Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects

Dean Eckles\textsuperscript{a,b} and Eytan Bakshy\textsuperscript{c}

\textsuperscript{a}Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA; \textsuperscript{b}Institute for Data, Systems & Society, Massachusetts Institute of Technology, Cambridge, MA; \textsuperscript{c}Facebook, Menlo Park, CA

ABSTRACT

Peer effects, in which an individual’s behavior is affected by peers’ behavior, are posited by multiple theories in the social sciences. Randomized field experiments that identify peer effects, however, are often expensive or infeasible, so many studies of peer effects use observational data, which is expected to suffer from confounding. Here we show, in the context of information and media diffusion, that high-dimensional adjustment of a nonexperimental control group (660 million observations) using propensity score models produces estimates of peer effects statistically indistinguishable from those using a large randomized experiment (215 million observations). Compared with the experiment, naïve observational estimators overstate peer effects by over 300% and commonly available variables (e.g., demographics) offer little bias reduction. Adjusting for a measure of prior behaviors closely related to the focal behavior reduces this bias by 91%, while models adjusting for over 3700 past behaviors provide additional bias reduction, reducing bias by over 97%, which is statistically indistinguishable from unbiasedness. This demonstrates how detailed records of behavior can improve studies of social influence, information diffusion, and imitation; these results are encouraging for the credibility of some studies but also cautionary for studies of peer effects in rare or new behaviors. More generally, these results show how large, high-dimensional datasets and statistical learning can be used to improve causal inference. Supplementary materials for this article are available online.

1. Introduction

Understanding how the behavior of individuals is affected by the behavior of their peers is of central importance for the social and behavioral sciences, and many theories suggest that positive peer effects, in which a peer performing a behavior causes an individual to perform the same behavior, are ubiquitous (Granovetter 1978; Blume 1995; Manski 2000; Centola and Macy 2007; Montanari and Saberi 2010). However, it has been difficult to identify and estimate peer effects in situ. Much of the most credible evidence about peer effects comes from small experiments in artificial social environments (Sherif 1936; Asch 1956; Whiten, Horner, and De Waal 2005; Herbst and Mas 2015). In some cases, researchers have conducted field experiments modulating tie formation and group membership (Sacerdote 2001; Zimmerman 2003; Lyle 2007; Carrell, Fullerton, and West 2009; Centola 2010; Firth, Sheldon, and Farine 2016), shocks to group or peer behavior (Bond et al. 2012; Banerjee et al. 2013; van de Waal, Borgeaud, and Whiten 2013; Aplin et al. 2015; Cai, De Janvry, and Sadoulet 2015; Eckles, Kizilcec, and Bakshy 2016), or subsequent exposure to peer behaviors via particular mechanisms (Salganik, Dodds, and Watts 2006; Aral and Walker 2011; Bakshy, Eckles et al. 2012). But in many cases these experimental designs are infeasible; for example, it may not be possible or ethical to randomize exposure to peer behaviors (see Eckles, Kizilcec, and Bakshy 2016; Taylor and Eckles 2018). Or feasible versions of these designs do not identify the quantities of interest; for example, randomizations that affect tie formation or group membership may not be informative about the counterfactuals of interest (Carrell, Sacerdote, and West 2013; Angrist 2014), while exposure to a peer behavior only via some particular mechanism (e.g., personalized social information in advertisements, Bakshy, Eckles et al. 2012) may be less interesting than total peer effects.

Thus, much recent work on peer effects uses observational data from new large-scale measurement of behavior (Aral et al. 2009; Bakshy et al. 2011; Ugander et al. 2012; Allen et al. 2013; Friggeri et al. 2014) or longitudinal surveys (Christakis and Fowler 2007; Iyengar, Van den Bulte, and Valente 2011; Banerjee et al. 2013; Card and Giuliano 2013; Christakis and Fowler 2013; Fortin and Yarbeck 2015). Many of these studies are expected to suffer from substantial confounding of peer effects with other processes that also produce clustering of behavior in social networks, such as homophily and external causes common to network neighbors. Individuals are typically more likely to form relationships and interact with similar others (McPherson, Smith-Lovin, and Cook 2001; Currarini, Jackson, and Pin 2010), network neighbors are often subject to common external causes (e.g., marketing campaigns, Van den Bulte and Lilien 2001), and current behaviors may have been produced by peer effects in previous periods. Thus, even when issues of simultaneity (including “reflection,” Manski 1993) can be avoided, it is generally not possible to identify peer effects...
using observational data without the often suspect assumption that adjusting for available covariates is sufficient to make peer behavior unconfounded (i.e., as-if randomly assigned, Shalizi and Thomas 2011; Angrist 2014). This may lead researchers to reasonably expect that observational studies of peer effects often suffer from confounding to a degree that makes them misleading for many purposes. However, even if these assumptions are not strictly satisfied, some observational estimators, especially those that adjust for numerous or particularly relevant behavioral variables, may have relatively small bias in practice, such that the bias is small compared with other sources of error (e.g., sampling error) or is small enough to not change choices of theories or policies.

Using a massive field experiment as a “gold standard,” we conduct a constructed observational study by adding a nonexperimental control group to a randomized experiment (LaLonde 1986; Dehejia and Wahba 1999, 2002; Hill, Reiter, and Zanutto 2004) to assess bias in observational estimators of peer effects. The randomized experiment modulates exposure to a peer behavior, such that individuals who would have been exposed to a peer performing the behavior are sometimes randomly assigned to not be exposed. On the other hand, the nonexperimental control group consists of cases where individuals do not have a peer who performs the behavior. The present work is an advance over prior experimental evaluations of observational causal inference methods, which have often suffered from both bias (e.g., problems of incomparability of the experimental data, and estimators used in this study. Section 3 compares the resulting experimental and observational estimates of peer effects, where we find substantial variation among the performance of the observational estimators. Section 4 concludes, highlighting advances over prior work but also limitations.

2. Data and Method

We analyze a large experiment that randomly modulated the primary mechanism of peer effects in information and media sharing behaviors on Facebook: the Facebook News Feed (Bakshy, Rosenn et al. 2012). Facebook users can share links to particular Web pages (URLs), which is a common way of disseminating news, entertainment, and other information and media. Using a cryptographic hash function, a small fraction of user–URL pairs were randomly assigned to a no feed condition in which News Feed stories about a peer sharing that URL were not displayed to that focal user. Deliveries of these stories and held out deliveries (i.e., those for pairs in the no feed condition) were recorded. We study in-kind peer effects whereby individuals are caused to perform a behavior by their peer performing the same behavior; that is, the potential outcomes of interest are binary indicators of whether user $i$ would share a URL $u$ if exposed $Y_{iu}(1)$ or not exposed $Y_{iu}(0)$ to a peer sharing that URL. The observed outcome is thus $Y_{iu} = E_{iu}Y_{iu}(1) + (1 - E_{iu})Y_{iu}(0)$, where $E_{iu}$ indicates that user $i$ is exposed to a peer sharing URL $u$. Taking exposure to a peer sharing a URL as the treatment, this experiment produces random variation in treatment among those who would otherwise be treated; thus, it identifies and enables straightforward estimation of average peer effects for users who would have been exposed to peer sharing. More specifically, key causal quantities identified by the experiment include the relative risk of sharing, $RR = \frac{p(1)}{p(0)}$, and the risk difference of sharing (i.e., the average treatment effect on the treated), $\delta = p(1) - p(0)$, where $p(1) = Pr(Y_{iu}(1) = 1 \mid E_{iu} = 1)$ is the probability of sharing a particular URL when exposed to a peer sharing that URL for those that would be exposed and $p(0) = Pr(Y_{iu}(0) = 1 \mid E_{iu} = 1)$ is the probability of sharing a particular URL when not exposed to a peer sharing URL for those that would be exposed. Note that $p(0)$, but not $p(1)$, involves a counterfactual, so our estimate $\hat{p}(1)$ is common to the experimental and observational analyses.

We combine the exposed user–URL pairs from the experiment with a sample of 660 million user–URL pairs that were observed to be unexposed because the user lacked peers who shared the URL. These form the nonexperimental control group (NECG); see Section 1.2 in the supplementary materials.

2.1. Propensity Score Modeling and Stratification

All of the observational estimators evaluated here are the result of post-stratification (i.e., subclassification) by the domain name of the URL (e.g., for the URL http://www.cnn.com/article_x, the domain is www.cnn.com); that is, per-domain estimates are combined weighting by the number of exposed observations per domain. The adjusted estimators additionally use granular stratification on estimated propensity scores (Rosenbaum and Rubin 1983, 1984; Rubin 1997).

The propensity score $\epsilon(X_{iu}) = Pr(E_{iu} = 1 \mid X_{iu})$ is the probability that user $i$ is exposed to a peer sharing URL $u$, where $X_{iu}$ are variables describing that user–URL pair. In an observational study, researchers typically rely on the following assumptions. Under conditional unconfoundedness, the potential outcomes are independent of the exposure, $(Y_{iu}(0), Y_{iu}(1)) \perp E_{iu} \mid X_{iu}$. Under overlap or positivity, units have positive conditional probability of exposure and non-exposure, $\epsilon(X_{iu}) \in (0, 1)$. These assumptions imply that exposure is also unconfounded conditional on $\epsilon(X_{iu})$ (Rosenbaum and Rubin 1983). In observational studies, propensity scores are estimated using available covariates and
conditioned on using regression adjustment, matching, weighting, or post-stratification.

2.1.1. Propensity Score Modeling

Since the true model for peer and ego behavior is expected to be highly heterogeneous across very different behaviors (i.e., URLs), we fit a separate model for each domain. This also facilitates a form of internal replication. So for model \( m \), the estimated propensity score for user \( i \) being exposed to a URL \( u \) from domain \( d \) is

\[
\hat{e}_d(X_{iu}) = \log^{-1}(X_{iu}\hat{\beta}_d).
\]

This procedure is conducted for each of the models described in the next section. (We suppress indication of which model \( m \) is used from notation except where needed.)

We estimate propensity scores using logistic regression using different sets of predictors. To provide numerical stability for large domains and some shrinkage for small domains, we use \( L_2 \)-penalized logistic regression (i.e., logistic ridge regression) with a penalty of \( \lambda = 0.5 \). As the estimated scores are used solely for stratification, only the rank of the scores matters for the analysis. (Results were not substantially changed by use of other penalties \( \lambda \in (0.1, 0.5, 5, 50) \); see Section 2.5 in the supplementary materials.)

2.1.2. Post-Stratification

The resulting estimated propensity scores can then be used in three closely related ways—to construct weights for each unit, to match exposed and unexposed units, or to divide the sample into strata (i.e., subclasses). We use post-stratification (i.e., subclassification) on the estimated propensity scores. Such stratification can also be regarded as form of nonparametric weighting or many-to-many matching, sometimes called “blocking” (Imbens 2004) or “interval matching” (Morgan and Harding 2006), that does not impose a particular ratio of treated to control units, as one-to-one matching methods do. For very large datasets, such as the current study, stratification has computational advantages over matching, easily supports a much larger control group than treatment group, and the larger sample sizes afford using more strata than is otherwise common—following Rosenbaum and Rubin (1984), stratification by quantiles is typical (Lunceford and Davidian 2004). It can also avoid some potential problems with one-to-one matching, such as variance increases due to discarding observations (King and Nielsen 2019) and failure of our bootstrap-based inferential strategy (Abadie and Imbens 2008); see Section 2.1.3.

The boundaries of the strata are given by quantiles of the estimated propensity scores for each user–URL pair within each domain. For each domain, we use a number of strata \( J_d \) proportional to the square-root of the number of exposed user–URL pairs \( n_d^{(1)} \), though any large number of strata yields similar results (see Section 2.1 in the supplementary materials). So for each model \( m \), domain \( d \), and \( j \in \{1, 2, \ldots, J_d\} \) we have an interval \( \hat{Q}_{dj} \subset [0, 1] \) of the scores between the \( j - 1 \) to \( j \)th quantiles. (For some model–domain pairs, discreteness in the estimates mean there are not \( J_d \) unique quantiles.) The stratum-specific probability of sharing is estimated with a simple average of the outcomes for all the unexposed pairs in that stratum,

\[
\hat{p}^{(0)}_{dj} = \frac{1}{n_{dj}^{(0)}} \sum_{(i,u) \in C(d)} Y_{iu} 1[\hat{e}_d(X_{iu}) \in \hat{Q}_{dj}],
\]

where \( C(d) \) is the set of user–URL pairs in the NECG from domain \( d \) and

\[
n_{dj}^{(0)} = \sum_{(i,u) \in C(d)} 1[\hat{e}_d(X_{iu}) \in \hat{Q}_{dj}].
\]

The estimate for a particular domain \( d \) for model \( m \) is an average of the estimates for each stratum weighted by the number of exposed pairs within that stratum

\[
\hat{p}^{(0)} = \frac{1}{J_d} \sum_{d} \frac{n_{dj}^{(1)}}{n_{dj}^{(0)}} \hat{p}^{(0)}_{dj},
\]

where \( n_d^{(1)} = \sum_{j=1}^{J_d} n_{dj}^{(1)} \) is the number of exposed pairs for domain \( d \). Propensity score stratification thus results in weighting outcomes for unexposed individuals in the NECG according to the number of exposed units with similar propensity scores. This process is illustrated with \( J_d = 100 \) strata for a single domain in Figure 1.

Estimates from multiple domains are combined in the same way by weighting the estimate for each domain by the number of exposed pairs for that domain

\[
\hat{p}^{(0)} = \frac{1}{C} \sum_{d} \frac{n_{d}^{(1)}}{n_{d}^{(0)}} \hat{p}^{(0)}_{d},
\]

where \( n_d^{(1)} = \sum_d n_{d}^{(1)} \) is the total number of exposed pairs. This weighted average of domain-specific estimates is then used to estimate the other quantities of interest (e.g., \( \delta \), RR) in combination with the estimate of \( \hat{p}^{(0)} \) common to both the experimental and observational analyses.

2.1.3. Dependence and Statistical Inference

Observations (i.e., user–URL pairs) that share either a user or URL are plausibly dependent for multiple reasons. Note that while the causal inference literature sometimes regards the potential outcomes as fixed and only exposures (i.e., treatments) as random, here exposure of a user to a URL results from the sharing decisions of other users, so it is perhaps more felicitous to allow that sharing decisions are random and perhaps involving common user- and URL-specific shocks. To address this dependence, all statistical inference in this article employs a nonparametric bootstrap strategy for data with this crossed structure (Brennan, Harris, and Hanson 1987; Owen and Eckles 2012); see Section 2.3 in the supplementary materials. For each bootstrap replicate, observations are weighted according to an online half-sampling bootstrap (Owen and Eckles 2012) clustered on both user and URL, according to the following procedure. For the rth replicate, each user is assigned a Bernoulli(1/2) draw and similarly for each URL. Each user–URL pair is then assigned the product of the corresponding draws as its weight; that is, a user–URL pair appears in a bootstrap replicate if and only if both the user and the URL are in the replicate. This is used to conservatively
Figure 1. High-dimensional propensity score modeling and stratification for the sharing probabilities. Illustration for a popular domain (www.nytimes.com) with \( J = 100 \) strata. A propensity score model is fit predicting the probability that a given user–URL pair is exposed (i.e., the user is exposed to that URL by a peer sharing it), using thousands of covariates (corresponding to model AMs). Observations are mapped to 100 strata based on the ECDF of the estimated propensity scores of exposed observations and unexposed observations in the NECG (A); there is variation around the grand mean (dashed line), and this variation is greater than for the models with fewer covariates (see Section 2.4 in the supplementary materials). As expected, exposed pairs are more common among higher strata (B). The probability of sharing a URL is greater for higher strata, both for exposed and unexposed users (C). The naïve observational estimate of \( p(0) \) (gray line) weights the probability of sharing for the unexposed in each stratum by their relative frequencies (as does ignoring the stratification). Propensity score stratification instead weights sharing probabilities by the relative frequency of exposed pairs. This estimate of \( p(0) \) (magenta line) is much closer to the gold-standard, experimental estimate of \( p(0) \) (blue line). The estimate of \( p(1) \) common to both the experimental and observational analysis is shown superimposed (green line).

Figure 2. Comparison of experimental and observational estimates of peer effects. (A) The experiment estimates that users are 6.7 times as likely to share when exposed to a peer sharing, while the observational point estimates are larger. (B) Treating the experimental estimate as the truth, the naïve observational estimate overestimates peer effects by 319%. This bias is substantially reduced by adjusting for prior same domain sharing (magenta) and prior sharing for 3703 other domains (squares). (C) All discrepancies in the estimates of relative risk are due to underestimating \( p(0) \) when using observational data. Error bars are 95% confidence intervals. Brief descriptions of the estimators with number of covariates in parentheses are shown for reference.
estimate sampling variances and covariances, which are used to construct confidence intervals and compute p-values. All confidence intervals reported in this article are 95% standard bootstrap confidence intervals; that is, $\pm 1.96\hat{se}$ where $\hat{se}$ is estimated using this bootstrap.

There could also be additional dependence (i.e., for user-URL pairs that share neither a user nor URL). For example, two URLs that are news stories about the same event might be substitutes and a user’s decision to share one could be affected by peers sharing the other. We do not directly account for such dependence in our statistical inference, instead employing the provisional assumption of the absence of such dependence. One could allow for some such dependence without substantially changing the interpretation of results. For example, if there is dependence due to interference, we can regard the present research as estimating regime-specific quantities that average over other units’ assignments (Sävje, Aronow, and Hudgens 2017).

### 2.1.4. Efficiency

Even under overlap and conditional unconfoundedness assumptions, this estimator will only achieve the asymptotic semiparametric efficiency bound for the ATT under stronger assumptions. First, we generally lack results on asymptotic efficiency under potential interference and other network dependence, though see Sävje, Aronow, and Hudgens (2017) and Chin (2018). Settings aside those concerns, if $\epsilon(X_{ni})$ were estimated nonparametrically, then weighting by the inverse of the estimated propensity scores would be efficient (Hirano, Imbens, and Ridder 2003), and similar arguments apply to post-stratification on the estimated propensity scores with a growing number of strata (Imbens 2004). The sparse, high-dimensional covariates available here motivate instead using a parametric $L_2$-penalized logistic regression, at the cost of this asymptotic property. Alternative methods could combine this regularized propensity score model with an outcome model to achieve efficient estimation (Robins and Rotnitzky 1995; Hahn 1998; Imbens 2004; Belloni, Chernozhukov, and Hansen 2014; Chernozhukov et al. 2016; Athey, Imbens, and Wager 2018). In the present and similar settings, there are reasons to think the potential variance reductions from modeling the outcome are small. In particular, with a rare (<0.2%) binary outcome, there is little information in the outcome for such methods to exploit. We selected the present methods in part because of their appealing computational properties for large datasets, and because of the widespread use and advocacy of methods that do not model outcomes, but only selection into treatment (Imbens and Rubin 2015, chap. 13).

### 2.2. Sets of Covariates

Traditional advice for observational causal inference in general is to adjust for measures of causes of selection into treatment (Imbens and Rubin 2015, chap. 13), particularly those that are also expected to be causes of the outcome, as when not all confounders are unobserved, adjusting for, for example, instrumental variables can increase bias by removing some of the as-good-as-random variation in the treatment (cf. Austin, Grootendorst, and Anderson 2007; VanderWeele and Shpitser 2011; Ding and Miratrix 2015; Middleton et al. 2016). In the case of peer effects, the analyst would select variables believed to be related to causes of ego behaviors (the outcome) and peer behaviors (the treatment). Because of homophily, common external causes, and prior influence, it is expected that peer and ego variables are correlated.

Table 1 lists the variables we computed for use as covariates. These variables are each included in at least one of the model specifications, which are designed to correspond to selections of variables that an analyst might make and to evaluate the contribution of different sets of variables to bias reduction. Model D includes demographic variables only. At least some of these variables, or similar measurements, would likely be available in many other settings. These are all expected to be associated with consuming content from particular sources. Model D can also be seen as a relatively minimal convenience selection of covariates. Model A includes additional variables describing general behaviors on Facebook and degree (i.e., number of peers).

We consider two additional sets of covariates of interest. These are prior behaviors that can be regarded as measuring an individual’s latent interest in and likelihood of performing the focal behavior (i.e., sharing a particular URL) in the absence of exposure to a peer performing that behavior. First, we expected that variables describing prior interactions with nominally related URLs could result in substantial bias reduction. In particular, for some user–URL pair $iu$, let same domain shares count the number of URLs that $i$ shared in the 6 months prior to the experiment that have the same domain name as $u$ (e.g., for the URL http://www.cnn.com/article_x, this is how many URLs the user shared with the domain www.cnn.com). Models that add this variable are indicated with $s$; for example, model $A_s$ adds same domain shares to model $A$. This allows for straightforward evaluation of the consequences of including this variable.

Second, we regard same domain shares as an example of information about prior behaviors that can be readily identified as related. In some cases, such information will be available to analysts. In other cases, this information may not be available, or the related behaviors may not be sufficiently common to be useful. In particular, if the focal behavior is new (e.g., a new product launch) or only newly popular, then this information may be limited. In the present case, very few users may have shared any URLs from some domains during the prior 6 months; that is, same domain shares can be 0 for most or even all users for some domains. For this reason, we also consider adjusting for many other prior behaviors, which may or may not be related to the focal behavior. In particular, we evaluate models that include the number of times a user shared URLs from each of the other domain names that have any prior sharing; we indicate the inclusion of this large sparse matrix of counts of other domain shares with $M$. (This matrix has 3703 or 3704 columns, depending on whether the focal domain is among the 3704 domains that have prior sharing or not, and whether that column is already included as same domain shares. We henceforth sometimes simplify and describe it as contributing 3703 columns.)

### 2.3. Comparisons of Estimators

To evaluate observational estimators of the relative risk $RR$ and risk difference $\delta$, we use the NECG, as described above, to
produce estimates of $p(0)$ that make no use of the control group from the randomized experiment. Recall that the experimental and observational estimates of $p(1)$ are identical, as they are both the proportion of exposed user–URL pairs that resulted in sharing; thus, all discrepancies are due to differences in estimating $p(0)$.

We compute the discrepancy between each of the resulting observational estimates and the experimental estimates. Our focus is primarily on estimates of the relative risk RR, as this has been the estimand emphasized by much prior work (e.g., Aral et al. 2009; Bakshy, Eckles et al. 2012; Bakshy, Rosenn et al. 2012). We also consider the risk difference $\delta$ (i.e., the average treatment effect on the treated, ATT). For each observational model $m$, we have two estimators, $\hat{RR}_m$ and $\hat{\delta}_m$. We generally take the experimental estimates as the gold standard—as unbiased for the causal estimand of interest. This motivates the description of these discrepancies as estimates of bias.

For the relative risk, we can compute the absolute discrepancy in the estimates, $\hat{RR}_m - \hat{RR}_{exp}$. To put this in relative terms, we can compute the percent bias in the relative risk:

$$100 \frac{\hat{RR}_m - \hat{RR}_{exp}}{\hat{RR}_{exp}}.$$

We can similarly compute absolute and percent bias for the risk difference. Since the risk difference $\delta = p(1) - p(0)$ is bounded from above by $p(0)$ (i.e., when the behavior cannot occur without exposure), the maximum possible overestimate of $\delta$ is too large by exactly $p(0)$. Thus, we can also characterize error in terms of this maximum possible overestimate, so when $\hat{\delta}_m \geq \hat{\delta}_{exp}$, we have percent bias of the maximum possible overestimate:

$$100 \frac{\hat{\delta}_m - \hat{\delta}_{exp}}{p(0)}.$$

### 3. Results

We begin with the analysis of the randomized experiment. On average a user exposed to a peer sharing a URL (i.e., a user–URL pair in the feed condition) goes on to share that URL 0.132% of the time, while a user who was not exposed to a URL because that user–URL pair was randomly assigned to the no feed condition goes on to share that URL 0.020% of the time. That is, exposure to a peer sharing a URL causes sharing for $0.112\%$ of pairs (CI = [0.110, 0.115]), and causes them to be 6.7 times as likely to share a URL (CI = [6.5, 7.0]). These are the experimental estimates of peer effects to which we compare observational estimates.

The naive observational analysis, which makes no adjustment for observed covariates, concludes that exposure makes sharing 28.2 times as likely (CI = [23.8, 32.5]). That is, it overestimates peer effects by 319% (Figure 2(A, B)). These results are also displayed in Table 2. Because $p(0)$ is bounded from below by zero (Figure 2(C)), $\hat{p}(1) = 0.00132$ is the maximum possible estimate of $\delta$ (i.e., if all sharing is attributed to peer effects). The naive observational analysis yields $\hat{\delta}_{naive} = 0.00127$ and so thus overstates $\delta$ by 76% of this maximum; in this sense it barely improves on attributing all sharing to peer effects.

We then evaluate observational estimators that use propensity score models with varied sets of covariates, as described above. Only adjusting for demographics (model D) or basic individual-level covariates (model A) resulted in similarly large overestimates by 229% and 194% (Figure 2(B)).

One covariate measuring prior, highly related behaviors was of particular interest and expected to substantially reduce bias a priori: For each user–domain-name pair, same domain shares counts how many URLs from that domain name that user shared in the 6 month pre-experiment period. Estimators that additionally adjusted for same domain shares (models Ds and As)
Table 2. Experimental and observational estimates of peer effects.

| Model | Demographics | General behavior, degree | Same domain shares | Other domain shares | \( p(0) \times 10^4 \) | \( \hat{RR} \) | \( \hat{\delta} \times 10^4 \) |
|-------|--------------|--------------------------|-------------------|-------------------|----------------|-------------|----------------|
| exp   |              |                          |                   |                   | 1.96          | 6.73        | 11.21         |
| naive |              |                          |                   |                   | 0.47          | 28.15       | 12.70         |
| D     | Yes          |                          |                   |                   | [0.39, 0.54]  | [23.80, 32.51] | [12.45, 12.96] |
| A     | Yes          | Yes                      |                   |                   | 0.66          | 19.81       | 12.51         |
| Ds    | Yes          |                          | Yes               |                   | 1.51          | 8.72        | 11.66         |
| As    | Yes          | Yes                      | Yes               |                   | 1.52          | 8.67        | 11.65         |
| M     | Yes          |                          | Yes               |                   | 1.02          | 12.86       | 12.15         |
| AM    | Yes          |                          | Yes               |                   | 1.14          | 11.55       | 12.03         |
| Ms    | Yes          |                          | Yes               | Yes              | 1.72          | 7.64        | 11.45         |
| AMs   | Yes          |                          | Yes               | Yes              | 1.77          | 7.46        | 11.41         |

NOTE: Estimates of the probability of sharing if not exposed \( p(0) \), relative risk (RR), and the risk difference (\( \delta \)) for each model with 95% bootstrap standard confidence intervals in brackets. \( p(0) \) and \( \delta \) are shown in basis points for readability.

conclude that exposure makes sharing 8.7 times as likely—both overestimating the relative risk by only 30% and 29% and thus reducing bias by 91% compared with the naive observational analysis.

Rather than selecting a single measure of prior sharing a priori (e.g., from the same domain), we can use a high-dimensional model including measures of prior sharing for other 3703 domains. The model with all 3719 covariates (AMs), including prior sharing for all domains, eliminated 97% of the naive estimator’s bias for the relative risk. This corresponds to concluding that exposure makes sharing 7.5 times as likely (CI = [6.7, 8.3]). This point estimate is less than 11% larger than and statistically equivalent with (\( p > 0.05 \)) the experimental estimate of 6.7.

Instead of supplementing an a priori related covariate, as in this full model, such high-dimensional adjustment might be used as a substitute when no such covariate is measured. In the absence of prior same domain sharing, the sparse matrix of prior sharing for all other domains reduces bias (i.e., AM compared with A, or AM compared with the naive estimator), but does not fully substitute for also adjusting for same domain sharing (i.e., AM compared with As); for statistical tests, see Table S1 in the supplementary materials.

3.1. Heterogeneity by Prior Popularity

These results show that striking levels of bias reduction are possible when pooling across all domains in our dataset, which aggregates over 10 million distinct behaviors (i.e., sharing a particular URL). Much of this bias reduction results from adjusting for same domain sharing, and over half of the exposures come from the top 10% of the domains, such that the bias reduction obtained may largely be due to well-established, highly popular classes of behaviors. However, much of the peer effects literature focuses on the early stages of the spread of new behaviors (e.g., adoption of new products, opinions, or adaptive behaviors, Iyengar, Van den Bulte, and Valente 2011; Myers et al. 2012; Ugander et al. 2012; Aplin et al. 2015).

How much bias reduction is possible for less common or only newly popular classes of behaviors? We analyze the estimates of peer effects for each of the domains with respect to how popular those domains are for the 6 months prior to the study. Simple
subgroup analyses on groups of domains defined by quintiles of prior popularity (specifically, the number of unique sharing users in the prior 6 months) show that the observational estimators remain significantly biased for the least previously popular domains (Figure 3). For these previously unpopular (i.e., bottom quintile) domains, all of the observational estimates are substantively similar—and dissimilar from the experimental analysis. But as prior popularity increases, substantial differences among the estimates emerge, with some becoming quite similar to the experimental estimates for all other quintiles (Section 3.3 in the supplementary materials). In particular, the domains that were most popular before the study are also popular during the study (Section 3.3 in the supplementary materials). Thus, much of bias reduction for the overall estimates (Figure 2) can be attributed to bias reduction for the domains with the greatest prior popularity.

4. Discussion

The study of peer effects, while central to the social sciences, has been limited by biases that have been difficult to quantify. If an individual performs a behavior after being exposed to a peer doing so, this may be because the exposure induced them to do so—or simply that the individual and peer have similar preferences and dispositions and are subject to common external causes. Field experiments are a promising solution to these problems, but restricting scientific inquiry about peer effects to questions answerable with field experiments would severely limit research in this area: few organizations are able to run experiments with sufficient statistical power to precisely estimate peer effects, it is often impractical to run real-world experiments, and it is not possible to run experiments to retrospectively study the contribution of peer effects to important events. We conducted an evaluation of observational studies of peer effects using a massive field experiment as its comparison point. Treating the experimental results as the “gold standard,” we find substantial variation in how well observational estimators perform: analyses that only adjust for a small set of common demographic and variables suffer from nearly as much bias as unadjusted, naive estimates; but estimates adjusting for a relevant prior behavior selected a priori or thousands of potentially relevant behaviors are able to remove the majority of bias. Our estimate adjusting for the full set of variables was statistically indistinguishable from the experimental estimate. With or without this measure selected a priori, high-dimensional adjustment using thousands of potentially relevant prior behaviors reduced bias. We note that this bias reduction may depend on the presence of meaningful variation in the relevant prior behaviors; we found that newer classes of behavior (i.e., sharing URLs from domains not previously popular) do not exhibit this substantial bias reduction.

These results show how causal inference with routinely collected data can be improved through both substantive knowledge and high-dimensional statistical learning techniques. Scientists studying peer effects can use existing knowledge to inform the selection and construction of relevant measures for adjustment. Our specific methods—granular stratification on propensity scores estimated using high-dimensional regularized regression—are directly applicable to a number of datasets where measures of prior behaviors like those used in this study are available, including online communication behaviors, purchase history data, history of drug prescriptions by individual doctors (Iyengar, Van den Bulte, and Valente 2011), or data from passive transponder tags on animals (Allen et al. 2013; Aplin et al. 2015). The success of high-dimensional adjustment here should encourage scientists to measure a larger number of prior behaviors and to employ modern statistical learning techniques with large datasets for causal inference (Varian 2016), rather than just description and prediction.

4.1. Comparison to Related Evaluations

Prior evaluations of observational estimators of peer effects have lacked comparison with an experiment that identifies the same quantity; these evaluations have instead relied on sensitivity analysis (VanderWeele 2011), simulations (Thomas 2013), and analyses when the absence of peer effects is assumed (Cohen-Cole and Fletcher 2008). One exception is in studies making use of a randomized encouragement as an instrumental variable and comparing this to an observational analysis (e.g., Eckles, Kizilcec, and Bakshy 2016); however, without additional strong assumptions, the estimands for the instrumental variable and the observational analyses need not coincide due to heterogeneous treatment effects (Angrist and Imbens 1995). We also lack meta-analyses comparing observational and experimental estimates of peer effects. Meta-analysis is especially difficult here because of the small number of field experiments previously conducted. One recent meta-analytic study of peer effects in worker productivity (Herbst and Mas 2015) is able to compare lab experiments with field studies, but the field studies are largely not randomized experiments; that study does not attempt a comparison of field experiments and observational studies. Thus, the present results address this absence of credible experimental evaluation of observational estimators of peer effects.

While causal inference for peer effects faces distinctive threats to validity (e.g., homophily), the present study also has methodological advantages over many other evaluations of observational estimators in other settings, such as for educational, medical, and public policy interventions (e.g., LaLonde 1986; Heckman, Ichimura, and Todd 1997; Dehejia and Wahba 1999, 2002; Hill, Reiter, and Zanutto 2004; Michalopoulos, Bloom, and Hill 2004; Diaz and Handa 2006; Shadish, Clark, and Steiner 2008).

First, the present study makes the observational and experimental analyses directly comparable. Unlike meta-analyses comparing observational and experimental studies (e.g., Schuemie et al. 2014; Hemkens, Contopoulos-Ioannidis, and Ioannidis 2016), the units contributing to the observational and experimental analyses are drawn from the same population. Though constructed observational studies aim to avoid the incomparability of observational and experimental estimates that occurs in meta-analyses, Shadish, Clark, and Steiner (2008) argued that many constructed observational studies “confound assignment method with other study features” (p. 1335), such as the places or times data is sampled from, the implementation of the treatments, the version of the measures used as covariates or outcomes, and different patterns of missing data. For
example, while Diaz and Handa (2006) found experimental-observational discrepancies in estimated effects of a conditional cash transfer program, for many of the outcomes these can be at least partially attributed to differences in the survey measures used for the experimental and observational data. Despite these limitations, of the six observational–experimental comparisons they review, Cook, Shadish, and Wong (2008) categorized Diaz and Handa (2006) as one of the two less ambiguous comparisons. On the other hand, in the present study the experimental control group and the nonexperimental control group are drawn from the same population and all of the variables are measured in the same way. Thus, the main explanation of any observational–experimental discrepancies in this study is a process that causes users to be selected into exposure to URLs they are more (or less) likely to share—that is, confounding by, for example, homophily, common external causes, and past influence.

Second, the present study has advantages in external and statistical validity over existing studies that do make the analyses comparable: doubly randomized preference trials (DRPT) (Long, Little, and Lin 2008; Shadish, Clark, and Steiner 2008) randomly assign units to whether they will be randomly assigned to treatment or whether they self-select into treatment. DRPTs can lack external validity and relevance to use of routinely collected data for social science. For example, DRPTs generally present subjects with an explicit choice between treatments, removing any biasing role of, for example, awareness of treatments. By comparison, the present study examines exposure to peer behaviors in situ, where individuals have heterogeneous awareness of the behaviors.

Papers reporting on DRPTs have argued they provide evidence for bias of observational estimators and about which types of covariates and analysis methods most reduce that bias (Shadish, Clark, and Steiner 2008; Pohl et al. 2009; Steiner et al. 2010). However, they employ relatively small samples (e.g., n = 445, n = 202), and these claims are apparently not based on formal statistical inference: for Shadish, Clark, and Steiner (2008) and Steiner et al. (2010) the experimental and unadjusted observational estimates are statistically indistinguishable, so their results are “basically descriptive” (Steiner et al. 2010, p. 256); see Section 5 in the supplementary materials for reanalysis of these results. The same is true of the smaller replication by Pohl et al. (2009). In fact, this problem is not unique to DRPTs: constructed observational studies (e.g., Griffen and Todd 2017) are often similarly underpowered to detect any bias to remove. Thus, an important contribution of the present study is that it has sufficient precision to detect bias and distinguish among the bias of estimators adjusting for different sets of covariates. Furthermore, we were able to conduct this evaluation for thousands of classes of behaviors (i.e., URLs from the same domain) and detect differences in bias reduction by prior popularity.

4.2. Limitations

This study should not be understood as justifying all observational studies of peer effects that adjust for prior behaviors. First, while this experiment includes millions of specific behaviors, they are similar to each other in important ways, and spread via the same communication platform. Other behaviors may differ in their prevalence, size of peer effects, costs of adoption, and the time-scales at which they occur. For example, the posited processes producing correlations of obesity in social networks (Christakis and Fowler 2007) occur over a long period of time and, to the extent they are attributable to peer effects, are presumably the result of peer effects in many contributing behaviors (e.g., diet, exercise). Second, evaluating observational methods requires considering how the resulting estimates are used by scientists and decision-makers. Many of the best-performing observational estimators in this study still overestimated peer effects. How problematic such bias is depends on the specific decisions made with such estimates (e.g., retaining or rejecting a theory, making a change to a marketing strategy). Qualitatively, an exposed individual being 7.5 or 8.7 times as likely to perform a behavior is similar to 6.7 times as likely, but such a difference could matter for computing and comparing return on investment (cf. Lewis and Rao 2015) from, for example, public health marketing campaigns.

Supplementary Materials

The supplementary materials consist of: (a) a document with additional details about the data, the methods, and additional tests comparing the estimators, and (b) code and aggregated data for replicating the results in the main text.

Acknowledgments

We are grateful to L. Adamic, S. Aral, J. Bailenson, J. H. Fowler, W. H. Hobbs, D. Holtz, G. W. Imbens, S. Messing, C. Nass, M. Nowak, A. B. Owen, A. Pelyakhov, B. Reeves, D. Rogosa, J. Sekhon, A. C. Thomas, J. Ugander, and participants in seminars at New York University Stern School of Business, Stanford University Graduate School of Business, UC Berkeley Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, University of Chicago Booth School of Business, Columbia University Department of Statistics, and UC Davis Department of Statistics, and anonymous referees for comments on this work.

Disclosure Statement

D.E. was previously an employee of Facebook while contributing to this research. E.B. has significant financial interests in Facebook, as did D.E. during writing earlier versions of this article. While revisions of this article were under editorial review, D.E. received a grant from Facebook for other research.

References

Abadie, A., and Imbens, G. W. (2008), “On the Failure of the Bootstrap for Matching Estimators,” Econometrica, 76, 1537–1557. [509]
Alcott, H., and Gentzkow, M. (2017), “Social Media and Fake News in the 2016 Election,” Journal of Economic Perspectives, 31, 211–236. [508]
Allen, J., Weinrich, M., Hoppitt, W., and Rendell, L. (2013), “Network-Based Diffusion Analysis Reveals Cultural Transmission of Loltib Feeding in Humpback Whales,” Science, 340, 485–488. [507,514]
Angrist, J. D. (2014), “The Perils of Peer Effects,” Labour Economics, 30, 98–108. [507,508]
Angrist, J. D., and Imbens, G. W. (1995), “Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity,” Journal of the American Statistical Association, 90, 431–442. [514]
