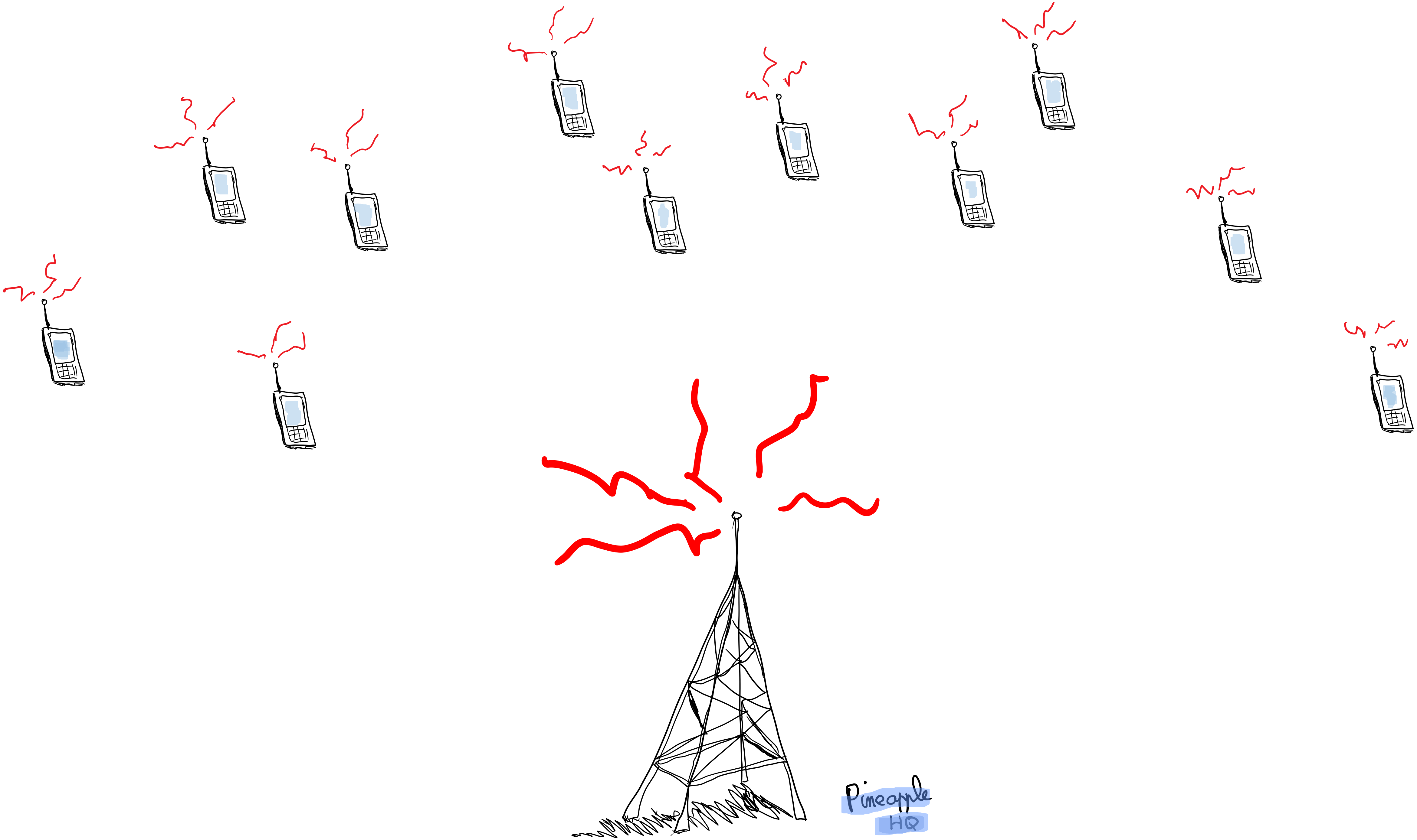
Distributed goodness-of-fit: when you can't talk much and have little in common

Jayadev Acharya (Cornell), Clément Canonne (Stanford), Yanjun Han (Stanford), Ziteng Sun (Cornell), and Himanshu Tyagi (IISc Bangalore)

Insert motivating story here.

(~~something about boats?~~)


Pineapple
HQ

Goal

Perform statistical inference in this distributed setting

Goal

Perform statistical inference
in this distributed setting

SMP

identity testing (goodness-of-fit)

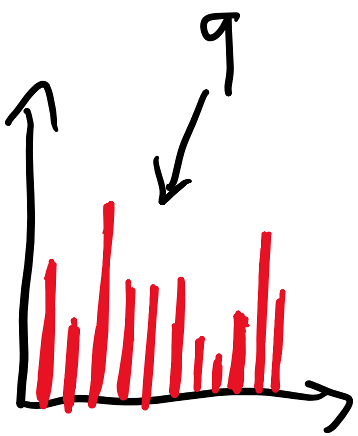
Goal

Perform ~~statistical inference~~
in this distributed setting

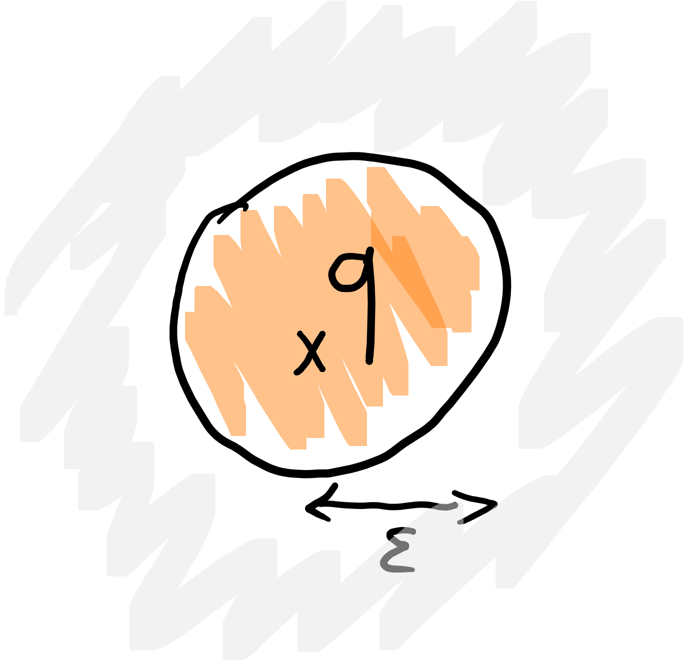
SMP

Identity testing

Known distribution q on $[k] = \{1, 2, \dots, k\}$



Distance parameter ε



H_0: $p = q$

vs.

H_1: $TV(p, q) > \varepsilon$

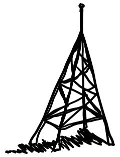
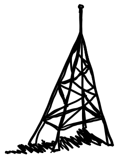
What's known

- Centralized: $\Theta\left(\frac{\sqrt{k}}{\varepsilon^2}\right)$ [Pan'08, VV'17]
- Our setting: $\Theta\left(\frac{k}{2^{\ell/2}\varepsilon^2}\right)$ public-coin [ACT'19$_{1,2}$]

  $\Theta\left(\frac{k^{3/2}}{2^{\ell}\varepsilon^2}\right)$ private-coin [ACT'19$_{1,2}$]

ℓ: # bits each player can send

Oh yes, about that...

Public-coin:  and all 's share common random seed
(hardcoded or broadcast to all by )

Private-coin: every  for itself (indep^{t} randomness)

(discuss both settings here)

Moreover, the optimal public-coin protocol only requires $O(\log k)$ shared random bits.

Yay.

However...

What happens when we have only few shared random bits? Say, $\sqrt{\log k}$, or 15?

Let's say s.

Theorem. For any $\ell \geq 1$, $s \geq 0$ s.t. $\ell + s \leq \log k$, there is a protocol for identity testing w/

$$\frac{\sqrt{k}}{\varepsilon^2}\sqrt{\frac{k}{2^{\ell}}}\sqrt{\frac{k}{2^{s+\ell}}}$$

phones. Moreover, this is tight.

~~Proof~~. Ideas.

Lower bound

private-coin

public-coin

Generalize the maxmin and minmax decoupled χ^2-fluctuations notions from [ACT'19] to limited randomness: semimaxmin

$$\underline{\overline{\chi}}(\mathcal{W}^n, \varepsilon, s) = \sup_{\substack{\mathcal{W}_s \subseteq \mathcal{W}^n \\ |\mathcal{W}_s| \leq 2^s}} \inf_{P \in \Upsilon_\varepsilon} \mathbb{E}_{W^n}\left[\chi^{(2)}(W^n \mid P)\right]$$

Upper bound

"Derandomization" of key anticoncentration lemma in [ACT'19]

<u>Theorem.</u> $\forall n$, $\forall x \in \mathbb{R}^n$, $\exists$ $m \lesssim n$ subsets $S_1, \dots, S_m \subseteq [n]$ w/ $|S_j| = \frac{n}{2}$
s.t.

$$\mathbb{P}_{j \sim [m]} \left\{ \left| \sum_{i \in S_j} x_i \right|^2 \gtrsim \|x\|_2^2 \right\} \gtrsim 1$$

Apply this with $n := 2^s$ to reduce the problem to private-coin identity testing over a domain of size $L \asymp \frac{k}{2^s}$, w/ distance $\varepsilon' \asymp \frac{\varepsilon}{\sqrt{2^s}}$.

Using the protocol from [ACT'19]:

$$\frac{L^{3/2}}{2^\ell \varepsilon'^2} \asymp \frac{k^{3/2}}{2^\ell \sqrt{2^s}\,\varepsilon^2} = \frac{\sqrt{k}}{\varepsilon^2} \cdot \sqrt{\frac{k}{2^\ell}} \cdot \sqrt{\frac{k}{2^{s+\ell}}}$$

phones suffice, as claimed*.

*OK, with a catch.

This gives a tester w/ low soundness (big Type-II error).
And we cannot amplify by repetition: our s bits are gone!

*OK, with a catch.

This gives a tester w/ low soundness (big Type-II error).
And we cannot amplify by repetition: our s bits are gone!

Solution: boost using the same randomness (but 100x more phones)!

Details omitted. See board.

Open questions

More uses of this "derandomization" lemma?

Of this probability amplification technique?

Extend to other local constraints?

Is my handwriting that terrible?

Thank you.