Challenges in RF Data Acquisition, Management, Dissemination

Session 3: Mobile and Wireless Measurement Challenges

Kaushik Chowdhury
Professor and Associate Director for Institute for the Wireless IoT
Northeastern University
Community Pain Points [for the Wireless Community]

- No representative datasets for the wireless community
  - MNIST for image processing community

- Growing interest in ML/AI, hampered by lack of diverse datasets
  - No clear categorization of technology options that intuitively leads to discovery of resources
  - Need for a centralized “focal point” for the wireless community

- Lack of tools to create datasets on wireless testbeds
  - External testbeds are complex, need time investment
  - Privacy concerns

- Lack of tools to share, disseminate datasets
  - Lack of community leadership for standardized metadata representation and sharing
  - Lack of education and awareness for wireless engineers
Sources of Wireless Datasets

- NTIA
- NRAO
- NIST
- FCC
- User Testbeds & Simulation
- Colosseum
- NSF PAWR Platforms

```
<table>
<thead>
<tr>
<th>Institutional, Agency Controlled</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTIA</td>
</tr>
<tr>
<td>NIST</td>
</tr>
</tbody>
</table>

```

```
<table>
<thead>
<tr>
<th>“Virtual world” emulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colosseum</td>
</tr>
</tbody>
</table>

```

```
<table>
<thead>
<tr>
<th>“Community-scale” platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF PAWR Platforms</td>
</tr>
</tbody>
</table>

```

New Opportunities
NSF PAWR Platforms – Rich Sources of Datasets

**NSF PAWR PLATFORMS**

**POWDER**
Salt Lake City, UT
Software defined networks and mMIMO

**COSMOS**
West Harlem, NY
mmWave, full duplex and backhaul research

**AERPAW**
Raleigh, NC
Unmanned aerial vehicles, mmWave and mobility

**COLOSSEUM**
Boston, MA
Spectrum sharing, mMIMO, mobility, IoT, mesh networks
Example Problem: RF Fingerprinting

Learn unique RF impairments (DC offset, IQ imbalance etc.)

Changed waveform

Hardware or software

Trust
Example Measurement and Dataset Generation

POWDER PAWR platform University of Utah campus, in Salt Lake City, spanning 6 km², composed of SDRs and open-source software stacks.

Reference: https://powderwireless.net
Datasets for RF Fingerprinting on the POWDER Platform

Download Datasets:
Please use below link to download the dataset:

Dataset#1: Raw IQ samples of over-the-air transmissions from 4 Base Stations deployed in the POWDER platform, in Salt Lake City, Utah, USA. These datasets were used for the paper "Trust in 5G Open RANs through Machine Learning: RF Fingerprinting on the POWDER PAWR Platform", IEEE GLOBECOM 2020. Any use of this dataset, which results in an academic publication or other publication that includes a bibliography, should contain a citation to our paper. Here is the reference for the work:

Conference version: PDF

https://genesys-lab.org/powder
Colosseum

- 900 TB of Network Attached Storage (NAS)
- 171 high-performance servers
- 256 USRP X310s (128 as communications devices, 128 as part of the channel emulator)
- 18 10G switches
- 19 clock distribution systems
- 25.6 GHz total instantaneous bandwidth
- 52 TB/s of digital RF data
- 320 FPGAs
- Hundreds of high-speed optical connections
- Software-based traffic generation solutions hosted on a pool of dedicated servers
- Full-mesh networking capability
- 21 racks of radios, FPGAs, servers and support equipment
Example Experiment Design

Legend

- Management network
- Internet emulation
- Traffic network
- External connection
- Other
- Reservation A - Channels 0, 1, 4, 5
- Reservation B - Channels 2, 3, 6, 7, 8, 9
- Connectivity Disabled

Not Pictured: High-Speed Inter-Quad Connections
Example Measurements & Dataset

- Cellular network w/ srsLTE: 6 interfering base stations w/ 24 users
- Downlink video streaming
- Pedestrian user mobility
- Real-world scenario with base station locations in Boston Public Garden
Open Challenges: Technical

Create representative community dataset(s)
- MNIST for wireless
- One central stop for the community

Accelerate 6G research
- Real-time and offline dataset generation, processing tools

Influence design of future experimental testbeds
- Show feasibility of native support/APIs for community testbeds
## Open Challenges: Education and Workforce Dev

### Train “tomorrow’s” wireless engineer
- Train wireless/networking workforce on how to build, store, share datasets
- Impact broad curriculum changes

### Bridge communities
- Provide resources, competitions for wireless engineers to build collaborations with ML/AI experts
- Engage with industry as they roll out new tools

### Democratize access to wireless datasets
- Enable anyone to create datasets with minimal expert knowledge
- New hands on activities K-12 and beyond
Community Needs

Needs?

Building a centralized, searchable repository, taxonomy of real-world static datasets
- PAWR
- NTIA/NIST
- User Testbeds

Creating, curating static datasets

Dataset generation tools
- APIs for PAWR
- APIs for Colosseum
- Offline- Python/GNURadio based tools that run on CPU

Dataset storage, pre-processing resources
- AWS
- CloudLab
- Automated scripts for spectrum activity detection/filtration

Dataset validation/processing tools
- Privacy preserving methods (MACID removal etc)
- SigMF metadata representation
- "Strength of dataset" outcomes before generation
- Online- CuSignal-based tools from NVIDIA that uses GPU acceleration
Thank You