MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Information Disclosure in Dynamic Innovation Contests

Jussi Keppo
NUS Business School and Institute of Operations Research and Analytics
National University of Singapore, Singapore
keppo@nus.edu.sg

Linsheng Zhuang
Institute of Operations Research and Analytics
National University of Singapore, Singapore
linsheng.z@u.nus.edu

...

Key words: Digital Economy, Data Protection Regulation, Innovation Contest

1. Introduction

- Kaggle[1] ...
- Meta-kaggle dataset Risdal and Bozsolik (2022).

1.1. Literature Review

This paper focuses on the two players innovation contest with a continuous time where the players' relative position is public information throughout the game. This is closely related to tug-or-war contest, which, to our knowledge, was first formally given by Harris and Vickers (1987) as a one-dimensional simplification of the multi-stage R&D race. The output processes are model by Brownian motions drifted with effort inputs, which is followed by Budd et al. (1993) who model the state of a dynamic competition of two innovative duopoly firms by a Brownian motion drifted by the effort gap, and solve the equilibrium approximately. Furthermore, Moscarini and Smith (2007) model the tug-of-war state as the gap of the two outputs directly, and draw an analytical equilibrium of the pure strategies.

[1] https://www.kaggle.com

...

Information disclosure in contest - Bimpikis et al. (2019).

...

Closest paper - Ryvkin (2022).

...

2. The Model

...

2.1. Settings

We assume two players, i and j, compete for a prize $\theta > 0$ in a contest. Winner gets the prize and loser gets nothing. The contest starts at time zero. At every time $t \geq 0$, the representative player i chooses an effort level $q_{i,t}$ and burdens a quadratic cost $C_i(q_{i,t}) = c_i q_{i,t}^2/2$, with a lower c_i corresponding to higher ability. The dynamic of player i's output follows

$$dx_{i,t} = q_{i,t}dt + \sigma_i dW_{i,t} \tag{1}$$

Here $W_{i,t}$ is a standard Brownian motion and $\sigma_i > 0$ measures her production risk. Moreover, denoted by y_t the *output gap* of player i and j at time t.

2.1.1. Kalman Filter. We assume that the contest is equipped with a submission system that allows participants to upload their algorithms at any time and receive immediate feedback. For simplicity, we further assume that agents submit their intermediate results whenever they make progress. This setup enables the contest organizers to monitor all players' progress $x_{i,t}$ and $x_{j,t}$ in real time (whenever there is a submission). The true outputs, evaluated by the system, is only known by the game designer but not the two players.

Let' suppose the submission events of the representative player i occur at times $(t_1^i, t_2^i, ...)$ following an inhomogeneous Poisson process driven by the intensity function $\tau_i(t)$.

After each submission, the contest designer emits a *public* signal of the real output x_{i,t_k}. The signal is ambiguous and the game holder controls the ambiguity. The dynamic of signal is

$$\hat{x}_{i,k} = x_{i,t_k} + \frac{v_{i,k}}{\sqrt{\lambda}}, \quad k = 1, 2, ... \tag{2}$$

where $v_{i,k}$ follows standard normal distribution and is independent with $(W_{i,t})$ and $(W_{j,t})$, and the parameter λ is set by the game holder to control the precision of signals. The larger the λ, the more accurate the signal would be.

The information set of both players at time $t \geq 0$ is $I_t := \{\hat{x}_{i,k}, \hat{x}_{j,k} : 0 \leq t_k \leq t\}$. Both players estimate the unknown outputs $x_{i,t}$ and $x_{j,t}$ purely based on the information set I_t and hidden

actions $q_{i,t}$ and $q_{j,t}$. Let $\tilde{x}_{i,t} \equiv E(x_{i,t}|I_t)$ be the estimated output gap and $S_{i,t} \equiv E[(\tilde{x}_{i,t} - x_{i,t})^2|I_t]$ be the estimation variance. The conditional distribution $y_t|I_t \sim \mathcal{N}(\tilde{y}_t, S_t|I_t)$ is fully captured by the mean $\tilde{y}_t$ and variance S_t.

The evolution of $\tilde{x}_{i,t}$ and $S_{i,t}$ is characterized by a continuous-discrete Kalman Filter (CD-KF, Barrau and Bonnabel 2017, Frogerais et al. 2012), with the measurement of each step $k = 1, 2, ...$ consisting of two phases: in (1) prediction phase during time interval (t_{k-1}, t_k), equations are derived from those of Kalman-Bucy filter (Bensoussan 1992) without considering the Kalman gain:

$$d\tilde{x}_{i,t} = q_{i,t}dt \tag{3}$$

$$dS_{i,t} = \sigma^2 dt \tag{4}$$

and in (2) updation phase $t = t_k$, equations are

$$\tilde{x}^+_{i,k} = \tilde{x}^-_{i,k} + \frac{\lambda S^-_{i,t_k}}{\lambda S^-_{i,t_k} + 1}\left(\hat{x}_{i,k} - \tilde{x}^-_{i,t_k}\right) \tag{5}$$

$$S^+_{i,t_k} = \frac{S^-_{i,t_k}}{\lambda S^-_{i,t_k} + 1} \tag{6}$$

If $\lambda = 0$, we have $S_t = S_0 + \sigma^2 t$, i.e., the estimation variance is increasing in time linearly. If $\lambda > 0$, the estimation error decreases after each update of the public leaderboard, i.e., $S^+_{i,t_k} < S^-_{i,t_k}$.

2.1.2. Dynamic Contest. Following Ryvkin (2022), let's consider a dynamic contest of two players with a fixed deadline. Suppose the contest is terminated when time $t = T > 0$. Since the steady state estimation variance $\bar{S}$ is fixed as displayed above, the state of the game is fully characterized by a tuple $(\tilde{y}_t, t)$. At any time $0 \le t < T$, player i optimizes her effort level $q_{i,\tau}$ in the remaining contest period $\tau \in [t, T)$ according to the following optimization problem,

$$V^i(\tilde{y}_t, t; q_{j,t}, \Theta_i) = \max_{\{q_{i,\tau}\}_{\tau=t}^T} \mathbb{E}\left(\theta \cdot 1_{\tilde{y}_T > 0} - \int_t^T C_i(q_{i,\tau})d\tau \middle| I_t\right) \tag{7}$$

where $\Theta_i \equiv \{\theta, \lambda, \sigma, c_i\}$, subject to constraints (3), (4) and $q_{i,\tau} \ge 0$ for all $\tau \in [t, T)$. The optimization problem for player j is just symmetric to that of player i as $V^j(\tilde{y}_t, t) = V^i(-\tilde{y}_t, t)$. The corresponding Hamilton-Jacobi-Bellman (HJB) equation for player i is

$$0 = \max_{q_{i,t} \ge 0}\left[-\frac{c_i q_i^2}{2} + V^i_y \cdot (q_{i,t} - q_{j,t}) + V^i_t + \frac{V^i_{yy}}{2}\lambda\bar{S}^2\right]$$

By definition, we have $\lambda\bar{S}^2 = \sigma^2$. Under the assumption of inner solution, we plug into the first order conditions $q_{i,t} = V^i_y/c_i$ and $q_{j,t} = -V^j_y/c_j$, we have the system of equations

$$\frac{1}{2c_i}(V^i_y)^2 + \frac{1}{c_j}V^i_y V^j_y + V^i_t + V^i_{yy}\frac{\sigma^2}{2} = 0$$

$$\frac{1}{2c_j}(V^j_y)^2 + \frac{1}{c_i}V^j_y V^i_y + V^j_t + V^j_{yy}\frac{\sigma^2}{2} = 0$$

subject to boundary conditions $V^i(-\infty,t)=0,\ V^i(+\infty,t)=\theta,\ V^j(-\infty,t)=\theta,\ V^j(+\infty,t)=0,$ $V^i(\tilde{y}_T,T)=\theta\cdot 1_{\tilde{y}_T>0}$ and $V^j(\tilde{y}_T,T)=\theta\cdot 1_{\tilde{y}_T<0}$.

The Nash equilibrium is summarized in the following lemma. We include a simplified version of the proof in the appendix:

LEMMA 1 (**Ryvkin 2022**). *In the Markov perfect equilibrium, the players' efforts in state* $(y,t)\in\mathbb{R}\times[0,T)$ *are given by*

$$m_{ij}(y,t)=\frac{e^{-z^2/2}}{\sqrt{2\pi\sigma^2(T-t)}}\cdot\frac{\sigma^2}{2}\left[\gamma(\rho_i)+\gamma(\rho_j)\right]\left[1-\rho(z)^2\right]\left[1\pm\rho(z)\right] \tag{8}$$

where $z=y/(\sigma\sqrt{T-t})$, $\rho(z)=\gamma^{-1}\left(\Phi(z)\left[\gamma(\rho_i)+\gamma(\rho_j)\right]-\gamma(\rho_j)\right)$ *and*

$$\gamma(u)=\frac{u}{1-u^2}+\frac{1}{2}\ln\frac{1+u}{1-u},\quad u\in(-1,1)$$

$$\rho_i=\frac{e^{w_i}+e^{-w_j}-2}{e^{w_i}-e^{-w_j}},\quad \rho_j=\frac{e^{w_j}+e^{-w_i}-2}{e^{w_j}-e^{-w_i}},\quad w_{i(j)}=\frac{\theta}{\sigma^2 c_{i(j)}}.$$

The variables $w_{i(j)}$ represent the abilities of two players, while $\rho_{i(j)}$ normalizes $w_{i(j)}$ into the interval $(-1,1)$. It is not hard to see that $\gamma(\cdot)$ is strictly increasing on $(-1,1)$, ranging from $-\infty$ to $+\infty$. Moreover, the equilibrium effort $m_{i(j)}$ can be represented to the product of $\phi(y;0,\sigma^2(T-t))$, the probability density of normal distribution with mean zero and variance $\sigma^2(T-t)$ at the state y and an amplitude factor that only depends on the composite variable z.

Figure...

3. Model Estimation

In this section, we describe the estimation procedure. We first outline the data generation process, establishing the connection between the empirical data and the theoretical model discussed previously. Then, we introduce a structural estimation method using Bayesian framework.

3.1. Data Generating Process

For each contest on Kaggle, the observable information can be classified into three primary components:

The first component consists of essential contest details, including the contest duration, prize structure, information disclosure, and other governing rules. Contrary to the assumptions of our model, a typical contest usually involves multiple teams rather than just two. We index the participating teams of the contest by $i\in\{1,2,...,n\}$. To fully leverage the potential of the data and establish a connection with our theoretical model, let's assume that each participant perceives a competitor they are playing against at every moment, denoted as j. This perceived competitor is typically understood as the most prominent individual on the leaderboard, i.e., the person ranked

first. When team i themselves hold the top position, their perceived competitor is the individual who poses the greatest threat, namely the person ranked second. Furthermore, the contests may feature intricate prize structures, such as the provision of multiple awards, rather than adhering to a winner-takes-all format. The issue will be discussed in Section 4.2.

The second component captures the submission events of each player i to the system, denoted by the sequence $\{\hat{t}_k^i\}_{k=1}^{N_i}$. Here, N_i represents for the total number of submissions by player i, and t represents for the time of each submission. We understand the submission events of player i and j as two conditional independent inhomogeneous Poisson processes, driven by the intensity functions $\tau_i(t)$ and $\tau_j(t)$.[2] Then, during any time interval $\mathcal{S}$ of the contest duration $\mathcal{T}$, the Poisson arrival rate of submissions of the representative player i is given by $\int_{s\in\mathcal{S}} \tau_i(s)ds$. We assume the submission intensity $\tau_i(t)$ is proportional to the effort level $m_i(\tilde{y}_t, t)$. More specifically,

$$\tau_i(t) = r \cdot m_i(\tilde{y}_t, t) \tag{9}$$

where $r > 0$ is the common ratio of submission intensity and effort.

The third component of the observed data is an open leaderboard that records the real-time rankings and scores of each participant, denoted by $\hat{x}_t^i$. We interpret the difference in scores between i and j displayed on the leaderboard as the signal Z_t (defined in (2)) released by the contest organizer. Specifically, let's denote $\hat{y}_t = \hat{x}_t^i - \hat{x}_t^j$ the gap between displayed scores and

$$dZ_t = \hat{y}_t dt \tag{10}$$

As indicated in (2), we assume that $\hat{y}_t$ represents a noisy signal, with its precision controlled by λ.

In practice, the leaderboard signals are inherently noisy, as organizers deliberately disclose only a subset of the full dataset to participants to mitigate the risk of overfitting. The proportion of the released data is generally known to all participants. In addition to the public leaderboard, most competitions hosted on Kaggle also maintain a private leaderboard, where organizers evaluate the true predictive performance of participants' models using the full dataset.

Beyond the uncertainty introduced by partial data disclosure, a second source of signal noise arises from the timing of leaderboard updates: rankings are refreshed only after model submissions, creating a delay relative to the true standings. It should be noted that such lag would bias our model estimates only if participants strategically timed their submissions. Although such strategic behaviour is theoretically possible, we abstract from it in this study. We assume that each submission follows a period of substantive effort, allowing the leaderboard rankings to broadly reflect the relative performance of algorithms based on the publicly available data.

[2] That is, given the intensity functions $\tau_i(t)$ and $\tau_j(t)$, the submission events $\{\hat{t}_k^i\}_{k=1}^{N_i}$ and $\{\hat{t}_k^j\}_{k=1}^{N_i}$ are mutually independent.

3.2. Bayesian Framework

The likelihood function of any realization of this point process $\{\hat{t}_k^i\}_{k=1}^{N_i}$ is given by

$$p\left(\{\hat{t}_k^i\}_{k=1}^{N_i}|\tau_i\right) = \exp\left\{-\int_{s\in\mathcal{T}} \tau_i(s)ds\right\} \prod_{k=1}^{N_i} \tau_i(\hat{t}_k^i) \tag{11}$$

By (3) and (10), the estimated gap $\tilde{y}_t$ by the two players follows

$$d\tilde{y}_t = (q_{i,t} - q_{j,t})dt + \sqrt{\lambda}\sigma(\hat{y}_t - \tilde{y}_t)dt \tag{12}$$

Unknown parameters to be estimated:

- σ (Total innovation risk): prior
- c_i and c_j (Capacities)

3.3. Estimation Procedure

...

3.3.1. Discrete time. ...

4. Application

...

4.1. Synthetic Data

Before applying our estimation procedure to real-world contest data, we evaluate its potential on synthetically generated data.

Based on the following assumptions, we will next generate the submission data for two players in a data analysis competition:

- Total innovation uncertainty of the two players per day: $\sigma = 5$
- Capacities: $c_i = 1.5$ and $c_j = 2$
- Contest duration: from 2025-01-01 to 2025-03-31
- Starting point $\tilde{y}_0 = 0$

A central challenge in generating synthetic data lies in dynamically constructing a point process that conforms to an inhomogeneous Poisson process. Inspired by the thinning technique (Lewis and Shedler 1979), we first generate candidate events according to a homogeneous Poisson process within each discrete time interval, using a fixed intensity $\tau_i^\star \geq \sup_t \tau_i(t)$, and distribute them uniformly across the interval. We then apply the thinning procedure to determine whether each candidate event is accepted.

4.2. Case Study

...

Algorithm 1 Synthetic Data Simulation

Input: c_i, c_j, σ, λ, r, $\tau^\star$, y_0, $\tilde{y}_0$

Sample $\{s_k^{i(j)}\}_{k=1}^{N_{i(j)}}$ from homogeneous Poisson process $(\tau^\star_{i(j)})$ on $[0,T)$

Sample $\{u_k^{i(j)}\}_{k=1}^{N_{i(j)}}$ from uniform distribution on $[0,1]$

Initialize $\ell=0$, $\ell^{i(j)}=0$, $\hat{y}_0=0$

for $t=0$ to T **do**

$m_{i(j)}(\tilde{y}_t,t) \leftarrow (8)$, $\tau^{i(j)}(t) \leftarrow (9)$

$\tilde{y}_{t+\Delta t} \leftarrow (12)$

for each $s_k^{i(j)} \subset [t,t+\Delta t)$ **do**

if $u_k^{i(j)} > \tau_{i(j)}(s_k^{i(j)})/\tau^\star_{i(j)}$ **then**

$\hat{t}_\ell^{i(j)} \leftarrow s_k^{i(j)}$; $\ell^{i(j)} \leftarrow \ell^{i(j)}+1$ {Accept the submission event}

$\hat{y}_\ell \leftarrow$; $\ell \leftarrow \ell+1$ {Update public leaderboard}

end if

end for

end for

Output: (y_t), $(\tilde{y}_t)$, $\{\hat{t}_k^{i(j)}\}_{k=1}^{N_{i(j)}}$, $\{\hat{y}_k\}_{k=1}^{N_i+N_j}$

5. Conclusion

References

Barrau A, Bonnabel S (2017) The invariant extended kalman filter as a stable observer. IEEE Transactions on Automatic Control 62(4):1797–1812, ISSN 0018-9286, 1558-2523, URL http://dx.doi.org/10.1109/TAC.2016.2594085.

Bensoussan A (1992) Stochastic Control of Partially Observed Systems (Cambridge University Press), 1 edition, ISBN 9780511526503, URL http://dx.doi.org/10.1017/CBO9780511526503.

Bimpikis K, Ehsani S, Mostagir M (2019) Designing dynamic contests. Operations Research 67(2):339–356, URL http://dx.doi.org/10.1287/opre.2018.1823.

Budd C, Harris C, Vickers J (1993) A model of the evolution of duopoly: Does the asymmetry between firms tend to increase or decrease? The Review of Economic Studies 60(3):543–573.

Frogerais P, Bellanger JJ, Senhadji L (2012) Various ways to compute the continuous-discrete extended kalman filter. IEEE Transactions on Automatic Control 57(4):1000–1004, ISSN 0018-9286, 1558-2523, URL http://dx.doi.org/10.1109/TAC.2011.2168129.

Harris C, Vickers J (1987) Racing with uncertainty. The Review of Economic Studies 54(1):1–21, ISSN 00346527, 1467937X, URL http://www.jstor.org/stable/2297442.

Lewis PAW, Shedler GS (1979) Simulation of nonhomogeneous poisson processes by thinning. Naval Research Logistics Quarterly 26(3):403–413, ISSN 0028-1441, 1931-9193, URL http://dx.doi.org/10.1002/nav.3800260304.

Moscarini G, Smith L (2007) Optimal dynamic contests. Working Paper 1–18, science (June).

Risdal M, Bozsolik T (2022) Meta kaggle. URL http://dx.doi.org/10.34740/KAGGLE/DS/9.

Ryvkin D (2022) To fight or to give up? dynamic contests with a deadline. Management Science 68(11):8144–8165, ISSN 0025-1909, 1526-5501, URL http://dx.doi.org/10.1287/mnsc.2021.4206.