# Incentivizing Truthful Data Quality for Quality-Aware Mobile Data Crowdsourcing

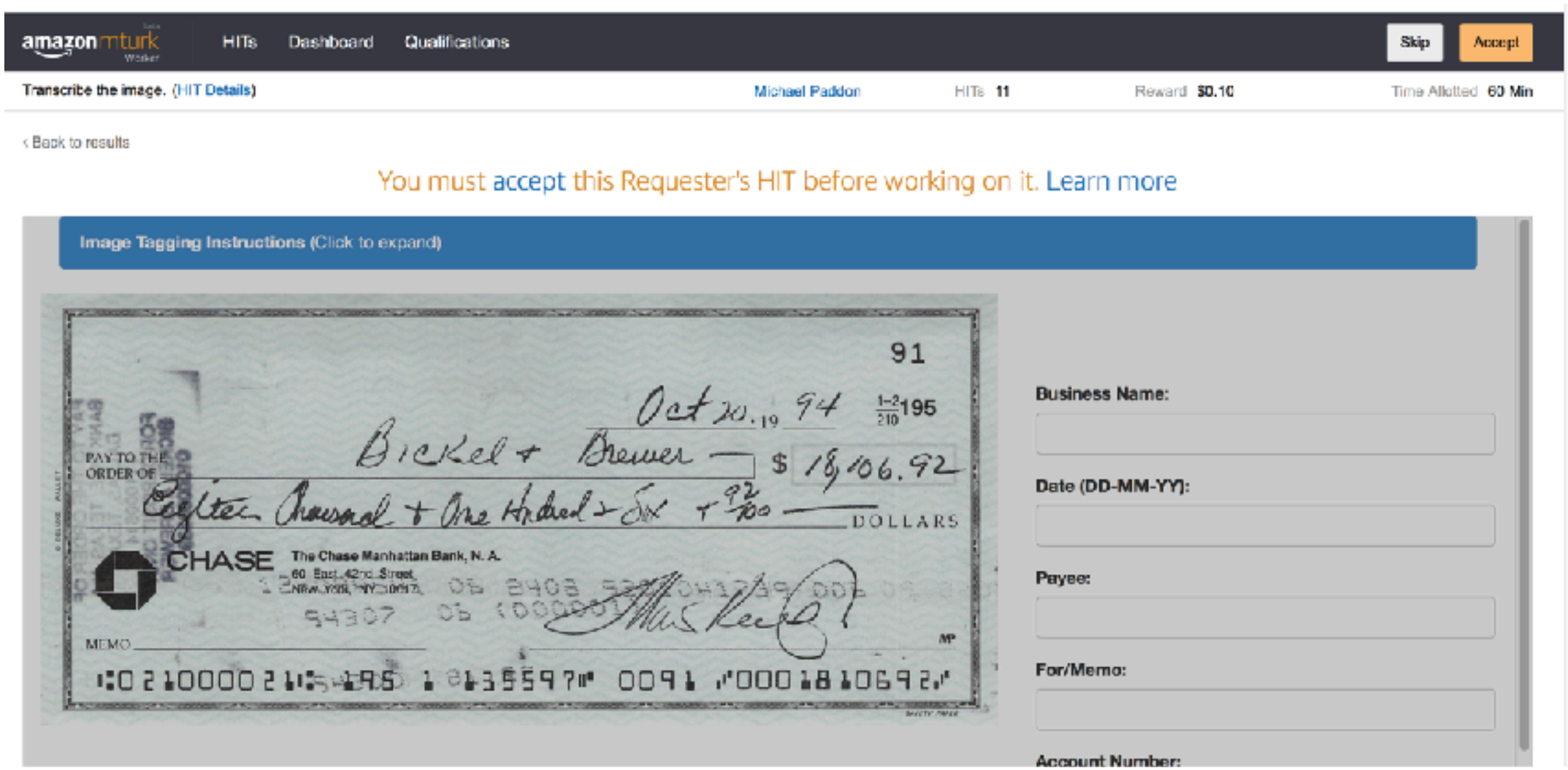
Xiaowen Gong, Ness Shroff

Department of Electrical and Computer Engineering
Auburn University
Department of Electrical and Computer Engineering &
Department of Computer Science and Engineering
The Ohio State University

MobiHoc'18, Los Angeles, CA
June 26th, 2018

# Data Crowdsourcing

- Data crowdsourcing: leverage "wisdom" of a crowd of users by collecting their data
  - ✓ Enabled by powerful mobile devices and pervasive connectivity
  - ✓ Wide range of applications: physical sensing (environmental monitoring, spectrum sensing…), human intelligence (image classification, text transcribing…)
  - ✓ Provide enormous potential via machine learning/data mining tools




Amazon mturk (image recognition)

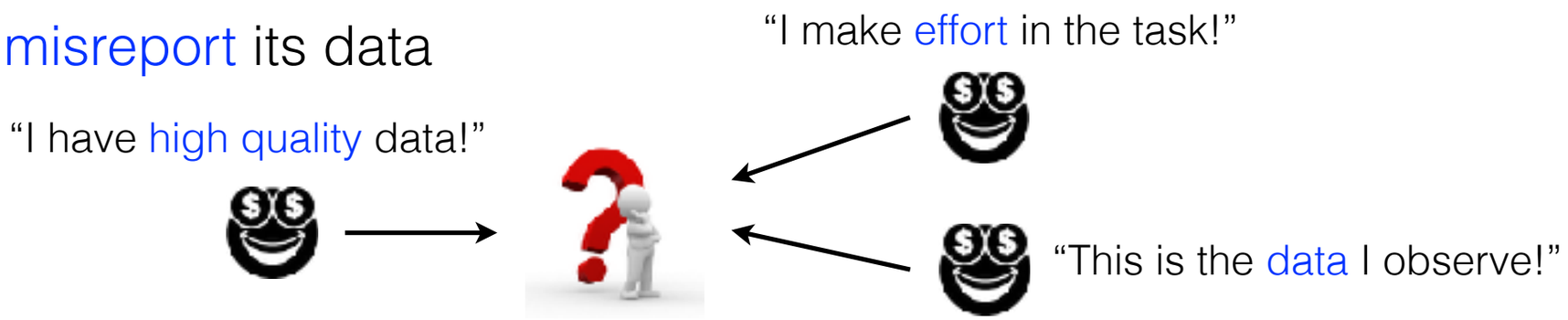
# Quality-Aware Crowdsourcing

- Data quality captures how accurate is a user's data compared to ground truth
  - ✓ Data accuracy is a key performance metric for crowdsourcing
    - Data observed by individual users are inherently inaccurate due to noise, interference, error…
    - Exploit diversity gain by aggregating data from the crowd to improve data accuracy
  - ✓ E.g., to measure transmit signal strength of a transmitter, the received SNR determines data quality
  - ✓ Users have diverse data quality depending on their specific situations (e.g., location)

- Information of users' data quality is important for the crowdsourcing requester
  - ✓ Assign more work to users with high quality data
  - ✓ Assign larger weights to high quality data in data aggregation
  - ✓ Know the accuracy of aggregated data

# Challenges in Quality-Aware Crowdsourcing

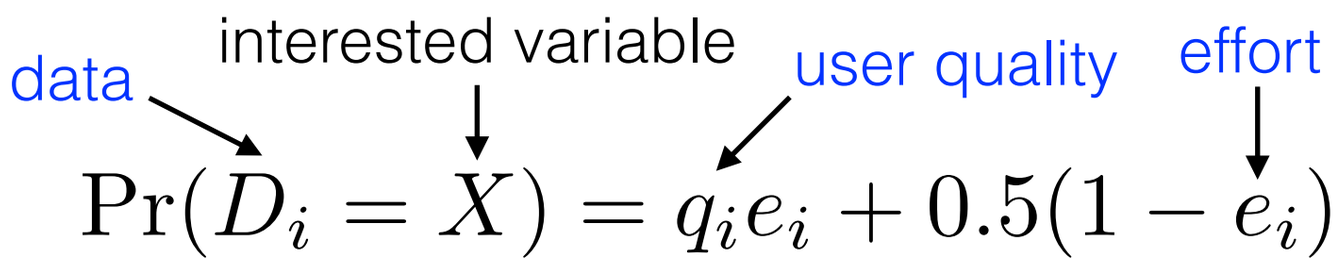
- A user's data quality can be its private information
  - ✓ A user can learn its data quality based on its relevant private information (e.g., location)
  - ✓ Cannot be known or verified by the requester
  - ✓ A user may have incentive to manipulate its data quality revealed to the requester
- A user's effort exerted in the crowdsourcing task can be its hidden action
  - ✓ Effort captures how much work is devoted to the task
  - ✓ A user can control its effort which affects its data quality and its cost in the task
  - ✓ Cannot be known or verified by the requester
  - ✓ A user may have incentive to manipulate its effort made in the task
- A user's data can be its private information
  - ✓ A user may misreport its data




How to incentivize users to behave truthfully?

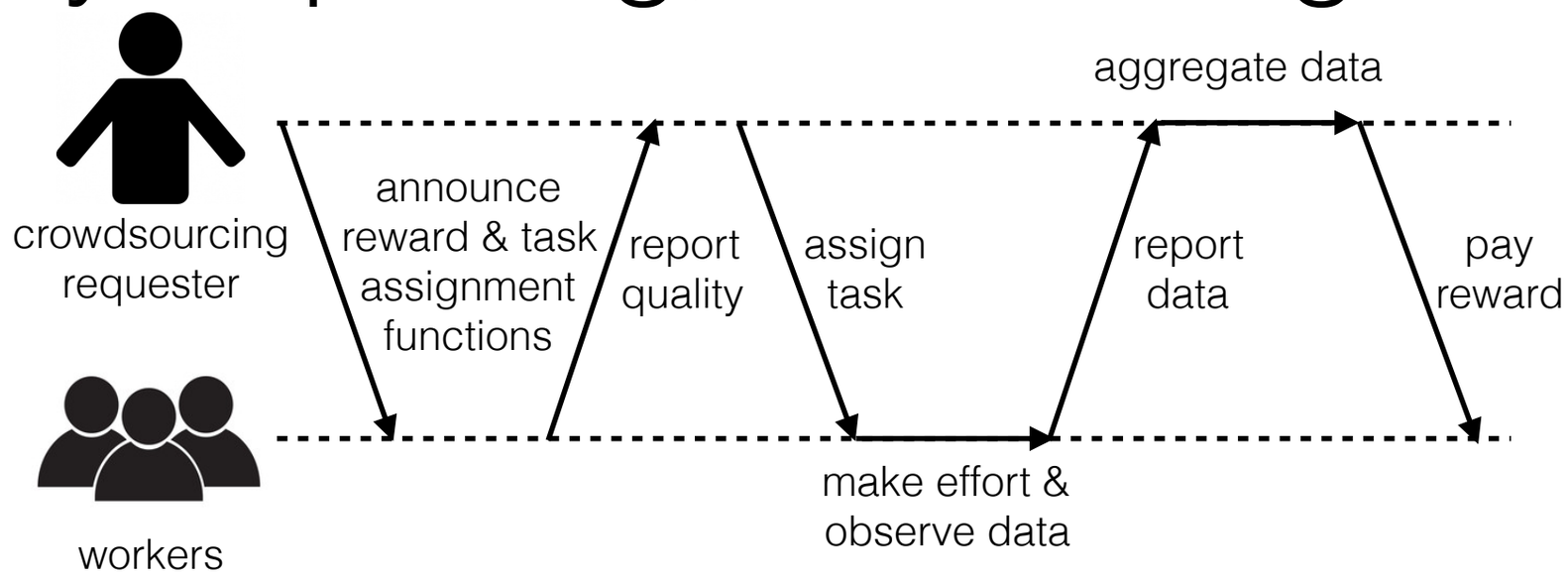
# Data Quality, User Quality and Effort

- Most human intelligence tasks and some physical sensing tasks have discrete data
  - ✓ e.g., image classification, sensing for spectrum occupancy




- Data quality is quantified by correct probability
- User quality $q_i$ quantifies how accurate is a user's data given its effort
  - ✓ Capture a user's intrinsic capability for the task
  - ✓ Private information of the user, unknown to the requester
- Effort $e_i \in \{0, 1\}$ indicates whether a user makes effort or not in the task
  - ✓ Hidden action of the user, unknown to the requester

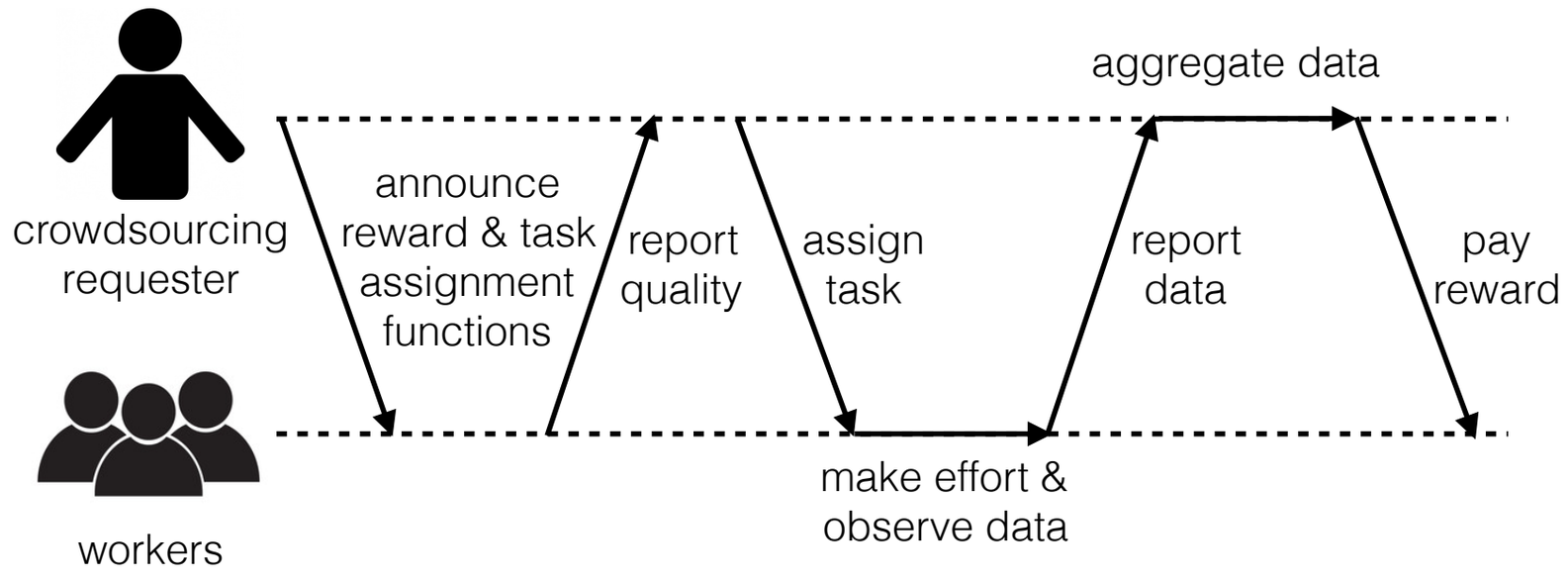
# Quality Reporting, Effort Assignment




- Each user reports its quality $q_i'$ to the requester
  - ✓ May have incentive to misreport $q_i' \neq q_i$
- The requester assigns a desired effort $e_i'(\boldsymbol{q}')$ to each user based on the reported quality
  - ✓ The requester may have desired effort assignments (e.g., to maximize social welfare)
  - ✓ A user may have incentive to not make the desired effort $e_i \neq e_i'$
  - ✓ Effort assignment function is pre-defined and announced to users first
- Each user observes data and reports $d_i'$ to the requester
  - ✓ May misreport $d_i' = \overline{d}_i \neq d_i$ ( $\overline{d}_i$ is complementary of $d_i$ )

$$\Pr(\overline{D}_i = X) = 1 - \Pr(D_i = X)$$

# Data Estimation, Reward Payment




- The requester obtains an estimate based on the reported quality and data, and assigned efforts

$$x_0(\boldsymbol{q}', \boldsymbol{e}', \boldsymbol{d}') \triangleq \arg \max_{d \in \{0,1\}} E_{X|\boldsymbol{d}'(\boldsymbol{q}', \boldsymbol{e}')}[\mathbf{1}_{X=d}]$$

- The requester's utility is the estimation loss

$$p_c(\boldsymbol{q}', \boldsymbol{e}', \boldsymbol{d}') \triangleq E_{X|\boldsymbol{d}(\boldsymbol{q}, \boldsymbol{e})}\left[\mathbf{1}_{X=x_0(\boldsymbol{q}', \boldsymbol{e}', \boldsymbol{d}')}\right]$$

- The requester pays a reward $r_i(\boldsymbol{q}', e_i', d_i', d_0)$ to each user based on its own observed data $d_0$, the reported quality and data, and assigned effort
  - ✓ Reward function is pre-defined and announced to users first
    - Can only depend on information known to the requester
    - Ground truth of interested variable $X$ is unknown

# Mechanism Design Objective

- Each user's payoff $u_i(\boldsymbol{q}', e_i, d_i', d_0) \triangleq r_i(\boldsymbol{q}', e_i', d_i', d_0) - c_i e_i$
  - ✓ Cost coefficient $c_i$ quantifies how much resource is consumed
    - Assume that it is known and uniform (can be relaxed to be diverse)
- Requester's payoff $u_0(\boldsymbol{q}', \boldsymbol{e}', \boldsymbol{d}') \triangleq p_c(\boldsymbol{q}', \boldsymbol{e}', \boldsymbol{d}') - \sum_{i \in \mathcal{N}} r_i(\boldsymbol{q}', e_i', d_i', d_0)$
- Dominant incentive compatible (DIC): Given any quality reported by other users, the optimal strategy of each user for maximizing its expected payoff is to truthfully report its quality and data, and make the effort desired by the requester

$$E_{D_0|d_i(q_i,e_i)}\left[u_i(q_i, \boldsymbol{q}'_{-i}, e_i, d_i, D_0)\right] \geq E_{D_0|d_i(q_i,e_i)}\left[u_i(q_i', \boldsymbol{q}'_{-i}, e_i', d_i', D_0)\right], \forall (q_i', e_i, d_i'), \forall \boldsymbol{q}'_{-i}$$

  - ✓ If users are not truthful, the actual effort assignment is not desired, data estimation is not optimal, and its information of data accuracy is incorrect!!
  - ✓ Dominant is strong and desirable
- Individual rational (IR): Given that a user truthfully reports its quality and makes the effort desired by the requester, its expected payoff is nonnegative
  - ✓ Each user's reward can compensate its cost

# Quality, Effort, and Data Elicitation Mechanism

Quality, Effort, and Data Elicitation (QEDE) mechanism: A pair of any effort assignment function $e'_i(\boldsymbol{q}')$ satisfying

$$e'_i(q'_i, \boldsymbol{q}'_{-i}) \geq e'_i(q''_i, \boldsymbol{q}'_{-i}),\ \forall q'_i \leq q''_i,\ \forall \boldsymbol{q}'_{-i}$$

and a reward function based on that $e'_i(\boldsymbol{q}')$ given by

$$r_i(d_0, d'_i, \boldsymbol{q}', e'_i) = kq'_i e'_i(\boldsymbol{q}') \left[\frac{\mathbf{1}_{d_0=d'_i} + q_0 - 1}{2q_0 - 1}\right] + ce'_i(\boldsymbol{q}') + \int_{\underline{q}}^{q'_i} kqe'_i(q, \boldsymbol{q}'_{-i})dq - kq'_i e'_i(\boldsymbol{q}')\left[0.5 + (q'_i - 0.5)e'_i(\boldsymbol{q}')\right]$$

where $k \geq \dfrac{c}{\underline{q}(\underline{q} - 0.5)}$

- ✓ The requester has quality $q_0$ and data $d_0$, and makes effort $e_0$
- ✓ The condition on the effort assignment function: assigns more effort for higher quality
  - General, natural (e.g. satisfied for maximizing social welfare)

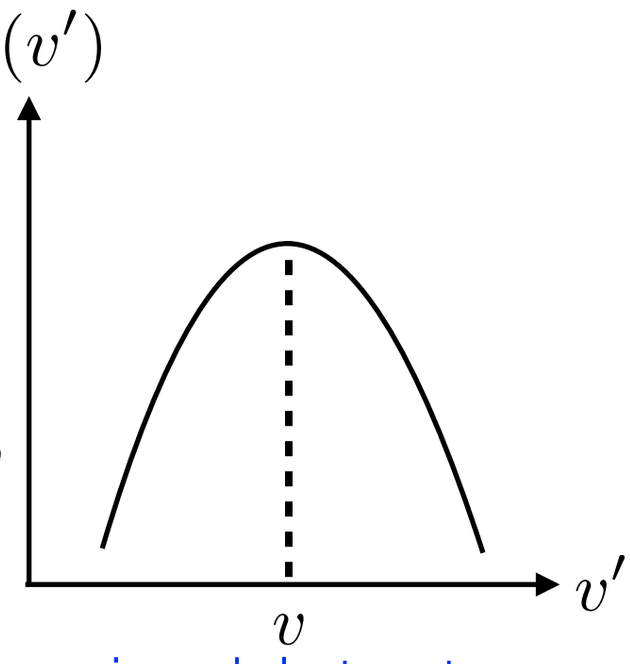
# Truthful Single Parameter Mechanism

- ✓ A bidder has a private scalar value $v$ for a divisible resource
- ✓ A fraction of the resource $a(v')$ is allocated based on the reported value $v'$
- ✓ The bidder pays $p(v')$ for the allocated resource
- ✓ The bidder's payoff

$$u(v, v') = va(v') - p(v')$$

If $\frac{\partial u}{\partial v'}\big|_{v'=v} = 0, \ \frac{\partial u}{\partial v'}\big|_{v'<v} > 0, \ \frac{\partial u}{\partial v'}\big|_{v'>v} < 0, \ \forall v$

then $\arg\max_{v'} u(v, v') = v$



unimodal structure

Truthful single parameter mechanism: A pair of any allocation function $a(v')$ that is increasing in $v'$ and a payment function based on that $a(v')$

$$p(v') = v'a(v') - \int_{\underline{v}}^{v'} a(x)dx$$

# Design of QEDE Mechanisms

- ✓ The user's payoff should depend on true quality $q_i$ and actual effort $e_i$

$$r_i(d_0, d_i', \boldsymbol{q}', e_i') = kq_i'e_i'(\boldsymbol{q}')\left[\frac{\mathbf{1}_{d_0=d_i'} + q_0 - 1}{2q_0 - 1}\right] + ce_i'(\boldsymbol{q}') + \int_{\underline{q}}^{q_i'} kqe_i'(q, \boldsymbol{q}'_{-i})dq - kq_i'e_i'(\boldsymbol{q}')\left[0.5 + (q_i' - 0.5)e_i'(\boldsymbol{q}')\right]$$

$$= 0.5 + (q_i - 0.5)e_i$$

- ✓ Similar to peer prediction method [Miller'05]

Lemma: Under the QEDE mechanisms, given that any user reports any quality $q_i'$ and makes any effort $e_i'$, it's optimal to report is true data $d_i' = d_i$, with its expected payoff:

$$E_{D_0|d_i(q_i,e_i)}[u_i(\boldsymbol{q}', e_i', d_i, D_0)] = kq_i'e_i'(\boldsymbol{q}')\left[0.5 + (q_i - 0.5)e_i\right] + \int_{q}^{q_i'} kqe_i'(q, \boldsymbol{q}'_{-i})dq - kq_i'e_i'(\boldsymbol{q}')\left[0.5 + (q_i' - 0.5)e_i'(\boldsymbol{q}')\right] + ce_i'(\boldsymbol{q}') - ce_i.$$

# Design of QEDE Mechanisms

$$r_i(d_0, d_i', \boldsymbol{q}', e_i') = kq_i'e_i'(\boldsymbol{q}') \left[ \frac{\mathbf{1}_{d_0=d_i'} + q_0 - 1}{2q_0 - 1} \right] + ce_i'(\boldsymbol{q}') + \int_{\underline{q}}^{q_i'} kqe_i'(q, \boldsymbol{q}'_{-i})dq - kq_i'e_i'(\boldsymbol{q}') \left[0.5 + (q_i' - 0.5)e_i'(\boldsymbol{q}')\right]$$

$$k \geq \frac{c}{\underline{q}(\underline{q} - 0.5)}$$

Lemma: Under the QEDE mechanisms, given that any user reports any quality $q_i'$ and its true data $d_i$, it's optimal to make the desired effort $e_i = e_i'$ .

# Design of QEDE Mechanisms

Plugging in $e_i = e_i'$

$$\hat{u}_i(\boldsymbol{q}', q_i, e_i') = kq_i'e_i'(\boldsymbol{q}')(q_i - q_i') + \int_{\underline{q}}^{q_i'} kqe_i'(q, \boldsymbol{q}'_{-i})dq$$

We can design $\hat{u}_i(\boldsymbol{q}', q_i, e_i')$ such that

$$\frac{\partial \hat{u}_i}{\partial q_i'}|_{q_i'=q_i} = 0,\ \frac{\partial \hat{u}_i}{\partial q_i'}|_{q_i'<q_i} > 0,\ \frac{\partial \hat{u}_i}{\partial q_i'}|_{q_i'>q_i} < 0,\ \forall e_i'(q_i', \boldsymbol{q}'_{-i}) \text{ decreasing in } q_i',\ \forall \boldsymbol{q}'_{-i}$$

$$r_i(d_0, d_i', \boldsymbol{q}', e_i') = kq_i'e_i'(\boldsymbol{q}') \left[\frac{\mathbf{1}_{d_0=d_i'} + q_0 - 1}{2q_0 - 1}\right] + ce_i'(\boldsymbol{q}') + \int_{\underline{q}}^{q_i'} kqe_i'(q, \boldsymbol{q}'_{-i})dq - kq_i'e_i'(\boldsymbol{q}')\left[0.5 + (q_i' - 0.5)e_i'(\boldsymbol{q}')\right]$$

Lemma: Under the QEDE mechanisms, given that any user reports its true data $d_i$ and makes its desired effort $e_i'$ , it's optimal to reported its true quality $q_i' = q_i$.

# Design of QEDE Mechanisms

Theorem: The QEDE mechanisms are dominant incentive compatible and individual rational.

$$\text{optimal } e_i = e_i'$$

$$r_i(d_0, d_i', \boldsymbol{q}', e_i') = kq_i' e_i'(\boldsymbol{q}') \left[\frac{\mathbf{1}_{d_0=d_i} + q_0 - 1}{2q_0 - 1}\right] + ce_i'(\boldsymbol{q}') + \int_{\underline{q}}^{q_i'} kqe_i'(q, \boldsymbol{q}'_{-i})dq - kq_i' e_i'(\boldsymbol{q}') \left[0.5 + (q_i' - 0.5)e_i'(\boldsymbol{q}')\right]$$

$$= 0.5 + (q_i - 0.5)e_i$$

$$\text{optimal } d_i' = d_i$$

$$\text{optimal } q_i' = q_i$$

$$e_i'(q_i', \boldsymbol{q}'_{-i}) \geq e_i'(q_i'', \boldsymbol{q}'_{-i}),\ \forall q_i' \leq q_i'',\ \forall \boldsymbol{q}'_{-i}$$

# No Reference Data from Requester

- Use reference data from other users (peers)

$$r_i(d'_j, d'_i, \boldsymbol{q}', e'_i) = kq'_i e'_i(\boldsymbol{q}') \left[ \frac{\mathbf{1}_{d'_j = d'_i} + q'_j - 1}{2q'_j - 1} \right] + ce'_i(\boldsymbol{q}')$$

$$+ \int_{\underline{q}}^{q'_i} kqe'_i(q, \boldsymbol{q}'_{-i}) dq - kq'_i e'_i(\boldsymbol{q}') \left[ 0.5 + (q'_i - 0.5) e'_i(\boldsymbol{q}') \right]$$

Theorem: The QEDE mechanisms achieve truthful strategies as a Nash equilibrium and is individual rational.

# Properties of QEDE Mechanisms

- Users' expected payoff when it truthfully reports its quality and make the desired effort

$$\sum_{i \in \mathcal{N}} \left( k \int_{\underline{q}}^{q_i} q e_i'(q, \boldsymbol{q}'_{-i}) dq + c e_i'(\boldsymbol{q}') \right)$$

✓ Always nonnegative

- The requester pays "information rent" due to uncertainty about users' quality

✓ Decreases when the lower bound of quality decreases

- Less "information rent" when having less uncertainty

# Optimal Effort Assignment for QEDE Mechanisms

- Social welfare is the crowdsensing utility minus the total cost of all users

$$v(\boldsymbol{q}, \boldsymbol{e}(\boldsymbol{q})) \triangleq E_{\boldsymbol{D}(\boldsymbol{q},\boldsymbol{e})}[p_c(\boldsymbol{q}, \boldsymbol{e}, \boldsymbol{D})] - \sum_{i \in \mathcal{N}} ce_i$$

- The socially optimal (SO) effort assignment maximizes the social welfare among all the QEDE mechanisms

$$\{\boldsymbol{e}^{so}(\boldsymbol{q}), \forall \boldsymbol{q}\} \triangleq \arg \max_{\{\boldsymbol{e}(\boldsymbol{q}), \forall \boldsymbol{q}\}} E_{\boldsymbol{Q}}[v(\boldsymbol{Q}, \boldsymbol{e}(\boldsymbol{q}))].$$

$$\text{s.t. } e_i'(q_i', \boldsymbol{q}'_{-i}) \geq e_i'(q_i'', \boldsymbol{q}'_{-i}),\ \forall q_i' \leq q_i'',\ \forall \boldsymbol{q}'_{-i}$$

- Exhaustive search: Find and compare the top $k$ user workers that have the highest quality, for all $k = 0, 1, \cdots, N$

Proposition: The socially optimal effort assignment is given by the exhaustive search.

- The proof is to show that the output of the exhaustive search satisfies the monotonicity condition

# Optimal Effort Assignment for QEDE Mechanisms

- The requester's optimal (RO) effort assignment maximizes the requester's expected payoff among all the QEDE mechanisms

$$\{\boldsymbol{e}^*(\boldsymbol{q}), \forall \boldsymbol{q}\} \triangleq \arg \max_{\{\boldsymbol{e}(\boldsymbol{q}), \forall \boldsymbol{q}\}} E_{\boldsymbol{D}(\boldsymbol{Q},\boldsymbol{e})}[u_0(\boldsymbol{Q}, \boldsymbol{e}, \boldsymbol{D})]$$

$$\text{s.t. } e_i'(q_i', \boldsymbol{q}'_{-i}) \geq e_i'(q_i'', \boldsymbol{q}'_{-i}), \ \forall q_i' \leq q_i'', \ \forall \boldsymbol{q}'_{-i}$$

Proposition: For the single-worker assignment, when

$$\alpha(q) \triangleq q + kq\frac{F(q) - 1}{f(q)}$$

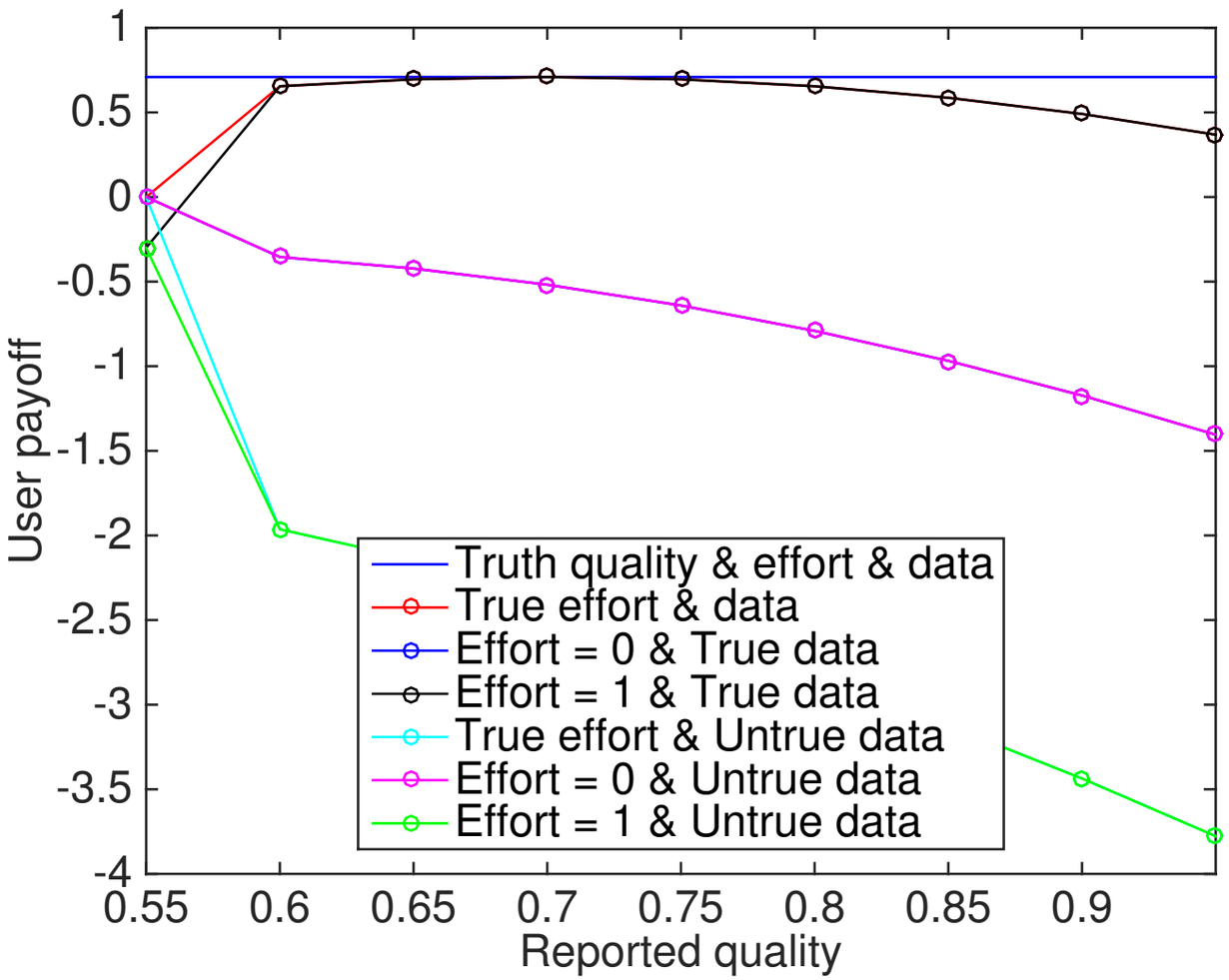
is an increasing function, the requester's optimal effort assignment is given by

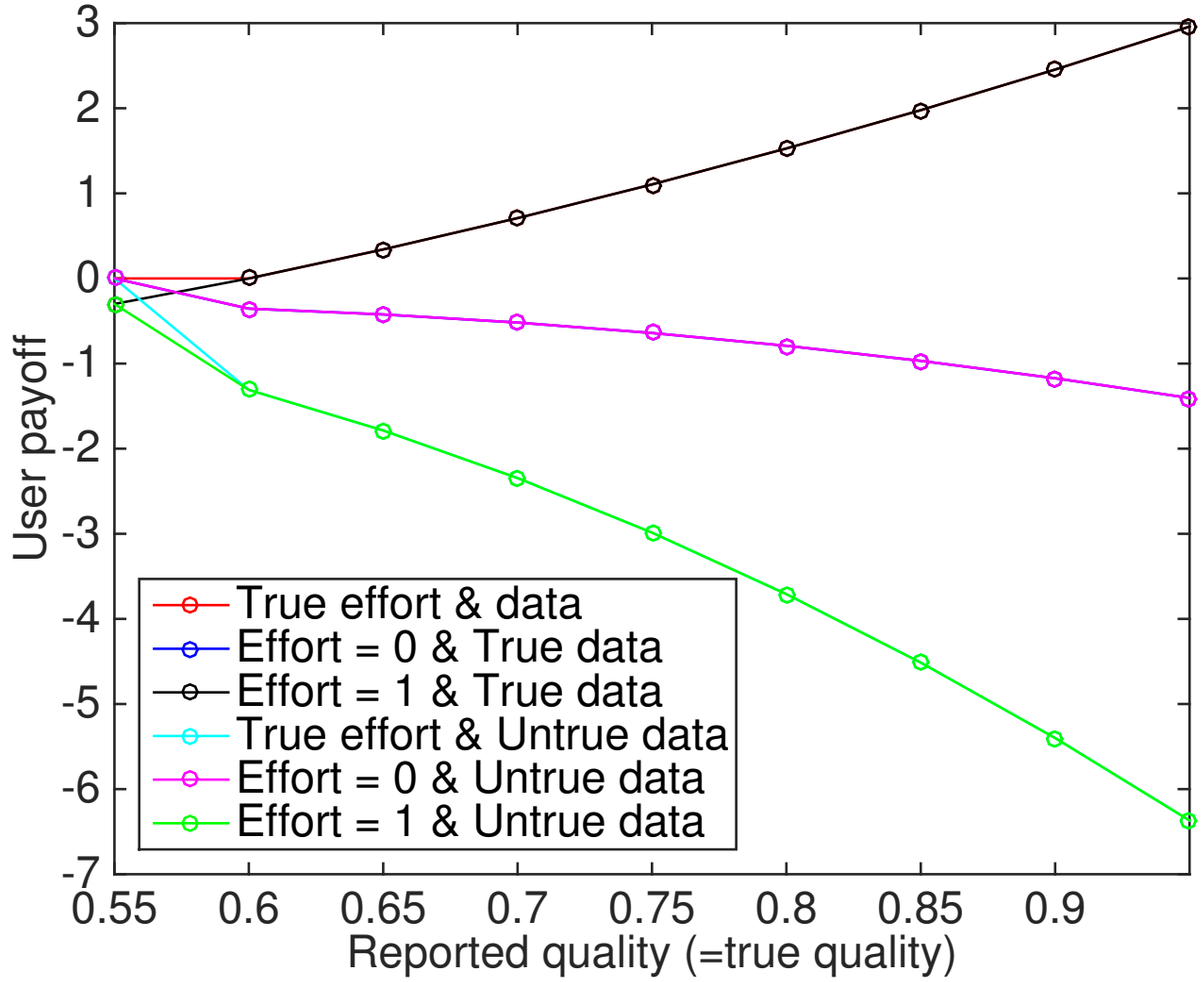
$$e_i^*(\boldsymbol{q}) = \begin{cases} 1, & i = \arg\max_j \alpha(q_j) \ and \ \alpha(q_i) \geq c \\ 0, & otherwise \end{cases}$$

✓ The best user has the smallest "virtual valuation" $\alpha(q_i)$ rather than the highest quality

- Depend on both the quality and the quality distribution
- In the same spirit as the virtual valuation introduced by [Myerson'84]
- The requester's optimal effort assignment is not socially optimal, due to the requester's uncertainty about users' quality

# Impact of Untruthful Behavior




- ✓ Each user's payoff is maximized by truthfully reporting its quality and data and making the desired effort

# Related Work

- Resource allocation in crowdsourcing (e.g., [He, et al, Infocom'14])
- Incentive mechanisms for crowdsourcing
  - ✓ Many works on incentivizing users to truthfully reveal their cost (quality elicitation has not been studied)
  - ✓ Existing solutions for cost elicitation (e.g., VCG auction, the characterization of truthful mechanisms) cannot work for quality elicitation
  - ✓ Few works (e.g., [Luo et al, NetEcon'15]) proposed incentive mechanisms for joint elicitation of effort, data, and/or cost
  - ✓ Joint elicitation of quality, effort and data involves coupling
- Quality-aware crowdsourcing
  - ✓ Incentive mechanisms for users with diverse quality and private cost in crowdsensing [Jin, et al, Mobihoc'15,'16, Infocom'17]
  - ✓ Learning quality of users' data in crowdsourcing [Karger et al, Allerton'11, Liu et al, Sigmetrics'15] for discrete data
  - ✓ Quality and effort elicitation [Gong and Shroff, WiOpt'17] for continuous data
    - Data quality has different structure for discrete data
    - Data elicitation is not considered

# Conclusion

- Summary
  - ✓ Quality-aware crowdsourcing framework that assigns tasks to users based on users' data quality and performs data estimation based on the data quality
  - ✓ Truthful mechanisms that incentive users to truthfully report quality and data, and make desired effort for discrete data
  - ✓ Analysis of optimal effort assignment for the requester's payoff and social welfare

- Highlights
  - ✓ The first study of quality elicitation for crowdsourcing
  - ✓ Untruthful quality reporting and effort exertion lead to undesired effort assignment, undesired data estimation, and incorrect information of data accuracy!!
  - ✓ Overcome coupling between quality elicitation, effort elicitation, and data elicitation in truthful mechanism design

# Future Work

- Quality-aware crowdsourcing
  - ✓ Joint elicitation of quality, effort, data, and cost
  - ✓ Online quality learning with effort and data elicitation [Gong, GameData'18]
  - ✓ Implement quality-aware crowdsourcing system that evaluates and experiments the proposed mechanisms/algorithms in practice
- Privacy control in crowdsourcing
  - ✓ Add random noise to users' data to protect privacy, e.g., differential privacy
  - ✓ Tradeoff between privacy and quality
  - ✓ Mechanisms that incentive users to add random noise to data at desired privacy level [Wang, et al, Sigmetrics'16, Yang, et al, Mobihoc'18]
- Crowdsourcing in machine learning
  - ✓ Crowdsourcing has various applications for machine learning, e.g, localization
  - ✓ Assign different but correlated tasks to users
  - ✓ Effort and data elicitation in crowdsourcing for machine learning [Liu, et al NIPS'16, EC'17]

*Thank you!*

Questions?

# Optimal Effort Assignment for QEE Mechanisms

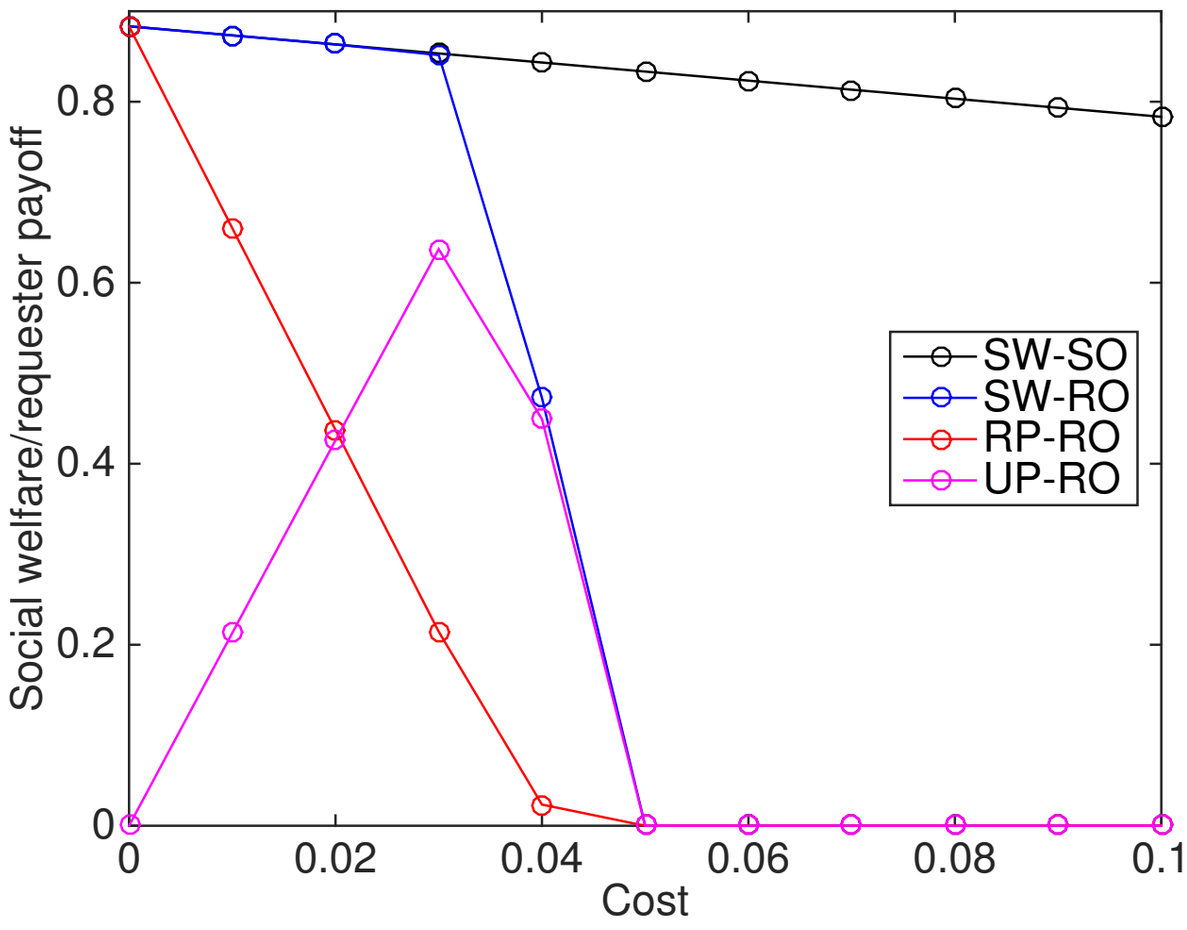
Proposition: The requester's optimal payoff $E_{\boldsymbol{Q}}[u_0(\boldsymbol{e}^*(\boldsymbol{Q}))]$ , the socially optimal social welfare $E_{\boldsymbol{Q}}[v(e_1^{so}(\boldsymbol{Q}))]$ , and the social welfare $E_{\boldsymbol{Q}}[v(e_1^*(\boldsymbol{Q}))]$ attained by the requester's optimal effort assignment increase when the number of users $N$ increase, or the cost coefficient $c$ decreases.

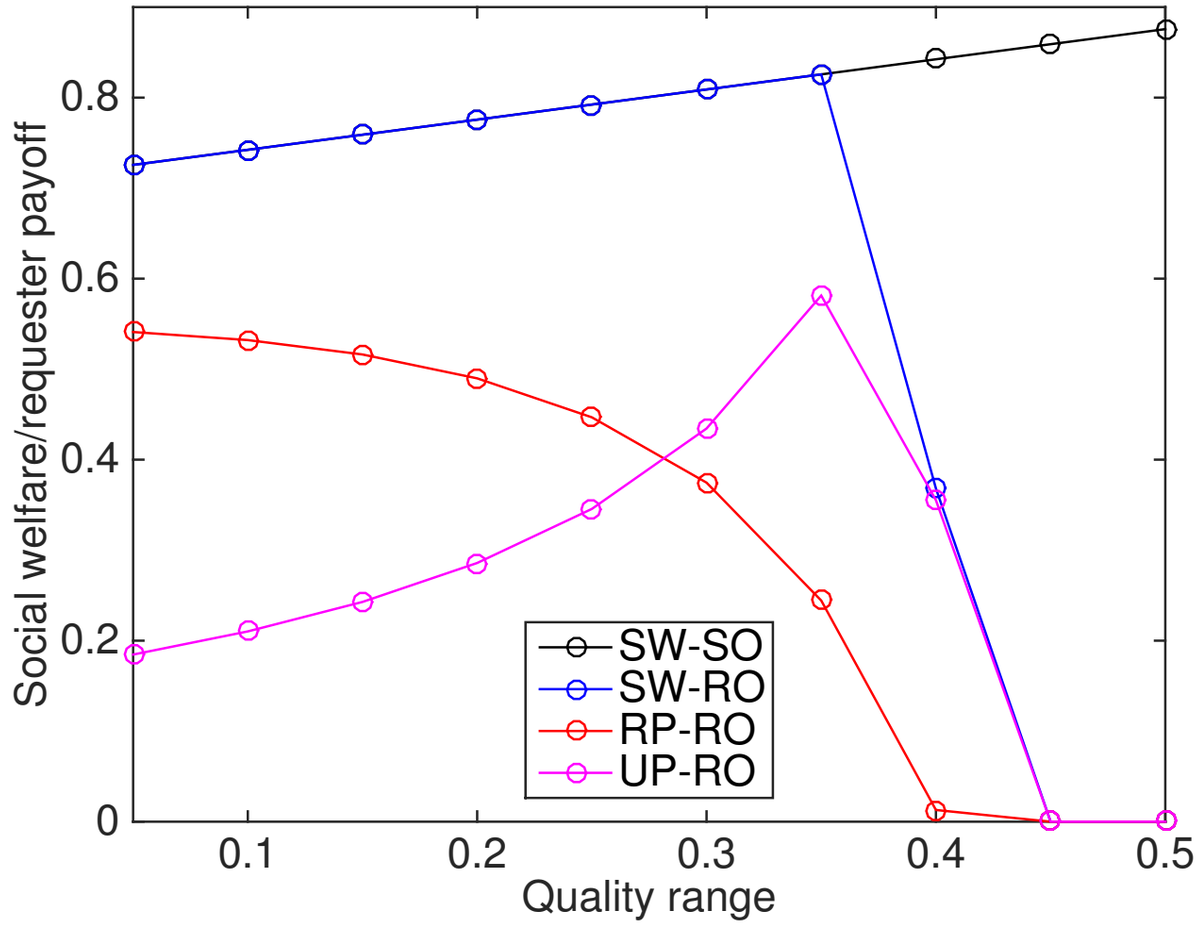
- ✓ The requester's payoff and social welfare both benefit from a greater diversity gain
  - When there are more users, the quality of the best user probably improves

Proposition: The gap between the social welfare attained by the requester's optimal effort assignment and the social optimal effort assignment $E_{\boldsymbol{Q}}[v(e_1^{so}(\boldsymbol{Q}))] - E_{\boldsymbol{Q}}[v(e_1^*(\boldsymbol{Q}))]$ decreases when the number of users $N$ increase, and converges to 0 as $N$ goes to infinity.

- ✓ The performance gap decreases to 0 asymptotically as the number of users increases

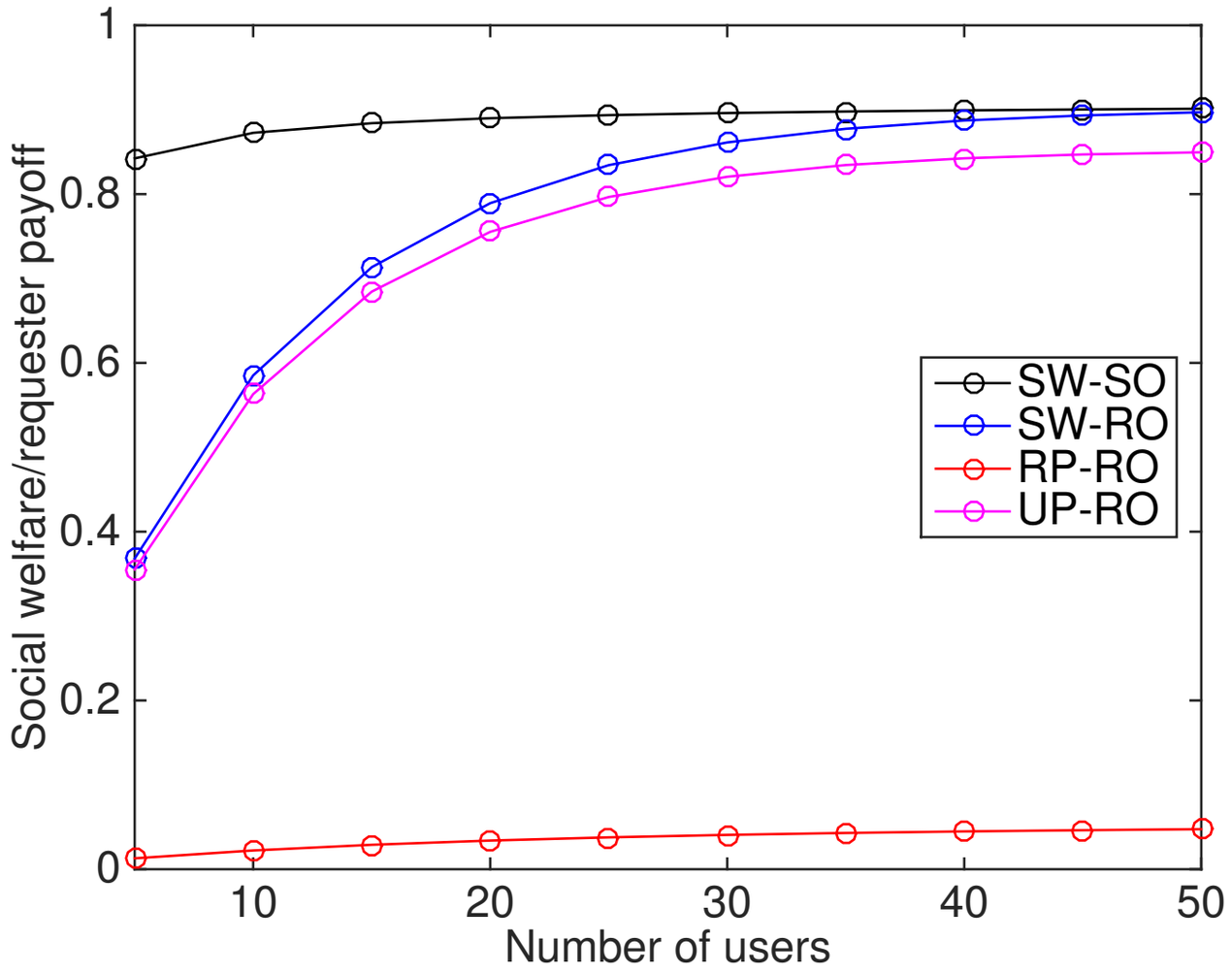
# Impact on Requester's Payoff




- ✓ The gap between the requester's optimal (CO) and socially optimal (SO) effort assignments decreases as the quality range decreases
  - When less uncertain about users' quality assign more effort which is closer to the SO effort assignment

# Impact on Requester/Users' Payoffs




- ✓ Social welfare/requester's/users' payoff increase when the number of users increases