Descendant distributions for the impact of mutant contagion on networks

Jonas S. Juul$^{1,2,*}$ and Steven H. Strogatz$^{2,†}$

$^{1}$Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, Copenhagen 2100-DK, Denmark
$^{2}$Center for Applied Mathematics, Cornell University, Ithaca, New York 14853, USA

(Received 7 January 2020; accepted 3 June 2020; published 1 July 2020)

Contagion, broadly construed, refers to anything that can spread infectiously from peer to peer. Examples include communicable diseases, rumors, misinformation, ideas, innovations, bank failures, and electrical blackouts. Sometimes, as in the 1918 Spanish flu epidemic, a contagion mutates at some point as it spreads through a network. Here, using a simple susceptible-infected model of contagion, we explore the downstream impact of a single mutation event. Assuming that this mutation occurs at a random node in the contact network, we calculate the distribution of the number of “descendants,” $d$, downstream from the initial “patient zero” mutant. We find that the tail of the distribution decays as $d^{-2}$ for complete graphs, random graphs, small-world networks, networks with block-like structure, and other infinite-dimensional networks. This prediction agrees with the observed statistics of memes propagating and mutating on Facebook and is expected to hold for other effectively infinite-dimensional networks, such as the global human contact network. In a wider context, our approach suggests a possible starting point for a mesoscopic theory of contagion. Such a theory would focus on the paths traced by a spreading contagion, thereby furnishing an intermediate level of description between that of individual nodes and the total infected population. We anticipate that contagion pathways will hold valuable lessons, given their role as the conduits through which single mutations, innovations, or failures can sweep through a network as a whole.

DOI: 10.1103/PhysRevResearch.2.033005

I. INTRODUCTION

The concept of contagion began in epidemiology, where it was used to describe the spread of disease between people in close contact. Nowadays contagion has taken on a broader meaning; it refers to any sort of process that can spread infectiously from node to node through a network [1–7]. Along with communicable diseases [8–15], examples of contagions include rumors [16], misinformation [17], ideas [18], innovations [19–21], bank failures [22], and electrical blackouts [23].

When a contagion spreads, it propagates from one or more “parent” nodes to a number of “descendant” nodes. Enumerating the descendants in all the paths stemming from a parent can reveal important and useful information. In particular, suppose a contagion mutates into a more pernicious form as it travels. Then counting its descendants would tell us how many nodes will be confronted by this nastier strain. A mutation event of this sort occurred in 1918 and gave rise to the Spanish flu epidemic that killed millions of people worldwide [24]. Similar (but less consequential) mutations happen online when users modify memes to make them funnier or stickier before sharing them with their peers [25].

Here, we derive exact results for the impact of a single mutation event, assuming the contagion dynamics are governed by the so-called susceptible-infected (SI) model. Our goals are to understand, in a statistical sense, how many nodes will ultimately get infected by the mutant strain and to clarify how the results depend on the structure of the underlying contact network.

II. DESCENDANT DISTRIBUTIONS

To make analytical progress, we consider an extremely simplified model in which each node is either susceptible or permanently infected (Fig. 1). This SI model effectively assumes infinite transmissibility of the contagion and ignores the possibility of recovery, death, migration, vaccination, temporary immunity, latency periods, heterogeneity of susceptibility and infectiousness, and many other realistic considerations. All of these would make for interesting extensions of our work.

As the contagion spreads [Fig. 1(a)], we record which nodes caught it from which, and plot the resulting paths of infection as an epidemic tree [Fig. 1(b)]. Then we count how many nodes would be affected by a mutation occurring at a random “patient zero” node. In the example shown in Fig. 1(c), the mutant infection occurs at node $B$ and is passed along to the two nodes below it. Of course, if the mutation had occurred elsewhere, it could have produced either more descendants (e.g., three descendants, had the mutation occurred at $A$) or fewer (zero descendants, had it occurred at $C$). Thus, the natural statistical quantity to study is the distribution...
of the number of descendants, aggregated over all possible patient-zero nodes.

In one sense, the dynamics assumed here are trivial: one node after another gets infected until no susceptibles remain. But what is not trivial are the descendant distributions implied by the model, because they also depend on the network’s structure.

One limiting case is already understood. In a completely structureless, well-mixed population, the impact of a single mutation can be quantified by the classical stochastic process known as the Yule process. In that case, the probability that a mutant generates exactly \( d \) descendants is

\[
P_d = \frac{1}{(d+2)(d+1)}. \quad (1)
\]

To the best of our knowledge, however, it has been an open problem to extend this result to structured populations.

To learn what to expect, we first compute descendant distributions numerically from Monte Carlo simulations [26]. For a given random realization of the SI contagion process on a given network, like the one shown in Fig. 1(b), we count the number of descendants of each node and compile a histogram. This histogram, however, merely gives the descendant distribution for one realization of the stochastic dynamics. To extract a more robust statistical measurement, we average over the random location of the initially infected seed node, as well as the random decisions of whom to infect at each step, to obtain an average descendant distribution.

Figure 2 shows the average descendant distribution for the simplest possible network structure: a complete graph, in which each node is connected to all the others. The downward slope of the plot indicates that many nodes have few descendants, and a few nodes have many descendants. Of course, the seed \( O \) has every other node as its descendant, as an artifact of the assumed initial conditions. Its corresponding data point in Fig. 2 lies off the curve for this reason.

The most striking feature of the descendant distribution in Fig. 2 is its apparent power-law decay for \( d \gg 1 \). To explain this scaling law intuitively, recall that one way of getting power-law distributions is through rich-get-richer effects [27–30], and observe:

(i) if node \( i \) infected node \( j \), the ancestors of \( j \) will be \( i \) and all the ancestors of \( i \);

(ii) a node \( i \) can acquire a new descendant \( j \) if it passes the infection on to \( j \), or if one of its descendants passes the infection on to \( j \).

The first point means that our model contagion process is equivalent to a network that grows by node copying [31].
The second point suggests that the probability of a node acquiring more descendants should grow, loosely speaking, in proportion to the number of descendants it already has, thereby making the rich richer.

To sharpen this intuition, we calculate the descendant distribution $P_d$ analytically for some exactly solvable networks [32]. First, for a complete graph in the limit $N \to \infty$, we recover the classical result of Yule,

$$P_d = \frac{1}{(d+2)(d+1)}. \quad (2)$$

This result was also found by Krapivsky and Redner for the in-degree distribution of networks growing by node copying [31]. Figure 2 shows that this result agrees well with our simulation data. For further discussion of the connection between the Yule process and rich-get-richer effects, see Ref. [33].

Likewise, for several classes of random networks, the descendant distributions can be derived in the limit of infinite network size. For $z$-regular configuration models and Erdős–Rényi random graphs with average degree $z$, we obtain [32] the infinite-$N$ solution

$$P_d = \frac{z-1}{z-2} B\left(\frac{z-1}{z-2} + d, 2\right). \quad (3)$$

where $B(a, b)$ denotes the beta function. Importantly, this expression reduces to the complete-graph solution for large values of $z$.

More complicated network structures yield similar results [32]. For networks consisting of equally sized Erdős–Rényi “blocks” with mean degree $z$, and with probability $p$ of connecting each node to a node chosen uniformly at random from nodes located in other blocks, we obtain the solution

$$P_d = \frac{z-1 + 2p}{z-2 + 2p} B\left(\frac{z-1 + 2p}{z-2 + 2p} + d, 2\right). \quad (4)$$

Figure 3 shows the simulation results for $z$-regular configuration models, Erdős–Rényi random graphs, and modular networks with block structure, all of size $N = 10^4$. When plotted in a manner suggested by Eqs. (3) and (4), the simulation data for the different random networks collapse onto a single curve (Fig. 3), consistent with the analytical approximation.

Finally, for a small-world network created by inserting random shortcuts in a ring lattice, with probability $p$ of connecting each node with another node chosen uniformly at random [34], the analytical solution [32] is

$$P_d = \frac{4p + 1}{4p} B\left(\frac{4p + 1}{4p} + d, 2\right). \quad (5)$$

This result agrees well with simulations; see Fig. 1 in the Supplemental Material [32].

III. SCALING LAW FOR THE TAIL

All the descendant distributions we have calculated so far turn out to decay asymptotically according to the same power law:

$$P_d \propto d^{-2} \quad (6)$$
over 10^3 simulations for each cascade size. The smallness of the predefined cascade sizes 100, 400, and 2000 and are averaged uniformly at random. The descendant distributions shown here have the realization, and started a new simulation with a seed chosen we stopped the spreading, obtained the descendant distribution for until exactly a predefined number of nodes were infected. Then started a contagion at a random seed node and ran the simulation together, this subnetwork contains 4039 nodes and 88,234 edges. We dimensional as

\[
\Delta \rightarrow \infty
\]

on a network of merged Facebook ego networks [35,36]. Taken together, this subnetwork contains 4039 nodes and 88,234 edges. We started a contagion at a random seed node and ran the simulation until exactly a predefined number of nodes were infected. Then we stopped the spreading, obtained the descendant distribution for the realization, and started a new simulation with a seed chosen uniformly at random. The descendant distributions shown here have predefined cascade sizes 100, 400, and 2000 and are averaged over 10^3 simulations for each cascade size. The smallness of the subnetwork gives rise to conspicuous finite-size effects in the tail of the distribution. Apart from these effects, the descendant distribution falls on the curve expected for highly connected infinite-dimensional networks, here illustrated with the analytical solution for the descendant distribution for contagion on the complete graph.

for \(d \gg 1\). Further analysis [32] indicates that this inverse-square scaling follows from a property that the complete graph shares with the random networks: they all become infinite dimensional as \(N \to \infty\). (Here, we consider a network to be infinite dimensional if the area and volume of a ball of a given radius \(s\) grow equally fast with \(s\); in this context, a ball of radius \(s\) is defined as the set of all nodes within \(s\) hops of a given node. See Sec. V in the Supplemental Material for details [32] as well as Refs. [37,38] for further discussion of the concept of network dimension.)

On this basis, we expect that the same \(d^{-2}\) scaling should hold for other infinite-dimensional networks, but not for one-dimensional chains, two-dimensional grids, three-dimensional lattices, or other networks whose dimensionality remains finite and sufficiently small as the number of nodes tends to infinity. In some sense, this expectation is natural: there are well-known analogies between epidemic models and percolation models, and for many of these, the critical properties vary with dimension for a range of intermediate dimensions and then agree with mean-field theory above some upper critical dimension [11,13,39]. Simulations of the model contagion on two-dimensional square grids support this predicted dependence on dimension: descendant distributions deviate significantly from the \(d^{-2}\) scaling [32]. Interestingly, scale-free networks also show a departure from the scaling predicted above, but their descendant distributions merge with our predictions past a crossover, producing the same inverse-square decay in the tail [32].

Conveniently, many real-world networks are effectively infinite dimensional. Consider the social network Facebook, which as of June 2019 had more than 2.4 billion active users. In a fascinating study, Adamic et al. [25] examined memes spreading from friend to friend on the Facebook social graph. Typically, memes would propagate from one user to another without being altered, but occasionally a user would change the content of the meme before resharing it. This would make a new variant of the meme, which would then spread on the network along with previously existing copies. Adamic et al. [25] examined the frequency of different variants of rarely changing memes and found that the frequency distribution of the most widely shared variants followed an inverse-square law. Specifically, they found the exponent to be \(-2.01 \pm 0.15\). This exponent matched the prediction of a mean-field model (the Yule process), but it remained unclear why a model without any underlying network structure could account for the exponent obtained from the actual Facebook network.

Our work suggests that the observed exponent of \(-2\) is a consequence of the approximate infinite-dimensionality of the Facebook network. Indeed, Fig. 4 shows that when we simulate our simple contagion process on a small subnetwork of Facebook [35,36], the resulting descendant distributions match what we would expect for highly connected infinite-dimensional networks. In particular, apart from effects caused by the small size of the subnetwork, an approximate power-law tail with a slope close to \(-2\) emerges.

IV. DISCUSSION

The epidemic trees analyzed in this paper, along with their associated pathways of contagion, have been studied previously in diverse disciplines. They have been called adoption paths [20], dissemination trees [40,41], spreading patterns [42], causal trees of disease transmission [43], diffusion structure patterns [44], the structure of diffusion events [45], and epidemic trees [46]. We have chosen to adopt the term “epidemic trees,” although it comes with a significant caveat: Generally, the graph of the propagation paths for a contagion need not be a directed tree; in the case of a complex contagion [47], where each child node has two or more parents, the graph could be a directed graph with no cycles. But for the simple contagions studied here, where each child is assumed to have only one parent, the graph of the propagation paths is always a tree.

Although epidemic trees have been examined previously in specific data sets, their statistical properties have not been analyzed theoretically until now. We regard our results in that direction as among the main contributions of this paper. In a wider context, our approach suggests a possible starting point for a mesoscopic theory of contagion, in which infection pathways, epidemic trees, and descendant distributions would play the leading role, operating at a scale in between the local level of individual nodes and the global level of the entire network.
To clarify these distinctions among the microscopic, mesoscopic, and macroscopic scales, consider the transition to a giant component in a susceptible-infected-removed (SIR) model of contagion on a network [1,13]. Above the transition, there exists a giant infected component of size proportional to $N$. Such macroscopic phenomena have been extensively and fruitfully studied in the literature on network contagion [1,2,7,13]. But giant component sizes and other macroscopic quantities lump all infected nodes together and thus discard information about which nodes infected which. Such causal information is retained in epidemic trees, which show the transmission pathways of contagion and thereby shed light on phenomena operating at the mesoscopic level.

These mesoscopic considerations inescapably come into play (at least for mutant contagions on infinite-dimensional networks) because the descendant distribution is a beta function with a $d^{-2}$ tail, as we have shown above. A consequence of this inverse-square scaling is that the expected size of the mutant infected component is of mesoscopic size comparable to $\log N$ for $N \gg 1$ and hence is intermediate in a precise sense; it is large compared with the $O(1)$ scale of individual nodes, but small compared with the $O(N)$ scale of the network itself, and of the giant infected (but nonmutated) component. Note, however, that the variance of this smaller network itself, and of the giant infected (but nonmutated) component, even more so. One practical implication is that we expect mutant strains of contagion to infect large fractions of network nodes occasionally.

We expect that notions like contagion pathways, epidemic trees, and descendant distributions are just the beginning of a mesoscopic theory of contagion. Much remains to be discovered about the geometry and statistics of these and other quantities, both empirically for real contagions and theoretically for a wide range of infection dynamics and network structures. Understanding this middle ground might also have practical benefits for the control of contagion processes, in contexts ranging from vaccination strategies for communicable diseases to methods for combating the spread of misinformation on social media.

But before such practical benefits can be realized, any future mesoscopic theory will also need to incorporate several realistic features that we have left out of the current model and analytical treatment. These include the extension to heterogeneities in degree, susceptibility, infectiousness, latency period, and so on. Such heterogeneities have shown themselves to be important in the COVID-19 outbreak [48,49] and are also thought to play a crucial role in the spread of many other infectious diseases [50,51]. Handling these heterogeneities theoretically will require extending the analytical treatment to a more sophisticated framework, like quenched mean-field theory [8,15].

**ACKNOWLEDGMENTS**

J.S.J. acknowledges funding through the University of Copenhagen UCPH 2016 Excellence Programme for Interdisciplinary Research and the Danish Council for Independent Research and thanks the Center for Applied Mathematics at Cornell University for hospitality while this work was carried out. S.H.S. was supported by NSF Grant CCF-1522054.

[1] M. E. J. Newman, Networks (Oxford University Press, New York, 2018).
[2] D. J. Watts, Proc. Natl. Acad. Sci. USA 99, 5766 (2002).
[3] P. S. Dodds and D. J. Watts, Phys. Rev. Lett. 92, 218701 (2004).
[4] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. 81, 591 (2009).
[5] N. A. Christakis and J. H. Fowler, N. Engl. J. Med. 357, 370 (2007).
[6] H. Sayama and R. Sinatra, Phys. Rev. E 91, 032809 (2015).
[7] S. Lehmann and Y.-Y. Ahn, Complex Spreading Phenomena in Social Systems (Springer International Publishing, Cham, Switzerland, 2018).
[8] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. 86, 3200 (2001).
[9] R. M. Anderson and R. M. May, Infectious Diseases of Humans: Dynamics and Control (Oxford University Press, West Sussex, Great Britain, 1991).
[10] M. J. Keeling and P. Rohani, Modeling Infectious Diseases in Humans and Animals (Princeton University Press, Princeton, New Jersey, 2011).
[11] M. De Domenico, C. Granell, M. A. Porter, and A. Arenas, Nat. Phys. 12, 901 (2016).
[12] P. G. Fennell and J. P. Gleeson, SIAM Rev. 61, 92 (2019).
[13] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. 87, 925 (2015).
[14] C. Castellano and R. Pastor-Satorras, Phys. Rev. Lett. 105, 218701 (2010).
[15] I. Z. Kiss, J. C. Miller, and P. L. Simon, Mathematics of Epidemics on Networks (Springer International Publishing, Cham, Switzerland, 2017).
[16] D. J. Daley and D. G. Kendall, IMA J. Appl. Math. 1, 42 (1965).
[17] S. Vosoughi, D. Roy, and S. Aral, Science 359, 1146 (2018).
[18] L. M. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, and C. Castillo-Chávez, Physica A (Amsterdam, Neth.) 364, 513 (2006).
[19] A. Mellor, M. Mobilia, S. Redner, A. M.ucklidge, and J. A. Ward, Phys. Rev. E 92, 012806 (2015).
[20] J. S. Juul and M. A. Porter, Phys. Rev. E 99, 022313 (2019).
[21] P. L. Krapivsky, S. Redner, and D. Volovik, J. Stat. Mech. (2011) P12003.
[22] A. G. Haldane and R. M. May, Nature (London) 469, 351 (2011).
[23] C. D. Brummitt, R. M. D’Souza, and E. A. Leicht, Proc. Natl. Acad. Sci. USA 109, E680 (2012).

033005-5
[24] V. Andreasen, C. Viboud, and L. Simonsen, *J. Infect. Dis.* **197**, 270 (2008).
[25] L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng, in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (ACM Press, New York, 2016), pp. 473–482.
[26] All scripts necessary to reproduce the simulated results are available at https://sid.erda.dk/wsgi-bin/ls.py?share_id=F8JmKnQryb.
[27] G. U. Yule, *Philos. Trans. R. Soc. Lond. B* **213**, 21 (1925).
[28] H. A. Simon, *Biometrika* **42**, 425 (1955).
[29] D. De Solla Price, *J. Am. Soc. Inf. Sci.* **27**, 292 (1976).
[30] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
[31] P. L. Krapivsky and S. Redner, *Phys. Rev. E* **71**, 036118 (2005).
[32] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.2.033005 for details of the analytical derivations, additional simulations, and further discussion of the breakdown of the assumption of infinite-dimensionality.
[33] A. Pachon, F. Polito, and L. Sacerdote, *J. Stat. Phys.* **162**, 1608 (2016).
[34] M. E. J. Newman and D. J. Watts, *Phys. Lett. A* **263**, 341 (1999).
[35] J. Leskovec and J. J. McAuley, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, 2012), pp. 539–547.
[36] J. Leskovec and A. Krevl, SNAP Datasets: Stanford Large Network Dataset Collection, http://snap.stanford.edu/data (2014).
[37] T. Nowotny and M. Requardt, *J. Phys. A: Math. Gen.* **31**, 2447 (1998).
[38] O. Shanker, *Mod. Phys. Lett. B* **21**, 639 (2007).
[39] D. Ben-Avraham and S. Havlin, *Diffusion and Reactions in Fractals and Disordered Systems* (Cambridge University Press, Cambridge, 2000).
[40] S.-W. Oh and M. A. Porter, *Chaos* **28**, 033101 (2018).
[41] D. Liben-Nowell and J. Kleinberg, *Proc. Natl. Acad. Sci. USA* **105**, 4633 (2008).
[42] S. M. Jang, T. Geng, J.-Y. Q. Li, R. Xia, C.-T. Huang, H. Kim, and J. Tang, *Comput. Hum. Behav.* **64**, 103 (2018).
[43] A. Vázquez, in *Discrete Methods in Epidemiology* (DIMACS, Rutgers University, Piscataway, NJ, 2004), pp. 163–180.
[44] Z.-K. Zhang, C. Liu, X.-X. Zhan, X. Lu, C.-X. Zhang, and Y.-C. Zhang, *Phys. Rep.* **651**, 1 (2016).
[45] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, *Manag. Sci.* **62**, 180 (2015).
[46] D. T. Haydon, M. Chase-Topping, D. Shaw, L. Matthews, J. Friar, J. Wilesmith, and M. Woolhouse, *Proc. R. Soc. London Ser. B* **270**, 121 (2003).
[47] B. Mønsted, P. Sapieżyński, E. Ferrara, and S. Lehmann, *PLoS ONE* **12**, e0184148 (2017).
[48] Q. J. Leclerc, N. M. Fuller, L. E. Knight, S. Funk, G. M. Knight, C. C.-W. Group et al., *Wellcome Open Res.* **5**, 83 (2020).
[49] D. Miller, M. A. Martin, N. Harel, T. Kustin, O. Tiros, M. Meir, N. Sorek, S. Gefen-Halevi, S. Amit, O. Vorontsov et al., medRxiv (2020).
[50] A. P. Galvani and R. M. May, *Nature (London)* **438**, 293 (2005).
[51] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, *Nature (London)* **438**, 355 (2005).