NBER WORKING PAPER SERIES

RESEARCH FUNDING AND COLLABORATION

Benjamin Davies
Jason Gush
Shaun C. Hendy
Adam B. Jaffe

Working Paper 27916
http://www.nber.org/papers/w27916

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020

The research discussed in this paper was supported by Te Pūnaha Matatini. We thank Ben Jones, Dave Maré and seminar participants at Motu for helpful suggestions. Jason Gush is employed by the administrator of the Fund being studied, i.e., the Royal Society of New Zealand, as their "Programme Manager - Insights and Evaluation". The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Research Funding and Collaboration
Benjamin Davies, Jason Gush, Shaun C. Hendy, and Adam B. Jaffe
NBER Working Paper No. 27916
October 2020
JEL No. O31,O38

ABSTRACT

We analyse whether research funding contests promote co-authorship. Our analysis combines Scopus publication records with data on applications to the Marsden Fund, the premiere source of funding for basic research in New Zealand. On average, and after controlling for observable and unobservable heterogeneity, applicant pairs were 13.8 percentage points more likely to co-author in a given year if they co-proposed during the previous ten years than if they did not. This co-authorship rate was not significantly higher among funded pairs. However, when we increase post-proposal publication lags towards the length of a typical award, we find that funding, rather than participation, promotes co-authorship.

Benjamin Davies
Stanford University
bldavies@stanford.edu

Jason Gush
Royal Society of New Zealand
11 Turnbull St, Thorndon
Wellington 6011
New Zealand
jason.gush@royalsociety.org.nz

Shaun C. Hendy
University of Auckland
Auckland 1010
New Zealand
shaun.hendy@auckland.ac.nz

Adam B. Jaffe
188 Brookline Avenue
Apartment 26A
Boston, MA 02215
and Brandeis University
and Queensland University of Technology
and also NBER
adam.jaffe@motu.org.nz

1 Introduction

Research is increasingly conducted by teams (Adams et al., 2005; Wuchty et al., 2007).[1] Collaboration allows researchers to divide labour and overcome the rising "burden of knowledge" required to generate new knowledge (Jones, 2009). Therefore, collaboration can boost productivity (Ductor, 2015). However, the extent to which teamwork leads to more impactful research depends on team composition (Ahmadpoor and Jones, 2019). Such dependence motivates studies of team formation processes and the mechanisms underlying those processes.[2] This paper explores one potential mechanism: participation in research funding contests (Ayoubi et al., 2018; Defazio et al., 2009; Ubfal and Maffioli, 2011).

Research funding contests may promote collaboration through several channels. First, preparing and submitting funding proposals requires proposal team members to develop joint plans for their proposed research and to invest resources in the pursuit of a shared idea. Second, funding application processes allow researchers to discover complementarities among their sets of knowledge and skills, and to screen for productive collaborators. Finally, if the most promising proposals win funding then shared proposal success signals that teams' ideas are worth pursuing and provides resources that may foster further collaboration.

In this paper, we study a specific form of collaboration: co-authorship of research papers. We analyse the relationship between co-authorship and proposal outcomes empirically, using data from New Zealand.[3] Our empirical strategy allows us to control for observable and unobservable factors that may confound the relationship between co-authorship and proposal outcomes. Our data include Scopus publication records on New Zealand researchers and their international co-authors. We link these records to data on applications to the Marsden Fund, the "premiere funding mechanism for basic research in New Zealand" (Gush et al., 2018, p. 227). The substantial financial and reputational rewards offered by the Fund encourage most serious New Zealand researchers to apply at least once during their careers. Consequently, our linked data provide representative information about the collaborative behaviour of New Zealand researchers during our period of study.

[1] Explanations for this trend include increasing knowledge specialisation (Agrawal et al., 2016; Jones, 2009; McDowell and Melvin, 1983), changing institutional incentives (Barnett et al., 1988; Bikard et al., 2015; Hamermesh, 2013), and decreasing communication and travel costs (Agrawal and Goldfarb, 2008; Catalini et al., 2019; Katz and Martin,1997; Kim et al., 2009; Rosenblat and Mobius, 2004).
[2] See, e.g., Boudreau et al. (2017), Fafchamps et al. (2010), Guimerà et al. (2005), and Rivera et al. (2010).
[3] The code used to conduct our analysis is available at https://doi.org/10.5281/zenodo.4041257.

We use our linked data to construct a sequence of co-authorship networks among Marsden Fund applicants. These networks capture the birth and decay of active collaborations over time. The mean degree in these networks grew between 2009 and 2018, consistent with the rise in co-authorship shown in previous studies. Researchers with more successful Marsden Fund applications tended to have more co-authors. However, this relationship may be driven by confounding factors, such as researchers' willingness and ability to generate publishable research.

We control for such factors by analysing co-authorship dynamics econometrically. We construct panel data on pairs of New Zealand researchers who collaborated as co-authors or as proposal team co-members between the years 2000 and 2018. We use these data to model pairwise co-authorship rates as functions of pairs' time-varying characteristics. Our models capture the relationship between co-authorship and proposal outcomes, conditional upon pairs' match qualities and assortative preferences. We control for selection bias and unobservable heterogeneity by including pair fixed effects.

On average, and after controlling for observable and unobservable heterogeneity, researcher pairs were 13.8 percentage points more likely to co-author in a given year if they had co-proposed during the previous ten years than if they had not. This co-authorship rate was not significantly larger among applicant pairs who received funding. However, increasing the lag between our dependent and independent variables delivers the opposite result: funding, rather than mere participation, promotes co-authorship. These patterns suggest that the "treatment effect" of research funding contest participation on co-authorship is limited to successful participants only. Our estimates are robust to dyadic clustering (Aronow et al., 2017; Graham, 2020) and to relaxing our pair selection criteria.

Our estimates suggest three behaviours among pairs of New Zealand researchers. First, pairs who were innately suited to co-authorship were more likely to submit Marsden Fund proposals and to receive funding during our period of study. Second, pairs were more likely to submit proposals at times when they had fruitful ideas and research success. Third, funding receipt increased the probability of co-authorship several years later, either because it increased the publication output of existing collaborations or because it encouraged new publication-focused collaborations.

This paper contributes to the literature on the impact of research funding on collaboration and output. Ebadi and Schiffauerova (2013) survey this literature, noting that "knowledge about the effects of funding on collaboration is very limited." Ayoubi et al. (2018), Defazio et al. (2009), and Ubfal and Maffioli (2011) analyse the impact of funding on collaboration, with generally positive results. However, none of these studies analyse the extent to which funding contests themselves

foster collaboration. We fill this gap in the literature by estimating how different stages of the grant application process—submitting proposals and receiving funding—contribute to subsequent co-authorship rates. Consequently, our analysis informs research and science funding system design by showing which features of the funding application process appear to be the most successful at encouraging collaboration.

This paper also contributes to the broader literature on research team formation. This literature draws upon sociological theories of network formation, and large bibliometric databases, to explain and analyse how teams form. For example, Ahmadpoor and Jones (2019) show that the quality of teams (measured by the citation intensity of team outputs) is heavily influenced by the *minimum* of the quality of teams' members (measured by previous team-member citation intensity). This leads teams to assemble among researchers with similar citation profiles, reflecting the assortative mechanisms through which collaborative networks evolve (Rivera et al., 2010). We analyse these mechanisms in an econometric setting, using dyadic regression (Graham, 2020) to estimate the determinants of pairwise co-authorship. We describe the empirical challenges associated with constructing and analysing pairwise co-authorship data, offer solutions to these challenges, and discuss how our solutions affect inference.

2 The Marsden Fund

The New Zealand Government established the Marsden Fund in 1994. The Fund "invests in excellent, investigator-led research aimed at generating new knowledge, with long-term benefit to New Zealand" (Royal Society of New Zealand, 2017). The Royal Society of New Zealand (RSNZ) administers the Fund on behalf of the Marsden Fund Council, appointed by the Minister of Science and Innovation. The Council oversees proposal assessment, and recommends to the RSNZ which proposals to fund and how much funding to award. Funding rounds occur annually. Typical awards last up to three years.

Proposal teams may contain one to eight researchers, who may be Principal Investigators (PIs), Associate Investigators (AIs), post-doctoral researchers, post-graduate students, or research assistants. Every team must contain a PI, who must be New Zealand-based. PIs and AIs cannot belong to more than two proposal teams per funding round. Applicants cannot be PIs on more than one proposal per funding round or, since 2011, more than one active grant. For example, if an applicant receives funding as a PI then they cannot submit another proposal as a PI during the following two years.

The Marsden Fund application process comprises two stages. In the first stage, teams submit a one-page abstract of their proposed research. These abstracts are reviewed by discipline-specific assessment panels. Each panel comprises six to ten panelists appointed by the RSNZ. The number of proposals received by each panel determines the panel's share of the overall allocable budget. Panels rank proposals on their potential scholarly impact, their rigour, and their ability to enhance New Zealand's research capacity. Based on these rankings, each panel selects about 20% of their received proposals to progress to the second stage.

In the second stage, teams submit "full" proposals detailing their research methods and objectives. Panels rank full proposals based on external reviews, applicants' responses to those reviews, and panelists' discussions and judgments. The highest ranked proposals within panels' budgets receive funding. Historically, about half of the full proposals submitted to assessment panels have received funding, implying an overall application success rate of about 10%.

The two stages of the Marsden Fund application process imply four levels of interaction among applicants: not applying, submitting a first round proposal, submitting a second round proposal, and receiving funding. Later levels require more intellectual engagement among researchers. To the extent that such engagement leads to co-authorship, researchers with more successful Marsden Fund applications may co-author more often, all else equal. Funded applicants may also feel obliged to publish together so as to demonstrate the outputs of their grant.

Marsden Fund application success may also correlate positively with co-authorship due to selection. Assessment panels rank proposals according to their expected impact, which may depend on proposal team members' perceived ability to work together productively. However, Gush et al. (2018) show that assessment panels' rankings in the second round are uncorrelated with proposal teams' subsequent performance, conditional on their past performance. Thus, even if panels indirectly attempt to select teams more likely to co-author, there is no evidence that these attempts succeed.

3 Publication and proposal data

We use Scopus data on publications generated by New Zealand researchers and their international co-authors between the years 1996 and 2018. These data contain 7,854,938 publications matched to 7,141,834 Scopus author IDs. However, this author ID count is unlikely to equal the true number of unique authors in the data due to author name disambiguation issues, which Scopus' internal systems appear to resolve only partially. Some author IDs errantly merge the records of multiple authors, while some authors have multiple author IDs.

We link our Scopus data to data on Marsden proposals submitted between the years 2000 and 2018. Our Marsden data describe 18,811 unique proposals, of which 22% proceeded to the second round and 10% received funding. The research teams on these proposals drew from a set of 16,400 applicants from New Zealand and overseas.

We link Marsden applicants to Scopus author IDs by constructing two sets of applicant-author pairs. Our first set comprises pairs with common email addresses and ORCID profiles. Our second set comprises pairs for which the applicant and author have full names that share a cosine similarity no smaller than 0.95.[4] We restrict this second set to pairs with (i) an exact full name match or (ii) at least two proposal team co-members whose matched authors have co-authored with the same matched author.[5]

We use our sets of applicant-author matches to construct data on researchers with observable proposal and publication behaviour. First, we take the two sets' union and discard all matches with multiple Marsden applicants linked to the same Scopus author. This leaves 15,557 matches, 11,428 of which are one-to-one. Second, we remove all matches where an applicant is linked to 10 or more authors. This leaves a set of 13,193 researchers, whom we associate with 15,438 unique Scopus author IDs. We refer to this set as our "linked data" for the remainder of this paper.

Our linking procedure doubles as a data cleaning procedure. Our applicant-author matching criteria disambiguate researcher profiles by cross-referencing the characteristics of researchers in our publication and proposal data sets against each other. Researchers outside our linked data do not have the same opportunities to have their characteristics cleaned. Therefore, we restrict our analysis to the researchers in our linked data.

Although we have publication data from 1996 to 2018, we ignore publications during the years 1996 through 1999 so that the years covered by our linked publication and proposal data are consistent. Thus, our period study comprises the years 2000 through 2018. Appendix Table 1 counts the unique publications, authors, proposals, and applicants in our linked data for each of these years. Overall, the researchers in our linked data generated 613,899 publications and 18,065 proposals during our period of study.

[4] This threshold is small enough to catch slight mis-spellings and middle initials, and large enough for us to efficiently compute the number of potential common co-authors between matched applicants. We use cosine similarities, rather than other string metrics (e.g., Levenshtein distances), for its low computational cost and its ability to handle character block rearrangements (e.g., mis-labelled forenames and surnames).

[5] Applicant-author pairs selected using criterion (ii) must have at least two proposal team co-members and at least two co-authors, potentially biasing the observed extent of collaboration upwards due to selecting on researchers with observable collaborators. Excluding such pairs from our analysis leaves our econometric results unchanged.

4 Co-authorship network trends

Our ultimate goal to analyse whether, and to what extent, joint participation in the Marsden Fund application process promotes co-authorship. We begin by describing how the co-authorship network among Marsden Fund applicants evolved during our period of study. This description helps us demonstrate the aggregate behaviour of the researchers discussed in this paper.

We construct a sequence of co-authorship networks among the 13,193 researchers in our linked data. For each year $t \in \{2009, 2010, \ldots, 2018\}$, we define the set V_t of researchers who published during the years $(t - 9)$ through t and the set $E_t \subseteq V_t \times V_t$ of pairs of researchers who co-authored at least one publication during those years. We then define a co-authorship network $\mathcal{N}_t$ with node set equal to V_t and edge set equal to E_t. Thus, nodes in $\mathcal{N}_t$ represent researchers while edges join "recent" co-authors. Each researcher $i \in V_t$ has degree

$$\deg_t(i) = |\{j \in V_t : \{i, j\} \in E_t\}|,$$

which counts researcher i's co-authors during the years $(t - 9)$ through t. Since each edge in E_t joins exactly two researchers, the nodes in any non-empty subset U of V_t have mean degree

$$\frac{1}{|U|}\sum_{i \in U} \deg_t(i) = \frac{2|E_t|}{|U|}.$$

Our 'rolling window' definition of $\mathcal{N}_t$ allows us to capture the birth and decay of active collaborations over time. We could instead define $\mathcal{N}_t$ 'cumulatively' as the network among researchers who had ever co-authored before or during year t. However, this cumulative definition has two drawbacks. First, we do not observe publications prior to the year 1996, meaning that the cumulative network created using our data may exclude some edges non-randomly. Second, defining $\mathcal{N}_t$ cumulatively may create spurious relationships between observed co-authorship patterns and proposal outcomes due to older researchers having more time to generate publications and submit Marsden Fund proposals. Our rolling definition of $\mathcal{N}_t$ avoids both of these issues.

We partition the set of researchers in $\mathcal{N}_t$ based on whether they belonged to a funded, second round, first round, or no proposal team(s) between the years $(t - 9)$ and t. The resulting partition $\mathcal{P}_t$ comprises four disjoint parts—one for each 'best' round reached. We then compute the mean degree of researchers in each part, allowing us to determine whether researchers with different proposal outcomes tended to have different co-authorship propensities among the researchers in our linked data.

The number of researchers in the co-authorship network $\mathcal{N}_t$ grew from 11,808 in the year 2009 to 12,823 in the year 2018. Figure 1 shows the distribution of these researchers across parts $P \in \mathcal{P}_t$. The percentage of researchers in $\mathcal{N}_t$ who never interacted with the Marsden Fund during the years $(t-9)$ through t fell from 42.2% in 2009 to 23.3% in 2018. In contrast, the percentage of researchers in $\mathcal{N}_t$ who had worked on second round or funded proposals remained relatively constant.

Co-authorship rates increased during our period of study. Figure 2 plots the normalised mean degree

$$\frac{100}{|V_t|-1}\left(\frac{1}{|P|}\sum_{i\in P}\deg_t(i)\right)=\frac{200|E_t|}{|P|(|V_t|-1)} \tag{1}$$

among researchers in each part $P \in \mathcal{P}_t$ across years $t \in \{2009, 2010, \ldots, 2018\}$. Multiplying by $100/(|V_t|-1)$ converts node degrees from counts to percentages of possible co-authors. Thus, the normalised mean degree (1) equals the mean percentage of nodes in V_t with whom researchers in P co-author. This percentage grew for each part $P \in \mathcal{P}_t$ during our period of study, consistent with the rise in co-authorship shown in other studies (e.g., Adams et al., 2005; Wuchty et al., 2007).

More successful Marsden Fund applicants tended to have more co-authors in our data during our period of study. For example, researchers who worked on funded proposals during the years 2009 through 2018 had, on average, 16.9 co-authors within our linked data during that period, compared to 13.8 for researchers who progressed to the second round but never received funding and 9.1 for researchers who submitted proposals but never progressed beyond the first round. These differences suggest that researchers with different proposal outcomes have different co-authorship patterns. The remainder of this paper analyses these patterns econometrically, allowing us to control for observable and unobservable factors that may influence both proposal outcomes and co-authorship.

5 Researcher pair panel construction

We use our linked publication and proposal data to construct panel data on researcher pairs. Observations in these data correspond to researcher pairs in a given year. We use our panel data to estimate models of the form

$$\Pr\left(coauth_{ijt}=1\right)=\Lambda^{-1}\left(x_{ijt}\beta+u_{ijt}\right) \tag{2}$$

where

$$coauth_{ijt}=\begin{cases}1 & \text{if researchers } i \text{ and } j \text{ co-authored in year } t\\ 0 & \text{otherwise,}\end{cases}$$

$\Lambda(x) \equiv \ln\big(x/(1-x)\big)$ is the logit link function, x_{ijt} is a row vector of pair $\{i,j\}$'s time-varying characteristics, β is a vector of coefficients to be estimated, and u_{ijt} is an error term. The following subsections describe the variables included in x_{ijt} and the criteria we use to select pair panel observations.

5.1 Covariate definitions

5.1.1 Proposal outcomes

Our primary interest is in how $coauth_{ijt}$ covaries with the indicator variables $first_{ijt}$, $second_{ijt}$, and $funded_{ijt}$ for the events in which researchers i and j were co-members on a first-round, second round, and funded proposal team during the years $(t-10)$ through $(t-1)$. These three variables capture pairs' shared proposal outcomes during the ten years prior to possible co-authorship. The variables are cumulative in the sense that pairs who worked together on a funded proposal during the first and second rounds have $first_{ijt} = second_{ijt} = funded_{ijt} = 1$. Hence, the coefficients on these variables capture the marginal increase in the probability of co-authorship associated with progressing to each stage of the Marsden Fund application process. This increase may arise from two distinct effects. First, it could reflect an increase in pair $\{i,j\}$'s propensity to collaborate with each other. Second, it could reflect an increase in the frequency at which pair $\{i,j\}$'s collaborative outputs are accepted for publication. We do not attempt to isolate these two effects but encourage further research on methods for such isolation.

Co-authorship rates may also covary with other pair-level factors, such as research field overlaps, prior co-authorship, collaborative propensities, and citation impacts. We control for these factors by including additional covariates in the vector x_{ijt}. We describe these covariates below.

5.1.2 Research field overlaps

First, we control for the amount of overlap among the sets of fields in which researchers publish. This overlap reveals similarity in research interest, which is "probably the single most important factor in determining the likelihood of collaboration" (Fafchamps et al., 2010, p. 217).

We identify research fields using the All Science Journal Classification (ASJC) system used by Scopus. This system matches each Scopus publication with one or more of 334 unique fields, such as Organic Chemistry, Logic, and Numerical Analysis. Each field belongs to one of 27 field groups, such as Chemistry and Mathematics. Appendix Table 2 presents the proportion of publications in our linked data matched to each of these field groups.

Let p_{if} denote the number of publications by researcher i in field f, 'fractionally' counted so that publications with n matched fields contribute $1/n$ to the count for each researcher-field pair. Then

$$\pi_{if} = \frac{p_{if}}{\sum_g p_{ig}}$$

is the proportion of researcher i's publications in field f. We measure the research field overlap for researchers i and j via the cosine similarity[6]

$$overlap_{ij} = \frac{\sum_f \pi_{if}\pi_{jf}}{\sqrt{\left(\sum_f \pi_{if}^2\right)\left(\sum_f \pi_{jf}^2\right)}}.$$

This variable equals zero if researchers i and j never publish in the same field, and equals one if researchers i and j publish in the same fields in the exact same proportions. Following Fafchamps et al. (2010), we include both $overlap_{ij}$ and its square in the vector x_{ijt}. Researchers may prefer coauthors with overlaps large enough to enable communication about field-specific ideas but small enough to allow for knowledge and skill complementarities. Including a quadratic term allows us to model this hypothesised inverted U-shaped relationship between co-authorship and overlap.

5.1.3 Prior co-authorship

Second, we control for whether researcher pairs co-authored recently by including in x_{ijt} the indicator variable

$$adjacent_{ijt} = \begin{cases} 1 & \text{if } \{i,j\} \in E_{t-1} \\ 0 & \text{otherwise} \end{cases}$$

for the event in which researchers i and j were adjacent in the co-authorship network $\mathcal{N}_{t-1}$. The coefficient on $adjacent_{ijt}$ captures pairs' tendencies to co-author again if they co-authored during the previous ten years.

5.1.4 Collaborative propensities

Third, we control for researchers' collaborative propensities. The more researchers collaborate, the more likely we are to observe them co-authoring, independently of their Marsden Fund proposal outcomes. Moreover, Figure 2 shows that more successful Marsden Fund applicants tend to have more co-authors. Therefore, including collaborate propensities in x_{ijt} controls for base rates, and mitigates omitted variable bias when we estimate coefficients on $first_{ijt}$, $second_{ijt}$, and $funded_{ijt}$.

We measure researchers' collaborative propensities using their co-authorship network degrees. In particular, we define

[6] Researchers' interests may change over time. However, the time-varying components of pairs' research overlaps are small—that is, indistinguishable from noise—during the 19-year period spanned by our data.

$$degree_{it} = \begin{cases} \deg_{t-1}(i) & \text{if researcher } i \text{ is a node in } \mathcal{N}_{t-1} \\ 0 & \text{otherwise} \end{cases}$$

for each researcher i and year t, where $\deg_{t-1}(i)$ denotes researcher i's degree in the co-authorship network $\mathcal{N}_{t-1}$. The variable $degree_{it}$ counts the New Zealand and international researchers in our linked data with whom researcher i co-authored during the years $(t - 10)$ through $(t - 1)$. This count varies at the researcher-year level. However, our model specification (2) includes pair-year level covariates only. We generate such covariates by computing the (adjusted) mean

$$\overline{degree}_{ijt} = \frac{degree_{it} + degree_{jt}}{2} - adjacent_{ijt} \tag{3}$$

and absolute difference

$$\Delta degree_{ijt} = \left| degree_{it} - degree_{jt} \right|$$

in co-author counts for each pair of researchers $\{i, j\}$ and year t. Subtracting $adjacent_{ijt}$ in (3) yields the mean number of co-authors among researchers i and j, excluding each other, during the years $(t - 10)$ through $(t - 1)$. Subtracting $adjacent_{ijt}$ also allows us to include it in x_{ijt} separately without introducing collinearity.

Including $\overline{degree}_{ijt}$ in (2) controls for researcher i and j's baseline propensities to co-author. Consequently, we expect the coefficient on $\overline{degree}_{ijt}$ to be positive and significant. In contrast, including $\Delta degree_{ijt}$ in (2) controls for the extent to which researchers i and j have different collaborative propensities. Consequently, the coefficient on $\Delta degree_{ijt}$ may be positive or negative, depending on whether the co-authorship network exhibits negative or positive assortative mixing with respect to node degrees (Newman, 2002).

5.1.5 Citation impacts

Last, we control for pairs' citation impacts. Researchers with more highly cited publications may be more able to attract new collaborators due to having demonstrated willingness and ability to conduct high quality research. Moreover, researchers with more highly cited publications tend to have more publications overall, increasing the probability that we observe them co-author with existing collaborators. Controlling for citation impacts helps us control for these sorting and base rate effects.

We capture citation impacts as follows. First, we divide the number of citations accrued to each publication in our data by the mean number of citations accrued to publications in the same year

and ASJC field globally.[7] This division makes citation counts comparable across years and fields (Waltman et al., 2011). Then, for each researcher i and year t, we sum the mean-normalised citations accrued to researcher i's publications during the years $(t - 10)$ through $(t - 1)$. We fractionalise these sums so that publications with n authors and c mean-normalised citations add c/n mean-normalised citations to each author's sum. This fractional summing procedure delivers a mean-normalised citation score $MNCS_{it}$ for each researcher i and year t. This score is greater for researchers who published more highly cited papers during the years $(t - 10)$ through $(t - 1)$. We then compute the mean

$$\overline{MNCS}_{ijt} = \frac{MNCS_{it} + MNCS_{jt}}{2}$$

and absolute difference

$$\Delta MNCS_{ijt} = \left|MNCS_{it} - MNCS_{jt}\right|,$$

and include both $\overline{MNCS}_{ijt}$ and $\Delta MNCS_{ijt}$ in the covariate vector x_{ijt}.

The coefficients on $\overline{MNCS}_{ijt}$ and $\Delta MNCS_{ijt}$ in (2) capture the extent to which researchers consider citation impact when forming co-authorship teams. For example, more impactful researchers may be more likely to become co-authors because they are more attractive to each other as productive, career-enhancing collaborators, in which case we would expect a positive coefficient on $\overline{MNCS}_{ijt}$. Likewise, more impactful researchers may avoid co-authoring with less impactful researchers, in which case we would expect a negative coefficient on $\Delta MNCS_{ijt}$.

5.1.6 Geographic proximity

Several studies examine the geography of research team formation processes. Teams may be more likely to form among researchers at the same or geographically proximate institutions due to relatively low communication and travel costs (Agrawal and Goldfarb, 2008; Catalini et al., 2019). Likewise, institutional activities and facilities, such as departmental seminars and shared meeting spaces, create environments for sharing ideas and forging collaborations.

Nevertheless, we do not include institutional co-location or geographic proximity as covariates in (2). Our justifications are two-fold. First, the geographic information in our data is of very low quality. Scopus often assigns all of a publication's authors to a single institution, which is seldom accurate. In other cases, institutional affiliations are missing or out-of-date. Where affiliations are available and

[7] For publications matched to multiple ASJC fields, we use the mean of the field-specific means.

correct, they are not always associated with an address, preventing calculations of geographic proximity.

Second, New Zealand is geographically small: all research institutions in New Zealand share the same time zone and are within a few hours travel from each other. Indeed, Aref et al. (2008) document widespread cross-institutional collaboration in New Zealand. Therefore, the effect of geography on collaboration is likely to be less important in New Zealand than in other contexts.

5.2 Restrictions on panel data observations

Our linked data contain 13,193 researchers, implying

$$\frac{13{,}193 \times (13{,}193 - 1)}{2} = 87{,}021{,}028$$

potential researcher pairs. It is computationally infeasible to analyse a set of this size. Moreover, most pairs never collaborated: 82,240 (0.09%) were co-authors and 40,541 (0.05%) were proposal team co-members during our period of study.[8] Such sparsity prevents reliable estimation of (2) because the heterogeneity among pairs who never co-author likely exceeds the mean differences between pairs who did and did not co-author. Sampling pairs randomly would lower the computational burden but maintain our data's sparsity.

The researchers in our data also face heterogeneous incentives to submit, and collaborate on, Marsden Fund proposals. Our data contain New Zealand and international researchers who collaborated on Marsden Fund proposals. However, international researchers are not able to receive direct funding support for their time or institutional costs spent collaborating on Marsden funded projects (Royal Society of New Zealand, 2017). If individual financial incentives contribute to the relationship between co-authorship and proposal team co-membership, then pooling researchers who can and cannot benefit from Marsden funding may bias our estimate of the strength of that relationship.

We overcome these challenges—large and sparse data, and heterogeneous incentives—by restricting our analysis to the 67,569 pairs of New Zealand researchers who collaborated at least once during our period of study, either as co-authors or as proposal team co-members.[9] We denote the set of such pairs by C.

[8] 17,494 pairs were both co-authors and proposal team co-members. Among these pairs, 41% were co-authors first, 46% were co-members first, and 13% were first-time co-authors and co-members in the same year.
[9] Lack of data prevents us from considering pairs who attempted joint work but neither published nor submitted Marsden Fund proposals jointly.

We also restrict our panel data to the years during which both researchers in each pair $\{i,j\} \in C$ were 'active'. We identify researchers as active during the years between their first publication or proposal, and their last publication or proposal.

We construct our observation set as follows. First, we let $T = \{2000, 2001, \ldots, 2018\}$ be the set of years for which we have publication and proposal data, and let A_t be the set of researchers in our data who were active in year $t \in T$. The Cartesian product

$$A_t \times A_t = \{\{i,j\}: i \in A_t \text{ and } j \in A_t\}$$

of A_t with itself contains all pairs of researchers who were both active in year t. By definition, every pair $\{i,j\} \in C$ must belong to the product $A_t \times A_t$ for some year $t \in T$. However, we want our panel data to include all years during which both i and j were active. Such years belong to the intersection

$$T_i \cap T_j = \{t \in T: \{i,j\} \in A_t \times A_t\}$$

of the sets $T_i = \{t \in T: i \in A_t\}$ and $T_j = \{t \in T: j \in A_t\}$ of years in which researchers i and j were active. Thus, our panel data contain observations of $coauth_{ijt}$ and x_{ijt} for each pair-year tuple $(\{i,j\}, t)$ belonging to the set

$$\{(\{i,j\}, t): \{i,j\} \in C \text{ and } t \in T_i \cap T_j \text{ and } t \geq 2010\}.$$

We identify 105,287 pairs of researchers who collaborated as co-authors or proposal team co-members during our period of study. Restricting our analysis to these pairs may bias our coefficient estimates if collaborating pairs' characteristics differ systematically from non-collaborating pairs'. Further restricting to the 67,569 pairs of New Zealand researchers who collaborated may exacerbate this bias. We investigate this possibility by comparing the mean and standard deviation of $coauth_{ijt}$ and our covariates within three panels: random pairs of researchers, pairs of researchers who ever collaborated, and pairs of New Zealand researchers who ever collaborated. We generate the panel of random pairs by uniformly sampling 105,287 pairs from the set of pairs of researchers who were ever active concurrently during our period of study. This random panel provides an unbiased estimate of our covariates' distributions among all pairs in our linked data who were active concurrently.

Table 1 presents the sample means and standard deviations of the variables in our random pair, collaborator, and New Zealand collaborator panels. We lose the first ten years of each panel as starting values for x_{ijt}. Consequently, pairs containing researchers who were inactive after 2009 fall out of our analysis. The variables $coauth_{ijt}$, $first_{ijt}$, $second_{ijt}$, and $funded_{ijt}$ have zero means in our panel of random pairs because most pairs of researchers in our data never collaborated.

Restricting to collaborators increases all four of these means, and restricting to New Zealand collaborators increases them further. These restrictions also increase the mean of $\overline{degree}_{ijt}$ by increasing the representation of researchers with high collaborative propensities. The means of $\overline{MNCS}_{ijt}$ suggest that New Zealand researchers in our data had lower citation impacts than non-New Zealanders. This could be because non-New Zealanders appear in our data only if they have collaborated with a New Zealander. Such international collaboration reveals a commitment to research and to pursuing topics of global interest, which may lead to an increase in mean-normalised citation scores.[10]

Table 1 shows that our panel of New Zealand researchers who collaborated differs statistically from the panel of all pairs in our data who were active concurrently. This difference arises due to non-random selection. However, the "both are New Zealanders" and "ever collaborated" criteria we use to select pairs are time-invariant. Therefore, we can control for selection bias by including pair-level fixed effects in the error term u_{ijt} in (2) because such effects are perfectly collinear with all time-invariant pair characteristics. We assess the robustness of our results to our panel selection criteria in subsection 6.3.

6 Logistic regression estimates

We estimate (2) using the panel of collaborating New Zealander pairs described in the right-most column of Table 1. For each variable $\theta_{ijt} \in \left\{\overline{degree}_{ijt}, \Delta degree_{ijt}, \overline{MNCS}_{ijt}, \Delta MNCS_{ijt}\right\}$, we define the zero-indicator variable

$$\phi(\theta_{ijt}) = \begin{cases} 1 & \text{if } \theta_{ijt} = 0 \\ 0 & \text{otherwise} \end{cases}$$

and the adjusted natural logarithm

$$\psi(\theta_{ijt}) = \begin{cases} 0 & \text{if } \theta_{ijt} = 0 \\ \ln(\theta_{ijt}) & \text{otherwise,} \end{cases} \tag{4}$$

which is well-defined because $\theta_{ijt} \geq 0$. We include $\phi(\theta_{ijt})$ and $\psi(\theta_{ijt})$ in x_{ijt} in place of θ_{ijt}. Thus, we have

$$\begin{aligned} x_{ijt}\beta = \beta_0 \\ +\beta_1 \cdot first_{ijt} + \beta_2 \cdot second_{ijt} + \beta_3 \cdot funded_{ijt} \end{aligned}$$

[10] Alternatively, international collaborators may have access to larger networks of colleagues among which to disseminate their research, leading to greater visibility of such research and more opportunities to attract citations.

$$
\begin{aligned}
&+\beta_4 \cdot adjacent_{ijt} \\
&+\beta_5 \cdot \phi\left(\overline{degree}_{ijt}\right) + \beta_6 \cdot \psi\left(\overline{degree}_{ijt}\right) \\
&+\beta_7 \cdot \phi\left(\Delta degree_{ijt}\right) + \beta_8 \cdot \psi\left(\Delta degree_{ijt}\right) \\
&+\beta_9 \cdot \phi\left(\overline{MNCS}_{ijt}\right) + \beta_{10} \cdot \psi\left(\overline{MNCS}_{ijt}\right) \\
&+\beta_{11} \cdot \phi\left(\Delta MNCS_{ijt}\right) + \beta_{12} \cdot \psi\left(\Delta MNCS_{ijt}\right) \\
&+\beta_{13} \cdot overlap_{ij} + \beta_{14} \cdot overlap_{ij}^2
\end{aligned}
$$

for each pair $\{i, j\}$ and year t, where $\beta_0, \beta_1, \ldots, \beta_{14}$ are coefficients to be estimated. Replacing our degree and citation covariates with their zero-indicators and adjusted natural logarithms allows us to capture the positive skew in the distributions of these covariates. Such skewness may arise from a cumulative advantage process (de Solla Price, 1976; Merton, 1968) wherein 'success breeds success.'

We lose the first ten years of our data as starting values for x_{ijt}. This leaves 247,110 observations among 46,052 pairs of New Zealand researchers between 2010 and 2018. Table 2 summarises the distributions of $coauth_{ijt}$ and our covariates across these observations. The table excludes the zero-indicator variables for our co-authorship propensity and citation covariates. However, the means of these excluded variables are captured by the "% zero" statistic for the corresponding adjusted natural logarithms.

Table 3 summarises our regression estimates. Column (1) shows how $coauth_{ijt}$ covaries with $first_{ijt}$, $second_{ijt}$, and $funded_{ijt}$. Pairs who were co-members on Marsden Fund proposal teams during the previous ten years were more likely to co-author than pairs who were not co-members during that period. Pairs who worked together on funded proposals were most likely to co-author. On average, pairs who worked on funded proposals were 8.1 percentage points more likely to co-author than pairs who worked on first and second round proposals but did not receive funding.[11]

Column (2) controls for pairs' collaborative propensities, citation impacts, and research field overlaps. Adding these controls changes the estimated coefficients on our proposal outcome variables, implying that our controls capture determinants of co-authorship that covary with proposal outcomes. In particular, our controls remove the co-authorship rate gain from working together on second round proposals and reduce the gain from working together on funded proposals. This could be because proposal outcomes are driven partially by pairs' match qualities, the observable components of which are captured by our control variables.

[11] We obtain this estimate by computing the average partial effect of a one unit increase in $funded_{ijt}$ using the model and data in column (1) of Table 3.

The strongest predictor of whether pairs co-authored in year t was whether they co-authored during the years $(t - 10)$ through $(t - 1)$, captured by the variable $adjacent_{ijt}$. The positive coefficient on $\psi\left(\overline{degree}_{ijt}\right)$ reflects a base rate phenomenon: pairs were more likely to co-author with each other if they co-authored more often in general. Likewise, the positive coefficient in $\psi(\overline{MNCS}_{ijt})$ implies that pairs were more likely to co-author if they were more productive.

The positive coefficient on $overlap_{ij}$ in column (2) suggests that pairs who published in more similar fields were more likely to co-author. We estimate a convex relationship between $coauth_{ijt}$ and $overlap_{ij}$, contradicting our hypothesised inverted U-shaped relationship. This may be due to our panel selection criteria, which truncate the distribution of research field overlaps among the pairs we analyse relative to the set of all pairs in our linked data.

Similarly, column (2) suggests a positive and significant relationship between co-authorship and $\Delta MNCS_{ijt}$, in contrast to Ahmadpoor and Jones' (2019) claim that teams tend to assemble among researchers with similar citation impacts. The positive coefficient on $\psi(\Delta MNCS_{ijt})$ implies negative assortative matching: pairs were more likely to co-author if their citation impacts differed than if they were equal. This result likely reflects the inter-generational nature of Marsden Fund proposal teams, which often comprise seasoned academics working with post-graduate students and post-docs.[12]

The estimates in columns (1) and (2) of Table 3 may be biased by unobservable pair- and year-specific factors that are correlated with our covariates of interest. They may also be biased from restricting our data to pairs of New Zealand researchers who ever collaborated during our period of study. We control for such biases by including pair and year fixed effects. However, our fixed effect estimator requires a dependent variable that, for each pair of researchers, varies across time. Thus, including fixed effects removes all pairs of researchers $\{i, j\}$ with no variation in $coauth_{ijt}$. This leaves 124,619 observations among 19,091 pairs.[13] Table 2 reports descriptive statistics for this restricted panel. The means of $coauth_{ijt}$ and $adjacent_{ijt}$ are larger in our restricted panel than in our full panel because our restricted panel excludes pairs who never co-author.

Including fixed effects changes both the observations *and* the source of variation used to identify coefficient estimates. We present the intermediate effect of restricting to pairs with variation in $coauth_{ijt}$ in columns (3) and (4) of Table 3, which re-estimate the models in columns (1) and (2)

[12] During our period of study, 61% of funding contracts awarded had budget for post-graduate students and 36% had budget for post-docs.

[13] Among the 26,961 pairs with no variation in $coauth_{ijt}$, 1,625 (6%) co-authored every year in which they were active concurrently. Among these 1,625 pairs, 726 (44.7%) were active concurrently for one year only.

using our restricted panel. Columns (2) and (4) present qualitatively similar patterns. Pairs who were co-members on first round or funded proposal teams were more likely to co-author. In contrast, pairs who were co-members on second round teams, but not funded teams, were not significantly more likely to co-author. These patterns may reflect successful identification of productive teams by assessment panels. If more productive teams were more likely to receive funding then pairs with $second_{ijt} = 1$ but $funded_{ijt} = 0$ will, all else equal, be less likely to co-author, driving a negative coefficient on $second_{ijt}$. However, it is unclear whether panels can successfully identify productive teams in the second round (Gush et al., 2018).

Column (5a) introduces pair and year fixed effects. This allows us to identify the effects of cross-sectional variation in pairs' characteristics, controlling for the components of those characteristics that do not vary over time. However, we lose the ability to identify coefficients on $overlap_{ij}$ and $overlap_{ij}^2$ separately because these covariates are perfectly collinear with our pair fixed effects. Including fixed effects may bias our coefficient estimates due to the incidental parameters problem (Neyman and Scott, 1948): our data contain at most nine observations with which to estimate each pair fixed effect. We correct for such bias using the analytical procedure suggested by Fernández-Val and Weidner (2016).

To ease interpretation, column (5b) reports average partial affects (APEs) using the estimated model and data in column (5a). These APEs estimate the mean change in $coauth_{ijt}$ resulting from a one-unit increase in each covariate while holding other covariates constant. For example, the APE of 0.138 associated with $first_{ijt}$ implies that, on average and holding all else constant, pairs who co-submitted first round proposals during the years $(t - 10)$ through $(t - 1)$ were 13.8 percentage points more likely to co-author in year t than pairs who did not co-submit such proposals during those years. In contrast, $second_{ijt}$ and $funded_{ijt}$ have small and insignificant APEs. These patterns suggest that applying for Marsden Funding promoted co-authorship, but more successful applicants were not significantly more likely to co-author than less successful applicants when we control for observable and unobservable heterogeneity. This result may reflect a self-selection effect, wherein pairs with greater innate match qualities were more likely to submit proposals and receive funding. Including pair fixed effects controls this effect, washing out the association of co-authorship with proposal submission and funding receipt. The positive coefficient on $first_{ijt}$ would then capture another self-selection effect: pairs were more likely to submit proposals when they shared research ideas worth pursuing to publication independently of their proposals' outcomes.

Columns (5a) and (5b) in Table 3 also show how $coauth_{ijt}$ covaries with co-authorship propensities and citation impacts after controlling for unobservable pair- and year-specific factors. Pairs who co-

authored during the years $(t - 10)$ through $(t - 1)$ were 32.8 percentage points less likely to co-author in year t than pairs who did not co-author during those years, on average and holding our other covariates constant. In contrast, pairs with greater mean collaboration propensities and citation impacts were more likely to co-author. Together, these patterns suggest that pairs were more likely to co-author if they had not co-authored recently, and if they were at a stage in their careers when they were co-authoring widely and publishing impactfully.

While suggestive, the results presented in Table 3 do not establish a causal link between proposal outcomes and co-authorship. First, the Marsden Fund application process may simply act as a screen that filters out unsuccessful collaborations. Preparing proposals allows researchers to trial collaborations before committing to research projects. These trials allow researchers to learn about their match qualities and, consequently, whether they would be likely to collaborate productively. Teams among researchers with low quality matches may choose not to submit proposals, leaving only those collaborations likely to generate co-authored publications. In this way, our finding that proposal submission covaries positively and significantly with co-authorship rates may represent a self-selection effect rather than a treatment effect. Controlling for citation impacts, collaborative propensities, research overlaps, and pair fixed effects helps control for variation in match qualities, and isolate the variation in whether pairs submitted proposals and received funding. However, this isolation may be partial only.

Second, even if our covariates isolate the variation in proposal outcomes fully, we cannot rule out reverse causality. Pairs may co-author to demonstrate their ability to collaborate productively and, consequently, improve their chances of receiving funding. Alternatively, pairs may apply for funding for ongoing research projects (or derivatives of those projects) that generate publications soon after the application process ends, independently of the application's outcome. If these behaviours are systematic among the researchers in our panel data then our coefficient estimates will be biased (upwards) due to endogeneity. We address this issue in the following subsection.

6.1 Varying publication lags

Our estimates in Table 3 come from modelling $coauth_{ijt}$ as a function of pair $\{i, j\}$'s characteristics in year $(t - 1)$. However, decisions to co-author may be made several years before we observe such co-authorship in our data. Research projects can take years to complete, submit to journals, and pass through peer review. Thus, to the extent that our covariates x_{ijt} capture factors relevant to researchers' co-authorship decisions, we may obtain better estimates of these factors' strengths by lagging our independent variables by more than one year.

Table 4 reports coefficient estimates for the model in column (5a) of Table 3 when we lag our covariates by one, two, three, and four years. Each additional lag drops an additional year of data. Moreover, because our fixed effect estimator requires variation in $coauth_{ijt}$ across years t, each lag also drops pairs with no such variation during the remaining years.[14]

Table 4 shows that $coauth_{ijt}$ covaries strongly with $first_{ijt}$ for short publication lags only. In contrast, $coauth_{ijt}$ covaries strongly with $funded_{ijt}$ when we assume three or four year publication lags. We attribute this shift in strength to two effects. First, increasing the lag between $coauth_{ijt}$ and our proposal outcome dummies reduces the impact of endogeneity bias because most collaborations pursued prior to submitting proposals would be complete before we observe $coauth_{ijt}$'s value. Second, increasing the lag between $coauth_{ijt}$ and x_{ijt} more accurately captures the time it takes to conduct and publish the research outlined in researchers' Marsden Fund proposals.[15] For these two reasons, we believe that the right-most columns in Table 4 provide our closest estimates of the 'treatment effect' of different proposal outcomes on co-authorship rates.

The patterns in Table 4 suggest that our results are unlikely to reflect self-selection into the Marsden Fund application process by pairs with high innate match qualities. If such selection was systematic in our data, then the relationship between co-authorship and proposal outcomes, conditional on innate match quality (captured by our pair fixed effects), would not change when we introduce additional lags in our independent variables. However, Table 4 shows precisely such change. Likewise, if the positive coefficient on $first_{ijt}$ in columns (5a) and (5b) of Table 3 represents only the fact that productive collaborators submitted proposals when they had ideas to pursue, then that positive coefficient should persist when we allow more time for pairs' ideas to spawn co-authored publications. In contrast, Table 4 shows that the relationship between co-authorship and proposal submission disappears when we allow for publication delays.

However, we cannot rule out selection effects entirely. To receive Marsden Funding, proposal teams must have a high-quality research idea, and demonstrate serious commitment to working together on that idea and pursuing it to publication. Thus, one might expect that funded teams may have been more likely to co-author independently of funding receipt. This "selectivity problem" (Jaffe, 2002, p. 22) is common to evaluations of all research funding mechanisms in which funding is awarded to proposals judged in advance most likely to succeed. However, we find no persistent

[14] Re-estimating the model with a one-year lag among pairs remaining after taking each additional lag generates quantitatively and qualitatively similar estimates. Therefore, the patterns shown in Table 4 are unlikely to be an artefact of restricting the set of pairs we analyse.

[15] Most Marsden grants are for three years, so if funding itself facilitates publication then four years makes sense as a publication lag.

'effect' of proposal submission or progression to the second round, but a significant longer-term effect of funding receipt. Viewing our results in combination with Gush et al. (2018), who find that second round assessment panels show no tendency to award higher ranks to proposals that subsequently have the most publication success, there appears to be little room for selection processes to explain away our results.

6.2 Comparison with previous studies

Ayoubi et al. (2018) find that Swiss scientists are more likely to co-author with research grant co-applicants if their proposals receive funding. However, the authors do not control for whether scientists applied for funding when estimating this effect. Thus, their estimate compares funded applicants to the combined pool of unfunded applicants and non-applicants, rather than the more relevant comparison of funded to unfunded applicants that we present in Tables 3 and 4. Table 4 shows that Ayoubi et al.'s finding is consistent with the patterns in our data when we allow more time for proposal outcomes to influence co-authorship rates.

Using Web of Science data on journal articles published between 1945 and 2005, Ahmadpoor and Jones (2019) find that teams tend to assemble among researchers with similar citation impacts. If this were true for researchers in our data then we would expect a negative and significant coefficient on $\psi(\Delta MNCS_{ijt})$. In contrast, we estimate a positive coefficient on $\psi(\Delta MNCS_{ijt})$ using the in column (2) of Table 3, which is closest of our models to the analysis of time-invariant characteristics that Ahmadpoor and Jones conduct.[16] However, our estimate is economically small: it implies that, on average and holding all else constant, doubling the difference in pairs' citation impacts corresponds to a 0.4 percentage point rise in the probability of co-authorship.

Ahmadpoor and Jones (2019) do not control for pair-level factors that covary with similarities in citation impact. One such factor is whether pairs co-authored a highly cited paper. If this occurs often then pairs with many citations will appear to work together, but only because they attracted their citations *while* working together rather than because prior citations were an attractive force. We control for this scenario by including $adjacent_{ijt}$ as a covariate in (2). Fafchamps et al. (2010) control for pair characteristics in a similar logistic regression setting to ours and estimate similarly weak sorting with respect to citation impacts.

Finally, the coefficient on $\psi(\Delta MNCS_{ijt})$ is insignificant in all the models presented in Table 4. Such insignificance suggests that researchers in our data may sort into teams based on the time-constant

[16] Introducing pair fixed effects controls for time-invariant characteristics, meaning that our coefficients are identified using within-pair variation in researchers' characteristics over time.

component of citation impacts but not the time-varying component, consistent with Ahmadpoor and Jones' analysis of time-invariant characteristics. However, our estimates imply negative sorting with respect to such characteristics in our data, whereas Ahmadpoor and Jones (2019) find positive sorting in their Web of Science data. We encourage further research on this issue to advance our understanding of the assortative mechanisms through which researchers form teams.

6.3 Robustness tests

We believe that our coefficient estimates in the right-most column of Table 4 are the closest to the coefficients in the true data generating process. Therefore, we focus on these estimates throughout our robustness tests.

6.3.1 Adjusting standard errors for dyadic clustering

The standard errors in Tables 3 and 4 may be biased by non-independence among observations in our data. Such non-independence arises due to dyadic clustering (Aronow et al., 2017; Graham, 2020). The co-authorship rates for pairs $\{i, j\}$ and $\{j, k\}$ covary because they share researcher j in common, whose collaboration choices affect both $coauth_{ijt}$ and $coauth_{jkt}$. Thus, our estimation errors are likely to be correlated across pair observations. Failing to control for such correlation may lead us to under-estimate our standard errors and, consequently, over-estimate the statistical significance of our coefficients (Cameron and Miller, 2014).

Estimating dyadic cluster-robust standard errors requires estimating the variance of our coefficient estimates under dyadic clustering. Study of such variance estimators began only recently (Graham, 2020). Aronow et al. (2017) propose the sandwich-type estimator

$$\hat{V} = (X^T M X)^{-1}(X^T(P * \varepsilon\varepsilon^T)X)(X^T M X)^{-1},$$

where X is the design matrix, X^T is the transpose of X, $M = \text{diag}(p_{ijt}(1 - p_{ijt}))$ is a diagonal matrix computed using the predicted probabilities $p_{ijt} = \Lambda^{-1}(x_{ijt}\hat{\beta})$, P is a square matrix with rs^{th} entry

$$P_{rs} = \begin{cases} 1 & \text{if the pairs associated with observations } r \text{ and } s \text{ share a common researcher} \\ 0 & \text{otherwise,} \end{cases}$$

ε is the vector of residuals, and $P * \varepsilon\varepsilon^T$ is the element-wise product of the matrices P and $\varepsilon\varepsilon^T$. This product captures the covariances of the errors associated with observations of pairs with common researchers. Aronow et al. show that $\hat{V}$ is a consistent estimator for the variance of the vector $\hat{\beta}$ of

coefficient estimates under dyadic clustering. Thus, we can obtain dyadic cluster-robust standard error estimates by computing $\hat{V}$ and taking the square roots of its diagonal entries.[17]

Table 5 compares the standard errors on the coefficient estimates in the right-most column of Table 4 before and after adjusting for dyadic clustering. Adjusting for dyadic clustering increases our standard errors by a factor of about 1.25–2. However, these increases barely affect our inferences because the unadjusted standard errors are small.

6.3.2 Relaxing selection criteria

Our estimates in Table 4 come from analysing pairs of New Zealand researchers who ever collaborated during our period of study. However, New Zealanders may differ systematically in their co-authorship patterns to non-New Zealanders. Likewise, restricting our data to pairs who collaborate may lead us to draw different inferences than we would from analysing the entire set of potential pairs if such analysis were computationally feasible.

We investigate these possibilities as follows. First, we re-estimate the model in the right-most column of Table 4 among the researcher pairs in the "Collaborators" panel described in Table 1, before and after including fixed effects. We present the resulting coefficient estimates in columns (3) and (4) of Table 6. To ease comparison, we present coefficients estimated using the "NZ collaborators" panel in columns (1) and (2). Expanding our data to include pairs containing non-New Zealand researchers produces quantitatively and qualitatively similar estimates. Thus, it appears that our inferences are not sensitive to restricting to pairs of New Zealand researchers.

Second, we re-estimate the model in the right-most column of Table 4 among the pooled set of researcher pairs in the "Random pairs" and "Collaborators" panels described in Table 1. We report the resulting coefficient estimates in column (5) of Table 6. Including pairs from the "Random pairs" panel weakens the selection on observable collaboration while maintaining computational feasibility.[18] Pooling the "Random pairs" and "Collaborators" panels preserves the positive association between co-authorship and funding receipt. Relaxing our selection criteria increases the

[17] Computing $\hat{V}$ is very computer-intensive when there are many observations and covariates. With n observations and k covariates, computing the $k \times k$ "bread" matrix $(X^T M X)^{-1}$ requires multiplying the $n \times n$ matrix M by the $n \times k$ matrix X, pre-multiplying the result by the $k \times n$ matrix X^T, and inverting the resulting matrix. Computing the $k \times k$ "meat" matrix $X^T(P * \varepsilon\varepsilon^T)X$ requires similar multiplications but no inversion. The bread and meat matrices can be computed efficiently when n and k are small, or when X is sparse. But the bread and meat matrices are generally not sparse themselves. Therefore, computing $\hat{V}$ requires multiplying three dense $k \times k$ matrices, which is computer-intensive when k is large (e.g., when there are many fixed effects).

[18] We do not estimate a fixed effects model among pairs in the pooled panel because we would have to drop pairs with no variation in $coauth_{ijt}$, which includes all pairs who never collaborate. Therefore, we would obtain the same coefficient estimates as in column (4) of Table 6.

coefficient on $first_{ijt}$ and decreases the coefficient on $second_{ijt}$. Both coefficients are significant at the 5% level using our pooled panel data. However, this significance disappears when we control for pair and year fixed effects.

The coefficients on $adjacent_{ijt}$, $\psi\left(\overline{degree}_{ijt}\right)$, and $\psi\left(\overline{MNCS}_{ijt}\right)$ are larger in column (5) than in columns (1) and (3). This is likely because the co-authorship network among researchers in the pooled pair panel is relatively sparse, so variables that covary with local network density provide relatively strong signals of which pairs are more likely to co-author. Column (5) shows a concave relationship between co-authorship and research overlap, consistent with our initial hypothesis. Thus, the convex relationships we estimate in columns (1) and (3) appear to arise from restricting our data to collaborating pairs, which truncates the distribution of research field overlaps.

7 Conclusion

Scientific collaboration has grown in incidence and importance. Consequently, there has been an increasing interest in understanding its determinants and consequences. Researchers' decisions to collaborate are inextricably linked with their decisions regarding research topics and efforts to secure funding for their research. In this paper, we focus on the interactions among the seeking of research funding, success in winning grants, and the publication of co-authored research papers. Because each of these events is related positively to the quality of research ideas, to inclinations to collaborate, and to the strength or success of collaborations, the events are all positively associated with each other. We demonstrate that such associations exist, and we use the structure of our data and of the New Zealand research funding process to tease out the extent to which participation in that process may increase co-authorship causally.

New Zealand provides a useful laboratory for studying these issues. This is, in part, because the RSNZ has kept excellent data on all Marsden Fund applications and makes those data available for research purposes. Further, because the Marsden Fund is central to the research enterprise in New Zealand, the set of researchers who interacted with the Fund during our period of study provides a representative sample of New Zealand's research system overall. Finally, New Zealand is a small, isolated country, which provides all of its researchers and scientists roughly equal access to collaborators. To exploit this situation, we construct a dataset comprised of almost all researchers who interacted with the Marsden Fund between the years 2000 and 2018, their proposals and publications during that period, and their collaborators on such proposals and publications.

Co-authorship rose among the researchers in our data during our period of study. The extent of co-authorship was highest among researchers who received Marsden Funding, lowest among

researchers who did not apply for funding, and at an intermediate level among researchers who applied for, but did not receive, funding.

We analyse pairs of collaborators over time, allowing us to identify both (i) innate attributes of pairs that affect their propensity to co-author and (ii) time-varying factors associated with the probability that they co-author in a given year. Pairs were more likely to co-author in a given year if they had co-authored previously, if they co-authored with others often, if they published in similar fields, or if their prior publications were more highly cited. Interestingly, pairs were also more likely to co-author if their prior citation impacts differed, which cuts against an expectation of assortative matching of researchers of similar prestige into teams.

Turning to the relationship between co-authorship and Marsden Fund proposal outcomes, we find that pairs who co-submitted proposals were more likely to co-author and that more successful co-submissions were associated with greater co-authorship rates. When we control for observable and unobservable researcher and pair characteristics using pair fixed effects, we find that there remains an association between co-authorship and proposal submission, but no further effect of funding receipt. In contrast, when we increase the delay between proposal outcomes and potential co-authorship, we find no association with submission but a significant association with funding.

Our results suggest that funding receipt increased co-authorship rates causally. We justify this claim by noting (i) the positive association between funding and delayed co-authorship, (ii) the lack of association between funding and contemporaneous co-authorship, and (iii) evidence from previous analysis (Gush et al., 2018) that assessment panels cannot identify which second round proposal teams are most likely to generate successful publications. However, we cannot determine whether this 'funding effect' comes directly from the resources conveyed, or indirectly through the generic benefits associated with the signal and prestige of winning a Marsden grant.[19]

Our results complement and amplify those of Ayoubi et al. (2018), who find that funded grant applicants are more likely to co-author than other researchers, but do not compare co-authorship among funded teams with that of unfunded teams. Our results also connect to those of Gush et al (2018), who find that Marsden Funding increases the overall publication and citation rates of teams relative to that of unsuccessful applicant teams, but do not examine the collaboration patterns underlying that effect. Together, these findings raise the question of whether funding, in addition to

[19] We do not analyse whether the funding effect raises co-authorship overall. Funded proposal team co-members may co-author with each other instead of potential collaborators outside the Marsden Fund application process. Such potential substitution effects are intrinsic to the partial equilibrium analysis that we conduct.

increasing the likelihood of co-authored publication, also increases the *quality* of co-authored publications (e.g., as measured by citations or journal impact factor). We leave this question for future research.

We think of co-authorship as situating researchers and scientists within a complex network of collaborative relationships. However, we do not analyse in detail how this network evolves. It is interesting that the pairwise propensity to co-author is associated with co-authorship overall, even after controlling for pair fixed effects. This hints at a preferential attachment process in which experience with collaboration breeds further collaboration. However, such a process is difficult to distinguish from one in which unobserved propensity to collaborate changes over time for other reasons. Data of the kind constructed in this paper could, in principle, permit investigation of these more complex dynamics.

Finally, we say nothing about exactly why pairs of researchers (or larger teams) decide to collaborate. Presumably, collaboration facilitates the combination of complementary knowledge and skills (in addition, perhaps, to simply making research more enjoyable). Understanding the nature and significance of these complementarities would require granular and complex information about individual researchers that is hard to conceive, let alone collect. However, at some point, further advancing our understanding of team formation processes will require taking on this data challenge.

References

Adams, J. D., Black, G. C., Clemmons, J. R., and Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999. *Research Policy*, 34(3):259–285.

Agrawal, A. and Goldfarb, A. (2008). Restructuring research: communication costs and the democratization of university innovation. *American Economic Review*, 98(4):1578–1590.

Agrawal, A., Goldfarb, A., and Teodoridis, F. (2016). Does knowledge accumulation increase the returns to collaboration? *American Economic Journal: Applied Economics*, 8(1):100–128.

Ahmadpoor, M. and Jones, B. F. (2019). Decoding team and individual impact in science and invention. *Proceedings of the National Academy of Sciences*, 116(28):13885–13890.

Aref, S., Friggens, D., and Hendy, S. (2008). Analysing scientific collaborations of New Zealand institutions using Scopus bibliometric data. Proceedings of the Australasian Computer Science Week Multiconference.

Aronow, P. M., Samii, C., and Assenova, V. A. (2017). Cluster-robust variance estimation of dyadic data. *Political Analysis*, 23(4):564–577.

Ayoubi, C., Pezzoni, M., and Visentin, F. (2018). The important thing is not to win, it is to take part: What if scientists benefit from participating in research grant competitions? *Research Policy*, 48(1):84–97.

Barnett, A. H., Ault, R. W., and Kaserman, D. L. (1988). The rising incidence of co-authorship in economics: Further evidence. *Review of Economics and Statistics*, 70(3):539–543.

Bikard, M., Murray, F., and Gans, J. S. (2015). Exploring trade-offs in the organisation of scientific work: Collaboration and scientific reward. *Management Science*, 61(7):1473–1495.

Boudreau, K. J., Brady, T., Ganguli, I., Gaule, P., Guinan, E., Hollenberg, A., and Lakhani, K. R. (2017). A field experiment on search costs and the formation of scientific collaborations. *Journal of Economics and Statistics*, 99(4):565–576.

Cameron, A. C. and Miller, D. L. (2014). Robust inference for dyadic data. Technical report, University of California, Davis.

Catalini, C., Fons-Rosen, C., and Gaulé, P. (2019). How do travel costs shape collaboration? *Management Science*, 66(8):3295–3798.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.

Defazio, D., Lockett, A., and Wright, M. (2009). Funding incentives, collaborative dynamics and scientific productivity: Evidence from the EU framework program. *Research Policy*, 38(2):293–305.

Ductor, L. (2015). Does co-authorship lead to higher academic productivity? *Oxford Bulletin of Economics and Statistics*, 77(3):385–407.

Ebadi, A. and Schiffauerova, A. (2013). Impact of funding on scientific output and collaboration: A survey of literature. *Journal of Information and Knowledge Management*, 12(4):1350037.

Fafchamps, M., Goyal, S., and van der Leij, M. J. (2010). Matching and network effects. *Journal of the European Economic Association*, 8(1):203–231.

Fernández-Val, I. and Weidner, M. (2016). Individual and time effects in nonlinear panel models with large N, T. *Journal of Econometrics*, 192(1):291–213.

Graham, B. S. (2020). Dyadic regression. In Graham, B. S. and de Paula, A., editors, *The Econometric Analysis of Network Data*. Academic Press, Amsterdam.

Guimerà, R., Uzzi, B., Spiro, J., and Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702.

Gush, J., Jaffe, A., Larsen, V., and Laws, A. (2018). The effect of public funding on research output: the New Zealand Marsden Fund. *New Zealand Economic Papers*, 52(2):227–248.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1):162–172.

Jaffe, A. B. (2002). Building program evaluation into the design of public research-support programmes. *Oxford Review of Economic Policy*, 18:22–34.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": is innovation getting harder? *Review of Economic Studies*, 76(1):283–317.

Katz, J. S. and Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1):1–18.

Kim, E. H., Morse, A., and Zingales, L. (2009). Are elite universities losing their competitive edge? *Journal of Financial Economics*, 93(3):353–381.

McDowell, J. M. and Melvin, M. (1983). The determinants of co-authorship: An analysis of the economics literature. *Review of Economics and Statistics*, 65(1):155–160.

Merton, R. K. (1968). The Matthew effect in science. *Science*, 159(3810):56–63.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20):208701.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

Rivera, M. T., Soderstrom, S. B., and Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Review of Sociology*, 36:91–115.

Rosenblat, T. S. and Mobius, M. M. (2004). Getting closer or drifting apart? *Quarterly Journal of Economics*, 119(3):971–1009.

Royal Society of New Zealand (2017). Marsden Fund Terms of Reference. Retrieved from https://www.royalsociety.org.nz/what-we-do/funds-and-opportunities/marsden/about/tor/ on June 11, 2020.

Ubfal, D. and Maffioli, A. (2011). The impact of funding on research collaboration: Evidence from a developing country. *Research Policy*, 4(9):1269–1279.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., and van Raan, A. F. J. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics*, 87(3):467–481.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

Figures and tables

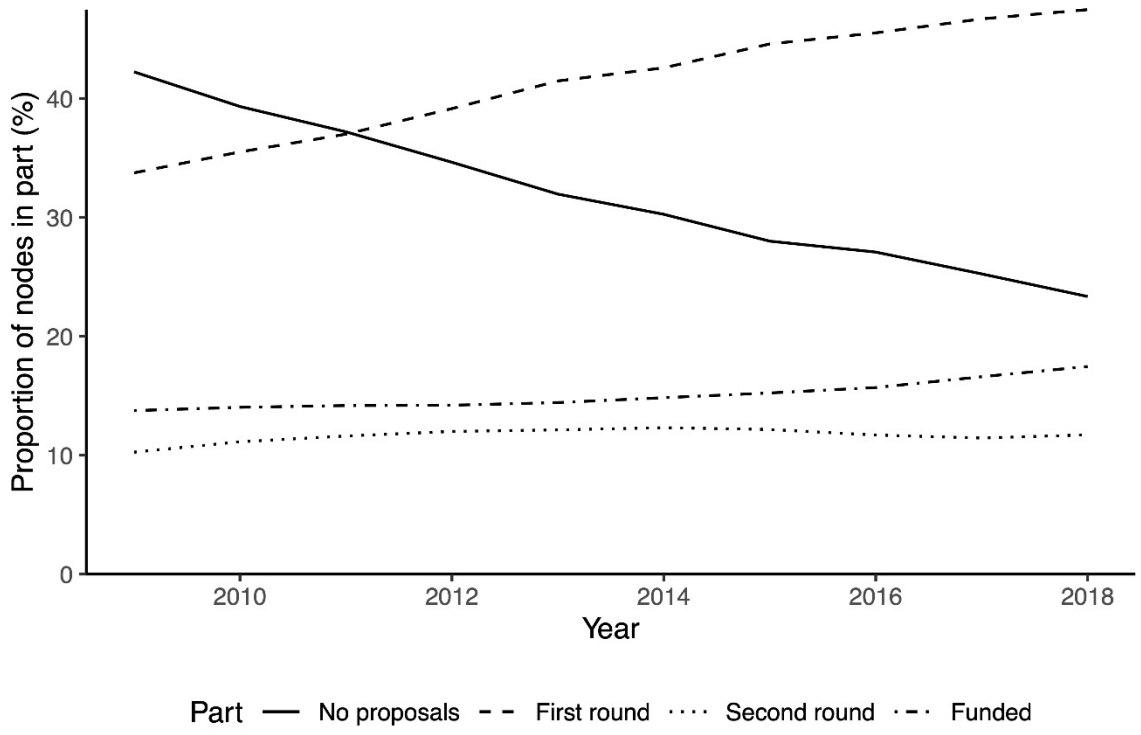

Figure 1: Proportion $|P|/|V_t|$ of researchers in each part $P \in \mathcal{P}_t$

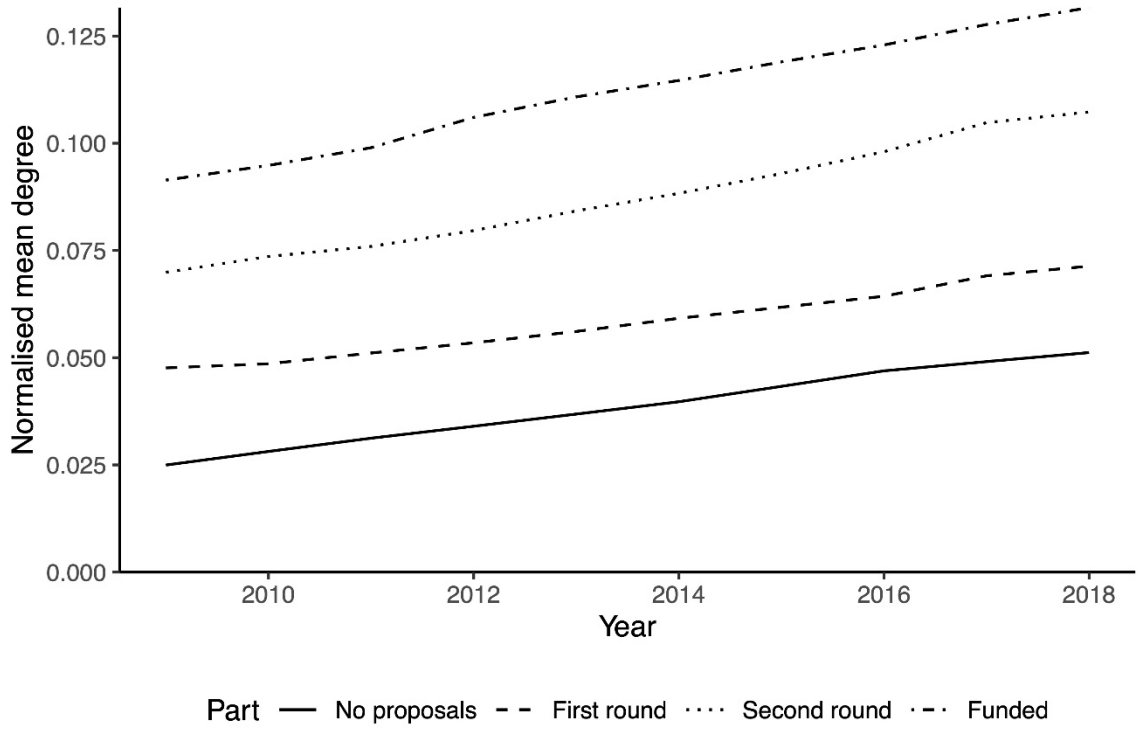

Figure 2: Normalised mean degree among researchers in each part $P \in \mathcal{P}_t$

Table 1: Variable means in random pair, collaborator, and NZ collaborator panels

Variable	Random pairs	Collaborators	NZ collaborators
Co-authored in year t ($coauth_{ijt}$)	0.000 (0.016)	0.142 (0.349)	0.153 (0.360)
First round co-members ($first_{ijt}$)	0.000 (0.015)	0.160 (0.367)	0.187 (0.390)
Second round co-members ($second_{ijt}$)	0.000 (0.008)	0.068 (0.251)	0.077 (0.267)
Funded co-members ($funded_{ijt}$)	0.000 (0.007)	0.037 (0.188)	0.043 (0.203)
Adjacent in co-auth. network ($adjacent_{ijt}$)	0.001 (0.023)	0.351 (0.477)	0.379 (0.485)
Mean degree ($\overline{degree}_{ijt}$)	5.355 (5.663)	17.341 (14.231)	19.610 (14.816)
Diff. in degrees ($\Delta degree_{ijt}$)	6.499 (8.419)	15.101 (15.965)	16.079 (16.488)
Mean citation impact ($\overline{MNCS}_{ijt}$)	5.330 (8.771)	17.009 (22.073)	14.303 (18.555)
Diff. in citation impacts ($\Delta MNCS_{ijt}$)	6.897 (15.892)	19.532 (34.797)	16.569 (30.100)
Research field overlap ($overlap_{ij}$)	0.049 (0.122)	0.513 (0.293)	0.481 (0.291)
Observations	312,279	355,911	247,110
Pairs	74,529	68,367	46,052
Researchers	3,398	8,997	5,823

Notes: Sample standard deviations in parentheses. NZ = New Zealand.

Table 2: Researcher pair panel descriptive statistics

Variable	Full panel		Restricted panel	
	Mean (s.d.)	% zero	Mean (s.d.)	% zero
Co-authored in year t ($coauth_{ijt}$)	0.153 (0.360)	84.749	0.266 (0.442)	73.431
First round co-members ($first_{ijt}$)	0.187 (0.390)	81.312	0.156 (0.363)	84.392
Second round co-members ($second_{ijt}$)	0.077 (0.267)	92.257	0.078 (0.268)	92.240
Funded co-members ($funded_{ijt}$)	0.043 (0.203)	95.681	0.050 (0.217)	95.034
Adjacent in co-auth. network ($adjacent_{ijt}$)	0.379 (0.485)	62.092	0.509 (0.500)	49.131
Log mean degree ($\psi(\overline{degree}_{ijt})$)	2.632 (0.947)	1.843	2.760 (0.893)	1.136
Log diff. in degrees ($\psi(\Delta degree_{ijt})$)	2.236 (1.153)	9.946	2.315 (1.141)	8.744
Log mean citation impact ($\psi(\overline{MNCS}_{ijt})$)	2.168 (1.024)	0.045	2.237 (0.978)	0.017
Log diff. in citation impacts ($\psi(\Delta MNCS_{ijt})$)	1.902 (1.485)	0.057	1.949 (1.467)	0.032
Research field overlap ($overlap_{ij}$)	0.481 (0.291)	0.492	0.519 (0.284)	0.000
Observations	247,110		124,619	
Pairs	46,052		19,091	

Notes: Sample standard deviations in parentheses. ψ denotes adjusted natural logarithm (4).

Table 3: Logistic regression estimates

Dependent variable: Co-authored in year t ($coauth_{ijt}$)						
	All pairs (coefficients)		Pairs with variation in $coauth_{ijt}$			
			Coefficients			APEs
	(1)	(2)	(3)	(4)	(5a)	(5b)
First round co-members	0.248^{***}	0.223^{***}	0.728^{***}	0.650^{***}	0.715^{***}	0.138^{***}
	(0.017)	(0.018)	(0.022)	(0.022)	(0.057)	(0.012)
Second round co-members	0.137^{***}	0.013	-0.017	-0.046	-0.087	-0.015
	(0.029)	(0.030)	(0.037)	(0.037)	(0.099)	(0.016)
Funded co-members	0.631^{***}	0.409^{***}	0.129^{**}	0.101^{*}	0.086	0.015
	(0.031)	(0.032)	(0.040)	(0.040)	(0.125)	(0.021)
Adjacent in co-auth. network		0.881^{***}		0.067^{***}	-2.193^{***}	-0.328^{***}
		(0.012)		(0.014)	(0.031)	(0.015)
Log mean degree		0.062^{***}		0.046^{***}	0.944^{***}	0.163^{***}
		(0.009)		(0.010)	(0.044)	(0.008)
Log diff. in degrees		0.016^{*}		-0.001	-0.016	-0.003
		(0.007)		(0.008)	(0.017)	(0.003)
Log mean citation impact		0.032^{***}		0.019	0.184^{***}	0.032^{***}
		(0.010)		(0.011)	(0.042)	(0.007)
Log diff. in citation impacts		0.032^{***}		0.033^{***}	0.014	0.002
		(0.006)		(0.007)	(0.014)	(0.002)
Research field overlap		0.581^{***}		0.205^{*}		
		(0.086)		(0.099)		
Squared research field overlap		0.596^{***}		0.522^{***}		
		(0.081)		(0.093)		
Pair and year fixed effects					Yes	Yes
Observations	247,110	247,110	124,619	124,619	124,619	124,619
Pairs	46,052	46,052	19,091	19,091	19,091	19,091
Log-likelihood	-104,531.110	-98,515.945	-71,096.536	-70,469.881	-56,206.543	-56,206.543

Notes: Standard errors in parentheses ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Intercept and zero-indicators suppressed. Column (5b) reports average partial effects for the model in (5a).

Table 4: Coefficient estimates with varying publication lags

Dependent variable: Co-authored in year t ($coauth_{ijt}$)				
	Publication lags (years)			
	One	Two	Three	Four
First round co-members	0.715^{***}	0.324^{***}	0.046	-0.173
	(0.057)	(0.066)	(0.078)	(0.097)
Second round co-members	-0.087	-0.218	-0.141	-0.128
	(0.099)	(0.114)	(0.132)	(0.159)
Funded co-members	0.086	0.268	0.501^{**}	0.635^{***}
	(0.125)	(0.137)	(0.157)	(0.191)
Adjacent in co-auth. network	-2.193^{***}	-1.902^{***}	-1.612^{***}	-1.155^{***}
	(0.031)	(0.034)	(0.041)	(0.054)
Log mean degree	0.944^{***}	0.703^{***}	0.383^{***}	0.313^{***}
	(0.044)	(0.048)	(0.053)	(0.063)
Log diff. in degrees	-0.016	-0.001	0.035	0.006
	(0.017)	(0.019)	(0.022)	(0.027)
Log mean citation impact	0.184^{***}	0.279^{***}	0.302^{***}	0.457^{***}
	(0.042)	(0.047)	(0.054)	(0.065)
Log diff. in citation impacts	0.014	-0.019	-0.014	-0.024
	(0.014)	(0.015)	(0.017)	(0.020)
Observations	124,619	97,450	73,263	51,841
Pairs	19,091	16,620	14,060	11,408
Years	9	8	7	6
Log-likelihood	-56,206.543	-46,741.286	-37,486.585	-28,170.943

Notes: Standard errors in parentheses ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Intercept and zero-indicators suppressed. All columns include pair and year fixed effects.

Table 5: Adjusting standard errors for dyadic clustering

		Standard errors	
Covariate	Coefficients	Unadjusted	Adjusted
First round co-members	−0.173	0.097	0.132
Second round co-members	−0.128	0.159	0.234
Funded co-members	0.635	0.191***	0.297*
Adjacent in co-auth. network	−1.155	0.054***	0.080***
Log mean degree	0.313	0.063***	0.109**
Log diff. in degrees	0.006	0.027	0.035
Log mean citation impact	0.457	0.065***	0.125***
Log diff. in citation impacts	−0.024	0.020	0.025

Notes: Based on model in right-most column of Table 4. One, two, and three stars denote significance at the 5%, 1%, and 0.1% levels. Intercept and zero-indicators suppressed.

Table 6: Coefficient estimates with relaxed selection criteria

Dependent variable: Co-authored in year t ($coauth_{ijt}$)					
	NZ collaborators		Collaborators		Pooled pairs
	(1)	(2)	(3)	(4)	(5)
First round co-members	-0.017	-0.173	0.052^{*}	-0.170	0.163^{***}
	(0.028)	(0.097)	(0.026)	(0.090)	(0.027)
Second round co-members	-0.031	-0.128	-0.039	-0.103	-0.113^{*}
	(0.047)	(0.159)	(0.044)	(0.147)	(0.045)
Funded co-members	0.627^{***}	0.635^{***}	0.578^{***}	0.516^{**}	0.574^{***}
	(0.047)	(0.191)	(0.045)	(0.170)	(0.046)
Adjacent in co-auth. network	0.238^{***}	-1.155^{***}	0.380^{***}	-1.115^{***}	0.443^{***}
	(0.017)	(0.054)	(0.015)	(0.048)	(0.015)
Log mean degree	-0.015	0.313^{***}	0.045^{***}	0.260^{***}	0.146^{***}
	(0.012)	(0.063)	(0.010)	(0.052)	(0.010)
Log diff. in degrees	0.050^{***}	0.006	0.018^{*}	0.024	0.017^{*}
	(0.009)	(0.027)	(0.008)	(0.024)	(0.008)
Log mean citation impact	0.024	0.457^{***}	-0.002	0.446^{***}	0.138^{***}
	(0.013)	(0.065)	(0.010)	(0.058)	(0.010)
Log diff. in citation impacts	0.028^{***}	-0.024	0.032^{***}	-0.023	0.018^{*}
	(0.008)	(0.020)	(0.007)	(0.018)	(0.007)
Research field overlap	0.768^{***}		0.554^{***}		5.195^{***}
	(0.113)		(0.099)		(0.090)
Sq. research field overlap	0.474^{***}		0.523^{***}		-3.007^{***}
	(0.107)		(0.093)		(0.086)
Pair fixed effects		Yes		Yes	
Year fixed effects		Yes		Yes	
Observations	121,581	121,581	170,773	67,661	293,457
Pairs	33,093	11,408	47,917	15,111	89,441
Years	6	6	6	6	6
Log-likelihood	-54,239.964	-28,170.943	-73,444.022	-36,911.377	-78,214.470

Notes: Four-year lag between dependent and independent variables. Standard errors in parentheses ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Intercept and zero-indicators suppressed. "NZ collaborators" and "Collaborators" correspond to panels described in Table 1. "Pooled pairs" corresponds to union of panels described in Table 1.

Appendix

Appendix Table 1: Publication, author, proposal, and applicant counts in our linked data

Year	Publications	Authors	Proposals	Applicants
2000	18,455	6,052	714	1,234
2001	19,812	6,389	844	1,435
2002	21,230	6,694	763	1,441
2003	22,554	7,030	714	1,569
2004	25,299	7,520	910	1,984
2005	27,472	7,888	858	1,843
2006	30,218	8,302	885	1,933
2007	31,940	8,639	864	1,900
2008	33,838	8,881	784	1,784
2009	35,447	9,111	887	1,959
2010	36,553	9,408	1,061	2,403
2011	38,807	9,665	1,053	2,403
2012	38,987	9,709	1,080	2,499
2013	39,984	9,824	1,114	2,638
2014	40,172	9,754	1,178	2,749
2015	39,499	9,719	1,166	2,767
2016	39,814	9,729	1,069	2,531
2017	39,062	9,652	1,070	2,585
2018	34,746	9,090	1,051	2,510
Total	613,889		18,065	

Notes: Researchers can be authors or applicants in multiple years.

Appendix Table 2: Distribution of linked publications across ASJC field groups

Field group	Publications (000)	Share (%)
Medicine	83.731	13.648
Agricultural and Biological Sciences	72.380	11.798
Biochemistry, Genetics and Molecular Biology	68.092	11.099
Earth and Planetary Sciences	57.758	9.414
Physics and Astronomy	40.628	6.622
Engineering	35.921	5.855
Computer Science	35.704	5.820
Environmental Science	31.684	5.164
Chemistry	26.128	4.259
Social Sciences	20.760	3.384
Mathematics	19.471	3.174
Materials Science	18.372	2.995
Immunology and Microbiology	14.931	2.434
Multidisciplinary	13.837	2.255
Neuroscience	12.442	2.028
Psychology	10.866	1.771
Pharmacology, Toxicology and Pharmaceutics	8.017	1.307
Chemical Engineering	7.735	1.261
Business, Management and Accounting	6.793	1.107
Arts and Humanities	6.631	1.081
Economics, Econometrics and Finance	4.621	0.753
Nursing	3.996	0.651
Energy	3.917	0.638
Veterinary	3.365	0.548
Health Professions	2.766	0.451
Decision Sciences	2.228	0.363
Dentistry	0.736	0.120
Total	613.509	100.000

Notes: Counts are fractional and exclude publications with no matched ASJC fields (of which there are 300 among the 613,889 publications in our linked data).