Local structure can identify and quantify influential global spreaders in large scale social networks

Yanqing Hu[a,1,2], Shenggong Ji[b,1], Yuliang Jin[c,1], Ling Feng[d,e,1], H. Eugene Stanley[f,1,2], and Shlomo Havlin[g,1]

[a]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China; [b]School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China; [c]Key Laboratory for Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China; [d]Computing Science, Institute of High Performance Computing, Agency for Science, Technology, and Research, Singapore 138632; [e]Department of Physics, National University of Singapore, Singapore 117551; [f]Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215; and [g]Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel

Contributed by H. Eugene Stanley, December 31, 2017 (sent for review August 31, 2017; reviewed by Marc Barthelemy and Zoltán Toroczkai)

Measuring and optimizing the influence of nodes in big-data online social networks are important for many practical applications, such as the viral marketing and the adoption of new products. As the viral spreading on a social network is a global process, it is commonly believed that measuring the influence of nodes inevitably requires the knowledge of the entire network. Using percolation theory, we show that the spreading process displays a nucleation behavior: Once a piece of information spreads from the seeds to more than a small characteristic number of nodes, it reaches a point of no return and will quickly reach the percolation cluster, regardless of the entire network structure; otherwise the spreading will be contained locally. Thus, we find that, without the knowledge of the entire network, any node's global influence can be accurately measured using this characteristic number, which is independent of the network size. This motivates an efficient algorithm with constant time complexity on the long-standing problem of best seed spreaders selection, with performance remarkably close to the true optimum.

social media | complex network | percolation | influence | viral marketing

Modern online social platforms are replacing traditional media (1) for the spreading of information and communication of opinions (2–6). A common feature of today's online social networks (OSNs) is their gigantic sizes—for example, as of the second quarter of 2016, there are about 1.5 billion monthly active users on Facebook. Notably, multiplicative explosions of some information may take place at a global scale in such gigantic OSNs, which is the foundation of viral marketing strategies (7). Because of this, quantification of viral spreading is traditionally believed to need global network information. Indeed, most measures, such as k-shell (2), degree discount (8), cost-effective lazy forward (9), betweenness (10), closeness (11), and Katz index (12), evaluate the influence of nodes based on the knowledge of global network structures. In general, these methods become impractical for giant OSNs, because either the full network structural data are unavailable or the computational time is nonscalable. On the other hand, based on massive social experiments, Christakis and Fowler (13, 14) proposed the so-called three degrees of influence (TDI) theory, which states that any individual's social influence ceases beyond three degrees (friends' friends' friends) and therefore suggests the existence of an unknown yet local effect. A recent study also shows that a local approximation works fairly well for a qualitative global measure of collective influence (4). The above situation reveals an apparent paradox, which inspires us to ask a fundamental question: Could local network structure accurately determine the size of global spreading?

Results

Here we recover a local characteristic infection size s^* of the spreading process. It determines the key influence size in the stochastic spreading process described by the susceptible–infected–recovered (SIR) family models (15–21), which well describe the information spreading process in social media (22–25). We find a ubiquitous and well-separated, bimodal behavior in the supercritical spreading regime—the spreading either extends globally, reaching a finite fraction of the total population irrespective of the initial condition, or diminishes quickly beyond the local characteristic infection size (Fig. 1 *A* and *C*). The global and local phases are unambiguously separated. Using the mapping between the SIR family model and bond percolation (18, 26), we provide a concrete physical understanding of these two well-separated phases. We show that the local phase can be used to accurately quantify the node's (or nodes') spreading power (Fig. 2*A*). In particular, the statistical properties of infected cluster size distribution allow us to use solely local network structural information for selecting the best seed spreaders in significantly short constant time complexity.

Methods

Our study is carried out for an SIR spreading mechanism on connected networks. The central quantity of interest in the spreading model is the final number of activated nodes or the spreading influence (17). A common definition of the spreading influence of node i is the expected number of active nodes that originated from i,

$$S(i) \equiv \sum_{s=1}^{N} s\, g(i, s), \qquad [1]$$

Significance

Identification and quantification of influential spreaders in social networks are challenging due to the gigantic network sizes and limited availability of the entire structure. Here we show that such difficulty can be overcome by reducing the problem scale to a local one, which is essentially independent of the entire network. This is because in viral spreading the characteristic spreading size does not depend on network structure outside the local environment of the seed spreaders. Our approach may open the door to solve various big data problems such as false information surveillance and control, viral marketing, epidemic control, and network protection.

Author contributions: Y.H., L.F., and S.H. designed research; Y.H., S.J., and Y.J. performed research; Y.H., S.J., H.E.S., and S.H. analyzed data; and Y.H., L.F., H.E.S., and S.H. wrote the paper.

Reviewers: M.B., Centre Commissariat à l'Energie Atomique; and Z.T., University of Notre Dame.

The authors declare no conflict of interest.

[1]Y.H., S.J., Y.J., L.F., H.E.S., and S.H. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: huyanq@mail.sysu.edu.cn or hes@bu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1710547115/-/DCSupplemental.

Published online July 3, 2018.

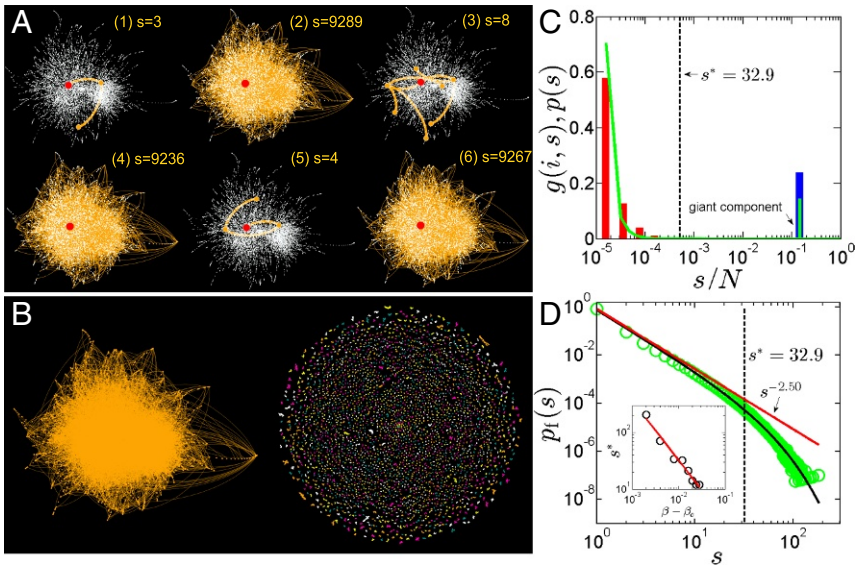

Fig. 1. Two phases phenomena. (*A*) Examples of simulated local (1, 3, 5) and viral (2, 4, 6) SIR spreadings in the NOLA Facebook network ($\beta = 0.02$, $\beta_c = 0.01$). We start the simulation from a randomly chosen node (red, $k = 27$). The active and nonactive nodes are colored in orange and white, respectively. (*B*) An illustration of giant (*Left*) and finite (*Right*) clusters in a bond percolation process. (*C*) The spreading probability distribution $g(i, s)$ (columns) is plotted together with the cluster size distribution function $p(s)$ (green line) obtained from percolation. Note that $p(s)$ is the average of $g(i, s)$ over all nodes. In this example, we use the same seed node i as in *A*, but other randomly chosen nodes give similar bimodal distributions, with the same viral peak at s^∞ (*SI Appendix*, section II). (*D*) The finite part $p_f(s)$ (green circles) of $p(s)$ is fitted to Eq. **5** (black solid line) to obtain the characteristic size $s^* = 32.9 \pm 0.6$ and the exponent $\tau = 2.50$. (*D*, *Inset*) The characteristic size s^* is fitted to a power-law divergence near the critical point β_c, with a non–mean-field exponent $\sigma = 1.05$. The same network and the same β are used in *A*–*D*.

where $g(i, s)$ is the probability that a total of s nodes are eventually activated by node i in a network of N nodes, and the probability of an active node to activate a neighboring node is β. In information spreading, an activated node corresponds to a spreader. We find that the probability distribution function $g(i, s)$ has two prominent features: (*i*) It consists of two peaks, which correspond to local and viral spreadings. The local peak is located at small s, while the viral peak is centered at significantly larger s (Fig. 1 *A* and *C*). Furthermore, the viral peak is a δ-like function, whose location is independent of node i and different stochastic realizations (*SI Appendix*, section II). (*ii*) The two peaks are separated by a wide gap, which implies that one may introduce a small filtering size,s^*, to distinguish between the two phases.

The statistical properties underlying these two features can be explored and better understood using the framework of percolation theory. This can be done through mapping the SIR process to bond percolation (18, 19), where every link (bond) has a probability $1 - \beta$ to be removed from the network (see *SI Appendix*, section XII for the more general case where β is link dependent). The final network forms many connected clusters of different sizes. It has been proved that the probability distribution function $g(i, s)$ in Eq. **1** is exactly equivalent to the cluster size distribution function $p(i, s)$, where s is the size of the cluster that node i belongs to (18, 27, 28). According to percolation theory, a giant component of size s^∞ emerges above the percolation transition threshold β_c (Fig. 1*B*), where $p(i, s)$ is split into a finite (nongiant) part $p_f(i, s)$ and a giant part $p(i, s^\infty)$ (Fig. 1*C*). The size of s^∞ is proportional to N and depends on β. Because $\sum s\, p_f(i, s) \ll s^\infty p(i, s^\infty)$, we may approximate Eq. **1** as

$$S(i) \approx \hat{S}(i) \equiv s^\infty\, p(i, s^\infty), \qquad [2]$$

where $s^\infty = \sum_{i=1}^{N} p(i, s^\infty)$. In other words, the spreading power of one node is the product of the giant component size and the probability that this node is in the giant component. In information spreading, a broader type of definition for "influence" exists by including the nodes that are linked to the spreaders but do not spread the information further, called "listeners" (29). Since the underlying two-phase behavior is essentially the same, the total number of listeners and spreaders increases monotonically with the total number of spreaders approximated in our percolation-based algorithm. This implies that maximizing the influence including listeners is equivalent to the problem of maximizing the number of spreaders (*SI Appendix*, section III).

In artificial random networks with structure purely determined by the degree distributions, we can give the analytical solution for this influence quantity $\hat{S}(i)$ in Eq. **1**, with

$$p(i, s^\infty) = 1 - (1 - q)^{k_i}, \qquad [3]$$

where k_i is the degree of node i, and $s^\infty = N \sum_{k=1}^{\infty} P(k)[1 - (1 - q)^k]$. Here, q is the probability of a random link to be connected to the giant component and is determined from the self-consistent equation $q = \beta \sum_{k=1}^{\infty} \frac{kP(k)}{\langle k \rangle}[1 - (1 - q)^{k-1}]$, with average degree $\langle k \rangle$ and arbitrary degree distribution $P(k)$ (30). The theoretical considerations and details for undirected, directed, and degree–degree correlated are presented in *SI Appendix*, sections IV and V.

For real networks whose structures are much more complex than random networks, an exact solution to the spreading influence is not possible. But the critical phenomenon and the statistical properties of the two phases remain the same (Fig. 1*C*). We can leverage on these properties, in particular the wide gap between these two phases to distinguish between viral and local spreadings, and construct methods to estimate the spreading influence of nodes in the network. In SIR processes, once the number of activated nodes reaches a threshold parameter m, the simulation could be terminated since this process is known to become most likely viral. We thus obtain a second approximated form for the node spreading power—the truncated spreading power,

$$S(i) \approx \tilde{S}(i) \equiv \tilde{s}^\infty\, \tilde{p}(i, \tilde{s}^\infty), \qquad [4]$$

where $\tilde{p}(i, \tilde{s}^\infty) \equiv \sum_{s=m}^{N} p(i, s)$, and $\tilde{s}^\infty \equiv \sum_{i=1}^{N} \tilde{p}(i, \tilde{s}^\infty)$. It turns out that percolation theory provides a fundamental insight into determining the threshold value m. According to the theory, the distribution $p_f(i, s)$ has a fast decay tail e^{-s/s^*}, where s^* gives a characteristic size of the finite components (27, 31). For any $m \geq s^*$, the error introduced in $\tilde{S}(i)$ by truncating this tail is small [see Fig. 2*A* for a comparison between the real $S(i)$ and $\tilde{S}(i)$ in real networks]. Fig. 2*B* shows that the relative error $E^r(i, m) \equiv [\tilde{S}(i) - S(i)]/S(i)$ decays quickly with m and becomes negligible for $m \geq s^*$ (see *SI Appendix*, sections IV and V for a theoretical calculation of the error in random networks). Similar to the giant component size, the characteristic component size s^* is intrinsic to the whole network and independent of the seed node i. Hence the characteristic size s^* has an important implication: Once it is determined either theoretically or numerically, it can be used as a threshold for the parameter m. As long as m is chosen to be above s^*, the truncated

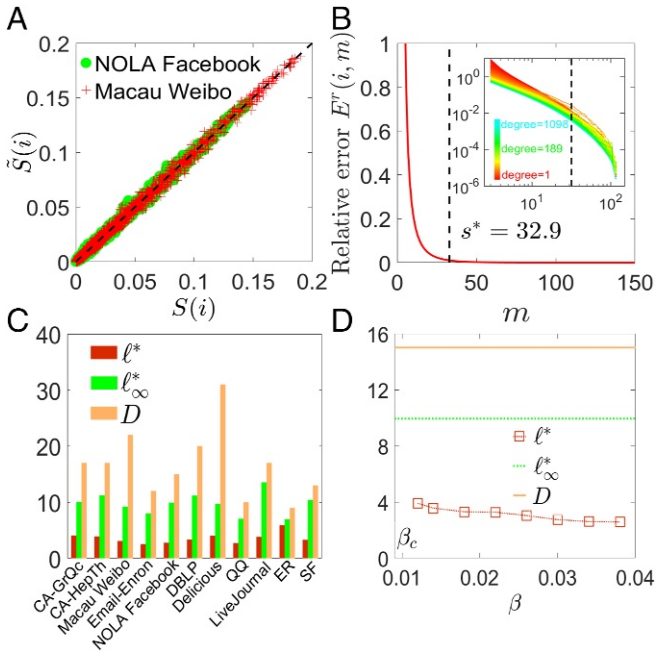

Fig. 2. Spreading power. (*A*) Comparison between the truncated spreading power $\tilde{S}(i)$ (Eq. **4**) and the real exact spreading power $S(i)$ (Eq. **1**) in NOLA Facebook and Macau Weibo ($\beta_c = 0.05$) networks, where each point represents one node. (*B*) The m dependence of the relative error $E^r(i,m)$ of nodes whose degrees are equal to the average degree $\langle k\rangle$, in the NOLA Facebook ($\beta = 0.02$) network. (*B*, *Inset*) The m dependence of the relative error $E^r(i,m)$ of nodes with different degrees. The relative error decreases quickly with m and becomes smaller than 1% when $m > s^*$. (*C*) Comparison among the influence radius ℓ^*, the average distance of the farthest nodes from the seed nodes ℓ^*_∞, and the network diameter D in nine OSNs and two random networks [an ER network with $N = 50{,}000$, $\langle k\rangle = 10$ and a scale-free (SF) network with $N = 50{,}000$, $P(k) \sim k^{-2.5}$]. We choose β in different networks such that the fraction of the giant component is the same; i.e., $s^\infty = 0.3N$ (see *SI Appendix*, section I for the real networks description). (*D*) The NOLA Facebook influence radius ℓ^* is smaller than both ℓ^*_∞ and D for any $\beta > \beta_c$.

spreading power $\tilde{S}(i)$ is an excellent approximation for $S(i)$, and its error is well controlled (*SI Appendix*, section VI).

The average of the cluster size distribution, $p(i,s)$, from seed node i gives the global cluster distribution function $p(s) = \frac{1}{N}\sum_{i=1}^{N} p(i,s)$. Excluding the giant component, its finite part $p_f(s)$ has the same tail as that of $p_f(i,s)$ (26, 27, 32) (Fig. 1*D*),

$$p_f(s) \sim s^{-\tau} e^{-s/s^*}, \qquad [5]$$

which can be used to obtain s^* theoretically in random networks (27). For example, in an Erdos–Renyi (ER) network, we obtain $s^*_{ER} = \frac{1}{\beta\langle k\rangle - 1 - \ln\beta - \ln\langle k\rangle}$ (*SI Appendix*, section IV). An expansion of this expression around the percolation transition β_c gives the critical scaling $s^* \sim |\beta - \beta_c|^{-1/\sigma}$, with the mean-field exponent $\sigma = 0.5$. For real OSNs, s^* is obtained by fitting the simulation data to the exponential tail in Eq. **5** (Fig. 1*D* and *SI Appendix*, section VII). Fig. 1*D*, *Inset* shows that s^* in real Facebook OSN also satisfies the critical power-law scaling.

To reveal the topological meaning of the characteristic size s^*, we define an influence hopping radius ℓ^* associated to s^*. We perform SIR simulations until s^* nodes are activated and assign the maximum hopping distance (shortest path) between the seed and active nodes, averaged over all realizations and nodes, to be the influence radius ℓ^*. For a typical β such that $s^\infty = 0.3N$, we find that $\ell^* \sim 3$–4 in all OSNs studied, which is significantly smaller than the average distance and diameter of the network as shown in Fig. 2*C*. This result shows that if an SIR spreading is local, then it would vanish within three to four steps; otherwise, it will spread to about $s^\infty = 0.3N$ nodes. Note that ℓ increases when $\beta \to \beta_c$ (Fig. 2*D*), whose scaling is discussed in *SI Appendix*, section IV. This behavior is analogous to a critical phenomenon of a continuous phase transition: At the critical point, the correlation length diverges, but as long as it moves beyond the critical point, a characteristic scale appears.

The above analysis explains the following paradox: While it is shown that the information spreading is in general a global process due to the viral spreading in the supercritical phase, the influence of any node basically depends only on its local network environment. While the computation time for $S(i)$ in Eq. **1** grows linearly with N, it is reduced to an N-independent constant $O(m)$ for the truncated spreading power $\tilde{S}(i)$ in Eq. **4**. An important extension of this finding is that the method can be combined with many search algorithms for detecting the best spreaders and reduce their time complexity by one order of N.

Next, we aim to find the best M spreaders $\mathcal{V} = \{v_1, v_2, \cdots, v_M\}$ from a given set $\mathcal{W}$ of L candidates, to maximize their collective spreading power $S(\mathcal{V}) = \sum_{s=1}^{N} s\, p(\mathcal{V}, s)$, where $p(\mathcal{V}, s)$ is the probability that a total of s nodes are activated by the selected spreaders in $\mathcal{V}$. Because it is usually more cost-effective to target a large set of less influential nodes, rather than a small set of globally most influential nodes (33), we choose nodes with average properties (around average degree) as candidates. In practice, it is usually extremely difficult to obtain the full network structural information. Therefore, unlike many other studies which select best seeds from the whole network, we only focus on a subset of candidate nodes. Extending from the formulation of a single node spreading power $\tilde{S}(i)$, we introduce a truncated collective spreading power $\tilde{S}(\mathcal{V}) \equiv \tilde{s}^\infty\, \tilde{p}(\mathcal{V}, \tilde{s}^\infty)$, where $\tilde{p}(\mathcal{V}, \tilde{s}^\infty)$ is the probability that at least one cluster of at least m nodes are activated by the M seed spreaders. While the computation time for the collective influence $S(\mathcal{V})$ increases linearly with N given any $\mathcal{V}$, it becomes N-independent for the estimator $\tilde{S}(\mathcal{V})$.

Now we demonstrate one example of how to improve other algorithms and introduce quantification capabilities through the combination of our approach with the natural greedy algorithm (NGA) (8, 17). We call this algorithm the percolation-based greedy algorithm (PBGA): We (*i*) first find the best spreader $\tilde{v}_1$ with the maximal individual spreading power based on the estimator $\tilde{S}(\tilde{v}_1)$, (*ii*) then fix $\tilde{v}_1$ and find the second best spreader $\tilde{v}_2$ that maximizes the collective spreading power $\tilde{S}(\tilde{\mathcal{V}})$ for $\tilde{\mathcal{V}} = \{\tilde{v}_1, \tilde{v}_2\}$, and (*iii*) repeat this process M times until M spreaders $\tilde{\mathcal{V}} = \{\tilde{v}_1, \tilde{v}_2, \ldots \tilde{v}_M\}$ are selected. As a greedy algorithm, the PBGA maximizes the marginal gain in the objective function $\tilde{S}(\tilde{\mathcal{V}})$ at each step. Note that replacing the objective function by the real spreading power in the above procedure would basically recover the NGA (see *SI Appendix*, section IX for more details).

As expected, the simulation results show that the computational time in terms of execution count of PBGA is independent of network size N (Fig. 3). This reduction becomes significant for a worldwide online social network with billions of nodes. In *SI Appendix*, Table S1, we compare and summarize the theoretical time complexities of our PBGA, the NGA, and other widely used algorithms, including brute-force search, genetic

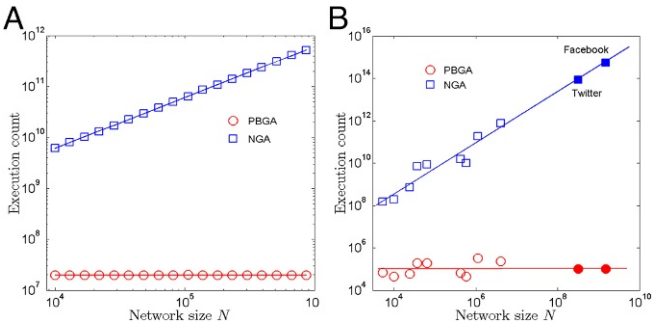

Fig. 3. Algorithm time complexity. (*A*) Comparison of the computational execution count of percolation-based greedy algorithm(PBGA) and natural greedy algorithm (NGA) (8) in ER networks with $\beta = 0.2$ ($\beta_c = 0.1$). The algorithms select the set of $M = 10$ most influential nodes from $L = 1{,}000$ candidates with degree at $\langle k\rangle = 10$. Unlike the NGA, PBGA's computational complexities are independent of network size. (*B*) Comparison of the computational execution count (rescaled by $\langle k\rangle$ and m) of the same algorithms in real OSNs (open symbols, from left to right: CA-GrQc, CA-HepTh, Macau Weibo, Email-Enron, NOLA Facebook, DBLP, Delicious, QQ, and LiveJournal; see *SI Appendix*, section I for the real networks description). The solid symbols are values of extrapolated execution count based on the size of whole Twitter and Facebook networks. The value of β is chosen such that the giant component size is 30% of the network size; i.e., $s^\infty = 0.3N$ in each OSN.

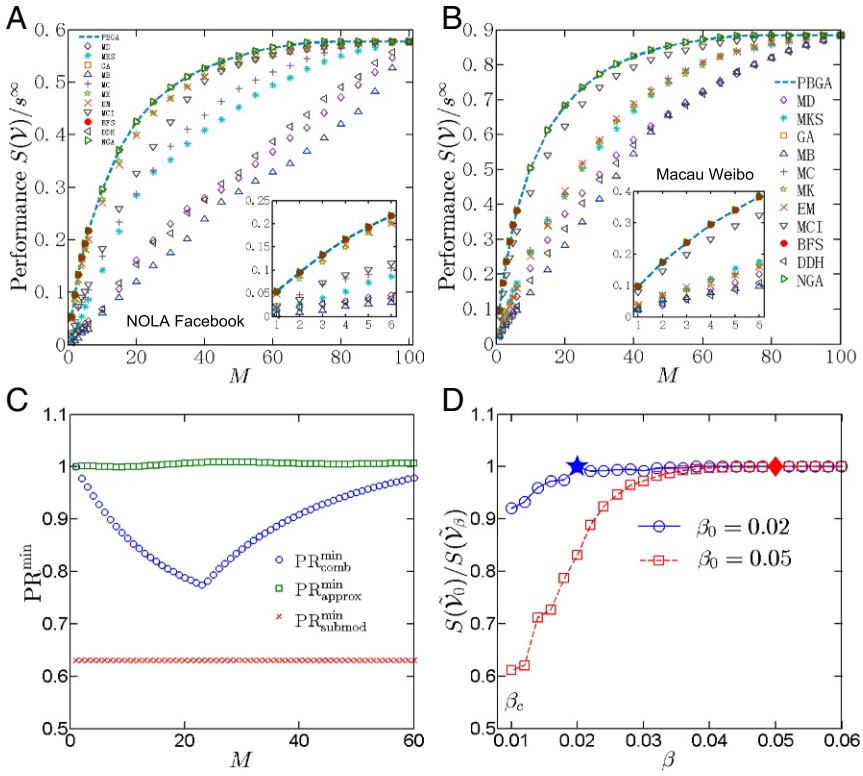

Fig. 4. Algorithm performance on real online social networks. For (*A*) NOLA Facebook ($\beta = 0.012$) and (*B*) Macau Weibo ($\beta = 0.055$) networks, we compare the algorithm performance of the PBGA with that of other algorithms: brute-force search (BFS), degree discount heuristic (DDH) (8), eigenvector method (EM), genetic algorithm (GA), maximum betweenness (MB) (10), maximum closeness (MC) (11), maximum degree (MD), maximum Katz (MK) index (12), maximum k-shell (MKS) (2), NGA, and MCI (4). The candidate nodes are randomly selected from the nodes with median degree nodes: Degree is 10 for nodes in Facebook and out-degree is 3 for nodes from Weibo. Since the candidate nodes have the same degree, the MD method is equivalent to the random selection of seed nodes. $S(\mathcal{V})$ is normalized by dividing the giant component size s^∞. Here $L = 100$ candidates and M varies from 1 to 100. (*A* and *B*, *Insets*) The regime $1 \le M \le 6$ is enlarged, where the rigorous optimum obtained from BFS is available. (*C*) On the Facebook network, the combined rigorous lower bound $\mathrm{PR}^{\min}_{\mathrm{comb}}$ and the approximated lower bound $\mathrm{PR}^{\min}_{\mathrm{approx}}$ are plotted together with the submodular lower bound $\mathrm{PR}^{\min}_{\mathrm{submod}} = 0.63$ (17), as functions of M. (*D*) The relative performance between the PBGA solution based on $\beta = 0.02$ ($\beta = 0.05$) and the PBGA solution based on the other β values, with both performances $\tilde{\mathcal{V}}_0$ and $\tilde{\mathcal{V}}_\beta$ estimated upon the same spreading rate β. The solid symbols (blue star and red diamond) label the performance of 1 when $\beta = \beta_0$. We see that as long as $\beta > \beta_0$, the solution at β_0 can be used at β since their performances are almost the same, as long as both β_0 and β are larger than the critical point $\beta_c \approx 0.01$.

algorithm, maximum degree, maximum k-shell (2), degree discount heuristic (8), maximum betweenness (10), maximum closeness (11), maximum Katz index (12), eigenvector method, and maximal collective influence (MCI) (4). Although maximum degree, degree discount, and MCI have N-independent theoretical computational complexities, the maximum degree and degree discount performances are much less than that of the PBGA and MCI is much slower than PBGA. This is because MCI needs the information of the nodes up to a distance ℓ of the seed nodes. In real networks which are small world, a small ℓ would lead to thousands or more nodes. On the other hand, PBGA's complexity depends on s^*, which is independent of the small-world effect. *SI Appendix*, Fig. S16 presents a graphical illustration of this difference.

We quantify the algorithm performance by comparing the collective spreading power $S(\mathcal{V})$ of the solution set $\mathcal{V}$ from different algorithms (Fig. 4 *A* and *B*). Our results show that for the entire range of studied M, the three algorithms, PBGA, NGA, and genetic algorithm (GA), have the best performances. Remarkably, the three algorithms give solutions indistinguishable from the true optimum obtained by the brute-force algorithm, when M is small (Fig. 4 *A* and *B*, *Insets*). In particular, comparing the performance of the PBGA and MCI in Fig. 5, we see that the PBGA significantly outperforms MCI when the number M of seed nodes is small. This can be understood since the original CI method (4) deals with best nodes for breaking down the network, which are not necessarily the best spreaders. This is likely the reason behind the relatively lower performance of MCI (more detailed discussion in *SI Appendix*, section IX, B9). When M becomes large, the performance difference diminishes, similar to the performance of any other algorithms as seen in Fig. 4*A*. In fact, we conjecture that for any M, the solution of the PBGA should be nearly optimal.

Another important aspect of this maximization problem is to have a sense of how good the solution is compared with the true optimal solution $\mathcal{V}^*$, which is usually unknown (9, 17). We give two lower bounds of the performance ratio $\mathrm{PR} \equiv S(\tilde{\mathcal{V}})/S(\mathcal{V}^*)$ between the PBGA performance $S(\tilde{\mathcal{V}})$ and the exact optimal performance $S(\mathcal{V}^*)$ (Fig. 4*C*): (*i*) a combined bound $\mathrm{PR}^{\min}_{\mathrm{comb}} \equiv \max\{\frac{p(\tilde{\mathcal{V}},s^\infty)}{\sum_{i\in\mathcal{U}^*} p(i,s^\infty)}, \frac{p(\tilde{\mathcal{V}},s^\infty)}{p(\mathcal{W},s^\infty)}\}$, where $\mathcal{U}^*$ is the set of M nodes with the maximum individual probability $p(i, s^\infty)$ (it is rigorous for any networks), and (*ii*) an approximated bound $\mathrm{PR}^{\min}_{\mathrm{approx}} \equiv \frac{p(\tilde{\mathcal{V}},s^\infty)}{1-\prod_{i\in\mathcal{U}^*}[1-p(i,s^\infty)]}$ that becomes rigorous in random networks (*SI Appendix*, section X). The rigorous boundaries, $\frac{p(\tilde{\mathcal{V}},s^\infty)}{\sum_{i\in\mathcal{U}^*} p(i,s^\infty)}$ and $\frac{p(\tilde{\mathcal{V}},s^\infty)}{p(\mathcal{W},s^\infty)}$, work well in the small and large M limits, respectively, where they both approach one, and the approximated bound $\mathrm{PR}^{\min}_{\mathrm{approx}} \approx 1$ for any M value considered. Considering the above analysis, we argue that the PBGA gives a nearly optimized solution for an arbitrarily given number M of spreaders.

In practical situations, the information spreading rate β is usually unknown. However, our PBGA method could find close to optimal solutions without knowing the exact β value, as long as the information spreading is viral, i.e., a supercritical region with $\beta > \beta_c$ (see *SI Appendix*, section XI for the nonviral subcritical regions $\beta < \beta_c$). As illustrated in Fig. 4*D*, the solutions found at an arbitrary spreading rate β_0 perform nearly optimally at higher spreading rate $\beta > \beta_0$. Thus, without knowing the exact spreading rate, one can use a spreading rate slightly above the critical value β_c, such that the solutions perform optimally at higher β values. On a related note, it has been observed that information spreading could exhibit bursty behaviors with different spreading speeds (29). Such mechanism could be mapped to dynamics having different beta values over the course of spreading,

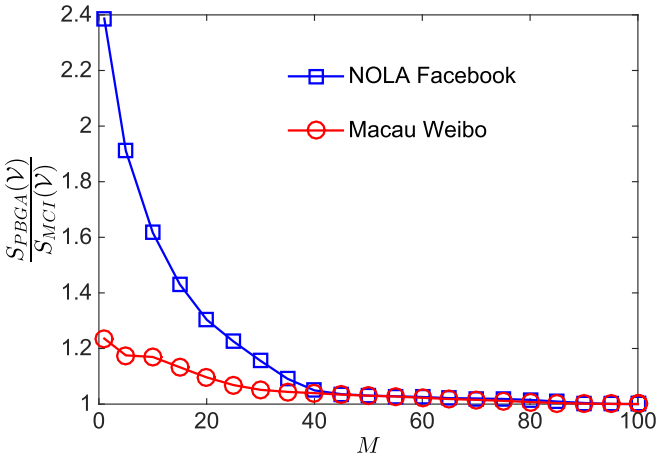

Fig. 5. Performance comparison between the PBGA and MCI. The vertical axis is the ratio between the seeds' influence of the PBGA and MCI. For a small number M of seed nodes, the PBGA significantly outperforms MCI. The difference diminishes as M increases, when both solutions approach the theoretical maximum of giant component size. The simulation is carried out on the Facebook network with $\beta = 0.012$.

possibly from small to large. They still belong to the SIR family. A more detailed analysis of this process is presented in *SI Appendix*, section XII.

Summary

In this work, we show from first principles that any node's influence can be quantified purely from its local network environment, based on the nature of the spreading dynamics. Our approach is distinct from other local attempts, which usually use some distance truncation strategies to approximate a relative global measure without the ability to quantify the actual influence. Although our framework is demonstrated on the basic SIR model, its applicability can be extended to several other spreading models if the following two properties hold: (*i*) For a collection of seed spreaders, the final steady states have two different outcomes of being either a localized outbreak with a small and finite number of infections or a global epidemic with the infectious/recovered population being proportional to the network size. (*ii*) When it is in the global outbreak, the size of the influence does not correlate with that of the initial spreader. See *SI Appendix*, section XII for discussions on a more general family of SIR models as well as for more complex models that include stiflers (34) and the susceptible–infected–susceptible model (22).

ACKNOWLEDGMENTS. Y.H., S.J., and L.F. are supported by The National Nature Science Foundation of China Grants 61773412, U1711265, 71731002; Guangzhou Science and Technology Project Grant 201804010473; and Three Big Constructions Supercomputing Application Cultivation Projects sponsored by National Supercomputer Center in Guangzhou. Y.J. is supported by Chinese Academy of Sciences Hundred-Talent Program. S.H. acknowledges the Israel Science Foundation, the Israel Ministry of Science and Technology (MOST) with the Italy Ministry of Foreign Affairs, MOST with the Japan Science and Technology Agency, the Bar-Ilan University Center for Research in Applied Cryptography and Cyber Security, and Defense Threat Reduction Agency Grants HDTRA-1-10-1-0014 for financial support. The Boston University Center for Polymer Studies is supported by National Science Foundation Grants PHY-1505000, CMMI-1125290, and CHE-1213217, and by Defense Threat Reduction Agency Grant HDTRA1-14-1-0017.

1. Rust RT, Oliver RW (1994) The death of advertising. *J Advert* 23:71–77.
2. Kitsak M, et al. (2010) Identification of influential spreaders in complex networks. *Nat Phys* 6:888–893.
3. Wang P, et al. (2009) Understanding the spreading patterns of mobile phone viruses. *Science* 324:1071–1076.
4. Morone F, Makse HA (2015) Influence maximization in complex networks through optimal percolation. *Nature* 524:65–68.
5. Aral S, Walker D (2012) Identifying influential and susceptible members of social networks. *Science* 337:337–341.
6. Bond RM, et al. (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489:295–298.
7. Ferguson R (2008) Word of mouth and viral marketing: Taking the temperature of the hottest trends in marketing. *J Consum Mark* 25:179–182.
8. Chen W, Wang Y, Yang S (2009) Efficient influence maximization in social networks. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09* (ACM, New York), pp 199–208.
9. Leskovec J, et al. (2007) Cost-effective outbreak detection in networks. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 420–429.
10. Newman ME (2001) Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Phys Rev E* 64:016132.
11. Freeman LC (1979) Centrality in social networks conceptual clarification. *Soc Network* 1:215–239.
12. Katz L (1953) A new status index derived from sociometric analysis. *Psychometrika* 18:39–43.
13. Christakis N, Fowler J (2007) The spread of obesity in a large social network over 32 years. *New Engl J Med* 357:370–379.
14. Christakis N, Fowler J (2013) Social contagion theory: Examining dynamic social networks and human behavior. *Stat Med* 32:556–577.
15. Barrat A, Barthelemy M, Vespignani A (2008) *Dynamical Processes on Complex Networks* (Cambridge Univ Press, New York).
16. Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci USA* 103:2015–2020.
17. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 137–146.
18. Newman MEJ (2002) Spread of epidemic disease on networks. *Phys Rev E* 66:016128.
19. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. *Rev Mod Phys* 81:591–646.
20. Meloni S, Arenas A, Moreno Y (2009) Traffic-driven epidemic spreading in finite-size scale-free networks. *Proc Natl Acad Sci USA* 106:16897–16902.
21. Eubank S, et al. (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429:180–184.
22. Pastor-Satorras R, Castellano C, Mieghem PV, Vespignani A (2015) Epidemic processes in complex networks. *Rev Mod Phys* 87:925–979.
23. Daley DJ, Kendall DG (1964) Epidemics and rumours. *Nature* 204:1118.
24. Iribarren JL, Moro E (2011) Branching dynamics of viral information spreading. *Phys Rev E* 84:046116.
25. Grabowski A, Kruszewska N, Kosiński RA (2008) Dynamic phenomena and human activity in an artificial society. *Phys Rev E* 78:066110.
26. Dorogovtsev SN, Goltsev AV, Mendes JFF (2008) Critical phenomena in complex networks. *Rev Mod Phys* 80:1275–1335.
27. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. *Phys Rev E* 64:026118.
28. Yuan X, Hu Y, Stanley HE, Havlin S (2007) Eradicating catastrophic collapse in interdependent networks via reinforced nodes. *Proc Natl Acad Sci USA* 114:3311–3315.
29. Borge-Holthoefer J, Rivero A, Yamir Moreno Y (2012) Locating privileged spreaders on an online social network. *Phys Rev E* 85:066123.
30. Feng L, Monterola CP, Hu Y (2015) The simplified self-consistent probabilities method for percolation and its application to interdependent networks. *New J Phys* 17:063025.
31. Bunde A, Havlin S (1991) *Fractals and Disordered Systems* (Springer, New York).
32. Cohen R, Shlomo Havlin S (2010) *Complex Networks: Structure, Robustness and Function* (Cambridge Univ Press, New York).
33. Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: Quantifying influence on twitter. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (ACM, New York), pp 65–74.
34. Borge-Holthoefer J, Moreno Y (2012) Absence of influential spreaders in rumor dynamics. *Phys Rev E* 85:026116.