PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES

Screening *p*-hackers: Dissemination noise as bait

Federico Echenique[a,1,2] and Kevin He[b,1]

Edited by Jose Scheinkman, Columbia University, New York, NY; received January 13, 2024; accepted April 7, 2024

We show that adding noise before publishing data effectively screens *p*-hacked findings: spurious explanations produced by fitting many statistical models (data mining). Noise creates "baits" that affect two types of researchers differently. Uninformed *p*-hackers, who are fully ignorant of the true mechanism and engage in data mining, often fall for baits. Informed researchers, who start with an ex ante hypothesis, are minimally affected. We show that as the number of observations grows large, dissemination noise asymptotically achieves optimal screening. In a tractable special case where the informed researchers' theory can identify the true causal mechanism with very few data, we characterize the optimal level of dissemination noise and highlight the relevant trade-offs. Dissemination noise is a tool that statistical agencies currently use to protect privacy. We argue this existing practice can be repurposed to screen *p*-hackers and thus improve research credibility.

p-hacking | research integrity | dissemination noise | privacy

Significance

Motivated by recent problems with research integrity in the behavioral sciences, we develop a model of researcher incentives and propose "dissemination noise" as a way to screen *p*-hacked findings that arise from data mining. In our model, *p*-hackers use observational data to uncover spurious explanatory mechanisms, while honest researchers use the same data to test ex ante hypotheses. We find that intentionally adding noise to data before making data public helps distinguish spurious correlations from genuine causal mechanisms. We characterize the optimal noise level in a tractable special case. This approach repurposes a privacy-protection technique currently used by data producers (e.g., the US Census Bureau) to help improve research credibility.

In the past 15 y, academics have become increasingly concerned with the harms of *p-hacking*: researchers' degrees of freedom that lead to spurious empirical findings. For the observational studies that are common in economics and other social sciences, *p*-hacking often takes the form of multiple testing: attempting many regression specifications on the same data with different explanatory variables, without an ex ante hypothesis, and then selectively reporting the results that appear statistically significant. Such *p*-hacked results can lead to misguided and harmful policies, based on a mistaken understanding of the causal relationships between different variables. Recent developments in data and technology have also made *p*-hacking easier: Today's rich datasets often contain a large number of covariates that can be potentially correlated with a given outcome of interest, while powerful computers enable faster and easier specification-searching.

In this paper, we propose to use dissemination noise to address and mitigate the negative effects of *p*-hacking. Dissemination noise is pure white noise that is intentionally added to raw data before the dataset is made public. Statistical agencies, such as the US Census Bureau, already use dissemination noise to protect respondents' privacy. Our paper suggests that dissemination noise may be repurposed to screen out *p*-hackers. Noise can limit the ability of *p*-hackers to "game" standards of evidence by presenting spurious but statistically significant results as genuine causal mechanisms. We show that the right amount of noise can serve as an impediment to *p*-hacking, while minimally impacting honest researchers who use data to test an ex ante hypothesis.

p-Hacking. Spurious results in many areas of science have been ascribed to the ability of researchers to, consciously or not, vary procedures and models to achieve statistically significant results. The reproducibility crisis in psychology has been blamed to a large extent on *p*-hacking (1–3). (4) evaluate experiments in economics and find a significant number of experiments that do not replicate.[*]

Most empirical work in economics and other social sciences are observational studies that use existing field data, not experiments that produce new data. Observational studies lead to a different sort of challenge for research credibility, where *p*-hacking stems mostly from discretion in choosing explanatory variables and econometric specifications. In experimental work, one remedy for *p*-hacking is preregistration: Researchers must describe their methods and procedures before data are collected. But this solution is not applicable for observational studies because researchers may have already accessed the public dataset before preregistering.

Dissemination Noise. Dissemination noise is currently used by major statistical agencies to protect people's privacy. The US Census Bureau, for instance, only disseminates a noisy version of the data from the 2020 Census. The practice is not new. Previously, the

[*] See also refs. 5 and 6. Imai et al. (7) find evidence against *p*-hacking in experimental economics.

Author affiliations: [a]Department of Economics, University of California, Berkeley, CA 94720; and [b]Department of Economics, University of Pennsylvania, Philadelphia, PA 19104

Author contributions: F.E. and K.H. designed research; performed research; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] F.E. and K.H. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: fede@econ.berkeley.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2400787121/-/DCSupplemental.

Published May 17, 2024.

Bureau has released a tool called "On the map" whose underlying data was infused with noise. Even earlier technologies for preserving respondent confidentiality like swapping data and imputing data can also be interpreted as noisy data releases. The contribution of our paper is to propose an alternative use for dissemination noise.

Setup and Key Results. We consider a society that wants to learn the true cause behind an outcome variable. Researchers differ in their expertise: Some are *mavens* whose domain knowledge narrows down the true cause to a small set of candidates, and others are *hackers* with no prior information about the true cause. Researchers derive utility both from reporting the true cause and from influencing policy decisions. So uninformed hackers have an incentive to game the system by using the data to fish for a covariate that would convince the policymaker.

We show that dissemination noise can help screen researcher expertise by introducing spurious correlations that can be proven to be spurious. These noise-induced correlations act like *baits* for p-hackers. But at the same time, they also make the data less useful for the informed mavens who use the data to test a specific ex ante hypothesis.

We explore this trade-off in our model. We show that as the number of observations grows large, dissemination noise asymptotically achieves optimal screening. In a tractable special case where the informed researchers' theory can identify the true cause with very few data, we characterize the optimal level of dissemination noise and derive comparative statics. The key intuition is that a small amount of noise hurts hackers more than mavens. All researchers act strategically to maximize their expected payoffs, but their optimal behavior differs. Mavens only entertain a small number of hypotheses, so a small amount of noise does not greatly affect their chances of detecting the truth. Hackers, by contrast, rationally try out a very large number of covariates because they have no private information about the true cause. The hackers' data mining amplifies the effect of even a small amount of noise, making them more likely to fall for a bait and get screened out. So, adding noise grants an extra informational advantage to the mavens, whose prior knowledge pinpoints a few candidate covariates. The hackers get screened out precisely because they (rationally) p-hack out of complete ignorance about the true cause.

We focus on a setting where the types of researchers primarily differ in terms of their expertise, not their incentives or biases. While our main results do allow the mavens and the hackers to assign different weights to correctly reporting the true cause versus influencing policy-making, our results are mainly driven by the fact that only the mavens have private information about the true cause. In terms of the classification of different kinds of p-hacking practices (see, for example, ref. 8), we focus on the problem of deterring capitalization on chance, where the researcher has no preconceived story but fishes around for anything that appears statistically significant in the data. We are not studying confirmation bias, where a researcher with a preconceived story looks for evidence that supports the story while discarding or downplaying evidence to the contrary.

We use a stylized model to represent researchers analyzing existing observational data for associations. Our intention is to explore a channel for screening researcher expertise in a simple and tractable setup. Of course, the practical usefulness of dissemination noise will need to be evaluated in more specific and realistic domains. Also, our focus is on simple correlational studies that use existing data: other research designs such as experiments that acquire new data or sophisticated econometric methods that exploit special structure of the data to credibly infer causation are outside of the scope of this work.

Alternative Solutions to p-Hacking. As already mentioned, the most common proposal to remedy p-hacking is preregistration. While it is a very good idea in many scientific areas, it is of limited use for observational studies, which are ubiquitous in the social sciences. Not only does it preclude useful exploratory work, it is also impossible to audit or enforce because publicly available data can be privately accessed by a researcher before preregistration.

A second solution is to change statistical conventions and make p-hacking more difficult. An extreme example is banning the use of statistical inference altogether (9). A less drastic idea is contained in ref. 10, which proposes to lower the P-value threshold for statistical significance by an order of magnitude—from 5% to 0.5%. Of course, this makes p-hacking harder, but a p-hacker armed with a sufficiently "wide" dataset and cheap enough computation power can discover spurious correlations that satisfy any significance threshold. We address this idea within our model and argue that our proposed use of dissemination noise is largely complementary to requiring more demanding statistical significance.

An idea related to our proposal is simply to reserve data for out-of-sample testing. Typically, the observations are partitioned into two portions. One portion is released publicly, and the rest is a "hold-out" dataset reserved for out-of-sample testing. We instead focus on a model of noise where each observation of each covariate is independently perturbed, which more closely resembles the kind of dissemination noise currently in use for privacy purposes. Our central message is that the current implementation of noise can be repurposed to screen out p-hacking. In addition, the kind of dissemination noise we study may be more applicable for datasets where the observations are not generated from an i.i.d. process, and thus it is less reasonable to designate some observations as a hold-out dataset (e.g., observations are different days in a time series). In an earlier version of the paper,[†] we show that our result about the benefit of a small amount of noise when mavens' theory can identify the true model with very few data continues to hold in the non-i.i.d. setting.

The out-of-sample approach is the focus of ref. 12. We differ in that we consider a world with two kinds of researchers and the dissemination noise here serves a screening role to separate the two types who act strategically to maximize their expected payoffs.

Related Literature. In economics, there is a recent strand of literature that seeks to understand the incentives and trade-offs behind p-hacking. Refs. 13–18 all study different games between a researcher (an agent) and a receiver (a principal). The agent has access to some p-hacking technology, which takes various forms such as repeatedly taking private samples and then selectively reporting a subset of favorable results to the principal, or sampling publicly but strategically stopping when ahead. These papers seek to better understand the equilibrium interaction between p-hacking agents and their principals, and study how such interactions are affected by variations in the hacking technology.

This literature differs from our work in two ways. First, they consider the case where hacking is costly. On this dimension, these papers about p-hacking are related to the broader literature

[†] See ref. 11, available at https://arxiv.org/pdf/2103.09164.pdf.

on "gaming," where agents can undertake costly effort to improve an observable signal (here, the P-value) beyond its natural level (e.g., ref. 19). We instead consider hackers who incur zero cost from p-hacking, motivated by our focus on researchers who data mine an existing dataset (which is essentially free with powerful computers). Absent any interventions, equilibria with zero hacking or gaming cost would be uninteresting. Our focus is instead on a specific intervention, dissemination noise, that can help screen out the p-hackers even though they face no hacking costs. The second difference is that these papers do not consider the problem of expertise screening. In our world, the principal's main problem is to provide sufficiently informative data to agents who have expertise while distorting the data enough to mislead another type of agent who tries to make up for their lack of expertise with p-hacking.

Di Tillio et al. (20) also study a game between a p-hacker and a principal, but give the agent some private information and the ability to select an area to do research in. This is a mechanism for hacking that is outside of the scope of our paper.

1. Model and Asymptotically Optimal Screening Using Dissemination Noise

1.1. The Baseline Model. We propose a model that captures the essence of how dissemination noise allows for expertise screening in an environment where nonexpert agents can p-hack, while keeping the model tractable enough to allow for analytic solutions.

1.1.1. The raw dataset. Consider an environment where each unit of observation is associated with an outcome Y and a set A of potential causal covariates $(X^a)_{a\in A}$. The outcome variable and all the covariates are binary. Suppose the dataset is wide, so the set of potential causes for the outcome is large relative to the number of observations. In fact, we assume a continuum of covariates; so $A = [0, 1]$. For instance, the covariates may indicate the presence or absence of different SNPs in a person's genetic sequence, while the outcome refers to the presence or absence of a certain disease.

There is one covariate $a^* \in A$, the *true cause*, with $\mathbb{P}[X^{a^*} = Y] = \psi$ for some $\psi \in (1/2, 1]$. So the true cause is positively correlated with the outcome, but it may not be perfectly correlated. For instance, a^* is the one SNP that causes the disease in question. There is also a *red herring* covariate $a^r \in A$ that is independent of Y. The red herring represents a theoretically plausible mechanism for the outcome Y that can only be disproved with data. For instance, a^r might be a SNP that seems as likely to cause the disease as a^* based on a biological theory about the roles of different SNPs.

Nature draws the true cause a^* and the red herring a^r, independently and uniformly from A. Then Nature generates the raw dataset $(Y_n, (X_n^a)_{a\in[0,1]})$ for observations $1 \le n \le N$. First, the values of the true cause in the N observations $(X_n^{a^*})_{1\le n\le N}$ are generated independently, each equally likely to be 0 or 1. Then, each Y_n is generated to match $X_n^{a^*}$ with probability ψ for $1 \le n \le N$, independently across n. Finally, covariates X_n^a for $a \neq a^*$, $1 \le n \le N$ are generated, each equally likely to be 0 or 1, independent of each other and of all other random variables. (So there is a continuum of independent Bernoulli random variables.) Equivalently, once a^* and a^r are drawn, we have fixed a joint distribution between Y and the covariates $(X^a)_{a\in A}$, and the raw dataset consists of N independent draws from this joint distribution. For instance, this may represent a dataset that shows the complete genetic sequences of N individuals and whether each person suffers from the disease.

1.1.2. Players and their incentives. There are three players in the model: a principal, an agent, and a policymaker. The *principal* owns the raw dataset, but lacks the ability to analyze the data and cannot influence policy-making norms. The principal disseminates a noisy version of the dataset, which we describe below. The *agent* uses the disseminated data to propose a covariate, $\hat{a}$. Finally, a *policymaker* evaluates the agent's proposal on the raw dataset using an exogenous test. We think of the agent as proposing an intervention: If this proposal passes, the policymaker will implement a policy that changes $X^{\hat{a}}$ in order to affect the value of Y. In the background, we implicitly assume that the principal grants the policymaker access to the raw data to conduct the test. [Alternatively, an earlier version of this paper (11) supposes that the principal periodically publishes noisy versions of the raw data for these tests. Such data releases will diminish the principal's ability to screen out p-hackers in the future.]

The policymaker's role is mechanical, and restricted to deciding whether the agent's proposal passes an exogenous test. We say that the proposal *a passes* if the covariate X^a equals the outcome Y in $M = \lfloor \gamma \cdot N \rfloor$ out of N observations, and that it *fails* otherwise. The parameter γ is an exogenous passing threshold with $1/2 < \gamma < \psi$. The policymaker will adopt a policy proposal if and only if it passes the test on the raw data. Passing the test does not require a to be the true cause of Y, for we could have some covariate $a \neq a^*$ where $Y_n = X_n^a$ for at least M observations by random chance.[‡]

The agent is either a maven (with probability $1 - h$) or a hacker (with probability h). Mavens and hackers differ in their expertise. A maven knows that the true cause is either a^* or a^r, and assigns them equal probabilities, but a hacker is ignorant about the realizations of a^* and a^r. The idea is that a maven uses domain knowledge (e.g., biological theory about the disease Y) to narrow down the true cause to the set $\{a^*, a^r\}$. A hacker, by contrast, is completely uninformed about the mechanism causing Y.

The agent's payoffs reflect both a desire for reporting the true cause and a desire for policy impact. If a type θ agent proposes a when the true cause is a^*, then his payoff is

$$w_\theta \cdot \mathbf{1}_{\{a=a^*\}} + (1 - w_\theta) \cdot \mathbf{1}_{\{\text{at least } \lfloor \gamma\cdot N \rfloor \text{ observations } n \text{ have } Y_n = X_n^a\}}.$$

Here, we interpret $\mathbf{1}_{\{a=a^*\}}$ as the effect of proposing a on the agent's long-run reputation when the true cause a^* of the outcome Y eventually becomes known. The other summand models the agent's gain from proposing a policy that passes the policymaker's test and gets implemented. The relative weight $w_\theta \in [0, 1]$ on these two components may differ for the two types of agent. Our main results in this section are valid for any values of w_{maven} and w_{hacker} in $[0, 1]$, but some later results in Section 2 will require restrictions on w_{maven}.

The principal obtains a payoff of 1 if a true cause passes, a payoff of -1 if any other $a \neq a^*$ passes, and a payoff of 0 if the agent's proposal is rejected. The principal's payoff reflects an objective of maximizing the positive policy impact of the research done on their data.

1.1.3. Dissemination noise. The principal releases a noisy dataset $\mathcal{D}(q)$ by perturbing the raw data. Specifically, they choose a *level of noise* $q \in [0, 1/2]$, and every binary realization of each covariate is flipped independently with probability q. So the noisy dataset $\mathcal{D}(q)$ is $(Y_n, (\hat{X}_n^a)_{a\in A})$, where $\hat{X}_n^a = X_n^a$ with probability $1 - q$, and $\hat{X}_n^a = 1 - X_n^a$ with probability q. The principal's choice of q is

[‡] There is no reward in our model for disproving a hypothesis.

common knowledge. A covariate a that matches the outcome in at least M observations in the noisy dataset but would not pass the policymaker's test—that is $\hat{X}_n^a = Y$ for at least M observations but $X_n^a = Y$ for fewer than M observations—is called a *bait*.

The form of noise in our model is motivated by the dissemination noise currently in use by statistical agencies, like the US Census Bureau. One could imagine other ways of generating a "noisy" dataset, such as selecting a random subset of the observations and making them fully uninformative, which corresponds to reserving the selected observations as a hold-out dataset for out-of-sample testing. Our analysis explores the possibility of repurposing the existing practice of adding dissemination noise, which more closely resembles perturbing each data entry independently than withholding some rows of the dataset altogether.[§]

1.1.4. Remarks about the model. We comment on our assumptions regarding the agents and the data in our model.

First, our model features very powerful p-hackers. A fraction h of researchers are totally ignorant about the true cause, but they are incentivized to game the system by fishing for some covariate that plausibly explains the outcome variable and passes the policymaker's test. This kind of p-hacking by multiple hypothesis testing is made easy by the fact that hackers have a continuum of covariates to search over and incur no cost from data mining. Our assumptions represent today's wide datasets and powerful computers that enable ever easier p-hacking. Our analysis suggests that dissemination noise can improve social welfare, even in settings where p-hacking is costless.

Second, the principal is an entity that wishes to maximize the positive social impact of the research done using their data but has limited power in influencing the institutional conventions surrounding how research results are evaluated and implemented into policies. In the model, the principal cannot change the policymaker's test. Examples include private firms like 23andMe that possess a unique dataset but have little say in government policy-making, and agencies like the US Census Bureau that are charged with data collection and data stewardship but do not directly evaluate research conclusions. Such organizations already introduce intentional noise in the data they release for the purpose of protecting individual privacy, so they may be willing to use the same tool to improve the quality of policy interventions guided by studies done on their data. In line with this interpretation of the principal, they cannot influence the research process or the policymaker's decision, except through changing the quality of the disseminated data. In particular, the principal cannot impose a cost on the agent to submit a proposal to the policymaker, write a contract to punish an agent who proposes a misguided policy, or change the protocols surrounding how proposals get tested and turned into policies.

Third, the dataset in our model contains just one outcome variable, but in reality a typical dataset (e.g., the US Census data) contains many outcome variables and can be used to address many different questions. We can extend our model to allow for a countably infinite number of outcome variables $Y^1, Y^2, \ldots$, with each outcome associated with an independently drawn true cause and red herring. After the principal releases a noisy version of the data, one random outcome becomes relevant and the agent proposes a model for this specific outcome. Our analysis remains unchanged in this world. This more realistic setting provides a foundation for the principal not being able to screen the agent types by eliciting their private information about the true cause without giving them any data. Who is a maven depends on the research question and the outcome variable being studied, and it is infeasible to test a researcher's domain expertise with respect to every conceivable future research question.

Fourth, the policymaker's exogenous test only evaluates how well the agent's proposal explains the raw dataset and does not provide the agent any other way to communicate his domain expertise. Such a convention may arise if domain expertise is complex and difficult to convey credibly: for instance, an uninformed hacker who has found a strong association in the data can always invent a plausible-sounding story to justify why a certain covariate causes the outcome. We also assume that the policymaker's test is mechanically set and does not adjust to the presence of p-hackers. This represents a short-run stasis in the science advocacy process or publication norms—for instance, while we know how to deal with multiple hypotheses testing, a vast majority of academic journals today still treat $P < 0.05$ as a canonical cutoff for statistical significance. Our analysis suggests that dissemination noise can help screen out misguided policies in the short run, when the principal must take as given a policymaking environment that has not adapted to the possibility of p-hacking.

To conclude, our basic model is meant to isolate the tradeoff between the cost imposed by noise on honest researchers and the benefit of screening p-hackers. A discussion of how the results are affected by relaxing our assumptions is in Section 3.

1.2. Screening Using Noise. We first derive the optimal behavior of the hacker and the maven.

Lemma 1. *For any $q \in [0, 1/2)$, it is optimal for the hacker to propose any $a \in A$ that satisfies $\hat{X}_n^a = Y_n$ for every $1 \leq n \leq N$. It is optimal for the maven to either propose $a \in \{a^*, a^r\}$ that maximizes the number of observations n for which $\hat{X}_n^a = Y_n$ (and randomize uniformly between the two covariates if there is a tie) or to propose any $a \in A$ that satisfies $\hat{X}_n^a = Y_n$ for every $1 \leq n \leq N$.*

Given the policymaker's exogenous test, hackers find it optimal to "maximally p-hack." Depending on the relative weight w_{maven} that mavens put on reporting the true cause, they will either use the noisy data to decide between their two true-cause candidates or engage in p-hacking. If the principal releases data without noise, then hackers will propose a covariate that is perfectly correlated with Y in the raw data. This covariate passes the policymaker's test, but it leads to a misguided policy with probability 1. The payoff to the principal from releasing the data without noise is therefore no larger than $1 - 2h$.

In fact, the principal cannot hope for an expected payoff higher than $1 - h$. This first-best benchmark corresponds to the policymaker always implementing the correct policy when the agent is a maven and not implementing any policy when the agent is a hacker. We show that with an appropriate level of dissemination noise, the principal's expected payoff approaches this first-best benchmark as the number of observations grows large.

Proposition 1. *For every q with $1-\gamma < q < 1/2$, the principal's payoff from using dissemination noise q converges to $1 - h$ as $N \to \infty$.*

[§] For instance, the Bureau publishes the annual Statistics of U.S. Businesses that contains payroll and employee data of small U.S. businesses. Statisticians at the Bureau say that separately adding noise to each business establishment's survey response provides "an alternative to cell suppression that would allow us to publish more data and to fulfill more requests for special tabulations" (21). The dataset has been released with this form of noise since 2007 (22). For the 2020 Census data, the Bureau will add noise through the new differentially private TopDown Algorithm that replaces the previous methods of data suppression and data swapping (23).

That is to say, dissemination noise is asymptotically optimal among all mechanisms for screening the two agent types, including mechanisms that take on more complex forms that we have not considered in our analysis.

The intuition is that noise does not prevent the agent from finding a policy that passes the policymaker's test whether his private information narrows down the true covariate to a small handful of candidates. But if the agent has a very large set of candidate covariates, then there is a good chance that the noise turns several covariates from this large set into baits. For example, if $N = 100$, $\psi = 0.95$, $\gamma = 0.9$, and $q = 0.15$, a covariate that perfectly correlates with Y in the noisy dataset has a 90% probability of being a bait. In the same environment, a maven who restricts attention to only two covariates (a^* and a^r) and proposes the covariate that correlates more with the outcome only fails the policymaker's test about 1% of the time. (As N grows for a fixed value of q in the range given by Proposition 1, the probability of a maven proposing a covariate other than a^* or a^r under his optimal strategy converges to zero). Hackers fall for baits at a higher rate than mavens because they engage in p-hacking and try out multiple hypotheses. Yet p-hacking is the hackers' best response, even though they know that the dataset contains baits.

While Proposition 1 applies asymptotically, the next result gives a lower bound on the number of observations such that a given level of noise is better than not adding any noise.

Proposition 2. *For every q with $1 - \gamma < q < 1/2$, the principal gets higher expected payoff with q level of noise than with zero noise whenever*

$$N \geq \max\left\{\frac{-\ln(h/8)}{2(q+\gamma-1)^2}, \frac{-2\ln(h/32)}{(\psi(1-q)+(1-\psi)q-0.5)^2}, \frac{-\ln(h/16)}{2(\psi-\gamma)^2}\right\}.$$

For example, when $\psi = 0.95$, $\gamma = 0.9$, $h = 0.1$, this result says $q = 0.15$ is better than $q = 0$ whenever $N \geq 1{,}016$.

2. Optimal Dissemination Noise in a Special Case

We now turn to a tractable special case where we can characterize the optimal level of noise with any finite number N of observations. We make two modifications relative to the baseline model discussed before.

First, we suppose the environment is such that the maven's theory only requires a minimal amount of data to distinguish a^* from a^r. Specifically, suppose we always have $Y_n = X_n^{a^*}$ and $X_n^{a^r} = 1 - X_n^{a^*}$ for every observation n. Unlike in the baseline model, the true cause is now perfectly correlated with the outcome Y and perfectly negatively correlated with the red herring X^{a^r}. We think of X^{a^*} as the causal covariate that determines the values of both X^{a^r} and Y. As before, the principal gets 1 if a proposal targeting a^* passes, -1 if any other proposal passes, and 0 if the proposal is rejected. Note that even though X^{a^r} is perfectly negatively correlated with the outcome, it does not cause the outcome. So a policy intervention that changes X^{a^r} is as ineffective at changing the outcome as a policy targeting any other covariate $a \neq a^*$.

Second, we suppose the policymaker uses the most stringent test. The proposal a passes if and only if $Y_n = X_n^a$ for all $1 \leq n \leq N$. (The principal can only do worse if the policymaker uses a more lenient test, as we will later show.)

As before, suppose agents maximize a weighted sum between reporting the true cause and passing the policymaker's test. Given the form of the test, the type θ agent's utility from proposing a when the true cause is a^* is:

$$w_\theta \cdot \mathbf{1}_{\{a=a^*\}} + (1 - w_\theta) \cdot \mathbf{1}_{\{Y_n = X_n^a \text{ for every } 1 \leq n \leq N\}}.$$

We suppose $0 \leq w_{\text{hacker}} \leq 1$ and $1/2 < w_{\text{maven}} \leq 1$.

Lemma 2. *For any $q \in [0, 1/2]$, it is optimal for the hacker to propose any $a \in A$ that satisfies $\hat{X}_n^a = Y_n$ for every $1 \leq n \leq N$, and it is optimal for the maven to propose $a \in \{a^*, a^r\}$ that maximizes the number of observations n for which $\hat{X}_n^a = Y_n$ (and randomize uniformly between the two covariates if there is a tie).*

When the agents follow the optimal behavior described in Lemma 2, the principal's expected utility from choosing noise level q is $-hV_{\text{hacker}}(q) + (1 - h)V_{\text{maven}}(q)$, where $V_\theta(q)$ is the probability that an agent of type θ's proposal passes the policymaker's test in the raw data, when the noise level is q. The next result formalizes the core idea that a small amount of noise harms the hackers more than the mavens.

Lemma 3. $V'_{maven}(q) = -\binom{2N-1}{N}Nq^{N-1}(1 - q)^{N-1}$ *and* $V'_{hacker}(q) = -N(1-q)^{N-1}$. *In particular,* $V'_{maven}(0) = 0$ *while* $V'_{hacker}(0) = -N$.

We can show that the principal's overall objective $-hV_{\text{hacker}}(q)+(1-h)V_{\text{maven}}(q)$ is strictly concave, and therefore the first-order condition characterizes the optimal q, provided the solution is interior:

Proposition 3. *If* $\frac{h}{1-h} \leq \binom{2N-1}{N}(1/2)^{N-1}$ *then the optimal noise level is* $q^* = \left(\frac{h}{1-h}\frac{1}{\binom{2N-1}{N}}\right)^{1/(N-1)}$. *More noise is optimal when there are more hackers and less is optimal when there are more observations. If* $\frac{h}{1-h} \geq \binom{2N-1}{N}(1/2)^{N-1}$ *then the optimal noise level is* $q^* = 1/2$.

Proposition 3 gives the optimal dissemination noise in closed form. With more hackers, screening out their misguided policies becomes more important, so the optimal noise level increases. With more observations, the same level of noise can create more baits, so the principal can dial back the noise to provide more accurate data to help the mavens.

2.1. Dissemination Noise and *P*-Value Thresholds.

Now suppose the principal can choose both the level of noise $q \in [0, 1/2]$ and a passing threshold $\underline{N} \in \{1, \ldots, N\}$ for the test, so that a proposal passes whenever $X_n^a = Y_n$ for at least $\underline{N}$ out of the N observations.

Proposition 4. *When the principal can optimize over both the passing threshold and the noise level, the optimal threshold is $\underline{N} = N$, and the optimal noise level is the same as in Proposition 3.*

We can interpret this result to say that stringent P-value thresholds and dissemination noise are *complementary tools* for screening out p-hackers and misguided policies. Think of different passing thresholds as different P-value thresholds, with the threshold $N = \underline{N}$ as the most stringent P-value criterion that one could impose in this environment. (10)'s article about lowering the "statistical significance" P-value threshold for new findings includes the following discussion:

> “The proposal does not address multiple-hypothesis testing, P-hacking, [...] Reducing the P value threshold complements—but does not substitute for—solutions to these other problems.”

Our result formalizes the sense in which reducing P-value thresholds complements dissemination noise in improving social welfare from research.

3. Concluding Discussion

We argue that infusing data with noise before making data public has benefits beyond the privacy protection guarantees for which the practice is currently being used. Noise baits uninformed p-hackers into reporting correlations that can be shown to be spurious. The paper investigates these ideas in a simple model that captures the trade-off between preventing hackers from passing off false findings as true and enabling legitimate research that seeks to test an ex ante hypothesis.

In an earlier version of the paper (11), we discuss extensions that relax the simplifying assumptions of our model.

1. We consider a situation where the N observations of each covariate are not i.i.d. We find that a small amount of dissemination noise still strictly improves the principal’s expected payoff in this setting.
2. We relax the assumption that there is a continuum of covariates. We find that fixing the number of observations, the same result goes through whenever the number of covariates is finite but large enough.
3. We suppose there is some chance that none of the covariates is a true cause for the outcome, so agents are asked to either report a cause or to say that no true cause exists. Our results are also robust to this extension.

The earlier paper (11) also contains a numerical simulation showing that the idea and the basic trade-offs for dissemination noise continue to hold in a more realistic empirical setting that is richer than our simple theoretical model. The simulation also shows that adding noise to the outcome variable may result in an overall smaller optimal amount of noise.

Finally, this earlier version considers a dynamic model with periodic noisy releases of a single dataset, where a finding submitted for validation in a given month is tested against the next month’s release of noisy data. We show that it remains optimal to release data with a strictly positive amount of noise, but over time the hackers’ access to all past data releases diminishes the effectiveness of noise.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. This research was made possible through the support of the Linde Institute at Caltech. Echenique also thanks the NSF’s support through the Grants SES-1558757 and CNS-1518941. We are grateful for comments from Sylvain Chassang, Eva Jin, Albert Ma, Pascal Michaillat, Marco Ottaviani, Nathan Yoder, and the audiences at the University of Pennsylvania, Caltech, Columbia University, Universidad de la República, Boston University, Allied Social Science Associations (ASSA) 2022, and Association for Computing Machinery Economics and Computation (ACM EC) ’22. Alfonso Maselli provided excellent research assistance.

1. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
2. U. Simonsohn, L. D. Nelson, J. P. Simmons, P-curve: A key to the file-drawer. *J. Exp. Psychol.: Gen.* **143**, 534 (2014).
3. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
4. C. F. Camerer *et al*., Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
5. C. F. Camerer *et al*., Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
6. A. Altmejd *et al*., Predicting the replicability of social science lab experiments. *PLoS ONE* **14**, e0225826 (2019).
7. T. Imai, K. Zemlianova, N. Kotecha, C. F. Camerer, How common are false positives in laboratory economics experiments? Evidence from the *p*-curve method (Working Paper, 2017).
8. R. J. MacCoun, S. Perlmutter, “Blind analysis as a correction for confirmatory bias in physics and in psychology” in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, S. O. Lilienfeld, I. D. Waldman, Eds. (John Wiley and Sons, 2017), pp. 295–322.
9. C. Woolston, Psychology journal bans *p* values. *Nature* **519**, 9 (2015).
10. D. J. Benjamin *et al*., Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
11. F. Echenique, K. He, Screening *p*-hackers: Dissemination noise as bait. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2103.09164 (Accessed 29 April 2024).
12. C. Dwork *et al*., The reusable holdout: Preserving validity in adaptive data analysis. *Science* **349**, 636–638 (2015).
13. E. Henry, Strategic disclosure of research results: The cost of proving your honesty. *Econ. J.* **119**, 1036–1064 (2009).
14. M. Felgenhauer, E. Schulte, Strategic private experimentation. *Am. Econ. J.: Microecon.* **6**, 74–105 (2014).
15. M. Felgenhauer, P. Loerke, Bayesian persuasion with private experimentation. *Int. Econ. Rev.* **58**, 829–856 (2017).
16. A. Di Tillio, M. Ottaviani, P. N. Sørensen, Strategic sample selection. *Econometrica* **89**, 911–953 (2021).
17. E. Henry, M. Ottaviani, Research and the approval process: The organization of persuasion. *Am. Econ. Rev.* **109**, 911–955 (2019).
18. A. McCloskey, P. Michaillat, Critical values robust to *p*-hacking. *Rev. Econ. Stat.*, in press (2024).
19. A. Frankel, N. Kartik, Muddled information. *J. Polit. Econ.* **127**, 1739–1776 (2019).
20. A. Di Tillio, M. Ottaviani, P. N. Sørensen, Persuasion bias in science: Can economics help? *Econ. J.* **127**, F266–F304 (2017).
21. T. Evans, L. Zayatz, J. Slanta, Using noise for disclosure limitation of establishment tabular data. *J. Off. Stat.* **14**, 537–551 (1998).
22. US Census Bureau, Technical documentation for statistics of U.S. businesses (2021). https://www.census.gov/programs-surveys/susb.html. Accessed 29 April 2024.
23. M. Hawes, R. Rodriguez, Determining the privacy-loss budget: Research into alternatives to differential privacy (2021). https://www2.census.gov/about/partners/cac/sac/meetings/2021-05/presentation-research-on-alternatives-to-differential-privacy.pdf. Accessed 29 April 2024.