Featured in Physics

Predicting the Speed of Epidemics Spreading in Networks

Sam Moore[*] and Tim Rogers[†]

*Centre for Networks and Collective Behaviour, Department of Mathematical Sciences, University of Bath,
Bath, England BA2 7AY, United Kingdom*

Global transport and communication networks enable information, ideas, and infectious diseases to now spread at speeds far beyond what has historically been possible. To effectively monitor, design, or intervene in such epidemic-like processes, there is a need to predict the speed of a particular contagion in a particular network, and to distinguish between nodes that are more likely to become infected sooner or later during an outbreak. Here, we study these quantities using a message-passing approach to derive simple and effective predictions that are validated against epidemic simulations on a variety of real-world networks with good agreement. In addition to individualized predictions for different nodes, we find an overall sudden transition from low density to almost full network saturation as the contagion progresses in time. Our theory is developed and explained in the setting of simple contagions on treelike networks, but we are also able to show how the method extends remarkably well to complex contagions and highly clustered networks.

It took more than nine years for the Black Death to spread across Europe. Progress of this devastating outbreak of bubonic plague was limited by 14th century travel networks to an average daily dispersion of approximately 1.5 km [1]. In frightening contrast, the recent Zika outbreak in South America was found to spread with an average daily dispersion of 42 km, rising as high as 634 km in the most densely populated parts of Brazil [2]. This extraordinary difference is indicative of a mobile society that is no longer rigidly bound by spatial structure, making the relevant notion of distance network based rather than geographic. Similarly, in the highly connected domain of social media, the spread of concepts, memes, and hashtags can be explosive. One recent empirical study of the dynamics of online rumour cascades—often reaching tens of thousands of users in a matter of days—made the worrying finding that false information spreads faster than true [3]. It takes little imagination to see how an understanding of propagation speeds in modern networks would have, in the digital case, great commercial and political benefit and, in the physical case, be invaluable in planning outbreak prevention, monitoring, and response.

The field of network epidemiology [4–7] has developed a wide spectrum of techniques for the analysis of spreading processes. One approach to the problem of spreading speed is through numerical simulations (see, e.g., [8]), which yield useful results on small scales but, for increasingly large complex networks, may prove slow and impractical. Alternative approximations have been made by considering only the most probable path between a given target node and the source [9]. It is known that this shortest-path approach can significantly overestimate the infection arrival times [10], but to take into account all possible

paths would soon be infeasible because their numbers typically grow exponentially with the number of vertices in the network. One promising idea is a conjectured connection between centrality measures and infection arrival time [11], which so far has only been tested numerically.

Although global networks of interest are highly connected, they are also typically sparse in the sense that individuals usually interact with a number of others that is very small relative to the total population size. Exploitation of this sparse network structure has been a key tool in network epidemiology: in particular, via the message-passing approach pioneered in [12]. This technique has allowed for efficient characterization of the epidemic (percolation) threshold [13,14], and it gave rise to the new notion of non-backtracking centrality [15]. In [16,17], a message-passing approach was used to make individualized predictions for node responses to spreading processes, giving a physical interpretation of non-backtracking centrality as the probability for a node to appear in the percolating cluster. None of these works has yet addressed the important questions of how fast an epidemic will spread in a given network and which nodes may fall victim first.

Here, we seek to assess the full time dependence of an epidemic outbreak in order to characterize the speed of spread in a given network by calculating the mean delay in infection between nodes at different graph distances from the source. Technically, we achieve this through a saddle-point analysis of the left tail of the distribution of time to infection expressed via the message-passing equations. This method enables us to find the overall speed of an infection in a network and to show that the arrival time at a node is accurately predicted by the logarithm of its non-backtracking centrality.

Our theoretical predictions for both spreading speed and arrival times show excellent agreement with numerical simulations performed on real-world networks, even in the case of highly clustered contact networks with heavy-tailed degree distributions. Remarkably, we show that the method can also be extended to complex threshold models of contagions in which a node must be exposed to multiple infective neighbors before acquiring the contagion itself. We finish by observing that the time for the infection to spread through the bulk of the network is independent of network size, implying an almost instantaneous jump from a low to a high density of infection when time is properly scaled; a property that we show to be common to time-ordered percolation in general.

Speed of spread.—We begin by considering a simple susceptible-infected contagion spreading on a sparse network starting from a single infected node (details of the extension to other models are found in the Supplemental Material [18]). When node i becomes infectious, it transmits the infection to a neighbor j after a delay $X_{i \to j}$: a random variable drawn from a distribution with density $f(x)$, independent from any other event. The choice of an exponential distribution for f would correspond to Markov disease dynamics, although it has been shown that real-world contagion dynamics differ substantially from this simple case [19–22]; hence, we study general distributions of the transmission time.

Write T_i^n for the length of the shortest (temporal) path to a node at distance n from i, and $T_{i \to j}^n$ for the shortest such path for which the first step is to node j. It follows that $T_i^n = \min_{j \in \partial i} T_{i \to j}^n$, where ∂i denotes the set of neighbors of i. More generally, $T_{i \to j}^n$ decomposes as

$$T_{i \to j}^n = X_{i \to j} + \min_{k \in \partial j \setminus i} T_{j \to k}^{n-1}. \tag{1}$$

Writing $F_{i \to j}^n(t)$ for the probability that $T_{i \to j}^n$ is less than t, we arrive at the message-passing equation

$$F_{i \to j}^n(t) = \int_0^t f(x) \left(1 - \prod_{k \in \partial j \setminus i} [1 - F_{j \to k}^{n-1}(t - x)] \right) dx. \tag{2}$$

In writing the above, we have assumed independence between the variables $\{T_{j \to k}^{n-1}\}$; although this technically only holds for tree graphs, we will see that the approximation is effective for a broad class of real-world networks.

Equation (2) represents a nested hierarchy of expressions that could in principle be solved numerically for a given network, infection, and source node. However, this process is computationally intensive, and the results are not generalizable. We will pursue a different path and investigate the structure of the dynamics described by Eq. (2) to reveal useful general insights.

At first glance, it appears that the spreading process depends in a complicated way on the precise layout of the network; however, we find that the system possesses a
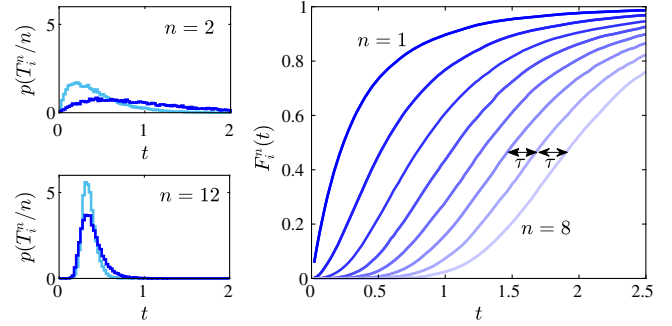

FIG. 1. Left panels: simulation of distribution of scaled time T_i^n/n for an epidemic to reach distance n from a source node i chosen to have degree 1 (dark) or degree 3 (pale); as $n \to \infty$, these distributions will converge to delta functions at some value τ. Right panel: simulation of $F_i^n(t)$ for time to reach distance n from a source node i chosen to have degree 1, showing convergence to a standard form with a fixed offset τ. In both cases, node-to-node transmission times are standard exponentials and the network is an Erdős-Rényi graph with mean degree 3 on $N = 10^4$ nodes.

regularity that emerges after a few iterations. In a network of $N \gg 1$ nodes, for $1 \ll n \ll N$, we observe the convergence $T_i^n/n \to \tau$ for some constant τ, describing the delay between spreading $n - 1$ steps from the source to n. In this sense, $1/\tau$ can be interpreted as the speed of spreading in the network. This effect is illustrated in the left panels of Fig. 1, showing the convergence and reduction of variance in simulated histograms of T_i^n/n for different source nodes i as n grows.

To compute the characteristic delay τ, we examine the left tails of $F_{i \to j}^n$ for large n. Our rationale for this approach is that, as illustrated in Fig. 1, the offset is the same across the whole distribution; and we will show that the left tails are amenable to a linear analysis. For $t \ll n\tau$, we linearize Eq, (2) to obtain

$$F_{i \to j}^n(t) \approx \int_0^t f(x) \sum_{k \in \partial j \setminus i} F_{j \to k}^{n-1}(t - x) dx. \tag{3}$$

This problem is mathematically analogous to that of front propagation, and we therefore follow the standard method described in [23]. The trivial solution $F^0(t) \equiv 0$ is linearly unstable with increasing n, and the dominant rate of growth will determine τ. The two-sided Laplace transform of Eq. (3) reads

$$\tilde{F}_{i \to j}^n(k) = \tilde{f}(k) \sum_{\ell \in \partial j \setminus i} \tilde{F}_{j \to \ell}^{n-1}(k), \tag{4}$$

where $\tilde{f}(k) = \int e^{-kx} f(x) dx$ is the Laplace transform of f. Viewing $\tilde{F}$ as a vector with entries indexed by directed edges, Eq. (4) describes an iterative process of multiplying by a matrix that encodes the entries of the sum, and then by the scalar $\tilde{f}(k)$. Thus, for large n, we can expect

$$\tilde{F}^n_{i \to j}(k) \propto v_{i \to j} e^{-\omega(k)n}, \qquad (5)$$

where the coefficient $v_{i \to j}$ contains the edge-specific information, and the function $\omega(k)$ determines the overall exponential growth rate. Substituting this ansatz into Eq. (4), we find $\mathbf{v} = \tilde{f}(k) e^{\omega(k)} B \mathbf{v}$, where B is the non-backtracking matrix [15]. This is an eigenvalue equation for B with a non-negative eigenvector $\mathbf{v}$; according to the Perron-Frobenius theorem, for a connected network, there is a unique maximum eigenvalue λ, which is real and positive. Thus, the growth rate is found as $\omega(k) = -\log[\lambda \tilde{f}(k)]$. Note that $1/\lambda = \rho_c$ is the percolation threshold of the network [13,16,17].

Examining the inverse transform at time $t + n\tau$, one finds (full details are in the Supplemental Material [18]) physically meaningful results in the limit of large n only when

$$\tau = \max_k \left\{ \frac{1}{k} [\log \rho_c - \log \tilde{f}(k)] \right\}. \qquad (6)$$

This is our first main result, showing how the speed of spread is determined by the network via its percolation threshold ρ_c and by the infection itself via the Laplace transform of its transmission time distribution. It is important to note that this result is derived from making a treelike assumption for the underlying network, and our calculation holds in the limit of the large distance from the source. In this sense, it describes the fastest spreading regime: the mid-outbreak phase of exponential growth.

In practical applications, however, most networks of interest are not treelike; and finite size effects mean the infection is unlikely to be able fully accelerate to the stable regime we have calculated. Nonetheless, our result still provides high-quality predictions. Figure 2 demonstrates the effectiveness of this measure on a variety of real-world networks from the Stanford Large Network Dataset Collection (SNAP) [24]: many with heavy-tailed degree distributions and high clustering; Table S.I in the Supplemental Material [18] gives full details. To further test the reliance of our method on the treelike assumption made in writing Eq. (2), we have simulated spreading processes in Watts-Strogatz random graphs with varying rewiring probabilities. Included in Fig. 2, the results for these networks show that our method performs better for higher rewiring probability, but is still very successful for highly clustered networks with low rewiring.

As well as the network, our measure of speed also depends on properties of the infection. One might expect the time delay τ to be scaled by the mean delay time; but, beyond this, it is difficult to discern from Eq. (6) how the shape of the distribution should affect the global speed of spread. To explore this aspect, we show in Fig. 3 the observed and predicted spreading speeds for Weibull distributed delays, interpolating between heavy tailed and Dirac distributed. Crucially, we find that the shape
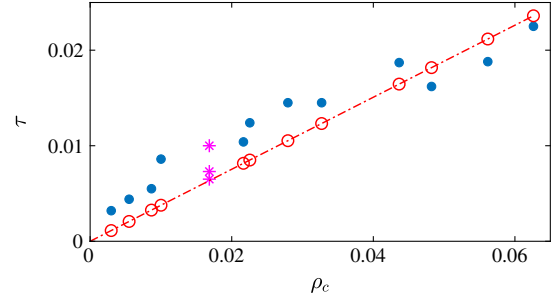

FIG. 2. Predicted and observed values of spreading delay τ for a unit rate exponential infection spreading on a variety of social and communication networks [25–32] with different percolation thresholds ρ_c. Predicted values (red circles) are calculated using Eq. (6). Observed values (blue dots) show the average over 10^3 simulations with random source nodes. Stars show results for Watts-Strogatz random graphs on 10^4 nodes with degree 30 and rewiring probabilities of 0.1, 0.5, and 1. Full details of all simulations are given in the Supplemental Material [18].

of the distribution of the transmission time has a substantial effect on the speed of spread in a network. If there is mass near zero, then delays are minimal due to the presence of extremely fast transmission routes. Conversely, if the transmission time is close to deterministic, then spreading is determined entirely by graph distance, meaning $\tau \approx 1$. In the Supplemental Material [18], we prove that τ is always less than the mean delay time, with equality only for Dirac-delta distributions.

The time taken to receive the infection.—As well as predicting the overall spreading speed, our approach also allows us to rank nodes in the network by their expected time to become infected. Write Δ_{ij} for the offset in infection time between nodes i and j, which for large n should satisfy $F^n_i(n\tau) = F^n_j(n\tau + \Delta_{ij})$. Inverting the transform in Eq. (5) for large n by steepest descent and comparing with the above (details in the Supplemental Material [18]), we find that
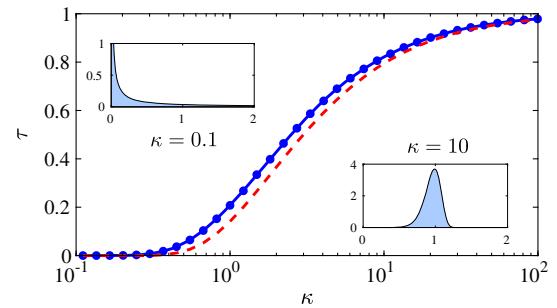

FIG. 3. Simulated (blue marked line) and predicted (red dashed line) spreading delays τ for infections with Weibull delay times with varying shape parameters κ and a fixed mean 1 (example delay distributions shown in insets). Simulations follow the same method as Fig. 2, averaged over 100 samples on an Erdős-Rényi graph of 10^4 nodes and a mean degree 3.
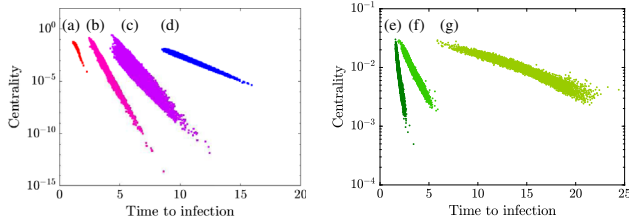
FIG. 4. Non-backtracking centrality predicts time to infection. Left: scatter plot of centrality and average arrival time for nodes in a selection of networks using a contagion with Weibull ($\kappa = 10$) infection times. Right: results for simulations of complex contagions with threshold θ in an Erdős-Rényi graph with 10^4 nodes and a mean degree of six. Key: (a) an interpersonal contact network in an American high school [32], (b) "Epinions" social media, (c) "Deezer" Romanian social network [25], (d) an Erdős-Rényi graph with $N = 10^5$ nodes, (e) $\theta = 1$, (f) $\theta = 2$, and (g) $\theta = 3$.

$$\Delta_{ij} = \frac{1}{k^\star} \log\left(\frac{c_i}{c_j}\right) + \mathcal{O}(1/n), \qquad (7)$$

where $k^\star = \operatorname{argmax}_k\{\omega(k)/k\}$, and

$$c_i = \sum_{j \in \partial i} v_{i \to j}$$

is the non-backtracking centrality of node i. This log-linear relationship is demonstrated numerically in Fig. 4 for nodes in a selection of networks from the SNAP. This result is important because it resolves the open question of exactly how network centrality measures may be used to estimate epidemic arrival time, and it provides a robust theoretical justification for the use of non-backtracking centrality (see Supplemental Material [18] for a comparison to other centrality measures).

Going further, many realistic models of network contagion require the number of infected neighbors of a node to reach some threshold $\theta \geq 1$ before the infection is passed on. In the Supplemental Material [18], we show how a variation of our theory, building on results from [33], extends to these complex contagion models by considering the θ-shortest temporal paths from a node. Remarkably, the log-linear relationship derived above continues to hold in this more complex setting, as illustrated in Fig. 4. In addition to this visual demonstration, we present in Table I the Pearson correlation coefficients between log-non-backtracking centrality and the infection arrival time for various disease dynamics in various networks. These results show that our theory, which is physically justified and cheap to compute, provides excellent predictions of the relative delay between nodes in a wide variety of spreading processes.

Because the non-backtracking centrality of a node is mainly a property of its local environment, the result of Eq. (7) means that we should expect the vast majority of infections to occur during a time window for which the duration is independent of the total size of the network. However, it can be shown that, in a network of size N, the time needed for an infection to take hold grows like $\log(N)/(\lambda - 1)$. Taken together, these results imply that, on the timescale of the spreading contagion in a large network, one will observe an almost instantaneous jump between a vanishing fraction of nodes infected to almost complete infection. We illustrate this result in Fig. 5 for Erdős-Rényi graphs of increasing size, and we provide precise theoretical derivations in the Supplemental Material [18], where we show that this property holds for models of temporal percolation in both sparse and dense networks.

Discussion.—We have presented here a theoretical framework for determining the speed of contagion processes in large networks. Analyzing the spreading front of the contagion probability, we derived Eq. (6), showing how network topology and infection dynamics affect speed via, respectively, the network percolation threshold and the Laplace transform of the transmission time law. Our theory also reveals in Eq. (7) a surprisingly simple relationship between contagion arrival times and the non-backtracking centrality of nodes. Finally, we have observed that these results imply that the spreading process in large networks

TABLE I. Correlation coefficient between contagion arrival time (measured from 10^3 simulated spreading processes with random sources) and the logarithm of non-backtracking centrality, for various networks. Values close to theoretical limit -1 correlation imply strong prediction quality.

| Network name	Exponential			Weibull ($\kappa = 10$)	
	Simple	2-core	3-core	Simple	2-core
Erdős-Rényi	-0.9487	-0.9693	-0.9502	-0.9721	-0.9691
Epinions [31]	-0.8765	-0.7769	-0.7428	-0.9946	-0.8661
Deezer Croatia [25]	-0.8265	-0.8094	-0.8088	-0.9049	-0.8697
Facebook Artists [25]	-0.7943	-0.7481	-0.7322	-0.9586	-0.8778
ArXiv Condensed Matter [28]	-0.8712	-0.7999	-0.7939	-0.9513	-0.8296
Facebook Companies [25]	-0.8439	-0.7203	-0.6685	-0.9138	-0.7606
School contact [32]	-0.7667	-0.7532	-0.7199	-0.9661	-0.9371
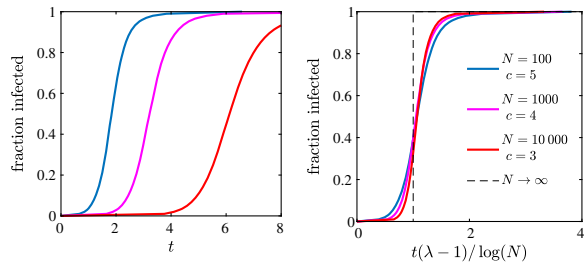
FIG. 5. Fractional size of the cluster of infected node as a function of time in various Erdős-Rényi graphs of different sizes N and mean degree c, averaged over 100 simulations from random seed nodes, with standard exponential infection times. The left panel shows real time, and the right has rescaled time showing convergence to a step function in limit $N \to \infty$, implying "instantaneous" spread to the bulk of the network.

undergoes an almost instantaneous expansion in their reach when time is properly scaled.

The setting for our theoretical derivation has been that of simple epidemics spreading on large treelike networks. However, we have shown that the key results hold remarkably well for a broad class of networks, including those with high clustering, and for contagion models including non-Markov dynamics and complex threshold models. Further development of rigorous mathematical results for these models is a challenging problem worthy of considerable future efforts. Excitingly, our results suggest possible routes for the development of monitoring and intervention protocols for real-world contagions using message-passing methods. Progress in this direction may require the consideration of even more detailed models including temporally varying and multilayered networks: both promising avenues for future research.

[*]s.moore@bath.ac.uk
[†]t.c.rogers@bath.ac.uk

[1] C. McEvedy, Sci. Am. **258**, 118 (1988).
[2] K. Zinszer, K. Morrison, J. S. Brownstein, F. Marinho, S. F. Alexandre, and E. O. Nsoesie, Emerg Infect Dis. **23**, 91 (2017).
[3] S. Vosoughi, D. Roy, and S. Aral, Science **359**, 1146 (2018).
[4] C. Moore and M. E. J. Newman, Phys. Rev. E **61**, 5678 (2000).
[5] M. J. Keeling and K. T. Eames, J. R. Soc. Interface **2**, 295 (2005).
[6] L. Danon, A. P. Ford, T. House, C. P. Jewell, M. J. Keeling, G. O. Roberts, J. V. Ross, and M. C. Vernon, Interdiscip. Perspect. Infect. Dis. **2011**, 1 (2011).
[7] R. Pastor-Satorras, C. Castellano, P. van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).
[8] W. Van den Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, BMC Infect. Dis. **11**, 37 (2011).
[9] A. Gautreau, A. Barrat, and M. Barthelemy, J. Theor. Biol. **251**, 509 (2008).
[10] A. Gautreau, A. Barrat, and M. Barthélemy, J. Stat. Mech. (2007) L09001.
[11] S. P. Borgatti, Soc. Networks **27**, 55 (2005).
[12] B. Karrer and M. E. J. Newman, Phys. Rev. E **82**, 016101 (2010).
[13] K. E. Hamilton and L. P. Pryadko, Phys. Rev. Lett. **113**, 208701 (2014).
[14] B. Karrer, M. E. J. Newman, and L. Zdeborová, Phys. Rev. Lett. **113**, 208702 (2014).
[15] T. Martin, X. Zhang, and M. E. J. Newman, Phys. Rev. E **90**, 052808 (2014).
[16] T. Rogers, Europhys. Lett. **109**, 28005 (2015).
[17] R. Kühn and T. Rogers, Europhys. Lett. **118**, 68003 (2017).
[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.124.068301 for additional details of calculations and simulations, and extensions of our results to other models.
[19] K. Gough, Biometrika **64**, 559 (1977).
[20] D. Anderson and R. Watson, Biometrika **67**, 191 (1980).
[21] A. L. Lloyd, Theor. Popul. Biol. **60**, 59 (2001).
[22] M. Eichner and K. Dietz, American Journal of Epidemiology **158**, 110 (2003).
[23] W. Van Saarloos, Phys. Rep. **386**, 29 (2003).
[24] J. Leskovec and A. Krevl, SNAP datasets: Stanford large network dataset collection, http://snap.stanford.edu/data (2014).
[25] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, arXiv:1802.03997.
[26] J. Leskovec, L. A. Adamic, and B. A. Huberman, ACM Trans. Web (TWEB) **1**, 5 (2007).
[27] J. Yang and J. Leskovec, Knowledge Inf. Syst. **42**, 181 (2015).
[28] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM Trans. Knowledge Discovery Data (TKDD) **1**, 2 (2007).
[29] E. Cho, S. A. Myers, and J. Leskovec, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, 2011), pp. 1082–1090.
[30] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Math. **6**, 29 (2009).
[31] M. Richardson, R. Agrawal, and P. Domingos, in *International Semantic Web Conference* (Springer, New York, 2003) pp. 351–368.
[32] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, Proc. Natl. Acad. Sci. U.S.A. **107**, 22020 (2010).
[33] M. Shrestha and C. Moore, Phys. Rev. E **89**, 022805 (2014).