Topology of Contact Network of Twitter Users Based On Filtered Degree PDF

Libao Jin

December 17, 2019

Abstract

In this paper, we develop a scheme to check whether ellipses overlap, based on which we can determine the connectivity of the corresponding vertices to construct a network. Then we get the degree distribution of vertices using filtered PDF to analyze the network topology.

1 Introduction

The spread of epidemic diseases has been of a great interest in past decades. Numerous epidemiological models, e.g., susceptible/infective/removed (SIR) model, susceptible/exposed/infective/removed (SEIR) model, and etc. are developed to reveal how the epidemic diseases are spreaded. However, these models are based on very unrealistic assumptions such as the infective individuals have the same probability to have disease-causing contacts with all the susceptible individuals. In other words, the above models are only applicable to the individuals on the fully-mixed network. Until 2002, Newman proposes an innovative way to generalize the SIR model on fully-mixed networks to general networks using percolation theory and generating function [2]. The key to solve the susceptible/infective/removed (SIR) model on general networks is to find the degree distribution of its vertices.

In this paper, we construct a contact network based on the overlapping of activity space of Twitter users which is represented using ellipses. Then we are interested in finding the degree distribution of the vertices. Recall that the degree of a vertex is the number of adjacent vertices of which each is directly connected to that vertex through an edge. It is known that finding the true degree probablity density function (PDF) or probability mass function is unrealistic. Fortunately, we can use the filtered PDF [1] to approximate the true PDF, from which we analyze the network topology of the contact network of Twitter users.

2 Method

2.1 Overlapping of Ellipses

Let us express a point (x, y) on a ellipse using homogeneous coordinates as $(x, y, 1)$. Every ellipse can be obtained by applying affine transformations

- scaling $S : (x, y, 1) \mapsto (ax, by, 1)$;
- rotation $R : (x, y, 1) \mapsto (x\cos\varphi - y\sin\varphi, x\sin\varphi + y\cos\varphi, 1)$;
- translation $T : (x, y, 1) \mapsto (x + x^*, y + y^*, 1)$;

to a unit circle C centered at the origin, whose parametrization is $x = x(t) = \cos t, y = y(t) = \sin t, t \in [0, 2\pi]$. The affine transformations can also be expressed in matrix form as follows:

$$S = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} \cos\varphi & -\sin\varphi & 0 \\ \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad T = \begin{bmatrix} 1 & 0 & x^* \\ 0 & 1 & y^* \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence, we can compose affine transformations using matrix multiplication. For instance, given any two ellipses E_1, E_2, we can get corresponding parametrizations as below,

$$E_1 = M_1 C = T_1 R_1 S_1 C = \begin{bmatrix} 1 & 0 & x_1 \\ 0 & 1 & y_1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\varphi_1 & -\sin\varphi_1 & 0 \\ \sin\varphi_1 & \cos\varphi_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 & 0 & 0 \\ 0 & b_1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos t \\ \sin t \\ 1 \end{bmatrix},$$

$$E_2 = M_2 C = T_2 R_2 S_2 C = \begin{bmatrix} 1 & 0 & x_2 \\ 0 & 1 & y_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\varphi_2 & -\sin\varphi_2 & 0 \\ \sin\varphi_2 & \cos\varphi_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_2 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos t \\ \sin t \\ 1 \end{bmatrix}.$$

Assume that $\max\{|a_2|, |b_2|\} \geq \max\{|a_1|, |b_1|\}$, i.e., ellipse E_2 has larger scaling parameter. Applying the inverse transformation M_2^{-1} to E_2 and E_1,

$$\begin{aligned} M_2^{-1} E_2 &= M_2^{-1} M_2 C = C, \\ M_2^{-1} E_1 &= M_2^{-1} M_1 C = (T_2 R_2 S_2)^{-1} T_1 R_1 S_1 C = S_2^{-1} R_2^{-1} T_2^{-1} T_1 R_1 S_1 C = E_1', \end{aligned}$$

where

$$S^{-1} = \begin{bmatrix} 1/a & 0 & 0 \\ 0 & 1/b & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad T^{-1} = \begin{bmatrix} 1 & 0 & -x^* \\ 0 & 1 & -y^* \\ 0 & 0 & 1 \end{bmatrix}.$$

It is shown that E_2, E_1 are transformed to a unit circle C centered at the origin, and another ellipse E_1', respectively. Thus, checking whether E_1 and E_2 overlap is equivalent to see whether E_1' and C have intersection, that is, $\exists p = (\hat{x}, \hat{y}) \in E_1'$ such that $\|p\|_2^2 = \hat{x}^2 + \hat{y}^2 \leq 1$. To be more exact, E_1' has the following parametrization,

$$\begin{aligned} E_1' = \; & S_2^{-1} R_2^{-1} T_2^{-1} T_1 R_1 S_1 C \\ = \; & \begin{bmatrix} \frac{a_1}{a_2}\cos(\varphi_1 - \varphi_2) & -\frac{b_1}{a_2}\sin(\varphi_1 - \varphi_2) & \frac{(x_1 - x_2)\cos\varphi_2 + (y_1 - y_2)\sin\varphi_2}{a_2} \\ \frac{a_1}{b_2}\sin(\varphi_1 - \varphi_2) & \frac{b_1}{b_2}\cos(\varphi_1 - \varphi_2) & \frac{-(x_1 - x_2)\sin\varphi_2 + (y_1 - y_2)\cos\varphi_2}{b_2} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos t \\ \sin t \\ 1 \end{bmatrix} \\ = \; & \begin{bmatrix} c_1 & c_2 & c_3 \\ d_1 & d_2 & d_3 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos t \\ \sin t \\ 1 \end{bmatrix} = \begin{bmatrix} c_1\cos t + c_2\sin t + c_3 \\ d_1\cos t + d_2\sin t + d_3 \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{x}(t) \\ \hat{y}(t) \\ 1 \end{bmatrix}. \end{aligned}$$

We can find the closest point $p^* = p(t^*) = (\hat{x}(t^*), \hat{y}(t^*))$ on E_1' to the origin by solving

$$\min_{t\in[0,2\pi]} \|p(t)\|_2^2 = \min_{t\in[0,2\pi]} \hat{x}^2(t) + \hat{y}^2(t).$$

It suffices to find the critical point of $\|p(t)\|_2^2$ by finding zeros of the derivative of $\|p(t)\|_2^2$ with respect to t using either Bisection method or Newton's method:

$$\begin{aligned} \frac{d}{dt}\|p(t)\|_2^2 = \; & \frac{d}{dt}[\hat{x}^2(t) + \hat{y}^2(t)] \\ = \; & 2\hat{x}(t)\hat{x}'(t) + 2\hat{y}(t)\hat{y}'(t) \\ = \; & 2(c_1\cos t + c_2\sin t + c_3)(-c_1\sin t + c_2\cos t) + \\ & 2(d_1\cos t + d_2\sin t + d_3)(-d_1\sin t + d_2\cos t) \\ = \; & 0. \end{aligned}$$

Note that there exist more than one critical points, and Newton's method can only find one of them for a given initial guess. In order to find all the critical points and pick the minimizer, we need to provide initial guesses with fixed difference. In our numerical experiment, we can use Newton's method four times with initial guesses 0, $\pi/2$, π and $3\pi/2$. Suppose two zeros are t_1 and t_2, we will get

$$\begin{cases} \min_{i=1,2} \|p(t_i)\|_2^2 \leq 1 & \text{if } E_1' \text{ intersects } C \text{ or } E_1' \text{ is inside of } C, \\ \min_{i=1,2} \|p(t_i)\|_2^2 > 1 & \text{otherwise.} \end{cases}$$

Example 2.1. Given two ellipses E_1 and E_2 with following parameters:

$$E_1 : x_1 = -4.5, y_1 = 3, a_1 = 2, b_1 = 1, \varphi_1 = \frac{\pi}{4},$$
$$E_2 : x_2 = -7, y_2 = 4, a_2 = 3, b_2 = 1.5, \varphi_2 = \frac{\pi}{2}.$$

Applying the inverse transformation to both E_1 and E_2, we can obtain E_1' and E_2' as shown in Figure 1.

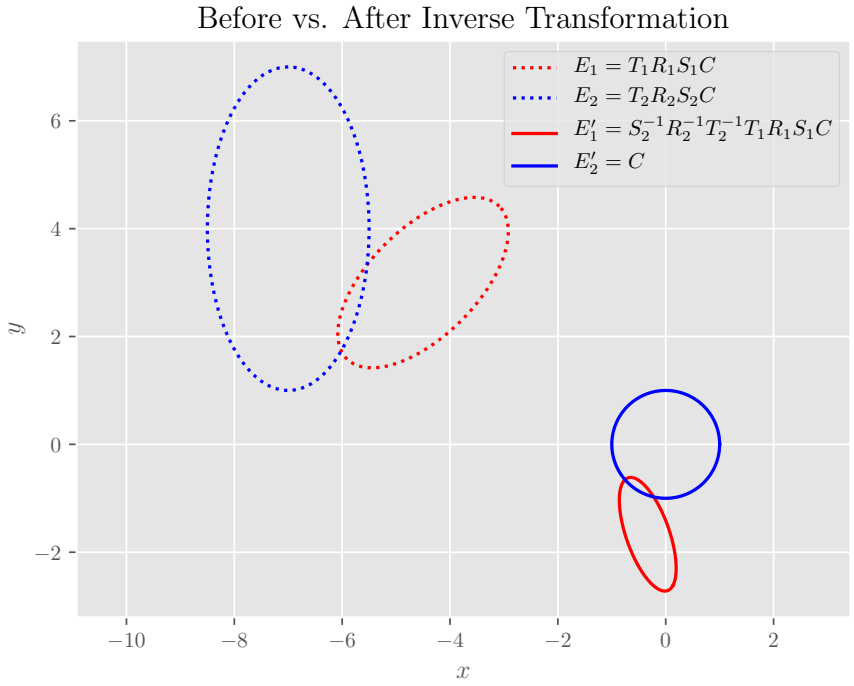

Figure 1: An example of affine transformation of two ellipses

2.2 Contact Network and Degree of Vertices

Given ellipses $E_1, E_2, \ldots, E_n$, we can construct a graph that reflects the overlapping between ellipses: Regard each ellipse, say E_i as a vertex v_i in the graph. Then if E_i and E_j overlap, we put an edge between v_i and v_j. Therefore, we can use an adjacency matrix A to represent the graph by setting A_{ij} as follows,

$$A_{ij} = \begin{cases} 1 & \text{if } E_i \text{ and } E_j \text{ overlap,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Note: $A_{ii} = 0, i = 1, 2, \ldots, n$. Thus the degree for vertice v_i is simply the sum of row i of A as follows,

$$d_i = \deg(v_i) = \sum_{j=1}^{n} A_{ij}. \tag{2.2}$$

2.3 Filtered PDF

The filtered PDF [1] over the interval Δx is given below,

$$\begin{aligned} f_\Delta(x) = \frac{1}{\Delta x}\frac{\Delta N}{N} &= \frac{1}{N\Delta x}\sum_{i=1}^{N}\left[\theta\left(x + \frac{\Delta x}{2} - X_i\right) - \theta\left(x - \frac{\Delta x}{2} - X_i\right)\right] \\ &= \frac{1}{N\Delta x}\sum_{i=1}^{N}\begin{cases} 1 & \text{if } x - \Delta x/2 \le X_i \le x + \Delta x/2, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{2.3}$$

where ΔN is the number of samples that are found in the interval $x - \Delta x/2 \le X_i \le x + \Delta x/2$. The filtered PDF has the following properties:

1. For any function $g(x)$, the filtered PDF $f_\Delta(x)$ has the property
$$\int_{-\infty}^{\infty} g(x) f_\Delta(x)\, dx = \langle g_\Delta(X)\rangle.$$
2. By setting $g = 1$, we find that $f_\Delta(x)$ represents indeed a PDF because it integrates to one,
$$\int_{-\infty}^{\infty} f_\Delta(x)\, dx = \langle 1\rangle = 1.$$
3. By setting $g(x) = x$, we have the mean $\langle X\rangle$,
$$\int_{-\infty}^{\infty} x f_\Delta(x)\, dx = \langle X\rangle.$$
4. By setting $g(x) = (x - \langle X\rangle)^2$, we have the variance $\langle \widetilde{X}^2\rangle$,
$$\int_{-\infty}^{\infty} (x - \langle X\rangle)^2 f_\Delta(x)\, dx = \langle \widetilde{X}^2\rangle.$$

Now given N sample points $\{X_i\}_{i=1}^{N}$, we have the moments of nth-order as follows,

$$\begin{aligned} \langle X^n\rangle &= \int_{-\infty}^{\infty} x^n f_{\Delta(x)}\, dx \\ &= \frac{1}{\Delta x}\left\langle \int_{X-\Delta x/2}^{X+\Delta x/2} x^n\, dx \right\rangle \\ &= \frac{1}{\Delta x}\left\langle \frac{1}{n+1}\left(X + \frac{\Delta x}{2}\right)^{n+1} - \frac{1}{n+1}\left(X - \frac{\Delta x}{2}\right)^{n+1} \right\rangle \\ &= \frac{1}{\Delta x}\frac{1}{n+1}\frac{1}{N}\sum_{i=1}^{N}\left[\left(X_i + \frac{\Delta x}{2}\right)^{n+1} - \left(X_i - \frac{\Delta x}{2}\right)^{n+1}\right]. \end{aligned}$$

Likewise, we can find the central moments of nth-order as below,

$$\begin{aligned}\langle \widetilde{X}^n \rangle &= \int_{-\infty}^{\infty} (x - \langle X \rangle)^n f_\Delta(x)\, dx \\ &= \frac{1}{\Delta x} \left\langle \int_{X-\Delta x/2}^{X+\Delta x/2} (x - \langle X \rangle)^n \, dx \right\rangle \\ &= \frac{1}{\Delta x} \left\langle \frac{1}{n+1} \left(X + \frac{\Delta x}{2} - \langle X \rangle \right)^{n+1} - \frac{1}{n+1} \left(X - \frac{\Delta x}{2} - \langle X \rangle \right)^{n+1} \right\rangle \\ &= \frac{1}{\Delta x} \frac{1}{n+1} \frac{1}{N} \sum_{i=1}^{N} \left[\left(X_i + \frac{\Delta x}{2} - \langle X \rangle \right)^{n+1} - \left(X_i - \frac{\Delta x}{2} - \langle X \rangle \right)^{n+1} \right].\end{aligned}$$

Then we can use the moments and central moments to obtain the mean $\langle X \rangle$, the variance $\langle \widetilde{X}^2 \rangle$, the skewness $\langle \widetilde{X}^3 \rangle / \langle \widetilde{X}^2 \rangle^{3/2}$, and the flatness $\langle \widetilde{X}^4 \rangle / \langle \widetilde{X}^2 \rangle^2$.

3 Results

The preprocessed Twitter data consists of 57688 entries which removed the abnormal data such as ellipses with area of the ellipses being zero (the major/minor axis being zero). The Twitter users are mainly distributed in the East Coast (Figure 2). Then we use (2.1) to construct the adjacency matrix A of the contact network, and thus obtain the degree of each vertex by (2.2). Then we apply (2.3) with $\Delta x = 4489.50, 561.19, 224.47$ and obtain the plots of the following PDFs (Figure 3) and the corresponding degree statistics (Table 1).

Table 1: Degree Statistics (No Vertices Excluded)

Δx	Mean	Variance	Skewness	Flatness
4489.50	1663.15	7055546.37	2.61	17.96
561.19	1663.15	5402156.47	3.89	28.63
224.47	1663.15	5380111.27	3.92	28.84

As shown in Figure 3, the filtered degree PDF with $\Delta x = 561.83$ is the smoothest one. However, the PDF is not smooth at all when the degree is close to zero. The possible cause is that there are some Twitter users have extremely large activity area, which causes the vertices of small degrees have higher degrees. To avoid the impact of these outliers, we exclude the Twitter users with top 5 % area size and repeat the above steps to obtain the PDFs (Figure 4) and statistics (Table 2).

Table 2: Degree Statistics (Vertices of Top 5 % Area Excluded)

Δx	Mean	Variance	Skewness	Flatness
1421.00	1185.25	2899211.24	2.56	11.84
177.62	1185.25	2733570.38	2.80	12.95
71.05	1185.25	2731361.84	2.80	12.97

Figure 4 shows that the filtered PDF has exponential decay on the interval $(0, \infty)$, hence we use maximum likelihood estimation (MLE) to obtain the Gamma PDF with parameters $k = 0.6274$

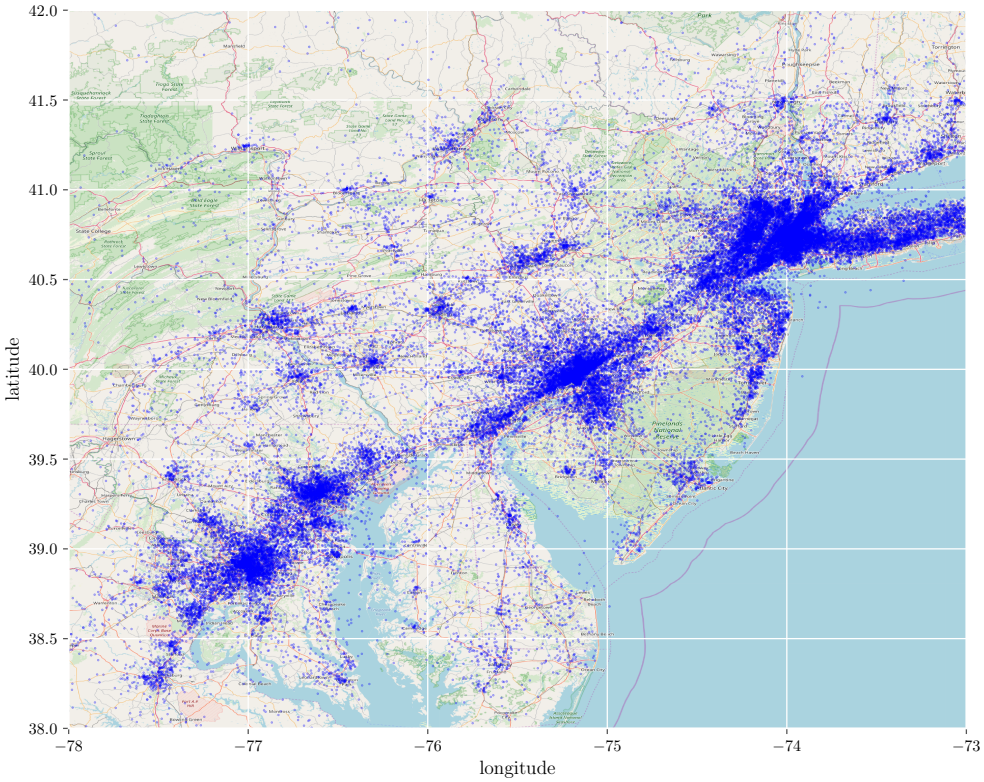

Figure 2: Spatial Distribution of Twitter Users (Center of the Ellipse)

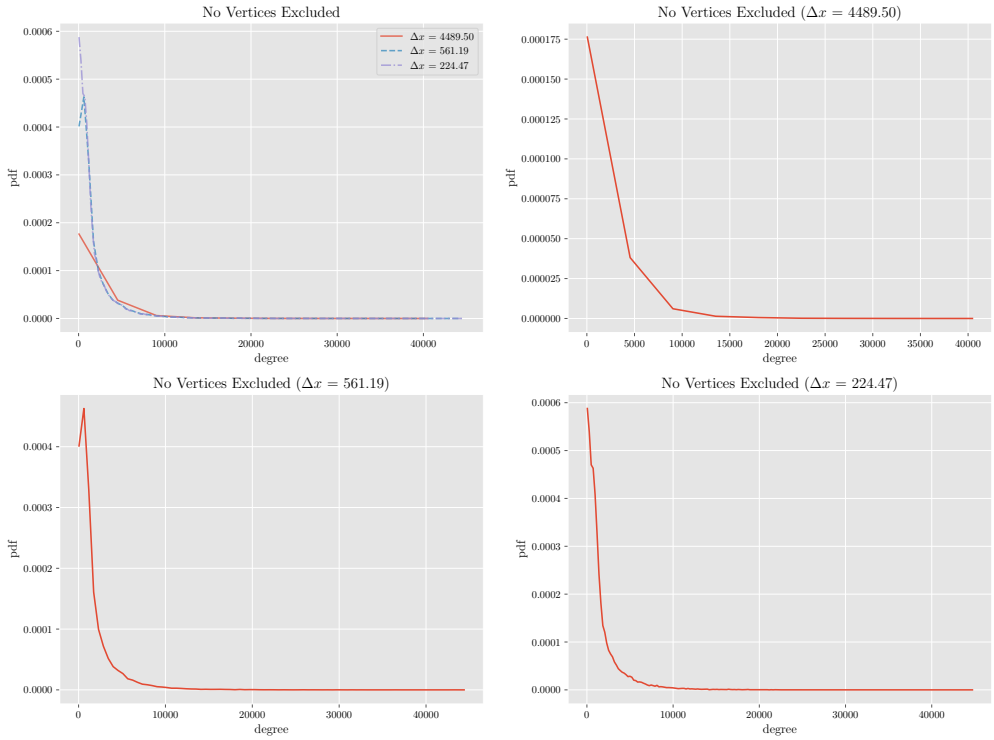

Figure 3: Filtered Degree PDF of Network without Vertices Excluded

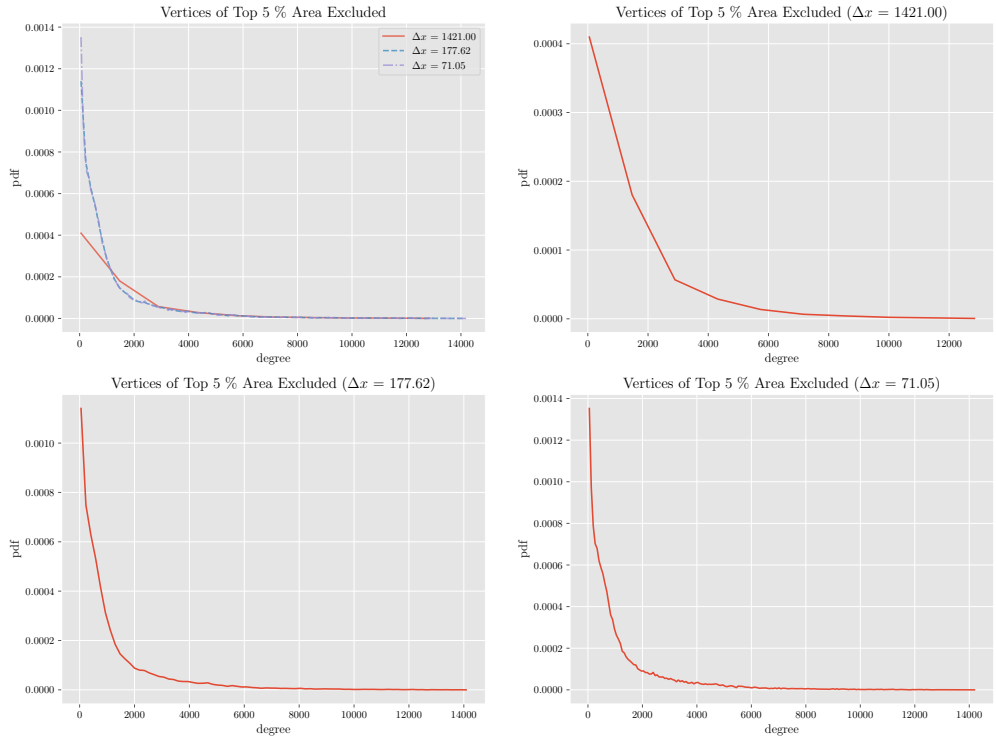

Figure 4: Filtered Degree PDF of Network with Vertices of Top 5 % Area Excluded

(shape parameter) and $\theta = 1889.0135$ (scale parameter). Then we compare the filtered PDF with the Gamma PDF in the same plot (Figure 5) and calculate the statistics of the corresponding Gamma PDF (Table 3).

Table 3: Statistics of Gamma PDF with $\hat{k} = 0.6274$ and $\hat{\theta} = 1889.0135$

Mean	Variance	Skewness	Flatness
1185.25	2238952.06	2.52	9.56

4 Discussion

1. The spatial distribution of the Twitter users (Figure 2) does reflect the population density in the east coastal area. To some extent, it is reasonable that the Twitter users can represent the population in that area (though the popluation movement dynamics might differ).

2. The exponential decay is realistic, because only a very small fraction of the population have large degrees due to the traveling. In other words, the contact network we constructed is not densely connected compared to the fully-mixed network whose degree of each vertex is exactly $n - 1$.

3. It is both shown in Figure 5 and Table 3 that the degree distribution follow the Gamma distribution.

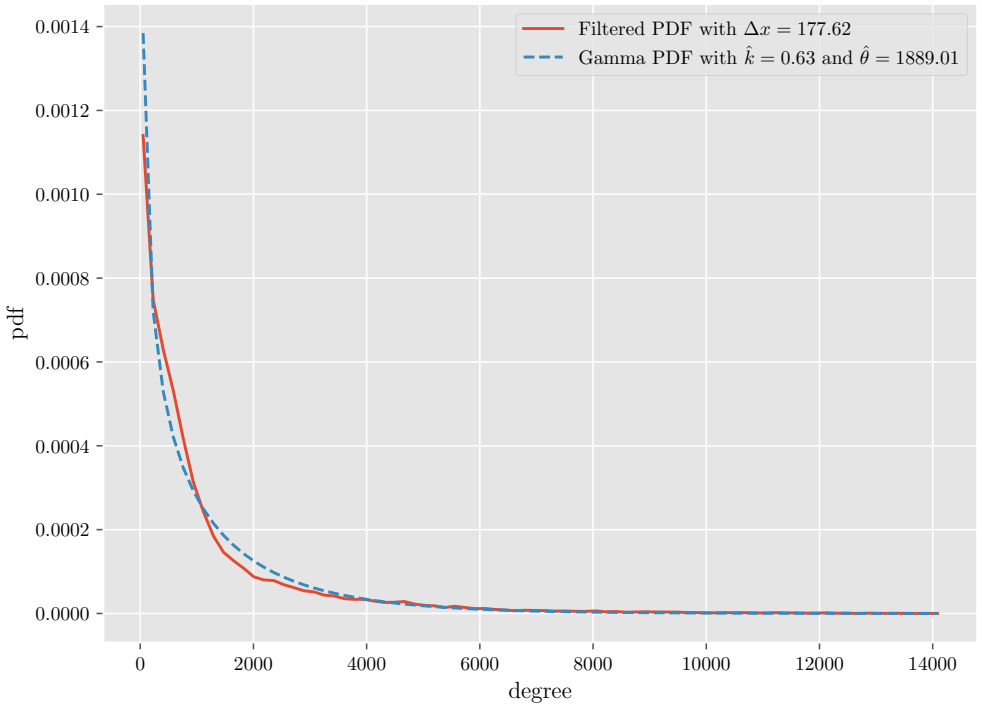

Figure 5: Filtered Degree PDF of Contact Network with Vertices Top 5 % Area Excluded vs. Gamma PDF with $\hat{k} = 0.6274$ and $\hat{\theta} = 1889.0135$

4. Therefore, the degree distribution we obtained can be used for solving the SIR model on the contact network.

References

[1] Heinz, S. *Mathematical Modeling*. Springer-Verlag Berlin Heidelberg, Department of Mathematics and Statistics, University of Wyoming, 2011.

[2] Newman, M. E. J. Spread of epidemic disease on networks. *Physical Review E 66*, 1 (Jul 2002).