Trust and Fairness as Incentives for Compliance with Information Security Policies

Alok Gupta, Dmitry Zhdanov
Information and Decision Sciences Department
Carlson School of Management
University of Minnesota

Abstract.
We consider the problem of enforcing compliance with information security policies in organizations in order to mitigate insider threat. We show that compliance with security policies may be enforced even for myopic, self-interested, agents by providing them proper economic incentives for compliance. Our approach includes several variations of a compliance game between the organization and its inside users in which a bonus is paid for compliance with security policies. We show that compliance may be sustained by emphasizing the continuous, repeated nature of security-related decisions. Alternatively, compliance is more likely to emerge when costs and benefits of increased protection are shared in a fair manner. Our results emphasize the need to build trust between organizational entities, as well as suggest a way to determine compliance bonus in a fair manner.

1. Introduction
Widespread adoption of new technologies such as web services and methods of business operations such as telework and virtual teams has led to the disappearance of the traditional data security perimeter around modern firms. For example, an estimated 30 percent of enterprise software purchased in 2007 will support Web services standards (McCoy et al 2004). Access to sensitive data and computing infrastructure is an issue even when a firm does not extend its operational business environment beyond its traditional boundaries.

Furthermore, the severity of insider threat continues to grow; for example, in 2004, employee misconduct involving information systems was named as second major security concern behind viruses, trojans and worms (E&Y 2004). The problem is becoming even more challenging with the growing trend of firms outsourcing their information systems development and management, since the number and variety of individuals with “insider” system knowledge increases.

In this paper, we show that compliance with security policies may be enforced even for myopic agents by providing them proper economic incentives for compliance. We consider a game between an organization and a representative internal user of its security system, and show how compliance may be achieved with help of two different approaches: i) a repeated game setting where the organization and its users interact in an infinite time horizon; and ii) a novel approach in information systems research -- games with fairness consideration. We show that both the approaches considered in this paper can result in compliance with security policies even for “economically rational” agents, who tend to avoid compliance due to cost and loss of productivity. We then consider the bargaining power of both parties and show how appropriate economic incentives may be determined in a fair manner.

2. Model Constructs: Role of Trust in Assuring Security Policy Compliance
First, let us develop a game theoretic model for a situation where an organization wants to ensure that its users comply with its security policies. Lack of compliance by the users increases the organization’s exposure to security breaches, resulting in expected losses of $z$. This amount can be envisioned as the amount that the organization has to reserve to implement additional security measures or to be prepared to use this money in clean-up and repair activities after a security breach. We assume that non compliance with security policies always leads to organization suffering a monetary loss of $z$. On the contrary, in case of the user compliance with security policies security risks are countered and the organization can re-deploy previously reserved resources. Therefore, in this case the organization experiences a gain of $z$. 
For a given user, however, compliance with security policies is costly because it requires commitment of additional monetary and human resources which ultimately may reduce its own productivity. We represent the cost associated with compliance to security policies as $c$. The users may not see immediate ways to redistribute increased protection to everyone, i.e., they anticipate no direct gain from improved security. However, when users do not comply with security policies, from a myopic perspective, there is no cost of non compliance with the organization’s security policies.

Because of the conflicting interests and different perceptions of value for security, the organization may need to offer additional incentives to the user to ensure compliance with security policies. It is generally possible to provide a performance incentive in form of a bonus payment for policy compliance, or fine for lack of compliance, or both. As Gupta and Zhdanov (2006) show, in many cases these incentives are equivalent in their effect. Therefore, we focus on just one of them – a bonus payment.

Let us now formally define the compliance game between a user and an organization. In this section we assume that the organization and the user are in a single-period simultaneous move game. The user may choose to comply with security policy at cost $c$ or chooses not to comply at zero cost; its strategy set can be represented by $S_u=(\text{Compliance, No Compliance})$. The organization, on the other hand, may choose to provide bonus of $b$ to the user for expected compliance or no bonus if it expects non-compliance; therefore, its strategy set can be represented by $S_o=(\text{Bonus, No Bonus})$. To ensure individual rationality, we require that double inequality $c < b < 2z$ holds, i.e., bonus received by the user must exceed its cost of compliance; however, the bonus cannot exceed the benefits retained by the organization. We can then represent the payoff vector, $S_o \times S_u \rightarrow \mathbb{R}$, in the Normal form game between the user and the organization as follows (payoffs of the user; payoff of the organization):

<table>
<thead>
<tr>
<th>User\Org.</th>
<th>Bonus</th>
<th>No Bonus</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Compliance</td>
<td>$+b$; $-z-b$</td>
<td>$0$; $-z$</td>
</tr>
<tr>
<td>Compliance</td>
<td>$+b-c$; $z-b$</td>
<td>$-c$; $z$</td>
</tr>
</tbody>
</table>

It is easy to verify that the Nash Equilibrium of this game is strategy pair $(\text{No Compliance, No Bonus})$. While it is a choice that should be made by rational players, it is obvious that this is not the most preferred alternative. It is easy to verify that if the players achieve an outcome $(\text{Compliance, Bonus})$, both of them will be better off, but this outcome cannot be achieved in a single-period game. However, security-related interactions rarely take place just once. Real-world organizations periodically monitor its partners' behavior and change its own security policies to adapt to new threats and solutions. Suppose that compliance game is played repeatedly between the user and the organization. Then, if both players use the same “time value of money” factor $g$ to discount the payoffs from future game periods, conditions for voluntary compliance with security policies are specified in Theorem 1.

**Theorem 1.** (Characterization of feasible bonus amount) Suppose that providing a bonus in the amount of $b$ is feasible. Then, if the choice of reward $b$ is inducing incentive compatibility (voluntary adherence with security policy), it may be characterized as follows:

\[
\frac{c}{g} < b < 2zg.
\]

**Corollary 1A.** (Existence of guaranteed discount rate that induces incentive compatibility).

Suppose that provisioning of bonus is feasible. Then, there exist a lower bound on time value of money factor $g$ that induces mutual incentive compatibility. This critical value of $g$ is $\frac{1}{\sqrt{2}}$.

Proof of Theorem 1 comes from the analysis of trigger strategies (Friedman 1971). Results of Theorem 1 and its Corollary 1A suggest that despite all uncertainty and threat of opportunistic behavior,

---

1 This may be an individual employee, structural unit of a company or its business partner. We focus on the relationship between formal entities (departments, firms).

2 The proofs of theorems are in the appendix.
there is a practically significant range of potential values of discount factors for which it is possible to build trust relationships between the user and the organization and insure the provision of appropriate level of security. Moreover, such cutoff level is independent of a particular configuration of agents’ payoffs.

However, assumption of infinite horizon of the compliance game may be questionable. Thus, it is important to find alternative mechanisms of compliance that are rooted in the real world actions of decision makers. One such mechanism concerns with issues of fairness and reciprocity. It has been shown in experimental economics as well as in decision science literature (see Rabin, 1993 for an extended discussion) that people deviate from predictions of conventional game theory and consider the attitude of other players in creating their strategies. In the next section, we provide an overview of fairness in game theory and apply these concepts to the problem of security compliance.

3. Constructs of Fairness

Since fairness is typically not a well known construct in analyzing games in IS literature (except for Bapna et al., 2005 who do not look at fairness in game theoretic sense), we first introduce some modeling constructs that are unique to games with fairness equilibrium. Formally, concept of fairness equilibrium (Rabin 1993) is somewhat similar to Bayesian equilibrium, as it includes beliefs about other players. However, fairness equilibrium directly includes beliefs in players’ payoffs. In fairness equilibrium such belief is about how kindly the other player is treating the player in question. Belief about the attitude of a counterpart may change behavior of people as well as organizations (for example, an employee who expects to be fired any day may behave differently from one hired for a lifetime). Clearly, such beliefs play a prominent role in information security settings, as majority of security incidents are originated from inside organizations (33% by an employee and another 20% by a former employee – Berinato 2005)

Suppose there is a game between the organization and the user with strategy sets $S_o$ and $S_u$ respectively. Let $\pi_i$ be player $i$’s material payoff from playing the game: $\pi_i: S_o \times S_u \rightarrow \mathbb{R}$. To design fairness equilibrium, it is important to consider three components: player’s own strategy (A), her opponent’s belief about her strategy (B), and her belief about opponent’s belief about her strategy (C). Indexes refer to strategy sets of players. At equilibrium, all beliefs must match the actual behavior of players. If in the compliance game described in previous section, fairness equilibrium is (Compliance, Bonus), then $A_o^* = B_o^* = C_o^* = \text{Compliance}$ and $A_u^* = B_u^* = C_u^* = \text{Bonus}$, with indexes “o” and “u” denoting the organization and the user, respectively.

However, confirmation of beliefs is only part of fairness equilibrium; in the same way as in traditional games, actual strategies played must maximize each player’s utility. That utility, in turn, depends partially on each player’s material payoff and partially on the kindness of the user and the organization to each other. In order to define player’s kindness, it is necessary to consider the range of possible payoffs to each player. Then the following payoffs are important:

- $\pi_i^{H}(a_i)$ – player $i$’s highest possible payoff from playing strategy $a_i$
- $\pi_i^{L}(a_i)$ – player $i$’s lowest payoff from playing strategy $a_i$ among those outcomes which are not Pareto-dominated
- $\pi_i^{E}(a_i)$ – player $i$’s equitable payoff from playing strategy $a_i$. It is defined as an average of $\pi_i^{H}(a_i)$ and $\pi_i^{L}(a_i)$:

$$\pi_i^{E}(a_i) = (\pi_i^{H}(a_i) + \pi_i^{L}(a_i))/2$$

- $\pi_i(a_i, b_j)$ – player $i$’s payoff if she plays strategy $a_i \in S_i$ and player $j$ plays strategy $b_j \in S_j$.

Following Rabin (1993), we define the equitable payoff of a compliance game as an average of highest and lowest payoffs, pertaining to each player. For example, an equitable payoff for the user may be thought of as a payoff that a she will get while randomly playing against ultimately kind or ultimately unkind organization. Equitable payoff serves as a benchmark of kindness – any payoff above it is
considered kind, and below it is considered unkind. Actual values of characteristic payoffs of the compliance game are presented in Tables 1 and 2 below.

Let \( f_j(a_i, b_j) \) define player \( j \)’s kindness with respect to player \( i \)’s chosen strategy \( a_i \in S_i \). If player \( j \) chooses strategy \( b_j \in S_j \), then her kindness function is

\[
f_j(a_i, b_j) = \frac{\pi_j(b_j, a_i) - \pi_j^E(a_i)}{\pi_i^H(a_j) - \pi_i^\min(a_j)}
\]

(1)

Negative values correspond to cases when a player is being hostile to the opponent, while positive values represent some degree of benevolence. Zero value of kindness function is achieved when the player is giving the opponent her equitable payoff (the average of highest and lowest realistic payoff expectation for the opponent). For example, if the organization is always providing a bonus to the user, it is increasing the user’s payoff and is being kind. Therefore, the values of the organization’s kindness function associated with Bonus strategy should be positive, e.g., \( f_o (Bonus, Compliance) = f_o (Bonus, No Compliance) > 0 \)

If we replace actual strategies with players’ beliefs about opponent strategies in kindness function, we will obtain the “kindness belief function” Let \( \tilde{f}_j(b_i, c_j) \) define player \( j \)'s belief about kindness of player \( j \) to her. We may then define each player’s utility function as follows:

\[
U_i = \pi_i(a_i, a_j) + \alpha_i \beta_i \pi_j (a_j, a_i),
\]

(2)

Where

\( \pi_i(a_i, a_j) \) – player’s own material payoff

\( \pi_j (a_j, a_i) \) – opponent’s material payoff

\( \alpha_i \) – measure of opponent’s helpfulness

\( \beta_i \) – degree of player’s concern about opponent’s payoff (\( \beta_i > 0 \))

\( a_j, a_i \) – players’ strategies; \( i, j \) – player index (the user or the organization)

In the next subsection, we revisit the compliance game between the user and organization and analyze it from the perspective of fairness.

4. Fairness Analysis of Compliance with Security Policies

To perform the fairness analysis of compliance game, we need to identify the characteristic payoffs of players as described above. These payoffs directly follow from the game matrix and are presented in the Tables 1 and 2:

<table>
<thead>
<tr>
<th>Table 1. Characteristic payoffs for the organization</th>
<th>Table 2. Characteristic payoffs for the user</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy</td>
<td>( \pi_o^H )</td>
</tr>
<tr>
<td>Bonus</td>
<td>( z-b )</td>
</tr>
<tr>
<td>No bonus</td>
<td>( z )</td>
</tr>
</tbody>
</table>

It may be easily verified that players are either identically kind or identically unkind to each other. Therefore, we assume that \( \alpha \) is either 1 or -1, depending on choice of strategies. With these simplifications, we can identify conditions necessary for voluntary compliance with security policies when fairness is considered. This result is presented in Theorem 2.

**Theorem 2.** Strategy choice \( (Compliance, Bonus) \) may be sustained as a fairness equilibrium of the compliance game if \( \beta_o > 1 \) and \( \beta_u > c/(2z) \).
We should note that bonus $b$ is not part of the equilibrium solution. Fairness equilibrium is possible for any feasible bonus amount; what matters is the proportion of players’ costs and benefits that are inherent in the game. Still, the amount of bonus paid to the user by the organization may serve as an indication of how fairly costs and benefits are shared. From the perspective of security practitioners, amount of bonus can not be directly tied to the outcomes of security efforts: on one hand, even the most compliant of users is not completely guaranteed from security incidents beyond her control (e.g., due to operating system vulnerabilities); on another hand, absence of detected incidents does not mean there is full compliance with security policies. Therefore, there is room for negotiation about what the appropriate level of effort should be on the part of the user and how it should be compensated by the organization. Next, we consider the problem of bargaining between players about the amount of bonus.

We analyze a simultaneous-move game based on original compliance game, which proceeds as follows. First, the organization selects amount of bonus $b$ that it will offer ($c < b < z$). At the same time, the user selects “reservation level” $r$, which is the minimum bonus amount it will accept. Both players announce their choices simultaneously. If $r > b$, then outcome (No Compliance, No Bonus) is achieved – payoffs to players (the user, the organization) are $(0, -z)$. If $r < b$, outcome (Compliance, Bonus) is achieved and payoffs to players are $(b-c, z-b)$.

Let $q$ denote a Nash equilibrium of this game. It is easy to verify that multiple Nash equilibria are possible in this game in the form of $(b=r=q)$, where $c < q < z$. However, not all of them will be fairness-based. To see this, we need to construct kindness functions for both players. First, kindness of the user towards the organization may be either: a) $-1$ when the user does not accept the contract; or b) zero, when the user accepts the contract and essentially allows the organization to keep the surplus in excess of bonus amount, decided by the organization. Organization’s kindness is a function of bonus $b$ and is defined in Theorem 3.

**Theorem 3.** Organization’s kindness function is defined as $f_o = (2b+c-z)/(2z-2c)$.

Note that $f_o$ is equal to zero when bonus is an equal split of mutual costs and benefits, i.e., $b = (z-c)/2$. Then, by analyzing the fairness-based utility for the user, we obtain the following characterization of a fair bonus amount:

**Theorem 4.** Lowest “fair” bonus amount that ensures the fairness equilibrium in bargaining game between a user and an organization is characterized as

$$b^* = \max(c; \frac{z-c+2zc-2c^2}{2(z-c+1)} = \begin{cases} \frac{c}{c \geq z/3} \\
\frac{z-c+2zc-2c^2}{2(z-c+1)} : c < z/3 \end{cases}$$

**Corollary 4A.** Larger values of the organization’s benefit $z$ will result in higher values of fair bonus $b^*$.

**Corollary 4B.** Values of the user’s cost $c$ that are relatively small compared to the organization’s benefit $z$ will result in higher values of fair bonus $b^*$, relative to the user’s cost $c$.

These corollaries have straightforward practical implications. On one hand, higher values of the organization’s benefit $z$ set a higher standard to which the user compares its bonus offer. On the other hand, magnitude of the user’s own cost compared to the organization’s benefit also matter for the user. When the difference between organizational benefit $z$ and the user’s cost $c$ is large (or $c$ is relatively small), the user may feel compelled to forgo a bonus that is a small increment over her cost because she considers this amount unfair. The analysis in this section indicates that the organization cannot drive the user surplus to zero.
5. Conclusions

In this paper, we have considered the problem of enforcing internal compliance with information security policies in organizations. Compliance with security policies is a pervasive modern day problem, which cannot be fully addressed using technical solutions. It is important to study economical and organizational incentives that can reduce the potential of “non-compliant insider” threat to information systems. Our results indicate that while compliance is possible to achieve when multiple period games are considered. In other words, if the entities involved in information security interactions are assured that these interactions will be continued in the future, it is possible to enforce compliance due to multiple payoffs in the future. We also consider the issues of fairness and find that compliance with information security policies may be sustained if organization and its agents are treating each other in a fair manner. We also find that the organization and its users have incentives to bargain over the amount of bonus that should be given for compliance; however, bargaining power of parties is also restricted when fairness is considered.

A potential limitation of this work is the choice of payoff structure. We should note that the current structure of payoffs is an appropriate representation of the situation when the user has no malicious intent. In this case, the non-compliance with security policies may be a result of factors such as carelessness or cognitive overload of the user. Should the user have explicit intention to do damage to the organization, it would be more appropriately modeled with much higher costs (e.g., prosecution) as well as payoffs (e.g., obtaining confidential information for sale). Also, the security breaches may have negative impact on individual users rather than organization as a whole; we are currently developing the analysis of this case.

References
Appendix. Proofs of Theorems.

**Theorem 1. (Characterization of feasible bonus amount)** Suppose that providing a bonus in the amount of $b$ is feasible. Then, if the choice of reward $b$ is inducing incentive compatibility (voluntary adherence with security policy), it may be characterized as follows: $\frac{c}{g} < b < 2zg$.

**Proof.**

In order to insure feasibility, we need to identify conditions under which neither player has an incentive to deviate from (Compliance, Bonus) strategies. Let $t$ be the current period of the game (i.e., it has already been played $t-1$ times). In Nash reversion (trigger) strategy, both players will choose mutually beneficial outcome (Compliance, Bonus) in current period $t$, if both of them played this strategy in previous period $t-1$. If one of the players deviates from this choice in current period, in period $t+1$ the opponent will start “punishing” this player by forcing the mutually inefficient outcome (No Compliance, No Bonus) thereafter.

First, consider the strategies and payoff for the user. If she does not deviate from Compliance strategy, then (Compliance, Bonus) outcome occurs infinitely, and her payment stream is:

$$(b-c) \cdot (1+g+g^2+g^3+\ldots) = \frac{(b-c)}{(1-g)} \quad (A1)$$

If she deviates and chooses No Compliance strategy, then (No Compliance, Bonus) outcome occurs once, giving the user a single-period payoff of $b$ instead of $b-c$. In the next period, the organization does not believe in good will of the user any longer, and plays No Bonus strategy infinitely. The user’s best response to the organization’s anticipated strategy is No Compliance, and her payment stream is:

$$b + 0 \cdot (g+g^2+g^3+\ldots) = b \quad (A2)$$

Therefore, the user has no incentive to deviate from Compliance strategy if her payoff from deviation is too low:

$$\frac{(b-c)}{(1-g)} > b, \text{ or } g > \frac{c}{b}. \quad (A3)$$

Thus, as the user’s cost of implementing high security is close to the amount of bonus, she will not deviate from Compliance strategy if sustainability of future payments is high.

On the other hand, from the organization’s perspective, payoff from non-deviation and infinitely repeated (Compliance, Bonus) outcome is:

$$(z-b) \cdot (1+g+g^2+g^3+\ldots) = \frac{(z-b)}{(1-g)} \quad (A4)$$

If the organization chooses to deviate from Bonus strategy, then outcome (Compliance, No Bonus) occurs once, giving the organization a single-period payoff of $z$ instead of $z-b$. In the following periods, the user reverts to playing No Compliance strategy, thus forcing the organization to play No Bonus strategy. Therefore, the organization’s payoff from a single-period deviation is:

$$z + (-z) \cdot (g+g^2+g^3+\ldots) = z \cdot \frac{(1-2g)}{(1-g)} \quad (A5)$$

Therefore, the organization will not deviate from Bonus strategy if payoff from deviation is too low:

$$\frac{(z-b)}{(1-g)} > z \cdot \frac{(1-2g)}{(1-g)}; \text{ or } g > \frac{b}{2z} \quad (A6)$$

Assumption of incentive compatibility implies that there is no need to deviate from equilibrium outcome for either the user or the organization. Thus, conditions (A3) and (A6) hold.

Rearranging those inequalities and combining them, we obtain $\frac{c}{g} < b < 2zg$.

**Q.E.D.**
Corollary 1A. (Existence of guaranteed discount rate that induces incentive compatibility). Suppose that provisioning of bonus is feasible. Then, there exist a lower bound on time value of money factor \( g \) that induces mutual incentive compatibility. This critical value of \( g \) is \( 1/\sqrt{2} \).

**Proof.**
If provision of bonus is feasible, then double inequality \( c/g < b < 2zg \) holds. From it, we can write that
\[
\frac{c}{g} < 2zg, \text{ or } g^2 > \frac{c}{2z}. \tag{A7}
\]
However, by construction, \( c \) is smaller than \( z \) to induce feasibility. Thus, \( c/(2z) < \frac{1}{2} \).
Substituting this result into (A7), we conclude that all time value of money factors that satisfy the inequality \( g^2 \geq \frac{1}{2} \) (or, equivalently, \( g \geq \frac{1}{\sqrt{2}} \)), automatically induce incentive compatibility.

Q.E.D.

Theorem 2. Strategy choice (Compliance, Bonus) may be sustained as a fairness equilibrium of the compliance game if \( \beta_o > 1 \) and \( \beta_u > c/(2z) \).

**Proof.**
Since the strategy pair (Compliance, Bonus) is characterized by positive kindness of both players, \( \alpha_o = \alpha_u = 1 \). Therefore, the organization’s utility is
\[
U_o (Compliance, Bonus) = z - b + 1 \cdot \beta_o (b - c) = z - b + \beta_o b - \beta_o c \tag{A8}
\]
If organization unilaterally deviates, its utility will be
\[
U_o (Compliance, No Bonus) = z - b + 1 \cdot \beta_o (b - c) = z - \beta_o c \tag{A9}
\]
And, the organization has no incentive to deviate if
\[
U_o (Compliance, Bonus) > U_o (Compliance, No Bonus) \equiv \beta_o > 1. \tag{A10}
\]
Similarly for the user,
\[
U_u (Compliance, Bonus) > U_u (No Compliance, Bonus) \equiv \beta_u > c/(2z). \tag{A11}
\]
Q.E.D.

Lemma 3. Organization’s kindness function is defined as \( f_o = (2b + c - z)/(2z - 2c) \)

**Proof.**
To define kindness function for the organization, we need to consider all relevant payoffs for the user. They are:
\[
\pi_u^H (r = b = q) = \pi_u (b = z) = z - c \tag{A12}
\]
Thus, the highest payoff is when the organization gives its entire benefit as bonus
\[
\pi_u^E (r = b = q) = \pi_u \min (r = b = q) = \pi_u (b = c) = 0 \tag{A13}
\]
In other words, lowest payoff is when the organization only covers the user’s cost as bonus
\[
\pi_u^E (r = b = q) = (z - c)/2 \tag{A14}
\]
Therefore, kindness of the organization on the Nash equilibrium path may be defined as:
\[
f_o(q, q) = \frac{\pi_u(q, q) - \pi_u^E (r = b)}{\pi_u^H (r = b) - \pi_u \min (r = b)} = \frac{2b + c - z}{2(z - c)} \tag{A15}
\]
Q.E.D.
Theorem 4. Lowest “fair” bonus amount that ensures the fairness equilibrium in bargaining game between a user and an organization is characterized as

\[
b^* = \max(c; \frac{z-c+2zc-2c^2}{2(z-c+1)}) = \begin{cases} 
  c : c \geq z/3 \\
  \frac{z-c+2zc-2c^2}{2(z-c+1)} : c < z/3 
\end{cases}
\]

Proof.
Suppose that the user’s fairness-based utility function is

\[
U_u(r,b) = \pi_u(r,b) + f_o(r,b)[1 + f_u(r,b)]
\]

(A16) 1

Recall that the user’s kindness is equal to -1 in case of no compliance and 0 in case of compliance, and kindness of the organization is defined in Lemma 3. Therefore, if the user adopts No Compliance strategy, her utility is:

\[
U_u(r < b) = \pi_u(r < b) + f_o(r < b)[1 + f_u(r < b)] = 0 + f_o(r = b) \cdot [1 - 1] = 0
\]

(A17)

Her utility in case of compliance is

\[
U_u(r = b) = \pi_o(r = b) + f_o(r = b)[1 + f_u(r = b)] = b - c + f_o(r = b) \cdot [1 + 0] = b - c + \frac{2b + c - z}{2(z-c)}
\]

(A18)

To ensure that fairness equilibrium (Compliance, Bonus) holds, utility for the user must exceed that of playing the No Compliance strategy:

\[
U_u(r = b) > 0
\]

(A19)

Substituting the actual value of the user’s utility from (A18), we obtain:

\[
b - c + \frac{2b + c - z}{2(z-c)} > 0
\]

(A20)

By multiplying through by a positive number 2(z-c) and rearranging to solve for b, we obtain:

\[
b^* > \frac{z - c + 2zc - 2c^2}{2(z-c+1)}
\]

(A21)

Next, we need to make sure that providing this bonus is feasible (c < b* holds). Substituting value for b* from (A21) and solving for c, we obtain that bonus b* is feasible if:

\[
c < \frac{z}{3}
\]

(A22)

Otherwise, the minimum acceptable bonus is the user’s cost c.

Q.E.D.

Corollary 4A. Larger values of the organization’s benefit z will result in higher values of fair bonus b*.

Corollary 4B. Values of the user’s cost c that are relatively small compared to the organization’s benefit z will result in higher values of fair bonus b*, relative to the user’s cost c.

Proofs of both corollaries are straightforward from analysis of function

\[
b(z, c) = \frac{z - c + 2zc - 2c^2}{2(z-c+1)}
\]

(A23)

restricted for cases when c < z. Its partial derivatives are:

\[
\frac{\partial b}{\partial z} = \frac{1 + 2c}{2(z-c+1)^2}
\]

(A24)
\[
\frac{\partial b}{\partial c} = \frac{2z - 4c - 1 + 2z^2 - 4zc + 2c^2}{2(z - c + 1)^2}
\]  
(A25)

Partial derivative in \( z \) is positive, proving Corollary 4A.
Partial derivative in \( c \) suggests presence of two possible extremum points:
\[ c = 1 + z + \frac{1}{2}\sqrt{6 + 4z} \quad \text{and} \quad c = 1 + z - \frac{1}{2}\sqrt{6 + 4z}. \]
First of these points is a local minimum, but it lies outside of domain \( c \leq z \). Second is a local maximum lying within the domain \( c < z \). Thus, as \( c \) approaches \( z \), values of \( b^* \) eventually start to decrease.

Q.E.D.

---

1 This is one of possible fairness-based utility functions as defined by Rabin (1993). We choose this particular shape to emphasize the role of kindness itself and abstract from how players weigh the importance of opponent’s payoffs.

The general form of such utility is \( U_1(a_1, b_2, c_1) = \pi_1(a_1, b_2) + G(X) \cdot f_2(b_2, c_1) \cdot (1 + f_1(a_1, b_2)) \), where \( G(X) \) is positive and increasing in \( X \), and \( X \) does not influence players’ strategies.