New Challenges of Data Privacy in a Socially Connected World

Jie Gao

Stony Brook University
http://www.cs.stonybrook.edu/~jgao

July 20th 2018.
New Challenges in Data Privacy

- Big Data;
- Internet of Things
New Challenges in Data Privacy

- Big Data;
- Internet of Things

New law enforcements:
New Challenges in Data Privacy

- Big Data;
- Internet of Things

New law enforcements:
- Fair Infomation Practice Principles (FIPPs): collection limitation, purpose specification, use limitation, accountability, security, notice and choice.
- General Data Protection Regulation (GPDR), effective 25 May 2018;
Problem A: De-anonymize A Social Network.
Network Alignment – De-anonymization

Consider the same group of people who participate in two social network platforms:

- Private network: identities not revealed, e.g., Facebook.
Network Alignment – De-anonymization

Consider the same group of people who participate in two social network platforms:

- Private network: identities not revealed, e.g., Facebook.
- Public network: identities shown in public, e.g., LinkedIn.
Network Alignment – De-anonymization

Consider the same group of people who participate in two social network platforms:

- Private network: identities not revealed, e.g., Facebook.
- Public network: identities shown in public, e.g., LinkedIn.

The two networks have almost the same topology.
Network Alignment – De-anonymization

Consider the same group of people who participate in two social network platforms:

- Private network: identities not revealed, e.g., Facebook.
- Public network: identities shown in public, e.g., LinkedIn.

The two networks have almost the same topology.

De-anonymization: If we can align the two networks by vertex correspondences, the identities of the private network are thus revealed.
Graph Isomorphism

Given a pair of graphs $G_1$, $G_2$, find a one-to-one correspondence of the vertices in $G_1$ to vertices in $G_2$ such that $(u, v)$ is an edge in $G_1$ if and only if their corresponding nodes $f(u), f(v)$ are connected in $G_2$. 
Graph Isomorphism

Given a pair of graphs $G_1, G_2$, find a one-to-one correspondence of the vertices in $G_1$ to vertices in $G_2$ such that $(u, v)$ is an edge in $G_1$ if and only if their corresponding nodes $f(u), f(v)$ are connected in $G_2$. 
Graph Isomorphism

One of the most fundamental problems in theoretical computer science.

- In NP.
Graph Isomorphism

One of the most fundamental problems in theoretical computer science.

- In NP.
- Nov 2015/Jan 2017, László Babai claimed quasi-polynomial time algorithm: $O(\exp(\log^{O(1)} n))$. 
Graph Isomorphism

One of the most fundamental problems in theoretical computer science.

- In NP.
- Nov 2015/Jan 2017, László Babai claimed quasi-polynomial time algorithm: \( O(\exp(\log^{O(1)} n)) \).
- Many practical algorithms: e.g., NAUTY.
Graph Isomorphism

One of the most fundamental problems in theoretical computer science.

- In NP.
- Nov 2015/Jan 2017, László Babai claimed quasi-polynomial time algorithm: $O(\exp(\log^{O(1)} n))$.
- Many practical algorithms: e.g., NAUTY.
- Subgraph isomorphism is NP-complete.
Graph Isomorphism

One of the most fundamental problems in theoretical computer science.

- In NP.
- Nov 2015/Jan 2017, László Babai claimed quasi-polynomial time algorithm: $O(\exp(\log^{O(1)} n))$.
- Many practical algorithms: e.g., NAUTY.
- Subgraph isomorphism is NP-complete.
- **Approximate graph isomorphism**: find the best correspondence between vertices in $G_1$ and $G_2$ s.t. if $u, v$ are connected in $G_1$ their corresponding nodes are likely connected in $G_2$. 
Our Solution: A Geometric Approach

How to align two sets of points in the plane, assuming that some landmarks $\ell_i$ are already aligned?

$p = (d_1, d_2, d_3)$

$p' = (d'_1, d'_2, d'_3)$
Our Solution: A Geometric Approach

How to align two sets of points in the plane, assuming that some landmarks $\ell_i$ are already aligned?

- Any point $p$ can be represented by the barycentric coordinates $(d_1, d_2, d_3)$, $d_i$ is distance to $\ell_i$.
- If the barycentric coordinates of $p$ and $p'$ are similar, we match $p$ and $p'$.
Quantify the ‘Position’ of a Node in a Network

In a social network there are often nodes that can be easily identified as landmarks. Define the position of a node wrt landmarks.
Quantify the ‘Position’ of a Node in a Network

In a social network there are often nodes that can be easily identified as landmarks. Define the position of a node wrt landmarks.

Q: What distance to use?
Quantify the ‘Position’ of a Node in a Network

In a social network there are often nodes that can be easily identified as *landmarks*. Define the position of a node wrt landmarks.

Q: What distance to use?

- Tie strength – Trouble: not easy to measure.
Quantify the ‘Position’ of a Node in a Network

In a social network there are often nodes that can be easily identified as *landmarks*. Define the position of a node wrt landmarks.

Q: What distance to use?

- **Tie strength** – Trouble: not easy to measure.
- **Count # hops to these landmarks** – Trouble: small world property;
Quantify the ‘Position’ of a Node in a Network

In a social network there are often nodes that can be easily identified as *landmarks*. Define the position of a node wrt landmarks.

Q: What distance to use?

- Tie strength – Trouble: not easy to measure.
- Count # hops to these landmarks – Trouble: small world property;
- Distances from some geometric embedding (spectral embedding, Tutte embedding).
Quantify the ‘Position’ of a Node in a Network

In a social network there are often nodes that can be easily identified as *landmarks*. Define the position of a node wrt landmarks.

Q: What distance to use?
- Tie strength – Trouble: not easy to measure.
- Count # hops to these landmarks – Trouble: small world property;
- Distances from some geometric embedding (spectral embedding, Tutte embedding).

Q: Robust to noises (edge insertion/deletion)?
Robustness: Remove Two Edges

Left: Spectral embedding; Right: Tutte/Spring embedding.
Robustness: Remove Two Edges

Left: Hop count; Right: our metric.

Q: How is our metric defined?
Robustness: Remove Two Edges

Left: Hop count; Right: our metric.

Q: How is our metric defined?
Discrete Ricci Curvature

Take the analog: for an edge $xy$, consider the distances from $x$’s neighbors to $y$’s neighbors and compare it with the length of $xy$. 

Issue: how to match $x$’s neighbors to $y$’s neighbors?

Assign uniform distribution $\mu_1$, $\mu_2$ on $x$’ and $y$’s neighbors.

Use optimal transportation distance (earth-mover distance) from $\mu_1$ to $\mu_2$: the matching that minimize the total transport distance.
Discrete Ricci Curvature

Take the analog: for an edge $xy$, consider the distances from $x$’s neighbors to $y$’s neighbors and compare it with the length of $xy$.

- Issue: how to match $x$’s neighbors to $y$’s neighbors?
Discrete Ricci Curvature

Take the analog: for an edge $xy$, consider the distances from $x$’s neighbors to $y$’s neighbors and compare it with the length of $xy$.

- Issue: how to match $x$’s neighbors to $y$’s neighbors?
- Assign uniform distribution $\mu_1$, $\mu_2$ on $x$’ and $y$’s neighbors.
- Use optimal transportation distance (earth-mover distance) from $\mu_1$ to $\mu_2$: the matching that minimize the total transport distance.
Discrete Ricci Curvature

Definition (Ollivier)

Let $(X, d)$ be a metric space and let $m_1, m_2$ be two probability measures on $X$. For any two distinct points $x, y \in X$, the (Ollivier-) Ricci curvature along $xy$ is defined as

$$\kappa(x, y) := 1 - \frac{W_1(m_x, m_y)}{d(x, y)},$$

where $m_x$ ($m_y$) is a probability distribution defined on $x$ ($y$) and its neighbors, $W_1(\mu_1, \mu_2)$ is the $L_1$ optimal transportation distance between two probability measure $\mu_1$ and $\mu_2$ on $X$:

$$W_1(\mu_1, \mu_2) := \inf_{\psi \in \Pi(\mu_1, \mu_2)} \int d(u, v)\psi(u, v)$$
Examples

Zero curvature: 2D grid.
Examples

Negative curvature: tree: $\kappa(x, y) = 1/d_x + 1/d_y - 1$, $d_x$ is degree of $x$. 

![Diagram of a tree with nodes and edges labeled with 0.33 and -0.167]
Examples

Positive curvature: complete graph.
Edge Weights Generated by Ricci flow

Given a graph $G$ in which $d(x, y)$ is the weight of the edge $xy$ and $\kappa(x, y)$ is the discrete Ricci curvature, we run

$$d_{i+1}(x, y) = (d_i(x, y) - \varepsilon \cdot \kappa_i(x, y) \cdot d_i(x, y)) \cdot N$$

Until convergence, where $N$ is to rescale to make sure total edge weights remain the same.

At the limit, $W(x, y)/d(x, y)$ is the same for all edges.
Ricci Flow Metric

Intuition: flatten the network – shrink an edge if it is within a well connected community; stretch an edge if otherwise, s.t., the network curvature is uniform everywhere.
Evaluation on Resilience

Randomly remove 10 edges in a random regular graph.

Random Regular (1000 nodes, 6000 edges) with 10 edges removed

Histogram of RF Metric with ATD

Histogram of Hop Count

Histogram of Spring

Histogram of Spectral

Histogram of RF Metric with OTD

Histogram of Shortest Path Stretch Ratio

−2.0 −1.5 −1.0 −0.5 0 0.5 1.0
Evaluation on Matching Performance

- Randomly remove one node in a random regular graph w/ degree 12.
- Right: remove randomly 10 edges in a protein-protein network.
Problem B: Location Privacy in Mobility Data.
Populated by Wireless Devices

Smart building, smart city, smart communities: an increasing number of wireless devices, both static and mobile, populating the physical space.
Populated by Wireless Devices

Smart building, smart city, smart communities: an increasing number of wireless devices, both static and mobile, populating the physical space.
Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Frequently visited locations: home or work addresses.
- Frequently co-located pairs: social ties.
- Unique signatures: 4 spatial temporal data points can be used to identify a user from a database of 3 million users.

Our setting: Mobility data of privacy-aware/sensitive and privacy-indifferent users.

Question: if privacy insensitive users publish their whereabouts, how much information can we infer for privacy sensitive users?
Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Frequently visited locations: home or work addresses.
Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Frequently visited locations: home or work addresses.
- Frequently co-located pairs: social ties.
Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Frequently visited locations: home or work addresses.
- Frequently co-located pairs: social ties.
- Unique signatures: 4 spatial temporal data points can be used to identify a user from a database of 1.3 million users.
Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Frequently visited locations: home or work addresses.
- Frequently co-located pairs: social ties.
- Unique signatures: 4 spatial temporal data points can be used to identify a user from a database of 1.3 million users.

Our setting: Mobility data of privacy-aware/sensitive and privacy-indifferent users.
Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Frequently visited locations: home or work addresses.
- Frequently co-located pairs: social ties.
- Unique signatures: 4 spatial temporal data points can be used to identify a user from a database of 1.3 million users.

Our setting: Mobility data of privacy-aware/sensitive and privacy-indifferent users.
Question: if privacy insensitive users publish their whereabouts, how much information can we infer for privacy sensitive users?
Location Privacy

Time: 9am; Location: North Hall

Time: 9:30am; Location unknown

Time: 10am; Location: CS building
Problem Definition

Consider a mixture of privacy aware users and privacy insensitive users in motion. Privacy insensitive users may occasionally report

- GPS event \((i, \tau, p)\): user \(i\) is at location \(p\) at time \(\tau\).
Problem Definition

Consider a mixture of privacy aware users and privacy insensitive users in motion. Privacy insensitive users may occasionally report

- GPS event \((i, \tau, p)\): user \(i\) is at location \(p\) at time \(\tau\).
- Meeting events \(\chi = (i, j, \tau)\): user \(i, j\) appear at the same location at time \(\tau\).
Problem Definition

Consider a mixture of privacy aware users and privacy insensitive users in motion. Privacy insensitive users may occasionally report

- GPS event \((i, \tau, p)\): user \(i\) is at location \(p\) at time \(\tau\).
- Meeting events \(\chi = (i, j, \tau)\): user \(i, j\) appear at the same location at time \(\tau\).

A user \(i\) has a maximum travel speed \(v_i\).
Problem Definition

Consider a mixture of privacy aware users and privacy insensitive users in motion. Privacy insensitive users may occasionally report

- GPS event \( (i, \tau, p) \): user \( i \) is at location \( p \) at time \( \tau \).
- Meeting events \( \chi = (i, j, \tau) \): user \( i, j \) appear at the same location at time \( \tau \).

A user \( i \) has a maximum travel speed \( v_i \).

Two events \( (i, j, \tau) \) and \( (i, j', \tau') \) that share one user is at most \( |\tau - \tau'|v_i \) away.
Problem Definition

Consider a mixture of privacy aware users and privacy insensitive users in motion. Privacy insensitive users may occasionally report

- GPS event \((i, \tau, p)\): user \(i\) is at location \(p\) at time \(\tau\).
- Meeting events \(\chi = (i, j, \tau)\): user \(i, j\) appear at the same location at time \(\tau\).

A user \(i\) has a maximum travel speed \(v_i\).

Two events \((i, j, \tau)\) and \((i, j', \tau')\) that share one user is at most \(|\tau - \tau'|v_i\) away.

Q: Can we infer feasibility region of the meeting events \(R = \{R(\chi), \forall \chi\}\) (which will imply location information of privacy aware users)?
Our Results

- Solving $R$ is a convex problem – all distance constraints are convex.
Our Results

- Solving $R$ is a convex problem – all distance constraints are convex.
- If we measure distance by $\ell_\infty$, we can separate different dimensions and obtain a nearly linear running time algorithm.
Our Results

- Solving $R$ is a convex problem – all distance constraints are convex.
- If we measure distance by $\ell_\infty$, we can separate different dimensions and obtain a nearly linear running time algorithm.
- If we use $\ell_2$ distance, the region boundary may have high complexity, e.g., depending on $O(n)$ parameters.
Our Results

- Solving $R$ is a convex problem – all distance constraints are convex.
- If we measure distance by $\ell_\infty$, we can separate different dimensions and obtain a nearly linear running time algorithm.
- If we use $\ell_2$ distance, the region boundary may have high complexity, e.g., depending on $O(n)$ parameters.
- But we can compute $\varepsilon$-approximations in nearly linear running time.
Our Results

- Solving $R$ is a convex problem – all distance constraints are convex.
- If we measure distance by $\ell_\infty$, we can separate different dimensions and obtain a nearly linear running time algorithm.
- If we use $\ell_2$ distance, the region boundary may have high complexity, e.g., depending on $O(n)$ parameters.
- But we can compute $\varepsilon$-approximations in nearly linear running time.
- If we enforce speed lower bound, the problem is $\exists R$-complete.
Our Results

- Solving $R$ is a convex problem – all distance constraints are convex.
- If we measure distance by $\ell_\infty$, we can separate different dimensions and obtain a nearly linear running time algorithm.
- If we use $\ell_2$ distance, the region boundary may have high complexity, e.g., depending on $O(n)$ parameters.
- But we can compute $\varepsilon$-approximations in nearly linear running time.
- If we enforce speed lower bound, the problem is $\exists R$-complete.
- If the domain has holes, the problem is NP-hard.
Simulations

6,099 taxis in a region of area $1,847\, km^2$ in one hour, and 14,534 meeting events.
Simulations

6,099 taxis in a region of area 1,847 km$^2$ in one hour, and 14,534 meeting events.

- If each taxi reports beginning and the end of the hour, we can bound feasibility region of each meeting event by a 10km by 12km region.
Simulations

6,099 taxis in a region of area $1,847 \text{km}^2$ in one hour, and 14,534 meeting events.

- If each taxi reports beginning and the end of the hour, we can bound feasibility region of each meeting event by a 10km by 12km region.

- If 1/3 taxis do not report their locations while others report every 5mins, feasibility region has height about 1.6km.
Conclusions

Protecting privacy is not easy.
Conclusions

Protecting privacy is not easy.

- Current privacy regulations do not rule out social attacks.
- Tradeoff between utility and privacy.
- More technical solutions are needed.
Acknowledgement

- Chien-Chun Ni, Yu-Yao Lin, Jie Gao, Xianfeng David Gu, Graph Drawing 2018.
- Boris Aronov, Alon Efrat, Ming Li, Jie Gao, Joseph S. B. Mitchell, Valentin Polishchuk, Boyang Wang, Hanyu Quan, Jiaxin Ding, MobiHoc’18.
- Questions and comments?