6. APPENDIX

6.1 Modeling the Organizer's Objective using Taylor Expansion

We describe how to model the organizer's objective function based on the Taylor approximation of the sample variance, which we refer to as (`OptIKA-L2T`). Let $v_{j,i}$ be a binary variable, which is 1 if and only if $l_i \in L_j$. In this case, $V_i = \sum_{j=1}^{m} v_{j,i} = Y_i - X_{0,i}$. Because $\mathbf{X}$ is the total number of historical visits, and $\mathbf{V} = \mathbf{Y} - \mathbf{X}$ is the net amount of visits in a short time window, it is often the case that $||\mathbf{X}||_2 \gg ||\mathbf{V}||_2$. In this case, we further approximate the sample variance using first order Taylor expansion. Let $\sigma(\mathbf{Y})$ be the sample variance: $\sigma(\mathbf{Y}) = \frac{1}{n}||\mathbf{Y} - \overline{\mathbf{Y}}||_2^2$. Expanding $\sigma(\mathbf{Y})$ at the point $\mathbf{Y} = \mathbf{X}$, we obtain: $\sigma(\mathbf{Y}) \approx \sigma(\mathbf{X}) + \bigtriangledown\sigma(\mathbf{X})^T \cdot (\mathbf{Y} - \mathbf{X})$,where $\bigtriangledown\sigma = \left(\frac{\partial\sigma}{\partial V_1}, \frac{\partial\sigma}{\partial V_2}, \ldots, \frac{\partial\sigma}{\partial V_n}\right)$, $\frac{\partial\sigma}{\partial V_i}(\mathbf{Y}) = \frac{2}{n}\left(Y_i - \overline{Y}\right)$.

With $\mathbf{V} = \mathbf{Y} - \mathbf{X}$ and $V_i = \sum_{j=1}^{m} v_{j,i}$, we have: $\sigma(\mathbf{Y}) \approx \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{2}{n}\left(X_{0,i} - \overline{X_0}\right) \cdot v_{j,i} + constant$.

Let $s_i = \frac{2}{n}\left(X_{0,i} - \overline{X_0}\right)$, then the organizer's problem is to minimize $S = \sum_{j=1}^{m}\sum_{i=1}^{n} s_i \cdot v_{j,i} = \sum_{i=1}^{n} s_i V_i$.

When the organizer's objective is taking this specific form, the optimal incentive allocation problem is related to the Stackelberg pricing games [1, 18], in which a monopoly (or *leader*) sets the price of multiple products so as to maximize its revenue from various consumers (or *followers*). In [18], it has been proven that the optimal incentives problem is APX-hard when each follower can purchase at most one item, and the seller can only set two possible prices.

6.2 Proofs for Theorems in Main Text

THEOREM 2. *(Weak Duality) The optimal objective value to the problem* $\max_{\{\boldsymbol{\lambda}_j\}}\ g(\{\boldsymbol{\lambda}_j\})$ *is a lower bound on the optimal value of the global problem (**OptIKA**).*

PROOF. Let OPT be the optimal value of the original global problem. Fix a set of $\{\boldsymbol{\lambda}_j | j = 1 \ldots m\}$, we must have $g(\{\boldsymbol{\lambda}_j\}) \leq OPT$, since by definition, g is the minimal value for $L_\rho(.)$, subject to the constraint $(\mathbf{v}_j, \mathbf{r}_j) \in \Sigma_j$ for $j = 1 \ldots m$, while OPT is also the minimal value for $L_\rho(.)$, but subject to $(\mathbf{v}_j, \mathbf{r}_j) \in \Sigma_j$ and an extra constraint $\mathbf{r}_1 = \mathbf{r}_2 = \cdots = \mathbf{r}_m = \mathbf{r}$. Because $g(\{\boldsymbol{\lambda}_j\}) \leq OPT$ for all possible choices of $\{\boldsymbol{\lambda}_j\}$, $\max_{\{\boldsymbol{\lambda}_j\}} g(\{\boldsymbol{\lambda}_j\}) \leq OPT$. □

THEOREM 3. $g(\{\boldsymbol{\lambda}_j\})$ *is concave for* $\{\boldsymbol{\lambda}_j | j = 1 \ldots m\}$.

PROOF. Fix $\mathbf{r}$ and $\mathbf{r}_j$, $\mathbf{v}_j$ is determined by $\mathbf{r}_j$ and the constraint $(\mathbf{v}_j, \mathbf{r}_j) \in \Sigma_j$, thus $L_\rho(.)$ becomes an affine function w.r.t. $\boldsymbol{\lambda}_j$. Let $\mathbf{r}$ and $\mathbf{r}_j$ range over its domain, $g(.)$ is a pointwise infimum of affine functions, thus $g(.)$ is concave. □

THEOREM 4. *If* `OptIKA-L2T-ADMM` *converges, then* $\mathbf{r}_1, \ldots \mathbf{r}_m$ *and* $\mathbf{r}$ *all converge to the same vector.*

PROOF. The proof follows from the fact that the gradient for all the $\boldsymbol{\lambda}_j$ will converge to zero. □

6.3 The Power of Uniform Sampling

To show the benefit of having a uniform training dataset, we choose one flag species: the White-throated Sparrow, a migrating species that mostly appears in Southern counties of the New York State during the winter, and migrates North as the weather becomes warmer. More interestingly, White-throated Sparrows stay in the Adirondack Mountains mostly during the summer, but there are few *eBird* submissions in that region, since it is not easily accessible. This biased sampling effort makes it challenging for a machine learning model to capture the distribution of this species.

We learn the monthly distribution of this species using *eBird* submissions in New York State. We fit a machine learning model to predict a binary label on the existence of this species, based on a set of environmental covariates. We set aside 10% of observations as the test set, and use the remaining observations as training set (the *Complete* set). To show the impact of uniform sampling, we artificially create three smaller training sets by subsampling about 5% of the *Complete* set in the following way: *(1) Grid*: Spatially-uniform sampling, i.e. lay down a grid on the study area and sample one submission from each spatial grid cell; *(2) Urban*: Sampling among submissions with a higher-than-median urban environmental component; and *(3) Subsample*: Random sampling among all submissions.

The *Grid* dataset demonstrates the *ideal* case, in which citizen scientists' effort is uniformly spread out geographically, while the *Urban* dataset is used to reflect the *worst* case, in which citizen scientists' effort is restricted by their daily activities to the urban areas. For comparison purposes, we restrict the sizes of the three datasets to be equal. We then train a random forest model with 1,000 trees with maximal depth of 10 on these three datasets as well as on the *Complete* training set. Table 4 reports the *Averaged Log-loss* score on the same separate test set. The averaged log-loss is defined as: $-\sum_i \log Pr(y_i^t | y_i^p) = -\sum_i y_i^t \log(y_i^p) + (1 - y_i^t)\log(1 - y_i^p)$, in which y_i^t is the true label for $\mathbf{x}_i$, and y_i^p is the predicted probability for $\mathbf{x}_i$ to be 1. The smaller the log-loss, the better fit of the machine learning model.

In addition, we plot the distributions learned from the four datasets as heatmaps in Fig. 6. These heatmaps show the predicted probabilities at the different locations of the study area. As depicted in the figure, there are significant differences in the learned spatial distributions of the White-throated Sparrow. As we can see, the performance of the *Urban* dataset is always the worst. This is in line with our expectations, because most observations in this dataset come from the urban area, while the White-throated Sparrow mostly resides in forests. Also, even though the *Grid* dataset has only 5% of the data compared to the *Complete* dataset, their performance is comparable. Moreover, the model trained on the *Grid* dataset performs better than the one trained on the *Subsample* dataset. This clearly suggests a potentially huge saving in citizen scientists' effort, if they can follow a more uniform protocol to sample the entire area, just as in the *Grid* case. In other words, if the total effort citizen scientists can devote to the current project is bounded by their participation rate, then their effort should be spent more wisely to focus on under-sampled areas.

6.4 Simulation

We present additional simulation results based on the current *Avicache* participants, whose behavior parameters are learned based on the first two months' field data (spring migration period). In one round, we act as the organizer, who uses **OptIKA** to set the rewards, and a set of virtual agents in turn act to optimize for their own Knapsack problems to visit different locations. We simulate this process for multiple rounds, and observe how effective the incentive mechanism is in terms of driving people to under-sampled

Month	Num Obs (*Grid*, *Urban*, *Subsample*)	Log-loss (*Grid*)	Log-loss (*Urban*)	Log-loss (*Subsample*)	Num Obs (*Complete*)	Log-loss (*Complete*)
April	796	0.40	0.55	0.44	34778	0.35
May	937	0.37	0.44	0.39	38075	0.34
June	901	0.18	0.30	0.20	22664	0.16
July	831	0.21	0.94	0.28	19112	0.18
August	774	0.10	0.13	0.15	21719	0.09
September	737	0.25	0.35	0.27	25218	0.23

Table 4: The performance of the random forest classifier for different training sets.

areas.

Moreover, once an agent decides to visit a location during the simulation, we sample one observation from the real *eBird* dataset, as if the agent submits this observation into the database. At the end of each round, we fit a predictive model based on all the data virtually collected so far, to see how much the species distribution model can be affected by agents' shifts of exploration efforts.

As we are restricted to subsample from the current *eBird* dataset, we start with a small dataset of 100 observations, subsampled with bias from the popular birding sites. This resembles the very early stage of *eBird*, where we have very few and highly biased data. Our agents in the simulation are all participants currently in *avicache*.

The left plot of Fig. 5 illustrates the sample variance D_2 as a function of the number of iterations with and without extra incentives. We offer four levels of incentives: 0, 4, 8, 12 points. We observe that, once incentives are introduced, people are pushed to under-sampled areas. The right plot of Fig. 5 shows the variation of the Log-loss of the predictive model at the end of the first few iterations. Here we use the collected data to predict the occurrence of the Horned Lark in spring. As we can see, without incentives (blue line) the Logloss of the predictive model stays at a high level, despite more data is collected by citizen scientists. On the other hand, the predicted performance quickly improves once we introduce external rewards (red line). The green dashed line shows the Log-loss of the predictive model fit with all available data in these two counties, which should be thought as the best we can do given the currently available data.

In summary, our simulation shows that our incentive scheme steers *eBird* participants to under-sampled area, and at the same time improves the prediction performance of the occupancy model for an interesting species.