Beyond network centrality: Individual-level behavioural traits for predicting and seeding information spreading in social media

Supplementary Material

Fang Zhou[1,2], Linyuan Lü[1*], Jianguo Liu[3], and Manuel Sebastian Mariani[1,4,*]

[1] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, P. R. China
[2] Yangtze Delta Region Institute (Huzhou) , University of Electronic Science and Technology of China, Huzhou 313001, P. R. China
[3] Institute of Accounting and Finance, Shanghai University of Finance and Economics, 200433, P. R. China
[4] URPP Social Networks, Universität Zürich, Switzerland
[*] e-mail:linyuan.lv@uestc.edu.cn; manuel.mariani@business.uzh.ch

Supplementary Table

datasets	N	L	$\langle k \rangle$	σ	URL
Email	$1,133$	$5,451$	9.6	0.078	https://github.com/zervel3/small-scale-social-datasets
Facebook	$2,888$	$2,981$	2.1	-0.668	https://github.com/zervel3/small-scale-social-datasets
HepTh	$9,875$	$25,973$	5.3	0.268	https://github.com/zervel3/small-scale-social-datasets
Hamster	$1,858$	$12,534$	13.5	-0.080	https://github.com/zervel3/small-scale-social-datasets

Table S1 Details of the four small-scale empirical networks. We use four small-scale datasets with synthetic spreading events to validate the IS algorithm's performance, the Table shows the details of the four datasets, N is the size of the network, L is the sum of edges, $\langle k \rangle$ is the average degree, σ is the degree assortativity, the four datasets can be downloaded from the URL.

Supplementary Method 1: Link prediction metrics

In the link prediction section, the similarity metrics used are based on individuals' common neighbours, including common neighbours, Jaccard index [1], Sørensen index [2], Adamic-Adar index [3], Resource allocation index [4], because the large-scale empirical network we used is directed, there are two types of similarity metrics, which are based on in- and out- common neighbours. For a node i, let $\Gamma(i)$ as the set of its nearest neighbours, following we will give the definition of the metrics used in the main text, the common neighbours score between x and y is:

$$s_{xy}^{cn} = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

The Jaccard index is based on common neighbours, its definition as following:

$$s_{xy}^{jaccard} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2}$$

Sørensen index is mainly applied in ecology field, the definition is:

$$s_{xy}^{Sørensen} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y} \tag{3}$$

Here the k_x and k_y are the degree of nodes x and y.

For the Adamic-Adar index, it tends to assign less-connected neighbours more weight, and it defines as:

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log k_z} \tag{4}$$

The Resource Allocation index also depress the high degree neighbours, but with the following form:

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{5}$$

Supplementary Note 1: Validation in synthetic spreading data.

To test the IS algorithm's performance in reconstructing the individual influence and susceptibility, based on the known ground-truth influence and susceptibility scores, we simulate spreading processes with the empirical influence model on a given network to obtain its counterpart diffusion network. Specifically, we start by drawing, for each individual i, the ground-truth influence and susceptibility scores, I_i and S_i, from a uniform distribution between 0 and 1; these scores determine the ground-truth contagion probability between any two individuals i and j: $p_{ij} = A_{ij}\, I_i\, S_j$. We simulate $100\, N$ spreading processes: each individual $i \in \{1, \ldots, N\}$ is the seed individual of 100 cascades. In a simulated spreading process, at each time step, an

individual i who already reshared the piece of information ($r_i = 1$) attempts once to propagate the information to each of her neighbours to reshare; such a propagation event has a probability p_{ij}. Each spreading process stops when no individuals can be influenced by their neighbours. For simplicity, we simulate the spreading on four small-scale empirical networks used in prior works [5–7] (see Supplementary Table S1 for details). After having simulated the spreading process, a diffusion network is constructed for each dataset, where the IS algorithm can be applied to reconstruct the influence ($\hat{I}$) and susceptibility ($\hat{S}$) scores of all individuals.

Datasets	$r(\hat{I}, I)$	$r(\hat{I}, f)$	$r(\hat{I}, k)$	$r(\hat{S}, S)$	$r(\hat{S}, g)$	$r(\hat{S}, k)$
Email	$\mathbf{0.999}$	0.468	0.040	$\mathbf{1.000}$	0.434	-0.021
Facebook	$\mathbf{0.791}$	0.147	0.120	$\mathbf{0.912}$	0.135	0.095
HepTh	$\mathbf{0.970}$	0.388	0.016	$\mathbf{0.968}$	0.391	0.011
Hamster	$\mathbf{0.982}$	0.327	0.018	$\mathbf{0.986}$	0.308	0.003

Table S2 IS algorithm's performance on four small-scale networks, based on synthetic spreading events. I, f, S, g, k denote individuals' ground-truth influence, outgoing contagion rate, susceptibility, incoming contagion rate, and degree, respectively. $\hat{I}$ and $\hat{S}$ denote the reconstructed influence and susceptibility, respectively. $r(a, b)$ denotes the Pearson's correlation coefficient between a given pair (a, b) of variables.

Table S1 illustrates the reconstruction accuracy of the IS algorithm on the four small-scale networks, which reveals three key insights. First, the results show that degree is uncorrelated with reconstructed influence ($\hat{I}$), indicating that in this setting, the social hubs are not guaranteed to be highly influential. Second, although the ground-truth outgoing contagion rate f is positively correlated with the reconstructed influence, the linear correlation is moderate at best – ranging from $r = 0.15$ in *Facebook* to $r = 0.47$ in *Email*. Third, the reconstructed influence derived by IS algorithm achieves a remarkably high correlation with the ground-truth influence, ranging from 0.79 in *Facebook* to 1.00 in *Email*. Similar insights can be obtained for the reconstruction of susceptibility (see Supplementary Figs. S7-S10 for the density plots of the four small-scale datasets). Taken together, these observations indicate that in synthetic spreading data, the IS algorithm effectively reconstructs the ground-truth influence and susceptibility scores.

Supplementary Note 2: Robustness with respect to incomplete data.

In real-world applications, we may have only access to data on a limited number of spreading events. To ascertain the robustness of the algorithm's performance with respect to incomplete data, we study the performance of the algorithm as we progressively remove randomly-selected spreading events from the generated synthetic data.

Specifically, on the four small-scale datasets, we set each individual as seed node and run 5 times Independent Cascade spreading process [8] to get a series of $(5N)$ spreading events, and calculate the edge weight p_{ij} by the whole spreading events, where p_{ij} is the fraction of tweets that j retweets from i, then we input the p_{ij} and network structure to IS algorithm to reconstruct the I and S and take them as the baseline, then remove a fraction of spreading events, calculate the edge weight $\hat{p}_{ij}$ by remaining spreading events, finally, estimate individuals $\hat{I}$ and $\hat{S}$, here we use the Pearson's correlation r and mean absolute error (MAE) to measure how the IS algorithm performs when the fraction of spreading events are removed. The results are averaged by 10 realisations.

We find that the accuracy of the IS algorithm in reconstructing the ground-truth influence and susceptibility scores decrease linearly with the number of removed spreading events (see Supplementary Note 1 for details). However, a relatively high correlation can be maintained even if a substantial portion of the spreading events is removed. For example, when 30% spreading events are removed from $Email$, the scores produced by the IS algorithm still exhibit high correlation with the ground-truth influence and susceptibility scores ($r = 0.69$ and $r = 0.83$, respectively; see Supplementary Note 1 for details). These results indicate that the IS algorithm can not only accurately reconstruct ground-truth influence and susceptibility scores, but also has remarkable robustness against data incompleteness.

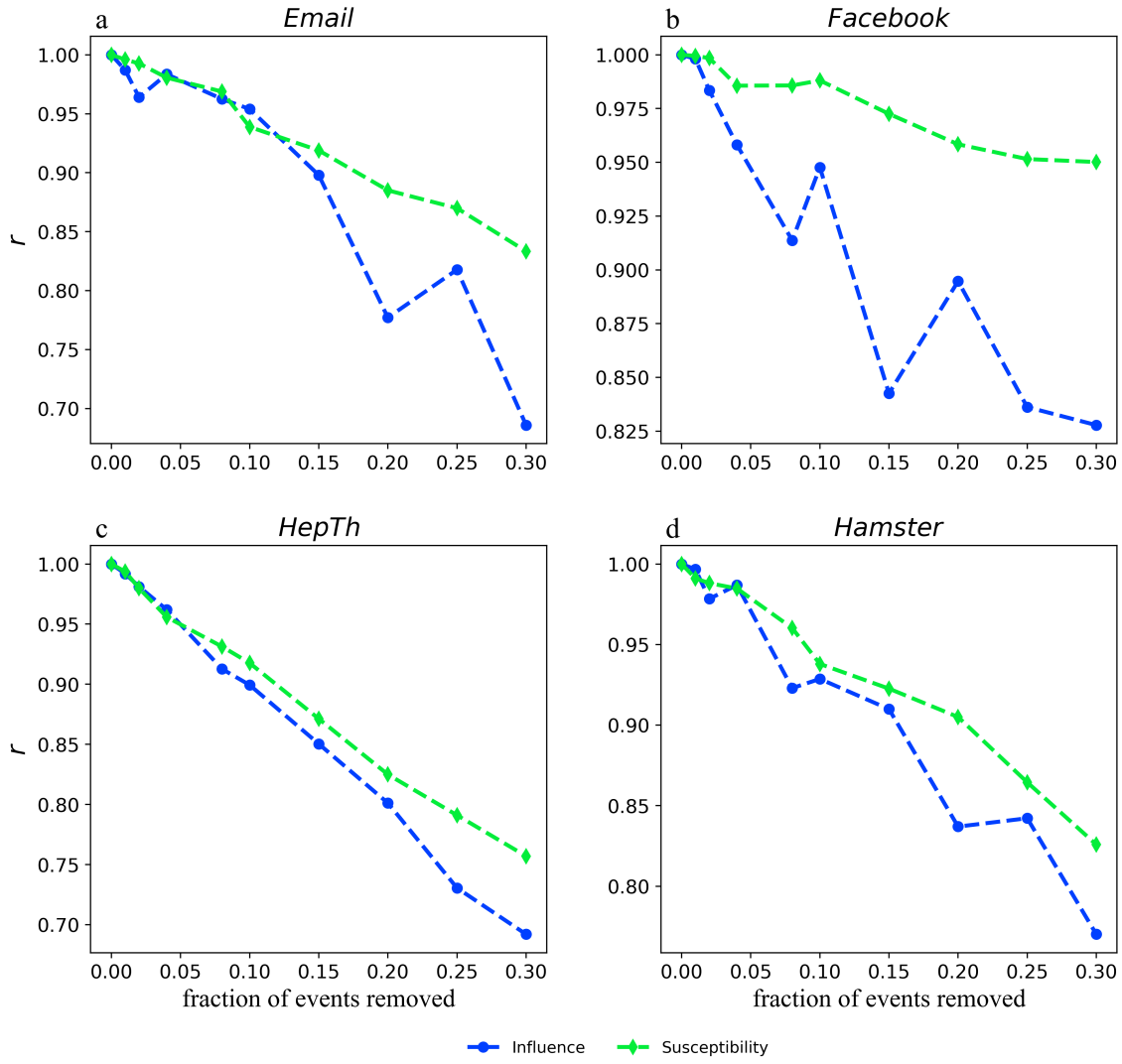

Fig. S1 Robustness of IS algorithm measured by Pearson's coefficient. The x-axis denotes the fraction of spreading events removed, ranging from 0.01 to 0.3, r is the Pearson's coefficient between the original I (S) and the $\hat{I}$ ($\hat{S}$) estimated by the incomplete spreading events.

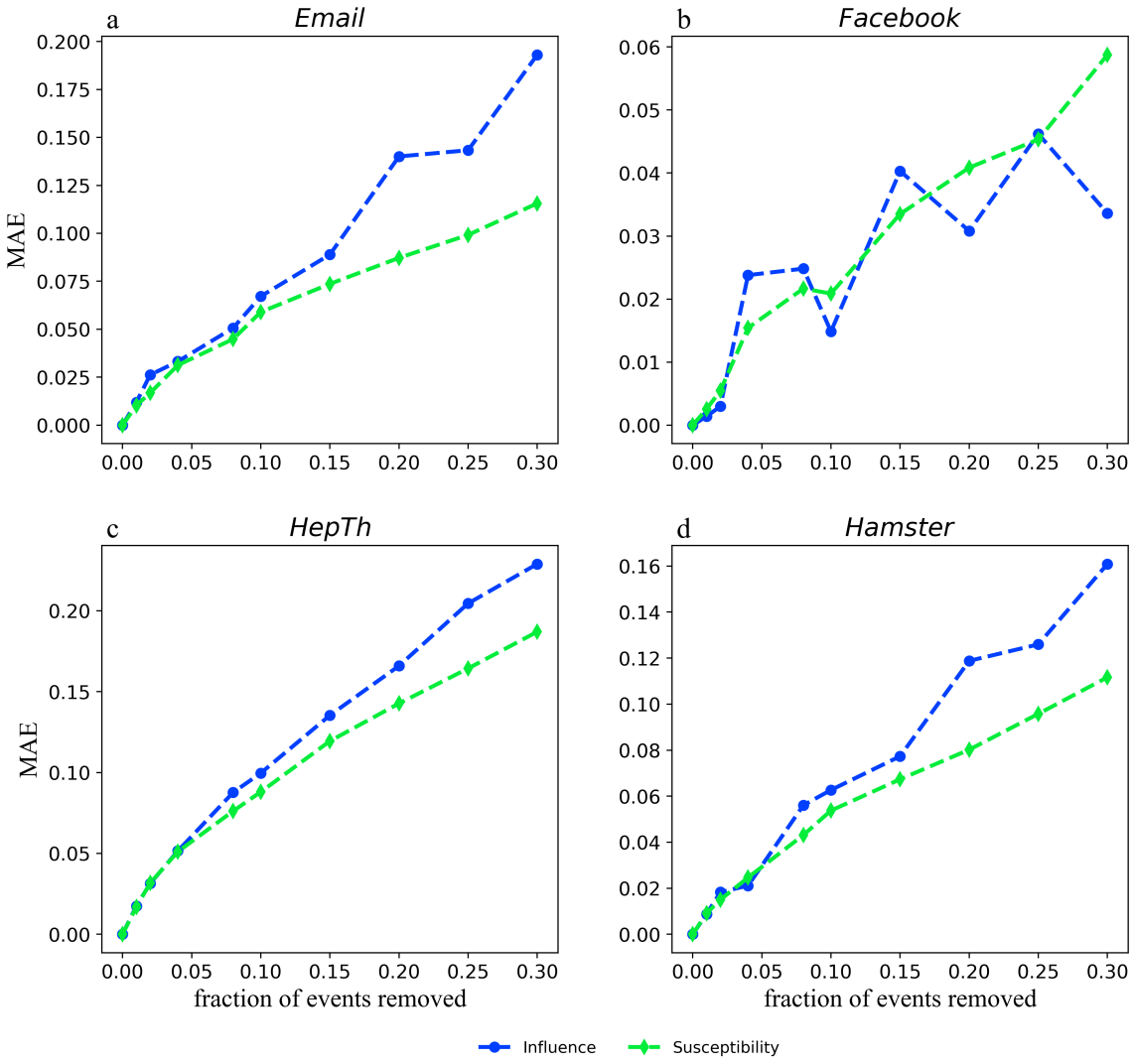

Fig. S2 Robustness of IS algorithm measured by mean absolute error. The x-axis denotes the fraction of spreading events removed, ranging from 0.01 to 0.3, MAE is the mean absolute error between the original I (S) and the $\hat{I}$ ($\hat{S}$) estimated by the incomplete spreading events.

We also applied to the four large-scale empirical datasets to validate the robustness of the IS algorithm, specifically, we set the I and S scores reconstructed by complete spreading events as the baseline, and randomly remove the fraction of retweets (spreading events), then calculate the Pearson's correlation and mean absolute error between the baseline and the scores ($\hat{I}$ and $\hat{S}$) reconstructed by incomplete spreading events.

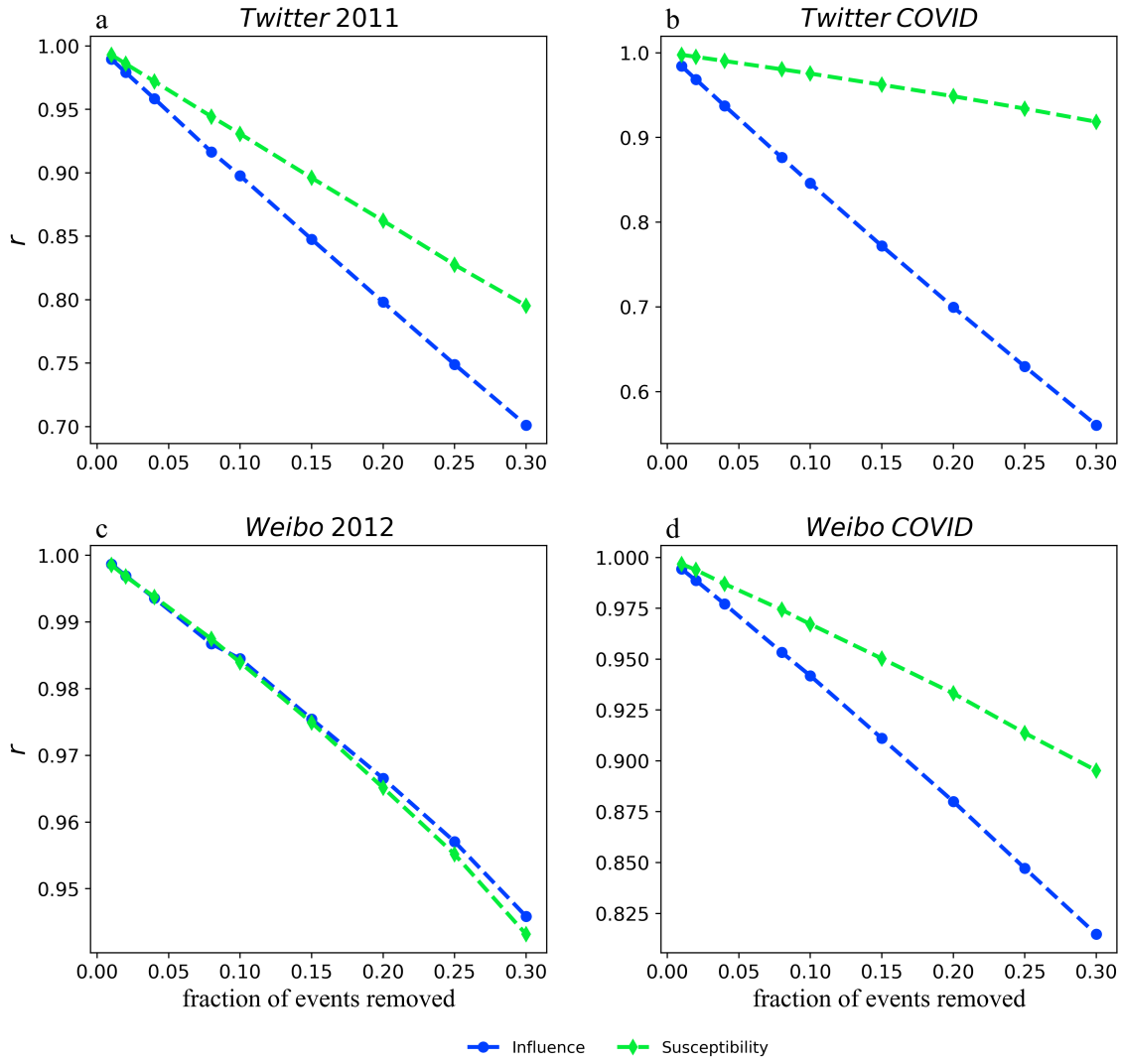

Fig. S3 Robustness of IS algorithm measured by Pearson's coefficient. The x-axis denotes the fraction of spreading events removed, r is the Pearson's coefficient between the original I (S) and the $\hat{I}$ ($\hat{S}$) estimated by the incomplete spreading events.

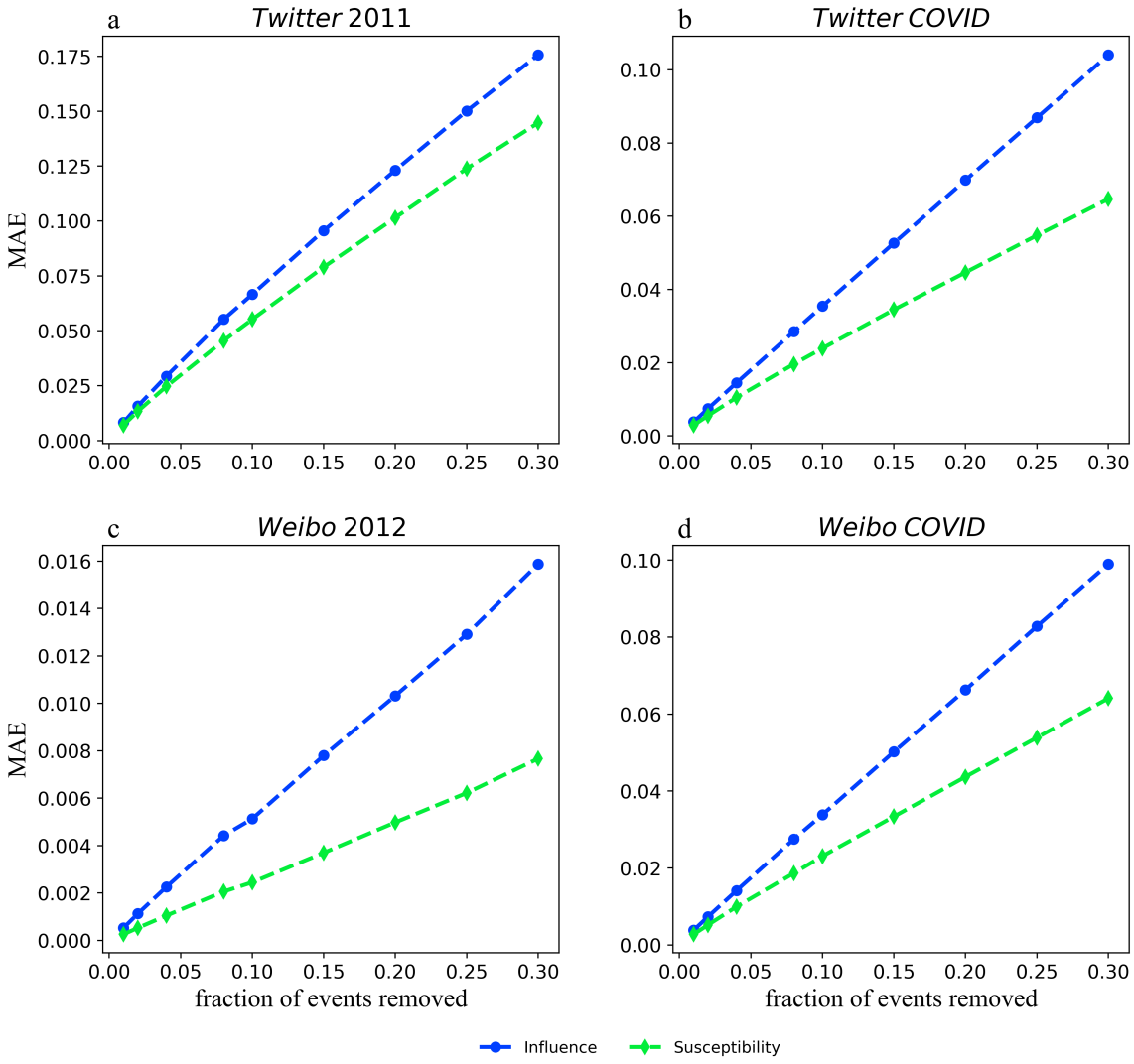

Fig. S4 Robustness of IS algorithm measured by mean absolute error. The x-axis denotes the fraction of spreading events removed, MAE is the mean absolute error between the original I (S) and the $\hat{I}$ ($\hat{S}$) estimated by the incomplete spreading events.

Supplementary Note 3: Verifying that the empirical IS scores are consistent with the empirical results by Aral and Walker [9]

Validating the algorithm in empirical spreading data faces the natural challenge that we lack access to the individuals' ground-truth influence and susceptibility scores. To bypass this challenge, we only validate the aggregate properties of the scores. Specifically, we verify whether, when applied to empirical spreading data from social media, the algorithm produces scores that respect the four stylised facts about influence and susceptibility identified in Aral and Walker's experiment [9]: (i) highly influential individuals tend not to be susceptible, highly susceptible individuals tend not to be influential, and almost no one is both highly influential and highly susceptible to influence; (ii) both influential individuals and non-influential individuals have approximately the same distribution of susceptibility to influence among their peers; (iii) there are more people with high influence scores than high susceptibility scores; (iv) influentials cluster in the network.

To this end, we apply the algorithm to four large-scale empirical datasets: *Twitter 2011*, *Twitter COVID*, *Weibo 2012*, and *Weibo COVID* (see Methods). In empirical social networks, the diffusion network is constructed from the spreading events among users, which is different from the social networks that connect followers and followees [10]. For this reason, we run the IS algorithm on the diffusion network [10] whose directed adjacency matrix's element $A_{ij} = 1$ if i retweeted at least once from j. With this choice, i's indegree $k^{in} = \sum_j A_{ji}$ denotes the number of individuals i retweeted from; her outdegree $k^{out} = \sum_j A_{ij}$ denotes the number of individuals who retweeted from i. Remarkably, for all the four analysed large-scale empirical datasets, the above-mentioned stylised facts (i–iii) identified by Aral and Walker [9] hold, whereas the validity of stylised fact (iv) is dataset-dependent – we refer to Supplementary Note 2 for all the results. This indicates that in empirical spreading data, the algorithm produces scores that respect the properties we would expect for influence and susceptibility scores from the experiment [9].

First, we test whether **highly influential individuals tend not to be susceptible, highly susceptible individuals tend not to be influential, and almost no one is both highly influential and highly susceptible to influence.** For the first two claims, it's proved by the negative and low correlation between I and S for *Twitter 2011*, *Twitter COVID*, *Weibo 2012* and *Weibo COVID* , they are $r(I, S) = -0.56$, $r(I, S) = -0.11$, $r(I, S) = -0.09$ and $r(I, S) = -0.52$ respectively, for the third claim, we choose top 0.5% of individuals by their influence and susceptibility score respectively, for *Twitter 2011*, *Twitter COVID*, *Weibo 2012* and *Weibo COVID*, the overlap of the above two types of individuals are 0.2%, 0.3%, 0.9% and 0.1%, such a value is very low. When we increase the fraction of individuals chosen to top 1%, the overlap are 0.5%, 0.5%, 1.5%, 2% respectively.

Second, we test whether **both influential individuals and noninfluential individuals have approximately the same distribution of susceptibility to influence among their peers.** To test this section, we rank individuals by their influence score, and choose the top 0.5% of influentials' (individuals with high influence score) susceptibility score, then do the Mann-Whitney test between influentials susceptibility score and remaining individuals' susceptibility score, the result shows, for *Twitter 2011*, *Twitter COVID*, *Weibo 2012* and *Weibo COVID*, their influentials' susceptibility score and remaining 99.5% individuals' susceptibility score are from the same distribution with the probability 99%.

Third, we test whether **there are more people with high influence scores than high susceptibility scores.** We need to define what having a high score means. We make a conservative choice, and consider an influence score $\hat{I}$ as high if and only if $\hat{I} > \mu_I + 3\,\sigma_I$, where μ_I and σ_I are the mean and standard deviation of the influence score across the whole population. Analogously, we consider a susceptibility score $\hat{S}$ as high if and only if $\hat{S} > \mu_S + 3\,\sigma_S$, where μ_S and σ_S are the mean and standard deviation of the susceptibility score across the whole population. We denote the fraction of high influence score and high susceptibility score individuals as F_{inf} and F_{sus}, respectively. For *Twitter 2011*, the F_{inf} and F_{sus} are 0.45% and 0.4%. For *Twitter COVID*, the F_{inf} and F_{sus} are 2.7%, 1.8%, respectively. For *Weibo 2012*, the F_{inf} and F_{sus} are 2.0%, 0.1%, respectively. For *Weibo COVID*, under the condition $\hat{I} > \mu_I + 2\,\sigma_I$ and $\hat{S} > \mu_S + 2\,\sigma_S$, the F_{inf} and F_{sus} are 4.3%, 1.9% respectively.

Finally, we test whether **influentials cluster in the network.** We choose the top 0.5% individuals by their influence score as influentials, for *Twitter 2011*, we find 1.6% of influentials have at least an edge to other influentials, which is significantly larger than the same fraction in networks randomised with the directed configuration model is $0.51\% + 0.11\%$. For *Twitter COVID*, 1.98% fraction of influentials have at least an edge to other influentials, the corresponding value in the directed configuration model is $1.12\% \pm 0.07\%$. For *weibo 2012*, 6.41% of influentials have at least an edge to other influentials, the corresponding value in the directed configuration model is $5.73\% \pm 0.31\%$. But for *weibo COVID*, 0.06% of influentials have at least an edge to other influentials, the corresponding value in the directed configuration model is $0.14\% \pm 0.01\%$.

Supplementary Note 4: Analytic convergence condition for the IS algorithm in undirected network

Part of the following proof borrows the idea from paper [11], we find a sufficient and unnecessary condition that guarantees the convergence of equations (5) in the main text. Since the equations (5) are non-linear equations, to prove its convergence, we can convert the equations (3), (4) in the main text to the fixed-point problem, which with the type $x = G(x)$.

For the following two equations (The equations (3), (4) in the main text)

$$I_i = \frac{f_i}{\sum_j A_{ij} S_j} \tag{6}$$

$$S_j = \frac{g_j}{\sum_i A_{ij} I_i} \tag{7}$$

We set $I_i = x_i^{-1}$, $S_j = y_j^{-1}$. The above two equations are transformed as

$$x_i = \frac{\sum_j A_{ij} y_j^{-1}}{f_i} \tag{8}$$

$$y_j = \frac{\sum_i A_{ij} x_i^{-1}}{g_j} \tag{9}$$

Following we will construct a block matrix that with the form $x = G(x)$. After applying the equations (3) and (4) to the whole network, we get the following $2N$-dimensional (N is the size of the network) vector equations.

$$\begin{pmatrix} x_1 \\ \vdots \\ x_N \\ y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{FA} \\ \mathbf{EA^T} & \mathbf{0} \end{pmatrix} \begin{pmatrix} x_1^{-1} \\ \vdots \\ x_N^{-1} \\ y_1^{-1} \\ \vdots \\ y_N^{-1} \end{pmatrix} \tag{10}$$

Where

$$\mathbf{F} = \begin{pmatrix} \frac{1}{f_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{f_N} \end{pmatrix}, \mathbf{E} = \begin{pmatrix} \frac{1}{g_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{g_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{g_N} \end{pmatrix}$$

$\mathbf{A}$ is the adjacent matrix. The equation (10) with the type $x = G(x)$ is what we need, it can be seen as a fixed-point iterative problem, to prove its convergence, we need to prove at one given initial condition x^*, the spectral radius of its Jacobian matrix is small than 1, that is $\rho(\frac{\partial G(x^*)}{\partial x}) < 1$.

Following we will calculate the Jacobian matrix of $G(x)$.

$$\mathbf{J} = \frac{\partial G(x)}{\partial x} = [\frac{\partial G(x)}{\partial x_1}, \cdots, \frac{\partial G(x)}{\partial x_N}, \frac{\partial G(x)}{\partial y_1}, \cdots, \frac{\partial G(x)}{\partial y_N}] \tag{11}$$

The $G(x)$ can be written as $G(x) = \mathbf{M}v(x)$, where

$$\mathbf{M} = \begin{pmatrix} \mathbf{0} & \mathbf{FA} \\ \mathbf{EA^T} & \mathbf{0} \end{pmatrix}, v(x) = \begin{pmatrix} x_1^{-1} \\ \vdots \\ x_N^{-1} \\ y_1^{-1} \\ \vdots \\ y_N^{-1} \end{pmatrix}$$

$\mathbf{M}$ is a constant matrix, so we only need to calculate the derivative of $v(x)$, finally we got

$$\mathbf{J} = \frac{\partial G(x)}{\partial x} = \begin{pmatrix} \mathbf{0} & \mathbf{-FAY^{-2}} \\ \mathbf{-EA^T X^{-2}} & \mathbf{0} \end{pmatrix} \tag{12}$$

$$\mathbf{Y^{-2}} = \begin{pmatrix} y_1^{-2} & 0 & \cdots & 0 \\ 0 & y_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_N^{-2} \end{pmatrix}, \mathbf{X^{-2}} = \begin{pmatrix} x_1^{-2} & 0 & \cdots & 0 \\ 0 & x_2^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_N^{-2} \end{pmatrix}$$

It's hard to calculate the spectral radius of $\mathbf{J}$ directly when applied to a large-scale network. We turn to estimate its spectral radius. The square of $\mathbf{J}$ can be written as

$$\mathbf{J^2} = \begin{pmatrix} \mathbf{FAY^{-2}EA^TX^{-2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{EA^TX^{-2}FAY^{-2}} \end{pmatrix} \tag{13}$$

We noticed that the two blocks $\mathbf{FAY^{-2}EA^TX^{-2}}$ and $\mathbf{EA^TX^{-2}FAY^{-2}}$ with same eigenvalues.
$\mathbf{FAY^{-2}EA^TX^{-2}}$ is

$$\begin{pmatrix} \frac{1}{f_1x_1^2}\sum_j A_{1j}A_{1j}\frac{1}{g_jy_j^2} & \frac{1}{f_1x_2^2}\sum_j A_{1j}A_{2j}\frac{1}{g_jy_j^2} & \cdots & \frac{1}{f_1x_N^2}\sum_j A_{1j}A_{Nj}\frac{1}{g_jy_j^2} \\ \frac{1}{f_2x_1^2}\sum_j A_{2j}A_{1j}\frac{1}{g_jy_j^2} & \frac{1}{f_2x_2^2}\sum_j A_{2j}A_{2j}\frac{1}{g_jy_j^2} & \cdots & \frac{1}{f_2x_N^2}\sum_j A_{2j}A_{Nj}\frac{1}{g_jy_j^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{f_Nx_1^2}\sum_j A_{Nj}A_{1j}\frac{1}{g_jy_j^2} & \frac{1}{f_Nx_2^2}\sum_j A_{Nj}A_{2j}\frac{1}{g_jy_j^2} & \cdots & \frac{1}{f_Nx_N^2}\sum_j A_{Nj}A_{Nj}\frac{1}{g_jy_j^2} \end{pmatrix} \tag{14}$$

By setting x_i^*, y_i^* both equal to non-negative I_0 as the initial condition, the above matrix is simplified as

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{k_1}\sum_j A_{1j}A_{1j}\frac{1}{k_j} & \frac{1}{k_1}\sum_j A_{1j}A_{2j}\frac{1}{k_j} & \cdots & \frac{1}{k_1}\sum_j A_{1j}A_{Nj}\frac{1}{k_j} \\ \frac{1}{k_2}\sum_j A_{2j}A_{1j}\frac{1}{k_j} & \frac{1}{k_2}\sum_j A_{2j}A_{2j}\frac{1}{k_j} & \cdots & \frac{1}{k_2}\sum_j A_{2j}A_{Nj}\frac{1}{k_j} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{k_N}\sum_j A_{Nj}A_{1j}\frac{1}{k_j} & \frac{1}{k_N}\sum_j A_{Nj}A_{2j}\frac{1}{k_j} & \cdots & \frac{1}{k_N}\sum_j A_{Nj}A_{Nj}\frac{1}{k_j} \end{pmatrix} \tag{15}$$

We noticed that the matrix $\mathbf{J^2}$ is the block matrix, that is

$$\mathbf{J^2} = \begin{pmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q^T} \end{pmatrix} \tag{16}$$

To estimate the spectral radius of $\mathbf{J^2}$, we turn to estimate the spectral radius of $\mathbf{Q}$ by the Gershgorin circle theorem [12], which is equivalent to prove the modulo of the maximum eigenvalue (spectral radius) of $\mathbf{Q}$ is small than or equal to 1. According to the deduction of Gershgorin circle theorem, for the matrix $\mathbf{Q}$, $max_i|\lambda_i| \leq max_i(|q_{ii}| + \sum_{j=1,j\neq i}^{n}|q_{ij}|)$ (here q_{ij} is the element of $\mathbf{Q}$), because the element of the matrix $\mathbf{Q}$ is always the positive real number, proving the spectral radius small than or equal to 1 is equivalent to prove the sum of arbitrary row's element is small than or equal to 1, which is to prove $\sum_{j=1}^{n} q_{ij} = \frac{1}{k_i}\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j} \leq 1$ hold. When k_i is 1 and i's neighbour with degree 1, we get $\frac{1}{k_i}\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j} = 1$ hold, otherwise, it's small than 1. When k_i is big than 1, to maximise $\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j}$, we get two cases, first is all i's neighbours don't connected together and with degree 1, which means i's neighbour only connect with i, in such a case, we get $\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j} = k_i$, second is all i's neighbours are fully connected (We can say its clustering coefficient $c_i = 1$), in such a case, $\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j} = k_i$ as well, except for the above two cases, $\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j} < k_i$ always hold, so we proved for matrix $\mathbf{Q}$, the sum of each row's elements small than or equal to 1 when we set x_i^* and y_i^* both equal to nonnegative I_0, according to the Gershgorin circle theorem [12], we know the spectral radius $\rho(\mathbf{Q}) \leq 1$, finally we proved $\rho(\mathbf{J}) \leq 1$. Besides, the case $\rho(\mathbf{J}) = 1$ holds if and only if for arbitrary i, its clustering coefficient c_i is either 0 or 1 (In this case, sum of each row of matrix J^2 is 1, getting $\mathbf{J^2}$ with a eigenvalue 1, then $\rho(\mathbf{J}) = 1$), In other case, we have $\rho(\mathbf{J}) < 1$. Such a result is checked by numerical simulation, the concrete process as follows, for the above four small-scale networks, we set the initial condition $I^{(0)}$ and $S^{(0)}$ from 0.001 to 10, 000 to iterate the algorithm (We chose eight initial conditions, they are 0.001, 0.01, 0.1, 1, 10, 100, 1, 000, 10, 000), and find the final results (I, S) get from different initial condition keep consistent.

Analytic convergence condition in directed network

For the directed network, the process of proof is similar to the undirected network, the difference is that for individual i, only its out-neighbours contribute to f_i, and only its in-neighbours contribute to g_i, similarly, we convert it to the fixed-point problem $x = G(x)$, and get the square of the Jacobian matrix as following:

$$\mathbf{J^2} = \begin{pmatrix} \mathbf{FAY^{-2}EA^TX^{-2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{EA^TX^{-2}FAY^{-2}} \end{pmatrix} \tag{17}$$

The two blocks $\mathbf{FAY^{-2}EA^TX^{-2}}$ and $\mathbf{EA^TX^{-2}FAY^{-2}}$ with same eigenvalues. $\mathbf{FAY^{-2}EA^TX^{-2}}$ equals to:

$$\begin{pmatrix} \frac{1}{f_1x_1^2}\sum_j A_{1j}A_{1j}\frac{1}{g_jy_j^2} & \frac{1}{f_1x_2^2}\sum_j A_{1j}A_{2j}\frac{1}{g_jy_j^2} & \cdots & \frac{1}{f_1x_N^2}\sum_j A_{1j}A_{Nj}\frac{1}{g_jy_j^2} \\ \frac{1}{f_2x_1^2}\sum_j A_{2j}A_{1j}\frac{1}{g_jy_j^2} & \frac{1}{f_2x_2^2}\sum_j A_{2j}A_{2j}\frac{1}{g_jy_j^2} & \cdots & \frac{1}{f_2x_N^2}\sum_j A_{2j}A_{Nj}\frac{1}{g_jy_j^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{f_Nx_1^2}\sum_j A_{Nj}A_{1j}\frac{1}{g_jy_j^2} & \frac{1}{f_Nx_2^2}\sum_j A_{Nj}A_{2j}\frac{1}{g_jy_j^2} & \cdots & \frac{1}{f_Nx_N^2}\sum_j A_{Nj}A_{Nj}\frac{1}{g_jy_j^2} \end{pmatrix} \tag{18}$$

We set x_i^* and y_i^* both equal to I_0 as the initial condition, then the above matrix is simplified as

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{k_1^{out}}\sum_j A_{1j}A_{1j}\frac{1}{k_j^{in}} & \frac{1}{k_1^{out}}\sum_j A_{1j}A_{2j}\frac{1}{k_j^{in}} & \cdots & \frac{1}{k_1^{out}}\sum_j A_{1j}A_{Nj}\frac{1}{k_j^{in}} \\ \frac{1}{k_2^{out}}\sum_j A_{2j}A_{1j}\frac{1}{k_j^{in}} & \frac{1}{k_2^{out}}\sum_j A_{2j}A_{2j}\frac{1}{k_j^{in}} & \cdots & \frac{1}{k_2^{out}}\sum_j A_{2j}A_{Nj}\frac{1}{k_j^{in}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{k_N^{out}}\sum_j A_{Nj}A_{1j}\frac{1}{k_j^{in}} & \frac{1}{k_N^{out}}\sum_j A_{Nj}A_{2j}\frac{1}{k_j^{in}} & \cdots & \frac{1}{k_N^{out}}\sum_j A_{Nj}A_{Nj}\frac{1}{k_j^{in}} \end{pmatrix} \tag{19}$$

Similarly, we use Gershgorin circle theorem [12] to prove that the modulo of the maximum eigenvalue (spectral radius) of the matrix $\mathbf{Q}$ is small than or equal to 1, which can turn to prove the sum of arbitrary row's element is small than or equal to 1, that is to prove $\frac{1}{k_i^{out}}\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j^{in}} \leq 1$ holds. When the out-degree of i is 1 and i's out-neighbour only connects with i, we get $\frac{1}{k_i^{out}}\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j^{in}} = 1$ holds, otherwise, it's small than 1. When k_i^{out} fixed and big than 1, to maximum $\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j^{in}}$, we get two cases, first is all i's out-neighbours don't connect mutually and with $k_{in} = 1$, in such a case, we get $\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j^{in}} = k_i^{out}$, second is all i's out-neighbours are fully connected (We can say its out clustering coefficient $c_i^{out} = 1$), in such a case, $\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j^{in}} = k_i^{out}$ as well, except for the above two cases, in any other condition, we get $\sum_j\sum_l A_{ij}A_{lj}\frac{1}{k_j^{in}} < k_i^{out}$, so we proved for matrix $\mathbf{Q}$, the sum of each row's elements small than or equal to 1 when we set x_i^* and y_i^* both equal to nonnegative I_0, according to the Gershgorin circle theorem, we know the spectral radius $\rho(\mathbf{Q}) \leq 1$, finally we proved $\rho(\mathbf{J}) \leq 1$. Besides, the case $\rho(\mathbf{J}) = 1$ if and only if for arbitrary i, its out clustering coefficient c_i^{out} is either 0 or 1 (In this case, sum of each row of matrix $\mathbf{Q}$ is 1, getting $\mathbf{Q}$ with maximum eigenvalue 1, which finally lead to $\rho(\mathbf{J}) = 1$), In other case, we have $\rho(\mathbf{J}) < 1$. One more thing we need to notice is that if there are some individuals' k_{out} is 0, the above process can also hold, what we need to do is to divide the whole directed network into several directed sub-networks and prove their convergence respectively. The above proof is also checked by numerical simulation, we set the initial condition $I^{(0)}$ and $S^{(0)}$ from 0.001 to $10,000$ (same as un-directed network) to iterate the algorithm, and find the results keep consistent.

Supplementary Note 6: Numeric results on the convergence of the IS algorithm

For both the small-scale and large-scale datasets, the convergence criteria are the error $\sum_i^N |I_i^{(n+1)} - I_i^{(n)}|/N < 10^{-5}$ and $\sum_i^N |S_i^{(n+1)} - S_i^{(n)}|/N < 10^{-5}$ hold, here N is the size of dataset, and $I_i^{(n)}$ ($S_i^{(n)}$) is the influence (susceptibility) value of individuals i in n-th iteration.

Supplementary Note 5: Convergence plots

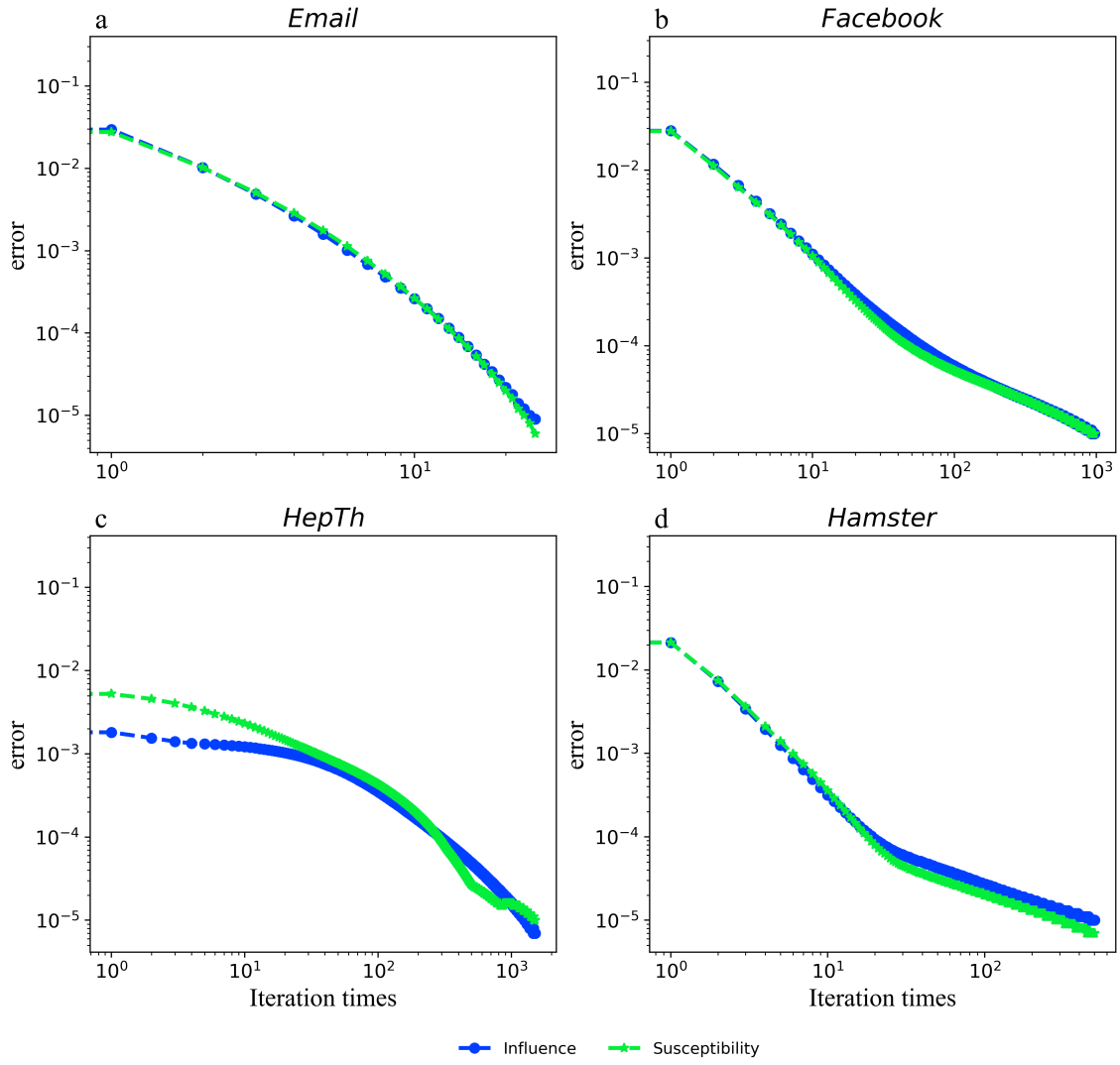

Fig. S5 Convergence plots of the four small-scale datasets. The correlation between iteration times and the error of I (S) of the four small-scale datasets.

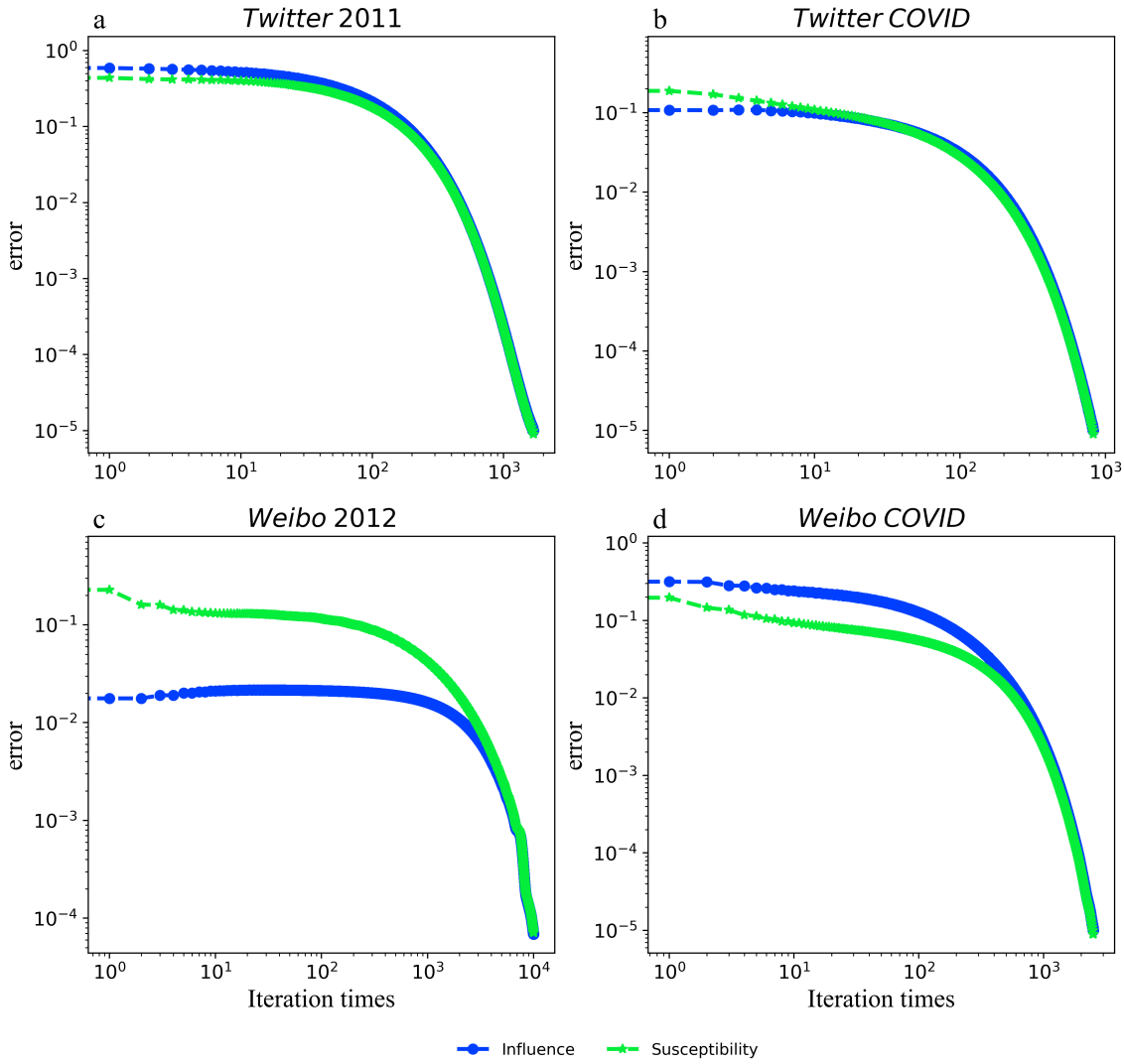

Fig. S6 Convergence plots of the four large-scale datasets. The correlation between iteration times and the error of I (S) of the four large-scale datasets with empirical spreading events.

Supplementary Figures

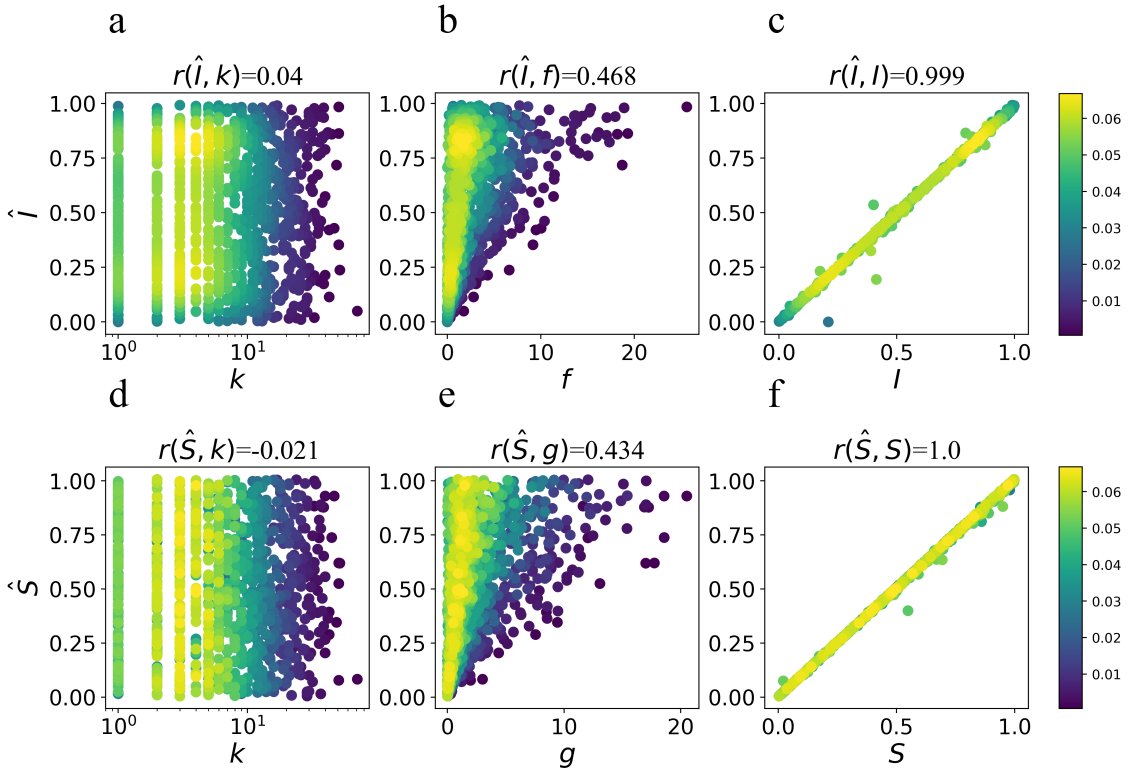

Fig. S7 Reconstructing individuals' ground-truth influence and susceptibility from synthetic spreading events on *Email* network. I, f, S, g, k are individuals' ground-truth influence, outgoing contagion rate, susceptibility, incoming contagion rate, and degree, respectively. $\hat{I}$ and $\hat{S}$ are the reconstructed influence and susceptibility, respectively. $r(\hat{I}, I)$ is the Pearson's coefficient between $\hat{I}$ and I, others are similar. The value of the colour bar denotes the normalised probability density (normalised by the maximum probability density of each row) estimated by Gaussian kernel method [13].

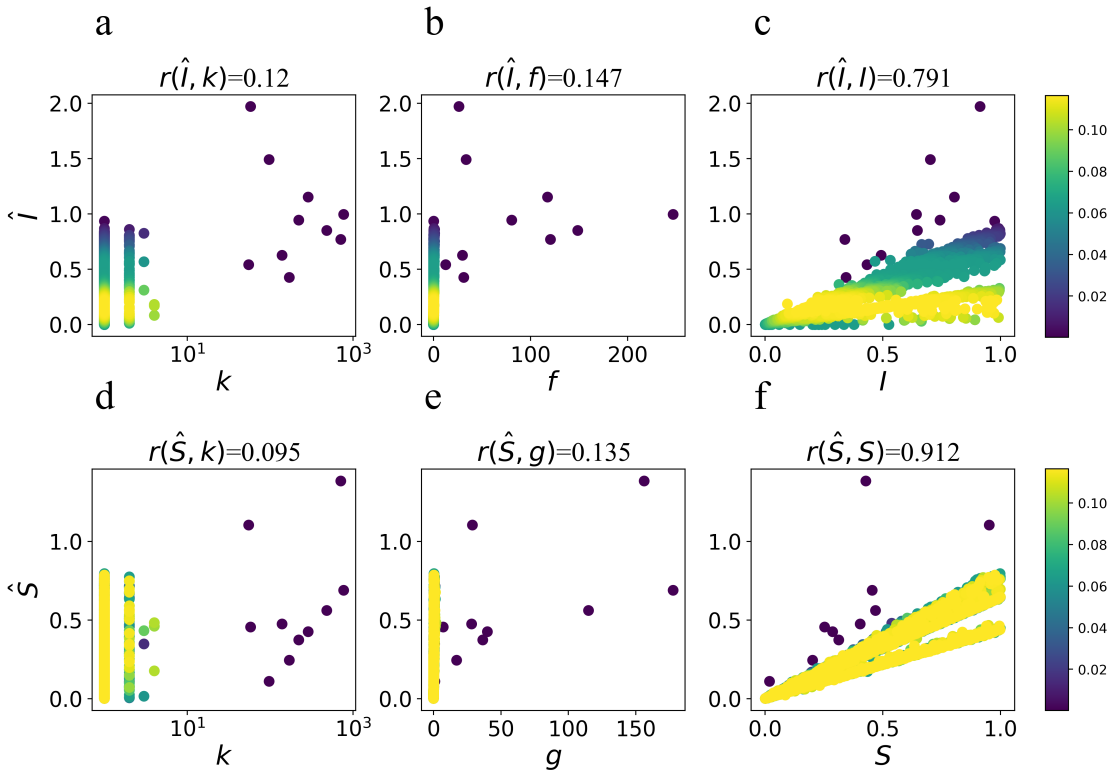

Fig. S8 Reconstructing individuals' ground-truth influence and susceptibility from synthetic spreading events on *Facebook* network. I, f, S, g, k are individuals' ground-truth influence, outgoing contagion rate, susceptibility, incoming contagion rate, and degree, respectively. $\hat{I}$ and $\hat{S}$ are the reconstructed influence and susceptibility, respectively. $r(\hat{I}, I)$ is the Pearson's coefficient between $\hat{I}$ and I, others are similar.

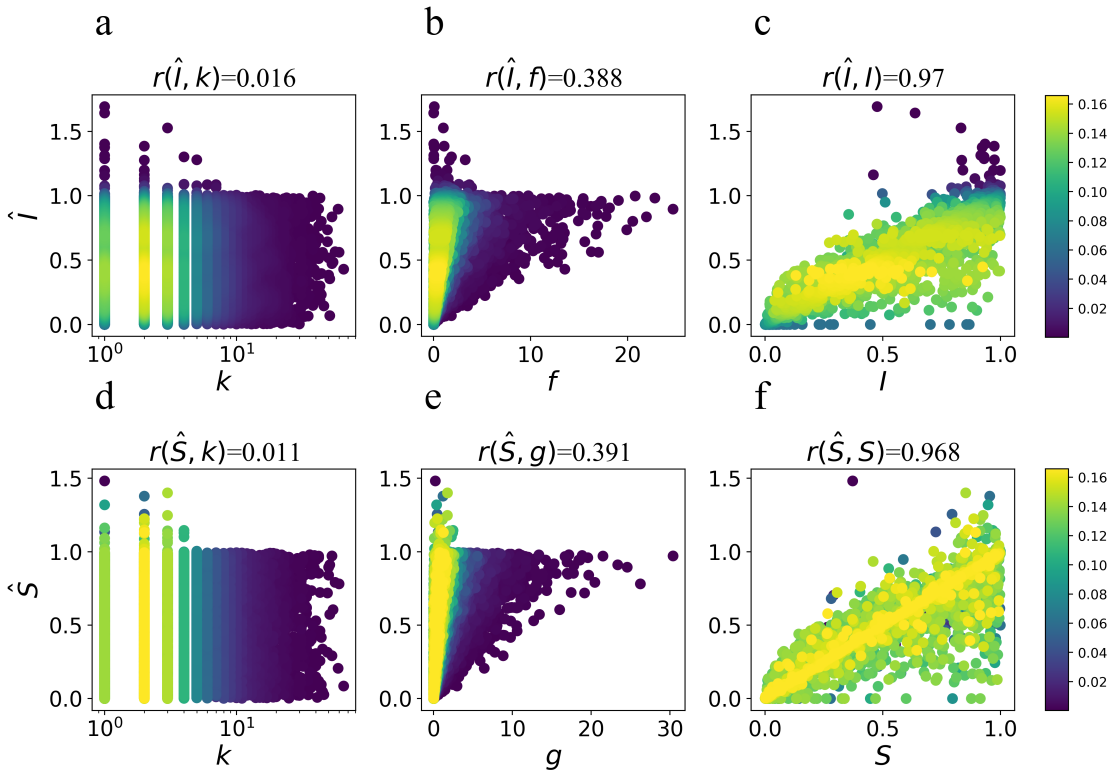

Fig. S9 Reconstructing individuals' ground-truth influence and susceptibility from synthetic spreading events on *HepTh* network. I, f, S, g, k are individuals' ground-truth influence, outgoing contagion rate, susceptibility, incoming contagion rate, and degree, respectively. $\hat{I}$ and $\hat{S}$ are the reconstructed influence and susceptibility, respectively. $r(\hat{I}, I)$ is the Pearson's coefficient between $\hat{I}$ and I, others are similar.

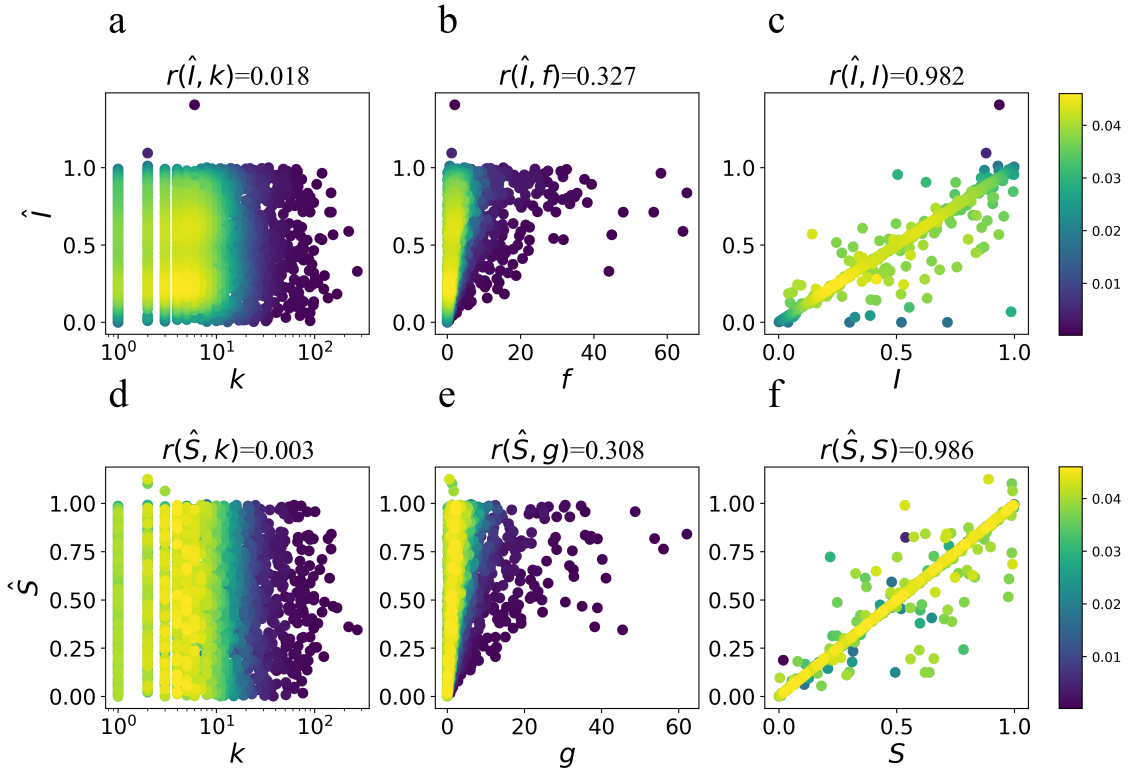

Fig. S10 Reconstructing individuals' ground-truth influence and susceptibility from synthetic spreading events on *Hamster* network. I, f, S, g, k are individuals' ground-truth influence, outgoing contagion rate, susceptibility, incoming contagion rate, and degree, respectively. $\hat{I}$ and $\hat{S}$ are the reconstructed influence and susceptibility, respectively. $r(\hat{I}, I)$ is the Pearson's coefficient between $\hat{I}$ and I, others are similar.

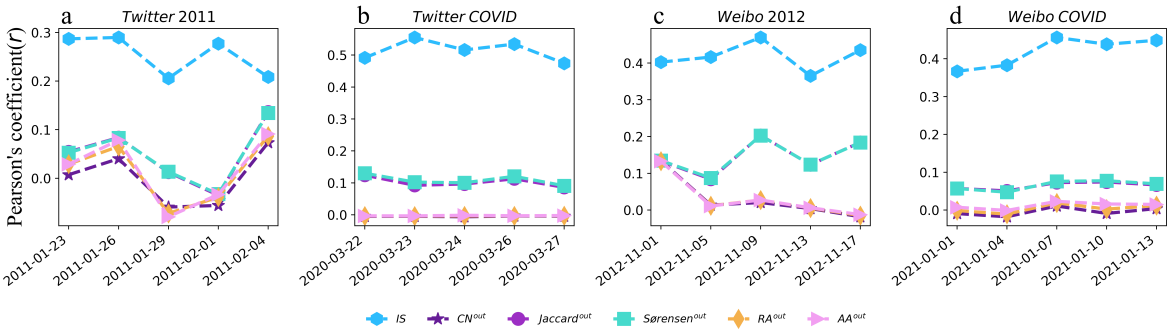

Fig. S11 The performance of IS-based predictor against out-coming common neighbours based metrics on predicting peer-to-peer contagion rate on four large-scale empirical datasets. In panels **a, b, c, d**, we compare the performance of the IS-based predictor against the influence-based, susceptibility-based, indegree- and outdegree-based predictors. We find that to effectively predict pairwise contagion, both individual influence and susceptibility are key properties; focusing on network centrality leads to sub-optimal predictions.

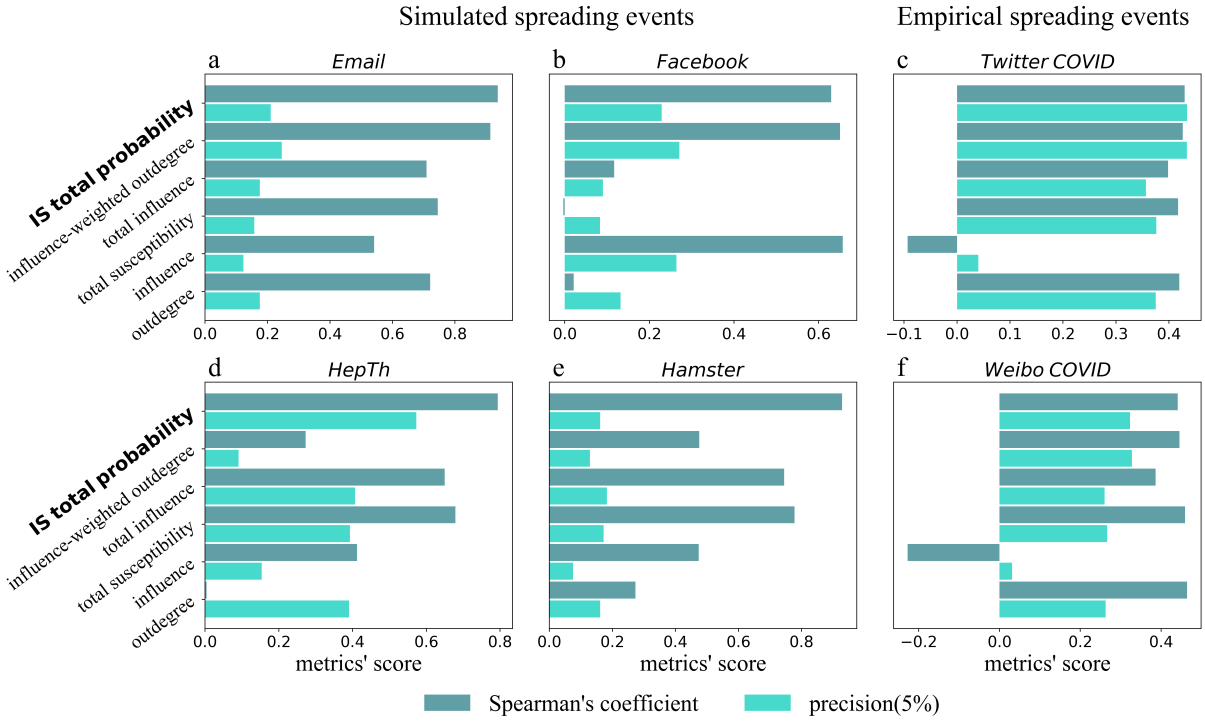

Fig. S12 Superspreader detection in presence of influence and susceptibility. The performance (Spearman's coefficient, top 5%) of six metrics to identify superspreaders. In both simulated and empirical spreading event, we use 80% of spreading events to reconstruct individuals' influence (I) and susceptibility (S) scores, then use the six metrics to identify these superspreaders in the following 20% spreading events.

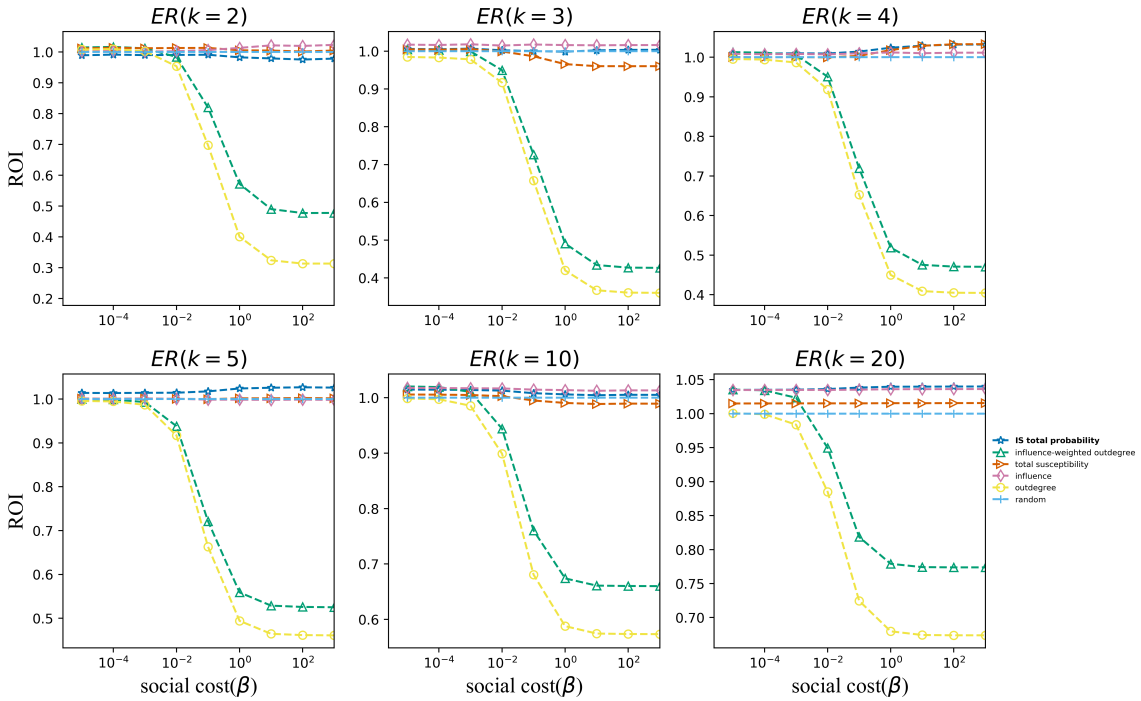

Fig. S13 Implications of the IS algorithm for seeding policies in ER networks. In artificial networks of different density generated with the ER model, we consider the ROI on seeding policies based on six different metrics as a function of the social cost parameter, β. The results for the random metric are averaged over 100 realisations, and all the six metrics' ROI values are normalised by the average ROI of the random metric. The results confirm the empirical finding that the IS total probability tends to outperform the random policy, especially for higher network density. However the ROI improvements are smaller compared to the empirical data, which suggests that the IS-based policy benefits from degree heterogeneity (see also the next Figure).

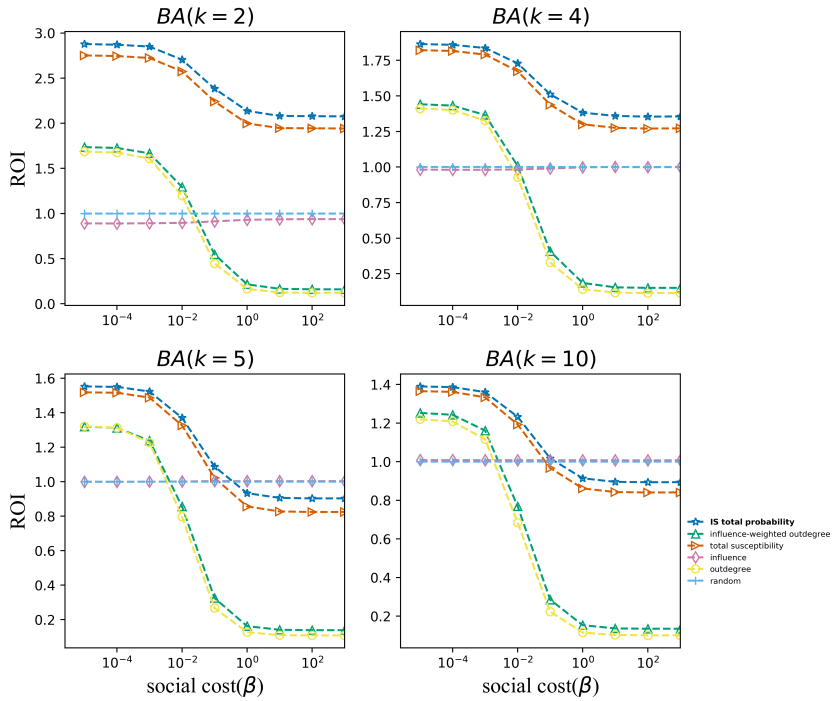

Fig. S14 Implications of the IS algorithm for seeding policies in BA networks. In artificial networks of different density generated with the BA model, we consider the ROI on seeding policies based on six different metrics as a function of the social cost parameter, β. The results for the random metric are averaged over 100 realisations, and all the six metrics' ROI values are normalised by the average ROI of the random metric. The results confirm the empirical finding that ROI gains from the IS total probability critically depend on the social cost parameter. For higher-density networks, the results are in qualitative agreement with the empirical ones in Fig. 6 of the main text: when $\beta < \alpha = 1$, the IS total probability outperforms the random policy, whereas when the social cost is high, the random policy performs best. We further observe that for low densities, the IS total probability outperforms the random policy for all values of the social cost. Taken together, these findings indicate that the relateve ROI of the seeding policies depend not only on the degree heterogeneity, but also on the network density of the underlying diffusion network. Finally, the findings further confirm the result that under no parameter values are the social hubs (i.e., high-outdegree nodes) the best targets.

References

1. Etude, P. J. Comparative de la distribution florale dans une portion des alpes et des jura. *Bull. Soc. Vaud. Sci. Nat* **37**, 547 (1901).
2. Sørensen, T. J. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons* (I kommission hos E. Munksgaard, 1948).
3. Adamic, L. A. & Adar, E. Friends and neighbors on the web. *Social Networks* **25**, 211–230 (2003).
4. Zhou, T., Lü, L. & Zhang, Y.-C. Predicting missing links via local information. *The European Physical Journal B* **71**, 623–630 (2009).
5. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data* **1**, 2–es (2007).
6. Zhou, F., Lü, L. & Mariani, M. S. Fast influencers in complex networks. *Communications in Nonlinear Science and Numerical Simulation* **74**, 69–83 (2019).
7. McAuley, J. J. & Leskovec, J. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems*, 539–547 (2012).
8. Aral, S. & Dhillon, P. S. Social influence maximization under empirical influence models. *Nature Human Behaviour* **2**, 375–382 (2018).
9. Aral, S. & Walker, D. Identifying influential and susceptible members of social networks. *Science* **337**, 337–341 (2012).
10. Pei, S., Muchnik, L., Andrade Jr, J. S., Zheng, Z. & Makse, H. A. Searching for superspreaders of information in real-world social media. *Scientific Reports* **4**, 1–12 (2014).
11. Servedio, V. D., Buttà, P., Mazzilli, D., Tacchella, A. & Pietronero, L. A new and stable estimation method of country economic fitness and product complexity. *Entropy* **20**, 783 (2018).
12. Gerschgorin, S. Uber die abgrenzung der eigenwerte einer matrix. *Izv. Akad. Nauk. USSR. Otd. Fiz-Mat. Nauk* **7**, 749–754 (1931).
13. Silverman, B. W. *Density estimation for statistics and data analysis* (Routledge, 2018).