bioRxiv preprint doi: https://doi.org/10.1101/093518; this version posted December 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The sharing of open data: a game-theoretic approach

Bradly Alicea
Orthogonal Research, http://orthogonal-research.weebly.com

ABSTRACT

Participation in Open Data initiatives require two semi-independent actions: the sharing of data produced by a researcher or group, and a consumer of shared data. Consumers of shared data range from people interested in validating the results of a given study to transformers of the data. These transformers can add value to the dataset by extracting new relationships and information. The relationship between producers and consumers can be modeled in a game-theoretic context, namely by using a Prisoners' Dilemma (PD) model to better understand potential barriers and benefits of sharing. In this paper, we will introduce the problem of data sharing, consider assumptions about economic versus social payoffs, and provide a game-theoretic model of data sharing. Several variations on the payoff matrix are given for different institutional scenarios, ranging from the ubiquitous acceptance of Open Science principles to a context where the standard is entirely non-cooperative. Implications for building a CC-BY economy are then discussed in context.

Keywords: Game Theory, Open Science, Infonomics, Econosemantics

Introduction

In a rather short period of time, the culture and practice of Open Science [1] has contributed to substantial changes in academic practices. One way to assess the advantages of openly sharing datasets, both in terms of scientific progress and diffusion of scientific knowledge is to use a formal model. Specifically, the process of sharing and consuming open data as a Prisoners' Dilemma (PD) game. The practice of openness and its associated benefits must be consistent across both the producers of primary data and the transformers (e.g. active consumers) of secondary data (Figure 1). While these acts may be independent of each other, this type of system must involve shared objectives to be sustainable.

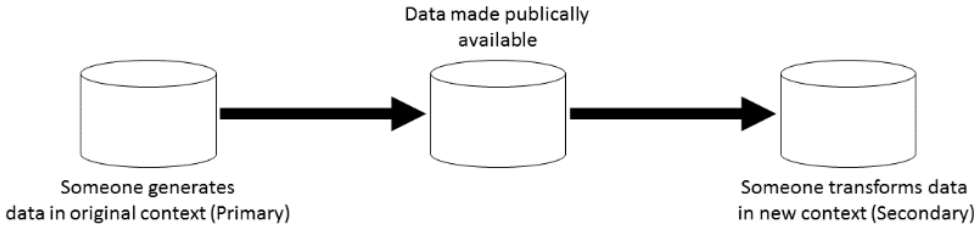

Figure 1. Typical pipeline of shared data between primary generation and secondary transformation.

To formulate this set of interactions as a game-theoretic model, we must determine the payoff structure of resulting from Open Science (and specifically Open Data) practices. Is this simply contingent on the type of license (e.g. CC-BY) chosen by the producer of primary data? To understand this better, we will assume that the enablers of openness are social rather than financial transactions. Social transactions are quantifiable using the concept of social capital [2]. Favorable social interactions result in increased social capital, while discouraged social interactions result in

bioRxiv preprint doi: https://doi.org/10.1101/093518; this version posted December 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a loss of social capital [3]. Conforming to the prevailing social norms, or producing something that helps others within that set of norms, can produce social capital-derived utility for the transactor. This is distinct from financial capital, but operates in conjunction with social capital in order to determine the utility of engaging in data sharing and reuse.

Assumptions

Before we can calculate our payoffs, we must clarify a few assumptions. The first assumption is one of openness: what is the incentive for a producer to share data? While there are an increasing number of institutional encouragements for making access to research materials open [4], let us instead focus on the relationship between time, financial gain, and utility. Sharing data requires both time spent on annotation and archiving, as well as the presence of an easy-to-use repository that allows for the assignation of credit. In particular, professional credit can be thought of as social capital, and plays a role in the utility of sharing data. This includes (but is not limited to): recognition by tenure and promotion committees, normative recognition of Altmetric milestones, and an increased recognition of reanalysis and replication studies as a valid form of scientific discovery [5].

One way to remove the effects of immediate financial incentive, which may or may not influence data sharing but will definitely confound our payoff structure, is to assume that all data are completely open access. In this case, completely open access means free of subscription and publication (e.g. APC) fees. To assume that open data is superior to paywalled data, the following relation must hold: the utility (social capital) gained from sharing data must exceed the amount of money gained from putting a dataset behind a paywall. While this is a difficult calculation to make, sharing data not only leads to immediate prestige, but cumulative payoffs as well. This comes in the form of citations, successful replications, and the remixing of datasets. A paywall is not only likely to limit the diffusion of the dataset in the scientific community [6], but limit the number of derivative works as well.

Scenarios and Payoffs

In modeling different scenarios for the use of open data, we must consider how the players (producer and transformer) interact. There are four basic interactions related to payoffs: producer gains and transformer loses; producer loses and transformer loses; producer gains and transformer gains; producer loses and transformer gains. Accumulation of social capital by both data producer and data transformer results in a mutual benefit for both parties, while mutual loss means that a loss of social capital by both data producer and data transformer. Sharing data for reuse is potentially a win-win transaction, but only if both parties recognize the payoff in social capital. As we will see in the three different scenarios highlighted in the analysis, payoffs are highly contingent on motivations and behaviors of both the data producer and the data transformer. The greatest potential payoff occurs when these are coordinated.

Data sharing as PD Game

We can model data sharing and recycling as an instance of the PD game. The basic payoff matrix (Table 1) can include various payoff structures for different access arrangements, but the basic set of interactions are the same. If both the producer and transformer agree to a CC-BY-NC (non-commercial) arrangement, then in principle their transactions result in a Nash equilibrium between players [7] and provides the basis for a sustainable open access community. As we will

bioRxiv preprint doi: https://doi.org/10.1101/093518; this version posted December 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

see, however, it is only one of many possible outcomes that depends on the culture in which both producer and transformer are embedded.

Table 1. Basic structure of the payoff matrix for interactions between producer and transformer. "Gain" and "Lose" payoff categories represent the social and financial utility for the act of engaging in the open exchange of data.

		Transformer	
		Gain	**Lose**
Producer	**Gain**	Gain/Gain	Gain/Lose
	Lose	Lose/Gain	Lose/Lose

Intuitively, one might assume that the relationship between producer and transformer is one that favors the transformer only. A more cynical worldview would describe the relationship as a parasitic one, with producer being the only creator of value. In light of this, let us consider the 2x2 payoff matrix as being representative of the potential for mutual gain (Table 2). As in the typical Prisoner's Dilemma scenario, each player act independently but according to a common set of principles. In this case, cooperation proceeds by both producer and transformer having been acculturated into the Open Science community.

Table 2. Payoff matrix if both the producer and transformer decide to share and reuse data, respectively. "Gain" and "Lose" payoff categories represent the social and financial utility for the act of engaging in the open exchange of data.

		Transformer	
		Gain	Lose
Producer	Gain	1	0
	Lose	0	-1

As is shown in Table 2, an asymmetric scenario in which one player loses utility for engaging in data sharing results in a payoff of zero. Figure 2 demonstrates the payoff matrix structure shown in Table 2. The most critical component of this payoff matrix involves the negative payoff when both producer and transformer lose utility by not participating in data sharing.

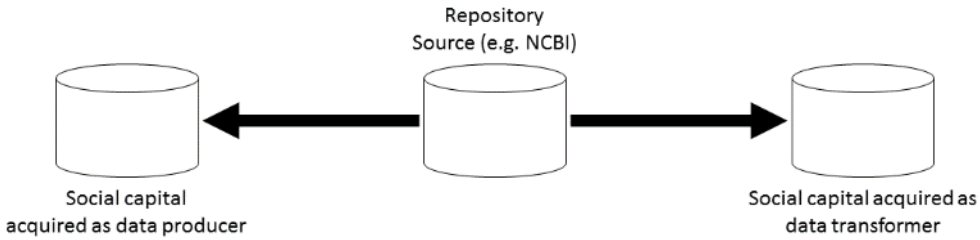

Figure 2. A payoff structure where data contributed to an open repository by the data producer can benefit both the primary producer and secondary transformer.

bioRxiv preprint doi: https://doi.org/10.1101/093518; this version posted December 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

While shared principles and easy-to-use infrastructure can allow for a Nash equilibrium, institutional skepticism about the Open Science movement can produce other outcomes. For example, let us suppose that while data sharing is considered a valuable activity, secondary use is discouraged. This is similar to the cynical view of data sharing mentioned earlier. Table 3 demonstrates the payoff matrix for such a scenario. In this scenario, there is a negative payoff in utility whenever the producer loses, and negative to no gain whenever the transformer gains.

Table 4 demonstrates what happens when sharing is discouraged entirely. As is the case in Table 3, there is no opportunity for mutual gain. In fact, the payoff matrix suggests that mutual loss is the only outcome with a payoff in utility (e.g. gain in social capital by the primary data producer). Of course, this is only true in a world where decision-makers and other people do not value Open Science. As with the other scenarios, the utility of data sharing stems from the value assigned to both the act of sharing data and rewarding the products of data sharing. These include published datasets, meta-analysis, and the reanalysis of data.

Table 3. Payoff matrix if both the producer and transformer decide to share and reuse data, but gains in utility by the transformer are discouraged. "Gain" and "Lose" payoff categories represent the social and financial utility for the act of engaging in the open exchange of data.

		Transformer	
		Gain	Lose
Producer	Gain	0	1
	Lose	-1	-1

Another feature of Tables 3 and 4 is the positive payoff for outcomes that discourage data sharing and reuse. As mentioned previously, specific payoffs are mediated by community standards, particularly when it comes to investments of time. When time is not allotted to activities such as data annotation, it can provide a motivation not to share information with the broader scientific community. In Table 3, this results in a gain for the producer (as time saved without motivation to share equals gained social capital) and a loss of utility for the transformer (as they cannot engage in their niche activity without data nor annotation information). In Table 4, this loss due to lack of proper social support results in a social capital payoff for a mutual loss of utility. In such a scenario, the transformer role is technically removed from the equation.

Table 4. Payoff matrix if both the producer is discouraged from sharing data (both in terms of institutional and infrastructural support). "Gain" and "Lose" payoff categories represent the social and financial utility for the act of engaging in the open exchange of data.

		Transformer	
		Gain	Lose
Producer	Gain	0	0
	Lose	0	1

bioRxiv preprint doi: https://doi.org/10.1101/093518; this version posted December 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Towards a CC-BY Economy

This paper contributes to our understanding of how a CC-BY (Creative Commons Attribution) economy might become sustainable, particularly with respect to properly valuing and rewarding openly-available research outputs. While some CC licenses limit the ability to remix and use for commercial purposes [8], the purpose of our model is to demonstrate the role of social capital in driving interactions between research parties. A more pessimistic worldview suggests substantial cultural change needs to occur before data sharing and secondary use becomes a dominant strategy. Indeed, institutional culture and technical infrastructure can affect the outcomes predicted by this analysis.

The greater question is what a CC-BY economy would look like, particularly in a time when the value of research outputs are being reconsidered. The cornerstone of a CC-BY economy rests upon the convertibility of social capital and intangible benefits of communal maintenance to some form of financial reward. A related factor is explicitly encouraging investments in annotating and the analysis of data. Another critical component of a CC-BY economy is greater recognition of the symbolic and semantic precursors of value. This ranges from recognizing the benefits of open access on the diffusion of research to changing the language of science to reflect desirable but undervalued contributions. Typically, components of the traditional academic economy come imbued with values that discourage open data, and thus the existing economic rewards and social practices within this economy have an inherent bias against Open Science practices [9, 10]. All too often, this leads us to the scenario in Tables 3 and 4 than the one in Table 2. It is up to us to shape the future, and the principles of Open Access enable that way forward.

REFERENCES

[1] Suber, P. (2012). Open Access. MIT Press, Cambridge, MA.

[2] Grix, J. (2010). Social Capital as a Concept in the Social Sciences: The Current State of the Debate. Democratization, 3, 189-210.

[3] Antocia, A., Saccob, P.L., Vanin, P. (2007). Social capital accumulation and the evolution of social participation. The Journal of Socio-Economics, 36(1), 128–143.

[4] Karsten, J. and West, D.M. (2016). The impact of open access scientific knowledge. Brookings Institute, January 11.

[5] Ball, P. (2015). The trouble with scientists. Nautil.us, May 14.

[6] Teplitskiy, M., Lu, G., Duede, E. (2016). Amplifying the Impact of Open Access: Wikipedia and the Diffusion of Science. Journal of the Association for Information Science and Technology, doi:10.1002/asi.23687.

[7] Moreno, D. and Wooders, J. (1996). Coalition-Proof Equilibrium. Games and Economic Behavior, 17(1), 80-112.

bioRxiv preprint doi: https://doi.org/10.1101/093518; this version posted December 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[8] Best Practices for Attribution. Creative Commons Wiki, https://wiki.creativecommons.org/wiki/Best_practices_for_attribution. Last accessed, December 6, 2016.

[9] Munafo, M. (2016). Open Science and Research Reproducibility. Ecancermedicalscience, 10, ed56.

[10] Nuzzo, R. (2015). How scientists fool themselves, and how they can stop. Nature News, October 7.