Online Appendix for "When Big Data Enables Behavioral Manipulation"

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, Asuman Ozdaglar

For completeness, this Appendix includes the omitted derivations and proofs from the text. Belief evolutions are standard, and the details are provided next.

Proof of the belief evolution in (2) and the belief trajectory

Before deriving belief evolution, we highlight that throughout, we follow the standard practice of dropping terms of order $o(dt)$. If $x_{i,t}b_{i,t} = 0$, the user belief about θ_i does not change. Next, we consider $x_{i,t}b_{i,t} = 1$ and, for notational convenience, let us suppress subscript i. Then, by using Bayes' rule, the probability of $\theta_i = 1$ is

$$\begin{aligned}
\mu_{t+dt} &= \frac{\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 1]\mathbb{P}[\theta = 1]}{\mathbb{P}[\mathrm{NBN}_{t+dt}]} \\
&= \mu_t \frac{\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \mathrm{NBN}_t, \theta = 1]}{\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 1, \mathrm{NBN}_t]\mathbb{P}[\theta = 1 \mid \mathrm{NBN}_t] + \mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t]\mathbb{P}[\theta = 0 \mid \mathrm{NBN}_t]} \\
&\overset{(a)}{=} \frac{\mu_t}{\mu_t + (1-\mu_t)\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t]} \\
&= \frac{\mu_t}{1 - (1-\mu_t)(1-\lambda_t)\gamma dt} \overset{(b)}{=} \mu_t \left(1 + (1-\mu_t)(1-\lambda_t)\gamma dt\right)
\end{aligned} \tag{A1}$$

where (a) follows from

$$\begin{aligned}
\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t] =& \mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t, \alpha_t = 1]\mathbb{P}[\alpha_t = 1 \mid \theta = 0, \mathrm{NBN}_t] \\
&+ \mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t, \alpha_t = 0]\mathbb{P}[\alpha_t = 0 \mid \theta = 0, \mathrm{NBN}_t]
\end{aligned}$$

and (b) follows by using Taylor expansion of $1/(1-x)$ around 0 and dropping the terms of the order $(dt)^2$. From (A1), we then have

$$d\mu_t = \mu_{t+dt} - \mu_t = \mu_t(1-\mu_t)(1-\lambda_t)\gamma dt.$$

Again, by using Bayes' rule,

$$\begin{aligned}
\lambda_{t+dt} &= \mathbb{P}[\alpha_{t+dt} = 1 \mid \theta = 0, \mathrm{NBN}_{t+dt}] \\
&= \frac{\lambda_t \mathbb{P}[\alpha_{t+dt} = 1 \mid \alpha_t = 1, \theta = 0]\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t, \alpha_{t+dt} = 1]}{\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t]} \\
&\overset{(a)}{=} \frac{\lambda_t(1-\rho dt)}{\lambda_t + (1-\lambda_t)(1-\gamma dt)} \overset{(b)}{=} \lambda_t - \lambda_t \rho dt + \lambda_t(1-\lambda_t)\gamma dt
\end{aligned} \tag{A2}$$

where (a) follows from

$$\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t] = \mathbb{P}[\mathrm{NBN}_{t+dt} \mid \alpha_t = 1, \theta = 0, \mathrm{NBN}_t]\mathbb{P}[\alpha_t = 1 \mid \theta = 0, \mathrm{NBN}_t]$$

$$+ \mathbb{P}[\text{NBN}_{t+dt} \mid \alpha_t = 0, \theta = 0, \text{NBN}_t]\mathbb{P}[\alpha_t = 0 \mid \theta =, \text{NBN}_t]$$

and (b) follows by dropping the terms of the order $(dt)^2$. From (A2), we now have

$$d\lambda_t = \lambda_t((1-\lambda_t)\gamma - \rho)dt.$$

Finally, given $x_{i,t}b_{i,t} = 1$, we have

$$\mathbb{P}[I_{i,t+dt} = 0 \mid I_{i,t} = 1] = (1-\mu_{i,t})\,(1-\lambda_{i,t})\,\gamma dt.$$

This completes the proof of the evolution.

We next state the belief trajectory and some monotonicity properties in the pre-AI environment.

For any product $i \in \mathcal{N}$ that has been offered for $[0,t)$ with $\text{NBN}_{i,t}$ (no bad news), we have

$$\lambda_{i,t} = \frac{(\gamma-\rho)\lambda}{\lambda\gamma + e^{(\rho-\gamma)t}((1-\lambda)\gamma - \rho))},$$

and

$$\mu_{i,t} = \frac{\mu_{i,0}(\gamma-\rho)}{\mu_{i,0}(\gamma-\rho) + (1-\mu_{i,0})\,(e^{-\rho t}\lambda\gamma + e^{-\gamma t}\,((1-\lambda)\gamma - \rho))}.$$

Moreover, $\mu_{i,t}$ is increasing in $\mu_{i,0}$ and t and converges to 1 as $t \to \infty$.

The above trajectories directly follow from solving the differential equations and evaluating its first-order derivative. ■

Proof of the belief evolution in (3) and the belief trajectory

Similar to the proof of belief evolution in (2), if $x_{i,t}b_{i,t} = 0$, the user belief about θ_i does not change, and when $x_{i,t}b_{i,t} = 1$, we let

$$\mu_{i,t}^{(P)} = \mathbb{P}\left[\theta = 1 \mid \alpha_{i,0} = 0, \text{NBN}_t\right]$$

be the probability of a product being high quality if the initial glossiness state is zero, and no bad news has arrived by time t, and

$$\lambda_{i,t}^{(P)} = \mathbb{P}\left[\alpha_{i,t} = 1 \mid \alpha_{i,0} = 1, \text{NBN}_t\right]$$

be the probability of the glossiness state being 1 if the initial glossiness state is 1, and no bad news has arrived by time t. Again, to make the notation easier, in this proof, we drop the subscript i. Notice that conditioning on $\alpha_0 = 1$, the platform knows the product is of low quality and therefore $\mu_t^{(P)} = 0$ for all t. Also, conditioning on $\alpha_0 = 0$, the glossiness state remains at zero, and therefore $\lambda_t^{(P)} = 0$. We next evaluate $\mu_t^{(P)}$ conditioning on $\alpha_0 = 0$ and $\lambda_t^{(P)}$ conditioning on $\alpha_0 = 1$.

By using Bayes' rule, for the probability of $\theta_i = 1$ when $\alpha_0 = 0$ we have

$$\mu_{t+dt}^{(P)} = \frac{\mathbb{P}[\text{NBN}_{t+dt} \mid \theta = 1, \alpha_0 = 0]\mathbb{P}[\theta = 1 \mid \alpha_0 = 0]}{\mathbb{P}[\text{NBN}_{t+dt} \mid \alpha_0 = 0]}$$

$$\overset{(a)}{=} \frac{\mu_t^{(P)}}{\mu_t^{(P)} + (1-\mu_t^{(P)})(1-\gamma dt)} \overset{(b)}{=} \mu_t^{(P)}\left(1 + (1-\mu_t^{(P)})\gamma dt\right) \tag{A3}$$

where (a) follows from

$$\begin{aligned}\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \mathrm{NBN}_t, \alpha_0 = 0] =& \mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 1, \mathrm{NBN}_t, \alpha_0 = 0]\mathbb{P}[\theta = 1 \mid \mathrm{NBN}_t, \alpha_0 = 0] \\ &+ \mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t, \alpha_0 = 0]\mathbb{P}[\theta = 0 \mid \mathrm{NBN}_t, \alpha_0 = 0] \\ &= \mu_t^{(P)} + \left(1 - \mu_t^{(P)}\right)(1 - \gamma dt)\end{aligned}$$

and (b) follows by dropping the terms of the order $(dt)^2$. By using (A3),

$$d\mu_t^{(P)} = \mu_{t+dt}^{(P)} - \mu_t^{(P)} = \mu_t^{(P)}(1-\mu_t^{(P)})\gamma dt.$$

By using Bayes' rule, when $\alpha_0 = 1$ we obtain

$$\begin{aligned}\lambda_{t+dt}^{(P)} &= \mathbb{P}[\alpha_{t+dt} = 1 \mid \theta = 0,\ \mathrm{NBN}_{t+dt}, \alpha_0 = 1] \\ &\overset{(a)}{=} \frac{\mathbb{P}[\alpha_t = 1 \mid \theta = 0, \alpha_0 = 1]\mathbb{P}[\alpha_{t+dt} = 1 \mid \alpha_t = 1, \theta = 0, \alpha_0 = 1]\mathbb{P}[\mathrm{NBN}_t \mid \theta = 0, \alpha_{t+dt} = 1, \alpha_0 = 1]}{\mathbb{P}[\mathrm{NBN}_t \mid \theta = 0, \alpha_0 = 1]\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \mathrm{NBN}_t, \theta = 0, \alpha_0 = 1]} \\ &\overset{(b)}{=} \frac{\lambda_t^{(P)}(1-\rho dt)}{\lambda_t^{(P)} + (1-\lambda_t^{(P)})(1-\gamma dt)} \\ &\overset{(c)}{=} \lambda_t^{(P)} - \lambda_t^{(P)}\rho dt + \lambda_t^{(P)}(1-\lambda_t)\gamma dt,\end{aligned} \tag{A4}$$

where (a) follows from $\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t, \alpha_{t+dt} = 1, \alpha_0 = 1] = 1$, (b) follows from

$$\begin{aligned}&\mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t, \alpha_0 = 1] \\ &= \mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t, \alpha_0 = 1, \alpha_t = 1]\mathbb{P}[\alpha_t = 1 \mid \theta = 0, \mathrm{NBN}_t, \alpha_0 = 1] \\ &+ \mathbb{P}[\mathrm{NBN}_{t+dt} \mid \theta = 0, \mathrm{NBN}_t, \alpha_0 = 1, \alpha_t = 0]\mathbb{P}[\alpha_t = 0 \mid \theta = 0, \mathrm{NBN}_t, \alpha_0 = 1],\end{aligned}$$

and (c) follows by dropping the terms of the order $(dt)^2$. By using (A4), we obtain

$$d\lambda_t^{(P)} = \lambda_t^{(P)}((1-\lambda_t^{(P)})\gamma - \rho)dt.$$

Finally, we have

$$\begin{aligned}\mathbb{P}[\text{ bad news at } t \mid \alpha_0 = 0, \mathrm{NBN}_t] =& \mathbb{P}[\text{ bad news at } t \mid \alpha_0 = 0, \theta = 0, \mathrm{NBN}_t]\mathbb{P}[\theta = 0 \mid \alpha_0 = 0, \mathrm{NBN}_t] \\ &+ \mathbb{P}[\text{ bad news at } t \mid \alpha_0 = 0, \theta = 1, \mathrm{NBN}_t]\mathbb{P}[\theta = 1 \mid \alpha_0 = 0, \mathrm{NBN}_t] \\ =& \gamma dt\left(1 - \mu_t^{(P)}\right)\end{aligned}$$

and

$$\mathbb{P}[\text{ bad news at } t \mid \alpha_0 = 1, \mathrm{NBN}_t] = \mathbb{P}[\text{ bad news at } t \mid \alpha_0 = 1, \mathrm{NBN}_t, \theta = 0] = \gamma dt \left(1 - \lambda_t^{(P)}\right).$$

The initializations follow by using Bayes' rule, completing the proof of the belief evolution.

We next state the belief trajectory and some monotonicity properties in the post-AI environment.

For any product $i \in \mathcal{N}$ that has been offered for $[0, t)$ with $\mathrm{NBN}_{i,t}$ (no bad news), the dynamics of $\mu_{i,t}$ and $\lambda_{i,t}$ are the same as the pre-AI environment, and additionally:

- If $\alpha_{i,0} = 0$, then

$$\mu_{i,t}^{(P)} = \frac{\mu_{i,0}}{\mu_{i,0} + e^{-\gamma t}(1-\lambda)(1-\mu_{i,0})}, \;\; \lambda_{i,t}^{(P)} = 0, \;\; \mathbb{P}[\mathrm{NBN}_{i,t} \mid \alpha_{i,0} = 0] = \frac{\mu_{i,0} + e^{-\gamma t}(1-\lambda)(1-\mu_{i,0})}{1 - \lambda(1-\mu_{i,0})},$$

  and

$$\mathbb{P}[I_{i,t+dt} = 0, I_{i,t} = 1 \mid \alpha_{i,0} = 0] = \frac{e^{-\gamma t}(1-\lambda)(1-\mu_{i,0})}{1 - \lambda(1-\mu_{i,0})} \gamma dt.$$

- If $\alpha_{i,0} = 1$, then

$$\mu_{i,t}^{(P)} = 0, \;\; \lambda_{i,t}^{(P)} = \frac{\gamma - \rho}{\gamma - e^{(\rho-\gamma)t}\rho}, \;\; \mathbb{P}[\mathrm{NBN}_{i,t} \mid \alpha_{i,0} = 1] = \frac{e^{-\rho t}\gamma - e^{-\gamma t}\rho}{\gamma - \rho},$$

  and

$$\mathbb{P}[I_{i,t+dt} = 0, I_{i,t} = 1 \mid \alpha_{i,0} = 1] = \frac{\left(e^{-\rho t} - e^{-\gamma t}\right)\rho}{\gamma - \rho} \gamma dt.$$

The above statements directly follow from solving the differential equations and evaluating its first-order derivative. ■

An additional lemma

We next state and prove a lemma we used in the proof of Theorem 4.

Lemma A1. *The expected user utility increases for large enough ρ after performing a helpful swap.*

Proof: It suffices to prove that in the limit of $\rho \to \infty$, the expected user utility after performing a helpful swap increases by some quantity which is strictly positive (and does not depend on ρ). This establishes the existence of large enough ρ for which the statement holds. For any $j \in \mathcal{N}$, we let τ_j be the stochastic time at which bad news occurs for product j given $\theta_j = 0$. Notice that in the limit of $\rho \to \infty$, for both a product with $\alpha_{j,0} = 1$ and a low-quality product with $\alpha_{j,0} = 0$, the platform knows that bad news arrive with rate γ. We let

$$A_j \triangleq \mathbb{E}\left[-\int_0^{\tau_j} r e^{-rt} \mu_{j,t} dt\right]$$

for $j = n, n-1, i$. Using this notation, and that for $j = n, n-1, \ldots, i+1$, $\delta \triangleq \mathbb{E}_{t\sim\tau_j}[e^{-rt}] < 1$, the expected utility before the swap is

$$\begin{aligned} U_1 \triangleq & \sum_{j=n}^{i+1} A_j \delta^{n-j} + \delta^{n-i}\left(\mu_{i,0}^{(P)} \int_0^\infty r e^{-rt}(1-\mu_{i,t})dt + (1-\mu_{i,0}^{(P)})A_i\right) \\ & + \delta^{n-i+1}(1-\mu_{i,0}^{(P)})\,(\text{Expected utility from products } i-1,\ldots,0)\,, \end{aligned} \tag{A5}$$

where $\mu_{i,0}^{(P)} = \frac{\mu_{i,0}}{\mu_{i,0}+(1-\lambda)(1-\mu_{i,0})}$ when $\alpha_{i,0} = 0$. The expected utility after the swap is

$$\begin{aligned} U_2 \triangleq & \mu_{i,0}^{(P)} \int_0^\infty r e^{-rt}(1-\mu_{i,t})dt + (1-\mu_{i,0}^{(P)})A_i + \delta(1-\mu_{i,0}^{(P)}) \sum_{j=n}^{i+1} A_j \delta^{n-j} \\ & + \delta^{n-i+1}(1-\mu_{i,0}^{(P)})\,(\text{Expected utility from products } i-1,\ldots,0) \\ \overset{(a)}{>} & \delta^{n-i}\mu_{i,0}^{(P)} \int_0^\infty r e^{-rt}(1-\mu_{i,t})dt + (1-\mu_{i,0}^{(P)})A_i + \delta(1-\mu_{i,0}^{(P)}) \sum_{j=n}^{i+1} A_j \delta^{n-j} \\ & + \delta^{n-i+1}(1-\mu_{i,0}^{(P)})\,(\text{Expected utility from products } i-1,\ldots,0) \\ = & U_1 + A_i\left((1-\mu_{i,0}^{(P)}) - \delta^{n-i}(1-\mu_{i,0}^{(P)})\right) + \sum_{j=n}^{i+1} A_j\left(\delta^{n+1-j}(1-\mu_{i,0}^{(P)}) - \delta^{n-j}\right) \\ \overset{(b)}{\geq} & U_1 + A_i\left((1-\mu_{i,0}^{(P)}) - \delta^{n-i}(1-\mu_{i,0}^{(P)})\right) + \sum_{j=n}^{i+1} A_i\left(\delta^{n+1-j}(1-\mu_{i,0}^{(P)}) - \delta^{n-j}\right) \\ = & U_1 - \mu_{i,0}^{(P)} A_i \left(\sum_{j=0}^{n-i+1} \delta^j\right) \overset{(c)}{>} U_1 \end{aligned} \tag{A6}$$

where (a) follow from $\mu_{i,0}^{(P)} \int_0^\infty r e^{-rt}(1-\mu_{i,t})dt > 0$ and $1 > \delta^{n-i}$, (b) follows from $A_j < A_i$ and $\delta^{n+1-j}(1-\mu_{i,0}^{(P)}) - \delta^{n-j} < 0$ for $j = n, n-1, i+1$, and (c) follows from $A_i < 0$, completing the proof of lemma. ■