Efficient Localization of Multiple Intruders in Shared Spectrum System

Caitao Zhan
Stony Brook University
cbzhan@cs.stonybrook.edu

Himanshu Gupta
Stony Brook University
hgupta@cs.stonybrook.edu

Arani Bhattacharya
KTH Royal Institute of Technology
aranib@kth.se

Mohammad Ghaderibaneh
Stony Brook University
mghaderibane@cs.stonybrook.edu

ABSTRACT

We address the problem of localizing multiple intruders (unauthorized transmitters) using a distributed set of sensors in the context of a shared spectrum system. In contrast to single transmitter localization, multiple transmitter localization (MTL) has not been thoroughly studied. In shared spectrum systems, it is important to be able to localize simultaneously present multiple intruders to effectively protect a shared spectrum from malware-based, jamming, or other multi-device unauthorized-usage attacks. The key challenge in solving the MTL problem comes from the need to "separate" an aggregated signal received from multiple intruders into separate signals from individual intruders. Furthermore, in a shared spectrum paradigm, presence of an evolving set of authorized users (e.g., primary and secondary users) adds to the challenge.

In this paper, we propose an efficient algorithm for the MTL problem based on the hypothesis-based Bayesian approach called MAP. Direct application of the MAP approach to the MTL problem incurs prohibitive computational and training cost. In this work, we develop optimized techniques based on MAP with significantly improved computational and training costs. In particular, we develop a novel interpolation method, ILDW, which helps minimize the training cost. We generalize our techniques via online-learning to the setting wherein there may be a set of dynamically-changing authorized users present in the background. We evaluate our developed techniques on large-scale simulations as well as on small-scale indoor and outdoor testbeds. Our experiments demonstrate that our technique outperforms the prior approaches by significant margins, i.e., error up to 74% less in large-scale simulations and 30% less in real-world testbeds.

CCS CONCEPTS

• **Networks** → **Location based services**; • **Security and privacy** → *Mobile and wireless security*;

KEYWORDS

localization, RF sensing, crowdsourced, shared spectrum

1 INTRODUCTION

The RF spectrum is a natural resource in great demand due to the unabated increase in mobile (and hence, wireless) data consumption [3]. The research community has addressed this capacity crunch via development of *shared spectrum paradigms*, wherein the spectrum is made available to unlicensed users (secondaries) as long as they do not interfere with the transmission of licensed incumbents (primaries). E.g., in the recent years, the FCC has made available the CBRS band, i.e., the 3550-3700 MHz band within the 3.5 GHz band, for shared commercial use to allow other users to utilize the otherwise low-usage band which was previously reserved for incumbent users including US Navy radar operators.

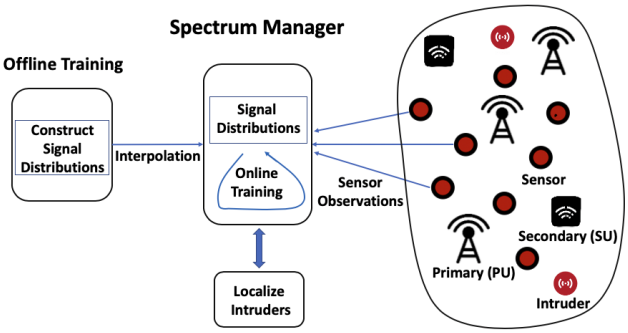


Figure 1: Overall approach to localize intruders in a shared spectrum system.

The increasing affordability of the software-defined radio (SDR) technologies makes the shared spectrums particularly prone to unauthorized usage or security attacks. With easy access to SDR devices [1, 2], it is easy for selfish users to transmit data on shared spectrum without any authorization and potentially causing harmful interference to the incumbent users. Such illegal spectrum usage could also happen as a result of infiltration of computer virus or malware on SDR devices. As the fundamental objective behind such shared spectrum paradigms is to maximize spectrum utilization, the viability of such systems depends on the ability to effectively guard the shared spectrum against unauthorized usage. The current mechanisms however to locate such unauthorized users (intruders) are human-intensive and time-consuming, involving FCC enforcement bureau which detects violations via complaints and manual investigation [18]. Motivated by above, we seek for an effective technique that is able to accurately localize multiple simultaneous intruders and even in the presence of dynamically changing set of authorized users. In the following, we begin with describing the multiple transmitter localization problem.

Multiple-Transmitter Localization (MTL). The transmitter localization problem has been well-studied, but most of the focus has been on localizing a *single* intruder at a time. However, it is important to localize *multiple* transmitters *simultaneously* to effectively guard a shared spectrum system. E.g., a malware or virus-based attachment could simultaneously cause many devices to violate spectrum allocation rules; spectrum jamming attacks would typically involve multiple transmitters. More importantly, a technique limited by localization of a single intruder could then be easily circumvented by an offender by using multiple devices. The key

challenge in solving the MTL problem comes from the fact that the deployed sensor would receive only a sum of the signals from multiple transmitters, and separating the signals may be impossible. In addition, the other challenge that MTL in the context of shared spectrum system poses is the presence of authorized users—e.g., the incumbent users and the dynamic set of secondary users that have been allocated spectrum by the manager. To the best our knowledge, no prior localization work has considered the presence of authorized users.

The state-of-the-art technique for the MTL problem is the recent work [18], which essentially decomposes the MTL problem to multiple single-transmitter localization problems based on the sensors with the highest power readings in a neighborhood. However, the technique has a few shortcomings: (i) it implicitly assumes a propagation model, and thus, may not work effectively in areas with complex propagation characteristics, (ii) it is not effective in the case of transmitters being located close-by, a key challenging scenario for MTL problem, and (iii) most importantly, it can't be extended effectively to incorporate background authorized users, a key requirement in the context of shared spectrum systems.

Our Approach. Transmitter localization is generally done based on observations at deployed sensors. In particular, as in prior works [7, 18], we assume a crowdsourced sensing architecture wherein relatively low-cost spectrum sensors are available for gathering signal strength in the form of received power. Our approach is a hypothesis-driven Bayesian approach, viz. *maximum a posteriori* (MAP) approach, wherein each hypothesis is a configuration (i.e. a combination of $\langle location, power \rangle$ pair) of the potential intruders, and the goal is to determine the hypothesis that best explains the sensor observations. This determination is done based on the distributions (gathered during a training phase) of sensor observations for each hypothesis. The MAP approach is known to have optimal classification accuracy, but (i) incurs prohibitive computation cost—exponential in number of potential intruders—when applied to the MTL problem, and (ii) requires significant amount of training cost. The focus of our work is to address these challenges, and design a viable MAP-based approach. In particular, using MAP as a building block, we develop an optimized approach that runs in polynomial time with minimized training cost. We extend our technique to work in presence of authorized users by incorporating online (real-time) training.

Motivation for MAP. Our motivation for using a MAP-based approach is multifold: First, with sufficient training data, MAP is known to deliver optimal classification accuracy for the MTL problem [11]. Second, the MAP approach doesn't assume any propagation model and thus works for arbitrary signal propagation characteristics. Third, it allows us to also estimate the intruder's transmit power, which can be very useful in some applications, e.g., where the penalty is proportional to the extent of violation. Last but not the least, it naturally extends to being able to handle a presence of an evolving set of authorized users.

Training Cost and Optimization. The benefits of a MAP-based approach come at a cost: the MAP framework requires prior training to build probability distributions (PDs) of sensor observations for each hypothesis. However, most of the training occurs offline, one-time, and can be automated e.g. via drones or robots. In our work, we develop strategies to minimize the training cost; in particular, we reduce the number of PDs to be constructed via a *novel interpolation scheme* suited to our unique setting, and evaluate the impact of reduced training on the localization accuracy. We note that the online training to incorporate presence of authorized users is needed only for the prevailing setting (of authorized transmitters and deployed sensors) and hence incurs minimal cost (see §4).

Overall Contributions. The goal of our work is to develop an efficient technique for accurate localization of simultaneously present multiple intruders in a shared spectrum system. The raw data are available at https://github.com/Wings-Lab/IPSN-2020-data. In this context, we make the following four specific contributions.

(1) Design an efficient localization algorithm (MAP*) for the MTL problem, based on an optimal hypotheses-driven Bayesian approach. The designed approach predicts both locations and transmit powers of the intruders, and does not assume any propagation model and thus, works for arbitrary signal propagation characteristics.
(2) Extend the designed algorithm (MAP**) to localize effectively in the presence of background authorized users, i.e., primaries with possibly unknown parameters (e.g., location and transmit power) and an evolving set of secondary users.
(3) Develop an effective interpolation scheme (ILDW) for our unique setting to reduce the one-time training cost of our scheme, without impacting the localization accuracy much.
(4) Evaluate our techniques via large-scale simulations as well as over two developed testbeds (indoor and outdoor), and demonstrate the effectiveness of our developed techniques and their superior performance compared to the best-known techniques.

2 PROBLEM, RELATED WORK, AND METHODOLOGY

In this section, we describe our model of the shared spectrum systems, formulate the MTL problem, and discuss related work. We also describe the building block of our approach, viz., a hypothesis-drived Bayesian localization approach (MAP).

Shared Spectrum System. In a shared spectrum paradigm, the spectrum is shared among licensed users (primary users, PUs) and unlicensed users (secondary users, SUs) in such a way that the transmission from secondaries does not interfere with that of the primaries (or secondaries from a higher-tier, in case of a multi-tier shared spectrum system [29]). In some shared spectrum systems, the location and transmit power of the primary users may be unavailable, as is the case with military or navy radars in the CBRS band [29]. Such sharing of spectrum is generally orchestrated by a centralized entity called *spectrum manager*, such as a spectrum database in TV white space [19] or a central spectrum access system in the CBRS 3.5GHz shared band [16]. The spectrum manager allocates spectrum to requesting secondaries (i.e., permission to transmit up to a certain transmit power at their location) based on their location, spectrum demand, configurations of the primaries, other active secondaries, prevailing channel conditions, etc.

Authorized and Unauthorized Users. Secondary users that have been explicitly given permission to transmit at their location are

termed as *authorized users*; the primaries users are also considered as authorized users. Note that the set of authorized users evolve over time, as more and more SUs are allocated spectrum and as some SUs stop using the spectrum after a while. We can assume that each SU is allocated spectrum for a certain duration of time, after which it stops using the spectrum. Other users that transmit without explicit permission (for that given time) are referred to as *unauthorized users* or *intruders*.

Problem Setting and Formal Definition. Consider a geographic area with a shared spectrum. Without loss of generality, we assume a single channel throughout this paper (multiple channels are handled similarly). For localization of unauthorized users, we assume available crowdsourced sensors that can observe received signal in the channel of interest, and compute (totel) received signal strength indicator (RSSI)[1]. These sensors, being crowdsourced, may be at different locations at different times. At any given instant, the shared spectrum area has some licensed primary users and some active secondary users; the PU configurations may not be known as can be the case for military users. The centralized spectrum manager is aware of the set of active SUs at any time, as each SU request is granted for a certain period of time. In addition to the authorized users, there may be a set of intruders present in the area with each intruder in a certain "configuration" (see §2.2).

The MTL problem is to determine the set of intruders with their configurations at each instant of time, based on the set of sensor observations at that instant. See Figure 1. The basic MTL problem assumes no other transmissions (of authorized users) in the background. The more general MTL problem, where there may be an evolving set of authorized users in the background, is referred to as the MTL-SS problem. We address the MTL problem in §3, and then address the more general MTL-SS problem in §4.

2.1 Related Work

Localization of an intruder in a field using sensor observations has been widely studied, but most of the works have focused on localization of a single intruder [6, 12]. In general, to localize multiple intruders, the main challenge comes from the need to "separate" powers at the sensors [24], i.e., to divide the total received power into power received from individual intruders. Blind source separation is a very challenging problem; only very limited settings allow for known techniques [20, 28] using sophisticated receivers. In our context of hypotheses-driven approach, the challenge of source separation manifests in terms of a large number of hypotheses, a challenge addressed in §3. We note that (indoor) localization of a device [4] based on signals received from multiple reference points (e.g, WiFi access points) is a quite different problem (see [30] for a recent survey), as the signals from reference points remain separate, and localization or tracking of multiple devices can be done independently. Recent works on multi-target localization/tracking are different in the way that targets are passive [9, 15, 17], instead of active transmitters in this work.

In absence of blind separation methods, to the best of our knowledge, only a few works have addressed multiple intruder(s) localization, and none of these consider it in the presence of a dynamically changing set of authorized transmitters. In particular, (i) [18] decomposes the multi-transmitter localization problem to multiple single-transmitter localization problems based on the sensors with highest of readings in a neighbohood, (ii) [22] works by clustering the sensors with readings above a certain threshold and then localizing intruders at the centers of these clusters, (iii) [23] uses an EM-based approach. The techniques of [18, 23] assume a propagation model, while that of [22, 23] require a priori knowledge of the number of intruders present. We have compared our approach with [18, 22] in §5, while [23] has high computational cost and has also been shown to be inferior in performance to [18, 22] even for a small number of intruders. Other related works include (i) [13] that addresses the challenge of handling time-skewed sensors observations in the MTL problem, and (ii) [5] that addresses the sensor selection optimization problem for our proposed hypotheses-based localization approach.

[1]We do not use angle-of-arrival (AoA) measurements [32] as they require additional and complex RF hardware.

2.2 MAP: Bayesian Approach for Localization

We localize intruders based on observations from a set of sensors. Each sensor communicates its observation to a centralized entity, the spectrum manager, which runs an appropriate localization algorithm to localize the intruders. In particular, we use a hypotheses-driven Bayesian approach, as described below, where intruders are localized by determining the most-likely prevailing hypothesis; this is done based on joint probability distributions of the sensors' observations (constructed during a priori training). Below, we formalize the above concepts, and the basic localization approach.

Observation; Observation Vector. Throughout this paper, we use the term *observation* at an individual sensor to mean the received power over a time window of certain duration, in the frequency channel of interest (we assume only one channel). In particular, received power is computed from the FFT of the I/Q samples in the time window [6]. We use the term *observation vector* $\mathbf{x}$ to denote a vector of observations from a given set of distributed sensors, with each vector dimension corresponding to a unique sensor.

Hypotheses. Let H_0, H_1, ..., H_m be the set of all hypotheses, where each hypothesis H_j represents a "configuration" of potential intruders. In this paper, we largely assume an intruder's configuration to be comprised of just its location and transmit power, but the concept of configuration is quite general and could include any attributes (e.g., height, antenna direction, etc.) that affects how its transmitted signal is received at other locations. Moreover, for simplicity, we assume that each intruder transmits at a fixed power (which may be different for different intruders). Thus, in our context, a configuration is simply the set of (location, transmit power) pairs of the potential intruders. We assume a bounded number of intruders. We use H_0 to represent the hypothesis with no intruders. See Figure 2.

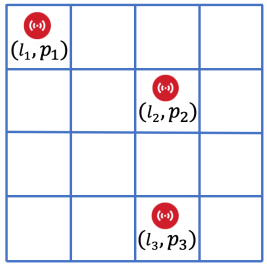


Figure 2: Illustration of a hypothesis formed of three transmitters.

If there is only one intruder, then each hypothesis represents the location and transmit power combination of the intruder, and

determining the hypothesis is equivalent to localizing the intruder and estimating its power. If we allow multiple intruders at a time, the number of possible hypotheses can be exponential in the number of intruders; we will address this challenge in §3.

Inputs. For a given set of sensors deployed over an area, we assume the following available inputs, obtained via a priori training, data gathering and/or analysis:

- Prior probabilities of the hypotheses, i.e. $P(H_i)$, for each hypothesis H_i. Prior probabilities come from known knowledge about area, intruder's behavior, etc., and can be assumed to be uniform in absence of better knowledge.
- Joint probability distribution (JPD) of sensors' observations for each hypothesis. More formally, for each hypothesis H_j, we assume $P(\mathbf{x}|H_j)$ to be known for each observation $\mathbf{x}$ for the set of deployed sensors. The JPDs can be obtained from prior training, a combination of training and interpolation (§3.3), or even by assuming a propagation model to remove the training cost completely.

Maximum a Posteriori (MAP) Localization Algorithm. We use Bayes rule to compute the likelihood probability of each hypothesis, from a given observation vector $\mathbf{x}$:

$$P(H_i|\mathbf{x}) = \frac{P(\mathbf{x}|H_i)P(H_i)}{\sum_{j=0}^{m} P(\mathbf{x}|H_j)P(H_j)} \quad (1)$$

We select the hypothesis that has the highest probability, for given observations of a set of sensors. That is, the MAP Algorithm returns the hypotheses based on the following equation:

$$\arg\max_{i=0}^{m} P(H_i|\mathbf{x}) \quad (2)$$

The above MAP algorithm to determine the prevailing hypothesis is known to be *optimal* [11], i.e., it yields minimum probability of (misclassification) error. The above hypothesis-based approach to localization works for arbitrary signal propagation characteristics, and in particular, obviates the need to assume a propagation model. However, the above MAP algorithm does incur a *one-time* training cost to construct the JPDs.

3 MAP*: OPTIMIZING MAP FOR MTL

The MAP algorithm of §2.2 can be directly applied to localize multiple intruders with optimal localization accuracy. However, MAP incurs prohibitive computational cost especially for a large number of potential intruders. In particular, note that if there are L potential locations, up to T potential intruders, and W possible discrete transmit-power levels, then the hypotheses-driven MAP algorithm needs to consider $(LW)^T$ hypotheses—making its runtime complexity exponential in number of potential intruders, and thus, making it impractical for localizing even a moderate number of intruders present simultaneously. In addition, MAP also incurs a high training cost. In the following subsections, we develop an optimized algorithm called MAP* based on MAP but with significantly improved computational and training cost. We start with optimizing the computation cost in §3.1. In the following subsection §3.2, we derive a closed-form expression to efficiently estimate intruder's power in the *continuous* domain. Finally, we discuss optimizing the training cost via a novel interpolation scheme ILDW.

3.1 Optimizing Computation Time

Basic Idea. Note that the MAP's exponential time complexity is due to the exponential number of *combinations* of locations and/or powers of the potential intruders. To motivate our proposed optimized approach, consider a simple example of 2 intruders with fixed power p in a large area. Assume that the "transmission radius" r for power p is much smaller than the area; we define the *transmission radius* as the range till which the received signal is more than a certain noise floor. The key observation is that if the intruders are far away (isolated) from each other (specifically, more than $2r$ distance away), then they could be localized independently. If the intruders are closer, then there is a need to separate aggregated signal at some of the sensors and hence we must apply the standard MAP algorithm *within that "subarea"*; however, since each such subarea is small (a disk of $2r$ radius around each possible location), the computation time is reduced significantly. However, since we do not a priori know the configurations of intruders, we need to consider appropriate possibilities.

In essence, our optimized approach is a divide-and-conquer approach, consisting of a sequence of two procedures each of which is executed iteratively. The first procedure focuses on localizing "isolated" intruders (if any) independently, while the second procedure localizes the remaining intruders—by considering all possible subareas as suggested above. The challenge lies in modifying the MAP algorithm for each iteration of the above procedures—as the hypotheses to consider across iterations of the procedures are not disjoint. We now describe each of the procedures.

Procedure 1. Localize Isolated Intruders. Informally, in this procedure, we localize intruders that are sufficiently separated from other intruders. In other words, we localize intruders x that are surrounded by sensors that receive most of their received power from x. More formally, we localize an intruder x at location l if (i) l's "neighborhood" has at least 3 sensors that receive most of their power from x, and (ii) there are no other intruders in the "vicinity" of l. In essence, we iterate over all locations l, and localize an intruder at l if the above conditions are satisfied with high enough probability, based on the readings of sensors around l. The precise definition of neighborhood above must depend on x's transmission radius which depends on its transmit power; however, as x's transmit power is unknown, we iterate over smaller and smaller neighborhoods.

We now formally describe the procedure. Let R_p denote the transmission radius for a transmit power of p. Let R denote the maximum transmission radius, i.e.,

$$\max_p R_p.$$

In the below description, we use a fractional value f to define a neighborhood and vicinity size. We start f equal to 1, use a disk of radius fR_p as a neighborhood and $R + fR_p$ as the vicinity, and iterate over the procedure for reduced values of f.

(a) Let $f = 1$.
(b) For each location and power pair (l, p), compute $P(\mathcal{H}_{l,p}|\mathbf{x}_{l,p})$ using a form of Equation 1 over appropriate JPDs. Here:
   - $\mathcal{H}_{l,p}$ represents the hypothesis that an intruder is at location l and using p transmit power. We also implicitly assume that there is no other intruder present within a distance of

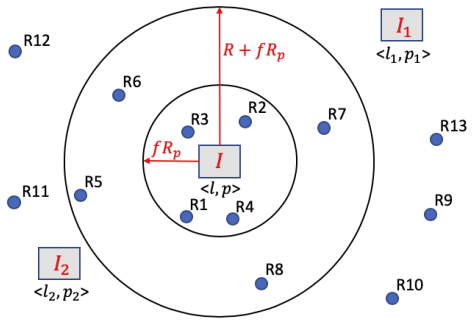


Figure 3: Illustration of Hypothesis $\mathcal{H}_{l,p}$ in Step (b) of Procedure 1. Here, the intruder I at location l is transmitting at power p, with no other intruder within a distance of $R + fR_p$ from I. The observation vector $\mathbf{x}_{l,p}$ consists of residual received powers from $R1$ to $R4$, and "noise floor" from the remaining sensors.

$R + fR_p$ from l; this ensures that the observations in $\mathbf{x}_{l,p}$ are only due to the intruder at l. See Figure 3.

- $\mathbf{x}_{l,p}$ represents the observation vector for all sensors, but the sensors that are within a radius of fR_p around l use an observation of "residual" received powers, as defined below, while the remaining sensors (outside the radius of fR_p around l) use an observation of the "noise floor" (in essence, we are "zeroing" the observations of the far-away sensors). See Figure 3.

(c) Denote (l, p) pairs that have $P(\mathcal{H}_{l,p}|\mathbf{x}_{l,p})$ higher than a certain threshold as *peaks*. If a location l is a peak and there are no other peaks within a distance of $R + fR_p$, then **localize an intruder at l with transmit power p.**

(d) For each sensor s, define its *residual received power* (RRP) as the total received power reduced by the sum of mean powers received from already localized intruders; the desired mean values are available from the given JPDs.

(e) Reduce f and go back to step #2 above, unless no new intruders were localized in (c) above. In our experiments, we used $f = 1, 1/2, 1/4$ and $1/8$.

The above procedure is partly inspired by the recent localization work [22]. However, instead of discarding sensors based on their individual power and clustering the rest as in [22], we "discard" sensors based on their neighborhood readings (i.e., likelihood $P(\mathbf{x}|H_i)$ values) and then "cluster" the remaining sensors. Also, we "cluster" iteratively, for smaller and smaller neighborhoods.

Procedure 2. Localize Intruders Situated Close-By. Once we have localized separated intruders as above, we now localize remaining intruders, if any, by applying the general MAP algorithm independently over "subareas" that still have some sensors with high-enough RRP (residual received power), but no intruder localized in the "vicinity." Formally, the procedure is as follows. Let T be the maximum number of intruders allowed within a disk of radius R, the maximum transmission radius.

(a) Let s be the sensor with highest RRP; if s's RRP is below a certain threshold (tantamount to noise), then quit.

(b) For $t = 2$ to T: Use MAP (from §2.2) to try to localize t transmitters within a disk of radius R around s, using observations of sensors within a radius of $2R$ from s. We use a certain threshold for a posterior probability, in a similar way as for Procedure 1.

(c) Update RRP of each sensor, and go to step (a) above.

Time Complexity. The worst-case time complexity of the first procedure is $O(LWG_R \log(G_R))$, where L and W are the number of potential locations (total grid cells) and transmit power levels respectively, and G_R is the maximum number of grid cells within a transmission range of an intruder. Here, the first term $O(LWG_R)$ is the time to compute the likelihood values in each iteration, since the number of sensors involved in each computation is at most G_R. Note that the number of iterations is bounded by $\log(G_R)$, as f is reduced by a constant multiplicative factor. The worst-case time complexity of the second procedure is $O(G_R(G_R)^T)$ where T is the maximum number of intruders allowed/possible in a transmission region (i.e., a circle of radius at most R). Thus, the overall time complexity of the above localization algorithm is $O(L.W.G_R.\log(G_R) + G_R.(G_R)^T)$. Generally, we would expect T to be a small constant, as more than 3 intruders in a R-radius region with a R transmission range would interfere with each other. If we also consider G_R as a small constant, the overall time complexity can be considered to be $O(L.W)$. In the following subsection, we further reduce the time complexity by removing the factor of W.

3.2 Intruder Power Estimation in the Continuous Domain

In this subsection, we derive a *closed-form* expression to estimate an intruder's power in the continuous domain, for the special case of single intruder and Gaussian probability distributions [14]. The derived result essentially removes the assumption of discrete power levels, and reduces the number of hypotheses to consider by a factor of W. We use this result within Procedure 1 of previous subsection to further optimize its time complexity and performance.

Estimating Intruder Power, Given a Location. Consider the special case of a single intruder in an area. In this case, each hypothesis can be represented as $\mathcal{H}_{l,p}$, for each location l and power p of the potential intruder. Let us focus on a particular location l^* and the corresponding hypotheses $\mathcal{H}_{l^*,p}$. For a given observation vector $\mathbf{x}$, we wish to estimate the power P that corresponds to the hypothesis with maximum likelihood among the hypotheses $\mathcal{H}_{l^*,p}$.

$$P = \arg max_p P(\mathcal{H}_{l^*,p}|\mathbf{x})$$

The value P can be computed by computing $P(\mathcal{H}_{l^*,p}|\mathbf{x})$ for each p, but our goal is to derive a closed-form expression for P from the given JPDs; such an expression yield power estimate in continuous domain without computing $P(\mathcal{H}_{l^*,p}|\mathbf{x})$ for each possible discrete p.

For each sensor (location) j, let $\mathcal{P}(\mathbf{x}_j|\mathcal{H}_{l^*,p^*})$ represent the probability distribution (PD) of j's observations $\mathbf{x}_j$ when the intruder is at l^* transmitting with power p^*, the power used at training. For a fixed l^* and p^*, the set of PDs $\mathcal{P}(\mathbf{x}_j|\mathcal{H}_{l^*,p^*})$ are equivalent to the JPDs defined in §2 under the assumption of conditional independence[2]. Let us assume that the above PDs are Gaussian distributions [14], and thus, can be represented as $\mathcal{P}(\mathbf{x}_j|\mathcal{H}_{l^*,p^*}) = N(\mu_j, \sigma_j^2)$ for a given l^* and p^*. In the above setting, the power value P that

[2]PD $\mathcal{P}(\mathbf{x}_j|\mathcal{H}_{l^*,p})$ can be computed $\mathcal{P}(\mathbf{x}_j|\mathcal{H}_{l^*,p^*})$ for any p, as the path-loss can be assumed to be independent of the transmit power, and JPD $\mathcal{P}(\mathbf{x}|\mathcal{H}_{l^*,p})$ can be computed as product of PDs $\mathcal{P}(\mathbf{x}_j|\mathcal{H}_{l^*,p})$ due to the conditional independence assumption.

maximizes $P(\mathcal{H}_{l^*,p}|\mathbf{x})$ can actually be derived as a closed-form expression; we state the result formally in the below lemma.

Lemma 1. *Consider the special case of a single intruder in an area. For a specific location l^* and power p^* (the only power used during training), let $\mathcal{P}(\mathbf{x}_j|\mathcal{H}_{l^*,p^*})$ represent the PDs of the sensor obversations at location j. Now, given the above PDs for various j and an observation vector $\mathbf{x}$, the power value $P = \arg max_p P(\mathcal{H}_{l^*,p}|\mathbf{x})$ is given by:*

$$p^* + \frac{\sum_{j=1}^{S} \frac{\gamma}{\sigma_j^2}(x_j - \mu_j)}{\sum_{j=1}^{S} \frac{\gamma}{\sigma_j^2}},$$

where $\gamma = \prod_{j=1}^{S} \sigma_j^2$ and S equals to the number of sensors in the neighborhood of l^*. ■

We omit the proof here, but give its intuition based on a special case. Consider the special case wherein each σ_j is 1 for all j. In this special case, the Lemma's equation reduces to $P = p^* + \frac{\sum_{j=1}^{S}(x_j-\mu_j)}{|S|}$, which implies that if each observation x_j is c more than its mean μ_j then P is also c more than p^*. We note that the above result does *not* extend to the case of multiple intruders. In short, the proof is a process of solving maximum likelihood esitmaion and multiple intruders introduce transcendental functions, thus cannot derive a closed-form solution.

Use of Lemma 1 in MAP*. For localization of multiple intruders, Lemma 1 can only be used in Procedure 1 of §3.1, due to its assumption of a single intruder. In particular, we can Procedure 1 of §3.1 as follows.

- We replace R_p by R, the maximum transmission radius.
- For each location l, using Lemma 1, we first compute the power $p(l)$ such that the hypothesis $\mathcal{H}_{l,p(l)}$ has the most likelihood (among the hypotheses at l) using the observations from sensors within a radius of R.
- Then, in the rest of the procedure, we only consider the (location, power) pairs of the type $(l, p(l))$ for any l.

Rest of the Procedure 1 remains unchanged. The above change has two benefits. First, the powers predicted in Procedure 1 are now continuous rather than discrete. Second, the above removes the factor of W from the time complexity of MAP* and reduces it to $O(LG_R \log(G_R) + G_R(G_R)^T)$ which becomes $O(L)$ if we consider G_R and T to be relatively small constants.

3.3 ILDW: Optimizing Training Cost

As in supervised machine learning algorithms, our Bayesian approach also needs training data. We use the term *training* to denote the process of collecting data and building up the JPDs for the hypotheses. Note that this training phase is done only one-time,[3] and hence, a certain cost is acceptable. The training cost incurred during such data gathering depends greatly on the exact mechanism used for such purposes, e.g., drones with appropriate routes can be used to gather such data [26]. In general, the cost of training would depend on the number of JPDs that need to be constructed, with the cost reduced with reduction in the number of JPDs needed. In this subsection, we design effective *interpolation* schemes that are useful in reducing the number of JPDs gathered which in turn will reduce the overall training cost. Note that reduction in JPDs constructed from raw data is bound to negatively impact the accuracy—we will evaluate this trade-off in our evaluations and show that impact on accuracy is minimal even with significant reduction in training cost.

Probability Distributions. First, we note that making the following reasonable assumptions and observations can greatly reduce the number of JPDs/PDs to be constructed.

- If we assume conditional independence of sensor observations, then JPDs can be computed from independently constructed probability distributions (PDs) of received powers at *individual sensors*.
- Since received power at a sensor location x due to multiple transmitters is merely a sum of received powers [18, 27] due to individual transmitters, we can compute PD at x for a particular hypothesis involving a set S of intruders from PDs due to each individual intruder in S.
- Lastly, we need to only construct a PD for one transmit power for each transmitter and sensor location pair, since path-loss is independent of transmit power.

Based on the above observations, if there are L discrete locations in an area for sensors or intruders, then a MAP-based approach requires L^2 PDs. Below, we propose to minimize the number of PDs to be constructed via data gathering/training, by estimating the remaining unconstructed PDs via interpolation.

Minimizing Training Cost with ILDW. Consider a particular location l^* of a potential intruder. Our eventual goal is to compute the PD for each of the L possible sensor locations for this location l^* of a potential intruder; a PD may be computed either by constructing it directly from gathered sensor observations or by estimation via interpolation from the constructed PDs. In particular, for effective interpolation, we construct PDs at coarser-grid sensor locations, and estimate via interpolation the PDs at the remaining finer-grid locations. See Figure 4. The exact coarseness at which the PDs are constructed is determined by the accuracy of the interpolation scheme for a given area and/or the impact on localization accuracy due to estimated PDs. Below, we describe the interpolation scheme that we use for our purposes.

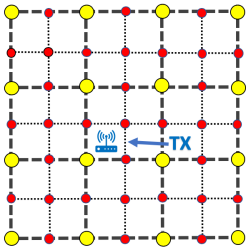


Figure 4: Training for PDs at coarse-grained locations (yellow bigger dots), while estimating PDs using interpolation at the remaining fine-grained locations (red smaller dots).

ILDW Interpolation Scheme. Consider a fixed transmitter location l^*, and let us assume locations $R_1, R_t, \cdots, R_n$ for which we know the path loss from l^*. Now, consider a new point R_0 for which we wish to estimate the path-loss from l^*. This is a traditional interpolation problem and well-known schemes such as inverse

[3] JPDs depend on the channel state and hence, must be updated periodically to account for any changes in the environment (e.g., terrain, buildings, etc.); however, such environment changes are infrequent. Also, note that the online-training of §4 is done repeatedly, but only for specific sensors and authorized users, and thus incurs minimal cost. See [31] for spectrum sensing in both spatio and temporal domains.

distance weighting (IDW), Ordinary Kriging (OK), k-NN, etc. have been evaluated even in the special context of signal strength or received power [7]. However, our specific context has an unique element. We *know* the location l^* of the transmitter from which the path-loss is being estimated—as we are in the training phase wherein we are gathering observations with transmitter at l^*. In light of the above unique element of our setting, and the observation of wireless signal characteristics, we use a custom interpolation technique which is a nontrivial modification of the IDW scheme, called *inverse log-distance weighting* (ILDW). The traditional IDW interpolation scheme estimates the path loss at R_0 by taking a weighted average of the path-losses at $R_1, R_t, \cdots, R_n$, with the weight being the inverse of the distance from R_0.

In our proposed ILDW scheme, we still estimate the path loss at R_0 as a weighted average of values at R_i's, but assign weights differently. In particular, we assign the weight for the point R_i as the inverse of the "distance" between R_0 and R_i in the domain where each point is represented merely by its logarithmic distance from l^*, the known transmitter's location—i.e., each point R_i is mapped to a point $\log d(R_i, l^*)$ on a line. This mapping is motivated by the expectation that the actual path loss would be somewhat similar to the log-distance path loss. Thus, the weight for the point R_i is assigned to be

$$w_i = \frac{1}{|\log d(R_i, l^*) - \log d(R_0, l^*)|},$$

where $d()$ is the Euclidean distance function and the path loss at R_0 is estimated as:

$$\mathbf{u_0} = \frac{\sum_{i=1}^{n} w_i \mathbf{u_i}}{\sum_{i=1}^{n} w_i},$$

where $\mathbf{u_i}$ denotes the path loss at point R_i from l^*. In the above equation for weights, if denominator is zero, then we assign w_i to be equal to the maximum of the weights among the given points (and if all denominators are 0, each weight is assigned to be 1). For an illustration of the above scheme, see Figure 5. In the IDW scheme, R_1 and R_2 will get equal weights, but under the ILDW scheme they will get weights of 5.57 and 8.00 respectively. More importantly, it can be easily shown that, for log-distance path loss, ILDW estimates the path loss for R_0 accurately from two unknown points R_1 and R_2, if $d(R_1, l^*) < d(R_0, l^*) < d(R_2, l^*)$.

The above discussion has been on using ILDW for estimating path-loss values. In general, it can be easily used to estimate PDs from the PDs at neighboring points—essentially, we can use ILDW to estimate both the mean and standard deviation of a Gaussian PD from other means and standard deviations respectively.

4 MAP**: LOCALIZING IN PRESENCE OF AUTHORIZED USERS

We have implicitly assumed till now that the only transmitters present in the area are the intruders which need to be localized. In this section, we adapt our MAP* approach described in the previous section to the setting wherein there may be authorized transmitters in the background and the localization technique must take their presence into account. In particular, in a shared spectrum paradigm, there are primary users and an evolving set of active secondary users transmitting in the background. The key challenge comes

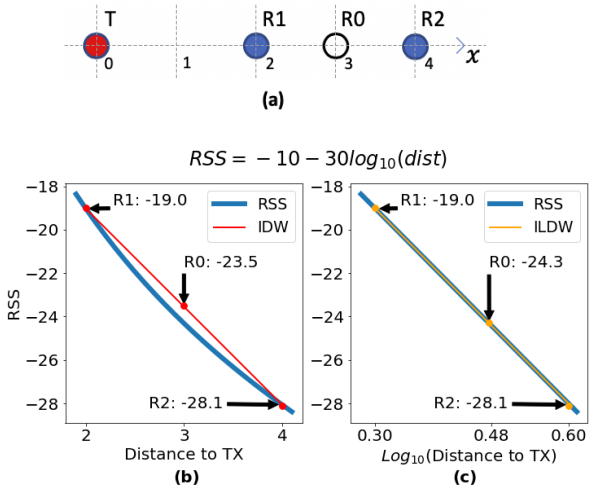


Figure 5: Illustration of ILDW vs. IDW. (a) Transmitter (T), points with known (R1 and R2) and unknown (R0) received signal strength (RSS) values. (b) Log-normal RSS function (= -10 - 30$\log_{10}$(distance)) plotted for varying distance from the transmitter T, along with IDW-estimated RSS value at a point between R1 and R2. (c) Log-normal RSS function and ILDW-estimated RSS value at a point between R1 and R2, plotted on a logarithmic distance scale.

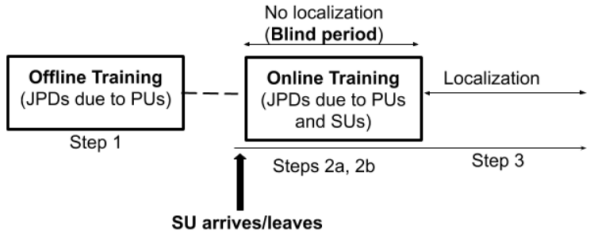


Figure 6: MAP's overall approach**

from the fact that the set of authorized users is not static and changes over time as allocation requests are granted and/or active secondary users become inactive over time.

One simple way to handle background users is to just localize every transmitter, and then remove the authorized users. However, any localization approach (including ours) is susceptible to performance degradation with increase in number of transmitters to be localized, especially if some of them are situated close together. Thus, this simple approach of localizing every transmitter is unlikely to be effective, as shown in our evaluations, especially when the number of primaries and active secondaries can be large. Thus, here, we develop an approach based on learning PDs in real-time in response to changes in the set of secondary users.

MAP**: Localizing with Authorized Users. Our problem is to localize intruders in a shared spectrum system with fixed primaries and changing set of secondaries. Our MAP** approach uses a combination of a priori (offline) and online training to construct JPDs for appropriate hypotheses based on gathered observations, and then use these JPDs to localize intruders in real-time using the MAP*

approach described in the previous section. We start with defining a few useful notations.

We use $\mathcal{R}$ to denote the set of (fixed) primaries, and $\mathcal{K}$ to denote the set of secondaries at a given instant, and $\mathcal{I}_j$ to denote the j^{th} configuration of intruders (we can assume the zero-th configuration to represent no intruders). We use $\tau = \mathcal{R} \cup \mathcal{K} \cup \mathcal{I}_j$ to denote the set to all transmitters (authorized and unauthorized) at a given instant. Finally, we use $\mathcal{P}(\mathbf{x}|(\tau = X))$ to denote the joint probability distribution (JPD) of observation vectors from the deployed sensors when the prevailing hypothesis is that the set τ of transmitters is X. MAP** is the sequence of following steps.

1. (Offline Step.) Construct JPDs $\mathcal{P}(\mathbf{x}|\mathcal{R})$ and $\mathcal{P}(\mathbf{x}|\tau = (\mathcal{I}_j \cup \mathcal{R}))$ for all j. Since these JPDs are independent of the secondaries, they do not change and can be done once a priori.
2. (Online Steps.) Whenever $\mathcal{K}$ (set of secondaries) changes:
   (a) Construct JPD $\mathcal{P}(\mathbf{x}|\tau = (\mathcal{R} \cup \mathcal{K}))$.
   (b) Compute $\mathcal{P}(\mathbf{x}|\tau = (\mathcal{R} \cup \mathcal{I}_j \cup \mathcal{K}))$ for all j, from above constructed JPDs, viz., $\mathcal{P}(\mathbf{x}|\mathcal{R})$, $\mathcal{P}(\mathbf{x}|\tau = (\mathcal{I}_j \cup \mathcal{R}))$, and $\mathcal{P}(\mathbf{x}|\tau = (\mathcal{R} \cup \mathcal{K}))$. See the below observation.
3. (Real-time Localization.) Periodically, each sensor sends its observation to a centralized entity (spectrum manager) which uses MAP* to localize any intruders present. Here, localization essentially means determining the most likely prevailing hypothesis among the hypotheses $\tau = (\mathcal{R} \cup \mathcal{I}_j \cup \mathcal{K})$, based on the JPDs $\mathcal{P}(\mathbf{x}|\tau = (\mathcal{R} \cup \mathcal{I}_j \cup \mathcal{K}))$ constructed in earlier steps.

Note that steps 1 and 2a are essentially learning the authorized users' signal charecteristics and view them as the "background signals". If there are no authorized users, then the background signals are "quite". Else, then the background signals have some "sound". We now state the observation that forms the basis of JPD computation in Steps 2b; note that the noise due to sensor's hardware gets duplicated when "adding" two JPDs, but can be easily removed.

Observation 1. *The JPD $\mathcal{P}(\mathbf{x}|(\tau = A \cup B))$ and be computed from JPDs $\mathcal{P}(\mathbf{x}|(\tau = A))$ and $\mathcal{P}(\mathbf{x}|(\tau = B))$. Similarly, JPD $\mathcal{P}(\mathbf{x}|(\tau = A))$ can be computed from the JPDs $\mathcal{P}(\mathbf{x}|(\tau = A \cup B))$ and $\mathcal{P}(\mathbf{x}|(\tau = B))$.*

Blind Period due to Step 2. Note that the steps 2a and 2b construct or compute the JPDs needed for localization, and thus, during their execution, the localization cannot be done. Thus, it is important that the duration of this "blind period" in minimal. Fortunately, step 2b being a simple mathematic computation takes only in the order of milliseconds under efficient implementation, while 2a merely entails gathering a sufficient number of observations to construct the desired JPD which could take anywhere from milliseconds to a few seconds, as an observation takes only a fraction of a millisecond [6].

Mobility of Users and Sensors. We note that MAP* works seamlessly for mobile intruders and sensors, due to the constructed PDs. However, MAP** has the following limitation: the sensors must remain static in between two consecutive online-training periods (i.e., step 2 of above). If a sensor X moves, then either X's observation must be ignored, or that X needs to online-train itself in its new location (and there should be no intruders during this individual online-training phase). Note that active SUs are expected to remain static anyway, as they are allocated spectrum for a specific location.

5 LARGE-SCALE SIMULATION RESULTS

To evaluate our techniques in a large scale area (a few kms square), we conducted simulations over a geographic area using path-loss values from the Longley-Rice propagation model generated by open sourse software SPLAT! [21]. We describe the simulation setting below and discuss the results.

5.1 Settings

Generating Probability Distributions. To evaluate our techniques over a large area with 100s of sensor nodes, we need to run simulations with an assumed propagation model. We use the well-known Longley-Rice [8] Irregular Terrain With Obstruction Model (ITWOM), which is a complex model of wireless propagation based on many parameters including locations, terrain data, obstructions and soil condition etc. and such. We consider an area of 4km × 4km in the NY state and use the 800 MHz band for SPLAT! We discretize the area using 40 vertical and 40 horizontal grid lines—yielding 1600 cells each of size 100m × 100m. To generate a probability distribution (PD) at a sensor location x due to a transmitter at location l transmitting at power p^*, we compute the received power at x using transmit power minus path-loss from SPLAT!, and use it as the mean of the probability distribution. For the complete PD, we assume Gaussian distributions and use a standard deviation between 1 and 3, with higher values for pairs (x, l) with smaller distance. As mentioned before, the PD due to multiple simultaneous transmitters can be computed as just a "sum" of the Gaussian distributions due to individual transmitters [18, 27].

Algorithms Compared. For the MTL problem, we compare our MAP* algorithm with SPLOT [18] and CLUS [22] (see §2.1). As mentioned before, [23] has been shown to be inferior in performance to both SPLOT and CLUS in their respective works, and thus, not evaluated here. CLUS uses k-means [25] for clustering, and needs to be provided with the number of clusters. To do a somewhat fair comparison, we provide CLUS with a *range* of the number of intruders and use the elbow-point method to pick the best number of clusters/intruders. In particular, the range of intruders passed to CLUS is 1 to $2x$, where x is the actual number of intruders present. For SPLOT, we use the same set of parameters values as in [18]

Table 1: Simulation Evaluation Parameters.

Param.	Value	Description
$Q_1^{'}$	0.6	Threshold for Procedure 1's hypothesis posterior
$Q_2^{'}$	0.1	Threshold for Procedure 2's hypothesis posterior
R	1000	Transmission radius when power is p^*, (m)
p^*	30	Transmit power during training, (dBm)
δ_p	2	Range of intruders' power is $[p^* - \delta_p,\ p^* + \delta_p]$

except that we use the confined area radius to be 800m for our large area setting ([18] only considered small 15m × 15m areas; 800m is roughly the maximum transmission radius in our large-scale setting and other values yielded worse results). Table 1 gives the main parameters of MAP* used in our evaluations. Recall that the transmission radius is the distance between the TX and RX for which the RX's RSS is at the noise floor (we use -80dBm).

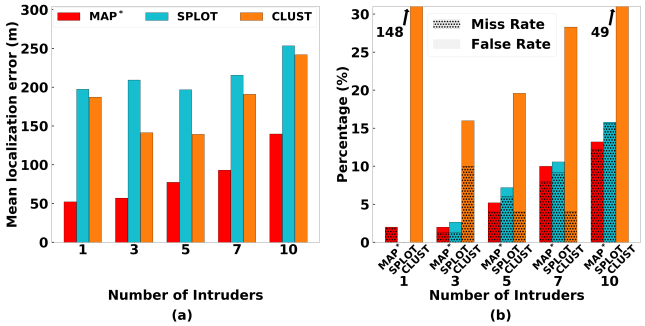


Figure 7: Localization performance of various algorithms in a large scale area, for varying number of intruders

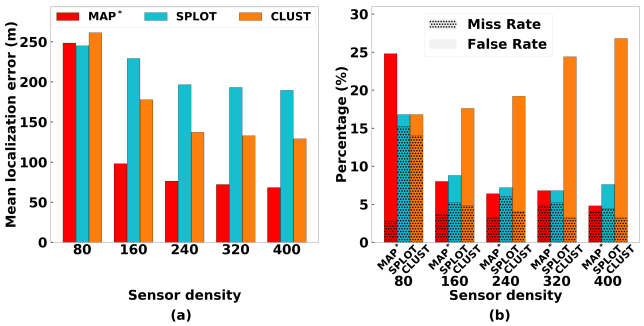


Figure 8: Localization performance of various algorithms in a large scale area, for varying sensor density

5.2 Five Evaluation Metrics.

We use the following metrics to evaluate the localization methods.

(1) Localization error (L_{err}).
(2) Miss rate (M_r).
(3) False alarm rate (F_r).
(4) Power error (P_{err}).

The above metrics are best explained using a simple example. Given a multi-intruder localization solution, we first compute the L_{err} as the minimum-cost matching in the bi-partite graph over the ground-truth and the solution's locations, where the cost of each edge in the graph is the Euclidean distance. We use a simple greedy algorithm to compute the min-cost matching. The unmatched nodes are regarded as false alarms or misses. E.g., if there are 4 intruders in reality, but the algorithm predits 6 intruders then it is said to incur 0 misses and 2 false alarms and if it predicts 3 intruders then it incurs 1 miss and 0 false alarms. The M_r and F_r metrics are on a per-intruder basis, so in the above two examples: M_r is 0 and 1/4 and F_r is 2/4 and 0. In the plots, we stack miss rate and false alarm rate together to show the overall difference between the true number of intruders and predicted number of intruders. P_{err} is the average difference between the predicted power and the actual power of the matched pair in the above bi-partite graph.

Finally for interpolation schemes, we use the metric (5) interpolation error (I_{err}) defined as the estimated path-loss minus the ground-truth path-loss value.

5.3 Results

In this subsection, we evaluate the performance of our techniques for varying parameter values, viz., number of intruders and sensors in the field, and training cost. Here, the training cost is defined relative (specifically, as a percentage of) to the full training scenario wherein we construct each of the 1600×1600 PDs (one for each pair of transmitter and sensor locations) directly from observations. E.g., x% training cost indicates that we construct $1600 \times (16x)$ PDs directly, and interpolate the remaining $1600 \times (1600 - 16x)$ PDs; our proposed interpolation scheme only interpolates for sensor locations. In general, when we vary a specific parameter, the other parameters are set to their default values which are: 9% for training cost, 5 for number of intruders, and 240 for number of sensors. For each experiment, the said number of sensors and intruders are deployed randomly in the field, with the intruders deployed in the continuous location domain while the sensors deployed only at the centers of the grid cells. Each data point in the plots is an average of 50 experiments.

Varying Number of Intruders. First, we compare the localization accuracy of various algorithms for varying number of intruders. See Figure 7. We vary the number of intruders from 1 to 10. We observe that the localization error of MAP* is the minimum across the three algorithm. The localization error is 45% – 74% less than SPLOT. In terms of the M_r and F_r, MAP* also performs others which confirms the overall performance of MAP* to be the best among the algorithms compared. In terms of absolute performance, note that the localization error of 50-150m indicates an error of 1-2 grid cells, and thus is minimal in the context of the large area of 4km by 4km with 1600 cells and a sensor population of 240. Investigating further, we observe that misses in MAP* are mostly due to the interpolated PDs (note that only 9% of the PDs are constructed from the actual sensor observations, and the remaining 91% are interpolated), while SPLOT's misses are mainly from the case of two or more intruders being close to each other. This demonstrates the superior ability of MAP* to localize intruders that are close-by via the designed sequence of Procedures 1 and 2.

Table 2: MAP* Power Error (dB)

# Intru.	MAE	ME
1	0.56	-0.07
3	1.02	0.89
5	1.31	0.97
7	1.52	1.16
10	1.47	1.04

Table 3: Running time (s)

# Intru.	MAP*	SPLOT	CLUS
1	0.55	0.56	0.03
3	1.07	1.02	0.11
5	5.74	1.35	0.23
7	8.14	1.63	0.30
10	16.50	1.89	0.41

Intruder Power Estimation, and Computation Time. Table 2 shows the mean absolute error (MAE) and mean error (ME) of the intruder's predicted power by MAP*. Note that CLUS and SPLOT do not predict intruder's power, and hence, not shown. We observe that MAP* is able to predict intuder's power quite accurately. The errors increase with the increase in number of intruders. Also, the mean error begins at near zero and then turns positive. Table 3 shows the running time of various algorithms over an Intel i7-8700 3.2 GHz processor. We see that CLUS is the fastest, and the running times of MAP* and SPLOT are comparable for small number of intruders, but for larger number of intruders, MAP* takes longer time than SPLOT mainly because of more number of iterations of the computationally-intensive Procedure 2.

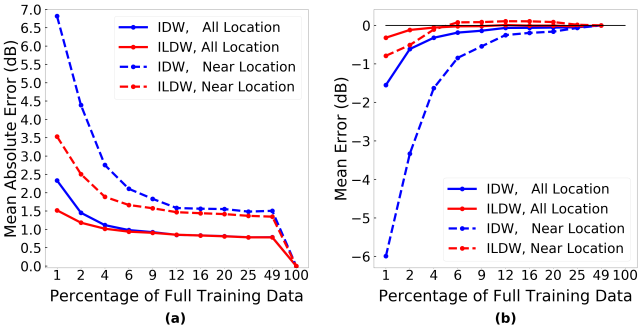


Figure 9: Estimation errors for interpolation schemes for varying training data

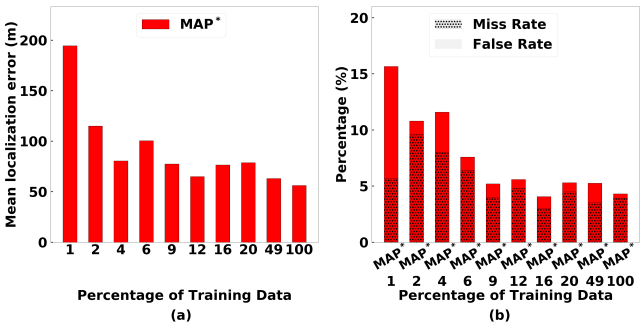


Figure 10: Localization performance of MAP* in a large scale area, for varying training data

Varying Sensor Density. We now vary the total number of sensors in the field, and observe the impact on the performance of various algorithms. See Figure 8, where the number of sensors is varied from 80 to 400. We see that all algorithms perform better with increasing number of sensors as expected, with MAP* performance improving significantly (in both L_{err} as well as $f_r + m_r$) as number of sensors is increased from 80 to 160. More importantly, except for very low number of sensors (i.e., 80), MAP* handily outperforms the other two algorithms.

Varying Training Cost. Finally, we now investigate how the training cost (i.e., number of PDs constructed from raw observations) affects the performance of our MAP* algorithm. Note that the other algorithms do not depend on the training data, hence not shown. We first evaluate the interpolation error of our ILDW scheme for varying training cost (number of known PDs) by comparing with the traditional IDW scheme on which it is based. See Figure 9, which plots the mean absolute error (MAE) as well as mean error (ME). As the interpolation error is substantially higher for points that are closer to the transmitter, we plot MAE and ME as averaged over all interpolated points as well as over just the points close (less than 800m away) to the transmitter. Note that the PDs at sensor locations closer to the transmitter would have a stronger bearing on the localization accuracy, and thus, the MAE and ME values for points closer to the transmitter are of more significance. We observe here that as expected both MAE and (absolute value of) ME decrease with increase in the training cost for both IDW and ILDW, but MAE and ME of ILDW is significantly lower than that of IDW especially for low percentages of training cost and when the points are close to the transmitter.

We now plot the performance of MAP* for varying training data; see Figure 10. As expected, the performance metrics show general improvement with increase in amount of training. More importantly, we note that with 5-10% of training, MAP* achieves performance comparable to that with 100% training, suggesting that our interpolation scheme is largely effective as long as 5-10% of PDs are constructed from raw observations.

In Presence of Authorized Users (MAP).** We now evaluate the performance of our MAP** approach which is tailored to work in the presence of authorized users. To evaluate MAP**, we place 5 authorized users in the area—with 2 primary and 3 secondary users. The primary users are placed at fixed locations, while the secondaries are put at random locations. We assign each authorized user a random power in the range of 30 to 32dBm, while, as before, a random power between 28 and 32dBm to the intruders. To ensure that these 5 authorized users do not "interfere" with each other, we ensure that the distance between any two of these authorized users is at least 1000m. We compare MAP** with the simpler approach called MAP*+ that uses MAP* to localize all transmitters (authorized as well as intruders) and then removes the predicted transmitters that are closest to the authorized users. See Figure 11, which shows that MAP** easily outperforms MAP*+ for varying number of intruders.

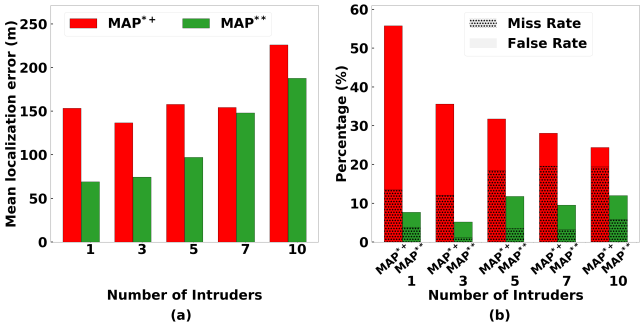
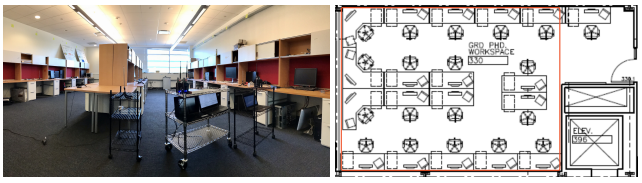


Figure 11: Localization performance of MAP*+ and MAP in large-scale simulations with authorized users present, for varying number of intruders**

6 TESTBED IMPLEMENTATION

In this section, we implement our techniques over commodity devices and evaluate them over two small-scale testbeds—one indoor and one outdoor. Outdoor environment is a realistic setting for our target application of shared spectrum systems, while the indoor environment provides more challenging signal attenuation characteristics due to walls and other obstacles.

Sensor and Transmitters Used. Our low-cost (sub $100, see [10] for a measurement study of low-cost spectrum sensors) sensing device is composed of a single-board computer Odroid-C2 with an RTL-SDR dongle which connects to a dipole antenna. We deploy 18 of these sensing devices in our indoor and outdoor testbeds, and configure them for low gain. For transmitters/intruders, we use USRP B210 and HackRF devices powered by laptops; we place these on a cart for mobility. These transmiter devices are uncalibrated, and there is no way to assign a specific transmit power. However,



(a) Indoor lab environment (b) Floor plan

Figure 12: Indoor testbed. (a) Our lab used for the indoor testbed, (b) The lab's floor plan.

(a) Outdoor parking lot environment (b) Satellite

Figure 13: Outdoor testbed. (a) Parking lot picture, (b) Satellite image of the parking lot; the red box is the area of the experiment, and the stars are the locations of sensing devices during evaluation.

they have a configurable parameter called *gain* which is almost perfectly correlated to power when the gain is in a specific range, i.e., when the transmitter's gain is increased by 1, the receiver's signal strength increases by 1dB. We thus use the gain parameter to adjust transmit power in the USRP devices. For indoor experiments, the location is manually derived, while for outdoor experiments, we use GPS dongles connected to the laptops. For collecting sensor observations, we implemented a Python repository in Linux that measures spectrum in real time at 915MHz ISM band and 2.4Msps sample rate. The repository collects I/Q samples fetched from the RTL-SDR dongle and computes the RSS value, then record the RSS along with timestamp and location. These three pieces of information are sent to a server that runs the localization algorithms.

Testbeds. The **indoor** testbed is built in a lab of our Computer Science building. Figure 12 depicts the lab with its floor plan. The red box in the floor plan is the area where experiments are conducted. The area is $9.6 \times 7.2\ m^2$ (or 2177 square feet) large, with four rows of desks. The middle two rows are separated by a wooden board. The area is imagined to be divided into 48 grid cells each of size $1.2\text{m} \times 1.2m$, with the help of ceiling tiles each of which is $0.6\text{m} \times 0.6\ m$. The **outdoor** testbed is over an open space parking lot. See Figure 13. The area is $32\text{m} \times 32\text{m}$. We divide the area into 100 grid cells with each cell representing an area of $3.2\text{m} \times 3.2\text{m}$. The GPS device returns location in (latitude, longitude) and the program converts it into coordinates. We use an outdoor WiFi router and long power cords for network and electrical connection respectively. During the evaluation, the 18 sensing devices are placed on the ground and are randomly spread out.

Training. In both the testbeds, for training (i.e., constructing non-interpolated PDs), we first pick 18 random grid cells and place sensors in their approximate centers. Then, we manually move the transmitter around in a cart through each of the grid cells.

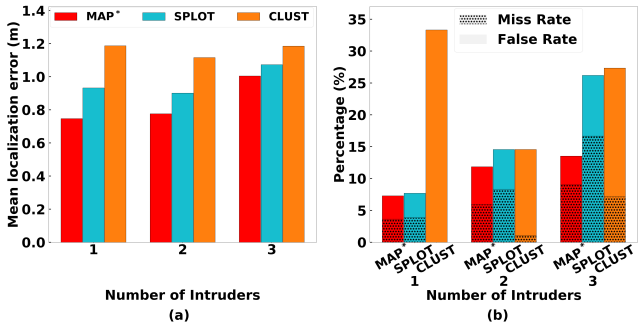


Figure 14: Localization performance of varies algorithms in an indoor testbed

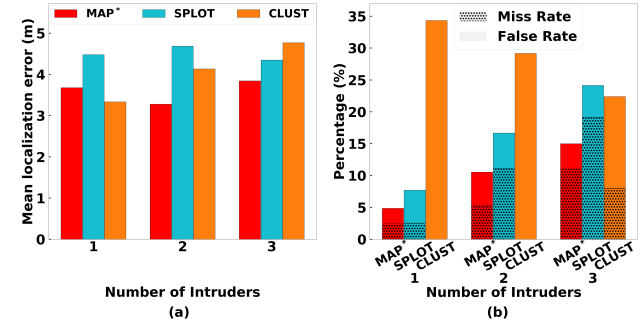


Figure 15: Localization performance of varies algorithms in an outdoor testbed

For the USRP transmitter, we use a gain value of 45 in the indoor environment and 58 in the outdoor testbed. We use a higher gain for outdoors to allow the transmitter to have a larger transmission range in a larger area. With each grid cell, the transmitter transmits from 3 to 4 different points within each grid cell, and for each such location of the transmitter, the sensors (at the 18 picked locations) gather tens of signal strength readings. From these readings, we construct a Gaussian probability distribution from each grid cell location of the transmitter. More specifically, for a particular grid cell location of the transmitter, we average over the readings from multiple TX positions within that particular grid cell—this process of averaging different positions of the TX inside a grid cell makes the Gaussian distributions more robust to multipath fading and shadowing. The overall training process takes an hour for indoors, and about two and a half hours for outdoors.

Evaluation. For evaluation, in both testbeds, we place the 18 sensors at centers of grid cells that are randomly chosen and are different from the cells chosen for training above. The chosen locations for the outdoor tested are shown in Fig. 13(b). We choose the intruder's gain/power to be in the range of $[p^* - 1, p^* + 1]$, where p^* is the gain/power used during the training phase as mentioned above. Roughly half of our experiments involve close-by (in the same or adjacent grid cells) intruders. Localization is done on a laptop which listens to HTTP requests containing the sensors' observations.

6.1 Results

Localization Metrics. Figure 14-15 show the localization results for the indoor and outdoor testbeds respectively. Overall, the results indicate that MAP* performs the best across all metrics, with the

overall performance gap between MAP* and SPLOT increasing with the increase in number of intruders. When the number of intruders is 3, the performance of SPLOT is significantly worse than MAP* due to a significantly higher (84% for indoors and 53% for outdoors) sum of miss and false-alarm rates and 43% higher localization error. The CLUS algorithm generally performs the worst, but its performance doesn't have a strong correlation with the increase in the number of intruders; recall that CLUS is given the range of number of intruders as an extra piece of information compared to the other algorithms. In terms of absolute performance, we see that the localization error of MAP* is roughly around 1 or less grid cell, and the sum of miss-rate and false-alarm is between 5-15%.

Table 4: Interpolation Mean Absolute Error (MAE) and Mean Error (ME) in dB for IDW and ILDW

Environment	IDW (MAE)	ILDW (MAE)	IDW (ME)	ILDW (ME)
Indoor	2.6	1.7	1.7	0.25
Outdoor	6.2	2.7	5.8	0.48

Interpolation Error. Table 4 show the interpolation mean absolute error (MEA) as well as mean error (ME) of IDW and ILDW when the transmitter and receiver are close by (i.e., within a distance of 3 grid cells). When the transmitter and receiver are far away, the difference of IDW and ILDW is small and thus not shown. We see that when compared with IDW, our ILDW interpolation scheme decreased the mean absolute error by 35 percent in the indoor environment and 56 percent in the outdoor environment. In terms of mean error, ILDW reduced the error compared to IDW by as large as 86 percent and 92 percent respectively. This is because IDW mostly tends to estimate the value to be larger than the ground truth, while ILDW's estimates are more even across the ground truth.

Table 5: Power Prediction Mean Absolute Error (MAE) and Mean Error (ME) in dB for indoor and outdoor testbed

# Intruder	Indoor (MAE)	Outdoor (MAE)	Indoor (ME)	Outdoor (ME)
1	0.34	0.50	-0.02	0.02
2	0.57	0.63	0.10	0.54
3	0.77	0.90	0.49	0.76

Intruder Power. Table 5 show the errors in the predicted powers of the intruders in MAP*. We see that the outdoors have a slightly higher power prediction error, likely because of a larger number of grid cells. We also note that with the increase in the number of intruders, the error in predicted power increases.

7 CONCLUSIONS

In this paper, we have developed an efficient Bayesian approach with a noval interpolation scheme to localize multiple transmitters in presence of authorized users, and demonstrate its superior power over large-scale simulations and smaller scale indoor and outdoor testbeds. In our future work, we wish to extend our techniques to allow a continuous location domain and design methods to further minimize training cost. In addition, we will consider alternate signal measurements such as angle-of-arrival (AoA).

8 ACKNOWLEDGMENTS

This work is supported by NSF grants CNS-1642965 and CNS-1815306. The authors would like to thank the anonymous reviewers and the shepherd for their valuable comments and helpful suggestions. The authors thank Salman Qavi and Jayesh Ranjan for their help during the testbed. Also thank Samir Das, Petar Djuric and Peter Milder for providing advice during group meetings.

REFERENCES

[1] https://www.ettus.com/all-products/ub210-kit/.
[2] https://greatscottgadgets.com/hackrf/one/.
[3] J. G. Andrews et al. What will 5G be? *IEEE JSAC*, 32(6), 2014.
[4] P. Bahl and V. N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. In *IEEE INFOCOM*, 2000.
[5] A. Bhattacharya, C. Zhan, H. Gupta, S. R. Das, and P. M. Djuric. Selection of sensors for efficient transmitter localization. In *IEEE INFOCOM*, 2020.
[6] A. Chakraborty, A. Bhattacharya, S. Kamal, S.R. Das, H. Gupta, et al. Spectrum patrolling with crowdsourced spectrum sensors. In *IEEE INFOCOM*, 2018.
[7] A. Chakraborty, Md S. Rahman, H. Gupta, and S.R. Das. SpecSense: Crowdsensing for efficient querying of spectrum occupancy. In *IEEE INFOCOM*, 2017.
[8] K. Chamberlin and R. Luebbers. An evaluation of longley-rice and gtd propagation models. *IEEE Transactions on Antennas and Propagation*, 30(6), 1982.
[9] P. Corbalan, G.P. Picco, and S. Palipana. Chorus: UWB concurrent transmissions for GPS-like passive localization of countless targets. In *ACM/IEEE IPSN*, 2019.
[10] M. Dasari, M. B. Atigue, A. Bhattacharya, and S.R. Das. Spectrum protection from micro-transmissions using distributed spectrum patrolling. In *PAM*, 2019.
[11] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.
[12] A. Dutta and M. Chiang. "see something, say something" crowdsourced enforcement of spectrum policies. *IEEE Trans. on Wireless Comm.*, 15(1), 2016.
[13] M. Ghaderibaneh, M. Dasari, and H. Gupta. Multiple transmitter localization under time-skewed observations. In *IEEE DySpan*, 2019.
[14] A. Goswami, L.E. Ortiz, and S.R. Das. Wigem: A learning-based approach for indoor localization. In *ACM CONEXT*, 2011.
[15] B. Grobwindhager, M. Stocker, et al. SnapLoc: An ultra-fast UWB-based indoor localization system for an unlimited number of tags. In *ACM/IEEE IPSN*, 2019.
[16] L. Hartung and M. Milind. Policy driven multi-band spectrum aggregation for ultra-broadband wireless networks. In *IEEE DySpan*, 2015.
[17] C. Karanam, B. Korany, and Y. Mostofi. Tracking from one side – multi-person passive tracking with wifi magnitude measurements. In *ACM/IEEE IPSN*, 2019.
[18] M. Khaledi et al. Simultaneous power-based localization of transmitters for crowdsourced spectrum monitoring. In *ACM MobiCom*, 2017.
[19] C. W. Kim, J. Ryoo, and M. M. Buddhikot. Design and implementation of an end-to-end architecture for 3.5 ghz shared spectrum. In *IEEE DySPAN*, 2015.
[20] Zhijing Li et al. Scaling deep learning models for spectrum anomaly detection. In *ACM MobiHoc*, 2019.
[21] J. A. Magliacane. SPLAT! a terrestrial RF path analysis application for Linux/Unix. https://www.qsl.net/kd2bd/splat.html, 2008.
[22] J. Nelson, M. Hazen, and M. Gupta. Global optimization for multiple transmitter localization. In *IEEE MILCOM*, 2006.
[23] J. K. Nelson et al. A quasi EM method for estimating multiple transmitter locations. *IEEE Signal Processing Letters*, 16(5), 2009.
[24] Neal Patwari et al. Locating the nodes: cooperative localization in wireless sensor networks. *IEEE Signal processing magazine*, 22(4), 2005.
[25] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
[26] P. K. Penumarthi et al. Multirobot exploration for building communication maps with prior from communication models. In *Intl. Symp. on Multi-Robot and Multi-Agent Systems*, 2017.
[27] Theodore Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 2001.
[28] M. Schmidt et al. Wireless interference identification with convolutional neural networks. In *IEEE Intl. Conf. on Industrial Informatics (INDIN)*, 2017.
[29] David Wright and Andrew Clegg. The 3.5 GHz citizens broadband radio service (CBRS), 2016.
[30] F. Zafari, A. Gkelias, and K. K. Leung. A survey of indoor localization systems and technologies. *IEEE Communications Surveys Tutorials*, 21(3), 2019.
[31] Y. Zeng, V. Chandrasekaran, S. Banerjee, et al. A framework for analyzing spectrum characteristics in large spatio-temporal scales. In *ACM MobiCom*, 2019.
[32] R. Zhang, J. Liu, X. Du, B. Li, and M. Guizani M. AOA-based three-dimensional multi-target localization in industrial WSNs for LOS conditions. *Sensors*, 2018.