6), pp.712-731.

[22]. Deng, K., Zhang, J.P. and Yang, J., 2015. AN EFFICIENT MULTI-OBJECTIVE COMMUNITY DETECTION ALGORITHM IN COMPLEX NETWORKS. Tehnicki vjesnik/Technical Gazette, 22(2).

[23]. Derényi, I., Palla, G. and Vicsek, T., 2005. Clique percolation in random networks. Physical review letters, 94(16), p.160202.

[24]. Macropol, K. and Singh, A., 2010. Scalable discovery of best clusters on large graphs. Proceedings of the VLDB Endowment, 3(1-2), pp.693-702.

[25]. Raghavan, U.N., Albert, R. and Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Physical review E, 76(3), p.036106.

[26]. Tang, L., Liu, X., Zhang, J., & Wang, X. (2020). A novel framework for dynamic community detection in evolving bipartite networks. IEEE Transactions on Knowledge and Data Engineering, 32(6), 1222-1235.

[27]. Li, J., Zhu, A., Wang, H., & Zhang, J. (2019). Label propagation for attributed bipartite graph community detection. IEEE Transactions on Cybernetics, 49(11), 4340-4351.

[28]. Wang, H., Zhang, J., Wang, Z., & Zhang, F. (2020). GNN-BIPO: Graph neural network for bipartite link prediction. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2285)

[29]. Liu, Bin; Chen, Zihan; Du, Hongmin W. Streaming Algorithms for Maximizing DRSubmodular Functions with d-Knapsack Constraints. In: Algorithmic Aspects in Information and Management - 15th International Conference, AAIM. Springer, 2021,vol. 13153, pp. 159–169.

[30]. Zhang, Zhenning; Guo, Longkun; Wang, Yishui; Xu, Dachuan; Zhang, Dongmei. - Streaming Algorithms for Maximizing Monotone DR-Submodular Functions with a Cardinality Constraint on the Integer Lattice. Asia Pac. J. Oper. Res. (2021), vol. 38, no. 5, 2140004:1–2140004:14.

[31]. Gong, Suning; Nong, Qingqin; Bao, Shuyu; Fang, Qizhi; Du, Ding-Zhu - A fast and deterministic algorithm for Knapsack-constrained monotone DR-submodular maximization over an integer lattice. Journal of Global Optimization. (2022), 1–24

[32]. Leskovec, J., & Sosič, R. - SNAP: A General-Purpose Network Analysis and Graph-Mining Library. ACM Transactions on Intelligent Systems and Technology (TIST), 8(1), (2016),1.

[33]. Skiena, S. "Coloring Bipartite Graphs." §5.5.2 in Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. Reading, MA: Addison-Wesley, p. 213, 1990.

[34]. Duch, J. and Arenas, A., 2005. Community detection in complex networks using extremal optimization. Physical review E, 72(2), p.027104.

[35]. Li, S., Chen, Y., Du, H. and Feldman, M.W., 2010. A genetic algorithm with local search strategy for improved detection of community structure. Complexity, 15(4), pp.53-60.

[36]. Arenas, A., Fernandez, A. and Gomez, S., 2008. Analysis of the structure of complex networks at different resolution levels. New journal of physics, 10(5), p.053039.