Computer Networks 182 (2020) 107566

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Data article

Massive-Scale I/Q Datasets for WiFi Radio Fingerprinting

Amani Al-shawabka *, Francesco Restuccia, Salvatore D'Oro, Tommaso Melodia

Institute for the Wireless Internet of Things, Northeastern University, Boston, MA 02115, USA

ARTICLE INFO

Keywords:
Software-Defined Radios
Radio Fingerprinting
Wireless Channel Conditions
Internet of Things

ABSTRACT

Recent research has proved the effectiveness of neural networks (NNs) in "fingerprinting" (*i.e.*, identifying) wireless radios, by determining the hardware impairments emitted from the transmitter during the waveform transmission process. The artificial neurons of the NN layers are employed to identify and track the radios' unique impairments by training a large amount of raw data released from these radios. Today, the radio fingerprinting field lacks such a large-scale waveform database that can provide a standard benchmark for researchers working on this field. In this paper, we publicly share 2TB of IEEE 802.11 a/g (WiFi) data obtained from 20 bit-similar Software-Defined-Radios (SDRs).

Specifications table

Subject	Artificial Intelligence.
Specific subject area	Radio fingerprinting techniques based on deep learning algorithms.
Type of data	IEEE 802.11 a/g (WiFi) I/Q datasets.
How data were acquired	Hardware: SDR running through Gnuradio. Our datasets were collected using 12 NI N-210 and 8 X-310 transmitters, as well as 1 N-210 receiver, each SDR equipped with a CBX daughterboard.
Data format	Raw and equalized I/Q samples. Each recorded transmission consists of two files: (i) a binary file of the recorded digital samples and (ii) a metafile that contains information describing each dataset in plain-text JSON format. Our binary and meta format is an extension of, and compatible with the SigMF specifications and [1].
Parameters for data collection	Our campaign was carried out over (i) several days, (ii) diverse environments (Arena [2] "in-the-wild" and anechoic chamber testbeds), and (iii) examining different channel conditions (wireless with different antennas, wireless with a single antenna, and wired connection).
Description of data collection	We used Bloessl et al. [3] model to stream a WiFi baseband signal with 2.432 GHz center frequency, 20 MS/s sampling rate, BPSK modulation, and 1/2 coding scheme. The WiFi frame is repeated over and over again for 30 s. Fig. 1 shows the data collection process methodology at the receiver side through Gnuradio. The received I/Q samples are passed through a Short Training Sequence (STS) and Long Training Sequence (LTS) processes used to detect the received WiFi frame and to accomplish the time synchronization operation, respectively. The received I/Q samples stored at three different demodulation stages.
Data source location	Institution: Northeastern University Testbed City: Two locations (i) Boston and (ii) Burlington Country: USA
Data accessibility	Repository name: https://repository.library.northeastern.edu/ Direct URL to data: To download our datasets please use the following links: • **"Arena Wireless Different Antennas":** "Setup 1" • **"Arena Wireless Single Antenna":** "Setup 2" • **"Arena Wired":** "Setup 3" • **"Anechoic Chamber Wireless Single Antenna":** "Setup 4"

(*continued on next page*)

* Corresponding author.
E-mail address: al-shawabka.a@northeastern.edu (A. Al-shawabka).

https://doi.org/10.1016/j.comnet.2020.107566
Received 19 August 2020; Received in revised form 14 September 2020; Accepted 19 September 2020
Available online 23 September 2020

(*continued*)

Subject	Artificial Intelligence.
Related research article	Author's name: Al-Shawabka, Amani and Restuccia, Francesco and D'Oro, Salvatore and Jian, Tong and Rendon, Bruno Costa and Soltani, Nasim and Dy, Jennifer and Chowdhury, Kaushik and Ioannidis, Stratis and Melodia, Tommaso. Title: Exposing the Fingerprint: Dissecting the Impact of the Wireless Channel on Radio Fingerprinting. Journal: Proc. of IEEE Conference on Computer Communications (INFOCOM).
Related project	Project name: RFMLS program. Funding body: Defense Advanced Research Projects Agency (DARPA). Project duration: 2 years.

Value of the data

- **Why are these data useful?** The radio fingerprinting process includes: (i) capturing large-scale labeled waveforms for each radio, (ii) extracting the unique radio's features, and (iii) employing the captured features in identifying the transmitter when a new waveform is received [4]. Before this work, the wireless community in the radio fingerprinting domain lacked a rich and diverse large-scale database for research benchmark activities. Sankhe et al. [1] released datasets close to ours. However, our datasets differ in many aspects (i) we employed 20 different transmitters, (ii) our datasets collected from different demodulation stages simultaneously, and (iii) we examined different combinations of channel conditions and environments.
- **Who can benefit from these data?** Radio fingerprinting based on deep learning algorithms research community.
- **How can these data be used for further insights and development of experiments?** Radio fingerprinting developers must assess and evaluate their deep learning algorithms based on standard and large-scale datasets. Our datasets are incredibly beneficial to perform such benchmarking activities.
- **What is the additional value of these data?** Our datasets are composed of several I/Q samples collected at different demodulation stages, which assist the radio fingerprinting researchers in understanding each stage's impact in boosting the radio fingerprinting accuracy. This massive data collection campaign was carried out (i) over several days, (ii) diverse environments, and (iii) different channel conditions. By considering bit-similar devices transmit the same signal repeatedly, researchers can examine the worst-case scenario and determine the conditions where the hardware impairments are still noticeable by their algorithm.

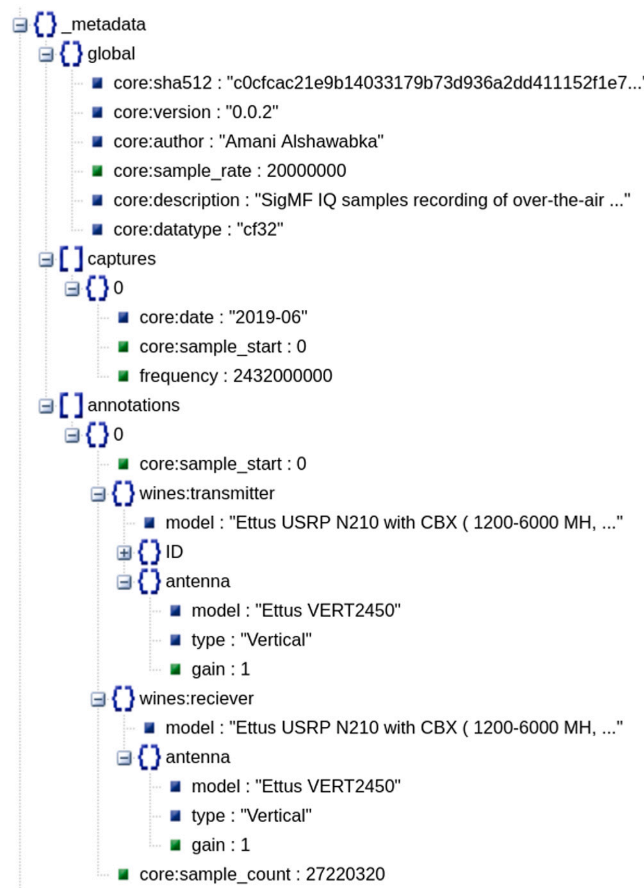

Fig. 2. SigMF's file sample viewed by http://jsonviewer.stack.hu/.

1. Data

Fig. 1 summarizes the I/Q data collection process at the receiver side through Gnuradio. Reader may refer to [5] for more details about the data collection process. Fig. 2 shows a sample of our SigMF representations. It displays the metadata fields that describe each of our binary files.

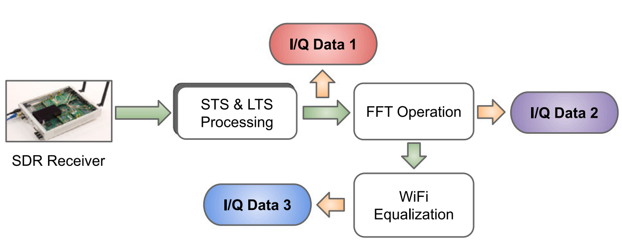

Fig. 1. I/Q data collection methodology [5].

2. Experimental design, materials, and methods

This massive campaign was carried out over (i) several days, (ii) diverse environments, and (iii) different channel conditions. To explore all the circumstances associated with our experiments and clearly understand the data collection process methodology, the testbeds characteristics, and other detailed descriptions, the reader may refer to [5].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Defense Advanced Research Projects Agency (DARPA) under RFMLS program contract N00164-18-R-WQ80.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of DARPA or the U.S. Government.

References

[1] K. Sankhe, M. Belgiovine, F. Zhou, S. Riyaz, S. Ioannidis, K. Chowdhury, ORACLE: Optimized Radio clAssification through Convolutional neuraL nEtworks, in: IEEE INFOCOM 2019-IEEE Conference on Computer Communications, IEEE, 2019, pp. 370–378.

[2] L. Bertizzolo, L. Bonati, E. Demirors, T. Melodia, Arena: A 64-antenna SDR-based Ceiling Grid Testbed for Sub-6 GHz Radio Spectrum Research, in: Proceedings of the 13th International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization, 2019, pp. 5–12.

[3] B. Bloessl, M. Segata, C. Sommer, F. Dressler, An IEEE 802.11 a/g/p OFDM receiver for GNU radio, in: Proceedings of the Second Workshop on Software Radio Implementation Forum, ACM, 2013, pp. 9–16.

[4] N. Soltanieh, Y. Norouzi, Y. Yang, N.C. Karmakar, A review of radio frequency fingerprinting techniques, IEEE J. Radio Freq. Identif. (2020).

[5] A. Al-Shawabka, F. Restuccia, S. D'Oro, T. Jian, B.C. Rendon, N. Soltani, J. Dy, K. Chowdhury, S. Ioannidis, T. Melodia, Exposing the Fingerprint: Dissecting the Impact of the Wireless Channel on Radio Fingerprinting, in: Proc. of IEEE Conference on Computer Communications, INFOCOM, 2020.

Amani Al-shawabka is currently a Ph.D. candidate at Northeastern University in the Electrical and Computer Engineering department. She received her Master's degree in Computer Engineering from Northeastern University in 2019. She started her professional career in mobile network companies (Orange and Umniah-Batelco groups) after receiving her B.S. degree in Communication Engineering from Yarmouk University, Jordan. Her research interests are wireless networks, wireless fingerprinting, embedded systems and network security.

Francesco Restuccia received his Ph.D. in Computer Science from Missouri S&T, Rolla, MO, USA, in 2016, and his M.S. and B.S. in Computer Science and Engineering with highest honors from the University of Pisa, Italy in 2011 and 2009, respectively. Currently, he is an Assistant Professor of Electrical and Computer Engineering at Northeastern University, USA. His research interests lie in the modeling, analysis, and experimental evaluation of wireless networked systems. Dr. Restuccia has published over 25 papers in top venues, as well as co-authoring 8 pending US patents and 2 book chapters. He regularly serves as a TPC Member for conferences such as IEEE INFOCOM and ACM MobiHoc, and is a reviewer for several ACM and IEEE conferences and journals. Dr. Restuccia is the recipient of the 2019 ISSNAF Mario Gerla Award for Young Investigators in Computer Science. He is a Member of the IEEE and the ACM.

Salvatore D'Oro received his Ph.D. degree from the University of Catania in 2015. He is currently a Research Associate Professor at Northeastern University. In 2015, 2016 and 2017 he organized the 1st, 2nd and 3rd Workshops on COmpetitive and COoperative Approaches for 5G networks (COCOA). He also served on the Technical Program Committee (TPC) of the IEEE Conference on Standards for Communications and Networking (CSCN'18), Med-Hoc-Net 2018 and the CoCoNet8 workshop at IEEE ICC 2016. He serves on the TPC of Elsevier Computer Communications journal. Dr. D'Oro is also a reviewer for major IEEE and ACM journals and conferences. Dr. D'Oro's research interests include game-theory, optimization, learning and their applications to telecommunication networks. He is a Member of the IEEE.

Tommaso Melodia is the William Lincoln Smith Professor with the Department of Electrical and Computer Engineering at Northeastern University in Boston. He is also the Founding Director of the Institute for the Wireless Internet of Things and the Director of Research for the PAWR Project Office. He received his Ph.D. in Electrical and Computer Engineering from the Georgia Institute of Technology in 2007. He is a recipient of the National Science Foundation CAREER award. Prof. Melodia is the Editor in Chief of Computer Networks. He has served as Technical Program Committee Chair for IEEE Infocom 2018, General Chair for IEEE SECON 2019, ACM Nanocom 2019, and ACM WUWnet 2014. Prof. Melodia is the Director of Research for the Platforms for Advanced Wireless Research (PAWR) Project Office, a $100M public–private partnership to establish 4 city-scale platforms for wireless research to advance the US wireless ecosystem in years to come. Prof. Melodia's research on modeling, optimization, and experimental evaluation of Internet-of-Things and wireless networked systems has been funded by the National Science Foundation, the Air Force Research Laboratory the Office of Naval Research, DARPA, and the Army Research Laboratory. Prof. Melodia is a Fellow of the IEEE and a Senior Member of the ACM.