Research Methods in computer science

Spring 2025

Lecture 22

Omprakash Gnawali
April 9, 2025

Agenda

Conference updates

Paper feedback

Idea generation

Artifact comparison experiments

HW10

Paper review template

Summary

Strengths

Weaknesses

Detailed Comments

Paper feedback

Baseline for comparison

What questions need to be answered?

Clarify Claims/contributions

Generating Research Ideas

“Standing on the shoulders of giants”

Most ideas may not be new

New may be subjective

- Adding a layer to an existing deep learning architecture
  - When is it new?
  - When is it not new?

Idea Generator Heuristics

- Combination / Hybrid techniques
  - From the same discipline
    - (e.g., ….)
  - From a different discipline
    - (e.g., ….)
- Address Gap/limitation (Incremental?)
  - Handle some cases that were not handled
  - Improve some (partial) aspects of dimension
- Apply different datasets / settings / contexts

In-class group activity

Pick a paper

Generate at least two derivative ideas

Present: original and derivative ideas

CS Experiments Today

Artifact Comparison Experiments

- Run the new artifact
- Run best-known prior work
- Compare

Simulations + “Real” experiments

Wireless Experiments Today

Protocol Comparison Experiments

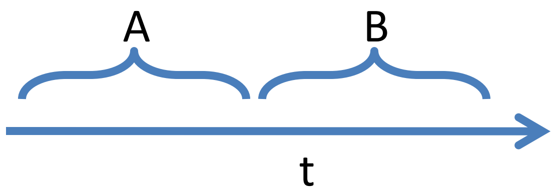
- Run the new protocol
- Run best-known prior work
- Compare

Simulations + Testbed experiments

Serial Experiments

Run one protocol at a time

Compare the results



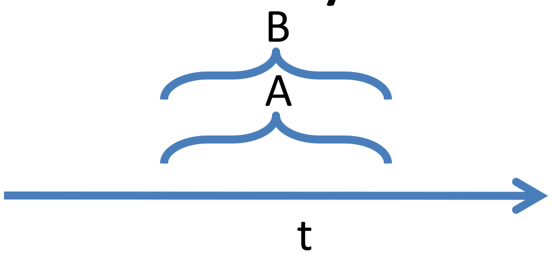
Difficult to distinguish the contribution of these these variables

- Environment
- Protocol mechanisms

Concurrent Experiments

Run multiple protocols concurrently

Compare the results



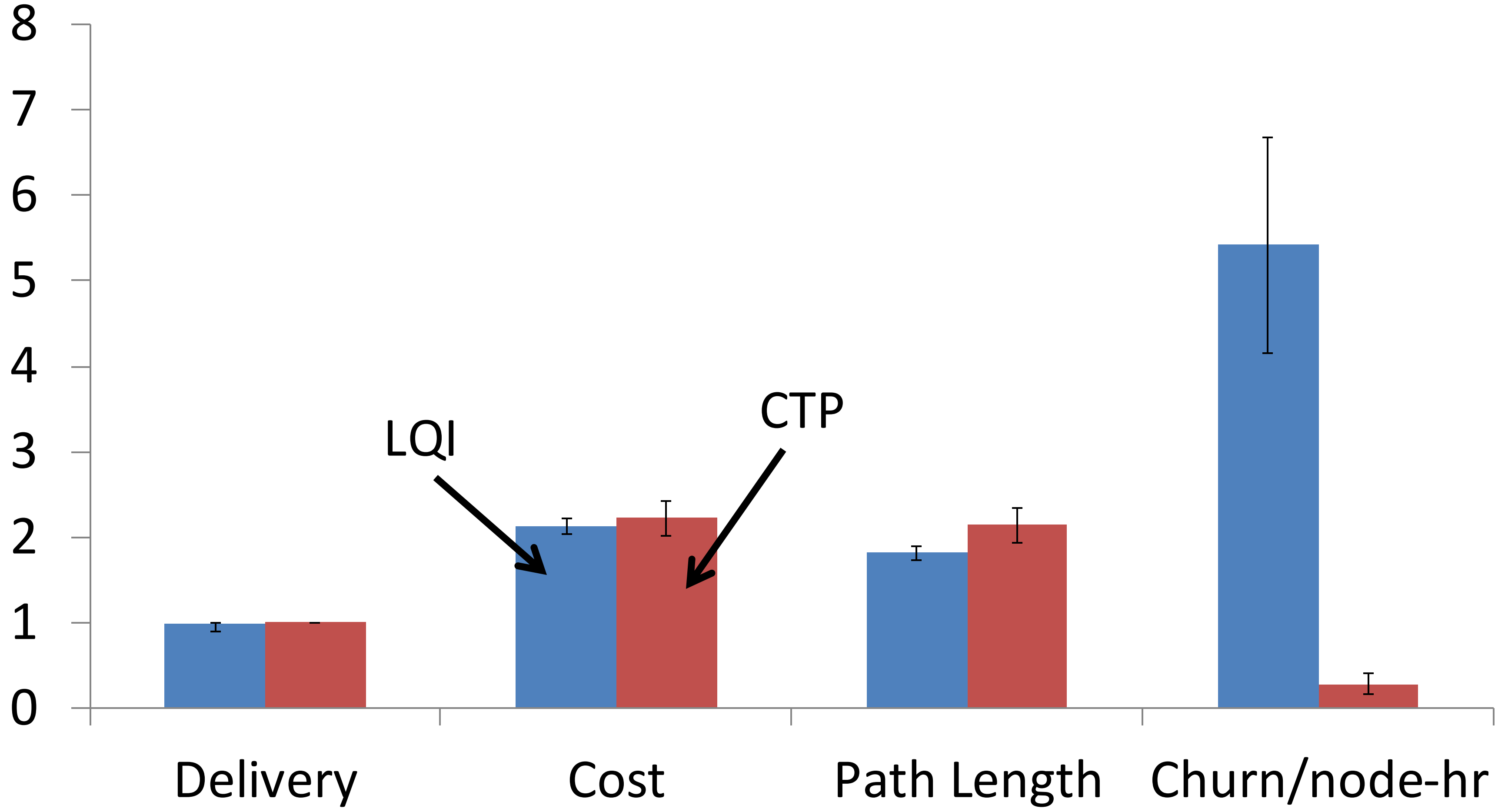
Advantages

Consistent environment for both the protocols

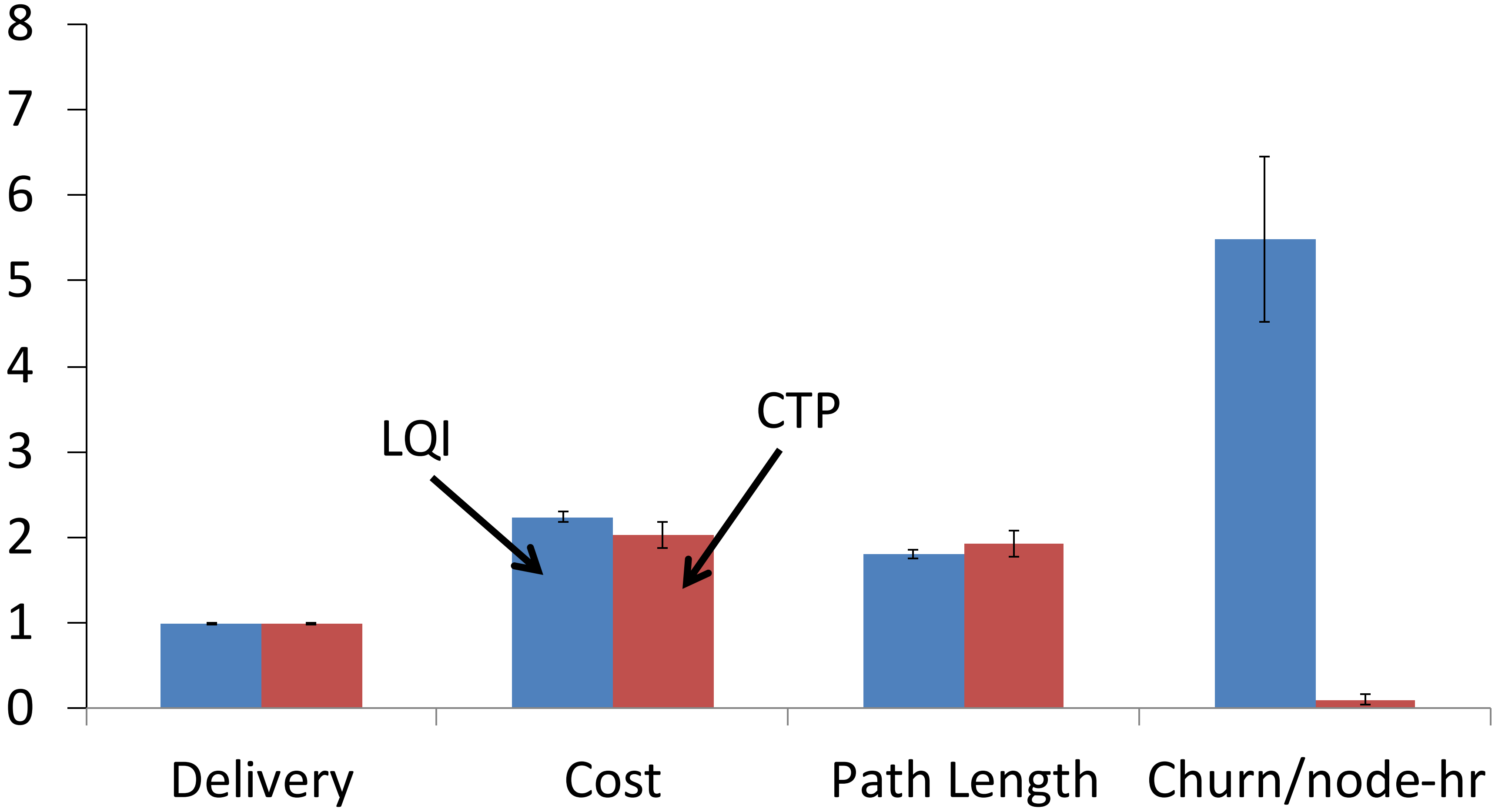
Concerns

Contention of different types

Results from Serial CTP vs LQI Experiment on Tutornet

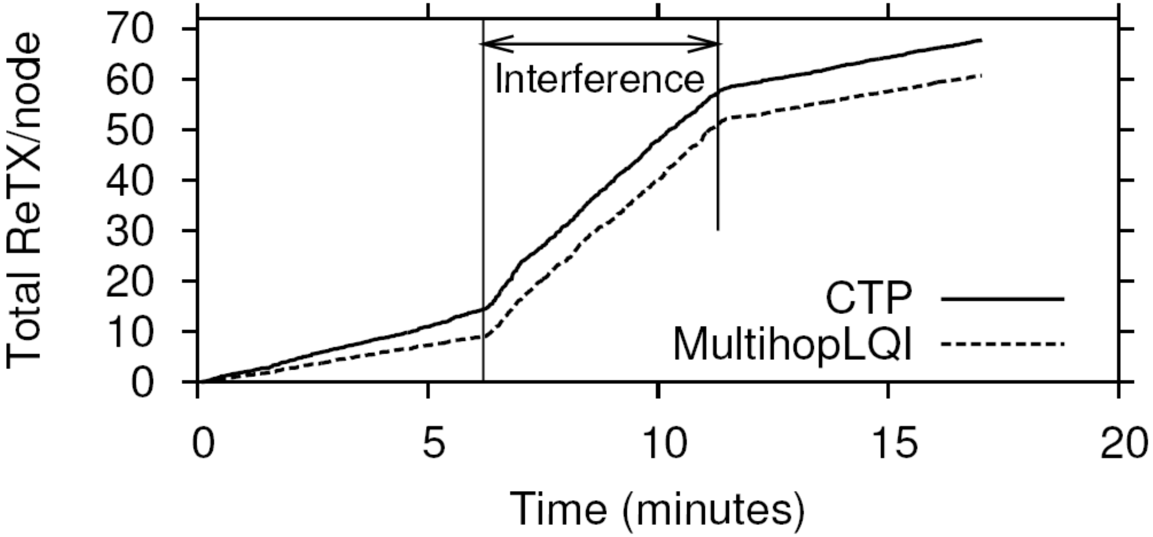


Results from Concurrent CTP vs LQI Experiment on Tutornet



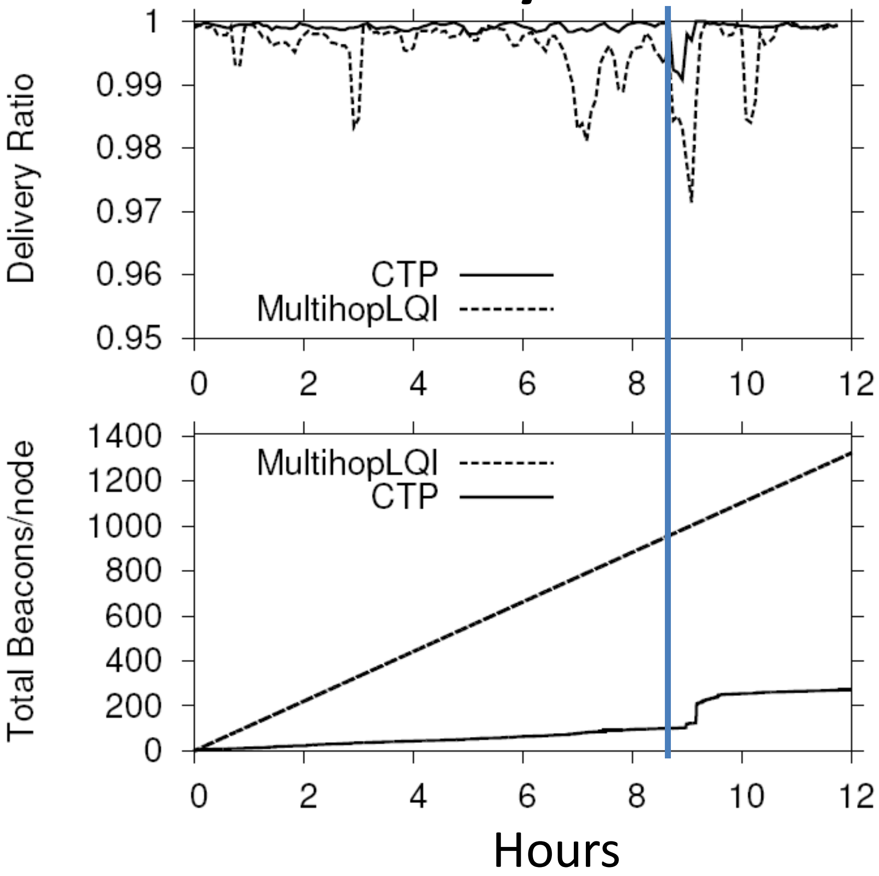
Putting Concurrent Methodology to Use: Expts. with External Interference

Engineered Scenario



Both protocols *struggle* in the same environment.

Putting Concurrent Methodology to Use: Experiments in a Dynamic Network



CTP and LQI react differently to dynamics.

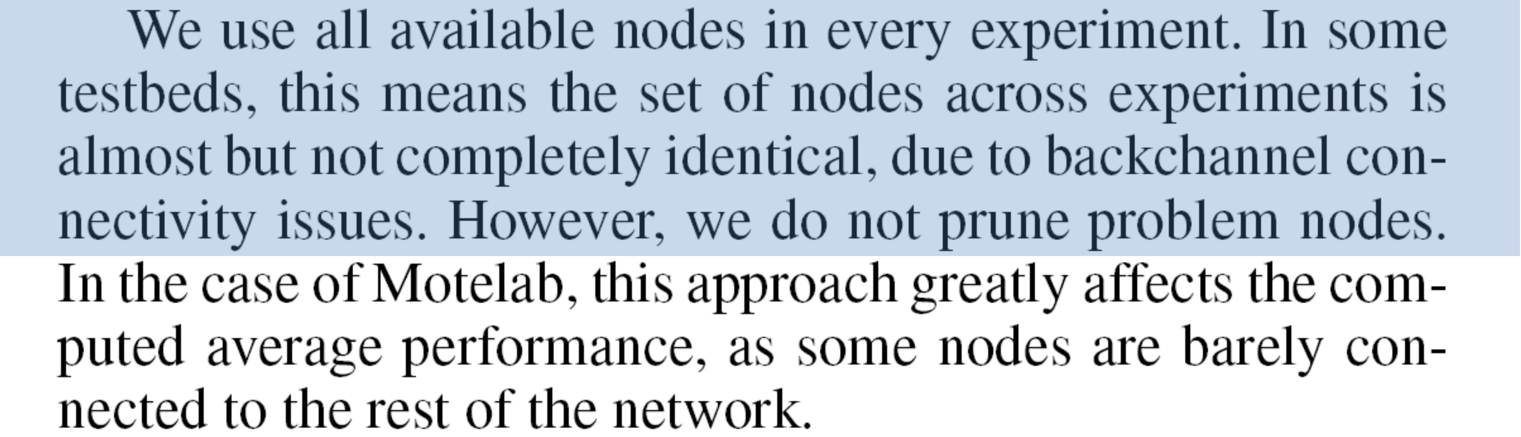
Uncontrolled environment does not imply we cannot do fair comparisons

Level of Details

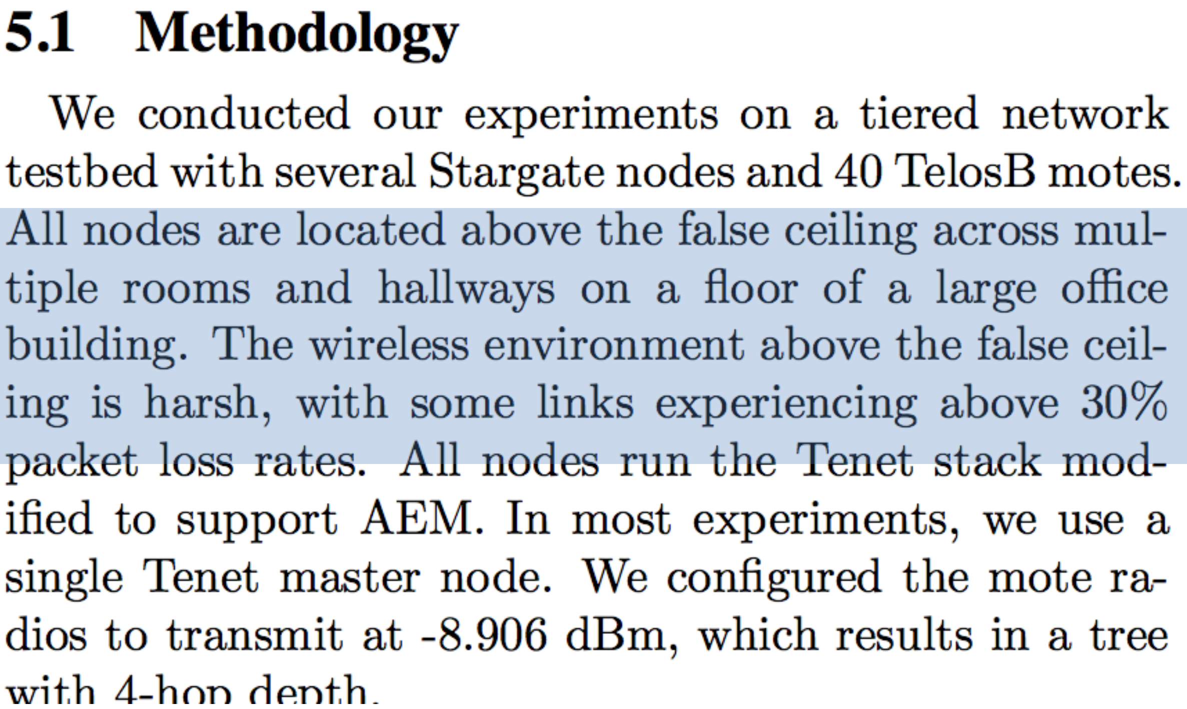
At What Level of Detail?

Descriptions

- System and algorithm
- Experiments
- Datasets
- Results

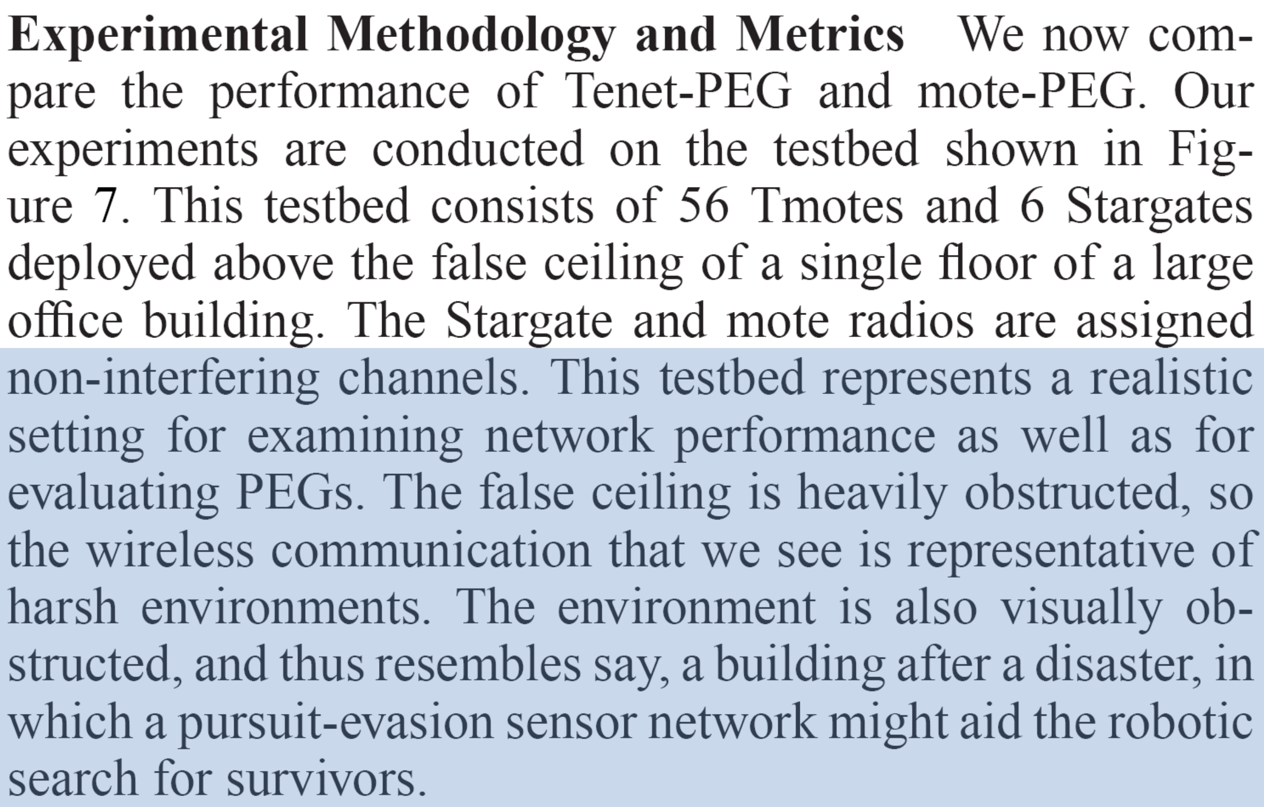


We use all available nodes in every experiment. In some testbeds, this means the set of nodes across experiments is almost but not completely identical, due to backchannel connectivity issues. However, we do not prune problem nodes. In the case of Motelab, this approach greatly affects the computed average performance, as some nodes are barely connected to the rest of the network.



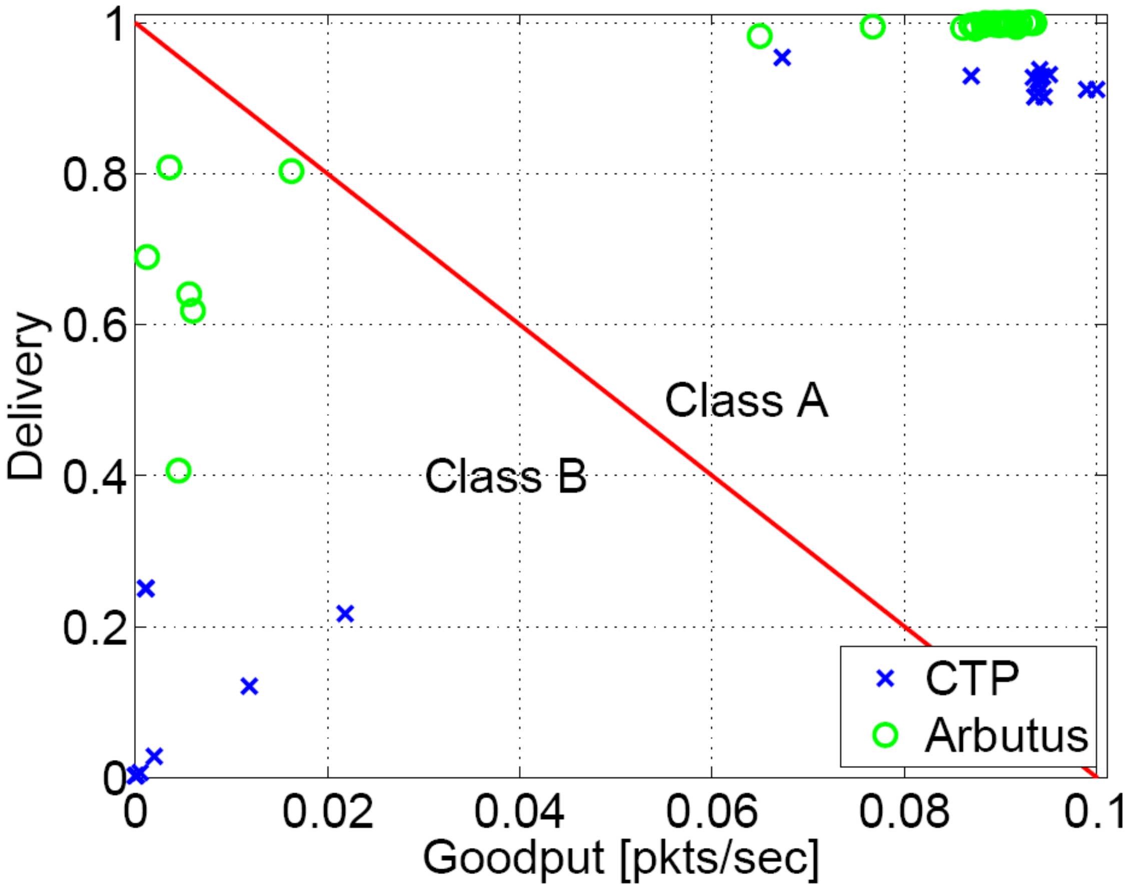
5.1 Methodology

We conducted our experiments on a tiered network testbed with several Stargate nodes and 40 TelosB motes. All nodes are located above the false ceiling across multiple rooms and hallways on a floor of a large office building. The wireless environment above the false ceiling is harsh, with some links experiencing above 30% packet loss rates. All nodes run the Tenet stack modified to support AEM. In most experiments, we use a single Tenet master node. We configured the mote radios to transmit at -8.906 dBm, which results in a tree with 4-hop depth.



Experimental Methodology and Metrics We now compare the performance of Tenet-PEG and mote-PEG. Our experiments are conducted on the testbed shown in Figure 7. This testbed consists of 56 Tmotes and 6 Stargates deployed above the false ceiling of a single floor of a large office building. The Stargate and mote radios are assigned non-interfering channels. This testbed represents a realistic setting for examining network performance as well as for evaluating PEGs. The false ceiling is heavily obstructed, so the wireless communication that we see is representative of harsh environments. The environment is also visually obstructed, and thus resembles say, a building after a disaster, in which a pursuit-evasion sensor network might aid the robotic search for survivors.

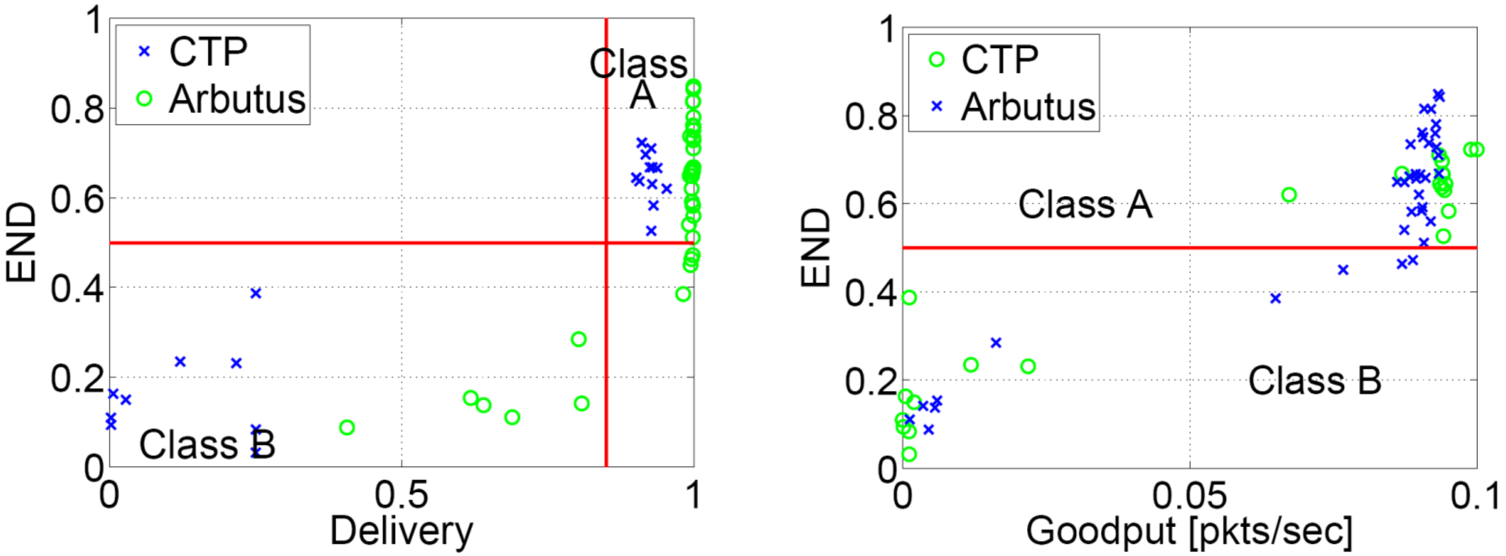
Results from the same Testbed



Network Metric

Converting these subjective descriptions to a more quantitative description

END and CTP Performance



“We evaluate the throughput and delay benefits of CQIC using the Google Nexus device to download content from a Google server via a popular cellular network provider. Reflecting a common CDN scenario, this server is located near the network of the mobile carrier such that the cellular channel is the bottleneck link…”

[Lu 2015]

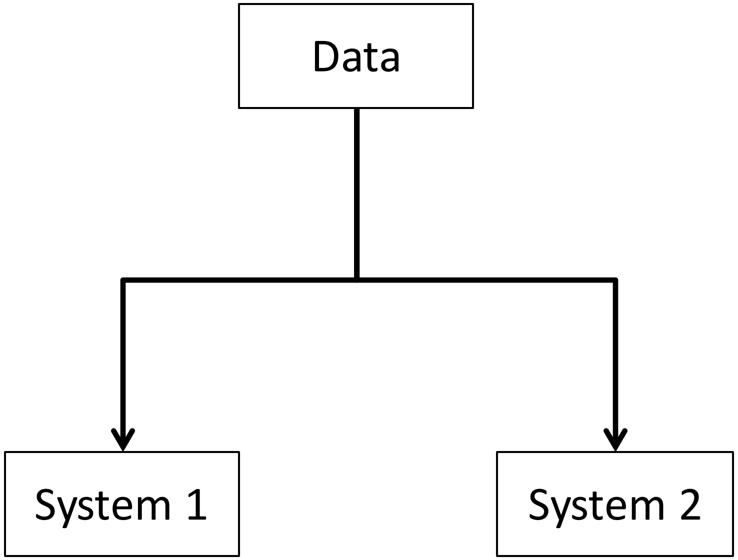
AI/ML/NLP

Many times standardized datasets or tasks

Compare systems in the same dataset

Tradition of shared notebooks/repo online

Faithful implementation of prior work often less challenging in systems areas but not entirely if related to operations/systems aspect of AI

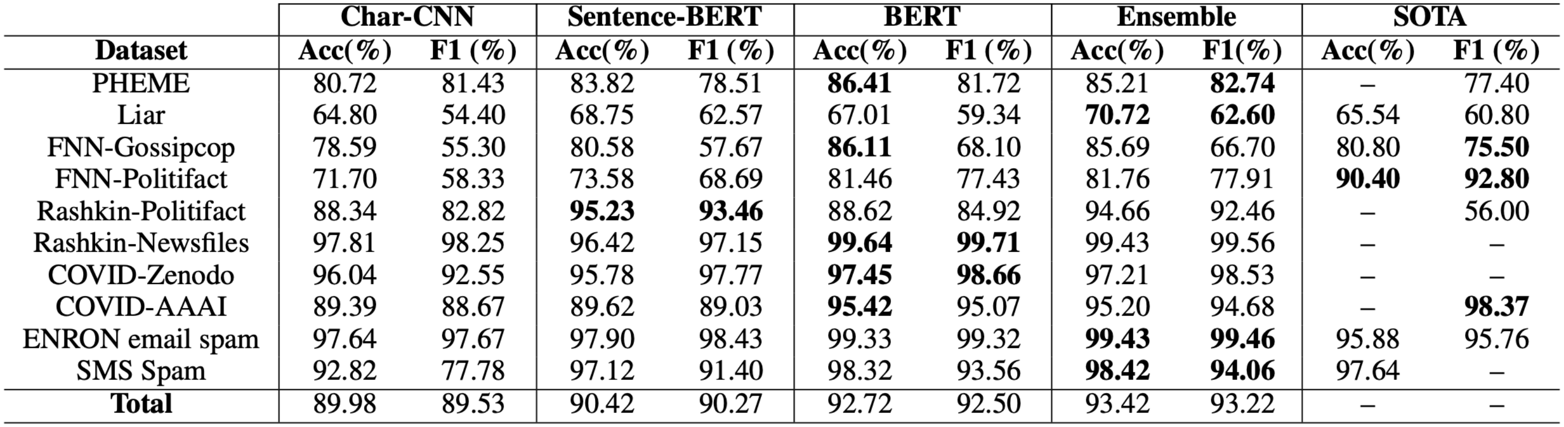


Data

System 1

System 2

Typical Expt. in NLP-related areas



	Char-CNN		Sentence-BERT		BERT		Ensemble		SOTA	
Dataset	**Acc(%)**	**F1 (%)**	**Acc(%)**	**F1 (%)**	**Acc(%)**	**F1 (%)**	**Acc(%)**	**F1(%)**	**Acc(%)**	**F1 (%)**
PHEME	80.72	81.43	83.82	78.51	**86.41**	81.72	85.21	**82.74**	–	77.40
Liar	64.80	54.40	68.75	62.57	67.01	59.34	**70.72**	**62.60**	65.54	60.80
FNN-Gossipcop	78.59	55.30	80.58	57.67	**86.11**	68.10	85.69	66.70	80.80	**75.50**
FNN-Politifact	71.70	58.33	73.58	68.69	81.46	77.43	81.76	77.91	**90.40**	**92.80**
Rashkin-Politifact	88.34	82.82	**95.23**	**93.46**	88.62	84.92	94.66	92.46	–	56.00
Rashkin-Newsfiles	97.81	98.25	96.42	97.15	**99.64**	**99.71**	99.43	99.56	–	–
COVID-Zenodo	96.04	92.55	95.78	97.77	**97.45**	**98.66**	97.21	98.53	–	–
COVID-AAAI	89.39	88.67	89.62	89.03	**95.42**	95.07	95.20	94.68	–	**98.37**
ENRON email spam	97.64	97.67	97.90	98.43	99.33	99.32	**99.43**	**99.46**	95.88	95.76
SMS Spam	92.82	77.78	97.12	91.40	98.32	93.56	**98.42**	**94.06**	97.64	–
Total	89.98	89.53	90.42	90.27	92.72	92.50	93.42	93.22	–	–

Datasets not always standardized

Describe the data in enough detail even if the dataset cannot be released to the public

DeepFace: Closing the Gap to Human-Level Performance in Face Verification

[Taigman 2014]

The SFC dataset includes 4.4 million labeled faces from 4,030 people each with 800 to 1200 faces, where the most recent 5% of face images of each identity are left out for testing. This is done according to the images' time-stamp in order to simulate continuous identification through aging. The large number of images per person provides a unique opportunity for learning the invariance needed for the core problem of face recognition…

"See the supplementary material for more details about SFC."

Supplementary Material:
DeepFace: Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman Ming Yang Marc'Aurelio Ranzato Lior Wolf

Facebook AI Research
Menlo Park, CA, USA
{yaniv, mingyang, ranzato}@fb.com

Tel Aviv University
Tel Aviv, Israel
wolf@cs.tau.ac.il

HW10

Full paper submission

Due: April 13