BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

VŨ CHÍ QUANG

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP GIẢI BÀI TOÁN CỰC ĐẠI ẢNH HƯỞNG TRÊN MẠNG XÃ HỘI VỚI RÀNG BUỘC ƯU TIÊN VÀ CHI PHÍ

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Mã số: 9 48 01 04

Hà Nội – Năm 2024

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Người hướng dẫn khoa học:

1. Người hướng dẫn khoa học: TS Nguyễn Như Sơn - Viện Công nghệ TT
2. Người hướng dẫn khoa học: PGS. TS Ngô Quốc Dũng - HV Công nghệ bưu chính viễn thông

Phản biện 1: .....……

Phản biện 2:.....……..

Phản biện 3: .....…….

Luận án sẽ được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện, họp tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi ... giờ ...', ngày ... tháng ... năm 2024

Có thể tìm hiểu luận án tại:

- Thư viện Học viện Khoa học và Công nghệ
- Thư viện Quốc gia Việt Nam

MỞ ĐẦU

1. Tính cấp thiết của luận án

- *Về mặt thực tiễn*: Với số lượng người dùng lớn mạng xã hội (Social Network - SN) đã và đang mang lại nhiều lợi ích thiết thực với người dùng. Có thể nói, SN đã và đang trở thành một công cụ hữu ích trong đời sống của con người, đồng thời là một kho tri thức khổng lồ mà mọi người có thể dễ dàng tiếp cận. SN đã mang lại những lợi ích to lớn về chính trị, về kinh tế cho toàn xã hội. Do đó cần nghiên cứu để tối đa hóa thông tin lan truyền trên SN ngày càng hiệu quả hơn.

- *Về mặt khoa học*: Nghiên cứu bài toán Cực đại ảnh hưởng trên SN là một hướng nghiên cứu được nhiều nhà khoa học quan tâm, thuộc nhóm các bài toán lan truyền thông tin (Spread Information - SI), Bên cạnh đó, SN có khối dữ liệu khổng lồ, phân tán và quá trình lan truyền thông tin ngẫu nhiên, cấu trúc mạng phức tạp, không đồng nhất và liên tục biến động do vậy cần phải đưa các giải pháp hiệu quả về mặt thời gian và bộ nhớ.

2. Mục tiêu nghiên cứu của luận án

- Nghiên cứu các bài toán cực đại ảnh hưởng trên các mô hình lan truyền thông tin. Qua đó đề xuất các biến thể mới có tính ứng dụng trong thực tiễn.

- Đề xuất các mô hình giải quyết các bài toán trên, nghiên cứu độ phức tạp của chúng trên các mô hình lan truyền thông tin.

- 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 khả năng ứng dụng với các mạng cỡ lớn hàng trăm nghìn cho tới hàng triệu, hàng tỷ cạnh hoặc đỉnh.

3. Các nội dung nghiên cứu chính của luận án

Chương 1: Cơ sở lý thuyết của luận án và các nghiên cứu liên quan. Trong chương này, luận án giới thiệu về SN, các thành phần cơ bản, một số đặc trưng cũng những lợi ích và mặt trái của SN; Giới thiệu các mô hình và

một số bài toán SI phổ biến trên SN. Những kiến thức tổng quan, mang tính nền tảng cho các nghiên cứu trong các chương sau của luận án.

Chương 2: Cực đại ảnh hưởng với ràng buộc ưu tiên trên mạng xã hội. Chương này, luận án đặt vấn đề và định nghĩa bài toán IMP trên mô hình lan truyền thông tin; đề xuất thuật toán tham lam tích hợp (IG) và thuật toán lấy mẫu dựa trên tham lam tích hợp (IGS) cho bài toán IMP; chứng minh hiệu suất thuật toán đạt xấp xỉ so với phương án tối ưu; phân tích lý thuyết và đánh giá thuật toán dựa trên thực nghiệm với các bộ dữ liệu của SN .

Chương 3: Cực đại ảnh hưởng lan truyền thông tin nhiều chủ đề với chi phí giới hạn. Luận án đề xuất mô hình mới cho bài toán lan truyền thông tin nhiều chủ đề, định nghĩa bài toán BkIM, đề xuất hai thuật toán luồng duyệt dữ liệu một lần cung cấp giới hạn lý thuyết của bài toán. Để xem xét hiệu suất của các thuật toán đề xuất trong thực tế, luận án tiến hành thử nghiệm trên ứng dụng Cực đại ảnh hưởng với k chủ đề trong điều kiện chi phí hạn chế.

CHƯƠNG 1
CƠ SỞ LÝ THUYẾT CỦA LUẬN ÁN VÀ CÁC NGHIÊN CỨU LIÊN QUAN

1.1 Giới thiệu về mạng xã hội

Khái niệm "mạng xã hội" lần đầu được đề cập và sử dụng bởi Barnes từ năm 1954. Từ đó đến nay có hàng trăm nghìn SN được xây dựng với hàng tỷ người dùng trên khắp thế giới. Mỗi mạng đều có cấu trúc và mục đích riêng, nhưng chúng đều có 04 thành phần cơ bản đó là: Người dùng, liên kết giữa các người dùng, thông tin lan truyền trên mạng và tương tác của người dùng với nhau. Ngoài ra SN còn có 04 đặc trưng chung đó là: Đặc trưng thế giới nhỏ, đặc trưng tập nhân, đặc trưng cấu trúc cộng đồng và đặc trưng phân bố lũy thừa.

Với số lượng người dùng lớn SN đã và đang mang lại nhiều lợi ích thiết thực đối với người dùng. Bên cạnh đó, nó cũng cho phép lan truyền nhanh chóng thông tin sai lệch, gây ra những thiệt hại đáng kể đối với đời sống con

người. Để SN ngày càng hữu ích hơn với cộng đồng, chúng ta cần tìm ra những giải pháp hiệu quả để phát huy lợi ích và hạn chế mặt trái của SN.

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

Mô hình hóa các bài toán lan truyền thông tin trên SN đóng vai trò quan trọng trong việc giải quyết các bài toán SI. Giúp các nhà nghiên cứu có cái nhìn tổng quan và ngắn gọn nhất về SN. Để từ đó đưa ra các giải pháp hiệu quả giải quyết các bài toán trên mô hình và từng bước áp dụng vào thực tiễn. Mô hình lan truyền rời rạc được sử dụng rộng rãi trong các nghiên cứu. Điển hình là mô hình *Ngưỡng tuyến tính LT (Linear Threshold)* và *Bậc độc lập IC (Independent Cascade),* đây được xem là những mô hình lan truyền rời rạc được sử dụng trong luận án.

1.2.1 Mô hình Ngưỡng tuyến tính (LT)

Một mạng xã hội được biểu diễn bởi đồ thị $G(V, E)$, mỗi cạnh có trọng số $w(u, v)$ là một số thực dương thỏa mãn điều kiện $\sum_{u \in N_{in}(v)} w(u, v) \leq 1$. $N_{in}(u)$, $N_{out}(u)$ là tập nút vào và tập nút ra của đỉnh u. Mỗi nút có trạng thái *kích hoạt* hoặc *không kích hoạt* và có ngưỡng kích hoạt $\theta_u \in [0,1]$. Gọi S là tập nguồn (tập hạt giống), S_t là tập nút bị *kích hoạt* bởi S tại thời điểm t. Khi $t = 0$, các nút trong tập S_0 đều có trạng thái *kích hoạt*; Khi $t \geq 1$, mỗi nút v sẽ bị *kích hoạt* nếu: $\sum_{u \in N_{in(v)} \cap S_{t-1}} w(u, v) \geq \theta_u$. Quá trình lan truyền kết thúc khi sau mỗi bước không có nút nào được kích hoạt thêm.

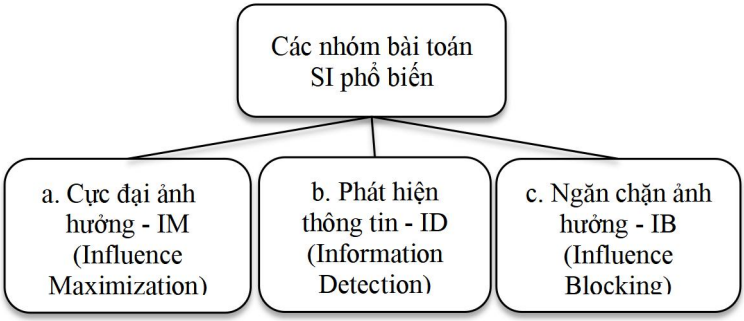
1.2.2 Mô hình Bậc độc lập (IC)

Khác với mô hình LT, trên mô hình IC mỗi cạnh được gán một xác suất ảnh hưởng $p(u, v) \in [0, 1]$. Gọi S_t là tập các nút bị kích hoạt bởi S tại thời điểm t. Khi $t = 0$, các nút trong tập nguồn S_0 đều có trạng thái *kích hoạt*. Tại thời điểm $t \geq 1$, mỗi nút $u \in S_0$ có một cơ hội duy nhất kích hoạt đến nút $v \in N_{out}(u)$ với xác suất thành công là $p(u, v)$. Quá trình lan truyền kết thúc khi giữa hai bước không có nút nào bị kích hoạt thêm.

Gọi $\sigma(S)$ là hàm ảnh hưởng trên mô hình LT, IC giá trị này là kỳ vọng số nút bị kích hoạt khi kết thúc lan truyền. Tính hàm $\sigma(S)$ được D. Kemp chứng minh là #P-khó, để giải quyết vấn đề này họ đề xuất mô hình cạnh trực tuyến LE (Live Edge) và chứng minh nó tương đương với LT và IC.

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

Bài toán lan truyền thông tin được nảy sinh từ nhu cầu của thực tiễn, các nhà phát triển mạng, người dùng mạng và các nhà khoa học luôn muốn tìm ra các giải pháp tối ưu để khai thác những thế mạnh của SN nhằm phục vụ cho các nhu cầu cần thiết của con người và hạn chế những ảnh hưởng tiêu cực không mong muốn. Xét về mục đích nghiên cứu, có thể phân bài toán SI thành 03 nhóm chủ yếu, đó là: Cực đại ảnh hưởng, Phát hiện thông tin và Ngăn chặn ảnh hưởng.



1.3.1 Cực đại ảnh hưởng (Influence Maximization - IM)

Bài toán này xuất phát từ yêu cầu chọn một tập người dùng để bắt đầu SI sao cho số người bị ảnh hưởng bởi thông tin đó trên SN đạt cực đại. IM có ứng dụng trong lan truyền tiếp thị sản phẩm (viral marketing), ngăn chặn thông tin sai lệch MI, phân tích ảnh hưởng trên SN, vv... Mục tiêu của bài toán là chọn một tập hạt giống để bắt đầu quá trình phát tán thông tin sao cho nó ảnh hưởng được nhiều người dùng nhất.

Các thách thức khi giải quyết bài toán này là chúng thuộc lớp NP-Khó và tính toán chính xác hàm mục tiêu thuộc lớp bài toán #P-Khó.

1.3.2 Phát hiện thông tin (Information Detection - ID)

Giả sử rằng đã biết trước một tập người dùng bị nghi ngờ lan truyền thông tin, mục tiêu của bài toán là tìm tập A để đặt giám sát sao cho khả năng phát hiện thông tin từ tập người dùng là lớn nhất. Bài toán này có ứng dụng trong phát hiện thông tin sai lệch (MisInformation - MI) và phát hiện nguồn lan truyền MI, đánh giá xu hướng quan điểm người dùng trên SN.

1.3.3 Ngăn chặn ảnh hưởng (Influence Blocking - IB)

Ngược lại với IM, bài toán ngăn chặn ảnh hưởng (Influence Blocking) 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 SN bao gồm: tin xấu, thông tin sai lệch, hoặc sự phát tán của virus, các tư tưởng cực đoan, v.v..

Các phương pháp có thể hạn chế ảnh hưởng của một nguồn phát tán cho trước được đề xuất bao gồm:

- Vô hiệu hóa người dùng hoặc tập liên kết: loại bỏ tập đỉnh hoặc cạnh để miễn nhiễm với ảnh hưởng.
- 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.

1.4 Bài toán tối ưu tổ hợp và một số phương pháp giải các bài toán tối ưu tổ hợp.

Như đã giới thiệu ở phần trước, nhóm bài toán SI phổ biến như IM, ID, IB thường được xây dựng dưới dạng bài toán Tối ưu tổ hợp (Combination Optimization - CO) thuộc lớp bài toán NP-khó. Hai bài toán được đề xuất trong luận án cũng được cho dưới dạng bài toán CO. Vì vậy để đưa ra phương pháp giải quyết các bài toán này, luận án nghiên cứu một số kiến thức cơ bản về CO. Đây là những kiến thức sử dụng trong các nghiên cứu tiếp theo của luận án.

Định nghĩa: (CO): Mỗi bài toán CO ứng với một bộ ba (S, f, Ω), trong đó S là tập hữu hạn trạng thái (lời giải tiềm năng), 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: max(min): $f(s)$: $s \in S$.

Các bài toán trên SN thường có kích thước lớn, vì vậy các phương pháp giải phổ biến là: Xấp xỉ, Monte Carlo, Heuristic.

- *Phương pháp xấp xỉ:* Phương pháp xấp xỉ là phương pháp đưa ra thuật toán đạt kết quả xấp xỉ một tỷ lệ nào đó so với lời giải tốt nhất. Giả sử ta cần tìm lời giải tối ưu bài toán lan truyền thông tin dưới dạng CO thuộc lớp NP-khó, NP-đầy đủ với mục tiêu tìm hàm cực đại f: $S \to \mathbb{R}_+$, trong đó S là không gian lời giải của bài toán. Gọi OPT(Optimal) là lời giải tối ưu của bài toán. Thuật toán xấp xỉ được định nghĩa như sau:

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

- *Phương pháp Monte Carlo (MC)*: Phương pháp này còn gọi là phương pháp mô phỏng hay còn gọi là phương pháp thử thống kê. Ý tưởng chính của phương pháp Monte Carlo (MC) là xấp xỉ một kỳ vọng μ của X bởi trung bình cộng kết quả của nhiều lần thử nghiệm độc lập, trong đó các biến ngẫu nhiên X có cùng phân phối. Trong nhiều trường hợp, bài toán có hàm mục tiêu phức tạp và không gian tìm kiếm không giới hạn thì không thể áp dụng các phương pháp xấp xỉ, lúc này MC là một phương pháp hiệu quả.

- *Phương pháp Heuristic:* Đây là một phương pháp được thiết kế dựa trên kinh nghiệm để giải một bài toán nhanh hơn khi các phương pháp trước đó quá chậm hoặc để tìm ta một giải pháp gần đúng khi các phương pháp trước không tìm được giải pháp chính xác nào.

- *Thuật toán luồng:* Trong khoa học máy tính, thuật toán luồng là một lớp các thuật toán được thiết kế để xử lý dữ liệu trong môi trường dữ liệu

được tiếp nhận lần lượt. Trong môi trường này, dữ liệu được xử lý dưới dạng chuỗi liên tục, không thể lưu trữ toàn bộ dữ liệu vào bộ nhớ và thường không thể truy cập lại dữ liệu đã xử lý. Thuật toán luồng thường được áp dụng trong các ứng dụng xử lý dữ liệu lớn, trong đó dữ liệu được tạo ra liên tục và cần được xử lý ngay lập tức để đưa ra kết quả trong thời gian thực.

Các tính chất quan trọng của thuật toán luồng bao gồm: *xử lý dữ liệu liên tục, bộ nhớ giới hạn, độ chính xác, cập nhật dữ liệu.*

1.5 Các nghiên cứu liên quan

- *Các nghiên cứu liên quan trong nước:*

Tác giả Phạm Văn Cảnh đã nghiên cứu các bài toán: Ngăn chặn thông tin sai lệch với ràng buộc về ngân sách và thời gian (MMR), Ngăn chặn thông tin sai lệch với mục tiêu cho trước (TMB), Tối đa ảnh hưởng cạnh tranh với ràng buộc về thời gian và ngân sách (BCIM) và Phát hiện thông tin sai lệch tổng quát (GMD).

Tác giả Phạm Văn Dũng đã nghiên cứu các bài toán: Phát hiện nguồn thông tin sai lệch trên mạng xã hội với ngân sách tối thiểu (MBD) và Ngăn chặn thông tin sai lệch nhiều chủ đề trên mạng xã hội có ràng buộc về ngân sách (MBMT).

- *Các nghiên cứu liên quan bài toán cực đại ảnh hưởng:*

Kempe và cộng sự [3] là những người đầu tiên phát biểu bài toán IM trên hai mô hình (IC) và (LT). Chứng minh bài toán IM là NP-Khó và hàm mục tiêu của bài toán IM là #P-Khó.

Chen và cộng sự [97] đã nghiên cứu khái quát về các bài toán IM và BIM.

Borgs và cộng sự [46] đề xuất thuật toán xấp xỉ 1-1/e-ϵ với xác suất là 1-δ, bằng cách giới thiệu mô hình Lấy mẫu ảnh hưởng ngược RR (Reverse Reachable).

Các tác giả trong tài liệu tham khảo [9]-[16] đã nghiên cứu các biến thể bài toán IM theo thời gian, chi phí, khoảng cách và theo các chủ đề.

- *Các nghiên cứu liên quan bài toán cực đại ảnh hưởng lan truyền thông tin nhiều chủ đề.*

Các tác giả trong tài liệu tham khảo [29] nghiên cứu đầu tiên về hàm *k*-Submodular.

Các tác giả trong tài liệu tham khảo [25] -[30], [106] - [110] nghiên cứu về tối ưu hàm *k*-Submodular với các biến thể khác nhau như: không ràng buộc, ràng buộc kích thước, ràng buộc chi phí, ràng buộc matroid, ràng buộc ba lô, ...

Tuy nhiên, các thuật toán của các tác giả chỉ áp dụng được cho trường hợp hàm *f* đơn điệu, trong trường hợp hàm *f* không đơn điệu cho ra được lời giải không như mong đợi.

1.6 Kết luận chương

Chương này luận án giới thiệu những kiến thức chung về SN, mô hình hóa các bài toán SI trên SN, mô hình SI rời rạc và 03 mô hình LT, IC và LE; đây là các mô hình được sử dụng trong các công bố của luận án. Tiếp theo luận án giới thiệu tổng quan về bài toán tối ưu tổ hợp và các phương pháp giải bài toán CO. Những nghiên cứu này là nền tảng quan trọng để luận án đề xuất các bài toán IMP, BkIM trong các chương sau của luận án.

CHƯƠNG 2
CỰC ĐẠI ẢNH HƯỞNG VỚI RÀNG BUỘC ƯU TIÊN TRÊN MẠNG XÃ HỘI

Bài toán cực đại ảnh hưởng (IM) yêu cầu tìm tập hợp *k* nút trong một mạng xã hội để bắt đầu lan truyền ảnh hưởng sao cho số lượng nút ảnh hưởng sau quá trình lan truyền thông tin là tối đa. Tuy nhiên, các nghiên cứu trước đây đã bỏ qua hạn chế về ràng buộc ưu tiên dẫn đến việc thu thập tập hạt giống không hiệu quả.

Để giải quyết vấn đề này luận án đề xuất một bài toán mới có tên là cực đại ảnh hưởng với ràng buộc ưu tiên (IMP), với mục tiêu tìm ra một nhóm gồm *k* nút trong SN để có thể tác động đến số lượng các nút lớn nhất trong khi ảnh hưởng đến một tập người dùng ưu tiên *U* không nhỏ hơn một ngưỡng *T*. NCS chỉ ra rằng bài toán này là NP-Khó và các thuật toán hiện có cho IM không thể áp dụng được với bài toán này. Để tìm ra

giải pháp NCS đề xuất 02 thuật toán hiệu quả, được gọi là Tham lam tích hợp (Integrated Greedy - IG) và thuật toán lấy mẫu dựa trên tham lam tích hợp (Integrated Greedy-based Sampling - IGS) với các đảm bảo tỷ lệ xấp xỉ của lời giải.

2.1 Phát biểu bài toán IMP

Định nghĩa: (Bài toán IMP). Cho đồ thị G = (V, E) theo mô hình IC, một số nguyên dương k (chi phí), tập ưu tiên $U \subset V$ và ngưỡng T với $T \leq k$, $T \leq |U|$. Bài toán IMP yêu cầu tìm tập hạt giống $S \subset V$, với $|S| \leq k$ và $\sigma_U(S) \geq T$ sao cho mức độ lan truyền ảnh hưởng $\sigma(S)$ là cực đại, tức là tìm S là giải pháp cho bài toán tối ưu hóa sau:

maximize: $\sigma(S)$; subject to: $|S| \leq k$; $\sigma_U(S) \geq T$.

IMP trở thành bài toán IM khi U là rỗng. Do đó, IM là một trường hợp đặc biệt của IMP và IMP cũng là NP-Khó. Ngoài ra, việc tính toán hàm ảnh hưởng từ tập hạt giống được chứng minh là # P-Khó.

2.2 Đề xuất thuật toán

Luận án đề xuất hai thuật toán: Thuật toán tham lam tích hợp IG và Thuật toán lấy mẫu dựa trên tham lam tích hợp IGS.

2.2.1 Thuật toán tham lam tích hợp IG

Thuật toán tham lam tích hợp (IG), dựa trên việc thay đổi thuật toán tham lam truyền thống để giải quyết các vấn đề đơn điệu và *submodular* đảm bảo tỷ lệ xấp xỉ cho lời giải.

Thuật toán 2.1: Thuật toán tham lam tích hợp IG

Input: Đồ thị G = (V, E), U ⊂ V, k, T
Output: Tập hạt giống S, và t

1. $S_1 \leftarrow \emptyset$, $S_2 \leftarrow \emptyset$
   /* Đoạn 1: Chiến lược tham lam cho tập ưu tiên */
2. **while** $\sigma_U(S_1) < T$ **do**
3. $u \leftarrow \mathbf{argmax}_{v \in V \setminus S1}(\sigma_U(S_1 \cup \{v\}) - \sigma_U(S_1))$
4. $S_1 \leftarrow S_1 \cup \{u\}$
5. **end**
6. $t \leftarrow k - |S_1|$, $i \leftarrow 0$
   /* Đoạn 2: Tham lam cho IM với ngân sách còn lại*/

7. **while** i < t **do**
8. u ← **argmax**$_{v \in V \setminus S2 \setminus S1}(\sigma(S_2 \cup \{v\}) - \sigma(S_2))$
9. **if** u ∈ S_1 **then**
10. t ← t + 1
11. **end**
12. $S_2 \leftarrow S_2 \cup \{u\}$, i ← i + 1
13. **end**
14. $S \leftarrow S_1 \cup S_2$
15. **return** S, t;

Đánh giá thuật toán 2.1. *Thuật toán IG trả về (S, t), trong đó S là nghiệm khả thi và t ≥ 1, thỏa mãn:*

$$\sigma(S) \geq \left(1 - \left(1 - \frac{1}{k}\right)^t\right)\sigma(S^*)$$

Tỷ lệ xấp xỉ trong trường hợp xấu nhất $1/k$ khi $t = 1$.

2.2.2 Thuật toán lấy mẫu dựa trên tham lam tích hợp IGS

Mặc dù Thuật toán 2.1 có thể cung cấp một đảm bảo gần đúng, nhưng nó không thể hoạt động với mạng xã hội thực vì việc tính hàm ảnh hưởng $\sigma(S)$ là #P-Khó. Để vượt qua thách thức này, luận án đề xuất một thuật toán ngẫu nhiên với đảm bảo xấp xỉ dựa trên việc kết hợp IG với kỹ thuật lấy mẫu.

Ý tưởng của IGS là tạo ra tập hợp các bộ N_u TRR $\mathcal{R}_1$ và đặt hai giải pháp ứng viên S_1, S_2 rỗng. Phần thân của IGS chia thành hai giai đoạn.

Giai đoạn 1, thuật toán tìm ra giải pháp ứng viên S_1 với kích thước nhỏ nhất sao cho $\hat{\sigma}(S) \geq (1 + \alpha)T$ bằng cách sử dụng chiến lược tham lam với hàm tiềm năng $\hat{\sigma}$ trên $\mathcal{R}_1$. Giải pháp ứng viên S_1 thu được trong giai đoạn này thỏa mãn ràng buộc ưu tiên $\sigma_U(S_1) \geq T$ với xác suất ít nhất là 1 - δ.

Giai đoạn 2, chọn một giải pháp ứng viên S_2 với ngân sách còn lại ($t = k - |S_1|$) để mức độ lan truyền ảnh hưởng σ(·) là cực đại. Giai đoạn này, thuật toán thiết lập các tham số ϵ_1, t_{max}, N_{max} và tạo ra N_1 tập hợp mẫu RR $\mathcal{R}_2$. Trong mỗi vòng lặp IGS tìm thấy một giải pháp ứng viên S_2 bằng một chiến lược tham lam. Thuật toán chọn một nút u có ảnh hưởng xấp xỉ tăng dần tối đa $\hat{\sigma}(\cdot)$ trên $\mathcal{R}_2$ cho đến khi t nút được chọn. Sau đó, thuật toán

kiểm tra chất lượng của giải pháp ứng viên S_2. Tiếp theo thuật toán tính toán các hàm $F_l(S_2, \mathcal{R}_2, \delta)$- cận dưới của $\sigma(S_2)$, và $F_u(S_2, \mathcal{R}_2, \delta)$- cận trên của một giải pháp tối ưu đối với bài toán IMP.

Thuật toán 2.2: Thuật toán lấy mẫu dựa trên tham lam tích hợp (IGS)

Input: Đồ thị $G = (V, E)$, $U \subset V$, k, T, ϵ , α, $\delta \in (0, 1)$

Output: Tập hạt giống S

1. Tạo một tập các bộ $N_U = (2+\frac{2}{3}\alpha)|U|\frac{\ln\left(\binom{|U|}{\lfloor |U|/2 \rfloor}/\delta\right)}{(T+T\alpha)\alpha^2}$TRIS sets $\mathcal{R}_1$
2. $S_1 \leftarrow \emptyset$, $S_2 \leftarrow \emptyset$

/ Đoạn 1 */*

3. **While** $\hat{\sigma}_u(S_1) < T + \alpha T$ **do**
4. $\quad u \leftarrow \arg\max_{v \in V \setminus S_1}\left(\hat{\sigma}_u(S_1 \cup \{v\}) - \hat{\sigma}_u(S_1)\right)$
5. $\quad S_1 \leftarrow S_1 \cup \{u\}$
6. **End**

/ Đoạn 2 */*

7. $\epsilon_1 \leftarrow \frac{\epsilon}{2(1-1/e)-\epsilon}$
8. $t_0 \leftarrow k - |S_1|$, $\delta_1 \leftarrow \frac{\delta}{6}$, $t_{max} \leftarrow \arg\max_{j \in \{t, t+1, t+2, \ldots, k\}} \ln(\binom{n}{j}/\delta_1)/j$
9. $N_1 \leftarrow \frac{\ln(1/\delta_1)}{\epsilon_1^2}$, $N_{max} \leftarrow \frac{(2+\frac{2}{3}\epsilon_1)\, n \ln\left(\binom{n}{t_{max}}/\delta_1\right)}{t_0 \epsilon_1^2}$
10. $i_{max} = \left\lceil \frac{N_{max}}{N_1} \right\rceil$, $\delta_2 \leftarrow \frac{\delta}{3 i_{max}}$
11. Tạo ra N_1 tập hợp mẫu RR $\mathcal{R}_2$
12. **Repeat**
13. $\quad t \leftarrow t_0$, $i \leftarrow 0$
14. $\quad$ **While** $i < t$ **do**
15. $\quad\quad u \leftarrow \arg\max_{v \in V \setminus S_2 \setminus S_1}\left(\hat{\sigma}(S_2 \cup \{v\}) - \hat{\sigma}(S_2)\right)$
16. $\quad\quad$ **If** $u \in S_1$ **then**
17. $\quad\quad\quad t \leftarrow t + 1$
18. $\quad\quad$ **End**
19. $\quad\quad S_2 \leftarrow S_2 \cup \{u\}$, $i \leftarrow i + 1$
20. $\quad$ **End**
21. $\quad$ Tính toán $F_l(S_2, \mathcal{R}_2, \delta_2)$ và $F_u(S_2, \mathcal{R}_2, \delta_2)$
22. $\quad$ **If** $\frac{F_l(S_2, \mathcal{R}_2, \delta_2)}{F_u(S_2, \mathcal{R}_2, \delta_2)} \geq 1 - \left(1 - \frac{1}{k}\right)^t - \epsilon$ **then**

23. **Return** S_2
24. **Else**
25. Tạo $|\mathcal{R}_2|$ mẫu RR và thêm chúng vào $\mathcal{R}_2$
26. **End**
27. **Until** $|\mathcal{R}_2| \geq N_{max}$
28. $S \leftarrow S_1 \cup S_2$
29. **Return** S;

Đánh giá thuật toán 2.2. Thuật toán IGS cung cấp nghiệm S và một số nguyên t, thỏa mãn: - $Pr[\sigma_U(S) \geq T] \geq 1 - \delta$.

- $Pr[\ \sigma(S) \geq \left(1 - \left(1 - \frac{1}{k}\right)\right)^t OPT\ ] \geq 1 - \delta$.

2.3 Thực nghiệm và đánh giá kết quả

2.3.1 Thực nghiệm

Dữ liệu thực nghiệm: NCS thực hiện trên 05 bộ dữ liệu của SN. Thuật toán đề xuất được so sánh với các thuật toán DSSA[110], BCT[97], OPIM-C[116] và Degree (thuật toán tham lam cơ sở). Đánh giá kết quả dựa trên các tiêu chí: Giá trị hàm ảnh hưởng, thời gian chạy và sử dụng bộ nhớ.

Bảng 2.1. Thống kê của bộ dữ liệu.

Cơ sở dữ liệu	Số đỉnh	Số cạnh	Loại	Bậc TB
netHEPT	15K	59K	Có hướng	4.1
ENRON	37K	184K	Có hướng	5
netPHY	37K	181K	Có hướng	13.4
DBLP	655K	2M	Có hướng	6.1
Twitter Retweet	1M	2M	Có hướng	4

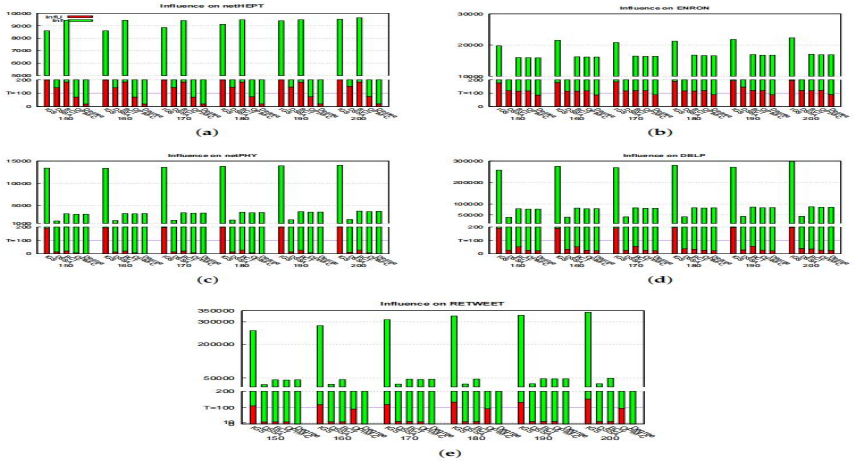
2.4.2 Đánh giá kết quả

- Giá trị hàm ảnh hưởng: Hình 2.1 và Bảng 2.2 cho thấy thuật toán IGS hoạt động tốt hơn các thuật toán khác khi tác động đến các nút ưu tiên theo một ngưỡng T nhất định.

Hình 2.1. Nhìn vào các thanh màu đỏ, ta có thể thấy IGS ảnh hưởng đến tập hợp U xấp xỉ gấp đôi giá trị của ngưỡng T trên hầu hết các cơ sở dữ liệu ngoại trừ Re-Tweet nhưng vẫn cao hơn T.

Bảng 2.2 Nhìn vào các giá trị in đậm, ta có thể thấy mặc dù *U* và *S* đều lớn và *T* tăng dần, ảnh hưởng đến *U* của IGS luôn cao hơn đáng kể so với *T*, thậm chí lên đến hơn chục lần.

Từ Hình 2.1 và Bảng 2.2, ta có thể thấy $\sigma_U(S)$ của IGS cao hơn đáng kể so với *T* và tạo ra kết quả tốt hơn so với các thuật toán hiện đại.



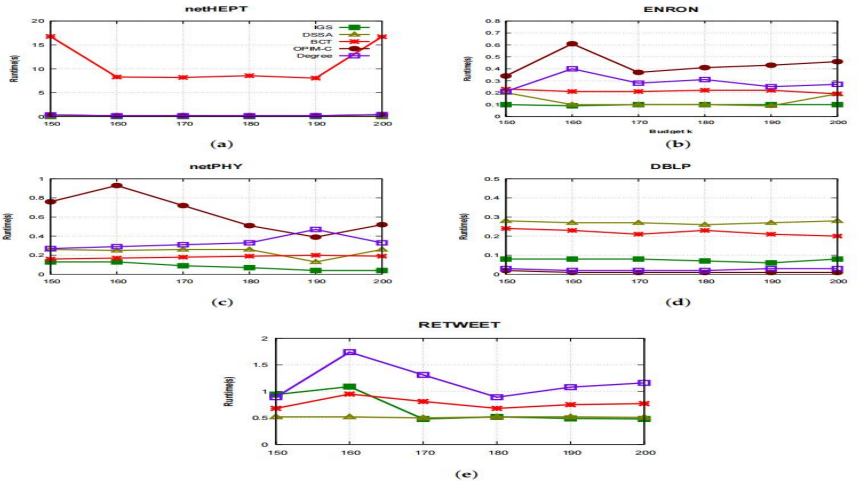
Hình 2.1. So sánh mức độ lan truyền ảnh hưởng với k=100→500, T=100 và U=200

- Thời gian chạy: Hình 2.2 so sánh thời gian chạy của các thuật toán. Thời gian chạy thuật toán IGS cho giá trị thấp nhất trên cơ sở dữ liệu netHEPT, ENRON và netPHY. Tuy nhiên IGS vẫn ở top 3 trên DBLP trong khi tốn thời gian chạy cao nhất trên phần còn lại của tập dữ liệu để tìm tập hạt giống 150 và 160 nhưng quay lại top 3 ở các giá trị khác của ngân sách *k*. IGS chỉ mất khoảng 0,1 giây để tìm ra tập hạt giống trong hầu hết các trường hợp ngoại trừ RETWEET. Nhìn chung, thời gian chạy của IGS cho kết quả ổn định nhất và thường chạy quanh mốc 0,1 giây.

Thời gian của IGS nhanh và ổn định vì lập trình song song và thuật toán này tốn phần lớn thời gian để tìm ra S_1 trong khi vòng lặp để tính S_2 thường dừng lại ở 1–2 vòng. Kỹ thuật lấy mẫu TRR cũng giúp nhanh chóng xác định hạt giống nào sẽ ảnh hưởng đến *U*.

Bảng 2.2. So sánh về σ(*S*) và σ$_U$(*S*) giữa IGS và các thuật toán khác với *k* = 500, *U* size = 1, *K* và *T* = 100 → 500.

			Cơ sở dữ liệu				
	T		**NetHept**	**Enron**	**netPHY**	**DBLP**	**RETWEET**
IGS	100	σ(*S*)	5,666.16	14,267.40	1,865.92	54,033.50	17,307.70
		σ$_U$(*S*)	**1,482.04**	**1,075.77**	**1,192.84**	**1,271.62**	**511.08**
	200	σ(*S*)	5,581.34	14,162.20	1805.26	53,553.90	18,581.50
		σ$_U$(*S*)	**1,478.93**	**1,079.74**	**1,175.32**	**1,267.52**	**491.35**
	300	σ(*S*)	5,645.40	14,284.80	1,773.33	53,240.50	19,459.10
		σ$_U$(*S*)	**1,476.08**	**1,074.30**	**1,153.32**	**1,264.79**	**492.39**
	400	σ(*S*)	5,640.21	14,196.50	1,688.53	52,918.80	18,832.20
		σ$_U$(*S*)	**1,468.48**	**1,075.68**	**1,125.69**	**1,260.31**	**490.46**
	500	σ(*S*)	5,039.45	14,245.50	1,593.66	52,130.90	228,801.00
		σ$_U$(*S*)	**1,238.54**	**1,079.28**	**1,104.20**	**1,252.70**	**994.40**
DSSA		σ(*S*)	4,098.63	9960.35	3230.27	58,197.70	38,253.70
		σ$_U$(*S*)	*1,093.70*	*857.61*	*174.48*	*474.64*	*168.09*
BCT		σ(*S*)	11,088.10	19,901.70	6,675.95	117,197.00	77,316.90
		σ$_U$(*S*)	*1,280.54*	*1,701.60*	*386.49*	*474.64*	*159.77*
OPIM-C		σ(*S*)	3,779.09	19,326.30	6,262.50	112,334	72,026.10
		σ$_U$(*S*)	*600.93*	*894.18*	*194.04*	*459.80*	*173.41*
Degree		σ(*S*)	3824.44	19,349.10	6,345.86	114,249.00	73,936.00
		σ$_U$(*S*)	*292.82*	*779.84*	*164.00*	*260.94*	*22.77*



Hình 2.2. So sánh về thời gian chạy (s) với *k* thay đổi từ 150 đến 200 giữa IGS và các thuật toán khác.

- Sử dụng bộ nhớ: Bảng 2.3 minh họa mức tiêu thụ bộ nhớ của thuật toán IGS và các phương pháp hiện đại. Các số nhỏ nhất được tô đậm trong khi các số lớn nhất được tô màu đỏ. Kết quả cho thấy IGS vượt trội hơn những thuật toán khác, tiêu tốn bộ nhớ ít hơn khoảng bốn lần.

Bảng 2.3. So sánh mức sử dụng bộ nhớ (MB) giữa IGS và các thuật toán khác

Cơ sở dữ liệu	Thuật toán	Ngân sách k					
		150	160	170	180	190	200
NetHEPT	**IGS**	**9.90**	**9.90**	**9.90**	**9.89**	**9.89**	**9.95**
	DSSA	22.84	22.84	22.84	22.84	22.84	22.84
	BCT	1023.79	1017.52	1021.60	1012.21	1020.18	1020.74
	OPIM-C	47.76	47.91	48.03	48.11	48.30	48.46
	Degree	49.14	49.18	49.48	49.68	49.86	50.13
ENRON	**IGS**	**16.82**	**16.79**	**16.81**	**16.81**	**16.82**	**16.82**
	DSSA	30.48	28.07	28.07	28.07	28.07	30.48
	BCT	30.35	30.35	30.39	30.39	30.39	30.39
	OPIM-C	27.16	27.20	42.00	27.22	27.25	27.30
	Degree	27.98	28.08	43.77	28.19	28.27	28.41
NetPHY	**IGS**	**15.18**	**15.18**	**15.18**	**15.18**	**15.18**	**15.04**
	DSSA	52.12	52.12	52.12	52.12	38.50	52.14
	BCT	34.82	34.82	34.82	34.82	34.82	34.80
	OPIM-C	87.88	88.39	88.92	89.31	90.26	90.51
	Degree	92.26	92.71	93.33	93.88	94.68	94.98
DBLP	**IGS**	**138.66**	**138.66**	**138.66**	**138.66**	**138.66**	**138.66**
	DSSA	152.90	152.87	152.87	152.91	152.91	152.83
	BCT	162.88	162.87	162.87	162.88	162.88	162.89
	OPIM-C	475.05	373.72	373.78	373.95	477.18	477.51
	Degree	500.87	395.00	394.26	395.35	504.52	505.26
RETWEET	**IGS**	**214.67**	**214.67**	**214.67**	**214.67**	**214.67**	**214.67**
	DSSA	253.14	253.14	253.14	253.14	253.14	253.14
	BCT	282.50	282.50	282.50	282.47	282.50	282.48
	OPIM-C	877.31	874.20	722.91	876.99	886.78	877.80
	Degree	918.53	916.23	756.93	920.00	930.33	921.95

Kỹ thuật lấy mẫu TRR tập trung vào việc tìm ra các hạt giống ảnh hưởng đến mức độ ưu tiên U trước, sau đó Thuật toán 2.2 khám phá các hạt giống khác để đẩy lên tập hợp hạt giống. Do đó thuật toán 2.2 tiết kiệm bộ nhớ để chạy vòng lặp hơn các thuật toán khác vì không phải kiểm tra xem một nút hạt giống có ảnh hưởng đến U hay không.

Hơn nữa, điều kiện $\frac{F_l(S_2, \mathcal{R}_2, \delta_2)}{F_u(S_2, \mathcal{R}_2, \delta_2)} \geq \left(1 - \left(1 - \frac{1}{\mathrm{k}}\right)\right)^{\mathrm{t}} - \epsilon$ giúp S_2 sinh ra sớm mà không cần chờ điều kiện dừng của vòng lặp.

Cuối cùng, thuật toán đề xuất IGS được thiết kế tốt để có được sự cân bằng giữa mục tiêu ảnh hưởng đến tập hợp ưu tiên nhất định và ảnh hưởng lan truyền đến số lượng nút lớn nhất. Do đó, thời gian chạy, bộ nhớ được sử dụng và ảnh hưởng của IGS cho kết quả cao hơn đáng kể và thậm chí ổn định hơn so với các kết quả của các thuật toán khác.

2.6 Kết luận chương

Trong chương này luận án nghiên cứu bài toán IMP với ràng buộc ưu tiên phát sinh trong một kịch bản thực tế. Mục tiêu của bài toán IMP là chọn một tập nguồn có k nút có thể ảnh hưởng của tập hợp ưu tiên nhất định U lớn hơn ngưỡng T và tổng ảnh hưởng đến các nút đạt cực đại.

Để giải quyết thách thức này luận án mô hình hóa bài toán IMP và đề xuất hai thuật toán IG và IGS với các đảm bảo lý thuyết có thể chứng minh được. Luận án chỉ ra rằng IG cung cấp một nghiệm gần đúng $1 - \left(1 - \frac{1}{k}\right)^t$; IGS là một thuật toán xấp xỉ ngẫu nhiên hiệu quả dựa trên phương pháp lấy mẫu trả về một nghiệm gần đúng $\left(1 - \left(1 - \frac{1}{\mathrm{k}}\right)^{\mathrm{t}} - \epsilon\right)$ với xác suất ít nhất là 1 - δ với $\epsilon > 0$, $\delta \in (0, 1)$ làm tham số đầu vào của bài toán. Các thực nghiệm trên mạng xã hội trong thế giới thực cho thấy các thuật toán đề xuất vượt trội hơn các thuật toán DSSA [110], BCT [97] , OPIM [116] và thuật toán tham lam cở sở về mặt giá trị hàm ảnh hưởng, thời gian chạy và bộ nhớ được sử dụng.

CHƯƠNG 3
CỰC ĐẠI ẢNH HƯỞNG LAN TRUYỀN THÔNG TIN NHIỀU CHỦ ĐỀ VỚI CHI PHÍ GIỚI HẠN

Chương này luận án tiếp tục nghiên cứu về bài toán cực đại ảnh hưởng lan truyền thông tin nhiều chủ đề với chi phí giới hạn.

Để giải quyết vấn đề này, tác giả đề xuất hai thuật toán luồng duyệt dữ liệu một lần với các đảm bảo gần đúng: Một cho trường hợp phần tử *e* chỉ có một giá trị chi phí khi được thêm vào tất cả các bộ thứ *i* và một cho trường hợp tổng quát với các giá trị khác nhau của $c_i(e)$. Kết quả thử nghiệm chỉ ra rằng các thuật toán đề xuất có thể trả lại kết quả cạnh tranh nhưng yêu cầu số lượng truy vấn, độ phức tạp ít hơn đáng kể và thời gian chạy ít hơn các phương pháp hiện tại.

3.1 Đặt vấn đề

Việc tối đa hóa hàm *k-submodular* đã thu hút rất nhiều sự quan tâm vì nó có tiềm năng trong việc giải quyết các vấn đề tối ưu hóa tổ hợp khác nhau, chẳng hạn như cực đại ảnh hưởng, vị trí cảm biến, lựa chọn tính năng và thông tin tối đa hóa phạm vi bảo hiểm. Ngoài trường hợp không bị hạn chế, các nhà nghiên cứu cũng giải quyết vấn đề dưới sự hạn chế về kích thước, ràng buộc matroid và ràng buộc knapsack. Tuy nhiên, những vấn đề này không bao gồm một số ứng dụng thực tế tùy chỉnh từng phần tử theo yêu cầu chi phí của nó cũng như giới hạn chi phí. Luận án sẽ thảo luận về ứng dụng sau:

Cực đại hưởng với k chủ đề trong điều kiện chi phí hạn chế. Trong SN theo một mô hình truyền bá thông tin và *k* chủ đề. Mỗi người dùng có một chi phí để bắt đầu ảnh hưởng của một chủ đề cho thấy mức độ khó tác động ban đầu đến người tương ứng cho chủ đề đó. Với chi phí *B*, NCS xem xét vấn đề tìm một tập hợp người dùng, mỗi người ban đầu

chấp nhận một chủ đề với tổng chi phí là tối đa B để tối đa hóa số lượng người dùng dự kiến được kích hoạt bởi ít nhất một chủ đề.

Trong ứng dụng trên, các hàm mục tiêu là *k-submodular*. Mặc dù đã cố gắng tìm ra một giải pháp cực đại hàm *k-submodular*, các nhà khoa học đã không đề cập đến trường hợp mỗi phần tử sẽ có chi phí khác nhau khi được thêm vào các bộ giải pháp khác nhau với chi phí hạn chế như được trình bày trong ví dụ trên. Được thúc đẩy bởi thực tế đó trong chương này, NCS nghiên cứu một bài toán mới có tên là Cực đại ảnh hưởng lan truyền thông tin nhiều chủ đề với chi phí giới hạn (BkIM), bài toán BkIM được định nghĩa như sau.

***Định nghĩa*:** (Bài toán BkIM) Cho một tập hữu hạn V, một chi phí B và một hàm *k-submodular* f: $(k + 1)^V \mapsto \mathbb{R}_+$ Bài toán yêu cầu tìm lời giải $s = (S_1, S_2, \dots, S_k) \in (k + 1)^V$, trong đó phần tử $e \in V$ có chi phí $c_i(e) > 0$ khi được thêm vào S_i, với tổng chi phí $c(s) = \sum_{i \in [k]} \sum_{e \in S_i} c_i(e) \leq B$ sao cho $f(s)$ đạt cực đại.

3.2 Đề xuất thuật toán

Với sự gia tăng liên tục của dữ liệu đầu vào làm cho dữ liệu đầu vào không thể được lưu trữ trong bộ nhớ máy tính. Do đó điều quan trọng là phải đưa ra các thuật toán luồng (streaming algorithm) cho bài toán BkIM.

Luận án đề xuất hai thuật toán luồng bao gồm: Thuật toán luồng tất định cho trường hợp chi phí cho tất cả các tập con $c_i(e)=c_j(e)$, $\forall e \in V$, $i \neq j$. và Thuật toán luồng ngẫu nhiên cho trường hợp tổng quát.

3.2.1 Thuật toán luồng tất định

Về cơ bản, thuật toán luồng tất định dựa trên ý tưởng tính xấp xỉ giá trị tối ưu opt của bài toán. Sau đó vận dụng giá trị opt để có thể lựa chọn được lời giải tối ưu.

Thuật toán 3.1: Thuật toán luồng tất định

Input : Một hàm f k-$submodular$, $B > 0$, $\epsilon \in (0,1/5)$

Output : một giải pháp **s**

if f đơn điệu **then**
  $\alpha \leftarrow \frac{1}{2}, \epsilon' \leftarrow 4\,\epsilon$
else
  $\alpha \leftarrow \frac{2}{5}, \epsilon' \leftarrow 5\,\epsilon$
end
$(e_{max}, i_{max}) \leftarrow (\emptyset,1)$, $t_j \leftarrow 0\ \forall\, j \in \mathbb{Z}^+$
foreach $e \in V$ **do**
  $i_e \leftarrow \arg max_{i\in[k]} f((e,i))$
  $(e_{max}, i_{max}) \leftarrow \arg max_{(e_1,i_1)\,\in\,\{(e_{max},\, i_{max}),(e,\, i_e\}} f((e_1, i_1))$
  $O \leftarrow \{j \mid f((e_{max}, i_{max})) \le \left(1 + \epsilon'\right)^j \le Bf((e_{max}, i_{max})), j \in \mathbb{Z}^+\}$
  for $j \in O$ **do**
    if $c(\mathbf{s}_j^{t_j}) + c(e) \le B$ **then**
      $i' \leftarrow \arg max_{i\in[k]} f(\mathbf{s}_j^{t_j} \sqcup (e,i))$
      if $\frac{f(\mathbf{s}_j^{t_j} \sqcup (e,i'))}{c(\mathbf{s}_j^{t_j}) + c(e)} \ge \frac{\alpha(1+\epsilon')^j}{B}$ **then**
        $\mathbf{s}_j^{t_j+1} \leftarrow s^{t_j} \sqcup (e, i')$
        $t_j \leftarrow t_j + 1$
      end
    end
  end
end
return $\arg max_{s\in\{(\mathbf{s}_j^{t_j}; j\in O),(e_{max},\, i_{max})\}} f(\mathbf{s})$ nếu f đơn điệu,
  $\arg max_{\mathbf{s}\in\{(\mathbf{s}_j^{t_j}: i\le t, j\in \mathbf{O}),(e_{max},\, i_{max})\}} f(\mathbf{s})$ nếu f không đơn điệu

Đánh giá thuật toán 3.1: Thuật toán luồng tất định là một thuật toán luồng duyệt dữ liệu một lần có độ phức tạp truy vấn $O\left(\frac{kn}{\epsilon}\log n\right)$, độ phức tạp không gian $O\left(\frac{n}{\epsilon}\log n\right)$ và cung cấp tỉ lệ gần đúng là:

- $1/4-\epsilon$ khi f đơn điệu;
- $1/5-\epsilon$ khi f không đơn điệu.

3.2.2 Thuật toán luồng ngẫu nhiên

Trường hợp này, mỗi phần tử có chi phí khác nhau làm cho bài toán trở nên khó khăn hơn và không thể áp dụng các thuật toán trước đây. NCS giới thiệu thuật toán luồng duyệt dữ liệu một lần cung cấp tỷ lệ gần đúng như mong đợi cho bài toán BkIM.

Thuật toán 3.2: Thuật toán luồng ngẫu nhiên

Input : Một hàm f *k-submodular* , $B > 0$, $\epsilon \in (0, 1)$, $\alpha \in (0, 1]$
Output : một giải pháp **s**

$\mathbf{s}^o \leftarrow \mathbf{0}$
$(e_{max}, i_{max}) \leftarrow (\emptyset,1)$, $t_j \leftarrow 0\ \forall j \in \mathbb{Z}_+$
foreach $e \in V$ **do**
  $i_e \leftarrow \arg max_{i\in[k]} f((e,i))$
  $(e_{max}, i_{max}) \leftarrow \arg max_{(e_1,i_1) \in \{(e_{max}, i_{max}),(e, i_e\}} f((e_1, i_1))$
  $O \leftarrow \{j \mid f((e_{max}, i_{max})) \le (1 + \epsilon)^j \le Bf((e_{max}, i_{max})), j \in \mathbb{Z}^+\}$
  foreach $j \in O$ **do**
    $j \leftarrow \emptyset$
    foreach $j \in O$ **do**
      if $c(\mathbf{s}_j^{t_j}) + c_i(e) \le B$ *and* $\frac{f(\mathbf{s}_j^{t_j} \sqcup (e,i)) - f(\mathbf{s}_j^{t_j})}{c_i(e)} \ge \frac{\alpha(1+\epsilon)^j}{B}$ ***then***
        $p_j \leftarrow \frac{f(\mathbf{s}_j^{t_j} \sqcup (e,i)) - f(\mathbf{s}_j^{t_j})}{c_i(e)}$
        $j \leftarrow j \cup \{i\}$
      end
    end
    $T \leftarrow \sum_{i \in J} p_i^{|J|-1}$
    Chọn vị trí $i \in J$ với xác suất $\frac{p_I^{|J|-1}}{T}$
    $\mathbf{s}_j^{t_j+1} \leftarrow \mathbf{s}_j^{t_j} \sqcup (e, i)$
    $t_j \leftarrow t_j + 1$
  end
end
return $\arg max_{s\in\{(\mathbf{s}_j^{t_j}; j\in O),(e_{max}, i_{max})\}} f(\mathbf{s})$ nếu f đơn điệu,
  $\arg max_{\mathbf{s}\in\{(\mathbf{s}_j^i: j\in\mathbf{0},\, i\le t),(e_{max}, i_{max})\}} f(\mathbf{s})$ nếu f không đơn điệu.

Đánh giá thuật toán 3.2: Thuật toán luồng ngâu nhiên là thuật toán luồng duyệt dữ liệu một lần có độ phức tạp truy vấn là $O(\frac{kn}{\epsilon}\log n)$, độ phức tạp không gian là $O(\frac{\mathrm{n}}{\epsilon}\log n)$ và

- Nếu f đơn điệu thì $\mathbb{E}[f(s)] \geq (min\left\{\frac{\alpha}{2}, \frac{(1-\alpha)k}{(1+\beta)k-\beta}\right\} - \epsilon)$opt. Vế phải được cực đại thành $(\frac{v}{3+\beta-\frac{\beta}{k}} - \epsilon)$opt khi $\alpha = \frac{2}{3+\beta-\frac{\beta}{k}}$.

- Nếu f không đơn điệu thì $\mathbb{E}[f(s)] \geq (min\left\{\frac{\alpha}{2}, \frac{(1-\alpha)k}{(1+\beta)k-\beta}\right\} - \epsilon)$opt. Vế phải được cực đại thành $(\frac{v}{3+2\beta-\frac{2\beta}{k}} - \epsilon)$opt khi $\alpha = \frac{2}{3+2\beta-\frac{2\beta}{k}}$.

3.3 Thực nghiệm và đánh giá kết quả

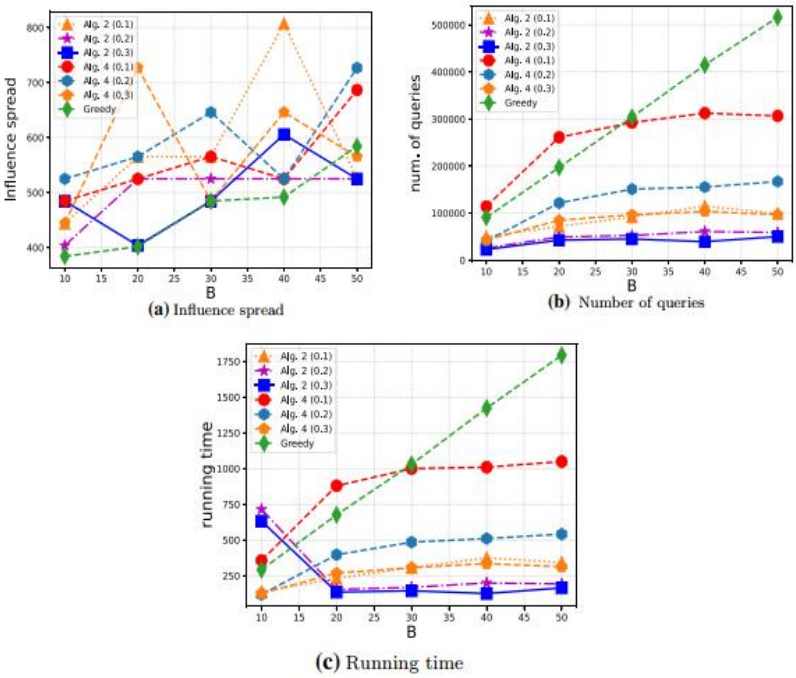
Trong phần này, NCS so sánh hiệu suất của các thuật toán đề xuất với thuật toán tham lam [26] trong ứng dụng của BkIM: Cực đại ảnh hưởng với *k* chủ đề bị hạn chế về chi phí trên ba chỉ số chính: giá trị của hàm mục tiêu, số lượng truy vấn và thời gian chạy.

Bài toán cực đại ảnh hưởng với k chủ đề bị hạn chế về chi phí (Influence Maximization with k topics subject to the budget constraint - IMkB)**.**

Định nghĩa 3.2: (*Bài toán* IMkB) Giả sử rằng mỗi người dùng *u* có một chi phí $c_i(u)$ cho chủ đề thứ *i*, điều này cho thấy mức độ khó ban đầu để gây ảnh hưởng đến người tương ứng đối với chủ đề đó. Với chi phí *B*, bài toán yêu cầu tìm một tập hạt giống *s* với $c(\boldsymbol{s}) \leq B$ sao cho $\sigma(\boldsymbol{s})$ là cực đại.

- *Bộ dữ liệu thực nghiệm*: mạng xã hội Facebook từ SNAP (2020)**.** Mạng chứa 4.039 nút và 88.234 cạnh.

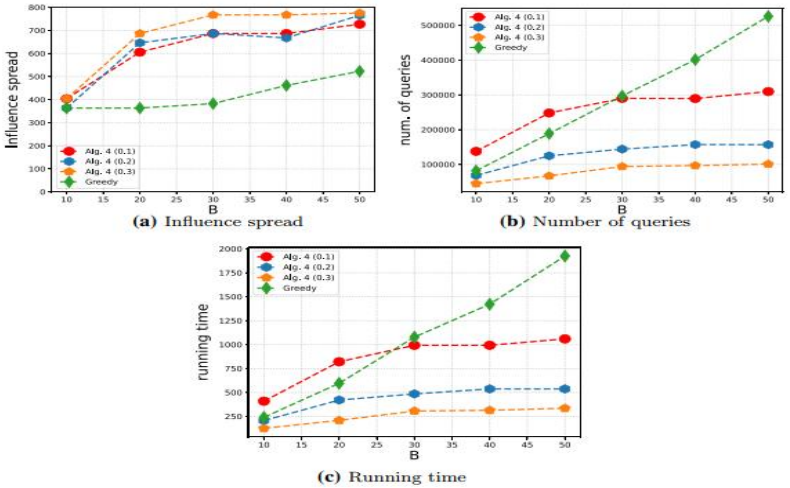
- *Đánh giá kết quả:* NCS chia thực nghiệm thành hai trường hợp: trường hợp đặc biệt khi β=1 và trường hợp tổng quát.



Hình 3.1 Hiệu suất của các thuật toán cho IM*k*B khi $\beta=1$

Hình 3.1 cho thấy hiệu suất của các thuật toán với $\boldsymbol{\beta}$=1. Chất lượng giải pháp, các thuật toán đề xuất vượt trội hơn so với Thuật toán tham lam [26] trong hầu hết các trường hợp. Đối với độ phức tạp truy vấn, các thuật toán đề xuất hoàn toàn vượt trội so với thuật toán tham lam một khoảng cách lớn. Chúng yêu cầu số truy vấn ít hơn tới 10 lần so với thuật toán tham lam cơ sở.

Chúng ta có thể thấy rằng Thuật toán 3.2 cho giải pháp tốt hơn Thuật toán 3.1 với cùng giá trị của cho hầu hết các trường hợp. Điều này phù hợp với phân tích lý thuyết. Tuy nhiên, Thuật toán 3.2 yêu cầu nhiều truy vấn và thời gian chạy hơn Thuật toán 3.1. Vì thế Thuật toán 3.1 đạt đến chi phí *B* nhanh hơn Thuật toán 3.2.



Hình 3.2: Hiệu suất của các thuật toán cho IMkB trong trường hợp chung

3.4 Kết luận chương

Chương này nghiên cứu bài toán BkIM, tổng quát hóa bài toán "Cực đại ảnh hưởng lan truyền thông tin nhiều chủ đề với chi phí giới hạn". NCS đề xuất hai thuật toán luồng duyệt dữ liệu một lần với đảm bảo có thể chứng minh được. Cốt lõi của các thuật toán là khai thác mối quan hệ giữa các giải pháp đề xuất và giải pháp tối ưu bằng cách phân tích các đại lượng trung gian và sử dụng phân phối xác suất mới sau đó so sánh giá trị đóng góp (mục tiêu cận biên trên mỗi chi phí) đến một ngưỡng thích hợp nhất định.

Để xem xét hiệu suất của các thuật toán đề xuất trong thực tế, NCS tiến hành một số thực nghiệm trên ứng dụng "Cực đại ảnh hưởng với k chủ đề bị hạn chế về chi phí". Kết quả thực nghiệm đã chỉ ra rằng các thuật toán đề xuất không chỉ trả về các giải pháp tốt về yêu cầu chất lượng mà còn có số lượng truy vấn nhỏ hơn đáng kể so với thuật toán tham lam tiên tiến nhất.

KẾT LUẬN

Luận án đã hoàn thành mục tiêu nghiên cứu đó là: Nghiên cứu các bài toán cực đại ảnh hưởng với ràng buộc ưu tiên **(IMP)** và Cực đại ảnh hưởng lan truyền thông tin nhiều chủ đề với chi phí giới hạn (**BkIM)**. Luận án đã đề xuất 04 thuật toán hiệu quả để giải quyết các bài toán đặt ra. Các nghiên cứu được công bố trên các kỷ yếu hội thảo, các tạp chí uy tín thuộc chuyên ngành Tối ưu hóa, Công nghệ thông tin trong nước và quốc tế, một số công bố thuộc danh mục SCIE, SCOPUS.

Các bài toán cực đại ảnh hưởng lan truyền thông tin trên SN là các bài toán khó, dữ liệu lớn và có sự ràng buộc phức tạp về không gian, thời gian và chi phí.

Những đóng góp chính của luận án bao gồm:

1. Đề xuất bài toán *"*Cực đại ảnh hưởng với ràng buộc ưu tiên*"* **-IMP**. Mặc dù hàm mục tiêu vẫn là một hàm đơn điệu và hàm *Submodular*, nhưng khi xem xét giới hạn ưu tiên, các thuật toán IM mới nhất không áp dụng được. Để giải quyết thách thức này, luận án đề xuất hai thuật toán IG và IGS với các đảm bảo lý thuyết có thể chứng minh được.

2. Đề xuất bài toán "Cực đại ảnh hưởng lan truyền thông tin nhiều chủ đề với chi phí giới hạn"-**BkIM**. Luận án đề xuất thuật toán luồng tất định và thuật toán luồng ngẫu nhiên duyệt dữ liệu một lần cung cấp giới hạn lý thuyết của bài toán BkIM.

Hướng nghiên cứu của luận án: NCS tiếp tục nghiên cứu và đề xuất các giải pháp hiệu quả hơn cho các bài toán đề xuất. Phát triển các ứng dụng của kết quả nghiên cứu để áp dụng được trên SN. Nghiên cứu các biến thể có tính ứng dụng thực tiễn đối với các bài toán IM, ID và IB. Đề xuất các thuật toán đủ mạnh để xử lý các SN hàng tỷ nút và cạnh trong thời gian chấp nhận được.

DANH MỤC CÁC BÀI BÁO ĐÃ XUẤT BẢN LIÊN QUAN ĐẾN LUẬN ÁN

1. Canh V. Pham, Dung K. T. Ha, **Quang C. Vu**, Anh N. Su and Huan X. Hoang (2020). *Influence Maximization with Priority in Online Social Networks* Algorithms 2020, tập 13, số 183; doi:10.3390/a13080183 (Tác giả liên hệ).
2. **Vũ Chí Quang**, Phạm Văn Cảnh, Hà Thị Kim Dung, Phạm Văn Dũng, Nguyễn Như Sơn. *"Tối đa ảnh hưởng với sự ưu tiên trên mạng xã hội trực tuyến"*, Hội thảo quốc gia lần thứ XXIII **(@)**: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, Quảng Ninh, ngày 5-6/11/2020.
3. Pham, C.V., **Vu, Q. C.**, Ha, D.K.T., Nguyen, T.T. (2021). *Streaming Algorithms for Budgeted k-Submodular Maximization Problem*. In: Mohaisen, D., Jin, R. (eds) Computational Data and Social Networks. CSoNet 2021. Lecture Notes in Computer Science(), vol 13116. Springer, Cham.
4. Pham, C.V., **Vu, Q. C.**, Ha, D.K.T., Nguyen, T.T. et al (2022). *Maximizing k-submodular functions under budget constraint: applications and streaming algorithms*. Journal of Combinatorial Optimization tập 44, trang 723–751.
5. **Vũ Chí Quang**, Phạm Văn Dũng, Nguyễn Thị Tuyết Trinh, Nguyễn Như Sơn.*"Tối ưu hàm k-submodular ứng dụng trong bài toán Cực đại ảnh hưởng của thông tin nhiều chủ đề lan truyền trên mạng xã hội"* Hội thảo quốc gia lần thứ XXVI: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông – Bắc Ninh, 5-6/10/2023.