Confidence Weighted Marginal Utility Analyses of Internet Mapping Techniques

Craig Prince and Danny Wyatt

December 6, 2004, CSE561 Networks

Internet Mapping

- What is it?
  - Figure out what the internet looks like
  - Find routers and their interconnections
  - Discern a topology
- What is it good for?
  - Research
  - Simulations
  - Problem diagnosis
  - Routing in overlay networks
  - Spying on competing ISPs

How do you map the internet?

- Cannot directly observe it
- Have to send traceroutes through it
- What you see depends on
  - Source
  - Target
  - Routing policies
  - ... and the topology itself!
- Can only control source and target...
- Errors can occur that do not reflect true topology...
- Things change over time...

How do you map the internet?

- Cannot directly observe it
- Have to send traceroutes through it
- What you see depends on
  - Source
  - Target
  - Routing policies
  - ... and the topology itself!
- Can only control source and target...
- Errors can occur that do not reflect true topology...
- Things change over time...
- ...so just add as many as you can

How do you map the internet?

- Cannot directly observe it
- Have to send traceroutes through it
- What you see depends on
  - Source
  - Target
  - Routing policies
  - ... and the topology itself!
- Can only control source and target...
- Errors can occur that do not reflect true topology...
- Things change over time...
- ...so just add as many as you can
- Is more really better?

Aims

- How do different mapping tools compare in their efficient use of data?
- Are some kinds of measurements more valuable than others?
- If we are uncertain of our observations, how would different methods address that uncertainty?

The Data: 3 Mapping Tools

- Skitter
  - 24 distributed sources
  - Each uses 1 or more of 4 lists of preselected target
  - Continually loop through lists
  - We use 3 days: 12/18-20, 2002

The Data: 3 Mapping Tools

- Skitter
  - 24 distributed sources
  - Each uses 1 or more of 4 lists of preselected target
  - Continually loop through lists
  - We use 3 days: 12/18-20, 2002
- Scriptroute
  - 70 distributed PlanetLab nodes
  - Each used same list of 125,000 address prefixes
  - Attempted all traces once a day for three days (same as above)

The Data: 3 Mapping Tools

- Skitter
  - 24 distributed sources
  - Each uses 1 or more of 4 lists of preselected target
  - Continually loop through lists
  - We use 3 days: 12/18-20, 2002
- Scriptroute
  - 70 distributed PlanetLab nodes
  - Each used same list of 125,000 address prefixes
  - Attempted all traces once a day for three days (same as above)
- Rocketfuel
  - 837 distributed public traceroute servers
  - $\approx 60,000$ targets
  - Heuristic pruning of source-target pairs to maximize coverage
  - Data collected over January, 2002

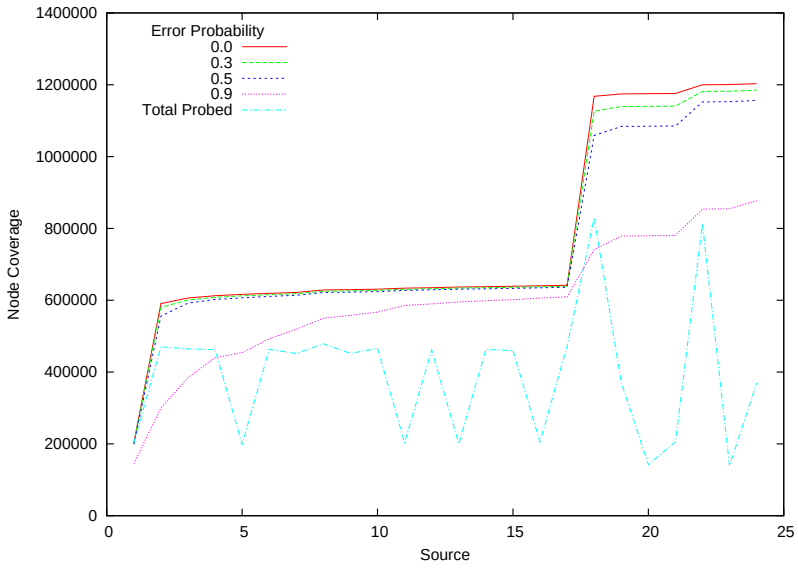
Some Definitions

- A map is a directed graph $G = (V, E)$
- There is some impossible, true map $\hat{G} = (\hat{V}, \hat{E})$ with 100% perfect coverage
- A map is made by aggregating many *measurements*
  - Sources
  - Targets
- *Coverage* is how well one map approximates another
- *Marginal coverage* is how much each measurement contributes to its map
- We evaluate the marginal coverage of each of the three tools

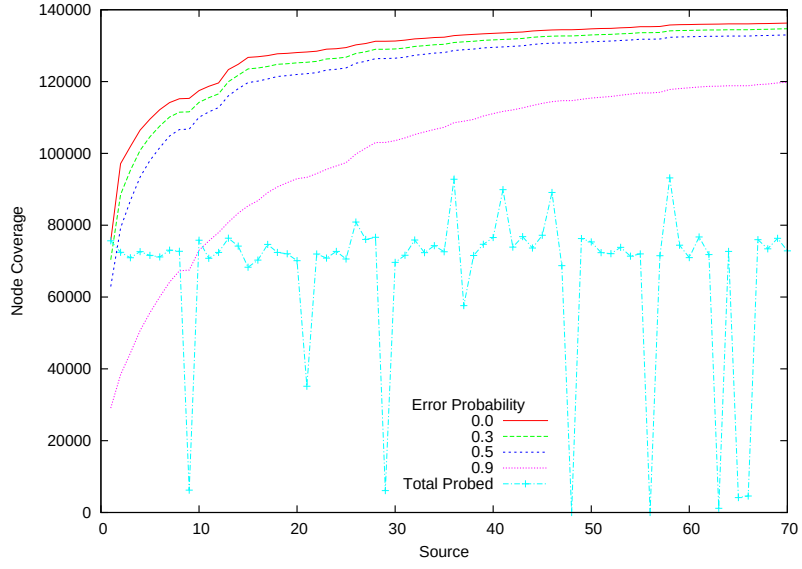
More Definitions: Confidence Weighting

- Traceroutes are noisy sensors with probability of error d
- $n(e)$ is number of observations of edge $e \in E$
- Probability that e exists is $P(e) = 1 - d^{n(e)}$
- *Edge coverage* of G is mean probability of all edges: $\frac{\sum_{e \in E} P(e)}{|E|}$
- *Node coverage* is defined similarly
- For each analysis, also consider how it compares according to different values of d

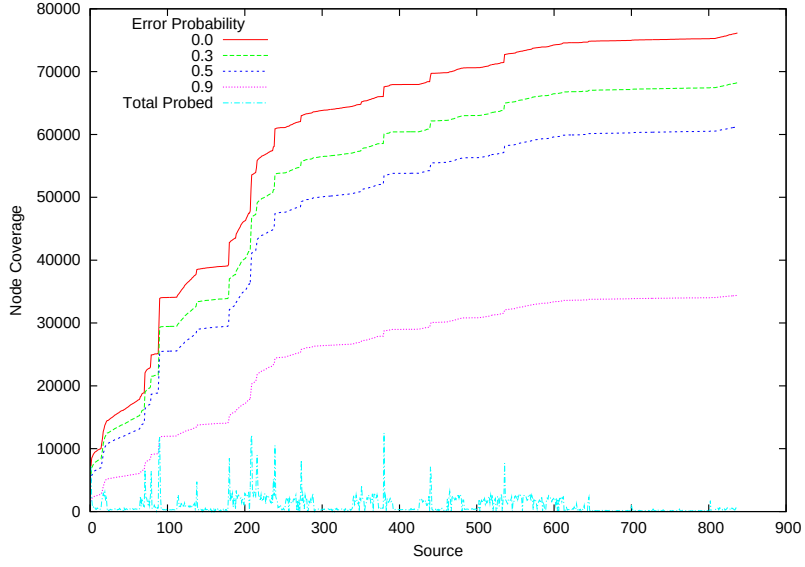
Node Coverage per Source



Skitter



Scriptroute



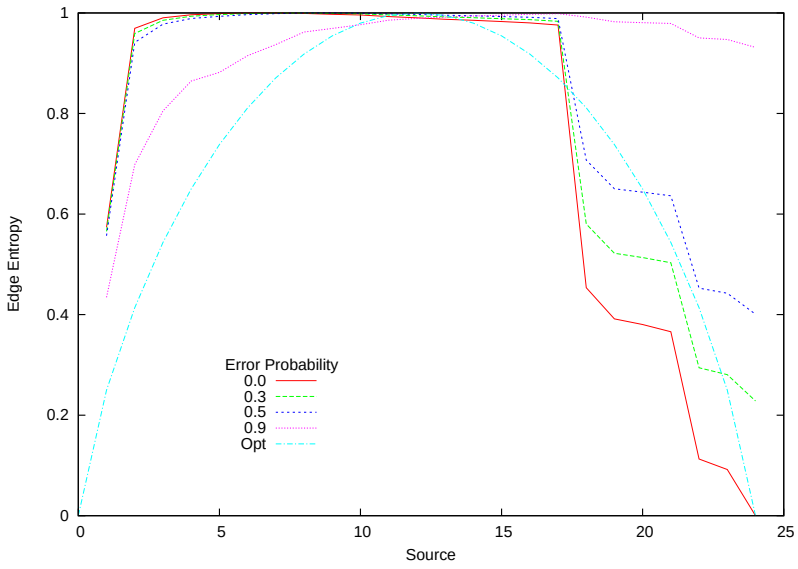
Rocketfuel

Entropy

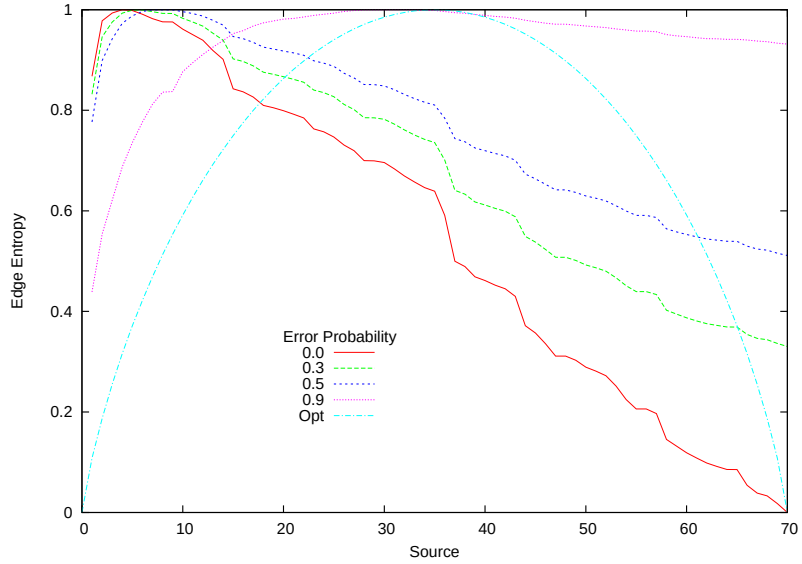
$$H(A) = \sum_{a \in A} -P(a) \log(P(a))$$

- Average number of bits needed to encode each event a
- We take the entropy of the mean node and edge distributions
- Should always be changing

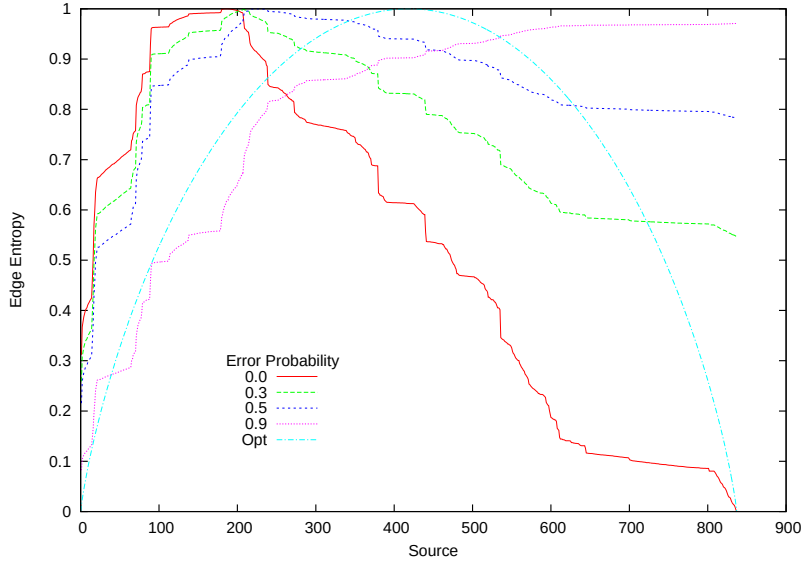
Edge Entropy per Source



Skitter



Scriptroute



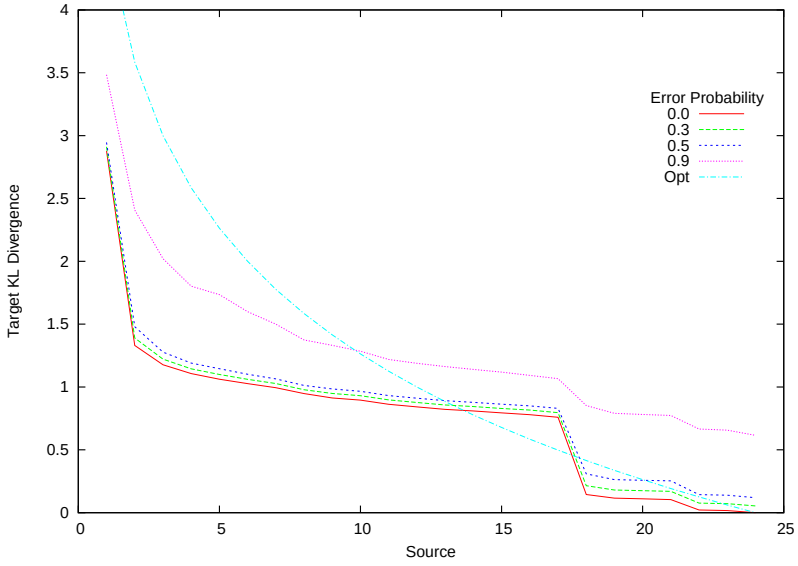
Rocketfuel

Kullback-Leibler Divergence

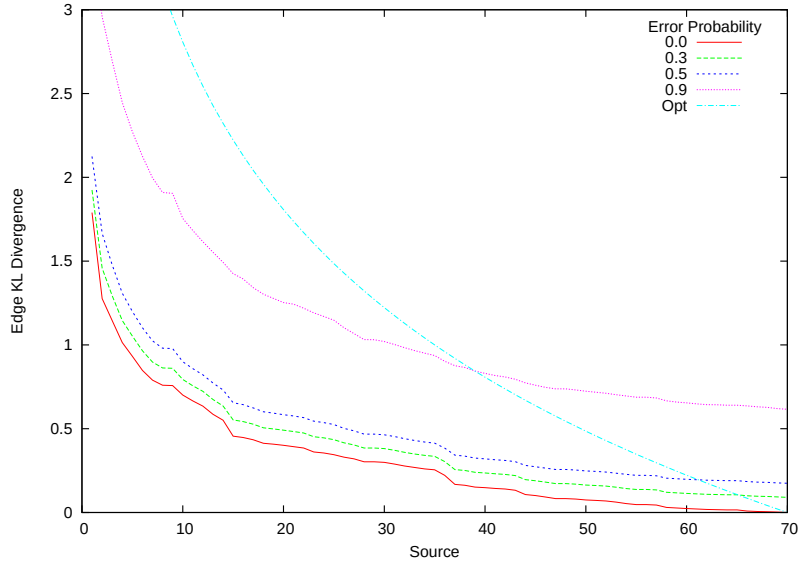
$$KL(A||B) = \sum_a p_A(a) \log \left(\frac{p_A(a)}{p_B(a)} \right)$$

- Also known as relative entropy
- Average extra bits per event for encoding according to the wrong distribution
- We measure divergence between coverage up to a measurement and final coverage
- *Marginal utility* is the decrease in K-L divergence between measurements

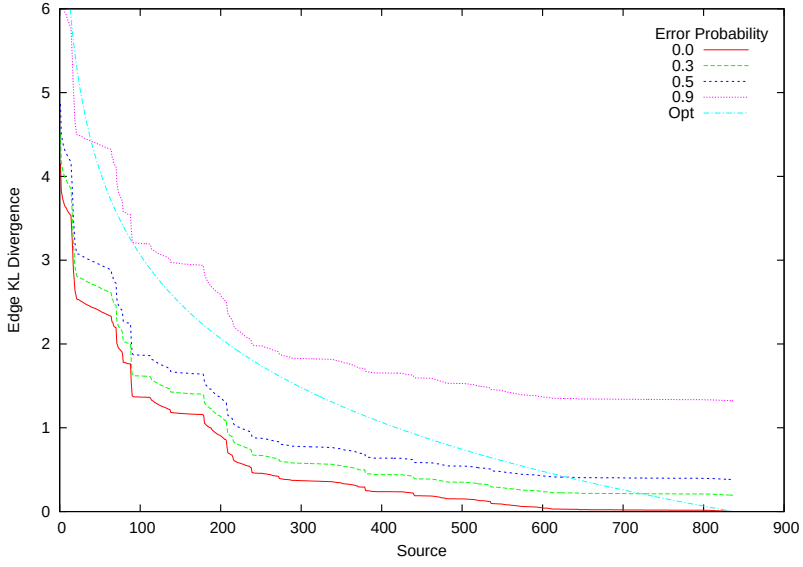
K-L Divergence per Source



Skitter

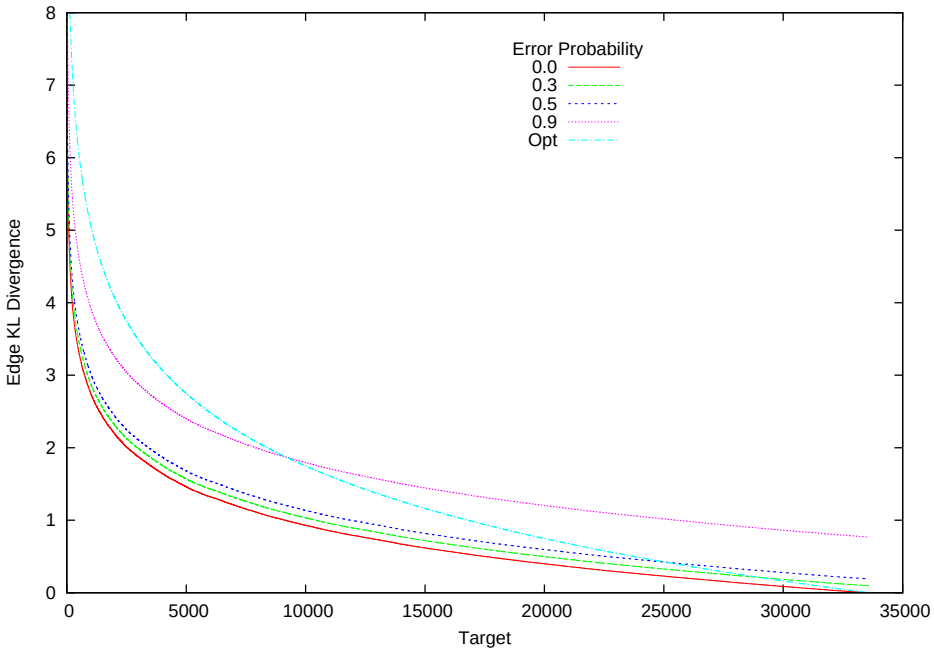


Scriptroute

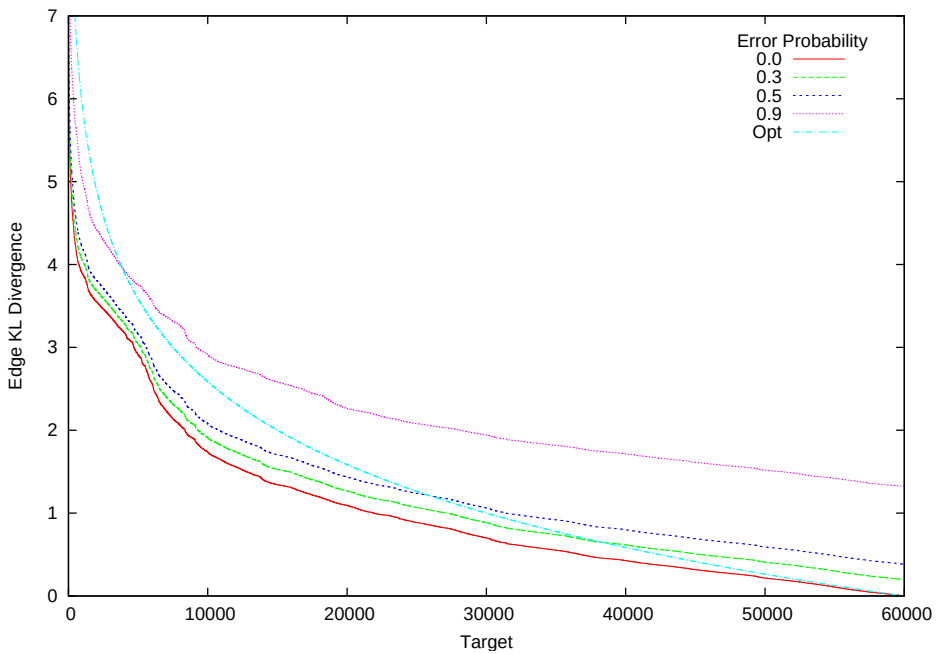


Rocketfuel

K-L Divergence per Target



Scriptroute



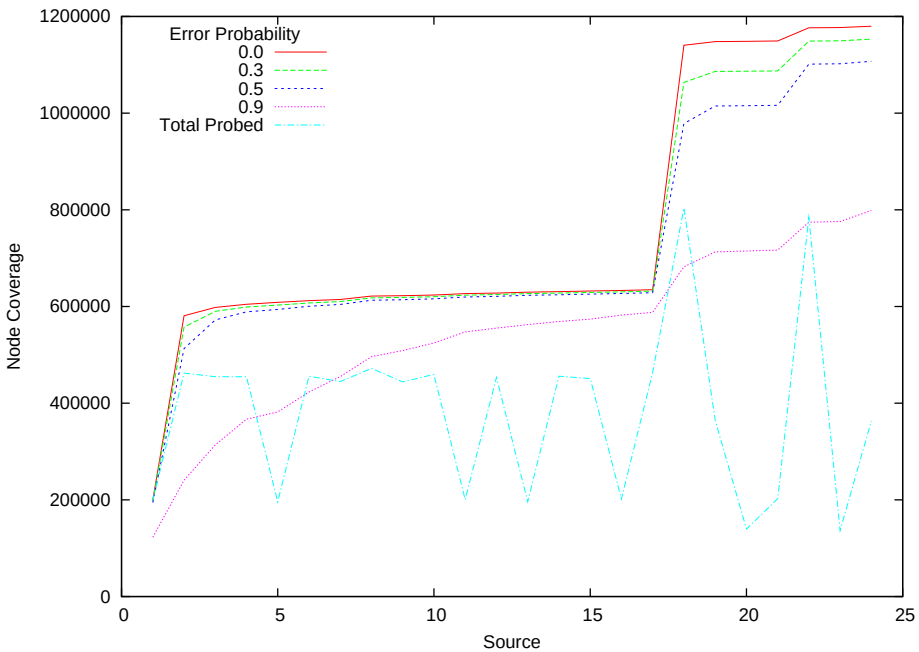
Rocketfuel

Conclusions

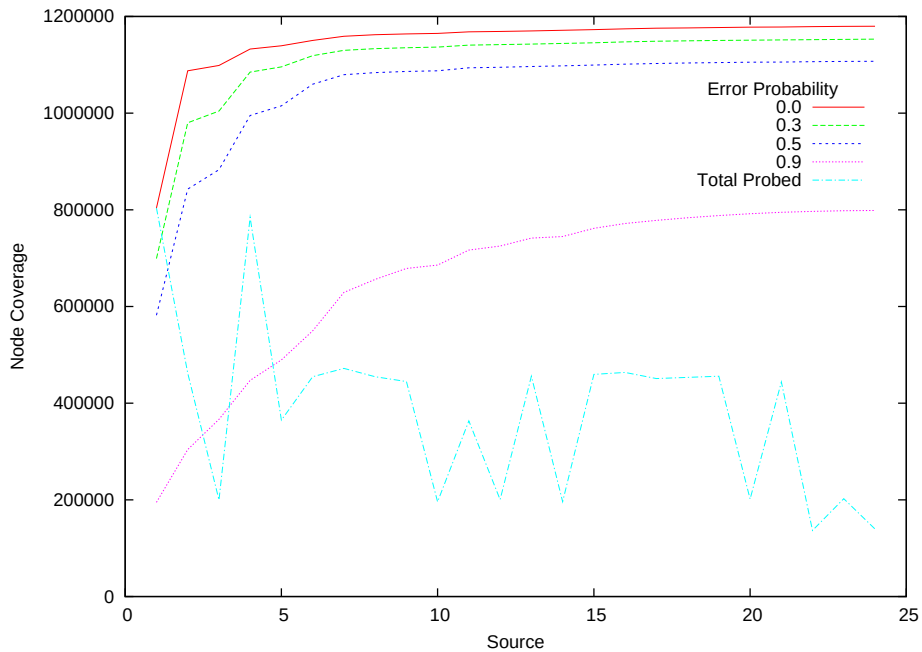
- Adding targets is more useful than adding sources
  - Half of all coverage comes from the first few sources
- Rocketfuel does increase its per measurement return
  - More targets always yield more information
  - More sources have diminished returns, but higher than other tools
- There is a pronounced trade off in confidence
  - Rocketfuel has more divergence between different error probabilities
  - More redundant tools are less effected

Conclusions

- These metrics can be used as heuristics for quicker mapping
- Reordering the second two days of Skitter data according to the first day:



Day 1



Day 2 and 3, reordered

Conclusions

Questions?