

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



NGUYỄN THỊ NGỌC HUYỀN

**TỐI ĐA HÓA ẢNH HƯỞNG TRÊN MẠNG ĐA KÊNH KẾT HỢP GIỮA
MÔ HÌNH SINH VÀ HỌC TĂNG CƯỜNG**

Chuyên ngành: Hệ thống thông tin

Mã số: 8.48.01.04

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(HỆ THỐNG THÔNG TIN)**

(Theo định hướng nghiên cứu)

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. PHẠM VĂN CƯỜNG

HÀ NỘI - NĂM 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



NGUYỄN THỊ NGỌC HUYỀN

**TỐI ĐA HÓA ẢNH HƯỞNG TRÊN MẠNG ĐA KÊNH KẾT HỢP GIỮA
MÔ HÌNH SINH VÀ HỌC TĂNG CƯỜNG**

**Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04**

**ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(HỆ THỐNG THÔNG TIN)
(Theo định hướng nghiên cứu)**

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. PHẠM VĂN CƯỜNG

HÀ NỘI - NĂM 2025

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong đề án tốt nghiệp là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Hà Nội, tháng 05 năm 2025

Tác giả đề án



Nguyễn Thị Ngọc Huyền

LỜI CẢM ƠN

Lời đầu tiên, em xin trân trọng gửi lời cảm ơn sâu sắc đến các thầy, cô giáo thuộc Khoa Công nghệ thông tin 1, Học viện Công nghệ Bưu Chính Viễn thông. Trong suốt hai năm học tập và rèn luyện dưới mái trường, các thầy cô không chỉ truyền đạt những kiến thức chuyên môn nền tảng mà còn chia sẻ những kinh nghiệm thực tiễn vô cùng quý báu, tạo dựng hành trang vững chắc để em tự tin bước vào tương lai.

Em xin bày tỏ lòng biết ơn chân thành và sâu sắc nhất tới PGS.TS Phạm Văn Cường, người thầy đã trực tiếp hướng dẫn, dìu dắt em trong suốt quá trình thực hiện đề án tốt nghiệp này. Sự tận tâm, nhiệt tình chỉ bảo, những góp ý khoa học sắc bén và sự động viên kịp thời của thầy đã giúp em vượt qua nhiều khó khăn, định hướng đúng đắn và hoàn thiện nghiên cứu của mình. Thầy không chỉ là người hướng dẫn về mặt học thuật mà còn là tấm gương về tinh thần làm việc nghiêm túc và say mê nghiên cứu.

Em cũng xin gửi lời cảm ơn đặc biệt tới anh Đỗ Hoàng Khôi Nguyên. Anh đã luôn sẵn lòng chia sẻ những kinh nghiệm quý giá, đưa ra những lời khuyên thực tế và hỗ trợ em về mặt kỹ thuật cũng như tinh thần trong những giai đoạn thử thách nhất của đề án. Sự giúp đỡ nhiệt tình của anh là nguồn động viên lớn lao đối với em.

Em cũng xin gửi lời cảm ơn đến các anh chị, các bạn và các em trong lớp M21CQIS01-B đã tạo ra một môi trường học tập chuyên nghiệp, sáng tạo. Cuối cùng, em xin cảm ơn gia đình, đồng nghiệp và bạn bè, những người đã luôn ở cạnh, quan tâm, chia sẻ và tạo điều kiện tốt nhất để em có thể hoàn thành đề án này.

Hà Nội, tháng 05 năm 2025

Tác giả đề án

Nguyễn Thị Ngọc Huyền

MỤC LỤC

LỜI CAM ĐOAN	i
LỜI CẢM ƠN	ii
MỤC LỤC	iii
DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT	vi
DANH SÁCH BẢNG.....	viii
DANH SÁCH HÌNH VẼ	ix
MỞ ĐẦU.....	1
1. Tính cấp thiết của đề tài	1
2. Tổng quan về vấn đề nghiên cứu	3
3. Mục đích nghiên cứu.....	5
4. Đối tượng và phạm vi nghiên cứu.....	5
5. Phương pháp nghiên cứu.....	5
CHƯƠNG 1 - GIỚI THIỆU TỔNG QUAN VỀ CÁC BÀI TOÁN LAN TRUYỀN THÔNG TIN	8
1.1. Giới thiệu về mạng xã hội	8
1.1.1. Những đặc điểm chung của mạng xã hội	8
1.1.2. Lợi ích của mạng xã hội.....	9
1.1.3. Tác hại của mạng xã hội.....	11
1.2. Mô hình hóa lan truyền thông tin trên mạng xã hội	13
1.2.1. Mô hình Ngưỡng tuyến tính (LT)	13
1.2.2. Mô hình bậc độc lập (IC)	15
1.2.3. Mô hình ngăn trong dịch tễ học (SIS).....	15
1.3. Một số bài toán lan truyền thông tin trên mạng xã hội.....	16

1.3.1. Tối đa hóa ảnh hưởng (IM)	16
1.3.2. Ngăn chặn ảnh hưởng (IB).....	17
1.4. Kết luận chương	18
CHƯƠNG 2 - TỐI ĐA HÓA ẢNH HƯỞNG ĐA TẦNG VỚI CÁC RÀNG BUỘC VỀ NGÂN SÁCH TRÊN MẠNG XÃ HỘI.....	19
2.1. Đặt vấn đề và phát biểu bài toán tối đa hóa ảnh hưởng đa tầng trên mạng xã hội.....	19
2.1.1. Đặt vấn đề.....	19
2.1.2. Phát biểu bài toán.....	27
2.2. Khung mô hình REM cho bài toán tối đa hóa ảnh hưởng trên mạng đa tầng	30
2.2.1. Mô hình học biểu diễn vector cho tập hạt giống phức tạp	32
2.2.2. Khám phá tập hạt giống tiềm ẩn.....	35
2.2.3. Lan truyền hỗn hợp các chuyên gia.....	37
2.3. Kết luận chương	41
CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ	43
3.1. Thiết lập thực nghiệm	43
3.1.1. Dữ liệu và tham số.....	43
3.1.2. Phương pháp so sánh	45
3.2. Kết quả và đánh giá.....	46
3.2.1. Khả năng lan truyền thông tin	46
3.2.2. So sánh thời gian chạy.....	55
3.2.3. Ảnh hưởng của số lần khám phá môi trường	56
3.2.4. Kết quả các biến thể khác của mô hình	58

3.3. Kết luận chương	59
KẾT LUẬN	60
CÔNG BỐ KHOA HỌC	62
TÀI LIỆU THAM KHẢO	63

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
IM	Influence Maximization	Tối đa hóa Ảnh hưởng
OSNs	Online Social Networks	Mạng Xã hội Trực tuyến
MIM	Multiplex Influence Maximization	Tối đa hóa Ảnh hưởng Đa tầng
CO	Combinatorial Optimization	Tối ưu hóa Tổ hợp
RL	Reinforcement Learning	Học tăng cường
GNN	Graph Neural Network	Mạng nơ-ron đồ thị
VAE	Variational Autoencoder	Bộ tự mã hóa biến phân
MoE	Mixture of Experts	Hỗn hợp Chuyên gia
PMoE	Propagation Mixture of Experts	Hỗn hợp Chuyên gia Lan truyền
LT	Linear Threshold	Ngưỡng tuyến tính
IC	Independent Cascade	Lan truyền độc lập / Bậc độc lập
SIS	Susceptible-Infectious-Susceptible	Mô hình ngăn trong dịch tễ học
IB	Influence Blocking	Ngăn chặn ảnh hưởng
MXH		Mạng xã hội
CELF	Cost-Effective Lazy Forward	Đánh giá lười hiệu quả về chi phí
PRM	Priority Replay Memory	Priority Replay Memory
ELBO	Evidence Lower Bound	Giới hạn Dưới của Bằng chứng
KL	Kullback–Leibler (Divergence)	Độ phân kỳ KL
MSE	Mean Squared Error	Tổn thất Bình phương Trung bình

GCN	Graph Convolutional Network	Mạng Tích chập Đồ thị
GAT	Graph Attention Network	Mạng Chú ý Đồ thị
ISF	Influential Seed Finder	Công cụ tìm hạt giống có ảnh hưởng
KSN	Knapsack Seeding of Networks	Gieo hạt mạng dựa trên bài toán ba lô

DANH SÁCH BẢNG

Bảng 3. 1: Giá trị các siêu tham số của mô hình.....	44
Bảng 3. 2: Thống kê các bộ dữ liệu sử dụng.....	45
Bảng 3. 3: So sánh hiệu suất dưới mô hình lan truyền IC (Cora-ML).....	47
Bảng 3. 4: So sánh hiệu suất dưới mô hình lan truyền IC (Celegans).	48
Bảng 3. 5: So sánh hiệu suất dưới mô hình lan truyền IC (Arabidopsis).....	48
Bảng 3. 6: So sánh hiệu suất dưới mô hình lan truyền IC. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (NYClimateMarch2014)	49
Bảng 3. 7: So sánh hiệu suất dưới mô hình lan truyền IC. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (ParisAttack2015)	50
Bảng 3. 8: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (Coral-ML).....	52
Bảng 3. 9: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (Celegans)	52
Bảng 3. 10: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (Arabidopsis).....	53
Bảng 3. 11: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (NYClimateMarch2014)	54
Bảng 3. 12: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (ParisAttack2015)	54
Bảng 3. 13: Thời gian suy luận trung bình (tính bằng giây) theo sự gia tăng số lượng nút với 10% số nút làm các nút khởi đầu.	55
Bảng 3. 14: So sánh giữa việc chọn GCN và GAT làm kiến trúc chuyên gia trong PMoE, với 10% số nút làm ngân sách, dưới cả hai mô hình LT và IC.....	58

DANH SÁCH HÌNH VẼ

Hình 1. 1: Ví dụ về mô hình Ngưỡng Tuyến tính (LT).....	14
Hình 1. 2: Minh họa mô hình bậc độc lập (IC) trong quá trình lan truyền ảnh hưởng trên mạng lưới	15
 Hình 2. 1: Ví dụ về Tối đa hóa Ảnh hưởng: Chọn tập hạt giống S (vòng đỏ) để tối đa hóa sự lan tỏa thông tin (vùng $\sigma(S)$ - vòng vàng) cho một chiến dịch.....	20
Hình 2. 2: Mô hình tương tác giữa Tác nhân (Agent) và Môi trường (Environment) trong Học Tăng Cường (RL).....	25
Hình 2. 3: Một ví dụ minh họa tính chất "kích hoạt trùng lặp" độc đáo của sự lan truyền ảnh hưởng trong một mạng đa lớp.	28
Hình 2. 4: Sơ đồ mô tả quy trình của REM trong việc giải quyết bài toán MIM. ...	30
 Hình 3. 1: Sự khác biệt trong mức độ lan truyền ảnh hưởng (trục y) của đầu ra REM trên các tập dữ liệu và ngân sách khác nhau khi tăng số bước khám phá (trục x). Hình 3a - 3e và Hình 3f - 3j được đánh giá dưới mô hình IC và LT, tương ứng.....	57

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Dữ liệu đồ thị đã tìm thấy nhiều ứng dụng trong các lĩnh vực như mạng xã hội và khai thác dữ liệu [23] [54] [14] [17]. Một ứng dụng phổ biến là Tối đa hóa Ảnh hưởng (Influence Maximization - IM), nhằm xác định một tập hợp các cá nhân có thể tối đa hóa sự lan truyền ảnh hưởng trong một mạng xã hội dưới một mô hình khuếch tán cụ thể. Vấn đề này được biết đến là NP-hard và đã được nghiên cứu rộng rãi trong các lĩnh vực như marketing lan truyền (viral marketing) [35] [13]. Với sự đa dạng hóa của các nền tảng xã hội, nhiều người dùng trên các Mạng Xã hội Trực tuyến (Online Social Networks - OSNs) như Facebook và Twitter đang liên kết tài khoản của họ qua nhiều nền tảng. Những OSNs này được kết nối với nhau với các người dùng chung, được gọi là Mạng Đa Tầng (Multiplex Networks). Cấu trúc của các mạng đa tầng cho phép người dùng đăng thông tin trên nhiều OSNs đồng thời, mang lại giá trị đáng kể cho các chiến dịch marketing [36] [20] [29]. Các mô hình lan truyền thông tin bên trong mỗi mạng xã hội trực tuyến (OSN) có thể khác nhau, dẫn đến sự khác biệt trong cách thông tin lan truyền và ảnh hưởng người dùng trên các nền tảng. Do đó, việc tùy chỉnh các chiến lược tối đa hóa ảnh hưởng sao cho hiệu quả trên nhiều nền tảng là rất quan trọng. Điều này được gọi là Tối đa hóa Ảnh hưởng Đa tầng (Multiplex Influence Maximization - MIM).

Cho đến nay, các thuật toán Tối ưu hóa Tổ hợp (Combinatorial Optimization - CO) cho MIM [39] [13] [34] [43] [7] còn nhiều hạn chế so với các phương pháp dựa trên học máy. Các thuật toán CO gặp khó khăn trong việc tổng quát hóa đối với các đồ thị chưa thấy trước và xử lý các mạng đa tầng phức tạp. Chúng cũng gặp vấn đề về khả năng mở rộng khi làm việc với các mạng lớn. Hơn nữa, các thuật toán CO phụ thuộc vào các quy tắc hoặc thủ thuật định sẵn, làm hạn chế khả năng nắm bắt các mẫu phức tạp và các mối quan hệ phi tuyến tính trong các mạng đa tầng. Những hạn chế này làm giảm đáng kể hiệu quả của chúng trong việc tối ưu hóa việc lựa chọn các nút có ảnh hưởng. Ngược lại, các phương pháp dựa trên học máy [34] [53] [45] [43]

[37] mang lại lợi thế về khả năng tổng quát hóa, khả năng mở rộng và nắm bắt các mẫu phức tạp. Tuy nhiên, chúng vẫn gặp phải một số yếu điểm quan trọng trong MIM như sau:

a. *Tối ưu hóa không hiệu quả.* Vì MIM là một vấn đề NP-hard và mỗi lớp có thể mở rộng lên hàng tỷ, nên việc huấn luyện hiệu quả là rất cần thiết. Các phương pháp học tăng cường (RL), như những phương pháp được đề xuất trong [41] [45] [53] [34], tối ưu hóa tập hợp các nút hạt giống trong các không gian rời rạc, nhiều bằng cách tận dụng việc khám phá, cải tiến giải pháp từng bước mà không cần đầu vào ban đầu. Tuy nhiên, những phương pháp này phụ thuộc vào việc lấy mẫu ngẫu nhiên ở quy mô lớn, thường dẫn đến thời gian huấn luyện đáng kể trước khi xác định được giải pháp phù hợp và có nguy cơ bị mắc kẹt trong các tối ưu cục bộ. Để khắc phục những hạn chế này, các phương pháp dựa trên dữ liệu [7] đã được giới thiệu. Mô hình sinh này học từ các tập dữ liệu đa dạng để tạo ra các giải pháp hiệu quả. Tuy nhiên, sự thành công của nó phụ thuộc nhiều vào chất lượng của dữ liệu huấn luyện vì chỉ khi tập dữ liệu đủ đa dạng, mô hình mới có thể nắm bắt được các đặc điểm chính để tối ưu hóa. Nhìn chung, việc tạo ra một mô hình có thể tối ưu hóa các giải pháp của nó với độ phức tạp thấp vẫn là một thách thức lớn.

b. *Các mô hình ước lượng lan truyền không chính xác.* Đo lường giá trị lan truyền của một tập hợp các nút hạt giống nhất định là rất quan trọng vì nó giúp chúng ta xác định được hiệu quả của tập hợp nút hiện tại. Các phương pháp như trong [41] [34] [53] tiêu tốn thời gian đáng kể trong các mô phỏng lan truyền, trong khi các phương pháp khác như [45] [7] xây dựng các mô hình dựa trên mạng nơ-ron đồ thị (GNN) để dự đoán tổng lan truyền khi cho một tập hợp hạt giống. Các phương pháp trước đây gặp khó khăn khi mở rộng với các đồ thị lớn, trong khi các phương pháp sau gặp phải vấn đề về độ chính xác do hiện tượng làm mịn quá mức (oversmoothing) [9] trong các mô hình GNN. Thách thức này càng trở nên phức tạp hơn khi mỗi lớp trong một mạng đa tầng có thể có một mô hình lan truyền cơ bản khác nhau và có thể mở rộng đến hàng tỷ nút, làm cho việc dự đoán chính xác trở nên khó khăn hơn.

Chính vì vậy, đề án này tập trung vào việc nghiên cứu và xây dựng một mô hình học sâu tiên tiến nhằm giải quyết hiệu quả bài toán Tối đa hóa Ảnh hưởng Đa tầng. Mô hình được đề xuất hướng tới việc khắc phục các nhược điểm về tối ưu hóa và ước lượng lan truyền đã nêu, thông qua việc kết hợp các kỹ thuật học biểu diễn, học tăng cường và kiến trúc mạng nơ-ron đồ thị chuyên biệt cho mạng đa tầng. Mục tiêu là tạo ra một giải pháp có khả năng xác định tập hạt giống tối ưu một cách hiệu quả và chính xác hơn, ngay cả trên các mạng xã hội đa tầng quy mô lớn và phức tạp.

Để trình bày nội dung một cách hệ thống và khoa học, đề án sẽ được cấu trúc thành các chương chính như sau:

- **Chương 1: Tổng quan về bài toán tối đa hóa ảnh hưởng trên mạng xã hội.** Chương này sẽ giới thiệu các khái niệm cơ bản về mạng xã hội, các mô hình lan truyền thông tin phổ biến, và đi sâu vào bài toán tối đa hóa ảnh hưởng (IM) trên mạng xã hội đơn lớp. Các nghiên cứu liên quan cũng sẽ được tổng hợp trong chương này.
- **Chương 2: Phương pháp đề xuất: Tối đa hóa ảnh hưởng trên mạng xã hội đa lớp.** Chương này trình bày chi tiết phương pháp được đề xuất để giải quyết bài toán tối đa hóa ảnh hưởng trên mạng đa tầng (MIM).
- **Chương 3: Thực nghiệm và đánh giá.** Chương này mô tả quá trình thiết lập thực nghiệm để đánh giá hiệu quả của mô hình REM đã đề xuất.

2. Tổng quan về vấn đề nghiên cứu

Tối đa tổ hợp cho bài toán IM. Bài toán Tối đa Ảnh hưởng (Influence Maximization - IM) thực chất là một trường hợp đơn giản hóa của Tối đa Ảnh hưởng trên Mạng Đa Lớp (Multiplex Influence Maximization - MIM), khi chỉ giới hạn trong một mạng đơn thay vì nhiều mạng liên kết chặt chẽ. Mặc dù bài toán IM truyền thống đã đạt được nhiều tiến bộ đáng kể, nhưng MIM lại đặt ra các thách thức đặc biệt do sự tương tác phức tạp giữa các lớp mạng liên kết. Các phương pháp tiếp cận IM ban đầu chủ yếu dựa trên mô phỏng [21], thực hiện quá trình khuếch tán ảnh hưởng trên

mạng lưới nhiều lần để ước lượng phạm vi ảnh hưởng. Dù trực quan, phương pháp này thường đòi hỏi tài nguyên tính toán lớn, đặc biệt với các mạng lưới lớn. Nhằm giải quyết vấn đề này, các phương pháp dựa trên proxy [31] [46] [47] đã ra đời, sử dụng các chỉ số đơn giản hơn để xấp xỉ phạm vi ảnh hưởng. Dựa trên tính chất con-mô-đun (submodularity) của quá trình khuếch tán, các thuật toán xấp xỉ như CELF++ [2] và UBLF [11] giúp chọn tập hạt giống hiệu quả với tỉ lệ xấp xỉ đảm bảo $(1 - 1/e)$. Gần đây, thuật toán Tiptop [50] đã tạo nên bước ngoặt khi cung cấp lời giải gần chính xác cho bài toán IM với tỉ lệ tối ưu $(1 - \epsilon)$ cho bất kỳ $\epsilon > 0$. Tuy nhiên, bài toán MIM lại yêu cầu các hướng tiếp cận hoàn toàn mới do sự phức tạp bổ sung từ nhiều mạng lưới liên kết. Dù đã xuất hiện các phương pháp đầy hứa hẹn sử dụng thuật toán xấp xỉ tổ hợp [19] [3], MIM vẫn là một lĩnh vực nghiên cứu sôi động. Các hướng nghiên cứu trong tương lai bao gồm việc tích hợp học máy và khai thác các đặc điểm riêng của mạng đa lớp để đạt được các giải pháp hiệu quả và chính xác hơn.

Ứng dụng Học Máy vào bài toán IM. Các phương pháp học máy, đặc biệt là học sâu (deep learning), đã xuất hiện để vượt qua những hạn chế của các phương pháp IM truyền thống, vốn thiếu khả năng tổng quát hóa. Việc kết hợp học tăng cường (Reinforcement Learning - RL) với IM đã cho thấy tiềm năng lớn [40] [26], với các tiến bộ gần đây tập trung vào việc học biểu diễn ẩn (latent embedding) của các nút hoặc mạng để lựa chọn các nút hạt giống [41] [45] [18]. Ngoài ra, các mạng nơ-ron đồ thị (Graph Neural Networks - GNNs) cũng được nghiên cứu để mã hóa ảnh hưởng xã hội và hướng dẫn việc lựa chọn nút trong IM [6] [10]. Đặc biệt, mô hình DeepIM [53] sử dụng cách tiếp cận tạo sinh cho bài toán IM, đạt được hiệu suất hàng đầu hiện nay. Tuy nhiên, các phương pháp này chủ yếu được thiết kế cho bài toán IM trên mạng đơn và gặp khó khăn khi mở rộng sang bài toán MIM phức tạp hơn. Nguyên nhân chính đến từ các hạn chế về khả năng xử lý quy mô lớn với nhiều mạng, cũng như khó khăn trong việc nắm bắt hiệu quả các mối quan hệ khuếch tán phức tạp giữa các lớp mạng (intra-propagation và inter-propagation). Bên cạnh đó, các phương pháp này thường bị giới hạn bởi sự phụ thuộc vào mẫu quan sát. Một số công trình gần đây [34] đã kết hợp mô hình đồ thị xác suất để biểu diễn quá trình khuếch tán ảnh hưởng

trong mạng đa chiều, cùng với học tăng cường để tìm tập hạt giống tối ưu. Tuy vậy, những phương pháp này vẫn đối mặt với vấn đề về khả năng mở rộng, do sự phụ thuộc vào các phương pháp truyền thống ở giai đoạn khởi đầu.

3. Mục đích nghiên cứu

Giới thiệu một khung làm việc mới mang tên Tối đa hóa Chuyên gia Tăng cường (REM) để giải quyết các thách thức trong Tối đa hóa Ảnh hưởng Đa tầng (MIM). REM sử dụng mô hình VAE (Variational autoencoders) để chuyển không gian đầu vào nhiều thành không gian tiềm ẩn liên tục (Seed2Vec), giúp tối ưu hóa tập hạt giống hiệu quả hơn. Để khắc phục hạn chế về dữ liệu huấn luyện, Seed2Vec được áp dụng như một chính sách Học Tăng cường (RL), cho phép khám phá và tạo ra các tập hạt giống tổng hợp có khả năng lan truyền cao. REM cũng đề xuất mô hình Hỗn hợp Chuyên gia Lan truyền (PMoE) dựa trên GNN để ước lượng chính xác các mẫu lan truyền phức tạp mà không tăng độ phức tạp của mô hình. Các đánh giá cho thấy REM vượt trội về hiệu suất lan truyền, khả năng mở rộng và thời gian suy luận so với các phương pháp tiên tiến khác.

4. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu: các mạng xã hội đa lớp
- Phạm vi nghiên cứu: tập trung vào việc tối ưu hóa việc lựa chọn các đỉnh ảnh hưởng trên mạng đa lớp để lan truyền thông tin hiệu quả trên các nền tảng xã hội.

5. Phương pháp nghiên cứu

Nghiên cứu này được tiến hành thông qua một phương pháp kết hợp, bao gồm giai đoạn nghiên cứu lý thuyết sâu rộng và giai đoạn nghiên cứu thực nghiệm để phát triển và đánh giá khung làm việc Tối đa hóa Chuyên gia Tăng cường (REM).

Ban đầu, công tác nghiên cứu lý thuyết sẽ tập trung vào việc tổng quan toàn diện các tài liệu khoa học hiện có. Quá trình này bao gồm việc tìm hiểu sâu về các khái niệm, thuật toán và thách thức cốt lõi trong lĩnh vực Tối đa hóa Ảnh hưởng (IM) nói chung và Tối đa hóa Ảnh hưởng Đa tầng (MIM) nói riêng. Cụ thể, đề tài sẽ phân

tích các mô hình lan truyền ảnh hưởng phổ biến (như Independent Cascade, Linear Threshold) và cách chúng được điều chỉnh cho mạng đa tầng. Các phương pháp Tối ưu hóa Tổ hợp (CO) truyền thống [39] [13] [34] [43] [7] cho MIM sẽ được xem xét kỹ lưỡng về thuật toán, ưu điểm và các hạn chế đã được ghi nhận như khả năng tổng quát hóa kém, vấn đề về khả năng mở rộng và sự phụ thuộc vào các quy tắc định sẵn. Song song đó, các phương pháp dựa trên học máy [34] [53] [45] [43] [37], đặc biệt là Học tăng cường (RL) và Mạng nơ-ron đồ thị (GNN), sẽ được nghiên cứu để hiểu rõ cơ chế hoạt động, lợi thế về khả năng tổng quát hóa và nắm bắt mẫu phức tạp, cũng như các nhược điểm còn tồn tại như tối ưu hóa không hiệu quả trong RL hoặc ước lượng lan truyền thiếu chính xác do hiện tượng làm mịn quá mức (oversmoothing) trong GNN. Các kỹ thuật liên quan như Variational Autoencoders (VAE) cho việc học biểu diễn trong không gian tiềm ẩn liên tục và Mixture of Experts (MoE) cho việc kết hợp nhiều mô hình chuyên biệt cũng sẽ được khảo sát. Từ nền tảng lý thuyết này, đề án sẽ phân tích và xác định rõ hơn các vấn đề mà REM cần giải quyết, bao gồm thách thức tối ưu hóa tập hạt giống hiệu quả trong không gian lớn, nhiễu của mạng đa tầng và việc ước lượng chính xác sự lan truyền khi có nhiều mô hình lan truyền khác nhau giữa các tầng mạng.

Để đánh giá hiệu quả của REM, một quy trình thực nghiệm chi tiết sẽ được thiết kế. Điều này bắt đầu bằng việc chuẩn bị các bộ dữ liệu mạng xã hội đa tầng công khai và đa dạng (ví dụ: Facebook-Twitter, DBLP), có thể bao gồm cả dữ liệu tổng hợp, để kiểm tra tính tổng quát. Chúng tôi sẽ lựa chọn các phương pháp so sánh (baseline methods) bao gồm cả thuật toán CO và các phương pháp học máy tiên tiến khác cho. Các độ đo đánh giá chính bao gồm hiệu suất lan truyền (số nút bị ảnh hưởng), thời gian huấn luyện, thời gian suy luận/chọn hạt giống, và khả năng mở rộng khi kích thước đồ thị và số tầng tăng. Môi trường thực nghiệm sẽ được thiết lập cẩn thận với việc phân chia dữ liệu (nếu cần), cấu hình và tinh chỉnh siêu tham số một cách công bằng. Sau khi thực hiện các thí nghiệm, kết quả sẽ được phân tích kỹ lưỡng thông qua so sánh định lượng với các baseline để đánh giá đóng góp của từng thành phần trong REM, và thảo luận về các phát hiện, ý nghĩa cũng như hạn chế của nghiên

cứu. Thông qua phương pháp tiếp cận toàn diện này, nghiên cứu hướng tới việc cung cấp một giải pháp mới, hiệu quả cho bài toán MIM, đồng thời làm sâu sắc thêm hiểu biết về việc ứng dụng học máy trong tối ưu hóa trên các cấu trúc đồ thị phức tạp.

CHƯƠNG 1 - GIỚI THIỆU TỔNG QUAN VỀ CÁC BÀI TOÁN LAN TRUYỀN THÔNG TIN

Trong kỷ nguyên số, mạng xã hội đã nổi lên như những hệ sinh thái phức tạp, nơi thông tin không chỉ được chia sẻ mà còn lan tỏa với tốc độ và quy mô chưa từng có, định hình mạnh mẽ dư luận và hành vi. Việc thấu hiểu và mô hình hóa những động lực lan truyền phức tạp này không chỉ là một mục tiêu khoa học hấp dẫn mà còn là chìa khóa cho vô số ứng dụng thực tiễn, từ chiến lược truyền thông đến an ninh thông tin. Do đó, chương này sẽ đặt nền móng bằng cách giới thiệu tổng quan về mạng xã hội, đi sâu vào các mô hình toán học kinh điển mô tả sự lan truyền, và phân tích những bài toán cốt lõi như tối đa hóa và ngăn chặn ảnh hưởng, tạo tiền đề cho các phân tích và giải pháp chuyên sâu được trình bày.

1.1. Giới thiệu về mạng xã hội

1.1.1. Những đặc điểm chung của mạng xã hội

Mạng xã hội là một tập hợp các nền tảng trực tuyến cho phép con người kết nối, giao tiếp và chia sẻ thông tin một cách dễ dàng, nhanh chóng, vượt qua mọi rào cản về không gian và thời gian. Các nền tảng nổi bật như Facebook, Twitter, Instagram, TikTok và LinkedIn đã trở thành một phần không thể thiếu trong cuộc sống hiện đại, tạo nên một không gian số hóa năng động và liên tục phát triển. Một trong những đặc điểm quan trọng nhất của mạng xã hội là khả năng kết nối và tương tác giữa các cá nhân, tổ chức, và cộng đồng. Người dùng có thể dễ dàng thiết lập liên lạc với bạn bè, người thân, đồng nghiệp hoặc những người có chung sở thích mà họ có thể chưa từng gặp ngoài đời thực. Các hình thức giao tiếp trên mạng xã hội rất phong phú, bao gồm tin nhắn trực tiếp, bình luận, bài đăng công khai, và thậm chí cả các cuộc gọi video hay phát trực tiếp, từ đó mang lại sự linh hoạt trong cách mọi người xây dựng và duy trì mối quan hệ. Mạng xã hội không chỉ đơn thuần là một công cụ liên lạc mà còn là nơi chia sẻ những câu chuyện, những trải nghiệm và cảm xúc một cách tự do. Việc truyền tải nội dung như văn bản, hình ảnh, video hay liên

kết thông qua các nền tảng này diễn ra nhanh chóng và dễ dàng, cho phép thông tin được lan tỏa tới hàng ngàn, thậm chí hàng triệu người chỉ trong vài giây. Đặc điểm này đã biến mạng xã hội trở thành một kênh truyền thông mạnh mẽ, không chỉ cho cá nhân mà còn cho các tổ chức, doanh nghiệp, và phong trào cộng đồng.

Bên cạnh đó, mạng xã hội còn là một môi trường tự do và tùy biến cao, nơi người dùng có thể xây dựng hình ảnh cá nhân và thể hiện bản thân một cách sáng tạo. Hồ sơ cá nhân, các bài đăng, và cách tương tác trên mạng xã hội đều có thể phản ánh rõ nét sở thích, quan điểm, và phong cách sống của mỗi người, tạo nên một "bản sắc số" độc đáo. Không dừng lại ở việc kết nối hay chia sẻ nội dung, các nền tảng mạng xã hội còn tích hợp nhiều tính năng hữu ích khác như thương mại điện tử, quảng cáo, giáo dục trực tuyến, hay quản lý công việc, từ đó mở rộng vai trò của chúng trong các hoạt động kinh tế và xã hội. Tính chất mở và khả năng truy cập dễ dàng của mạng xã hội cũng là một yếu tố quan trọng, khi bất kỳ ai có kết nối Internet đều có thể tham gia vào hệ sinh thái này. Điều này tạo nên một mạng lưới toàn cầu khổng lồ, nơi thông tin có thể lan truyền với tốc độ nhanh chóng và tác động mạnh mẽ đến đời sống xã hội. Tuy nhiên, chính những đặc điểm này cũng khiến mạng xã hội trở thành một công cụ hai mặt, vừa mang lại nhiều lợi ích đáng kể trong việc kết nối và thúc đẩy giao lưu thông tin, vừa tiềm ẩn những thách thức và rủi ro trong việc kiểm soát và quản lý nội dung.

1.1.2. Lợi ích của mạng xã hội

Mạng xã hội mang lại rất nhiều lợi ích vượt bậc, ảnh hưởng sâu rộng đến nhiều khía cạnh của cuộc sống cá nhân, xã hội và kinh tế. Trước hết, mạng xã hội đã làm thay đổi cách con người kết nối và giao tiếp với nhau. Nó giúp mọi người có thể dễ dàng duy trì mối quan hệ với bạn bè, gia đình và đồng nghiệp, dù ở bất kỳ đâu trên thế giới. Chỉ cần một thiết bị kết nối Internet, khoảng cách địa lý không còn là rào cản, người dùng có thể gọi điện, nhắn tin, hay thậm chí tham gia các cuộc trò chuyện video với những người thân yêu. Không chỉ dừng lại ở kết nối cá nhân, mạng xã hội còn giúp mọi người mở rộng mối quan hệ của mình, gặp gỡ và tương tác với những

người có chung sở thích, ý tưởng hoặc mục tiêu trong cuộc sống. Các nhóm cộng đồng trên mạng xã hội là nơi mọi người cùng nhau chia sẻ kiến thức, kinh nghiệm và xây dựng những mối quan hệ bền chặt, tạo ra giá trị tích cực không chỉ trong phạm vi cá nhân mà còn ở quy mô cộng đồng lớn hơn.

Bên cạnh khả năng kết nối mạnh mẽ, mạng xã hội còn trở thành một công cụ quan trọng trong lĩnh vực giáo dục và trao đổi tri thức. Nó cung cấp một không gian học tập không giới hạn, nơi mọi người có thể tiếp cận nguồn tài nguyên phong phú từ khắp nơi trên thế giới. Học sinh, sinh viên, giáo viên và cả các chuyên gia đều có thể sử dụng mạng xã hội để tìm kiếm tài liệu, trao đổi kiến thức, hoặc tham gia các nhóm học tập trực tuyến. Những khóa học, hội thảo và buổi trò chuyện trực tuyến được tổ chức thông qua mạng xã hội đã mở ra cơ hội học tập đa dạng, linh hoạt, giúp mọi người nâng cao trình độ mà không cần phải đến lớp học truyền thống. Không chỉ vậy, các nền tảng này còn giúp người dùng cập nhật các xu hướng, công nghệ mới nhất trong lĩnh vực chuyên môn, thúc đẩy sự học hỏi và sáng tạo liên tục trong một thế giới không ngừng thay đổi.

Một lợi ích khác rất đáng kể của mạng xã hội là tác động tích cực trong lĩnh vực kinh doanh. Các doanh nghiệp hiện đại đã tận dụng mạng xã hội như một kênh quảng bá hiệu quả để đưa sản phẩm và dịch vụ của mình đến với khách hàng. Thông qua các bài viết, hình ảnh, video và quảng cáo được nhắm mục tiêu, doanh nghiệp có thể tiếp cận chính xác nhóm khách hàng tiềm năng dựa trên sở thích, thói quen và hành vi tiêu dùng. Không chỉ giúp tiết kiệm chi phí quảng cáo, mạng xã hội còn cho phép doanh nghiệp giao tiếp trực tiếp với khách hàng, nhận phản hồi nhanh chóng và kịp thời giải quyết các vấn đề phát sinh. Hơn thế nữa, mạng xã hội còn tạo ra một môi trường để các doanh nghiệp nhỏ và cá nhân khởi nghiệp có thể tiếp cận thị trường mà trước đây khó có thể thực hiện được, từ đó thúc đẩy sự phát triển của nền kinh tế số và tăng cường tính cạnh tranh trên thị trường.

Mạng xã hội không chỉ dừng lại ở việc phục vụ mục đích cá nhân hay kinh doanh, mà còn góp phần quan trọng trong việc thúc đẩy ý thức cộng đồng và lan tỏa

các giá trị tích cực trong xã hội. Đây là nơi các phong trào thiện nguyện, chiến dịch bảo vệ môi trường, hay các dự án vì cộng đồng được phát động và lan truyền nhanh chóng. Nhờ mạng xã hội, các câu chuyện truyền cảm hứng, các sáng kiến vì cộng đồng, hay những hoạt động xã hội ý nghĩa có thể tiếp cận đến hàng triệu người chỉ trong thời gian ngắn, khơi dậy ý thức trách nhiệm và sự chung tay đóng góp của nhiều người. Điều này đã tạo ra sự kết nối mạnh mẽ giữa các cá nhân và tổ chức, giúp giải quyết nhiều vấn đề xã hội và thúc đẩy sự phát triển bền vững.

Cuối cùng, mạng xã hội còn là một nguồn giải trí phong phú và đa dạng, mang lại niềm vui và sự thư giãn cho người dùng. Từ việc xem video, nghe nhạc, đọc truyện đến tham gia các trò chơi trực tuyến, người dùng có vô vàn lựa chọn để giải trí ngay trên các nền tảng này. Không chỉ vậy, mạng xã hội còn là nơi để mọi người chia sẻ những khoảnh khắc đáng nhớ trong cuộc sống, nhận được sự động viên, ủng hộ từ bạn bè và cộng đồng. Đây cũng là công cụ để khám phá những nền văn hóa mới, kết nối với những ý tưởng sáng tạo và tìm kiếm nguồn cảm hứng.

Nhìn chung, lợi ích của mạng xã hội rất đa dạng và sâu rộng, từ việc cải thiện khả năng kết nối, nâng cao tri thức, hỗ trợ kinh doanh đến tạo ra sự gắn kết xã hội và mang lại niềm vui trong cuộc sống. Tuy nhiên, để tối ưu hóa những lợi ích này, người dùng cần sử dụng mạng xã hội một cách tỉnh táo và có trách nhiệm, đảm bảo rằng nó thực sự trở thành một công cụ mang lại giá trị tích cực và ý nghĩa lâu dài.

1.1.3. Tác hại của mạng xã hội

Mặc dù mạng xã hội mang lại nhiều lợi ích, nó cũng tiềm ẩn những tác hại đáng kể, ảnh hưởng tiêu cực đến đời sống cá nhân, xã hội và sức khỏe tinh thần của con người. Các vấn đề này không chỉ giới hạn trong phạm vi cá nhân mà còn lan rộng ra cộng đồng, gây ra nhiều hệ lụy khó lường nếu không được quản lý và sử dụng một cách hợp lý.

Trước hết, mạng xã hội là nguyên nhân dẫn đến việc lãng phí thời gian và gây ra tình trạng nghiện mạng xã hội. Nhiều người dành quá nhiều thời gian trên các nền tảng như Facebook, Instagram, TikTok hay Twitter để lướt xem thông tin, tương tác,

hoặc giải trí mà không nhận ra rằng họ đang bỏ lỡ các hoạt động thực tế quan trọng. Việc "nghiện" mạng xã hội không chỉ ảnh hưởng đến năng suất làm việc và học tập mà còn gây mất cân bằng trong cuộc sống, dẫn đến giảm chất lượng mối quan hệ ngoài đời thực. Nhiều nghiên cứu đã chỉ ra rằng thời gian sử dụng mạng xã hội quá mức có liên quan đến sự suy giảm khả năng tập trung và hiệu quả làm việc.

Ngoài ra, mạng xã hội còn gây ra những tác động tiêu cực đến sức khỏe tinh thần. Các nền tảng này thường tạo ra áp lực về hình ảnh và phong cách sống hoàn hảo, khiến nhiều người cảm thấy tự ti hoặc không hài lòng với bản thân. Việc so sánh cuộc sống của mình với những hình ảnh "lung linh" trên mạng xã hội có thể dẫn đến cảm giác ghen tị, trầm cảm, và lo âu. Thêm vào đó, sự hiện diện của các bình luận tiêu cực, bạo lực mạng, hay các hành vi "body shaming" và "cyberbullying" càng làm gia tăng mức độ căng thẳng, đặc biệt là ở giới trẻ. Nhiều người đã phải đối mặt với những tổn thương tâm lý nặng nề khi trở thành nạn nhân của những cuộc tấn công trên mạng xã hội.

Mạng xã hội cũng góp phần lan truyền thông tin sai lệch và tin giả (fake news) với tốc độ nhanh chóng. Bản chất mở và không kiểm soát chặt chẽ của các nền tảng mạng xã hội khiến cho thông tin sai lệch dễ dàng được lan truyền, gây ảnh hưởng tiêu cực đến nhận thức và hành động của cộng đồng. Những thông tin giả mạo về các sự kiện, dịch bệnh, hoặc chính trị không chỉ làm xáo trộn xã hội mà còn gây mất niềm tin vào các nguồn thông tin chính thống. Điều này đặc biệt nguy hiểm trong những thời điểm khủng hoảng, khi tin giả có thể dẫn đến những hậu quả nghiêm trọng về kinh tế, sức khỏe và an ninh xã hội.

Bên cạnh đó, mạng xã hội còn ảnh hưởng tiêu cực đến sự riêng tư và bảo mật thông tin cá nhân. Khi tham gia mạng xã hội, người dùng thường phải chia sẻ thông tin cá nhân, hình ảnh, và các dữ liệu nhạy cảm khác. Điều này khiến họ dễ trở thành mục tiêu của các hành vi xâm phạm quyền riêng tư, lừa đảo, hoặc đánh cắp thông tin. Các vụ rò rỉ dữ liệu lớn trên mạng xã hội đã gây ra những hậu quả nghiêm trọng, ảnh hưởng đến hàng triệu người dùng trên toàn cầu. Không chỉ vậy, việc các nền tảng

mạng xã hội sử dụng dữ liệu cá nhân để phục vụ quảng cáo hoặc mục đích thương mại cũng đặt ra nhiều lo ngại về quyền riêng tư.

Mạng xã hội còn có tác động tiêu cực đến các mối quan hệ trong đời thực. Thay vì dành thời gian trò chuyện trực tiếp hoặc tham gia các hoạt động chung, nhiều người lại bị cuốn vào thế giới ảo trên mạng xã hội. Điều này dẫn đến sự xa cách trong gia đình, bạn bè và đồng nghiệp, làm suy giảm chất lượng của các mối quan hệ. Hơn nữa, mạng xã hội còn là nguyên nhân dẫn đến nhiều xung đột do sự khác biệt về quan điểm hoặc những hiểu lầm trong giao tiếp, đặc biệt khi các cuộc tranh luận trên mạng dễ bị đẩy lên mức căng thẳng và tiêu cực.

Ngoài các tác hại trên, mạng xã hội còn gây ra các vấn đề nghiêm trọng về xã hội như giảm sự tập trung vào các giá trị văn hóa truyền thống, cổ vũ lối sống chạy theo xu hướng, và làm gia tăng tình trạng phân cực xã hội. Thay vì thúc đẩy sự đoàn kết và hiểu biết lẫn nhau, mạng xã hội đôi khi tạo ra các "vùng ý kiến cộng hưởng" (echo chamber), nơi mà người dùng chỉ tiếp cận những thông tin phù hợp với quan điểm cá nhân, dẫn đến sự thiếu đa dạng trong tư duy và khả năng chấp nhận ý kiến trái chiều.

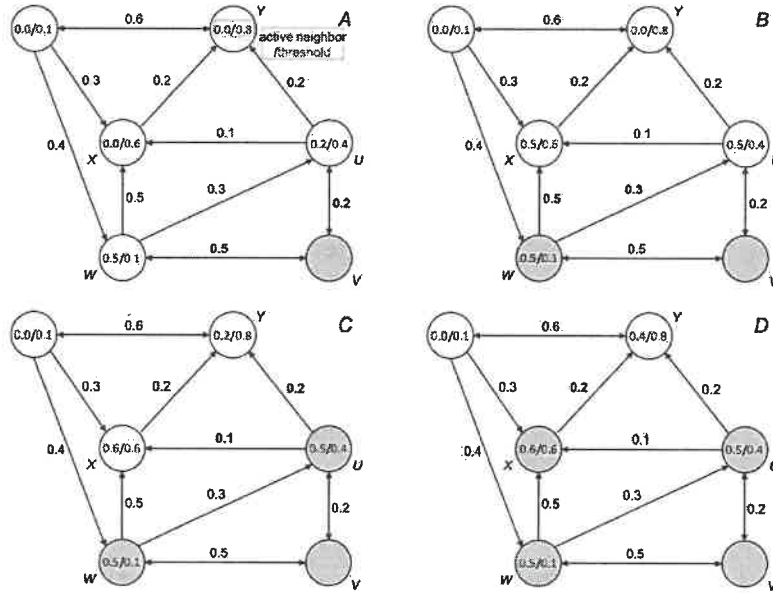
Tóm lại, mạng xã hội là một con dao hai lưỡi. Bên cạnh những lợi ích vượt trội mà nó mang lại, những tác hại mà mạng xã hội gây ra cũng rất đáng lo ngại. Việc sử dụng mạng xã hội một cách thiếu kiểm soát không chỉ ảnh hưởng đến cuộc sống cá nhân mà còn gây ra nhiều vấn đề lớn cho xã hội. Do đó, cần có sự cân nhắc, tỉnh táo và ý thức rõ ràng khi sử dụng mạng xã hội để tận dụng tốt nhất các giá trị tích cực mà nó mang lại, đồng thời hạn chế tối đa những rủi ro và tác động tiêu cực.

1.2. Mô hình hóa lan truyền thông tin trên mạng xã hội

1.2.1. Mô hình Ngưỡng tuyến tính (LT)

Mô hình Ngưỡng tuyến tính (LT) là một trường hợp đặc biệt của lan truyền thông tin trên đồ thị rời rạc, trong đó mỗi cạnh $e = (u, v)$ được gán một trọng số $w(u, v)$, thể hiện mức độ tương tác giữa hai đỉnh, với điều kiện tổng trọng số của tất

cả các cạnh đi vào một đỉnh không vượt quá 1, tức $\sum_{u \in N_{in}(v)} w(u, v) \leq 1$. Mỗi đỉnh u có một ngưỡng kích hoạt θ_u , được chọn ngẫu nhiên trong khoảng $[0,1]$, biểu thị mức tác động cần thiết để kích hoạt đỉnh đó. Quá trình lan truyền bắt đầu tại bước $t = 0$, khi các đỉnh trong tập khởi tạo S được kích hoạt ($S_0 = S$). Ở các bước tiếp theo ($t \geq 1$), một đỉnh v chưa kích hoạt sẽ được kích hoạt nếu tổng trọng số của các cạnh đi vào từ các đỉnh đã kích hoạt vượt ngưỡng θ_v , tức $\sum_{u \in N_{in}(v) \cap S_{t-1}} w(u, v) \geq \theta_v$. Một khi kích hoạt, trạng thái của đỉnh không thay đổi. Quá trình lan truyền kết thúc khi không có đỉnh nào mới được kích hoạt trong hai bước liên tiếp. Mô hình này mô phỏng cách thông tin lan tỏa trong mạng xã hội, kết hợp ảnh hưởng từ các mối quan hệ tương tác và mức độ nhạy cảm cá nhân của từng đỉnh.

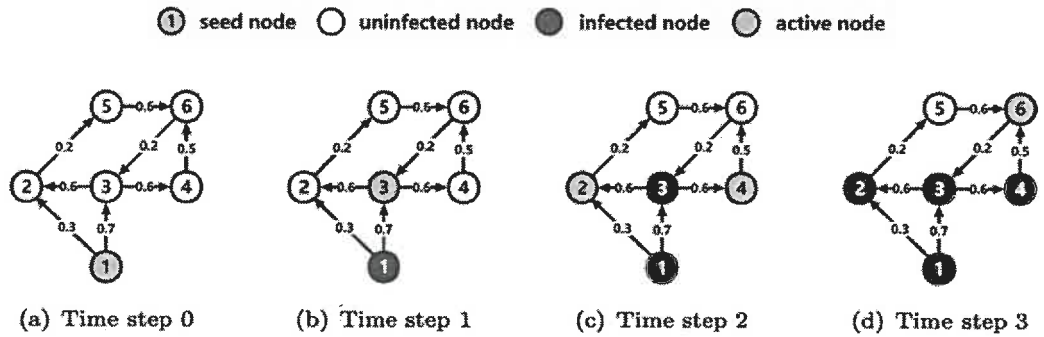


Hình 1. 1: Ví dụ về mô hình Ngưỡng Tuyến tính (LT).

(A) Nút V được kích hoạt và ảnh hưởng đến W và U lần lượt với cường độ 0.5 và 0.2; (B) W được kích hoạt và ảnh hưởng đến X và U lần lượt với cường độ 0.5 và 0.3; (C) U được kích hoạt và ảnh hưởng đến X và Y lần lượt với cường độ 0.1 và 0.2; (D) X được kích hoạt và ảnh hưởng đến Y với cường độ 0.2; không còn nút nào có thể được kích hoạt thêm; quá trình dừng lại.

1.2.2. Mô hình bậc độc lập (IC)

Mô hình IC (Independent Cascade) gán cho mỗi cạnh $(u, v) \in E$ một xác suất ảnh hưởng $p(u, v) \in [0, 1]$, đại diện cho khả năng mà đỉnh u có thể kích hoạt đỉnh v . Cụ thể, xác suất $p(u, v)$ biểu thị mức độ mà đỉnh u tác động đến trạng thái của đỉnh v trong quá trình lan truyền. Theo mô hình này, nếu một đỉnh u đã được kích hoạt tại thời điểm $t \geq 0$, thì nó sẽ có một cơ hội duy nhất để cố gắng kích hoạt các đỉnh hàng xóm của mình chưa được kích hoạt tại bước $t + 1$. Quá trình lan truyền dừng lại khi không có bất kỳ đỉnh nào mới được kích hoạt trong hai bước liên tiếp. Mô hình này mô phỏng một cách trực quan sự lan tỏa ảnh hưởng trong mạng lưới, chẳng hạn như sự lan truyền của thông tin, hành vi xã hội hoặc sản phẩm trong cộng đồng.



Hình 1. 2: Minh họa mô hình bậc độc lập (IC) trong quá trình lan truyền ảnh hưởng trên mạng lưới

1.2.3. Mô hình ngăn trong dịch tễ học (SIS)

Mô hình SIS (Susceptible-Infectious-Susceptible) là một trong những mô hình cơ bản trong dịch tễ học, được sử dụng để mô phỏng sự lây lan của bệnh truyền nhiễm trong quần thể. Trong mô hình này, mỗi cá thể có thể tồn tại ở một trong hai trạng thái: S (dễ nhiễm) và I (đang nhiễm bệnh). Khác với nhiều mô hình khác, cá thể trong trạng thái I sau khi hồi phục sẽ quay lại trạng thái S , tức không phát triển khả năng miễn dịch và có thể bị tái nhiễm. Quá trình lây lan được điều khiển bởi hai tham số chính: tỷ lệ lây nhiễm β (xác suất một cá thể dễ nhiễm bị nhiễm bệnh khi tiếp xúc với cá thể đang bệnh) và tỷ lệ hồi phục γ (xác suất cá thể bệnh hồi phục về trạng thái dễ

nhiễm). Quá trình này được mô tả bằng hệ phương trình vi phân: $\frac{dS}{dt} = \gamma I - \beta SI$ và $\frac{dI}{dt} = \beta SI - \gamma I$ trong đó số lượng cá thể ở mỗi trạng thái thay đổi theo thời gian. Một yếu tố quan trọng trong mô hình là số sinh sản cơ bản $R_0 = \frac{\beta}{\gamma}$, giúp đánh giá khả năng bùng phát của bệnh: nếu $R_0 > 1$, bệnh sẽ tiếp tục lan rộng; ngược lại, nếu $R_0 \leq 1$, bệnh sẽ dần biến mất. Mô hình SIS được ứng dụng để phân tích các bệnh truyền nhiễm không tạo miễn dịch lâu dài, như cảm cúm hoặc các bệnh nhiễm trùng tái phát.

1.3. Một số bài toán lan truyền thông tin trên mạng xã hội

Trong phần này, luận đề tập trung trình bày ba bài toán chính liên quan đến lan truyền thông tin: Tối đa hóa ảnh hưởng (IM) và Tối đa hóa ngân sách ảnh hưởng (IB).

1.3.1. Tối đa hóa ảnh hưởng (IM)

Bài toán tối đa hóa ảnh hưởng (Influence Maximization - IM) là một trong những bài toán quan trọng trong nghiên cứu mạng xã hội và có ý nghĩa to lớn trong các hoạt động tiếp thị (marketing) và kinh doanh trên mạng xã hội (MXH) ngày nay. Mục tiêu của bài toán là tìm ra cách chọn một nhóm cá thể (hoặc nút) có khả năng lan tỏa thông tin mạnh mẽ nhất trong mạng. Cụ thể, bài toán được mô tả như sau: Cho một mạng xã hội $G = (V, E)$, nơi V là tập các đỉnh (người dùng) và E là tập các cạnh (quan hệ giữa các người dùng), cùng với một mô hình lan tỏa thông tin $\mathcal{M}$. Với một số nguyên dương $k > 0$ (đại diện cho ngân sách hoặc số lượng cá thể được chọn), tìm tập hợp $S \subseteq V$, sao cho $|S| = k$ và ảnh hưởng lan tỏa từ tập S là lớn nhất. Bài toán này thuộc lớp các bài toán NP-hard và việc tính toán chính xác hàm ảnh hưởng thuộc nhóm bài toán #P-hard. Vì vậy, các thuật toán thường tiếp cận bài toán thông qua các phương pháp xấp xỉ hoặc dựa trên lý thuyết đồ thị. Hai nhóm chính của các thuật toán bao gồm:

1. Thuật toán xấp xỉ: Đảm bảo lời giải có tính hiệu quả về mặt lý thuyết, thường dựa trên các đặc tính của mô hình lan tỏa.

2. Thuật toán dựa trên cấu trúc mạng: Sử dụng các đặc trưng như độ đo trung tâm, độ đo cộng đồng hoặc đường đi để tối ưu hóa quá trình chọn tập S.

Ngoài ra, các bài toán biến thể của IM cũng được nghiên cứu rộng rãi, tập trung vào các yếu tố thực tiễn như chi phí và lợi ích của lan truyền, đặc điểm chủ đề, khoảng cách, thời gian, và địa điểm. Những biến thể này giúp bài toán trở nên linh hoạt hơn và ứng dụng hiệu quả vào nhiều ngữ cảnh khác nhau, từ quảng cáo, lan tỏa sản phẩm, đến chiến lược truyền thông xã hội.

1.3.2. Ngăn chặn ảnh hưởng (IB)

Trái ngược với bài toán tối đa hóa ảnh hưởng (IM), bài toán ngăn chặn ảnh hưởng (Influence Blocking - IB) tập trung vào việc hạn chế sự lan tỏa và phát tán thông tin không mong muốn trong mạng xã hội. Mục tiêu chính của bài toán này là giảm thiểu tác động tiêu cực từ các yếu tố xấu trên mạng xã hội, bao gồm: tin đồn sai sự thật, nội dung kích động, thông tin lệch lạc, các tư tưởng cực đoan, hoặc sự lây lan của virus (kỹ thuật số hoặc sinh học).

Để đạt được mục tiêu này, bài toán IB tìm kiếm các chiến lược tối ưu nhằm hạn chế tác động từ các nguồn phát tán. Hai phương pháp chính thường được áp dụng bao gồm:

1. **Loại bỏ hoặc bảo vệ các nút/cạnh nhạy cảm:** Phương pháp này tương tự như "tiêm vắc-xin" trong dịch tễ học, bằng cách chọn một tập các đỉnh (hoặc cạnh) cụ thể để cách ly hoặc bảo vệ, giúp chúng trở nên miễn nhiễm với tác động tiêu cực.
2. **Tẩy nhiễm thông tin bằng cách thúc đẩy các ảnh hưởng tích cực:** Chọn các nút hoặc vị trí chiến lược để phát tán thông tin tích cực nhằm đối trọng và vô hiệu hóa ảnh hưởng của các thông tin tiêu cực.

Với tính chất phức tạp và thực tế, bài toán IB không chỉ có ý nghĩa trong việc bảo vệ môi trường mạng xã hội mà còn góp phần duy trì sự an toàn thông tin, xây

dựng cộng đồng mạng lành mạnh, và chống lại các mối đe dọa như tin giả hoặc nội dung độc hại

1.4. Kết luận chương

Chương 1 đã hoàn thành vai trò cung cấp một bức tranh tổng quan và nền tảng kiến thức vững chắc về các khía cạnh then chốt của sự lan truyền thông tin trên mạng xã hội. Chúng ta đã bắt đầu bằng việc khám phá bản chất đa diện của mạng xã hội, từ những lợi ích kết nối to lớn đến những rủi ro tiềm ẩn, qua đó nhấn mạnh tầm quan trọng của việc nghiên cứu các hiện tượng diễn ra trên đó. Tiếp theo, các mô hình lan truyền thông tin kinh điển như Ngưỡng Tuyến tính (LT), Bậc Độc lập (IC) và mô hình SIS trong dịch tễ học đã được giới thiệu, cung cấp bộ công cụ lý thuyết cần thiết để mô phỏng và lý giải động lực học phức tạp của sự lan tỏa.

Trên cơ sở hiểu biết về các mô hình này, các bài toán thực tiễn và mang tính học thuật cao như Tối đa hóa Ảnh hưởng (IM) và Ngăn chặn Ảnh hưởng (IB) đã được định nghĩa. Những bài toán này không chỉ làm nổi bật các thách thức kỹ thuật mà còn phản ánh nhu cầu thực tiễn trong việc quản lý, khai thác và kiểm soát dòng chảy thông tin trong thế giới kết nối ngày nay. Việc điểu qua các nghiên cứu liên quan cũng đã cho thấy sự phát triển không ngừng và tính đa dạng của các hướng tiếp cận trong lĩnh vực này.

Với những kiến thức nền tảng về mạng xã hội, các mô hình lan truyền và các bài toán liên quan đã được trang bị, chúng ta đã sẵn sàng để đi sâu hơn vào các phương pháp và kỹ thuật cụ thể nhằm giải quyết bài toán tối đa hóa ảnh hưởng – một trong những thách thức trung tâm và có nhiều ứng dụng nhất. Chương tiếp theo sẽ tập trung vào việc phân tích chi tiết hơn bài toán này và giới thiệu một số phương pháp giải quyết tiêu biểu, đặc biệt là trong bối cảnh các mạng xã hội ngày càng trở nên phức tạp và đa tầng.

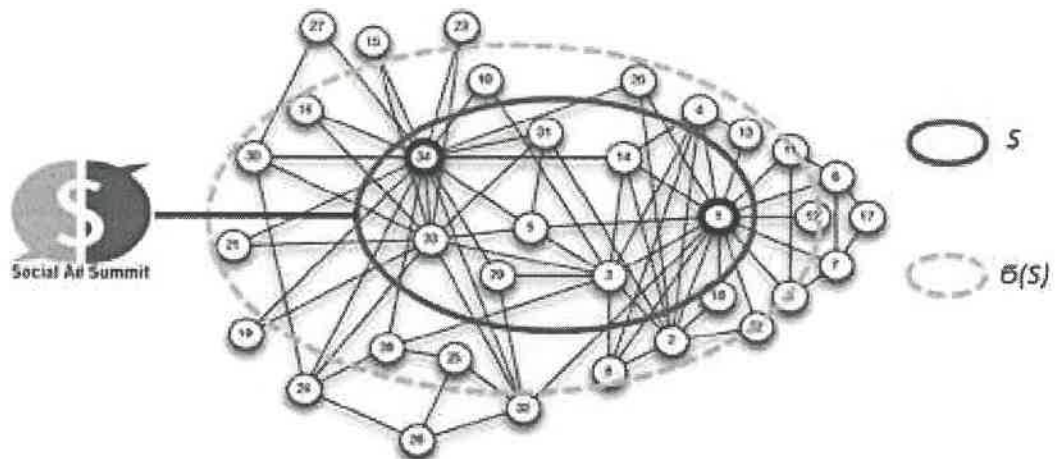
CHƯƠNG 2 - TỐI ĐA HÓA ẢNH HƯỞNG ĐA TẦNG VỚI CÁC RÀNG BUỘC VỀ NGÂN SÁCH TRÊN MẠNG XÃ HỘI

Tiếp nối những khái niệm nền tảng về lan truyền thông tin và bài toán tối đa hóa ảnh hưởng (IM) trên mạng đơn lớp từ Chương 1, chương này sẽ tập trung vào một kịch bản phức tạp và thực tế hơn: Tối đa hóa Ảnh hưởng Đa tầng (Multiplex Influence Maximization - MIM). Trong bối cảnh người dùng ngày càng tương tác trên nhiều nền tảng, việc hiểu và khai thác sự lan tỏa qua các lớp mạng liên kết, đặc biệt khi đối mặt với những ràng buộc về ngân sách, trở thành một thách thức quan trọng. Chương này sẽ chính thức định nghĩa bài toán MIM, phân tích các đặc thù và độ phức tạp của nó, đồng thời khám phá các hướng tiếp cận và phương pháp giải quyết, qua đó đặt nền móng cho việc phát triển giải pháp được đề xuất trong đề án.

2.1. Đặt vấn đề và phát biểu bài toán tối đa hóa ảnh hưởng đa tầng trên mạng xã hội

2.1.1. Đặt vấn đề

Tối đa hóa ảnh hưởng trên mạng xã hội là bài toán tìm cách chọn một tập hợp nhỏ các cá nhân (hoặc nút) trong mạng xã hội để lan truyền thông tin, sao cho thông tin đó tiếp cận được đến số lượng người dùng lớn nhất có thể. Đây là một bài toán quan trọng trong lĩnh vực truyền thông xã hội, tiếp thị, và quản lý khủng hoảng, vì nó giúp tối ưu hóa hiệu quả của các chiến dịch quảng cáo hoặc lan truyền thông điệp.



Hình 2. 1: Ví dụ về Tối đa hóa Ảnh hưởng: Chọn tập hạt giống S (vòng đỏ) để tối đa hóa sự lan tỏa thông tin (vùng $\sigma(S)$ - vòng vàng) cho một chiến dịch.

Bài toán tối đa hóa ảnh hưởng được coi là một bài toán tối ưu tổ hợp vì nó liên quan đến việc chọn ra một tập hợp con tối ưu từ một tập hợp lớn các lựa chọn có thể (ví dụ: chọn ra k người dùng có ảnh hưởng lớn nhất trong mạng xã hội). Cụ thể, bài toán này có thể được mô tả bằng bộ ba (S, f, Ω) , trong đó:

- S : Tập hợp hữu hạn các trạng thái (các tập hợp con có thể của người dùng).
- f : Hàm mục tiêu, đo lường mức độ ảnh hưởng của một tập hợp người dùng (ví dụ: số lượng người dùng được tiếp cận thông tin).
- Ω : Tập các ràng buộc, chẳng hạn như giới hạn về số lượng người dùng được chọn (ví dụ: chỉ được chọn tối đa k người).

Mục tiêu của bài toán là tìm ra tập hợp người dùng $s^* \in S$ sao cho hàm f đạt giá trị cực đại, tức là tối đa hóa phạm vi lan truyền thông tin trên mạng xã hội.

a) Phân loại các lớp bài toán trong tối ưu tổ hợp

Trong lý thuyết tối ưu tổ hợp, việc phân loại các lớp bài toán dựa trên độ phức tạp tính toán là một khía cạnh quan trọng giúp hiểu rõ hơn về bản chất và khả năng giải quyết của các bài toán. Hai lớp bài toán cơ bản và được nghiên cứu rộng rãi nhất là **P**, **NP** và **#P**.

- Lớp P (Polynomial-time): là lớp các bài toán quyết định có thể được giải quyết bằng một thuật toán đơn định (deterministic algorithm) trong thời gian đa thức theo kích thước của đầu vào. Nói cách khác, nếu tồn tại một thuật toán có thể giải bài toán trong thời gian $O(n^k)$, với n là kích thước đầu vào và k là một hằng số không phụ thuộc vào n , thì bài toán đó thuộc lớp P. Các bài toán trong lớp P được coi là "dễ giải" vì thời gian giải quyết chúng tăng chậm khi kích thước đầu vào tăng lên.
- Lớp NP (Non-Deterministic Polynomial-time): bao gồm các bài toán quyết định mà nếu có một lời giải, thì lời giải đó có thể được "*kiểm tra*" bằng một thuật toán đơn định trong thời gian đa thức. Ví dụ, bài toán "kiểm tra xem một số có phải là số nguyên tố không" thuộc lớp P vì ta có thể kiểm tra nó trong thời gian đa thức. Tuy nhiên, bài toán "tìm một chu trình Hamilton trong đồ thị" thuộc lớp NP vì nếu ai đó đưa ra một chu trình cụ thể, ta có thể kiểm tra xem chu trình đó có thỏa mãn điều kiện hay không trong thời gian đa thức. Một cách hiểu khác, lớp NP cũng bao gồm các bài toán có thể được giải bằng một thuật toán không đơn định (non-deterministic algorithm) trong thời gian đa thức. Thuật toán không đơn định có khả năng "đoán" được lời giải đúng trong một bước tính toán, và sau đó kiểm tra lời giải đó trong thời gian đa thức. Một câu hỏi mở nổi tiếng trong khoa học máy tính là liệu $P = NP$ hay không, tức là liệu mọi bài toán trong NP có thể được giải trong thời gian đa thức hay không.
- Lớp #P: là lớp các bài toán đếm (counting problems), trong đó mục tiêu là xác định số lượng các lời giải hợp lệ cho một bài toán quyết định thuộc lớp NP. Cụ thể, một bài toán thuộc lớp #P yêu cầu tính toán hàm $f(x)$, với $f(x)$ là số lượng các đường đi từ cấu hình ban đầu đến một cấu hình chấp nhận được trong máy Turing không đơn định, và việc tính toán này phải được thực hiện trong thời gian đa thức theo kích thước đầu vào x . Các bài toán trong lớp #P thường khó hơn các bài toán trong lớp NP, vì chúng không chỉ yêu cầu kiểm tra sự tồn tại của một lời giải mà còn phải đếm số lượng lời giải đó.

Bài toán tối đa hóa ảnh hưởng trên mạng xã hội là một bài toán tối ưu tổ hợp thuộc lớp **NP-hard**. Điều này có nghĩa là không có thuật toán đơn định nào có thể giải quyết bài toán này trong thời gian đa thức đối với trường hợp tổng quát, trừ khi $P = NP$. Cụ thể, bài toán này yêu cầu chọn ra một tập hợp con tối ưu các nút (người dùng) trong mạng xã hội để tối đa hóa phạm vi lan truyền thông tin, và việc duyệt qua tất cả các tập hợp con có thể là không khả thi do độ phức tạp tính toán tăng theo hàm mũ. Do tính chất phức tạp của bài toán, các phương pháp giải quyết thường tập trung vào việc sử dụng các thuật toán gần đúng (approximation algorithms) hoặc heuristic để tìm ra các lời giải chấp nhận được trong thời gian hợp lý. Các phương pháp này tuy không đảm bảo tìm được lời giải tối ưu toàn cục, nhưng chúng có thể cung cấp các giải pháp hiệu quả và thiết thực trong các ứng dụng thực tế.

b) Một số phương pháp giải bài toán tối đa hóa ảnh hưởng

❖ Thuật toán tham lam

Thuật toán Greedy (tham lam) là một trong những phương pháp tiếp cận phổ biến và hiệu quả để giải quyết bài toán tối đa hóa ảnh hưởng trên mạng xã hội. Thuật toán này dựa trên nguyên tắc lựa chọn các quyết định tối ưu cục bộ tại mỗi bước, với hy vọng rằng các lựa chọn này sẽ dẫn đến một lời giải tối ưu toàn cục hoặc gần tối ưu. Mặc dù thuật toán Greedy không đảm bảo luôn tìm ra lời giải tối ưu toàn cục, nhưng nó thường cung cấp các giải pháp chấp nhận được với độ phức tạp tính toán thấp, phù hợp với các bài toán có kích thước lớn. Thuật toán Greedy giải quyết bài toán tối đa hóa ảnh hưởng bằng cách lặp lại các bước sau:

1. **Khởi tạo:** Bắt đầu với một tập hợp rỗng $S = \emptyset$, đại diện cho tập hợp các nút (người dùng) được chọn để lan truyền thông tin.
2. **Lựa chọn tham lam:** Tại mỗi bước, thuật toán sẽ chọn một nút v chưa được chọn sao cho việc thêm v vào tập hợp S sẽ làm tăng giá trị hàm mục tiêu $f(S)$ một cách tối đa. Nói cách khác, tại mỗi bước, thuật toán chọn nút có ảnh hưởng lớn nhất đến phạm vi lan truyền thông tin.

3. **Cập nhật tập hợp:** Sau khi chọn được nút v , thuật toán thêm v vào tập hợp S và cập nhật giá trị hàm mục tiêu $f(S)$.
4. **Lặp lại:** Quá trình lựa chọn tham lam được lặp lại cho đến khi tập hợp S đạt được kích thước mong muốn (ví dụ: $|S| = k$) hoặc không thể cải thiện thêm giá trị hàm mục tiêu.

Thuật toán Greedy có nhiều ưu điểm như đơn giản, dễ triển khai, hiệu quả về thời gian và linh hoạt trong việc áp dụng cho các bài toán khác nhau. Tuy nhiên, nó không đảm bảo tối ưu toàn cục do chỉ tập trung vào lựa chọn cục bộ và kết quả có thể phụ thuộc vào thứ tự lựa chọn, dẫn đến sự không ổn định. Dù vậy, với ưu điểm vượt trội, thuật toán Greedy vẫn là một phương pháp phổ biến và hiệu quả trong thực tế.

❖ Thuật toán CELF

Thuật toán CELF (Cost-Effective Lazy Forward) là một phương pháp hiệu quả để giải bài toán tối đa hóa ảnh hưởng trên mạng xã hội. Thuật toán này được đề xuất nhằm cải thiện hiệu suất của thuật toán tham lam (Greedy Algorithm) truyền thống bằng cách tận dụng tính chất submodular của hàm ảnh hưởng và áp dụng kỹ thuật lazy evaluation (đánh giá lười). CELF giúp giảm đáng kể số lần tính toán ảnh hưởng gia tăng (marginal gain) mà vẫn duy trì được chất lượng lời giải gần tối ưu.

Ý tưởng chính của thuật toán CELF dựa trên việc thuật toán tham lam truyền thống yêu cầu tính toán ảnh hưởng gia tăng của tất cả các nút trong mỗi bước, dẫn đến độ phức tạp tính toán cao, đặc biệt với các mạng xã hội lớn. CELF khắc phục hạn chế này bằng cách:

1. **Tận dụng tính chất submodular:** Hàm ảnh hưởng $f(S)$ có tính chất submodular, nghĩa là ảnh hưởng gia tăng của một nút sẽ giảm dần khi tập hợp S được mở rộng. Nhờ đó, CELF có thể loại bỏ các nút không cần thiết khỏi quá trình tính toán.

2. **Lazy evaluation (Đánh giá lười):** Thay vì tính toán lại ảnh hưởng gia tăng của tất cả các nút trong mỗi bước, CELF chỉ tính toán lại ảnh hưởng gia tăng của các nút có khả năng cao nhất để trở thành lựa chọn tiếp theo.

Thuật toán CELF được thực hiện qua các bước sau:

Bước 1: Khởi tạo

- Đầu vào: Đồ thị mạng xã hội $G = (V, E)$, số lượng nút cần chọn k .
- Đầu ra: Tập hợp S gồm k nút có ảnh hưởng lớn nhất.
- Khởi tạo: $S = \emptyset$, và một hàng đợi ưu tiên (priority queue) Q chứa tất cả các nút trong mạng, với ưu tiên ban đầu là ảnh hưởng gia tăng của từng nút khi thêm vào tập S .

Bước 2: Tính toán ảnh hưởng gia tăng ban đầu

- Với mỗi nút $u \in V$, tính ảnh hưởng gia tăng $\Delta(u) = f(S \cup \{u\}) - f(\emptyset)$, trong đó f là hàm ảnh hưởng.
- Sắp xếp các nút trong hàng đợi Q theo thứ tự giảm dần của $\Delta(u)$.

Bước 3: Lặp để chọn k nút

Trong mỗi bước i (từ 1 đến k):

1. Lấy nút u có ảnh hưởng gia tăng lớn nhất từ hàng đợi Q .
2. Tính toán lại ảnh hưởng gia tăng của u : $\Delta(u) = f(S \cup \{u\}) - f(S)$.
3. So sánh $\Delta(u)$ với ảnh hưởng gia tăng của nút tiếp theo trong hàng đợi:
 - Nếu $\Delta(u)$ vẫn lớn nhất, thêm u vào tập S .
 - Ngược lại, đẩy u trở lại hàng đợi với giá trị $\Delta(u)$ mới và lặp lại bước 3.

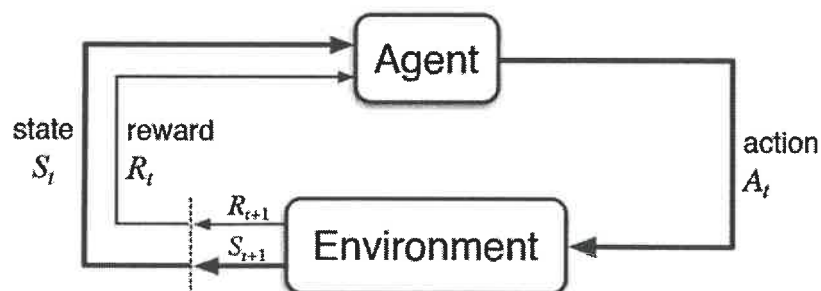
Bước 4: Trả về kết quả

Kết thúc khi đã chọn được k nút. Tập hợp S là kết quả cuối cùng.

Thuật toán CELF mang lại nhiều ưu điểm đáng kể, đặc biệt trong việc giảm thời gian tính toán nhờ kỹ thuật lazy evaluation, giúp giảm đáng kể số lần tính toán ảnh hưởng gia tăng so với thuật toán tham lam truyền thống. Đồng thời, CELF vẫn duy trì được chất lượng lời giải với tỷ lệ xấp xỉ $1 - 1/e$ (khoảng 63%), tương tự như thuật toán tham lam. Ngoài ra, CELF rất hiệu quả khi áp dụng trên các mạng xã hội có quy mô lớn nhờ khả năng loại bỏ các nút không cần thiết khỏi quá trình tính toán. Tuy nhiên, thuật toán này cũng có một số hạn chế. Hiệu quả của CELF phụ thuộc nhiều vào việc tính toán chính xác hàm ảnh hưởng $f(S)$, điều này có thể tốn kém về thời gian, đặc biệt khi sử dụng các mô hình lan truyền phức tạp như Independent Cascade hoặc Linear Threshold. Bên cạnh đó, mặc dù CELF giảm số lần tính toán, nó vẫn yêu cầu tính toán lại ảnh hưởng gia tăng của một số nút trong mỗi bước, do đó chưa tối ưu hóa triệt để.

❖ Mô hình Học Tăng Cường cho Bài toán Tối đa hóa Ảnh hưởng

Học Tăng Cường (Reinforcement Learning - RL) là một trong những phương pháp học máy tiên tiến, nơi một tác nhân (agent) học cách thực hiện các hành động tối ưu thông qua việc tương tác liên tục với môi trường. Mục tiêu của tác nhân là tối đa hóa một phần thưởng tích lũy, đại diện cho hiệu suất tổng thể của nó trong quá trình thực hiện nhiệm vụ. RL đã chứng minh được hiệu quả vượt trội trong nhiều bài toán phức tạp, từ điều khiển robot đến tối ưu hóa hệ thống giao thông, và gần đây là bài toán tối đa hóa ảnh hưởng trên mạng xã hội.



Hình 2. 2: Mô hình tương tác giữa Tác nhân (Agent) và Môi trường (Environment) trong Học Tăng Cường (RL).

Trong bối cảnh của bài toán tối đa hóa ảnh hưởng, môi trường được mô hình hóa như một đồ thị xã hội, nơi các nút đại diện cho người dùng và các cạnh biểu thị mối quan hệ hoặc sự tương tác giữa họ. Tác nhân cần học cách chọn ra một tập hợp con các nút (người dùng) để lan truyền thông tin sao cho phạm vi lan truyền là lớn nhất. Đây là một bài toán có tính chất tổ hợp cao và phức tạp, đòi hỏi sự kết hợp giữa lý thuyết đồ thị và các kỹ thuật học máy. Các Thành Phần Chính Trong Mô Hình RL Cho Bài Toán Tối Đa Hóa Ảnh Hưởng:

- **Trạng thái (State):** Trạng thái tại mỗi bước của quá trình ra quyết định được biểu diễn bằng tập hợp các nút đã được chọn để lan truyền thông tin, cấu trúc mạng và các nút đã được kích hoạt (tức là các nút đã tiếp nhận thông tin) tại thời điểm đó. Trạng thái này phản ánh tình hình hiện tại của mạng xã hội, bao gồm cả những người dùng đã bị ảnh hưởng và những người dùng tiềm năng có thể bị ảnh hưởng trong tương lai.
- **Hành động (Action):** Hành động của tác nhân là chọn một nút chưa được chọn để thêm vào tập hợp các nút lan truyền. Mỗi hành động được thực hiện sẽ làm thay đổi trạng thái của môi trường, dẫn đến việc lan truyền thông tin đến các nút mới. Việc lựa chọn hành động phải dựa trên chiến lược tối ưu để đảm bảo rằng thông tin lan truyền một cách hiệu quả nhất.
- **Phần thưởng (Reward):** Phần thưởng là một yếu tố quan trọng trong RL, đóng vai trò hướng dẫn tác nhân học cách thực hiện các hành động tối ưu. Trong bài toán tối đa hóa ảnh hưởng, phần thưởng có thể được tính dựa trên số lượng nút mới được kích hoạt sau mỗi hành động. Mục tiêu của tác nhân là tối đa hóa tổng phần thưởng tích lũy, tương ứng với việc tối đa hóa phạm vi lan truyền thông tin trên mạng xã hội.

Học Tăng Cường (Reinforcement Learning - RL) mang lại nhiều ưu điểm như khả năng tối ưu hóa hành động dựa trên phần thưởng, xử lý môi trường phức tạp và thích nghi với thay đổi trong mạng xã hội. Tuy nhiên, RL cũng đối mặt với thách thức như yêu cầu lượng dữ liệu lớn, thời gian huấn luyện dài, khó thiết kế phần thưởng

phù hợp và khó hội tụ đến chính sách tối ưu trong môi trường phức tạp. Những thách thức này đòi hỏi sự cân nhắc kỹ lưỡng để áp dụng RL hiệu quả.

2.1.2. Phát biểu bài toán

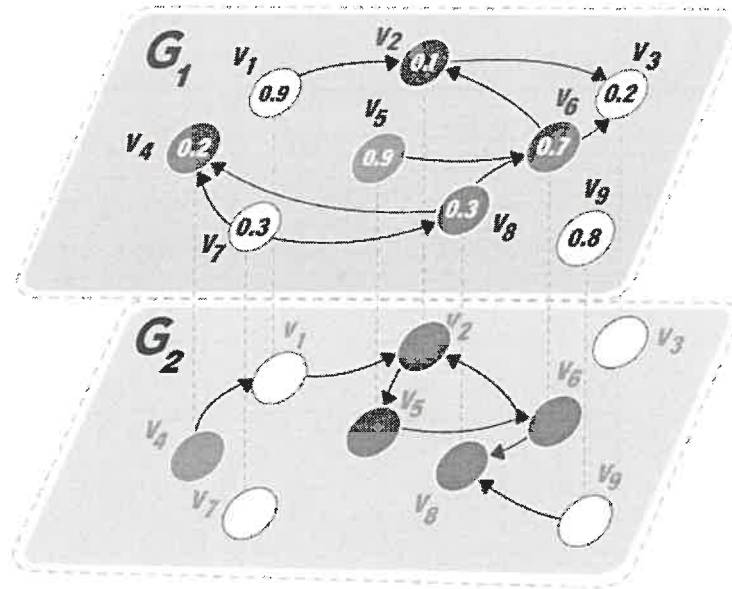
Một mạng đa lớp với l lớp được biểu diễn bởi $\mathcal{G} = \{G_1 = (V_1, E_1), G_2 = (V_2, E_2), \dots, G_l = (V_l, E_l)\}$, trong đó mỗi phần tử bao gồm một đồ thị có hướng $G_i = (V_i, E_i)$. Nếu một nút tồn tại ở nhiều lớp, nút này được thêm vào tập hợp các nút trùng lặp của mạng đa lớp $\mathcal{G}$. Để không mất tính tổng quát, chúng tôi giả sử mỗi lớp của mạng đa lớp có cùng số lượng nút. Do đó, nếu một nút $v \in G_i$ không thuộc $G_j (i \neq j)$, chúng tôi thêm nút này vào G_j như một nút cô lập. Sau đó, đối với mỗi nút, các cạnh liên lớp được thêm vào để kết nối các bản sao liên lớp liền kề của nó trên tất cả các mạng đa lớp. Cuối cùng, chúng tôi coi tập hợp tất cả các nút của mạng đa lớp là $V = \bigcup_{i=1}^l V_i$. Trong đề án này, vì chúng tôi cho phép các lớp khác nhau của mạng đa lớp tuân theo các mô hình lan truyền ảnh hưởng khác biệt, nên cần thiết phải định nghĩa một mô hình toán học cho sự lan truyền trên mạng $\mathcal{G}$.

Định nghĩa 1 (Lan tỏa ảnh hưởng): Cho một lớp $G_i = (V, E_i)$, chúng tôi định nghĩa một tập hợp hạt giống $S \subseteq V$. Hàm δ_i đại diện cho một mô hình lan truyền ảnh hưởng trong G_i , ánh xạ từ tập lũy thừa của V tới các số thực không âm, $\delta_i: 2^V \rightarrow \mathbb{R}_{\geq 0}$. Lan tỏa ảnh hưởng, được định nghĩa là số nút kỳ vọng bị ảnh hưởng bởi tập hợp hạt giống S , được ký hiệu là $\delta_i(S)$ và được tính như sau:

$$\delta_i(S) = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{j=1}^m |T_j| \quad (2.1)$$

trong đó T_j đại diện cho tập hợp các nút đã được kích hoạt cuối cùng $T_j \subset V$ với tập hạt giống S tại bước mô phỏng thứ j . Mô phỏng tiếp tục cho đến khi không còn nút nào được kích hoạt hoặc cho đến khi đạt đến số bước Monte Carlo tối đa m . Tăng giá trị m sẽ cải thiện độ chính xác của việc ước lượng số nút bị ảnh hưởng. Phương pháp này áp dụng được cho hầu hết các mô hình lan truyền, bao gồm mô

hình Lan truyền độc lập (Independent Cascade - IC) và Ngưỡng tuyến tính (Linear Threshold - LT)[47].



Hình 2. 3: Một ví dụ minh họa tính chất "kích hoạt trùng lặp" độc đáo của sự lan truyền ảnh hưởng trong một mạng đa lớp.

Hai lớp G_1 và G_2 có các mô hình khuếch tán tương ứng của chúng là LT và IC. Các nút màu cam là nút hạt giống, các nút màu hồng bị lây nhiễm, và các nút màu xanh lá cây là các nút được kích hoạt do trùng lặp. Nếu lan truyền diễn ra độc lập trong mỗi lớp, nút v_6 của G_1 sẽ không hoạt động do ngưỡng cao của nó là 0.7, nghĩa là nó cần ít nhất 70% hàng xóm bị kích hoạt để trở nên hoạt động. Tuy nhiên, trong mạng đa lớp, nút trùng lặp v_5 của G_1 cũng được kích hoạt do kích hoạt xác định từ G_2 , đáp ứng yêu cầu kích hoạt của nút v_6 trong G_1 . Do đó, tập hợp các nút bị lây nhiễm tổng thể của mạng đa lớp, không tính các nút giống nhau ở các lớp khác nhau, giờ đây là $(v_2, v_4, v_5, v_6, v_8)$.

Tiếp theo, chúng ta định nghĩa mô hình lan truyền ảnh hưởng tổng thể δ trong mạng đa lớp $\mathcal{G}$. Đầu tiên, khi một nút trùng lặp v được kích hoạt trong một lớp đồ thị G_i , các bản sao liên lớp tương ứng của nó trong các lớp khác cũng sẽ được kích hoạt một cách xác định. Hiện tượng này được gọi là "kích hoạt trùng lặp" [3] [34] và được

minh họa trong Hình 2.3. Thứ hai, khi tính số lượng kỳ vọng các nút bị ảnh hưởng trong toàn bộ mạng đa lớp $\mathcal{G}$, chúng ta coi các nút trùng lặp là một thực thể duy nhất thay vì đếm chúng nhiều lần. Điều này có nghĩa là khi đếm các nút bị ảnh hưởng trên tất cả các lớp, chúng ta không cộng thêm các bản sao trùng lặp do kích hoạt trùng lặp gây ra. Do đó, ảnh hưởng tổng thể $\delta(S)$ kết hợp ảnh hưởng độc lập từ mỗi lớp đồng thời tính đến các kích hoạt trùng lặp:

$$\delta(S) = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{j=1}^m \left| \bigcup_{i=1}^l T_{ij}(S) \right| \quad (2.2)$$

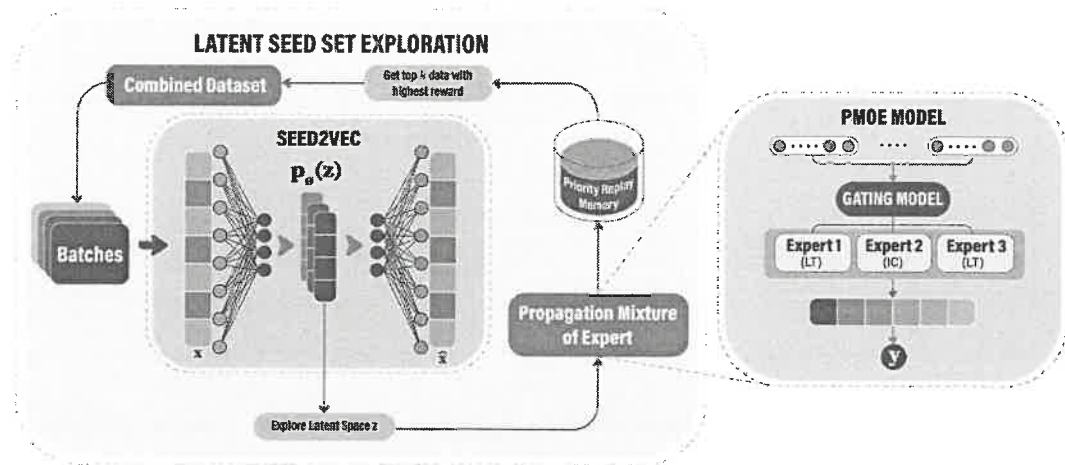
trong đó $T_{ij} \subset V$ đại diện cho tập các nút đã được kích hoạt cuối cùng trong lớp i với tập hạt giống S tại bước mô phỏng thứ j . Giờ đây, chúng ta đã sẵn sàng định nghĩa bài toán MIM như sau:

Định nghĩa 2 (Tối đa hóa Lan tỏa Ảnh hưởng trong Mạng Đa Lớp (MIM)): Cho một đồ thị đa lớp $\mathcal{G} = (G_1 = (V, E_1), \delta_1), \dots, (G_l = (V, E_l), \delta_l)$ và một ngân sách $b \in \mathbb{N}$. Cụ thể, tập hạt giống S được biểu diễn dưới dạng một vector nhị phân $\mathbf{x} \in \mathbb{R}^{1 \times |V|}$, trong đó mỗi phần tử x_j tương ứng với một nút v_j trong V . Cụ thể, $x_j = 1$ nếu v_j được đưa vào tập hạt giống, và $x_j = 0$ nếu không. Giả sử chúng ta có một tập dữ liệu huấn luyện gồm các cặp chỉ số tập hạt giống (x, y) , trong đó x đại diện cho một tập hạt giống và $y = \delta(\mathbf{x})$ là số nút bị ảnh hưởng tổng cộng tương ứng. Bài toán MIM yêu cầu chúng ta tìm một tập hợp nút hạt giống tối ưu $\tilde{\mathbf{x}}$ có kích thước không vượt quá b để tối đa hóa lan tỏa ảnh hưởng tổng thể $\delta(\mathbf{x})$ được tính trong mạng đa lớp. Bài toán này được mô tả như sau:

$$\tilde{\mathbf{x}} = \arg \max_{|\mathbf{x}| \leq b} \delta(\mathbf{x}) \quad (2.3)$$

Đối với mỗi lớp $G_i \in \mathcal{G}$, nhiều thuật toán tham lam [21] [2] [52] đã đạt được bảo đảm hiệu năng $(1 - 1/e)$, nếu δ_i là hàm con mô đun và đơn điệu tăng [13]. Nếu tất cả δ_i của mọi G_i thỏa mãn tính chất Tổng quát Định hình con mô đun (GDS), thì δ là hàm con mô đun [3].

2.2. Khung mô hình REM cho bài toán tối đa hóa ảnh hưởng trên mạng đa tầng



Hình 2. 4: Sơ đồ mô tả quy trình của REM trong việc giải quyết bài toán MIM.

Đầu tiên, REM sử dụng Seed2Vec để nhúng biểu diễn phức tạp của tập hạt giống vào không gian liên tục và ít nhiễu hơn. Sau đó, REM khám phá và tạo ra các tập hạt giống đa dạng từ không gian ẩn này. REM kiểm soát chất lượng của việc tạo tập hạt giống thông qua PMoE, một mô hình có khả năng học và dự đoán chính xác sự lan truyền của tập hạt giống trong mạng multiplex quy mô lớn. Sau khi các tập tổng hợp được tạo ra, chúng được lưu trữ trong bộ nhớ ưu tiên PRM. Để tránh hiện tượng sụp đổ mô hình hoặc quên lãng thảm khốc, các tập hạt giống hàng đầu, được PMoE dự đoán có đóng góp lớn nhất cho sự lan truyền trong mạng multiplex, được kết hợp với tập dữ liệu ban đầu để tạo thành tập dữ liệu mới nhằm huấn luyện lại mô hình. Quá trình này củng cố khả năng sản xuất các tập hạt giống chất lượng cao hơn trong các lần lặp lại sau.

Mô hình REM giải quyết các thách thức đã nêu bằng cách áp dụng các khái niệm được minh họa trong Hình 2.4. Đầu tiên, thay vì tối ưu hóa tập hạt giống (seed set) trong một không gian phức tạp và rời rạc, REM sử dụng Seed2Vec, một mô hình dựa trên Variational Autoencoder (VAE) [15]. VAE là một khung tạo sinh giúp mã hóa dữ liệu vào không gian liên tục trong khi bảo toàn cấu trúc ý nghĩa, cho phép biểu

diễn các tập hạt giống phức tạp dưới dạng ít nhiễu hơn. Điều này giúp tối ưu hóa và tạo ra các giải pháp tiềm năng mới trong không gian đó. Nhận ra rằng Seed2Vec chỉ nắm bắt và tạo ra các giải pháp trong phân phối đặc trưng của dữ liệu huấn luyện ban đầu, khung làm việc của chúng tôi coi Seed2Vec như một tác nhân RL (Reinforcement Learning). Tác nhân này khám phá và khai thác các biểu diễn ẩn đa dạng trong mỗi giai đoạn huấn luyện. Với mỗi mẫu ẩn được Seed2Vec tạo ra, chúng tôi áp dụng Propagation Mixture of Experts (PMoE) để dự đoán khả năng lan truyền của nó với độ chính xác cao, sắp xếp và lưu trữ nó trong Priority Replay Memory (PRM) [12], một cấu trúc được thiết kế để ưu tiên các mẫu quan trọng dựa trên giá trị dự đoán nhằm nâng cao hiệu quả học. Sau đó, chúng tôi chọn ra k mẫu tốt nhất từ PRM và kết hợp chúng với tập dữ liệu ban đầu để tạo thành một tập dữ liệu mới. Cuối cùng, REM sử dụng tập dữ liệu kết hợp này để huấn luyện lại mô hình Seed2Vec, cải thiện khả năng tạo ra tập hạt giống chất lượng cao. Chi tiết thuật toán được trình bày như sau:

Algorithm 1: REM Framework

Input: Multiplex graph G , budget b and the original seed set dataset X_0 with N instances.

Output: Optimal seed set $\tilde{x}$.

- 1: Initialize Seed2Vec model $\mathcal{F}_\theta$ with encoder $\mathcal{E}_\psi$ and decoder $\mathcal{D}_\phi$.
- 2: Initialize PMoE model $\mathcal{P}(\mathbf{x}, \mathcal{G}; \xi)$.
- 3: Initialize the number of steps T and the number of iterations η , learning rate α .
- 4: Initialize Priority Replay Memory (PRM).
- 5: **for** $t = 0, \dots, T$ **do**
- 6: $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- 7: **for** $i = 0, \dots, \eta$ **do**
- 8: $\mathbf{z} \leftarrow \mathbf{z} - \alpha \cdot \nabla \mathcal{L}^{Explore}(\mathbf{z})$.
- 9: Reconstruct seed set: $\hat{\mathbf{x}} = \mathcal{D}_\phi(\mathbf{z})$.

```

10:   Store  $(\hat{\mathbf{x}}, \mathcal{P}(\mathcal{D}_\phi(\mathbf{z}), \mathcal{G}; \xi))$  in PRM.
11:   end for
12:   Sample top  $k$  seed sets  $S_t^{<k}$  from PRM based on  $\mathcal{P}$ .
13:   Combined dataset  $\mathbf{X}_t = S_t^{<k} \cup \mathbf{X}_0$ 
14:   Retrain  $\mathcal{F}_\theta$  and  $\mathcal{P}$  on combined dataset.
15: end for
16:  $\tilde{\mathbf{x}} := \text{REM PREDICTION}(\phi, \xi, \mathcal{G})$ 
17: Return:  $\tilde{\mathbf{x}}$ 

```

Algorithm 2: REM PREDICTION

Input: Decoder parameters ϕ , PMoE parameters ξ , Multiplex graph $\mathcal{G}$.

Output: Optimal seed set $\tilde{\mathbf{x}}$.

```

1: Initialize the number of iterations  $\eta$ , learning rate  $\beta$ .
2: Initialize random latent representation  $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .
3: for  $i = 0, \dots, \eta$  do
4:    $\mathbf{z} \leftarrow \mathbf{z} + \beta \cdot \nabla_{\mathbf{z}} \mathcal{P}(\mathcal{D}_\phi(\mathbf{z}), \mathcal{G}; \xi)$ 
5: end for
6:  $\tilde{\mathbf{x}} = \mathcal{D}_\phi(\mathbf{z})$ 
7: Return:  $\tilde{\mathbf{x}}$ 

```

2.2.1. Mô hình học biểu diễn vector cho tập hạt giống phức tạp

Để tối ưu hóa và xác định tập hạt giống chất lượng trong mạng multiplex, chúng tôi đề xuất đặc trưng hóa xác suất của một tập hạt giống, ký hiệu là $p_\phi(\mathbf{x})$, dựa trên đồ thị multiplex $\mathcal{G}$. Việc học $p_\phi(\mathbf{x})$ có thể cung cấp những hiểu biết sâu sắc về bản chất của tập hạt giống, từ đó tạo điều kiện thuận lợi cho việc khám phá các tập hạt giống trong không gian tìm kiếm. Tuy nhiên, học xác suất này là một thách thức vì các mối liên kết giữa các nút trong mỗi tập hạt giống và sự tương quan cao của chúng dựa trên cấu trúc mạng $\mathcal{G}$. Những kết nối phức tạp này khiến việc giải mã các

mối quan hệ giữa các nút trong tập hạt giống khó khăn hơn so với các bài toán tổ hợp tương tự khác.

Do đó, thay vì học trực tiếp biểu diễn phức tạp của $\mathbf{x}$, chúng tôi học một biểu diễn ẩn $\mathbf{z}$ bằng cách sử dụng Variational Autoencoder (VAE) [12], được ký hiệu là $\mathcal{F}_\theta$. Để tiện lợi, chúng tôi chia nhỏ mô hình VAE $\mathcal{F}_\theta$ thành hai mô hình: Bộ mã hóa (Encoder) ký hiệu là $\mathcal{E}_\psi$ và Bộ giải mã (Decoder) ký hiệu là $\mathcal{D}_\phi$. Cụ thể:

$$\mathcal{F}_\theta = \mathcal{E}_\psi \circ \mathcal{D}_\phi, \quad \hat{\mathbf{x}} = \mathcal{F}_\theta(\mathbf{x}) = \mathcal{D}_\phi(\mathcal{E}_\psi(\mathbf{x})) = \mathcal{D}_\phi(\mathbf{z}) \quad (2.4)$$

trong đó $\hat{\mathbf{x}} \in [0,1]^{1 \times |\mathcal{V}|}$ là tập hạt giống được tái tạo.

Cụ thể, để sinh $\mathbf{x}$, $\mathcal{F}_\theta$ giả định sự tồn tại của một biến ngẫu nhiên ẩn $\mathbf{z} \in R^{1 \times s}$, trong đó s đại diện cho chiều của các biến trong $\mathbf{z}$. Biến ẩn này nắm bắt các đặc trưng của tập hạt giống ban đầu và tuân theo phân phối ẩn $p_\phi(\mathbf{z})$. Quá trình tạo sinh hoàn chỉnh được mô tả bởi phương trình:

$$p_\phi(\mathbf{z} | \mathbf{x}) = \frac{p_\phi(\mathbf{x} | \mathbf{z})p_\phi(\mathbf{z})}{p_\phi(\mathbf{x})} \quad (2.5)$$

Tuy nhiên, việc tính toán giá trị chính xác của $p_\phi(\mathbf{x}) = \int \dots \int p_\phi(\mathbf{x}, \mathbf{z}), d_{\mathbf{z}_0} \dots d_{\mathbf{z}_v}$ không khả thi, làm cho phương trình trở nên thách thức về tính toán. Để giải quyết vấn đề này, $\mathcal{E}_\psi$ sẽ học p_ϕ , một phân phối hậu nghiệm xấp xỉ của $p_\phi(\mathbf{z} | \mathbf{x})$. Mục tiêu là xấp xỉ phân phối hậu nghiệm không khả thi bằng một phân phối đơn giản hơn $q_\psi(\mathbf{z} | \mathbf{x})$ với tập hạt giống $\mathbf{x}$. Nói cách khác, mục tiêu là có $p_\phi(\mathbf{z} | \mathbf{x}) \approx q_\psi(\mathbf{z} | \mathbf{x})$.

Điều này dẫn đến giới hạn dưới của bằng chứng (ELBO) để huấn luyện mô hình bằng reparameterization trick và SGD [15]:

$$\mathcal{L}^{ELBO} = \mathbb{E}_{q_\psi}[\log p_\phi(\mathbf{z} | \mathbf{x})] - \mathbb{E}_{q_\psi}[\log q_\psi(\mathbf{z} | \mathbf{x})] \quad (2.6)$$

$$\begin{aligned}
&= \mathbb{E}_{q_\psi}[\log p_\phi(\mathbf{x} | \mathbf{z})] - \mathbb{E}_{q_\psi}[\log p_\phi(\mathbf{z})] - \mathbb{E}_{q_\psi}[\log q_\psi(\mathbf{z} | \mathbf{x})] \\
&= \mathbb{E}_{q_\psi}[\log p_\phi(\mathbf{x} | \mathbf{z})] - \mathbb{E}_{q_\psi} \left[\log \frac{q_\psi(\mathbf{z} | \mathbf{x})}{p_\phi(\mathbf{z})} \right]
\end{aligned}$$

Lưu ý rằng chúng tôi mô hình hóa $p_\phi(\mathbf{z})$ dưới dạng phân phối Gaussian $\mathcal{N}(\mu, \sigma^2)$, trong đó μ và σ là các siêu tham số được xác định.

Trong khung làm việc của Variational Autoencoder (VAE), mục tiêu chính là tối đa hóa Evidence Lower Bound (ELBO) vốn đóng vai trò như một đại lượng xấp xỉ cho xác suất hợp lý của dữ liệu. ELBO bao gồm hai thành phần chính: Tổn thất tái tạo (Reconstruction Loss) và Độ phân kỳ KL (KL Divergence).

$$\mathcal{L}^{ELBO} = \mathbb{E}_{q_\psi}[\log p_\phi(\mathbf{x} | \mathbf{z})] - \mathbb{E}_{q_\psi} \left[\log \frac{q_\psi(\mathbf{z} | \mathbf{x})}{p_\phi(\mathbf{z})} \right] \quad (2.7)$$

Trong đó:

- Tổn thất tái tạo đo lường sự khác biệt giữa tập dữ liệu ban đầu $\mathbf{x}$ và kết quả tái tạo $\hat{\mathbf{x}}$.
- Độ phân kỳ KL điều chỉnh phân phối không gian tiềm ẩn $q_\psi(\mathbf{z} | \mathbf{x})$ gần với phân phối tiên nghiệm $p_\phi(\mathbf{z})$.

Quá trình tối thiểu hóa tổn thất tái tạo là nhiệm vụ của Decoder (bộ giải mã) được tham số hóa bởi ϕ . Decoder nhận đầu vào là $\mathbf{z}$ và cố gắng tạo ra tái tạo $\hat{\mathbf{x}}$, sao cho càng gần với dữ liệu gốc càng tốt, do đó giảm thiểu tổn thất tái tạo. Để huấn luyện hiệu quả VAE, Tổn thất bình phương trung bình (MSE) được sử dụng làm tổn thất tái tạo. MSE trực tiếp định lượng sự khác biệt giữa đầu vào gốc $\mathbf{x}$ và đầu ra tái tạo $\hat{\mathbf{x}}$, làm cho nó trở thành một hàm tổn thất đơn giản và được sử dụng phổ biến. Công thức MSE như sau:

$$\text{MSE Loss} = \frac{1}{N} \sum_{i=1}^N |\hat{\mathbf{x}}_i - \mathbf{x}_i|^2 \quad (2.8)$$

Việc giảm thiểu tổn thất MSE tương ứng với việc tối đa hóa thành phần khả năng $\log p_\phi(\mathbf{x} | \mathbf{z})$ trong ELBO. Điều này là do MSE nhỏ hơn biểu thị rằng đầu ra tái tạo $\hat{\mathbf{x}}$ gần hơn với đầu vào gốc $\mathbf{x}$, dẫn đến xác suất cao hơn của dữ liệu dưới mô hình. Cuối cùng, ELBO có thể được biểu diễn như sau:

$$\mathcal{L}^{ELBO}(\mathbf{x}; \psi, \phi) = \frac{1}{N} \sum_{i=1}^N \text{MSE}(\hat{\mathbf{x}}_i, \mathbf{x}_i) - \text{KL}(q_\psi(\mathbf{z} | \mathbf{x}) | p_\phi(\mathbf{z})) \quad (2.9)$$

Ở đây, tổn thất MSE được tích hợp trực tiếp như là thành phần tái tạo trong ELBO, dẫn dắt quá trình tối ưu hóa nhằm cải thiện khả năng của VAE trong việc tái tạo dữ liệu đầu vào một cách chính xác. Trong khi đó, thành phần phân kỳ KL đảm bảo không gian tiềm ẩn được điều chỉnh về phân phối tiên nghiệm, duy trì sự cân bằng giữa chất lượng tái tạo và điều chỉnh không gian tiềm ẩn.

2.2.2. Khám phá tập hạt giống tiềm ẩn

Là một mô hình sinh, Seed2Vec chỉ có thể tạo ra các tập hạt giống chất lượng nếu dữ liệu huấn luyện ban đầu giàu đặc trưng. Nếu dữ liệu thiên lệch về các đặc trưng chiếm ưu thế hoặc thiếu sự đa dạng, Seed2Vec có thể bỏ lỡ những đặc trưng quan trọng nhưng ít xuất hiện. Khi mạng đa kênh trở nên phức tạp hơn và số lượng nút tăng lên, mô hình có xu hướng ưu tiên các nút hạt giống chiếm ưu thế trong tập dữ liệu, thường bỏ qua các nút ít phổ biến nhưng có thể quan trọng. REM khắc phục điều này bằng cách xem Seed2Vec như một tác nhân RL, chủ động khám phá các tập hạt giống mới và tiềm năng có ảnh hưởng lớn, tối đa hóa sự lan tỏa để huấn luyện lại và củng cố chính nó thông qua định lý sau:

Bổ đề 1 (Tương đương Tối đa Hóa Entropy Tiềm Ẩn): Giả sử mô hình Seed2Vec đã hội tụ, ta có $\arg \max_{\mathbf{z}} \mathcal{H}(\mathcal{D}_\phi(\mathbf{z})) \propto \arg \max_{\mathbf{x}} \mathcal{H}(\mathbf{x})$

Entropy của một biến ngẫu nhiên $\mathbf{x}$ được định nghĩa như sau:

$$\mathcal{H}(\mathbf{x}) = - \sum_{i=1}^{|\mathbf{x}|} p(\mathbf{x}_i) \log p(\mathbf{x}_i) \quad (2.10)$$

Tương tự, entropy của biến tiềm ẩn $\mathcal{D}_\phi(\mathbf{z})$ được viết là:

$$\mathcal{H}(\mathcal{D}_\phi(\mathbf{z})) = - \sum_{i=1}^{|\mathbf{x}|} p(\hat{\mathbf{x}}_i) \log p(\hat{\mathbf{x}}_i) \quad (2.11)$$

Trong đó, $\hat{\mathbf{x}} = \mathcal{D}_\phi(\mathbf{z})$ đại diện cho dữ liệu được tái tạo từ biến tiềm ẩn $\mathbf{z}$.

Khi $\mathcal{F}_\theta$ hội tụ, dữ liệu gốc $\mathbf{x}$ và dữ liệu tái tạo $\hat{\mathbf{x}}$ gần như giống nhau, nghĩa là $\mathbf{x} \approx \hat{\mathbf{x}}$. Do $\mathbf{x}$ và $\hat{\mathbf{x}}$ gần như giống nhau, entropy của chúng cũng gần như bằng nhau, $\mathcal{H}(\mathbf{x}) \approx \mathcal{H}(\mathcal{D}_\phi(\mathbf{z}))$. Do đó, việc tối đa hóa entropy của biến tiềm ẩn $\mathbf{z}$ tương đương với việc tối đa hóa entropy của dữ liệu gốc $\mathbf{x}$:

$$\arg \max_{\mathbf{z}} \mathcal{H}(\mathcal{D}_\phi(\mathbf{z})) \propto \arg \max_{\mathbf{x}} \mathcal{H}(\mathbf{x}) \quad (2.12)$$

Theo Định lý 1, việc khám phá không gian tiềm ẩn $\mathbf{z}$ để tìm tập hạt giống mới $\mathcal{S}_t$, với $t = 1, 2, 3, \dots$ là số tập huấn luyện, tỷ lệ thuận với việc khám phá không gian rời rạc $\mathbf{x}$. Sự tương quan này xuất hiện vì mô hình *Seed2Vec* đã được huấn luyện tốt với tập hạt giống ban đầu X_0 , đảm bảo tính liên tục (các điểm gần nhau trong không gian tiềm ẩn giải mã ra nội dung tương tự) và tính hoàn thiện (mọi điểm được lấy mẫu từ phân phối không gian tiềm ẩn đều sinh ra nội dung "có ý nghĩa"). Tại điểm này, $p_\phi(\mathbf{z} | \mathbf{x}) \approx q_\psi(\mathbf{z} | \mathbf{x})$, với $q_\psi(\mathbf{z} | \mathbf{x})$ hội tụ về phân phối Gaussian $\mathcal{N}(\mu, \sigma^2)$ như được chỉ ra trong thuật ngữ thứ hai của Phương trình (9). Một tác nhân RL có thể khám phá các đặc trưng tiềm ẩn khác nhau bằng cách lấy mẫu $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ và tái tạo tập hạt giống $\hat{\mathbf{x}}$ qua Bộ giải mã (tức $\hat{\mathbf{x}} = \mathcal{D}_\phi(\mathbf{z})$). Tuy nhiên, do $q_\psi(\mathbf{z} | \mathbf{x})$ hội tụ về một hàm liên tục có đạo hàm theo $\mathbf{z}$, thay vì lấy mẫu ngẫu nhiên, ta sử dụng Gradient Descent trực tiếp trên $\mathbf{z}$ để giảm thiểu hàm mục tiêu sau:

$$\mathcal{L}^{Explore}(\mathbf{z}) = \mathbb{E} \left(c \cdot \mathcal{H} \left(\mathcal{D}_\phi(\mathbf{z}) \right) + \exp \left(- \mathcal{P} \left(\mathcal{D}_\phi(\mathbf{z}) \right) \right) \right) \quad (2.13)$$

trong đó c là các hệ số. Thuật ngữ $\mathcal{H} \left(\mathcal{D}_\phi(\mathbf{z}) \right) = - \sum_{i=1}^{|\hat{\mathbf{x}}|} p(\hat{\mathbf{x}}_i) \log p(\hat{\mathbf{x}}_i)$ biểu diễn entropy của biến tiềm ẩn, nhằm khuyến khích khám phá các vùng mới trong không gian tiềm ẩn. Hàm $\mathcal{P} \left(\mathcal{D}_\phi(\mathbf{z}) \right)$ chỉ Mixture of Experts (PMoE), được mô tả trong phần tiếp theo, và được dùng để dự đoán tập hạt giống tái tạo $\hat{\mathbf{x}} = \mathcal{D}_\phi(\mathbf{z})$. Để giảm thiểu hàm mục tiêu, ta sử dụng hàm mũ $\exp(\cdot)$ nhằm giảm tác động của $\mathcal{P} \left(\mathcal{D}_\phi(\mathbf{z}) \right)$ khi giá trị của nó tăng lên. Với tập hạt giống mới $\mathcal{S}_t$ (lưu trữ bằng bộ nhớ Replay Ưu tiên [12]) được tối ưu từ Phương trình (9), ta lấy mẫu k giá trị tốt nhất và kết hợp với tập dữ liệu gốc $\mathbf{X}_0$ để tạo Tập dữ liệu Kết hợp $\mathbf{X}_t = \mathcal{S}_t^{(<\mathcal{R})} \cup \mathbf{X}_0$. Do đó, khi tập huấn luyện t tiến triển, ta có thể sử dụng $\mathbf{X}_t$ để huấn luyện lại mô hình Seed2Vec $\mathcal{F}_\theta$. Phương pháp này cho phép $\mathcal{F}_\theta$ tạo ra các tập hạt giống cải thiện trong các lần lặp tiếp theo.

2.2.3. Lan truyền hỗn hợp các chuyên gia

Việc áp dụng Mạng Nơ-ron Đồ Thị (Graph Neural Networks - GNNs) để dự đoán lan truyền trong các mạng đa lớp quy mô lớn với hàng tỷ nút gặp nhiều thách thức do hiện tượng làm mịn quá mức (oversmoothing) [9]. Bên cạnh đó, khi sử dụng một GNN đơn với h lớp, các nút tổng hợp thông tin từ hàng xóm ở khoảng cách h -hop, có khả năng trộn lẫn dữ liệu từ các lớp khác nhau, dẫn đến thiếu chính xác. Để giải quyết vấn đề này, chúng tôi đề xuất Hỗn Hợp Chuyên Gia Lan Truyền (PMoE). Cách tiếp cận này sử dụng nhiều mô hình GNN với độ sâu lớp khác nhau nhằm nắm bắt động học lan truyền một cách hiệu quả. Các nút được định tuyến tới chuyên gia phù hợp nhất dựa trên đặc tính của chúng và độ sâu lan truyền mong muốn, giúp mô hình tập trung vào thông tin liên quan và giảm nhiễu. Phương pháp này cho phép dự đoán lan truyền chính xác và hiệu quả trong các mạng đa lớp quy mô lớn.

Khung PMoE của chúng tôi mô phỏng quá trình lan truyền với tập hạt giống $\mathbf{x}$ và đồ thị đa lớp $\mathcal{G}$. Trong khung này, chúng tôi định nghĩa một tập gồm C "mạng chuyên gia," ký hiệu là $e_1, e_2, \dots, e_C$. Mỗi chuyên gia e_i được triển khai dưới dạng một GNN với độ sâu lớp khác nhau, đầu ra là $e_i(\mathbf{x}, \mathcal{G}, \xi_i) \in [0, 1]^{1 \times |V|}$, một vector biểu diễn xác suất nhiễm bệnh ước tính của từng nút trong $\mathcal{G}$, với ξ_i là tham số của chuyên gia thứ i . Để tận dụng hiệu quả kiến thức đa dạng từ các chuyên gia, chúng tôi sử dụng mạng định tuyến R , cho đầu ra là một phân phối xác suất trên các chuyên gia $R(\mathbf{x}) \in \mathbb{R}^{1 \times C}$ dựa trên tập hạt giống $\mathbf{x}$. Mỗi phần tử trong phân phối này tương ứng với xác suất liên quan của một chuyên gia cụ thể đối với đầu vào.

Lấy cảm hứng từ cơ chế định tuyến top m nhiều được đề xuất bởi [33], chúng tôi chọn m chuyên gia liên quan nhất cho mỗi đầu vào. Cơ chế này hoạt động như sau:

$$Q(\mathbf{x}) = \mathbf{x}\xi_g + \epsilon \cdot \text{Softplus}(\mathbf{x}\xi_n), \quad (2.14)$$

$$R(\mathbf{x}) = \text{Softmax}(\text{TopM}(Q(\mathbf{x}), m)), \quad (2.15)$$

Trong phương trình trên, $\epsilon \sim \mathcal{N}(0, 1)$ là nhiễu Gaussian chuẩn. Các tham số ξ_g và ξ_n là trọng số học được, điều chỉnh đóng góp của các điểm số sạch và nhiễu, tương ứng. Giá trị kỳ vọng $\mathcal{M}(\mathbf{x}, \mathcal{G}; \xi)$, trong đó $\xi = [\xi_g, \xi_n, \xi_1, \dots, \xi_C]$ biểu diễn tham số của mô hình PMoE $\mathcal{M}$, được tính dựa trên đầu ra của tất cả các chuyên gia và có thể được biểu diễn như sau:

$$\mathcal{M}(\mathbf{x}, \mathcal{G}; \xi) = e(\mathbf{x}, \mathcal{G}; \xi_e, \xi_g, \xi_n) = \sum_{i=1}^C R_i(\mathbf{x}) e_i(\mathbf{x}, \mathcal{G}; \xi_i) \quad (2.16)$$

Ở đây, $R_i(\mathbf{x})$ là phần tử thứ i của mạng định tuyến $R(\mathbf{x})$, biểu diễn xác suất liên quan của chuyên gia thứ i trong việc dự đoán ảnh hưởng của tập hạt giống $\mathbf{x}$. Trong trường hợp này, tổng số nút bị nhiễm, ký hiệu là $\hat{y} \in \mathbb{R}_+$, được tính là $\hat{y} =$

$\mathcal{P}(\mathbf{x}, \mathcal{G}; \xi) = g(\mathcal{M}(\mathbf{x}, \mathcal{G}; \xi); \zeta)$. Ở đây, $g(\cdot)$ là một hàm chuẩn hóa (ví dụ: chuẩn $l-1$) và ζ là ngưỡng để chuyển đổi xác suất thành giá trị rời rạc.

Bổ đề 2 (Tính đơn điệu của Mô hình PMoE): *Giả sử mô hình PMoE được huấn luyện tới hội tụ và trong giai đoạn suy luận, các điểm số nhiễu ξ_n không được xem xét, đối với bất kỳ GNN nào, $\mathcal{P}$ là đơn điệu nếu hàm tổng hợp và hàm kết hợp trong GNN không giảm.*

Chứng minh:

Giả sử chúng ta có một mạng nơ-ron đồ thị (Graph Neural Network - GNN) với H tầng, trong đó $\mathcal{A}^h$ và C^h là các hàm không giảm, ký hiệu là $e(\cdot)$. Đầu vào là một vector $\mathbf{x}$, và chúng ta áp dụng GNN lên $\mathbf{x}$ qua H tầng như sau:

1. *Định nghĩa đầu vào:* Ban đầu, xét đầu vào $r_v^{(0)}$ bằng $\mathbf{x}$ cho mỗi nút v trong đồ thị, nghĩa là tất cả các nút bắt đầu với vector đặc trưng ban đầu $\mathbf{x}$.

2. *Lặp qua các tầng:* Với mỗi tầng h từ 1 đến H , hàm tổng hợp được áp dụng cho mỗi nút v như sau:

$$e(\mathbf{x}) = \mathcal{A}^1 \circ (C^1 \circ \mathcal{A}^2 \circ C^2 \dots \circ \mathcal{A}^H \circ C^H) \quad (2.17)$$

Vì $\mathcal{A}^h$ và C^h là các hàm không giảm, nên $\mathcal{A}^1 \circ C^1 \dots \circ \mathcal{A}^H \circ C^H$, tức là $e(\mathbf{x})$, cũng là một hàm không giảm. Do đó, ta có $e(\mathbf{x})$ là một hàm không giảm.

Tiếp theo, chúng ta sẽ chứng minh rằng quá trình lan truyền được tổng hợp từ một mô hình Mixture of Experts (PMoE) cũng là một hàm không giảm, với điều kiện mô hình đã hội tụ và không có nhiễu trong việc chọn chuyên gia. Nhớ lại rằng quá trình lan truyền $\mathcal{M}(\mathbf{x}, \mathcal{G}; \xi)$ được tính cho bất kỳ tập hạt giống $\mathbf{x}$ nào bằng Eq. 10. Trong thiết lập của chúng ta, khi chỉ xét các chuyên gia không nhiễu, thì $\mathcal{M}(\mathbf{x}, \mathcal{G}; \xi)$ không phụ thuộc vào ξ_n . Ta có thể viết lại đầu ra của mô hình như sau:

$$Q(\mathbf{x}, \xi_g, \xi_n) = \mathbf{x}\xi_g + \epsilon \cdot \text{Softplus}(\mathbf{x}\xi_n) \quad (2.18)$$

$$\begin{aligned}
&= \mathbf{x}\xi_g + 0 \cdot \text{Softplus}(\mathbf{x}\xi_n) \\
&= \mathbf{x}\xi_g \\
\Rightarrow R(\mathbf{x}, \xi_g, \xi_n) &= R(\mathbf{x}, \xi_g) \\
&= \text{Softmax}\left(\text{TopM}(Q(\mathbf{x}, \xi_g), m)\right)
\end{aligned} \tag{2.19}$$

Do đó, ta có:

$$\Rightarrow \mathcal{M}(\mathbf{x}, \mathcal{G}; \xi) = \sum_{i=1}^c R_i(\mathbf{x}, \xi_g) e_i(\xi_i) \tag{2.20}$$

Vì $R(\cdot)$ là toán tử softmax, một hàm không giảm, và $e(\cdot)$ cũng là một hàm không giảm, suy ra mô hình PMoE $\mathcal{M}(\mathbf{x}, \mathcal{G}; \xi)$ là hàm đơn điệu. Kết quả là, vì hàm $g(\cdot)$ cũng không giảm, hàm lan truyền ảnh hưởng $\mathcal{P}(\mathbf{x}, \mathcal{G}; \xi)$ cũng là hàm đơn điệu.

Bổ đề 1 chỉ ra rằng mô hình PMoE $\mathcal{P}(\mathbf{x}, \mathcal{G}; \xi)$ có bảo đảm lý thuyết để duy trì tính đơn điệu, và mục tiêu học của mô hình $\mathcal{P}(\mathbf{x}, \mathcal{G}; \xi)$ được đưa ra như sau:

$$\max_{\xi} \mathbb{E}[p_{\xi}(y|\mathbf{x}, \mathcal{G})] \tag{2.21}$$

Mục tiêu Học Tập Tổng Thể

Cuối cùng, để kết nối học biểu diễn, khám phá tập hạt giống tiềm ẩn và huấn luyện mô hình khuếch tán, ta giảm thiểu hàm mục tiêu tổng thể sau, kết hợp Eq. (9), (13), (21):

$$\mathcal{L}_{train} = \mathbb{E}[\mathcal{L}^{ELBO}(\theta) + \mathcal{L}^{PMoE}(\xi) + \mathcal{L}^{Explore}(\mathbf{z})] \tag{2.22}$$

trong đó $\mathcal{L}^{PMoE} = (\hat{y} - y)^2$

Suy Diễn Tập Hạt Giống

Cuối cùng, phương pháp của chúng tôi kết thúc với việc suy diễn tập hạt giống từ không gian tiềm ẩn liên tục. Cụ thể, ta sử dụng gradient ascent để tìm đại diện tiềm

ẩn $\tilde{\mathbf{z}}$ tối đa hóa sự lan tỏa ảnh hưởng dự đoán, dựa trên ước lượng do mô hình PMoE cung cấp. Đại diện $\tilde{\mathbf{z}}$ được giải mã qua mạng giải mã của Seed2Vec để thu được tập hạt giống tối ưu $\tilde{\mathbf{x}}$.

Định lý 3 (Tính nhất quán của Ước lượng Ảnh hưởng): Cho hai tập hạt giống $\mathbf{x}^{(i)}$ và $\mathbf{x}^{(j)}$, với đại diện tiềm ẩn tương ứng là $\mathbf{z}^{(i)}$ và $\mathbf{z}^{(j)}$ được mã hóa bởi Seed2Vec. Nếu lỗi tái tạo được giảm thiểu trong quá trình huấn luyện và $\mathcal{P}(p_\phi(\mathbf{z}^{(i)}), \mathcal{G}; \xi) > \mathcal{P}(p_\phi(\mathbf{z}^{(j)}), \mathcal{G}; \xi)$, thì suy ra rằng $\mathcal{P}(\mathbf{x}^{(i)}, \mathcal{G}; \xi) > \mathcal{P}(\mathbf{x}^{(j)}, \mathcal{G}; \xi)$.

Chứng minh:

Dựa trên các giả định từ Bổ đề 2, hàm ước lượng ảnh hưởng $\mathcal{P}$ của mô hình PMoE là đơn điệu, nghĩa là với bất kỳ hai giá trị $\mathbf{x}^{(i)} > \mathbf{x}^{(j)}$, thì $\mathcal{P}(\mathbf{x}^{(i)}, \mathcal{G}; \xi) \geq \mathcal{P}(\mathbf{x}^{(j)}, \mathcal{G}; \xi)$. Vì mô hình Seed2Vec tối thiểu hóa lỗi tái tạo, các biểu diễn tiềm ẩn $\mathbf{z}^{(i)}$ và $\mathbf{z}^{(j)}$ bảo toàn các đặc tính quan trọng của tập hạt giống ban đầu tương ứng $\mathbf{x}^{(i)}$ và $\mathbf{x}^{(j)}$. Do đó, nếu $\mathcal{P}(p_\phi(\mathbf{z}^{(i)}), \mathcal{G}; \xi) > \mathcal{P}(p_\phi(\mathbf{z}^{(j)}), \mathcal{G}; \xi)$ trong không gian tiềm ẩn, thứ tự này cũng phải giữ nguyên trong không gian ban đầu, dẫn đến $\mathcal{P}(\mathbf{x}^{(i)}, \mathcal{G}; \xi) > \mathcal{P}(\mathbf{x}^{(j)}, \mathcal{G}; \xi)$. Điều này hoàn thành chứng minh của Định lý 1.

Theo Định lý 1, tập hạt giống tối ưu tối đa hóa ảnh hưởng có thể được tìm thấy bằng cách tối ưu $\mathbf{z}$.

2.3. Kết luận chương

Chương 2 đã tập trung đi sâu vào việc phân tích và định nghĩa bài toán Tối đa hóa Ảnh hưởng Đa tầng (MIM), một vấn đề phức tạp và ngày càng trở nên quan trọng trong bối cảnh người dùng tương tác trên nhiều nền tảng mạng xã hội. Chương đã làm rõ những thách thức đặc thù của MIM so với bài toán trên mạng đơn lớp, bao gồm sự cần thiết phải mô hình hóa động lực lan truyền riêng biệt trên từng tầng mạng và sự tương tác giữa các tầng, điển hình là hiện tượng "kích hoạt trùng lặp".

Để giải quyết những thách thức này, chương đã giới thiệu chi tiết khung mô hình học sâu tiên tiến REM (Reinforced Expert Maximization). REM giải quyết bài

toán thông qua sự kết hợp sáng tạo của nhiều cơ chế. Đầu tiên, Seed2Vec, một mô hình dựa trên Bộ tự mã hóa biến phân, học cách biểu diễn hiệu quả các tập hạt giống phức tạp trong một không gian tiềm ẩn liên tục, ít nhiễu hơn, tạo điều kiện thuận lợi cho việc tối ưu hóa. Vượt ra ngoài việc chỉ học từ dữ liệu huấn luyện ban đầu, Seed2Vec được coi như một tác nhân học tăng cường, chủ động khám phá không gian tiềm ẩn này để tạo ra các tập hạt giống tổng hợp, đa dạng và có khả năng lan truyền cao, một quá trình được hỗ trợ bởi Bổ đề 2 về sự tương đương trong tối đa hóa entropy tiềm ẩn. Để dẫn dắt quá trình khám phá và đánh giá chất lượng của các tập hạt giống, REM giới thiệu Hỗn hợp Chuyên gia Lan truyền (PMoE). Đây là một kiến trúc Mạng nơ-ron đồ thị đa chuyên gia, kết hợp với một mạng định tuyến thông minh, được thiết kế để ước lượng chính xác và hiệu quả các mẫu lan truyền phức tạp trong mạng đa tầng quy mô lớn, với các đảm bảo lý thuyết về tính đơn điệu và nhất quán của ước lượng ảnh hưởng như được trình bày trong Bổ đề 1 và Định lý 3.

Thông qua sự kết hợp chặt chẽ của các thành phần này, cùng với việc sử dụng bộ nhớ ưu tiên (PRM) để lưu trữ và tái sử dụng các tập hạt giống chất lượng cao, và một hàm mục tiêu tổng thể tích hợp các khía cạnh học tập, REM hướng đến việc tối ưu hóa hiệu quả và tạo ra các giải pháp chất lượng cao cho bài toán MIM. Với nền tảng lý thuyết vững chắc và kiến trúc mô hình REM đã được trình bày chi tiết, chương tiếp theo sẽ tập trung vào việc đánh giá thực nghiệm hiệu suất của REM, kiểm chứng tính hiệu quả, khả năng mở rộng và các ưu điểm của giải pháp được đề xuất.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

Tiếp nối việc trình bày chi tiết khung mô hình REM từ Chương 2, Chương 3 này tập trung đánh giá hiệu quả thực nghiệm của phương pháp. Chúng tôi sẽ mô tả thiết lập thí nghiệm, bao gồm dữ liệu, các phương pháp so sánh và độ đo. Sau đó, kết quả về khả năng lan truyền, hiệu quả tính toán sẽ được trình bày và phân tích, cùng với việc đánh giá tác động của các tham số và thành phần cốt lõi của REM, nhằm minh chứng tính ưu việt của mô hình trong bài toán Tối đa hóa Ảnh hưởng Đa tầng.

3.1. Thiết lập thực nghiệm

3.1.1. Dữ liệu và tham số

Mục tiêu chính của chúng tôi là đánh giá tác động của sự lan tỏa ảnh hưởng trong các kịch bản khác nhau trong bài toán Tối đa Hóa Ảnh Hưởng (Influence Maximization - IM). Các thí nghiệm của chúng tôi tập trung vào hai mô hình lan truyền chủ đạo trong IM: mô hình Ngưỡng Tuyến Tính (Linear Threshold - LT) và mô hình Cascade Độc Lập (Independent Cascade - IC).

Hyperparameter Setting. Chúng tôi đã thực hiện các thí nghiệm trên một hệ thống được trang bị bộ xử lý Intel(R) Core i9-13900k, RAM 128 GB và hai GPU Nvidia RTX 4090 với 24GB VRAM mỗi chiếc. Đối với mỗi phương pháp so sánh, chúng tôi thiết lập các siêu tham số theo bài báo gốc của chúng và tinh chỉnh trên từng tập dữ liệu. Đối với cấu hình của mỗi mô hình khuếch tán, chúng tôi sử dụng phiên bản cascade có trọng số của mô hình IC, cụ thể là xác suất lan truyền $p_{u,v} = 1/d_v^{in}$ (d_v^{in} là bậc trong của nút v) cho mỗi cạnh $e = (u, v)$ trên đồ thị G . Đối với mô hình LT, ngưỡng θ được đặt là 0.5 cho mỗi nút v .

Phần này trình bày việc lựa chọn siêu tham số cho REM (Bảng 3.1), tập trung vào hiệu suất mô hình, sự ổn định trong huấn luyện, và hiệu quả. Tốc độ học được điều chỉnh theo độ phức tạp của mô hình: 0.003 cho Variational Autoencoder (VAE) để tăng tốc hội tụ, và 0.001 cho mô hình Propagation Mixture of Experts (PMOE) có tham số hóa cao nhằm đảm bảo sự ổn định. Hệ số trọng số KL divergence của VAE

được đặt là 0.55, cân bằng giữa độ chính xác tái tạo và việc chuẩn hóa không gian tiềm ẩn. Chúng tôi đặt số lượng chuyên gia trong PMOE là 20, nhằm đảm bảo việc nắm bắt toàn diện các quá trình liên tầng, đồng thời giảm thiểu rủi ro bỏ sót các đường lan truyền tiềm năng. Kích thước minibatch được đặt là 256 để cân bằng sự ổn định trong huấn luyện của cả hai mô hình.

Quá trình khám phá tập hạt giống tiềm ẩn được cấu hình cho 30 tập, mỗi tập gồm 400 bước, dẫn đến tổng cộng 12.000 điểm dữ liệu được tạo ra. Khối lượng này đủ để tăng cường bất kỳ kiến trúc đồ thị nào được nghiên cứu. Phương pháp Proximal Policy Optimization (PPO) được sử dụng để huấn luyện chính sách, kết hợp một hệ số entropy và dropout nhằm khuyến khích khám phá và ngăn ngừa hiện tượng overfitting.

Bảng 3. 1: Giá trị các siêu tham số của mô hình.

Tham số	Giá trị
Tốc độ học cho Mô hình VAE	0.003
Tốc độ học cho Mô hình PMOE	0.001
Bộ tối ưu	Adam
Số bước mỗi tập	400
Số tập	30
Kích thước minibatch	256
Trọng số KL	0.55
Số lượng chuyên gia	8
Tỷ lệ dropout	0.2
Hệ số entropy	0.1

Dữ liệu. Các thí nghiệm của chúng tôi sử dụng nhiều tập dữ liệu mạng đa lớp (multiplex network) với các loại tương tác và hệ thống đa dạng. Mạng Celegans Multiplex GPI từ BioGRID [3] (phiên bản 3.2.108) bao gồm các tương tác di truyền trong *Caenorhabditis elegans*, với 6 lớp, 3.879 nút và 8.181 cạnh. Mạng Arabidopsis Multiplex cũng từ BioGRID [38] chi tiết hóa các tương tác di truyền và protein của *Arabidopsis thaliana*, gồm 7 lớp, 6.980 nút và 18.654 cạnh.

Đối với động lực học trên mạng xã hội, Mạng Twitter NYClimateMarch2014 [31] ghi nhận các hoạt động retweet, đề cập và trả lời trong sự kiện People's Climate March, với 3 lớp, 102.439 nút và 353.495 cạnh. Mạng Twitter ParisAttack2015 [39] ghi nhận các tương tác tương tự trong vụ tấn công Paris năm 2015, với 3 lớp, 1.896.221 nút và 4.163.947 cạnh. Chúng tôi cũng sử dụng tập dữ liệu Cora [40], một mạng trích dẫn của 2.708 bài báo khoa học và 7.981 cạnh, để phân tích sự lan tỏa ảnh hưởng trong xuất bản học thuật.

Bảng 3. 2: Thống kê các bộ dữ liệu sử dụng.

Dữ liệu	Số nút	Số cạnh
Cora-ML	2.708	7.981
Celegans	3.879	8.181
Arabidopsis	6.980	18.654
NYClimateMarch2014	102.439	353.495
ParisAttack2015	1.896.221	4.163.947

3.1.2. Phương pháp so sánh

Chúng tôi đánh giá hiệu suất của REM bằng cách so sánh nó với hai loại kỹ thuật tối ưu hóa ảnh hưởng. 1) Các phương pháp truyền thống: ISF (Influential Seed Finder) [3] là một thuật toán tham lam được thiết kế cho tối đa hóa ảnh hưởng trên mạng đa lớp; KSN (Knapsack Seeding of Networks) [3] sử dụng phương pháp ba lô

(knapsack) để tìm các người dùng hạt giống tốt nhất trong một mạng đa lớp. 2) Các phương pháp học sâu: ToupleGDD [45], GCOMB [41], DeepIM [7] là các giải pháp tối đa hóa ảnh hưởng mạng đơn tiên tiến. Đối với mạng đa lớp, phương pháp MIM-Reasoner [34] sử dụng mô hình đồ thị xác suất để nắm bắt động lực học trong mạng đa lớp, sau đó xác định các bộ hạt giống tốt nhất bằng một giải pháp học tăng cường. Chúng tôi cũng đánh giá hiệu suất của 2 biến thể REM khác nhau để chứng minh hiệu quả của phương pháp của chúng tôi. Một phương pháp là REM-NonRL, không sử dụng khám phá các bộ hạt giống và chỉ dựa vào bộ dữ liệu ban đầu để cung cấp giải pháp. Biến thể này cung cấp quan sát về hiệu quả của cấu trúc học tăng cường mà chúng tôi đề xuất. Biến thể còn lại, REM-NonMixture, bỏ qua cấu trúc Mixture of Expert của chúng tôi, nắm bắt sự lan tỏa phức tạp trong mạng đa lớp với một mô hình GNN duy nhất. Biến thể này sẽ nhấn mạnh những lợi thế của các cấu hình phức tạp hơn của chúng tôi. So sánh được thực hiện dựa trên ba chỉ số: tổng ảnh hưởng lan tỏa (số nút được kích hoạt) và thời gian suy diễn (thời gian thực, tính bằng giây).

3.2. Kết quả và đánh giá

3.2.1. Khả năng lan truyền thông tin

Chúng tôi đánh giá hiệu suất của phương pháp REM so với các chiến lược IM khác bằng cách so sánh khả năng tối đa hóa ảnh hưởng trên các tập dữ liệu khác nhau. Trong mỗi trường hợp, các mô hình xác định các nút hạt giống đại diện cho 1%, 5%, 10% và 20% tổng số nút. Chúng tôi mô phỏng quá trình lan tỏa cho đến khi hoàn tất và xác định ảnh hưởng lan tỏa trung bình qua 100 lần lặp. Chúng tôi báo cáo số nút bị nhiễm cuối cùng.

IM dưới mô hình IC. Chúng tôi tiến hành đánh giá so sánh hiệu suất của các phương pháp trên năm bộ dữ liệu đa dạng, bao gồm Cora-ML, Celegans, Arabidopsis, NYClimatMarch2014, và ParisAttack2015, dưới mô hình lan truyền Independent Cascade (IC). Ngân sách cho việc lựa chọn nút hạt giống được thiết lập ở bốn mức khác nhau, tương ứng với 1%, 5%, 10% và 20% tổng số nút trong mỗi mạng, nhằm khảo sát hiệu quả của các phương pháp ở các quy mô ảnh hưởng khác nhau.

Kết quả, được tổng hợp chi tiết trong các Bảng 3.3 đến 3.7, cho thấy một cách nhất quán rằng REM vượt trội hơn tất cả các phương pháp so sánh khác về khả năng lan tỏa ảnh hưởng dưới mô hình IC. Sự ưu việt của REM trở nên đặc biệt rõ rệt khi xét đến các bộ dữ liệu quy mô lớn hơn như NYClimateMarch2014 (minh họa trong Bảng 3.6) và ParisAttack2015 (minh họa trong Bảng 3.7). Trên các mạng này, với ngân sách hạt giống tăng lên, chẳng hạn từ 10% đến 20%, REM không chỉ duy trì vị trí dẫn đầu mà còn nới rộng khoảng cách đáng kể so với các đối thủ cạnh tranh. Điều này minh chứng cho khả năng mở rộng và hiệu quả vượt trội của REM trong các kịch bản phức tạp và quy mô lớn, nơi việc xác định các nút ảnh hưởng tối ưu trở nên đặc biệt thách thức.

Khi phân tích sâu hơn về hiệu suất của các nhóm phương pháp, những hạn chế cố hữu của các giải pháp truyền thống như ISF và KSN trở nên đặc biệt rõ nét. Dù các phương pháp này tỏ ra tương đối hiệu quả trên các bộ dữ liệu đồ thị có quy mô nhỏ hơn, ví dụ như Cora-ML và Celegans, nhưng chúng nhanh chóng bộc lộ những điểm yếu rõ rệt khi phải xử lý các mạng lưới lớn hơn và với ngân sách hạt giống cao hơn. Minh chứng cụ thể là trên các tập dữ liệu NYClimateMarch2014 và ParisAttack2015, ISF và KSN không chỉ thường xuyên gặp phải tình trạng lỗi hết bộ nhớ (được ghi nhận bằng ký hiệu "-"), mà ngay cả khi hoàn thành, kết quả lan truyền của chúng cũng kém hơn đáng kể so với các phương pháp dựa trên học máy. Thực trạng này phản ánh một cách trực tiếp những thách thức căn bản về khả năng mở rộng (scalability) và năng lực mô hình hóa các tương tác lan truyền phức tạp vốn có trong cấu trúc đa tầng mà các thuật toán dựa trên heuristic truyền thống thường khó lòng vượt qua.

Bảng 3. 3: So sánh hiệu suất dưới mô hình lan truyền IC (Cora-ML).

Cora-ML				
	1%	5%	10%	20%
ISF	398.34	778.62	979.87	1368.56

KSN	398.31	778.62	979.10	1366.03
GCOMB	347.11	766.02	976.14	1251.52
ToupleGDD	349.01	721.42	862.42	1132.66
DeepIM	311.52	606.82	826.41	1179.45
MIM-Reasoner	398.22	778.02	978.95	1363.35
REM-NoneRL	321.24	732.84	880.51	1151.68
REM-NonMixture	343.16	736.42	921.24	1301.55
REM	347.34	765.48	965.04	1404.14

Bảng 3. 4: So sánh hiệu suất dưới mô hình lan truyền IC (Celegans).

Celegans				
	1%	5%	10%	20%
ISF	1465.86	2298.01	2571.92	2819.26
KSN	1382.86	2176.32	2335.44	2620.10
GCOMB	1389.63	1896.86	2237.37	2550.61
ToupleGDD	1279.28	1905.23	2117.74	2411.14
DeepIM	1275.66	1527.25	1938.58	2251.53
MIM-Reasoner	1432.39	2199.26	2389.91	2645.24
REM-NoneRL	1301.22	1884.95	2098.16	2451.87
REM-NonMixture	1387.67	2062.79	2304.15	2695.28
REM	1445.16	2278.07	2585.06	2904.03

Bảng 3. 5: So sánh hiệu suất dưới mô hình lan truyền IC (Arabidopsis).

Arabidopsis				
	1%	5%	10%	20%
ISF	2415.04	3140.58	3871.10	4694.16
KSN	2282.72	2941.54	3621.26	4641.34
GCOMB	2315.44	3097.97	3622.08	4547.63
ToupleGDD	2044.67	2856.37	3487.41	4483.27
DeepIM	1993.37	2397.89	3328.33	4073.50
MIM-Reasoner	2396.95	2989.15	3729.39	4621.54
REM-NoneRL	2076.31	2421.13	3399.34	4120.51
REM-NonMixture	2317.39	2982.67	3712.35	4611.27
REM	2430.84	3181.26	3964.45	4701.75

Bảng 3. 6: So sánh hiệu suất dưới mô hình lan truyền IC. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (NYClimateMarch2014)

NYClimateMarch2014				
	1%	5%	10%	20%
ISF	-	-	-	-
KSN	-	-	-	-
GCOMB	2093.32	7228.02	11780.29	16933.89
ToupleGDD	1821.43	6714.98	10231.81	17822.21
DeepIM	1893.02	6409.32	8064.77	14269.49
MIM-Reasoner	2101.86	7387.91	11984.55	21062.87

REM-NoneRL	2057.82	6835.44	8564.63	16021.63
REM-NonMixture	2215.82	7124.29	10034.71	18932.40
REM	2162.83	7465.96	12834.45	23142.83

Bảng 3. 7: So sánh hiệu suất dưới mô hình lan truyền IC. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (ParisAttack2015)

ParisAttack2015				
	1%	5%	10%	20%
ISF	-	-	-	-
KSN	-	-	-	-
GCOMB	114672.50	180977.09	356187.81	587891.16
ToupleGDD	102872.11	171992.43	335298.85	563387.05
DeepIM	83972.39	149237.59	281298.85	480393.42
MIM-Reasoner	129650.48	217291.02	379932.17	608192.57
REM-NoneRL	87109.27	152621.23	298923.41	515237.59
REM-NonMixture	125098.55	207621.48	380274.03	605875.67
REM	147193.96	229769.76	402372.32	637621.59

IM dưới mô hình LT. Chúng tôi đánh giá các phương pháp dưới mô hình lan tỏa LT, với kết quả trong Bảng 3.8-3.12 cho thấy REM liên tục vượt trội hơn các kỹ thuật khác trong việc tối đa hóa số nút bị nhiễm. Điều đáng chú ý là sự vượt trội này của REM không chỉ ổn định trên các bộ dữ liệu khác nhau mà còn trở nên đặc biệt nổi bật khi áp dụng trên các mạng lưới quy mô lớn và với ngân sách hạt giống ở mức cao nhất (20%). Trong những kịch bản thử thách nhất này, khả năng của REM trong

việc xác định các nút hạt giống chiến lược và dự đoán chính xác động lực lan truyền phức tạp của mô hình LT được thể hiện rõ ràng. Cụ thể, trên hai bộ dữ liệu lớn là NYClimateMarch2014 (Bảng 3.11) và ParisAttack2015 (Bảng 3.12), khi ngân sách hạt giống là 20%, REM đã tạo ra một khoảng cách đáng kể. Nó đạt được mức độ lan tỏa ảnh hưởng cao hơn lần lượt khoảng 10% (ví dụ, 98976 của REM so với 96239 của MIM-Reasoner trên NYClimateMarch2014) và 15% (ví dụ, 804469 của REM so với 786819 của MIM-Reasoner trên ParisAttack2015) so với phương pháp cạnh tranh tốt nhất ngay sau đó, điển hình là MIM-Reasoner.

Thành công ấn tượng này không chỉ nhấn mạnh hiệu quả vượt trội của REM trong việc xử lý các mạng lưới phức tạp và điều kiện ngân sách dồi dào dưới mô hình LT, mà quan trọng hơn, nó là một minh chứng rõ ràng cho khả năng tổng quát hóa vượt trội của mô hình. Việc REM duy trì hiệu suất ấn tượng và nhất quán trên cả hai mô hình lan truyền cơ bản và khác biệt nhau là Independent Cascade (IC) và Linear Threshold (LT) cho thấy sự linh hoạt và mạnh mẽ của kiến trúc được đề xuất. Điều này ngụ ý rằng REM không bị "overfit" vào một cơ chế lan truyền cụ thể mà có khả năng thích ứng và tối ưu hóa hiệu quả với các động lực lan truyền đa dạng, một yếu tố then chốt cho ứng dụng thực tiễn trong các mạng xã hội nơi các mô hình lan truyền thực tế có thể không hoàn toàn tuân theo một lý thuyết duy nhất.

Bảng 3. 8: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (Coral-ML)

Coral-ML				
	1%	5%	10%	20%
ISF	381.0	907.0	1392.0	2145.0
KSN	381.0	907.0	1392.0	2145.0
GCOMB	379.0	1005.0	1297.0	2026.0
ToupleGDD	375.0	998.0	1192.0	1989.0
DeepIM	291.0	708.0	1023.0	1881.0
MIM-Reasoner	381.0	907.0	1392.0	2145.0
REM-NoneRL	321.0	811.0	1101.0	1964.0
REM-NonMixture	361.0	1173.2	1212.0	2102.0
REM	376.0	883.0	1281.0	2141.0

Bảng 3. 9: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (Celegans)

Celegans				
	1%	5%	10%	20%
ISF	1530.0	2643.0	3274.0	3857.0
KSN	1272.0	2340.0	2959.0	3637.0
GCOMB	1376.0	2327.0	3006.0	3633.0
ToupleGDD	1289.0	2153.0	2932.0	3569.0

DeepIM	929.0	1762.0	2012.0	2598.0
MIM-Reasoner	1356.0	2340.0	3019.0	3676.0
REM-NoneRL	1043.0	2176.0	2781.0	3246.0
REM-NonMixture	1332.0	2347.0	3118.0	3688.0
REM	1514.0	2668.0	3251.0	3877.0

Bảng 3. 10: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (Arabidopsis)

Arabidopsis				
	1%	5%	10%	20%
ISF	2901.0	4571.0	5686.0	6855.0
KSN	2440.0	3969.0	4934.0	6147.0
GCOMB	2501.0	4247.0	4993.0	6286.0
ToupleGDD	2479.0	4147.0	4725.0	5841.0
DeepIM	1772.0	3232.0	3211.0	3968.0
MIM-Reasoner	2554.0	4133.0	5027.0	6191.0
REM-NoneRL	2176.0	3548.0	4185.0	5742.0
REM-NonMixture	2632.0	4586.0	5089.0	6197.0
REM	2943.0	4894.0	5705.0	6890.0

Bảng 3. 11: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (NYClimateMarch2014)

NYClimateMarch2014				
	1%	5%	10%	20%
ISF	-	-	-	-
KSN	-	-	-	-
GCOMB	5281.0	29974.0	63187.0	74865.0
ToupleGDD	4866.0	26987.0	61827.0	71865.0
DeepIM	3582.0	21246.0	51721.0	57977.0
MIM-Reasoner	6872.0	35833.0	77925.0	96239.0
REM-NoneRL	3976.0	26920.0	55315.0	59764.0
REM-NonMixture	5352.0	30187.0	66872.0	77359.0
REM	7111.0	37417.0	81255.0	98976.0

Bảng 3. 12: So sánh hiệu suất dưới mô hình lan truyền LT. – biểu thị lỗi hết bộ nhớ. (Giá trị tốt nhất được in đậm.) (ParisAttack2015)

ParisAttack2015				
	1%	5%	10%	20%
ISF	-	-	-	-
KSN	-	-	-	-
GCOMB	429978.0	498301.0	571193.0	729745.0
ToupleGDD	420086.0	487925.0	568872.0	739712.0

DeepIM	387129.0	460273.0	538160.0	578724.0
MIM-Reasoner	487871.0	550864.0	629830.0	786819.0
REM-NoneRL	408149.0	468621.0	540824.0	581872.0
REM-NonMixture	449862.0	529562.0	592055.0	692974.0
REM	503994.0	604890.0	651100.0	804469.0

3.2.2. So sánh thời gian chạy

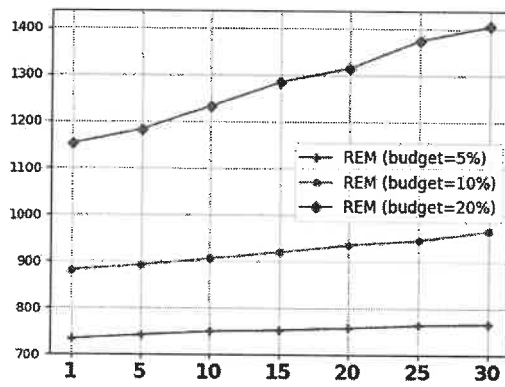
Chúng tôi nghiên cứu thời gian chạy của các bộ hạt giống khi kích thước đồ thị tăng lên của REM so với các giải pháp IM dựa trên học máy khác. Như có thể thấy trong Bảng 3.13, REM thể hiện sự tăng trưởng gần như tuyến tính về thời gian chạy khi kích thước đồ thị tăng lên. Ngoài ra, nó đạt được thời gian suy diễn ngắn hơn chung (trung bình, REM nhanh hơn 10% so với MIM-Reasoner, phương pháp nhanh thứ hai, và nhanh hơn 20% so với DeepIM, phương pháp nhanh thứ ba).

Bảng 3. 13: Thời gian suy luận trung bình (tính bằng giây) theo sự gia tăng số lượng nút với 10% số nút làm các nút khởi đầu.

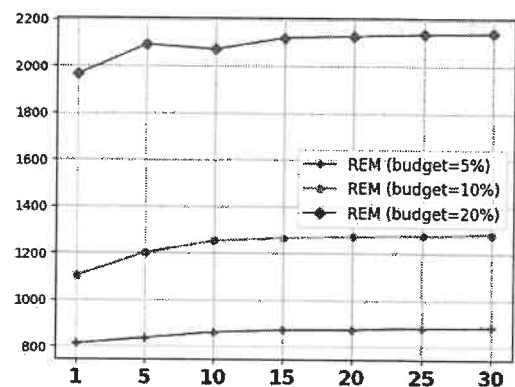
	10,000	20,000	30,000	50,000
GCCOMB	17.894s	30.831s	46.275s	73.983s
ToupleGDD	15.873s	25.321s	37.882s	58.985s
MIM-Reasoner	7.948s	12.532s	26.575s	36.437s
DeepIM	10.321s	19.325s	32.185s	44.871s
REM	8.873s	10.198s	23.404s	33.482s

3.2.3. Ảnh hưởng của số lần khám phá môi trường

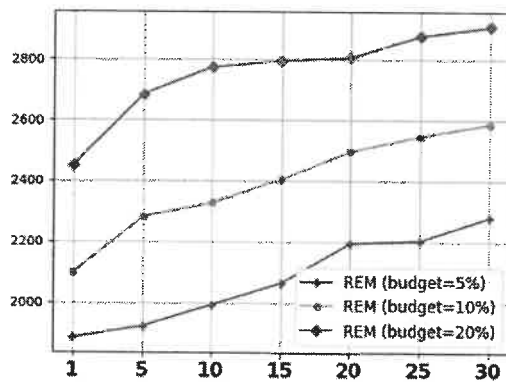
IM với số bước khám phá. Chúng tôi so sánh hiệu quả của việc tăng số bước khám phá dưới các mô hình IC và LT trong điều kiện ngân sách. Như được chỉ ra trong Hình 3.1, việc tăng số bước khám phá thường cải thiện kết quả trên các mạng, đặc biệt là đối với các tập dữ liệu lớn hơn. Đối với các tập dữ liệu nhỏ như Cora-ML, sự khác biệt về hiệu suất là không đáng kể, trong khi đối với các tập dữ liệu lớn hơn, khoảng cách này mở rộng đáng kể khi số bước tăng lên.



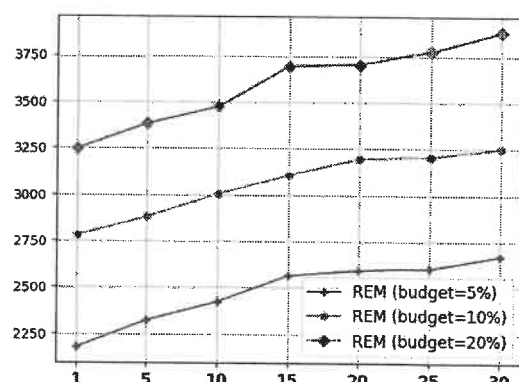
(a) Cora ML-IC



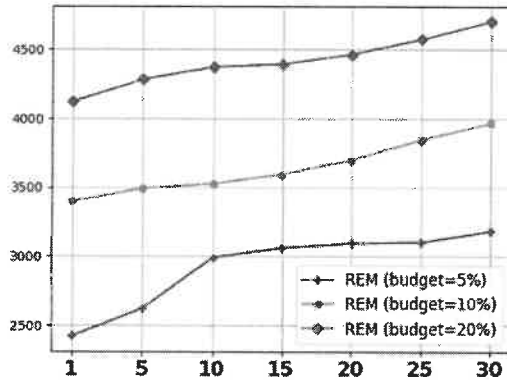
(f) Cora ML-LT



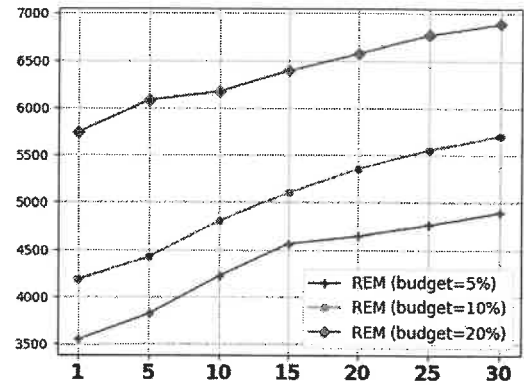
(b) Celegans-IC



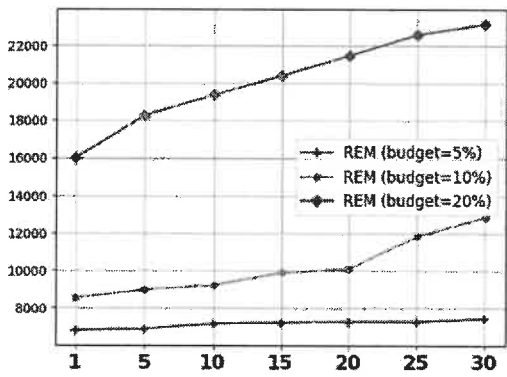
(g) Celegans-LT



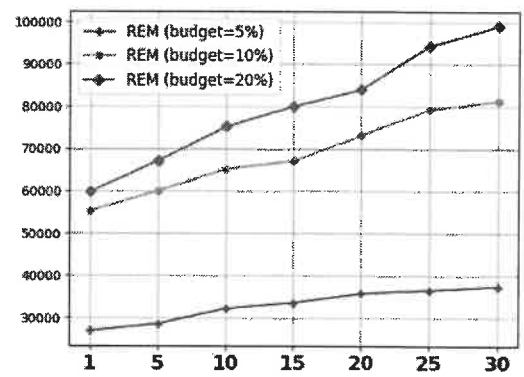
(c) Arabidopsis-IC



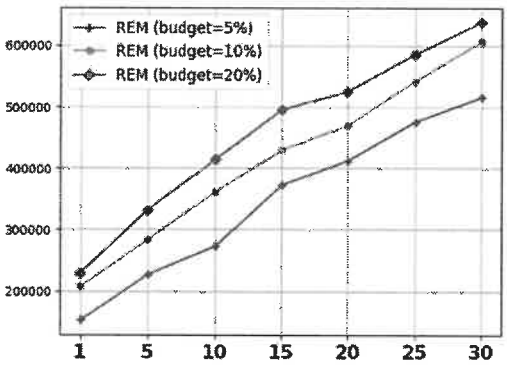
(h) Arabidopsis-LT



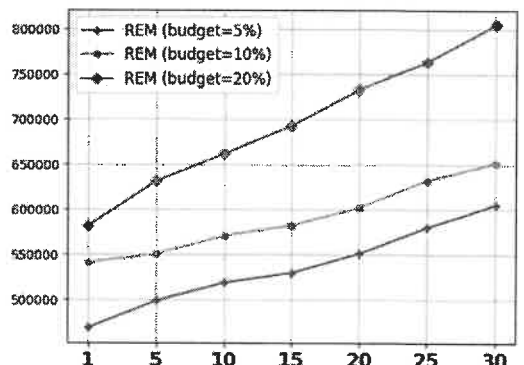
(d) NYClimateMarch-IC



(i) NYClimateMarch-LT



(e) ParisAttack2015-IC



(j) ParisAttack2015-LT

Hình 3. 1: Sự khác biệt trong mức độ lan truyền ảnh hưởng (trục y) của đầu ra REM trên các tập dữ liệu và ngân sách khác nhau khi tăng số bước khám phá (trục x). Hình 3a - 3e và Hình 3f - 3j được đánh giá dưới mô hình IC và LT, tương ứng.

3.2.4. Kết quả các biến thể khác của mô hình

Bảng 3. 14: So sánh giữa việc chọn GCN và GAT làm kiến trúc chuyên gia trong PMoE, với 10% số nút làm ngân sách, dưới cả hai mô hình LT và IC.

	IC		LT	
Methods	GCN	GAT	GCN	GAT
Cora-ML	947.3	965.04	1235.0	1281.0
Celegans	2419.2	2585.06	3182.0	3251.0
Arabidopsis	3802.4	3964.45	5398.0	5705.0
NYClimateMarch2014	11198.3	12834.45	76912.0	81255.0
ParisAttack2015	397672.1	402372.32	603127.0	651100.0

Phần này so sánh hiệu suất giữa hai kiến trúc Mạng Nơ-ron Đồ thị (GNN) nổi bật, Mạng Tích chập Đồ thị (GCN) và Mạng Chú ý Đồ thị (GAT), trong khuôn khổ PMoE của chúng tôi. Bảng 3.14 trình bày sự lan tỏa ảnh hưởng mà REM đạt được khi áp dụng mỗi biến thể GNN làm kiến trúc chuyên gia PMoE của nó trên 5 bộ dữ liệu đã đề cập trước, dưới cả hai mô hình lan truyền LT và IC. Đáng chú ý, chúng tôi quan sát thấy hiệu suất vượt trội của GAT so với GCN trong tất cả các kịch bản. Sự khác biệt về hiệu suất này xuất phát từ khả năng của GAT trong việc gán các mức độ quan trọng khác nhau cho các nút láng giềng trong quá trình tổng hợp, điều mà GCN không có, vì GCN đối xử với tất cả các láng giềng như nhau. Việc chọn kiến trúc GAT cho các chuyên gia là cần thiết, đặc biệt là trong môi trường multiplex, nơi mà tính đa dạng và phức tạp của mối quan hệ giữa các nút đòi hỏi một quá trình tổng hợp thích ứng và chọn lọc hơn để đạt được hiệu suất tối ưu.

3.3. Kết luận chương

Chương 3 đã trình bày chi tiết quá trình thiết lập và kết quả thực nghiệm nhằm đánh giá hiệu quả của khung mô hình REM được đề xuất cho bài toán Tối đa hóa Ảnh hưởng Đa tầng (MIM). Thông qua các thử nghiệm trên nhiều bộ dữ liệu đa dạng (Cora-ML, Celegans, Arabidopsis, NYClimatMarch2014, ParisAttack2015) và dưới cả hai mô hình lan truyền phổ biến (IC và LT), kết quả đã chứng minh một cách nhất quán rằng REM vượt trội hơn các phương pháp so sánh, bao gồm cả các thuật toán truyền thống (ISF, KSN) và các phương pháp học sâu tiên tiến (GCOMB, ToupleGDD, DeepIM, MIM-Reasoner).

Sự vượt trội này đặc biệt rõ rệt trên các mạng quy mô lớn và với ngân sách hạt giống lớn hơn, nơi REM đạt được mức độ lan tỏa ảnh hưởng cao hơn đáng kể. Việc đánh giá các biến thể REM-NonRL và REM-NonMixture cũng cho thấy hiệu suất giảm sút, khẳng định tầm quan trọng của cả cơ chế khám phá dựa trên học tăng cường (RL) và kiến trúc Hỗn hợp Chuyên gia Lan truyền (PMoE) trong việc đạt được hiệu quả tối ưu. Các thành phần này giúp REM không chỉ học từ dữ liệu ban đầu mà còn chủ động khám phá các giải pháp tiềm năng tốt hơn và mô hình hóa chính xác động lực lan truyền phức tạp trong mạng đa tầng.

Về mặt hiệu quả tính toán, REM cho thấy khả năng mở rộng tốt với thời gian suy luận tăng gần như tuyến tính theo kích thước đồ thị và thường nhanh hơn các phương pháp dựa trên học máy cạnh tranh khác, làm cho nó trở thành một giải pháp khả thi cho các ứng dụng thực tế. Các phân tích sâu hơn về ảnh hưởng của số bước khám phá môi trường đã cho thấy lợi ích của việc tăng cường khám phá, đặc biệt trên các mạng lớn. Đồng thời, việc so sánh GCN và GAT trong PMoE đã làm nổi bật ưu thế của GAT, nhấn mạnh sự cần thiết của cơ chế chú ý trong việc xử lý các tương tác phức tạp của mạng đa tầng.

KẾT LUẬN

Đề án này đã tập trung nghiên cứu và giải quyết bài toán Tối đa hóa Ảnh hưởng Đa tầng (Multiplex Influence Maximization - MIM), một vấn đề phức tạp và có ý nghĩa thực tiễn cao trong bối cảnh mạng xã hội đa nền tảng ngày càng phát triển. Bài toán MIM đặt ra nhiều thách thức hơn so với Tối đa hóa Ảnh hưởng (IM) truyền thống do sự tương tác phức tạp và động lực lan truyền độc đáo giữa các lớp mạng, điển hình là hiện tượng "kích hoạt trùng lặp". Các phương pháp hiện có, bao gồm cả thuật toán tối ưu tổ hợp và các kỹ thuật học máy ban đầu, thường gặp khó khăn về khả năng mở rộng, độ chính xác trong ước lượng lan truyền và hiệu quả tối ưu hóa trong không gian giải pháp lớn và nhiễu của mạng đa tầng.

Để giải quyết những thách thức này, đề án đã đề xuất một khung làm việc học sâu đa tầng mới mang tên Tối đa hóa Chuyên gia Tăng cường (Reinforced Expert Maximization - REM). Đóng góp cốt lõi của REM nằm ở việc tích hợp sáng tạo các cơ chế học sâu và học tăng cường. REM bắt đầu bằng việc sử dụng Seed2Vec, một mô hình dựa trên Bộ tự mã hóa biến phân (VAE), để học biểu diễn hiệu quả cho các tập hạt giống phức tạp, ánh xạ chúng vào một không gian tiềm ẩn liên tục, ít nhiễu hơn, tạo điều kiện thuận lợi cho việc tối ưu hóa. Nhận thấy giới hạn của việc chỉ học từ dữ liệu ban đầu, REM tiếp tục coi Seed2Vec như một tác nhân học tăng cường (RL), chủ động khám phá không gian tiềm ẩn này để tạo ra các tập hạt giống tổng hợp, đa dạng và có khả năng lan truyền cao, vượt qua giới hạn của dữ liệu huấn luyện và tránh bị mắc kẹt ở các cực tiểu cục bộ. Để dẫn dắt quá trình khám phá và đánh giá chất lượng các tập hạt giống tiềm năng, REM giới thiệu kiến trúc Hỗn hợp Chuyên gia Lan truyền (PMoE). Kiến trúc này sử dụng một tập hợp các Mạng nơ-ron đồ thị (GNN) chuyên biệt với độ sâu khác nhau, kết hợp với một mạng định tuyến thông minh, nhằm ước lượng chính xác và hiệu quả các mẫu lan truyền phức tạp trong mạng đa tầng quy mô lớn, giảm thiểu hiện tượng làm mịn quá mức và tăng cường khả năng mô hình hóa động lực học tinh vi của mạng. Sự kết hợp giữa việc biểu diễn hiệu quả, khám phá thông minh và ước lượng lan truyền chính xác tạo nên sức mạnh tổng thể của khung REM.

Quá trình thực nghiệm trên nhiều bộ dữ liệu mạng đa lớp thực tế (Cora-ML, Celegans, Arabidopsis, NYClimatMarch2014, ParisAttack2015) dưới cả hai mô hình lan truyền phổ biến (IC và LT) đã chứng minh hiệu quả vượt trội của REM. Kết quả cho thấy REM liên tục đạt được mức độ lan tỏa ảnh hưởng cao hơn đáng kể so với các phương pháp tiên tiến hiện có, bao gồm cả các thuật toán tối ưu tổ hợp chuyên biệt cho MIM và các phương pháp học sâu khác. Đặc biệt, REM thể hiện khả năng mở rộng tốt trên các mạng quy mô lớn và duy trì thời gian suy luận cạnh tranh. Các phân tích sâu hơn về các biến thể mô hình và lựa chọn kiến trúc cũng đã khẳng định vai trò quan trọng của từng thành phần được đề xuất trong việc đạt được hiệu suất cao. Đáng chú ý, những đóng góp và kết quả nghiên cứu chính của luận văn đã được chấp nhận đăng và sẽ được trình bày tại Hội nghị AAAI Conference on Artificial Intelligence (AAAI) 2025, tổ chức tại Philadelphia, Mỹ, minh chứng cho tính mới và ý nghĩa khoa học của công trình (có đính kèm phụ lục bài báo ở cuối đề án).

Mặc dù đạt được những kết quả đáng khích lệ, nghiên cứu này vẫn còn một số hạn chế. Hiệu quả của REM vẫn có thể phụ thuộc vào chất lượng của tập dữ liệu hạt giống ban đầu, mặc dù cơ chế RL đã giúp giảm thiểu phần nào sự phụ thuộc này. Chi phí tính toán cho quá trình huấn luyện, đặc biệt là giai đoạn khám phá bằng RL và huấn luyện lại mô hình, có thể còn cao và cần được tối ưu hóa thêm.

Hướng nghiên cứu trong tương lai có thể tập trung vào việc cải thiện hiệu quả huấn luyện của REM, khám phá các chiến lược RL tiên tiến hơn để tối ưu hóa quá trình khám phá tập hạt giống. Việc mở rộng REM để xử lý các mạng đa tầng động, nơi cấu trúc mạng và mối quan hệ thay đổi theo thời gian, cũng là một hướng đi đầy hứa hẹn, phản ánh đúng hơn bản chất của các mạng xã hội thực tế. Ngoài ra, việc tích hợp các thuộc tính ngữ nghĩa của nút hoặc nội dung thông tin lan truyền vào mô hình có thể làm tăng thêm độ chính xác và tính ứng dụng thực tế của giải pháp. Cuối cùng, việc áp dụng các ý tưởng cốt lõi của REM, đặc biệt là sự kết hợp giữa học biểu diễn, học tăng cường và mô hình dự đoán chuyên biệt, cho các bài toán tối ưu tổ hợp khác trên đồ thị phức tạp cũng là một tiềm năng cần được khai thác.

CÔNG BỐ KHOA HỌC

Tiêu đề bài báo: REM: A SCALABLE REINFORCED MULTI-EXPERT
FRAMEWORK FOR MULTIPLEX INFLUENCE MAXIMIZATION

Hội nghị: Hội nghị AAAI 2025

Điểm tác động nghiên cứu: 32.10

Ngày nhận bài: 15 tháng 8 năm 2024

Ngày chấp nhận: 9 tháng 12 năm 2024

Ngày nộp bản chỉnh sửa: 19 tháng 12 năm 2024

Tác giả: Huyen Nguyen, Hieu Dam, Nguyen Do, Cong Tran, Cuong Pham

Vị trí: Tác giả chính

TÀI LIỆU THAM KHẢO

- [1] A. Antelmi, G. Cordasco, C. Spagnuolo, and P. Szufel. Social influence maximization in hypergraphs. *Entropy*, 23(7):796, 2021.
- [2] A. Goyal, W. Lu, and L. Lakshmanan. CELF++: Optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th International Conference Companion on World Wide Web, WWW 2011*, pages 47–48, 2011.
- [3] A. Kuhnle, M. A. Alim, X. Li, H. Zhang, and M. T. Thai. Multiplex influence maximization in online social networks with heterogeneous diffusion models. *IEEE Transactions on Computational Social Systems*, 5(2):418–429, 2018.
- [4] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.
- [5] B. W. Dolhansky and J. A. Bilmes. Deep submodular functions: Definitions and learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.
- [6] C. Ling, J. Jiang, J. Wang, and L. Zhao. Source localization of graph diffusion via variational autoencoders for graph inverse problems. In *Proc. of the KDD*, 2022.
- [7] C. Ling, J. Jiang, J. Wang, M. T. Thai, R. Xue, J. Song, M. Qiu, and L. Zhao. Deep graph representation learning and optimization for influence maximization. In *International Conference on Machine Learning*, pages 21350–21361. PMLR, 2023.
- [8] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: A general repository for interaction datasets. *Nucleic acids research*, 34:D535–9, 2006.
- [9] C. Cai and Y. Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv: 2006.13318*, 2020.
- [10] C. Ling, T. Chowdhury, J. Jiang, J. Wang, X. Zhang, H. Chen, and L. Zhao. DeepGAR: Deep graph learning for analogical reasoning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 1065–1070. IEEE, 2022.
- [11] C. Zhou, P. Zhang, W. Zang, and L. Guo. On the upper bounds of spread for greedy algorithms in social network influence maximization. *IEEE Transactions on Knowledge and Data Engineering*, 27(10):2770–2783, 2015.
- [12] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. Van Hasselt, and D. Silver. Distributed prioritized experience replay. *arXiv preprint arXiv:1803.00933*, 2018. D. P. Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc. of the KDD*, 2003.

- [14] D. F. Nettleton. Data mining of social networks represented as graphs. *Computer Science Review*, 7:1–34, 2013.
- [15] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2013.
- [16] E. Omodei, M. D. De Domenico, and A. Arenas. Characterizing interactions in online social networks during exceptional events. *Frontiers in Physics*, 3, 2015.
- [17] F. Bonchi. Influence propagation in social networks: A data mining perspective. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 1:2–2. IEEE, 2011.
- [18] H. Li, M. Xu, S. S. Bhowmick, J. S. Rayhan, C. Sun, and J. Cui. PIANO: Influence maximization meets deep reinforcement learning. *IEEE Transactions on Computational Social Systems*, 2022.
- [19] H. Zhang, D. T. Nguyen, H. Zhang, and M. T. Thai. Least cost influence maximization across multiple social networks. *IEEE/ACM Transactions on Networking*, 24(2):929–939, 2016.
- [20] H. Zhang, D. T. Nguyen, H. Zhang, and M. T. Thai. Least cost influence maximization across multiple social networks. *IEEE/ACM Transactions on Networking*, 24(2):929–939, 2016.
- [21] H. Zhang, X. Chen, Y. Peng, G. Kou, and R. Wang. The interaction of multiple information on multiplex social networks. *Information Sciences*, 605:366–380, 2022.
- [22] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, 2007.
- [23] J. Wu and D. Li. Modeling and maximizing information diffusion over hypergraphs based on deep reinforcement learning. *Physica A: Statistical Mechanics and its Applications*, 629:129193, 2023.
- [24] J. S. Lim, S. Y. Ri, B. D. Egan, and F. A. Biocca. The cross-platform synergies of digital video advertising: Implications for cross-media campaigns in television, Internet and mobile TV. *Computers in Human Behavior*, 48:463–472, 2015.
- [25] J. Zhu, S. Ghosh, W. Wu, and C. Gao. Profit maximization under group influence model in social networks. In *International Conference on Computational Data and Social Networks*, pages 108–119. Springer, 2019.
- [26] J. Zhu, J. Zhu, S. Ghosh, W. Wu, and J. Yuan. Social influence maximization in hypergraph in social networks. *IEEE Transactions on Network Science and Engineering*, 6(4):801–811, 2019.

- [27] K. Ali, C.-Y. Wang, and Y.-S. Chen. Boosting reinforcement learning in competitive influence maximization with transfer learning. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pages 395–400. IEEE, 2018.
- [28] M. Contisciani, E. A. Power, and C. De Bacco. Community detection with node attributes in multilayer networks. *Scientific Reports*, 2020.
- [29] M. De Domenico and E. G. Altmann. Unraveling the origin of social bursts in collective attention. *Scientific reports*, 10(1):4629, 2020.
- [30] M. Jalili, Y. Orouskhani, M. A. Mehrabadi, N. Alipourfard, and M. Perc. Link prediction in multiplex online social networks. *Royal Society Open Science*, 4(2):160863, 2017.
- [31] M. Katukuri, M. Jagarapu, and others. CIM: clique-based heuristic for finding influential nodes in multilayer networks. *Applied Intelligence*, 52(5):5173–5184, 2022.
- [32] M. Kimura and K. Saito. Tractable models for information diffusion in social networks. In *Knowledge Discovery in Databases: PKDD 2006*, pages 259–271. Springer Berlin Heidelberg, 2006.
- [33] M. Xie, X.-X. Zhan, C. Liu, and Z.-K. Zhang. Influence maximization in hypergraphs. *arXiv preprint arXiv:2206.01394*, 2022.
- [34] N. M. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.
- [35] N. H. K. Do, T. Chowdhury, C. Ling, L. Zhao, and M. T. Thai. MIM-Reasoner: Learning with theoretical guarantees for multiplex influence maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 2296–2304, 2024.
- [36] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66. ACM, 2001.
- [37] P. Vikatos, P. Gryllos, and C. Makris. Marketing campaign targeting using bridge extraction in multiplex social network. *Artificial Intelligence Review*, 53:2335–2360, 2020.
- [38] Q. Li, Z. Han, and X.-M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2018.
- [39] Q. Zhan, J. Zhang, S. Wang, P. S. Yu, and J. Xie. Influence maximization across partially aligned heterogeneous social networks. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 58–69. Springer, 2015.
- [40] S. S. Singh, A. Kumar, K. Singh, and B. Biswas. C2IM: Community based context-aware influence maximization in social networks. *Physica A: Statistical Mechanics and its Applications*, 514:796–818, 2019.

- [41] S.-C. Lin, S.-D. Lin, and M.-S. Chen. A learning-based framework to handle multi-round multi-party influence maximization on social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 695–704, 2015.
- [42] S. Manchanda, A. Mittal, A. Dhawan, S. Medya, S. Ranu, and A. Singh. GCOMB: Learning budget-constrained combinatorial algorithms over billion-sized graphs. In *Advances in Neural Information Processing Systems*, pages 20000–20011, 2020.
- [43] S. Banerjee, M. Jenamani, and D. K. Pratihari. ComBIM: A community-based solution approach for the Budgeted Influence Maximization Problem. *Expert Systems with Applications*, 125:1–13, 2019.
- [44] S. S. Singh, K. Singh, A. Kumar, and B. Biswas. MIM2: Multiple influence maximization across multiple social networks. *Physica A: Statistical Mechanics and its Applications*, 526:120902, 2019.
- [45] S. Bai, F. Zhang, and P. H. S. Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognit.*, 110:107637, 2021.
- [46] T. Chen, S. Yan, J. Guo, and W. Wu. ToupleGDD: A finedesigned solution of influence maximization by deep reinforcement learning. *arXiv preprint arXiv:2210.07500*, 2022.
- [47] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1029–1038. ACM, 2010b.
- [48] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In *2010 IEEE International Conference on Data Mining*, pages 88–97. IEEE, 2010a.
- [49] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.
- [50] X. Chen, L. Deng, Y. Zhao, X. Zhou, and K. Zheng. Community-based influence maximization in location-based social network. *WWWJ*, 24(6):1903–1928, 2021.
- [51] X. Li, J. D. Smith, T. N. Dinh, and M. T. Thai. Tiptop:(almost) exact solutions for influence maximization in billion-scale networks. *IEEE/ACM Transactions on Networking*, 27(2):649–661, 2019.
- [52] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox. DeepIM: Deep iterative matching for 6D pose estimation. *International Journal of Computer Vision*, 2018.
- [53] Y. Tang, X. Xiao, and Y. Shi. Influence maximization: Nearoptimal time complexity meets practical efficiency. In *Proc. of the SIGMOD*, pages 75–86, 2014.

- [54] Z. Yuan, M. Shao, and Z. Chen. Graph bayesian optimization for multiplex influence maximization. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 22475–22483. AAAI Press, 2024.
- [55] Z. Liaghat, A. Rasekh, and A. Mahdavi. Application of data mining methods for link prediction in social networks. *Social Network Analysis and Mining*, 3(1):97–111, 2013.



BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

Tên tài liệu: Đề án_Nguyễn Thị Ngọc Huyền_B23CHIS012
Tác giả: Huyền Nguyễn
Điểm trùng lặp: 7
Thời gian tải lên: 00:46 19/05/2025
Thời gian sinh báo cáo: 00:49 19/05/2025
Các trang kiểm tra: 81/81 trang



Kết quả kiểm tra trùng lặp



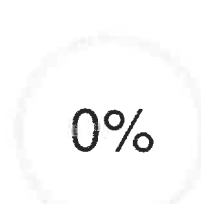
Có 7% nội dung trùng
lặp



Có 93% nội
dung không
trùng lặp



Có 0% nội dung
người dùng loại
trừ



Có 0% nội dung
hệ thống bỏ qua

Nguồn trùng lặp tiêu biểu

arxiv.org 123docz.net tailieu.vn

Người hướng dẫn
(Ký tên)

PGS.TS. Phạm Văn Cường

Tác giả thực hiện
(Ký tên)

Nguyễn Thị Ngọc Huyền

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIÊN ĐỀ ÁN TỐT NGHIỆP

Họ và tên học viên: Nguyễn Thị Ngọc Huyền

Chuyên ngành: HTTT

Khóa: 2023 đợt 1

Tên đề tài: Tối đa hóa ảnh hưởng trên mạng đa kênh kết hợp giữa mô hình sinh và học tăng cường.

Người hướng dẫn khoa học: PGS.TS. Phạm Văn Cường

Ngày bảo vệ: 19/07/2025

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Gắn kết logic các nội dung viết trong trình bày rõ hơn	Học viên đã rà soát, chỉnh sửa các lỗi logic các nội dung viết trong trình bày của quyền đồ án.

Hà Nội, ngày 4 tháng 8 năm 2025

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG
CHẤM ĐỀ ÁN

GS.TS. Từ Minh Phương

THƯ KÝ HỘI ĐỒNG

TS. Đỗ Thị Liên

NGƯỜI HƯỚNG
DẪN KHOA HỌC

PGS.TS. Phạm Văn
Cường

HỌC VIÊN

Nguyễn Thị Ngọc
Huyền

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 1098/QĐ-HV ngày 26 tháng 06 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi 7.....giờ 40 phút, ngày 19 tháng 07 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Nguyễn Thị Ngọc Huyền**

Tên đề án tốt nghiệp: **Tối đa hóa ảnh hưởng trên mạng đa kênh kết hợp giữa mô hình sinh và học tăng cường**

Chuyên ngành: **Hệ thống thông tin**

Mã số: **8480104**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: 05...../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	GS.TS. Từ Minh Phương	Chủ tịch	
2	TS. Đỗ Thị Liên	Thư ký	
3	PGS. TS. Nguyễn Mạnh Hùng	Phản biện 1	
4	TS. Nguyễn Văn Vinh	Phản biện 2	
5	PGS.TS. Trần Đăng Hưng	Ủy viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên. (có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)

6. Các câu hỏi của thành viên Hội đồng:

- làm thế nào để hình thức luận được?
- Ngân sách lương được ngân sách đến mạng xã hội được hình như thế nào?
- Có số đơn tham số như nghiệm như thế nào?
- Kiểm tra định hình của mô hình thuật toán ASec d2 Vec?

7. Trả lời của học viên:

Học viên giải trình và tiếp thu làm rõ hơn trong trình bày.....

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên (có văn bản kèm theo).

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:

1. Trưởng Ban kiểm phiếu: ...*Trần Đăng Hùng*.....

2. Ủy viên Ban kiểm phiếu: ...*Đỗ Thị Liên*.....

3. Ủy viên Ban kiểm phiếu: ...*Nguyễn Mạnh Hùng*.....

- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.

- Ban kiểm phiếu làm việc:

- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu (có Biên bản họp Ban kiểm phiếu kèm theo)

- Điểm trung bình của đề án tốt nghiệp:*10*.....

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

.....*Cần kết logic các nội dung viết trong trình bày rõ hơn*.....

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên:

.....*Đề nghị học viện công nhận và cấp bằng thạc sĩ cho học viên*.....

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho

NCS.....*6*.....

Buổi làm việc kết thúc vào.....*8h25*..... cùng ngày.

Chủ tịch



GS.TS. Từ Minh Phương

Thư ký



TS. Đỗ Thị Liên

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc
-----o0o-----

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề án: ***Tối đa hóa ảnh hưởng trên mạng đa kênh kết hợp giữa mô hình sinh và học tăng cường***

Chuyên ngành: Hệ thống thông tin

Mã chuyên ngành: 8.48.01.04

Tên học viên: ***Nguyễn Thị Ngọc Huyền***

Họ và tên người phản biện: ***Nguyễn Mạnh Hùng***

Học hàm, học vị: PGS.TS.

Nơi công tác: Khoa Công nghệ thông tin 1 - Học viện Công nghệ Bưu chính viễn thông.

Số điện thoại liên hệ:

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Bài toán tối đa hóa ảnh hưởng trên MXH là một trong các bài toán có ý nghĩa thực tiễn hiện nay.

II/ Nội dung của đề án, các kết quả đã đạt được:

Nội dung đề án bao gồm 3 chương:

Chương 1 giới thiệu bài toán và các công cụ liên quan.

Chương 2 trình bày chiến lược tối ưu hóa ảnh hưởng với các ràng buộc về ngân sách trên MXH.

Chương 3 trình bày thực nghiệm và đánh giá.

Nhìn chung, đề án đạt yêu cầu của một đề án thạc sĩ kỹ thuật.

Tuy nhiên, đề án vẫn còn một số vấn đề:

- Tên chương 2 có nói về các ràng buộc về ngân sách trên MXH nhưng nội dung không thấy định nghĩa thế nào là các ràng buộc ngân sách trên MXH, cũng không thấy chỗ nào trình bày giải pháp thỏa mãn hoặc tối ưu các ràng buộc ngân sách này.
- Chương 3, các thực nghiệm so sánh tham số là hiệu suất lan truyền của các mô hình, nhưng không thấy chỗ nào định nghĩa cách tính hiệu suất này.
- Bảng 3.1 trình bày giá trị được chọn của các tham số trong mô hình, nhưng không thấy có minh chứng nào, cơ sở nào để biết được những giá trị này là tối ưu cho mô hình.

- Định nghĩa 1 trang 27 lại đi tham khảo từ tài liệu số 47, nếu không phải do mình tự định nghĩa thì chỉ nên dùng dẫn nguồn là được.
- Về trình bày, còn nhiều lỗi sơ đẳng: Hình 2.2 (trang 25.26) hình ở một trang, caption ở một trang. Một số đoạn trang 28.30 tự nhiên in đậm lên cả đoạn.

III/ Những vấn đề cần giải thích thêm:

- Tham số hiệu suất tính như thế nào?
- Cơ sở nào để chọn các tham số ở bảng 3.1?
- Việc thỏa mãn các ràng buộc ngân sách được tính thế nào trong mô hình, và được đánh giá thế nào trong thực nghiệm?

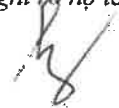
IV/ Kết luận:

Đề án đáp ứng được yêu cầu của một đề án thạc sĩ kỹ thuật.

Đồng ý cho phép học viên được bảo vệ đề án tốt nghiệp.

Hà Nội, ngày 15 tháng 07 năm 2025

Người nhận xét
(Ký ghi rõ họ tên)



Nguyễn Mạnh Hùng

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ
(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: **TỐI ĐA HÓA ẢNH HƯỞNG TRÊN MẠNG ĐA KÊNH
KẾT HỢP GIỮA MÔ HÌNH SINH VÀ HỌC TĂNG CƯỜNG**

Chuyên ngành: Hệ thống thông tin

Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: Nguyễn Thị Ngọc Huyền

Họ và tên người nhận xét: Nguyễn Văn Vinh

Học hàm, học vị: TS

Chuyên ngành: KHMT

Cơ quan công tác: Trường ĐH Công Nghệ, ĐHQG Hà Nội

Số điện thoại:.....E-mail:.....

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Tối ưu hóa ảnh hưởng trên mạng đa kênh, đặc biệt là mạng xã hội là bài toán quan trọng trong lĩnh vực truyền thông trực tuyến. Vì vậy việc nghiên cứu và xây dựng mô hình sinh và kết hợp với học tăng cường để giải quyết bài toán tối ưu ảnh hưởng trên mạng đa kênh là có ý nghĩa về mặt thực tiễn cũng như ứng dụng.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

- Nghiên cứu về bài toán tối đa hóa ảnh hưởng trên mạng xã hội
- Đề xuất tối đa hóa ảnh hưởng mạng xã hội đa lớp dựa vào mô hình sinh và học tăng cường.
- Thực nghiệm phương pháp đề xuất với nhiều bộ dữ liệu mạng đa lớp thực tế (Cora-ML, Celegans, Arabidopsis, NYClimatMarch2014, ParisAttack2015) dưới cả hai mô hình lan truyền phổ biến IC và LT cho thấy phương pháp REM tốt hơn so với các phương pháp trước đó. Kết quả chính đã được công bố trên hội nghị A* về AI (AAAI 2025).

III/ Những vấn đề cần giải thích thêm:

- Một số hình vẽ nhảy trang như hình 2.2 trang 26 nên chỉnh lại sang trang mới. Thứ tự bỏ đề trong chương 2 nên để bỏ đề 1 trước bỏ đề 2.
- Việc đánh số các công thức nên đánh theo tên chương ví dụ (2.1, ...)

- Nên phân tích rõ hơn về thành phần Seed2vec, đây là thành phần rất quan trọng trong phương pháp đề xuất. Ví dụ bổ đề 2, Seed2Vec hội tụ nghĩa là thế nào? Tại sao lại biểu diễn học dưới dạng VAE lại hiệu quả?
- Các thuật ngữ nên nhất quán: (Tối ưu, tối đa), Một số lỗi chính tả cần được chỉnh sửa như các Mikitiuk .

IV/ Kết luận:

Đề án tốt nghiệp đáp ứng tốt đề án tốt nghiệp hướng nghiên cứu. Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày: 14...tháng...07...năm 2025

NGƯỜI NHẬN XÉT



NGUYỄN VĂN VINH

