Toward Privacy Definitions for Anonymized Network Data

Scott E. Coull
University of North Carolina
Chapel Hill, NC
scoull@cs.unc.edu

Joint work with: Michael Bailey, Fabian Monrose, Michael Reiter

Network Data Anonymization

- Network data is useful in research projects:
  - Network measurements
  - Evaluation of security techniques
- Network data contains sensitive info:
  - Usernames & passwords
  - Web sites
  - Network security posture

Network Data Anonymization

- **Network Data Anonymization**
  - Alters data to remove sensitive info
  - Maintains general utility to researchers

- Ad-hoc process with no privacy guarantees
  - Remove payloads
  - Replace IP addresses with pseudonyms
  - Quantize timestamps

Network Data Anonymization

- No rigorous way of evaluating whether specific anonymization procedures “work”
  - Primary goal: maximize utility/minimize changes

- Leads to so-called **inference attacks**:
  - Attackers use external info to infer sensitive info that was removed
  - Previous work has shown that host identities and user web browsing activities can be revealed

Relation to Microdata Privacy

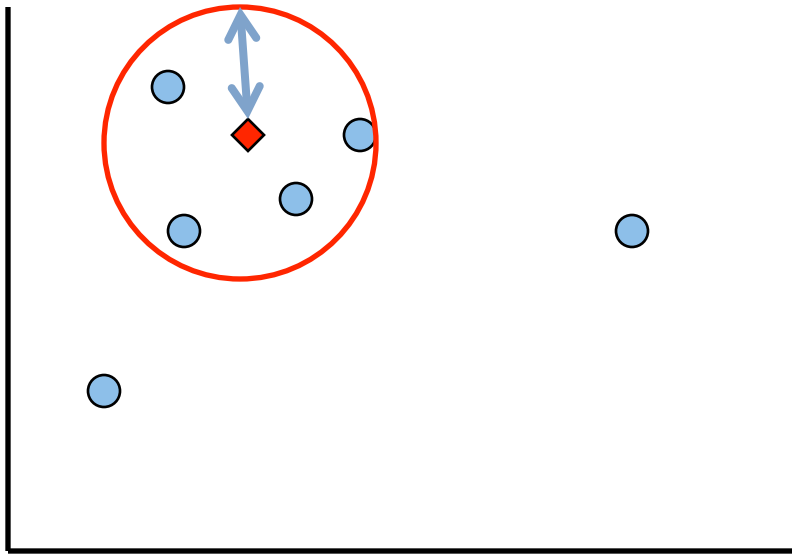
- **Microdata**:
  - Database of attributes for individuals
  - Ex: Census data – location, sex, income, etc.
- Long history of privacy research
  - k-Anonymity, (c,t)-isolation, etc.
- Superficially similar to network data
  - Database of records, desire to hide sensitive info, etc.

k-Anonymity Example

First Name	Last Name	Height	Weight	Religion
----	----	5’ 2” - 6’ 2”	110-175lbs	Muslim
----	----	5’ 2” - 6’ 2”	110-175lbs	Catholic
----	----	5’ 2” - 6’ 2”	110-175lbs	Atheist

Generalize/remove attributes to ensure all records are same as k-1 other records

(c,t)-isolation Example



Ball of radius of c around the point,
with t points within it

Relation to Microdata Privacy

- In reality, network data is far more complex
  - Multiple "objects" to protect
    - Ex: hosts, web pages, users
  - Temporal data
    - Many observations of the objects over time
    - Often containing tens or hundreds of millions of records
  - Strong semantics due to network protocols
    - Not possible to simply add noise or create equivalence classes without breaking semantics

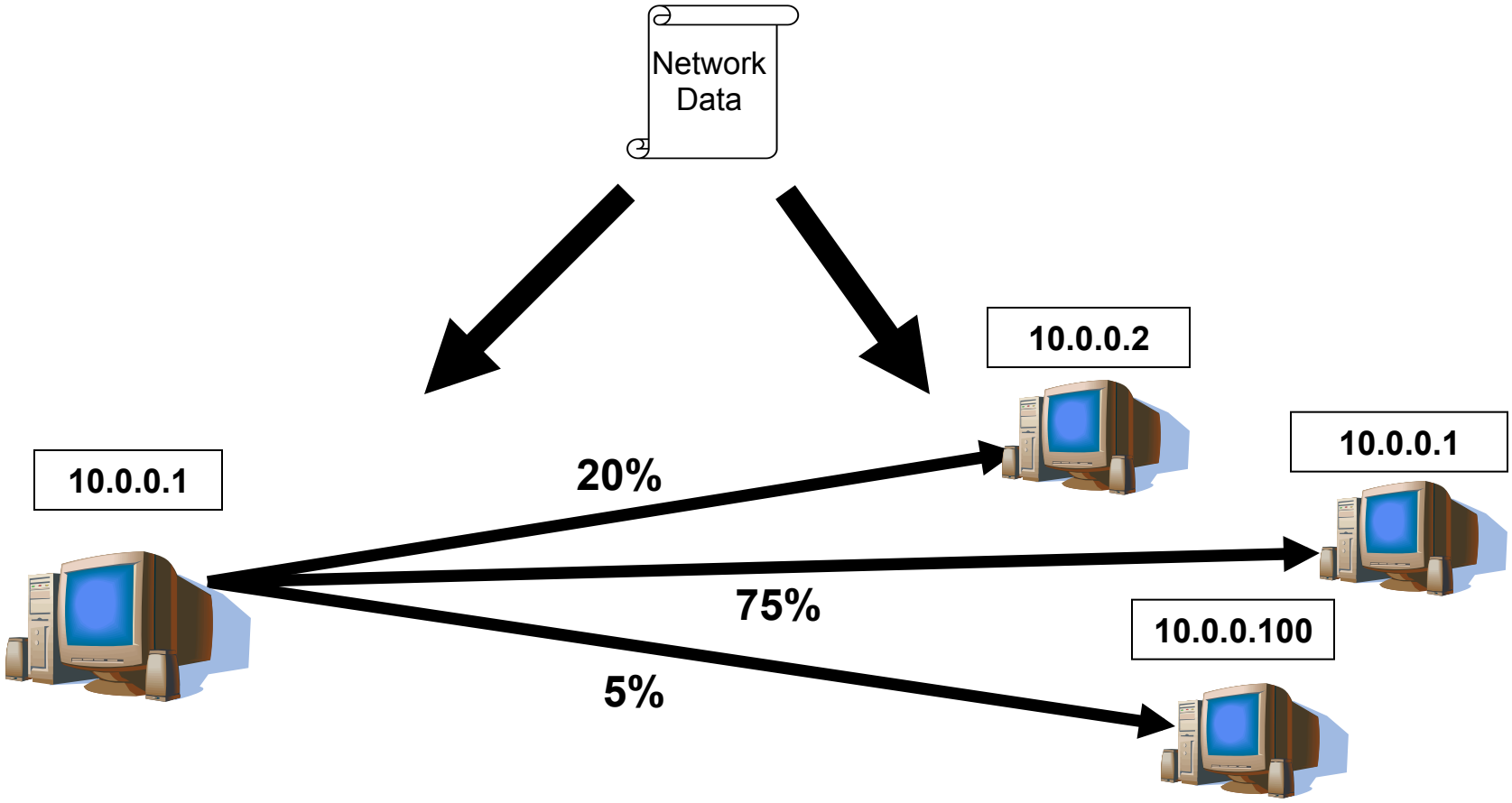
Outline

- Methodology for measuring anonymity
  - General method of calculating object similarity
    - From an adversary's perspective
  - Connect similarity to privacy definitions
- Evaluation on flow data from U. of Michigan
  - Overall anonymity
  - Anonymity of specific features
  - Comparison between anon. methods

Goal

- Quantify anonymity of objects based on adversary-aware notion of similarity
  - Similarity underlies all microdata privacy def' ns
- Provide analysis tools to determine efficacy of anonymization policies on data
  - Place anonymity in context of microdata def' ns

Methodology



Methodology

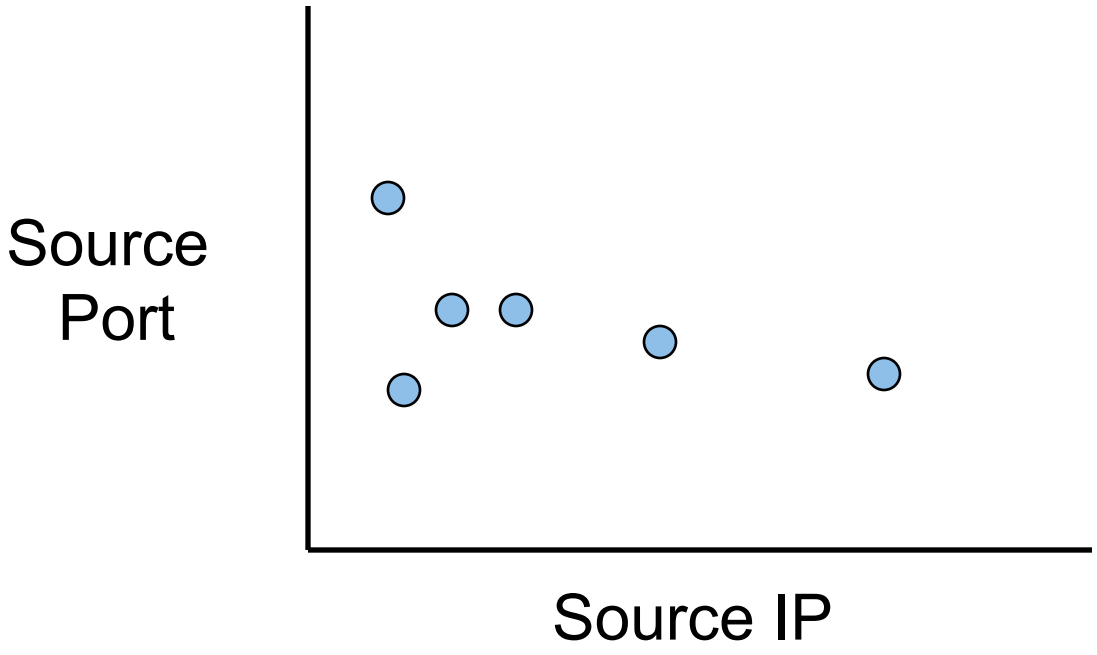
- **Idea:** Examine unanonymized data and compare objects in a pairwise fashion
  1. Define objects
  2. Use a similarity measure to compare objects
  3. Induce a probability distribution over identities
  4. Calculate entropy of distribution

Calculating Similarity

1. Treat records as n-dimensional points
   - Each field is a dimension
   - User-defined distance measures based on intended anonymization policy
     - Capture information available to adversary and general semantic meaning
     - Intuition: Distance between points indicates adversary's belief that they are the same

Calculating Similarity

1. Treat records as n-dimensional points

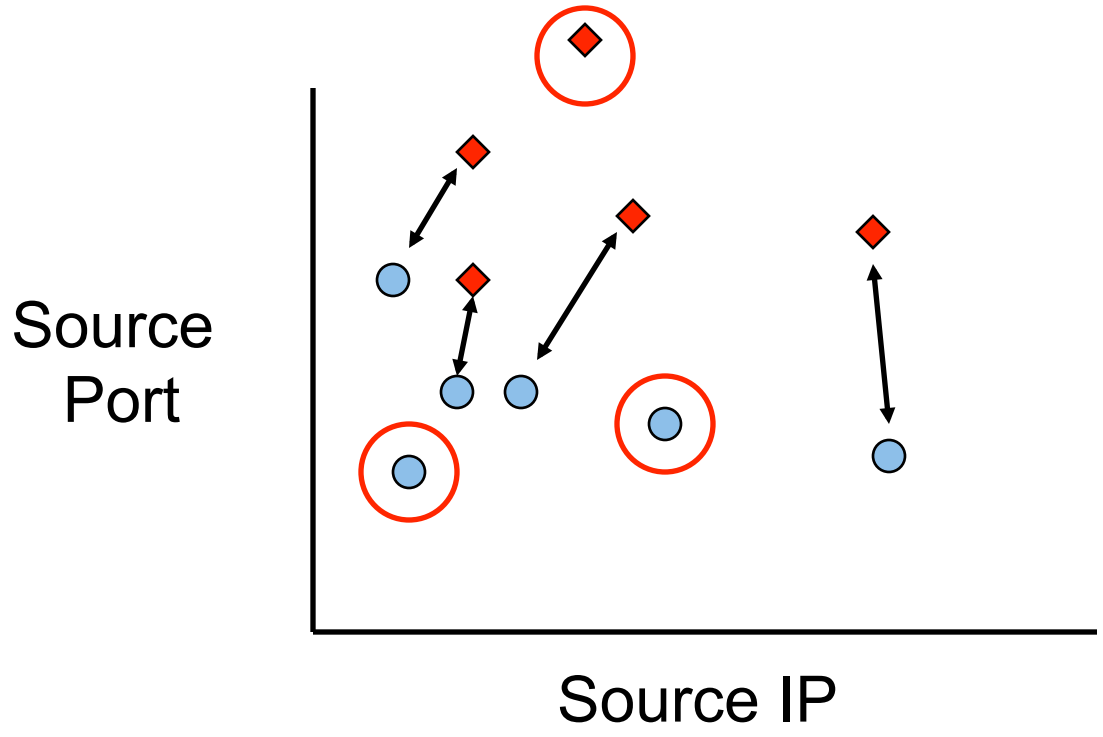


Calculating Similarity

2. Calculate pairwise similarity among objects
   - A given object is compared to all others
   - Treat each object as a sequence of points
   - Use sequence alignment to find minimal distance
   - Convert distance to normalized similarity by subtracting the maximum and dividing

Calculating Similarity

2. Calculate pairwise similarity among objects



Calculating Similarity

3. Induce a probability distribution over potential identities for the object being evaluated using similarity scores
   - Recall: distance between points based on adversary's belief in their similarity
   - Belief in identities based on relative similarity

$$P(X = Y) = \frac{S(X,Y)}{\sum_{\forall Y} S(X,Y)}$$

Calculating Similarity

4. Entropy of distribution indicates anonymity
   - Low entropy means distribution is peaked with similarity only to the object itself
   - High entropy means that many objects share an equivalent level of similarity
     - Note: this does not guarantee high similarity

Privacy Definitions

- Analog to k-Anonymity:
  - Recall: Guarantees k records must be similar
  - Entropy can be thought of as expected number of similar objects
  - Calculated as: $k = 2^{H(X)}$
  - Simply set a lower bound on entropy

Privacy Definitions

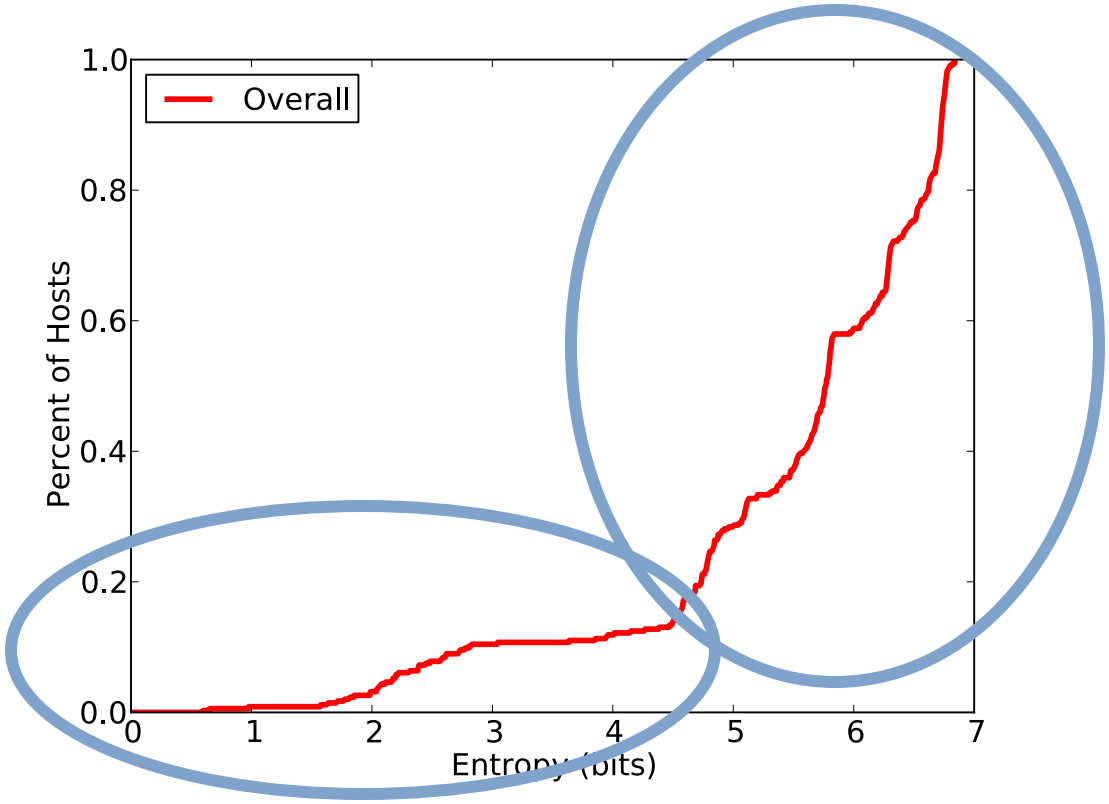
- Analog to (c,t)-isolation:
  - Recall: Guarantees t points within a radius of c
  - Alignment distance is equivalent to c
  - Number of objects within alignment distance c is equivalent to t
  - Provides a notion of both quality *and* quantity of anonymity for objects

Evaluation

- One day of flow data collected at the University of Michigan
  - 5 /24 Networks
  - Computer Science Dept.
    - 31,157,107 flows, 345 hosts
    - Processing time: 8 days 14 hours
  - Engineering School
    - 11,995,395 flows, 1,280 hosts
    - Processing time: 5 days 20 hours

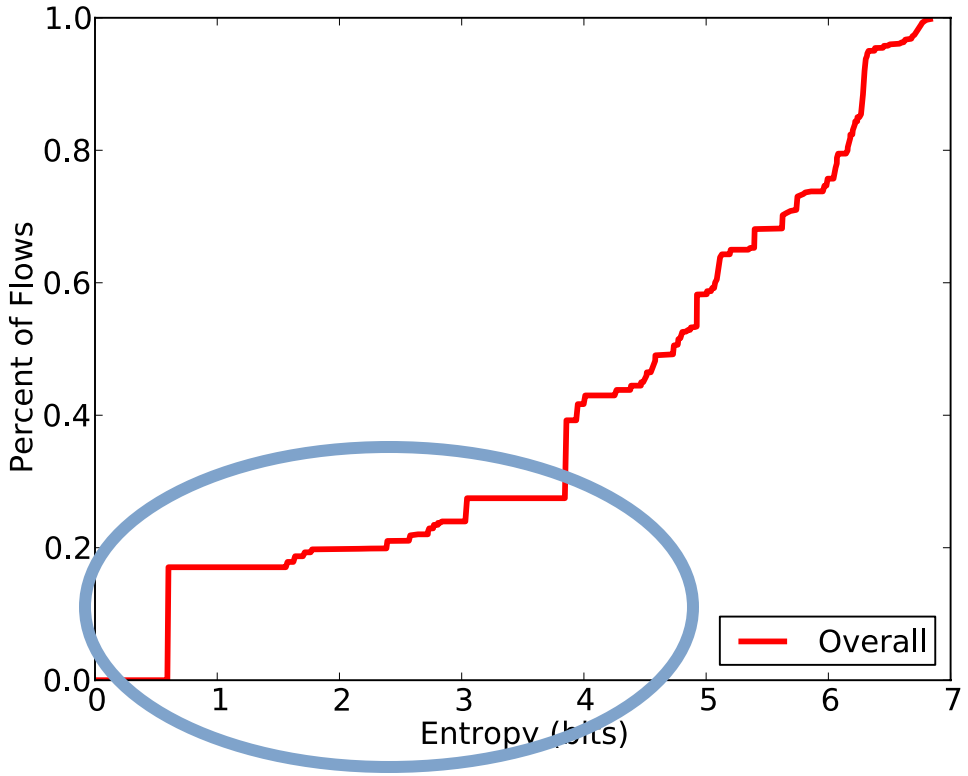
Evaluation

- CDF of host entropy for CS dept:



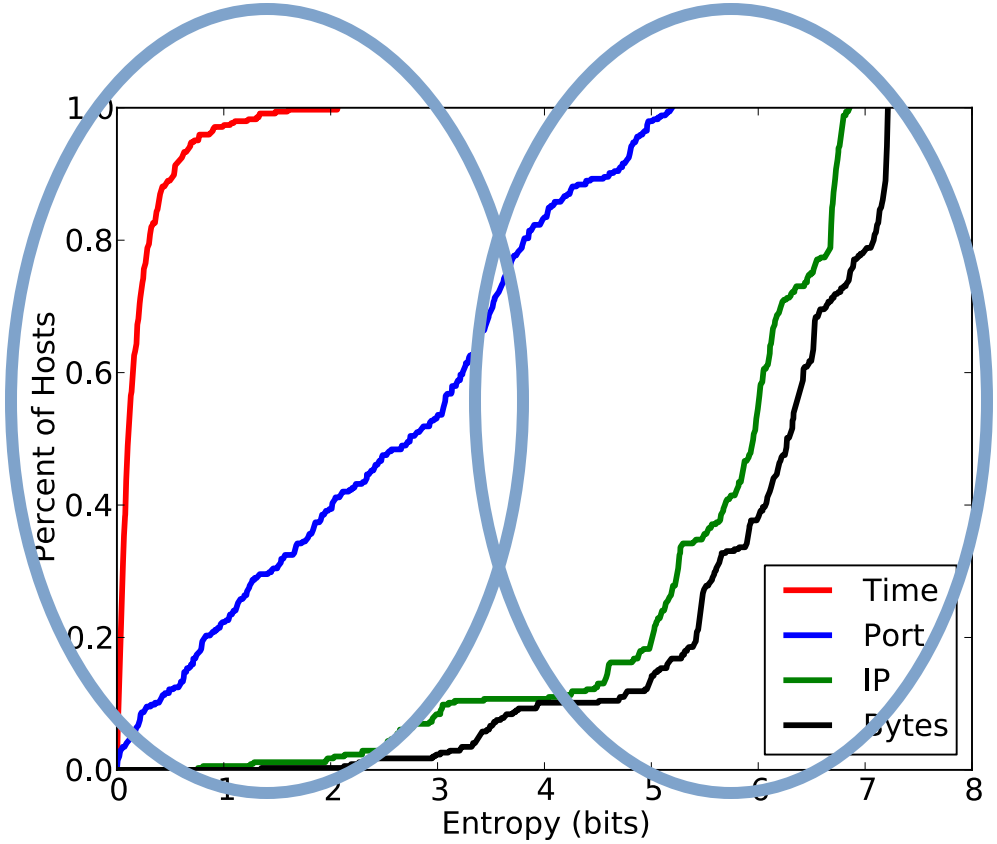
Evaluation

- CDF as percent of flows for CS dept:



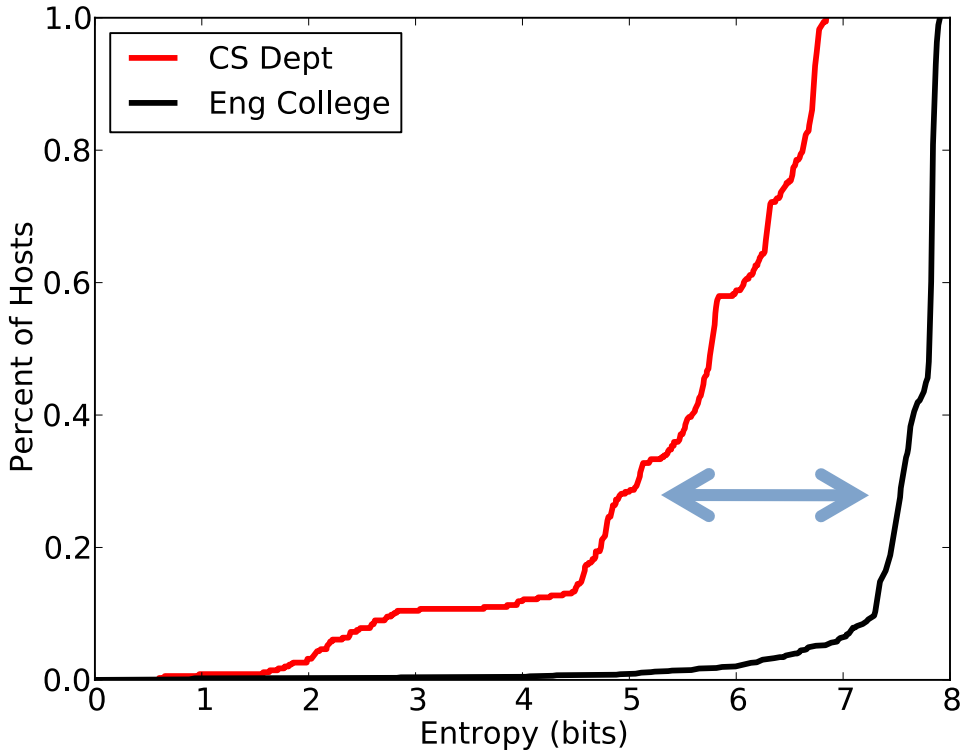
Evaluation

- CDF of each dimension in isolation:



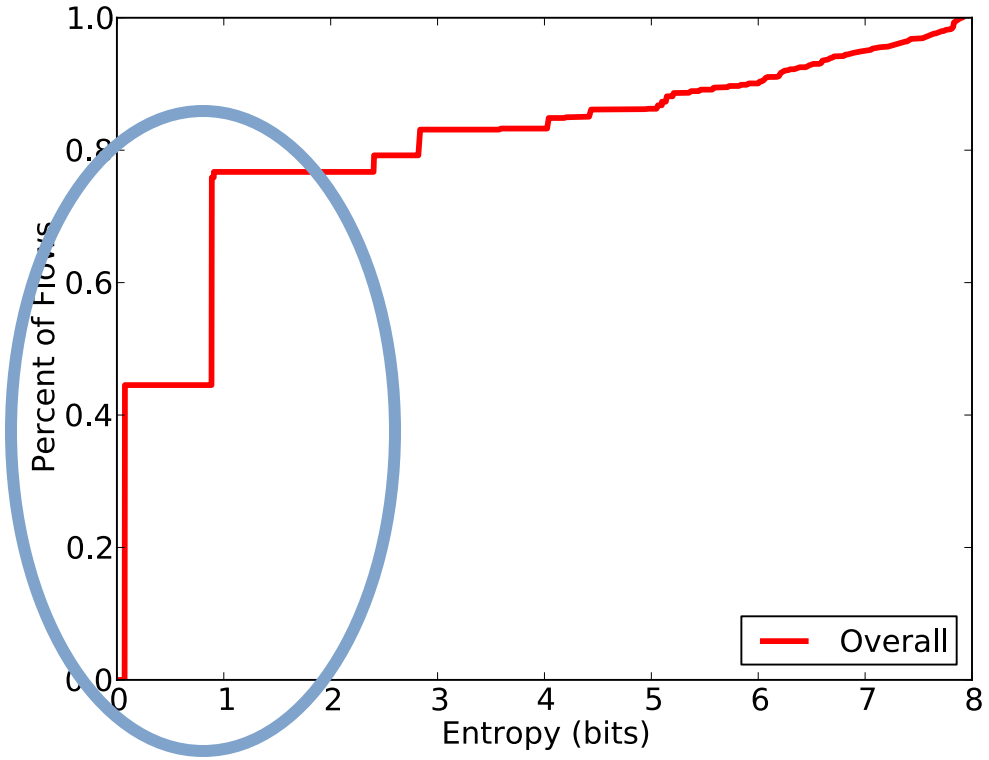
Evaluation

- CDF comparing CS to Engineering:



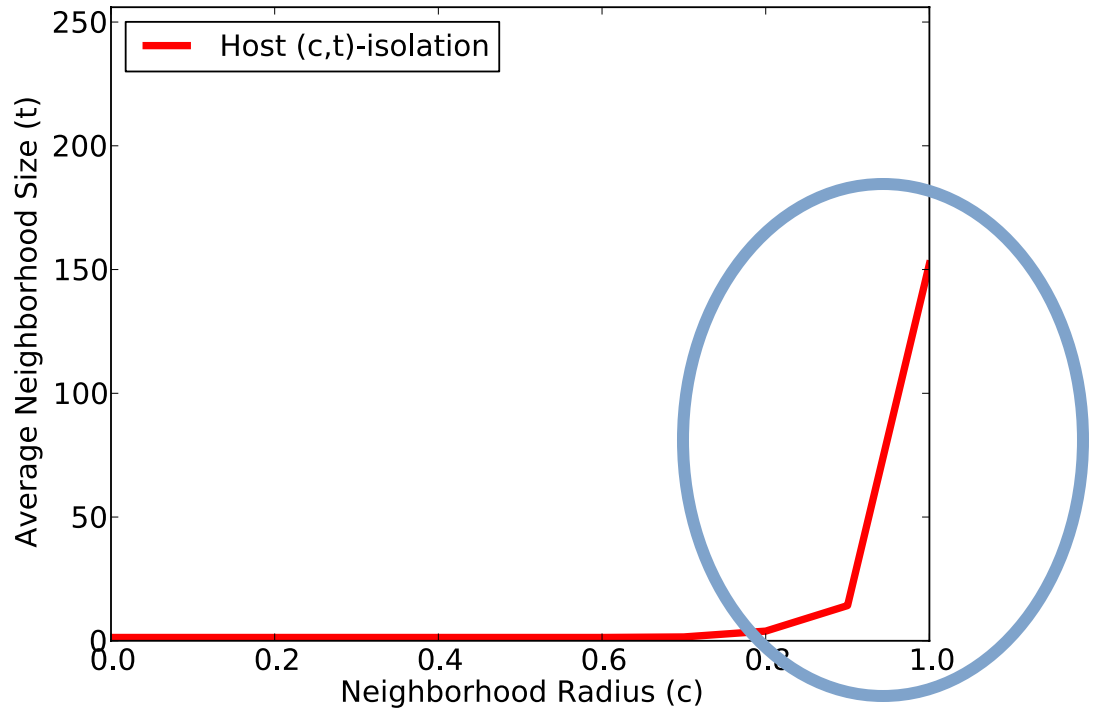
Evaluation

- CDF as percent of flows for Engineering:



Evaluation

- Average neighborhood size for fixed distance in Comp Sci. network:



Evaluation Summary

- Examining overall dataset privacy
  - CDF of hosts or flows
- Verifying anonymization policy efficacy
  - CDF of individual field types
- Comparing anonymization methods or datasets
  - What type of IP pseudonyms to use?
  - Where to place network monitors?
- Quality of anonymity
  - (c,t)-isolation notion of anonymity

Conclusion

- Presented rigorous method for analyzing anonymity based on similarity of objects
- Similarity notion for network objects allows for connection to well-known privacy definitions
- Provable anonymization methods?
  - Need a definition of utility to allow “blending” of objects that is semantically meaningful
  - (c,t)-isolation provides an interesting avenue