Quantifying Privacy Loss of Human Mobility Graph Topology

Dionysis Manousakas[*], Cecilia Mascolo[*,†], Alastair R. Beresford[*], Dennis Chan[*], Nikhil Sharma[‡]

[*]University of Cambridge
[†]The Alan Turing Institute
[‡]UCL

The 18th Privacy Enhancing Technologies Symposium
July 24–27, 2018

Mobility data privacy vs. utility

- Information sharing for data-driven customization and large-scale analytics
  - context-awareness
  - transportation management, health studies, urban development
- **Utility**-preserving anonymized data representations
  - timestamped GPS, CDR, etc. measurements
  - histograms
  - heatmaps
  - **graphs**
- How **privacy** conscientious they are?
  - often poorly understood, leading to privacy breaches

Mobility data privacy vs. utility

- Information sharing for data-driven customization and large-scale analytics
  - context-awareness
  - transportation management, health studies, urban development

- **Utility**-preserving anonymized data representations
  - timestamped GPS, CDR, etc. measurements
  - histograms
  - heatmaps
  - **graphs**

- How **privacy** conscientious they are?
  - often poorly understood, leading to privacy breaches

Mobility data privacy vs. utility

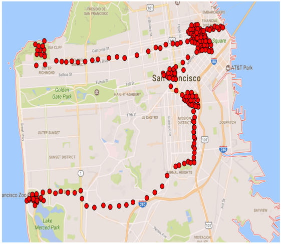
- Information sharing for data-driven customization and large-scale analytics
  - context-awareness
  - transportation management, health studies, urban development
- **Utility**-preserving anonymized data representations
  - timestamped GPS, CDR, etc. measurements
  - histograms
  - heatmaps
  - **graphs**
- How **privacy** conscientious they are?
  - often poorly understood, leading to privacy breaches

Mobility data privacy vs. utility

- Information sharing for data-driven customization and large-scale analytics
  - context-awareness
  - transportation management, health studies, urban development
- **Utility**-preserving anonymized data representations
  - timestamped GPS, CDR, etc. measurements
  - histograms
  - heatmaps
  - **graphs**
- How **privacy** conscientious they are?
  - often poorly understood, leading to privacy breaches

Deanonymizing mobility

Raw mobility data



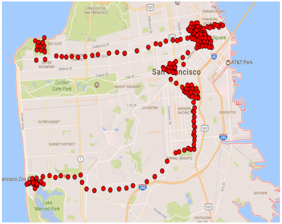
Inference on **individual** traces information

(1) Sparsity and regularity-based

- "top-*N*" location attacks [Zang and Bolot, 2011]
- unicity of spatio-temporal points [de Montjoye et al., 2013]
- matching of individual mobility histograms [Naini et al., 2016]

Deanonymizing mobility

Raw mobility data



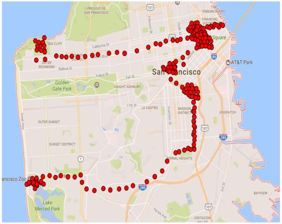
Inference on **individual** traces information

(2) Probabilistic models

- Markovian mobility models [De Mulder et al., 2008]
- Mobility Markov chains [Gambs et al., 2014]

Deanonymizing mobility

Raw mobility data

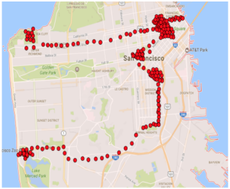


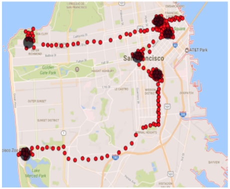
Inference on **population** statistics

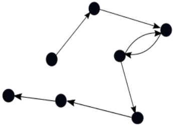
(3) On aggregate information

- Individual trajectory recovery from aggregated mobility data [Xu et al., 2017]
- Probabilistic inference on aggregated location time-series [Pyrgelis et al., 2017]

Mobility representations





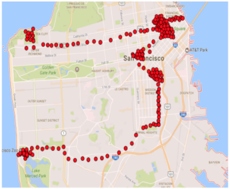


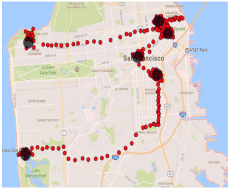
raw mobility data

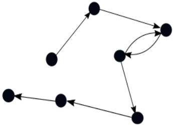
sequences of pseudonymised regions of interest

e.g. MDC research track, Device Analyzer

Mobility representations



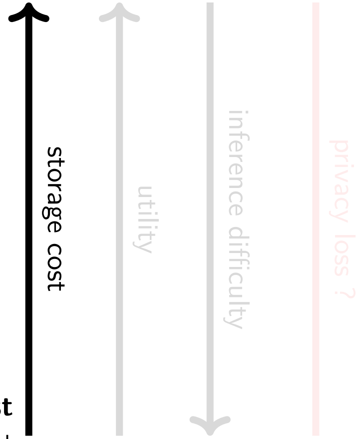




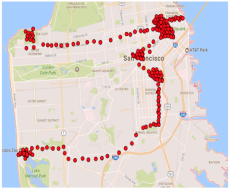
raw mobility data

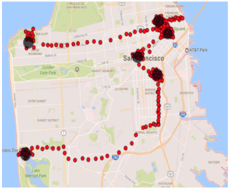
sequences of pseudonymised regions of interest

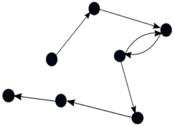
e.g. MDC research track, Device Analyzer



Mobility representations







raw mobility data

sequences of pseudonymised regions of interest

e.g. MDC research track, Device Analyzer

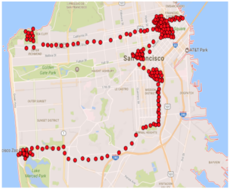
storage cost

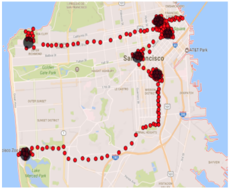
utility

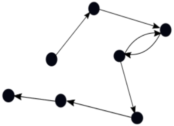
inference difficulty

privacy loss ?

Mobility representations







raw mobility data

sequences of pseudonymised regions of interest

e.g. MDC research track, Device Analyzer

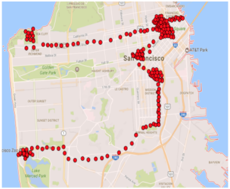
storage cost

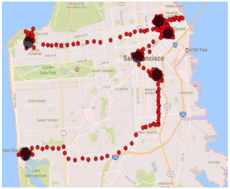
utility

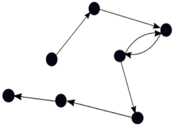
inference difficulty

privacy loss ?

Mobility representations







raw mobility data

sequences of pseudonymised regions of interest

e.g. MDC research track, Device Analyzer

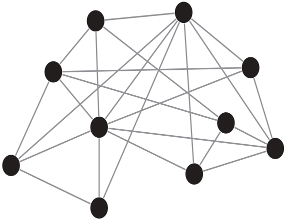
storage cost

utility

inference difficulty

privacy loss ?

Motivation

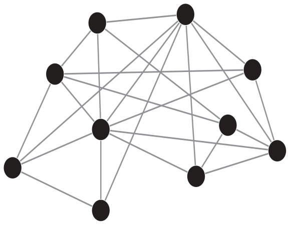


Let's remove

- temporal (except from *ordering* of states)
- geographic, and
- cross-referencing information

– What is the privacy leakage of this representation?
– Does *topology* still bear identifiable information?
– Can an adversary exploit it in a deanonymization attack?

Motivation

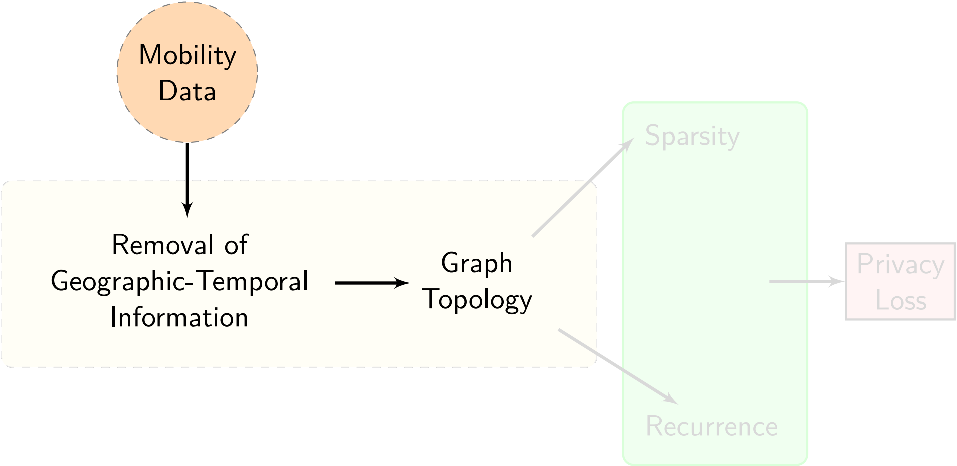


Let's remove

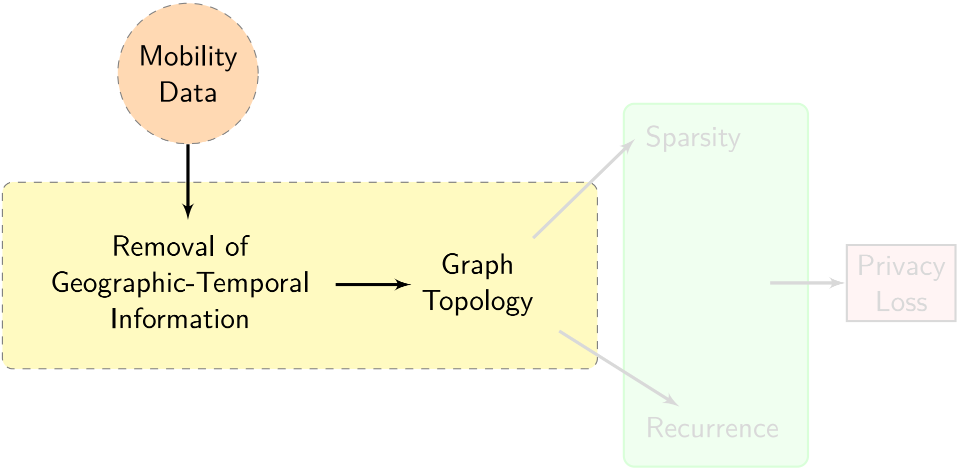
- temporal (except from *ordering* of states)
- geographic, and
- cross-referencing information

– What is the privacy leakage of this representation?
– Does *topology* still bear identifiable information?
– Can an adversary exploit it in a deanonymization attack?

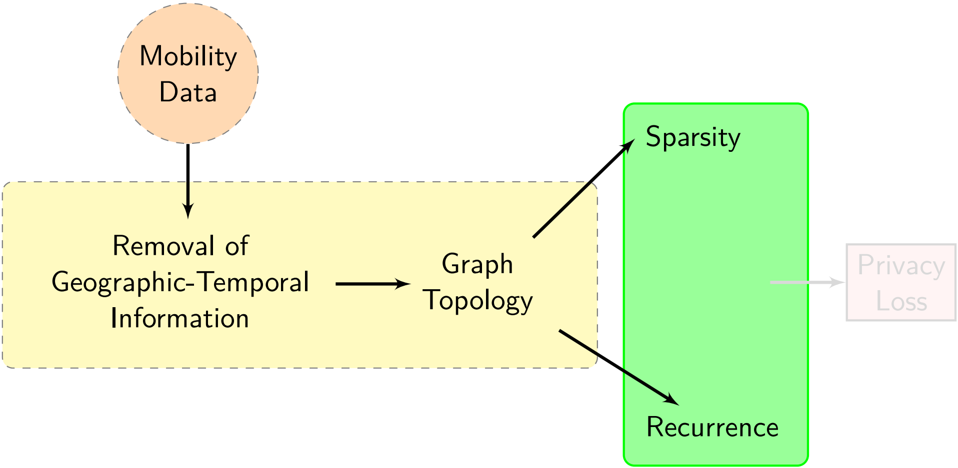
Mobility information flow



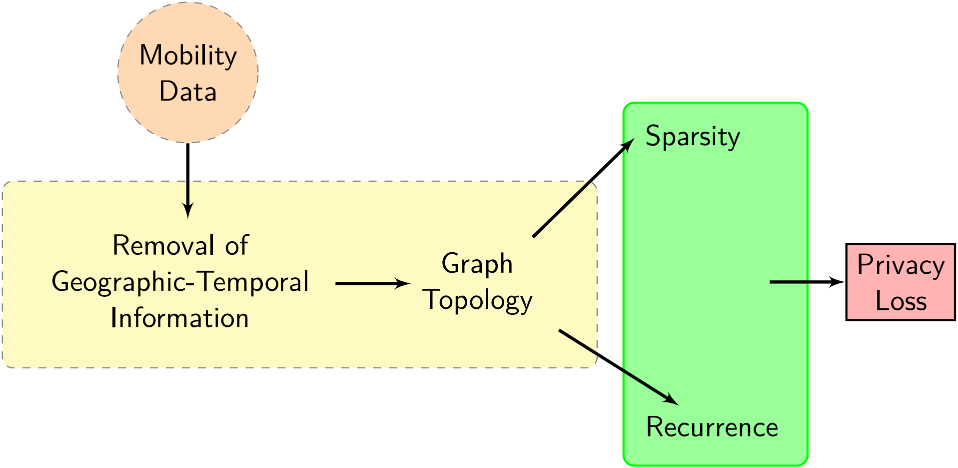
Mobility information flow



Mobility information flow



Mobility information flow



Differences of our approach

Mobility deanonymization

- **No cross-referencing** between locations
- **No fine-grained temporal information** (as opposed to [Lin et al., 2015])

Privacy on graphs

- **Each user**'s information is an **entire graph**: No need for node matching [Narayanan and Shmatikov, 2008, Sharad and Danezis, 2014]
- **No social network information**

Data

- **Device Analyzer** : global dataset from mobile devices with system information, cellular and wireless location
- **1500 users** with the most cid location datapoints
  - average of 430 days of observation,
  - 200 regions of interest
- cids pseudonymized per handset

Mobility networks

Graphs with nodes corresponding to ROIs and edges to recorded transitions between ROIs

- **Network Order Selection** via Markov chain modeling of sequential data [Scholtes, 2017]
- **Node attributes** with no temporal/geographic information
- **Edge weights** corresponding to frequency of transitions
- Location pruning to **top**$-N$ **networks** by keeping the most frequently visited regions in user's routine

Empirical statistics

Graphs with:

- heavy-tailed degree distributions
- large number of rarely repeated transitions
- small number of frequent transitions
- high recurrence rate

Privacy framework

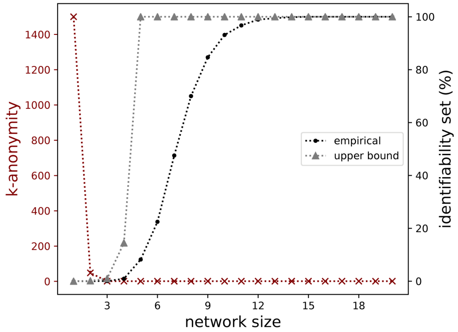
$k-$***anonymity*** **via** ***graph isomorphism***

Graph $k-$anonymity

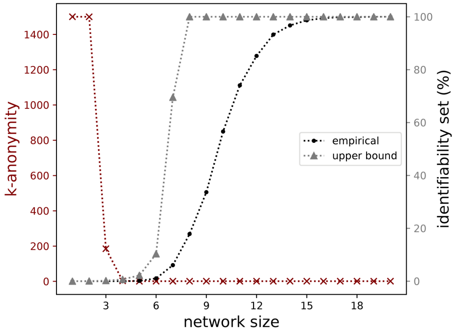
is the minimum cardinality of isomorphism classes within a population of graphs

[Sweeney, 2002]

Identifiability of top−N mobility networks



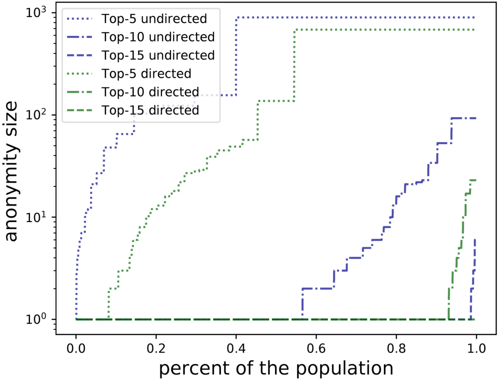
directed



undirected

- $\mathbf{15}$ and $\mathbf{19}$ locations suffice to form uniquely identifiable **directed** and **undirected** networks
- $\mathbf{5}$ and $\mathbf{8}$ are the corresponding theoretical upper bounds

Anonymity size of top$-N$ mobility networks



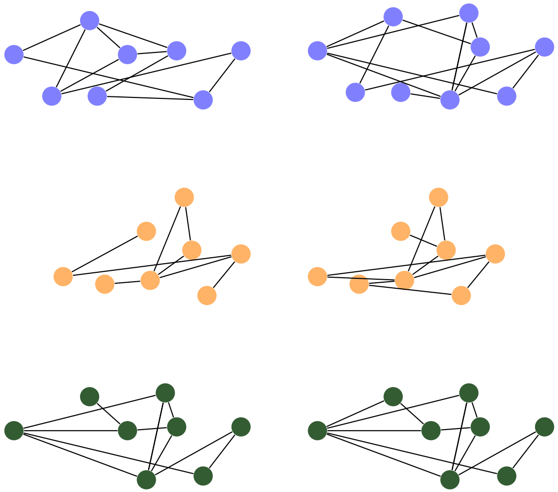
- small isomorphism clusters for even very few locations
- median anonymity becomes one for network sizes of 5 and 8 in directed and undirected networks respectively

Recurring patterns in typical user's mobility

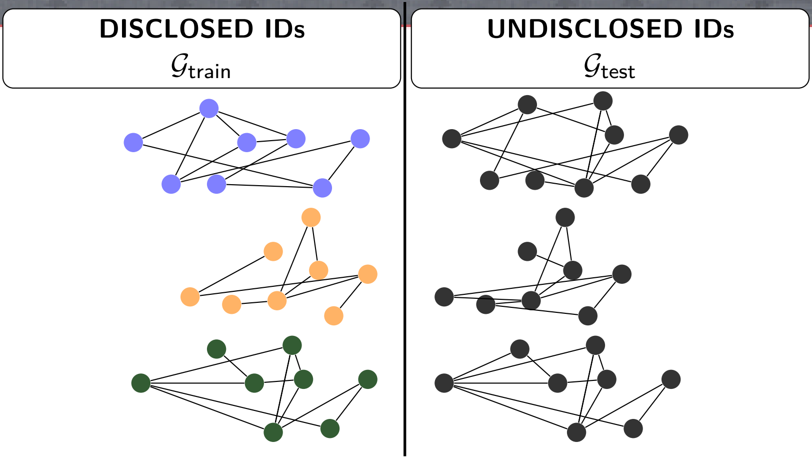
1st half of the observation period   2nd half of the observation period

shown edges correspond to the 10% most frequent transitions in the respective observation window

Threat Model

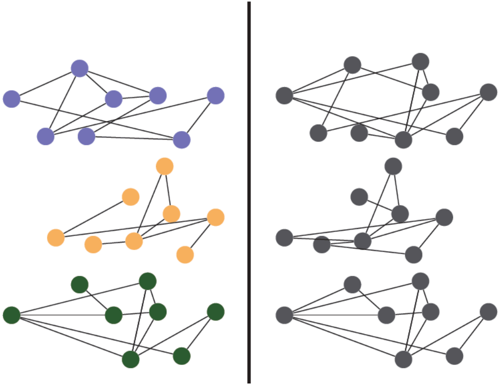


Threat Model

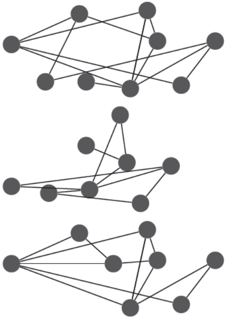


- closed-world
- partition point for each user randomly $\in (0.3, 0.7)$ of total obs. period
- state frequency information

Threat Model

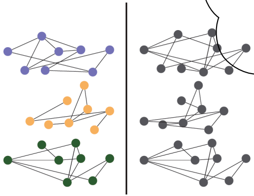


Attacks: Uninformed Adversary



$P(l_{G'} = l_{G_i}) = 1/|\mathcal{L}|,$
for every $\mathrm{G}_i \in \mathcal{G}_{\text{train}}$
expected rank$=|\mathcal{L}|/\mathbf{2}$

Attacks: Informed Adversary



$$P\big(l_{G'} = l_{G_i} \mid \mathcal{G}_{\text{train}}, K\big) \propto f\big(K(G_i, G')\big),$$

for every $G_i \in \mathcal{G}_{\text{train}}$

K : graph similarity metric,

f : non-decreasing

Attacks: Informed Adversary

- Posterior probability
  $P\big(I_{G'} = I_{G_i} | \mathcal{G}_{\text{train}}, K\big) \propto f\big(K(G_i, G')\big)$, for every $G_i \in \mathcal{G}_{\text{train}}$
- Privacy Loss

$$PL\big(G'; \mathcal{G}_{\text{train}}, K\big) = \frac{P\big(I_{G'} = I_{G'_{\text{true}}} | \mathcal{G}_{\text{train}}, K\big)}{P\big(I_{G'} = I_{G'_{\text{true}}}\big)} - 1$$

Graph Similarity Functions

Graph Kernels

Express similarity as inner product of vectors with graph statistics [Vishwanathan et al., 2010]

- on **Atomic Substructures** (e.g. Shortest-Paths, Weisfeiler-Lehman subtrees)

$$K(G, G') = \left\langle \frac{\phi(G)}{||\phi(G)||}, \frac{\phi(G')}{||\phi(G')||} \right\rangle$$

- **Deep Kernels** [Yanardag and Vishwanathan, 2015]

$$K(G, G') = \phi(G)^T \mathcal{M} \phi(G')$$

$\mathcal{M}$: encodes similarities between substructures

Graph Similarity Functions

Graph Kernels

Express similarity as inner product of vectors with graph statistics [Vishwanathan et al., 2010]

- on **Atomic Substructures** (e.g. Shortest-Paths, Weisfeiler-Lehman subtrees)

$$K(G, G') = \left\langle \frac{\phi(G)}{||\phi(G)||}, \frac{\phi(G')}{||\phi(G')||} \right\rangle$$

- **Deep Kernels** [Yanardag and Vishwanathan, 2015]

$$K(G, G') = \phi(G)^T \mathcal{M} \phi(G')$$

$\mathcal{M}$: encodes similarities between substructures

Graph Similarity Functions

Graph Kernels

Express similarity as inner product of vectors with graph statistics [Vishwanathan et al., 2010]

- on **Atomic Substructures** (e.g. Shortest-Paths, Weisfeiler-Lehman subtrees)

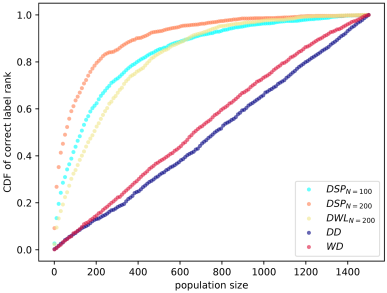
$$K(G, G') = \left\langle \frac{\phi(G)}{||\phi(G)||}, \frac{\phi(G')}{||\phi(G')||} \right\rangle$$

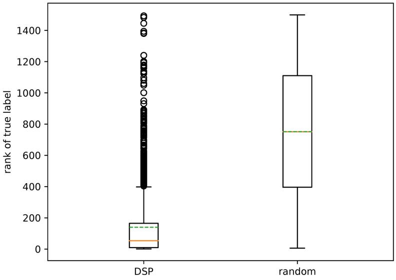
- **Deep Kernels** [Yanardag and Vishwanathan, 2015]

$$K(G, G') = \phi(G)^T \mathcal{M} \phi(G')$$

$\mathcal{M}$: encodes similarities between substructures

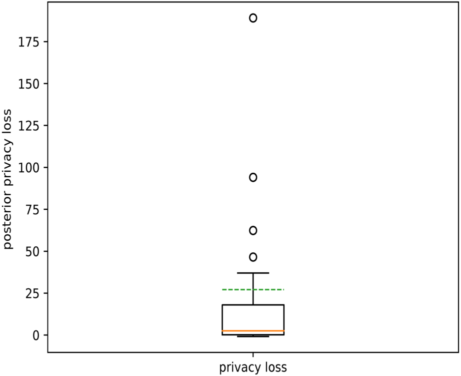
Kernel-assisted Ranking





- $f(\cdot) = \frac{1}{rank(\cdot)}$
- mean correct rank under **DSP** (random) at $\mathbf{140}$ (750)

Privacy Loss



- mean $= 27$
- median $= 2.52$

Takeaways

- **Location pruning** does not necessarily make network more privacy-preserving
- Including **rare transitions** in longitudinal mobility did **not** add discriminative information
- Deanonymization is assisted by **frequency of locations**, **directionality of transitions**

Future Directions

- **Geometry of kernel feature spaces**: high dimensional space with meaningful neighborhood relations
- **Other graph similarity techniques**: network alignment, persistent cascades, frequent/discriminative substructure mining, anonymous walks, spectral representations
- Application to **other categories of sequential datasets**: web browsing behaviour, smartphone app usage
- **Formal privacy guarantees** for mobility networks
- **Utility preserving defense mechanisms**: kernel-agnostic defense, randomisation of node
- **Generative mechanisms** for synthetic traces with anonymity guarantees attributes, perturbations of edges, node removal

Summary of findings

We investigated privacy properties of **graph representations** of longitudinal mobility

- New deanonymization attack on mobility data using **structural similarity** with historical information
- Evaluation on **large dataset of cell-tower location traces**
  - network representations of mobility display **distinct structure**, **even for small number of nodes**
  - $<$ **20 locations** are enough to identify uniquely a population of **1500 users**
  - **kernel-based distance functions** can quantify similarity in absence of location semantics and fine-grained temporal information
  - probabilistic deanonymization using similarity with historical data can achieve **median success probability** $3.5\times$ **higher than a random mechanism**

References I

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013).
Unique in the Crowd: The privacy bounds of human mobility.
Scientific reports, 3(1):1376.

De Mulder, Y., Danezis, G., Batina, L., and Preneel, B. (2008).
Identification via location-profiling in GSM networks.
In *Proceedings of the 2008 ACM Workshop on Privacy in the Electronic Society, WPES 2008, Alexandria, VA, USA, October 27, 2008*, pages 23–32.

Gambs, S., Killijian, M.-O., and Núñez Del Prado Cortez, M. (2014).
De-anonymization attack on geolocated data.
J. Comput. Syst. Sci., 80(8):1597–1614.

Lin, M., Cao, H., Zheng, V. W., Chang, K. C., and Krishnaswamy, S. (2015).
Mobile user verification/identification using statistical mobility profile.
In *2015 International Conference on Big Data and Smart Computing, BIGCOMP 2015, Jeju, South Korea, February 9-11, 2015*, pages 15–18.

Naini, F. M., Unnikrishnan, J., Thiran, P., and Vetterli, M. (2016).
Where You Are Is Who You Are: User Identification by Matching Statistics.
IEEE Transactions on Information Forensics and Security, 11(2):358–372.

Narayanan, A. and Shmatikov, V. (2008).
Robust de-anonymization of large sparse datasets.
In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, SP '08, pages 111–125, Washington, DC, USA. IEEE Computer Society.

References II

Pyrgelis, A., Troncoso, C., and De Cristofaro, E. (2017).
What does the crowd say about you? evaluating aggregation-based location privacy.
Proceedings on Privacy Enhancing Technologies, 2017(4):156–176.

Scholtes, I. (2017).
When is a network a network?: Multi-order graphical model selection in pathways and temporal networks.
In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 1037–1046, New York, NY, USA. ACM.

Sharad, K. and Danezis, G. (2014).
An automated social graph de-anonymization technique.
In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, WPES '14, pages 47–58, New York, NY, USA. ACM.

Sweeney, L. (2002).
k-anonymity: A model for protecting privacy.
International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570.

Vishwanathan, S., Schraudolph, N., Kondor, R., and Borgwardt, K. (2010).
Graph Kenrels.
Journal of Machine Learning Research, 11:1201–1242.

Xu, F., Tu, Z., Li, Y., Zhang, P., Fu, X., and Jin, D. (2017).
Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data.
In *Proceedings of the 26th International Conference on World Wide Web*, pages 1241–1250. International World Wide Web Conferences Steering Committee.

References III

Yanardag, P. and Vishwanathan, S. (2015).
Deep graph kernels.
In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1365–1374, New York, NY, USA. ACM.

Zang, H. and Bolot, J. (2011).
Anonymization of location data does not work: A large-scale measurement study.
In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, MobiCom '11, pages 145–156, New York, NY, USA. ACM.

Thanks!

Any Questions?

dm754@cam.ac.uk