Lecture #04

January 13, 2010

Today's Objectives

- Evaluation techniques

Notes

- Evaluation is one of the three critical aspects of a paper… it may even be *the* most critical aspect
  - Other two: significant/interesting problem and novel solution
- Goal of evaluation is to demonstrate some objective of your choosing
  - Good performance
    - Or “better” performance if comparing to another solution
    - Or simply that something works
  - Solution to a newly discovered problem
  - Understand behavior of an existing system

How Much Evaluation?

- Depends on how many other existing solutions there are

- *Worst case*: developing a new version of TCP congestion control
  - Tons of existing solutions
  - Hard to evaluate because TCP CC algorithm has to be robust and work in *many* different scenarios

- One way to make evaluation easier is to change the problem enough so that other solutions don't apply
  - Scenario: look at performance of wireless video susceptibility to loss but in a 802.11a environment as opposed to 802.11b environment (argue that loss conditions are different)
  - Not always possible, reasonable, or justifiable

How Much Evaluation?

- Doing a comparative evaluation is typically hard…
  - Have to implement one or more other systems
  - Time consuming
  - May not be details on how the system really works
  - May not even be accurate representation of other system

- …but necessary/worthwhile
  - If there are other proposed systems how can you claim to be “better” without some sort of comparative evaluation?
  - A good comparative evaluation is really effective

Evaluation Can Be Hard to Plan

- Issue #1: what questions *can* be answered?
  - Limits of evaluation techniques
  - Watch for authors that over-reach in reaching conclusions based on what they've done or what their evaluation shows
    - Ex: a proposed encoding technique that is robust to packet loss, but the evaluation is limited to certain kinds of loss or only small amounts of loss
    - Ex: a routing protocol that works (or is better than something else) when only a few experiments have been run

- Issue #2: what questions are *compelling*?
  - Just because the authors successfully demonstrate a point, question whether the point was compelling in the first place
  - Ex: showing that a routing protocol has less overhead than one (or more) of several existing protocols is not compelling
    - Issues aren't necessarily overhead related, but deployment related

What’s “Good” Changes Over Time

- If authors can justify their type of evaluation is more accurate, that earns “contribution points”

- There is even a branch of research that looks at developing more accurate methods of evaluation
  - Typically big projects that get government funding to develop and release a simulation package
  - Sometimes, there is an even an effort to start a company to support the software (keep it working over time and make improvements)

- Also work to show that a particular method is flawed
  - Works for both methods of evaluation and parameter choices

Venues Have Expectations

- The more competitive venues have an implied type of evaluation (sometimes more than one)
  - Ex: ACM Mobisys: must have an implemented system

- If/As you get into the serious business of publishing at highly competitive conferences, closely study the kinds of evaluation that are performed for accepted papers

- Even now, pay attention to what papers do for their evaluation
  - What questions they answer
  - What methods they use

Types of Evaluation

- Analysis
  - Particularly good for algorithms
- Simulation
  - Develop your own
  - Use an existing software package (ns-2, OPNET, MATLAB, Qualnet, GloMoSim)
- Prototype
  - Develop a stand alone application
  - Use an infrastructure like PlanetLab
- Measurement
  - Monitor existing system (e.g., traffic statistics/characteristics)
- Emulation
  - Hybrid of simulation, prototype, and possibly measurement

Analysis

- Good for evaluating algorithm
  - Metrics like complexity, memory requirements, overhead, etc.
- Advantages
  - A wide range of conditions can be applied and tested
  - Typically requires no (specialized) hardware
  - Can provide good insight into underlying system behavior
- Disadvantages
  - Typically have to make lots of assumptions
  - Unrealistic assumptions lead to unrealistic results
- Be aware if papers try and add some analysis in combination with another evaluation technique
  - Does the analysis add anything?

Simulation

- Good for studying characteristics of a system
  - Not so good for generating absolute performance values

- Advantages
  - Can give greater realism than simulation
  - Experiments are repeatable
  - Can test wide variations of scenarios
    - Test lots of different factors

- Disadvantages
  - Even the most realistic simulation makes assumptions
  - If used incorrectly, can lead to incorrect conclusions
  - Tradeoff between fidelity and complexity
    - The more sophisticated the simulation, the more time it takes to run

Types of Simulations

- Discrete Event
  - Like the Infocom paper for today
  - Given a set of inputs to the system, they generate events, events update system state, and measurements are taken
  - Time steps can be large (order of minutes) or small (protocols)

- Monte Carlo
  - Does not have a time component
  - Basically probabilistic (flip a coin to see what is likely to happen)

- Trace Driven
  - Inputs are taken from observed behavior and followed according
  - Ex: track user locations, feed into mobility simulator

Evaluation Terminology

- Applies generally, but particularly to simulations
- Parameters
  - Characteristics of a system that affect its performance but aren't varied (assume one static value)
  - Ex: the range of a wireless transmitter
- Factors
  - The characteristics of the system that are varied
  - Factors are varied over a *range*
  - Have a nominal value: baseline value when not being changed
  - Ex: a video source (high-motion video, talking head, scenery)
- Metrics
  - The basis for evaluating a system and drawing conclusions
  - Ex: delay, loss, throughput, SNR, transactions completed, etc.
  - Be careful of the metrics used, for example, "averages" can hide critical behavior

Evaluation Terminology

- Parameters, Factors, and Metrics have to be justified as either being realistic or useful
  - Each has to be justified!
  - Pay attention to corner cases
  - Why parameters weren’t factors
  - Why factor ranges and nominal values were selected
  - Why metrics are the right choice to get to a particular conclusion
- Not always reported in the paper
  - Would use a lot of space
  - Can be a tedious process of explaining *everything*
  - Often times, authors will justify some things but not others
- Remember: embedded in evaluation choices are assumptions. Pay attention to when making choices and when reading about the choices of others

Inputs to a Simulator

- Inputs are typically a parameter or factor
- Probabilistic
  - Ex: Arrivals of events (Poisson, Zipf, etc.)
  - Ex: Probability of some user action
- Trace Driven
  - Inputs are taken from observed behavior and followed according
  - Ex: track user locations, feed into mobility simulator

Prototypes

- Implement a (scaled-down) version of the system

- Can either be:
  - A stand-alone prototype of a system (e.g., a phone app)
  - A testbed with multiple components (e.g., a mesh network)

- A powerful combination is to simulate something that has been implemented, show the simulator is accurate, and then use simulator to test broad range of factors

Prototypes

- Advantages
  - Useful to demonstrate proof-of-concept
  - Useful when combined with other evaluation types
  - Accurate to the extent the prototype is complete
- Disadvantages
  - Hard to repeat experiments
    - Can't control background process load or environment conditions
  - Limited combinations of factors
  - Sometimes hard to stress a system or test scalability
  - Requires time to write software and build testbed
  - Hard to maintain
  - Can be expensive

Measurement

- Only in the last 10 years has measurement really taken off as a way of evaluating systems
- Became popular when trying to understand behavior in complex systems (like the Internet)
  - Ex: route stability
  - Ex: network traffic characteristics
  - Ex: user behavior (e.g., web page requests)
- "Just" measuring a system is no longer sufficient
  - Measure system, find problem, propose solution, evaluate solution
- Measurement is now also being used to collect inputs into simulators

Emulation

- Combination of simulation and prototyping
  - Ex: evaluate performance of protocols over a satellite link
    - Actually send traffic over a link (just not a satellite link)
    - Use link emulator so that link behaves like a satellite link

- Advantages
  - Best of both simulation and prototype: can be made to be more accurate but still has element of environment control

- Disadvantages
  - Requires testbed environment (and all of the associated disads)

Graphs and Tables

- As mentioned before, there is a ton of evaluation work that goes into evaluating a system, but only a limited space to present the work
- Graphs and tables should try to have an intuitive message
  - The less explanation required, the more intuitive and compelling the message
- Generally graphs are better, tables are good when you need to report specific values
- Graphs should be properly formatted
  - Fonts should be as large as fonts of surrounding text
  - Lines should be bold and clear

Paper Analysis Review

- What are the questions being answered?
  - Is the translation between hypotheses/conclusions and metrics reasonable?
- Are the questions worth answering?
- What kind of evaluation was used?
  - Is it the right evaluation?
- What are the implicit assumptions in the evaluation’s parameters, factors, and metrics?
  - Are they justifiable (either because logic says so or the authors provide a cogent explanation)
- Is the evaluation understandable and compelling?