A game theoretic analysis of research data sharing

Tessa E Pronk, Paulien H Wiersma, Anne van Weerden, Feike Schieving

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. In this paper we analyse, within a game theoretical framework, how sharing and reuse of research data affect individuals who share or do not share their datasets. We construct a model in which there is a cost associated with sharing datasets whereas reusing such sets implies a benefit. In our calculations conflicting interests appear for researchers. Individual researchers are *always* better off not sharing and omitting the sharing cost, at the same time both sharing and not sharing researchers are better off if (almost) all researchers share. Namely, the more researchers share, the more benefit can be gained by the reuse of those datasets. We simulated several policy measures to increase benefits for researchers sharing or reusing datasets. Results point out that, although policies should be able to increase the rate of sharing researchers, and increased discoverability and dataset quality could partly compensate for costs, a better measure would be to directly lower the cost for sharing, or even turn it into a (citation-) benefit. Making data available would in that case become the most profitable, and therefore stable, strategy. This means researchers would willingly make their datasets available, and arguably in the best possible way to enable reuse.

A GAME THEORETIC ANALYSIS OF RESEARCH DATA SHARING

Tessa E. Pronk, [1]*
Paulien H. Wiersma*
Anne van Weerden*
Feike Schieving[#]

*Utrecht University Library , Utrecht University, Heidelberglaan 3, Utrecht, the Netherlands

[#] Ecology and Biodiversity, Utrecht University, Padualaan 8, Utrecht, the Netherlands

[1] Corresponding author: T.E.Pronk@uu.nl

Introduction

While sharing datasets has group benefits for the scientific community and society as a whole, decisions to archive datasets are made by individual researchers. It is less obvious that the benefits of sharing outweigh the costs for all individuals [Tenopir et al., 2011; Roche et al., 2014]. Many researchers are reluctant to share their dataset publicly because of real or perceived individual costs [Pitt and Tang, 2013]. This probably explains why sharing datasets is no daily practice [Roche et al., 2014], especially when compared to sharing knowledge and information in the form of a scientific paper. Costs to individual researchers include time investment, money, the chance of being scooped by others on any future publications on the dataset, a chance that results from published papers will be over-scrutinized, 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, datasets are perceived as intellectual property and researchers simply do not want others to benefit from it [Vickers, 2011].

In contrast, the act of sharing research data could have advantageous consequences. Scientific outreach might be extended into other than the original research areas [Chao, 2011], and researchers' reputations could grow by the publicity of good sharing practices, possibly initiating new collaborations. In genetics [Botstein, 2010; Piwowar and Vision, 2013] it was calculated that papers with open data were cited more than studies without the data available. This citation advantage was also found in other disciplines like astronomy [Henneken E.A., 2011; Dorch, 2012] and oceanography [Sears, 2011]. As citations to papers for many disciplines are a the key metric by which impact of researchers is measured, this could mean a very important incentive to researchers for sharing their data. Moreover, there is a tendency to regard datasets as research output that can be used as a citeable reference or source in their own right [Costello et al., 2013; Neumann and Brase, 2014]. For the field of oceanography it was found that datasets can be cited even more than most papers [Belter, 2014]. This would mean that sharing datasets in the near future could have a direct positive influence on a researcher's scientific impact.

On the other side of the coin, a researcher who reuses a dataset that was shared can gain several advantages. Time is saved in not having to collect or produce the data, which can be put to use to produce more papers. Papers can be enhanced with a comparison or meta-analysis based on an extra dataset. If the added dataset merits publication in a higher impact journal, the paper could be cited more often. In more general terms, the scientific community can benefit from reuse of datasets. Sharing data enables open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, facilitates the education of new researchers, enables novel applications to data not envisioned by the initial investigators, permits the creation of new datasets when data from multiple sources are combined, and provides a basis for new experiments [Ascoli, 2007; Kim, 2013; Pitt and Tang, 2013]. It also is a way to prevent scientific fraud; with the dataset provided one should be able to reproduce scientific results.

To summarize, data sharing implies costs and/or benefits for the individual researcher, but are of clear benefit to researchers that reuse the dataset, and to the scientific community as a whole. In this context, the problem of data sharing can be studied as a game theoretic problem. The strength of game theory lies in the methodology it provides for structuring and analysing problems of strategic choice. The players, their strategic options, the external factors of influence on those decisions, all have to be made explicit. With the model we show how research data sharing fits the definition of a typical 'tragedy of the commons', in which cooperating is the best strategy but cheating is the evolutionary stable strategy. In addition, we assess measures for altering costs and benefits with sharing and reuse and analyse how each measure would turn the balance towards *more* sharing and *more* benefits from sharing, benefitting the community, society and the individual researcher.

Methods

A Model for Impact

We assume a community of researchers who publish papers. We consider two types of researchers: those sharing and not sharing research data associated with those papers. We make the simplifying assumption that the goal for both types of researchers is to perform well by making a significant contribution to science, i.e. to have a large impact on science. We assume that produced papers, P_s for sharers and P_{ns} for non sharers, create impact by getting cited a number of times c. We assume c is constant, which means we do not distinguish between low and highly cited papers. To increase their performance, researchers need to be efficient, i.e. they should try to minimize the time spent on producing a paper, so more papers can be produced within the same timeframe. Papers from which the dataset is shared gain an extra citation advantage, increasing the impact of that paper by a factor b. In our model we consider only papers with a dataset as a basis, i.e. no review or opinion papers. So, the performance of researchers is expressed as an impact rate, in terms of citations per year, i.e. the impact for sharing and non-sharing researchers is defined as

$$E_s = P_s \cdot c \cdot (1+b) \qquad E_{ns} = P_{ns} \cdot c \tag{1}$$

From the above expressions it is clear that the difference in impact between sharing and not sharing researchers is to a large extent dependent on the number of publications P per year. These publications can be expressed in terms of an average time to write a paper T_s for sharers and T_{ns} for not sharers.

$$P_s = \frac{1}{T_s} \qquad P_{ns} = \frac{1}{T_{ns}} \tag{2}$$

The time *T* consists of several elements that we make explicit here. Each paper costs time t_a to produce. Producing the associated dataset costs a certain time t_d. Sharing a dataset implies a time cost t_c. We do not distinguish between large and small efforts to prepare a dataset for sharing; all datasets take the same amount of time. We assume there is a certain probability *f* to find an appropriate dataset for a paper from the pool of shared datasets *X*, in which case the time needed to produce a dataset t_d is avoided. We do acknowledge that some time is needed for a good 'getting to know' the external dataset and to process it, resembled in the time cost t_r. We calculate the time to produce a paper by

$$T_s = t_a + \frac{t_d}{1+f\cdot X} + \left(t_r - \frac{t_r}{1+f\cdot X}\right) + t_c \qquad T_{ns} = t_a + \frac{t_d}{1+f\cdot X} + \left(t_r - \frac{t_r}{1+f\cdot X}\right) \tag{3}$$

In these formulae, the pool of available datasets X determines the value of the terms with t_d and t_r. When *X* is close to zero, the term with t_d approaches t_d. This implies that everybody has to produce their own dataset with time cost t_d. In contrast, when *X* is very large the term approaches zero, implying almost everyone can reuse a dataset and almost no time is spent in the community to produce datasets. Between these two extremes, the term first rapidly declines with increasing *X* and then ever more slowly approaches zero (see the plots in the last column in the figure in Appendix 2). This is under the assumption that at a small number of available datasets, adding datasets will have a profound influence on the reuse possibilities. If datasets are already superfluous, adding extra datasets will have less influence on the reuse rate. The term representing the effort to reuse a paper t_r works opposite to the term representing t_d. When *X* is close to zero, the term approaches zero, implying nobody spends time to prepare a set for reuse. When *X* is very large the term approaches t_r; everyone spends this time because everyone has found a set for reuse.

While the pool of datasets *X* determines the values of the terms with t_d and t_r and with that the number of shared datasets, at the same time the shared datasets accumulate in the pool of shared datasets *X*. To come to a specification of this pool size *X* we formulate a differential equation for the pool size. A change in the pool of available, shared datasets *X* depends on adding datasets belonging to papers P_s from sharing researchers Y_s , minus the decay $q_x \cdot X$ of the datasets. Such a decay rate could be a result from a fixed storage time after which datasets would be disposed of or by a loss of data value, for instance by outdated techniques.

$$\frac{dX}{dt} = Y_s \cdot P_s - q_x \cdot X \tag{4}$$

Using Formula (2) and (3) with the system at steady state i.e. $dX/dt = 0$, the pool size X as function of the publication parameters and the size of the group of sharing researchers is given by

$$X = \frac{-\left(q_x\left(t_a + t_c + t_d\right) - Y_s f\right) + \sqrt{\left(q_x\left(t_a + t_c + t_d\right) - Y_s f\right)^2 - 4\left(q_x \cdot f\left(t_a + t_c + t_r\right)\right)\cdot\left(-Y_s\right)}}{2\left(q_x f\left(t_a + t_c + t_r\right)\right)} \quad (5)$$

(Formula (5) is derived in Appendix 1). So, for each parameter setting, we calculate X, and consequently, we calculate the impact in terms of citation rates E_s and E_{ns} with Formulae (1-3). Table 1 gives the default parameter settings that we use for our simulations.

An Individual Based Model

In addition to the model for impact we set up an individual based model to assess the impact for individual researchers depending on their personal publication rate, sharing and reuse habits, rather than to work with averages. We use the 'model for impact' as a basis for the calculations and then assign characteristics to individuals. First, a publication rate P_r per researcher is assigned at random to individual researchers. P_r is based on the distribution as seen in Figure 1, fitted with the function

$$P_r = Y \cdot e^{-(t_a + t_d)} \quad (6)$$

As a next step we introduce parameters that have to do with sharing. The percentage of sharing researchers is a fixed parameter in this model. The researchers sharing type is assigned at random to individuals. The actual reuse of a dataset, based on the probability to find an appropriate dataset for a paper, is assigned at random to publications. The portion of papers R for which an appropriate dataset for reuse is found is calculated as

$$R = 1 - \frac{1}{1 + f \cdot X} \quad (7)$$

We now have a mix of individual researchers that share or do not share, find a dataset for reuse or not for any of their papers, and publish different number of papers in a year. Based on the parameters in Table 1 we assign costs and benefits with these traits. These factors determine the performance of researchers in terms of impact by citations.

To determine the publication rate distribution in Figure 1, 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 in 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. For the data see [Pronk et al., 2015]. This distribution, based on our sampling, implies that most researchers publish one- and a few researchers publish many papers in a given year. We fitted an exponential distribution through the sampled population (Formula 6). The average for the distribution is close to three papers per researcher in a given year.

Simulations

For the R-scripts to generate the plots for all simulations, see [Pronk et al., 2015].

We start with a set of simulations regarding performances per sharing type, with the model for impact. We calculate the impact for the two types of researchers over a range of sharing from zero to a hundred percent of all researchers. In addition to the default values (see Table 1), we change parameters to assess their influence on the publication rate and associated impact by citations for sharing and not sharing researchers. In Table 2 we list the parameters changed in the simulations and a score of the measures that would have these effects in a 'real world' scientific community [Chan et al., 2014].

To have a closer look on individual performance, we perform the same set of simulations with the individual based model. For each setting we calculate the difference between the publication rate assigned in Formula (6) at no costs or benefits with sharing or reuse, and a new, calculated publication rate based on sharing and reuse traits per researcher under the assumption that half of the researchers share. So, again we change the parameters in Table 2 and assess their influence, as in the first simulation.

We end by zooming out to community performance with the model for impact. We calculate the average impact over all researchers in the community, now at more extreme settings of the citation benefit b and in a second simulation at even higher cost t_c for preparing a dataset for sharing. This is to provide a broader range of results. Citation benefit b and the sharing rate are changed within their range in one hundred equal steps.

Results

Shown in Figure 2 are the simulations with the model for impact (Formulae 1-5). The simulation in (a) is at default parameter values (Table 1). In (b-f) we simulated measures to improve upon impact. There are two important observations. First, in all (but the last) subfigure of Figure 2 (a-e) the average impact of not sharing researchers exceeds that of sharing researchers irrespective of how many sharing researchers there are. This means that *not sharing* is the best option, at all percentages sharing researchers. In this scenario it would be logical if all individual researchers would choose not to share and eventually end up getting the average impact by citations depicted at zero percent sharing. So we see here a classical example of the tragedy of the commons or prisoners dilemma phenomenon. What is important to note though is that the measures in (b) (c) (d) and (e) ascertain a key effect when compared to the default in (a). The average impact of sharing researchers at the highest percentage sharing researchers (straight horizontal light-grey line; stripes) is increasingly higher with the measures than the average impact for not sharing researchers at zero percentage sharing researchers (straight horizontal dark-grey line; dots-stripes). Should a policy enforce the sharing, or all would agree to cooperate and share, a higher gain is achieved than in the case that researchers would all choose not to share. This illustrates the conflicting interest for individual researchers, who are better off not sharing, while they would do better if all of them did share. Subfigure (f) of Figure 2 shows the potential of the citation benefit with sharing. In the picture it is profitable to share at low sharing rates, and profitable not to share at high sharing rates, leading to a stable coexistence of sharing and not sharing researchers. This means that the community would exist of researchers from both strategies. Hypothetically, should the citation benefit be even higher, the sharing strategy would outperform the not sharing strategy at all sharing percentages. Researchers would in this case choose to share even without measures to promote sharing, simply because it directly increases their impact.

Second, it can be noted that in some subfigures of Figure 2 (a, b, c, e) the average citations are the highest at intermediate sharing. This means that if sharing increases further, it has a detrimental effect on average community impact. This is because the model is formulated in Formula (3) in a way that total costs for sharing increase for the community as more researchers share, whereas total benefits cease to increase at high sharing rate. The extra datasets do not contribute much to the benefits, or in other words, the research community has become saturated with datasets. Compared to the average community citations, which are highest at intermediate sharing, for both sharing and not sharing researchers the highest impact by citations is at the point at which everyone is sharing.

Results from the individual based in Figure 3 model show that the individual researchers have various gains depending on their publication rate, reuse, and dataset sharing habits. In (a) are the gains and losses in impact, at default parameter values (Table 1). In (b-f) we simulated measures to improve gains or limit losses. A possible desired effect of sharing of datasets would be that every individual researcher can benefit, sharing or not sharing. It can be observed that in Figure 3 (a-e) most of the sharing researchers have lower benefits or even costs compared to not sharing researchers. This logically is in line with the lower averages for sharing researchers in Figure 2. Also, it can be noted in all subfigures of Figure 3 that there are always sharing researchers that do not benefit from the availability of datasets by the reuse of datasets. These researchers were not (fully) able to compensate for the cost to share their data. It is notable that in (b) individual researchers are left with lower costs than in (c). This is because in (b) the probability of finding an appropriate dataset for reuse f is set higher, compensating the sharing costs for many of the researchers. In (c) the time cost t_r with reuse per paper is lower, benefitting only those few researchers that do find a reusable set. In (d) the lowering of the time cost t_c for preparing a dataset for sharing improves the situation for *all* researchers compared to the default in (a), but still some researchers are not fully compensated. In (e) the introduction of the citation benefit *b* does not help much to improve the benefits for sharing researchers. Only when in (f) a substantial citation benefit *b* is introduced for sharing researchers, the costs associated with sharing are (more than) compensated for, for all sharing researchers.

When simulating community impact in Figure 4 (a) and (b) it can be seen that, as the benefits *b* for sharing increase towards the right of the plot, the average community impact increasingly starts to rise with more sharing in both plots. Even the drop after the initial increase at increased sharing caused by the datasets saturation is eventually compensated for with the increase of the citation benefit with sharing. In subfigure (b) at the left side of the plot, without a citation benefit and with the very high cost for sharing t_c, there appears an alarming effect. At these parameter values the average impact becomes lower at high sharing than at no sharing at all. Policies increasing sharing would, if successful, in this case backfire and reduce scientific community impact.

Discussion

We analysed the effect of sharing and not sharing research data on scientific community impact. We found that there is a conflicting interest for individual researchers, who are *always* better off not sharing and omitting the sharing cost while they would have higher impact when sharing as a community. With our model we assessed some measures to improve the costs and benefits with sharing and reuse of data, to make most researchers profit from the sharing of datasets. We simulated policies to increase sharing, measures to stimulate reuse by reducing reuse costs or increasing discoverability of datasets, and measures to stimulate sharing by lowering costs associated with sharing or introducing a citation benefit with each shared dataset. These simulations concretize the notion in literature that improving spontaneous participation in sharing datasets will require lowering costs and/or increasing benefits for sharing [Smith, 2009; Roche et al., 2014] and values different measures to do so.

A policy is a straightforward measure to increase community impact simply by enforcing higher percentages of sharing researchers. Moreover, policies are pivotal for establishing acceptable data sharing practices and community-level standards. Such policies can be enforced on the level of institutions, funders, or journals. In the model these do increase community impact, as long as the community is not already saturated with datasets. In real life, at least for journals, policies have not been enough to convince researchers to actually make their dataset publicly available [Wicherts et al., 2006; Savage and Vickers, 2009; Alsheikh-Ali et al., 2011; Wicherts and Bakker, 2012; Vines et al., 2013]. This could be exemplary for the reluctance of individual researchers to share datasets because of real, or perceived costs. The inequality in costs between sharing researchers and not sharing researchers remains with mandated sharing, and the researcher that does not share a dataset but does reuse a dataset will have the highest impact compared to all others. Of course there are many factors for researchers to decide to share data or not, but simply said this could predispose a researcher towards not sharing. 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' datasets but only few wanted to make their dataset available to others, for a variety of reasons [Tenopir et al., 2011; Fecher et al., 2015].

Stimulating reuse by reducing reuse costs or increasing discoverability of datasets in the model increases average community impact, though not equally for all individuals within the community. Only the researchers that actually reuse a dataset profit from these measures, and the costs for those who share, although partly compensated, still exist. Again, although helpful, the inequality in costs between sharing and not sharing researchers is not addressed with such measures.

A direct reduction of the time costs with sharing a dataset in our model improved the situation for all sharing researchers. Only a small inequality between sharing and not sharing researchers remained. The best solution is however to introduce a 'citation benefit' for papers with the dataset shared, to directly balance the costs of sharing individuals. The citation benefit in real life can not only come from increased citations to the paper [Botstein, 2010; Sears, 2011; Dorch, 2012; Piwowar and Vision, 2013] but also from citations to the shared dataset itself [Costello et al., 2013; Belter, 2014; Neumann and Brase, 2014]. With a relatively high citation benefit, sharing datasets even becomes more profitable than not sharing, at any percentage of sharing researchers. Sharing then is not only optimal for maximizing community impact, but also for the individual researcher.

All in all, enhancement of the citation benefit would bring about better incentives to share datasets than imposing an obligation to share by funders, institutes or journals, or partly compensating for costs by enabling reuse. Better incentives arguably also lead to better sharing practices as researchers would strive to present their dataset as such that its reuse potential is optimal.

All models come with simplifications and assumptions. A central assumption of the model is the gain of scientific impact by citations to papers, and implicitly datasets. For some communities the concept of impact by citations is less applicable overall [Krell, 2002]. These fall outside the scope of this model. It also should be noted that there are other ways to count scientific impact such as Altmetrics [Roemer and Borchardt, 2012]. Additionally, we derived general phenomena for the scientific community, whereas (perceived) costs and benefits with sharing will differ between scientific communities [Vickers, 2011; Tenopir et al., 2011; Kim, 2013] and attitudes towards sharing can differ largely between disciplines [Kirwan, 1997; Huang et al., 2012; Pitt and Tang, 2013; Anagnostou et al., 2013]. This means that the measures taken to make sharing worthwhile will have to differ in their focus in each scientific community [Borgman et al., 2007; Acord and Harley, 2013]. To apply the current model to any specific situation or community, parameter values for that community should be carefully determined and, where necessary, the model should be adjusted or expanded. Additional factors that may influence the outcome of this model and that could possibly be incorporated in community specific versions or future refinements of this model include: differences in quality of papers leading to differences in citation rates, heterogeneity in the costs of sharing (small and easy versus big and complicated datasets to document), heterogeneity in the contribution of a papers' dataset to the available pool of datasets, introducing and allowing for heterogeneity in search time for datasets, feedback between the number of times a dataset is reused and the citation benefit for that dataset. A focal point to assess in the current model would also be the pool of available datasets. What is the relation between available datasets and reuse rate for researchers, do these datasets overlap in content, will all new datasets contribute to science, does the pool become saturated, are all datasets reused, what is the decay rate of datasets in the pool for that specific community?

Lastly, it is clear that 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, high time costs to prepare the data for reuse, the wish to publish later perhaps, patents pending, privacy sensitive data) [Kim, 2013; Cronin, 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 impact. 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. In the simulations there appeared a conflicting interest for individual researchers, who are *always* better off not sharing and omitting the sharing cost, while they are ultimately better off all sharing as a community. Although policies are indispensable and should be able to increase the rate of sharing researchers, and increased discoverability and dataset quality could partly compensate for costs, a better measure to promote sharing would be to lower the cost for sharing, or even turn it into a (citation-) benefit.

Acknowledgements

We thank Dorinne Raaimakers, Jeroen Bosman, Jan Molendijk, Conny van Bezu from Utrecht University Library and Mark van Oorschot from PBL, RIVM for their constructive ideas concerning the manuscript and initial concept. We thank two anonymous reviewers and Patricia Soranno, **Gyuri Barabas**, and **Timothée Poisot** for pointing out possibilities for improvement in previous versions of the manuscript.

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.

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.

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

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

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.

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

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.

Botstein, D. (2010), It's the data! Mol. Biol. Cell, 21(1), 4-6.

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.

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.

Dorch, B. (2012), On the Citation Advantage of linking to data. hprints-00714715, version 2, Hprints, http://hprints.org/hprints-00714715.

Fecher, B., S. Friesike, M. Hebing, S. Linek, and A. Sauermann (2015), A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing. DIW Berlin Discussion Paper No. 1454., doi:http://dx.doi.org/10.2139/ssrn.2568693.

Henneken E.A., A. A. (2011), Linking to data - effect on citation rates in astronomy. arXiv:1111.3618v1, http://arxiv.org/abs/1111.3618.

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.

Kirwan, J. R. (1997), Making original data from clinical studies available for alternative analysis, J. Rheumatol., 24(5), 822-825.

Krell, F. T. (2002), Why impact factors don't work for taxonomy. Nature, 415(6875), 957.

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.

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.

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

Pronk, T. E., P. H. Wiersma, and v. A. Weerden (2015), Replication data for: GAMES WITH RESEARCH DATA SHARING", http://hdl.handle.net/10411/20328 V4 [Version], Dutch Dataverse Network.

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.

Roemer, R. C. and R. Borchardt (2012), From bibliometrics to altmetrics. A changing scholarly landscape. College & Research Libraries News, 73, 596.

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.

Sears, J. R. L. (2011), Data Sharing Effect on Article Citation Rate in Paleoceanography, Fall Meeting, AGU, San Francisco, Calif., 5-9 Dec., Abstract IN53B-1628,http://adsabs.harvard.edu/abs/2011AGUFMIN53B1628S.

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.

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.

Vickers, A. J. (2011), Making raw data more widely available, BMJ, 342, d2323, doi:10.1136/bmj.d2323.

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.

Wicherts, J. M. and M. Bakker (2012), Publish (your data) or (let the data) perish! Why not publish your data too? Intelligence, 40(2), 73-76, doi:10.1016/j.intell.2012.01.004.

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.

Table 1(on next page)

Parameters, variables, and their values.

Table 1. Overview of parameters, variables, and their standard values used in the model. Grey rows indicate the parameters that are varied in the model to assess their influence (examples for real-world measures to change these are explained in Table 2).

Table 1.

Table 1. Overview of parameters, variables, and their standard values used in the model. Grey rows indicate the parameters that are varied in the model to assess their influence (examples for real-world measures to change these are explained in Table 2).

Parameter	Meaning	Value	Source	Unit
t_a	Time-cost to produce a paper	0.13	Derived: t_a+t_d amount to 121 days; leading to ~3 papers a year (similar to the average in Figure 1)	Year/ Paper
t_d	Time-cost to produce a dataset	0.2	Derived: t_a+t_d amount to 121 days; leading to ~3 papers a year (similar to the average in Figure 1)	Year/ Paper
t_c	Time-cost to prepare a dataset for sharing	0.1	Estimated: 36.5 days	Year/ Paper
t_r	Time-cost to prepare a dataset to reuse	0.05	Estimated: 18.25 days	Year/ Paper
q_x	Decay rate of shared datasets	0.1	Derived: based on a storage time of 10 years	1 / Year
b	Citation benefit (sharing researcher)	0	Estimated: percent extra citations	Percent
f	Probability to find an appropriate dataset	0.00001	Fitted	1 / Dataset
c	Citations per paper produced	3.4	Derived: approximate from 'baselines'; average citation rate by year three, Thompson Reuters	Citation / Paper
State Variables	**Meaning**	**Value**		**Unit**
E	Efficiency of researchers	See formula (1)	Calculated	Citation / Year
P	Number of papers	See formula (2)	Calculated	Paper / Year
T	Time for a publication	See formula (3)	Calculated	Year / Paper
X	Pool of shared datasets	See formula (5)	Calculated	Dataset
Y	Number of researchers	10000	Defined	n.a.

Table 2(on next page)

Changed parameters and associated measures

Table 2. Overview of considered parameters determining reuse and sharing habits of researchers, and possible measures to improve these in a realistic setting.

Table 2.

Table 2. Overview of considered parameters determining reuse and sharing habits of researchers, and possible measures to improve these in a realistic setting.

Parameters investigated in the model	**Possible associated measures to improve this**
Time 't_r' spent to assess and include an external dataset	• Improve data quality, for instance by the use of data journals [Costello et al., 2013; Atici et al., 2013; Gorgolewski et al., 2013], or peer review of datasets (i.e. a 'comment' field in data repositories). • Offer techniques or tools for easy assessment of dataset quality [Eijssen et al., 2013], faster pre-processing or data cleaning (i.e. 'OpenRefine' or 'R statistical language').
Chance 'f' to find an external dataset	• Harvest databases through data portals to reduce 'scattering' of datasets. • Standardization of metadata and documentation • Advanced community and project-specific databases • Library assistance in finding and using appropriate datasets.
Time 't_c' associated with sharing of research data	• Offer a good storing & sharing IT infrastructure. • Assistance with good data management planning at the early stages of a research project.
Benefit in citation per paper 'b' associated with sharing of research data	• Provide a permanent link between paper and dataset. • Increase attribution to datasets by citation rules . • Establish impact metrics for datasets.
Percentage 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.

Figure 1(on next page)

Publication distribution

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 published papers' distribution for the simulated community.

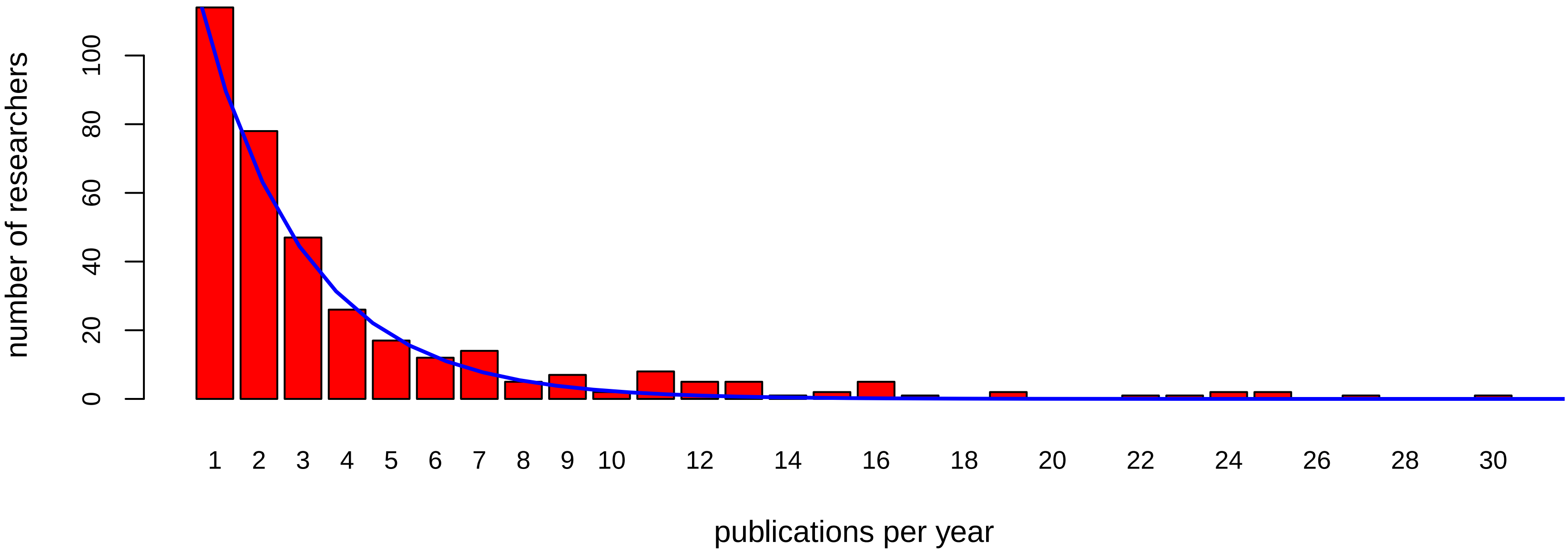

number of researchers
0
20
40
60
80
100
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30
publications per year

Figure 2 (on next page)

Impact per sharing type

Figure 2. Citations ('impact') per year for researchers sharing and not sharing, at different percentages of sharing researchers. The simulations are done at parameter settings a) default (see Table 1), b) default but with *f* increased threefold c) default but with t_r decreased threefold d) default but with t_c decreased threefold e) default but with *b* set to 0.1 f) default but with *b* set to 0.4. The curved light-grey line depicts the impact of the sharing researchers . The curved dark-grey line depicts the impact of the not sharing researchers. The thin dotted curved black line is the averaged community impact. The straight black vertical dotted line depicts the percentage of sharing researchers at which community impact is maximized. The straight horizontal lines respectively depict the impact at zero percent researchers sharing (dark-grey line; dots-stripes) and hundred percent sharing researchers (light-grey line; stripes).

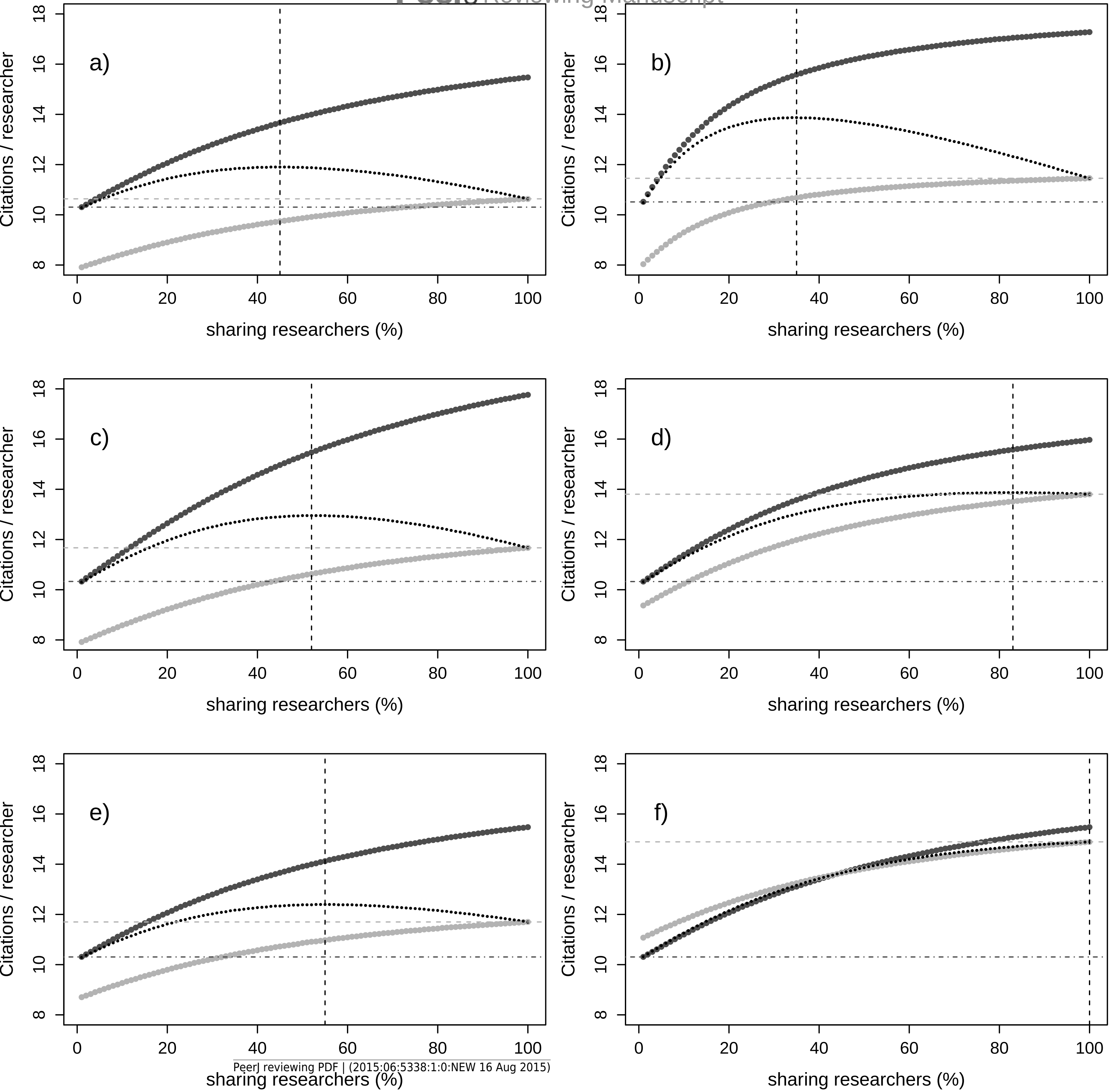

a)
b)
c)
d)
e)
f)
Citations / researcher
sharing researchers (%)
8
10
12
14
16
18
0
20
40
60
80
100

Figure 3 (on next page)

Individual gains with sharing

Figure 3. Gains from sharing in number of citations per individual researcher. These gains are calculated for the situation with fifty percent sharing researchers compared to the same situation without sharing researchers. For visualization purposes the researchers are sorted according to sharing habitat: not sharing researchers (dark grey circles) to the left, sharing researchers (light grey circles) to the right. See the legend of Figure 2 for parameter settings in all subfigures.

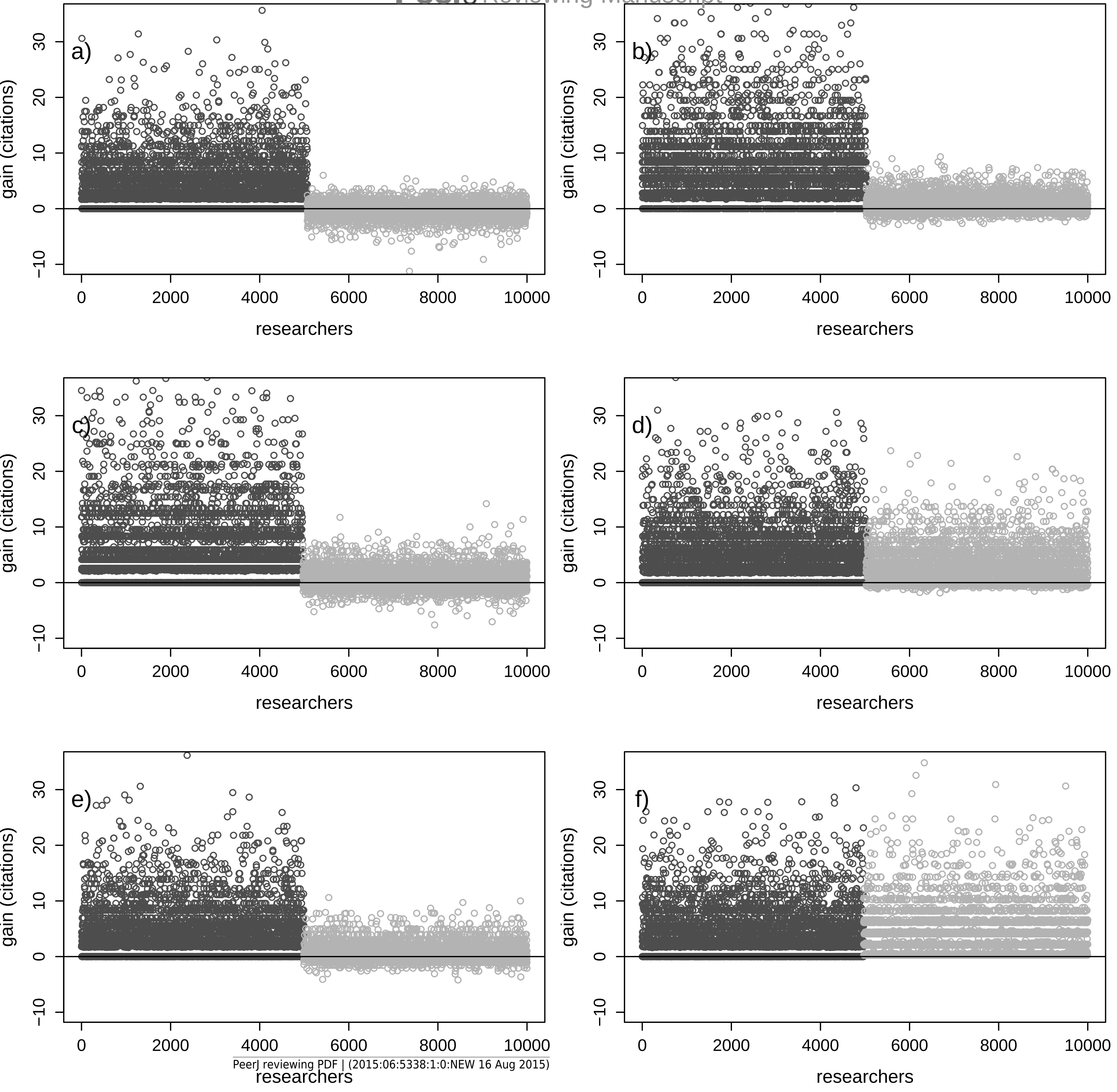
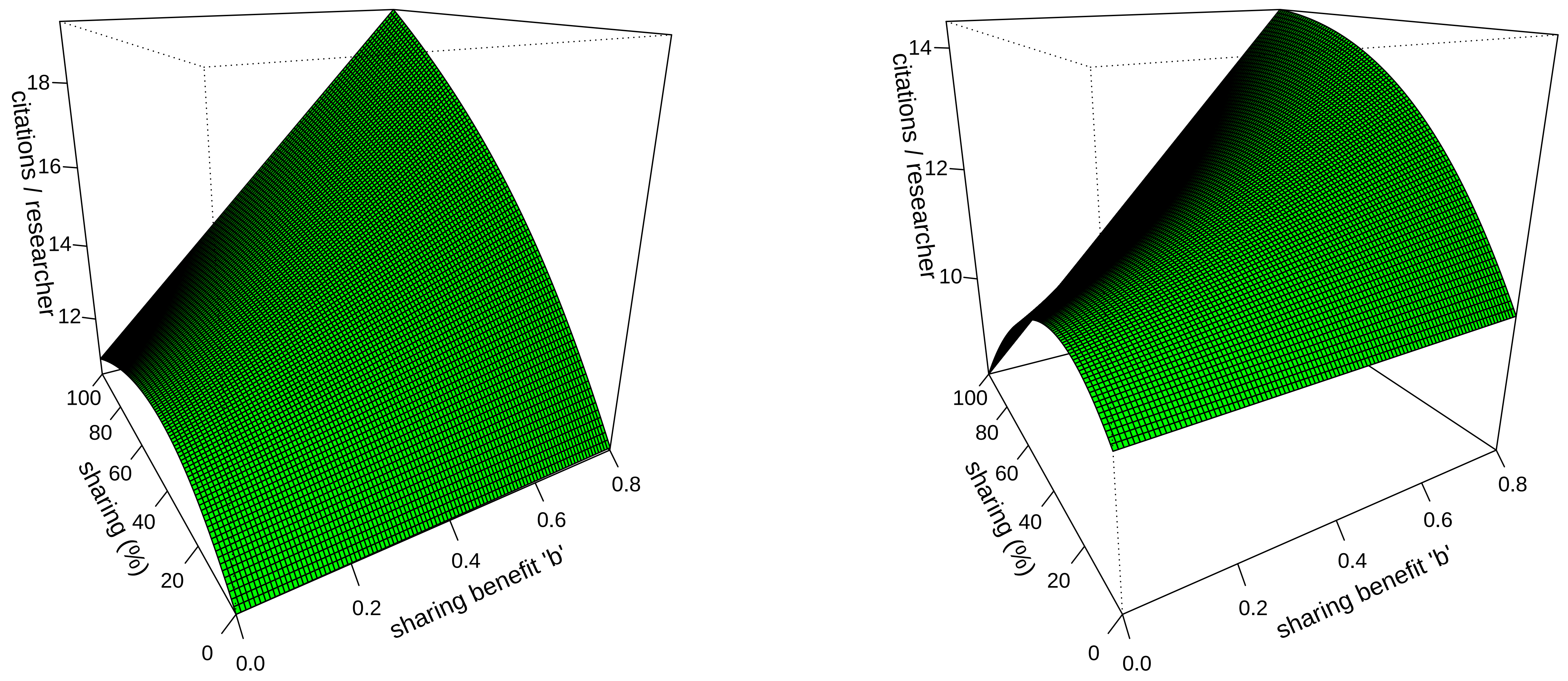

a)
b)
c)
d)
e)
f)
gain (citations)
researchers
−10
0
10
20
30
0
2000
4000
6000
8000
10000

Figure 4 (on next page)

Community impact

Figure 4. Average community impact with varying percentage of sharing researchers and varying sharing benefit *b*. Figures are calculated at default parameter values (see Table 1) with the exception of *b* which is varied, and for subplot (b) t_c, of which the value was set from 0.1 to 0.2. On the z-axis is t he average community impact. On the x and y axes respectively increasing benefits *b* for sharing from 0 to 0.8 (0 to 80% citation benefit with sharing) and increasing percentage of sharing researchers from 0 to100%.


citations / researcher
18
16
14
12
sharing (%)
100
80
60
40
20
0
sharing benefit 'b'
0.0
0.2
0.4
0.6
0.8
citations / researcher
14
12
10
sharing (%)
100
80
60
40
20
0
sharing benefit 'b'
0.0
0.2
0.4
0.6
0.8