User Activity Detection in Grant-Free Massive Random Access

Prof. Chandra R Murthy
Dept. of ECE, Indian Institute of Science

Algorithmic Structures for Uncoordinated Communications and Statistical Inference in Exceedingly Large Spaces
Mar. 12, 2024

Disclaimer

See these talks

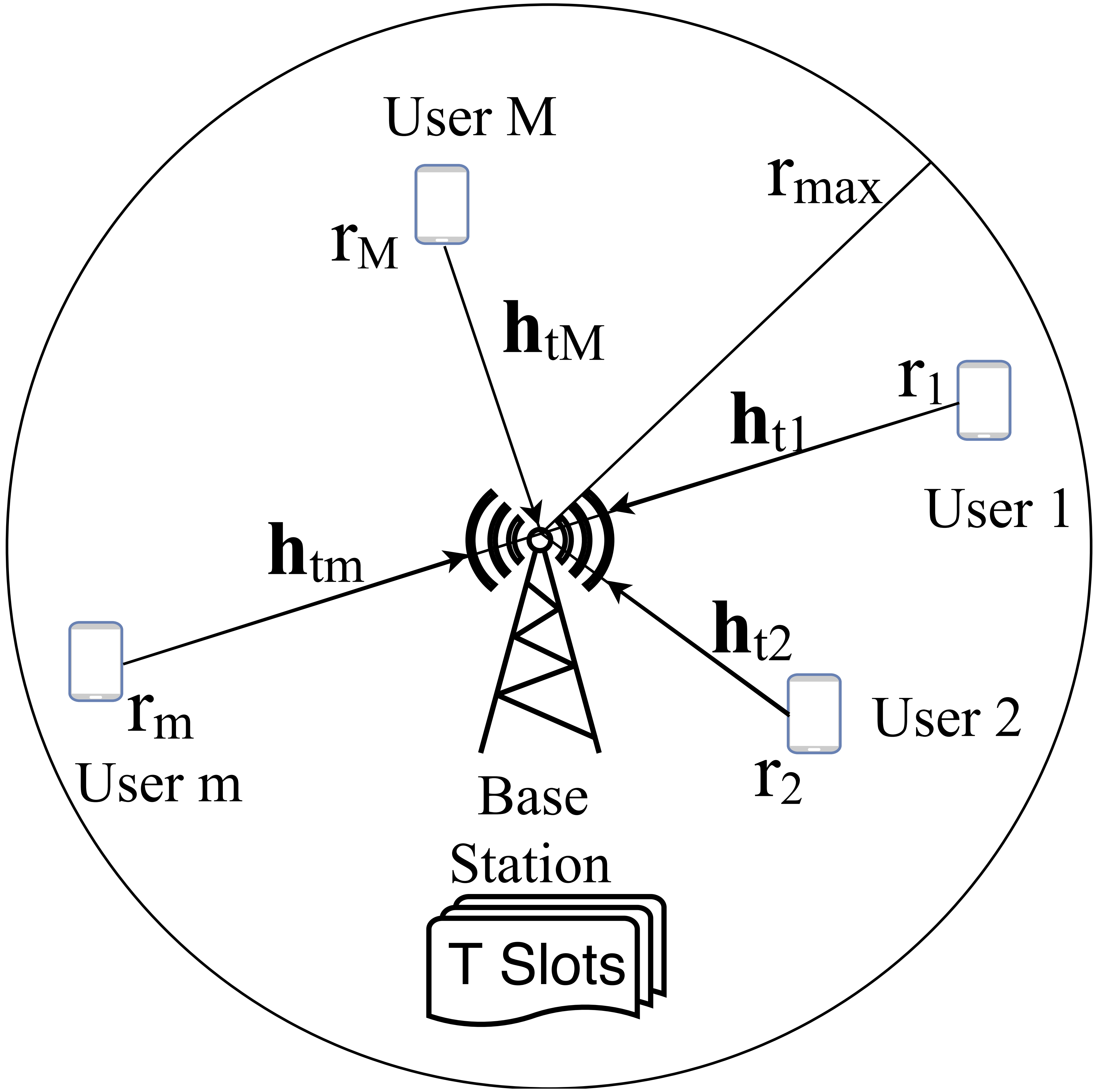
- Gianluigi Liva
- Alex Fengler
- Wei Yu
- Ya-Feng Liu
- And others to follow …

Massive Random Access (mRA)

- 5G, 6G, and beyond-5G applications: eMBB, URLLC, and Massive Random Access (mRA)
- mRA expected to serve millions of devices/km^2
  - mRA devices are sporadically active and transmit short packets
- Grant-free random access (GFRA) protocols can efficiently serve mRA
  - Advantages: Low control overhead, non-orthogonal use of channel

Challenges in mMTC

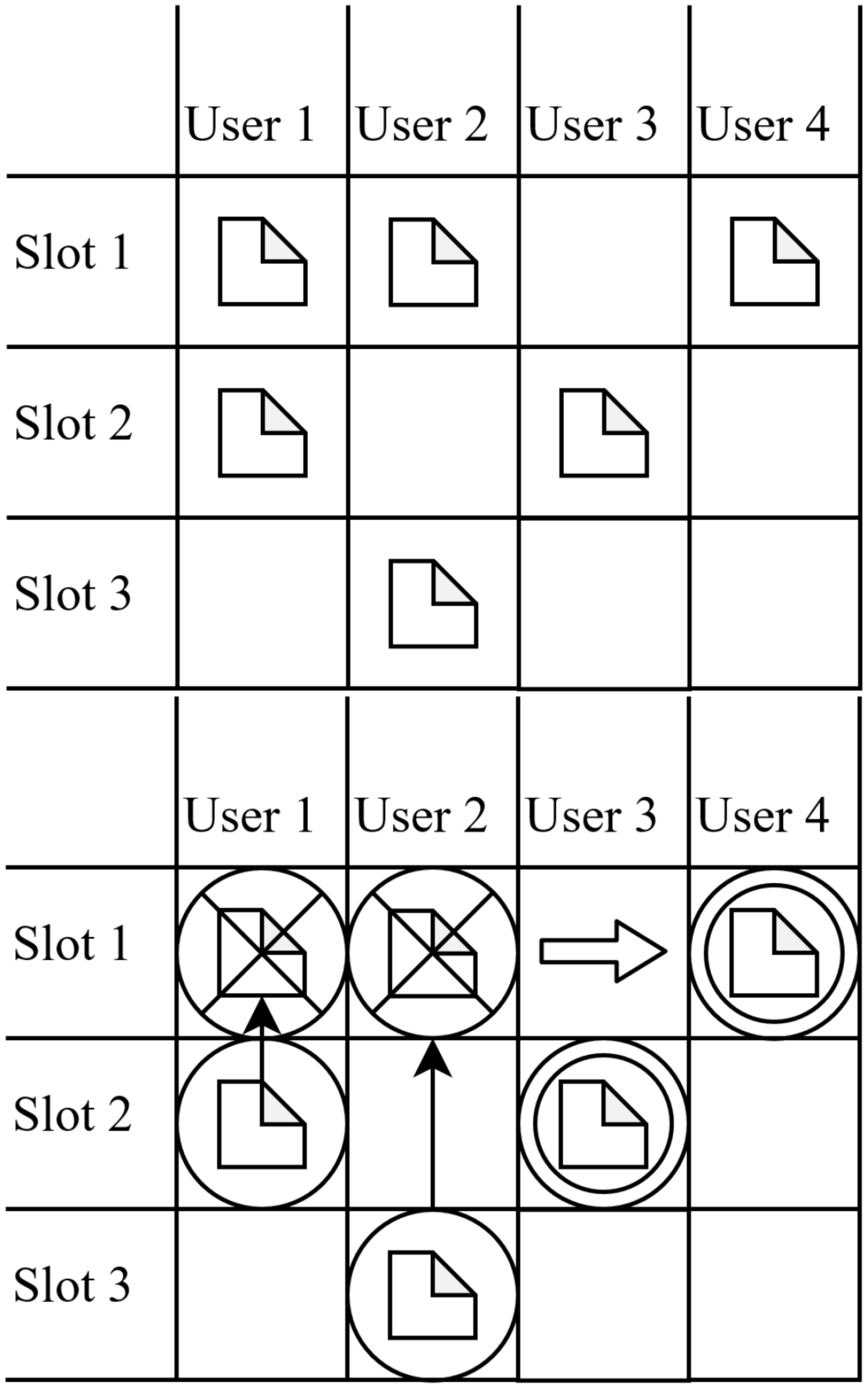
- User activity detection
- Channel estimation
- Non-orthogonal pilot sequences, leading to pilot contamination
  - Special cases: Orthogonal pilot reuse, quasi-orthogonal pilots, random pilots
- Multi-user interference
- Practical aspects: Path loss, fading, MIMO, short packets, time & frequency synchronization



Irregular Repetition Slotted Aloha (IRSA)

- In IRSA, each user transmits several replicas of the packet in randomly chosen slots
  - Each replica has pilot, data & error check
  - Repetition distribution governs slot indices for the transmission of packet replicas
- Decoding via successive interference cancellation (SIC)
  - Continue SIC iteratively until no more users can be decoded

K. R. Narayanan and H. D. Pfister, “Iterative collision resolution for slotted ALOHA: An optimal uncoordinated transmission policy,” in Proc. ISTC, Aug 2012, pp. 136–139.



System Model

- M single-antenna users, N antenna BS
- T resource blocks, τ-length pilots
- Binary access pattern matrix $\mathbf{G}_{T\times M}$ governs replica transmission
- Received pilot signal:

Access coefficient

$$\mathbf{Y}_t^{\mathrm{p}} = \sum_{m=1}^{M} a_m g_{tm} \mathbf{h}_{tm} \mathbf{p}_m^H + \mathbf{N}_t^{\mathrm{p}}$$

Pilot

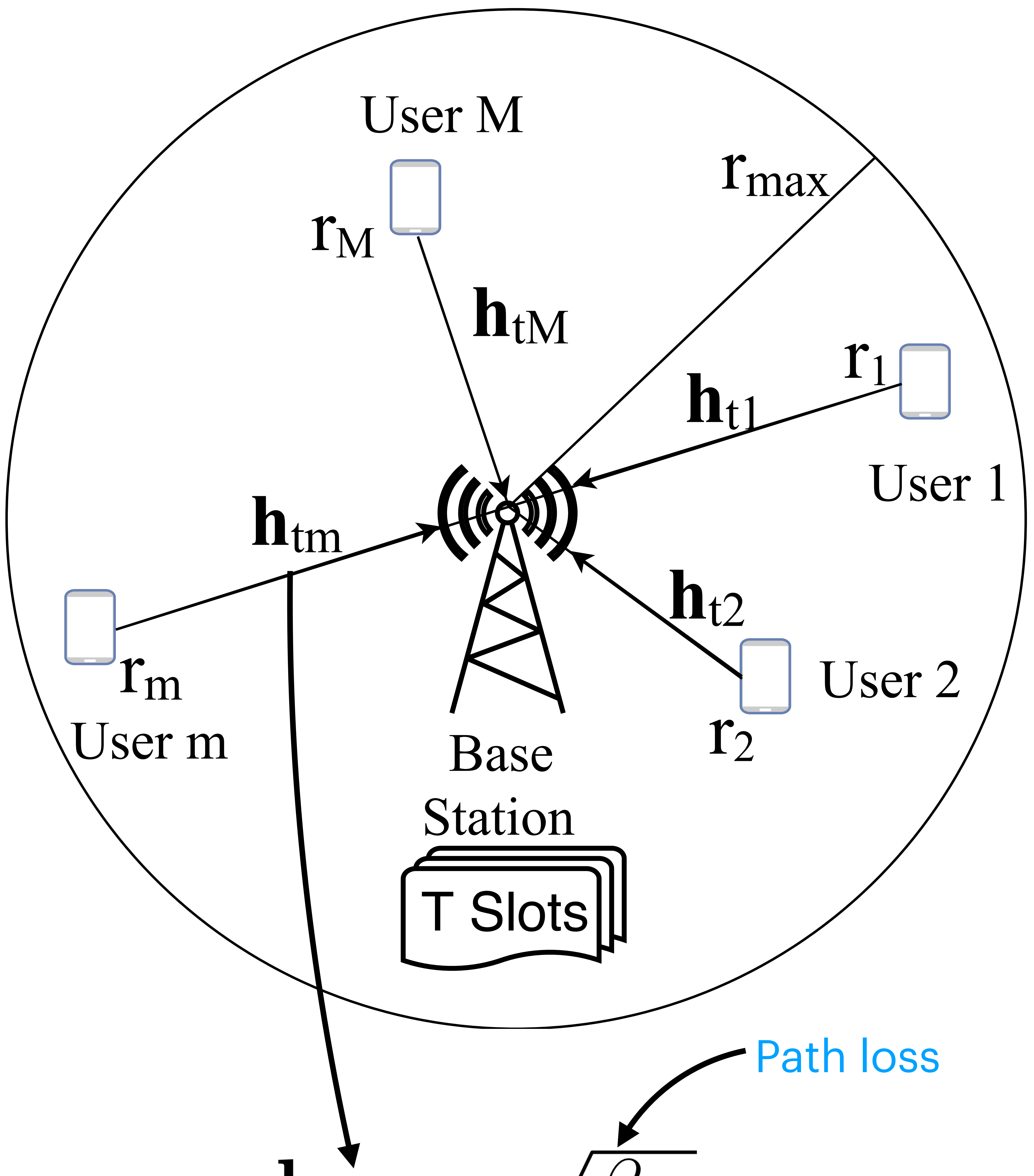
Noise

- Received data signal:

$$\mathbf{y}_t = \sum_{m=1}^{M} a_m g_{tm} \mathbf{h}_{tm} x_m + \mathbf{n}_t$$

Activity coefficient

Data



Path loss

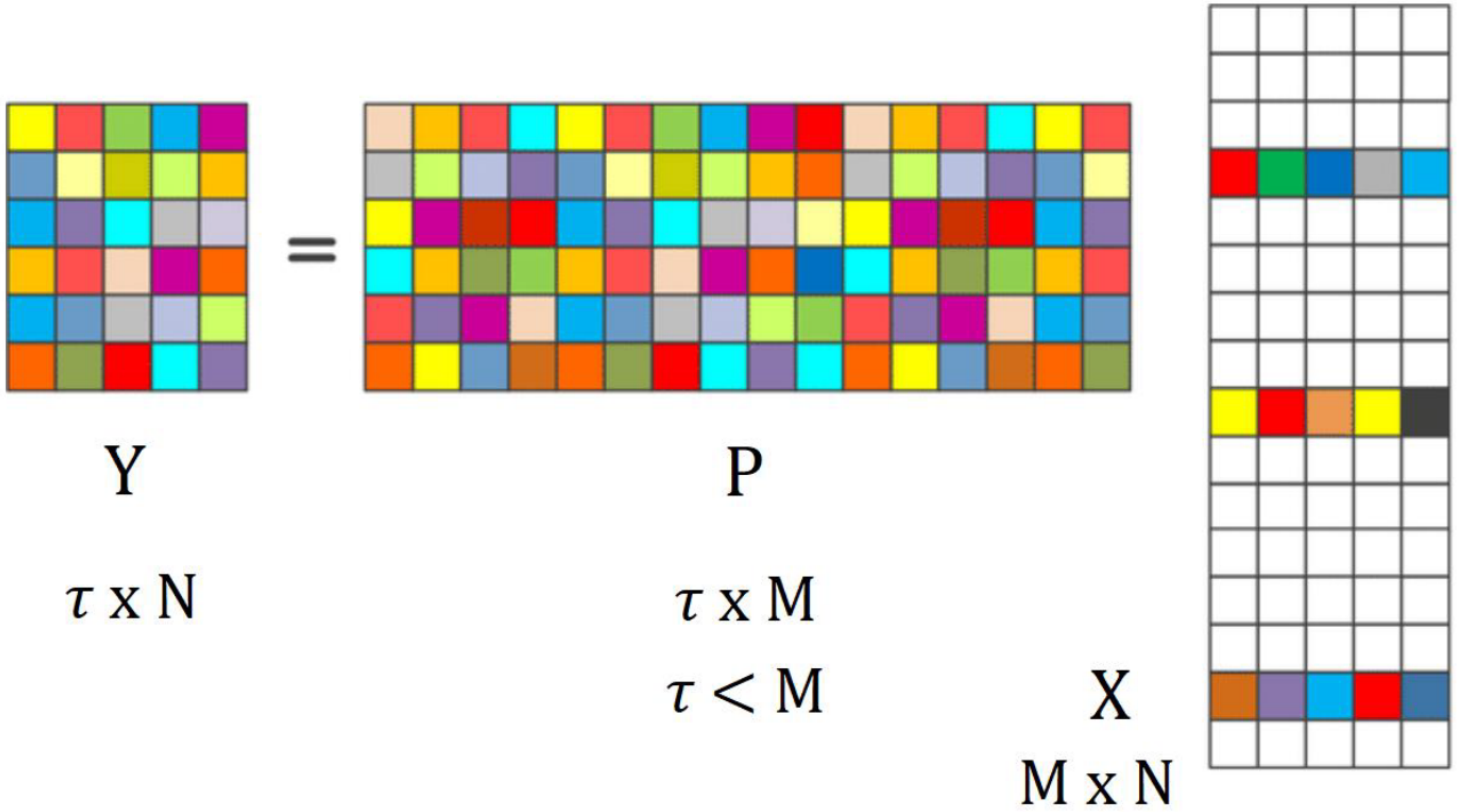
$$\mathbf{h}_{tm} = \sqrt{\beta_m}\mathbf{v}_{tm}$$

Channel

Fading

User Activity Detection (UAD) in IRSA

- Less than 1% of mRA devices are active at any instant
- Support recovery problem in compressed sensing (CS)
- Underdetermined system of equations with a sparse vector/matrix to be estimated
- Multiple measurement vector (MMV): Columns of X share a common support



- τ - Pilot length
- M - Number of users
- N - Number of antennas

How to choose pilot length?

- Setup: MMV compressed sensing problem where the columns of $\mathbf{X}$ are jointly k-sparse
- Model: $\mathbf{Y}_{\tau \times N} = \mathbf{P}_{\tau \times M} \mathbf{X}_{M \times N} + \mathbf{Z}_{\tau \times N}$
- Sufficient condition for support recovery of $\mathbf{X}_{M \times N}$:
  - Choosing $\tau = \Omega(k \log(M/k))$ yields a vanishing support recovery error rate as $M \to \infty$, when $N \gg \log M / \log \log M$
- Practically, this translates to choosing $\tau = ck \log(M/k)$

Reference: G. Tang and A. Nehorai, "**Performance analysis for sparse support recovery**," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1383–1399, Mar. 2010.

Pilot length scaling laws

- Suppose the sensing matrix **P** has i.i.d. sub-Gaussian entries
- M-SBL recovers the true support with vanishing prob. error, provided

$$\tau = \Theta(k \log M) \text{ and } N = \Omega\left(\frac{M}{k} \log M + M \log k + M \log\log M\right)$$

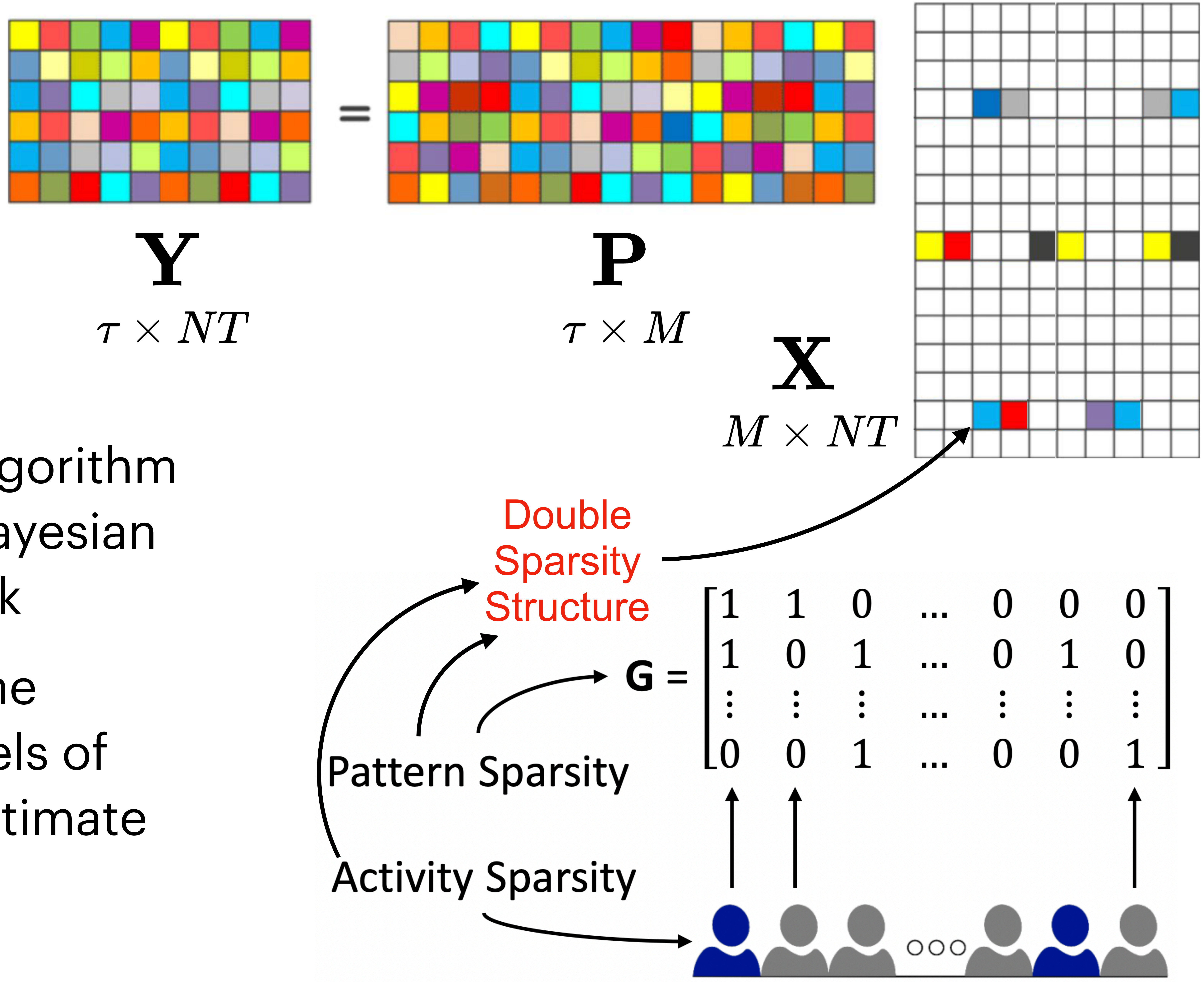
- Or

$$\tau = \Theta(\sqrt{k} \log M) \text{ and } N = \Omega\left(\frac{M}{\sqrt{k}} \log M + M\sqrt{k} \log k + M\sqrt{k} \log\log M\right)$$

- Here, k = num. of active users, M = total num. users, N = num. antennas

S. Khanna and M., T-IT Nov. 2022

- Due to the structure of IRSA, for an active user, the row entries are nonzero in chunks
- Row-chunk sparsity
- Our solution: Bayesian algorithm inspired by the Sparse Bayesian Learning (SBL) framework
- SBL: Impose a prior on the covariance of the channels of the users and use it to estimate the activity coefficients



$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 0 & 1 & \ldots & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & \ldots & 0 & 0 & 1 \end{bmatrix}$$

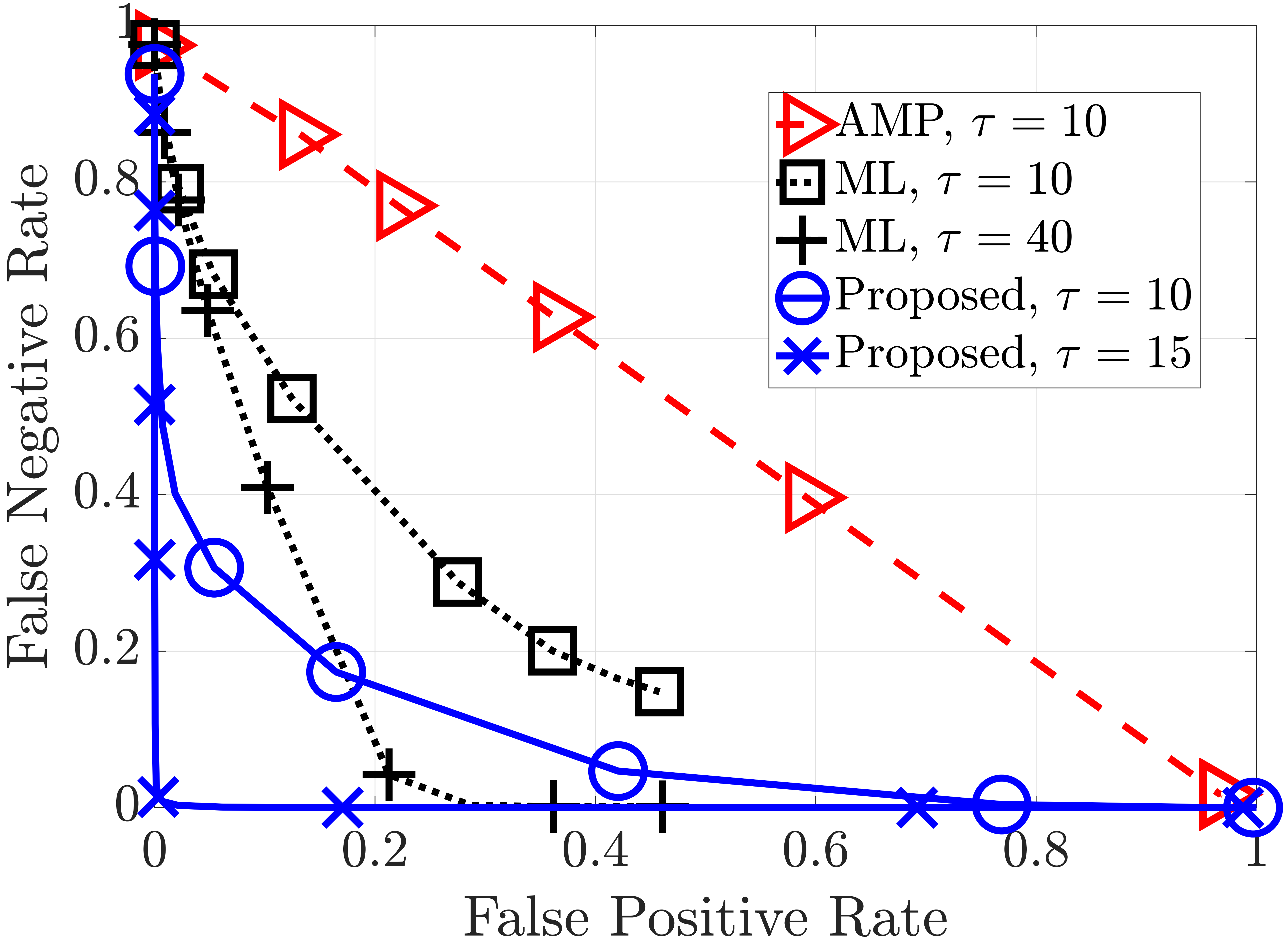
UAD Algorithm

- SBL: Finds the MAP estimate of the user's activity coefficients by using a fictitious, sparsity-promoting hierarchical prior
- Idea: Estimate the channel covariance - goes to zero for inactive users!
- Objective: $\log(p(\mathbf{Y}_t^{\mathrm{p}};\boldsymbol{\gamma}_t)) \propto -N\log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_t}| - \mathrm{Tr}(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_t}^{-1}\mathbf{Y}_t^{\mathrm{p}}\mathbf{Y}_t^{\mathrm{p}H})$
- Solution: Expectation maximization to iteratively find MAP estimate
- The first algorithm for UAD specifically for IRSA!
  - UAD algorithm can be applied to all variants of IRSA
  - (Much) faster versions of M-SBL exist, and our modification to M-SBL can be applied to the faster versions also

Publication: C. R. Srivatsa and C. R. Murthy, "**User Activity Detection for Irregular Repetition Slotted Aloha Based MMTC**," in *IEEE Transactions on Signal Processing*, vol. 70, pp. 3616-3631, 2022, doi: 10.1109/TSP.2022.3185891.

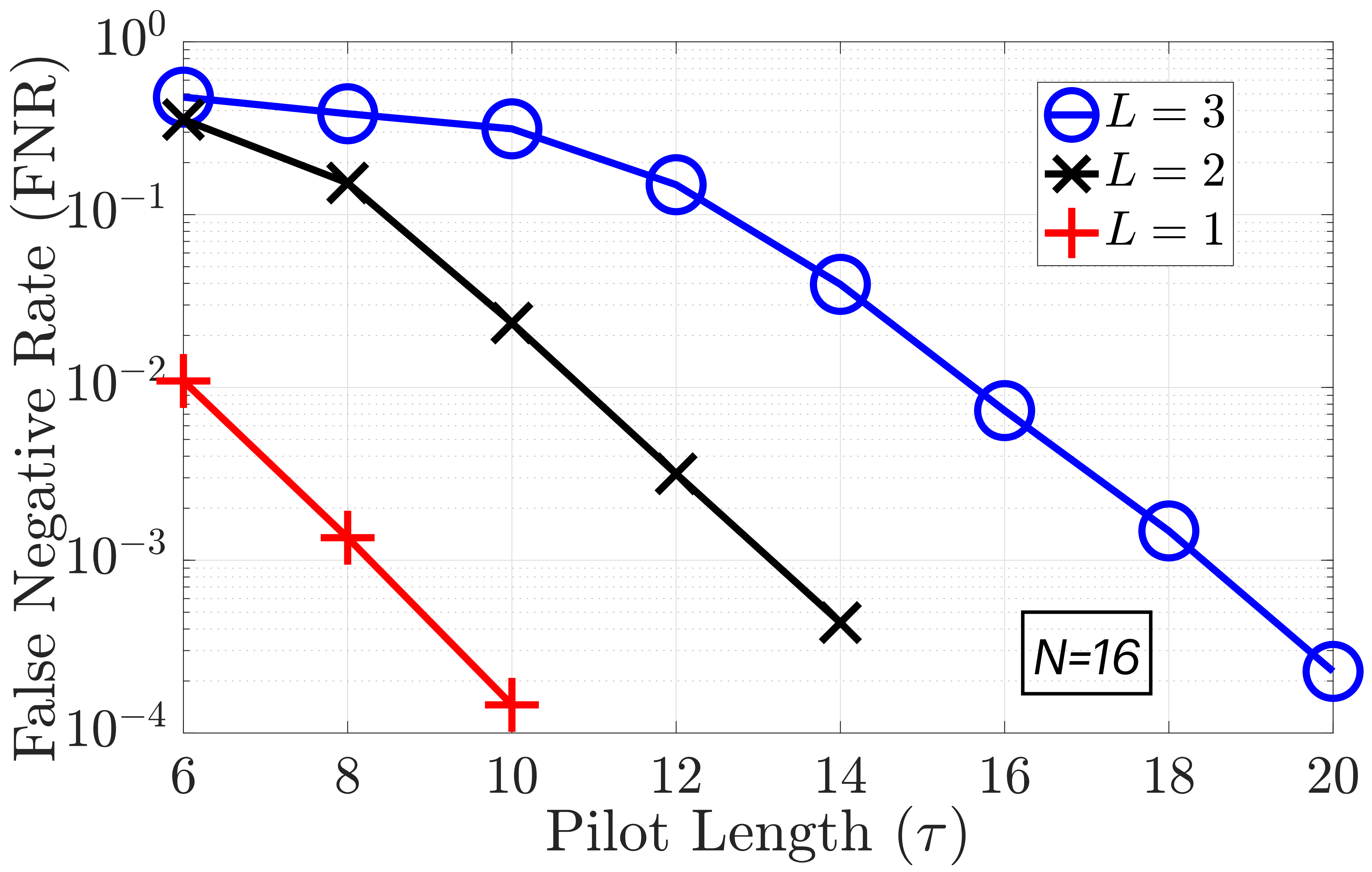
Numerical Results

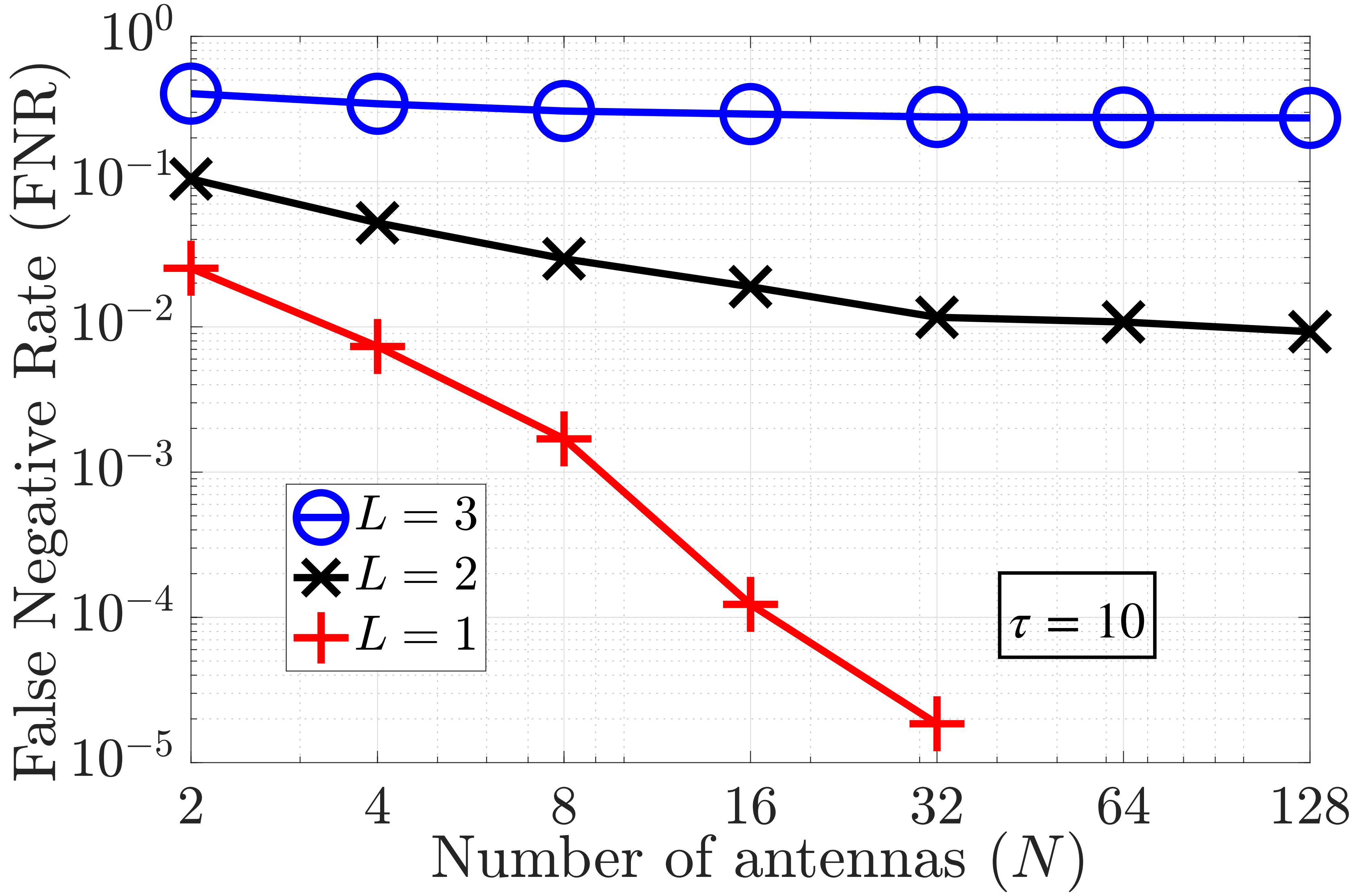
- Metrics: False negative rate and false positive rate
- Setup: T = 50 slots, M = 1500 users, N = 4 antennas, 1% active
- For other algorithms, we perform $\hat{a}_i = 1\{\sum_{t=1}^{T} \hat{a}_i^{(t)} \geq 1\}$
- At FNR=0.2: 4-fold reduction in τ compared to classical detection techniques (ML, AMP) since we exploit structure



How to improve UAD performance?

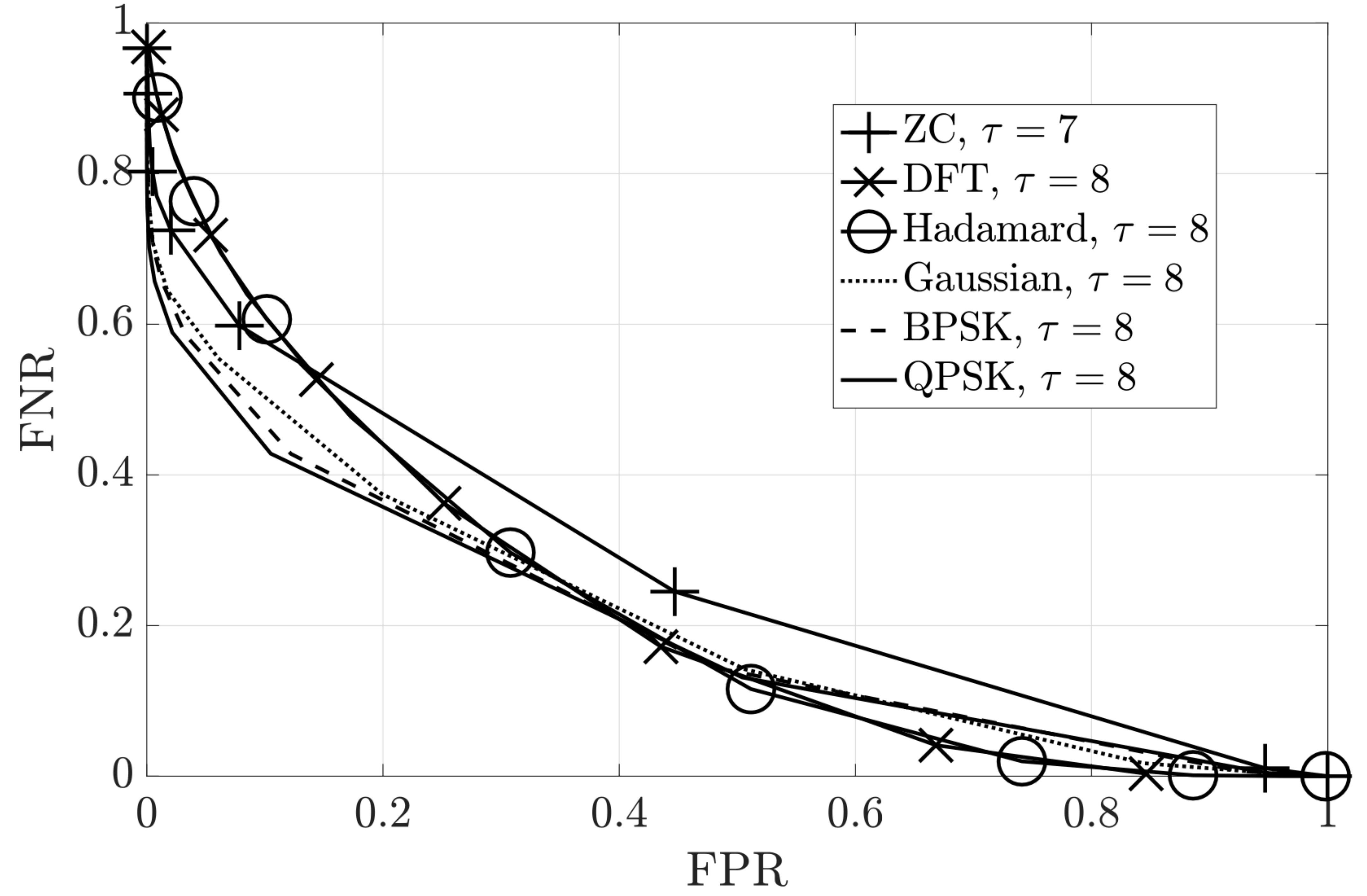
- Pilot length of $\tau = 20$ is sufficient for very low error rates for 1500 users!
- Conventional MMV CS algorithms would need $\tau = k\ln(M/k) = 350$ to achieve a similar performance
- Insight: Very low τ sufficient: Huge reduction in overhead for mRA
- Takeaway: Increasing pilot length is more beneficial than increasing the number of antennas

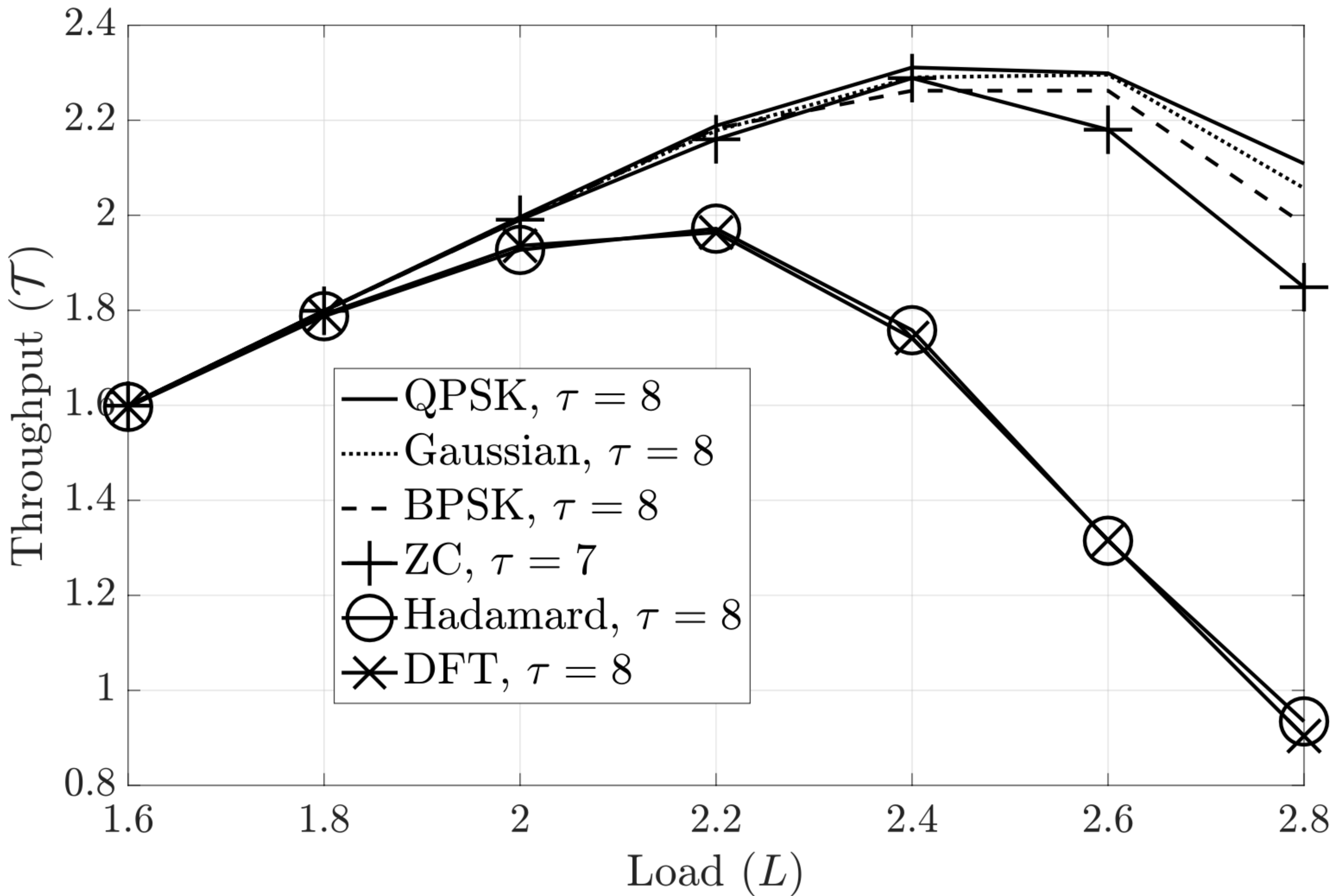




Choice of Pilot Sequences

- Orthogonal pilot reuse (OPR) performed with Hadamard and DFT pilots
- All pilot sequences yield similar UAD performance
- Non-orthogonal pilots (MPSK/Gaussian) yield higher throughput than OPR due to diversity
    - Pr(Two users choosing identical pilot sequences within a slot in IRSA) is lower for non-orthogonal pilots





Pilot length scaling with load for UAD

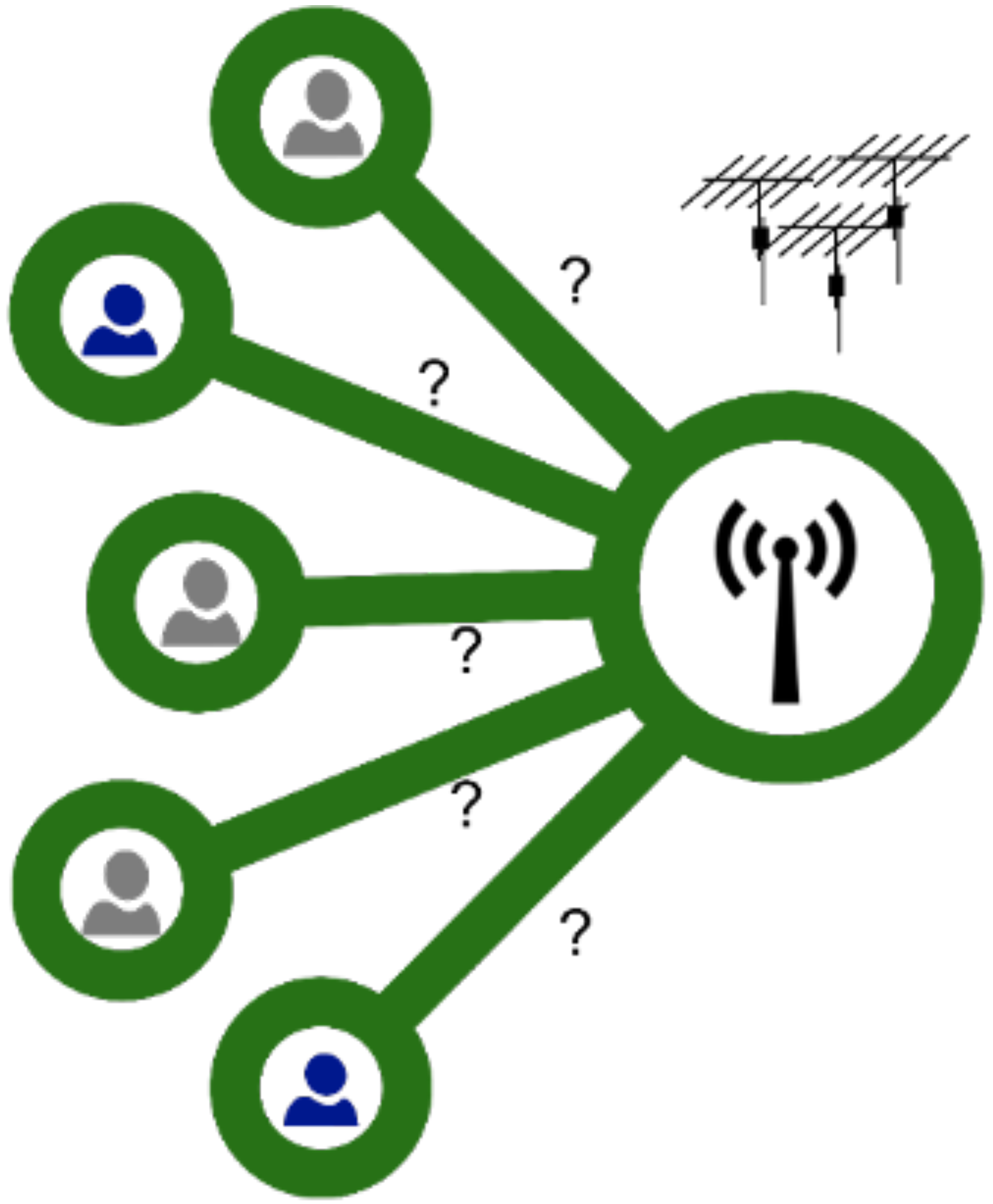
- For the UAD problem in IRSA, we need to choose the pilot length as $\tau = cM_a \log(M/M_a)$, where M is the total number of users, M_a is the number of active users, and T is the number of slots
  - We use the average number of active users $\mathbb{E}[M_a] = Mp_a$ instead of M_a in the above, where p_a is the per-user activity probability
    - Under typical mRA settings, p_a and T are fixed, and $M_a \approx \mathbb{E}[M_a] = Mp_a$
- Effectively, the pilot length needs to scale logarithmically with the load

Pilot length scaling with load for UAD

- In IRSA, in each slot, only $L\bar{d}$ users collide on average; where L is the IRSA system load, and $\bar{d}$ is the average number of replicas transmitted by users
- Effectively, much fewer users collide in each slot, i.e., $L\bar{d} \ll M$
  - This implies that low pilot lengths are sufficient for accurate UAD
  - $\tau = cM_a \log(M/M_a) = cL\bar{d}(-p_a \log(p_a))$
- Under typical mRA settings, with $M = 10^5, p_a = 0.01,\ T = 100,\ \bar{d} = 3$, pilot length of the order 30 is sufficient
  - With $\tau \geq 30$, we observe vanishing error rates in this regime

Interesting directions

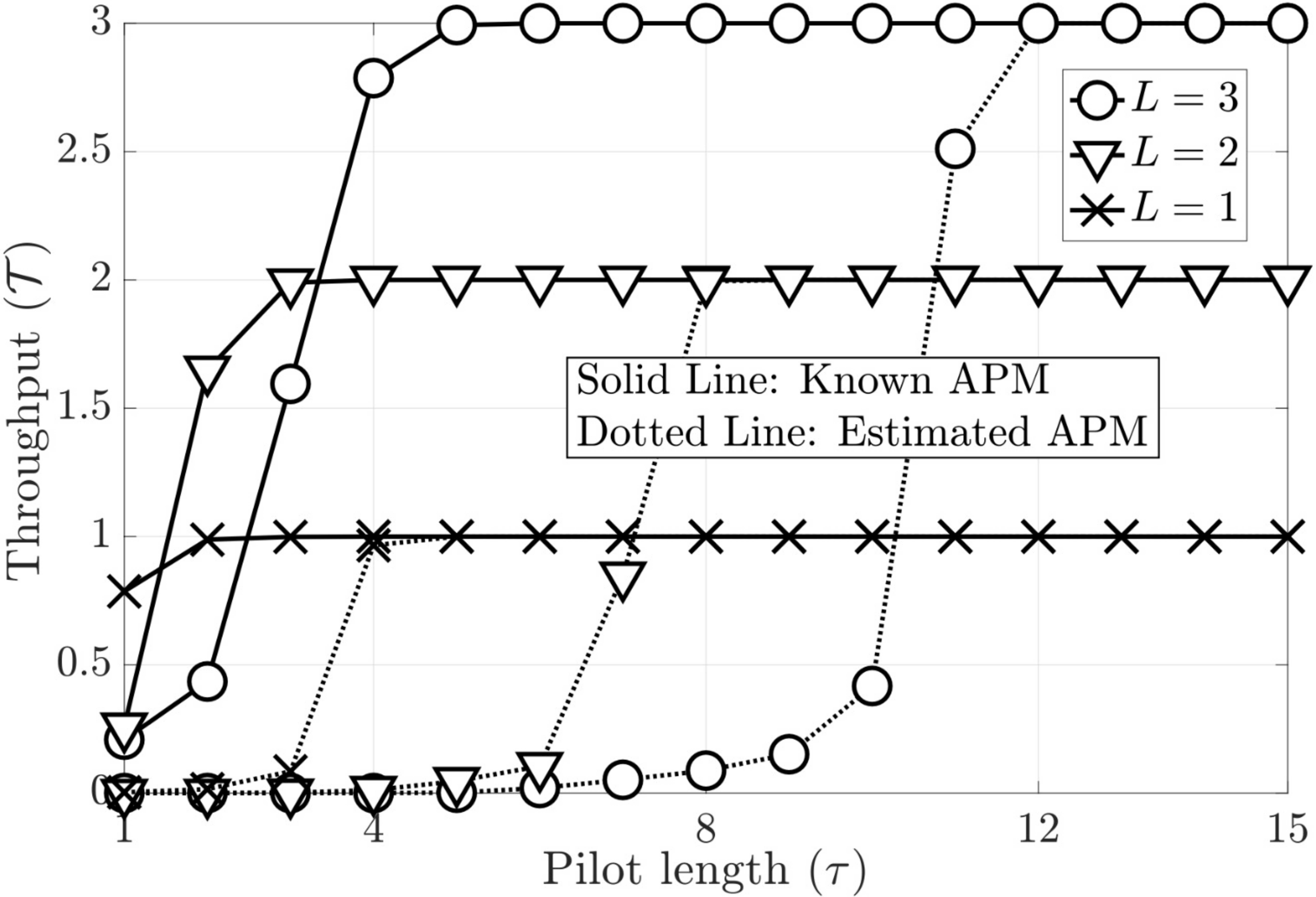
- Short packet communications (e.g., using Polyanskiy's results)
- Time and frequency synchronization errors
- Energy efficiency, latency, age-of-information
- Power control and performance improvement



Shameless ad: Lekshmi Ramesh's talk at 14.00 today

Thank you!

Email: cmurthy@iisc.ac.in


3
2.5
2
1.5
1
0.5
0
Throughput ($\mathcal{T}$)
1
4
8
12
15
Pilot length (τ)
$L = 3$
$L = 2$
$L = 1$
Solid Line: Known APM
Dotted Line: Estimated APM