PRADA-TF Privacy-Diversity-Aware Online Team Formation

IEEE International Conference on Web Services (ICWS) 2021

Yash Mahajan (Presenter) and Jin-Hee Cho

Department of Computer Science
Virginia Tech, Falls Church, VA, USA

9th September, 2021
Table of Contents

1 Motivation
2 Research Questions
3 Problem Statement
4 Key Contributions
5 Preliminaries
6 Design
7 Experimental Setup
8 Overview of the PRADA-TF
9 Numerical Results and Analysis
10 Conclusion
Motivation

- You can imagine an online crowdsourcing system, such as Amazon Mechanical Turk (MTurk) and Fiverr, where a requestor wants to form a team to execute a given task.
Motivation

- You can imagine an online crowdsourcing system, such as Amazon Mechanical Turk (MTurk) and Fiverr, where a requestor wants to form a team to execute a given task.

Scenario for Team Formation
Research Questions

**RQ1** What is the relationship between team performance and team members’ privacy preserving preferences?
Research Questions

RQ1 What is the relationship between team performance and team members’ privacy preserving preferences?

RQ2 What are the effects of team diversity on the team performance?
Research Questions

RQ1 What is the relationship between team performance and team members’ privacy preserving preferences?

RQ2 What are the effects of team diversity on the team performance?

RQ3 How does the team behave in an actual real world scenario?
Problem Statement

- The mechanism designer (MD), as a team leader, aims to identify an effective team composition based on diverse domain expertise of team members and effective communications allowing to share quality information.

\[
\text{arg max}_{x \in X} \sum_{i \in T} x u_i(x, \hat{\theta}_i, \hat{\theta}_{-i} | \theta_i), \forall \theta_i \in \Theta_i
\]

where \( \theta_i \) is player \( i \)'s true preferences in privacy preservation and expertise level where \( \Theta_i \) refers to a set of preferences. \( \hat{\theta}_i \) is player \( i \)'s revealed preference to the MD and \( \hat{\theta}_{-i} \) refers to revealed preferences of all other players, except player \( i \).
**Problem Statement**

- The mechanism designer (MD), as a team leader, aims to identify an effective team composition based on diverse domain expertise of team members and effective communications allowing to share quality information.

- Given a player $i$’s utility, $u_i$ formulated to maximize his/her contribution to the team performance and privacy preservation while minimizing potential privacy loss, the MD will aim to achieve:

$$\arg \max_{x \in X} \sum_{i \in T_x} u_i(x, \hat{\theta}_i, \hat{\theta}_{-i}|\theta_i), \ \forall \theta_i \in \Theta_i$$

where

- $\theta_i$ is player $i$’s true preferences in privacy preservation and expertise level where $\Theta_i$ refers to a set of preferences.
- $\hat{\theta}_i$ is player $i$’s revealed preference to the MD and
- $\hat{\theta}_{-i}$ refers to revealed preferences of all other players, except player $i$. 
Key Contributions

- Developed the PRADA-TF to identify a set of team members forming a team that can maximize social welfare of the selected team based on the concept of *mechanism design*.
Key Contributions

- Developed the PRADA-TF to identify a set of team members forming a team that can maximize social welfare of the selected team based on the concept of *mechanism design*.

- Investigated the social welfare of a selected team in both expected and actual task scenarios, unlike existing team formation studies.
Key Contributions

- Developed the PRADA-TF to identify a set of team members forming a team that can maximize social welfare of the selected team based on the concept of *mechanism design*.

- Investigated the social welfare of a selected team in both expected and actual task scenarios, unlike existing team formation studies.

- Conducted extensive experiments to evaluate the performance of the proposed PRADA-TF in terms of expected and actual social welfare, expected and actual potential privacy leakout, and team diversity of a selected team.
Key Contributions

- Developed the PRADA-TF to identify a set of team members forming a team that can maximize social welfare of the selected team based on the concept of *mechanism design*.

- Investigated the social welfare of a selected team in both expected and actual task scenarios, unlike existing team formation studies.

- Conducted extensive experiments to evaluate the performance of the proposed PRADA-TF in terms of expected and actual social welfare, expected and actual potential privacy leakout, and team diversity of a selected team.

- Developed a semi-synthetic data based on *Netscience* to reflect a realistic scenario of the online expert social network used in this work.
Preliminaries

- Task Model
- Information Model
- Adversarial Model
Task Model

- **Number of Team Members** ($m$)

- **Set of Required Domain Expertise** ($E$): The successful completion of a given task requires that a given team has expertise in a set of domain knowledge to perform the task, denoted by $E = \{e_1, e_2, \ldots, e_l\}$, where $l \leq m$.

$L$ is used to represent the set of required domain expertise, where $L(i)$ returns the extent of knowledge required in expertise domain $i$ in $E$ as a real number in $[0, 1]$. 
Information Model

Information model is based on Value-of-Information (VoI) where VoI refers to how valuable given information is to support a given task based on the following criteria:

- Credibility ($crd_{ih}$): $crd_{ih} = \theta_{ih}$,
- Usefulness ($usf_{ih}$): $uf_{ih} = \min \left[ 1, \frac{\theta_{ih}}{L(e_h)} \right]$,
- Novelty ($nov_{ih}$): $nov_{ih} = \sum_{j \in T, j \neq i} \max \left[ 0, \sqrt{\theta_{ih}^e} - \sqrt{\theta_{jh}^e} \right] \frac{|T|}{|T|}$

$$Vol_{ih} = w_{crd}.crd_{ih} + w_{usf}.usf_{ih} + w_{nov}.nov_{ih}$$

where each weight is ranged in [0, 1] as a real number with $w_{crd} + w_{usf} + w_{nov} = 1$ and represents how much each component of $Vol_{ih}$ is weighed.
Adversarial Model

- We consider possible private information leakout by team members based on their level of distrust \((1 - P_j^{MD})\).

\[ p_l^i = \exp\left(-\left|E\right| \lambda \left(\sum_{h \in E} (1 - \hat{\theta} P_l^h)\right)\right) \]

Member \(i\)'s shared information \((1 - \prod_{j \in M, i \neq j} P_{MD}^j)\),

\[ \text{Probability of any other member } j \text{ leaking out information} \]

\(^1\text{The number of domain expertise}\)
Adversarial Model

- We consider possible private information leakout by team members based on their level of distrust \((1 - P_{j}^{MD})\).
- Given a candidate team chosen by the MD, we formulate the extent of member \(i\)'s private information that can be potentially leaked out to outside of the team (i.e., unauthorized parties) by other team members \(j\)'s by:

\[
pl_i = \exp \left( - \frac{|E|^{1}}{\lambda} \left( \sum_{h \in E} (1 - \hat{\theta}_{ih}^{p}) \right) \left( 1 - \prod_{j \in M, i \neq j} P_{j}^{MD} \right) \right),
\]

\(1\) The number of domain expertise
Design

- Player Types
- Player’s Payoff
- Player’s Preference Revelation
- Team Selection Process
- Proposed Candidate Team Selection Methods
Player Types

Each player $i$ has its own type $\theta_i$ with two private signals, $\theta_i = \{\theta^e_i, \theta^p_i\}$:

- **Expertise preference** ($\theta^e_i$): We assume that each player has a set of values representing expertise in $|M|$ knowledge domains where player $i$’s expert domains and their strengths are indicated in a vector $\theta^e_i$ with $o$ elements for $h = 1, \ldots, o$. $\theta^e_{ih}$ has a real number in $[0, 1]$, given $\sum_{h \in M} \theta^e_{ih} \leq |M|$. 

- **Privacy preference** ($\theta^p_i$): Each player $i$ has a different level of privacy-preserving preference when sharing information, denoted by $\theta^p_i$. Higher $\theta^p_i$ means player $i$ is less willing to share information to minimize privacy exposure.
Player Types

Each player $i$ has its own type $\theta_i$ with two private signals, $\theta_i = \{\theta^e_i, \theta^p_i\}$:

- **Expertise preference** ($\theta^e_i$): We assume that each player has a set of values representing expertise in $|M|$ knowledge domains where player $i$’s expert domains and their strengths are indicated in a vector $\theta^e_i$ with $o$ elements for $h = 1, \ldots, o$. $\theta^e_{ih}$ has a real number in $[0, 1]$, given $\sum_{h \in M} \theta^e_{ih} \leq |M|$.

- **Privacy preference** ($\theta^p_i$): Each player $i$ has a different level of privacy preserving preference when sharing information, denoted by $\theta^p_i$. Higher $\theta^p_i$ means player $i$ is less willing to share information to minimize privacy exposure.
Player’s Payoff

Player $i$’s payoff is estimated by:

$$u_i(x, \hat{\theta}_i, \hat{\theta}_-|\theta_i) = u_{i\text{team}}(x, \hat{\theta}_i, \hat{\theta}_-|\theta_i) + u_{i\text{priv}}(x, \hat{\theta}_i, \hat{\theta}_-|\theta_i) - pl_i,$$

where

- $u_{i\text{team}}(x, \hat{\theta}_i, \hat{\theta}_-|\theta_i)$ models the decrease of novelty when more information is shared as discussed in [Trapido, 2015; Briggs, 2009] and is given by:

$$u_{i\text{team}}(x, \hat{\theta}_i, \hat{\theta}_-|\theta_i) = \sum_{h \in E} \text{Vol}_{ih} \cdot (1 - \hat{\theta}_{ih}^p)^2$$

- $u_{i\text{priv}}(x, \hat{\theta}_i, \hat{\theta}_-|\theta_i)$ reflects how much an individual’s privacy preference is preserved.

$$u_{i\text{priv}}(x, \hat{\theta}_i, \hat{\theta}_-|\theta_i) = \sum_{h \in E} (1 - \text{Vol}_{ih}) \cdot (\hat{\theta}_{ih}^p)^2,$$

- $pl_i$ refers to the loss caused by user $i$’s private information leakout.
Player’s Preference Revelation

Unlike other TF research, we additionally validate the quality of TF algorithms when a task is actually executed by the selected team.

- **Case 1:** $\theta_p^i = \hat{\theta}_p^i$, where player $i$’s revealed privacy type is the same as his/her truthful type.
- **Case 2:** $\theta_p^i \neq \hat{\theta}_p^i$, where player $i$’s revealed type is not the same as his/her truthful type.

Player $i$’s exhibited privacy preference level in an actual task execution, denoted by $\theta_p^{i'}$, is determined by:

$$
\theta_p^{i'} = \begin{cases} 
\hat{\theta}_p^i & \text{if } u_i(x^*, \hat{\theta}_i, \hat{\theta}_{-i}|\theta_i) > u_i(x, \theta_i, \hat{\theta}_{-i}|\theta_i), \\
\theta_p^i & \text{otherwise.}
\end{cases}
$$

We consider $pc_i$ as a real number in $[0, 1]$ to indicate how much player $i$ can compromise its revealed type, $\hat{\theta}_p^i$, which can be selected as a real number in $[pc_i \cdot \theta_p^i, \theta_p^i]$. Higher $pc_i$ refers to lower privacy compromise.
MD is a team leader aiming to form a team in an OSN. The MD will recruit candidate team members from the MD’s ego network, which is defined as a social network consisting of all users (or players) within k-hop distances from the MD.
Team Selection Process 2/2

Among all users within the k-hop distances from the MD, the MD will select a set of team members via two rounds:

- Select a set of member candidates from the users within the k-hop distances based on the trust relationships between users. From the users in the MD’s k-hop trust network, the MD will select top $\phi$ number of member candidates based on the candidate team selection schemes; and

- From the top $\phi$ number of member candidates selected based on the MD’s $k$-hop trust network, the MD runs the social welfare function to select a final set of $m$ team members.
Proposed Candidate Team Selection Methods

After the top $\phi$ players are selected based on the MD’s k-hop trust network, the MD further cuts down prospective team members by applying the different heuristic candidate team selection (CTS) methods to avoid high complexity.

Our proposed PRADA-TF scheme can have the following variants:

- **Utility-based Serial Dictatorship (USD)**
- **Expertise Diversity-based CTS (ED-CTS)**
- **Vol-based CTS (Vol-CTS)**
- **Information Sharing (IF-CTS)**
Experimental Setup

- Dataset
- Parameterization
- Metrics
- Comparing Schemes
A semi-synthetic dataset was developed by leveraging the *Netscience* dataset for expertise preference, which contains a coauthorship network of 1,590 scientists working on Network Theory and Experiments compiled from the bibliographies of two review articles on networks.

To convert the sparse network into a small world network, and make authors more reachable from the MD, additional edges are added when the cosine similarity value between two authors is no less than 0.9 in expertise level and the subject areas.

Each author’s privacy preference is drawn from a Gaussian distribution with mean $\mu = 0.5$ and standard deviation $\sigma = 0.3$. 
## Parameterization

<table>
<thead>
<tr>
<th>Param.</th>
<th>Meaning</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{ik}$</td>
<td>Strength in expertise in domain $k$</td>
<td>[0, 1]</td>
</tr>
<tr>
<td>$\theta_p^i$</td>
<td>Privacy preserving preference in domain $k$</td>
<td>[0, 1]</td>
</tr>
<tr>
<td>$\hat{\theta}_p^i$</td>
<td>Revealed privacy preserving preference in domain $k$</td>
<td>[0, 1]</td>
</tr>
<tr>
<td>$pc_i$</td>
<td>Extent to which a player can lie about its privacy preserving preference</td>
<td>[0, 1]</td>
</tr>
<tr>
<td>$w_e, w_s$</td>
<td>Weights for expertise and privacy preserving, respectively</td>
<td>[0, 1]</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Number of candidates selected from the trust network</td>
<td>200</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>Number of participants selected using CTS schemes</td>
<td>40</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of team members</td>
<td>20</td>
</tr>
<tr>
<td>$</td>
<td>E</td>
<td>$</td>
</tr>
<tr>
<td>$w_{crd}, w_{uf}, w_{nov}$</td>
<td>Weights for the three components of VoI</td>
<td>[0, 1]</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Sum of domain expertise levels required by a given task (i.e., $\sum_{e \in E} e_i = \epsilon$)</td>
<td>5</td>
</tr>
<tr>
<td>$L$</td>
<td>A vector of domain expertise levels in a given task,</td>
<td>[1, 1, 1, 1]</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>A constant to scale $pl_i$ in Eq. (3.3)</td>
<td>$</td>
</tr>
</tbody>
</table>
Metrics

- **Team Diversity** ($\mathcal{T}D$): This metric refers to the extent of diversity in team members’ expertise background and is given by:

  $$\mathcal{H}(\theta^e_i, \theta^e_j) = \sqrt{\sum_{h \in \mathcal{E}} D^e_{ij}}, \quad (1)$$

  where

  $$D^e_{ij} = \frac{\sum_{h \in \mathcal{E}} \max \left[ 0, \theta^e_{ih} - \theta^e_{jh} \right]}{|\mathcal{E}|}$$

- **Social Welfare** ($SW_T$): This refers to a team’s expected social welfare estimated and is given by:

  $$SW_T = \sum_{i \in T} u_i(x, \hat{\theta}_i, \hat{\theta}_{-i}| \theta_i)$$

- **Potential Privacy Loss** ($PPL$): This metric refers to the amount of penalty a player may have because of potential privacy leakout by other players, given by $pl_i$. 
Comparing Schemes

The variants of the proposed PRADA-TF scheme (i.e., USD, ED-CTS, VoI-CTS, and IF-CTS) are compared against the following two existing counterparts and one baseline model in terms of metrics described above:

- **Homophily-based CTS (H-CTS)** selects top $\zeta$ number of candidates based on the degree of players’ homophily in terms of their expertise required in a given task, based on a cosine-similarity metric. - **Counterpart**

- **Centrality-based CTS (C-CTS)** selects top $\zeta$ number of candidates based on players’ betweenness to identify the influential members for the team. - **Counterpart**

- **Random** selects $m$ number of candidate members at random from users in a given trust network. - **Baseline**
Overview of the PRADA-TF

Strength in Expertise $\theta_i^e$

Revealed Privacy Preserving Preference $\theta_i^p$

True Privacy Preserving Preference $\theta_i^{p_t}$

Actual Task Execution

Prospective Team Members

E-SW, E-PPL, Team Diversity

A-SW, A-PPL, Team Diversity

Building k-hop Trust Network

k-hop Trust Network

Candidate Team Selection (CTS) Method

Top m

Maximum Social Welfare

Top $\phi$

$\theta_i^{p_t} = \begin{cases} \hat{\theta}_i^p & \text{if } u_i(x^*, \hat{\theta}_i, \hat{\theta}_{-i}|\theta_i) > u_i(x, \theta_i, \hat{\theta}_{-i}|\theta_i), \\ \theta_i^p & \text{otherwise.} \end{cases}$
Numerical Analysis and Results
Effect of Different Task Types on Social Welfare

The lower bound weight of a revealed privacy preference \((pc_i)\) is set to \(= 0.8\), the number of hops in an online trust network \((k\text{-hop})\) is set to 5, and the error bar represents the standard deviation.

E-SW vs. A-SW under \(|E| = 5\)

E-SW vs. A-SW under \(|E| = 3\)

E-SW vs. A-SW under \(|E| = 1\)
Numerical Analysis and Results
Effect of Different Task Types on Social Welfare

The lower bound weight of a revealed privacy preference \((pc_i)\) is set to \(= 0.8\), the number of hops in an online trust network \((k\text{-hop})\) is set to 5, and the error bar represents the standard deviation.

**Key Observations**

- A certain level of team diversity is critical in promoting team performance.
- When a task requires high expertise in a less number of domain, the uniqueness of a member in a single expertise introduces more critical impact to improve team diversity.
Effect of Different Task Types on Potential Privacy Loss

Key Observations

There exists a trade-off between information sharing and privacy. Sharing more information regardless of its value (VoI) will naturally lead to higher privacy loss.

All players act similarly in a selfish manner to protect their privacy irrespective of the scheme used.
Effect of Different Task Types on Potential Privacy Loss

Key Observations

- There exists a trade-off between information sharing and privacy. Sharing more information regardless of its value (VoI) will naturally lead to higher privacy loss.

- All players act similarly in a selfish manner to protect their privacy irrespective of the scheme used.
Effect of Different Task Types on Team Diversity

Key Observations
From this observation, we say that team diversity is important when the task requires subject-area specific expertise. However, under a task requiring a diverse domain expertise, it is unclear that having a certain level of diversity is closely related to high team performance (i.e., high SW).

Yash Mahajan (Presenter) and Jin-Hee Cho
PRADA-TF
9th September, 2021
**Key Observations**

- From this observation, we say that team diversity is important when the task requires subject-area specific expertise.

- However, under a task requiring a diverse domain expertise, it is unclear that having a certain level of diversity is closely related to high team performance (i.e., high SW).
The key findings from our study are summarized as:

- **RQ1**: What is the relationship between team performance and team members’ privacy preserving preferences?
The **key findings** from our study are summarized as:

- **RQ1**: What is the relationship between team performance and team members’ privacy preserving preferences?
  
  Team members are more likely to compromise their privacy preference only when keeping their truthful preference significantly hurts their utilities in team performance.
Conclusion 1/2

The key findings from our study are summarized as:

- **RQ1**: What is the relationship between team performance and team members’ privacy preserving preferences? Team members are more likely to compromise their privacy preference only when keeping their truthful preference significantly hurts their utilities in team performance.

- **RQ2**: What are the effects of team diversity on the team performance?
The **key findings** from our study are summarized as:

- **RQ1**: What is the relationship between team performance and team members’ privacy preserving preferences?
  Team members are more likely to compromise their privacy preference only when keeping their truthful preference significantly hurts their utilities in team performance.

- **RQ2**: What are the effects of team diversity on the team performance?
  Team diversity is important when the task requires subject area specific expertise. However, under a task requiring a single domain expertise, it is unclear that having a certain level of diversity is closely related to high team performance (i.e., high SW).
RQ3: How does the team behave in an actual real world scenario?
Conclusion 2/2

- **RQ3**: How does the team behave in an actual real world scenario?

We observed that A-SW is likely to be higher than E-SW because team members are more likely to compromise their privacy preferences, aiming to increasing information sharing and accordingly higher utility.

The actual PPL (A-PPL) is lower than the expected PPL (E-PPL) because E-PPL is calculated using the revealed type of the team members whereas in an actual task execution, a player might either stick to its revealed type or revert back to its true type, which would be higher, leading to less privacy loss.
RQ3: How does the team behave in an actual real world scenario?
We observed that A-SW is likely to be higher than E-SW because team members are more likely to compromise their privacy preferences, aiming to increasing information sharing and accordingly higher utility.
The actual PPL (A-PPL) is lower than the expected PPL (E-PPL) because E-PPL is calculated using the revealed type of the team members whereas in an actual task execution, a player might either stick to its revealed type or revert back to its true type, which would be higher, leading to less privacy loss.
Any Questions?

Thank you!

Feel free to reach out to Yash Mahajan at
yashmahajan@vt.edu