SEISMIC: Mô hình quy trình điểm tự hào hứng
để Dự đoán mức độ phổ biến của Tweet

Qingyuan Zhao
Đại học Stanford
qyzhao@stanford.edu

Murat A. Erdogdu
Đại học Stanford
erdogdu@stanford.edu

Hera Y. He
Đại học Stanford
yhe1@stanford.edu

Anand Rajaraman
Đại học Stanford
anand@cs.stanford.edu

Jure Leskovec
Đại học Stanford
jure@cs.stanford.edu

TRỪU TƯỢNG

Các trang web mạng xã hội cho phép người dùng tạo và chia sẻ nội dung. Các dòng thông tin lớn chia sẻ lại bài đăng có thể hình thành khi người dùng các trang web này chia sẻ lại bài đăng của người khác với bạn bè và người theo dõi của họ. Một trong những thách thức chính trong việc hiểu các hành vi xếp tầng như vậy là trong việc dự báo bùng phát thông tin, nơi một bài đăng duy nhất trở nên phổ biến rộng rãi do được nhiều người dùng chia sẻ lại.

Trong bài báo này, chúng tôi tập trung vào việc dự đoán số lượt chia sẻ lại cuối cùng của một bài đăng nhất định. Chúng tôi xây dựng dựa trên lý thuyết về các quá trình điểm tự động để phát triển một mô hình thống kê cho phép chúng tôi đưa ra các dự đoán chính xác. Mô hình của chúng tôi không yêu cầu đào tạo hoặc kỹ thuật tính năng đắt tiền. Nó dẫn đến một công thức tính toán đơn giản và hiệu quả cho phép chúng tôi trả lời các câu hỏi, trong thời gian thực, chẳng hạn như: Với lịch sử chia sẻ lại của một bài đăng cho đến nay, ước tính hiện tại của chúng tôi về số lượng chia sẻ lại cuối cùng của bài đăng đó là bao nhiêu? Dòng chia sẻ lại bài đăng có đang trong giai đoạn tăng trưởng bùng nổ ban đầu không? Và, những bài đăng nào sẽ được chia sẻ lại nhiều nhất trong tương lai?

Chúng tôi xác thực mô hình của mình bằng cách sử dụng một tháng dữ liệu Twitter hoàn chỉnh và chứng minh sự cải thiện mạnh mẽ về độ chính xác dự đoán so với các phương pháp hiện có. Mô hình của chúng tôi chỉ đưa ra sai số tương đối 15% trong việc dự đoán kích thước cuối cùng của dòng thông tin trung bình sau khi quan sát nó chỉ trong một giờ.

Trình mô tả danh mục và chủ đề: H.2.8 [Quản lý cơ sở dữ liệu]: Ứng dụng cơ sở dữ liệu—*Khai thác dữ liệu*
Điều khoản chung: Các thuật toán; Thử nghiệm.
Từ khóa: sự lan tỏa thông tin; dự đoán thác; quá trình tích điểm tự thú vị; sự lây lan; truyền thông xã hội.

1. GIỚI THIỆU

Các dịch vụ mạng xã hội trực tuyến, chẳng hạn như Facebook, Youtube và Twitter, cho phép người dùng của họ đăng và chia sẻ nội dung dưới dạng bài đăng, hình ảnh và video [9, 17, 21, 30]. Khi người dùng tiếp xúc với bài đăng của những người khác mà cô ấy theo dõi, đến lượt người dùng có thể chia sẻ lại bài đăng với những người theo dõi của chính cô ấy, họ có thể chia sẻ lại bài đăng đó với nhóm người theo dõi tương ứng của họ. Bằng cách này, hàng loạt thông tin lớn của các bài chia sẻ lại lan truyền qua mạng.

KDD'15, Ngày 10-13 tháng 8 năm 2015, Sydney, NSW, Úc.

DOI: http://dx.doi.org/10.1145/2783258.2783401.

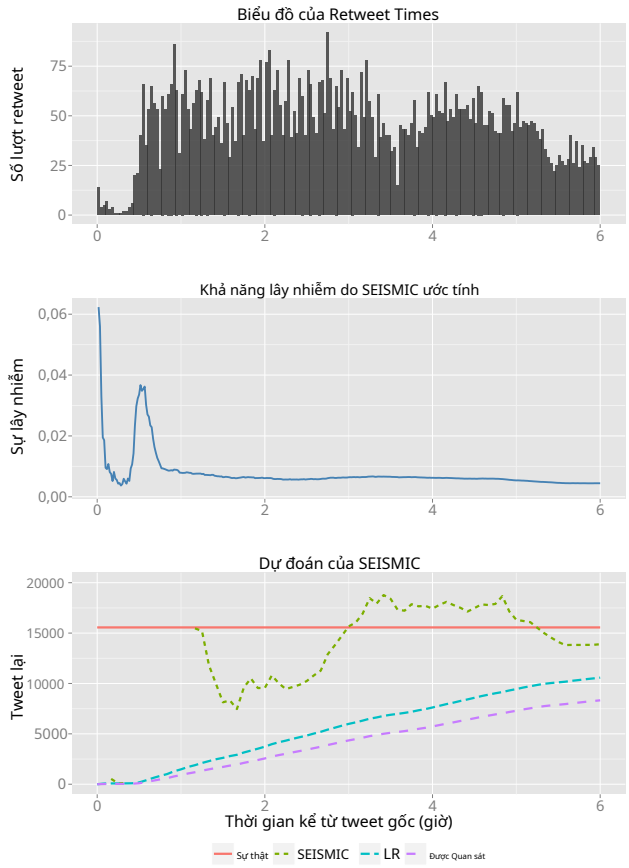
Một câu hỏi cơ bản trong mô hình hóa các tầng thông tin là dự đoán sự tiến hóa trong tương lai của chúng. Có thể cho rằng cách trực tiếp nhất để hình thành câu hỏi này là xem xét việc dự đoán kích thước cuối cùng của một tầng thông tin. Đó là, để dự đoán cuối cùng sẽ nhận được bao nhiêu lượt chia sẻ lại một bài đăng nhất định.

Dự đoán mức độ phổ biến cuối cùng của một bài đăng rất quan trọng đối với xếp hạng và tổng hợp nội dung. Ví dụ, Twitter tràn ngập các bài đăng và người dùng gặp khó khăn trong việc theo dõi tất cả chúng. Do đó, nhiều nội dung bị bỏ sót và cuối cùng bị mất. Dự đoán chính xác sẽ cho phép Twitter xếp hạng nội dung tốt hơn, khám phá các bài đăng thịnh hành nhanh hơn và cải thiện mạng phân phối nội dung của nó. Hơn nữa, dự đoán các dòng thông tin cho phép chúng ta có được những hiểu biết cơ bản về khả năng dự đoán của các hành vi tập thể trong đó các hành động không phối hợp của nhiều cá nhân dẫn đến kết quả tự phát, ví dụ như bùng phát thông tin lớn.

Hầu hết các nghiên cứu về dự đoán các tầng thông tin liên quan đến việc trích xuất một tập hợp đầy đủ các tính năng mô tả quá trình phát triển trong quá khứ của một tầng và sau đó sử dụng các tính năng này trong một bộ phân loại học máy đơn giản để đưa ra dự đoán về sự tăng trưởng trong tương lai [4, 6, 17, 20, 26, 30]. Tuy nhiên, việc trích xuất tính năng có thể tốn kém và cồng kềnh, và người ta không bao giờ chắc chắn liệu có thể trích xuất các tính năng hiệu quả hơn hay không. Câu hỏi vẫn là làm thế nào để thiết kế một mô hình từ dưới lên đơn giản và có nguyên tắc về hành vi xếp tầng. Thách thức nằm ở việc xác định một mô hình cho hành vi của một cá nhân và sau đó tổng hợp các tác động của các cá nhân để đưa ra một dự đoán toàn cầu chính xác.

Công việc hiện tại. Ở đây chúng tôi tập trung vào việc dự đoán kích thước cuối cùng của một dòng thông tin lan truyền qua mạng. Chúng tôi phát triển một mô hình thống kê dựa trên lý thuyết về *quy trình điểm tự thú vị*. Một quá trình điểm được lập chỉ mục theo thời gian được gọi là *quá trình đếm* khi nó đếm số lượng phiên bản (chia sẻ lại, trong trường hợp của chúng tôi) theo thời gian. Trái ngược với các quá trình Poisson đồng nhất giả định cường độ không đổi theo thời gian, các quá trình tự kích thích giả định rằng tất cả các trường hợp trước đó (*I E*, chia sẻ lại) ảnh hưởng đến sự phát triển trong tương lai của quy trình. Các quy trình điểm tự hào hứng thường được sử dụng để mô hình hóa các hiện tượng "giàu càng giàu" [22, 23, 33, 36]. Chúng lý tưởng để lập mô hình các dòng thông tin trong mạng vì mỗi lượt chia sẻ lại bài đăng mới không chỉ làm tăng số lượt chia sẻ lại tích lũy của nó lên một lượt mà còn cho thấy những người theo dõi mới có thể chia sẻ lại bài đăng đó.

Chúng tôi phát triển SEISMIC (*Mô hình tự thú vị của các tầng thông tin*) để dự đoán tổng số lượt chia sẻ lại của một bài đăng nhất định. Trong mô hình của chúng tôi, mỗi bài đăng đều được đặc trưng bởi *sự lây nhiễm* đo xác suất chia sẻ lại. Chúng tôi cho phép khả năng lây nhiễm thay đổi tự do theo thời gian với quan sát rằng khả năng lây nhiễm có thể giảm khi nội dung bị cũ (xem Hình 1).



Hình 1: 6 giờ đầu tiên hoạt động retweet của một tweet phổ biến [1] (trên cùng). Dòng tweet gây tranh cãi nói về cái chết tươi mới của nhà độc tài Muammar Gaddafi và đề cập đến ca sĩ Justin Bieber. Điều thú vị là, tài khoản Twitter của nhà sản xuất ô tô Chevrolet đã đăng lại tweet một cách không thích hợp khoảng 30 phút sau tweet ban đầu, điều này có thể dẫn đến sự phổ biến lâu dài của tweet. Khả năng lây nhiễm của Tweet theo thời gian được ước tính bởi NSEISMIC (ở giữa). Dự đoán về số lượt retweet cuối cùng của tweet (được biểu thị là "Sự thật") như một hàm của thời gian (dưới cùng). Chúng tôi so sánhNSEISMIC với hồi quy tuyến tính chuỗi thời gian (LR), "Đã quan sát" vẽ biểu đồ số lượt retweet tích lũy được quan sát trong một thời gian nhất định. Lưu ýNSEISMIC nhanh chóng tìm ra ước tính chính xác về số lượt retweet cuối cùng của tweet.

Hơn nữa, mô hình của chúng tôi có thể xác định tại mỗi thời điểm xem liệu thác có trong *siêu tới hạn* hoặc *cận tới hạn* dựa trên mức độ lây nhiễm của nó trên hay dưới ngưỡng tới hạn. Một thác nước ở trạng thái siêu tới hạn đang trải qua thời kỳ "bùng nổ" và kích thước cuối cùng của nó không thể được dự đoán chính xác tại thời điểm hiện tại. Ngược lại, một dòng thác có thể điều chỉnh được nếu nó ở trạng thái dưới tới hạn. Trong trường hợp này, chúng ta có thể dự đoán chính xác mức độ phổ biến cuối cùng của nó bằng cách lập mô hình hành vi xếp tầng trong tương lai của cây Galton-Watson.

Của chúng tôiEISMIC phương pháp tiếp cận có một số đóng góp:

- Mô hình tạo: NSEISMIC không áp đặt các giả định tham số và không yêu cầu kỹ thuật tính năng đắt tiền. Hơn nữa, vì cấu trúc mạng xã hội hoàn chỉnh có thể khó có được, SEISMIC giả định kiến thức tối thiểu về mạng: Đầu vào bắt buộc duy nhất là lịch sử thời gian của các lần chia sẻ lại và mức độ của các nút chia sẻ lại.
- Tính toán có thể mở rộng: Đưa ra dự đoán bằng SEISMIC chỉ yêu cầu tuyến tính thời gian tính toán về số lượt chia sẻ lại được quan sát. Vì các dự đoán cho các bài đăng riêng lẻ có thể được thực hiện độc lập, nên thuật toán của chúng tôi cũng có thể dễ dàng song song hóa.
- Dễ hiểu: Đối với một tầng riêng lẻ, mô hình tổng hợp tất cả lịch sử quá khứ của nó thành một thông số lây nhiễm duy nhất. Thông số lây nhiễm này có ý nghĩa rõ ràng và có thể dùng làm đầu vào cho các ứng dụng khác.

Chúng tôi đánh giá SEISMIC trong một tháng toàn bộ dữ liệu Twitter, nơi người dùng đăng các tweet mà sau đó người khác có thể chia sẻ lại bằng cách đăng lại chúng. Chúng tôi chứng minh rằng SEISMIC có thể dự đoán số lượt retweet cuối cùng của một tweet nhất định với độ chính xác tốt hơn 30% so với các phương pháp hiện đại (*ví dụ*, [12]). Đối với các tweet phổ biến hợp lý, mô hình của chúng tôi đạt được lỗi tương đối 15% khi dự đoán số lượt tweet lại cuối cùng sau khi quan sát tweet trong 1 giờ và lỗi 25% sau khi quan sát tweet chỉ trong 10 phút. Hơn nữa, chúng tôi cũng chứng minh cách SEISMIC có thể xác định các tweet sẽ "lan truyền" và nằm trong số các tweet phổ biến nhất trong tương lai. Bằng cách duy trì danh sách 500 tweet động theo thời gian, chúng tôi có thể xác định 78 trong số 100 tweet được chia sẻ lại nhiều nhất và 281 trong số 500 tweet được chia sẻ lại nhiều nhất chỉ trong 10 phút sau khi chúng được đăng.

Phần còn lại của bài báo được sắp xếp như sau: Phần 2 khảo sát các công việc liên quan. Phần 3 mô tả SEISMICvà Phần 4 chỉ ra cách mô hình có thể được sử dụng để dự đoán kích thước cuối cùng của một tầng thông tin. Chúng tôi đánh giá phương pháp của mình và so sánh hiệu quả của nó với một số đường cơ sở cũng như các phương pháp tiếp cận hiện đại trong Phần 5. Cuối cùng, trong Phần 6, chúng tôi kết luận và thảo luận về các hướng nghiên cứu trong tương lai.

2. LIÊN QUAN

Nghiên cứu về các tầng thông tin là một lĩnh vực phong phú và tích cực [27]. Các mô hình gần đây để dự đoán kích thước của các tầng thông tin thường được đặc trưng bởi hai loại phương pháp tiếp cận, phương pháp dựa trên tính năng và phương pháp dựa trên quy trình điểm.

Các phương pháp dựa trên tính năng trước tiên trích xuất một danh sách đầy đủ các tính năng có thể có liên quan, bao gồm các tính năng nội dung, các tính năng áp phích gốc, các đặc điểm cấu trúc mạng và các đặc điểm thời gian [6]. Sau đó, các thuật toán học tập khác nhau được áp dụng, chẳng hạn như mô hình hồi quy đơn giản [2, 6], lọc cộng tác xác suất [35], cây hồi quy [3], mô hình dựa trên nội dung [24] và thuật toán tích cực thụ động [26]. Có một số vấn đề với các phương pháp tiếp cận như vậy: kỹ thuật tính năng tốn nhiều công sức và đào tạo chuyên sâu là rất quan trọng cho sự thành công của họ và hiệu suất rất nhạy cảm với chất lượng của các tính năng [4, 30]. Các cách tiếp cận như vậy cũng có khả năng áp dụng hạn chế vì chúng không thể được sử dụng trong cài đặt trực tuyến thời gian thực — với số lượng lớn bài đăng được tạo ra mỗi giây, thực tế là không thể trích xuất tất cả các tính năng cần thiết cho mọi bài đăng và sau đó áp dụng các quy tắc dự đoán phức tạp. Ngược lại, SEISMIC không yêu cầu kỹ thuật tính năng và kết quả trong một công thức có thể tính toán hiệu quả cho phép nó dự đoán mức độ phổ biến cuối cùng của hàng triệu bài đăng khi chúng được lan truyền qua mạng.

Loại tiếp cận thứ hai dựa trên các quy trình điểm, mô hình hóa trực tiếp sự hình thành của một tầng thông tin trong một mạng. Các mô hình như vậy chủ yếu được phát triển cho vấn đề bổ sung của suy luận mạng, trong đó người ta quan sát một số tầng thông tin và cố gắng suy ra cấu trúc của mạng cơ sở mà các tầng truyền qua đó [8, 10, 13, 14, 15, 18, 33, 36]. Các phương pháp này đã được áp dụng thành công để nghiên cứu sự lan truyền của meme trên web [10, 14, 32, 33] cũng như thẻ bắt đầu bằng # trên

Biểu tượng	Sự miêu tả
w	Dòng thác bài đăng / thông tin Sự lây nhiễm
P_{NS}	của w ở thời điểm NS (Phần 3.2) Nhân bộ nhớ
$\varphi(NS)$	(Phần 3.1)
$tôi$	Nút đã đóng góp $tôi_{NS}$ chia sẻ lại.
$_{NS_{tôi}}$	$tôi = 0$ tương ứng với người khởi tạo bài đăng. Thời gian của $tôi_{NS}$ chia sẻ lại liên quan đến bài đăng gốc. Ngoài mức độ
$n_{tôi}$	của $tôi_{NS}$ nút Mức độ phổ biến tích lũy theo thời gian NS: $\lvert \{$
NS_{NS}	$tôi > 0;\ NS_{tôi} \le t\} \rvert$
NS_{∞}	Mức độ phổ biến cuối cùng (số lượt chia sẻ lại cuối cùng) $\sum : \lvert \{tôi > 0\} \rvert$
n_{NS}	Mức độ tích lũy của người chia sẻ lại theo thời gian NS: $_{tôi:NS_{tôi} \le NS}\, n_{tôi}$
$n_{\eta_{NS}}$	Mức độ tích lũy hiệu quả của người chia sẻ lại theo thời gian NS: $n_{\eta_{NS}} = \sum_{tôi=0}^{NS} n_{tôi} \int_{N S_{tôi}}^{NS} \varphi(s - t_{tôi})\, ds$
λ_{NS}	Cường độ phổ biến tích lũy NS_{NS}
P_{NS}	Ước tính của mô hình về khả năng lây nhiễm P_{NS} ở thời điểm NS
$NS_{\infty}(NS)$	Ước tính của mô hình tại thời điểm NS mức độ phổ biến cuối cùng NS_{∞}

Bảng 1: Bảng ký hiệu.

Twitter [36]. Ngược lại, mục tiêu của chúng ta không phải là suy ra mạng mà là dự đoán kích thước cuối cùng của một tầng trong một mạng được quan sát.

Sự khác biệt chính giữa mô hình của chúng tôi và các phương pháp hiện có dựa trên các quy trình của Hawkes (*ví dụ*, [22, 23, 33, 34, 36]) là chúng tôi giả định cường độ quá trình λ_{NS} phụ thuộc vào một process P_{NS}, tính truyền nhiễm sau. Nói cách khác, chúng tôi cho phép khả năng lây nhiễm thay đổi theo thời gian. Hơn nữa, một số meth ... ods [34] dựa trên suy luận Bayes về mặt tính toán đắt tiền, trong khi phương pháp của chúng tôi có độ phức tạp thời gian tuyến tính. Một công việc liên quan khác được đề xuất gần đây là [12], cũng sử dụng phương pháp tiếp cận quy trình điểm và trực tiếp nhằm mục đích dự đoán mức độ phổ biến của tweet. Tuy nhiên, phương pháp của họ đưa ra các giả định tham số hạn chế và không xem xét cấu trúc mạng, điều này làm hạn chế khả năng dự đoán của nó. Chúng tôi so sánh SEISMIC với [12] trong Phần 5 và thể hiện sự cải thiện 30%.

3. MÔ HÌNH THÔNG TIN CASCADES

Trong phần này, chúng tôi mô tả SEISMIC và thảo luận về cách nó có thể được sử dụng để:

1. Ước tính tốc độ lan truyền của một dòng thông tin nhất định, mà chúng tôi định lượng bằng mức độ lan truyền của bài đăng.
2. Xác định xem thác đang ở trạng thái siêu tới hạn (nổ) hay dưới tới hạn (sắp chết).
3. Dự đoán kích thước cuối cùng của một tầng thông tin, được đo bằng số lượt chia sẻ lại cuối cùng mà bài đăng bắt đầu tầng đó nhận được.

Các đại lượng quan trọng trong mô hình của chúng tôi là tổng số re-chia sẻ NS_{NS} của một bài nhất định cho đến thời điểm NS và tốc độ lan truyền theo tầng λ_{NS}. Trong mô hình của chúng tôi, λ_{NS} được xác định bởi bài viết infec- sự mệt mỏi P_{NS} và thời gian phản ứng của con người. Mục tiêu của chúng tôi là dự đoán NS_{∞}, số lượt chia sẻ lại cuối cùng.

Một số lượng quan trọng khác trong mô hình của chúng tôi là *nhân bộ nhớ* $\varphi(NS)$, định lượng độ trễ giữa một bài đăng đến nguồn cấp dữ liệu của người dùng và người dùng chia sẻ lại nó. Một cách trực quan, khả năng lây nhiễm xác định xác suất mà một người dùng nhất định sẽ chia sẻ lại một bài đăng nhất định và nhân bộ nhớ mô hình thời gian phản ứng của người dùng. Bằng cách kết hợp cả hai, chúng tôi có thể lập mô hình chính xác tốc độ mà bài đăng sẽ lan truyền qua mạng. Bảng 1 tóm tắt ký hiệu.

3.1 Thời gian phản ứng của con người

Để dự đoán kích thước tầng, chúng ta cần biết mất bao lâu để một người chia sẻ lại bài đăng. Biết được độ trễ cho phép chúng tôi lập mô hình chính xác tốc độ của một dòng thác lan qua mạng lưới. Chúng tôi coi đó là thời điểm NS giữa sự xuất hiện của một bài đăng trong dòng thời gian của người dùng và việc người dùng chia sẻ lại bài đăng được phân phối với mật độ $\varphi(NS)$. Mật độ xác suất $\varphi(NS)$ còn được gọi là nhân bộ nhớ vì nó đo lường bộ nhớ của hệ thống vật lý / xã hội về các kích thích [7].

Sự phân bố thời gian phản hồi của con người $\varphi(NS)$ đã được chứng minh là có nhiều quan điểm trong các mạng xã hội [5]. Thường là đuôi của $\varphi(NS)$ được giả định tuân theo luật lũy thừa với số mũ giữa 1 và 2 hoặc phân phối chuẩn log [7, 34]. Tuy nhiên, do tính chất nhanh chóng của việc chia sẻ thông tin trên Twitter, việc mong đợi nhiều lần phản ứng tức thì cũng là điều đương nhiên. Trên thực tế, phân tích dữ liệu khám phá của chúng tôi trong Phần 5.2 xác nhận rằng trong Twitter, $\varphi(NS)$ xấp xỉ không đổi trong 5 phút đầu tiên và sau đó là sự phân rã theo định luật lũy thừa. Các mạng xã hội khác nhau có thể có sự phân bổ thời gian phản ứng của con người khác nhau. Tuy vậy, $\varphi(NS)$ chỉ cần được ước tính một lần cho mỗi mạng và do đó chúng ta có thể cho rằng nó được đưa ra một cách an toàn. Chúng tôi mô tả quy trình ước tính chi tiết về $\varphi(NS)$ trong Phần 5.2.

3.2 Khả năng lây nhiễm sau

Thành phần thứ hai của mô hình của chúng tôi là khả năng lây nhiễm sau. Chúng tôi giả sử mỗi bài w được liên kết với một phụ thuộc thời gian, intrin- thông số lây nhiễm sic $P_{NS}(w)$. Nói cách khác, $P_{NS}(w)$ mô hình về khả năng bài đăng w sẽ được chia sẻ lại vào thời điểm NS. Sự lây nhiễm của một bài đăng có thể phụ thuộc vào sự kết hợp của nhiều yếu tố, bao gồm nhưng không giới hạn ở chất lượng nội dung của bài đăng, cấu trúc mạng xã hội, giờ địa phương hiện tại và vị trí địa lý. Trong- thay vì giả định một dạng tham số của P_{NS}, chúng tôi mô hình hóa nó một cách linh hoạt theo cách phi tham số, giải thích ngầm cho tất cả các yếu tố này.

Hầu hết các phương pháp hiện có nghiên cứu các quá trình điểm tự thú vị như- sume P_{NS} được cố định theo thời gian. Do đó, một khái niệm quan trọng là *sự nghiêm trọng* của quá trình NS_{NS}. Trong một quá trình tích điểm tự thú vị với tính lây nhiễm liên tục $P_{NS} \equiv P$, tồn tại hiện tượng chuyển pha ở ngưỡng tới hạn nhất định P_* sao cho [11]:

1. Nếu $p > p_*$, sau đó $NS_{NS} \to \infty$ như $NS \to \infty$ gần như chắc chắn và nhanh chóng theo cấp số nhân. Đây được gọi là *siêu tới hạn* chế độ.
2. Nếu $p < p_*$, sau đó $\sup_{NS} NS_{NS} < \infty$ gần như chắc chắn. Đây được gọi là *cận tới hạn* chế độ.

Thực tế, NS_{NS} luôn bị giới hạn do kích thước hữu hạn của mạng- công việc. Do đó, không có thác siêu tới hạn nào có thể tồn tại nếu P_{NS} được giả định là một hằng số. Điều này không đủ để mô hình hóa các tweet rất dễ lây lan và giả định của chúng tôi về khả năng lây nhiễm không liên tục cũng giải quyết được vấn đề này. Hơn nữa, khi bài viết cũ hơn, thông tin sẽ trở nên lỗi thời và sức lan truyền (tính truyền nhiễm) của nó có thể giảm xuống. Hiệu ứng này cũng có thể được quan sát thấy khi bài đăng lan ra xa hơn so với áp phích gốc [34]. Ngoài ra, việc chia sẻ lại bởi một người dùng có ảnh hưởng lớn có thể làm tăng khả năng lây lan của bài đăng. Vì vậy, thay vì giả định một mô hình tiến hóa chung của P_{NS} đối với tất cả các tweet, chúng tôi chỉ giả định rằng nó thay đổi trơn tru theo thời gian và sử dụng các phương pháp phi tham số để ước tính P_{NS} cho mỗi tweet.

3.3 Mô hình SEISMIC

Chúng tôi kết hợp thời gian phản ứng của con người và sau khả năng lây nhiễm để loại bỏ rive SEISMIC. Để liên kết P_{NS} đến quy trình chia sẻ lại bài đăng NS_{NS}, chúng tôi làm mẫu NS_{NS} như một *quy trình điểm tự thú vị kép ngẫu nhiên* . Đây là một phần mở rộng cho quy trình tiêu chuẩn về điểm tự thú vị (cũng được gọi là quá trình Hawkes [16]) ban đầu được sử dụng để lập mô hình động đất [25].

Đầu tiên chúng ta xác định cường độ λ_{NS} của NS_{NS}, chỉ đơn giản là đo lường tỷ lệ có được một lượt chia sẻ lại bổ sung tại thời điểm NS. Chính thức hơn:

$$\lambda_{NS} = \lim_{\Delta \downarrow 0} \frac{\mathrm{P}\,(NS_{NS+\Delta} - NS_{NS} = 1)}{\Delta}.$$

Trong SEISMIC, cường độ λ_{NS} ở thời điểm NS được xác định bởi infec-
sự mệt mỏi P_{NS}, số lần chia sẻ lại $NS_{tôi}$, độ nút $n_{tôi}$ và phân phối thời gian phản
ứng của con người $\varphi(NS)$. Mối quan hệ chính xác được mô tả trong Eq. (1)
được lấy cảm hứng từ lý thuyết về các quá trình của Hawkes [16]:

$$\lambda_{NS} = P_{NS} \cdot \sum_{NS_{tôi} \le t,\ tôi \ge 0} n_{tôi} \varphi(t - t_{tôi}), \qquad NS \ge NS_0. \tag{1}$$

Một cách trực quan, $\sum_{NS_{tôi} \le t,\ tôi \ge 0} n_{tôi} \varphi(t - t_{tôi})$ là cường độ của sự xuất hiện
người dùng mới tiếp xúc tại thời điểm NS và sản phẩm của nó với tính năng chia sẻ lại
xác suất P_{NS} cung cấp cường độ chia sẻ lại tại thời điểm NS. Lưu ý
rằng quá trình điểm trên được gọi là *tự hào* tại vì
từng quan sát trước đó *tôi* như vậy mà $NS_{tôi} \le NS$ góp phần vào-
mật độ λ_{NS}, hoặc tương đương, mỗi lần quan sát sẽ tăng cường độ trong
tương lai. Nó còn xa hơn *ngẫu nhiên kép* (hoặc một *Quy trình Cox*) thì là ở-
gây ra sự lây nhiễm P_{NS} bản thân nó là một quá trình ngẫu nhiên.

Ngoài ra, chúng tôi giả định độ nút $\{n_{tôi}\}$ được phân phối độc lập và giống
hệt nhau với giá trị trung bình n_*. Mức độ trung bình n_* có liên quan đến
ngưỡng quan trọng P_* mà đã được thảo luận trong Phần 3.2.
Ngưỡng lây nhiễm quan trọng có giá trị $P_* = 1/n_*$. Chúng tôi đưa ra
bằng chứng về thực tế này trong Đề xuất 4.1.

4. THÔNG TIN DỰ BÁO

Trong phần này, chúng tôi mô tả cách thực hiện suy luận thống
kê cho mô hình tự động của các tầng thông tin đã được giới thiệu
trong phần trước. Cụ thể, chúng tôi thảo luận về cách SEISMIC esti-
giao phối thông số lây nhiễm P_{NS} và sau đó dự đoán điều cuối cùng
kích thước của thác NS_∞.

Trong suốt phần này, chúng tôi đưa ra giả định kỹ thuật rằng
những người theo dõi tất cả những người chia sẻ lại là rời rạc, vì vậy chúng ta có thể
sử dụng cấu trúc cây để mô tả sự lan tỏa thông tin (Hình 2). Các kết luận được đưa ra
trong phần này vẫn có giá trị ngay cả khi người chia sẻ lại không
phân chia. Trong trường hợp này, chúng ta có thể thay thế mức độ nút $n_{tôi}$ với
tổng số lân cận mới được tiếp xúc của nút *tôi* (những người theo dõi
của *tôi*-người chia sẻ lại thứ không theo dõi người đầu tiên *tôi* - 1 người chia sẻ).

4.1 Ước tính khả năng lây nhiễm sau lây nhiễm

Đầu tiên chúng tôi xác định *mật độ chức năng mẫu*, đóng vai trò trung tâm
vai trò trong các quá trình điểm kích thích es (timating self-e) [29]. Hãy biểu thị
$NS_{NS} = \sigma\{(n_{tôi}, NS_{tôi})\}_{NS_{tôi} \ge 0}$ như là σ-algebra được tạo bởi tất cả trong-
hình thành có sẵn theo thời gian NS: thời gian $NS_{tôi}$ trong số tất cả các lượt chia sẻ lại
đến lúc NS và số lượng người theo dõi (IE, mức độ nút) $n_{tôi}$ sau đó
tôi-người dùng thứ để chia sẻ lại. Mật độ chức năng mẫu được định nghĩa là khớp
xác suất của số lượt chia sẻ lại trong khoảng thời gian $[NS_0, NS)$ và
mật độ số lần xuất hiện của chúng.

Để thúc đẩy công cụ ước tính của chúng tôi P_{NS}, trước tiên chúng ta xem
xét trường hợp thông số lây nhiễm không đổi theo thời gian, IE, $P_{NS} \equiv P$. Sau đó, chúng tôi sẽ nới lỏng giả định này và cho phép P_{NS} thay đổi theo thời gian.

Trong SEISMIC, mật độ hàm mẫu có thể được biểu thị bằng cách sử dụng
cường độ λ_{NS} như [29, Thm. 6.2.2]

$$P(NS_{NS} = r, t_1, \ldots, NS_{NS}) = \prod_{tôi=1}^{NS_{NS}} \lambda_{NS_{tôi}} \cdot \exp\left\{ - \int_{NS_0}^{NS} \lambda_{NS}\, ds. \right\} \tag{2}$$

Bằng cách lấy đạo hàm của nhật ký Eq. (2) và kết hợp nó với Eq. (1),
chúng tôi thu được ước tính khả năng xảy ra tối đa (MLE) của P_{NS}:

$$P_{NS} = \sum_{tôi=0}^{NS_{NS}} \frac{n_{tôi}}{\int_{NS_0}^{NS} \varphi(s - t_{tôi})\, ds} \tag{3}$$

Phương trình trên là cơ sở của SEISMIC vì nó cho phép chúng tôi
để ước tính khả năng lây nhiễm P_{NS} tại bất kỳ thời điểm nào NS. Hơn
nữa, khoảng tin cậy của P_{NS} cũng có thể được [29].

Mẫu số trong Eq. (3), ký hiệu là $n_{e\ NS}$ sau đây, có thể trong-
được hiểu là số lượng người dùng tiếp xúc "hiệu quả" tích lũy
đến bài đăng. Tử số NS_{NS} là số lượt chia sẻ lại bài đăng hiện tại. Để làm sáng tỏ
hơn về công cụ ước tính của chúng tôi, chúng tôi sử dụng $NS \to \infty$,
dẫn đến:

$$P(\infty) = \frac{1}{\frac{1}{NS_\infty} \sum_{NS=0}^{NS_\infty} n_{NS}} \approx \frac{1}{n_*} \tag{4}$$

Do đó, bằng cách giả định tính lây nhiễm P_{NS} là một hằng số theo thời gian, về
cơ bản người ta sẽ cho rằng hầu hết các bài đăng đều có cùng một thông tin
sự mệt mỏi $1/n_*$. Tuy nhiên, giả định như vậy là không thực tế vì nó không thể
giải thích các dòng thông tin động lực học bùng nổ và dễ bay hơi
(*ví dụ*, Hình 1).

Hệ quả không mong muốn này của việc giả định là không đổi P_{NS} là một cái khác
động lực để cho phép P_{NS} thay đổi theo thời gian. Ước tính, ước lượng P_{NS} trong trường hợp
này, chúng tôi làm mịn MLE trong Eq. (3) bằng cách chỉ sử dụng các quan sát
gần với thời gian NS ước tính, ước lượng P_{NS}. Đặc biệt, chúng tôi dựa trên một trình tự
của các hạt nhân một phía $K_{NS}(NS)$, $s > 0$, được lập chỉ mục theo thời gian NS. Chúng tôi sử dụng các
hạt nhân này để xác định trọng số của các lượt chia sẻ lại và ước tính có trọng số của
P_{NS} được đưa ra bởi

$$P_{NS} = \frac{\int_{NS_0}^{NS} K_{NS}(t - s) N_{NS}}{\int_{NS_0}^{NS} K_{NS}(t - s)\, dN_{e\ NS}} = \sum_{tôi=0}^{NS_{NS}} \frac{n_{tôi}}{\int_{NS_0}^{NS} K_{NS}(t - s) \varphi(s - t_{tôi})\, ds}. \tag{5}$$

Chú ý rằng khi $K_{NS}(NS) \equiv 1$ công cụ ước tính giảm xuống MLE mà chúng tôi thu
được trong Eq. (3). Trong SEISMIC chúng tôi sử dụng một nhân tam giác với
kích thước cửa sổ đang phát triển $NS/2$ như trọng lượng hạt nhân $K_{NS}(NS)$:

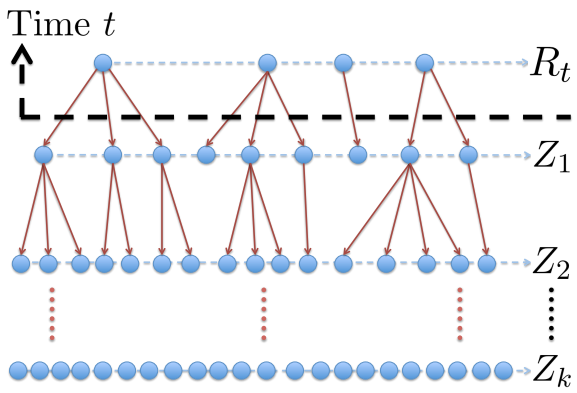
$$K_{NS}(NS) = \text{tối đa}\left\{0,\ 1 - \frac{2}{NS - NS_0},\ s > 0.\right. \tag{6}$$

Chúng tôi chọn nhân hình tam giác vì nó có các thuộc tính quan trọng
đối với ứng dụng của chúng tôi. Đầu tiên, hạt nhân loại bỏ tất cả các bài
đăng cũ hơn $NS/2$. Đặc biệt, nó nhanh chóng loại bỏ giai đoạn không ổn
định và có khả năng bùng nổ lúc đầu, nếu bao gồm,
sẽ giới thiệu một khuynh hướng đi lên đối với P_{NS}. Thứ hai, hạt nhân tính đến các bài
đăng trong kích thước cửa sổ lớn hơn theo thời gian NS tăng. Phù hợp-
trong các thử nghiệm của chúng tôi, kích thước cửa sổ ngày càng tăng giúp ổn định
$P(NS)$ so với kích thước cửa sổ cố định. Thứ ba, để chia sẻ lại trong cửa sổ, hạt
nhân tăng trọng số các bài đăng gần đây nhất và giảm dần các bài đăng cũ
hơn. Điều này giữ cho công cụ ước tính của chúng tôi $P(NS)$ gần hơn
đến từng-chan∫ging đúng P_{NS}. Và cuối cùng, là $K_{NS}(NS)$ là tuyến tính mảnh, tích
phân $K_{NS}(t-s)\varphi(s-t_{tôi})ds$ có một hình thức đóng cho nhiều người
các chức năng khác nhau $\varphi(NS)$ bao gồm cả cái mà chúng tôi sử dụng cho Twitter trong các
thử nghiệm của mình, xem Phần 5.

4.2 Dự đoán mức độ phổ biến cuối cùng

Sau khi mô tả quy trình để suy ra mức độ lây nhiễm của bài đăng,
bây giờ chúng ta cần tính đến cấu trúc mạng để dự đoán mức độ
lan truyền của bài đăng trên mạng. Để đơn giản, chúng tôi giả sử
bài viết được đăng lần đầu tiên vào thời điểm 0, IE,
$NS_0 = 0$. Hãy xem xét chúng tôi đã quan sát bài đăng cho NS đơn vị thời gian và mục tiêu của
chúng tôi bây giờ là dự đoán số lượt chia sẻ lại cuối cùng của bài đăng, NS_∞, dựa trên
về thông tin chúng tôi đã quan sát được cho đến nay, NS_{NS}.

Mệnh đề sau đây chỉ ra cách tính toán
số lượt chia sẻ lại cuối cùng của một bài đăng. Ý tưởng chính là mô hình hóa
một chuỗi thông tin lan truyền trên mạng với quy trình phân nhánh đếm số
lượt chia sẻ lại của một bài đăng, như được minh họa trong Hình 2.
Dự đoán cho NS_∞ được sử dụng bởi SEISMIC có thể được phát biểu như sau:



Hình 2: Hình minh họa cây khuếch tán thông tin. Chúng tôi quan sát dòng thác theo thời gian NS (được biểu thị bằng một đường đứt nét) và câu hỏi đặt ra là làm thế nào mà cây tầng sẽ phát triển trong tương lai. Chúng tôi xác định các biến Z_k biểu thị số lượt chia sẻ lại do k_{NS} thế hệ con cháu. Sử dụng biến Z_k số lượt chia sẻ lại cuối cùng NS_∞ sau đó có thể đơn giản máy tính như $NS_{NS} + \sum_{k=1}^{\infty} Z_k$.

PVỊ TRÍ 4.1. *Giả sử độ (ngoài) trong mạng là iid với sự mong đợi n và thông số lây nhiễm P là một hằng số P vì $NS \geq NS$. Sau đó chúng tôi có*

$$\mathrm{E}[NS_\infty / NS_{NS}] = \begin{cases} NS_{NS} + \dfrac{P(N_{NS} - N_e)}{1 - pn_*}, & \text{nếu như } p < \dfrac{1}{n_*}, \\ \infty, & \text{nếu như } P \geq \dfrac{1}{n_*}. \end{cases} \quad (7)$$

PMÁI NHÀ. Đầu tiên, chúng tôi xem xét trường hợp $p < 1/n_*$. Chúng tôi xác định một chuỗi các biến ngẫu nhiên $\{Z_1, Z_2, Z_3, \ldots\}$ mô hình hóa cây khuếch tán thông tin trong tương lai, như được minh họa trong Hình 2. Trong cây, Z_k biểu thị số lượt chia sẻ lại được thực hiện bởi k_{NS} thế hệ con cháu (tính từ thế hệ NS_{NS} trở đi). Do đó, 1_{NS} thế hệ con cháu Z_1 đề cập đến số lượt chia sẻ lại mới được tạo bởi các bài đăng được tạo trước thời gian NS, 2_{NS} thế hệ hậu duệ Z_2 đề cập đến lượt chia sẻ lại các bài đăng của 1_{NS} hạ xuống-dants, và như vậy. Lưu ý rằng triệu hồi∑ tion qua Z_k của đưa ra số lượt chia sẻ lại cuối cùng của bài đăng $NS_\infty = NS_{NS} + \sum_{k=1}^{\infty} Z_k$. Trong những điều sau đây chúng tôi sử dụng con cháu Z_k chỉ để tính toán Eq. (7) và nhấn mạnh rằng công cụ ước tính cuối cùng của chúng tôi không yêu cầu cấu trúc mạng rõ ràng thông tin.

Được cho Z_1, chuỗi các biến ngẫu nhiên Z_k xác định một Galton-Cây Watson với *kỳ vọng con cái* $\mu = n_* P$ [11]. Ở đây, μ biểu thị số lượt chia sẻ lại dự kiến mà bài đăng nhận được. Sử dụng một kết quả quá trình phân nhánh tiêu chuẩn, chúng tôi có $Z_{tôi} / \mu_{tôi}$ là một martingale. Vì vậy, $\forall k > 1$, $\mathrm{E}[Z_{k+1} / Z_k] = \mu$ Z_k, và,

$$\mathrm{E}\left[\sum_{k=1}^{\infty} Z \middle| Z_{k-1}\right] = \frac{Z_1}{(1-\mu)} = \frac{Z_1}{(1 - n_* P)}.$$

Do đó, chúng tôi có được

$$\mathrm{E}[NS_\infty / F_{NS}] = NS_{NS} + \mathrm{E}\left[\sum_{k=1}^{\infty} Z_k\right] = NS_{NS} + \frac{\mathrm{E}[Z_1]}{(1 - n_* P)},$$

mà cuối cùng trở thành bên phải trong Eq. (7) bởi vìE $[Z_1] = P(n_{NS} - n_{e\,NS})$ theo định nghĩa của Z_1 và n_{eNS}.

Thuật toán 1 NSEISMIC: Dự đoán kích thước tầng cuối cùng

Mục đích: Đối với một bài đăng nhất định tại thời điểm NS, dự đoán số lượt chia sẻ lại cuối cùng

Đầu vào: Đăng thông tin chia sẻ lại: $NS_{tôi}$ và $n_{tôi}$ vì $tôi = 0, \ldots, NS_{NS}$.

Thuật toán:

$n_{NS} = 0$, $n_{eNS} = 0$

vì $tôi = 0, \ldots, NS_{NS}$ làm

$$n_{NS+} = n_{tôi} \int_{NS_{tôi}}^{NS} \varphi(s - t_{tôi})\, ds \quad \text{(Phần 3.1)}$$

kết thúc cho

$$NS_\infty(NS) = NS_{NS} + \alpha_{NS} P_{NS}(n_{NS} - n_{e\,NS}) / (1 - \gamma_{NS} P_{NS} n_*) \quad \text{(Bài 2)}$$

Giao: $NS_\infty(NS)$

Tiếp theo, hãy xem xét trường hợp $P = P_{NS} \geq 1/n_*$. Trong chế độ này, quá trình điểm là siêu tới hạn và vẫn bùng nổ. Về mặt cây Galton-Watson đã thảo luận ở trên, kỳ vọng con cái $\mu = n_* P \geq 1$, vì thế $\mathrm{E}[Z_k \sum_{k=1}^{\infty} 1] \geq \mathrm{E}[Z_k] \geq \cdots \geq \mathrm{E}[Z_1]$. Vì vậy tổng số lượt chia sẻ lại trong tương lai $\sum_{k=1}^{\infty} Z_k$ có kỳ vọng vô hạn và Số lượt chia sẻ lại cuối cùng không thể được dự đoán một cách đáng tin cậy. □

Lưu ý rằng Dự luật 4.1 giả định rằng tính lây nhiễm sau khi vẫn còn không đổi trong tương lai ($P_{NS} = P_{NS}$ vì $NS \geq NS$), điều này có thể không thực tế đối với một số dòng thông tin. Chúng tôi sửa lỗi này bằng cách thay đổi công thức dự đoán trong Eq. (7) bằng cách thêm hai hằng số tỷ lệ α_{NS}, γ_{NS} điều chỉnh dự đoán cuối cùng:

$$\hat{NS}_\infty(NS) = NS_{NS} + \alpha_{NS} \frac{P(N_{NS} - N_{e\,NS})}{1 - \gamma_{NS} P_{NS} n_*}, \quad 0 < \alpha_{NS}, \gamma_{NS} < 1\,. \quad \text{(số 8)}$$

Chúng tôi giới thiệu các yếu tố điều chỉnh này dựa trên trực giác sau: ý tưởng. Chúng tôi mong đợi α_{NS} giảm dần theo thời gian NS vì vậy, nó giảm tỷ lệ lây nhiễm ước tính trong tương lai, điều này giải thích cho bài đăng trở nên cũ kỹ và lỗi thời. Tương tự, γ_{NS} giải thích cho sự trùng lặp trong các vùng lân cận của những người theo dõi những người đăng lại. Theo thời gian như bài lan rộng hơn trong mạng, chúng tôi mong đợi γ_{NS} để tăng lên khi có nhiều nút được hiển thị nhiều lần, có nghĩa là tỷ lệ đến các nút mới (các nút chưa được khai thác trước đó) giảm dần theo thời gian.

Chúng tôi sử dụng các giá trị giống nhau của α_{NS} và γ_{NS} cho tất cả các bài đăng nhưng cho phép chúng thay đổi theo thời gian. Các giá trị của α_{NS} và γ_{NS} được chọn để giảm thiểu Lỗi tỷ lệ phần trăm tuyệt đối trung bình (tham khảo Phần 5.4 để biết định nghĩa) trên tập dữ liệu huấn luyện. Như được mô tả trong Phần 5.2, chúng tôi thấy α_{NS} quan trọng hơn γ_{NS} trong thực tế.

4.3 Thuật toán SEISMIC

Cuối cùng, chúng tôi tập hợp tất cả các thành phần được mô tả cho đến nay và tổng hợp chúng trong SEISMIC thuật toán. Chữ SEISMIC thuật toán để dự đoán $NS_\infty(NS)$ được mô tả trong Thuật toán 1, sử dụng thuật toán cho máy tính P_{NS} (Thuật toán 2) như một chương trình con. Các thuật toán này dựa trên Eqs. (5) và (8). Chúng tôi giả định các thông số $K_{NS}(NS)$, α_{NS}, γ_{NS}, n_* được tặng *tiên nghiệm* hoặc ước tính từ dữ liệu.

Độ phức tạp tính toán của NSEISMIC. Đối với bất kỳ sự lựa chọn nào của $\varphi(NS)$ và $K_{NS}(NS)$, chi phí tính toán của SEISMIC Là $O(NS_{NS})$ cho cả hai tính toán P_{NS} và dự đoán $NS_\infty(NS)$. Của đồng∫urse, thời gian tính toán thực tế phụ thuộc nhiều vào việc tích hợp $\int_{NS_{tôi}}^{NS} K(t-s)\varphi(s-t)\, ds$ $_{tôi}$ và $\int_{NS_{tôi}}^{NS} \varphi(s - t_{tôi})\, ds$. Tuy nhiên, nói quá $_{tôi}$ đồng chi phí đầu ra của NSEISMIC Là *tuyến tính* trong số lượng lượt chia sẻ lại được quan sát NS_{NS} của một bài nhất định theo thời gian NS.

Độ phức tạp thời gian tuyến tính một phần cũng do hình dạng của nhân bộ nhớ của chúng ta. Trong Phần 5.2, chúng ta sẽ ước tính hạt nhân bộ nhớ $\varphi(NS)$ cho Twitter có biểu mẫu sau (đối với một số $s > 0$):

$$\varphi(NS) = \begin{cases} NS & \text{nếu như } 0 < NS \leq NS_0, \\ NS(NS_0)^{-(1+\theta)} & \text{nếu như } s > s_0. \end{cases} \quad (9)$$

Thuật toán 2 Tính toán khả năng lây nhiễm theo thời gian thực $P(NS)$

Mục đích: Đối với một bài nhất định w, tính toán khả năng lây nhiễm P_{NS} với thông tin về w trước thời gian NS

Đầu vào: Đăng thông tin chia sẻ lại: $NS_{tôi}$ và $n_{tôi}$ vì $tôi = 0,\ldots, NS_{NS}$.

Thuật toán:

$NS_{NS} = 0$, $N_{eNS} = 0$

vì $tôi = 0,\ldots, NS_{NS}$ làm

$NS_{NS} += K_{NS}(t - t_{tôi})$

kết thúc cho

vì $tôi = 0,\ldots,\int NS_{NS}$ làm

$$N_{eNS} += n_{tôi} \int_{NS_{tôi}}^{NS} K_{NS}(t - s)\varphi(s - t_{tôi})ds \quad \text{(Phần 4.1)}$$

kết thúc cho

$P_{NS} = NS_{NS}/N_{eNS}$

Giao: P_{NS}

Điều này có nghĩa là với nhân bộ nhớ $\varphi(NS)$ trong Eq. (9) và tri-nhân trọng số góc $K_{NS}(NS)$ trong Eq. (6), cả hai tích phân đều có thể được đánh giá ở dạng đóng vì chúng là đa thức từng mảnh (đa thức với số mũ có thể không phải là số nguyên), làm giảm đáng kể chi phí tính toán của SEISMIC.

5. THÍ NGHIỆM

Trong phần này, chúng tôi mô tả tập dữ liệu Twitter, quy trình ước tính tham số của chúng tôi và so sánh hiệu suất của SEISMIC đến các phương pháp tiếp cận hiện đại.

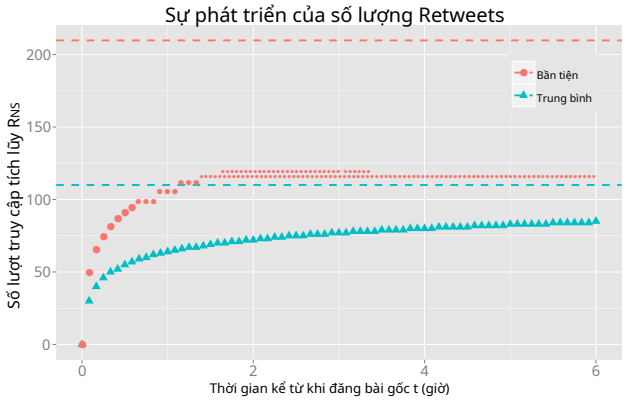
5.1 Mô tả dữ liệu và xử lý dữ liệu

Dữ liệu của chúng tôi là tập hợp đầy đủ hơn 3,2 tỷ lượt tweet và retweet trên Twitter từ ngày 7 tháng 10 đến ngày 7 tháng 11 năm 2011. Đối với mỗi lượt tweet lại, tập dữ liệu bao gồm id tweet, thời gian đăng, thời gian tweet lại và số lượng người theo dõi người đăng / người đăng lại. . Lưu ý, tập dữ liệu thiếu thông tin mạng Twitter. Phần thông tin mạng duy nhất có sẵn cho chúng tôi là số lượng người theo dõi một nút.

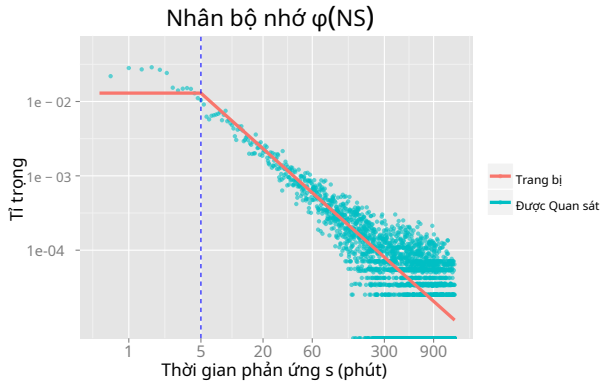
Chúng tôi tập trung vào một tập hợp con các tweet phổ biến hợp lý với ít nhất 50 lượt tweet lại, để mô hình của chúng tôi cho phép dự đoán ngay khi có đủ số lượt tweet lại. Lưu ý rằng nếu nhiều người dùng Twitter đăng cùng một tweet một cách độc lập, sau đó được tweet lại, thì mỗi bài đăng ban đầu sẽ tạo ra một tầng độc lập của riêng nó. Tổng cộng có 166.076 tweet đáp ứng tiêu chí này trong 15 ngày đầu tiên. Chúng tôi hình thành nhóm đào tạo bằng cách sử dụng các tweet trong 7 ngày đầu tiên và nhóm thử nghiệm sử dụng các tweet trong 8 ngày tiếp theo. Chúng tôi sử dụng 14 ngày còn lại để các tầng retweet diễn ra và phát triển. Đối với một đợt retweet cụ thể, chúng tôi nhận được tất cả các lượt retweet đã đăng trong vòng 14 ngày kể từ thời điểm đăng bài gốc, $I E$, chúng tôi chấp thuận-
bạn NS_∞ qua $NS_{14\text{ ngày}}$. Chúng tôi ước tính các thông số $\varphi(NS)$, α_{NS}, γ_{NS} và n_* với bộ đào tạo và đánh giá hiệu suất của công cụ ước tính NS_∞ trên bộ thử nghiệm. Đối với các tweet trong tập huấn luyện của chúng tôi, $NS_{14\text{ ngày}}$ có giá trị trung bình là 209,8 và trung vị là 110. Sự phát triển theo thời gian của giá trị trung bình và trung bình của NS_{NS} cũng được thể hiện trong Hình 3.

5.2 Ước tính tham số SEISMIC

Đầu tiên, chúng tôi mô tả cách điều chỉnh hạt nhân bộ nhớ $\varphi(NS)$ (Mục 3.1). Chúng tôi cẩn thận chọn 15 tweet trong tập huấn luyện và sử dụng phân bổ tất cả thời gian retweet của chúng làm $\varphi(NS)$ (Hình 4). Biểu đồ của 15 chuỗi thời gian retweet đều hiển thị hình dạng rõ ràng của phân rã dưới tới hạn. Hơn nữa, tất cả các áp phích gốc đều có số lượng người theo dõi áp đảo. Do đó, hầu hết các lượt retweet, nếu không phải là tất cả, đều đến từ những người theo dõi ngay người đăng gốc. Do đó, phân bố thời gian phản ứng của con người có thể gần đúng với thời gian retweet của



Hình 3: Sự hội tụ của giá trị trung bình và phương tiện tích lũy số lượt retweet NS_{NS} như một hàm của thời gian. Các đường ngang có nghĩa là phản hồi với số lượt retweet cuối cùng trung bình và trung bình $NS_{14\text{ ngày}}$. Trung bình, một tweet nhận được 75% lượt retweet trong 6 giờ đầu tiên.



Hình 4: Phân bố thời gian phản ứng và nhân bộ nhớ ước tính $\varphi(NS)$. Thời gian phản ứng được vẽ trên trục logarit, do đó xu hướng tuyến tính gợi ý một quy luật lũy thừa.

15 tweet này. Ước tính của $\varphi(NS)$ có thể được cải thiện hơn nữa nếu cấu trúc mạng có sẵn.

Sự phân bố thời gian phản ứng quan sát được (Hình 4) gợi ý một dạng của Eq. (9) đối với nhân bộ nhớ: không đổi trong 5 phút đầu tiên, sau đó là sự phân rã theo luật lũy thừa. Sau khi thiết lập hằng số khoảng thời gian NS_0 đến 5 phút, chúng tôi ước tính tham số phân rã luật lũy thừa $\theta = 0{,}242$ với d tích lũy miễn phíJhàm istribution (ccdf) và đã chọn $NS = 6.27 \times 10^{-4}$ để làm cho $\int_0^\infty \varphi(NS)ds = 1$.
Nhân bộ nhớ là một tham số toàn mạng và chỉ cần được ước lượng một lần. Hạt nhân bộ nhớ vừa vặn được vẽ trong Hình 4.

Cuối cùng, chúng tôi nhận xét ngắn gọn về các yếu tố hiệu chỉnh α_{NS} và γ_{NS} được giới thiệu trong Eq. (7). Chúng tôi sử dụng các giá trị giống nhau của α_{NS} và γ_{NS} cho tất cả các tweet. Thông báo rằng γ_{NS} và n_* chỉ ảnh hưởng đến các dự đoán thông qua sản phẩm của họ $\gamma_{NS}n_*$. Nhìn chung, chúng tôi nhận thấy giá trị của $\gamma_{NS}n_*$ có rất ít ảnh hưởng đến hiệu suất của thuật toán của chúng tôi. Trong các thử nghiệm của chúng tôi, chúng tôi đơn giản là thiết lập $\gamma_{NS}n_* = 20$ cho tất cả NS. Chúng tôi chọn giá trị của α_{NS} sao cho nó giảm thiểu lỗi phần trăm tuyệt đối trung bình đào tạo (Mục 5.4). Chúng tôi báo cáo giá trị của α_{NS} trong Bảng 2. α_{NS} có giá trị đặc biệt nhỏ ở $NS = 5$ phút, có thể là kết quả của đánh giá quá cao P_{NS}, khi nhân tam giác vẫn chưa di chuyển khỏi thời kỳ đầu không ổn định. Sau đó α_{NS} bắt đầu một

thời gian (phút)	5	10	15	20	30
α	0,389	0,803	0,772	0,709	0,680
thời gian (phút)	60	120	180	240	360
α	0,562	0,454	0,378	0,352	0,326

Bảng 2: Giá trị α_{NS} được sử dụng trong Thuật toán 1.

phân rã chậm và nhất quán là do thông tin ngày càng trở nên cũ kỹ và lỗi thời theo thời gian.

Với tất cả các tham số ước tính trong SEISMIC, chúng tôi đã sẵn sàng áp dụng nó (Thuật toán 1 và 2) vào tập dữ liệu Twitter. Đối với một tweet nhất định w và cứ cách 5 phút một lần NS, chúng tôi đưa ra ước tính của mình $NS_\infty(t, w)$ tổng số lượt retweet cuối cùng của tweet $NS_\infty(w)$.

5.3 Cơ sở để so sánh

Chúng tôi xem xét bốn phương pháp dự đoán khác nhau để so sánh. Hai phần đầu là dựa trên hồi quy và hai phần tiếp theo là dựa trên quy trình điểm.

- Hồi quy tuyến tính (LR) [31]: Mô hình có thể được định nghĩa là

  khúc gỗ $NS_\infty = \alpha_{NS}$ + khúc gỗ $NS_{NS} + \varepsilon$,

  ở đâu ε biểu thị tiếng ồn Gaussian. Đây cũng là công cụ ước lượng đường cơ sở thứ hai được sử dụng trong [34]. Lưu ý rằng tất cả các tweet nhận được cùng một hằng số nhân trong một thời gian nhất định.
- Hồi quy tuyến tính với mức độ (LR-D) [31]: Mô hình này có thể được viết là

  khúc gỗ $NS_\infty = \alpha_{NS} + \beta_{1,NS}$ khúc gỗ $NS_{NS} + \beta_{2,NS}$ khúc gỗ $n_{NS} + \beta_{3,NS}$ khúc gỗ $n_0 + \varepsilon$

  ở đâu ε biểu thị, như trước đây, tiếng ồn Gaussian. LR-D là hơn linh hoạt hơn LR, vì nó cho phép khúc gỗ NS_{NS} có độ dốc không bằng 1 và sử dụng các tính năng bổ sung.
- Mô hình Poisson (DPM) [2, 7]: Nó mô hình hóa thời gian retweet $\{NS_k\}$ như một quá trình điểm với tỷ lệ

  $$\lambda_{NS} = \lambda_{NS_{\text{đỉnh cao}}}(t - t_{\text{đỉnh cao}})^y$$

  ở đâu $NS_{\text{đỉnh cao}} = \text{argmax}_{s<t} \lambda_{NS}$. Tham số luật lũy thừa y được ước tính riêng cho mỗi tweet. Để tùy chỉnh mô hình, chúng tôi bỏ thời gian retweet vào $NS \nexists$ Các khoảng thời gian 5 phút. Ghi chú rằng khi $y > -1$, tích phân $\int_{NS_{\text{đỉnh}}+NS}^{\infty} \lambda_{NS} dt$ là vô hạn. Trong những trường hợp như vậy, chúng tôi di chuyển $NS_{\text{đỉnh cao}}$ chuyển tiếp đến thùng tối đa thứ hai.
- Mô hình Poisson được gia cố (RPM) [12]: Phương pháp tiếp cận hiện đại được công bố gần đây này mô hình hóa tỷ lệ chia sẻ lại là

  $$\lambda_{NS} = cf_1(NS)NS_\alpha(NS_{NS})$$

  tham số ở đâu NS đo lường mức độ hấp dẫn của Hiền nhân, $NS_1(NS) \propto NS^{-\gamma}(\gamma > 0)$ mô hình hóa hiệu ứng lão hóa và $NS_\alpha(NS_{NS})$ $(\alpha > 0)$ là chức năng củng cố mô tả hiện tượng "giàu càng giàu". Đưa ra một tweet cụ thể, những thông số c, γ, α được tìm thấy bằng cách tối đa hóa hàm khả năng, trong đó các giá trị tối ưu được chiếu vào các tập khả thi của chúng bất cứ khi nào chúng nằm ngoài phạm vi.

5,4 Các chỉ số đánh giá

Đối với một tweet cụ thể, giả sử rằng dự đoán cho NS_∞ ở thời điểm NS được ký hiệu bởi $NS_\infty(NS)$. Chúng tôi sử dụng các số liệu đánh giá sau trong thử nghiệm của mình:

- Lỗi tỷ lệ phần trăm tuyệt đối (APE): Đối với một tweet nhất định w và thời gian dự đoán NS, chỉ số APE được định nghĩa là,

  $$\text{APE}(w, t) = \frac{|\hat{R}_\infty(w, t) - NS_\infty(w)|}{NS_\infty(w)}.$$

  Khi chỉ số APE được sử dụng cho mục đích đánh giá, các lượng tử APE khác nhau trên các tweet (tất cả đều có thể w) trong tập dữ liệu thử nghiệm sẽ được báo cáo tại mỗi thời điểm NS.
- Kendall-τ Tương quan thứ hạng: Đây là thước đo tương quan thứ hạng [19], tính toán mối tương quan giữa cấp bậc của $NS_\infty(NS)$ và NS_∞ cho tất cả các tweet thử nghiệm. Số liệu này thường mạnh mẽ hơn mối tương quan của Pearson về các giá trị của $NS_\infty(NS)$ và NS_∞. Giá trị tương quan thứ hạng cao có nghĩa là tổng số lượt retweet được dự đoán và cuối cùng có tương quan chặt chẽ với nhau.
- Phạm vi Tweet đột phá: Chúng tôi tạo ra một danh sách sự thật cơ bản về đứng đầu-k tweet có số lượt retweet cuối cùng cao nhất. Chúng tôi gọi những tweet này là tweet "đột phá". Sử dụng mô hình của chúng tôi, chúng tôi cũng có thể tạo ra một k danh sách dựa trên số lượt retweet cuối cùng được dự đoán. Chúng tôi đánh giá các phương pháp bằng cách định lượng mức độ tốt nhất k danh sách bao gồm top-sự thật cơ bản-k danh sách. Chúng tôi cung cấp thêm chi tiết trong Phần 5.5.3.

5.5 Kết quả thực nghiệm

Trong phần này, chúng tôi đánh giá hiệu suất của SEISMIC và bốn đối thủ được mô tả trong Phần 5.3. Tất cả các phương pháp bắt đầu đưa ra dự đoán ngay sau khi một tweet nhất định được tweet lại 50 lần.

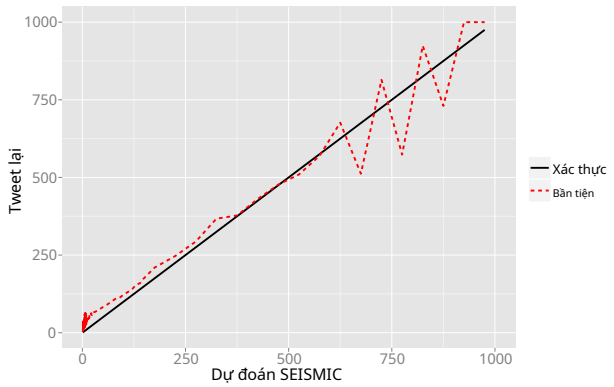
5.5.1 Xác nhận mô hình SEISMIC

Đầu tiên, chúng tôi xác thực theo kinh nghiệm SEISMIC. Trong Đề xuất 4.1, chúng tôi có được công thức cho số lượt retweet cuối cùng dự kiến của thông số lây nhiễm P_{NS}. Mục tiêu của chúng tôi ở đây là cho thấy rằng Đề xuất 4.1 cung cấp một ước tính không thiên vị về lượt retweet thực sự cuối cùng đếm. Chúng tôi tiến hành như sau. Chúng tôi sử dụng SEISMIC để đưa ra dự đoán sau khi quan sát mỗi tweet trong 1 giờ và sau đó lập biểu đồ dự đoán so với số lượng tweet cuối cùng thực sự. Nếu SEISMIC đưa ra một ước tính không thiên vị, sau đó chúng tôi mong đợi một đường cong chéo $y = NS$, đó là, dự đoán được mong đợi NS_∞ phù hợp với mong đợi thực sự NS_∞. Hình 5 cho thấy mức trung bình thực nghiệm gần như hoàn hảo cides với SEISMICdự đoán của. Điều này cho thấy rằng SEISMIC công cụ ước tính trong Eq. (7) là không thiên vị và chúng ta có thể sử dụng nó một cách an toàn để dự đoán số lượt retweet cuối cùng dự kiến. Tuy nhiên, như đã đề cập trước đó, trong thực tế, người ta thường muốn thu nhỏ dự đoán để ổn định công cụ ước tính và đạt được hiệu suất tổng thể tốt hơn. Do đó, chúng tôi sử dụng công thức dự đoán đã hiệu chuẩn Eq. (8) cho phần còn lại của các thí nghiệm.

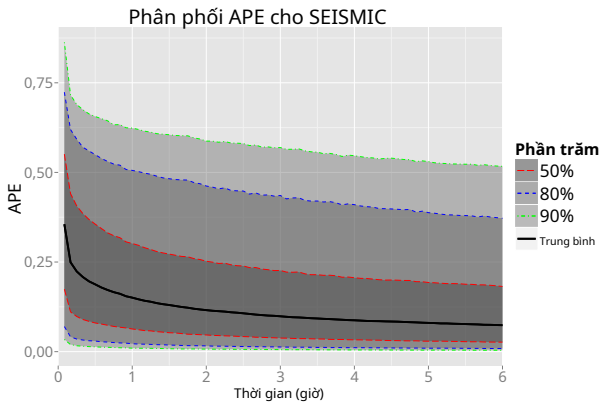
5.5.2 Dự đoán số lượt retweet cuối cùng

Chúng tôi chạy S của chúng tôiEISMIC cho mỗi tweet và tính toán phần trăm lỗi tuyệt đối (APE) như một hàm của thời gian. Chúng tôi vẽ biểu đồ lượng tử phân bố APE của SEISMIC trong Hình 6. Sau khi quan sát dòng thác trong 10 phút (NS =10 phút), phần trăm thứ 95, 75 và 50 của APE lần lượt nhỏ hơn 71%, 44% và 25%. Điều này có nghĩa là sau 10 phút, sai số trung bình là dưới 25% đối với 50% số tweet và dưới 71% đối với 95% số tweet. Sau 1 giờ, lỗi thậm chí còn thấp hơn — APE cho 95%, 75% và 50% tweet giảm xuống lần lượt là 62%, 30% và 15%.

Phương pháp đề xuất, SEISMIC, thể hiện sự cải thiện rõ ràng so với các đường cơ sở và hiện đại như được thể hiện trong Hình 7 và 8. Bảng bên trái của Hình 7 và 8 cho thấy APE trung bình của các phương pháp khác nhau theo thời gian khi ngày càng có nhiều dòng retweet được tiết lộ. Đường cơ sở LR và LR-D có hiệu suất rất giống nhau, cho thấy các tính năng bổ sung được sử dụng bởi LR-D không nhiều thông tin. DPM hoạt động kém trong toàn bộ thời gian tồn tại của tweet, trong khi cách tiếp cận quy trình điểm khác RPM kém hơn LR và LR-D trong giai đoạn đầu nhưng trở nên tốt hơn sau khoảng 2 giờ. Nói chung, xét về điểm APE trung bình SEISMIC nói về



Hình 5: Số lượt retweet cuối cùng được dự đoán theo đúng số lượt retweet dựa trên sự thật, điều này cho thấy NSEISMIC cung cấp ước tính không thiên vị về số lượt retweet cuối cùng. Đường cong màu đỏ gạch ngang có được bằng cách sắp xếp các tweet theo dự đoán và sau đó tính toán số lượt retweet trung bình trong mỗi thùng.



Hình 6: Sai số phần trăm tuyệt đối (APE) của NSEISMIC trên bộ thử nghiệm. Chúng tôi vẽ biểu đồ phần trăm trung bình và trung bình thứ 50, 80, 90 của phân phối APE trên các tweet.

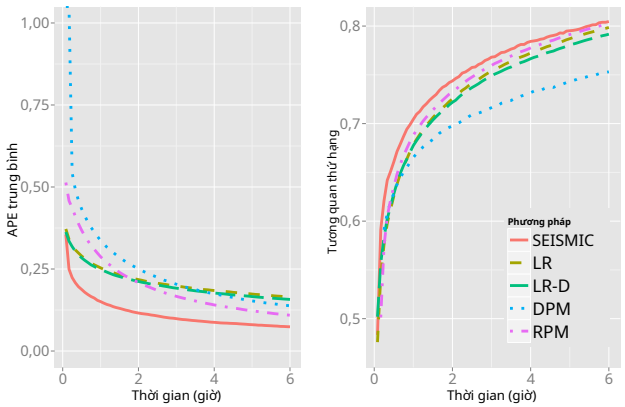
30% chính xác hơn tất cả các đối thủ cạnh tranh trong toàn bộ thời gian hoạt động của twee.

Tương tự, các bảng bên phải của Hình 7 và 8 cho thấy Kendall-τ tương quan thứ hạng giữa xếp hạng dự đoán của các tweet được tweet lại nhiều nhất và xếp hạng trung thực của các tweet. Một lần nữa SEISMIC Là cho tôi uch xếp hạng chính xác hơn các phương pháp khác.

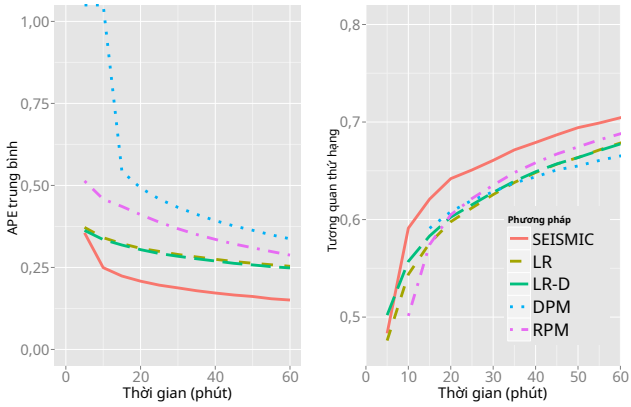
5.5.3 *Xác định các tweet đột phá*

Chúng ta có thể xác định một tweet đột phá trước khi nó nhận được hầu hết các tweet của nó không? Câu hỏi này nảy sinh từ các ứng dụng khác nhau như dự báo xu hướng hoặc phát hiện tin đồn. Mục tiêu của nhiệm vụ dự đoán này là xác định càng sớm càng tốt các tweet "đột phá", có số lượt retweet cuối cùng cao nhất. Chúng tôi định lượng hiệu suất của các mô hình khác nhau trong việc phát hiện các tweet đột phá bằng cách sử dụng các dự đoán của mô hình về số lượt tweet cuối cùng của các tweet.

Đầu tiên, chúng tôi hình thành một tập hợp sự thật nền tảng L^{NS} có kích thước NS. Bộ $L^{\star}_{NS}$ chứa hàng đầu-NS tweet có số lượt retweet cuối cùng cao nhất. Sau đó, với mỗi phương pháp dự đoán, chúng tôi tạo ra một chuỗi kích thước



Hình 7: Lỗi tỷ lệ phần trăm tuyệt đối trung bình (APE) và Tương quan thứ hạng của Kendall về NSEISMIC và các đường cơ sở như một hàm của thời gian. NSEISMIC luôn mang lại hiệu quả tốt nhất.



Hình 8: Zo om-in của Figur e 7: M edian APE và Rank Corre-Lation trong 60 phút đầu tiên sau khi tweet được đăng. NSEIS-MIC hoạt động đặc biệt tốt so với các đường cơ sở ban đầu trong thời gian tồn tại của tweet.

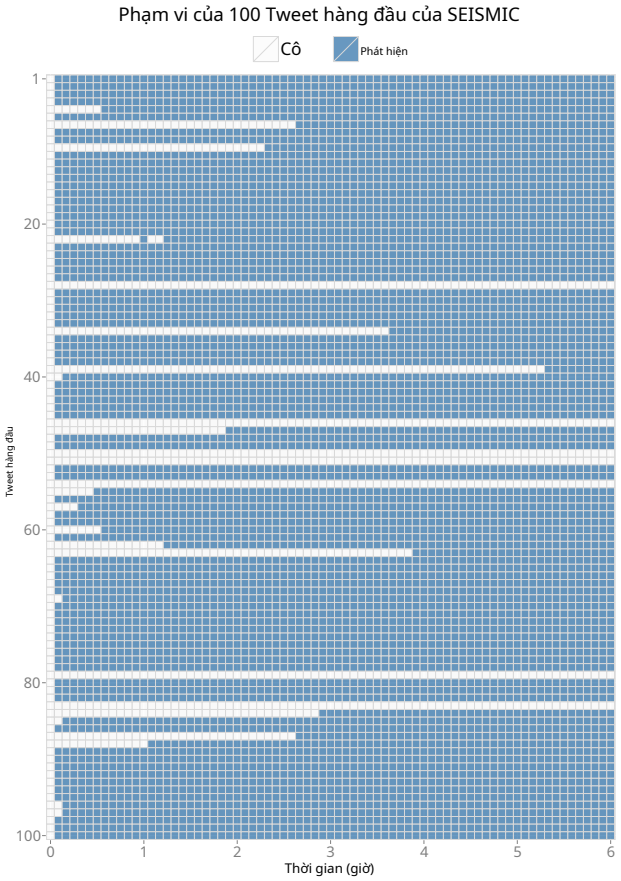
NS danh sách, $\hat{L}^{NS}(NS)$. Tại mỗi thời điểm NS danh sách $\hat{L}^{NS}(NS)$ chứa trên cùng-NS tweet có số lượt retweet được dự đoán cao nhất tại thời điểm NS.

Như được mô tả trong Phần 5.4, sau đó chúng tôi so sánh từng $\hat{L}^{NS}(NS)$ với $L^{\star}_{NS}$ và tính toán *Phạm vi Tweet đột phá*, được xác định như tỷ lệ tweet trong $L^{\star}_{NS}$ được bao phủ bởi $\hat{L}^{NS}(NS)$.

Hình 9 cho thấy hiệu suất của SEISMIC trong việc phát hiện top 100 hầu hết các tweet được tweet lại ($L^{\star}_{100}$) như một hàm của thời gian. NSEISMIC Là có thể bao gồm 82 tweet trong 1 giờ đầu tiên và 93 tweet trong lần đầu tiên 6 giờ.

Tweet được tweet lại nhiều thứ năm trong âm mưu này thực sự là tweet mà chúng tôi đã hiển thị trước đó trong Hình 1. Chúng tôi quan sát thấy rằng SEISMIC phát hiện tweet này 30 phút sau khi nó được đăng, trong khi cả LR và LR-D đều mất hơn một giờ. DPM không phát hiện ra sự đột phá này trong 6 giờ đầu tiên (các biểu đồ không hiển thị ngắn gọn).

Để so sánh SEISMIC với các phương pháp khác, chúng tôi giữ cho kích thước của các danh sách được dự đoán là $NS = 500$ và sử dụng danh sách mục tiêu lớn hơn $L^{\star}_{500}$, đó là một nhiệm vụ khó khăn hơn việc tìm kiếm $L^{\star}_{100}$. Hình 10 com-so sánh mức độ phù hợp của các phương pháp khác nhau với tỷ lệ lượt retweet được nhìn thấy. Sau khi thấy 20% lượt retweet, SEISMIC chiếm 65% danh sách rút gọn, trong khi cả LR-D và LR đều chỉ chiếm 50%. Trong



Hình 9: Phạm vi của 100 tweet được tweet lại nhiều nhất. Mỗi hàng đại diện cho một tweet. Các khối màu trắng cho biết rằng một tweet nhất định không bị che bởiNSEISMICdanh sách dự đoán 500 tweet hàng đầu tại thời điểm *NS*và màu xanh lam cho biết phạm vi bảo hiểm thành công.

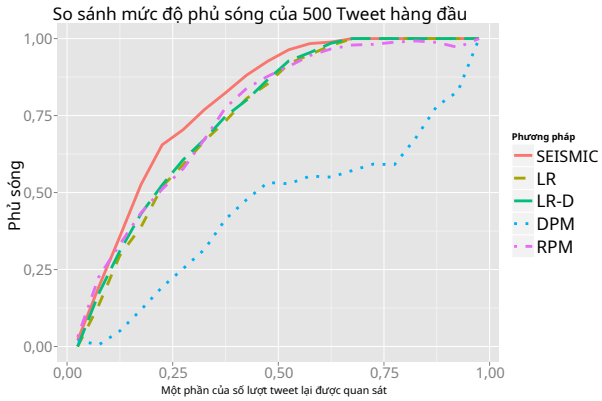
nói chung, mô hình Poisson động không cung cấp các dự đoán chính xác và nhận dạng đột phá.

Nhìn chung, SEISMIC cho phép phát hiện hiệu quả các tweet đột phá. Ví dụ: sau khi thấy khoảng 25% tổng số lượt retweet của một tweet nhất định (nói cách khác, sau khi quan sát một tweet trong khoảng 5 phút), SEISMIC có thể xác định 60% trong số 100 tweet hàng đầu theo số lượt retweet cuối cùng.

5.6 Thảo luận về độ bền của mô hình

NSEISMIC thể hiện tính mạnh mẽ hơn so với hai phương pháp dựa trên quy trình điểm khác - DPM và RPM. Trong khi SEISMIC không thể đưa ra dự đoán cho các tweet ở trạng thái siêu tới hạn, DPM và RPM không thể đưa ra dự đoán khi thông số phân rã nằm ngoài tập hợp khả thi (y <-1 cho DPM và $y < 0$ hoặc $\alpha < 0$ cho RPM). Ví dụ, trong Hình 1, SEISMIC mô tả tweet là siêu tới hạn trong 70 phút đầu tiên, DPM không đưa ra dự đoán trong 6 giờ đầu tiên và RPM chỉ có thể đưa ra dự đoán từ 30 đến 80 phút.

Nhìn chung, chúng tôi thấy rằng các tweet chỉ ở chế độ siêu tới hạn trong một thời gian rất ngắn và SEISMIC có thể đưa ra dự đoán cho hầu hết các tweet trong hầu hết thời gian. Chúng tôi thấy rằng trung bình, SEISMIC không thể đưa ra dự đoán cho 1,80% số tweet sau khi quan sát chúng trong 15 phút. Nói cách khác, sau 15 phút, 1,80% số tweet vẫn ở chế độ siêu tới hạn



Hình 10: Mức độ phủ sóng của 500 tweet hàng đầu (↓) bởi va rious meth-ods. NSEISMIC cho thấy sự cải thiện rõ ràng so với tất cả các phương pháp sau khi quan sát thấy khoảng 10% lượt retweet. Tất cả các phương pháp ngoại trừ DPM đều đạt được độ phủ hoàn hảo sau khi quan sát thấy 65% lượt retweet.

(trên tất cả các tweet có ít nhất 50 lượt retweet). Con số này giảm xuống 1,29% (0,67%) sau 1 giờ (6 giờ). Để so sánh, chúng tôi cũng lưu ý rằng các phương pháp khác không thể đưa ra dự đoán cho một phần lớn hơn nhiều tweet: DPM không đưa ra dự đoán cho 6,77%, 5,79% và 1,45% và RPM không thành công cho 3,45%, 5,69% và 15,43% số tweet sau 15 phút, 1 giờ và 6 giờ. Của chúng tôiEISMIC phương pháp này cũng nhanh hơn đáng kể so với mô hình RPM [28], mô hình này yêu cầu giải một bài toán tối ưu hóa phi tuyến mỗi khi nó dự đoán. Trong quá trình triển khai của chúng tôi, thời gian chạy trung bình trên mỗi tweet để dự đoán cứ sau 5 phút trong 6 giờ là 0,02 giây đối với S EISMIC và 3,6 giây cho RPM. Thời gian chạy được báo cáo bao gồm cả học tham số và dự đoán.

6. KẾT LUẬN VÀ TƯƠNG LAI

Trong bài báo này, chúng tôi đề xuất SEISMIC, một khuôn khổ linh hoạt để mô hình hóa các tầng thông tin và dự đoán kích thước cuối cùng của một tầng thông tin. Những đóng góp của chúng tôi như sau:

- Chúng tôi lập mô hình các tầng thông tin dưới dạng các quá trình điểm tự thú vị trên cây Galton-Watson. Cách tiếp cận của chúng tôi cung cấp một khung lý thuyết để giải thích các mô hình thời gian của các tầng thông tin.
- NSEISMIC vừa có thể mở rộng vừa chính xác. Mô hình không yêu cầu kỹ thuật tính năng và chia tỷ lệ tuyến tính với số lượt chia sẻ lại được quan sát của một bài đăng nhất định. Điều này cung cấp một cách để dự đoán thông tin lan truyền cho hàng triệu bài đăng trong một thiết lập thời gian thực trực tuyến.
- NSEISMIC mang lại tính linh hoạt cao hơn cho các nhiệm vụ ước tính và dự đoán vì nó yêu cầu kiến thức tối thiểu về dòng thông tin cũng như cấu trúc mạng bên dưới.

Có rất nhiều địa điểm thú vị cho công việc trong tương lai và mô hình đề xuất của chúng tôi có thể được mở rộng theo nhiều hướng khác nhau. Ví dụ: nếu cấu trúc mạng có sẵn, người ta có thể thay thế mức độ nút $n_{tới}$bởi số lượng người theo dõi mới được tiếp xúc. Nếu các tính năng dựa trên nội dung hoặc các tính năng của bài đăng gốc có sẵn, người ta có thể phát triển một nội dung dựa trên P_{NS}cho mỗi bài viết. Nếu các tính năng tạm thời như múi giờ của người dùng khả dụng, người ta có thể trực tiếp sử dụng chúng để sửa đổi công cụ ước tính P_{NS}. Theo nghĩa này, mô hình được đề xuất cung cấp một khuôn khổ có thể mở rộng để dự đoán các tầng thông tin.

NSEISMIC là một mô hình từ dưới lên rõ ràng về mặt thống kê và có thể mở rộng của các tầng thông tin cho phép dự đoán kích thước tầng cuối cùng là

thác mở ra trên mạng. Chúng tôi hy vọng rằng khuôn khổ của chúng tôi sẽ hữu ích để phát triển sự hiểu biết phong phú hơn về các hành vi xếp tầng trong mạng trực tuyến, mở đường cho việc quản lý nội dung được chia sẻ tốt hơn.

Dữ liệu và Phần mềm

Chữ SEISMIC phần mềm và tập dữ liệu chúng tôi sử dụng trong Phần 5 có thể được tìm thấy trong http://snap.stanford.edu/seismic/. Gói R của thuật toán của chúng tôi cũng có sẵn trên http: // cran. r-project.org/web/packages/seismic.

Sự nhìn nhận

Các tác giả xin cảm ơn David O. Siegmund vì những góp ý mang tính xây dựng của ông và Austin Benson, Bhaswar B. Bhattacharya, Joshua Loftus vì những nhận xét hữu ích của họ. Nghiên cứu này đã được hỗ trợ một phần bởi NSF IIS-1016909, CNS-1010921, IIS-1149837, IIS-1159679, ARO MURI, DARPA SMISC, DARPA SIMPLEX, Stanford Data Science Initiative, Boeing, Facebook, Volkswagen và Yahoo.

7. NGƯỜI GIỚI THIỆU

[1] https://twitter.com/mottbollomy/status/127001313513967616.

[2] D. Agarwal, B.-C. Chen và P. Elango. Mô hình không gian-thời gian để ước tính tỷ lệ nhấp. Trong *WWW '09*, Năm 2009.

[3] E. Bakshy, JM Hofman, WA Mason và DJ Watts. Mọi người đều là người có ảnh hưởng: định lượng ảnh hưởng trên twitter. Trong *WSDM '11*, 2011.

[4] R. Bandari, S. Asur, và BA Huberman. Nhịp độ tin tức trên mạng xã hội: Dự báo mức độ phổ biến. Trong *ICWSM '12*, trang 26–33, 2012.

[5] A.-L. Barabasi. Nguồn gốc của sự bùng nổ và những cái đuôi nặng nề trong động lực học của con người. *Thiên nhiên*, 435: 207, 2005.

[6] J. Cheng, L. Adamic, PA Dow, JM Kleinberg, và J. Leskovec. Các thác có thể được dự đoán? Trong *WWW '14*, 2014.

[7] R. Crane và D. Sornette. Các lớp năng động mạnh mẽ được tiết lộ bằng cách đo lường chức năng phản ứng của một hệ thống xã hội. *PNAS*, 105 (41), 2008.

[8] H. Daneshmand, M. Gomez-Rodriguez, L. Song, và B. Schölkopf. Ước tính cấu trúc mạng khuếch tán: Điều kiện khôi phục, độ phức tạp của mẫu và thuật toán ngưỡng mềm. Trong *ICML '14*, 2014.

[9] PA Dow, LA Adamic và A. Friggeri. Giải phẫu các thác lớn facebook. Trong *ICWSM '13*, 2013.

[10] N. Du, L. Song, M. Yuan, và AJ Smola. Học mạng ảnh hưởng không đồng nhất. Trong *NIPS '12*, 2012.

[11] R. Durrett. *Xác suất: lý thuyết và ví dụ*. Báo chí đại học Cambridge, 2010.

[12] S. Gao, J. Ma, và Z. Chen. Lập mô hình và dự đoán động lực đăng lại trên nền tảng tiểu blog. Trong *WSDM '15*, 2015.

[13] M. Gomez-Rodriguez, J. Leskovec, D. Balduzzi, và B. Schölkopf. Khám phá cấu trúc và động lực thời gian của truyền thông tin. *Khoa học mạng*, 2: 26–65, 4 năm 2014.

[14] M. Gomez-Rodriguez, J. Leskovec và B. Schölkopf. Mô hình hóa sự truyền bá thông tin với lý thuyết sinh tồn. Trong *ICML '13*, 2013.

[15] M. Gomez-Rodriguez, J. Leskovec và B. Schölkopf. Cấu trúc và Động lực của Đường dẫn Thông tin trong Phương tiện Trực tuyến. Trong *WSDM '13*, 2013.

[16] AG Hawkes. Quang phổ của một số quá trình điểm tự thú vị và thú vị lẫn nhau. *Biometrika*, 58 (1), năm 1971.

[17] L. Hong, O. Dan, và BD Davison. Dự đoán các tin nhắn phổ biến trong twitter. Trong *WWW '11*, 2011.

[18] D. Hunter, P. Smyth, DQ Vu, và AU Asuncion. Mô hình tập trung động cho mạng trích dẫn. Trong *ICML '11*, 2011.

[19] MG Kendall. Một thước đo mới về tương quan thứ hạng. *Biometrika*, trang 81–93, 1938.

[20] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev và A. Kustarev. Dự đoán về kích thước tầng retweet theo thời gian. Trong *CIKM '12*, 2012.

[21] D. Liben-Nowell và J. Kleinberg. Truy tìm luồng thông tin trên quy mô toàn cầu bằng cách sử dụng dữ liệu chuỗi thư trên Internet. *Kỷ yếu của Viện Hàn lâm Khoa học Quốc gia*, 105 (12): 4633–4638, 2008.

[22] Y. Matsubara, Y. Sakurai, BA Prakash, L. Li, và C. Faloutsos. Các mô hình phổ biến thông tin tăng và giảm: Mô hình và hàm ý. Trong *KDD '12*, 2012.

[23] GO Mohler, MB Short, PJ Brantingham, FP Schoenberg và GE Tita. Mô hình quy trình điểm tự thú vị của tội phạm. *Tạp chí của Hiệp hội Thống kê Hoa Kỳ*, 106 (493): 100–108, 2011.

[24] N. Naveed, T. Gottron, J. Kunegis, và AC Alhadi. Tin xấu đi nhanh: Phân tích dựa trên nội dung về mức độ thú vị trên twitter. Trong *ACM WebSci '11*, 2011.

[25] Y. Ogata. Mô hình thống kê cho các lần xuất hiện động đất và phân tích dư cho các quá trình điểm. *Tạp chí của Hiệp hội Thống kê Hoa Kỳ*, 83, 1988.

[26] S. Petrovic, M. Osborne và V. Lavrenko. RT để giành chiến thắng! Dự đoán truyền bá thông điệp trong Twitter. Trong *ICWSM '11*, 2011.

[27] EM Rogers. *Sự lan tỏa của những đổi mới*. Simon và Schuster, Năm 2010.

[28] H.-W. Shen, D. Wang, C. Song, và A.-L. Barabási. Mô hình hóa và dự đoán động lực học phổ biến thông qua các quá trình poisson tăng cường. *arXiv: 1401.0778*, 2014.

[29] DL Snyder và MI Miller. *Các quá trình điểm ngẫu nhiên trong thời gian và không gian*. Springer, 2011.

[30] B. Suh, L. Hong, P. Pirolli, và EH Chi. Bạn muốn được tweet lại? phân tích quy mô lớn về các yếu tố ảnh hưởng đến lượt retweet trong mạng twitter. Trong *SOCIALCOM '10*, Năm 2010.

[31] G. Szabo và BA Huberman. Dự đoán mức độ phổ biến của nội dung trực tuyến. *Thông tin liên lạc của ACM*, 53 (8): 80–88, tháng 8 năm 2010.

[32] J. Yang và J. Leskovec. Các mô hình biến đổi theo thời gian trong phương tiện truyền thông trực tuyến. Trong *WSDM '11*, 2011.

[33] S.-H. Yang và H. Zha. Hỗn hợp các quá trình thú vị lẫn nhau để lan truyền virus. Trong *ICML '13*, 2013.

[34] T. Zaman, E. Fox và E. Bradlow. Một cách tiếp cận của Bayes để dự đoán mức độ phổ biến của các tweet. *Biên niên sử của Thống kê Ứng dụng*, 8 (3): 1583–1611, 2014.

[35] TR Zaman, R. Herbrich, J. Van Gael và D. Stern. Dự đoán thông tin lan truyền trên twitter. Trong *Hội thảo Khoa học xã hội tính toán và Trí tuệ của đám đông, NIPS*, Năm 2010.

[36] K. Zhou, H. Zha, và L. Song. Tìm hiểu tính lây nhiễm xã hội trong các mạng xếp hạng thấp thưa thớt bằng cách sử dụng quy trình hawkes đa chiều. Trong *AISTATS '13*, 2013.