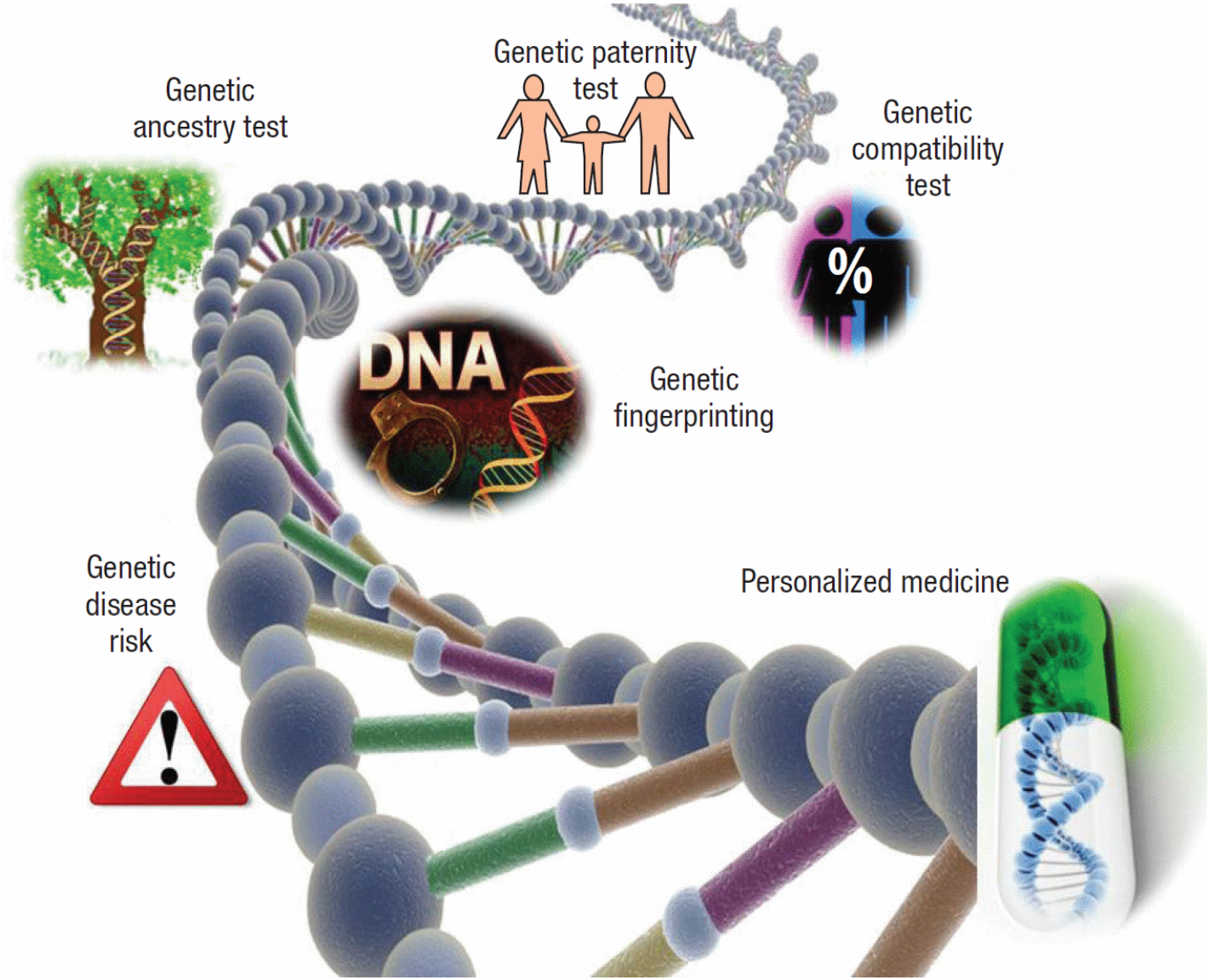
Classifying Large Graphs with Differential Privacy

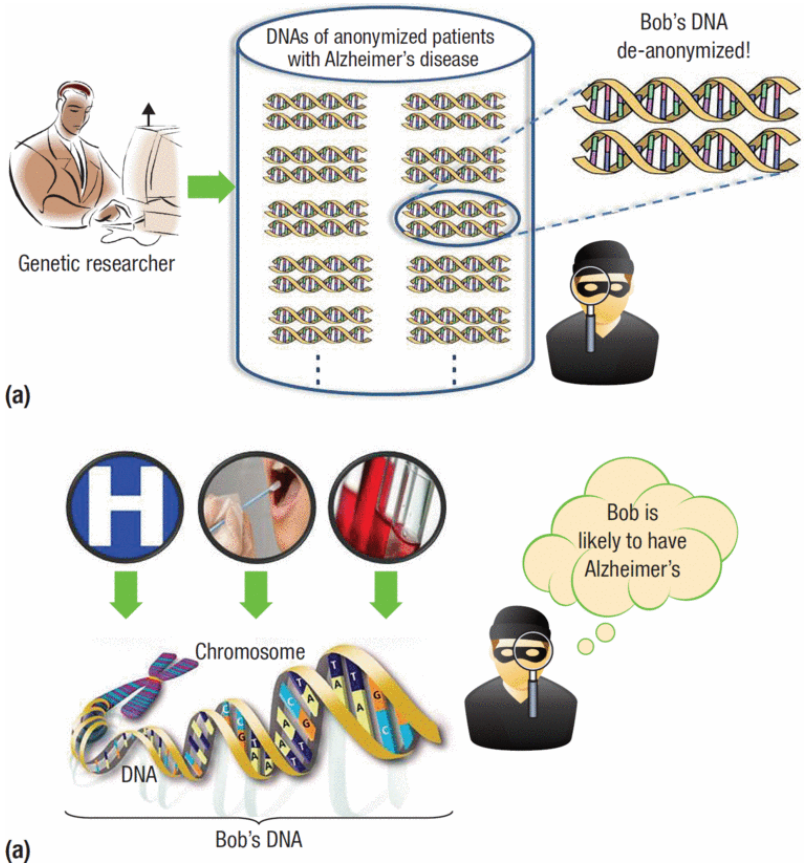
Devdatt Dubhashi
Joint work with Fredrik Johansson, Otto Frost, Carl Retzner

MDAI, Skövde Sept.22 2015

Sharing of Data in Genomics



Privacy Risks



E. Ayday, E. De Cristofaro, J-P. Hubaux, G. Tsudik: "Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare?" *IEEE Computer* 48(2): 58-66 (2015)

Simple Anonymization Fails!

Netflix Challenge

- Netflix released data for subsets of movies and users with users anonymized with random ids.
- Linkage attack: Deidentified by using public incomplete IMDB data: on average 4 movies uniquely identified a user.

Many other examples including genetic (GWAS) data and advertsing systems.

Differential Privacy: Motivation

How many of your online friends are dogs? (Google's RAPPOR)

"On the Internet, nobody knows you're a dog."

Differential Privacy: Randomization

Friend's answer

Flip a coin.

H Answer truthfully.

T Always say "yes".

- Get a good estimate of the true count from the greater-than-half fraction of your friends that answered "Yes".
- However, you still wouldn't know which of your friends was a dog: each answer "Yes" would most likely be due to that friend's coin flip coming up tails.

Differential Privacy

Differential Privacy

A randomized algorithm $\mathcal{A}$ is (ϵ, δ)–differentially private if for all events S in the output space of $\mathcal{A}$ and for all neighboring databases D, D' that differ only in a single individual,

$$\Pr\left(\mathcal{A}(D) \in S\right) \leq \exp(\epsilon) \times \Pr\left(\mathcal{A}(D') \in S\right) + \delta \ .$$

Intuitively, the output of the algorithm is statistically indistinguishable whether or not a particular individual is included in the database or not.

Diffenrential Privacy Promise

You will not be affected adversely by allowing your data to be used in any study of analysis, no matter what other studies, data sets or information sources are used, and no matter how powerful or malign an algorithm is used. Nothing is compromised about an individual while useful analysis is done on the population as a whole.

Global Sensitivity and Laplace Mechanism

Global Sensitivity

The global sensitivity S_f of a function $f : D \to \mathbb{R}$ is,

$$S_f = \max_{d(D,D')=1} |f(D) - f(D')|$$

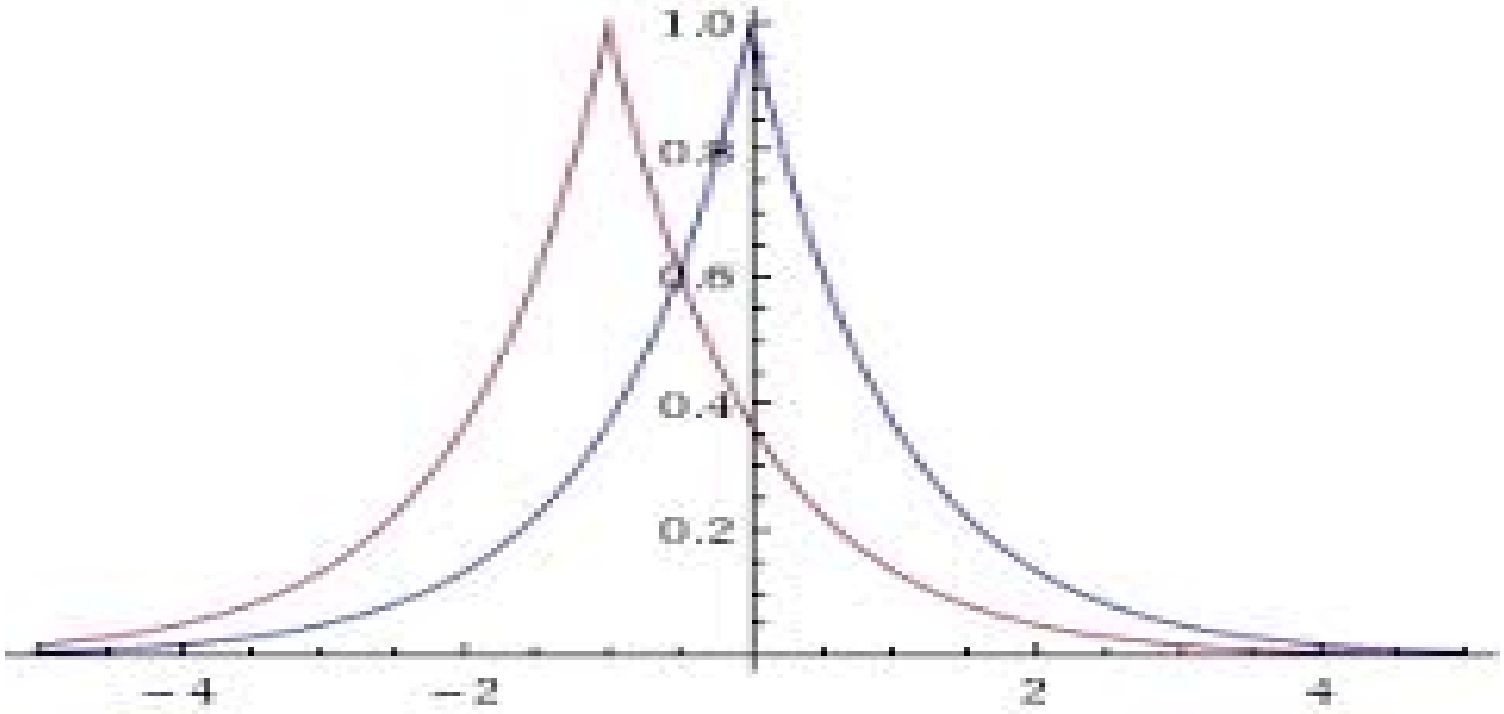
Laplace Mechanism

The mechanism

$$\mathcal{A}(f, D) = f(D) + \mathsf{Lap}\left(GS_f/\epsilon\right)$$

is $(\epsilon, 0)$–differentially private.

Laplace Distribution



$$\text{Lap}\,(b)\,(x) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$$

Counts

Summary statistics

Suppose we want to release a summary e.g. the fraction of diabetics in the database

$$f(D) := \frac{1}{|D|} \sum_{i \in D} \text{diabetic}(i)$$

Easy to see

$$GS_f = \frac{1}{|D|}$$

Hence release

$$A(D) = \text{proportion} \pm \frac{1}{\epsilon |D|}.$$

Properties of Differential Privacy

Post-processing invariance

The composition of a data independent mapping with an (ϵ, δ)–differentially private algorithm is also (ϵ, δ)–differentially private.

Sequential Composition

The sequential composition of a (ϵ_1, δ_1) and a (ϵ_2, δ_2)-differentially private algorithm satisfies $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$–differential privacy.

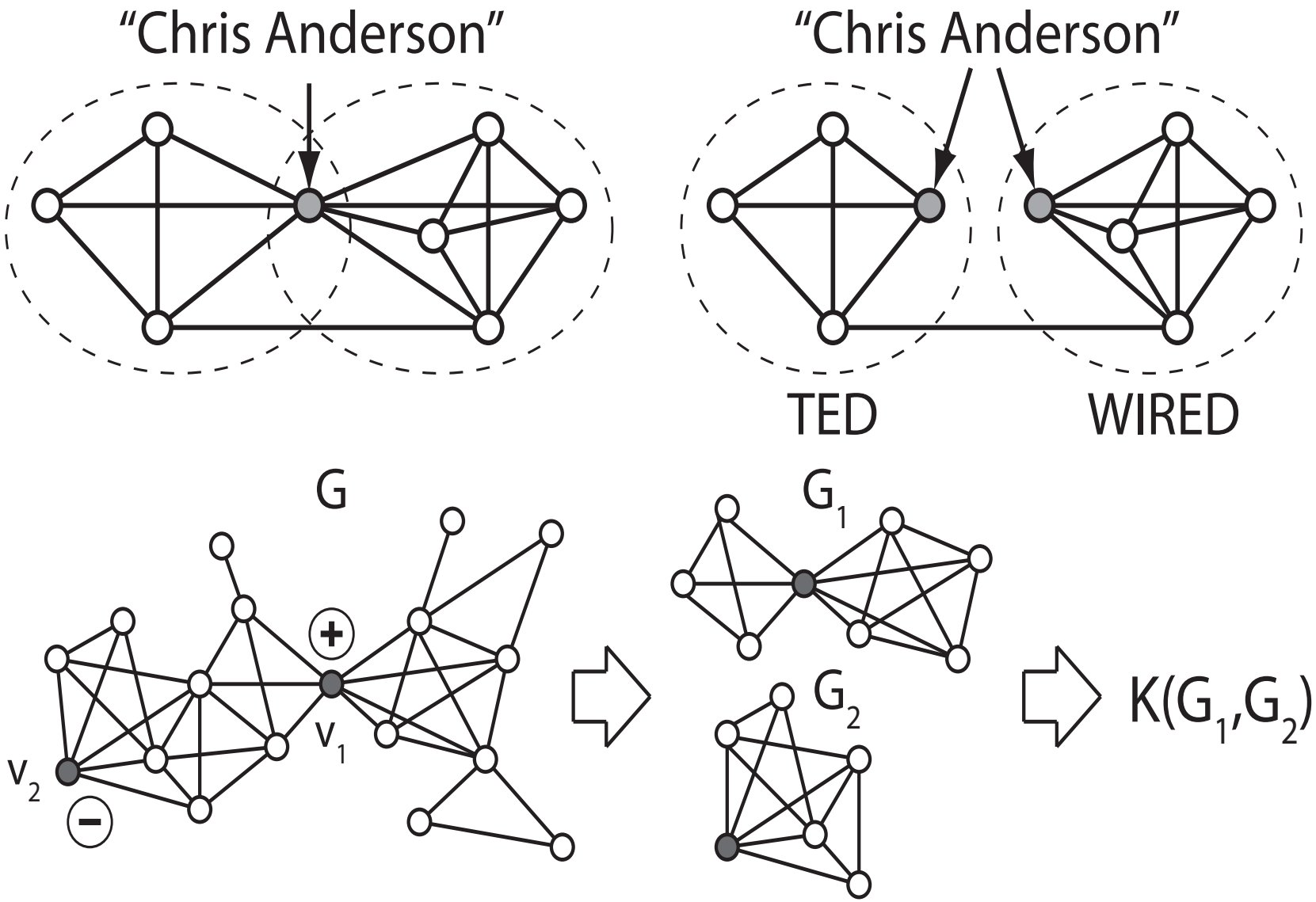
C. Dwork and A. Roth: The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* 9(3-4): 211-407 (2014)

Privacy in Social Networks

Facebook Patent

If the average credit rating of applicant's neighbours is at least a minimum credit score, the lender continues to process the loan application. Otherwise, the loan application is rejected.

Graph Classification for Entity Disambiguation



Graphs and Differential Privacy

Edge/Node-differential privacy

A randomized algorithm $\mathcal{A}$ is edge(node)-differentially private if for all events S in the output space of $\mathcal{A}$ and for all neighbor graphs G, G',

$$\Pr\left(\mathcal{A}(G) \in S\right) \leq \exp(\epsilon) \times \Pr\left(\mathcal{A}(G') \in S\right) + \delta \ .$$

G, G' are neighbours if:

node they differ in a node (and its neighbouring edges)

edge they differ in a single edge.

Graph Kernels

- A way to compare graphs for similarity
- Can be achieved by describing a graph by a feature vector

$$\mathbf{f}_G := [f_G(1), \ldots, f_G(a)]^\top.$$

Laplace Mechanism for Private Graph Kernels

Input: $G = (V, E)$.
Input: Privacy level ϵ
Input: Queries $f_G(i) : \mathcal{G} \to \mathbb{R}, \quad i = 1, ..., a$

$\mathbf{f}_G := [f_G(1), \ldots, f_G(a)]^\top$
$\tilde{\mathbf{f}}_G := \mathbf{f}_G + \mathbf{e}, \quad e(l) \sim \mathsf{Lap}(a \cdot GS_{f(i)}(\mathcal{G}_G)/\epsilon)$

Output: Private counts $\tilde{\mathbf{f}}_G$

Private Graph Kernels

The output of Algorithm is $(\epsilon, 0)$-differentially private.

High sensitivity of queries

High sensitivity

What is the maximum size of the common neighbourhood of a pair of connected vertices?

G (u, v) and both vertices connected to all other $n-2$: $n-2$

G' G minus edge (u, v): 0

Global sensitivity is $\Omega(n)$: problem is high degree vertices.

Restricted Sensitivity

Restrict to a subset $\mathcal{G}^H \subset \mathcal{G}$.

Restricted sensitivity. Blocki et al 2013

For a query f over $\mathcal{G}^H \subset \mathcal{G}$, with distance metric $d(G, G')$, the restricted sensitivity is

$$RS_f(\mathcal{G}) = \max_{G,G' \in \mathcal{G}^H} \left(\frac{|f(G) - f(G')|}{d(G, G')} \right) .$$

Example: set of graphs $\mathcal{G}_D$ of max degree at most D e.g. on Facebook, max degree may be $D = 5000 << n \approx 10^9$. For a graph $\mu_D(G)$ is a pruning of G to have max degree D.

Laplace Mechanism for Restricted Private Graph Kernels

Input: $G = (V, E)$, truncation level D
Input: Privacy level ϵ
Input: Queries $f_G(i) : \mathcal{G} \to \mathbb{R}, \quad i = 1, ..., a$

$G_D := \mu_D(G)$
$\mathbf{f}_{G_D} := [f_{G_D}(1), \ldots, f_{G_D}(a)]^\top$
$\tilde{\mathbf{f}}_G := \mathbf{f}_{G_D} + \mathbf{e}, \quad e(l) \sim \mathsf{Lap}(a \cdot RS_{f(i)}(\mathcal{G}_{G_D})/\epsilon)$

Output: Private counts $\tilde{\mathbf{f}}_G$

Restricted Sensitivity of Kernels

Shortest Paths

- $f_G(k) =$ no. of shortest paths of length k
- $RS_f = \Omega(n)$ even with $D = 2$.

Random Walks

- $f_G(k) =$ no. of walks of length k
- $RS_f \leq 2kD^{k-1}$.

Graphlets

- $f_G(k) =$ no. of graphlets of size k
- $RS_f \leq kD^{k-1}$.

"A good graph kernel need not be a good *private* graph kernel."

Graphlet Kernel: Sampling Large Graphs

Input: $G = (V, E)$, sample size s

$\hat{\mathbf{f}} = \mathbf{0}$, Approximate count for each graphlet type
for $j = 1 \ldots, s$ **do**
  Sample $e_j \in E$ uniformly at random
  $\hat{\mathbf{f}} \leftarrow \hat{\mathbf{f}}+$ counts of graphlets containing e
end for
$\hat{f}(i) \leftarrow \frac{\hat{f}(i)}{s} \frac{m}{|m_{H_i}|}$ for all i

Approximation of Graphlet Kernel

Approximation error of edge graphlet sampling

Consider $G = (V, E)$ with degree bounded by D, and let $Z_e(i)$ be the number of graphlets in G of type i that contains $e \in E$. For any $\gamma > 0, 0 < \rho < 1$, the counts, $\hat{\mathbf{f}} = [\hat{f}(1), \ldots, \hat{f}(a)]^\top$ produced by Algorithm have the following property.

$$\Pr\left(\left|\hat{f}(i) - \mathbb{E}[\hat{f}(i)]\right| \geq \gamma \mathbb{E}[\hat{f}(i)]\right) \leq \rho, \quad i \in [a] \tag{1}$$

using $s_i = 3\alpha_i \frac{\log \frac{2}{\rho}}{\gamma^2}$ samples with $\alpha_i = \frac{\max_e Z_e(i)}{\mathbb{E}[Z_e(i)]}$.

Complexity: computing number of graphlets containing an edge can be done in time $O(D^{k-2})$ and hence overall complexity is $O(sD^{k-2})$. Although α could be as high as m, *often it is much smaller* independent of m and so is the complexity.

Sampling and Privacy for Graphlet Kernel

Sampling increases privacy, hence well suited for large graphs, since it also approximates utility well.

Increasing privacy of sampled graphlet counts

Let $G = (V, E)$ and $m = |E|$. The mechanism $\mathcal{A}(\hat{\mathbf{f}})$ first sampling followed by the Laplace mechanism is an (ϵ_2, δ_2)-private algorithm for graphlet counts, for $\epsilon_2 \geq \log(1 + \beta_u e^{\epsilon_1} - \beta_l)$ and $\delta_2 \geq \beta_u \delta_1$, with $\beta_u = 1 - \left(\frac{(m-1)(m-2(D-1)^{t-1})}{m^2}\right)^s$ and $\beta_l = 1 - (1 - \frac{1}{m})^s$.

Datasets

PROTO	Synthetic population networks from Portland and Montgomery County (NDSSL Virginia Tech)
ROADS	Road network graphs of Texas and California, (Stanford SNAP database)
SOCIAL	Twitter and Google+ graphs (Stanford SNAP database)
D&D	Benchmark protein dataset for graph classification.

Results

Table: Statistics of datasets. Number of graphs N, number of nodes n, number of edges m and with α_k^* the maximum α_i over k-node graphlets, with α_i as in Theorem **??**. ‡Computation did not finish within 2 days.

N	Pos./Neg.	α_3^*	α_4^*	m_{max}	m_{avg}	n_{max}	n_{avg}	d_{m}
1178	691/487	4.3	50.6	14267	715.7	5748	284.3	
200	100/100	14.9	1689	10308	4321.4	1000	1000	1
200	100/100	33.9	2291	13973	13283.7	10000	10000	
262	132/132	‡	‡	1473709	104026.1	4938	1072.2	29

Kernel	D & D	PROTO	SOCIAL
prw	75.4 ± 0.6	83.8 ± 1.2	83.0 ± 0.4
wl	74.9 ± 0.6	93.7 ± 5.1	79.8 ± 1.8
gk-c3	74.4 ± 1.0	98.4 ± 1.1	89.0 ± 0.7
gk-c4	73.3 ± 1.0	99.9 ± 0.2	‡
gk-c5	74.1 ± 0.7	‡	‡
gk-a3	74.4 ± 0.4	74.0 ± 1.1	71.8 ± 1.7
gk-a4	74.7 ± 0.5	83.8 ± 0.9	76.2 ± 2.0
gk-a5	74.6 ± 0.5	85.0 ± 1.9	81.5 ± 1.8
prw	75.2 ± 0.9	89.0 ± 1.1	82.7 ± 0.9
gk-c3	74.2 ± 1.3	97.8 ± 2.9	88.7 ± 0.7
gk-c4	73.8 ± 0.7	99.5 ± 0.0	‡
gk-c5	73.6 ± 1.2	‡	‡
dpprw	68.4 ± 1.1	86.9 ± 3.0	68.9 ± 1.9
dpgk-c3	59.3 ± 0.5	73.3 ± 1.6	77.0 ± 0.5
dpgk-c4	58.6 ± 0.1	51.0 ± 3.1	‡
dpgk-s3	58.8 ± 0.1	74.0 ± 2.6	77.2 ± 0.7
dpgk-s4	58.7 ± 0.1	53.0 ± 2.6	52.7 ± 2.6

Sampling

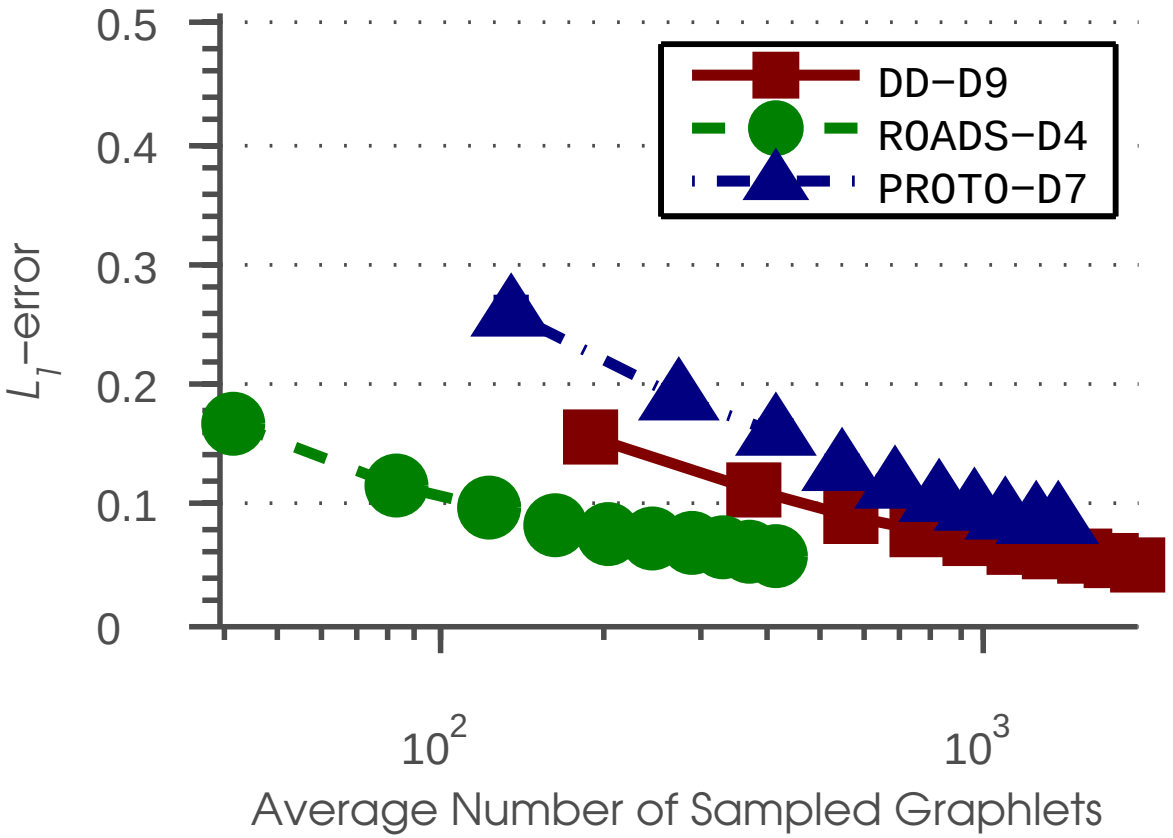
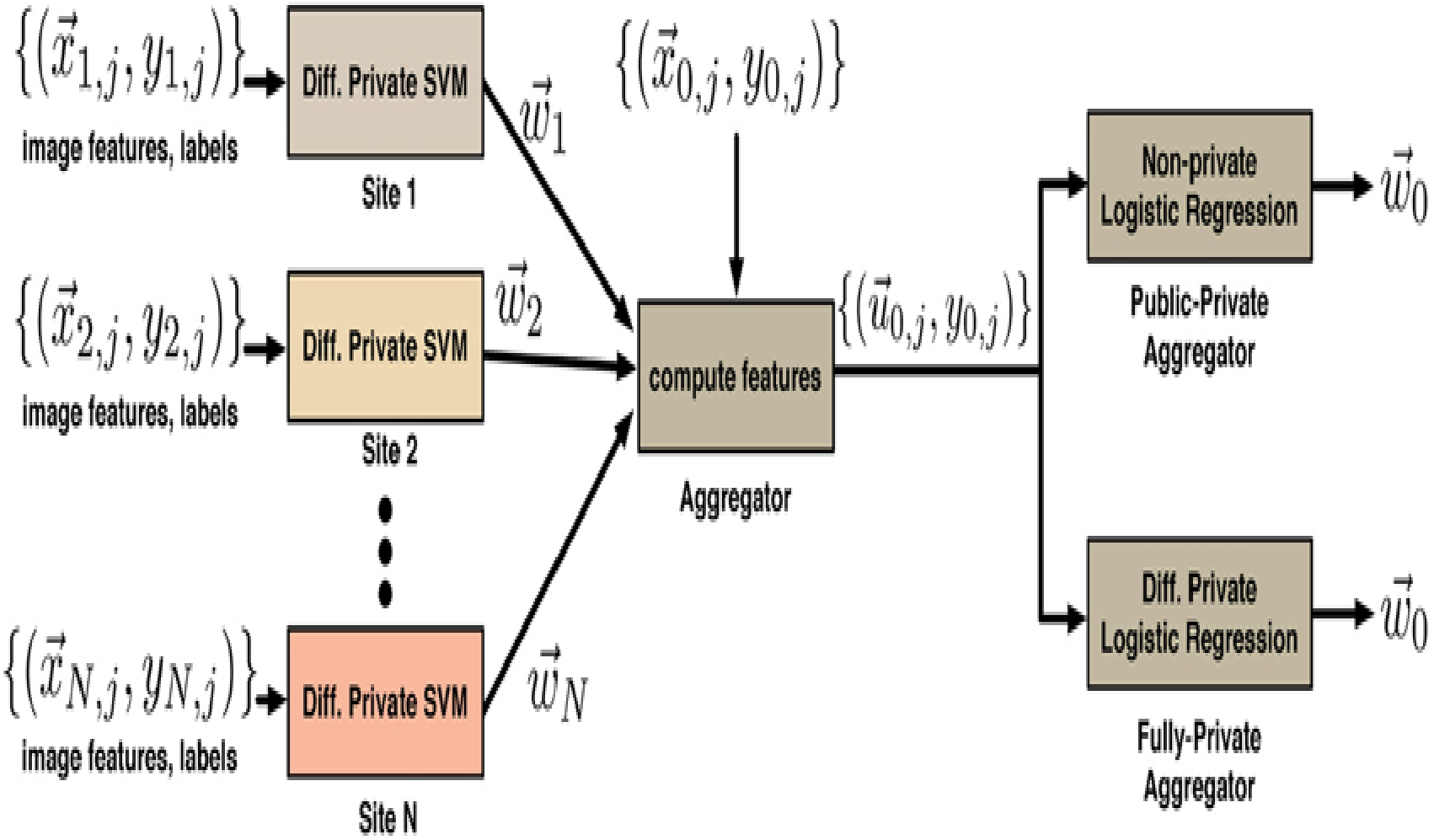


Figure: Average L_1-error in estimated 4-graphlet distributions of three datasets, for varying numbers of sampled graphlets. Each marker corresponds to an addition of 10 sampled edges.

Differential privacy for Neuroimaging Data

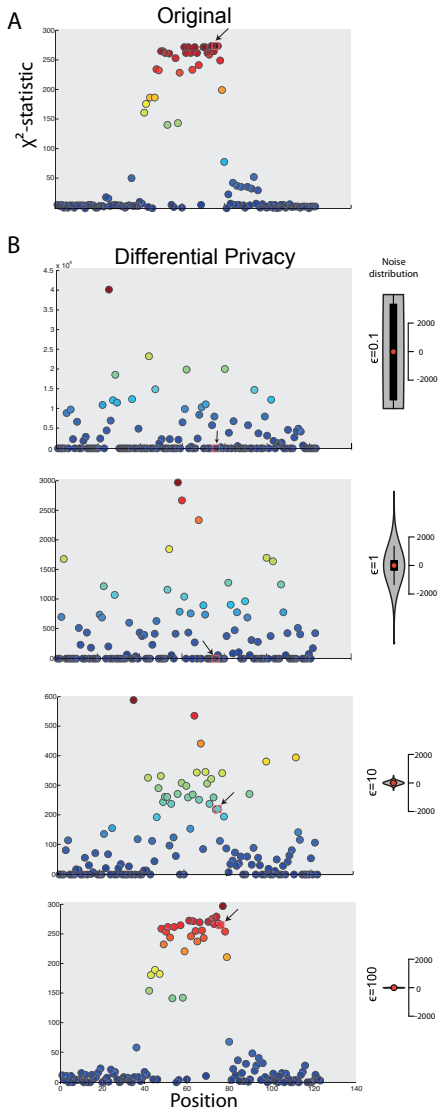


A. D. Sarwate *et al* "Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation". *Front. Neuroinform.* 2014.

Differential privacy for Genomic Data

SUPPLEMENTARY INFORMATION

In format provided by Erlich & Narayanan (JUNE 2014)



Supplementary information S1 (figure) | **Differential privacy statistic of an association study.** In the context of genetic privacy, several studies have explored differential private release of common summary statistics of GWAS data, such as the allele frequencies of cases and controls, χ^2-statistic, and p-values[1,2] or shifting the original locations of variants[3]. Currently, these techniques require a large amount of noise even for the release of a GWAS statistics from a small number of SNPs, which renders these differential private measures impractical. **a** | The χ^2-statistic of the association of single nucleotide polymorphisms (SNPs) in the IL28B region with Hepatitis-C outcome. We simulated association signals based on a previous study that examined genetic markers in the IL28B region for predictions of Hepatitis-C treatment outcomes in individuals from Japan[4]. To match the original study results, we tuned the simulation such that the genetic variant rs8099917 was the best indicator for treatment outcome with identical effect size to the original study. The arrow points to rs8099917, the top SNP identified in the original study. **b** | Differential private versions of the χ^2-statistic. We then tested varying levels of differential privacy perturbation to the χ^2-statistic following the procedure by Uhler and colleagues[1]. The χ^2-statistic of rs8099917 and the association signals showed very different patterns in the differential private results. Top: most stringent sanitization (ϵ = 0.1); bottom: least stringent sanitization (ϵ = 100). The two bottom subfigures show santiziation below the magnitude recommended in the literature. Right: violin plots of the noise probablity density functions (the ϵ = 0.1 probability density function was trimmed to fit the figure). Only when we significantly relaxed the perturbation far below the values recommended in the literature[5] did the results resemble the original association.

1. Uhler, C., Slavkovic, A. B. & Fienberg, S. E. Privacy-preserving data sharing for genome-wide association studies. *arXiv preprint arXiv:1205.0739* (2012).
2. Yu, F., Fienberg, S. E., Slavkovič, A. & Uhler, C. Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies. *arXiv preprint arXiv:1401.5193* (2014).
3. Johnson, A. & Shmatikov, V. in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* 1079-1087 (ACM, Chicago, Illinois, USA, 2013).
4. Ochi, H. *et al.* IL-28B predicts response to chronic hepatitis C therapy--fine-mapping and replication study in Asian populations. *Journal of General Virology* **92**, 1071-1081 (2011).
5. Hsu, J. *et al.* Differential Privacy: An Economic Method for Choosing Epsilon. *arXiv preprint arXiv:1402.3329* (2014).

Summmary

- Differential privacy (DP) is a principled approach to privacy.
- DP can be naturally adapted to graph structured data.
- Classification problems on graphs can be achieved via differential privacy, sometimes with reasonable utility performance.
- Can one make it more effective in practice?