Delay-Predictability Tradeoffs in Reaching a Secret Goal

John N. Tsitsiklis
LIDS, Massachusetts Institute of Technology, Cambridge, MA 02139, USA jnt@mit.edu

Kuang Xu
Graduate School of Business, Stanford University, Stanford, CA 94305, kuangxu@stanford.edu

We formulate a model of sequential decision-making, dubbed the Goal Prediction game, to study the extent to which an overseeing adversary can predict the final goal of an agent who tries to reach that goal quickly, through a sequence of intermediate actions. Our formulation is motivated by the increasing ubiquity of large-scale surveillance and data collection infrastructures, which can be used to predict an agent's intentions and future actions, despite the agent's desire for privacy.

Our main result shows that with a carefully chosen agent strategy, the probability that the agent's goal is correctly predicted by an adversary can be made *inversely proportional* to the time that the agent is willing to spend in reaching the goal, but cannot be made any smaller than that. Moreover, this characterization depends on the topology of the agent's state space only through its diameter.*

Key words: privacy, secrecy, goal reaching

1. Introduction

Information technologies and large-scale surveillance infrastructures have become ubiquitous and continue to expand at a rapid pace. It is conceivable that in the near future most of an agent's virtual or physical actions will be measured, monitored, and recorded by private enterprises or governmental entities. These measurements will potentially enable data collectors to make powerful predictions of an agent's intentions or future behavior, based on knowledge of her past actions.

In an environment where past actions are increasingly difficult to conceal, can we still keep our intentions unpredictable? If so, what is the additional effort required? Conversely, can a data collector design reliable prediction methods that are robust even against a sophisticated agent who carefully engineers a sequence of actions to hide her true goals? In this paper, we aim to study such issues in the context of a simple model of sequential decision-making, which we call the *Goal Prediction game*, and to characterize an agent's intrinsic level of *predictability* as she *approaches*

* This version: August, 2017. The research is supported in part by NSF Grant CMMI-1234062 and a Stanford Cyber Initiative Research Grant. The authors would like to thank the anonymous referees for their detailed, constructive feedback.

the final goal. Our main result shows, in a fairly general setting, that the predictability of the agent's final goal can be made *inversely proportional* to the time the agent is willing to spend in reaching it. While our model is highly stylized, it is intended to provide insights on the general tradeoff between predictability and the concealment effort.

1.1. Preview of the Model

We begin by informally describing our model; precise definitions will be given in Section 2. The Goal Prediction game is played between an agent (e.g., an individual) and an *adversary* (e.g., a data collector or a law enforcement agency), in discrete time. The agent's *state* at any time t, belongs to a finite set $\mathcal{V}$. At time $t = 1$, the agent is at an initial state $x_1 \in \mathcal{V}$, and has a goal D, drawn randomly from $\mathcal{V}$ according to some prior distribution. The goal is unknown to the adversary. The agent's objective is to approach and eventually reach the goal D, through a sequence of state transitions. We assume that the agent's state transitions are constrained to lie along the edges of a given undirected graph G with vertex set $\mathcal{V}$.

The objective of the adversary is to guess the identity of the goal D, by the time it is reached by the agent. We assume that the adversary can make at most one attempt to guess the goal, at a time of her own choosing. In particular, at any given time, the adversary can either wait or make a guess, based on the knowledge of the graph G, the agent's decision making strategy, and *all* of the agent's past actions (i.e., state transitions); the only information that the adversary does not possess is D itself, and any internal randomness that the agent might use when choosing her transitions. The adversary wins the game if she correctly guesses D by the time that it is reached by the agent, and loses, otherwise.

1.2. Motivating Examples

The structure of the model is easiest to understand in the context of a highly abstracted version of a *law enforcement* scenario. The agent, who is engaged in illegal activities, moves between different physical locations (vertices in $\mathcal{V}$). One of these locations (the goal state) is special, in that it stores certain incriminating evidence (e.g., drugs or explosives). At each stage of the process, the agent can move to a new location, as allowed by the underlying graph G. Similarly, at each stage, the adversary (the authorities) has three options:

(a) do nothing;
(b) perform a raid at the current location of the agent;
(c) perform a raid at some other location.

Here, a raid corresponds to guessing the special location and acting on this guess. The agent loses if the special location is raided at the time that the agent reaches that location, or earlier. If on

the other hand, the agent reaches the special location and no raid takes place by that time, then the agent has succeeded in her illegal activities, and wins. Our restriction that the adversary can make at most one attempt to guess the goal, reflects an assumption that once law enforcement performs a raid, its cover is exposed, the illegal agent can change its mode of operations, and hence law enforcement loses because it fails to arrest her.

Remark: It can be shown that under our model, there is no loss of optimality on the part of the adversary if at each time they are only allowed to raid the agent's current location; that is, if they never exercise option (c) above. However, we will still allow option (c) in the authorities' strategy space. This additional flexibility can become relevant in extensions of our model in which the adversary's objective incorporates a preference for an early decision/raid.

Our model is not limited to the law enforcement application, and may also be used to capture, say, the strategic interactions between a small firm trying to conceal its long-term product strategy against a competitor who tries to predict the firm's future products.

1.3. The Tradeoffs

In the example in the preceding subsection, it is natural to assume that the agent prefers to reach the goal sooner rather than later. This leads to a tradeoff between *delay* and *predictability*: if the agent is eager to reach the goal quickly by traversing a shortest path, then the adversary, by observing a few initial actions of the agent, may be able to quickly eliminate many potential goals, namely those that are inconsistent with the path taken, leading to a high probability of a correct guess. On the other hand, an extremely patient agent, with a large time budget, can start along a path that traverses the entire state space (e.g., a Hamiltonian path) and stop when her goal is reached. Under this strategy for the agent, her past actions reveal essentially nothing about the identity of the goal, making prediction difficult. A general strategy for the agent will typically lie somewhere in the middle, by combining a fast, time-efficient path, with some wasteful but obfuscating steps. Our main focus is to understand the resulting tradeoff, between predictability and time wasted in obfuscating actions.

Example: Complete graph. Suppose that G is a complete graph with n nodes, and let k be an integer between 1 and n. The agent can use the following simple strategy. Generate a set of k nodes that consists of the goal and another $k-1$ randomly chosen nodes. Then, visit those k nodes in a random order. Even if this k-element set is revealed to the adversary, the adversary can do no better than a random guess, so that the probability of a correct guess is $1/k$. At the same time, the agent reaches the goal in approximately $k/2$ time units, in expectation.

In the above example, and for the particular agent strategy that we discussed, we see a tradeoff, in the form of an inverse relation between the probability that the adversary wins and the expected time to reach the goal. Our subsequent results show a similar tradeoff for Pareto optimal agent strategies, and also for arbitrary graphs.

1.4. Preview of Main Result

In this subsection, we introduce some terminology, and summarize our results. The key quantities of interest are the *delay* of the agent, defined as the expected number of steps until the goal is reached, and the *prediction risk*, defined as the probability that the adversary makes a correct guess by the time that the goal state is reached. Our main result (Theorem 1) characterizes the minimax prediction risk for the agent, for any connected undirected graph, G. They are informally stated below, where, for simplicity of exposition, we assume that the agent's goal is distributed uniformly at random in $\mathcal{V}$.

1. Let d be the diameter of a graph G with n vertices. For any w that satisfies $d < w \leq n$, we show that there exists an agent strategy with a delay of at most w, under which the prediction risk, denoted by q, satisfies the upper bound

$$q \leq \frac{2}{w-d}, \tag{1}$$

   against *any* adversary strategy. Note that the diameter, d, is essentially the unavoidable worst-case delay in reaching the goal, even in the absence of a secrecy constraint.
2. Conversely, given any agent strategy that incurs a delay of at most w, we show that there exists a strategy for the adversary which correctly predicts the goal with probability

$$q \geq \frac{1}{2w+1}. \tag{2}$$

These results, taken together, establish that the intrinsic prediction risk faced by the agent is *inversely proportional* to the delay she is willing to sustain:

$$\text{prediction risk} = \Theta\left(\frac{1}{\text{delay}}\right), \tag{3}$$

and that this holds regardless of the detailed topology of the agent's state space. Our proof is constructive and provides concrete strategies for the two players that achieve the upper and lower bounds on the prediction risk.

1.5. Related Work

To the best of our knowledge, our formulation is new. Our model is related, in spirit, to the literature on search games (cf. Alpern and Gal (2003), Garnaev (2000)), in which a searcher tries to identify the hidden location of an evader, and the decision maker is concerned with finding search strategies that minimize the time until the evader is found. More broadly, researchers have also investigated how to efficiently detect terror plots (Kaplan (2015)), or uncover consumer choices from behavioral data Cummings et al. (2016). While the details differ, the majority of these models focus on uncovering some present or past state of an opponent; in contrast, our work focuses on the predictability of the agent's future actions, and it brings about very different strategic considerations.

The challenges faced by the adversary in our model are reminiscent of those of an inspector in inspection games (Dresher (1962), von Stengel (2016), Avenhaus et al. (2002)), whose objective is to allocate a small number of inspections across a finite number of time slots so as to efficiently detect any violations that an inspectee may commit during this period. Among other things, a crucial difference between our model and this literature is that inspection games do not possess a spatial dimension, while in our case the agent must traverse an underlying graph and reach a specific node in order for a "violation" to occur. As a result, the actions available to the inspectee in an inspection game, i.e., to violate or act lawfully, remain the same throughout the game, whereas those of our agent evolve over time as she traverses the graph.

There is also a large body of computer science literature on information security, which is concerned with secure communication or computation protocols that can prevent an adversary from accessing information (cf. Pfleeger and Pfleeger (2002)). More recently, there has been a growing interest in designing privacy-preserving data release protocols, which aim to protect individual identities when releasing statistical summaries of a data set, typically by injecting noise in the outputs (cf. Dwork and Roth (2013)). Compared to these areas of research, our formulation assumes a much stronger adversary who observes all past actions of the agent and her strategy. This could arise when the security or privacy-preserving mechanisms employed by an individual have already been compromised, which is not difficult to imagine when the adversary has a vast technological superiority over the individual, e.g., a large Internet provider versus an average user. As a result, in our model, the agent can only conceal her goal by *doing*, not by *hiding*. It would be interesting to study the possible improvement of the agent's performance if she is capable of also hiding some of her past actions, but this is beyond the scope of the current paper.

1.6. Organization

The remainder of the paper is organized as follows. We formally define the Goal Prediction game in Section 2, along with the performance metrics. Our main result is stated in Section 3. Its proof is given in Section 4, after an overview of the main steps. Section 5 presents some numerical results. Section 6 discusses an alternative measure of prediction risk. We conclude in Section 7, together with a discussion of possible model variations.

2. The Goal Prediction Game

In this section, we provide a formal definition of the Goal Prediction game. The game is played between two players, the *agent* and the *adversary*, and will be defined in terms of the following elements.

Definition 1 (The elements of the Goal Prediction Game). The game is specified by a quintuple $(G, x_1, \pi, \mathcal{R}_A, \mathcal{R}_D)$, consisting of:

(a) an undirected graph, $G = (\mathcal{V}, \mathcal{E})$, with n vertices;

(b) an initial agent state $x_1 \in \mathcal{V}$;

(c) a probability distribution π, used to generate a $\mathcal{V}$-valued random variable D, with components $\pi_v = \mathbb{P}(D = v)$.

(d) an auxiliary collection of independent random variables $\mathcal{R}_A$;

(e) an auxiliary collection of independent random variables $\mathcal{R}_D$.

For an interpretation of the different elements of the game, $\mathcal{V}$ represents the possible states of the agent, $\mathcal{E}$ the transitions that are allowed at each step, and D the agent's goal. Finally, $\mathcal{R}_A$ and $\mathcal{R}_D$ are independent internal random variables that the agent or the adversary, respectively, can use for the purpose of randomization.

2.1. Agent Strategies and Trajectories

An agent **trajectory** is defined as a sequence of random variables $\{(X_t, \Gamma_t)\}_{t\in\mathbb{N}}$, where the X_t take values in the set $\mathcal{V}$ and satisfy the constraints $X_1 = x_1$ and $(X_t, X_{t+1}) \in \mathcal{E}$, for all $t \geq 1$, and the Γ_t take values in some arbitrary set.

We interpret X_t as the *state* of the agent at time t. Furthermore, Γ_t encodes any *side information* that the agent is willing to provide to the adversary at the beginning of time slot t. The use of Γ_t is mainly intended to simplify our analysis and does not change the nature of the game, since the agent can always choose to not disclose any side information, by setting Γ_t to a fixed symbol for all $t \in \mathbb{N}$.

An *agent strategy*, generically denoted by ψ, is a mapping that takes G, x_1, and the realized values of D and $\mathcal{R}_A$ as inputs, and generates the agent's trajectory. For any agent strategy ψ, we define the time that the goal is reached (to be referred to as the *goal-reaching time*) as

$$T_\psi = \min\{t \in \mathbb{N} : X_t = D\}., \tag{4}$$

with the convention that $T_\psi = \infty$ if the goal is never reached. We will refer to $\mathbb{E}(T_\psi)$ as the *delay* of agent strategy ψ, where the expectation is taken with respect to the randomness in D and $\mathcal{R}_A$.

2.2. Adversary Actions and Strategies, and the Prediction Risk

In contrast to the agent's strategy, which generates the entire trajectory at once, the adversary's strategy, generically denoted by χ, operates sequentially, because it also takes into account the agent's past actions.

At each time t, the adversary's strategy has access to G, x_1, $\mathcal{R}_D$, and the history of the agent trajectory up to an including time t, namely, $X_1, \ldots, X_t$, and $\Gamma_1, \ldots \Gamma_t$. On the basis of this information, the adversary determines the realized value of an associated decision random variable, $\widehat{D}_t \in \mathcal{V} \cup \{0\}$.[1] Under a given pair (ψ, χ) of strategies for the agent and the adversary, respectively, denote by $U_{\psi,\chi}$ the first time that the adversary lets $\widehat{D}_t$ be an element of $\mathcal{V}$, i.e.,

$$U_{\psi,\chi} = \inf\{t : \widehat{D}_t \in \mathcal{V}\}. \tag{5}$$

We interpret $\widehat{D}_{U_{\psi,\chi}}$ as the prediction made by the adversary. See Figure 1 for an illustration.

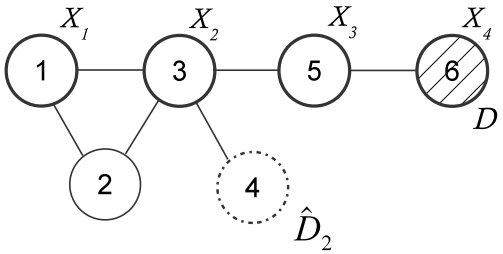

Figure 1 An example of the Goal Prediction game. The agent starts at the initial state 1, and the (secret) goal is state 6. The agent plans to follow the trajectory $1 \to 3 \to 5 \to 6$. Suppose that the adversary makes a prediction at time $t = 2$ by letting $\widehat{D}_2 = 4$. At that time, the game terminates, and since $\widehat{D}_2 \neq D$, the agent wins. If on the other hand, the adversary makes a prediction $\widehat{D}_t = 6$ at any time $t \leq 4$, then the adversary wins.

[1] We are assuming here that 0 is not one of the elements of $\mathcal{V}$.

The adversary wins the game if it makes a correct prediction of D, no later than the goal-reaching time. We define the *prediction risk* to be the probability of this event, namely,

$$q(\psi,\chi) = \mathbb{P}\left(\widehat{D}_{U_{\psi,\chi}} = D \text{ and } U_{\psi,\chi} \le T_\psi\right), \tag{6}$$

where the the probabilities are calculated with respect to the randomness in D, $\mathcal{R}_A$, and $\mathcal{R}_D$.

2.3. The Minimax Prediction Risk

If the agent is to use strategy ψ, she will be concerned about all possible adversary strategies. By focusing on the most unfavorable adversary strategy χ, we are led to define the *maximal prediction risk of* ψ by

$$q^*(\psi) = \sup_\chi q(\psi,\chi). \tag{7}$$

As already discussed, we are interested in the tradeoff between the prediction risk and the expected time to reach the goal. Accordingly, for any given time budget $w \in \mathbb{R}_+$, we define Ψ_w as the set of all agent strategies with a delay of at most w:

$$\Psi_w = \{\psi : \mathbb{E}(T_\psi) \le w\}. \tag{8}$$

The agent is interested in minimizing the maximal prediction risk, subject to a given time budget. Accordingly, we define the *minimax prediction risk* associated with a given time budget w as

$$\mathcal{Q}(w) = \inf_{\psi\in\Psi_w} q^*(\psi) = \inf_{\psi\in\Psi_w} \sup_\chi q(\psi,\chi). \tag{9}$$

3. Main Result: Characterization of Minimax Prediction Risk

We denote by d_G the diameter of a graph G, i.e., $d_G = \max_{u,v\in\mathcal{V}} d(u,v)$, where $d(u,v)$ is the number of edges on a shortest path from u to v. The following is our main result.

THEOREM 1. *Fix $n \in \mathbb{N}$ and let $G = (\mathcal{V},\mathcal{E})$ be a connected undirected graph with n vertices. Fix $x_1 \in \mathcal{V}$ and $w \in \{1,\dots,n\}$, such that $w - d_G$ is a positive even integer.*[2] *Let c^* be the maximum entry of π, i.e., $c^* = \max_{v\in\mathcal{V}} \pi_v$. Then, the minimax prediction risk under a time budget of w satisfies*

$$\max\left\{\frac{1}{2w+1}, c^*\right\} \le \mathcal{Q}(w) \le \frac{2nc^*}{w-d_G}. \tag{10}$$

In the special case where the prior distribution π is uniform over $\mathcal{V}$, Theorem 1 yields a fairly tight characterization of the minimax prediction risk: as long as d_G is relatively small compared to w, the upper and lower bounds agree within a factor of 4. Indeed, the following corollary follows from Theorem 1 by replacing c^* with $1/n$.

[2] The assumption of $w - d_G$ being an even integer helps simplify notation by avoiding floors and ceilings throughout the proof, and can be easily relaxed.

COROLLARY 1. *Suppose that π is uniform over $\mathcal{V}$. Then, the minimax prediction risk in Theorem 1 satisfies:*

$$\frac{1}{2w+1} \leq \mathcal{Q}(w) \leq \frac{2}{w-d_G}. \tag{11}$$

4. Proof of Theorem 1

We present in this section the proof of Theorem 1, starting with an overview of the main steps involved.

4.1. Overview of the Proof

For the upper bound on the minimax prediction risk in Eq. (10), we focus on a specific family of agent strategies, which we call *segment-based strategies*, and show that the prediction risk under an appropriately chosen segment-based strategy always satisfies the upper bound, against any adversary strategy. A segment-based strategy can be roughly described as follows. For a given goal D, the agent first generates a certain random path in the graph G, referred to as a segment, which contains D, and immediately reveals the segment to the adversary. The agent then proceeds to traverse the segment in a deterministic manner, until D is reached. The key idea is to have the agent generate the segment so that conditional on the realized segment, all of its member vertices are nearly equally likely to be the goal, in which case, it is difficult for the adversary to make an accurate prediction.

The performance analysis of segment-based strategies is carried out in two parts. We first express the maximal prediction risk of a given segment-based strategy as a function of some basic structural properties of the strategy, which roughly correspond to the total number of possible segments, as well as the the degree of uniformity with which different segments are spread over the underlying graph G (Section 4.3). Subsequently (Section 4.4), we show how to construct a segment-based strategy with the desirable structure just alluded to. Combining these results, the upper bound in Theorem 1 then follows from some straightforward calculations.

The lower bound in Theorem 1 is proved in Section 4.6. We will show that for any agent strategy ψ, there exists a simple adversary strategy that results in a prediction risk of at least $1/(2w+1)$. The main insight is that, in order to have delay of at most w, the distribution of the agent's goal-reaching time T_ψ needs to be somewhat concentrated. As a consequence, there will exist a time $t^* \in \mathbb{N}$, such that $T_\psi = t^*$ with probability at least $1/(2w+1)$. The adversary can then achieve a prediction risk of $1/(2w+1)$ by waiting until t^*, and setting the prediction $\widehat{D}_{t^*}$ to the value of the current agent state X_{t^*}.

4.2. Segment-based Strategies

The upper bound in Eq. (10) is proved by focusing on a restricted class of agent strategies, which we now proceed to define. We say that $s = (s_1, \ldots, s_r)$ is a *segment*, of length $|s| = r$, if $(s_i, s_{i+1}) \in \mathcal{E}$, for $i = 1, \ldots, r-1$. (A segment may have repeated vertices.)

Definition 2 (Covering Index). Fix a set $\mathcal{S}$ of segments and a vertex $v \in \mathcal{V}$. We define $\mathcal{S}_v$ as the set of segments in $\mathcal{S}$ that contain v, i.e.,

$$\mathcal{S}_v = \{s \in \mathcal{S} : v \in s\}. \tag{12}$$

We define the *covering index* of v with respect to $\mathcal{S}$, denoted by $c_{\mathcal{S}}(v)$, as the number of segments that belong to $\mathcal{S}_v$:

$$c_{\mathcal{S}}(v) = |\mathcal{S}_v|. \tag{13}$$

We now describe segment-based strategies.

Definition 3 (Segment-based Strategies). Fix $r \in \mathbb{N}$. Let $\mathcal{S}$ be a set of segments, with each segment having the same length r, and with $c_{\mathcal{S}}(v) > 0$ for all vertices v. A *segment-based strategy based on* $\mathcal{S}$ is defined as follows.

1. Recall that $\mathcal{S}_v$ is the set of segments in $\mathcal{S}$ that cover the vertex v (cf. Eq. (12)). The agent chooses a segment S_D, uniformly at random from $\mathcal{S}_D$.
2. The agent then travels to the goal, in two stages:
   (a) Stage 1. She travels to the first vertex, S_D^1, of the segment S_D along a shortest path.
   (b) Stage 2. She then travels along the segment S_D.
3. The agent, at time $t = 1$, announces to the adversary (through the side-information variable Γ_1) the segment S_D, as well as the shortest path to be followed in Stage 1. She also sets Γ_t, for $t \geq 2$ to a fixed (hence uninformative) symbol.

Note that the goal D will be reached either during Stage 1, if D happens to lie on the shortest path from x_1 to L_D^1, or, otherwise, during Stage 2. In either case, the time at which the goal is reached is upper bounded by the diameter of the graph, plus the length r of of the chosen segment, so that

$$T_\psi \leq d_G + r. \tag{14}$$

In particular, given a time budget $w > d_G$, it suffices to set $r = w - d_G$.

4.3. Characterization of the Maximal Prediction Risk under a Segment-Based Strategy

In this section we develop a characterization of the maximal prediction risk under a segment-based strategy, in terms of the covering indices associated with $\mathcal{S}$, the size of the set $\mathcal{S}$ of segments, and the prior distribution π. This result will be used in the next subsection to derive an upper bound on the maximal prediction risk.

PROPOSITION 1. *Let ψ be a segment-based strategy, based on a family $\mathcal{S}$ of segments. Its maximal prediction risk is given by*

$$q^*(\psi) = \sup_{\chi} q(\psi, \chi) = \sum_{s \in \mathcal{S}} \max_{v \in s} \frac{\pi_v}{c_{\mathcal{S}}(v)}. \tag{15}$$

Proof. We begin the proof with a simple observation: given that the trajectory to be followed by the agent is announced in the beginning (through Γ_1) to the adversary, the adversary's decisions $\widehat{D}_t$ are all determined at time $t = 1$. That is, there is an optimal adversary strategy in which $U_{\psi,\chi} = 1$. Let $\widehat{D} = \widehat{D}_1$ be the adversary's prediction of D, under a strategy χ of this type.

Because of the above observation, the adversary should use a Maximum a Posteriori Probability rule, and set $\hat{D}$ to be equal to a node v for which the probability $\mathbb{P}(D = v \mid \Gamma_1)$ is largest. Note that the value of Γ_1 provides information to the adversary that the target node D belongs to a certain segment s. In particular, the adversary should simply maximize, over v, the probability

$$\mathbb{P}(D = v \mid S_D = s).$$

It follows that the conditional maximal prediction risk, given $S = s$ is $\max_{v \in s} \mathbb{P}(D = v \mid v \in s)$. We then have

$$\begin{aligned}
\sup_{\chi} q(\psi_{\mathrm{d}}, \chi) &\overset{(a)}{=} \sum_{s \in \mathcal{S}} \mathbb{P}(\widehat{D}_1 = D \,\big|\, S_D = s)\, \mathbb{P}(S_D = s) \\
&\overset{(b)}{=} \sum_{s \in \mathcal{S}} \Big(\max_{v \in s} \mathbb{P}(D = v \,\big|\, S_D = s) \Big)\, \mathbb{P}(S_D = s) \\
&= \sum_{s \in \mathcal{S}} \max_{v \in s} \big(\mathbb{P}(S_D = s \,\big|\, D = v)\, \mathbb{P}(D = v) \big) \\
&\overset{(c)}{=} \sum_{s \in \mathcal{S}} \max_{v \in s} \frac{\pi_v}{c_{\mathcal{S}}(v)},
\end{aligned}$$

where step (a) follows from the definition of $q(\psi, \chi)$, step (b) from our earlier discussion, and step (c) from the fact that S_D is drawn uniformly at random from $\mathcal{S}_v$, the set of segments in $\mathcal{S}$ that cover v, and whose number is $c_{\mathcal{S}}(v)$. This completes the proof of Proposition 1. Q.E.D.

4.4. Constructing a Good Segment Family

In this subsection, we show how to construct a "good" segment family $\mathcal{S}$, so that the resulting maximal prediction risk, in Eq. (15), is small. Before presenting the details, let us first discuss, heuristically, the properties that a good segment family should satisfy. From Eq. (15), we see an inverse dependence on the covering indices, $c_{\mathcal{S}}(v)$. Thus, a good segment family should ensure that most of the vertices of G are covered by a large number of segments. At the same time, we want to ensure that the number of summands on the right-hand side of Eq. (15) is not too large, i.e., we do not want to have too many segments. In the remainder of this subsection, we show how to satisfy both of the above requirements. Specifically, we show that there exists a segment family containing no more than $2(n-1)$ segments, under which all vertices are covered by at least r segments.

PROPOSITION 2. *Let r be an even integer that satisfies $1 \leq r \leq n-1$. There exists a family $\mathcal{S}$ of segments of length r that satisfies*

$$|\mathcal{S}| = 2(n-1), \tag{16}$$

and

$$c_{\mathcal{S}}(v) \geq r, \quad \forall v \in \mathcal{V}. \tag{17}$$

Proof. We construct the segment family by exploiting some elementary properties of the depth-first traversal of a spanning tree. Let $H = (\mathcal{V}, \mathcal{E}_H)$ be a spanning tree of G, i.e., a connected subgraph of G that contains all vertices and has no cycles. In particular, the number of edges of the spanning tree, $|\mathcal{E}_H|$, is $n-1$. We say that a path (in which vertices may be repeated) is a *loop* if its first and last vertices coincide. Pick an arbitrary node $v_0 \in \mathcal{V}$ as the root of H, and let h be a loop generated by a depth-first-traversal of H that starts and ends on the vertex v_0 (cf. Section 22.3 of Cormen et al. (2009)); i.e., h traverses H by exploring as far as possible along each sub-tree of H before returning to the parent vertex. It is not difficult to see that h traverses each edge in H exactly twice, and hence we have that

$$|h| = 2\,|\mathcal{E}_H| + 1 = 2(n-1) + 1 = 2n - 1. \tag{18}$$

Let $h = (h_1, \ldots, h_{2n-1})$ be the above constructed loop. We extend the loop, by omitting the last vertex on the loop (which coincides with the initial vertex) and then continuing along the same loop, for another $r-1$ steps, resulting in a path of the form

$$\hat{h} = (h_1, h_2, \ldots, h_{2n-2}, h_1, h_2, \ldots h_{r-1}).$$

We now use $\hat{h}$ to construct a segment family $\mathcal{S}$, which consists of $|h|-1$ segments, where the ith segment, $s^{(i)}$, is of the form

$$s^{(i)} = \left(\hat{h}_i, \hat{h}_{i+1} \ldots, \hat{h}_{i+r-1}\right), \quad i = 1, \ldots, |h| - 1. \tag{19}$$

Note that since $|h|-1=2n-2$, the requirement in Eq. (16) is automatically satisfied. For a concrete example, if $n=3$, $h=(a,b,c,b,a)$, and $r=3$, then $\hat{h}=(a,b,c,b,a,b)$, the set $\mathcal{S}$ will consist of the $|h|-1=2n-2=4$ paths (a,b,c), (b,c,b), (c,b,a), and (b,a,b).

Next, we show that with the segment family $\mathcal{S}$ defined in Eq. (19), the covering index $c_{\mathcal{S}}(v)$ of every vertex v is at least r. Fix $i\in\{1,\ldots,|h|-1\}$. Since $\hat{h}$ is a continuation of h which repeats the first $r-1$ vertices of h, it follows that:

1. If $i\geq r$, then h_i belongs to $s^{(i-r+1)},\ldots,s^{(i)}$.
2. If $i<r$, then h_i belongs to the first i segments and the last $r-i$ segments of $\mathcal{S}$.

In particular, h_i belongs to at least r different segments, and $c_{\mathcal{S}}(h_i)\geq r$, which establishes Eq. (17).

4.5. Completing the Proof of the Upper Bound

Given a time budget $w>d_G$, we let $r=w-d_G$, and construct a segment-based strategy, with segments of length r, with the properties in Proposition 2. From Eq. (14), the strategy satisfies the budget constraint:

$$\mathbb{E}(T_\psi)\leq d_G+r=w.$$

Furthermore, by combining Propositions 1 and 2, we obtain

$$q^*(\psi)\overset{(a)}{=}\sum_{s\in\mathcal{S}}\max_{v\in s}\frac{\pi_v}{c_{\mathcal{S}}(v)}\overset{(b)}{\leq}\frac{c^*}{r}|\mathcal{S}|\overset{(c)}{=}\frac{2(n-1)c^*}{r}\leq\frac{2nc^*}{r}, \tag{20}$$

where $c^*=\max_{v\in\mathcal{V}}\pi_v$. Step (a) follows from Eq. (15) in Proposition 1. Steps (b) and (c) follow from the properties $c_{\mathcal{S}}(v)\geq r$ and $|\mathcal{S}|=2(n-1)$ in Proposition 2.

This completes the proof of the upper bound in Theorem 1.

4.6. Proof of the Lower Bound

We now turn to the lower bound in Theorem 1. We fix an agent strategy ψ for which $\mathbb{E}(T_\psi)\leq w$. We will show that a simple strategy for the adversary will perform well.

Definition 4 (Follow-and-Predict). Fix an agent strategy ψ, and let

$$t(\psi)\in\arg\max_{t\in\mathbb{N}}\mathbb{P}(T_\psi=t). \tag{21}$$

The adversary's strategy consists of making a prediction at time $t=t(\psi)$, equal to the agent's current state, i.e., $\widehat{D}_{t(\psi)}=X_{t(\psi)}$.

Let us denote by χ_F the Follow-and-Predict strategy for the adversary. To analyze its performance, we make use of the following fact. The proof involves an elementary application of the Markov's inequality and is given in Appendix A.1.

LEMMA 1. *Let Y be a random variable taking values in $\mathbb{N}$. Then, there exists $y \in \mathbb{N}$ such that*

$$\mathbb{P}(Y = y) \geq \frac{1}{2\mathbb{E}(Y) + 1}. \tag{22}$$

We observe that if $T_\psi = t(\psi)$, then the adversary makes a correct prediction, and this happens by the time that the agent reaches the goal, so that the adversary wins. Therefore, applying also Lemma 1 to T_ψ, we obtain

$$q(\psi, \chi_F) \geq \mathbb{P}(T_\psi = t(\psi)) = \max_{t \in \mathbb{N}} \mathbb{P}(T_\psi = t) \geq \frac{1}{2\mathbb{E}(T_\psi) + 1} \geq \frac{1}{2w + 1}. \tag{23}$$

Finally, note that the adversary could also make a prediction at time $t = 1$ based on the prior π alone, by choosing a v that maximizes π_v, which yields a probability of success of c^*. Therefore, we conclude that

$$\mathcal{Q}(w) \geq \max\left\{q(\psi, \chi_F), c^*\right\} \geq \max\left\{\frac{1}{2w+1}, c^*\right\}. \tag{24}$$

This proves the lower bound in Eq. (10), and completes the proof of Theorem 1.

5. Numerical Examples

Figure 2 illustrates simulation results on graphs with 100 vertices and a uniform prior distribution for the goal vertex. The agent in these experiments uses the segment-based strategy outlined in Section 4.2, while the length of the segment, r, varies. Each marker in the figure corresponds to a fixed value of r, where the delay and risk are calculated by averaging over 100 Erdős-Rény random graphs with edge probability $p = 0.2$, conditional on the graph being connected. The solid curve with diamond markers corresponds to the scenario where the adversary uses the Follow-and-Predict strategy (Definition 4), and the dashed curve with triangle markers corresponds to a simpler strategy, dubbed *naive*, where the adversary makes a prediction on a time slot chosen uniformly at random from the first w slots, where w is equal to the agent's delay. The dash-dot line is the lower bound in Eq. (11), i.e., $1/(2w + 1)$. The Follow-and-Predict strategy appears to outperform the naive strategy (from the point of view of the adversary) when the delay is small, and the gap between the two diminishes as the delay grows. Notably, the solid-diamond curve (Follow-and-Predict) in the log-log scale plot of Figure 2 is nearly linear with a slope of approximately -0.99. This suggests that the prediction risk scales inversely proportionally with respect to the agent's delay, which is consistent with the theoretical results in Corollary 1.

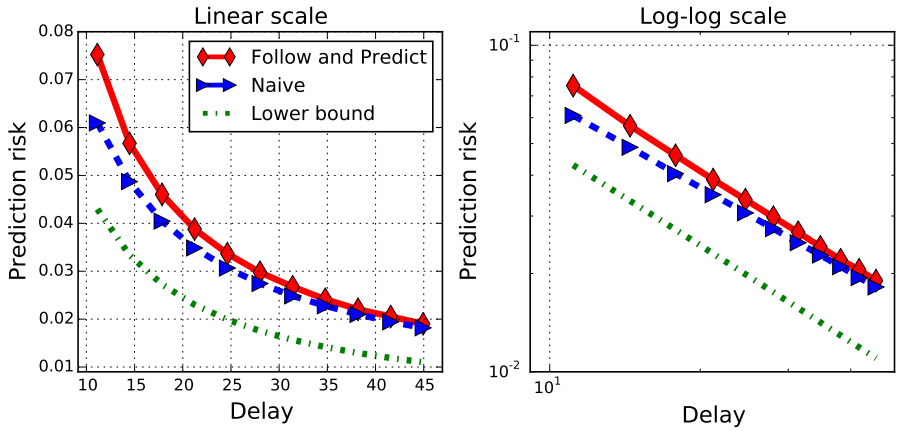

Figure 2 An illustration of the prediction risks as a function of the agent's delay.

6. The Maximin Prediction Risk

The minimax prediction risk defined in Eq. (9) can be interpreted as the minimax value of a zero-sum game played between an adversary and the agent, whose payoffs are the prediction risk (Eq. (6)) and its negative, respectively. It is therefore natural to consider, similar to Eq. (9), the *maximin prediction risk* associated with a time budget w:

$$\overline{\mathcal{Q}}(w) = \sup_{\chi} \inf_{\psi \in \Psi_w} q(\psi, \chi). \tag{25}$$

By the max-min inequality (cf. Section 5.4.1 of Boyd and Vandenberghe (2004)), we have

$$\overline{\mathcal{Q}}(w) = \sup_{\chi} \inf_{\psi \in \Psi_w} q(\psi, \chi) \le \inf_{\psi \in \Psi_w} \sup_{\chi} q(\psi, \chi) = \mathcal{Q}(w), \tag{26}$$

and hence the upper bound on the minimax prediction risk in Theorem 1 also applies to $\overline{\mathcal{Q}}(w)$. The result that follows, stated in a form that parallels Theorem 1, provides a lower bound on $\overline{\mathcal{Q}}(w)$ which is weaker than the lower bound in Theorem 1 by a factor of 2. The proof is based on an adversary strategy which is similar to but simpler than the Follow-and-Predict strategy (Definition 4), and is given in Appendix A.2.

THEOREM 2. *Fix $n \in \mathbb{N}$ and let $G = (\mathcal{V}, \mathcal{E})$ be a connected undirected graph with n vertices. Fix $x_1 \in \mathcal{V}$ and $w \in \{1, \ldots, n\}$, such that $w - d_G$ is a positive even integer. Let c^* be the maximum entry of π, i.e., $c^* = \max_{v \in \mathcal{V}} \pi_v$. Then, the maximin prediction risks under a time budget of w satisfies*

$$\max\left\{\frac{1}{4w}, c^*\right\} \le \overline{\mathcal{Q}}(w) \le \frac{2nc^*}{w - d_G}. \tag{27}$$

It may well be the case that $\mathcal{Q}(w) = \overline{\mathcal{Q}}(w)$, i.e., that the zero-sum game has a value.[3] But even if this were to be the case, an exact expression for the value is unlikely to become available (since it

[3] Unfortunately, such a result does not follow directly from the usual minimax theorems, because our strategy spaces are complicated enough, and do not readily satisfy the usual topological assumptions.

would depend in a complicated way on the detailed topology of the graph), and we do not expect to be able to state any results stronger than what is already implied by Theorems 1 and 2.

On a pragmatic level, we have focused on the minimax prediction risk because we consider it more relevant for the applications that we have in mind: the minimax formulation captures the thought process of an agent who wishes to protect herself against a powerful adversarial data collector. In contrast, the maximin formulation applies to the less realistic situation where the agent is sophisticated, experienced, and possessing information on the data collector's prediction strategy.

7. Conclusions and Model Variations

We have proposed in this paper a framework for quantifying the tradeoff between the predictability of an agent's goal and the additional effort that the agent is willing to sustain in order to hide its goal from an adversary who oversees the agent's actions. Our main result establishes that the probability of a correct prediction by the adversary scales in inverse proportion to the additional time that the agent is willing to spend. Furthermore, this result holds independently of the detailed topology of the individual's state space.

We now discuss a number of potentially interesting variations and extensions of our model and results, some of which may be better suited for modeling more realistic situations.

Far-from-uniform distributions of the final goal. We note that there remains a gap between the upper and lower bounds in Theorem 1, which becomes pronounced if the prior distribution of the goal is highly non-uniform. We believe that this gap can be reduced with an improved construction of the agent's strategy, which that takes into account the variability in the entries of π_D, although it is not clear how to do so within the framework of segment-based strategies. An interesting special case to consider is one where the goal is (approximately) uniformly chosen within a *proper* subset of nodes.

Directed graphs. In this variation, the underlying graph, G, has directed edges, i.e., it may be possible to transit from vertex v to u, but not the other way around. Our method of generating a family of segments (Section 4.4) cannot be applied directly, since one may not be able to traverse an edge of the spanning tree in both directions. On the positive side, as long as one is able to construct a family of segment that admits a lower bound on the covering index, as in Proposition 2, the proof of Theorem 1 should carry over to the directed case with little difficulty.

Weighted states or edges. We have quantified the agent's obfuscation effort in terms of the delay in reaching the goal. In more realistic settings, it may be more expensive to visit some states than others. It is thus natural to consider a generalization where each state is associated with a cost and the agent's cost is defined as the expected value sum of the total cost until the goal is reached.

If the cost at different states differs by at most a constant factor, segment-based strategies still apply, and results similar to Theorem 1 will again hold (with different constants involved). On the other hand, if some states are significantly more costly than others, the picture is less clear, since the agent may prefer to choose the segments in a way that depends on the costs of the states in a complex manner. A similar variation, involving different edge costs, may also be of interest.

Multiple executions or predictions. We may consider a scenario where the adversary is allowed to make more than one predictions. Here, the agent will encounter a new strategic dimension as she should now take into account the adversary's past predictions. We may consider yet another variation where the agent would like to execute more than one goal, and we should expect the adversary's strategy to become more nuanced by considering the goals already executed by the agent.

References

Alpern, S. and Gal, S. (2003). *The Theory of Search Games and Rendezvous*, volume 55. Springer.

Avenhaus, R., Von Stengel, B., and Zamir, S. (2002). Inspection games. *Handbook of Game Theory with Economic Applications*, 3:1947–1987.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. MIT press.

Cummings, R., Echenique, F., and Wierman, A. (2016). The empirical implications of privacy-aware choice. *Operations Research*, 64(1):67–78.

Dresher, M. (1962). A sampling inspection problem in arms control agreements: A game-theoretic analysis. Technical report, The RAND Corporation.

Dwork, C. and Roth, A. (2013). The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407.

Garnaev, A. (2000). *Search Games and Other Applications of Game Theory*, volume 485. Springer.

Kaplan, E. H. (2015). Socially efficient detection of terror plots. *Oxford Economic Papers*, 67(1):104–115.

Pfleeger, C. P. and Pfleeger, S. L. (2002). *Security in Computing*. Prentice Hall Professional Technical Reference.

von Stengel, B. (2016). Recursive inspection games. *Mathematics of Operations Research*, 41(3):935–952.

Appendix A: Proofs

A.1. Proof of Lemma 1

Proof. Let $\mu = \mathbb{E}(Y)$. Suppose, for the sake of contradiction, that $\mathbb{P}(Y = y) < 1/(2\mu+1)$, for all $y \in \mathbb{N}$. We then have

$$\mathbb{P}(Y \geq i) = 1 - \mathbb{P}(Y < i) = 1 - \sum_{j=1}^{i-1} \mathbb{P}(Y = y) > 1 - (i-1) \cdot \frac{1}{2\mu+1}. \tag{28}$$

We then obtain

$$\begin{aligned}
\mu &= \sum_{i=1}^{\infty} \mathbb{P}(Y \geq i) \\
&\geq \sum_{i=1}^{\lfloor 2\mu+1 \rfloor} \mathbb{P}(Y \geq i) \\
&> \sum_{i=1}^{\lfloor 2\mu+1 \rfloor} \Big(1 - (i-1) \cdot \frac{1}{2\mu+1}\Big) \\
&= \lfloor 2\mu+1 \rfloor - \frac{1}{2\mu+1} \cdot \frac{\lfloor 2\mu+1 \rfloor (\lfloor 2\mu+1 \rfloor - 1)}{2} \\
&\geq \mu,
\end{aligned}$$

which is a contradiction, and proves the desired result. Q.E.D.

A.2. Proof of Theorem 2

Proof. The upper bound follows directly from Theorem 1 and the max-min inequality, as discussed in the text preceding Theorem 2. We will hence focus on the lower bound. Same as in Theorem 1, the term c^* in the lower bound reflects the fact that the adversary can simply make a prediction at time $t=1$ on a vertex with the highest prior probability, resulting in a prediction risk of c^*. It therefore suffices to construct an adversary strategy, denoted by $\tilde{\chi}$, which guarantees a prediction risk of at least $1/4w$ against any agent strategy in Ψ_w. The adversary's strategy $\tilde{\chi}$ works as follows. The adversary first generates a random variable, $\tilde{T}$, uniformly distributed over $\{1, 2, \ldots, 2w\}$, and independent of everything else. She then makes a prediction at time $t = \tilde{T}$, equal to the agent's current state, i.e., $\widehat{D}_{\tilde{T}} = X_{\tilde{T}}$.

Fix an agent strategy $\psi \in \Psi_w$. We now show that the strategy $\tilde{\chi}$ leads to a prediction risk of at least $1/4w$ when deployed against ψ. Intuitively, because the agent's expected goal-reaching time, $\mathbb{E}(T_\psi)$, is constrained to be at most w, T_ψ must be no greater than $2w$ with probability of at least $1/2$. This ensures that an adversary using $\tilde{\chi}$ will make a correct prediction at $t = T_\psi$ with probability at least $1/4w$. To make this precise, we observe that the adversary wins the game if the event $\{\tilde{T} = T_\psi\}$ occurs, implying that

$$q(\psi, \tilde{\chi}) \geq \sum_{t=1}^{2w} \mathbb{P}(\tilde{T} = T_\psi = t) \overset{(a)}{=} \sum_{t=1}^{2w} \mathbb{P}(\tilde{T} = t)\mathbb{P}(T_\psi = t) \overset{(b)}{=} \frac{1}{2w}\mathbb{P}(T_\psi \leq 2w) \overset{(c)}{\geq} \frac{1}{4w}. \tag{29}$$

Step (a) follows from $\tilde{T}$ being independent from T_ψ, and (b) from $\tilde{T}$ being uniformly distributed. Step (c) is based on the assumption that $\psi \in \Psi_w$ and hence $\mathbb{E}(T_\psi) \leq w$:

$$\mathbb{P}(T_\psi \leq 2w) \geq 1 - \mathbb{P}(T_\psi \geq 2w) \geq 1 - \frac{\mathbb{E}(T_\psi)}{2w} \geq 1 - \frac{w}{2w} = \frac{1}{2}, \tag{30}$$

where the second inequality follows from Markov's inequality and the fact that T_ψ is non-negative. Because Eq. (29) holds for any $\psi \in \Psi_w$, we have that

$$\overline{\mathcal{Q}}(w) = \sup_{\chi} \inf_{\psi \in \Psi_w} q(\psi, \chi) \geq \inf_{\psi \in \Psi_w} q(\psi, \tilde{\chi}) \geq \frac{1}{4w}. \tag{31}$$

This completes the proof of Theorem 2. Q.E.D.