New Challenges of Data Privacy in a Socially Connected World

Jie Gao

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

Fall Workshop on Computational Geometry
October 27th 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;

Protecting Data Privacy

Most work on protecting data privacy focus on databases.

- Medical records.

Protecting Data Privacy

Most work on protecting data privacy focus on databases.

- Medical records.

Privacy models:

- k-anonymity (for ID association) [Sweeney02]: each item is identical to $k-1$ other items.

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Protecting Data Privacy

Most work on protecting data privacy focus on databases.

- Medical records.

Privacy models:

- k-anonymity (for ID association) [Sweeney02]: each item in D is identical to $k-1$ other items.
- Differential privacy [Dwork06]: add noises to query results $Q[D]$ s.t. $\forall p \in D$

$$\frac{\text{Prob}\{Q[D] = r\}}{\text{Prob}\{Q[D \setminus p] = r\}} \leq e^{\varepsilon}$$

Protecting Data Privacy

Most work on protecting data privacy focus on databases.

- Medical records.

Privacy models:

- k-anonymity (for ID association) [Sweeney02]: each item in D is identical to $k-1$ other items.
- Differential privacy [Dwork06]: add noises to query results $Q[D]$ s.t. $\forall p \in D$

$$\frac{\text{Prob}\{Q[D] = r\}}{\text{Prob}\{Q[D \setminus p] = r\}} \leq e^{\varepsilon}$$

Our focus: **structured data**.

Privacy of Structured Data

Strutured data: sensor data, location/trajectory data.

- Strong spatial/temporal correlations.
- Data items are not independent, unlike database records.

Privacy of Structured Data

Strutured data: sensor data, location/trajectory data.

- Strong spatial/temporal correlations.
- Data items are not independent, unlike database records.

Implications: personally identifying information could be revealed through innocent records.

- Sparse sensing: reduce the amount of data collected yet still support target applications.

Privacy of Structured Data

Strutured data: sensor data, location/trajectory data.

- Strong spatial/temporal correlations.
- Data items are not independent, unlike database records.

Implications: personally identifying information could be revealed through innocent records.

- Sparse sensing: reduce the amount of data collected yet still support target applications.
- Advanced network-based attacks: using explicitly the connections of data items.

Outline

- Privacy issues in trajectory data: 3 clustering algorithms.
  - Distributed r-gather problem.
  - Sensing path topology.
  - Sensing popular paths.
- Network based attacks.
  - Trajectory inference through colocation events.
  - Network de-anonymization.

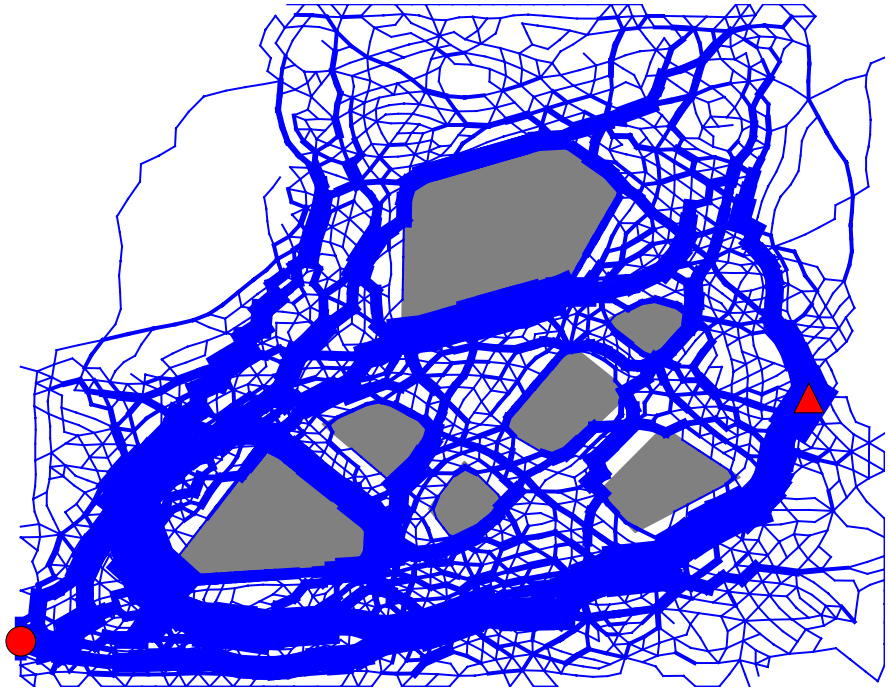
Part I: Location Privacy in Mobility Data.

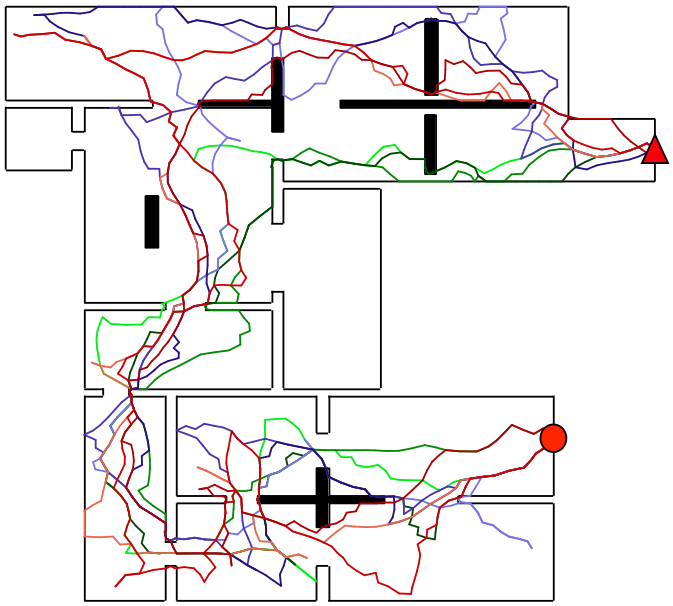
Populated by Wireless Devices

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.

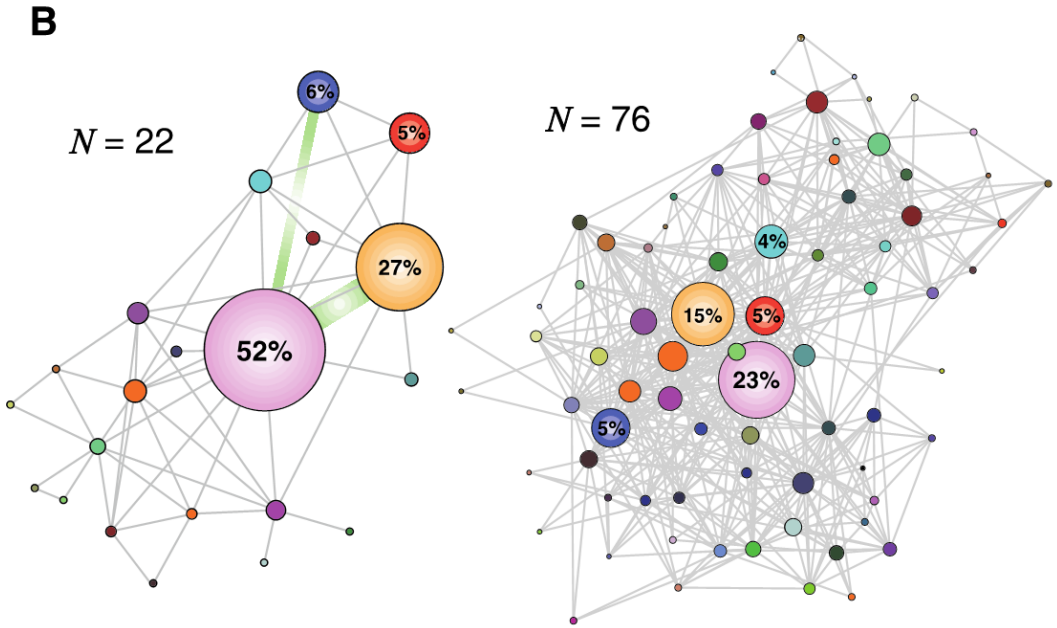




Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Frequently visited locations: home or work addresses. Predictability of location $\geq 93\%$

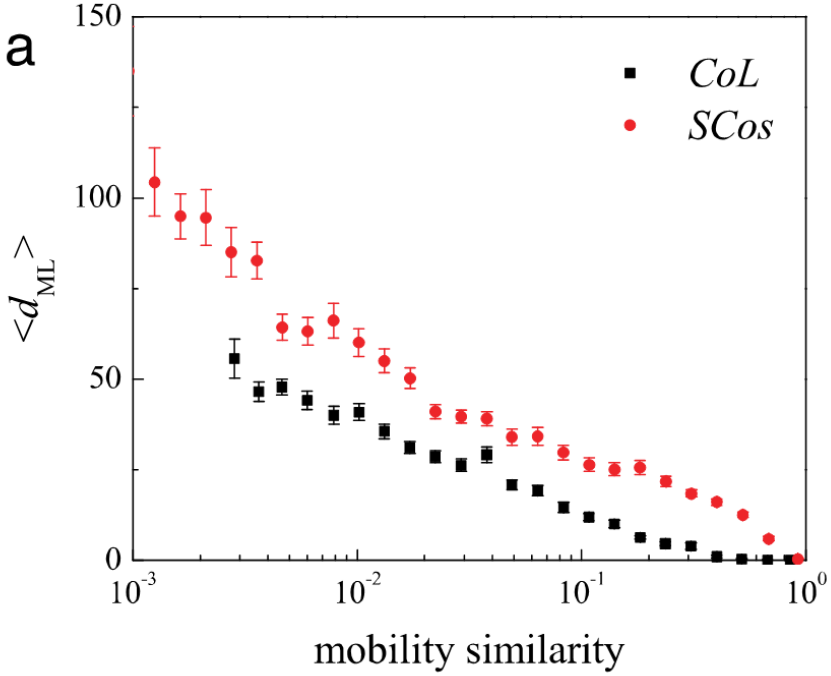


Limits of Predictability in Human Mobility, Science, 2010.

Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Frequently co-located pairs: social ties.

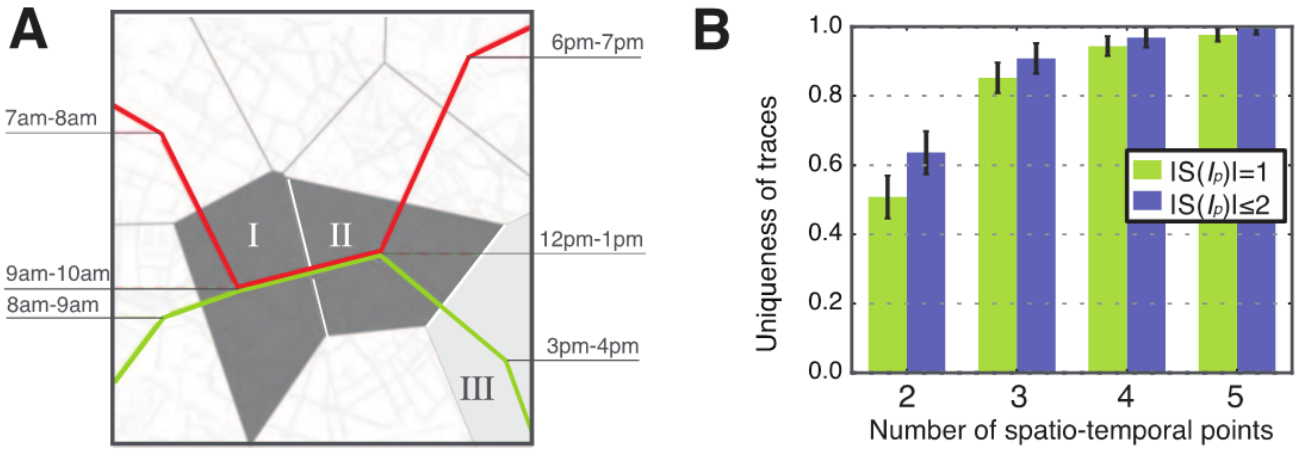


Human Mobility, Social Ties, and Link Prediction, KDD'11.

Location Privacy

Location and trajectory data can reveal personal sensitive information:

- Unique signatures: 4 spatial temporal data points can be used to identify a user from a database of 1.3 million users.

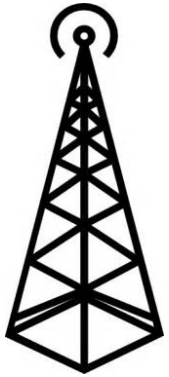


Unique in the Crowd: The privacy bounds of human mobility, Nature, 2013.

Our Approach: Application Dependent Sketches

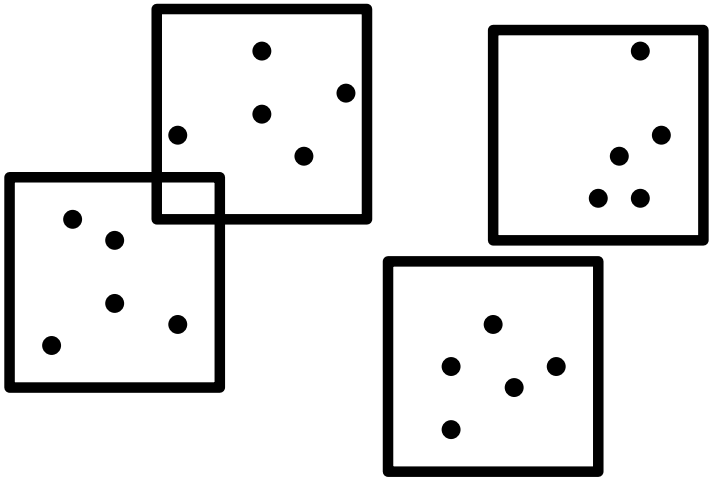
Smart city environment: many checkpoints (WiFi APs, cell towers, surveillance cameras) that record user appearances.
Goal: Collect data sketches for data mining applications.

- Distributed;
- Low cost;
- Privacy guarantee.



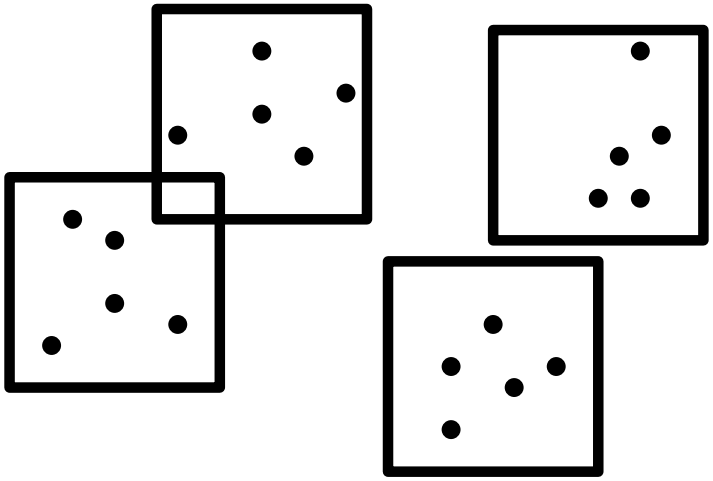
Clustering Mobile Nodes with k-anonymity

Spatial cloaking for location based queries: Report a box which contains $\geq k$ users.



Clustering Mobile Nodes with k-anonymity

Spatial cloaking for location based queries: Report a box which contains $\geq k$ users.



r-gather problem.

- each cluster has at least r nodes;
- max radius of the cluster is minimized.

Clustering Mobile Nodes with k-anonymity

r-gather in the metric setting [Aggarwal et al., Armon]:

- $r = 2$, in P, matching.
- $r > 2$, NP-hard to approximate better than 2.
- 2-approximation using network flow.

Clustering Mobile Nodes with k-anonymity

r-gather in the metric setting [Aggarwal et al., Armon]:

- $r = 2$, in P, matching.
- $r > 2$, NP-hard to approximate better than 2.
- 2-approximation using network flow.

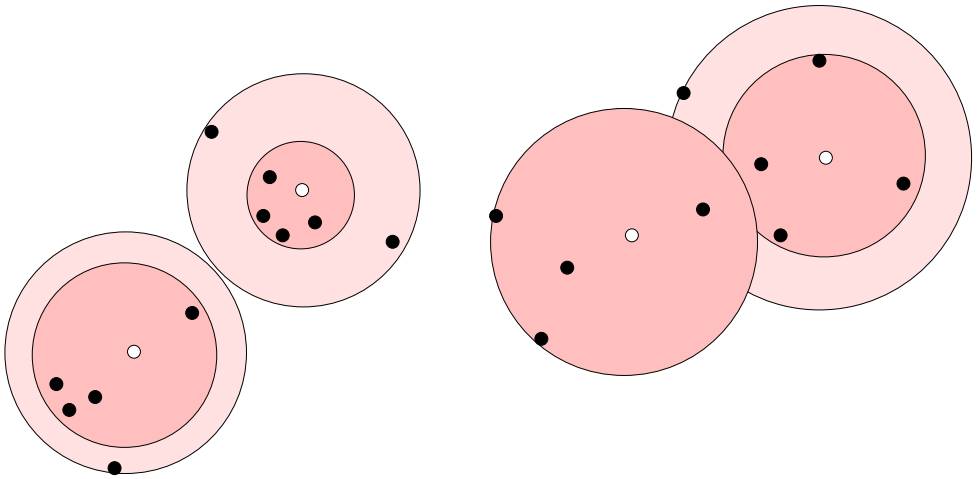
Euclidean setting:

- $r > 2$, NP-hard to app better than 1.932 for max diameter, and 1.802 for $r = 3$ ($\sqrt{35}/4 + \sqrt{3}/4$ for $r \geq 5$) for minimum enclosing ball radius. [Zeng et.al'17]
- MEB hardness: 1.938 ($r = 4$) and 1.939 ($r \geq 5$). [Johnson'17]

Clustering Mobile Nodes with k-anonymity

Kinetic r-gather:

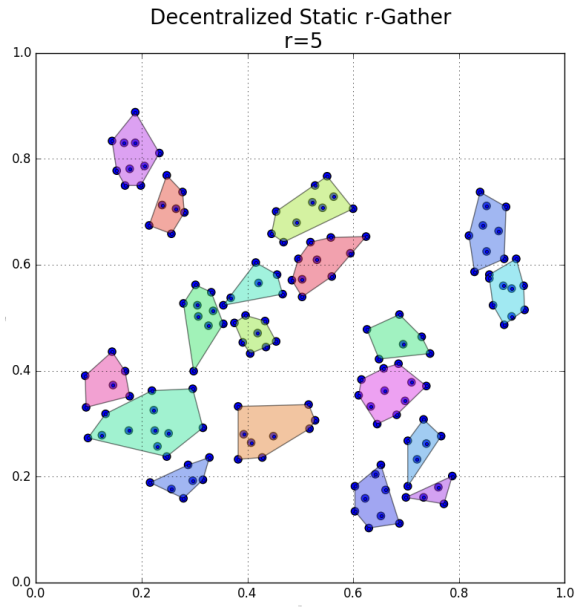
- Compute r-NN graph.
- Find maximal independent set
- Assign remaining nodes to nearest cluster.

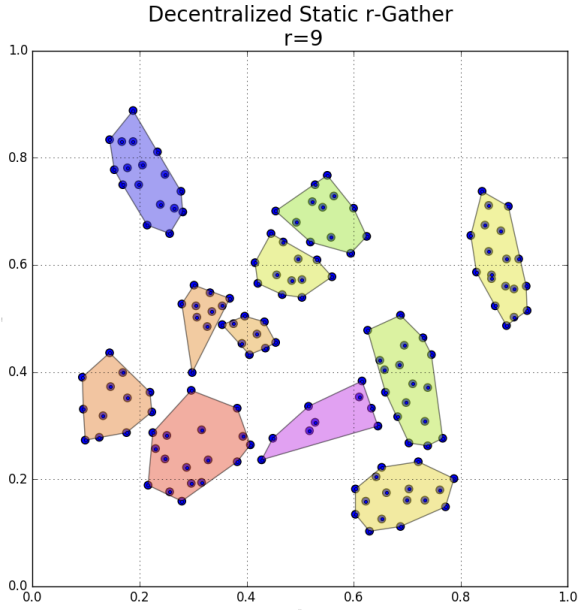


Clustering Mobile Nodes with k-anonymity

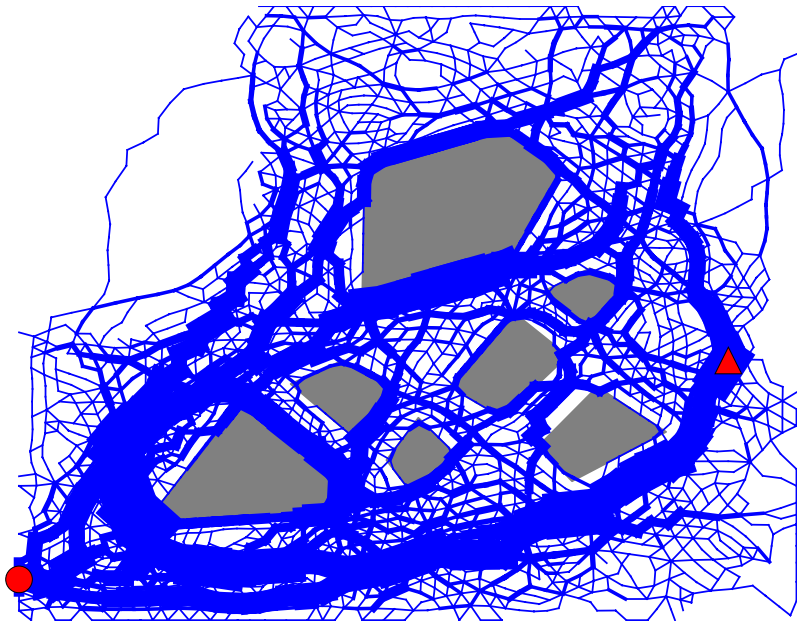
Kinetic r-gather:

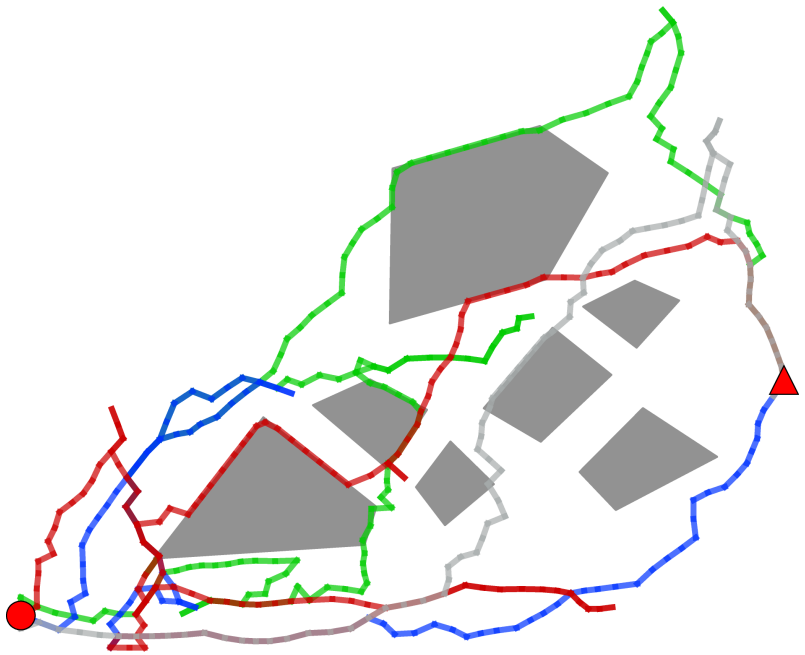
- 4-approximation: $d(p, C(p)) \leq 2d(p, \mathrm{rNN}(p))$.
- # changes: $O(n^2)$ for poly motion.
- Distributed algorithm.





Clustering Trajectories by Topology

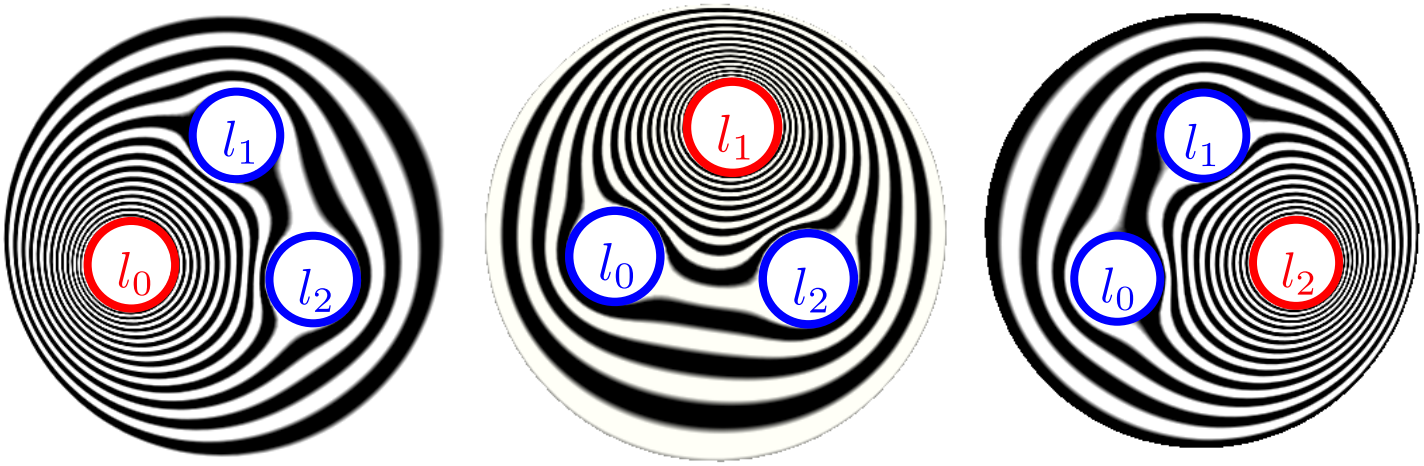




Clustering Trajectories by Topology [Yin et. al'15]

Directly sense topology by using harmonic differential forms:

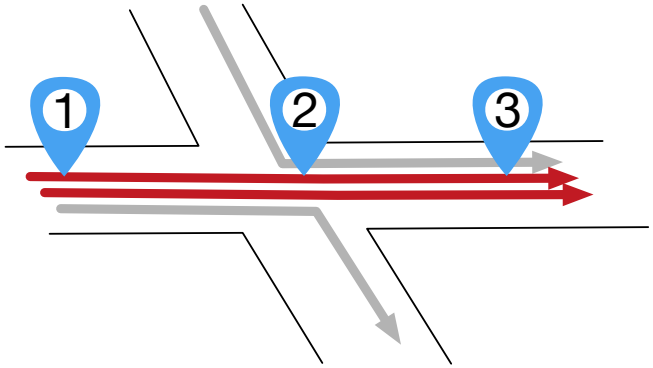
- Checkpoints maintain a triangulation.
- Edges are given directed weights – "differential forms"
- Summation of weights along any cycle not enclosing holes $= 0$
- Trajectories that sum up to different values have different homology.



Mining Popular Paths in Traffic

Popular Paths: paths travelled by ϕ-fraction of all vehicles that appear on the path.

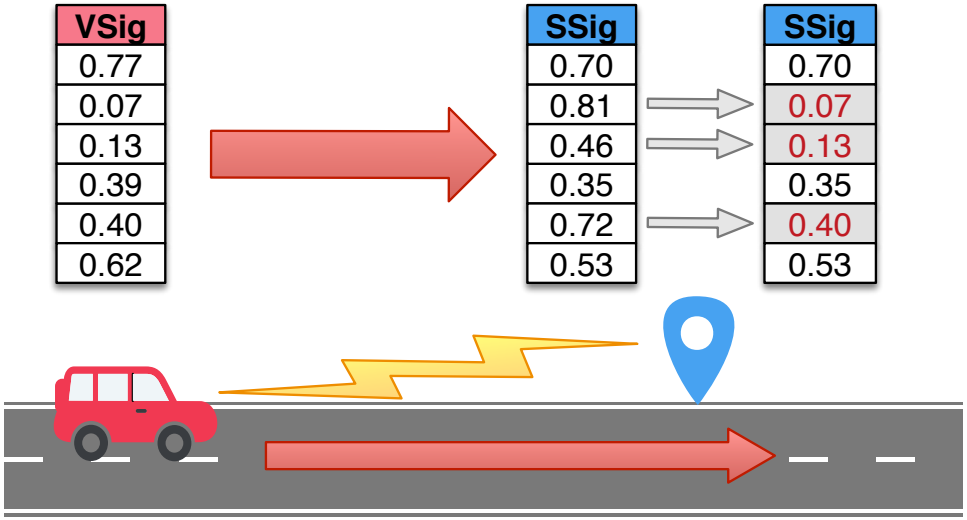
- Popular paths are user prefered 'high quality' paths.
- Opportunity for public transportation.



Question: Can we learn popular paths without collecting all traffic data?

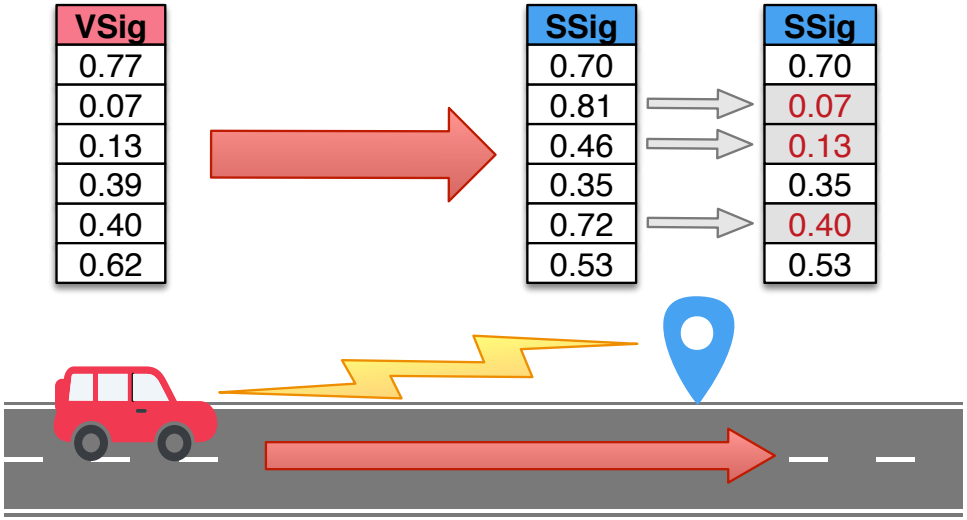
MinHash Signature [Ding et. al'17]

Sensor i sees a set of vehicles V, and stores the min hash value $h_i(V)$, for k hash functions $\{h_i\}$.



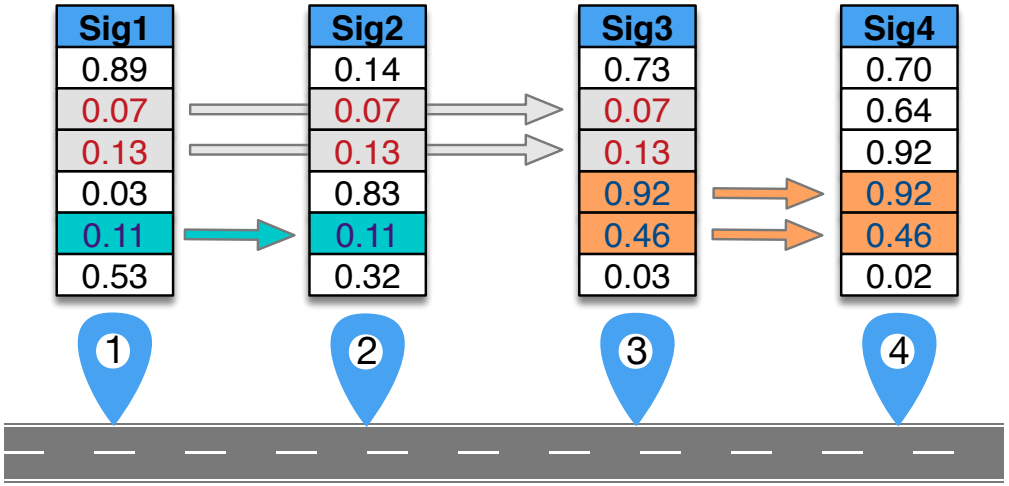
MinHash Signature [Ding et. al'17]

Sensor i sees a set of vehicles V, and stores the min hash value $h_i(V)$, for k hash functions $\{h_i\}$.

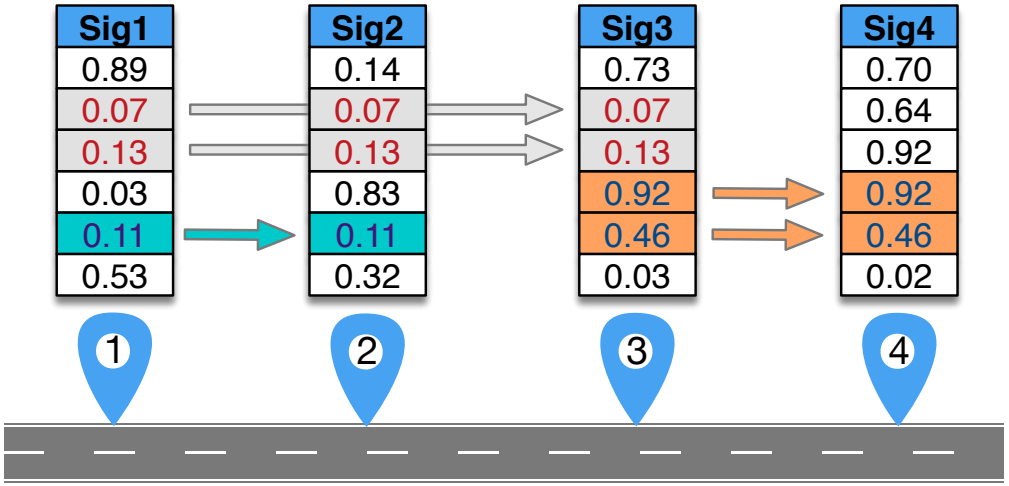


- Storage per sensor: $O(k)$.
- Easy to update.

MinHash Signature

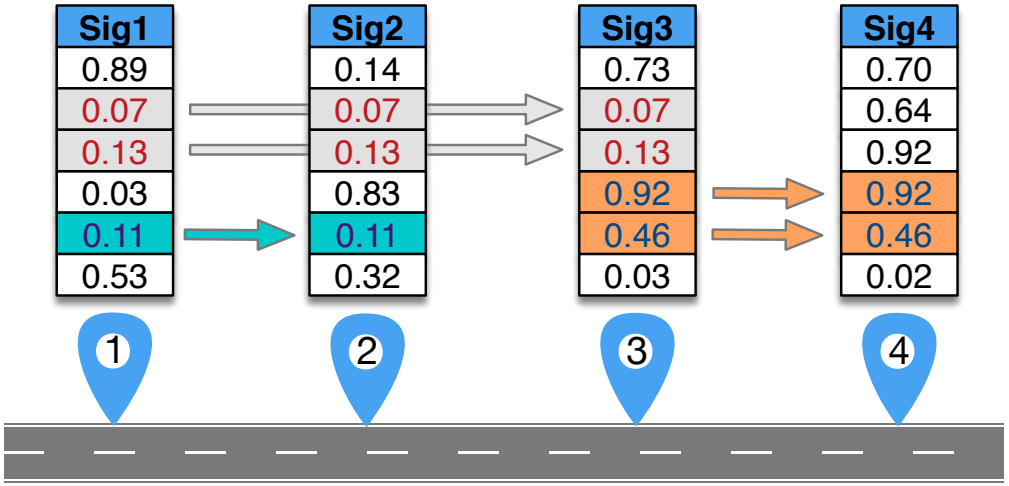


MinHash Signature



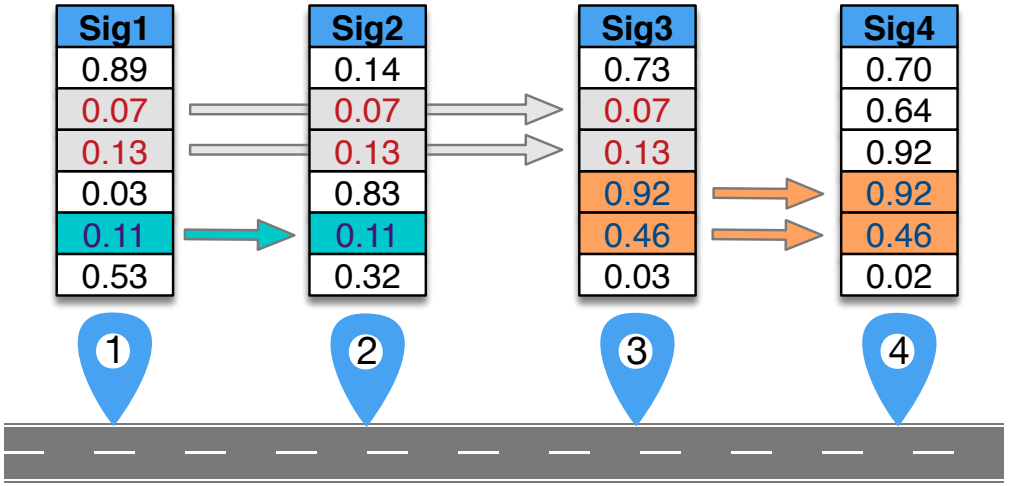
- Magnitude of min hash value implies $|V|$.

MinHash Signature



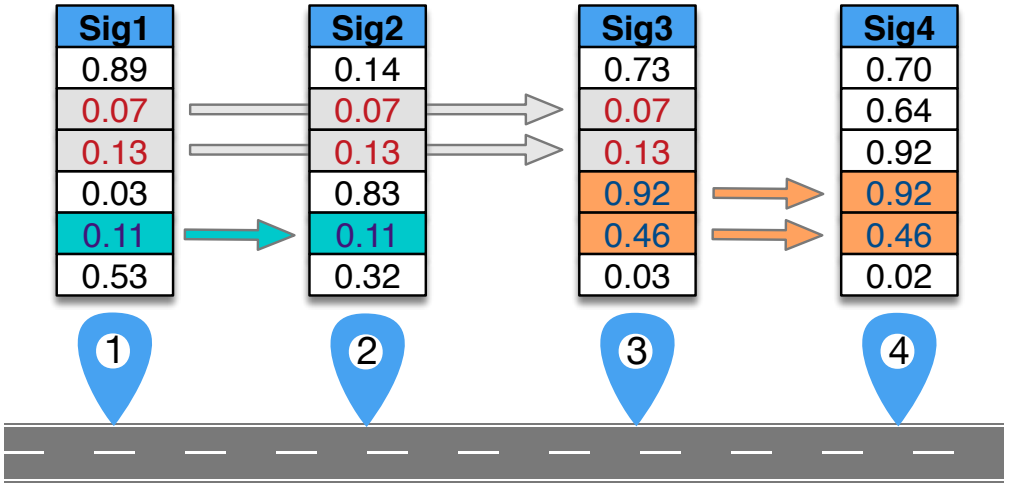
- Magnitude of min hash value implies $|V|$.
- Estimate set intersection $|V_1 \cap V_2|$ by # shared min hash entries.

MinHash Signature



- Magnitude of min hash value implies $|V|$.
- Estimate set intersection $|V_1 \cap V_2|$ by # shared min hash entries.
- Privacy protection: removal of a vehicle likely does not change the min hash signatures, upon random seeds.

MinHash Signature



- Magnitude of min hash value implies $|V|$.
- Estimate set intersection $|V_1 \cap V_2|$ by # shared min hash entries.
- Privacy protection: removal of a vehicle likely does not change the min hash signatures, upon random seeds.
- Support efficient (polylogarithmic) queries for popular paths.

Part II: Network based Privacy Attacks.

Network Based Location Attacks [Aronov et.al.'18]

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

Network Based Location Attacks [Aronov et.al.'18]

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

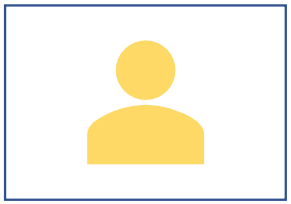
Question: if privacy insensitive users publish their whereabouts, how much information can we infer for privacy sensitive users?

Location Inference

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, τ, p): user i is at location p at time τ.

Problem Definition

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

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

Problem Definition

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

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

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

A user i has a maximum travel speed v_i.

Two events (i, j, τ) and (i, j', τ') 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, τ, p): user i is at location p at time τ.
- Meeting events $\chi = (i, j, \tau)$: user i, j appear at the same location at time τ.

A user i has a maximum travel speed v_i.

Two events (i, j, τ) and (i, j', τ') 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)?

Results

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

Results

- Solving R is a convex problem – all distance constraints are convex.
- If we measure distane by ℓ_∞, we can separate different dimensions and obtain a nearly linear running time algorithm.

Results

- Solving R is a convex problem – all distance constraints are convex.
- If we measure distane by ℓ_∞, we can separate different dimensions and obtain a nearly linear running time algorithm.
- If we use ℓ_2 distance, the region boundary may have high complexity, e.g., depending on $O(n)$ parameters.

Results

- Solving R is a convex problem – all distance constraints are convex.
- If we measure distane by ℓ_∞, we can separate different dimensions and obtain a nearly linear running time algorithm.
- If we use ℓ_2 distance, the region boundary may have high complexity, e.g., depending on $O(n)$ parameters.
- But we can compute ε-approximations in nearly linear running time.

Results

- Solving R is a convex problem – all distance constraints are convex.
- If we measure distane by ℓ_∞, we can separate different dimensions and obtain a nearly linear running time algorithm.
- If we use ℓ_2 distance, the region boundary may have high complexity, e.g., depending on $O(n)$ parameters.
- But we can compute ε-approximations in nearly linear running time.
- If we enforce speed lower bound, the problem is $\exists R$-complete.

Results

- Solving R is a convex problem – all distance constraints are convex.
- If we measure distane by ℓ_∞, we can separate different dimensions and obtain a nearly linear running time algorithm.
- If we use ℓ_2 distance, the region boundary may have high complexity, e.g., depending on $O(n)$ parameters.
- But we can compute ε-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 $1,847 km^2$ 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, feasiblity region has height about 1.6km.

Network Alignment – De-anonymization [Ni et. al 18]

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 [Ni et. al 18]

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 [Ni et. al 18]

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 [Ni et. al 18]

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.

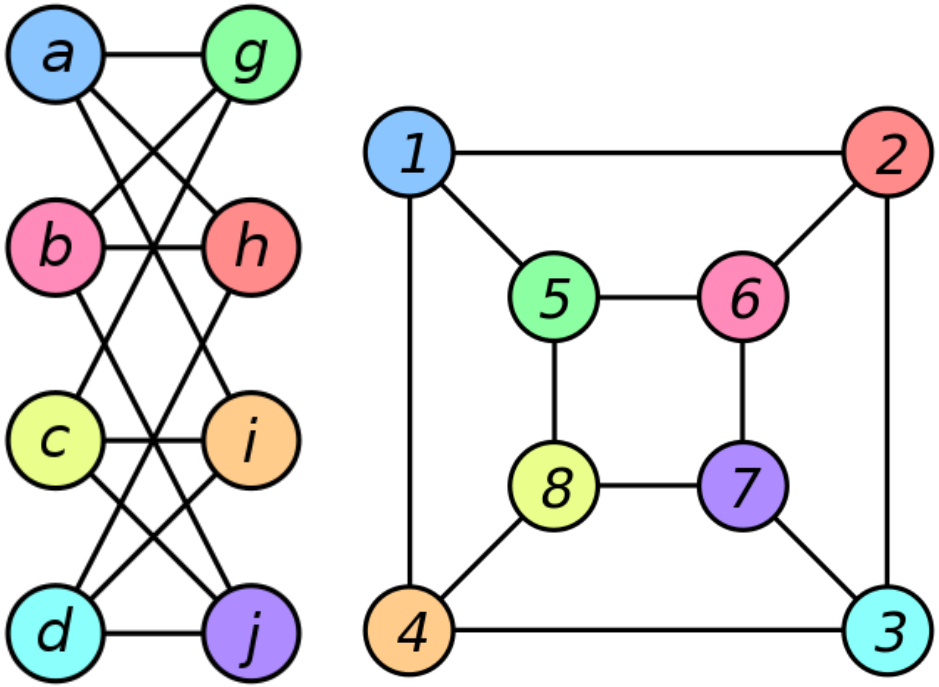
De-anonymization: If we can align the two networks by vertex correspondences, the identifies of the private network are thus revealed.

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.

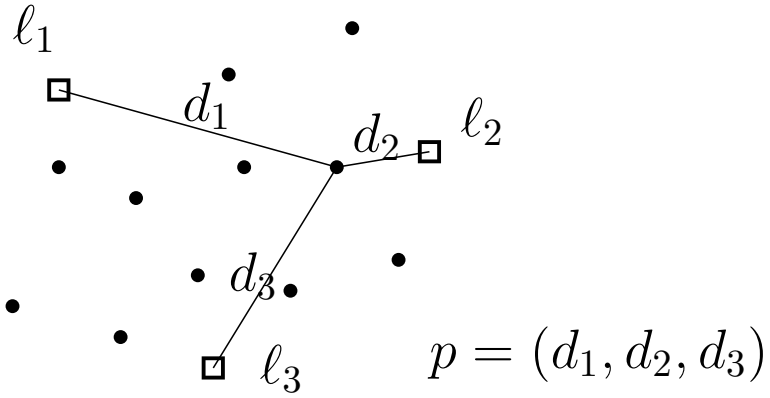
Graph Isomorphism

One of the most fundamental problems in theoretical computer science.

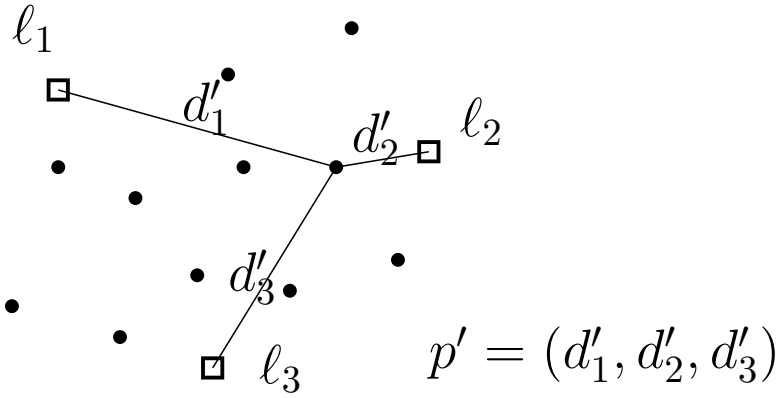
- In NP.
- Nov 2015/Jan 2017, László Babai claimed quasi-polynomial time algorithm: $O(\exp(\log^{O(1)} n))$.
- Many practical algorithms: e.g., NAUTY.
- Subgraph isomorphism is NP-complete.
- Approximate graph isomorphism: find the best correspondence between vertices in G_1 and G_2 s.t. if u, v are 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 ℓ_i are already aligned?



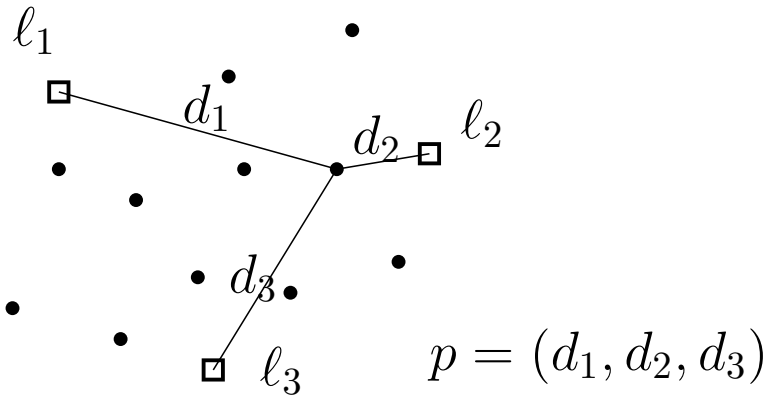
$p = (d_1, d_2, d_3)$

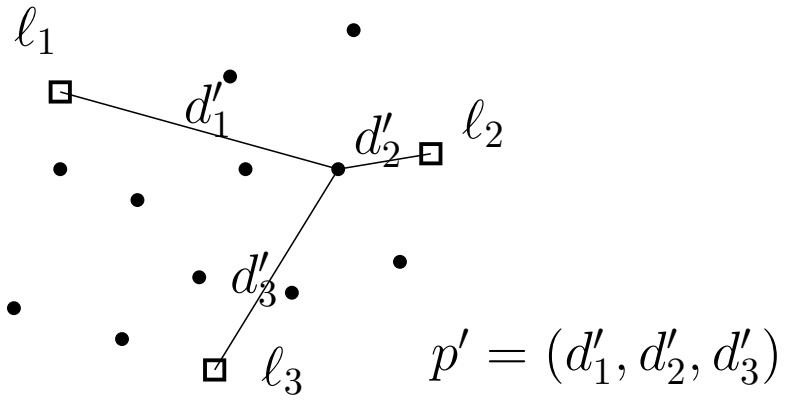


$p' = (d'_1, d'_2, d'_3)$

Our Solution: A Geometric Approach

How to align two sets of points in the plane, assuming that some landmarks ℓ_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 ℓ_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.

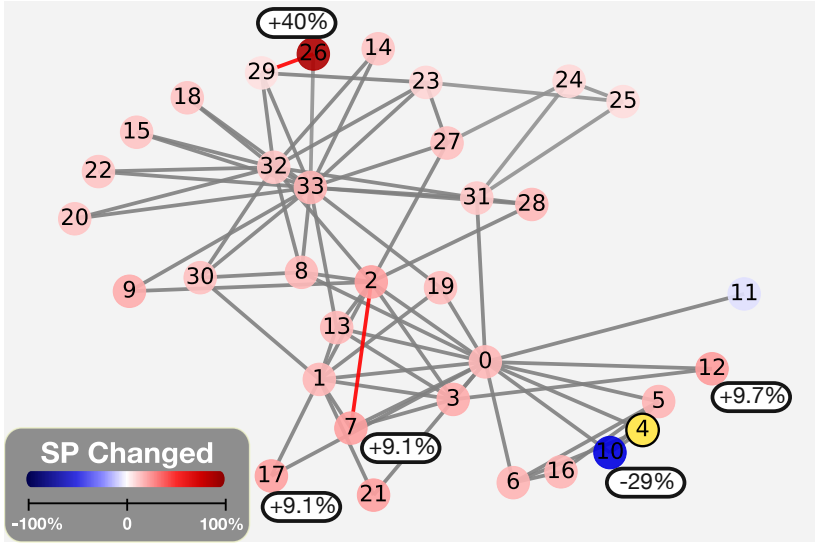
Q: What distance to use?

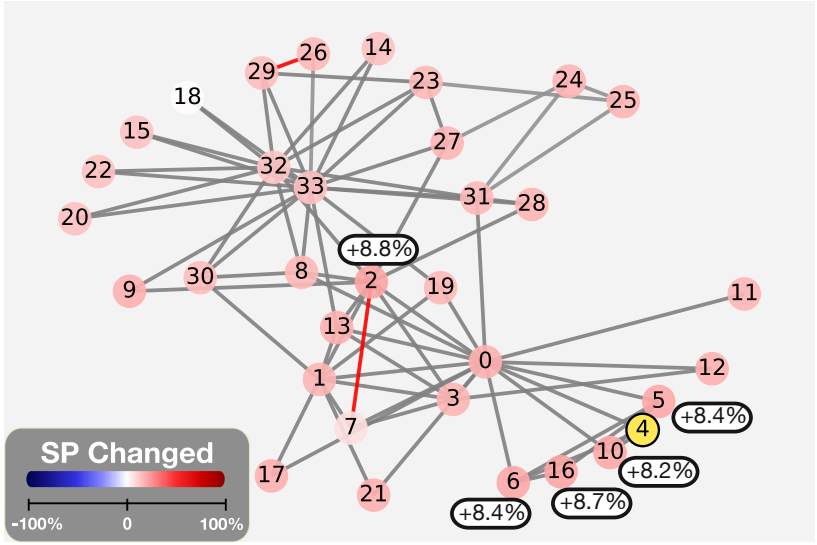
- Tie strength – Trouble: not easy to measure.
- Count # hops to these landmarks – Trouble: small world property;
- Distances from some geometric embedding (spectral embedding, Tutte embedding).

Q: Robust to noises (edge insertion/deletion)?

Robustness: Remove Two Edges

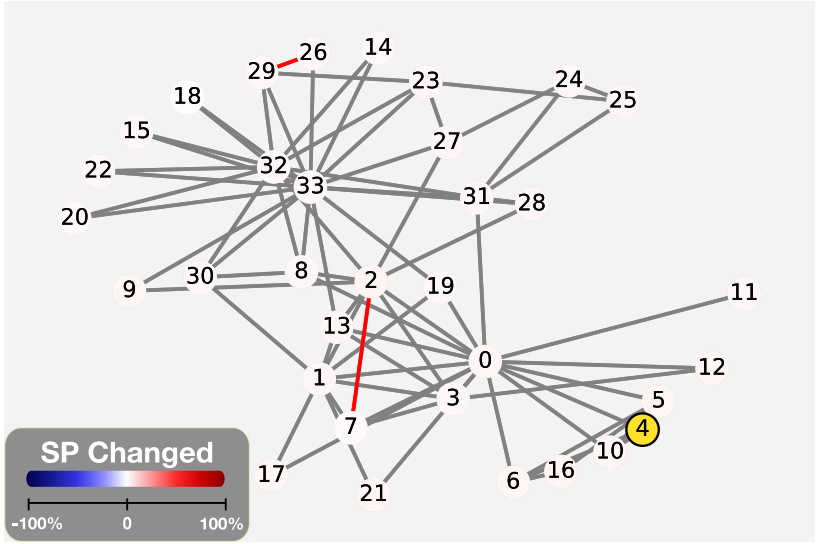
Left: Spectral embedding; Right: Tutte/Spring embedding.

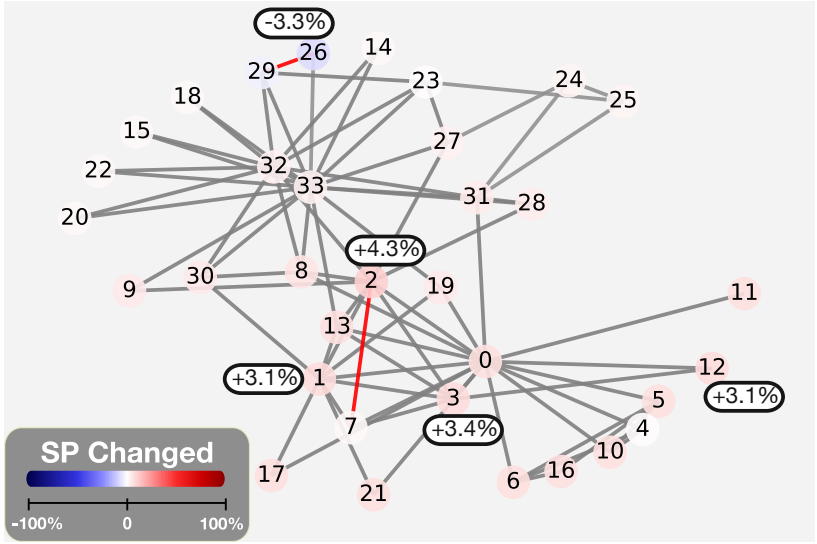




Robustness: Remove Two Edges

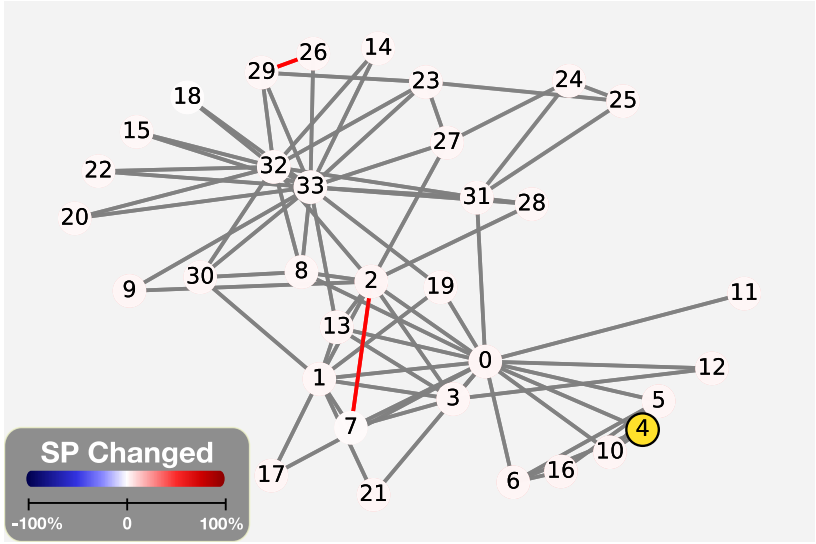
Left: Hop count; Right: our metric.

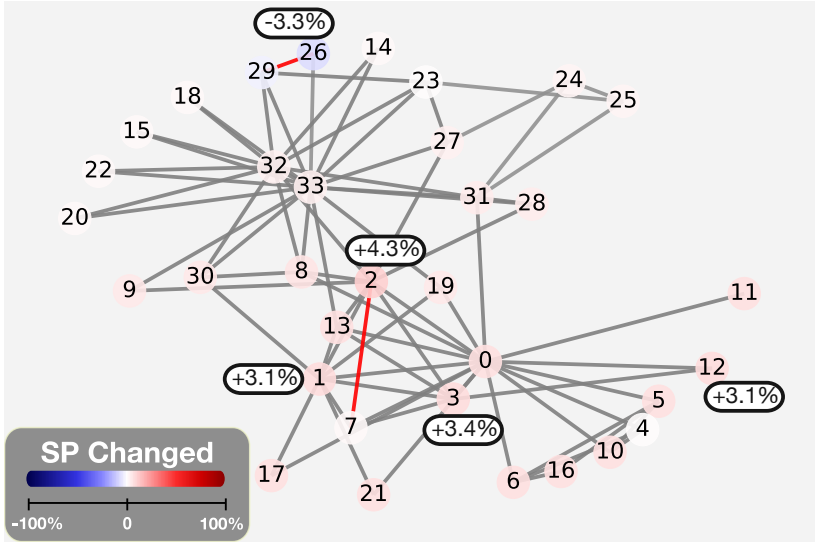




Robustness: Remove Two Edges

Left: Hop count; Right: our metric.





Q: How is our metric defined?

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 μ_1 and μ_2 on X:

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

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

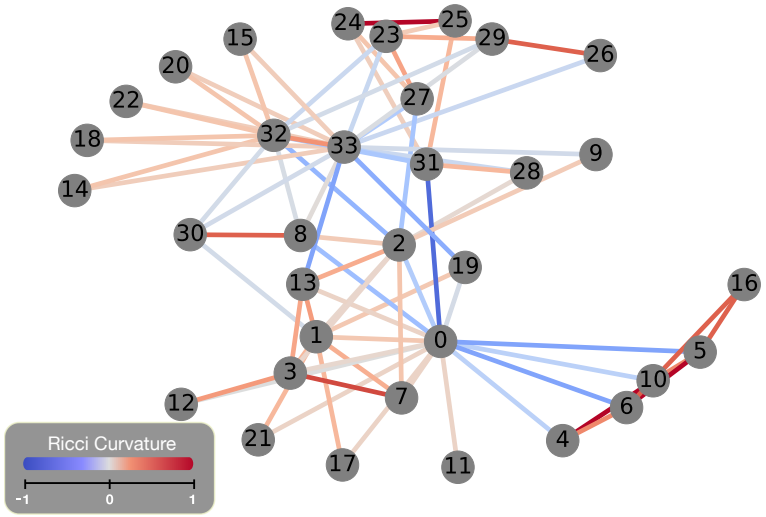
$$d_{i+1}(x, y) = (d_i(x, y) - \varepsilon \cdot \kappa_i(x, y) \cdot d_i(x, y)) \cdot N$$

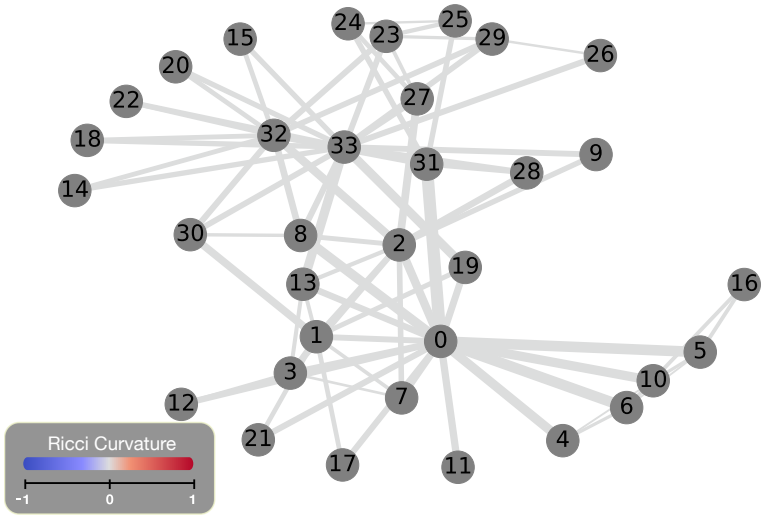
Until convergence, where N is to rescale to make sure total edge weights remain the same.

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

Ricci Flow Metric

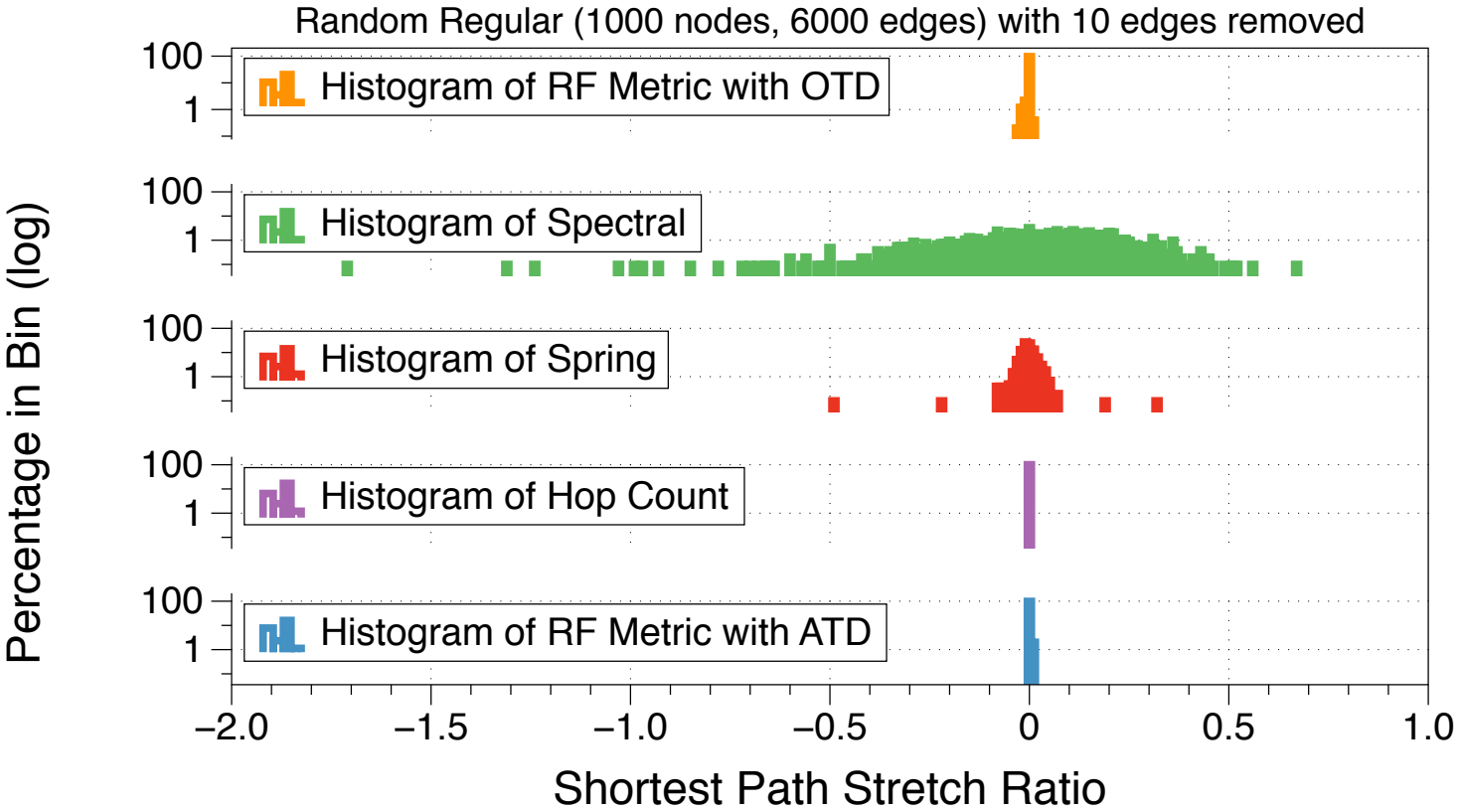
Intuition: flatten the network – shrink an edge if it is within a well connected community; stretch an edge if otherwise, s.t., the network curvature is uniform everywhere.





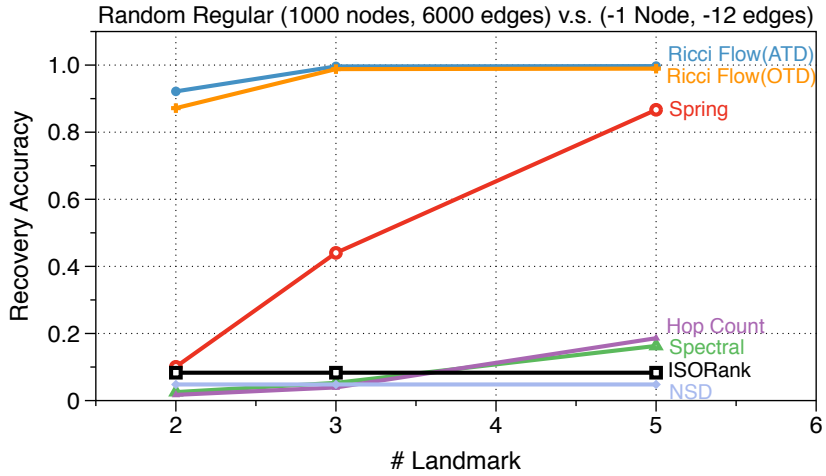
Evaluation on Resilience

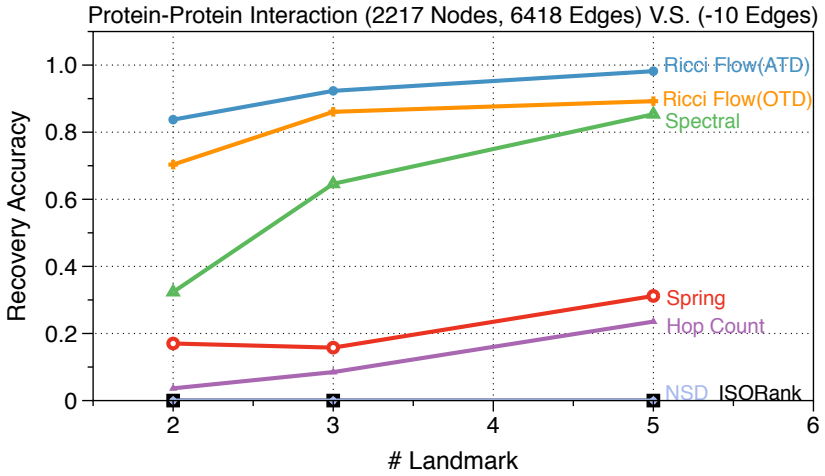
Randomly remove 10 edges in a random regular graph.



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.





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

- Boris Aronov, Alon Efrat, Ming Li, Joseph S. B. Mitchell, Valentin Polishchuk, Boyang Wang, Hanyu Quan, Xianfeng David Gu, Rik Sarkar, Matt Johnson.
- Chien-Chun Ni, Yu-Yao Lin, Jiaxin Ding, Jiemin Zeng, Gaurish Telang, Mengyu Zhou, Dengpan Zhou, Xiaotian Yin.

- Questions and comments?