DIMACS Working Group on Measuring Anonymity
Notes from Session 4: Using Differential Privacy

Scribe: Matthew Wright

In this session, we had three 15-minute talks based on submitted abstracts and about 45 minutes of "panel" discussion with the three speakers as panelists. The focus of the session was on differential privacy and whether and how it could be used for the study of anonymity systems.

Talk #1: **Catuscia Palamidessi**. Differential privacy and anonymity. (15 min.)

Quick overview of differential privacy
- Privacy in statistical databases: side information makes it hard.
- Differential privacy: gives a measure anyway
  - For every possible distribution of the value of an individual, the answer doesn't change (much) if you take the individual out of the aggregate query.
  - Independent from the a-priori info
  - Composable: loss of privacy can be controlled
  - Focus on the individual and on the worst case
- epsilon-DP

Treating anonymity as a channel matrix
- Differential privacy induces a bound on the Shannon entropy and on the min-capacity

Much prior work: translate other privacy problems into data problems
- Extension [PETS '13]: d(x, x') *level of distinguishability* between x and x' -- is the same as DP for some settings. *d-privacy*.
  - http://freehaven.net/anonbib/papers/pets2013/paper_57.pdf
- Universal optimality for more cases.
- Maybe this would work for anonymity, too?

Talk #2: **George Danezis**. Measuring anonymity: a few thoughts and a differentially private bound. (15 min.)
- Multiple applications leak anonymity.
  - We have been pretty poor at characterizing this.
  - Can we use something from DP for this?
  - Not a standard application of DP...
- The heart of a DP theorem: a bound on the likelihood ratios
  - For all Obs., Pr[Obs | Hidden state, constraints] < $e^{\backslash epsilon}$

Left World	Right World
Mix, Alice and other sender,	P_{AB}' -- different prob. A to B

Bob and other receivers. | Otherwise the same
Prior: a prob. of A to B, |
a prob. of others to B. |

What is the Pr[volume to B | Left] vs. Pr[vol. to B | Right]?

- The bound on the likelihood ratios holds in *most* cases -- not all.
  - Add \delta (This makes \epsilon-\delta DP)
  - \epsilon-\delta DP captures the notion that there are rare, extreme cases in which a lot is revealed, so we discount those if rare enough
- This is not the traditional DP model. The prob. of others to B should not be assumed usually. We can try to make it hold w/ some dummies.
- It's a bit weird that we assume the worst case every time -- maybe we can get a bit more realistic.

[A set of questions and answers from the following panel discussion]
Why do you not say the number of other users is not a security parameter?

- A: epsilon doesn't depend on the number of users; delta does. In the case of a big leak, the number of users doesn't matter.

Left and Right don't seem like a very interesting case if Alice always talks to Bob.

- A: If P_{AB} is modest, it's quite interesting.

Do you really care about the upper bound for Alice's rate to Bob?

- A: If you have no bound from above, then you have to assume that the others have a high sending rate to Bob too.

The hidden state P_{AB} is the prior?

- Not really.

What kind of practical environments could this be applied to?

- Tails + ISDN-mix. Reason on the privacy.

Talk #3: **Scott E. Coull**. How (Not) to Apply Differential Privacy in Anonymity Networks.(15 min.)

- Main point: DP can't be applied (directly)
  - See "No Free Lunch in Data Privacy" http://www.cse.psu.edu/~dkifer/papers/nflprivacy.pdf
- Popular interpretations of DP (ALL NOT TRUE)
  - No assumptions on the data
  - Super strong (attacker knows all but one)
  - Robust to arbitrary background knowledge
- Key: DP assumes independence between records

Network information

- Objects in computer networks: users, PCs, websites, packets, etc.
- Objects influence other objects: social group affects the user, user affects the sites selected, sites affect the packets/timing/etc....

- Objects contain information about both "lower" and "higher" objects
- Adversary may have a more complete view than we do

Breaking Independence Assumptions: no DP

- Knowledge of correlations leads to failure (no semantic value of DP)
  - e.g. social network graphs: one edge alters the growth of the graph
  - A single record (edge) has a huge impact on the data
- Fix?: k*epsilon noise, which is monstrous.
- Network data: Same issue.
- e.g. remove one TCP packet (handshake, slow start, etc.), screw up the flow a lot
  - Single packet privacy at best

Discussion in Talk #3

- A: DP doesn't promise any privacy.
- B: DP promises no assumptions about the data, but independence is one.
- C: Doesn't that mean the notion of neighbor is ill-defined?
- D: Sure, but it's solved at extreme cost.
- E: It seems that another way of looking at this is that DP correctly defined would capture this, but getting the guarantees you want is really expensive

Pufferfish Framework

- Generalization of DP that includes more assumptions
- Check against all distributions of [inputs?]

Applying this to anonymity

- Challenge to define the attacks
- Think about the multiple attacker models at once in a single metric
- Maybe Bayesian approaches? Most semantically useful information.
- DP can be an inspiration, due to *possibility* of a strong guarantee.

Panel Discussion (45 min.)

Note: Discussion participants are labeled 'A' to 'Z' for each question.

Will DP ever be useful in a real system?

- A: I think so. It tells us something about the system. It forces us to think deeply about the system. Any kind of modeling will uncover bugs.
- B: in a small number of cases, where things can be done cleanly, it will be very useful. The context is built into it. But not much faith that we can create distributions for these things. Even in the Pufferfish paper, there are well-defined distributions that don't reflect reality.
- B: Also, it ties overhead and security together. Epsilon param.
- C: DP is a very interesting approach here.