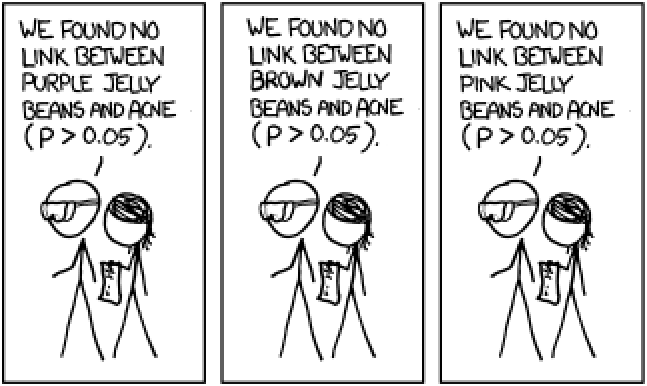
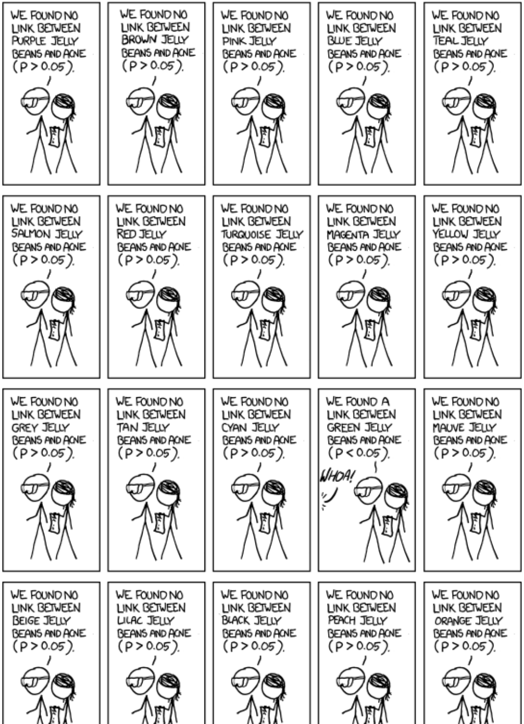
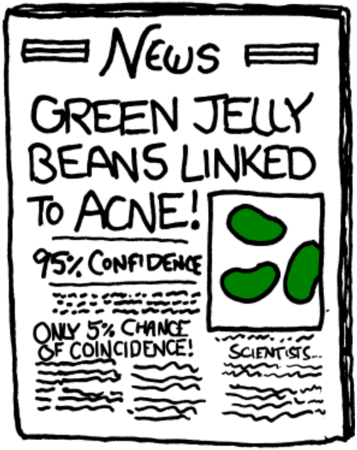
Screening p-Hackers: Dissemination Noise as Bait

Federico Echenique   Kevin He

Columbia   April 8, 2022


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

In Economics

I Just Ran Two Million Regressions

By XAVIER X. SALA-I-MARTIN*

Following the seminal work of Robert Barro (1991), the recent empirical literature on economic growth has identified a substantial number of variables that are partially correlated with the rate of economic growth. The basic methodology consists of running cross-sectional regressions of the form

An initial answer to this question
by Ross Levine and David Renelt
They applied Edward Leamer's
extreme-bounds test to identify "rol
pirical relations in the economi
literature. In short, the extreme-bo
works as follows. Imagine that there

Forensic use of DNA evidence

"Puckett's defense lawyer contacted the Arizona lab for more information about their findings, but the head of the lab denied the request. After a court issued a subpoena to compel the lab to disclose its findings, the analyst who had found the matching nine-locus pair testified that she had actually found ninety others within the database. When the lab offered no explanation for why 1 in 1 trillion events were happening regularly, the court ordered them to conduct a full search of the known-offender database and report back all matching pairs."

"The Dark Side of DNA Databases," Erin Murphy *The Atlantic* 2015.

p-Hacking

- Coined by Simmons, Nelson, and Simonsohn (2011), **p-hacking** = researcher degrees of freedom that lead to false statistical significance
- Attempt multiple covariates or econometric specifications, then selectively report the most significant one

p-Hacking

- *p*-hacking found in various disciplines (including economics): Gerber and Malhotra (2008a, 2008b), Brodeur et al. (2016, 2020), Christensen and Miguel (2018), Vivalt (2019)
- Pressing problem today as scope of *p*-hacking expands
  - Number of covariates explodes (e.g., 300 million SNPs in genomic data can be correlated with socioeconomic outcomes)
  - Specification-searching easier with more powerful computers
- How to mitigate harms of *p*-hacked results on policymaking, in a world that implements policies based on *p*-values and with technology that enables ever easier *p*-hacking?

Datasets Infused with Dissemination Noise

Seemingly unrelated news:

- 2020 US Census will feature Disclosure Avoidance System
  - inject noise into responses before releasing to public
  - goal: protect confidentiality of respondents
- Census Bureau has been using various kinds of dissemination noise since 1920's, including suppressing all data tables from small areas, imputing data, swapping data, etc

Key observation: though intended to protect respondent privacy, dissemination noise may also help prevent p-hacking

Main Ideas of This Project

Two kinds of researchers:

- *p*-Hackers
- Mavens

Main Ideas of This Project

- Dissemination noise turns some covariates into **"baits"** that appear correlated with an outcome variable in noisy data, but not in original data
  - Researchers analyze noisy data to propose policies
  - Policy proposal then checked using original data
  - This screens out p-hackers who fall for baits
- Trade-off for noise: Dissemination noise also degrades policy proposals from honest agents with legitimate use of data
- **This project**: how the steward of a unique dataset (e.g., US Census Bureau, 23andMe, ...) maximizes positive policy impact using the right amount of dissemination noise

Main Ideas of This Project

The key intuition for why dissemination noise can help screen out p-hackers is that a small amount of noise hurts hackers more than mavens.

Mavens entertain only a small number of hypotheses, so a small amount of noise does not interfere too much with their chances of detecting the truth.

Hackers, by contrast, rationally try out a very large number of model specifications because they have no private information about the true cause behind the outcome variable.

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, a strictly positive amount of noise is optimal.

Outline

1. Motivating numerical example and related literature
2. Basic model
3. Reusing the dataset — dynamic model
4. Extensions

Motivating Numerical Example

Data-generating process:

- One dependent variable; 20 covariates: $X_1, ..., X_{20} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.
- Principal gets 20 independent observations of $(Y, X_1, ..., X_{20})$ from their joint distribution
- The dataset is wide in the sense that there is a large number of possible models for the number of observations. Indeed, there are $\binom{20}{3} = 1140$ linear models of the form $Y = X^{i_1} + X^{i_2} + X^{i_3} + \epsilon$
- Enormous scope for data mining. p-hacker has $> 70\%$ chance of finding a regression that passes statistical muster.
- Reality: $Y = X_1 + X_2 + X_3 + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 4)$

Motivating Numerical Example

A policymaker, a data steward (principal), and a researcher (agent)

Policymaking procedure

- Uncertain which three covariates (i_1, i_2, i_3) generate Y
- Policy = a guess about the data-generating triplet (i_1, i_2, i_3)
  - 1 from correct guess, -1 from wrong guess, 0 from not guessing
- Agent proposes a triplet $(\hat{\imath}_1, \hat{\imath}_2, \hat{\imath}_3)$
- Policymaker guesses $(\hat{\imath}_1, \hat{\imath}_2, \hat{\imath}_3)$ if $Y = X_{\hat{\imath}_1} + X_{\hat{\imath}_2} + X_{\hat{\imath}_3} + \epsilon$ exceeds a critical R^2 threshold in original data, and makes no guess otherwise

Motivating Numerical Example

Principal's problem

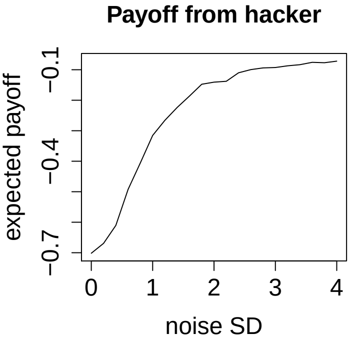
- Principal cannot affect policymaking procedure (fixed institutional norms outside of data steward's control)
- Disseminates noisy data to agent where $\mathcal{N}(0, \sigma^2_{noise})$ added to each realization of each covariate
- Wants to maximize expected utility from policy

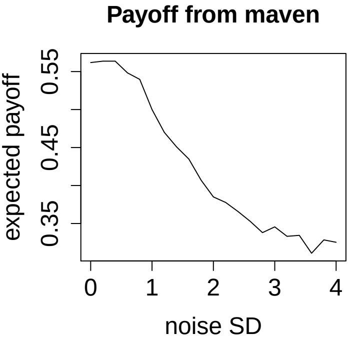
Agent's behavior

- Agent is either a maven or a hacker
- Maven knows correct policy either (1, 2, 3) or (4, 5, 6). Runs two regressions and reports triplet with higher R^2 in noisy data
- Hacker has no idea about correct policy, runs all 1140 possible regressions, reports triplet with highest R^2 in noisy data

Policymaker naively uses the $p < 0.05$ critical value that assumes no p-hacking: one of 1140 regressions chosen uniformly at random

Payoff Conditional on Agent Type

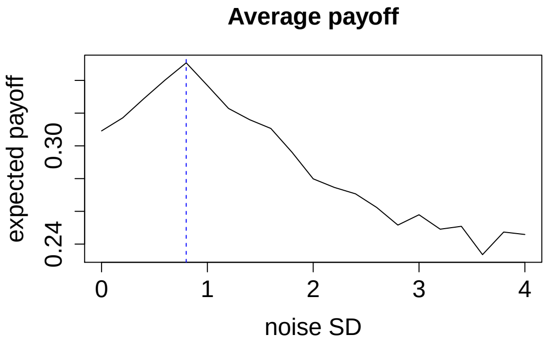




- Without noise, hacker easily finds (mostly wrong) triplet that passes critical threshold
- With noise, highest R^2 triplet in noisy data often a **bait** that fails to replicate

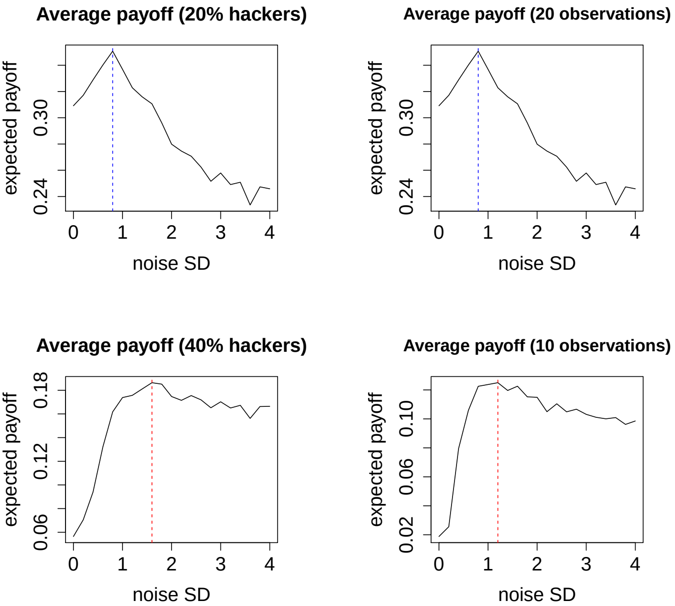
- Maven needs data to compare two policy candidates
- Noisy data makes it hard to figure out the correct one

Average Payoff and Optimal Noise



- Suppose 20% hackers, 80% mavens
- Optimal dissemination noise trades off screening out hackers via bait triplets VS preserving data quality for mavens
- Small σ_{noise} hurts hacker more than maven: likely that some baits are created, but unlikely that one bait happens to be $(4, 5, 6)$. Hackers screened out precisely because they p-hack.

Comparative Statics in Motivating Example



Key features of the model

Dataset is “wide.” Many possible potential explanations.

This means powerful hackers, who are likely to find a spurious correlation that passes statistical criteria.

Maven considers a small number of possible hypothesis.

Statistical standards are exogenously fixed.

Related Literature

Costly data acquisition with strategic disclosure: Henry (2009), Felgenhauer and Schulte (2014), Felgenhauer and Schulte (2017), Di Tillio, Ottaviani, and Sørensen (2017, 2021), McCloskey and Michaillat (2020 WP)

- We consider hackers who incur **no cost** from *p*-hacking: mining existing data, not collecting new data
- The "equilibrium" in these papers uninteresting with free hacking. Instead, we focus on an intervention that can screen out *p*-hackers even when they face no costs

Other approaches to increase research credibility

- Lower sig. threshold to $p < 0.005$: Benjamin et al. (2018)
  - Enough covariates and free data mining beats any threshold
  - Low *p*-value and adding noise are complements (more later)
- Pre-registration: Abrams, Libgober, List (2021 WP) find <5% of field experiments in top econ outlets have pre-analysis plan. Almost no pre-registration for observational studies due to credibility problems (Christensen and Miguel (2018)).

Model

Assume a simple DGP with infinitely many superfluous and irrelevant covariates.

Raw data is obtained as a finite sample of the DGP.

Hence a wide dataset.

Three players:

- Principal (think of US Census Bureau or 23andMe): Disseminates data.
- Agent: analyzes disseminated data to propose "policy" $\hat{a} \in [0, 1]$ for j.
- Policymaker: mechanical agent.

Model: Data-Generating Process

Ideas we seek to capture:

- “Wide” data set: many covariates, leading to powerful hackers.
- A principal – data steward – disseminates data.
- Decision rule is exogenous and fixed (think $p < 0.05$ rule).
- Principal doesn’t know the questions a priori.

Model

Principal (data steward) has limited scope in influencing how policies are implemented.

- Unmodeled stasis in publication norm / science advocacy process
- Principal has no power to elicit agent's domain expertise, reward accuracy, etc.
- Choosing a constrained-optimal mechanism by pulling one lever: the quality of data disseminated

Model: Data-Generating Process

Wide data set: many covariates, leading to powerful hackers.

- Continuum of binary **covariates** $X^{(a)}$ for $a \in [0,1]$
- Finite or countable binary **outcome variables** $Y^{(1)}, Y^{(2)}, \ldots$
- Each outcome $j \in \{1, 2, \ldots\}$ is associated with:
  - a **true cause** $a_j^* \in [0,1]$, so $Y^{(j)} = X^{(a_j^*)}$
  - a **red herring** $a_j^{\varnothing} \in [0,1]$ — plausible mechanism for j that can only be disproved with data.
  - For now, suppose $Y^{(j)} = 1 - X^{(a_j^{\varnothing})}$

Model: Data-Generating Process

- True cause and red herring drawn i.i.d. from Unif$[0, 1]$ for each outcome, and this fixes a joint distribution of (Y, X)
  - Each $Y^{(j)}$ i.i.d., equally likely 0 or 1, also determines $X^{(a_j^*)}, X^{(a_j^{\varnothing})}$
  - Other $X^{(a)}$ are i.i.d. equally likely 0 or 1

Model: Players and Policymaking Procedure

- Principal:
  - Gets raw data with N i.i.d. obs $(Y_n, X_n)_{n=1}^N$
  - Disseminates data. Then some $Y^{(j)}$ becomes relevant
- Agent: analyzes disseminated data to propose "policy" $\hat{a} \in [0, 1]$ for j
- Policymaker: mechanical, sets an exogenous policy implementation procedure
  - Implements $\hat{a}$ if $Y_n^{(j)} = X_n^{(\hat{a})}$ for all n

Agents: Hackers and Mavens

The true cause a^* and red herring a^r are drawn independently from the uniform distribution on A.

The maven knows that the true specification is either $Y = X^{a^*}$ or $Y = X^{a^r}$, and assigns them equal probabilities, but the hacker is ignorant about the realizations of a^* and a^r.

The idea is that the maven uses domain expertise (e.g., theory about the outcome Y) to narrow down the true cause to the set $\{a^*, a^r\}$.

The hacker, in contrast, is completely uninformed about the mechanism causing Y.

Model: Incentives and Information of Agents Types

Agent is a **maven** or a **hacker**:

$0 < h < 1$ fraction hackers;
$m = 1 - h$ fraction mavens.

- When $Y^{(j)}$ becomes relevant, maven uses domain expertise ("theory") to narrow down true cause to the set $\{a_j^*, a_j^\varnothing\}$
- Hacker has no information about true cause
- Agent cares about being right and being implemented

$$wU_{\text{right}} + (1-w)U_{\text{implemented}}$$

- Weight $w \in [0,1]$ possibly differs across agent types
- Any w works for hacker. Assume $w > 0.5$ for maven.

Model: Incentives and Information of Agents Types

Remark: A model with very powerful *p*-hackers

- Continuum of covariates to search over, no data-mining cost
- Represents today's "wide" datasets and fast computers

Remark: Theory and data are complements for learning true cause

- $\varphi =$ prob. that best guess about true cause is right
- $\varphi(\varnothing) = 0$, $\varphi(\text{data}) = 0$, $\varphi(\text{theory}) = \frac{1}{2}$, $\varphi(\text{theory} + \text{data}) = 1$

Model: Dissemination Noise

- Principal gets 1 if correct proposal implemented, -1 if wrong proposal implemented, 0 if proposal rejected
- Principal releases noisy dataset $(Y, \hat{X})$ where

$$\hat{X}_n^{(a)} = \begin{cases} 1 - X_n^{(a)} & \text{w.p. } q \\ X_n^{(a)} & \text{w.p. } 1 - q \end{cases}$$

  independently across a, n.
- $q \in [0, 1/2]$ the **noise level** is common knowledge

Marginal Impact of Noise on Different Types

Can derive behavior of hacker and maven from their utilities

- Hacker: proposes some $\hat{a}$ such that $\hat{X}_n^{(\hat{a})} = Y_n^{(j)}$ for every n
  - Wrong policy with probability 1, but may get accepted
- Maven: proposes a_j^* or $a_j^\varnothing$ depending on whether $\hat{X}^{(a_j^*)}$ or $\hat{X}^{(a_j^\varnothing)}$ matches $Y^{(j)}$ in more observations (randomize if tie)
  - Proposal accepted if and only if it is the true cause.

$V_i(q) =$ probability type i's proposal accepted with noise level q.

So principal maximizes

$$hV_{\text{hacker}}(q) \times (-1) + mV_{\text{maven}}(q) \times (1)$$

Marginal Impact of Noise on Different Types

Lemma

$V'_{maven}(0) = 0$ *but* $V'_{hacker}(0) < 0$.

- Low amount of noise does not prevent agent from finding a policy that gets accepted starting from small set of candidates
- But, high chance of baits in a very large set of candidates
- When $N = 100$, $q = 0.01$, $\mathbb{P}[a \text{ is bait} \mid \hat{X}^{(a)} = Y^{(j)}] > 63\%$
- But, $\mathbb{P}[a \text{ is bait} \mid \hat{X}^{(a)} = Y^{(j)}, \text{ and } a \in \{a_j^*, a_j^{\varnothing}\}] \approx 0\%$

Optimal Noise Level and Comparative Statics

Proposition 1

The optimal level of noise is

$$q^* = \min\{\frac{1}{2}, \left(\frac{h}{m\binom{2N-1}{N}}\right)^{1/(N-1)}\}$$

- More noise is optimal when there are more hackers and less is optimal when there are more observations.
- With more hackers, screening out their wrong policies becomes more important
- With more observations, same level of noise creates more baits
- With too many hackers we hit a boundary: optimal to not release data at all (equivalent to $q = 1/2$)

Optimal Noise Level and Comparative Statics

Proposition 2

The principal's expected payoff under the optimal noise level approaches $1 - h$ *as* $N \to \infty$.

That is, injecting the optimal level of noise is asymptotically optimal, among all mechanisms for screening the two agent types, including mechanisms that involve a hold-out dataset, or take more complex forms.

Extensions

In the paper:

- Non iid observations (eg time series).
- Red herrings that are "less wrong."
- Finite number of variables (covariates).
- No true cause.

When Principal Controls the Acceptance Threshold

- Assumed policymaker accepts $\hat{a}$ when $Y_n^{(j)} = X_n^{(\hat{a})}$ for every n
- Now let principal choose both q and threshold $\underline{N} \in \{1, ..., N\}$, proposal $\hat{a}$ implemented when $Y_n^{(j)} = X_n^{(\hat{a})}$ for at least $\underline{N}$ obs

When Principal Controls the Acceptance Threshold

Proposition

Principal's optimal acceptance threshold is $\underline{N} = N$.

- Choice of $\underline{N}$ does not affect analysis when facing maven
- Can show for any $\underline{N}$, hacker still proposes $\hat{a}$ with $Y^{(j)} = \hat{X}^{(\hat{a})}$
- Interpretation: stringent p-value threshold and dissemination noise are complementary tools for accurate policymaking

  Benjamin et al. (2018), Redefine statistical significance: "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 model formalizes the sense in which they are complements

Reusing Data - Dynamic Model of Noise

Think monthly releases of noisy data.

Each finding validated against next month's release.

All releases are public.

Reusing Data - Dynamic Model of Noise

- Transparency may demand that the principal publishes validation datasets
- Static model: check on original data since no future use
- Realistically, same dataset reused for multiple research questions over different years
- Once data exposed, can no longer screen out *p*-hackers
- To reuse dataset, must validate proposals on noisy data
    - Degrades accuracy, but retains some defense against future hackers
    - As more noisy versions of the data made public, hackers figure out true data values as **stock of randomness** depletes
- We illustrate principal's dynamic incentives in a simple setup

Reusing Data - Dynamic Model of Noise

- Time discrete and infinite, $t = 0, 1, 2, \ldots$
- Principal's data realizes once at $t = 0$. Assume $N = 1$.
- In period $t \geq 1$, outcome $Y^{(t)}$ is relevant and a short-lived agent arrives, uses all disseminated data in past to propose $\hat{a}$
- Principal releases a dataset with q_t noise level $(Y, \hat{X})$ to validate proposal, accept when $Y^{(t)} = \hat{X}^{(\hat{a})}$
- Assume (unlike before) maven always proposes true cause
- Principal maximizes δ-discounted expected utility, $0 < \delta < 1$
- Dwork et al. (2015) also embodies an intertemporal trade-off in exhausting the stock of randomness in a dataset
    - Different use of randomness: evaluate adaptively generated queries about DGP
    - Another difference: we characterize optimal solution to dynamic problem

Intertemporal Consumption of Randomness

- Hacker proposes $\hat{a}$ with $Y^{(t)} = \hat{X}^{(\hat{a})}$ in all past datasets
- Suppose such $\hat{a}$ has b_t chance of being bait ($Y^{(t)} \neq X^{(\hat{a})}$)
- Principal's utility today from noisy level q is

$$u(q_t; b_t) := m \times \underbrace{(1 - q_t)}_{X^{(a_t^*)} \text{ not flipped}} + h \times (- \underbrace{(1 - b_t)(1 - q_t)}_{\text{false positive}} - \underbrace{b_t q_t}_{\text{bait validates}})$$

- Think of $u(q; b)$ as "utility from consuming $\frac{1}{2} - q$ in state b" where stock of randomness left is b and state evolves:

$$b_{t+1} = \frac{b_t q_t}{(1 - b_t)(1 - q_t) + b_t q_t}$$

- We have $\frac{\partial u}{\partial q} < 0$, $\frac{\partial u}{\partial b} > 0$, and $\frac{\partial b_{t+1}}{\partial q_t} > 0$
  - More noise = less consumption (less accurate validation)
  - But, less noise depletes stock of randomness faster, easier to hack later ($b \downarrow$)

Principal Eventually Abandons Dissemination Noise

Principal's Bellman equation:

$$V(b) := \max_{q \in [0,1/2]} \left\{ u(q; b) + \delta V\left(\frac{bq}{(1-b)(1-q)+bq}\right) \right\}$$

Proposition

Suppose $h < 1/2$. In any solution to the principal's problem, there exists finite t^* such that:

- *If $t < t^*$, then $0 < q_t < 1/2$ and $b_{t+1} < b_t$*
- *If $t \geq t^*$, then $q_t = 0$ and $b_{t+1} = 0$*

- Principal disseminates partly noisy datasets for $t^* - 1$ periods
- In period t^*, gives up and publishes original data without noise
- From then on, data fully exposed and hackers uninhibited
- Why? More noise needed to slow decline of b when b lower

Extension: Non-i.i.d. Observations

- For a given outcome variable $Y^{(j)}$ or covariate $X^{(a)}$, assumed so far that its N observations are i.i.d.
- Relax this assumption: unconditional distribution of each $X^{(a)}$ is any full-support $\mu \in \Delta(\{0,1\}^N)$
  - Time-series data on different economic indicators ($n =$ year)
  - Characteristics of N individuals on a social network, where network neighbors more likely to be similar
- After true cause and red herring drawn for each outcome j, draw $Y^{(j)} \sim \mu$ and let $X^{(a_j^*)} = Y^{(j)} = 1 - X^{(a_j^\varnothing)}$
- Generate all other covariates $X^{(a)} \sim \mu$
- Unreasonable to release only a subset of observations
- But, small amount of i.i.d. dissemination noise still improves principal's expected payoffs

Proposition

For any μ, there exists $\bar{q} > 0$ s.t. the principal gets strictly higher expected payoff with any noise level $0 < q \leq \bar{q}$ than with $q = 0$.

Extension: More Misleading Red Herring

- Have focused on a story with the strongest possible complementarity between theory and data
- A single observation disproves the red herring since $Y^{(j)} = X^{(a_j^*)}$, $Y^{(j)} = 1 - X^{(a_j^{\varnothing})}$
- Small amount of noise still helps in more general settings
- Suppose for each outcome j, $Y^{(j)}$ and $X^{(a_j^{\varnothing})}$ are independent (like with any covariate other than j's true cause)
- Harder for maven to find true cause, also principal might implement red herring

Proposition

Provided $\frac{h}{m} > \frac{N+1}{2^{N+1}}$, there exists $\bar{q} > 0$ s.t. the principal gets strictly higher expected payoff with any noise level $0 < q \leq \bar{q}$ than with $q = 0$.

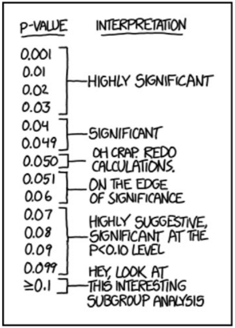
If $N = 10$, noise helps whenever more than 0.53% of agents p-hack

Takeaway Messages

- Dissemination noise is a data stewardship tool already in use that can serve the additional purpose of preventing p-hacking
- Noise creates baits that attract and screen out uninformed hackers, but minimally impact researchers with ex-ante theory
- Complements other approaches to research credibility, like lower p-value
- Stock of randomness in a new dataset that defends against p-hacking depletes as different noisy versions are made public. Principal solves intertemporal consumption of randomness.

Takeaway Messages

What is p-hacking?



https://imgs.xkcd.com/comics/p_values.png

Thank you!