DESIGNING EFFECTIVE CELEBRITY MESSAGING: RESULTS FROM A NATIONWIDE TWITTER EXPERIMENT PROMOTING VACCINATION IN INDONESIA

VIVI ALATAS[⋆], ARUN G. CHANDRASEKHAR[†], MARKUS MOBIUS[§], BENJAMIN A. OLKEN[‡], AND CINDY PALADINES[⋆⋆]

Abstract. Do celebrity endorsements matter? And if so, how can celebrities communicate effectively? We conduct a nationwide Twitter experiment in Indonesia promoting vaccination. Celebrity messages are 72 percent more likely to be passed or liked than similar messages without a celebrity's imprimatur. Decomposing this, 79 percent of the celebrity effect comes from authorship, compared to passing on messages. Citing external medical sources decreases retweets by 27 percent. Phone surveys show that those randomly exposed to messaging have fewer incorrect beliefs and report more vaccination among friends and neighbors. The results can inform public health campaigns, such as for COVID-19 vaccination.

Date: February 25, 2022.

We thank Ran Abramitsky, Marcella Alsan, Nancy Baym, Emily Breza, Leo Bursztyn, Rebecca Diamond, Dean Eckles, Paul Goldsmith-Pinkham, Ben Golub, Mary Gray, Rema Hanna, Johannes Haushofer, Matt Jackson, Tyler McCormick, Madeline McKelway, Matthew Wai-Poi, Alex Wolitsky, and participants at various seminars for helpful discussions. Aaron Kaye, Nurzanty Khadijah, Devika Lakhote, Eva Lyubich, Sinead Maguire, Lina Marliani, Sebastian Steffen, Vincent Tanutama provided excellent research assistance. We thank Nila Moeloek, then Indonesian Special Envoy for Sustainable Development Goals, Diah Saminarsih, and their team for providing support for this project. This study was approved by IRBs at MIT (Protocol #1406006433) and Stanford (Protocol #31451), and registered in the AEA Social Science Registry (AEARCTR-0000757). Funding for this project came from the Australian Government Department of Foreign Affairs and Trade. The views expressed here are those of the authors only and do not represent those of any of the institutions or individuals acknowledged here.

[⋆]Asa Kreativita.

[†]Department of Economics, Stanford University; NBER; J-PAL.

[§]Microsoft Research, New England.

[‡]Department of Economics, MIT; NBER; J-PAL.

[⋆⋆]World Bank.

1. Introduction

Social media has allowed celebrities to take an increasing role in social discourse. With millions of followers, celebrities have a channel to spread messages on many issues, including some far removed from their original reason for fame. Their participation in ongoing discussions can make issues prominent and shape the zeitgeist.

Examples abound, from #BlackLivesMatter, for racial justice, to the #IceBucketChallenge, promoting awareness of Lou Gehrig's disease. Each of these campaigns was initiated by a less-well-known activist, but made prominent in part through celebrity participation. Celebrities are now increasingly being recruited as public health messengers.

Using celebrities effectively, however, depends on what features of the celebrity messaging spur diffusion. Beyond the direct *reach* that celebrities have – they have numerous followers – there are several dimensions that could affect people's decision to further spread or follow a celebrity's message. First, how much does celebrity *endorsement*, meaning involvement with the messaging beyond the simple fact that they have numerous followers, matter? Second, how much of an endorsement premium comes from a direct *authorship* effect, as opposed to relaying the message of others? Third, how much does inclusion of *credible sourcing* matter, particularly when celebrities are speaking on a topic removed from their core area of expertise, such as public health?

While celebrity endorsements may have effects for commercial products, whether their endorsement matters in the context of public health remains a matter of public debate, especially since public health issues are often far from their main area of expertise. For instance, in a May 2021 New York Times article (Ives, 2021), epidemiologists and psychologists, based on focus groups, argue that celebrity endorsements may not address COVID-19 vaccine indifference and hesitancy.[1] Despite widespread interest in using celebrities to promote vaccination, there remains little rigorous empirical evidence on whether this matters, and if so, how to effectively design celebrity outreach campaigns.

Measuring and decomposing the endorsement, authorship, and credible sourcing effects, as well as measuring whether any of these campaigns have offline effects, is challenging, as celebrities' decisions are typically endogenous, and because people consume such a wide range of information that it is challenging to isolate their impact.

To study these issues, we conducted an experiment through a nationwide immunization campaign on Twitter from 2015-2016 in Indonesia, in collaboration with the Indonesian Government's Special Ambassador to the United Nations for Millennium Development Goals. Working with the Special Ambassador, we recruited 46 high-profile celebrities and organizations, with a total of over 11 million followers, each of whom gave us access to send up to 33 tweets or retweets promoting immunization from their accounts. The content and timing

[1] See https://www.nytimes.com/2021/05/01/health/vaccinated-celebrities.html

of these tweets was randomly chosen from a bank approved by the Indonesian Ministry of Health, all of which featured a campaign hashtag #AyoImunisasi ("Let's Immunize").

To isolate the role of celebrity endorsement and sourcing, we randomly varied (1) Did the celebrity / organization send the tweet directly, or did they retweet a message (drawn randomly from the same tweet library) sent by us from a non-celebrity user's account?; (2) Did the tweet explicitly cite a public health source? We also randomly varied (3) When did the celebrity tweet?

We study the effects of this induced variation in two ways. First, we use online reactions, i.e., likes and retweets, so we can observe the online reactions of every follower to each tweet, and trace out which randomized characteristics of the tweet lead to more approval and diffusion.[2] Second, we also study the offline effects of exposure to the campaign by conducting phone surveys of Twitter users. By randomly allocating celebrity activity into one of several phases – either before or after the survey – we can examine whether individuals who follow celebrities randomized to tweet before the survey are more likely to have heard of the campaign, updated their beliefs, discussed immunization status with their friends and neighbors, and observed changes in immunization behavior among their friends, relatives, and neighbors.

Our study builds on several facets of the literature. First, we build on the recent literature on diffusion of information for public policy and the computer science literature on generating online cascades (e.g, Leskovec, Adamic, and Huberman, 2007; Katona, Zubcsek, and Sarvary, 2011; Bakshy, Hofman, Mason, and Watts, 2011; Banerjee, Chandrasekhar, Duflo, and Jackson, 2013; Beaman, BenYishay, Magruder, and Mobarak, 2021; Beaman and Dillon, 2018). While this literature has studied the flow of information over social networks, and how network position affects the flow of information, it has typically been silent on what aspects of the message matter.

Second, we build on the extensive literature that has studied celebrity endorsements, primarily in commercial advertising. This literature has examined the impacts of celebrity endorsements of commercial products on outcomes such as stock prices (e.g., Agrawal and Kamakura, 1995), sales (e.g., Elberse and Verleun, 2012; Garthwaite, 2014), and brand evaluations, and studying various aspects of the celebrity's identity (e.g., gender, attractiveness); see Bergkvist and Zhou (2016) for a comprehensive review. Our study is one of the first to study celebrity effects in the online space through a real-world, large-scale field experiment, and the first large-scale field experiment we know of to study public health messaging of any type. Indeed, the only study of a similar magnitude we know of is a marketing study by Gong, Zhang, Zhao, and Jiang (2017), who experimentally vary tweets in China on Sina Weibo about TV programs, randomizing whether these tweets were retweeted by influencers.

[2]Note that on Twitter, a "like" is not pushed to one's followers, while a "retweet" subsequently passes on the tweet to all of one's followers.

Our study builds on this by *decomposing* the celebrity effect, identifying the value of authorship *per se* as opposed to relaying others' messages, and identifying the value of credible sources, i.e., health authorities in a public health context.

2. Experiment

2.1. **Setting and Sample.** Our study took place in Indonesia in 2015 and 2016 on Twitter, which is one of the most important mediums of information exchange in the world, with over 1 billion users. Indonesia is very active on social media, ranking third worldwide with 130 million Facebook accounts[3] in 2020 (about half the population), and ranking eighth with 10.6 million Twitter accounts (about 6.4 percent of the population).[4]

The experiment focused on immunization. At the time, Indonesia was trying to improve immunization as part of its drive towards the Millennium Development Goals. A set of 550 tweets was developed in coordination with the Ministry of Health that sought to improve information about immunization. The tweets included information about access (e.g., immunizations are free, available at government clinics, and so on); information about immunization's importance (e.g., immunizations are crucial to combat child diseases); and information designed to combat common myths about immunization (e.g., vaccines are made domestically in Indonesia, rather than imported). For each tweet, we identified a source (either a specific link or an organization's Twitter handle). All tweets were approved by the Ministry of Health, and included a common hashtag, #AyoImunisasi ("Let's Immunize").

With help from the Indonesian Special Ambassador to the United Nations for Millennium Development Goals, we recruited 37 high-profile Twitter users, whom we denote "celebrities," with a total of 11 million Twitter followers. These "celebrities" come from many backgrounds, including music stars, TV personalities, actors and actresses, motivational speakers, government officials, and public intellectuals. They have a mean of 262,647 Twitter followers each, with several having more than one million. While these celebrities primarily tweet about things pertaining to their reason for fame, they also comment occasionally on public issues, so tweets about immunization would not necessarily have been unusual.[5] We also recruited 9 organizations involved in public advocacy and/or health issues in Indonesia with a mean of 132,300 followers each.

---

[3] https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/

[4] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[5] For example, three celebrities in our sample (a musician, a TV personality, and a musician's agent) had recently tweeted about the importance of breakfast, including a link to an article about the health benefits of children's breakfast; an athlete tweeted about supporting education for young children; and a musician tweeted in support of Asia Against AIDS.

In addition, we recruited 1,032 ordinary citizens, primarily Indonesian university students, whom we call "Joes and Janes". Their role will be to allow us to have everyday individuals compose tweets that are then retweeted by celebrities. Their Twitter profiles are far more typical, with a mean of 511 followers.

Every participant (celebrities and Joes/Janes) consented to signing up with our app that (1) lets us tweet content from their account (13, 23, or 33 times), (2) randomizes the content of the tweets from a large list of 550 immunization tweets approved by the Ministry of Health, and (3) has no scope for editing.[6] Participants were given two choices: (1) the maximum number of tweets (13, 23, or 33), and (2) a choice of formal or casual Indonesian language (to better approximate their normal writing style).

2.2. **Experimental Design.** Our experiment is designed to understand which aspects of social media campaigns are important for disseminating a message. Ex ante it may seem obvious, for instance, that sources are better (after all, the information is more credible) and celebrity involvement is better (after all, for a variety of reasons the information may be viewed as more credible). But thinking carefully about the information sharing process demonstrates that, in fact, the effect of each of these design options is actually theoretically ambiguous.[7]

We focus on two main interventions: (1) whether a tweet was tweeted directly by a celebrity, or tweeted by a Joe/Jane and then retweeted by a celebrity; and, (2) for a subset of tweets, whether the tweet included a credible source (i.e., the source link or referring organization's Twitter handle).[8] We also randomized the timing of tweets (matching the empirical frequency of local time-of-day of Indonesian tweets), and the content (i.e., which tweet from our pre-prepared bank of approved tweets was tweeted by whom and when).[9]

Randomizing whether a tweet was directly tweeted by a celebrity or retweeted – in combination with the particular way Twitter messages are shown – allows us to develop a novel test for isolating celebrity effects *per se*. Specifically, we use the fact that Twitter messages show the identity of only two people: the originator who wrote the tweet, and the person whom you follow who directly passed it to you (call this person F). Other steps along the chain are omitted.

[6]Celebrities could veto a tweet if they did not want it sent from their account, though this in fact never happened.

[7]Appendix F presents an application of a simple model developed independently by Chandrasekhar, Golub, and Yang (2018) to demonstrate the ambiguity, though certainly other models can be used and this is inessential for the empirical analysis.

[8]A small subset of tweets on topics deemed 'sensitive' by the Government always included a source; these are excluded from the analysis of sourcing.

[9]Note that in the period we study, a Twitter user saw all tweets and retweets from the users they follow in strict reverse chronological order (i.e., newest tweets appeared first, and so on). Twitter subsequently (in March 2016) applied an algorithm to prioritize the ordering of the tweets, but since in the period we study (July 2015 through February 2016) tweets appeared in strictly chronological order, nothing in our experimental design affects the ordering of tweets in a user's Twitter feed.

Consider what happens when some celebrity followers (whom we denote F_1) retweet a message to their followers, whom we denote F_2, depicted in Figure 1. If the celebrity authored the original message (Panel A), the followers-of-followers (F_2s) observe that the celebrity authored the message and that F_1 retweeted it, as seen in Panel B. But if the celebrity retweeted a message from a Joe/Jane (Panel C), the followers-of-followers of the celebrity (F_2s) observe only that the Joe/Jane tweeted and that F_1 then retweeted for F_2 to see, as displayed in Panel D. The F_2 does not observe the celebrity involvement at all in the second case.

We therefore can study the retweet behavior of F_2s – i.e., followers-of-followers-of-celebrities – across these two cases to identify the endorsement effect premium, i.e., the additional effect of their knowing that C originated the message, holding F_2's network position fixed.[10]

We can then further decompose the endorsement effect to understand the impact of celebrities speaking in their own voice (an 'authorship' effect). We use the same experimental variation, but look at behavior of the direct followers of celebrities (F_1s), who see both the celebrities' directly-authored tweets and the celebrities' retweets.

To identify the impact of sources, we explicitly randomize whether a source was included in the tweet, and examine the behavior of the F_1s.

Finally, to measure offline effects, celebrities were randomized into two phases, with half tweeting in the first phase (Phase I, July and August 2015) and half in the second phase (November 2015 - February 2016, Phases II and III). In addition, towards the end of the last phase, all tweets / retweets by a celebrity were then retweeted by a randomly selected number of Joes/Janes (Phase III). We conducted a survey between phases (August - October 2015) of a subset of followers of our celebrities and we use this between-celebrity randomization to estimate the impact of the Twitter campaign on offline beliefs and behaviors.

2.3. **Data.**

2.3.1. *Online data.* We collected detailed data via the Twitter Firehose and API. Before the experiment began, in early 2015, we collected a baseline image of the publicly available Twitter network, including the list of followers of any celebrity participating in our study.

On Twitter, followers can take two primary actions: "likes" and "retweets". While likes are public, likes are not automatically pushed out as tweets to a user's followers.

For each the of the 672 total tweets originated by our experiment, we tracked each time the tweet was liked or retweeted by any of the over 7.8 million unique users who followed at least one of the participants in our study. When the tweet was retweeted by a celebrity's follower, we also scraped all of this follower's followers and their liking and retweeting behavior.

[10]A challenge is that the F_1 decision to retweet may be endogenous. We discuss this issue in detail in Section 3.1.1, and show that the results are largely similar in the subset of cases where F_1s were also study participants whom we randomly selected and had retweet exogenously, and hence the sample of exposed F_2s is identical.

We denote those retweets / likes coming from a direct follower of a celebrity as F_1 events, and those retweets / likes coming from a follower of a follower of a celebrity as F_2 events. Appendix Table A.1 reports descriptive statistics.

2.3.2. *Offline data.* To measure whether online conversations led to offline behavioral changes, we conducted a phone survey on a sample of 2,441 subjects, all of whom followed at least one of our study participants on Twitter. These subjects were recruited primarily via ads placed on the Twitter platform. The phone survey was designed to capture information about immunization (including beliefs in some of the various myths our tweets were intended to counteract), immunization history for children in the family and knowledge of recent immunizations of children of close friends, and questions about immunization and Twitter.

To recruit this sample, we advertised with promoted tweets on Twitter a recruitment to participate in a healthcare survey, targeted to the 7.8 million unique users who followed participants in our study. This process resulted in 2,441 total subjects. All respondents were surveyed by phone during the endline period; we also contacted a subsample of these respondents (approximately 73 percent) by phone for a baseline survey prior to the beginning of Phase I tweets.

Appendix Table A.1, Panel B reports demographics of our offline survey sample. To gauge the sample selection in our sample, we also present comparable data from the 2014 wave of the SUSENAS, the annual representative Indonesian national household survey. Relative to the nationally representative SUSENAS sample, we see that our demographic is more urban, slightly younger, and has a similar gender composition.

Panel C reports baseline statistics for beliefs about vaccinations. We see that there is considerable confusion about the nature and value of vaccines. For instance, only 56 percent of individuals thought that vaccines are domestically made (they are), and only 38.5 percent thought that vaccines are free of cost (they are). This suggests substantial room for improvement on immunization knowledge in our study sample.

2.4. **Estimation.** We estimate three models. First, to estimate the overall effect of endorsement, we focus on the behavior of followers-of-followers-of-celebrities – i.e., F_2s – and estimate by Poisson regression, the equation

$$\mathrm{E}[y_{trcmp}|\mathbf{x}_{trcmp}] = \exp\left(\alpha \cdot \mathrm{Celeb}_{tcm} + \beta \cdot \log(\mathrm{Followers})_r + \omega_c + \omega_m\right) \tag{2.1}$$

where t indexes tweets, r indexes retweeters (i.e., an F_1 who retweeted the tweet t), c indexes celebrities, m indexes the type of message content, and p indexes phase. The variable Celeb_{tcm} is a dummy that takes 1 if the celebrity authored the tweet herself (and hence her identity is visible to F_2), and 0 otherwise (and hence her identity is not visible to F_2). Each observation is a retweet of one of our original tweets, and the dependent variable y_{trcmp} is a

count of how many times this retweet was itself either liked or retweeted again by an F_2. Since y is a count, we estimate a Poisson regression, with robust standard errors clustered at the original tweet (t) level. We control for the log number of followers of F_1, and for dummies (ω_m) for different types of messages (e.g., dummies for it being about a fact, importance of immunization, etc.). All regressions include celebrity fixed effects (ω_c), which absorb variation in casual/formal style, etc. The coefficient of interest is α, which measures the differential impact of the tweet having been written by the celebrity (as compared to being written by a Joe/Jane) and this being observable to the F_2 deciding whether to retweet.

Second, to decompose the celebrity effect further, we restrict attention to direct followers of the celebrity (F_1 individuals), and estimate

$$\mathrm{E}[y_{tcmp}|\mathbf{x}_{tcmp}] = \exp\left(\alpha \cdot \mathrm{Celeb}_{tcm} + \omega_c + \omega_m\right). \tag{2.2}$$

We now have one observation per tweet, and look at the number of retweets/likes, retweets, or likes by F_1s who are distance 1 from the celebrity. We continue to include celebrity (ω_c), and message-type (ω_m) fixed effects. We run an analogous regression replacing *Celeb* with *Source* to study the impact on F_1 behavior of including public health authority sources.

Finally, to examine offline effects, we turn to our phone survey data. We define Exposure to Tweets$_i$ as the number of campaign tweets that i is randomized to see through Phase I (normalized to have standard deviation 1). Potential Exposure$_i$ is the total number of campaign tweets that i could potentially see through the campaign given the celebrities she follows at baseline (i.e., had all the celebrities she followed been randomzied to tweet in Phase I). The experimental design of randomizing celebrities into phases means that, while individuals i may differ in the number of our celebrities they follow, Exposure to Tweets$_i$ is random conditional on Potential Exposure$_i$.

We therefore run logistic regressions of the form

$$f(y_i) = \alpha + \beta \cdot \text{Exposure to Tweets}_i + \gamma \cdot \text{Potential Exposure}_i + \delta' X_i, \tag{2.3}$$

where y_i is the outcome for respondent i and $f(\cdot)$ is log-odds, i.e., $\log\left(\frac{\mathrm{P}(y_i=1|\mathbf{x}_i)}{1-\mathrm{P}(y_i=1|\mathbf{x}_i)}\right)$. X are controls, such as the number of celebrities followed by i, the log of the number of followers of celebrities by i, survey dates, and (in some specifications) demographics and baseline beliefs, selected here and in subsequent regressions by double post-LASSO (Belloni, Chernozhukov, and Hansen, 2014a,b). We report standard errors and p-values clustered at the level of the combination of celebrities followed; further, because of the complex nature of the potential correlation in Exposure to Tweets$_i$ across individuals i induced by partial overlap in which celebrities our survey respondents follow, we present randomization-inference (RI) p-values as well.

3. Results

3.1. **Decomposing The Endorsement Effect: Authorship versus Relaying Messages, and Citing Public Health Authorities.**

3.1.1. *Measuring the Total Endorsement Effect.* We begin by measuring the size of the endorsement effect, and then decompose it into the value that comes from direct authorship of a message as opposed to a more passive involvement, and examine the value of citing public health officials.

To do so, we begin by examining the behavior of F_2s, i.e., followers-of-followers. As described above and shown in Figure 1, this allows us to test the net effect of F_2s being blinded to the celebrities' involvement altogether.

We analyze this by estimating equation (2.1). Table 1 presents the results. We present (1) likes and retweets combined, and then each separately. Columns 1, 3, and 5 present the results on the full sample for each variable.

We see large endorsement effects. Having a celebrity compose and tweet the message relative to having a Joe/Jane compose the message and the celebrity retweet it leads to a 72 percent (0.54 log point) increase in the retweet or like rate (column 1, $p = 0.001$; note that since this is a Poisson model, the coefficients are interpretable as the change in log number of retweets/likes) by followers-of-followers (F_2s). The results are qualitatively similar for retweets and likes separately. We document similar effects for the 9 organizations in our sample being the originator rather than a Joe/Jane in Table D.1 of Appendix D.

These results imply that, holding the content of the tweet constant (since it is randomized across tweets) and holding the F_2 position in the network constant (since they are all followers-of-followers of the celebrity), having the F_2 be aware of the celebrity's involvement in passing on the message substantially increases the likelihood that the F_2 responds online.

The results allow us to begin to decompose the reason for retweeting. Specifically, the estimates imply that 63-72 percent[11] of retweets come from the fact that the celebrity is involved (in this case having written the tweet), with the remainder coming from the intrinsic interest in the content of the tweet itself.

There are two main potential challenges. First, whether a given F_2 agent sees a retweet from his or her F_1s may be endogenous and respond to our treatment, i.e., which F_1s choose to retweet the message may be affected by the fact that the celebrity composed the message. In equation (2.1), we control for the log number of followers of the F_1 who retweeted the message, and hence the number of F_2s who could potentially retweet it, so there is no mechanical reason for a bias in equation (2.1). But there may nevertheless be a *compositional* difference in which F_1s retweet it, which could potentially lead to selection bias of which F_2s are more likely to see the retweet.

[11] $\frac{\exp(\alpha)}{1+\exp(\alpha)}$ for coefficient α.

To address this issue, in the last phase of the experiment, we added an additional randomization. We use the subset of Joes/Janes who are also F_1s, and so direct followers of our celebrities. For some of these Joes/Janes, we randomly had their accounts retweet our celebrities' tweets and retweets in the experiment; that is, we created exogenous $F_1 s$. For this sample, we can look at how *their* followers – that is, the followers of F_1 Joes/Janes we exogenously forced to retweet a particular tweet – responded as we randomly vary whether the celebrity, an organization, or a Joe/Jane composes the message. We analyze this experiment by estimating equation (2.1) just as we did for the full sample, but here we know that whether an F_2 sees the tweet is exogenous by construction.

Columns 2, 4, and 6 present these results. The point estimates are if anything somewhat larger than in the full sample, and we cannot reject equality. Statistical significance is reduced somewhat (p-values of 0.119, 0.111, and 0.107 in columns 2, 4, and 6 respectively), but the fact that results are broadly similar to the overall effects in columns 1, 3, and 5 suggests that the possible endogenous selection of F_1s in our full sample is not leading to substantial bias.

The second potential confound comes from the fact that a retweet shows how many times the original tweet has been retweeted or liked when the user views it (see Figure E.1).Since our treatment assignment affects the retweet count, this itself could spur further changes in the likelihood of retweeting. The same randomization of forced Joe/Jane retweets also helps address this issue, because we randomly varied the number of Joes/Janes who retweet a particular tweet. Being randomly assigned one, five, ten, or even fifteen extra retweets makes no impact on the number of F_1 or F_2 retweets that the given tweet experiences (see Appendix C, Table C.1).

3.1.2. *Decomposing Endorsement: Authorship vs. Sending Messages.* In the preceding analysis, we isolated whether the F_2 knew the celebrity was involved in the message thread at all. But it groups together celebrity *authorship* with the celebrity *sending* the message, since the F_2 either saw both – a celebrity authored and sent message – or no celebrity involvement altogether.

To decompose this, we can examine the direct followers of the celebrities (i.e., F_1s). For F_1s, they know the celebrity is involved either way, but the randomization changes whether the message was directly authored by the celebrity or retweeted.

We estimate equation (2.2), and present results in Panel A of Table 2. We find that authorship matters: tweets authored by celebrities are 200 percent more likely to be retweeted/liked than those where the celebrity retweets (column 1, $p < 0.001$), with large effects on both likes and retweets.

This fact allows us to further decompose the impacts of celebrities. These estimates imply that 79 percent of the endorsement effect estimated earlier comes from authorship per se.

Combining these estimates with those in the previous section suggests that, on net, 56 percent of the celebrity effect comes from authorship, 14 percent from endorsement, with the remainder attributable to the intrinsic interest of the message.

3.1.3. *Citing Public Health Authorities.* Finally, we examine the impact of citing sources. Every tweet in our databank was paired with a source, but we randomized whether this source was included in each tweet. The sources come in several forms. In some cases, the tweet refers to the website or Twitter handle of a trusted authority who has issued that statement. For example, one tweet says "Polio vaccine should be given 4 times at months 1, 2, 3, 4. Are your baby's polio vaccines complete? @puskomdepkes" where @puskomdepkes is a link to the Twitter handle of the Ministry of Health (known as *DepKes* in Indonesian). In other cases, explicit sources are cited, with a Google shortened link provided.[12]

We re-estimate equation (2.2) at the F_1 level, adding a variable for whether the tweet was randomized to include a source.[13] Panel B of Table 2 presents the results.

On average, we find that citing a public health authority reduces the retweet and liking rate by 26.3 percent (-0.306 log points; $p = 0.051$, column 1), with similar effects for likes and retweets. This result shows that when a message is relayed by citing a public health authority, willingness to pass it on downstream, and perhaps liking the message, declines.

3.2. **Does Online Discussion Have Offline Effects?**

3.2.1. *Did people hear about the campaign?* We next examine whether an online celebrity endorsement campaign can have measurable offline effects. To investigate this, we estimate equation (2.3). We use the fact that our offline survey was conducted between Phases I and II, so that conditional on the number of our celebrities a user followed, exposure to our campaign as of the time of the phone survey was randomly assigned.

We begin with what can be thought of as akin to a first-stage in Panel A of Table 3. We ask whether respondents were more likely to have heard of our hashtag (#AyoImunisasi) (column 1) or heard about immunization discussions from Twitter if they were randomly more exposed to campaign tweets, conditional on their potential exposure (column 2).

We find that a one-standard deviation increase in exposure to the campaign (15 tweets) corresponds to a 16.75 percent increase in the probability that the respondent had heard of our hashtag relative to a mean of 7.7 percent (clustered $p = 0.044$, RI $p = 0.107$).[14]Further,

[12]Note that Twitter automatically produces a short preview of the content if the site linked to has Twitter cards set up. There is one non-Google shortened link used when citing IDAI (Ikatan Dokter Anak Indonesia, the Indonesian Pediatric Society).

[13]Note that the number of observations is smaller, because some tweets on topics deemed 'sensitive' by the Government always included a source, as noted above. We restrict analysis to tweets for which we randomized whether the source was included.

[14]Note that we report impacts on log-odds; we translate these into percent increases in the text.

a one-standard deviation increase in exposure corresponds to an 8.3 percent increase in the probability they heard about immunization in general from Twitter relative to a mean of 18.1 percent (clustered $p = 0.106$, RI $p = 0.046$).

3.2.2. *Did people then increase their knowledge about immunization facts?* We then ask whether exposure to the campaign led to increased knowledge about immunization, particularly about common vaccine 'myths', which were the focus of the campaign.[15]

Our survey asks questions about several categories of knowledge which were covered by the campaign. We ask whether people know that vaccines are domestically produced, to combat the common rumor in Indonesia that they contain pig products (which would make them unacceptable for Muslims, who represent the bulk of Indonesia's population; domestic products are known to be halal). We ask whether they believe that natural alternatives (breastfeeding, herbal supplements, alternative supplements) replace the need for immunization. And, we examine knowledge that typical symptoms (mild fevers or swelling) are normal and not a cause for alarm. We also ask about "access" information; in particular, we test whether they know that immunizations are free at government health centers.

Table 3, Panel B, presents the results for each of these categories of information. We find knowledge effects about domestic production (column 1)—though not on rumors about substitutability, side-effects, nor free access (columns 2-4). Seeing 15 campaign tweets in general corresponded to an increase of 5 percent in the probability of correctly answering the domestic question on a base of 57.6 percent (clustered $p = 0.042$, RI $p = 0.028$; Bonferroni p-values adjusting for the fact that we asked about 4 types of information are 0.168 and 0.112, respectively).

3.2.3. *Communication about immunization.* We next ask whether individuals were more likely to know about their neighbors, friends, and relatives' immunization behavior, which would be a byproduct of offline conversations. We ask about knowledge of immunizations since June 2015 to capture the period since the start of the campaign. Immunizations in Indonesia take place at monthly *posyandu* meetings, which occur each month in each neighborhood (usually hamlets, or *dusun*, in rural areas, and neighborhoods known as *rukun warga*, or *RW*, in urban areas; see Olken, Onishi, and Wong (2014)), so if knowledge would increase, one might expect it to be the knowledge about immunization practices of ones' neighbors.[16]

Panel A of Table 4 presents the results. Column 1 shows that being exposed to 15 more tweets corresponds to a 5.2 percent increase in the probability of knowing the neighbor's status (clustered $p = 0.004$, RI $p = 0.088$). We do not consistently see significant effects on

[15]Myth-dispelling facts comprised 36.7 percent of tweets, and 82.4 percent of all fact-related tweets sent out. Table B.1 in Appendix B shows that tweets concerning myths also diffused more widely than other facts.

[16]Note that here the sample is restricted to respondents who know friends, relatives, or neighbors with at least one child (ages 0-5) respectively as this is the relevant set.

non-neighbor friends. With relatives the point estimates are comparably large, though the estimates are noisier.

3.2.4. *Did exposure lead to changes in reported immunizations?* Finally, we ask whether individuals have more knowledge of actual immunizations among friends, neighbors, and relatives. That is, does the campaign appear to change immunization behavior as reported by our survey respondents?

Panel B of Table 4 presents the results, where the dependent variable is whether the respondent knows of anyone among their friends, neighbors, or family who immunized a child since June 2015. We condition the sample when we look at knowledge of immunization among neighbors, friends, and family to those who knew whether the vaccination status of the children of members of these individuals, so this effect is above and beyond the effects reported above.[17]

Column 1 shows that when looking at neighbors, an increased exposure by 15 tweets corresponds to a 12.5 percent increase in the number of reported vaccinations (clustered $p = 0.071$, RI $p = 0.132$) relative to a mean of 0.356. For friends, shown in column 2, we find an increased exposure of 15 tweets corresponds to a 16.0 percent increase in the number of reproted vaccinations (clustered $p = 0.001$, RI $p = 0.071$) relative to a mean of 0.353. Column 3 presents results looking at relatives. An increased exposure by 15 tweets corresponds to a 9.6 percent increase in the number of reported vaccinations among relatives (clustered $p = 0.159$, RI $p = 0.06$) relative to a mean of 0.314. Finally, column 4 looks at own behavior, and our estimate is not statistically different from zero. On net, the results in this section are suggestive that an online Twitter campaign can have offline effects on knowledge and, potentially, behavior.

4. Conclusion

We study how to design a public health campaign to deploy celebrities effectively by conducting a large-scale, online field experiment. Using our design, we are able to decompose to what extent celebrity authorship *per se* as opposed to passing on messages of others matters, and to show whether citing public health officials amplifies or reduces these effects.

We find the celebrity endorsement matters. Moreover, the vast majority (79%) of the 70% increase in retweet rate due to celebrity endorsement comes from authorship *per se*, rather than merely passing on messages.

[17]An important caveat is that knowledge of neighbors', friends', and family members' children's vaccination status itself is affected by treatment (as discussed above), so these results should be interpreted with some caution. Interestingly, the fact that we see impacts on reported vaccinations of friends, but knowledge of friends' behavior is unchanged, suggests that perhaps this is less of an issue. More generally, however, this could reflect some combination of changes in immunization and changes in discussions about it.

We also find that explicitly referring to public health sources has an adverse effect. This result may seem surprising, since one might expect that a sourced message may be more reliable. There are, however, several possible explanations for this finding. One possibility is that for an F_1, passing on a message has both instrumental value (delivering a good message), as well as a signaling value (conveying to followers that the F_1 is able to discern which information is good). We illustrate this in Appendix F, building on prior work by the authors (Chandrasekhar et al., 2018), though other mechanisms are certainly possible as well. At a broader level, these results are consistent with the results on celebrity authorship: messages are most likely to be passed on when they come from the celebrity speaking directly, rather than passing on messages from others.

We also find offline effects: messages are heard, certain myths (about imported vaccines, in particular) become less prevalent, offline communication about vaccination increases, and we see more reported immunization behavior among respondents' neighbors, friends, and relatives.

These findings – particularly the idea that celebrity messaging works best when celebrities speak directly, themselves, rather than merely pass on recommendations from others – are applicable in a wide variety of settings, including perhaps in designing information campaigns to encourage vaccination during the COVID-19 crisis pandemic.

References

Agrawal, J. and W. A. Kamakura (1995): "The economic worth of celebrity endorsers: An event study analysis," *Journal of Marketing*, 59, 56–62. 1

Bakshy, E., J. M. Hofman, W. A. Mason, and D. J. Watts (2011): "Everyone's an influencer: quantifying influence on Twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 65–74. 1

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The diffusion of microfinance," *Science*, 341, 1236498. 1

Banerjee, A. V., E. Breza, A. G. Chandrasekhar, and B. Golub (2018): "When Less is More: Experimental Evidence on Information Delivery During India's Demonetization," Working Paper 24679, National Bureau of Economic Research. F.1

Beaman, L., A. BenYishay, J. Magruder, and A. M. Mobarak (2021): "Can network theory-based targeting increase technology adoption?" *American Economic Review*, 111, 1918–43. 1

Beaman, L. and A. Dillon (2018): "Diffusion of agricultural information within social networks: Evidence on gender inequalities from Mali," *Journal of Development Economics*, 133, 147–161. 1

Belloni, A., V. Chernozhukov, and C. Hansen (2014a): "High-dimensional methods and inference on structural and treatment effects," *The Journal of Economic Perspectives*, 29–50. 2.4

——— (2014b): "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81, 608–650. 2.4

Bergkvist, L. and K. Q. Zhou (2016): "Celebrity endorsements: a literature review and research agenda," *International Journal of Advertising*, 35, 642–663. 1

Bursztyn, L., G. Egorov, and R. Jensen (2017): "Cool to be smart or smart to be cool? Understanding peer pressure in education," *The Review of Economic Studies*. F.1

Bursztyn, L. and R. Jensen (2015): "How does peer pressure affect educational investments?" *Quarterly Journal of Economics*, 130, 1329–1367. F.1

Chandrasekhar, A. G., B. Golub, and H. Yang (2018): "Signaling, Shame, and Silence in Social Learning," Tech. rep., National Bureau of Economic Research. 7, 4, F.1, F.3

Elberse, A. and J. Verleun (2012): "The economic value of celebrity endorsements," *Journal of Advertising Research*, 52, 149–165. 1

Garthwaite, C. L. (2014): "Demand spillovers, combative advertising, and celebrity endorsements," *American Economic Journal: Applied Economics*, 6, 76–104. 1

Gong, S., J. Zhang, P. Zhao, and X. Jiang (2017): "Tweeting as a marketing tool: A field experiment in the TV industry," *Journal of Marketing Research*, 54, 833–850. 1

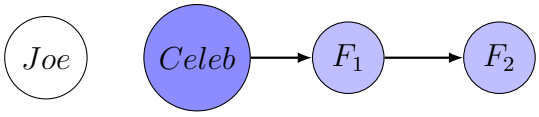
Ives, M. (2021): "Celebrities Are Endorsing Covid Vaccines. Does It Help?" *The New York Times*. 1

Katona, Z., P. P. Zubcsek, and M. Sarvary (2011): "Network Effects and Personal Influences: The Diffusion of an Online Social Network," *Journal of Marketing Research*, 48:3, 425–443. 1

Leskovec, J., L. A. Adamic, and B. A. Huberman (2007): "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, 1, 5. 1

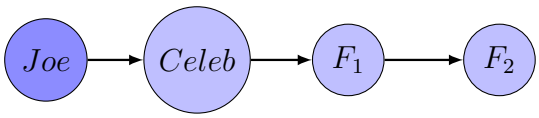
Olken, B. A., J. Onishi, and S. Wong (2014): "Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia," *American Economic Journal: Applied Economics*, 6, 1–34. 3.2.3

Figures



(A) Message M originated by Celeb

(B) F_2's observation



(C) Message M originated by Joe

(D) F_2's observation

Figure 1. Identification of the value of endorsement of celebrity involvement.

Tables

Table 1. Estimating the average value of celebrity involvement using followers-of-followers' (F_2) behavior

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Pooled	(3) Poisson # Retweets	(4) Poisson # Retweets	(5) Poisson # Likes	(6) Poisson # Likes
Celeb writes and tweets	0.544	0.788	0.518	0.931	0.664	1.109
	(0.166)	(0.505)	(0.166)	(0.584)	(0.482)	(0.687)
	[0.00105]	[0.119]	[0.00175]	[0.111]	[0.168]	[0.107]
Observations	1,997	911	1,997	911	1,997	911
Joe/Jane writes mean	0.0417	0.00915	0.0417	0.00686	0.00745	0.00229
Forced Joes/Janes only		✓		✓		✓

Notes: Standard errors (clustered at the original tweet level) are reported in parentheses. p-values are reported in brackets. Sample conditions on all tweets originated by Joes/Janes or celebrities. All regressions control for phase, celebrity fixed effects, content fixed effects, and the log number of followers of the F_1.

TABLE 2. Estimating the value of authorship and citing public health officials using followers' (F_1) behavior

Panel A: Decomposing what % of the involvement effect comes from authorship

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes
Celeb writes and tweets	1.101	1.329	0.803
	(0.0840)	(0.0910)	(0.105)
	[0]	[0]	[0]
Observations	451	451	451
Joe/Jane writes and Celeb retweets mean	2.058	1.045	1.013

Panel B: Measuring the effect of citing public health sources

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes
Source cited	-0.306	-0.318	-0.277
	(0.157)	(0.161)	(0.183)
	[0.0513]	[0.0478]	[0.130]
Observations	492	492	492
Depvar Mean	3.644	3.644	3.644

Notes: In Panel A, sample conditions on all tweets originated by Joes/Janes or celebrities. In Panel B, the sample conditions on non-sensitive tweets. All regressions control for phase, celebrity fixed effects, content fixed effects. Standard errors (clustered at the celebrity/organization level) are reported in parentheses.

TABLE 3. Knowledge of Campaign and Facts

Panel A: Did people offline hear about the campaign?

VARIABLES	(1) Logit Heard of *#Ayoimunisasi*	(2) Logit Heard of immunization from Twitter
Std. Exposure to tweets	0.197	0.108
	(0.0980)	(0.0666)
	[0.0443]	[0.106]
	{.107}	{.046}
Observations	2,164	2,404
Potential exposure control	✓	✓
Double Post-LASSO	✓	✓
Depvar Mean	0.0776	0.181

Panel B: Did people offline increase knowledge?

VARIABLES	(1) Logit Domestic	(2) Logit Substitutes	(3) Logit Side-effects	(4) Logit Free
Std. Exposure to tweets	0.120	-0.0391	0.0305	0.0549
	(0.0591)	(0.0589)	(0.0624)	(0.0687)
	[0.0424]	[0.506]	[0.625]	[0.424]
	{.028}	{.891}	{.751}	{.629}
Observations	2,434	2,440	2,440	2,440
Potential exposure control	✓	✓	✓	✓
Double Post-LASSO	✓	✓	✓	✓
Depvar Mean	0.576	0.527	0.486	0.680

Notes: Standard errors (clustered at the combination of celebs followed level) are reported in parentheses. Clustered p-values are reported in brackets. Randomization inference (RI) p-values are reported in braces. Demographic controls include age, sex, province, dummy for urban area and dummy for having children. One standard deviation of exposure is 14.96 tweets.

TABLE 4. Communication With and Behavior of Neighbors, Friends, and Relatives

Panel A: Communication: Knowledge of immunization behavior of others

VARIABLES	(1) Logit Neighbor	(2) Logit Friend	(3) Logit Relative
Std. Exposure to tweets	0.231	0.0156	0.214
	(0.0814)	(0.0826)	(0.132)
	[0.00449]	[0.850]	[0.105]
	{.088}	{.778}	{.462}
Observations	1,642	1,626	1,564
Potential exposure control	✓	✓	✓
Double Post-LASSO	✓	✓	✓
Depvar Mean	0.775	0.813	0.923

Panel B: Immunization behavior of others and self

VARIABLES	(1) Logit Neighbor	(2) Logit Friend	(3) Logit Relative	(4) Logit Own
Std. Exposure to tweets	0.194	0.246	0.140	-0.0840
	(0.107)	(0.0955)	(0.0997)	(0.0886)
	[0.0707]	[0.00994]	[0.159]	[0.343]
	{.132}	{.071}	{.06}	{.66}
Observations	682	682	682	634
Potential exposure control	✓	✓	✓	✓
Double post-LASSO	✓	✓	✓	✓
Depvar Mean	0.356	0.353	0.314	0.486

Notes: In both panels, standard errors (clustered at the combination of celebs followed level) are reported in parentheses. Clustered p-values are reported in brackets. Randomization inference (RI) p-values are reported in braces. Demographic controls include age, sex, province, dummy for urban area and dummy for having children. One standard deviation of exposure is 14.96 tweets. In Panel A, the sample is restricted to respondents who know friends/relatives/neighbors with at least one child (ages 0-5) respectively. In Panel B, the sample when looking at network members' behaviors (columns 1-3) is restricted to respondents who know the behavior of their network. When looking at own behavior in column 4, sample restricted to respondents with children younger than age 2.

Appendix A. Summary Statistics and Balance

Table A.1. User summary stats

Panel A: Online user summary

	mean	obs
Followers of celebrities	262648	37
Followers of organizations	145300	9
Followers of Joes/Janes	574	134
Followers of forced Joes/Janes	502	898
Followers of celeb followers	1379	1073

Panel B: Offline user summary

	Sample mean	National Average (SUSENAS) mean
Age	28.416	29.959
Female	0.504	0.499
City (*kota*)	0.610	0.200

Panel C: Baseline beliefs

	mean	obs
Immunization is important	0.988	886
Immunization is safe	0.944	886
Immunization is beneficial	0.983	886
Breastfeeding can't replace immunization	0.650	886
Supplements can't replace immunization	0.872	886
Herbal supplements can't replace immunization	0.832	886
BCG is a basic vaccine	0.452	622
Hepatitis B is a basic vaccine	0.291	622
DPT is a basic vaccine	0.399	622
HIB is a basic vaccine	0.084	622
Polio is a basic vaccine	0.712	622
Measles is a basic vaccine	0.611	622
Immunization does not cause swelling	0.561	886
Immunization does not cause fever	0.647	886
Vaccines are domestically made	0.561	886
Vaccines are free of cost	0.385	886

TABLE A.2. Balance Check

VARIABLES	(1) OLS Facts	(2) OLS Importance Info	(3) OLS Access Info	(4) OLS Myth-busting Facts	(5) OLS Other Facts	(6) OLS Source cited
Celeb writes and tweets	0.0302	-0.0580	0.0278	0.0139	0.0163	-0.0184
	(0.0449)	(0.0430)	(0.0318)	(0.0451)	(0.0255)	(0.0469)
	[0.501]	[0.178]	[0.383]	[0.758]	[0.524]	[0.695]
Observations	451	451	451	451	451	451
Phase control	✓	✓	✓	✓	✓	✓
Log #followers control	✓	✓	✓	✓	✓	✓
Message style control	✓	✓	✓	✓	✓	✓

Notes: Standard errors (clustered at the original tweet level) are reported in parentheses. p-values are reported in brackets. Sample conditions on all tweets originated by Joes/Janes or celebrities. All regressions control for phase, formality, and exception status.

Appendix B. Does content affect retweeting?

Table B.1. How Content Affects Retweeting by F_1 likes/retweets

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes	(4) Poisson # Pooled	(5) Poisson # Retweets
Myth-busting Facts	0.588	0.627	0.518	-0.0481	0.136
	(0.319)	(0.346)	(0.381)	(0.296)	(0.413)
	[0.0654]	[0.0698]	[0.174]	[0.871]	[0.742]
Access Info	0.402	0.319	0.530	0.315	0.477
	(0.258)	(0.292)	(0.309)	(0.294)	(0.366)
	[0.118]	[0.275]	[0.0863]	[0.284]	[0.192]
Importance Info	0.543	0.526	0.565	0.466	0.442
	(0.229)	(0.267)	(0.290)	(0.246)	(0.343)
	[0.0178]	[0.0487]	[0.0516]	[0.0578]	[0.197]
Celeb writes and tweets				1.040	1.327
				(0.283)	(0.374)
				[0.000242]	[0.000382]
Myth × Celeb Direct				0.652	0.432
				(0.381)	(0.485)
				[0.0871]	[0.374]
Access × Celeb Direct				-0.0101	-0.299
				(0.374)	(0.451)
				[0.979]	[0.508]
Importance × Celeb Direct				0.0558	0.00945
				(0.314)	(0.406)
				[0.859]	[0.981]
Myth × Celeb RT Org				0.250	0.147
				(0.290)	(0.384)
				[0.389]	[0.701]
Access × Celeb RT Org				0.103	-0.0418
				(0.244)	(0.256)
				[0.672]	[0.870]
Importance × Celeb RT Org				0.135	0.402
				(0.215)	(0.190)
				[0.531]	[0.0348]
Observations	492	492	492	492	492
Depvar Mean	3.644	3.644	3.644	3.644	3.644

Notes: Standard errors (clustered at the celebrity/organization level) are reported in parentheses. p-values are reported in brackets. All columns include fixed effects for number of non-exception tweets assigned and condition on non-exception tweets. The omitted category is non-myth facts.

Appendix C. Does RT count affect retweeting?

Table C.1. Impact of No. of Forced Joe RTs on F_2 and F_1 likes/retweets

VARIABLES	(1) F2 Poisson # Retweets	(2) F1 Poisson # Retweets
5 Forced Joe RTs assigned	0.0399	0.444
	(0.346)	(0.388)
	[0.908]	[0.252]
10 Forced Joe RTs assigned	0.244	0.0395
	(0.414)	(0.440)
	[0.556]	[0.928]
15 Forced Joe RTs assigned	0.256	0.207
	(0.407)	(0.359)
	[0.529]	[0.565]
Observations	505	184
Phase Control	✓	✓
Log #followers control	✓	✓
Message style control	✓	✓
Depvar Mean	0.184	2.707
1 Forced Joe RT assigned log mean	-2.331	0.870

Notes: Robust standard errors are reported in parentheses. p-values are reported in brackets.

Appendix D. Effect of celebrity retweeting organizations

Table D.1. Identifying the Role of Celebrity and Organization Endorsement

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes
Celeb writes and tweets	0.423	0.421	0.574
	(0.182)	(0.179)	(0.520)
	[0.0201]	[0.0185]	[0.269]
Org writes and Celeb retweets	0.564	0.600	0.255
	(0.221)	(0.258)	(0.520)
	[0.0107]	[0.0200]	[0.624]
Observations	1,791	1,791	1,791
Joe writes mean	0.0417	0.0343	0.00745

Notes: Standard errors (clustered at the original tweet level) are reported in parentheses. p-values are reported in brackets. The sample conditions on tweets that are not sensitive and includes tweets originated by Joes, organizations, and celebrities. All regressions control for phase, celebrity fixed effects, and content fixed effects.

Appendix E. Sample Tweet

(A) Celebrity Tweet: casual with credibility boost

(B) Celebrity retweeting an Organization: casual with credibility boost

(C) Celebrity retweeting a Joe/Jane: formal without credibility boost

Figure E.1. Sample tweets and retweets from the campaign

Appendix F. Model

F.1. **Overview.** We study the decision by individuals on Twitter to pass on information to their followers by "retweeting" it. Before proceding to our empirical analysis, we begin by discussing a simple framework to think through how individuals make the decision to pass on information. The framework is standard, developed in Chandrasekhar, Golub, and Yang (2018) and also previously applied in Banerjee, Breza, Chandrasekhar, and Golub (2018).

In our framework, individuals pass on information for two reasons. First, individuals may care that others are informed about a topic. Second, as retweeting is intrinsically a social activity, individuals can be motivated by how they are viewed by their followers. In this case, individuals may choose to retweet certain topics as a function of how the act of sharing the information changes how they are perceived by others. For example, individuals on Twitter may be trying to gather more followers, and it is plausible that people are more likely to keep following someone whom they believe is sharing high-quality information.

This second observation – that people may share information with a view to how it affects how others perceive them – turns out to have subtle ramifications for how we think about a dissemination strategy. Whether information is more likely to spread more widely if originated by a celebrity or an ordinary Joe/Jane, or whether messages cite credible sources or simply consist of assertions, turn out to be ambiguous questions once we include the fact that these features of messages change the degree to which sharing the message provides information in equilibrium about the likely quality of the person deciding whether to share it.

In particular, the standard intuition is that more and credible information is simply better, and hence more likely to be retweeted. This comes from a standard model in which individuals only base their decisions to pass on information based on the first factor, namely the quality of that information. In this case, if a message has more credibility and has a verified source, then more retweeting should happen. This generates an intuition that, for instance, sourced tweets or celebrity tweets should be retweeted more.

However, when we consider the fact that retweeting has a social component—that individuals certainly care about how they are perceived and that is likely a key component of their motivation to retweet—we see that these conclusions change. Assume that an individual F follows an originator of a tweet, o. Suppose that F is more willing to pass on information if he is more certain about the state of the world. Also assume that F can be one of two private types: a high type (greater ability or social consciousness for the sake of discussion) and a low type. Individuals desire to be perceived of as a high type by their followers, so part of the motivation to retweet is for this social perception payoff. It is commonly known that high types are better able to assess the state of the world rather than low types (i.e.,

imagine that in addition to the tweet, individual F gets a private signal as to the state of the world, and the high types' signal is more informative). When F sees a tweet by o, he needs to glean the state of the world using both the tweet and his own private signal, and decide whether or not to retweet.

To illustrate ideas, let us compare the case where o's tweet contains no source versus cites a credible source about the topic. Inclusion of a source has multiple effects. First, the source citation should make the state of the world even more evident. This should encourage retweeting through increasing certainty. Second, and more subtly, if social perception is important enough, source citation can have a discouraging effect on retweeting. Specifically, if a source makes it very clear what is true, then there is no room for signaling remaining: high types are no better able to assess things than low types and therefore ability does not really matter. We show below that which effect dominates on net—the increased direct effect of the source on quality, or the fact that the source decreases the ability of F to use the tweet to signal quality—turns out to be ambiguous.

To show this more formally, we adapt the endogenous communication model developed by Chandrasekhar, Golub, and Yang (2018) to our context of retweeting on Twitter (see also Banerjee et al. (2018) for another such prior application of this model). Such image concerns have also been looked at both theoretically and empirically in both Bursztyn and Jensen (2015); Bursztyn, Egorov, and Jensen (2017) who study whether peer perceptions inhibit the seeking of education. We look at individuals who have payoffs from passing on information and who are concerned with social perception as well the direct value of the information they pass. We show how sourcing, originator identity, exposure, and content all can have ambiguous effects on the amount of retweeting, and explore when we might expect which policies to work well.

It is important to note that we are not claiming of course that these are the only motives for retweeting. After all, there can be more mundane motivations: it is just more fun to retweet anything by a celebrity, it is just frivolous to retweet anything by a celebrity, one likes to retweet something that he/she anticipates will not be otherwise widely spread, among other explanations. But without hardcoding anything else into the model, in the simplest interpretation of dynamics on Twitter, we can demonstrate and motivate why the questions we study are ultimately empirical issues.

F.2. **Setup.**

F.2.1. *Environment.* The state of the world is given by $\eta \in \{0, 1\}$, with each state equally likely. There is an originator o (she) who writes an initial message about the idea with probability $q \in (0, 1]$, which is received by her follower F (he). With probability $1 - q$ nothing happens. The message is a binary signal about the state of the world, which is

accurate with probability α, i.e.

$$m = \begin{cases} \eta & \text{w.p. } \alpha \geq \frac{1}{2} \\ 1-\eta & \text{o.w.} \end{cases}.$$

The message may or may not cite a source, designated by $z \in \{S, NS\}$ respectively. We allow the quality of the signal to depend on source, so $\alpha = \alpha_z$, discussed below.

Further, there are two types of originators: ordinary Janes/Joes and celebrities, given by $o \in \{J, C\}$ respectively. We allow the quality of the signal to depend on originator, so $\alpha = \alpha_o$, discussed below.

Finally, followers come in two varieties: $\theta \in \{H, L\}$ represents F's privately known type, and one's type is drawn with equal odds. High types have better private information about the state of the world. This can represent ability in a loose way such as intelligence, social accumen, taste-making ability, or any trait which allows F to better discern the state of the world if he is of type H rather than L. We model this by supposing that F draws an auxiliary signal, x, with $x = \eta$ with probability π_θ and $x = 1-\eta$ with probability $1-\pi_\theta$. We assume $\pi_H \geq \pi_L$ which reflects that H-types can better discern whether the idea is valuable. As discussed below, it is socially desirable to be perceived as $\theta = H$.

This environment captures our basic experimental setting. We randomly vary originator $o \in \{J, C\}$ and whether the message is sourced, $z \in \{S, NS\}$.

F.2.2. *Bayesian Updating.* F is assumed to be Bayesian. Let $\alpha = \alpha_{o,z}$ be the quality of the signal depending on originator and source. Therefore given message m and private signal x, we can compute the likelihood ratio that F believes the state of the world being good versus bad as

$$\begin{aligned} LR\left(\eta | m, x; o, z, \theta\right) &= \frac{\mathrm{P}\left(\eta = 1 | m, x\right)}{\mathrm{P}\left(\eta = 0 | m, x\right)} = \frac{\mathrm{P}\left(m, x | \eta = 1\right)}{\mathrm{P}\left(m, x | \eta = 0\right)} \\ &= \left(\frac{\alpha_{o,z}}{1-\alpha_{o,z}}\right)^m \left(\frac{1-\alpha_{o,z}}{\alpha_{o,z}}\right)^{1-m} \left(\frac{\pi_\theta}{1-\pi_\theta}\right)^x \left(\frac{1-\pi_\theta}{\pi_\theta}\right)^{1-x}. \end{aligned}$$

Note that as α or π tend to 1 or $\frac{1}{2}$, the likelihood ratio tends to $+\infty$ (the signal reveals the state) or 1 (the signal has no content), respectively.

F.2.3. *Payoffs.* The utility of F depends on two components. The first is the instrumental payoff: it is a payoff from retweeting when the state of the world is more clear: that is when $LR\left(\eta\right)$ is more extreme. Thus we assume that the instrumental payoff when you do not retweet, i.e., when $r = 0$, is 0 and when you do retweet, i.e., $r = 1$, is $\varphi\left(LR\left(\eta | m, x; o, z, \theta\right)\right)$ for some smooth increasing in distance function from 1, $\varphi\left(\cdot\right)$. What this captures is that there is more instrumental value in passing on a message the greater certainty in the state

of the world. For instance if we set

$$\varphi(x) = f\left(\left|\frac{x}{1+x} - 1\right|\right)$$

for a smooth increasing function $f(\cdot)$ on $\left[0, \frac{1}{2}\right]$, the instrumental value is a monotone function in the probability the state of the world is high, but other functions φ will also work.[18] Further, due to taste or cost heterogeneity, there is a shock ϵ to the instrumental payoff of retweeting, where ϵ is a mean-zero random variable drawn from a continuous CDF with full support, such as the logit CDF $\Lambda(\cdot)$. Altogether, the instrumental payoff V^r is given by

$$V^r = \begin{cases} \varphi(LR(\eta|m,x;o,z,\theta)) - \epsilon & \text{if } r = 1 \\ 0 & \text{if } r = 0. \end{cases}$$

The second is the social perception payoff. Specifically F is concerned with the posterior that his followers have about his type given his decision to retweet: $\psi(\mathrm{P}(\theta = H|r))$ where $\psi(\cdot)$ is a monotonically increasing function. The perception in equilibrium is simply a function of the retweet decision itself. The idea here is that someone who is more able is more likely to be able to discern valuable topics and therefore the equilibrium decision to retweet itself has a signaling component. [19]

F's total utility is given by

$$U(r|m,x) = \underbrace{V^r}_{\text{instrumental}} + \underbrace{\lambda\psi(\mathrm{P}(\theta = H|r))}_{\text{perception}}$$

where $\lambda \geq 0$ is a parameter that tunes the strength of the perception payoff. [20]

Correspondingly, the marginal utility of choosing $r = 1$ versus $r = 0$ is given by

$$MU(r|m,x) = \underbrace{\varphi(LR(\eta|m,x;o,z,\theta)) - \epsilon}_{\text{change in instrumental}} + \underbrace{\lambda\Delta_r\psi(\mathrm{P}(\theta = H|r))}_{\text{change in perception}}.$$

Let $Q_H(\cdot)$ be the CDF of $\varphi(LR(\eta|m,x;o,z,H)) - \epsilon$ and $Q_L(\cdot)$ be the CDF of $\varphi(LR(\eta|m,x;o,z,L)) - \epsilon$.[21] It immediately follows that $Q_H \succ_{\text{FOSD}} Q_L$. This can be seen by inspection, where the likelihood ratio under type H first order stochastically dominates that of type L when $\eta = 1$ and the inverse of the ratio first order stochastically dominates when $\eta = 0$. It will be useful

[18] To see this, note that

$$\varphi(LR(\eta|m,x;o,z,\theta)) = f\left(\left|\frac{LR}{1+LR} - 1\right|\right) = f\left(\left|\mathrm{P}(\eta = 1|m,x;o,z,\theta) - \frac{1}{2}\right|\right)$$

which is just a smooth function of distance from pure uncertainty of a belief of $\frac{1}{2}$.

[19] For simplicity we abstract from F's followers interpretation of m and their own subsequent private signals. The reason is that we can demonstrate interesting non-monotonicities in retweeting behavior as a function of message quality without such additions, which would only serve to complicate matters.

[20] While λ could be absorbed into $\psi(\cdot)$, it is useful for exposition to keep it separate.

[21] This holds fixed o and z.

below to denote by G_θ the complementary CDF, $G_\theta := 1 - Q_\theta$, i.e., $G_\theta(v)$ is the fraction of types θ with a (net-of-costs) instrumental value of passing greater than or equal to v.

F.3. **Analysis.** F decides to retweet if and only if $MU(r|m,x) \geq 0$. This decision trades off two components. On the one hand is the relative instrumental benefit (or cost) of passing on the message, which is an increasing function of the likelihood that the state of the world $\eta = 1$, and is given by $\varphi(LR(m,x|o,z,\theta))$. On the other hand, retweeting itself changes the perception of F by his followers, given by$\Delta_r \psi(\mathrm{P}(\theta = H|r))$, and so the (equilibrium) relative gain/loss of reputation must be taken into account.

The model is formally characterized in Proposition 1 of Chandrasekhar et al. (2018), and we refer the interested reader to that paper for proofs. Chandrasekhar et al. (2018) show that under the above assumptions, an equilibrium exists, and will be in cutoff strategies where F chooses to retweet if and only if $\varphi(LR(\eta|m,x;o,z,\theta)) - \epsilon \geq v$ for some v. An equilibrium is characterized by a cutoff $\underline{v} < 0$, which is used by all Fs irrespective of type θ, where it is the solution to

$$\underline{v} = \lambda\psi(\mathrm{P}(\theta = H|r=0)) - \lambda\psi(\mathrm{P}(\theta = H|r=1)).$$

Here the equilibrium posteriors are determined by:

$$\frac{\mathrm{P}(\theta = H|r=0)}{1-\mathrm{P}(\theta = H|r=0)} = \frac{1-qG_H(v)}{1-qG_L(v)} \text{ and } \frac{\mathrm{P}(\theta = H|r=1)}{1-\mathrm{P}(\theta = H|r=1)} = \frac{G_H(v)}{G_L(v)}.$$

The intuition for the equilibrium is as follows. First, note that F's type does not matter for the decision he makes conditional on the draw v. That is, while θ affects the distribution of the instrumental value, once F knows his instrumental value, he is trading off that against the change in reputation due to his behavior. Therefore the cutoff (in utility space) will not depend on θ's type.

At the cutoff $\underline{v}$ in equilibrium the marginal benefit of retweeting (which is a way to gain reputation by being viewed as more likely to be a high type) must be equal to the marginal cost of retweeting (which in this case is the instrumental benefit of passing the information relative to the stochastic cost). The reason $\underline{v} < 0$ is because here retweeting is a signal of being the high type, and therefore some low types will opt into retweeting despite having a negative net instrumental cost.

Holding fixed o, z as we have been doing above, we can compute the retweeting share in equilibrium:

$$\frac{1}{2}G_H(\underline{v}) + \frac{1}{2}G_L(\underline{v}).$$

We can also look at several contrasting situations. In the first, assume that $\lambda = 0$ with the same setup as above, so there is no interest in social concerns. Then only positive

instrumental values are retweeted, so the share retweeting is given by

$$\frac{1}{2}G_H(0) + \frac{1}{2}G_L(0).$$

Clearly the retweeting share is lower than when there is also a signaling motive, which featured an equilibrium cutoff $\underline{v} < 0$.

A second contrasting situation is one in which, while individuals would potentially care about signaling, neither party is better at discerning the state of the world. That is, $\widehat{G}_H = \widehat{G}_L =: \widehat{G}$. In this case the share retweeted again is only determined by positive instrumental values and therefore is given by

$$\widehat{G}(0).$$

Whether $\widehat{G}(0) \lesseqgtr \frac{1}{2}G_H(\underline{v}) + \frac{1}{2}G_L(\underline{v})$ depends on how $\widehat{G}$ compares to G_H and G_L.

A subtle feature of the model is the fact that the retweet share is not necessarily monotonically increasing in the quality of the message, α. Intuitively, there are two effects of increasing α of retweeting. First, as α increases, the message becomes more informative. This increases the instrumental value of retweeting, and hence retweeting increases with α. Second, as α increases, the m signal becomes more informative relative to the private x signal. This makes the act of retweeting more about m than x, and hence lowers the signaling value of retweeting. Indeed, in the limit where $\alpha = 1$, there is no signaling value whatsoever. Thus, the signaling effect leads to a reduction in the amount of retweeting as α increases. Which effect dominates depends on parameters, and as we show now, in fact the effect of α on retweeting can be non-monotonic under some configurations of parameters.

Figure B.1 presents simulation results to further illustrate these intuitions. First consider the case when there is no reputation considerations ($\lambda = 0$). In this case, as the message's quality increases, the share retweeting must increase clearly because the value of information on average increases.

Next let us consider the case where neither H nor L are particularly able types, with $\pi_H = 0.53$ and $\pi_L = 0.5$. In this case, there is limited scope for signaling because the priors are quite poor: both types heavily lean on the message's signal m rather than their personal signals x. As such, like in the case with $\lambda = 0$, the quality of the message increases the share to retweet.

In contrast, consider the case where both types are expert, but H-types are somewhat better ($\pi_H = 0.95$, $\pi_L = 0.9$). In this case, with low α, since the predominant component of instrumental value comes from type to begin with, and because high types are much more likely to receive correct signals than low types but both have typically good signals about the state of the world (so m and x will agree), many more L types will also find it worthwhile to essentially "pool" with H types despite negative instrumental values due to reputation concerns. This leads to a monotonic decline in the retweet rate as α increases, since there

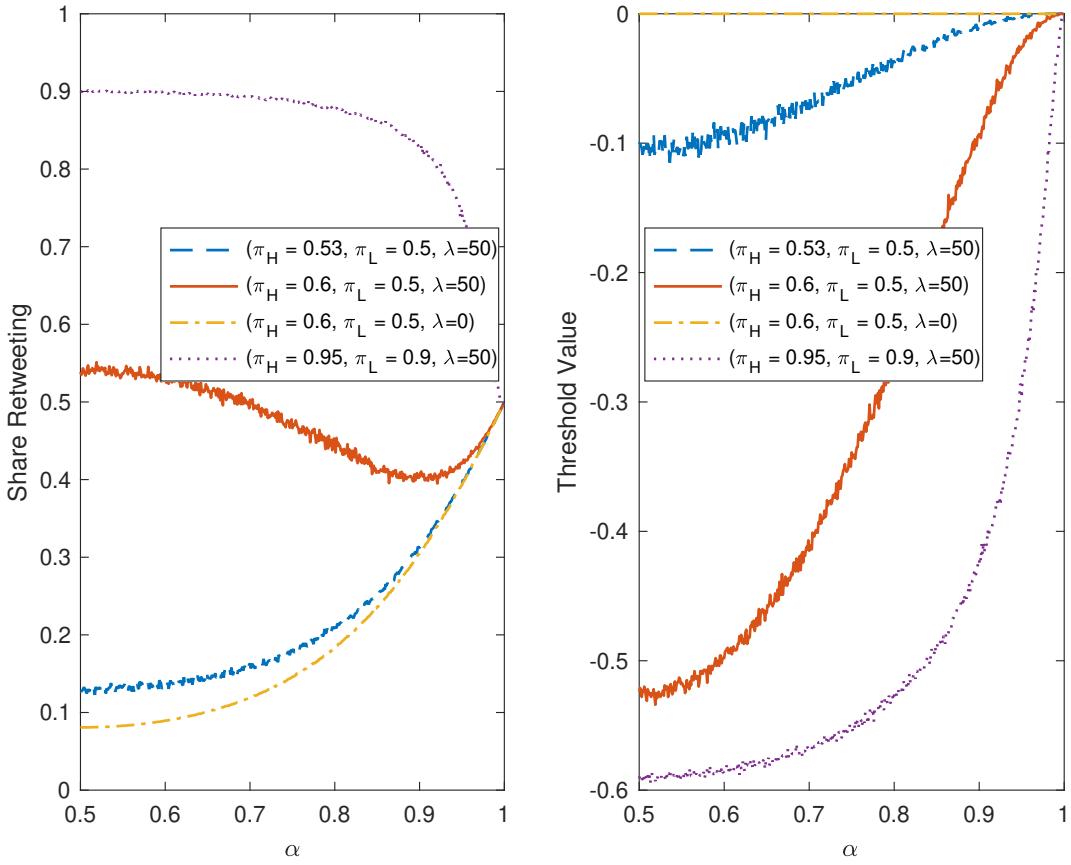


FIGURE B.1. Retweet share for various combinations of (π_H, π_L, λ).

is increasing reliance on the m-signal. What this means in practice is that it is possible to improve the quality of the message and yet reduce the overall share of retweeting, contrary to the naive intuition without a social perception payoff component.

The final case we show is the intermediate one, with $\pi_H = 0.65$ and $\pi_L = 0.5$. The signaling effect at this parameter level dominates initially, and hence increasing α initially decreases retweeting, but then eventually is dwarfed by the instrumental effect as the m-signal is considerably better than the gap in quality for the x-signal across types.

The fact that the relationship between α and retweeting is non-monotonic means that it is possible that mild increases in informativeness can reduce retweeting whereas dramatic increases in informativeness can increase it.

F.4. **Application to Experiment.** In what follows, we use the above framework to consider possible implications of our experimental variations, i.e., (1) whether the originator is a celebrity or a Jane/Joe and (2) whether the tweet has a source or not.

F.4.1. *Celebrity versus Jane/Joe.* Celebrities and Joes/Janes can vary in the quality of their messaging. As such, we consider α_C versus α_J. Ex ante it may be possible for these to have any relationship, though we might think that celebrities tend to generate higher-quality signals. This could be, for instance, because celebrities' messages reach many more individuals and therefore they need to be more cautious in their messaging, or it could be because they have better access to information in general.

Assuming $\alpha_C \geq \alpha_J$ and since $\eta = 1$ for an experimental topic (since all our messages are sent about true beneficial effects of immunization),

$$\mathrm{E}_{m,x}\left[\varphi\left(LR\left(\eta|m,x;C,\theta\right)\right)\right] \geq \mathrm{E}_{m,x}\left[\varphi\left(LR\left(\eta|m,x;J,\theta\right)\right)\right]$$

and therefore the distribution of instrumental payoffs $Q_{C,\theta} \succ Q_{J,\theta}$ for each θ. Note that this depends both on the originator and the type of the individual.

To see the effect, consider the case when $\alpha_C \to 1$. In this case, following the intuition discussed above, $Q_{C,H} \to Q_{C,L}$ and let $\widehat{Q}_C(\cdot)$ be the resulting CDF of the instrumental value, so there is nothing to signal at all. Thus $\underline{v}^C = 0$ and so anyone with any positive instrumental value immediately retweets. In contrast, with Joes/Janes, as above there is some negative $\underline{v}^J < 0$ that sets the equilibrium.

Consequently, the retweeting share is given by

- $\widehat{G}_C(0)$ under Celebrity origination and
- $\frac{1}{2}G_{J,H}\left(\underline{v}^J\right) + \frac{1}{2}G_{J,L}\left(\underline{v}^J\right)$ under Joe origination.

Notice that it is not clear which dominates. On the one hand, since $\eta = 1$ is essentially revealed as $\alpha_C \to 1$, $\widehat{G}_C$ has a higher mean than $G_{J,\theta}$ for either θ. On the other hand, the cutoff $\underline{v}^J$ can be considerably below 0 making the point of evaluating the $G_{J,\theta}$ CDFs at a lower point. This is because the likelihood ratio distribution of knowing that we are in a "good" world is not the same under celebrities (where it is substantially more likely) and Joes/Janes (where it is less likely, but there is a signaling effect reason to retweet).

Remark **1.** *The total endorsement effect we identify in the experiment can be thought of being comprised of (a) a shift in instrumental value and (b) a shift in the threshold to retweet due to the signaling effect. To see this*

$$\begin{aligned}\frac{1}{2}\left[\widehat{G}_C(0) - G_{J,H}\left(\underline{v}^J\right)\right] + \frac{1}{2}\left[\widehat{G}_C(0) - G_{J,L}\left(\underline{v}^J\right)\right] = \frac{1}{2}\left[\widehat{G}_C(0) - G_{J,H}(0)\right] + \frac{1}{2}\left[G_{J,z,H}(0) - G_{J,H}\left(\underline{v}^J\right)\right] \\ + \frac{1}{2}\left[\widehat{G}_C(0) - G_{J,L}(0)\right] + \frac{1}{2}\left[G_{J,z,L}(0) - G_{J,L}\left(\underline{v}^J\right)\right].\end{aligned}$$

In this expression, the $\widehat{G}_C(0) - G_{J,\theta}(0)$ term measures how for a given cutoff of 0, the amount of retweets increases when a message is originated by the celebrity, and the $G_{J,\theta}(0) - G_{J,\theta}\left(\underline{v}^J\right)$ term measures the change in the share of retweets when we move the cutoff to the left due to the signaling effect, holding the distribution fixed. When the signaling impetus is dominant,

this second term can overtake the prior term, making even a celebrity originator generate a lower volume of retweets.

F.4.2. *Sourcing.* In this case, holding originator fixed, we study the effect of adding a source. The analysis is identical to the case with celebrities. Ex-ante it seems reasonable to model sourcing as having direct positive effect on the likelihood of the signal being true: $\alpha_S \geq \alpha_{NS}$. Consequently

$$\mathrm{E}_{m,x}\left[\varphi\left(LR\left(\eta|m,x;S,\theta\right)\right)\right] \geq \mathrm{E}_{m,x}\left[\varphi\left(LR\left(\eta|m,x;NS,\theta\right)\right)\right].$$

This comes from the fact that a sourced tweet is just more likely to be right, so the likelihood ratio will be higher in distribution so for every originator and type of F, sourced tweets have more value in distribution so $Q_S \succ Q_{NS}$.

Again, if we assume sources are fully revealing $\alpha_S \to 1$ but without a source we have $\underline{v}^{NS} < 0$. Retweeting shares are given by $\widehat{G}_S\left(0\right)$ and $\frac{1}{2}G_{NS,H}\left(\underline{v}^{NS}\right) + \frac{1}{2}G_{NS,L}\left(\underline{v}^{NS}\right)$ under sourcing and no sourcing, respectively.

Crucially, even assuming sources are intrinsically good, retweeting can be reduced. This comes from the fact that the perception payoff effect can simply outweigh the gains in quality. If there is a source there is nothing to signal, whereas if there is no source F has a signaling motivation that is traded off against quality.

Remark **2.** *A natural question to ask is whether, since the arguments for celebrity versus Joe/Jane and sourced versus unsourced are identical, if anything seemingly relabeling, then the effects of sourced messaging and celebrity origination must have the same sign. But more careful reflection demonstrates that this is not true. Recall that retweeting share can be non-monotonic in α in this model. That is, given an initial α, a move to some $\alpha' > \alpha$ can lead to a decline in retweeting share and whether this is the case can depend on (π_H, π_L, λ). Concretely, recall the case of $(\pi_H = 0.65, \pi_L = 0.5, \lambda = 50)$ in Figure B.1 where the retweet share is non-monotonic with α. Thus, the increase due to a celebrity versus the increase due to adding a source need not be the same and in fact can generate different signs on retweeting behavior.*