DOI: 10.1002/pra2.252

LONG PAPERS

COVID-19 pandemic and information diffusion analysis on Twitter

Ly Dinh† | Nikolaus Parulian†

School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

Correspondence
Ly Dinh, School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL 61820.
Email: dinh4@illinois.edu

Abstract

The COVID-19 pandemic has impacted all aspects of our lives, including the information spread on social media. Prior literature has found that information diffusion dynamics on social networks mirror that of a virus, but applying the epidemic Susceptible-Infected-Removed model (SIR) model to examine how information spread is not sufficient to claim that information spreads like a virus. In this study, we explore whether there are similarities in the simulated SIR model (*SIRsim*), observed SIR model based on actual COVID-19 cases (*SIRemp*), and observed information cascades on Twitter about the virus (*INFOcas*) by using network analysis and diffusion modeling. We propose three primary research questions: (a) What are the diffusion patterns of COVID-19 virus spread, based on *SIRsim* and *SIRemp*? (b) What are the diffusion patterns of information cascades on Twitter (*INFOcas*), with respect to retweets, quote tweets, and replies? and (c) What are the major differences in diffusion patterns between *SIRsim*, *SIRemp*, and *INFOcas*? Our study makes a contribution to the information sciences community by showing how epidemic modeling of virus and information diffusion analysis of online social media are distinct but interrelated concepts.

KEYWORDS

COVID-19, epidemic modeling, information diffusion, network analysis, social media

1 | INTRODUCTION

On December 31, 2019, The World Health Organization (WHO) reported the first confirmed case of SARS-CoV-2 virus, frequently known as "COVID-19". To date, the virus has spread to more than 150 countries, with over three million confirmed cases globally. Modeling and examining diffusion dynamics of COVID-19 pandemic network is critical to provide information that health professionals and associated stakeholders can leverage to make effective decisions (Xie et al., 2020).

Extant literature has found that information diffusion dynamics on social networks mirror that of a virus (Abdullah & Wu, 2011; Lerman, 2016; Seki & Nakamura, 2016; Ver Steeg, Ghosh, & Lerman, 2011). (Abdullah & Wu, 2011) show that trending topics on Twitter spread in similar patterns with an epidemic *Susceptible-Infected-Removed* model (*SIR*), in which *I* and *R* both started at 0 but as I began to increase at a certain reproductive rate, *S* and *R* would be impacted.

† These authors contributed equally to this work

83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.

(Seki & Nakamura, 2016) assert that the *SIR* model can be applied to examine the decline of diffusion activities on Friendster online social network, and found that the decline started when popular users left Friendster (labeled as *R* in *SIR* model). However, applying the epidemic *SIR* model to examine how information spread is not sufficient to claim that information spreads like a virus (Lerman, 2016; Wu, Huberman, Adamic, & Tyler, 2004). There are different mechanisms that influences how information spread from one user to another, but does not influence how a virus spread from one person to another, and vice versa (Lerman & Ghosh, 2010; Mønsted, Sapieżyński, Ferrara, & Lehmann, 2017).

In this study, we examine in parallel the epidemic and information diffusion, and the mechanisms by which both diffusion processes contribute to COVID-19's spread. Specifically, we compare COVID-19 virus's (a) *SIR* -modeled and (b) empirically observed diffusion patterns with (c) information cascades of retweeting, quote tweeting, and replying behaviors on Twitter social network to understand the relationships between information and virus diffusion. To do this, first, we create an *SIR* simulation (we call this *SIRsim*) of COVID-19's diffusion with respect to empirically validated parameters such as reproductive rate (*R0*), incubation period, and symptom length range. Secondly, we create an SIR model from actual confirmed cases with data gathered from Johns Hopkins University (JHU-CSSE, 2020) (we call this *SIRemp*). Thirdly, we construct information cascades from on our collected Twitter data (we call this *INFOcas*) based on three dimensions: retweets, quote tweets, same as retweets but with comment included), and replies to tweets. For the information cascades, we also categorize each piece of information to either *Susceptible* (new tweet about the virus), *Infected* (retweets, quoting of retweets, or replying to tweets), or *Removed* (tweet not shared by others after a period of time). Consistent with the aspects of the study, we propose three primary research questions:

RQ1: What are the diffusion patterns of COVID-19 virus spread, based on *SIRsim* and *SIRemp*?

RQ2: What are the diffusion patterns of information cascades on Twitter (*INFOcas*), with respect to retweets, quote tweets, and replies?

RQ3: What are the major differences in diffusion patterns between *SIRsim*, *SIRemp*, and *INFOcas*?

Our study makes a contribution to the information sciences community by showing how epidemic modeling of virus and information diffusion analysis of online social media are distinct, but interrelated concepts.

2 | RELATED WORK

2.1 | Information diffusion on social networks

With the advent of social networking sites and online microblogs such as Twitter, individuals can create and exchange information with larger amounts of people in lesser amounts of time. These online social networks are thus instrumental for researchers to examine what types of information diffuses between individuals and what underlying mechanisms facilitate the diffusion. In the context of social networks, information diffusion is formally defined as a process by which a piece of information is passed down from one node to another node through an edge (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004; Guille, Hacid, Favre, & Zighed, 2013). Two seminal models have been widely adopted to examine diffusion dynamics with network structure considered, namely independent cascade models (Goldenberg, Libai, & Muller, 2001) and linear threshold models (Granovetter, 1978). Independent cascade models assume that each node has a certain fixed probability to spread, or "infect" a piece of information to a neighboring node. On the other hand, linear threshold models posit that a node would be "infected" by a piece of information if a certain threshold of neighboring nodes have also been infected by that information. Both models have been widely used to detect influential topics (Gruhl et al., 2004) and influential users (Yang & Leskovec, 2010) in online social networks and the impacts they have on diffusion rate. (Gruhl et al., 2004) focus on the spread of topics on blogs based on RSS (rich site summary) feeds and found that topics were either consistently popular (called "chatter") or only popular for a short time (called "spikes"). The authors also observed that topics with high chatter also contained larger and more frequent spikes. (Yang & Leskovec, 2010) demonstrate that an influential node can be detected with respect to how many nodes have been influenced by that particular node before.

2.2 | Epidemic models for information diffusion

In addition to the independent cascade model and linear threshold model, scholars studying information diffusion from a wide range of disciplines have also found the utility of modeling diffusion as an epidemic process. In particular, the SIR model has been frequently used to explain how information in an online social network becomes "infectious" and passes from one node to another. SIR is known as a compartmental model. Because it categorizes an individual to be in one of three states at a certain point in time, susceptible (S), infected (I), or removed (R) (Kermack & McKendrick, 1927). An individual may transition their state due to influence from another individual in the same network, in which the transition is linear ($S \rightarrow I$, $I \rightarrow R$). At the first transition point, $S \rightarrow I$ occurs because a susceptible individual was in contact with an infected individual and therefore got the virus. The infection assumed at this transition point is at a constant rate of β per time unit. At second transition point, $I \rightarrow R$ transition occurs when an infected individual either recovered from the virus and got immunity from it, or has been removed (i.e., has died). At this transition, the model assumes that recovery rate is fixed at γ per time unit. These assumptions are stated in the following set of equations of (S), (I), (R) at time (t):

$$\frac{dS}{dt} = -\beta \cdot S(t)$$

$$\frac{dI}{dt} = \beta S(t) - \gamma \cdot I(t)$$

$$\frac{dR}{dt} = \gamma \cdot I(t)$$

(Abdullah & Wu, 2011) examine how trending news spread on Twitter by sorting users into three compartments, S for users who saw tweets from an infected user, I for users who tweet about a news topic, and R for users who no longer tweet about a topic after a predefined timeframe of 4 h. The authors also assume fixed infection rate β and recovery rate γ in their epidemic simulation and observed model with Twitter data, and found a strong fit between the models. In addition to news, scholars have also examined whether false rumors and disinformation diffuse on social networks in a manner similar to how an infectious disease spread (Jin, Dougherty, Saraf, Cao, & Ramakrishnan, 2013; Nekovee, Moreno, Bianconi, & Marsili, 2007). Research by (Nekovee et al., 2007) conceptualizes rumor spreading as a epidemic transition process between ignorants, spreaders, and stiflers. They found that rumor spread rate is higher in scale-free networks than in random graphs. Their finding is consistent with (Lerman & Ghosh, 2010)'s observation that information cascades on Twitter follow a power-law distribution. (Jin et al., 2013) also refine the SIR model to examine rumor diffusion by adding exposed E and skeptical Z individuals, and found that the rate of rumor infection (I) increases as the rate of E decreases, and the susceptible (S) rate decreases as Z increases. Other works have also found SIR models to be useful in explaining diffusion of content on other social networking platforms such as Flickr (Cha, Mislove, Adams, & Gummadi, 2008) and Digg (Ver Steeg et al., 2011).

On the other hand, several studies observe that there are clear differences in SIR epidemic model and information diffusion process. (Goel, Munagala, Sharma, & Zhang, 2015) do not find strong correlation between the SIR model and observed retweet cascades as the epidemic model do not take into account users' characteristics. Similarly, (Liu & Zhang, 2014) point out that information diffusion process includes variables not in SIR model such as content of the information, strength of ties among individuals, and other social factors. In light of diverse findings on the extent to which SIR models can explain information diffusion on social networks, we examine whether there are similarities in our simulated SIR model (*SIRsim*), observed SIR model based on actual COVID-19 cases (*SIRemp*), and observed information cascades on Twitter about the virus (*INFOcas*).

3 | OUR FRAMEWORK AND METHODOLOGY

We empirically test whether there are similarities between the information diffusion process on Twitter about COVID-19 topics and the diffusion of the virus itself between individuals. To do this, we develop three different networks. The first two networks are created to capture the diffusion of the COVID-19 virus in the entire population, via an SIR simulated model (*SIRsim*) and an observed model based on reported data about infected (I), and removed (R) cases (*SIRemp*). The third network is constructed from information cascades on Twitter (we call this *INFOcas*), where infected (I) are tweets that interacted with the original tweets about COVID-19 by either retweeting, quoting, or replying, and removed (R) include tweets that are no longer interacted with for a defined period. We describe the datasets used and the process of constructing each network in the following sections. All data collected and code used in this work are available on FigShare (Dinh & Parulian, 2020).

3.1 | SIR simulation model (*SIRsim)*

We implement a SIR simulation model of COVID-19 on NetLogo,[1] an open-source environment for agent-based modeling. We extended an existing model[2] on virus spread on Netlogo, and refined model parameters based on official sources' information about COVID-19 spread and shown in Table 1. We keep the parameters constant throughout the simulation, and set the duration of the simulation to 88 days. We choose the duration of 88 days to reflect the timeframe between December 17, 2019 to

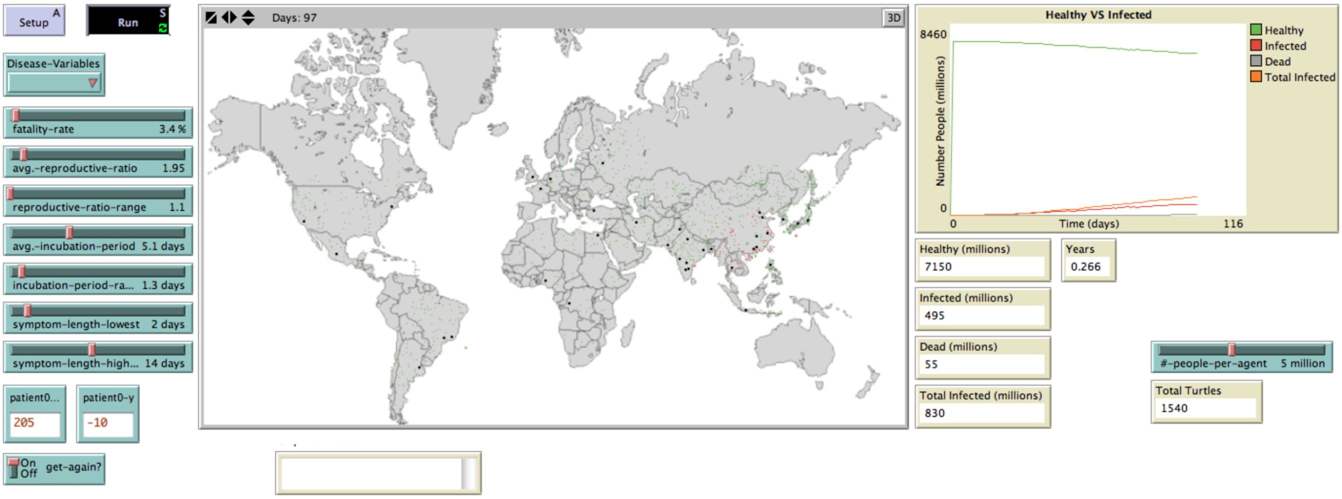
March 14, 2020. We choose December 17, as opposed to December 31, as the first date of COVID-19 to take into account the 14 days (see Table 1 for virus symptom length) of symptoms leading up to the confirmation of the infected case. The initial population for our model includes the entire world population, at 7.7 billion people.[3]

Figure 1 shows the NetLogo interface of our *SIRsim* model, with additional parameters included to simulate the transitions of agents from ($S \rightarrow I$), and $I \rightarrow R$). Adhering to the SIR model, S agents represent the carriers of the virus, I agents are those infected by the carriers, and R are agents who are removed due to death. Due to computational limitations that poses difficulty to represent each individual as an agent, we group 5 million people in each agent (#-people-per-agent setting). Thus, our model contains 1,540 agents interacting with one another. The first agent represents patient zero, and is originated the city of Wuhan in our world map (x-axis: 205, y-axis: -10). We assign agents to move around 36 major cities across the world (e.g., New York City, Paris, Tokyo, Moscow) (see Table A1 in Appendix). All agents initially started in S state, except for patient zero, who then spreads the disease by contacting with agents from other cities through two modes of traveling: driving (parameter *mode* = "human") or flying (parameter *mode* = "plane"). We set these parameters through the use of *patches* (*pixel)* feature, enabling each agent to move certain distances depending on the patch size. The circumference of our simulated "world" is 711 pixels, and with the given circumference of 24,901 miles,[4] each patch covers about 35 miles in our model. To simulate driving, we calculate the average mileage driven per day[5] (36.9 miles), and then derive a movement of 1.05 patches per day for each agent. To simulate flying, each agent has a random chance to create an airplane and fly to any other major cities. While our model accounts for many parameters that are reflective of actual virus spread dynamics, we do not take into account any virus control strategies such as quarantine or social distancing.

TABLE 1 Parameter settings for *SIRsim*

Parameter	Setting	Source
Fatality Rate	3.4%	WHO Director-General's media briefing on COVID-19 (Ghebreyesus, 2020)
Avg. Reproductive Ratio (R_0)	1.95%	(Ghebreyesus, 2020)
Avg. R_0 Range	1.1	(Ghebreyesus, 2020)
Avg. Incubation Period	5.1	(Lauer et al., 2020)
Incubation Period Range	1.3	(Lauer et al., 2020)
Symptom Length (Lowest)	2 days	(CDC, 2020); (Lauer et al., 2020)
Symptom Length (Highest)	14 days	(CDC, 2020); (Lauer et al., 2020)
Duration of Simulation	97 days	Virus started from Dec. 8, 2019 (Wu & McGoogan, 2020)

We repeat the simulation over 100 iterations to ensure reliability of experimental results. Each iteration result is presented as a network that contains multiple types of nodes, *susceptible*, *infected*, and *removed*. An edge can form between any two node types, and node type can change over time (e.g., from *susceptible to infected* if there is an edge between the two nodes), except for when a node has been labeled as *removed*.



fatality-rate 3.4 %
avg.-reproductive-ratio 1.95
reproductive-ratio-range 1.1
avg.-incubation-period 5.1 days
incubation-period-ra... 1.3 days
symptom-length-lowest 2 days
symptom-length-high... 14 days
patient0... 205
patient0-y -10
Healthy (millions) 7150
Years 0.266
Infected (millions) 495
Dead (millions) 55
Total Infected (millions) 830
#-people-per-agent 5 million
Total Turtles 1540

FIGURE 1 NetLogo simulation interface for *SIRsim*

3.2 | SIR model from empirically-validated cases (*SIRemp*)

We gather actual cumulative cases of COVID-19 from Johns Hopkins Center for Systems Science and Engineering (JHU CSSE)'s data repository. This repository contains global confirmed cases, death cases, and recovered cases from January 22 to March 14, 2020, for over 185 countries (JHU-CSSE,- 2020). To our knowledge, this data repository is the most comprehensive so far, with triangulation of cases counts from 18 sources (e.g., WHO, China CDC, Italy Ministry of Health, WorldoMeters). We analyze this dataset within the assumptions of SIR model, where S are individuals in the population that are not yet infected nor immune to the virus, I is equivalent to "confirmed cases" in the dataset, and R is equivalent to "deaths cases". We do not include the "recovered" cases in our model as the data does provide whether these cases are re-entered into the "confirmed cases" in latter time-frames. In the original dataset, there is no inclusion of S, given that susceptible nodes include all members of the world population.

3.3 | Information diffusion on Twitter (*INFOcas*)

The third dataset we use for this research is Twitter data that contains information about COVID-19. We collect tweets during the period of December 31, 2019 to March 14, 2020 with a maximum of 10,000 samples (limit set by firehose) for each day from Crimson Hexagon firehose.[6] We collect 675,228 tweets that include either or all of the hashtags #coronavirus, #covid19, #ncov. We construct information cascades based on three primary behaviors that occurs between tweets in our dataset: (1) retweet, (2) quote tweet, and (3) reply. We exclude all tweets content originated from European countries, in recognition of General Data Protection Regulation (GDPR).[7] Based on the *SIR* model, we define the conditions for *infected* nodes, and *removed* nodes below. Our approach does not consider *susceptible* nodes because in this context, susceptible tweets are all tweets that exist on Twitter.

3.3.1 | New information

An original tweet that has yet to be retweeted or interacted with is counted in this category. There are 14,139 tweets in this category.

3.3.2 | Infected (I)

If a tweet interacts with an S tweet through either retweeting, quoting, or replying, the tweet is counted in this category. (a) Retweet is an action of reposting an original tweet, and without changing the original tweet content. Our sample contains 419,739 retweets. (b) Quote tweet is an action to forward the message with additional information related to the original tweet. Quoting is usually used if the user wants to add a new comment about the related event but still preserving the original content. Our sample contains 17,569 quote tweets. (c) Replying entails commenting on the original tweet, using Twitter's "reply" function. Reply tweets are aggregated into a "thread" of discussions under the original tweet. There are 22,594 tweets that are in the replies category. Cascade statistics are shown in Table 2.

3.3.3 | Removed (R)

If an original tweet has not been retweeted, quoted, or replied to by other tweets in a defined period. We used the average delta time between each activity on the original tweet as our incubation period. Therefore, if there is no user interaction with the tweet between the average time frame from the latest spread, we consider the tweet is removed. Average delta time statistics for each type of cascade (retweet, quote tweet, and reply tweet) can be seen in Table 2. There are 69,216 tweets in total that are in this category.

An information cascade is determined by the period other tweets (I) interact with an original tweet (S) on this dataset. Given an original tweet (T_0) on time $t0$ the cascade c on time $t1$ (c_{t1}) is equal to:

TABLE 2 *INFOcas* cascades and network descriptives

		Retweet	Quote tweet	Reply tweet
Cascades statistics	# of cascades	419,739	17,569	22,594
	Avg. Δ time	0 day-04:54:29	0 day-14:55:33	0 day-08:42:30
	S.D	1 day-00:42:14	2 days-17:55:19	1 day-19:27:15
Network statistics	# of nodes	303,486	15,962	19,016
	# of edges	389,717	15,651	17,712
	Density	4.23e-06	6.14e-0.6	4.89e-05

$$c_{t1} = t1 - t0$$

For each type of information cascade, we analyze the cascade growth by aggregating the *S*, *I*, and *R* tweet for each day.

4 | RESULTS

4.1 | *SIRsim* and *SIRemp*

Our first research question asks about the diffusion patterns of COVID-19 based on both a simulated SIR model (*SIRsim)* and actual number of cases from empirically-validated sources (*SIRemp*). For *SIRsim*, across 100 iterations of our simulation, we find the average counts of *susceptible* agents to be 7,299.4 million, average counts of *infected* to be 384.9 million, and average counts of *removed* to be 15.0 million. Thus, the proportion of healthy, but susceptible agents is 94.8% (*S*) in our model. There are only 5% (*I*) of agents that are infected by the virus, and only 0.19% (*R*) are removed due to death. As shown in Figure 2, the distribution of *infected* (blue line, left) and *removed* (red line, left) agents per day, non-cumulatively, and find an increasing pattern for both trendlines. The proportions of removed cases is much lower than infected cases, and this is shown in the network visualization in Figure 3.

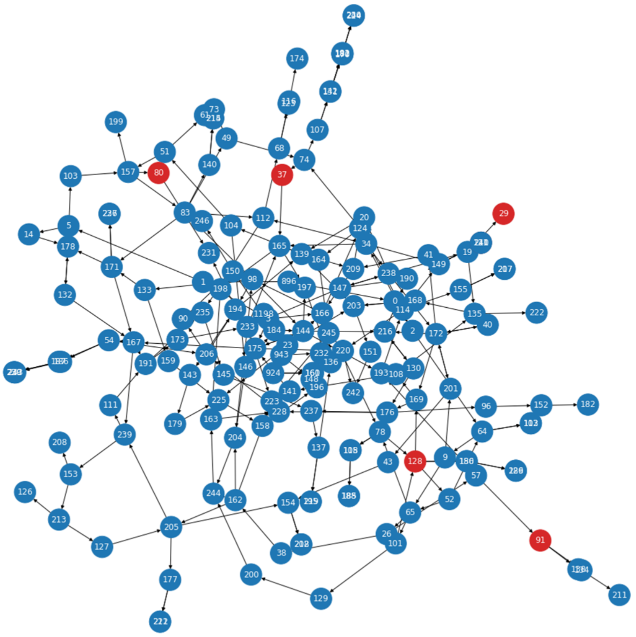


FIGURE 2 Distribution of *Infected* and *Removed* agents for *SIRsim* (left) and *SIRemp* (right) models

We then compare these results to *SIRemp*, which finds that as of March 14, 2019, there were 156,102 *infected* cases, and 5,819 *removed* cases (deaths only). By proportion with the world population, therefore, *infected* cases is 0.002%, and *removed* cases is a minimal percent. By comparison, the empirically-validated results show substantially lower proportions of *infected* and *removed* agents, and in turn, higher proportion of *susceptible* agents. We also analyze the distribution of *infected* (blue line, right) and *removed* cases (red line, right) for *SIRemp*, and finds multiple spikes in the blue line, but flat distribution for the red line. The spikes in *infected* counts are due to inclusion of cases from countries such as the U.S, South Korea, Italy. In comparison to the distributions from *SIRsim*, the distribution of *removed* cases in *SIRemp* is relatively static throughout.

4.2 | Twitter cascades: *INFOcas*

Table 2 (network statistics) shows the sizes of the three network cascades within *INFOcas*, retweet, quote tweet, and reply tweets. We find that retweet cascade is 19 times larger in size than the quote tweet cascades, and 16 times larger than the reply cascades. This finding is consistent with the notable differences in the number of cascades

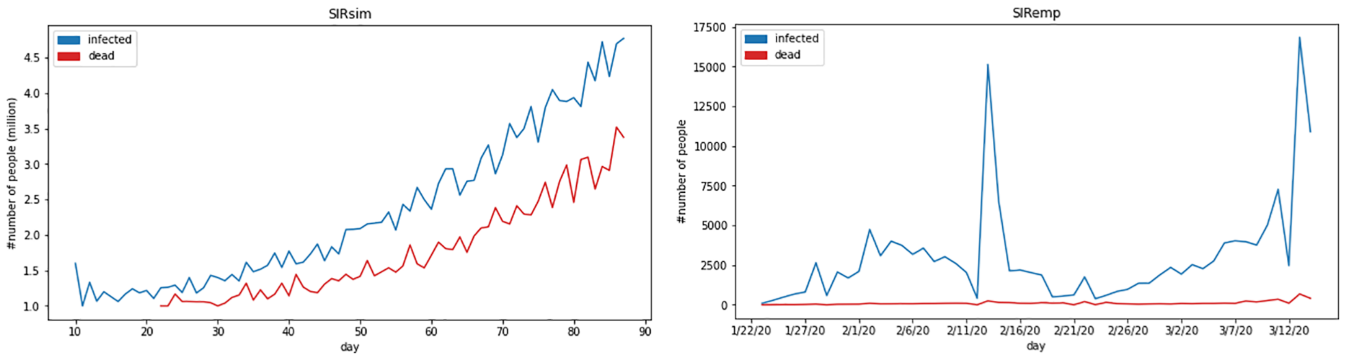


FIGURE 3 *SIRsim* network. *Blue* nodes = infected cases, *Red* nodes = removed (death) cases

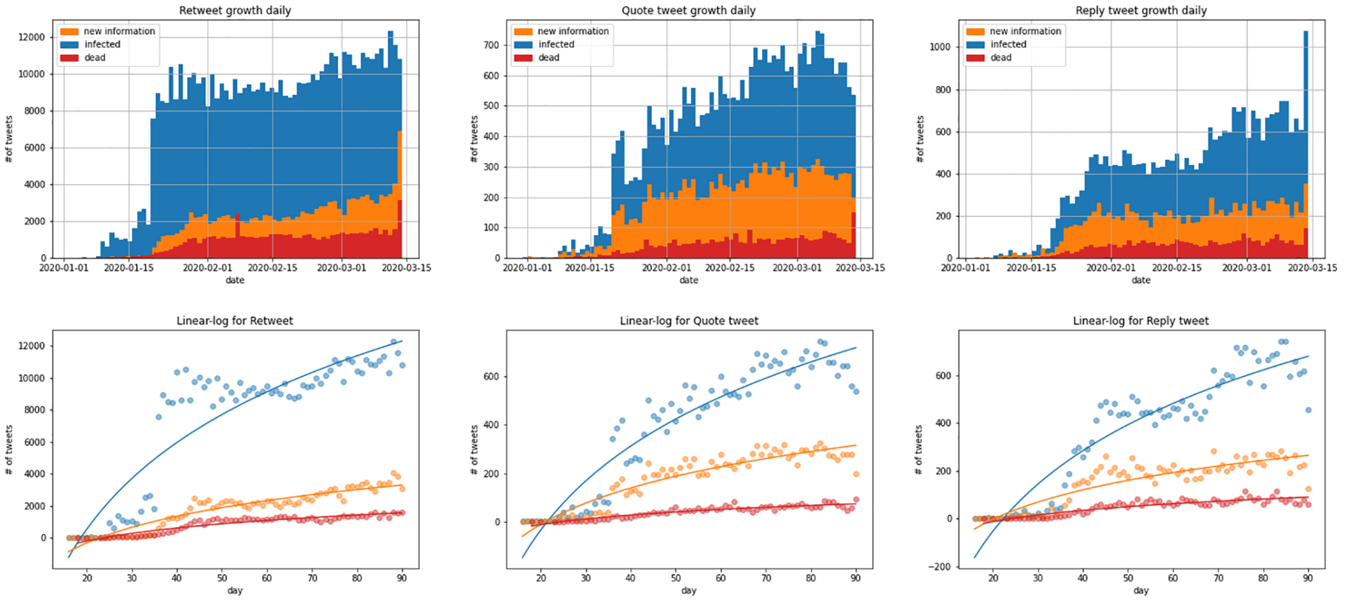


FIGURE 4 Retweet, Quote tweet, and Reply tweet growth for each day during the COVID-19 outbreak period. x-axes represent the day, y-axes represent the number of tweets. *New information* represents the original source of information, *infected* represents an interaction with another user, and *removed* represents the end of the information spread after a defined period

present in each network, in which retweet network has 23 times more cascades than quote tweet network, and 19 times more cascades than reply tweet network.

Figure 4 presents the rapid growth in tweet activities, with stark increase in retweets, quote tweets, and reply tweets during mid-January. We find that the growth distributions for all three tweet types follow a logarithmic curve. In addition, the number of infected users, equivalent to individuals spreading the information, is much higher compared to the new information consistent on the three observations. We also observe that the cascade growth for retweets is substantially higher than growth for quote tweets and reply tweets.

Table 3 shows the coefficients and parameters for each linear fit of the number of tweets to the day-period. As we can see from the table, the slope of a retweet is the highest, followed by the quote tweet and reply tweet. The slope for *removed* information is the lowest compared to the infected and new information and consistent for all cascade types. This indicates that as the number of *new information* is introduced each day, some portion of the information stops spreading.

4.3 | Correlations between *SIRsim*, *SIRemp*, and *INFOcas*

We aggregate the data from SIR-simulation over 100 iterations (*SIRsim*) and CSSE's real-infection data (*SIRemp*) and analyze correlation with Twitter's information growth (*INFOcas*) for the same time period. Table 4 shows the correlational values in terms of Pearson's correlation, for each SIR state.

TABLE 3 Linear-log regression summary for *INFOcas*

	Intercept (β_0)	Coefficient (β_1)	r^2
Retweets			
New Information	−7570.42	2414.83	0.89
Infected	−23000	7847.59	0.82
Removed	−3736.46	1176.49	0.87
Quote tweets			
New Information	−660.72	216.79	0.87
Infected	−1533.84	500.18	0.89
Removed	−181.96	56.84	0.85
Reply tweets			
New Information	−537.23	178.34	0.79
Infected	−1513.02	487.25	0.88
Removed	−217.06	68.34	0.79

For the cascades of *infected* nodes, we find the highest correlation between *SIRsim* and *INFOcas* -retweets (r = 0.86). The second-highest correlation is between retweets and quote tweets (r = 0.83). Another notable correlation is between *SIRsim* and quote tweets (r = 0.76). *SIRemp* has low correlations with all other types of cascades, with correlations ranging from 0.31 to 0.47.

TABLE 4 Correlations between *INFOcas*, *SIRsim*, *SIRemp* in terms of *Infected* and *Removed* nodes

Infected nodes	**Retweet**	**Quote tweet**	**Reply tweet**	***SIRsim***	***SIRemp***
Retweet	1	0.83	0.75	0.86	0.41
Quote tweet	0.83	1	0.65	0.76	0.39
Reply tweet	0.75	0.65	1	0.58	0.31
SIRsim	0.86	0.76	0.58	1	0.47
SIRemp	0.41	0.39	0.31	0.47	1
Removed nodes	**Retweet**	**Quote tweet**	**Reply tweet**	***SIRsim***	***SIRemp***
Retweet	1	0.83	0.79	0.66	0.53
Quote tweet	0.83	1	0.74	0.69	0.51
Reply tweet	0.79	0.74	1	0.58	0.28
SIRsim	0.66	0.69	0.58	1	0.57
SIRemp	0.53	0.51	0.28	0.57	1

In terms of the *removed* nodes cascades, there is also high correlation observed between *INFOcas* -retweets and quote tweets (r = 0.83). Retweets also have high correlation with reply tweets (r = 0.74). These two correlations show that retweet cascades are most correlated to quote tweets and reply tweets with respect to tweets that are no longer interacted with, and thus can no longer spread that particular tweet content in the network. Correlation between *INFOcas* and *SIRsim* is relatively lower (r = 0.58-0.69), showing that there is a weaker relationship between the simulated and observed Twitter's *removed* cascades. Similarly, there is a weak relationship between *SIRemp* and all *INFOcas* cascades, especially with reply tweets (r = 0.28).

5 | DISCUSSION AND CONCLUSION

Our study focuses on the diffusion patterns of COVID-19 virus itself and the information shared online about the virus. To capture the diffusion patterns of the virus, we create an SIR model (*SIRsim*) based on empirically-validated transmission dynamics of COVID-19 (e.g., reproductive ratio, incubation period), and then compare with actual confirmed cases of COVID-19 from January 22 to March 14, 2020 (*SIRemp*). To examine diffusion patterns of information discussed online about COVID-19, we construct three cascades (*INFOcas*) based on retweets, quote tweets, and reply tweets on Twitter that mentioned COVID-19 from the period of December 31st to March 14, 2020.

Our first research question asks about the diffusion patterns of COVID-19 virus, based on epidemiological assumptions of SIR. From our *SIRsim* model, we find the proportions of infected cases to be only 5% of the entire world population, and the proportions of removed (dead) cases is only 0.19% of the population. Our model accounts for 88 days since the first case of the virus, and the upward trajectory beyond linear growth suggests to us that rate of infection and deaths may increase logarithmically. This is consistent to current findings on COVID-19 that finds the distributions of infected cases follow a logarithmic distribution (Cao et al., 2020; Maier & Brockmann, 2020). (Cao et al., 2020) finds the logarithmic growth rate is suitable considering that COVID-19 is relatively in the early stage, and thus growth is slowly increasing. We also find notable differences in the simulated model and the actual confirmed cases of COVID-19 (from *SIRemp*). In fact, the distribution of *removed* cases in *SIRemp* is flat, as opposed to the increasing distribution observed in *SIRsim*. There are two reasons for the mismatch in simulated and actual distributions of *SIR* cases. The first is that our model does not take into account preventive measures such as social distancing, self-quarantine, and shelter-in-place which are found to be effective in "flattening the curve" (Lewnard & Lo, 2020; Parmet & Sinha, 2020). The second reason may be that the quantification of infection and death rates need further modifications, specifically because there is still limited testing (Ioannidis, 2020), and reporting delays (Gardner, Zlojutro, & Rey, 2020).

The second research question asks about the diffusion patterns of information cascades on Twitter about COVID-19. We construct retweet cascade, quote tweet cascade, and reply cascade (we call these *INFOcas*) to fully capture the different types of interactions between users on Twitter. All three cascades show strong fit with linear-log distribution, suggesting a power-law decay in the diffusion of new information about COVID-19 over time. With this finding along with the cascade length of each tweet type, we expect that retweet cascade decays at the fastest rate, given that its cascade length is only approximately 4 hours. On the other hand, we find quote tweets' average cascade length to be about 3 days, which means that each original tweet that has been interacted with via quotes has longer duration in terms of activity. This is also observed for reply tweets, where the average cascade length is about 2 days.

The third research question focuses on the correlation in diffusion patterns between *SIRsim*, *SIRemp*, and *INFOcas* to address the connection between epidemic and information diffusion dynamics. Based on the examination of *infected* cascades, we find the stronger positive correlation between *SIRsim* and *INFOcas* -retweets (r = 0.86), and quote tweets (r = 0.76). On the other hand, we observe low correlations between *SIRemp* and all three *INFOcas* types (r = 0.31-0.41). This shows that the distribution of infected agents are more correlated between *INFOcas* and *SIRsim*, and not so much with *SIRemp*. With the rapid spread dynamics seen in *SIRsim*, this correlation shows that tweets about COVID-19 gets retweeted most quickly, then followed by quote tweets, and then reply tweets. The correlation between *SIRsim* and *SIRemp* is relatively low (r = 0.47), which may indicate that either the simulated model potentially overestimates the *infection* rate, or that the actual reported cases may underestimate the *infection* rate. For the *removed* cascades, we find strongest correlations between *INFOcas* cascades, specifically between retweets and quote tweets (r = 0.83), retweets and reply tweets (r = 0.79), and quote tweets and reply tweets (r = 0.74). We find weaker correlations between *INFOcas* and *SIRsim* (r = 0.58-0.69), and weakest correlations between *INFOcas* and *SIRemp* (r = 0.28-0.53). This result is consistent with our observation that the *removed* distribution on *SIRemp* is more uniform and flat compared to other distributions. It is also expected that the *removed* distribution for *INFOcas* would be different from *SIRsim*, given that the likelihood of tweets to transition from *infected* to *removed* is notably higher.

Overall, we find complex relationships between diffusion dynamics about COVID-19 from the simulated virus spread model, the actual reported cases of the virus spread, and the information shared and discussed online. Our study demonstrates how epidemic modeling, in combination with examining information cascades about the virus can help capture the many activities surrounding the COVID-19 pandemic. In future work, we hope to expand our data collection to more recent dates, given the constantly-changing nature of the pandemic. Additionally, we aim to improve our simulated epidemic model (*SIRsim*) to include additional control variables that reflects prevention strategies, namely social distancing, self-quarantine, and shelter-in-place.

ENDNOTES

[1] http://ccl.northwestern.edu/netlogo

[2] Netlogo model #4286

[3] World population, https://www.worldometers.info/world-population/

[4] https://www.space.com/17638-how-big-is-earth.html

[5] https://www.fhwa.dot.gov/ohim/onh00/bar8.htm, updated March 29, 2018

[6] Crimson Hexagon, https://forsight.crimsonhexagon.com/

[7] General Data Protection Regulation (GDPR), https://gdpr-info.eu/

REFERENCES

Abdullah, S. & Wu, X. (2011). An epidemic model for news spreading on twitter. In *2011 IEEE 23rd International Conference on Tools With Artificial Intelligence* (pp. 163–169). IEEE.

Cao, Z., Zhang, Q., Lu, X., Pfeiffer, D., Jia, Z., Song, H., & Zeng, D. D. (2020). Estimating the effective reproduction number of the 2019-ncov in China. *medRxiv*.

CDC (2020). *Coronavirus Disease 2019*. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

Cha, M., Mislove, A., Adams, B., & Gummadi, K. P. (2008). Characterizing social cascades in Flickr. In *Proceedings of the First Workshop on Online Social Networks* (pp. 13–18).

Dinh, L. & Parulian, N. (2020). *COVID19 datasets to examine diffusion patterns*. doi: https://doi.org/10.6084/m9.figshare.12465383

Gardner, L., Zlojutro, A., & Rey, D. (2020). *Modeling the spread- ing risk of 2019-ncov*. Baltimore, MD: Center for Systems Science and Engineering, Johns Hopkins University.

Ghebreyesus, T. A. (2020, March 3). WHO Director-General's opening remarks at the media briefing on COVID-19.

Goel, A., Munagala, K., Sharma, A., & Zhang, H. (2015). A note on modeling retweet cascades on Twitter. In International Workshop on Algorithms and Models for the Web-Graph (pp. 119–131). Springer.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, *12*(3), 211–223.

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, *83*(6), 1420–1443.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 491–501).

Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM Sigmod Record*, *42*(2), 17–28.

Ioannidis, J. P. (2020). A fiasco in the making? as the coronavirus pandemic takes hold, we are making decisions without reliable data. *Stat 17*.

JHU-CSSE (2020). *Coronavirus 2019-nCoV Global Cases by Johns Hopkins CSSE*. Retrieved from https://github.com/CSSEGISandData/COVID-19.

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis* (pp. 1–9).

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, *115*(772), 700–721.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., … Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, *172*(9), 577–582.

Lerman, K. (2016). Information is not a virus, and other consequences of human cognitive limits. *Future Internet*, *8*(2), 21.

Lerman, K. & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Lewnard, J. A., & Lo, N. C. (2020). Scientific and ethical basis for social-distancing interventions against Covid-19. *The Lancet*, *20* (6), 631.

Liu, C., & Zhang, Z.-K. (2014). Information spreading on dynamic social networks. *Communications in Nonlinear Science and Numerical Simulation*, *19*(4), 896–904.

Maier, B. F. & Brockmann, D. (2020). Effective containment explains sub-exponential growth in confirmed cases of recent Covid-19 outbreak in mainland china. *arXiv preprint arXiv:2002.07572*.

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using twitter bots. *PloS One*, *12*(9), 1–12.

Nekovee, M., Moreno, Y., Bianconi, G., & Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, *374*(1), 457–470.

Parmet, W. E., & Sinha, M. S. (2020). Covid-19—the law and limits of quarantine. *New England Journal of Medicine*, *382*(15), e28.

Seki, K. & Nakamura, M. (2016). The collapse of the friendster network started from the center of the core. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 477–484). IEEE.

Ver Steeg, G., Ghosh, R., & Lerman, K. (2011). What stops social epidemics? In *Fifth International AAAI Conference on Weblogs and Social Media*.

Wu, F., Huberman, B. A., Adamic, L. A., & Tyler, J. R. (2004). Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, *337*(1-2), 327–335.

Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in China: summary of a report of 72 314 cases from the Chinese center for disease control and prevention. *JAMA*, *323*(13), 1239–1242.

Xie, B., He, D., Mercer, T., Wang, Y., Wu, D., Fleischmann, K. R., et al. (2020). Global health crises are also information crises: A call to action. *Journal of the Association for Information Science and Technology*, 1–5.

Yang, J. & Leskovec, J. (2010). Modeling information diffusion in implicit networks. In 2010 IEEE International Conference on Data Mining (pp. 599–608). IEEE.

How to cite this article: Dinh L, Parulian N. COVID-19 pandemic and information diffusion analysis on Twitter. *Proc Assoc Inf Sci Technol*. 2020;57:e252. https://doi.org/10.1002/pra2.252

APPENDIX A

TABLE A1 List of 36 major cities used in *SIRsim* model, and their associated coordinates (in pixels)

Major City	x-coordinate (in pixels)	x-coordinate (in pixels)
Tokyo	257	6
New Delhi	135	-13
Seoul	232	7
Shanghai	216	-7
Mumbai	127	-33
Mexico City	-221	-28
Beijing	208	14
Sao Paulo	-112	-113
Jakarta	194	-85
New York City	-165	20
Karachi	115	-19
Osaka	247	3
Manila	219	-39
Cairo	44	-9
Dhaka	159	-23
Los Angeles	-254	5
Moscow	49	64
Buenos Aires	-137	-143
Kolkata	151	-24
London	-22	50
Bangkok	180	-42
Lagos	-9	-55
Istanbul	40	16
Rio de Janeiro	-104	-112
Tehran	83	4
Guangzhou	205	-21
Kinshasa	15	-78
Shenzhen	202	-23
Lahore	127	-3
Rhine-Ruhr	-4	48
Tianjin	211	9
Bengaluru	133	-44
Paris	-14	38
Chennai	136	-43
Hyderabad	134	-37
Wuhan	205	-10