TxMiner: Identifying transmitters in real-world spectrum measurements

Mariya Zheleva
Computer Science
University at Albany, SUNY
mzheleva@albany.edu

Ranveer Chandra, Aakanksha Chowdhery,
Ashish Kapoor, Paul Garnett
Microsoft Research
{ranveer, ac, akapoor, paulgar}@microsoft.com

***Abstract*—Knowledge about active radio transmitters is critical for multiple applications: spectrum regulators can use this information to assign spectrum, licensees can identify spectrum usage patterns and better provision their future needs, and dynamic spectrum access applications can leverage such knowledge to pick operating frequency. Despite the importance of transmitter identification the current work in this space is limited and requires prior knowledge of transmitter signatures to identify active radio transmitters. More naive approaches are limited to detecting power levels and do not identify characteristics of the active transmitter. To address these challenges we propose *TxMiner*; a system that identifies transmitters from raw spectrum measurements without prior knowledge of transmitter signatures. TxMiner harnesses the observation that wireless signal fading follows a Rayleigh distribution and applies a novel machine learning algorithm to mine transmitters. We evaluate TxMiner on real-world spectrum measurements between 30MHz and 6GHz. The evaluation results show that TxMiner identifies transmitters robustly. We then make use of TxMiner to map the number of active transmitters and their frequency and temporal characteristics over 30MHz-6GHz, we detect rogue transmitters and identify opportunities for dynamic spectrum access.**

I. Introduction

There is an increased demand for additional RF spectrum to support mobile data communication. However, nearly all the RF spectrum has been allocated for different purposes, e.g. TV, radio, cellular, radars, satellites, etc. Therefore, spectrum regulators worldwide are investigating the use of Dynamic Spectrum Access (DSA) techniques, such as in the TV white spaces or tiered access in 3.5 GHz of spectrum, to meet the additional demand. Using these techniques, mobile devices can send and receive packets over a frequency as long as they do not interfere with the licensed user of that frequency.

To identify new spectrum for DSA, the government and spectrum regulators worldwide have expressed a desire to create large-scale spectrum inventory in order to determine spectrum usage at different locations over long periods of time [6]. For example, in the US, the goal of the Spectrum Inventory Bill [2] is to create a nationwide footprint of spectrum usage over time. Based on these measurements, spectrum regulators can open new portions of the spectrum for DSA [5]. Furthermore, new DSA technologies can be designed taking into account the characteristics of these bands.

Creating such national spectrum inventory is aimed at answering various questions including (i) how much spectrum is occupied and how much is idle, (ii) how many transmitters occupy a given frequency band, and (iii) are they authorized to operate in this band. While the first question can be approached by simple estimation of power level in a given band, the other two questions require more elaborate analysis of spectrum occupancy. Such analysis needs to answer questions such as are there more than one transmitters in a given band, and what are their received powers, operating frequencies, bandwidth and temporal characteristics. Learning these characteristics from raw spectrum measurements is critical for improved policing and technological advancements in the DSA domain.

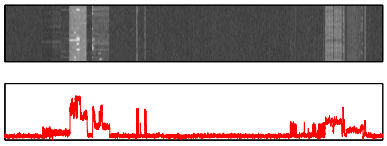


Fig. 1. Example of overlapping transmitters.

Despite the need for deep understanding of spectrum occupancy, there does not exist a platform to create such nationwide spectrum usage footprint. This is primarily due to lack of scalable infrastructure for collection and processing of RF spectrum measurements. Traditionally, spectrum occupancy is analyzed via spectrum analyzers that capture large amounts of data. The latter poses challenges in scalable data storage. Furthermore, the current approaches to mining and summarizing spectrum measurements are very limited, making it hard to evaluate the collected spectrum data.

Current approaches to spectrum summarization require prior knowledge of transmitter signatures and fine-grained spectrum measurements [9, 14], both of which are difficult to obtain in wide-band sweep-based spectrum sensing. Other, more general approaches, make use of measured power level in order to determine spectrum occupancy [17], however they can be very inaccurate in creating a spectrum inventory. To illustrate the problems related to spectrum traces summarization let us consider the example in Figure 1. The top part of the figure plots power spectral density (PSD) measured over the course of 90 seconds between 700 and 900MHz. The bottom figure plots a maxhold of PSD in the entire frequency range, that is the maximum measured PSD value in each frequency bin. Arguably, in some parts of the spectrum there exist more than one transmitters that occupy the same band in a time-division fashion. Direct analysis of the data in time-frequency domain is prone to errors due to the noisy nature of raw spectrum signals. Analysis of the maxhold, on another hand, can provide intuition of occupied fractions of this spectrum

978-1-4799-7452-8/15/$31.00 ©2015 IEEE

but hides the time-frequency characteristics of the signal; that is if multiple transmitters share the same frequency band they will be considered as a single transmitter. Furthermore, using maxhold-based analysis, one cannot determine the temporal patterns of the signal.

To address these challenges we design a new technique, called TxMiner, that identifies transmitters from raw spectrum measurements, even when the transmitter characteristics are not known and the spectrum sensing resolution is low. TxMiner leverages the phenomenon that fading of non-line-of-sight wireless signals typically follows a Rayleigh distribution, while noise follows a Gaussian distribution [7]. Thus, the raw spectrum samples can be modeled as a mixture of Rayleigh distributions capturing the ongoing transmissions, and a Gaussian representing the noise. Based on this observation we design a machine learning algorithm that extracts Rayleigh and Gaussian sub-populations from a given RF signal population. There are two challenges associated with such approach to transmitter characterization. First, the performance of our Rayleigh-Gaussian mixture model is dependent on the initialization of the model. To address this challenge, we design a multi-scale initialization scheme, presented in detail in Section II-D. Second, in order to extract frequency and temporal transmitter characteristics we need to perform time-frequency analysis of the collected data. Towards this end, we design a post-processing technique detailed in Section II-E. Thus, TxMiner is comprised of three critical components: (i) multi-scale initialization, (ii) Rayleigh-Gaussian representation of raw spectrum measurements and (iii) post-processing for actual transmitter identification.

We evaluate TxMiner on spectrum measurements collected by the Spectrum Observatory[1], as well as on several controlled transmissions, and we have found that it can accurately identify transmitters of different types including WiMax, TV & FM broadcasts, as well as proprietary DSA protocols. We demonstrate TxMiner's ability to map the number of active transmitters and their bandwidths over a wide band from 30MHz to 6GHz, recognize rogue transmitters and identify opportunities for dynamic spectrum access. This paper makes several key contributions:

- We design the first of its kind mechanism, called TxMiner, that can identify transmitters and their characteristics in raw spectrum measurements.
- We harness TxMiner to create a spectrum inventory through longitudinal, wideband analysis of traces collected by the Spectrum Observatory in the course of a year between 30MHz and 6GHz.
- We demonstrate TxMiner's ability to detect rogue transmitters in raw spectrum scans and to quantify the opportunity for secondary transmitters in licensed bands.

This paper is organized as follows. In Section II we detail the physical phenomenon that backs our modeling technique. We then present in detail the challenges associated with mining transmitter characteristics and outline our solutions to those challenges. We continue with evaluation in Section III. Finally, we apply TxMiner on real world spectrum traces in Section IV in order to demonstrate TxMiner's capabilities to create a nationwide spectrum inventory and thus benefit both regulators as well as DSA technology designers. We finalize the paper with discussion and conclusion in Section VI.

[1]The Spectrum Observatory is comprised of distributed spectrum analyzers that perform wide-band measurements between 30MHz and 6GHz. The collected data is stored centrally for further analysis, summarization and presentation. More information at http://observatory.microsoftspectrum.com

II. TxMiner: Identifying Transmitters in Spectrum Observatory Data

Traditionally, spectrum occupancy is analyzed manually by the use of tools, such as spectrograms of power spectral density. While such tools are informative, they are not very actionable. Particularly, they do not allow automated, fine-grained, long-term observation of spectrum occupancy patterns that are needed to inform DSA system design and policy. We propose TxMiner to solve the above problems by identifying transmitters in raw Spectrum Observatory data without prior knowledge of transmitter characteristics. TxMiner enables several new and exciting applications including mapping spectrum occupancy, identifying rogue transmitters, DSA beyond TV white spaces and spectrum management.

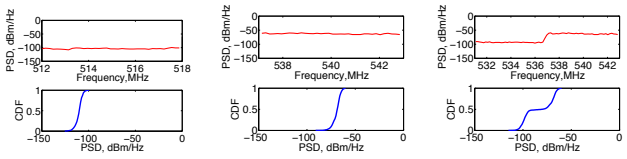
Applications of TxMiner: The problem of spectrum mapping and management is relevant worldwide. In the US the FCC has been mandated by Congress to create a spectrum usage map to be included in the Spectrum Inventory Bill [1]. In developing countries, spectrum regulators often do not know how spectrum is being used[2]. TxMiner can be applied in both scenarios for advanced *mapping of spectrum occupancy*, which in turn enables effective spectrum use and regulation. Furthermore, it can inform *spectrum management* by answering questions such as (i) how many types of transmitters are using the channel? (ii) how many transmitters of each type are present? and (iii) what is the noise floor of the channel when these transmissions are not present? TxMiner can also be useful in *identifying rogue transmitters* by detecting discrepancies between expected and detected transmitters in a given band. This capability enables spectrum licensees and regulators to identify and remove spectrum squatters.

Beyond analysis of spectrum use, TxMiner can be applied in *support of DSA technologies*. The concept of DSA is often applied in the TV bands, where incumbents have fairly static transmission patterns. Frequency ranges beyond TV bands provide vast opportunity for DSA access, however the dynamic nature of transmitters in non-TV bands poses challenges for the operation of secondary devices. TxMiner can help by providing historical information of spectrum occupancy, which in turn can inform DSA users about the transmission opportunity in spectrum bands beyond TV white spaces.

A. Key Insights

The key insight behind TxMiner is that the *probability distributions* of measured Power Spectral Density (PSD) reveal a lot about channel occupancy. To illustrate this observation we study the probability distributions of different spectrum occupancy scenarios in the TV bands. Note that these observations are valid in other bands as well.

[2]The authors have been approached by representatives of the Kenyan, Moroccan and Philippines government asking for help with analysis of spectrum occupancy.



(a) Idle (b) Occupied (c) Idle + occupied

Fig. 2. Probability Distributions of Power Spectral Density for different occupancy scenarios. The figure demonstrates how differences in measured signal distributions can inform transmitter characterization.

Figure 2 presents the probability distributions for the studied spectrum occupancy scenarios. The top graphs present a max-hold of PSD over a time window of 100 seconds, while the bottom graphs present the CDF of all values measured in this window over frequency and time. We note that a max-hold of a signal in frequency and time captures the highest measured value in a given frequency over all time samples. We see that the distributions of one occupied and one idle TV channel (Figure 2(a) and 2(b)) are very similar in shape, however, the mean of the occupied channel is higher than that of the idle channel. In a frequency band, which is in part occupied and in part idle (Figure 2(c)), the probability distribution we observe is bimodal, reflecting on the two spectrum activities. The means of the two modes correspond to the mean received power levels during the spectrum measurements.

In an urban or indoor environment, the transmitter's radio signal will attenuate with distance and encounter multiple objects in the environment that produce additional reflected, diffracted or scattered copies of the signal known as multipath signal components. Thus, the amplitude of the received signal can be characterized by a Rayleigh distribution while the phase can be characterized by a uniform distribution if we assume narrowband fading (i.e. different multipath components are not resolvable) [7]. In mathematical notation, the amplitude of the received signal $s(t)$ can be characterized by Rayleigh distribution as follows

$$R(s;\mu) = \frac{\pi s}{2\mu^2} \exp -\frac{\pi s^2}{4\mu^2},$$

where μ is the mean of Rayleigh distribution and $4\frac{\mu^2}{\pi}$ is the average received power of the signal based on the attenuation resulting from path-loss and shadowing alone. Along with active transmitters, a spectrum scan might also capture noise. This background noise can be modeled as white noise, which follows a Gaussian probability distribution [7].

So far we observed that measured transmission signals follow a Rayleigh distribution, while measured noise follows a Gaussian distribution. Thus, power values from spectrum measurements can be modeled as a mixture of Rayleigh distributions, one for each measured transmitter, and a Gaussian representing the noise. Following this intuition we develop a custom machine learning algorithm that models spectrum measurements as a mixture of Rayleighs and a Gaussian distribution. We dub this method *RGMM* (for Rayleigh-Gaussian Mixture Model). In the remainder of this section we first outline the challenges of using such an approach to characterize transmitters. We then describe how we address these challenges and present our RGMM algorithm in details.

B. Challenges

There are several challenges associated with unsupervised learning of transmitters related to (i) mixture extraction, (ii) mixture initialization and (iii) post-processing to mine for transmitters. We describe these challenges in turn.

Mixture extraction. The goal of our analysis is, given a spectrum scan over time and frequency, to identify the number and characteristics of the transmitters that occupy the measured spectrum. We assume no prior knowledge for our spectrum data, thus this problem requires an *unsupervised machine learning technique*. As already established in Section II-A, a population of radio signals can be represented as a mixture of Rayleigh and Gaussian distributions, however, there does not exist an off-the-shelf machine learning technique to fit such a mixture over unlabeled data. Towards this end we develop a custom machine learning algorithm dubbed Rayleigh-Gaussian Mixture Model (RGMM) that fits a mixture of multiple Rayleigh and one Gaussian distributions over unlabeled data. We present RGMM in detail in Section II-C.

Mixture initialization. While RGMM successfully models the power distribution of raw spectrum scans, obtaining a robust fit in a large time-frequency scan is a challenge. RGMM uses unsupervised machine learning and therefore requires a good initialization approach to extract a representative mixture model. To this end, we need a rough estimation of the signal distributions in a raw spectrum scan before running RGMM. There is a plethora of off-the-shelf data clustering algorithms that can be helpful in this step. TxMiner makes use of Gaussian Mixture Models for mixture initialization. We develop two mixture initialization techniques that are described and compared in Section II-D.

Post-processing. Obtaining a robust mixture model that represents our raw data can help answer questions such as how many transmitters do we observe and what are their approximate power levels. This mixture model, however, hides time-frequency properties of the signal that answer more interesting questions such as what is the transmitter bandwidth and what are its temporal characteristics. In order to answer these questions we need a post-processing procedure that brings together the extracted mixture model and the time-frequency characteristics of the measured spectrum scan. We design a post-processing technique that (i) calculates the association probability of each measured power value with each of the distributions in the mixture model and (ii) smooths these associations to facilitate time-frequency analysis of the raw spectrum traces. We detail our post-processing algorithm in Section II-E.

C. Rayleigh-Gaussian Mixture Models

The key feature of TxMiner that enables transmitter analysis is its ability to represent raw spectrum measurements as a mixture of Rayleigh and Gaussian distributions. This is enabled by our custom machine learning technique called Rayleigh-Gaussian Mixture Model (RGMM) that represents raw spectrum measurements as a mixture of several Rayleigh distributions – one for each sensed transmitter, and a Gaussian for the noise. We use this approach to identify sub-populations in the raw data that correspond to individual transmissions.

A mixture model is a representation of any probability distribution in terms of a weighted sum of individual probability distributions (densities). In our case, these individual densities correspond to Rayleigh densities (one for each transmitter) and one Gaussian density (noise). Each Rayleigh component in the mixture model is characterized via its mean. Furthermore, each of the components is associated with a weight that characterizes its contribution to the mixture. The Gaussian density on the other hand has two parameters: mean and variance. Formally, the Rayleigh-Gaussian mixture model $p_{MM}(s)$ can be represented as

$$p_{MM}(s) = \sum_{i=1}^{k} w_i \cdot R(s; \mu_i) + w_{\mathrm{n}} \cdot N(s; \mu_{\mathrm{n}}, \sigma_{\mathrm{n}}^2) \quad (1)$$

Here, $R(s;\mu)$ denotes the Rayleigh density with mean μ. Similarly, $N(s; \mu_{\mathrm{n}}, \sigma_{\mathrm{n}}^2)$ is the Gaussian distribution with mean μ_{n} and variance σ_{n}^2. The weights $(w_1, .., w_{\mathrm{n}})$, means $(\mu_1, .., \mu_{\mathrm{n}})$ and the variance σ_{n}^2 comprise the parameters of the mixture model, which are discovered via the Expectation-Maximization (EM) algorithm. EM aims to discover the parameters that maximize the likelihood of the statistical model (i.e. the mixture) to represent the raw data. Formally, EM is an iterative procedure that starts with a random initial assignment of the parameters and keeps refining them by alternating between the E and the M step. The E and the M step for our application are defined as follows:

$$\text{E-Step:} p(s \in j) = \frac{R(s; \mu_j)}{\sum_{i=1}^{k} R(s; \mu_i) + N(s; \mu_{\mathrm{n}}, \sigma_{\mathrm{n}}^2)}$$

$$p(s \in \mathrm{n}) = \frac{N(s; \mu_{\mathrm{n}}, \sigma_{\mathrm{n}}^2)}{\sum_{i=1}^{k} R(s; \mu_i) + N(s; \mu_{\mathrm{n}}, \sigma_{\mathrm{n}}^2)}$$

$$\text{M-Step:} \mu_j = \frac{\sum_s s \cdot p(s \in j)}{\sum_s p(s \in j)}$$

$$\sigma_j^2 = \frac{\sum_s (s - \mu_j)^2 \cdot p(s \in j)}{\sum_s p(s \in j)}$$

$$w_j = \frac{\sum_s p(s \in j)}{|s|},$$

where $s \in j$ refers to s belongs to signal component j. The EM steps are repeated until convergence (change in parameters is less than a threshold).

Once we have learned the model that best represents the raw spectrum data we can calculate the likelihood of each original data sample to be generated by each of the components in our learned mixture model. We call these likelihoods *association probabilities* and note that they are essential in our further analysis of transmitter characteristics. We now explain our approach to calculating these association probabilities.

Let us denote the matrix of raw spectrum measurements over time and frequency with S. Each element of the matrix is s_{tf}, where t is the the row of the matrix (representing a time sample) and f is the column (representing a frequency sample). The association probability with each Rayleigh component R_i can be calculated using the probability density function (PDF) of a Rayleigh distribution as follows:

$$R_i(s_{tf}, \mu_i) = \frac{\pi s_{tf}}{2\mu_i^2} exp - \frac{\pi s_{tf}^2}{4\mu_i^2} \quad (2)$$

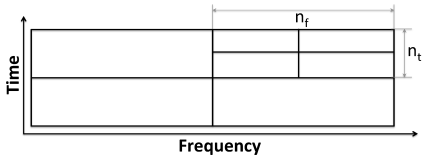


Fig. 3. An example of MultiScale with maximum resolution $l_{max} = 3$.

where μ_i is the mean of the i-th Rayleigh distribution. Similarly, the association probability with the Gaussian component N can be calculated using the PDF of a Gaussian distribution:

$$N(s_{tf}, \mu_{\mathrm{n}}, \sigma_{\mathrm{n}}^2) = \frac{1}{\sigma_{\mathrm{n}}\sqrt{2\pi}} exp \frac{(s_{tf} - \mu_{\mathrm{n}})^2}{2\sigma_{\mathrm{n}}^2} \quad (3)$$

We use the so calculated association probabilities in our further analysis of transmitter characteristics.

D. Mixture initialization

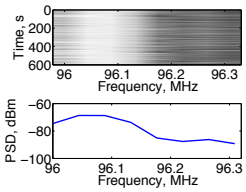
Unsupervised machine learning methods such as RGMM enable us to analyze transmitter characteristics without prior knowledge of signatures. Obtaining a robust mixture model to represent raw spectrum measurements, however, is not trivial. The robustness of the Rayleigh-Gaussian Mixture Model depends on the initialization of our RGMM algorithm. To initialize RGMM we need a rough estimation of the distributions of signal in a raw spectrum scan. To this end we use a generic clustering algorithm called Gaussian Mixture Models (GMM) to find an estimate of the signal distributions. The input to GMM is the raw data and a guess of the number of distributions k to be found. In our implementation we increase k from 1 to 5 and evaluate the best fit based on the Bayesian Information Criterion (BIC). The output of GMM clustering is a set of normal distributions characterized with a mean, standard deviation and mixing weights. We use the mean of these distributions to initialize RGMM.

We propose two initialization techniques, both of which are based on GMM. The first initialization technique takes all the raw data of interest as an input, runs GMM and uses the means of the fitted distributions to initialize our RGMM algorithm. We dub this initialization method *OnePass*. The key benefit of this initialization approach is fast calculation of the seed values for RGMM. The drawback, however, is that if we consider a spectrum scan that features multiple transmitters, some of these transmitters might either be omitted or more components than the existing transmitters might be discovered. The reason for such deviations is that it is harder to model data with a large number of generating processes (i.e. transmitters).

To reduce the number of generating processes and achieve robust initialization we design a divide-and-conquer approach dubbed *MultiScale* that calculates the initialization in a bottom-up fashion. MultiScale divides the raw data in sub-spaces with increasing resolution as illustrated in Figure 3. At the highest resolution MultiScale runs GMM in each sub-space to find the representative distributions. MultiScale then groups the discovered distributions in decreasing resolution until it produces a single set of initialization values.

Algorithm 1: MultiScale

```
1  Input: S data, (n_f, n_t) # partitions, l level, l_max - max level
2  Output: ({λ_i}, μ_n, σ_n) Rayleigh transmitters and noise parameters
3  if l = l_max then
4      ({λ_i}, μ_n, σ_n) ←Rayleigh-Fit(S, {λ_{l_max}})
5      return ({λ_i}, μ_n, σ_n)
6  if l < l_max then
7      Λ ← ∅
8      Partition S into n_f × n_t regions S_{f,t}
9      for ∀S_{f,t} do
10         ({λ_i}, μ_n, σ_n) ← MultiScale(S_{f,t}, (n_f, n_t), l + 1, l_max)
11         Λ ← Λ ⋃ {λ_i}
12     {λ_i} ← Cluster(Λ)
13     if l < l_max then
14         return ({λ_i}, 0, 0)
15     else if l = 0 then
16         ({λ_i}, μ_n, σ_n) ←Rayleigh-Fit(S, {λ_i})
17         return ({λ_i}, μ_n, σ_n)
```



Initialization	Components, dBm
MultiScale (l=1)	-79.2
MultiScale (l=2)	-71.4 -87
MultiScale (l=3)	-71.4 -87
MultiScale (l=4)	-71 -87
OnePass	-68.4 -73.1 -86.6

Fig. 4. Illustration of the benefits of MultiScale over OnePass. MultiScale discovers a robust initialization.

MultiScale is presented in Alg. 1. It is a recursive function that computes model fits in sub-regions of the frequency-time domain and aggregates the parameters up to obtain a fit for the whole space. The input to our function consists of the power measurement data S, the number of partitions in which the domain is to be recursively split in time n_t and frequency n_f, the current resolution level l initialized at 0 and the maximum level l_{max}. The maximum level parameter l_{max} controls the maximum resolution at which we will be obtaining the fit. The output of MultiScale is the set of Rayleigh parameters $\{\lambda_i\}$ and the mean μ_n and standard deviation σ_n used to initialize RGMM.

MultiScale begins with $l = 0$ and increases l until $l = l_{max}$ at which point the function reaches the base case of the recursion. The base case of the recursion ($l = l_{max}$) corresponds to the highest resolution of the frequency-time space (Lines 3-5). We perform a Rayleigh mixture fit with our default initialization $\{\lambda_{l_{max}}\}$ based on GMM (Line 4) and return the obtained model parameters. The internal recursion levels are described in Lines 6-17. At each internal level we initialize an empty set of Rayleigh parameters Λ (Line 7) and partition the current time-frequency space S into $n_f \times n_t$ regions $S_{f,t}$ (Line 8). Next, we recursively invoke MultiScale for each of the subspace regions while incrementing l, and add the parameters of obtained Rayleigh components to Λ (Lines 9-11). We cluster the set of all Rayleigh parameters returned from the higher resolution using a threshold-based approach that groups all components less than 2dBm apart. Finally, if we are at an internal recursion level (i.e. non-zero level), we return the clustered set of Rayleigh parameters (Lines 13-14), while at level 0 we perform one fit over the whole data initiating with the aggregated parameters from higher resolutions and return the final fit including the noise component (Lines 15-17).

We demonstrate the benefits of MultiScale over OnePass in Figure 4. The figure presents results using a spectrum scan of two FM channels, one sensed at -71dBm and the other at -87dBm, collected over the course of 600 seconds. To the left is an illustration of the raw data. The top figure presents a heatmap of the raw signal over time and frequency, while the bottom figure presents the average values measured in each frequency bin. The table to the right presents results from MultiScale with increasing resolution and OnePass initialization. OnePass identifies one extra transmitter, while MultiScale extracts exactly two transmitters when run with resolution $l_{max} = 2$ or higher.

Naturally, one might ask what is an appropriate maximum resolution l_{max} with which to run MultiScale. As seen in our example, the number of components discovered by MultiScale plateaus after a certain resolution (arguably, at the resolution equal to the numbed of transmitters). Using this observation, we can initialize using MultiScale with increasing l_{max} until the number of discovered components stops increasing.

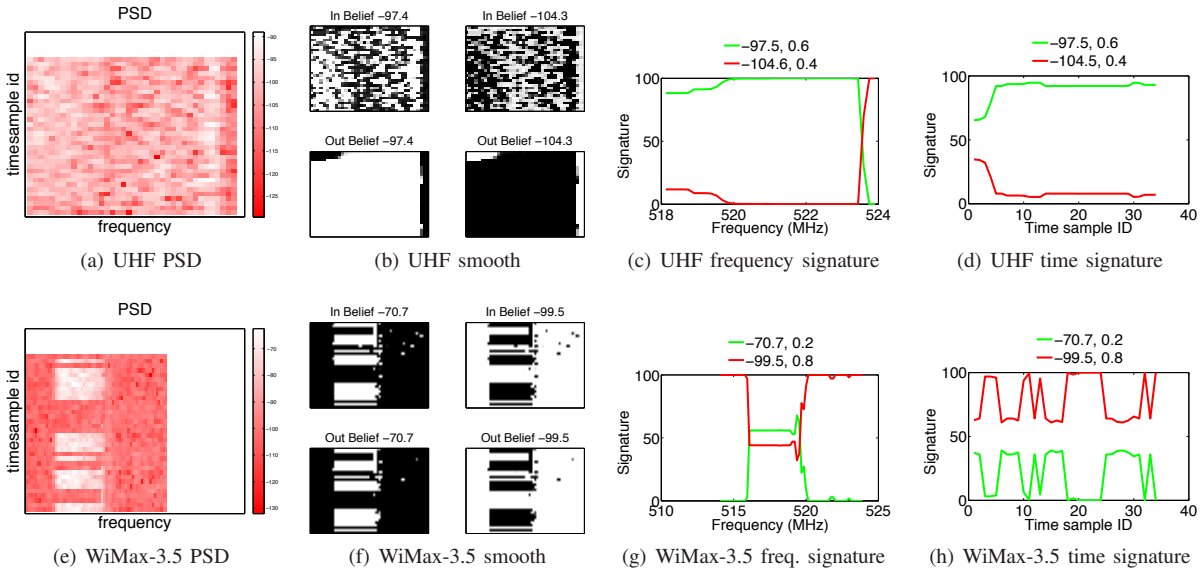
The drawback of MultiScale over OnePass is that MultiScale takes more time to obtain an initialization. We observe, however, that initialization does not need to be run every time TxMiner is ran, rather we can use the same initialization until RGMM obtains models with which the raw data values are poorly associated. Such poor association will be an indicator that a new initialization should be computed.

E. Post-processing

While RGMM allows mining of the number of transmitters and their power level, it does not allow time-frequency analysis of the collected data. Such time-frequency analysis enables characterization of other important properties such as bandwidth and temporal behavior. In order to mine time-frequency properties we implement a post-processing procedure that makes use of the association probabilities we calculate following the fitted mixture model (Equation (2) and (3)).

The association probabilities provide intuition about the time-frequency properties of sensed transmitters, however, the inherently noisy nature of spectrum scans makes it hard to mine transmitter characteristics directly from the association probability matrices. Towards this end we make the following observation. Since transmitters occupy adjacent time and frequency samples, transmitter scans are coherent in the time-frequency domain. That is, adjacent values that are of similar magnitude are likely to be due to the same transmission. This observation allows us to apply spatial regularization to smooth the association probabilities and reduce the noisiness of the post-processed signal. For the purpose of spatial regularization we use a machine learning technique called Belief Propagation.

In the remainder of this section we detail our spatial regularization approach and describe how we use the regularized data to extract transmitter characteristics. Along with our methodology, in Figure 5 we present an illustrative example of mining transmitter characteristics in two transmitter scenarios: a TV broadcast and a WiMax TDMA. Our RGMM method has fitted two components in each transmission: one representing the power of the transmitter and one capturing the noise. We detail each of these figures as we describe our post-processing technique.



(a) UHF PSD (b) UHF smooth (c) UHF frequency signature (d) UHF time signature

(e) WiMax-3.5 PSD (f) WiMax-3.5 smooth (g) WiMax-3.5 freq. signature (h) WiMax-3.5 time signature

Fig. 5. Illustrative example of TxMiner post-processing. The importance of belief propagation in "salt-and-pepper" signals such as the TV-UHF transmission is well emphasized.

Time-frequency regularization using Belief Propagation. The inherently noisy nature of RF signals causes our association probability matrices to suffer from salt-pepper noise. We can see the effect of salt-pepper noise in our example on Figure 5. The first column from left to right represents the original PSD data over frequency and time. The more white the color is, the higher the measured power. The second column presents results before and after the regularization. The "In Belief" plots are the association probabilities before smoothing, while the "Out Belief" plots are the resulting smoothed association probabilities. Darker colors represent lower values. We see that the "In Belief" suffers from salt-pepper noise, whereby neighboring cells differ in their values. The latter makes it hard to determine if adjacent values belong to the same transmitter, which in turn makes it hard to detect transmitter characteristics.

We propose to alleviate this problem via spatial regularization using a machine learning technique popular in the image segmentation literature [3]. In particular, we formulate an energy minimization problem where we consider adjacent cells in the PSD matrix (both in frequency and time) as neighbors. The goal of the energy minimization problem is to determine a solution that aligns with the mixture model available from the previous step and is spatially smooth. Formally, let us use $x_i \in \{1, ..k, \text{noise}\}$ to denote the index of the mixture distribution, with which the data s_i is associated. Then we consider the following form of the energy:

$$E(\mathbf{X}) = \sum_i -\log p_{MM}(s_i \in x_i) + \sum_{ij} V(x_i, x_j, s_i, s_j). \quad (4)$$

Here, $p_{MM}(s_i \in x_i)$ is a unary term and simply depends upon the output association probabilities from the mixture model. Intuitively, this term favors assignments that are obtained from the inference when fitting the model. The second term considers all pairs of neighbors (i and j), and smooths the data by using a function $V(\cdot)$ that depends upon the corresponding observations s_i and s_j in the PSD matrix $\mathbf{S}$:

$$V(x_i, x_j, s_i, s_j) = \begin{cases} -\log e^{-\beta|s_i - s_j|} & \text{if } x_i = x_j \\ -\log[1 - e^{-\beta|s_i - s_j|}] & \text{Otherwise} \end{cases}$$

Note that the pairwise term favors similar assignments to s_i and s_j only when the values x_i and x_j are similar. Intuitively, the pairwise term will favor dissimilar assignments to adjacent cells only when there is a large difference in observations in the PSD matrix.

An assignment that minimizes the above energy would aim to provide a solution that is coherent in time and frequency and aligned with the solution provided from the mixture model procedure. However, determining the minimum energy assignment for such energies has been determined to be NP-complete. Reasonable approximation can be computed via message passing schemes such as loopy Belief Propagation [18]. In this paper we specifically, use the sum-product version of loopy belief propagation, where given the mixture model inferences, we formulate the energy and obtain a solution via loopy message passing until convergence.

The "Out Belief" plots in Figures 5(b) and 5(f) show the result after running the loopy BP. We can observe that the resulting signal is more regularized in the time-frequency domain and does not suffer from salt-pepper noise.

Mining transmitter characteristics. The smoothed association probabilities obtained in the previous step enable efficient extraction of transmitter signatures in order to mine transmitter characteristics. In this analysis we determine key transmitter properties including: bandwidth, active time and type (including TDMA, FDMA, broadcast and frequency hopping). Towards this end we compact the association probabilities from the time-frequency domain in one-dimensional space in either frequency or time. We call these compacted probabilities *temporal* and *frequency* transmitter *signatures* and denote them as P^t and P^f. A temporal P_i^t and frequency P_i^f signature is

TABLE I. RULES FOR DETERMINING TRANSMITTER TYPE.

Type	Rule
Broadcast	$\sigma_T < THR_T$ and $\sigma_F < THR_F$
TDMA	$\sigma_T > THR_T$ and $\sigma_F < THR_F$
FDMA	$\sigma_T < THR_T$ and $\sigma_F > THR_F$
Hopping	$\sigma_T > THR_T$ and $\sigma_F > THR_F$

calculated for each Rayleigh distribution i fitted onto the raw spectrum measurements. We calculate P_i^t and P_i^f as follows:

$$P_i^t = \frac{\sum_f^F R_i(s_{ft}, \mu_i)}{F} \text{ and } P_i^f = \frac{\sum_t^T R_i(s_{ft}, \mu_i)}{T} \quad (5)$$

Here F and T denote the total number of frequency and time samples in our time-frequency matrix.

Our illustrative example in Figure 5 presents the time and frequency signatures of a TV broadcast and WiMax TDMA transmitter. Since a broadcast channel occupies all the time-frequency samples, we see that the signatures of such transmitters have low variance over time and frequency(Figure 5(c) and 5(d)). In contrast, a TDMA transmitter such as WiMax occupies a fixed bandwidth, however, its active time is non-contiguous. The latter is reflected in its signatures (Figure 5(g) and 5(h)): the frequency signature has low variability, whereas the time signature varies, capturing the intermittent presence of this transmitter over time.

We use these observations to design our detection of transmitter bandwidth, active time and type. Specifically, we calculate the transmitter bandwidth by determining the span of non-zero frequency signature. Similarly, we determine the transmitter active time by calculating the span of non-zero time signature. Lastly, we use the variance of transmitter signature to determine the transmitter type. Let us denote the variance of the time signature by σ_T and the variance of the frequency signature by σ_F. We can then determine the transmitter type by following the rules in Table II-E. THR_T and THR_F denote thresholds of time and frequency signature variance against which we decide the type of transmitter. Since the magnitude of the variance depends on the magnitude of the transmitter signature, we pick a percentage of the maximum signature value as our threshold. Of note here is that the threshold can be adjusted if we had prior knowledge about the expected transmission. Since we assume no such knowledge, we use 20% on the maximum signature as a threshold for the following evaluation of TxMiner.

III. EVALUATION

In this section we evaluate TxMiner. We start by describing the datasets we utilize in our evaluation. We begin our evaluation by looking at accuracy of detecting active radio transmitters. Following, we evaluate TxMiner's ability to extract individual transmitter characteristics such as transmitter count, bandwidth and type. We compare TxMiner with a state-of-the-art algorithm for occupancy detection called edge detection [17]. Our evaluation shows that TxMiner outperforms edge detection in both controlled settings as well as in real world measurements. We show that TxMiner has high accuracy in detecting occupancy of individual transmitters and their bandwidths. Furthermore, TxMiner is capable of detecting transmitter count and bandwidth in multi-transmitter scenarios.

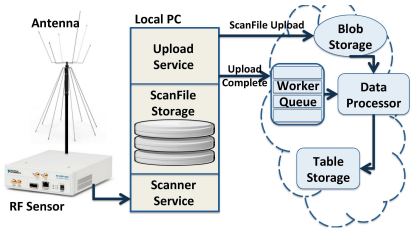


Fig. 6. Spectrum Observatory architecture.

A. Measurement Setup and Data

Our evaluation makes use of data collected by the Spectrum Observatory, which provides a distributed spectrum monitoring platform. The Spectrum Observatory continuously scans the spectrum from 30MHz to 6GHz and consists of two components as shown in Figure 6: (i) local spectrum measurement equipment at various locations, and (ii) the storage and analysis component in the cloud. The observatory makes use of heterogeneous RF sensors at the various locations ranging from high-end CRFS RfEyes[3], to the commonplace USRPs, to low-end RF Explorers such as RTL-SDR dongles. Each station's RF sensor is equipped with a multi-polarized antenna that covers the entire 25MHz-6GHz band. The RF sensor is connected to a local PC that runs a scanner service controlling the speed and granularity of spectrum scans. Spectrum data is stored locally on the PC and uploaded to the cloud via the upload service. Data is then stored and summarized in the cloud in order to be displayed by the Spectrum Observatory's frontend.

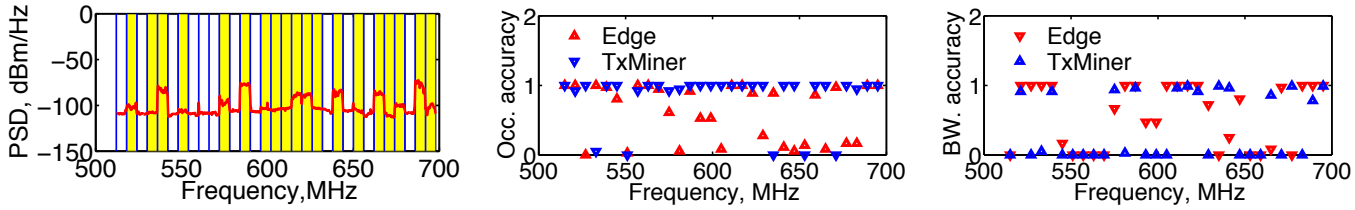
We use raw data collected by the observatory point in Microsoft's campus in Redmond, WA. The spectrum data is collected with variable frequency granularity depending on frequency band and constant speed of one sweep every three seconds. We make use of several datasets to evaluate TxMiner.

Ground truth. We use measurements in the TV-UHF band as our ground truth. We establish our TV-UHF ground truth by combining spectrum measurements in this band with information from the FCC CDBS, AntennaWeb, TVFool, and white space databases from Spectrum Bridge and iConnectiv, which indicate the channels that are idle or occupied. The TV-UHF traces were collected using a stationary RfEye sensor scanning the spectrum every 3 seconds with a frequency step of 160kHz. We make use of this dataset to evaluate accuracy in occupancy detection and detection of transmitter bandwidth. We also use this data to create an artificially mixed dataset (see below) that provides basis for more elaborate evaluation of transmitter count and bandwidth.

Controlled. We utilize a few controlled transmissions to evaluate TxMiner's ability to detect custom transmitters. We record traces from three modes of wide-range outdoors WiMax transmission: one using 1.75MHz bandwidth, a second using 3.5MHz and a third transmitting at 7MHz. We also performed spectrum scans during on-campus wide-range outdoor trials of FCC-certified white spaces radios running proprietary DSA protocols. Both the WiMax and the trial traces were collected with a stationary RfEye scanning every 3 seconds with a frequency step of 160kHz.

Artificially mixed. We generate artificially-mixed signals drawn from our TV-UHF ground truth. We intertwine over

[3]http://www.crfs.com/products/rf-sensor-rfeye-node/



(a) UHF spectrum (b) UHF occupancy accuracy (c) UHF bandwidth accuracy

Fig. 7. Occupancy and bandwidth detection. (a) Ground truth, (b) occupancy accuracy and (c) bandwidth accuracy. TxMiner outperforms edge detection in both occupancy and bandwidth accuracy. Edge detection fails in nearly 50% of the cases to accurately detect an occupied channel. Furthermore, it often detects bandwidth where there is no active transmitter or does not detect anything where there is an active transmitter.

the same frequency band different transmissions or alternate transmission with idle period. By doing so we can emulate single- or multiple-transmitter TDMA schemes, which allows us to establish a ground truth set and quantitatively evaluate TxMiner's ability to tease out multiple transmitters and their bandwidths.

B. TxMiner Performance

Occupancy accuracy. We begin our evaluation by analysing accuracy in detecting occupancy status. For this experiment we run TxMiner over the entire TV-UHF band (512-698MHz) in 6MHz steps and calculate the accuracy of occupancy detection. In each 6MHz bin there are F samples, depending on the scan configuration. For each of these samples we find if it is occupied or idle. Our accuracy metric then captures the fraction of correctly-detected samples divided by the total number of samples F. Intuitively, an accuracy of 1 corresponds to correct detection of all frequency samples in a given bin. For some cases, however, our measurements do not agree with the ground truth. Particularly, we measure 5 of the 31 channels in TV-UHF as idle, where they are supposed to be occupied according to the ground truth. In such cases, our accuracy metric would be 0, however this is still a good indicator that TxMiner can persistently detect the occupancy status. In a nutshell, for good accuracy prediction, we want our occupancy accuracy to be either 0 or 1; anything in between indicates weak prediction.

Figure 7 presents our results for detection accuracy in the UHF band. Figure 7(a) plots the measured occupancy as an average PSD over the capture period. Channels that are supposed to be occupied are designated with yellow. Figure 7(b) presents our occupancy results, where the blue markers correspond to TxMiner and the red ones represent Edge Detection. As we can see, TxMiner typically has a prediction accuracy of 0 or 1 and outperforms Edge Detection in nearly 50% of the cases. For example, channel 23 (the third channel in Figure 7(b)) is idle but is surrounded by two low-power channels, thus edge detection fails to recognize it, while TxMiner detects it successfully. The reason for the poor performance of edge detection is that it often fails to recognize a rising or falling edge, which forces longer frequency spans to be incorrectly recognized as idle or occupied.

Bandwidth detection. Next we evaluate TxMiner's ability to detect transmitters' bandwidths. First, we focus on our TV-UHF data where we run an experiment in the entire band from 512 to 698MHz in 6MHz steps. At each step we calculate the bandwidth of the detected transmitter. Figure 7(c) presents

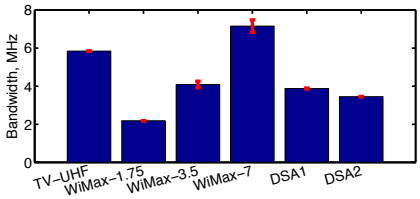


Fig. 8. Bandwidth detection of different transmitters. TxMiner is persistently able to detect the bandwidth of different transmitter and the detected values are very close to the expected ones.

a comparison between TxMiner and Edge Detection. The y-axis on the graph presents the ratio between detected and expected bandwidth, where expected bandwidth in this case is equal to that of a TV channel in the US – 6MHz. As we can see, TxMiner successfully detects the bandwidth of active transmissions and detects a bandwidth of 0MHz where we have measured no transmission or where there is no expected transmission. At the same time Edge Detection often fails to detect the bandwidth of active transmitters, or detects a 6MHz transmitter in channels that are not occupied. The reason for the poor performance of Edge Detection is that it often times fails to account for a rising or falling edge. The latter results in larger areas being detected as idle or occupied than there actually exist.

Next we evaluate TxMiner's capability to persistently detect transmitter bandwidth in different transmission scenarios. Particularly we look at the TV-UHF band, three TDMA WiMax transmissions with known bandwidths of 1.75MHz, 3.5MHz and 7MHz and two proprietary TDMA DSA transmissions with bandwidths of 4MHz and 3.5MHz. For TV-UHF we present average and standard deviation of detected bandwidth across all the channels we identify as occupied. For all the WiMax and DSA transmissions we present average and standard deviation across five distinct periods from the captured traces. All but the DSA2 scan periods are of 100s duration. For DSA2 we use a 300s scan duration because the TDMA nature of this transmission makes it so we cannot capture enough transmission samples within 100s. Figure 8 presents our results. As we can see, TxMiner is persistently able to recognize the bandwidth of each transmitter type. Furthermore, the detected bandwidths are very close to the expected bandwidths.

We also evaluate TxMiner's performance in cases with narrow-band transmissions such as those in the radio FM band. TxMiner detects this entire band as occupied. Figure 9 presents our results for accuracy of bandwidth detection. In this experiment we ran TxMiner over the entire FM band from 88MHz to 108MHz in steps of 400kHz. The graph presents

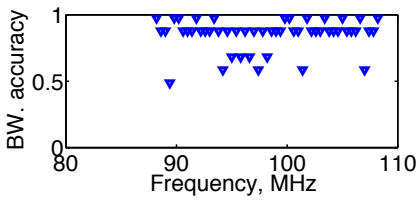


Fig. 9. Bandwidth detection in the radio FM band. TxMiner performs is highly-accurate in detecting narrow-band transmissions.

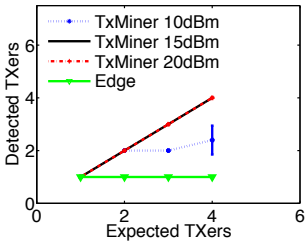


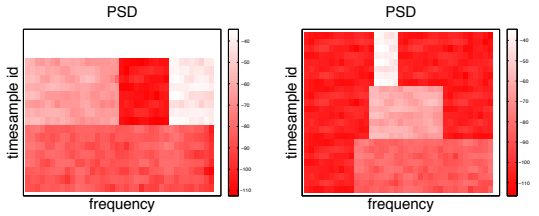
Fig. 10. Transmitter detection with increasing number of transmitters. TxMiner is able to detect the number of transmitters as they increase and clearly outperforms edge detection, which cannot identify more than one transmitter.

for each 400kHz chunk the bandwidth accuracy expressed as the ratio between detected bandwidth and step size (in this case 400kHz). As we can see, majority of the detected channels have bandwidth accuracy of either 0.98 or 0.88, which corresponds to a bandwidth of 392kHz and 352kHz, respectively. The 392kHz bandwidths likely correspond to HD radio transmissions, which by specification occupy wider bands. The 352kHz transmissions correspond to stations that were sensed with very strong signal, in which case we would see the squelch tones as a separate peak. Finally, we see transmissions whose bandwidth accuracy is lower. Those are likely to be radio transmissions that were sensed with low power, thus their bandwidth does not span the entire 400kHz band.

Transmitter type detection. As detailed in Section II-E we make use of the variance of the time and frequency signatures of a transmitter to determine the transmitter type. We now demonstrate TxMiner's ability to determine the transmitter type of our ground truth transmissions. We focus on a TV broadcast operating on channel 22 (518-524MHz). We use 20% of the maximum signature to determine the variance thresholds. For the TV broadcast $THR_F = 20$ and $THR_T = 9.66$. The calculated variance of this transmitter's signatures are 3.73 and 18.31 for time and frequency, respectively. Both the variances are lower than the respective thresholds and thus the transmitter is correctly identified as a broadcast.

Detection of multiple transmitters. Next we evaluate TxMiner's performance in scenarios where multiple transmitters are present. To emulate such scenarios we artificially mix and amplify measured signals.

Our first evaluation focuses on TxMiner's ability to detect an increasing number of transmitters of the same bandwidth. For this experiment we mix over time measured signals from the TV-UHF band and artificially amplify them (by adding 10, 15 or 20dBm) to make the difference between transmitters more pronounced. We then run TxMiner and count the number of detected transmitters. Figure 10 plots the number of detected transmitters as a function of the number of expected transmitters. We present three results for TxMiner averaged



(a) Case 1 (b) Case 2

Fig. 11. Evaluation cases of multiple transmitters with different bandwidths. (a) 6MHz transmission followed by two simultaneous transmissions at 1.4MHz and 3MHz; (b) three consecutive transmissions with bandwidths of 4.375MHz, 2.34MHz and 0.78MHz.

TABLE II. DETECTION OF MULTIPLE TRANSMITTERS.

	TX 1		TX 2		TX 3	
	E.BW (MHz)	D.BW (MHz)	E.BW (MHz)	D.BW (MHz)	E.BW (MHz)	D.BW (MHz)
Case1	6	5.84	3	2.84	1.4	1.26
Case2	4.375	4.26	2.34	2.68	0.78	0.63

over five runs and compare TxMiner's performance with Edge Detection. As we can see, TxMiner clearly outperforms edge detection. The reason for the poor performance of Edge Detection is that it only considers an average of the measured signal and unlike TxMiner, does not take into account the time-frequency properties of the signal. In contrast, TxMiner is capable of detecting the number of transmitters with high accuracy. We see that the accuracy of TxMiner is lower in the 10dBm margin scenario, where the algorithm sometimes fails to differentiate between signals.

Next we evaluate TxMiner's ability to extract multiple transmitters with variable bandwidths. For this experiment too we use artificially mixed and amplified signals. We study two cases of spectrum occupancy presented in Figure 11. Each of these cases includes a different configuration of three transmitters. In case 1 we have a 25 second transmission with 6MHz bandwidth, followed by two concurrent transmissions, one 3MHz wide and one 1.4 MHz wide and separated by an idle zone. The second case features three consecutive transmissions each of 25 seconds. Table II presents for each case and each transmitter the expected and the detected bandwidth (E.BW and D.BW, respectively) for each case and transmitter. As we can see, TxMiner successfully detects all the expected transmissions and is also accurate in detecting their bandwidths.

C. Impact of Scan Duration

In this section we evaluate the impact of scan duration on the accuracy of occupancy detection. The presented results indicate how quickly can TxMiner begin detecting transmitters after a spectrum scan is initiated. To this end, we run TxMiner on all the channels in the TV UHF band while changing the number of time samples we consider. We start with a scan duration of 3 seconds, which in our setup corresponds to two sweeps, and double the scan duration up to 192 seconds (65 sweeps). Figure 12 presents average and standard deviation of accuracy (as calculated in section III-B) over all the TV channels for each scan duration. As we can see, even for small scan durations the average accuracy is high which indicates that TxMiner can detect transmitters successfully even after

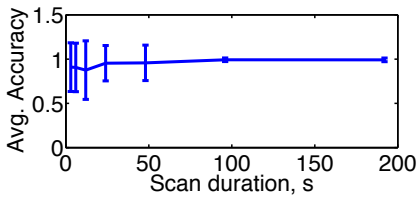


Fig. 12. Accuracy with changing scan duration. TxMiner achieves high detection accuracy with scan durations as short as 3 seconds (2 sweeps). The stability of transmitter detection across different channels, regardless how noisy they are, is guaranteed at a scan duration of 96 seconds (33 sweeps).

two frequency sweeps. Notably, the deviation across channels for small scan times is high as well, which would not be desirable for stable performance across various scenarios. This deviation depends on how noisy the channel is: intuitively the more noisy the channel the more samples TxMiner needs in order to perform accurate transmitter detection. As the scan duration increases up to 96 seconds (33 sweeps), we see that the standard deviation becomes minimal, which indicates that Txminer can persistently achieve high accuracy in about 33 sweeps across different transmission scenarios.

IV. TxMiner in the Wild

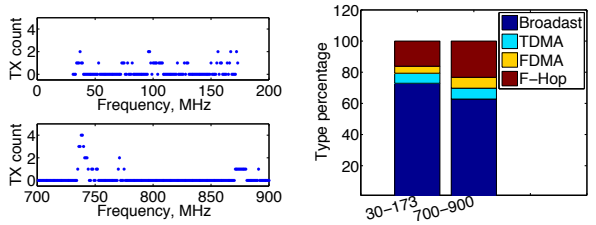
So far we have shown that TxMiner is capable of detecting transmitter characteristics in various RF bands and occupancy scenarios. This capability can be harnessed for creation of a RF inventory that gives information about transmitter characteristics over frequency, time and space. To demonstrate this capability we utilize data collected by the Spectrum Observatory over a day at a single location and seek transmitter patterns. In this section we present results from wide-band and long-term analysis of spectrum occupancy using TxMiner. First, we map spectrum occupancy by analyzing the number of transmitters and their type over wide frequency band. We then propose a technique to detect rogue transmitters and utilize it to detect a rogue transmitter in the Spectrum Observatory traces. Finally, we make a case for TxMiner-based support of DSA systems through longitudinal analysis of the DSA opportunity in parts of the UHF band.

A. Mapping spectrum occupancy.

When mapping spectrum occupancy, it is important to look at occupancy states both over a wide frequency range as well as over long time. We now demonstrate TxMiner's capability to support such analysis by drawing a map of transmitters in a wide frequency range.

Mapping number of transmitters. Our analysis of number of transmitters focuses on two frequency bands including 30-173MHz and 700-900MHz. We choose these bands to demonstrate TxMiner's ability to detect the number of transmitters in bands that are typically occupied by narrow-band transmitters such as 30-173MHz and parts of 700-900MHz and other characterized with wide-band transmitters such as portions of 700-900MHz band.

Figure 13(a) plots the number of transmitters detected in each 1MHz chunk. In ranges that are characterized with narrow-band transmissions TxMiner detects up to 4 transmitters in a single 1MHz chunk. In contrast, where wide-band transmitters are present, TxMiner detects contiguous 1MHz chunks as occupied by a single transmitter. Further analysis



(a) Number of transmitters (b) Transmitter type

Fig. 13. Number of transmitters (a) and transmitter type (b) detected over a wide frequency range. TxMiner successfully detects multiple transmitters in a single 1MHz chunk in bands that are characterized with narrow-band transmissions. Simultaneously, TxMiner detects wide-band transmitters by extracting a single transmitter in each 1MHz chunk of a contiguous band. Lastly, TxMiner identifies transmitter type in bands occupied by a single transmitter.

of the powers of the detected transmitters can indicate which 1MHz chunks are occupied by a particular transmitter. In interest of space we omit results of such analysis.

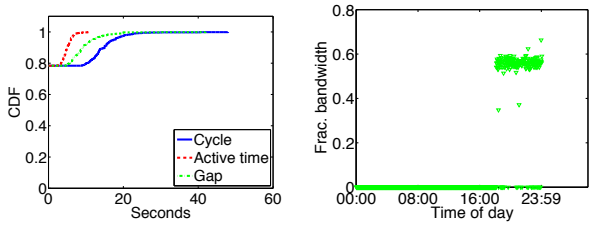
Transmitter type detection. Along with transmitter count we utilize TxMiner to detect transmitter type in a wide frequency band. Figure 13(b) presents a bar-graph with detected transmitter types in 1MHz chunks occupied by a single transmitter. Each bar presents the percentage of transmitter types detected in the two frequency bands of interest. As we can see, majority of the transmitters in both bands are broadcast. We observe a higher percentage of TDMA, FDMA and frequency hopping transmitters in the 700-900MHz band in comparison with the 30-173MHz band. The latter can be explained with the nature of the incumbent transmitters in these bands: while 30-173MHz is characterized with narrow-band broadcast transmissions such as FM radio, the 700-900MHz band hosts technologies such as public safety land mobile communicationsthat are non-broadcast transmissions.

B. Identifying rogue transmitters

To illustrate TxMiner's capability to detect rogue transmitters we define a *rogue coefficients* C_β and $C_\mathcal{T}$ that capture the likelihood that the transmitter sensed in a time-frequency chunk is rogue by analyzing the bandwidth β and active time $\mathcal{T}$ of the detected transmitter. Towards this end we require prior knowledge of the characteristics of the transmitter that is expected to operate in a given band. We note that such prior knowledge can be obtained by considering the previous transmitter characteristics discovered by TxMiner. Thus, our rogue coefficients captures the difference between the expected and the detected transmitter characteristics as follows:

$$C_\beta = \frac{\beta_d}{\beta_e} \text{ and } C_\mathcal{T} = \frac{\mathcal{T}_d}{\mathcal{T}_e} \tag{6}$$

where β_d and β_e are the detected and expected transmitter bandwidth, while $\mathcal{T}_d$ and $\mathcal{T}_e$ are the detected and expected active time. These coefficients vary between 0 and 1, where 1 indicates that the detected and expected transmitters are the same, 0 indicates that there is no transmitter, where a transmitter is expected and anything between 0 and 1 indicates that the detected and expected transmitters are different. Of note is that since this method requires prior observations of incumbent characteristics it will fail detecting rogues if (i) the rogue is spoofing an incumbent, or (ii) if the rogue has the same active time pattern as the incumbent but transmits



(a) Active time (b) Bandwidth

Fig. 14. Characteristics of a single TDMA transmitter over 24 hours. TxMiner successfully identifies transmitter bandwidth and active time over a long period, and can thus inform DSA technologies about the transmission opportunity in a given frequency band.

at different times. We leave a more robust rogue transmitter detection technique as a future work.

We calculate the rogue coefficient for all TV bands and identify that one TV channel is occupied by a non-TV transmitter. The channel in question is channel 20 (506-512MHz), for which TxMiner calculates rogue coefficients $C_\beta = 0.61$ and $C_\mathcal{T} = 0.22$. While the expected transmitter here is a TV broadcast with 6MHz bandwidth and continuous active time, the detected transmitter exhibits different characteristics as captured by the rogue coefficients. A closer look at the occupant indicates that the transmitter has a bandwidth of 4MHz and transmits in a TDMA fashion.

C. Support for DSA systems.

Dynamic spectrum access is a concept most often applied in the context of TV White Spaces, where the primary transmitters have static behavior. Thus these bands are well-suited for database-driven management. There are plethora of radio bands such as radar and satellite bands that are seldom used by their incumbents, which provide a great opportunity for dynamic spectrum access beyond TV White Spaces. However, these bands pose challenges in operation of secondary users due to the highly-dynamic nature of incumbents. In order for secondary users to fully utilize the potential of these bands they need a mechanism to evaluate the transmission opportunity in both frequency and time by assessing not only if there is an incumbent but also how much bandwidth and time does it occupy and whether the temporal occupancy patterns are predictable or not. TxMiner can provide such information. To illustrate how, we analyze one proprietary DSA transmission that exhibits TDMA behavior.

Figure 14 presents our analysis of a 6MHz band (506-512MHz) over 24 hours. TxMiner identifies a single transmitter in this band that is sensed at -94dBm and is active for about 20% of the entire 24-hour period. We analyze the frequency and temporal characteristics of this transmitter in Figure 14(a) and 14(b). In analyzing the temporal characteristics we consider three metrics: (i) the *active time* duration, (ii) the active time *cycle*, that is the time from the beginning of one active period to the beginning of the next active period, and (iii) the *gap* between consecutive active times, that is the time between the end of one active period and the beginning of the next. In Figure 14(a) we plot a CDF of the average active time, cycle and gap in intervals of 100s over the 24-hour period. Since the transmitter is active only 20% of the time, 80% of the values are zero. Based on the values that correspond to transmitter activity we can see that the average duration has a median of 5 seconds and does not vary much over different 100 seconds snapshot. In contrast, the gap has a median of 9 seconds, which is larger than the active time, and varies significantly (from 5 to 42 seconds). Lastly, the cycle has a large variation (between 9 and 48 seconds). These temporal characteristics indicate that the observed transmission is a-periodic and the transmitter is inactive for a larger fraction of the time. Finally, we analyze occupied bandwidth. Figure 14(b) plots the ratio of detected bandwidth vs. analyzed bandwidth (which is 6MHz in this analysis) in each 100s period. As we can see, the fraction of occupied bandwidth is persistently around 0.6, which indicates that 40% of the analyzed band is idle.

This analysis can inform a secondary DSA transmission as follows. If the 80/20 ratio of incumbent presence persists longer than 24 hours, the secondary transmitter can use the entire band for transmission in 80% of the day. In periods where the incumbent is active, due to the a-periodic nature of the incumbent it would be hard to predict opportunities for secondary transmission without real-time sensing. Depending on the sensing efficiency of the secondary transmitter, it can decide whether to opt for sensing and transmission based on the average gap duration supplied by TxMiner. Finally, 2MHz of the 6MHz analyzed band is persistently available, thus the secondary transmitter can decide to utilize this portion continuously without the need of complex sensing techniques if this would satisfy the application requirements.

V. Related Work

Prior work on spectrum analysis can be classified into four categories: wide band spectrum occupancy analysis, envelope detection for identifying unknown signals, multiple signal classification, and detection of transmitters with known signatures.

Several studies have analyzed large-scale spectrum measurements to identify portions of spectrum that are not used [16, 12], or identify patterns of primary users that allow opportunistic spectrum reuse [11, 4]. This body of work assumes no knowledge about the transmitter. They typically apply a threshold for noise, and any signal above this threshold is assumed to be occupied, anything below is assumed to be free. [4] analyses spectrum from China, and models the arrival of users in the cellular bands. [11] analyses spectrum from 30 MHz to 6 GHz, and studies opportunities for dynamic spectrum access in these bands. However, none of these analyses share the goal of TxMiner, and are unable to predict transmitter characteristics from a wideband spectrum trace.

Another set of techniques, which is primarily used by practitioners, is to tease apart unknown transmissions from known transmitters. This is frequently used to identify interferers in the spectrum, for example, in the wireless carrier spectrum. The most common technique is that of envelope detection [8]. A circuit (or these days software) tries to fit a curve around the max-hold (or mean) of signals. Although this technique is useful in determining anomalies in the curves, it does not provide much insight into the distributions that make up the max-hold or mean.

Some classic techniques from the signal processing literature such as MUSIC [15] are also able to detect the number of signals arriving at an antenna. Unlike TxMiner, however, they

do not tackle detailed transmitter characterization, including bandwidth, temporal characteristics and type.

Most closely related to our work are SpecNet [10], DoF [9], AirShark [14] and DECLOAK [13]. SpecNet is a system for large-scale spectrum measurements, which harnesses high-end spectrum analyzers contributed by SpecNet participants and provides basic functionality for SNR-driven occupancy detection. In contrast, our Spectrum Observatory makes use of lower-end spectrum sensors and incorporates TxMiner for advanced transmitter characterization. DoF builds cyclostationary signatures for different transmitters in 2.4 GHz, such as Wi-Fi, Bluetooth, etc., and mines spectrum data for these signatures to determine the users of the spectrum. AirShark tried to solve a similar problem, but using commodity Wi-Fi chipsets. DECLOAK focuses on OFDM transmissions only and uses a combination of cyclostationary features with Gaussian Mixture Models to extract transmitter characteristics. While all three techniques are useful, they only work when the transmitter patterns are known. TxMiner takes the next step, and identifies transmitters when their patterns are not known.

VI. Discussion & Future Work

To summarize, in this paper we have presented the first system, called TxMiner, that is able to mine raw spectrum measurement data and identify properties of transmitters operating in that spectrum. TxMiner is based on a simple observation from signal propagation theory that fading follows a Rayleigh distribution. We use this principle to build machine learning algorithms (RGMM) that attributes spectrum measurements to different transmitters. We use TxMiner to create a spectrum map that features transmitter count and characteristics. We demonstrate detection of rogue transmitters and analysis of DSA opportunity in licensed bands.

Although the knowledge gleaned by TxMiner is very useful, it is still the first step. We believe that many more details can be learnt about transmitters, which will enable several additional applications of spectrum analysis. We list some of our research efforts in this direction below.

Collocated transmitters: Since TxMiner looks at power profiles of transmitters, it is unable to distinguish two collocated transmitters with similar power profiles. In such cases the two transmitters together will be classified as a single transmitter. To this end we can use prior knowledge of the occupants' characteristics to determine the number of active transmitters.

Mobile transmitters: TxMiner is currently unable to detect transmitter mobility. We note that the properties of signal distributions can be applied to this problem as well. Particularly, the signal distributions of mobile transmitters are different than those of static in that they change over time depending on the speed and direction of the transmitter with respect to the RF sensor. Using this observation, we are designing methods for identification of mobility and speed.

Integration with Known Transmitter Signatures: Prior work [9, 14] has looked at identifying transmitters with known temporal signatures. TxMiner can leverage such techniques to eliminate known transmitters from spectrum scans and focus on unknown transmitters, thus improving detection time and accuracy. Furthermore, such knowledge can enhance identification of more complex transmitters such as a mobile performing power control in which case our basic algorithm will detect multiple transmitters.

Despite these limitations, the current implementation of Txminer revolutionizes spectrum mapping by allowing extraction of transmitter count and characteristics, detection of rogue transmitters and identification of opportunities for dynamic spectrum access. Our future research efforts in this direction will open doors toward efficient use and better understanding of spectrum bands nationwide.

References

[1] FCC Notice of Inquiry, ET Docket No. 10-237, 25 FCC Rcd 16632 (2010).

[2] S.3433 - Radio Spectrum Inventory Act of 2012 . https://www.congress.gov/bill/112th-congress/senate-bill/3433.

[3] A. Blake, P. Kohli, and C. Rother. Advances in Markov Random Fields for Vision and Image Processing. MIT Press, 2011.

[4] D. Chen, S. Yin, Q. Zhang, M. Liu, and S. Li. Mining spectrum usage data: A large-scale spectrum measurement study. MobiCom '09, Beijing, China, 2009.

[5] A. Chowdhery, R. Chandra, P. Garnett, and P. Mitchell. Characterizing Spectrum Goodness for Dynamic Spectrum Access. 50th Allerton Conference on Communication, Control and Computing, Monticello, Illinois, October, 2012.

[6] L. Downes. Snowe, Kerry introduce spectrum inventory bill. www.cnet.com/news/snowe-kerry-introduce-spectrum-inventory-bill/.

[7] A. Goldsmith. *Wireless communications*. Cambridge Univ Pr, 2005.

[8] J. Gorin. Detector selection for spectrum analyzer measurements. http://mobiledevdesign.com/, February 2003.

[9] S. S. Hong and S. R. Katti. DOF: A local wireless information plane. SIGCOMM, Toronto, Canada, August, 2011.

[10] A. Iyer, K. K. Chintalapudi, V. Navda, R. Ramjee, V. Padmanabhan, and C. Murthy. Specnet: Spectrum sensing sans frontières. NSDI, Boston, MA, March, 2011.

[11] V. Kone, L. Yang, X. Yang, B. Y. Zhao, and H. Zheng. On the feasbility of effective opportunistic spectrum access. IMC, Melbourne, Australia, November, 2010.

[12] M. A. McHenry, P. A. Tenhula, D. McCloskey, D. A. Roberson, and C. S. Hood. Chicago spectrum occupancy measurements and analysis and a long-term studies proposal. TAPAS, Boston, MA, August, 2006.

[13] N. T. Nguyen, R. Zheng, and Z. Han. On identifying primary user emulation attacks in cognitive radio systems using nonparametric bayesian classification. *Signal Processing, IEEE Transactions on*, 60(3):1432–1445, 2012.

[14] S. Rayanchu, A. Patro, and S. Banerjee. Airshark: Detecting non-WiFi RF Devices Using Commodity WiFi Hardware. IMC '11, Berlin, Germany, 2011.

[15] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, 34(3):276–280, 1986.

[16] M. Wellens and P. Mahonen. Lessons learned from an extensive spectrum occupancy measurement campaign and a stochastic duty cycle model. TRIDENTCOM, Washington D.C., April, 2009.

[17] L. Yang, W. Hou, L. Cao, B. Y. Zhao, and H. Zheng. Supporting Demanding Wireless Applications with Frequency-Agile Radios. NSDI'10, San Jose, California, 2010.

[18] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. IEEE Transactions on Information Theory 51 (7), pages 2282—-2312, July 2005.