Some simulation models for the epidemic

Nguyễn Tiến Dũng - CTTN Toán Tin - K62

Ngày 24 tháng 12 năm 2020

Lời cảm ơn

Em xin bày tỏ lòng kính trọng và cảm ơn chân thành đến TS. Nguyễn Thị Thanh Huyền, TS. Lê Chí Ngọc – Viện Toán ứng dụng và Tin học – Trường Đại học Bách khoa Hà Nội, người đã định hướng đề tài, tận tình giúp đỡ và hướng dẫn em trong suốt quá trình thực hiện đồ án.

Em xin chân thành cảm ơn các thầy giáo, cô giáo – Viện Toán ứng dụng và Tin học – Trường Đại học Bách khoa Hà Nội, những người đã tận tình truyền đạt các kiến thức cho em trong suốt thời gian em học tập và nghiên cứu tại trường.

Em cũng xin gửi lời cảm ơn tới gia đình đã ủng hộ, động viên em trong suốt quá trình học tập vừa qua.

Cuối cùng, em xin cảm ơn các bạn học cùng lớp CTTN – Toán Tin K62 đã giúp đỡ em trong suốt quá tình học tập và thực hiện đồ án

Trong quá trình nghiên cứu, tìm hiểu và thực nghiệm đồ án chắc chắn không thể tránh khỏi sai sót, em rất mong nhận được sự góp ý của thầy, cô giáo và các bạn để đồ án được hoàn chỉnh hơn.

Em xin trân trọng cảm ơn!

Lời cam đoan

Em xin cam đoan đồ án *Mô phỏng mô hình lan truyền dịch bệnh trên mạng* này là công trình nghiên cứu của em dưới sự hướng dẫn khoa học của TS. Nguyễn Thị Thanh Huyền và TS. Lê Chí Ngọc. Tất cả những tài liệu tham khảo em đã liệt kê rõ ở phần cuối của đồ án. Các nội dung công bố và kết quả trình bày trong báo cáo này là trung thực và không có sự sao chép của người khác. Nếu phát hiện có bất kỳ sự gian lận nào, em xin chịu hoàn toàn trách nhiệm trước hội đồng, cũng như kết quả môn học của mình.

Mục lục

Danh sách hình vẽ

Danh sách bảng

1 Giới thiệu bài toán

1.1 Đặt vấn đề

Thời điểm cuối năm 2019, đầu năm 2020, dịch COVID19 bùng phát với hậu quả để lại ảnh hưởng sâu sắc trên tất cả các lĩnh vực và vẫn chưa có dấu hiệu giảm xuống.

1.2 Mục tiêu

Mô phỏng lại quá trình bùng phát dịch Covid19 tại Đà Nẵng.

Đồ án này sẽ gồm một số nội dung chính như sau:

2 Kiến thức cơ sở

2.1 Đồ thị

Bên cạnh hướng tiếp cận bằng các mô hình dịch bệnh cổ điển, trong báo cáo này, ta sử dụng đồ thị để mô phỏng lại dịch bệnh tại Đà Nẵng diễn ra trong khoảng thời gian từ 23/6/2020 đến 10/8/2020

Definition 2.1. Đồ thị $G = (V, E)$ là một cặp có thứ tự, trong đó:

- V là tập các đỉnh
- E là tập các cạnh

Đồ thị thường được sử dụng trong việc biểu diễn các đối tượng và quan hệ của chúng, ví dụ như quan hệ giữa mọi người trên mạng xã hội, hoặc liên kết giữa các thành phần hóa học,...

Trong báo cáo này, ta sẽ sử dụng đồ thị để mô hình hóa lại quá trình diễn tiến của dịch bệnh theo thời gian.

2.1.1 Các độ đo trung tâm trên đồ thị

Đối với một đồ thị $G = (V, E)$ cho trước, một câu hỏi rất tự nhiên được đưa ra là:

"Những đỉnh v nào trong tập đỉnh V quan trọng nhất G?"

Một khái niệm khác cần được làm rõ cho câu hỏi trên là "thế nào là *quan trọng?*". Quan trọng có thể được xem xét dưới nhiều góc độ khác nhau, và với một định nghĩa cho *quan trọng* cho ta một độ đo trung tâm khác nhau trên đồ thị. Hiện tại có rất nhiều các độ đo trung tâm khác nhau dựa trên các tiêu chí khác nhau mà ta có thể kể đến như:

1. Độ đo theo bậc (Degree Centrality)

2. Độ đo theo khoảng cách (Betweeness Centrality)

3. Độ đo theo trị riêng (Eigenvector Centrality)

4. Pagerank

   ...

Một phát biểu toán học của độ đo được đưa ra bởi định nghĩa 2.2.

Definition 2.2 (Centrality Measure)**.** Độ đo trung tâm[3] là một ánh xạ c

$$c : G(n) \to \mathbb{R}^n$$

trong đó $c_i(g)$ là độ đo của đỉnh i trong đồ thị g, hay nói cách khác, $c_i(g)$ thể hiện cho độ quan trọng của đỉnh i trong đồ thị g.

Trong báo cáo này, pagerank[4] được lựa chọn bởi [4, 6, 5].

2.1.2 Pagerank

Thuật toán pagerank lần đầu tiên được đưa ra bởi Larry Page và Sergei Brin vào năm 1998, tuy nhiên ý tưởng ban đầu có thể được

2.2 Kiểm định phân phối

Với bộ dữ liệu quan sát và một mô hình hiện có, ta mong muốn xác định được mức độ phù hợp của mô hình với dữ liệu hiện có, để từ đó xác định được mức độ phù hợp của mô hình với dữ liệu trong tương lai.

Để xác định được tính phù hợp của một mô hình thống kê với bộ dữ liệu cho trước, năm 1892, K. Pearson [7] đã lần đầu tiên đưa ra khái niệm “Goodness-of-Fit" nhằm giải quyết vấn đề này. Ý tưởng chung của phương pháp tương đối đơn giản, bằng cách so sánh hàm phân phối tích lũy theo lý thuyết (cummulative distribution function) với hàm phân phối thực nghiệm (empirical distribution function).

Hàm phân phối tích lũy lý thuyết của biến ngẫu nhiên X là phân phối xác suất được cho bởi định nghĩa dưới đây:

Definition 2.3 (Cummulative Distribution Function)**.** Hàm phân phối tích lũy F của biến ngẫu nhiên X là xác suất để $X \leq x$:

$$\Pr\{X \leq x\} = F(x), -\infty < x < \infty$$

và hàm phân phối thực nghiệm (empirical distribution function) được cho như dưới đây

Definition 2.4 (Empirical Distribution Function)**.** Hàm edf được định nghĩa trên một mẫu quan sát $X = \{x_1, x_2, ..., x_n\}$ như sau:

$$F_n(x) = \frac{\text{Card}\{\{x_i \mid x_i \in X, x_i < x\}\}}{n}$$

Khi đó, bằng việc so sánh sai khác giữa F và F_n mà ta có thể đưa ra kết luận liệu các quan sát $x_1, x_2, ..., x_n$ có được sinh ra từ phân phối F hay không.

Ta đưa ra 2 giả thuyết sau:

$$H_0 : F_n = F$$
$$H_A : F_n \neq F$$

Giả thuyết H_0 sẽ bị bác bỏ nếu F và F_n sai khác nhau đủ nhiều. Một độ đo độ sai khác giữa $F(.)$ và $F_n(.)$ được đưa ra bởi Cramér và von Mises [2] như sau:

$$\Delta(F, F_n) = \int_{-\infty}^{\infty} [F(x) - F_n(x)]^2 \psi[F(x)] dF(x) \tag{2.1}$$

Một độ đo khác được đưa ra bởi Kolmogorov - Smirnov [1]:

$$K_n = \sup_{-\infty<x<\infty} \sqrt{n} |F_n(x) - F(x)| \sqrt{\psi[F(x)]}$$

Anderson - Darling đưa ra một cách chọn cho hàm trọng số $\psi(t) = \frac{1}{t(1-t)}$, và qua đó ta có kiểm định Anderson - Darling sẽ được trình bày ở dưới đây

2.2.1 Kiểm định Anderson - Darling

Anderson - Darling [1] lựa chọn hàm trọng số $\psi(t) = \frac{1}{t(1-t)}$ và cho ta thống kê sau:

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x)[1 - F(x)]} dF(x) \tag{2.2}$$

Đối với phân phối chuẩn, M. A. Stephen[8] đã chỉ ra rằng trong hầu hết các trường hợp, thống kê A_n^2 cho kết quả tốt hơn về mặt kiểm định, do đó trong báo cáo này, ta sẽ sử dụng kiểm định được đưa ra bởi Anderson - Darling.

2.3 Ước lượng tham số

Để ước lượng các tham số μ, σ ta có thể sử dụng ước lượng hợp lí cực đại (MLE - Maximum Likelihood Estimation) trên tập quan sát $X = \{x_1, x_2, ..., x_n\}$

Giả sử rằng các $X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$ và iid, khi đó, ta có hàm likelihood như sau:

$$L\left(\mu, \sigma\right) = L_n\left(\mu, \sigma, X\right) = f_n\left(X, \mu, \sigma\right) \tag{2.3}$$

Ta cần tìm các giá trị $\hat{\mu}, \hat{\sigma}$ để hàm likelihood L đạt giá trị cực đại. Với giả thiết rằng các X_i là i.i.d, (2.3) có thể được viết lại như sau:

$$\begin{aligned} L\left(\mu, \sigma\right) &= \prod_{i=1}^{n} f\left(x_i, \mu, \sigma\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\} \end{aligned} \tag{2.4}$$

Lấy log vào 2 vế của (2.4) ta có:

$$\begin{aligned} \ln L\left(\mu, \sigma\right) &= \ln\left[\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right\}\right] \\ &= -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \end{aligned} \tag{2.5}$$

2.3.1 Ước lượng μ

Dễ thấy rằng $g\left(\mu, \sigma\right) = \ln L\left(\mu, \sigma\right)$ là hàm lõm (concave) với μ nên ước lượng MLE cho μcó thể thu được bằng cách xác định $\hat{\mu}$ sao cho

$$\left.\frac{\partial \ln L\left(\mu, \sigma\right)}{\partial \mu}\right|_{\hat{\mu}} = 0$$

Ta có:

$$\frac{\partial \ln L\left(\mu, \sigma\right)}{\partial \mu} = \sum_{i=1}^{n} \mu\left(\mu - x_i\right)$$

Thấy rằng $\hat{\mu} = 0$ là một nghiệm tầm thường, ta bỏ qua nghiệm này. Ước lượng còn lại là $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$.

2.3.2 Ước lượng σ

Ta có

$$\begin{aligned} \frac{\partial \ln L\left(\mu, \sigma\right)}{\partial \sigma} &= \frac{-n}{2}\frac{4\sigma}{\sigma^2} + \frac{2}{\sigma^3}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \\ &= \frac{-2n\sigma^2 + 2\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{\sigma^3} \end{aligned} \tag{2.6}$$

Từ (2.6) ta có ước lượng MLE cho σ như sau

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

3 Xử lí dữ liệu

Với dữ liệu về các ca COVID19 trong đợt bùng phát thứ 2, ta quan tâm đến ngày bệnh nhân bắt đầu khởi phát triệu chứng của bệnh. Nói cách khác, tuy một số bệnh nhân không được báo cáo có triệu chứng khởi phát, tuy nhiên việc ngày bệnh nhân đạt đến trạng thái lây nhiễm cao nhất là xác định, và ta cần thiết phải xử lí những thông tin này. Tuy nhiên, trong số 389 bệnh nhân được ghi nhận trong đợt bùng phát dịch lần thứ hai, chỉ có 112 bệnh nhân được báo cáo là có thông tin về ngày khởi phát, do đó ta sẽ cần phải ước lượng ngày khởi phát cho các bệnh nhân còn lại.

Ta sẽ ước lượng thông tin về ngày khởi phát dựa theo ngày tiếp xúc cuối cùng của bệnh nhân. Có 325/389 bệnh nhân có dữ liệu về ngày tiếp xúc cuối cùng với các ca bệnh.

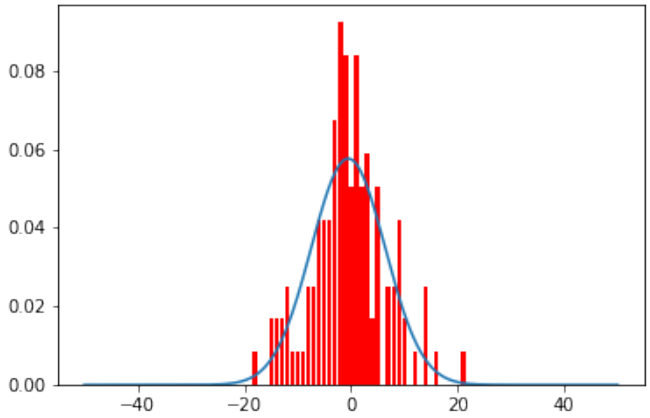
Dưới đây là bảng số liệu về khoảng cách giữa ngày tiếp xúc cuối cùng và ngày khởi phát triệu chứng theo tần suất.

Δ	f
-18	1
-15	2
-14	2
-13	2
-12	3
-11	1
-10	1
-9	1
-8	3
-7	3
-6	5
-5	5
-4	5
-3	8
-2	11

Δ	f
-1	10
0	6
1	10
2	6
3	7
4	2
5	6
7	3
8	3
9	5
10	2
12	1
14	3
16	1
21	1

Bảng 3.1: Bảng tần suất

Minh họa dưới dạng đồ thị của bảng số liệu trên được cho bởi hình vẽ 3.1

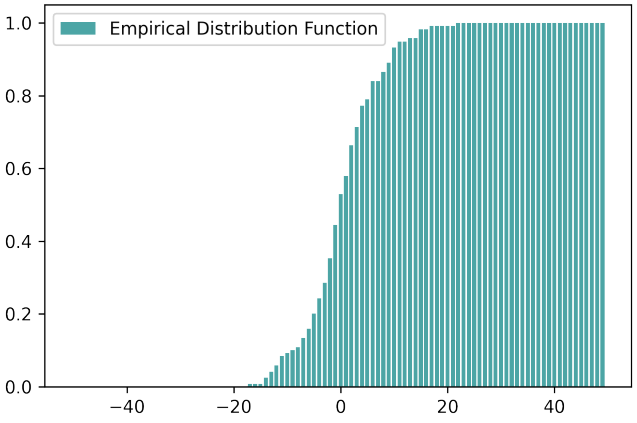


Hình 3.1: Khoảng thời gian giữa ngày tiếp xúc cuối cùng và ngày phát bệnh

Từ đồ thị trên, ta thấy rằng các quan sát Δ có thể được sinh ra từ phân phối chuẩn $\mathcal{N}\left(\mu, \sigma^2\right)$. Để kiểm chứng giả thuyết này, ta sẽ sử dụng kiểm định Anderson - Darling như đã trình bày ở mục 2.2.1

Theo (điền vào đây ref công thức tính thống kê Anderson - Darling), ta có $A = 0.747 < 0.763$ tương ứng với mức ý nghĩa $\alpha = 0.05$. Căn cứ vào kết quả này, ta có thể chấp nhận giả thuyết.

Đồ thị minh họa cho hàm phân phối thực nghiệm cho bởi đồ thị 3



Hình 3.2: Hàm phân phối thực nghiệm của quan sát X

Tính độc lập của các quan sát có thể được kiểm tra nếu có thêm thông tin về thứ tự công bố của các bệnh nhân, do đó ở đây ta bỏ qua biệc kiểm tra tính độc lập của các quan sát mà thay vào đó công nhận thẳng tính độc lập của dữ liệu.

Tính độc lập của các quan sát có thể được ước lượng qua các kiểm định von Neumann hoặc các mô hình tự hồi quy (AR).

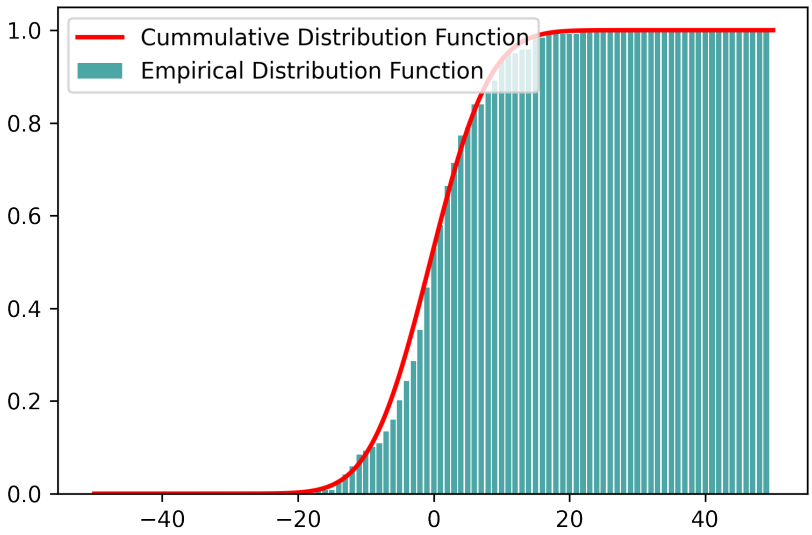
3.1 Trên dữ liệu thực tế

Từ dữ liệu đã cho với 389 bệnh nhân, ta thấy có 112 bệnh nhân có thông tin về ngày khởi phát.

Thông qua kiểm định Anderson - Darling ta thấy rằng có thể chấp nhận được giả thuyết các quan sát tuân theo phân phối chuẩn, và giả thiết rằng các quan sát độc lập lẫn nhau, ta có ước lượng cho phân phối như sau

$$\begin{cases} \mu & = -0.521 \\ \sigma^2 & = 48.370 \end{cases}$$

So sánh tương quan giữa EDF và CDF:



Hình 3.3: Đồ thị tương quan giữa EDF và CDF

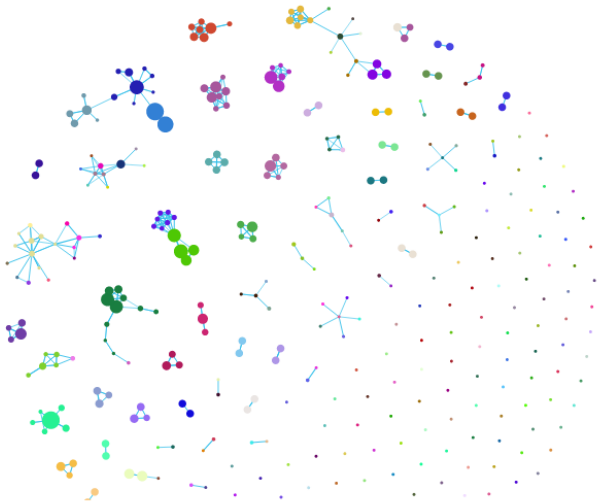
Thực hiện sinh các giá trị ngẫu nhiên từ phân phối chuẩn đưa ra như trên

4 Xây dựng đồ thị

4.1 Xác định cấu trúc đồ thị

Dựa trên dữ liệu đã thu thập được về đợt dịch lần thứ 2 diễn ra tại Đà Nẵng, ta tiến hành xây dựng mạng thể hiện cho dịch bệnh.

Đồ thị được xây dựng bằng cách liên kết các ca bệnh có liên quan với nhau được biểu diễn bởi hình 4.1

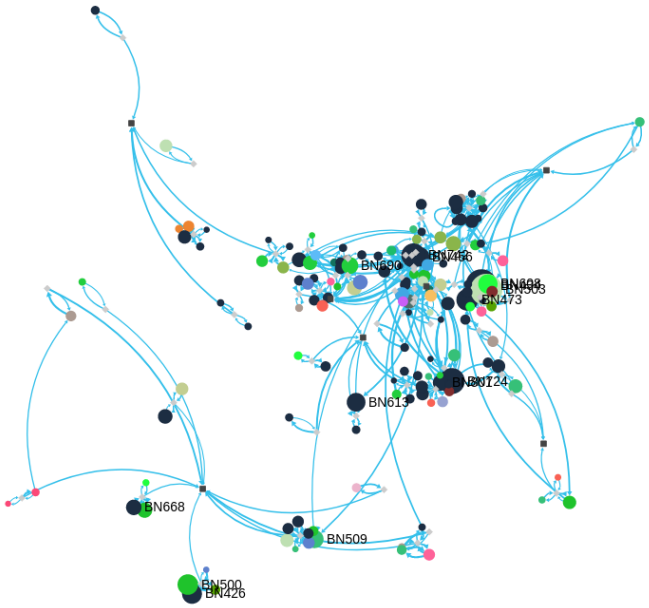


Hình 4.1: Đồ thị không có địa điểm

Nhìn vào đồ thị, ta có thể thấy rằng với việc có tới 277 thành phần liên thông, trong đó có tới một nửa là các thành phần chỉ gồm có một đỉnh. Với đồ thị thưa như trên ta thấy rằng không có nhiều yếu tố có thể khai thác ở đây.

Để giải quyết vấn đề này, ta giả định sự tồn tại sự lây bệnh qua các đồ vật trung gian, như tay nắm cửa, nút bấm thang máy,...cũng như các nguồn lây không xác định. Các nguồn lây không xác định được thể hiện qua các đỉnh biểu diễn cho địa điểm, rằng những bệnh nhân cùng nơi sinh sống có thể lây nhiễm qua nhau.

Hình 4.1 dưới đây minh họa cho đồ thị được xây dựng theo cách trên.



Hình 4.2: Đồ thị sau khi thêm địa điểm

Đồ thị lúc này đã có liên kết tốt hơn giữa các địa điểm cũng như các bệnh nhân, số thành phần liên thông lúc này cũng đã giảm xuống và số đỉnh thuộc mỗi thành phần liên thông cũng tăng lên rõ rệt. Sau khi đã xây dựng được đồ thị như trên, ta sẽ tiến hành việc ước lượng trọng số cho các cạnh của đồ thị.

5 Sinh trọng số cho đồ thị

5.1 Các yếu tố cần xem xét

Các yếu tố được xem xét trong mô hình là

1. Nhóm tuổi
2. Nghề nghiệp
3. Mối quan hệ giữa các ca bệnh

5.1.1 Nhóm tuổi

Nhóm tuổi càng xa nhau càng ít tiếp xúc với nhau.

Dữ liệu hiện tại có thể phân ra thành 4 nhóm tuổi, cụ thể như sau:

1. Dưới 18 tuổi
2. Từ 18 đến 40 tuổi
3. Từ 40 - 60 tuổi
4. Trên 60 tuổi

Để biểu diễn được mối liên hệ này, ta xây dựng một ma trận $P = \left[p_{ij}\right]_{n\times n}$, trong đó các phần tử p_{ij} được xác định là quan hệ giữa người thuộc nhóm tuổi i và nhóm tuổi j.

Dễ thấy P là ma trận đối xứng.

Giả sử rằng những người cùng nhóm tuổi có tiếp xúc nhiều nhất, ta gán các giá trị $p_{ii} = 1$. Các phần tử còn lại giảm dần theo khoảng cách $|i - j|$

Giả định

$$p_{ij} = 1 - 0.2\,|i - j|$$

Biểu diễn một cách rõ ràng hơn, ma trận P như sau:

$$P = \begin{bmatrix} 1 & 0.8 & 0.6 & 0.4 \\ 0.8 & 1 & 0.8 & 0.6 \\ 0.6 & 0.8 & 1 & 0.8 \\ 0.4 & 0.6 & 0.8 & 1 \end{bmatrix}$$

5.1.2 Quan hệ

Những người có quan hệ càng gần gũi càng có hệ số cao

Điều này cần kiểm chứng, bởi các cụ già thì thích chơi với trẻ em ???

Những người có quan hệ trong gia đình có tiếp xúc nhiều và gần gũi hơn đồng nghiệp, đồng nghiệp hơn những người có quan hệ xã giao. Những trường hợp không xác định được quan hệ sẽ có trọng số được sinh tự do.

Có 5 loại quan hệ có thể được tìm thấy trong dữ liệu

1. Không xác định.
    - Dạng quan hệ này có thể là 1 trong 4 dạng dưới, hoặc là dạng nào đó không được đề cập đến, trọng số có thể xác định bằng cách sinh ngẫu nhiên.
    - Trọng số này có thẻ cao hoặc không, bởi có nhiều dạng quan hệ không được đề cập đến ở đây như chỉ là xã giao, họ hàng xa, người yêu, bồ bịch,....
2. Bệnh nhân - nhân viên
3. Đồng nghiệp
4. Xã giao
5. Họ hàng

Dựa vào thông tin trên ta có thể giả sử trọng số được cho bởi bảng dưới.

Trọng số này được sinh theo cảm tính, không dựa trên cơ sở nào

Quan hệ	Trọng số
Không xác định	uniform$(0.2, 0.7)$
Bệnh nhân - nhân viên	0.4
Đồng nghiệp	0.5
Người thân	0.9
Xã giao	0.3

Bảng 5.1: Bảng trọng số

5.1.3 Các quyết định của chính phủ

Đây là một yếu tố rất quan trọng ảnh hưởng đến cách thức dịch bệnh phát tán. Điển hình nhất là quả giãn cách xã hội, biết mùi nhau ngay.

Có 4 động thái chính của chính quyền tác động đến Đà Nẵng nhằm ngăn chặn sự lây lan của dịch theo các mốc thời gian và sức tác động của chúng như sau:

Mốc thời gian	Hoạt động	Trọng số
27/07/2020	Thực hiện giãn cách xã hội	3
28/07		1.1
31/07		1.5
02/08		2

Bảng 5.2: Các quyết định được ban hành nhằm ngăn cản dịch bệnh lây lan

Có nhiều yếu tố có thể xét thêm, ví dụ như về nghề nghiệp,..., tuy nhiên tạm thời ta chưa xét đến do vấn đề về dữ liệu, nghề nghiệp thì không ước lượng được.

Từ các yếu tố trên, ta có thể đưa ra một ước lượng cho trọng số cạnh như mục dưới đây.

6 Xây dựng đồ thị

6.1 Ước lượng trọng số cạnh cho đồ thị

6.1.1 Cạnh giữa bệnh nhân - bệnh nhân

Xét một cạnh có hướng e thể hiện quan hệ giữa bệnh nhân s và ca liên quan t vào ngày d.

Giả sử rằng ngày phát bệnh của ca s là ngày d_{o} và ngày bị bế đi cách li là d_{q}. Ta thấy s sẽ không tồn tại trong đồ thị nếu một trong 2 điều kiện sau được thỏa mãn:

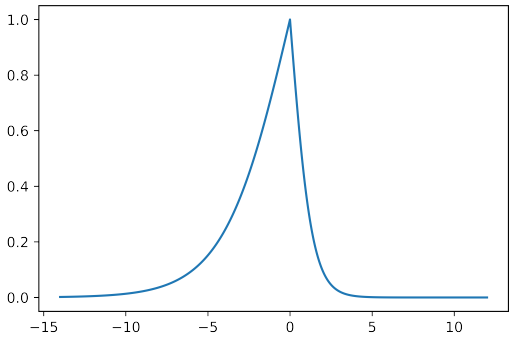
- $d - d_q > 0$
- $d_o - d > 14$

Xét một ngày d nào đó mà cả 2 bệnh nhân s và t cùng tồn tại trong đồ thị, ta sinh trọng số w_e cho cạnh e nối từ t đến s, cạnh e này thể hiện cho mức độ ảnh hưởng của bệnh nhân s lên bệnh nhân t.

Ta mong muốn ước lượng một hàm sinh trọng số đáp ứng được một số yêu cầu sau:

1. Rất nhỏ và tăng rất chậm vào thời điểm mới bắt đầu nhiễm bệnh
2. Tăng nhanh vào thời điểm sau đó, có thể là từ 3, 4, 5,...whatever ngày sau thời điểm bắt đầu nhiễm bệnh
3. Đạt đỉnh vào ngày phát bệnh
4. Sau khi phát bệnh, trọng số giảm dần do bệnh nhân đã có nhận thức về việc bất bình thường trong sức khỏe của mình

Với các tiêu chí như trên, đồ thị ta mong muốn có thể sẽ có dạng như sau:



Hình 6.1: Đồ thị mong muốn

Một dạng đồ thị có thể rất dễ dàng nhìn thấy được là ông thần e^x. (có thể thử sigmoid, ông này nhìn cũng rất hợp lí)

Cho rằng:

$$f(d) = \exp\left\{-|d - d_o| P_{ij} r_{st} \gamma\right\}$$

trong đó:

$$P_{ij} = P_{ij}(d)$$

$$r_{st}$$

$$\gamma = \gamma(d)$$

với P_{ij} là trọng số thể hiện tương tác theo lứa tuổi, r_{st} thể hiện mối quan hệ giữa s và t và γ là trọng số thể hiện cho sức ảnh hưởng của các quyết định.

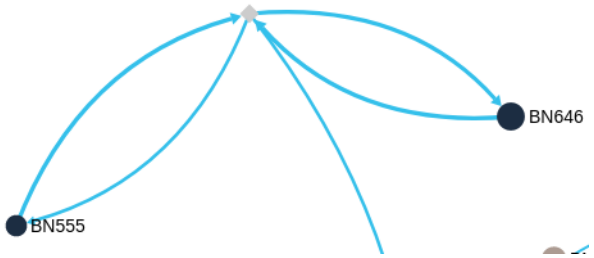
6.1.2 Cạnh giữa bệnh nhân và địa điểm sinh sống

Nguồn bệnh tồn tại trên đồ vật là không lâu, ta có thể cho rằng sau 3 ngày kể từ khi 1 người phát bệnh có tiếp xúc vào đồ vật đó thì nguồn lây gần như không còn khả năng lây bệnh.

Giả sử người s sống tại xã/phường x, vào ngày d, trọng số của cạnh e nối từ x sang s là w xác định như sau:

$$w = \begin{cases} w(d) & \text{trong trường hợp còn lại} \\ 0 & \text{nếu } d - d_o > 3 \text{ hoặc } d < d_o \end{cases}$$

Xét 2 bệnh nhân s_1, s_2 cùng sống tại x và không có liên hệ trực tiếp với nhau. Giả sử rằng quan hệ của các nodes này được biểu diễn như đồ thị dưới đây:



Hình 6.2: Quan hệ P - L - P

Giả sử s_1 tượng trưng cho BN555 và s_2 tượng trưng cho BN546, và trong ngày d, s_1 phát bệnh. Khi đó, cạnh e_{s_1} nối x với s_1 sẽ có trọng số $w_{s_1} = \gamma_{s_1}$, điều này dẫn tới node s_2 có nguy cơ lây nhiễm từ s_1 là e_{s_2} được xác định bởi

$$\gamma_{s_2} = \gamma_{s_1} \times \zeta$$

trong đó ζ là một hệ số suy giảm theo thời gian.

Hiện tại thông tin về nghề nghiệp của các bệnh nhân chưa đầy đủ, do đó việc phân ra khu vực tiếp xúc là trong nhà hay ngoài trời vẫn chưa được thực hiện, chỉ mới áp dụng trên cụm Đà Nẵng.

Trong nhà dễ lây hơn ngoài trời nên trọng số cao hơn, và hệ số suy giảm cũng nhỏ hơn khi ở ngoài trời.

Với cách tiếp cận như trên, ta đi đến việc thử nghiệm trên dữ liệu thật.

6.1.3 Truyền thông

Hầu hết việc sinh trọng số được xác định theo ngày, truyền thông ngày nào cũng đưa tin dẫn đến yếu tố truyền thông nên để trở thành 1 trọng số (hằng số) làm cho trọng số cạnh giảm nhanh hơn (đều đặn từng ngày). Đây được gọi là hiện tượng “mưa dầm thấm lâu".

Ta tạm gọi trọng số này là η.

Tính từ ngày đầu tiên bùng dịch đến ngày d hiện tại là δ ngày, tác động của truyền thông đến trọng số cạnh được tính là $w = \frac{w_0}{\eta^\delta}$ trong đó w_0 là trọng số gốc của cạnh nếu không có sự tác động của truyền thông.

Remark 6.1. Chia như trên sẽ làm w giảm rất nhanh, trong khi thực tế tác động của truyền thông chỉ có tác động mạnh mẽ trong những ngày đầu, rồi sau đó giảm dần. Đến những ngày cuối dịch, tinh thần bà con đang hưng phấn và lạc quan nên có thể sẽ lại làm tăng trọng số lên một chút.

Có thể cải thiện tiếp

6.1.4 Đợt bùng phát dịch đầu tiên

Đợt dịch đầu tiên bắt đầu từ cuối tháng 12/2019, VN sẵn sàng chuẩn bị chống dịch với sự chuẩn bị kỹ lưỡng đến từ mọi mặt, điều này tạo tiền đề cho việc đợt dịch đầu tiên qua đi với gần 300 bệnh nhân và không có ca nào tử vong. Đây là một kết quả rất đáng khích lệ.

Tuy nhiên, kết quả này cũng gián tiếp tạo tâm lý chủ quan cho người dân, mà bằng chứng là đợt dịch lần 2 bùng phát, mặc dù đã có kinh nghiệm từ lần 1, nhưng việc thiếu nghiêm túc chấp hành các quy định về an toàn dẫn đến nguy cơ cao xảy ra bùng phát dịch.

Đây sẽ là trọng số (hằng) làm dịch bùng phát nhanh hơn.

Kinh nghiệm chống dịch cũng đc xem xét, 2 ông thần này tỉ lệ với nhau.

Tạm đặt trọng số thể hiện cho tác động của đợt dịch đầu tiên này là λ

Nhân thẳng λ vào trọng số gốc.

7 Các vấn đề

Ở đây sẽ chạy các thuật toán về xác định cộng đồng (community detection) và centrality (pagerank, eigenvec)

8 Các vấn đề kỹ thuật

8.1 Cơ sở dữ liệu

Neo4j

Tham khảo ở Neo4j

8.2 Tương tác với đồ thị

python-igraph

8.3 Visualization

Sigmajs

Tại sao lại là sigmajs thì đơn giản là ông thần này làm cho ăn sẵn rồi, ít phải làm việc nhiều, thay vào đó có thể có nhiều thời gian để làm việc với networks hơn là chỉ chăm chăm vào visualizes

Tài liệu

[1] T. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit"criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1952.

[2] L. Baringhaus and Norbert Henze. Cramér–von mises distance: probabilistic interpretation, confidence intervals, and neighbourhood-of-model validation. *Journal of Nonparametric Statistics*, 29:1–22, 02 2017.

[3] Francis Bloch, Matthew O. Jackson, and Pietro Tebaldi. Centrality measures in networks, 2017.

[4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117, 1998.

[5] Doina Bucur and Petter Holme. Beyond ranking nodes: Predicting epidemic outbreak sizes by network centralities. *PLOS Computational Biology*, 16(7):e1008052, Jul 2020.

[6] Mohammadreza Doostmohammadian, Hamid Rabiee, and Usman Khan. Centrality based epidemic control in complex social networks. *Social Network Analysis and Mining*, 10, 12 2020.

[7] Karl Pearson. *On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling*, pages 11–28. Springer New York, New York, NY, 1992.

[8] M. A. Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, 1974.