The Opacity Problem in Social Contagion

George Berry, Christopher J. Cameron, Patrick Park, and Michael Macy

Department of Sociology, Cornell University, 323 Uris Hall, Ithaca NY 14853, USA
SONIC Research Group, Northwestern University, Frances Searle Building, Evanston, IL 60208, USA
Department of Information Science, Cornell University Gates Hall, Ithaca NY 14853, USA

*To whom correspondence should be addressed; E-mail: geb97@cornell.edu.

September 26, 2017

Abstract

Fads, product adoption, mobs, rumors, memes, and emergent norms are diverse social contagions that have been modeled as network cascades. Empirical study of these cascades is vulnerable to what we describe as the “opacity problem”: the inability to observe the critical level of peer influence required to trigger an individual’s behavioral change. Even with maximal information, network cascades reveal intervals that bound critical levels of peer exposure, rather than critical values themselves. Existing practice uses interval maxima, which systematically over-estimates the social influence required for behavioral change. Simulations reveal that the over-estimation is likely common and large in magnitude. This is confirmed by an empirical study of hashtag cascades among 3.2 million Twitter users: one in five hashtag adoptions suffers critical value uncertainty due to the opacity problem. Different assumptions about these intervals lead to qualitatively different conclusions about the role of peer reinforcement in diffusion. We introduce a solution that combines identifying tightly bounded intervals with predicting uncertain critical values using node-level information.

Keywords: social contagion, social influence, peer effects, measurement
Like epidemic diseases [30, 45, 11], social contagions are ubiquitous, highly consequential, and widely studied [24, 39, 40, 13, 32, 42, 43, 16, 44]. Unlike nearly all diseases, social contagions often require multiple sources of infection, especially if adoption is costly, risky, motivated by affect, or entails positive network externalities [13].

We discover a challenge confronting observational studies of social contagion, which we term the “opacity problem”. Data generated by network cascades often fail to precisely indicate the critical value of social influence required for a node to change behavior. Instead, cascades reveal intervals in which critical values lie. Researchers typically infer the amount of social reinforcement required to trigger activation by recording node exposure (number of active neighbors) at activation time. This procedure is equivalent to taking the maximum of the critical value interval for each node. We note that this exposure-at-activation (EAA) rule can be found in diverse studies across disciplines, including sociology [22, 43, 20], economics [19, 10, 6], medicine [16], and information science [11, 37, 38, 23, 41, 9, 48]. It is even present in “snapshot” studies [4, 18] which compare a network observed at two or more time intervals.

The implications of opacity and the EAA rule are potentially far-reaching: the data used for empirical estimates of peer influence likely overstate the level of influence required to change behavior. This suggests that estimates of peer influence suffer bias, including the possibility of estimating peer effects when choices are in fact independent.

Simulated cascades on a variety of network topologies, using both deterministic and probabilistic activation rules, demonstrate that opacity and the EAA rule often produce over-estimates of critical values. An empirical study of hashtag cascades among 3.2 million users on Twitter confirms this intuition: different responses to the opacity problem produce qualitatively different conclusions about whether peer reinforcement promotes diffusion.

Unlike Manski’s famous reflection problem [33], opacity cannot be solved by alternative model specification or more fine-grained data (such as more frequent samples of node behavior). It is part of the data generated by cascades.

Fortunately, opacity can be addressed by: 1) identifying which node critical values are measured precisely based on the width of their activation intervals; and 2) using models trained on the precisely measured nodes to predict for nodes whose critical values are uncertain. In addition, we expect future research to develop novel ways to further reduce measurement error introduced by opacity and the EAA rule.

The importance of cascades and the difficulty of performing experiments on graphs means that cascade dynamics are often studied observationally. This is true even in settings where higher-order units are randomized, such as villages [6] or health discussion groups [14]. In cases where ego-networks are randomized [5, 2], the opacity problem we highlight can be avoided. In particular, this method solves the opacity problem by turning a large, complex network into many much simpler...
structures where dependencies between nodes are limited (see [21, 3] for a discussion of how to design experiments in networks while avoiding interference).

1 Intuition

1.1 Example

Suppose we wish to measure the level of social influence required for a focal node ("ego") to adopt a hashtag on social media. Ego logs in without any neighbors using the hashtag (exposure 0) and then logs out for a period of time. While ego is logged out and “not looking”, three of ego’s network neighbors use the hashtag. Ego then logs in again, sees that three neighbors have adopted the hashtag, and adopts the hashtag (exposure 3). Given the available information, ego’s critical exposure could be 1, 2, or 3. Ego therefore is measured as having a critical value interval [1, 3] rather than a point.

Standard empirical practice in observational studies of network cascades is to record ego’s exposure-at-activation (EAA) as the “critical value”, which in this example is 3. Yet we cannot rule out the possibility that ego would have activated at exposure 1 or 2, in which case the EAA rule over-estimates the required level of social influence.

The opacity problem refers to uncertainty about the precise critical value within an interval. Put differently, opacity is the inability to observe a counter-factual: the number of activated neighbors at which ego would have activated had ego logged in sooner. The problem is not limited to cascades observed on social media, nor is the problem solved if ego were to continuously update the status of her neighbors. For example, suppose we are studying a cascade in real time on a small face-to-face network such as a workplace or classroom. We observe that ego adopted only after three of ego’s neighbors simultaneously adopted. There is no way to know if ego might have adopted if only one or two neighbors had adopted and not all three. The critical value is still uncertain in an interval.

We may conjecture that if only ego had checked (“updated”) more frequently, the opacity problem could be avoided. However, simple graphs (Figure 1-B) demonstrate that the opacity problem is often inevitable for at least one node. In an enumeration of all two-, three-, and four-node graphs with all viable critical value combinations (Figure 2), we find that 75% always suffer from opacity for at least one node.

1.2 Implications

The “worst-case” outcome of the opacity problem occurs when an “innovator” (critical value 0) activates after all of its neighbors have activated. In this case, the node has a large critical value interval, and the EAA rule has error equal to the node degree. This raises the alarming possibility that much prior research on diffusion greatly overstates the level of social reinforcement required for behavioral change.
Yet the seriousness of the opacity problem may be greatest when the error magnitude is minimal. Consider the case of an innovator who has exposure 1 at activation (Figure 1A). The EAA rule assigns this node a critical value of 1. This minimum possible over-estimation leads to a qualitatively incorrect conclusion that social influence is responsible for behavioral change that is actually independent. The opacity problem therefore creates an additional problem for researchers seeking to determine the role of influence as opposed to homophily in empirical cascades [1].

The opacity problem also impacts graph-level analyses. In large hashtag cascades on Twitter, one in five nodes was found to suffer from uncertainty. The effects of this uncertainty for cascade prediction and seed selection [15, 28] could be dire due to the sensitivity of cascades to even very small gaps in the distribution of critical values, as demonstrated by Granovetter [24]. Even a seemingly trivial over-estimation of the critical value of a single node in a large population could make the difference between predicting early cascade failure and global adoption.
Figure 1: Graphs over three time periods. Nodes are labeled with the critical exposure required to activate. Intervals below the node indicate the observed critical value interval for each node. A dash (“-”) means we have no observation for the bound in question. Transparent nodes are inactive, blue nodes are active and precisely measured, red nodes are active and imprecisely measured. A dashed border indicates a node updates (and potentially activates) in the given time period. We call “innovators” which activate with 0 active neighbors precisely measured. In order for a node with critical value $> 0$ to be precisely measured, its interval must contain a single point, which is then the critical value. 

A) Opacity creates problems even in two node graphs. The right node activates with exposure 1 yet has critical value 0. Note that simultaneous updating would allow precise measurement of both nodes. 

B) The EAA rule incorrectly records the left node’s critical value as 2, when it is in fact 1. Examining the critical value interval interval $[1, 2]$ indicates that the true critical value could be 1 or 2, and suggests that EAA may introduce error. Note that any update strategy for $B$ produces at least one imprecise measurement.
2 Findings

2.1 Terminology and notation

In a graph, each node $i$ has an activation status (also called adoption status) $y_i \in \{0, 1\}$, where 1 indicates active and 0 indicates inactive. We denote $i$’s exposure as $k_i$, or as $k_{ti}$ for exposure at time $t$. We assume that once a node has activated, it remains active. The critical value $h_i$ is the minimum exposure $k_i$ that will trigger $i$’s activation. For each $i$, cascades reveal a critical value interval $N_i$ which contains the true critical value $h_i$. An interval such as $N_i = [1, 3]$ indicates that $h_i \in \{1, 2, 3\}$. The maximum of $N_i$ is obtained from $i$’s exposure at activation. The minimum of $N_i$ is $i$’s exposure at the most recent update before activation, plus 1. In the case where $\max N_i \neq h_i$, there is measurement error denoted $m_i$.

We say a node “updates” when it checks neighbor activation statuses and makes an activation decision. Nodes are assumed to check all neighbor statuses when updating, which is important for establishing the correct lower bound on $N_i$.

To briefly recap our example above, assume $i$’s true critical value $h_i = 2$. Then, $i$ updates at $t$ but remains inactive with $k_{ti} = 0$, and updates and activates at some later $t'$ with $k_{t'i} = 3$. In this case, $N_i = [1, 3]$ and we know $h_i \in \{1, 2, 3\}$. We add one to $k_i$ because $i$ updated yet remained inactive. The lowest possible exposure that could trigger activation is therefore the next integer $k_{ti} + 1$. With the knowledge that $h_i = 2$, using the EAA rule and taking the maximum of $N_i$ results in measurement error $\max N_i - h_i = m_i = 1$.

Note that we do not assume a probabilistic [28] or threshold [46] model since our results are consistent with both. In the probabilistic model, the critical value is the first “coin flip” which comes up heads and activates the node. In the threshold model, the critical value is the threshold.

2.2 Precise measurement condition

A simple condition determines whether a node’s critical value is precisely measured, or whether the critical value is uncertain within an interval. This condition can be applied to cascades where both updates and activations are observed. If only activations are observed, it is impossible to obtain lower bounds on critical value intervals.

Consider a node $i$ which updates at $t$ and some later $t'$. Assume $i$ is inactive at $t$ ($y_{ti} = 0$) and active at $t'$ ($y_{t'i} = 1$), giving critical value interval $N_i = [k_{ti} + 1, k_{t'i}]$.

**Condition 1.** The critical value for $i$ is precisely measured if the upper and lower bounds of $N_i = [k_{ti} + 1, k_{t'i}]$ are equal, $k_{ti} + 1 = k_{t'i}$.

Intuitively, $i$ updates at $t$ but does not activate, and then updates at $t'$ and activates. In that time, $i$’s exposure has changed by exactly one. In this case, $i$’s critical value is $k_{t'i}$, which is the exposure at activation. Intervals that contain more than a
point indicate uncertainty about where the true critical value lies. Condition 1 thus provides a quantification of the measurement uncertainty in addition to identifying precisely measured nodes.

An edge case occurs for innovators, or nodes which activate with 0 exposure. If we assume that critical values cannot be negative, this provides a lower bound and innovators are precisely measured. However, if critical values can be negative (e.g. “super eager” to activate as found in latent dependent variable models [20]), then innovators are not precisely measured. In a given case, correctly measuring innovators relies on an assumption about the underlying distribution of critical values.

Empirically, nodes often update many times with equal exposure before activating. This corresponds to cases where factors aside from exposure affect activation. For instance a node may go from 0 exposure to 2 exposure, and have many updates at 2 exposure before adopting. In these cases, we focus on the [1, 2] interval rather than the [2, 2] interval. Even if other factors change to encourage activation at 2 exposure, uncertainty still exists whether those factors would have caused activation at 1 exposure as well.
Figure 2: The severity of the opacity problem in all connected small graphs that support diffusion. The x-axis contains all connected graphs with between 2 and 4 nodes. Each point represents a critical value assignment for the graph in question. The proportion of precisely measured nodes is plotted on the y-axis. To generate each point, 1000 cascades are simulated by having nodes update in random order and activate if exposure is greater than critical value. We apply Condition 1 to determine which nodes are precisely measured, assuming that 0 exposure nodes are. Transparent triangles indicate combinations where one or more nodes always suffer from opacity. Solid circles indicate combinations where it is possible for all nodes to be precisely measured. In 86/115 = 75% of these cases, measurement uncertainty due to opacity is inevitable.
2.3 Distributional consequences of the EAA rule

The two main approaches to studying social cascades theoretically are probabilistic [26, 28, 34] and threshold [24, 46, 13, 12] approaches. Probabilistic approaches treat each active neighbor as having an independent chance to activate ego. Threshold approaches encode an assumption that nodes require a specific level of reinforcement to activate.

In both cases, there is a distribution of critical values that is of interest to researchers. In the probabilistic case, understanding this distribution allows estimating the transmissibility factor of the cascade. In the threshold case, this distribution is the distribution of thresholds.

While the substantive interpretations differ, opacity creates similar issues for recovering true critical value distributions in both cases. Figure 3 shows consequences of using the EAA rule for probabilistic (A,B), integer threshold (C,D), and fractional threshold (E,F) models. Integer threshold models assume only active neighbors have influence, while fractional threshold models assume that both active and inactive neighbors have influence [13].

In five out of six cases, the EAA rule does not recover the basic critical value distribution shape. For probabilistic models (A,B), the true distribution is geometric, yet the EAA distribution looks vaguely normal. In the fractional threshold cases (E,F), the measurements produced by the EAA rule seem to have no relationship to the underlying distribution. In the case where all nodes have fractional threshold 0.2 except for 5% seed nodes (F), the EAA rule records 10% of nodes as having “threshold” of 1.0. In the one case where the EAA rule roughly replicates the true distribution shape (normal distribution in panel D), it creates a long tail of “high threshold” nodes not present in the true distribution.

These results indicate that opacity presents substantial challenges for estimating the effects of peer influence. Regression-based approaches often use the count or proportion of active neighbors as an independent variable, which does not reliably replicate true node sensitivity to peer behavior (see SI for further discussion of the implications of opacity for regression models).

2.4 Hashtag cascades

Social media provides an ideal setting to examine the empirical implications of the opacity problem. This section demonstrates that opacity can qualitatively affect conclusions about whether and to what extent social influence/reinforcement facilitates diffusion. For about 2 in 5, different responses to opacity yield opposite conclusions about the effectiveness of social reinforcement.

The effects of peer influence can be plotted using $p(k)$ curves [18, 37, 31] of the probability of first activation with exposure $k$, given ever being $k$-exposed. Work using this methodology has found that higher exposure levels substantially increase the probability of hashtag adoption, particularly for controversial issues like politics.
More recent work has found that hashtags spread more like simple contagion where nodes are immunized after the first exposure [31]. Since complex contagions have different cascade dynamics than simple contagions [12], understanding the effects of reinforcement is consequential for predictions and interventions in online networks.

To that end, we tracked cascades for 50 hashtags (see SI for list and selection procedure) with between 40,000 and 360,000 unique adopters and 41,000 to 1.3 million total usages per tag. A total of 3.2 million users (“active users”) tweeted using one or more of these hashtags. These active users are connected by 45 million bidirected @mention links, which is a common measure of ties on Twitter [37]. An additional 105 million bidirected @mention edges connect active users to inactive users. In these data, active users tweeted a total of 7 billion times.

2.4.1 Computing critical value intervals

To construct critical value intervals for all users, we assume updates (checking alter status) correspond with tweet times. A user is engaged with the platform when tweeting, and may check the tweets of friends in the seconds before or after sending a tweet. A user’s Twitter timeline therefore provides an update record. For each node \( i \) first adopting each hashtag \( f \), there is a first usage time denoted \( t^f_i \). The time of tweets prior to \( t^f_i \) are denoted in descending order as \( t^f_i - 1, t^f_i - 2, \text{ etc.} \) where \( t^f_i - 1 \) is the time of the tweet immediately prior to \( i \)’s first usage of \( f \).

Iterating through pairs of timestamps in reverse chronological order gives descending time intervals \( (t^f_i - 1, t^f_i) \), \( (t^f_i - 2, t^f_i - 1) \), etc. These time intervals correspond to periods of time when \( i \) was “not looking”, bookended by updates. We find the most recent time interval where at least one neighbor first used \( f \), denoted \( T^* \), and count the total number of neighbor’s first usages of \( f \) in \( T^* \). If exactly one neighbor adopted in \( T^* \), then by Condition 1, \( i \)’s critical value for using \( f \) is precisely measured. If more then one neighbor activated in \( T^* \), then \( i \) has an interval of size larger than one. For instance if \( i \) had 2 neighbors activate in \( T^* \) and 4 neighbors activate before \( T^* \), then \( i \)’s critical value interval would be \([5, 6] \), with a width of two.

2.4.2 Empirical measurement rates

Measurement uncertainty affects 21% of activations, although there is substantial variation both by hashtag and exposure level. Note that we exclude critical value 0 nodes from this analysis because measuring these nodes precisely relies on an assumption about the critical value distribution. Nodes which adopt with higher exposure are more likely to suffer from measurement uncertainty (Figure 4). Among nodes that activate with exposure 1, 90.9% are measured precisely. Among nodes that activate with exposure 2, only 65.7% are precisely measured.
Variation also exists in measurement rates between tags. Among nodes that activate with exposure 2, 73.6% are precisely measured among the 13 easiest-to-measure tags, while only 49.1% are precisely measured among the 13 hardest-to-measure tags.

### 2.4.3 Social reinforcement and p(k) curves

Based on critical value intervals for each node, we construct “lower” and “upper” $p(k)$ curves by taking the minima and maxima of intervals, respectively. The lower curve $p_L(k)$ assumes that the EAA rule is maximally wrong. While taking interval minima may at first seem like an unrealistic assumption, it avoids introducing a long-tailed error into the data and can therefore have lower error than maxima in some cases. The upper curve $p_U(k)$ is equivalent to using the EAA rule as applied in existing research.

For some tags, upper and lower curves show similar patterns (Figure 5, red lines), while for others upper and lower curves diverge sharply (blue lines). For nearly all hashtags, the upper curve has $p_U(2) > p_U(1)$, corresponding to social reinforcement increasing adoption probability. For 19 out of 50 hashtags, the lower curve reverses this pattern, $p_L(1) > p_L(2)$, suggesting that these tags do not benefit from reinforcement. We refer to these tags as having a “drop” for the lower curve.

Examining characteristics of cascades, we find that tags with high clustering coefficients among adopters tended to have a drop (Figure 5B). Other factors, such as the overall measurement rate of the hashtag or the temporal concentration of adoptions (Gini coefficient of adoptions across days) were poor predictors of a drop.

Taken together, these results indicate that the opacity problem creates more uncertainty about cascade dynamics when adopters are clustered in dense networks. Two similarly plausible assumptions for dealing with opacity produce opposite qualitative results about the effects of social reinforcement for nearly 2 in 5 hashtags. This finding paradoxically makes it particularly difficult to determine the effects of social reinforcement when adopters have high clustering, even though contagion which requires social reinforcement is most likely to spread in clustered networks.
Figure 3: True individual critical values (blue) compared to measurements using the exposure-at-activation (EAA) rule (red). Graphs contain 1000 nodes, edges are introduced using a Barabási-Albert procedure with each node having 6 links. Five percent of nodes are chosen randomly as seed nodes. Each plot averages over 100 runs, critical values over 10 are set to 10. The top row shows results for the independent cascade model using “push” (A) and “random update” (B) strategies. The middle row shows results for an integer threshold model using an exponential distribution with $1/\lambda = \beta = 3$ (C), and a normal distribution with $\mu = 5, \sigma = 1$ (D). The bottom row shows a fractional threshold model using an exponential distribution with $1/\lambda = \beta = 3$ normalized to the $[0, 1]$ interval (E), and a uniform threshold of 0.2 for all nodes except seed nodes (F).
Figure 4: Precise measurement rates for cascades of 50 hashtags on Twitter among 3.2m unique users. We assume that nodes update when tweeting, which provides critical value intervals for each activation. Applying condition 1 allows determining which nodes are precisely measured. Plotting the proportion of precisely measured nodes at each exposure shows that measurement is more difficult when nodes activate with higher exposure, creating problems for understanding the effects of social reinforcement. Average precise measurement rates drop from around 90% at exposure 1 to slightly under 70% at exposure 2. Among the quarter of most difficult to measure tags (blue line), measurement rates are about 50% at exposure 2.
Figure 5: Different responses to the opacity problem produce different conclusions about the importance of peer reinforcement. We visualize this with $p(k)$ curves, which plot the probability of first activation at $k$-exposure given ever being $k$-exposed. Solid lines indicate the upper curve $p_U(k)$, which uses critical value interval maxima (equivalent to EAA rule), while dashed lines indicate the lower curve $p_L(k)$ which uses interval minima. A) shows 19/50 hashtags (blue, dashed line) have lower curves with a “drop”, indicating that exposures after the first do not facilitate adoption. Upper curves for these tags follow the usual pattern of increasing adoption probability with additional exposure. B) shows the “drop” pattern can be replicated by taking the quarter of tags where adopters have highest clustering (blue, dashed line).
2.5 Reducing measurement error with predictive models

In empirical cases, many nodes have uncertain critical values. Simulation provides a way to quantify measurement error since critical values are known. Using simulated cascades with integer thresholds on random graphs, we find that measurement error is substantial in a variety of cases, with the EAA rule having an average root-mean-squared error (RMSE) of 3.1 to 8.1 times baseline in cases we study. We simulate a scenario where node-level characteristics contribute to the node threshold. Training a predictive model on the precisely measured nodes using this information reduces this error to 1.15 to 1.35 times baseline.

2.5.1 Simulation details

Integer thresholds are generated by the function \( h_i = 5 + 3x_i + \epsilon_i \), where \( x_i, \epsilon_i \sim \mathcal{N}(0, 1) \). The covariate \( x_i \) represents some feature of \( i \) (e.g., age, wealth) that provides information about how much social reinforcement \( i \) needs before activation. The error variance represents the inherent unpredictability in thresholds and provides a natural baseline for comparing error-reduction methods. An unbiased ordinary least squares (OLS) model trained on the true data will have a root-mean-square error (RMSE) equal to the error variance, which here is set to 1.

Graphs are generated using a power law with clustering algorithm [27] with clustering parameter 0.1. Graphs have 1000 nodes and mean degree varies between 12, 16, and 20, with 1000 replications for each parameter set. Power law with clustering graphs exhibit high heterogeneity of node degree found in online settings, while also building in clustering that is typical of social networks. See the SI for a description of cascade dynamics.

2.5.2 Simulation results

As graph mean degree increases, more nodes activate but fewer are precisely measured (Table 1). For a fixed integer threshold distribution, adding links simulates a more “viral” cascade which spreads further in less time. When cascades spread quickly relative to node update pace, there is a higher chance that a node is “not looking” while multiple neighbors adopt, leading to large threshold intervals and

| Mean degree | Nodes | Activations | Precisely measured |
|-------------|-------|-------------|--------------------|
| 12          | 1000  | 720.7       | 113.4              |
| 16          | 1000  | 924.1       | 43.0               |
| 20          | 1000  | 983.1       | 13.0               |

Table 1: Descriptive statistics on simulations, averaged over 1000 runs. The average number of activated nodes increases as degree increases. However, it becomes more difficult to precisely measure nodes in higher mean degree graphs.
increased measurement uncertainty.

Using the EAA rule produces a substantially different distribution from the true threshold distribution, including introducing a long right tail. The maximum threshold in the true distribution is around 20, but the EAA rule records many nodes with “thresholds” of over 30.

Despite the small number of precisely measured nodes, they have an important advantage: we are certain their threshold is measured without error. This allows estimating a model relating $h_i$ to node characteristic $x_i$. In these simulations, $E[\epsilon_i]$ for precisely measured nodes is often slightly less than 0, indicating selection on the error due to lower threshold nodes being more likely to be measured precisely. Nevertheless, models estimated with the precisely measured nodes (Figure 6, blue) reliably outperform two natural alternatives: using the EAA rule directly (red) and estimating a model to predict thresholds using all active nodes (green). Compared to a baseline RMSE of 1, a model using the precisely measured nodes increases error by 14% to 35%, while using all active nodes increases error by 203% to 507%, and the EAA rule increases error by 310% to 807%. The low-error predictions using the precisely measured nodes are obtained despite using, on average, only 13 to 113 observations. The other two approaches use, on average, between 721 and 983 observations.
Figure 6: A model trained on the precisely measured subset (under 15% of the data) predicts thresholds for all nodes with low error (blue), while the EAA rule produces the highest error (red). The dashed black line is the baseline error of the true model for predicting thresholds equal to RMSE 1. When threshold interval maxima are substituted for thresholds (EAA rule), error rates are 3.1 (mean degree 12) to 8.1 (mean degree 20) times baseline. Taking the intermediate case of mean degree 16 graphs, the EAA rule has RMSE 5.3 for predicting true thresholds, which is larger than the mean of the true threshold distribution 5. Since the EAA rule takes maxima, all error results in over-estimating thresholds. For mean degree 16 graphs, modeling with the correctly measured subset (43 nodes on average) reduces RMSE from 5.3 to 1.25.

3 Conclusion

Studying the opacity problem highlights the dangers of using exposure at activation to measure the critical level of social influence needed to trigger activation. To our knowledge, we are the first to recognize the potential danger of this problem for research on social influence and network cascades. Even under optimal conditions, the EAA rule inflates measurements of critical values, leading to over-estimation of influence and incorrectly diagnosing as social contagions what are actually social allergens.

Although opacity presents substantial challenges, it can be addressed when researchers have access to node update behavior. In these cases, the magnitude of the opacity problem can be quantified at the node level using Condition 1. When covariates exist to predict uncertain critical values, the error introduced by opacity
can be reduced dramatically.

Employing this response to opacity has the potential to yield more precise answers to important questions. For instance, models of node critical values can be used to better study complex contagion [13] empirically. Results from such an analysis can then be applied to problems such as influence maximization [28, 29].

Our approach also facilitates refinements of the cascade prediction problem [15], where a model can be trained on active nodes while the contagion is still unfolding. Such a model can be used to identify susceptible nodes on the cascade fringe.

Additional work is needed to refine the method we propose for reducing measurement error. We consider only precisely measured critical values, yet there are many critical values that fall within a narrow interval which may provide useful information. Future work should address how to better incorporate this information, as well as investigating the ability to predict critical values in different domains.

4 Acknowledgements

We thank David Strang, Thomas Davidson, participants of INSNA Sunbelt 2016, and members of the Social Dynamics Lab for helpful discussions and comments. This research has been supported in part by National Science Foundation grant SES-1357442 and the Department of Defense Minerva Initiative grant FA9550-15-1-0036.

5 Appendix

5.1 Simulation details

We simulate cascades using the NetworkX package in Python 3 [25]. All code is available at https://github.com/georgeberry/thresholds. Statistical analyses are done using Scikit Learn [35] in Python 3, and in R.

We use 4 different graph generation routines at various points: Barabási-Albert [7], power-law with clustering [27], Watts-Strogatz [47], and an “atlas” of all small graphs [36]. All four are built-in to the NetworkX package.

5.1.1 Small graphs

NetworkX provides a full enumeration of all small graphs (“graph atlas”). We take all such graphs with between two and four nodes. Then, we filter out graphs which have more than one component, giving a set of all connected graphs with between two and four vertices. Call this set of graphs $G$.

For each $g \in G$, we generate all possible critical value assignments given that the critical value is less than or equal to node degree. For node $i$ in graph $G$ with degree $d_i$, $i$ has the critical value set $H_i = \{h_i : h_i \leq d_i\}$. Taking the product of $H_i$ for all $i \in g$ gives the set of critical value assignments for the graph $H(g)$. 

18
For each critical value assignment, we simulate cascades by updating nodes randomly and activating nodes immediately if exposure is greater than critical value, $k_i \geq h_i$. We filter out graphs where at least one node never activates, which occurs when each inactive node has updated yet none activates.

For these set of “admitted” graphs $G^*$ where all nodes eventually activate, we simulate 100 cascades per graph. For each simulated cascade, we record exposure at each node update. This allows applying the exposure-at-activation (EAA) rule to determine if nodes are precisely measured or not. If node $i$ updates at $t$ and some later $t'$, and $k_i^t + 1 = k_i^{t'}$, then the node is precisely measured. If there is no lower bound (e.g. no update before activation) and the exposure at activation is greater than zero, we call the node imprecisely measured. Innovators which adopt with 0 active neighbors are considered precisely measured here. As we discuss in the main text, this assumption about innovators may not be appropriate in all cases.

5.1.2 Distributional simulations

We use a Barabási-Albert graph generation process for these simulations. Graphs have 1000 nodes and each node has 6 links to attach preferentially to other nodes. In each case 5% of nodes are “seed” nodes and are either randomly activated (in the probabilistic case) or have threshold of 0 (in the threshold case). Each simulation is repeated 100 times.

For independent cascade model runs, we use two different methods to spread the cascade, which we call “push” and “random updates”.

The “push” version randomly selects an active node and iterates through its neighbors. Each neighbor flips a coin that comes up heads with probability $p$ and activates immediately if the coin comes up heads. When a node flips a coin, we consider it to be updating and so record its exposure. We also record the number of coins it has flipped. For instance, if a node flips a coin that comes up tails and then flips a coin that comes up heads, it has critical value two. At activation time, however, it may have more than two active neighbors.

The push model has a quirk not found in other models studied: exposure at the most recent update before activation may not be a lower bound of the true critical value. This occurs because nodes are “compelled” to update by an active node but effectively ignore other active neighbors. The EAA rule still over-estimates critical values, however.

The “random updates” model works similarly to the threshold model. Nodes randomly update and flip a number of coins equal to the number of neighbors that have activated since last update time. Exposure is recorded at each update time, plus the outcome of the coin flips. The first “heads” flip is the critical value, and the node activates immediately.

Integer and fractional threshold simulations work by randomly updating nodes and checking if exposure is greater than threshold. Nodes activate immediately if
the threshold is crossed. For plotting, fractional thresholds are binned to the nearest 0.05.

5.1.3 Error reduction simulations

Cascade dynamics are simulated by updating inactive nodes in random order, with nodes activating immediately if exposure is greater than or equal to the threshold, \( k_i \geq h_i \). Nodes record exposure before and after activation, allowing identification of precisely measured nodes by applying Condition 1. Cascades run until either all nodes have activated, or each inactive node has updated without activating. This process produces graphs where some nodes remain inactive (nodes with threshold greater than degree), which replicates the common empirical setting where not everyone adopts.

5.2 Twitter analysis

The Twitter data used for this analysis was collected for another project supported by an NSF grant (SES 1226483) between November 2013 and October 2014. Tweets were localized to a country using the method described in [17]. Users in Anglophone countries (US, UK, CA, AU, NZ, SG) were extracted and retweets were filtered out.

We selected users who had used one or more of 50 hashtags with moderate-to-high usage. We extracted the entire timeline for these users.

5.2.1 Hashtag selection

Hashtags were selected for first occurring between 2012 and 2014. We manually examined the top 1000 hashtags by usage among the Anglophone users and selected tags that we believed related to specific events or social media phenomena. Two project members independently examined the top 1000 tags, nominated tags, and then discussed disagreements.

The full list of (lowercased) tags is: obama2012, election2012, kony2012, romney, rippaulwalker, teamobama, bringbackourgirls, trayvonmartin, hurricanesandy, cantbreathe, miley, olympics2014, prayfornewtown, goodbyebreakingbad, governmentshutdown, riprobinwilliams, romneyryan2012, harlemshake, euro2012, marriageequality, benghazi, debate2012, newtown, linsanity, zimmerman, betawards2014, justicefortrayvon, samelove, worldcupfinal, prayersforboston, nobama, ferguson, springbreak2014, drawsomething, nfldraft2014, romney2012, snowden, replaceashowtitlewithwerk, inaugh2013, ivoted, trayvon, ios6, voteobama, jodiarias, windows8, mentionsomebodyyourethankfulfor, sharknado2, gop2012, whatdoesthefoxxsay, firstvine.
5.3 Effects of opacity and the EAA rule on estimates of social influence

Our results indicate that models of social influence and peer effects suffer from substantial error due to opacity and the EAA rule.

Here we present the particular problem the EAA rule causes when estimating parametric models of peer effects commonly found in the social sciences. Importantly, even though opacity and the EAA rule cause critical values to be over-estimated, this does not necessarily imply a particular type of bias for regression models. Estimates of the effect of social influence may be too large or too small, depending on the particular form bias takes.

This occurs because of two different biases introduced by opacity + the EAA rule: 1) the weak influence effect where nodes which would have activated at low exposure are instead measured as requiring high exposure; and 2) the phantom influence effect where nodes which would have activated with no active neighbors are instead measured as requiring some amount of social influence. The weak influence effect makes influence appear to be weaker than it really is (although, as we show below, it can increase peer effect model coefficients), while the phantom influence effect creates the illusion of influence where none exists. The worst case for the weak influence effect occurs when a node which requires one neighbor to activate is instead measured as requiring all neighbors to activate. The worst case for the phantom influence effect occurs when a node which requires zero neighbors to activate is instead measured as requiring one neighbor to activate.

The particular combination of these two effects may be specific to the dataset. To better understand these effects, we show how measurement error impacts nonparametric probability estimates of activation at some exposure level, and then consider how measurement error affects a simple regression model of social influence similar to models found in the literature.

Assume there is a binary outcome $y_i \in \{0, 1\}$ and call $k_i$ the exposure of $i$. $i$'s exposure at activation $k^*_i$ may contain error: $k^*_i = h_i + m_i$, where $h_i$ denotes the true critical value and $m_i$ denotes error. If we assume that nodes activate immediately, the error affects the data in a peculiar way: active nodes have measurement error since the EAA rule produces estimated critical values above the true critical value, but the exposure measurements for inactive nodes do not have error.

To see this more clearly, consider the effect of $m_i$ on estimating adoption probability at some particular exposure level $k_i = k$. The true adoption probability at this $k$ is

$$P_{true}(y_i = 1 | k_i = k) = \frac{\#(h_i = k)}{\#(k_i = k)}$$  

(1)

where $\#(h_i = k)$ denotes the count of nodes whose true critical value $h_i$ is $k$ and $\#(k_i = k)$ denotes the count of nodes who were ever $k$-exposed. With measurement
error, this becomes

\[ P_{err}(y_i = 1|k_i = k) = \frac{\#(h_i + m_i = k)}{\#(h_i + m_i = k) + \#(k_i = k|y_i = 0)} \] (2)

where \( \#(h_i + m_i = k) \) indicates the count of nodes whose true critical value \( h_i \) plus measurement error \( m_i \) is \( k \), and \( \#(k_i = k|y_i = 0) \) indicates the count of nodes which were \( k \)-exposed and remained inactive. This error means that for some low \( k \) values, the estimated probability will be too small, while for some large \( k \) the estimated probability will be too large.

We can write down some properties of \( m_i \) with this in mind: \( m_i \geq 0 \forall i \), \( E[m_i|h_i] > 0 \), and \( m_i \) may be correlated with \( h_i \). Additionally, \( E[m_i|h_i, y_i = 0] = 0 \) while \( E[m_i|h_i, y_i = 1] > 0 \). In other words, there is measurement error for active nodes but not for inactive nodes, on the condition that we assume immediate activation.

We can examine further examine the impacts of measurement error on a simple regression model similar to those used in empirical work. Assume that each of \( i \)'s active neighbors increases \( i \)'s probability to adopt by \( \beta \). If \( k_i \) is \( i \)'s exposure and \( \epsilon_i \) is idiosyncratic error, we can then estimate a linear probability model of form

\[ y_i = \alpha + \beta k_i + \epsilon_i. \] (3)

First, we examine the weak influence effect. We choose \( k_i \sim \text{Pois}(4) \) and \( \beta = 0.1 \). Since observations with \( k_i \geq 10 \) are active with probability 1, the range of \( k_i \) is restricted to \([0, 10]\). Three different types of error are added to the active nodes only: 1) exponentially distributed error with \( \lambda = 1/2 \); 2) normally distributed error \( \mathcal{N}(0, 2) \) if the realization of the random variable is greater than zero; 3) binomial error where 3 is added to the observation with probability 1/2.
Figure 7: Predicted values from OLS (A) and logistic (B) models estimated on true data (purple) and data with three types of error. Data is generated by setting nodes to active according to function $y_i = 0.1k_i$, indicating zero probability of adoption at exposure 0 and unity probability of adopting at exposure 10. Focusing on the OLS results (A, pattern similar for logistic), we see that the true slope (purple) is flatter than when models are estimated on data with error, while the true y-intercept is larger than in models estimated on data with error. Data with error causes models to overstate the importance of influence (steeper slope). This happens paradoxically because error causes adoption probability at each exposure (points) to be underestimated at low exposure values and over-estimated at high exposure values. The regression line is therefore steepened and moved to the right.
While nonparametric estimates of activation probability at \( k \) give the result above (underestimates at some low \( k \), overestimates at some high \( k \)), effects on estimated \( \hat{\beta} \) are somewhat more complicated. The types of error we consider combined with our assumed distribution of \( k_i \) lead to \( \hat{\beta} > \beta \) while \( \hat{\alpha} < \alpha \). This is counterintuitive: since we under-estimate the effectiveness of social influence at low levels, we over-estimate the social influence model coefficient. The reason for this can be seen clearly in Figure 7: models on data with error suggest that higher exposures are disproportionately effective at encouraging activation. This increases the slope of \( \hat{\beta} \) while shifting the entire line to the right, leading to a low \( \hat{\alpha} \). This is present in both the linear probability model and fitting the same data with a logistic model.

This result shows that the weak influence effect has two consequences: 1) underestimating activation probability at low \( k \) while over-estimating it at high \( k \); and 2) introducing bias into \( \hat{\beta} \), which can paradoxically make the coefficient for social influence larger. Note that different assumptions about \( k_i \) (e.g. uniformly distributed) can reverse this effect and cause \( \hat{\beta} \) to be too small.

The results of the phantom influence effect are more straightforward and are shown in Table 2. We set \( k_i = 0 \) for all \( i \) (therefore \( \beta = 0 \)), while retaining the same \( y \) distribution as above. This means that the the constant \( \alpha = 0.4 \). We add the same error to the active observations as above. As predicted, when adding error, we find an effect of influence much larger than zero.

| Dependent variable: | True | Exp error | Norm > 0 error | Binom error |
|---------------------|------|-----------|----------------|-------------|
| \( \alpha \)        | 0.398*** (0.388, 0.407) | 0.196*** (0.189, 0.204) | 0.289*** (0.280, 0.298) | 0.247*** (0.239, 0.256) |
| \( \beta \)         | 0    | 0.199*** (0.195, 0.203) | 0.259*** (0.251, 0.268) | 0.251*** (0.245, 0.257) |

This demonstration is not meant to be realistic but rather to give an idea of the effects of the opacity problem plus the EAA rule on estimated coefficients. Where possible, it is helpful to examine estimates of activation probabilities at various exposure levels directly. Both weak influence and phantom influence effects are likely present in empirical data.

Table 2: Phantom influence effect
5.4 When is opacity not a problem?

The main text focuses on the case where there are no activation delays, and where the researcher does not control node update ordering. In this case, opacity is a common problem even in small graphs.

We can imagine other such processes:

1. Researcher control over node update ordering
2. Researcher control over node activation delays
3. Researcher can observe the activation delay
4. Researcher can observe when a node’s critical value is exceeded before public activation has taken place

We assume that in all cases, researchers do not control graph structure or the node critical values themselves. We present two results in this section: first, a restatement in greater detail of the finding in the main text that opacity is a problem in the case normally studied by social scientists. Second, we present a case where opacity is not a problem. This requires the researcher to have control over node updates, node activation delays, and to observe when a node’s critical value is exceeded before public activation occurs. This case does not occur in any domain we are aware of, but we present it to give an idea of the challenges posed by the opacity problem.

5.4.1 Notation

Individual nodes in a graph are indexed with $i$, and time periods are indexed with $t$. We use subscripts for individuals $i$ and superscripts for times $t$.

- $G = (V, E)$—the graph structure composed of sets of nodes $V$ and edges $E$. Vertices are labeled $\{1, 2, 3, \ldots, |V|\}$.
- $H$—the vector of length $|V|$ containing critical values for each vertex, with a value for $i$ denoted $h_i$.
- $U$—the node update ordering. $u^t$, specifies the set of nodes that update simultaneously at time $t$. $u_i$ denotes the time periods where node $i$ updates.
- $\Delta$—the vector of length $|V|$ containing node activation delays, with $\delta_i$ corresponding to the activation delay for node $i$.
- $a^t$—the vector of length $|V|$ indicating which nodes have publicly activated at time $t$. If $i$ has publicly activated at $t$, then we write $a^t_i = 1$, else $a^t_i = 0$. 
• $s^t$—the vector of length $|V|$ indicating which node critical values have been met at time $t$, but that have not necessarily publicly activated due to an activation delay. We write $s^t_i = 1$ if $i$’s critical value has been exceeded at $t$, otherwise $s^t_i = 0$.

• $k^t$—the vector of length $|V|$ indicating each node’s exposure at time $t$. In other words, how many active neighbors does $i$ have at time $t$.

5.4.2 Assumptions

Assumption 1. **Granularity**: The researcher sees every state the contagion process takes on.

Assumption 2. **Saturation**: All nodes in $V$ will activate eventually if they continue to update.

Assumption 3. **No updates during activation delays**: Nodes do not continue to update while already satisfied and waiting to activate.

5.4.3 Definitions

Definition 1. An element is **exogenous** when the researcher does not have control over it. $G$ and $H$ are always exogenous, while $U$ and $\Delta$ may or may not be exogenous.

Definition 2. Node $i$’s critical value is **precisely measured** if, for some $t, t' \in u_i, t' > t$, we have $k^t_{i} - k^{t'}_{i} = 1$ and $a^{t'}_{i} = 1$ while $a^t_{i} = 0$. In this case, the node’s critical value $h_i = k^t_{i}$.

Note: If we can observe $s^t$, simply replace references to $a^t$ with $s^t$ throughout. This condition is similar to the equivalent condition in the main text.

This node-level condition allows formulating a straightforward process-level condition.

Definition 3. A contagion process is **measurable** if we are guaranteed to precisely measure all critical values, regardless of the values of the exogenous elements.

In other words, if $(G, H, U, \Delta)$ are exogenous, then we must show that regardless of the values of these elements, all nodes are guaranteed to be correctly measured. To prove it is not measurable, we need only provide a counterexample. Note that this means that some $(G, H, U, \Delta)$ may produce all correctly measured nodes, but that this will not be certain.

5.4.4 Theorems

Theorem 1. If $(G, U, H, \Delta)$ are exogenous, the process is not measurable, even if we can observe $\Delta$ and $s^t$. 

26
Proof. Since \((G, U, H, \Delta)\) are exogenous, we can pick any values for the counterexample. Assume that \(V = (i, j, k)\), \(E = ((i, j), (j, k), (k, i))\), and \(h_i = h_j = 0, h_k = 1\). Let \(U = (\{k\}, \{i, j\}, \{k\})\) so that \(k\) updates at time 1, \(i\) and \(j\) update at time 2, and \(k\) updates again at time 3. Let \(\Delta = (0, 0, 0)\), indicating no lags in public activation.

Then node \(k\)'s critical value can be 1 or 2. This process can be seen visually in Figure 1-B.

\[\square\]

**Theorem 2.** A contagion process is measurable if \((G, H)\) are exogenous, \((U, \Delta)\) are controlled by the researcher, and the researcher observes \(s^t\) and \(\Delta\).

Proof. Intuitively, we can choose node update ordering and activation delays such that at each even time period \(t_{\text{even}} = \{0, 2, 4, \ldots\}\) all nodes update (except those in the queue \(z\) defined below), and at each odd time period \(t_{\text{odd}} = \{1, 3, 5, \ldots\}\) exactly one node publicly activates. We use private satisfaction status \(s^t\) to determine when individual node critical values are exceeded apart from public activation \(a^t\).

There is a “cold start” problem here, so we can randomly update nodes one at a time until one activates at critical value zero. If the cascade stalls we repeat this cold start process, and by Assumption 2 above we’re confident that the cascade will start again.

Let \(z\) be a first-in-first-out queue of satisfied but unactivated nodes, where \(s_i^t = 1\) but \(a_i^t = 0\) for all \(i \in z\). We add node \(i\) to the end of \(z\) if, at some time \(t \in t_{\text{even}}\), \(i\) updates and \(s_i^t = s_i^{t-2} = 0\). At each odd time period, we remove the first element of \(z\) and activate it publicly. This defines an update ordering \(U\) where each node updates simultaneously at each time in \(t_{\text{even}}\) (unless it is already in \(z\)). It also defines a \(\Delta\) where exactly one node activates publicly at each \(t_{\text{odd}}\).

To see that such a process is measurable, consider any \(i \in V\). Since \(i\) updates at every time \(t = \{0, 2, 4, 6, \ldots\}\) until it is put in the queue \(z\), and we see \(s_i^t\) at each \(t\), find \(t_i^* = \min(t) : s_i^t = 1\). By construction, exactly one node activated publicly at \(t_i^* - 1\), which implies that \(i\)'s true critical value is \(h_i = k_i^{t_i^*}\), \(i\)'s exposure at \(t_i^*\).

\[\square\]

We require the ability to construct \(U\) and \(\Delta\) as the process is unfolding. This type of centralized control does not exist in social networks, which are decentralized by definition.

We note that there are likely other constructions of \(U\) and \(\Delta\) that satisfy the measurability condition. And, as we noted above, there may be other contagion processes not discussed here that allow measurability.

27
References

[1] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[2] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[3] Susan Athey, Dean Eckles, and Guido Imbens. Exact P-values for Network Interference. *arXiv:1506.02084v1*, (June), 2015.

[4] Lars Backstrom. Group formation in large social networks: membership, growth, and evolution. *KDD*, pages 44–54, 2006.

[5] Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. Social Influence in Social Advertising: Evidence from Field Experiments. In *EC ’12*, pages 146–161, 2012.

[6] Abhijit Banerjee, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. The diffusion of microfinance. *Science*, 341(July):1236498, 2013.

[7] Albert-Laszlo Barabasi and Reka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, October 1999.

[8] Vladimir Barash, Christopher Cameron, and Michael Macy. Critical phenomena in complex contagions. *Social Networks*, 34(4):451–461, October 2012.

[9] Javier Borge-Holthoefer, Raquel A. Banos, Sandra Gonzalez-Bailon, and Yamir Moreno. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, 1(1):3–24, June 2013.

[10] Yann Bramoulle, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, May 2009.

[11] Ellsworth Campbell and Marcel Salath. Complex social contagion makes networks more vulnerable to disease outbreaks. *Scientific Reports*, 3:1905, 2013.

[12] Damon Centola, Victor M. Eguiluz, and Michael W. Macy. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, 374(1):449–456, 2007.

[13] Damon Centola and Michael Macy. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, 113(3):702–734, 2007.

[14] Damon M Centola. The Spread of Behavior in an Online Social Network Experiment. *Science*, 1194(September):1194–1198, 2010.
[15] Justin Cheng, Lada Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? WWW, pages 925–936, 2014.

[16] Nicholas A. Christakis and James H. Fowler. The Spread of Obesity in a Large Social Network Over 32 Years. New England Journal of Medicine, 357(4):370–9, 2007.

[17] Ryan Compton, David Jurgens, and David Allen. Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization. arXiv Preprint. arXiv1404.7152, 2014.

[18] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. KDD, page 160, 2008.

[19] Giacomo De Giorgi, Michele Pellizzari, and Silvia Redaelli. Identification of Social Interactions through Partially Overlapping Peer Groups. American Economic Journal, 2:241–275, 2010.

[20] Paul J. DiMaggio and Filiz Garip. Network Effects and Social Inequality. Annual Review of Sociology, 38(1):93–118, 2012.

[21] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. arXiv:1404.7530, page 29, 2014.

[22] Noah E. Friedkin and Eugene C. Johnsen. Social Influence and Opinions. Journal of Mathematical Sociology, 15(3-4):193–205, 1990.

[23] Sandra Gonzalez-Bailon, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The Dynamics of Protest Recruitment through an Online Network. Scientific Reports, 1(1), December 2011.

[24] Mark Granovetter. Threshold models of collective behavior. American Journal of Sociology, 83(6):1420–1443, 1978.

[25] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. SciPy, pages 11–16, 2008.

[26] Herbert W. Hethcote. The mathematics of infectious diseases. SIAM review, 42(4):599–653, 2000.

[27] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. Physical review E, 65(2):026107, 2002.

[28] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. KDD, page 137, 2003.
[29] David Kempe, Jon Kleinberg, and Eva Tardos. Influential Nodes in a Diffusion Model for Social Networks. *ICALP*, 3580:1127–1138, 2005.

[30] A.S. Klovdahl, J.J. Potterat, D.E. Woodhouse, J.B. Muth, S.Q. Muth, and W.W. Darrow. Social networks and infectious disease: The Colorado Springs study. *Social Science & Medicine*, 38(1):79–88, January 1994.

[31] Kristina Lerman. Information Is Not a Virus, and Other Consequences of Human Cognitive Limits. *Future Internet*, 8(2):21, May 2016.

[32] Michael W. Macy. Chains of Cooperation: Threshold Effects in Collective Action. *American Sociological Review*, 56(6):730–747, 1991.

[33] Charles F. Manski. Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies*, 60(3):531, 1993.

[34] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[35] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, VincentDubourg, and others. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[36] Ronald Read, C. and Robin J. Wilson. *An Atlas of Graphs*. Clarendon Press, 1998.

[37] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. *WWW*, pages 695–704, 2011.

[38] Greg Ver Steeg, Rumi Ghosh, and Kristina Lerman. What stops social epidemics? *arXiv preprint arXiv:1102.1985*, 2011.

[39] David Strang and Sarah A. Soule. Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills. *Annual Review of Sociology*, 24(1):265–290, 1998.

[40] Steven H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.

[41] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, April 2012.

[42] Thomas W. Valente. *Network Models of the Diffusion of Innovations*. Cresskill New Jersey Hampton Press, 1995.
[43] Thomas W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks, 18*(1):69–89, 1996.

[44] Thomas W. Valente. Network Interventions. *Science, 337*(6090):49–53, 2012.

[45] Susan van den Hof, Christine M. A. Meffre, Marina A. E. Conyn-van Spaendonck, Frits Woonink, Hester E. de Melker, and Rob S. van Binnendijk. Measles outbreak in a community with very low vaccine coverage, the Netherlands. *Emerging Infectious Diseases, 7*(3):593–597, 2001.

[46] Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences, 99*(9):5766–5771, 2002.

[47] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ’small-world’ networks. *Nature, 393*(6684):440–2, June 1998.

[48] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality Prediction and Community Structure in Social Networks. *Scientific Reports, 3*, August 2013.