Bias and high-dimensional adjustment in observational studies of social contagion

Dean Eckles[1*] and Eytan Bakshy[1]
[1]Facebook, Inc.

*To whom correspondence should be addressed; E-mail: deaneckles@fb.com.

Social contagion is the spread of behavior that occurs when there are peer effects in the adoption of a behavior, but other processes can also produce correlated behavior in social networks thereby raising concerns about the credibility of observational studies of social contagion. We use a large randomized experiment that identifies peer effects in information and media sharing among 67 million users of Facebook as a "gold standard" for assessing the bias of observational estimators. Here we show, using comparisons with the experiment, how naive observational estimators can hugely overstate contagion, estimating peer effects to be over three times larger than the true effects. We evaluate widely-used propensity score methods with both small convenience sets of variables and high-dimensional models adjusting for thousands of prior behaviors. Commonly used variables (e.g., demographics) offer little bias reduction. Additionally adjusting for a measure of prior behaviors closely related to the focal behavior reduces bias by over 50%. High-dimensional models adjusting for over 3,700 past behaviors provide additional bias reduction, such that the full model reduces bias by over 70%, and the resulting estimates overstate peer ef-

fects by less than 0.3 times. These results are cautionary when such variables may not be measured, but also demonstrate how large data sets and statistical learning techniques can be used to substantially improve causal inference.

Understanding how the behavior of individuals is affected by the behavior of their peers is of central importance for the social and behavioral sciences, as well as decision-making in public policy, health care, product design, and marketing. Many theories predict there will be positive peer effects in many behaviors, such that increasing the number of peers adopting the behavior makes an individual more likely to adopt (*1–4*).

Since these behaviors spread through social networks, they are said to exhibit social contagion. Much of the most credible evidence about social contagion comes from small experiments in artificial social environments (*5, 6*). In some cases, large field experiments manipulating tie formation (*7–9*) or exposure to peer behaviors (*10, 11*) have, in rare cases, been possible, but in many cases these experimental designs are not feasible or ethical. Thus, much of the recent evidence about social contagion comes from observational studies making use of new large-scale measurement of human behavior (*12–17*) or longitudinal surveys (*18, 19*). 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 (*20*) and external causes common to network neighbors. It is thus generally not possible to identify peer effects using observational data without the implausible assumption that the available covariates are sufficient to make peer behavior unconfounded (i.e., conditionally ignorable) (*21*).

To some statisticians and experimentalists, these results may suggest that observational studies of social contagion are more misleading than informative. However, even if these assumptions are not strictly satisfied, some observational estimators may have relatively small bias in practice. For example, the bias may be small enough that it would not substantially affect policy decisions, or the bias could be small in comparison to other sources of error (e.g., sampling

error). This motivates characterizing the performance of observational estimators of peer effects by using realistic simulations or real data for which the true peer effects are known. Prior work has evaluated observational methods using sensitivity analysis (*22*), simulations (*23*), and analyses when the absence of peer effects is assumed (*24*).

Using a massive field experiment as a "gold standard", we conduct a constructed observational study (*25–27*) to assess bias in observational estimates of peer effects in sharing links to Web pages. Since sharing links is a common way of disseminating news, entertainment, and other information, these behaviors are of substantial social-scientific and practical importance. Along with other online behaviors, link sharing has been the subject of a number of studies of social contagion (*13, 15, 17*). The present study is an analysis of peer effects in millions of distinct sharing behaviors, with important differences among them (e.g., sharing one news article versus another).

We find that while some observational analyses overstate positive peer effects by almost as much as is possible, other observational analyses that adjust for numerous prior sharing behaviors eliminate the most of this bias. These results provide evidence of substantial confounding bias in these studies, but also show that state-of-the-art methods using often-available variables can produce much more informative estimates. The analyses we evaluate are appropriate for very large data sets, and so can be widely used by many studies of social contagion of online behaviors.

Data and Methods

We analyze a large experiment (*28*) that manipulated the primary mechanism of peer effects in information and media sharing behaviors on Facebook: the Facebook News Feed. Users can share links to particular Web pages, whose location on the Web is identified by a URL. A small percentage 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 user. Deliveries and held out deliveries were recorded. Taking exposure to a peer sharing a URL as the treatment, this experiment identifies peer effects for users who would have been exposed to peer sharing. For this population, it identifies the relative risk,

$$RR = p^{(1)} / \, p^{(0)},$$

and the risk difference of sharing (i.e., the average treatment effect on the treated),

$$\delta = p^{(1)} - \, p^{(0)},$$

where $p^{(1)}$ is the average probability of sharing a particular URL when exposed to a peer sharing that URL for those that would be exposed and $p^{(0)}$ is the average probability of sharing a particular URL when not exposed to a peer sharing that URL for those that would be exposed.

To evaluate observational estimates of RR and δ, we construct a nonexperimental control group (NECG) of user–URL pairs where that user happens to have not been exposed to that URL (because, e.g., no peers shared that URL). These pairs are used to produce estimates of $p^{(0)}$ that make no use of the experimental control group. 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.

This study considers observational analyses using propensity score modeling, which have been used in many of the most credible observational studies of social contagion (*12*). The propensity score e is the probability that a given case (i.e., a user–URL pair) is exposed to the treatment (i.e., a peer sharing the URL). Conditioning on e is sufficient to identify average treatment effects (*29*). In observational studies, propensity scores are estimated using available covariates and conditioned on using regression adjustment, matching, weighting, or post-stratification. The limited set of available covariates can mean that conditioning on the estimated scores $\hat{e}$ does not identify treatment effects, but may substantially reduce bias.

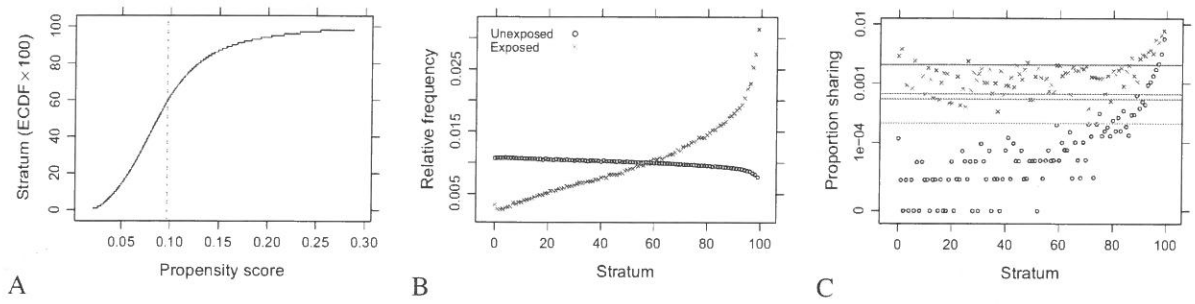

Figure 1: High-dimensional propensity score modeling and stratification for the sharing probabilities of unexposed users from a popular domain (www.nytimes.com). A propensity score model is fit predicting the probability a given user–URL pair is exposed, using thousands of covariates (corresponding to the AMs model). User–URL pairs are mapped to strata (i.e., percentiles) based on the ECDF of the modeled propensity scores of exposed cases and unexposed users in the NECG (A); there is variation around the grand mean (dashed line). 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 naive observational estimate of $p^{(0)}$ (grey 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 common estimate of $p^{(1)}$ is shown is superimposed (green line).

We evaluate methods using propensity score models making use of different sets of demographic and behavioral variables estimated with L_2-penalized logistic regression. This includes sets that reflect common practice and state-of-the-art high-dimensional propensity score modeling. Basic models include only user demographics (D) or 15 user-level variables (A). We group URLs by their domain name, and for each user–domain pair, we compute the number of URLs from that domain shared by that user in a 6 month period prior to the experiment. These behavioral variables are of particular interest, since they were expected to be useful for reducing bias and similar variables can often be collected by social scientists studying peer effects in online behavior. In particular, prior sharing from the same domain is a measure of a users' disposition to share URLs from that same domain (models Ds and As). For example, given the

URL `http://www.cnn.com/article_x`, a propensity score model that includes same domain sharing would include a count variable with the number of times the user had previously shared any URLs with the domain `www.cnn.com`. In place or in addition to such a measure, prior sharing of URLs from other domains can be used as measures of other, potentially related dispositions, as in recommendation systems. Models M, AM, and AMs include the sparse matrix whose columns are these 3,703 variables. Descriptions of all variables are included in *Supporting Information*.

Estimates of $p^{(0)}$ from each model are the result of post-stratification (i.e., subclassfication) on the estimated propensity scores (*30*). For very large data sets with a much larger NECG than treatment group, such as ours, stratification has computational and statistical advantages over other methods, such as one-to-one matching methods. Propensity score stratification 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 in Figure 1, and specified in greater detail in *Materials and Methods*.

Taking the experimental estimates as the gold standard, for each model m, the discrepancy between $\widehat{RR}_{\text{exp}}$ and $\widehat{RR}_m$ is a measure of bias of the observational estimator (and similarly for $\hat{\delta}_{\text{exp}}$ and $\hat{\delta}_m$). To account for dependence among observations of the same user or same URL, all confidence intervals reported in this paper are 95% multiway bootstrapped confidence intervals (*31*) clustered on both users and URLs. See *Supporting Information* for details.

Results

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 0.130% 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 0.019% of the time. That is, exposure to a peer sharing a URL causes sharing for

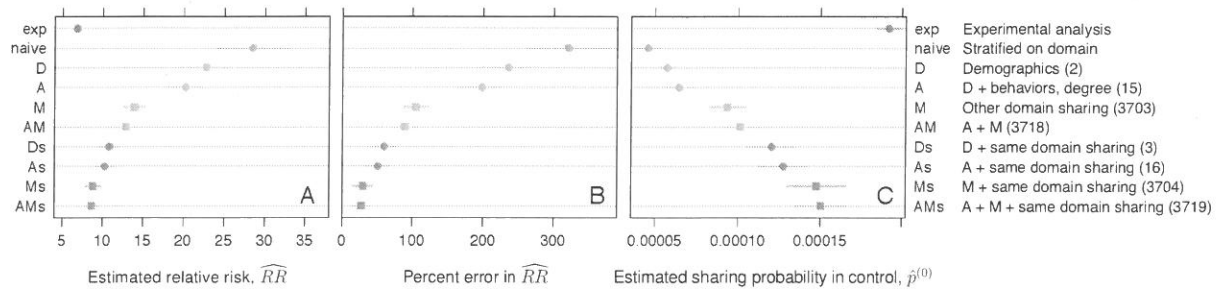

Figure 2: Comparison of experimental and observational estimates of peer effects. (A) The experiment estimates users are 6.79 times as likely to share when exposed to a peer sharing, while the observational estimates are larger. (B) Treating the experimental estimate as the truth, the naive observational estimate overestimates peer effects by 320%. This bias is substantially reduced by adjusting for prior same domain sharing (magenta) and prior sharing for 3,703 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.

0.111% of pairs (CI = [0.109, 0.114]). Users are 6.79 times as likely to share a URL in the *feed* condition compared to those in the *no feed* condition (CI = [6.54, 7.04]). These are the experimental estimates of peer effects to which we compare observational estimates.

Many studies of social contagion quantify peer effects in terms of relative risk (Figure 2A), so it is striking that the naive analysis, which makes no adjustment for observed covariates, concludes that exposure makes uses sharing 28.5 times as likely (CI = [24.1, 33.0]). That is, it overestimates peer effects by 320%. Estimates adjusting for covariates through propensity score modeling and stratification significantly reduced bias, though by how much varied substantially with the inclusion of prior behaviors (Figure 2B). Only including demographic (D) or basic individual-level covariates (A) reduced error in $\hat{\delta}$ by only 7.8% and 12.9% over the naive estimate. That is, these analyses overestimated peer effects by 235% and 197%.

Including related prior behaviors results in substantial bias reduction. Those that additionally adjusted for prior same domain sharing behavior reduced error by 51.6% (Ds) and 55.9% (As), thus overestimating peer effects by 60.0% and 50.1%. Finally, the model with all 3,719 co-

variates (AMs), including prior sharing for all domains, reduced error by 71.4% from the naive estimate. This corresponds to concluding that exposure makes sharing 8.68 times as likely (CI = [7.81, 9.54]), which is much closer to the experimental estimate of 6.79. That is, this full model overestimates peer effects by 27.8%.

Estimates of average peer effects in terms of the risk difference δ are bounded from above by $\hat{p}^{(1)} = 0.00130$ (i.e., if $p^{(0)}$ is estimated to be zero). Thus, bias in $\hat{\delta}_m$ can be measured relative to this maximum possible overestimate. The naive analysis overstates δ by 76.2% of what is possible (CI = [75.9, 79.5]). High-dimensional propensity score modeling with the full model AMs overestimates δ by 21.8% of this upper bound (CI = [13.7, 29.8]).

Error by popularity of related behaviors

The multiplicity of behaviors in this study allows examining how observational estimators perform for different behaviors; in particular, since most of the bias reduction results from measures of closely related prior behavior, we expect this bias reduction to depend on the prior prevalence of these behaviors. Many studies aim to estimate peer effects' contribution to the early stages of the spread of new behaviors (e.g., new product adoption) (*16, 19*), so it is important to evaluate methods in these cases. We analyze domain-specific estimates of peer effects according to how many unique users shared URLs from that domain during the six months prior to the experiment. A local linear regression is fit to the domain-specific estimates of $p^{(0)}$ for each model, yielding estimates of relative risk of sharing as a function of prior popularity (Figure 3). For previously unpopular domains (e.g., 0–20th percentiles), all of the observational estimators are similarly biased, but as prior popularity increases, substantial differences among the estimates emerge. The domains that were most popular before the study are also popular during the study: the top 5% of domains by unique prior sharing users contribute 33.8% of exposed user–URL pairs. Thus, much of bias reduction for the overall estimates (Figure 2) can be attributed to bias

reduction for the most popular domains.

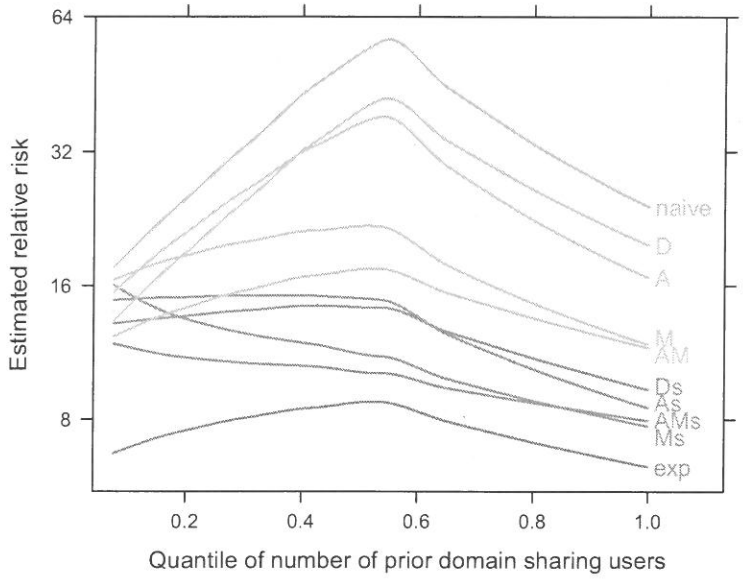

Figure 3: Estimates of relative risk of sharing by quantiles of the number of unique users sharing URLs from that domain in the prior six months. Domain-specific estimates were combined using local linear regression (loess) with bandwidth selected by 10-fold cross-validation. For previously unpopular domains, all of the observational analyses are similarly biased. Estimates for URLs from domains popular prior to the experiment exhibit more error reduction when using prior same domain sharing or prior sharing from other domains.

Discussion

The study of social contagion, while central to the social sciences, has been limited by biases that have been difficult to quantify. Field experiments are a promising solution to these problems (*11,28,32*), but restricting scientific inquiry about social contagion to questions answerable with field experiments would severely limit research in this area (*33*): few organizations are able to run experiments with sufficient 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 social contagion to important events. Previous work has noted that naive estimates of peer effects are often much larger than adjusted estimates using propensity score models (*12*), but lacked a standard against which to then evaluate these adjusted estimates. We conducted the first evaluation of observational studies of social contagion to use a large field experiment as its comparison point. Treating 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. We note that this bias reduction depends on the presence of meaningful variation in this prior behavior; 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 can be improved through both substantive knowledge and high-dimensional statistical learning techniques. Scientists studying social contagion can use existing knowledge to inform the selection and construction of relevant measures for adjustment. In many large data sets, measures of prior behaviors like those used in this study are available. Most directly, many studies of social contagion in online communication (e.g., peer effects in word use, hashtags, or link sharing) can collect similar data for prior behaviors. These methods are also applicable to a number data sets relating to offline behavior, such as purchase history data in studies using data from retail loyalty cards (i.e., "scanner data") or history of drug prescriptions by individual doctors (*19*). Scalable methods for adjusting for many variables, such as high-dimensional propensity score modeling and stratification, should encourage scientists to measure a larger number of prior behaviors. In the present study, adding thousands of additional variables reduced bias beyond simpler models with more stringently curated variables. This highlights the value of modern statistical learning techniques with large data sets for

causal inference, rather than just description and prediction. Scientists can use these methods to reduce bias in other large studies of social contagion for a wide range of behaviors of interest.

This study is not without limitations. 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. We also presented results only for "in-kind" peer effects, where exposure to one behavior causes that behavior; these results depend on the individuation of the behaviors (e.g., unique URLs), and other research may wish to estimate other cross-URL peer effects. 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 social contagion of obesity (*18*) takes place over a long period of time and is the result of peer effects in many contributing behaviors (e.g., diet, exercise). Some behaviors may not be as predictable from prior behaviors, and even when a behavior is predictable from prior behaviors, it may be difficult for investigators to determine *a priori* what prior behaviors they should measure. Finally, we limited our analysis to the average effect of a single adopting peer, in large part because these cases constituted most of the observed exposures. Observational methods may have different performance when estimating other quantities, such as the effects of multiple peers adopting a behavior, which would be useful in distinguishing simple and complex contagion (*2, 4, 9*) or studying effects of the structural diversity of adopting peers (*16*), or peer effects for specific influencers (*13, 32*).

Given these sources of variation in bias and bias reduction, others who are able to run field experiments should use them to evaluate observational methods, thereby building up more knowledge about the credibility of observational studies of social contagion. Likewise, evaluating observational methods requires considering how the resulting estimates of peer effects are used. Even the best-performing observational estimators in this study still overestimated peer effects. How problematic this bias is depends on the specific actions to take as a result of these estimates (i.e., retaining or rejecting a theory, making a change to a marketing strategy).

Qualitatively, an exposed individual being 8.68 times as likely to propagate a link is similar to 6.79 times as likely, but such a difference could matter for computing return on investment from, e.g., a public health marketing campaign.

References and Notes

1. L. E. Blume, *Games and Economic Behavior* **11**, 111 (1995).
2. D. Centola, M. Macy, *American Journal of Sociology* **113**, 702 (2007).
3. M. Granovetter, *American Journal of Sociology* **83**, 1420 (1978).
4. A. Montanari, A. Saberi, *Proceedings of the National Academy of Sciences* **107**, 20196 (2010).
5. S. E. Asch, *Psychological Monographs: General and Applied* **70**, 1 (1956).
6. M. Sherif, *The Psychology of Social Norms* (Harper, New York, 1936).
7. B. Sacerdote, *Quarterly Journal of Economics* **116**, 681 (2001).
8. S. E. Carrell, R. L. Fullerton, J. E. West, *Journal of Labor Economics* **27**, 439 (2009).
9. D. Centola, *Science* **329**, 1194 (2010).
10. S. Aral, D. Walker, *Management Science* **57**, 1623 (2011).
11. E. Bakshy, D. Eckles, R. Yan, I. Rosenn, *Proceedings of the ACM conference on Electronic Commerce* (ACM, 2012).
12. S. Aral, L. Muchnik, A. Sundararajan, *Proceedings of the National Academy of Sciences* **106**, 21544 (2009).

13. E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts, *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11 (ACM, 2011), pp. 65–74.

14. L. Coviello, *et al.*, *PLoS ONE* **9**, e90315 (2014).

15. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)* .

16. J. Ugander, L. Backstrom, C. Marlow, J. Kleinberg, *Proceedings of the National Academy of Sciences* **109**, 5962 (2012).

17. S. Wu, J. M. Hofman, W. A. Mason, D. J. Watts, *Proceedings of the 20th international conference on World wide web*, WWW '11 (ACM, 2011), pp. 705–714.

18. N. A. Christakis, J. H. Fowler, *N Engl J Med* **357**, 370 (2007).

19. R. Iyengar, C. Van den Bulte, T. W. Valente, *Marketing Science* **30**, 195 (2011).

20. M. McPherson, L. Smith-Lovin, J. M. Cook, *Annual Review of Sociology* **27**, 415 (2001).

21. C. R. Shalizi, A. C. Thomas, *Sociological Methods & Research* **40**, 211 (2011).

22. T. J. VanderWeele, *Sociological Methods & Research* **40**, 240 (2011).

23. A. C. Thomas, *Statistics in Medicine* **32**, 581 (2013).

24. E. Cohen-Cole, J. M. Fletcher, *BMJ* **337**, 2533 (2008).

25. R. H. Dehejia, S. Wahba, *Review of Economics and Statistics* **84**, 151 (2002).

26. R. J. LaLonde, *The American Economic Review* **76**, 604 (1986).

27. J. Hill, *Journal of the American Statistical Association* **103**, 1346 (2008).

28. E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, *Proceedings of the 21st international conference on World Wide Web*, WWW '12 (ACM, 2012), pp. 519–528.

29. P. Rosenbaum, D. B. Rubin, *Biometrika* **70**, 41 (1983).

30. P. R. Rosenbaum, D. B. Rubin, *Journal of the American Statistical Association* **79**, 516 (1984).

31. A. B. Owen, D. Eckles, *The Annals of Applied Statistics* **6**, 895 (2012).

32. S. Aral, D. Walker, *Science* **337**, 337 (2012).

33. N. A. Christakis, J. H. Fowler, *Statistics in Medicine* **32**, 556 (2013).

34. This work benefitted from comments by L. Adamic, S. Aral, J. Bailenson, J. H. Fowler, W. H. Hobbs, G. Imbens, S. Messing, C. Nass, M. Nowak, A. B. Owen, B. Reeves, D. Rogosa, A. C. Thomas, and J. Ugander.

Supporting information for "Bias and high-dimensional adjustment in observational studies of social contagion"

Dean Eckles & Eytan Bakshy

1 Data

1.1 Original experiment

We analyze a large experiment that manipulated a primary mechanism of peer effects in information and media sharing behaviors. Bakshy et al. (2012) randomly assigned some user–URL pairs to be prevented from being exposed to their peers sharing that URL on Facebook. In particular, for user–URL pairs assigned to the *no feed* condition, those individuals will not see that URL in their Facebook News Feed or the feed on the profile of peers, whether or not any of their peers have shared it. On the other hand, for user–URL pairs assigned to the *feed* (i.e., control) condition, those individuals can see that URL and associated comments by their peers; of course, if their peers do not share the URL, they still will not see it. Less than 1% of all user–URL pairs that would have resulted in exposure are assigned to the no feed condition. Even for pairs in the no feed condition, individuals could still see that their peer shared the URL if, e.g., the peer sent it to them in a message or posted it to the individual's profile. We refer readers to Bakshy et al. (2012) for further details about the experiment and other analyses of the experimental data.

This experiment identifies the average effect of exposure to peer URL sharing on Facebook for user–URL pairs for which that individual would have been exposed; this quantity can be described as the average treatment effect on the treated (ATET), for each treatment of exposure to one through six peers sharing the URL. We restrict our analysis to a single peer sharing the URL. More formally, for individuals who would have been exposed to a peer sharing a URL, the experiment identifies

$$\delta = \mathrm{P}(Y_{iu} = 1 \mid E_{iu} = 1) - \mathrm{P}(Y_{iu} = 1 \mid E_{iu} = 1,\ \mathrm{do}(E_{iu} = 0)) \tag{1}$$

$$= p^{(1)} - p^{(0)} \tag{2}$$

where $E_{iu} = 1$ if and only if i would have been exposed to a peer sharing u, $Y_{iu} = 1$ if and only if i shares u, and we use the do operator to indicate setting E_{iu} to something other than the value that would have been observed.[1]

[1] If one assumes that exposure via Facebook News Feed and profile feeds is the exhaustive, deterministic mechanism by which a peer sharing a URL on Facebook Z_{iu} affects whether the ego shares that behavior, then this experiment would also identify

$$\mathrm{P}(Y_{iu} = 1 \mid Z_{iu} = z) - \mathrm{P}(Y_{iu} = 1 \mid Z_{iu} = z,\ \mathrm{do}(Z_{iu} = 0))$$

because we always have $E_{iu} = Z_{iu}$ and thus this is equal to (1). We can be sure that this assumption is not strictly true. Some individuals can fail to be exposed even when peers share a URL, so the relationship between Z and E is stochastic, rather than deterministic. There can also be other ways that peer sharing can affect ego sharing besides exposure; however, it may be the case that, especially for weak ties, exposure via News Feed and profile feeds is almost an exhaustive mechanism of peer effects in URL sharing.

We restrict our analysis to Facebook users believed to be located in the United States and using Facebook in American English and to the 3,704 domains with at least 10,000 individual–URL pairs in the experimental data set.[2] This results in an experimental data set with 35 million users, 7.1 million URLs, and 71 million user–URL pairs exposed to a peer sharing the URL; this is the treated group used in both the experimental and observational analyses. The experimental control group has 49 million users, 9.5 million URLs, and 142 million pairs. The full non-experimental control group includes 67 million individuals, 11 million URLs, and 652 million pairs.

1.2 Nonexperimental control group

We constructed a nonexperimental control group (NECG)[3] with approximately ten-times the number of user–URL pairs in the experimental data set.[4] The full NECG is constructed so as to have a similar marginal distribution of individuals and URLs as the exposed group. That is, URLs appear in the NECG a number of times proportional to how many times each appears in the experimental data set. To form user–URL pairs from this set of repeated URLs, individuals were then sampled with probability proportional to the number of times they appear in the experiment. In expectation, this procedure produces a NECG with users and URLs with the same marginal distribution of characteristics as the exposed group. Thus, the potential source of bias in the observational estimates is in the pairing of users and URLs, not, e.g., in marginal distribution of user characteristics.

We constructed a NECG to be approximately 10 times the size of the combined treated and experimental control group. We did this since subsequent analysis using propensity scores would result in substantially down-weighting many of these user–URL pairs. The NECG includes 652 million user–URL pairs, compared with 70.9 million pairs in the experimental treatment (feed) group and 142 million in the experimental control (no feed) group.

2 Analysis

2.1 Stratification on estimated propensity scores and domain

The primary method used is granular stratification on estimated propensity scores and domain (Rosenbaum and Rubin, 1983, 1984; Rubin, 1997). The unknown true propensity score

$$e(X_{iu}) = \mathrm{P}(E_{iu} = 1 \mid X_{iu})$$

is the probability of an individual i being exposed to URL u, where X_{iu} are variables describing that user–URL pair. We estimate propensity scores using L_2-penalized logistic regression (i.e., logistic ridge regression) using different sets of predictors. Since the true model for peer and ego behavior is expected to be heterogeneous across very different URLs, we fit a separate model for each domain. So for model m, the estimated propensity score for i being exposed to a URL u from domain d is

$$\hat{e}_{md}(X_{iu}) = \mathrm{logit}^{-1}\left(X'_{iu}\hat{\beta}_{md}\right).$$

The is done for each of the models described in Section 2.3.

[2] This is a subset of the data used by Bakshy et al. (2012), which included data for users from all countries and language settings.

[3] In the context of methods in which individual treated and control units are matched with each other, this is sometimes called a *reservoir*.

[4] This is approximate because, for computational reasons, the sampling method used waited until the final step to filter out pairs that were actually exposed.

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 or subclasses. We use post-stratification (i.e., subclassfication) on the estimated propensity scores. Such stratification can also be regarded as form of nonparametric weighting or a form of 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 data sets, 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.[5] Additionally, as discussed below in Section 2.5, the best available method for producing confidence intervals — a bootstrap strategy — is inconsistent for nearest neighbor matching, rather than stratification, on propensity scores. For smaller data sets without this dependence structure, other recent developments, such as direct matching on many covariates could be preferable (Diamond and Sekhon, 2013).

To construct the strata, we compute the percentiles of the estimated propensity scores for each user–URL pair within each domain. These percentiles are then used as the boundaries for stratification of pairs for each domain. So for each model m, domain d, and $j \in \{1, 2, ..., 100\}$ we have an interval $\hat{Q}_{mdj} \subset [0, 1]$ of the scores in the jth percentile.[6] The strata-specific probability of sharing is estimated with a simple average of the outcomes for all the unexposed pairs in that strata

$$\hat{p}^{(0)}_{dmj} = \frac{1}{n^{(0)}_{mdj}} \sum_{\langle i,u \rangle \in \mathrm{C(d)}} Y_{iu} \mathbf{1}[\hat{e}_{md}(X_{iu}) \in \hat{Q}_{mdj}]$$

where $C(d)$ is the set of user–URL pairs in the NECG from domain d. The estimate for a particular domain d for model m is an average of the estimates for each strata weighted by the number of exposed pairs within that strata

$$\hat{p}^{(0)}_{dm} = \sum_{j=1}^{100} \frac{n^{(1)}_{mdj}}{n^{(1)}_{d}} \hat{p}^{(0)}_{dmj}.$$

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

$$\hat{p}^{(0)}_{m} = \sum_{d} \frac{n^{(1)}_{d}}{n^{(1)}} \hat{p}^{(0)}_{dm}.$$

This weighted average of domain-specific estimates[7] is then used to estimate the other quantities of interest (e.g., δ, RR) in combination with the estimate of $p^{(0)}$ common to both the experimental and observational analyses.

We set the L_2 penalty $\lambda = 0.5$.[8] The effect of the penalty on the resulting estimates is expected to be small for two reasons. First, the estimated scores are used for stratification, so only the rank of

[5] Rosenbaum and Rubin's (1984) original presentation of stratification on estimated propensity scores used quintiles, as have many applications since. As the precision of the estimated propensity scores increase (i.e., with more observations), there can be substantial within-strata confounding (Lunceford et al., 2004) that can be reduced by using a larger number of strata.

[6] For some of the simpler models, discreteness in the estimate mean there are not 100 unique percentiles.

[7] We write this estimate as a weighted average of domain-specific estimates, but it can also be written as a weighted average of domain-strata-specific estimates or a weighted average of the individual observations

[8] These models were fit with LIBLINEAR (Fan et al., 2008), in which λ is specified by setting a parameter C, where $\lambda = 1/2C$.

the scores matters for the analysis; thus, the size of the penalty primarily serves to control how much more small principal components are shrunk than the larger principal components.[9] Second, most of the models have many more observations than input dimensions; even for models with M, since this matrix is sparse, for domains with a small number of observations, only some of the columns have any non-zero values. Thus, changes to λ are expected to produce only small changes to the estimates. Analysis of other penalties $\lambda \in \{0.1, 0.5, 5, 50\}$ (not shown) was consistent with this expectation.

2.2 Naive analysis

For the sake of comparison, we also conduct a more basic analysis that does not utilize propensity scores or other adjustment. To estimate the probability of sharing for unexposed user–URL pairs, we simply compute the proportion of user–URL pairs in the NECG that shared the URL for each domain. For analyses of multiple domains, we average these estimates, weighting each by the number of exposed user–URL pairs for that domain. Because the method by which the NECG was constructed approximated the marginal distribution of users from the exposed group, this approach can be seen as finding unexposed individuals similar to the exposed individuals, but without any adjustment for propensity to be exposed to different URLs. In the subsequent analysis, we refer to the resulting estimates as the *naive observational estimates*.

2.3 Variable selection and model specification

There are numerous variables available for the propensity score model.[10] It is not possible or desirable to include all the variables that an analyst could construct because of the work involved in defining variables, the costs of increasing dimensionality for precision, and computational challenges in using all of them in an analysis. Furthermore, many situations may require that investigators decide in advance what variables are worth measuring. In both of these cases, it is standard practice to use theory and other domain knowledge to select variables. In the case of peer effects in URL sharing, the analyst would select variables believed to be related to causes of sharing a URL and to be associated with network structure (i.e., peer and ego variables are associated because of homophily, common external causes, and prior influence).[11] Table 1 lists the variables we computed based on these expectations. These variables are each included in at least one of 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 A includes all of the base variables. This model is expected to have the largest potential for bias reduction but to also suffer from increased sampling variance. In other settings, many of these variables might not be available to analysts.

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. D can also be seen as a relatively minimal convenience selection of covariates.

We consider two additional sets of predictors that can be combined with these base models. First, we expected that, by virtue of serving as measures of a user's latent interest in and likelihood

[9] Note that with a univariate linear ridge regression, the penalty has no effect on the ranks of the scores. Similarly for linear regression on an orthonormal basis.

[10] For example, an analyst could construct the *individual–term matrix* counting all the words used by each individual in their Facebook communications; each of these thousands of variables could be used as a covariate.

[11] This is not to say that the analyst must think each variable is a likely cause of sharing behaviors, but simply that they are causes of sharing behaviors *or* are descendants of these causes.

Table 1: Variables included in models predicting exposure. The final column indicates which base model specification include that variable. Some variables are transformed and/or contribute multiple inputs (columns) to a model. †: Includes untransformed and squared terms, x and x^2; *: Transformed with $\log(x+1)$; ‡: Includes binary indicator, $1\{x > 0\}$. All other variables are untransformed; if categorical, one indicator (dummy) for each level is included in the model matrix.

Category	Name	Description	Columns	Models
Demographics	Age†	As indicated on profile	2	A, D
	Gender	Indicated or inferred: female, male, or unknown	2	A, D
Facebook	Friend count	Number of extant friendships	1	A
	Friend initiation	Number and proportion of extant friendships initiated	2	A
	Tenure*	Days since registration of account	1	A
	Profile picture	Whether the user has a profile picture	1	A
	Visitation freq.	Days active in prior 30, 91, and 182 day periods	3	A
Communication	Action count*	Number of posts (including URLs), comments, and likes in a prior one month period	1	A
	Post count*	Number of posts (including URLs) in a prior one month period	1	A
	Comment count*	Number of comments on posts in a prior one month period	1	A
	Like count*	Number of posts and comments "liked" in a prior one month period	1	A
Total link sharing	Shares*‡	Number of URLs shared in a one month period	2	A
	Unique domains*	Number of unique domains of URLs shared in a six month period	1	A
Focal domain sharing	Same domain shares*‡	Number of times shared any URL with the same domain as outcome URL in six month period	2	s
Other domains sharing	Other domain shares*	Number of times shared any URL in six month period for each of the other domains	3,703	M

of independently encountering a URL, variables describing prior interactions with 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 six months prior to the experiment that have the same domain name as u. Models that add this variable are indicated with s; for example, Model As adds same domain shares to Model A. This allows for straightforward evaluation of the consequences of using this variable to the observational analysis.

We regard same domain shares as an example of more specific information about related prior behaviors. 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 recently popular, then this information may be limited. In the present case, very few users may have shared any URLs from a particular domain during the prior six months; that is, same domain shares can be 0 for most or all users for some domains.

For this reason, we also evaluate models that include the number of times a user shared URLs from each of the other 3,703 domain names; we indicate the presence of these predictors with M, as this corresponds to the addition of a large sparse matrix of (log-transformed) counts. These models have important similarities with the use of low-rank matrix decomposition methods in, e.g., recommendation systems: the L_2 penalty results in shrinking larger principal components of the training data less, where many matrix decomposition methods would simply select a small number of components to use to represent the tastes of individuals (Hastie et al., 2008, §3.4.1).

2.4 Evaluation

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 and the risk difference δ (i.e., the average treatment effect on the treated, ATET). For each analysis m, we have two estimators, $\widehat{RR}_m$ and $\hat{\delta}_m$. We can characterize discrepancies with the experiment in multiple ways. 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:

$$\widehat{RR}_m - \widehat{RR}_{\text{exp}}.$$

To put this is relative terms, we can compute the relative percent bias in the relative risk:

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

Percent changes in error from, e.g., from adjusting for covariates by going from m to m' are given by

$$\frac{|\widehat{RR}_m - \widehat{RR}_{\text{exp}}| - |\widehat{RR}_{m'} - \widehat{RR}_{\text{exp}}|}{\widehat{RR}_{\text{exp}}}$$

For the risk difference, we can also compute absolute and percent bias, similarly to above. Since the risk difference $\delta = p^{(1)} - p^{(0)}$ is bounded from above by $p^{(1)}$, the maximum possible overestimate of δ is too large by $p^{(0)}$. Thus, we can also characterize error in terms of this maximum possible overestimate, percent bias of the maximum possible

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

where we assume $\hat{\delta}_m \geq \hat{\delta}_{\text{exp}}$.

2.5 Confidence intervals

Our observations of both exposure and sharing are not independent and identically distributed (IID). Individuals vary in their probabilities of exposure and sharing, as do URLs. Exposure and sharing events are dependent, since an individual using Facebook at a particular time can often result in exposure to multiple URLs, and one person sharing a URL affects multiple others' exposure status. Methods for computing confidence intervals that neglect this dependence structure are expected to be substantially anti-conservative; that is, they would substantially overstate our confidence about the probability limit of each estimator.

To address this issue, all statistical inference in this paper employs a bootstrap strategy for data with this crossed structure (Brennan et al., 1987; Owen, 2007; Owen and Eckles, 2012). For each of $R = 100$ bootstrap replicates, we reweight observations according to the following procedure (Owen and Eckles, 2012).[12] For the rth replicate, each individual is assigned a Bernoulli(0.5) draw, and each URL is assigned a binary random variable, a Bernoulli(0.5) draw. 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. All procedures are applied to the original data set and each of the replicates, such that each propensity score model is fit $R + 1 = 101$ times, percentiles of estimated propensity scores for each domain are computed 101 times, etc. Under general conditions, this strategy is known to be conservative when estimating the variance of means (Owen, 2007; Owen and Eckles, 2012). Throughout, we report 95% bootstrap standard confidence intervals, which are expected to have at least 95% coverage due to variable-level duplication (Owen and Eckles, 2012). The bootstrap distribution of matching and resulting estimates is generally inconsistent because matching estimators do not satisfy required smoothness conditions for bootstrap validity, resulting in mildly incorrect confidence intervals (Abadie and Imbens, 2008). This is one motivation for using stratification, rather than one-to-one matching.

Note that all of the comparisons of interest are not entirely between-units. For example, the observational and experimental estimates share individuals, URLs, and even user–URL pairs. Observing that confidence intervals for two quantities overlap does not indicate that their difference (or ratio) is not statistically significantly different from zero (or one). This is one reason why we include figures showing estimates and intervals for relevant differences and ratios themselves (i.e., the quantities in Section 2.4).

2.6 Analysis by prior popularity

We present analyses of how the domain-specific estimates vary with the prior popularity of the domain. To summarize this relationship, we used local linear regression (loess) to estimate expected $p^{(0)}$ for all levels of prior popularity and model choice and $p^{(1)}$ for all levels of prior popularity. This smoother was implemented using `loess` in R with a tricubic kernel. The width of the kernel was selected through the following cross-validation procedure applied to the fits to estimates of $p^{(0)}$. Domains were randomly partitioned into 10 folds. The model was fit to 10 subsets of the data, each leaving out one fold. Mean-squared error (MSE), weighted by the number of exposed user–URL pairs, was computed for the out-of-sample domains for each value of the parameter $\alpha \in \{0.3, 0.4, 0.5, ..., 1.5\}$. The selected kernel width $\alpha = 1.0$, minimizes this weighted MSE averaged over all models (Figure 1).

[12] Software implementing this multiway bootstrap in R is available at `https://github.com/deaneckles/multiway_bootstrap`. The results in this paper were produced by similar software for Apache Hive.

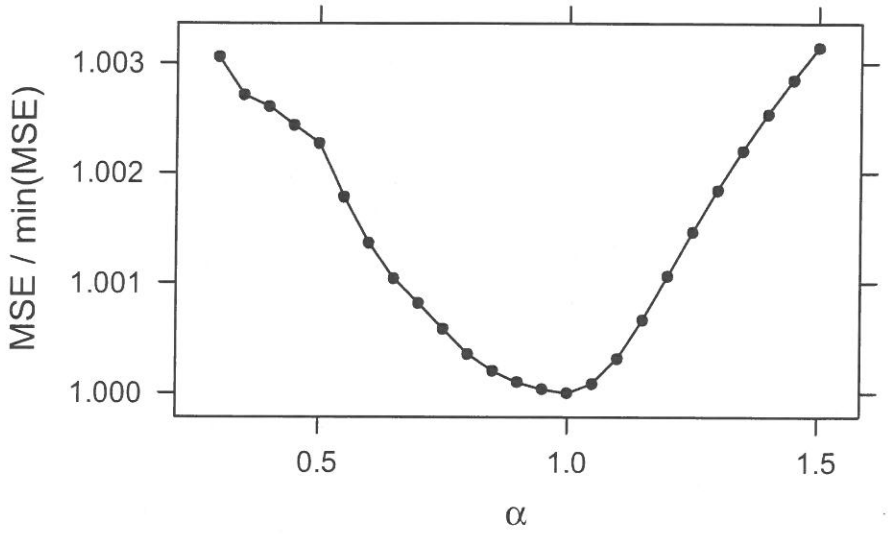

Figure 1: Mean-squared error in 10-fold cross-validation for model-specific, domain-specific estimates $\hat{p}_{dm}^{(0)}$ averaged over all models.

3 Supplementary results

3.1 Estimates for individual domains

We now consider the estimates of peer effects via News Feed for individual domains of interest. The subgroup analysis by domain popularity showed how important the most popular domains are for the overall estimates, so this motivates further scrutiny of estimates for the these domains. In particular, the following analysis examines our estimates for the 15 domains with the largest number of exposed user–URL pairs during the experiment.[13] We can treat these domains individually because the amount of available data means that we can produce precise peer effect estimates for each, making comparison of experimental and observational estimates possible. They also provide a useful variety of types of URLs that individuals can share.

The probabilities of sharing a URL for user–URL pairs in the feed and no feed conditions are shown in Figure 2. These top domains exhibit substantial heterogeneity in both probabilities. This is, at least in part, attributable to differences in other opportunities for discovering these URLs. For example, URLs at `www.nytimes.com` have a comparatively high probability of being shared, even without exposure via News Feed to a peer sharing the URL; in fact, this proportion is larger than the proportions *with* exposure for several of the other domains. This variation suggests one source for subsequent differences in estimates of peer effects.

We now compare the observational estimates peer effects for each domain (Figure 3 and 4). The pattern of estimates for most of the domains is consistent with the expectation that the observational

[13] For this analysis, we exclude two domains that are URL shortening services, as these domains are a very different sort of grouping of URLs than the other domains. The observational methods generally fail to achieve much error reduction for these URLs, when regarding the experimental estimates as the gold standard. However, for reasons related to those discussed in Section 3.2, the experimental estimators for peer effects for pairs with these domains probably do not estimate quantities of primary interest.

Table 2: Estimates for each model with 95% bootstrap standard confidence intervals in brackets.

Model	$\hat{p}^{(0)}$	$\widehat{RR}$	$\hat{\delta}$
AMs	1.502E-04 [1.343E-04, 1.661E-04]	8.68 [7.82, 9.53]	1.153E-03 [1.123E-03, 1.183E-03]
Ms	1.477E-04 [1.295E-04, 1.658E-04]	8.83 [7.85, 9.80]	1.155E-03 [1.127E-03, 1.184E-03]
AM	1.014E-04 [8.971E-05, 1.132E-04]	12.85 [11.63, 14.07]	1.202E-03 [1.175E-03, 1.228E-03]
M	9.368E-05 [8.317E-05, 1.042E-04]	13.91 [12.59, 15.23]	1.209E-03 [1.185E-03, 1.234E-03]
As	1.274E-04 [1.128E-04, 1.421E-04]	10.23 [9.08, 11.38]	1.176E-03 [1.150E-03, 1.202E-03]
Ds	1.203E-04 [1.052E-04, 1.353E-04]	10.83 [9.56, 12.11]	1.182E-03 [1.157E-03, 1.208E-03]
A	6.460E-05 [5.722E-05, 7.198E-05]	20.17 [17.93, 22.42]	1.238E-03 [1.214E-03, 1.263E-03]
D	5.727E-05 [4.938E-05, 6.515E-05]	22.76 [19.77, 25.75]	1.246E-03 [1.221E-03, 1.270E-03]
naive	4.567E-05 [3.842E-05, 5.291E-05]	28.54 [24.12, 32.95]	1.257E-03 [1.233E-03, 1.282E-03]
exp	1.920E-04 [1.845E-04, 1.995E-04]	6.79 [6.54, 7.04]	1.111E-03 [1.086E-03, 1.136E-03]

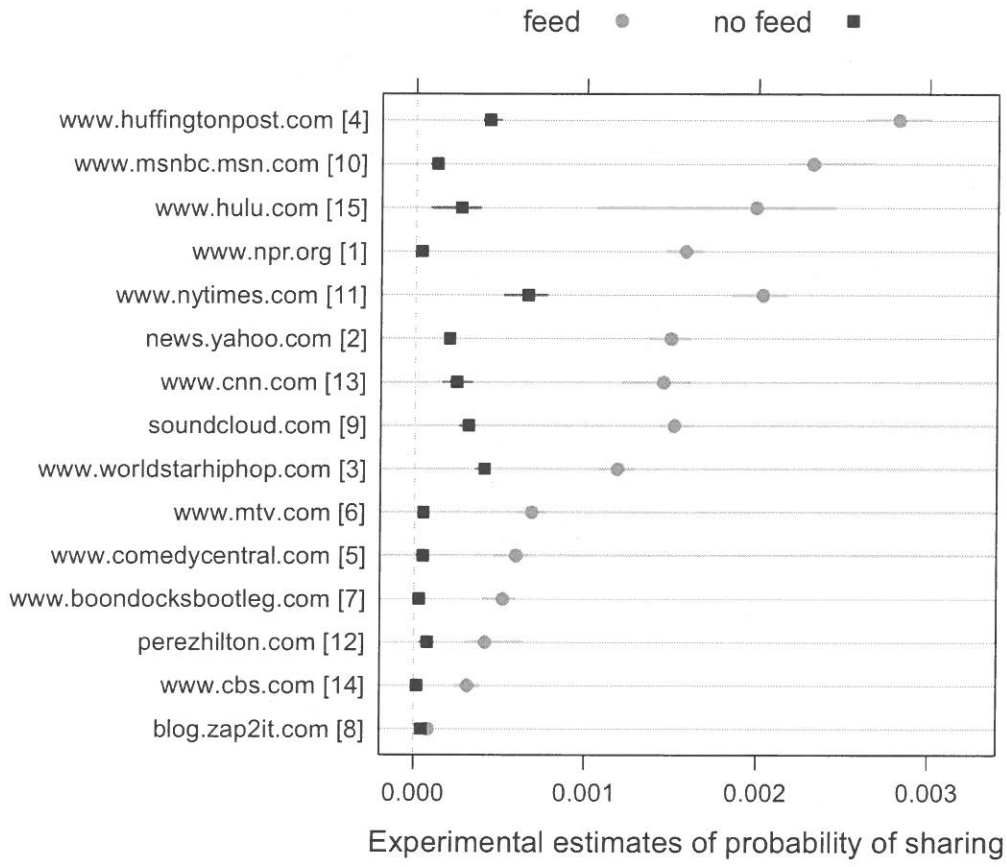

Figure 2: Probability of sharing a specific URL in the feed and no feed experimental conditions for the 15 domains with the largest number of exposed user–URL pairs; numbers in brackets indicate this rank. Domains are sorted by the difference between the two conditions (i.e., the experimental estimate of the risk difference). Error bars are 95% bootstrap confidence intervals from reweighting both users and URLs.

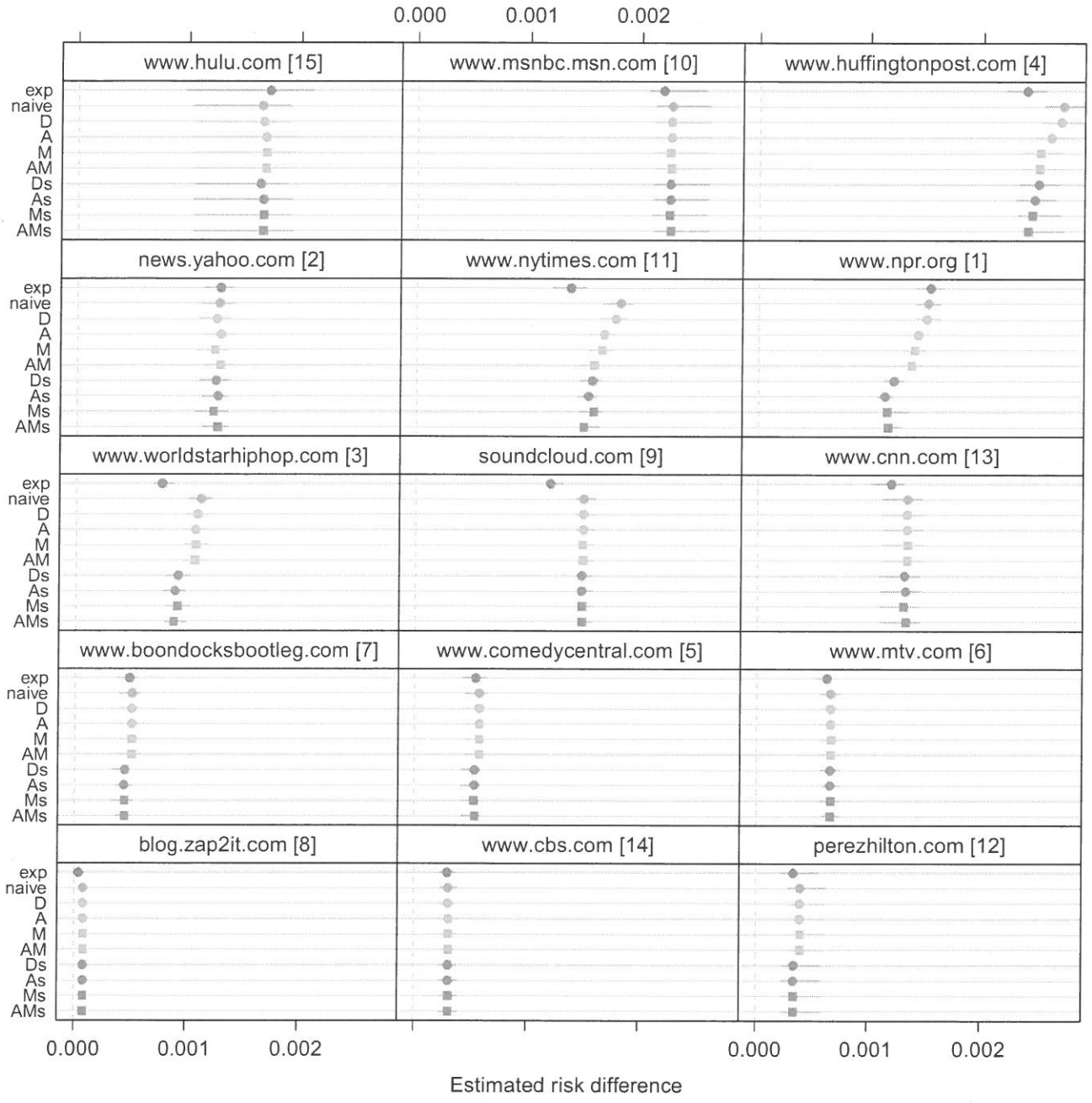

Figure 3: Estimated peer effects via News Feed using stratification on propensity scores for the 15 domains with the largest number of exposed user–URL pairs; numbers in brackets indicate this rank. Domains are sorted by the experimental estimate. Error bars are 95% bootstrap confidence intervals from reweighting both users and URLs.

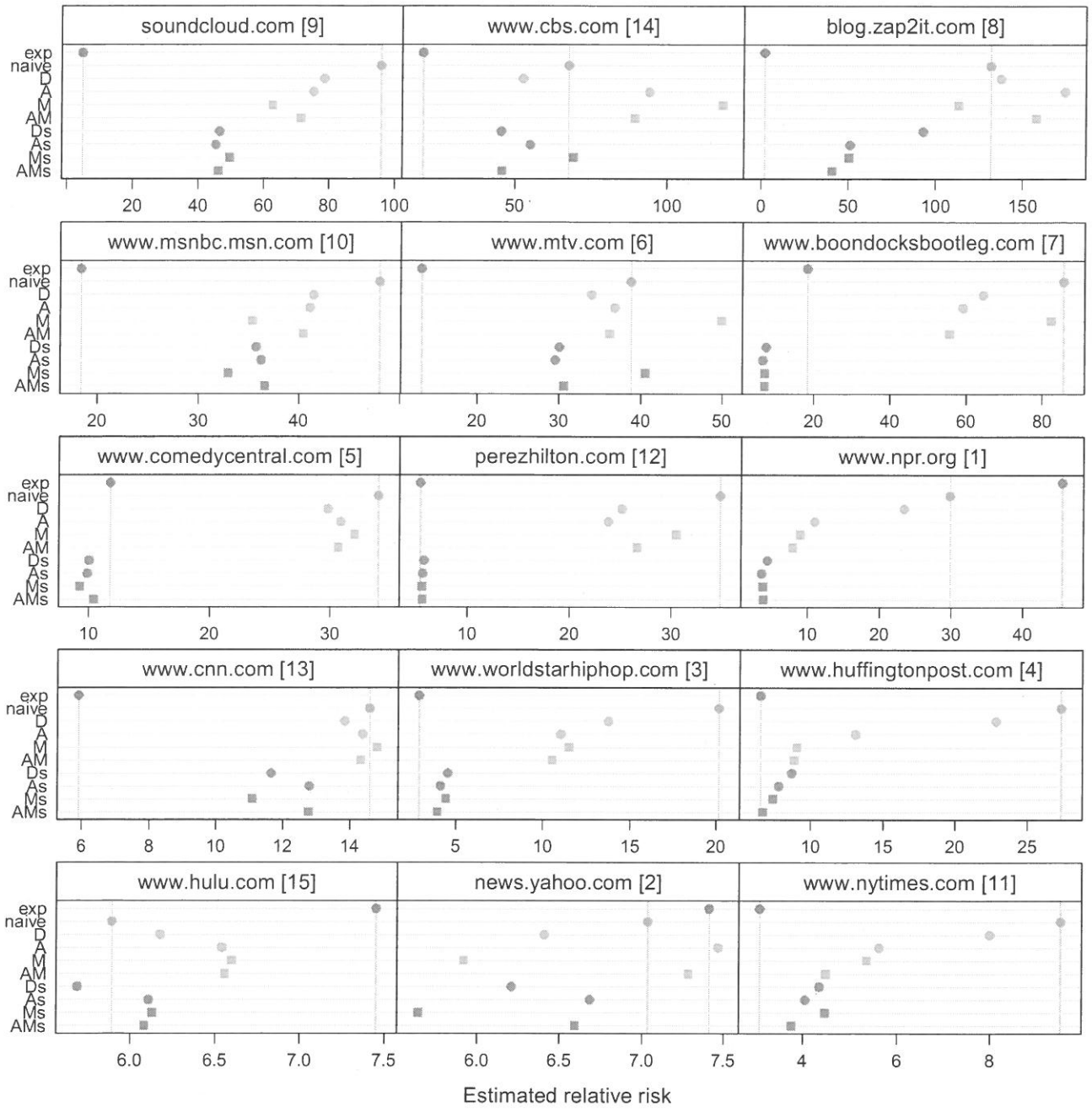

Figure 4: Estimated peer effects in terms of relative risk for the 15 domains with the largest number of exposed user–URL pairs; numbers in brackets indicate this rank. Note the very large experimental estimate for www.npr.org.

analyses will produce larger estimates of peer effects. Compared to the experimental estimates, we find that the naive and estimates of the risk difference are significantly larger. On the other hand, there some domains for which the point estimates of this difference suggest that the observational estimates might exhibit *negative* bias. In particular, for `www.npr.org` the naive estimate is marginally smaller than the experimental estimate, and the point estimates from propensity score stratification are also smaller.[14] This latter point is consistent with the observed pattern that these models, and propensity score stratification more generally, produce larger estimates of $p^{(0)}$ than the naive estimator; however, it is surprising both that the difference between the experimental and observational estimates has a negative sign and that the propensity score stratification apparently increases this difference.

3.2 Alternatives explanations of differences from experimental estimates

For much of the above discussion, we have regarded the experimental estimates as the gold standard. That is, we have regarded the experiment as identifying the average peer effects of interest for those who would be exposed. Thus, discrepancies between the experimental and observational estimates are then attributable to sampling variance in either and bias in the observational estimates. We expected the observational estimates to suffer from confounding bias because of selective tie formation and dissolution (i.e., homophily and heterophily), common external causes, and prior influence. Except for heterophily, these would all make it more likely for peers to share the same URLs, even in the absence of peer effects, so we anticipated that the naive observational analysis would overestimate peer effects, and that the estimators using propensity score stratification would reduce, but not eliminate or reverse, this bias. This is the primary explanation of differences between the the experimental and observational estimates. In this section, we consider two other explanations of differences between these estimates.

3.2.1 Total peer effects versus peer effects of exposure for the exposed

Even if total average peer effects are conditionally unconfounded given the covariates used in our propensity score models, the observational and experimental estimates can differ if the former consistently estimate total peer effects (i.e., effects of peer sharing via all mechanisms) and the latter consistently estimate peer effects of exposure through News Feed and profile feeds. This places an important limitation on what we can learn from this constructed observational study. We nonetheless regard studies such as this as one of the best available tools for better understanding the performance of observational methods for estimating peer effects.

We expect that while exposure via News Feed and profile feeds is not an exhaustive mechanism for peer effects in URL sharing on Facebook, it may be nearly exhaustive, since the other primary mechanism is exposure through that peer sharing the URL on Facebook and then, *because of this prior sharing decision*, sharing with the ego via some other method, such as via email, in person, or through Facebook chat. While sharing via other methods may be common, and this may be associated with sharing on Facebook, we expect that doing so as a result of having also done so on Facebook is relatively rare.

[14] This unexpected difference might be a false positive, given that we are conducting many tests here.

3.2.2 Problems in meaningfully individuating URLs

One additional reason to doubt that the experimental estimates should uniformly be treated as the gold standard concerns how URLs are individuated. While some attempts were made in the original experiment to canonicalize URLs — that is, to identify multiple URLs that correspond to the same online resource or page — this is not perfect for many uses. In particular, there are many cases where there are multiple URLs that point to essentially the same resource. At the extreme, there can be variations in the URL that are directly related to whether a user arrived at it via Facebook. For example, many `www.npr.org` URLs have the form `http://www.npr.org/templates/story/story.php?storyId=[n]` where `[n]` is a numeric identifier for the story. Some of these URLs also appear, or appear exclusively, appended with the query parameters `&sc=fb&cc=fm`. This variation in the URL does not change the primary content shown to visitors to the URL. Rather, it is apparently used by `ww.npr.org` to track whether visitors are coming to the site from Facebook. For the purpose of studying peer effects in information diffusion, media consumption and sharing, etc., treating these two URLs as distinct, such that an individual is only counted as sharing the same URL if they share a version that matches this appended set of query parameters exactly, is likely undesirable. Consider an individuals who would be exposed to a peer sharing a URL with the query parameters (i.e., $M = 1$). They might encounter the same content through other means, in which case the URL would likely not have these parameters, or have different ones, and share that URL. Under the experimental analysis in Bakshy et al. (2012) and in this chapter, they would not be counted as sharing the URL. If we would prefer to consider these to be the same URL, then this results in underestimating $p^{(0)}$ and $p^{(1)}$ and likely overestimating their difference and ratio. We regard this as likely a primary cause of the unexpected negative difference between the experimental and observational estimates for `www.npr.org`.

References

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 519–528. ACM.

Brennan, R. L., Harris, D. J., and Hanson, B. A. (1987). The bootstrap and other procedures for examining the variability of estimated variance components in testing contexts. Technical report, American College Testing Program.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.

Lunceford, J. K., Davidian, M., Lunceford, J. K., and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Morgan, S. L. and Harding, D. J. (2006). Matching estimators of causal effects prospects and pitfalls in theory and practice. *Sociological Methods & Research*, 35(1):3–60.

Owen, A. B. (2007). The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411.

Owen, A. B. and Eckles, D. (2012). Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927.

Rosenbaum, P. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41 –55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8 Part 2):757–763.