The effectiveness of contact tracing in heterogeneous networks

Sadamori Kojaku,[1] Laurent Hébert-Dufresne,[2,3] and Yong-Yeol Ahn[1,4,*]

[1] *Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN 47408, USA*
[2] *Vermont Complex Systems Center, University of Vermont, Burlington, VT 05405, USA*
[3] *Department of Computer Science, University of Vermont, Burlington, VT 05405, USA*
[4] *Network Science Institute, Indiana University, Bloomington, IN 47408, USA*

(Dated: May 4, 2020)

Case isolation and contact tracing is a widely-used intervention method for controlling epidemic outbreaks. Here, we argue that the effectiveness of contact tracing and isolation is likely underestimated by existing studies because they do not take into account the different forms of heterogeneity and sampling biases from the network structure. Specifically, we show that contact tracing can be even more effective than acquaintance sampling at locating hubs. Our results call for the need for contact tracing to go both backward and forward, in multiple steps, to leverage all forms of positive biases. Using simulations on networks with a power-law degree distribution, we show that this deep contact tracing can potentially prevent almost all further transmissions even at a small probability of detecting infected individuals. We also show that, when the number of traced contacts is small, the number of prevented transmission per traced node is even higher—although most traced individuals are healthy—than that from case isolation without contact tracing. Our results also have important consequences for new implementations of digital contact tracing and we argue backward and deep tracing can be incorporated without the important sacrificing privacy-preserving requirements of these new platforms.

INTRODUCTION

Contact tracing, combined with case isolation, is one of the most intuitive and oldest methods for controlling epidemic outbreaks [1–3]. Contact tracing is considered a highly effective strategy *if* the contact network is concretely defined (e.g. sexually-transmitted diseases) or when the outbreak is still at the early stage. For instance, it was possible to control the severe acute respiratory syndrome (SARS) outbreak in 2003 with case isolation and contact tracing [4]. However, contact tracing, which involves interviewing and tracing contacts, is a labor-intensive and costly process. When there are many cases or the route of spreading is not clearly defined (e.g. airborne diseases), the traditional contact tracing can thus become cost inhibitive. Therefore, a critical question in contact tracing has been assessing its viability [5] or the cost-benefit tradeoffs [6].

Meanwhile, recent studies on *digital contact tracing* have demonstrated that leveraging smart mobile devices may allow much more swift and efficient contact tracing [7], asking an important question of whether digital contact tracing can overcome the existing limitations of contact tracing. Yet, existing studies do not fully leverage insights from network science about the underlying contact structure, particularly regrading the imapct of its heterogeneity.

Heterogeneous networks, where the number of contacts varies significantly among individuals, have been of great interest in network epidemiology because such heterogeneity can alter the fundamental nature of the dynamics in the form of, for instance, vanishing epidemic threshold [8], hierarchical spreading [9], and large variance in individual's reproductive number [10] as well as the final outbreak size [11]. Under the assumption of heterogeneous contact network structure, epidemic dynamics is dominated by the hubs and the superpreading events caused by them.

Superspreading events can be easily found during an epidemic outbreak. For instance, a famous example from the COVID-19 pandemic would be the "Patient 31" in South Korea [12]. The patient was the first positive case from a church that was later identified, via contact tracing, as the single biggest super-spreading event in South Korea, linked to more than $5,000$ cases, which account for *more than half* of South Korea's total cases at the time [12]. Super-spreading events are common in the past epidemic outbreaks [10], and they are often discovered through contact tracing efforts [13, 14]. Contact tracing's ability to identify super-spreading events raise important questions: how effective is the contact tracing at identifying superspreading events? How does the heterogeneity in the contact network affect the effectiveness of contact tracing?

Here, we show that the effectiveness of contact tracing strongly depends on the degree heterogeneity of the networks and can become extremely efficient in heterogeneous networks. Importantly for new applications of digital contact tracing, we show that this heterogeneity is not fully leveraged by the most common contact tracing implementations that are limited to future contacts or to a single step in the contact network.

RESULTS

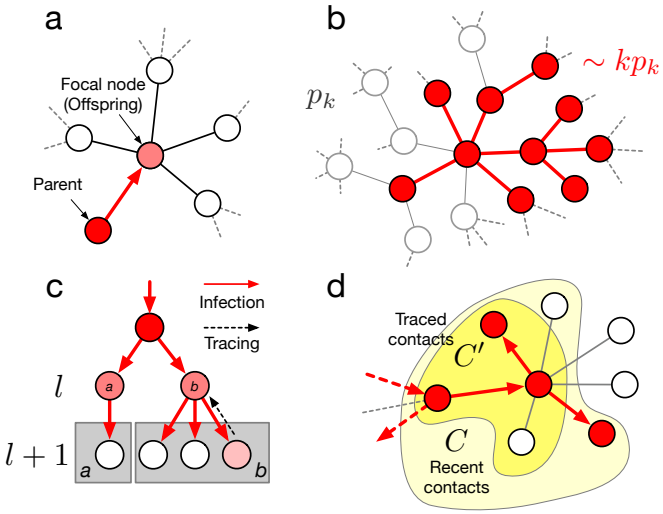

FIG. 1. (a) A transmission occurs from a 'parent' to a 'focal node' (or an offspring). (b) Because a node is infected through an edge, at the early stage of the epidemic, the degree distribution of the infected nodes is roughly proportional to kp_k. Because the degree of a node in the transmission tree is also proportional to the original degree, sampling from the transmission tree via an edge will be even more biased; it samples nodes proportional to k^2p_k. (c) If contact tracing perfectly identifies the parents, we sample nodes based on k^2p_k. (d) The true list of recent contacts (C) always includes the parent, along with other neighbors. The actual list of contacts that get traced (C') is a subset of C.

Let us first sketch the central idea. Given that the dynamics of epidemic spreading is dominated by hubs in static heterogeneous networks [8, 9, 15], let us first focus on the degree distribution of the traced nodes. Consider a random network with degree distribution p_k. The probability generating function for this network's degree distribution can be written as

$$G_0(x) = \sum_k p_k x^k. \tag{1}$$

First, note that the spreading process, at the early stage, can be considered as sampling nodes by following edges (see Fig. 1a). Because nodes are sampled proportionally to their degree, the degree distribution of the sampled (infected) nodes would follow, not p_k, but $q_k = \frac{kp_k}{\sum_k kp_k}$. Second, assume that the disease spread through an edge with a probability T, which we call *transmissibility* of the disease, and this assumption allows us to approximate the disease spreading by a bond percolation process [15–17]. If we focus on a transmission tree—a connected component from the bond percolation process with probability T (see Fig. 1b)—we can consider it as another network where the *original* degree of the nodes is distributed according to q_k because the nodes in this tree are sampled by following edges. Now, an idealized and simplified contact tracing can be thought as a process of (i) identifying an infected node from the transmission tree and (ii) following the edges of the node to further identify other infected nodes in the transmission tree. Because the degree of a node in the transmission tree is proportional to its original degree and because following an edge in the transmission tree is *once again* biased by the degree of a node, the degree distribution of the *contact-traced* infected nodes would be $\sim k^2p_k$ rather than $\sim kp_k$. In other words, contact tracing can be highly efficient—even more efficient than acquaintance sampling [18, 19]—at identifying high-degree nodes, or super-spreading events. Simply put, this is because we can leverage the work required in identifying an infected node in the first place, and then do biased sampling over the already biased subset of infected nodes.

We can also explicitly take the directionality into account by considering a directional transmission tree (see Fig. 1c). For the sake of simplicity, let us consider a case shown in Fig. 1c and assume that tracing only occurs from offsprings to parents because the nodes at $l-1$ are more likely to be asymptomatic or already recovered when the disease spread further down. Because the nodes in the tree are already sampled by following an edge, the node with original degree k is k times more likely to be in the transmission tree in the first place. Then the generating function for the excess degree distribution can be written as

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{1}{\sum_k kp_k} \sum_k kp_k x^{k-1}. \tag{2}$$

An extreme case of perfect tracing would be the case where we can immediately identify the true parent node once we identify an infected node. If we compare nodes at level $l+1$, as illustrated in Fig. 1c, node b is three times more likely to be traced than a; that is, for a node with original degree k, the probability of being traced from an offspring is proportional to $k-1$. Combining these two factors, the probability that a node with degree k gets contact-traced is proportional to $k(k-1)$ or roughly to k^2. The remaining degree that can be potentially blocked by the contact tracing (e.g. two remaining white nodes in the case of b in Fig. 1c) can be captured by the following probability generating function:

$$G_2(x) = \frac{G_1'(x)}{G_1'(1)} = \frac{1}{\sum_k k(k-1)p_k} \sum_k k(k-1)p_k x^{k-2}. \tag{3}$$

This intuition can be easily checked by running a SIR model on a network and sampling the parents of a randomly chosen infected node (see Fig. 2a). As expected, if we sample infected nodes and trace their parents, their degree distribution closely matches the distribution described by $G_2(x)$.

Needless to say, it is impossible to perfectly identify the disease pathway in practice; we can only obtain an incomplete list of recent contacts of a node. Note that, however, the *parent* node who is responsible for the infection of the focal node is, by definition, always belongs to the *true* list of recent contacts (denoted as C in Fig. 1d). Therefore, when we obtain C', the set of nodes to be traced (e.g. via an interview or through digital contact tracing), the parent node is more likely to be included, and thus more likely to be traced. Moreover, if the parent node is a super-spreader, then it would appear in many recent contact lists (from different nodes) and becomes much more likely to be contact-traced. Imagine an infected node i that has spread the disease to n of its neighbors. If there is a constant probability (p) for i to be traced (included in C') then the probability of i being contact-traced is $1-(1-p)^n$. In other words, super-spreaders are more likely to show up, potentially multiple times, in the list of recent contacts to be traced even in practical scenarios. Based on this idea, we hypothesize that contact tracing is highly effective in heterogeneous networks, especially if it is allowed to move at least two steps deep—immediately identifying and isolating the offsprings of the identified parents—and both backward and forward, in the contact network.

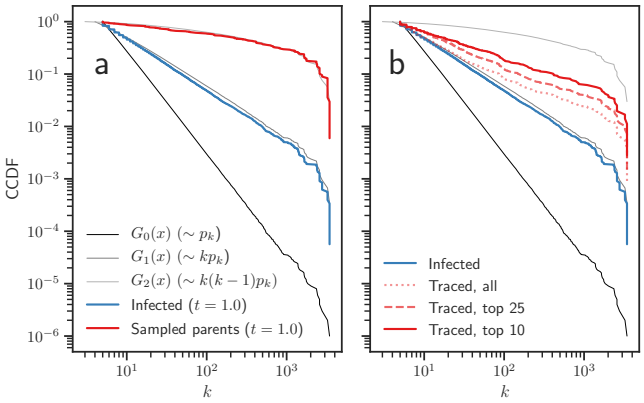

FIG. 2. (a) The degree distribution of the infected nodes and that of the parents of randomly sampled (1%) infected nodes at $t = 1.0$. They closely follow the distribution G_1 and G_2 respectively. (b) The effectiveness of more realistic contact tracing in terms of hub-sampling. Even under more realistic scenarios where we cannot accurately identify the true parents, contact tracing reaches hubs effectively. Such tendency increases as we sort the list of nodes to be traced based on how many times they appear in the list and pick the most frequent ones. Parameters used in this simulation are: $p_s = 0.01$, $p_t = 0.5$, and $p_r = 0.5$. We run the simulation 50 times and sample the nodes of each type for each simulation. Then, we compute the degree distributions for the nodes.

We test this hypothesis with simulations. First, we examine the degree distribution of the contact-traced nodes. We create a network with Barabási-Albert model [20], with $N = 10^6$ and $m = 5$. We then run the SIR model with transmission rate $\tau = 0.1$, recovery rate $\gamma = 1$, and initial seed fraction $\rho = 0.0001$, by using the `EoN` package [21]. At the early stage ($t = 1.0$), we examine the degree distribution of all infected nodes (blue) and that of the parents of sampled infected nodes (red). As shown in Fig. 2a, the empirical distributions closely follow the expected probability distribution.

We then simulate the contact tracing. Let us assume that there exists a true list of recent contacts, which always contains the parent. The actual contacts to be traced is obtained by sampling from this true recent contact list. This process can be described as following: (i) at time t, a fraction (p_s) of infected nodes will be identified and isolated (no further infection); (ii) for each identified node, we add the parent to the true recent-contact list and add others with probability p_t; (iii) we then sample nodes from this list with probability p_r to create the actual list of nodes to be traced. As we merge the lists from many identified nodes, we keep the counts of each node's occurrences; if there was a super-spreading event, the super-spreader is likely to appear multiple times in the list. This procedure cannot be as effective as the idealized perfect contact tracing. Yet, this scenario still samples hubs much better than acquaintance sampling (see Fig. 2b). As we focus more on those who ranked high in the contact list, the effectiveness increases.

Let us now measure the effectiveness of contact tracing more directly. We focus on a single intervention at time t. At t, we perform a single course of the contact tracing procedure as described above (with varying p_s) and measure how many further infections can be prevented by contact tracing and isolation. We run the SIR model simulation on the network and examine the transmission events between t and $t+\Delta$. Formally, we measure

$$\phi(t, t+\Delta) = \frac{\hat{T}(t, t+\Delta)}{T(t, t+\Delta)}, \tag{4}$$

where $T(t, t+\Delta)$ is the total number of transmission events that would have happened between t and $t+\Delta$ if there were no intervention at t, and $\hat{T}(t, t+\Delta)$ is the number of transmission events that would not happen because they were from the contact-traced nodes. As can be seen in Fig. 3, even a tiny amount of contact tracing is highly effective at preventing further spreading. Even if we only detect 1% of currently infected nodes and trace their contacts, it is enough to prevent about half of the total near-future transmissions. The ϕ saturates close to 1 when only 5% of contact tracing (see Fig. 3 inset). In other words, almost all further transmissions can be prevented by 5–10% contact tracing. This is primarily because contact tracing can efficiently identify super-spreaders; if we randomly sample infected nodes, they are much more likely infected from a super-spreader and thus likely lead us to the super-spreader.

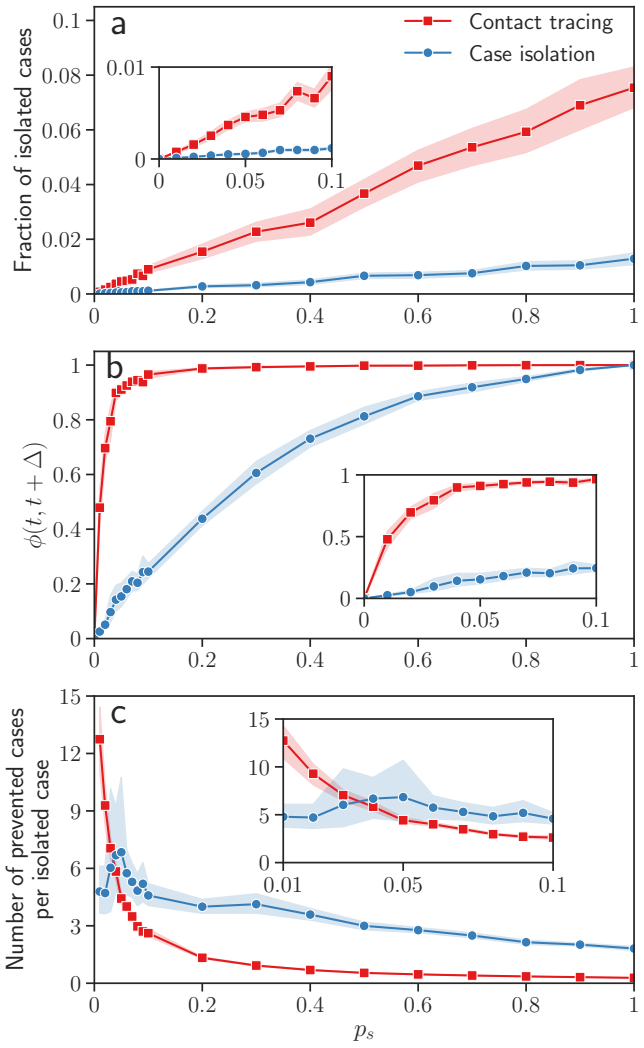

FIG. 3. (a) Fraction of isolated cases by case isolation (blue) and contact tracing (red). We generate 30 networks with Barabási-Albert model with the same parameter used in Fig. 2 and $p_s \in [0, 1]$. Each point indicates the value averaged over the SIR model simulations for the generated networks. The translucent band indicates the 95% confidence interval estimated by a bootstrapping with 10^4 resamples. (b) The fraction of transmissions that are prevented (ϕ) between t and $t + \Delta$ ($t = 3$ and $\Delta = 0.5$). (c) Number of prevented cases per isolated case. A small fraction of contact tracing $p_s \ll 1$ prevents a majority of transmission events more efficiently than case isolation.

We also examine the efficiency of contact tracing by measuring the number of prevented future cases per traced & isolated individuals (see Fig. 3c). The case isolation baseline refers to the intervention where we isolate only the infected individuals who have been discovered (with probability p_s from all infected individuals). Unlike the isolation-alone strategy where every isolation is almost guaranteed to prevent further spreading, the contact tracing inevitably has to examine many healthy individuals. Therefore, contact tracing is almost always considered as a costly strategy unless the outbreak is at its very early stage. However, our results demonstrate that contact tracing can be even more efficient than case isolation even if the epidemic is well on the way. Because contact tracing is highly efficient at locating super-spreading events and super-spreaders, it can prevent a huge number of further transmissions even with a small amount of resource.

Our results indicate that contact tracing may be highly effective and efficient strategy particularly if it were to be performed at all. Contrary to the usual findings that suggest a substantial amount of contact tracing is necessary to control the epidemic [5], our results suggest that—if strong heterogeneity in contact structure is present—*any* amount of contact tracing is effective. Our results also indicate that the 'cheaper' contact tracing offered by digital contact tracing may hold even greater potential than previous suggested.

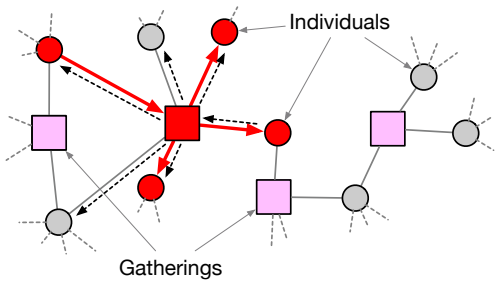

FIG. 4. The biased sampling is also in play in a bipartite network of people and gatherings. A high-degree gathering is more likely be "infected" and it is also more likely to be traced. Both sampling biases are roughly proportional to the gathering's degree in the bipartite network.

The same reasoning can be extended to the bipartite network of people and gatherings (see Fig. 4; e.g. churches, grocery markets, or any spontaneous gatherings) [22]. The recently-proposed privacy-preserving contact tracing protocols such as DP-3T [23] also works in terms of 'gatherings' by marking each temporal segment with a unique code (gathering) and propagating the identification of infected individuals through the past gatherings. Existing studies also support the formalism of people-gathering bipartite network. It has been shown that the human mobility follow regular routines, making it highly predictable [24, 25]. An important implication of these studies is that the biparite network between people and gatherings—such as public transportation system, groceries, religious services, or even ad-hoc corridor encounters—would be less dynamic than the network of human contacts and proximate encounters, making a static network model more plausible. Moreover, studies have shown that high temporal resolution proximity detection can clearly reveal the people-gathering structure [22].

In this bipartite network, the degree of an individual roughly captures how mobile the person is across diverse sets of places and gatherings, and the degree of a gathering captures how large the gathering is. Let us use $G_0(x)$ again to denote the generating function of the individuals' degree distribution, and $F_0(x)$ to denote the generating function of the gatherings' degree distribution, where

$$G_0(x) = \sum_k p_k x^k, \tag{5}$$

$$F_0(x) = \sum_k q_k x^k. \tag{6}$$

The transmission event happens from a person to others *via* a gathering. When we sample a gathering from an individual, the excess size of the gathering is generated by

$$F_1(x) = \frac{F_0'(x)}{F_0'(1)} = \frac{1}{\sum_k k q_k} \sum_k k q_k x^{k-1}, \tag{7}$$

because we are sampling an edge. Therefore, the probability distribution of the number of one's neighbors through gatherings is captured by $G_0(F_1(x))$.

Let us assume that a gathering can be 'infected' by an infected individual (see Fig. 4). If we were to sample from all infected gatherings, the probability distribution of the size of the gathering would be generated by $F_1(x)$. Yet again, because larger gatherings would produce more infections and thus more likely to be traced, the probability generating function for the remaining contacts for a gathering (except the original spreader and the identified individual from which the contact tracing is initiated) would be written as

$$F_2(x) = \frac{F_1'(x)}{F_1'(1)} = \frac{1}{\sum_k (k^2 - k) q_k} \sum_k k(k-1) q_k x^{k-2}, \tag{8}$$

in the case of perfect contact tracing. Applying the same logic above, we again expect that, even in more realistic scenarios, even a small amount of contact tracing would be able to identify super-spreading events and prevent numerous further disease transmission events. Our results imply that the proximity-based digital contact tracing (e.g. DP-3T) can be exceptionally effective at identifying super-spreading events and subsequently preventing further transmissions from those who attended those super-spreading events, much more so than currently assumed.

notify other recent contacts of this identified parent node. While this will result in a lot more notifications, it is a feature and not a problem per se: The goal is to identify potential super-spreading events that, by definition, imply many contacts. Moreover, given that multiple signals is particularly indicative of high-risk as shown here, such features can be potentially leveraged for better intervention strategies.

Current implementations of digital contact tracing, including the Apple and Google partnership [26] and the DP-3T proposal [23], are focused on a one-step process: Notifying previous contacts of an infected individual to take appropriate safety measures and get tested. They do not explicitly consider the fact that one of these previous contact is likely the parent node of the infected individual that might be infecting other. Considering the requirements for testing in many regions, it is unlikely for the parent node to be tested in time for the app to then notify its heavily-biased number of excess neighbours. Therefore, we urge the consideration of multi-step notification feature that can fully leverage the biases from the contact network structure.

Importantly, an implementation of our model does not necessarily necessitate any compromise in terms of privacy or decentralization of the contact tracing protocol itself [27]. One could also imagine a hybrid where previous notifications pushed to different devices are stored in a decentralized database and deep contact tracing is only undertaken once a given device has been notified more than a certain amount of time, thereby increasing its chance of being a parent node. The benefits of such network-based contact-tracing could be great, especially if accompanied by serious educational efforts for users to comprehend the science behind the intervention and the importance of their own role in our social network.

ACKNOWLEDGMENTS

The authors would like to thank M. Girvan, J. Lovato, and other organizers of Net-COVID program, which initiated the project. We also thank A. Allard, S. Lehmann, C. Moore, E. Moro, A. S. Pentland, and S. V. Scarpino for helpful discussions. L.H.-D. acknowledges support from the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award.

* yyahn@iu.edu

[1] K. T. Eames and M. J. Keeling, Proceedings of the Royal Society of London. Series B: Biological Sciences **270**, 2565 (2003).
[2] D. Klinkenberg, C. Fraser, and H. Heesterbeek, PLoS ONE **1** (2006).
[3] C. M. Peak, L. M. Childs, Y. H. Grad, and C. O. Buckee, Proceedings of the National Academy of Sciences **114**, 4023 (2017).
[4] J. W. Glasser, N. Hupert, M. M. McCauley, and R. Hatchett, Epidemics **3**, 32 (2011).
[5] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun, *et al.*, The Lancet Global Health (2020).

[6] B. Armbruster and M. L. Brandeau, Health Care Management Science **10**, 341 (2007).

[7] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser, Science (2020).

[8] R. Pastor-Satorras and A. Vespignani, Physical Review Letters **86**, 3200 (2001).

[9] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, Physical Review Letters **92**, 178701 (2004).

[10] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, Nature **438**, 355 (2005).

[11] L. Hébert-Dufresne, B. M. Althouse, S. V. Scarpino, and A. Allard, arXiv preprint arXiv:2002.04004 (2020).

[12] Y. Shin, B. Berkowitz, and M. J. Kim, The Washington Post (2020).

[13] M. Andre, K. Ijaz, J. D. Tillinghast, V. E. Krebs, L. A. Diem, B. Metchock, T. Crisp, and P. D. McElroy, American Journal of Public Health **97**, 470 (2007).

[14] S. Park, Y. Kim, S. Yi, S. Lee, B. Na, C. Kim, J. Kim, H. Kim, Y. Kim, Y. Park, *et al.*, Emerging Infectious Diseases **26** (2020).

[15] M. E. Newman, Physical Review Letters **95**, 108701 (2005).

[16] M. E. Newman, Physical Review E **66**, 016128 (2002).

[17] E. Kenah and J. M. Robins, Physical Review E **76**, 036113 (2007).

[18] R. Cohen, S. Havlin, and D. Ben-Avraham, Physical Review Letters **91**, 247901 (2003).

[19] N. A. Christakis and J. H. Fowler, PLoS ONE **5** (2010).

[20] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[21] I. Z. Kiss, J. C. Miller, P. L. Simon, *et al.*, *Mathematics of epidemics on networks*, Vol. 598 (Springer, 2017).

[22] V. Sekara, A. Stopczynski, and S. Lehmann, Proceedings of the National Academy of Sciences **113**, 9977 (2016).

[23] C. Troncoso, M. Payer, J.-P. Hubaux, M. Salathé, J. Larus, E. Bugnion, W. Lueks, T. Stadler, A. Pyrgelis, D. Antonioli, L. Barman, S. Chatel, K. Paterson, S. Capkun, D. Basin, J. Beutel, D. Jackson, B. Preneel, N. Smart, D. Singelee, A. Abidin, S. Gürses, M. Veale, C. Cremers, M. Backes, R. Binns, C. Cattuto, G. Persiano, D. Fiore, M. Barbosa, and D. Boneh, "Decentralized privacy-preserving proximity tracing," `https://github.com/DP-3T` (2020).

[24] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, Science **327**, 1018 (2010).

[25] J. P. Bagrow and Y.-R. Lin, PLoS ONE **7** (2012).

[26] Apple and Google, "Exposure notification," (2020).

[27] H. Cho, D. Ippolito, and Y. W. Yu, arXiv:2003.11511 (2020).