SIAM J. COMPUT.
Vol. 46, No. 6, pp. 1785–1826

NONSTOCHASTIC MULTI-ARMED BANDITS WITH GRAPH-STRUCTURED FEEDBACK*

NOGA ALON†, NICOLÒ CESA-BIANCHI‡, CLAUDIO GENTILE§, SHIE MANNOR¶, YISHAY MANSOUR‖, AND OHAD SHAMIR#

Abstract. We introduce and study a partial-information model of online learning, where a decision maker repeatedly chooses from a finite set of actions and observes some subset of the associated losses. This setting naturally models several situations where knowing the loss of one action provides information on the loss of other actions. Moreover, it generalizes and interpolates between the well-studied full-information setting (where all losses are revealed) and the bandit setting (where only the loss of the action chosen by the player is revealed). We provide several algorithms addressing different variants of our setting and provide tight regret bounds depending on combinatorial properties of the information feedback structure.

Key words. online learning, multi-armed bandits, learning from experts, learning with partial feedback, graph theory

AMS subject classifications. 68T05, 68Q32

DOI. 10.1137/140989455

1. Introduction.

Prediction with expert advice—see, e.g., [15, 17, 24, 35, 45]—is a general abstract framework for studying sequential decision problems. For example, consider a weather forecasting problem where each day we receive predictions from various experts and we need to devise our forecast. At the end of the day, we observe how well each expert did, and we can use this information to improve our forecasting in the future. Our goal is that over time, our performance converges to that of the best expert in hindsight. More formally, such problems are often modeled

*Received by the editors September 29, 2014; accepted for publication (in revised form) September 14, 2017; published electronically November 28, 2017. Preliminary versions of this manuscript appeared in [3, 37].

http://www.siam.org/journals/sicomp/46-6/98945.html

Funding: The first author was supported in part by a USA-Israeli BSF grant, by an ISF grant, by the Israeli I-Core program, and by the Oswald Veblen Fund. The second author was supported in part by MIUR (project ARS TechnoMedia, PRIN 2010-2011, grant 2010N5K7EB_003). The fourth author was supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 306638 (SUPREL). The fifth author was supported in part by a grant from the Israel Science Foundation, a grant from the United States-Israel Binational Science Foundation (BSF), a grant by Israel Ministry of Science and Technology, and the Israeli Centers for Research Excellence (I-CORE) program (Center 4/11). The sixth author was supported in part by a grant from the Israel Science Foundation (425/13) and a Marie-Curie Career Integration Grant.

†Department of Mathematics, Tel Aviv University, Tel Aviv, 6997801, Israel, Institute for Advanced Study, Princeton, NJ 08540, and Microsoft Research, Herzliya, 4672513, Israel (nogaa@tau.ac.il).

‡Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, 20135 Milano, Italy (nicolo.cesa-bianchi@unimi.it).

§Department of Information and Communication, University of Insubria, 21100 Varese, Italy (claudio.gentile@uninsubria.it).

¶Department of Electrical Engineering, The Technion, Haifa, 3200003, Israel (shie@ee.technion.ac.il).

‖Google Research, Israel, and School of Computer Science, Tel Aviv University, Tel Aviv, 6997801, Israel (mansour@tau.ac.il). Work was done in part while the author was at Microsoft Research, Hertzelia.

#Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 7610001, Israel (ohad.shamir@weizmann.ac.il).

as a repeated game between a player and an adversary, where in each round, the adversary privately assigns a loss value to each action in a fixed set (in the example above, the discrepancy in the forecast if we follow a given expert's advice). Then the player chooses an action (possibly using randomization) and incurs the corresponding loss. The goal of the player is to control regret, which is defined as the cumulative excess loss incurred by the player as compared to the best fixed action over a sequence of rounds.

In some situations, however, the player only gets partial feedback on the loss associated with each action. For example, consider a web advertising problem, where every day one can choose one ad to display to a user out of a fixed set of ads. As in the forecasting problem, we sequentially choose actions from a given set and may wish to control our regret with respect to the best fixed ad in hindsight. However, while we can observe whether a displayed ad was clicked on, we do not know what would have happened had we chosen a different ad to display. In our abstract framework, this corresponds to the player observing the loss of the action picked, but not the losses of other actions. This well-known setting is referred to as the (nonstochastic) multi-armed bandit problem, which in this paper we denote as the *bandit* setting. In contrast, we refer to the previous setting, where the player observes the losses of all actions, as the *expert* setting. In this work, our main goal is to bridge between these two feedback settings and explore the spectrum of models in between.

We start by recalling the performance attainable in the expert and the bandit settings, assuming bounded losses (for example, in the interval $[0, 1]$). Letting K be the number of available actions and T be the number of played rounds, the best possible regret for the expert setting is of order $\sqrt{\ln(K)\, T}$. This optimal rate is achieved by the Hedge algorithm [24] or the Follow the Perturbed Leader algorithm [28]. In the bandit setting, the optimal regret is of order $\sqrt{KT}$, achieved by the Implicitly Normalized Forecaster [5]. A bandit variant of Hedge, called Exp3 [7], achieves a regret with a slightly worse bound of order $\sqrt{K \ln(K)\, T}$. Thus, switching from the full-information expert setting to the partial-information bandit setting increases the attainable regret by a multiplicative factor of $\sqrt{K}$, up to extra logarithmic factors. This exponential difference in terms of the dependence on K can be crucial in problems with large action sets. The intuition for this difference in performance has long been that in the bandit setting, we only get $1/K$ of the information obtained in the expert setting (as we observe just a single loss, rather than all K losses at each round) and hence the additional K factor under the square root in the bound.

While the bandit setting received much interest, it can be criticized for not capturing additional side-information we often have on the losses of the different actions. As a motivating example, consider the problem of web advertising mentioned earlier. In the standard multi-armed bandit setting, we assume that we have no information whatsoever on whether undisplayed ads would have been clicked on. However, in many relevant cases, the semantic relationship among actions (ads) implies that we do indeed have some side-information. For instance, the user's reaction to a displayed ad might allow us to infer what behavior other related ads would have elicited from the same user. This sort of side-information is not captured by the standard bandit setting. A similar type of side-information arises in product recommendation systems hosted on online social networks, in which users can befriend each other. In this case, it has been observed that social relationships reveal similarities in tastes and interests [41]. Hence, a product liked/disliked by some user may also be liked/disliked by the user's friends.

Online auctions provide another motivating scenario for side-information in bandits. In online advertisement, individual impressions are sold to advertisers through programmatic instantaneous second-price auctions between a seller (the publisher) and the buyers (the advertisers). Both parties face a bandit problem: in each auction, the seller needs to set the reserve price (i.e., the smallest price below which the seller is not willing to sell) and the buyers need to determine their bids. The side-information is provided in part by the ad exchange (which, for instance, discloses the highest bid) and in part because buyer and seller profits are determined by known functions of the reserve price and the two highest bids. Therefore, depending on whether the impression was sold or not, buyers and seller may be able to compute ex-post the profits of bids/reserves higher or lower than those actually used in the auction—see, e.g., the regret analysis of Algorithm Exp3-RTB in [16], where the set of available actions for the seller is (a suitable discretized version of) the set of all possible reserve prices, and a natural graph-based feedback system can be defined over this set, where the loss of the played reserve price i also reveals to the seller the loss of all prices j such that $j \geq i$.

A further example, not in the marketing domain, is route selection, where we are given a graph of possible routes connecting cities. When we select a route connecting two cities, we observe the cost (say, driving time or fuel consumption) of the "edges" along that route, and, in addition, we have complete information on subroutes including any subset of the edges.[1]

Note that our framework also accommodates more general scenarios in which no dependencies (like the semantic relationship in the ad example) among losses have to be assumed. Namely, the adversary assigning losses to actions and the mechanism governing the release of side-information can be fully oblivious to each other. The study of settings in which the loss assignment and the feedback model are required to depend on each other in a certain way (e.g., observability implies similarity, or observability depends on the value of the player's loss) are definitely worth studying but transcend the scope of this paper—see also item 1 in the list of open problems of section 5.

In this paper, we present and study a setting which captures this type of side-information and in fact interpolates between the bandit setting and the expert setting. This is done by defining a *feedback system* under which choosing a given action also reveals the losses of some subset of the other actions. This feedback system can be viewed as a directed and time-changing graph G_t over actions: an arc (directed edge) from action i to action j implies that when playing action i at round t we get information also about the loss of action j at round t. Thus, the expert setting is obtained by choosing a complete graph over actions (playing any action reveals all losses), and the bandit setting is obtained by choosing an empty edge set (playing an action only reveals the loss of that action). The attainable regret turns out to depend on nontrivial combinatorial properties of this graph. To describe our results, we need to make some distinctions in the setting that we consider.

Directed versus symmetric setting. In some situations, the side-information between two actions is symmetric—for example, if we know that both actions will have a similar loss. In that case, we can model our feedback system G_t as an undirected graph. In contrast, there are situations where the side-information is not symmetric. For example, consider the side-information gained from asymmetric social links,

[1] Though this example may also be viewed as an instance of combinatorial bandits [18], the model we propose is more general. For example, it does not assume linear losses, which could arise in the routing example from the partial ordering of subroutes.

such as followers of celebrities. In such cases, it might be more likely that followers will shape their preferences after the person they follow rather than the other way around. Hence, a product liked by a celebrity is probably also liked by his/her followers, whereas a preference expressed by a follower is more often specific to that person. Another example in the context of ad placement is when a person buying a video game console might also buy a high-def cable to connect it to the TV set. On the other hand, interest in high-def cables need not indicate an interest in game consoles. In such situations, modeling the feedback system via a directed graph G_t is more suitable. Note that the symmetric setting is a special case of the directed setting, and therefore handling the symmetric case is easier than the directed case.

Informed versus uninformed setting. In some cases, the feedback system is known to the player before each round and can be utilized for choosing actions. For example, we may know beforehand which pairs of ads are related, or we may know the users who are friends of another user. We denote this setting as the informed setting. In contrast, there might be cases where the player does not have full knowledge of the feedback system before choosing an action, and we denote this harder setting as the uninformed setting. For example, consider a firm recommending products to users of an online social network. If the network is owned by a third party, and therefore not fully visible, the system may still be able to run its recommendation policy by only accessing small portions of the social graph around each chosen action (i.e., around each user to whom a recommendation is sent).

Generally speaking, our contribution lies in both characterizing the regret bounds that can be achieved in the above settings as a function of combinatorial properties of the feedback systems, as well as providing efficient sequential decision algorithms working in those settings. More specifically, our contributions can be summarized as follows (see section 2 for a brief review of the relevant combinatorial properties of graphs).

Uninformed setting. We present an algorithm (Exp3-SET) that achieves $\widetilde{\mathcal{O}}\Big(\sqrt{\ln(K)\sum_{t=1}^{T}\mathtt{mas}(G_t)}\Big)$ regret in expectation, where $\mathtt{mas}(G_t)$ is the size of the maximal acyclic graph in G_t, and the $\widetilde{\mathcal{O}}(\cdot)$ notation hides constants and logarithmic factors. In the symmetric setting, $\mathtt{mas}(G_t) = \alpha(G_t)$ (the independence number of G_t), and we prove that the resulting regret bound is optimal up to logarithmic factors, when $G_t = G$ is fixed for all rounds. Moreover, we show that Exp3-SET attains $\mathcal{O}\big(\sqrt{\ln(K)\,T}\big)$ regret when the feedback graphs G_t are random graphs generated from a standard Erdős–Renyi model.

Informed setting. We present an algorithm (Exp3-DOM) that achieves $\mathcal{O}\Big(\ln(K)\sqrt{\ln(KT)\sum_{t=1}^{T}\alpha(G_t)}\Big)$ regret in expectation, for both the symmetric and directed cases. Since our lower bound also applies to the informed setting, this characterizes the attainable regret in the informed setting, up to logarithmic factors. Moreover, we present another algorithm (ELP.P) that achieves $\mathcal{O}\Big(\sqrt{\ln(K/\delta)\sum_{t=1}^{T}\mathtt{mas}(G_t)}\Big)$ regret with probability at least $1-\delta$ over the algorithm's internal randomness. Such a high-probability guarantee (which relies on the bounded loss assumption) is stronger than the guarantee for Exp3-DOM, which holds just in expectation, and turns out to be of the same order in the symmetric case. However, in the directed case, the regret bound may be weaker since $\mathtt{mas}(G_t)$ may be larger than $\alpha(G_t)$. Moreover, ELP.P requires us to solve a linear program at each round, whereas Exp3-DOM only requires finding an approximately minimal dominating set, which can be done by a standard greedy set cover algorithm.

Our results interpolate between the bandit and expert settings; when G_t is a full graph for all t (which means that the player always gets to see all losses, as in the expert setting), then $\mathtt{mas}(G_t) = \alpha(G_t) = 1$, and we recover the standard guarantees for the expert setting: $\sqrt{T}$ up to logarithmic factors. In contrast, when G_t is the empty graph for all t (which means that the player only observes the loss of the action played, as in the bandit setting), then $\mathtt{mas}(G_t) = \alpha(G_t) = K$, and we recover the standard $\sqrt{KT}$ guarantees for the bandit setting, up to logarithmic factors. In between are regret bounds scaling like $\sqrt{BT}$, where B lies between 1 and K, depending on the graph structure (again, up to log-factors).

Our results are based on the algorithmic framework for the nonstochastic bandit setting introduced in [7]. In this framework, the full-information Hedge algorithm is combined with unbiased estimates of the full loss vectors in each round. The key challenge is designing an appropriate randomized scheme for choosing actions, which correctly balances exploration and exploitation or, more specifically, ensures small regret while simultaneously controlling the variance of the loss estimates. In our setting, this variance is subtly intertwined with the structure of the feedback system. For example, a key quantity emerging in the analysis of Exp3-DOM can be upper bounded in terms of the independence number of the graphs. This bound (Lemma 16 in Appendix B) is based on a combinatorial construction which may be of independent interest.

Related work. The notion of side-information in bandits has been formalized in different ways. Multivariate bandits, where the side-information takes the form of an i.i.d. sequence of random variables (and rewards are smooth functions of these variables), were initially studied in [46, 47] (in the univariate case) and further investigated in [39, 40] (in the multivariate case). A similar setting is that of contextual bandits [34], where regret is defined relative to the best policy (mapping side-information to actions) in a given set of policies; see also algorithm Exp4 in [7] for a fully nonstochastic formulation. In [42] the side-information takes the form of a sequence of elements of an arbitrary metric space. Another model in which side-information belongs to a finite, nonstructured set is investigated in [36].

Many follow-up papers have appeared since the notion of side-information used here was initially published [3, 37]. In particular, our work was improved by [33], where they introduced and analyzed Exp3-IX, a bandit algorithm using a novel "implicit exploration" technique. The algorithm can be applied in our uninformed and directed setting and can be shown to have a regret bound matching our minimax lower bounds up to logarithmic factors. In [38] Exp3-IX is modified to obtain high-probability (as opposed to expected) regret bounds which have a better dependence on the graph parameters than the bounds we show here. Further improvements and extensions were obtained in [1], where the authors strengthened and simplified the techniques in our work obtaining optimal results (up to logarithmic factors) in the uninformed directed/undirected setting. In the same paper, optimal bounds for the case of feedback systems in which the loss of the played action is not necessarily observed were also obtained. See also [21] for recent additional results. Further recent work on bandits with side-information, in both the stochastic and nonstochastic setting, include [10, 12, 31, 32, 49].

The related setting in which the loss assignment is a smooth function over the graph (but no side-information is available; hence we know that losses of neighboring actions are similar, but we do not observe them unless we play them) was studied in [44]—see also [27]. A different notion of side-information was explored in [14], where each play of an action triggered a stochastic cascade according to the independent cascading model of [29, 30]. Here the goal is to compete against the node triggering the largest expected cascade. Our side-information setting was also used to derive algorithms for related problems, such as the work [50] on online learning with costly access to attributes.

The setting of online learning with feedback graphs is also closely related to the more general setting of partial monitoring—see, e.g., [17, section 6.4], where the player's feedback is specified by a feedback matrix rather than a feedback graph. Under mild conditions on the loss values, it can be shown that the problem of learning with graph-structured feedback can be reduced to the partial monitoring setting (see [2] for a proof). Nevertheless, the analysis presented in this paper has several clear advantages over the more general analysis [8] of partial monitoring games—see [1] for a discussion.

Paper organization. In the next section, we formally define our learning protocols, introduce our main notation, and recall the combinatorial properties of graphs that we require. In section 3, we tackle the uninformed setting by introducing Exp3-SET, with upper and lower bounds on regret based on both the size of the maximal acyclic subgraph (general directed case) and the independence number (symmetric case). In section 4, we study the informed setting by analyzing two algorithms: Exp3-DOM (section 4.1), for which we prove regret bounds in expectation, and ELP.P (section 4.2), whose bounds are shown to hold not only in expectation but also with high probability. We conclude the main text with section 5, where we discuss open questions and possible directions for future research. All technical proofs are provided in the appendices. We organized such proofs based on which section of the main text the corresponding theoretical claims occur.

2. Learning protocol, notation, and preliminaries. As stated in the introduction, we consider adversarial decision problems with a finite action set $V = \{1, \dots, K\}$. At each time $t = 1, 2, \dots$, a player (the learning algorithm) picks some action $I_t \in V$ and incurs a bounded loss $\ell_{I_t,t} \in [0, 1]$. Unlike the adversarial bandit problem [7, 17], where only the played action I_t reveals its loss $\ell_{I_t,t}$, here we assume all the losses in a subset $S_{I_t,t} \subseteq V$ of actions are revealed after I_t is played. More formally, the player observes the pairs $(i, \ell_{i,t})$ for each $i \in S_{I_t,t}$. We also assume $i \in S_{i,t}$ for any i and t; that is, any action reveals its own loss when played. Note that the bandit setting ($S_{i,t} \equiv \{i\}$) and the expert setting ($S_{i,t} \equiv V$) are both special cases of this framework. We call $S_{i,t}$ the *feedback set* of action i at time t, and we write $i \xrightarrow{t} j$ when playing action i at time t reveals the loss of action j. (We sometimes write $i \to j$ when time t plays no role in the surrounding context.) With this notation, $S_{i,t} = \{j \in V : i \xrightarrow{t} j\}$. The family of feedback sets $\{S_{i,t}\}_{i \in V}$ we collectively call the *feedback system* at time t.

The adversaries we consider are nonoblivious. Namely, each loss $\ell_{i,t}$ and feedback set $S_{i,t}$ at time t can be both arbitrary functions of the past player's actions $I_1, \dots, I_{t-1}$ (note, though, that the regret is measured with respect to a fixed action assuming the adversary would have chosen the same losses, so our results do not extend to truly adaptive adversaries in the sense of [22]). The performance of a player A is measured through the expected regret

$$\max_{k \in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big],$$

where $L_{A,T} = \ell_{I_1,1} + \cdots + \ell_{I_T,T}$ and $L_{k,T} = \ell_{k,1} + \cdots + \ell_{k,T}$ are the cumulative losses of the player and of action k, respectively.[2] The expectation is taken with respect to the player's internal randomization (since losses are allowed to depend on the player's past random actions, $L_{k,T}$ may also be random). In section 3 we also consider a variant in which the feedback system is randomly generated according to a specific stochastic model. For simplicity, we focus on a finite horizon setting, where the number of rounds T is known in advance. This can be easily relaxed using a standard doubling trick.

We also consider the harder setting where the goal is to bound the actual regret

$$L_{A,T} - \max_{k \in V} L_{k,T}$$

with probability at least $1-\delta$ with respect to the player's internal randomization, and where the regret bound depends logarithmically on $1/\delta$. Clearly, a high-probability bound on the actual regret implies a similar bound on the expected regret.

Whereas some of our algorithms need to know the feedback system at the beginning of each step t, others need it only at the end of each step. We thus consider two online learning settings: the *informed* setting, where the full feedback system $\{S_{i,t}\}_{i \in V}$ selected by the adversary is made available to the learner *before* making the choice I_t; and the *uninformed setting*, where the information regarding the time-t feedback system (with the associated information about the losses) is given to the learner only after the prediction at time t.

We find it convenient at this point to adopt a graph-theoretic interpretation of feedback systems. At each step $t = 1, 2, \ldots, T$, the feedback system $\{S_{i,t}\}_{i \in V}$ defines a directed graph $G_t = (V, D_t)$, the feedback graph, where V is the set of actions and D_t is the set of arcs (i.e., ordered pairs of nodes). For $j \neq i$, arc (i, j) belongs to D_t if and only if $i \xrightarrow{t} j$ (the self-loops created by $i \xrightarrow{t} i$ are intentionally ignored). Hence, we can equivalently define $\{S_{i,t}\}_{i \in V}$ in terms of G_t (plus all self-loops). Observe that the outdegree $d^+_{i,t}$ of any $i \in V$ equals $|S_{i,t}| - 1$. Similarly, the indegree $d^-_{i,t}$ of i is the number of actions $j \neq i$ such that $i \in S_{j,t}$ (i.e., such that $j \xrightarrow{t} i$ with $j \neq i$). A notable special case of the above is when the feedback system is symmetric: $j \in S_{i,t}$ if and only if $i \in S_{j,t}$ for all i, j, and t. In words, playing i at time t reveals the loss of j if and only if playing j at time t reveals the loss of i. A symmetric feedback system for time t defines an undirected graph G_t or, more precisely, a directed graph having, for every pair of nodes $i, j \in V$, either no arcs or length-two directed cycles. Thus, from the point of view of the symmetry of the feedback system, we also distinguish between the *directed* case (G_t is a general directed graph) and the *symmetric* case (G_t is an undirected graph for all t).

The analysis of our algorithms depends on certain properties of the sequence of graphs G_t. Two graph-theoretic notions playing an important role here are those of *independent sets* and *dominating sets*. Given an undirected graph $G = (V, E)$, an independent set of G is any subset $T \subseteq V$ such that no two $i, j \in T$ are connected by an edge in E, i.e., $(i, j) \notin E$. An independent set is *maximal* if no proper superset thereof is itself an independent set. The size of any largest (and thus maximal) independent set is the *independence number* of G, denoted by $\alpha(G)$. If G is directed, we can still associate with it an independence number: we simply view G as undirected by ignoring arc orientation. If $G = (V, D)$ is a directed graph, a subset $R \subseteq V$ is a dominating set for G if for all $j \notin R$ there exists some $i \in R$ such that $(i, j) \in D$. In our bandit

[2]Although we defined the problem in terms of losses, our analysis can be applied to the case when actions return rewards $g_{i,t} \in [0, 1]$ via the transformation $\ell_{i,t} = 1 - g_{i,t}$.

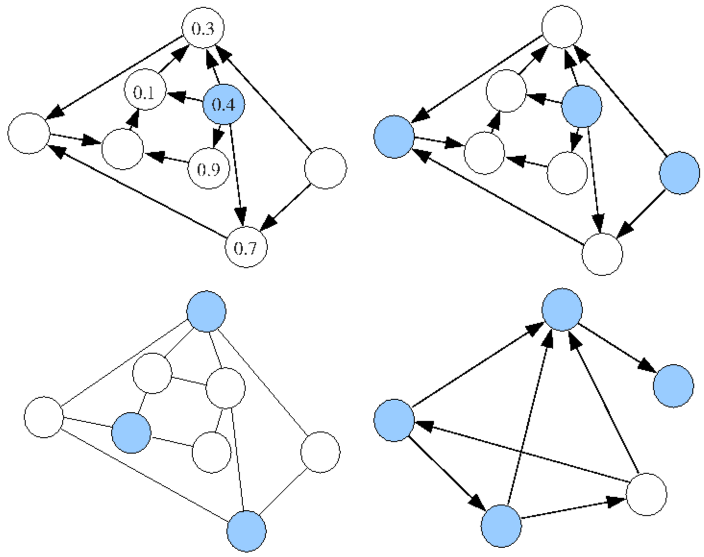

FIG. 1. *An example for some graph-theoretic concepts. Top left: A feedback graph with* $K = 8$ *actions (recall that self-loops are implicit in this representation). The light blue (shaded) action reveals its loss* 0.4*, as well as the losses of the other four actions it points to. Top right: The same graph as before, where the light blue nodes are a minimal dominating set. Recall that each action "dominates itself" through the self-loops. In this example, the rightmost action is included in any dominating set, since no other action is dominating it. Bottom left: The same graph as before where edge orientation has been removed. This gives rise to a symmetric feedback system in which the light blue nodes are a maximal independent set. Bottom right: The light blue nodes are a maximum acyclic subgraph of the depicted* 5*-action graph.*

setting, a time-t dominating set R_t is a subset of actions with the property that the loss of any remaining action in round t can be observed by playing some action in R_t. A dominating set is *minimal* if no proper subset thereof is itself a dominating set. The domination number of a directed graph G, denoted by $\gamma(G)$, is the size of any smallest (and therefore minimal) dominating set for G; see Figure 1 for examples.

Computing a minimum dominating set for an arbitrary directed graph G_t is equivalent to solving a minimum set cover problem on the associated feedback system $\{S_{i,t}\}_{i \in V}$. Although minimum set cover is NP-hard, the well-known Greedy Set Cover algorithm [20], which repeatedly selects from $\{S_{i,t}\}_{i \in V}$ the set containing the largest number of uncovered elements so far, computes a dominating set R_t such that $|R_t| \le \gamma(G_t)\,(1 + \ln K)$.

We can also lift the notion of independence number of an undirected graph to directed graphs through the notion of *maximum acyclic subgraphs*. Given a directed graph $G = (V, D)$, an acyclic subgraph of G is any graph $G' = (V', D')$ such that $V' \subseteq V$, and $D' \equiv D \cap \big(V' \times V'\big)$, with no (directed) cycles. We denote by $\mathtt{mas}(G) = |V'|$ the maximum size of such V'. Note that when G is undirected (more precisely, as above, when G is a directed graph having for every pair of nodes $i, j \in V$ either no arcs or length-two cycles), then $\mathtt{mas}(G) = \alpha(G)$; otherwise $\mathtt{mas}(G) \ge \alpha(G)$. In particular, when G is itself a directed acyclic graph, then $\mathtt{mas}(G) = |V|$. See Figure 1 (bottom right) for a simple example. Finally, we let $\mathbb{I}\{A\}$ denote the indicator function of event A.

Algorithm 1. The Exp3-SET algorithm (for the uninformed setting).

Input: $\eta \in [0,1]$.
Initialization $w_{i,1} = 1$ for all $i \in V \equiv \{1, \ldots, K\}$.
For $t = 1, 2, \ldots$:

1. Feedback system $\{S_{i,t}\}_{i \in V}$ and losses $\{\ell_{i,t}\}_{i \in V}$ are generated but not disclosed;
2. Set $p_{i,t} = \frac{w_{i,t}}{W_t}$ for each $i \in V$, where $W_t = \sum_{j \in V} w_{j,t}$;
3. Play action I_t drawn according to distribution $p_t = (p_{1,t}, \ldots, p_{K,t})$;
4. Observe:
   (a) pairs $(i, \ell_{i,t})$ for all $i \in S_{I_t,t}$;
   (b) Feedback system $\{S_{i,t}\}_{i \in V}$ is disclosed;
5. For any $i \in V$ set $w_{i,t+1} = w_{i,t} \exp\bigl(-\eta\, \widehat{\ell}_{i,t}\bigr)$, where

$$\widehat{\ell}_{i,t} = \frac{\ell_{i,t}}{q_{i,t}} \mathbb{I}\{i \in S_{I_t,t}\} \qquad \text{and} \qquad q_{i,t} = \sum_{j \,:\, j \xrightarrow{t} i} p_{j,t} \,.$$

3. The uninformed setting. In this section we investigate the setting in which the learner must select an action without any knowledge of the current feedback system. We introduce a simple general algorithm, Exp3-SET (Algorithm 1), that works in both the directed and symmetric cases. In the symmetric case, we show that the regret bound achieved by the algorithm is optimal to within logarithmic factors.

When the feedback graph G_t is a fixed clique or a fixed edgeless graph, Exp3-SET reduces to the Hedge algorithm or, respectively, to the Exp3 algorithm. Correspondingly, the regret bound for Exp3-SET yields the regret bound of Hedge and that of Exp3 as special cases.

Similar to Exp3, Exp3-SET uses importance sampling loss estimates $\widehat{\ell}_{i,t}$ that divide each observed loss $\ell_{i,t}$ by the probability $q_{i,t}$ of observing it. This probability $q_{i,t}$ is the probability of observing the loss of action i at time t; i.e., it is simply the sum of all $p_{j,t}$ (the probability of selecting action j at time t) such that $j \xrightarrow{t} i$ (recall that this sum always includes $p_{i,t}$).

In the expert setting, we have $q_{i,t} = 1$ for all i and t, and we recover the Hedge algorithm. In the bandit setting, $q_{i,t} = p_{i,t}$ for all i and t, and we recover the Exp3 algorithm (more precisely, we recover the variant Exp3Light of Exp3 that does not have an explicit exploration term; see [19] and also [43, Theorem 2.7]).

In what follows, we show that the regret of Exp3-SET can be bounded in terms of the key quantity

$$Q_t = \sum_{i \in V} \frac{p_{i,t}}{q_{i,t}} = \sum_{i \in V} \frac{p_{i,t}}{\sum_{j \,:\, j \xrightarrow{t} i} p_{j,t}} \,. \tag{1}$$

Each term $p_{i,t}/q_{i,t}$ can be viewed as the probability of drawing i from p_t conditioned on the event that $\ell_{i,t}$ was observed. A key aspect of our analysis is the ability to deterministically and nonvacuously[3] upper bound Q_t in terms of certain quantities defined on $\{S_{i,t}\}_{i \in V}$. We do so in two ways: either irrespective of how small each $p_{i,t}$ may be (this section), or depending on suitable lower bounds on the probabilities

[3]An obvious upper bound on Q_t is K, since $p_{i,t}/q_{i,t} \le 1$.

$p_{i,t}$ (section 4). In fact, forcing lower bounds on $p_{i,t}$ can be viewed as inducing the algorithm to perform exploration, and performing it in the best way (according to the analysis) requires knowing $\{S_{i,t}\}_{i\in V}$ before each prediction (hence the informed setting). The following result, whose proof is in Appendix A.2, is the building block for all subsequent results in the uninformed setting.

LEMMA 1. *The regret of Exp3-SET satisfies*

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbb{E}[Q_t]\ . \tag{2}$$

In the expert setting, $q_{i,t} = 1$ for all i and t implies $Q_t = 1$ deterministically for all t. Hence, the right-hand side of (2) becomes $(\ln K)/\eta + (\eta/2)\,T$, corresponding to the Hedge bound with a slightly larger constant in the second term; see, e.g., [17, page 72]. In the bandit setting, $q_{i,t} = p_{i,t}$ for all i and t implies $Q_t = K$ deterministically for all t. Hence, the right-hand side of (2) takes the form $(\ln K)/\eta + (\eta/2)\,KT$, equivalent to the Exp3 bound; see, e.g., [9, equation 3.4].

We now move on to the case of general feedback systems, for which we can prove the following result (proof is in Appendix A.3).

THEOREM 2. *The regret of Exp3-SET satisfies*

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbb{E}[\mathtt{mas}(G_t)]\ .$$

If $\mathtt{mas}(G_t) \le m_t$ *for* $t = 1,\ldots,T$, *then setting* $\eta = \sqrt{(2\ln K)/\sum_{t=1}^{T} m_t}$ *gives*

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \sqrt{2(\ln K)\sum_{t=1}^{T} m_t}\ .$$

As we pointed out in section 2, $\mathtt{mas}(G_t) \ge \alpha(G_t)$, with equality holding when G_t is an undirected graph. Hence, in the special case when G_t is undirected (i.e., in the symmetric setting), we obtain the following result.

COROLLARY 3. *In the symmetric setting, the regret of Exp3-SET satisfies*

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbb{E}[\alpha(G_t)]\ .$$

If $\alpha(G_t) \le \alpha_t$ *for* $t = 1,\ldots,T$, *then setting* $\eta = \sqrt{(2\ln K)/\sum_{t=1}^{T} \alpha_t}$ *gives*

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \sqrt{2(\ln K)\sum_{t=1}^{T} \alpha_t}\ .$$

Note that both Theorem 2 and Corollary 3 require the algorithm to know upper bounds on $\mathtt{mas}(G_t)$ and $\alpha(G_t)$, which may be computationally nontrivial—we return to and expand on this issue in section 4.2.

In light of Corollary 3, one may wonder whether Lemma 1 is powerful enough to allow a control of regret in terms of the independence number even in the directed case. Unfortunately, the next result shows that, in the directed case, Q_t cannot be controlled unless specific properties of p_t are assumed. More precisely, we show that even for simple directed graphs, there exist distributions p_t on the vertices such that Q_t is linear in the number of nodes while the independence number[4] is 1.

FACT 4. *Let $G = (V, D)$ be a total order on $V = \{1, \dots, K\}$, i.e., such that for all $i \in V$, arc $(j, i) \in D$ for all $j = i+1, \dots, K$. Let $p = (p_1, \dots, p_K)$ be a distribution on V such that $p_i = 2^{-i}$ for $i < K$ and $p_k = 2^{-K+1}$. Then*

$$Q = \sum_{i=1}^{K} \frac{p_i}{\sum_{j\,:\,j \to i} p_j} = \sum_{i=1}^{K} \frac{p_i}{\sum_{j=i}^{K} p_j} = \frac{K+1}{2}.$$

The simple proof can be found at the beginning of Appendix A. Next, we discuss lower bounds on the achievable regret for arbitrary algorithms. The following theorem provides a lower bound on the regret in terms of the independence number $\alpha(G)$ for a constant graph $G_t = G$ (which may be directed or undirected).

THEOREM 5. *Suppose $G_t = G$ for all t with $\alpha(G) > 1$. There exist two constants $C_1, C_2 > 0$ such that whenever $T \geq C_1 (\alpha(G))^3$, for any algorithm there exists an adversarial strategy for which the expected regret of the algorithm is at least $C_2 \sqrt{\alpha(G) T}$.*

The intuition of the proof (provided in Appendix A.4) is the following: if the graph G has $\alpha(G)$ nonadjacent vertices, then an adversary can make this problem as hard as a standard bandit problem, played on $\alpha(G)$ actions. Since for bandits on K actions there is an $\Omega(\sqrt{KT})$ lower bound on the expected regret, a variant of the proof technique leads to an $\Omega(\sqrt{\alpha(G)T})$ lower bound in our case.

One may wonder whether a sharper lower bound exists which applies to the general directed adversarial setting and involves the larger quantity $\mathtt{mas}(G)$. Unfortunately, the above measure does not seem to be related to the optimal regret: using Lemma 11 in Appendix A.5 (see proof of Theorem 6 below) one can exhibit a sequence of graphs, each having a large acyclic subgraph, on which the regret of Exp3-SET is still small.

3.1. Random feedback systems. We conclude this section with a study of Lemma 1 in a setting where the feedback system is stochastically generated via the Erdős–Renyi model. This is a standard model for random directed graphs $G = (V, D)$, where we are given a density parameter $r \in [0, 1]$ and, for any pair $i, j \in V$, arc $(i, j) \in D$ with independent probability r (self-loops, i.e., arcs (i, i), are included by default here). We have the following result.

THEOREM 6. *For $t = 1, 2, \dots,$ let G_t be an independent draw from the Erdős–Renyi model with fixed parameter $r \in [0, 1]$. Then the regret of Exp3-SET satisfies*

$$\max_{k \in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \leq \frac{\ln K}{\eta} + \frac{\eta T}{2r}\Big(1 - (1-r)^K\Big)\ .$$

In the above, expectations are computed with respect to both the algorithm's randomization and the random generation of G_t occurring at each round. In particular, setting $\eta = \sqrt{\frac{2r \ln K}{T(1-(1-r)^K)}}$ gives

$$\max_{k \in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \leq \sqrt{\frac{2(\ln K)T\big(1-(1-r)^K\big)}{r}}\ . \tag{3}$$

[4]In this specific example, the maximum acyclic subgraph has size K, which confirms the looseness of Theorem 2.

Note that as r ranges in $[0,1]$ we interpolate between the multi-armed bandit[5] ($r=0$) and the expert ($r=1$) regret bounds.

Finally, it is worth noticing that standard results from the theory of Erdős–Renyi graphs—at least in the symmetric case (see, e.g., [26, 25])—show that when $Kr \to \infty$ as $K \to \infty$, the independence number α of the resulting graph satisfies

$$\alpha \sim \frac{2\log(Kr)}{\log(\frac{1}{1-r})}$$

with high probability, as $K \to \infty$. In particular, when r is such that $0 < c \le Kr = o(K)$ as $K \to \infty$, for some constant c, then $\alpha = \Omega(\frac{1}{r})$ with high probability as $K \to \infty$. This fact, combined with the lower bound in Theorem 5, gives a lower bound on the regret of the order $\sqrt{T/r}$. It is then easy to see that this lower bound matches (up to logarithmic factors) the upper bound of Theorem 6, since the right-hand side of (3) is upper bounded anyway by $\sqrt{\frac{2(\ln K)T}{r}}$.

4. The informed setting. The lack of a lower bound matching the upper bound provided by Theorem 2 is a good indication that something more sophisticated has to be done in order to upper bound the key quantity Q_t defined in (1). This leads us to consider more refined ways of allocating probabilities $p_{i,t}$ to nodes. We do so by taking advantage of the informed setting, in which the learner can access G_t before selecting the action I_t. The algorithm Exp3-DOM, introduced in this section, exploits the knowledge of G_t in order to achieve an optimal (up to logarithmic factors) regret bound.

Recall the problem uncovered by Fact 4: when the graph induced by the feedback system is directed, Q_t cannot be upper bounded, in a nonvacuous way, independent of the choice of probabilities $p_{i,t}$. The new algorithm Exp3-DOM controls these probabilities by adding an exploration term to the distribution p_t. This exploration term is supported on a dominating set of the current graph G_t, and computing such a dominating set before selection of the action at time t can only be done in the informed setting. Intuitively, exploration on a dominating set allows us to control Q_t by increasing the probability $q_{i,t}$ that each action i is observed. If the dominating set is also minimal, then the variance caused by exploration can be bounded in terms of the independence number (and additional logarithmic factors) just like the undirected case. In order to optimize the regret bound, the exploration rate must be tuned according to the size of the dominating set in G_t, which is possibly changing at every round. We take this into account by having Exp3-DOM run a logarithmic (in K) number of instances of Exp3, each tuned to a geometrically increasing value of the dominating set size. Each instance is also internally running a doubling trick in order to further tune its local exploration rate to the specific set of rounds t which are managed by that instance (based on the size of the dominating set in G_t).

Finally, when proving high-probability results on the regret in the informed setting (algorithm ELP.P in section 4.2), we also assume the feedback system is known beforehand.

4.1. Bounds in expectation: The Exp3-DOM algorithm. The Exp3-DOM algorithm (Algorithm 2) for the informed setting runs $\mathcal{O}(\log K)$ variants of Exp3 (with explicit exploration) indexed by $b = 0, 1, \ldots, \lfloor \log_2 K \rfloor$. At time t the algorithm is given the current feedback system $\{S_{i,t}\}_{i \in V}$ and computes a dominating set R_t of the

[5]Observe that $\lim_{r \to 0^+} \frac{1-(1-r)^K}{r} = K$.

Algorithm 2. The Exp3-DOM algorithm (for the informed setting).

Input: Exploration parameters $\nu^{(b)} \in (0,1]$ for $b \in \{0,1,\dots,\lfloor \log_2 K \rfloor\}$
Initialization: $w_{i,1}^{(b)} = 1$ for all $i \in V \equiv \{1,\dots,K\}$ and $b \in \{0,1,\dots,\lfloor \log_2 K \rfloor\}$
For $t = 1,2,\dots$:

1. Feedback system $\{S_{i,t}\}_{i\in V}$ is generated *and disclosed* (losses $\{\ell_{i,t}\}_{i\in V}$ are generated and not disclosed);
2. Compute a dominating set R_t for G_t associated with $\{S_{i,t}\}_{i\in V}$;
3. Let b_t be such that $|R_t| \in [2^{b_t}, 2^{b_t+1}-1]$;
4. Set $W_t^{(b_t)} = \sum_{i\in V} w_{i,t}^{(b_t)}$;
5. Set $p_{i,t}^{(b_t)} = \left(1-\nu^{(b_t)}\right)\frac{w_{i,t}^{(b_t)}}{W_t^{(b_t)}} + \frac{\nu^{(b_t)}}{|R_t|}\mathbb{I}\{i \in R_t\}$;
6. Play action I_t drawn from distribution $p_t^{(b_t)} = \left(p_{1,t}^{(b_t)},\dots,p_{K,t}^{(b_t)}\right)$;
7. Observe pairs $(i,\ell_{i,t})$ for all $i \in S_{I_t,t}$;
8. For any $i \in V$ set $w_{i,t+1}^{(b_t)} = w_{i,t}^{(b_t)} \exp\left(-\nu^{(b_t)}\,\widehat{\ell}_{i,t}^{(b_t)}/2^{b_t}\right)$, where

$$\widehat{\ell}_{i,t}^{(b_t)} = \frac{\ell_{i,t}}{q_{i,t}^{(b_t)}}\,\mathbb{I}\{i \in S_{I_t,t}\} \qquad \text{and} \qquad q_{i,t}^{(b_t)} = \sum_{j\,:\,j \xrightarrow{t} i} p_{j,t}^{(b_t)}.$$

directed graph G_t induced by $\{S_{i,t}\}_{i\in V}$. Based on the size $|R_t|$ of R_t, the algorithm uses instance $b_t = \lfloor \log_2 |R_t| \rfloor$ to draw action I_t. We use a superscript b to denote the quantities relevant to the variant of Exp3 indexed by b. Similarly to the analysis of Exp3-SET, the key quantities are

$$q_{i,t}^{(b)} = \sum_{j\,:\,i\in S_{j,t}} p_{j,t}^{(b)} = \sum_{j\,:\,j\xrightarrow{t} i} p_{j,t}^{(b)} \qquad \text{and} \qquad Q_t^{(b)} = \sum_{i\in V} \frac{p_{i,t}^{(b)}}{q_{i,t}^{(b)}}\,, \quad b = 0,1,\dots,\lfloor \log_2 K \rfloor.$$

Let $T^{(b)} \equiv \left\{t = 1,\dots,T \,:\, |R_t| \in [2^b, 2^{b+1}-1]\right\}$. Clearly, the sets $T^{(b)}$ are a partition of the time steps $\{1,\dots,T\}$, so that $\sum_b |T^{(b)}| = T$. Since the adversary adaptively chooses the dominating sets R_t (through the adaptive choice of the feedback system at time t), the sets $T^{(b)}$ are random variables. This causes a problem in tuning the parameters $\nu^{(b)}$, the exploration parameters of Exp3-DOM. For this reason, we do not prove a regret bound directly for Exp3-DOM, where each instance uses a fixed $\nu^{(b)}$, but for a slight variant of it, where each $\nu^{(b)}$ is separately set through a doubling trick. In fact, a good choice of $\nu^{(b)}$ depends on the *unknown* random quantity

$$\overline{Q}^{(b)} = \sum_{t\in T^{(b)}} \overline{Q}_t^{(b)},$$

where $\overline{Q}_t^{(b)} = 1 + \frac{Q_t^{(b)}}{2^{b+1}}$. To overcome this problem, we slightly modify Exp3-DOM by applying a doubling trick to guess $\overline{Q}^{(b)}$ for each b. Specifically, for each $b = 0,1,\dots,\lfloor \log_2 K \rfloor$, we use a sequence $\nu_r^{(b)} = \sqrt{(2^b \ln K)/2^r}$ for $r = 0,1,\dots$. We initially run the algorithm with $\nu_0^{(b)}$. Whenever the algorithm is running with $\nu_r^{(b)}$ and observes that $\sum_s \overline{Q}_s^{(b)} > 2^r$, where[6] the sum is over all s so far in $T^{(b)}$, then we

[6]Notice that $\sum_s \overline{Q}_s^{(b)}$ is an observable quantity.

restart the algorithm with $\nu_{r+1}^{(b)}$.

LEMMA 7. *In the directed case, the regret of Exp3-DOM satisfies*

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \sum_{b=0}^{\lfloor \log_2 K \rfloor} \left(\frac{2^b \ln K}{\nu^{(b)}} + \nu^{(b)} \mathbb{E}\left[\sum_{t\in T^{(b)}} \left(1 + \frac{Q_t^{(b)}}{2^{b+1}} \right) \right] \right). \tag{4}$$

Moreover, if we use the above doubling trick to choose $\nu^{(b)}$ for each $b = 0, \ldots, \lfloor \log_2 K \rfloor$, then

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] = \mathcal{O}\left((\ln K)\, \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \left(|R_t| + Q_t^{(b_t)} \right)} \right] + (\ln K) \ln(KT) \right). \tag{5}$$

Importantly, the next result (proof in Appendix B.2) shows how bound (5) of Lemma 7 can be expressed in terms of the sequence $\alpha(G_t)$ of independence numbers of graphs G_t whenever the Greedy Set Cover algorithm [20] (see section 2) is used to compute the dominating set R_t of the feedback system at time t.

THEOREM 8. *If Step 2 of Exp3-DOM uses the Greedy Set Cover algorithm to compute the dominating sets R_t, then the regret of Exp-DOM using the doubling trick satisfies*

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] = \mathcal{O}\left(\ln(K) \sqrt{\ln(KT) \sum_{t=1}^{T} \mathbb{E}[\alpha(G_t)]} + \ln(K) \ln(KT) \right).$$

Combining the upper bound of Theorem 8 with the lower bound of Theorem 5, we see that the attainable expected regret in the informed setting is characterized by the independence numbers of the graphs. Moreover, a quick comparison of Corollary 3 and Theorem 8 reveals that a symmetric feedback system overcomes the advantage of working in an informed setting: The bound we obtained for the uninformed symmetric setting (Corollary 3) is sharper by logarithmic factors than the one we derived for the informed—but more general, i.e., directed—setting (Theorem 8).

4.2. High-probability bounds: The ELP.P algorithm. We now turn to present an algorithm working in the informed setting for which we can also prove high-probability regret bounds.[7] We call this algorithm ELP.P (Exponentially weighted algorithm with Linear Programming, with high Probability). Like Exp3-DOM, the exploration component is not uniform over the actions, but is chosen carefully to reflect the graph structure at each round. In fact, the optimal choice of the exploration for ELP.P requires us to solve a simple linear program—hence the name of the algorithm.[8] The pseudocode appears as Algorithm 3. Note that unlike the previous algorithms, this algorithm utilizes the "rewards" formulation of the problem; i.e., instead of using the losses $\ell_{i,t}$ directly, it uses the rewards $g_{i,t} = 1-\ell_{i,t}$ and boosts the weight of actions for which $g_{i,t}$ is estimated to be large, as opposed to decreasing the weight of actions for which $\ell_{i,t}$ is estimated to be large. This is done merely for technical convenience and does not affect the complexity of the algorithm or the regret guarantee. The form of

[7]We have been unable to prove high-probability bounds for Exp3-DOM or variants of it.

[8]We note that this algorithm improves over the basic ELP algorithm initially presented in [37], in that its regret is bounded with high probability and not just in expectation, and applies in the directed case as well as the symmetric case.

Algorithm 3. The ELP.P algorithm (for the informed setting).

Input: Confidence parameter $\delta \in (0,1)$, learning rate $\eta > 0$;
Initialization: $w_{i,1} = 1$ for all $i \in V = \{1, \dots, K\}$;
For $t = 1, 2, \dots$:

1. Feedback system $\{S_{i,t}\}_{i \in V}$ is generated *and disclosed* (losses $\{\ell_{i,t}\}_{i \in V}$ are generated and not disclosed);
2. Let Δ_K be the K-dimensional probability simplex and $s_t = (s_{1,t}, \dots, s_{K,t})$ be a solution to the linear program
$$\max_{(s_1,\dots,s_K) \in \Delta_K} \quad \min_{i \in V} \sum_{j \,:\, j \xrightarrow{t} i} s_j;$$
3. Set $p_{i,t} := (1-\nu_t)\frac{w_{i,t}}{W_t} + \nu_t s_{i,t}$ where $W_t = \sum_{i \in V} w_{i,t}$,
$$\nu_t = \frac{(1+\beta)\,\eta}{\min_{i \in V} \sum_{j \,:\, j \xrightarrow{t} i} s_{j,t}}, \qquad \text{and} \qquad \beta = 2\eta\sqrt{\frac{\ln(5K/\delta)}{\ln K}};$$
4. Play action I_t drawn according to distribution $p_t = \big(p_{1,t}, \dots, p_{K,t}\big)$;
5. Observe pairs $(i, \ell_{i,t})$ for all $i \in S_{I_t,t}$;
6. For any $i \in V$ set $g_{i,t} = 1 - \ell_{i,t}$ and $w_{i,t+1} = w_{i,t}\, \exp\big(\eta\, \widehat{g}_{i,t}\big)$, where
$$\widehat{g}_{i,t} = \frac{g_{i,t}\mathbb{I}\{i \in S_{I_t,t}\} + \beta}{q_{i,t}} \qquad \text{and} \qquad q_{i,t} = \sum_{j \,:\, j \xrightarrow{t} i} p_{j,t}.$$

the reward estimator is similar to the one used in Exp3.P [7] (see also [17, section 6.8]). Different estimators can also be used; see, for example, [6] and the recent work [38], where an improved high-probability regret bound depending on $\alpha(G_t)$ rather than on $\mathtt{mas}(G_t)$ has been shown. This recent improvement also brings the computational advantage of avoiding solving linear programs when performing exploration.

THEOREM 9. *Let algorithm ELP.P be run with learning rate $\eta \le 1/(3K)$ sufficiently small such that $\beta \le 1/4$. Then, with probability at least $1-\delta$ we have*
$$\begin{aligned} L_{A,T} - \max_{k \in V} L_{k,T} \ \le \ & \sqrt{5 \ln\left(\frac{5}{\delta}\right) \sum_{t=1}^{T} \mathtt{mas}(G_t)} + \frac{2\ln(5K/\delta)}{\eta} \\ & + 12\eta \sqrt{\frac{\ln(5K/\delta)}{\ln K}} \sum_{t=1}^{T} \mathtt{mas}(G_t) \\ & + \widetilde{\mathcal{O}}\left(1 + \sqrt{T\eta} + T\eta^2\right)\left(\max_{t=1\dots T} \mathtt{mas}^2(G_t)\right), \end{aligned}$$
where the $\widetilde{\mathcal{O}}$ notation hides only numerical constants and factors logarithmic in K and $1/\delta$. In particular, if for constants $m_1, \dots, m_T$ we have $\mathtt{mas}(G_t) \le m_t$, $t = 1, \dots, T$, and we pick $\eta \le 1/(3K)$ such that
$$\eta^2 = \frac{1}{6}\,\frac{\sqrt{\ln(5K/\delta)\,(\ln K)}}{\sum_{t=1}^{T} m_t},$$

then we get with probability at least $1-\delta$ *that*

$$L_{A,T} - \max_{k \in V} L_{k,T} \;\le\; 10\,\frac{\ln^{1/4}(5K/\delta)}{\ln^{1/4} K}\sqrt{\ln\left(\frac{5K}{\delta}\right)\sum_{t=1}^{T} m_t} + \widetilde{\mathcal{O}}(T^{1/4})\left(\max_{t=1\ldots T}\mathtt{mas}^2(G_t)\right).$$

This theorem essentially tells us that the regret of the ELP.P algorithm, up to second-order factors, is quantified by $\sqrt{\sum_{t=1}^{T}\mathtt{mas}(G_t)}$. Recall that, in the special case when G_t is symmetric, we have $\mathtt{mas}(G_t) = \alpha(G_t)$.

One computational issue to bear in mind is that this theorem (as well as Theorem 2 and Corollary 3) holds under an optimal choice of η. In turn, this value depends on upper bounds on $\sum_{t=1}^{T}\mathtt{mas}(G_t)$ (or on $\sum_{t=1}^{T}\alpha(G_t)$ in the symmetric case). Unfortunately, in the worst case, computing the maximal acyclic subgraph or the independence number of a given graph is NP-hard, so implementing such algorithms is not *always* computationally tractable.[9] However, it is easy to see that the algorithm is robust to approximate computation of this value; misspecifying the average independence number $\frac{1}{T}\sum_{t=1}^{T}\alpha(G_t)$ by a factor of v entails an additional $\sqrt{v}$ factor in the bound. Thus, one might use standard heuristics resulting in a reasonable approximation of the independence number. Although computing the independence number is also NP-hard to approximate, it is unlikely for intricate graphs with hard-to-approximate independence numbers to appear in relevant applications. Moreover, by setting the approximation to be either K or 1, we trivially obtain an approximation factor of at most either K or $\frac{1}{T}\sum_{t=1}^{T}\alpha(G_t)$. The former leads to an $\widetilde{\mathcal{O}}(\sqrt{KT})$ regret bound similar to the standard bandits setting, while the latter leads to an $\widetilde{\mathcal{O}}(\frac{1}{T}\sum_{t=1}^{T}\alpha(G_T)\sqrt{T})$ regret bound, which is better than the regret for the bandits setting if the average independence number is less than $\sqrt{K}$. In contrast, this computational issue does not show up in Exp3-DOM, whose tuning relies only on efficiently computable quantities.

5. Conclusions and open questions. In this paper we investigated online prediction problems in partial-information regimes that interpolate between the classical bandit and expert settings. We provided algorithms, as well as upper and lower bounds on the attainable regret, with a nontrivial dependence on the information feedback structure. In particular, we have shown a number of results characterizing prediction performance in terms of the structure of the feedback system, the amount of information available before prediction, and the nature (adversarial or fully random) of the process generating the feedback system.

There are many open questions that warrant further study, some of which are briefly mentioned below:

1. It would be interesting to study adaptations of our results to the case when the feedback system $\{S_{i,t}\}_{i\in V}$ may depend on the loss $\ell_{I_t,t}$ of player's action I_t. Note that this would prevent a direct construction of an unbiased estimator for unobserved losses, which many worst-case bandit algorithms (including ours—see the appendix) hinge upon.
2. Even in the uninformed setting, our algorithms need to observe the feedback system at the end of each round (more precisely, Exp3-SET needs to know the feedback graph up to and including the second neighborhood of the chosen action). An interesting question would be to investigate the best possible

[9] [37] proposed a generic mechanism to circumvent this, but the justification has a flaw which is not clear how to fix.

regret rates when only the neighborhood of the chosen action is revealed after each round. An answer to this question was recently given in [21], where they showed that no improvements over the standard bandit bound are possible in this case.

3. The upper bound contained in Theorem 2, expressed in terms of $\mathtt{mas}(\cdot)$, is almost certainly suboptimal, even in the uninformed setting, and it would be nice to see if more adequate graph complexity measures can be used instead.
4. Our lower bound in Theorem 5 refers to a constant graph sequence. We would like to provide a more complete characterization applying to sequences of adversarially generated graphs $G_1, G_2, \ldots, G_T$ in terms of sequences of their corresponding independence numbers $\alpha(G_1), \alpha(G_2), \ldots, \alpha(G_T)$ (or variants thereof) in both the uninformed and the informed settings. Moreover, the adversary strategy achieving our lower bound is computationally hard to implement in the worst case (the adversary needs to identify the largest independent set in a given graph). What is the achievable regret if the adversary is assumed to be computationally bounded?
5. The information feedback models we used are natural and simple. They assume that the action at a given time period only affects rewards and observations for that period. In some settings, the reward observation may be delayed. In such settings, the action taken at a given stage may affect what is observed in subsequent stages. We leave the issue of modeling and analyzing such settings to future work.
6. Finally, we would like to see what the achievable performance is in the special case of stochastic rewards, which are assumed to be drawn i.i.d. from some unknown distributions. This was recently considered in [13], with results depending on the graph clique structure. However, the tightness of these results remains to be ascertained.

Appendix A. Technical lemmas and proofs from section 3. This appendix contains the proofs of all technical results occurring in section 3, along with ancillary graph-theoretic lemmas. Throughout this appendix, $\mathbb{E}_t[\cdot]$ is a shorthand for $\mathbb{E}\big[\cdot \mid \{\ell_{i,r}\}_{i\in V}, \{S_{i,r}\}_{i\in V}, r = 1,\ldots,t,\, I_1,\ldots,I_{t-1}\big]$. Also, for ease of exposition, we implicitly first condition on the history, i.e., $\{\ell_{i,r}\}_{i\in V}, \{S_{i,r}\}_{i\in V}, r = 1,\ldots,t,\, I_1,\ldots,I_{t-1}$, and later take an expectation with respect to that history. This implies that, given that conditioning, we can treat random variables such as $p_{i,t}$, $\ell_{i,t}$, and $S_{i,t}$ as constants, and we can later take an expectation over history so as to remove the conditioning.

A.1. Proof of Fact 4. Using standard properties of geometric sums, one can immediately see that

$$\sum_{i=1}^{K} \frac{p_i}{\sum_{j=i}^{K} p_j} = \sum_{i=1}^{K-1} \frac{2^{-i}}{2^{-i+1}} + \frac{2^{-K+1}}{2^{-K+1}} = \frac{K-1}{2} + 1 = \frac{K+1}{2};$$

hence the claimed result.

A.2. Proof of Lemma 1. Following the proof of Exp3 [7], we have

$$
\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i \in V} \frac{w_{i,t+1}}{W_t} \\
&= \sum_{i \in V} \frac{w_{i,t} \exp(-\eta\, \widehat{\ell}_{i,t})}{W_t} \\
&= \sum_{i \in V} p_{i,t} \exp(-\eta\, \widehat{\ell}_{i,t}) \\
&\le \sum_{i \in V} p_{i,t} \left(1 - \eta \widehat{\ell}_{i,t} + \frac{1}{2}\, \eta^2 (\widehat{\ell}_{i,t})^2\right) \quad \text{using } e^{-x} \le 1 - x + x^2/2 \text{ for all } x \ge 0 \\
&= 1 - \eta \sum_{i \in V} p_{i,t} \widehat{\ell}_{i,t} + \frac{\eta^2}{2} \sum_{i \in V} p_{i,t} (\widehat{\ell}_{i,t})^2.
\end{aligned}
$$

Taking logs, using $\ln(1+x) \le x$ for all $x > -1$, and summing over $t = 1, \ldots, T$ yields

$$
\ln \frac{W_{T+1}}{W_1} \le -\eta \sum_{t=1}^{T} \sum_{i \in V} p_{i,t} \widehat{\ell}_{i,t} + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i \in V} p_{i,t} (\widehat{\ell}_{i,t})^2.
$$

Moreover, for any fixed comparison action k, we also have

$$
\ln \frac{W_{T+1}}{W_1} \ge \ln \frac{w_{k,T+1}}{W_1} = -\eta \sum_{t=1}^{T} \widehat{\ell}_{k,t} - \ln K.
$$

Putting this together and rearranging gives

$$
\sum_{t=1}^{T} \sum_{i \in V} p_{i,t} \widehat{\ell}_{i,t} \le \sum_{t=1}^{T} \widehat{\ell}_{k,t} + \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i \in V} p_{i,t} (\widehat{\ell}_{i,t})^2. \tag{6}
$$

Note that, for all $i \in V$,

$$
\mathbb{E}_t[\widehat{\ell}_{i,t}] = \sum_{j \,:\, i \in S_{j,t}} p_{j,t} \frac{\ell_{i,t}}{q_{i,t}} = \sum_{j \,:\, j \xrightarrow{t} i} p_{j,t} \frac{\ell_{i,t}}{q_{i,t}} = \frac{\ell_{i,t}}{q_{i,t}} \sum_{j \,:\, j \xrightarrow{t} i} p_{j,t} = \ell_{i,t}.
$$

Moreover,

$$
\mathbb{E}_t\big[(\widehat{\ell}_{i,t})^2\big] = \sum_{j \,:\, i \in S_{j,t}} p_{j,t} \frac{\ell_{i,t}^2}{q_{i,t}^2} = \frac{\ell_{i,t}^2}{q_{i,t}^2} \sum_{j \,:\, j \xrightarrow{t} i} p_{j,t} \le \frac{1}{q_{i,t}^2} \sum_{j \,:\, j \xrightarrow{t} i} p_{j,t} = \frac{1}{q_{i,t}}.
$$

Hence, taking expectations on both sides of (6), and recalling the definition of Q_t, we get

$$
\mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}[Q_t] \tag{7}
$$

as claimed.

A.3. Proof of Theorem 2. We first need the following lemma.

LEMMA 10. *Let $G = (V, D)$ be a directed graph with vertex set $V = \{1, \ldots, K\}$ and arc set D. Then, for any distribution p over V we have*

$$\sum_{i=1}^{K} \frac{p_i}{\sum_{j\,:\,j \to i} p_j} \le \mathtt{mas}(G).$$

Proof. Let $N^-_{i,0}$ be the in-neighbors of node i in the initial graph G, i.e., the set of nodes j such that $(j, i) \in D$. We show that there is a subset of vertices V' such that the induced graph is acyclic and $|V'| \ge \sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j \in N^-_{i,0}} p_j}$. This is proven by adding elements to an initially empty set V'. Let

$$\Phi_0 = \sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j \in N^-_{i,0}} p_j},$$

and let i_1 be the vertex which minimizes the denominator $p_i + \sum_{j \in N^-_{i,0}} p_j$ over $i \in V$. We now delete i_1 from the graph, along with all its incoming neighbors (set $N^-_{i_1,0}$) and all edges which are incident (both departing and incoming) to these nodes, and then iterate on the remaining graph. The contribution of all the deleted vertices to Φ_0 is

$$\sum_{r \in N^-_{i_1,0} \cup \{i_1\}} \frac{p_r}{p_r + \sum_{j \in N^-_{r,0}} p_j} \le \sum_{r \in N^-_{i_1,0} \cup \{i_1\}} \frac{p_r}{p_{i_1} + \sum_{j \in N^-_{i_1,0}} p_j} = 1,$$

where the inequality follows from the minimality of i_1.

Let $V' \leftarrow V' \cup \{i_1\}$, $V_1 = V \setminus (N^-_{i_1,0} \cup \{i_1\})$, and let $N^-_{i,1}$ be the in-neighbors of node i in the graph after the first step. Then, defining

$$\Phi_1 = \sum_{i \in V_1} \frac{p_i}{p_i + \sum_{j \in N^-_{i,1}} p_j},$$

we can write

$$\Phi_1 \ge \sum_{i \in V_1} \frac{p_i}{p_i + \sum_{j \in N^-_{i,0}} p_j} \ge \Phi_0 - 1.$$

We apply the very same argument to Φ_1 with node i_2 (minimizing $p_i + \sum_{j \in N^-_{i,1}} p_j$ over $i \in V_1$), to Φ_2 with node $i_3, \ldots,$ to Φ_{s-1} with node i_s, up until $\Phi_s = 0$, i.e., until no nodes are left in the reduced graph. This gives $\Phi_0 \le s = |V'|$, where $V' = \{i_1, i_2, \ldots, i_s\}$. Moreover, since in each step $r = 1, \ldots, s$ we remove all remaining arcs incoming to i_r, the graph induced by set V' cannot contain cycles. ∎

The claim of Theorem 2 follows from a direct combination of Lemma 1 with Lemma 10.

A.4. Proof of Theorem 5. The proof uses a variant of the standard multi-armed bandit lower bound [17]. The intuition is that when we have $\alpha(G)$ nonadjacent nodes, the problem reduces to an instance of the standard multi-armed bandit (where information beyond the loss of the action chosen is observed) on $\alpha(G)$ actions.

By Yao's minimax principle, in order to establish the lower bound, it is enough to demonstrate some probabilistic adversary strategy, on which the expected regret of any *deterministic* algorithm A is bounded from below by $C\sqrt{\alpha(G)T}$ for some constant C.

Specifically, suppose without loss of generality that we number the nodes in some largest independent set of G by $1, 2, \ldots, \alpha(G)$, and all the other nodes in the graph by $\alpha(G)+1, \ldots, |V|$. Let ϵ be a parameter to be determined later, and consider the following joint distribution over stochastic loss sequences:

- Let Z be uniformly distributed on $1, 2, \ldots, \alpha(G)$.
- Conditioned on $Z = i$, each loss $\ell_{j,t}$ is independent Bernoulli with parameter $1/2$ if $j \neq i$ and $j < \alpha(G)$, is independent Bernoulli with parameter $1/2 - \epsilon$ if $j = i$, and is 1 with probability 1 otherwise.

For each $i = 1, \ldots, \alpha(G)$, let T_i be the number of times the node i was chosen by the algorithm after T rounds. Also, let T_Δ denote the number of times some node whose index is larger than $\alpha(G)$ is chosen after T rounds. Finally, let $\mathbb{E}_i$ denote expectation conditioned on $Z = i$, and let $\mathbb{P}_i$ denote the probability over loss sequences conditioned on $Z = i$. We have

$$\begin{aligned}
&\max_{k\in V}\mathbb{E}[L_{A,T} - L_{k,T}] \\
&\quad = \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} \mathbb{E}_i\left[L_{A,T} - \left(\frac{1}{2}-\epsilon\right)T\right] \\
&\quad = \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} \mathbb{E}_i\left[\sum_{j\in\{1\ldots\alpha(G)\}\setminus i}\frac{1}{2}T_j + \left(\frac{1}{2}-\epsilon\right)T_i + T_\Delta - \left(\frac{1}{2}-\epsilon\right)T\right] \\
&\quad = \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} \mathbb{E}_i\left[\frac{1}{2}\sum_{j=1}^{\alpha(G)}T_j + \frac{1}{2}T_\Delta + \frac{1}{2}T_\Delta - \epsilon T_i - \left(\frac{1}{2}-\epsilon\right)T\right].
\end{aligned}$$

Since $\sum_{j=1}^{\alpha(G)} T_j + T_\Delta = T$, this expression equals

$$\text{(8)} \qquad \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} \mathbb{E}_i\left[\frac{1}{2}T_\Delta + \epsilon(T - T_i)\right] \;\geq\; \epsilon\left(T - \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)}\mathbb{E}_i[T_i]\right).$$

Now, consider another distribution $\mathbb{P}_0$ over the loss sequence, which corresponds to the distribution above but with $\epsilon = 0$ (namely, all nodes $1, \ldots, \alpha(G)$ have losses which are ± 1 independently and with equal probability, and all nodes whose index is larger than $\alpha(G)$ have losses of 1), and denote by $\mathbb{E}_0$ the corresponding expectation. We upper bound the difference between $\mathbb{E}_i[T_i]$ and $\mathbb{E}_0[T_i]$, using information-theoretic arguments. Let λ_t be the collection of loss values observed at round t, and $\lambda^t = (\lambda_1, \ldots, \lambda_t)$. Note that since the algorithm is deterministic, λ^{t-1} determines the algorithm's choice of action I_t at each round t, and hence T_i is determined by λ^T, and thus $\mathbb{E}_0[T_i \mid \lambda^T] = \mathbb{E}_i[T_i \mid \lambda^T]$. We have

$$\begin{aligned}\mathbb{E}_i[T_i] - \mathbb{E}_0[T_i] &= \sum_{\lambda^T} \mathbb{P}_i(\lambda^T)\mathbb{E}_i[T_i \mid \lambda^T] - \sum_{\lambda^T} \mathbb{P}_0(\lambda^T)\mathbb{E}_0[T_i \mid \lambda^T] \\ &= \sum_{\lambda^T} \mathbb{P}_i(\lambda^T)\mathbb{E}_i[T_i \mid \lambda^T] - \sum_{\lambda^T} \mathbb{P}_0(\lambda^T)\mathbb{E}_i[T_i|\lambda^T] \\ &\leq \sum_{\lambda^T : \mathbb{P}_i(\lambda^T) > \mathbb{P}_0(\lambda^T)} \big(\mathbb{P}_i(\lambda^T) - \mathbb{P}_0(\lambda^T)\big)\, \mathbb{E}_i[T_i \mid \lambda^T] \\ &\leq T \sum_{\lambda^T : \mathbb{P}_i(\lambda^T) > \mathbb{P}_0(\lambda^T)} \big(\mathbb{P}_i(\lambda^T) - \mathbb{P}_0(\lambda^T)\big)\ .\end{aligned}$$

Using Pinsker's inequality, this is at most

$$T\sqrt{\frac{1}{2} D_{\mathrm{kl}}(\mathbb{P}_0(\lambda^T) \,\|\, \mathbb{P}_i(\lambda^T))},$$

where D_{kl} is the Kullback–Leibler divergence (or relative entropy) between the distributions $\mathbb{P}_i$ and $\mathbb{P}_0$. Using the chain rule for relative entropy, this equals

$$T\sqrt{\frac{1}{2}\sum_{t=1}^{T}\sum_{\lambda^{t-1}} \mathbb{P}_0(\lambda^{t-1}) D_{\mathrm{kl}}\big(\mathbb{P}_0(\lambda_t|\lambda^{t-1}) \,\|\, \mathbb{P}_i(\lambda_t|\lambda^{t-1})\big)}.$$

Let us consider any single relative entropy term above. Recall that λ^{t-1} determines the node I_t picked at round t. If this node is not i or adjacent to i, then λ_t is going to have the same distribution under both $\mathbb{P}_i$ and $\mathbb{P}_0$, and the relative entropy is zero. Otherwise, the coordinate of λ_t corresponding to node i (and that coordinate only) will have a different distribution: Bernoulli with parameter $\frac{1}{2} - \epsilon$ under $\mathbb{P}_i$, and Bernoulli with parameter $\frac{1}{2}$ under $\mathbb{P}_0$. The relative entropy term in this case is easily shown to be $-\frac{1}{2}\log(1-4\epsilon^2) \leq 8\log(4/3)\,\epsilon^2$. Therefore, letting S_{I_t} denote the feedback set at time t, we can upper bound the above by

$$\begin{aligned} T\sqrt{\frac{1}{2}\sum_{t=1}^{T} \mathbb{P}_0(i \in S_{I_t})(8\log(4/3)\epsilon^2)} &= 2T\epsilon\sqrt{\log\left(\frac{4}{3}\right)\mathbb{E}_0\big[|\{t : i \in S_{I_t}\}|\big]} \\ &\leq 2T\epsilon\sqrt{\log\left(\frac{4}{3}\right)\mathbb{E}_0\left[T_i + T_\Delta\right]}. \end{aligned} \tag{9}$$

We now claim that we can assume $\mathbb{E}_0[T_\Delta] \leq 0.08\sqrt{\alpha(G)T}$. To see why, note that if

$\mathbb{E}_0[T_\Delta] > 0.08\sqrt{\alpha(G)T}$, then the expected regret under $\mathbb{E}_0$ would have been at least

$$\begin{aligned}
\max_{k\in V} \mathbb{E}_0[L_{A,T} - L_{k,T}] &= \mathbb{E}_0\left[T_\Delta + \frac{1}{2}\sum_{j=1}^{\alpha(G)} T_j\right] - \frac{1}{2}T \\
&= \mathbb{E}_0\left[\frac{1}{2}T_\Delta + \frac{1}{2}\left(T_\Delta + \sum_{j=1}^{\alpha(G)} T_j\right)\right] - \frac{1}{2}T \\
&= \mathbb{E}_0\left[\frac{1}{2}T_\Delta + \frac{1}{2}T\right] - \frac{1}{2}T \\
&= \frac{1}{2}\mathbb{E}_0[T_\Delta] \\
&> 0.04\sqrt{\alpha(G)T}.
\end{aligned}$$

So for the adversary strategy defined by the distribution $\mathbb{P}_0$, we would get an expected regret lower bound as required. Thus, it only remains to treat the case where $\mathbb{E}_0[T_\Delta] \leq 0.08\sqrt{\alpha(G)T}$. Plugging this upper bound into (9), we get overall that

$$\mathbb{E}_i[T_i] - \mathbb{E}_0[T_i] \leq 2T\epsilon\sqrt{\log\left(\frac{4}{3}\right)\mathbb{E}_0\left[T_i + 0.08\sqrt{\alpha(G)T}\right]}.$$

Therefore, the expected regret lower bound in (8) is at least

$$\begin{aligned}
&\epsilon\left(T - \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)}\mathbb{E}_0[T_i] - \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} 2T\epsilon\sqrt{\log\left(\frac{4}{3}\right)\mathbb{E}_0\left[T_i + 0.08\sqrt{\alpha(G)T}\right]}\right) \\
&\geq \epsilon\left(T - \frac{T}{\alpha(G)} - 2T\epsilon\sqrt{\log\left(\frac{4}{3}\right)\frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)}\mathbb{E}_0\left[T_i + 0.08\sqrt{\alpha(G)T}\right]}\right) \\
&\geq \epsilon T\left(1 - \frac{1}{\alpha(G)} - 2\epsilon\sqrt{\log\left(\frac{4}{3}\right)\left(\frac{T}{\alpha(G)} + 0.08\sqrt{\alpha(G)T}\right)}\right).
\end{aligned}$$

Since $\alpha(G) > 1$, we have $1 - \frac{1}{\alpha(G)} \geq \frac{1}{2}$, and since $T \geq 0.0064\alpha^3(G)$, we have $0.08\sqrt{\alpha(G)T} \leq \frac{T}{\alpha(G)}$. Overall, we can lower bound the expression above by

$$\epsilon T\left(\frac{1}{2} - 2\epsilon\sqrt{2\log\left(\frac{4}{3}\right)\frac{T}{\alpha(G)}}\right).$$

Picking $\epsilon = \frac{1}{8\sqrt{2\log(4/3)T/\alpha(G)}}$, the expression above is

$$\frac{T}{8\sqrt{2\log\left(\frac{4}{3}\right)\frac{T}{\alpha(G)}}}\frac{1}{4} \geq 0.04\sqrt{\alpha(G)T}.$$

This constitutes a lower bound on the expected regret, from which the result follows.

A.5. Proof of Theorem 6. Fix round t, and let $G = (V, D)$ be the Erdős–Renyi random graph generated at time t, let N_i^- be the in-neighborhood of node i, i.e., the set of nodes j such that $(j, i) \in D$, and denote by d_i^- the indegree of i. We need the following lemmas.

LEMMA 11. *Fix a directed graph $G = (V, D)$. Let $p_1, \ldots, p_K$ be an arbitrary probability distribution defined over V, $f : V \to V$ be an arbitrary permutation of V, and $\mathbb{E}_f$ denote the expectation w.r.t. a random permutation f. Then, for any $i \in V$, we have*

$$\mathbb{E}_f\left[\frac{p_{f(i)}}{\sum_{j\,:\,j\longrightarrow i} p_{f(j)}}\right] = \frac{1}{1+d_i^-}.$$

Proof. Consider selecting an *ordered* sequence S made up of $1 + d_i^-$ nodes of V, where $S = \{f(i)\} \cup \{f(j) \,:\, j \in N_i^-\}$. The expectation $\mathbb{E}_f[\frac{p_{f(i)}}{\sum_{j\,:\,j\longrightarrow i} p_{f(j)}}]$ is an average over the $K(K-1)\cdots(K-d_i^-+1)$ terms corresponding to selecting such S uniformly at random. We can write

$$\begin{aligned}
&\mathbb{E}_f\left[\frac{p_{f(i)}}{\sum_{j\,:\,j\longrightarrow i} p_{f(j)}}\right]\\
&\quad= \mathbb{E}_f\left[\frac{p_{f(i)}}{p_{f(i)} + \sum_{j\in N_i^-} p_{f(j)}}\right]\\
&\quad= \frac{1}{K(K-1)\ldots(K-d_i^-+1)} \sum_{S\subset V, |S|=1+d_i^-} \frac{1}{1+d_i^-} \sum_{i\in S} \frac{p_i}{p_i + \sum_{j\in S, j\neq i} p_j}\\
&\quad= \frac{1}{K(K-1)\ldots(K-d_i^-+1)} \sum_{S\subset V, |S|=1+d_i^-} \frac{1}{1+d_i^-} \sum_{i\in S} \frac{p_i}{\sum_{j\in S} p_j}\\
&\quad= \frac{1}{K(K-1)\ldots(K-d_i^-+1)} \sum_{S\subset V, |S|=1+d_i^-} \frac{1}{1+d_i^-}\\
&\quad= \frac{1}{1+d_i^-}. \qquad \square
\end{aligned}$$

LEMMA 12. *Let $p_1, \ldots, p_K$ be an arbitrary probability distribution defined over V, and let $\mathbb{E}$ denote the expectation w.r.t. the Erdős-Renyi random draw of arcs at time t. Then, for any fixed $i \in V$, we have*

$$\mathbb{E}\left[\frac{p_i}{\sum_{j\,:\,j\longrightarrow i} p_j}\right] = \frac{1}{rK}\left(1-(1-r)^K\right).$$

Proof. For the given $i \in V$ and time t, consider the Bernoulli random variables $X_j, j \in V \setminus \{i\}$, where $X_j = 1$ if arc (j, i) is generated, and $X_j = 0$ otherwise, and denote by $\mathbb{E}_{j\,:\,j\neq i}$ the expectation w.r.t. all of them. We symmetrize $\mathbb{E}[\frac{p_i}{\sum_{j\,:\,j\longrightarrow i} p_j}]$ by means of a random permutation f, as in Lemma 11. We can write

$$\begin{aligned}
\mathbb{E}\left[\frac{p_i}{\sum_{j\,:\,j\to i} p_j}\right] &= \mathbb{E}_{j\,:\,j\neq i}\left[\frac{p_i}{p_i + \sum_{j\,:\,j\neq i} X_j p_j}\right] \\
&= \mathbb{E}_{j\,:\,j\neq i}\mathbb{E}_f\left[\frac{p_{f(i)}}{p_{f(i)} + \sum_{j\,:\,j\neq i} X_{f(j)} p_{f(j)}}\right] \qquad \text{(by symmetry)} \\
&= \mathbb{E}_{j\,:\,j\neq i}\left[\frac{1}{1 + \sum_{j\,:\,j\neq i} X_j}\right] \qquad \text{(from Lemma 11)} \\
&= \sum_{i=0}^{K-1}\binom{K-1}{i} r^i (1-r)^{K-1-i}\frac{1}{i+1} \\
&= \frac{1}{rK}\sum_{i=0}^{K-1}\binom{K}{i+1} r^{i+1}(1-r)^{K-1-i} \\
&= \frac{1}{rK}\left(1-(1-r)^K\right). \qquad \square
\end{aligned}$$

At this point, we follow the proof of Lemma 1 up until (7). We take an expectation $\mathbb{E}_{G_1,\dots,G_T}$ w.r.t. the randomness in generating the sequence of graphs $G_1,\dots,G_T$. This yields

$$\sum_{t=1}^{T}\mathbb{E}_{G_1,\dots,G_T}\left[\sum_{i\in V} p_{i,t}\,\ell_{i,t}\right] \le \sum_{t=1}^{T}\ell_{k,t} + \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}_{G_1,\dots,G_T}\left[Q_t\right].$$

We use Lemma 12 to upper bound $\mathbb{E}_{G_1,\dots,G_T}[Q_t]$ by $\frac{1}{r}\left(1-(1-r)^K\right)$, and we take the outer expectation to remove conditioning, as in the proof of Lemma 1. This concludes the proof.

Appendix B. Technical lemmas and proofs from section 4.1. Again, throughout this appendix, $\mathbb{E}_t[\,\cdot\,]$ is a shorthand for the conditional expectation $\mathbb{E}\big[\,\cdot \mid \{\ell_{i,r}\}_{i\in V}, \{S_{i,r}\}_{i\in V}, r=1,\dots,t,\ I_1,\dots,I_{t-1}\big]$. Moreover, as we did in Appendix A, in round t we first condition on the history $\{\ell_{i,r}\}_{i\in V}, \{S_{i,r}\}_{i\in V}, r=1,\dots,t,\ I_1,\dots,I_{t-1}$, and then we take an outer expectation with respect to that history.

B.1. Proof of Lemma 7. We start to bound the contribution to the overall regret of an instance indexed by b. When clear from the context, we remove the superscript b from $\nu^{(b)}$, $w_{i,t}^{(b)}$, $p_{i,t}^{(b)}$, and other related quantities. For any $t \in T^{(b)}$ we have

$$
\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i\in V} \frac{w_{i,t+1}}{W_t} \\
&= \sum_{i\in V} \frac{w_{i,t}}{W_t} \exp\bigl(-(\nu/2^b)\,\widehat{\ell}_{i,t}\bigr) \\
&= \sum_{i\in R_t} \frac{p_{i,t}-\nu/|R_t|}{1-\nu} \exp\bigl(-(\nu/2^b)\,\widehat{\ell}_{i,t}\bigr) + \sum_{i\notin R_t} \frac{p_{i,t}}{1-\nu} \exp\bigl(-(\nu/2^b)\,\widehat{\ell}_{i,t}\bigr) \\
&\le \sum_{i\in R_t} \frac{p_{i,t}-\nu/|R_t|}{1-\nu} \left(1-\frac{\nu}{2^b}\widehat{\ell}_{i,t}+\frac{1}{2}\left(\frac{\nu}{2^b}\widehat{\ell}_{i,t}\right)^2\right) \\
&\quad + \sum_{i\notin R_t} \frac{p_{i,t}}{1-\nu} \left(1-\frac{\nu}{2^b}\widehat{\ell}_{i,t}+\frac{1}{2}\left(\frac{\nu}{2^b}\widehat{\ell}_{i,t}\right)^2\right) \\
&\quad (\text{using } e^{-x}\le 1-x+x^2/2 \text{ for all } x\ge 0) \\
&\le 1-\frac{\nu/2^b}{1-\nu}\sum_{i\in V} p_{i,t}\widehat{\ell}_{i,t} + \frac{\nu^2/2^b}{1-\nu}\sum_{i\in R_t}\frac{\widehat{\ell}_{i,t}}{|R_t|} + \frac{1}{2}\frac{(\nu/2^b)^2}{1-\nu}\sum_{i\in V} p_{i,t}\bigl(\widehat{\ell}_{i,t}\bigr)^2.
\end{aligned}
$$

Taking logs, upper bounding, and summing over $t\in T^{(b)}$ yields

$$
\begin{aligned}
\ln\frac{W_{|T^{(b)}|+1}}{W_1} &\le -\frac{\nu/2^b}{1-\nu}\sum_{t\in T^{(b)}}\sum_{i\in V} p_{i,t}\widehat{\ell}_{i,t} + \frac{\nu^2/2^b}{1-\nu}\sum_{t\in T^{(b)}}\sum_{i\in R_t}\frac{\widehat{\ell}_{i,t}}{|R_t|} \\
&\quad + \frac{1}{2}\frac{(\nu/2^b)^2}{1-\nu}\sum_{t\in T^{(b)}}\sum_{i\in V} p_{i,t}\bigl(\widehat{\ell}_{i,t}\bigr)^2.
\end{aligned}
$$

Moreover, we also have

$$
\ln\frac{W_{|T^{(b)}|+1}}{W_1} \ge \ln\frac{w_{k,|T^{(b)}|+1}}{W_1} = -\frac{\nu}{2^b}\sum_{t\in T^{(b)}}\widehat{\ell}_{k,t} - \ln K.
$$

Putting this together, rearranging, and using $1-\nu\le 1$ gives

$$
\sum_{t\in T^{(b)}}\sum_{i\in V} p_{i,t}\widehat{\ell}_{i,t} \le \sum_{t\in T^{(b)}}\widehat{\ell}_{k,t} + \frac{2^b\ln K}{\nu} + \nu\sum_{t\in T^{(b)}}\sum_{i\in R_t}\frac{\widehat{\ell}_{i,t}}{|R_t|} + \frac{\nu}{2^{b+1}}\sum_{t\in T^{(b)}}\sum_{i\in V} p_{i,t}\bigl(\widehat{\ell}_{i,t}\bigr)^2.
$$

Reintroducing the notation $\nu^{(b)}$ and summing over $b=0,1,\dots,\lfloor\log_2 K\rfloor$ gives

$$
\begin{aligned}
&\sum_{t=1}^{T}\left(\sum_{i\in V} p^{(b_t)}_{i,t}\widehat{\ell}^{(b_t)}_{i,t} - \widehat{\ell}_{k,t}\right) \\
&\quad\le \sum_{b=0}^{\lfloor\log_2 K\rfloor}\frac{2^b\ln K}{\nu^{(b)}} + \sum_{t=1}^{T}\sum_{i\in R_t}\frac{\nu^{(b_t)}\widehat{\ell}^{(b_t)}_{i,t}}{|R_t|} + \sum_{t=1}^{T}\frac{\nu^{(b_t)}}{2^{b_t+1}}\sum_{i\in V} p^{(b_t)}_{i,t}\bigl(\widehat{\ell}^{(b_t)}_{i,t}\bigr)^2.
\end{aligned}
\tag{10}
$$

Now, similarly to the proof of Lemma 1, we have that $\mathbb{E}_t\bigl[\widehat{\ell}^{(b_t)}_{i,t}\bigr]=\ell_{i,t}$ and $\mathbb{E}_t\bigl[(\widehat{\ell}^{(b_t)}_{i,t})^2\bigr]\le\frac{1}{q^{(b_t)}_{i,t}}$ for any i and t. Moreover,

$$
\sum_{t=1}^{T}\sum_{i\in R_t}\frac{\nu^{(b_t)}\ell_{i,t}}{|R_t|} \le \sum_{t=1}^{T}\sum_{i\in R_t}\frac{\nu^{(b_t)}}{|R_t|} = \sum_{t=1}^{T}\nu^{(b_t)} = \sum_{b=0}^{\lfloor\log_2 K\rfloor}\nu^{(b)}|T^{(b)}|
$$

and

$$\sum_{t=1}^{T} \frac{\nu^{(b_t)}}{2^{b_t+1}} Q_t^{(b_t)} = \sum_{b=0}^{\lfloor \log_2 K \rfloor} \frac{\nu^{(b)}}{2^{b+1}} \sum_{t \in T^{(b)}} Q_t^{(b)}.$$

Therefore, by taking expectations on both sides of (10), and recalling that $T^{(b)}$ is a random variable (since the adversary adaptively decides which steps t fall into $T^{(b)}$), we get

$$\begin{aligned} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] &\le \sum_{b=0}^{\lfloor \log_2 K \rfloor} \mathbb{E}\left[\frac{2^b \ln K}{\nu^{(b)}} + \nu^{(b)} |T^{(b)}| + \frac{\nu^{(b)}}{2^{b+1}} \sum_{t \in T^{(b)}} Q_t^{(b)}\right] \\ &= \sum_{b=0}^{\lfloor \log_2 K \rfloor} \left(\frac{2^b \ln K}{\nu^{(b)}} + \nu^{(b)} \mathbb{E}\left[\sum_{t \in T^{(b)}} \left(1 + \frac{Q_t^{(b)}}{2^{b+1}}\right)\right]\right). \end{aligned} \tag{11}$$

This establishes (4).

In order to prove inequality (5), we tune each $\nu^{(b)}$ separately through a doubling trick. More specifically, for each $b = 0, 1, \ldots, \lfloor \log_2 K \rfloor$, we use a sequence $\nu_r^{(b)} = \sqrt{(2^b \ln K)/2^r}$ for $r = 0, 1, \ldots$. We initially run the algorithm with $\nu_0^{(b)}$. Whenever the algorithm is running with $\nu_r^{(b)}$ and detects that the condition $\sum_s \overline{Q}_s^{(b)} > 2^r$ is satisfied, the sum being over all s so far in $T^{(b)}$, then we restart the algorithm with $\nu_{r+1}^{(b)}$.

Now, because the contribution of instance b to (11) is

$$\frac{2^b \ln K}{\nu^{(b)}} + \nu^{(b)} \sum_{t \in T^{(b)}} \overline{Q}_t^{(b)},$$

the regret we pay when using any $\nu_r^{(b)}$ is at most $2\sqrt{(2^b \ln K) 2^r}$. The largest r we need is $\lceil \log_2 \overline{Q}^{(b)} \rceil$ and

$$\sum_{r=0}^{\lceil \log_2 \overline{Q}^{(b)} \rceil} 2^{r/2} < 5\sqrt{\overline{Q}^{(b)}}.$$

Since we pay regret at most 1 for each restart (more precisely, during the last round before each restart takes place), we get

$$\mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le c \sum_{b=0}^{\lfloor \log_2 K \rfloor} \mathbb{E}\left[\sqrt{(\ln K)\left(2^b |T^{(b)}| + \frac{1}{2} \sum_{t \in T^{(b)}} Q_t^{(b)}\right)} + \left\lceil \log_2 \overline{Q}^{(b)} \right\rceil\right]$$

for some positive constant c. Taking into account that

$$\begin{aligned} \sum_{b=0}^{\lfloor \log_2 K \rfloor} 2^b |T^{(b)}| &\le 2 \sum_{t=1}^{T} |R_t| \\ \sum_{b=0}^{\lfloor \log_2 K \rfloor} \sum_{t \in T^{(b)}} Q_t^{(b)} &= \sum_{t=1}^{T} Q_t^{(b_t)} \\ \sum_{b=0}^{\lfloor \log_2 K \rfloor} \left\lceil \log_2 \overline{Q}^{(b)} \right\rceil &= \mathcal{O}\big((\ln K) \ln(KT)\big), \end{aligned}$$

we obtain

$$\begin{aligned}
&\mathbb{E}\big[L_{A,T} - L_{k,T}\big] \\
&\quad \le c \sum_{b=0}^{\lfloor \log_2 K \rfloor} \mathbb{E}\left[\sqrt{(\ln K)\left(2^b|T^{(b)}| + \frac{1}{2}\sum_{t\in T^{(b)}} Q_t^{(b)}\right)}\right] + \mathcal{O}\big((\ln K)\ln(KT)\big) \\
&\quad \le c\,\lfloor \log_2 K \rfloor \mathbb{E}\left[\sqrt{\frac{\ln K}{\lfloor \log_2 K \rfloor}\sum_{t=1}^{T}\left(2|R_t| + \frac{1}{2}Q_t^{(b_t)}\right)}\right] + \mathcal{O}\big((\ln K)\ln(KT)\big) \\
&\quad = \mathcal{O}\left((\ln K)\,\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\left(4|R_t| + Q_t^{(b_t)}\right)}\right] + (\ln K)\ln(KT)\right),
\end{aligned}$$

as desired.

B.2. Proof of Theorem 8. The following graph-theoretic lemma turns out to be fairly useful for analyzing directed settings. It is a directed-graph counterpart to a well-known result [11, 48] holding for undirected graphs.

LEMMA 13. *Let $G = (V, D)$ be a directed graph, with $V = \{1, \ldots, K\}$. Let d_i^- be the indegree of node i, and let $\alpha = \alpha(G)$ be the independence number of G. Then*

$$\sum_{i=1}^{K} \frac{1}{1 + d_i^-} \le 2\alpha \,\ln\left(1 + \frac{K}{\alpha}\right).$$

Proof. We proceed by induction, starting from the original K-node graph $G = G_K$ with indegrees $\{d_i^-\}_{i=1}^K = \{d_{i,K}^-\}_{i=1}^K$ and independence number $\alpha = \alpha_K$, and then progressively reduce G by eliminating nodes and incident (both departing and incoming) arcs, thereby obtaining a sequence of smaller and smaller graphs $G_K, G_{K-1}, G_{K-2}, \ldots,$ associated indegrees $\{d_{i,K}^-\}_{i=1}^K$, $\{d_{i,K-1}^-\}_{i=1}^{K-1}$, $\{d_{i,K-2}^-\}_{i=1}^{K-2}, \ldots,$ and independence numbers $\alpha_K, \alpha_{K-1}, \alpha_{K-2}, \ldots.$ Specifically, in step s we sort nodes $i = 1, \ldots, s$ of G_s in nonincreasing value of $d_{i,s}^-$, and we obtain G_{s-1} from G_s by eliminating node 1 (i.e., the one having the largest indegree among the nodes of G_s), along with its incident arcs. On all such graphs, we use the classical Turan theorem (e.g., [4]) stating that any *undirected* graph with n_s nodes and m_s edges has an independent set of size at least $\frac{n_s}{\frac{2m_s}{n_s}+1}$. This implies that if $G_s = (V_s, D_s)$, then α_s satisfies[10]

$$\frac{|D_s|}{|V_s|} \ge \frac{|V_s|}{2\alpha_s} - \frac{1}{2}. \tag{12}$$

We then start from G_K. We can write

$$d_{1,K}^- = \max_{i=1\ldots K} d_{i,K}^- \ge \frac{1}{K}\sum_{i=1}^{K} d_{i,K}^- = \frac{|D_K|}{|V_K|} \ge \frac{|V_K|}{2\alpha_K} - \frac{1}{2}.$$

[10]Note that $|D_s|$ is at least as large as the number of edges of the undirected version of G_s which the independence number α_s actually refers to.

Hence,

$$\begin{aligned}\sum_{i=1}^{K}\frac{1}{1+d^-_{i,K}} &= \frac{1}{1+d^-_{1,K}}+\sum_{i=2}^{K}\frac{1}{1+d^-_{i,K}}\\ &\le \frac{2\alpha_K}{\alpha_K+K}+\sum_{i=2}^{K}\frac{1}{1+d^-_{i,K}}\\ &\le \frac{2\alpha_K}{\alpha_K+K}+\sum_{i=1}^{K-1}\frac{1}{1+d^-_{i,K-1}},\end{aligned}$$

where the last inequality follows from $d^-_{i+1,K} \ge d^-_{i,K-1}$, $i = 1, \ldots, K-1$, due to the arc elimination trasforming G_K into G_{K-1}. Recursively applying the same argument to G_{K-1} (i.e., to the sum $\sum_{i=1}^{K-1}\frac{1}{1+d^-_{i,K-1}}$) and then iterating all the way to G_1 yields the upper bound

$$\sum_{i=1}^{K}\frac{1}{1+d^-_{i,K}} \le \sum_{i=1}^{K}\frac{2\alpha_i}{\alpha_i+i}.$$

Combining with $\alpha_i \le \alpha_K = \alpha$ and the fact that $\frac{\alpha}{\alpha+i}$ increases in α, this is at most $2\alpha\sum_{i=1}^{K}\frac{1}{\alpha+i} \;\le\; 2\alpha\ln\left(1+\frac{K}{\alpha}\right)$, concluding the proof. $\square$

The next lemma relates the size $|R_t|$ of the dominating set R_t computed by the Greedy Set Cover algorithm of [20], operating on the time-t feedback system $\{S_{i,t}\}_{i\in V}$, to the independence number $\alpha(G_t)$ and the domination number $\gamma(G_t)$ of G_t.

Lemma 14. *Let $\{S_i\}_{i\in V}$ be a feedback system, and let $G = (V, D)$ be the induced directed graph, with vertex set $V = \{1, \ldots, K\}$, independence number $\alpha = \alpha(G)$, and domination number $\gamma = \gamma(G)$. Then the dominating set R constructed by the Greedy Set Cover algorithm (see section 2) satisfies*

$$|R| \le \min\bigl\{\gamma(1+\ln K), \lceil 2\alpha\ln K\rceil + 1\bigr\}.$$

Proof. As recalled in section 2, the Greedy Set Cover algorithm of [20] achieves $|R| \le \gamma(1+\ln K)$. In order to prove the other bound, consider the sequence of graphs $G = G_1, G_2, \ldots,$ where each $G_{s+1} = (V_{s+1}, D_{s+1})$ is obtained by removing from G_s the vertex i_s selected by the Greedy Set Cover algorithm, together with all the vertices in G_s that are dominated by i_s, and all arcs incident to these vertices. By definition of the algorithm, the outdegree d^+_s of i_s in G_s is largest in G_s. Hence,

$$d^+_s \ge \frac{|D_s|}{|V_s|} \ge \frac{|V_s|}{2\alpha_s} - \frac{1}{2} \ge \frac{|V_s|}{2\alpha} - \frac{1}{2}$$

by Turan's theorem (e.g., [4]), where α_s is the independence number of G_s and $\alpha \ge \alpha_s$. This shows that

$$|V_{s+1}| = |V_s| - d^+_s - 1 \le |V_s|\left(1-\frac{1}{2\alpha}\right) \le |V_s|e^{-1/(2\alpha)}.$$

Iterating, we obtain $|V_s| \le K\,e^{-(s-1)/(2\alpha)}$. Choosing $s = \lceil 2\alpha\ln K\rceil + 1$ gives $|V_s| < 1$, thereby covering all nodes. Hence the dominating set $R = \{i_1, \ldots, i_s\}$ so constructed satisfies $|R| \le \lceil 2\alpha\ln K\rceil + 1$. $\square$

LEMMA 15. *If $a, b \geq 0$ and $a + b \geq B > A > 0$, then*

$$\frac{a}{a+b-A} \leq \frac{a}{a+b} + \frac{A}{B-A}.$$

Proof.

$$\frac{a}{a+b-A} - \frac{a}{a+b} = \frac{aA}{(a+b)(a+b-A)} \leq \frac{A}{a+b-A} \leq \frac{A}{B-A}. \qquad \square$$

We now lift Lemma 13 to a more general statement.

LEMMA 16. *Let $G = (V, D)$ be a directed graph, with vertex set $V = \{1, \dots, K\}$, and arc set D. Let α be the independence number of G, $R \subseteq V$ be a dominating set for G of size $r = |R|$, and $p_1, \dots, p_K$ be a probability distribution defined over V, such that $p_i \geq \beta > 0$, for $i \in R$. Then*

$$\sum_{i=1}^{K} \frac{p_i}{\sum_{j\,:\,j \longrightarrow i} \; p_j} \leq 2\alpha \, \ln\left(1 + \frac{\lceil \frac{K^2}{r\beta} \rceil + K}{\alpha}\right) + 2r.$$

Proof. The idea is to appropriately discretize the probability values p_i, and then upper bound the discretized counterpart of $\sum_{i=1}^{K} \frac{p_i}{\sum_{j\,:\,j \longrightarrow i} \; p_j}$ by reducing to an expression that can be handled by Lemma 13. In order to make this discretization effective, we need to single out the terms $\frac{p_i}{\sum_{j\,:\,j \longrightarrow i} \; p_j}$ corresponding to nodes $i \in R$. We first write

$$\begin{aligned} \sum_{i=1}^{K} \frac{p_i}{\sum_{j\,:\,j \longrightarrow i} \; p_j} &= \sum_{i \in R} \frac{p_i}{\sum_{j\,:\,j \longrightarrow i} \; p_j} + \sum_{i \notin R} \frac{p_i}{\sum_{j\,:\,j \longrightarrow i} \; p_j} \\ &\leq r + \sum_{i \notin R} \frac{p_i}{\sum_{j\,:\,j \longrightarrow i} \; p_j} \end{aligned} \tag{13}$$

and then focus on (13).

Let us discretize the unit interval[11] $(0, 1]$ into subintervals $\left(\frac{j-1}{M}, \frac{j}{M}\right]$, $j = 1, \dots, M$, where $M = \lceil \frac{K^2}{r\beta} \rceil$. Let $\widehat{p}_i = j/M$ be the discretized version of p_i, where j is the unique integer such that $\widehat{p}_i - 1/M < p_i \leq \widehat{p}_i$. We focus on a single node $i \notin R$ with in-neighborhood N_i^- and indegree $d_i^- = |N_i^-|$. Introduce the shorthand notation $P_i = \sum_{j\,:\,j \longrightarrow i, j \neq i} \; p_j$ and $\widehat{P}_i = \sum_{j\,:\,j \longrightarrow i, j \neq i} \; \widehat{p}_j$. We have that $\widehat{P}_i \geq P_i \geq \beta$, since i is dominated by some node $j \in R \cap N_i^-$ such that $p_j \geq \beta$. Moreover, $P_i > \widehat{P}_i - \frac{d_i^-}{M} \geq$

[11]The zero value is not of concern here, because if $p_i = 0$, then the corresponding term in (13) can be disregarded.

$\beta - \frac{d_i^-}{M} > 0$, and $\widehat{p}_i + \widehat{P}_i \geq \beta$. Hence, for any fixed node $i \notin R$, we can write

$$\begin{aligned} \frac{p_i}{p_i + P_i} &\leq \frac{\widehat{p}_i}{\widehat{p}_i + P_i} \\ &< \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i - \frac{d_i^-}{M}} \\ &\leq \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + \frac{d_i^-/M}{\beta - d_i^-/M} \\ &= \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + \frac{d_i^-}{\beta M - d_i^-} \\ &< \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + \frac{r}{K - r}, \end{aligned}$$

where in the second-to-last inequality we used Lemma 15 with $a = \widehat{p}_i$, $b = \widehat{P}_i$, $A = d_i^-/M$, and $B = \beta > d_i^-/M$. Recalling (13) and summing over i then gives

$$\sum_{i=1}^{K} \frac{p_i}{p_i + P_i} \leq r + \sum_{i \notin R} \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + r = \sum_{i \notin R} \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + 2r. \tag{14}$$

Therefore, we continue by bounding from above the right-hand side of (14). We first observe that

$$\sum_{i \notin R} \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} = \sum_{i \notin R} \frac{\widehat{s}_i}{\widehat{s}_i + \widehat{S}_i} \qquad \text{and} \qquad \widehat{S}_i = \sum_{j \,:\, j \to i, j \neq i} \widehat{s}_j, \tag{15}$$

where $\widehat{s}_i = M\widehat{p}_i$, $i = 1, \ldots, K$, are integers. Based on the original graph G, we construct a new graph $\widehat{G}$ made up of connected cliques. In particular:

- Each node i of G is replaced in $\widehat{G}$ by a clique C_i of size $\widehat{s}_i$; nodes within C_i are connected by length-two directed cycles.
- If arc (i, j) is in G, then for each node of C_i draw an arc toward each node of C_j.

We would like to apply Lemma 13 to $\widehat{G}$. Note that, by the above construction,

- the independence number of $\widehat{G}$ is the same as that of G,
- the indegree $\widehat{d}_k^-$ of each node k in clique C_i satisfies $\widehat{d}_k^- = \widehat{s}_i - 1 + \widehat{S}_i$, and
- the total number of nodes of $\widehat{G}$ is

$$\sum_{i=1}^{K} \widehat{s}_i = M \sum_{i=1}^{K} \widehat{p}_i < M \sum_{i=1}^{K} \left(p_i + \frac{1}{M} \right) = M + K.$$

Hence, we can apply Lemma 13 to $\widehat{G}$ with indegrees $\widehat{d}_k^-$ and find that

$$\sum_{i \notin R} \frac{\widehat{s}_i}{\widehat{s}_i + \widehat{S}_i} = \sum_{i \notin R} \sum_{k \in C_i} \frac{1}{1 + \widehat{d}_k^-} \leq \sum_{i=1}^{K} \sum_{k \in C_i} \frac{1}{1 + \widehat{d}_k^-} \leq 2\alpha \ln \left(1 + \frac{M + K}{\alpha} \right).$$

Combining (14) and (15) and recalling the value of M gives the claimed result. $\square$

Proof of Theorem 8. We are now ready to derive the proof of the theorem. We start from the upper bound (5) in the statement of Lemma 7. We want to bound the quantities $|R_t|$ and $Q_t^{(b_t)}$ occurring therein at any step t in which a restart does not occur—the regret for the time steps when a restart occurs is already accounted for by the term $\mathcal{O}\big((\ln K)\ln(KT)\big)$ in (5). Now, Lemma 14 gives

$$|R_t| = \mathcal{O}\big(\alpha(G_t)\ln K\big).$$

If $\nu_t = \nu_t^{(b_t)}$ for any time t when a restart does not occur, it is not hard to see that $\nu_t = \Omega\big(\sqrt{(\ln K)/(KT)}\big)$. Moreover, Lemma 16 states that

$$Q_t = \mathcal{O}\big(\alpha(G_t)\ln(K^2/\nu_t) + |R_t|\big) = \mathcal{O}\big(\alpha(G_t)\ln(K/\nu_t)\big).$$

Hence,

$$Q_t = \mathcal{O}\big(\alpha(G_t)\ln(KT)\big).$$

Putting this together as in (5) and moving the expectation inside the square root by Jensen's inequality gives the desired result.

Appendix C. Technical lemmas and proofs from section 4.2. As in the previous appendices, we will use here $\mathbb{E}_t[\,\cdot\,]$ as shorthand for the conditional expectation $\mathbb{E}\big[\,\cdot \mid \{\ell_{i,r}\}_{i\in V}, \{S_{i,r}\}_{i\in V}, r = 1,\ldots,t,\ I_1,\ldots,I_{t-1}\big]$. Under this conditioning, we can treat random variables such as $p_{i,t}$, $\ell_{i,t}$, and $S_{i,t}$ as fixed, and only consider randomness with respect to the algorithm's play at round t.

C.1. Proof of Theorem 9. The following lemmas are of preliminary importance in order to understand the behavior of the ELP.P algorithm. Recall that for a directed graph $G = (V, D)$, with vertex set $V = \{1,\ldots,K\}$ and arc set D, we write $\{j : j \to i\}$ to denote the set of nodes j which are in-neighbors of node i, including node i itself. Similarly, $\{j : i \to j\}$ is the set of out-neighbors of node i where, again, node i is included in this set. Let Δ_K be the K-dimensional probability simplex.

LEMMA 17. *Consider a directed graph $G = (V, D)$, with vertex set $V = \{1,\ldots,K\}$ and arc set D. Let $\mathtt{mas}(G)$ be the size of a largest acyclic subgraph of G. If $s_1,\ldots,s_K$ is a solution to the linear program*

$$\max_{(s_1,\ldots,s_K)\in\Delta_K}\ \min_{i\in V}\left(\sum_{j\,:\,j\longrightarrow i} s_j\right), \tag{16}$$

then we have

$$\max_{i\in V}\frac{1}{\sum_{j\,:\,j\longrightarrow i} s_j} \le \mathtt{mas}(G).$$

Proof. We first show that the above inequality holds when the right-hand side is replaced by $\gamma(G)$, the domination number of G. Then let R be a smallest (minimal) dominating set of G, so that $|R| = \gamma(G)$. Consider the valid assignment $s_i = \mathbb{I}\{i \in R\}/\gamma(G)$ for all $i \in V$. This implies that for all i, $\sum_{j\,:\,j\longrightarrow i} s_j \ge 1/\gamma(G)$, because any $i \in V$ either is in R or is dominated by a node in R. Therefore, for this particular assignment, we have

$$\max_{i\in V}\frac{1}{\sum_{j\,:\,j\longrightarrow i} s_j} \le \gamma(G).$$

The assignment returned by the linear program might be different, but it can only make the left-hand side above smaller,[12] so the inequality still holds. Finally, $\gamma(G) \le \mathtt{mas}(G)$ because any set $M \subseteq V$ of nodes belonging to a maximal acyclic subgraph of G is itself a dominating set for G. In fact, assuming the contrary, let j be any node such that $j \notin M$. Then, including j in M would create a cycle (because of the maximality of M), implying that j is already dominated by some other node in M. ∎

LEMMA 18. *Consider a directed graph $G = (V, D)$, with vertex set $V = \{1, \dots, K\}$ and arc set D. Let $\mathtt{mas}(G)$ be the size of a largest acyclic subgraph of G. Let $(p_1, \dots, p_K) \in \Delta_K$ and $(s_1, \dots, s_K) \in \Delta_K$ satisfy*

$$\sum_{i=1}^{K} \frac{p_i}{\sum_{j\,:\,j \longrightarrow i} p_j} \le \mathtt{mas}(G) \qquad \textit{and} \qquad \max_{i \in V} \frac{1}{\sum_{j\,:\,j \longrightarrow i} s_j} \le \mathtt{mas}(G)$$

with $p_i \ge \nu s_i$, $i \in V$, for some $\nu > 0$. Finally, introduce the shorthand $q_i = \sum_{j\,:\,j \longrightarrow i} p_j$ for $i \in V$. Then the following relations hold:

1. $\sum_{i=1}^{K} \frac{p_i}{q_j^2} \le \frac{\mathtt{mas}^2(G)}{\nu}$;
2. $\sum_{i=1}^{K} p_i \sum_{j\,:\,i \longrightarrow j} \frac{p_j}{q_j} = 1$;
3. $\sum_{i=1}^{K} p_i \sum_{j\,:\,i \longrightarrow j} \frac{p_j}{q_j^2} \le \mathtt{mas}(G)$;
4. $\sum_{i=1}^{K} p_i \left(\sum_{j\,:\,i \longrightarrow j} \frac{p_j}{q_j}\right)^2 \le \mathtt{mas}(G)$;
5. $\sum_{i=1}^{K} p_i \left(\sum_{j\,:\,i \longrightarrow j} \frac{p_j}{q_j^2}\right)^2 \le \frac{\mathtt{mas}^3(G)}{\nu}$.

Proof.

1. Applying the assumptions in the lemma, we obtain

$$\begin{aligned}
\sum_{i=1}^{K} \frac{p_i}{q_i^2} &= \sum_{i=1}^{K} \left(\frac{p_i}{q_i}\right) \left(\frac{1}{q_i}\right) \\
&\le \left(\sum_{i=1}^{K} \frac{p_i}{q_i}\right) \left(\max_{i \in V} \frac{1}{q_i}\right) \\
&= \left(\sum_{i=1}^{K} \frac{p_i}{q_i}\right) \left(\max_{i \in V} \frac{1}{\sum_{j\,:\,j \longrightarrow i} p_j}\right) \\
&\le \mathtt{mas}(G) \max_{i \in V} \frac{1}{\nu \left(\sum_{j\,:\,j \longrightarrow i} s_j\right)} \\
&\le \frac{\mathtt{mas}^2(G)}{\nu}.
\end{aligned}$$

2. We have

$$\sum_{i=1}^{K} \sum_{j\,:\,i \longrightarrow j} \frac{p_i\, p_j}{q_j} = \sum_{j=1}^{K} \frac{p_j\, q_j}{q_j} = \sum_{j=1}^{K} p_j = 1.$$

3. Similar to the previous item, we can write

$$\sum_{i=1}^{K} \sum_{j\,:\,i \longrightarrow j} \frac{p_i\, p_j}{q_j^2} = \sum_{j=1}^{K} \frac{p_j\, q_j}{q_j^2} = \sum_{j=1}^{K} \frac{p_j}{q_j} \le \mathtt{mas}(G).$$

[12]This can be seen by noting that (16) is equivalent to $\min_{(s_1,\dots,s_K) \in \Delta_K} \max_{i \in V} \frac{1}{\sum_{j\,:\,j \longrightarrow i} s_j}$.

4. From item 2, and the assumptions of this lemma, we can write

$$\begin{aligned}\sum_{i=1}^{K} p_i \left(\sum_{j\,:\,i\longrightarrow j} \frac{p_j}{q_j}\right)^2 &= \sum_{i=1}^{K}\left(p_i \sum_{j\,:\,i\longrightarrow j} \frac{p_j}{q_j}\right)\left(\sum_{j\,:\,i\longrightarrow j} \frac{p_j}{q_j}\right)\\ &\le \left(\sum_{i=1}^{K} p_i \sum_{j\,:\,i\longrightarrow j} \frac{p_j}{q_j}\right)\left(\sum_{j=1}^{K} \frac{p_j}{q_j}\right)\\ &= \sum_{j=1}^{K} \frac{p_j}{q_j} \le \texttt{mas}(G).\end{aligned}$$

5. From items 1 and 3 above, we can write

$$\begin{aligned}\sum_{i=1}^{K} p_i \left(\sum_{j\,:\,i\longrightarrow j} \frac{p_j}{q_j^2}\right)^2 &= \sum_{i=1}^{K}\left(p_i \sum_{j\,:\,i\longrightarrow j} \frac{p_j}{q_j^2}\right)\left(\sum_{j\,:\,i\longrightarrow j} \frac{p_j}{q_j^2}\right)\\ &\le \left(\sum_{i=1}^{K} p_i \sum_{j\,:\,i\longrightarrow j} \frac{p_j}{q_j^2}\right)\left(\sum_{i=1}^{K} \frac{p_i}{q_i^2}\right)\\ &\le \texttt{mas}(G)\,\frac{\texttt{mas}^2(G)}{\nu}\\ &= \frac{\texttt{mas}^3(G)}{\nu},\end{aligned}$$

concluding the proof. ∎

Lemma 18 applies, in particular, to the distributions $s_t = (s_{1,t}, \dots, s_{K,t})$ and $p_t = (p_{1,t}, \dots, p_{K,t})$ computed by ELP.P at round t. The condition for p_t follows from Lemma 10, while the condition for s_t follows from Lemma 17. In other words, putting together Lemmas 10 and 17 establishes the following corollary.

COROLLARY 19. *Let $p_t = (p_{1,t}, \dots, p_{K,t}) \in \Delta_K$ and $s_t = (s_{1,t}, \dots, s_{K,t}) \in \Delta_K$ be the distributions generated by ELP.P at round t. Then,*

$$\sum_{i=1}^{K} \frac{p_{i,t}}{\sum_{j\,:\,j\xrightarrow{t}i} p_{j,t}} \le \texttt{mas}(G) \qquad \text{and} \qquad \max_{i\in V} \frac{1}{\sum_{j\,:\,j\xrightarrow{t}i} s_{j,t}} \le \texttt{mas}(G),$$

with $p_{i,t} \ge \nu_t\, s_{i,t}$, for all $i = 1, \dots, K$.

For the next result, we need the following version of Freedman's inequality [23] (see also [17, Lemma A.8]).

LEMMA 20. *Let $X_1, \dots, X_T$ be a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}_{t=1,\dots,T}$, and with $|X_i| \le B$ almost surely for all i. Also, let $V > 0$ be a fixed upper bound on $\sum_{t=1}^{T} \mathbb{E}\big[X_t^2 \mid \mathcal{F}_{t-1}\big]$. Then for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$\sum_{t=1}^{T} X_t \le \sqrt{2\ln\left(\frac{1}{\delta}\right) V} + \frac{B}{2}\ln\left(\frac{1}{\delta}\right).$$

LEMMA 21. *Let $\{a_t\}_{t=1}^T$ be an arbitrary sequence of positive numbers, and let $s_t = (s_{1,t}, \dots, s_{K,t})$ and $p_t = (p_{1,t}, \dots, p_{K,t})$ be the probability distributions computed by ELP.P at the tth round. Then, with probability at least $1-\delta$,*

$$\sum_{t=1}^T \sum_{i=1}^K a_t p_{i,t}(\widehat{g}_{i,t} - g_{i,t}) \le \sqrt{2 \ln\left(\frac{1}{\delta}\right) \sum_{t=1}^T a_t^2 \, \mathtt{mas}(G_t)}$$

$$+ \frac{1}{2} \ln\left(\frac{1}{\delta}\right) \max_{t=1,\dots,T} \big(a_t \, \mathtt{mas}(G_t)\big) + \beta \sum_{t=1}^T a_t \, \mathtt{mas}(G_t). \tag{17}$$

Proof. Recall that $q_{i,t} = \sum_{j \,:\, j \xrightarrow{t} i} p_{j,t}$ for $i \in V$, and let

$$\tilde{g}_{i,t} = \frac{g_{i,t} \, \mathbb{I}\{i \in S_{I_t,t}\}}{q_{i,t}}$$

with $g_{i,t} = 1 - \ell_{i,t}$. Note that $\widehat{g}_{i,t}$ in Algorithm 3 satisfies $\widehat{g}_{i,t} = \tilde{g}_{i,t} + \frac{\beta}{q_{i,t}}$, so the left-hand side of (17) equals

$$\sum_{t=1}^T \sum_{i=1}^K a_t p_{i,t}(\tilde{g}_{i,t} - g_{i,t}) + \beta \sum_{t=1}^T \sum_{i=1}^K \frac{a_t p_{i,t}}{q_{i,t}},$$

which by Corollary 19 is at most

$$\sum_{t=1}^T \sum_{i=1}^K a_t p_{j,t}(\tilde{g}_{i,t} - g_{i,t}) + \beta \sum_{t=1}^T a_t \, \mathtt{mas}(G_t). \tag{18}$$

It is easy to verify that $\sum_{i=1}^K a_t p_{i,t}(\tilde{g}_{i,t} - g_{i,t})$ is a martingale difference sequence (indexed by t), because $q_{i,t} > 0$ and therefore $\mathbb{E}_t[\tilde{g}_{i,t}] = g_{i,t}$. Moreover,

$$\begin{aligned}
\sum_{i=1}^K a_t p_{i,t}\left(\tilde{g}_{i,t} - g_{i,t}\right) &= \sum_{i=1}^K a_t p_{i,t} \left(\frac{\mathbb{I}\{i \in S_{I_t,t}\}}{q_{i,t}} - 1\right) g_{i,t} \\
&\le a_t \sum_{i=1}^K \frac{p_{i,t}}{q_{i,t}} \\
&\le \max_{t=1,\dots,T} a_t \, \mathtt{mas}(G_t),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_t\left[\left(\sum_{i=1}^K a_t p_{i,t}(\tilde{g}_{i,t} - g_{i,t})\right)^2\right] &\le a_t^2 \, \mathbb{E}_t\left[\left(\sum_{i=1}^K p_{i,t} \tilde{g}_{i,t}\right)^2\right] \\
&\le a_t^2 \sum_{i=1}^K p_{i,t} \left(\sum_{j \,:\, i \xrightarrow{t} j} p_{j,t} \frac{1}{q_{j,t}}\right)^2 \\
&\le a_t^2 \, \mathtt{mas}(G_t)
\end{aligned}$$

by Lemma 18, item 4. Therefore, by invoking Lemma 20, we get that with probability at least $1-\delta$,

$$\sum_{t=1}^{T}\sum_{j=1}^{K} a_t p_{j,t}(\tilde{g}_{j,t}-g_{j,t}) \le \sqrt{2\ln\left(\frac{1}{\delta}\right)\sum_{t=1}^{T} a_t^2\,\mathtt{mas}(G_t)} + \frac{1}{2}\ln\left(\frac{1}{\delta}\right)\max_{t=1,\dots,T} a_t\,\mathtt{mas}(G_t).$$

Substituting into (18), the lemma follows. ∎

LEMMA 22. *Let $s_t = (s_{1,t},\dots,s_{K,t})$ and $p_t = (p_{1,t},\dots,p_{K,t})$ be the probability distributions computed by ELP.P, run with $\beta \le 1/4$, at round t. Then, with probability at least $1-\delta$,*

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}\widehat{g}_{i,t}^2 \le \sum_{t=1}^{T}&\left(\frac{\beta^2\,\mathtt{mas}^2(G_t)}{\nu_t} + 2\,\mathtt{mas}(G_t)\right)\\ &+\sqrt{2\ln\left(\frac{1}{\delta}\right)\sum_{t=1}^{T}\left(\frac{4\beta^2\,\mathtt{mas}^4(G_t)}{\nu_t^2}+\frac{3\,\mathtt{mas}^3(G_t)}{\nu_t}\right)}\\ &+\ln\left(\frac{1}{\delta}\right)\max_{t=1,\dots,T}\frac{\mathtt{mas}^2(G_t)}{\nu_t}.\end{aligned}$$

Proof. Recall that $q_{i,t} = \sum_{j\,:\,j\xrightarrow{t}i} p_{j,t}$ for $i \in V$. By the way we defined $\widehat{g}_{i,t}$ and Lemma 18, item 1, we have that

$$\sum_{i=1}^{K} p_{i,t}\widehat{g}_{i,t}^2 \;\le\; \sum_{i=1}^{K} p_{i,t}\left(\frac{1+\beta}{q_{i,t}}\right)^2 \;\le\; \frac{(1+\beta)^2\mathtt{mas}^2(G_t)}{\nu_t}.$$

Moreover, from $g_{i,t}\le 1$, and again using Lemma 18, item 1, we can write

$$\begin{aligned}\mathbb{E}_t\left[\left(\sum_{j=1}^{K} p_{j,t}\widehat{g}_{j,t}^2\right)^2\right] \;&\le\; \sum_{i=1}^{K} p_{i,t}\left(\sum_{j=1}^{K}\frac{p_{j,t}}{(q_{j,t})^2}\left(\mathbb{I}\{i\xrightarrow{t}j\}+\beta\right)^2\right)^2\\ &= \sum_{i=1}^{K} p_{i,t}\left(\beta^2\sum_{j=1}^{K}\frac{p_{j,t}}{(q_{j,t})^2} + (2\beta+1)\sum_{j\,:\,i\xrightarrow{t}j}\frac{p_{j,t}}{(q_{j,t})^2}\right)^2\\ &\le \sum_{i=1}^{K} p_{i,t}\left(\frac{\beta^2\,\mathtt{mas}^2(G_t)}{\nu_t} + (2\beta+1)\sum_{j\,:\,i\xrightarrow{t}j}\frac{p_{j,t}}{(q_{j,t})^2}\right)^2,\end{aligned}$$

which, by expanding, using Lemma 18, items 3 and 5, and slightly simplifying, is at most

$$\frac{(\beta^4+2\beta^2(2\beta+1))\,\mathtt{mas}^4(G_t)}{\nu_t^2} + \frac{(2\beta+1)^2\,\mathtt{mas}^3(G_t)}{\nu_t} \le \frac{4\beta^2\,\mathtt{mas}^4(G_t)}{\nu_t^2} + \frac{3\,\mathtt{mas}^3(G_t)}{\nu_t},$$

the last inequality exploiting the assumption $\beta \le 1/4$. Invoking Lemma 20 we get that with probability at least $1-\delta$

$$\sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}\widehat{g}_{i,t}^2 \le \sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}\mathbb{E}_t[\widehat{g}_{i,t}^2] + \sqrt{2\ln\left(\frac{1}{\delta}\right)\sum_{t=1}^{T}\left(\frac{4\beta^2\,\mathtt{mas}^4(G_t)}{\nu_t^2}+\frac{3\,\mathtt{mas}^3(G_t)}{\nu_t}\right)} + \frac{(1+\beta)^2}{2}\ln\left(\frac{1}{\delta}\right)\max_{t=1\ldots T}\frac{\mathtt{mas}^2(G_t)}{\nu_t}. \tag{19}$$

Finally, from $g_{i,t}\le 1$, Lemma 18, item 1, and the assumptions of this lemma, we have

$$\begin{aligned}
\sum_{i=1}^{K} p_{i,t}\mathbb{E}_t[\widehat{g}_{i,t}^2] &\le \sum_{i=1}^{K} p_{i,t}\sum_{j=1}^{K} p_{j,t}\left(\frac{\mathbb{I}\{j\xrightarrow{t} i\}+\beta}{q_{i,t}}\right)^2 \\
&= \beta^2\sum_{i=1}^{K}\frac{p_{i,t}}{(q_{i,t})^2} + (2\beta+1)\sum_{i=1}^{K} p_{i,t}\sum_{j\,:\,j\xrightarrow{t} i}\frac{p_{j,t}}{(q_{i,t})^2} \\
&= \beta^2\sum_{i=1}^{K}\frac{p_{i,t}}{(q_{i,t})^2} + (2\beta+1)\sum_{i=1}^{K}\frac{p_{i,t}}{q_{i,t}} \\
&\le \frac{\beta^2\,\mathtt{mas}^2(G_t)}{\nu_t} + (2\beta+1)\,\mathtt{mas}(G_t) \\
&\le \frac{\beta^2\,\mathtt{mas}^2(G_t)}{\nu_t} + 2\,\mathtt{mas}(G_t),
\end{aligned}$$

where we used again $\beta\le 1/4$. Plugging this back into (19) the result follows. $\square$

LEMMA 23. *Suppose that the ELP.P algorithm is run with $\beta\le 1/4$. Then it holds with probability at least $1-\delta$ that for any $i=1,\ldots,K$,*

$$\sum_{t=1}^{T}\widehat{g}_{i,t} \ge \sum_{t=1}^{T} g_{i,t} - \frac{\ln(K/\delta)}{\beta}.$$

Proof. The lemma, including its proof, is very similar to the one used in the analysis of the standard Exp3.P algorithm (see [17, Lemma 6.7]) and is provided here for completeness. Let $\lambda>0$ be a parameter to be specified later. Since $\exp(x)\le$

$1 + x + x^2$ for $x \le 1$, we have by definition of $\widehat{g}_{i,t}$ that

$$\begin{aligned}
&\mathbb{E}_t\Big[\exp\left(\lambda(g_{i,t} - \widehat{g}_{i,t})\right)\Big] \\
&\quad = \mathbb{E}_t\left[\exp\left(\lambda\left(g_{i,t} - \frac{g_{i,t}\mathbb{I}\{I_t \xrightarrow{t} i\}}{q_{i,t}}\right) - \frac{\beta\lambda}{q_{i,t}}\right)\right] \\
&\quad \le \left(1 + \mathbb{E}_t\left[\lambda\left(g_{i,t} - \frac{g_{i,t}\mathbb{I}\{I_t \xrightarrow{t} i\}}{q_{i,t}}\right)\right] + \mathbb{E}_t\left[\left(\lambda\left(g_{i,t} - \frac{g_{i,t}\mathbb{I}\{I_t \xrightarrow{t} i\}}{q_{i,t}}\right)\right)^2\right]\right) \\
&\qquad\qquad \times \exp\left(-\frac{\beta\lambda}{q_{i,t}}\right) \\
&\quad \le \left(1 + 0 + \lambda^2\mathbb{E}_t\left[\left(\frac{g_{i,t}\mathbb{I}\{I_t \xrightarrow{t} i\}}{q_{i,t}}\right)^2\right]\right)\exp\left(-\frac{\beta\lambda}{q_{i,t}}\right) \\
&\quad \le \left(1 + \lambda^2 \sum_{j\,:\,j \xrightarrow{t} i} \frac{p_{j,t}}{(q_{i,t})^2}\right)\exp\left(-\frac{\beta\lambda}{q_{i,t}}\right) \\
&\quad = \left(1 + \frac{\lambda^2}{q_{i,t}}\right)\exp\left(-\frac{\beta\lambda}{q_{i,t}}\right).
\end{aligned}$$

Picking $\lambda = \beta$ and using the fact that $(1+x)\exp(-x) \le 1$, we get that this expression is at most 1. As a result, we have

$$\mathbb{E}_t\left[\exp\left(\lambda\sum_{t=1}^{T}(g_{i,t} - \widehat{g}_{i,t})\right)\right] \le 1.$$

This holds for the conditional expectation $\mathbb{E}_t$. Taking expectations, we can remove the conditioning and get that

$$\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(g_{i,t} - \widehat{g}_{i,t})\right)\right] \le 1.$$

Now, by a standard Chernoff technique, we know that for any $\lambda > 0$,

$$\mathbb{P}\left(\sum_{t=1}^{T}(g_{i,t} - \widehat{g}_{i,t}) > \epsilon\right) \le \exp(-\lambda\epsilon)\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T}(g_{i,t} - \widehat{g}_{i,t})\right)\right].$$

In particular, for our choice of λ, we get the bound

$$\mathbb{P}\left(\sum_{t=1}^{T}(g_{i,t} - \widehat{g}_{i,t}) > \epsilon\right) \le \exp\left(-\beta\epsilon\right).$$

Substituting $\delta = \exp(-\beta\epsilon)$, solving for ϵ, and using a union bound to make the result hold simultaneously for all i, the result follows. ∎

Proof of Theorem 9. With these key lemmas in hand, we can now prove Theorem 9. We have

$$\frac{W_{t+1}}{W_t} = \sum_{i\in V}\frac{w_{i,t+1}}{W_t} = \sum_{i\in V}\frac{w_{i,t}}{W_t}\exp(\eta\widehat{g}_{i,t}). \tag{20}$$

Now, by definition of $q_{i,t}$ and ν_t in Algorithm 3 we have

$$q_{i,t} \geq \nu_t \sum_{j\,:\,j \xrightarrow{t} i} s_{j,t} \geq (1+\beta)\,\eta$$

for all $i \in V$, so that

$$\eta \widehat{g}_{j,t} \;\leq\; \eta \max_{i \in V} \left(\frac{1+\beta}{q_{i,t}} \right) \;\leq\; 1.$$

Using the definition of $p_{i,t}$ and the inequality $\exp(x) \leq 1 + x + x^2$ for any $x \leq 1$, we can then upper bound the right-hand side of (20) by

$$\sum_{i \in V} \frac{p_{i,t} - \nu_t s_{i,t}}{1-\nu_t} \left(1 + \eta \widehat{g}_{i,t} + \eta^2 \widehat{g}_{i,t}^2\right) \;\leq\; 1 + \frac{\eta}{1-\nu_t} \sum_{i \in V} p_{i,t} \widehat{g}_{i,t} + \frac{\eta^2}{1-\nu_t} \sum_{i=1}^{K} p_{i,t} \widehat{g}_{i,t}^2.$$

Taking logarithms and using the fact that $\ln(1+x) \leq x$, we get

$$\ln \left(\frac{W_{t+1}}{W_t} \right) \;\leq\; \frac{\eta}{1-\nu_t} \sum_{i \in V} p_{i,t} \widehat{g}_{i,t} + \frac{\eta^2}{1-\nu_t} \sum_{i \in V} p_{i,t} \widehat{g}_{i,t}^2.$$

Summing over all t, and canceling the resulting telescopic series, we get

$$\ln \left(\frac{W_{T+1}}{W_1} \right) \;\leq\; \sum_{t=1}^{T} \sum_{i \in V} \frac{\eta}{1-\nu_t} p_{i,t} \widehat{g}_{i,t} + \sum_{t=1}^{T} \sum_{i \in V} \frac{\eta^2}{1-\nu_t} p_{i,t} \widehat{g}_{i,t}^2. \tag{21}$$

Also, we have

$$\ln \left(\frac{W_{T+1}}{W_1} \right) \geq \ln \left(\frac{\max_k w_{k,T+1}}{W_1} \right) = \eta \cdot \max_k \sum_{t=1}^{T} \widehat{g}_{k,t} - \ln K. \tag{22}$$

Combining (21) with (22) and slightly rearranging and simplifying, we get

$$\begin{aligned} &\max_k \sum_{t=1}^{T} \widehat{g}_{k,t} - \sum_{t=1}^{T} \sum_{i \in V} p_{i,t} \widehat{g}_{i,t} \\ &\leq \frac{\ln K}{\eta} + \frac{\eta}{1 - \max\limits_{t=1,\ldots,T} \nu_t} \sum_{t=1}^{T} \sum_{i \in V} p_{i,t} \widehat{g}_{i,t}^2 + \frac{1}{1 - \max\limits_{t=1,\ldots,T} \nu_t} \sum_{t=1}^{T} \sum_{i \in V} \nu_t p_{i,t} \widehat{g}_{i,t}. \end{aligned} \tag{23}$$

In what follows, we apply the various lemmas, using a union bound. To keep things manageable, we will use asymptotic notation to deal with second-order terms. In particular, we will use $\widetilde{\mathcal{O}}$ notation to hide numerical constants and logarithmic factors.[13] Note that by definition of β and ν_t, as well as Corollary 19, it is easy to verify that

$$\beta = \widetilde{\mathcal{O}}(\eta), \qquad \nu_t = \widetilde{\mathcal{O}}(\eta\, \mathtt{mas}(G_t)), \qquad \nu_t \in \left[\eta, \frac{1}{2}\right]. \tag{24}$$

[13] Technically, $\widetilde{\mathcal{O}}(f) = O(f \log^{O(1)} f)$. In our $\widetilde{\mathcal{O}}$ we also ignore factors that depend logarithmically on K and $1/\delta$.

Specifically, the bound for β is by definition, and the bound for ν_t holds since by Lemma 17 and the assumptions that $\eta \le 1/(3K)$ and $\beta \le 1/4$ we have

$$\begin{aligned}\nu_t &= \frac{(1+\beta)\eta}{\min_{i\in V}\sum_{j\,:\,j\xrightarrow{t}i} s_{j,t}}\\ &\le (1+\beta)\eta\,\mathtt{mas}(G_t)\\ &\le \frac{(1+\beta)\max(G_t)}{3K}\\ &\le \frac{1+1/4}{3} < 1/2.\end{aligned}$$

We now collect the main components required for the proof. First, by Lemma 21, we have with probability at least $1-\delta$ that

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}\widehat{g}_{i,t} \le{}& \sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}g_{i,t}\\ &+\sqrt{2\ln\left(\frac{1}{\delta}\right)\sum_{t=1}^{T}\mathtt{mas}(G_t)}\\ &+\beta\sum_{t=1}^{T}\mathtt{mas}(G_t)\\ &+\widetilde{\mathcal{O}}\left(\max_{t=1\ldots T}\ \mathtt{mas}(G_t)\right).\end{aligned} \tag{25}$$

Moreover, by Azuma's inequality, we have with probability at least $1-\delta$ that

$$\sum_{t=1}^{T}\sum_{i=1}^{K} p_{i,t}g_{i,t} \le \sum_{t=1}^{T} g_{I_t,t} + \sqrt{\frac{\ln(1/\delta)}{2}\,T}. \tag{26}$$

Second, again by Lemma 21 and the conditions (24), we have with probability at least $1-\delta$ that

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{K}\nu_t p_{i,t}\widehat{g}_{i,t} &\le \sum_{t=1}^{T}\sum_{i=1}^{K}\nu_t p_{i,t}g_{i,t} + \widetilde{\mathcal{O}}\left(\max_{t=1,\ldots,T}\mathtt{mas}^2(G_t)(1+\sqrt{T}\eta+T\eta^2)\right)\\ &\le \sum_{t=1}^{T}\nu_t + \widetilde{\mathcal{O}}\left(\max_{t=1,\ldots,T}\mathtt{mas}^2(G_t)(1+\sqrt{T}\eta+T\eta^2)\right).\end{aligned} \tag{27}$$

Third, by Lemma 22, and conditions (24), we have with probability at least $1-\delta$ that for all i,

$$\sum_{t=1}^{T} p_{i,t}\widehat{g}_{i,t}^2 \le 2\sum_{t=1}^{T}\mathtt{mas}(G_t) + \left(\max_{t=1,\ldots,T}(\mathtt{mas}^2(G_t))\right)\widetilde{\mathcal{O}}\left(T\eta+\frac{1}{\eta}+\sqrt{T\left(1+\frac{1}{\eta}\right)}\right). \tag{28}$$

Fourth, by Lemma 23, we have with probability at least $1-\delta$ that

$$\max_k \sum_{t=1}^{T}\widehat{g}_{k,t} \ge \max_k \sum_{t=1}^{T} g_{k,t} - \frac{\ln(K/\delta)}{\beta}. \tag{29}$$

Combining (25), (26), (27), (28), and (29) with a union bound (i.e., replacing δ by $\delta/5$), substituting back into (23), and slightly simplifying, we get that with probability at least $1-\delta$, $\max_k \sum_{t=1}^T (g_{k,t} - g_{I_t,t})$ is at most

$$\sqrt{2\ln\left(\frac{5}{\delta}\right)\sum_{t=1}^T \mathtt{mas}(G_t)} + \beta\sum_{t=1}^T \mathtt{mas}(G_t) + \sqrt{\frac{\ln(5/\delta)}{2}T} + \frac{\ln(5K/\delta)}{\beta} + \frac{\ln K}{\eta}$$
$$+ 4\eta\sum_{t=1}^T \mathtt{mas}(G_t) + 2\sum_{t=1}^T \nu_t + (1+\sqrt{T\eta}+T\eta^2)\,\widetilde{\mathcal{O}}\left(\max_{t=1,\dots,T}(\mathtt{mas}^2(G_t))\right).$$

Substituting in the values of β and ν_t, overapproximating, and simplifying (in particular, using Corollary 19 to upper bound ν_t by $(1+\beta)\,\eta\,\mathtt{mas}(G_t))$, we get the upper bound

$$\sqrt{5\ln\left(\frac{5}{\delta}\right)\sum_{t=1}^T \mathtt{mas}(G_t)} + \frac{2\ln(5K/\delta)}{\eta} + 12\eta\sqrt{\frac{\ln(5K/\delta)}{\ln K}}\sum_{t=1}^T \mathtt{mas}(G_t)$$
$$+ \widetilde{\mathcal{O}}(1+\sqrt{T\eta}+T\eta^2)\left(\max_{t=1,\dots,T}(\mathtt{mas}^2(G_t))\right).$$

In particular, by picking η such that

$$\eta^2 = \frac{1}{6}\,\frac{\sqrt{\ln(5K/\delta)\,(\ln K)}}{\sum_{t=1}^T m_t},$$

noting that this implies $\eta = \widetilde{\mathcal{O}}(1/\sqrt{T})$, and overapproximating once more, we get the following bound on $\max_k \sum_{t=1}^T (g_{k,t} - g_{I_t,t})$:

$$10\,\frac{\ln^{1/4}(5K/\delta)}{\ln^{1/4} K}\sqrt{\ln\left(\frac{5K}{\delta}\right)\sum_{t=1}^T m_t} + \widetilde{\mathcal{O}}(T^{1/4})\left(\max_{t=1,\dots,T}\mathtt{mas}^2(G_t)\right).$$

To conclude, we simply plug in $\ell_{i,t} = 1 - g_{i,t}$ for all i and t, thereby obtaining the claimed results.

Acknowledgment. We warmly thank the anonymous reviewers for their detailed and constructive comments, which greatly helped to improve the presentation of this paper.

REFERENCES

[1] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren, *Online learning with feedback graphs: Beyond bandits*, in Proceedings of the 28th Conference on Learning Theory, COLT 2015, Paris, France, 2015, pp. 23–25.

[2] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren, *Online Learning with Feedback Graphs: Beyond Bandits*, preprint, https://arxiv.org/abs/1502.07617, 2015.

[3] N. Alon, N. Cesa-Bianchi, C. Gentile, and Y. Mansour, *From bandits to experts: A tale of domination and independence*, in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS 2012), Lake Tahoe, NV, 2013, pp. 1610–1618.

[4] N. Alon and J. H. Spencer, *The Probabilistic Method*, John Wiley and Sons, New York, 2004.

[5] J.-Y. Audibert and S. Bubeck, *Minimax policies for adversarial and stochastic bandits*, in Proceedings of COLT 2009, Montreal, Canada, 2009, online proceedings.

[6] J.-Y. Audibert and S. Bubeck, *Regret bounds and minimax policies under partial monitoring*, J. Mach. Learn. Res., 11 (2010), pp. 2785–2836.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, *The nonstochastic multiarmed bandit problem*, SIAM J. Comput., 32 (2002), pp. 48–77, https://doi.org/10.1137/S0097539701398375.

[8] G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári, *Partial monitoring—classification, regret bounds, and algorithms*, Math. Oper. Res., 39 (2014), pp. 967–997.

[9] S. Bubeck and N. Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, Found. Trends Mach. Learn., 5 (2012), pp. 1–122.

[10] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff, *Stochastic bandits with side observations on networks*, ACM SIGMETRICS Perform. Eval. Rev., 42 (2014), pp. 289–300.

[11] Y. Caro, *New Results on the Independence Number*, Tech. report, Tel Aviv University, Tel Aviv, Israel, 1979.

[12] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat, *Leveraging Side Observations in Stochastic Bandits*, preprint, https://arxiv.org/abs/1210.4839, 2012.

[13] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat, *Leveraging side observations in stochastic bandits*, in Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI), Catalina Island, CA, 2012, pp. 142–151.

[14] A. Carpentier and M. Valko, *Revealing graph bandits for maximizing local influence*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, Cadiz, Spain, 2016, pp. 10–18.

[15] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, *How to use expert advice*, J. ACM, 44 (1997), pp. 427–485.

[16] N. Cesa-Bianchi, P. Gaillard, C. Gentile, and S. Gerchinovitz, *Algorithmic chaining and the role of partial feedback in online nonparametric learning*, in Proceedings of the 30th Conference on Learning Theory (COLT), JMLR Workshop and Conference Proceedings Volume 65, Amsterdam, The Netherlands, 2017, pp. 465–481.

[17] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, Cambridge, UK, 2006.

[18] N. Cesa-Bianchi and G. Lugosi, *Combinatorial bandits*, J. Comput. System Sci., 78 (2012), pp. 1404–1422.

[19] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz, *Improved second-order bounds for prediction with expert advice*, in Proceedings of the 18th Annual Conference on Learning Theory, Bertinoro, Italy, 2005, pp. 217–232.

[20] V. Chvatal, *A greedy heuristic for the set-covering problem*, Math. Oper. Res., 4 (1979), pp. 233–235.

[21] A. Cohen, T. Hazan, and T. Koren, *Online Learning with Feedback Graphs without the Graphs*, preprint, https://arxiv.org/abs/1605.07018, 2016.

[22] O. Dekel, A. Tewari, and R. Arora, *Online bandit learning against an adaptive adversary: From regret to policy regret*, in Proceedings of ICML 2012, Edinburgh, Scotland, 2012, pp. 1503–1510.

[23] D. Freedman, *On tail probabilities for martingales*, Ann. Probab., 3 (1975), pp. 100–118.

[24] Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, in Proceedings of the 2nd Annual Conference on Computational Learning Theory (Euro-COLT '95) (Barcelona, 1995), J. Comput. System Sci., 55 (1997), pp. 119–139.

[25] A. Frieze and C. McDiarmid, *Algorithmic theory of random graphs*, Random Structures Algorithms, 10 (1997), pp. 5–42.

[26] A. M. Frieze, *On the independence number of random graphs*, Discrete Math., 81 (1990), pp. 171–175.

[27] M. K. Hanawal and V. Saligrama, *Cost effective algorithms for spectral bandits*, in Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton 2015), Monticello, IL, 2015, pp. 1323–1329.

[28] A. Kalai and S. Vempala, *Efficient algorithms for online decision problems*, J. Comput. System Sci., 71 (2005), pp. 291–307.

[29] D. Kempe, J. Kleinberg, and É. Tardos, *Maximizing the spread of influence through a social network*, in Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 2003, pp. 137–146.

[30] D. Kempe, J. Kleinberg, and É. Tardos, *Maximizing the spread of influence through a social network*, Theory Comput., 11 (2015), pp. 105–147.

[31] T. Kocák, G. Neu, and M. Valko, *Online learning with Erdős-Rényi side-observation graphs*, in Proceedings of the Conference on Uncertainty in Artificial Intelligence, New York, 2016, pp. 339–346.

[32] T. Kocák, G. Neu, and M. Valko, *Online learning with noisy side observations*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, Cadiz, Spain, 2016, pp. 1186–1194.

[33] T. Kocák, G. Neu, M. Valko, and R. Munos, *Efficient learning by implicit exploration in bandit problems with side observations*, in Proceedings of the Conference on Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 613–621.

[34] J. Langford and T. Zhang, *The epoch-greedy algorithm for multi-armed bandits with side information*, in Proceedings of the Conference on Advances in Neural Information Processing Systems, Vancouver, Canada, 2008, pp. 817–824.

[35] N. Littlestone and M. K. Warmuth, *The weighted majority algorithm*, Inform. Comput., 108 (1994), pp. 212–261.

[36] O. Maillard and R. Munos, *Adaptive bandits: Towards the best history-dependent strategy*, in Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR Workshop and Conference Proceedings Volume 15, Fort Lauderdale, FL, 2011, pp. 570–578.

[37] S. Mannor and O. Shamir, *From bandits to experts: On the value of side-observations*, in Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS 2011), Granada, Spain, 2011, pp. 684–692.

[38] G. Neu, *Explore no more: Improved high-probability regret bounds for non-stochastic bandits*, in Proceedings of the Conference on Advances in Neural Information Processing Systems, Montreal, Canada, 2015, pp. 3150–3158.

[39] V. Perchet and P. Rigollet, *The multi-armed bandit problem with covariates*, Ann. Statist., 41 (2013), pp. 693–721.

[40] P. Rusmevichientong and J. Tsitsiklis, *Linearly parameterized bandits*, Math. Oper. Res., 35 (2010), pp. 395–411.

[41] A. Said, E. W. De Luca, and S. Albayrak, *How social relationships affect user similarities*, in Proceedings of the International Conference on Intelligent User Interfaces Workshop on Social Recommender Systems, Hong Kong, 2010, online proceedings.

[42] A. Slivkins, *Contextual bandits with similarity information*, in Proceedings of the 24th Annual Conference on Learning Theory (COLT), JMLR Workshop and Conference Proceedings Volume 19, Budapest, Hungary, 2011, pp. 679–702.

[43] G. Stoltz, *Information Incomplète et Regret Interne en Prédiction de Suites Individuelles*, Ph.D. thesis, Université Paris-XI Orsay, Orsay, France, 2005.

[44] M. Valko, R. Munos, B. Kveton, and T. Kocák, *Spectral bandits for smooth graph functions*, in Proceedings of the International Conference on Machine Learning, Beijing, China, 2014, pp. 46–54.

[45] V. G. Vovk, *Aggregating strategies*, in Proceedings of the Conference on Learning Theory (COLT), Rochester, NY, 1990, pp. 371–386.

[46] C. Wang, S. Kulkarni, and H. Poor, *Arbitrary side observations in bandit problems*, Adv. Appl. Math., 34 (2005), pp. 903–938.

[47] C. Wang, S. Kulkarni, and H. Poor, *Bandit problems with side observations*, IEEE Trans. Automat. Control, 50 (2005), pp. 338–355.

[48] V. K. Wey, *A Lower Bound on the Stability Number of a Simple Graph*, Bell Lab. Tech. Memo 81-11217-9, 1981.

[49] Y. Wu, A. György, and C. Szepesvári, *Online learning with Gaussian payoffs and side observations*, in Proceedings of the Conference on Advances in Neural Information Processing Systems, Montreal, Canada, 2015, pp. 1360–1368.

[50] N. Zolghadr, G. Bartók, R. Greiner, A. György, and C. Szepesvári, *Online learning with costly features and labels*, in Proceedings of the Conference on Advances in Neural Information Processing Systems, Lake Tahoe, NV, 2013, pp. 1241–1249.