ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Phạm Văn Cảnh

MẠNG XÃ HỘI VÀ BÀI TOÁN TỐI ƯU TỔ HỢP

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Hà Nội – 2019

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học:

1. GS. TS Thái Trà My
2. PGS. TS Hoàng Xuân Huấn

Phản biện:.....

.....

Phản biện:.....

.....

Phản biện:.....

.....

Luận án sẽ được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại .....

vào hồi giờ ngày tháng năm

Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội

MỤC LỤC

MỞ ĐẦU

Các bài toán lan truyền thông tin (information diffusion problem) trên các Mạng xã hội (MXH) được quan tâm nghiên cứu trong thời gian gần đây xuất phát từ thực tiễn cần có những giải pháp hiệu quả trong việc quản lý những thông tin trên MXH, bao gồm các nhiệm vụ: phát tán thông tin cần thiết, theo dõi, giám sát, ngăn chặn những thông tin xấu một cách hiệu quả. Việc giải quyết những bài toán này cũng góp phần nâng cao sự phục vụ, độ tin cậy của MXH đối với cộng đồng người dùng. Các bài toán này được xây dựng dưới dạng tối ưu tổ hợp và được phân loại thành 03 nhóm bài toán quan trọng là:

1. Tối đa hóa ảnh hưởng (Influence Maximization - IM*)* . Bài toán này yêu cầu chọn một tập hợp nhỏ người dùng (ngân sách giới hạn) để bắt đầu lan truyền thông tin sao cho số người bị ảnh hưởng bởi thông tin đó trên một mạng xã hội đạt cực đại.

2. Ngăn chặn thông tin (Influence Blocking - IB*)*. Mục tiêu của bài toán này là tìm một tập người dùng để loại bỏ, hoặc cách ly, hoặc bắt đầu lan truyền thông tin tốt sao cho ảnh hưởng của thông tin xấu (hoặc thông tin đối lập) đạt giá trị cực tiểu.

3. Phát hiện và giám sát thông tin (Information Detection - ID*)*: Mục tiêu của bài toán này đưa ra những giải pháp nhằm giám sát các thông tin trên MXH một cách hiệu quả.

Tuy vậy, việc giải quyết và áp dụng ba nhóm bài toán trên trong thực tiễn gặp một số thách thức chính là:

1. Lớp bài toán này thường thuộc lớp bài toán tối ưu tổ hợp NP-Khó, NP-đầy đủ. Thêm vào đó, các mô hình lan truyền thông tin đã được đề xuất cho lớp bài toán lan truyền thông tin thường là các mô hình xác suất nên việc tính toán hàm mục tiêu thường là #P-Khó. Do vậy, cần những thuật toán hiệu quả để tìm lời giải tốt trong thời gian cho phép.

2. Với sự mở rộng của quy mô các MXH (hàng triệu, tỷ người dùng), cần có những thuật toán hoặc cách tiếp cận hiệu quả hơn nữa cho những bài toán trên để nâng cao tính thực tiễn của chúng.

3. Để nâng cao hơn nữa tính ứng dụng của mỗi bài toán, cần nghiên cứu những biến thể phù hợp với thực tế đối theo các khía cạnh khác nhau như: thời gian, khoảng cách, chi phí, lợi ích, tính cạnh tranh vv...

Để nghiên cứu và tìm cách giải quyết các thách thức đặt ra, tác giả cùng các cộng sự đã chọn chủ đề nghiên cứu "***Mạng xã hội và bài toán tối ưu tổ hợp***" với mục tiêu như sau:

1. Nghiên cứu bài toán IM, IB, ID các mô hình lan truyền thông tin. Qua đó đề xuất nghiên cứu các bài toán biến thể của hai bài toán trên có tính ứng dụng trong thực tiễn.

2. Đề xuất các thuật toán hiệu quả để giải quyết các bài toán trên, trong đó đặc biệt chú trọng tới việc nâng cao chất lượng lời giải cũng như áp dụng với các mạng cỡ lớn hàng trăm nghìn cho tới hàng triệu, tỷ cạnh hoặc đỉnh.

Trong thời gian nghiên cứu, tác giả luận án đã có đóng góp sau.

1. Nghiên cứu bài toán Hạn chế tối đa thông tin sai lệch (Maximizing Misinformation Restriction-MMR) trong đó có xem xét ngân sách và thời gian hạn chế trên một số mô hình lan truyền thông tin. Tác giả chỉ ra độ phức tạp của bài toán và đề xuất các thuật toán hiệu quả cho bài toán bao gồm các thuật toán xấp xỉ và thuật toán heuristic. Luận án cũng mở rộng kết quả MMR trên mô hình ngưỡng tuyến tính xác định CLT.

2. Trong một kịch bản khác, để hạn chế sự phát tán của thông tin sai lệch đảm bảo số người bị ảnh hưởng bởi thông tin sai lệch lớn hơn một ngưỡng xác đinh, tác giả nghiên cứu bài toán Hạn chế thông tin sai lệch có chủ đích (Targeted Misinformation Blocking-TMB). Ngoài việc chỉ ra độ khó của bài toán trên các mô hình lan truyền thông tin phổ biến, tác giả đã đề xuất các thuật toán hiệu quả đối với bài toán này trên hai mô hình phổ biến.

3. Đề xuất nghiên cứu bài toán Tối đa ảnh hưởng cạnh tranh tổng quát (Budgeted Competitive Influence Maximization - BCIM) là một biến thể của IM với mục tiêu tối đa hóa ảnh hưởng trong trường hợp có sự cạnh tranh trên một số mô hình lan truyền thông tin cạnh tranh với ngân sách và thời gian hạn chế. Luận án đề xuất một thuật toán xấp xỉ hiệu quả cho bài toán BCIM. Ngoài ra, luận án cũng mở rộng nghiên cứu bài toán BCIM trên mô hình Ngưỡng tuyến tính cạnh tranh xác định (TCLT).

4. Phát triển thuật toán hiệu xấp xỉ hiệu quả cho bài toán Phát hiện thông tin sai lệch tổng quát (GMD). Luận án đề xuất SBMD (Sampling-based for Billion Scale Misinformation Detection) có tỷ lệ xấp xỉ là $1-1/e-\epsilon$ với xác xuất $1-\delta$ với $\epsilon, \delta \in (0,1)$.

Ngoài phần mở đầu và kết luận, bố cục của luận án được chia thành 06 chương như sau:

Chương 1 trình bày các kiến thức cơ bản về cơ chế lan truyền thông tin trên MXH và tình hình nghiên cứu các bài toán IM, IB, và ID.

Chương 2 trình bày kiến thức cơ bản về các bài toán tối ưu tổ hợp.

Chương 3 trình bày các kết quả nghiên cứu đối với bài toán MMR

Chương 4 trình bày các kết quả nghiên cứu đối với bài toán TMB

Chương 5 trình bày các kết quả nghiên cứu đối với bài toán BCIM

Chương 6 trình bày kết quả nghiên cứu thuật toán SBMD có tỷ lệ xấp xỉ là $1-1/e-\epsilon$ với xác xuất $1-\delta$ với $\epsilon, \delta \in (0,1)$ cho bài toán GMD

CHƯƠNG 1
TỔNG QUAN VỀ CÁC BÀI TOÁN LAN TRUYỀN THÔNG TIN TRÊN MẠNG XÃ HỘI

Sự phát tán, lan truyền thông tin trên một Mạng xã hội (MXH) được các nhà khoa học biểu diễn lại dưới dạng các mô hình phát tán thông tin. Các bài toán về lan truyền thông tin được xây dựng dưới dạng các bài toán tối ưu tổ hợp (TƯTH) trên các mô hình đó.

1.1. Các mô hình phát tán thông tin trên mạng xã hội

Sự *phát tán, khuếch tán* là một quá trình mà một sự đổi mới được truyền đạt qua các kênh nhất định theo thời gian giữa các thành viên của một hệ thống xã hội. Có ba yếu tố quan trọng trong quá trình này là: thành viên trong hệ thống xã hội, sự tương tác lẫn nhau và các kênh truyền thông. Sự phát tán thông tin trên MXH được các nhà khoa học nghiên cứu và mô hình lại dưới dạng các mô hình phát tán thông tin. Theo đó, một MXH được mô tả lại theo các thành. V là tập hợp các đỉnh của đồ thị biểu diễn tập hợp tất cả người dùng trên MXH với số đỉnh $|V| = n$. E là tập hợp các cạnh của đồ thị, biểu diễn *liên kết* giữa người dùng trong MXH.

Ngoài ra đối với đồ thị $G = (V, E)$, ta dùng các ký hiệu $N_{out}(u)$ và $N_{in}(u)$ tương ứng là tập hợp các đỉnh hàng xóm đi ra và đi vào đỉnh u, $d_{out}(u)$ và $d_{in}(u)$ tương ứng với bậc đi ra và đi vào của đỉnh u. Trong luận án này, để tiện lợi trong cách gọi tên ta coi một MXH như một đồ thị.

1.1.1. Mô hình Ngưỡng tuyến tính (LT)

Mô hình này là một trường hợp của mô hình phát tán thông tin rời rạc. Trong mô hình này, mỗi cạnh $e = (u, v) \in E$ có một trọng số $w(u, v)$ là một số thực dương biểu diễn cho các tần số tương tác, trao đổi giữa hai người dùng. Các trọng số thỏa mãn: $\sum_{u \in N_{in}(v)} w(u, v) \leq 1$. Quá trình lan truyền thông tin theo các bước rời rạc $t = 0, 1, 2, ...$ Mỗi một đỉnh u có một ngưỡng kích hoạt θ_u được chọn *ngẫu nhiên* trong khoảng $[0, 1]$. Quá trình phát tán thông tin diễn ra như sau: Tại bước $t = 0$, tất cả các đỉnh thuộc S đều bị kích hoạt, tức là $S_0 = S$. Tại bước $t \geq 1$, tất đỉnh u ở trạng thái không kích hoạt sẽ bị kích hoạt nếu tổng trọng số của các cạnh đến với đỉnh đầu được kích hoạt ở các bước trước đó lớn hơn ngưỡng kích hoạt θ_u, tức là: $\sum_{v \in N_{in}(u) \cap S_{t-1}} w(v, u) \geq \theta_u$. Khi một đỉnh ở trạng thái kích hoạt, nó sẽ giữ nguyên trạng thái. Quá trình lan truyền kết thúc khi giữa hai bước không có thêm đỉnh nào bị kích hoạt.

1.1.2. Mô hình Bậc độc lập (IC)

Trong mô hình IC, mỗi cạnh $(u, v) \in E$ được gán một *xác suất ảnh hưởng* (*influence probability*) $p(u, v) \in [0, 1]$ biểu diễn mức độ ảnh hưởng của đỉnh u với đỉnh v. Trong mô hình này mỗi đỉnh u đã bị kích hoạt tại bước $t \geq 0$ có một cơ hội duy nhất để kích hoạt các đỉnh hàng xóm chưa kích hoạt ở bước $t + 1$. Quá trình lan truyền kết thúc khi giữa hai bước không có thêm đỉnh nào bị kích hoạt.

1.1.3. Mô hình cạnh trực tuyến (live-edge)

Để thuận tiện trong việc tính toán hàm mục tiêu và thiết kế các thuật toán trong các bài toán lan truyền thông tin. Mô hình này sinh ra các đồ thị mẫu g từ đồ thị ban đầu. Tuy nhiên việc sinh đồ thị mẫu này ứng với mỗi mô hình là khác nhau. Với mô hình LT. Gọi $\Pr[g \sim G]$ là xác suất sinh ra đồ thị mẫu g từ G. Ảnh hưởng của tập hạt giống S trên cả hai mô hình là

$$\sigma(S) = \sum_{g \sim G} \Pr[g \sim G] R(g, S) \tag{1.1}$$

Trong đó $R(g, S)$ là tập các đỉnh có thể đi tới từ S trên đồ thị g.

1.2. Một số bài toán lan truyền thông tin trên MXH

Trong phần này, luận án trình bày một các bài toán IM, IB và ID.

1.2.1. Tối đa ảnh hưởng (IM)

Bài toán tối đa hóa ảnh hưởng *(Influence Maximization-IM)* có ý nghĩa lớn trong hoạt động tiếp thị (marketing) đối với các hoạt động kinh doanh trên MXH hiện nay. Bài toán được phát biểu cụ thể như sau: Cho một MXH $G = (V, E)$ trên mô hình phát tán thông tin $\mathcal{M}$. Cho trước số nguyên dương $k > 0$ (ngân sách), tìm tập hạt giống $S \subseteq V, |S| = k$ sao cho *ảnh hưởng* của S là lớn nhất ?

Đây là bài toán thuộc lớp NP-Khó và việc tính toán hàm ảnh hưởng là #P-Khó. Về thuật toán có hai hướng tiếp cận chính là: thuật toán xấp xỉ đảm bảo lời giải về mặt lý thuyết và các thuật toán gần đúng dựa theo: đường đi, độ đo trong mạng, và cấu trúc cộng đồng. Các bài toán biến thể của IM được quan tâm nghiên bao gồm: chi phí và lợi ích, chủ đề, khoảng cách, thời gian, địa điểm.

1.2.2. Ngăn chặn ảnh hưởng (IB)

Ngược lại với IM, bài toán IB nhằm mục đích hạn chế sự phát tán, lan truyền thông tin của một nguồn tin cho trước. Mục tiêu của các bài toán này nhằm hạn chế sự phát tán của các yếu tố xấu trên MXH, bao gồm: tin xấu, thông tin sai lệch, hoặc sự phát tán của virus, các tư tưởng cực đoan, vv.. Các phương pháp có thể hạn chế ảnh hưởng của một nguồn phát tán cho trước được đề xuất bao gồm (1) Loại bỏ tập đỉnh hoặc cạnh hoặc tiêm vắc-xin (theo ngôn ngữ dịch tễ học) vào tập đỉnh hoặc cạnh để miễn nhiễm với ảnh hưởng.(2) Tẩy nhiễm thông tin: chọn tập đỉnh để bắt đầu phát tán các ảnh hưởng tích cực để chống lại ảnh hưởng của thông tin tiêu cực.

1.2.3. Phát hiện thông tin (ID)

Bài toán này được nghiên cứu sau hai bài toán IM và IB tuy nhiên vai trò của nó vô cùng quan trọng trong việc phân tích, quản lý kịp thời các thông tin xấu trên MXH. Ứng dụng to lớn của bài toán này là phát hiện thông tin sai lệch, tin giả mạo, tin đồn trên các MXH. Mục tiêu của bài toán này là tìm tập các đỉnh để đặt giám sát sao cho khả năng phát hiện thông tin sai lệch là lớn nhất.

CHƯƠNG 2
BÀI TOÁN TỐI ƯU TỔ HỢP VÀ MỘT SỐ PHƯƠNG PHÁP GIẢI CÁC BÀI TOÁN TỐI ƯU TỔ HỢP

2.1. Bài toán TƯTH

Mỗi bài toán TƯTH ứng với một bộ ba (S, f, Ω), trong đó S là tập hữu hạn trạng thái (lời giải tiềm năng hay phương án), f là hàm mục tiêu xác định trên S, còn Ω là tập các ràng buộc. Mục tiêu của các bài toàn này là tìm cực đại hoặc cực tiểu hàm số f trên tập S

2.2. Phân loại các lớp bài toán trong TƯTH

Định nghĩa 2.1. Lớp bài toán P, và NP được định nghĩa như sau **P** (Polynomial-time): là lớp các bài toán giải được bằng thuật toán đơn định trong thời gian đa thức.
NP (Non-Deterministic Polynomial-time): là lớp tất cả các bài toán giải được bằng thuật toán không đơn định trong thời gian đa thức.

Định nghĩa 2.2. Lớp bài toán #P là lớp bài toán xác định các hàm $f(x)$ bằng với số đường đi từ cấu hình ban đầu tới một cấu hình chấp nhận được trong máy Turing không đơn định trong thời gian đa thức theo kích cỡ của đầu vào x.

2.3. Một số phương pháp giải bài toán TƯTH

2.3.1. Thuật toán xấp xỉ

Định nghĩa 2.3. Ta nói thuật toán xấp xỉ $\mathcal{A}$ cho lời giải là $s \subseteq S$ có tỷ lệ xấp xỉ (approximation ratio) thuật toán này là $\rho > 0$ nếu nó thực hiện trong thời gian đa thức theo kích cỡ của thể hiện đầu vào của bài toán và thỏa mãn $\frac{f(s)}{\mathsf{OPT}} \geq \rho$ Trong trường hợp cần tìm hàm f cực tiểu (tìm giá trị nhỏ nhất), thì tỷ lệ tối ưu được định nghĩa là: $\frac{f(s)}{\mathsf{OPT}} \leq \rho$

Trong trường hợp bài toán tìm cực đại $\rho < 1$, còn bài toán tìm cực tiểu thì $\rho > 1$.

Thuật toán tham lam (Greedy Algorithm) là một trong những thuật toán phổ biến và có tính ứng dụng cao bởi tính đơn giản và độ phức tạp về thời gian thấp. Nếu hàm tham lam của một thuật toán tham lam có tính chất submodular thì việc phân tích tỉ lệ xấp xỉ trở nên đơn giản hơn nhiều.

Ngoài ra để ước lượng kỳ vọng của một biến ngẫu nhiên X trong không gian mẫu Ω rất lớn, người ta thường dùng phương pháp này để đưa về một giá trị ước lượng đủ tốt.

Định nghĩa 2.4 ((δ, ϵ)-xấp xỉ)**.** Cho biến ngẫu nhiên X trên không gian mẫu Ω, μ là kỳ vọng của X. Ta nói $\hat{\mu}$ là một (δ, ϵ)-xấp xỉ của nếu thỏa mãn:

$$\Pr[(1-\epsilon)\hat{\mu} \leq \mu \leq (1+\epsilon)\hat{\mu}] \geq 1 - \delta \tag{2.1}$$

2.3.2. Thuật toán heuristic cấu trúc

Một phương pháp rất được ưa chuộng trong việc giải các bài toán NP-Khó là các thuật toán heuristic. Những thuật toán này cho kết quả gần đúng trong thời gian chấp nhận được.

CHƯƠNG 3
NGĂN CHẶN THÔNG TIN SAI LỆCH VỚI RÀNG BUỘC VỀ NGÂN SÁCH VÀ THỜI GIAN

3.1. Đặt vấn đề và phát biểu bài toán

3.1.1. Đặt vấn đề

Dù các nghiên cứu trước giải quyết vấn đề ngặn chặn ảnh hưởng của nguồn tin cho trước trong nhiều trường hợp và mô hình khác nhau. Tuy nhiên, một số thách thức đặt ra mà các nghiên cứu trước còn bỏ qua là:

1. Chưa xem xét yếu tố thời gian trong quá trình lan truyền. Việc ngăn chặn sự phát tán của nguồn tin càng sớm thì hậu quả, thiệt hại càng nhỏ.

2. Chưa xem xét chi phí trong ngăn chặn thông tin sai lệch. Để đảm bảo tính tự do ngôn luận cho các MXH, không thể loại bỏ quá nhiều nút và việc loại bỏ cũng như miễn nhiễm thông tin với mỗi đỉnh khác nhau là khác nhau, do vậy công việc này đối với mỗi đỉnh cần có những chi phí khác nhau.

3. Chưa thực hiện việc ngăn chặn trên mô hình LT.

Để giải quyết những thách thức trên, luận án đề xuất nghiên cứu bài toán Ngăn chặn tối đa thông tin sai lệch với ràng buộc về ngân sách và thời gian (MMR) như sau:

3.1.2. Phát biểu bài toán

Trước hết để xử lý được ràng buộc thời gian hạn chế (Time contraint Linear Threshold - TLT), chúng tôi đề xuất một mô hình phát tán thông tin có ràng buộc thời gian dựa trên việc mở rộng mô hình truyền thống LT tổng quát.

*Mô hình ngưỡng tuyến tín ràng buộc thời gian (*TLT*).* Mô hình này xét sự lan truyền của nguồn thông tin sai lệch có hạn chế thời bước lan truyền. Ta tạm thời đồng nhất thời gian lan truyền với bước lan truyền với giả thuyết rằng thời gian lan truyền thông tin từ người dùng này tới người dùng khác là như nhau.

Cho một MXH $G = (V, E)$, mô hình TLT cơ bản giống với mô hình LT tuy nhiên sự khác nhau là số *bước lan truyền được giới hạn* trước là một số nguyên dương d. Cụ thể như sau: Quá trình lan truyền thông tin theo các bước thời gian rời rạc, với thời gian $t = 0, 1, 2, \ldots, d$. Ảnh hưởng của S ở thời gian t là:

$$\sigma_d(S) = \sum_{g \sim G} \Pr[g \sim G] R_d(g, S) \tag{3.1}$$

Gọi ảnh hưởng của S sau khi loại bỏ A sau thời gian d là $\sigma_t(S, A)$, ta có

$$\sigma_d(S, A) = \sum_{g \sim G[V \setminus A])} \Pr[g \sim G] R_d(g, S) \tag{3.2}$$

Hàm mục tiêu là giá trị độ giảm của ảnh hưởng khi loại đi tập đỉnh A:

$$h(A) = \sigma_d(S, \emptyset) - \sigma_d(S, A) \tag{3.3}$$

Giả sử mỗi đỉnh $u \in V$ có một chi phí để loại bỏ là $c(u) \geq 0$, $v \in V \setminus S$ và một ngân sách giới hạn $L > 0$. Bài toán MMR được phát biểu như sau

Định nghĩa 3.1. Bài toán MMR

- **Input:** Một MXH $G = (V, E, w)$ trên mô hình TLT, nguồn phát TTSL $S \subset V$, thời gian giới hạn d, chi phí giới hạn $L > 0$
- **Output:** Tập $A \subseteq V \setminus S$ với tổng chi phí $c(A) = \sum_{u \in A} c(u) \leq L$ sao cho $h(A)$ đạt cực đại?

3.2. Độ phức tạp của bài toán

Định lý 3.1. MMR *là NP-Khó trong mô hình* TLT *kể cả trong trường hợp đồ thị* G *là cây có gốc.*

Định lý 3.2. *Tính toán hàm mục tiêu* $h(A)$ *là bài toán* $\#P$*-Khó trên mô hình* TLT *kể cả trong trường hợp* A *chỉ có một đỉnh.*

3.3. Các thuật toán cho MMR

Trong mục này, luận án đề xuất hai hướng tiếp cận cho bài toán: thiết kế thuật toán xấp xỉ cho bài toán, thiết kế thuật toán heuristic hiệu quả với thời gian chạy đủ tốt.

3.3.1. Thuật toán xấp xỉ

a. Thuật toán FPTAS trong trường hợp cây. Xét bài toán MMR trong trường hợp đồ thị G có dạng một cây có gốc tại duy nhất một đỉnh nguồn $S = \{I\}$ (gọi là TMMR). Thuật toán chia làm hai giai đoạn, chi tiết được mô tả trong Thuật toán 1.

Algorithm 1: Thuật toán FPTAS cho bài toán TMMR

Input: $G = (V, E, w), I, d, \epsilon > 0$.
Output: A
`// Phase 1. Preprocessing`
1 Find sub-tree T_I of G root I has depth d.
2 CalBen$(T_I, u), \forall u \in T_I$
3 $M = \max\{h(v) | v \in V, c(v) \leq L\}$, $K = \frac{\epsilon M}{n}$
4 Let $h'(u) = \lfloor \frac{h(u)}{K} \rfloor$
`// Phase 2: Dynamic Programming algorithm`
Compute $F^u(p), F_i^u(p)$ using the recursions.
Find an optimal solution, call A' , by tracing from $\max\{p | F^u(p) \leq L\}$
return A'

Định lý 3.3. *Thuật toán 1 là một FPTAS cho bài toán* T-MMR.

b. Thuật toán xấp xỉ trong trường hợp tổng quát. Trong trường hợp này, hàm mục tiêu có các tính chất sau

Định lý 3.4. $h(\cdot)$ *là hàm đơn điệu tăng và submodular*

Dựa trên kết quả này, luận án đề xuất thuật toán IGA cho tỷ lệ xấp xỉ là $1 - \frac{1}{\sqrt{e}}$. Chi tiết của phương pháp này được trình bày ở thuật toán 2. Gọi R thời gian tính toán hàm

Algorithm 2: Thuật toán tham lam cải tiến (IGA)

Input: $G = (V, E, w), L, d, S$.

Output: A

1 $U \leftarrow$ remove all nodes having cost greater than L from V

2 A_1= Result of Greedy;

3 $v_{max} = \arg\max_{u \in U} h(v)$

4 $A = \arg\max\{h(A_1), h(v_{max})\}$

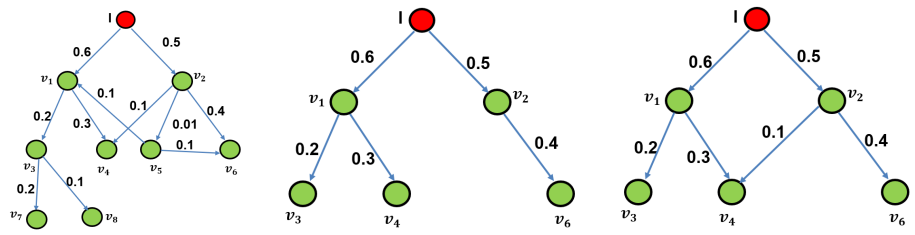
5 **return** A;

$h(A), \forall A \subseteq V$, độ phức tạp của thuật toán 2 trong thời gian $\mathcal{O}(n^2 R)$.

c. Thuật toán tham lam tăng tốc (*SG*) Để áp dụng được thuật toán IGA trên dữ liệu thực, luận án đề xuất một phương pháp nhằm để tăng tốc thuật toán tham lam, gọi là Thuật toán tham lam mở rộng (Scalable Greedy-SG). Ý tưởng chính của thuật toán này là đề *ước lượng hàm mục tiêu* trên một tập mẫu xác định.

3.3.2. Thuật toán Heuristic

Xây dựng DAG từ đồ thị ban đầu. Hình 3.1 là một ví dụ mô tả lại các bước xây dựng DAG với trên G với $d = 2, \theta = 0.051$. Hình 3.1(a) là đồ thị G, hình 3.1(b) là kết quả xây dựng $MIOA(G, I, d, \theta)$. Tại Hình 3.1(c), DAG được tạo tành bằng cách thêm một cạnh hợp lệ với quy tắc trên là (v_2, v_4). . Trên DAG, luận án đề xuất một độ đo gọi là



(a) (b) (c)

Hình 3.1: Ví dụ xây dựng DAG từ G

vai trò lan truyền (propagation role) nhằm ước lượng hàm mục tiêu. Độ đo vai trò lan truyền của đỉnh u dựa trên hai yếu tố. Ảnh hưởng từ nguồn I đến u (ký hiệu là $f_{in}(u)$):

Bảng 3.1: Thời gian chạy (giây) của các thuật toán với chi phí tổng quát và $L = 100$

	$d = 3$		$d = 4$		$d = 5$	
Dataset	PR-DAG	SG	PR-DAG	SG	PR-DAG	SG
Oregon	800.30	20556.32	880.92	26585.70	839.97	27290.34
Epinions	9255.00	18421.07	10084.91	24359.07	9984.81	26665.14
Gnutella	172.53	1152.47	440.92	1721.73	676.95	1996.49
EU Email	19973.13	-				

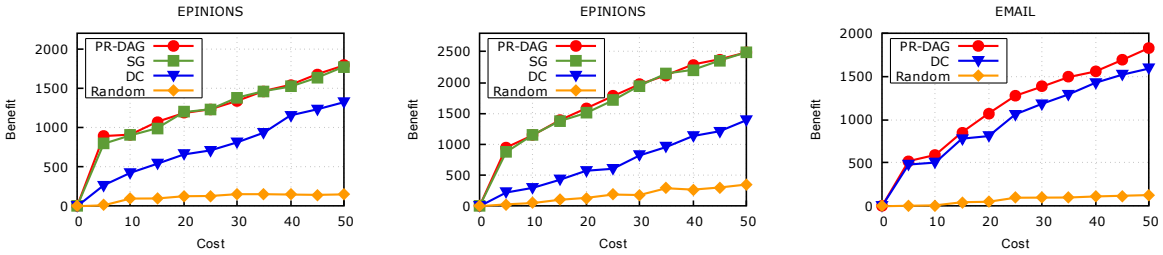
$f_{in}(u) = \sum_{P \in \mathsf{P}(\mathcal{D},I,u)} \mathsf{Inf}(P)$. Ảnh hưởng từ u đến các đỉnh khác (ký hiệu là $f_{out}(u)$). $f_{out}(u) = \sum_{v \in U} \sum_{P \in \mathsf{P}(\mathcal{D},u,v)} \mathsf{Inf}(P)$. Trong đó $\mathsf{P}(\mathcal{D}, u, v)$ là tập các đường đi từ u đến v trên DAG $\mathcal{D}$. Vai trò lan truyền của u được tính như sau: $r(u) = f_{in}(u) \cdot f_{out}(u)$. Thuật toán PR-DAG hoạt động dựa trên các bước của IGA. Trong đó, ảnh hưởng của I đến các đỉnh khác được ước lượng bởi $\sigma(I) \approx \mathsf{EstInf}(\mathcal{D}, I) = \sum_{u \in \mathcal{D}} f_{in}(u)$. Độ phức tạp của PR-DAG là $\mathcal{O}(k_1(m_\theta + n_\theta \log n_\theta))$.

3.3.3. Thực nghiệm và kết quả

Luận án tiến hành thực nghiệm để so sánh các các thuật toán đề xuất cho MMR bao gồm: SG và PR-DAG với các thuật toán cơ sở thường được dùng trong các bài toán về lan truyền thông tin được liệt kê dưới đây. Random: Lựa chọn ngẫu nhiên các tập đỉnh A với ngân sách nhỏ hơn L. DC Degree Centrality)

3.3.3.1. Kết quả thực nghiệm

Luận án đánh giá sự hiệu quả của các thuật toán thông qua hai tiêu chí: Chất lượng lời giải (hàm mục tiêu) và thời gian chạy của các thuật toán. Để đánh giá toàn diện và đầy đủ, hiệu quả của các thuật toán được đánh giá trong hai trường hợp: *Chi phí tổng quát (general cost)* và *Chi phí đồng nhất (unit cost)* Các thuật toán đề xuất PR-DAG và



Epinions, $d = 3$ Epinions, $d = 4$ Email, $d = 3$

Hình 3.2: Chất lượng lời giải của các thuật toán với chi phí đồng nhất

SG cho kết quả vượt trội so với các thuật toán cơ sở. SG và PR-DAG cho kết quả tương tự nhau trên hầu hết các bộ dữ liệu. Điều này cho thấy hiệu quả của việc xây dựng DAG nhằm xấp xỉ hóa hàm mục tiêu cùng như hàm ảnh hưởng trong PR-DAG. Thời gian chạy

của PR-DAG nhanh hơn so với SG từ 32.5 đến 45 lần. Khả năng của SG bị giới hạn trên các bộ dữ liệu lớn trong khi PR-DAG có khả năng mở rộng trên các bộ dữ liệu này.

3.3.4. Ngăn chặn thông tin sai lệch trên mô hình ngưỡng tuyến tính xác định

Luận án mở rộng các kết quả nghiên cứu cho bài toán MMR trên mô hình Ngưỡng tuyến tính xác đinh DLT (Deterministic Linear Threshold) gọi (là bài toán $\mathsf{MMR_D}$).

3.3.4.1. Định nghĩa bài toán và độ phức tạp

Trên mô hình này, quá trình lan truyền cũng được giới hạn trong thời gian d giống TLT. Sự khác giữa hai mô hình là các ngưỡng kích hoạt $\theta_v, v \in V$ trong TDLT *được cho trước*.

Định nghĩa 3.1. (Bài toán $\mathsf{MMR_D}$) Cho MXH $G = (V, E)$, ngân sách k, tập nguồn S trên mô hình DTLT. Bài toán yêu cầu Tìm $A, |A| = k$ sao cho $h(A)$ lớn nhất?

Định lý 3.5. *Không có thuật toán xấp xỉ trong thời gian đa thức có tỷ lệ* $n^{1-\epsilon}$ *cho bài toán* $\mathsf{MMR_D}$ *trên mô hình* TDLT *với* $0 < \epsilon < 1$.

3.3.4.2. Các thuật toán đề xuất cho $\mathsf{MMR_D}$

a. Thuật toán tham lam. Một giải pháp đơn giản cho việc tìm lời giải cho các bài toán lan truyền thông tin là thuật toán tham lam. Luận án đề xuất thuật toán Tham lam bằng việc lần lượt chọn các đỉnh u có làm cho hàm mục tiêu $\delta(A, u)$ $\delta(A, u) = h(A \cup \{u\}) - h(A)$. Độ phức tạp của thuật toán này là $\mathcal{O}(kn_d(m_d + n_d)$.

b. Thuật toán FLE. Luận án đề xuất một thuật toán mới có tên là FLE *(Fast And Effective Limiting Epidemics)*. Thuật toán này dựa trên tư tưởng tham lam nhưng có sự cập nhật nhanh và tính toán gần đúng hàm $\delta(A, u)$ qua việc tính toán nhanh các tham số $\alpha(u), \beta(u)$ Tham số $\alpha(u)$ đánh giá khả năng cứu được các đỉnh khác khi đỉnh đã bị kích u bị loại bỏ, tham số $\beta(u)$ có thể ước lượng thay thế cho $\delta(A, u)$. Ý tưởng chính của thuật toán là chọn ra các đỉnh một cách lần lượt theo đánh giá của hai hàm α và β. Ban đầu, tập được khởi tạo $A = \emptyset$ và $U = V_d$. Trong mỗi bước, ta chọn đỉnh u $\beta(u)$ lớn nhất trong đồ thị còn lại. Trường hợp tất cả các đỉnh đều có giá trị $\beta(u)$ là 0, ta chọn đỉnh u có $\alpha(u)$ cực đại. Độ phức tạp chung của Thuật toán FLE là $O(k(m_d + n_d)$.

3.3.4.3. Kết quả thực nghiệm với $\mathsf{MMR_D}$

Các kết quả chỉ ra thuật toán FLE cho kết quả hàm mục tiêu gần như tương tự với Greedy tuy nhiên thời gian nhanh hơn gấp nhiều lần. Hai thuật toán đề xuất cũng cho kết quả hơn hẳn các thuật toán cơ sở.

CHƯƠNG 4
NGĂN CHẶN THÔNG TIN SAI LỆCH CÓ CHỦ ĐÍCH

Một vấn đề phát sinh trong thực tế mà các nghiên cứu trước bỏ qua là ta không biết phải *loại bỏ bao nhiêu đỉnh hoặc cạnh (ngân sách) để ngăn chặn được đáng kể hoặc sự phát tán TTSL trên diện rộng?* Ví dụ: Cần loại bỏ bao nhiêu tài khoản hoặc liên kết trong một MXH để số người dùng không bị ảnh hưởng bởi nguồn TTSL là 5,000. Điều này có ý nghĩa lớn để bảo vệ sự tin cậy của các MXH vì nếu tỷ lệ số đỉnh bị ảnh hưởng bởi TTSL càng lớn thì tính chính xác của thông tin cũng như tính đáng tin cậy của MXH đó càng giảm.

Thúc đẩy bởi yêu cầu này, nghiên cứu sinh cùng các cộng sự đã nghiên cứu bài toán *Ngăn chặn TTSL với mục tiêu cho trước (Targeted Misinformation Blocking-*TMB*)* nhằm mục đích tìm tập đỉnh S có số đỉnh nhỏ nhất để loại bỏ khỏi một MXH sao cho ảnh hưởng của nguồn thông tin cho trước giảm đi một lượng lớn hơn ngưỡng γ cho trước.

4.1. Phát biểu bài toán và độ phức tạp của bài toán

Định nghĩa 4.1. (Bài toán ngăn chặn TTSL có chủ đích-TMB])

- **Input:** MXH $G = (V, E)$ trên mô hình phát tán thông tin $\mathcal{M}$, ngưỡng $\gamma \in (0, |V|)$.
- **Output:** Tìm tập đỉnh $A \subseteq V \setminus S$ sao cho $h(A) \geq \gamma$.

Định lý 4.1. *Bài toán* TMB *thuộc lớp #P-Khó trên mô hình* LT *ngay cả trong trường hợp S của một đỉnh duy nhất.*

Định lý 4.2. TMB *thuộc lớp NP-Khó trên mô hình* IC *ngay cả trong trường hợp G là đồ thị không có chu trình.*

4.2. Các thuật toán đề xuất cho TMB trên mô hình LT

Trên mô hình này, hàm mục tiêu được chứng minh có tính chất *đơn điệu tăng* và *submodular*.

4.2.1. Thuật toán tham lam

Dựa trên kết quả của định lý việc chứng minh hàm $h()$ là đơn điệu tăng và submodular, luận án đề xuất thuật toán tham lam có tỷ lệ xấp xỉ là $1 + \ln \frac{\gamma}{\epsilon}$.

4.2.2. Thuật toán STMB-LT

Áp dụng ý tưởng của thuật toán tham lam, phương pháp mô phỏng Monte Carlo cũng nhu việc cập nhật nhanh giá trị hàm mục tiêu sau mỗi vòng lặp, luận án đề xuất thuật toán mới có tính thực tiễn cũng như khả năng tìm kiếm lời giải đối với dữ liệu lớn có tên là STMB-LT *(Scalable Targeted Misinformation Blocking)*. Thuật toán STMB-LT có độ phức tạp là $\mathcal{O}(\eta(m + qn))$.

Algorithm 3: Thuật toán STMB-LT

Input: Graph $G = (V, E, w)$, $S = \{s_1, s_2, .., s_q\}$, $\gamma > 0$

Output: set of nodes A

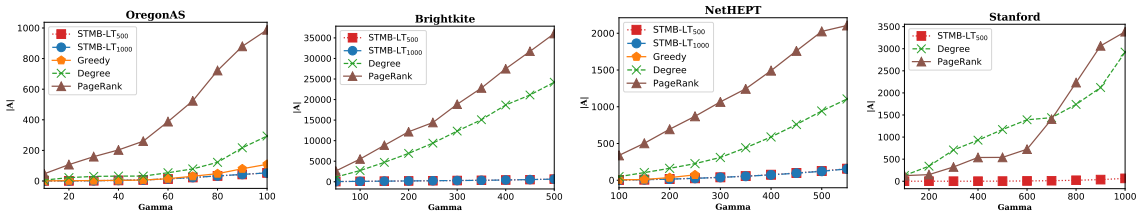
$A \leftarrow \emptyset$; $(G', I) \leftarrow \texttt{Merge}(G, \mathcal{S})$.
Remove all node, I can't reach in G.
Generate η sample graphs and set η trees $\mathcal{L} = \{T_I^1, T_I^2, \ldots, T_I^\eta\}$
For each $T_I \in \mathcal{L}$, calculate $h(u, T_I)$ for all $u \in T_I$ (by using DFS algorithm).
for $u \in V$ **do**
    $u.\delta(u) \leftarrow \frac{1}{\eta} \sum_{T_I \in \mathcal{L}} h(u, T_I)$; $u.cur \leftarrow 1$
    Insert element u into Q with $u.\delta(u)$ as the key
end
$h_{max} \leftarrow 0$; $iteration \leftarrow 1$
while $h_{max} < \gamma - \epsilon$ **do**
    $u_{max} \leftarrow dequence\ Q$
    if $u_{max}.cur = iteration$ **then**
        $A \leftarrow A \cup \{u_{max}\}$; $iteration \leftarrow iteration + 1$
        foreach $T_I \in \mathcal{L}_c$ **do**
            If $u_{max} \in T_I$, remove node u_{max} and update $h(v, T_I)$, $\forall v \in T_I$.
        end
        $h_{max} \leftarrow h_{max} + u_{max}.\delta(u_{max})$
    else
        $u_{max}.\delta(u_{max}) \leftarrow \frac{1}{\eta}\big(\sum_{T_I \in \mathcal{L}} h(I, T_I) - \sum_{T_I \in \mathcal{L}} h(I, T_I \setminus u_{max})\big)$
        $u_{max}.cur = iteration$; re-insert u_{max} into Q
    end
end
return A;

4.2.3. Thực nghiệm và kết quả

Luận án tiến hành các thực nghiệm để so sánh các thuật toán đề xuất cho TMB với các thuật toán cơ sở . Các kết quả chỉ ra thuật toán STMB-LT cho kết quả tốt nhất trong các thuật toán, tập đỉnh cần loại bỏ có số lượng ít nhất trong cùng một ngưỡng γ. Hai phiên bản của STMB-LT là STMB-LT$_{500}$ và STMB-LT$_{1000}$ gần như cho kết quả tương tự nhau. STMB-LT$_{500}$ chạy nhanh hơn Greedy đến 203.9 lần còn STMB-LT$_{1000}$ chạy nhanh hơn Greedy đến 96.1 lần.

4.3. Thuật toán cho TMB trên mô hình IC

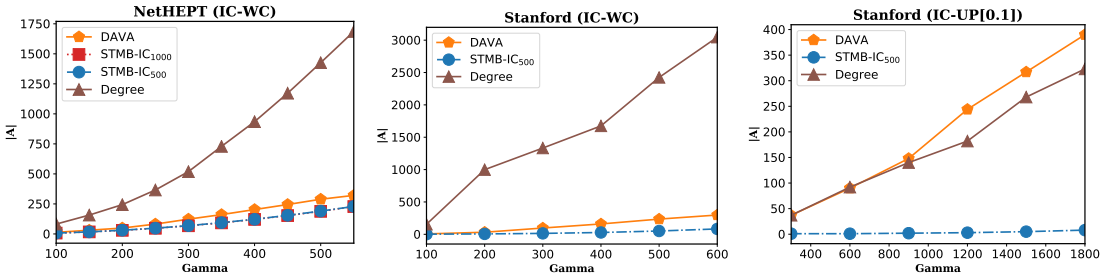
Đặc tính của mô hình IC khác với LT do đó hàm mục tiêu có tính chất khác so với mô hình LT. Cụ thể, hàm mục tiêu trên mô hình này không có tính chất submodular và supermodular. Do vậy, không thể áp dụng trực tiếp thuật toán tham lam để đạt được tỷ



Hình 4.1: So sánh chất lượng lời giải của các thuật toán cho TMB trên mô hình LT

lệ xấp xỉ. Trong mục này, luận án đề xuất các thuật toán cho bài toán TMB trên mô hình IC bao gồm: (1) Xây dựng hệ quy hoạch tuyến tính cung cấp một cách tiếp cận lý thuyết cho việc tìm lời giải tối ưu của bài toán. Nó có thể được áp dụng như một công cụ cho các thuật toán tìm lời giải khác. (2) Thuật toán Heuristics STMB-IC. Thuật toán này dựa trên việc thay đổi STMB-LT trong đó có sự cải tiến và thay đổi để phù hợp với IC.

4.3.1. Thực nghiệm và kết quả



Hình 4.2: So sánh chất lượng lời giải của các thuật toán trên mô hình IC

STMB-IC_{500} cho kết quả tương tự với STMB-IC_{1000} trong tất cả các trường hợp. Nói chung, STMB-IC cho kết quả tốt nhất trong các thuật toán.

Trong tất cả các trường hợp, STMB-IC trả về tập đỉnh A với số đỉnh nhỏ hơn DAVA và Degree. Thuật toán DAVA hoạt động không tốt trên các tập dữ liệu Brightkite và Stanford trên mô hình IC-UP[0.1]. Thuật toán STMB-IC_{500} cho thời gian chạy nhanh nhất trong tất cả các thuật toán. Trung bình, STMB-IC_{500} chạy nhanh gấp hai lần so với STMB-IC_{1000} và nhanh gấp 15.7 lần so với DAVA.

CHƯƠNG 5
TỐI ĐA ẢNH HƯỞNG CẠNH TRANH VỚI RÀNG BUỘC VỀ THỜI GIAN VÀ NGÂN SÁCH

Bài toán Tối đa ảnh hưởng cạnh tranh (CIM) được quan tâm nghiên cứu trong thời gian gần đây do tính ứng dụng của nó trong hoạt động lan truyền tiếp thị sản phẩm trên các MXH. Các nghiên trên đã tập trung nghiên cứu bài toán CIM với nhiều mục tiêu khác nhau. Tuy nhiên có một số hạn chế sau:

- Các nghiên cứu thường bỏ qua sự ràng buộc vê thời gian và ngân sách (chi phí khác nhau để bắt đầu quá trình lan truyền) trong việc giải quyết bài toán.
- Các thuật toán đề xuất cho các trường hợp khả năng mở rộng còn hạn chế, chưa áp dụng được với các mạng cỡ lớn hàng trăm nghìn và triệu đỉnh.
- Việc giải quyết sự cạnh tranh trong mô hình chưa phù hợp với thực trạng cạnh tranh trong các MXH thực.

Trong chương này, luận án nghiên cứu bài toán Tối đa ảnh hưởng cạnh tranh với ràng buộc về thời gian và ngân sách (BCIM). Đây là bài toán tổng quát của CIM trong đó có xét đến chi phí chọn một người dùng vào tập hạt giống và thời gian lan truyền giới hạn. Thêm vào đó, trong việc nghiên cứu BCIM, luận án cũng đề xuất luật TP-PP phản ánh sự cạnh tranh công bằng trong lan truyền ảnh hưởng.

5.1. Phát biểu bài toán

5.1.1. Mô hình ảnh hưởng cạnh tranh

Để giải quyết bài toán BCIM, trước hết luận án đề xuất mô hình Ngưỡng tuyến tính cạnh tranh rằng buộc thời gian TCLT bằng việc mở rộng mô hình CLT trong đó có thêm yếu tố bước thời gian. Ngoài ra, luận án đề xuất một luật tie-breaking mới để phản ảnh đúng thực tế hơn sự cạnh tranh trong tiếp thị sản phẩm trên MXH.

Mô hình này hoạt động cơ bản giống với CLT, thời gian lan truyền được đơn giản hóa thành các bước lan truyền (mỗi bước lan truyền ứng với một đơn vị thời gian). Với bước lan truyền giới hạn $\tau \geq 1$, quá trình lan truyền xảy ra theo các bước rời rạc $t = 0, 1, \ldots, \tau$ như sau:

- Ở bước $t = 0$, $A_0 = S_A, B_0 = S_B$.
- Ở bước $t \geq 1$, gán $A_t = A_{t-1}$, $B_t = B_{t-1}$. Mỗi đỉnh $v \notin A_{t-1} \cup B_{t-1}$ chuyển sang trạng thái A-active nếu thỏa mãn:

$$\sum_{u \in N_-(v) \cap A_{t-1}} w_A(u, v) \geq \theta_A(v) \text{ và } \sum_{u \in N_-(v) \cap B_{t-1}} w_B(u, v) < \theta_B(v) \tag{5.1}$$

  Đỉnh v chuyển sang B-active nếu

$$\sum_{u \in N_-(v) \cap B_{t-1}} w_B(u, v) \geq \theta_B(v) \text{ và } \sum_{u \in N_-(v) \cap A_{t-1}} w_A(u, v) < \theta_A(v) \tag{5.2}$$

- Nếu tại bước t, một đỉnh v có các trọng số thỏa mãn các tổng ảnh hưởng lớn hơn ngưỡng tương ứng, luận án đề xuất luật *tie-breaking với trọng số tỷ lệ* (*weight proportional probability tie-breaking rule* (TB-WPP)) để xác định trạng thái của đỉnh v như sau: v bị kích hoạt bởi A với xác suất

$$p_A(v|A_{t-1}, B_{t-1}) = \frac{\sum_{u \in N_-(v) \cap A_{t-1}} w_A(u, v)}{\sum_{u \in N_-(v) \cap A_{t-1}} w_A(u, v) + \sum_{u \in N_-(v) \cap B_{t-1}} w_B(u, v)} \quad (5.3)$$

  v bị kích hoạt bởi B với xác suất:

$$p_B(v|A_{t-1}, B_{t-1}) = \frac{\sum_{u \in N_-(v) \cap B_{t-1}} w_B(u, v)}{\sum_{u \in N_-(v) \cap A_{t-1}} w_A(u, v) + \sum_{u \in N_-(v) \cap B_{t-1}} w_B(u, v)} \quad (5.4)$$

- Khi một đỉnh bị kích hoạt (A-active hoặc B-active) nó sẽ giữ nguyên trạng thái ở các bước tiếp theo. Quá trình lan truyền dừng lại khi không còn đỉnh nào được kích hoạt thêm.

Luật TB-WPP ta xem xét tổng trọng số của các hàng xóm trong việc đưa ra các xác xuất kích hoạt. Luận án xây dựng mô hình Cạnh tranh cạnh trực tuyến (Competitve live-edge - CLE) tương đương với mô hình TCLT. Những lợi ích của tính chất này là:

- Có thể sử dụng mô hình CLE cho việc ước lượng giá trị hàm mục tiêu
- Nhờ có thể ước lượng được hàm mục tiêu, mô hình CLE làm cơ sở cho các thuật toán đề xuất trong luận án cho bài toán BCIM.

Định lý 5.1. *Với tập hạt giống S_A và S_B cho trước, phân bố của tạp đỉnh A-active và B-active tại mỗi bước $t = 1, 2.., \tau$ trên hai mô hình* TCLT *và* CLE *là như nhau.*

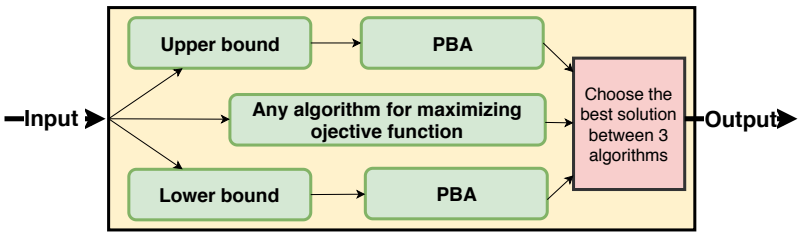
Định nghĩa $\mathbb{I}(S_A)$ là kỳ vọng của tập đỉnh có trạng thái A-active sau τ bước, với S_B là cho trước. Dựa vào Định lý 5.1, ta có:

$$\mathbb{I}(S_A) = \sum_{v \in V \setminus S_B} \sum_{g \in X_G} \Pr[g \sim G] \gamma_g^v(S) \quad (5.5)$$

trong đó $\gamma_g^v(S_A)$ là biến ngẫu nhiên được định nghĩa như sau:

$$\gamma_g^v(S_A) = \begin{cases} 1, \text{ Nếu } v \text{ là } A\text{-active trên mô hình CLE với đồ thị } g \\ 0, \text{Trường hợp ngược lại} \end{cases} \quad (5.6)$$

Bổ đề 5.1 (Ước lượng hàm mục tiêu)**.** *Cho trước tập hạt giống S_B, với tập hạt giống $S_A \subset V \setminus S_B$, ta có $\mathbb{I}(S_A) = n_0 \cdot \mathbb{E}[\gamma(S_A)]$, trong đó $\gamma(S_A)$ là giá trị kỳ vọng của $\gamma_g^v(A)$ trên tất cả các đỉnh nguồn được chọn ngẫu nhiên và đồ thị mẫu được sinh ra ngẫu nhiên từ G*



Hình 5.1: Thành phần của thuật toán SPBA

5.1.1.1. Bài toán BCIM

Định nghĩa 5.1. (Bài toán BCIM)

- **Input:** MXH $G = (V, E)$ trên mô hình TCLT, tập hạt giống $S_B \subseteq V$, ngân sách giới hạn $L > 0$ và thời gian ràng buộc là $\tau > 0$.
- **Output:** Tìm tập hạt giống $S_A \subseteq V \setminus S_B$ với tổng chi phí $\sum_{u \in A} c(u) \leq L$ để cực đại hàm ảnh hưởng $\mathbb{I}(S_A)$.

Định lý 5.2. BCIM *là bài toán NP-Khó và việc tính hàm mục tiêu* $\mathbb{I}(\cdot)$ *là #P-Khó.*

Định lý 5.3. *Hàm mục tiêu* $\mathbb{I}(\cdot)$ *không phải là submodular và supermodular dưới mô hình* TCLT.

5.2. Thuật toán xấp xỉ cho bài toán BCIM

Mô tả khái quát. Thuật toán SPBA chia thành các bước chính sau:

- Tác giả thiết kế các hàm *xấp xỉ dưới* $\mathbb{L}(\cdot)$ và *xấp xỉ trên* $\mathbb{U}(\cdot)$ của hàm mục tiêu. Các hàm này đều có tính chất submodular. Luận án đề xuất thuật toán xấp xỉ dựa trên phương pháp bỏ phiếu (Polling-based Algorithm- PBA) cho bài toán tìm cực đại của các hàm xấp xỉ trên và dưới. Thuật toán PBA có tỷ lệ xấp xỉ là $(1 - 1/\sqrt{e} - \epsilon)$ với xác suất ít nhất là $1 - \delta$, trong đó $\delta, \epsilon \in (0, 1)$ là tham số cho trước.
- Tác giả áp dụng phương pháp SA trong đó các 3 thành phần: lời giải của thuật toán PBA cho bài toán tìm cực đại hàm $\mathbb{L}$ và $\mathbb{U}$ và lời giải của thuật toán bất kỳ cho bài toán BCIM. Thuật toán chính trả về lời giải có kết quả tốt nhất. Cấu trúc của phương pháp SA được mô tả trong Hình 5.1

5.2.1. Thuật toán PBA cho bài toán cực đại các hàm xấp xỉ

PBA sinh ra tập $\mathcal{R}_1$ gồm Λ tập R_j. Thành phần chính của PBA là các vòng lặp (số vòng lặp tối đa là t_{max}) (dòng 3-11).

Trong mỗi vòng lặp, thuật toán tìm tập lời giải ứng viên trên tập $\mathcal{R}_t$ là S_A Bằng việt sử dụng thuật toán Tham lam Greedy (dòng 6). Thuật toán này cho tỷ lệ xấp xỉ là $(1 - \frac{1}{\sqrt{e}})$. Ở bước sau, S_A được kiểm tra chất lượng lời giả qua CheckQS. Thuật toán này sinh ra một tập $\mathcal{R}_c$ bao gồm tập $\mathcal{R}_t$ trước đó và thêm $|\mathcal{R}_t|$ các mẫu R_j sau đó tính toán

Algorithm 4: Thuật toán PBA

Input: Graph $G = (V, E, w_A, w_B)$, budget $L > 0$, and $\epsilon, \delta \in (0, 1)$

Output: A-seed set S_A

1. $\Lambda = \frac{nN_{max}\epsilon^2}{k_{max}}$, $t \leftarrow 1$, $N_{max} \leftarrow N(\epsilon, \frac{\delta}{3})\frac{\mathsf{OPT}_u}{k_{max}}$, $t_{max} = \left\lceil \log_2 \frac{N_{max}}{\Lambda} \right\rceil$
2. Generate Λ URR sets and add them into $\mathcal{R}_1$
3. **repeat**
4. $\quad < S, \mathsf{Cov}_{\mathcal{R}_t}(S) > \leftarrow \mathsf{Greedy}(\mathcal{R}_t, L)$
5. $\quad$ **if** $\mathsf{CheckQS}(\mathcal{R}_t, \mathsf{Cov}_{\mathcal{R}_t}(S_A), \delta, \epsilon) = True$ *or* $|\mathcal{R}_t| \geq N_{max}$ **then**
6. $\quad\quad$ **return** S
7. $\quad$ **else**
8. $\quad\quad$ $t \leftarrow t + 1$, $\mathcal{R}_t \leftarrow \mathsf{CheckQS}(\mathcal{R}_t, \mathsf{Cov}_{\mathcal{R}_t}(S_A), \delta, \epsilon)$
9. $\quad$ **end**
10. **until** $|\mathcal{R}_t| \geq N_{max}$;
11. **return** S_A;

giá trị $\mathsf{Cov}_{\mathcal{R}_c}(S_A) = \sum_{R_j \in \mathcal{R}_c} \min\{|S_A \cap R_j|, 1\}$, $\mathsf{Cov}_{\mathcal{R}_c}(S_A)$ cho biết số tập R_j trong $\mathcal{R}_c$ được phủ bởi S_A. Giá trị này được sử dụng để tính các tham số ϵ_1, ϵ_2 và tính các hàm xấp xỉ trên của lời giải tối ưu và xấp xỉ dưới của giá trị $\mathbb{I}(S_A)$ $f_l(S, \mathcal{R}_c, \epsilon_1)$ và xấp xỉ trên của giá trị lời tối ưu $f_u(\mathsf{OPT}_u, \mathcal{R}_t, \epsilon_2)$. Nếu lời giải hiện tại S_A thỏa mãn điều kiện: $\frac{f_l(S_A, \mathcal{R}_c, \epsilon_1)}{f_u(\mathsf{OPT}_u, \mathcal{R}_t, \epsilon_2)} \geq 1 - \frac{1}{\sqrt{e}} - \epsilon$, thuật toán trả về lời giải là S_A. Nếu không, CheckQS trả về tập $\mathcal{R}_c$ để làm tập mẫu ở bước sau (bước $t + 1$) (dòng 16), sau đó PBA chuyển tiếp sang vòng lặp tiếp theo và dừng lại đến khi số tập $\mathcal{R}_t \geq N_{max}$.

Định lý 5.4. *Với* $0 \leq \epsilon, \delta \leq 1$*, Thuật toán* PBA *trả về lời giải* S_A *thỏa mãn*

$$\Pr[\mathbb{U}(S_A) \geq (1 - 1/\sqrt{e} - \epsilon)\mathbb{U}(S_u^*)] \geq 1 - \delta \tag{5.7}$$

5.2.2. Thuật toán xấp xỉ Sandwich cho BCIM

Luận án đề xuất thuật toán SPBA dựa trên việc áp dụng phương pháp xấp xỉ Sandwich. Chi tiết của thuật toán được mô tả Thuật toán 5.

Algorithm 5: Thuật toán SPBA

Input: Graph $G = (V, E)$, budget $L > 0$, and $\epsilon, \delta, \epsilon', \delta' \in (0, 1)$

Output: seed A

1. $S_U \leftarrow \mathsf{PBA}(\mathbb{L}, G, L, \epsilon, \delta)$
2. $S_L \leftarrow \mathsf{PBA}(\mathbb{U}, G, L, \epsilon, \delta)$
3. $S' \leftarrow$ a solution for maximizing $\mathbb{I}$ by any algorithm.
4. $S \leftarrow \arg\max_{S \in \{S_U, S_L, S'\}} \hat{\mathbb{I}}(S)$
5. **return** S;

Định lý 5.5. *Goi* S_A^* *là lời giải tối ưu của* BCIM, S_{sa} *là lời giải của thuật toán* SPBA, *và*

$$\alpha = \max\left\{\frac{\mathbb{I}(S_U)}{\mathbb{U}(S_U)}, \frac{\mathbb{L}(S_L^*)}{\mathbb{I}(S^*)}\right\}\frac{(1-\epsilon')}{(1+\epsilon')}\left(1-\frac{1}{\sqrt{e}}-\epsilon\right) \tag{5.8}$$

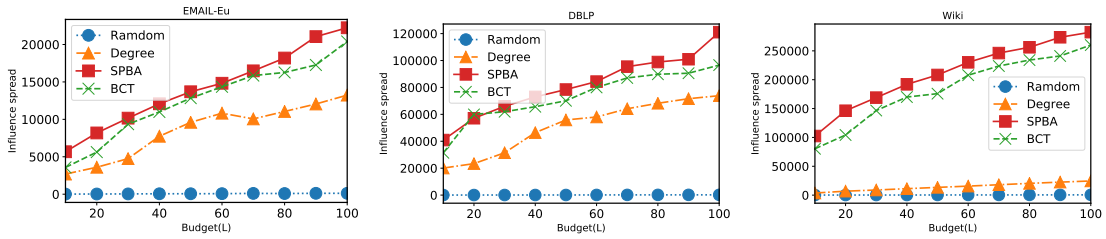
ta có $\Pr[\mathbb{I}(S_{sa}) \geq \alpha \cdot \mathsf{OPT}] \geq 1 - 2\delta$

5.3. Thực nghiệm và kết quả

Luận án so sánh thuật toán SPBA với các thuật toán khác trên nhiều bộ dữ liệu khác nhau để đánh giá hiệu quả của SPBA. Các thuật toán được so sánh bao gồm: BCT: là thuật toán cho bài toán Tối đa ảnh hưởng với chi phí và giới hạn. Lý do luận án sử dụng BCT để so sánh vì BCIM là một biến thể của IM và cũng xét chi phí khác nhau. Các thuật toán cơ sở: Degree và Random.

5.3.1. Kết quả thực nghiệm

Luận án tiến hành đánh giá các thuật toán theo hai trường hợp: chi phí tổng quát (mỗi đỉnh có chi phí khác nhau) và chi phí đồng nhất (các đỉnh có chi phí giống nhau). Thuật toán SPBA luôn cho kết quả tốt nhất. Cụ thể SPBA tốt hơn từ 10% đến 30% so



Hình 5.2: So sánh các thuật toán trong trường hợp chi phí tổng quát

với BCT. SPBA tốt hơn tới 7.7 lần so với Degree. SPBA cho thời gian chạy lâu nhất. Tuy nhiên, SPBA cho thấy thời gian chạy tương đối nhanh với các bộ dữ liệu. Đặc biệt, với mạng Wiki (với 1.79 triệu đỉnh và 28.5 triệu cạnh) SPBA có thể hoàn thành trong 90 giây.

5.4. Bài toán tối đa ảnh hưởng cạnh tranh trên mô hình cạnh tranh ngưỡng tuyến tính xác định

Trong mục này, luận án mở rộng các kết quả nghiên cứu cho bài toán CIM trên mô hình Cạnh tranh ngưỡng tuyến tính xác đinh DCLT. Trên mô hình này, hàm ảnh hưởng cạnh tranh có thể tính toán được trong thời gian $O(n^2)$.

5.4.1. Mô hình và định nghĩa bài toán

Trong mô hình này, mỗi cạnh (u, v) cũng có hai trọng số $w_A(u, v)$ và $w_B(u, v)$ biểu diễn ảnh hưởng của A và B trên cạnh (u, v). Sự khác nhau giữa mô hình CLT với mô hình DCLT là: mỗi đỉnh v có hai ngưỡng kích hoạt $\theta_A(v)$ và $\theta_B(v)$ được cho trước, và (ii) bước lan truyền giới hạn là (steps) d. Quá trình lan truyền diễn ra theo các bước rời rạc $t = 1, \ldots, d$ như sau: Tại bước $t = 0$, $A_0 = A$, $B_0 = B$. Tại bước $t \geq 1$, mỗi đỉnh v sẽ bị

kích hoạt bởi A nếu thõa mãn điều kiện sau:

$$\sum_{u \in N_{in}(v)} w_A(u,v) \geq \theta_A(v) \text{ và } \sum_{u \in N_{in}(v)} w_B(u,v) < \theta_B(v) \tag{5.9}$$

$$\text{hoặc } \sum_{u \in N_{in}(v)} w_A(u,v) \geq \sum_{u \in N_{in}(v)} w_B(u,v) + \alpha_A \tag{5.10}$$

Tương tự, đỉnh v bị kích hoạt bở B nếu:

$$\sum_{u \in N_{in}(v)} w_A(u,v) < \theta_A(v) \text{ và } \sum_{u \in N_{in}(v)} w_B(u,v) \geq \theta_B(v) \tag{5.11}$$

$$\text{hoặc } \sum_{u \in N_{in}(v)} w_B(u,v) \geq \sum_{u \in N_{in}(v)} w_A(u,v) + \alpha_B \tag{5.12}$$

Trong đó α_A, α_B được gọi *hệ số kiềm chế* của A với B và B với A.

Định nghĩa 5.2. (Tối đa ảnh hưởng cạnh tranh -CIM trên DCLT)

- **Input:** Cho một MXH $G = (V, E)$ dưới một mô hình ảnh hưởng cạnh tranh DCLT. Có hai đối thủ cạnh tranh là A và B cùng thực hiện chiến lược lan truyền cạnh tranh dưới mô hình DCLT với bước thời gian giới hạn là d. Cho trước ngân sách k.
- **Output:** Tìm tập hạt giống của A là S_A với $S_A \in V \setminus S_B, |S_A| \leq k$ sao cho số người dùng bị ảnh hưởng bởi S_A là lớn nhất.

Định lý 5.6. CIM *là bài toán NP-Khó và không thể xấp xỉ được trong thời gian đa thức với tỷ lệ* $n^{1-\epsilon}$ *trên mô hình* DCLT

5.4.2. Các thuật toán cho CIM trên mô hình DCLT

Trong phần này, tác giả đề xuất hai thuật toán cho bài toán CIM bao gồm Thuật toán tham lam (Greedy) và Thuật toán tham lam nâng cao (Greedy++). Vì hàm mục tiêu không có tính chất submodular nên hai thuật toán này không cho tỷ lệ xấp xỉ. Tại mỗi bước nó chọn đỉnh u có hàm đo lợi ích $\delta(S_A, u) = \sigma_A(S_A + \{u\}) - \sigma_A(S_A)$. đến khi lựa chọn được k đỉnh. Độ phức tạp của thuật toán này là $O(kn(m+n))$.

Thuật toán Greedy++ là một phiên bản cải tiến của thuật toán tham lam trong đó sử dụng kỹ thuật "lazy evaluation". Kỹ thuật này sẽ không xem xét các đỉnh có lợi ích thấp trong các vòng lặp sau.

5.4.3. Thực nghiệm

Luận án tiến hành thực nghiệm Greedy, Greedy++ với hai thuật toan cơ sở bao gồm thuật toán Random và thuật Degree. Greedy cho kết quả tốt nhất nhưng không áp dụng được với các mạng cỡ vừa và lớn. Thuật toán Greedy++ cho kết quả xấp xỉ với Greedy nhưng có thời gian chạy nhanh hơn từ 17.5 đến 103 lần. Các thuật toán cơ sở cho kết quả không tốt, Greedy++ và Greedy cho kết quả hơn Degree tới 1.65 lần.

CHƯƠNG 6
PHÁT TRIỂN THUẬT TOÁN XẤP XỈ CHO BÀI TOÁN PHÁT HIỆN THÔNG TIN SAI LỆCH

6.1. Đặt vấn đề và phát biểu bài toán

Bài toán phát hiện thông tin sai lệch MD được quan tâm và nghiên cứu gần đây

Định nghĩa 6.1. (Phát hiện thông tin sai lệch -MD)

- **Input:** Cho một MXH $G = (V, E)$ trên mô hình phát tán thông tin $\mathcal{M}$, ngân sách k, (k nguyên dương), mỗi đỉnh $v \in V$ có xác xuất là nguồn TTSL là $\gamma(u)$.
- **Output:** Tìm tập đỉnh $A, |A| = k$ để đặt các giám sát sao cho khả năng phát hiện TTSL là lớn nhất?

6.1.1. Phát biểu bài toán

Luận án nghiên cứu bài toán phát hiện thông tin sai lệch tổng quát GMD, hàm mục tiêu và phát biểu bài toán được nêu chi tiết dưới các mục sau:

6.1.2. Mô hình và hàm mục tiêu

Để xây dựng bài toán GMD với việc xem xét thời gian trễ trong lan truyền thông tin sai lệch, tác giả sử dụng mô hình Independent Cascade Edge Delay (ICED) là một biến thể của mô hình IC, trong đó mỗi cạnh $e = (u, v) \in E$ có xác suất truyền tin là $p(e) \geq 0$ và độ trễ trong lan truyền thông tin là $t(e) \geq 0$.

Quá trình phát tán thông tin sai lệch diễn ra như sau: khi đỉnh u bị kích hoạt bởi thông tin sai lệch sau thời gian $t(u, v)$ nó sẽ có một cơ hội duy nhất để kích hoạt v với xác suất thành công là $p(u, v)$. Thông tin từ đỉnh u tới v có thể được lan truyền thông qua đường đi có *tổng thời gian* là ngắn nhất u đến v trong g gọi là $t_g(u, v)$. Gọi A là tập đỉnh giám sát, thời gian thông tin từ u có thể lan truyền đến A là

$$t_g(u, A) = \min_{v \in A} t_g(u, v) \tag{6.1}$$

Trên đồ thị g, ta định nghĩa biến $D(A, g, u)$ là khả năng phát hiện thông tin sai lệch từ u của A như sau:

$$D(g, A, u) = \begin{cases} 1 & \text{, nếu } t_g(u, A) \leq t \\ 0 & \text{, nếu } t_g(u, A) > t \end{cases} \tag{6.2}$$

Khả năng phát hiện thông tin sai lệch được lượng hóa thông qua *hàm phát hiện* như sau:

$$\mathbb{D}(A) = \sum_{u \in V} \gamma(u) \sum_{g \sim G} \Pr[g \sim G] D(g, A, u) \tag{6.3}$$

Định nghĩa 6.2. (Phát hiện thông tin sai lệch tổng quát-GMD)

- **Input:** Cho MXH $G = (V, E)$ trên mô hình phát tán thông tin ICED. Cho tập $C \subseteq V$ là các đỉnh có thể đặt các giám sát trong phát hiện TTSL, và số nguyên dương k (ngân sách).
- **Output:** Tìm tập các tìm tập $A \subseteq C, |A| = k$ sao cho $\mathbb{D}(A)$ đạt cực đại?

Bài toán GMD là trường hợp tổng quát của bài toán MD. Kế thừa các tính chất của MD, suy ra tính toán hàm mục tiêu $\mathbb{D}()$ là #P-Khó và GMD là bài toán thuộc lớp NP-Khó, không thể xấp xỉ được với tỷ lệ $1 - 1/e + \epsilon, \epsilon > 0$.

6.2. Thuật toán đề xuất cho bài toán GMD

6.2.1. Tính chất và ước lượng hàm mục tiêu

Luận án chỉ ra rằng, việc ước lượng hàm mục tiêu có thể thực hiện qua việc sinh ra các tập *phát hiện ngẫu nhiên* Random Detection (RD) (gọi là tập mẫu) được định nghĩa như sau:

Định nghĩa 6.3 (RD set)**.** Cho đồ thị $G = (V, E)$ dưới mô hình ICED, một tập RD R_j được sinh ra từ G theo các bước như sau. Chọn một *đỉnh nguồn* $u \in V$ với xác suất $\frac{\gamma(u)}{\Gamma}$, $\Gamma = \sum_{v \in V} \gamma(v)$. Sinh ra đồ thị mẫu g từ G, thêm các đỉnh v thõa mãn $t_g(u, v) \leq t$ vào R_j và trả về tập R_j.

Với tập $A \subseteq C$, ta định nghĩa biến ngẫu nhiên $X_j(A)$ như sau:

$$X_j(A) = \begin{cases} 1, \text{ If } R_j \cap A \neq \emptyset \\ 0, \text{ Trong trường hợp ngược lại} \end{cases} \tag{6.4}$$

Bổ đề 6.1. *Với mọi $A \subseteq V$, ta có:* $\mathbb{D}(A) = \Gamma \cdot \mathbb{E}[X_j(A)]$

Bổ đề 6.2. *Hàm $\mathbb{D}(A)$ là đơn điệu tăng và submodular*

Gọi $\hat{\mathbb{D}}(A)$ là ước lượng của $\mathbb{D}$ trên tập $\mathcal{R}$ chưa các tập RD. Dựa trên Bổ đề 6.1, ta có

$$\hat{\mathbb{D}}(A) = \Gamma \cdot \frac{\mathsf{Cov}_{\mathcal{R}}(A)}{|\mathcal{R}|} = \frac{1}{|\mathcal{R}|} \sum_{i=j}^{|\mathcal{R}|} X_j(A) \tag{6.5}$$

Dựa trên kết quả này chúng ta có thể áp dụng các thuật toán dựa theo mô hình RIS bao gồm IMM D-SSA và mới đây nhất OPIM.

6.2.2. Thuật toán SBMD

Thuật toán SBMD bao gồm 2 thành phần chính: 1) Luận án đề xuất việc sinh các tập RD quan trọng trong việc ước lượng hàm phát hiện; 2) Sử dụng lý thuyết Martingle để giảm bớt số mẫu trong việc ước lượng hàm mục tiêu.

Với đỉnh nguồn u, gọi Ω_u là tập các tập RD có đỉnh nguồn là u, ta chia Ω_u thành 2 thành phần như sau:

- Tập phát hiện tầm thường (Trivial Random Detection): gồm duy nhất một đỉnh u, gọi là Ω_u^0
- Ảnh hưởng ngẫu nhiên không tầm thường (Non-trivial Random Detection -NRD): ký hiệu $\Omega_u^n = \Omega_u \setminus \Omega_u^0$

Bổ đề 6.3. *Với mọi tập $A \subseteq V$, ta có:*

$$\mathbb{D}(A) = \Phi \cdot \mathbb{E}[Z_j(A)] + \sum_{v \in A}(1 - \varphi(v))\gamma(v) = \Gamma \cdot \mathbb{E}[Y_j(A)] \tag{6.6}$$

Dựa trên Bổ đề 6.3, ta có một ước lược của $\mathbb{D}(A)$ trên $\mathcal{R}$ là

$$\hat{\mathbb{D}}(A) = \Phi \cdot \frac{\mathsf{Cov}_{\mathcal{R}}(A)}{|\mathcal{R}|} + \sum_{u \in A}\gamma(u)(1 - \varphi(u)) \tag{6.7}$$

Thuật toán SBMD được mô tả chi tiết trong 6.

Algorithm 6: Thuật toán SBMD

Input: Graph $G = (V, E)$, budget $k > 0$, a query (q, t), and $\epsilon, \delta \in (0, 1)$
Output: seed A

$N_{max} \leftarrow N(\epsilon, \frac{\delta}{3}) \cdot \frac{\mathsf{OPT}}{est\mathsf{OPT}}$, $N_1 = N_{max} \cdot k_{max}\epsilon^2/n$, $t \leftarrow 1$,
$t_{max} \leftarrow \left\lceil \log_2 \frac{N_{max}}{\Lambda} \right\rceil$, $\delta_1 \leftarrow \frac{\delta}{3t_{max}}$
Generate N_1 NRD sets and add them into $\mathcal{R}_t$, $\mathcal{R}_c \leftarrow \emptyset$;
repeat
    Add $\mathcal{R}_c$ into $\mathcal{R}_t$, $\mathcal{R}_c \leftarrow \emptyset$, $< S, \mathsf{Cov}(\mathcal{R}_t, A) > \leftarrow \mathsf{Greedy}(\mathcal{R}_t, k)$
    Generate $|\mathcal{R}_t|$ NRD sets and add it into $\mathcal{R}_c$
    Calculate $f_l(A, \mathcal{R}_c, \delta_1)$ and calculate $f_u(\mathsf{OPT}, \mathcal{R}_t, \delta_1)$
    if $\frac{f_l(A,\mathcal{R}_c,\delta_1)}{f_u(\mathsf{OPT},\mathcal{R}_t,\delta_1)} \geq 1 - 1/e - \epsilon$ *or* $|R_t| \geq N_{max}$ **then**
        return A
    end
until $|\mathcal{R}_t| \geq N_{max}$;
return A;

Bổ đề 6.4 (Hàm chặn dưới)**.** *Với mọi $\delta \in (0, 1)$, Tập các tập* NRD *$\mathcal{R}$, và $\hat{\mathbb{D}}(A)$ là ước lượng của $\mathbb{D}(A)$ trên $\mathcal{R}$ được tính bởi* (6.7)*. Đặt $c = \ln(\frac{1}{\delta})$, $a = \beta - \alpha$, ta có*

$$f_l(A, \mathcal{R}, \delta) = \min\left\{\hat{\mathbb{D}}(A) - \frac{ac\Gamma}{3T}, \hat{\mathbb{D}}(A) - \frac{\Gamma}{T}\left(\frac{ac}{3} - cp + \sqrt{\left(\frac{ac}{3} - cp\right)^2 + 2Tpc\frac{\hat{\mathbb{D}}(A)}{\Gamma}}\right)\right\}$$

ta có $\Pr[\mathbb{D}(A) \geq f_l(A, \mathcal{R}, \delta)] \geq 1 - \delta$

Bổ đề 6.5 (Hàm chặn trên). *Với $\delta \in (0,1)$, tập $\mathcal{R}$, A_G là lời giải của thuật toán tham lam với dữ liệu đầu vào là $(\mathcal{R}, k)$, $\hat{\mathbb{D}}(A_G)$ là ước lượng của $\mathbb{D}(A)$ trên $\mathcal{R}$ được tính bởi* (6.3)*, đặt*

$$f_u(\mathsf{OPT}, \mathcal{R}, \delta) = \frac{\hat{\mathbb{D}}(A_G)}{1-1/e} + \frac{\Gamma}{T}\left(-cp + \sqrt{c^2p^2 + 2Tcp\frac{\hat{\mathbb{D}}(A_G)}{(1-1/e)\Gamma}}\right) \tag{6.8}$$

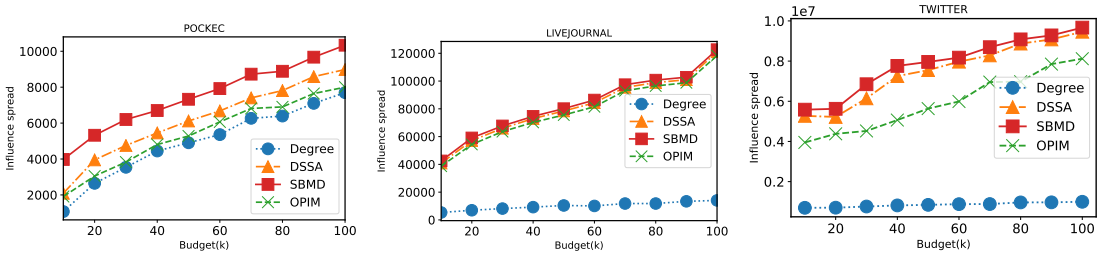
ta có $\Pr[\mathsf{OPT} \leq \hat{\mathbb{D}}(A_G)] \geq 1-\delta$

Định lý 6.1. *Với $\epsilon, \delta \in (0,1)$ là các tham số đầu vào, thuật toán* SBMD *cho lời giải A thỏa mãn* $\Pr[\mathbb{D}(A) \geq (1-1/e-\epsilon)\mathsf{OPT}] \geq 1-\delta$

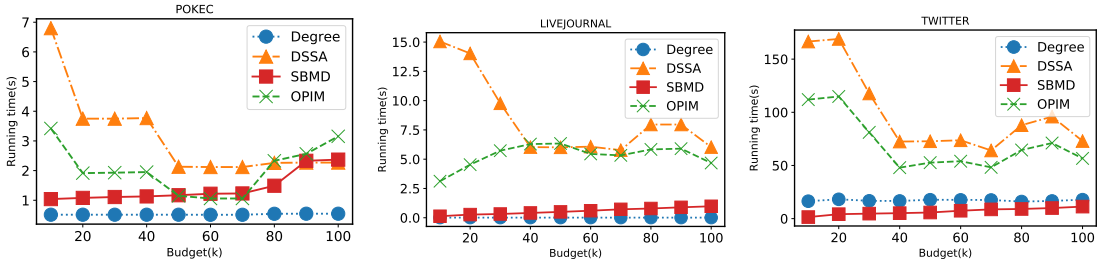
6.3. Thực nghiệm và kết quả

Luận án tiến hành chạy thực nghiệm so sánh kết quả của thuật toán SBMD với các thuật toán khác bao gồm các thuật toán mới nhất hiện nay bao gồm D-SSA và OPIM. Như đã chỉ ra trong phần trước, do tính chất tương đồng giữa IM và GMD, nên các thuật toán này có thể áp dụng cho GMD. Các thuật toán này cùng cho tỷ lệ xấp xỉ là $1-1/e-\epsilon$. Ngoài ra luận án còn so sánh với thuật toán cơ sở là Degree.

Thuật toán SBMD cho kết quả tốt nhất so với các thuật toán còn lại. Với hàm mục tiêu, SBMD cho kết quả tốt hơn hẳn so với các thuật toán khác có cùng tỷ lệ xấp xỉ. Về thời gian, SBMD tỏ ra hơn hẳn các thuật toán còn lại.



Hình 6.1: So sánh hàm mục tiêu đối của các thuật toán



Hình 6.2: So sánh thời gian của các thuật toán

KẾT LUẬN

Luận án nghiên cứu một số bài toán lan về truyền thông tin được quan tâm nghiên cứu trong những năm gần đây, bao gồm: Bài toán ngăn chặn thông tin sai lệch với ràng buộc về ngân sách và thời gian (MMR), Bài toán ngăn chặn thông tin sai lệch với mục tiêu cho trước (TMB), Bài toán Tối đa ảnh hưởng cạnh tranh với ràng buộc về thời gian và ngân sách (BCIM) và Bài toán phát hiện thông tin sai lệch tổng quát (GMD). Các đóng góp của Luận án bao gồm:

1. Nghiên cứu các tính chất, độ phức tạp của bài toán MMR trên mô hình LT, mô hình DTLT. Phát triển các thuật toán hiệu quả cho bài toán MMR bao gồm các thuật toán xấp xỉ, thuật toán heuristic.

2. Nghiên cứu các tính chất, độ phức tạp của toán TMB trên hai mô hình IC và LT. Phát triển các thuật toán hiệu quả cho bài toán TBM trên hai mô hình này.

3. Nghiên cứu bài toán BCIM là bài toán tổng quát của CIM. Đề xuất thuật toán xấp xỉ cho bài toán BCIM trên mô hình TCLT. Mở rộng kết quả nghiên cứu CIM trên mô hình DTLT.

4. Đề xuất thuật toán SBMD có tỷ lệ xấp xỉ $1 - 1/e - \epsilon$ với xác suất ít nhất bằng $1 - \delta$, $\epsilon, \delta \in (0, 1)$ cho bài toán GMD. Các thực nghiệm trên dữ liệu thực chỉ ra hiệu quả nổi trội của thuật toán đề xuất với các thuật toán mới nhất hiện nay.

Trong tương lai, Luận án tiếp tục mở rộng nghên cứu các bài toán trong nhóm các bài toán lan truyền thông tin và tiếp tục phát triển các thuật toán hiệu quả hơn có thể mở rộng cho các mạng hàng tỷ đỉnh để bắt kịp xu hướng mở rộng liên tục của các MXH. Các vấn đề có thể mở rộng nghiên cứu bao gồm

1. Cải tiến các thuật toán đã đề xuất cho các bài toán MMR, TBM cho các mạng cỡ lớn.

2. Nghiên cứu bài toán xác định nguồn phát thông tin ban đầu.

3. Nghiên cứu phát triển thuật toán hiệu quả cho bài toán IM theo cách tiếp cận thuật toán xấp xỉ giảm thiểu số mẫu cần dùng.

4. Nghiên cứu các bài toán biến thể có tính ứng dụng của các bài toán IM, IB và ID.

DANH MỤC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

1. **Canh V. Pham**, Quat V. Phu, Huan X. Hoang, Jun Pei, My T. Thai *Minimum budget for Misinformation Blocking in Online Social Networks*. Journal of Combinatorial Optimization (2019) (**SCI-E**)

2. **Canh V. Pham**, Hieu V. Duong, Huan X. Hoang, and My T. Thai *Competitive Influence Maximization within time and budget constraints in Online Social Networks: An algorithmic approach*. Applied Sciences (2019), 9(11) (**SCI-E**)

3. **Canh V. Pham**, Van Nam Nguyen, Xuan Tuan Le and Xuan Huan Hoang. *Competitive Influence maximization on Online Social Networks: A deterministic modeling approach*. In: Proceeding of IEEE RIVF International Conference on Computing and Comunication Technologies 2019 (RIVF 2019), Danang, Vietnam, March 2019 **(SCOPUS)**.

4. **Canh V. Pham**, Hieu V. Duong, Bui Q. Bao and My T. Thai. *Budgeted Competitive Influence Maximization on Online Social Networks*. In: Proceeding of 7th Conference on Computational Data and Social Networks (CSoNet 2018), pp. 13-24, Shanghai, China, December 2018 (**SCOPUS**)

5. **Canh V. Pham**, My T Thai, Hieu V Duong, Bao Q Bui, Huan X. Hoang *Maximizing misinformation restriction within time and budget constraints*. Journal of Combinatorial Optimization (2018), 35 (4), 1202-1240 (**SCI-E**)

6. **Canh V. Pham**, Quat V. Phu, Huan X. Hoang. *Targeted Misinformation Blocking on Online Social Networks*. In: proceeding of 10 th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2018), pp. 107-116, Quang Binh, Vietnam, March 2018 (**SCOPUS**)

7. **Canh V. Pham**, Hoang M. Dinh, Hoa D. Nguyen, Huyen T. Dang, Huan X. Hoang. *Limiting the Spread of Epidemics within Time Constraint on Online Social Networks*. In: proceeding of the Eighth International Symposium on Information and Communication Technology (SoICT 2017), pp. 262-269, Nha Trang, Vietnam, December 2017 (**SCOPUS**)