

Influence maximization in complex networks through optimal percolation

Flaviano Morone¹ & Hernán A. Makse¹

The whole frame of interconnections in complex networks hinges on a specific set of structural nodes, much smaller than the total size, which, if activated, would cause the spread of information to the whole network¹, or, if immunized, would prevent the diffusion of a large scale epidemic^{2,3}. Localizing this optimal, that is, minimal, set of structural nodes, called influencers, is one of the most important problems in network science^{4,5}. Despite the vast use of heuristic strategies to identify influential spreaders^{6–14}, the problem remains unsolved. Here we map the problem onto optimal percolation in random networks to identify the minimal set of influencers, which arises by minimizing the energy of a many-body system, where the form of the interactions is fixed by the non-backtracking matrix¹⁵ of the network. Big data analyses reveal that the set of optimal influencers is much smaller than the one predicted by previous heuristic centralities. Remarkably, a large number of previously neglected weakly connected nodes emerges among the optimal influencers. These are topologically tagged as low-degree nodes surrounded by hierarchical coronas of hubs, and are uncovered only through the optimal collective interplay of all the influencers in the network. The present theoretical framework may hold a larger degree of universality, being applicable to other hard optimization problems exhibiting a continuous transition from a known phase¹⁶.

The optimal influence problem was initially introduced in the context of viral marketing¹, and its solution was shown to be NP-hard⁴ for a generic class of linear threshold models of information spreading^{17,18}. Indeed, finding the optimal set of influencers is a many-body problem in which the topological interactions between them play a crucial role^{13,14}. On the other hand, there has been an abundant production of heuristic rankings to identify influential nodes and ‘superspreaders’ in networks^{6–12,19}. The main problem is that heuristic methods do not optimize a global function of influence. As a consequence, there is no guarantee of their performance.

Here we address the problem of quantifying nodes’ influence by finding the optimal (that is, minimal) set of structural influencers. After defining a unified mathematical framework for both immunization and spreading, we provide its optimal solution in random networks by mapping the problem onto optimal percolation. In addition, we present CI (Collective Influence), a scalable algorithm to solve the optimization problem in large-scale real data sets. The thorough comparison with competing methods (Supplementary Information section I²⁰) ultimately establishes the better performance of our algorithm. By taking into account collective influence effects, our optimization theory identifies a new class of strategic influencers, called ‘weak nodes’, which outrank the hubs in the network. Thus, the top influencers are highly counterintuitive: low-degree nodes play a major broker role in the network, and despite being weakly connected, can be powerful influencers.

The problem of finding the minimal set of activated nodes^{17,18} to spread information to the whole network⁴ or to optimally immunize a network against epidemics¹¹ can be exactly mapped onto optimal percolation (see Supplementary Information section IIB). This mapping

provides the mathematical support to the intuitive relation between influence and the concept of cohesion of a network: the most influential nodes are the ones forming the minimal set that guarantees a global connection of the network^{5,9,10}. We call this minimal set the ‘optimal influencers’ of the network. At a general level, the optimal influence problem can be stated as follows: find the minimal set of nodes which, if removed, would break down the network into many disconnected pieces. The natural measure of influence is, therefore, the size of the largest (giant) connected component as the influencers are removed from the network.

We consider a network composed of N nodes tied with M links with an arbitrary-degree distribution. Let us suppose we remove a certain fraction q of the total number of nodes. It is well known from percolation theory²¹ that, if we choose these nodes randomly, the network undergoes a structural collapse at a certain critical fraction where the probability of existence of the giant connected component vanishes, $G = 0$. The optimal influence problem corresponds to finding the minimum fraction q_c of influencers to fragment the network: $q_c = \min\{q \in [0, 1]: G(q) = 0\}$.

Let the vector $\mathbf{n} = (n_1, \dots, n_N)$ represent which node is removed ($n_i = 0$, influencer) or left ($n_i = 1$, the rest) in the network ($q = 1 - 1/N \sum_i n_i$), and consider a link from i to j ($i \rightarrow j$). The order parameter of the influence problem is the probability that i belongs to the giant component in a modified network where j is absent, $v_{i \rightarrow j}$ (refs 22, 23). Clearly, in the absence of a giant component we find $\{v_{i \rightarrow j} = 0\}$ for all $i \rightarrow j$. The stability of the solution $\{v_{i \rightarrow j} = 0\}$ is controlled by the largest eigenvalue $\lambda(\mathbf{n}; q)$ of the linear operator $\hat{\mathcal{M}}$,

defined on the $2M \times 2M$ directed edges as $\mathcal{M}_{k \rightarrow \ell, i \rightarrow j} \equiv \frac{\partial v_{i \rightarrow j}}{\partial v_{k \rightarrow \ell}} \Big|_{\{v_{i \rightarrow j} = 0\}}$.

We find for locally tree-like random graphs (see Fig. 1a and Supplementary Information section II):

$$\mathcal{M}_{k \rightarrow \ell, i \rightarrow j} = n_i \mathcal{B}_{k \rightarrow \ell, i \rightarrow j} \quad (1)$$

where $\mathcal{B}_{k \rightarrow \ell, i \rightarrow j}$ is the non-backtracking matrix of the network^{15,24}. The matrix $\mathcal{B}_{k \rightarrow \ell, i \rightarrow j}$ has non-zero entries only when $(k \rightarrow \ell, i \rightarrow j)$ form a pair of consecutive non-backtracking directed edges, that is, $(k \rightarrow \ell, \ell \rightarrow j)$ with $k \neq j$. In this case $\mathcal{B}_{k \rightarrow \ell, \ell \rightarrow j} = 1$ (equation (13) in Supplementary Information). Powers of the matrix $\hat{\mathcal{B}}$ count the number of non-backtracking walks of a given length in the network (Fig. 1b)²⁴, much in the same way as powers of the adjacency matrix count the number of paths⁵. Operator $\hat{\mathcal{B}}$ has recently received a lot of attention thanks to its high performance in the problem of community detection^{25,26}. We show its topological power in the problem of optimal percolation.

Stability of the solution $\{v_{i \rightarrow j} = 0\}$ requires $\lambda(\mathbf{n}; q) \leq 1$. The optimal influence problem for a given q ($\geq q_c$) can be rephrased as finding the optimal configuration \mathbf{n} that minimizes the largest eigenvalue $\lambda(\mathbf{n}; q)$ (Fig. 1c). The optimal set \mathbf{n}^* of Nq_c influencers is obtained when the minimum of the largest eigenvalue reaches the critical threshold:

$$\lambda(\mathbf{n}^*; q_c) = 1 \quad (2)$$

¹Levich Institute and Physics Department, City College of New York, New York, New York 10031, USA.

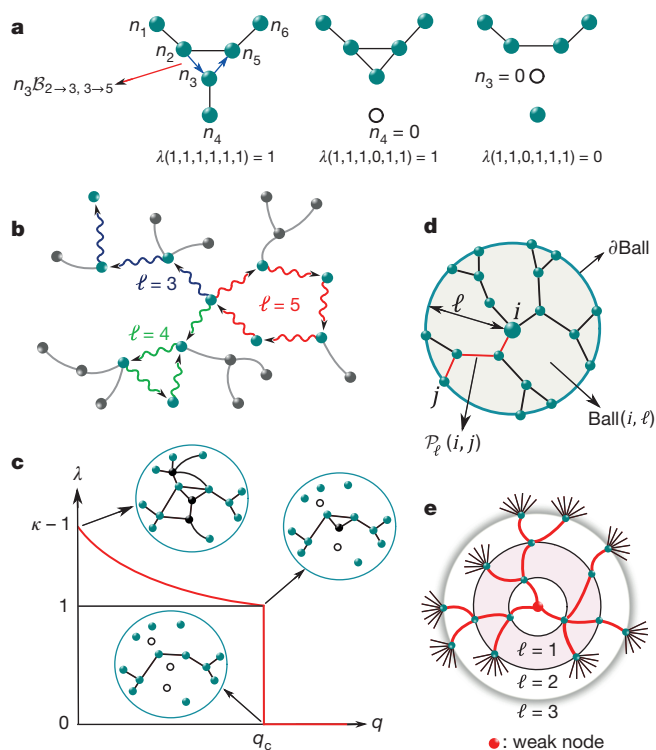


Figure 1 | The non-backtracking (NB) matrix and weak nodes. **a**, The largest eigenvalue λ of $\hat{\mathcal{M}}$ exemplified on a simple network. The optimal strategy for immunization and spreading minimizes λ by removing the minimum number of nodes (optimal influencers) that destroys all the loops. Left panel, the action of the matrix $\hat{\mathcal{M}}$ is on the directed edges of the network. The entry $\mathcal{M}_{2 \rightarrow 3, 3 \rightarrow 5} = n_3 \mathcal{B}_{2 \rightarrow 3, 3 \rightarrow 5} = n_3$ encodes the occupancy ($n_3 = 1$) or vacancy ($n_3 = 0$) of node 3. In this particular case, the largest eigenvalue is $\lambda = 1$. Centre panel, non-optimal removal of a leaf, $n_4 = 0$, which does not decrease λ . Right panel, optimal removal of a loop, $n_3 = 0$, which decreases λ to zero. **b**, A NB walk is a random walk that is not allowed to return back along the edge that it just traversed. We show a NB open walk ($\ell = 3$), a NB closed walk with a tail ($\ell = 4$), and a NB closed walk with no tails ($\ell = 5$). The NB walks are the building blocks of the diagrammatic expansion to calculate λ . **c**, Representation of the global minimum over \mathbf{n} of the largest eigenvalue λ of $\hat{\mathcal{M}}$ versus q . When $q \geq q_c$, the minimum is at $\lambda = 0$. Then, $G = 0$ is stable (still, non-optimal configurations exist with $\lambda > 1$ for which $G > 0$). When $q < q_c$, the minimum of the largest eigenvalue is always $\lambda > 1$, the solution $G = 0$ is unstable, and then $G > 0$. At the optimal percolation transition, the minimum is at \mathbf{n}^* with $\lambda(\mathbf{n}^*, q_c) = 1$. For $q = 0$, we find $\lambda = \kappa - 1$ ($\kappa = \langle k^2 \rangle / \langle k \rangle$), where k is the node degree) which is the largest eigenvalue of $\hat{\mathcal{B}}$ for random networks²⁵ with all nodes present ($n_i = 1$). When $\lambda = 1$, the giant component is reduced to a tree plus one single loop (unicyclic graph), which is suddenly destroyed at the transition q_c to become a tree, causing the abrupt fall of λ to zero. **d**, $\text{Ball}(i, \ell)$ of radius ℓ around node i is the set of nodes at distance ℓ from i , and $\partial \text{Ball}(i, \ell)$ is the set of nodes on the boundary. The shortest path from i to j is shown in red. **e**, Example of a weak node: a node with a small number of connections surrounded by hierarchical coronas of hubs at different ℓ levels.

The formal mathematical mapping of the optimal influence problem to the minimization of the largest eigenvalue of the modified non-backtracking matrix for random networks, equation (2), represents our first main result.

An example of a non-optimized solution corresponds to choosing n_i at random and decoupled from the non-backtracking matrix^{23,27} (random percolation²¹, Supplementary Information section IID). In the optimized case, we seek to derandomize the selection of the set $n_i = 0$ and optimally choose them to find the best configuration \mathbf{n}^* with the lowest q_c according to equation (2). The eigenvalue $\lambda(\mathbf{n})$ (from now on we omit q in $\lambda(\mathbf{n}; q) \equiv \lambda(\mathbf{n})$, which is always kept fixed) determines the growth rate of an arbitrary vector \mathbf{w}_0 with $2M$ entries after ℓ iterations of the matrix

$\hat{\mathcal{M}}: |\mathbf{w}_\ell(\mathbf{n})| = \langle \mathbf{w}_\ell | \mathbf{w}_\ell \rangle^{\frac{1}{2}} = |\hat{\mathcal{M}}^\ell \mathbf{w}_0| = \left\langle \mathbf{w}_0 \left| (\hat{\mathcal{M}}^\ell)^\dagger \hat{\mathcal{M}}^\ell \right| \mathbf{w}_0 \right\rangle^{\frac{1}{2}} \sim e^{\ell \log \lambda(\mathbf{n})}$. The largest eigenvalue is then calculated by the power method:

$$\lambda(\mathbf{n}) = \lim_{\ell \rightarrow \infty} \left[\frac{|\mathbf{w}_\ell(\mathbf{n})|}{|\mathbf{w}_0|} \right]^{1/\ell} \quad (3)$$

Equation (3) is the starting point of an (infinite) perturbation series that provides the exact solution to the many-body influence problem in random networks and therefore contains all physical effects, including the collective influence. In practice, we minimize the cost energy function of influence $|\mathbf{w}_\ell(\mathbf{n})|$ in equation (3) for a finite ℓ . The solution rapidly converges to the exact value as $\ell \rightarrow \infty$, the faster the larger the spectral gap. We find for $\ell \geq 1$, to leading order in $1/N$ (Supplementary Information section IIE):

$$|\mathbf{w}_\ell(\mathbf{n})|^2 = \sum_{i=1}^N (k_i - 1) \sum_{j \in \partial \text{Ball}(i, 2\ell - 1)} \left(\prod_{k \in \mathcal{P}_{2\ell - 1}(i, j)} n_k \right) (k_j - 1) \quad (4)$$

where $\text{Ball}(i, \ell)$ is the set of nodes inside a ball of radius ℓ (defined as the shortest path) around node i , $\partial \text{Ball}(i, \ell)$ is the frontier of the ball, $\mathcal{P}_\ell(i, j)$ is the shortest path of length ℓ connecting i and j (Fig. 1d), and k_i is the degree of node i .

The first collective optimization in equation (4) is $\ell = 1$. We find $|\mathbf{w}_1(\mathbf{n})|^2 = \sum_{i,j=1}^N A_{ij} (k_i - 1)(k_j - 1) n_i n_j$, where A_{ij} is the adjacency matrix (equation (39) in Supplementary Information). This term is interpreted as the energy of an antiferromagnetic Ising model with random bonds in a random external field at fixed magnetization, which is an example of a pair-wise NP-complete spin-glass whose solution is found in Supplementary Information section III with the cavity method²⁸ (Extended Data Fig. 2).

For $\ell \geq 2$, the problem can be mapped exactly to a statistical mechanical system with many-body interactions which can be recast in terms of a diagrammatic expansion, equations (41)–(49) in Supplementary Information. For example, $|\mathbf{w}_2(\mathbf{n})|^2$ leads to 4-body interactions (equation (45) in Supplementary Information), and, in general, the energy cost $|\mathbf{w}_\ell(\mathbf{n})|^2$ contains 2ℓ -body interactions. As soon as $\ell \geq 2$, the cavity method becomes much more complicated to implement and we use another suitable method, called extremal optimization (EO)²⁹ (Supplementary Information section IV). This method estimates the true optimal value of the threshold by finite-size scaling following extrapolation to $\ell \rightarrow \infty$ (Extended Data Figs 3, 4). However, EO is not scalable to find the optimal configuration in large networks. Therefore, we develop an adaptive method, which performs excellently in practice, preserves the features of EO, and is highly scalable to present-day big data.

The idea is to remove the nodes causing the biggest drop in the energy function, equation (4). First, we define a ball of radius ℓ around every node (Fig. 1d). Then, we consider the nodes belonging to the frontier $\partial \text{Ball}(i, \ell)$ and assign to node i the collective influence (CI) strength at level ℓ following equation (4):

$$\text{CI}_\ell(i) = (k_i - 1) \sum_{j \in \partial \text{Ball}(i, \ell)} (k_j - 1) \quad (5)$$

We notice that, while equation (4) is valid only for odd radii of the ball, $\text{CI}_\ell(i)$ is defined also for even radii. This generalization is possible by considering an energy function for even radii analogous to equation (4), as explained in Supplementary Information section IIG. The case of one-body interaction with zero radius $\ell = 0$ (equation (59) in Supplementary Information) leads to the high-degree (HD) ranking (equation (62) in Supplementary Information)¹⁰.

The collective influence, equation (5), is our second and most important result since it is the basis for the highly scalable and optimized CI algorithm which follows. In the beginning, all the nodes are present: $n_i = 1$ for all i . Then, we remove node i^* with highest CI_ℓ and set $n_{i^*} = 0$. The degree of each neighbour of i^* is decreased by one, and the procedure is repeated to find the new top CI node to remove. The algorithm is terminated when the giant component is zero (see Supplementary Information section V for implementation, and

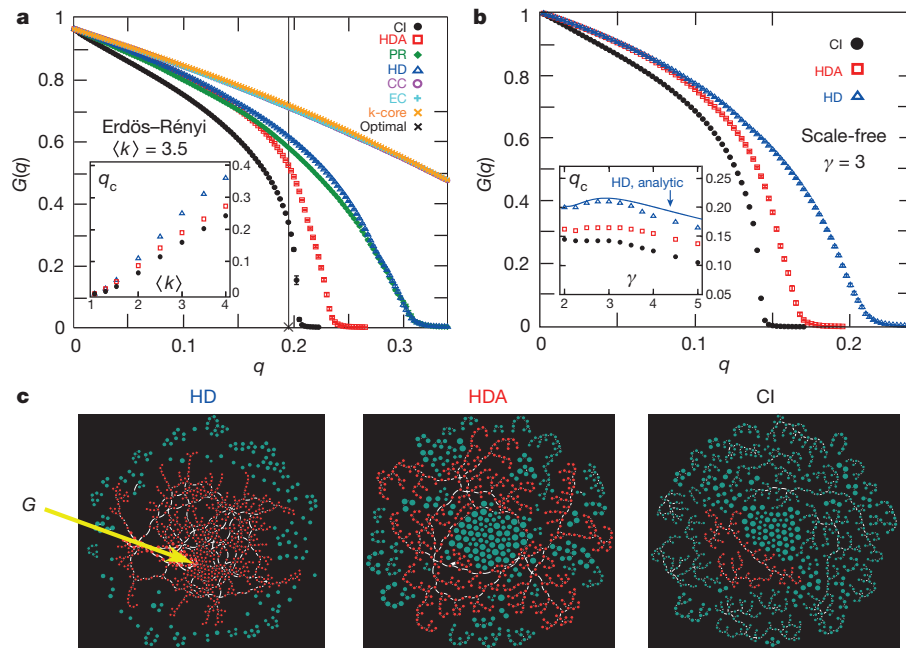


Figure 2 | Exact optimal solution and performance of CI in synthetic networks. **a**, $G(q)$ in an ER network ($N = 2 \times 10^5$, $\langle k \rangle = 3.5$, error bars are s.e.m. over 20 realizations). We show the true optimal solution found with EO ('x' symbol), and also using CI, HDA, PR, HD, CC, EC and k-core methods. The other methods are not scalable and perform worse than HDA and are treated in Supplementary Information sections VI and VII (Extended Data Figs 8, 9). CI is close to the optimal $q_c^{\text{opt}} = 0.192(9)$ obtained with EO in Supplementary Information section IV. Note that EO can estimate the extrapolated optimal value of q_c , but it cannot provide the optimal

configuration for large systems. Inset, q_c (obtained at the peak of the second-largest cluster) for the three best methods versus $\langle k \rangle$. **b**, $G(q)$ for a SF network with degree exponent $\gamma = 3$, maximum degree $k_{\text{max}} = 10^3$, minimum degree $k_{\text{min}} = 2$ and $N = 2 \times 10^5$ (error bars are s.e.m. over 20 realizations). Inset, q_c versus γ . The continuous blue line is the HD analytical result computed in Supplementary Information section IIG (Extended Data Fig. 1b). **c**, Example of SF network with $\gamma = 3$ after the removal of 15% of nodes, using the three methods HD, HDA and CI. CI produces a much reduced giant component G (red nodes).

Supplementary Information section VA for minimizing $G(q) \neq 0$. By increasing the radius ℓ of the ball we obtain better and better approximations of the optimal exact solution as $\ell \rightarrow \infty$ (for finite networks, ℓ does not exceed the network diameter).

The collective influence CI_ℓ for $\ell \geq 1$ has a rich topological content, and consequently tells us more about the role played by nodes in the network than the non-interacting high-degree hub-removal strategy at $\ell = 0$, CI_0 . The augmented information comes from the sum in the right hand side of equation (5), which is absent in the naive high-degree rank. This sum contains the contribution of the nodes living on the surface of the ball surrounding the central vertex i , each node weighted by the factor $k_j - 1$. This means that a node placed at the centre of a corona irradiating many links—the structure hierarchically emerging at different ℓ levels as seen in Fig. 1e—can have a very large collective influence, even if it has a moderate or low degree. Such ‘weak nodes’ can outrank nodes with larger degree that occupy mediocre peripheral locations in the network. The commonly used word ‘weak’ in this context sounds particularly paradoxical. It is, indeed, usually used as a synonym for a low-degree node with an additional bridging property, which has resisted a quantitative formulation. We provide this definition through equation (5), according to which weak nodes are, de facto, quite strong. Paraphrasing Granovetter’s conundrum³⁰, equation (5) quantifies the “strength of weak nodes”.

The CI-algorithm scales as $\sim O(N \log N)$ by removing a finite fraction of nodes at each step (Supplementary Information section VB). This high scalability allows us to find top influencers in current big-data social media and the minimal set of people to immunize in large-scale populations at the country level. The applications are investigated next.

Figure 2a shows the optimal threshold q_c for a random Erdős-Rényi (ER) network⁵ (marked by the vertical line) obtained by extrapolating the EO solution to $N \rightarrow \infty$ and $\ell \rightarrow \infty$ (Supplementary Information section IV). In the same figure we compare the optimal threshold against the heuristic centrality measures: high-degree (HD)⁹, high-degree

adaptive (HDA), PageRank (PR)⁷, closeness centrality (CC)⁶, eigenvector centrality (EC)⁶, and k-core¹² (see Supplementary Information section I for definitions). Supplementary Information sections VI and VII show the comparison with the remaining heuristics^{6,11} and the Belief Propagation method of ref. 14, respectively, which have worse computational complexity (and optimality), and cannot be applied to the network sizes used here. Remarkably, at the optimal value q_c predicted by our theory, the best among the heuristic methods (HDA, PR and HD) still predict a giant component ~ 50 – 60% of the whole original network. Furthermore, the influencer threshold predicted by CI approximates very well the optimal one, and, notably, CI outperforms the other strategies. Figure 2b compares CI in scale-free (SF) networks⁵ against the best heuristic methods, that is, HDA and HD. In all cases, CI produces a smaller threshold and a smaller giant component (Fig. 2c).

As an example of an information spreading network, we consider the web of Twitter users (Supplementary Information section VIII¹⁹). Figure 3a shows the giant component of Twitter when a fraction q of its influencers is removed following CI. It is surprising that a lot of Twitter users with a large number of contacts have a mild influence on the network. This is witnessed by the fact that, when CI (at $\ell = 5$) predicts a zero giant component (and so it exhausts the number of optimal influencers), the scalable heuristic ranks (HD, HDA, PR and k-core) still give a substantial giant component of the order of 30–70% of the entire network. These heuristics also, inevitably, find a remarkably large number of (fake) influencers, which is at least 50% larger than that predicted by CI (Fig. 3b and Supplementary Information section VIII). One cause for the poor performance of the high-degree-based ranks is that most of the hubs are clustered, which gives a mediocre importance to their contacts. As a consequence, hubs are outranked by nodes with lower degree surrounded by coronas of hubs (shown in detail in Fig. 3c), that is, the weak nodes predicted by the theory (Fig. 1e).

Finally, we simulate an immunization scheme on a personal contact network built from the phone calls performed by 14 million people in

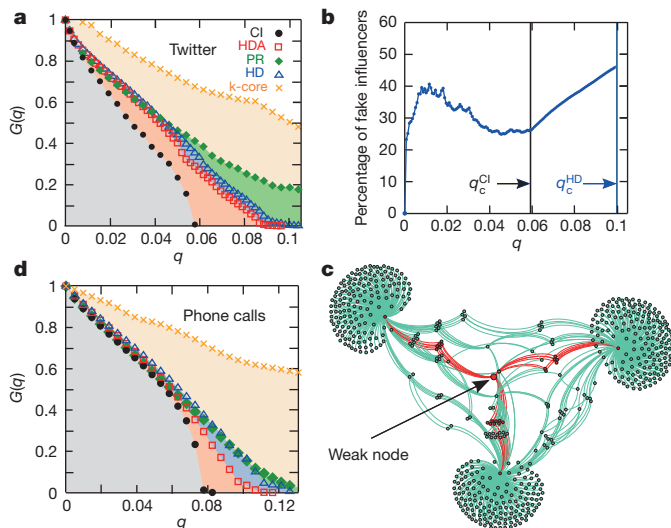


Figure 3 | Performance of CI in large-scale real social networks. **a**, Giant component $G(q)$ of Twitter users¹⁹ ($N = 469,013$) computed using CI, HDA, PR, HD and k-core strategies (other heuristics have prohibitive running times for this system size). **b**, Percentage of fake influencers or false positives (PFI, equation (120) in Supplementary Information) in Twitter as a function of q , defined as the percentage of non-optimal influencers identified by the HD algorithm in comparison with CI. Below q_c^{CI} , PFI reaches as much as $\sim 40\%$, indicating the failure of HD in optimally finding the top influencers. Indeed, to obtain $G = 0$, HD has to remove a much larger number of fake influencers, which at q_c^{HD} reaches PFI $\approx 48\%$. **c**, An example of the many weak nodes found in Twitter. These crucial influencers were missed by all heuristic strategies. **d**, $G(q)$ for a social network of 1.4×10^7 mobile phone users in Mexico representing an example of big data to test the scalability and performance of the algorithm in real networks. CI immunizes this social network using half a million fewer people than the best heuristic strategy (HDA), saving $\sim 35\%$ of the vaccine stockpile.

Mexico (Supplementary Information section IX). Figure 3d shows that our method saves a large number of vaccines or, equivalently, finds the smallest possible set of people to quarantine; our method therefore also outranks the scalable heuristics in large real networks. Thus, while the mapping of the influencer identification problem onto optimal percolation is strictly valid for locally tree-like random networks, our results may apply also to real loopy networks, provided the density of loops is not excessively large.

Our solution to the optimal influence problem shows its importance in that it helps to unveil hitherto hidden relations between people, as witnessed by the weak-node effect. This, in turn, is the by-product of a broader notion of influence, lifted from the individual non-interacting point of view^{6–12,19,20} to the collective sphere: influence is an emergent property of collectivity, and top influencers arise from the optimization of the complex interactions they stipulate.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 February; accepted 20 May 2015.

Published online 1 July 2015.

- Domingos, P. & Richardson, M. Mining knowledge-sharing sites for viral marketing. In *Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 61–70 (ACM, 2002); <http://dx.doi.org/10.1145/775047.775057>.

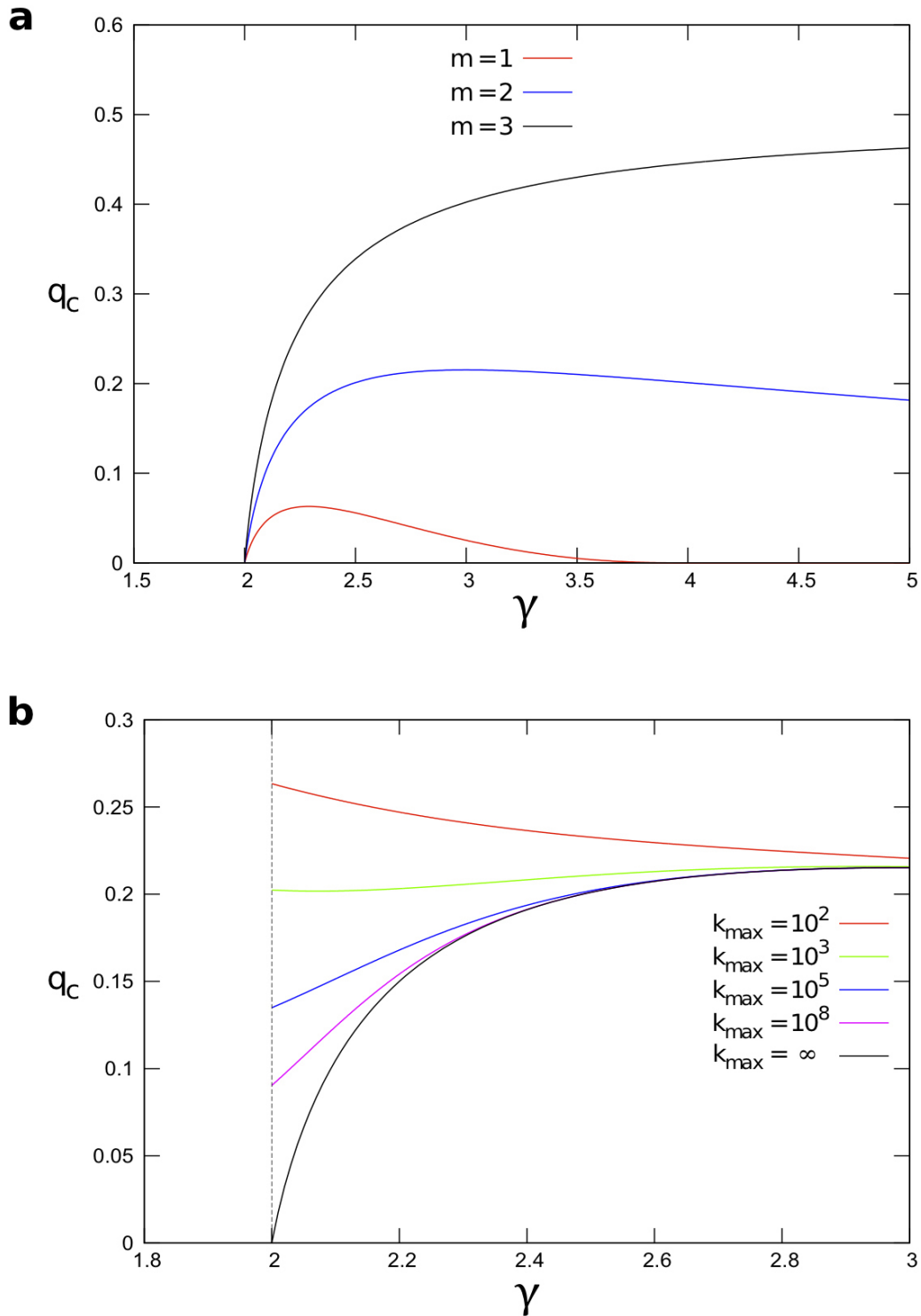
- Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001).
- Newman, M. E. J. Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128 (2002).
- Kempe, D., Kleinberg, J. & Tardos, E. Maximizing the spread of influence through a social network. In *Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 137–143 (ACM, 2003); <http://dx.doi.org/10.1145/956750.956769>.
- Newman, M. E. J. *Networks: An Introduction* (Oxford Univ. Press, 2010).
- Freeman, L. C. Centrality in social networks: conceptual clarification. *Soc. Networks* **1**, 215–239 (1978).
- Brin, S. & Page, L. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks ISDN Systems* **30**, 107–117 (1998).
- Kleinberg, J. Authoritative sources in a hyperlinked environment. In *Proc. 9th ACM-SIAM Symp. on Discrete Algorithms* (1998); *J. Assoc. Comput. Machinery* **46**, 604–632 (1999).
- Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
- Cohen, R., Erez, K., ben-Avraham, D. & Havlin, S. Breakdown of the Internet under intentional attack. *Phys. Rev. Lett.* **86**, 3682–3685 (2001).
- Chen, Y., Paul, G., Havlin, S., Liljeros, F. & Stanley, H. E. Finding a better immunization strategy. *Phys. Rev. Lett.* **101**, 058701 (2008).
- Kitsak, M. et al. Identification of influential spreaders in complex networks. *Nature Phys.* **6**, 888–893 (2010).
- Altarelli, F., Braunstein, A., Dall'Asta, L. & Zecchina, R. Optimizing spread dynamics on graphs by message passing. *J. Stat. Mech.* P09011 (2013).
- Altarelli, F., Braunstein, A., Dall'Asta, L., Wakeling, J. R. & Zecchina, R. Containing epidemic outbreaks by message-passing techniques. *Phys. Rev. X* **4**, 021024 (2014).
- Hashimoto, K. Zeta functions of finite graphs and representations of p-adic groups. *Adv. Stud. Pure Math.* **15**, 211–280 (1989).
- Coja-Oghlan, A., Mossel, E. & Vilenchik, D. A spectral approach to analyzing belief propagation for 3-coloring. *Combin. Probab. Comput.* **18**, 881–912 (2009).
- Granovetter, M. Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443 (1978).
- Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99**, 5766–5771 (2002).
- Pei, S., Muchnik, L., Andrade, J. S. Jr, Zheng, Z. & Makse, H. A. Searching for superspreaders of information in real-world social media. *Sci. Rep.* **4**, 5547 (2014).
- Pei, S. & Makse, H. A. Spreading dynamics in complex networks. *J. Stat. Mech.* P12002 (2013).
- Bollobás, B. & Riordan, O. *Percolation* (Cambridge Univ. Press, 2006).
- Bianconi, G. & Dorogovtsev, S. N. Multiple percolation transitions in a configuration model of network of networks. *Phys. Rev. E* **89**, 062814 (2014).
- Karrer, B., Newman, M. E. J. & Zdeborová, L. Percolation on sparse networks. *Phys. Rev. Lett.* **113**, 208702 (2014).
- Angel, O., Friedman, J. & Hoory, S. The non-backtracking spectrum of the universal cover of a graph. *Trans. Am. Math. Soc.* **367**, 4287–4318 (2015).
- Krzakala, F. et al. Spectral redemption in clustering sparse networks. *Proc. Natl Acad. Sci. USA* **110**, 20935–20940 (2013).
- Newman, M. E. J. Spectral methods for community detection and graph partitioning. *Phys. Rev. E* **88**, 042822 (2013).
- Radicchi, F. Predicting percolation thresholds in networks. *Phys. Rev. E* **91**, 010801(R) (2015).
- Mézard, M. & Parisi, G. The cavity method at zero temperature. *J. Stat. Phys.* **111**, 1–34 (2003).
- Boettcher, S. & Percus, A. G. Optimization with extremal dynamics. *Phys. Rev. Lett.* **86**, 5211–5214 (2001).
- Granovetter, M. The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380 (1973).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by NIH-NIGMS 1R21GM107641 and NSF-PoLS PHY-1305476. Additional support was provided by Army Research Laboratory Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). We thank L. Bo, S. Havlin and R. Mari for discussions and Grandata for providing the data on mobile phone calls.

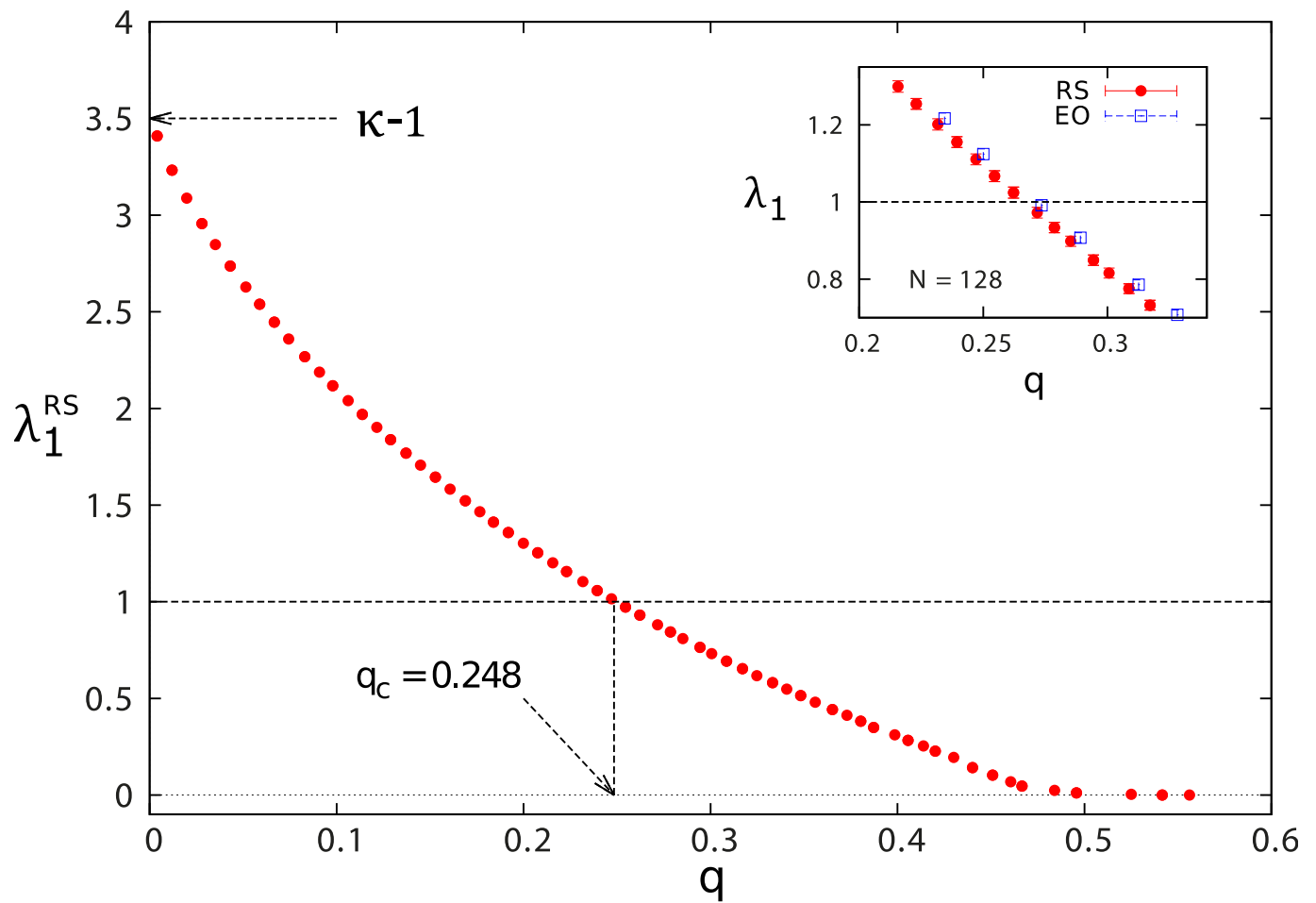
Author Contributions Both authors contributed equally to the work presented in this paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.A.M. (hmake@lev.cuny.cuny.edu).



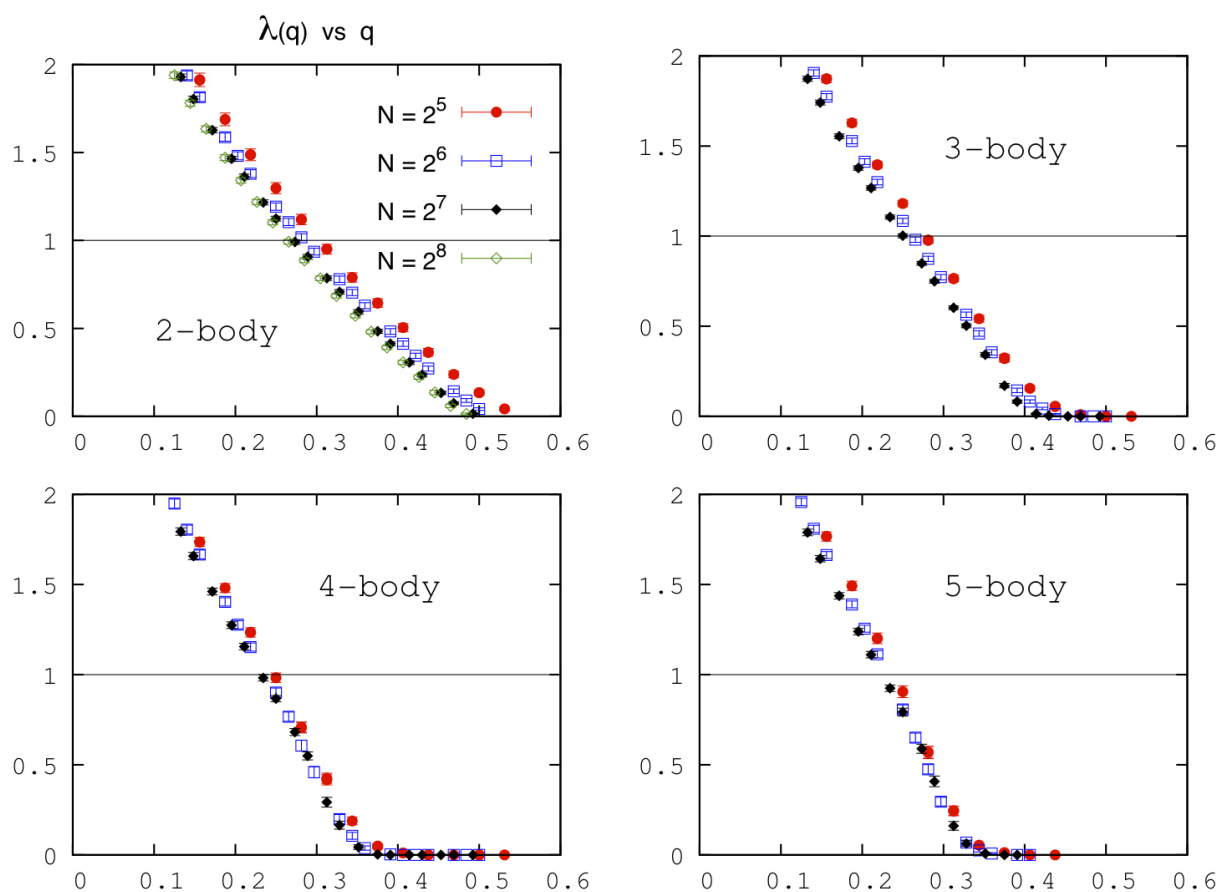
Extended Data Figure 1 | High-degree (HD) threshold. **a**, HD influence threshold q_c as a function of the degree distribution exponent γ of scale-free networks in the ensemble with $k_{\max} = mN^{1/(\gamma-1)}$ and $N \rightarrow \infty$. The curves refer to different values of the minimum degree m : 1 (red), 2 (blue), 3 (black). The fragility of SF networks (small q_c) is notable for $m=1$ (the case calculated in ref. 10). In this case ($m=1$), the network contains many leaves, and reduces to a star at $\gamma=2$, which is trivially destroyed by removing the only single hub, explaining the general fragility in this case. Furthermore, in this same case, the network becomes a collection of dimers with $k=1$ when $\gamma \rightarrow \infty$, which is still trivially fragile. This also explains why $q_c \rightarrow 0$ for $\gamma \geq 4$. Therefore, the fragility in the case $m=1$ has its roots in these two limiting trivial cases. Removing the leaves ($m=2$) results in a 2-core, which is already more robust.

For the 3-core $m=3$, $q_c \approx 0.4-0.5$ provides a quite robust network, and has the expected asymptotic limit to a non-zero q_c of a random regular graph with $k=3$ as $\gamma \rightarrow \infty$, $q_c \rightarrow (k-2)/(k-1) = 0.5$. Thus, SF networks become robust in these more realistic cases, and the search for other attack strategies becomes even more important. **b**, HD influence threshold q_c as a function of the degree distribution exponent of scale-free networks with minimum degree $m=2$ in the ensemble where k_{\max} is fixed and does not scale with N . The curves refer to different values of the cut-off k_{\max} : 10^2 (red), 10^3 (green), 10^5 (blue), 10^8 (magenta), and $k_{\max} = \infty$ (black), and show that for a typical k_{\max} degree of 10^3 , for instance in social networks, the network is fairly robust with $q_c \approx 0.2$ for all γ . The curve with $m=2$ and $k_{\max} = 10^3$ is replotted in the inset of Fig. 2b.



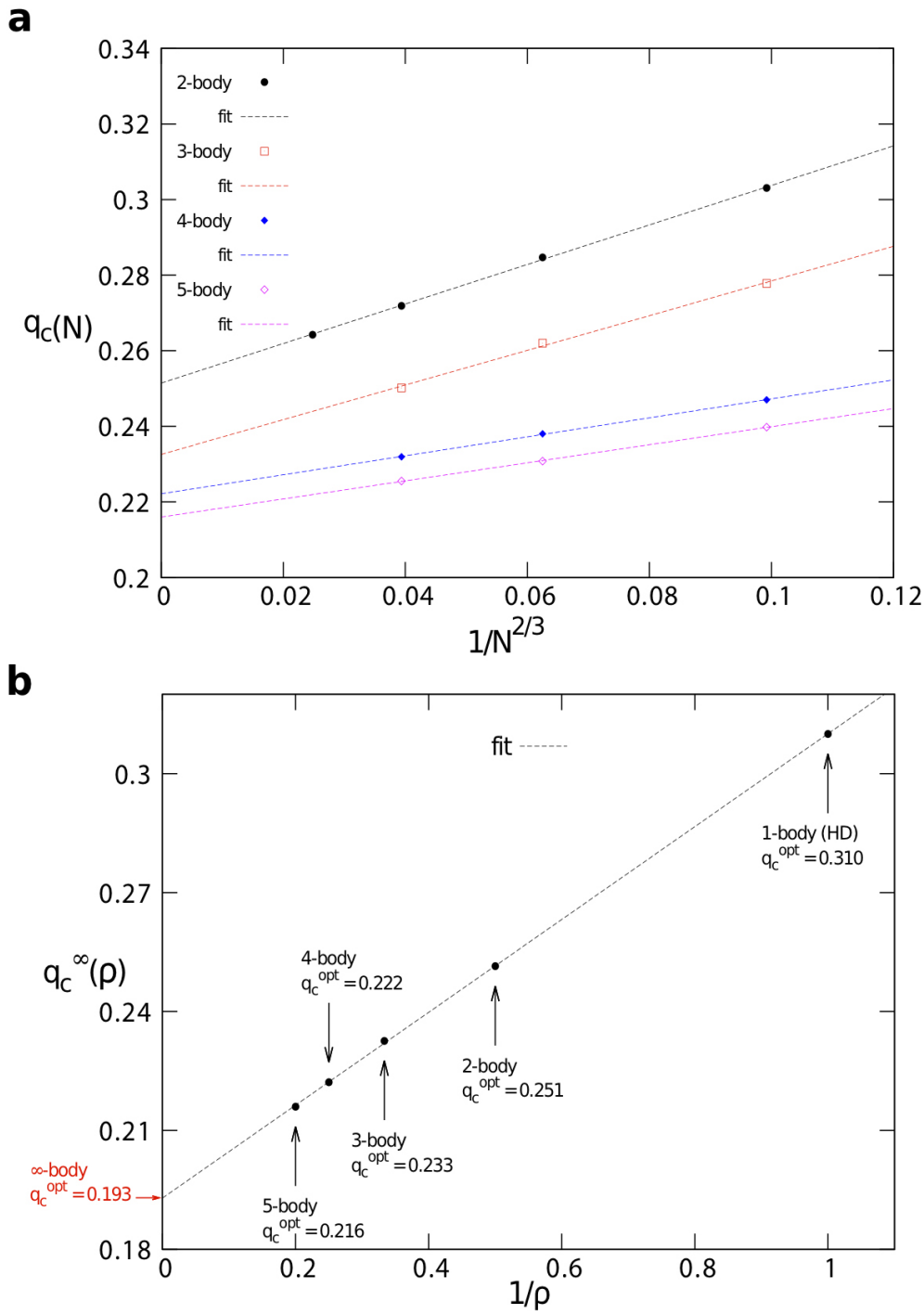
Extended Data Figure 2 | Replica Symmetry (RS) estimation of the maximum eigenvalue. Main panel, the eigenvalue $\lambda_1^{\text{RS}}(q)$, equation (92) in Supplementary Information for the two-body interaction $\ell = 1$, obtained by minimizing the energy function $\mathcal{E}(s)$ with the RS cavity method. The curve was computed on an ER graph of $N = 10,000$ nodes and average degree $\langle k \rangle = 3.5$

and then averaged over 40 realizations of the network (error bars are s.e.m.). Inset, comparison between the RS cavity method and EO (extremal optimization) for an ER graph of $\langle k \rangle = 3.5$ and $N = 128$. The curves are averaged over 200 realizations (error bars are s.e.m.).



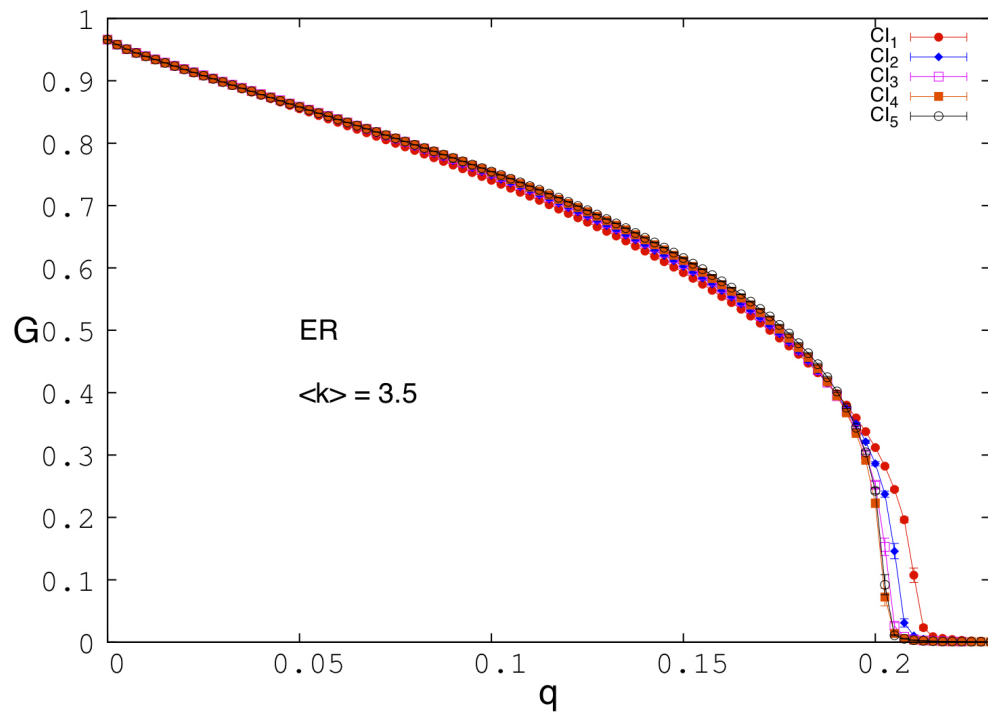
Extended Data Figure 3 | EO estimation of the maximum eigenvalue. Eigenvalue $\lambda(q)$ obtained by minimizing the energy function $\mathcal{E}(\mathbf{n})$ with τ EO (τ -extremal optimization), plotted as a function of the fraction of removed nodes q . The panels are for different orders of the interactions. The curves in

each panel refer to different sizes of ER networks with average connectivity $\langle k \rangle = 3.5$. Each curve is an average over 200 instances (error bars are s.e.m.). The value q_c where $\lambda(q_c) = 1$ is the threshold for a particular N and many-body interaction.



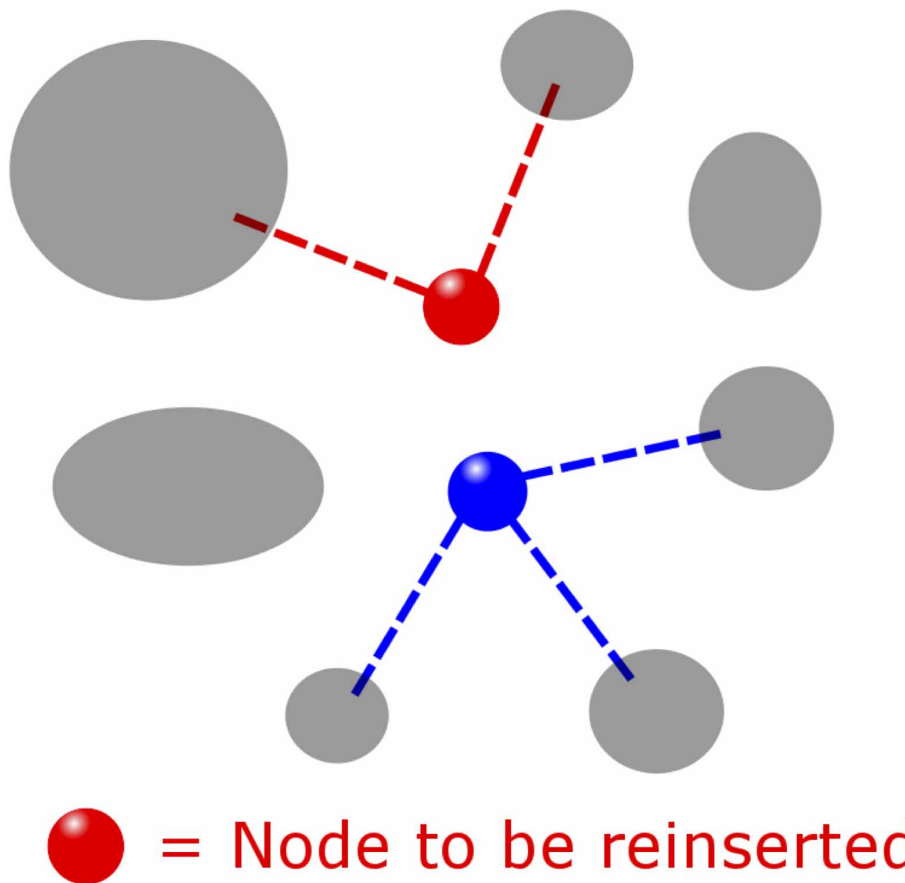
Extended Data Figure 4 | Estimation of optimal threshold q_c^{opt} with EO.
a, Critical threshold q_c as a function of the system size N , obtained with EO from Extended Data Fig. 3, of ER networks with $\langle k \rangle = 3.5$ and varying size. The curves refer to different orders of the many-body interactions. The data show a linear behaviour as a function of $N^{-2/3}$, typical of spin glasses, for each many-

body interaction ρ . The extrapolated value $q_c^{\infty}(\rho)$ is obtained at the y intercept.
b, Thermodynamic critical threshold $q_c^{\infty}(\rho)$ as a function of the order of the interactions ρ from **a**. The data scale linearly with $1/\rho$. From the y intercept of the linear fit we obtain the thermodynamic limit of the infinite-body optimal value $q_c^{\text{opt}} = q_c^{\infty}(\rho \rightarrow \infty) = 0.192(9)$.



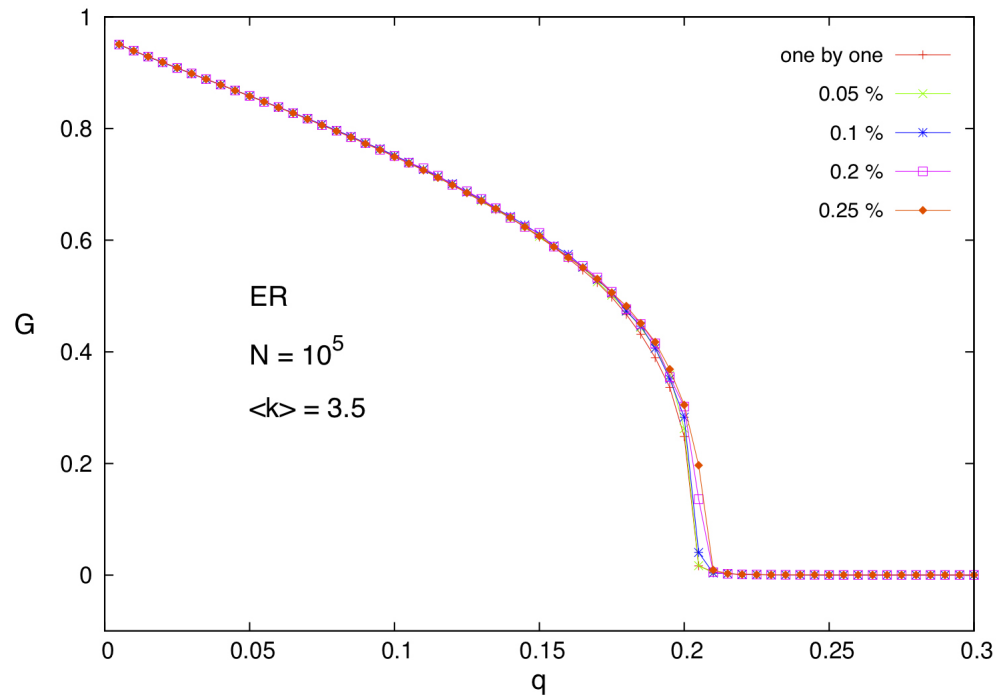
Extended Data Figure 5 | Comparison of the CI algorithm for different radii ℓ of the Ball(ℓ). We use $\ell = 1, 2, 3, 4, 5$, on a ER graph with average degree $\langle k \rangle = 3.5$ and $N = 10^5$ (the average is taken over 20 realizations of the network, error bars are s.e.m.). For $\ell = 3$ the performance is already practically indistinguishable from $\ell = 4, 5$. The stability analysis we developed to minimize q_c is strictly valid only when $G = 0$, since the largest eigenvalue of the

modified NB matrix controls the stability of the solution $G = 0$, and not the stability of the solution $G > 0$. In the region where $G > 0$ we use a simple and fast procedure to minimize G explained in Supplementary Information section VA. This explains why there is a small dependence on having a slightly larger G for larger ℓ , when $G > 0$ in the region $q \approx 0.15$.

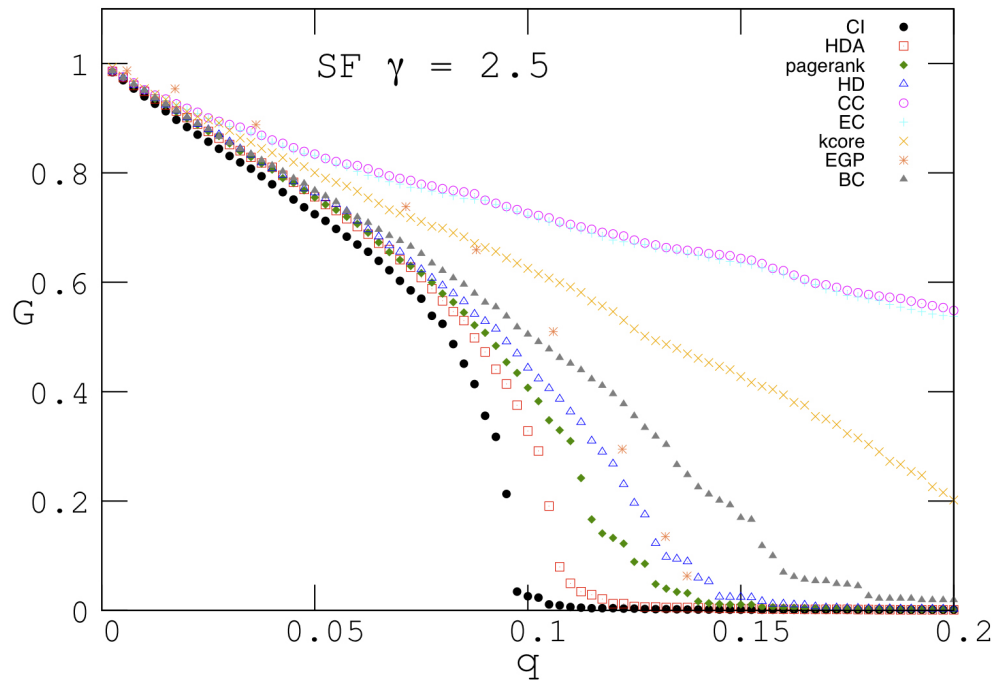


Extended Data Figure 6 | Illustration of the algorithm used to minimize $G(q)$ for $q < q_c$. Starting from the completely fragmented network at $q = q_c$, the Nq_c influencers are reinserted with their original degree and connected to their original neighbours with the following criterion: each node is assigned and index $c(i)$ given by the number of clusters it would join if it were reinserted

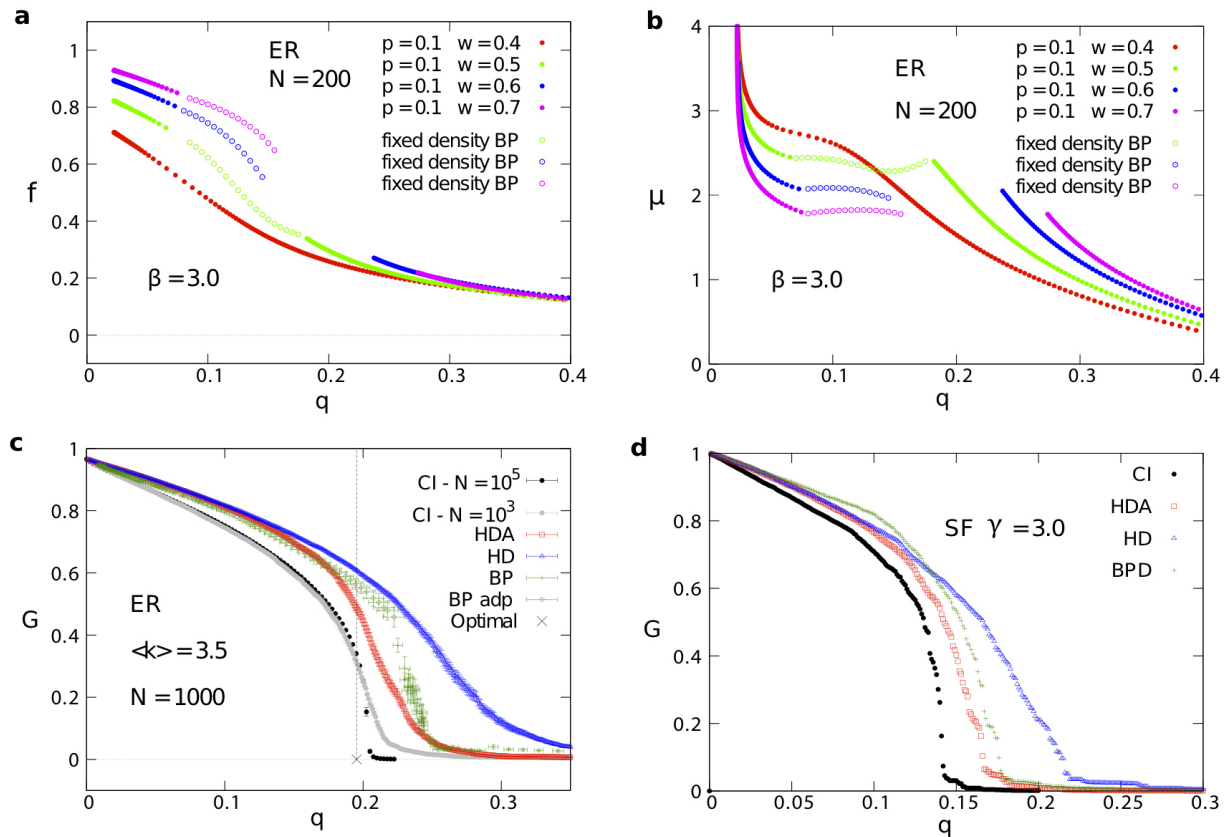
in the network. For example, the red node has $c(\text{red}) = 2$, while the blue one has $c(\text{blue}) = 3$. The node with the smallest $c(i)$ is reinserted in the network: in this case the red node. Then the $c(i)$ s are recalculated and the new node with the smallest $c(i)$ is found and reinserted. These steps are repeated until all the removed nodes are reinserted in the network.



Extended Data Figure 7 | Test of the decimation fraction. Giant component G as a function of the fraction of removed nodes q using CI, for an ER network of $N = 10^5$ nodes and average degree $\langle k \rangle = 3.5$. The profiles of the curves are drawn for different percentages of nodes fixed at each step of the decimation algorithm.



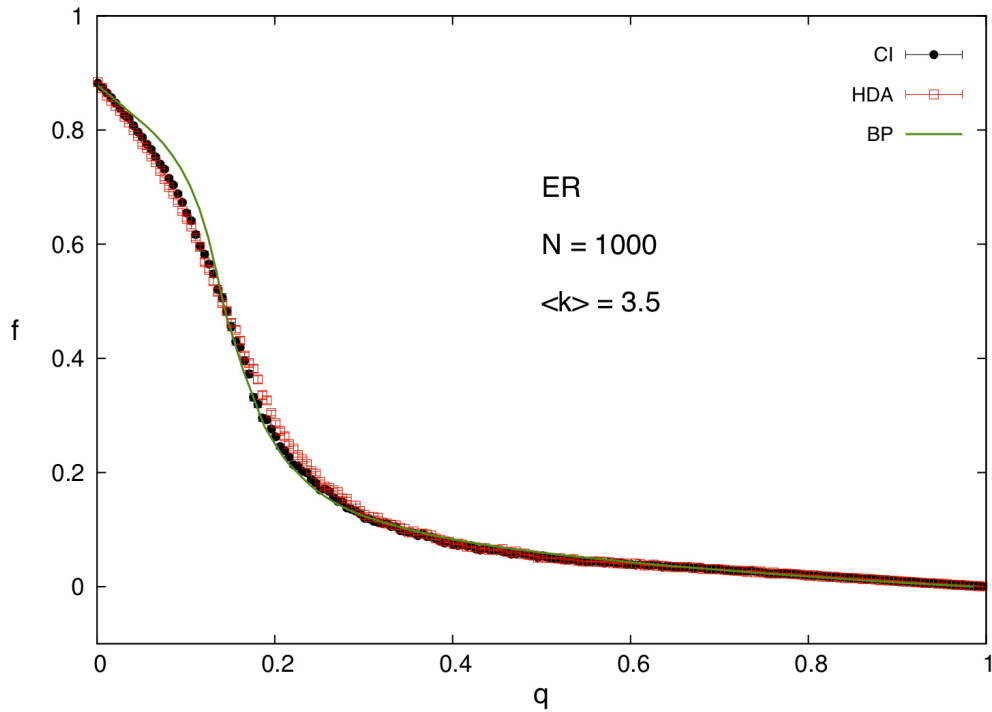
Extended Data Figure 8 | Comparison of the performance of CI, BC and EGP in destroying G . We also include HD, HDA, EC, CC, k-core and PR. We use a scale-free (SF) network with degree exponent $\gamma = 2.5$, average degree $\langle k \rangle = 4.68$, and $N = 10^4$. We use the same parameters as in ref. 11.



Extended Data Figure 9 | Comparison with BP for a network

immunization. **a**, Fraction of infected nodes f as a function of the fraction of immunized nodes q in the susceptible-infected-removed (SIR) model from the BP solution. We use an ER random graph of $N = 200$ nodes and average degree $\langle k \rangle = 3.5$. The fraction of initially infected nodes is $p = 0.1$ and the inverse temperature $\beta = 3.0$. The profiles are drawn for different values of the transmission probability w : 0.4 (red curve), 0.5 (green), 0.6 (blue), 0.7 (magenta). Also shown are the results of the fixed density BP algorithm (open circles). **b**, Chemical potential μ as a function of the immunized nodes q from BP. We use an ER random graph of $N = 200$ nodes and average degree $\langle k \rangle = 3.5$. The fraction of the initially infected nodes is $p = 0.1$ and the

inverse temperature $\beta = 3.0$. The profiles are drawn for different values of the transmission probability w : 0.4 (red curve), 0.5 (green), 0.6 (blue), 0.7 (magenta). Also shown are the results of the fixed density BP algorithm (open circles) for the region where the chemical potential is non-convex. **c**, Comparison between the giant components obtained with CI, HDA, HD and BP. We use an ER network of $N = 10^3$ and $\langle k \rangle = 3.5$. We also show the solution of CI from Fig. 2a for $N = 10^5$. We find in order of performance: CI, HDA, BP and HD. (The average is taken over 20 realizations of the network, error bars are s.e.m.) **d**, Comparison between the giant components obtained with CI, HDA, HD and BPD. We use a SF network with degree exponent $\gamma = 3.0$, minimum degree $k_{\min} = 2$, and $N = 10^4$ nodes.



Extended Data Figure 10 | Fraction of infected nodes $f(q)$ as a function of the fraction of immunized nodes q in SIR from BP. We use the following parameters: initial fraction of infected people $p = 0.1$, and transmission probability $w = 0.5$. We use an ER network of $N = 10^3$ nodes and $\langle k \rangle = 3.5$.

We compare CI, HDA and BP. All strategies give similar performance, owing to the large value of the initial infection p , which washes out the optimization performed by any sensible strategy, in agreement with the results shown in figure 12a of ref. 14.

CORRECTIONS & AMENDMENTS

CORRIGENDUM

doi:10.1038/nature15737

Corrigendum: Influence maximization in complex networks through optimal percolation

Flaviano Morone & Hernán A. Makse

Nature **524**, 65–68 (2015); doi:10.1038/nature14604

In the Acknowledgements section of this Letter, 'ARL' should read 'Army Research Laboratory Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA)'. This has been corrected in the online versions of the paper.