Estimating Threshold Distributions From Observational Data

George E. Berry and Christopher J. Cameron

June 1, 2016

Abstract

We identify a property of contagion processes on networks that prevents the correct measurement of node activation thresholds for most individuals. Previous approaches have used the number of activated alters at ego's activation time, which we demonstrate creates a huge upward bias in the estimation of thresholds across a wide range of threshold distributions and network configurations. We formalize the elements of contagion processes on networks and present an analytic argument that specifies when node thresholds can be measured correctly. We show that if both pre- and post- activation exposure level can be measured for each node then nodes with incorrectly measured thresholds can be identified and excluded from analysis. Using simulation, we find that the rate of correct measurement is a function of the distribution of thresholds and the network structure. Finally, we consider the case where activation thresholds are a function of the observable node attributes and find that statistical models fit on correctly measured nodes can be used to estimate thresholds for nodes with incorrectly measured thresholds. When good predictors for thresholds exist, coefficient bias and prediction errors are modest. Coefficient bias is found to be primarily a function of the level of unexplained variance in the model of thresholds.

1 Introduction

Threshold approaches have long been central to understanding the diffusion of innovations, information cascades, and behavioral change in networks (Rogers 1962; Granovetter 1978; Valente 1996; Kempe, J. Kleinberg, and Tardos 2003; M. W. Macy 1990; Duncan J Watts 2002). When individuals in a social network face uncertain adoption decisions or when adoption decisions exhibit strategic complimentarities, the behavior of network neighbors can play a crucial role in decision making. Threshold approaches provide a useful abstraction to model these types of interdependent decisions: nodes require a certain number of network neighbors to adopt before themselves adopting. Threshold models can be used to explore how social structure and individual heterogeneity affect macro-level outcomes, such as the development of riots or the diffusion of innovations (Granovetter 1978; Strang and Soule 1998).

Despite wide applicability, threshold approaches have been largely confined to developing theoretical models of social cascades Few studies have attempted to empirically measure activation thresholds, with a few notable exceptions

(Valente 1995; Valente 1996; Ludemann 1999). To our knowledge, *statistical estimation* rather than *modeling* of thresholds as a function of observables has not been attempted, despite the intuitive appeal of such a method.

In this article, we demonstrate that existing approaches to *measuring* thresholds provide upwardly biased estimates, we develop a new measurement approach to address this bias, and we *model* thresholds as a function of observables to both vastly reduce this bias and explain adoption thresholds.

Our work is motivated by a desire to directly test theoretical propositions with data. Without treating thresholds as a subject for empirical investigation, we cannot examine important theories about the properties of social cascades. For instance, Centola and M. Macy 2007 propose the theory of *complex contagion*, where individuals require social reinforcement before adopting a behavior. If we can measure thresholds in an empirical cascade, we can determine its level of "complexity" and examine the path this contagion takes through the network. We can then see if behaviors that are theorized to spread as a complex contagion are indeed complex, and compare the diffusion process of behaviors of different complexity.

A second theoretical proposition that can only be addressed by measuring individual thresholds is Granovetter's (1978) result that the threshold distribution matters for the ultimate outcome of cascades. If we can measure thresholds of cascades that spread differently, we can use the differing distributions of thresholds as a factor explaining the behavior of the cascades. If we can accurately estimate the threshold distribution early in the cascade, we can predict cascade behaviors in a novel way. However, if we cannot measure thresholds and threshold distributions, we are unable to assess either the threshold distribution of a particular empirical contagion or the effect of the threshold distribution on cascade dynamics.

1.1 Threshold Definition

Thresholds have been defined as both the number (e.g. Granovetter 1978) and fraction (e.g. Valente 1996; Duncan J Watts 2002; Kempe, J. Kleinberg, and Tardos 2003; Kempe, J. Kleinberg, and Tardos 2005) of alters that are required to induce ego to activate. The difference between absolute and fractional thresholds encodes our assumption about the influence of non-adopters. In the absolute threshold case, only those alters adopting exert influence, whereas in the fractional threshold case, both adopters and non-adopters exert influence.

In this paper, we employ absolute rather than fractional thresholds for ease of presentation. However, results presented here apply to the fractional case as well, with suitable changes to the modeling procedure.

Throughout this paper we will use *ego* to refer to a focal node, and *alters* to refer to that focal node's network neighbors. Ego's *threshold* will be the minimum number of alters that need to adopt to induce ego to adopt the contagion. At a given point in time, the number of activated nodes among ego's alters is ego's *exposure*.

As an example, consider a doctor's decision to begin prescribing a new drug and assume the doctor relies on peer behavior to inform her own practice. A doctor may have a threshold of four, meaning that she needs four alters to adopt a drug before she finds it worthwhile to start prescribing the drug. At a given point in time, we measure the doctor and the activation status of her alters. If three of her alters are activated, we say that she has exposure 3 at this point in time. At this exposure, she is *unactivated*, meaning she has not adopted the innovation yet.

1.2 Existing Empirical Approaches

Despite the seeming simplicity of the task, we find that complex topologies of social networks prevent correct threshold measurement for a large fraction of nodes, even in ideal measurement circumstances. Previous work with observational data has generally assumed that the "adoption threshold is the exposure at time-of-adoption" (Valente 1996). We find that using this rule leads to upwardly biased estimates of adoption thresholds in almost any conceivable case [1], regardless of whether individuals update synchronously or asynchronously, whether individuals have a lag time between perception and adoption, or whether time is considered discrete or continuous. Work in computer science using the $p(k)$-curve method [2] suffers from a similar bias (Crandall et al. 2008; Backstrom 2006; Romero, Meeder, and J. Kleinberg 2011).

We find that this upward bias does not affect all nodes equally and a condition that allows correct measurement for a subset of nodes. To estimate thresholds for the nodes that we could not measure correctly, we employ a modeling approach. We find that in certain circumstances, this estimation procedure works extremely well. In several situations that we study here, using the exposure-at-activation-time rule for every node in the graph leads to larger levels of error than using only 5% of the nodes, but which are correctly measured.

Because thresholds determined by the exposure-at-activation-time rule are often upwardly biased, probabilistic or hazard rate models of adoption which incorporate the number of activated neighbors as an independent variable systematically underestimate the effect of peer influence on adoption decisions by using an upwardly-biased independent variable. By using our threshold estimation method before running such models, scholars interested in time-to-adoption or adoption probabilities can make their models more accurate.

2 Threshold Measurability

In this section, we present a formal argument that details the conditions under which all thresholds in a diffusion process can be correctly measured. We use the concept of a *diffusion process* to refer to the mechanics of diffusion, regardless of the specific network topology, node update order, or assignment of thresholds. This representation allows us to investigate which *types* of processes allow correct measurement of adoption thresholds. In practice, we do not usually have control over the graph structure, when nodes update, or the specific assignment of thresholds to nodes.

A diffusion process happens on a network $G = (V, E)$ where V represents the nodes and E represents the edges. Each node $i \in V$ has threshold $h_i \in \mathbb{R}$. e_i

[1] As shown below, to avoid bias one has to construct the cascade according to a stringent set of rules that are virtually impossible to satisfy in a real social network.

[2] Authors using this method do not call it a threshold-estimation method, but it serves a similar purpose: estimating the probability of adoption at a given exposure level. If this is the goal, the method provides a biased estimate of this probability for the same reason that existing approaches systematically overestimate adoption thresholds.

is a node's *exposure*, or its number of active neighbors. Each $i \in V$ has a public *activation status* $a_i \in \{0, 1\}$, where 1 indicates that i is publicly active. Each i also has a private *threshold satisfaction* status $s_i \in \{0, 1\}$. Whenever $e_i \geq h_i$, $s_i = 1$, indicating the threshold has been satisfied; however, public activation ($a_i = 1$) may not have taken place yet.

There is an *update ordering* C that specifies the set of nodes checked for threshold satisfaction at each update time[3]. We subscript quantities with t to indicate the value at a specific time.

A diffusion process is characterized by four features that affect threshold measurability:

1. Public versus private information—Diffusion processes may be *public information*, where threshold satisfaction s_i is only observed indirectly when nodes activate publicly ($a_i = 1 \implies s_i = 1$), or *private information* where threshold satisfaction s_i can be observed separate from public activation (e.g. we can observe $s_i = 1$ but $a_i = 0$).

2. Synchronous versus asynchronous updating—When more than one node updates in at least one update period, the process is *synchronous*, otherwise it is *asynchronous*. For instance $C = (\{i\}, \{j, k\})$ represents a synchronous updating process, since j and k update together at time 2.

3. Instantaneous versus delayed activation—Nodes may have an individual-specific delay δ_i between threshold satisfaction ($s_i = 1$) and public activation ($a_i = 1$). In cases where $\delta_i = 0$ for all i, then a diffusion process has *instantaneous activation*, whereas if $\delta_i > 0$ for at least one i, then we have *delayed activation*.

4. "Fast" versus "slow" updating—When every node has a chance to update between each public activation, we call updating "fast," otherwise it is slow. Note that fast updating requires that there be a lag time $\delta_i > 0$ for all nodes.

Now that we have characterized diffusion processes, we have the following definitions:

Definition 1 *A threshold for node i is* ***correctly measured*** *when, for i's update times $u = \{1, 2, 3, ...\}$, for some $t \in u$, we have $e_{i,t} - e_{i,t-1} = 1$ and $a_{i,t} = 1$ while $a_{i,t-1} = 0$.*

Definition 1 captures the idea that node i activates at some time t with $e_{i,t}$ active neighbors, while at i's last update, i was inactivate with exactly one less active neighbor. The final adopting neighbor can be said to trigger i's adoption in this case. This means that the *threshold interval* $R = [e_{i,t-1}, e_{i,t}]$ has size 1, and we can be certain of where i's threshold lies. If the size of R is greater than 1, we do not have a principled way of determining where the

[3] C is of the form $(\{i\}, \{j, k\})$, indicating that i updates at time 1 and j and k update at time 2. Note that this includes both discrete and continuous time processes. In a continuous time process, we draw a random variable for a node's next update time, which then gives us an ordering with one node updating per update time. In a discrete time process, we assign nodes to update at a certain time period, and all nodes updating at each time period are checked simultaneously.

most likely threshold in R without additional assumptions, which we leave for future research. Contrasting Definition 1 with prior research, the exposure-at-activation-time rule always takes the max of R, which leads to upwardly biased measurements.

Definition 2 *A diffusion process is* ***threshold measurable*** *if we can correctly measure thresholds for all nodes in the diffusion process according to Definition 1, regardless of the graph topology, node update order, or specific assignment of thresholds.*

The strict nature of Definition 2 reflects the parts of diffusion that we do not generally have control over: the network structure, the update behavior of nodes, and the assignment of thresholds to nodes. This definition also allows us to determine whether a process is *not* threshold measurable by providing a counterexample for that specific process. In creating this counterexample, we have control over the graph structure, the update ordering, and the threshold assignments.

2.1 An Example

To see this more intuitively, consider the following hypothetical laboratory experiment: we wish to see how many alter adoptions are required for an individual to commit to purchasing a product. We show her a computer screen indicating that zero friends have adopted and ask her whether she would like to purchase the product. We repeat this process for zero, one, two, and three neighbors. After the second alter activation, she indicates that she would purchase the product. Since we know she did not adopt with zero or one alters active, but did adopt when two alters were active, we correctly measure her adoption threshold as two.

Consider the case of a lazy research assistant, who forgets to check with ego after the first and second alters have adopted. The assistant checks with ego with zero active neighbors, and again when she has three active neighbors, where ego indicates that she would adopt. In this case, we do not have a measurement of ego's activation status with 1 and 2 active alters. Therefore, the most precise statement we can make based on this data is that her true threshold lies in the interval $(0, 3]$.

In other words, if we do not have a record of ego's activation status after some alter activations, we can fail to correctly measure ego's adoption threshold. This happens in instances where nodes check their alter status at intervals, so that we do not know the minimum exposure level that would induce ego to adopt.

2.2 Measurability in Diffusion Processes

Table 2.2 shows the eight types of diffusion processes, and indicates that only two out of the eight permit threshold measurability in certain circumstances.

To see why this is so, consider the hypothetical laboratory experiment above. In this experiment, we are able to check with the subject after each alter activation. In this case we apply Definition 1 and see that ego's threshold is correctly measured. In order to approximate this level of granularity with an empirical

Information	Update Process	Activation Delay	Fast Updating	Measurable
Public	Async	Yes	Yes	No
Public	Async	Yes	No	No
Public	Async	No	Yes	No
Public	Async	No	No	No
Public	Sync	Yes	Yes	No
Public	Sync	Yes	No	No
Public	Sync	No	Yes	No
Public	Sync	No	No	No
Private	Async	Yes	Yes	Yes
Private	Async	Yes	No	No
Private	Async	No	Yes	No
Private	Async	No	No	No
Private	Sync	Yes	Yes	Yes
Private	Sync	Yes	No	No
Private	Sync	No	Yes	No
Private	Sync	No	No	No

Table 1: Exhaustive list of the 8 possible diffusion processes. Only certain private information processes with an activation delay are threshold measurable. Gray indicates measurable.

social contagion, we require a measurement of node i unactivated with $h_i - 1$ active neighbors, and again with node i activated with h_i active neighbors. If this fails for at least one node, then the process will not be threshold measurable.

Since we do not know thresholds in advance, this implies that every node must update after each public activation, and we must be able to record s_i (private information) at these updates. It also implies that there must be an activation delay for all nodes, since if there were no delay, additional nodes could adopt during this update step, before we had recorded s_i for all nodes. This motivates the following condition, which must hold for the two diffusion processes identified above to be threshold measurable.

Condition 1 *Assume a node has just activated publicly at time t. Consider the public activation times of all nodes with satisfied thresholds ($s_i = 1$), S_t. Denote $\underline{s_t} = min(S_t)$ the next node scheduled to publicly activate.*

Denote U_t node update times at time t, with $\star u_t$ the time at which each node has updated at least once, starting at time t.

If $\underline{s_t} < \star u_t$, the condition fails, since a second node will activate before each node has had a chance to update, violating Definition 1.

If $\underline{s_t} > \star u_t$, then cycle through the update times up to $\star u_t$ and check each node for its threshold satisfaction status. If a node has $s_j = 1$, then add it to S_t according to its activation delay, δ_j. If such an $s_j < \underline{s_t}$, then the condition fails.

If all nodes with newly satisfied thresholds have $s_j > \bar{u}_t$, then the condition holds for this particular activation.

*If the condition holds for all activations in a diffusion process, then we say that the diffusion process satisfies **fast updating**.*

Condition 1 says that, after every activation, each node must get a chance to update before the next activation. And this must hold after each public activation.

From this discussion, we determine that the following conditions are needed for threshold measurability:

1. Private information—s_i can be observed for all nodes immediately
2. Activation delay—$\delta_i > 0$ for all i, and we know δ_i for all i
3. Fast node updating—After each public activation (a_i set to 1), *all* inactive nodes must update before the next scheduled public activation.

Importantly for social scientists, all threshold measurable processes are private information processes. If we cannot directly observe s_i, but instead measure activation behaviorally via a_i, our process is not threshold measurable. This finding means that exposure-at-activation-time methods, which rely on public activation, do not correctly measure adoption thresholds. Further, activation delays must be long relative to node update frequency and we must have precise information about activation delays. If all nodes do not update before the next public activation, a node may have two neighbors activate between updates, violating Definition 1 and therefore Definition 2. This requirement—very fast updates with slow activation—is seldom satisfied in the social world.

Since the conditions under which threshold measurability holds are almost never encountered, we conclude that Definition 2 is rarely satisfied empirically. In fact, Definition 2 is rarely satisfied even in simulation studies. We note that any simulation that allows the possibility of multiple alters adopting between ego updates is not threshold measurable. For instance, a simulation that employs random updating is not threshold measurable, since the possibility exists that two alters may activate in between ego updates.

We spend the rest of this paper arguing that measuring *all* thresholds in a diffusion process is not necessary to usefully analyze thresholds in empirical situations.

2.3 Failures of Threshold Measurability

The arguments provided above are abstract. It is useful to examine small graphs to understand how, even in simple cases, correctly measuring thresholds fails. We examine two-, three-, and four-node graphs in the section. We limit ourselves to public information processes with no activation delay, since this is often the model of diffusion employed in empirical research. In this section, nodes are named according to their position in the graph (e.g. the topmost node is called *top*).

2.3.1 Dyad

We see a dyad in Figure 2.3.1 where both the *top* and *bottom* nodes have threshold 0. Three update orderings are possible here: $(\{top\}, \{bottom\})$, $(\{bottom\}, \{top\})$, and $(\{top, bottom\})$. The first two of these update orderings are asynchronous. In the asynchronous case, we fail to correctly measure one of the two thresholds. This is indicated by the triangle in Figure 2.3.1.

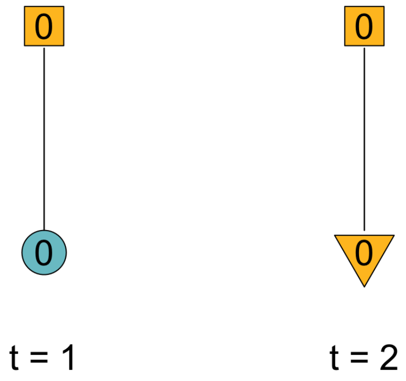

Figure 1: Social contagion in a dyad across two time periods. Blue indicates inactive nodes, orange indicates active nodes; squares represent nodes with correctly measured thresholds, triangles represent nodes with incorrectly measured thresholds. In this particular process, both nodes have threshold 0. Top updates first and is correctly measured. *Bottom* then updates and activates with 1 active neighbor, and so *bottom*'s threshold is incorrectly measured as 1.

2.3.2 Triad

Figure 2.3.2 displays a triad with update order $C = (\{top\}, \{left\}, \{right\})$. We correctly measure the thresholds of *top* and *left*, but incorrectly measure the threshold of $right$. For $right$, we measure a threshold interval of $(0, 2]$.

2.3.3 Tetrad

Figure 2.3.3 demonstrates a diffusion process with update order $C = (\{left\}, \{top, bottom\}, \{right\})$. The reader may have noticed that in the dyadic and triadic cases, a synchronous updating process would allow correct measurement of all thresholds. However, the four-node case demonstrates that a synchronous updating process fails to correctly measure thresholds in simple graph structures as well.

It is easy to check that any asynchronous updating order will also produce at least one incorrectly measured node in the absence of activation delays plus private signals. For instance, the update order $C = (\{left\}, \{top\}, \{bottom\}, \{right\})$ will incorrectly measure the thresholds of *bottom* and $right$.

3 Simulation Evidence

To better understand the prevalence and implications of incorrect measurement, we conduct simulations. We simplify by assuming that our diffusion process is an asynchronous public information process with no activation delays. This type of process is not threshold measurable, so we should not expect to correctly measure all thresholds. However, the rate of incorrect measurement isn't clear from results already presented. Through simulation, we wish to address four questions:

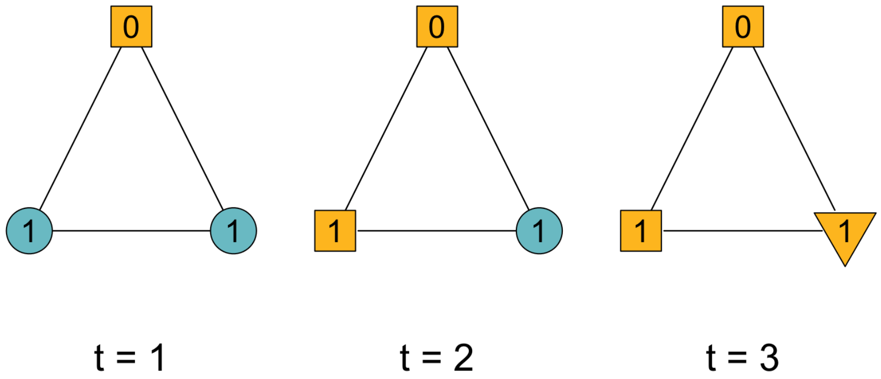

Figure 2: Social contagion in a triad across three time periods. Blue indicates inactive nodes, orange indicates active nodes; squares represent nodes with correctly measured thresholds, triangles represent nodes with incorrectly measured thresholds. We update one node per time period. We see that *right* is incorrectly measured as having threshold of 2, when it is in fact 1.

1. The rate of incorrect measurement
2. The level of mismeasurement introduced by using the exposure-at-activation rule
3. The prediction bias in estimating the threshold distribution from the observed subset
4. The prediction bias when predicting thresholds from the first k adopters

3.1 Simulation Details

We conduct simulations on four graph topologies: random graph, Watts–Strogatz (D J Watts and Strogatz 1998), Barabási—Albert (Barabási and Albert 1999), and power law with clustering (Holme and Kim 2002). Graphs have 1000 nodes, with mean degree of 12, 16, or 20. We run simulations 100 times per graph parameterization. We present results only from Watts–Strogatz and power law with clustering graphs here, as results from the other topologies are not markedly different.

For each simulation trial, we use the following simulation algorithm:

1. Generate a random asynchronous update order for all nodes C, which has one node per interval.

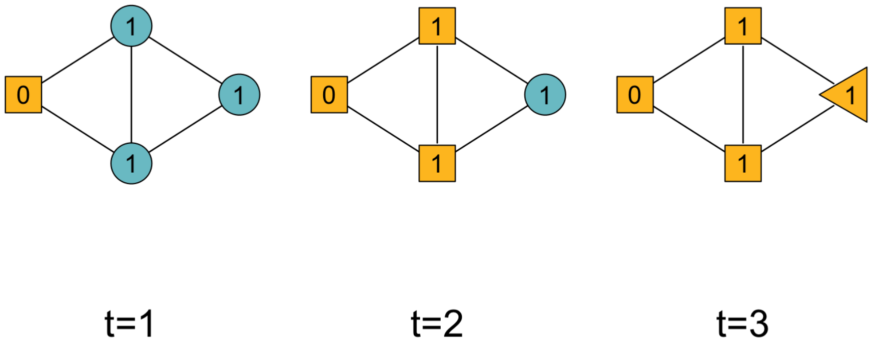

Figure 3: Social contagion in a tetrad across three time periods. Blue indicates inactive nodes, orange indicates active nodes; squares represent nodes with correctly measured thresholds, triangles represent nodes with incorrectly measured thresholds. *Right* has threshold 1, yet node has 2 active neighbors at activation time. Note that any synchronous or asynchronous updating order fails to correctly measure the threshold one of at least one of the three nodes with threshold 1. For instance, if we were to activate $left$, *top*, and *right*, *bottom* would have 3 active neighbors at activation time and therefore be incorrectly measured.

2. For each node in C:
   (a) If the node is active, do nothing
   (b) If the node is inactive:
      i. If the node's threshold is satisfied: activate it
      ii. If the node's threshold is not satisfied: do not activate, and record the number of active neighbors
3. If not all nodes in the graph are active and at least one node was activated in the last pass through C, generate a new C and repeat the process
4. Stop if all nodes have been iterated through without any activations

This algorithm allows the social contagion to diffuse maximally, since if all nodes are checked at random and none update, no alternate update ordering will change the outcome. We record the number of active neighbors at each node update time, even if the node does not activate. This allows us to construct a threshold interval from the data as found in Definition 1. If a node has a threshold interval of size 1 or had 0 neighbors active at activation time, it is correctly measured. Otherwise, it is incorrectly measured.

3.2 Thresholds from Covariates

Granovetter 1978 generates thresholds by simply drawing from a uniform or Gaussian distribution, while Duncan J Watts 2002; Kempe, J. Kleinberg, and

Tardos 2003 use a uniform distribution. In empirical research, however, we often employ actor-level characteristics to explain outcomes. If we treat a node's threshold as the outcome to be explained, we can construct a model of ego's threshold as a function of ego-level covariates, including ego-network characteristics (e.g. degree), and ego's network-level characteristics (e.g. ego's betweenness). As with all regression-style models, as long as we can measure the characteristic, we can include it a model of thresholds.

To our knowledge, thresholds have not been treated as an outcome variable, and have traditionally been used as an explanatory variable. In addition to the theoretical usefulness of explaining thresholds, we we need to develop a model with thresholds as the outcome in order to predict thresholds for the incorrectly measured nodes. Recasting thresholds in a regression framework makes Granovetter's model a special case of ours, where we suppress the importance of actor-level covariates in favor of studying other theoretical questions.

For our simulations, we generate thresholds from a simple model. Call y_i node i's threshold, $x_i \sim \mathcal{N}(0,1)$ an explanatory variable, and $\epsilon_i \sim \mathcal{N}(0,\sigma)$ the unmeasured idiosyncratic error of i. We treat σ as a parameter to explore and vary it from $[0.5, 0.8, 1.0, 1.5, 2.0]$. Then we have *threshold equation*

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

For simulation purposes, assume the parameters of this model to be

$$y_i = 5 + 3x_i + \epsilon_i \tag{2}$$

Treating the error standard deviation σ as a parameter allows varying the maximum explained variance in the model. A high value of σ means that even a perfect model will have modest explanatory power.

3.3 Which Thresholds Are Correctly Measured?

We choose one run of the simulation with a power law with clustering graph with mean degree 20 for exposition. We see in Figure 3.3 that measured thresholds tend to be lower-valued thresholds. As shown in Figure 3.3, the correct measurement rate for the simulations we study ranges between 6% and 19%.

3.4 Measurement Bias

Only a small fraction of thresholds are correctly measured. Threshold intervals with size greater than 1 prevent us from being certain where an individual's threshold lies. The exposure-at-activation-time rule uses the maximum of the threshold interval as an individual's threshold, which always produces upward bias in measuring thresholds.

The amount of upward bias is a relevant question, however. In Figures 3.4 and 3.4 we plot the exposure-at-activation-time measurement against the true threshold. At all true threshold levels, upward measurement bias is substantial. In two particularly extreme cases, this simulation run results in a node of threshold 6 being measured with threshold 43, and a node with threshold 13 being measured with threshold 46. These represent high degree nodes that go long periods of time between updates.

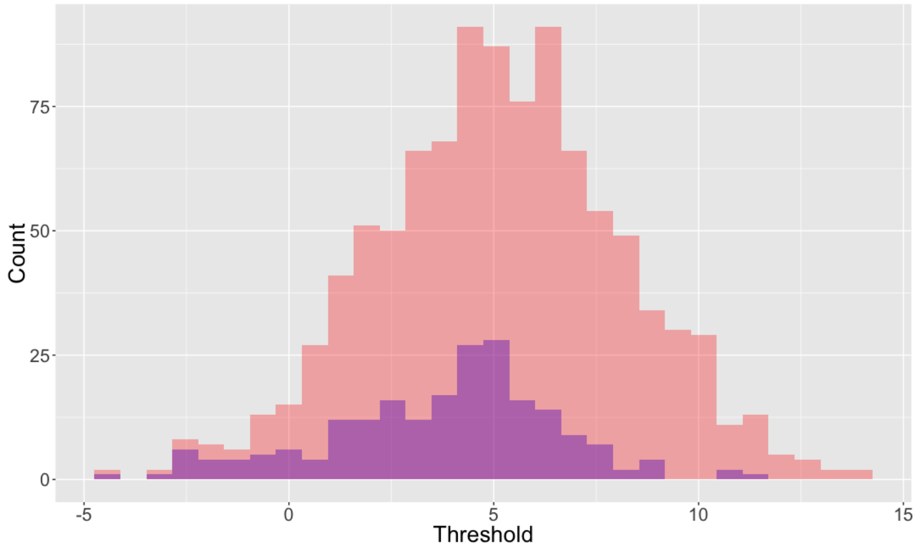

Figure 4: Correctly measured thresholds for one run of the power law with clustering graph with mean degree 20. Correctly measured thresholds are in purple, while all measured thresholds are in red. For instance, if at a given threshold value, the purple bar reaches 40% of the height of the red bar, it means that we measure 40% of the thresholds at that value correctly.

This example is not anomolous: such outliers occur in every simulation we have examined visually. Without incorporating information on node update behavior and applying our measurement condition, any node observed with a high exposure at activation can be rationalized as a low-threshold node with a long update interval. In Figure 3.4 we replot Figure 3.4 with axes to scale, in order to give a clear visual indication of the level of mis-measurement induced by employing the exposure-at-activation-time rule.

3.5 Selection Into Correct Measurement

We use the correctly measured subset of nodes to estimate the relationship between Y and X. In graphs with 1000 nodes, this means using between 50-200 observations to estimate the model. If selection into correct measurement were random, then we would expect $\hat{\beta}$ to be unbiased, but with more variance due the reduction in sample size.

Figure 3.5 demonstrates the result of using OLS on a representative simulation run. The slope of the line generated from the correctly measured subset slightly understates the relationship between X and Y. Using replications across our parameter space, we find that using the correctly measured subset to estimate β does not recover the true relationship β_1 between X and Y[4].

Biased parameters imply selection on the error term. Figure **??** presents the error term in the correctly measured subset as a function of model parameters. Note that, on average, the error term is negative for all parameter values, while

[4] This is not due to left-censoring. Tobit models are also biased.

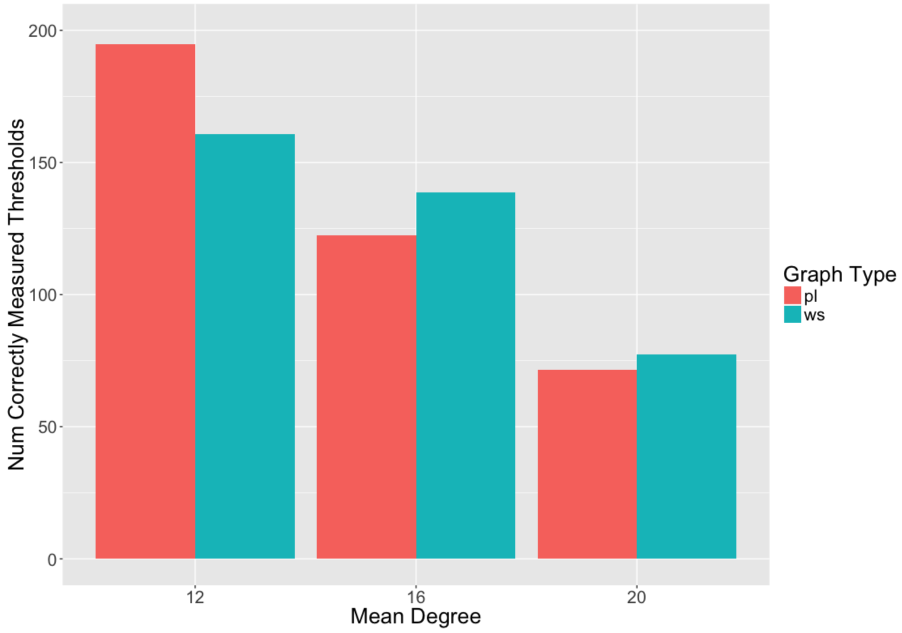

Figure 5: Number of correctly measured thresholds by graph type and mean degree. Graphs have 1000 nodes. Each bar is averaged over the 5 different σ values employed for the error term, comprising 5 different simulations with 100 replications each.

the true error is mean 0 by construction. This demonstrates the presence of selection bias. Returning to Figure 3.3, we are more likely to correctly measure lower thresholds, meaning that having a negative error term makes it more likely that we correctly measure a node's threshold.

The standard response to selection bias is to model it as a function of observables using a Heckman procedure Since we know the true model (Equation 2), the only other variables that we can include are network measures. We hypothesized that nodes in certain network positions would be more likely to be selected into correct measurement. For instance, perhaps lower degree nodes would be more likely to be correctly measured. In our simulations, using ego's degree, closeness, betweenness, and eigenvector centrality in the first stage of a Heckit procedure did not reduce selection bias. We leave further investigation of network-level factors leading to selection into correct measurement for future research.

Since we are unable to model the selection bias, we have *selection on the error*, which results in the correctly measured subset having a *systematically different* relationship between X and Y in the correctly measured subset than the population. In other words, $\hat{\beta} \neq \beta$.

3.6 The Effects of Selection Bias

There are two reasons to estimate Equation 2: to recover an unbiased $\hat{\beta}$ to explain causes of individual thresholds; or to use the estimated model to predict

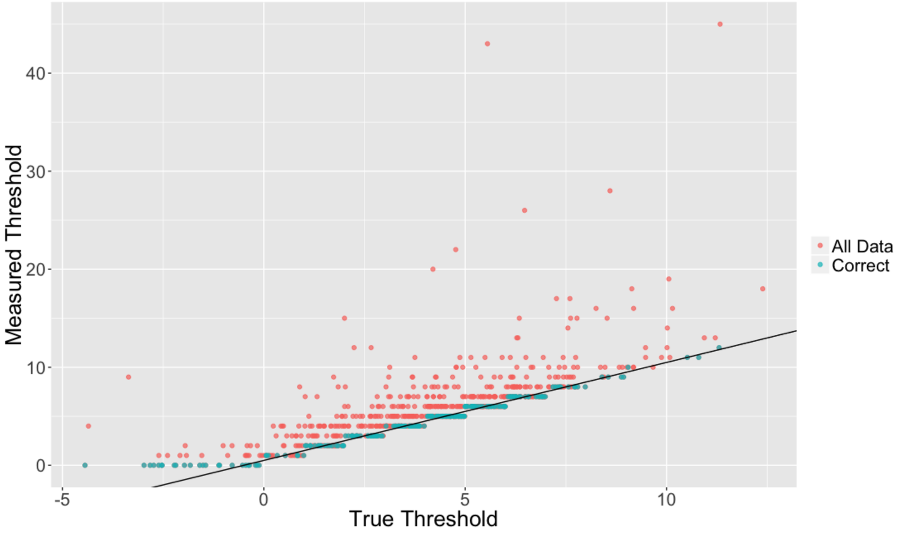

Figure 6: Plot of the thresholds versus measured thresholds. As we go across the x-axis, we see that at each true threshold value, measured thresholds are above the $x = y$ line, except for cases identified by our correct measurement condition. Note the different scales in the axes.

thresholds for the incorrectly measured nodes in order to recover the threshold distribution. This is an *explanation* versus *prediction* problem, as discussed in B. J. Kleinberg et al. 2015.

Figure **??** displays the average value of the error term in the correctly measured subset. As we increase the unexplained variance in the true model, selection on the error becomes worse. We see in Figure 3.5 the level of bias in $\hat{\beta}$ for increasing σ. When X explains much of the variance in Y, $\hat{\beta}$ is close to β, although still biased. When σ is larger, nodes with larger negative error terms are selected into correct measurement, which increases the bias on $\hat{\beta}$. This indicates that using our methodology to explain factors that contribute to thresholds must be done carefully, and researchers must have strong theoretical reasons to believe that the correct variables are included in the model. We note that the R^2 of the regression on the correctly measured subset is *not* the same as the R^2 on the true model in the population, and an argument that good predictors are included in the model is primarily theoretical.

The increasing level of bias on $\hat{\beta}$ as σ increases creates specific conditions for which $\hat{\beta}$ may be interpreted for *explanation*. However, we find that at any level of σ, our method performs well when using X for *prediction* in order to recover the threshold distribution Y for all nodes. In this context, bias in $\hat{\beta}$ is far less problematic. Figures 3.7 and 3.7 demonstrates the RMSE for predicting Y for all nodes in the graph, using only the correctly measured subset. We use the RMSE of the exposure-at-activation-time rule as a baseline[5].

We see that, even as we increase σ, we do not substantially increase the prediction error. If we treat the total RMSE as composed of *bias* and *variance*, we see that the vast majority of RMSE at high levels of σ is due to variance. This

[5] This compares $e_{i,t}$ for the first t at which i is active to h_i, the true threshold of i.

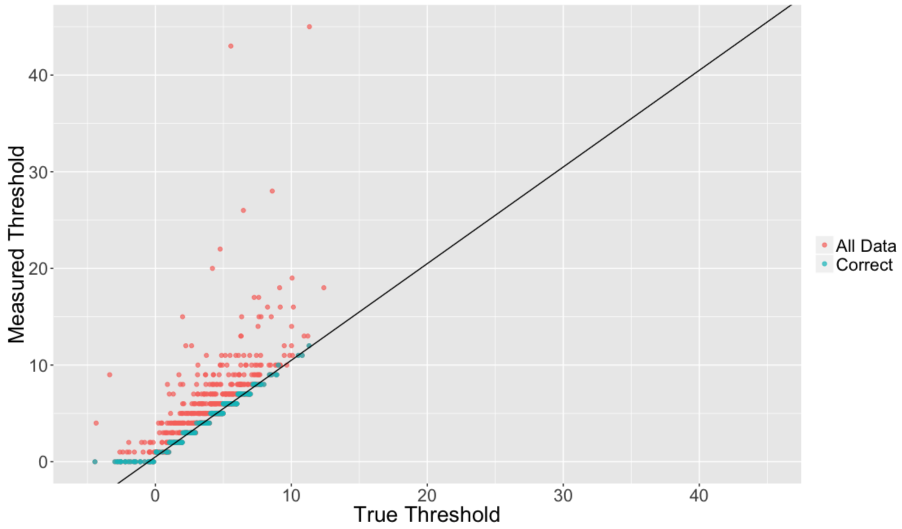

Figure 7: A restatement of 3.4, except with equally scaled axes. We see clearly here that the level of upward bias for some nodes is extreme. For instance, there is a node with a threshold around 6 that is measured as 40+. Using the rule that thresholds are the number of active neighbors at activation time stretches the range of the threshold distribution by a factor of two.

implies that, even if coefficients are biased, simple regression methods provide good predictions of Y, and allow recovering the true threshold distribution with high accuracy.

3.7 Prediction Error Using k Correct Measurements

Prediction performance is good when using the entire correctly measured subset, but in many empirical applications we wish to predict future diffusion performance from its initial behavior (Cheng et al. 2014). We address the ability to predict node thresholds using only the first k correctly measured nodes.

4 Conclusion

In this paper, we have identified the conditions under which node thresholds are correctly measured. By analogy with a laboratory experiment, we have shown that the private threshold satisfaction state of nodes must be measured after each alter adoption for diffusion processes to be threshold measurable. Since this condition is strict, in virtually all empirical and simulation cases, some thresholds will be incorrectly measured.

Despite this incorrect measurement, we show that employing a simple regression technique provides low-bias parameter estimates when observables explain a large fraction of the variance in thresholds. Even when explained variance is low, parameter bias adds relatively little prediction error beyond the prediction error from variance. We view this result as encouraging for work that wishes to recover threshold distributions in empirical settings. However, when a model of

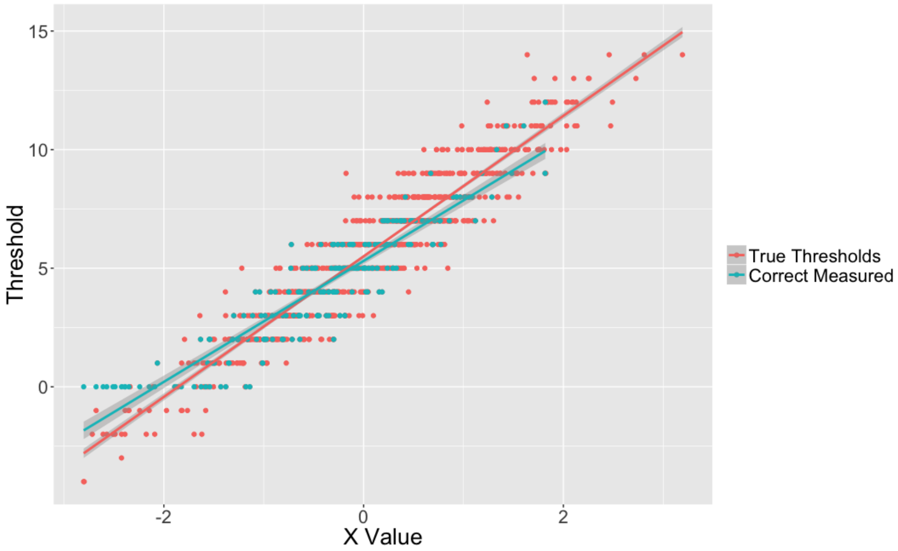

Figure 8: A comparison of OLS estimates for the relationship between x and y using the correctly measured subset and the true relationship in the data.

thresholds is interpreted as explanatory, researchers must carefully craft theoretical arguments that the incorporated variables have both explanatory power and causal relevance.

We argue that this paper facilitates categorizing social contagion by its *threshold distribution*, which has been theoretically relevant since Granovetter 1978 but has not been operationalized. In contrast to structural categorizations of diffusion (Goel, Duncan J Watts, and Goldstein 2012), a threshold distribution categorization of diffusion facilitates new explanations for the success or failures of diffusion. We look forward to future work contrasting diffusion processes on the basis of their threshold distributions[6].

[6] For example, we reanalyzed the Coleman, Katz, and Menzel 1957 dataset, and found no evidence that doctor-level characteristics affected the doctors' thresholds. This suggests the absence of peer influence, since theoretical work on diffusion suggests that certain individuals are predisposed towards adopting new and risky products.

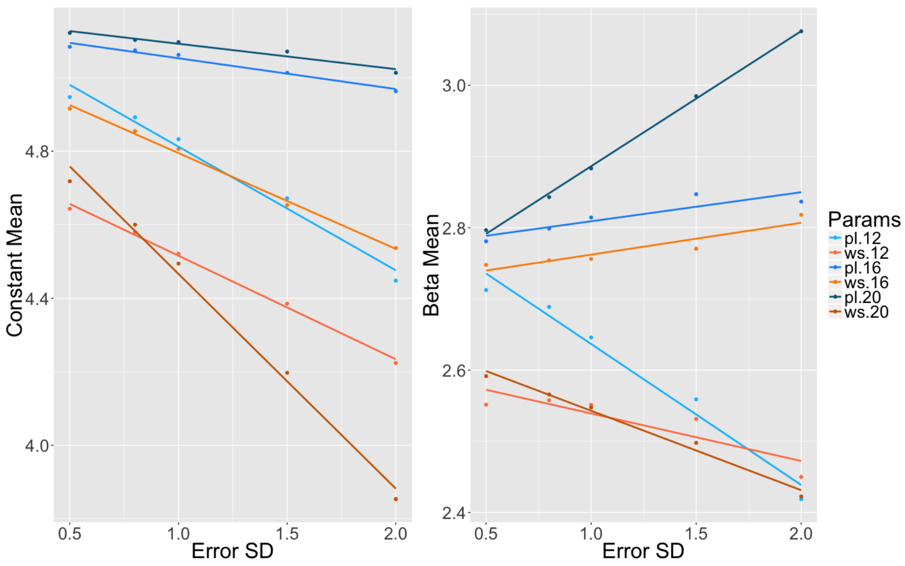

Figure 9: The true values from Equation 2 are a constant of 5 and a beta coefficient of 3. We see here that, as we increase the unexplained variance in the model in the form of error standard deviation, we create more variance in our predictions and, in most cases, introduce more bias as well. Colors correspond to graph type, with shades of blue representing power law graphs with clustering, and shades of orange representing Watts-Strogatz graphs.

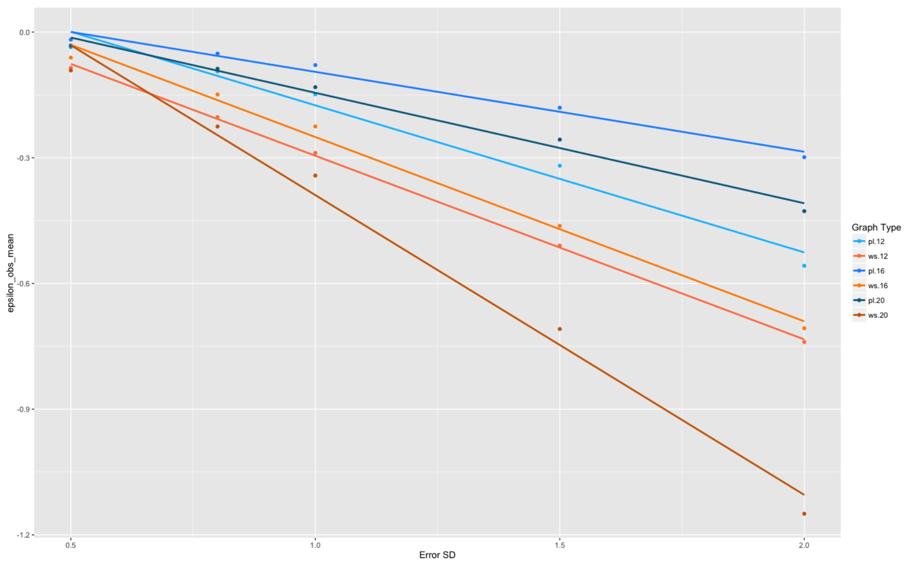

Figure 10: The value of epsilon in the correctly measured subset for values of the error standard deviation.

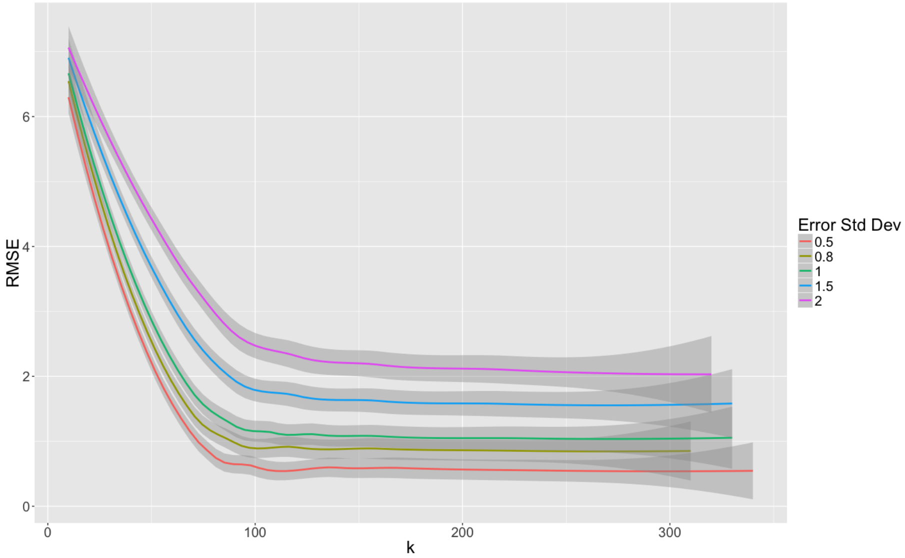

Figure 11: We see the RMSE using the first k correctly measured nodes to predict thresholds for the whole graph. Here we break down RMSE by error SD, and see that higher error SD corresponds to higher RMSE. This does not disentangle bias and variance.

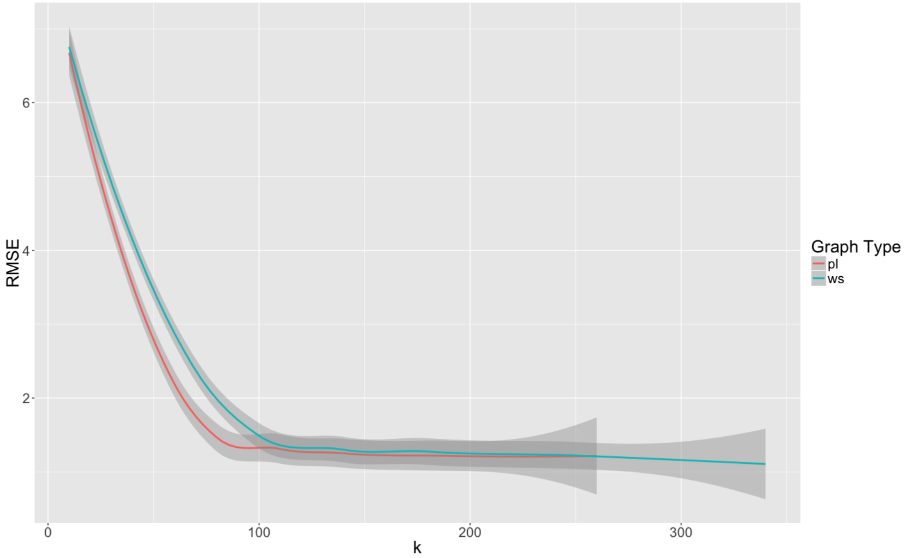

Figure 12: We see the RMSE using the first k correctly measured nodes to predict thresholds for the whole graph. Here we break down RMSE by graph type, and see that both graphs have about the same prediction error when $k > 100$ (or 10%). This does not disentangle bias and variance.

References

Backstrom, Lars (2006). "Group formation in large social networks: membership, growth, and evolution". In: *In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 44–54. DOI: `10.1145/1150402.1150412`.

Barabási, Albert-László and Reka Albert (1999). "Emergence of scaling in random networks". In: *Science* 286.5439, p. 11. DOI: `10.1126/science.286.5439.509`.

Centola, Damon and Michael Macy (2007). "Complex Contagions and the Weakness of Long Ties". In: *American Journal of Sociology* 113.3, pp. 702–734.

Cheng, Justin et al. (2014). "Can cascades be predicted?" In: *Proceedings of the 23rd international conference on World wide web*, pp. 925–936. DOI: `10.1145/2566486.2567997`.

Coleman, James, Elihu Katz, and Herbert Menzel (1957). "The Diffusion of an Innovation Among Physicians". In: *Sociometry* 20.4, pp. 253–270.

Crandall, David et al. (2008). "Feedback effects between similarity and social influence in online communities". In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, p. 160. DOI: `10.1145/1401890.1401914`.

Goel, Sharad, Duncan J Watts, and Daniel G Goldstein (2012). "The Structure of Online Diffusion Networks". In: *Proceedings of the 13th ACM Conference on Electronic Commerce* 1.212, pp. 623–638. DOI: `10.1145/0000000.0000000`.

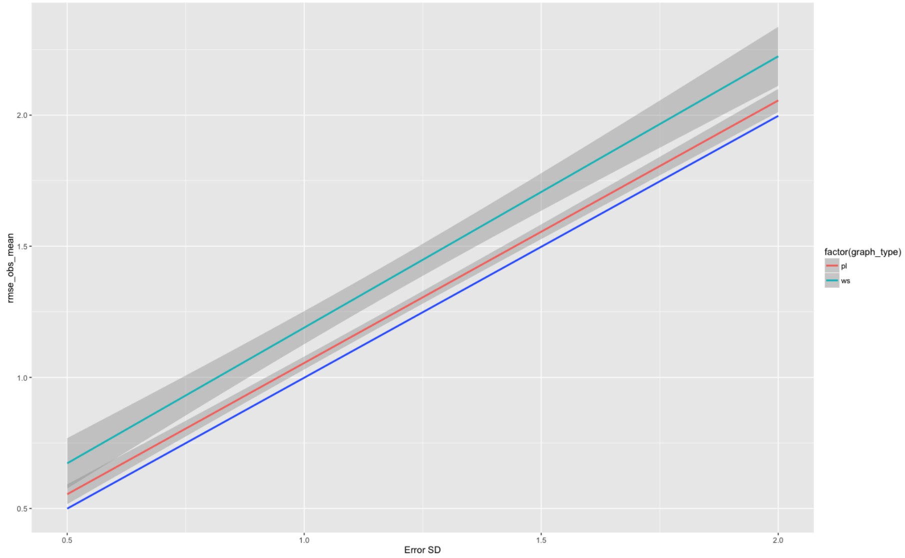

Figure 13: This figure plots the ideal RMSE at a given error standard deviation versus the RMSE from the model estimated with the correctly measured subset. The gap between the baseline (dark blue line) the cyan and red lines represents the RMSE due to bias. We see that RMSE due to bias does not increase dramatically as the unexplained variance increases.

Granovetter, M. (1978). “Threshold models of collective behavior”. In: *American journal of sociology* 83.6, pp. 1420–1443. DOI: 10.1086/226707.

Holme, Petter and Beom Jun Kim (2002). “Growing scale-free networks with tunable clustering”. In: *Physical Review E* 65.2, pp. 1–4.

Kempe, David, Jon Kleinberg, and Éva Tardos (2003). “Maximizing the spread of influence through a social network”. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’03*, p. 137. DOI: 10.1145/956755.956769.

— (2005). “Influential Nodes in a Diffusion Model for Social Networks”. In: *Automata, Languages and Programming* 3580, pp. 1127–1138. DOI: 10.1007/11523468_91.

Kleinberg, By Jon et al. (2015). “Prediction Policy Problems”. In: 105.5, pp. 491–495.

Ludemann, C. (1999). “Subjective Expected Utility, Thresholds, and Recycling”. In: *Environment and Behavior* 31.5, pp. 613–629. DOI: 10.1177/00139169921972263.

Macy, Michael W (1990). “Learning Theory and the Logic of Critical Mass”. In: *American Sociological Review* 55.6, pp. 809–826.

Rogers, Everett M. (1962). “Diffusion of Innovations”. In:

Romero, Daniel M, Brendan Meeder, and Jon Kleinberg (2011). “Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter”. In: *Proceedings of the 20th international conference on World wide web*, pp. 695–704. DOI: 10.1145/1963405.1963503.

Strang, David and Sarah a. Soule (1998). “Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills”. In: *Annual Review of Sociology* 24.1, pp. 265–290. DOI: 10.1146/annurev.soc.24.1.265.

Valente, Thomas W (1995). *Network Models of the DIffusion of Innovations*, p. 184.

— (1996). “Social Network tresholds in the diffusion of innovations”. In: *Social Networks* 18.95, pp. 69–89.

Watts, D J and S H Strogatz (1998). “Collective dynamics of 'small-world' networks.” In: *Nature* 393.6684, pp. 440–2. DOI: 10.1038/30918. URL: http://www.ncbi.nlm.nih.gov/pubmed/9623998.

Watts, Duncan J (2002). “A simple model of global cascades on random networks”. In: *Proc. Natl. Acad. Sci.* 99.9, pp. 5766–5771.