Differential Privacy in the Wild: Challenges and Open Questions

Ashwin Machanavajjhala

ashwin @ cs.duke.edu
World according to DP research

Figure 1: Two alternative models for the workflow of private data analysis: (a) depicts an idealized workflow where private algorithm $A$ performs task $T$ on the private input; (b) presents a more realistic but more complex workflow for typical data mining tasks. Our proposed research focuses on the shaded boxes.

(a) Simple, idealized workflow.

(b) More complex, recommended workflow.

Addressing the issues outlined above is essential if the rich array of private data mining algorithms developed in the research community are going to see widespread adoption. Although our vision poses new challenges for private data analysis, it does not necessarily mean the utility of the output must be reduced.

Proposed research

We aim to achieve three broad objectives: (a) systematically formalize a complete workflow for differentially private data analysis, (b) develop solutions for key components of the workflow that are currently unknown, and (c) improve the empirical evaluation and dissemination of differentially private algorithms under realistic working conditions. Our plan includes the following outcomes:

- Private algorithms for cleaning and transformation (Research Aims 1-4; Sec. 2.2) In preliminary work, we find that we can improve the accuracy of differentially private classifiers through preprocessing. In addition, we see that some standard preprocessing techniques that traditionally work well break down in the private setting. Thus, we plan to comprehensively re-design preprocessing algorithms so they satisfy privacy conditions and achieve maximum utility.

- Improved model selection and support for iteration. (Research Aims 5-6; Sec. 2.3) We will develop improved model selection algorithms by characterizing how privacy impacts bias-variance tradeoffs in model selection. While these tradeoffs are well-studied in the non-private setting, the noise introduced for privacy alters the equation. For example, preliminary work shows that some query answering strategies can be almost universally improved by imposing a data-dependent uniformity bias. We plan to expand this work to a broader range of tasks. In addition, we propose to develop algorithms that support the specific kinds of iterative processes that arise in model selection, including ensemble methods.

- Principled empirical evaluation of private workflows (Research Aims 7-10; Sec. 2.4) To improve empirical evaluation we plan to design and build a public web-based platform, DP COMP, to support principled evaluation of private data mining workflows and to encourage dissemination of related code and data. In conjunction with DP COMP, we will develop new principles for determining appropriate privacy policies as well as new standards for making meaningful judgements of utility. These privacy and
Figure 1: Two alternative models for the workflow of private data analysis: (a) depicts an idealized workflow where private algorithm $A$ performs task $T$ on the private input; (b) presents a more realistic but more complex workflow for typical data mining tasks. Our proposed research focuses on the shaded boxes.

- **Data Owner**
- **Analyst**

### Problem Statement

- **Task $T$**
- **Private Input**
- **Output $\circ o \pm error$**

#### Issues

- **No guidance on setting privacy policy.** While prior works have given semantic interpretations to differential privacy, none provide concrete guidance on how to set parameters. Setting policy is even more complex with emerging variants of differential privacy [22, 75, 113], which give the data owner potentially overwhelming flexibility in the protected properties of the private input. An algorithm may technically satisfy differential privacy, but an inappropriate policy can lead to blatant disclosure [73].

- **Inadequate empirical evaluation.** A final barrier is that the research community has not made a convincing argument that algorithms are ready for deployment. Experiments are done under idealized conditions with an emphasis on comparative error rates (is $A_1$ better than $A_2$?) rather than absolute utility (is the error in the output of $A$ acceptable?). Emerging algorithms have complex dependencies on the input, making it difficult to extrapolate performance from published results. There has been little attention paid to the methodology of empirical evaluation and no benchmarks have been established. As a result, a data owner cannot easily assess the state-of-the-art methods for their task.

#### Proposed Research

- **Three Broad Objectives:**
  1. **Systematically formalize a complete workflow for differentially private data analysis.**
  2. **Develop solutions for key components of the workflow that are currently unknown.**
  3. **Improve the empirical evaluation and dissemination of differentially private algorithms under realistic working conditions.**

#### Outcomes

- **Private algorithms for cleaning and transformation (Research Aims 1-4; Sec. 2.2)**
  - In preliminary work, we find that we can improve the accuracy of differentially private classifiers through preprocessing.
  - In addition, we see that some standard preprocessing techniques that traditionally work well break down in the private setting. Thus, we plan to comprehensively re-design preprocessing algorithms so they satisfy privacy conditions and achieve maximum utility.

- **Improved model selection and support for iteration. (Research Aims 5-6; Sec. 2.3)**
  - We will develop improved model selection algorithms by characterizing how privacy impacts bias-variance tradeoffs in model selection.
  - While these tradeoffs are well-studied in the non-private setting, the noise introduced for privacy alters the equation. For example, preliminary work shows that some query answering strategies can be almost universally improved by imposing a data-dependent uniformity bias.
  - We plan to expand this work to a broader range of tasks.
  - In addition, we propose to develop algorithms that support the specific kinds of iterative processes that arise in model selection, including ensemble methods.

- **Principled empirical evaluation of private workflows (Research Aims 7-10; Sec. 2.4)**
  - To improve empirical evaluation we plan to design and build a public web-based platform, DP COMP, to support principled evaluation of private data mining workflows and to encourage dissemination of related code and data.
  - In conjunction with DP COMP, we will develop new principles for determining appropriate privacy policies as well as new standards for making meaningful judgements of utility.
Gaps between research & reality

Choose a privacy policy

Formulate a task

Preprocessing

Model Selection & Iteration

Data Owner

Analyst

Private Input

Clean

Transform

Apply $A_k$

Assessment and selection

Iteration

Task $T$

Output $o$ ± error

Choose algorithms for task

$A_1, A_2, \ldots, A_k$

Charles River Workshop: Privacy & Social Networks, 5/19/2013
This talk ...

- For a practitioner who wants to use DP:
  - *Tips & caveats on how to effectively utilize the wealth of literature.*

- For a DP researcher:
  - *Identify open questions that help bridge the gap*
Outline

• Real world applications
  – Synthetic Data Generation w/ US Census (Relational Data)
  – Human Mobility Traces w/ AT&T, Duke Medicine (Streaming/Spatial Data)
  – Private recommendations on graphs (Social Networks)

• Differential Privacy in the Wild
  – Formulating a task
  – Choosing a privacy policy
  – Preprocessing
  – Model selection & Iteration
Outline

• Real world applications
  – Synthetic Data Generation of Census data
  – Human Mobility Traces
  – Private recommendations on graphs

• Differential Privacy in the Wild
  – Formulating a task
  – Choosing a privacy policy
  – Preprocessing
  – Model selection & Iteration
Longitudinal Employer-Household Dynamics (LEHD)

“Release public use data by combining federal, state, and Census Bureau data on employers and employees ...”
Schema

- Worker
  - Age
  - Sex
  - Race & Ethnicity
  - Education
  - Home location (Census block)

- Workplace
  - Geography (Census blocks)
  - Industry
  - Ownership (Public vs Private)

- Job
  - Start date
  - End date
  - Worker & Workplace IDs
  - Earnings
Goal: Release Synthetic Data

• Sample from a model built using a set of lower order marginals

• Measures:
  – Average Employment
  – CDF/quantiles over Earnings

• Stratifying variables:
  – Age, Sex, Race & Ethnicity, Education, Home location
  – Work location, Industry, Ownership
Application: OnTheMap

http://onthemap.ces.census.gov/
Application: QWI

• To compute Quarterly Workforce Indicators
  – Total employment
  – Average Earnings
  – New Hires & Separations
  – Unemployment Statistics

E.g., Missouri state used this data to formulate a method allowing **QWI to suggest industrial sectors where transitional training might be most effective** ... to proactively reduce time spent on unemployment insurance ...
What is Sensitive?

- **PII:**
  - Name, SSN, DoB, Biometrics
  - Educational & Medical Records, Financial transactions
  - Criminal or Employment history

- **BII:**
  - Business name, address, industry (NAICS)
  - Payroll, assets, sales, financial data
State of the Art

• < 2008
  – Ad hoc protection measures used to add noise to the marginals before building a model

• since 2008
  – Workplace characteristics protected using ad hoc perturbation schemes
  – Worker characteristics (at one time snapshot) protected using algorithms that provably satisfy (probabilistic) differential privacy

  • Only a subset of the characteristics were used to build the model to regulate for sparsity

[M et al ICDE 2008]
Outline

• Real world applications
  – Synthetic Data Generation
  – Human Mobility Traces
  – Private recommendations on graphs

• Differential Privacy in the Wild
  – Formulating a task
  – Choosing a privacy policy
  – Preprocessing
  – Model selection & Iteration
Problem

• Identify smoking hotspots
• Identify environmental determinants of smoking
• Predict whether a person is likely to smoke based on their current location
• ...

• ... but these analyses should preserve differential privacy.
Differentially private synthesis of mobility traces

• Compute differentially private counts of short sequences (or k-grams)

A tree of counts of 1-grams, 2-grams, ... K-grams.

Differentially private counts (Laplace noise)

Pruned trees with consistent, non-negative counts
Differentially private synthesis of mobility traces

- Compute differentially private counts of short sequences (or k-grams)

- Fit a semi-Markov model using noisy counts

\[
\Pr[e_{n+1} = j \mid e_1, \ldots, e_n] = \frac{\hat{c}(e_{n+1}, \ldots, e_{n-k})}{\hat{c}(e_n, \ldots, e_{n-k})}
\]
Differentially private synthesis of mobility traces

- Compute differentially private counts of short sequences (or k-grams)

- Fit a semi-markov model using noisy counts

\[
\Pr[e_{n+1} = j \mid e_1, \ldots, e_n] = \Pr[e_{n+1} = j \mid e_{n-k}, \ldots, e_n] = \frac{\hat{c}(e_{n+1}, \ldots, e_{n-k})}{\hat{c}(e_n, \ldots, e_{n-k})}
\]

- Sample synthetic trajectories from the noisy semi-markov model.
Outline

• Real world applications
  – Synthetic Data Generation
  – Human Mobility Traces
  – Private recommendations on graphs

• Differential Privacy in the Wild
  – Formulating a task
  – Choosing a privacy policy
  – Preprocessing
  – Model selection & Iteration
Personalized Social Recommendations

Recommend keywords/ads/competitors based on private bids of other companies.
Social Recommendations... privacy problem

Only the items (products/people) liked by Alice’s friends are recommendations for Alice

Fact that “Betty” liked “VistaPrint” is leaked to “Alice”
Social Recommenders

Utility Function – $u(a, i)$
utility of recommending candidate $i$ to target $a$

2-hop neighborhood
• Common Neighbors
• Adamic/Adar

Holistic
• Katz (weighted paths)
• Personalized PageRank
Social Recommenders

Exponential Mechanism:
Randomly pick a candidate with probability proportional to $\exp(\varepsilon \cdot u(a,i) / \Delta)$
($\Delta$ is maximum change in utilities by changing one edge)

Utility Function – $u(a, i)$
utility of recommending candidate $i$ to target $a$
Negative Result [M et al VLDB 2009]

- Theorem: Under edge privacy, in order to achieve $\Omega(1)$ accuracy for a single recommendation, for Common Neighbors, Adamic/Adar and Katz utility functions ...

$$\varepsilon > \Omega\left(\frac{\log n}{\text{degree}(a)}\right)$$

95% of users have accuracy < 5%
Outline

• Real world applications
  – Synthetic Data Generation
  – Human Mobility Traces

• Differential Privacy in the Wild
  – Formulating a task
  – Choosing a privacy policy
  – Preprocessing
  – Model selection & Iteration
Formulating a Task

• *SN Practitioner:* Try to formulate a task as:

  – A workload of queries $W$ on the data
  – An error metric
    
    *(distance between DP answer and true answer)*
Example from Census

• Let R be a set of contiguous census blocks.

Query 1: Average earnings of individuals working in R cross-tabulated by all the stratifying variables.

Query 2: Average employment of businesses in R cross-tabulated by all the stratifying variables.

Query 3: Histogram of residences of individuals working in R cross-tabulated by all the stratifying variables.

Error Measure: support weighted root mean squared error
Why formulate a task?

• Everyone knows the Laplace mechanism ...
  ... but it can have high sensitivity (and thus high error)

• Workload W can be answered with lower error by choosing to answer a different strategy workload A
  – Each query in W can be answered using a small number of queries in A
  – A has low sensitivity

• Recent work:
  Algorithms for finding A given linear workloads W
  – (K-Norm mechanism [Hardt-Talwar], Matrix Mechanism [Li et al])

Much work on identifying strategies for specific workloads.
Workloads & Strategies

• **SN Practitioner:**
  What are interesting workloads for social network analysis?

• **DP Researcher:**
  Much of the theoretical work in differential privacy focuses on asymptotic bounds (for sufficiently large data)
  – Number of tuples usually much larger than size of domain.

• **Question:**
  What are mechanisms with optimal error in sparse finite datasets?
Outline

• Real world applications
  – Synthetic Data Generation
  – Human Mobility Traces

• Differential Privacy in the Wild
  – Formulating a task
  – Choosing a privacy policy
  – Preprocessing
  – Model selection & Iteration
Considerations for Privacy Policy

• What is epsilon?

• What are neighboring databases?

• Are there any constraints or correlations (known to adversary)?
Some values of epsilon don’t make sense.

• **SN Practitioner:**
  Beware of non-private epsilon:
  – Histogram release with Laplace mechanism
  – With sufficiently large epsilon, significant number of counts will not change with high probability

• **DP Researcher:**
  Beware of “useless” epsilon:
  – For sufficiently small epsilon, most differentially private mechanisms may have higher error than “useless” algorithms
Considerations for Privacy Policy

• What is epsilon?

• What are neighboring databases?

• Are there any constraints or correlations (known to adversary)?
Choosing Neighboring Databases

• Mobility Traces:

  User: Add or remove all trajectories of one user
  Event: Change one location of a user’s trajectory
  Window: Change w consecutive locations of a user’s trajectory

• Choice should depend on what is secret.

  Event & Window: Disclose the home location of an individual.
Beyond just adding or removing a row ...

• Adding or removing one row in Census data leads to high sensitivity
  – Average earnings for an individual
  – Some individuals earn in billions
  – Removing outliers (billionaires) is not the answer

• What should be secret?
  – Precise estimate of earnings
  – NOT billions vs thousands
Blowfish: Formalizing neighbors

- **Secret**: Boolean predicate over the domain
  - Bob’s home location is Boston
  - Bob was in Boston on May 19
  - Bob earns $100,000

- **Discriminative pairs**: Pairs of mutually exclusive secrets that adversary should not distinguish between
  - Bob’s home is Boston vs Bob’s home is Durham
  - Bob was in Boston on May 19 vs Bob was in NYC on May 19
  - Bob earns $100,000 vs Bob earns $90,000

[He et al SIGMOD 14]
Discriminative Secret Graph

• $G = (V, E)$
  – Nodes: values in the domain
  – Edges: $(s_1, s_2)$ is a discriminative secret
    v_1 satisfies secret s_1
    v_2 satisfies secret s_2,
    then $(v_1, v_2)$ is an edge in G.

• G-Neighbors:
  Databases that differ in one row, and the row takes values $v_1$ and $v_2$ in the databases, where $(v_1, v_2)$ is an edge in G.
Examples

• Secrets: \(\{s_x: \text{Bob earns } x \mid x \text{ is a natural number}\}\)
  Discriminative Secret Graph:
  \[E = \{(s_x, s_y) \mid x/c < y < xc\}\]

*Intuition: Can’t tell earnings within a multiplicative factor \(c\).*

Mobility: Neighbors by Events

• Secrets: \(\{s_{i,x}: \text{Bob’s location is } x \text{ at time } i\}\)
  Discriminative Secret Graph:
  \[E = \{(x, y) \mid \text{trajectories } x \text{ and } y \text{ are all same except for one location}\}\]
Blowfish

[Haney et al 2014]

• Answering a set of queries $W$ under Blowfish (with discriminative secret graph $G$)

is equivalent (in terms of error*) to

Answering a transformed set of queries $W_G = f(W, G)$ under differential privacy.

* Under the Matrix Mechanism framework
Considerations for Privacy Policy

- What is epsilon?

- What are neighboring databases?

- Are there any constraints or correlations (known to adversary)?
Constraints & Correlations

• Certain constraints/correlation in the data may be publicly known
  – Exact marginal counts released by other agencies
  – Constraints on mobility (e.g., speed limits)
  – Homophily in social networks

• Participation of an individual in the data is not hidden by differential privacy in non-iid data.
Correlations in social networks

- Want to release the number of edges between blue and green communities.
- Should not disclose the presence/absence of Bob-Alice edge.
Adversary knows how social networks evolve

- Depending on the social network evolution model, 
  \((d_2 - d_1)\) is linear or even super-linear in the size of the network.
Differential privacy fails to avoid breach

Output \( (d_1 + \delta) \)

\( \delta \sim \text{Laplace}(1/\varepsilon) \)

Output \( (d_2 + \delta) \)

Adversary can distinguish between the two worlds if \( d_2 - d_1 \) is large.
Privacy in non-iid data

• An area of active privacy

• Include constraints on adversary’s (non-iid) prior about the data

• Counterfactual approach:
  Pufferfish [Kifer-M PODS 2012]
  Noiseless privacy [Bhaskar et al 2011]

• Simulation based approach:
  Coupled Worlds Privacy [Raef et al FOCS 2013]
Outline

• Real world applications
  – Synthetic Data Generation
  – Human Mobility Traces

• Differential Privacy in the Wild
  – Formulating a task
  – Choosing a privacy policy
  – Preprocessing
  – Model selection & Iteration
Preprocessing is essential for messy data

IP Aliasing Problem
[Willinger et al. 2009]

Figure 2. The IP alias resolution problem. Paraphrasing Fig. 4 of [50], traceroute does not list routers (boxes) along paths but IP addresses of input interfaces (circles), and alias resolution refers to the correct mapping of interfaces to routers to reveal the actual topology. In the case where interfaces 1 and 2 are aliases, (b) depicts the actual topology while (a) yields an “inflated” topology with more routers and links than the real one.

Figure 3. The IP alias resolution problem in practice. This is reproduced from [48] and shows a comparison between the Abilene/Internet2 topology inferred by Rocketfuel (left) and the actual topology (top right). Rectangles represent routers with interior ovals denoting interfaces. The histograms of the corresponding node degrees are shown in the bottom right plot. © 2008 ACM.
Preprocessing is essential for messy data

- Multiple Imputation for missing values  [Rubin 1987]
  - Build a model for missing values (based on existing ones)
  - Impute missing value by sampling from the model
  - Construct multiple imputations (by sampling many times)
  - Run analysis on all imputations
  - Combining formulae quantify error due to imputation

```plaintext
Incomplete data matrix
(missing data = #)

A    B    C    D    E
1    3    #    2    #
4    #    5    #    1
7    7    3    9    1
#    #    1    #    3

Imputed data matrix
(1) Sum the imputed values

Imputed data matrix
(2) Generate multiple imputed datasets from the imputation model, using data augmentation.

Imputed data matrix
(3) Fit multiple regression, or other model, to each imputed dataset. Save the parameter values and standard errors.

Imputed data matrix
(4) Calculate mean parameter estimates, and total standard errors that combine within-imputation and between-imputation parameter uncertainty, to calculate proper hypothesis tests.

Charles River Workshop: Privacy & Social Networks, 5/19/2013
Preprocessing costs privacy

- *SN Practitioner:* Unless the preprocessing step looks at each “row” independently, this step costs privacy budget.

- Example of free preprocessing: Picking a subset of $k$ locations visited by a user’s trajectory.

- Example of costly preprocessing: Ignoring all census blocks in the US with count $= 0$
Preprocessing

- DP Researcher: Most interesting preprocessing steps look at more than one row.

Need DP algorithms for effective preprocessing.
Outline

• Real world applications
  – Synthetic Data Generation
  – Human Mobility Traces

• Differential Privacy in the Wild
  – Formulating a task
  – Choosing a privacy policy
  – Preprocessing
  – Model selection & Iteration
Model Selection & Iteration

- **Practitioner:**
  Building $K$ models on the same data (each with epsilon) and choosing the best does not imply privacy budget is $K \times \epsilon$.

- **Sparse Vector Technique:** [Hardt 2011, Roth] Can test whether an unbounded number of queries have answers larger or smaller than a threshold, and ensure $2 \epsilon$ DP.
  - Use epsilon to perturb the threshold
  - Use epsilon to answer the query and compare with noisy threshold.
Model Selection & Iteration

• *DP Researcher*: Iteration is not as well understood.

Under what conditions do iterative algorithms not consume privacy budget proportional to number of iterations?
Summary

• Vast literature on differential privacy
  – Theoretical upper and lower bounds
  – Sophisticated algorithms

• But very few real world applications of differential privacy
  – Gaps between real applications and idealized differential privacy workflow.

• Recommendations:
  *SN Practitioners*: Think like a *DP researcher*
  *DP Researcher*: Think like a *SN Practitioner*
Thank you 😊