Hardness of Preventing False Discovery

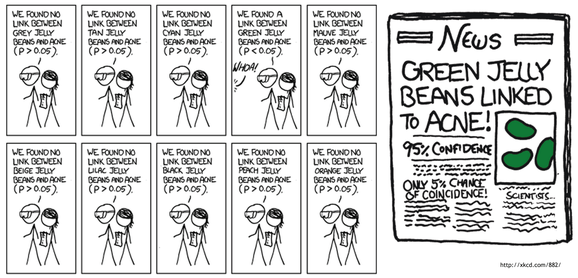
Thomas Steinke & Jonathan Ullman

Harvard & Columbia

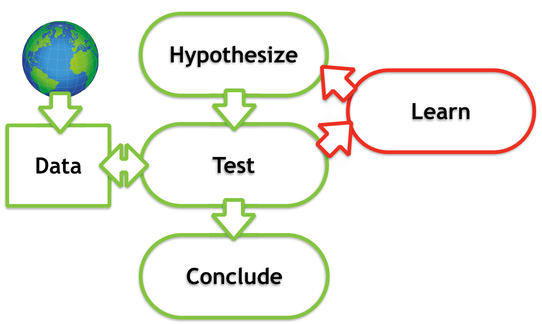
Privacy Tools for Sharing Research Data

A National Science Foundation Secure and Trustworthy Cyberspace Project

The Problem of False Discovery

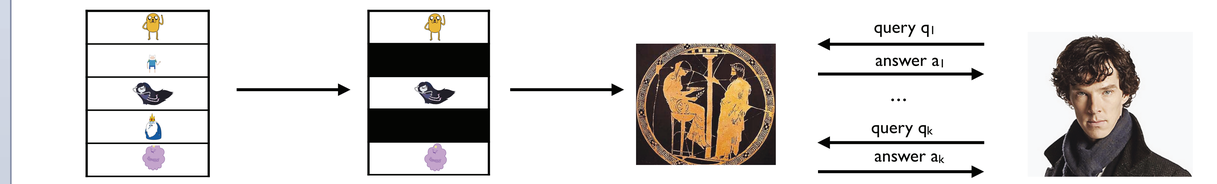


Science is Interactive



Question: How much data is needed to ensure conclusions are valid?

False Discovery and Statistical Queries



Distribution D over $\{0,1\}^d$ — Sample S of size n — Oracle — Interactive Analyst

False Discovery ≡ Inaccurate Answers

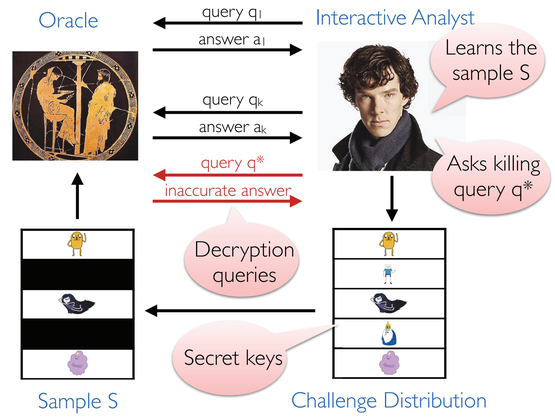
Oracle's goal: Ensure each answer a is accurate - $|a-q(D)|<0.01$

Question: How many samples does the oracle need?

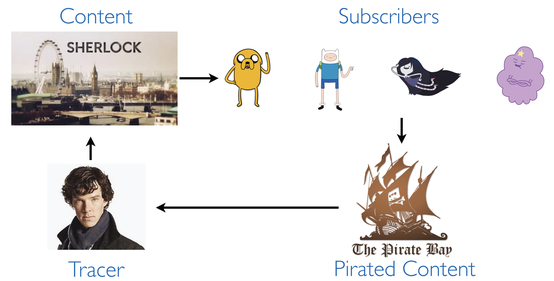
To answer k non-adaptive statistical queries, $n=O(\log k)$ samples suffices.

Directly answering k adaptive queries requires $n=\tilde{\Theta}(k)$ [DN03].

Techniques

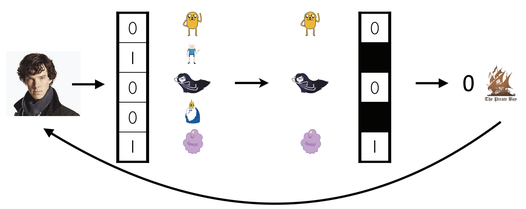


Interactive Fingerprinting Codes



Tracer's goal: Identify and disconnect pirates.

Tracer creates two versions of each episode.

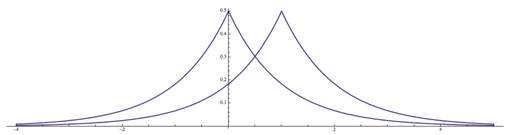


If all the pirates receive the same version, then that version appears online.

If there are n pirates, after $O(n^2)$ episodes all pirates are identified [BS95, FT01, T03, LDRSdW13, HU14, This work].

Preventing False Discovery

Differential Privacy [DMNS06]



An oracle is differentially private if it does not depend "too much" on any single sample item.

Theorem [DFHPRR14]: If the oracle is differentially private, then it prevents false discovery.

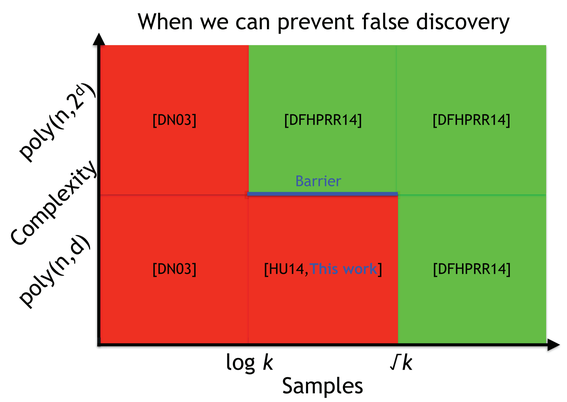
Corollary: There exists an **efficient** oracle that answers k queries, prevents false discovery, and uses $n=\tilde{O}(\sqrt{k})$ samples.

Corollary: There exists an **inefficient** oracle that answers k queries, prevents false discovery, and uses $n=\tilde{O}(d \log k)$ samples.

Question [HU14]: Is the gap between efficient and inefficient oracles inherent?

Our Results

Preventing false discovery has an inherent computational barrier.



Theorem [This work]: Any efficient oracle that answers k queries and prevents false discovery must have use $n=\Omega(\sqrt{k})$ samples.

Our result is a nearly-optimal strengthening of [HU14] (which proves $n=\tilde{\Omega}(k^{1/3})$) and holds under the assumption of strong OWFs and $d \gg \log n$.

Author contacts: tsteinke@seas.harvard.edu jullman@cs.columbia.edu

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences

BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The Institute for Quantitative Social Science at Harvard University

DATA PRIVACY LAB