FHSS Signal Separation Using Constrained Clustering

Parker D. White*, R. Michael Buehrer†, and William C. Headley‡
Bradley Department of Electrical and Computer Engineering
Hume Center for National Security and Technology
Virginia Polytechnic Institute and State University
Email: {*pwhite18,†buehrer, ‡cheadley}@vt.edu

Abstract—Frequency Hopping Spread Spectrum (FHSS) signaling is used across many devices operating in both regulated and unregulated bands. In either situation, if there is a malicious device operating within these bands, or more simply a user operating out of the required specifications, the identification and separation of this user is important to insure communication link integrity and interference mitigation. Previous signal separation methods often require difficult to obtain hardware fingerprinting characteristics or rough geolocation estimations. This work will consider the detection based characteristics of FHSS signals in addition to background knowledge that is more freely available as a result of spectrum sensing. From estimates of these hopping characteristics alone, novel source separation with classic clustering algorithms can be performed. Background knowledge derived from temporal properties of received waveforms can improve these clustering methods with the novel application of cannot-link pairwise constraints to signal separation. For equivalent clustering accuracy, constraint-based clustering tolerates higher parameter estimation error, caused by diminishing received signal to noise ratio.

Index Terms—FHSS, signal separation, constrained clustering, source identification

I. Introduction

Frequency Hopping Spread Spectrum (FHSS) signaling is a digital communication technique commonly used for its narrow band interference avoidance and its frequency selective fading resistance. FHSS is used both in regulated commercial environments, as well as unregulated commercial and military environments. In each case, if a malicious spectrum user is present, the identification of this user is crucial for threat analysis and communication link integrity. From a commercial point of view, the Industrial, Scientific, and Medical Radio Band (ISM) bands are high-traffic regulated bands. If a device using one of these openly shared bands is violating the terms of it's operating regulations defined in [1] or [2], this device should be identified and separated for further analysis. As cognitive radio grows, and unlicensed band usage increases, this type of spectrum sensing and source identification will be important in maintaining cooperative and collective use of multi-user frequency bands. From a military aspect, knowledge of spectrum use can provide vital information such as non cooperative use of spectrum or environment intelligence in terms of active communicators.

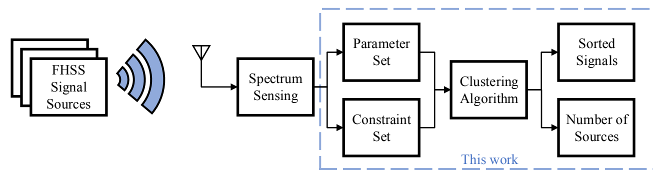

Fig. 1. The process of FHSS signal separation with the use of clustering algorithms as evaluated in this work.

This separation of signal sources can be approached in several different ways. Hardware fingerprinting, or specific emitter identification (SEI), uses RF transmitting chain imperfections such as a transient behavior, quadrature errors [3], amplitude clipping, and/or self-interference [4] to separate sources. While this hardware specific information may be useful in user identification, these hardware imperfections are difficult to calculate and obtain [5]. The current work relies on much simpler generalized parameter sets that can be estimated with just the reception of the source signals, and is not dependent on specific selection of separable criteria as in many device fingerprinting applications. Other techniques, such as TDOA, DOA, or RSS location involves multiple receiving elements cooperating to determine possible source locations in 3-D space as shown in [6]–[8]. However, nearby spectrum users and multi-path effects complicate source separation based strictly on location. Alternatively, the current work will use both hopping characteristics shown in Table I, as well as Time of Arrival (ToA) domain knowledge, to provide a novel and improved solution to the source separation problem, as shown in Fig. 1. Hopping characteristics alone can be used in classic clustering algorithms, but can be improved with the addition of this ToA background information in the form of pairwise cannot link constraints. This improvement not only provides more accurate identification of a malicious user, but also provides a higher error tolerance with similar performance in comparison with classic clustering algorithms using parameter sets alone. In this application, increased error tolerance allows for cheaper receivers, or lower receive Signal to Noise Ratio (SNR).

The rest of this paper is organized as follows. Section

TABLE I
Detection Based Hopping Characteristics

	L-Bound	U-Bound	Clustered	Refs
Start (s)	0.0	-	N	[13]-[18]
Stop (s)	-	0.5	N	[13]-[18]
Dwell (s)	3.125E-4	0.4	Y	[13]-[18]
B_n (kHz)	50	250	Y	[18],[19]
B (kHz)	250	500	Y	[18],[19]
RSS (dBm)	-10	10	Y	[20],[21]

II will cover previous work in FHSS signal separation. In Section III, the simulation model will be discussed. In Section IV, clustering methods are described. In Section V, signal separation results are presented. Finally, future work and concluding remarks are covered in VI and VII, respectively.

II. Background

Previous works have used mixtures of similar hopping parameters to solve different tasks. The work in [9] uses center frequency, bandwidth, power, and duration to classify the type of hopping device currently using the unlicensed ISM band. The method in [9] requires prior knowledge of typical hopping characteristics and is treated much like a classification problem, where there are a set number of source classes. In high traffic unregulated frequency bands, the knowledge of specific device characteristics is typically unavailable, and is not necessary prior knowledge for the method in this current work. The work in [10] detects non-WiFi interfering devices using parameters such as duration, bandwidth, center frequency, spectral signature, timing signature, duty cycle, and pulse spread. This work strictly uses common WiFi hardware for signal detection, but does not account for possible availability of higher quality signal detectors. Alternatively, the current work will measure the impact of device capability on source separation by clustering across a range of receive SNRs. The work in [11] uses frequency offset measurements as well as precise start and stop times to gain information about hopping signals in a band of interest. In addition to accurate timing estimation this work also requires equal power and dwell times from every source, as well as knowledge of frame periods or dwell times of the transmitted signals. In a realistic scenario, consistent receive powers and dwell times from all sources is not likely. While the current work also requires high fidelity in timing estimation, it does not require a reference to specific or ideal center frequencies that were attempted to be transmitted, or equal power and dwell times for all sources. The current work relies strictly on the estimated parameters listed in Table I to perform source separation. The current work does not require prior knowledge of device specific hopping behaviors, and can also separate a variable number of sources and signals. These advantages come in the novel application of pairwise constraint-based clustering. Unlike prior work, the effectiveness of this clustering will be demonstrated with various receive SNRs to demonstrate the impact of variable detector capabilities on source separation.

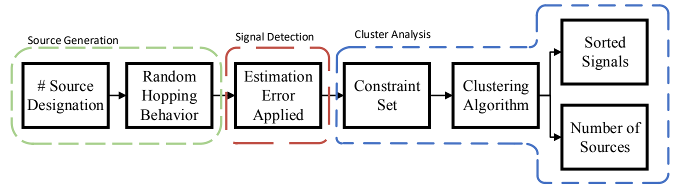

Fig. 2. The simulation of FHSS signal behavior, and the associated detection error prior to the novel application of clustering algorithms for signal separation.

III. System Model

The method of analysis follows the diagram in Fig. 2. First, the number of sources and their respective hopping patterns are randomly designated. Next, hop detection is assumed to be performed on every hop with some amount of parameter estimation error. The parameter sets of each hop is then passed to the constraint-based clustering method proposed in this current work.

A. Hopping Behavior

In this scenario, 3 to 8 hopping sources will be active at a time, each with randomly selected characteristic parameter sets as defined in Table I. These transmit parameters will be selected within the Bluetooth specifications, [1] and [2], and other than dwell time, will not change over consecutive hops or packet arrivals. The dwell time for each hop will vary up to twice the length of the uniformly selected dwell time for each source. The dwell time, RSS, and bandwidth are all selected uniformly in between the lower bound labeled "L-Bound", and the upper bound labeled "U-Bound", in Table I. The initial arrival time will be modeled as a Poisson process with an independent arrival modeled with an exponential distribution [12]. The number of consecutive hops for each source is chosen randomly between 5 and 15. After the number of consecutive hops is defined, the next burst from a source will be modeled with an exponential Random Variable (RV). The center frequency will be pseudo-randomly chosen between 80 channels as defined in [2].

B. Parameter Collection

Prior work demonstrates that parameter estimation accuracy is highly dependent on receive SNR. In this work, estimation accuracy will be swept to represent decreasing receive SNR caused by degrading reception environments. In the current work, this estimation error is modeled with a Gaussian distribution, in the form of a coefficient of variation, defined as $E = \sigma/\mu$. This coefficient of variation E applies error relative to the mean μ by increasing the variance σ of the Gaussian distribution. This ensures that error is applied uniformly across all estimation parameters, listed in Table I. Realistically, the estimation error for each parameter will largely depend on the receiving device, and estimation algorithm used [10], but for ease of analysis, a uniform error is applied to all parameters.

Detection and parameter estimation of FHSS signals used in this work has been previously performed in the literature as follows. The works [13]–[15] show estimation of hop timing and center frequency when periodic hops are present, aperiodic hops are considred in [16], [17]. Spectral properties can be revealed with bandwidth measurements such as total bandwidth or occupied bandwidth, typically measured by -20dB of signal energy [1], and necessary bandwidth, B_n, or 3dB bandwidth which describes the bandwidth of the signal at 3dB below the peak energy. Bandwidth and similarly symbol rate are estimated in the works [18], and [19]. While power estimation is at risk of many channel effects such as multi-path and shadow fading, it may still be helpful in a relatively constant environment. [20] discusses this power estimate and gives a typical range based on [21].

IV. Clustering Techniques

Clustering algorithms are typically robust to variable amounts of input data sizes, as well as output cluster numbers. For this reason, cluster analysis fits this blind problem of source separation well considering the number of hops observed in some time window can vary widely, and the number of sources active in that time window may vary as well. Cluster analysis can be used to partition like hopping signals from one source into one group, while separating dissimilar hops from other sources into other groups. The similarity measure is typically expressed as distance, where minimum distance between points in a cluster is desired for similar points, and maximum distance between points in one cluster to points in another cluster is desired for dissimilar clusters. Some common but very different clustering methods will be shown below.

A. Classic Clustering Algorithms

The clustering algorithms investigated in the current work have been chosen to cover a variety of clustering methods, including density-based DBSCAN [22], graph-based Spectral [23], hierarchical [24], and centroid based [25] clustering. These algorithms will be applied in the typical unsupervised domain where there is an unspecified number of inputs (individual received hops), and a determined number of clusters (signal sources). The performance of each can provide an indication of the best algorithm for this application and can also serve as a baseline for further improvement. In addition to classic clustering algorithms, there have been some modifications that make clustering algorithms slightly more intelligent. An example of this is the introduction of constraints to the K-Means algorithm in [26].

B. Constraint-Based Clustering

In the signal separation domain, background knowledge may come in the form of temporal orthogonality, direction of arrival, or modulation class. Incorporating this information in the form of constraints in the clustering process can improve results when parameter sets become more erroneous. There are several variations of constraint based clustering, however the majority of these algorithms apply constraints during the clustering process in the form of cannot-link or must-link pairwise constraints. Among the previously listed classical clustering algorithms, K-means provides the most elegant transition to incorporating these constraints into the clustering process [26].

Algorithm 1 Modified COP-Kmeans Algorithm

procedure COP-KMEANS(D, Con, K) ▷ Data points D, cannot-link constraints $Con_{\neq} \subseteq D$xD, Clusters K

1: Initialize Cluster Centers $C_1 \dots C_k$ with K-means++.
2: Randomly shuffle all $d \in D$
3: for all points $d \in D$:
   Assign d_i to nearest cluster c_j s.t. VIOLATE-CONSTRAINTS is false.
   If all clusters fail, assign point to left out set.
4: for each cluster $c \in C$:
   compute mean of $D \in c_j$ and assign cluster center to new location μ_j.
5: Add left out set to D and repeat steps (4) and (5) until convergence: $\sum_{n=1}^{j} \Delta\mu_n < threshold$

procedure VIOLATE-CONSTRAINTS$(d_i, C_j, Con_{\neq})$ ▷ Data point d_i, cluster C_j, cannot-link constraints $Con_{\neq} \subseteq D$xD

1: for each $(d_i, d_{\neq}) \in Con_{\neq}$:
2:   if $d_{\neq} \in C$, return true.
3:   else return false.

1) Probabilistic Cannot-Link Constraints: The constraints in the current work are assigned on a pairwise, or point to point, basis. These constraints can require two points to be assigned to the same cluster, as in must-link, or restrict assignment of one point to another cluster, as in cannot-link. In this work, must-link constraint assignment requires absolute certainty that two signals are from the same source. This level of certainty is not available due to the erroneous parameter estimation error used in this work. Cannot link constraints however can be assigned more freely based on the assumption that frequency hopping sources only transmit one hopping signal at a time. If two hopping signals occur in the observation bandwidth at the same time and/or frequency, these signals can be strictly assigned to separate sources with a cannot-link constraint. A hard requirement on constraints is enforced in this work, so no two individual signals can be assigned to the same cluster if there is a cannot-link constraint between them. As a result, as estimation error of parameters increases, points may not be validly assigned due to a constraint existing in every other group. If this happens, the point will be assigned into a null set and will be re-included in the next clustering iteration.

C. COP-Kmeans Clustering

The COP-Kmeans (Constrained K-Means) [26] clustering algorithm, was applied on three general University of California, Irvine datasets, and in a real world GPS lane finding application. The current work makes novel modifications to the K-means algorithm to incorporate instance level pairwise constraints in the application of signal separation. The algorithm with modifications is shown in Algorithm 1.

1) Modifications: There are some additional considerations that can improve the classification performance and reduce the number of unassigned points in the algorithm proposed in [26]. First, the order of points can influence the quality of point assignment when a pairwise constraint in violated. This is due to the lack of decision making when a constraint is violated. Shuffling points at the beginning of each algorithm iteration can reduce the impact of this ordering. Also, increasing initial cluster center distances can improve clustering convergence time and overall performance. Kmeans++ initialization [27], achieves this by assigning other cluster centers with probability $P(d)^2$ where $P(d)$ is the distance from each point d, to the current cluster center. This method reduces the random bias caused by random assignment of cluster centers as in the original algorithm. The work [26] also does not consider the number of points that are in violation with every other cluster. This work simply states that the points are assigned to a null set which could skew performance results if a large number of data points are unassigned. In contrast, the current work includes the null set back into the assignment stage in step 5 of Algorithm 1. This inclusion provides more chances for every point to be validly assigned to a cluster, and thus reduces the number of points unassigned prior to algorithm convergence. In the case of convergence with points remaining in the null set, these points can either be forcefully assigned to the nearest cluster center, or left out for further evaluation.

2) Execution: This algorithm functions with inputs D, cannot link constraints Con, and number of clusters K. Each input d_i is a signal hop observed in the time-frequency window. Each pairwise cannot-link constraint is derived from temporal overlaps in the time-frequency observation interval. In this application, very accurate estimates of start and stop time are assumed, and therefore can be used to define these cannot-link constraints between hops. With input signals D, constraints Con, and number of clusters C defined, clustering as in Algorithm 1 is executed. Once the cluster centers do not change, or no points are assigned to alternate clusters, the algorithm has converged with K clusters containing all points not including the null set. The K cluster assignment represents separate sources in this application, and can be used as a means of identifying typical hopping characteristics of every independent source. As will be shown, accurate cluster assignment increases the uncertainty about an individual source's hopping behavior which is improved overall with the addition of time based constraints.

V. Simulation Results

The analysis in this section will be performed with respect to the system model outlined in Section III, and outlined in Fig. 2. This model simulates what an observable time and frequency window may look like in a realistic scenario. With this in mind, a variable amount of hopping sources may be active at a time, each with independent hopping behaviors. Signals from each of these hopping sources are assumed to be detected, and the spectrum sensing based characteristics defined in Table I are estimated with some error outlined in Section III-B. This section will demonstrate the effectiveness of clustering these hopping signals based on the similarity and differences of each hops set of estimated hopping parameters. Classic clustering techniques and modified constraint based clustering are the methods used for this signal separation, as outlined in Section IV.

A. Performance Metrics

1) Normalized Mutual Information: Normalized mutual information (NMI) [28] is commonly used for clustering performance evaluation [29]. This measure is defined as

$$NMI = \frac{MI(L,P)}{E[H(L),H(P)]} \tag{1}$$

where L is the vector of true labels of the data set, P is a vector of arbitrarily assigned labels as a result of cluster groupings, MI is the mutual information between L and P and H is the entropy of either set. Mutual information can be described as a measure of information, the clustering labels, contains about the true labels. While NMI provides a mathematical evaluation of clustering performance, an additional measure of accuracy is included in this work to give a relatable measure of the percent of correctly assigned labels.

2) Accuracy: A measure of raw accuracy, or purity [30], is used as another performance benchmark. With a set of clusters C, classes L, and N total data points D, accuracy is defined as follows.

$$Accuracy(C,D,L) = \frac{1}{N}\sum_{c\in C}\max_{l\in L}|c\cap d| \tag{2}$$

For every cluster, the maximum number of one sources signals in that cluster, is assigned to be that clusters true label. With this assignment, a percentage can be derived by the number of correct labels and total number of labels.

B. Clustering Results

The figures shown in this section are all results of a Monte Carlo Simulation. For Fig. 3, 4, and 5, K was swept from 3 to 8, and the best result out of 20 algorithm runs was chosen for each Monte Carlo iteration based on the minimum number of unassigned points. The following sections will discuss the takeaways of these results.

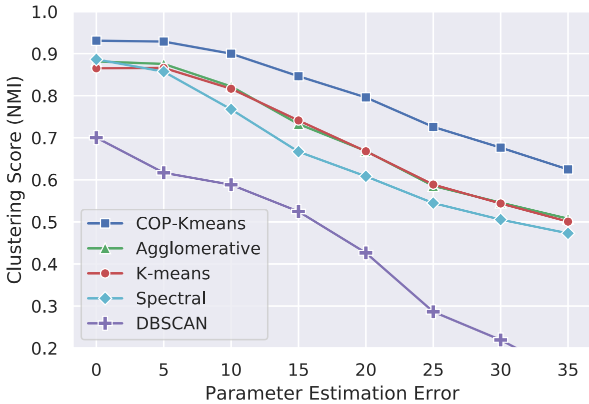

Fig. 3. Comparison of Normalized Mutual Information across clustering algorithms as a function of increasing coefficient of variation to represent degrading FHSS signal receive SNR.

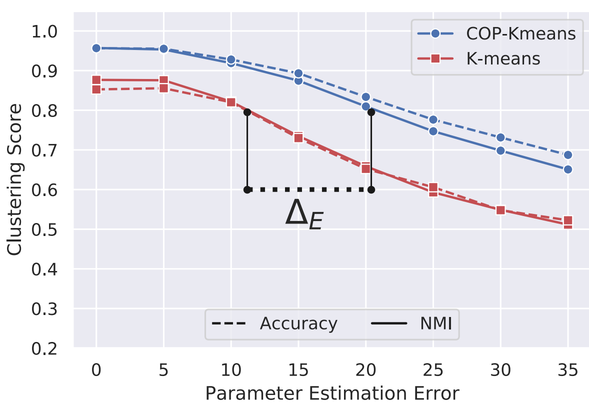

Fig. 4. The increase in estimation error tolerance of the K-means algorithm with pairwise constraints applied in the COP-Kmeans algorithm.

1) Algorithm Comparison: The clustering NMI of algorithms are shown in Fig. 3. As can be seen, the best performing were agglomerative and K-means. For this reason, the K-Means algorithm was chosen as the baseline in clustering performance. The addition of pairwise constraints, shown with the COP-Kmeans result demonstrates the benefit of incorporating background information in the form of constraints. A closer comparison of the performance benefit is shown in Fig. 4, and is discussed in the next section.

2) Accuracy and NMI vs Error: The performance gain in this analysis is expressed in terms of NMI and Accuracy as previously stated in Section V-A. For this analysis, the values of K clusters will be correctly assigned prior to clustering. The results in Fig. 4 demonstrate the improvement in using pairwise cannot link constraints in clustering rather than strictly a classic clustering algorithm. The separation distance Δ_E demonstrates the tolerance of parameter estimation error with equivalent performance in signal clustering. This separation demonstrates an error

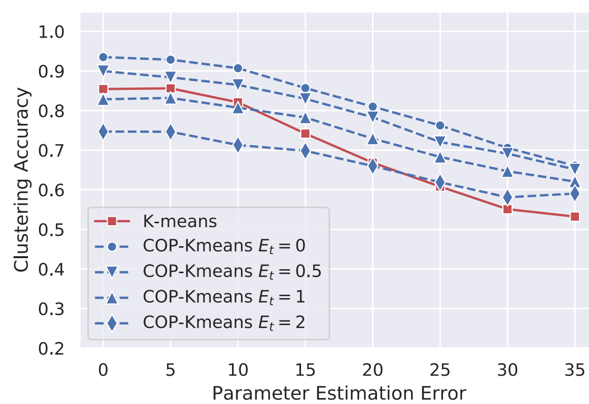

Fig. 5. The impact of erroneous cannot-link constraints on clustering performance (NMI), caused by hop timing estimation error E_t.

tolerance of nearly 10 standard deviations when using constraints in the clustering process. In a scenario where receive signal quality is essential in clustering performance, increased error tolerance will allow for lower quality measurements that are typically available in cheaper receivers, as well as lower receive SNRs for equivalent clustering performance.

3) Impact of Erroneous Constraints: Up to this point, the results in this current work considered the pairwise constraints as hard and completely accurate. These hard constraints are dependent on very accurate hop timing estimation, which will not be perfect in a realistic scenario. Fig. 5 shows the impact of increased start and stop time estimation error which causes false cannot-link constraints assignments that. As seen in the figure, the approach in this current work can tolerate some amount of timing errors but with increased amounts of timing error, K-means becomes the better performing algorithm as expected. In scenarios where GPS level accuracy is available for timing estimates, the COP-Kmeans algorithm can tolerate higher parameter estimation error with equal clustering performance compared to classic clustering algorithms.

VI. Future Work

The constraints used in the current work are dependent on accurate hop timing estimates. As these estimates become more erroneous, hard constraints can get falsely assigned and decrease clustering performance as shown in Fig. 5. Rather than removing constraints altogether when accurate hop timing estimates are unavailable, soft constraints could be used instead. Unlike the strict requirement that constraints must be satisfied in the COP-Kmeans algorithm, soft constraints could be violated with a weighted penalty. This penalty would be applied in the point assignment stage in Algorithm 1, and would be balanced with point to point similarity. The work [26] is expanded in [31] which incorporates soft constraints to the previous COP-Kmeans algorithm.

VII. Conclusions

This work has highlighted detection based FHSS signal characteristics that can be estimated and used to separate signals. These characteristics include bandwidth, dwell time, hop instants, Time of Arrival, and Received Signal Strength. Other works have estimated and used these parameters for signal classification, but have been limited to a set number of known sources. This work also removed some prior knowledge in signal characteristics and demonstrated that classic clustering methods can be used to separate signals. DBSCAN, Spectral, agglomerative, and K-means clustering were compared for this problem data. Since K-means outperformed others, K-means was chosen as a baseline and further improved with the novel application of pairwise cannot-link constraints derived from accurate knowledge of time-frequency spectral properties of hopping signals. Improvement in clustering accuracy allows for higher parameter estimation error which is a result of lower receive SNR. Overall, with accurate measurements of hop timing and erroneous measurements of other parameters listed in Table I, clustering methods can be used to separate sources, and improved with the addition of background knowledge.

References

[1] "Federal communications commission (2019). electronic code of federal regulations. pp.title 47 chapter 1 subchapter a part 2 subpart c 2.202.," 2019.

[2] Bluetooth Special Interest Group, Bluetooth Core Specification 5.1. 2019.

[3] L. J. Wong, W. C. Headley, S. Andrews, R. M. Gerdes, and A. J. Michaels, "Clustering learned cnn features from raw i/q data for emitter identification," in MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM), pp. 26–33, Oct 2018.

[4] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, "Wireless device identification with radiometric signatures," in Proceedings of the 14th ACM International Conference on Mobile Computing and Networking, MobiCom '08, (New York, NY, USA), pp. 116–127, ACM, 2008.

[5] K. Talbot, P. Duley, and M. Hyatt, "Specific emitter identification and verification," 01 2003.

[6] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal, "Locating the nodes: cooperative localization in wireless sensor networks," IEEE Signal Processing Magazine, vol. 22, pp. 54–69, July 2005.

[7] R. M. Vaghefi, M. R. Gholami, and E. G. Ström, "Rss-based sensor localization with unknown transmit power," in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2480–2483, May 2011.

[8] S. Tomic, M. Beko, and R. Dinis, "Rss-based localization in wireless sensor networks using convex relaxation: Noncooperative and cooperative schemes," IEEE Transactions on Vehicular Technology, vol. 64, pp. 2037–2050, May 2015.

[9] K. M. Neil Diener, "System and method for spectrum management of a shared frequency band," September 2007.

[10] S. Rayanchu, Wireless and Systems for Understanding Wireless Interference. PhD thesis, University of Wisconsin-Madison, 2012.

[11] A. Gök, S. Joshi, J. Villasenor, and D. Cabric, "Estimating the number of frequency hopping interferers using spectral sensing with time and frequency offset measurements," in MILCOM 2009 - 2009 IEEE Military Communications Conference, pp. 1–7, Oct 2009.

[12] V. S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," IEEE Communications Magazine, vol. 32, pp. 70–81, March 1994.

[13] S. Barbarossa and A. Scaglione, "Parameter estimation of spread spectrum frequency-hopping signals using time-frequency distributions," in First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications, pp. 213–216, April 1997.

[14] M. K. Simon, U. Cheng, L. Aydin, A. Polydoros, and B. K. Levitt, "Hop timing estimation for noncoherent frequency-hopped m-fsk intercept receivers," IEEE Transactions on Communications, vol. 43, pp. 1144–1154, Feb 1995.

[15] L. Aydin and A. Polydoros, "Hop-timing estimation for fh signals using a coarsely channelized receiver," IEEE Transactions on Communications, vol. 44, pp. 516–526, April 1996.

[16] D. Angelosante, G. B. Giannakis, and N. D. Sidiropoulos, "Estimating multiple frequency-hopping signal parameters via sparse linear regression," IEEE Transactions on Signal Processing, vol. 58, pp. 5044–5056, Oct 2010.

[17] X. Liu, J. Li, and X. Ma, "An em algorithm for blind hop timing estimation of multiple fh signals using an array system with bandwidth mismatch," IEEE Transactions on Vehicular Technology, vol. 56, pp. 2545–2554, Sep. 2007.

[18] D. Cohen, L. Pollak, and Y. C. Eldar, "Carrier frequency and bandwidth estimation of cyclostationary multiband signals," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3716–3720, March 2016.

[19] X. Jun, W. Fu-ping, and W. Zan-ji, "The improvement of symbol rate estimation by the wavelet transform," in Proceedings. 2005 International Conference on Communications, Circuits and Systems, 2005., vol. 1, pp. 100–103 Vol. 1, May 2005.

[20] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal, "Locating the nodes: cooperative localization in wireless sensor networks," IEEE Signal Processing Magazine, vol. 22, pp. 54–69, July 2005.

[21] T. Rappaport, Wireless Communications: Principles and Practice. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2nd ed., 2001.

[22] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise.," in Kdd, vol. 96, pp. 226–231, 1996.

[23] U. von Luxburg, "A tutorial on spectral clustering," Statistics and Computing, vol. 17, pp. 395–416, Dec 2007.

[24] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press, 2008.

[25] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 881–892, July 2002.

[26] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, (San Francisco, CA, USA), pp. 577–584, Morgan Kaufmann Publishers Inc., 2001.

[27] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07, (Philadelphia, PA, USA), pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[28] T. M. Cover and J. A. Thomas, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). New York, NY, USA: Wiley-Interscience, 2006.

[29] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," J. Mach. Learn. Res., vol. 11, pp. 2837–2854, Dec. 2010.

[30] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," Natural Language Engineering, vol. 16, no. 1, pp. 100–103, 2010.

[31] K. Wagstaff, Intelligent Clustering with Instance-Level Constraints. PhD thesis, Cornell University, 2002.