The Effect of Decentralized Behavioral Decision Making on System-Level Risk

Kim Kaivanto
Department of Economics
Lancaster University

In computer networks, system-level risk depends on the actions and choices of a collection of ‘lay’ users.

Q. How should we model the decision making of these lay users ?

A. PT-SDT with psychology of deception effects.

Q. Does it matter whether our modeling assumptions reflect normative rationality or heuristics & biases?

A. YES. (See comparative statics and simulation results)

Outline

Classical SDT under normative rationality

Behavioral factors

- From the decision-making literature: CPT
- From the phishing & deception literatures

Re-derivation of optimal cutoff threshold under CPT-SDT

- Using T&K92 probability weighting function
- Using neo-additive probability weighting function
- Incorporating the psychology of deception

Beyond comparative statics: comparative simulation results

Implications

- Spam filtering
- Education & training

End

Classical SDT

Elements:

- a 'score' variable $\theta = \Gamma(\mathcal{I})$, $\theta \in \Theta = [\underline{\theta}, \overline{\theta}] \subset \mathbb{R}$
- known distributions of the score under D and $\neg D$
- as the cutoff threshold θ' is varied, traces out the Receiver Operating Characteristics (ROC) curve

Task:

identify optimal cutoff threshold $\theta^* \in \Theta = [\underline{\theta}, \overline{\theta}] \subset \mathbb{R}$
such that the observed score $\theta = \Gamma(\mathcal{I})$
is either in the acceptance interval $(\underline{\theta} \leq \theta \leq \theta^*)$
or in the rejection interval $(\theta^* < \theta \leq \overline{\theta})$
where the null hypothesis $H_0 : \neg D \Leftrightarrow H_0 : \theta_{\Gamma(\mathcal{I})} \leq \theta^*$
and the alt. hypothesis $H_1 : D \Leftrightarrow H_1 : \theta_{\Gamma(\mathcal{I})} > \theta^*$

Problem:

choose the optimal cutoff threshold θ^* by solving

$$\min_{\theta'} E(C) \qquad \text{s.t.} \qquad \text{TPL} = G(\text{FPL})$$

Classical SDT

The expected cost of using the SDT mechanism is:

$$\begin{aligned} E(C) &= C + C_{\mathrm{TP}} P(\mathrm{TP}) + C_{\mathrm{FN}} P(\mathrm{FN}) + C_{\mathrm{TN}} P(\mathrm{TN}) + C_{\mathrm{FP}} P(\mathrm{FP}) \\ &= -[C_{\mathrm{FN}} - C_{\mathrm{TP}}] P(D) \mathrm{TPL} + [C_{\mathrm{FP}} - C_{\mathrm{TN}}] P(\neg D) \mathrm{FPL} \\ &\quad + C + C_{\mathrm{TN}} P(\neg D) + C_{\mathrm{FN}} P(D) \quad . \end{aligned}$$

Setting the total differential of expected cost to zero

$$\mathrm{d}E(C) = -[C_{\mathrm{FN}} - C_{\mathrm{TP}}] P(D)\, \mathrm{dTPL} + [C_{\mathrm{FP}} - C_{\mathrm{TN}}] P(\neg D)\, \mathrm{dFPL} = 0$$

it follows that the slope of each iso-*E*(*C*) line is the probability weighted ratio of the incremental cost of misclassifying a non-malicious email to the incremental cost of misclassifying a malicious email

$$\left(\frac{\mathrm{dTPL}}{\mathrm{dFPL}} \right)_{\bar{C}} = \frac{P(\neg D)}{P(D)} \left[\frac{C_{\mathrm{FP}} - C_{\mathrm{TN}}}{C_{\mathrm{FN}} - C_{\mathrm{TP}}} \right] = \left(\frac{\mathrm{dTPL}}{\mathrm{dFPL}} \right)_{\bar{C}^*}$$

Classical SDT

The optimal cutoff threshold θ^* is a function of

- a misclassification cost matrix
- the baserate odds of (email) being non-malicious
- risk preferences

n.b. There is no reason for the misclassification costs to be the same for the user as for the organization

n.b. Classical SDT admits that costs can be replaced by their utilities, but in fact proceeds (exclusively) with minimizing expected cost, i.e. assuming *risk neutrality*.

n.b. In the literature, the 'optimal classifier' is computed under risk neutrality (!)

Behavioral factors

Descriptively, behavioral decision makers display

- -reference dependence, framing effects
- -non-linear probability weighting
- -loss aversion
- -ambiguity aversion
- -four-fold pattern of risk aversion

All incorporated in (cumulative) Prospect Theory (PT)

Deception deploys employ

- -peripheral-route persuasion
  - -authority, scarcity, similarity & identification, reciprocation, consistency, social proof
- -visceral emotions
- -urgency
- -contextual cues

PT extension of SDT

Using a conventional (cumulative) PT specification

$$x_{-m} < \cdots < x_0 < \cdots < x_n$$

$$V(\boldsymbol{x}, \boldsymbol{p}) = V^{+}(\boldsymbol{x}, \boldsymbol{p}) + V^{-}(\boldsymbol{x}, \boldsymbol{p})$$

$$V^{-}(\boldsymbol{x}, \boldsymbol{p}) = w^{-}(p_{-m})v^{-}(x_{-m}) + \sum_{k=1}^{m}\left[w^{-}\left(\sum_{j=0}^{k} p_{-(m-j)}\right) - w^{-}\left(\sum_{j=0}^{k-1} p_{-(m-j)}\right)\right] v^{-}(x_{-(m-k)})$$

$$v(x) = \begin{cases} x^{\phi^{+}} & x \geq 0 \\ -\lambda(-x)^{\phi^{-}} & x < 0 \end{cases} \qquad \begin{array}{l} \phi^{-} = 0.88 \\ \lambda = 2.25 \end{array}$$

$$w^{-}(p) = \frac{p^{\delta}}{(p^{\delta} + (1-p^{\delta}))^{1/\delta}} \qquad \delta = 0.69$$

PT extension of SDT

Assume $C_{\mathrm{FN}} > C_{\mathrm{TP}} > C_{\mathrm{FP}} > C_{\mathrm{TN}} \geq 0.$

Further, wlog $C_{\mathrm{TN}} = 0$

- this becomes the PT reference point

Then the PT value function, in terms of α and β:

$$\begin{aligned} V^{-}(C) \;=\; & -\; w^{-}(p\beta)\lambda v(C_{\mathrm{FN}}) \\ & -\; \left[w^{-}(p) - w^{-}(p\beta)\right]\lambda v(C_{\mathrm{TP}}) \\ & -\; \left[w^{-}(p+(1-p)\alpha) - w^{-}(p)\right]\lambda v(C_{\mathrm{FP}}) \\ & -\; \left[w^{-}(1) - w^{-}(p+(1-p)\alpha)\right]\lambda v(C_{\mathrm{TN}}) \end{aligned}$$

PT extension of SDT with neo-additive pwf

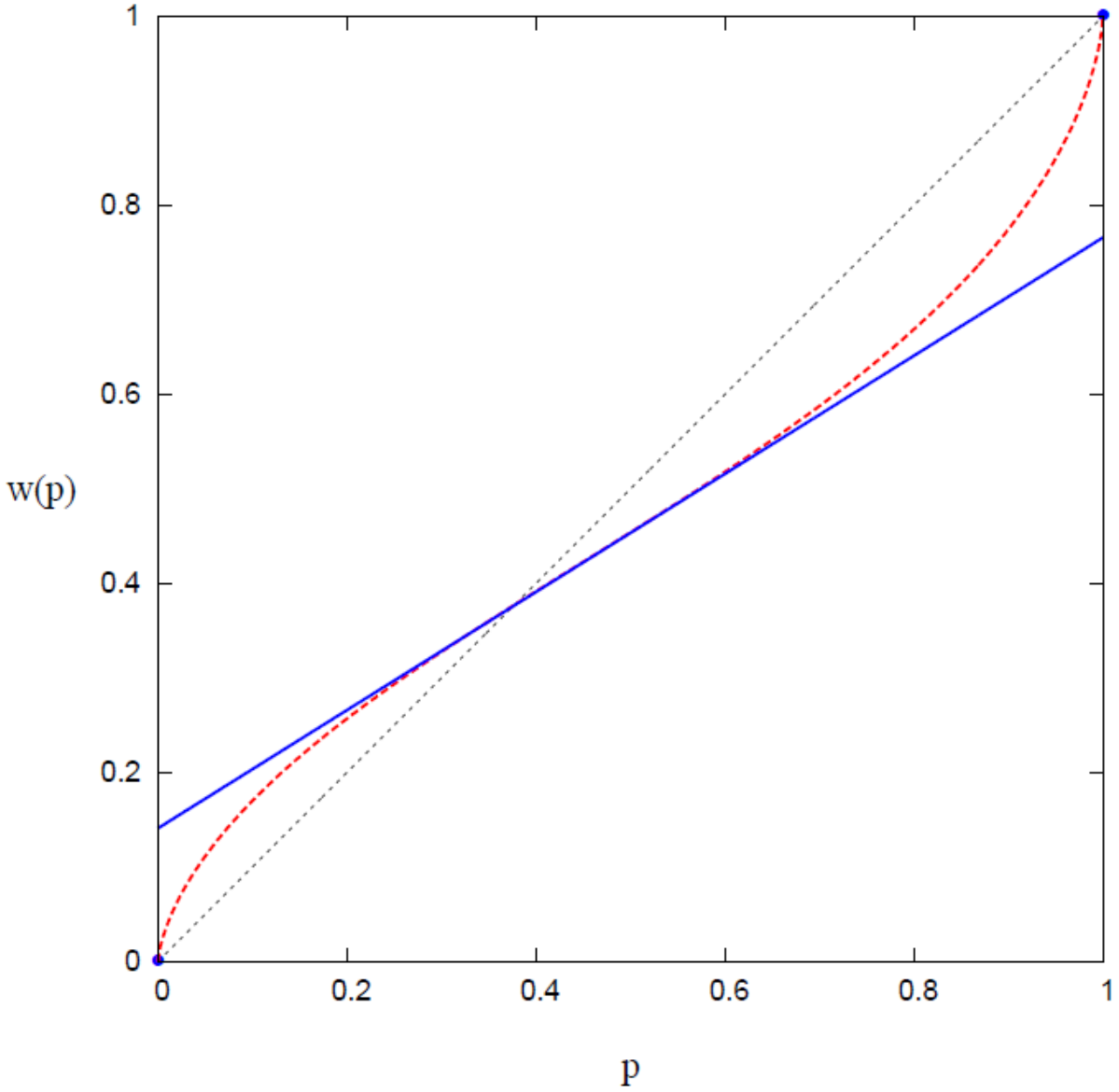
Consider the piece-wise linear *neo-additive* pwf

$$w_{\text{n-a}}(p) = \begin{cases} 0 & \text{for } p = 0 \\ \mathsf{a}\,p + \mathsf{b} & \text{for } 0 < p < 1 \\ 1 & \text{for } p = 1 \end{cases} \qquad 0 \leq \mathsf{b} < 1, \;\; 0 < \mathsf{a} \leq 1 - \mathsf{b}$$

"among the most promising candidates regarding the optimal tradeoff of parsimony and fit" (Wakker, 2010).

Captures the *possibility effect*, the *certainty effect*, the *overweighting of small probabilities*, and *the underweighting of large probabililties.*

PT extension of SDT with neo-additive pwf



PT extension of SDT with neo-additive pwf

Solving for the slope of the iso-$V^{-}_{\text{n-a}}(C)$ contours in ROC space

$$\frac{\text{dTPL}}{\text{dFPL}} = \left[\frac{(C_{\text{FP}})^{\phi^-} - (C_{\text{TN}})^{\phi^-}}{(C_{\text{FN}})^{\phi^-} - (C_{\text{TP}})^{\phi^-}}\right] \cdot \left(\frac{1-p}{p}\right)$$

The iso-$V^{-}_{\text{n-a}}(C)$ contours are staight lines, just as in classical SDT, but they are *steeper*:

$$\left[\frac{(C_{\text{FP}})^{\phi^-} - (C_{\text{TN}})^{\phi^-}}{(C_{\text{FN}})^{\phi^-} - (C_{\text{TP}})^{\phi^-}}\right] > \left[\frac{C_{\text{FP}} - C_{\text{TN}}}{C_{\text{FN}} - C_{\text{TP}}}\right]$$

PT-SDT is more *conservative* than classical SDT!

PT-SDT with psychology of deception

Successful deception deploys:

-peripheral-route persuasion
-visceral emotions
-urgency
-contextual cues

The deception-perpetrator's skill $K \in \mathbb{R}_+$ and effort $e \in \mathbb{R}_+$.

Mark *i*'s ploy-specific discriminability at time *t*:

$$d' = d_{it}(K, e)$$

$$\frac{\mathrm{dAUC}}{\mathrm{d}d'} \frac{\partial d'}{\partial K} \leq 0 \quad , \quad \frac{\mathrm{dAUC}}{\mathrm{d}d'} \frac{\partial d'}{\partial e} \leq 0$$

PT-SDT comparative statics

The difference between classical SDT and PT-SDT optimal trade-offs entails that the bias of incorrectly assuming normative rationality is larger for agents with a lower *d'*, i.e. a lower ROC curvature and AUC.

PT-SDT shifts the optimal cutoff and the optimal operating point more for agents with a lower ROC curvature and AUC.

The psychology of deception magnifies the effect of behavioral decision making under risk and uncertainty.

Comparative simulation results

Are the individual-level behavioral effects quantitatively consequential at the level of the whole network?

M0	Classical SDT model
M1	PT-SDT model
M2	PT-SDT with psych of deception,

We simulate (ABM, NetLogo) a 3-week spear-phishing attack on an organization with 100 users.

Each user receives 250 emails per working week.

1/250=0.004 of emails are malicious.

During an attack, a user may be fooled at most once. The users *learn* from their mistakes.

Comparative simulation results

Are the individual-level behavioral effects quantitatively consequential at the level of the whole network?

M0 Classical SDT model
$\overline{d'} = 3.0\ (\mathrm{AUC} = 0.983)$

M1 PT-SDT model
$\phi^{-} = 0.88$

M2 PT-SDT with psych of deception
$\underline{d}' = 0.5\ (\mathrm{AUC} = 0.638)$ with probability $\pi = 0.05$

Comparative simulation results

Distribution of security breaches in 10,000 repetitions

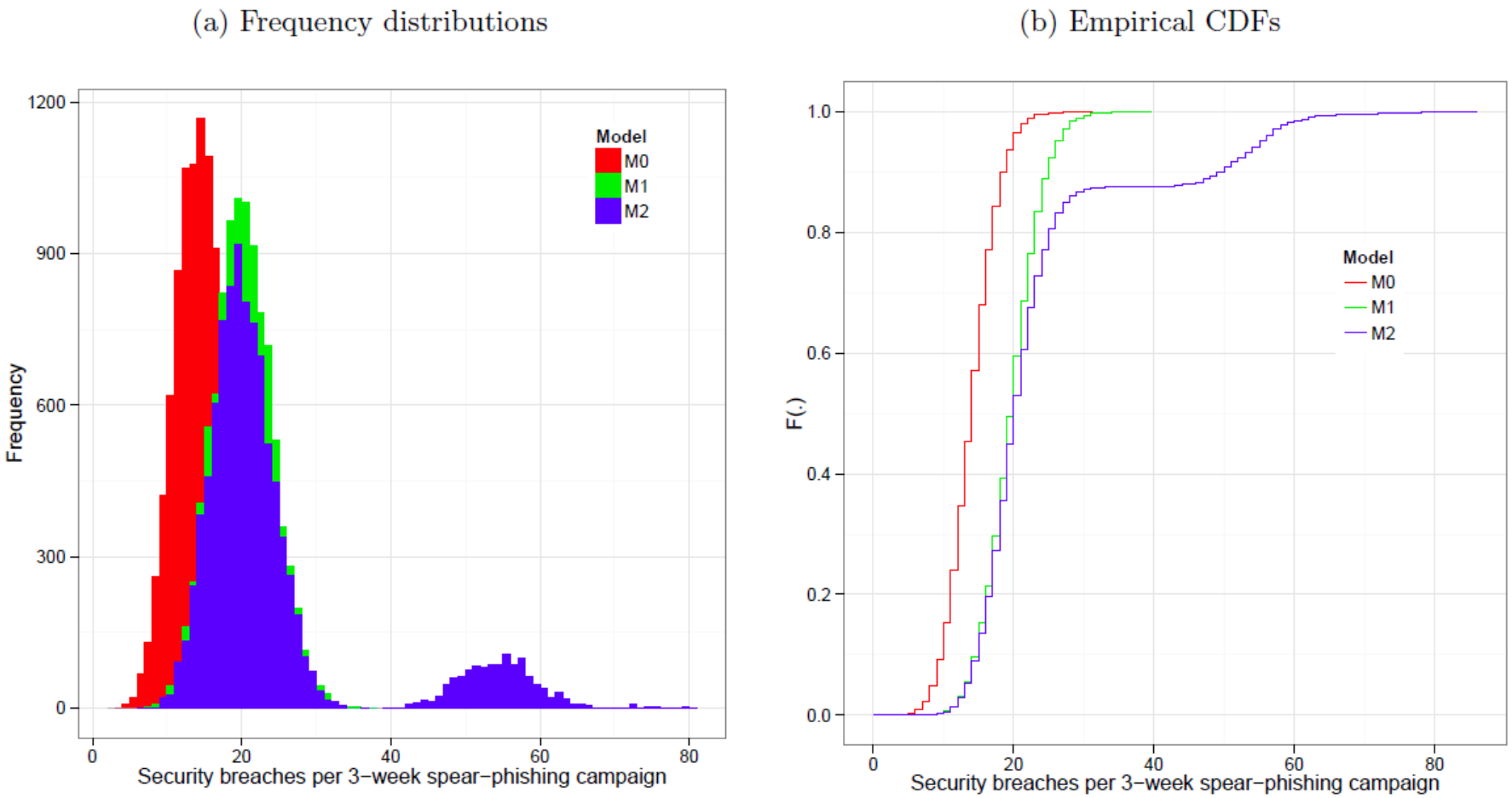
	M0	**M1**	**M2**
Min.	3.0	6.0	6.0
Q_1	12.0	17.0	17.0
Q_2	14.0	20.0	20.0
$\hat{\mu}$	14.0	19.7	23.9
Q_3	16.0	22.0	24.0
Max.	29.0	37.0	80.0

$$Q_j^{\mathrm{M0}} < Q_j^{\mathrm{M1}} \leq Q_j^{\mathrm{M2}} \qquad \forall j \in \{1, 2, 3\}$$

$$\hat{\mu}^{\mathrm{M0}} < \hat{\mu}^{\mathrm{M1}} < \hat{\mu}^{\mathrm{M2}}$$

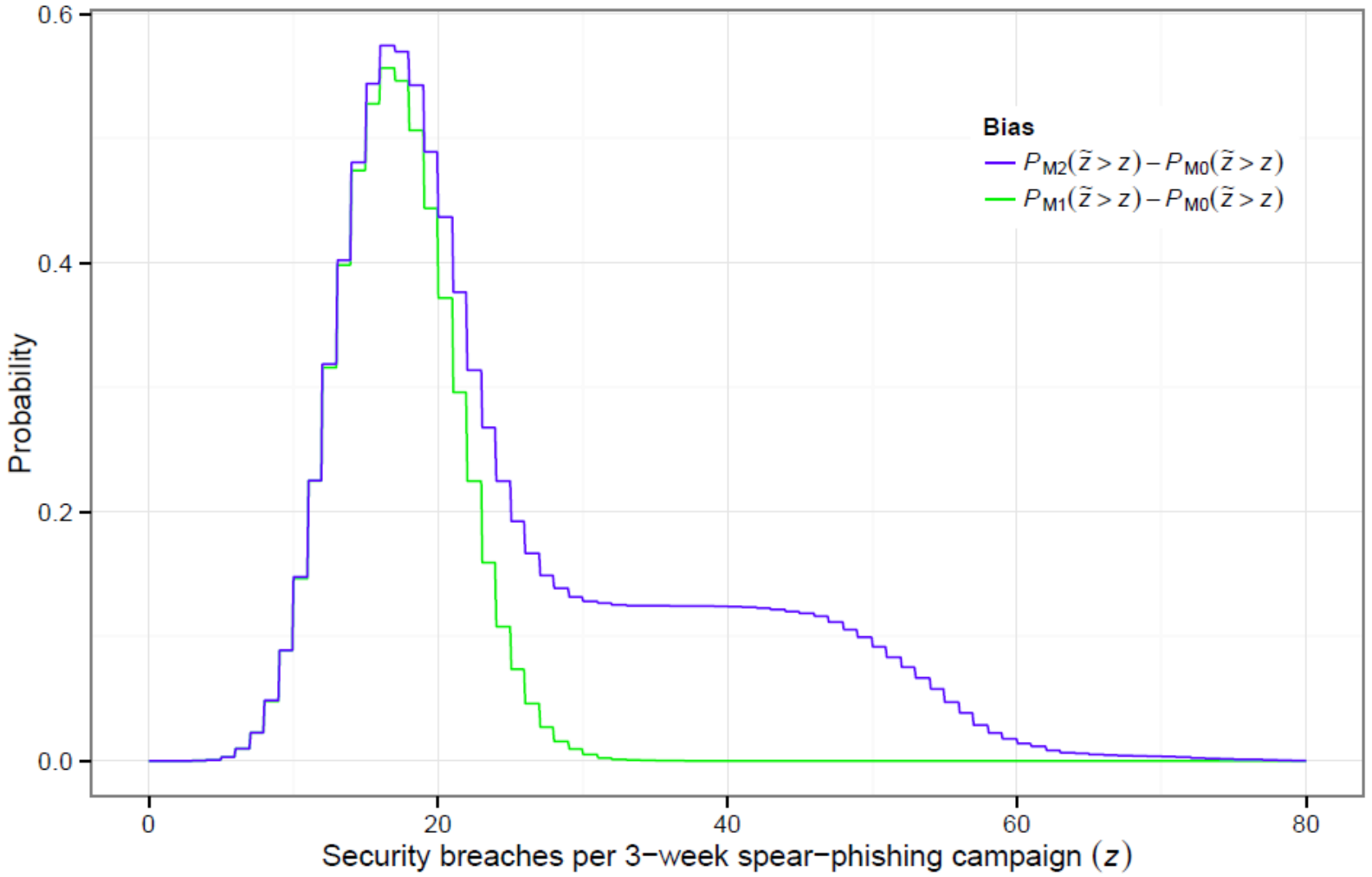
Comparative simulation results

Distribution of security breaches in 10,000 repetitions

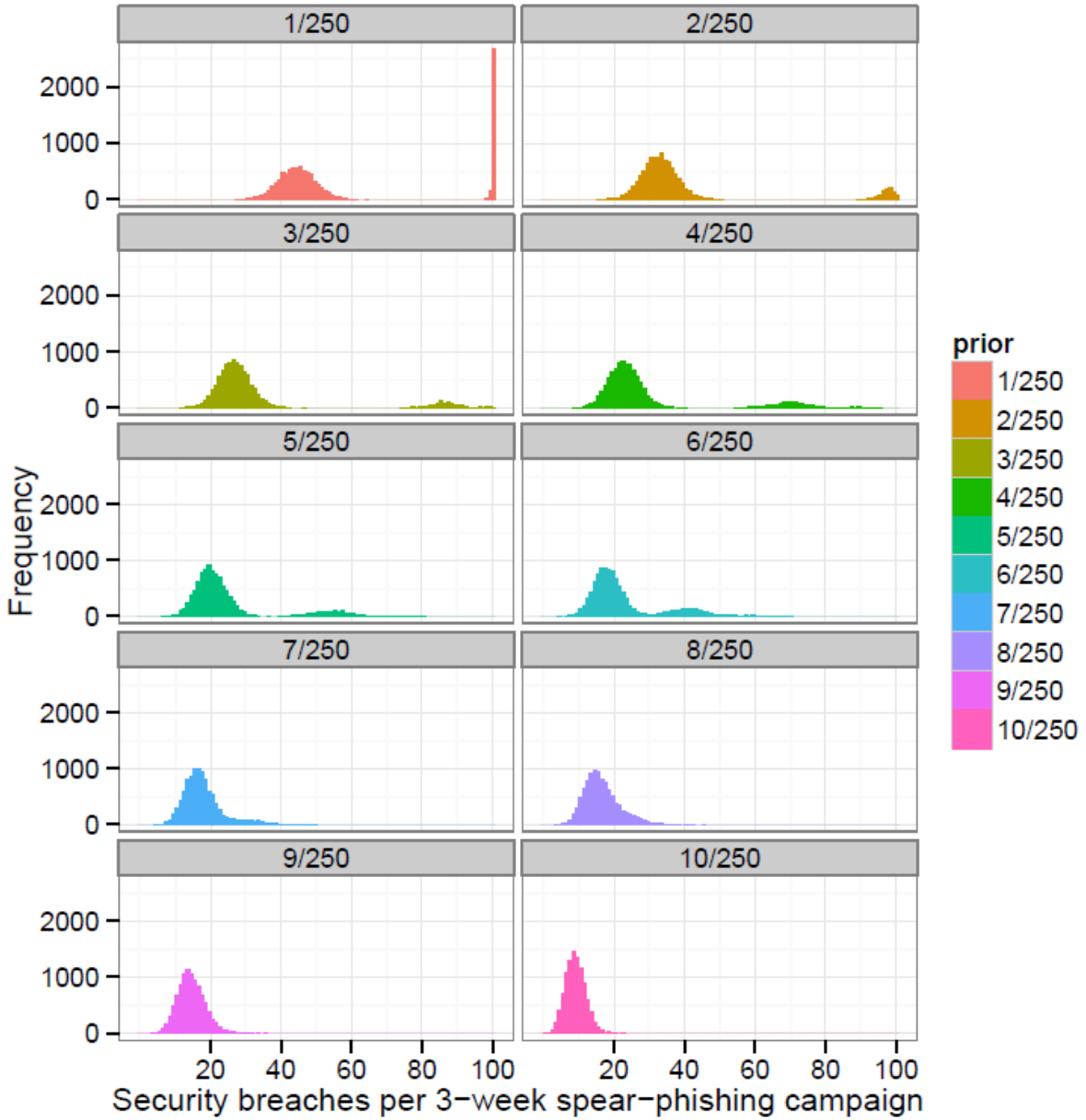


Comparative simulation results

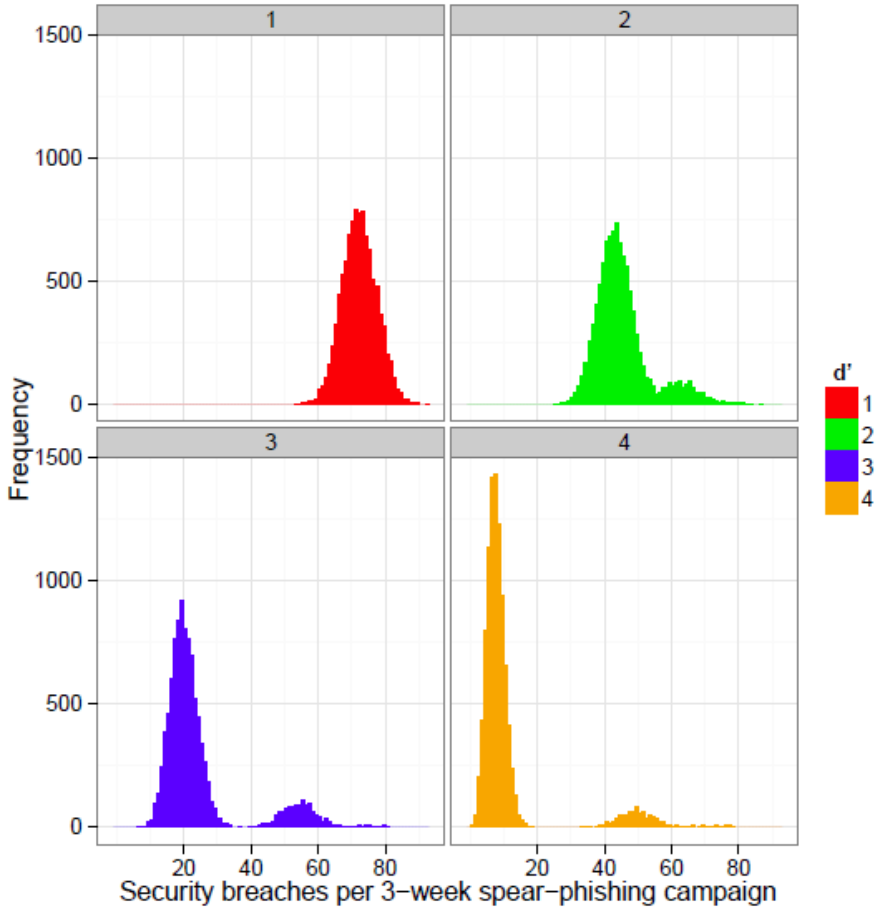
Figure 4: Magnitude of under-estimate (bias) in calculating $P_{M0}(\tilde{z} > z)$ when in fact the descriptively accurate model is M1 (green line) or M2 (blue line).

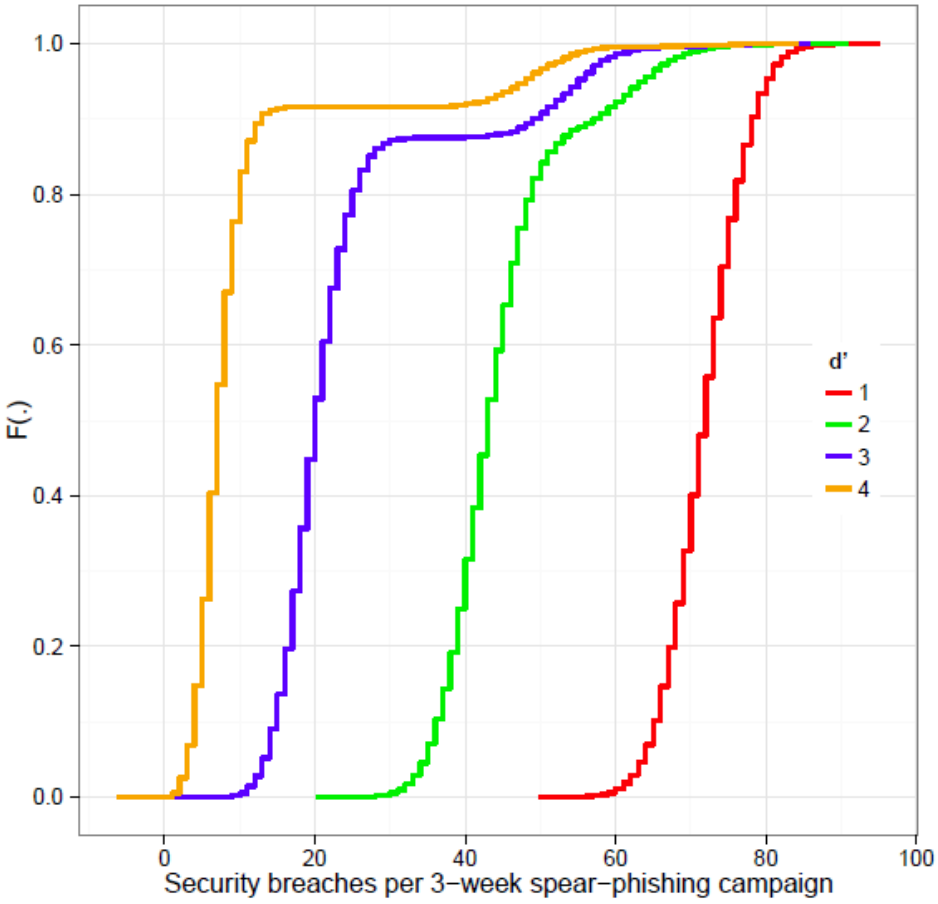


Comparative simulation results: M2, varying prior probability



Comparative simulation results: M2, varying discrimination d’





Implications for education, training and administration

Target: discriminability
prior probability
susceptibility to deception

Note: unintended consequence: spam filters lower *p* !

Idea #1: two-stage classification

Idea #2: short periods of high-prevalence training

Conclusion

Individual-level behavioral effects matter for system-level risk!