Estimating peer effects in networks with peer encouragement designs

Dean Eckles[a,b,1], René F. Kizilcec[b,c], and Eytan Bakshy[b]

[a]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Facebook, Menlo Park, CA 94025; and [c]Department of Communication, Stanford University, Stanford, CA 94305

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved April 15, 2016 (received for review June 15, 2015)

Peer effects, in which the behavior of an individual is affected by the behavior of their peers, are central to social science. Because peer effects are often confounded with homophily and common external causes, recent work has used randomized experiments to estimate effects of specific peer behaviors. These experiments have often relied on the experimenter being able to randomly modulate mechanisms by which peer behavior is transmitted to a focal individual. We describe experimental designs that instead randomly assign individuals' peers to encouragements to behaviors that directly affect those individuals. We illustrate this method with a large peer encouragement design on Facebook for estimating the effects of receiving feedback from peers on posts shared by focal individuals. We find evidence for substantial effects of receiving marginal feedback on multiple behaviors, including giving feedback to others and continued posting. These findings provide experimental evidence for the role of behaviors directed at specific individuals in the adoption and continued use of communication technologies. In comparison, observational estimates differ substantially, both underestimating and overestimating effects, suggesting that researchers and policy makers should be cautious in relying on them.

social interactions | social networks | causal inference | experimental design

Social interactions among people enable the spread of information, preferences, and behavior, including technology adoption. Despite the unprecedented availability of detailed information on human interactions, credible identification of how individuals affect each other has been difficult. Many of the empirical studies that estimate these peer effects rely on analyzing observational (i.e., nonexperimental) data (e.g., refs. 1 and 2). These methods can incorrectly "detect" peer effects in their absence (3–5) and can substantially overestimate them (6). There are many causes of correlated behaviors in networks that make it difficult to identify peer effects, including selective tie formation [i.e., homophily (7)], unobserved correlated external causes, and prior peer effects (4, 8, 9). Faced with these challenges, observational studies of peer effects are sometimes described as tentatively providing evidence of peer effects (cf. refs. 10 and 11) or as providing upper bounds, rather than point estimates, for peer effects (6).

This paper presents designs for randomized experiments for estimating peer effects in social networks that overcome common challenges to credible identification. We conducted a large field experiment on Facebook that implements a peer encouragement design to estimate peer effects in the use of communication technologies. In particular, many people share information, personal media, or other content via online social networks. Most of these services allow them to receive feedback from their peers in the form of comments on their post and expressions of approval (or disapproval). How does receiving more or less of this feedback from peers affect use of these technologies? Decision makers benefit from knowing the value of receiving social feedback, relative to other potential actions, as this informs the design of interfaces for giving feedback. For social scientists, precisely estimating the effects of feedback is important for, e.g., understanding network effects in the adoption and continued use of communication technologies.

There is some theoretical and empirical support for expecting substantial peer effects in initial adoption and use of communication technologies. Individuals' utilities from using such technologies usually depend on peer adoption decisions, as this determines who can be communicated with and the consequences of communication. Prior work on Facebook specifically (12, 13), and other related technologies (14, 15) has provided observational and quasi-experimental evidence for peer effects in initial adoption, content production, and sustained use. Other prior observational research has found that receiving feedback (e.g., comments, "likes") is associated with higher rates of sharing; in particular, new users who receive comments on their photos tend to share more photos in the future (10). However, in the presence of confounding due to homophily and common external causes, these prior observational results are expected to overstate (or otherwise misstate) the relationship between receiving feedback and subsequent behavior if interpreted causally.

Peer Encouragement Designs

Randomized experiments are one appealing way to identify peer effects in the presence of unknown confounding (16–18). Although directly randomizing the behavior of existing peers in realistic settings is generally not possible or desirable, multiple experimental designs for learning about peer effects appear in the literature. Social psychologists have used inauthentic, confederate peers since the 1950s (19, 20), often in artificial (e.g., laboratory) settings. Other studies have induced random variation in the process of tie or group formation (21–25). Although these approaches have been successful at answering some important questions in the social sciences, it is often not possible for such designs to credibly answer questions about either existing peers or effects of specific peer behaviors.

The widespread adoption of online social networks has facilitated in situ studies of the effects of peers' behaviors on individual behavior. Much of the experimental work in this area has used mechanism designs, which directly modulate mechanisms (or channels) by which information about peer behavior is optionally or nondeterministically

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Drawing Causal Inference from Big Data," held March 26–27, 2015, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Big-data.

Author contributions: D.E., R.F.K., and E.B. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

Conflict of interest statement: The authors were all employed by Facebook while conducting this research. D.E. and E.B. have significant financial interests in Facebook.

This article is a PNAS Direct Submission.

Data deposition: Analysis code and aggregate statistics for reproducing the main results (Figs. 3–5) are archived at the Harvard Dataverse Network (dx.doi.org/10.7910/DVN/ELUQVD).

[1]To whom correspondence should be addressed. Email: dean@deaneckles.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1511201113/-/DCSupplemental.

transmitted to a focal individual (ego) through the network (18, 26–28). For example, Aral and Walker (26) randomize which peers are sent viral messages to adopt a product, and Bakshy et al. (18) randomize the number of personalized social cues in advertisements. The causal directed acyclic graph (DAG) (29) shown in Fig. 1*A* illustrates a mechanism design with binary peer behaviors and ego behavior. When these designs involve enabling/disabling a mechanism of peer effects, they allow estimating an average treatment effect on the treated (ATT)—the effect of exposure to a peer behavior for those who would normally be exposed; if the mechanism is normally deterministic, this is also an ATT for the peer behavior, not just for exposure. Despite their advantages, mechanism designs are often not possible or practical in many empirical settings, such as when the mechanism is deterministic (i.e., information about a peer's behavior is always transmitted to the ego, such as feedback in an online social network).

Encouragement Designs. We develop and illustrate a variation on randomized encouragement designs for identifying effects in networks. Encouragement designs (30) are widely used by social and biomedical scientists when interested in the effects of behaviors not directly controlled by the experimenter. Units are randomly assigned to an encouragement Z_i, and the endogenous behavior of interest D_i and the outcome Y_i are measured. For example, in educational contexts, one may encourage children to watch a particular educational program (31) or prepare for tests (32). Not all parents or children may follow through with such interventions, but it is still possible to analyze the causal effect of

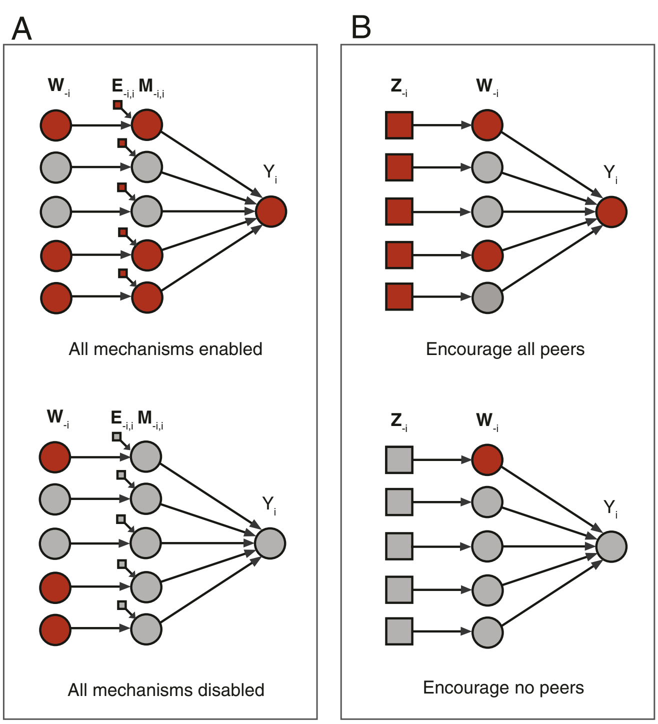


Fig. 1. Mechanism designs and peer encouragement designs for estimating peer effects, illustrated with binary variables. W_{-i} indicate peers' behaviors, and Y_i represent the behavior of a focal individual (ego). Variables are colored to represent example values under different random assignments (red = 1, gray = 0). (*A*) Mechanism designs modulate a channel by which peer effects occur, for example, by randomly enabling or disabling ($E_{-i,i}$) a particular mechanism ($M_{-i,i}$) by which a focal individual (i) is exposed to peer behavior ($-i$). (*B*) Peer encouragement designs use randomized encouragements to peers (Z_{-i}). All variables represented by circles may have other common causes not shown. Variables represented by squares are root nodes and are determined by random assignment.

the intervention for those who are induced to use the educational materials by the randomized encouragement (i.e., for compliers). For this purpose, the encouragement is treated as an instrumental variable (IV); that is, it is assumed that the encouragement only affects outcomes by affecting the intermediate, endogenous behavior of interest. This complete mediation or exclusion restriction can be stated as follows. Define the potential outcomes for Y_i and D_i as functions of the encouragement and the behavior, $Y_i : \mathbb{D} \times \mathbb{Z} \to \mathbb{Y}$ and $D_i : \mathbb{Z} \to \mathbb{D}$.

Assumption 1. (Exclusion restriction). *Suppose* $Y_i(d_i, z_i) = Y_i(d_i, z_i')$ *for all* $d_i \in \mathbb{D}$ *and* $z_i, z_i' \in \mathbb{Z}$, *so we can uniquely define* $Y_i(d_i)$.

[Combining this assumption with the random assignment of Z_i, some authors (e.g., ref. 33), write $Y_i(d_i), D_i(z_i) \perp\!\!\!\perp Z_i$ for all $d_i \in \mathbb{D}$ and $z_i, Z_i \in \mathbb{Z}$. The exclusion restriction is then combined with either parametric assumptions about $Y_i(\cdot)$ or nonparametric assumptions about $D_i(\cdot)$ to identify effects of D_i on Y_i (34, 35); both are discussed in *Model*. Here and elsewhere, we use capital letters for random variables and lowercase letters for fixed values. We retain subscripts for units even in the latter case, as those without subscripts denote n vectors.

Encouragement Designs in Groups and Networks. Peer encouragement designs randomize an individual's peers to conditions that increase or decrease the probability of those peers performing a specific behavior. One may then examine how this shock to peer behaviors "spills over" to the behaviors of focal individuals. Furthermore, these designs can provide point estimates of the effect of peer behavior on ego behavior (i.e., peer effects) by using encouragements to a specific behavior and assuming that the only effect of peer assignment to these encouragements on ego behavior is via that specific peer behavior. The causal DAG in Fig. 1*B* illustrates a peer encouragement design with binary encouragements, peer behaviors, and ego behavior. In this example, the encouragement causes one peer to adopt the specific behavior, which in turn causes the ego to adopt. Given the assumptions encoded in this DAG, peer encouragement is an IV, and we can estimate the effect of the behavior of peers, as caused by the encouragement to adopt, on ego behavior.

Plausibility of the Exclusion Restriction. This DAG encodes an exclusion restriction (Assumption 1): All effects of the peer encouragement on ego behavior occur via changes to peer behavior. In standard encouragement designs, the randomized encouragement Z_i, endogenous behavior D_i, and outcome Y_i are all defined as direct interventions on or measurements of the same individual, often making this assumption implausible because the encouragement may affect that individual in many ways (30, 36, 37). For example, parents encouraged to watch Sesame Street with their children (31) may modify their child-rearing in many ways. In peer encouragement designs, however, the exclusion restriction assumption is often particularly plausible for a structural reason: For a given ego whose outcome is observed, the encouragement is applied to other units—their peers. The ego usually does not directly observe or experience the encouragement; instead, it only affects the ego through peer behavior and, the researcher hopes to ensure, primarily through a small number of measured peer behaviors. We make the following two design recommendations—implemented in our empirical example—that can increase the plausibility of the exclusion restriction and increase statistical power.

First, provided the sample is sufficiently large, selecting a peer encouragement that is minimal may reduce the potential for reactance; we illustrate this point by comparison. Many designs that randomly assign treatment and estimate "spillovers" (i.e., interference or exogenous peer effects) (38, 39) can be understood and analyzed as peer encouragement designs. Recent work by economists and political scientists has examined such spillover effects within groups (40–44) or in a social network (45–47). In some cases, researchers have attributed the estimated spillovers from treatment assignment to a specific peer behavior: In one study (41), employees were randomly assigned to encouragements to attend a retirement benefits fair. Among other analyses, Duflo and Saez (41) attribute spillover effects on retirement plan enrollment to effects of peer attendance at the fair. However, as the authors note, this encouragement may have directly affected always-attenders [e.g., via self-perception or crowd-out effects (48)], never-attenders (e.g., via salience of benefits), and their peers (e.g., via increased discussion of benefits). If, instead, the encouragement was unlikely to be remembered or even consciously perceived as an inducement, perhaps such violations of the exclusion restriction would be less likely to occur. Thus, peer encouragement designs could provide more credible peer effect estimates if the encouragement is a minimal "nudge" that may not warrant much conscious consideration.

A second design recommendation is, when appropriate for the research question, to use encouragements that are specific to particular directed edges, rather than encouraging a general, undirected behavior in peers. The experiments mentioned above generally use the number or fraction of assigned peers as the instrument. This instrument is then necessarily correlated for all egos in the same group or, more generally, who share peers. On the other hand, it is sometimes possible to encourage directed behaviors on particular edges; that is, an encouragement that induces a behavior from an alter j to an ego i. Such an encouragement could be randomly assigned at the level of the directed edge, or at the level of the target (i.e., the ego). In the latter ego-specific design, an ego i is randomly assigned to a peer encouragement condition Z_i, according to which all edges from any alter j to ego i are treated. That is, egos are randomly assigned to conditions that encourage their peers to engage in directed behaviors toward them; those same peers might be assigned to a different condition with respect to their other peers. In this ego-specific design, the instrument is no longer correlated within groups or in the network. This design choice can substantially change power; simulations on small-world networks demonstrate the ego-specific design reducing true SEs by 20% to over 90% (*SI Appendix, Simulations with Ego-Specific and General Designs*). Here we report on a large experiment in which the peer encouragement is a minimal change that causes a specific behavior directed at a particular ego.

Empirical Context and Data

Our empirical study examines the effects of receiving feedback from peers on Facebook. In particular, we examine feedback on socially shared content (posts) such as text, photos, videos, and links shared by egos. This content appears in the News Feeds of peers (friends), who may, in turn, provide feedback on these posts by providing comments on the post or clicking on the "Like" button. Individuals who receive feedback on a post may receive notifications immediately on Facebook, or via mobile notifications or email.

The design of a feedback interface poses a complex tradeoff: An interface that causes an ego's post to occupy more space in their peers' News Feeds may increase the likelihood that peers will provide feedback on the post; at the same time, such an interface may cut into peers' limited time and attention to view and interact with others' posts. To choose among interfaces, Facebook product teams frequently randomly assign some users to receive an alternative (often new) version of an interface to evaluate these alternatives; the data presented here arise from one such trial. In particular, we implemented a peer encouragement design that enables estimating the effects of receiving feedback when sharing content in social media.

Peers' responses to an ego's content (i.e., liking and commenting) are expected to vary with the user interface associated with that content when seen by peers. Egos were randomly assigned to conditions that encouraged their peers to provide feedback under different circumstances. There were two experimental factors that independently governed the display of egos' posts in peers' News Feeds (Fig. 2). First, the "encourage initiation" factor was relevant for posts without any feedback, and it determined whether the viewer would need to click "Comment" to display the textbox in which to write a comment or whether this textbox would be already visible. Second, the "conversation salience" factor was relevant for posts that had already received feedback, and it determined whether this existing feedback would be summarized numerically and displayed after a click or would already be visible (up to three comments shown by default). Thus, the encourage initiation factor should primarily cause the first feedback to occur at all or earlier, whereas the conversation salience factor should cause additional feedback. There were six possible conditions egos could be assigned to, resulting in a three (encourage initiation: always, sometimes, never) by two (conversation salience: high, low) design. (For the encourage initiation factor, the level sometimes was the default interface at the time: Posts displayed in the first position in News Feed would have the textbox shown, but posts appearing in other positions would not.)

This experiment thus is a peer encouragement design in which directed edges are treated according to an ego-specific random assignment: A particular person viewing their News Feed could see posts from multiple egos, which would be displayed according to the conditions to which each of those egos were assigned. We use this experiment to examine the effects of receiving feedback on how many posts egos make and how much feedback they give on others' posts. To establish a baseline for comparing effect sizes, we also estimate effects on how much they respond to feedback on their own posts. Feedback received is measured as the mean daily number of comments and likes received during the experimental period. All analysis is of deidentified data primarily consisting of counts of behaviors.

Model

For the main analysis, we work with log-transformations of the count variables (see *SI Appendix, Transformed and Untransformed Count Variables*). Let D_i be the logarithm of feedback received (likes and comments) by i during the experimental period and Y_i be the logarithm of one of the ego behaviors of interest. We aim to estimate effects of D_i on Y_i by using the random variation in D_i caused by assignment to the peer encouragement, Z_i. That is, we aim to summarize contrasts between potential

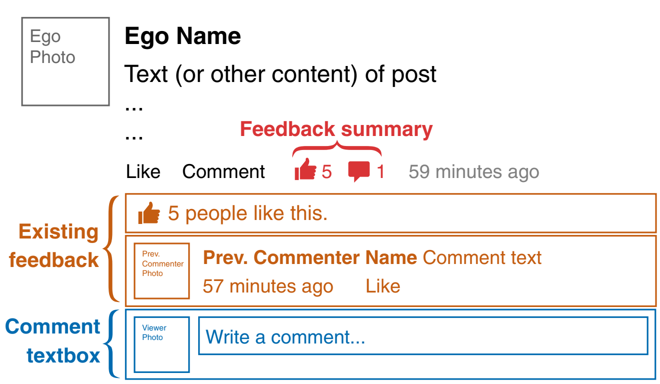


Fig. 2. Illustration of the feedback interfaces that would be used to display an ego's post to their peers according to which condition the ego was randomly assigned. For posts with feedback, the conversation salience factor determines whether the feedback is summarized numerically (red) or, instead, the existing feedback is shown (orange) with a textbox for writing a new comment (blue). In the low-salience case, a click on "Comment" or the feedback summary would display the existing feedback and comment textbox. For posts without feedback, the encourage initiation factor determines whether the comment textbox is shown by default (blue) or whether a click on "Comment" is needed to display it.

outcomes for different levels of feedback received, for example, some summary of $Y_i(d_i) - Y_i(d_i')$. Writing i's potential outcomes as functions only of the ith elements of an n-vector of feedback received requires two assumptions that specify the potential outcomes are constant in some inputs (i.e., specify level sets). First, this requires the exclusion restriction for IVs (Assumption 1). The minimal nature of the encouragement makes it plausible that it only affects egos by causing them to receive additional feedback; however, there may be effects of the feedback not captured by its quantity (e.g., content of comments, timing).

Additionally, already in writing $Y_i(d_i, z_i)$, we assume that the behaviors and assignments of all other units can be safely ignored—a "no interference" (49) or "individualistic treatment response" (50) assumption.

Assumption 2. (No interference). *Suppose that* $Y_i(d_i, d_{-i}, z_i, \mathbf{z}_{-i}) = Y_i(d_i, d'_{-i}, z_i, \mathbf{z}'_{-i})$ *for all* $d, d' \in \mathbb{D}^n$, $\mathbf{z}, \mathbf{z}' \in \mathbb{Z}^n$ *so that we can uniquely define* $Y_i(d_i, z_i)$.

This assumption is expected to be violated in this setting, even in our finite population. First, the units are interacting and make up a substantial portion of a single network. Second, the peer encouragement conditions would have different effects under a different global policy such that, e.g., peers were seeing all posts displayed according to the same interface rule. However, methods for statistical and causal inference in the presence of interference remain somewhat underdeveloped, especially for interference in a single network rather than within many isolated groups. We therefore work with the assumption that relevant nuisance interference is small compared with the effects of interest. For example, consider the assumption that this nuisance interference is no larger than than the effect of an increase c to feedback received.

Assumption 3. (Direct-effect-bounded interference). *Suppose that*

$$|Y_i(d_i, d_{-i}, z_i, \mathbf{z}_{-i}) - Y_i(d_i, d'_{-i}, z_i, \mathbf{z}'_{-i})| \leq |Y_i(d''_i + c, d''_{-i}, z''_i, \mathbf{z}''_{-i}) - Y_i(d''_i, d''_{-i}, z''_i, \mathbf{z}''_{-i})|$$

for all $d, d', d'' \in \mathbb{D}^n$, $\mathbf{z}, \mathbf{z}', \mathbf{z}'' \in \mathbb{Z}^n$.

If, as in our main analysis, feedback received is modeled on a log scale, then, for $c = 1$, this assumes that any interference is smaller than the effect of multiplying feedback received by e (i.e., increasing feedback received by 172%); thus, sensitivity analysis based on such an assumption allows for very substantial interference. In *SI Appendix*, we combine this assumption with a specific model of local interference (50, 51) to conduct analyses quantifying the sensitivity of our results to nuisance interference. For simplicity, we now proceed with a model without nuisance interference.

In addition to Assumptions 1 and 2, there are multiple sets of assumptions that allow identification and estimation using peer encouragement conditions as IVs. One such assumption is that the effects of feedback received are (log–log) linear and constant; that is,

$$Y_i(d_i) - Y_i(0) = \gamma d_i.$$

In this case, two-stage least squares (TSLS) with multiple instruments simply increases precision in estimating γ; because both Y_i and D_i are on a logarithmic scale, γ is approximately the effect of a 1% increase in D_i in terms of percent change in Y_i. To estimate γ, we estimate the following two regression equations using TSLS:

$$Y = \mathbf{X}\mu + \gamma D + \varepsilon_i$$

$$D = \mathbf{X}\alpha + \mathbf{Z}\beta + \eta_i$$

where $\mathbf{X} = [\mathbf{S}\ \mathbf{C}]$ is a sparse $n \times 80{,}065$ matrix of (i) binary indicators for 64 strata formed by the quartiles of preexperiment feedback received, number of peers active on the web interface to Facebook, and preexperiment posting and (ii) binary indicators for 80,001 network clusters formed by graph partitioning, and $\mathbf{Z}$ is an $n \times k$ matrix of instruments, which are each binary indicators derived from the peer encouragement factors.

We expect the effects of feedback to be somewhat heterogeneous. "Marginal feedback," feedback that occurs (or does not occur) because of small changes, may be different from other feedback. Additionally, there may be heterogeneous effects of marginal feedback. For these reasons, we could adopt a nonparametric assumption on $D_i(\cdot)$ rather than a parametric assumption on $Y_i(\cdot)$: Each encouragement does not reduce feedback received for any egos.

Assumption 4. (Monotonicity). *Define* $h(\cdot): \mathbb{Z} \to \{1, \ldots, k\}$ *to order the* k *values in* $\mathbb{Z}$ *such that* $j < l$ *implies* $\mathrm{E}[D_i | h(Z_i) = j] < \mathrm{E}[D_i | h(Z_i) = l]$. *With probability 1,* $D_i(z_i) - D_i(z_i') \geq 0$ *for all* $i \in \mathbb{P}_{egos}$, *where* $h(z_i) > h(z_i')$.

Then TSLS using binary indicators formed from the levels of $h(Z_i)$ estimates a weighted average of estimators using a single binary indicator (ref. 33, theorem 2), each of which estimates an average causal response (ACR), which is a weighted average of effects of changes in increments of D_i. Because $D_i = g(D_i^*)$ is transformed from its original, skewed count distribution, this means that the weights for a change to $g(d_i^*)$ from $g(d_i^* - 1)$ in this average are the normalized product of a difference in cumulative distribution functions for D_i at $g(d_i^* - 1)$ for that instrument and $g(d_i^*) - g(d_i^* - 1)$; see *SI Appendix, Transformed and Untransformed Count Variables*. Our main results use a first stage without interactions between the two factors, so this simple theorem does not directly apply to that model. However, Lochner and Moretti (ref. 52, proposition 2) show that TSLS nonetheless estimates a weighted average of the single instrument estimands. This weighting function is shown in *SI Appendix*, Fig. S7. We test the choice of this first-stage model and show in Fig. 4 that the results are not affected by instead using data-driven shrinkage and selection with the lasso.

Results

We first examine the effects of the peer encouragements on feedback received (i.e., first-stage effects). Both encouragement factors cause peers to comment on and like posts by egos, such that these factors increase (geometric) mean feedback received by 0.2–1.3% (Fig. 3*A*), $F(3, 4.9e7) = 519$, $p <$ 1e-12. Adding the two interaction terms for these factors did not significantly improve fit, $F(2, 4.9e7) = 0.23$, $p = 0.80$. As expected, the encourage initiation factor shifts the lower end of the distribution of feedback received more, compared with the conversation salience factor (Fig. 3*B*).

This randomly induced variation in feedback received allows us to estimate effects of receiving feedback on multiple ego behaviors. We focus on results from a first-stage specification as in Fig. 3*A*, with all three main effects (black points in Fig. 4).

Receiving additional feedback is expected to have the largest effects on "reply" behaviors by the ego, such as commenting on their own posts and liking comments on their posts. We estimate large effects of receiving feedback on both of these ego behaviors, such that a 10% increase in feedback received causes a 9.6% increase in comments (self) and a 10.5% increase in likes (self). Although unsurprising, these estimates can help put the magnitude of effects on other ego behaviors in perspective.

Effects on other ego behaviors are more important for understanding the spread of feedback and sharing behaviors. Receiving additional feedback also causes egos to give others more feedback, in terms of both likes and comments separately: Receiving 10% more feedback causes egos to give others 1.1% more likes and 1.1% more comments. Thus, causing one individual to receive more feedback will cause them to give more feedback to their peers, potentially creating desirable feedback loops. As expected, these effects are substantially smaller than

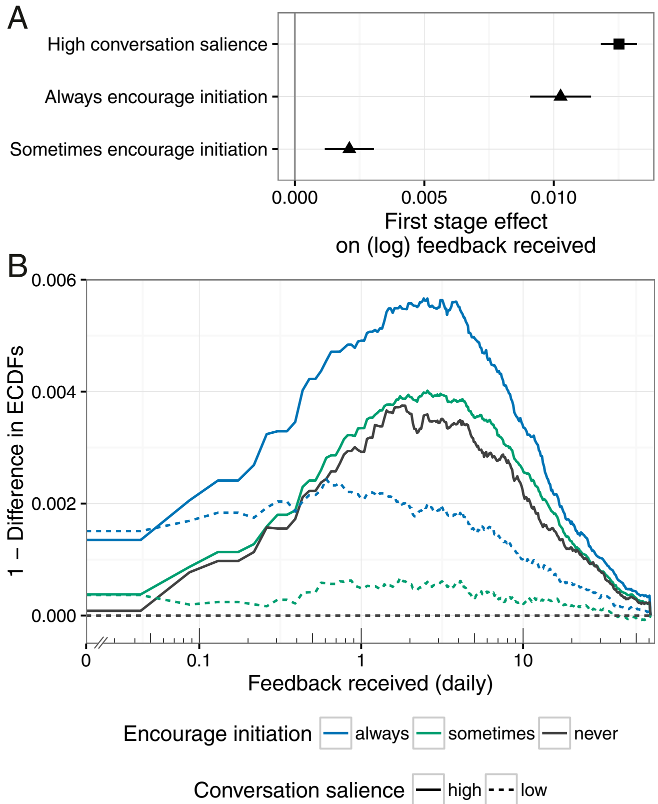


Fig. 3. Effects of the encouragements on feedback received (first stage). (*A*) First-stage average effects. Points are coefficient estimates for the effects of the conversation salience (circle) and encourage initiation (triangles) factors on (log) feedback received, where the base condition is low conversation salience and never encourage initiation. Error bars are 95% network adjacency- and cluster-robust confidence intervals. (*B*) Effects on the distribution of feedback received, computed as a difference in the empirical cumulative distribution functions (ECDFs) of feedback received. Again, with the lowest-feedback condition (never/low) as the baseline, each line represents the difference in probability that daily feedback received is at least the value on the *x* axis. The encourage initiation factor, which has its immediate effects only when a post has no feedback, has larger effects at the low end of the feedback distribution, whereas the conversation salience factor produces shifts at the high end. These differences use poststratification on quartiles of prior feedback received.

on the reply behaviors, but are less than an order of magnitude smaller. Furthermore, when egos receive more feedback, they also share more new posts during the experiment: A 10% increase in feedback causes a 0.7% increase in creating new posts.

We also computed estimates with other first-stage specifications: only the conversation salience factor, only the two encourage initiation factors, and a high-dimensional specification. Specifically, to potentially use heterogeneity in the true first-stage model, we fit a lasso (i.e., L1 penalized) first-stage model (53, 54) with both factors, interactions, and interactions with the stratum-defining variables, with the selected model having 23 nonzero coefficients. The results (Fig. 4) for feedback to peers and posting are statistically indistinguishable for all four models, whereas the two single-factor models differ for effects on the reply behaviors.

Comparison with Observational Estimates. In the absence of this peer encouragement design, scientists and decision makers could instead rely on observational data to study the effects of receiving feedback (10, 55). We thus evaluate how observational estimates compare with our experimental results. We regress each of the ego outcomes on observed feedback received, adjusting for strata and network clusters, as in the IV analysis, but ignoring assignment to peer encouragement conditions. (This analysis includes some variation in feedback received caused by the experiment, but this is a very small fraction of the variance, and it does not materially affect the results.) For all outcomes, these observational estimates of the effect of receiving feedback are substantially different from IV estimates from the peer encouragement design (Fig. 5). For the main outcomes of interest (posting and feedback to others), the observational coefficient estimates are 317–498% larger. On the other hand, they appear to underestimate reply behaviors by 36% and 68% for comments and likes, respectively. That is, in contrast to claims that observational estimates can upper bound true peer effects (6), the sign of the implied large-sample bias of the observational estimators varies across outcomes. These differences could be attributed to confounding, simultaneity, or the fact that IV and observational analyses often estimate different causal quantities (52).

Robustness to Dependence and Nuisance Interference. The preceding inferential results use a network adjacency- and cluster-robust estimator of the variance–covariance matrix (56, 57) to compute SEs; see *SI Appendix*, *Randomization Inference with Sensitivity Analysis*. To further examine the robustness of the results to nuisance interference, we used Fisherian randomization inference for the effect of feedback received on posting (which was the least statistically significant with $p = 0.0013$), while allowing for inference according to Assumption 3 with $c = 1$ under a model whereby an ego's outcome depends the assignments of their peers. This estimate remained statistically significant in the

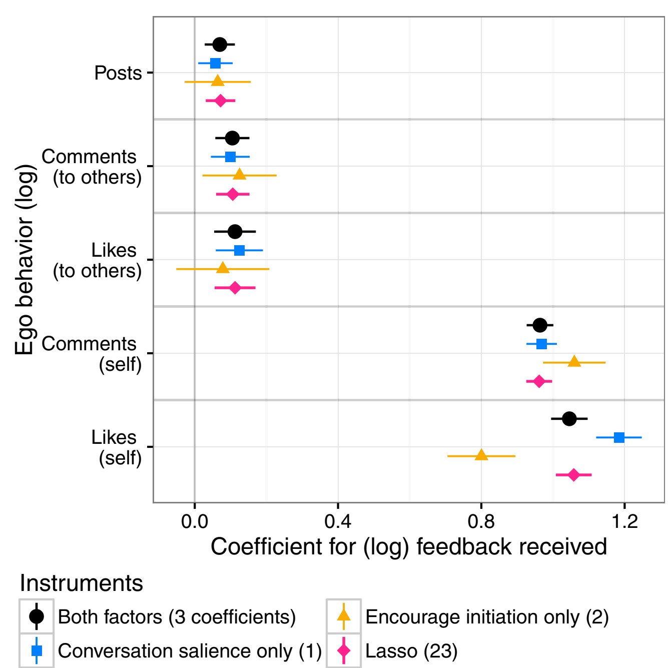


Fig. 4. Effects of receiving feedback on five ego behaviors, as estimated using IV analysis of the peer encouragement design with both peer encouragement factors (black), with only one of the factors (blue and yellow), and using a lasso (L1 penalized) first stage. Points are coefficient estimates from a log–log model. Comments (self) and likes (self) are the ego's comments on their own posts and likes of comments on their posts; these outcomes were expected to be most directly affected by receiving more feedback. The other ego behaviors involve giving more feedback to others and making new posts. Number of instruments is shown in parentheses; for the lasso, this counts only nonzero coefficients. Error bars are 95% network adjacency- and cluster-robust confidence intervals. *SI Appendix*, Table S2 displays these results in tabular form.

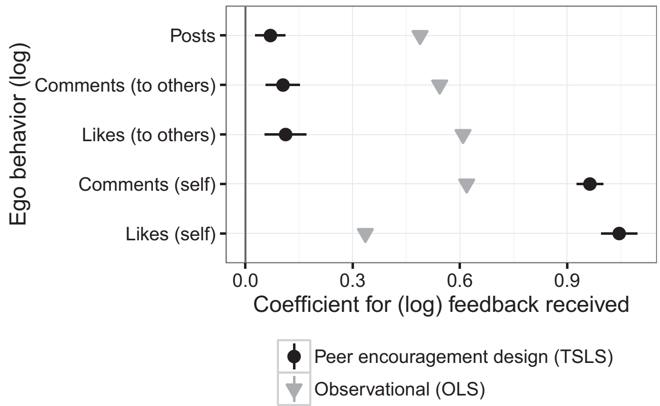


Fig. 5. Comparison of estimates using IV analysis of the peer encouragement design (black) and observational analysis using ordinary least squares (OLS) regression (gray). Points are coefficient estimates from a log–log model. Compared with the IV estimates, the observational analysis of ego behaviors on feedback received either substantially overestimates or underestimates peer effects. Error bars are 95% network adjacency- and cluster-robust confidence intervals and, for the observational (OLS) estimates, are smaller than the points.

presence of additive interference (maximum $p = 0.012$) or interactive interference (maximum $p = 0.017$) from peers; most of this difference in inference arises from the use of randomization inference with the rank sum test statistic, rather than allowing for interference per se (without interference, $p = 0.009$).

Discussion

Peer encouragement designs can be an effective strategy for estimating peer effects in networks: By randomly encouraging peers to specific behaviors, researchers can learn about the effects of those behaviors on egos. In this paper, we reviewed this class of experimental designs and demonstrated the potential to use a minimal encouragement (here, a small change to the user interface for giving feedback) to an ego-specific behavior. We found that receiving additional feedback causes individuals to give feedback to others and to share new posts. Compared with direct reply behaviors, these effects are smaller but still very substantial. This provides new evidence for the influence of peer effects in the use of communication technologies. It also informs our understanding of the value of social feedback to its recipients, as reflected in recipients' decisions to continue using a medium. In particular, receiving more feedback causes individuals to more frequently repeat the same behavior (posting content) that made them able to receive feedback in the first place. These results are informative about the role of directed behaviors in the adoption of technologies that enable both undirected (broadcast) and directed communications.

One limitation of this experiment is that it does not elucidate the mechanisms by which receiving feedback affects egos or distinguish different types of feedback. The observed effects are expected to occur for many reasons. For example, effects on giving feedback to others could be due to a psychological response (e.g., generalized reciprocity), or occur simply because receiving feedback causes users to return to Facebook more often, and therefore creates more opportunities to comment on peers' posts. Distinguishing these and other mechanisms would be difficult, but additional studies could test alternative explanations. For simplicity, we have focused on an experiment that identifies undifferentiated effects of feedback. Additional peer encouragement designs could also distinguish among different types of feedback.

A peer encouragement design identifies an encouragement-specific quantity: the effect of receiving additional feedback for egos' whose peers are induced by the encouragement to provide more feedback. This quantity is an ACR, the generalization of a local average treatment effect (LATE) for a multivalued treatment, or a weighted combination of ACRs. In this study, these are weighted average effects of feedback that would occur (or not) depending on small changes to the user interface. The conventional wisdom (cf. refs. 58–60) is that a LATE or ACR is less relevant than quantities that average over other, larger sets of potential outcomes, such as an average treatment effect or ATT. We argue that an ACR, in fact, averages over differences in peer behavior that are realistic under many relevant alternative policies. Researchers, marketers, or policy makers may be particularly interested in the average effects of incremental peer behaviors—behaviors that will occur or not depending on realistic changes to the environment, policy, or marketing campaign. In some standard economic models, the LATE is a piecewise constant approximation to this marginal treatment effect (61). Thus, if design, policy, or marketing decisions are expected to produce shifts in peer behaviors similar to the encouragement design, then the LATE or ACR may be of greater substantive relevance (cf. ref. 62). Of course, even different encouragements might define quite different ACRs. This experiment included two different peer encouragements that are expected to cause feedback at different times in the lifecycle of an ego's post: The encourage initiation factor should primarily cause the first feedback to occur at all or earlier, and the conversation salience factor should generate additional feedback on posts with existing feedback. We find that, despite this difference, these two factors identify quite similar ACRs, especially for the primary outcomes (*SI Appendix*, Fig. S8). This could increase our confidence that the present results may be informative about other attempts to cause people to receive more feedback on their posts. In this particular case, it is unclear what other averages would be preferable, because the natural generalization of the ATT to a multivalued variable like feedback received would average over contrasts comparing outcomes when egos receive their status quo levels of feedback and if they were to receive no feedback [i.e., $Y_i(D_i) - Y_i(0)$]. This thus includes contrasts in which very active, high-degree egos who currently receive large amounts of feedback receive none, which is perhaps unlikely to occur under policies being considered.

The current work highlights the advantages of large data sets and novel experimental designs for causal inference about how people affect each other. Our peer encouragement design provides credible causal estimates for the effects of receiving social feedback on Facebook; this is, to our knowledge, the first experimental evidence for these effects. The plausibility of a key assumption in our model, the exclusion restriction, partly depends on encouragements being minimal. However, encouragements that produce minimal variation can result in imprecise IV estimates; even studies with hundreds of thousands of observations will often suffer from the instruments being too weak (63). Peer encouragement designs with such minimal encouragements thus require a very large sample size and careful design (e.g., the ego-specific design used here) to estimate peer effects precisely. When feasible, however, peer encouragement designs can provide valuable insights into real-world social dynamics that can inform social science and policy decisions.

Materials and Methods

The peer encouragement design ran for 3 wk between September and October 2012. The egos in the data analyzed are 48.9 million Facebook users globally who had created at least one status update in the 4 d before the start of the experiment or during the experiment, who had at least one friend frequently using Facebook via the web interface, and who reported their age as at least 18. Note that 52% of these egos are randomly assigned to the peer encouragement condition reflecting the status quo at the time. Approximately 905 million Facebook users were peers of the egos and used the web interface. Further details about the sample and covariate balance are reported in *SI Appendix*, Table S1. The primary results we report are adjusted with a set of sparse binary covariates (i.e., dummies) for quartiles of three pretreatment variables (forming $4^3 = 64$ strata) and 80,001 clusters formed by graph partitioning (see *SI Appendix*).

This study uses data from an experiment conducted for routine product improvement purposes and that posed no more than minimal risk. D.E. and E.B. designed and conducted the experiment as part of product development while employees of Facebook in 2012. Research using this data is consistent with the Data Policy that people accept when they choose to use the Facebook service. Accordingly, we did not separately notify users of this specific product test, nor did we obtain written informed consent. R.F.K. later contributed to this research using this existing data while an employee of Facebook in 2014 and 2015. Because he intended to use his university affiliation in reports on this study, R.F.K. asked the Stanford University institutional review board (IRB) to review a protocol for use of this previously collected anonymized data; the Stanford IRB approved this protocol. Similarly, when D.E. became a member of the Massachusetts Institute of Technology (MIT) faculty in 2015, the MIT IRB determined that a protocol for use of this previously collected anonymized data was exempt, and approved the protocol.

ACKNOWLEDGMENTS. This work benefited from comments by A. Fradkin, G. W. Imbens, S. Messing, A. Peysakhovich, J. Sekhon, S. J. Taylor, members of the Facebook Core Data Science team, and anonymous reviewers. We thank L. Backstrom, K. Deeter, and D. Vickrey for assistance conducting this experiment.

1. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. *N Engl J Med* 357(4):370–379.
2. Christakis NA, Fowler JH (2008) The collective dynamics of smoking in a large social network. *N Engl J Med* 358(21):2249–2258.
3. Cohen-Cole E, Fletcher JM (2008) Detecting implausible social network effects in acne, height, and headaches: Longitudinal analysis *BMJ* 337:a2533.
4. Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociol Methods Res* 40(2):211–239.
5. Thomas AC (2013) The social contagion hypothesis: Comment on 'Social contagion theory: Examining dynamic social networks and human behavior.' *Stat Med* 32(4): 581–590, and discussion (2013) 32(4):597–599.
6. Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc Natl Acad Sci USA* 106(51):21544–21549.
7. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annu Rev Sociol* 27:415–444.
8. Manski CF (2000) Economic analysis of social interactions. *J Econ Perspect* 14(3): 115–136.
9. Ogburn EL, van der Weele TJ (2014) Causal diagrams for interference. *Stat Sci* 29(4): 559–578.
10. Burke M, Marlow C, Lento T (2009) Feed me: Motivating newcomer contribution in social network sites. *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (Assoc Comput Machinery, New York), pp 945–954.
11. Christakis NA, Fowler JH (2013) Social contagion theory: Examining dynamic social networks and human behavior. *Stat Med* 32(4):556–577.
12. Ugander J, Backstrom L, Marlow C, Kleinberg J (2012) Structural diversity in social contagion. *Proc Natl Acad Sci USA* 109(16):5962–5966.
13. Jacobs AZ, Way SF, Ugander J, Clauset A (2015) Assembling thefacebook: Using heterogeneity to understand online social network assembly. arXiv:1503.06772.
14. Shriver SK, Nair HS, Hofstetter R (2013) Social ties and user-generated content: Evidence from an online social network. *Manage Sci* 59(6):1425–1443.
15. Tucker C (2008) Identifying formal and informal influence in technology adoption with network externalities. *Manage Sci* 54(12):2024–2038.
16. Moffitt RA (2001) Policy interventions, low-level equilibria, and social interactions. *Social Dynamics*, eds Durlauf SN, Young HP (MIT Press, Cambridge, MA), pp 45–82.
17. Walker D, Muchnik L (2014) Design of randomized experiments in networks. *Proc IEEE* 102(12):1940–1951.
18. Bakshy E, Eckles D, Yan R, Rosenn I (2012) Social influence in social advertising: Evidence from field experiments. *Proceedings of the 13th ACM Conference on Electronic Commerce* (Assoc Comput Machinery, New York), pp 146–161.
19. Asch SE (1956) Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychol Monogr* 70(9):1–70.
20. Deutsch M, Gerard HB (1955) A study of normative and informational social influences upon individual judgment. *J Abnorm Soc Psychol* 51(3):629–636.
21. Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.
22. Sacerdote B (2001) Peer effects with random assignment: Results for Dartmouth roommates. *Q J Econ* 116(2):681–704.
23. Carrell SE, Fullerton RL, West JE (2009) Does your cohort matter? Measuring peer effects in college achievement. *J Labor Econ* 27(3):439–464.
24. Kling JR, Liebman JB, Katz LF (2007) Experimental analysis of neighborhood effects. *Econometrica* 75(1):83–119.
25. Centola D (2010) The spread of behavior in an online social network experiment. *Science* 329(5996):1194–1197.
26. Aral S, Walker D (2011) Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Manage Sci* 57(9):1623–1639.
27. Aral S, Walker D (2012) Identifying influential and susceptible members of social networks. *Science* 337(6092):337–341.
28. Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. *Proceedings of the 21st International Conference on World Wide Web* (Assoc Comput Machinery, New York), pp 519–528.
29. Pearl J (2009) Causal inference in statistics: An overview. *Stat Surv* 3:96–146.
30. Holland PW (1988) Causal inference and path analysis. *Sociol Methodol* 18:449–484.
31. Bogatz GA, Ball S (1971) *The Second Year of Sesame Street: A Continuing Evaluation: A Report to the Children's Television Workshop* (Educ Test Serv, Princeton), Vol 1.
32. Powers DE, Swinton SS (1984) Effects of self-study for coachable test item types. *J Educ Psychol* 76(2):266–278.
33. Angrist J, Imbens G (1995) Two-stage lease squares estimation of average causal effects in models with variable treatment intensity. *J Am Stat Assoc* 90(430):431–442.
34. Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–475.
35. Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. *J Am Stat Assoc* 91(434):444–455.
36. Imbens GW (2014) Instrumental variables: An econometrician's perspective. *Stat Sci* 29(3):323–358.
37. Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge Univ Press, Cambridge, UK).
38. Manski CF (1993) Identification of endogenous social effects: The reflection problem. *Rev Econ Stud* 60(3):531–542.
39. Athey S, Eckles D, Imbens GW (2015) Exact p-values for network interference. arXiv: 1506:02084.
40. Angelucci M, De Giorgi G (2009) Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *Am Econ Rev* 99(1):486–508.
41. Duflo E, Saez E (2003) The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *Q J Econ* 118(3):815–842.
42. Forastiere L, Mealli F, VanderWeele TJ (December 22, 2015) Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification. *J Am Stat Assoc*, 10.1080/01621459.2015.1125788.
43. Miguel E, Kremer M (2004) Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72(1):159–217.
44. Nickerson DW (2008) Is voting contagious? Evidence from two field experiments. *Am Polit Sci Rev* 102(1):49–57.
45. Bond RM, et al. (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298.
46. Cai J, De Janvry A, Sadoulet E (2012) Social networks and the decision to insure. *Am Econ J Appl Econ* 7(2):81–108.
47. Coppock A, Guess A, Ternovski J (2016) When treatments are tweets: A network mobilization experiment over Twitter. *Polit Behav* 38(1):105–128.
48. Frey BS, Jegen R (2001) Motivation crowding theory. *J Econ Surv* 15(5):589–611.
49. Cox DR (1958) *Planning of Experiments* (Wiley, New York).
50. Manski CF (2013) Identification of treatment response with social interactions. *Econom J* 16(1):S1–S23.
51. Aronow P, Samii C (2014) Estimating average causal effects under general interference. Working paper (Washington Univ St. Louis). Available at polmeth.wustl.edu/node/233.
52. Lochner L, Moretti E (2015) Estimating and testing models with many treatment levels and limited instruments. *Rev Econ Stat* 97(2):387–397.
53. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* 58(1):267–288.
54. Belloni A, Chen D, Chernozhukov V, Hansen C (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80(6): 2369–2429.
55. Burke M, Kraut RE (2014) Growing closer on Facebook: Changes in tie strength through social network site use. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Assoc Comput Machinery, New York), pp 4187–4196.
56. Conley TG (1999) GMM estimation with cross sectional dependence. *J Econom* 92(1): 1–45.
57. Cameron AC, Gelbach JB, Miller DL (2011) Robust inference with multiway clustering. *J Bus Econ Stat* 29(2):238–249.
58. Aronow PM, Carnegie A (2013) Beyond LATE: Estimation of the average treatment effect with an instrumental variable. *Polit Anal* 21(4):492–506.
59. Deaton A (2010) Instruments, randomization, and learning about development. *J Econ Lit* 48(2):424–455.
60. Imbens GW (2010) Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *J Econ Lit* 48(2):399–423.
61. Heckman JJ, Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proc Natl Acad Sci USA* 96(8): 4730–4734.
62. Dunning T (2012) *Natural Experiments in the Social Sciences: A Design-Based Approach* (Cambridge Univ Press, Cambridge, UK).
63. Bound J, Jaeger DA, Baker RM (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc* 90(430):443–450.

Supporting information for "Estimating peer effects in networks with peer encouragement designs"

Dean Eckles, René F. Kizilcec, Eytan Bakshy

Contents

1 Description of experimental conditions

The experiment consisted of two design factors: encourage initiation and conversation salience (see details in the main text). Both factors only affected the user interface when users were viewing News Feed in the Web interface for Facebook (i.e., not interfaces for mobile phones). The encourage initiation factor has three levels that determine how often the existing feedback and textbox for making a comment are shown in News Feed by default, rather than requiring a click to see. The *always* and *never* levels correspond to either always or never automatically showing existing feedback when displaying the shared content. The *sometimes* condition shows existing feedback only when the shared content appears in the first position in News Feed.

2 Summaries by condition

We provide summary statistics for the peer encouragement conditions in Table S1, including the number of assigned egos, demographics, and a set of pre-experimental covariates. Analysis of variance for these pre-experiment covariates by condition were all non-significant, consistent with successful randomization.

A simplified version of the IV analysis of the effect of feedback received on ego behavior can be presented visually. Fig. S1 shows condition-level summaries of feedback received and content production.

Table S1: Peer encouragement conditions with Ns and summaries of pre-experiment covariates. Comparisons of means and quartiles (in brackets) of variables across conditions. Analysis of variance for these pre-experiment covariates by condition were all non-significant, consistent with successful randomization. Prior posts and prior feedback received are both skewed enough to have means larger than upper quartiles. Variation between the size of conditions is intentional and reflective of the default presentation at the time of the experiment.

Conversation salience	Encourage initiation	N	Female	Age	Active peers	Prior posts	Prior feedback received
high	always	2328489	0.50	30.42	189.25	0.85	5.40
					[46, 109, 233]	[0.06, 0.22, 0.78]	[0.22, 1.44, 5.11]
high	sometimes	25618796	0.50	30.42	189.35	0.85	5.41
					[46, 109, 233]	[0.06, 0.22, 0.78]	[0.22, 1.44, 5.11]
high	never	2328194	0.50	30.44	189.16	0.85	5.40
					[46, 110, 233]	[0.06, 0.22, 0.78]	[0.22, 1.44, 5.11]
low	always	4658871	0.50	30.42	189.37	0.85	5.41
					[46, 110, 233]	[0.06, 0.22, 0.78]	[0.22, 1.44, 5.11]
low	sometimes	9313677	0.50	30.43	189.41	0.85	5.41
					[46, 109, 233]	[0.06, 0.22, 0.78]	[0.22, 1.44, 5.11]
low	never	4656022	0.50	30.43	189.48	0.85	5.39
					[46, 110, 233]	[0.06, 0.22, 0.78]	[0.22, 1.44, 5.11]

3 Covariates

As described in Materials and Methods, we used relevant covariates to increase the precision of the estimates reported in the main text. In particular, we used dummies for strata defined by three discrete covariates:

1. Prior feedback received: Quartile buckets of the number of likes and comments on the ego's posts in the 18 days prior to the experiment.
2. Active peers: Quartile buckets of the number of peers (Facebook friends) who had used the Web interface to Facebook at least 7 of the 28 days prior to the experiment.
3. Prior sharing: Quartile buckets of the number of posts made by the ego in the 18 days prior to the experiment.

Note that since the peer encouragement only affected peers viewing a post in Facebook News Feed via the Web interface, the second variable only counts peers who are active via the Web interface.

Additionally, as discussed in Section 5.1 below, the main analysis allows for dependence within each of the 80,001 clusters defined by graph partitioning. Therefore, we also included dummies for each of these clusters to eliminate any mean dependence within clusters and potentially increase precision.

4 Model

This section provides additional information about the models and estimands that can motivate the design and analysis of the peer encouragement design. Let $\mathbb{P}_{\text{egos}}$ be the set of n egos and $\mathbb{P}_{\text{peers}}$ be the set of m_{p} peers. These sets are not disjoint: nearly all units in $\mathbb{P}_{\text{egos}}$ are in $\mathbb{P}_{\text{peers}}$. Let $\mathbb{P}_{\text{e}\cup\text{p}} = \mathbb{P}_{\text{egos}} \cup \mathbb{P}_{\text{peers}}$ be the union of $m_{\text{e}\cup\text{p}}$ units. We use bold letters for matrices and capital letters for random variables.

For the question of how receiving additional feedback affects ego behaviors, key quantities of interest are contrasts between ego behaviors with different amounts of received feedback. For all egos $i \in \mathbb{E}$, we parameterize received feedback as as a function of the sum of the feedback (likes L and comments C) on an ego's posts:

$$D_i \equiv g\Big(\sum_{j \in \mathbb{P}_{\text{peers}}} L_{ji} + C_{ji} \Big).$$

For substantive and data analytic reasons discussed in Section 9, we take $g(\cdot)$ to be a logarithmic function in our preferred analysis.

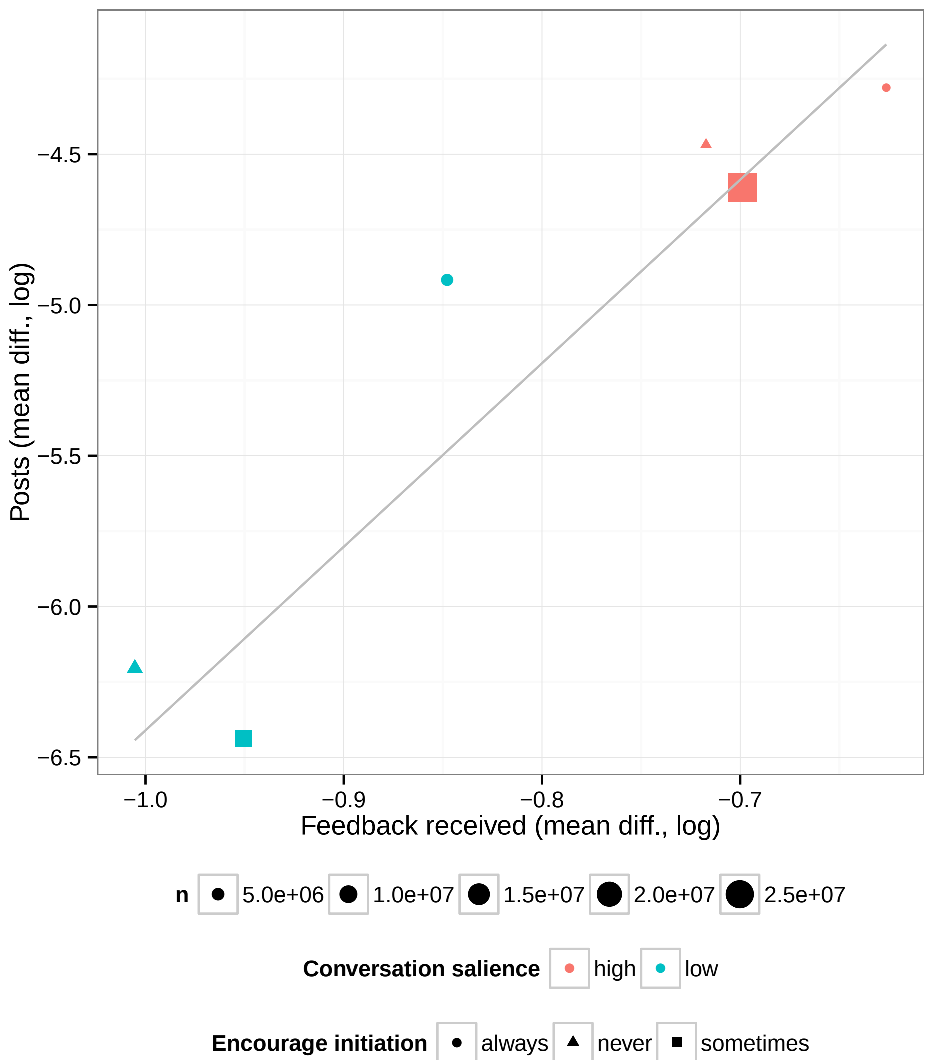


Figure S1: Summary of feedback received and posts by peer encouragement condition. Each point represents a single combination of the the two factors. We compute the mean (log) feedback received and (log) posts and subtract the same variable from the pre-experiment period. Each point is then the mean of this pre–post difference for that condition. A linear best fit line (weighted by n) is shown.

Let $Y_i(d_i, d_{-i}, z_i, \mathbf{z}_{-i})$ denote the potential outcome (e.g., posts shared during the experiment) for ego i if they received d_i feedback and were assigned to z_i and all other units receive feedback and have assignments according the rest of the $m_{\mathrm{e}\cup\mathrm{p}}$-vectors d and z.[1]

Without making further assumptions about other peer effects, these quantities of interest are contrasts between potential outcomes that may depend on vectors of all units' assignments and behaviors. For example,

$$Y_i(d_i, d_{-i}, z_i, \mathbf{z}_{-i}) - Y_i(d_i', d_{-i}, z_i, \mathbf{z}_{-i})$$

contrasts ego i's outcomes under two different levels of how much feedback i receives (d_i and d_i') while holding constant the other determinates of i's outcome, including their assignment, the assignment of all other units, and feedback received by other units. These two quantities are not simultaneously observable. Multiple sets of assumptions can be used to justify estimating summaries (e.g., averages) of these contrasts from data. We describe three sets of these assumptions.

4.1 Homogeneous effects

Standard treatments of instrumental variable methods and widespread practice in econometrics has, until recently, worked primarily with models in which the estimand is a coefficient in a linear model. In the present case, this model would have the form

$$Y_i(d_i) - Y_i(0) = \gamma d_i, \tag{1}$$

where we then observe the function $Y_i(\cdot)$ for the value of the observed level of the endogenous directed peer behavior $Y_i^{obs} = Y_i(D_i^{obs})$ (in this case, feedback received). The "dose" itself (feedback received) is function of the randomly assigned instrument, for which we observe $D_i^{obs} = D_i(Z_i^{obs})$.[2]

In this setting, for the finite population $\mathbb{P}_{\mathrm{egos}}$, exact inference for γ is possible (17) by inverting a hypothesis test for $\gamma = \gamma_0$. We conduct this analysis in Section 5.2.

One important shortcoming of this model is that it implicitly makes assumptions about any interference. By writing the outcomes as a function only of a level of directed peer behaviors, this would typically require that outcomes are constant in many of the arguments previously posited (i.e. other behaviors and random assignment); that is, we have for all i that

$$Y_i(d_i) = Y_i(d_i, d_{-i}, z_i, \mathbf{z}_{-i}) = Y_i(d_i, d_{-i}', z_i', \mathbf{z}_{-i}')$$

[1]This specification of potential outcomes rules out certain kinds of simultaneity or feedback loops, such as the ego's outcome (e.g., posting) in an initial period causing them to receive more feedback subsequently, and then this causing the outcome in a later period.

[2]Writing $D_i(\cdot)$ as a function only of Z_i assumes that D_i has no other parents among our variables; this, along with the exclusion restriction, rules out some cases of simultaneity or aggregation of multiple time periods. For example, say an encouragement causes feedback to occur, which in turn causes an outcome, this in turn causes more feedback to occur, etc. Ogburn et al. (20) addresses the negative consequences for identification of endogeneous timing of the mediating behaviors.

for all $d \in \mathbb{D}^n$, $\mathbf{z}, \mathbf{z}' \in \mathbb{Z}^n$. This makes a hypothesis that $\gamma = \gamma_0$ a *sharp* hypothesis, in that all of the potential outcomes can be inferred from a single observation.

4.2 Heterogeneous effects

Without the assumption of Equation 1, effects may be heterogenous across units and across increments to the endogenous treatment. That is, the model might be linear but heterogeneous,

$$Y_i(d) - Y_i(0) = \gamma_i d_i, \tag{2}$$

or both heterogeneous and potentially non-linear.

In the absence of interference, the identification results from Angrist and Imbens (1) apply. That is, for each instrument there is a parameter γ that is a weighted average per-unit treatment effect, where the weights are determined by the shift in the distribution of D_i caused by the instrument. Angrist and Imbens (1) call this the *average causal response* (ACR). The weighting functions for all pairs of conditions are given by Fig. 3B in the main text. The weighting function for the main results are shown in Fig. S7.

4.3 With interference

The models above make a unit's potential outcomes invariant in changes to other units' assignments, but we expect units to be interacting.

In some other work, an explicit goal is to contrast very different treatment vectors. Hudgens and Halloran (16) consider a population average overall causal effect that compares two arbitrary distributions of treatment assignments ϕ and ϕ'. When ϕ is deterministic assignment to treatment and ϕ' is deterministic treatment assignment to control; Eckles et al. (15) call this the global ATE. This is not our goal here; rather, the present work aims to estimate effects on an individual under more-or-less the current regime, as this corresponds to questions about how egos are affected by marginal feedback and what effects small, targeted changes might have.

Following Hudgens and Halloran (16), one can define individual-level average potential outcomes in terms of a unit's assignment for some distribution ϕ of global assignment vectors. For the total effects of assignment, we have

$$\bar{Y}_i(z_i; \phi) \equiv E_\phi[Y_i(D_i(\mathbf{z}), d_{-i}, z_i, \mathbf{z}_{-i})].$$

Then define individual-level average effects by

$$\bar{\tau}_i(z_i, z_i'; \phi) \equiv \bar{Y}_i(z_i; \phi) - \bar{Y}_i(z_i'; \phi).$$

These can be summarized as the finite population average of individual-level effect for, e.g., assignment to treatment versus control:

$$\bar{\tau}(1, 0; \phi) \equiv \sum_{i \in \mathbb{P}_{\text{egos}}} \bar{\tau}_i(1, 0; \phi).$$

If units' probability of assignment under ϕ is constant in the population (as it is in our design), then the sample difference in means is an unbiased estimator. Thus, standard intent-to-treat estimators remain unbiased, but simply for a quantity that may depend on the distribution the assignment vector is drawn from. Similarly, by considering the analogous averages of the D_i, also for the first-stage effects. This is sufficient for both the numerator and denominator of the Wald estimator for instrumental variables to remain unbiased. Of course, as in the absence of interference, the estimator itself remains biased in finite samples (9).

However, interference can still affect the sampling distribution of these estimators and so affect the operating characteristics of standard methods for statistical inference; see Section 5 below. More specifically, Section 5.2 conducts a sensitivity analysis in the presence of interference; we find that this does not substantially affect our inferences.

4.3.1 Additive interference

More technically, Equation 1 does not exclude all forms of interference; in particular, if interference is additive, then this constant effects model could hold. More generally (i.e., allowing for heterogeneous effects), if there is additive interference then differences between outcomes for different peer behaviors are invariant in the other inputs; that is, that for all i there is some

$$\tau_i(d_i, d_i') = Y_i(d_i, d_{-i}, z_i, \mathbf{z}_{-i}) - Y_i(d_i', d_{-i}', z_i, \mathbf{z}_{-i})$$

for all $d, d' \in \mathbb{D}^n$, $\mathbf{z}, \mathbf{z}' \in \mathbb{Z}^n$. In this case, the individual-level average effects $\bar{\tau}_i(d_i, d_i'; \phi)$ do not depend on the choice of ϕ (as long as ϕ puts positive probability on the relevant values of D). As discussed in Section 5.2, this interference can affect inference, but in our sensitivity analysis, it does so only slightly.

5 Statistical inference

General statistical inference in networks remains a relatively open area of research, such that available methods involve strong substantive assumptions, un-scalable computation, low statistical power, and/or unclear asymptotics. Two distinct but related potential violations of standard independence assumptions apply to the present case: interference (i.e., spillovers, effects of other units' assignments) and network-correlated errors. We employ and combine multiple established methods with the aim of evaluating the robustness of the results to the specific assumptions of each. We expected that inference would not be substantially affected by accounting for potential interference and network-correlated errors since the peer encouragement is ego-specific and not correlated in the network.

5.1 Asymptotic inference with adjacency- and cluster-robust SEs

Following Conley (12), spatial econometrics has made use of estimators for variance–covariance matrices that are consistent in the presence of spatially correlated errors. These are Huber–White "sandwich" estimators of the form

$$\widehat{\mathrm{Var}}[\hat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\hat{u}\hat{u}' \odot \mathbf{B})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{3}$$

where $\hat{u}$ is the vector of residuals, $\odot$ is element-wise multiplication, and $\mathbf{S}$ is a $n \times n$ matrix that selects and/or weights pairs of observations. In versions of this estimator robust to one-way clustering, $\mathbf{B}$ is a block diagonal matrix with $b_{ij} = 1$ if and only if i and j are in the same cluster. In the spatial case, $b_{ij} = K(s_{ij})$ where s_{ij} is the distance between units i and j and $K(\cdot)$ is a kernel.

We use this estimator with network distance, such that $\mathbf{B}^{\text{adj}} = \mathbf{I} + \mathbf{A}$, where $\mathbf{A}$ is the adjacency matrix. That is, for each i, $B_{ii}^{\text{adj}} = 1$ and for each $i \neq j$, $B_{ij}^{\text{adj}} = A_{ij}$.

We note that while this method has been widely applied to networks and non-spatial measures of distance, including by Conley (12), the relevant asymptotics for networks are underdeveloped. The Conley (12) results use a metric space embedding, though other results use other sets of assumptions (18). This method also coincides with the use of multi-way clustering for dyadic data (3). To illustrate the performance of these methods in networks with local interference, we present some simple simulation results in Section 11 below.

The adjacency matrix for this analysis is prohibitively large. We use a sample of 1.7 million of its rows, including all columns, thus maintaining the full dependence structure for the egos included in the sample. In a smaller sample size, this would simply change the degrees of freedom for relevant t- and F-statistics, but this is without consequence at this sample size.

There may be some dependence between egos that are not each others' peers, but have mutual neighbors. The number of paths of length two is prohibitively large (i.e., over 5 billion such paths originating from the subsample of 1.7 million egos) to readily incorporate into the above estimator. However, we additionally computed estimators of the variance–covariance matrix using cluster-robust sandwich estimators. Here clusters are defined according to a conveniently available partitioning of the friendship graph for other purposes (e.g., for graph cluster randomized experiments, as in Ugander et al. (23) and Eckles et al. (15)) into 80,000 clusters using balanced label propagation (22); egos not in any of these clusters were assigned to another cluster. This analysis is partially motivated by recent results in spatial econometrics on the application of cluster-robust variance–covariance estimators to spatially dependent data (8). On its own, this cluster-robust variance–covariance estimator has the disadvantage that only approximately 30% of the edges between egos are within the same cluster, so clearly there is potential for between-cluster dependence.

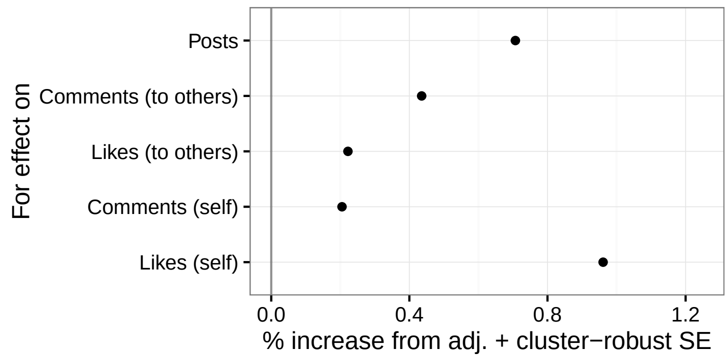


Figure S2: Comparison of adjacency- and spatial-robust SEs and standard heteroskedastic-robust SEs for the main estimates from Fig. 4 in the main text. The increases are less than 1%.

For these reasons, we use a variance–covariance estimator that combines both the adjacency- and cluster-robust estimators. The cluster-robust estimator consists of Equation 3 with a different "selector" matrix $\mathbf{B}^{\mathrm{clu}}$ with $B_{ij}^{\mathrm{clu}} = 1$ if and only if egos i and j are in the same cluster. This can be interpreted as arising from an alternative measurement of distance between i and j. Following, in the spatial literature, Conley and Molinari (13), Kelejian and Prucha (18), or, in work on multi-way clustering, Cameron et al. (11), one could use a weighting matrix $\mathbf{B}^{\mathrm{both}}$ that is the element-wise maximum of the two, such that $B_{ij}^{\mathrm{both}} = \max(B_{ij}^{\mathrm{adj}}, B_{ij}^{\mathrm{clu}})$. This estimator, call it $\widehat{\mathrm{Var}}^{\mathrm{both}}[\hat{\beta}]$, can be computed as a linear combination of estimators,

$$\widehat{\mathrm{Var}}^{\mathrm{m}}[\hat{\beta}] = \widehat{\mathrm{Var}}^{\mathrm{adj}}[\hat{\beta}] + \widehat{\mathrm{Var}}^{\mathrm{clu}}[\hat{\beta}] - \widehat{\mathrm{Var}}^{\mathrm{adj}\times\mathrm{clu}}[\hat{\beta}],$$

where the last term is based on a selector matrix that is the element-wise product of the first two. Since $\widehat{\mathrm{Var}}^{\mathrm{adj}}[\hat{\beta}]$ is computed on a sample of rows and to avoid construction of the rest of these matrices, we replace $\widehat{\mathrm{Var}}^{\mathrm{adj}\times\mathrm{clu}}[\hat{\beta}]$ with an estimator that is smaller in expectation, the standard heteroskedasticity-robust sandwich estimator for independent data, which is equivalent to using a selector matrix with ones on the diagonal.

The primary results in the main text make use of this combined adjacency- and cluster-robust sandwich estimator. Compared with not accounting for these potential sources of the dependence, this results in a quite small increase in SEs for the coefficients of interest. For the main IV estimates, the increase is less than 1% (Fig. S2). This was expected given that the instruments are not correlated in the network and interference was expected to be small.

5.2 Randomization inference with sensitivity analysis

We use Fisherian randomization inference to further examine the robustness of our results. First, we repeat the main analysis using the randomization inference method of Imbens and Rosenbaum (17). Second, we extend this method to conduct a sensitivity analysis to certain types of interference. We conduct this analysis for the effect of feedback received on content production (i.e., sharing posts), which is perhaps the most substantial outcome and the outcome with the largest p-value (and therefore expected to be most sensitive to alternative inferential methods).

Imbens and Rosenbaum (17) apply Fisherian randomization inference to an instrumental variables design, yielding exact tests of a constant effects model. Their simulations illustrate this method's robustness to weak instruments and higher power with long-tailed distributions. For the model in Equation 1, one can use an instrument to test hypotheses of the form $\gamma = \gamma_0$ as follows. Compute the residuals $r_0 = Y^{obs} - \gamma_0 D^{obs}$ and some statistic that is a function of the r_0 and instruments $\mathbf{Z}$. Then compare this observed statistic T^{obs} to the known null distribution of T under repeated sampling of $\mathbf{Z}$ from the distribution of treatment assignments ϕ. We construct a confidence set for γ by inverting this test. If this set is nonempty, then there are some values of γ for which this model is consistent with the data, at least with respect to the alternatives against which the choice of T has power. However, as is common to methods that invert Fisher's exact hypothesis tests of this kind, this confidence set could exclude some values of γ that are consistent with the data under a less restrictive model (i.e., one with interference).

We thus extend this method to allow us to examine the sensitivity of these results to a limited form of interference. Following Assumption 3 in the main text (i.e., direct-effect-bounded interference), we assume that the interference is smaller than the effect of some increment to D_i. In particular, we consider a model in which egos' outcomes depend linearly on the fraction of treated peers,

$$Y(d, \mathbf{Z}) = Y(0, 0) + \gamma d + \tilde{\mathbf{A}}\mathbf{Z}\zeta, \tag{4}$$

where d is an n-vector, $\mathbf{Z}$ is a $n \times 2$ matrix of indicators for whether each unit has conversation salience high and always encourage initiation, and $\tilde{\mathbf{A}}$ is the row-normalized adjacency matrix. It is possible to test joint hypotheses of the form $\gamma = \gamma_0$ and $\zeta = \zeta_0$ against some alternatives. As with the no-interference case, we compute the residuals $r_0 = Y^{obs} - \gamma_0 D^{obs} - \zeta_0 \tilde{\mathbf{A}}\mathbf{Z}^{obs}$ and some statistic that is a function of residuals r_0 and instruments $\mathbf{Z}$. This statistic is then compared with the known null distribution.

This general procedure is the same as in Bowers et al. (10), except that we apply it to inference with instruments and we treat the interference parameter as a nuisance, rather than trying to do inference for it. As we are primarily interested in γ, we can conduct sensitivity analysis by testing sets of hypotheses of the form $\gamma = \gamma_0$ and $\zeta_k \in (\zeta^-, \zeta^+)$ for $k \in \{1, 2\}$ for the two instruments we use.

We expect that any spillovers should be small compared with the direct effects (i.e., Assumption 3). To be conservative we allow the spillovers from each of the two columns of $\mathbf{Z}$ to be as large as γ. Since D_i is on a log scale, this corresponds to the assumption that the interference from each factor is less than the effect of a 172% increase in feedback received. Of course, we do not know γ so we examine sensitivity to spillovers as large as γ in two ways. In both cases, we set $\zeta^- = -\zeta^+$ for symmetry. First, we set $\zeta^+ = \gamma_0$, a hypothetical value of γ, such that as we test different values of γ we also test different values of ζ. This has the consequence that when testing $\gamma = 0$, we also have $\zeta = 0$, meaning that inference is not affected by interference at this point. So we also do a sensitivity analysis with $\zeta^+ = \hat{\gamma}$, such that there are the same levels of interference for all tested values of γ.

As our test statistic, we use the sum of ranks of r_0 within the four groups formed by the binary factors high conversation salience and always encourage initiation. We selected this test statistic because (a) the null distribution can be approximated without actually computing permutations (as this corresponds to the Kruskal–Wallis rank-sum test) and (b) it is expected to be sensitive to changes in γ.

We find that inference for γ is largely unaffected by these levels of linear interference; that is, the resulting confidence set is of a similar size and location as the confidence intervals from our asymptotic inference. Fig. S3 shows p-values as a function of γ. Across all settings of ζ for $\gamma = 0$, the largest p-value is 0.012, so we still reject $\gamma = 0$.

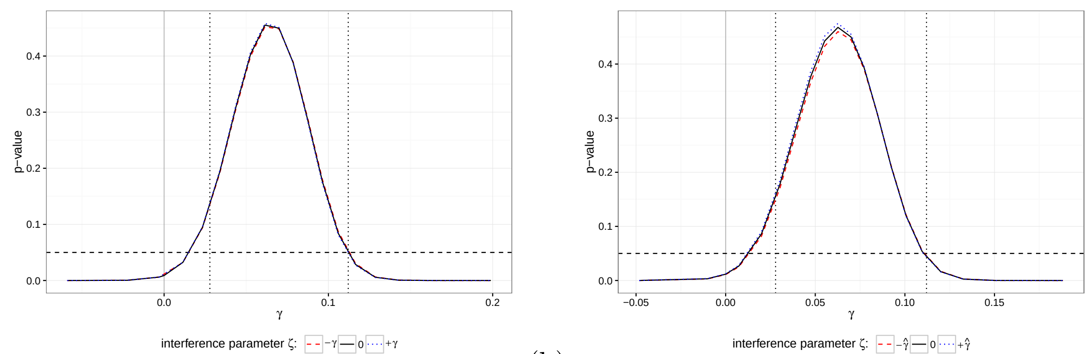
Note that because the confidence set for γ is non-empty when $\zeta = 0$, the data are consistent with there being no interference from the fraction of treated neighbors — at least to the extent deviations from this no interference model are detectable with these test statistics. That is, compared with the conditional randomization methods in Aronow (2) and Athey et al. (4), which condition on elements of the observed assignment vector, the procedure here tests a more specific model for effects of a unit's own treatment (and fails to reject it).

We also conducted the same sensitivity analysis, but with interactions between ego treatment and peer treatment. This allows for the effect of peers' assignment to have *their peers* encouraged to affect egos differently depending on each ego's assignment. This model is

$$Y_i(d_i, \mathbf{Z}) = Y_i(0,0) + \gamma d_i + \big(\tilde{A}_i \mathbf{Z} z_i - \frac{1}{n} \sum_{j \in \mathbb{P}_{\text{egos}}} \tilde{A}_j \mathbf{Z}\big)\zeta, \tag{5}$$

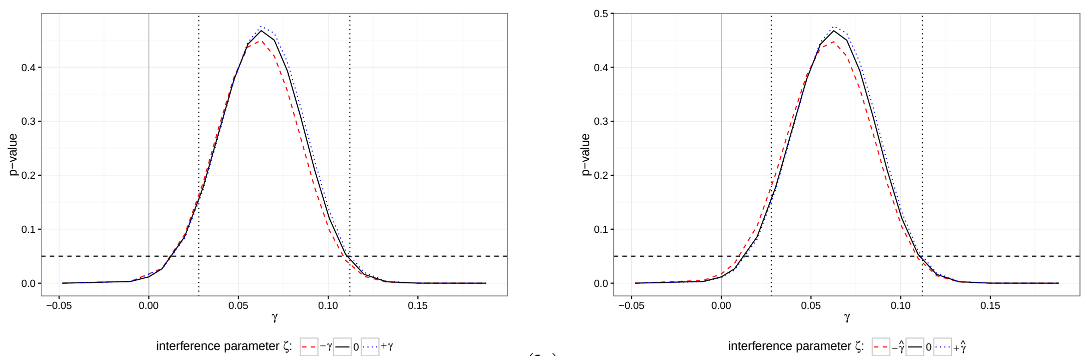
where $\tilde{\mathbf{A}}_i$ is the ith row of the row-normalized adjacency matrix $\tilde{\mathbf{A}}$, and ζ takes on the same values as before. Note that γ has the interpretation of being both the effect of feedback at the mean level of the fraction of ego-peers assigned and also, due to linearity, the average effect under a distribution of ego-peers assigned that is centered at this value. In this sense, in the framework of Hudgens and Halloran (16), inference for γ is inference for the average of average individual-level direct effects.

This interactive interference appears to affect inference somewhat more than the additive interference, but still does not substantially change the results (Fig. S4). Across all settings of ζ, the largest p-value for the hypothesis that $\gamma = 0$ is 0.017.



(a) (b)

Figure S3: Sensitivity analysis for additive spillovers using Fisherian randomization inference for the effect of feedback received on posting using the Kruskal–Wallis rank-sum test statistic. Values of γ with a p-value greater than 0.05 (dashed horizontal line) for all values of ζ are included in the 95% confidence set. Here, the confidence set is simply an interval defined by a start and end point. In (a), ζ takes on values that depend on the posited value of γ shown on the x-axis: $\zeta \in \{-\gamma, 0, \gamma\}$. In (b), ζ is constant for all tested values of γ based on our estimated value for γ: $\zeta \in \{-\hat{\gamma}, 0, \hat{\gamma}\}$. In neither case does this appreciably affect inference for the effect of feedback received on posting, γ. There is nonetheless some difference between the randomization inference confidence set and the 95% confidence interval from asymptotic adjacency- and cluster-robust inference (limits are shown as dotted vertical lines).



(a) (b)

Figure S4: Sensitivity analysis for interactive spillovers using Fisherian randomization inference for the effect of feedback received on posting, γ, using the Kruskal–Wallis rank-sum test statistic. Values of γ with a p-value greater than 0.05 (dashed horizontal line) for all values of ζ are included in the 95% confidence set. Here, the confidence set is simply an interval defined by a start and end point. In (a), ζ takes on values that depend on the posited value of γ shown on the x-axis: $\zeta \in \{-\gamma, 0, \gamma\}$. In (b), ζ is constant for all tested values of γ based on our estimated value for γ: $\zeta \in \{-\hat{\gamma}, 0, \hat{\gamma}\}$. In neither case does changing ζ substantially affect inference for the effect of feedback received on posting, γ. The limits of the 95% confidence interval from asymptotic adjacency- and cluster-robust inference are shown as dotted vertical lines.

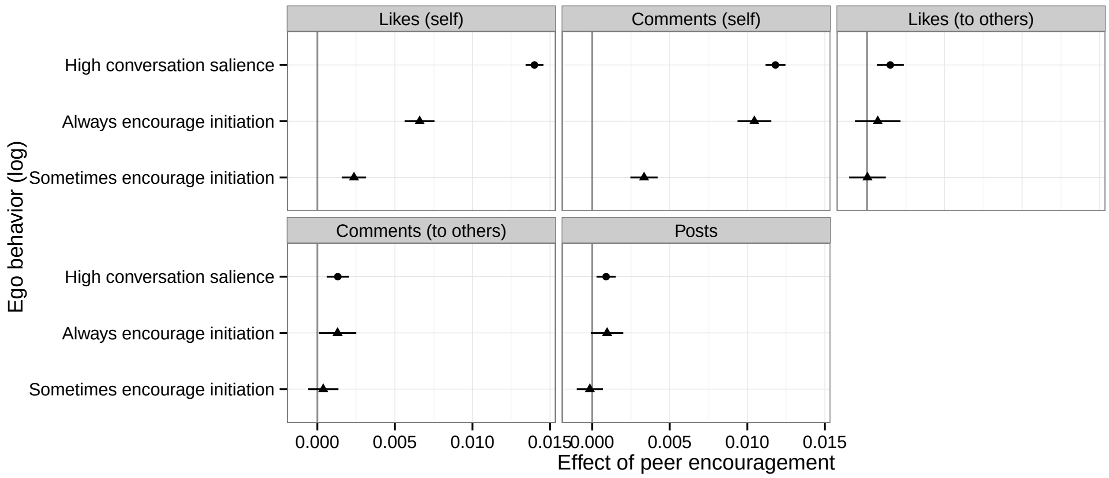


Figure S5: Effects of peer encouragements on (log) ego behaviors. These estimates are based on post-stratification on quartiles of prior feedback, friend activity, and prior sharing. Error bars are 95% heteroscedasticity-robust confidence intervals.

6 Intent-to-treat effects

When analyzing encouragement designs, it is common to report the total effects of random assignment to the encouragement on the outcomes. Figure S5 shows these "intent-to-treat" effects.

7 First-stage distributional effects

In addition to the effects on mean feedback received reported in the main text, we can compare the distributions of feedback received in different encouragement conditions. This shows what changes in feedback received are caused by the peer encouragement and thus what changes the TSLS analysis is averaging over. Figure 3 in the main text illustrates the difference in these distributions for all egos, and Figure S6 shows these distributional effects separately for egos who previously received differing levels of feedback; these are combined to produce the results in the main text.

In particular, in the case of exclusive, binary instruments, the differences in CDFs are the weights that define the ACR for that instrument or the weighted combination of ACRs. Using results extending Angrist and Imbens (1, Th. 2) to TSLS estimation with other TSLS specifications Lochner and Moretti (19, Prop. 2) and accounting for the log transformation, we compute the combined weights for the primary set of three binary

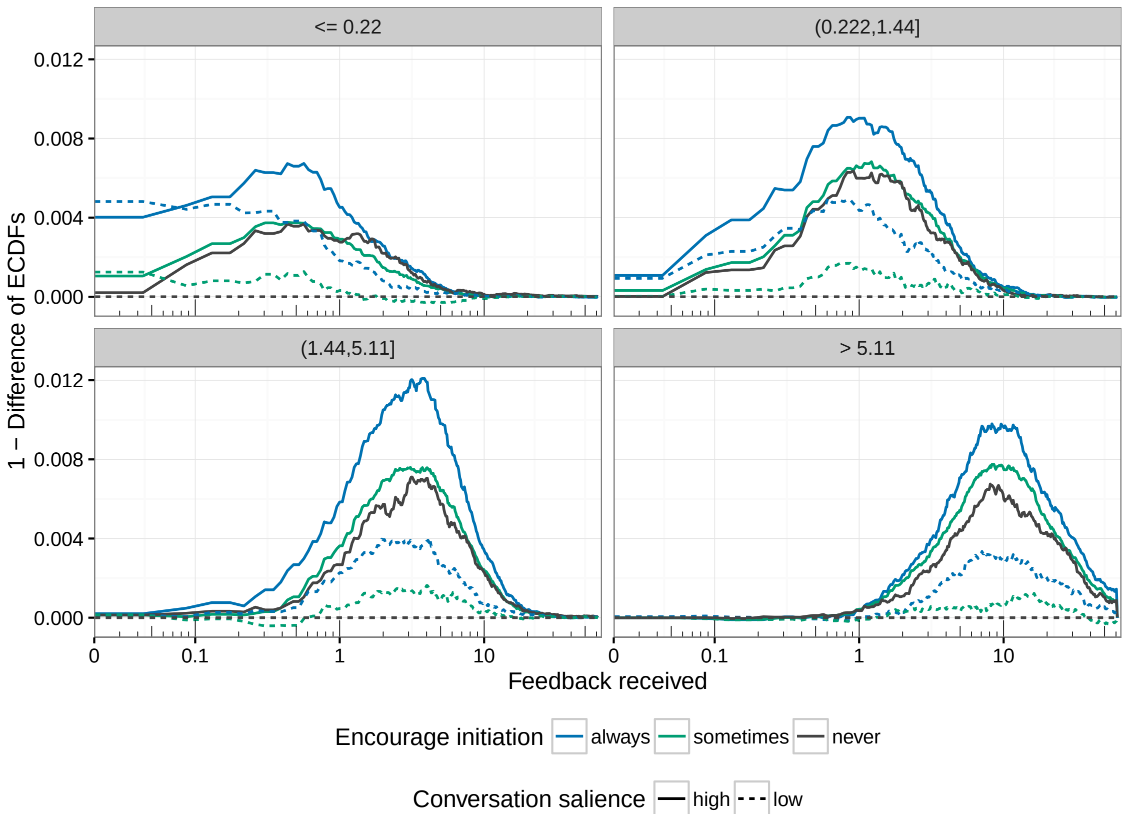
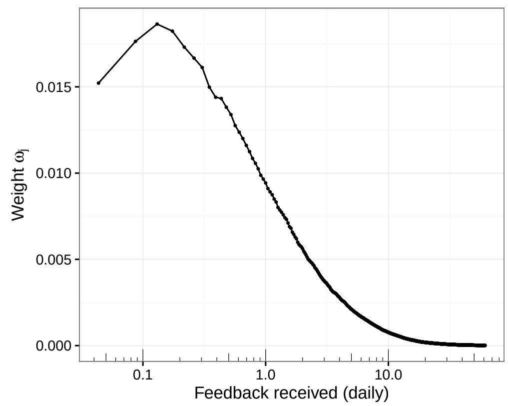
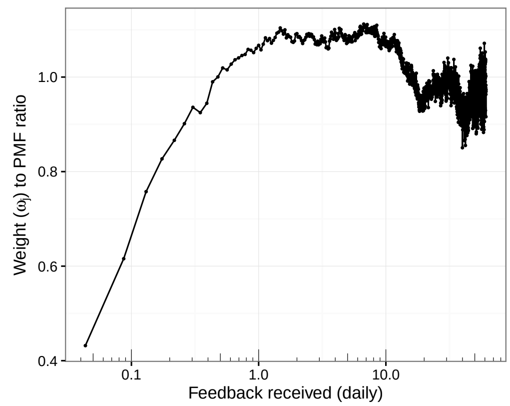


Figure S6: Effect of the encouragements on feedback received, by quartiles of prior feedback received. Using the lowest-feedback condition (never encourage initiation, low conversation salience) as the baseline, the lines represent the difference in probability that feedback received is at least the value on the x-axis. As expected, for egos who received less feedback prior to the experiment (top-left panel), the encourage initiation factor has larger effects relative to the conversation salience factor.



(a)



(b)

Figure S7: (a) Weighting function for TSLS with primary set of three instruments. Each point is the weight on changes from the $(k-1)$th value to the kth value. (b) Ratio of weighting function to probability mass function (PMF) for feedback received (excluding non-zero and maximum, winsorized value).

instruments (the main effects). These are displayed in Fig. S7. The weights are largest for small values of D_i, but this only partially reflects the greater probability mass on these smaller values; in fact, compared with the probability mass function, larger values of D_i are given more weight.

8 Alternative selection of instruments

We also produced estimates from multiple first-stage specifications: models with both factors and models with only the conversation salience factor and only the encourage initiation factor. We also include, following Belloni et al. (7), a lasso (i.e., L1 penalized, (21)) model. The matrix of potential instruments for this model has 325 columns with both factors (3 columns), interactions (2), and interactions with the strata-defining variables ($64 \times 5 = 320$). Note that this is intentionally overparameterized in that terms for all 64 strata (not 63) are included. The selected penalty $\lambda_s = 2.14 \times 10^{-5}$, which minimizes MSE in 10-fold cross-validation, results in a model with 23 of a possible 325 non-zero coefficients (Fig. S8), including the 3 main effects and 20 strata-specific terms.

Figure 4 in the main text and Table 8 present results from these four models. The estimates for most outcomes are statistically indistinguishable. For the "reply" behaviors, the estimates from the two factors are statistically significantly different. This could reflect that these encouragements may produce different types of feedback (e.g., comments vs. likes, or comments with different content), thus affecting the number of targets for these reply behaviors (note, e.g., that only comments, not likes, on the ego's post can themselves

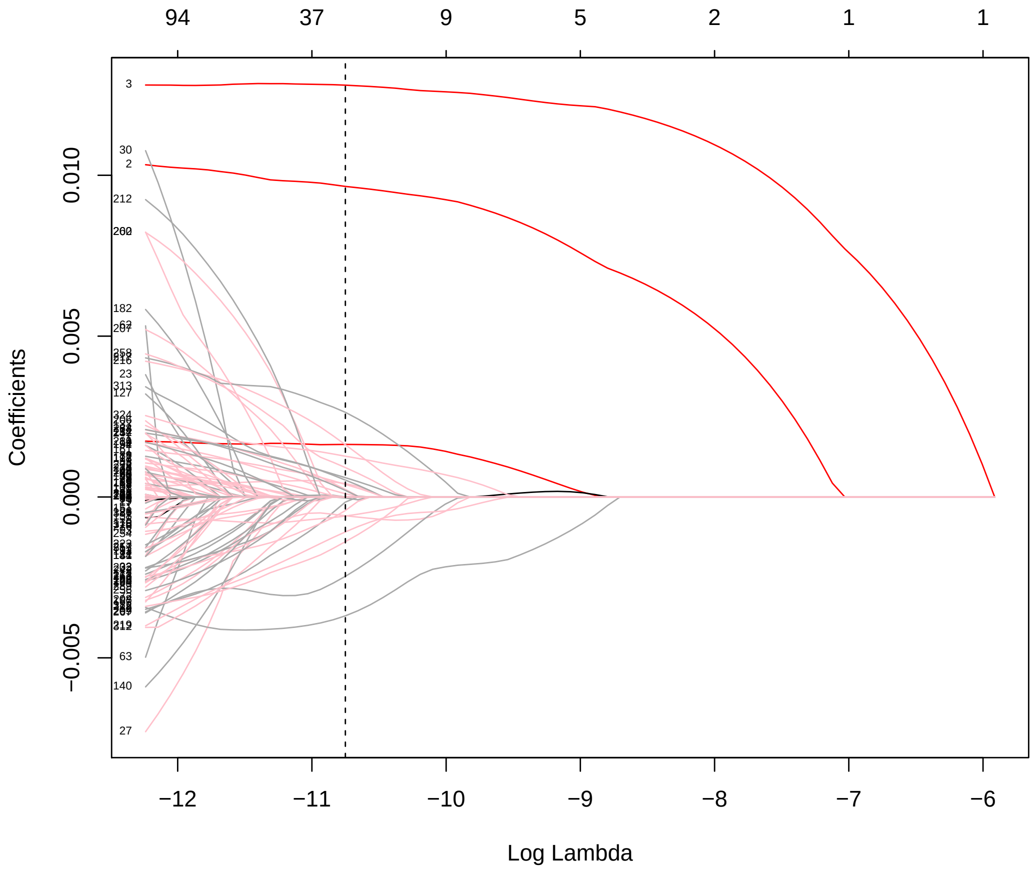


Figure S8: Regularization path for the lasso (L1 penalized) first-stage model. The selected value of the penalty λ is indicated with a dashed vertical line. The coefficient paths are numbered and colored according to: (1–3) main effects (red), (4–5) interactions of the factors (black), (6–325) stratum-specific effects, with main effects (pink) and interactions (grey). As λ is decreased, the first non-zero coefficient is for high conversation salience and the second is for always encourage initiation.

be liked by the ego). The lasso estimates do not significantly differ from the simpler model, suggesting that, at least on the log-transformed scale, there is little heterogeneity in the first stage between the 64 strata, at least to the extent that it is associated with heterogeneous effects of feedback.

Table S2: Effects of receiving feedback on five ego behaviors, as estimated using IV analysis of the peer encouragement design with four different first-stage specifications. These are coefficient estimates from a TSLS log–log model with network adjacency- and cluster-robust standard errors (in parentheses).

Outcome	Main effects	Salience only	Initiation only	Lasso
Likes (self)	1.046	1.184	0.800	1.058
	(0.026)	(0.032)	(0.048)	(0.025)
	z = 40.17	z = 36.47	z = 16.51	z = 41.50
	p < 1e-12	p < 1e-12	p < 1e-12	p < 1e-12
Comments (self)	0.964	0.968	1.060	0.961
	(0.019)	(0.022)	(0.045)	(0.018)
	z = 50.72	z = 44.72	z = 23.78	z = 52.22
	p < 1e-12	p < 1e-12	p < 1e-12	p < 1e-12
Likes (to others)	0.112	0.125	0.078	0.113
	(0.030)	(0.034)	(0.066)	(0.029)
	z = 3.78	z = 3.71	z = 1.18	z = 3.87
	p = 1.6e-04	p = 2.1e-04	p = 2.4e-01	p = 1.1e-04
Comments (to others)	0.105	0.099	0.125	0.106
	(0.024)	(0.028)	(0.053)	(0.024)
	z = 4.33	z = 3.59	z = 2.37	z = 4.46
	p = 1.5e-05	p = 3.3e-04	p = 1.8e-02	p = 8.1e-06
Posts	0.070	0.058	0.064	0.072
	(0.021)	(0.025)	(0.047)	(0.021)
	z = 3.26	z = 2.35	z = 1.36	z = 3.42
	p = 1.1e-03	p = 1.9e-02	p = 1.7e-01	p = 6.2e-04

9 Transformed and untransformed count variables

As with many behaviors in social media, counts of behaviors on Facebook are highly skewed. To guard against extreme values, all quantitative variables counting behaviors were winsorized: we computed the 99th percentile of the non-zero values of the variable, and all values above that were replaced with that value.

Besides the standard motivations for log-transforming thick-tailed count variables, the log-transformed variables lead to intuitively appealing models. We expected the data-generating process in the first stage to be better approximated by a multiplicative model, instead of an additive model. First, people who receive larger amounts of feedback will often have more peers who would be affected by encouragements and more posts to which it would apply. Other aspects of the News Feed system also suggest a multiplicative model. For example, the amount of feedback a story receives is among one of the top signals used to rank items in the News Feed (5), and it is well understood that content in high positions are more likely to be attended to (6, 14). Combined, it is easy to see how additional feedback could increases the likelihood that a post receives more feedback, thus introducing a multiplicative data generating process.

In the second stage model, similar general data analytic considerations apply. Furthermore, even if one expected that one additional like or comment would have the same effect for egos who receive different amounts of feedback, this is also consistent with the log-log model when the baseline levels of outcomes and feedback received are highly correlated. We therefore used a model with a log-log parameterization in our primary analyses.

We therefore used log-transformations of the quantitative variables counting behaviors in our primary analyses (we also provide estimates from linear models below). In particular, we transform these variables by adding one prior to dividing by the number of days[3] in the corresponding period and then take the natural log:

$$y = \log((y^* + 1)/n_{\text{days}}).$$

How the transformation of the outcome changes the estimand is straightforward. We can also state how the estimand of TSLS is changed by the transformation of the endogenous variable (feedback received). Let D^* be the untransformed endogenous variable. Following Angrist and Imbens (1, Th. 1), the TSLS estimand for a single binary instrument is

$$\frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\mathrm{E}[D^*|Z=1] - \mathrm{E}[D^*|Z=0]} = \sum_{j=1}^{J} w_j \mathrm{E}[Y(j) - Y(j-1)|D^*(1) \geq j > D^*(0)]$$

where

$$w_j = \frac{\Pr(D^*(1) \geq j > D^*(0))}{\sum_{i=1}^{J} \Pr(D^*(0) \geq i > D^*(0))}$$

are weights that correspond to normalized differences in CDFs between $D^*(1)$ and $D^*(0)$. We define $D = g(D^*)$. D, like D^*, still takes on J values. The numerator is unchanged by working with D instead of D^*. However, the denominator changes so that the weights

[3]This rescaling is only of consequence for the non-log-transformed results in this section.

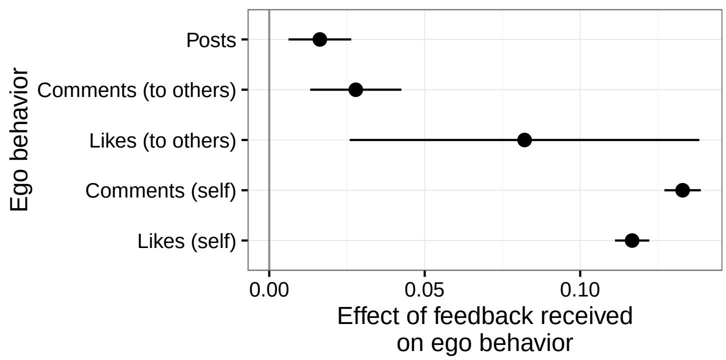


Figure S9: Effects of receiving feedback on five ego behaviors, as estimated using TSLS. Unlike in the main text, these results derive from variables that are not log-transformed. Error bars are 95% adjacency- and cluster-robust confidence intervals.

change:

$$\frac{\mathrm{E}[Y|Z=1]-\mathrm{E}[Y|Z=0]}{\mathrm{E}[D|Z=1]-\mathrm{E}[D|Z=0]}=\sum_{j=1}^{J} w_j \mathrm{E}[Y(j)-Y(j-1)|D(1)\geq g(j)>D(0)]$$

where

$$w_j=\frac{[(g(j)-g(j-1)]\Pr(D\geq g(j)>D(0))}{\sum_{i=1}^{J}[g(i)-g(i-1)]\Pr(D\geq g(i)>D(0))}.$$

As would be expected, with $g(x)=log(x+c)$, then $g(j)-g(j-1)$ decreases with j.

Similar modification apply to the other results in Angrist and Imbens (1), as this similarly affects how TSLS with multiple instruments combines the individual Wald estimates.

For comparison, Fig. S9 presents TSLS results with untransformed versions of these variables (i.e., with $y=y^*/n_{\text{days}}$ and $d=d^*/n_{\text{days}}$).

10 Simulations with ego-specific and general designs

We compare the statistical properties of the current ego-specific encouragement design with the properties of the more common general encouragement design. Based on simulated random graphs, we compute the true SE of the TSLS estimator under various conditions. To compare the two designs, we vary the random assignment and the specification of the first-stage model while keeping the second stage constant. In the ego-specific design, we randomly assign half of the nodes to be treated — implicitly assigning all of an assigned ego's peers. The ego-specific design is specified as

$$D_i=\beta Z_i+\eta_i$$

$$Y_i = \gamma D_i + 0.5\eta_i + \epsilon_i$$

with the first-stage effect size β, target parameter γ, ego-specific assignment indicator z_i (i.e., the instrument), and noise from a standard normal distribution in the first stage, η_i, and second stage, ϵ_i; the common error term η_i in the first and second stage results in confounding bias in the absence of the instrument.

Since the ego-specific design results in all peers being encouraged to a behavior directed at the ego, the equivalent in the general (non-ego-specific) design is when all peers happen to be assigned to be encouraged to that behavior (towards all of their neighbors). To achieve the same sized shock in the first stage in the general design with everyone assigned as in the specific design with $Z_i = 1$, we used the fraction of peers assigned to the encouragement as the instrument in the general design. The general design is specified as

$$D_i = \beta W_i + \eta_i$$

$$Y_i = \gamma D_i + 0.5\eta_i + \epsilon_i$$

with the proportion of an ego's assigned peers $W = \tilde{\mathbf{A}}Z$ as the instrument, but otherwise unchanged from the ego-specific specification.

We simulated 5,000 TSLS estimates based on the Watts–Strogatz small-world network model (24) for different numbers of units ($\log_2 n \in \{7, ..., 12\}$), rewiring probabilities $p_{\mathrm{rw}} \in \{0.00, 0.01, 0.10\}$, neighborhood sizes (nei $\in \{1, 2, 5\}$, corresponding to average degree of 1, 4, and 10), and effect sizes ($\beta = 1$ and $\gamma \in \{0.0, 0.5, 1.0\}$). We use common random numbers for η and ϵ so that the randomness within the 5,000 replicates of each configuration arises from the random assignment of Z; that is, the potential outcomes are fixed.

Across all settings and as expected, the ego-specific design resulted in increased precision of $\hat{\gamma}$ and power to detect non-zero γ, compared with the general encouragement, as shown in Figs. S11 and S11.

11 Simulations with interference: Type I error rates of tests

Using simulations very similar to those in the previous section for ego-specific designs, we illustrate the performance of four methods for constructing tests for $\gamma = 0$. We use the same generative model as in the previous section, but add local interference, as in the model (Equation 4) posited by our sensitivity analysis in Section 5.2. The generative model is:

$$D_i = \beta Z_i + \eta_i$$

$$Y_i = \gamma D_i + (\tilde{\mathbf{A}}_i \mathbf{Z})\zeta + 0.5\eta_i + \epsilon_i.$$

The methods used are two tests that were expected to not be robust to interference and the two related methods we used in the main text and in the sensitivity analysis:

1. Heteroskedasticity-robust sandwich estimator for independent data,

2. Adjacency-robust sandwich estimator,
3. Wilcoxon rank-sum test,
4. Wilcoxon rank-sum test with the interference model of Equation 4.

We simulated 5,000 TSLS estimates and associated tests for each combination of $\log_2 n \in \{7, ..., 12\}$, $p_{\mathrm{rw}} = 0.01$, neighborhood sizes (nei = 2), effect sizes ($\beta \in \{0.1, 0.5, 1.0\}$ and $\gamma = 0$), and interference ($\zeta \in \{0, 1, 2\}$). That is, when $\zeta > 0$, these simulations use very large interference, larger than even the first-stage effects.

For the Wilcoxon rank-sum test with the interference model, we use $\zeta_0 \in \{-\zeta, 0, \zeta\}$; that is, zero, the true ζ, and its negation. We select the maximum p-value for $\gamma = 0$, as in Section 5.2. In contrast to our sensitivity analysis, where we set $\zeta \in \{-\gamma, 0, \gamma\}$, we here set $\gamma = 0$.

The results are shown in Fig. S12. In the absence of interference ($\zeta = 0$), all tests have size (Type I error rates) close to the nominal size of $\alpha = 0.05$ for these settings. However, in the presence of interference, the two tests for independent data have larger-than-nominal size. On the other hand, the test using the adjacency-robust sandwich estimator and the Wilcoxon rank-sum test with the interference model both exhibit size $\leq \alpha$.

References

[1] Angrist, J. and Imbens, G. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

[2] Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16.

[3] Aronow, P. M., Samii, C., and Assenova, V. A. (2015). Cluster-robust variance estimation for dyadic data. *Political Analysis*, 23(4):564–577.

[4] Athey, S., Eckles, D., and Imbens, G. W. (2015). Exact p-values for network interference. NBER Working Paper 21313. http://arxiv.org/abs/1506.02084.

[5] Backstrom, L. (2013). News feed fyi: A window into news feed. https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed.

[6] Bakshy, E., Messing, S., and Adamic, L. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

[7] Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

[8] Bester, C. A., Conley, T. G., and Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, 165(2):137–151.

[9] Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

[10] Bowers, J., Fredrickson, M. M., and Panagopoulos, C. (2013). Reasoning about interference between units: A general framework. *Political Analysis*, 21(1):97–124.

[11] Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2).

[12] Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of econometrics*, 92(1):1–45.

[13] Conley, T. G. and Molinari, F. (2007). Spatial correlation robust inference with errors in location or distance. *Journal of Econometrics*, 140(1):76–96.

[14] Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 87–94. ACM.

[15] Eckles, D., Karrer, B., and Ugander, J. (2015). Design and analysis of experiments in networks: Reducing bias from interference. http://arxiv.org/abs/1404.7530.

[16] Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482).

[17] Imbens, G. W. and Rosenbaum, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):109–126.

[18] Kelejian, H. H. and Prucha, I. R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics*, 140(1):131–154.

[19] Lochner, L. and Moretti, E. (2015). Estimating and testing models with many treatment levels and limited instruments. *Review of Economics and Statistics*, 97(2):387–397.

[20] Ogburn, E. L., VanderWeele, T. J., et al. (2014). Causal diagrams for interference. *Statistical Science*, 29(4):559–578.

[21] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[22] Ugander, J. and Backstrom, L. (2013). Balanced label propagation for partitioning massive graphs. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 507–516. ACM.

[23] Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. M. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proc. of KDD*. ACM.

[24] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2.

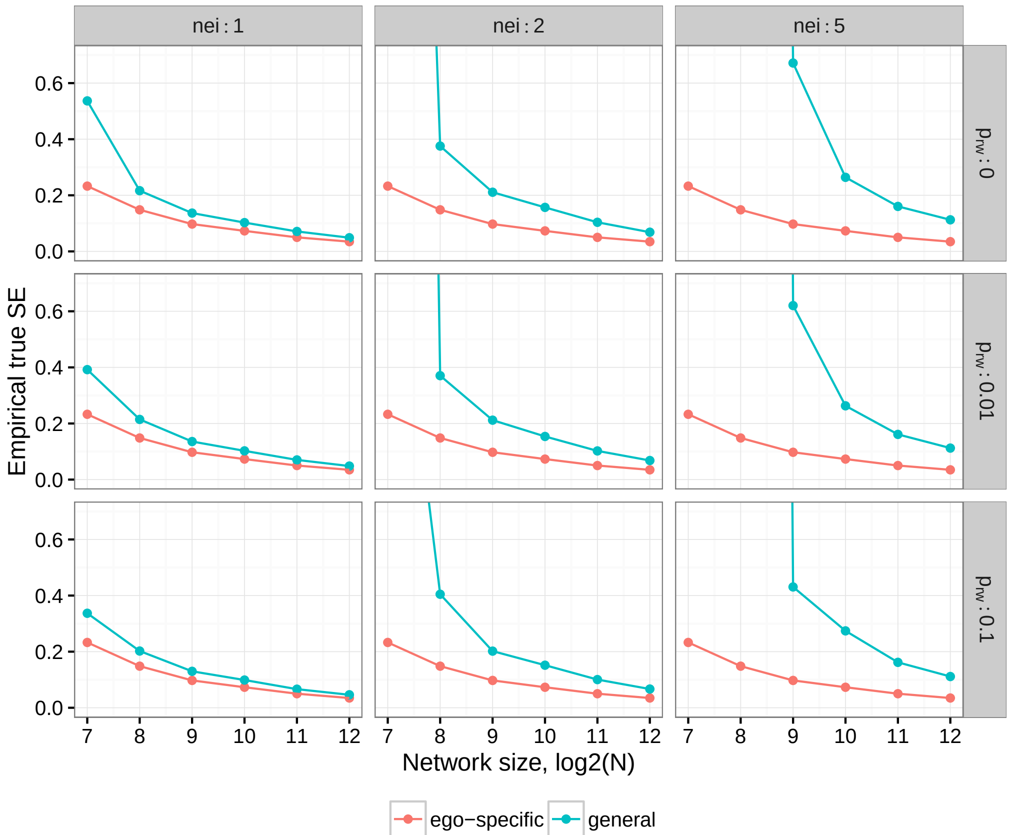


Figure S10: True SE for estimates of γ in ego-specific and general peer encouragement designs from simulations with small-world networks of different size (n), varying number of neighbors (nei), and different rewiring probabilities. The true standard error is estimated with the standard deviation of $\hat{\gamma}$ over 5,000 draws of Z. These results do not change with γ; results for $\gamma = 1$ are shown.

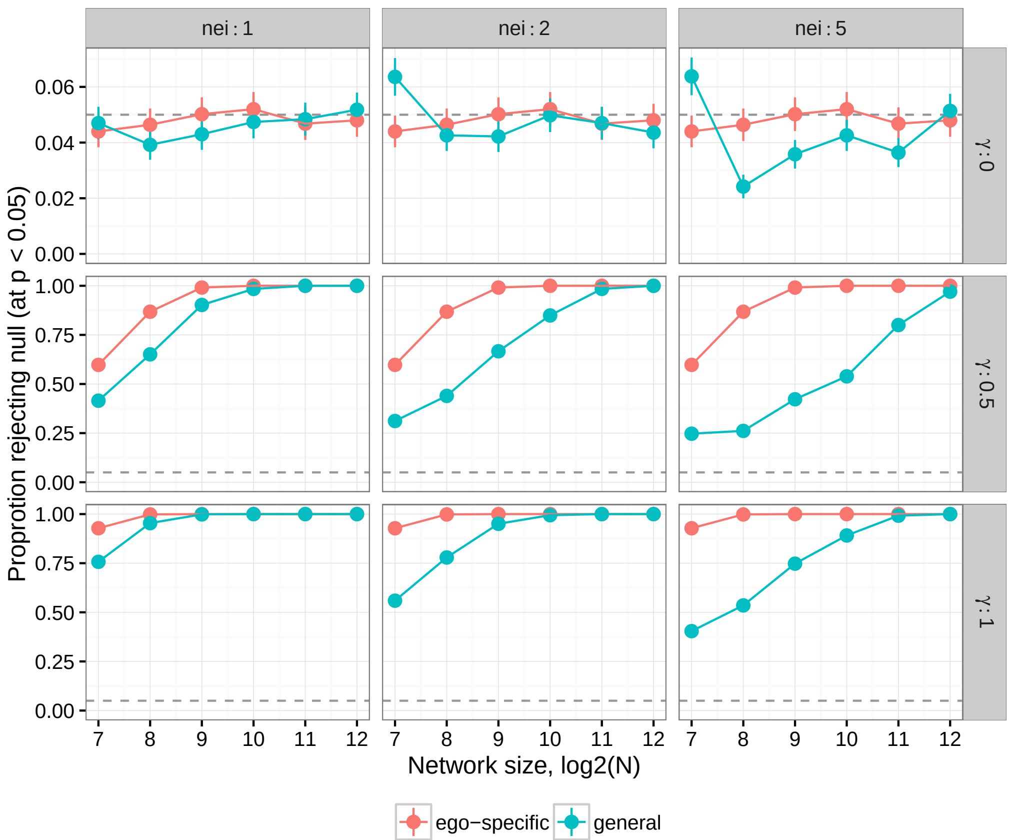


Figure S11: Rate of rejecting the null of $\gamma = 0$ in ego-specific and general peer encouragement designs from simulations with small-world networks of different size (n), varying number of neighbors (nei), and with different true effect sizes (γ). When $\gamma = 0$, this is the empirical size (Type I error rate) of the test, which appears to have size less than $\alpha = 0.05$, except for small n with the general design. When $\gamma \neq 0$, this is the power of the test. The p-values were computed using asymptotic adjacency-robust sandwich standard errors. Error bars are 95% confidence intervals for a proportion.

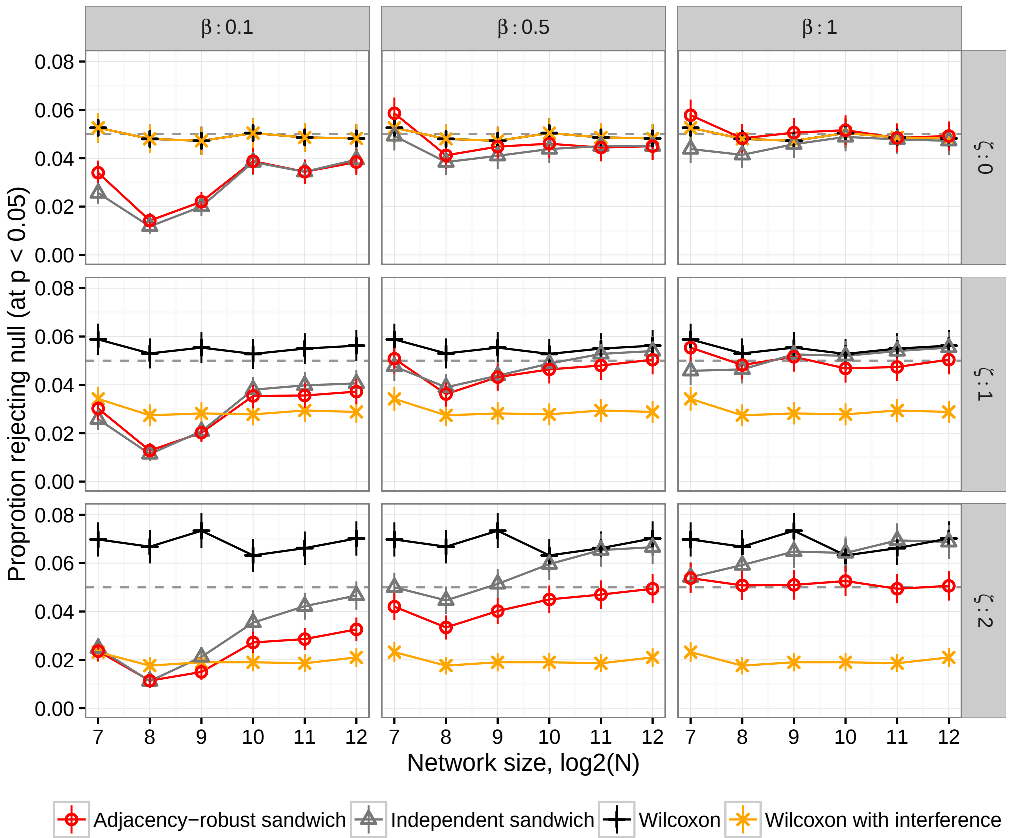


Figure S12: Size (Type I error rate) of tests for empirical size (Type I error rate) for tests of $\gamma = 0$ in the presence of varying levels of interference, ζ. In the case of $\zeta = 0$, the two Wilcoxon tests are identical by construction.