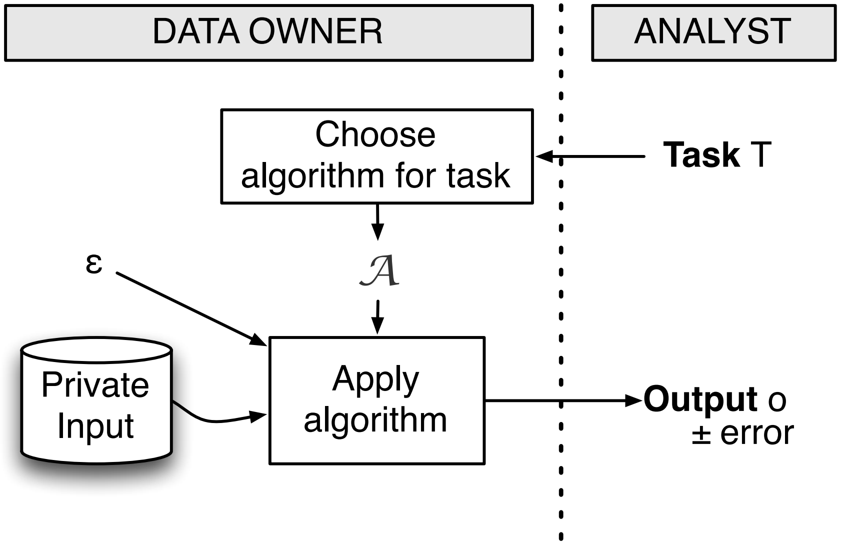
Differential Privacy in the Wild: Challenges and Open Questions

Ashwin Machanavajjhala

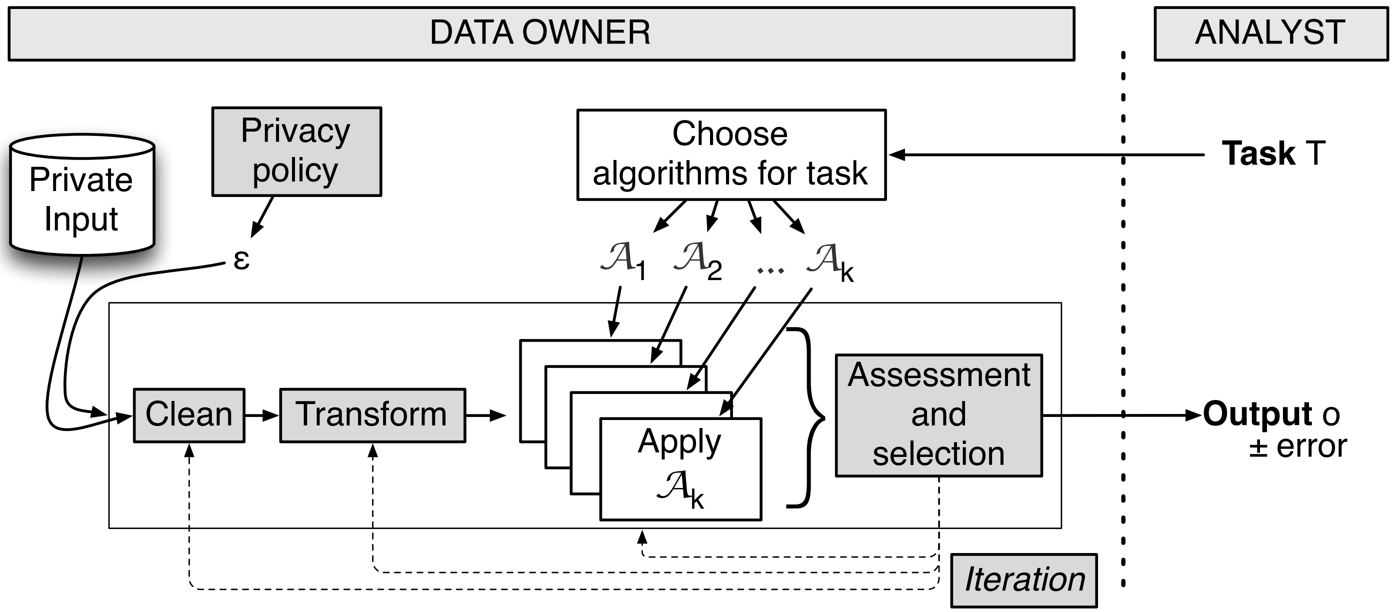
ashwin @ cs.duke.edu

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

World according to DP research

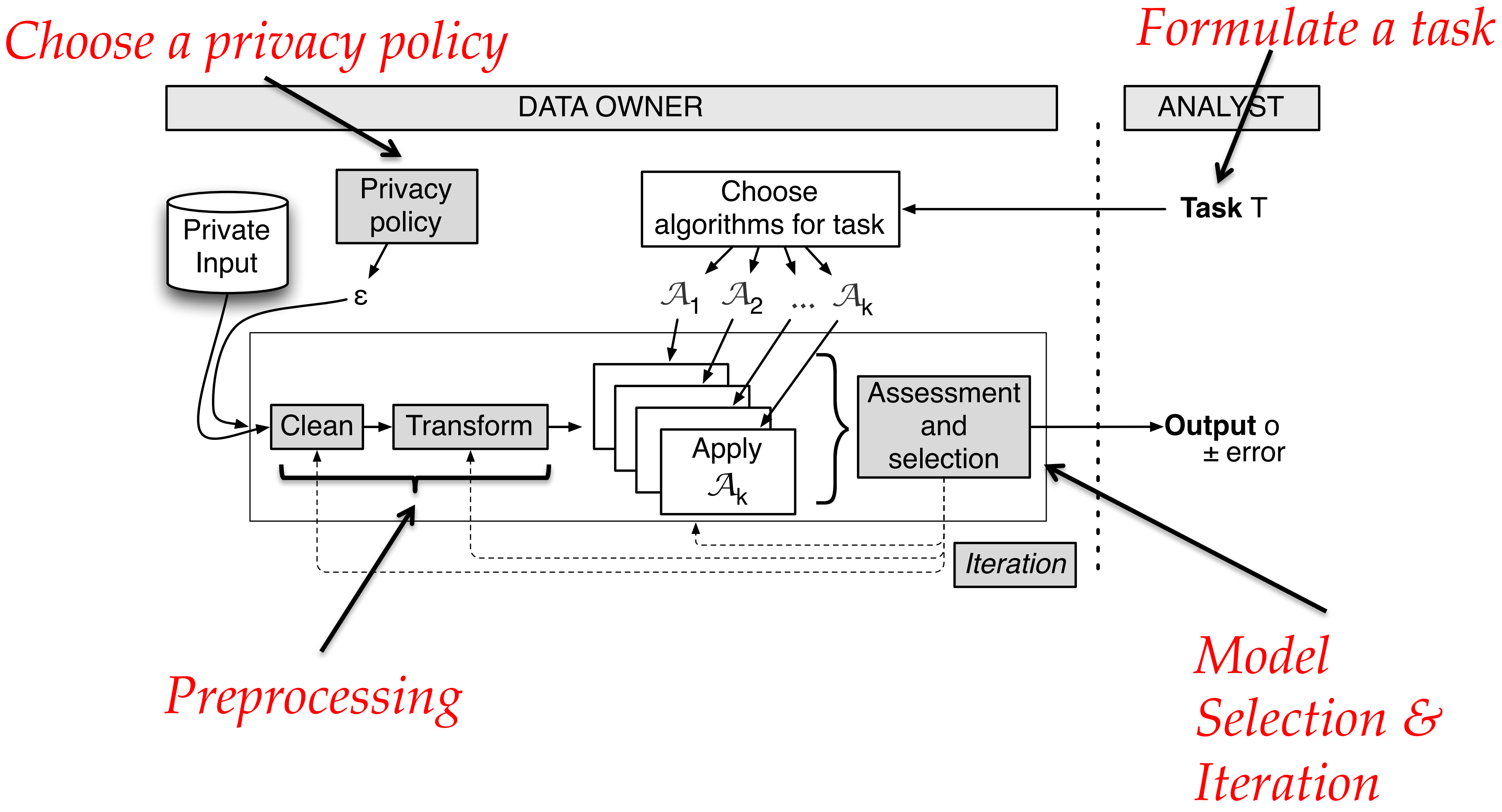


Real World



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

Gaps between research & reality



Choose a privacy policy

Formulate a task

DATA OWNER

ANALYST

Private Input

Privacy policy

ε

Choose algorithms for task

Task T

$\mathcal{A}_1$ $\mathcal{A}_2$ ... $\mathcal{A}_k$

Clean

Transform

Apply $\mathcal{A}_k$

Assessment and selection

Output o ± error

Iteration

Preprocessing

Model Selection & Iteration

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

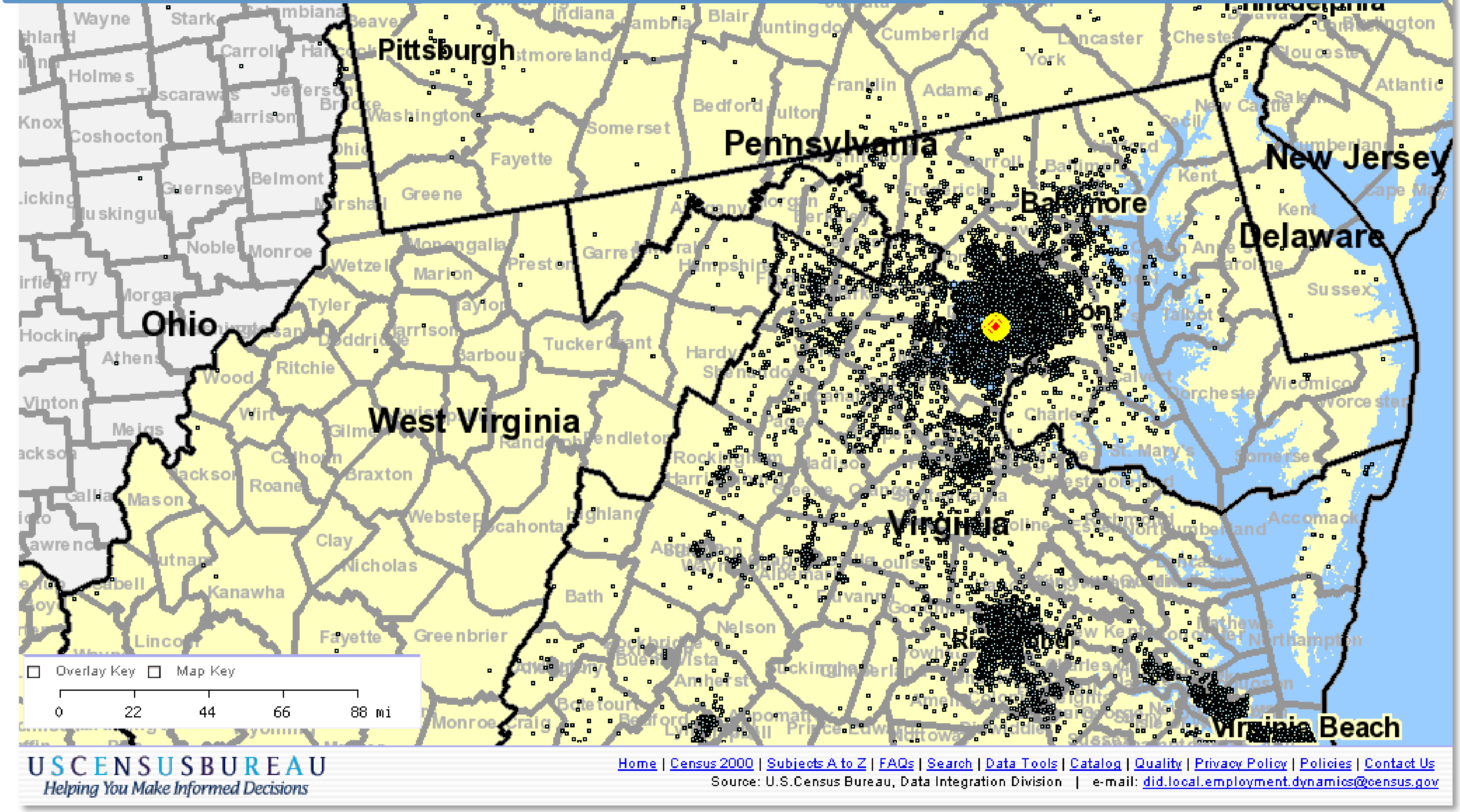
- Sample from a model built using a set of lower order marginals

- 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/



Source: U.S.Census Bureau, Data Integration Division | e-mail: did.local.employment.dynamics@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

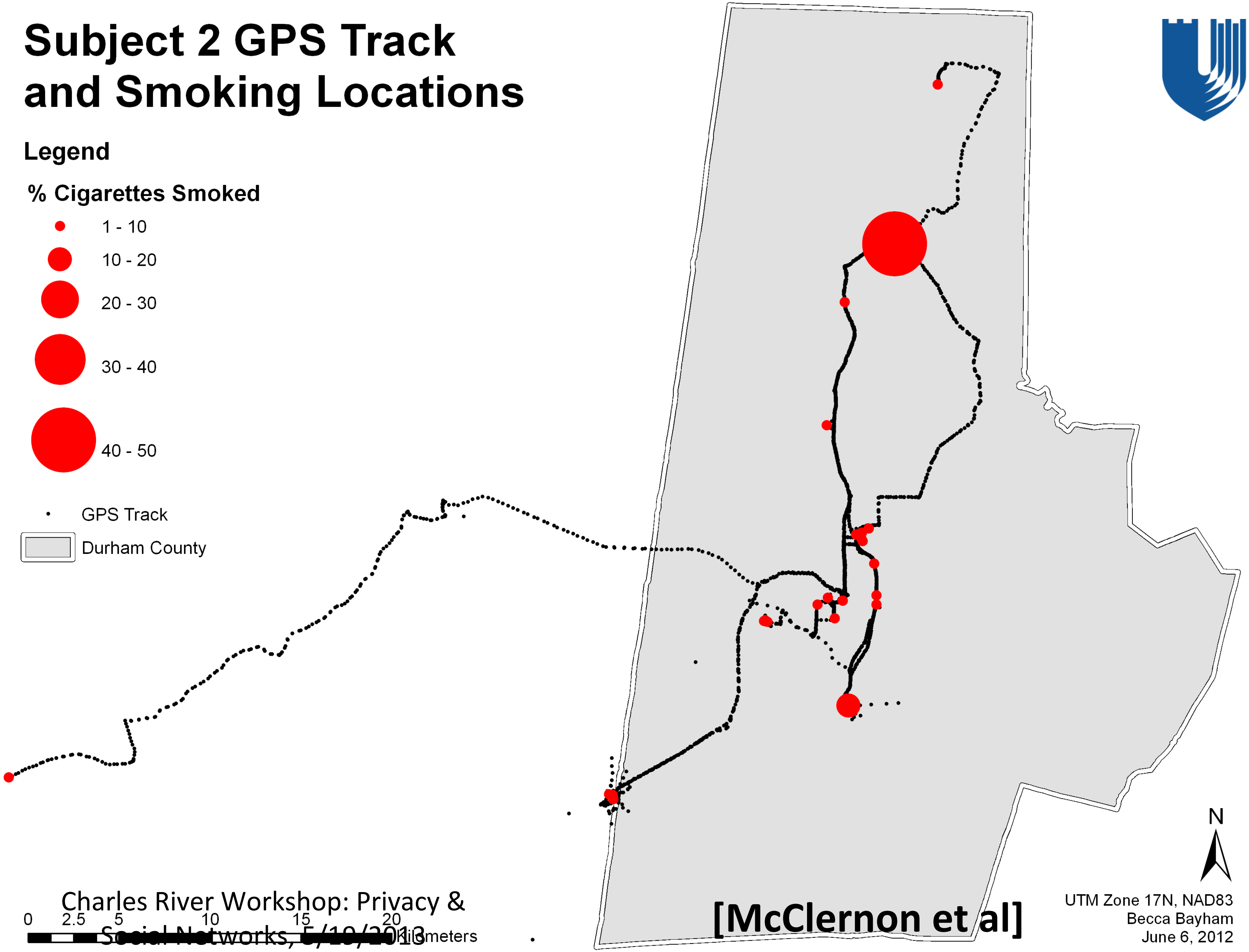
- since 2008 [M et al ICDE 2008]
  - Workplace characteristics protected using ad hoc perturbation schemes
  - Worker characteristics (at one time snapshot) protected using algorithms that provably satisfy (probabilistic) differential privacy
    - Only a subset of the characteristics were used to build the model to regulate for sparsity

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

Subject 2 GPS Track and Smoking Locations



UTM Zone 17N, NAD83
Becca Bayham
June 6, 2012

[McClernon et al]

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)

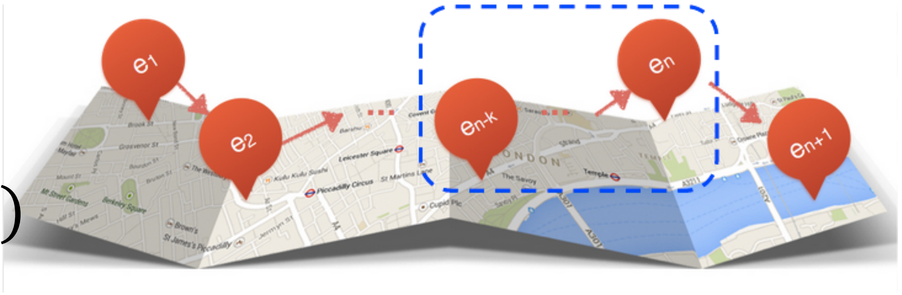
Pruned trees with consistent, non-negative counts

Differentially private synthesis of mobility traces

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

- Fit a semi-markov model using noisy counts

$$\begin{aligned}\Pr[e_{n+1}=j \mid e_1,\ldots,e_n] &= \Pr[e_{n+1}=j \mid e_{n-k},\ldots,e_n] \\ &= \hat{c}(e_{n+1},\ldots,e_{n-k}) \,/\, \hat{c}(e_n,\ldots,e_{n-k})\end{aligned}$$

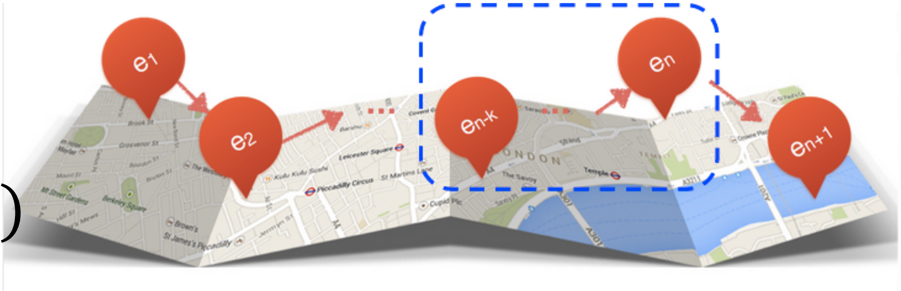


Differentially private synthesis of mobility traces

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

- Fit a semi-markov model using noisy counts

$$
\begin{aligned}
&\Pr[e_{n+1}=j \mid e_1,\ldots,e_n] \\
&\quad = \Pr[e_{n+1} = j \mid e_{n-k},\ldots,e_n] \\
&\quad = \hat{c}(e_{n+1}, \ldots, e_{n-k}) \,/\, \hat{c}(e_n, \ldots, e_{n-k})
\end{aligned}
$$



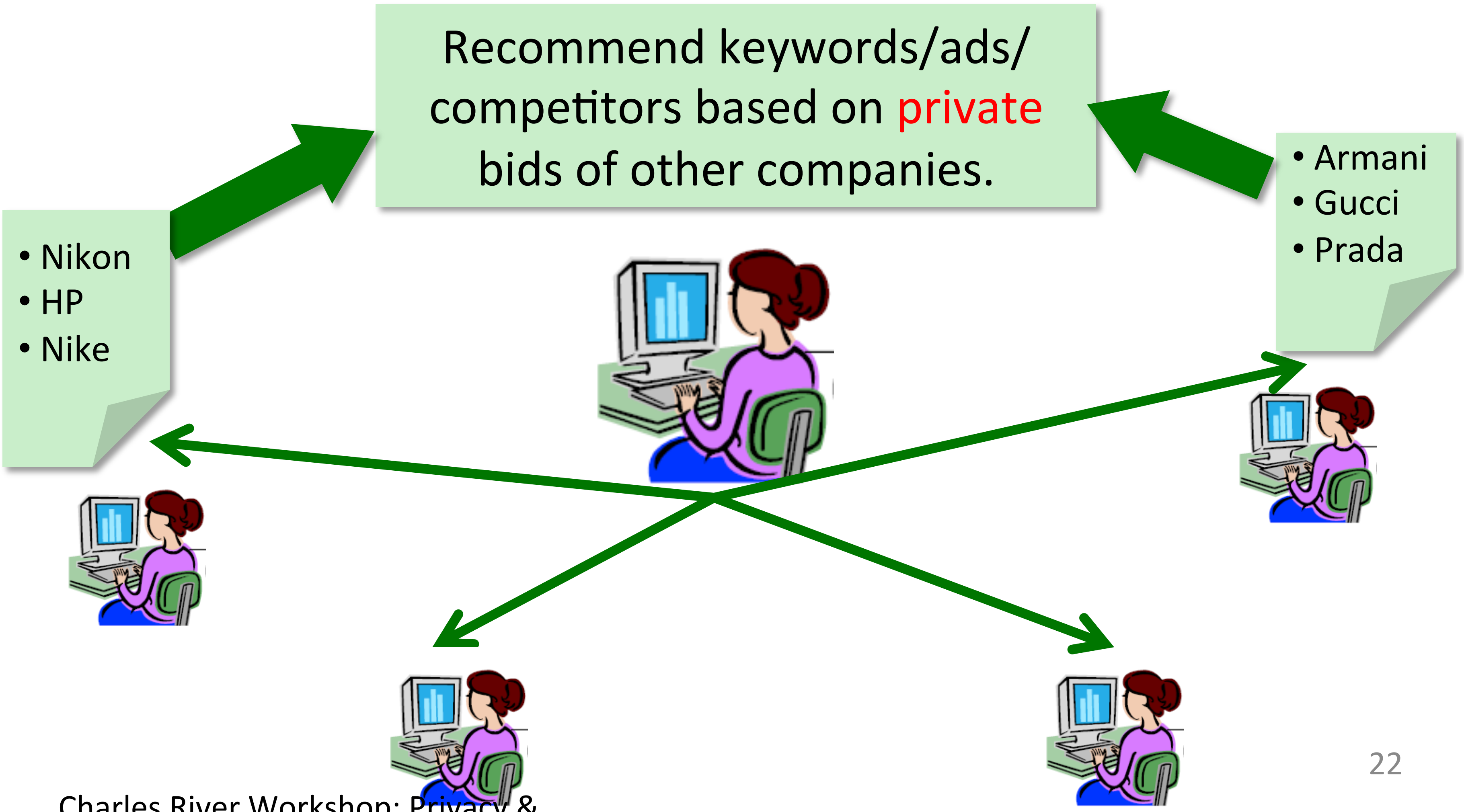
- Sample synthetic trajectories from the *noisy* semi-markov model.

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.

- Nikon
- HP
- Nike

- Armani
- Gucci
- Prada

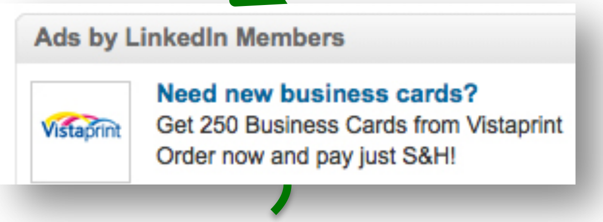
Social Recommendations… privacy problem

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

Alice

Betty

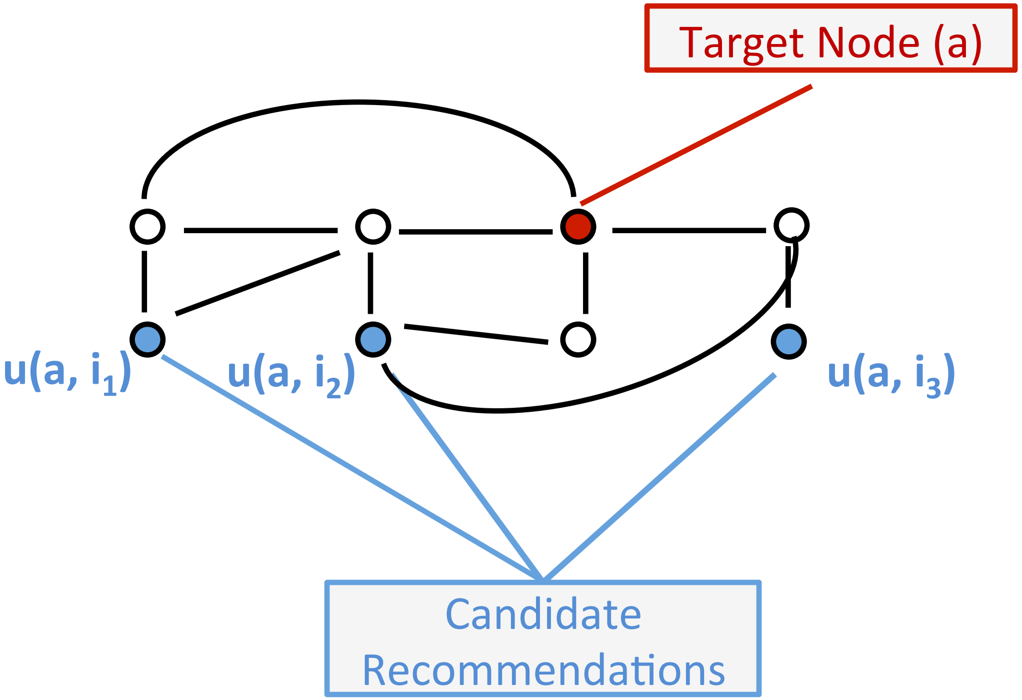
Fact that "Betty" liked "VistaPrint" is leaked to "Alice"



Ads by LinkedIn Members

Need new business cards?
Get 250 Business Cards from Vistaprint
Order now and pay just S&H!

Social Recommenders



Target Node (a)

$u(a, i_1)$

$u(a, i_2)$

$u(a, i_3)$

Candidate Recommendations

Utility Function – *u(a, i)*

utility of recommending candidate *i* to target *a*

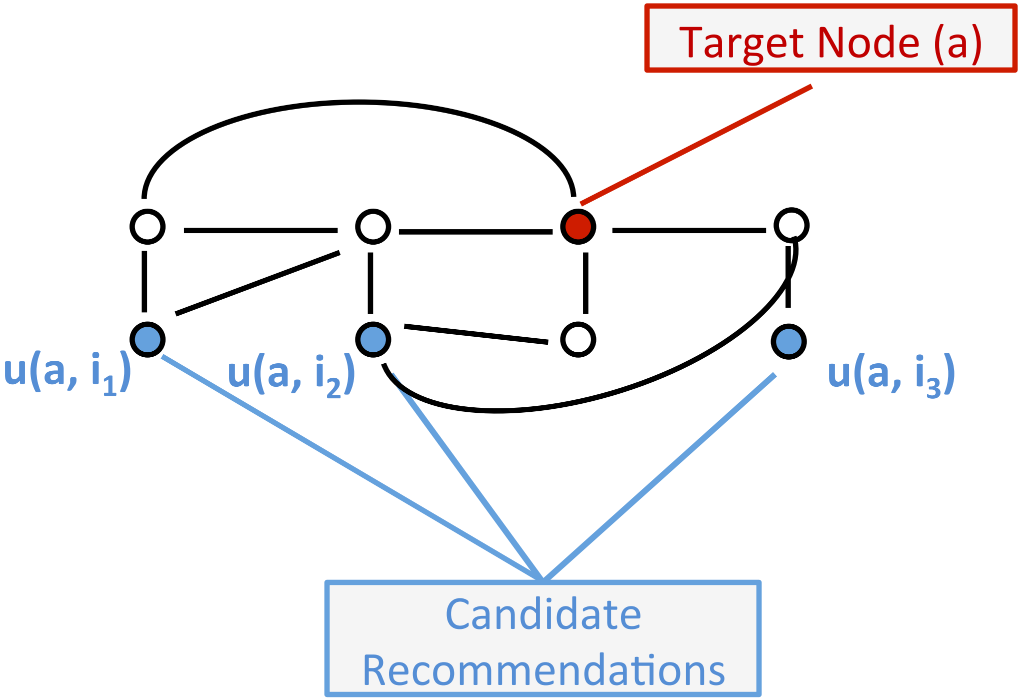
2-hop neighborhood

- Common Neighbors
- Adamic/Adar

Holistic

- Katz (weighted paths)
- Personalized PageRank

Social Recommenders



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

Exponential Mechanism:

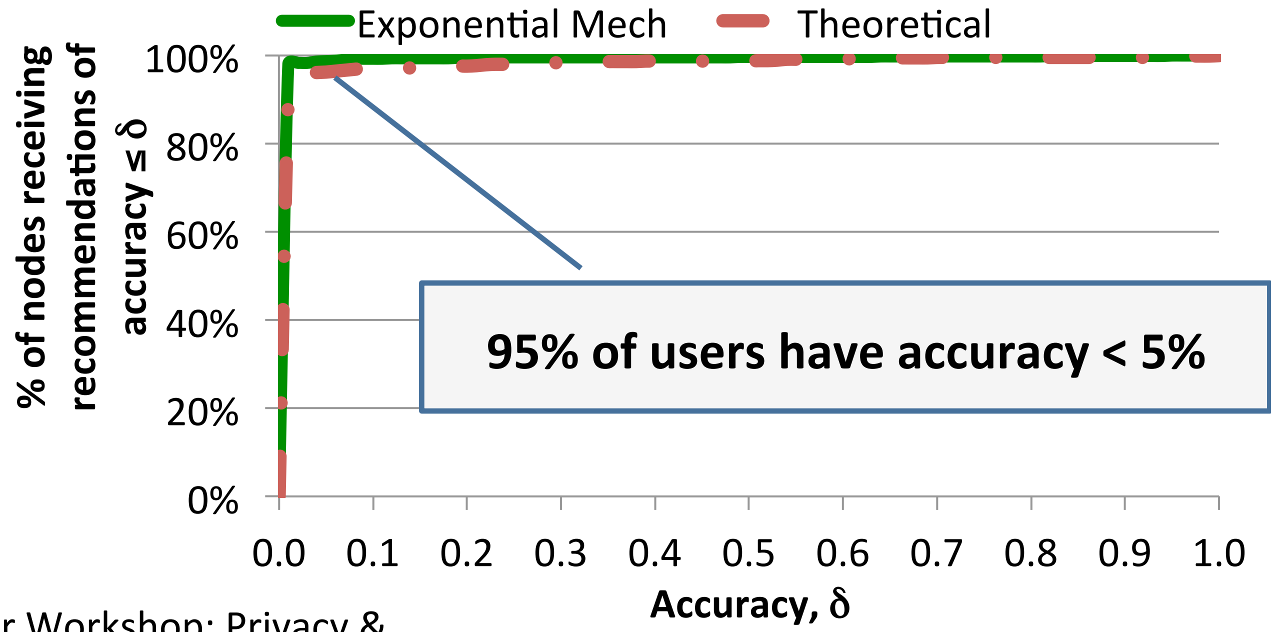
Randomly pick a candidate with probability proportional to **$\exp(\varepsilon \cdot u(a,i) / \Delta)$**

(Δ is maximum change in utilities by changing one edge)

Negative Result [M et al VLDB 2009]

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

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



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

[He et al SIGMOD 14]

- 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

Discriminative Secret Graph

- G = (V, E)
  - Nodes: values in the domain
  - Edges: (s1, s2) is a discriminative secret
    v1 satisfies secret s1
    v2 satisfies secret s2,
    then (v1, v2) is an edge in G.

- G-Neighbors:
  Databases that differ in one row, and the row takes values v1 and v2 in the databases, where (v1, v2) is an edge in G.

Examples

- Secrets: $\{s_x$: Bob earns x | x 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: $\{s_{i,x}$: Bob's location is x at time i$\}$
  Discriminative Secret Graph:

$$E = \{(\mathbf{x}, \mathbf{y}) \mid \text{trajectories x and y are all same except for one location}\}$$

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)?*

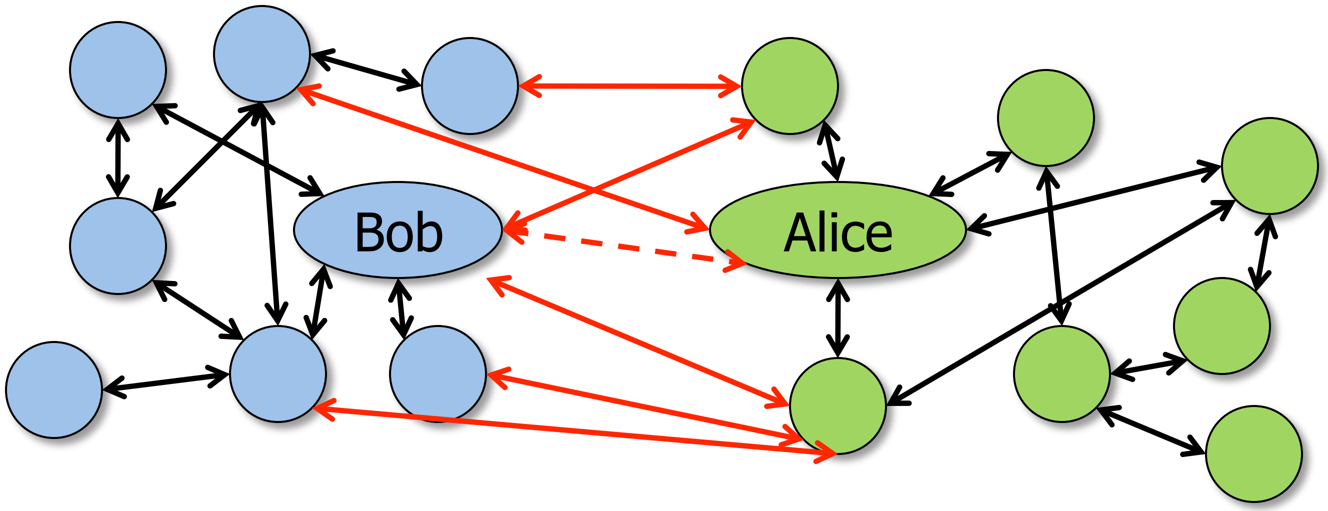
Constraints & Correlations

- 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.

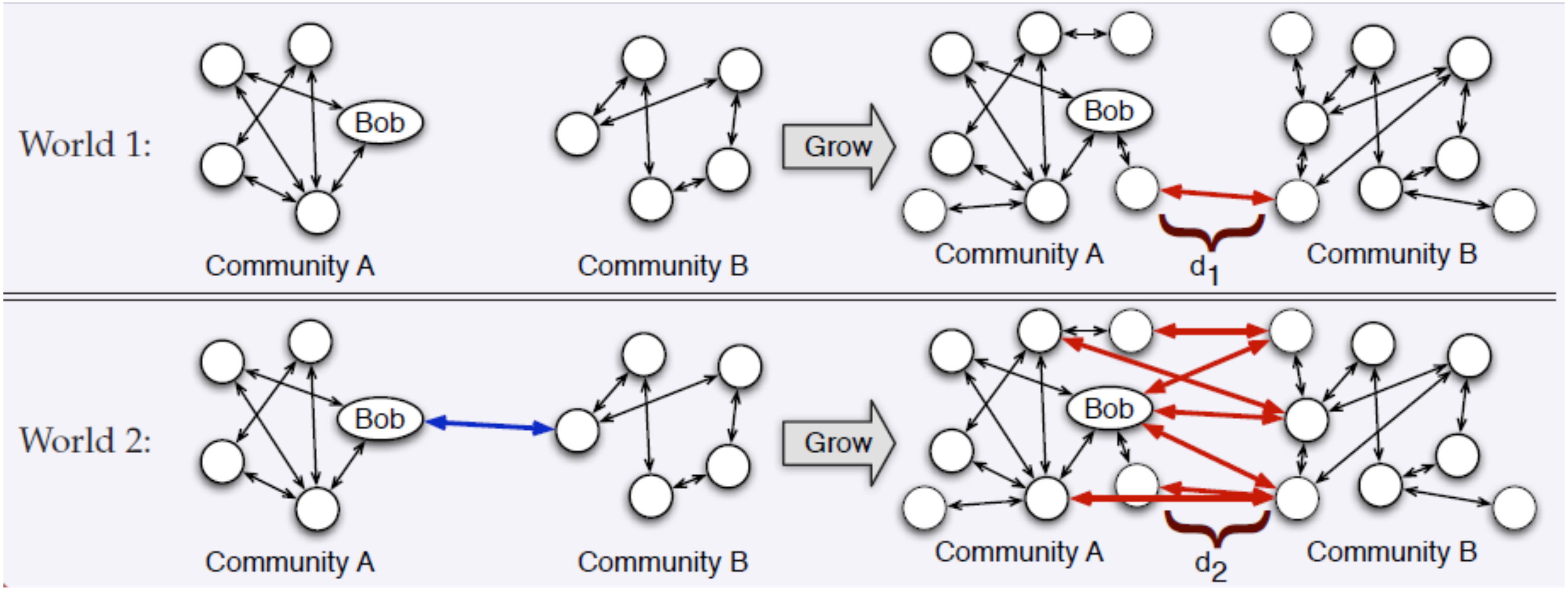
Correlations in social networks

[Kifer-M SIGMOD 2011]



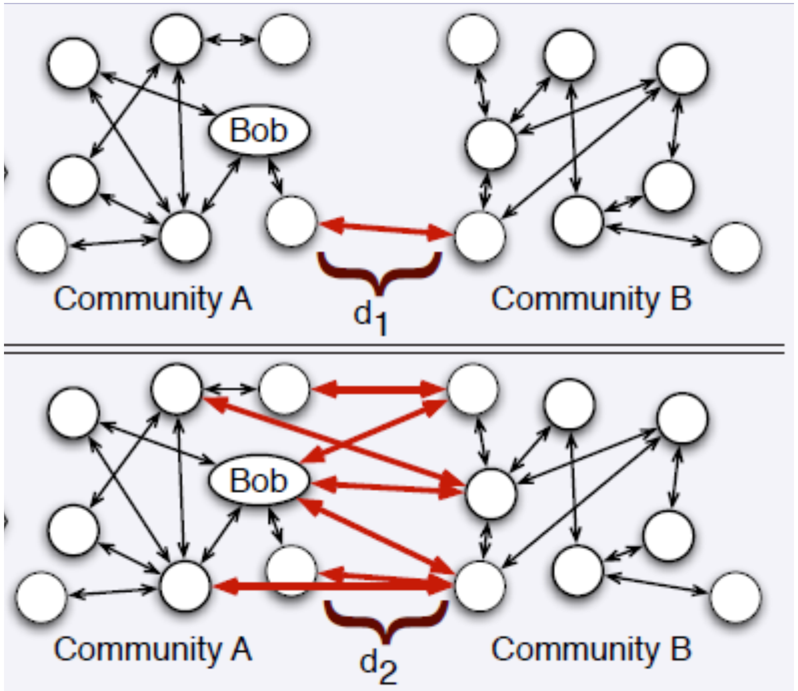
- Want to release the number of edges between **blue** and **green** communities.
- Should not disclose the presence/absence of Bob-Alice edge.

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]

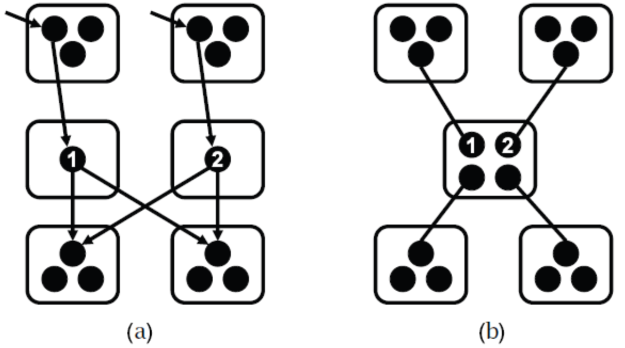


Figure 2. The IP alias resolution problem. Paraphrasing Fig. 4 of [50], traceroute does not list routers (boxes) along paths but IP addresses of input interfaces (circles), and alias resolution refers to the correct mapping of interfaces to routers to reveal the actual topology. In the case where interfaces 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.

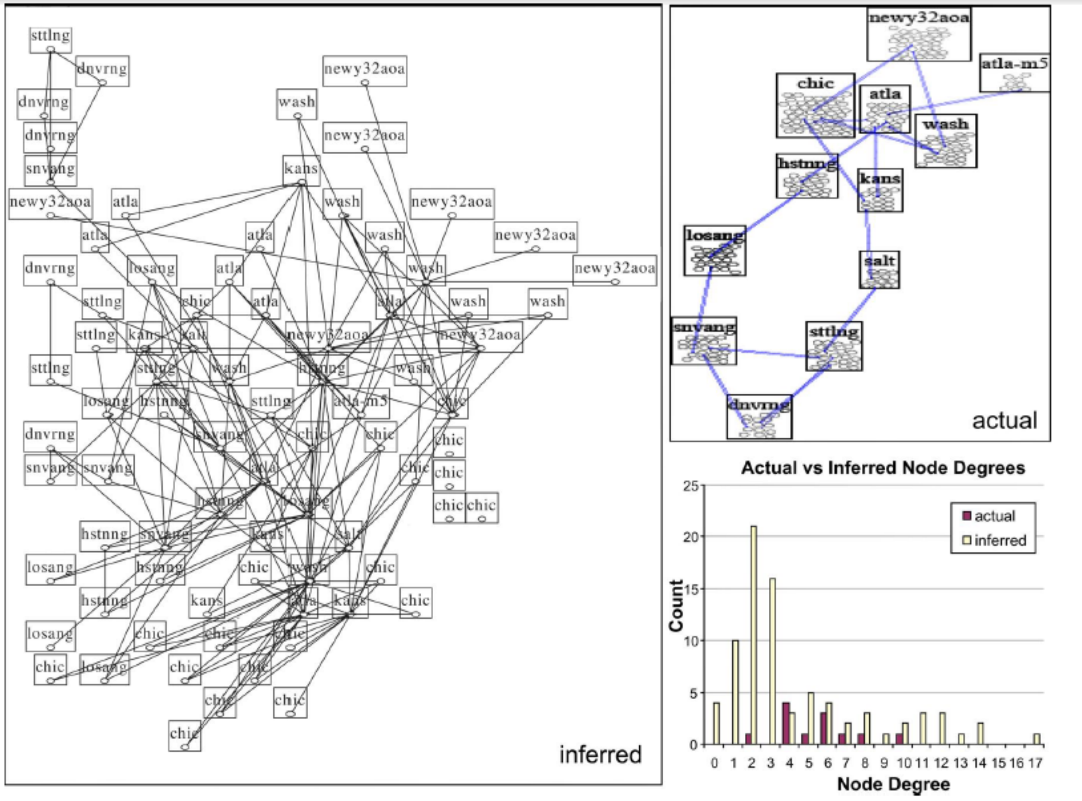
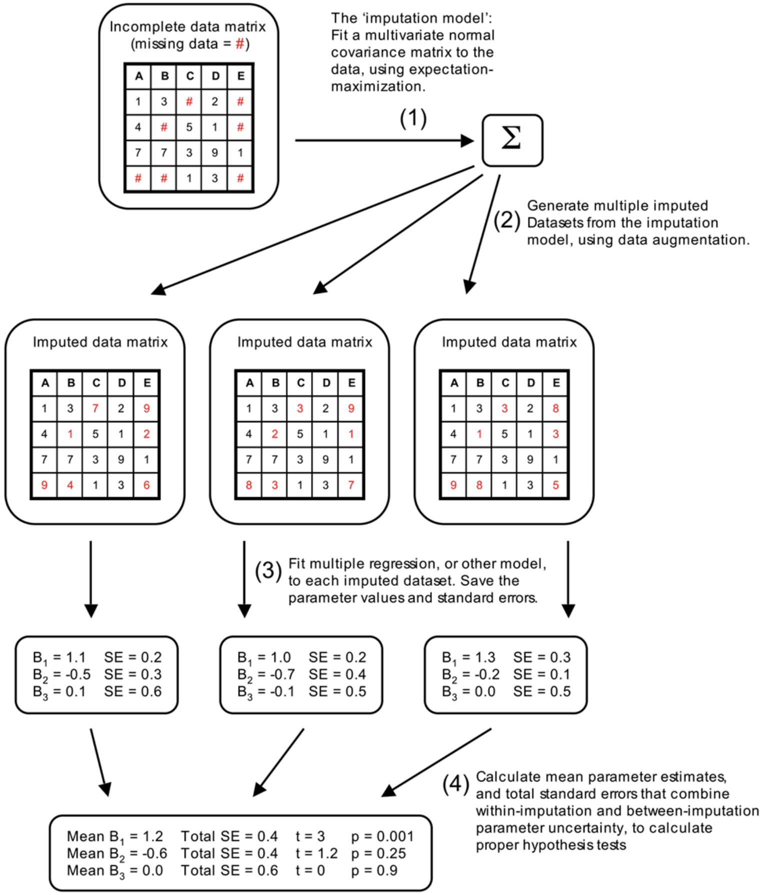


Figure 3. The IP alias resolution problem in practice. This is re-produced from [48] and shows a comparison between the Abilene/Internet2 topology inferred by Rocketfuel (left) and the actual topology (top right). Rectangles represent routers with interior ovals denoting interfaces. The histograms of 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



Incomplete data matrix (missing data = #)

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

The 'imputation model': Fit a multivariate normal covariance matrix to the data, using expectation-maximization.

(1)

(2) Generate multiple imputed Datasets from the imputation model, using data augmentation.

Imputed data matrix

A	B	C	D	E
1	3	7	2	9
4	1	5	1	2
7	7	3	9	1
9	4	1	3	6

Imputed data matrix

A	B	C	D	E
1	3	3	2	9
4	2	5	1	1
7	7	3	9	1
8	3	1	3	7

Imputed data matrix

A	B	C	D	E
1	3	3	2	8
4	1	5	1	3
7	7	3	9	1
9	8	1	3	5

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

B_1 = 1.1 SE = 0.2
B_2 = -0.5 SE = 0.3
B_3 = 0.1 SE = 0.6

B_1 = 1.0 SE = 0.2
B_2 = -0.7 SE = 0.4
B_3 = -0.1 SE = 0.5

B_1 = 1.3 SE = 0.3
B_2 = -0.2 SE = 0.1
B_3 = 0.0 SE = 0.5

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

Mean B_1 = 1.2 Total SE = 0.4 t = 3 p = 0.001
Mean B_2 = -0.6 Total SE = 0.4 t = 1.2 p = 0.25
Mean B_3 = 0.0 Total SE = 0.6 t = 0 p = 0.9

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 x 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 ☺

[Bhaskara 2011] Bhaskara et al, “Noiseless Database Privacy”, Asiacrypt 2011

[Hardt 2011] Hardt “A Study of Privacy and Fairness in Sensitive Data Analysis”, PhD Thesis 2011

[Haney 2014] Haney et al, “Answering Query Workloads with Optimal Error under Blowfish Privacy”, Manuscript

[Hardt-Talwar] Hardt et al, “On the Geometry of Differential Privacy”, STOC 2010

[He 2014] He et al, “Blowfish Privacy”, SIGMOD 2014

[Kifer 2011] Kifer et al, “No Free Lunch in Differential Privacy”, SIGMOD 2011

[Kifer 2012] Kifer et al, “A rigorous and customizable framework for privacy” PODS 2012

[Li 2010] Li et al, “Optimizing linear counting queries under differential privacy”, PODS 2010

[M 2008] Machanavajjhala et al, “Privacy: From theory to practice on the map”, ICDE 2008

[M 2009] Machanavajjhala et al “Personalized Social Recommendations: Accurate or Private”, VLDB 2009

[Raef 2013] Raef et al, “Coupled Worlds Privacy”, FOCS 2013

[Roth-notes] http://www.cis.upenn.edu/~aaroth/courses/slides/Lecture11.pdf

[Rubin 1987] Rubin, “Multiple Imputation for Nonresponse in Surveys”, Wiley & Sons 1987

[Willinger 2009] W. Willinger et al, “Mathematics and the Internet: A Source of Enormous Confusion and Great Potential”, Notices of the AMS 56(5), 2009