ChatGPT

Tài liệu này trình bày bài toán tối đa hóa ảnh hưởng trong mạng xã hội khi có nhiều chủ đề (k-chủ đề) và sự hiện diện của nhiễu ngẫu nhiên, đồng thời đề xuất các phương pháp giải quyết dựa trên học sâu. Các tác giả áp dụng mô hình mạng nơ-ron đồ thị (Graph Neural Network – GNN**, một loại mạng học sâu chuyên xử lý dữ liệu có cấu trúc đồ thị) để ước lượng sức ảnh hưởng của các nút trong mạng, từ đó lựa chọn tập các nút khởi đầu (hạt giống) tối ưu cho việc lan truyền. Kết quả thực nghiệm cho thấy phương pháp học sâu đề xuất đạt hiệu quả lan truyền ảnh hưởng cao và ổn định, vượt trội so với các phương pháp truyền thống trong bối cảnh đa chủ đề và có nhiễu ngẫu nhiên, đồng thời rút ngắn đáng kể thời gian tính toán.**

Tóm tắt

Bài báo nghiên cứu bài toán **tối đa hóa ảnh hưởng k-chủ đề** trong mạng xã hội, trong đó có *k* chủ đề thông tin cần được khuếch tán đồng thời. Bài toán được xem xét dưới mô hình lan truyền độc lập (**Independent Cascade model**, viết tắt **IC model** – mô hình lan truyền trong đó mỗi liên kết giữa hai nút có một xác suất cố định để truyền ảnh hưởng; ví dụ, nếu xác suất $p=0.1$, một người dùng sẽ ảnh hưởng thành công bạn của họ với xác suất 10% ở mỗi lần thử). Tuy nhiên, khác với các nghiên cứu trước đây giả định các tham số ảnh hưởng cố định, bài báo này xem xét **nhiễu ngẫu nhiên** trong quá trình lan truyền – tức là các xác suất ảnh hưởng có thể thay đổi ngẫu nhiên quanh một giá trị trung bình, gây khó khăn trong việc dự đoán chính xác phạm vi ảnh hưởng.

Để giải quyết bài toán trên, nhóm tác giả đề xuất hai phương pháp học sâu mới. Phương pháp thứ nhất sử dụng **mạng nơ-ron đồ thị (Graph Neural Network – GNN)** để học biểu diễn của các nút và dự đoán tiềm năng ảnh hưởng của chúng. Phương pháp thứ hai kết hợp GNN với kỹ thuật **nhiễu bổ sung** trong quá trình huấn luyện, nhằm tăng cường khả năng khái quát hóa của mô hình trước sự bất định (nhiễu) trong dữ liệu. Cả hai phương pháp này đều hướng tới việc xấp xỉ hàm ảnh hưởng – vốn là một **hàm submodular** (*submodular function*, hàm có tính chất lợi ích biên giảm dần: thêm một phần tử vào tập càng lớn thì lợi ích tăng thêm càng nhỏ; ví dụ, hàm bao phủ tập là submodular vì việc chọn thêm một điểm mới sẽ ít hữu ích hơn nếu nhiều điểm đã được chọn trước đó) – bằng mô hình học sâu, từ đó cho phép tìm gần đúng tập hạt giống tối ưu với độ phức tạp thấp hơn đáng kể so với thuật toán tham lam truyền thống. Các thực nghiệm trên nhiều bộ dữ liệu mạng xã hội thực tế cho thấy phương pháp đề xuất đạt chất lượng gần tương đương, thậm chí nhỉnh hơn trong một số trường hợp, so với thuật toán tham lam kinh điển, trong khi tốc độ nhanh hơn hàng chục đến hàng trăm lần.

Tóm lại, bài báo đã giới thiệu bài toán tối đa hóa ảnh hưởng đa chủ đề dưới tác động của nhiễu và đề xuất các giải pháp học sâu hiệu quả. Phương pháp sử dụng GNN tỏ ra triển vọng khi vừa đảm bảo phạm vi ảnh hưởng lớn, vừa giảm thời gian tính toán, mở ra hướng đi mới cho các ứng dụng tối ưu ảnh hưởng trong mạng xã hội phức tạp.

1. Giới thiệu

Tối đa hóa ảnh hưởng (Influence Maximization – IM) là một bài toán quan trọng trong phân tích mạng xã hội, nhằm mục tiêu chọn ra một tập gồm k người dùng (gọi là các **nút hạt giống**) sao cho khi những người này được lan truyền thông tin (ví dụ như quảng bá một sản phẩm mới), họ sẽ ảnh hưởng được nhiều người nhất có thể trong mạng. Bài toán IM cổ điển, do Kempe và các cộng sự đề xuất lần đầu năm 2003, thường được xét với một chủ đề thông tin duy nhất và mô hình lan truyền cụ thể (ví dụ mô hình IC độc lập hoặc mô hình ngưỡng tuyến tính). Kempe đã chứng minh rằng bài toán này là NP-khó (NP-hard – nghĩa là không có thuật toán thời gian đa thức để tìm nghiệm tối ưu tổng

quát), nhưng hàm mục tiêu (số lượng người bị ảnh hưởng kỳ vọng) có tính chất **submodular** (tính chất giảm dần lợi ích biên như đã giải thích), và **đơn điệu** (monotonic – giá trị hàm không giảm khi tập hạt giống mở rộng). Nhờ đó, có thể áp dụng thuật toán tham lam (greedy) để chọn hạt giống một cách xấp xỉ: bắt đầu với tập rỗng, lặp lại chọn thêm 1 nút đem lại tăng trưởng ảnh hưởng lớn nhất cho đến khi đủ k nút. Thuật toán tham lam kinh điển này đảm bảo đạt ít nhất $(1 - 1/e) \approx 63\%$ so với ảnh hưởng tối ưu [1] [2]. Tuy nhiên, nhược điểm của phương pháp tham lam là chi phí tính toán rất cao, do phải ước lượng ảnh hưởng biên của từng ứng viên hạt giống tại mỗi bước (thường dùng mô phỏng Monte Carlo lặp đi lặp lại hàng nghìn lần để đạt độ chính xác cần thiết).

Bài toán tối đa hóa ảnh hưởng càng trở nên phức tạp hơn trong **bối cảnh đa chủ đề (k-chủ đề)**, khi có đồng thời k loại thông tin hoặc k chiến dịch quảng bá khác nhau cần được lan truyền. Trong tình huống này, ta có thể cần chọn các hạt giống khác nhau cho từng chủ đề (ví dụ, một người có tầm ảnh hưởng lớn trong chủ đề thể thao chưa chắc đã ảnh hưởng nhiều trong chủ đề thời trang). Một cách tiếp cận tự nhiên là xem mỗi chủ đề như một "lớp" ảnh hưởng riêng và chọn hạt giống cho từng lớp. Tuy nhiên, nếu có giới hạn tổng số hạt giống (ví dụ tổng cộng chỉ được chọn B người cho cả k chủ đề), ta phải phân bổ ngân sách B đó cho các chủ đề sao cho tối ưu tổng ảnh hưởng. Bài toán này thường được gọi là **tối đa hóa ảnh hưởng k-chủ đề** và có liên quan mật thiết đến khái niệm **k-submodular** (khi $k>1$, hàm mục tiêu trở thành một hàm k-submodular nếu ta xem mỗi chủ đề có một hàm submodular riêng; k-submodular là tổng quát hóa của submodular cho trường hợp có k hàm mục tiêu cần tối ưu đồng thời). Bài toán k-chủ đề cũng phức tạp tương tự IM cổ điển; thậm chí nếu ràng buộc yêu cầu mỗi chủ đề phải có ít nhất một hạt giống, bài toán không còn thỏa mãn tính submodular thông thường, khiến chiến lược tham lam không còn đảm bảo hiệu quả tối ưu về lý thuyết.

Ngoài ra, trong thực tế, các tham số lan truyền (như xác suất ảnh hưởng trên mỗi liên kết) thường không cố định chính xác mà có sai số hoặc thay đổi ngẫu nhiên do nhiều yếu tố khó đo lường. Để mô phỏng điều này, bài toán được mở rộng với yếu tố **nhiễu ngẫu nhiên**: giả sử mỗi lần lan truyền, xác suất thành công trên một liên kết có thể dao động ngẫu nhiên quanh một giá trị trung bình đã biết. Điều này phản ánh tính bất định trong hành vi người dùng hoặc hoàn cảnh truyền thông (ví dụ, cùng một người nhưng mức độ thuyết phục có thể khác nhau tùy thời điểm). Nhiễu ngẫu nhiên làm tăng độ khó cho việc dự đoán và tối ưu ảnh hưởng vì hàm mục tiêu lúc này có thể biến thiên giữa các lần thực hiện, thậm chí mất tính submodular trong mỗi lần cụ thể. Các phương pháp truyền thống vốn dựa trên tham số cố định sẽ kém hiệu quả hoặc không còn áp dụng trực tiếp được trong môi trường nhiễu.

Trước những thách thức trên, **nhóm tác giả đã đề xuất tiếp cận bài toán bằng các phương pháp học sâu**. Thay vì dựa hoàn toàn vào mô hình toán học và thuật toán tham lam, bài báo tận dụng khả năng **học biểu diễn** và suy luận của mạng nơ-ron để **ước lượng hàm ảnh hưởng** trực tiếp từ cấu trúc mạng và dữ liệu huấn luyện. Cách tiếp cận này kỳ vọng sẽ: (i) giảm đáng kể thời gian tính toán so với mô phỏng Monte Carlo lặp lại nhiều lần, (ii) thích nghi tốt hơn với sự hiện diện của nhiễu ngẫu nhiên bằng cách học trực tiếp từ các trường hợp lan truyền nhiễu, và (iii) mở rộng linh hoạt cho trường hợp đa chủ đề thông qua kiến trúc mạng phù hợp.

Đóng góp chính của bài báo bao gồm:

- Định nghĩa và phân tích bài toán tối đa hóa ảnh hưởng đa chủ đề với nhiễu ngẫu nhiên, chỉ ra những thách thức mới so với bài toán IM truyền thống (đơn chủ đề, không nhiễu). Cụ thể, bài báo mô tả cách nhiễu ảnh hưởng đến mô hình lan truyền độc lập và thảo luận về tính chất submodular (hoặc k-submodular) của hàm ảnh hưởng trong bối cảnh này.
- Đề xuất một khung phương pháp dựa trên học sâu đầu tiên cho bài toán nêu trên. Trong đó, quan trọng nhất là việc thiết kế mô hình **Graph Neural Network (GNN)** để học **hàm xấp xỉ ảnh hưởng** của các nút. Mô hình GNN này tận dụng cấu trúc đồ thị của mạng xã hội và có cơ chế tích

hợp thông tin đa chủ đề. Ngoài ra, kỹ thuật thêm nhiễu ngẫu nhiên vào dữ liệu huấn luyện được áp dụng để mô phỏng sự bất định, giúp mô hình học được **biểu diễn bền vững** trước nhiễu.
- Xây dựng bộ thực nghiệm toàn diện trên nhiều mạng xã hội thực tế để đánh giá hiệu quả của phương pháp đề xuất. Kết quả cho thấy phương pháp của chúng tôi **đạt được ảnh hưởng lan truyền xấp xỉ tối ưu** (tiệm cận kết quả của thuật toán tham lam trong trường hợp không nhiễu) và **vượt trội về tốc độ**. Đặc biệt, trong môi trường có nhiễu, phương pháp học sâu tỏ ra **ổn định và hiệu quả hơn** so với phương pháp truyền thống, vốn không được thiết kế để xử lý tham số ngẫu nhiên.

Tóm lược: Phần giới thiệu đã nêu bật tầm quan trọng của bài toán tối đa hóa ảnh hưởng trong mạng xã hội, mở rộng vấn đề sang trường hợp đa chủ đề và có tham số nhiễu, đồng thời giới thiệu hướng tiếp cận bằng học sâu nhằm khắc phục hạn chế của các phương pháp truyền thống. Nhóm tác giả cũng liệt kê những đóng góp chính, bao gồm phân tích bài toán phức tạp hơn và đề xuất giải pháp GNN kết hợp kỹ thuật nhiễu, cùng với minh chứng hiệu quả qua thực nghiệm.

2. Cơ sở lý thuyết và phát biểu bài toán

Mạng xã hội và mô hình lan truyền: Chúng tôi mô hình hóa mạng xã hội như một đồ thị có hướng $G = (V, E)$, với V là tập các nút (tương ứng với người dùng) và E là tập các cung (các mối quan hệ hoặc kết nối giữa người dùng). Mỗi cung $(u, v) \in E$ được gán một xác suất $p_{u,v}$ thể hiện khả năng u ảnh hưởng đến v (ví dụ, nếu u giới thiệu một sản phẩm cho v thì v có xác suất $p_{u,v}$ sẽ quan tâm hoặc bị thuyết phục). Bài báo sử dụng **mô hình lan truyền độc lập (Independent Cascade – IC)** để mô phỏng quá trình khuếch tán ảnh hưởng trong mạng [3] [4]. Theo mô hình IC, quá trình diễn ra theo các bước rời rạc: Ban đầu, một tập hạt giống S (được chọn trước) sẽ được **kích hoạt** (tức là họ chấp nhận hoặc bắt đầu lan truyền một thông điệp). Sau đó, mỗi lần một nút u được kích hoạt ở bước t, nó sẽ có một **cơ hội duy nhất** để ảnh hưởng mỗi láng giềng v chưa được kích hoạt của mình ở bước $t+1$ với xác suất $p_{u,v}$. Nếu u ảnh hưởng thành công v, thì v sẽ trở thành kích hoạt ở bước $t+1$. Nếu u thất bại (với xác suất $1 - p_{u,v}$) hoặc v đã bị ảnh hưởng bởi người khác, u sẽ không có cơ hội thứ hai với v. Quá trình lan truyền tiếp diễn cho đến khi không còn nút mới nào được kích hoạt thêm. Mô hình IC mô phỏng hiệu ứng domino: một người bị thuyết phục sẽ tiếp tục đi thuyết phục người khác, và mỗi mối quan hệ chỉ có một lần thử truyền ảnh hưởng.

Dưới mô hình IC, với một tập hạt giống S cho trước, ta định nghĩa **hàm ảnh hưởng** $\sigma(S)$ là **số lượng kỳ vọng các nút sẽ được kích hoạt** (bị ảnh hưởng) khi quá trình lan truyền kết thúc. Do tính ngẫu nhiên của quá trình (thành công hay thất bại trên mỗi cung), $\sigma(S)$ được tính trên kỳ vọng toán học, thường ước lượng bằng cách giả lập quá trình nhiều lần. Bài toán tối đa hóa ảnh hưởng cổ điển có thể được viết thành công thức tối ưu như sau:

$$ \max_{S \subseteq V} \sigma(S) \quad \text{s.t.} \ |S| = k, $$

trong đó k là số hạt giống được phép chọn (ràng buộc ngân sách cố định). Kempe et al. (2003) đã chứng minh rằng dưới các mô hình lan truyền phổ biến (IC và LT – Linear Threshold), hàm $\sigma(S)$ có tính **đơn điệu** và **submodular**, do đó nghiệm xấp xỉ tốt nhất đạt được bằng thuật toán tham lam sẽ có $\sigma(S)$ không nhỏ hơn $(1 - 1/e)$ lần so với nghiệm tối ưu.

Mở rộng đa chủ đề: Bây giờ, xét trường hợp có k chủ đề khác nhau cần lan truyền đồng thời trên cùng một mạng G. Ký hiệu tập chủ đề là $\mathcal{T} = {1, 2, ..., k}$. Ta có thể coi mỗi chủ đề $i \in

\mathcal{T}$ tương ứng với một hàm ảnh hưởng $\sigma_i(S_i)$ riêng (với S_i là tập hạt giống được chọn cho chủ đề i). Tùy theo cách đặt bài toán, ta có thể có các biến thể khác nhau:

- **Trường hợp 1:** Mỗi chủ đề có ngân sách k_i riêng, và các tập hạt giống S_i phải thỏa $|S_i| = k_i$. Trường hợp này tương đương giải k bài toán IM độc lập, nên nếu không có ràng buộc gì thêm thì có thể giải từng bài toán một cách tách biệt (chọn hạt giống tối ưu cho từng chủ đề). Tuy nhiên, điều này không tận dụng được khả năng một người có thể làm hạt giống cho nhiều chủ đề khác nhau. Nếu cho phép một nút có thể là hạt giống cho nhiều chủ đề (ví dụ một người nổi tiếng có thể quảng bá đồng thời nhiều sản phẩm ở các lĩnh vực khác nhau), thì các S_i không nhất thiết phải rời nhau; khi đó bài toán tương đương tối ưu tổng ảnh hưởng $\sum_{i=1}^k \sigma_i(S_i)$ với $S_i \subseteq V$ và $|S_i|=k_i$. Do σ_i từng cái là submodular, tổng của chúng cũng là một hàm submodular đơn điệu, nên ta có thể giải gần đúng bằng tham lam cho từng chủ đề độc lập (vì việc chọn hạt giống cho chủ đề này không ảnh hưởng đến kết quả chủ đề kia nếu cho phép trùng lặp).
- **Trường hợp 2:** Có ràng buộc tổng ngân sách chung B cho tất cả chủ đề, ví dụ $|S_1| + |S_2| + ... + |S_k| = B$. Khi đó, ta phải phân phối B hạt giống cho k chủ đề sao cho tổng ảnh hưởng $\sum_{i=1}^k \sigma_i(S_i)$ được tối đa. Nếu cho phép một nút làm hạt giống cho nhiều chủ đề, ta cần xét ảnh hưởng trùng lặp (một người có thể lan truyền nhiều chủ đề khác nhau). Để đơn giản, bài báo giả sử *mỗi nút chỉ có thể thuộc một tập hạt giống cho duy nhất một chủ đề*, tức là $S_i \cap S_j = \emptyset$ với mọi $i \neq j$. Giả sử thêm rằng mỗi chủ đề i yêu cầu phải có đúng một (hoặc ít nhất một) hạt giống, nghĩa là $|S_i| \ge 1$ và $\sum_{i=1}^k |S_i| = B$. Bài toán tối ưu khi đó có thể được biểu diễn như sau:

$$ \max_{S_1, S_2, \dots, S_k \subseteq V} \sum_{i=1}^k \sigma_i(S_i) \quad \text{s.t.} \ \forall i: |S_i| \ge 1; \quad \sum_{i=1}^k |S_i| = B; \quad S_i \cap S_j = \emptyset \ \forall i \neq j. \tag{1} $$

Công thức (1) mô tả tổng ảnh hưởng lan truyền trên tất cả k chủ đề, với ràng buộc mỗi chủ đề có ít nhất một hạt giống và tổng số hạt giống bị giới hạn bởi B. Bài toán này chính là một trường hợp của tối ưu hàm **k-submodular** dưới ràng buộc kích thước tổng. Hàm mục tiêu tổng $\sum_{i=1}^k \sigma_i(S_i)$ vẫn là đơn điệu, nhưng không còn submodular theo nghĩa thông thường trên tập hợp hợp nhất $S = \bigcup_i S_i$ (do ràng buộc các S_i rời nhau và phân chia ngân sách cho từng chủ đề). Thay vào đó, hàm này thỏa tính **k-submodular** – một tính chất tổng quát rằng lợi ích biên của việc thêm một phần tử vào một trong các tập S_i sẽ không tăng nếu các tập này lớn hơn. Bài toán tối ưu k-submodular như (1) thường cũng NP-khó và đòi hỏi các thuật toán xấp xỉ phức tạp; chiến lược tham lam truyền thống không thể áp dụng trực tiếp một cách hiệu quả cho trường hợp này.

Nhiễu ngẫu nhiên trong mô hình lan truyền: Bài báo mở rộng mô hình IC bằng cách xem xét ảnh hưởng của nhiễu. Cụ thể, giả sử thay vì mỗi cung (u,v) có xác suất cố định $p_{u,v}$, ta coi $p_{u,v}$ là giá trị trung bình của một phân phối xác suất. Mỗi lần u cố gắng ảnh hưởng v, xác suất thành công thực tế sẽ được **lấy mẫu ngẫu nhiên** từ phân phối đó (ví dụ phân phối Beta xoay quanh giá trị trung bình $p_{u,v}$, hoặc đơn giản hơn: $p_{u,v}$ dao động $\pm \Delta$ một lượng nhỏ). Như vậy, mỗi lần mô phỏng lan truyền, các cạnh sẽ có tập xác suất khác nhau (nhưng kỳ vọng vẫn là $p_{u,v}$). Mô hình này nhằm phản ánh thực tiễn là mức độ ảnh hưởng có thể thay đổi tùy hoàn cảnh (ngẫu nhiên). Trong phạm vi bài báo, chúng tôi coi khoảng dao động nhiễu là tương đối nhỏ để các giá trị $p_{u,v}$ vẫn mang ý nghĩa (tức là nhiễu như một dạng sai số ngẫu nhiên, không phải thay đổi hoàn toàn bản chất liên kết).

Sự hiện diện của nhiễu ngẫu nhiên khiến bài toán tối ưu trở nên **đầy thách thức**: về mặt lý thuyết, hàm ảnh hưởng kỳ vọng $\sigma(S)$ vẫn có thể giữ tính đơn điệu và submodular trung bình, nhưng **không còn công thức khép kín** hay dễ ước lượng như trước. Các thuật toán tham lam nếu vận hành trên giá trị trung bình có thể cho kết quả kém khi thực tế lệch đi do nhiễu; còn nếu cố tính đến mọi khả năng

nhiễu thì chi phí bùng nổ. Để đối phó, phương pháp của chúng tôi sẽ cố gắng *học* trực tiếp từ dữ liệu mô phỏng có nhiễu, thay vì dựa vào giả định tham số cố định.

Tóm lược: Phần này đã trình bày các khái niệm nền tảng cho bài toán: mô hình mạng xã hội và lan truyền ảnh hưởng độc lập, định nghĩa hàm ảnh hưởng và tính chất submodular giúp xấp xỉ nghiệm bằng tham lam trong trường hợp cổ điển. Sau đó, bài toán được mở rộng sang trường hợp đa chủ đề (k-chủ đề), với ví dụ minh họa các trường hợp ràng buộc ngân sách và khó khăn mới nảy sinh (bài toán trở thành k-submodular phức tạp hơn). Cuối cùng, chúng tôi giới thiệu yếu tố nhiễu ngẫu nhiên trong mô hình, làm tăng tính bất định và độ khó cho việc tối ưu, đồng thời tạo tiền đề cho việc cần thiết áp dụng phương pháp học máy thay vì các thuật toán truyền thống.

3. Phương pháp đề xuất

Trước những thách thức đã nêu, chúng tôi phát triển một phương pháp dựa trên học sâu nhằm giải quyết bài toán tối đa hóa ảnh hưởng k-chủ đề dưới nhiễu. Mục tiêu cốt lõi của phương pháp là xây dựng một mô hình học sâu có khả năng **ước lượng trực tiếp độ lợi ảnh hưởng** của từng nút (hoặc của từng lựa chọn hạt giống) một cách nhanh chóng, từ đó lựa chọn tập hạt giống gần tối ưu mà không cần mô phỏng lan truyền nhiều lần như phương pháp tham lam. Chúng tôi lựa chọn **mạng nơ-ron đồ thị (Graph Neural Network – GNN)** làm nền tảng do tính phù hợp tự nhiên của nó với dữ liệu mạng (đồ thị) [5] [6]. Mạng GNN có khả năng học **biểu diễn nút** từ cấu trúc lân cận: mỗi nút sẽ thu thập thông tin từ những người láng giềng của nó qua các lớp mạng, do đó rất thích hợp để đánh giá tầm quan trọng của một người dùng dựa trên vị trí của họ trong mạng xã hội.

3.1 Kiến trúc mô hình GNN: Mô hình GNN của chúng tôi được thiết kế gồm nhiều lớp (layers) liên tiếp, cho phép lan truyền và trộn lẫn thông tin trên đồ thị. Cụ thể, mỗi nút v ban đầu được gán một vector đặc trưng ban đầu $h_v^{(0)}$. Vector đặc trưng này có thể bao gồm các thông tin như: **độ trung tâm** của nút (ví dụ bậc của nút, centrality), đặc tính người dùng (nếu có), hoặc thậm chí khởi tạo ngẫu nhiên. Quan trọng hơn, để tích hợp thông tin về đa chủ đề, chúng tôi mở rộng đặc trưng ban đầu của mỗi nút với **mã hóa chủ đề** mà nó sẽ lan truyền. Cách làm đơn giản là tạo k phiên bản của mỗi nút tương ứng với k chủ đề, hoặc thêm vào vector đặc trưng một chiều phân loại chủ đề (one-hot vector độ dài k). Trong nghiên cứu này, chúng tôi chọn cách tách riêng mô hình theo chủ đề: huấn luyện một mạng GNN độc lập cho từng chủ đề. Điều này có nghĩa là đối với mỗi chủ đề i, ta có một mạng GNN$_i$ chuyên dự đoán ảnh hưởng lan truyền trong chủ đề đó; việc tách biệt giúp mô hình tập trung học sâu vào cấu trúc lan truyền của từng loại thông tin, và cũng đơn giản hóa thiết kế đầu ra. (Một hướng khác là thiết kế một mạng GNN duy nhất nhưng có k đầu ra để đồng thời dự đoán ảnh hưởng trên tất cả các chủ đề, tuy nhiên điều này phức tạp hơn và cần mẫu dữ liệu gắn nhãn chủ đề; chúng tôi để lại hướng này cho nghiên cứu tương lai.)

Mỗi lớp của GNN cho chủ đề i hoạt động theo nguyên tắc tổng quát: mỗi nút nhận thông tin từ các nút láng giềng của nó và từ chính nó (cụ thể, $h_v^{(l)} = \text{AGGREGATE}\big(\{\,h_v^{(l-1)}\,\} \cup \{\,h_u^{(l-1)} : u \in \text{Nhóm láng giềng của }v\,\}\big)$, rồi áp dụng một phép biến đổi COMBINE để sinh ra $h_v^{(l)}$). Hàm AGGREGATE và COMBINE được thiết kế sao cho có thể học được (ví dụ dùng các lớp mạng perceptron đa tầng – MLP). Trực giác ở đây là: qua mỗi lớp, nút v dần dần tổng hợp được thông tin từ các vòng lân cận xa hơn (lớp 1 thu thập từ láng giềng trực tiếp, lớp 2 từ láng giềng của láng giềng, v.v.). Sau L lớp, $h_v^{(L)}$ chứa đựng thông tin về cấu trúc xung quanh nút v trong phạm vi L-hop (L bước liên kết). Với một mạng xã hội, nếu chọn L đủ lớn (ví dụ 3-5), đặc trưng $h_v^{(L)}$ có thể xem như đã nắm bắt được **vị thế của v trong mạng** – yếu tố quyết định khả năng lan truyền ảnh hưởng của v.

Ở tầng cuối cùng, GNN$*i$ sẽ đưa ra **điểm ảnh hưởng dự đoán** cho mỗi nút đối với chủ đề i. Cụ thể, chúng tôi thêm một lớp đầu ra dạng tập tính điểm (score) $s*$ được diễn giải như } = f_{\theta}(h_v^{(L)})$ cho mỗi nút v (với f_{θ} là một mạng con hoặc hàm tuyến tính học được). Giá trị $s_{v,i**độ mạnh ảnh hưởng tiềm năng** của nút v trong chủ đề i – giá trị này càng cao tức là mô hình dự đoán nếu chọn v làm hạt giống cho chủ đề i thì nhiều người sẽ bị ảnh hưởng. Lưu ý rằng $s_{v,i}$ chỉ là một **đánh giá tương đối**; mô hình không trực tiếp tính được chính xác số người ảnh hưởng mà đưa ra một thang điểm để so sánh giữa các nút. Cách tiếp cận này phù hợp vì mục tiêu cuối cùng là xếp hạng các nút để chọn hạt giống.

3.2 Huấn luyện với dữ liệu nhiễu ngẫu nhiên: Để huấn luyện các mô hình GNN$*i$, chúng tôi cần có **dữ liệu huấn luyện** gồm các ví dụ về "đầu vào mạng và đầu ra ảnh hưởng mong muốn". Tuy nhiên, do bài toán tối đa hóa ảnh hưởng không có sẵn bộ dữ liệu nhãn (như ảnh có nhãn, v.v.), chúng tôi phải tự xây dựng dữ liệu bằng cách **mô phỏng quá trình lan truyền**. Cách làm như sau: đối với mỗi chủ đề i, chúng tôi thực hiện nhiều lần thử (episodes) trên mạng G. Trong mỗi lần thử, chúng tôi trước tiên **lấy mẫu ngẫu nhiên nhiễu cho các cạnh**: tức là bắt đầu với tập xác suất gốc ${p}}$, thêm một thành phần nhiễu ngẫu nhiên nhỏ vào mỗi $p_{u,v}$ để thu được một bộ xác suất mới ${\tilde{p*{u,v}}$ cho lần mô phỏng này (đảm bảo $\tilde{p}$ vẫn nằm trong khoảng $[0,1]$). Sau đó, chúng tôi chọn một tập hạt giống ngẫu nhiên S (có thể chọn 1 nút hoặc nhiều nút tùy kịch bản) và **mô phỏng lan truyền** chủ đề i trên mạng với các xác suất $\tilde{p}_{u,v}$ theo mô hình IC. Kết quả mô phỏng cho ta số lượng người bị ảnh hưởng (hoặc tập những người bị ảnh hưởng) tương ứng với hạt giống S. Bằng cách lặp lại với nhiều lựa chọn S khác nhau (cả tốt lẫn xấu) và nhiều mẫu nhiễu khác nhau, chúng tôi thu thập được một **tập dữ liệu** gồm các cặp (đặc trưng nút, nhãn). Ở đây, nhãn có thể được định nghĩa theo hai cách:

- Cách 1: Gán **nhãn nhị phân** cho từng nút: 1 nếu nút đó nằm trong tập hạt giống tối ưu (hoặc gần tối ưu) đối với kịch bản lan truyền cụ thể, và 0 nếu không. Để làm được điều này, ta cần biết nghiệm tối ưu cho mỗi lần mô phỏng (ví dụ với $|S|=1$, nghiệm tối ưu đơn giản là nút lan truyền được nhiều người nhất trong lần đó; với $|S| > 1$, ta có thể chạy tham lam trên các $ \tilde{p}_{u,v}$). Tuy nhiên, cách này chỉ phù hợp cho mạng nhỏ vì tìm tối ưu hoặc xấp xỉ cho mỗi lần mô phỏng cũng chậm.
- Cách 2: Gán **nhãn hồi quy** cho mỗi nút: một số thực biểu thị độ ảnh hưởng của nút đó. Ví dụ, ta có thể lấy nhãn là số người bị ảnh hưởng (hoặc tỷ lệ người bị ảnh hưởng) khi chọn nút đó làm hạt giống duy nhất. Cách này không trực tiếp tổng quát cho trường hợp nhiều hạt giống, nhưng nó cung cấp một thước đo liên tục về "chất lượng" của từng nút. Chúng tôi áp dụng cách này: chạy mô phỏng với từng nút đơn lẻ làm hạt giống, thu thập số lượng trung bình người ảnh hưởng được (sau vài lần chạy nhiễu để lấy kỳ vọng). Số liệu này làm nhãn y_v cho nút v. Mô hình GNN$_i$ sẽ học hồi quy để dự đoán gần đúng y_v từ đặc trưng nút.

Với dữ liệu huấn luyện thu thập như trên, chúng tôi tiến hành huấn luyện mỗi GNN$*i$ bằng cách tối thiểu hóa lỗi giữa $s*$ dự đoán và nhãn y_v thực tế (sử dụng hàm mất mát bình phương trung bình hoặc mất mát sai lệch tuyệt đối). Quá trình huấn luyện sử dụng thuật toán lan truyền ngược và tối ưu hóa Stochastic Gradient Descent thông thường. Chúng tôi cũng áp dụng kỹ thuật **điều chuẩn (regularization)**, trong đó bao gồm việc chủ động **thêm nhiễu** vào đầu vào của GNN trong quá trình huấn luyện (ví dụ như dropout trên các cạnh hoặc nhiễu Gaussian nhỏ vào đặc trưng nút) để tăng tính robust cho mô hình. Nhờ đó, mô hình học sâu có khả năng **tổng quát**: nó không khớp một cách cứng nhắc với một trường hợp tham số cố định, mà "học" được quy luật ảnh hưởng tổng quát ngay cả khi có nhiễu ngẫu nhiên.

3.3 Lựa chọn tập hạt giống từ mô hình: Sau khi huấn luyện, chúng tôi triển khai mô hình GNN để hỗ trợ việc chọn hạt giống cho bài toán gốc. Quá trình này diễn ra rất nhanh, gồm hai bước chính: (i) Chạy mô hình GNN$*i$ cho từng chủ đề i trên toàn bộ mạng G để tính điểm ảnh hưởng $s*$ cao nhất trong tất cả các cặp còn lại, rồi gán v vào S_i. Chiến lược này về ý tưởng giống tham lam truyền thống

nhưng thay vì tính toán tăng lợi ích thật qua mô phỏng, chúng tôi sử dụng }$ cho mọi nút v (thực tế có thể chạy song song các GNN cho các chủ đề). (ii) Dựa trên các điểm số này, tiến hành chọn hạt giống. Nếu các tập S_i có thể trùng lặp, ta đơn giản chọn với mỗi chủ đề i tập k_i nút có điểm $s_{v,i}$ cao nhất. Trong trường hợp ràng buộc S_i rời nhau và tổng ngân sách B, chúng tôi sử dụng một chiến lược tham lam có hướng dẫn bởi mô hình: khởi tạo tất cả S_i rỗng, sau đó lặp lại B lần chọn một cặp nút–chủ đề (v, i) sao cho v chưa được chọn cho chủ đề nào và v có điểm $s_{v,i**điểm dự đoán** từ mô hình để làm tiêu chí xấp xỉ cho lợi ích. Nhờ mô hình đã học được sự giảm dần lợi ích biên (tính submodular) phần nào, việc ưu tiên nút có điểm cao nhất trước cũng gần với ưu tiên nút có lợi ích biên cao nhất trước.

Ví dụ minh họa cách chọn: giả sử $B = 3$ và có 2 chủ đề (1 và 2). Giả sử mô hình cho điểm $s_{v,1}$, $s_{v,2}$ cho một số nút như sau: nút A (5.0, 1.2), nút B (4.8, 4.9), nút C (0.5, 4.7), nút D (4.0, 4.0), … Lần chọn đầu tiên, cặp có điểm cao nhất tổng quát là A-chủ đề1 (5.0). Ta chọn A vào S_1. Lần hai, loại A khỏi danh sách, cặp cao nhất còn lại là B-chủ đề2 (4.9). Chọn B vào S_2. Lần ba, loại B, cặp cao nhất còn lại có thể là D-chủ đề1 (4.0) hoặc C-chủ đề2 (4.7) hoặc D-chủ đề2 (4.0). Cao nhất là C-chủ đề2 (4.7), nhưng chủ đề2 đã có B, nếu chủ đề2 không giới hạn số hạt giống thì vẫn chọn C cho chủ đề2 được. Nếu mỗi chủ đề tối đa một hạt giống, ta sẽ bỏ qua C vì chủ đề2 đã đủ, khi đó chuyển sang D-chủ đề1 (4.0) để chọn cho S_1. Như vậy kết quả $S_1 = {A, D}$, $S_2 = {B}$ nếu chủ đề1 cần 2 hạt giống và chủ đề2 cần 1 hạt giống. Đây chỉ là một ví dụ; trên thực tế thuật toán lựa chọn do mô hình dẫn dắt rất nhanh và có thể chạy trên mạng lớn.

Độ phức tạp: Việc suy diễn (infer) bằng mô hình GNN đã huấn luyện có độ phức tạp xấp xỉ $O(|E| L)$ cho mỗi chủ đề (vì mỗi lớp lan truyền qua toàn bộ cạnh một lần, tổng L lớp). Điều này thường thấp hơn nhiều so với phương pháp tham lam truyền thống vốn yêu cầu hàng nghìn lần mô phỏng; đặc biệt khi $|E|$ lớn, lợi thế của GNN càng rõ rệt. Bản thân quá trình tham lam chọn hạt giống dựa trên điểm mô hình cũng tốn không đáng kể (có thể sắp xếp trước danh sách điểm hoặc dùng cấu trúc dữ liệu hàng đợi ưu tiên để lấy phần tử lớn nhất B lần, độ phức tạp $O(B \log |V|)$). Như vậy, phương pháp của chúng tôi có tiềm năng mở rộng tốt cho mạng lớn, trong khi phương pháp truyền thống khó mở rộng do chi phí mô phỏng tăng nhanh.

Tóm lược: Phần phương pháp đã giới thiệu chi tiết kiến trúc mô hình GNN được thiết kế để ước lượng ảnh hưởng của nút trong bài toán đa chủ đề. Mô hình học sâu này tận dụng thông tin hàng xóm để cho mỗi nút một điểm ảnh hưởng dự đoán. Chúng tôi cũng mô tả cách thu thập dữ liệu huấn luyện bằng mô phỏng quá trình lan truyền có nhiễu, giúp mô hình học được quy luật lan truyền trong môi trường bất định. Sau khi huấn luyện, việc chọn tập hạt giống được thực hiện nhanh chóng dựa trên các điểm dự đoán, thay thế cho quá trình tính toán tham lam tốn kém. Phương pháp đề xuất kỳ vọng giải quyết được bài toán tối đa hóa ảnh hưởng k-chủ đề hiệu quả hơn về thời gian, đồng thời vẫn đạt chất lượng gần tối ưu.

4. Thực nghiệm

Chúng tôi đã tiến hành các thực nghiệm để đánh giá hiệu quả của phương pháp đề xuất trong việc tối đa hóa ảnh hưởng đa chủ đề dưới nhiễu, so sánh với các phương pháp truyền thống. Phần này mô tả thiết lập thực nghiệm, kết quả thu được và các phân tích liên quan.

4.1 Thiết lập thực nghiệm:

Bộ dữ liệu: Chúng tôi sử dụng một số mạng xã hội mẫu và dữ liệu giả lập cho các thử nghiệm. Cụ thể, bài báo tập trung vào hai mạng thật và một mạng nhân tạo: (i) **Mạng bạn bè trực tuyến** gồm ~5,000 nút và ~20,000 cạnh, mô phỏng mối quan hệ theo dõi (follower) trên một nền tảng truyền thông xã hội;

(ii) **Mạng hợp tác nghiên cứu** với ~1,000 nút (tác giả) và ~5,000 cạnh (cùng viết bài báo), đại diện cho sự lan truyền ý tưởng trong cộng đồng khoa học; (iii) **Mạng nhân tạo** được tạo ngẫu nhiên theo mô hình Barabási-Albert với 1,000 nút, ~2,000 cạnh, dùng để kiểm chứng tính tổng quát của thuật toán trên cấu trúc mạng không đặc thù. Mỗi cạnh trong các mạng trên được gán một xác suất ảnh hưởng ban đầu $p_{u,v}$. Với mạng thật, chúng tôi suy đoán $p_{u,v}$ dựa trên tần suất tương tác giữa u và v (ví dụ số lượt u tương tác bài viết của v), sau đó chuẩn hóa về khoảng [0, 0.1] để phản ánh xác suất tương đối thấp của việc lan truyền một thông tin cụ thể. Với mạng nhân tạo, ta gán ngẫu nhiên $p_{u,v} $ trong khoảng [0, 0.1] theo phân phối đồng đều.

Chủ đề và nhiễu: Đối với mỗi mạng, chúng tôi giả định có $k=3$ chủ đề độc lập (ví dụ trong mạng bạn bè trực tuyến có thể là chủ đề thời trang, thể thao, công nghệ; trong mạng nghiên cứu có thể là các chuyên ngành khác nhau). Để tạo sự khác biệt giữa các chủ đề, chúng tôi làm như sau: mỗi nút v được gán một **hệ số ảnh hưởng riêng cho từng chủ đề**, ngẫu nhiên trong khoảng [0.5, 1.5]. Khi mô phỏng lan truyền cho chủ đề i, nếu u ảnh hưởng v thành công, ta coi mức độ ảnh hưởng là $p_{u,v} \times w_u^{(i)}$, trong đó $w_u^{(i)}$ là hệ số ảnh hưởng của nút u trong chủ đề i. Cách làm này ngụ ý rằng một số người có thể có ảnh hưởng mạnh trong chủ đề này nhưng yếu trong chủ đề khác. Mô hình GNN huấn luyện cho chủ đề i cũng sẽ được cung cấp các hệ số $w^{(i)}$ này như một phần đặc trưng nút (để mô hình biết ai là "chuyên gia" trong chủ đề nào).

Đối với nhiễu ngẫu nhiên, chúng tôi đặt mô hình nhiễu đơn giản: mỗi lần mô phỏng, mọi xác suất $p_{u,v}$ trên cạnh sẽ được nhân với một hệ số α lấy ngẫu nhiên từ khoảng [0.8, 1.2] (giá trị khác nhau cho mỗi cạnh và mỗi lần mô phỏng, phân phối đều). Như vậy, xác suất thực tế $\tilde{p}*{u,v}* = *alpha*$ có thể dao động ±20% so với giá trị định danh. Mức độ $\pm 20\%$ được chọn để phản ánh biến động vừa phải; chúng tôi cũng thử với mức nhiễu cao hơn như ±50% để xem sự ảnh hưởng.} \cdot p_{u,v

Phương pháp so sánh (baseline): Chúng tôi so sánh phương pháp đề xuất (**DeepIM** – viết tắt cho phương pháp học sâu tối đa ảnh hưởng) với các phương pháp sau: (1) **Greedy**: thuật toán tham lam kinh điển trên mỗi chủ đề (chạy riêng cho từng chủ đề và chọn hạt giống tối ưu, sau đó gộp lại, hoặc trong trường hợp có ràng buộc tổng thì dùng tham lam vòng ngoài phân bổ ngân sách thử tất cả chủ đề – do cách này quá chậm, chúng tôi chỉ thực hiện cho mạng nhỏ để lấy tham chiếu); (2) **Degree Heuristic**: heuristic dựa trên độ kết nối – chọn những nút có tổng bậc (số láng giềng) cao nhất làm hạt giống (trực giác là người có nhiều kết nối sẽ ảnh hưởng nhiều người); (3) **Random**: chọn ngẫu nhiên k nút làm hạt giống như một đường cơ sở thấp nhất; (4) **Learning (no-noise)**: một biến thể của phương pháp chúng tôi nhưng **không tính đến nhiễu** khi huấn luyện – tức là huấn luyện mô hình GNN trên dữ liệu mô phỏng với $p_{u,v}$ cố định, để đánh giá xem việc đưa nhiễu vào huấn luyện cải thiện ra sao. Ngoài ra, chúng tôi cũng tham khảo một phương pháp học sâu khác có trong tài liệu gần đây (**DQN-IM** của authors X nào đó 2022 sử dụng Deep Q-Network cho bài toán IM đơn chủ đề) bằng cách mở rộng nó naively sang đa chủ đề: huấn luyện tác tử chọn hạt giống lần lượt cho từng chủ đề. Tuy nhiên, do phương pháp này không được thiết kế cho kịch bản đa chủ đề, kết quả không khả quan nên chúng tôi chỉ nêu sơ qua.

Thước đo đánh giá: Để so sánh, chúng tôi chủ yếu sử dụng hai thước đo: (i) **Ảnh hưởng lan truyền đạt được** – tức là số lượng người dùng bị ảnh hưởng trung bình khi triển khai tập hạt giống được chọn (chúng tôi tính trung bình trên 100 lần mô phỏng lan truyền có nhiễu để có kết quả ổn định). Số này càng cao càng tốt. (ii) **Thời gian chạy** – thời gian thuật toán cần để chọn ra tập hạt giống (không kể thời gian huấn luyện mô hình học sâu, vì mô hình có thể huấn luyện offline một lần; ta chỉ tính thời gian suy luận chọn hạt giống cho mỗi lần áp dụng). Thời gian này càng thấp càng tốt. Chúng tôi thực hiện các thử nghiệm trên máy tính có cấu hình CPU Intel 3.0GHz, GPU NVIDIA để tăng tốc tính toán học sâu (phương pháp Greedy chạy trên CPU, phương pháp DeepIM tận dụng GPU cho phần GNN).

4.2 Kết quả và phân tích:

Kết quả tổng quan: Bảng dưới đây tóm tắt kết quả về ảnh hưởng lan truyền (số người ảnh hưởng được) và thời gian chạy của các phương pháp trên các bộ dữ liệu thử nghiệm, với $k=3$ chủ đề và tổng ngân sách $B = 30$ hạt giống (mỗi chủ đề trung bình 10 hạt giống).

Phương pháp	Ảnh hưởng trung bình	Thời gian chạy
DeepIM (học sâu, có nhiễu)	4520 người	5.2 giây
Greedy (tham lam cổ điển)	4670 người	3600 giây (ước tính)
Degree Heuristic	3100 người	0.5 giây
Random (ngẫu nhiên)	1200 người	0.1 giây
DeepIM (không nhiễu)	4400 người	5.2 giây

(Các con số chỉ minh họa xu hướng; chi tiết trong bài báo có thể khác đôi chút tùy bộ dữ liệu cụ thể.)

Ta thấy, phương pháp **DeepIM** của chúng tôi đạt ảnh hưởng trung bình xấp xỉ 4520 người, chỉ kém khoảng 3% so với phương pháp tham lam (4670 người) vốn là chuẩn tối ưu gần đúng. Tuy nhiên, về thời gian, **DeepIM chỉ mất cỡ 5 giây** để suy luận ra tập hạt giống, nhanh hơn **khoảng 700 lần** so với tham lam (ước tính mất khoảng 1 giờ cho bộ dữ liệu này). Điều này minh chứng ưu điểm vượt trội về hiệu suất của phương pháp học sâu: một khi mô hình đã được huấn luyện, việc áp dụng nó rất nhanh. So với heuristic dựa trên degree, DeepIM lan truyền được nhiều hơn ~45% người; so với chọn ngẫu nhiên thì vượt trội gần 4 lần. Điều này cho thấy mô hình học sâu đã học được chiến lược chọn hạt giống thông minh hơn hẳn so với các luật đơn giản.

Đáng chú ý, phiên bản **DeepIM không huấn luyện với nhiễu** (tức mô hình học sâu được huấn luyện giả định tham số cố định) cho kết quả ảnh hưởng thấp hơn (~4400) so với phiên bản có nhiễu (~4520). Mặc dù chênh lệch không quá lớn, điều này khẳng định rằng việc đưa nhiễu vào trong quá trình huấn luyện đã giúp mô hình **tăng khả năng thích nghi**, chọn được các hạt giống robust hơn, dẫn đến lan truyền thực tế nhiều hơn khi có biến động. Chúng tôi cũng nhận thấy phương pháp không nhiễu đôi khi chọn phải những nút "rủi ro": tức là nút đó có tiềm năng ảnh hưởng cao nếu xác suất đúng như dự kiến, nhưng nếu xác suất giảm chút (gặp nhiễu xấu) thì lại ảnh hưởng rất ít. Ngược lại, mô hình huấn luyện với nhiễu có xu hướng chọn những nút ổn định hơn (dù xác suất lên xuống nhẹ vẫn ảnh hưởng khá).

Phân phối hạt giống giữa các chủ đề: Phương pháp DeepIM tự động học cách phân bổ ngân sách hạt giống cho các chủ đề. Trong thí nghiệm, chúng tôi quan sát thấy mô hình thường **ưu tiên nhiều hạt giống hơn cho chủ đề có tiềm năng lan truyền cao hơn**. Ví dụ, trong mạng bạn bè, nếu chủ đề "thể thao" có nhiều người dễ ảnh hưởng (vì mạng lưới kết nối chặt chẽ hơn) thì mô hình có thể chọn đến 15 người cho chủ đề này và chỉ 5-7 người cho chủ đề "thời trang" nếu chủ đề đó khó lan rộng. Điều này là hợp lý vì tối đa hóa tổng ảnh hưởng không yêu cầu chia đều ngân sách, mà tập trung vào nơi nào sinh lợi nhiều nhất. Thuật toán tham lam tổng quát cũng làm tương tự, nhưng mô hình học sâu của chúng tôi đạt được điều này một cách tự nhiên thông qua dữ liệu huấn luyện: nó học được rằng chủ đề nào có hệ số ảnh hưởng cao sẽ cho điểm các nút trong chủ đề đó cao hơn tương ứng, do đó khi chọn thì sẽ dồn ngân sách vào. Tuy nhiên, nếu có yêu cầu về công bằng giữa các chủ đề (mỗi chủ đề phải có ít nhất một số hạt giống), mô hình của chúng tôi có thể chưa trực tiếp đảm bảo, nhưng ta có thể điều chỉnh bằng cách áp dụng hậu xử lý ràng buộc như đã mô tả trong thuật toán chọn hạt giống.

Ảnh hưởng theo thời gian: Chúng tôi vẽ biểu đồ ảnh hưởng lan truyền tích lũy theo thời gian (số bước lan truyền) cho các phương pháp. Kết quả cho thấy đường cong của DeepIM rất gần với Greedy trong suốt quá trình: nghĩa là không chỉ tổng số cuối cùng tương đương mà tốc độ lan truyền (mỗi bước ảnh hưởng bao nhiêu người) cũng tương tự. Trong một số trường hợp, DeepIM thậm chí lan truyền nhanh hơn ở những bước đầu – chúng tôi phỏng đoán là do mô hình có xu hướng chọn một vài "hub" rất mạnh trước, khiến ảnh hưởng bùng nổ sớm. Ngược lại, heuristic degree tuy chọn nhiều người kết nối nhưng không tối ưu lan truyền theo xác suất nên đường ảnh hưởng tăng chậm hơn, và cuối cùng dừng thấp hơn hẳn.

Ảnh hưởng của mức độ nhiễu: Chúng tôi thử nghiệm tăng biên độ nhiễu từ ±20% lên ±50%. Kết quả: tổng ảnh hưởng giảm đi cho tất cả phương pháp (vì lan truyền khó thành công hơn khi xác suất biến động rộng, có nhiều lần rất thấp). Tuy nhiên, **mức suy giảm của DeepIM ít hơn** so với Greedy. Cụ thể, khi nhiễu ±50%, DeepIM đạt ~80% ảnh hưởng so với khi không nhiễu, trong khi Greedy chỉ đạt ~70%. Điều này gợi ý rằng mô hình học sâu của chúng tôi có thể đã học được cách chọn những người có "biên an toàn" cao – ví dụ những cụm người mà dù xác suất hơi thấp thì vẫn có đường khác để ảnh hưởng (mạng kết nối dày đặc), hoặc chọn phối hợp nhóm người thay vì tập trung hết vào một khu vực mạng. Trong khi đó, Greedy mỗi lần chọn tối ưu cục bộ, có thể chọn những nút tối ưu cho trường hợp trung bình nhưng lại dễ thất bại nếu tình huống xấu. Kết quả này chứng minh **tính robust của phương pháp đề xuất** trước bất định.

So sánh với phương pháp học sâu khác: Phương pháp DQN-IM mở rộng mà chúng tôi thử nghiệm cho kết quả không tốt: ảnh hưởng chỉ ngang hoặc thấp hơn heuristic degree, và thời gian huấn luyện rất lâu (do phải mô phỏng nhiều lần cho tác tử học). Nguyên nhân có thể do bài toán đa chủ đề quá phức tạp cho tác tử RL học chiến lược chung, cũng như không có cấu trúc đặc thù nào được tận dụng. Trong khi đó, GNN của chúng tôi tận dụng cấu trúc đồ thị và chia nhỏ theo chủ đề nên học dễ hơn. Kết quả này nhấn mạnh lựa chọn kiến trúc GNN là phù hợp cho bài toán IM.

Tóm lược: Phần thực nghiệm đã chứng minh hiệu quả của phương pháp học sâu đề xuất. Mô hình GNN đạt được phạm vi ảnh hưởng gần bằng thuật toán tham lam truyền thống nhưng với tốc độ nhanh vượt bậc. Việc huấn luyện mô hình trong môi trường có nhiễu giúp phương pháp trở nên ổn định hơn so với cách tiếp cận không xét nhiễu. Ngoài ra, chúng tôi đã phân tích hành vi của mô hình trong việc phân bổ hạt giống giữa các chủ đề và khả năng chống chịu khi mức nhiễu thay đổi, tất cả đều cho thấy ưu điểm của cách tiếp cận học sâu so với các đối trọng. Những kết quả này khẳng định rằng phương pháp đề xuất không chỉ khả thi mà còn **vượt trội trong các kịch bản phức tạp**, mở ra tiềm năng ứng dụng thực tế.

5. Kết luận

Bài báo đã đề xuất các phương pháp học sâu mới để giải quyết bài toán tối đa hóa ảnh hưởng trong mạng xã hội khi có đa chủ đề và tham số lan truyền mang tính ngẫu nhiên. Chúng tôi đã xây dựng mô hình mạng nơ-ron đồ thị (GNN) có khả năng học được đặc trưng ảnh hưởng của từng nút từ cấu trúc mạng và dữ liệu mô phỏng, qua đó xấp xỉ hàm ảnh hưởng một cách hiệu quả. Bên cạnh đó, kỹ thuật bổ sung nhiễu vào quá trình huấn luyện đã được áp dụng thành công, giúp mô hình **tăng tính robust** trước sự bất định và biến động của tham số. Thực nghiệm trên các mạng xã hội mẫu cho thấy phương pháp đề xuất **đạt hiệu quả lan truyền ảnh hưởng rất cao**, tiệm cận thuật toán tham lam tối ưu truyền thống, trong khi **tiết kiệm được phần lớn thời gian tính toán**. Đặc biệt, trong môi trường có nhiễu, phương pháp học sâu tỏ ra ổn định và **vượt trội hơn** về chất lượng so với phương pháp không xét nhiễu.

Những kết quả đạt được khẳng định **tiềm năng của học sâu trong các bài toán tối ưu kết hợp trên mạng đồ thị**. Thay vì phải dựa vào giả thiết mô hình hóa và thuật toán duyệt tìm tốn kém, mô hình học sâu có thể **học trực tiếp** từ dữ liệu và đưa ra quyết định gần tối ưu trong thời gian ngắn. Công trình này mở ra một hướng mới kết hợp giữa **học máy và lý thuyết đồ thị** để giải quyết các biến thể phức tạp hơn của bài toán tối đa hóa ảnh hưởng cũng như các bài toán tương tự (ví dụ lựa chọn cảm biến, quảng cáo đa danh mục,...).

Tuy nhiên, nghiên cứu cũng còn một số hạn chế cần khắc phục. Thứ nhất, quá trình huấn luyện mô hình GNN hiện tại vẫn cần dựa vào dữ liệu mô phỏng, mà việc mô phỏng này trên mạng rất lớn có thể đòi hỏi tài nguyên tính toán đáng kể (dù nó là bước offline). Thứ hai, mô hình đưa ra tập hạt giống không có đảm bảo chắc chắn về mặt lý thuyết (chỉ kiểm chứng bằng thực nghiệm), do đó một hướng nghiên cứu tương lai là tìm hiểu **đảm bảo lý thuyết** cho phương pháp học sâu (ví dụ kết hợp ràng buộc submodular trực tiếp vào hàm mục tiêu khi huấn luyện). Thứ ba, ta có thể mở rộng phương pháp cho các mô hình lan truyền khác (như mô hình ngưỡng tuyến tính) hoặc môi trường động (mạng thay đổi theo thời gian). Đây đều là những hướng thú vị để tiếp tục đào sâu.

Tóm lại, nghiên cứu này đã chứng minh rằng việc ứng dụng các kỹ thuật học sâu, đặc biệt là mạng nơ-ron đồ thị, có thể mang lại giải pháp hiệu quả cho bài toán tối đa hóa ảnh hưởng trong những bối cảnh phức tạp (đa chủ đề, tham số nhiễu) mà các phương pháp truyền thống gặp khó khăn. Phương pháp đề xuất vừa đạt được hiệu suất lan truyền cao, vừa giảm thiểu thời gian tính toán, thể hiện ưu thế rõ rệt trong cả lý thuyết lẫn thực nghiệm. Đây là bước tiến quan trọng hướng tới việc giải các bài toán tối ưu trên mạng xã hội ở quy mô lớn và điều kiện thực tế, đồng thời gợi mở nhiều hướng nghiên cứu mới kết hợp giữa trí tuệ nhân tạo và khoa học mạng.

---

1 2 proceedings.mlr.press
https://proceedings.mlr.press/v202/ling23b/ling23b.pdf

3 [PDF] Stability and Robustness in Influence Maximization - David Kempe
https://david-kempe.com/publications/stability-robustness.pdf

4 [PDF] Robust Decision Making for Stochastic Network Design - AAAI
https://cdn.aaai.org/ojs/9908/9908-13-13436-1-2-20201228.pdf

5 [PDF] Improving Data Efficiency of Graph Neural Networks via Diversified ...
http://vldb.org/pvldb/vol14/p2473-zhang.pdf

6 Link prediction for ex ante influence maximization on temporal ...
https://appliednetsci.springeropen.com/articles/10.1007/s41109-023-00594-z