Strategic Data Access Management

George Charlson
Janeway Institute, Cambridge, United Kingdom, gc556@cam.ac.uk

Ruslan Momot
Ross School of Business, University of Michigan, United States, momot@umich.edu

Marat Salikhov
New Economic School and Moscow School of Management Skolkovo, Russia, msalikhov@nes.ru

December 5, 2022

A recent survey showed that 33% of businesses grant their employees access to all company data, with at least another 35% granting accesses to more data than is needed. Such overly permissive *data access strategy* allows the firms to run more efficiently, but at the same time, such strategies present growing cybersecurity risks. With work-from-home becoming more popular, remote employees are being increasingly exploited by the malicious adversaries to gain access to their organizations' data. To address this issue, we investigate the optimal design of *data access architectures – who should have access to what data*. Our economic model captures a firm managing a set of employees and a set of datasets. For each employee the firm chooses which datasets this employee should have access to. An employee may be attacked by a potentially sophisticated adversary whose goal is to steal all their data. Therefore, the firm trades off the efficiency benefit of the more permissive data access architecture with the adversarial risk it incurs. We characterize the firm's optimal data access architecture and investigate how it depends both on the adversarial environment and the firm's technology.

Key words: cybersecurity strategy, data access management, bipartite graphs.

1. Introduction

A recent survey showed that "*[33% of businesses] give [their] employees access to all company data. Another 35% [give] access to more data than is strictly necessary to perform their job (but not all company data)*" (GetApp 2022). Naturally, allowing employees to have access to more of company data can be help the company run more efficiently.[1] At the same time, such overly permissive data access policies present growing cybersecurity risks. This is because "*knowingly or unknowingly, the majority of cyber incidents are caused by an employee of the impacted organization*" (Deloitte 2020) and 50% of the data breaches among 7,385 recently studied by McKinsey & Company (Bailey et al. 2018) "*had a substantial insider [company employee] component*". Such insider threats via companies' own employees became especially pronounced with the start of COVID-19 pandemic

[1] "4 Reasons to Give Your Employees Instant Access to Business Information", *CIO*, Mar 2, 2018.

when the majority of businesses had to adopt work from home. The move to remote work explosively increased the number of access points (remote employees and their devices) that the adversaries eager to get to the company data can exploit. Together with the rapidly rising number of data breaches as well as the costs of these breaches reaching decades-high \$9.4M levels per incident (IBM Security 2022), these factors define a new cybersecurity landscape. Today, companies can not simply focus on "*fortifying the [physical] perimeter to deter outsiders from accessing corporate data, while implicitly trusting insiders*" (Deloitte 2021), but rather have to systematically control *who has access to what data*. The design of such *data access architectures* is the focus of this paper.

A standard recommendation (Stallings and Brown 2018, Section 4.5) is to ensure that employees only have access to the data required to fulfil their functions. However, this recommendation does not specify how these data requirements are to be determined. For example: should an organization have many flexible roles who use different datasets or should it rely on employees specializing in working with just a few datasets? Should some employees be privileged in having access to more data or should the organization be flat in that regard? We show that such questions should be answered by taking into account both operational and cybersecurity considerations, which we represent in a game-theoretic model. Our model features a firm and an adversary. The firm manages a number of employees and datasets. The firm's production function – *technology* – is such that, in the absence of a cybersecurity risk, it is in the firm's interest to grant employees access to as many datasets as possible, because making an extra dataset available to an employee is always economically beneficial for the firm (e.g., consider a data scientist who requires access to the historical data in order to estimate the demand and optimize a product ordering decisions). The matches between different datasets and different employees may differ in value representing, for example, the employees' talent for extracting value from data. The firm faces an adversary who is trying to obtain access to as many datasets as possible and it does so via attacking employees. In particular, the adversary chooses an employee to attack and steals all the datasets that this employee is granted access to. The adversary can be one of the two types (unknown to the firm) – *strategic* or *opportunistic*.[2] These types represent adversary's *sophistication* (e.g., see Fig. 1 in Deloitte 2017): an opportunistic adversary targets all employees with fixed probabilities – assumed to be equal in our model (e.g., via a malware attack), whereas a strategic adversary picks the target deliberately by taking into account the firm's data access allocation (e.g., via a well-researched targeted attack). When such adversarial risk is present, the firm faces a non-trivial decision on its *data access architecture*—a bipartite graph that encodes which employees have access to which datasets. This decision is non-trivial because the firm has to carefully balance the "production" benefit of a more permissive data access architecture against the adversarial risk it incurs.

[2] https://www.seic.com/cyber-protection/sphere-blog/opportunistic-vs-targeted-cyberattacks.

We first characterize the firm's optimal data access architecture. We find that a commonly implemented laissez-faire approach of giving employees access to every dataset is only optimal when the *strength of adversarial attacks*—measured by the damage the latter incur—is low. Otherwise the firm must limit access to data to some of its employees. To say more about the structural properties of the optimal data access architecture strategies, and explain them in intuitive terms, we describe the distributions over bipartite graphs that represent these strategies using the two fundamental summary statistics: the *average data access level* – the expected number of datasets given to each employee, which characterizes the overall level of data access in a company and *data access inequality* – a measure of disparity in data access across employees, which characterizes the firm's selectivity with which it picks what data to give to its employees. We investigate how these statistics of the optimal architecture depend on the factors *external* to the firm, namely, the adversary's strength and sophistication; as well as the *internal* ones: the distribution of the employee talent, and the firm's technology. We find that as the adversary's strength increases, the average data access level is decreasing while the access inequality is inverse U-shaped (first increasing and then decreasing); an increase in the adversary's sophistication decreases the access inequality; and as the firm's talent disparity increases, the firm provides less data access and higher access inequality. Finally, we consider the role of the firm's technology focusing on two particular cases. First, vertically differentiated technology – all datasets and employees are objectively ordered in terms of their value. Second, horizontally differentiated technology – employees specialize on handling certain datasets. We show that when the technology is vertically differentiated and the adversary is sophisticated enough, the optimal data access architecture is hierarchical. When the technology is horizontally differentiated and the adversary is strong enough, the optimal architecture exhibits a widely-discussed "Zero-Trust" structure (Deloitte 2021).

2. Literature Review

The vast literature in Computer Science explores issues of data access management (e.g., see Bell and La Padula 1973, Kain and Landwehr 1987, Ferraiolo and Kuhn 2009). These studies, however, typically focus on designing sound (e.g., data leakage-proof) data access systems, rather than deriving optimal and economically-informed choice of the data allocation (i.e., architecture) under adversarial pressure, as we do in this paper.

Our setting also resembles structurally the "Colonel Blotto" games (Roberson 2006), in which two (typically symmetric) players decide on allocating limited resources among several battlefields. The player that dominates the largest number of battlefields wins. Bimpikis et al. (2016) recently applied similar ideas to the design of targeted advertising strategies, with several competing advertisers engaged in a Blotto-like game. While our model shares some features with such games, it differs in (i) the asymmetry between the attacker and the defender; (ii) the payoff structure, with the attacker being able to capture any desired location and the defender willing to minimize its losses at said location and (iii) modeling the resources (datasets in our case) as heterogeneous.

Our representation of the data access architecture choice as a networked structure (i.e., a bipartite graph) is in line with similar approaches in technology management (Gokpinar et al. 2010) and process management (Gurvich and Van Mieghem 2015, Dawande et al. 2021). One of the closest papers to ours, in terms of methodology, is Charlson (2020) which considers management of an online platform and represents the platform's "observation architecture"—which buyers can observe which sellers—as a distribution over bipartite graphs. We differ in that we focus on how the architecture choices (and therefore the firm's operations decisions) should be shaped by the adversarial pressure.

A classic setting in the operations literature that is similar to our problem is capacity flexibility (Jordan and Graves 1995, Bassamboo et al. 2012, Wang et al. 2015, Chan and Fearing 2019). Similarly to this setting, our firm has to pick a bipartite graph optimally, trading off the costs of a new link against its benefits. The main difference in our setting, however, is the endogeneity of risks because the adversary is strategic.

Finally, our paper contributes to a growing field of managerial studies in cybersecurity (Kannan and Telang 2005, Arora et al. 2008, August et al. 2019, 2022) and data privacy and management (see the review of this stream of literature in Fainmesser et al. 2022). To the best of our knowledge, no such paper ever addressed the question of strategic data access management.

3. The Model

3.1. Players and Timeline

Our model features two players: a *firm* and an *adversary*. The firm manages a set $\mathcal{M}$ of m *employees* and a set $\mathcal{N}$ of n *datasets*. The firm chooses its *data access architecture*: for each employee i, the firm decides which datasets this employee should be granted access to. Thus, data access architecture can be represented as a bipartite graph edges of which connect the set of employees with the set of datasets, access to which these employees have. This graph can be described using an adjacency matrix g of size $m \times n$ with the rows and columns corresponding to the employees and datasets, correspondingly. An element $g_{ij} \in \{0,1\}$ of this matrix is such that $g_{ij} = 1$ if an employee i is granted access to the dataset j. In the remainder of the paper, we will use adjacency matrix terminology interchangeably with that of the bipartite graph and data access architecture. Denote $\mathcal{G}_{mn}$ – the set of all bipartite graphs of size $m \times n$. Then the firm's (potentially mixed) strategy G is a probability distribution over the set $\mathcal{G}_{mn}$, i.e., $G \in \Delta(\mathcal{G}_{mn})$.

As alluded to in the Introduction, the *adversary* can be one of the two types: the *strategic* type $\theta = s$, with probability $\sigma_s = \sigma$ or the *opportunistic* type $\theta = o$, with probability $\sigma_o = 1 - \sigma$. The types can be thought of as representing adversary's *sophistication* – the higher is σ, the more targeted and thought-through is the attack (see Fig. 1 in Deloitte 2017). The firm does not observe the adversary's type. After the firm chose data access architecture, the adversary picks an employee to attack and then steals every dataset this employee has access to. The opportunistic adversary picks a target uniformly at random, while the strategic adversary carefully selects an employee to attack by playing the best-response to the firm's architecture strategy G. For each such architecture strategy G we denote by $X_\theta(G)$ the potentially mixed strategy of adversary with type θ. This mixed strategy is a probability distribution over the set $\mathcal{M}$ of all employees, i.e., $X_\theta(G) \in \Delta(\mathcal{M})$. We note, though, that the opportunistic adversary's strategy is simply attacking each employee with probability $1/m$.

We denote by g and x the realizations of the firm's data access architecture (i.e., the realized bipartite graph described by an adjacency matrix g) and the adversary's target choice, correspondingly. After g and x are realized, the players receive their payoffs as described next.

3.2. Payoffs

We do not specify the opportunistic adversary's payoff since it chooses an employee to attack uniformly at random.

The strategic adversary's payoff depends on the realized architecture and its choice of the target. In particular, if the adversary attacked employee x in the architecture g, its realized payoff is $d_x(g)^2$, where $d_x(g) = \sum_{j\in\mathcal{N}} g_{xj}$ is the number of datasets employee x is granted access to (we will also call this number the employee's *degree*, following standard graph theory terminology). This quadratic specification can be justified either because the adversary may experience increasing returns from combining multiple datasets or it can be micro-founded through a stochastic model of selecting effort to obtain access to the datasets of the targeted employee. For example, suppose that once the adversary selects a target x, they exert effort $e \in [0,1]$ at a cost $e^2/2$ to obtain with probability e access to $d_x(g)$ datasets belonging to the employee x. Therefore, the adversary maximizes his expected payoff $ed_x(g) - \frac{1}{2}e^2$, with the resulting maximum expected payoff at the optimal effort level $e^* = d_x(g)$ being proportional to $d_x(g)^2$.

The firm's payoff also depends on the realized data access architecture g and the adversary's target employee. In particular, if employee x were targeted, the firm's payoff is:

$$\pi(x, g; \mu, \gamma) = \sum_{i\in\mathcal{M}, j\in\mathcal{N}} \gamma_{ij} g_{ij} - \mu d_x(g)^2. \tag{1}$$

Here, parameter $\mu > 0$ is the *strength* of the cybersecurity risk (or the attack) and it measures the firm's loss as a multiple of the adversary's gain from an attack. The parameters $\gamma_{ij} > 0$ specify, for each employee i and each dataset j, the value that this employee brings to the firm when having access to the dataset j. The first part of the firm's payoff above can, thus, be thought of as a production function of the firm defined on a bipartite graph g connecting employees with the datasets. Vector $\boldsymbol{\gamma}$ then describes how connections between employees and datasets generate value to the firm. We will call this vector *productivity vector* or the firm's *technology* interchangeably.

3.3. Equilibrium

We consider the subgame perfect Nash equilibria of the above game. The opportunistic adversary's equilibrium strategy $X_o^*(\cdot)$ always picks a target at random, thus, with slight abuse of notation, we omit from our notation for an equilibrium. The pair of strategies of the firm and strategic adversary, $(G^*, X_s^*(\cdot))$ is an equilibrium if the following two conditions hold:

1. The strategic adversary's equilibrium attack strategy $X_s^*(\cdot)$ is a best response to the firm's architecture choice G:
$$X_s^*(G) = \arg\max_X \mathbb{E}_X \left[\mathbb{E}_G \left[d_X(G)^2\right]\right]. \tag{2}$$
2. The platform's optimal architecture strategy G^* is a best response to the adversarial equilibrium attack strategies $X_o^*(\cdot)$ and $X_s^*(\cdot)$:
$$G^* = \arg\max_G \sum_{\theta\in\{s,o\}} \sigma_\theta \cdot \mathbb{E}_{G, X_\theta^*(G)} \left[\pi(G, X_\theta^*(G); \mu, \boldsymbol{\gamma})\right]. \tag{3}$$

To address the potential multiplicity of equilibria, in all the subsequent results we will focus on the equilibrium that delivers the largest ex-ante expected payoff to the firm.

3.4. Summary Statistics of a Data Access Architecture Strategy

To capture the essential properties of the data access architecture, we describe the bipartite graph representing such architecture using two types of summary statistics: (i) the employee-specific statistics that allow us to understand the data access patterns of individual employees and (ii) the aggregate statistics that characterize the firm's data access architecture as a whole.

We consider the following employee-specific summary statistics: the *expected degree* $\mathbb{E}_G[d_i(G)]$, capturing the number of datasets assigned to an employee i, and the *expected target attractiveness* $\mathbb{E}_G[d_i^2(G)]$, capturing the attractiveness of employee i as a target for the adversary. Understanding the behaviour of these summary statistics is helpful since it allows us to understand whether, for example, more productive employees should be granted access to more datasets.

Perhaps more importantly, we define the following aggregate statistics: the *average access level* $\bar{d}(G) = \mathbb{E}_G\left[\frac{1}{m}\sum_{i\in\mathcal{M}} d_i(G)\right]$, representing the average number of datasets assigned to all employees, and the *data access inequality* $w(G) = \mathbb{E}_G\left[\sum_{i,i'\in\mathcal{M}}(d_i(G) - d_{i'}(G))^2\right]$, capturing the average disparity in data access among employees. While the average access level is quite an intuitive metric, the data access inequality requires a slightly more detailed explanation. If data access inequality is high, then some employees obtain access to a large number of datasets while others only have access to a few of those (see Fig. 1 (a) for an illustration of one such architecture). On the other hand, when data access inequality is low, all employees have access to almost the same amount of datasets (see Fig. 1 (b) and (c)). These statistics are helpful because they can be thought of as the fundamental *instruments* that the firm employs to design optimal architectures. The firm may increase or decrease the average data access level for everyone in the organization as well as distribute access to data in equal or unequal fashion among the employees.

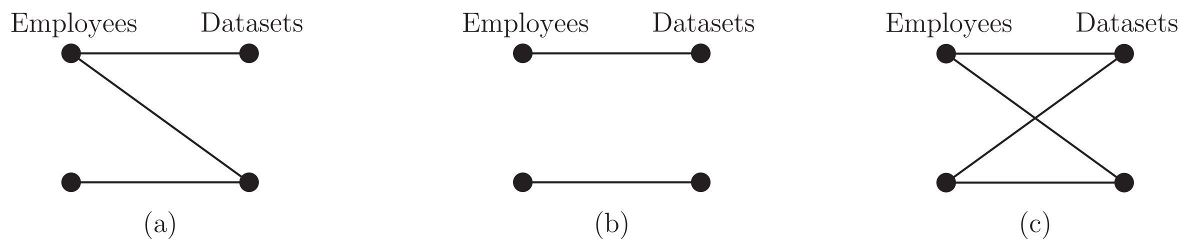

Figure 1 Examples of data access architectures on the sets of two employees and two datasets with **(a)** high data access inequality and moderate average data access level, **(b)** low data access inequality and low average data access level, **(c)** low data access inequality and high average data access level.

4. Optimal Design of Data Access Architecture

Our first result characterizes equilibrium of our game – the firm's data access architecture strategy, G^* and the strategic adversary's attack strategy, $X_s^*(\cdot)$ which is contingent on the firm's strategy. As we noted before, the opportunistic adversary's equilibrium strategy $X_o^*(\cdot)$ picks a target uniformly at random and is therefore omitted.

PROPOSITION 1. *There always exists an equilibrium* $(G^*, X_s^*(\cdot))$ *where architecture strategy* G^* *and attack strategy* $X_s^*(\cdot)$ *solve Eq.* (2), (3). *Furthermore, there exists a* $\bar{\mu} > 0$ *such that if adversary's strength is low,* $\mu \leq \bar{\mu}$ *then the firm grants all employees full data access with probability 1.*

Proof. Proofs of all the results are provided in the Online Appendices. □

This proposition shows that an equilibrium of the game always exists and any such equilibrium (of which there are potentially multiple) exhibits a set of basic properties. First, regardless of the adversary's sophistication, σ the firm's optimal architecture grants all employees access to all company data if the adversary's strength, μ is sufficiently small. If the latter is *not* the case (i.e., if μ is high), then at least some employees will not be guaranteed (with probability 1) to have access to all data. The second property underlines the importance of optimally choosing the data access architecture depending on the adversarial environment. As we mentioned in the introduction at minimum 33% of the businesses rely on the laissez-faire approach and grant their employees access to all data. The above proposition shows that such strategy is only optimal when the adversary is sufficiently weak. If this is not the case, this strategy may lead to suboptimal levels of adversarial damage inflicted on the company.

The above results follow from our specification of the firm's profit function (1). The first term of the profit function—production—increases linearly in any given employee's degree. However, the second term of the profit function—the adversarial damage—is quadratic and decreasing. Therefore, if the adversary's strength is sufficiently low, the linear term dominates and the full-access architecture is optimal. However, when the adversary's strength is sufficiently high, the firm optimally assigns higher probabilities to architectures that deny full data access to some employees because in this case the quadratic term dominates.

The next sections explore how the optimal data access architecture depends on the factors external to the firm, namely, the adversary's strength and sophistication; as well as the internal ones: the distribution of the employee talent, and the firm's technology.

4.1. Adversary: Effects of Strength and Sophistication

In order to highlight and investigate the impact of the adversarial environment on the optimal data access architecture, we make a simplifying assumption that all datasets are exchangeable (or homogeneous): i.e., $\gamma_{ij} = \gamma_i$ for all employees i and datasets j. We will relax this assumption in Section 4.3 when we study the impact of the heterogeneity in the value of the datasets on the equilibrium. Without further loss of generality, we will also assume that the vector $\boldsymbol{\gamma}$ is ordered such that $\gamma_i > \gamma_j$ if $i < j$. The following Proposition gives an account of how the distribution of degrees and damage change with attack strength μ and adversary sophistication, σ.

PROPOSITION 2. *For any employee i, the expected degree, $\mathbb{E}_G[d_i(G)]$, and target attractiveness, $\mathbb{E}_G[d_i^2(G)]$, are decreasing in adversary's strength, μ. Furthermore, with respect to adversary's sophistication, σ these measures are increasing for the most productive employee 1, decreasing for the least productive employee m, and first increasing and then decreasing for any other employee.*

As the adversary's strength increases, the firm has a stronger incentive to reduce the adversarial damage. It does so by reducing the degree and, therefore, the target attractiveness of its employees. An increase in the adversary's sophistication σ has a more subtle effect. When the adversary is fully opportunistic (i.e., $\sigma = 0$), the firm gives the most data to the most productive employees. When the adversary becomes more sophisticated (i.e., σ increases), these highly productive employees, who have more data, begin to attract the attacks from the sophisticated adversary. Because of that, the firm has an incentive to give less data to these highly productive employees and more data to the less productive ones. Eventually, once the adversary's sophistication, σ becomes very high, the firm will provide the same amount of data to all of its employees. An increase in adversarial sophistication, then, does not imply that the firm should optimally reduce access rights uniformly across employees. In fact, rather counterintuitively, most employees should receive more access to data in this case: the exception is highly productive employees, who instead should receive less access to data.

The above analysis also has implications for the aggregate architecture summary statistics introduced in Section 3.4. The following theorem provides comparative statics of the architecture summary statistics with respect to the parameters of the adversarial environment: the attack strength, μ and adversary sophistication, σ.

THEOREM 1. *For any equilibrium $(G^*, X_s^*(\cdot))$, the average access level $\bar{d}(G^*)$, and firm profit, $\mathbb{E}_{G,X_\theta^*(G)}[\pi(G^*, X_\theta^*(G); \mu, \gamma)]$ are weakly decreasing in the adversary's strength, μ. Furthermore, access inequality $w(G^*)$ is first weakly increasing and then weakly decreasing in the adversary's strength, μ and is decreasing in the adversary's sophistication, σ.*

The first part of the above Theorem shows that the firm chooses to grant its employees access to, on average,[3] less datasets when the adversary's strength is higher. The second part shows that the data access inequality, which measures disparity in data access across employees, first increases then decreases with the adversary's strength, μ. As we saw in Proposition 1, when this parameter is small, the firm is able to give to all of its employees access to every dataset. Thus, there is no inequality among the employees in terms of their degrees (the number of the datasets they have access to). When μ is very high, the firm provides almost no access to data. Thus, again, all employees are almost equal with regard to the data access rights and the level of data access inequality is low. Finally, when μ is intermediate, the firm finds it optimal to limit data access first to the employees who are less productive (i.e., such employees i that have low γ_i), thus, resulting in a non-trivial level of access inequality. As the adversary's sophistication σ increases, the firm chooses to give less data to the more productive employees, as they are more likely to be attacked. As σ approaches 1, in equilibrium the firm gives each employee the same amount of data, as employees with access to more datasets will be attacked with too high a probability. As a result, the access inequality is decreasing in σ.

To sum up, the firm has several instruments to confront different adversarial environments. When the adversary becomes stronger, the firm responds by lowering the data access of its employees. The firm, however, does not always do that uniformly across the employees. Certain employees (particularly, the less productive ones) may obtain access to less data than others, resulting in more unequal data access architectures. Such data access inequality, however, subsides when the

[3] Recall that for the measure $\bar{d}(G^*)$ the average is taken over all employees and over all architectures that the firm's (potentially mixed) strategy G^* prescribes with positive probability.

adversary becomes so strong that the firm has to restrict data access severely to all of its employees. Figure 2 summarizes our findings by showing that as the adversary becomes stronger, the firm transitions from the "full-access" (every employee has access to all datasets) to "access-managed" (employees have access to a curated set of datasets), and, finally, to the "encapsulated" (each employee has access to very few datasets) data access architectures.

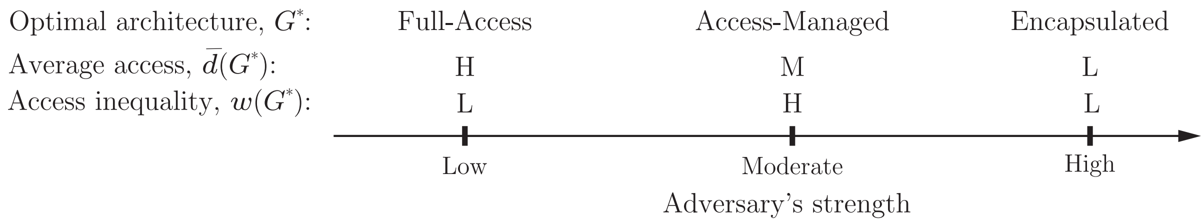

Figure 2 Optimal data access architecture and its aggregate summary statistics (see Section 3.4 for their definitions) depending on the adversary's strength, μ. Here, H = high, M = moderate, and L = low levels.

4.2. Employees: Effects of Talent Disparity

Companies differ not only in how productive or talented their employees are but also in how this talent is distributed among the employees. A recent report (McKinsey & Company 2017) shows that the gap in productivity between the top and the average performers within an organization varies from 50% to 800%, with the gap being higher for more "complex" occupations. In relationship to our study, talent or productivity of an employee lies in their ability to generate value to the firm when granted access to the datasets. This section investigates how distribution of employee talent affects firm's optimal data access architecture.

In order to keep our focus solely on the dimension of employee talent, we maintain the assumption that we made in Section 4.1 that the datasets are exchangeable (or homogeneous). As mentioned before, we will relax this assumption in the next section. With slight abuse of notation we use vector $\boldsymbol{\gamma}$ of size m (the number of employees) to denote value that each employee brings when granted access to one dataset (it does not matter to which one due to our assumption). Without loss of generality, we assume that all talent vectors $\boldsymbol{\gamma}$ are ordered with the first element being the highest. The following theorem describes how the three statistics of the optimal data access architecture depend on $\boldsymbol{\gamma}$ (we explicitly include this dependence in the notation of the equilibrium architecture $G^*(\boldsymbol{\gamma})$).

Theorem 2. *Consider an equilibrium architecture strategy G^* as a function of the employee productivity vector $\boldsymbol{\gamma}$. Assume that the adversary's strength μ is high enough to ensure that, in expectation, no employee is granted access to all datasets: $\mathbb{E}_{G^*}[d_i(G^*)] < n$ for all i. Then, the average access level $\bar{d}(G^*(\boldsymbol{\gamma}))$ is Schur-concave in $\boldsymbol{\gamma}$. Furthermore, there exists a $\bar{\sigma}$ such that if adversary's sophistication $\sigma \leq \bar{\sigma}$ then the access inequality $w(G^*(\boldsymbol{\gamma}))$ is Schur-convex in $\boldsymbol{\gamma}$.*

The implications of this theorem can be best understood by considering two firms with equal number of employees and datasets: Firm 1 and Firm 2. Let $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ be the corresponding employee talent vectors of these firms. As mentioned above, we will be interested in capturing the phenomenon highlighted in the McKinsey & Company (2017) survey – namely, that for some companies, in comparison to others, distribution of talent of their employees may be more non-uniform. Formally, we will assume that vector $\boldsymbol{\gamma}_2$ of Firm 2 *majorizes* vector $\boldsymbol{\gamma}_1$ of Firm 1, which, by definition of majorization, means that $\sum_{i=1}^{k}\gamma_{2i} \geq \sum_{i=1}^{k}\gamma_{1i}$ for all $k \in \{1,\ldots,m\}$ and $\sum_{i=1}^{m}\gamma_{2i} = \sum_{i=1}^{m}\gamma_{1i}$. Intuitively, Firm 2 then has a number of "star" (top) performing employees generating most of the value to this firm. In comparison, when granted access to the datasets, employees of Firm 1 generate closer to the the same value as one another and, thus, their distribution of talent is more uniform than that of Firm 2.

Schur-convexity (and concavity)[4] of the functions that we derived in the Theorem then directly implies that (i) $\bar{d}(G^*(\boldsymbol{\gamma_1})) \geq \bar{d}(G^*(\boldsymbol{\gamma_2}))$ – the average access level of Firm 1 is higher than that of Firm 2; (ii) $w(G^*(\boldsymbol{\gamma_2})) \geq w(G^*(\boldsymbol{\gamma_1}))$ – the access inequality of Firm 2 is higher than that of Firm 1. The intuition behind these results is as follows. When compared to Firm 1, Firm 2 sees more disparity between talent of its employees. In the presence of adversarial risk, Firm 2, thus, has a greater incentive to grant its best performing employees access to more datasets while reducing access to data to the worst performing employees. This naturally creates higher level of access inequality. As discussed above, when $\sigma > 0$, employees with more access to data are attacked with higher probability, limiting the extent to which their access increases as they get more productive. However, the firm always has an incentive and the possibility to reduce the data access level of their lower performing employees. Therefore, the average data access level of Firm 2 is lower than that of Firm 1.

4.3. Technology: Horizontal vs Vertical Differentiation

In this section we investigate the impact of the firm's technology on the optimal data access architecture. To do that, we relax the assumption that the datasets are homogeneous in their value to the firm. Recall from Section 3 that the firm's technology is described by the matrix $\{\gamma_{ij}\}_{i\in\mathcal{M},j\in\mathcal{N}}$, the element γ_{ij} of which captures the value to the firm that employee $i \in \mathcal{M}$ generates when being granted access to the dataset $j \in \mathcal{N}$. Analysis of a completely general specification of such technology matrix is unlikely to yield structural results or generate deep managerial insights. Thus, instead, we consider two particular structures of the firm's technology. *Vertically differentiated technology* is such that there is a well-defined (and objective) order of both datasets and employees

[4] A function $f : \mathbb{R}^m \to \mathbb{R}$ is said to be Schur-convex if for any $x, y \in \mathbb{R}^m$ such that x is majorized by y, $f(x) \leq f(y)$. A function $-f(x)$ is Schur-concave if a function $f(x)$ is Schur-convex. See https://en.wikipedia.org/wiki/Schur-convex_function.

with regard to the value that they bring to the firm. This is similar to the quality measure of the vertically differentiated products. Such technology function is also a natural extension of our previous specification with heterogeneously productive employees. *Horizontally differentiated technology* is such that different datasets and different experts belong to different functional areas, and for maximum value to be generated for the firm, the expert's area must match the dataset's area. This is similar to the horizontally differentiated market in which customers prefer to buy the product which is the closest to them on a Hotelling line (if the price is the same). We formalize vertically differentiated technology in Definition 1 and horizontally differentiated technology in Proposition 3. We then analyze the structural properties of the firm's optimal data access architecture in each of these two cases.

Definition 1. The firm's technology is vertically differentiated and there is an objective ordering of datasets and employees by value to the firm when the following two conditions hold simultaneously:

1) For each pair of employees i and i' either $\gamma_{ij} > \gamma_{i'j}$ or $\gamma_{ij} < \gamma_{i'j}$ for all datasets j.

2) For each pair of datasets j and j', either $\gamma_{ij} > \gamma_{ij'}$ or $\gamma_{ij} < \gamma_{ij'}$ for all employees i.

This definition introduces an order on the sets of employees and datasets. In particular, when comparing any two employees, the firm is able to say which of them is more valuable. Here "more valuable" means that, when granted access to *any* dataset, this employee brings higher value that the other employee. Similarly, the second part of the definition states that the firm can establish the order on the set of datasets.

Proposition 3. *Assume that the firm's technology is vertically differentiated or, equivalently, that there is an objective ordering of datasets and employees by value to the firm (Definition 1). Then any equilibrium architecture g^* (played with positive probability) has an hierarchical structure: if an employee i has access to more datasets than an employee i' has, then employee i also has access to all datasets of employee i'.*

This proposition states that when the firm's technology is vertically differentiated, the equilibrium data access architecture has a particular hierarchical (or nested) structure. There are employees who have access to the highest number of datasets. If we denote the set of datasets that these employees have access to by $\mathcal{J}$, then, all other employees with access to lower number of datasets should have access only to the datasets in a *subset* of $\mathcal{J}$ (and to no datasets outside of $\mathcal{J}$, i.e. in $\mathcal{N} \setminus \mathcal{J}$). The hierarchy then continues to the employees with even less data access. Anecdotal evidence suggests that such architectures, when the higher ranked employees (e.g., CEOs) have access to all the data, are widely used by the companies. However, as the proposition suggests, the hierarchical data access architecture is only optimal when the chance that the adversary is strategic is low or, in other words, when the attacks that the firm faces are mostly of opportunistic type (e.g., a malware attack). Naturally, this is rarely the case and, hence, the result of the proposition implies that such companies should rethink how they distribute access to data.[5] Intuitively, the result may be explained by the knapsack-like[6] structure of the optimal architecture: the bipartite graph that describes it would only contain the edges that are of sufficiently high value. This structure, combined with the objective ordering on employees and datasets, implies that a higher-productivity employee would always be granted access to any dataset of a lower-productivity employee, and a similar logic holds for the datasets.

PROPOSITION 4. *Let the sets of employees and datasets be split as follows:* $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$ *and* $\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2$*, where* $\mathcal{M}_1 \cap \mathcal{M}_2$ *and* $\mathcal{N}_1 \cap \mathcal{N}_2$ *are empty sets. Assume that the firm's technology is horizontally differentiated or, equivalently, that it exhibits specialization:* $\gamma_{ij} = \gamma_H$ *if both* $i \in \mathcal{M}_k$*, and* $j \in \mathcal{N}_k$ *for some* $k \in \{1, 2\}$ *and* $\gamma_{ij} = \gamma_L < \gamma_H$ *otherwise. Then for sufficiently high attack strength* μ *employees of group* $\mathcal{M}_k$ *have access to datasets exclusively from group* $\mathcal{N}_k$ *in any equilibrium architecture* g^* *that has positive probability.*

The above proposition considers a case of horizontal differentiation. That means that for each employee there are datasets on which this employee specializes: when granted access to these datasets, this employee generates more value as compared to when having access to any other dataset outside the set of specialization. If the attack strength is sufficiently high, then the firm, when choosing the data access architecture, will follow the specialization order and will only match employees to the datasets that they specialize on. In other words, the firm will only grant employees access to the datasets that are most needed to them from the firm's value perspective. When facing harsh adversarial environments (high attack strength, μ) this result can, thus, be aligned with the widely-discussed *Zero Trust* approach to cybersecurity and data access management. Zero Trust framework, in particular, prescribes to "*employ network level security controls to only expose the applications [data] a user needs, thereby preventing anybody from exploring any part of the network to which they don't need access.*" (Deloitte 2021) – the outcome achieved when designing a data access architecture while keeping an employee specialization in mind.

5. Discussion and Conclusion

This paper investigates design of the data access architecture in a company facing strategic adversaries. In particular, we ask "*who should be granted access to what data*" and explore how the answer to this question depends on the factors external to the firm: the adversary's strength and

[5] "Why the C-Suite Doesn't Need Access to All Corporate Data", *DarkReading*, Dec 6, 2021, https://www.darkreading.com/vulnerabilities-threats/why-the-c-suite-doesn-t-need-access-to-all-corporate-data

[6] See https://en.wikipedia.org/wiki/Knapsack_problem for the definition of the knapsack problem.

sophistication; as well as the internal ones: the distribution of the employee talent, and the firm's technology. Our findings indicate that, with respect to the adversarial environment, the firm should transition from the "full-access" to the "access-managed", and, finally, to the "encapsulated" data access architectures when the adversarial environment becomes more harsh. The firm should do so by consciously lowering the overall data access level in the organization while maintaining sufficient levels of access inequality which are instrumental when facing sophisticated adversaries. We show that the same instruments can also be effective for the firms that face an increased adversarial risk due to disparity in productivity (or talent) of their employees. Finally, our analysis shows emergence of two commonly observed in practice types of data access architectures – hierarchical and Zero-Trust structures. These particular data access architectures emerge as the optimal response to the harsh adversarial environments in cases when the firm's technology is horizontally or vertically differentiated respectively (as defined in Section 4.3).

Our results can also be interpreted in the language of production flexibility (Bassamboo et al. 2012): the degree of an employee in our setting corresponds to that employee's flexibility. While the existing results show that an increase in the demand-related risk (connected to its uncertainty) induces more flexibility, we show that the cybersecurity-related risks would actually lead to its reduction. More generally, the results of our study imply that the operational decisions—such as the assignment of the datasets to the employees—are closely intertwined with cybersecurity decisions–such as minimizing the losses from the data breaches. These decisions should not be made separately and rather must be coordinated in a coherent strategy. Such joint operations-and-cybersecurity strategies are understudied and so this gap motivates our analysis. Being a starting point in this direction, our paper makes a number of simplifying assumptions which allow us to only focus on the first-order effects and derive structural results and insights. For instance, we assume that the firm's payoff function (i.e., technology) is additive in the value of edges between employees and datasets. However, one may consider more complex payoff functions such that capture data-synergy effects – more value is being generated when multiple datasets are matched to the same employee. We expect that such modifications, while significantly complicating the analysis, will only have limited impact on the main driving forces of our results.

Finally, our analysis can also be extended in a number of directions that would require a major overhaul of our modeling framework. Those include: understanding how existing studies on operational architecture such as (Gokpinar et al. 2010, Gurvich and Van Mieghem 2015) could be augmented to integrate cybersecurity concerns; exploring the endogenous specification of the firm's technology, adversary's attack strength and sophistication; investigating the firm's uncertainty about its technology; and exploring the question of micro-founding the firm's technology (e.g., through potentially linking it to the queueing theory models). We leave such investigations for future follow-up work that, as we hope, will be inspired by our study.

References

Arora A, Telang R, Xu H (2008) Optimal policy for software vulnerability disclosure. *Management Science* 54(4):642–656.

August T, Dao D, Kim K (2019) Market segmentation and software security: Pricing patching rights. *Management Science* 65(10):4575–4597.

August T, Dao D, Niculescu MF (2022) Economics of ransomware: Risk interdependence and large-scale attacks. *Management Science* .

Bailey T, Kolo B, Rajagopalan K, Ware D (2018) Insider threat: The human element of cyberrisk. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/insider-threat-the-human-element-of-cyberrisk.

Bassamboo A, Randhawa RS, Van Mieghem JA (2012) A little flexibility is all you need: on the asymptotic value of flexible capacity in parallel queuing systems. *Operations Research* 60(6):1423–1435.

Bell DE, La Padula LJ (1973) Secure computer systems: Mathematical foundations. Technical report, MITRE Corp Bedford MA.

Bimpikis K, Ozdaglar A, Yildiz E (2016) Competitive targeted advertising over networks. *Oper. Res.* 64(3):705–720.

Chan TC, Fearing D (2019) Process flexibility in baseball: The value of positional flexibility. *Management Science* 65(4):1642–1666.

Charlson G (2020) Searching for results: Optimal platform design in a network setting. Faculty of Economics, University of Cambridge, URL https://www.inet.econ.cam.ac.uk/working-paper-pdfs/wp2052.pdf.

Dawande M, Feng Z, Janakiraman G (2021) On the structure of bottlenecks in processes. *Management Science* 67(6):3853–3870.

Deloitte (2017) Responding to cyber threats: A change in paradigm. https://www2.deloitte.com/content/dam/Deloitte/sg/Documents/risk/sea-risk-cyber-thought-leadership.pdf.

Deloitte (2020) Covid-19: Cyber and the remote workforce. https://www2.deloitte.com/global/en/pages/risk/cyber-strategic-risk/articles/covid-19-cyber-and-the-remote-workforce.html.

Deloitte (2021) Zero Trust: In the face of escalating cyber-attacks, companies target a new level of security. https://www2.deloitte.com/us/en/pages/finance/articles/zero-trust-in-the-face-of-escalating-cyber-attacks.html.

Fainmesser IP, Galeotti A, Momot R (2022) Digital privacy. *Management Science* .

Ferraiolo DF, Kuhn DR (2009) Role-based access controls. *CoRR* abs/0903.2171, URL http://arxiv.org/abs/0903.2171.

GetApp (2022) Getapp's 2022 data security report—seven startling statistics. https://www.getapp.com/resources/annual-data-security-report/.

Gokpinar B, Hopp WJ, Iravani SM (2010) The impact of misalignment of organizational structure and product architecture on quality in complex product development. *Management science* 56(3):468–484.

Gurvich I, Van Mieghem JA (2015) Collaboration and multitasking in networks: Architectures, bottlenecks, and capacity. *Manufacturing & Service Operations Management* 17(1):16–33.

IBM Security (2022) Cost of a data breach 2022: A million-dollar race to detect and respond. https://www.ibm.com/reports/data-breach.

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. *Management science* 41(4):577–594.

Kain RY, Landwehr CE (1987) On access checking in capability-based systems. *IEEE Transactions on Software Engineering* (2):202–207.

Kannan K, Telang R (2005) Market for software vulnerabilities? think again. *Management science* 51(5):726–740.

McKinsey & Company (2017) Attracting and retaining the right talent. https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/attracting-and-retaining-the-right-talent.

Roberson B (2006) The Colonel Blotto game. *Economic Theory* 29(1):1–24.

Stallings W, Brown L (2018) *Computer Security: Principles and Practice* (Pearson Education, Incorporated), ISBN 9780134794105, URL https://books.google.ru/books?id=YpjZtwEACAAJ.

Wang W, Yang L, Chen Y, Zhang Q (2015) A privacy-aware framework for targeted advertising. *Computer Networks* 79:17–29.

Online Appendices

Appendix A: Proofs of the Main Results

A.1. Preliminaries

To simplify notation in this appendix, we will write the expectation operator $\mathbb{E}_G[\cdot]$ defined in the main text as $\mathbb{E}[\cdot]$. We also let $\boldsymbol{\varrho}$ denote a $S\times 1$ vector of probabilities, where $|\mathcal{G}_{mn}|=S$ and such that ϱ_i is the probability that a $m-n$ bipartite graph, g, is realised given a probability distribution G.

A.2. Proof of Proposition 1 (Page 7).

Recall that $X_s\in\Delta(\mathcal{N})$ denotes a generic (potentially mixed) strategy of the sophisticated adversary. Then, let $\mathbb{E}[\pi^S(G,X_\theta(G))]$ denote the strategic adversary's expected payoff.

Noting that both $\Delta(\mathcal{M})$ and $\Delta(\mathcal{G}_{mn})$ are compact and convex, by the Debreu-Fan-Glicksberg Theorem, if the following two conditions hold, a (mixed) equilibrium of our game exists:

1. The adversary's payoff function $\mathbb{E}[\pi^S(G,X_\theta(G))]$ is: (a) continuous in G and (b) continuous and concave in X;
2. The firm's payoff function $\mathbb{E}_{G,X^*_\theta(G)}[\pi(X^*_\theta(G),G;\mu,\gamma)]$ is: (a) continuous in X and (b) continuous and concave in G;

It is clear (a) holds in both cases. For 1(b), $\mathbb{E}_{G,X^*_\theta(G)}[\pi(G,X;\mu,\gamma)]$ is continuous in X for any G. As $\mathbb{E}[\pi^S(G,X_\theta(G))]$ is linear in any probability $x_i\in X$, it follows that it is also concave in X.

For 2(b), let g_e represent the empty graph with $m+n$ nodes and ϱ_e denote the probability that a graph g_e is realised. Note that it is possible to rewrite the firm's maximisation problem such that they choose the difference between the probability, ϱ_i, that a bipartite graph, $g_i\in\mathcal{G}_{mn}$ is realised, and ϱ_e; that is they choose the vector of probabilities $\hat{\boldsymbol{\varrho}}$ whose ith entry is $\varrho_i-\varrho_e$ subject to the constraint $\sum_k\hat{\varrho}_k=1$. Given the functional form of $\mathbb{E}_{G,X^*_\theta(G)}[\pi(\hat{\boldsymbol{\varrho}},X;\mu,\gamma)]$, and the fact that the preceding constraint is linear, it follows that for all X, $\mathbb{E}_{G,X^*_\theta(G)}[\pi(\hat{\boldsymbol{\varrho}},X;\mu,\gamma)]$ is concave in each $\hat{\varrho}_i$, and so 2(b) holds. Hence, the Debreu-Fan-Glicksberg Theorem applies to our game and a Nash equilibrium (and, equally, a subgame perfect equilibrium) exists.

For the final claim, clearly if $\mu=0$, the firm has a dominant strategy such that $\varrho^*_c=1$, where g_c is the complete $n-m$ bipartite network for any $\sigma\in[0,1]$. As $\mathbb{E}_{G,X^*_\theta(G)}[\pi(\hat{\boldsymbol{\varrho}},X;\mu,\gamma)]$ is continuous in μ, it immediately follows that this also applies when $\mu<\bar{\mu}$, for some $\bar{\mu}\in\mathbb{R}_+$ and any $\sigma\in[0,1]$.

A.3. Proof of Theorem 1 (Page 9) and Proposition 2 (Page 8)

Let $\kappa_i(X_\theta(G),\sigma):=\sigma x_i(G)+(1-\sigma)\frac{1}{n}$, where $x_i(G)$ is the probability a sophisticated adversary attacks i given their strategy, $X^*_\theta(G)$.

We specify the adversary's optimal strategy. Noting the adversary's profit function, it must be that if $\mathbb{E}[d_i^2(G)]>\mathbb{E}[d_j^2(G)]$ for all i, then $x_i(G)=1$. If $\mathbb{E}[d_i^2(G)]=\mathbb{E}[d_j^2(G)]$ for some employee pair i,j, then the adversary is indifferent between i and j, and could potentially play a mixed strategy over (at least) i and j such that $x_i(G),x_j(G)\in(0,1)$.

Note that $\gamma_{ij}=\gamma_i$ for all j. We can thus restate the firm's profit function as follows:

$$\mathbb{E}_{G,X^*_\theta(G)}[\pi(\boldsymbol{\varrho},X^*_\theta(G);\mu,\gamma)]=\sum_i[\gamma_i\mathbb{E}[d_i(\boldsymbol{\varrho})]-\mu\kappa_i(X^*_\theta(G),\sigma)\mathbb{E}[d_i^2(\boldsymbol{\varrho})]].$$

Any equilibrium is a solution to $\max_{\boldsymbol{\varrho}} \mathbb{E}_{G,X^*_\theta(G)}[\pi(\boldsymbol{\varrho}, X;\mu,\gamma]$ subject to the constraints $\sum_k \varrho_k = 1$ and $\varrho_k \geq 0$ $\forall k$. Note that the profit function is separable in each $\mathbb{E}[d_i(\boldsymbol{\varrho})]$ and $\mathbb{E}[d_i^2(\boldsymbol{\varrho})]$ term and $\gamma_{ij} = \gamma_i$, and so we can consider the choice of $\mathbb{E}[d_i(\boldsymbol{\varrho})]$ and $\mathbb{E}[d_i^2(\boldsymbol{\varrho})]$ separably for each i. That is, letting $\mathbb{Z}_n$ denote the set of integers $\{0,1,...,n\}$, the firm's problem is equivalent to, without loss of generality, the case where the firm instead chooses, for each i, a distribution $Z_i \in \Delta\mathbb{Z}_n$. Any collection of such distributions $\{Z_i\}_i$ potentially corresponds to multiple probability distributions over $\mathcal{G}_{mn}$, but all of those probability distributions are payoff identical for both the firm and the adversary. Straightforwardly, a distribution over $\mathcal{G}_{mn}$, G, induces a distribution over degrees, for each agent i, $\{Z_i\}_i$.

Let z_{ik} be the probability that employee i receives a degree of k under some strategy Z_i. Then, the firm's optimisation problem can be reduced to choosing $\{Z_i\}_i$, to maximise:

$$\mathbb{E}_{G,X^*_\theta(G)}[\pi(\{Z_i\}_i, X^*_\theta(G);\mu,\gamma)] = \sum_i [\gamma_i \mathbb{E}[d_i(Z_i)] - \mu\kappa_i(X^*_\theta(G),\sigma)\mathbb{E}[d_i^2(Z_i)]].$$

subject to the constraints that $\sum_k z_{ik} = 1$ and $z_{ik} \geq 0$.

Now, we prove the following Lemma:

LEMMA 1. *Let $(G^*(X), X(G^*))$ be an equilibrium and let the set $\{Z_i^*\}_i$ be induced by $G^*(X)$. Then, for any graph, g, realised with positive probability, $d_i = \lfloor \mathbb{E}[d_i(Z_i)] \rfloor$ or $\lceil \mathbb{E}[d_i(Z_i)] \rceil$ for all i and σ.*

Proof. Suppose that, according to some G, there is a non-zero probability that $d_i = p \in \mathbb{Z}_n$ **and** $d_i = q$ for some $q > p+1$. Consider a strategy profile $X(G)$, with $x_i(G) \in [0,1]$ and the implied total attack probability, $\kappa_i(X(G),\sigma)$. Suppose first that $\sigma < 1$. Then, by definition $\kappa_i(X(G),\sigma) > 0$ for any $x_i(G)$.

Given the stated Z_i, there exists a strategy, $Z_i^{'}$, where $d_i = p^{'}$ or $d_i = q^{'}$, where $p^{'}, q^{'} \in \mathbb{Z}_n$ and $q^{'} = p^{'} + 1$ with probability 1, and $\mathbb{E}[d_i(G^{'})]^2 > \mathbb{E}[d_i(G)]^2$ (and so $\mathbb{E}[d_i(G^{'})] > \mathbb{E}[d_i(G)]$, as $\mathbb{E}[d_i(G^{'})], \mathbb{E}[d_i(G)] > 0$) and $\mathbb{E}[d_i^2(G^{'})] = \mathbb{E}[d_i^2(G)]$. The latter equality and the fact that $\mathbb{E}[d_i(G^{'})] > \mathbb{E}[d_i(G)]$ immediately implies the firm receives a higher payoff under $Z^{'}$ than under Z. The firm therefore prefers to deviate from Z_i to $Z_i^{'}$,and so G cannot form part of an equilibrium.

Now suppose $\sigma = 1$. Then, given the definition of $\kappa_i(X(G),\sigma)$, it is, in theory, possible that $\kappa_i(X(G),\sigma) = 0$, i.e. if $x_i(G) = 0$. If $\kappa_i(X(G),\sigma) = 0$, then for G to be a best response for the firm, it must be that $\mathbb{E}[d_i(Z_i)] = n$. This follows from the fact that if $\kappa_i(X(G),\sigma) = 0$, the firm's payoff is linearly increasing in $\mathbb{E}[d_i(G)]$ - they increase the degree of i with no cost. If $\mathbb{E}[d_i(G)] = n$, then the property in the Lemma holds trivially. $\square$

The above Lemma can be used to establish the results in Theorem 1.

Suppose first that $\sigma = 0$ and $\mu > \bar{\mu}$, . It follows from Lemma 1 that the solution to the firm's constrained optimization problem above can found by solving the first order conditions of a set of Kuhn-Tucker (K-T) problems, one for each employee i, whose ith equation can be written:

$$\mathcal{L}_i = \mathbb{E}_{G,X^*_\theta(G)}[\pi(\{Z_i\}_i, X^*_\theta(G);\mu,\boldsymbol{\gamma})] - \lambda_{i0}(\sum_k z_{ik} - 1) + \sum_k \lambda_{ik} z_{ik}. \tag{4}$$

The K-T conditions then characterise an optimal Z_i, and a corresponding $\mathbb{E}[d_i(Z_i)]$ with $\lambda_{ik} = 0$ for all $k \neq \lfloor \mathbb{E}[d_i(Z_i)] \rfloor, \lceil \mathbb{E}[d_i(Z_i)] \rceil$ (given that this is a non-negativity constraint, this then implies $z_{ik} = 0$). Note that, for all k, $\frac{\partial \mathbb{E}_{G,X^*_\theta(G)}[\pi(Z_i X^*_\theta(G);\mu,\boldsymbol{\gamma})]}{\partial z_{ik}} = \gamma_i k - \frac{1}{m}\mu k^2$,where $\gamma_i k - \frac{1}{m}\mu k^2 \geq 0$ if $k \leq \lfloor \frac{\gamma_i}{\frac{2}{m}\mu} \rfloor$ and $\gamma_i k - \frac{1}{m}\mu k^2 < 0$

if $k > \lfloor \frac{\gamma_i}{\frac{2}{m}\mu} \rfloor$. Then the solution to the above problem is such that $d_i^* = \lfloor \frac{\gamma_i}{\frac{2}{m}\mu} \rfloor$ for all g realised with positive probability. That is, $\mathbb{E}[d_i(Z_i^*)] = \lfloor \frac{\gamma_i}{\frac{2}{m}\mu} \rfloor$ and $\mathbb{E}[d_i^2(Z_i^*)] = \mathbb{E}[d_i(Z_i^*)]^2$.

To build some intuition (though it is not strictly necessary for the proof to be complete), consider next the case where $\sigma = 1$. Then, the following Lemma holds:

Lemma 2. *When $\sigma = 1$, $\mathbb{E}[d_i^2(Z_i^*)] = \mathbb{E}[d_j^2(Z_j^*)]$ for all i, j pairs.*

Proof. Suppose instead that $\mathbb{E}[d_i^2(Z_i^*)] > \mathbb{E}[d_j^2(Z_j^*)]$ for some potential equilibrium probability distribution $G^*(X)$. Then it must be that $x_j^* = 0$ for any best response to G^*,$X_\theta^*(G^*)$, which implies $\kappa_j(X_\theta^*(G^*), \sigma) = 0$. However, in this case, the firm has an incentive to deviate to some strategy $G^{'}$, which induces some distribution over degrees $\{Z_i\}_i$ such that $Z_k = Z_k^{'}$ for $k \neq j$ but $Z_j^{'}$ is such that $z_n^{'} = 1$ (i.e. employee j has maximum degree), and so $\mathbb{E}[d_i^2(Z_i^*)] \geq \mathbb{E}[d_j^2(Z_j^*)]$ for all $i \neq j$, as $\mathbb{E}[d_j^2(Z_j^*)] \leq n^2$ by definition. Hence, we have a contradiction. □

Lemmas 1 and 2 jointly imply that the following Lemma:

Lemma 3. *When $\sigma = 1$, $\mathbb{E}[d_i(Z_i^*)] = d \in (0, n]$ for all i, j pairs.*

Proof. Suppose instead that for some i, j pair, $\mathbb{E}[d_i(Z_i^*)] > \mathbb{E}[d_j(Z_j^*)]$ generated by some distribution over degrees $\{Z_i^*\}_i$ induced by a probability distribution $G^*(X^*)$, which is part of the an equilibrium $\{G^*(X^*), X^*(G^*)\}$. By Lemma 2, $\mathbb{E}[d_i^2(Z_i^*)] = \mathbb{E}[d_j^2(Z_j^*)]$. By Lemma 1, $d_i = \lfloor \mathbb{E}[d_i(Z_i^*)] \rfloor$ or $\lceil \mathbb{E}[d_i(Z_i^*)] \rceil$ for all i and graphs realised with positive probability. For both $\mathbb{E}[d_i(Z_i^*)] > \mathbb{E}[d_j(Z_j^*)]$ and $\mathbb{E}[d_i^2(Z_i^*)] = \mathbb{E}[d_j^2(Z_j^*)]$ to hold simultaneously, it must be that there is some g realised with positive probability where $d_j = \tilde{d}$ for some $\tilde{d} > \lceil \mathbb{E}[d_i(Z_i^*)] \rceil$ otherwise, $\mathbb{E}[d_i^2(Z_i^*)] > \mathbb{E}[d_j^2(Z_j^*)]$,yielding a contradiction. However, in this case, by definition, $\tilde{d} \neq \lfloor \mathbb{E}[d_j(Z_j^*)] \rfloor, \lceil \mathbb{E}[d_j(Z_j^*)] \rceil$, again, yielding a contradiction. □

We then have that, for any equilibrium, G^*, $\mathbb{E}[d_i(Z_i^*)] = \mathbb{E}[d_j(Z_j^*)]$ and $\mathbb{E}[d_i^2(Z_i^*)] = \mathbb{E}[d_j^2(Z_j^*)]$ for all i, j pairs. Hence, the firm's expected payoff from each employee is equal in any equilibrium. Given the functional form of the firm's profit function, it follows that to find the optimal pattern of degrees, it is possible to reduce the firm's optimisation problem to a choice of a single $Z \in \Delta\mathbb{Z}_n$, which maximises:

$$\mathbb{E}_{G, X_\theta^*(G)}[\pi(\{Z_i\}_i, X_\theta^*(G); \mu, \gamma)] = \sum_i [\gamma_i \mathbb{E}[d_i(Z)] - \mu\kappa_i(X(G), \sigma)\mathbb{E}[d_i^2(Z)]],$$

subject to the constraints that $\sum_k z_k = 1$ and $z_k \geq 0$. We solve this by considering a single modified Kuhn Tucker (MKT) function, where the firm chooses which is of the following form:

$$\mathcal{L}_m = \sum_i \gamma_i k - m\mu k^2 - \lambda_0(\sum_k z_k - 1) + \sum_k \lambda_k z_k.$$

Solving, we find that $\mathbb{E}[d_i(Z_i^*)] = \lfloor \frac{\sum_i \gamma_i}{2\mu} \rfloor$ for all i, and, moreover, $d_i = \lfloor \frac{\sum_i \gamma_i}{2\mu} \rfloor$ for any graph realised with positive probability in any equilibrium G^* and all i.

Finally, we find a probability distribution $X_\theta^*(G)$ that sustains such an equilibrium pattern of degrees, to ensure the proposed pattern of degrees can be sustained in an equilibrium (though, formally, we know this must be true by the equilibrium existence result in Proposition 1). For this to hold, it must be that

$\frac{\partial \mathbb{E}_{G,X_\theta^*(G)}[\pi(\{Z,X_\theta^*(G);\mu,\gamma)]}{\partial z_{ik^*}} \leq 0$ where $k^* = \lfloor \frac{\sum_i \gamma_i}{2\mu} \rfloor$: otherwise, the firm has a profitable deviation. This holds when it $x_i^*(G^*) = \frac{\gamma_i}{2\mu k^*}$ for all i, and by the definition of k^*, this is implementable. Hence, we have found an equilibrium to the above game.[7]

Now, consider the general case where $\sigma \in [0,1]$, but if $\mu > \bar{\mu}$: we will state the optimal distribution where $\mu \leq \bar{\mu}$ in the summary of the equilibrium below. Suppose, that the vector $\boldsymbol{\gamma}$ is ordered such that $\gamma_i > \gamma_j$ if $i > j$ and hence $\gamma_1 \in \max\{\gamma_i\}$ - we show later that this is without loss of generality.

Under the optimal pattern of degree distributions which solve the K-T function when $\sigma = 0$, $\mathbb{E}[(d_1^*(Z_1))^2] > \mathbb{E}[(d_i(Z_i^*))^2]$ for all $i \neq 1$. Hence, $x_i(G^*) = 1$ for all G^* at $\sigma = 0$ and so $\mathbb{E}[d_i(Z_i^*)] = \lfloor \frac{\gamma_i}{\frac{2}{m}\mu(1-\sigma)} \rfloor$ for all $i \neq 1$ and $\mathbb{E}[d_1(Z_1^*)] = \lfloor \frac{\gamma_1}{2\mu[\sigma + \frac{1}{m}(1-\sigma)]} \rfloor$.

As $\gamma_1 > \gamma_2$ by definition and, $\lim_{\sigma \to 1} \frac{\gamma_2}{2(1-\sigma)\frac{1}{m}\mu} = \infty$ for any $\gamma_2 > 0$, it follows that if $\mu > \bar{\mu}$, $\exists \tilde{\sigma}_2 \in [0,1]$ where:

1. if $\sigma < \tilde{\sigma}_2$, then $\lfloor \frac{\gamma_1}{2\mu[\sigma + \frac{1}{m}(1-\sigma)]} \rfloor > \lfloor \frac{\gamma_2}{\frac{2}{m}\mu(1-\sigma)} \rfloor$
2. $\lfloor \frac{\gamma_2}{2(1-\tilde{\sigma}_2)\frac{1}{m}\mu} \rfloor = \lfloor \frac{\gamma_1}{2\mu(\tilde{\sigma}_2 + (1-\tilde{\sigma}_2)\frac{1}{m})} \rfloor$.

Hence, when $\sigma \leq \tilde{\sigma}_2$ it follows that, in any equilibrium, $x_i(G^*) = 1$, and, in any graph, g, realised with positive probability $d_i = \lfloor \frac{\gamma_i}{\frac{2}{m}\mu(1-\sigma)} \rfloor$ for $i \neq 1$ and $d_1 = \lfloor \frac{\gamma_1}{2\mu(\sigma + (1-\sigma)\frac{1}{m})} \rfloor$.

Now, consider $\sigma > \tilde{\sigma}_2$. We state the following:

LEMMA 4. *Suppose* $\sigma > \tilde{\sigma}_2$. $\mathbb{E}[d_1^2(Z_1^*)] = \mathbb{E}[d_2^2(Z_2^*)]$ *and* $\mathbb{E}[d_1(Z_1^*)] = \mathbb{E}[d_2(Z_2^*)]$.

Proof. For the first claim, suppose instead that $\mathbb{E}[d_1^2(Z_1^*)] > \mathbb{E}[d_2^2(Z_2^*)]$. There are three subcases to consider. First, suppose there is some employee, $s \in \mathcal{N}$, such that $\mathbb{E}[d_s^2(Z_s^*)] > \mathbb{E}[d_1^2(Z_1^*)]$. We show this cannot hold. To this see, note that the inequality $\mathbb{E}[d_s(Z_s^*)] > \mathbb{E}[d_1(Z_1^*)]$ by Lemma 1: for $\mathbb{E}[d_s(Z_s^*)] < \mathbb{E}[d_1(Z_1^*)]$ but $\mathbb{E}[d_s^2(Z_s^*)] > \mathbb{E}[d_1^2(Z_1^*)]$ to hold simultaneously, it must be that there is some g realised with positive probability where $d_s = \tilde{d}$ for some $\tilde{d} > \lceil \mathbb{E}[d_s(Z_s^*)] \rceil$ otherwise, $\mathbb{E}[d_1^2(Z_1^*)] > \mathbb{E}[d_s^2(Z_s^*)]$,yielding a contradiction. But then, as $\gamma_1 > \gamma_s$, $\lfloor \frac{\gamma_1}{\frac{2}{m}\mu(1-\sigma)} \rfloor > \lfloor \frac{\gamma_s}{2\mu(\kappa_s(X_\theta(G),\sigma))} \rfloor$ for all $x_s(G) \in [0,1]$, and so there does not exist an employee $s \in \mathcal{M}$ where $\mathbb{E}[d_s^2(Z_s^*)] > \mathbb{E}[d_1^2(Z_1^*)]$ (note, this also holds for $s = 2$). This holds because $\lfloor \frac{\gamma_1}{\frac{2}{m}\mu(1-\sigma)} \rfloor > \lfloor \frac{\gamma_s}{2\mu(\kappa_s(X_\theta^*(G),\sigma))} \rfloor$ implies that, at any optimum, $\mathbb{E}[d_1(Z_1^*)] > \mathbb{E}[d_s(Z_s^*)]$, and by the previous argument this implies that $\mathbb{E}[d_1^2(Z_1^*)] > \mathbb{E}[d_s^2(Z_s^*)]$, yielding a contradiction.

Second, suppose that $\mathbb{E}[d_1^2(Z_1^*)] > \mathbb{E}[d_i^2(Z_i^*)]$ for all $i \neq 1$. Then it must be that $x_1^*(G) = 1$. But we know that $\lfloor \frac{\gamma_2}{2(1-\sigma)\frac{1}{m}\mu} \rfloor > \lfloor \frac{\gamma_1}{2\mu(\sigma + (1-\sigma)\frac{1}{m})} \rfloor$ by definition. Hence, in equilibrium it must be that $\mathbb{E}[d_2(Z_2^*)] > \mathbb{E}[d_1(Z_1^*)]$, and so $\mathbb{E}[d_2^2(Z_2^*)] > \mathbb{E}[d_1^2(Z_1^*)]$, yielding a contradiction.

Suppose now that $\mathbb{E}[d_1^2(Z_1^*)] = \mathbb{E}[d_i^2(Z_i^*)]$ for a single employee $i \neq 1, 2$. For this to hold in equilibrium, it must be that $X_\theta^*(G)$ is such that $\lfloor \frac{\gamma_1}{2\mu(\kappa_1(X_\theta^*(G),\sigma))} \rfloor = \lfloor \frac{\gamma_i}{2\mu(\kappa_i(X_\theta^*(G),\sigma))} \rfloor$, with $\kappa_1(X_\theta^*(G),\sigma) + \kappa_i(X_\theta^*(G),\sigma) = 1$ and as we are considering the case $\mathbb{E}[d_1^2(Z_1^*)] > \mathbb{E}[d_2^2(Z_2^*)]$, it must be that $x_2^*(G) = 0$. By the definition of $\kappa_1(X_\theta^*(G),\sigma))$, for $\lfloor \frac{\gamma_1}{2\mu(\kappa_1(X_\theta^*(G),\sigma))} \rfloor = \lfloor \frac{\gamma_i}{2\mu(\kappa_i(X_\theta^*(G),\sigma))} \rfloor$, there must then exist a pair of probabilities $q_1, q_2 \in [0,1]$ where $q_1 + q_2 = 1$ and $\lfloor \frac{\gamma_1}{2\mu((1-\sigma)\frac{1}{m} + \sigma q_1)} \rfloor = \lfloor \frac{\gamma_i}{2\mu((1-\sigma)\frac{1}{m} + \sigma q_2)} \rfloor$. Recall $\gamma_2 > \gamma_i$ by definition and so

[7] Due to the fact that the degree distribution of each employee is discrete, it is possible that other equilibria exist, in which $X_\theta^*(G)$ differs from the schema above. However, the above proofs show that the degree distribution, and the firm's profit, is the same in any such equilibria.

$\kappa_1(X^*_\theta(G),\sigma) > q_1$. But then, as $\sigma > \tilde{\sigma}_2$, $\lfloor \frac{\gamma_1}{2\mu\kappa_1(X^*_\theta(G),\sigma)} \rfloor < \lfloor \frac{\gamma_2}{2\mu\frac{1}{m}(1-\sigma)} \rfloor$, and hence $\mathbb{E}[d_1(Z^*_1)] < \mathbb{E}[d_2(Z^*_2)]$ and, by the above argument, it then follows that $\mathbb{E}[d^2_1(Z^*_1)] < \mathbb{E}[d^2_2(Z^*_2)]$ too, yielding a contradiction. The same proof strategy can be used when $\mathbb{E}[d^2_1(Z^*_1)] = \mathbb{E}[d^2_i(Z^*_i)]$ for all $i \in E \subset \mathcal{M}$.

We have then proved that $\mathbb{E}[d^2_1(Z^*_1)]$ cannot be strictly greater than $\mathbb{E}[d^2_2(Z^*_2)]$, and, equally, the reverse cannot be true either. So $\mathbb{E}[d^2_1(Z^*_1)] = \mathbb{E}[d^2_2(Z^*_2)]$ Furthermore, we have also shown that if $\mathbb{E}[d^2_1(Z^*_1)] = \mathbb{E}[d^2_2(Z^*_2)]$ Lemma 1 implies that $\mathbb{E}[d_1(Z^*_1)] = \mathbb{E}[d_2(Z^*_2)]$, so the second statement in the Lemma holds. □

Lemma 4 implies that at $\sigma = \tilde{\sigma}_2$, the firm's problem can be solved by considering $m-1$ Kuhn-Tucker functions. At $\sigma = \tilde{\sigma}_2$, $m-2$ of these functions are identical to the K-T problem stated above (see equation (1)). The remaining equation can be written as follows:

$$\mathcal{L}_{1+2} = \sum_{i=1,2} \gamma_i k - n\mu k^2 - \lambda_0(\sum_k z_k - 1) + \sum_k \lambda_k z_k.$$

Characterising the solutions to these problems, we find that $\mathbb{E}[d_1(Z^*_1)] = \mathbb{E}[d_2(Z^*_2)] = \lfloor \frac{\gamma_1+\gamma_2}{4\mu(\sigma+(1-\sigma)\frac{1}{m})} \rfloor$, $\mathbb{E}[d^2_1(Z^*_1)] = \mathbb{E}[d^2_2(Z^*_2)] = \mathbb{E}[d_1(Z^*_1)]^2$ and $\mathbb{E}[d_i(Z^*_i)] = \lfloor \frac{\gamma_i}{\frac{2}{m}\mu(1-\sigma)} \rfloor$ for $i \neq 1,2$.

Now, it should be clear we can iteratively define $\tilde{\sigma}_i$ for all $i \neq 1$. For example, as $\frac{\gamma_1+\gamma_2}{2} > \gamma_3$ by definition and, $\lim_{\sigma\to 1} \frac{\gamma_3}{2(1-\sigma)\frac{1}{m}\mu} = \infty$ for any $\gamma_3 > 0$, it follows that if $\mu > \bar{\mu}$, $\exists \tilde{\sigma}_3 \in [0,1]$ where:

1. if $\sigma < \tilde{\sigma}_3$, then $\lfloor \frac{\gamma_1+\gamma_2}{4\mu(\sigma+(1-\sigma)\frac{1}{m})} \rfloor > \lfloor \frac{\gamma_3}{\frac{2}{m}\mu(1-\sigma)} \rfloor$
2. $\lfloor \frac{\gamma_3}{2(1-\tilde{\sigma}_3)\frac{1}{m}\mu} \rfloor = \lfloor \frac{\gamma_1+\gamma_2}{4\mu(\tilde{\sigma}_3+(1-\tilde{\sigma}_3)\frac{1}{m})} \rfloor$.

We can repeat the proof above to find the equilibrium outcomes when $\tilde{\sigma}_3 = \sigma$ and so on for all $i \neq 1$, until we get to $\tilde{\sigma}_m < 1$.

We can then summarise results as follows. We have shown that for all i and σ, $\mathbb{E}[d_i(Z^*_i)]^2 = \mathbb{E}[d^2_i(Z^*_i)]$, and so we can characterise equilibrium outcomes solely in terms of $\mathbb{E}[d_i(Z^*_i)]$. The following results hold for any equilibrium probability distribution G^*:

1. If $\mu \leq \bar{\mu}$, then $\mathbb{E}[d_i(Z^*_i)] = n$ for all $i \in \mathcal{M}$;
2. If $\mu > \bar{\mu}$ and $\tilde{\sigma}_j \leq \sigma < \tilde{\sigma}_{j+1}$, then $\mathbb{E}[d_k(Z^*_k)] = \min\{\lfloor \frac{\sum_{i=1}^j \gamma_i}{2j\mu(\sigma+(1-\sigma)\frac{1}{m})} \rfloor, n\}$ for $k \leq j$ and $\mathbb{E}[d_k(Z^*_k)] = \min\{\lfloor \frac{\gamma_k}{2\mu(1-\sigma)\frac{1}{m})} \rfloor, n\}$ for $k > j$;
3. If $\mu > \bar{\mu}$ and $\sigma \geq \tilde{\sigma}_m$ then $\mathbb{E}[d_i(Z^*)] = \min\{\lfloor \frac{\sum_i \gamma_i}{2\mu} \rfloor, n\}$ for all $i \in \mathcal{M}$.

where $\tilde{\sigma}_j$ is defined implictly such that $\lfloor \frac{\sum_{i=1}^j \gamma_i}{2j\mu(\tilde{\sigma}_j+(1-\tilde{\sigma}_j)\frac{1}{m})} \rfloor = \lfloor \frac{\gamma_j}{2\mu(1-\tilde{\sigma}_j)\frac{1}{m})} \rfloor$ for $j = [2,m]$ and $\tilde{\sigma}_j = 0$ for $j = 1$.

Now, we can prove the statements in Proposition 2 and Theorem 1.

For Proposition 2, note that we have shown $\mathbb{E}[d_i(G^*)]^2 = \mathbb{E}[d^2_i(G^*)]$ for any equilibrium G^*. Furthermore, the summary of equilibrium above immediately implies $\mathbb{E}[d_i(G^*)]$ is decreasing in μ. The statement regarding σ also automatically follows from the above characterisation of the equilibrium.

Note that each $\mathbb{E}[d_i(Z^*_i)]$ is decreasing in μ for all σ. This is sufficient to prove statement (1) in Theorem 1.

The above shows that we can write $w(G^*) = w(G^*;\mu)$: inequality is a function of μ. For (2) in Theorem 1, first note that $\sigma = 1$, then we have shown that $\mathbb{E}[d_i(Z^*)] - \mathbb{E}[d_j(Z^*)] = 0$ for all i,j pairs and all σ and μ, and so $w(G^*) = 0$.

We will consider $w(G^*)$ for $\sigma < 1$. We know that if $\mu \leq \bar{\mu}$ then $\mathbb{E}[d_i(Z^*_i)] = n$ for all $i \in \mathcal{M}$, and hence $w(G^*) = 0$. It is also clear from the above analysis and the fact that $\lim_{\mu\to\infty} \mathbb{E}[d_i(Z^*)] = 0$ for all i and any equilibrium G^*, that $\exists \tilde{\mu}_i > 0$ for all i such that $\mathbb{E}[d_i(Z^*)] < n$. What this then implies is that for $\mu < \tilde{\mu}_m$,$\mathbb{E}[d_i(Z_i^*)] = n$ for all $i \neq m$ and $\mathbb{E}[d_m(Z_m^*)] < m$, and so $w(G^*;\mu)$ is increasing in μ for at least $\mu \in [0, \underline{\mu}]$, for some $\underline{\mu}$ where $\underline{\mu} \geq \tilde{\mu}_m$ and potentially for $\mu \in [0, \infty]$.

We now show that $w(G^*;\mu)$ is not increasing in μ for the entire interval $[0,\infty]$. Note that $\tilde{\sigma}_j$ is a decreasing function of μ and so we write $\tilde{\sigma}_j(\mu)$. Suppose $\mu^{'} > \mu$ and assume at first that if $\sigma < \tilde{\sigma}_j(\mu)$ for some employee j, then $\sigma < \tilde{\sigma}_j(\mu^{'})$ as well. Suppose $\tilde{\sigma}_k(\mu^{'}) \leq \sigma < \tilde{\sigma}_{k+1}(\mu^{'})$ and that $\mu > \tilde{\mu}_k$. If $\sigma > \tilde{\sigma}_t(\mu^{'})$ and $\sigma \geq \tilde{\sigma}_{t'}(\mu^{'})$ then $\mathbb{E}[d_t(Z_t^*, \mu)] = \mathbb{E}[d_{t'}(Z_{t'}^*, \mu)]$ and $\mathbb{E}[d_t(Z_t^*, \mu^{'})] = \mathbb{E}[d_{t'}(Z_t^*, \mu^{'})]$. If $\sigma < \tilde{\sigma}_t(\mu^{'})$ and $\sigma \geq \tilde{\sigma}_{t'}(\mu^{'})$ (which implies that $\mathbb{E}[d_{t'}(Z_{t'}^*, \mu)] > \mathbb{E}[d_t(Z_t^*, \mu)]$), then as $\frac{\partial \frac{\sum_{i=1}^{j} \gamma_i}{2j\mu(\sigma + (1-\sigma)\frac{1}{m})}}{\partial \mu} < \frac{\partial \frac{\gamma_t}{2\mu(1-\sigma)\frac{1}{m})}}{\partial \mu}$, it follows that $\mathbb{E}[d_{t'}(Z_{t'}^*, \mu)] - \mathbb{E}[d_t(Z_t^*, \mu)] > \mathbb{E}[d_{t'}(Z_{t'}^*, \mu^{'})] - \mathbb{E}[d_t(Z_t^*, \mu^{'})]$. If $\sigma < \tilde{\sigma}_t(\mu^{'})$ and $\sigma < \tilde{\sigma}_{t'}(\mu^{'})$ but $t^{'} < t$, then, as $\frac{\partial \frac{\gamma_{t'}}{2\mu(1-\sigma)\frac{1}{m})}}{\partial \mu} < \frac{\partial \frac{\gamma_t}{2\mu(1-\sigma)\frac{1}{m})}}{\partial \mu}$, $\mathbb{E}[d_{t'}(Z_{t'}^*, \mu)] - \mathbb{E}[d_t(Z_t^*, \mu)] > \mathbb{E}[d_{t'}(Z_{t'}^*, \mu^{'})] - \mathbb{E}[d_t(Z_t^*, \mu^{'})]$.

Now, consider the case where there exists at least one employee j such that $\sigma < \tilde{\sigma}_j(\mu)$ but $\sigma \geq \tilde{\sigma}_j(\mu^{'})$. For any $t < j$, $\mathbb{E}[d_t(Z_t^*, \mu^{'})] = \mathbb{E}[d_j(Z_j^*, \mu^{'})]$ and $\mathbb{E}[d_t(Z_t^*, \mu)] > \mathbb{E}[d_j(Z_j^*, \mu)]$ if $\mu > \tilde{\mu}_t$. For any $t > j$, and, $\mu > \tilde{\mu}_j$, either $\sigma \geq \tilde{\sigma}_t(\mu^{'})$, it which case again $\mathbb{E}[d_t(Z_t^*, \mu^{'})] = \mathbb{E}[d_j(Z_j^*, \mu^{'})]$ and $\mathbb{E}[d_t(Z_t^*, \mu)] > \mathbb{E}[d_j(Z_j^*, \mu)]$,or $\sigma < \tilde{\sigma}_t(\mu^{'})$, in which case, as $\frac{\partial \frac{\gamma_j}{2\mu(1-\sigma)\frac{1}{m})}}{\partial \mu}$, $\frac{\partial \frac{\sum_{i=1}^{j} \gamma_k}{2j\mu(\sigma + (1-\sigma)\frac{1}{m})}}{\partial \mu} < \frac{\partial \frac{\gamma_j}{2\mu(1-\sigma)\frac{1}{m})}}{\partial \mu}$, it follows that $\mathbb{E}[d_j(Z_j^*, \mu)] - \mathbb{E}[d_t(Z_t^*, \mu)] > \mathbb{E}[d_j(Z_j^*, \mu^{'})] - \mathbb{E}[d_t(Z_t^*, \mu^{'})]$. For all employees, k, where $\sigma < \tilde{\sigma}_k(\mu)$,$\tilde{\sigma}_k(\mu^{'})$, the inequalities in the previous analysis still hold.

The above analysis shows that there exists a $\hat{\mu} \in [\underline{\mu}, \tilde{\mu}_1]$ such that when $\mu \geq \hat{\mu}$:

$$\mathbb{E}[d_i(Z_i^*, \mu)] - \mathbb{E}[d_j(Z_j^*, \mu)] \geq \mathbb{E}[d_i(Z_i^*, \mu^{'})] - \mathbb{E}[d_j(Z_j^*, \mu^{'})]$$

for all i, j pairs such that $i < j$, and so $w(G;\mu)$ is decreasing in μ. For the statement in the proposition to be true, then, it is sufficient to show that $\underline{\mu} = \hat{\mu}$. Suppose $\mu < \mu^{'} < \mu^{''}$. The previous analysis immediately implies that if $w(G;\mu) > w(G;\mu^{'})$ then it must be that $w(G;\mu^{''}) > w(G;\mu^{'})$: this follows from the fact for two agents, i and j with $i < j$, if $\mu > \tilde{\mu}_i$ then we have shown that $w(G;\mu)$ is decreasing in μ and clearly if $\mu > \tilde{\mu}_i$ then $\mu^{'} > \tilde{\mu}_i$ and $\mu^{''} > \tilde{\mu}_i$ (and indeed, it may well be that there is some i such that $\mu < \tilde{\mu}_i$ but $\mu^{'} < \tilde{\mu}_i$ etc.)

Similarly, if if $w(G;\mu^{'}) < w(G;\mu^{''})$ then it must be that $w(G;\mu) < w(G;\mu^{'})$: this holds because, as above, if $\mu > \tilde{\mu}_i$ then $\mu^{'} > \tilde{\mu}_i$ and $\mu^{''} > \tilde{\mu}_i$, and so weakly more agents have a degree of n in equilibrium under a cost of μ than for $\mu^{'}$ and $\mu^{''}$. Hence, $\underline{\mu} = \hat{\mu}$ and the statement in (2) is proved.

Lastly, for the statement on profit, suppose $\mu < \mu^{'}$, and, noting that the equilibrium distribution is a function of μ, we write $G^* = G^*(\mu)$. Note that, trivially, $\mathbb{E}_{G, X_\theta^*(G)}[\pi(G^*(\mu^{'}), X_\theta^*(G^*); \mu, \gamma)] > \mathbb{E}_{G, X_\theta^*(G)}[\pi(G^*(\mu^{'}), X_\theta^*(G^*); \mu^{'}, \gamma)]$, and by definition, $\mathbb{E}_{G, X_\theta^*(G)}[\pi(G^*(\mu), X_\theta^*(G^*); \mu, \gamma)] > \mathbb{E}_{G, X_\theta^*(G)}[\pi(G^*(\mu^{'}), X_\theta^*(G^*); \mu, \gamma)]$. Hence, the statement in the theorem holds.

A.4. Proof of Theorem 2 (Page 10).

Noting that $\mathbb{E}[d_i(G^*)]$ is a function of $\boldsymbol{\gamma}$ and that $\mathbb{E}[d_i(G^*)] = d_i^*$ for any equilibrium G^*, we write $\mathbb{E}[d_i(G^*)] = d_i^*(\boldsymbol{\gamma})$.Note that, by our characterization of the equilibrium in the proof of theorem 1, $\exists \hat{\mu}$ such that if $\mu > \hat{\mu}$ then $\mathbb{E}[d_i(G^*)] < n$ for all $d_i^*(\boldsymbol{\gamma})$. We will assume throughout this proof that $\mu > \hat{\mu}$, as stated implicitly in the theorem.

For the statement regarding $\mathbb{E}[d_i(G^*)]$ and $\mathbb{E}[d_i^2(G^*)]$, first note that by the proof of Theorem 1, $\mathbb{E}[d_i^2(G^*)] = \mathbb{E}[d_i(G^*)]^2$. Suppose γ majorizes $\gamma^{'}$. Noting that both vectors are ordered such that $\gamma_i > \gamma_j$ if $i < j$, then this implies that:

$$\sum_{i=1}^{k} \gamma_i \leq \sum_{i=1}^{k} \gamma_i^{'} \forall k = 1, ..., m,$$

and $\sum_{i=1}^m \gamma_i = \sum_{i=1}^m \gamma_i^{'}$. Consider the case where $\sigma = 1$ first. In this case, as established in the proof of Theorem 1, $d_i^*(\boldsymbol{\gamma}) = \frac{\sum_i \gamma_i}{2\mu}$ for all i when $\mu > \hat{\mu}$. Hence, the statement trivially holds, as $d_i^*(\boldsymbol{\gamma}) = d_i^*(\boldsymbol{\gamma}^{'})$ for all i.

Now suppose $\sigma = 0$. In this case, we know from the proof of Theorem 1 that $d_i^*(\boldsymbol{\gamma}) = \frac{\gamma_i}{2\mu\frac{1}{m}}$ for all i. Suppose that for both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^{'}$,$d_i^*(\boldsymbol{\gamma}) < n$ for all i. Then, as $\sum_{i=1}^m \gamma_i = \sum_{i=1}^m \gamma_i^{'}$, it follows that if $\mu > \hat{\mu}$ then $\sum_i d_i^*(\boldsymbol{\gamma}) = \sum_i d_i^*(\boldsymbol{\gamma}^{'})$, and hence the statement holds in this case as well.

Now, consider the case where $\sigma \in (0,1)$. In this case, as in the proof of Theorem 1, we know that if $\mu > \hat{\mu}$ and $\tilde{\sigma}_j \leq \sigma < \tilde{\sigma}_{j+1}$, then $d_k^*(\boldsymbol{\gamma}) = \frac{\sum_{i=1}^j \gamma_i}{2j\mu(\sigma + (1-\sigma)\frac{1}{m})}$ for $k \leq j$ and $d_k^*(\boldsymbol{\gamma}) = \frac{\gamma_k}{2\mu(1-\sigma)\frac{1}{m})}$ for $k > j$. Furthermore, we know that for all $j \in \{1, ..., m\}$ and $m > 1$, $2j\mu(\sigma + (1-\sigma)\frac{1}{m}) > 2\mu(1-\sigma)\frac{1}{m})$. As $\sum_{i=1}^k \gamma_i \leq \sum_{i=1}^k \gamma_i^{'} \forall k = 1, m$ and $\sum_{i=1}^m \gamma_i = \sum_{i=1}^m \gamma_i^{'}$, it then immediately follows that $\sum_i d_i^*(\boldsymbol{\gamma}) > \sum_i d_i^*(\boldsymbol{\gamma}^{'})$, which again satisfies the statement in the Theorem. Thus, $\bar{d}(G^*, \boldsymbol{\gamma})$ is Schur-concave in $\boldsymbol{\gamma}$, and therefore so too $\bar{d}^2(G^*, \boldsymbol{\gamma})$.

For the statement regarding $w(G^*, \boldsymbol{\gamma})$, consider first $\sigma = 0$ when $\mu > \hat{\mu}$. We know from the proof of Theorem 1 that in this case, $d_i^*(\boldsymbol{\gamma}) = \frac{\gamma_i}{2\mu\frac{1}{m}}$ for all i. Noting that $\mathbb{E}[d_i(G^*, \boldsymbol{\gamma})] = d_i(\boldsymbol{\gamma})$ for all i and G^*, we can restate $w(G^*, \boldsymbol{\gamma})$ as follows:

$$w(G^*, \boldsymbol{\gamma}) := \sum_i \sum_j [d_i(\boldsymbol{\gamma}) - d_j(\boldsymbol{\gamma})]^2 = \sum_i \sum_j [2d_i^2(\boldsymbol{\gamma}) - 2d_i(\boldsymbol{\gamma})d_j(\boldsymbol{\gamma})].$$

When $\sigma = 0$, given the stated equilibrium degrees, $\sum_{i=1}^m \gamma_i = \sum_{i=1}^m \gamma_i^{'}$, then $d_i^2(\boldsymbol{\gamma}) = d_i^2(\boldsymbol{\gamma}^{'})$. Furthermore, as $\sum_{i=1}^k \gamma_i \leq \sum_{i=1}^k \gamma_i^{'} \forall k = 1, ..., m$,it follows that $\sum_i \sum_j d_i(\boldsymbol{\gamma})d_j(\boldsymbol{\gamma}) > \sum_i \sum_j d_i(\boldsymbol{\gamma}^{'})d_j(\boldsymbol{\gamma}^{'})$ by definition. This then implies that $w(G^*, \boldsymbol{\gamma}) < w(G^*, \boldsymbol{\gamma}^{'})$.

Now, suppose that $\sigma > 0$. In this case, we know that if $\tilde{\sigma}_j \leq \sigma < \tilde{\sigma}_{j+1}$, then $d_k^*(\boldsymbol{\gamma}) = \frac{\sum_{i=1}^j \gamma_i}{2j\mu(\sigma + (1-\sigma)\frac{1}{m})}$ for $k \leq j$ and $d_k^*(\boldsymbol{\gamma}) = \frac{\gamma_k}{2\mu(1-\sigma)\frac{1}{m})}$ for $k > j$. We write $d_i^*(\boldsymbol{\gamma}) = d_i^*(\boldsymbol{\gamma}, \sigma)$. Given these definitions, it is clear that as $\sigma \to z$ then $d_i^*(\boldsymbol{\gamma}, \sigma) \to d_i^*(\boldsymbol{\gamma}, z)$ for all i, that is $d_i^*(\boldsymbol{\gamma}, \sigma)$ is continuous in σ. Specifically, $d_i^*(\boldsymbol{\gamma}, \sigma) \to d_i^*(\boldsymbol{\gamma}, 0)$ as $\sigma \to 0$. Hence, as $w(G^*, \boldsymbol{\gamma}) < w(G^*, \boldsymbol{\gamma}^{'})$ for $\sigma = 0$, it follows that there exists some $\bar{\sigma}$ such that if $\sigma \leq \bar{\sigma}$, $w(G^*, \boldsymbol{\gamma}) \leq w(G^*, \boldsymbol{\gamma}^{'})$ and so the statement is proved.

A.5. Proof of Proposition 3 (Page 12).

First note that, in any equilibrium $(X_\theta^*(G^*), G^*)$, it must be the case that if $\gamma_{ij} > \gamma_{ik}$ if $ik \in g$ for some g realised with positive probability then $ij \in g$. To see this note that, suppose the converse, and there is some g_t realised with positive probability such that $ik \in g_t$ but $ij \notin g_t$. Then, for any $x_i^*(G)$, the firm can receive a higher profit by deviating to a new strategy, $G^{'}$, in which $\rho_t(G^{'}) = 0$ and $\rho_s(G^{'}) = \rho_t(G)$, which straightforwardly generates a higher payoff.

This statement is sufficient for us to begin to state the firm's optimal strategy in this case. Again, letting $\mathbb{Z}_n$ denote the set of integers $\{0, 1, ..., n\}$, the firm's problem is equivalent to, without loss of generality, the case where the firm instead chooses, for each i, a distribution $\tilde{Z}_i \in \Delta\mathbb{Z}_n$, where the datasets are ordered such that if a degree, $k < n$, is realised with positive probability according to a distribution $\tilde{Z}_i$, then the datasets of the k-highest value are the datasets i is connected to. We note then that $\frac{\partial \mathbb{E}_{G, X_\theta^*(G)}[\pi(\{\tilde{Z}_i\}_i, X_\theta^*(G);\mu,\gamma)]}{\partial \tilde{z}_{ik}} = \sum_{j=1}^{k} \gamma_{ij} - \kappa_i(X_\theta^*(G),\sigma)\mu k^2$.

Now, it is clear that the proofs of the Lemmas 1 and 4 (and 2 and 3, but they are not required here) above hold in the case where the firm chooses $\{\tilde{Z}_i\}_i$ rather than $\{Z_i\}_i$, as none of their proofs rely on the fact $\gamma_{ij} = \gamma_i$. Crucially, this fact establishes that: (a) $\mathbb{E}[d_i(Z_i)] \geq \mathbb{E}[d_j(Z_j)]$ if $i < j$ and all σ and; (b) $d_i(g) = d_i(g')$ for all g, g' realised with positive probability in any equilibrium.

Now suppose $jk \in g$ for some g realised with positive probability. Claims (a) and (b) above establish that $d_i(g) \geq d_j(g)$. Suppose that $\gamma_{is} > \gamma_{it}$, and so $\gamma_{js} > \gamma_{jt}$. Recall that if $it \in g$ for some g realised with positive probability then $is \in g$. The result immediately follows.

A.6. Proof of Proposition 4 (Page 13).

As we showed in the proof of Propostion 3, for any equilibrium $(X_\theta^*(G^*), G^*)$, it must be the case that if $\gamma_{ij} > \gamma_{ik}$ if $ik \in g$ for some g realised with positive probability then $ij \in g$.

This statement is sufficient for us to begin to state the firm's optimal strategy in this case. Again, letting $\mathbb{Z}_n$ denote the set of integers $\{0, 1, ..., n\}$, the firm's problem is equivalent to, without loss of generality, the case where the firm instead chooses, for each i, a distribution $\hat{Z}_i \in \Delta\mathbb{Z}_n$, where the datasets are ordered such that if a degree, $k < n$, is realised with positive probability according to a distribution $\hat{Z}_i$, then the datasets of the k-highest value are the datasets i is connected to. Under the horizontal differentiation assumption set out in the theorem, the above implies that if $i \in \mathcal{M}_l$, $s \in \mathcal{N}_l$ and $t \in \mathcal{N}_{l'}$ where $l \neq l'$, then when $g_{it} = 1$, $g_{is} = 1$ as well.

Let $|\mathcal{N}_l| = n_l$ and $i \in \mathcal{M}_l$. We note that $\frac{\partial \mathbb{E}_{G, X_\theta^*(G)}[\pi(\{\hat{Z}_i\}_i, X_\theta^*(G);\mu,\gamma)]}{\partial \hat{z}_{ik}} = f(k) - \kappa_i(X_\theta^*(G),\sigma)\mu k^2$, where:

$$f(k) = \begin{cases} k\gamma_H & \text{if } k \leq n_l \\ n_l\gamma_H + \gamma_L(k - n_l) & k > n_l \end{cases}.$$

Again, it is clear that the proofs of the Lemmas 1 and 4 (and 2 and 3, but they are not required here) above hold in the case where the firm chooses $\{\hat{Z}_i\}_i$ rather than $\{Z_i\}_i$, as none of their proofs rely on the fact $\gamma_{ij} = \gamma_i$. Specifically, we then know that for any agent i, $\exists \hat{\mu}_{ik}$ such that if $\mu > \hat{\mu}_{ik}$ then $\mathbb{E}[d_i(G^*)] < k \in [0, n]$. Hence, if, for all $i \in \mathcal{M}_l$, $\mu > \hat{\mu}_{ik^*}$ for $k^* = n_l$, then the result holds.