Reward and punishment mechanism for research data sharing

Jinfang Niu
University of Michigan
Concepts and Definitions

Research data

Data created by scientific researchers primarily for research purposes.

Data created primarily for commercial, clinical or administrative purposes, such as user purchase history, patient medical records, or government records are excluded.
Data sharing for collaboration

Very large data sets

A researcher has collected data, but he/she doesn’t have enough time or capability to analyze the data.

One researcher’s data can be used to address some research questions that he is not interested in.
Data sharing for reuse

Data have value beyond the purpose for which they were originally collected.

Reused for
- Verification
- Replication
- Different research questions
- Meta-analysis
- Decision making & information
- Teaching
Sharing through archives

- Producers (prepare & deposit)
- Archive (process & disseminate)
- Funding agency
- Users
Publicly funded research data are public goods

Non-rivalry: one individual’s use of the data does not reduce the amount available for other people.

Not-excluded: no one is excluded from the online data archive whether or not they deposited data.
Theories about public goods

Free-riding in the voluntary contribution to public goods tends to be the dominant strategy in a non-cooperative game.

The achievement of a pareto-optimal via decentralized methods in the presence of public goods is fundamentally incompatible with individual incentives.
Beyond public goods

For normal public goods
contributors benefit the same as free-riders.
Contributors are not harmed from their contributions

In data sharing
The benefit of data preparation largely goes to the users.
Assumption: It is not very likely that a data depositor uses his own data deposited into a data archive

Data producers might be harmed by sharing data.
Incentive mechanisms implemented

NIH
In the case of noncompliance, NIH can take various actions to protect the federal government's interests. In some instances, NIH may make data sharing an explicit term and condition of subsequent awards. It does not offer rewards for doing well in data-sharing plan.

ESRC
The final payment of an award will be withheld until data has been deposited in accordance with the requirements. The requirements of the data sharing policy are now a condition of funding.

Both: citation of data is not mandatory, co-authorship is not allowed as a condition for sharing data.
Producers’ expectations

Some sort of acknowledgments for data deposit, such as a certificate

If data deposit is mandatory to receive new funding, or a prerequisite for publishing a paper derived from the data …

Some grantees are strongly against punishment

---- From survey data
A simple math model

Purpose: to show the effects of punishment and reward.

Punishment: the data producer will be punished if the effort he spent on data preparation is lower than a threshold.

Reward: when the deposited data is used, the producer of the data gets a reward.
Model set-up

A data producer:
has total fund P
chooses $\theta$ for research
Chooses $e$ for data preparation ($P = \theta + e$)
benefits $\Omega(\theta)$ ($\Omega'(\theta) > 0$) from research

Assumptions
$\Omega(\theta)$ is concave, and $\Omega'(\theta) > 0$
The data producer always tries to maximize his benefit.
No reward & no punishment

Data producer:
Choice: $e = 0$
Maximized benefit: $\Omega(P)$

User:
no data to use, benefit: 0

Social benefit: $\Omega(P)$
Punishment only

Producer:

If $e \geq e'$, benefit: $\Omega(\theta) = \Omega(P-e)$, max: $\Omega(P - e')$

If $e < e'$, benefit: $\Omega(\theta) - f$, ($f > 0$), max: $\Omega(P) - f$
When $\Omega(P - e') > \Omega(P) - f$

The data producer would meet the threshold to avoid the punishment.

Two explanations:

- The punishment is severe enough.
- The threshold is easy to meet.

When $\Omega(P - e') = \Omega(P) - f$

The data depositor is indifferent between getting punished and meet the threshold.
When $\Omega(P - e') < \Omega(P) - f$

The data producer would choose not to prepare data and get punished.

Two explanations:
- the punishment is not severe enough
- the threshold is too costly to meet.

Conclusion:
In the punishment only scenario, data producer's benefit is $\max [\Omega(P - e'), \Omega(P) - f]$. 
when the data producer would rather be punished, benefit: 0.

Social benefit: $\Omega(P)$

$$\max [\Omega(P - e'), \Omega(P) - f] = \Omega(P) - f.$$ The data producer loses $f$, but the funding agency gets $f$. : $\Omega(P) - f + f + 0 = \Omega(P)$.

Too weak punishment or too high a threshold for punishment is not effective.
When the producer chooses to meet the threshold, 
The user would use the data with probability $\pi(e')$ 
($\pi'(e) > 0$, $\pi(0) = 0$). Expected benefit: $v * \pi(e')$

Social benefit: $\Omega(P - e') + v * \pi(e')$
Max $[\Omega(P - e'), \Omega(P) - f] = \Omega(P - e')$

when $\Omega(P) < \Omega(P - e') + v * \pi(e')$, the punishment 
and threshold are effective.
When $\Omega(P) > \Omega(P - e') + v * \pi(e')$, the punishment 
and threshold are not effective.
Reward only

When the deposited data is used, the producer gets a reward $r$. The deposited data has a probability $\pi(e)$ ($\pi'(e) > 0$, $\pi(0) = 0$) of being used.

Producer benefit: $\Omega(\theta) + r \ast \pi(e) \iff \Omega(P-e) + r \ast \pi(e)$. 
If there is a value of $e$ which makes $[\Omega(P-e) + r * \pi(e)]' = 0 \iff r * \pi'(e) = \Omega'(P-e)$, the producer's benefit is maximized when the marginal benefit of spending an additional amount of funding on research is equal to the product of reward and the marginal probability of being used.
If the reward \( r \) is so big that \( r \pi'(e) > \Omega'(P-e) \) is always true, the data depositor would spend all funding available on data preparation.

The utility is maximized when \( e = P \)

If the reward is so small that \( r \pi'(e) < \Omega'(P-e) \) is always true, the data producer would not deposit data and there is no data to use.

The producer’s benefit is maximized when \( e=0 \)

The social benefit is \( \Omega(P) \). The small reward is not effective.
Suppose the producer’s benefit is maximized at $e = e^*$, 

User’s benefit: $v^*\pi(e^*)$.
Producer’s benefit: $\Omega(P-e^*) + r^* \pi(e^*)$,
The funding agency: $- r^* \pi(e^*)$.
Social benefit: $\Omega(P-e^*) + v^*\pi(e^*)$.

$\Omega(P) \leq \Omega(P-e^*) + v^*\pi(e^*)$, reward causes at least as much social benefit than no reward & no punishment.

**Warning:** do not make the reward too big or too small.
Characterize punishment and reward

◆ Punishment

Coercive and uniform

Pros and cons:
- Makes all data accessible to the public.
- All data producers have to prepare and deposit data to avoid punishment even if their data sets are not likely to be used.
Cumulative and Average Monthly Access Rates for the 10 Most Frequently (Top 10) and 10 Least Frequently (Bottom 10) Accessed Data Sets.

<table>
<thead>
<tr>
<th>Cumulative</th>
<th>Average (per month)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 10</td>
<td>Bottom 10</td>
</tr>
<tr>
<td>5185</td>
<td>27</td>
</tr>
<tr>
<td>2345</td>
<td>25</td>
</tr>
<tr>
<td>2234</td>
<td>24</td>
</tr>
<tr>
<td>1651</td>
<td>24</td>
</tr>
<tr>
<td>1623</td>
<td>23</td>
</tr>
<tr>
<td>1328</td>
<td>21</td>
</tr>
<tr>
<td>1267</td>
<td>20</td>
</tr>
<tr>
<td>1229</td>
<td>16</td>
</tr>
<tr>
<td>932</td>
<td>14</td>
</tr>
<tr>
<td>851</td>
<td>11</td>
</tr>
</tbody>
</table>

5185: 1037 (Top 10), 27 (Bottom 10)
2345: 260 (Top 10), 25 (Bottom 10)
2234: 319 (Top 10), 24 (Bottom 10)
1651: 183 (Top 10), 24 (Bottom 10)
1623: 232 (Top 10), 23 (Bottom 10)
1328: 148 (Top 10), 21 (Bottom 10)
1267: 271 (Top 10), 20 (Bottom 10)
1229: 137 (Top 10), 16 (Bottom 10)
932: 104 (Top 10), 14 (Bottom 10)
851: 95 (Top 10), 11 (Bottom 10)
contrast of access

![Graph showing the contrast of access with data sets from 1 to 10. The graph compares 'top 10' and 'bottom' access metrics. The y-axis represents access values ranging from 0 to 1200, with a significant drop in access for the 'top 10' set after the first data set.]
 Rewards

Inducive and selective

Pros and cons:

Related to the actual use of data
Difficult to anticipate actual use.
Not all data are accessible to the public
Conclusions

Publicly funded research data are public goods.

They have some features beyond public goods that make free-riding even more attractive under voluntary contribution.

Incentive mechanisms should be available to encourage data sharing.

Punishments or rewards need to be carefully decided to avoid unintended consequences.
Thanks!

This research was supported by the National Science Foundation (NSF Award Number IIS-0456022) as part of a larger project entitled “Incentives for data producers to created archive-ready data sets.”