Differentially Private Data Analysis of Social Networks via Restricted Sensitivity

Jeremiah Blocki, Avrim Blum, Anupam Datta, Or Sheffet

Presentation by Eric Bannatyne

GRAPHS AND SOCIAL NETWORKS

Social Network: A graph G with labeling function

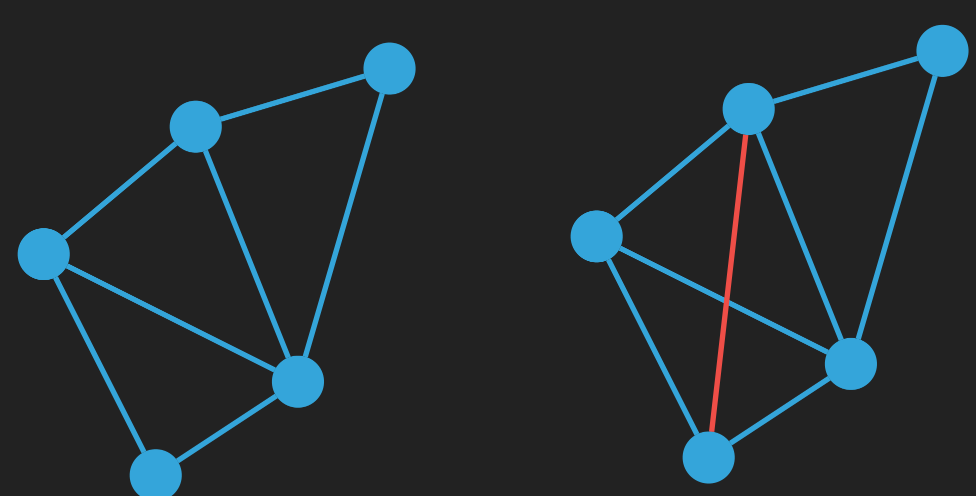
$$\ell : V(G) \to \mathbb{R}^m$$ (A person's age, occupation, etc.)

GRAPHS AND SOCIAL NETWORKS

Social Network: A graph G with labeling function

$$\ell : V(G) \to \mathbb{R}^m$$ (A person's age, occupation, etc.)

Edge Adjacency

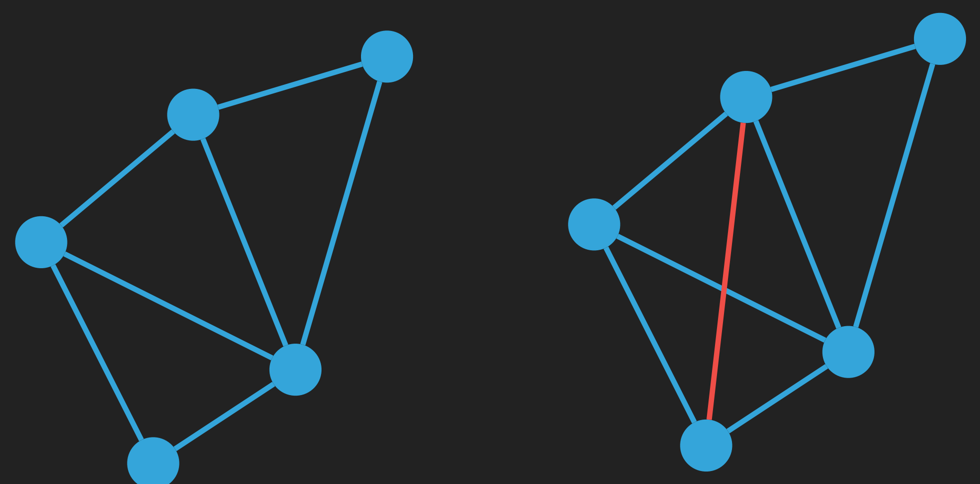


GRAPHS AND SOCIAL NETWORKS

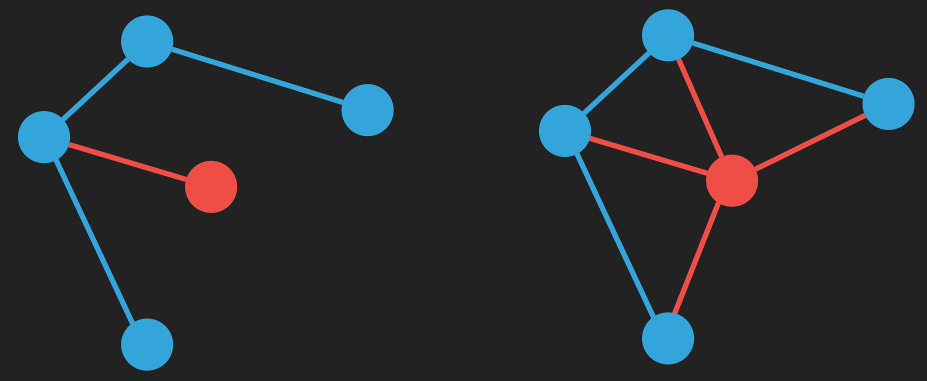
Social Network: A graph G with labeling function

$$\ell : V(G) \to \mathbb{R}^m$$ (A person's age, occupation, etc.)

Edge Adjacency



Vertex Adjacency



QUERYING SOCIAL NETWORKS

Local Profile Queries: How many people know two spies who don't know each other?

QUERYING SOCIAL NETWORKS

Local Profile Queries: How many people know two spies who don't know each other?

Subgraph Counting Queries: How many triangles are there involving at least one spy?

QUERIES HAVE HIGH GLOBAL SENSITIVITY

Vertex Adjacency: How many people are a doctor or are friends with a doctor?

QUERIES HAVE HIGH GLOBAL SENSITIVITY

Vertex Adjacency: How many people are a doctor or are friends with a doctor?

QUERIES HAVE HIGH GLOBAL SENSITIVITY

Vertex Adjacency: How many people are a doctor or are friends with a doctor?

Edge Adjacency: How many people are friends with two doctors who are friends with each other?

QUERIES HAVE HIGH GLOBAL SENSITIVITY

Vertex Adjacency: How many people are a doctor or are friends with a doctor?

Edge Adjacency: How many people are friends with two doctors who are friends with each other?

RESTRICTED SENSITIVITY

In any real social network, nobody is friends with everyone.

Hypothesis $\mathcal{H}$ encodes beliefs about the database.

e.g. Every node has degree at most $k = 5000$

RESTRICTED SENSITIVITY

In any real social network, nobody is friends with everyone.

Hypothesis $\mathcal{H}$ encodes beliefs about the database.

e.g. Every node has degree at most $k = 5000$

$$RS_f(\mathcal{H}) = \max_{D_1, D_2 \in \mathcal{H}} \left(\frac{|f(D_1) - f(D_2)|}{d(D_1, D_2)} \right)$$

Length of shortest chain of neighbouring databases between D_1 and D_2

RESTRICTED SENSITIVITY TO REDUCE NOISE

Restricted sensitivity is often much smaller than global sensitivity.

When possible: add noise proportional to $RS_f(\mathcal{H})$

- Achieve better accuracy when $\mathcal{H}$ is true.
- Still maintain privacy, even if $\mathcal{H}$ is false.

RESTRICTED SENSITIVITY TO REDUCE NOISE

Restricted sensitivity is often much smaller than global sensitivity.

When possible: add noise proportional to $RS_f(\mathcal{H})$

- Achieve better accuracy when $\mathcal{H}$ is true.
- Still maintain privacy, even if $\mathcal{H}$ is false.

Goal: Given a query $f : \mathcal{D} \to \mathbb{R}$
Define a new query $f_{\mathcal{H}}$ such that

$$f_{\mathcal{H}}(D) = f(D) \qquad \forall\, D \in \mathcal{H} \qquad \text{and}$$

$$GS_{f_{\mathcal{H}}} = RS_f(\mathcal{H}).$$

GENERAL CONSTRUCTION

For each $D \in \mathcal{H}$ set $f_{\mathcal{H}}(D) = f(D)$

Arbitrarily order elements of $\mathcal{D} \setminus \mathcal{H} = \{D_1, D_2, \ldots, D_m\}$

- Define $f_{\mathcal{H}}(D_i)$ inductively. $\mathcal{T}_i = \mathcal{H} \cup \{D_1, \ldots, D_i\}$.

GENERAL CONSTRUCTION

For each $D \in \mathcal{H}$ set $f_{\mathcal{H}}(D) = f(D)$

Arbitrarily order elements of $\mathcal{D} \setminus \mathcal{H} = \{D_1, D_2, \ldots, D_m\}$

Define $f_{\mathcal{H}}(D_i)$ inductively. $\mathcal{T}_i = \mathcal{H} \cup \{D_1, \ldots, D_i\}$.

Choose $f_{\mathcal{H}}(D_{i+1})$ such that

$$\frac{|f_{\mathcal{H}}(D) - f_{\mathcal{H}}(D_{i+1})|}{d(D, D_{i+1})} \leq RS_{f_{\mathcal{H}}}(\mathcal{T}_i) \qquad \forall D \in \mathcal{T}_i.$$

Need a bit of calculation to show that this exists.

GENERAL CONSTRUCTION CONT'D

Choose $f_{\mathcal{H}}(D_{i+1})$ such that

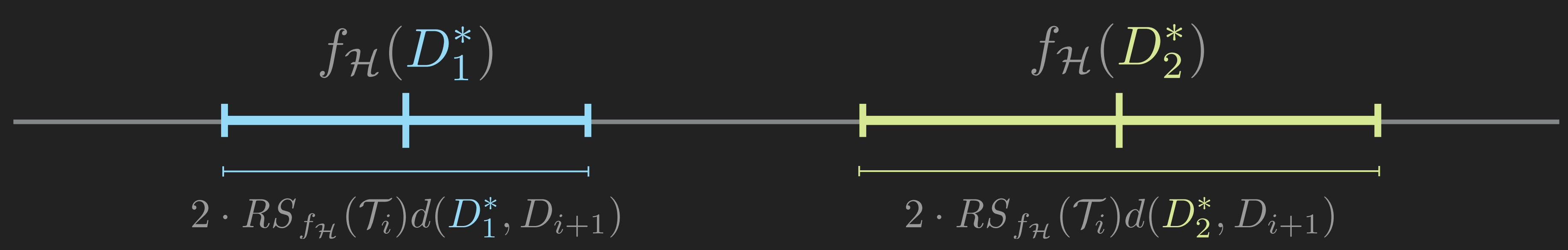
$$\frac{|f_{\mathcal{H}}(D) - f_{\mathcal{H}}(D_{i+1})|}{d(D, D_{i+1})} \leq RS_{f_{\mathcal{H}}}(\mathcal{T}_i) \qquad \forall D \in \mathcal{T}_i.$$

GENERAL CONSTRUCTION CONT'D

Choose $f_{\mathcal{H}}(D_{i+1})$ such that

$$\frac{|f_{\mathcal{H}}(D) - f_{\mathcal{H}}(D_{i+1})|}{d(D, D_{i+1})} \leq RS_{f_{\mathcal{H}}}(\mathcal{T}_i) \qquad \forall D \in \mathcal{T}_i.$$

If no such value exists, then there would be some $D_1^*, D_2^* \in \mathcal{T}_i$ such that

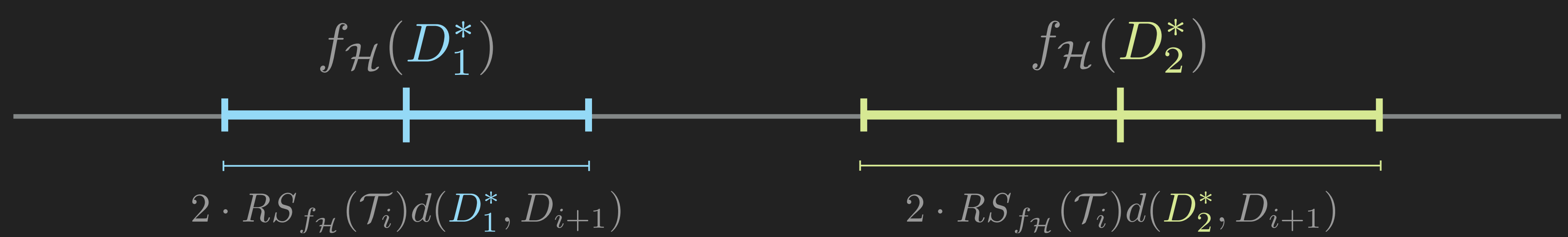


GENERAL CONSTRUCTION CONT'D

Choose $f_{\mathcal{H}}(D_{i+1})$ such that

$$\frac{|f_{\mathcal{H}}(D) - f_{\mathcal{H}}(D_{i+1})|}{d(D, D_{i+1})} \leq RS_{f_{\mathcal{H}}}(\mathcal{T}_i) \qquad \forall D \in \mathcal{T}_i.$$

If no such value exists, then there would be some $D_1^*, D_2^* \in \mathcal{T}_i$ such that



(Contradiction)

Then $$\frac{|f_{\mathcal{H}}(D_1^*) - f_{\mathcal{H}}(D_2^*)|}{d(D_1^*, D_2^*)} \geq \frac{|f_{\mathcal{H}}(D_1^*) - f_{\mathcal{H}}(D_2^*)|}{d(D_{i+1}, D_1^*) + d(D_{i+1}, D_2^*)} > RS_{f_{\mathcal{H}}}(\mathcal{T}_i).$$

RESTRICTED SENSITIVITY FOR $\mathcal{H}_k$

$\mathcal{H}_k = \{$graphs of degree at most $k\}$, where $k \ll n$.

RESTRICTED SENSITIVITY FOR $\mathcal{H}_k$

$\mathcal{H}_k = \{$graphs of degree at most $k\}$, where $k \ll n$.

Local profile $p(v, G_v) \in [0, 1]$ $\qquad f_p(G, \ell) = \sum_{v \in V(G)} p(v, G_v).$

Nbhd of v
(including v itself)

RESTRICTED SENSITIVITY FOR $\mathcal{H}_k$

$\mathcal{H}_k = \{$graphs of degree at most $k\}$, where $k \ll n$.

Local profile $p(v, G_v) \in [0,1]$ $\qquad f_p(G,\ell) = \sum_{v \in V(G)} p(v, G_v).$

Nbhd of v
(including v itself)

For local profile queries, $RS_f(\mathcal{H}_k) \leq 2k+1$ (Vertex adjacency)

$RS_f(\mathcal{H}_k) \leq k+1$ (Edge adjacency).

RESTRICTED SENSITIVITY FOR $\mathcal{H}_k$

$\mathcal{H}_k = \{$graphs of degree at most $k\}$, where $k \ll n$.

Local profile $p(v, G_v) \in [0,1]$ $\qquad f_p(G,\ell) = \sum_{v \in V(G)} p(v, G_v).$

Nbhd of v

(including v itself)

For local profile queries, $RS_f(\mathcal{H}_k) \leq 2k+1$ (Vertex adjacency)

$RS_f(\mathcal{H}_k) \leq k+1$ (Edge adjacency).

Subgraph counting: Given connected graph H, predicates $p_1, \ldots, p_t$,

$$f(G,\ell) = \left|\{\{v_1,\ldots,v_t\} : G[v_1,\ldots,v_t] = H \text{ and } \forall i, \ell(v_i) \in p_i\}\right|.$$

RESTRICTED SENSITIVITY FOR $\mathcal{H}_k$

$\mathcal{H}_k = \{$graphs of degree at most $k\}$, where $k \ll n$.

Local profile $p(v, G_v) \in [0,1]$ $\qquad f_p(G,\ell) = \sum_{v \in V(G)} p(v, G_v).$

Nbhd of v

(including v itself)

For local profile queries, $RS_f(\mathcal{H}_k) \leq 2k+1$ (Vertex adjacency)

$RS_f(\mathcal{H}_k) \leq k+1$ (Edge adjacency).

Subgraph counting: Given connected graph H, predicates $p_1, \ldots, p_t$,

$$f(G,\ell) = \left|\{\{v_1, \ldots, v_t\} : G[v_1, \ldots, v_t] = H \text{ and } \forall i, \ell(v_i) \in p_i\}\right|.$$

For subgraph counting queries, $RS_f(\mathcal{H}_k) \leq tk^{t-1}$.

SMOOTH PROJECTIONS

General construction is really inefficient, only works for one query at a time.

- Want a canonical projection $\mu : \mathcal{D} \to \mathcal{H}$ such that $\mu(D) = D \quad \forall\, D \in \mathcal{H}$.
- Then set $f_{\mathcal{H}} = f \circ \mu$.

SMOOTH PROJECTIONS

General construction is really inefficient, only works for one query at a time.

- Want a canonical projection $\mu : \mathcal{D} \to \mathcal{H}$ such that $\mu(D) = D \quad \forall\, D \in \mathcal{H}$.
- Then set $f_{\mathcal{H}} = f \circ \mu$.

Projection is c-smooth if D, D' neighbouring implies $d(\mu(D), \mu(D')) \leq c$.

SMOOTH PROJECTIONS

General construction is really inefficient, only works for one query at a time.

- Want a canonical projection $\mu : \mathcal{D} \to \mathcal{H}$ such that $\mu(D) = D \quad \forall D \in \mathcal{H}$.
- Then set $f_{\mathcal{H}} = f \circ \mu$.

Projection is c-smooth if D, D' neighbouring implies $d(\mu(D), \mu(D')) \leq c$.

Lemma. If μ is c-smooth, then $GS_{f_{\mathcal{H}}} \leq c \cdot RS_f(\mathcal{H})$.

SMOOTH PROJECTIONS

General construction is really inefficient, only works for one query at a time.

- Want a canonical projection $\mu : \mathcal{D} \to \mathcal{H}$ such that $\mu(D) = D \quad \forall\, D \in \mathcal{H}$.
- Then set $f_{\mathcal{H}} = f \circ \mu$.

Projection is c-smooth if D, D' neighbouring implies $d(\mu(D), \mu(D')) \leq c$.

Lemma. If μ is c-smooth, then $GS_{f_{\mathcal{H}}} \leq c \cdot RS_f(\mathcal{H})$.

Proof.

$$\begin{aligned}
GS_{f_{\mathcal{H}}} &= \max_{D_1 \sim D_2} |f(\mu(D_1)) - f(\mu(D_2))| \\
&\leq \max_{D_1 \sim D_2} |f(\mu(D_1)) - f(\mu(D_2))| \frac{c}{d(\mu(D_1), \mu(D_2))} \\
&\leq c \max_{D_1, D_2 \in \mathcal{H}} \frac{|f(D_1) - f(D_2)|}{d(D_1, D_2)} \\
&= c \cdot RS_f(\mathcal{H}).
\end{aligned}$$

PROJECTION SCHEME FOR $\mathcal{H}_k$ (In the Edge Adjacency Model)

Fix a canonical ordering over all possible edges.

For each node v with $\deg(v) > k$, delete all but the first k edges incident to v.

PROJECTION SCHEME FOR $\mathcal{H}_k$ (In the Edge Adjacency Model)

Fix a canonical ordering over all possible edges.

For each node v with $\deg(v) > k$, delete all but the first k edges incident to v.

Claim. This is a 3-smooth projection.

PROJECTION SCHEME FOR $\mathcal{H}_k$ (In the Edge Adjacency Model)

Fix a canonical ordering over all possible edges.

For each node v with $\deg(v) > k$, delete all but the first k edges incident to v.

Claim. This is a 3-smooth projection.

Proof. Suppose G_1, G_2 differ on a single edge $e = (x, y) \in E(G_1)$.

If μ deletes e then $\mu(G_1) = \mu(G_2)$. Otherwise,

PROJECTION SCHEME FOR $\mathcal{H}_k$ (In the Edge Adjacency Model)

Fix a canonical ordering over all possible edges.

For each node v with $\deg(v) > k$, delete all but the first k edges incident to v.

Claim. This is a 3-smooth projection.

Proof. Suppose G_1, G_2 differ on a single edge $e = (x, y) \in E(G_1)$.

If μ deletes e then $\mu(G_1) = \mu(G_2)$. Otherwise,

PROJECTION SCHEME FOR $\mathcal{H}_k$ (In the Edge Adjacency Model)

Fix a canonical ordering over all possible edges.

For each node v with $\deg(v) > k$, delete all but the first k edges incident to v.

Claim. This is a 3-smooth projection.

Proof. Suppose G_1, G_2 differ on a single edge $e = (x, y) \in E(G_1)$.

If μ deletes e then $\mu(G_1) = \mu(G_2)$. Otherwise,

PUTTING IT TOGETHER

For any query f, in the edge adjacency model, the mechanism

$$\mathcal{M}(G, f) = f(\mu(G)) + Lap\left(\frac{3 \cdot RS_f(\mathcal{H}_k)}{\varepsilon}\right)$$

satisfies $(\varepsilon, 0)$-differential privacy. (In the edge adjacency model)

For local profile queries, $RS_f(\mathcal{H}_k) \leq 2k + 1 \ll n$.

SUMMARY

Natural queries on social networks have high global sensitivity.

- Require lots of noise to preserve privacy.

SUMMARY

Natural queries on social networks have high global sensitivity.

- Require lots of noise to preserve privacy.

By choosing the right hypothesis, we can reduce the restricted sensitivity.

- We can add less noise to preserve privacy.
- Can achieve better accuracy when $\mathcal{H}$ is true.
- Still preserve privacy, even if $\mathcal{H}$ is false.

SUMMARY

Natural queries on social networks have high global sensitivity.

- Require lots of noise to preserve privacy.

By choosing the right hypothesis, we can reduce the restricted sensitivity.

- We can add less noise to preserve privacy.
- Can achieve better accuracy when $\mathcal{H}$ is true.
- Still preserve privacy, even if $\mathcal{H}$ is false.

For graphs of bounded degree, we can efficiently reduce the noise needed, using smooth projections.