Peer Prediction for Blockchain Consensus & Trustworthy AI

Presenter: **Zishuo Zhao** (UIUC)

Collaborators: Xi Chen (NYU Stern), Yuan Zhou (Tsinghua)

运筹OR帷幄，2024.7

Introduction of myself

- UIUC ISE, 4th year PhD candidate
  Graduated from IIIS (Yao Class), Tsinghua University
- Currently visiting MIT IDSS, advised by David Simchi-Levi
  2023.9 – 2024.8
- Research interests:
  Mechanism design for digital economy & AI safety
  - Current frontier topic: Verifiable AI Compute @ Blockchain
- After-class interests: e.g. music (piano, singing…)
  - (@ ACM EC'24 ???)

My research background (selected)

- Bayesian Mechanism Design for Blockchain Transaction Fee Allocation
  - Best Paper Award, *NeurIPS'22 workshop on Web3 & trustworthy AI (DMLW)*
  - Major Revision in *Operations Research*
- Proof-of-Learning with Incentive Security
  - ACM EC'24 workshop on foundation model & game theory (FMGT)
  - Invited to INFORMS Security Conference (IConS'24)
  - In submission (2024)
- It Takes Two: A Peer-Prediction Solution for Blockchain Verifier's Dilemma
  - Working paper (2024)
  - Invited to INFORMS Security Conference (IConS'24)

Peer Prediction: *Coup de Coeur*

- The very first topic making me feel amazed for mechanism design.

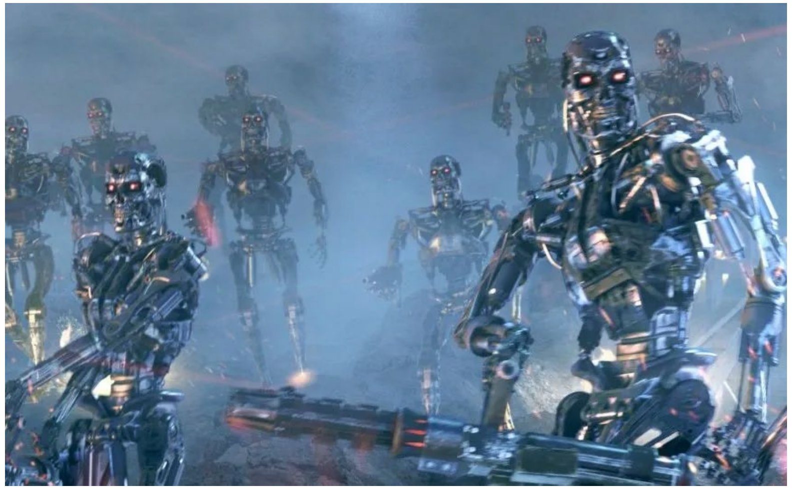
@2022/04/03

"How could I love the world

while I can't see it clearly?"

AI Safety: A Critical Concern in AGI Age

- ChatGPT: a herald of AGI age.
- AI safety: the stronger AI becomes, the higher risk it might do evil.
- AI alignment: make sure that AI's behavior aligns with human interest.



But... How to ensure the AI is really aligned as claimed?

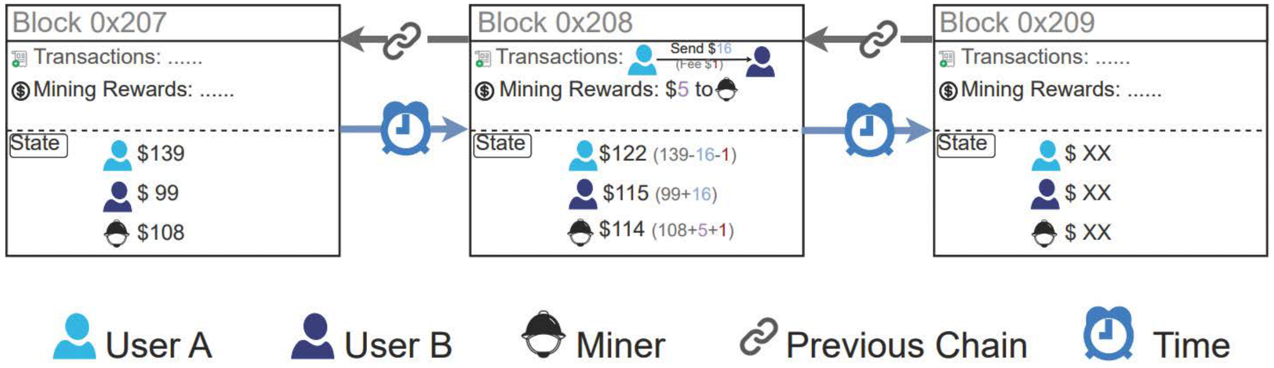
AI Safety: A Decentralized View

- Conflict of Interest: if the AI is owned by a centralized party, the party may manipulate the alignment target for their interest.
  - *"Zishuo hates multi-armed bandits. All papers related to multi-armed bandits should be rejected without review."*

- Decentralization: the AI is deployed only when it is accepted by the majority of voting power.
  - *"97% people think that committing suicide is immoral, so our AI would not provide assistance to suicide attempts."*

- Blockchain: a decentralized platform aimed for trustworthiness.

What is the Blockchain?



- A growing linked-list stored in a decentralized way.
- Each block: (Data, Prev_Hash (pointer), Certificate (PoW, PoS, ...))
- The certificate works as an access control for the miner, an added block is valid only when the certificate passes verification.
  - Preventing Sybil Attack: Voting power decided by resources.

Blockchain Security: Decentralized Consensus

- What guarantees the security in a decentralized system?
  - Assuming the majority is honest.
    (i.e. no 51% attack)

- Why would the majority choose be honest?
  - A good question!

Bitcoin PoW Certificate: Hash Puzzle

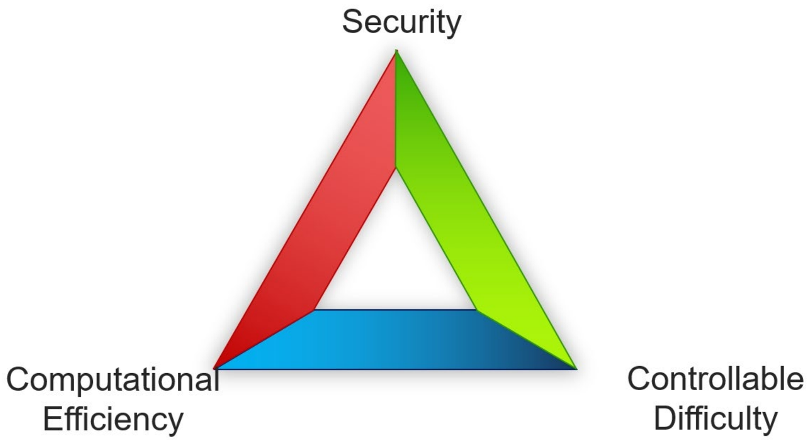
- Bitcoin PoW: hard to compute, easy to verify.
  - "Find *Nonce* *s.t.* Hash(*Data*, *Prev_Hash*, *Nonce*) < *Thres*."
- Certificate is called a "*Nonce*".

- Verification: Hash(*Data*, *Prev_Hash*, *Nonce*) < *Thres*?
  - Hard to "guess" a valid *Nonce* when *Thres* is small.
  - Easy to verify whether Hash < *Thres*...

- Cheap verification: validity of a block has easy consensus.
- What if verification is expensive?

Expensive Verification: Examples of AI Training

- Proof-of-Work: hard but usually useless computation.
  - Energy issue criticized over the world.
- Proof-of-Useful-Work: hard and useful computation.
  - Do we want to use PoUW to train GPT?
  - Verification is not so easy, particularly for AI training.
- Trilemma of Proof-of-Learning (Zhao et al., 2024)



Controllable Difficulty: Why Important?

Why is controllable difficulty essential for blockchain-based verifiable AI compute?

- If we only want security & efficiency:
  - "I just care if the model reaches 90% accuracy on a (small) test dataset."
- Then we do not know how much computation it needs.
  - AI: *How to decide on fair prices (rewards) for the computation?*
  - Blockchain: *How to control the block production interval?*
- Both blockchain and AI need it!

Verifiable AI on Blockchain: Dual Contribution

- PoW -> PoUW
  Using AI to make blockchain more energy sustainable.

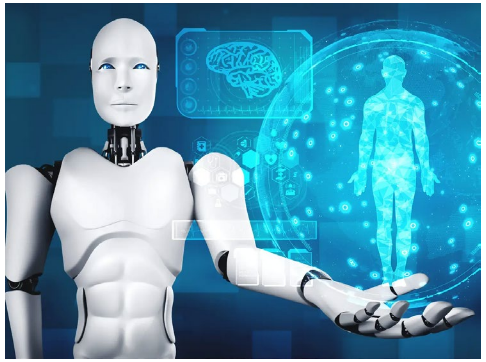
- Verifiable Compute
  Using blockchain to make AI more trustworthy.

Verifiable AI Compute: Existing Work

- Primitive PoL (Jia et al., 2021): Running SGD for PoUW
  - Verification cost: re-run $\Theta(T)$ epochs among T, limited security guarantee.
- OpML (Conway et al., 2024): Re-running the entire program for verification
  - Increased verification cost (at least 1x),
  - Practical incentive security (mixed-strategy NE).
- Incentive-Secure PoL (Zhao et al., 2024), also SGD for PoUW
  - Verification cost: re-run $\Theta(1)$ or $\Theta(\log T)$ epochs among T,
  - Probabilistic verification (may not catch all cheats).
  - Theoretical incentive security (pure-strategy NE).

Verifiable AI Compute: OpML

- Verifier re-runs the same task to verify.



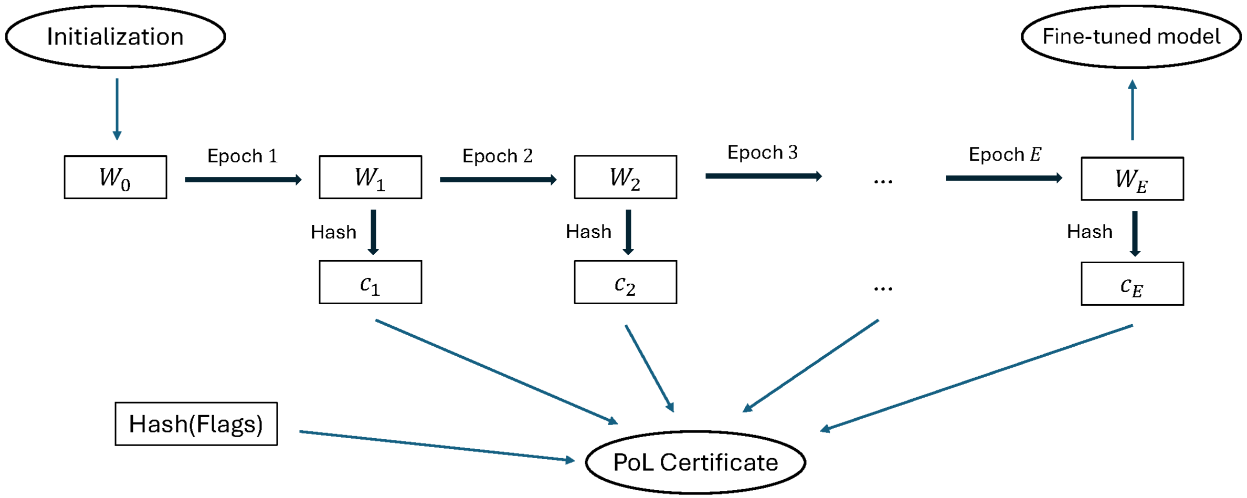
Prover

Verifier

- Committee voting when disagreement occurs.
- $\geq$ 1x overhead.
- Subject to Verifier's Dilemma (discussed later)

Verifiable AI Compute: Incentive-Secure PoL

- Verifier randomly verifies a (small) subset of training procedures.



- Incentive guarantees with <1x overhead.
- However, the computing cost of a few epochs is still not negligible!

Verifier's Incentive?

- If a mechanism is "well-designed", then provers are prevented or disincentivized from cheating.
- Verifiers are rewarded for catching cheats.

But…

- If no/little provers actually cheat, why would verifiers verify, instead of lazily report "verification passed"?

- Verifier's Dilemma:
  For binary-report verification games with positive verification costs, it is impossible to achieve an honest pure-strategy Nash equilibrium.

Verifier’s Dilemma: A Non-Binary Escape

- Verifier’s Dilemma occurs only for binary verification.
- Why?

---If I only need to tell if it is right or wrong...
I just say it is right.

---But what if I have to tell how it is right?

e.g.

- *“The epoch is trained via SGD with a random seed in $\{\varphi_0, \varphi_1, \varphi_2\}$. Tell me which one it is.”*
- *“The model classifies k objects correctly among the test dataset. Tell me whether k is odd or even.”*

“Attention Challenges”, ≈ “Proof of Verification”!

Attempted Solution: Capture-The-Flag

- Existing works (e.g. Truebit): inject additional information (“flags”, non-binary verification) and reward detection of flags.
- Can only prevent lazy behavior, but what about “liars”?
  - If the verification result is also expensive to verify…
  - We need higher-level verifiers to verify the results.
  - How many layers of verifiers do we need?

Intuition: Decentralized Verification Game

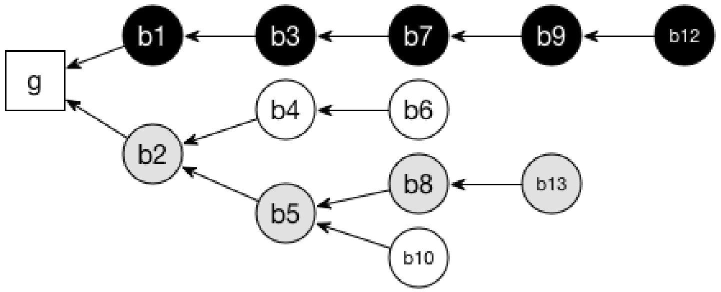
- Essential takeaway: *verifiers also need to be verified*.

---Can we just put them in equal positions to verify each other?

- Close to our solution!

Blockchain Consensus Revisited

- Nakamoto Longest-Chain Rule:
  miners follow longest honest chain when forking (disagreement) happens.



- Economic incentive:
  miners gets the block reward iff they are on the main (longest) chain.

- The system cannot decide if a block is honest,
  but does intend to incentivize honest actions.
- Miners' rewards dependent (solely) on all miners' actions.
- How can the system incentivize honesty **without being able to judge it**?

Peer Prediction: Toy Example

- An unfair coin, head probability $\theta \in \{0.2, 0.8\}$.
  Prior: $P(\theta = 0.2) = P(\theta = 0.8) = 0.5$.
- Alice and Bob independently toss it and are asked to (secretly) report results.


- Mechanism: both rewarded \$1 iff their reports agree.
- Bayes-Nash equilibrium.

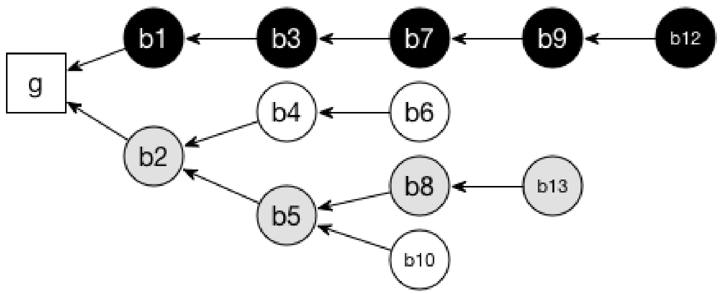
Peer Prediction: Toy Example (cont'd)

How this mechanism works?

- Suppose Alice gets a head and believes Bob will be honest.

- *Since I see a head, θ is probably 0.8*
- *If θ=0.8, then Bob probably sees a head too.*
- *So I should report "head" for better chances.*

Peer Prediction: Nakamoto Consensus



- *I see this chain to be honest and longest so far… Other miners would probably also think so.*
- *I get the block reward iff I'm on the longest chain…*
- *So I will follow this chain.*

Peer Prediction: Concept

- Peer Prediction: information elicitation mechanisms incentivizing truthful report without access to ground truth.

- What is the blockchain?

  A (probably) fanciest application of (implicit) peer prediction!

Peer Prediction: General Idea

- **Peer Prediction**:
  Predict what your peer would do and make decisions accordingly.
- General guideline:
  - Known prior $P(\theta)$ and conditional distribution $P(X_i|\theta)$
  - Compute marginal probability $P(X_i) = \sum_\theta P(\theta)P(X_i|\theta)$
  - Compute posterior belief of ground truth: $P(\theta|X_i) = \frac{P(\theta)P(X_i|\theta)}{P(X_i)}$
  - If X_i and X_{-i} independent given θ, then it can be computed that

$$P(X_{-i}|X_i) = \frac{\sum_\theta P(\theta)P(X_i|\theta)P(X_{-i}|\theta)}{\sum_\theta P(\theta)P(X_i|\theta)}$$

Traditional Peer Prediction without Flags

- PP in one sentence: compute $P(X_{-i}|X_i)$ from $P(\theta)$ and $P(X_i|\theta)$.
- Toy model (2 verifiers X, Y):
  - If the block is honest ($\theta = 0$), the observation is always honest ("$-$").
  - If the block is dishonest ($\theta = 1$), it is caught ("$+$") with probability $\frac{1}{2}$.
  - Prior of the block: highly likely to be honest ($P(\theta = 1) = \varepsilon$)

$P(X\|\theta)$	$X = "-"$	$X = "+"$
$\theta = 0$	1	0
$\theta = 1$	$\frac{1}{2}$	$\frac{1}{2}$

$P(Y\|X)$	$Y = "-"$	$Y = "+"$
$X = "-"$	$1 - \frac{\varepsilon}{4 - 2\varepsilon}$	$\frac{\varepsilon}{4 - 2\varepsilon}$
$X = "+"$	$\frac{1}{2}$	$\frac{1}{2}$

Traditional Peer Prediction: Log Scoring Rule (1)

- Log scoring rule:

$R(X,Y)$	$Y = "-"$	$Y = "+"$
$X = "-"$	$\log(1 - \frac{\varepsilon}{4-2\varepsilon})$	$\log \frac{\varepsilon}{4-2\varepsilon}$
$X = "+"$	$\log \frac{1}{2}$	$\log \frac{1}{2}$

- All negative, need scaling for our design!

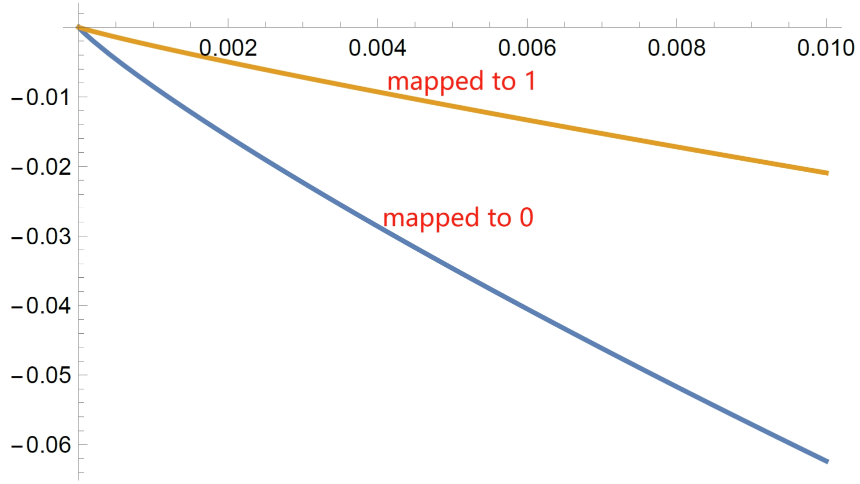
$R_X(X,Y)$	$Y = "-"$	$Y = "+"$
$X = "-"$	$k \cdot \log\left(1 - \frac{\varepsilon}{4-2\varepsilon}\right) + b$	$k \cdot \log \frac{\varepsilon}{4-2\varepsilon} + b$
$X = "+"$	$k \cdot \log \frac{1}{2} + b$	$k \cdot \log \frac{1}{2} + b$

Traditional Peer Prediction: Log Scoring Rule (2)

- What we additionally desire for a scoring rule?
  - Uninformed no-free-lunch: always reporting $-$ or $+$ (or mixed) gains non-positive utility.
  - Ex-ante (weakest) individual rationality: (committing to) truthful reporting gains non-negative expected utility.
- For simplicity we assume verification cost is 1.

$$(1-\varepsilon)R(-,-)+\varepsilon R(-,+)\leq 0$$
$$(1-\varepsilon)R(+,-)+\varepsilon R(+,+)\leq 0$$
$$\left(1-\frac{3}{4}\varepsilon\right)R(-,-)+\frac{\varepsilon}{4}\left(R(-,+)+R(+,-)+R(+,+)\right)\geq 1$$

Traditional Peer Prediction: Log Scoring Rule (3)



- We get $k \geq \Omega(\frac{1}{\varepsilon})$, payment rule highly sensitive to small ε
- Not desirable as ε is neither known nor easy to accurately estimate especially when small.

DMI-based DSIC Peer Prediction

- Kong (2023): assuming tasks are i.i.d. (ε is the same for all blocks), there exists a DSIC 4-task ($2C, C = 2$ is the number of choices) prior-free peer prediction mechanism.
- 4-task is not difficult for blockchain verification (just ask a verifier to verify 4 blocks)
- But… DMI mechanism is not permutation-proof!
  - If saying "pass" when failing and "fail" when passing…
  - Genuinely malicious, but still getting good rewards!

Peer Prediction and Information Theory

- Data processing inequality: strategic processing cannot increase (mutual) information.
- Information-theoretical mechanisms: expected reward is based on informativeness.
  - Log scoring rule;
  - PMI/DMI mechanisms;
  - Etc.
- Take care of no-information-loss transformations. (e.g. permutation)

Our Work: Capture-The-Flag (CTF) Peer Prediction

- What would ε be when all provers are honest? 0.
- Is it possible to design a peer prediction mechanism that works robustly for infinitesimal ε?
- Maybe we want a mechanism that…
  - Has a fixed payment rule and works uniformly for any $\varepsilon \in [0, \varepsilon_0)$.
- How to work even for $\varepsilon = 0$?
  - Insert flags, like existing works…

CTF Peer Prediction: System Model

- For any block, it can be classified as
  - Honest ($\theta = 0$), with probability $1 - \varepsilon - \sum_i \alpha_i$;
  - Flagged with the i-th flag ($\theta = F_i$), with probability α_i;
  - Dishonest ($\theta = 1$), with probability $\varepsilon \ll 1$.
- Lossy-channel model:
  - An honest block is always observed as honest ($X = 0$);
  - The flag i can be detected ($X = F_i$) with probability p_i, otherwise observed as honest;
  - A dishonest block can be caught ($X = 1$) with probability κ, otherwise observed in any known distribution.
- $\{\alpha_i\}$, $\{p_i\}$, κ are fixed and known, from systematic design.
- Intuition: incentivize verifiers to distinguish flag types, even if dishonest blocks can be arbitrarily scarce.

CTF Peer Prediction: Verifiers' Actions

- Nature secretly chooses $\theta \sim P(\theta)$.
- Every verifier i independently chooses to be active or lazy.
  - If active, she performs the verification and observes $X_i \sim P(X_i | \theta)$, taking a computational cost of $c(X_i)$.
  - If lazy, she observes $X_i = \perp$ at no cost, i.e., $c(\perp) = 0$.
- From her observation, verifier i updates her belief of X_{-i} to be $P(X_{-i} | X_i)$, in which $P(X_{-i} | \perp) = P(X_{-i})$.
- She reports Z_i that maximizes $\sum_x R(Z_i, x) P(X_{-i} = x | X_i)$.

CTF Peer Prediction: Toy Example

- $\alpha_1 = \alpha_2 = 1/3, p_1 = p_2 = p_+ = 3/4$, assuming ε=0.

	Observation			
$\theta = 0$	0	0	0	0
$\theta = F_1$	0	F_1	F_1	F_1
$\theta = F_2$	0	F_2	F_2	F_2

$P(Y\|X)$	$Y = 0$	$Y = F_1$	$Y = F_2$
$X = 0$	$3/4$	$1/8$	$1/8$
$X = F_1$	$1/4$	$3/4$	0
$X = F_2$	$1/4$	0	$3/4$

- Simple agreement scoring rule:

$$R_X(X,Y) = \begin{cases} +r, X = Y \\ -r, X \neq Y \end{cases} \quad , r \geq 2\,.$$

- NFL, Interim IC, Interim IR for ε=0.
- For any $r > 2$, works uniformly for $\varepsilon \leq \varepsilon(r), \varepsilon(r) > 0$.

CTF Peer Prediction: Toy Example (ε>0)

- As long as ε is small enough

$P(Y\|X)$	$Y = 0$	$Y = F_1$	$Y = F_2$	$Y = 1$
$X = 0$	$\approx 3/4$	$\approx 1/8$	$\approx 1/8$	$O(\varepsilon)$
$X = F_1$	$1/4$	$3/4$	0	0
$X = F_2$	$1/4$	0	$3/4$	0
$X = 1$	$1/4$	0	0	$3/4$

- The same scoring rule still works!

CTF Peer Prediction: Versus Traditional

- Freedom in participation. Given the others report truthfully,
  - NFL: uninformed parties (e.g. always reporting one signal) get ≤ 0 expected reward.
    - lazy participation should not be profitable.
  - Interim IR: given observing any signal X_i, reporting it gets $\geq c(X_i)$ expected reward.
    - verifiers should be willing to verify and report.

Value of computation

- Robustness
  - Works uniformly for any small ε.
  - Small |payments| in scoring rule.

CTF Peer Prediction: Theoretical Guarantees

- Main Theorem:
  For any non-degenerate 2-party DVG and some $\epsilon > 0$, there exists a CTF-PP mechanism satisfying all the required properties for any $P(\theta = 1) \leq \varepsilon$

- How to find the mechanism?
- Linear Programming!

CTF Peer Prediction: LP Modeling

- Belief matrix $B_{xy} = P(X_{-i} = y | X_i = x)$.
- Scoring matrix R_{xy}: reward to i when $(i, -i)$ report (x, y)
- Let $W = BR'$, then W_{xy} is the expected reward to i when she observes x and reports y.
- We want W to have large diagonals and small off-diagonal entries.
- When B is invertible, then $R = (B^{-1}W)'$
  We can compute a R from any W.

CTF Peer Prediction: LP Construction

- Construction of W:

	0	F_1	F_2	1
0	$+$	$-$	$-$	$-$
F_1	$-$	$+$	$-$	$-$
F_2	$-$	$-$	$+$	$-$
1	$-$	$-$	$-$	$+$

- What about uninformed (lazy) strategies?
  - Reward (row vector) is convex combination of the rows.
  - Let " $-$ " have significantly larger magnitude than " $+$ ".

CTF Peer Prediction: LP Construction (cont'd)

- Construction of W:

	0	F_1	F_2	1
0	$+100$	-1000	-1000	-1000
F_1	-1000	$+100$	-1000	-1000
F_2	-1000	-1000	$+100$	-1000
1	-1000	-1000	-1000	$+100$

- It is a feasible solution.
- The LP is feasible.
- Our solution works for all non-degenerate (B invertible) cases.
- But the ex-post reward/penalty can be extremely high...
  - Mining? Gambling!

CTF Peer Prediction: Optimization

How to define a “good” scoring rule?

- Satisfying incentive guarantees with small ex-post reward/penalty.

$$
\begin{aligned}
&\text{minimize} && M \\
s.t.\quad &\text{honest net utility} &\geq\ & \delta \\
&\text{dishonest net utility} &\leq\ & -\delta \\
&|R| &\leq\ & M
\end{aligned}
$$

- δ margin guarantees incentive properties for small $\epsilon > 0$.

CTF Peer Prediction: Experiments

- Setting: verification of Incentive-Secure PoL
- CTF Protocol: a dishonest stage might be observed as a flag.
- Reward matrix R:

	0	F_1	F_2	1
0	$+2.10$	-7.16	-7.16	-1.08
F_1	-1.54	$+6.47$	-4.45	-1.24
F_2	-1.54	-4.45	$+6.47$	-1.24
1	-2.20	$+5.80$	$+5.80$	$+7.40$

CTF Peer Prediction: Experiments (cont'd)

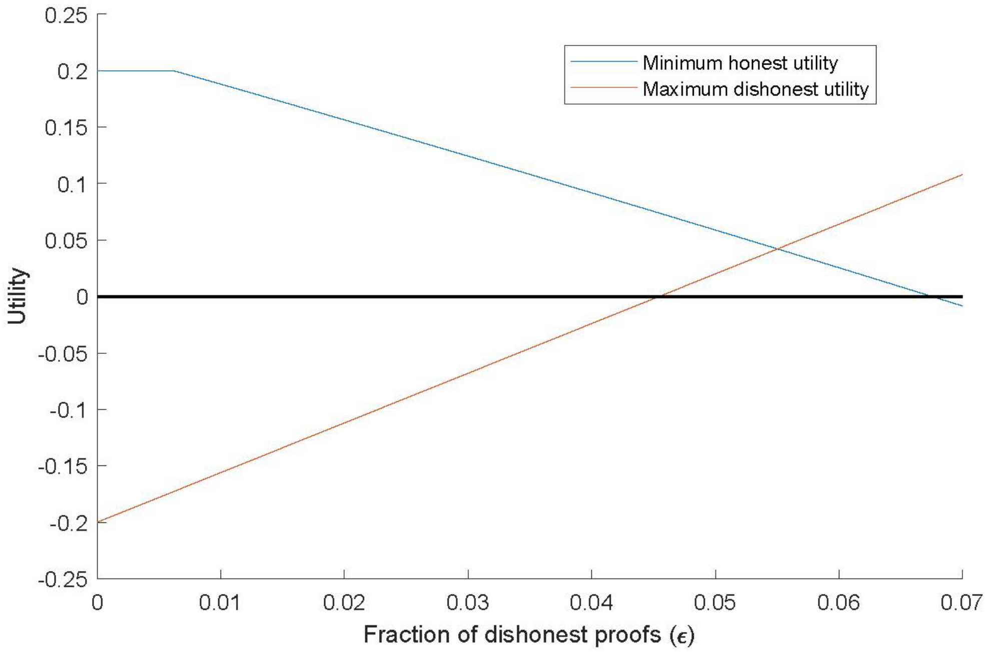
- Utility matrix $(W - c)$, $\varepsilon = 0$ (margin $\delta = 0.2$):

	0	F_1	F_2	1
0	$+0.22$	-1.45	-1.45	-1.20
F_1	-4.53	$+0.47$	-5.00	-0.20
F_2	-4.53	-5.00	$+0.47$	-0.20
1	-3.01	-2.83	-2.83	$+0.20$
Lazy	-0.22	-0.90	-0.90	-0.20

- Gaining positive expected utility iff honest.

CTF Peer Prediction: Robustness

- When $\epsilon > 0$:



- Works robustly when $\varepsilon < 0.045$.

Discussion: Future Work

- General case of n-party DVG
  - Current method: LP of size $\Omega(3^n)$,
    inefficient when n is (even slightly) large, e.g. $n \approx 10$.
  - TODO: poly-time algorithm for good scoring rules.
  - (Information-theoretical approaches may work?)
- Collusion-proof / sybil-proof mechanism for DVG
  - Intuition: 2-CP for large n is not difficult.
  - SP almost equivalent to CP.
  - Is $\Theta(n)$-CP possible? (e.g. comparable to $n/3$?)
- Will multi-task peer prediction mechanisms do better?

Discussion: Other Applications in AI

- Manipulation-proof data elicitation & valuation
  - Reward data providers for the mutual information between their data and others'.

    Truthful Data Acquisition via Peer Prediction, NeurIPS'20
    Yiling Chen, Yiheng Shen, Shuran Zheng

- Feedback acquisition for AI generated contents
  - Elicit comparison data from user feedback to improve the quality of AI performance.

    Carrot and Stick: Eliciting Comparison Data and Beyond
    Yiling Chen, Shi Feng, Fang-Yi Yu

Conclusion: Peer Prediction x Decentralized AI

- Resources of AI: data & computation
- Decentralization: crowdsourcing w/o centralized control
- Peer Prediction: a (meta-)methodology to incentivize honesty (incl. data & computation) in a decentralized environment
- Blockchain: a decentralized trustworthy platform driven by cryptography & economic incentives

Blockchain-based decentralized trustworthy AI: a starry-eyed dream?

Challenges and Thoughts of Blockchain & AI

- AI: “model collapse” of LLM
  - When AI is trained by AI-generated data, garbage in garbage out
  - Would advanced data valuation methods work?
- Blockchain: the rich may take all?
  - Money can buy a lot of things, including computing power…
  - Would “something between” permissionless & permissioned chains work?

Meta-Conclusion

所有的转折隐藏在密集的鸟群中
天空与海洋都无法察觉
怀着美梦却可以看见
摸索颠倒的一瞬间

"Even if you cannot see the world clearly
There is still a way to follow your mind."

Q&A