DOI: https://doi.org/10.62985/j.huit_ojs.vol25.no5.321

GIẢI BÀI TOÁN TỐI ĐA HÓA ẢNH HƯỞNG CỦA LAN TRUYỀN TIẾP THỊ TRÊN CÁC CỘNG ĐỒNG MẠNG XÃ HỘI DỰA TRÊN TỐI ƯU HÓA HÀM DR-SUBMODULAR TRONG LƯỚI NGUYÊN DƯƠNG

Nguyễn Thị Bích Ngân, Nguyễn Trường Phát, Đỗ Thế Sang,

Phạm Nguyễn Huy Phương*

Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: *phuongpnh@huit.edu.vn*

Ngày nhận bài: 09/4/2024; Ngày nhận bài sửa: 21/5/2024; Ngày chấp nhận đăng: 29/5/2024

TÓM TẮT

Trong bối cảnh xã hội phát triển ở rất nhiều lĩnh vực, con người phải đối mặt và giải quyết nhiều bài toán tối ưu hóa với hàm mục tiêu ngày càng phức tạp. Nổi bật trong số đó là họ các bài toán tối ưu hóa có hàm mục tiêu với tính chất lợi nhuận hiệu suất giảm dần, hay còn gọi là hàm DR-submodular (diminishing return submodular). Trong bài báo này, nhóm tác giả nghiên cứu một bài toán cụ thể thuộc họ bài toán trên, đó là tối đa hóa tầm ảnh hưởng cho việc lan truyền tiếp thị trên các cộng đồng của mạng xã hội. Nhóm tác giả áp dụng kỹ thuật duyệt dữ liệu theo luồng phát trực tiếp (streaming) để đề xuất thuật toán DR-SubOptStream cho bài toán và thu được kết quả khả quan cho cả dữ liệu lớn. Trong phần thực nghiệm, nhóm tác giả phải phân tích và tiền xử lý dữ liệu của mạng xã hội từ dạng đồ thị liên thông thông thường thành dạng dữ liệu đồ thị lưỡng cực. Sau đó, thuật toán DR-SubOptStream được chạy với một số bộ dữ liệu mạng xã hội dạng lưỡng cực đã được tiền xử lý. Kết quả thực nghiệm cho thấy thuật toán đề xuất có hàm mục tiêu đạt giá trị chấp nhận theo xấp xỉ và độ phức tạp tốt hơn thuật toán hiện có của dạng bài toán này.

Từ khóa: Hàm DR-submodular, bài toán tối ưu, kỹ thuật luồng phát trực tiếp, dữ liệu đồ thị lưỡng cực.

1. ĐẶT VẤN ĐỀ

Trên các mạng xã hội (MXH), mỗi tài khoản người dùng có thể tham gia nhiều nhóm, mỗi nhóm còn được gọi là cộng đồng. Mỗi cộng đồng chứa số tài khoản có mật độ liên kết "dày đặc" với nhau. Các nhà sản xuất muốn quảng bá sản phẩm thông qua nền tảng MXH, sẽ có nhiều chiến lược khác nhau. Một trong những chiến lược là *chọn và phân bổ "chi phí" tối thiểu cho các cộng đồng sao cho tác động lan truyền ảnh hưởng đến nhiều tài khoản người dùng nhất*, hay còn được gọi là bài toán "*tối đa hóa ảnh hưởng của lan truyền tiếp thị trên các cộng đồng mạng xã hội*". Để giải quyết bài toán này, cốt lõi không chỉ chọn các cộng đồng có nhiều thành viên là đủ, mà còn có nhiều yếu tố khác ràng buộc khác như: chi phí, thời gian chọn lựa cộng đồng, cộng đồng có chứa nhiều tài khoản là đối tượng khách hàng của sản phẩm quảng bá hay không, sự tương tác và ảnh hưởng của người dùng trong cộng đồng,… Nhóm tác giả gọi các ràng buộc trên là "chi phí ảnh hưởng" của cộng đồng mà nhà sản xuất cần bỏ ra để quảng bá sản phẩm tới người dùng, nhưng tổng chi phí không thể vượt một ngưỡng cố định. Nói cách khác, bài toán này cần tìm các cộng đồng mà có phân bổ "chi phí ảnh hưởng" nhỏ nhất sao cho tác động lan truyền tới nhiều người dùng nhất, nghĩa là "lợi nhuận" đạt tối đa. Bài toán này thuộc nhóm bài toán tối ưu hóa có hàm mục tiêu thuộc dạng hàm submodular, và để giải bài toán, nhóm tác giả đề xuất thuật toán dựa trên việc tối ưu hóa hàm DR-submodular trong lưới nguyên dương.

Hàm submodular là hàm số có tính chất lợi nhuận hiệu suất giảm dần, được áp dụng để giải quyết nhiều bài toán tối ưu hóa [1]. Định nghĩa hàm submodular như sau:

Một hàm số $f: 2^V \to \mathbb{R}_+$ *được gọi là hàm submodular nếu và chỉ nếu:*

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \qquad (1.1)$$

$\forall A, B \subseteq V$ và V là tập nền hữu hạn. Ngoài ra, có một định lý tương đương cho hàm submodular, đó là tính chất lợi nhuận hiệu suất giảm dần (diminishing return – DR) của hàm submodular f:

$$f(A \cup v) - f(A) \geq f(B \cup v) - f(B) \quad (1.2)$$

$\forall A \subseteq B \subseteq V$và $v \in V \backslash B$.

Trong thời gian qua, các bài toán tối ưu hóa có hàm mục tiêu thuộc dạng hàm submodular đã thu hút nhiều nhóm nghiên cứu, đặc biệt trong các lĩnh vực thuộc khoa học máy tính, kinh tế [2]. Bởi vì mô hình thực tế của các bài toán đó có tính hiệu suất giảm dần của hàm mục tiêu. Ví dụ một số bài toán phổ biến như là: tóm tắt tài liệu [3,4], thiết lập vị trí các cảm biến [5], phân bổ chi phí, ngân sách [6,7], hay tối đa hóa ảnh hưởng trong lan truyền tiếp thị trên các MXH [8, 9], thiết kế hệ thống mạng [10], và tối ưu hóa tầm ảnh hưởng trong phân tích MXH [11].

Bài toán tối ưu hóa hàm submodular có mục tiêu là chọn tập con S của tập nền V sao cho giá trị $f(S)$ đạt giá trị tối đa [8].

Phần lớn các nghiên cứu của bài toán tối ưu hóa này tập trung vào các hàm submodular trên một tập hợp. Nghĩa là, đầu vào của hàm mục tiêu là một tập con của tập nền và kết quả hàm trả về một giá trị xác định. Nhưng trong thực tế, có nhiều tình huống là không chỉ xác định một phần tử $v \in V$được chọn hay không, mà còn chọn bao nhiêu bản sao của phần tử để hàm mục tiêu đạt tối đa. Nói cách khác, bài toán xem xét các hàm submodular trên một đa tập hợp (multiset), hoặc còn được gọi là hàm submodular trên mạng lưới số nguyên (gọi tắt là hàm submodular trên lưới nguyên) [12].

Một hàm $f: \mathbb{Z}^V \rightarrow \mathbb{R}$ là hàm submodular trên lưới nguyên nếu $\forall x, y \in \mathbb{Z}^V$:

$$f(x) + f(y) \geq f(x \wedge y) + f(x \vee y) \quad (1.3)$$

với $x \wedge y$ hàm ý phép toán tối thiểu và $x \vee y$ hàm ý phép toán tối đa theo tọa độ của x và y.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Quá trình khảo sát các nghiên cứu liên quan cho thấy có nhiều phương pháp để giải quyết bài toán tối ưu hàm submodular nhưng nổi bật trong số đó là hai phương pháp dùng kỹ thuật tham lam (greedy) và luồng phát trực tiếp (streaming) [13].

Nhiều nghiên cứu cho thấy kỹ thuật tham lam thường được áp dụng giải các bài toán tối ưu hóa vì kết quả đầu ra của kỹ thuật này tốt hơn các kỹ thuật khác do tính chất hoạt động "tham lam". Thật vậy, do kỹ thuật tham lam duyệt tất cả dữ liệu, có thể duyệt nhiều lần để chọn phần tử cho kết quả tốt nhất. Tuy nhiên, đây cũng là yếu điểm của kỹ thuật tham lam, nó làm cho thuật toán chạy rất lâu, và do đó thậm chí kỹ thuật tham lam không thể áp dụng khả thi trên dữ liệu lớn. Ngược lại, kỹ thuật luồng phát trực tiếp chỉ duyệt dữ liệu một lần. Mỗi phần tử trong dữ liệu lần lượt được xét đến theo một trình tự nào đó (tùy vào bài toán), kỹ thuật luồng phát trực tiếp phải quyết định phần tử này được chọn hoặc không, trước khi xét đến phần tử tiếp theo. Do đó, kết quả đầu ra của thuật toán luồng phát trực tiếp có thể không tốt bằng kết quả của kỹ thuật tham lam vì các phần tử được chọn không phải là tốt nhất mà chỉ thỏa điều kiện để được chọn. Nhưng điểm mạnh vượt trội của kỹ thuật luồng phát trực tiếp là thời gian thực thi nhanh hơn kỹ thuật tham lam rất nhiều. Vì vậy, kỹ thuật luồng phát trực tiếp thường phù hợp với dữ liệu lớn [14].

Thời gian gần đây có nhiều nghiên cứu được công bố liên quan đến bài toán tối ưu hóa hàm DR-submodular trên lưới nguyên với nhiều ràng buộc khác nhau hoặc được xét trong ngữ cảnh khác nhau. Một số công trình tiêu biểu có thể kể đến như sau:

Năm 2018, Soma và Yoshida phát triển thuật toán tham lam có ngưỡng với kỹ thuật liệt kê một phần các phần tử để giải quyết bài toán tối đa hóa hàm DR-submodular đơn điệu dưới ràng buộc "bài toán ba-lô" trên lưới nguyên [12]. Năm 2020, Gu và cộng sự đề xuất thuật toán dùng kỹ thuật tham lam đôi (double greedy algorithm) để giải quyết bài toán tối đa hóa hàm DR-submodular không đơn điệu [15]. Năm 2021, Liu và cộng sự giải quyết bài toán tối đa hóa hàm DR-submodular dưới ràng buộc "bài toán ba-lô" bằng kỹ thuật luồng phát trực tiếp [16]. Cùng năm 2021, Zhang và các cộng sự đề xuất thuật toán luồng phát trực tuyến để giải bài toán tối đa hóa hàm DR-submodular tăng đơn điệu trên lưới nguyên với ràng buộc số lượng cho tập chọn phần tử [17]. Năm 2022, Gong và cộng sự nghiên cứu bài toán tối đa hóa hàm DR-submodular trên lưới nguyên dưới ràng buộc "bài toán ba-lô", và đã đề xuất thuật toán dùng kỹ thuật tham lam có ngưỡng khi xét duyệt các phần tử [18].

Trong bài báo này, nhóm tác giả tập trung nghiên cứu bài toán *tối đa hóa tầm ảnh hưởng của việc lan truyền tiếp thị trên các cộng đồng của mạng xã hội dựa trên bài toán tối đa hóa hàm DR-submodular trên lưới nguyên dương.* Đây là một biến thể thuộc họ bài toán tối ưu hóa hàm DR-submodular trên lưới nguyên [12]. Qua quá trình khảo sát các nghiên cứu liên quan, Zhang và các cộng sự đã xây dựng thuật toán mới để giải quyết một dạng bài toán cùng họ với bài toán mà nhóm tác giả nghiên cứu trong bài báo này. Đó là *tối đa hóa hàm DR-submodular đơn điệu trên lưới nguyên dưới ràng buộc về số lượng phần tử được chọn*. Thuật toán của Zhang đã sử dụng kỹ thuật luồng phát trực tuyến (Cardinality constraint/DR-submodular, gọi tắt là CaDR-sub). Độ phức tạp của CaDR-sub là $O\left(\frac{k}{\varepsilon}(log\ k)^2\right)$ [17]. Dựa vào ưu điểm của kỹ thuật luồng phát trực tiếp, nhóm tác giả đề xuất một thuật toán mới, gọi là thuật toán DR-SubOptStream. Thuật toán này cải tiến dựa trên kỹ thuật luồng phát trực tuyến Sieve cho bài toán nói trên [21]. Thuật toán mà nhóm tác giả đề xuất có độ phức tạp là $O\left(\frac{n}{\varepsilon} log\left(\frac{1}{\varepsilon} log\ T_{max}\right) log\ k\right)$. Để kiểm chứng hiệu quả của thuật toán, nhóm tác giả tiến hành thực nghiệm với một số dữ liệu của các MXH đã được tiền xử lý, bằng cách chuyển từ dữ liệu dạng đồ thị liên thông thông thường sang dạng dữ liệu lưỡng cực cho phù hợp với việc ứng dụng trong bài toán *tối đa hóa tầm ảnh hưởng của việc lan truyền tiếp thị trên các cộng đồng của MXH.*

Phần còn lại của bài báo được tổ chức gồm những nội dung như sau: phần 3 trình bày lý thuyết và định nghĩa của bài toán; phần 4 giới thiệu thuật toán đề xuất; phần 5 mô tả quá trình xử lý chuyển dữ liệu dạng đồ thị thông thường sang dạng đồ thị lưỡng cực; phần 6 phân tích kết quả thực nghiệm, và cuối cùng phần 7 tổng kết nội dung bài báo.

3. TỐI ĐA HÓA HÀM DR-SUBMODULAR TRÊN LƯỚI NGUYÊN DƯƠNG

3.1 Một số ký hiệu

Bài báo này sử dụng một số ký hiệu liên quan tập hợp và không gian véc-tơ trên tập hợp của mạng lưới số nguyên dương như sau [19]:

(1) Cho tập nền $\boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}\}$, quy ước $\boldsymbol{x(i)}$ là giá trị thành phần $\boldsymbol{i}$ trong véc-tơ $\boldsymbol{x}$, với $\boldsymbol{x} \in \mathbb{Z}_+^{\boldsymbol{V}}$, và $\forall \boldsymbol{v} \in \boldsymbol{V}$, quy ước véc-tơ đơn vị tại vị trí của $\boldsymbol{v}$ là $\boldsymbol{\gamma_v(u)}$, với $\boldsymbol{\gamma_v(u)} = \boldsymbol{1}$ nếu $\boldsymbol{v} = \boldsymbol{u}$, ngược lại $\boldsymbol{\gamma_v(u)} = \boldsymbol{0}$ nếu $\boldsymbol{u} \neq \boldsymbol{v}$.

(2) $[\boldsymbol{k}]$ là tập các số tự nhiên từ 1 đến $\boldsymbol{k}$.

(3) Cho véc-tơ $\boldsymbol{x} \in \mathbb{Z}_+^{\boldsymbol{V}}$, quy ước $\{\boldsymbol{x}\}$là đa tập hợp mà phần tử $\boldsymbol{v} \in \boldsymbol{V}$có giá trị thành phần tại $\boldsymbol{v}$ là $\boldsymbol{x(v)}$ lần.

(4) Cho $\boldsymbol{A} \subseteq \boldsymbol{V}, \boldsymbol{x(A)} = \sum_{\boldsymbol{a} \in \boldsymbol{A}} \boldsymbol{x(a)}$ và $\boldsymbol{supp^+(x)} = \{\boldsymbol{v} \in \boldsymbol{V} | \boldsymbol{x(v)} > \boldsymbol{0}\}$.

(5) Theo khái niệm chuẩn (norm) của véc-tơ, có các ký hiệu như sau:
$\|\boldsymbol{x}\|_\infty = \boldsymbol{max}_{\boldsymbol{v} \in \boldsymbol{V}} \boldsymbol{x(v)}$ và $\|\boldsymbol{x}\|_{\boldsymbol{1}} = \sum_{\boldsymbol{v} \in \boldsymbol{V}} \boldsymbol{x(v)}$.

(6) Cho 2 véc-tơ $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_+^{\boldsymbol{V}}$,

(6.1) $\boldsymbol{x} < \boldsymbol{y}$ có nghĩa là $\forall \boldsymbol{v} \in \boldsymbol{V}, \boldsymbol{x(v)} \leq \boldsymbol{y(v)}$.

(6.2) $(\boldsymbol{x} \wedge \boldsymbol{y})(\boldsymbol{v}) = \boldsymbol{min}\{\boldsymbol{x(v)}, \boldsymbol{y(v)}\}$

(6.3) $(\boldsymbol{x} \vee \boldsymbol{y})(\boldsymbol{v}) = \boldsymbol{max}\{\boldsymbol{x(v)}, \boldsymbol{y(v)}\}$

(6.4) $\boldsymbol{x} + \boldsymbol{y} = \{\boldsymbol{x} + \boldsymbol{y}\}$ là đa tập hợp mà phần tử $\boldsymbol{v} \in \boldsymbol{V}$có giá trị thành phần tại $\boldsymbol{v}$ là $\boldsymbol{x(v)} + \boldsymbol{y(v)}$ lần. Từ đó, suy ra: $\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{x} + (-\boldsymbol{y})$.

(6.5) Cho hàm $\boldsymbol{f}: \mathbb{Z}_+^{\boldsymbol{V}} \to \mathbb{R}_+, \boldsymbol{f(x|y)} = \boldsymbol{f(x+y)} - \boldsymbol{f(y)} = \boldsymbol{f(z)}$
với $\boldsymbol{z(v)} = \boldsymbol{0}$ nếu kết quả sau khi tính $\boldsymbol{f(x|y)}$ của $\boldsymbol{z(v)} < \boldsymbol{0}$.

Ngoài ra, trong Bảng 1, nhóm tác giả giải thích ý nghĩa thêm cho một số ký hiệu được dùng trong bài báo này.

Bảng 1. Ý nghĩa các ký hiệu dùng trong bài báo

Ký hiệu	Mô tả ý nghĩa
$\boldsymbol{V}$	Tập nền, $\boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}\}$
n	Số phần tử của tập nền $\boldsymbol{V}$
$\boldsymbol{2^V}$	Họ các tập con của tập nền $\boldsymbol{V}$
A,B	Các tập con bất kỳ của $\boldsymbol{V}$
x, y	Các véc-tơ bất kỳ thuộc không gian $\mathbb{Z}^V$
γ_v	Véc-tơ đơn vị tại tọa độ $v, v \in \boldsymbol{V}$
$\{x\}$	Đa tập hợp chứa các phần tử $v, v \in V$ trong véc-tơ $x$và mỗi phần tử v có thể được chọn nhiều lần.
$x(v), y(v)$	Giá trị tọa độ của v trong véc-tơ x, y với $v \in \boldsymbol{V}$
$x(\boldsymbol{V})$	Tổng số phần tử (tính số bản sao) của mọi phần tử trong V mà được chọn vào véc-tơ x, nói cách khác $\boldsymbol{x(V)} = \sum_{\boldsymbol{v \in V}} \boldsymbol{x(V)}$
o	Véc-tơ $\mathbf{0}$ với giá trị $o(v) = 0, \forall v \in V$
$\boldsymbol{T}$	Véc-tơ chặn trên của véc-tơ x trong bài toán đang xét ($o \leq x \leq \boldsymbol{T}$)
$\|x\|_\infty$	Chuẩn vô hạn (infinity norm) của véc-tơ x, $\|x\|_\infty = max_{v \in V} x(v)$
$\|x\|_1$	Chuẩn 1 (taxicab norm) của véc-tơ x, $\|x\|_1 = \sum_{v \in V} x(v)$
$\boldsymbol{T_{max}}$	Chuẩn vô hạn của véc-tơ $\boldsymbol{T}$, $\boldsymbol{T_{max}} = \|\boldsymbol{T}\|_\infty$
$\boldsymbol{Opt}$	Giá trị tối ưu nhất (tốt nhất) của hàm mục tiêu $\boldsymbol{f}$
$\boldsymbol{k}$	Chặn trên của tổng số phần tử trong véc-tơ x trên lưới nguyên dương $\mathbb{Z}_+^V$, $x(V) \leq k$

3.2 Định nghĩa bài toán

Định nghĩa 1. Hàm DR-submodular trên lưới nguyên dương

Cho hàm số $\boldsymbol{f}: \mathbb{Z}_+^{\boldsymbol{V}} \to \mathbb{R}_+$ *là hàm đơn điệu nếu* $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{Z}_+^{\boldsymbol{V}}$ *và* $\boldsymbol{x} \leq \boldsymbol{y}$ *thì* $\boldsymbol{f(x)} \leq \boldsymbol{f(y)}$, *và* $\boldsymbol{f}$ *có tính lợi nhuận hiệu suất giảm dần của hàm submodular trên lưới nguyên dương nếu*

$$f(x + \gamma_v) - f(x) \geq f(y + \gamma_v) - f(y) \qquad (1.4)$$

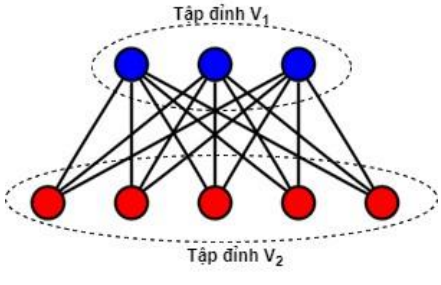
với $v \in V$ *và* γ_v *là véc-tơ đơn vị, có tọa độ của v là 1 và các phần tử khác là 0.*

Định nghĩa 2. Bài toán tối đa hóa hàm DR-submodular trên lưới nguyên dương (gọi tắt là bài toán ĐN2)

Cho f là hàm DR-submodular trên lưới nguyên dương, véc-tơ $\boldsymbol{T} \in \mathbb{Z}_+^{\boldsymbol{V}}$ *là véc-tơ chặn trên,* $\boldsymbol{T_{max}} = \|\boldsymbol{T}\|_\infty$ *và số nguyên* $\boldsymbol{k} \geq \mathbf{0}$, *bài toán ĐN2 cần tìm véc-tơ* $\boldsymbol{x} \in \mathbb{Z}_+^{\boldsymbol{V}}$ *thỏa điều kiện* $\boldsymbol{o} \leq \boldsymbol{x} \leq \boldsymbol{T}$ *và* $\boldsymbol{x(V)} \leq \boldsymbol{k}$ *sao cho* $\boldsymbol{f(x)}$ *đạt giá trị tối đa.*

Nhóm tác giả áp dụng bài toán ĐN2 vào một biến thể thực tế của nó, là *tối đa hóa tầm ảnh hưởng của việc lan truyền tiếp thị trên các cộng đồng của MXH.*

Vì bài toán ĐN2 thực hiện trên mạng lưới nguyên nên dữ liệu thực nghiệm phải có dạng đồ thị lưỡng cực [19]. Hình 1 minh họa một ví dụ về đồ thị lưỡng cực.



Hình 1. Đồ thị lưỡng cực với 2 tập nút V_1 có 3 đỉnh màu xanh và V_2 có 5 đỉnh màu đỏ, tập cạnh là các cạnh nối giữa các đỉnh thuộc V_1 và V_2

Đồ thị lưỡng cực (đồ thị hai phía - bipartite graph): *là đồ thị trong đó các đỉnh có thể được chia thành hai tập hợp rời nhau sao cho tất cả các cạnh đều nối một đỉnh trong tập hợp này với một đỉnh trong tập hợp khác, không có cạnh nào nối giữa các đỉnh trong các tập rời rạc.*[20]

Diễn giải bài toán ĐN2 ở dạng đồ thị cho việc phân tích và thực nghiệm như sau:

Cho đồ thị *G* dạng lưỡng cực thể hiện dữ liệu của một MXH, $G(V;E)$ với *V* là tập các đỉnh được chia thành 2 phần (V_1; V_2), V_1 được định nghĩa là tập các cộng đồng của MXH, V_2 là tập các người dùng trên MXH; $E \subseteq V_1 \times V_2$ là tập các cạnh. Mỗi nút $v_1 \in V_1$ có một giá trị $\tau_{v_1} \in \mathbb{Z}_+$ thể hiện "chi phí tối đa" có thể cấp cho cộng đồng v_1. Mỗi cạnh $v_1v_2 \in E$ được liên kết có kèm trọng số $p(v_1v_2) \in [0;1]$, có nghĩa là khi chọn cộng đồng v_1 sẽ có xác suất $p(v_1v_2)$ lan truyền ảnh hưởng đến người dùng v_2. Mỗi cộng đồng v_1 sẽ được phân bổ một chi phí $x(v_1) \in \{0,1,\dots,\tau_{v_1}\}$ sao cho $\sum_{v_1 \in V_1} x(v_1) \le k$. Hàm mục tiêu *f* của bài toán là tìm x chứa các cộng đồng v_1 sao cho tác động lan truyền đến số người dùng v_2 là tối đa theo công thức (1.5) đã được chứng minh trong nghiên cứu của Soma và cộng sự [19], như sau:

$$f: \mathbb{Z}_+^V \to \mathbb{R}_+ \text{với } f(x) = \sum_{v_2 \in V_2} \left(1 - \prod_{v_1v_2 \in E} (1 - p(v_1v_2))^{x(v_1)}\right) \quad (1.5)$$

4. ĐỀ XUẤT THUẬT TOÁN

Dựa vào ý tưởng của thuật toán luồng phát trực tiếp Sieve trong nghiên cứu của Badanidiyuru và các cộng sự trong nghiên cứu [21], nhóm tác giả của bài báo này đề xuất thuật toán luồng phát trực tiếp cải tiến để giải bài toán ĐN2, được gọi là thuật toán DR-SubOptStream.

Ý tưởng chính của thuật toán DR-SubOptStream là: *với mỗi phần tử v khi được duyệt, tìm tập các phương án x^μ dựa vào giá trị xấp xỉ ε, và tìm tập I chứa các giá trị có khả năng là số lượng bản sao của v chọn vào x^μ dựa vào ε và* **T.** *Sau đó với mỗi phương án x^μ, dựa vào I, tìm số lượng bản sao nhỏ nhất k' của v đưa vào x^μ sao cho giá trị hàm mục tiêu của v thỏa điều kiện xấp xỉ ε và chi phí k. Kết quả là $f(x^\mu)$ với $f(x^\mu) = argmax_{x^\mu, \mu \in O} f(x^\mu)$.*

Trong thuật toán này, nhóm tác giả cải tiến so với phương pháp Sieve đó là tìm trước tập I dựa theo ngưỡng **T và** ε để thu nhỏ phạm vi tìm giá trị số lượng bản sao được chọn của *v vào* x^μ. Cải tiến này giúp tiết kiệm thời gian tìm kiếm nhưng vẫn đảm bảo điều kiện xấp xỉ ε và chi phí *k.*

a. Thuật toán DR-SubOptStream

- **Đầu vào**: hàm *f,* **T**, *k* và ε, với ε là xấp xỉ tối đa của kết quả so với Opt.
- **Đầu ra**: một véc-tơ kết quả x có xấp xỉ theo ε.

```
1.  O := ∅, với O là tập hợp O = {(1+ε)^μ | μ ∈ ℤ+}
2.  x^μ := ∅, ∀μ ∈ O; mf := 0
3.  for v ∈ V do {
4.          mf := max(mf, f(γ_v))
5.          O := {(1+ε)^μ | μ ∈ ℤ+, mf ≤ (1+ε)^μ ≤ 2k.mf}
6.          J := {⌈T(v)(1-ε)^i⌉ | i ∈ ℤ+ sao cho 1 ≤ T(v)(1-ε)^i ≤ T(v)}
7.          I := {i_1, i_2, ..., i_|J|} với i_1 < i_2 ... < i_|J|
8.          for μ ∈ O do {
9.                  Tìm k_v = min(i ∈ I: f(i * γ_v | x^μ) < iμ/2k)
10.                 l := min(k_v, k − ||x^μ||_1)
11.                 if l ≠ 0 then
12.                         x^μ += l.γ_v
13.                 else break;
14.         }
15. }
16. return argmax_{x^μ, μ∈O}   f(x^μ)
```

b. Phân tích thuật toán

Dựa theo chứng minh về độ phức tạp của thuật toán luồng phát trực tiếp trong nghiên cứu [21] của Badanidiyuru và cộng sự, mỗi phần tử trong tập V chỉ cần duyệt một lần, nên lệnh **for** ở dòng 3 thực hiện n lần lặp.

Gọi t là số phần tử của O, $t = |O|$, ta có:

$$t \leq \frac{1}{\varepsilon} log\ k \tag{1.6}$$

Để tìm $k_v$trong mỗi phương án μ của ta có:

$$log\ |I| = O\left(log\left(\frac{1}{\varepsilon} log\ T_{max}\right)\right), \text{ với } T_{max} = \|T\|_\infty = max_{v \in V}\ T(v) \tag{1.7}$$

Như vậy độ phức tạp của thuật toán này là $O\left(\frac{n}{\varepsilon} log\left(\frac{1}{\varepsilon} log\ T_{max}\right) log\ k\right)$.

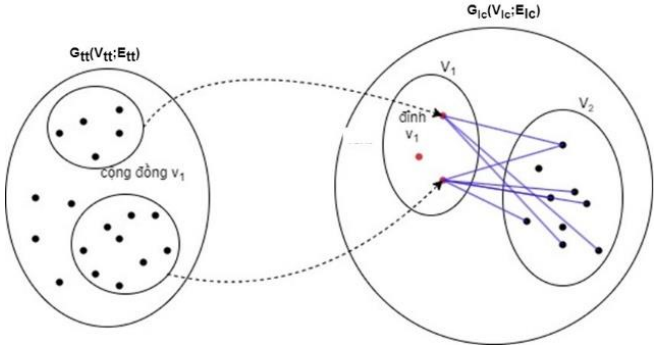
5. XỬ LÝ DỮ LIỆU

Theo diễn giải bài toán ở trên, dữ liệu để thực nghiệm cho bài toán này phải là dữ liệu dạng lưỡng cực của MXH. Tuy nhiên qua quá trình khảo sát, chưa có bộ dữ liệu lưỡng cực chuẩn nào của MXH, mà hầu hết các dữ liệu chuẩn của MXH đều ở dạng đồ thị thông thường (có tập đỉnh và nút liên thông nhau) [22]. Vì vậy, nghiên cứu này đã thực hiện một bước tiền xử lý dữ liệu trước khi thực nghiệm cho bài toán này. Đó là xây dựng dữ liệu đồ thị lưỡng cực cho MXH từ bộ dữ liệu đồ thị thông thường của nó. Các bộ dữ liệu của MXH được nhóm tác giả lấy từ SNAP (Stanford Network Analysis Project – thư viện khai thác đồ thị và phân tích mạng của Jure Leskovec và cộng sự, thuộc trường đại học Stanford, Hoa Kỳ [22]).

Quá trình thực hiện chuyển đồ thị thông thường $\boldsymbol{G_{tt}(V_{tt}; E_{tt})}$ thành đồ thị lưỡng cực $\boldsymbol{G_{lc}(V_{lc}; E_{lc})}$ (với $\boldsymbol{V_{lc} = (V_1; V_2)}$) như sau:

Dùng thuật toán phát hiện cộng đồng (detecting communities) để tìm các cộng đồng trong đồ thị của MXH.

- *Mỗi cộng đồng tìm được trong* $\boldsymbol{G_{tt}}$ *chuyển thành một đỉnh* $\boldsymbol{v_1}$ *trong tập* $\boldsymbol{V_1}$.
- *Các đỉnh* $\boldsymbol{v \in V_{tt}}$ *chuyển thành* $\boldsymbol{v_2 \in V_2}$
- *Cạnh* $\boldsymbol{v_1 v_2 \in E_{lc}}$ *nếu* $\boldsymbol{v_2}$ *có thuộc về cộng đồng* $\boldsymbol{v_1}$.
- *Trọng số* $\boldsymbol{p(v_1 v_2)}$ *của cạnh* $\boldsymbol{v_1 v_2}$ *được tính theo công thức* $\boldsymbol{p(v_1 v_2) = \frac{degree(v_2\ in\ v_1)}{max\_degree\ in\ v_1}}$ *với* $\boldsymbol{degree(v_2 in\ v_1)}$*: số bậc của* $\boldsymbol{v_2}$ *trong* $\boldsymbol{v_1}$*;* $\boldsymbol{max\_degree\ in\ v_1}$*: số bậc lớn nhất trong* $\boldsymbol{v_1}$. Công thức tính $\boldsymbol{p(v_1 v_2)}$được dựa theo nghiên cứu của Nicolas Dugu´e và Anthony Perez [23]. Hình 2 minh họa quá trình chuyển đồ thị thông thường G_{tt} thành đồ thị lưỡng cực G_{lc}.



Hình 2. Minh họa chuyển đồ thị thông thường G_{tt} thành đồ thị lưỡng cực G_{lc}.

Trong quá trình xử lý dữ liệu này, nhóm tác giả đã nghiên cứu và áp dụng hai thuật toán phát hiện cộng đồng, đó là Clauset-Newman-Moore greedy (gọi tắt là Greedy) [22] và Directed Louvain [23], để so sánh hiệu quả của chúng. Chi tiết của kết quả này được trình bày ở phần thực nghiệm.

6. KẾT QUẢ THỰC NGHIỆM

6.1 Thực nghiệm chuyển dữ liệu từ dạng đồ thị thông thường sang đồ thị lưỡng cực

Bài báo này chọn ba bộ dữ liệu MXH có kích thước số nút và số cạnh khác nhau từ hệ thống SNAP. Các thông tin của dữ liệu trình bày trong Bảng 2, gồm:

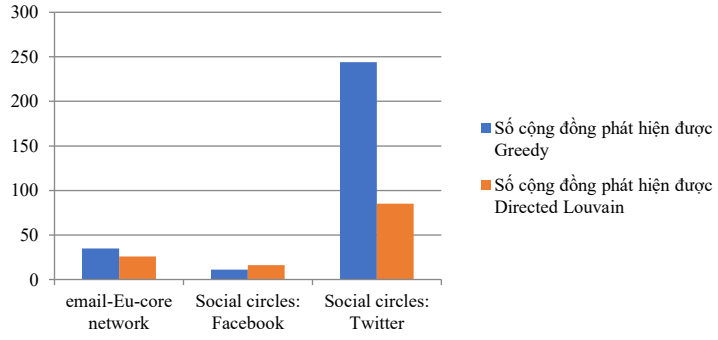
Email-Eu-core network: Mạng được tạo bằng dữ liệu email từ một tổ chức nghiên cứu lớn ở châu Âu, Có một cạnh *(u, v)* trong mạng nếu người *u* gửi cho người *v* ít nhất một email. Các e-mail chỉ thể hiện sự liên lạc giữa các thành viên của tổ chức và tập dữ liệu không chứa các tin nhắn đến hoặc đi đến phần các người không thuộc mạng [25].

Social circles- Facebook: Tập dữ liệu này bao gồm 'danh sách bạn bè' từ Facebook. Dữ liệu Facebook được thu thập từ những người tham gia khảo sát bằng ứng dụng Facebook. Tập dữ liệu bao gồm các tính năng nút (hồ sơ), vòng kết nối trong danh sách bạn bè [26].

Social circles-Twitter: Tập dữ liệu này bao gồm 'danh sách bạn bè' từ Twitter. Dữ liệu Twitter được thu thập từ các nguồn công cộng. Tập dữ liệu bao gồm các tính năng nút (hồ sơ), vòng kết nối [26].

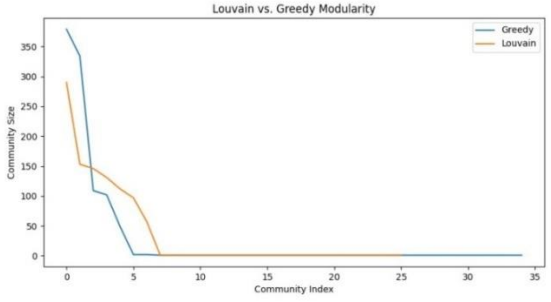
Bảng 2. Mô tả dữ liệu thực nghiệm và kết quả sau khi chuyển thành đồ thị lưỡng cực

Tên dữ liệu	Số nút	Số cạnh	Số cộng đồng phát hiện được	
			Greedy	Directed Louvain
Email-Eu-core network	1.005	25.571	35	26
Social circles-Facebook	4.039	88.234	11	16
Social circles-Twitter	81.306	1.768.149	244	85



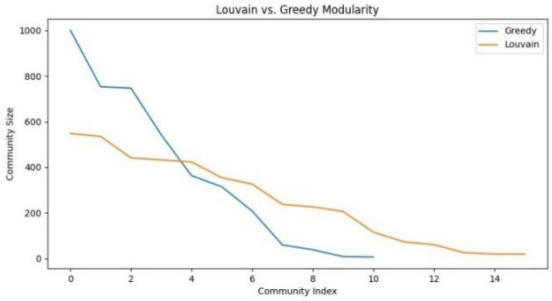
Hình 3. Kết quả số cộng đồng được phát hiện của 3 bộ dữ liệu tương ứng với 2 kỹ thuật Greedy và Directed Louvain

Sau quá trình thực hiện tiền xử lý dữ liệu, chuyển từ đồ thị thông thường sang đồ thị lưỡng cực bằng 2 kỹ thuật Greedy và Directed Louvain, kết quả thu được như ở Bảng 2 và các Hình 3, 4, 5 và 6.



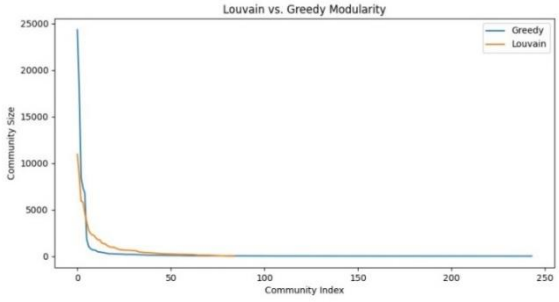
Hình 4. Kết quả số cộng đồng và mật độ kích thước cộng đồng phát hiện được của bộ dữ liệu email-Eu-core network

Thông qua thực nghiệm chuyển đổi dạng đồ thị (ở Bảng 2 và Hình 3, 4, 5, 6), kết quả có thể thấy: tùy theo đặc trưng của mỗi dữ liệu mà số cộng đồng được phát hiện của mỗi thuật toán sẽ khác nhau. Nhìn chung, Greedy thường phát hiện số cộng đồng nhiều hơn và mật độ "dày đặc" của các cộng đồng cao hơn thuật toán Directed Louvain. Điều này là hiển nhiên vì Greedy có tính "tham lam" sẽ phát hiện cộng đồng, số thành viên cùng cộng đồng sao cho nhiều nhất có thể, trong khi Directed Louvain chỉ cần phát hiện cộng đồng có độ "gắn kết" đạt ngưỡng yêu cầu là thuật toán dừng, để chuyển tìm cộng đồng tiếp theo. Đánh đổi cho kết quả đó, Greedy thường có thời gian thực thi lâu hơn Directed Louvain. Vì vậy tùy vào yêu cầu của dữ liệu đầu ra mà người dùng lựa chọn thuật toán cho phù hợp.



Hình 5. Kết quả số cộng đồng và mật độ kích thước cộng đồng phát hiện được của bộ dữ liệu Social circles-Facebook

Trong nghiên cứu này, nhóm tác giả chạy cả hai thuật toán và chọn kết quả nào phát hiện được nhiều cộng đồng thì chuyển thành dữ liệu lưỡng cực để thực nghiệm cho bài toán ĐN2.



Hình 6. Kết quả số cộng đồng và mật độ kích thước cộng đồng phát hiện được của bộ dữ liệu Social circles-Twitter

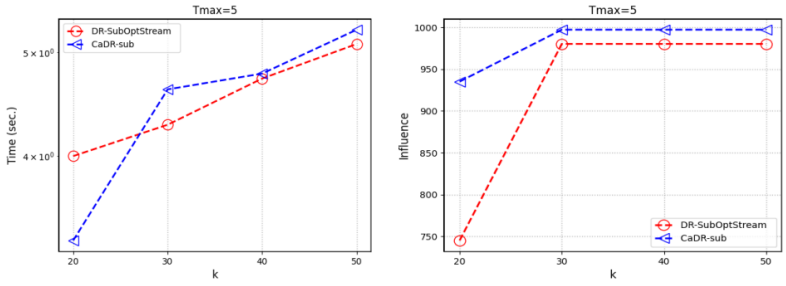
6.2 Thiết lập tham số cho quá trình thực nghiệm bài toán ĐN2

Nhằm kiểm chứng hiệu quả của thuật toán đề xuất, nhóm tác giả tiến hành thực nghiệm để so sánh kết quả thuật toán DR-SubOptStream và CaDR-sub (Zhang và cộng sự [17]). Để thực nghiệm được hiệu quả, nhóm tác giả chọn kết quả xử lý tiền dữ liệu nào có số cộng đồng phát hiện được nhiều hơn và mật độ liên kết trong cộng đồng vẫn đảm bảo ngưỡng cho phép. Các tham số trong thực nghiệm được thiết lập như sau: $\boldsymbol{\varepsilon = 0.1, T_{max} = 5}$ cho cả 3 bộ dữ liệu; dựa vào kích thước số nút và số cạnh của mỗi bộ mà chọn $\boldsymbol{k} \in \{20, \mathbf{30}, \mathbf{40}, \mathbf{50}\}$ cho bộ email-Eu-core network, $\boldsymbol{k} \in \{\mathbf{80}, \mathbf{100}, \mathbf{120}\}$ cho bộ Social circles-Facebook, và $\boldsymbol{k} \in \{180, 190, \mathbf{200}, \mathbf{250}, \mathbf{300}\}$ cho bộ Social circles Twitter.

6.3 Kết quả thực nghiệm

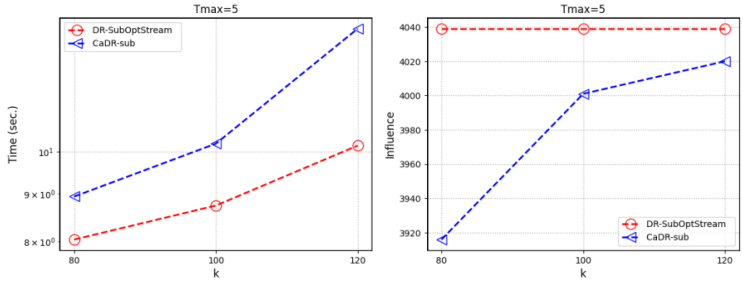
Phần này thảo luận và đánh giá các kết quả thực nghiệm để làm rõ những ưu điểm và hạn chế của hai thuật toán DR-SubOptStream và CaDR-sub thông qua hai giá trị quan trọng, là giá trị hàm mục tiêu $\boldsymbol{f(x)}$ (tầm ảnh hưởng lan truyền - influence) và thời gian thực thi (time). Kết quả thực nghiệm trên ba bộ dữ liệu trình bày rõ trong các Hình 7, 8, 9.

Nhận xét tổng thể, thuật toán CaDR-sub thu được giá trị tầm ảnh hưởng nhiều hơn thuật toán DR-SubOptStream nhưng đánh đổi là thời gian thực thi lâu hơn. Cụ thể, riêng với bộ dữ liệu Social circles: Facebook, có thể do cấu trúc của dữ liệu mà DR-SubOptStream chiếm ưu thế hoàn toàn, vừa có thời gian thực thi nhanh hơn từ 1,11 đến 1,33 lần và tầm ảnh hưởng lớn hơn 1,03 lần.

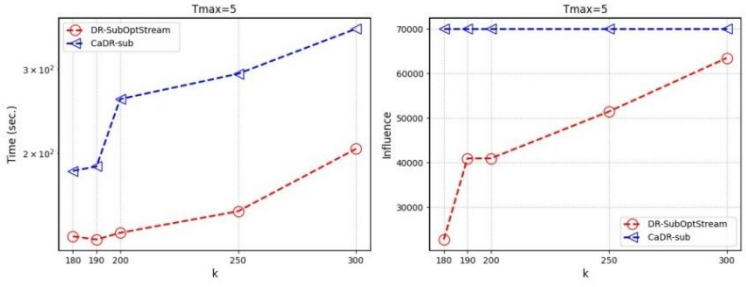


Hình 7. Kết quả thực nghiệm so sánh thời gian và tầm ảnh hưởng của 2 thuật toán trên bộ dữ liệu email-Eu-core network

Ngoài ra, một điều có thể dễ dàng nhận thấy, ở cả ba bộ dữ liệu, khi k càng tăng thì độ lớn về thời gian thực thi của CaDR-sub càng tăng so với DR-SubOptStream. Trong khi đó, khoảng cách giá trị tầm ảnh hưởng của DR-SubOptStream càng được tiệm cận với giá trị tầm ảnh hưởng của CaDR-sub. Do đó, lợi ích về thời gian của thuật toán DR-SubOptStream là điểm mạnh đáng kể để so sánh với sự chênh lệch về tầm ảnh hưởng.



Hình 8. Kết quả thực nghiệm so sánh thời gian và tầm ảnh hưởng của hai thuật toán trên bộ dữ liệu Social circles: Facebook



Hình 9. Kết quả thực nghiệm so sánh thời gian và tầm ảnh hưởng của hai thuật toán trên bộ dữ liệu Social circles: Twitter.

7. KẾT LUẬN

Trong bài báo này, nhóm tác giả nghiên cứu bài toán tối đa hóa hàm DR-submodular trên lưới nguyên dương và ứng dụng trong tối đa hóa tầm ảnh hưởng của việc lan truyền tiếp thị trên các cộng đồng của mạng xã hội. Đồng thời, nghiên cứu này cũng đề xuất một quy trình xử lý dữ liệu từ dạng đồ thị thông thường sang dạng đồ thị lưỡng cực. Qua quá trình thực nghiệm trên 3 bộ dữ liệu MXH, kết quả cho thấy ưu điểm của thuật toán đề xuất DR-SubOptStream có khả năng mở rộng cho dữ liệu lớn khi dữ liệu càng mở rộng, chi phí cấp càng tăng thì tầm ảnh hưởng và thời gian thực thi càng thể hiện tính vượt trội khi so sánh với thuật toán CaDR-sub.

Trong công việc nghiên cứu tiếp theo, nhóm tác giả sẽ tập trung vào nghiên cứu bài toán tối ưu hóa hàm DR-submodular đơn điệu và không đơn điệu ứng dụng trong các bài toán học máy, hoặc trong ngữ cảnh của ràng buộc về giới hạn thời gian và chi phí.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Trường Đại học Công Thương Thành phố Hồ Chí Minh, thuộc đề tài nghiên cứu khoa học và công nghệ cấp cơ sở hợp đồng số 82/HĐ-DCT ngày 15/08/2023.

TÀI LIỆU THAM KHẢO

1. Fujishige S. - Submodular Functions and Optimization. Elsevier (2005) 71-104.
2. Krause A., Golovin D. - Submodular function maximization. In Tractability: Practical approaches to hard problems, Cambridge University Press (2014). http://dx.doi.org/10.1017/CBO9781139177801.004
3. Lin H., Bilmes J.- Multi-document summarization via budgeted maximization of submodular functions. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (2010) 912–920. https://aclanthology.org/N10-1134.pdf
4. Lin H., Bilmes J. - A class of submodular functions for document summarization. In Proc.of NAACL (2011) 510-520. https://aclanthology.org/P11-1052.pdf
5. Krause A, Singh A., Guestrin C. - Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. The Journal of Machine Learning Research **9** (2008) 235-284. https://dl.acm.org/doi/10.5555/1390681.1390689
6. Hatano D., Fukunaga T., Maehara T., and Kawarabayashi K. - Lagrangian decomposition algorithm for 371 allocating marketing channels, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial 372 Intelligence, AAAI Press (2015) 1144–1150. https://cdn.aaai.org/ojs/9358/9358-13-12886-1-2-20201228.pdf
7. Feldman M., Nutov Z., Shoham E. - Practical budgeted submodular maximization. Algorithmica **85** (2023) 1332-1371. https://dl.acm.org/doi/abs/10.1007/s00453-022-01071-2

   8.Kempe D., Kleinberg J., Tardos E. - Maximizing the spread of influence through a social network. In Proc. of KDD (2003) 137–146. https://dl.acm.org/doi/10.1145/956750.956769
9. Nguyen, BNT., Pham PNH., Le VV., and Snášel V. - Influence maximization under fairness budget distribution in online social networks. Mathematics **10** (22) (2022) 4185. https://doi.org/10.3390/math10224185
10. DeValve L., Pekeč S., Wei Y. - Approximate submodularity in network design problems. Operations Research **71** (4) (2023) 1021-1039. https://dl.acm.org/doi/abs/10.1287/opre.2022.2408
11. Pham PNH., Nguyen BNT., Pham CV., Nghia DN., Snášel V. - Efficient algorithm for multiple benefit thresholds problem in online social networks. In 2021 RIVF International Conference on Computing and Communication Technologies (RIVF) (2021) 1-6. http://dx.doi.org/10.1109/RIVF51545.2021.9642099
12. Soma T., Yoshida Y. - Maximizing monotone submodular functions over the integer lattice, Mathematical Programming **172** (1) (2018) 539-563. https://doi.org/10.1007/s10107-018-1324-y
13. Iyer R., Khargonkar N., Bilmes J., Asnani H. - Generalized submodular optimization: Theory, algorithms and applications, Northwestern University, Dissertation of Doctor of Philosophy (2023) 31-34. https://doi.org/10.21985/n2-4ksz-9d55
14. Mitrovic S., Bogunovic I, Norouzi Fard A., Tarnawski J., Cevher V. - Streaming Robust Submodular Maximization: A Partitioned Thresholding Approach. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (2017) 4557–4566. https://dl.acm.org/doi/10.5555/3294996.3295209
15. Gu S., Shi G., Wu W., and Lu C. - A fast double greedy algorithm for non-monotone DR-submodular function maximization. Discrete Mathematics, Algorithms and Applications **12** (01) (2020) 2050007. https://doi.org/10.1142/S179383092050007X
16. Liu B., Chen Z., Du HW. - Streaming Algorithms for Maximizing DRSubmodular Functions with d-Knapsack Constraints. In: Algorithmic Aspects in Information and Management - 15th International Conference, AAIM. Springer, vol. **13153** (2021) 159-169. https://dl.acm.org/doi/10.1007/978-3-030-93176-6_14

17. Zhang Z., Guo L., Wang Y., Xu D., Zhang D. - Streaming algorithms for maximizing monotone dr-submodular functions with a cardinality constraint on the integer lattice. Asia Pac. J. Oper. Res. **38** (5) (2021) 2140004:1–2140004:14. https://doi.org/10.1142/S0217595921400042
18. Gong S., Nong Q., Bao S., Fang Q., Du DZ. - A fast and deterministic algorithm for Knapsack-constrained monotone DR-submodular maximization over an integer lattice. Journal of Global Optimization (2022) 1–24. https://doi.org/10.1007/s10898-022-01193-5
19. Soma T., Yoshida Y. - A generalization of submodular cover via the diminishing return property on the integer lattice. Advances in neural information processing systems 28. Annual Conference on Neural Information Processing Systems (2015) 847–855. https://dl.acm.org/doi/10.5555/2969239.2969334
20. Skiena S. - "Coloring Bipartite Graphs." §5.5.2 in Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. Reading, MA: Addison-Wesley, p. 213, 1990.
21. Badanidiyuru A., Mirzasoleiman B., Karbasi A., Krause A. - Streaming submodular maximization: Massive data summarization on the Fly. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery (2014) 671–680. https://dl.acm.org/doi/10.1145/2623330.2623637
22. Leskovec J., Sosic R. - SNAP: A general-purpose network analysis and graph-mining library. ACM Transactions on Intelligent Systems and Technology (TIST) **8** (1) (2016) 1-20. https://dl.acm.org/doi/10.1145/2898361
23. Clauset A., Newman M.E., Moore C. - Finding community structure in very large networks. Physical review E **70** (6) (2004) 066111. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.70.066111
24. Dugué N,. Perez A. - Directed Louvain: maximizing modularity in directed networks (Doctoral dissertation, Université d'Orléans) (2015). https://doi.org/10.1016/j.physa.2022.127798
25. Leskovec J., Kleinberg JM., Faloutsos C. - Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD) b (1) (2007). https://dl.acm.org/doi/10.1145/1217299.1217301
26. McAuley J., Leskovec J. - Learning to discover social circles in ego networks. Advances in Neural Information Processing Systems (2012) 25. https://dl.acm.org/doi/10.5555/2999134.2999195

ABSTRACT

OPTIMIZING THE DR-SUBMODULAR FUNCTION ON THE INTEGER LATTICE TO MAXIMIZE THE INFLUENCE OF VIRAL MARKETING ON COMMUNITIES IN SOCIAL NETWORKS

Nguyen Thi Bich Ngan, Nguyen Truong Phat, Do The Sang, Pham Nguyen Huy Phuong

Ho Chi Minh City University of Industry and Trade

*Email: *phuongpnh@huit.edu.vn*

As society continues to develop, individuals encounter increasingly complex optimization problems with multiple objectives to achieve. One such problem is optimizing a DR-submodular function, characterized by diminishing returns. In this article, we focus on maximizing the influence of marketing spread on social network communities, using a novel technique known as streaming data browsing. Our proposed DR-SubOptStream algorithm yields positive results, achieving an acceptable approximation of the objective function value and better complexity than existing algorithms. To conduct experiments, we transform social network data from a connected graph form to bipartite data form and then run the algorithm on preprocessed datasets. Overall, our findings demonstrate the effectiveness of our approach in solving this type of problem.

Keywords: DR-submodular function, optimization problem, streaming algorithm, Sieve streaming algorithm, bipartite graph.