Reward and Punishment Mechanism for Research Data Sharing

*by Jinfang Niu**

Abstract

Many funding agencies require grantees to deposit their data into an archive after they finish their research projects. The archive processes and disseminates the data for public use. These deposited data sets are public goods that benefit users and society. However, under voluntary contribution, public goods tend to be under-provided. For normal public goods, the contributors benefit from their own contributions as much as free-riders. Contributors are not harmed by their contributions. In the data sharing case, data producers make efforts to prepare the data for deposit, but the benefit of the data preparation largely goes to secondary users. In addition, data producers are at risk of being harmed by the misuse and misinterpretation of data by unqualified users, or by being charged with misconduct. That makes free-riding even more attractive. To motivate data producers to prepare and share data, there must be some incentive mechanisms. In this paper, I built a simple mathematic model to analyze the effects of punishment and reward. Hopefully it will help policy makers decide on incentive mechanisms for data sharing.

Background

Many funding agencies require grantees to deposit data into an archive after they finish their research projects, such as the National Institute of Justice (NIJ), the National Institutes of Health (NIH) in the United States, and the Medical Research Council (MRC) and the Economic and Social Research Council (ESRC) in the United Kingdom. The archive processes and disseminates data for public use. Data sharing benefits society in many ways. It saves funding and avoids repeated data collecting efforts, allows the verification and replication of research findings, facilitates scientific openness, deters scientific misconduct, and supports communication and progress

Before deposit, data depositors need to prepare their data according to the requirements of data archives. The purpose of the preparation is to help secondary use of data and protect the privacy of human research subjects. Data preparation includes three kinds of work: preparing data, creating documentation and processing confidential information. Data preparation includes checking the integrity[1] and consistency of data, careful naming of variables and choice of variable labels that will be easy for secondary users to understand, organizing the variables such as grouping them to enable secondary analysts to get an overview of the data quickly, etc. (ICPSR, 2005). Data documentation provides metadata about the data sets and research projects, such as the principal investigator of the project, when and where the data were collected, the methodology and procedures used to collect the data, details about codes, definitions of variables, frequencies, and the like (NIH, 2003). Even data collection instruments, such as questionnaires and interview guides are required parts of documentation. Documentation is indispensable for the searching, managing, preserving and re-using of data. In other words, without adequate documentation, secondary users of a data set will not be able to find the data, nor will they be able to interpret and analyze the data. As a result, the goal of data sharing will not be achieved. In addition, insufficient documentation might lead to the misuse of data or incorrect conclusions. To protect confidential information in data, all direct identifiers, such as names, addresses, telephone numbers, and Social Security Numbers, have to be removed. In addition, indirect identifiers and other information that could lead to "deductive disclosure" of participants' identities should also be removed or processed before the data are made public.

Data preparation involves a lot of work, and a fair amount of it is done only for secondary users. For example, data producers do not have to process the confidential information if they keep the data for their own use. Many data producers do document data for their own use. However, documentation created for the producer's own use are informal and biased toward short-term needs. To share with others, data producers have to take extra effort to shape the "public face" of their documentation (Markus, 2001). In addition, data producers should take the main responsibility for preparing data. Zimmerman (2003) found that both secondary data users and data managers (intermediaries or data archivists) agree that no one understands the data better than the scientists who gathered them, and that it is the data producers who must document data.

When publicly funded research data are disseminated to the public through the website of a data archive, no one is excluded from using them, and one individual's use of the data and documentation does not reduce the amount available for other people. Those data sets are public goods by definition (Mas-Colell, et al., 1995). Since no one is excluded from the online data archive whether or not they have deposited data, as with other public goods people have strong incentives to free ride in preparing and depositing data. From the game theory perspective, free-riding in voluntary contribution to public goods tends to be the dominant strategy in a non-cooperative game (Bergstrom, et al., 1986; Cornes & Sandler, 1986).

For normal public goods, the contributors benefit from their own contributions in the same way as free-riders, and they are not harmed by their contributions. For example, once a bridge is built, the contributors and free - riders get the same benefit. However, in the data sharing case, the depositor of a data set does not benefit from the data he deposited in the same way as secondary users. A data depositor is unlikely to use his own data deposited into a data archive, either because he has used it before, or because he keeps his own data for future use. People mostly benefit from others' contributions. The benefit of depositors' effort in data preparation largely goes to the users. In addition, data producers are at risk of being harmed by the misuse and misinterpretation of data by unqualified users, or by being charged with misconduct. That makes free - riding even more attractive. To change this situation, there must be some incentive mechanisms to motivate researchers to prepare and deposit data.

The incentive mechanisms that some funding agencies have implemented focus on punishment for non-compliance with data sharing requirements, and pay less attention to rewards. According to the policy of the National Institutes of Health (NIH), in the case of noncompliance (depending on its severity and duration), NIH can take various actions to protect the Federal Government's interests. In some instances, for example, the NIH may make data sharing an explicit term and condition of subsequent awards (NIH, 2003). Under the policy of the 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 ESRC research funding." (ESRC, 2000). The data sharing policies of NIH and ESRC do not mandate that users cite the data they use, and they are against the idea that data producers require co-authorship as a condition for sharing the data. NIH explicitly stated that they do not offer rewards for doing well in data-sharing. Data from a survey[2] of the grantees of a funding agency showed that some grantees expect rewards for data deposit. For example, one grantee said he would be more likely to deposit data if there were some sort of acknowledgment that he had deposited data, such as a certificate. Some other grantees claimed that some sort of punishment would make them more likely to deposit data, for example, if data deposit were mandatory to receive new funding from NIJ, or a prerequisite for publishing a paper derived from the data. One grantee was strongly against a punishment mechanism. He said: "Do you really want a system where archiving data prevents people from publishing or from doing new work? This would be a triumph of bureaucracy over common sense. If the funding agency becomes obsessed with bureaucratic requirements, they will drive away talented researchers."

I believe that either punishments or rewards would provide incentives for data producers to take more effort in data preparation. But when decide the punishment or reward mechanisms, the level of the punishment and reward should be carefully chosen. Otherwise, unintended consequences might occur. To help illustrate this, I have built a very simple mathematical model

The model

Three parties are involved in the model. They are the data producer, the data user and the funding agency. In reality, there are many data producers and users. To make the problem simple, I only consider one data producer and one data user. The data producer has total fund P. He chooses θ and e to spend on research and data preparation respectively ($P = \theta + e$). He benefits $\Omega(\theta)$ from spending θ on research. I assume that $\Omega(\theta)$ is concave, differentiable and $\Omega(\theta) > 0$, meaning that the more the data producer spends on research, the more he will benefit, but the increase rate of the benefit decreases[3]. See Figure 1 for the graph of $\Omega(\theta)$. I also assume that the data producer always tries to maximize his benefit when making decisions. I analyzed and compared the social benefits generated from three scenarios: no reward & no punishment, punishment only and reward only. Social benefit is defined as the sum of the benefit gained by the data producer, the user and the funding agency in each scenario.

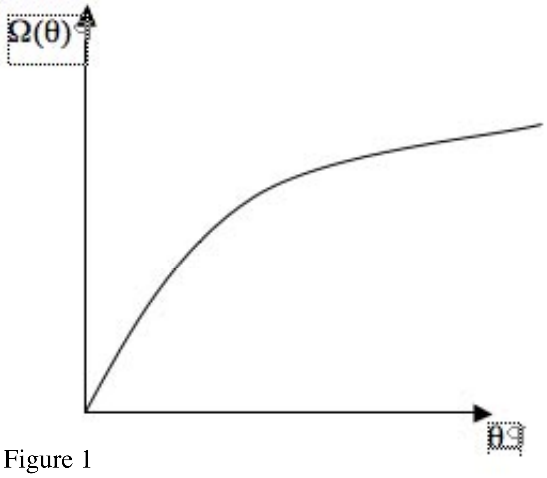

Figure 1

Scenario 1: No reward & no punishment

In the no reward & no punishment scenario, the data producer does not benefit from spending effort on data preparation, and he loses nothing if does not spend any effort on data preparation. To state this formally, the utility function of the data producer is $\Omega(\theta)$, $0 < \theta \leq P$. Since the more the data depositor spends on research, the more he benefits, the data depositor would spend $e = 0$ on data preparation to maximize his utility. His maximized utility is $\Omega(P)$. For the user, since the data depositor did not spend any effort on data preparation for deposit there is no data to use, so the user's utility is 0. The social benefit is the sum of the benefit of the depositor and the user: $\Omega(P)$.

Scenario 2: Punishment only

In this scenario, the data producer will be punished if the effort he spends on data preparation is lower than a threshold. To state this formally, the benefit function is: if $e \geq e'$, the data depositor's benefit is $\Omega(\theta)$, if $e < e'$, the data depositor's benefit is: $\Omega(\theta) - f$, $(f > 0)$. f is a fine that the data depositor has to pay to the funding agency if he is punished. e' is the threshold for punishment.

In this case, to maximize his benefit, the data producer needs to compare the highest possible benefit he could get if he passes the threshold versus if he does not. Mathematically, he needs to maximize a benefit function of two parts, and pick the one that is larger. When $e \geq e'$, the data depositor's benefit is $\Omega(\theta) = \Omega(P-e)$. Since $\Omega'(\theta) > 0$, to maximize $\Omega(\theta)$, we need to minimize e, the smallest value of e is e', and the maximized benefit of the data depositor is $\Omega(P - e')$. When $e < e'$, the data depositor's benefit is $\Omega(\theta) - f = \Omega(P-e) - f$, again we need to minimize e to maximize the data depositor's benefit. The smallest value of e is 0, so the maximized utility of the data depositor is $\Omega(P) - f$.

Now compare $(P - e')$ and $\Omega(P) - f$.

If $\Omega(P - e') > \Omega(P) - f \iff f > \Omega(P) - \Omega(P - e')$, the function is maximized at $e = e'$, which means that the data producer will benefit more by passing the threshold. So the data producer would choose to pass the threshold to avoid punishment. There are two explanations for this. First, if the threshold (e') is fixed, this means that the punishment is severe enough (f is big enough) to make $f > \Omega(P) - \Omega(P - e')$. Second, if the punishment level is fixed (keep f constant), this means the threshold is easy to meet (e' is low), so the data depositor would like to meet the threshold to avoid the punishment.

If $\Omega(P - e') = \Omega(P) - f$, the data depositor is indifferent between preparing data for deposit and getting punished.

If $\Omega(P - e') < \Omega(P) - f \iff f < \Omega(P) - \Omega(P - e')$, the data producer benefits more from being punished than from preparing and depositing data. To maximize his benefit, the data depositor will choose to spend nothing on data preparation and be punished. There are two explanations for this. First, if the threshold is fixed, it means the punishment is not severe enough to deter non-compliance behaviors. Second, if the punishment level is fixed, it means the threshold e' is too costly to meet, so the data depositor would rather be punished than meet the threshold.

Based on the analysis above, we can see that the data depositor's benefit in this scenario is max $[\Omega(P - e'), \Omega(P) - f]$.

For the user, when the data depositor would prefer to be punished than deposit data $(\Omega(P - e') < \Omega(P) - f)$, there is no data to use. So the user's benefit is 0. If Max $[\Omega(P - e'), \Omega(P) - f] = \Omega(P) - f$, the data depositor loses f, but the funding agency gets f[4]. The social benefit is the sum of the benefits of the data depositors, the users and the funding agency. So the social benefit $= \Omega(P) - f + f + 0 = \Omega(P)$. This is equal to the social benefit in the no punishment & no reward scenario. We can see that too weak a punishment or too high a standard for data preparation is not effective. The data depositor is punished, yet there is no gain in social benefits. This actually confirms the findings of existing literature that punishment is effective only when it is relatively harsh (Trevino & Ball, 1992).

When the data depositor chooses to meet the threshold for data preparation, there is data available to use. But deposited data sets are not always used. In reality, there are various reasons. For example, a user does not use a data set because it does not fit his research purpose, or because the documentations of the data is not sufficient. Here, I assume that the probability that the data is used depends on the fund that the data producer spends on data preparation. The more fund the data producer spends on data preparation, the more likely the data is used by the user. To state this formally, there is a probability $\pi(e')$ $(\pi'(e) > 0, \pi(0) = 0)$ that the user will use the data. If he uses the data, the user will benefit v, so the user's expected benefit of using data is $v * \pi(e')$. The social benefit is: $\Omega(P - e') + v * \pi(e')$. Remember the social benefit in the no punishment & no reward is $\Omega(P)$.

So when $\Omega(P) < \Omega(P - e') + v * \pi(e')$, it means that an appropriate punishment and a carefully selected threshold causes higher social benefit than no punishment & no reward. When $\Omega(P) > \Omega(P - e') + v * \pi(e')$, it means the reverse.

Scenario 3: Reward only

In this scenario, when the deposited data is used, the producer of the data gets a reward r, and the user of the data set benefits v. The deposited data has a probability $\pi(e)$ $(\pi'(e) > 0, \pi(0) = 0)$ of being used. So the depositor's expected benefit from the reward is $r * \pi(e)$, the user's expected benefit $v*\pi(e)$. The producer's total benefit is the expected benefit from the reward plus the benefit from doing research: $\Omega(\theta) + r * \pi(e) <=> \Omega(P-e) + r * \pi(e)$.

To maximize the benefit of the data producer, we need to check the first order condition of $\Omega(P-e) + r * \pi(e)$.

If there is a value of "e" which makes $[\Omega(P-e) + r * \pi(e)]' = 0 <=> r * \pi'(e) = \Omega'(P-e)$, then the benefit of a data producer 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 no matter how small the marginal probability of the data being used ($\pi'(e)$) is, the product ($r * \pi'(e)$) is always greater than the marginal benefit of doing research [$r * \pi'(e) > \Omega'(P-e)$], it means that the function $\Omega(P-e) + r * \pi(e)$ is monotonically increasing in the interval $e \in [0, P]$. In this case, utility is maximized when $e = P$, which means that to maximize his benefit, the data depositor should spend all funding available on data preparation. If the reward is so small that no matter how big the marginal probability of the data being used, the product ($r * \pi'(e)$) is always smaller than the marginal benefit of doing research [$r * \pi'(e) < \Omega'(P-e)$], the function $\Omega(P-e) + r * \pi(e)$ is monotonically decreasing in the interval $e \in [0,P]$. Here the benefit is maximized when $e = 0$, which means that to maximize his benefit, the data depositor should spend all funding on research. Then the data producer will not deposit data and there is no data to use. In this case, the data producer is not rewarded because he did not deposit data. His benefit is $\Omega(P)$. The user does not benfit because there is no data to use. The funding agency does not need to pay any reward to the producer. The social benefit = sum of benefit (producer, user and funding agency) = $\Omega(P)$. It is exactly the same as the case with no reward & no punishment. In this case, the small reward is not effective at all. This confirms the findings of other literature that rewards should be of sufficient value, as rewards of insufficient value are the same as no reward at all (Buhler, 1992). Neither of these two cases are what we want. So we need to be careful not to make the reward too big or too small.

Suppose the data depositor's benefit is maximized at $e = e^*$, and the data user's benefit is $v^*\pi(e^*)$. The data depositor's benefit is $\Omega(P-e^*) + r * \pi(e^*)$, but the $r * \pi(e^*)$ is from the funding agency. In other words, the funding agency loses $r * \pi(e^*)$ in rewarding the data depositor. So social benefit = sum of the benefit of (depositor, user and funding agency) = $\Omega(P-e^*) + v^*\pi(e^*)$.

$\Omega(P)$ is the special point for $\Omega(P-e^*) + v^*\pi(e^*)$ where $e = 0$. e^* is the maximized point, so $\Omega(P)$ cannot be greater than $\Omega(P-e^*) + v^*\pi(e^*)$. So reward causes at least as much social benefit as no reward and no punishment. But we need to find an appropriate reward to make sure that the data depositor does not choose $e = 0$ or $e = P$.

This simple model reveals the importance of choosing an appropriate level of punishment and reward, and an appropriate threshold for punishment. It does not deal with specific kinds of punishment or reward. For example, we do not consider whether we should punish non-compliers by withholding 10% of their final grant, or by factoring the quality of deposited data into consideration of future grants. I propose the following reward mechanism for data sharing policies: make the citation of data sets or the acknowledgement of data providers a mandatory requirement of publishing, the violation of which is treated in the same way as using but not citing published papers. Treat the citation of data the same as the citation of published papers in the performance evaluation of researchers.

As a complement for the model, here is a qualitative analysis of the punishment and reward mechanisms. Effective punishments force all data producers without plausible excuses to prepare and deposit data, which would make all data collected under public funding accessible to the public. This gives users chances to verify the research findings of data producers, which would deter scientific fraud and misconduct. On the other hand, not all data sets will be used heavily (Niu and Hedstrom, 2007). Under the punishment scenario, even if the data is very unlikely to be used in the future the data producer still needs to prepare and deposit data to avoid punishment. Also, the archive needs to process, disseminate and preserve the data. Enforcing uniform strong punishment on all data sets would cause the waste of resources. Unlike the coercive and uniform nature of punishments, rewards are inductive and selective. Rather than forcing researchers, rewards induce researchers to prepare and deposit data. Researchers who expect their data to be used by other people will be motivated to do better in data preparation. Data depositors who do not expect their data to be used will not prepare and deposit data, which may be a good choice. In this case, not all federally funded data sets will be made available to the public. The chance to verify some research is lost. Also, data producers decide their effort in data preparation based on the expected future use of their data, which might be hard to anticipate.

Acknowledgments

This research is funded by the National Science Foundation, Award # IIS 0456022 as part of the project entitled "Incentives for Data Producers to Create Archive-Ready Data Sets."

I gratefully thank Linhong Chen's help and instructions in building the model.

References

Bergstrom, T. C., Blume, L. & Varian, H. R. (1986). On the Private Provision of Public Goods. Journal of Public Economics, February 1986, 29(1), pp. 25– 49.

Buhler, Patricia M. (1992). The keys to shaping behavior. Supervision v. 53 (Jan. '92) pp. 18-20.

Cornes, R. C. & Sandler, T. (1986). The theory of externalities, public goods, and club goods. Cambridge: Cambridge University Press.

ESRC (Economic and Social Research Council), (2000). Economic and Social Research Council data policy. http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-12051.pdf

ICPSR (Inter-university Consortium for political and social research). (2005). Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle. http://www.icpsr.umich.edu/access/dataprep.pdf

Markus, M. L. (2001) Toward a theory of knowledge reuse: type of knowledge reuse situations and factors in reuse success. Journal of Management Information Systems. 18(1), pp. 57-93.

NIH (National Institute of Justice). (2003). NIH Data Sharing Policy and Implementation Guidance. http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Niu, J. & Hedstrom, M. (2007). Streamlining the "Producer/Archive" Interface: Mechanisms to Reduce Delays in Ingest and Release of Social Science Data. DigCCurr 2007. April 18-20, Chapel Hill, NC, USA.

Trevino, L. K. & Ball, G. A. (1992). "The social implications of punishing unethical behavior: observers' cognitive and affective reactions." Journal of Management, 18(4), pp. 751-768.

Zimmerman, A. (2003). Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. Unpublished Dissertation, Information and Library Studies, University of Michigan, Ann Arbor.

Endnotes:

1. Integrity means no wild codes or impossible values. For example, a respondent has 99 rather than 9 children. Consistency means the variable values are consistent, for example, a respondent doesn't work but reports earnings.

2. That survey was done in 2006 by the team of the NSF project "Incentives for Data Producers to Create Archive-Ready Data Sets."

3. P: the total fund available for the research project. θ: the amount of fund the data producer spent on research. e: the amount of fund the data producer spent on data preparation. $\Omega'(\theta)$: the first derivative of $\Omega(\theta)$.

4. The fine is paid by the data depositor to the funding agency. So when the data producer pays f to the funding agency, the data depositor loses f, and the funding agency gets f.

*Jinfang Niu is a PhD candidate at the School of Information, University of Michigan. She has a Master degree in Library Science and 3 years working experiences in Tsinghua University Library, China. Her research interests include data sharing, metadata, digital preservation and digital libraries. The paper was presented at the IASSIST 2007 conference in May in Montreal, Canada. Contact: niujf@umich.edu.