# Network Hubs Cease to be Influential in the Presence of Low Levels of Advertising

*Gabriel Rossman**
*Sociology, UCLA*
*rossman@soc.ucla.edu*
*264 Haines Hall LA, CA 90095*
*https://orcid.org/0000-0002-3556-0601*

*Jacob C. Fisher*
*Institute for Social Research, University of Michigan and*
*Social Science Research Institute, Duke University*
*https://orcid.org/0000-0003-0299-0346*

# Abstract

Attempts to find central “influencers,” “opinion leaders,” “hubs,” “optimal seeds,” or other important people who can hasten or slow diffusion or social contagion has long been a major research question in network science. We demonstrate that opinion leadership occurs only under conventional but implausible scope conditions. We demonstrate that a highly central node is a more effective seed for diffusion than a random node if nodes can only learn via the network. However, actors are also subject to external influences such as mass media and advertising. We find that diffusion is noticeably faster when it begins with a high centrality node, but that this advantage only occurs in the region of parameter space where external influence is constrained to zero and collapses catastrophically even at minimal levels of external influence. Importantly, nearly all prior agent-based research on choosing a seed or seeds implicitly occurs in the network influence only region of parameter space. We demonstrate this effect using preferential attachment, small world, and several empirical networks. These networks vary in how large the baseline opinion leadership effect is, but in all of them it collapses with the introduction of external influence. This implies that in marketing and public health that advertising broadly may be underrated as a strategy for promoting network-based diffusion.

# Acknowledgements

Simulations for this project were conducted using the Duke Compute Cluster and the University of Michigan - Institute for Social Research's Likert cluster. The authors thank Tom Milledge, Mark DeLong, and Nick Hinkle-Degroot for their assistance in using those clusters. Research for this paper was funded by the NSF (1535370, 1760609) and NIH (R25HD079352, R01HD075712, UL1TR002240). The authors are grateful for advice and feedback from two anonymous reviewers, Jacob Foster, Kieran Healy, Brayden King, Omar Lizardo, Aliza Luft, Pam Oliver, Matt Salganik, and participants at Sunbelt, UCI’s COR/Merage Colloquium, UCLA’s Markets, Organizations, and Movements Working Group, and the Princeton University miniconference “Celebrating Paul DiMaggio.”

Among the central theoretical and practical attractions of social network analysis is the promise that key nodes, known as "opinion leaders" or "influentials," hold structural power to change the ideas and behaviors of entire social systems(1–3). An extensive literature in sociology, physics, and network science centers on how best to measure network centrality. From the beginning, much of this literature takes as its motivation identifying a node or nodes that are optimal seeds for diffusion(4–8).[*] For instance, a seminal study of how doctors prescribe new drugs ascribed this behavior to key doctors in the advice network(11). In such applied contexts as "viral" marketing and public health outreach, opinion leadership suggests the promise that a structurally important node (and, by extension, the social network analyst who can identify that node) is the key to controlling the spread of a product, health behavior, or other idea or behavior(2, 3, 8, 12–14).

The influentials literature focuses on network sources of information, but in most realistic scenarios people have sources of information that transcend the network(15–17). Introducing these non-network sources of information may qualitatively change the nature of diffusion, and specifically the role of a highly central hub or hubs. In many theories and simulations, agents are constrained to only observe information through a social graph, but real people are not so myopic. Even if we are most attentive to word-of-mouth from our social ties, we also learn about new ideas and behaviors from mass media, advertising, government mandates, and even direct observation of events. If it begins raining and everyone opens her umbrella, the proximate cause of this behavior is a response to nature rather than information spreading through a social network(18). Some diffusion models meaningfully incorporate roles for external sources of information(15, 19, 20), but other models effectively assume an entirely word-of-mouth process even if their narrative theory allows for external influence(3).

The computational experiment we present in this article contributes to a large body of social networks literature on influentials and opinion leadership(7, 8), but takes as its micro-foundations a diffusion model from marketing that involves both network-based diffusion and external influence from sources like advertising(15). We conduct a large-scale computer simulation in which we seed diffusion with either the most central node or a node chosen at random in various empirical and algorithmically-generated networks.[†] We test the opinion leadership hypothesis for various points in parameter space where one axis is the strength of network-based diffusion (e.g., "word of mouth") and the other axis is the strength of an external force (e.g., advertising and mass media). We measure the strength of opinion leadership for each point in parameter space by how much faster diffusion occurs when the initial node is highly central versus chosen at random.

[*] Aside from measuring influence over diffusion, the other two major theoretical interpretations of network centrality are status and bargaining power(9, 10). These theoretical applications have their own associated centrality metrics.

[†] Replication code is available at https://osf.io/25rav/?view_only=509223b515e64c99a4b277d0bdd376fb

The experiment adapts a mixed-influence model outlined by Bass(15, 21–23) to test whether the effect of central nodes on diffusion is robust to the presence of external influence. In the Bass model, people are exposed to information about the innovation from two sources: interpersonal imitation (with a density dependent hazard) and external influence (with a constant hazard). Interpersonal influence represents the effect of word-of-mouth (or closely analogous processes like local network externalities or person-to-person spread)(24, 25). External influence represents the effect of advertising, mass media, internet search, or government mandates(15, 17, 23, 26). Traditionally, the Bass model is represented as a differential equation that measures diffusion in aggregate over time. The aggregate approach has the advantage of simplicity but makes it impossible to integrate network structure. We therefore adapt the Bass model to an agent-based model which allows for potential emergent properties of unequal influence between nodes based on their structural positions.

The Bass model defines the rate of new adoptions in aggregate as

$$\Delta N_t = (a + bN_t) \, * \, (N_{max} - N_t)$$

where $N_t$ is the cumulative number of people who have adopted as of time *t*, *a* is the coefficient of external influence, *b* is the coefficient of interpersonal influence, and $N_{max}$ is the asymptotic number of people who will ever adopt. To include the effect of network structure on individual adoption, we adapt this equation to an agent-based model. In the agent-based model, for each agent *i* at time *t*:

$$p(i\ adopts\ at\ time\ t \mid i\ has\ not\ adopted\ before\ t) \\ = \alpha + \beta(fraction\ of\ i's\ neighbors\ who\ adopted\ before\ time\ t)$$

α is a constant hazard of adoption, representing the weight given to advertising and other external influences on diffusion, and β is the weight given to social or network influence.[‡] To ensure that α and β are on comparable scales, we allow them to range between 0 and a maximum value that saturates the network with a consistent probability. We identify these maxima with a separate set of simulations, which identify the values at which α and β saturate the network in one hundred ticks or less in 50% of trials. We refer to these α and β maxima as "LD50," as a metaphor for the standard "lethal dose 50%" metric in toxicology. Full details of

[‡] More formally stated,

$$P(X_i = t | X_i \geq t) = \alpha + \beta\left(\frac{\sum_{j \in N(i)} X_j < t}{|N(i)|}\right)$$

where $X_i$ is the time when person *i* adopts, *N(i)* is the set of neighbors of person *i*, α is a constant hazard of adoption, representing the weight given to advertising and other external influences on diffusion, and β is the weight given to social or network influence.

estimating the LD50 values appear in the supplementary materials.[§] To highlight changes at the lowest end of parameter space, we explore both dimensions of the parameter space on a log scale. In both the aggregate and agent-based Bass models, once a person adopts, she cannot abandon the innovation, meaning the number of adopters increases monotonically.

Our experimental setup varies the seed, meaning the initial innovator in the simulation. In the simulation, innovations start at one person, the seed, and spread outward from that person.[**] Our control condition seeds the innovation with a randomly chosen person in the network. Our treatment condition seeds the innovation with the most central person in the network, as measured by betweenness. In most networks, betweenness is right-skewed so in our networks the most central node is anywhere from six to several hundred standard deviations above the mean.[††] We test the effects on preferential attachment networks (shown in Figure 1) and small world networks generated in igraph(27) as well as the giant components of the Democratic National Committee email network (548 nodes and 2,442 edges), Enron e-mail network (33,696 nodes and 180,811 edges), and a network of retweets and mentions on Twitter (532,325 nodes and 694,606 edges). We focus on preferential attachment networks in Figures 1 and 2 but show robustness of our key finding to all these networks in Figure 3 and the supplementary materials.

**FIGURE 1 ABOUT HERE**

Figure 2 shows the central tendencies of the cumulative distribution functions by random versus highest betweenness seed node given the assumption of peak social influence ($\alpha = 0$, $\beta = LD50$).[‡‡] Under those conditions, innovations that

[§] Parameter values are set to the same scale so that a one-unit change in $\alpha$ does not have a substantially different meaning than a one-unit change in $\beta$. We bracket the question of what position in parameter space is most realistic for what applications. However, neither the exact choice of maxima nor the exact position in parameter space matter for our findings, as the majority of the effect occurs with the introduction of *any* external influence.

[**] See the supplementary materials for specifications with multiple seeds targeted by keyplayer(5). More seeds results in faster diffusion, but the relative advantage of targeting multiple seeds with keyplayer versus an equal number of random seeds is an order of magnitude weaker than that of a single targeted seed versus a single random seed. However, the advantage of targeting multiple seeds is less fragile to the introduction of external influence.

[††] Centrality metrics tend to be correlated in the right-tail, implying that the analysis should be robust to the choice of metric. As an example, appendix Fig. S7 replicates our findings using closeness centrality instead of betweenness.

[‡‡] See Figure S1 in supplementary materials for spaghetti plots of individual CDFs.

start with the most central person spread to half of the people in the network over twice as fast.

## FIGURE 2 ABOUT HERE

As Figure 2 indicates, the gap between the conditions is approximately widest at time to 50% adoption (CDF = 500, displayed as a red horizontal line), making it the metric most favorable to opinion leadership. In addition, time to 50% adoption is much less vulnerable to right-censorship than time to saturation. We use this metric, average time to 50% adoption, to summarize the full parameter space. In the top panel of Figure 3, we demonstrate how diffusion speed on a preferential attachment network responds to varying the α and β parameters separately for random seeds and seeding at the highest betweenness node. The heat dimension shows the ratio of the mean time to 50% saturation for a random seed over that for a high centrality seed. (Supplementary materials figure S1 shows how this ratio is derived from Figure 2). Seeding with a highly central node has an advantage but only when α = 0. This advantage disappears quickly for all points in parameter space where α > 0, dropping precipitously at the next interval (α = 0.25% of LD50) and the advantage of a highly central seed node almost completely vanishes for points in parameter space where α > 3% of LD50.

## FIGURE 3 ABOUT HERE

The heat map in the top panel only illustrates results for preferential attachment networks, but in the bottom panel we provide sparklines summarizing several networks for the plane of parameter space where $\beta = LD50$ (i.e., the equivalent of the top row of the heat map). When α = 0, the effect of a highly central seed node varies substantially by the type of network, being trivial in a small world, but substantial in the three empirical networks. However, the finding from preferential attachment networks that the advantage of seeding with the peak betweenness node collapses rapidly when α > 0 replicates in all other networks, no matter how strong the highly central seed node effect is when $\alpha = 0$. Targeting the central node materially speeds adoption only in the region of parameter space where there is no external influence (α = 0). In all networks, there is a precipitous drop in the effect of highly central seeding as α goes from zero to 0.26% of the LD50 and the highly central seed effect is essentially absent when α reaches even a few percentage points of its LD50 value. The supplementary materials contain full heat maps for all networks listed in the Figure 3 sparklines plot.

These findings indicate that the positive effect of targeting the most central node as opinion leader is subject to a highly restrictive scope condition. Previous research has shown that opinion leadership requires substantial inequality in centrality(28) but many phenomena of interest meet that scope condition. Here we show the much more demanding scope condition of the absence of advertising or other forms of external influence. When no external influences are present, targeting a highly central person results in diffusion that can spread to half of the network faster than if a person were chosen at random, with the advantage being trivial for small world networks and an order of magnitude for the email networks. However, in the presence of external influences, even extremely weak external influences, identifying and seeding with an opinion leader does not lead to appreciably faster adoption of an innovation. This suggests that the simulation literature on optimal seeding to opinion leaders only applies under restrictive scope conditions that likely apply to few empirical scenarios. When diffusion follows the network strictly, as in the spread of a sexually transmitted disease(29) or clandestine communication with a cell structure, then centrality can have appreciable effects. However, the diffusion of a product, behavior, or belief, will normally involve some level of external influence and even if that external influence is dwarfed by network influence, there should be no effect of the seed node's network position so long as external influence exists at all.

Adding in even weak advertising effects nullifies the impact of seeding with the most central node. Advertising creates a non-zero probability that people can adopt without exposure from other adopters, conceptually similar to increasing the number of seeds. Our findings thus suggest that advertisers or public health officials who are planning a campaign should consider that advertising can also promote network-based spread and may do so more efficiently than identifying and recruiting a highly central seed node. This implies a return to the early "two step flow" model, in which most people adopt based on influence from numerous minor opinion leaders of purely local influence, who in turn got information from mass media(19, 20).

There is substantial evidence that ideas and behaviors spread via interpersonal influence, but this is neither the same thing as an emergent property of critical importance for a highly central node nor a practical upshot that seeding with a central node is important under realistic circumstances. While social connections remain important for the spread of ideas, products, and behaviors, our simulations highlight the importance of the context in which those networks are embedded. Our results imply that in studies of diffusion the effect of mass media and advertising on the spread of a trend changes the nature of network-based diffusion, even if mass media and advertising have a weak role in and of themselves. To understand the drivers behind a trend, it is not sufficient to understand how well positioned the initial adopter is to spread the trend. We must also understand if advertising or other broad forces like mass media, government mandates, or search engines seed the trend widely, and thereby render the choice of the initial adopter, no matter how central to the network, irrelevant.

# References

1. M. Gladwell, *The Tipping Point: How Little Things Can Make a Big Difference* (Little, Brown, and Company, 2000).

2. R. Iyengar, C. van den Bulte, T. W. Valente, Opinion Leadership and Social Contagion in New Product Diffusion. *Mark. Sci.* **30**, 195–212 (2011).

3. E. M. Rogers, *Diffusion of Innovations*, 5th Ed. (Free Press, 2003).

4. A. Bavelas, A Mathematical Model for Group Structures. *Appl. Anthropol.* **7**, 16–30 (1948).

5. S. P. Borgatti, Identifying sets of key players in a social network. *Comput. Math. Organ. Theory* **12**, 21–34 (2006).

6. L. C. Freeman, Centrality in social networks conceptual clarification. *Soc. Netw.* **1**, 215–239 (1978).

7. F. Morone, H. A. Makse, Influence maximization in complex networks through optimal percolation. *Nature* **524**, 65–68 (2015).

8. T. W. Valente, Network Interventions. *Science* **337**, 49–53 (2012).

9. P. Bonacich, Power and Centrality: A Family of Measures. *Am. J. Sociol.* **92**, 1170–1182 (1987).

10. P. Bonacich, P. Lloyd, Eigenvector-Like Measures of Centrality for Asymmetric Relations. *Soc. Netw.* **23**, 191–201 (2001).

11. J. S. Coleman, E. Katz, H. Menzel, *Medical Innovation: A Diffusion Study* (Bobbs-Merrill Co, 1966).

12. J. W. Dearing, Applying Diffusion of Innovation Theory to Intervention Development. *Res. Soc. Work Pract.* **19**, 503–518 (2009).

13. J. A. Kelly, *et al.*, Randomised, controlled, community-level HIV-prevention intervention for sexual-risk behaviour among homosexual men in US cities. *The Lancet* **350**, 1500–1505 (1997).

14. T. W. Valente, P. Pumpuang, Identifying Opinion Leaders to Promote Behavior Change. *Health Educ. Behav.* **34**, 881–896 (2007).

15. F. M. Bass, A New Product Growth for Model Consumer Durables. *Manag. Sci.* **15**, 215--227 (1969).

16. C. van den Bulte, G. L. Lilien, Medical Innovation Revisited: Social Contagion versus Marketing Effort. *Am. J. Sociol.* **106**, 1409–1435 (2001).

17. C. Riedl, *et al.*, Product diffusion through on-demand information-seeking behaviour. *J. R. Soc. Interface* **15**, 20170751 (2018).

18. M. Weber, *Economy and Society: An Outline of Interpretive Sociology* (University of California Press, 1978).

19. E. Katz, P. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications* (Free Press, 1955).

20. P. F. Lazarsfeld, B. Berelson, H. Gaudet, *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign* (Duell, Sloan, and Pearce, 1944).

21. F. M. Bass, Comments on "A New Product Growth for Model Consumer Durables The Bass Model." *Manag. Sci.* **50**, 1833–1840 (2004).

22. V. Mahajan, R. A. Peterson, *Models for Innovation Diffusion* (Sage Publications, 1985).

23. T. W. Valente, Diffusion of Innovations and Policy Decision-Making. *J. Commun.* **43**, 30–45 (1993).

24. N. A. Christakis, J. H. Fowler, The Spread of Obesity in a Large Social Network over 32 Years. *N. Engl. J. Med.* **357**, 370–379 (2007).

25. P. J. DiMaggio, F. Garip, How Network Externalities Can Exacerbate Intergroup Inequality. *Am. J. Sociol.* **116**, 1887--1933 (2011).

26. P. S. Tolbert, L. G. Zucker, Institutional Sources of Change in the Formal Structure of Organizations: The Diffusion of Civil Service Reform, 1880-1935. *Adm. Sci. Q.* **28**, 22 (1983).

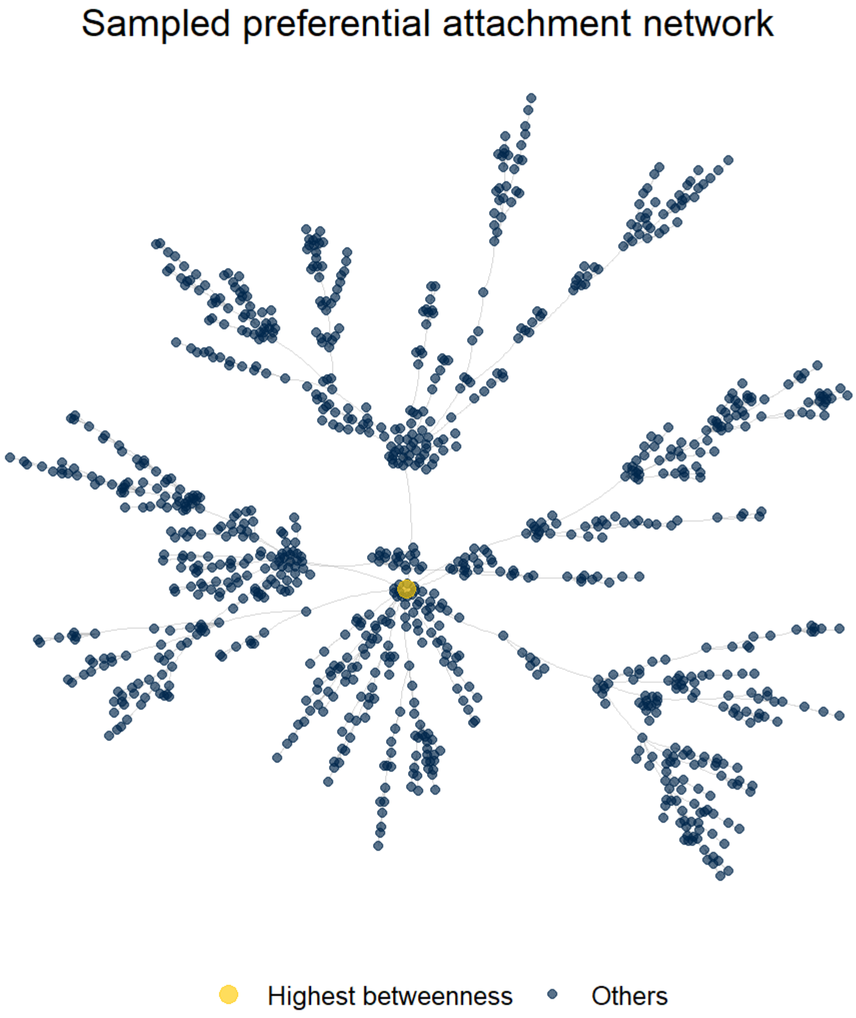
27. G. Csardi, T. Nepusz, The igraph Software Package for Complex Network Research. *Interjournal* **Complex Systems**, 1695 (2006).

28. D. J. Watts, P. S. Dodds, Influentials, Networks, and Public Opinion Formation. *J. Consum. Res.* **34**, 441–458 (2007).

29. J. Moody, The Importance of Relationship Timing for Diffusion. *Soc. Forces* **81**, 25–56 (2002).

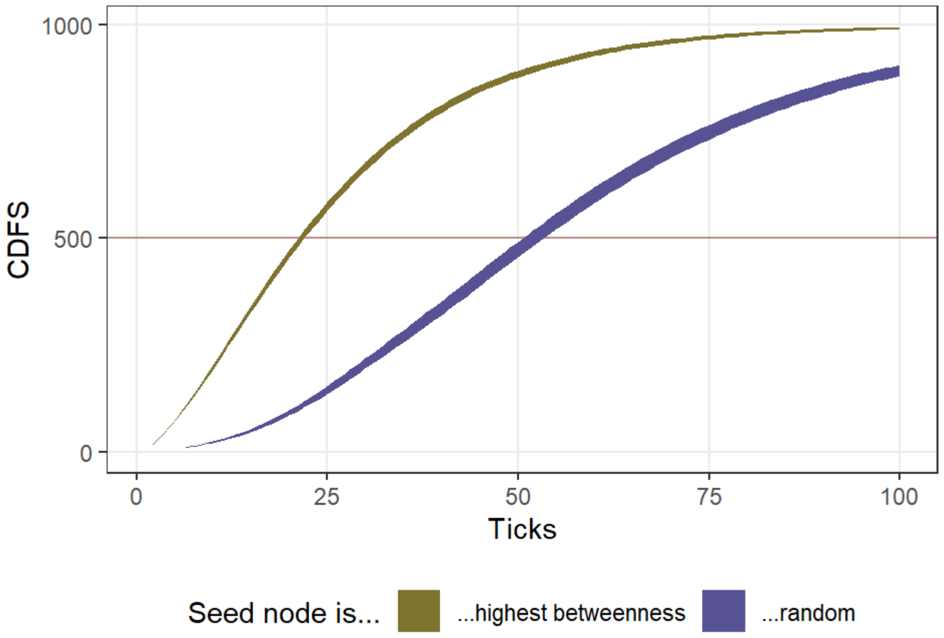
30. A.-L. Barabási, R. Albert, Emergence of Scaling in Random Networks. *Science* **286**, 509–512 (1999).

## Figure 1



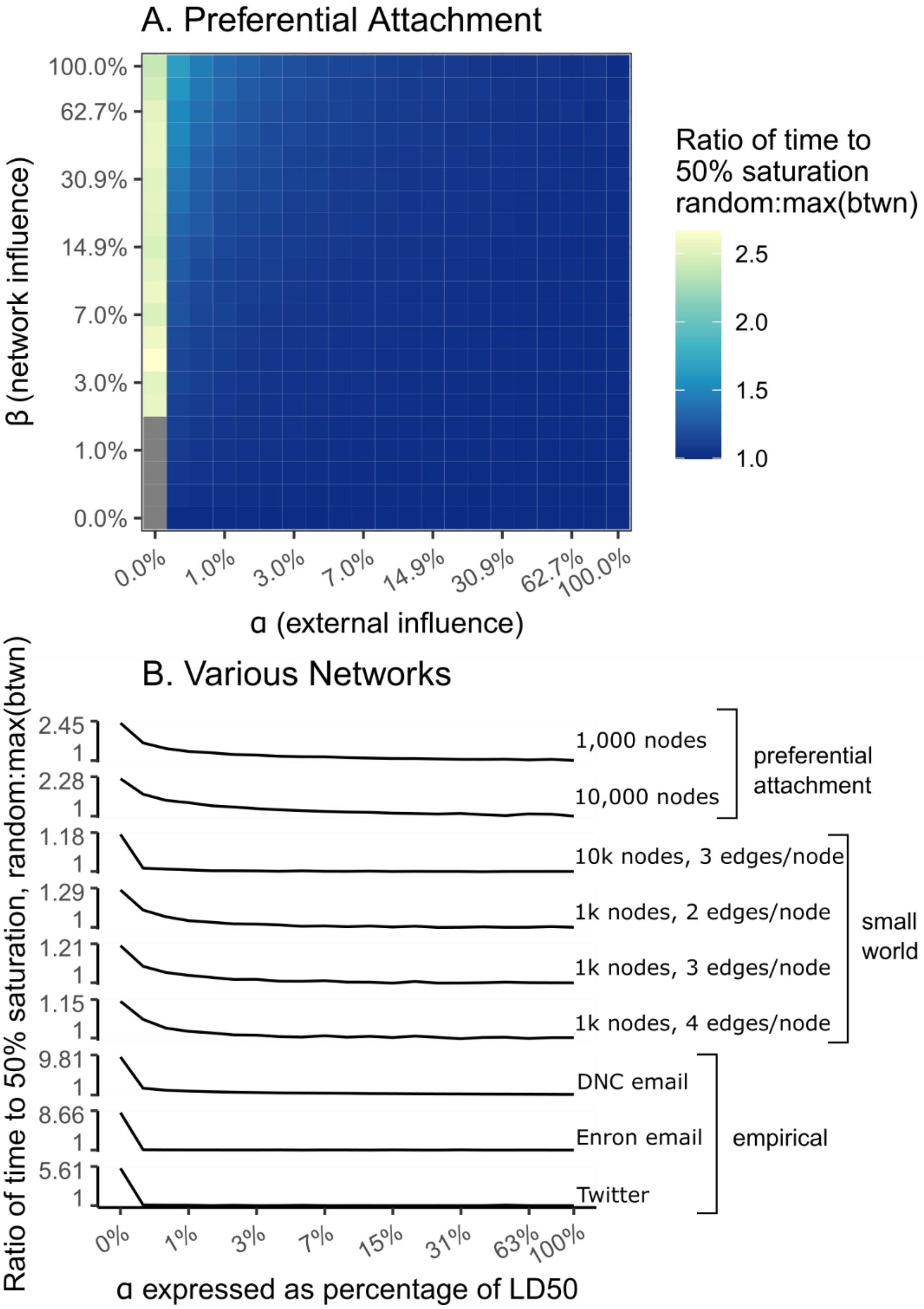
Example of a preferential attachment network generated with the Barabási-Albert algorithm(30) with 1000 nodes, one edge per node, and an exponent of one. We focus on this network as relatively favorable to opinion leadership but in figure 3 and the supplementary materials show other networks. The yellow node is the highest betweenness node used to test the effect of influentials.

## Figure 2



Cumulative number of adopters, denoted CDF, in simulations assuming only network diffusion ($\alpha = 0, \beta = LD50$) in a preferential attachment network (1000 nodes, 1 edge/node). The plot shows the confidence interval around the mean of both experimental conditions: simulations seeded with the highest betweenness node and simulations seeded with a randomly selected node. Seeding with the highest betweenness person saturates half the network (indicated by the red horizontal line) over twice as fast.

## Figure 3



Ratio of mean time to mid-saturation in simulations targeting a randomly chosen node in the network versus targeting the highest betweenness node. The top panel (A) shows the full parameter space for randomly generated preferential attachment networks (1000 nodes, 1 edge/node). Gray cells represent right censored cases. Targeting a highly central person results in adoption that is over twice as fast, but only when there is no effect of advertising ($\alpha = 0$). The bottom panel (B) shows a summary across several algorithmically generated and empirical networks as we assume high levels of network diffusion ($\beta = ld50$) but vary external influence ($\alpha$) as a percentage of each LD50 value, plotted on a logarithmic scale. This is the equivalent to the top row of cells in panel A, but substituting a y-axis for the heat dimension and showing more networks. Across all these networks, targeting a highly central person results in faster adoption, but only when there is no effect of advertising ($\alpha = 0$). The impact of highly central seeds approaches parity with random seeds at even very low positive levels of advertising.

# Network Hubs Cease to be Influential in the Presence of Low Levels of Advertising

*Supplementary Materials*

## Simulation procedure

### Agent based model

We develop an agent-based model version of the Bass(1) model of adoption. Our model takes the following steps.

1. Seed a network with a single initial adopter.
    - In some trials the seed node is chosen at random and in others the seed node is the highest centrality node in the network.
2. Calculate each person's probability of adopting at each discrete time interval using the following relationship: $P(X_i = t|X_i \geq t) = \alpha + \beta \left(\frac{\sum_{j \in N(i)} X_j < t}{|N(i)|}\right) = \alpha + \beta A Y_t$, where:
    - $\alpha$ and $\beta$ are scalar parameters expressing external influence and network influence, respectively. For any given cell in parameter space they are fixed.
    - $Y_t$ is an $n \times 1$ binary vector whose elements $y_{it}$ are 1 if $X_i < t$, meaning that $i$ adopted the innovation before time $t$, and 0 otherwise.
    - $A$ is the $n \times n$ row-normalized adjacency matrix of the network, whose cells $a_{ij}$ are defined:

    $$a_{ij} = z_{ij} / \sum_{j} z_{ij}$$

    where $z_{ij}$is 1 if $i$ and $j$ are connected and 0 otherwise.
3. Repeat step 2 until the network achieves the target level of saturation or until the maximum number of ticks has been completed.

We repeat this procedure for 1,000 trials for each cell in parameter space. For some networks, we expanded this to up to 5,000 trials to minimize right-censorship.

# Network used in the experiments

*Preferential attachment networks*(2) -- generated networks in R using igraph::sample_pa(). The networks vary only by number of nodes and all of them have edges/node and power exponent set to one. This results in a tree-like structure. To measure key network traits, we generated a thousand networks of each type. We report the average of their mean path lengths and the betweenness score for the highest betweenness node in each network expressed as a Z-score.

| Number of generated networks | Nodes | Average mean path length | Peak betweenness nodes (Z-score) |
|---|---|---|---|
| 1,000 | 1,000 | 8.4 | 18.7 |
| 1,000 | 10,000 | 11.5 | 58.6 |

*Small world networks*(3) -- generated in R using igraph::sample_smallworld(). We generated 1,000 examples of each type and varied number of nodes and edges/node. For these networks, the table below reports average mean path length and peak betweenness as a Z-score. The given parameters and measured attributes are summarized below.

| Number of generated networks | Parameters to generate the networks | | | Measured attributes of the networks | |
|---|---|---|---|---|---|
| | Nodes | edges/node | rewiring probability | average mean path length | Peak betweenness nodes (Z-score) |
| 1,000 | 10,000 | 3 | 2% | 12.4 | 10.8 |
| 1,000 | 1,000 | 2 | 2% | 13.3 | 7.1 |
| 1,000 | 1,000 | 3 | 2% | 8.5 | 6.8 |
| 1,000 | 1,000 | 4 | 2% | 6.6 | 6.5 |

*Empirical networks* – We tested three empirical examples of real-world communication networks: the DNC e-mail network, the Enron e-mail network, and a Twitter network. The *Democratic National Committee (DNC) e-mail network* is the set of Democratic party e-mails posted to Wikileaks in 2016, available online at http://konect.uni-koblenz.de/networks/dnc-temporalGraph. The *Enron e-mail network* consists of Enron e-mails collected by the Federal Energy Regulatory Commission in 2002, available online at http://snap.stanford.edu/data/email-Enron.html. The *Twitter* network contains users who mention or retweet others during January 23 - February 8, 2011 and is available online at http://www-levich.engr.ccny.cuny.edu/~min/retweetformat.txt.(4) We forced DNC and Twitter to be undirected but Enron was already undirected. For each network we eliminated redundant edges and isolated the giant component. This last step biases the analysis in favor of the hypothesis that hubs are influential. For the two networks that were originally directed, DNC and Twitter, we kept only nodes that both sent and received to identify the active core of the network before making them undirected. The attributes of these networks are summarized in the table.

| Network | Nodes | Edges | Mean path length | Peak betweenness node (Z-score) |
|---|---|---|---|---|
| DNC | 548 | 2,442 | 2.9 | 17.2 |
| Enron | 33,696 | 180,811 | 4.0 | 70.7 |
| Twitter | 532,325 | 694,606 | 9.08 | 391.1 |

# Determining parameter range

The $\alpha$ and $\beta$ parameters are not inherently on the same scale. Indeed, any given value of $\alpha$ implies much more rapid diffusion than the same value for $\beta$. We use a simulation approach to rescale the parameters to be on the same scale. Under our approach, we allow parameters to range from 0 to the parameter value at which 50% of trials for a given network will be completely saturated by 100 ticks. We refer to this parameter value as the LD50, following terminology from toxicology for the median lethal dose, meaning the dose at which 50% of the test population perished.

Note that by setting comparable maxima for $\alpha$ and $\beta$ we are not implying where the most realistic position is in parameter space and indeed this probably varies by empirical scenario.(5) For awareness, or learning about ideas or products, diffusion tends to be characterized by α. However, adoption, or actually embracing new ideas or purchasing new products, tends to skews towards β, especially when the new idea or behavior has low legitimacy or would be onerous to implement. There are exceptions though and adoption can skew towards α if there is aggressive promotion and the product being promoted is a good fit with the expectations of its target market.(6, 7)

We estimated these LD50 values using a separate set of simulations. We repeatedly run simulated diffusion processes on the networks that we use to determine the appropriate LD50 values through trial and error. We calculate LD50 values separately for each network formation heuristic or empirical network. For instance, preferential attachment with a thousand nodes has one set of LD50s, preferential attachment with ten thousand nodes another, and the Enron network a third.

In our main set of simulations, we use LD50 values to define the maximum for the range of each dimension of parameter space and we explore parameter space by taking 21 logarithmically spaced steps from 0 to the LD50 value for both $\alpha$ and $\beta$. We compute the logarithmically spaced percentage steps with the following R code:

```
expm1(seq(from = 0, to = log1p(100), length.out = 21)) / 100
```

and then multiply this vector by the LD50 value for each parameter of each network. This results in the following sequence:

```
0.0%, 0.3%, 0.6%, 1.0%, 1.5%, 2.2%, 3.0%, 4.0%, 5.3%, 7.0%, 9.0%,
11.7%, 14.9%, 19.1%, 24.3%, 30.9%, 39.1%, 49.5%, 62.7%, 79.2%, 100.0%
```

For example, DNC has an LD50 of 6.52 for $\alpha$, which results in the following intervals (rounded to two digits for clarity):

```
0.00, 0.02, 0.04, 0.07, 0.10, 0.14, 0.20, 0.26, 0.35, 0.46, 0.59, 0.76,
0.97, 1.24, 1.58, 2.01, 2.55, 3.23, 4.09, 5.16, 6.52
```

Where 0.02 = 6.52 * 0.003, 0.04 = 6.52 * 0.006, etc.

DNC's LD50 of 27.98 for $\beta$ results in:

```
0.00, 0.07  0.16  0.28  0.42  0.61, 0.84, 1.13, 1.49, 1.95, 2.53, 3.26,
4.18, 5.34, 6.80, 8.63, 10.95, 13.86, 17.53, 22.16, 27.98
```

R code for this process would be as follows:

```
alpha_ld50 <- 6.52
beta_ld50 <- 27.98
log_pcts <- expm1(seq(from = 0, to = log1p(100), length.out = 21)) /
100
alpha_intervals <- alpha_ld50 * log_pcts
beta_intervals <- beta_ld50 * log_pcts
```

At a conceptual level, we are taking logarithmically spaced steps of percentages then we multiply those logarithmically spaced steps by the relevant LD50 to establish the intervals. This gives us a fine-grained view of low values, which is important as the collapse of a special role for high centrality nodes occurs at very low values for $\alpha$; typically below 1% of $\alpha$'s LD50. We prefer the order of operations in which we take the logs of percentages then multiply by the LD50 over simply taking the logarithms of LD50s in that it results in scale intervals that are intuitively comparable regardless of the maximum. Since $\beta$ LD50s are always greater than $\alpha$ LD50s, our order of operations biases the analysis against our hypothesis that low levels of $\alpha$ qualitatively change the nature of diffusion.

# Right censorship

When both parameters were set to extremely low values, the simulation can take a very long time to reach mid-saturation in all 1,000 trials per cell. We allowed simulations to run for either 1,000 iterations (Twitter, preferential attachment with 10,000 nodes, and all small world networks) or 5,000 iterations (DNC email, Enron email, and preferential attachment with 1,000 nodes). Nonetheless, some cells still had right-censored trials. If less than 10% of trials in a cell were right-censored, we top-coded those right-censored cells at the maximum number of iterations the trial was allowed to run. If more than 10% of trials in a cell were right-censored, we treated that cell as missing. The level of missing data and the top-code for each treatment condition in each cell of parameter space can be found in the replication files ending ".rds" as the columns *missing_reps* and *max.ticks*, respectively. This top-coding has the effect of biasing the results slightly towards parity between experimental conditions, especially in the region of parameter space where $\alpha = 0$ and $\beta$ is close to 0. Since right-censorship is almost entirely an issue for cells where $\alpha = 0$, top-coding has the effect of making the plane where $\alpha = 0$ more similar to the rest of parameter space. This biases our results *against* our argument that opinion leadership is distinct to $\alpha = 0$.

Readers who prefer to see the results with differing tolerance for right-censored data may use the replication files and change the parameter *p_censored_ok* in the main RMarkdown file. For instance, to drop cells entirely that have any right-censored trials, please set *p_censored_ok* <- *0*. (Note that Twitter has some missing data for every cell in the crucial left column and so dropping *p_censored_ok* < .03 will drop the left column of the heat map with the result that the scale is radically compressed and the heat map will display noise among substantively similar cases.)

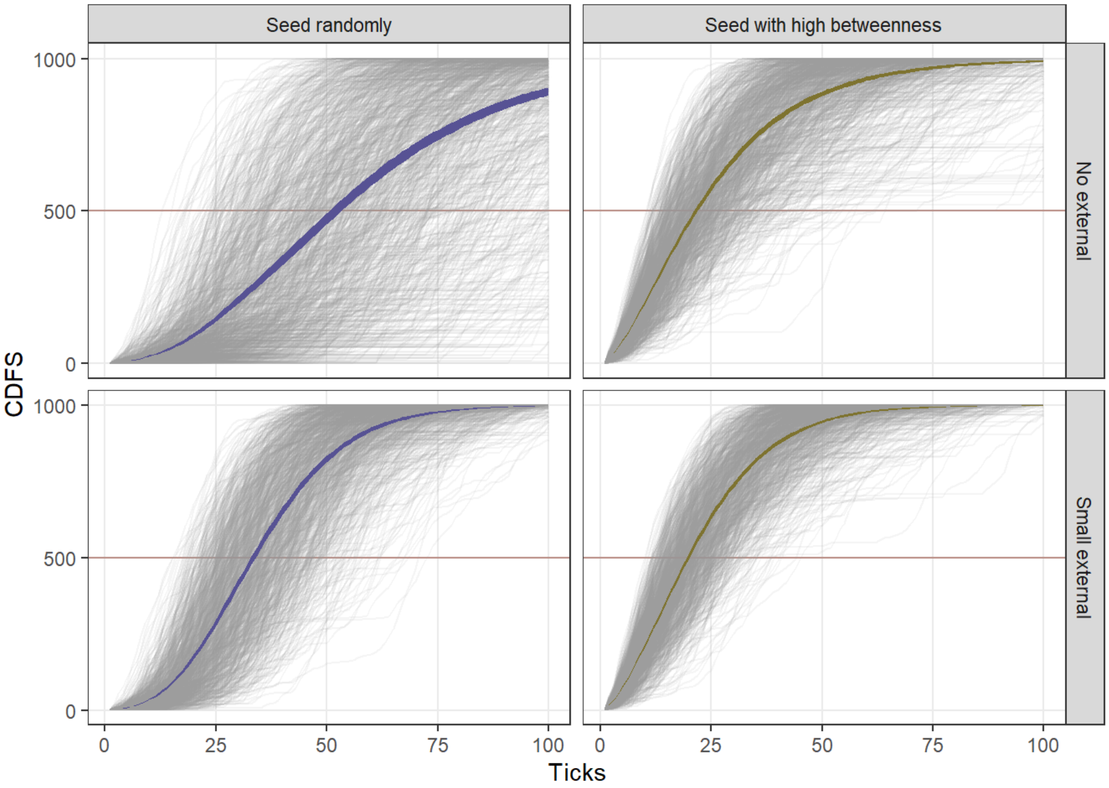
# CDF plots of $\alpha = 0, \beta = LD50$ and $\alpha = 0.0026LD50, \beta = LD50$ for random and high betweenness seeds

Figure 3 illustrates that the key transition in parameter space occurs when medium to high network influence (β) combines with either zero or a tiny bit of external influence (α). In Figure S1, we visualize the micro-dynamics of the rapid decay of the importance of central nodes with spaghetti plots showing individual CDFs with ribbon plots showing the central tendencies. Note that the top panels are a minor variation on figure 2 but the introduction of the spaghetti plots shows that the principal difference between random and high betweenness seeds is not how fast they can be, but how slow, with the fastest CDFs being similar, but random seeds having many glacial CDFs.

The bottom panels illustrate three things about the micro-level effect of introducing a modicum of external influence. First, as would be expected from the corresponding cell in the heat map, the two bottom panels are much more similar than are the top two panels. Second, the bottom-left panel (displaying randomly seeded simulations with low external influence) has CDFs that are still qualitatively s-curves, validating that this point in parameter space really is primarily an endogenous process and the external influence at this level is mostly a catalyst to change the nature of endogenous diffusion rather than an appreciable diffusion force itself. Third, the introduction of a small external influence makes the individual CDFs less stochastic and in particular nearly eliminates laggard outliers. In contrast, when α = 0, β = LD50 (top panels), there is enormous variation in diffusion trajectories, especially with a random seed node. At the micro-level, a small amount of external influence not only makes diffusion faster, it makes it more predictable.

This is all illustrated with a difference between α set to zero versus about one quarter of a percent of LD50 and even that tiny increment shows visually striking differences. At slightly larger levels of α of 1% or 2% of LD50, the CDFs based on random vs. high betweenness seeds are indistinguishable.

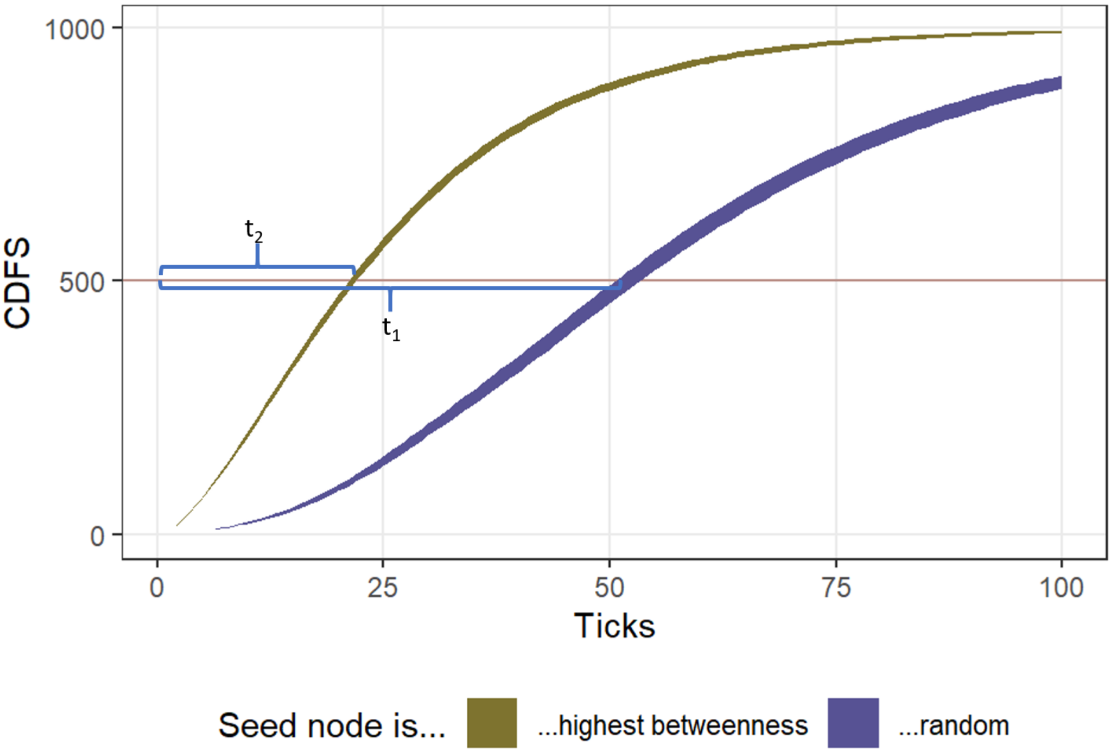
# Figure S1



CDFs for diffusion on a preferential attachment network with a thousand nodes. The mean trajectory with confidence intervals for each panel is shown as a ribbon plot. Individual trials are shown as spaghetti plots. The panels represent α = 0, β=LD50 (top) versus $\alpha = 0.26\% \times LD50, \beta = LD50$ (bottom) and random seeds (left) vs high betweenness seeds (right).

# Correspondence of Figure 2 to Figure 3

Figure S2

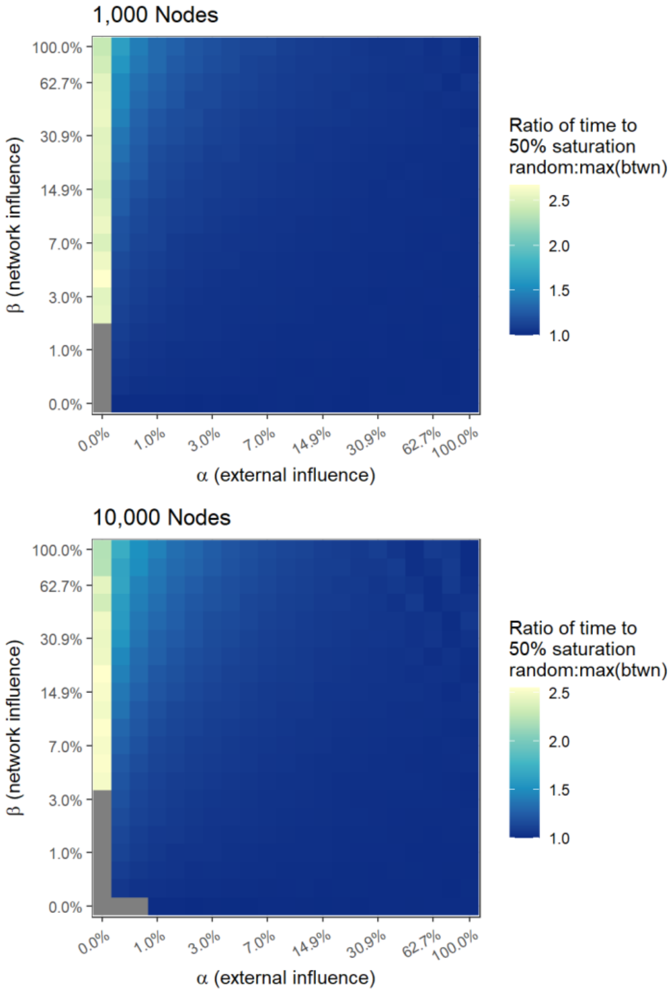


Same as Figure 2 but annotated with brackets to highlight the mean time to mid-saturation for randomly seeded trials ($t_1$) and trials seeded with the highest centrality node (t2). The heat dimension in all heat maps (e.g., the top panel of figure 3) represents the ratio of $t_1$:$t_2$. Likewise, the height of the sparklines in the bottom panel of figure 3b represents the ratio of $t_1$:$t_2$.

# Preferential attachment networks

In Figure S3, we reproduce the top panel of figure 3 (based on 1,000 node preferential attachment networks) and replicate it for networks with 10,000 nodes.

## Figure S3

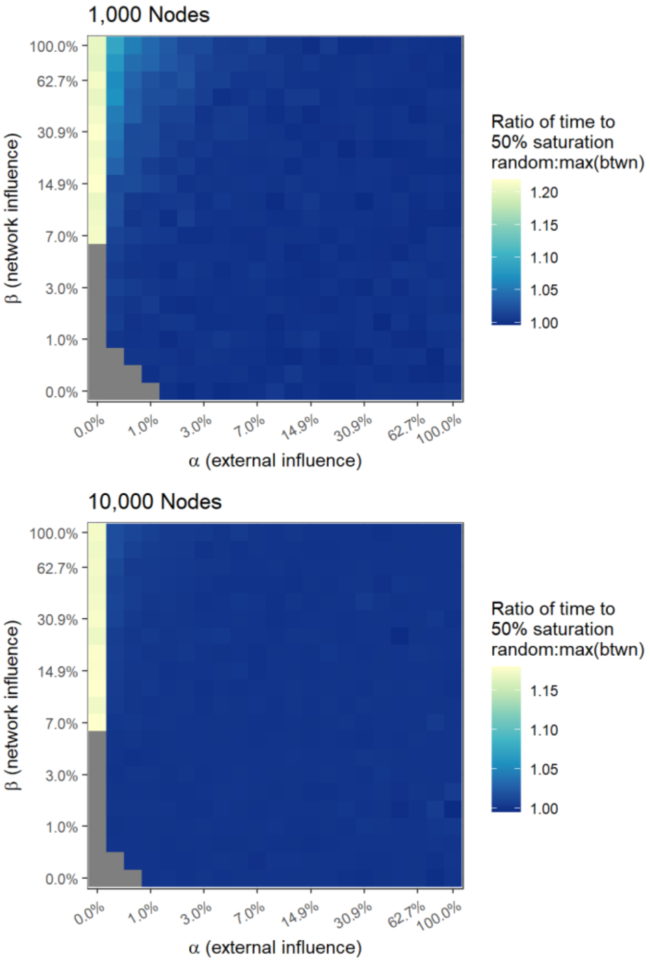


Heatmap of ratio of time to adoption. Gray cells represent right censored cases after 5,000 iterations (1,000 nodes) or 1,000 iterations (10,000 nodes). Ratio of time to adoption in simulations targeting the highest betweenness node versus targeting a randomly chosen node in the network. The sampled graphs follow a preferential attachment network (either 1,000 or 10,000 nodes, 1 edge/node, and an exponent of 1).

# Small world networks

In Figures S4 and S5, we rerun the simulation on Watts-Strogatz small world networks. A small number of generated small world networks had disconnected components. We replaced these networks with newly generated networks to bias the simulation in favor of opinion leadership. Aside from the different network (small world vs. preferential attachment), the figures are similar to the top panel of Figure 3.

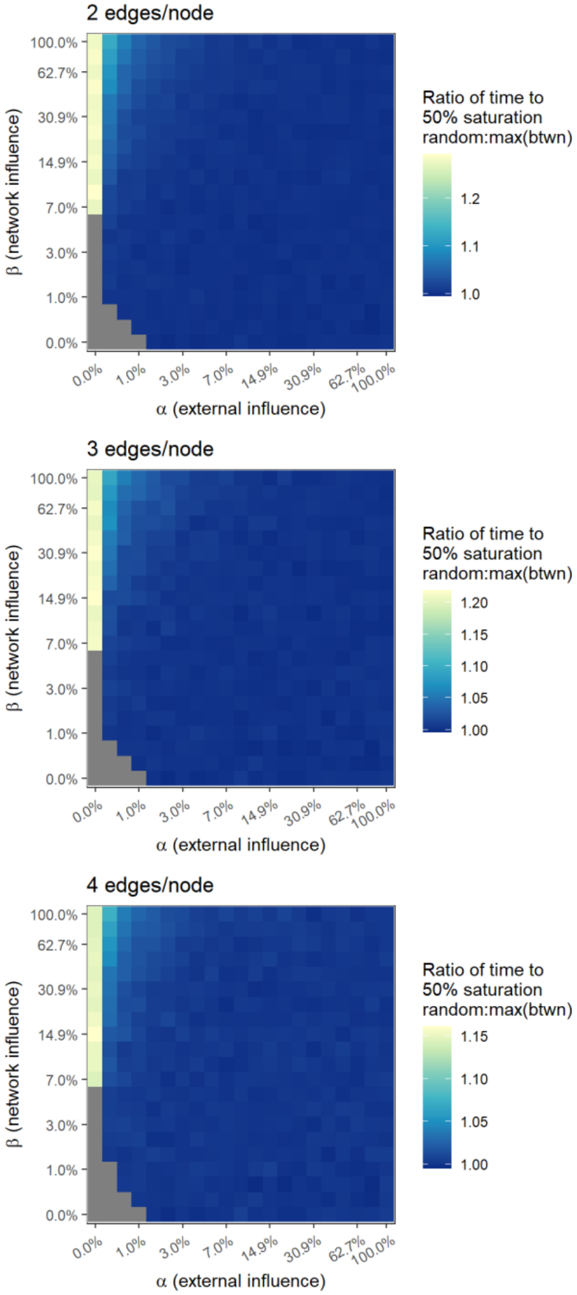
## Figure S4


1,000 Nodes

10,000 Nodes

Heatmap of ratio of time to adoption. Gray cells represent right censored cases after 1,000 iterations. Ratio of time to adoption in simulations targeting the highest betweenness node versus targeting a randomly chosen node in the network. The sampled graphs follow a Watts-Strogatz small world network (either 1,000 or 10,000 nodes, 3 edges/node, and 2% probability of random rewiring).

# Figure S5

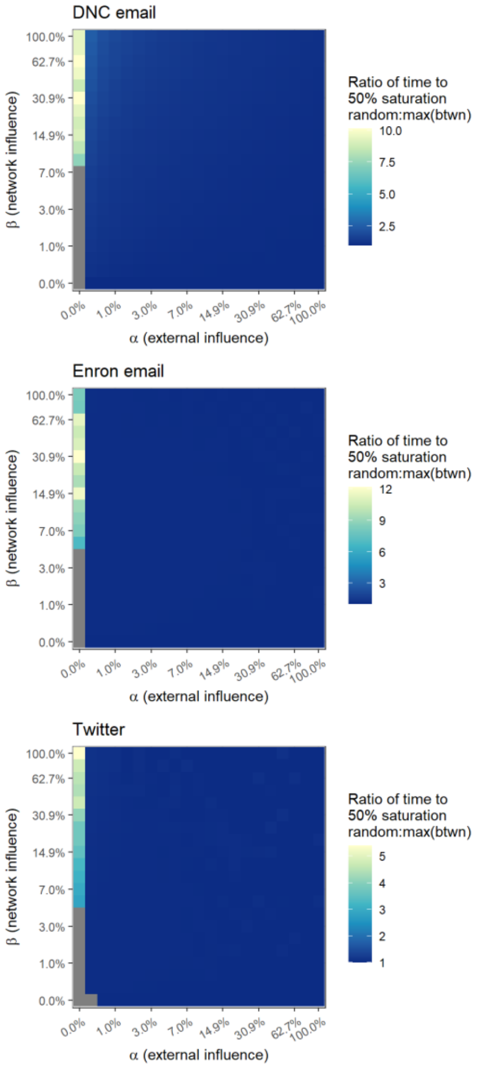


Heatmap of ratio of time to adoption. Gray cells represent right censored cases after 1,000 iterations. Ratio of time to adoption in simulations targeting the highest betweenness node versus targeting a randomly chosen node in the network. The sampled graphs follow a Watts-Strogatz small world network (1,000 nodes, 2-4 edges/node, and 2% probability of random rewiring).

# Empirical networks

In Figure S6, we rerun the simulation using three empirical networks. The specification is otherwise similar to the top panel of Figure 3.

## Figure S6

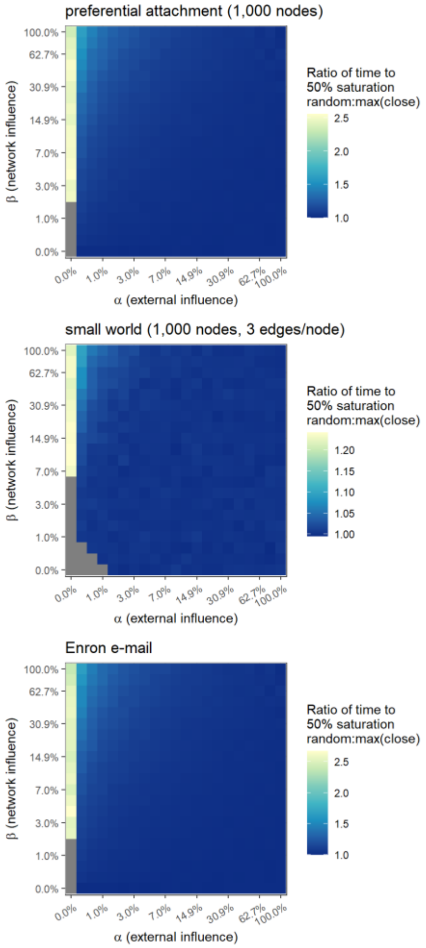


Heatmap of ratio of time to adoption. Gray cells represent points in parameter space where over 10% of trials were right censored after 5000 iterations (DNC and Enron) or 1000 iterations (Twitter). For cells with less than 10% right-censorship, right-censored trials were top-coded at 1000 iterations (Twitter) or 5000 iterations (DNC and Enron).

# Closeness centrality

We primarily use betweenness centrality to identify seed nodes, but in this alternate specification (figure S7) we experiment with closeness centrality as a demonstration that our findings are not limited to using betweenness to identify seed nodes.

## Figure S7



Heatmap of ratio of time to adoption comparing simulations seeded with the highest closeness node versus seeded at random. Gray cells represent >10% right censored cases after 1,000 iterations (small world) or 5,000 iterations (preferential attachment and Enron). Aside from using closeness centrality to choose the seed node, the specification is identical to that in Figures 3 (top panel), S5 (middle panel), and S6 (middle panel).

# Multiple seeds

We primarily use a single seed node. In figures S8, S9, and S10, we compare this to 1% seeds and 5% seeds. Figure S8 shows the ratio of mid-saturation diffusion times for random versus targeted nodes when $\beta = LD50$ and $\alpha$ varies along the usual log scale between 0 and LD50. LD50 values are re-estimated for 1% and 5% seeds. Figures S9 and S10 show heat maps for preferential attachment graphs (1,000 nodes) and DNC E-mail (548 nodes), respectively.

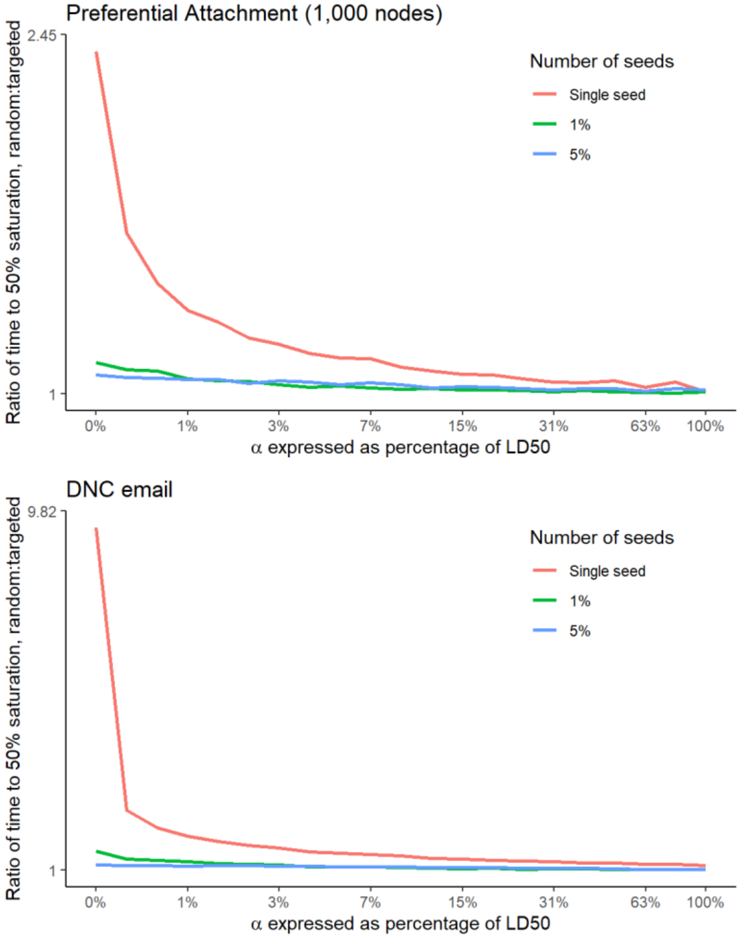
Because betweenness centrality scores a node's importance to the network irrespective of whatever nodes have already been identified, it is not well suited for identifying multiple seeds. We thus identify multiple seeds using Borgatti's keyplayer algorithm as implemented in the R library influenceR to identify a 5% seed.(8, 9).

The multiple seed results differ in two notable ways from results based on a single node. First, as shown in Figure S8, the baseline effect of targeted seeding in the absence of external influence is much smaller. In the region of parameter space that is ideal for opinion leadership ($\alpha = 0, \beta = LD50$), a single random seed takes 150% longer than a maximum betweenness seed to saturate half of a preferential attachment network and 881% longer to saturate the DNC e-mail network. In contrast, a 1% random seed versus a 1% seed identified by the keyplayer algorithm takes 15% longer to saturate half of a preferential attachment network and 47% for DNC. The comparable figures for a 5% seed are 8% on a preferential attachment and 15% longer on DNC e-mail. Thus, choosing an optimal 5% seed is at least an order of magnitude less important than choosing an optimal single seed.

Second, as shown in figures S9 and S10, the drop-off in the advantage of a highly targeted seed diminishes with multiple seeds. This is fairly minimal with a 1% seed as both preferential attachment and DNC e-mail still show something like the yellow stripe and blue field pattern familiar from this paper's other heat maps. The only difference is that the 1% seed heat maps show a gradient for values of $\alpha < 0.03 \times LD50$. Nonetheless, the 1% seed heat maps mostly resemble the single seed heat maps. The gradient extends much further with a 5% seed and now appears smooth. Note that the heat map has a log-log scale, such that even with a 5% seed, the baseline advantage ($\alpha = 0, \beta = LD50$) nearly disappears by the time $\alpha > 0.3 \times LD50$. However, this should not be discounted as this is a plausible region of parameter space for many empirical applications.

We must thus conclude that the effect of targeting multiple seeds is much smaller but also more robust to the introduction of external influence than is targeting a single seed and that both the decline in the baseline effect and the robustness to the introduction of external influence are roughly proportional to the size of the seed. Under many realistic scenarios, targeting multiple seeds may be practical so long as the cost of identifying and targeting specific seeds is low and external influence is very expensive and/or very ineffective. Outlining the contours of this trade-off is an issue for future research.

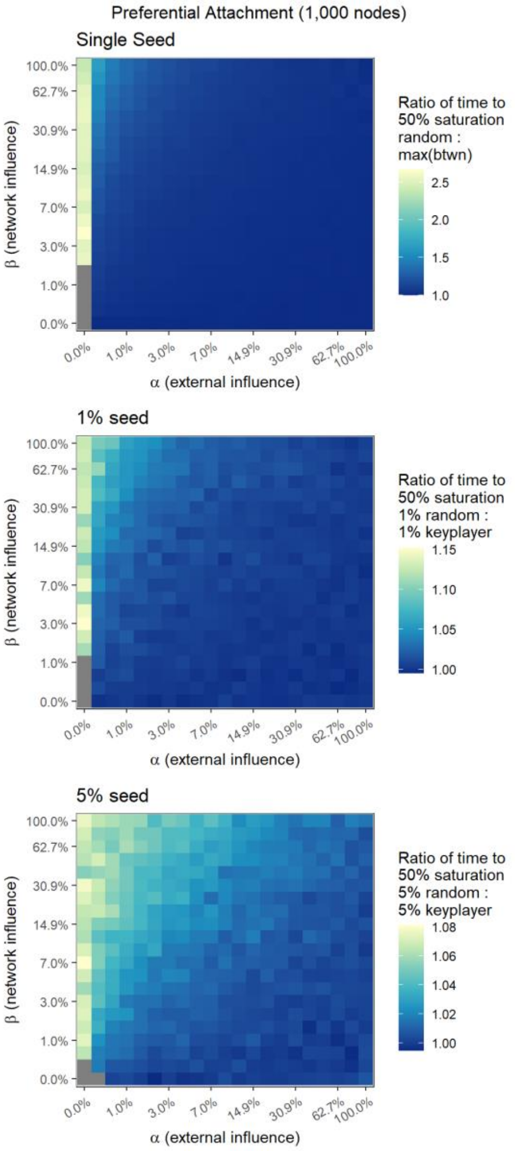
# Figure S8



Targeted seed is selected as max(betweenness) for single seed
and using keyplayer for 1% and 5% seeds

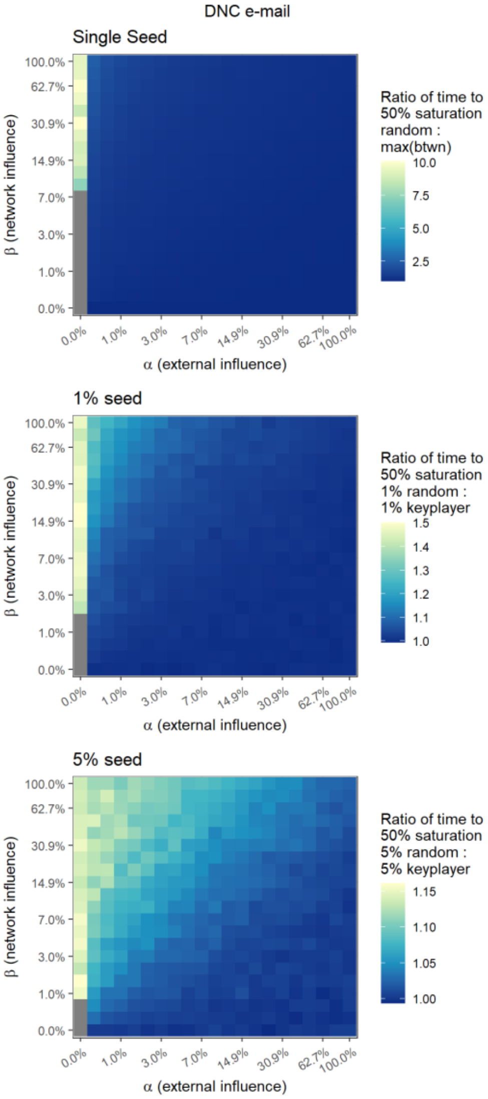
Line graph of ratio of time to adoption comparing simulations on 1,000 node preferential attachment networks (top panel) or DNC email network (bottom panel) with random seed(s) to targeted seed(s) identified by maximum betweenness for a single seed or keyplayer for 1% or 5% seeds. Both panels assume high levels of network diffusion ($\beta = LD50$) but vary external influence ($\alpha$) as a percentage of each LD50 value, plotted on a logarithmic scale. This is the equivalent to the top row of cells in Figures S9 and S10 but substituting a y-axis for the heat dimension. In both preferential attachment and DNC, the baseline effect of targeting a single node is higher but more sensitive to rising $\alpha$.

## Figure S9



Heat maps summarizing ratio of time to adoption for random seed(s) versus targeted seed(s) on preferential attachment networks with 1,000 nodes. Targeted seeds are identified by maximum betweenness for a single seed or keyplayer for 1% or 5% seeds. Gray cells represent >10% right censored cases after 5,000 iterations.

## Figure S10



Heat maps summarizing ratio of time to adoption for random seed(s) versus targeted seed(s) on DNC email network. Targeted seeds are identified by maximum betweenness for a single seed or keyplayer for 1% or 5% seeds. Gray cells represent >10% right censored cases after 5,000 iterations.

# Status-biased Diffusion

In most of our models we treat network-based diffusion pressure as proportional to the percentage of a node's alters who have already adopted. In Figure S11 we allow high degree nodes to have more influence. Theoretically this reflects a heuristic that nodes see their high degree neighbors as prestigious and especially worthy of emulation. In the status-biased model, the baseline effect ($\alpha = 0, \beta = LD50$) is slightly stronger than in the standard model (midpoint saturation is 3.3x vs 2.7x faster with a hub seed). However, both models show this paper's general pattern that the advantage for seeding with the most central node disappears with the introduction of any external influence.

A quick overview of status-biased diffusion in our model.

Take the following network:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Normally we row normalize and multiply by a column vector to get the percentage of one's alters who have adopted. Concretely, let's say person 1 (row / column 1) is the only adopter. We would do the following:

$$\begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1/2 \\ 1 \\ 0 \end{bmatrix}$$

Adoption is then proportional to the resulting column vector.

For figure S11, we bias in favor of high status people by multiplying each column by a function of its sum (the person's degree) before row normalizing. (Namely, we multiply but the log of degree in accordance with our intuition that one's status bias increases with the order of magnitude of one's popularity.) For example, the columns multiplied by their sums would give:

$$\begin{bmatrix} 0 & 1 & 2 & 1 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \end{bmatrix}$$

And the row-normalized multiplication would be:

$$\begin{bmatrix} 0 & 1/4 & 1/2 & 1/4 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 3/4 & 0 & 0 & 0 & 1/4 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3/4 \\ 1 \\ 0 \end{bmatrix}$$

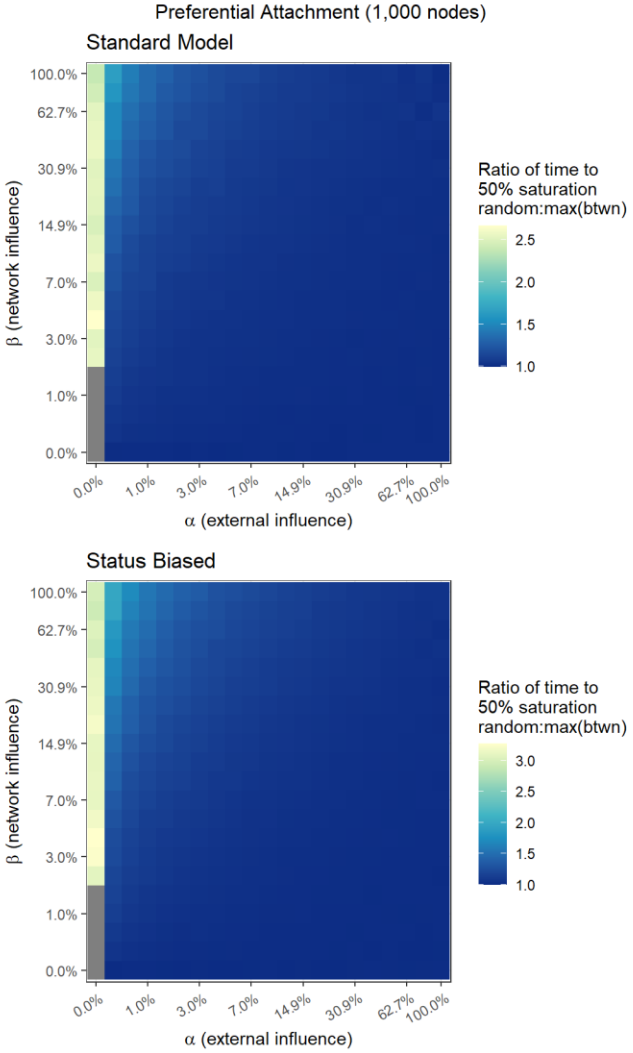
This is equivalent to taking the weighted percentage of one's alters who have adopted, where the weights are given by the degree. If the degree of the $i$-th node is $d_i$, then the weighted percentage of alters is

$$\left(\frac{d_i}{\sum_{i=1}^{n} d_i}\right) \mathbf{1}(i \text{ adopted})$$

When unweighted, $d_i = 1$ for all $i$, and this simplifies to

$$\left(\frac{1}{n}\right) \mathbf{1}(i \text{ adopted})$$

# Figure S11



Preferential Attachment (1,000 nodes)

Standard Model

Status Biased

Heat maps summarizing ratio of time to adoption for random seed(s) versus targeted seed(s) on preferential attachment network. The top panel assumes nodes are equally influential and the bottom panel assumes high degree nodes are more influential. Gray cells represent >10% right censored cases after 5,000 iterations.

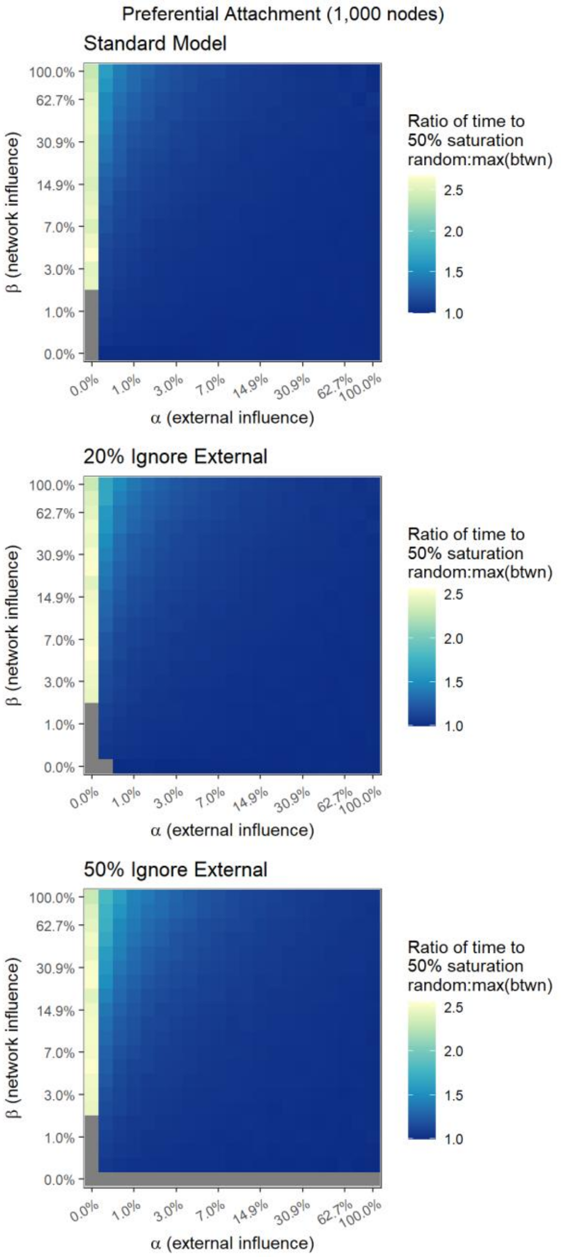
## Heterogeneity of Susceptibility to External Influence

In most of our simulations, we assume that all nodes are equally susceptible to external influence. However, we can imagine that some people are invulnerable to external influence. Perhaps they do not see television advertisements as they either do not own a television or only watch Netflix. Or in the case of government mandates, maybe some firms are exempt from the mandate by virtue of being below a certain threshold of employees or having no government contracts. (The employee threshold is particularly important in France and reliance on government contracts in the United States).(10, 11) Or maybe some people are simply intensely skeptical of external influence.

To account for these possibilities, in Figure S12 we relax the assumption of equal susceptibility to external influence by allowing for the possibility that 20% or 50% of nodes completely ignore external influence. Note that under this assumption the LD50 for $\alpha$ is undefined, however in a qualitative sense the LD50 is still a reasonable approximation (albeit a conservative one) so long as the fraction of nodes invulnerable to external influence is not too high.

The results with 20% of nodes being invulnerable to external influence are indistinguishable from those where all nodes are susceptible to external influence. The results with 50% of nodes being invulnerable are also broadly similar but have two properties of note. First, the bottom row of the heat map where $\alpha = 0$ is undefined. This is necessarily true as the measure is based on how many time periods it takes for adoptions to exceed 50% and this is impossible for these positions in parameter space. Second, there is a slightly longer gradient than usual as the effect of highly central seeds takes a bit higher values of $\alpha$ to fully fade out. However, this is easily explicable by the LD50 being undefined and us substituting a value that in a qualitative sense is probably about half as big as it ought to be. The model is remarkably robust to a small to moderate number of nodes being completely invulnerable to external influence. This reflects the classic two-step flow model where many people are completely inattentive to mass media on a particular subject but learn about new products, ideas, and behaviors from locally influential opinion leaders who themselves are attentive to mass media.(12)

# Figure S12



Heat maps summarizing ratio of time to adoption for random seed(s) versus targeted seed(s) on preferential attachment network. The top panel is the standard model. The middle panel and bottom panel show results when 20% or 50%, respectively, of nodes completely ignore external influence. Gray cells represent >10% right censored cases after 5,000 iterations.

# References

1. F. M. Bass, A New Product Growth for Model Consumer Durables. *Manag. Sci.* **15**, 215--227 (1969).

2. A.-L. Barabási, R. Albert, Emergence of Scaling in Random Networks. *Science* **286**, 509–512 (1999).

3. D. J. Watts, S. H. Strogatz, Collective Dynamics of "Small-World" Networks. *Nature* **393**, 440–442 (1998).

4. F. Morone, H. A. Makse, Influence maximization in complex networks through optimal percolation. *Nature* **524**, 65–68 (2015).

5. T. W. Valente, Diffusion of Innovations and Policy Decision-Making. *J. Commun.* **43**, 30–45 (1993).

6. G. Rossman, *Climbing the Charts* (Princeton University Press, 2012).

7. C. van den Bulte, G. L. Lilien, Medical Innovation Revisited: Social Contagion versus Marketing Effort. *Am. J. Sociol.* **106**, 1409–1435 (2001).

8. S. P. Borgatti, Identifying sets of key players in a social network. *Comput. Math. Organ. Theory* **12**, 21–34 (2006).

9. S. Jacobs, A. Khanna, *influenceR: Software tools to quantify structural importance of nodes in a network* (2015).

10. F. Dobbin, F. R. Sutton, The Strength of a Weak State: The Rights Revolution and the Rise of Human Resources Management Divisions. *Am. J. Sociol.* **104**, 441–476 (1998).

11. L. Garicano, C. Lelarge, J. Van Reenen, Firm Size Distortions and the Productivity Distribution: Evidence from France. *Am. Econ. Rev.* **106**, 3439–3479 (2016).

12. E. Katz, P. Lazarsfeld, *Personal Influence: The Part Played by People in the Flow of Mass Communications* (Free Press, 1955).