Draining the Water Hole: Mitigating Social Engineering Attacks with CyberTWEAK

Appendix

A Deferred Algorithms

A.1 Attacker's Better Response Heuristic

In light of the hardness of finding the adversary's best response, we consider a greedy heuristic. Leveraging Theorem 2, GREEDY (Alg. 3) allocates the adversary's budget to websites in decreasing order of the ratio $r_w = \frac{t_w(1-x_w)/t_w^{all}}{\alpha_w}$, where α_w is a tuning parameter. We replace the MILP for $\mathcal{P}_2(x)$ in CYBERTWEAK with Alg. 3 to find an adversary's better response. If it does not yield a new effort vector, the MILP is called. The column generation process terminates if the MILP again does not find a new effort vector. We refer to this entire procedure as GREEDYTWEAK. Note that GREEDYTWEAK also terminates with the optimal solution. Although GREEDY (Alg. 3) does not provide an approximation guarantee, it performs well in practice. As we show in the experiment section, in practice the accuracy of its solution improves as the size of the problem grows. We also considered a dynamic programming algorithm which is exact and runs in pseudo-polynomial time. However, its practical performance is unsatisfactory.

Algorithm 3: GREEDY

```
Sort the websites in decreasing order of
  $r_w = \frac{t_w(1-x_w)/t_w^{all}}{\alpha_w}$.
foreach website $w$ in the sorted order do
    if remaining attack budget $\geq$ attack cost $\pi_w$ then
        Attack this website $w$ with maximum effort
          allowed
    if running out of budget then break
```

A.2 Baseline Algorithm for $\mathcal{P}_1$

We show the details of one of our baseline algorithms, All Actions, in Alg. 4. Let $\mathcal{A}$ denote the set of actions available to the adversary such that the budget constraint is satisfied. Each action $a^* \in \mathcal{A}$ is a set of websites being compromised. According to Theorem 2, among all the websites w compromised in a^*, the adversary puts "partial" effort $e_w \in (0, t_w^{all}$ on at most one website w^*. Therefore, the action-website pairs (a^*, w^*) fully characterize the adversary's strategies. Alg. 4 works by finding the optimal defender strategy, assuming each action-website pair is the optimal strategy for the adversary.

Algorithm 4: ALL ACTIONS

```
foreach $(a^*, w^*) \in \mathcal{A} \times W$ where $w^* \in a^*$ do
    foreach website $w \in W$ do
        if $w = w^*$ then
            Define
              $z_w = \min\{B_e - \sum_{w \in a^*, w \neq w^*} t_w^{all}, t_{w^*}^{all}\}$
        else if $w \in a^*$ then
            Define $z_w = t_w^{all}$
        else
            Define $z_w = 0$
        Define $k_w = \frac{t_w}{t_w^{all}} z_w$
        foreach $(\hat{a}, \hat{w}) \in \mathcal{A} \times W$ where $\hat{w} \in \hat{a}$ do
            Define $\hat{k}_w$ similarly as above, for each
              $w \in W$.
            Add to $BR(a^*, w^*)$ the following linear
              constraint
              $\sum_{w \in a^*} k_w(1 - x_w) \geq \sum_{w \in \hat{a}} \hat{k}_w(1 - x_w)$
    Solve the following LP
        $\min_{x,v} \quad v$                                        (21)
        s.t. $v \geq \sum_{w \in W} k_w(1 - x_w)$                    (22)
             linear constraints in $BR(a^*, w^*)$                    (23)
             $\sum_{w \in W} c_w t_w x_w \leq B_d$                    (24)
             $x_w \in [0, 1], \quad \forall w \in W$                  (25)
Select the best solution out of all the LPs.
```

B Deferred Proofs

B.1 Proof of Theorem 1

We reduce from the knapsack problem. In the knapsack problem, we have a set W of items each with a weight ω_w and value p_w $\forall w \in N$, and aim to pick items of maximum possible value subject to a capacity B. We now create an instance of the SED problem. Create a website for each item $w \in W$ with organization traffic and total traffic $t_w = t_w^{all} = p_w$ and attack cost ω_w. Assume that $x = \mathbf{0}^T$. Next, set $B_a = B$ and $B_e = \infty$. Notice that the objective function becomes $\sum_{w \in W} e_w$ where $\sum_{w \in W} e_w \leq \infty$ and $e_w \leq p_w y_w$. Hence, $e_w = p_w$ whenever $y_w = 1$. Then, the adversary's best response problem is given by:

$$\max_{y} \quad \sum_{w \in W} p_w y_w \tag{26}$$

$$\text{s.t.} \quad \sum_{w \in W} \omega_w y_w \leq B \tag{27}$$

$$y_w \in \{0, 1\} \qquad \forall w \in W \tag{28}$$

This is exactly the knapsack problem described above. □

B.2 Proof of Theorem 2

For each $w \in W$, let $k_w = t_w(1 - x_w^*)/t_w^{all}$. Suppose there exist some $w_1, w_2 \in W_B$, and w.l.o.g assume $k_{w_1} \geq k_{w_2}$.

Let $\Delta e = \min\{e^*_{w_2}, t^{all}_{w_1} - e^*_{w_1}\}$. Consider the solution $(x^*, y^*, \hat{e})$ where $\hat{e}_{w_1} = e^*_{w_1} + \Delta e$, $\hat{e}_{w_2} = e^*_{w_2} - \Delta e$, and $\hat{e}_w = e^*_w$ for all other websites $w \in W$. This is a feasible solution, and the objective increases by $(k_{w_1} - k_{w_2})\Delta e \geq 0$ compared to (x^*, y^*, e^*). Furthermore, at least one of w_1 and w_2 is removed from W_B. We can apply this argument repeatedly until $|W_B| \leq 1$. □

B.3 Proof of Corollary 1

Since $B_e \leq t^{all}_w \ \forall w \in W$, we know $|W_F| \leq 1$ for any feasible solution. If $|W_F| = 1$, then we have $|W_Z| = n - 1$ and $|W_B| = 0$. If $|W_F| = 0$, by Theorem 2, we have $|W_B| = 1$ and $|W_Z| = n - 1$. In either case, there is only website w^* such that $e_{w^*} > 0$. It follows that $w^* \in \arg\max_{w \in W} \frac{t_w(1-x_w)B_e}{t^{all}_w}$ given a defender strategy x. The optimal defender strategy can be found by solving the following LP.

$$\min_{x,v} \quad v \tag{29}$$

$$\text{s.t.} \quad v \geq \frac{t_w(1-x_w)B_e}{t^{all}_w} \qquad \forall w \in W \tag{30}$$

$$\sum_{w \in W} c_w t_w x_w \leq B_d \tag{31}$$

$$x_w \in [0,1] \qquad \forall w \in W \quad \square \tag{32}$$

B.4 Proof of Theorem 3

Under these assumptions, the problem $\mathcal{P}_1$ becomes

$$\min_{x} \max_{y,e} \quad \sum_{w \in W} t_w(1-x_w)y_w \tag{33}$$

$$\text{s.t.} \quad \sum_{w \in W} y_w \leq B_a \tag{34}$$

$$\sum_{w \in W} c_w t_w x_w \leq B_d \tag{35}$$

$$x_w \in [0,1], y_w \in \{0,1\} \qquad \forall w \in W \tag{36}$$

The constraint $\sum_{w \in W} y_w \leq B_a$ must be satisfied with equality because $t_w(1-x_w) \geq 0$ for all $w \in W$. The defender's problem is to minimize the sum of B_a largest linear functions $t_w - t_w x_w$ among the $n = |W|$ of them, subject to the polyhedral constraints on x_w. This problem can be solved as a single LP (Ogryczak and Tamir 2003) as follows.

$$\min_{d^+,x,z} \quad B_a z + \sum_{w \in W} d^+_w \tag{37}$$

$$\text{s.t.} \quad d^+_w \geq t_w - t_w x_w - z \qquad \forall w \in W \tag{38}$$

$$\sum_{w \in W} c_w t_w x_w \leq B_d \tag{39}$$

$$x_w \in [0,1],\ d^+_w \geq 0 \qquad \forall w \in W \quad \square \tag{40}$$

B.5 Proof of Theorem 4

Let x^* be the optimal solution to $\mathcal{P}_1$. Consider the problem $\hat{\mathcal{P}}_2(x^*)$. At optimal solution, the inequality $e_w \leq t^{all}_w \cdot y_w$ in $\hat{\mathcal{P}}_2(x^*)$ is satisfied with equality, as if $e_w < t^{all}_w \cdot y_w$, then we can decrease y_w without changing the objective value and violating any constraints. Then, we can eliminate the variables e_w and $\hat{\mathcal{P}}_2(x^*)$ becomes a standard two-dimensional fractional knapsack problem $\hat{\mathcal{P}}_4(x^*)$. It is well-known that there exists an optimal solution to $\hat{\mathcal{P}}_4(x^*)$ which has at most 2 fractional values y_{w_1} and y_{w_2} (Kellerer, Pferschy, and Pisinger 2004). We have

$$\begin{aligned} &OPT(\hat{\mathcal{P}}_1) \leq OPT(\hat{\mathcal{P}}_2(x^*)) = OPT(\hat{\mathcal{P}}_4(x^*)) \\ &\leq OPT(\mathcal{P}_2(x^*)) + t_{w_1}(1-x^*_{w_1}) + t_{w_2}(1-x^*_{w_2}) \\ &\leq 3OPT(\mathcal{P}_2(x^*)) = 3OPT(\mathcal{P}_1) \end{aligned}$$

Note that if $B_e = \infty$, $\hat{\mathcal{P}}_1$ is a 2-approximation. □

B.6 Proof of Theorem 5

Since $\hat{x}^*$ and its best response calculated by $\mathcal{P}_2(\hat{x}^*)$ form a feasible solution to $\mathcal{P}_1$, the first inequality holds. For any defender strategy x, $\text{OPT}(\mathcal{P}_2(x)) \leq \text{OPT}(\hat{\mathcal{P}}_2(x))$ as adversary can choose fractional y_w's in $\hat{\mathcal{P}}_2(x)$. For $\hat{x}^*$ specifically, we have $\text{OPT}(\hat{\mathcal{P}}_2(\hat{x}^*)) = \text{OPT}(\hat{\mathcal{P}}_1)$, since $\hat{\mathcal{P}}_1$ is, by strong duality, equivalent to $\mathcal{P}_1$ except that the adversary is allowed to choose fractional y_w's. This establishes the second inequality. The last inequality holds because x^* and its fractional best response calculated by $\hat{\mathcal{P}}_2(x^*)$ form a feasible solution to $\hat{\mathcal{P}}_1$. □

B.7 Proof of Theorem 6

From conditions (1) and (2), we know that for the same amount of effort, the attacker will be better off attacking website u than w, regardless of the defender's strategy.

Suppose $e_w > 0$ and $e_u = 0$ (consequently $y_w = 1, y_u = 0$). Then we could let $e'_w = 0$ and $e'_u = e_w$. This is possible because from condition (4), $e_w \leq t^{all}_w \leq t^{all}_u$ so we have $e'_u \leq t^{all}_u$. Doing this does not increase the attack cost because now $y'_w = 0$ and $y'_u = 1$ and $\pi_w \geq \pi_u$ from condition (3).

Suppose $e_w > 0$ and $e_u > 0$ (consequently $y_w = y_u = 1$). Let $e'_w = e_w - \min\{e_w, t^{all}_u - e_u\}$ and $e'_u = e_u + \min\{e_w, t^{all}_u - e_u\}$. We know that if $e'_w > 0$, then $e'_u = t^{all}_u$. Of course, the attack cost does not increase as well. □

B.8 Proof of Claim 1

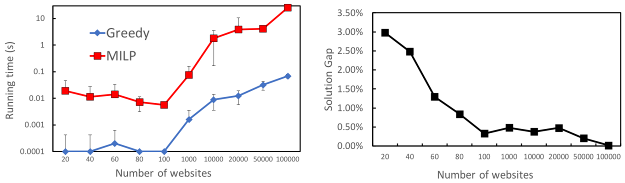
Suppose $(\hat{x}^*, OPT(P_2(\hat{x}^*)))$ is not an optimal solution for the LP $\mathcal{P}^{\text{LP}}_1(\hat{e}^{\mathcal{A}})$ which is equivalent to $\mathcal{P}_1$. Thus, equivalently $\hat{x}^*$ not optimal for $\mathcal{P}_1$. Any of its neighborhood with radius ϵ contains some $(\hat{x}', v')$ as a better solution, meaning $v' < OPT(P_2(\hat{x}^*))$. This solution $(\hat{x}', v')$ satisfies constraint (20), which is strictly stronger than constraint (17). Therefore $(\hat{x}', v')$ is feasible for $\tilde{P}_3(\hat{x}^*)$; this contradicts $OPT(\tilde{P}_3(\hat{x}^*)) \geq OPT(P_2(\hat{x}^*))$. □

B.9 Proof of Claim 2

Claim 1 has covered the case where CYBERTWEAK terminates after the optimality check on Line 3, Alg. 1. In the other case, CYBERTWEAK terminates when no new effort

α_w	$\frac{\text{OPT}-\text{OPT}_{\text{Greedy}}}{\text{OPT}}$
π_w	0.0079
$\pi_w/B_a + 1/B_e$	0.0285
1	0.0082

Table 2: Solution gaps of different greedy heuristics for the adversary best response problem. Results are averaged over 5 runs on different problem sizes $|W| = 100, 200, \ldots, 500$.



(a) Greedy running time (b) Greedy solution gap

vectors are found for the adversary. Suppose x is the optimal solution to the defender's optimization problem (Line 6, Alg. 1), and suppose now $\mathcal{P}_2(x)$ does not find a new effort vector (Line 7, Alg. 1). This implies x would still be feasible for the LP $\mathcal{P}_1^{\text{LP}}(e^{\mathcal{A}})$ even if $e^{\mathcal{A}}$ is replaced by the set of all max effort vectors $\hat{e}^{\mathcal{A}}$. Thus, x is an optimal solution. Indeed, at this point the optimal values of $\mathcal{P}_1^{\text{LP}}(e^{\mathcal{A}})$ and $\mathcal{P}_2(x)$ are equal. □

C Deferred Experiments

We present additional experiments on the adversary's best response problem. In the Greedy algorithm (Alg. 3), the adversary selects websites based on a decreasing order of $r_w = \frac{t_w(1-x_w)/t_w^{all}}{\alpha_w}$. Here, α_w is the tuning parameter. With different choices of α_w, we compare the output value $\text{OPT}_{\text{Greedy}}$ of Greedy with the optimal value OPT obtained by solving the MILP $\mathcal{P}_2(x)$. Table 2 shows the solution gap $\frac{\text{OPT}-\text{OPT}_{\text{Greedy}}}{\text{OPT}}$. We observe that $\alpha_w = \pi_w$ yields the smallest solution gap. We also tested other choices for α_w such as $(\pi_w/B_a)^p + (1/B_e)^q$ for different powers p and q, yet they do not yield better optimization gaps. Hence we fix $r_w = \frac{t_w(1-x_w)/t_w^{all}}{\pi_w}$ in subsequent experiments.

Fig. 4b shows Greedy's solution gap decreases to near zero as the problem size grows. In addition, Greedy typically runs within 1% of the time of the MILP.

D Experiment Parameters

Table 3 shows the distribution from which the parameters are generated in most of our experiments. In Table 4, we detail the parameters used in the experiment in Fig. 2e.

In addition, in the case of small effort budget, B_e is generated uniformly between 1 and $\min_{w\in W} t_{all}^w$.

For large scale instances, we set different websites to have different importance, motivated by the fact that people do not visit all websites with equal frequency. We split W into W_1, W_2 with $|W_1| : |W_2| = 1 : 9$. Websites in W_1 have a large portion of traffic from the organization and those in W_2 have a smaller portion. Thus, W_1 and W_2 follow different distributions (Table 4). The attacker has a uniform cost of attack. In less than 4 of the 20 instances DWE did not reduce the problem size by much. We report in Fig. 2e the majority group where DWE eliminated a significant number of websites. $|W_1|/|W|$ could be a lot smaller in reality, and our algorithms with DWE would run even faster.

Variable	Distribution
t_w^{all}	$U(350, 750)$
t_w	$U(50, 100)$
c_w	$U(1, 4)$
π_w	$U(30, 54)$
B_d	$U(0.11 \sum_{w\in W} c_w t_w, 0.71 \sum_{w\in W} c_w t_w)$
B_a	$U(0.1 \sum_{w\in W} \pi_w, 0.8 \sum_{w\in W} \pi_w)$
B_e	$U(0.2 \sum_{w\in W} t_w^{all}, 0.8 \sum_{w\in W} t_w^{all})$

Table 3: Parameter distribution

For $w \in W_1$		For $w \in W_2$	
Variable	Distribution	Variable	Distribution
t_w^{all}	$U(60, 110)$	t_w^{all}	$U(20, 70)$
t_w	$U(45, 55)$	t_w	$U(3, 8)$
c_w	$U(2, 6)$	c_w	$U(1, 3)$
π_w	3	π_w	3
B_d	$U(0, 10 \sum_{w\in W} c_w t_w/|W|)$		
B_a	$U(0.1 \sum_{w\in W} \pi_w, 0.8 \sum_{w\in W} \pi_w)$		
B_e	$U(0, 3 \sum_{w\in W} t_w^{all}/|W|)$		

Table 4: Parameter distributions for the experiment on large instances.

E Discussion

Assumptions and generality We assumed that the attack will succeed if and only if the network packet is unaltered. If the attacker can obtain the true system information with probability p_w even if the packet is altered, we may modify the objective in Eq. (1) to $\sum_w t_w(1 - x_w(1 - p_w))e_w/t_w^{all}$. If the organization has other countermeasures (e.g. Bromium browser VMs), the attack may fail with probability q_w even if the packet is unaltered, the objective then becomes $\sum_w t_w(1 - x_w)(1 - q_w)e_w/t_w^{all}$. Thus, our algorithm can account for different levels of adversary and defender sophistication.

We do not attempt to claim that altering the network packets is a panacea to all watering hole attacks. Cyber attackers have many tools to circumvent existing deception techniques. Nonetheless, the proposed deception technique increases their uncertainty about the true nature of the environment, which leads to more cost on them, e.g. technical complexity and increased exposure. This uncertainty ties into our consideration of the attacker's scanning effort e_w and budget B_e, as the attacker cannot easily obtain or trust the basic information in the network packets.

Limitations The generality notwithstanding, We acknowledge a few limitations of our work and potential problems in large-scale deployment. First, if an organization is the sole user of our method and if the attacker has (possibly imperfect) clue about the source of traffic from the start, randomizing network packet information might serve as an unintended signal to the attacker, reducing the effort needed

e_w to identify traffic from the targeted organization. Second, by manipulating the web traffic, the organization is effectively monitoring its employees' internet activities. Although in many jurisdictions this is allowed when doing properly, the potential ethical issues must be carefully addressed.

References

[Ogryczak and Tamir 2003] Ogryczak, W. and Tamir, A., 2003. Minimizing the sum of the k largest functions in linear time. Information Processing Letters, 85(3), pp.117-122.

[Kellerer, Pferschy, and Pisinger 2004] Kellerer, H., Pferschy, U. and Pisinger, D., 2004. Knapsack problems. Springer, Berlin, Heidelberg.