A RESEARCH DATA SHARING GAME

Pronk, T.E.[1], Wiersma, P.H., Weerden, A. van
University Library Utrecht, Heidelberglaan 3, Utrecht, the Netherlands
[1] Corresponding author: T.E.Pronk@uu.nl

Abstract

While reusing research data has evident benefits for the scientific community as a whole, decisions to archive and share these data are primarily made by individual researchers. For individuals, it is less obvious that the benefits of sharing data outweigh the associated costs, i.e. time and money. In this sense the problem of data sharing resembles a typical game in interactive decision theory, more commonly known as game theory.

Within this framework we analyse in this paper how different measures to promote sharing and reuse of research data affect sharing and not sharing individuals. We find that the scientific community can benefit from top-down policies to enhance sharing data even when the act of sharing itself implies a cost. Namely, if (almost) everyone shares, many individuals can gain a higher efficiency as datasets can be reused. Additionally, measures to ensure better data retrieval and quality can compensate for sharing costs by enabling reuse. Nevertheless, an individual researcher who decides not to share omits the costs of sharing. Assuming that the natural tendency will be to use a strategy that will lead to maximisation of individual efficiency it is seen that, as more individuals decide not to share, there is a point at which average efficiency for both sharing and non-sharing researchers becomes lower than was originally the case and scientific community efficiency steadily drops. With this in mind, we conclude that the key to motivate the researcher to share data lies in reducing the costs associated with sharing, or even better, turning it into a benefit.

Introduction

Science is driven by data and even more so now data collection has been enabled by new technologies. In addition, the use and reuse of data have been facilitated by techniques for data mining and analysis [Hanson et al., 2011; Levy et al., 2012]. Summing up all the reasons arguing that reuse of data is beneficial, it is obvious that making data widely available is an essential element of scientific research. Firstly, society relies on scientific data of diverse kinds; for example, in responding to disease outbreaks, managing resources, responding to climate change, and improving transportation [Hanson et al., 2011]. Secondly, sharing data enables the scientific community to benefit from a whole suite of novel possibilities. Sharing data opens access to and reinforces open scientific inquiry; encourages diversity of analysis and opinion; promotes new research; facilitates the education of new researchers; enables the exploration of topics not envisioned by the initial investigators; permits the creation of new data sets when data from multiple sources are combined; and it sets the stage for new experiments [Ascoli, 2007]. Thirdly, in terms of scientific quality and integrity, data underlying scientific publications can be assessed and replicated to check the scientific

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.599v1 | CC-BY 4.0 Open Access | rec: 8 Nov 2014, publ: 8 Nov 2014

results and conclusions [Hernan and Wilcox, 2009]. Lastly, if (re-)collection of data is minimized, use of resources is optimized and scientific efficiency is enhanced [Piwowar et al., 2011]. The efficiency of the scientific system is of key importance to ensure the competitiveness of a group, university, nation or region.

While sharing data has obvious group benefits for the scientific community and society, decisions to archive data are made by individual researchers, and it is less obvious that the benefits of sharing data outweigh the costs for all individuals [Tenopir et al., 2011]. Many researchers are reluctant to share their data publicly because of real or perceived individual costs [Roche et al., 2014; Pitt and Tang, 2013] which probably explains why sharing data is far from universal. Improving participation in sharing data will require lowering costs and/or increasing benefits for primary data collectors [Smith, 2009] [Roche et al., 2014]. Costs to individual researchers include the time investment, high costs (and lack of funding), the chance of being scooped by others on any future publications on the data, a chance on over-scrutinization of results from published papers, misinterpretation of data resulting in faulty conclusions [Atici et al., 2013], misuse [Bezuidenhout, 2013], and possible infringement of the privacy of test subjects [Antman, 2014]. Also, there is the perception that data is intellectual property and researchers simply don't want others to benefit from their hard-won data [Vickers, 2011]. In contrast, there are signs that sharing of research data confers an advantage. In a study of Piwowar and Vision [Piwowar and Vision, 2013] it was calculated that papers with open microarray data were cited, on average, nine percent more than studies without the data available. Belter [Belter, 2014] found an even higher number for three selected oceanographic datasets. Scientific reach might also be extended into other than the original research areas [Chao, 2011], and researchers' reputations could improve by good sharing practices, possibly initiating new collaborations. Moreover, there is a movement towards regarding datasets as full-fledged research output that can be cited in itself [Costello et al., 2013; Neumann and Brase, 2014]. This would mean that sharing data in the near future will have a direct positive influence on a researchers' scientific impact.

To summarize, the act of sharing data means either a benefit or a cost for the individual researcher, even though it could be of clear benefit to the scientific community as a whole in which, of course, the individual researcher also takes part. The problem of data sharing is therefore in essence a game-theoretical problem. Specifically, game theory is the study of mathematical models of conflict and cooperation between intelligent rational decision-makers. An assumption herein is that an individual will always try to maximize his or her gains *relative to the gains of others*. Here we have a framework to investigate the community gains versus the gains of the individual researcher in the competitive world of scientific research. For our analysis, we have constructed a simple model of a scientific community where researchers publish a certain amount of papers in a given year and have the habit either to share or not to share. With help of the model, we simulate the effect of sharing policies, explore several cost scenarios, and evaluate the overall benefits to the scientific community relative to the benefits of the individual researcher. Although this is a simple model, it enables us to assess these key principles. With the model we show how research data sharing fits in a game-theoretical framework. More importantly, we assess which measures to alter costs and benefits would turn the balance in a scientific community towards more sharing and more benefits from sharing, benefitting the community, society and the individual researcher.

Methods

The simulated scientific community

We construct a steady-state community of ten thousand researchers that each have published a certain amount of papers in a given year. To determine a distribution of published papers for an average scientific community we sampled the bibliographic database Scopus. We selected the first four papers for each of the 26 subject areas in Scopus-indexed papers, published in 2013. If a paper appeared within the first four in more than one subject area, it was replaced by the next paper in that subject area. For each of the selected papers we noted down all authors and checked how many papers each author (co-) authored in total in 2013. We came to 366 unique authors for our selected papers. Authors that were ambiguous, because they seemingly published many papers, were checked individually and excluded if it was a group of authors publishing under the same name with different affiliations between the papers. The distribution of papers that the selected authors published in 2013 is shown in Figure 1 (for the data see [Pronk et al., 2014]). This distribution, based on our sampling, implies that most researchers publish one paper in a year, declining fast down to a few researchers that publish many papers in a given year. We fitted an exponential distribution through the sampled population and take this as a basis for our simulated scientific community of 10.000 researchers.

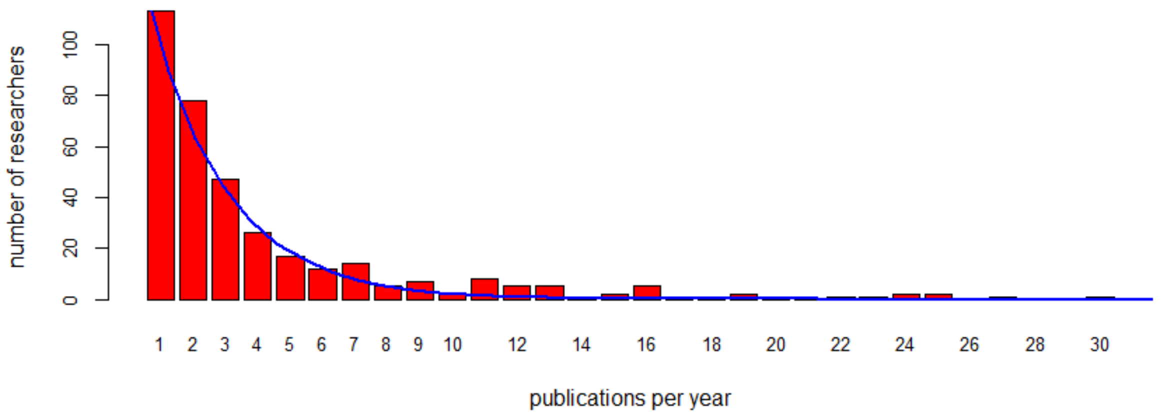


Figure 1. The sampled (bars) and fitted (line) distribution of published papers per researcher in a given year, in this case 2013. For reasons of visualisation the distribution is shown up to thirty publications, whereas the sampling sporadically included more publications per researcher. The fitted line is used as the publishing distribution for the simulated community.

Determinants for efficiency

We assume the goal for each researcher in our community is scientific efficiency and that this efficiency can be gained by (high-quality) publications. The researchers have the habit either to share or not to share the data (i.e. dataset) from all papers that they publish, as appointed by random selection of these researchers in the simulated community. We assume there is a certain probability for each researcher for each paper to find an appropriate dataset to improve that paper, resulting in higher efficiency papers. In the model these factors affecting efficiency are formalized into four different parameters (Table 1), namely the improved efficiency per paper '*e*' with the reuse of an external dataset, the chance of finding such an external dataset '*f*', the cost for sharing data per dataset '*c*', the percentage of sharing researchers '*r*' (Table 2). The standard values for these parameters, given in Table 2, are quite arbitrary as we do not know their true value. They are used here to resemble a situation in which the cost for sharing is relatively high compared to the possible efficiency gain with reuse of datasets. As such they function as a reference point for other, more profitable parameter settings that we will test. The following rules apply to the four determinants of efficiency:

- We assume that a paper is produced with a higher efficiency '*e*' in the case of reuse of external data. This is expressed as a percentage of improvement of efficiency per paper.
- We assume that there is a certain probability '*f*' that a researcher can find an appropriate external dataset that will be useable for his paper.
- We consider an offset '*c*', either a cost or benefit, when sharing the research data underlying a paper, as we want to simulate the consequences in both scenario's. Cost or benefit is expressed as a percentage offset from the total efficiency per researcher who shares data.
- We consider a range of percentages '*r*' of researchers sharing their datasets (ranging from 0 to 100%).

Table 1. Overview of parameters in the model determining scientific community efficiency and possible measures to improve this.

Parameters in the model	**Possible associated measures to improve the parameter in a real world situation**
Increased efficiency '*e*' of a paper with inclusion of an external dataset	• Improve data quality, for instance by the use of data journals, or peer review of datasets. • Offer techniques to easily assess the quality or other techniques to reuse datasets with less effort.
Chance '*f*' to find an external dataset	• Harvest databases through data portals to reduce 'scattering' of datasets. • Standardization of metadata-terms.

	• Advanced community and project-specific databases • Library assistance in finding and using appropriate datasets
Offset (cost or benefit) in efficiency '*c*' associated with sharing of research data	• Offer a good storing & sharing IT infrastructure. • Fund open data. • Increase attribution to datasets by citation rules and establish impact metrics for datasets.
Percentage '*r*' of scientists sharing their research data	• Promote sharing by a top down policy from an institute, funder, or journal. • Promote sharing bottom up by offering education on the benefits of sharing, to change researchers' mind set.

Table 2. Overview of all parameters and variables and their standard values in the model

Parameter	Meaning	Value
r	Percentage sharing researchers	From 0 to 1 (none to all sharing)
c	Sharing cost (efficiency offset per sharing researcher)	- 0.1
f	Probability of finding an appropriate dataset (per paper)	0.2
e	Improved efficiency (per paper)	0.2
P_r	Published papers (per researcher)	From distribution (Fig. 1)
P_t	Total number of published papers	~30130
E_s	Efficiency of sharing researchers	See Formula (1)
E_n	Efficiency of non-sharing researchers	See Formula (2)
E_t	Total efficiency of the scientific community	See Formula (3)

The actual efficiencies for researchers and the exact number of published papers in our simulation are subjected to some stochasticity from the random draw of the number of publications per researcher (from the exponential distribution, see Figure 1) and the random assignment of researchers who share their research data. The efficiency of any sharing researcher can on average be approached by

$$E_s = P_r \cdot (1 + f \cdot r \cdot e + c) \tag{1}$$

The efficiency of any non-sharing researcher can be approached by

$$E_n = P_r \cdot (1 + f \cdot r \cdot e) \tag{2}$$

since these researchers have no costs but do have the benefits of the shared datasets. Total efficiency of the scientific community is represented by

$$E_t = P_t \cdot (1 + r \cdot c + f \cdot r \cdot e) \quad (3)$$

If there is no sharing at all, the total efficiency E_t reduces to the amount of published papers P_t. In order to have benefits from sharing; we need to satisfy the following statement: $E_t > P_t$. In the case of a cost for sharing for individual researchers, a benefit from sharing (by reuse of datasets) for the community is achieved if:

$$f \cdot e > -c \quad (4)$$

In case of a benefit for sharing, i.e. '*c'* is positive, the efficiency will of course always increase with increased sharing of research.

Simulations

With the model as described in the previous paragraphs we simulate the efficiency of individual researchers at different cost scenarios, from which scientific community efficiency follows. First of all, we simulate a range of costs to benefits and sharing percentages, with offset (cost or benefit) ranging from -0.25 to 0.25 per shared dataset and sharing ranging from 0 to 100% of researchers, with otherwise standard parameters (Table 2). Secondly, we simulate the efficiencies at two levels of sharing: at low percentage ´*r*´ of sharing researchers (5%) and at high percentage ´*r*´ of sharing researchers (95%) and compare these contrasting scenarios. In these simulations, in addition to the standard parameter values, we assume a higher probability '*f*' of finding an appropriate paper, similarly a higher efficiency '*e*' per paper with a reused external dataset, and positive ´*c*´ with sharing a dataset. We show the results in different visualisations. For the R-scripts to generate these plots see [Pronk et al., 2014] .

Results

In Figure 2 we show results of the first simulation of the average efficiency for researchers over the community with different cost '*c*' (ranging from -0.25 to 0.25 per shared paper) and sharing rate '*r*' (ranging from 0 to 100% of researchers) with otherwise standard parameters (see Table 2). Cost '*c*' and sharing rate '*r*' are changed within their range in one hundred equal steps. It can be observed that the average efficiency for the community gradually goes up with costs changing from negative to positive. On the contrary, with an increase in percentage of sharing researchers, the increase or decrease of average community efficiency is dependent on the cost. If costs are relatively high the average community efficiency drops with more sharing instead of rises. Policies increasing sharing would in this case backfire and reduce scientific community efficieny. The point of balance between costs and benefits where there is no change in efficiency with a change in percentage of sharing researchers can, for any parameter setting, be deduced from Formula (4). For the parameters '*f*' and '*e*' as used in Figure 2 (see Table 2) this is at a cost of -0.04. It can be seen that with more profitable costs / benefits for sharing the average community efficiency increasingly starts to rise.

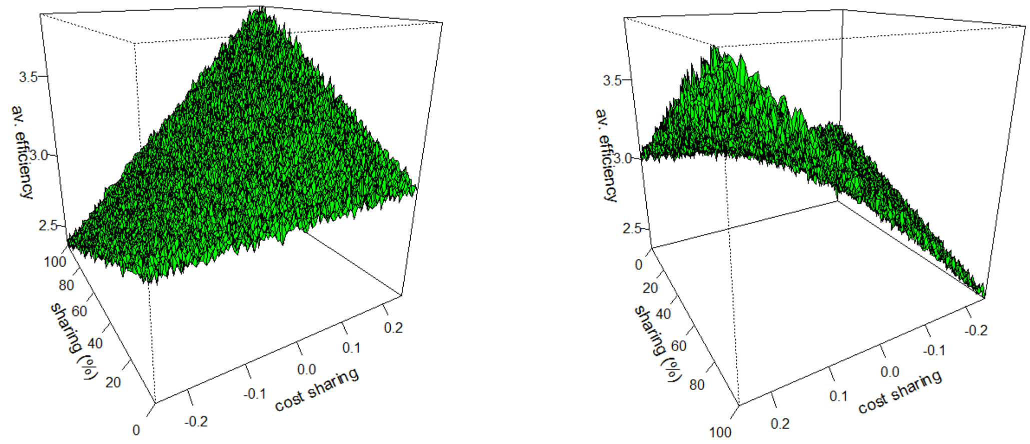


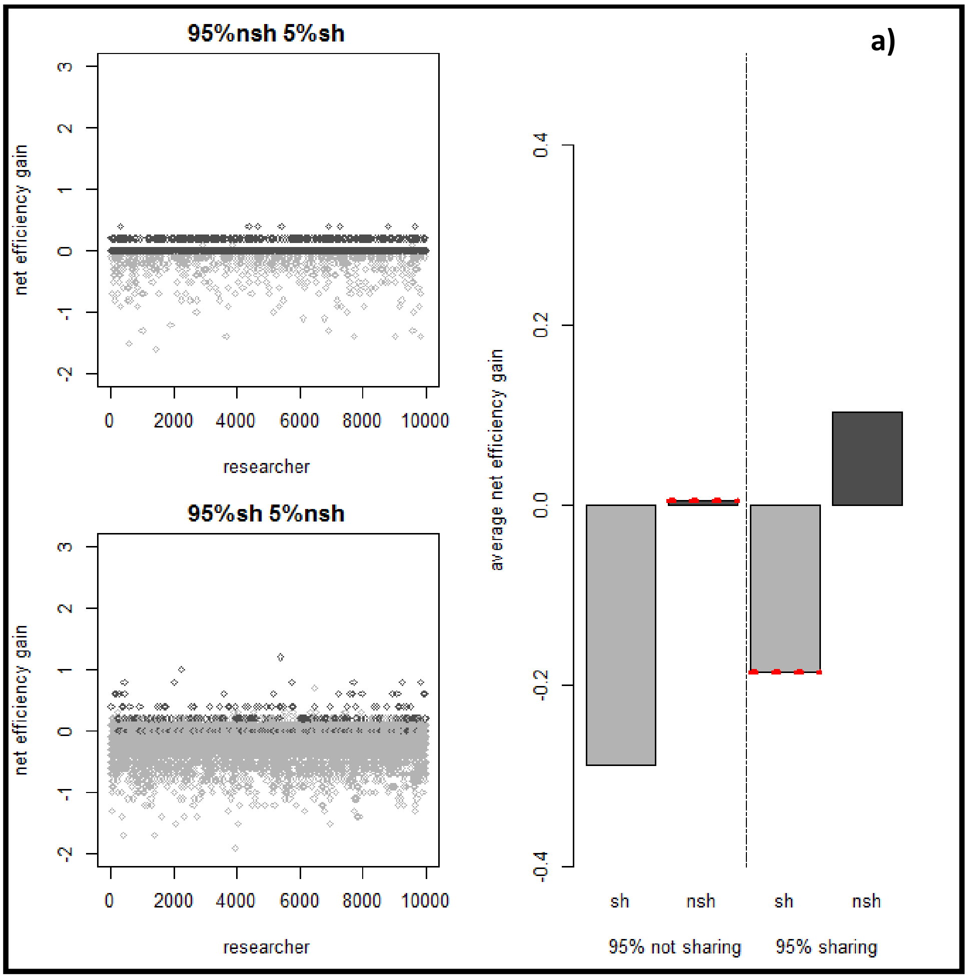
Figure 2. Shown here on the z-axis is the average efficiency per researcher in the simulated scientific community, simulated at standard parameter values (Table 2) and on the x and y axes changing costs for sharing (up to a benefit) from -0.25 to 0.25 and changing percentage of sharing researchers (sharing rate) from 0-100%. The same plot is shown from two perspectives: in the second plot rotated 180 degrees.

Of course, the community efficiency as depicted in Figure 2 is the average per researcher, while actually the simulated individual researchers have various efficiencies depending on their publication rate, reuse, and dataset sharing habits. In Figure 3 we show four simulations (a-d) that distinguish between sharing and non-sharing researchers in the community. Results for individual researchers are shown at 5% sharing (leaving 95% not sharing) (top left figure within each subfigure) and 95% sharing (leaving 5% not sharing) (bottom left figure within each subfigure). The bar plots within each subfigure provide the average community net efficiencies for sharing and not sharing researchers. Subfigure a) provides a reference at standard parameter values (Table 2). In subfigure (b) the efficiency per paper when reusing a dataset is increased from 0.2 to 0.8. In subfigure (c) the chance to find an appropriate dataset for reuse is increased from 0.2 to 0.8. In subfigure (d) the costs for sharing are turned into a benefit for sharing and is set from -0.1 to 0.1. In Table 1 we list a score of measures that could accomplish these effects in a 'real world' scientific community.

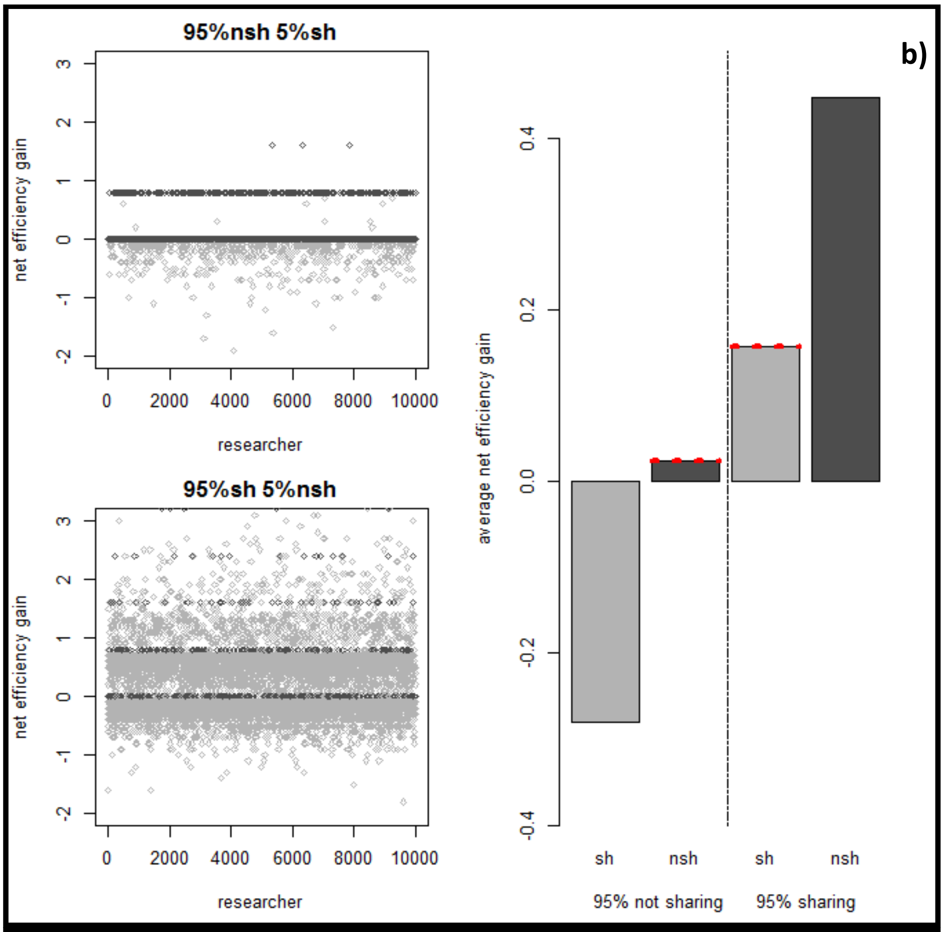
Subfigure a) shows that in a situation with costs higher than benefits, almost no individual sharing researcher has a higher net efficiency gain than a researcher that does not share. Subfigures b) and c) exemplify that, at low sharing levels, net efficiency gain for sharing researchers is negative for most of them. At high sharing levels, more have a positive net gain from the reuse of papers. It is notable that b) has more individual researchers with high costs than subfigure c), even though the average community average as seen in the bar plot, is the same. This is because in b) the gain in efficiency 'e' per paper is high, benefitting some, but for those that do not find a reusable set the costs for sharing remain uncompensated. In c) the probability of finding an appropriate dataset ‘f’ is very high, to compensate for the costs for sharing for more (almost all) of the sharing researchers.

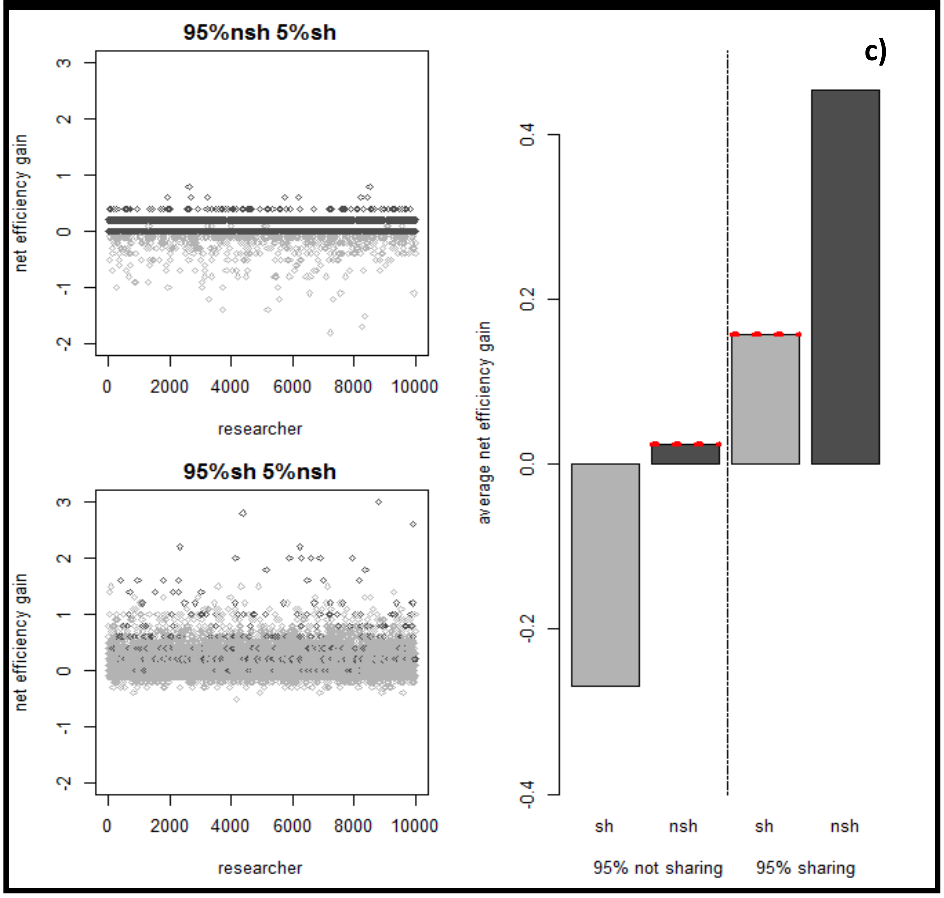
The bar plots in b) and c) indicate an intriguing result. The average efficiency of non-sharing researchers at low sharing drops below the average efficiency of sharing researchers at high sharing. This counterintuitive result implies that, even though *not sharing* is beneficial compared to sharing for the individual, there is a point after which *not sharing* can lead to a lower efficiency overall if more researchers adhere to this strategy.

In Appendix 1 more visualisations of these simulations are shown, with a focus on reusing and non-reusing researchers, high and low publishing researchers, the average costs and benefits for sharing researchers.



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.599v1 | CC-BY 4.0 Open Access | rec: 8 Nov 2014, publ: 8 Nov 2014


95%nsh 5%sh
b)
net efficiency gain
researcher
95%sh 5%nsh
net efficiency gain
researcher
average net efficiency gain
sh
nsh
sh
nsh
95% not sharing
95% sharing


95%nsh 5%sh
c)
net efficiency gain
researcher
95%sh 5%nsh
net efficiency gain
researcher
average net efficiency gain
sh
nsh
sh
nsh
95% not sharing
95% sharing

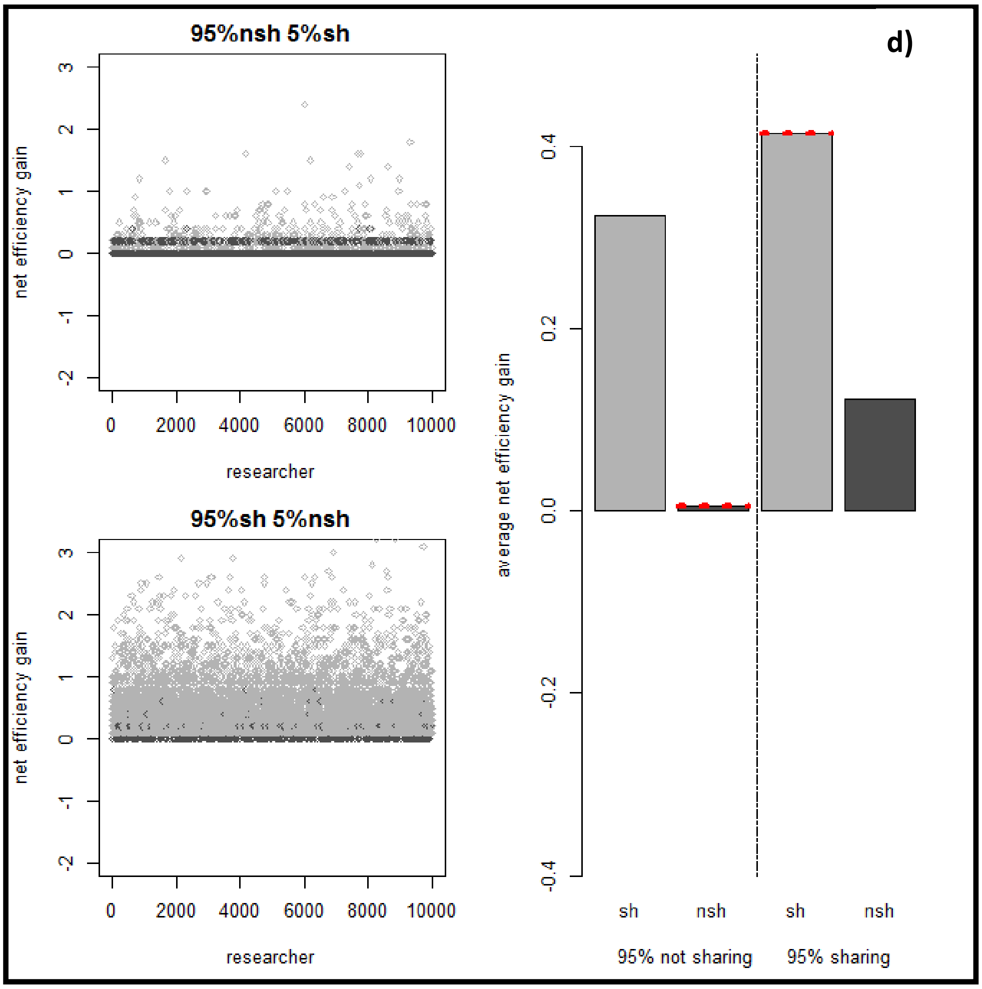


Figure 3. The net efficiency gain for individual researchers sharing and not sharing in the simulated community. Left in each subfigure a,b,c,d, are net gains per individual researcher at 5% sharing rate (top) and 95% sharing rate (bottom). Right in each subfigure are averaged net gains for sharing and non-sharing researchers at these sharing rates. Sharing researchers are light grey, not sharing researchers are dark grey. Dots at the top of a bar emphasise that the average is for **95%** of the researchers. a) Costs are relatively high compared to benefits (parameters as in Table 2). b) efficiency ´*e*´ from reusing data is raised to 0.8. c) The probability ´*f*´ to find an appropriate dataset is raised to 0.8. d) Cost ´*c*´ for sharing data is raised to 0.1, turning sharing to a benefit.

Discussion

The strength of game theory is the methodology it provides for structuring and analysing problems of strategic choice. Constructing such a model thus already has the potential of providing a clearer and broader view of the situation as the players, their strategic options, and the external factors of influence on those decisions have to be made explicit. In this paper we use game theory as a prescriptive application, with the goal of improved strategic decision making. This could help prioritizing measures that could accomplish advantageous effects for scientific efficiency in a 'real world' scientific community.

We analysed the effect of sharing and not sharing data on the scientific community efficiency, relative to the efficiency of the individual researcher. In our simulations we assume a number of parameters that can be of influence on share-rate and reuse-rate and, with that, on the efficiency of individual researchers and that of the community as a whole. These parameters are: the percentage of sharing researchers, the efficiency gain in producing a high quality paper when reusing a dataset, the probability of finding an appropriate dataset, and the costs associated with sharing data. In Table 1 of this paper we address these parameters and measures that could improve these parameters in a 'real world' scientific community [Chan et al., 2014]. With the result from our simulations we can assess and prioritize these measures.

Results show that in the case of moderate costs associated with sharing, sharing research data can still lead to a general higher community efficiency. This is because of the supposition that the more research data is shared, the more can be reused and as a result (high quality-) papers are more efficiently produced. However, an individual researcher can decide to reuse the datasets provided by others, and omit the sharing costs as indicated in the introduction of this paper. If *everyone* should adopt this strategy, *everyone* is worse off and average efficiency for both sharers and non-sharers declines. Efficiency at some point even drops below a level that was the efficiency when the researcher was sharing in the original situation. This means that in in the end, nobody benefits from the decision not to share. This counterintuitive result implies that for an individual, even though *not sharing* is beneficial compared to sharing, *not sharing* can lead to a lower efficiency overall if more researchers adhere to this strategy.

We show that policies to enforce higher percentages of sharing researchers could increase community efficiency. Policies can be enforced on the level of institutions, funders, or journals. In several studies on the public availability of published research data, journal policy stating data should be made available with a publication was (not yet) apt to convince researchers to actually make their data publicly available. Between different studies, the raw data availability rate differed from 9% to 41% of papers adhering to journal policies [Wicherts et al., 2006; Alsheikh-Ali et al., 2011; Vines et al., 2013; Savage and Vickers, 2009]. This could be exemplary for the reluctance of individual researchers to share data because of real or perceived costs. This could mean that, even though policy measures could increase community efficiency in theory, the problem of costs for sharing individuals and consequent reluctance to share are not addressed.

Therefore, another solution is to compensate for sharing costs for individuals. This can be done by increasing the benefits with reusing available data for individual researchers. In this way sharing costs are indirectly compensated for. We analysed these by two measures: increasing the data quality so datasets can be reused with less effort and increasing findability of datasets. To improve quality, many archives now provide the opportunity for researchers to comment on the deposited dataset. Data journals are another means to ensure high data quality by peer review and strict data preparation guidelines [Costello et al., 2013; Atici et al., 2013; Gorgolewski et al., 2013]. Although this is an important and valid measure, results show that as a single measure this has a lesser impact if only a few researchers can profit from this. It would be more important to take measures to improve the findability of datasets. Datasets are scattered across different archives and metadata is minimal and not standardized, making it difficult to retrieve appropriate datasets. Our results show that with an improved findability for datasets more sharing researchers acquire a net positive efficiency. This could be an effective means to compensate for sharing costs in a community where sharing is common.

Another simulated measure is to reduce the costs with sharing or even turning it into a benefit for the individual sharing researcher [He et al., 2013; Roche et al., 2014]. When it comes to sharing data, in practice researchers are hesitant because of real or perceived costs associated with sharing, as pointed out in our introduction. Not much effort has been done to quantify these costs [Roche et al., 2014]. Nevertheless, as long as there is a cost associated with sharing data, the researcher that has the strategy ‘reuse-don't share’ will have the highest efficiency in the scientific community. Especially the high-publishing scientists will fall under this category as they potentially have higher costs in sharing all their datasets. This is troublesome as these researchers have a relatively high influence on the reuse-rate within the community because of the high number of papers with underlying datasets that they themselves could make available. The 'reuse-don't share' strategy is a true current sentiment towards using: according to a survey in 2011 of about 1,300 scientists, more than 80 percent said they would use other researchers’ data sets. At the same time there were a relatively small number of scientists who wanted to make their data electronically available to others, for a variety of reasons [Tenopir et al., 2011]. In contrast, when data sharing incurs a benefit for the individual researcher, the researcher that has the strategy ‘reuse-share’ will have the highest efficiency in the scientific community. If we again assume that the natural tendency will be to use any strategy that will lead to maximisation of individual efficiency, a benefit with sharing data will automatically lead to a higher efficiency of the community as a whole. With the improvement of benefits and reduction of costs for the individual researchers, the balance will shift more naturally towards more sharing, benefitting the scientific community and therewith society. This would be a better mechanism to promote sharing than simply imposing an obligation to share by funders, institutes, or journals. Better incentives arguably also lead to better sharing practices.

With our model we derived general phenomena for the scientific community, whereas (perceived) costs and benefits with sharing in reality will differ between scientific communities. This means that the measures taken for each scientific community to make sharing worthwhile will have to differ in their focus between them [Borgman et al., 2007; Acord and Harley, 2013]. For instance, standardization of data and metadata is easier in some disciplines, such as genomics, then it is in others [Acord and Harley, 2013]. Moreover, attitudes towards sharing can differ between disciplines. For instance, surveys revealed that in pharmaceutical research, sharing is opposed by the larger part (75%) of researchers [Vickers, 2011], while in biodiversity research most researchers are positive towards sharing their article-related data [Huang et al., 2012]. Also forensic geneticists are more willing to make their data available than evolutionary or medical geneticists, there being quite a difference (6% and 23%, respectively) [Anagnostou et al., 2013]. Possible explanations given for this particular difference are the policies for data sharing by the two most important forensic journals. Plus, ''familiarity'' and collaborative spirit among investigators increase their predisposition towards sharing [Pitt and Tang, 2013; Anagnostou et al., 2013].

Lastly, not all data can or should be made fully or immediately publicly available for a variety of practical reasons (e.g., lack of interest, sheer volume and lack of storage, cheap-to-recreate data, the need of specialist software to access data, want to publish later perhaps, patents pending) [Cronin, 2013]. For instance, in some disciplines, the amount of data grows faster than the financial and technical means of sharing it, causing problems of scale and data deluge [Kim, 2013]. With our simulations we show that if costs for sharing are too high relative to the benefits of reuse, in theory sharing policies to increase sharing could even backfire and reduce scientific community efficiency. It should be carefully considered whether the alleged benefits of storage for the scientific community will outweigh the costs for each data type and set. For easily obtainable data such as the data underlying this paper, recreating it is probably cheaper than storing and interpreting the datasheet.

In conclusion, we performed a game-theoretic analysis to provide structure and to analyse problems of strategic data sharing. While increasing benefits with sharing will have the most positive influence on the efficiency of both the individual researcher and the scientific community, we showed that in the case of moderate costs, sharing research data can still lead to a general higher scientific community efficiency as a result of efficient data reuse. An intriguing result is that although for the individual researcher *not sharing* is beneficial compared to sharing, *not sharing* can lead to a lower efficiency for all researchers in the community if more than a certain ratio of all researchers adhere to this strategy. Although policies should be able to increase the rate of sharing researchers, and increased findability and data quality could partly compensate for costs, a better measure would be to lower the costs for sharing, or even turn them into a benefit.

Acknowledgements

We thank Dorinne Raaimakers, Jeroen Bosman, and Jan Molendijk from the University Library Utrecht and Mark van Oorschot from PBL, RIVM for their constructive ideas concerning the manuscript and initial concept.

REFERENCES

Acord, S. K. and D. Harley (2013), Credit, time, and personality: The human challenges to sharing scholarly work using Web 2.0, New Media and Society, 15(3), 379-397, doi:10.1177/1461444812465140.

Alsheikh-Ali, A. A., W. Qureshi, M. H. Al-Mallah, and J. P. Ioannidis (2011), Public availability of published research data in high-impact journals, PLoS One, 6(9), e24357, doi:10.1371/journal.pone.0024357 [doi].

Anagnostou, P., M. Capocasa, N. Milia, and G. D. Bisol (2013), Research data sharing: Lessons from forensic genetics, Forensic. Sci. Int. Genet., 7(6), e117-9, doi:10.1016/j.fsigen.2013.07.012 [doi].

Antman, E. (2014), Data sharing in research: benefits and risks for clinicians, BMJ, 348, g237, doi:10.1136/bmj.g237 [doi].

Ascoli, G. A. (2007), Successes and rewards in sharing digital reconstructions of neuronal morphology, Neuroinformatics, 5(3), 154-160, doi:NI:5:3:154 [pii].

Atici, L., S. W. Kansa, J. Lev-Tov, and E. C. Kansa (2013), Other People's Data: A Demonstration of the Imperative of Publishing Primary Data, J. Archaeol. Method and Theory, 20(4), 663-681, doi:10.1007/s10816-012-9132-9.

Belter, C. W. (2014), Measuring the value of research data: a citation analysis of oceanographic data sets, PLoS One, 9(3), e92590, doi:10.1371/journal.pone.0092590 [doi].

Bezuidenhout, L. (2013), Data sharing and dual-use issues, Sci. Eng. Ethics, 19(1), 83-92, doi:10.1007/s11948-011-9298-7 [doi].

Borgman, C. L., J. C. Wallis, and N. Enyedy (2007), Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries, Int. J. Digital Libr., 7(1-2), 17-30, doi:10.1007/s00799-007-0022-9.

Chan, A. W., F. Song, A. Vickers, T. Jefferson, K. Dickersin, P. C. Gotzsche, H. M. Krumholz, D. Ghersi, and H. B. van der Worp (2014), Increasing value and reducing waste: addressing inaccessible research, Lancet, 383(9913), 257-266, doi:10.1016/S0140-6736(13)62296-5 [doi].

Chao, T. C. (2011), Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences, Proc. ASIST Ann. Meet., 48, doi:10.1002/meet.2011.14504801125.

Costello, M. J., W. K. Michener, M. Gahegan, Z. -. Zhang, and P. E. Bourne (2013), Biodiversity data should be published, cited, and peer reviewed, Trends Ecol. Evol., 28(8), 454-461, doi:10.1016/j.tree.2013.05.002.

Cronin, B. (2013), Thinking about data, J. Am. Soc. Inf. Sci. Technol., 64(3), 435-436, doi:10.1002/asi.22928.

Gorgolewski, K. J., D. S. Margulies, and M. P. Milham (2013), Making data sharing count: a publication-based solution, Front. Neurosci., 7, 9, doi:10.3389/fnins.2013.00009 [doi].

Hanson, B., A. Sugden, and B. Alberts (2011), Making data maximally available, Science, 331(6018), 649, doi:10.1126/science.1203354.

He, S., M. Ganzinger, J. F. Hurdle, and P. Knaup (2013), Proposal for a data publication and citation framework when sharing biomedical research resources, Stud. Health Technol. Inform., 192, 1201.

Hernan, M. A. and A. J. Wilcox (2009), Epidemiology, data sharing, and the challenge of scientific replication, Epidemiology, 20(2), 167-168, doi:10.1097/EDE.0b013e318196784a [doi].

Huang, X., B. A. Hawkins, F. Lei, G. L. Miller, C. Favret, R. Zhang, and G. Qiao (2012), Willing or unwilling to share primary biodiversity data: Results and implications of an international survey, Conserv. Lett., 5(5), 399-406, doi:10.1111/j.1755-263X.2012.00259.x.

Kim, J. (2013), Data sharing and its implications for academic libraries, New Libr. World, 114(11), 494-506, doi:10.1108/NLW-06-2013-0051.

Levy, M. A., J. B. Freymann, J. S. Kirby, A. Fedorov, F. M. Fennessy, S. A. Eschrich, A. E. Berglund, D. A. Fenstermacher, Y. Tan, X. Guo, T. L. Casavant, B. J. Brown, T. A. Braun, A. Dekker, E. Roelofs, J. M. Mountz, F. Boada, C. Laymon, M. Oborski, and D. L. Rubin (2012), Informatics methods to enable sharing of quantitative imaging research data, Magn. Reson. Imaging, 30(9), 1249-1256, doi:10.1016/j.mri.2012.04.007 [doi].

Neumann, J. and J. Brase (2014), DataCite and DOI names for research data, J. Comput. Aided Mol. Des., doi:10.1007/s10822-014-9776-5 [doi].

Pitt, M. A. and Y. Tang (2013), What should be the data sharing policy of cognitive science? Top. Cogn. Sci., 5(1), 214-221, doi:10.1111/tops.12006 [doi].

Piwowar, H. A. and T. J. Vision (2013), Data reuse and the open data citation advantage, PeerJ, 1, e175, doi:10.7717/peerj.175 [doi].

Piwowar, H. A., T. J. Vision, and M. C. Whitlock (2011), Data archiving is a good investment, Nature, 473(7347), 285, doi:10.1038/473285a [doi].

Pronk, T.E., Wiersma, P.H., Weerden, van A., (2014) Replication data for: A RESEARCH DATA SHARING GAME, http://hdl.handle.net/10411/20328 [Version1]
Roche, D. G., R. Lanfear, S. A. Binning, T. M. Haff, L. E. Schwanz, K. E. Cain, H. Kokko, M. D. Jennions, and L. E. Kruuk (2014), Troubleshooting public data archiving: suggestions to increase participation, PLoS Biol., 12(1), e1001779, doi:10.1371/journal.pbio.1001779 [doi].
Savage, C. J. and A. J. Vickers (2009), Empirical study of data sharing by authors publishing in PLoS journals, PLoS ONE, 4(9), doi:10.1371/journal.pone.0007078.
Smith, V. S. (2009), Data publication: towards a database of everything, BMC Res. Notes, 2, 113-0500-2-113, doi:10.1186/1756-0500-2-113 [doi].
Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame (2011), Data sharing by scientists: practices and perceptions, PLoS One, 6(6), e21101, doi:10.1371/journal.pone.0021101 [doi].
Vickers, A. J. (2011), Making raw data more widely available, BMJ, 342, d2323, doi:10.1136/bmj.d2323 [doi].
Vines, T. H., R. L. Andrew, D. G. Bock, M. T. Franklin, K. J. Gilbert, N. C. Kane, J. S. Moore, B. T. Moyers, S. Renaut, D. J. Rennison, T. Veen, and S. Yeaman (2013), Mandated data archiving greatly improves access to research data, FASEB J., 27(4), 1304-1308, doi:10.1096/fj.12-218164 [doi].
Wicherts, J. M., D. Borsboom, J. Kats, and D. Molenaar (2006), The poor availability of psychological research data for reanalysis, Am. Psychol., 61(7), 726-728, doi:10.1037/0003-066X.61.7.726.

Appendix 1.

The figures in Appendix 1 are the results of simulations at several parameter values with sharing varied in each simulation from 0 to 100% researchers sharing. Other parameter settings are as in the simulations for Figure 2. The figure consists of four results in columns: 1) the community efficiency, 2) average efficiency per paper of researcher that did and did not find datasets to reuse, 3) average efficiency per paper of researchers that did find datasets to reuse, divided in high and low publishing researchers, 4) the average costs and benefits for a sharing researcher. For reasons of illustration for the point at which costs equal benefits, the cost is depicted positive where it is negative and vice versa.

Column 1: In the first simulation (a) we see the community efficiency decline with an increase in sharing. The costs for sharing outweigh the benefits and sharing has a negative impact on the whole. In the second (b) and third (c) and fourth (d) simulation, we see the community efficiency increase with sharing. This was accomplished in (b) by increasing the efficiency per paper when reusing a dataset. In (c) this was accomplished by increasing the chance to find an appropriate dataset for reuse. In (d) this was accomplished by turning the costs for sharing into a benefit for sharing. In Table 1 we list a score of measures that could accomplish both effects in a ‘real world’ scientific community.

Column 2: This column shows the efficiencies per publication for data reusing and non-data reusing researchers. To recall, in our model the papers for which a reusable set is found are appointed by chance. If ‘e’ is set to a high value in b), the average benefit of reuse is higher. The benefit increases relatively with more researchers sharing data. Efficiency of researchers who do not reuse data declines because part of these researchers do share their data, while there is no benefit of reuse.

Column 3: This column shows the efficiency, for data reusing researchers only. The high publishing researchers benefit the most from the availability of sets in any of the simulations. On average they have a higher efficiency per paper. This is because the probability of encountering a good set for any of their many publications is larger. Of course for non-reusing researchers, there is no difference between efficiency per paper for high and low publishing researchers so we do not show them.

Column 4: This column shows the costs and benefits for sharing researchers. In simulation b) and c) there is a point after which the benefits of reuse outweigh the costs for sharing. The benefits of reuse increase with the number of sharing researchers. There is no difference for sharing researchers between high and low publishing researchers, as both high and low publishing researchers have a cost or benefit as a percentage of their publications.

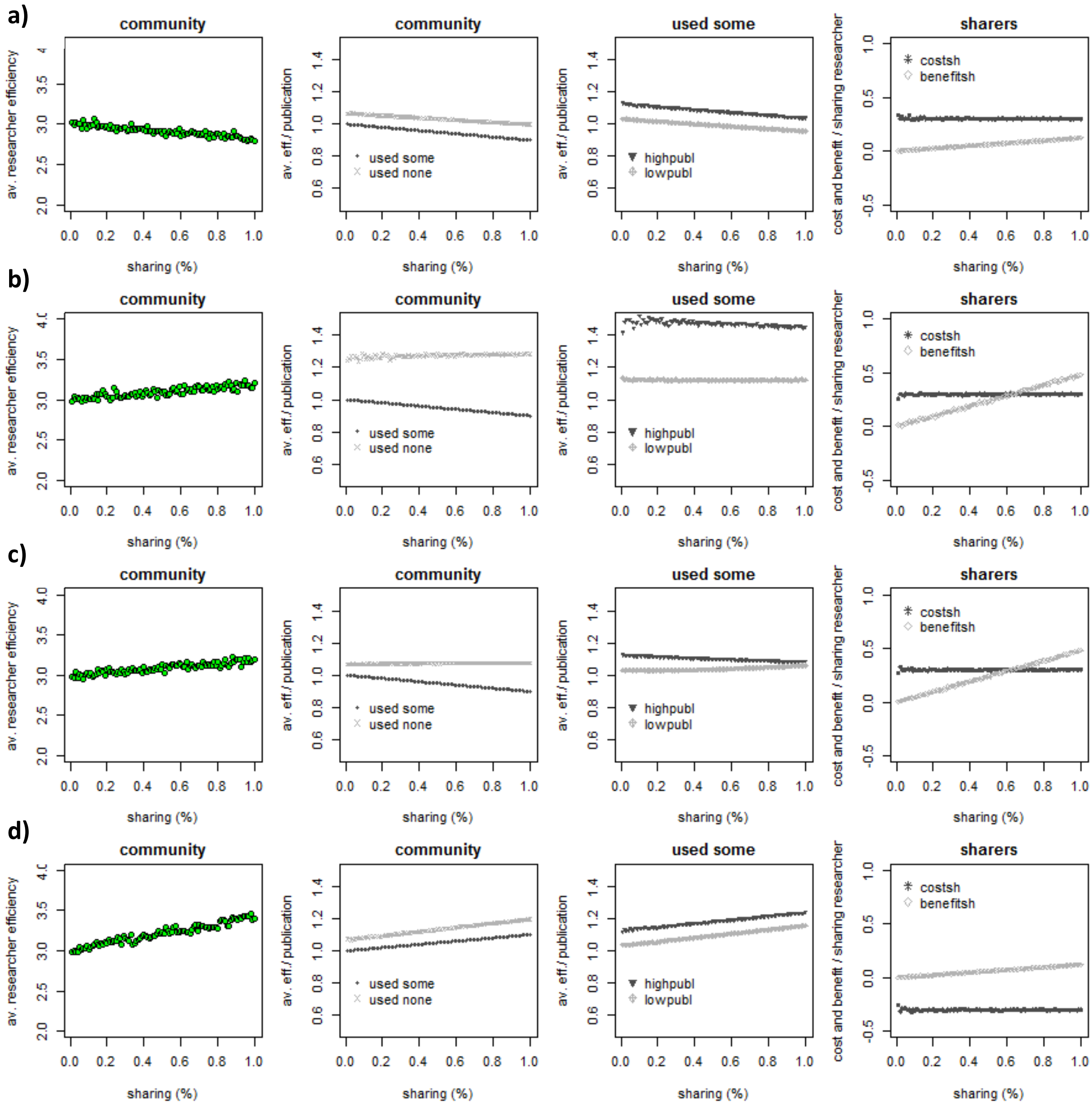


Figure 4. Simulation of average efficiencies per researcher in the scientific community with increased sharing (0 to 100% of researchers) with associated cost (a-c) and with associated benefit (d) to sharing. (a) gives the situation at default values (see Table 2). (b) with higher benefit attached to reuse of a dataset (c) with a higher probability of finding a dataset for reuse. (d) with a benefit to

sharing research data instead of a cost. Abbreviations: ‘sharers’ : researchers that share research data. 'community': all researchers belong to the scientific community. ‘used some’: a researcher that has reused at least one dataset to improve a paper. ‘used none’: a researcher that has not reused a dataset. ‘highpubl’: a researcher that has published 3 or more papers in a year. ‘lowpubl’: a researcher that has published less than 3 papers in a year. 'costsh': the costs for sharers. 'benefitsh': the gains (by data reuse) for sharing researches.