Network Structure and Performance

Ilse Lindenlaub[*] Anja Prummer[†]

May 13, 2020[‡]

Abstract

We develop a theory that links individuals' network structure to their productivity and earnings. While a higher degree leads to better access to information, more clustering leads to higher peer pressure. Both, information and peer pressure, affect effort in a model of team production, each beneficial in different environments. We find that information is particularly valuable under high uncertainty, whereas peer pressure is more valuable in the opposite case. We apply our theory to gender disparities in performance. We document that men establish more connections (a higher degree) whereas women possess denser networks (a higher clustering coefficient). We therefore expect men to outperform women in jobs that are characterised by high uncertainty in project outcomes and earnings. We provide suggestive evidence that support our predictions.

Keywords: Networks, Peer Pressure, Information, Labour Market Outcomes, Gender
JEL Classification: D85, Z13, J16

[*] Contact: Department of Economics, Yale University, 28 Hillhouse Avenue, 06520 New Haven, USA, ilse.lindenlaub@yale.edu

[†] Contact: School of Economics and Finance, Queen Mary University of London, Mile End Road, London, E1 4NS, UK, a.prummer@qmul.ac.uk

[‡] We thank the editor Rachel Kranton and the referees for excellent comments. We are grateful to our advisors Jérôme Adda, Jan Eeckhout, Massimo Morelli, Nicola Pavoni and Fernando Vega Redondo for helpful discussions and advice. We further would like to thank Carlos Alós-Ferrer, Alp Atakan, Estelle Cantillon, Hector Chade, Francesca Cornaglia, Giancarlo Corsetti, Raquel Fernandez, Edoardo Gallo, Sanjeev Goyal, François Maniquet, Andrea Mattozzi, Sujoy Mukerji, Francesco Nava, Brian Rogers, Carl Sanders, Karl Schlag, Klaus Schmidt, Jan-Peter Siedlarek, Adam Szeidl, Yves Zenou as well as seminar participants at the University of Cambridge, the University of Cologne, UC Louvain-CORE, Ludwig-Maximilians University Munich, the EUI Micro Working Group, the Berkeley Student Theory Group, EEA 2013, SAET 2013, CTN 2013 and PET 2013 for helpful comments. We thank Tianhao Wu and Mrinmoyee Chatterjee for very valuable research assistance. Additionally, we thank Unicredit who awarded us for this paper with the *UWIN Best Paper Award*. All remaining errors are ours. A previous draft was circulated under the title 'Gender, Social Networks and Performance'. First draft: November 2012.

Loose connections are the connections you need. It's the No. 1 rule of business.
Sallie Krawcheck, owner of the global women's network *85 Broads*[1]

1 Introduction

It is common wisdom that networks matter for labour market outcomes. But what network structures are particularly beneficial? Different network types are associated with distinct advantages. While the importance of loose connections (Granovetter (1973)) has been shown to be especially valuable in the context of job search as it grants access to information, it is far from obvious that these networks are still optimal *on the job*. There, different considerations come into play, making tight networks that generate peer pressure (Coleman (1988)) potentially more advantageous.

Our main contribution is to formalise the trade-off between social capital generated by loose versus tight networks and to illustrate the importance of the *structure* of social networks for labour market outcomes. We show that a loose network with more connections (a higher degree) allows for better performance in uncertain environments with potentially high but risky returns. In turn, a tight network (with greater clustering) will lead to better performance in stable environments. In a new application, we relate our theory on network differences to the gender gap in job performance. We first document a novel fact about gender disparities in network structure. We establish that women have on average fewer connections, a lower degree, while their connections tend to be linked, resulting in a higher clustering coefficient compared to men. We then link our theoretical predictions to the gender gap in performance, which is particularly pronounced in risky occupations, and to occupational sorting. The evidence suggests that our theory can help understand why gender disparities are more pronounced in risky work environments.

To formalize the trade-off between access to information and peer pressure and their impact on *performance on the job*, we develop a model in which workers differ regarding their network structure. They are repeatedly selected into partnerships to complete projects of uncertain output value. Project success positively depends on the partners' efforts, where effort is unobservable. If the project is completed successfully, the project payoff is shared between the team members. Because output is split but effort costs are not, there is a team moral hazard problem at work, inducing inefficiently low effort (as in Holmstrom (1982)).

Networks can ameliorate this problem and increase effort, with different network structures achieving this through distinct mechanisms. Agents with a higher *degree* receive more signals as they can observe not only their own signal about the value of the project, which can be high or low, but also the signals of their friends in the network. Therefore, a loose network characterised by a higher degree leads to better information, which allows workers to identify a valuable project on which to exert higher effort. In turn, workers with higher *clustering* face more peer pressure through the following mechanism: failure at the workplace leads to frictions

[1] Krawcheck at Marie Claire's luncheon for the *New Guard*, November 2013.

not only among project partners but also between them and their common connections, that is their disagreement spreads through the entire group – an idea based on the *structural balance theory*.[2] Since an intact relationship is necessary for a successful project, repercussions of a failure are especially bad for a worker with high clustering. Therefore, higher clustering leads to higher effort in order to be on good terms with future potential project partners.

We are interested in the effort levels of the project partners as a proxy for their performance and specifically in how valuable different network structures are in distinct environments. Our main theoretical findings are as follows: A higher degree is more beneficial for performance in volatile environments, where the uncertainty about the project value is considerable. This is true when (i) overall information (that is, information coming from sources unrelated to the network) is scarce, (ii) when signals are noisy, and (iii) when project rewards differ significantly across states. In these cases, uncertainty about the state of the world and associated rewards is large and the benefits of purely information-based, loose networks outweigh the benefits of closed networks that lead to more peer pressure. In turn, peer pressure leads to relatively higher effort and thus project completion in environments characterised by little uncertainty where additional information has no value. Note that workers facing high peer pressure exert extra effort even if the expected project reward is low. Thus, peer pressure induces agents to be undiscerning about when to put effort. Information, on the other hand, reduces effort if agents anticipate a low project value, thus fine-tuning their effort to the expected project reward. Even though information can reduce effort, it turns out that more information is still valuable in expectation, driven by the advantage of superior effort adjustment. We further show that degree and clustering are complementary: The marginal effect of clustering on effort is particularly large when degree is high (i.e. when information is abundant) and vice versa.

Effort choices then translate into wages. Someone with higher clustering earns more than someone with higher degree when uncertainty about the state is negligible. Such a worker has a comparative advantage in jobs whose outcomes are more certain compared to jobs with less certain outcomes. We also show that, in line with our result on effort, the marginal return to clustering is higher when the degree is high. Finally, due to the dynamic effect of clustering, there is a strong persistence of wage patterns across time, consolidating early career wage gaps.

We propose a novel application of this theory by connecting our predictions to gender differences in labour market outcomes. We proceed by (i) documenting gender differences in network structures in a variety of environments, (ii) showing how the gender wage gap varies with the uncertainty of the occupation, in line with our predictions, and (iii) directly connecting labour market outcomes and network patterns in two distinct settings.

Men and women differ in their average network structure. Women have fewer connections than men, that is they have a lower *degree*, but their peers are more likely to be connected among each other, implying a higher *clustering coefficient*. Thus, women have smaller but tighter networks, whereas men have larger but looser networks. This observation, for instance, holds

[2] This is a concept first proposed by Heider (1946) who has spawned a field of research that remains active until today. For an overview on structural balance theory, see Easley and Kleinberg (2010), chapters 3 and 5.

true for academic computer scientists (from the dblp computer science bibliography) where we build the scholars' networks based on co-authorships. We also find support for these patterns beyond Academia, namely from the Enron company where we construct the employees' network based on email exchanges. Finally, we document these network differences across gender in the AddHealth data set based on friendship nominations, where we focus on young adults that are about to enter the labour market. These environments – academia, private company and schools – vary considerably, highlighting the pervasiveness of these gender disparities in network structure. It is beyond the scope of this paper to analyse *the source* of network differences between men and women with the available data.[3] But even though we do not provide an explanation for why these network differences across gender arise, we believe the fact in itself is of interest and to the best our knowledge, this paper is the first to document it.

We then connect the key predictions of our model regarding network structures to gender differences in labour market outcomes: We first show that women perform worse in uncertain environments, while having a comparative advantage in low risk occupations. Using a representative data set for the US, the US Census, we show that women have particularly low earnings in high risk occupations such as legal occupation or management and are less likely to select into those occupations, where we measure risk by the earnings risk of an occupation.This is in line with our prediction given their tighter networks.

In order to assess the connection between network characteristics and labour market outcomes more directly, we turn to AddHealth, where we can relate network features and labour market outcomes. Our results suggest that tight networks with high clustering are indeed positively correlated with a comparative advantage in low risk occupations. We further relate networks in Computer Science to different, commonly used measures of academic performance, obtained from Google Scholar profiles. As research is characterised by complex and, especially, uncertain tasks, we view research in general and computer science in particular as an intrinsically uncertain occupation. Our correlations indicate that loose networks are associated with a better performance in this risky setting and that women could potentially improve their performance with a different network structure.

Based on these findings, we argue that network differences across gender *at work* may be an overlooked source of well-known gender gaps in the labour market, especially in risky environments where women perform particularly poorly.

Studies investigating gender differences in other academic disciplines corroborate our theory and findings. Ductor, Goyal, and Prummer (2018) first show that in Economics a higher degree, a more loose network, is correlated with higher research output across all performance measures considered. In turn, a higher clustering coefficient, a tighter network, is negatively related with output. Second, they also control for gender and show that women have a lower research output. Importantly, having a higher degree helps close the gender productivity gap in Economics, while

[3] These differences might be due to different patterns of socialisation or distinct preferences. To analyse the origins we would require more systematic data of children at younger ages, which to the best of our knowledge is not available at this point.

a higher clustering coefficient exacerbates it even further. Our conclusion is further supported by case studies from the film industry as well as patented research, highlighting the importance of loose networks in uncertain environments, also beyond academia.

Related Literature We contribute to a small, but distinguished literature on the relative advantages of different network structures. This literature goes back to seminal, but contradictory, work by Granovetter (1973) and Coleman (1988). While Granovetter (1973) emphasizes the importance of loose connections, Coleman (1988) postulates that tight networks are crucial in overcoming trust issues as well as free riding. This debate has spawned influential research focusing on the advantages of *one type* of network structure, which is translated into social capital (Putnam (2000), Burt (1992), Lin (1999)). Our paper reconciles these two strands of literature and resolves the conflict between different network advantages: loose networks are best in uncertain environments where information is crucial. In turn, tight networks are most beneficial if free riding is a greater concern than information acquisition.

This trade off has been addressed by Dixit (2003) for trading networks. Karlan, Mobius, Rosenblat, and Szeidl (2009) highlight the distinct advantages of different network structures in the context of borrowing through networks. In contrast, we focus on networks in the labour market. Networks first rose to prominence due to their explanatory power in the labour market, with a particular focus on referral networks, see Montgomery (1991), Marsden and Gorman (2001), Arrow and Borzekowski (2004), Calvó-Armengol and Jackson (2004, 2007), Calvó-Armengol and Zenou (2005). In this literature, agents who search for a job through their network face lower unemployment risk and receive higher wages. The reason is that they are more likely to hear about jobs and the associated wages. This allows them to extract a wage premium. In order to find a well-paying job, it is particularly beneficial to possess a loose network. Our work is complementary to this well-established literature: we investigate the importance of networks *on the job*.

We also provide a novel explanation for the gender gap in labour market outcomes — disparities in network structure. Common explanations of these gender differences are discrimination (Goldin and Rouse (2000)), differences in the number and length of career interruptions and overall labour force experience (Bertrand, Goldin, and Katz (2010), Gayle and Golan (2011)), differences in performing job tasks with low-promotability (Babcock, Recalde, Vesterlund, and Weingart (2017)), differences in competitiveness (Gneezy, Niederle, and Rustichini (2003), Niederle and Vesterlund (2007), Dohmen and Falk (2011)) or exogenous differences in hours worked at home, inducing women to choose low-hours-low-wage occupations (Erosa, Fuster, Kambourov, and Rogerson (2017)). Nevertheless, these factors are not sufficient to fully close the gender gaps. By focussing on a new disparity between men and women (their networks), we provide a novel angle to the ever-pressing question of what explains the gender gaps in labour market outcomes.

The paper proceeds as follows: In Section 2, we develop our model and Section 3 contains our main theoretical results. We provide evidence on gender disparities in networks, gender differences in labour market outcomes and how the two relate in Section 4. Section 5 concludes.

2 Model

We consider an undirected network g of N workers. Two of those workers, $i, j \in N$, are selected in each period t. We focus here on a two period model, $t \in \{1, 2\}$.[4] Once two workers are selected they have to complete a project. Whether they are successful depends on their exerted effort, which in turn depends on their network structure and past project outcomes. In order to highlight how each of these factors matters we first consider the game that is played in each period t.

1. Worker Selection. At the beginning of each period, two workers are randomly sampled (without replacement) from the set of workers to complete a project. Whenever two workers i and j have a direct link, denoted by $g_{ij} = g_{ji} = 1$, they have an informal connection. We assume that workers can only complete their project successfully if there exists a direct link between them. If there is no link between the two selected workers, their project fails and both workers receive a payoff of zero. The number of links of worker i, his degree, is denoted by D_i. Then, the joint probability of being selected for a project and being partnered with a directly connected worker is given by (see Appendix A for details)

$$s_i = \frac{2D_i}{N(N-1)}. \tag{1}$$

This probability is proportional to the degree of an individual. This implies that workers with higher degrees will be selected more often into potentially profitable projects.[5]

2. Information. Every period is characterized by a state of the world, θ, which is high or low

$$\theta = \begin{cases} \theta_h & \text{with probability} \quad q \\ \theta_l & \text{with probability} \quad 1-q \end{cases}$$

and iid. It is drawn after project teams are formed and is not observable to the workers. In the high (low) state, the project value is $2v_h$ ($2v_l$), with $v_h > v_l$. We assume that the payoff of the project is split equally among the project partners.[6]

In the following, we show how a worker's network structure affects his information about the state of the world. Each worker obtains a signal about the state (with a signal value of one (zero) indicating the high (low) state) but he can also observe the signals of those workers he is directly connected to. We denote the probability of a correct signal by p and assume that signals are informative with $p > 1/2$.

Since we focus on ego networks (i.e. the network of an individual), we distinguish between the number of signals a worker obtains internally, by himself and from his direct friends, $n_{int,i} =$

[4] We provide a discussion of the extension to an infinite horizon in the Online Appendix.

[5] This assumption is supported by Aral, Brynjolfsson, and Van Alstyne (2012), who study project performance in a recruiting firm. They find that peripheral nodes, i.e. nodes that are not well connected, do fewer projects per unit of time than central nodes.

[6] We impose the equal split assumption as we aim for a model in which agents are perfectly symmetric except for their network. This allows to analyse the effects of network structures in the cleanest way possible.

$D_i + 1$, and the signals he obtains from external sources (which can include the signals of a neighbor's friends), $n_{ext,i}$. This enables us to vary the baseline amount of information below. We denote by $n_i = n_{int,i} + n_{ext,i}$ the overall number of signals of worker i.
Based on his signals, a worker then computes a sufficient statistic $y_i = (x_i, n_i - x_i)$, where $x_i \in \{0, 1, \ldots, n_i\}$ is the number of high signals out of all observed signals and $n_i - x_i$ denotes the number of low signal. We further assume that co-workers share their information, which implies that two project partners always hold the same information.[7]

Based on y_i, the posterior probability of being in the high state, $P(\theta_h|y_i)$, is computed via Bayesian updating and thus having a higher number of signals gives a more precise posterior. The project value for agent i, $\pi(y_i)$, is then given by

$$\pi(y_i) = P(\theta_h|y_i)v_h + (1 - P(\theta_h|y_i))v_l. \tag{2}$$

To summarise, the network structure matters as a higher degree gives a higher number of internal signals, which in turn affects the expectation about the project value.

3. Choice of Effort. The paired workers simultaneously choose what effort, $e_i, e_j \geq 0$, to exert on the project. This effort is costly with all workers facing the same cost function $c(e)$, which we assume to be convex. We focus on the effort choice of two directly linked team mates. Effort makes project success more likely. The probability that the project is completed if effort choices are e_i and e_j is given by $f(e_i, e_j) \in [0, 1)$. To ensure that $f(e_i, e_j)$ is strictly smaller than one, we assume that effort is bounded.[8] This implies that success cannot be guaranteed. Further, we make some natural assumptions on the success function f, namely that it is twice continuously differentiable, increasing and concave in each argument, that it has constant returns to scale and is symmetric in both arguments. Moreover, we assume that f is strictly super-modular, $f_{12} = f_{21} > 0$, implying that effort levels of the workers are strategic complements. We focus on complements as the natural benchmark for a team problem: With substitutes a worker should complete the project by himself, circumventing the team moral hazard problem that stems from the individual team partner bearing the full cost of effort but only obtaining a share of the project value. Finally, if one team member chooses zero effort, the project fails for sure. After effort has been chosen, the project outcome – success or failure – is realised. A worker's payoff is his share of the expected project value minus the cost of effort.

These three stages – worker selection, information acquisition, and effort choice – occur in both periods. What differs across periods is information (i.e. the signals workers obtain) and the effect of peer pressure (which impacts effort only if today's project outcome matters for tomorrow's outcome). Effort depends on information through the sufficient statistic y. It depends on peer pressure because publicly observable past project outcomes affect the quality of current relation-

[7] This implies that two collaborators do not hide or falsify information. We discuss in Section 3.5 the implications if we relax this assumption and conjecture that our main results could be strengthened in this case.

[8] That is $e_i \in [0, e_{max}]$ where $f(e_{max}, e_{max}) < 1$. By choosing an appropriate bound on v_h, we can guarantee an interior solution $e \leq e_{max}$.

ships between workers, which in turn affects the success of collaboration. We describe the quality of the relationship by $\gamma \in \{\gamma_b, \gamma_g\}$, that is the relationship can be *b*ad or *g*ood. We outline this peer pressure channel here informally and defer the formal discussion to Appendix A:

Whether the relationship is good or bad depends on past outcomes, as a project failure leads to discord among project partners that negatively affects their friendship. We further argue that this discord between partners also spreads to common friends. This idea is based on the well-established *structural balance theory*: Triads of friends are only stable as long as the relationships are balanced. Suppose that i, j and l are all directly connected. Initially, all three relationships are intact. Then, i and j work on a project together that fails, affecting not only their link but rendering the entire triad unstable. This instability is resolved by the workers taking sides (here, l would side with i or j). To simplify our analysis, we assume that *all* relationships in a triad will turn bad after a project failure. Our assumption is a simplification of the following idea: When a project fails, a worker has a positive probability of ending up with *more than one* negative connection if he and the project partner had common friends. A project failure results in *only one* negative connection if the project failed with someone he does not have a common friend with. This is why project failures affect workers with high clustering more than those with low clustering: they are deprived of more future project opportunities. A relationship between i and j turns bad after a project failure if in the previous period either (1) i and j were teamed up or (2) i or j were teamed with a common friend.

In each period, a strategy of an agent maps his signals y and the relationship-status γ into an effort level, where we focus on pure strategies. Given that both the relationship-status and signals are observable for both team partners, our equilibrium notion is *perfect public equilibrium* (PPE).[9] This is a strategy profile that satisfies the usual requirements of being mutually best responses (Nash equilibrium) and sequentially rational. See Appendix A, for the formal definition of strategies and equilibrium (Definition 2).

3 Effort Choice, Wages & Network Structure

In our setting a higher degree leads to more signals, allowing for a more precise belief about the project value. Higher clustering, on the other hand, leads to a larger number of bad relationships after a project failure, which affects the success of future collaboration and therefore incentivises effort through peer pressure. This is the main trade-off we focus on. To flesh out how peer pressure influences effort choices, we focus on a dynamic setting.

3.1 Effort Choices & Wages

We first derive the effort choices and wages before analyzing how they are affected by an agent's network structure. In order to ease exposition, we begin with the static game and then extend it to the dynamic setting.

[9] A formal definition of this equilibrium concept is provided in Mailath and Samuelson (2006), p.231, Definition 7.1.3.

Static Game. In the static setting, worker i chooses effort to maximize his expected payoff,

$$\max_{e_i} f(e_i, e_j)\pi(y_i) + (1 - f(e_i, e_j))0 - c(e_i). \tag{3}$$

Recall that we have $y_i = y_j = y$ in any team where agents i and j are connected. Given our assumptions on $f(\cdot, \cdot)$ and $c(\cdot)$, the first order condition of (3) is both necessary and sufficient for a maximum. The problem is symmetric for worker j. Based on the first order approach, we determine the pure strategy public perfect equilibria of the static game and denote by $e(y)$ the optimal strategy based on signals y.

Lemma 1 (Static Game)**.**
1. Every public perfect equilibrium is symmetric: $e_i(y) = e_j(y) = e(y)\ \forall y$.
2. For each y, there exist exactly two pure public perfect equilibria.

(a) Zero effort:	$e(y) = 0$
(b) Strictly positive effort:	$e(y) > 0$

All proofs are collected in Appendix B. Given the symmetry of the problem, both workers exert the same effort in equilibrium of the static game. Moreover, there exist two pure strategy PPE. There always exists an equilibrium where both project partners exert zero effort independently of signal realisations. It is a best response to choose zero effort given the partner chooses zero effort as, by assumption, $f(e_i, 0) = f(0, e_j) = 0$. There also exists a PPE with strictly positive efforts. The uniqueness of the equilibrium with strictly positive effort follows from supermodularity and the constant returns to scale of $f(\cdot, \cdot)$, as well as the convexity of the cost function.

Dynamic Game. We now extend the static game by allowing each team partner to maximise his payoff with respect to effort across two periods. We assume that in period 1 each worker is in a good relationship with everyone he is connected to and thus omit the dependence on the relationship status. We further focus on a strategy profile where, for any realisation of signals, a worker puts strictly positive effort if the relationship to the project partner is good, and zero effort if it is bad.[10] Although clearly, there exist other equilibria in this model, in Appendix A (see *Equilibrium Selection*) we make a case for why this equilibrium is a reasonable one to focus on. As zero effort automatically leads to a project failure, payoffs for a team with a bad relationship are zero.

The dynamic problem of team partner i is then given by

$$\max_{e_i, e_i'} \quad f(e_i, e_j)\pi(y_i) + (1 - f(e_i, e_j))0 - c(e_i) + \beta s_i P_i(\gamma_g')\mathbb{E}\left[f(e_i', e_k')\pi(y_i') - c(e_i') \,\middle|\, \gamma_g'\right] \tag{4}$$

We index second period variables by *prime* and denote by β the discount factor. The expectation is taken with respect to the distribution of signals in period two, y_i'. By the same argument as

[10] Again, see Appendix A for the formal definition of these strategies.

in the static game, it is true that $y_i = y_j \equiv y$ and $y'_i = y'_k \equiv y'$.

The expected payoff in the second period depends on whether a worker is selected for the project, which occurs with probability s_i. If he is chosen, he can either be teamed with someone he has a good or someone he has a bad relationship with. The probability of a good relationship with future project partner k depends on past effort and network structure, and is given by

$$P_i(\gamma'_g) \equiv P(\gamma'_g|e_i, e_j, r_{ij}) = f\left(e_i, e_j\right) + (1 - r_{ij})(1 - f\left(e_i, e_j\right)), \tag{5}$$

where $r_{ij} = \frac{C_{ij}}{D_i}$ is the probability that the second period's team partners have a bad relationship after a first period failure between i and j, and where C_{ij} is a proxy for their common friends and thus for clustering.[11] Note that this probability is symmetric across first period's project partners, $r_{ij} = r_{ji}$. Thus, worker i has a good relationship with *all* his potential second period project partners only if his current project succeeds, which happens with probability $f(e_i, e_j)$. If it fails, then he only has a good relationship with his future partner, if this partner is not the same as the current one or a common friend, indicated by the joint probability $(1 - r_{ij})(1 - f\left(e_i, e_j\right))$.

We can now, based on the results from the static setting, turn to the dynamic problem in the first period, where not only the signals, but also considerations about the relationship state with *future* project partners matter.

We denote by $V_i^*(y')$ the maximised second period payoff when the relationship between i and k is good, that is $V_i^*(y') \equiv \max_{e'_i} f(e'_i, e'_k)\pi(y') - c(e'_i)$. Using this notation along with (4) and (5), the maximisation problem of agent i in the first period reads

$$\max_{e_i} \quad f(e_i, e_j)\pi(y) - c(e_i) + \beta s_i(f(e_i, e_j) + (1 - r_{ij})(1 - f(e_i, e_j)))\mathbb{E}[V_i^*(y')]. \tag{6}$$

Similar to the static problem, we show that there exists a unique PPE in which both team partners exert strictly positive effort. With some abuse of notation, we denote the optimal effort function in periods one and two by $e(y)$ and $e'(y')$, and omit the relationship-state γ as an argument, as effort is only strictly positive in case of a good relationship. We denote derivatives by subscripts, e.g. the first derivative of the cost function, which is the same in both periods, is denoted by $c_e(\cdot)$.

Proposition 1 (Dynamic Game)**.**

1. *Every PPE is symmetric:* $e_i(y) = e_j(y) = e(y)$ $\forall y$ *and* $e'_i(y') = e'_j(y') = e'(y')$ $\forall y'$.
2. *In both periods, there exists a unique PPE with strictly positive effort* $\forall y, y'$*, determined by*

$$c_e(e(y)) = f_1(e(y), e(y))(\pi(y) + \beta sr\mathbb{E}[V^*(y')]) \tag{7}$$

$$c_e(e'(y')) = f_1(e'(y'), e'(y'))\pi(y'). \tag{8}$$

Lemma 1 established that in the second period (which is identical to the static problem) there exists a unique equilibrium with strictly positive effort, which is symmetric across project part-

[11] Formally, $C_{ij} = 1 + \sum_{k, k \neq i, k \neq j} g_{ik} g_{jk}$ where $\sum_{k, k \neq i, k \neq j} g_{ik} g_{jk}$ gives the number of common friends of i and j. So, r_{ij} is the probability that in the second period, worker i is doing a project with someone who would be affected by a first period project failure, *given* that i and j are chosen for a project in the first period.

ners. This second period effort is implicitly defined by (8) in Proposition 1. The proposition further establishes that also in the first period, effort levels are symmetric. First, any two team workers have the same signals. Second, two workers must have the same number of common friends, $C_{ij} = C_{ji}$, and thus $s_i r_{ij} = s_j r_{ji} = sr$ (since $s_i r_{ij} = \frac{C_{ij}}{\frac{1}{2}(N-1)N}$). The effort does not depend on the selection probability s_i separately from r_{ij}. The intuition for this is that a worker is only punished when being paired with common friends or the same project partner and this is all that matters for adjusting effort. It is irrelevant whether a person is more likely to be selected again as the effort exerted does not increase the selection probability.

In what follows, we write $\beta sr\mathbb{E}[V^*(y')]$ for the expected second period value in the equation for first period effort (7). Note that in both periods effort increases in the contemporaneous project values, $\pi(y)$ or $\pi(y')$. But in the first period there is an additional factor at play, captured by $\beta sr\mathbb{E}[V^*(y')]$ in (7): the dynamic incentives of maintaining good relationships push first period effort up.

The equilibrium effort determines the first and second period wages (which we use interchangeably with *productivity*). We focus on wages for a given team member and conditional on the state, where we drop the subscript i when the wage is the same across team partners:

$$w(\theta) \equiv \mathbb{E}[f\left(e(y), e(y)\right) v|\theta] \tag{9}$$

$$w_i'(\theta, \theta') \equiv s_i P_i(\gamma_g'|e(y), r)\mathbb{E}[f\left(e'(y'), e'(y')\right) v'|\theta'], \tag{10}$$

where $\theta \in \{\theta_l, \theta_h\}$, $\theta' \in \{\theta_l', \theta_h'\}$ are the realised first and second period states and $v \in \{v_l, v_h\}$, $v' \in \{v_l', v_h'\}$ are the associated project values. The expectations are as usual taken over the signal realisations y and y'. We define these expected wages *given* that a certain state of the world has materialised. Recalling that q is the probability that the high state occurs, the expected wage across states can then be easily computed, e.g. $\mathbb{E}[w] = qw(\theta_h) + (1-q)w(\theta_l)$ for the first period.

Note that the structure of both periods' wages is the same in that agents obtain their share of output in case the project is successful. In the second period, however, one also has to take into account the joint probability of being selected and having a good friendship history with the project partner, given by $s_i P_i(\gamma_g'|e(y), r)$. Since friendship histories matter, the second period expected payoff not only depends on contemporaneous but also on first period effort. Both periods' wages are increasing in effort, highlighting the tight link between the agents' actions and their rewards.

3.2 Degree and Information

We now turn to the effect of information on effort and wages. All else equal, a worker with a higher degree receives more signals about the state of the world and thus more information. We want to know how effort varies with the number of signals and how this depends on the environment's underlying *uncertainty*.

Definition 1 (Uncertainty)**.**

*We call a setting uncertain if **all** of the following features are given:*

- *high and low project values differ, $v_l \neq v_h$*
- *signals are not completely informative $p \in \left(\frac{1}{2}, 1\right)$*
- *workers' prior about the state reflects some uncertainty $q \in (0,1)$*
- *overall information is bounded, $n_i < \infty$.*

In turn, by *vanishing uncertainty* we mean a situation in which any of the four requirements from Definition 1 is violated. We obtain the following result.

Proposition 2 (Degree, Effort & Wages)**.**
A higher degree leads to more information, which

1. *increases (decreases) expected second period effort iff the second period state is high (low).*
2. *increases expected first period effort if the first period state is high.*
3. *increases first and second period wages if the state in both periods is high.*

The impact of additional information on effort and wages in both periods vanishes as the underlying uncertainty vanishes.

Information impacts effort through the belief about the current project value: A high signal leads to a more optimistic belief and therefore to higher effort. Since signals are informative, the expected project value, $\mathbb{E}[\pi(y)]$ (with the expectation again taken over y), increases in the number of signals conditional on the realised state of the world being high $\theta = \theta_h$, and decreases in the number of signals conditional on the state being low, $\theta = \theta_l$. Therefore, the more signals are available due to a higher degree, that is the higher n_{int}, the more accurate is the worker's posterior belief about the state of the world. In the high state, he exerts on average higher effort compared to a worker with lower degree (where in the 'average' effort, the expectation is as usual taken over the sufficient statistic y). The opposite is true for the low state. As a result, in both periods $\mathbb{E}[e(y)|\theta_h] - \mathbb{E}[e(y)|\theta_l]$ is increasing in information. Intuitively, workers with more accurate information, i.e. more signals due to a higher degree, can better fine-tune their effort to the expected project reward. Based on this discussion, the second period effort increases in information if the state is high and decreases if the state is low (part 1.).

The first period effort (part 2.) does not only depend on the first period project value, but also on the expected second period payoff, see (7): A higher $\mathbb{E}[V^*(y')]$ translates into higher effort on average. We prove in Lemma 3 (Appendix B) that $\mathbb{E}[V^*(y')]$ is increasing in degree, that is the number of signals: Having more signals yields a more precise belief about the state and therefore allows each team to better adjust their efforts. Generally, being able to adjust effort optimally leads to higher payoffs, and this is why more signals lead to a higher value of the problem.

In sum, a higher *degree* improves information about the state of the world and is beneficial when the true state is high. In this case, additional signals induce the agents to put significantly more weight on the high state, translating into higher effort and project completion, and ultimately into higher productivity/wages (part 3.).

Notice that the effect of additional information on effort and thus wages is reinforced when the uncertainty of the underlying environment is considerable but dies out when uncertainty becomes small. The reason behind this result is that the expected project value becomes independent of the number of overall signals as uncertainty vanishes, that is if either (i) there is no difference between high and low project values; or (ii) signals are completely informative; or (iii) a worker's prior reflects complete certainty about the state of the world; or (iv) overall information due to an increase in the number of external signals becomes abundant. In any of these cases, an agent does not need to rely on his network to learn about the state of the world.

3.3 Clustering and Peer Pressure

We now analyse the effect of clustering on effort choices and wages. Clustering induces higher peer pressure, which attenuates the team moral hazard problem in the first period and thus affects first period effort and wages.

Proposition 3 (Clustering, Effort & Wages)**.**
Higher clustering increases peer pressure which leads to both higher expected first period effort and higher first period wages independently of the state of the world.

The effect of peer pressure (through *clustering* given by sr above) on first period expected effort is straightforward and unambiguously positive (see equation (7)). This channel is *independent* of both the true state of the world in period one and the underlying uncertainty. Peer pressure induces higher effort because a potential project failure today puts more friendships and thus future project opportunities in jeopardy. Since peer pressure works as a *dynamic* incentive, second period effort is unaffected by it. It then follows that peer pressure boosts the first period wage independently of the state and the underlying uncertainty. Only in the second period, the effect on wages is ambiguous: Peer pressure leads to higher first period effort (increasing $P(\gamma'_g|\theta)$ and thus pushing the second period wage up), but having many common friends also makes a non-intact relationship with the second period team partner more likely (lowering $P(\gamma'_g|\theta)$).

3.4 Peer Pressure versus Information

While the previous discussion has shown that clustering and degree impact effort in quite different ways, they can both have a positive effect on effort, depending on the level of uncertainty and the state of the world.

We now show that these two network characteristics are *complementary* as the effort and the wage in period one exhibit increasing differences in clustering and degree under the additional assumption that $c_{eee}(\cdot) \leq 0$.

Proposition 4 (Complementarity of Degree & Clustering)**.**
Let $c_{eee}(\cdot) \leq 0$. Then, higher peer pressure, i.e. more clustering, leads to a greater increase in expected first period effort and first period wage if the worker has more information, i.e. a higher degree.

More clustering has a greater positive impact on effort when the degree is high, i.e. when information is already abundant. This is equivalent to the degree having a higher impact on effort when there is already a high level of clustering. First-period effort depends on three factors, namely (i) the expected payoff in the first period, $\mathbb{E}\left[\pi(y)\right]$, (ii) the expected value in the second period, $\mathbb{E}\left[V^*(y')\right]$, and (iii) the probability of being punished for project failure, proxied by sr. A higher number of signals does not affect (i), the expected payoff in the first period, as it is a martingale (see Lemma 2 in Appendix B). However, an additional signal increases the value of the second period problem (ii). Particularly, this value increases the effort *proportionally* to the probability of being punished for failure (iii), since effort is affected through the expression $\beta sr\mathbb{E}\left[V^*(y')\right]$. This immediately implyies that information and clustering are complementary. Consequently, wages, which are a function of effort, also display complementarities in peer pressure and information. Our theory thus rationalizes why the different types of social capital that emerge from tight versus loose networks are complementary.

While the discussion so far has focused on comparative statics effects of a single network characteristic holding other network characteristics fixed, we now turn to the more interesting but also more involved case of comparing two types of workers: one with higher degree but lower clustering, called D-worker, and one with lower degree but more clustering, denoted as C-worker. This is an empirically relevant case as clustering and degree are generally negatively correlated, as is also evident in our datasets.[12] We are interested in their relative performance depending on the underlying uncertainty of the environment. We therefore define the notion of *comparative advantage* in our context: the C-worker holds a comparative advantage in environments with lower uncertainty if his relative expected wage, $\frac{\mathbb{E}[w^C]}{\mathbb{E}[w^D]} = \frac{qw^C(\theta_h)+(1-q)w^C(\theta_l)}{qw^D(\theta_h)+(1-q)w^D(\theta_l)}$, increases as uncertainty decreases.

Proposition 5 (Trade-Off Between Information and Peer Pressure)**.**
Assume the cost function is quadratic $c(e) = e^2/2$. Then:
1. Comparative Advantage: C-Workers hold a comparative advantage in environments with low uncertainty.
2. Wage Dynamics: If a C-worker has a weakly lower first period wage than a D-worker, then he also expects a lower wage in the second period.

The additional assumption imposed allows us to provide a closed form solution for effort, which simplifies our analysis significantly. Our model predicts that workers with higher clustering and lower degree have a *comparative advantage* in environments characterized by less uncertainty relative to workers with lower clustering and higher degree (part 1.). We establish that the ratio of expected wages, $\frac{\mathbb{E}[w^C]}{\mathbb{E}[w^D]}$ increases as the number of external signals n_{ext} grows, which captures an increase in the baseline information of a worker. We first show that there exists a number of external signals, such that an additional signal leads to a higher relative expected wage. We then establish that for a sufficiently high number of signals the comparative advantage will not be reversed, or put differently that this reversal is a probability zero event.

[12] See Appendix C, Tables 17, 18, 19.

The intuition is clear: Workers with higher degree obtain more information and thus have an advantage when information is valuable. But in environments with low uncertainty this is not the case and workers with higher clustering exert relatively more effort. This leads to higher wages for C-workers relative to D-workers, which underlines one of our key predictions: Clustering gains importance as uncertainty vanishes. Similarly, when the uncertainty is high, then additional information proves to be more valuable, giving an advantage to a D-worker.

Our model also predicts a strong impact of an early career wage gap on the future wage trajectory through peer pressure, which puts workers with high clustering but low information at a disadvantage (part 2.). As a result, if there is a wage gap in the first period, it persists even if they perform equally well in the second period (i.e. even if uncertainty vanishes in the second period). In expectation, a wage gap in the first period arises if and only if there is a difference in the exerted effort. Thus, a C-worker with lower first-period wage than a D-worker will have chosen lower effort. But a C-worker is more likely to be punished for inefficiently low effort that resulted in a project failure through missed opportunities in the second period. So, in expectation, the C-worker will also do worse in the second period and wage inequality persists. Moreover, second period wage gaps between C-workers and D-workers arise even if they exert the same effort in the first period, again through more forgone opportunities for the C-worker after a project failure.

3.5 Discussion Modelling Assumptions

Information Sharing. We assume throughout that two team partners have the same level of information. This implies that two collaborators share their information, although this may induce their co-worker to exert lower effort.[13] We abstract here from the possibility that agents hide or falsify signals. In our setting, successful collaborators work together repeatedly and have a good relationship which seems at odds with them omitting relevant information. However, even if information could be hidden or falsified by a team partner, then we conjecture this could strengthen our key result, namely that loose connections are more valuable than tighter ones under sufficient uncertainty. If agents could hide information or lie, then team partners would only rely on their own signal, signals from their direct neighbors (excluding their team partner) who have no incentive to lie, as well as signals from other external sources. Agents with loose networks and thus superior access to information could tailor their effort better to the expected project reward relative to their collaborator. This guarantees them a relatively high payoff in risky settings, increasing the output gap between two differentially informed agents even more.

Social versus Organizational Networks. We focus here on social networks (as opposed to organisational or formal ones) that are fixed (as opposed to endogenously formed). While firms try to optimise their organisational structure and hierarchy (i.e. the formal network) – a question at the heart of the literature on personnel economics (for an overview see Lazear and

[13] This would happen if someone with a high degree and thus more information knew that the quality of project was poor, but his partner with less information would exert a higher effort.

Oyer (2007)) – their ability to affect who gets along is much more limited. Thus, they are forced to take social networks as given, as fixed. However, these informal networks are essential to the success of a firm and may interact with the organisational structure. Understanding how they operate is therefore vital and can improve the operations of a firm. These social ties in a work setting are the focus of our analysis.

Exogenous Networks. We assume that networks are taken as given not only by firms but also by workers. Some workers simply get along and others do not. This is observationally equivalent to a random meeting process, which is the correct model for friendships according Jackson and Rogers (2007).[14] Rather than focusing on network origins, we analyze the impact of social networks on outcomes.

Note further that networks do not change across periods. We allow for two agents to (temporarily) have a bad relationship after a project failure. But even in this case, they are still connected – links are never cut.[15]

Team Composition. In our analysis, we focus on the network characteristics of an *individual*. However, it is worth noting that given full information sharing and the fact that team partners necessarily have the same number of common friends, our analysis carries over to network characteristics of the team. Suppose that the team is such that both workers have two signals. Then, the outcome is the same as when one worker had access to three signals and the other worker had access to one signal. Similarly, as clustering is given by the number of common friends, it is team specific and so we could interpret the network characteristics as features of the team.

External Signals. The notion of external signals is useful to vary the underlying uncertainty of the environment. The meaning of these external signals varies depending on the context of project. For example, in academia, external sources could be top publication records by field or topic which may influence whether to work on a topic. It could also be information about conferences or editors' tastes or rotations. In a law firm, this may be information on what kind of cases are likely to be 'career cases' due to their significance and reach while others would be insignificant. In a company, this may be information about its medium-run objectives and priorities of management, so that employees understand that projects advancing those objectives will receive more weight and are be thus more important for promotions.

[14] Jackson and Rogers (2007) aim to distinguish between two meeting processes, a random one and a process where agents meet through friends. They show empirically that a random process determines friendship networks.

[15] Note that links being cut, without links being added is not a sensible extension as at some point agents would run out of potential project partners.

4 Gender Differences in Network Structure and Labour Market Outcomes

Our framework shows that the impact of network structure on labour market outcomes (effort and wages/output) depends on the underlying uncertainty of the work environment. In this section, we offer an application of our theory to gender differences in labour market outcomes. First, we establish that in a variety of settings, networks of men and women differ: on average, men have a higher degree but lower clustering compared to women. Second, we provide evidence that the gender earnings gap is indeed larger in more uncertain occupations and moreover, that women are less likely to select themselves into riskier occupations. Finally, we aim to connect these two pieces and relate network structure to labour market outcomes. We provide correlations indicating that individuals with high clustering are less likely to select into risky occupations, potentially rationalising why women seem to have a comparative disadvantage in those occupations – in line with our theory.

4.1 Gender Differences in Network Structure

We document gender difference in network structure in three settings: we investigate co-authorship networks in academic computer science, email networks from the Enron company and social network patterns of high school students.

Network Measures. We begin with a formal definition of graphs that represent networks. A graph consists of a set of nodes N and a $n \times n$ matrix g, where g_{ij} represents the possibly directed relation between i and j. As we focus on unweighted graphs, g_{ij} equals either 0 or 1. For each node in the graph, we define two concepts that allow an assessment of agents' network structure: degree and clustering coefficient.

Degree. The degree is a measure of how connected an individual is. For a directed graph, there are three types of degree, in-degree (ID), out-degree (OD) and degree (D), denoted by

$$ID_i = \sum_j g_{ji}, \qquad OD_i = \sum_j g_{ij}, \qquad D_i = \sum_j \min\{1, g_{ij} + g_{ji}\}$$

The in-degree describes how many agents named or wrote an email to individual i. The out-degree provides information on how many agents individual i named or sent emails to. For an undirected network (like that of academic computer scientists), only the degree is defined, which in this case gives the number of coauthors.

Clustering Coefficient. The clustering coefficient is a measure of how close-knit or tight a network is. It is computed as the ratio of the actual number of links between a node's neighbors to the total possible number of links between the node's neighbors. This measure depends on whether the network is directed (superscript d) or undirected (superscript u):

$$CC_i^d = \frac{\sum_{j\neq i;k\neq j;k\neq i} g_{ij}g_{ik}g_{jk}}{D_i(D_i-1)}.$$

$$CC_i^u = \frac{2\sum_{j\neq i;k\neq j;k\neq i} g_{ij}g_{ik}g_{jk}}{D_i(D_i-1)}.$$

Co-Authorship Networks in Computer Science. We study collaboration networks of academic computer scientists. We obtain this data from the `dblp` computer science bibliography, a service providing open bibliographic information on all major computer science journal publications since 1995 (our sample includes academic papers from 1995-2016).[16]

The raw data set contains names of scholars and the names of co-authors for each publication listed on the platform, where based on publications we were able to extract 1,348,324 names.

After data cleaning, we are left with 585,360 unique names to whom we can assign a gender, based on their first names.[17] We have 438,531 men and 146,829 women in that sample and construct their co-authorship networks.

In the co-authorship network, nodes are authors and a link between two nodes exists if the corresponding authors have published at least one paper together. Note that this co-authorship network is an undirected network since collaborations are bilateral. Hence the network characteristics we compute are degree and undirected clustering coefficient. On average, a computer scientist has a degree of 7.7 (i.e. 7.7 coauthors) and a clustering coefficient of 0.15, see Table 1.

Our results show significant differences in collaboration networks across gender: While female computer scientists have a higher clustering coefficient, male computer scientists have a higher degree (Table 2). Further, these characteristics show a strong negative correlation (-0.46) in our sample of computer scientists (Table 17).

Email Networks at the Workplace. We also study networks in a private business, the Enron company. We reconstruct the network at Enron based on email communications that were made publicly available by the Federal Energy Regulatory Commission during its investigation of this company following its fraudulent bankruptcy.[18]

Our dataset contains about 400,000 emails, where we focus on a subset of emails that have *a single receiver* as group emails do not provide a good measure of whether employees have indeed a relationship. Doing so we obtain 26,298 distinct email addresses, either senders or receivers. One challenge is that 'gender' is not recorded. Fortunately, in many email addresses first and last name are separated, so we are able to extract the first name of the employee and assign a gender.[19] This procedure leaves us with 10,211 individuals whose gender we successfully

[16] See `http://dblp.uni-trier.de/`

[17] We remove duplicates and deal with outliers in the degree data by trimming the top and bottom 1% of the observations. Further, we use the package `Gender` in R to predict gender. The package provides a function to predict gender from names using historical data, a comprehensive description can be found via `https://cran.r-project.org/web/packages/gender/gender.pdf`.

[18] The Enron data is available at `http://www.cs.cmu.edu/~enron/`.

[19] We again use the package `Gender` in R to predict gender.

predicted.[20] Further, since our focus is on the network *at work*, we compute network measures for those individuals who have an "enron.com" email address. This reduces our final sample to 3,926 Enron employees with 1,628 women and 2,298 men, for whom we compute their network characteristics based on their email communication.[21]

Since emails are directed, we report both directed and undirected network characteristics. The summary statistics are in Table 3. Regarding gender differences, we find that women have a significantly higher clustering coefficient, both directed and undirected, and a significantly lower in- and overall degree compared to men (Table 4). The correlation between degree and clustering is again negative (Table 18). Thus, also in this very different context – a private business – we find that men and women have fundamentally different network structures, with women having smaller and denser networks.

Friendship Networks. We obtain the friendship networks from the AddHealth data set, which contains data on students in grades 7-12 from a nationally representative sample of roughly 140 US schools in 1994-95. Every student attending the sampled schools on the interview day is asked to compile a questionnaire (in-school data) on respondents' demographic and behavioural characteristics, education, family background and friendships. This sample contains information on 90,118 students. Students were asked to name up to 5 male and 5 female friends. The AddHealth website describes surveys and data in detail.[22]

The friendship network constructed from the AddHealth data is a directed network, based on friendship nominations. We compute both directed and undirected clustering coefficients as well as in-, out- and overall degree. We restrict attention to the individuals whose age and gender we can identify and those with an identifier. This leaves us with a dataset of 73,244 students. The descriptive statistics of the students are given in Table 5, those split by gender are provided in Table 6.

The results for the entire sample are given in Table 7, where we restrict the sample to those individuals for whom both degree and clustering are defined. We show that girls always have a higher clustering coefficient than boys across all ages. In turn, all measures of degree change with age. Younger girls have a higher degree than younger boys, whereas older boys (from age of 16 onwards) have a higher degree than older girls, which holds irrespective of our measure of degree.[23]

Further, we ensure that our results are not due to different shares of boys and girls at schools by restricting attention to schools where the share of boys and girls is balanced. This addresses the question of whether the differences in networks are driven by gender imbalances in

[20] While we lose a significant amount of individuals because their email addresses do not allow us to detect their gender, we think that it is unlikely that those missing employees are predominantly male or female.

[21] To address the issue of outliers we also trim here the top/bottom 1% of the degree data.

[22] For more details on the AddHealth data, see http://www.cpc.unc.edu/projects/addhealth.

[23] While there may be a concern that the out-degree is influenced by the constraint on friendship nominations, this is not a concern for in-degree. Moreover, if we take all out-degrees into account, the share of students who name 10 friends is limited to 13%.

the environment, which may matter if there is gender homophily.[24] Our results in the selected sample confirm that our findings are not due to differing gender shares (see Table 20).[25] Last, as a robustness check, we show that our results regarding the degree are not determined by restricting attention to those with fewer than two friends, see Table 21. As in our other datasets, degree and clustering are negatively related (see Table 19).

We thus find that clustering is unambiguously higher for girls, hinting at girls choosing denser and tighter networks. The number of friendships is much more sensitive to age, confirming the results of sociologists that do not find conclusive evidence for the number of nominated friends.[26] However, at older age, which is most relevant for the labour market, boys have larger networks in all of our specifications (see also Figure 1).[27]

In sum, across very different environments – academia, private sector and schools – we find that women network in smaller but tighter groups while men have larger but looser networks. To our best knowledge, this paper is the first to document gender differences in network structure, degree and clustering, in real world settings.

Gender Disparities in Networks: The Literature. Following up on our findings on gender differences in networks, Ductor et al. (2018) show that the described gender differences also emerge in Economics and Sociology. In economics, they use the Econlit database from 1970 until 2011, which comprises a large number of journals. In line with our finding, male economists have a higher number of distinct co-authors, that is a higher degree, whereas women's co-authors are more likely to be co-authors among each other, leading to a higher clustering coefficient. This finding is robust to including various controls, such as field, seniority, time trends and institution fixed effects.

In Sociology, based on data from the Sociological Abstract database from 1963 to 1999, they find the same patterns. Note that while men are strongly over-represented in Computer Science as well as Economics, Sociology is almost gender balanced in the 90s.

Despite this, the gender disparities found in Computer Science and Economics persist in Sociology, indicating once again that gender differences in networks are not driven by a gender imbalance in the environment.

While not measuring networks in the same way we do here, Friebel and Seabright (2011) and Friebel, Lalanne, Richter, Schwardmann, and Seabright (2017) also show in an experimental setting that men tend to have looser networks, whereas women's networks are tighter.[28]

[24] Gender homophily in referral networks has been shown in Beaman, Keleher, and Magruder (2018), Fernandez and Sosa (2005), Torres and Huffman (2002), Zeltzer (2020), Zhu (2018). Mengel (2020) shows gender homophily in networks in the lab.

[25] We further construct a measure for gender balance and use this as an additional control in our regression. Our results also hold for this alternative specification, see Table 1 in the Online Appendix.

[26] See for example Lee, Howes, and Chamberlain (2007), who show that girls have more friends than boys, Benenson (1990, 1993), Parker and Seal (1996) who show the opposite and Eder and Hallinan (1978), who find no conclusive evidence.

[27] Abstracting from network structure, David-Barrett, Rotkirch, Carney, Izquierdo, Krems, Townley, McDaniell, Byrne-Smith, and Dunbar (2015) show a preference for different types of networks across gender among adults.

[28] However, Mengel (2020) does not find gender differences in an experiment with very small networks.

This evidence along with our own leads us to conclude that there are significant differences in how men and women network, with men having a higher degree and women having a higher clustering coefficient. In this application of our theory, we want to raise the question whether these differences relate to differences in labour market outcomes across gender.

4.2 Gender Differences in Labour Market Outcomes and Uncertainty

Having established gender differences in networks, we now turn to gender differences in labour market outcomes and show that they are systematically related to the uncertainty of the environment. For this exercise and the measurement of uncertainty, we rely on data from the 2000 US Census as a large and representative sample of the US population, which matches the timing of Add Health.[29]

We first construct a measure of uncertainty at the occupational level, *occupational risk*, which for brevity we will refer to as risk in tables and graphs. We measure occupational risk by the standard deviation of residual earnings by occupation in a Mincer-type wage regression where we regress individual log yearly earnings on commonly used observable characteristics.[30] The standard deviation of the residual earnings in a given occupation is a measure for wage variation that is associated with this occupation and *cannot* be predicted based on commonly used observable controls. As such, this measure of unpredicted wage risk is closely related to the uncertainty in our model. In Figure 2(a), we plot this measure of occupational risk for 21 broad occupation groups.[31]

While occupations like legal occupations or management are high risk occupations, occupations in education or health support are considered low risk based on our measure. In Figure 2(b), we plot occupational risk against occupational mean earnings, which reveals an intuitive risk-return trade-off.

We then establish a connection between the gender earnings gap and occupational risk. We use the following regression to assess this relationship:

$$LogEarnings_{ij} = \beta_0 + \beta_1 Risk_j + \beta_2 Female_i + \beta_3 (Risk_j \times Female_i) + \mathbf{x_i}^T \gamma + \epsilon_{ij} \tag{11}$$

We regress log yearly earnings of individual i in occupation j in 2000 on occupational risk, $Risk_{ij}$, an indicator for female, $Female_i$, and their interaction (and some standard individual-level controls $\mathbf{x}_i$). The results are presented in Table 8. Column (1) is the baseline where we do *not* control for occupational risk. The gender earnings gap is 36%. In column (2), where we additionally control for occupational risk and the interaction between occupational risk and female, the gender gap considerably shrinks (by 17 percentage points or almost by 50%). Moreover, and important for this application, women's earnings disadvantage is particularly

[29] To be precise, the year 2000 matches the time frame of wave III of AddHealth that we focus on in the next section.

[30] These include education, work experience, race, gender, occupation and industry dummies, where we focus on full-time workers.

[31] These groups correspond to occupational categories available in AddHealth, wave III, see next Section.

pronounced in risky occupations, as indicated by the significant negative interaction between female and risk. Women in the least risky occupation (Administrative occupations, where $Risk = 0.44$) earn 33% less than their male counterparts (-0.19+(-0.31)*0.44=-0.33). In turn, women in the most risky occupations (Legal occupations, with $Risk = 0.72$) earn 41% less than their male counterparts (-0.19+(-0.31)*0.72=-0.41). Thus, the gender earnings gap is around 24% larger in the most risky compared to the least risky occupation.

Based on the results from the earnings regression, we expect women to be less likely to select into riskier occupations if they understand their earnings disadvantage in those settings. This is indeed what we find. We run an *ordered probit model*, which is a choice model where the occupational alternatives are ordered by risk. We regress the riskiness of occupation j in which individual i is employed, $Risk_{ij}$, on an indicator for female, $Female_i$ and some standard individual-level controls and occupational characteristics:

$$Risk_{ij} = \beta_0 + \beta_1 Female_i + \mathbf{x_{ij}}^T \gamma + \epsilon_{ij}. \tag{12}$$

Our main explanatory variable of interest is $Female_i$. We find that women are *less* likely to be employed in riskier occupations (Table 9).

Our results suggest that women do not thrive in the way that men do in risky environments, reflected by lower earnings in risky occupations and self-selection into safer environments, suggesting they have a comparative advantage in safer settings.[32]

4.3 Network Structure and Labour Market Outcomes

We now show a correlation between our two findings – gender differences in network structure and gender disparities in labour market outcomes that are particularly pronounced in risky occupations. Larger gender differences in risky occupations are related to gender differences in networks, suggesting that women's networks constitute a hindrance in riskier settings. We conduct this exercise for AddHealth, where we relate network patterns and labour market outcomes as well as for Computer Science, linking co-authorship networks to research output.[33]

Network Structure and Labour Market Outcomes: AddHealth. The network measures are collected in Wave I in AddHealth (1994/1995) when the respondents were still in high school. Table 7 demonstrates that the degree is changing across age and so we focus on the first wave in AddHealth that contains labour market information, wave 3 (with data collection in 2000/2001), to relate network structure to labour market outcomes. This way we minimise the time gap between the assessment of network characteristics and the assessment of labour market outcomes.

Wave III contains information on a subsample of 15,170 individuals that were already interviewed in wave I. We drop those with missing variables, that is those with missing network variables or occupations, as well as unemployed individuals. The resulting sample contains

[32] We conduct several robustness checks, including industry and occupation fixed effects, in the Online Appendix.

[33] The Enron e-mail data does not contain information on performance on the job.

around 4,000 individuals.[34]

We are interested in whether one of the key predictions of our model (Proposition 5.1) finds support in the data:

Prediction: Workers with a relatively higher clustering coefficient and lower degree have a comparative advantage in less risky environments.

We assess this hypothesis by investigating occupational choices, which should reflect individuals' comparative advantage in risky/non-risky work settings. This allows us to assess the value of our theory more cleanly compared to wages. The key finding of the literature on referral networks is that a higher degree leads to higher wages as agents obtain more job offers, see for example Calvó-Armengol and Jackson (2004, 2007). Our data does not allow us to disentangle the effects of referral networks and networks on-the-job on wages. Therefore, we focus on occupational choice, which the literature on referral networks is silent on.[35]

Our model predicts that workers with a relatively high clustering coefficient and/or relatively low degree should be more likely to select into less risky occupations. To provide support this prediction, we run an *ordered probit model.* We regress the riskiness of the occupation j that was chosen by an individual i, $Risk_{ij}$, on his/her network characteristics $Clustering_i$ and $Degree_i$,

$$Risk_{ij} = \beta_0 + \beta_1 Clustering_i + \beta_2 Degree_i + \mathbf{x_{ij}}^T \gamma + \epsilon_{ij} \tag{13}$$

where $\mathbf{x_{ij}}$ is a vector of individual and occupational controls. Our main explanatory variables of interest are $Clustering_i$ and $Degree_i$. Based on our theory, we expect that β_1 is negative (higher clustering leads to choosing less risky occupations) while β_2 is positive (higher degree encourages choosing riskier occupations). We find that a higher *clustering coefficient* is *negatively* related with selecting a riskier occupations, while *degree* does not significantly correlate with the occupational risk choices of individuals, see Table 10. Thus, there is a sense in which individuals self-select into the occupation that caters to their network characteristics. This finding is further illustrated using the marginal effects corresponding to the ordered probit regression for one low risk occupation, administrative support, and one high risk occupation, management, in Table 11. Column (1) shows that that a higher clustering coefficient is associated with a higher probability of choosing admin support as an occupation. In contrast, column (2) demonstrates that a higher clustering coefficient is related to a lower probability of selecting into a management occupation. This finding may help explain why women perform poorly in risky environments and avoid them, i.e. could help explain our findings from the US Census: This is potentially driven by women's higher clustering coefficient that is disadvantageous in risky settings.

[34] In more detail: We are able to merge around 13,000 individuals from Wave 3 with their network characteristics from Wave I. Focusing on those with valid occupational identifier (so that we can merge in our wage risk variable) leaves us with around 9,000 individuals. Dropping those with missing earnings results in 6,100 individuals. Further imposing our sample restrictions – our age restriction ($>$ 21 years old, so individuals in our sample are in the age range 21-27) and keeping only those individuals who work at least part time and earn at least half of the minimum wage income – leaves us with a final sample of around 4,000 individuals.

[35] Additionally, the earnings/wage variables in AddHealth are of rather poor quality.

Network Structure and Research Output: Computer Science. Research is characterised by complex and, especially, uncertain tasks. The success of a research project and patents is difficult to foresee at the time of production. Moreover, there is a considerable amount of uncertainty stemming from the lack of job security before tenure. We therefore view research and specifically computer science as intrinsically risky. In light of our theory (Proposition 5.1), we predict that higher degree leads to a higher research output. On the other hand, a higher clustering coefficient reduces research output.

Prediction: A higher degree increases research output, a higher clustering coefficient reduces it.

To evaluate our prediction, we require a measure of research output. A natural starting point are citations and number of papers, where the number of papers can be proxied by Google's i10-index, which counts the number of papers with more then 10 citations.[36] A more sophisticated measure is the h-index, developed in Hirsch (2005). The h-index equals the number h when the scholar has published h papers each of which has been cited in other papers at least h times.[37] It was developed as an improvement over the simpler measures of citations and number of papers to take into account both research quantity *and* quality. As such it may be the preferred measure.

In order to obtain these output measures, we obtain the Google Scholar profiles of a randomly selected subsample of the co-authors obtained from the `dblp` computer science bibliography, which results in a sample of 25 428 computer scientists.[38] This subsample has the same average clustering coefficient but a higher average degree than the full sample (potentially due to the fact that only more productive researchers have a google scholar profile). Regarding the purely quantitative output measures, citations and i10-index, women score better than men in our sample. In turn, men have a higher h-index. One interpretation is that women tend to have more publications of little impact compared to men (captured by a higher i10-index). We report the summary statistics of the Google Scholar sample in the Online Appendix.

To evaluate the prediction stemming from our theory, we analyse whether high degree and low clustering is associated with better performance in this risky environment. We regress each of the three performance measures on these network characteristics, controlling for gender. The results are reported in Table 14: For a given clustering coefficient, an increase of degree by one is associated with an *increase* in the h-index by 0.28, in the i10-index by 1.09 and in citations by 64. In turn, for a given degree, increasing the clustering coefficient from 0 to 1 is related to a *decrease* in the h-index by 9, in the i10-index by 61 and in citations by around 2000. These results lend support to our theory that a high degree as opposed to high clustering is advantageous in work environments that are characterised by high uncertainty.

Given our reported finding that computer scientists show large disparities in networks across gender, with male scientists having higher degrees but female scientists having higher clustering

[36] This is a low bar in computer science, see Table 12, as the average citation is above 6000.

[37] To give an example, suppose a researcher has 5 publications, A, B, C, D, and E. The citation count is $c(A) = 10$, $c(B) = 8$, $c(C) = 5$, $c(D) = 4$, $c(E) = 3$. This researcher has an h-index of 4. If the citation count was instead $c(A) = 10$, $c(B) = 8$, $c(C) = 5$, $c(D) = 3$, $c(E) = 3$, then the researcher would have an h-index of 3.

[38] We restrict attention to a reduced number of computer scientists due to the time consuming nature of the scraping process, which was done by a standard web crawler using Python.

coefficients, we can also ask whether these differences in networks are related to the gender productivity gap. To do so we compare the results from Table 14 that controls for both gender and network characteristics with a regression that omits the network characteristics, Table 15.

Regarding the h-index, comparing columns (1) across tables, we observe that controlling for network characteristics is related to a substantial decrease in the gender gap in performance (by around 25%). Similarly, when focussing on the i10-index or the number of citations, controlling for network characteristics is associated with even better female performance (compare columns (2) and (3) across tables).

Our results suggests that if female computer scientists did not have their disadvantageous network characteristics they might perform better according to our measures of performance. These findings from Computer Science complement our analysis in the US Census on the relation between risk and female labour market performance.

Network Structure, Gender and Labour Market Outcomes: The Literature. To substantiate our findings, we also turn to the literature, some of which has more suitable data to test our theory. There are several bits of evidence in the literature that corroborate our theory and findings above. First, we supplement our suggestive evidence regarding computer science with other findings for researchers in academia. We would expect, as with computer science, that a higher degree increases output, while a higher clustering coefficient reduces it. This is indeed what the the literature finds.

Ductor, Fafchamps, Goyal, and van der Leij (2014) show, without distinguishing or controlling for gender, that in Economics a higher degree is related to higher research output, while higher clustering decreases it, as predicted by our theory. Their measure takes into account the number of publications, the impact factor of the journal the article was published in as well as the number of co-authors on the paper. Ductor et al. (2018) also control for gender and show that women have a lower research output according to the described measures. The gender discrepancy in performance also holds in terms of the numbers of published papers or citations. This finding is robust to controlling for field, institution, seniority and time trends. The output gap between men and women is also affected by the network. Importantly, having a higher degree helps close the gender productivity gap in Economics, while a higher clustering coefficient exacerbates it even further (the magnitude of the gap closure by network characteristics is 5 -10%), see Table 6 in their paper. An Oaxaca-Blinder decomposition of the gender output gap confirms these effects.

Another sector where gender inequalities persist is the film industry (Lutter (2012) and Lutter (2013)), where women create lower box revenues from movies. This industry is highly project-based where tasks have uncertain outcomes. Ferriani, Cattani, and Baden-Fuller (2009) argue that the film market requires fast adjustment to new work environments since film ventures operate under constant uncertainty and have to foresee ex-ante whether the project opportunity is valuable. They argue that information is crucial to identify potentially successful scripts and to assemble the right project team. Based on the finding that producers who are more central in their network (i.e., have more access to information) are more likely to increase the box revenue from a movie, the authors conclude that social networks provide crucial access to information. In a similar vein, Lutter (2013) documents that women with loose information-based networks perform better in the film-industry than women with dense networks, supporting our hypothesis that information is the key to success in uncertain environments.

Another well-known area for gender disparities is the market for patents. Hunt, Garant, Herman, and Munroe (2012) document that women in the US are much less likely to be granted a patent than men, with women holding only 5.5% of commercialized patents. Gabbay and Zuckerman (1998) document that in basic research, which is typically characterised by complex, uncertain tasks, scientists benefit from sparse networks with many holes, whereas in applied research, which is typically characterised by non-complex, certain tasks, scientists benefit from dense networks. In line with this view, Ding, Murray, and Stuart (2006) argue that an important reason for the gender wage gap in patenting is that women's networks are less effective: In relying more on close relationships, they lack access to industry contacts.

All of this suggests that women's network characteristics hold them back in occupations that are characterised by uncertainty. We view these results of how differences in gender networks may account for productivity gaps and differential occupational choices an interesting implication of our theory.

5 Conclusion

We develop a novel theory that sheds light on the relative advantages of having a loose versus a tight social network at work. A loose network is particularly beneficial in an uncertain environment as it allows greater access to information. In turn, a tighter network generates peer pressure which leads workers to exert more effort, independently of the environment. These effects induce individuals with low clustering and high degree to have a comparative advantage in risky work settings where information is crucial.

We apply our theory to improve our understanding of the gender wage gap, which is particularly pronounced in high uncertainty settings, where we measure uncertainty by earnings risk on the occupational level. We first document a novel fact that male and female networks differ. On average, men have a higher degree and lower clustering coefficient, resulting in a looser network compared to women. Second, we show that women perform particularly poorly relative to men in high risk occupations. Finally, we connect the differences in network structure to differences in labour market outcomes, suggesting that tight networks are indeed more beneficial in low risk occupations. We argue that network differences across gender *at work* potentially are an overlooked source of well-known gender gaps in the labour market, especially in risky environments where women perform particularly poorly.

Table 1: Summary Statistics: Computer Scientists

	Min	Max	Mean	Std
Degree	2.00	67.00	7.65	8.74
Clustering Coefficient	0.00	0.25	0.15	0.06
Observations	585,360			

Note: Sample consists of 438,531 men and 146,829 women. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Table 2: Network Measures by Gender: Computer Science

	Male	Female	Difference
Degree	7.8810	6.9551	0.9259***
			(37.5695)
Clustering Coefficient	0.1511	0.1608	-0.0098***
			(-51.4112)
Observations	585,360		

t-statistics in parenthesis. ***p<0.01, **p<0.05, *p<0.1.
Note: Sample consists of 438,531 men and 146,829 women. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Table 3: Summary Statistics: Enron

	Min	Max	Mean	Std
Outdegree	0.00	157.00	5.87	13.96
Indegree	0.00	124.00	6.13	11.70
Degree	2.00	203.00	9.68	19.15
Clustering Coefficient Undirected	0.00	1.00	0.33	0.33
Clustering Coefficient Directed	0.00	1.00	0.23	0.27
Observations	3,926			

Note: Sample consists of 3,926 individuals, 1,628 women and 2,298 men. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Table 4: Network Measures by Gender: Enron

	Male	Female	Difference
Outdegree	5.9852	5.7021	0.2831
			(0.6270)
Indegree	6.4904	5.6130	0.8774***
			(2.4299)
Degree	10.0461	9.1566	0.8895*
			(1.4687)
Clustering Coefficient Undirected	0.3236	0.3500	-0.0264***
			(-2.4635)
Clustering Coefficient Directed	0.2241	0.2462	-0.0221***
			(-2.4645)
Observations		3,926	

Note: t-values in parenthesis. ***p<0.01, **p<0.05, *p<0.1.
The statistics are based on our sample of 3,926 observations, with 1,628 women and 2,298 men. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Table 5: Summary Statistics: Add Health

	Mean	Std. Dev.	Min	Max
Cl. Coeff. Dir.	0.124	0.143	0	1
Cl. Coeff. Undir.	0.182	0.184	0	1
Degree	8.306	4.305	2	39
In-Degree	4.496	3.642	0	37
Out-Degree	5.514	3.197	0	10
Age	14.952	1.708	10	19
Gender	0.515	0.5	0	1
Observations		73244		

Note: Network measures by gender, individuals that cannot be uniquely identified are omitted. For the number of men versus women, see Table 6. For gender, zero denotes men, one denotes women.

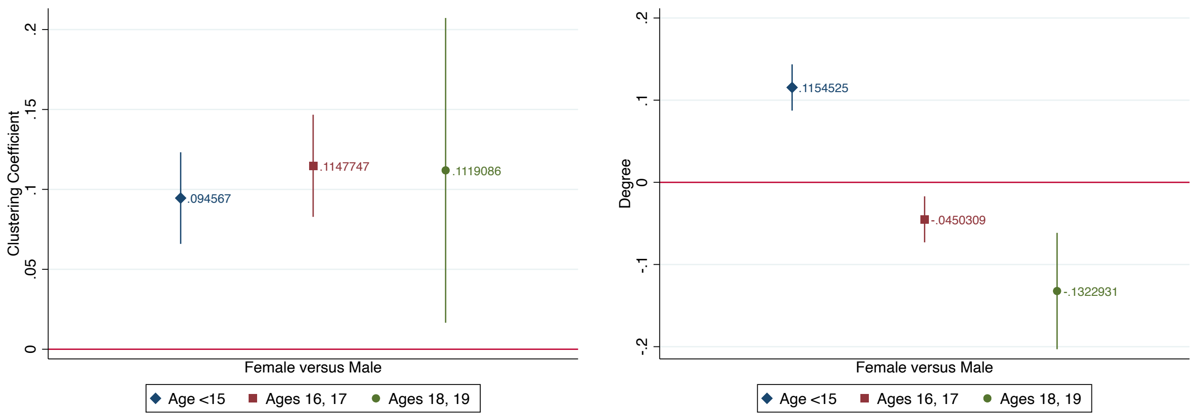
Table 6: Network Measures and Age by Gender: Add Health

	Male Students				Female Students				Difference
	Mean	Std Dev.	Min	Max	Mean	Std Dev.	Min	Max	t-test
Cl. Coeff. dir.	0.125	0.151	0	1	0.123	0.135	0	1	0.00251*
Cl. Coeff. undir.	0.175	0.187	0	1	0.188	0.182	0	1	-0.0124***
Degree	8.166	4.43	2	39	8.436	4.179	2	37	-0.272***
In-Degree	4.396	3.719	0	37	4.589	3.566	0	34	-0.194***
Out-Degree	5.252	3.331	0	10	5.761	3.044	0	10	-0.509***
Age	15.04	1.716	10	19	14.869	1.696	10	19	0.171***
Observations		35506				37738			

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Network measures by gender; individuals that cannot be uniquely identified are omitted.

Figure 1: Clustering Coefficient and Degree Across Ages



Note: The figure plots the coefficients on the indicator female of a regression of dependent variable on constant, age, school fixed effects and the female dummy for ages 10-15, 16/17 and 18/19, respectively. Dot gives the coefficient, line depicts 99% confidence interval. The graphs for the other network measures are given in the appendix.

Table 7: Differences in Degree & Clustering for Boys and Girls: Entire Sample Add Health

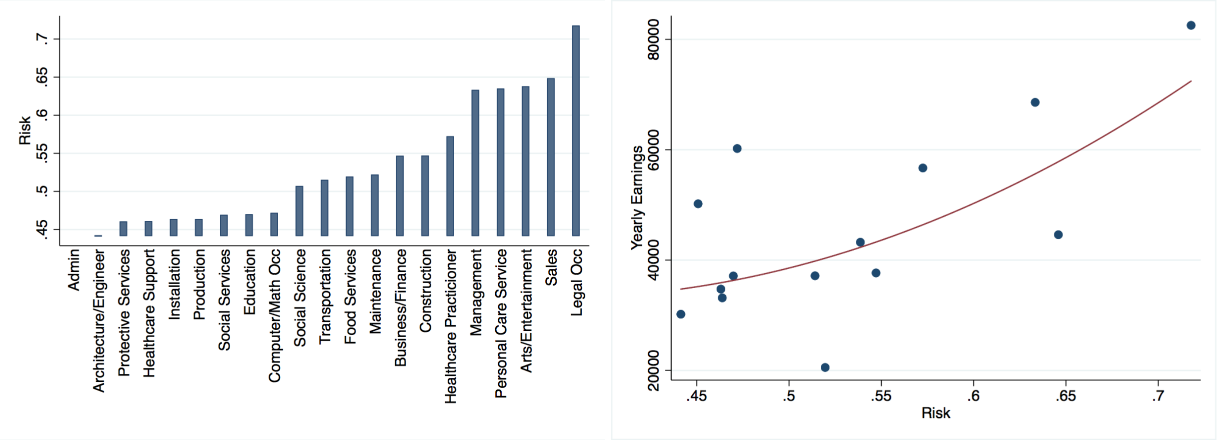
	Cl. Coeff. (dir)	Cl. Coeff. (dir)	Cl. Coeff.	Cl. Coeff.	Degree	Degree	In Degree	In Degree	Out Degree	Out Degree
Female	0.000190	0.00772	0.0985***	0.0923***	0.0434***	0.106***	0.0536***	0.114***	0.135***	0.164***
	(0.00735)	(0.00924)	(0.00718)	(0.00903)	(0.00663)	(0.00868)	(0.00724)	(0.00954)	(0.00683)	(0.00863)
Age	0.0102**		0.0115***		-0.0345***		-0.00679*		-0.0448***	
	(0.00315)		(0.00307)		(0.00266)		(0.00289)		(0.00279)	
Age 16-17		0.0347**		0.0424***		0.000560		0.0633***		-0.0475***
		(0.0126)		(0.0119)		(0.0107)		(0.0117)		(0.0114)
Age 18-19		0.0532*		0.0308		-0.113***		0.00486		-0.213***
		(0.0260)		(0.0236)		(0.0204)		(0.0216)		(0.0223)
Female*Age 16-17		-0.0248		0.0127		-0.148***		-0.137***		-0.0692***
		(0.0155)		(0.0154)		(0.0137)		(0.0149)		(0.0143)
Female*Age 18-19		0.0191		0.0254		-0.223***		-0.274***		-0.0724*
		(0.0395)		(0.0371)		(0.0282)		(0.0291)		(0.0319)
SCHOOL FIXED EFFECTS INCLUDED										
Observations	73244	73244	73244	73244	73244	73244	73244	73244	73244	73244
R^2	0.038	0.038	0.067	0.067	0.100	0.102	0.070	0.073	0.077	0.078

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Network characteristics are standardized and can be interpreted in terms of standard deviations, all regressions include school fixed effects. We omit students who have fewer than two friends as the clustering coefficient is not defined for them.

Figure 2: (a) Measured Risk by Occupation; (b) Occupational Risk and Earnings



Note: Line in (b) based on regression of yearly earnings on a constant, risk and risk squared.

Table 8: Log Earnings, Risk and Gender

	(1) Log Earnings	(2) Log Earnings
Female	-0.363***	-0.192***
	(0.000617)	(0.00428)
Risk		1.057***
		(0.00608)
Female*Risk		-0.310***
		(0.00836)
Years of Educ	0.115***	0.110***
	(0.000161)	(0.000159)
Experience	0.0274***	0.0275***
	(0.000136)	(0.000135)
Experience^2	-0.000383***	-0.000387***
	(0.00000277)	(0.00000274)
Black	-0.0672***	-0.0586***
	(0.00142)	(0.00142)
White	0.0555***	0.0508***
	(0.00111)	(0.00111)
Constant	8.605***	8.120***
	(0.00291)	(0.00411)
Observations	3558758	3558758
R^2	0.246	0.257

Sample: 2000 US Census, Full-time workers. Estimation by OLS.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Occupational Risk and Gender

	(1) Risk
Female	-0.166***
	(0.00111)
Years of Educ	-0.0164***
	(0.000244)
Experience	-0.00722***
	(0.000234)
Experience^2	0.000124***
	(0.00000459)
Black	-0.0619***
	(0.00256)
White	-0.0138***
	(0.00192)
Log Occ Mean Earnings	1.625***
	(0.00238)
Observations	3558758
Pseudo R^2	0.044

Sample: 2000 US Census, Full-time workers. Estimation method: ordered probit.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Occupational Risk and Networks

	(1) Risk
Clustering Coefficient	-0.0325**
	(0.0165)
Degree	-0.0114
	(0.0171)
Female	-0.0690**
	(0.0331)
Age	-0.517
	(0.447)
Age^2	0.00970
	(0.00976)
Log Occ Mean Earnings	0.987***
	(0.0478)
Race Dummies	Yes
Education Dummies	Yes
Work Exper 1999	Yes
Work Exper 2000	Yes
Observations	3976
Pseudo R^2	0.021

Standard errors in parentheses
Sample: AddHealth Wave III. Focus is on individuals who work at least part-time, who earn at least half of the minimum wage, $\geq$ 21 years old.
We include indicators for 3 education groups: $<$ high school, high school, $>$ high school.
Estimation method: ordered probit.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Occupational Risk and Networks (Marginal Effects)

	(1) Risk	(2) Risk
Clustering Coefficient	0.00571**	-0.00129*
	(0.00291)	(0.000660)
Degree	0.00200	-0.000450
	(0.00301)	(0.000679)
Female (d)	0.0121**	-0.00273**
	(0.00583)	(0.00132)
Age	0.0910	-0.0205
	(0.0785)	(0.0178)
Age^2	-0.00171	0.000384
	(0.00171)	(0.000389)
Log Occ Mean Earnings	-0.174***	0.0391***
	(0.00981)	(0.00333)
Race Dummies	Yes	Yes
Education Dummies	Yes	Yes
Work Exper 1999	Yes	Yes
Work Exper 2000	Yes	Yes
Observations	3976	3976

Marginal effects; Standard errors in parentheses
Sample: AddHealth Wave III. Focus is on individuals who work at least part-time,
who earn at least half of the minimum wage, $\geq$ 21 years old.
Column (1) reports the marginal effects corresponding to a low risk occupation (admin support).
Column (2) reports the marginal effects corresponding to a low high occupation (management).
The prob of choosing Management Occupations=0.054; the prob of choosing Admin Occupations=0.1 .
(d) for discrete change of dummy variable from 0 to 1
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Performance Measures: Computer Science (Google Scholar Sample)

	Min	Max	Mean	Std
Citations	0.00	679167.00	6135.92	21134.20
h-Index	0.00	262.00	20.84	24.74
i10-Index	0.00	1973.00	59.56	139.61

Note: Sample consists of 15,827 men and 9,601 women. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Table 13: Performance Measures by Gender: Computer Science (Google Scholar Sample)

	Male	Female	Difference
Citations	5925.6335	6482.5707	-556.9371**
			(-1.9417)
h-Index	21.4645	19.8201	1.6443***
			(5.1499)
i10-Index	58.3324	61.5721	-3.2397**
			(-1.7057)
Observations		25,428	

t-statistics in parenthesis. ***p<0.01, **p<0.05, *p<0.1.
Note: Sample consists of 15,827 men and 9,601 women. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Table 14: Performance of Computer Scientists: Network Characteristics and Gender

	Dependent variable:		
	h-Index	i10-Index	Citations
	(1)	(2)	(3)
Degree	0.28000***	1.09269***	63.81948***
	(0.01441)	(0.08180)	(12.45628)
Clustering.Coefficient	−8.73374***	−61.11535***	−1,989.62000
	(2.61046)	(14.82026)	(2,256.90300)
Female	−1.23644***	5.01811***	649.90050**
	(0.31642)	(1.79639)	(273.56360)
Constant	19.64833***	54.86391***	5,511.55300***
	(0.49005)	(2.78215)	(423.68020)
Observations	25,428	25,428	25,428
R^2	0.02538	0.01378	0.00190

*p<0.1; **p<0.05; ***p<0.01
Sample: Google Scholar Sample.

Table 15: Performance of Computer Scientists by Gender

	Dependent variable:		
	h-Index	i10-Index	Citations
	(1)	(2)	(3)
Female	−1.64434***	3.23972*	556.93710**
	(0.31985)	(1.80597)	(273.37420)
Constant	21.46446***	58.33241***	5,925.63400***
	(0.19654)	(1.10972)	(167.98080)
Observations	25,428	25,428	25,428
R^2	0.00104	0.00013	0.00016

*p<0.1; **p<0.05; ***p<0.01
Sample: Google Scholar Sample.

Appendix A: Omitted Details of Model and Derivations

Derivation of s_i

The probability that one agent is chosen is given by $P(K) = \frac{N-1}{\frac{1}{2}N(N-1)} = \frac{2}{N}$, and the probability that this agent i is linked to the suggested project partner j, given that he is selected by $P(g_{ij} = 1|K) = \frac{D_i}{N-1}$. Then, the probability of being chosen *and* being partnered with a friend is

$$s_i \equiv P(g_{ij} = 1 \wedge K) = P(g_{ij} = 1|K)P(K) = \frac{2D_i}{N(N-1)}.$$

Peer Pressure and Relationship Quality

We outline here formally how a project outcome affects the relationships of workers. As mentioned previously, whether the project of workers i and j was a success, S, or a failure, F is publicly observable and denoted by $\omega \in \Omega = \{S, F\} \times \{1, 2, \dots, N\}^2$. As an example, if $\omega = S12$, this means that a project was successfully completed by workers 1 and 2. We condition also on the workers who carried out the project as we do not only care about whether the project was successful but also about the workers who were involved. Each project failure induces some bad relationships in the network g. The network that contains the links that signify a bad relationship is denoted by $g_b \subset g$. The specific network g_b that arises after Fij, that is a project failure between workers i and j, where $g_{ij} = 1$, is given by $g_b(Fij) = \{\{ij, il, jl\}|g_{il} = 1 \wedge g_{jl} = 1, \forall l\}$. Workers i and j have a bad relationship with each other if their joint project fails. But a worker l, who is connected to both i and j also has a bad relationship with both of them. Denote by $g_g(Fij) = g \backslash g_b(Fij)$ the good relationships in the network g. Let $\gamma_g \in g_g$ and $\gamma_b \in g_b$. Further, for any i, j $g_g(Sij) = g$.

Perfect Public Equilibrium

The relationship quality between two directly connected workers constitutes a state, $\gamma \in \Gamma = \{\gamma_g, \gamma_b\}$. Also, recall the publicly observed signals $y \in Y$.

We can define a pure public strategy $\sigma : \Gamma \times Y \rightarrow E$, which maps from the relationship state and the signals into the action space.

Due to our restriction to public strategies, the equilibrium concept applied is that of a public perfect equilibrium, which is a sequential equilibrium with the further restriction that agents only condition on publicly observable outcomes, but not on privately observed actions. This implies we are not allowing agents to condition on their own effort, but only on whether projects failed or not. We index the variables in the second period by *prime*.

Definition 2. *A public perfect equilibrium (PPE) is a profile of public strategies σ that for any state $\gamma, \gamma' \in \Gamma$ and for any signal realization $y, y' \in Y$ specifies a Nash equilibrium for the repeated game, i.e. in the first period, $\sigma(\gamma_g, y)$ is a Nash equilibrium and in the second period $\sigma'(\gamma', y')$ is a Nash equilibrium.*

We restrict attention to the strategies according to which agents exert high effort if the relationship to the project partner is good and zero effort otherwise. This implies for period one and period two strategies:

$$\begin{aligned}
&\text{Period 1:} \quad \forall\ y \quad \sigma(\gamma_g, y) > 0 \\
&\text{Period 2:} \quad \forall\ y' \quad \sigma'(\gamma_g', y') > 0 \qquad \text{and} \qquad \sigma'(\gamma_b', y') = 0.
\end{aligned}$$

To simplify notation in the main text, we there denote the first and second period strategies by $e(y)$ and $e(y')$ (instead of $\sigma(\gamma_g, y)$ and $\sigma(\gamma_g', y')$), where we omit the relationship state as an argument. Note that a PPE is a sequential equilibrium

Equilibrium Selection

In our analysis, we have selected the equilibrium that induces workers to play high effort if their relationship is good and zero effort if their relationship is bad. Alternatively, agents could choose to play the static high effort PPE each period, independently of their relationship. Another possibility is to select zero effort independently of past project outcomes and signals.[39] We evaluate these different equilibria according to their expected payoffs. We find that if workers always choose the payoff maximizing equilibrium, then the zero effort equilibrium will never be played. Men will do even better in volatile environments, whereas women keep their advantage in environments with little uncertainty, leaving our conclusions of Section 2 unchanged.

In order to see this, we define the payoffs from choosing the static high effort PPE and from our proposed strategy, respectively:

$$\begin{aligned}
W_i^{stat} =& s_i(1+\beta)\mathbb{E}\left[f(e'(y), e'(y))\pi(y) - c(e'(y))\right], \qquad (14)\\
W_i^{dyn} =& s_i\mathbb{E}\left[f(e(y), e(y))\pi(y) - c(e(y))\right] \\
&+ s_i\beta\mathbb{E}\left[(1 - r(1 - f(e(y), e(y))))\right]\mathbb{E}\left[f(e'(y'), e'(y'))\pi(y') - c(e'(y'))\right]. \qquad (15)
\end{aligned}$$

The equilibrium we select yields a higher payoff than the static PPE whenever $W_i^{dyn} > W_i^{stat}$. To simplify notation, we let $\mathbb{E}[V_1] = \mathbb{E}\left[f(e(y), e(y))\pi(y) - c(e(y))\right]$ and $\mathbb{E}[V_2] = \mathbb{E}\left[f(e'(y'), e'(y'))\pi(y') - c(e'(y'))\right]$. Welfare under our strategy, W_i^{dyn}, is higher than welfare in the static high effort PPE, W_i^{stat}, whenever

$$\mathbb{E}[V_1] - \mathbb{E}[V_2] > \beta r_i(1 - \mathbb{E}[f(e(y), e(y))])\mathbb{E}[V_2] \qquad (16)$$

So, if $\mathbb{E}[V_1] - \mathbb{E}[V_2] > 0$ and $\mathbb{E}[f(e(y), e(y))]$ is sufficiently large, then welfare is higher under our strategy.[40] An example of parameter values for which equation (16) holds is given in Table

[39] Obviously, there are other equilibria, such as whenever a project fails, all relationships in the network turn bad and then all players choose zero effort. Another possibility is that a good relationship leads to zero effort and a bad relationship to positive effort. We find these equilibria hard to justify and therefore use the static PPE as a benchmark. Further, endogenizing the equilibrium selection is beyond the scope of this work.

[40] Note that $\mathbb{E}[V_1] - \mathbb{E}[V_2] > 0$ might not always be the case, although $e > e'$. To see this we consider the

16. We assume $f(e_i, e_j) = \sqrt{e_i, e_j}$ and $c(e_i) = \frac{1}{2}e_i^2$. In this example, men exert on average lower effort than women, in both states of the world. This is not surprising given that the project value in both states of the world is fairly similar.

Table 16: Welfare Parameters

v_l	v_h	p	q	β	d^W	d^M	C^W	C^M	N
1.5	1.6	0.75	0.5	0.9	2	3	2	1	4

Notice that $\mathbb{E}[f(e(y), e(y))]$ is large if effort is high under any signal realization. Effort does not vary greatly with the different signal realizations if the project values across states are similar, implying little uncertainty in the environment. We have shown that women exert higher effort than men in these environments, see Proposition (2).

If agents always play the strategy that yields the highest payoff, then in an environment with high uncertainty the static high effort PPE will be selected, whereas in an environment with low uncertainty and relatively high payoffs, our proposed strategy is implemented. But this implies that the differences between men and women, which we discussed in Section 2 , remain unchanged. Women would do even worse than men in uncertain environments than under our strategy and perform the same in situations with low uncertainty and high payoffs.

But the equilibrium that is payoff maximizing might not be selected. If a worker exerts positive effort, but his team partner shirks and only exerts zero effort, then he will face a loss. So, if there is a possibility of mis-coordination it might be better to always choose zero effort. Whether the expected payoff maximizing equilibrium or the zero effort equilibrium (that even under mis-coordination yields no losses) will be selected depends on whether payoff or risk dominant strategies should be played. The evidence for this is mixed at best (Van Huyck, Battalio, and Beil (1990), Cooper, DeJong, Forsythe, and Ross (1990), Cooper, DeJong, Forsythe, and Ross (1992)).

We believe that it is plausible to assume that workers might risk to choose the high effort which can potentially result in a loss (namely when they trust their project partner after a good history) and that they go for the strategy that ensures a nonnegative profit after a loss and thus bad history.

Appendix B: Proofs

Throughout, we make the following assumption on f:

Assumption 1. *The success probability function f satisfies:*

1. *Symmetry:* $f(e_i, e_j) = f(e_j, e_i)$
2. $f_1(e_i, e_j) > 0, f_2(e_j, e_i) > 0$
3. $f_{11}(e_i, e_j) = f_{22}(e_j, e_i) < 0$

example given in Table 3.5, where $\mathbb{E}[V_1] < \mathbb{E}[V_2]$. The reason is that workers choose very high effort in the first period even if the project does not yield a payoff in order to avoid having a bad relationship in the second period.

4. *Strict Supermodularity:* $f_{12}(e_i, e_j) = f_{21}(e_i, e_j) > 0$
5. $f(e_i, 0) = f(0, e_j) = 0$
6. $f(\lambda e_i, \lambda e_j) = \lambda f(e_i, e_j)$, $\lambda e_i, \lambda e_j \leq e_{max}$[41]

The cost function $c(e)$ *has the following properties:* $c(0) = 0$, $c_e(0) = 0$, $c_e(e_{max}) > \pi(y)$.

Proof of Lemma 1: Static Game

Given the Assumption 1, there always exists an equilibrium where both project partners exert zero effort. It therefore remains to show that this equilibrium is unique, with $e_i = e_j > 0$.

We first show symmetry. From the first order conditions we obtain

$$\frac{f_1(e_i, e_j)}{f_2(e_i, e_j)} = \frac{c_e(e_i)}{c_e(e_j)} \tag{17}$$

Suppose, by contradiction, that effort levels are not symmetric $e_j > e_i$. Due to convexity of the cost functions, the RHS of (17) is smaller than one. Due to concavity and supermodularity of the effort function, we have $f_1(e_i, e_j) > f_2(e_i, e_j)$, which is why the LHS is larger than one, which gives the contradiction.

Further, the equilibrium where both workers exert strictly positive effort is unique. It suffices to show that the FOCs (which under symmetry become a function of one variable) have one zero under the condition that effort is strictly positive.

$$f_1(e, e)\pi(y) = c_e(e) \tag{18}$$

Due to our assumption of constant returns to scale, $f_1(e, e)$ is constant in e. By our assumption of convex costs with $c(0) = 0$, the first derivative of the cost function $c_e(e)$ starts in the origin below $f_1(e, e)$ as $c_e(0) = 0$, is strictly increasing and at the maximum effort exceeds $\pi(y)$. Hence, the two functions have a unique intersection, implying a unique symmetric equilibrium with strictly positive effort.

Proof of Proposition 1:

The proof follows directly from maximization problem (6), Lemma 1 and the arguments given in the text.

Proof of Proposition 2:

Second Period Effort We first establish how second period effort is affected by additional information, depending on the state of the world. We know from equation (8) that the second period effort is a function of expected payoff, $\pi(y)$. To stress that a worker receives n signals, we adjust our notation and denote the project value by $\pi(y_n)$. We then establish in Lemma 2 that $\pi(y_n)$ increases in the number of signals in the high state and decreases in the number of signals in the

[41] We know that $e_i \in [0, e_{max}]$. If $\lambda \in [0, 1]$, then $\lambda e_i \leq e_{max}$, and for $\lambda > 1$ we impose the additional restriction that $\lambda e_i \leq e_{max}$, $\forall i$.

low state. It then follows immediately from equation (8) that a worker with more signals exerts higher effort in the high state and lower effort in the low state (proving claim 2. of Proposition 2).

Additionally, Lemma 2 characterizes the effect of vanishing uncertainty on the expected payoff.

Lemma 2 (Information and Expected Project Value)**.** *Project value $\pi(y_n)$ satisfies the martingale property: $\pi(y_n) = \mathbb{E}[\pi(y_{n+1})|y_n]$. However, given that the state is realized, a worker with more signals holds a more accurate posterior belief about the state of the world and thus about the project value:*

$$v_h > \mathbb{E}\left[\pi(y_{n+1})|\theta_h\right] > \mathbb{E}\left[\pi(y_n)|\theta_h\right] \qquad v_l < \mathbb{E}\left[\pi(y_{n+1})|\theta_l\right] < \mathbb{E}\left[\pi(y_n)|\theta_l\right].$$

The impact of an additional signal vanishes, if uncertainty vanishes, i.e. $\mathbb{E}\left[\pi(y_n)|\theta\right] = \mathbb{E}\left[\pi(y_{n+1})|\theta\right]$, if either (i) $v_l \to v_h$, (ii) $p \to 1$, (iii) $q \to 1$ if $\theta = \theta_h$, $q \to 0$ if $\theta = \theta_l$, or (iv) $n_{ext} \to \infty$.

Proof of Lemma 2:

We prove this Lemma in three steps.

First, we show the claim that $\pi(y)$ has the martingale property:

$$\pi(y_n) = P(\theta_h|y_n)v_h + (1 - P(\theta_h|y_n))v_l$$

Define $\psi_n \equiv P(\theta_h|y_n)$. We know that the stochastic process $\{\psi_n\}$ is a martingale as

$$\mathbb{E}[\psi_{n+1}|y_n] = \mathbb{E}[\mathbb{E}[\psi|y_{n+1}]|y_n] = \mathbb{E}[\psi|y_n] = \psi_n,$$

where the second equality follows from the *tower property* of conditional expectations. Then,

$$\begin{aligned}\mathbb{E}[\pi(y_{n+1})|y_n] &= \mathbb{E}[\psi_{n+1}v_h + (1-\psi_{n+1})v_l|y_n] = \mathbb{E}[\psi_{n+1}v_h|y_n] + \mathbb{E}[(1-\psi_{n+1})v_l|y_n] \\ &= \psi_n v_h + (1-\psi_n)v_l = \pi(y_n)\end{aligned}$$

Second, we prove the stated properties of $\mathbb{E}\left[\pi(y_n)\right]$ and $\mathbb{E}\left[\pi(y_n)|\theta\right]$. Some useful observations:

1. The number of signals do not impact $\mathbb{E}[\pi(y)]$ due to the martingale property of $\pi(y)$,

$$\mathbb{E}[\pi(y_{n+1})] = \mathbb{E}[\mathbb{E}[\pi(y_{n+1})|y_n]] = \mathbb{E}[\pi(y_n)].$$

2. We note that the posterior is given by

$$P(\theta_h|y) = \frac{P(y|\theta_h)P(\theta_h)}{P(\theta_h)P(y|\theta_h)+P(\theta_l)P(y|\theta_l)} = \frac{qp^x(1-p)^{n-x}}{qp^y(1-p)^{n-x}+(1-q)p^{n-x}(1-p)^x}$$
$$= \frac{1}{1+\frac{1-q}{q}\left(\frac{1-p}{p}\right)^{2x-n}} \tag{19}$$

To simplify notation we define $\tilde{p} \equiv \frac{1-p}{p}$, $\tilde{q} \equiv \frac{1-q}{q}$ and $\hat{y} \equiv 2x-n$. Then, $\psi_n = P(\theta_h|y) = \frac{1}{1+\tilde{q}\tilde{p}^{\hat{y}}}$.

3. We note that the expected project value conditional on state is given by:

$$\mathbb{E}\left[\pi(y_n)|\theta_h\right] = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!}\left(p^x(1-p)^{n-x}\right)\left(\frac{qp^x(1-p)^{n-x}v_h+(1-q)p^{n-x}(1-p)^x v_l}{qp^x(1-p)^{n-x}+(1-q)p^{n-x}(1-p)^x}\right)$$

We are interested in showing that

$$\mathbb{E}\left[\pi(y_{n+1})|\theta_h\right] > \mathbb{E}\left[\pi(y_n)|\theta_h\right] \tag{20}$$
$$\mathbb{E}\left[\pi(y_{n+1})|\theta_l\right] < \mathbb{E}\left[\pi(y_n)|\theta_l\right] \tag{21}$$

We will show that equation (20) holds and leave the proof of equation (21) to the reader. Using notation $\psi_n = P(\theta_h|y)$, we can rewrite equation (20) as

$$(v_h - v_l)\mathbb{E}\left[(\psi_{n+1}-\psi_n)|\theta_h\right] > 0$$

As $(v_h - v_l) > 0$, by assumption, it remains to be shown that $\mathbb{E}\left[\psi_{n+1}-\psi_n|\theta_h\right] > 0$. Given $\theta = \theta_h$, and a signal realization $\hat{y}$, $\psi_{n+1} = \frac{1}{1+\tilde{q}\tilde{p}^{\hat{y}+1}}$ with probability p and $\psi_{n+1} = \frac{1}{1+\tilde{q}\tilde{p}^{\hat{y}-1}}$, with probability $(1-p)$. Therefore,

$$\frac{1}{1+\tilde{q}\tilde{p}^{\hat{y}}} < \frac{p}{1+\tilde{q}\tilde{p}^{\hat{y}+1}} + \frac{1-p}{1+\tilde{q}\tilde{p}^{\hat{y}-1}}$$
$$\Leftrightarrow \quad p\tilde{p}^2 + (1-p) - \tilde{p} < \tilde{q}\tilde{p}^{\hat{y}}(p+(1-p)\tilde{p}^2 - \tilde{p})$$

which holds since $p\tilde{p}^2 + (1-p) - \tilde{p} = 0$ and $0 < \tilde{q}\tilde{p}^{\hat{y}}(p+(1-p)\tilde{p}^2-\tilde{p})$, as $p > \frac{1}{2}$ and thus

$$\mathbb{E}\left[\psi_n|\theta_h\right] < \mathbb{E}\left[\psi_{n+1}|\theta_h\right],$$

which concludes the proof.

Third, we show our last claim that additional signals do not matter as uncertainty vanishes which is true in any of the following cases:

(i) For $v_l \to v_h$,

$$\lim_{v_l \to v_h} \mathbb{E}\left[\pi(y_n)|\theta_h\right] = \sum_{x=0}^{n} \frac{(n)!}{x!(n-x)!} \left(p^x(1-p)^{n-x}\right) v_h = (p+1-p)^n v_h = v_h,$$

where the second step follows from the binomial formula. The expression is independent of n and therefore additional signals do not matter. An analogous argument holds for $\mathbb{E}\left[\pi(y)|\theta_l\right]$.

(ii) Assume $p \to 1$. Then,

$$\begin{aligned}\lim_{p\to 1} \mathbb{E}\left[\pi(y_n)|\theta_h\right] &= \lim_{p\to 1} \sum_{x=0}^{n} \frac{(n)!}{x!(n-x)!} \left(p^x(1-p)^{n-x}\right) \left(\frac{qp^x(1-p)^{n-x}v_h + (1-q)p^{n-x}(1-p)^x v_l}{qp^x(1-p)^{n-x} + (1-q)p^{n-x}(1-p)^x}\right) \\ &= \lim_{p\to 1} \frac{(n)!}{n!(n-n)!} \left(p^n(1-p)^{n-n}\right) \left(\frac{qp^n(1-p)^{n-n}v_h + (1-q)p^{n-n}(1-p)^n v_l}{qp^n(1-p)^{n-n} + (1-q)p^{n-n}(1-p)^n}\right) \\ &= \lim_{p\to 1} p^n \left(\frac{qp^n v_h + (1-q)(1-p)^n v_l}{qp^n + (1-q)(1-p)^n}\right) = v_h,\end{aligned}$$

which is independent of n; and analogous when conditioning on $\theta = \theta_l$.
(iii) Assume $q \to 1$. Then,

$$\lim_{q\to 1} \mathbb{E}\left[\pi(y_n)|\theta_h\right] = \sum_{x=0}^{n} \frac{(n)!}{x!(n-x)!} \left(p^x(1-p)^{n-x}\right) v_h = (p+1-p)^n v_h = v_h$$

which is independent of n. Similarly for $q \to 0$ and $\mathbb{E}\left[\pi(y)|\theta_l\right]$.
(iv) Note that $x \sim \text{Binomial}(np, np(1-p))$ if $\theta = \theta_h$ and $x \sim \text{Binomial}(n(1-p), np(1-p))$ if $\theta = \theta_l$. Then, $\lim_{n\to\infty}(x-(n-x)) = \infty$ if $\theta = \theta_h$ and $\lim_{n\to\infty}(x-(n-x)) = -\infty$ if $\theta = \theta_l$. To see this note that $x-(n-x) = 2x-n$. By the weak law of large numbers, as $n \to \infty$,

$$\begin{aligned} &\text{if} \quad \theta = \theta_h \quad x \xrightarrow{P} np \quad &&\Rightarrow \lim_{n\to\infty}(2np - n) = \infty \\ &\text{if} \quad \theta = \theta_l \quad x \xrightarrow{P} n(1-p) \quad &&\Rightarrow \lim_{n\to\infty}(2n(1-p) - n) = -\infty. \end{aligned}$$

Then, $\lim_{n\to\infty} P(\theta_h|y) = 1$ if the true state is $\theta = \theta_h$ and $\lim_{n\to\infty} P(\theta_h|y) = 0$ if the true state is $\theta = \theta_l$ as

$$\lim_{n\to\infty} P(\theta_h|y) = \lim_{n\to\infty} \frac{1}{1 + \frac{1-q}{q}\left(\frac{1-p}{p}\right)^{2x-n}}$$

We have already shown that $P(\theta_h|y)$ is increasing in n if $\theta = \theta_h$ and decreasing in n if $\theta = \theta_l$. Thus we can apply the Monotone Convergence Theorem, which implies that $\lim_{n\to\infty} \mathbb{E}[P(\theta_h|y)v_h] = \mathbb{E}[\lim_{n\to\infty} P(\theta_h|y)v_h]$. From this it follows that $\lim_{n\to\infty} \mathbb{E}\left[\pi(y)|\theta_h\right] = v_h$ and $\lim_{n\to\infty} \mathbb{E}\left[\pi(y)|\theta_l\right] = v_l$. ■

First Period Effort The first period effort is a function of both $\pi(y_n)$ and $\mathbb{E}[V^*(y')]$, see equation

(7). Thus, additional signals affect effort through their impact on $\pi(y_n)$ and $\mathbb{E}[V^*(y')]$. As we have already established the effect of additional signals on $\pi(y_n)$, we now turn to $\mathbb{E}[V^*(y')]$. In Lemma 3 we first show that $V^*(y')$ is a convex function of the second period project value, $\pi(y')$, which is a martingale, see Lemma 2. This establishes that additional signals lead to a higher expected value. As before we analyze the effect of vanishing uncertainty, now on the expected value.

Lemma 3 (Information and Second Period Expected Value)**.** *$V^*(y'_n)$ is a submartingale. Thus, a worker with more signals has a higher second period expected value:*

$$\mathbb{E}[V^*(y'_n)] < \mathbb{E}[V^*(y'_{n+1})],$$

The impact of an additional signal vanishes, if uncertainty vanishes, i.e. $\mathbb{E}[V^*(y'_n)] = \mathbb{E}[V^*(y'_{n+1})]$, if either (i) $v_l \to v_h$ (ii) $p \to 1$ (iii) $q \to 1$ if $\theta = \theta_h$, $q \to 0$ if $\theta = \theta_l$, or (iv) $n_{ext} \to \infty$.

Proof of Lemma 3:

First, we establish that $V^*(y')$ is a submartingale: We can express $V^*(y')$ as a function of $\pi(y')$, and write

$$V^*(y') \equiv g(\pi(y')) \tag{22}$$

As $\pi(y')$ is a martingale, we have that $g(\pi(y'))$ is a submartingale if g is a convex function, whenever $\mathbb{E}[V^*(y'_n)] < \infty$ which holds as $0 \leq \mathbb{E}[V^*(y'_n)] < v_h, \forall n$.

Note that equilibrium effort depends on the expected project payoff through the signals, or $e'(y')$. We mostly omit this dependence here in order to keep notation simple and write e'.

Applying the envelope theorem repeatedly, the first and second derivative of g are given by

$$\begin{aligned}
\frac{\partial g(\pi(y'))}{\partial \pi(y')} &= f_2(e',e')\pi(y')\frac{\partial e'}{\partial \pi(y')} + f(e',e') \\
\frac{\partial^2 g(\pi(y'))}{\partial \pi(y')^2} &= [f_{22}(e',e') + f_{12}(e',e')]\pi(y')\left(\frac{\partial e'}{\partial \pi(y')}\right)^2 + f_2(e',e')\pi(y')\frac{\partial^2 e'}{\partial \pi(y')^2} + f_2(e',e')\frac{\partial e'}{\partial \pi(y')} \\
&+ (f_1(e',e') + f_2(e',e'))\frac{\partial e'}{\partial \pi(y')} \\
&= f_2(e',e')\pi(y')\frac{\partial^2 e'}{\partial \pi(y')^2} + f_2(e',e')\frac{\partial e'}{\partial \pi(y')} + (f_1(e',e') + f_2(e',e'))\frac{\partial e'}{\partial \pi(y')}
\end{aligned}$$

From the first order condition of the static problem, evaluated at the equilibrium effort, we can compute

$$\begin{aligned}
\frac{\partial e'}{\partial \pi(y')} &= \frac{f_1(e',e')}{(\partial^2 c(e')/\partial e'^2)} > 0 \\
\frac{\partial^2 e'}{\partial \pi(y')^2} &= \frac{(f_{11}(e',e') + f_{21}(e',e'))\frac{\partial e'}{\partial \pi(y)}}{(\partial^2 c(e')/\partial e'^2)} = 0
\end{aligned}$$

It follows that

$$\frac{\partial^2 g(\pi(y'))}{\partial \pi(y')^2} = f_2(e',e')\frac{\partial e'}{\partial \pi(y')} + (f_1(e',e') + f_2(e',e'))\frac{\partial e'}{\partial \pi(y')} > 0,$$

which implies that $V^*(y_n')$ is a submartingale and therefore $\mathbb{E}[V^*(y_n')]$ is increasing in n.

Second, we prove the stated properties of

$$\mathbb{E}[V_n^*] = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!}(qp^x(1-p)^{n-x} + (1-q)p^{n-x}(1-p)^x)\left(f(e',e')\pi(y) - c(e')\right) \tag{23}$$

as uncertainty vanishes:

(i) Consider $v_l \to v_h$.
We are interested in

$$\lim_{v_l \to v_h} \mathbb{E}[V_n^*] = \lim_{v_l \to v_h} \sum_{x=0}^{n} \frac{n!}{(x)!(n-x)!}(qp^x(1-p)^{n-x} + (1-q)p^{n-x}(1-p)^x)\left(f(e'(y'),e'(y'))\pi(y') - c(e'(y'))\right),$$

where $e'(y')$ is the equilibrium effort for given y'. As the other terms are constant in v_l, all that matters is

$$\begin{aligned}\lim_{v_l \to v_h}\left(f(e'(y'),e'(y'))\pi(y') - c(e'(y'))\right) &= \lim_{v_l \to v_h} f(e'(y'),e'(y')) \lim_{v_l \to v_h} \pi(y') - \lim_{v_l \to v_h} c(e'(y')) \\ &= \lim_{v_l \to v_h} f(e'(y'),e'(y'))v_h - \lim_{v_l \to v_h} c(e'(y'))\end{aligned}$$

Note that $\lim_{\pi(y') \to v_h} e'(y') = e'_{v_h}$, i.e. the effort converges to some constant e'_{v_h} as $\pi(y') \to v_h$, since $e'(y')$ is a linear function of $\pi(y')$ as due to constant returns to scale of f $e(y) = f_1(1,1)\pi(y)$. Also, due to constant returns to scale, $f(e'(y'),e'(y')) = e'(y')f(1,1)$ and thus $\lim_{e'(y') \to e'_{v_h}} f(e'(y'),e'(y')) = e'_{v_h}f(1,1)$, which again is constant in n. As $f(e'_{v_h},e'_{v_h}) = e'_{v_h}f(1,1)$ is continuous, we know that $\lim_{\pi(y') \to v_h} f(e'(y'),e'(y')) = e'_{v_h}f(1,1)$. The argument is similar for $c(.)$. Then, we can write

$$\lim_{v_l \to v_h}\left(f(e'(y'),e'(y'))\pi(y') - c(e'(y'))\right) = b_{v_l},$$

where b_{v_l} is constant and thus independent of n. Therefore, as v_l converges to v_h, the expected second period value converges to a constant and is independent of the number of signals,

$$\lim_{v_l \to v_h} \mathbb{E}[V_n^*] = b_{v_l}.$$

(ii) Consider $p \to 1$ for $\theta \in \{\theta_h, \theta_l\}$.

Note that

$$\begin{aligned}
&\lim_{p\to 1}\pi(y) = v_h \quad \text{if} \quad n-2x<0\\
&\lim_{p\to 1}\pi(y) = qv_h + (1-q)v_l \quad \text{if} \quad n-2x=0\\
&\lim_{p\to 1}\pi(y) = v_l \quad \text{if} \quad n-2x>0
\end{aligned}$$

As $\pi(y)$ converges to some constant (and, of course, the same holds for $\pi(y')$), so does $f(e'(y'), e'(y'))\pi(y') - c(e')$. We denote by $V^*(v_h)$ $(V^*(v_l))$ $[V^*(v)]$ the limit of $f(e'(y'), e'(y'))\pi(y') - c(e')$ when $\pi(y)$ converges to v_h (v_l) $[qv_h + (1-q)v_l]$.

Note further that if $n-2x<0$, $\lim_{p\to 1}(qp^x(1-p)^{n-x} + (1-q)p^{n-x}(1-p)^x) = \lim_{p\to 1} qp^x(1-p)^{n-x}$. Then we know that

$$\lim_{p\to 1} qp^x(1-p)^{n-x} = \begin{cases} q & \text{if} \quad x=n \\ 0 & \text{otherwise} \end{cases}$$

If $n-2x>0$, $\lim_{p\to 1}(qp^x(1-p)^{n-x} + (1-q)p^{n-x}(1-p)^x) = \lim_{p\to 1}(1-q)p^{n-x}(1-p)^x$. It follows that

$$\lim_{p\to 1}(1-q)p^{n-x}(1-p)^x = \begin{cases} 1-q & \text{if} \quad x=0 \\ 0 & \text{otherwise} \end{cases}$$

Last, if $n-2x=0$, $\lim_{p\to 1}(qp^x(1-p)^{n-x} + (1-q)p^{n-x}(1-p)^x) = \lim_{p\to 1} p^x(1-p)^{n-x} = 0$, as $x, n > 0$. From this it then follows that

$$\begin{aligned}
\lim_{p\to 1}\mathbb{E}[V_n^*] &= \lim_{p\to 1}\left(\sum_{x=0}^{n}\frac{n!}{x!(n-x)!}(qp^x(1-p)^{n-x} + (1-q)p^{n-x}(1-p)^x)\left(f(e'(y'), e'(y'))\pi(y') - c(e'(y'))\right)\right)\\
&= qV^*(v_h) + (1-q)V^*(v_l),
\end{aligned}$$

which is independent of n.

(iii) Consider $q\to 1$. Notice that

$$\begin{aligned}
&\lim_{q\to 1}(qp^x(1-p)^{n-x} + (1-q)p^{n-x}(1-p)^x) = p^{n-x}(1-p)^x,\\
&\lim_{q\to 1}\pi(y) = v_h.
\end{aligned}$$

It follows that

$$\begin{aligned}
\lim_{q\to 1}\mathbb{E}[V_n^*] &= \lim_{q\to 1}\left(\sum_{x=0}^{n}\frac{n!}{x!(n-x)!}(qp^x(1-p)^{n-x}+(1-q)p^{n-x}(1-p)^x)\left(f(e'(y'),e'(y'))\pi(y')-c(e'(y'))\right)\right)\\
&= \lim_{q\to 1}\left(\sum_{x=0}^{n}\frac{n!}{x!(n-x)!}(p^{n-x}(1-p)^x)V^*(v_h)\right)\\
&= \lim_{q\to 1}V^*(v_h)(1-p+p)^n = V^*(v_h),
\end{aligned}$$

where the last step follows from the fact that $V^*(v_h)$ is a constant and the *binomial theorem*. Thus, the limit is a constant and independent of n.

Next, consider $q \to 0$.

$$\begin{aligned}
&\lim_{q\to 0}(qp^x(1-p)^{n-x}+(1-q)p^{n-x}(1-p)^x) = p^{n-x}(1-p)^x,\\
&\lim_{q\to 0}\pi(y) = v_l,
\end{aligned}$$

and by the same steps as previously it follows that $\lim_{q\to 0}\mathbb{E}[V_n^*]$ is constant.

(iv) Consider the case of abundance of information: $n_{ext} \to \infty$.
We want to show that

$$\lim_{n\to\infty}\mathbb{E}[V_n^*] = \mathbb{E}[V^*].$$

We know that for each n, $\mathbb{E}[V_n^*] \leq \mathbb{E}[V_{n+1}^*]$ as V_n^* is a submartingale and that $\mathbb{E}[V_n^*] \leq v_h$ for all n. By the monotone convergence theorem, we know that a finite limit exists, which we denote by $\mathbb{E}[V^*]$. ■

We have thus established that $\mathbb{E}[V_n^*]$ is increasing in the number of signals. We know from Lemma 2 that $\pi(y_n)$ can be increasing or decreasing in the number of signals, depending on the state of the world. Thus, if the state in the first period is high, first period effort is increasing in the number of signals (proving claim 1. in Proposition 2).

Wages The effect of information on wages follows immediately from wage functions (9) and (10), conditional on the high state $\theta = \theta_h$ (proving claim 3. in Proposition 2).

Vanishing Uncertainty Additional information does not affect second period effort if uncertainty is vanishing, see Lemma 2 (i)-(iv). Further, the result that the impact of degree on average first period effort vanishes as uncertainty vanishes is due to Lemma 2 (i)-(iv) and Lemma 3 (i)-(iv). Similarly, under vanishing uncertainty, the impact of a higher degree on wages vanishes since information affects wages through effort.

Proof of Proposition 3

The effect of clustering on expected first period effort follows from our expression of equilibrium effort (7), showing that it is increasing in sr. The effect on productivity/wages follows immediately from the wage function (9).

Proof of Proposition 4:

We show that clustering and degree are complementary for the expected first period effort. We can rewrite the first order condition (7) as follows

$$\left(\frac{c_e(e)}{f_1(e,e)}\right) = \pi(y) + \beta sr\mathbb{E}[V^*(y')]$$

Taking expectations yields

$$\mathbb{E}\left(\frac{c_e(e)}{f_1(e,e)}\right) = \mathbb{E}(\pi(y)) + \beta sr\mathbb{E}[V^*(y')]$$

$E(\pi(y))$ is independent of the number of first period signals and so an additional signal only increases $\mathbb{E}[V^*(y')]$. We define

$$F\left(e, \mathbb{E}[V^*(y')], sr\right) \equiv \mathbb{E}\left(\frac{c_e(e)}{f_1(e,e)}\right) - \mathbb{E}(\pi(y)) - \beta sr\mathbb{E}[V^*(y')] = 0$$

Then we calculate

$$\begin{aligned}\frac{\partial^2 F\left(e, \mathbb{E}[V^*(y')], sr\right)}{\partial\mathbb{E}[V^*(y')]\partial sr} &= \frac{\partial\left(\frac{\partial\mathbb{E}\left(\frac{c_e(e)}{f_1(e,e)}\right)}{\partial e}\frac{\partial e}{\partial\mathbb{E}[V^*(y')]}\right)}{\partial sr} - \beta \\ &= \frac{\partial^2\left(\mathbb{E}\left(\frac{c_e(e)}{f_1(e,e)}\right)\right)}{\partial e^2}\frac{\partial e}{\partial\mathbb{E}[V^*(y')]}\frac{\partial e}{\partial sr} + \frac{\partial\mathbb{E}\left(\frac{c_e(e)}{f_1(e,e)}\right)}{\partial e}\frac{\partial^2 e}{\partial\mathbb{E}[V^*(y')]\partial sr} - \beta = 0\end{aligned}$$

Note that

$$\frac{\partial\mathbb{E}\left(\frac{c_e(e)}{f_1(e,e)}\right)}{\partial e} = \mathbb{E}\left(\frac{c_{ee}(e)}{f_1(e,e)}\right) > 0; \qquad \frac{\partial^2\mathbb{E}\left(\frac{c_e(e)}{f_1(e,e)}\right)}{\partial^2 e} = \mathbb{E}\left(\frac{c_{eee}(e)}{f_1(e,e)}\right) \leq 0$$

Further,

$$\frac{\partial e}{\partial sr} = \frac{\beta\mathbb{E}[V^*(y')]}{\mathbb{E}\left(\frac{c_{ee}(e)}{f_1(e,e)}\right)} > 0; \qquad \frac{\partial e}{\partial\mathbb{E}[V^*(y')]} = \frac{\beta sr}{\mathbb{E}\left(\frac{c_{ee}(e)}{f_1(e,e)}\right)} > 0$$

It follows that

$$\frac{\partial^2 e}{\partial\mathbb{E}[V^*(y')]\partial sr} = \frac{1}{\mathbb{E}\left(\frac{c_{ee}(e)}{f_1(e,e)}\right)}\left(\beta - \frac{\beta sr}{\mathbb{E}\left(\frac{c_{ee}(e)}{f_1(e,e)}\right)}\frac{\beta\mathbb{E}[V^*(y')]}{\mathbb{E}\left(\frac{c_{ee}(e)}{f_1(e,e)}\right)}\mathbb{E}\left(\frac{c_{eee}(e)}{f_1(e,e)}\right)\right) > 0$$

This implies immediately that clustering and degree are complements, also if we allow for discrete changes. Given that effort increases more for a high level of information as clustering increases, it follows that the first period wage, which is an increasing function of effort displays increasing differences in clustering and degree:

$$\frac{\partial^2 \mathbb{E}[w]}{\partial \mathbb{E}[V^*(y')]\partial sr} = \mathbb{E}\left[(f_1(e,e)+f_2(e,e))\frac{\partial^2 e}{\partial \mathbb{E}[V^*(y')]\partial sr}\right] > 0 \tag{24}$$

Proof of Proposition 5: Trade-Off Between Information and Peer Pressure

We assume that a D-worker has a higher degree and hence more signals, n_{int}, and has clustering, $(sr)^D$. In turn, a C-worker has a lower degree and thus a lower number of signals (and therefore $s^D > s^C$) but higher clustering and therefore $(sr)^C > (sr)^D$.

1. Comparative Advantage:

We want to show that $\frac{\mathbb{E}[w^C]}{\mathbb{E}[w^D]}$ (where w^C indicates the first period wage of a C-worker and w^D indicates the first period wage of a D-worker) increases as the environment becomes more certain. First notice that,

$$\frac{\mathbb{E}[w^C]}{\mathbb{E}[w^D]} = \frac{qw^C(\theta_h)+(1-q)w^C(\theta_l)}{qw^D(\theta_h)+(1-q)w^D(\theta_l)} \tag{25}$$

Note that under the stated assumption of CRS of f, the first and second period efforts are given in closed form:

$$e(y) = f_1(1,1)(\pi(y)+\beta sr\mathbb{E}[V^*(y')]) \tag{26}$$

$$e'(y') = f_1(1,1)\pi(y'). \tag{27}$$

And so we obtain for the expected wage as:

$$\begin{aligned}\mathbb{E}[w] =& f(1,1)f_1(1,1)\left(qv_h+(1-q)v_l\right)\beta sr\mathbb{E}[V^*(y')] \\ &+ f(1,1)f_1(1,1)\left(qv_h\mathbb{E}[\pi(y|\theta_h)]+(1-q)v_l\mathbb{E}[\pi(y|\theta_l)]\right),\end{aligned} \tag{28}$$

which follows from substituting equation (26) into wage equation (9). To simplify notation we define $k_1 \equiv f(1,1)f_1(1,1)$ and $\overline{v} \equiv qv_h + (1-q)v_l$. By the law of total expectation it follows that $(1-q)\mathbb{E}[\pi(y|\theta_l)] = \mathbb{E}[\pi(y)] - q\mathbb{E}[\pi(y|\theta_h)]$. Then, equation (28) becomes

$$\mathbb{E}[w] = k_1\left(\overline{v}\beta sr\mathbb{E}[V^*(y')]+q(v_h-v_l)E[\pi(y|\theta_h)]+\mathbb{E}[\pi(y)]v_l\right) \tag{29}$$

The wage ratio (25) can then be expressed as

$$\frac{\mathbb{E}[w^C]}{\mathbb{E}[w^D]} = \frac{\overline{v}\beta(sr)^C\left(\mathbb{E}[V^*(y')]\right)^C+q(v_h-v_l)\left(\mathbb{E}[\pi(y|\theta_h)]\right)^C+\left(\mathbb{E}[\pi(y)]\right)^C v_l}{\overline{v}\beta(sr)^D\left(\mathbb{E}[V^*(y')]\right)^D+q(v_h-v_l)\left(\mathbb{E}[\pi(y|\theta_h)]\right)^D+\left(\mathbb{E}[\pi(y)]\right)^D v_l} \tag{30}$$

Note that $\mathbb{E}[\pi(y)]$ is independent of the number of signals as it is a martingale and thus, $(\mathbb{E}[\pi(y)])^C = (\mathbb{E}[\pi(y)])^D = \overline{v}$. Further, note that

$$\mathbb{E}[V^*(y')] = C_1(v_h - v_l)^2 q\mathbb{E}\left[\psi_n|\theta_h\right] + 2C_1 v_l(v_h - v_l)q + C_1 v_l^2 \tag{31}$$

$$\mathbb{E}[\pi(y)|\theta_h)] = (v_h - v_l)\mathbb{E}[\psi_n|\theta_h] + v_l \tag{32}$$

where $C_1 = f_1(1,1)\left(1 - \frac{1}{2}f_1(1,1)\right)$. C_1 is positive as the value of the problem is positive. To obtain the simplified expression for $\mathbb{E}[V^*(y')]$ in (31), we used equation (23) and substituted in the equilibrium first period effort (26). We then applied the binomial theorem and used the martingale property.

To make the notation more compact and to single out those variables that depend on information, we now introduce the variables a^i and b^i, $i \in \{C, D\}$ (which do not depend on information) and write the wage ratio (25) as (where we also use (31) and (32)):

$$\frac{\mathbb{E}[w^C]}{\mathbb{E}[w^D]} = \frac{a^C + b^C\left(\mathbb{E}[\psi|\theta_h]\right)^C}{a^D + b^D\left(\mathbb{E}[\psi|\theta_h]\right)^D} \tag{33}$$

For illustration, we now focus on the case where the D-worker has one more signal than the C-worker. Our exercise aims at analyzing (33) when reducing uncertainty, which we here achieve by letting n_{ext} and thus n grow. If (33) is increasing in the number of signals n, it must hold that

$$\frac{a^C + b^C\mathbb{E}[\psi_n|\theta_h]}{a^D + b^D\mathbb{E}[\psi_{n+1}|\theta_h]} > \frac{a^C + b^C\mathbb{E}[\psi_{n-1}|\theta_h]}{a^D + b^D\mathbb{E}[\psi_n|\theta_h]}$$

or

$$\begin{aligned}&\left(a^C b^D + a^D b^C\right)\mathbb{E}[\psi_n|\theta_h] + b^C b^D \mathbb{E}[\psi_n|\theta_h]^2 \\ &\quad > a^C b^D\mathbb{E}[\psi_{n+1}|\theta_h] + a^D b^C\mathbb{E}[\psi_{n-1}|\theta_h] + b^C b^D\mathbb{E}[\psi_{n+1}|\theta_h]\mathbb{E}[\psi_{n-1}|\theta_h]\end{aligned} \tag{34}$$

We focus first on showing that $b^C b^D\mathbb{E}[\psi_n|\theta_h]^2 > b^C b^D\mathbb{E}[\psi_{n+1}|\theta_h]\mathbb{E}[\psi_{n-1}|\theta_h]$, or

$$\mathbb{E}[\psi_n|\theta_h]^2 > \mathbb{E}[\psi_{n+1}|\theta_h]\mathbb{E}[\psi_{n-1}|\theta_h] \tag{35}$$

Thus, if we establish that $\mathbb{E}[\psi_n|\theta_h]$ is log-concave, then inequality (35) follows immediately. As concavity implies log-concavity it suffices to show that $\mathbb{E}[\psi_n|\theta_h]$ is concave.

Concavity of $\mathbb{E}[\psi_n|\theta_h]$*:*

It is helpful to express ψ_n in terms of its log-likelihood ratio (LLR). Without any signals the LLR, denoted by λ_0 is a function of the prior q:

$$\lambda_0 = \log\left(\frac{q}{1-q}\right), \tag{36}$$

Generally, the LLR is given by

$$\lambda_{n+1} = \lambda_n + 2\log\left(\frac{p}{1-p}\right)(x_n - \frac{1}{2}),$$

where we denote by x_n the signal realization of the nth observation. Further,

$$\log\left(\frac{\psi_n}{1-\psi_n}\right) = \lambda_n \qquad \Leftrightarrow \qquad \psi_n = \frac{e^{\lambda_n}}{1+e^{\lambda_n}}.$$

Taking expectations yields

$$\mathbb{E}(\psi_n|\theta_h) = \mathbb{E}\left(\frac{e^{\lambda_n}}{1+e^{\lambda_n}}|\theta_h\right)$$

Then, we take the first and second derivative with respect to n, which yields

$$\begin{aligned}\frac{\partial \mathbb{E}(\psi_n|\theta_h)}{\partial n} &= \mathbb{E}\left(\frac{\partial \psi_n}{\partial \lambda_n}\frac{\partial \lambda_n}{\partial n}\middle|\theta_h\right)\\ \frac{\partial^2 \mathbb{E}(\psi_n|\theta_h)}{\partial n^2} &= \mathbb{E}\left(\frac{\partial^2 \psi_n}{\partial \lambda_n^2}\left(\frac{\partial \lambda_n}{\partial n}\right)^2 + \frac{\partial \psi_n}{\partial \lambda_n}\frac{\partial^2 \lambda_n}{\partial n^2}\middle|\theta_h\right)\end{aligned}$$

Note that λ_n is a linear function in n. To see this note that with each signal, the LLR either increases or decreases by a constant. Thus, $\frac{\partial^2 \lambda_n}{\partial n^2} = 0$ and

$$sign\left(\frac{\partial^2 \mathbb{E}(\psi_n|\theta_h)}{\partial n^2}\right) = sign\left(\frac{\partial^2 \mathbb{E}(\psi_n|\theta_h)}{\partial \lambda_n^2}\right)$$

We therefore restrict attention to the derivative with respect to λ_n:

$$\begin{aligned}\frac{\partial \mathbb{E}(\psi_n|\theta_h)}{\partial \lambda_n} &= \mathbb{E}\left(\frac{(1+e^{\lambda_n})e^{\lambda_n} - e^{2\lambda_n}}{(1+e^{\lambda_n})^2}\middle|\theta_h\right) = \mathbb{E}\left(\frac{e^{\lambda_n}}{(1+e^{\lambda_n})^2}\middle|\theta_h\right)\\ \frac{\partial^2 \mathbb{E}(\psi_n|\theta_h)}{\partial \lambda_n^2} &= \mathbb{E}\left(\frac{(1+e^{\lambda_n})^2 e^{\lambda_n} - 2e^{2\lambda_n}(1+e^{\lambda_n})}{(1+e^{\lambda_n})^4}\middle|\theta_h\right) = \mathbb{E}\left(\frac{e^{\lambda_n}(1-e^{\lambda_n})}{(1+e^{\lambda_n})^3}\middle|\theta_h\right)\end{aligned}$$

The second derivative is negative (thereby implying that $\mathbb{E}(\psi_n|\theta_h)$ is concave) if

$$1 - e^{\lambda_n} < 0 \quad \Leftrightarrow \quad 0 < \lambda_n$$

This implies that if the LLR is negative, then the expected posterior belief $\mathbb{E}(\psi_n|\theta_h)$ is convex, otherwise, it is concave. The LLR is positive if the probability of the high state outweighs the probability of the low state, that is if sufficiently many signals have been positive. It remains to be shown that, given that the true state is $\theta = \theta_h$, λ_n is positive for some n and that once it is positive, the probability of it becoming negative again vanishes. We first show that λ_n becomes positive within a finite number of observations. To see this define a stopping time T over the set of all possible observations $\mathcal{P}$,

$$T = \inf\{n \geq 0 : \lambda_n^+ \in \mathcal{P}\},$$

where λ_n^+ is the sequence for which $\lambda_n > 0$. Then, Williams (1991), p. 101 establishes the following:

Lemma 4. *Suppose that T is a stopping time such that for some $N \in \mathbb{N}$ and some $\epsilon > 0$ we have for every $n \in \mathbb{N}$:*

$$P(T \leq n + N|\mathcal{F}_n) > \epsilon \qquad \textit{almost surely} \tag{37}$$

Then, $\mathbb{E}[T] < \infty$.

Note that $\mathcal{F}_n$ denotes the filtration with n observations. Inequality (37) is fulfilled as the probability that there are more positive than negative signals given any number of signals is strictly positive and thus there exists an ϵ that is smaller than this probability. This establishes that λ_n becomes positive for a finite number of signals.

Next, we want to establish that once λ_n is strictly positive, the probability of λ_n becoming negative converges to zero as n grows. We know from *Hoeffding's Inequality* that the number of high signals is concentrated around its mean, with exponentially small tail, formally

$$P(np - y \geq t) \leq e^{-2nt^2}$$

Given that some λ_n is positive, we know that the number of high signals must satisfy $y > \frac{n}{2}$ for n even and $y \geq \frac{n+1}{2}$ for n odd. We focus on the case where n is even, the case of n odd follows immediately. The probability of λ_n being negative is equivalent to having more than half of the signals indicating the low state. We therefore set $t = np - \frac{n}{2}$, which yields

$$P(np - y \geq np - \frac{n}{2}) \leq exp\left(-2n^3\left(p - \frac{1}{2}\right)^2\right) \tag{38}$$

It is evident that for n sufficiently high, the probability of having more low signals than high signals (which is the probability on the LHS of (38)) approaches zero quickly and thus we have established that λ_n is positive in finite time and remains positive for sufficiently large n with probability approaching one. Thus for a sufficiently high LLR, $E(\psi_n|\theta_h)$ is concave. While we focus here on the effect of an increase in signals n on the LLR, note that the LLR is also affected by q and p, where q is the prior probability of the high state and p is the probability of the signal being high given that the state is high. More precisely, λ_n is increasing in q and p. Thus, the LLR is influenced by all of our measures of uncertainty. Decreasing uncertainty by increasing q, p or n leads to a higher and, at some point, positive LLR, in which case $E(\psi_n|\theta_h)$ is concave.

Inequality (35) is thus fulfilled and for (34) to hold it remains to be shown that

$$\begin{aligned} & \left(a^C b^D + a^D b^C\right) \mathbb{E}(\psi_n|\theta_h) \geq a^C b^D \mathbb{E}(\psi_{n+1}|\theta_h) + a^D b^C \mathbb{E}(\psi_{n-1}|\theta_h) \\ \Leftrightarrow \qquad & a^D b^C \mathbb{E}(\psi_n - \psi_{n-1}|\theta_h) \geq a^C b^D \mathbb{E}(\psi_{n+1} - \psi_n|\theta_h) \end{aligned} \tag{39}$$

is fulfilled. We know that $\mathbb{E}(\psi_n|\theta_h)$ is concave and increasing for n sufficiently high, and thus

$$\mathbb{E}(\psi_n - \psi_{n-1}|\theta_h) > \mathbb{E}(\psi_{n+1} - \psi_n|\theta_h).$$

Further, we can show that $a^D b^C = a^C b^D$, which is equivalent to

$$\beta C_1 \left((sr)^C - (sr)^D\right) \overline{v}(\overline{v} - qv_h - (1-q)vl) = 0$$

as $\overline{v} = qv_h + (1-q)vl$. Thus, we have shown that (39) is positive for n sufficiently high, which establishes that C-workers have a comparative advantage as uncertainty vanishes. ■

2. Wage Dynamics:

Claim: $\quad w^D(\theta) \geq w^C(\theta) \quad \Rightarrow \quad \mathbb{E}[w'^D] > \mathbb{E}[w'^C]$.

From (10), it follows that the second period expected wage across states is defined as

$$\mathbb{E}[w'] = qw'(\theta, \theta_h') + (1-q)w'(\theta, \theta_l') = f(1,1)s_i P_i(\gamma_g'|\theta) \left(q\mathbb{E}[e'(y')|\theta_h']v_h + (1-q)\mathbb{E}[e'(y')|\theta_l']v_l\right)$$

Recall that $P(\gamma_g'|\theta) \equiv \mathbb{E}[f\left(e(y), e(y)\right) + (1-r)(1 - f\left(e(y), e(y)\right))|\theta] = \mathbb{E}[e(y)|\theta] r f(1,1) + 1 - r$.

Suppose that in the first period $w^D(\theta) \geq w^C(\theta)$, implying $\mathbb{E}[e(y)^D|\theta] \geq \mathbb{E}[e(y)^C|\theta]$. Moreover, by assumption, $s^C < s^D$ and $(sr)^C > (sr)^D$. Hence, $[sP(\gamma_g|\theta)]^D > [sP(\gamma_g|\theta)]^C$ since

$$[sP(\gamma_g|\theta)]^D = (sr)^D(\mathbb{E}[e(y)^D|\theta]f(1,1) - 1) + s^D > [sP(\gamma_g|\theta)]^C = (sr)^C(\mathbb{E}[e(y)^C|\theta]f(1,1) - 1) + s^C$$

where the expression in brackets, $\mathbb{E}[e(y)|\theta]f(1,1) - 1$, is negative but (weakly) less so for the D-worker. Last, we focus on

$$q\mathbb{E}[e'(y')|\theta_h']v_h + (1-q)\mathbb{E}[e'(y')|\theta_l']v_l = f_1(1,1)\left(q(v_h - v_l)\mathbb{E}[\pi(y')|\theta_h']v_h + \overline{v}v_l\right),$$

where we again denoted $\overline{v} \equiv qv_h + (1-q)v_l$ and where we used the law of total expectation $(1-q)\mathbb{E}[\pi(y|\theta_l)] = \mathbb{E}[\pi(y)] - q\mathbb{E}[\pi(y|\theta_h)]$. As $\mathbb{E}[\pi(y')|\theta_h']$ is the only variable here that depends on information and since it is increasing in the number of signals, it follows that $q\left(\mathbb{E}[e'(y')|\theta_h']\right)^D v_h + (1-q)\left(\mathbb{E}[e'(y')|\theta_l']\right)^D v_l > q\left(\mathbb{E}[e'(y')|\theta_h']\right)^C v_h + (1-q)\left(\mathbb{E}[e'(y')|\theta_l']\right)^C v_l$. Thus, $w^D(\theta) \geq w^C(\theta)$ implies $\mathbb{E}[w'^D] > \mathbb{E}[w'^C]$, which proves the claim. ■

Appendix C: Data and Additional Results

Computer Scientists

Table 17: Correlation of Network Measures: Computer Science

	Degree	Clustering Coefficient
Degree	1	
Clustering Coefficient	-0.46***	1

***p<0.01, **p<0.05, *p<0.1.

Note: Sample consists of 438,531 men and 146,829 women. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Enron

Table 18: Correlation of Network Measures: Enron

	Degree	Clustering Coefficient
Degree	1	
Clustering Coefficient	-0.21***	1

***p<0.01, **p<0.05, *p<0.1.

The statistic is based on our sample of 3,926 individuals, 1,628 women and 2,298 men. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

AddHealth

Table 19: Correlation of Network Measures: Add Health

	Degree	Clustering Coefficient
Degree	1	
Clustering Coefficient	-0.11***	1

***p<0.01, **p<0.05, *p<0.1.

Note: This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links. It also does not include individuals for whom we do not have gender or age. Observations : 73,244.

Table 20: Differences in Degree & Clustering for Boys and Girls: Age 12-18, Balanced Gender Ratio, Add Health

	Cl. Coeff. (dir)	Cl. Coeff. (dir)	Cl. Coeff.	Cl. Coeff.	Degree	Degree	In Degree	In Degree	Out Degree	Out Degree
Female	0.0104	0.0299	0.102***	0.109***	0.0538***	0.122***	0.0317**	0.0895***	0.156***	0.195***
	(0.0131)	(0.0158)	(0.0128)	(0.0155)	(0.0114)	(0.0144)	(0.0123)	(0.0156)	(0.0119)	(0.0145)
Age	0.00628		0.00524		-0.0344***		-0.00839		-0.0470***	
	(0.00572)		(0.00562)		(0.00488)		(0.00523)		(0.00509)	
Age 16-17		0.0464		0.0466*		0.00251		0.0419		-0.0346
		(0.0242)		(0.0226)		(0.0199)		(0.0214)		(0.0215)
Age 18-19		0.0583		0.0425		-0.0358		0.0687		-0.152***
		(0.0506)		(0.0454)		(0.0402)		(0.0423)		(0.0434)
Female*Age 16-17		-0.0621*		-0.0215		-0.174***		-0.139***		-0.0998***
		(0.0290)		(0.0285)		(0.0247)		(0.0263)		(0.0264)
Female*Age 18-19		-0.0245		0.0102		-0.350***		-0.378***		-0.168**
		(0.0715)		(0.0685)		(0.0528)		(0.0529)		(0.0595)
SCHOOL FIXED EFFECTS INCLUDED										
Observations	23671	23671	23671	23671	23671	23671	23671	23671	23671	23671
R^2	0.026	0.026	0.041	0.041	0.096	0.100	0.062	0.065	0.076	0.076

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Network characteristics are standardised and can be interpreted in terms of standard deviations, all regressions include school fixed effects, students below 12 and above 18 are dropped from the sample. The regression includes only schools which have a share of women between .49 and .51. Individuals with fewer than two friends are dropped from the sample.

Table 21: Differences in Degree: Add Health Sample All Degrees

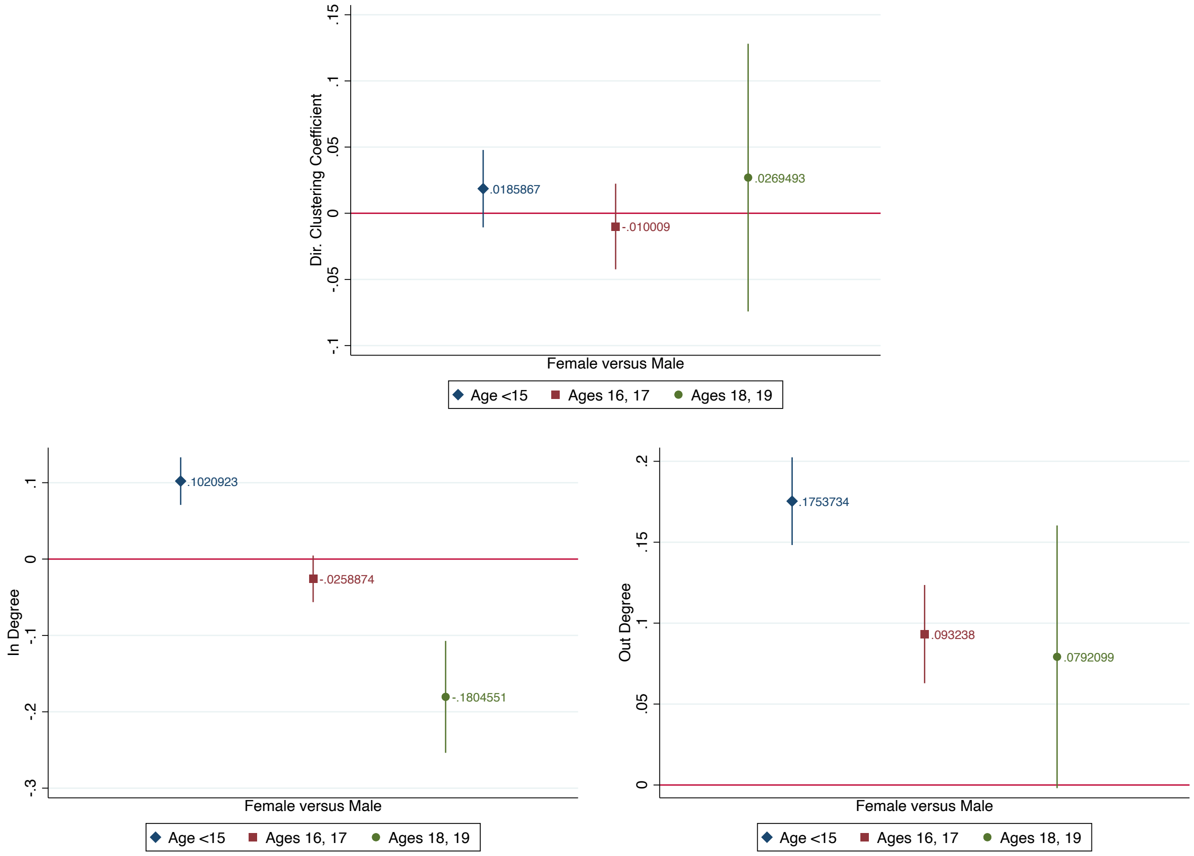
	Degree	Degree	In Degree	In Degree	Out Degree	Out Degree
Female	0.0892***	0.153***	0.0802***	0.139***	0.172***	0.204***
	(0.00661)	(0.00867)	(0.00682)	(0.00903)	(0.00673)	(0.00857)
Age	-0.0431***		-0.0140***		-0.0520***	
	(0.00265)		(0.00271)		(0.00274)	
Age 16-17		-0.00915		0.0510***		-0.0523***
		(0.0106)		(0.0109)		(0.0111)
Age 18-19		-0.160***		-0.0353		-0.243***
		(0.0201)		(0.0198)		(0.0211)
Female*Age 16-17		-0.144***		-0.131***		-0.0718***
		(0.0136)		(0.0141)		(0.0141)
Female*Age 18-19		-0.237***		-0.266***		-0.109***
		(0.0280)		(0.0267)		(0.0305)
	SCHOOL FIXED EFFECTS INCLUDED					
Observations	80333	80333	80333	80333	80333	80333
R^2	0.160	0.163	0.105	0.107	0.130	0.131

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Network characteristics are standardized and can be interpreted in terms of standard deviations, all regressions include school fixed effects. This sample also contains individuals with fewer than 2 connections.

Figure 3: Clustering Coefficient and Degree Across Ages



Note: Coefficients on the indicator female of a regression of dependent variable on constant, age, school fixed effects and the female dummy for ages 10-15, 16/17 and 18/19, respectively. Dot gives the coefficient, line depicts 99% confidence interval.

References

Aral, S., E. Brynjolfsson, and M. W. Van Alstyne (2012, September). Information, Technology and Information Worker Productivity. *Information Systems Research 23*(3), 849–867.

Arrow, K. and R. Borzekowski (2004). Limited Network Connections and the Distribution of Wages. *FEDS Working Paper No. 2004-41*.

Babcock, L., M. P. Recalde, L. Vesterlund, and L. Weingart (2017, March). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review 107*(3), 714–47.

Beaman, L., N. Keleher, and J. Magruder (2018). Do job networks disadvantage women? Evidence from a recruitment experiment in Malawi. *Journal of Labor Economics 36*(1), 121–157.

Benenson, J. F. (1990). Gender differences in social networks. *The Journal of Early Adolescence 10*(4), 472–495.

Benenson, J. F. (1993). Greater preference among females than males for dyadic interaction in early childhood. *Child Development 64*(2), 544–555.

Bertrand, M., C. Goldin, and L. F. Katz (2010). Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors. *American Economic Journal: Applied Economics*, 228–255.

Burt, R. (1992). *Structural Holes: The Social Structure of Competition.* Harvard University Press.

Calvó-Armengol, A. and M. Jackson (2004). Effects of Social Networks on Employment and Inequality. *The American Economic Review 94*(3), 426–454.

Calvó-Armengol, A. and M. Jackson (2007). Networks in Labor Markets: Wage and Employment Dynamics and Inequality. *Journal of Economic Theory 132*(1), 27–46.

Calvó-Armengol, A. and Y. Zenou (2005). Job matching, social network and word-of-mouth communication. *Journal of Urban Economics 57*(3), 500–522.

Coleman, J. (1988). Free riders and Zealots: The Role of Social Networks. *Sociological Theory 6*(1), 52–57.

Cooper, R., D. V. DeJong, R. Forsythe, and T. Ross (1990). Selection criteria in coordination games: Some experimental results. *American Economic Review 80*(1), 218–233.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1992). Communication in coordination games. *The Quarterly Journal of Economics 107*(2), 739–771.

David-Barrett, T., A. Rotkirch, J. Carney, I. B. Izquierdo, J. A. Krems, D. Townley, E. McDaniell, A. Byrne-Smith, and R. I. Dunbar (2015). Women favour dyadic relationships, but men prefer clubs: cross-cultural evidence from social networking. *PloS one 10*(3), e0118329.

Ding, W. W., F. Murray, and T. E. Stuart (2006). Gender differences in patenting in the academic life sciences. *Science 313*(5787), 665–667.

Dixit, A. (2003). Trade expansion and Contract Enforcement. *Journal of Political Economy 111*(6), 1293–1317.

Dohmen, T. and A. Falk (2011, April). Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. *American Economic Review 101*(2), 556–90.

Ductor, L., M. Fafchamps, S. Goyal, and M. J. van der Leij (2014). Social networks and research output. *Review of Economics and Statistics 96*(5), 936–948.

Ductor, L., S. Goyal, and A. Prummer (2018). Gender and collaboration. Working Paper.

Easley, D. and J. Kleinberg (2010). *Networks, Crowds, and Markets.* Cambridge Univ Press.

Eder, D. and M. Hallinan (1978). Sex Differences in Children's Friendships. *American Sociological Review*, 237–250.

Erosa, Fuster, Kambourov, and Rogerson (2017). Hours, occupations, and gender differences in labor market outcomes. *NBER Working Paper 23636.*

Fernandez, R. M. and M. L. Sosa (2005). Gendering the job: Networks and recruitment at a call center. *American Journal of Sociology 111*(3), 859–904.

Ferriani, S., G. Cattani, and C. Baden-Fuller (2009). The relational antecedents of project-entrepreneurship: Network centrality, team composition and project performance. *Research Policy 10*(38), 1545–1558.

Friebel, G., M. Lalanne, B. Richter, P. Schwardmann, and P. Seabright (2017). Women form social networks more selectively and less opportunistically than men. Working Paper.

Friebel, G. and P. Seabright (2011). Do women have longer conversations? Telephone evidence of gendered communication strategies. *Journal of Economic Psychology 32*(3), 348–356.

Gabbay, S. M. and E. W. Zuckerman (1998). Social capital and opportunity in corporate r&d: The contingent effect of contact density on mobility expectations. *Social Science Research 27*(2), 189 – 217.

Gayle, G.-L. and L. Golan (2011, 09). Estimating a Dynamic Adverse-Selection Model: Labour-Force Experience and the Changing Gender Earnings Gap 1968–1997. *The Review of Economic Studies 79*(1), 227–267.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics 118*(3), 1049–1074.

Goldin, C. and C. Rouse (2000, September). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review 90*(4), 715–741.

Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology*, 1360–1380.

Heider, F. (1946). Attitudes and Cognitive Organization. *The Journal of Psychology 21*(1), 107–112.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences 102*(46), 16569–16572.

Holmstrom, B. (1982). Moral Hazard in Teams. *The Bell Journal of Economics*, 324–340.

Hunt, J., J.-P. Garant, H. Herman, and D. J. Munroe (2012, March). Why Don't Women Patent? NBER Working Papers 17888, National Bureau of Economic Research, Inc.

Jackson, M. O. and B. W. Rogers (2007). Meeting strangers and friends of friends: How random are social networks? *The American Economic Review 97*(3), 890–915.

Karlan, D., M. Mobius, T. Rosenblat, and A. Szeidl (2009). Trust and Social Collateral. *The Quarterly Journal of Economics 124*(3), 1307–1361.

Lazear, E. P. and P. Oyer (2007). Personnel economics. Technical report, National Bureau of economic research.

Lee, L., C. Howes, and B. Chamberlain (2007). Ethnic heterogeneity of social networks and cross-ethnic friendships of elementary school boys and girls. *Merrill-Palmer Quarterly 53*(3), 325–346.

Lin, N. (1999). Building a network theory of social capital. *Connections 22*(1), 28–51.

Lutter, M. (2012). Anstieg oder Ausgleich? Die multiplikative Wirkung sozialer Ungleichheiten auf dem Arbeitsmarkt für Filmschauspieler. *Zeitschrift für Soziologie* (41), 435–457.

Lutter, M. (2013). Is there a Closure Penalty? Cohesive Network Structures, Diversity, and Gender Inequalities in Career Advancement. *MPIfG Discussion Paper* (13/9).

Mailath, G. J. and L. Samuelson (2006). *Repeated games and reputations: long-run relationships.* Oxford University Press.

Marsden, P. V. and E. H. Gorman (2001). Social networks, job changes, and recruitment. In *Sourcebook of Labor Markets*, pp. 467–502. Springer.

Mengel, F. (2020). Gender differences in networking. *The Economic Journal*.

Montgomery, J. D. (1991). Social networks and labor-market outcomes: Toward an economic analysis. *The American Economic Review 81*(5), 1408–1418.

Niederle, M. and L. Vesterlund (2007). Do Women Shy Away From Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

Parker, J. G. and J. Seal (1996). Forming, losing, renewing, and replacing friendships: Applying temporal parameters to the assessment of children's friendship experiences. *Child Development 67*(5), 2248–2268.

Putnam, R. (2000). *Bowling Alone: America's Declining Social Capital.* Simon and Schuster.

Torres, L. and M. L. Huffman (2002). Social networks and job search outcomes among male and female professional, technical, and managerial workers. *Sociological Focus 35*(1), 25–42.

Van Huyck, J. B., R. C. Battalio, and R. O. Beil (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review 80*(1), 234–248.

Williams, D. (1991). *Probability with martingales.* Cambridge University Press.

Zeltzer, D. (2020). Gender Homophily in Referral Networks: Consequences for the Medicare Physician Earnings Gap. *American Economic Journal: Applied Economics 12*(2), 169–97.

Zhu, M. (2018). Job networks through college classmates: Effects of referrals for men and women. Working Paper.

Network Structure and Performance: Online Appendix

Ilse Lindenlaub Anja Prummer

January 14, 2020

1 Theory

Selection Probability s_i. We demonstrate the robustness of our results by calculating the wage conditional on being selected. This is only relevant for the wages in the second period. To see this note that effort and wages in the first period are calculated conditional on being selected. For the *effort* in the second period, the probability of being selected does not matter per se, but only in combination with the share of common friends r_{ij}, that is the second period effort only depends on $s_i r_{ij} = sr$. However, the probability of being selected matters for *wages* in the second period, defined as $w_i'(\theta, \theta') \equiv s_i P_i(\gamma_g'|e(y), r)\mathbb{E}[f\,(e'(y'), e'(y'))\, v'|\theta']$. We now show that our results do not change if we consider the second period wage conditional on being selected, which we also refer to wages per project:

$$\omega_i'(\theta, \theta') \equiv P_i(\gamma_g'|e(y), r)\mathbb{E}[f\left(e'(y'), e'(y')\right) v'|\theta'] \tag{1}$$

The effect of information on second period wages is unchanged, as s_i is not affected by additional signals. Having common friends also does not have an impact on s_i, and so our results hold also for the wages per project.

Our result on wage dynamics also remains unchanged. We restate the result here.

Remark 1. *Wage Dynamics: If a C-worker has a weakly lower first period wage than a D-worker, then he also expects a lower wage per project in the second period.*

Proof. Claim: $w^D(\theta) \geq w^C(\theta) \quad \Rightarrow \quad \mathbb{E}[\omega'^D] > \mathbb{E}[\omega'^C]$.

As in the Proof of Proposition 5, we define

$$\mathbb{E}[\omega'] = q\omega'(\theta, \theta_h') + (1-q)\omega'(\theta, \theta_l') = f(1,1)P_i(\gamma_g'|\theta)\left(q\mathbb{E}[e'(y')|\theta_h']v_h + (1-q)\mathbb{E}[e'(y')|\theta_l']v_l\right),$$

which, using the new definition (1), no longer depends on s_i. As before, $P(\gamma_g'|\theta) \equiv \mathbb{E}[f\,(e(y), e(y)) + (1-r)(1 - f\,(e(y), e(y)))|\theta] = \mathbb{E}[e(y)|\theta]rf(1,1) + 1 - r$.

Again, by assumption, in the first period $w^D(\theta) \geq w^C(\theta)$, implying $\mathbb{E}[e(y)^D|\theta] \geq \mathbb{E}[e(y)^C|\theta]$. Moreover, $P(\gamma_g|\theta)^D > P(\gamma_g|\theta)^C$ since

$$[P(\gamma_g|\theta)]^D = r^D(\mathbb{E}[e(y)^D|\theta]f(1,1) - 1) + 1 > [P(\gamma_g|\theta)]^C = r^C(\mathbb{E}[e(y)^C|\theta]f(1,1) - 1) + 1$$

where the expression in brackets, $\mathbb{E}[e(y)|\theta]f(1,1) - 1$, is negative but (weakly) less so for the D-worker, exactly as before. The remainder of the proof of Proposition 5 applies as is. □

This demonstrates that our current definition of wages (according to which they depend on the probability of being selected) does not affect our results.

Infinite Horizon We extend our setting to allow for an infinite horizon. If a project fails in t, then relationships in the next period $t+1$ are bad with the current project partner and common friends. In $t+2$, all relationships are good again. Note that links are never cut.

Consider first the problem if an agent is in a bad state, that is he is paired with someone he has a bad relationship with. In the current period his payoff is zero, but in the following period he has a good relationship with everyone again. Denote the expected present value of a good relationship by $\mathbb{E}W_i(\gamma_g)$. In addition to the usual expectation regarding signals, the expectation is also taken over all potential pairings of partners as it can make a difference of whether i is matched with j or k even though he has a connection with both. Thus, agent i's discounted present value in the current period conditional on being matched with someone with whom the relationship is bad is given by:

$$W_{ij}(\gamma_b) = 0 + \beta s_i \mathbb{E}W_i(\gamma_g) \tag{2}$$

Note that this expression does not depend on who the partner is, beyond having a bad relationship with him, and further, that it is a constant. Therefore, $W_{ij}(\gamma_b) = W_i(\gamma_b) = \mathbb{E}W_i(\gamma_b)$.

The value of the problem in the good state if partners i and j are selected is then given by

$$W_{ij}(\gamma_g) = V_{ij}^*(\gamma_g) + \beta s_i \left(P_i(\gamma_g')\mathbb{E}W_i(\gamma_g') + (1 - P_i(\gamma_g'))\mathbb{E}W_i(\gamma_b')\right), \tag{3}$$

where $V_{ij}^*(\gamma_g)$ is the current value of the problem given that the team is in a good state and $P_i(\gamma_g')$ is the probability of a good state in the next period, as defined on p. 10, expression (5). We can simplify this expression to

$$\begin{aligned} W_{ij}(\gamma_g) &= V_{ij}^*(\gamma_g) + \beta s_i \left(P_i(\gamma_g')\mathbb{E}W_i(\gamma_g') + (1 - P_i(\gamma_g'))[0 + \beta s_i \mathbb{E}W_i(\gamma_g'')]\right) \\ &= V_i^*(\gamma_g) + \beta s_i P_i(\gamma_g')\mathbb{E}W_i(\gamma_g') + \beta^2 s_i^2 (1 - P_i(\gamma_g'))\mathbb{E}W_i(\gamma_g''), \end{aligned} \tag{4}$$

where double prime denotes two periods ahead. Noting that $EW_i(\gamma_g') = EW_i(\gamma_g'')$, we can simplify further to

$$W_{ij}(\gamma_g) = V_i^*(\gamma_g) + \beta(1-\beta s_i)s_i P_i(\gamma_g')\mathbb{E}W_i(\gamma_g') + \beta^2 s_i^2 \mathbb{E}W_i(\gamma_g') \tag{5}$$

This setting differs in some aspects, but the key dependence of the payoff on degree and clustering is the same as before. Note that clustering enters through $s_i P_i(\gamma_g')$, which is the same as in the main text. Therefore, higher clustering still yields higher effort. Similarly, $\mathbb{E}W_i(\gamma_g)$ depends on the number of signals. Note that $\mathbb{E}W_i(\gamma_g)$ is a weighted average of value of the problem for i with different agents and we have shown that the value of the problem for each team depends positively of the number of signals.

However, in this setting the probability of being selected matters independently of the clustering, so the symmetry of effort choices among project partners is broken. If an agent has a higher probability of being selected then his payoff is higher and therefore, he would exert higher effort. The baseline set up in the paper allows us to focus on the trade off between peer pressure and information, whereas here degree comes with an additional advantage–that of being selected for more projects in the future.

In sum, in the infinite horizon setting our two key forces – clustering and information – are still crucial determinants of effort. However, the trade-off we highlight is cleaner in our baseline model.

2 Data

2.1 Add Health: Robustness

AddHealth network measures with gender balance measured at the school level as control

Table 1: Controlling for Gender Balance

	Cl. Coeff. (dir)	Cl. Coeff. (dir)	Cl. Coeff.	Cl. Coeff.	Degree	Degree	In Degree	In Degree	Out Degree	Out Degree
Female	-0.0193**	0.00744	0.0657***	0.0897***	0.0526***	0.118***	0.0515***	0.124***	0.144***	0.172***
	(0.00742)	(0.00945)	(0.00741)	(0.00940)	(0.00682)	(0.00916)	(0.00734)	(0.00993)	(0.00693)	(0.00898)
Age	-0.0105***		-0.00788***		-0.0328***		-0.00733***		-0.0330***	
	(0.00221)		(0.00222)		(0.00199)		(0.00213)		(0.00202)	
Age 16-17		-0.00873		-0.00195		-0.00945		0.0623***		-0.0427***
		(0.0119)		(0.0114)		(0.0106)		(0.0114)		(0.0111)
Age 18-19		-0.00755		-0.0404		-0.138***		-0.0128		-0.219***
		(0.0257)		(0.0234)		(0.0208)		(0.0219)		(0.0223)
Gender Balance		-0.731***		-1.324***		1.313***		0.727***		1.015***
		(0.0734)		(0.0809)		(0.0307)		(0.0361)		(0.0336)
Female* Age 16-17		-0.0270		0.00959		-0.150***		-0.138***		-0.0732***
		(0.0157)		(0.0157)		(0.0142)		(0.0154)		(0.0147)
Female* Age 18-19		0.0102		0.0127		-0.239***		-0.291***		-0.0874**
		(0.0398)		(0.0376)		(0.0294)		(0.0301)		(0.0327)
Female* Balance		-0.193		-0.0841		-3.144***		-3.024***		-2.215***
		(0.180)		(0.186)		(0.133)		(0.139)		(0.139)
Observations	73244	73244	73244	73244	73244	73244	73244	73244	73244	73244
R^2	0.000	0.003	0.001	0.010	0.005	0.018	0.001	0.009	0.010	0.017

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Network characteristics are standardized and can be interpreted in terms of standard deviations. This sample contains individuals with more than 2 connections. Gender Balance is calculated as share of women minus .5, gender parity. Gender balance is positive if there are more women at a school, negative otherwise.

2.2 Census Data: Robustness

Earnings Regression with industry and occupational fixed effects.

Table 2: Log Earnings, Risk and Gender

	(1) Log Earnings	(2) Log Earnings
Female	-0.323*** (0.000718)	0.0601*** (0.00460)
Risk		4.255*** (0.0144)
Female*Risk		-0.712*** (0.00879)
Years of Educ	0.0846*** (0.000180)	0.0842*** (0.000180)
Experience	0.0258*** (0.000127)	0.0256*** (0.000127)
Experience^2	-0.000362*** (0.00000257)	-0.000361*** (0.00000257)
Black	-0.0437*** (0.00134)	-0.0438*** (0.00133)
White	0.0382*** (0.00103)	0.0380*** (0.00103)
Constant	9.030*** (0.00550)	6.362*** (0.00916)
Observations	3558758	3558758
R^2	0.332	0.334

Sample: 2000 US Census, Full-time Workers. Estimation by OLS.
Controls that are included but not reported: Industry and occupation fixed effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Occupational choice with industry fixed (ordered probit).

Table 3: Occupational Risk and Gender

	(1) Risk
Female	-0.198*** (0.00125)
Years of Educ	0.000240 (0.000267)
Experience	-0.00153*** (0.000237)
Experience^2	0.0000574*** (0.00000465)
Black	0.0234*** (0.00254)
White	0.00884*** (0.00190)
Log Occ Mean Earnings	1.861*** (0.00248)
Observations	3558758
Pseudo R^2	0.072

Sample: 2000 US Census, Full-time workers. Estimation method: ordered probit. Controls that are included but not reported: Industry fixed effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Occupational choice (linear probability model).

Table 4: Occupational Risk and Gender

	(1) Risk (binary)
Female	-0.0507***
	(0.000504)
Years of Educ	-0.0176***
	(0.000121)
Experience	-0.000833***
	(0.000110)
Experience^2	0.0000121***
	(0.00000216)
Black	-0.0149***
	(0.00124)
White	-0.00152
	(0.000936)
Log Occ Mean Earnings	0.492***
	(0.000942)
Constant	-4.405***
	(0.00962)
Observations	3558758
R^2	0.094

Sample: 2000 US Census, Full-time Workers. Estimation method: Linear probability model. Binary outcome is occupational risk below and above median.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Occupational choice with industry fixed (linear probability model).

Table 5: Occupational Risk and Gender

	(1) Risk (binary)
Female	-0.0542***
	(0.000508)
Years of Educ	-0.00644***
	(0.000117)
Experience	0.000746***
	(0.000100)
Experience^2	0.00000325
	(0.00000198)
Black	0.0139***
	(0.00113)
White	0.00237***
	(0.000838)
Log Occ Mean Earnings	0.525***
	(0.000876)
Constant	-4.697***
	(0.00940)
Observations	3558758
R^2	0.252

Sample: 2000 US Census, Full-time workers. Estimation method: Linear probability model.
Binary outcome is occupational risk below and above median.
Controls that are included but not reported: Industry fixed effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Computer Science: Robustness

Table 6: Summary Statistics: Computer Scientists (Google Scholar Sample)

	Min	Max	Mean	Std
Degree	2.00	95.00	10.19	12.54
Clustering Coefficient	0.00	0.25	0.14	0.07
Observations		25,428		

Note: Sample consists of 15,827 men and 9,601 women. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Table 7: Correlation of Network Measures: Computer Science (Google Scholar Sample)

	Degree	Clustering Coefficient
Degree	1	
Clustering Coefficient	-0.54***	1

t-statistics in parenthesis. ***p<0.01, **p<0.05, *p<0.1.
Note: Sample consists of 15,827 men and 9,601 women. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.

Table 8: Network Measures by Gender: Computer Science (Google Scholar Sample)

	Male	Female	Difference
Degree	10.6618	9.4204	1.2415***
			(7.8320)
Clustering Coefficient	0.1339	0.1408	-0.0069***
			(-7.7570)
Observations		25,428	

t-statistics in parenthesis. ***p<0.01, **p<0.05, *p<0.1.
Note: Sample consists of 15,827 men and 9,601 women. This sample does not contain individuals for whom the clustering coefficient is not defined, i.e. those with fewer than 2 links.