Social networks and research output*

Lorenzo Ductor[†] Marcel Fafchamps[‡] Sanjeev Goyal[§]
Marco J. van der Leij[¶]

February 2013

Abstract

We study how knowledge about the social network of an individual researcher – as embodied in his coauthor relations – helps us in developing a more accurate prediction of his future productivity. We find that incorporating information about coauthor networks leads to a modest improvement in the accuracy of forecasts on individual output, over and above what we can predict based on the knowledge of past individual output. Our second finding is that the signalling content of the network is quantitatively more important than the flow of ideas.

*We are grateful to Maria Dolores Collado, Markus Mobius and conference participants at SAEe (Vigo), Bristol, Stern (NYU), Microsoft Research, Cambridge, MIT, Alicante and Oxford for useful comments. Lorenzo gratefully acknowledges financial support from the Spanish Ministry of Education (Programa de Formacion del Profesorado Universitario).

[†]Massey University. E-mail: lductor@massey.ac.nz

[‡]University of Oxford. Email: marcel.fafchamps@economics.ox.ac.uk

[§]University of Cambridge. E-mail: sg472@econ.cam.ac.uk

[¶]E-mail: mvanderleij@gmail.com

1 Introduction

Good recruitment requires an accurate prediction of a candidate's potential future performance. Sports clubs, academic departments, and business firms routinely use past performance as a guide to predict the potential of applicants and to forecast their future performance. In this paper the focus is on researchers.

Social interaction is an important aspect of research activity: researchers discuss and comment on each other's work, they assess the work of others for publication and for prizes, and they join forces to coauthor publications. Scientific collaboration involves the exchange of opinions and ideas and facilitates the generation of new ideas. Access to new and original ideas may in turn help researchers be more productive. It follows that, other things being equal, individuals who are better connected and more 'central' in their professional network may be more productive in the future.

Network connectedness and centrality arise out of links created by individuals and so they reflect their individual characteristics – e.g., ability, sociability, and ambition. For instance, a collaboration with highly productive coauthors often reveal that these coauthors find such collaboration worthwhile. Since the ability of a researcher is imperfectly known, the existence of such ties may be informative.

The above considerations suggest that somenone's collaboration network is related to their research output in two ways: one, the network serves as a conduit of ideas and, two, the network signals their individual quality. The first channel suggests a causal relationship from network to research output, whereas the second does not. Determining causality would clarify the importance of the two channels. Unfortunately, as is known in the literature on social interactions (Manski, 1993; Moffit, 2001), identifying network effects in a causal sense is difficult in the absence of randomized experiments.

In this paper we take an alternative route: we focus on the *predictive power* of social networks in terms of future research output. That is, we investigate how much current and past information on collaboration networks contribute to forecasting future research output. "Causality" in the sense of prediction informativeness is known as Granger causality and is commonly analyzed in the macroeconometrics literature – see for example, Stock and Watson (1999) who investigate the predictive power of unemployment rate and other macroeconomics variables on forecasting inflation.[1] Finding that network variables Granger-cause future output does not constitute conclusive evidence of causal network effects in the traditional sense. Nonetheless, it implies that knowledge of a researcher's network can potentially be used by an academic department in making recruitment decisions.

We apply this methodology to evaluate the predictive power of collaboration networks on future research output, measured in terms of future economics publications. We first ask whether social network measures help predict future research output beyond the information contained in individual past performance. We then investigate which specific network variables are informative and how their informativeness varies over a researcher's career.

Our first set of findings are about the information value of networks. We find that including information about coauthor networks leads to an improvement in the accuracy of forecasts about individual output *over and above* what we can predict based on past individual output. The effect is significant but modest, e.g., the root mean squared error in predicting future productivity falls from 0.773 to 0.758 and the R^2 increases from 0.395 to 0.416. We also observe that several network variables – such as productivity of coauthors, closeness centrality, and the number of coauthors – have predictive power. Of those, the productivity of

[1] A few examples of applications that have determined the appropriateness of a model based on its ability to predict are Swanson and White (1997), Sullivan et al. (1999), Lettau and Ludvigson (2001), Rapach and Wohar (2002) and Hong and Lee (2003).

coauthors is the most informative network statistic among those we examine.

Secondly, the predictive power of network information varies over a researcher's career: it is more powerful for young researchers but declines systematically with career time. By contrast, information on recent past output remains a strong predictor of future output over an author's entire career. As a result, fourteen years after the onset of a researcher's publishing career, networks do not have any predictive value on future research output over and above what can be predicted using recent and past output alone.

Our third set of findings is about the relation between author ability and the predictive value of networks. We partition individual authors in terms of past productivity and examine the extent to which network variables predict their future productivity. We find that the predictive value of network variables is non-monotonic with respect to past productivity. Network variables do not predict the future productivity of individuals with below average initial productivity. They are somewhat informative for individuals in the highest past productivity tier group. But they are most informative about individuals in between. In fact, for these individuals, networks contain more information about their future productivity than recent research output. Taken together, these results predict that academics recruiters would benefit from gathering and analyzing information about the coauthor network of young researchers, especially for those who are relatively productive.

This paper is a contribution to the empirical study of social interactions. Traditionally, economists have studied the question of how social interactions affect behavior across well defined groups, paying special attention to the difficulty of empirically identifying social interaction effects. For an overview of this work, see for instance Moffitt (2001) and Glaeser and Scheinkman (2002). In recent years, interest has shifted to the ways by which the architecture of social networks influences behavior and outcomes.[2] Recent empirical papers on network effects include Bramoullé, Djebbari and Fortin (2009), Calvó-Armengol, Patacchini and Zenou (2008), Conley and Udry (2010), and Fafchamps, Goyal and van der Leij (2010).

This paper is also related to a more specialized literature on research productivity. Two recent papers, Azoulay et al. (2010) and Waldinger (2010), both use the 'unanticipated' removal of individuals as a natural experiment to measure network effects on researchers' productivity. Azoulay et al. (2010) study the effects of the unexpected death of 'superstar' life scientists. Their main finding is that coauthors of these superstars experience a 5% to 8% decline in their publication rate. Waldinger (2010) studies the dismissal of Jewish professors from Nazi Germany in 1933/34. His main finding is that a fall in the quality of a faculty has significant and long lasting effects on the outcomes of research students. Our paper quantifies the predictive power of network information over and above the information contained in past output.

The rest of the paper is organized as follows. Section 2 lays out the empirical framework. Section 3 describes the data and define the variables. Section 4 presents our findings. Section 5 checks the robustness of our main findings. Section 6 concludes.

2 Empirical framework

It is standard practice in most organizations to look at the past performance of job candidates as a guide to their future output. This is certainly true for the recruitment and promotion of researchers, possibly because research output – i.e., journal articles and books – is publicly observable.

[2]For a survey of the theoretical work on social networks see Goyal (2007), Jackson (2009) and Vega-Redondo (2007).

The practice of looking at past performance appears to rest on two ideas. The first is that a researcher's output largely depends on ability and effort. The second is that individuals are aware of the relationship between performance and reward and consequently exert effort consistent with their career goals and ambition. This potentially creates a stable relationship between ability and ambition on the one hand, and individual performance on the other hand. Given this relationship, it is possible to (imperfectly) predict future output on the basis of past output. In this paper we start by asking how well past performance predicts future output.

We then ask if future output can be better predicted if we include information about an individual's research network. Social interaction among researchers takes a variety of forms, some of which are more tangible than others. Our focus is on social interaction reflected in the coauthorship of a published paper. This is a concrete and quantifiable form of interaction. Coauthorship of academic articles in economics rarely involves more than 4 authors. So, it is likely that coauthorship entails personal interaction. Moreover, given the length of papers and the duration of the review process in economics, it is reasonable to suppose that collaboration entails communication over an extended period of time. These considerations – personal interaction and sustained communication – in turn suggest several ways by which someone's coauthorship network can reveal valuable information on their future productivity. We focus on two: research networks as a conduit of ideas; and coauthorship as a signal about unobserved ability and career objectives.

Consider first the role of research networks as a conduit for ideas. Communication in the course of research collaboration involves the exchange of ideas. So we expect that a researcher who is collaborating with highly creative and productive people has access to more new ideas. This, in turn, suggests that a researcher who is close to more productive researchers may have early access to new ideas. As early publication is a key element in the research process, early access to new ideas can lead to greater productivity. These considerations lead us to expect that, other things being equal, an individual who is in close proximity to highly productive authors will on average have greater future productivity.

Proximity need not be immediate, however: if A coauthors with B and B coauthors with C, then ideas may flow from A to C through their common collaborator B. The same argument can be extended to larger network neighborhoods. It follows that authors who are more central in the research network are expected to have earlier and better access to new research ideas.

As a first step we look at how the productivity of an individual, say i, varies with the productivity of his or her coauthors. We then examine whether i's future productivity depends on the past productivity of the coauthors of his or her coauthors. Finally we generalize this idea to i's centrality in the network – in terms of how close a researcher is to all other researchers (closeness) or how critical a researcher is to connections among other researchers (betweenness) – the idea being that centrality gives privileged access to ideas that can help a researcher's productivity.

Access to new ideas may open valuable opportunities but it takes ability and effort to turn a valuable idea into a publication in an academic journals. It is reasonable to suppose that the usefulness of new ideas varies with ability and effort. In particular, a more able researcher is probably better able to turn the ideas accessed through the network into publications than a less able researcher. Since ability and industriousness are reflected in past performance, we expect the value of a social network to vary with past performance. To investigate this possibility, we partition researchers into different tier groups based on their past performance and examine whether the predictive power of having productive coauthors and other related network variables varies systematically across tier groups.

The second way by which network information may help predict future output is because the quantity and quality of one's coauthors is correlated with – and thus can serve as a signal for – an individual's hidden ability and ambition. Given the commitment of time and effort involved in a research collaboration, it is reasonable to assume that researchers do not casually engage in a collaborative research venture. Hence when a highly productive researcher forms and maintains a collaboration with another – possibly more junior – researcher i, this link reveals positive attributes of i that could not be inferred from other observable data. Over time, however, evidence on i's performance accumulates, and residual uncertainty about i's ability and industriousness decreases. We therefore expect the signal value of network characteristics to be higher at the beginning of a researcher's career and to fall afterwards.

Our empirical strategy is based on the above ideas. Since our focus is on predictive power, we worry that overfitting may bias inference. To avoid this, we divide the sample into two halves, one of which is used to obtain parameter estimates, and the other to assess the out-of-sample predictive power of these estimates. We thus begin by randomly dividing the authors into two equal size groups. The first halve of the authors is used to estimate a regression model of researcher output. We then use the estimated coefficients obtained from the model fitted on the first halve of the authors to predict researcher output for the authors in the second halve of the data. We then compare these predictions with actual output.

The purpose of this procedure is to assess the out-of-sample prediction performance of the model. The reason for using out-of sample predictions is that in-sample errors are likely to understate forecasting errors. As stated by Fildes and Makridakis (1995) "the performance of a model on data outside that used in its construction remains the touchstone for its utility in all applications" regarding predictions. Other drawback of in-sample tests is that they tend to reject the null hypothesis of predictability. In other words, in-sample tests of predictability may spuriously indicate predictability when there is none.[3]

The rest of this section develops some terminology and presents the regressions more formally. We begin by describing the first step of our procedure and then we explain how we assess prediction performance. The dependent variable of interest is a measure y_{it} of the future output of author i, defined more in detail in the data section. This measure takes into account the number of articles published, the length of each of the articles, and the ranking of the journal where the article appears.

We first study predictions of y_{it} based on past output and a set of controls x_{it}. Control variables include: cumulative output since the start t_{i0} of i's career until $t-5$; career time dummies; year dummies; and the number of years since i's last publication. Career time dummies are included to capture career cycle effects, i.e., that researchers publish less as they approach retirement. We then examine by how much recent research output and network characteristics improve the prediction. We also compare the accuracy of the prediction when we use only past output and when we combine it with recent network characteristics.

The order of the regression models we estimate is as follows. We start with benchmark model 0 which examines the predictive power of the control variables x_{it}:

$$\text{Model 0} \qquad y_{i,t+1} = x_{it}\beta + \varepsilon_{it}$$

We then include past output $y_{i,t}$ as additional regressor. This yields Model 1:

[3]Arguments in favour of using out-of sample predictions can be found in Ashley et al. (1980) who state that "a sound and natural approach" to testing predictability "must rely primarily on the out-of-sample forecasting performance of models relating the original series of interest" (page 1149). Along with Fair and Shiller (1990), they also conjecture that out-of-sample inference is more robust to model selection biases and to overfitting or data mining.

$$\text{Model 1} \qquad y_{i,t+1} = x_{it}\beta + y_{it}\gamma_1 + \varepsilon_{it}$$

In Model 2 we investigate the predictive power of network variables $z_{i,t}$:

$$\text{Model 2} \qquad y_{i,t+1} = x_{it}\beta + z_{it}\gamma_2 + \varepsilon_{it}$$

Network variables include the number of i's coauthors up to time t, the productivity of these coauthors, and different network centrality measures detailed in the empirical section. We estimate Model 2 first with one network variable at a time, then including network variables simultaneously.

Finally, in Model 3 we ask if network variables z_{it} improve the prediction of future output over and above the prediction obtained from Model 1, that is, from past productivity:

$$\text{Model 3} \qquad y_{i,t+1} = x_{it}\beta + y_{it}\gamma_1 + z_{it}\gamma_2 + \varepsilon_{it}$$

Here too we first consider one network variable at a time to ascertain which network characteristic have more predictive power. We also estimate Model 3 with several networks variables together to evaluate the overall information contained in the network.

Models 0, 1 and 2 are nested in Model 3. A comparison of models 1 and 2 allows us to investigate the relative information content of recent individual output and recent social network. A comparison of models 2 and 3 examines whether social network variables have explanatory power over and above the information contained in recent individual output.

For Models 2 and 3 we consider both regressions with a single network variable, and regressions with multiple network variables. In the latter case, since our ultimate purpose is to predict research output, we need a criterion to select a parsimonious set of regressors so as to avoid overfitting. To select among social network regressors we use the Bayesian Information Criterion (BIC). We find that, in our case, the lowest values of the BIC criteria are obtained when all the network variables are included, which is why our final specification of the "multivariate" model includes them all.

The previous models are called restricted models because we are imposing the constraint that the lagged productivity variables from t_{i0} of i's career until $t-5$ have the same effect on future productivity. Moreover, in these models we only consider 5-year network variables, i.e. each network variable is computed assuming that a link between author i and her co-author has a predictive effect that lasts for five years. These restricted models are simple to estimate and allow us to compare the predictive power of network variables and recent output. But we may be able to improve the predictions of the restricted models by relaxing the constraint that productivity lags have the same coefficient. Similarly, the predictive power of the network variables might increase if we include several lags of the network variables.

To see whether this is the case, we also estimates versions of Models 1, 2 and 3 that include several lags of the productivity and network variables. The number of lags of the productivity and network variables are selected using the BIC criteria. We call these the unrestricted models. The benchmark unrestricted model 1, Model 1', contains thirteen lags of the productivity variable and a new set of control variables x_{it}: career dummies, time dummies and years since the last publication. This model examines the predictive power of past output:

$$\text{Model 1'} \qquad y_{i,t+1} = x_{it}\beta + \sum_{s=0}^{12} y_{it-s}\gamma_s + \varepsilon_{it}$$

We also consider a model with only network information, Model 2':

$$\text{Model 2'} \qquad y_{i,t+1} = x_{it}\beta + \sum_{s=1}^{T} z_{it-s}\theta_s + \varepsilon_{it}$$

where T is the maximum lag length of the network variable selected using the BIC criteria. For example, if $T = 15$ we include lags from z_{it-1} to z_{it-15} – z_{it-1} is the network variable computed using the joint publications at period t and z_{it-15} is the network variable obtained combining all joint publications from $t - 14$ to t. A comparison of Model 1' and 2' provides insights about the importance of past networks relative to past output.

The unrestricted Model 3, Model 3', combines all past output and past network information:

$$\text{Model 3'} \qquad y_{i,t+1} = x_{it}\beta + \sum_{s=0}^{12} y_{it-s}\gamma_s + \sum_{s=0}^{T} z_{it-s}\theta_s + \varepsilon_{it}$$

We also estimate models 2' and models 3' with multiple network variables. A comparison of Model 1' and Model 3' allow us to examine the explanatory power of network variables over and above knowledge of past output.

This describes the first step of our analysis. In the second step we evaluate the predictive accuracy of the different models. To this effect we compare, in the second halve of the data, the actual research output $y_{i,t+1}$ to the predictions $\widehat{y}_{i,t+1}$ obtained by applying to authors in the second halve of the data the regression coefficients of restricted models 0 to 3 and unrestricted models 1' to 3' obtained from the first halve of the data. To evaluate the prediction accuracy of $\widehat{y}_{i,t+1}$ we report the root mean squared errors (RMSE) defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i,t}(y_{i,t+1} - \widehat{y}_{i,t+1})^2}.$$

If the introduction of an explanatory variable in $\widehat{y}_{i,t+1}$ decreases the out-of-sample RMSE, this variable contains useful information that helps predict researchers future productivity.

In order to assess whether forecasts from two models are significantly different we use a test described by Diebold and Mariano (1995). This test is based on the loss differential of forecasting the future output of an individual i, $d_{i,t}$. As we measure the accuracy of each forecast by a squared error loss function (RMSE), we apply the Diebold-Mariano test to a squared loss differential, that is,

$$d_{i,t} = \varepsilon^2_{Ai,t} - \varepsilon^2_{Bi,t}.$$

where A is a competing model and B is the benchmark model.

To determine if one model predicts better we test the null hypothesis, H_0 : $E[d_{i,t}] = 0$, against the alternative, $H_1 : \; E[d_{i,t}] \neq 0$. Under the null hypothesis, the Diebold-Mariano test is

$$\frac{\overline{d}}{\sqrt{\hat{V}(\overline{d})/n}} \backsim N(0,1)$$

where $\overline{d} = n^{-1} \sum\limits_{i,t} d_{i,t}$, is the average loss differential, and $\hat{V}(\overline{d})$ is a consistent estimate of the asymptotic (long-rung) variance of $\sqrt{n}\overline{d}$. We adjust for serial correlation by using a Newey-West type estimator of $\hat{V}(\bar{d})$.[4]

3 Data

The data used for this paper are drawn from the EconLit database, a bibliography of journals in economics compiled by the editors of the *Journal of Economic Literature.* From this database we use information on all articles published between 1970 and 1999. These data are the same as those analyzed by Goyal, van der Leij,

[4]Formally, $\hat{V}(\bar{d}) = \sum_i (\hat{\gamma}_0 + 2 \sum\limits_{\tau=1}^{T-t} w_{m(T)} \hat{\gamma}_\tau)$, and $\hat{\gamma}_\tau = \hat{Cov}(d_{i,t}, d_{i,t-\tau})$, where $w_{m(T)}$ is the Bartlett Kernel function:

$$w_{m(T)} = \begin{cases} \left(1 - \frac{\tau}{m(T)}\right) & \text{if } \; 0 \leq \frac{\tau}{m(T)} \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

and $m(T)$ also known as the "truncation" lag is a number growing with T, the number of periods in the panel. The truncation lag has been chosen by the BIC.

and Moraga-González (2006), Fafchamps, Goyal and van der Leij (2010), van der Leij and Goyal (2011) and Ductor (2012).

3.1 Definition of variables

The output q_{it} of author i in year t is defined as:

$$q_{it} = \sum_{j \in S_{it}} \text{journal quality}_j \tag{1}$$

where S_{it} is the set of articles j of individual i published in year t. When available, the Journal quality variable is taken from the work of Kodrzycki and Yu (2006) – hereafter KY.[5] Unfortunately, KY do not include in their analysis all the journals in the EconLit database. To avoid losing information and minimize measurement error in research output, we construct a prediction of the KY quality index of journals not included in their list.[6] The actual KY journal quality index is used whenever available.

We are interested in predicting future output. In economics, the annual number of papers per author is small and affected by erratic publication lags. We therefore need a reasonable time window over which to aggregate output. The results presented here are based on a three year window, but our findings are insensitive

[5]We do not consider citations because they are often materialize long after a paper has been published. This means that authors at the beginning of their career often have a small citation record and hence, for them at least, citations have little predictive power.

[6]To do this, we regress the KY index on commonly available information of each journal listed in EconLit, such as the number of published articles per year, the impact factor, the immediacy index, the Tinbergen Institute Index, an economics dummy, interaction terms between the economics dummy and the impact factor, and various citation measures. Estimated coefficients from this regression are then used to obtain a predicted KY journal quality index for journals not in their list. Since most of the journals that KY omitted are not highly ranked, their predicted quality index is quite small.

to the use of alternative window length, e.g., five years. Our dependent variable of interest is thus the output of author i in years $t+1, t+2, t+3$:

$$q_{it}^f = q_{i,t+1} + q_{i,t+2} + q_{i,t+3} \tag{2}$$

Unsurprisingly q_i^f has a long upper tail. To avoid our results from being entirely driven by a handful of highly productive individuals, we log the dependent variable as follows:[7]

$$y_{i,t+1} = \ln\left(1 + q_{it}^f\right)$$

The analysis presented in the rest of the paper uses $y_{i,t+1}$ as dependent variable.

We expect recent productivity to better predict output over the next three years than ancient output. To capture this idea, we divide past output into two parts in the restricted models: cumulative output until period $t-5$, which captures i's historical production and is used as control variable; and output from $t-4$ until t, which represents i's recent productivity and is expected to be a strong predictor of future output. We define recent output q_{it}^r from t to $t-4$ as:

$$q_{it}^r = q_{it} + q_{i,t-1} + q_{i,t-2} + q_{i,t-3} + q_{i,t-4}$$

Control variables in the restricted model x_{it} include cumulative output q_{it}^c from the start t_{i0} of i's career until $t-5$:

$$q_{it}^c = q_{i,t_{i0}} + ...q_{i,t-6} + q_{i,t-5}$$

where t_{i0} is the year in which individual i obtained his or her first publication. We use $\ln(1+q_{i,t}^c)$ and $\ln(1+q_{i,t}^r)$ as regressors, since the distribution of both variables

[7]We have considered other functional forms such as y^{15},y^{50} ,y^{75}, log of log, Poisson, Non-negative Binomial, Zero inflated Non-Negative Binomial and Tobit. In terms of out-of-sample RMSE, the specification that provides the best forecast is $\ln(x+1)$, which is the one we report here. See the online appendix for more details.

presents fat tails. We also include the number of years r_{it} with no published article since i's last article was published:

$$r_{it} = \begin{cases} 0 & \text{if } q_{it} > 0 \\ r_{i,t-1} + 1 & \text{otherwise.} \end{cases} \quad \text{and}$$
$$r_{i,t_{i0}} = 0$$

Variable r_{it} is used as proxy for leave or retirement from academics: the longer someone has not published, the more likely he or she has retired or left research. Other controls include career time dummies c_{it}, and year dummies t. To summarize, $x_{it} = \{q^c_{it}, r_{it}, c_{it}, t\}$.

In the unrestricted models 1' and 3', we relax the constraint imposed in q^r_{it} and q^c_{it}. In these models, we consider thirteen lags of the productivity variable, where lags are obtained as:

$$y_{i,t-s} = \ln(1 + q_{i,t-s} + q_{i,t-s-1} + q_{i,t-s-2}) \qquad \forall s = 0, ..., 12.$$

Control variables in the unrestricted models are the same as in the restricted models but excluding past output.

Next we turn to the network variables. Given that we wish to investigate whether network characteristics have predictive power over and above that of recent productivity, network variables must be constructed in such a way that they do not contain information outside the time window of q^r_{it}. We therefore define the 5-year co-authorship network $G_{t,5}$ at time t over the same time window as q^r_{it} for the restricted models, that is, using all joint publications from year $t-4$ to t. At time t, two authors i and j are said have a link $g_{ij,t}$ in $G_{t,5}$ if they have published in an EconLit journal in years $t-4$ to t. Otherwise $g_{ij,t} = 0$.

For the unrestricted models 2' and 3' we introduce different co-authorship networks, $G_{t,s}$, where s determines the number of years that a link between author i

and her co-author j lasts. For example, in the network $G_{t,10}$, we assume that the effects from a collaboration last during 10 years, from $t-9$ to t.

The set of network statistics that we construct from $G_{t,s}$ is motivated by the theoretical discussion of Section 2. Some of the network statistics we include in our analysis are, on a priori grounds, more correlated with access to new scientific ideas; others are included because they are thought to have a high signalling potential. Measures of network topology such as centrality and degree reflect network proximity and thus belong primarily to the first category while. Other measures, such as the productivity of coauthors, are likely to have greater signalling potential.

Based on these observations, the list of network variables that we use in the analysis is as follows. We say that there is a path between i and j in $G_{t,s}$ if $g_{ij,t}=1$ at some period from $t-s$ to t or there exists a set of distinct nodes $j_1,\ldots,j_m$, such that $g_{ij_1,t}=g_{j_1j_2,t}=\ldots=g_{j_mj,t}=1$. The length of such a path is $m+1$. The distance $d(i,j;G_{t,s})$ is the length of the shortest path between i and j in $G_{t,s}$. We use the following standard definitions:

- *(First order) degree* is the number of coauthors that i has in period $t-s$ to t, $n_{1i,t}=|N_i(G_{t,s})|$, where $N_i(G_{t,s})=\{j:g_{ij,t}=1\}$.

- *Second order degree* is the number of nodes at distance 2 from i in period $t-s$ to t, $n_{2i,t}=|N_i^2(G_{t,s})|$, where $N_i^2(G_{t,s})=\{k:d(i,k;G_{t,s})=2\}$.

- *Giant component*: The giant component in $G_{t,s}$ is the largest subset of nodes such that there exist a path between each pair of nodes in the giant component, and no path to a node outside. We create a dummy variable which takes value 1 if an author belongs to the giant component and 0 otherwise.

Within the giant component we consider the following two global proximity measures.[8]

- *Closeness centrality* $C^c_{i,t}$ is the inverse of the average distance of a node to other nodes within the giant component and is defined as:

$$C^c_{i,t} = \frac{n_t - 1}{\sum_{j \neq i} d(i, j; G_{t,s})}$$

where n_t is the size of the giant component in year t in the co-authorship network $G_{t,s}$. Because $C^c_{i,t}$ has fat tails, we use $\ln(1 + C^c_{i,t})$ as regressor instead.

- *Betweenness centrality* $C^b_{i,t}$ is the frequency of shortest paths passing through node i and is calculated as:

$$C^b_{i,t} = \sum_{j \neq k : j,k \neq i} \frac{\tau^i_{j,k}(G_{t,s})}{\tau_{j,k}(G_{t,s})}$$

where $\tau^i_{j,k}(G_t)$ is the number of shortest paths between j and k in $G_{t,s}$ that pass through node i, and $\tau_{j,k}(G_{t,s})$ is the total number of shortest paths between j and k in $G_{t,s}$. In the regression analysis, we similarly use $\ln(1+C^b_{i,t})$ as regressor.

Next, we define regressors that capture the productivity of coauthors and that of coauthors of coauthors. We apply the $\ln(x + 1)$ transformation to them as well.

- *Productivity of coauthors*: is defined as the output of coauthors' of author i from $t - s$ to t

$$q^1_{it} = \sum_{j \in N_i(G_{t,s})} q^r_{jt}$$

where q^r_{jt} is the output of j from period $t - s$ to period t (excluding papers that are coauthored with i).

[8]For a careful discussion on the interpretation of centrality measures, see Wasserman and Faust (1994).

- *Productivity of coauthors of coauthors*: the output of coauthor of coauthors' of author i from $t-s$ to t,

$$q_{it}^2 = \sum_{k \in N_i^2(G_{t,s})} q_{kt}^r$$

  where q_{kt}^r is the output of k from $t-s$ to t excluding papers that are coauthored with the neighbors of i, $N_i(G_{t,s})$.

We also include a dummy variable that takes value 1 for author i if one of i's coauthors in $G_{t,s}$ has an output q_{jt}^r in the top 1% of the distribution of q_{it}^r.

In the restricted models, only five-year network variables are included. In contrast, in the unrestricted models the number of network periods is selected according to the BIC criteria.

3.2 Descriptive statistics

Table 1 provides summary statistics of the variables included in the analysis. Column 1 provides the mean value of each variable. Column 2 shows the standard deviation and column 3 provides correlations between the different variables and future productivity.

For the restricted model, we excluded observations relative to authors in the earliest stage of their career, i.e., for which $c_{it} < 6$. The reason is that these authors have not yet established a publication record and network so that there is little information on which to form predictions of future output. This assumption is relaxed in the unrestricted models, where we consider the full sample, 1,335,428 observations, after replacing the missing lagged productivity and network variables by zeros. The rationale for doing so is that authors who have just started their career have no past output and co-authorship, hence the value of their lagged productivity and network variables are truly zero.

We draw attention to some distinctive features of the data. First, we observe that the variance in future output q_{it}^f is large, with a standard deviation 2.41 times larger than the mean. There is a high positive correlation of 0.69 between recent output q_{jt}^r and future output q_{it}^f. Figure 1 shows a scatter plot and a linear regression line with confidence interval between q_{it}^f and q_{jt}^r for 1000 random selected observations. This visually confirms that, as anticipated, recent past output has a strong predictive power on future output.

Second, we observe a high correlation between q_{it}^f and several 5-year network variables such as coauthors' output q_{it}^1, author degree, and closeness and betweenness centrality. The network variable most highly correlated with future productivity is q_{it}^1, the productivity of i's coauthors with a correlation coefficient of 0.58. Other network variables such as degree, closeness, and betweenness centrality are also highly correlated with future output q_{it}^f. Figure 2 shows the relationship between some 5-year network variables and future output.

4 Empirical findings

We have seen that there is a reasonably strong correlation between future output and recent past output, but also between future output and the characteristics of i's recent coauthorship network. We now turn to a multivariate analysis and estimate the different models outlined in Section 2. We start by presenting the results on the predictive power of network information. This leads to a closer examination of the role of signalling and flow of ideas in networks. We then examine the relation between the productivity of an individual author and the predictive power of network variables.

4.1 Predicting future output

Table 2 presents the prediction results for Model 0, the baseline model with controls $x_{it} = \{q^c_{it}, r_{it}, c_{it}, t\}$, Model 1, that includes recent output q^r_{it}, and Model 2 that includes a network variable, one per regression. Column 1 presents the R^2 of the regression on the in-sample data for each model. Column 2 shows the out-of-sample RMSE for each model. Column 3 compares the RMSE of Model 1/Model 2 with the benchmark model, Model 0. Column 4 shows the coefficient of each regressor.

Recent output q^r_{jt} explains slightly less than half of the variation in future output q^f_{it}. Half of the variation in q^f_{it} – around 51% of the total variation – remains unexplained after we take q^r_{it} into account. The question is: can we improve upon this using network variables?

We begin by examining the predictive power of the different network variables when one network variable is added to controls x_{it}. This is achieved by comparing the results from the Model 2 regressions with Model 0. Results, presented in Table 2, show that coauthor productivity q^1_{it}, closeness centrality $C^c_{i,t}$, and the productivity q^2_{it} of coauthors of coauthors – are statistically significant and help predict future output. However, the predictive power is much less than recent output, for example, coauthors' productivity reduces the RMSE by 9.38% whereas recent output reduces the RMSE by 15.72%.

We then combine recent output q^r_{it} and network variables in Model 3. Results presented in Table 3 show that the same network variables remain significant once we include q^r_{it} as regressor. Being significant does not imply that network variables are very informative, however. For this we have to examine the improvement in prediction that they represent. We compare Multivariate Model 3, that is, with multiple network variables in the regression, to Model 1. Table 4 shows that the R^2 of Model 3 is greater than the R^2 obtained under Model 1. This means that network information taken in combination with recent output yields a more accurate prediction than a prediction based on past output alone. The gain in explanatory power is small, however: the R^2 rises from 0.49 in Model 1 to 0.51 in Model 3. In line with this, the RMSE declines from 0.67 down to 0.65 when we incorporate network information. This small difference is statistically significant, as shown by the Diebold-Mariano test.

Table 5 presents the prediction results for the benchmark unrestricted Model 1' and Model 2'. Model 1' contains thirteen lags of the productivity variable and the same control variables as in the restricted models except past output. Model 2' includes the control variables without past output and several lags of a network variable. Column 1 presents the lag length of each variable, the rest of Columns are analogous to Table 2. The predictions obtained from the unrestricted models are consistent with their restricted versions. The network variable with the highest predictive power is coauthors' productivity with a RMSE 7.76% greater than the past output model, Model 1'. As shown in table 7, the predictive power of network over and above information of past output is slightly higher when we consider the unrestricted version, that is, when we include several lags of the network variables. In the restricted multivariate models, the RMSE is reduced by 1.65% when we add network variables to past and recent output, while in the unrestricted version, the reduction is around 1.94%.

From this we conclude that network variables contain predictive information over and above what can be predicted on the basis of past output, but this information gain is modest.

4.2 Networks and career cycle

Next we estimate the predictive power of network variables for different career time c_{it}. The RMSE of restricted Models 0, 1 and Multivariate Models 2 and 3 (that is, with multiple network variables included in the regression) as well as the RMSE of unrestricted models 1' and Multivariate Models 2' and 3' are plotted in Figure 3 and Figure 3, respectively. Career age c_{it} is on the horizontal axis while RMSE is measured on the vertical axis. Unsurprisingly, the Figures shows that the predictive accuracy of all the models improves – reflected in the decline in RMSE – with career time. This is primarily because the control variables x_{it} – particularly cumulative output q^c_{it} – reveal more information about individual ability and preferences over time.

To examine whether the relative predictive gain of network variables varies with career time, we report in Figure 4 and Figure 6 the difference in RMSE between Multivariate Models 2 and 3 versus Model 1 and the difference in RMSE between their unrestricted versions, respectively. We note a marked decline in the difference between Models 1' and 3' over the course of a researcher's career. After time $t =$ 14, the prediction accuracy of models with or without network variables becomes virtually indistinguishable. The Diebold-Mariano test shows that the differences between Multivariate Model 3' and Model 1' are not statistically significant from $t = 14$ to $t = 20$. In the restricted models, Figure 4, the decline in the predictive power of network variables is not observed till $t = 15$. This indicates that, for senior researchers, network variables contains little information over and above the information contained in past and recent output.

What does this pattern in the data suggest about the relative importance of the two potential ways in which networks may matter: flow of ideas and signalling? As time passes, the publication record of a researcher builds up. Since ability, research ambition, and other personality traits are relatively stable over time, this accumulating evidence ought to provide a more accurate estimate of the 'type' of the person. Hence it should become easier to judge his or her ability and research ambition on the basis of the publication record alone. Based on this, we would expect that the signalling value of networks decreases over time, and hence that network variables have less and less additional predictive power.

Research networks can, however, be important conduits of valuable research ideas as well. Unlike the signalling value of networks, access to new research ideas remains important throughout a researcher's career. Thus if network variables help predict future output because they capture access to new ideas, their predictive value should remain relatively unchanged over a researcher's career. This is not what we observe, leaving signalling as a stronger contender as the possible channel by which network variables help predict future productivity.

4.3 Network information across productivity categories

In this section we examine whether the predictive power of network information varies systematically with recent output q_{it}^r. This analysis is predicated on the idea that it takes talent and dedication to transform the new ideas conveyed by the research network into publishable output. Consequently we expect the predictive power of network variables to increase with ability – and hence with q_{it}^r – at least over a certain range.

To investigate this possibility, we divided the observations into five tier groups on the basis of their recent output q_{it}^r. The top category includes authors in the top 1% in terms of q_{it}^r. The second top category includes authors in the 95-99 percentiles of q_{it}^r. The third category covers authors in the 90-94 percentiles, the fourth includes authors in the 80-89 percentiles, and the last category is for authors in the 50-79 percentiles.[9]

Figure 7 shows the RMSE % difference between Model 1 and Model 2 versus Model 0 across the different categories. For the most productive authors, those above the 99th percentile, network variables have predictive power in explaining

[9]We do not consider authors below the median because the median recent output is zero.

future research output but much less than recent output. For the next category of researchers, those in the 95-98 percentile range, network information has greater predictive power. Even more strikingly, for researchers in the third category, the 90-94 percentile range, network variables are better at predicting future research output than q^r_{it}! All the Models have statistically significant predictive power across the different tiers group.

By contrast, network information has little but significant predictive power for low productive individuals (those in the lower half of the distribution). This suggests that, for researchers with low ability or research ambition, having published with high quality coauthors has little informative content regarding their future output – perhaps because they are unable to take advantage of the access to information and research ideas that good coauthors provide.

Figure 8 presents the RMSE of unrestricted Model 2' versus Model 1'. The findings are consistent with the restricted version. For the most productive authors, the predictive power of network variable is small relative to past output. For the low productive authors, those in the 50-79 percentile range, the explanatory power of network variables is very close to past output but none of them are very useful to predict the future output of an individual. The RMSE of Model 1' is almost the same as the RMSE of Model 1' excluding all the lags productivity variables, suggesting that neither past output nor network variables help to predict the productivity of low productive authors.

5 Robustness

We have conducted an extensive investigation into the robustness of our results to various assumptions made in constructing the variables used in the estimation. The results of this analysis are summarized here; the details, not shown here to save space, are available in an online appendix.

In the analysis so far we have used average productivity from $t+1$ to $t+3$ as the variable q_{it} we seek to predict (see equation 2). The rationale for doing so is that the distant future is presumably harder to predict than the immediate future, and we want to give the model a fair chance. Yet, in economics there are long lags between the submission and publication of a paper, and wide variation in these lags across papers and journals. Publication lags thus introduce additional variation in the variable we are trying to predict, and may thus lead us to underestimate the predictive power of network information. To check whether this is affecting our results, we repeat the analysis using average future productivity over a five year window instead of three:

$$q_{it}^{f} = q_{i,t+1} + q_{i,t+2} + q_{i,t+3} + q_{i,t+4} + q_{i,t+5}.$$

and, as before, use $\ln\left(1 + q_{it}^{f}\right)$ as the variable we seek to predict. Results are similar to those reported here except that the predictive power of network variables is larger using a five year window. In particular, network variables are even more useful than past output to forecast the future performance of a researcher, i.e., Unrestricted Multivariate Model 2 outperforms Unrestricted Model 1.

Next we investigate whether results are sensitive to our definition of output q_{it}. We examine whether different results obtain if we correct for article length and number of co-authors. Results show that the predictive power of network variables is unaffected.[10]

Finally, the main specification used so far is a linear model estimated by OLS in which the dependent variable is a logarithmic transformation of future research output, $\ln(q_{it}^{f} + 1)$. We are concerned that the model might be misspecified by restricting ourselves to OLS applied to this particular functional form. We, there-

[10]See online appendix for more details.

fore, repeat the analysis with nonlinear regression models frequently used to study research output or citations, such as the Poisson model, the Negative Binomial model, and the Zero-inflated Negative Binomial model. Results show that the in-sample log-likelihood is higher for the (Zero-inflated) Negative Binomial model than for the linear model applied to the $\ln(y+1)$-transformation. However the out of sample RMSE is lowest for the linear model. As the linear model is also easy to interpret and to evaluate, we use it as our main specification.

We also consider panel data models. Fixed effect models are not useful to predict the productivity of junior researchers so we do not pursue them further.[11] We also investigate the predictive power of vector autoregressive (VARs) models where past network variables affect future output and past output influence future network variables. We estimate such VAR models using a seemingly unrelated regressions (SUR) approach, allowing for correlation in the error terms across the two equations. The lag length of each equation is selected using the BIC criteria. The SUR regressions should in principle lead to more efficient predictions as long as the two equations do not include the same set of lagged variables, a conditions that is fulfilled here. Results show that the predictions generated by the unrestricted SUR Model 3 using feasible generalized least squares (FGLS) hardly differ from the unrestricted Model 3 estimated using simple OLS. Therefore, the SUR model does not outperform, out-of-sample, the simple OLS.

6 Concluding remarks

In this paper we have examined whether information about a researcher's coauthor network reveals information that helps predict their future output. Underlying our study are two main ideas. The first idea is that a collaboration resulting

[11] Results from panel data regressions are available in the online appendix.

in a published article reveals valuable information about an author's ability and research ambitions. This is particularly true for junior researchers whose type cannot be fully assessed from their cumulative output. The second idea is that professional research networks provide access to new research ideas. These ideas can subsequently be turned into published papers provided the researcher possesses the necessary ability and dedication.

To investigate these ideas, we examine coauthorship in economics. Our focus is not on statistical significance or causality but rather on predictive power. For this reason, we adopt a methodology that eliminates data mining and minimizes the risk of pre-testing bias. To this effect we randomly divide the data into two halves. Parameter estimates are obtained with one halve and predictions are judged by how well they perform in the other half of the sample.

We find that information about someone's coauthor networks leads to a modest improvement in the forecast accuracy of their future output *over and above* what can be predicted from their past output. The network variables that have the most information content are the productivity of coauthors, closeness centrality, and the number of past coauthors. These results are robust to alternative specifications and variable definitions.

We investigate whether the predictive power of network variables is stronger for more talented researchers, as would be the case if taking advantage of new ideas requires talent and dedication. We find that the predictive value of network information is non-existent for less talented and dedicated researchers. We also find that for the most able researchers – those in the top 1% of the output distribution – network variables have no information content, possibly because these researchers are so talented and dedicated that they would succeed irrespective of their collaboration history.

The work presented here leaves many questions unanswered. In particular, we do not claim to have identified a causal effect of coauthorship or network quality on future output. If anything, the signalling hypothesis is based on a reverse causality argument, and it receives the most support from our analysis. We do, however, also find evidence that network connections are most useful to talented researchers. This result is consistent with a causal relationship between the flow of research ideas and future output, with the caveat that talent is needed to turn ideas into publishable papers. These issues deserve more research.

References

[1] Akerlof, G. (1970), The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics*, 84, 3, 488-500.

[2] Arellano, B. and S. Bond (1991), Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, *Review of Economic Studies*, 58, 2, 277-297.

[3] Ashley, R., C.W.J. Granger, and R. Schmalensee, (1980), Advertising and Aggregate Consumption: An Analysis of Causality, *Econometrica*, 48, 1149-1167.

[4] Azoulay, P., J. Graff Zivin, J. Wang (2010), Superstar Extinction, *Quarterly Journal of Economics*, 125, 2, 549–589.

[5] Bertrand, M., E.F.P. Luttmer, and S. Mullainathan (2000), Network effects and welfare cultures, *Quarterly Journal of Economics*, 115, 3, 1019-1055.

[6] Bramoullé, Y., H. Djebbari, and B. Fortin (2009), Identification of peer effects through social networks, *Journal of Econometrics*, 5, 1, 41–55.

[7] Brock, W. and S. Durlauf (2001), *Interaction-based Models*, in J. Heckman and E. Leamer (eds), *Handbook of Econometrics*, Volume 5, Amsterdam: North Holland.

[8] Calvó-Armengol, A., E. Patacchini and Y. Zenou (2009), Peer effects and social networks in education, *Review of Economic Studies*, 76, 1239- 1267.

[9] Conley, T. and C. Udry (2010), Learning about a new technology: Pineapple in Ghana, *The American Economic Review*, 100, 1, 35–69.

[10] Ductor, L. (2012). Does Co-authorship Lead to Higher Academic Productivity? Mimeo.

[11] Duflo, E. and E. Saez (2003), The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment, *Quarterly Journal of Economics*, 118, 3, 815–842.

[12] Inoue, A., Kilian, L., (2004), In-sample or out-of-sample tests of predictability? which one should we use? *Econometric Reviews* 23, 371–402.

[13] Fafchamps, M. (2004), Market institutions in Sub-Saharan Africa. *MIT Press*, Cambridge, Mass.

[14] Fafchamps, M., S. Goyal, and M. J. van der Leij (2010), Matching and Network Effects, *Journal of European Economic Association*, 8, 1, 203-231.

[15] Fair, R.C., Shiller, R., (1990), Comparing information in forecasts from econometric models. *American Economic Review* 80, 375–389.

[16] Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review* 63, 289–308.

[17] Glaeser, E. J. Scheinkman (2002), Non-market interactions. *Econometric Congress: advances in economic theory and econometrics.* edited by M. Dewatripont, L.P Hansen and S. Turnovsky. Cambridge University Press.

[18] Goyal, S., M.J. van der Leij and J.L. Moraga (2006), Economics: An emerging small world, *Journal of Political Economy*, 114, 2, 403-412.

[19] Goyal, S., M.J. van der Leij (2011), Strong Ties in a Small World. *Review of Network Economics*, 10, 2, 1446-9022.

[20] Goyal, S. (2007), *Connections: an introduction to the economics of networks.* Princeton University Press, Princeton, New Jersey.

[21] Granger, C.W.J., (1980), Testing for Causality: A Personal Viewpoint, *Journal of Economic Dynamics and Control*, 2, 329-352.

[22] Granovetter, M. (1985), Economic Action and Social Structure: The Problem of Embeddedness, *American Journal of Sociology,* 91, 3, 481-510.

[23] Harris, R.D.F. and E. Tzavalis (1999), Inference for unit roots in dynamic panels where the time dimension is fixed, *Journal of Econometrics*, 91, 201–226.

[24] Hong, Y., and T. Lee (2003), Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* 85, 1048–1062.

[25] Kennedy, P. (2008). A guide to econometrics. 6th Edition. *Blackwell Publishing.*

[26] Kalaitzidakis, P., T. Mamuneas, and T. Stengos (2003), Rankings of academic journals and institutions in economics, *Journal of European Economic Association*, 1, 6, 1346-1366.

[27] Kuha, J. (2004), AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, 33, 188-229.

[28] Lettau, M., Ludvigson, S., (2001), Consumption, aggregate wealth, and expected stock returns. *Journal of Finance* 3, 815–849.

[29] Manski, C. (1993), The Identification of endogenous social effects: the reflection problem, *Review of Economic Studies*, 60, 3, 531-542.

[30] Mayer, A. and S.L. Puller (2008), The old boy (and girl) network: Social network formation on university campuses, *Journal of Public Economics*, 92, 1-2, 329-347.

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

[32] Moffitt R. (2001), Policy interventions, low-level equilibria, and social interactions, in: S. Durlauf and P. Young (eds.), *Social Dynamics*, Cambridge: MIT Press.

[33] Montgomery, J. (1991), Social networks and labor-market outcomes: toward an economic analysis, *American Economic Review*, 81, 5, 1408-1418.

[34] Munshi, K. (2003), Networks in the modern economy: Mexican migrants in the U.S. labor market, *Quarterly Journal of Economics*, 118, 2, 549-599.

[35] Pesaran, M.H., Timmermann, A., (1995), Predictabilityof stock returns: robustness and economic significance. *Journal of Finance* 50, 1201–1228.

[36] Potscher, B.M., (1991), Effects of model selection on inference, *Econometric Theory* 7, 163-185.

[37] Rapach, D.E., Wohar, M.E., (2002), Testing the monetary model of exchange rate determination: new evidence from a century of data. *Journal of International Economics* 58, 359–385.

[38] Sauer, Raymond D., (1988). Estimates of the Returns to Quality and Coauthorship in Economic Academia, *Journal of Political Economy*, 96(4), pp. 855-66.

[39] Sullivan, R., Timmermann, A., White, H., (1999). Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance* 54, 1647–1691.

[40] Swanson, N.R., White, H., (1997). A model-selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *The Review of Economics and Statistics* 79, 265–275.

[41] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16, 437– 450.

[42] Topa, G. (2001), Social Interactions, Local Spillovers, and Unemployment, *Review of Economic Studies*, 68(2): 261-95

[43] Waldinger, F. (2010), Quality Matters: The expulsion of students and the consequences for Phd student outcomes, *Journal of Political Economy*, vol. 118, no. 4, pp. 787-831.

[44] Wasserman, S. and K. Faust (1994), Social Network Analysis. *Cambridge University Press.*

Table 1: Summary statistics

	Mean	Std. deviation	Correlations
Output			
Future productivity	.41	.99	1
Past stock output	1.62	1.44	.44
Recent past output	.65	1.23	.69
Network variables			
Degree	.55	1.15	.55
Degree of order two	.80	2.81	.46
Giant component	.10	.30	.47
Closeness centrality	.01	.02	.48
Betweenness centrality	.50	2.29	.48
Coauthors' productivity	.58	1.40	.58
Coauthors of Coauthors' prod.	.55	1.55	.54
Working with top 1%	.01	.11	.34
Number of observations	1697415	1697415	1697415
Number of authors	75109	75109	75109

Network variables are computed assuming that a link between two authors lasts during five years (5-year network variables). The number of observations used to obtain the statics for recent past output is 1230335 and for past stock output is 1132248. All the correlations coefficients are obtained using the same number of observations, 872344

Table 2: Prediction Accuracy: Restricted Models 1 and 2

	R^2	RMSE	RMSE Diff.	Coefficients
Model 0				
Past output	.28	.789	-	.22**
Model 1				
Recent past output	.49	.665	15.72%	.49**
Model 2				
Degree	.38	.728	7.73%**	.29**
Degree of order 2	.36	.744	5.70%**	.10**
Giant component	.35	.748	5.20%**	1.05**
Closeness	.36	.743	5.83%**	22.96**
Betweenness	.38	.734	6.97%**	.11**
Coauthors' productivity	.41	.715	9.38%**	.30**
Coauthors of Coauthors' prod.	.39	.727	7.86%**	.24**
Working with a top 1%	.36	.746	5.45%**	1.75**

** Significant at 1% level, * Significant at 5%. Model 0 includes career time dummies, year dummies, number of years since the last publication and cumulative productivity from the first publication till the $t-5$. Model 1 adds to Model 0 recent output. Model 2 adds to Model 0 one of the network variable. Each network variable is computed assuming that a link from a collaboration last during 5 years (5-year network variable). The number of in-sample observations is 436440.

Table 3: Prediction Accuracy: Restricted Models 1 and 3

	R^2	RMSE	RMSE Diff.	Coefficients
Model 0				
Past output	.28	.789	-	.22**
Model 1				
Recent past output	.49	.665	15.72%	.49**
Model 3				
Degree	.50	.660	16.35%**	.09**
Degree of order 2	.50	.660	16.35%**	.03**
Giant component	.50	.662	16.10%**	.27**
Closeness	.50	.660	16.35%**	13.89**
Betweenness	.50	.657	16.73%**	.06**
Coauthors' productivity	.50	.660	16.35%**	.09**
Coauthors of Coauthors' prod.	.39	.660	16.35%**	.07**
Working with a top 1%	.50	.660	16.35%**	.59**

** Significant at 1% level, * Significant at 5%. Model 0 includes career time dummies, year dummies, number of years since the last publication and cumulative productivity from the first publication till $t-5$. Model 1 adds to Model 0 recent output. Model 3 adds to Model 1 one of the network variable. Each network variable is computed assuming that the effects from a collaboration last during 5 years (5-year network variable). The number of in-sample observations is 436440.

Table 4: Prediction accuracy of the restricted multivariate models

	R^2	RMSE	RMSE Diff.
Model 0	.278	.789	-
Model 1	.493	.665	15.72%**
Multivariate Model 2	.433	.700	11.28%**
Multivariate Model 3	.509	.654	17.11%**

** Significant at 1% level. These restricted models only includes 5 year network variables. The number of in-sample observations is 436440.

Table 5: Prediction Accuracy: Unrestricted Models 1' and 2'

	Lag Length	R^2	RMSE	RMSE Diff.	Coefficients
Model 1'					
Recent past output	13	.39	.773	-	.44**
Model 2'					
Degree	15	.24	.861	-11.38%**	. .10**
Degree of order 2	14	.23	.867	-12.16%**	.05**
Giant component	15	.23	.868	-12.29%**	.96**
Closeness	15	.24	.862	-11.51%**	1.42
Betweenness	15	.26	.849	-9.83%**	.07*
Coauthors' productivity	12	.29	.833	-7.76%**	.11**
Coauthors of Coauthors' prod.	15	.27	.847	-9.57%**	.09**
Working with a top 1%	14	.24	.862	-11.51%**	.45**

** Significant at 1% level, * Significant at 5%. Model 1' includes career time dummies, year dummies, number of years since the last publication and thirteen lags of the productivity variable. Model 2' contains career time dummies, year dummies, number of years since the last publication and several lags of a network variable. The RMSE Diff. is the percentage RMSE difference between Model 1' and Model 2'. The maximum lag length is selected using the BIC criteria. For the network variables, the maximum possible lag length considered is 15. The coefficient presented in the table correspond to the first lag of the variable. The number of in-sample observations is 667423.

Table 6: Prediction Accuracy of Unrestricted Model 1' and 3'.

	Lag Length	R^2	RMSE	RMSE Diff.	Coefficients
Model 1'					
Past output	13	.39	.773	-	.44**
Model 3'					
Degree	6	.41	.768	.65%**	.14**
Degree of order 2	5	.40	.768	.65%**	.06**
Giant component	8	.40	.768	.65%**	.58**
Closeness	10	.40	.767	.78%**	2.35*
Betweenness	9	.40	.767	.78%**	.02
Coauthors' productivity	12	.41	.761	1.55%**	.09**
Coauthors of Coauthors' prod.	11	.41	.764	1.16%**	.07**
Working with a top 1%	13	.40	.767	.78%**	.39**

** Significant at 1% level, * Significant at 5%. Model 1' includes career time dummies, year dummies, number of years since the last publication and thirteen lags of the productivity variable. Model 3' adds to Model 1' several lags of a network variable. The RMSE Diff. is the percentage RMSE difference between Model 1 and Model 3. The maximum lag length is selected using the BIC criteria. For the network variables, the maximum possible lag length considered is 15. The coefficient presented in the table correspond to the first lag of the variable. The number of in-sample observations is 667423.

Table 7: Prediction accuracy of the unrestricted multivariate models

	Lags	R^2	RMSE	RMSE Diff.
Model 1'	13	0.395	0.773	-
Multivariate Model 2'	15	0.321	0.814	-5.30%**
Multivariate Model 3'	8	0.416	0.758	1.94%**

** Significant at 1% level. For Multivariate model 3, we considered 8 lags for each network variable and 13 lags of the output. The lag length is selected according to the BIC criteria, for the multivariate models we only considered as candidate models those where each network variable has the same number of lags. The number of in-sample observations is 667423

Figure 1: A scatter plot of future output and recent past output.

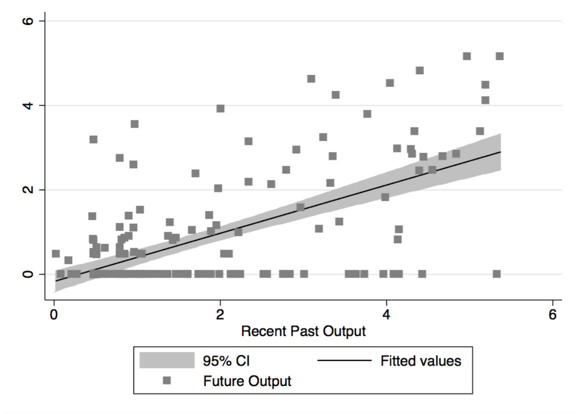


Figure 2: Scatter plots of future productivity on closeness centrality and coauthors' productivity.

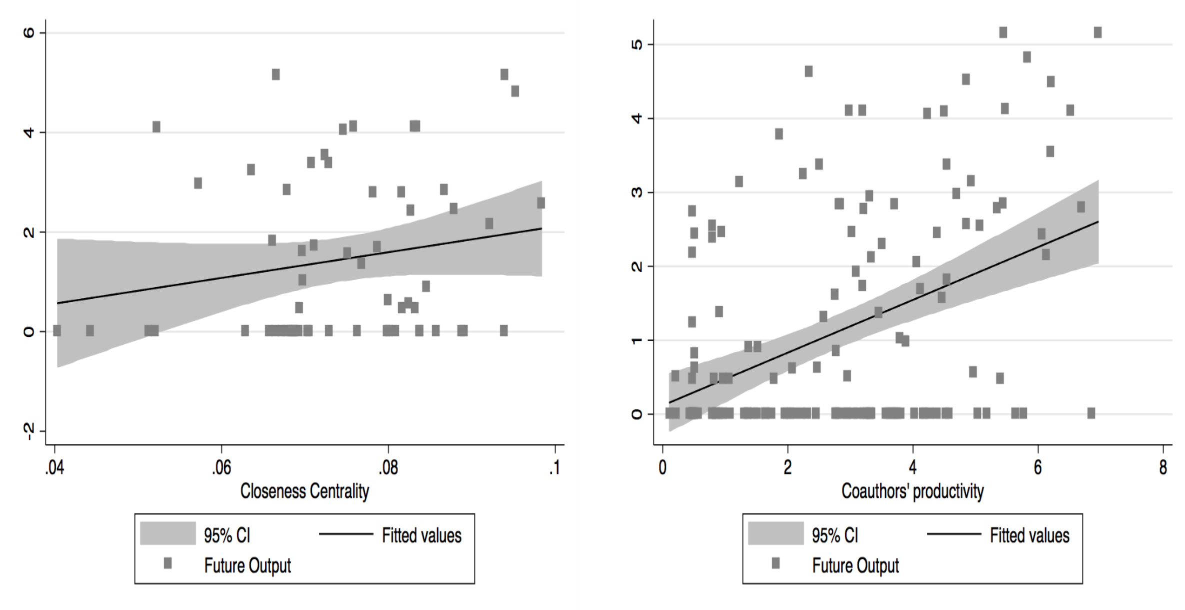


Figure 3: RMSE out-of-sample across Career time. Restricted Models

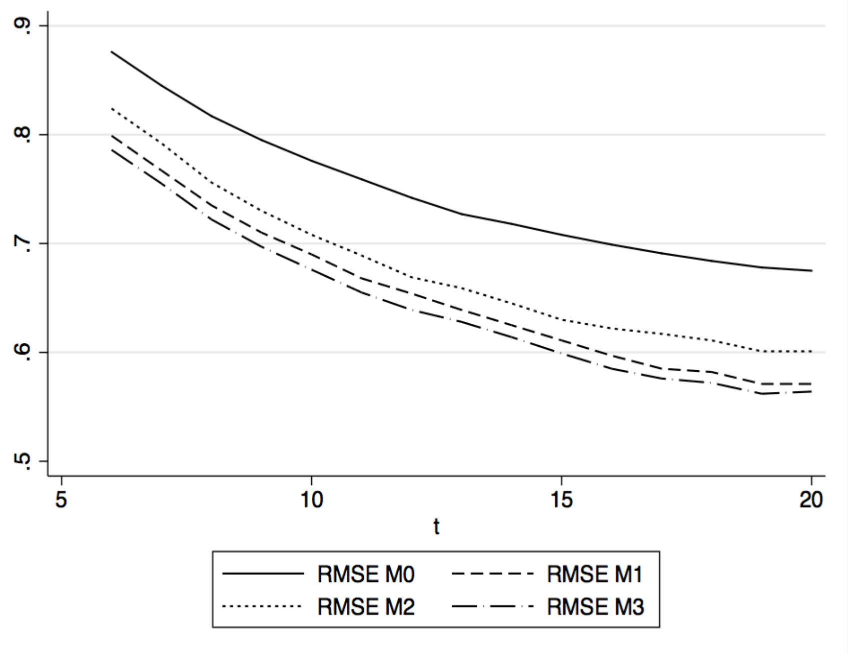


Figure 4: RMSE % Difference across Career time. Restricted Models

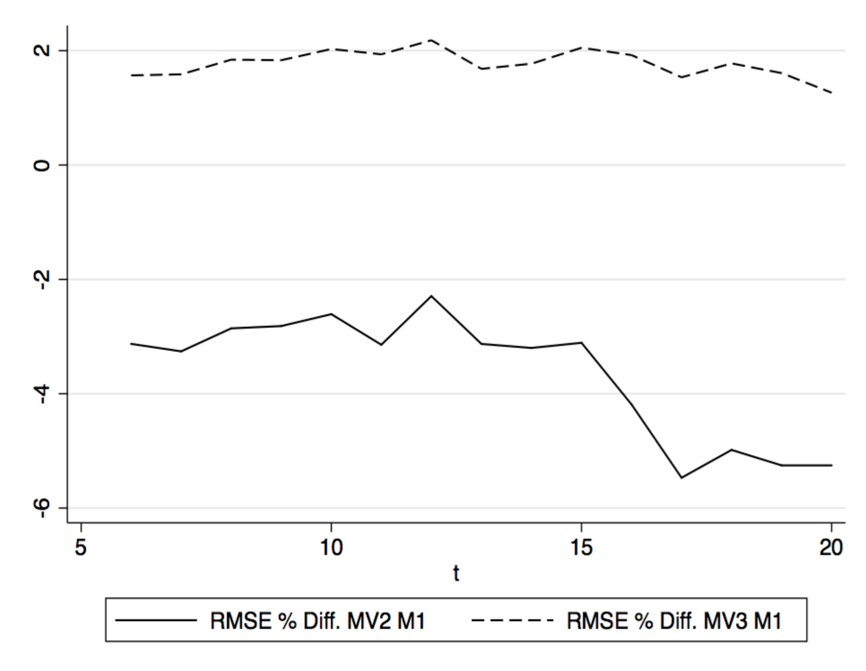
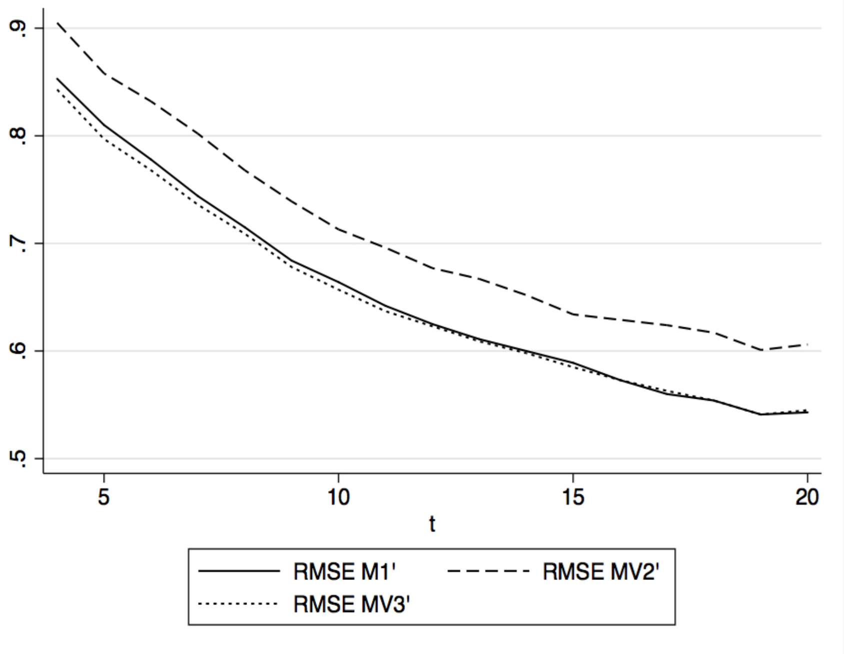
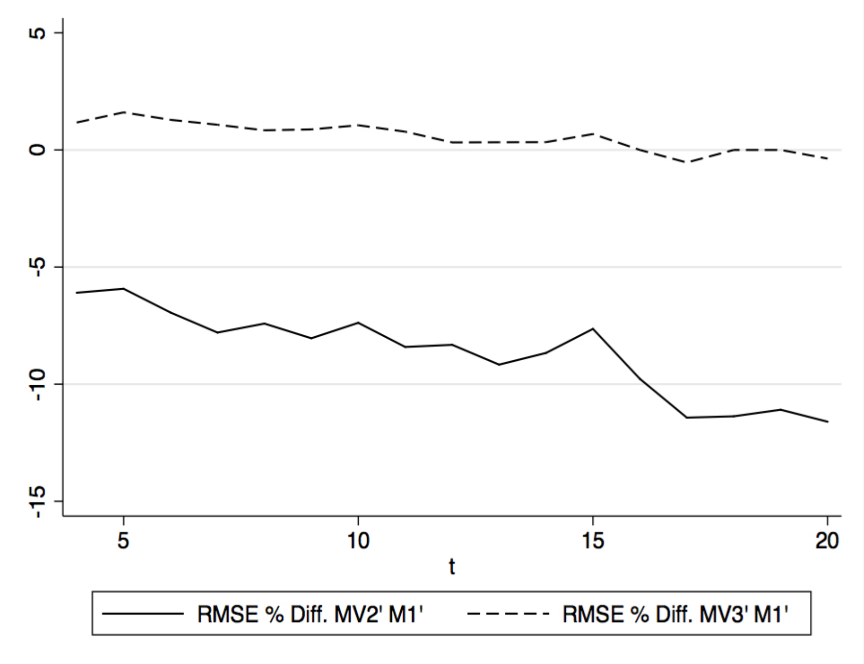


Figure 5: RMSE out-of-sample across Career time. Unrestricted Models



According to the Diebold-Mariano test, the RMSE % difference between Multivariate Model 3 and Model 1 are statistically significant for every career time.

Figure 6: RMSE % Difference across Career time. Unrestricted Models



According to the Diebold-Mariano test, the RMSE % difference between Multivariate Model 3 and Model 1 are insignificant for $t = 12$ and from $t = 14$ to $t = 20$.

Figure 7: RMSE % Difference between Restricted Models across productivity tiers.

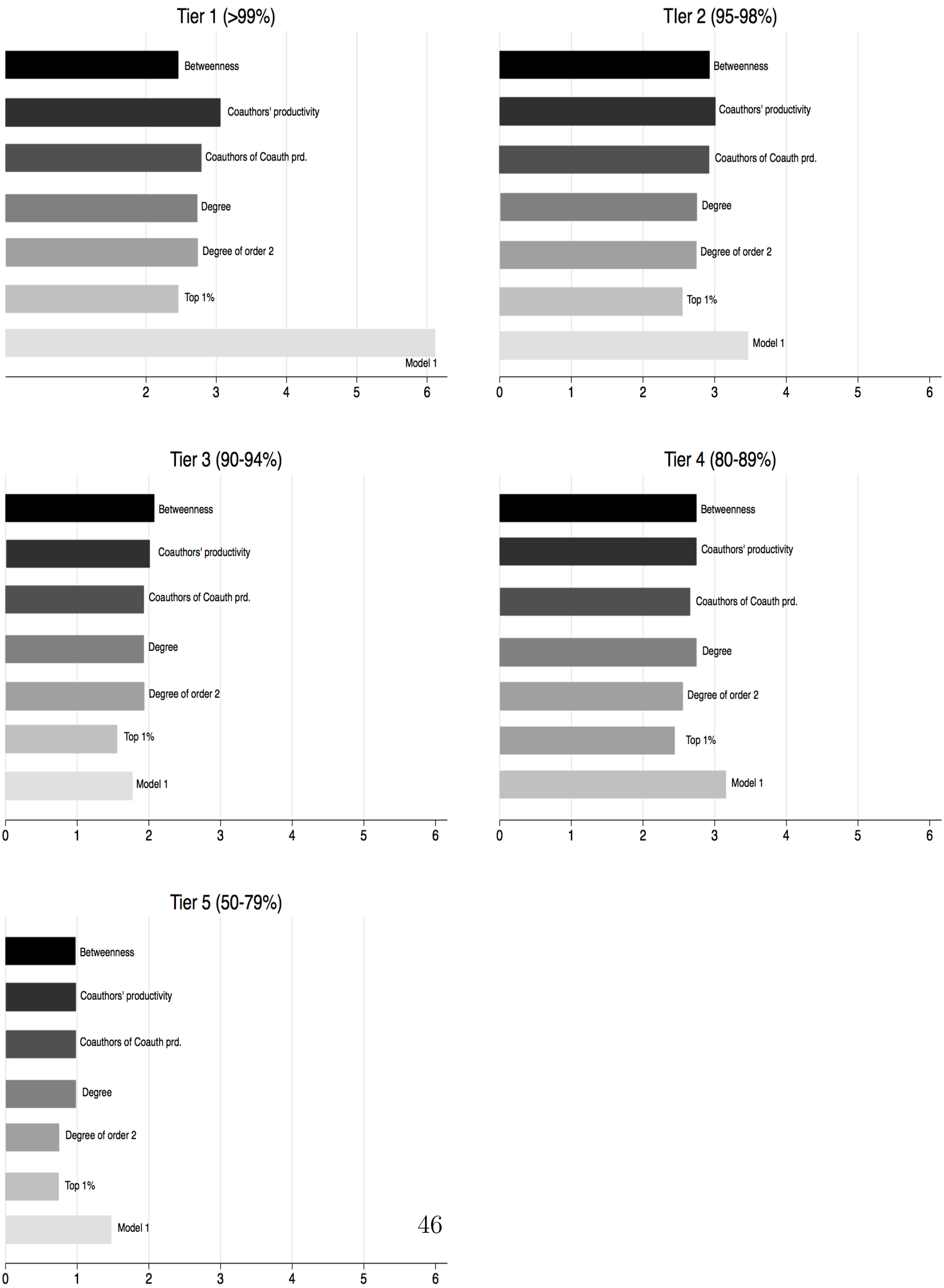
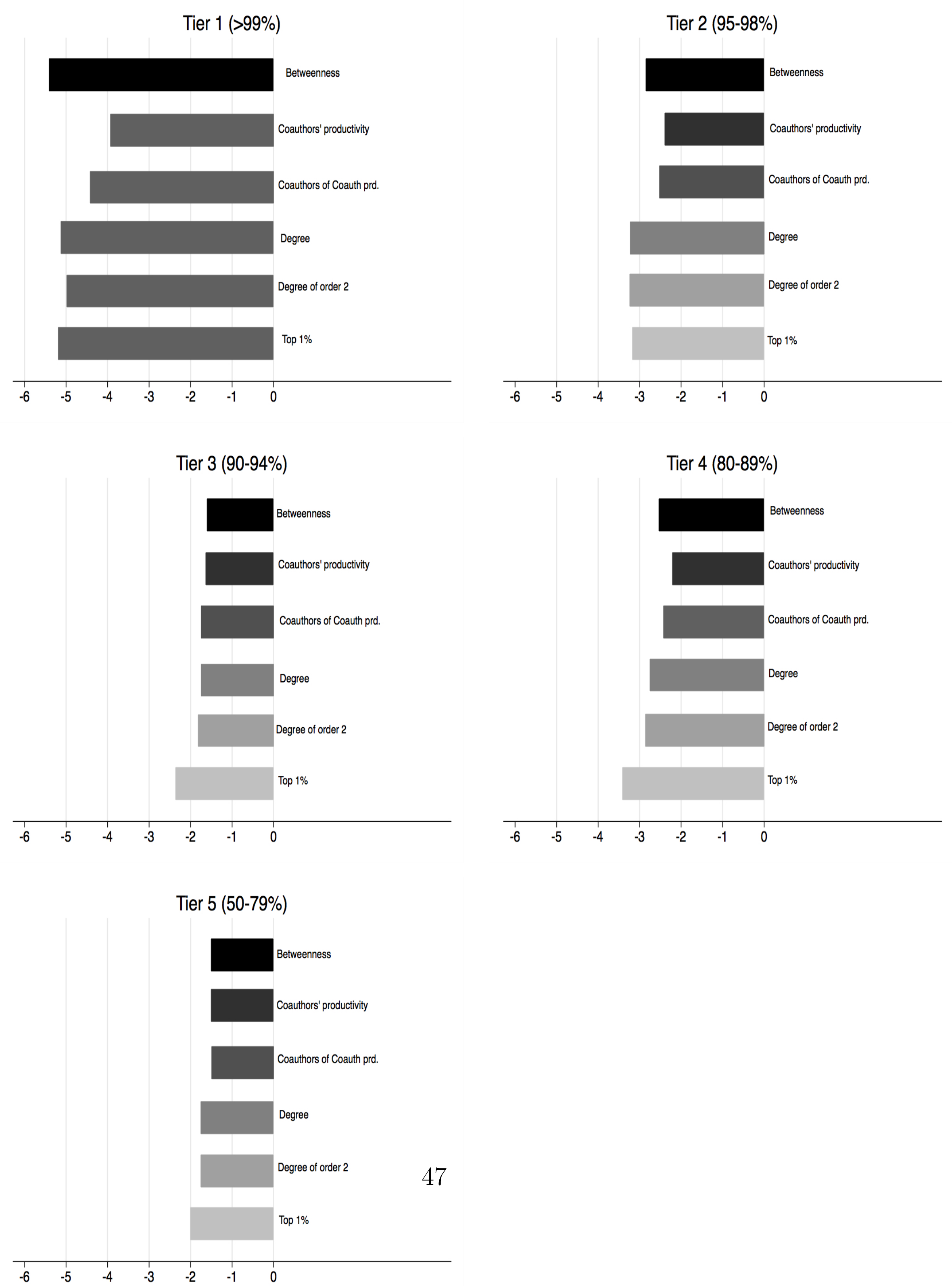


Figure 8: RMSE % Difference between Unrestricted Model 2 and Model 1 across productivity tiers.



7 Online Appendix: Detailed robustness analysis

In this appendix we report in detail on the analysis of the robustness of our results to various assumptions made in constructing the variables used in the estimation. We also provide details on how we derive the final specification of our model.

7.1 Model Specification

In this paper we estimate dynamic models that require stationary time series. A first concern is that the dependent variable, future research output, or the regressors may not be stationary, in which case differencing the series may lead to better predictions. To check stationarity, we test for the presence of unit roots in the productivity and network variables. We use the Harris-Tzavalis panel unit root test (Harris & Tzavalis, 1999) because it is suitable for panels with a large cross-sectional dimension and a fixed number of time periods. This test requires a strong balanced panel, a requirement that is not satisfied in our data. Each author enters the panel at the time of his or her first publication, and this timing naturally varies across authors. To implement the test, we divide the sample into sets of authors who are present over the same time window. More precisely, we first restrict our sample to authors who are in the panel for 15 years and we apply the Harris-Tzavalis test to this balanced sample. We then repeat the exercise but for authors who have been in the panel for 16 years, and so on.

Table 8 presents the results of the Harris-Tzavalis unit root test for all variables of interest for the set of authors who have been 15, 18, 21, 24 and 26 years in the panel. The results provide strong evidence of stationarity for almost all variables, except for the variables 'Degree' and 'Degree of order two', which show mixed results, in particular for shorter panel lengths. This may be due to a lack of power of the Harris-Tzavalis test. Based on this, we therefore decide not to difference the data.

Second, in the paper we consider as main specification a linear model in which the dependent variable is a logarithmic transformation of future research output, $\ln(q_{it}^f + 1)$, and we estimate this model by OLS. We are concerned that the model might be misspecified. To investigate this possibility, we estimate other commonly used models for research output and citation data, such as the Poisson, Negative Binomial, and Zero-inflated Negative Binomial models. Table 9 compares the performance of different linear and nonlinear models for the specification of 'unrestricted Model 1', that is, the specification without network variables. An optimal number of lagged dependent variables is included among the regressors. The results show that, although the in-sample log-likelihood is higher for the (Zero-inflated) Negative Binomial model, the out of sample RMSE is lowest for the linear model with an $\ln(y + 1)$-transformation.

Third, in the paper we report results from a pooled OLS regression. Can forecasts be improved by using a dynamic fixed effects model, either in first difference or using a system GMM estimator (Blundell & Bond, 1998)?[12] Table 10 presents the out-of-sample RMSE of different panel estimators applied to our data and shows that the pooled OLS performs best. This may be surprising to some, but it is rather understandable if one recognizes that our purpose is out-of-sample prediction rather than coefficient estimation. Panel data models are designed to reduce bias in coefficient estimates by filtering out distortions caused by individual effects. However, in order to predict the future performance of an individual author, one needs to predict the individual fixed effect as well. Out-of-sample predictions of

[12]The System GMM model deals with the dynamic panel bias introduced by the correlation between the lagged dependent variable and the error term in the fixed effect specification. Our specification includes as instruments the career time dummies, year dummies, years since the last publication, the network variables and the fourteen and fifteen lags of productivity.

individual effects are unavailable since there is no in-sample information on which to base their estimation. Hence we must rely on first-differencing the variables to filter out individual effects. It is well known that the correct estimation of first-differenced variables is difficult (Arellano & Bond, 1991). This is confirmed in our data: dynamic panel estimators do not lead to better out-of-sample predictions of future productivity in our data set.

Fourth, we estimate vector autoregressions (VARs) models that simultaneously allow past network variables to affect future output, and past output to influence future network variables. These VAR models are estimated using a seemingly unrelated regressions (SUR) approach, allowing for correlation in the error terms across the equations. The lag length of each equation is selected using the BIC criteria. The SUR regressions should lead to more efficient predictions, as long as the different equations do not always include the same set of lagged variables, a condition that is fulfilled here. Table 11 shows that the results from estimating the unrestricted SUR Model 3 using feasible generalized least squares (FGLS) hardly differ from the unrestricted Model 3 estimated using simple OLS. From this conclude that, in our data, the SUR VAR model does not outperform the OLS out of sample.

Finally, there remains a possible concern about functional form: if the predictive power of past output is non-linear and this non-linearity is correlated with network characteristics, this could generate a spurious predictive power for network variables. To investigate this possibility, we also consider an alternative specification with a quadratic term in each lagged productivity variable to capture a possible non-monotonic relationship between past and future output. The results, presented in Table 12, show that the coefficient of the first lag of the quadratic term is also positive and significant at the 1% level. But network variables remain significant and their predictive power is only slightly smaller than in the nonquadratic case. This confirms that network variables do not have predictive power simply because they are capturing a non-linearity in the effect of past output on future output.

7.2 Different definition of variables

In the analysis presented in the paper we have used average productivity from $t+1$ to $t+3$ as the variable q_{it} we seek to predict (see equation 2). The rationale for doing so is that the distant future is presumably harder to predict than the immediate future, and we want to give the model a fair chance. Yet, in economics there are long lags between the submission and publication of a paper, and wide variation in these lags across papers and journals. Variation in publication lags thus introduces additional variation in the variable we are trying to predict, and may thus lead us to underestimate the predictive power of network information. To check whether this is affecting our results, we repeat the analysis using average future productivity over a five year window instead of three:

$$q_{it}^{f} = q_{i,t+1} + q_{i,t+2} + q_{i,t+3} + q_{i,t+4} + q_{i,t+5}.$$

and, as before, use $\ln\left(1+q_{it}^{f}\right)$ as the variable we seek to predict.

Results, presented in Table 13, are similar to those reported in the paper except that the predictive power of network variables is larger using a five year window. In particular, network variables are even more useful than past output to forecast the future performance of a researcher.

Next we investigate whether results are sensitive to the definition of output q_{it}. It is customary for studies of author and departmental productivity in economics to correct for article length (e.g., Kodrzycki and Yu, 2005) and number of co-authors. To check whether this affects our results, we redo the analysis using a productivity measure that corrects for the number of published pages and number of co-authors, i.e., using:

$$q_{it} = \sum_{j \in S_{it}} \frac{\text{pages}_j * \text{journal quality}_j}{\text{Number of coauthors}_j} \tag{3}$$

The variable "pages$_j$" is the number of pages of article j divided by the average number of pages of articles published in the journal.[13] For comparison purposes, q_{it} assigns 28 points to a single-authored 9 page article or a 18 page two-author paper in the *American Economic Review.* As our definition of output. Results, shown in Table **??**, show that the predictive power of network variables is unaffected.

7.3 Active sample

In the results reported so far, we keep all the authors in the dataset and replace missing lagged productivity and network variables by zeros. The rationale for doing so is that authors who have just started their career have no past output and coauthorships, hence the value of their lagged productivity and network variables are truly zero. Without such replacement we would loose the first years of an author's career and this could bias results.

We nevertheless worry that this may introduce another kind of bias in the prediction of network variables. As an inactive author (i.e., an author without many publications) matures, future output and network variables are both zeros but lagged productivity is not. As a result network variables might have an increasing predictive power across time. To investigate whether such a bias affects our results, we redo the analysis dropping those observations where an author did not publish anything for five consecutive years or more.

[13]The number of pages is truncated above fifty pages to correct for a small number of unusually long published articles. Overly long papers are usually literature review articles. Hence not truncating above fifty pages would probably overrepresent their contribution.

Results are presented in Table 15. We observe that, once we restrict the sample to active authors, the RMSE is higher for all models compared to the full sample: not surprisingly perhaps, all models find it more difficult to predict the future productivity of authors who publish little. However, the relative RMSE improvement from the inclusion of network variables is similar across the two samples. Hence, our main conclusions are unaffected.

References

[1] Arellano, Manuel and Stephen Bond (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58(2): 277-97, April.

[2] Blundell, Richard and Stephen Bond (1998). "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 87(1): 115-143, August.

[3] Harris, R.D.F. and E. Tzavalis (1999), Inference for unit roots in dynamic panels where the time dimension is fixed, *Journal of Econometrics*, 91, 201–226.

[4] Kodrzycki, Yolanda K. and Pingkang David Yu (2005), "New approaches to ranking economics journals," Working Papers 05-12, Federal Reserve Bank of Boston.

Table 8: Harris-Tzavalis Unit Root Tests

	15 years	18 years	21 years	24 years	26 years
Future output	.5671 (.0000)	.5816 (.0000)	.5552 (.0000)	.6059(.0000)	.6095 (.0000)
Five Year Network Variables:					
Degree	.8198(.7457)	.8144(.0031)	.8322(.0007)	.8572(.0063)	.8118 (.0000)
Degree of order two	.8137 (.5432)	.8284(.0883)	.8475 (.0509)	.8777(.4015)	.8268 (.0000)
Giant component	.5801 (.0000)	.6004(.0000)	.6309(.0000)	.6278(.0000)	.6334 (.0000)
Betweenness centrality	.6444 (.0000)	.6582 (.0000)	.6842 (.0000)	.6972 (.0000)	.6698 (.0000)
Closeness centrality	.6551(.0000)	.6830(.0000)	.7025 (.0000)	.6970(.0000)	.6734 (.0000)
Coauthors' productivity	.6841 (.0000)	.6923(.0000)	.7094(.0000)	.6945 (.0000)	.7048 (.0000)
Coauthors' of Coauthors' prod.	.6797(.0000)	.6927(.0000)	.6803(.0000)	.6907(.0000)	.6981(.0000)
Working with a top 1%	.5327 (.0000)	.5465(.0000)	.5917 (.0000)	.5709 (.0000)	.5234 (.0000)

The first order autocorrelation coefficient is provided, p-values in parenthesis. Panel means are incorporated in all the tests, time trends are not. Each network variable is computed using links from the last 5 years. The Harris-Tzavalis (HZ) test requires strong balanced panel. Since our panel is unbalanced we report the test for authors with a career lifetime of 15 years, 18 years, 21 years, 24 years and 26 years, results for the rest of authors are available upon request.

Table 9: Prediction performance of unrestricted Model 1' using different functional forms

	Log-likelihood	AIC	BIC	RMSE
Log(y+1)	-782,223	1,564,580	1,565,344	16.69
Level	-2,861,612	5,723,358	5,724,122	17.16
Poisson	-3,777,972	7,556,078	7,556,842	19.67
Negative Binomial	-766,673	1,533,482	1,534,258	19.58
Zero inflated NB	-760,749	1,521,637	1,522,436	19.54

RMSE is obtained from out-of-sample level predictions. The $\ln(y+1)$ model is re-transformed using $\exp(\beta' X_{i,t})\exp(u_{i,t}) - 1$. The $E(u_{i,t})$ is estimated by the sample average $N^{-1}\sum \exp(\widehat{u_j})$ where N is the total number of observations.

Table 10: Prediction performance of Unrestricted Models using Panel data models

	RMSE Model 1	RMSE MV. Model 2	RMSE MV. Model 3
Log(y+1)	.680	.742	.673
First Difference	.711	.764	.713
System GMM	.780	-	.760

In the System GMM model, we use as instruments the career time dummies, year dummies, number of years since the last publication, the fourteen and fifteen lags of productivity and the network variables for the equation in differences. System GMM is not computed for MV. Model 2 since past output is not included.

Table 11: Prediction Accuracy of Unrestricted Model 1 and 3. Seemly Unrelated Regression.

	Lag Length	R^2	RMSE	RMSE Diff.	Coefficients
Model 1					
Past output	13	.39	.773	-	.36**
Model 3					
Degree	6	.40	.768	.65%**	.15**
Degree of order 2	5	.40	.768	.65%**	.06**
Giant component	8	.40	.768	.65%**	.58**
Closeness centrality	10	.40	.767	.78%**	2.35*
Betweenness centrality	9	.40	.767	.78%**	.02
Coauthors' productivity	12	.41	.761	1.55%**	.09**
Coauthors of Coauthors' prod.	11	.41	.764	1.16%**	.07**
Working with a top 1%	13	.40	.767	.78%**	.39**
Multivariate Model 3		.42	.759	1.81%**	

** Significant at 1% level, * Significant at 5%. We estimate future productivity and future network variables in the SUR model using the feasible generalized least squares method. The results from estimating future network variables are available upon request. The lag length for each model is selected using BIC. The coefficients presented in the table correspond to the first lag of each variable. The number of in-sample observations is 667423.

Table 12: Prediction accuracy of the unrestricted multivariate models. Including quadratic past output.

	Lags	R^2	RMSE	RMSE Diff.	Coefficients
Model 1'					
Past output	13	.40	.770		.23**
Past output squared					.05**
Model 3'					
Degree	6	.40	.765	.65%**	.13**
Degree of order 2	5	.39	.766	.52%**	.05**
Giant component	8	.40	.765	.65%**	.38**
Closeness centrality	10	.40	.764	.78%**	3.19**
Betweenness centrality	9	.40	.764	.78%**	-.02
Coauthors' productivity	12	.41	.759	1.43%**	.08**
Coauthors of Coauthors' prod.	11	.40	.762	1.04%**	.05**
Working with a top 1%	13	.39	.766	.52%**	.26**
Multivariate Model 3'	8	.42	.757	1.69%**	

** Significant at 1% level. Model 1' and Multivariate Model 3' include 13 lags of the productivity variable and their quadratic terms. Column 4 shows the coefficient of the first lag of each variable. We include 8 lags of the network variables and 11 lags of the output in Multivariate Model 3'. The lag length is selected according to the BIC criteria. The number of in-sample observations is 566040.

Table 13: Prediction Accuracy: Unrestricted Models 1' and 2'; using 5-period productivity variable.

	Lag Length	R^2	RMSE	RMSE Diff.	Coefficients
Model 1'					
Past output	11	.34	.946	-	.51***
Model 2'					
Degree	13	.26	.998	-5.50%***	.04***
Degree of order 2	13	.22	1.006	-6.34%***	.06***
Giant component	15	.25	1.005	-6.24%***	1.05***
Closeness centrality	15	.26	.998	-5.50%***	1.75
Betweenness centrality	12	.28	.985	-4.12%***	.08**
Coauthors' productivity	13	.31	.963	-1.80%***	.12***
Coauthors of Coauthors' prod.	12	.29	.980	-3.59%***	.09***
Working with a top 1%	14	.26	.999	-5.60%***	.50***
Multivariate Model 2'	13	.34	.942	.42%**	
Model 3'					
Degree	15	.37	.928	1.90%***	.16***
Degree of order 2	15	.36	.932	1.48%***	.07***
Giant component	15	.36	.931	1.59%**	.84***
Closeness centrality	15	.36	.929	1.80%***	2.04*
Betweenness centrality	15	.37	.927	2.01%***	.03
Coauthors' productivity	7	.39	.912	3.59%***	.11***
Coauthors of Coauthors' prod.	15	.38	.921	2.64%***	.08***
Working with a top 1%	9	.37	.928	1.90%***	.48***
Multivariate Model 3'	9	.40	.904	4.44%**	

*** Significant at 1% level, ** Significant at 5%, * Significant at 10%. The dependent variable is the future productivity from $t+1$ to $t+5$. For Multivariate Model 3', we include 9 lags of the network variables and 11 lags of the output. The lag length is selected according to the BIC criteria. The number of in-sample observations is 566040.

Table 14: Prediction Accuracy: Unrestricted Models 1', 2' and 3'; using discounted productivity.

	Lag Length	R^2	RMSE	RMSE Diff.	Coefficients
Model 1'					
Recent past output	14	.40	.670	-	.43***
Model 2'					
Degree	15	.21	.760	-13.43%***	.08***
Degree of order 2	15	.21	.762	-13.73%***	.04***
Giant component	15	.21	.762	-13.73%***	.86***
Closeness centrality	15	.22	.756	-12.84%***	.34
Betweenness centrality	15	.24	.747	-11.49%***	.07**
Coauthors' productivity	15	.28	.729	-8.81%***	.10***
Coauthors of Coauthors' prod.	15	.25	.742	-10.75%***	.08***
Working with a top 1%	14	.23	.754	-12.54%***	.38***
Multivariate Model 2'	15	.31	.714	-6.57%**	
Model 3'					
Degree	4	.40	.666	.60%***	.10***
Degree of order 2	5	.40	.666	.60%***	.05***
Giant component	8	.40	.666	.60%***	.50***
Closeness centrality	8	.40	.665	.75%***	1.68*
Betweenness centrality	9	.40	.665	.75%***	0.02
Coauthors' productivity	12	.41	.660	1.49%***	.08***
Coauthors of Coauthors' prod.	12	.41	.662	1.19%***	.06***
Working with a top 1%	12	.41	.665	.75%***	.32***
Multivariate Model 3'		.42	.658	1.79%**	

*** Significant at 1% level, ** Significant at 5%, * Significant at 10%. The coauthors' productivity, coauthors of coauthors' productivity and the dummy variable working with a top 1% have been obtained using the discounted productivity. The number of in-sample observations is 667423

Table 15: Prediction Accuracy: Unrestricted Models 1' and 2'. Active Sample

	Lag Length	R^2	RMSE	RMSE Diff.	Coefficients
Model 1'					
Recent past output	13	.35	.999	-	.36**
Model 2'					
Degree	12	.22	1.095	-9.61%**	.04**
Degree of order 2	11	.22	1.09	-9.11%**	.03**
Giant component	15	.22	1.091	-9.21%**	.86**
Closeness centrality	15	.24	1.080	-8.11%**	1.52**
Betweenness centrality	15	.25	1.075	-7.61%**	.07**
Coauthors' productivity	12	.27	1.058	-5.91%**	.10**
Coauthors of Coauthors' prod.	12	.25	1.070	-7.11%**	.08**
Working with a top 1%	14	.22	1.096	-9.71%**	.36**
Multivariate Model 2'		.30	1.039	-4.00%**	
Model 3'					
Degree	5	.36	.994	.50%**	.11**
Degree of order 2	5	.36	.993	.60%**	.05**
Giant component	8	.36	.993	.60%**	.58**
Closeness centrality	7	.36	.991	.80%**	2.21*
Betweenness centrality	9	.36	.991	.80%**	.02**
Coauthors' productivity	6	.37	.984	1.50%**	.09**
Coauthors of Coauthors' prod.	10	.37	.988	1.10%**	.06**
Working with a top 1%	6	.36	.993	.60%**	.44**
Multivariate Model 3'		.38	.980	1.90%**	

** Significant at 1% level, * Significant at 5%. Active sample: we dropped all observations where the recent output is zero. The number of in-sample observations is 357832.