Privacy and Ethics Issues

Mirco Musolesi
Department of Geography, UCL
m.musolesi@ucl.ac.uk

Example: Location-based Social Network Systems

Privacy, Ethics and the Law

- Mining social and geographic data raises a series of ethical concerns related to the privacy rights of the individuals.
- It is fundamental to consider the ethical implications of the various types of analysis we perform on the data.
- A typical example is related to the analysis of mobility patterns: we can easily extract not only home and work locations, but also religion (looking at religious places visited by the individual regularly), political affiliation (e.g., an individual attending a rally), etc.

Privacy Issues concerning Social and Geographic Data

- Privacy is a key concern for various reasons including:
  - Information are almost by definition of personal nature;
  - Information such a location can be linked to personal identity;
  - In general, data mining and data fusion techniques might be applied to infer information about the profile of the users;
  - Data might include health information (see sensor data extracted from Apple Watch, Fitbit devices, etc.)

Location&Privacy

- Possible solutions for preserving users' privacy include:
  - Obfuscation: the precision of the data is blurred
  - Data aggregation: data of individuals are aggregated and are presented together as a statistical sample
  - Anonimysation: the identity of the people is not revealed
    - Possible techniques include: encryption, mapping with keys that are not publicly available
    - Possible problem: linking different data sources (more later)

Location&Privacy

- Location has been investigated for long time by the research and industrial community.
- A good survey is the following:

Krumm, J., 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing*, *13*(6), pp.391-399.

Linked Data and Privacy Issues

- Another emerging problem is the availability of datasets collected by means of different sources:
  - Commercial data (fidelity cards, online shopping, etc.)
  - Financial data (bank transactions, etc.)
  - Governmental data (fiscal, etc.)
  - Sensor data (for example, CCTVs, card readers, but also mobile sensing data)

De-anonymisation

- By linking all the data sources, it might be possible to de-anonymise the data, revealing for example the identity of people or information about them (for example their locations)
- When you design a privacy-preserving system, you should keep in mind potential use of additional data sources for de-anonymise your information

Identification and Obfuscation

- Data can be used to determine the identity of an individual: few points might be sufficient to determine the identity of a person.
- By adding “noise” it is possible to avoid user identification (these are usually called *obfuscation techniques*).
- An interesting book on the topic is the following:

Brunton, Finn, and Helen Nissenbaum. *Obfuscation: A User's Guide for Privacy and Protest*. MIT Press, 2015.

Data Use and Sharing

- Another problem is the use of the personal data
- Usually, a consent from the user is required
  - See, for example, the “agreement” when you install a mobile app;
- Personal data collected must be stored securely:
  - For example, personal data collected by mobile apps must be stored in an encrypted way in a secure server;
- Sharing is usually not permitted if not regulated by the initial agreement.

So is Mining Big Data Good or Evil?

- Big opportunities but also potential issues especially related to privacy
- Many interesting applications:
  - Intelligent marketing
  - Personalisation
  - Transportation
  - Understanding groups, communities, cities, nations, etc.
  - …

Mirco Musolesi. Big Mobile Data Mining: Good or Evil? In IEEE Internet Computing. January-February 2014.

Further Readings

- de Montjoye, Yves-Alexandre, et al. "Unique in the crowd: The privacy bounds of human mobility." *Scientific reports* 3 (2013).
- Rossi, L., Walker, J., & Musolesi, M. (2015). Spatio-temporal techniques for user identification by means of GPS mobility data. *EPJ Data Science*, *4*(1), 1-16.
- Gross, Ralph, and Alessandro Acquisti. "Information revelation and privacy in online social networks." *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM, 2005.

Further Readings

- Zheleva, Elena, and Lise Getoor. "To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles." *Proceedings of the 18th international conference on World Wide Web*. ACM, 2009.
- Narayanan, Arvind, and Vitaly Shmatikov. "De-anonymizing social networks." *Security and Privacy.* IEEE, 2009 (very technical, but the finding is very interesting).
- boyd, danah. *It's complicated: The social lives of networked teens*. Yale University Press, 2014.

Location-Based Social Networks

- The proliferation of GPS enabled devices has led to the popularity of Location-Based Social Networks
- Foursquare: **> 45 million users** (beginning 2014)

Location-Based Social Networks

- Based on the concept of **check-in**
  - A user can register his/her presence at a certain location and share this information with social contacts, along with comments, recommendations, etc.

Location-Based Social Networks

- Based on the concept of **check-in**
  - A user can register his/her presence at a certain location and share this information with social contacts, along with comments, recommendations, etc.
- Users are **encouraged to disseminate** location information in the network

Location-Based Social Networks

- Based on the concept of **check-in**
  - A user can register his/her presence at a certain location and share this information with social contacts, along with comments, recommendations, etc.
- Users are **encouraged to disseminate** location information in the network
- Tagging can lead to release of location information of users that have **no control over the data**

Location-Based Social Networks

- Based on the concept of **check-in**
  - A user can register his/her presence at a certain location and share this information with social contacts, along with comments, recommendations, etc.
- Users are **encouraged to disseminate** location information in the network
- Tagging can lead to release of location information of users that have **no control over the data**

- Increasing concern about possibility of identifying users from geo-social media

A Toy Example

- The attacker has access to both unanonymised LBSN data and a source of anonymised location information
- The attacker's goal is that of revealing the identities of u_i by linking location information across the two databases
  - Along with potentially **sensitive information s_i**

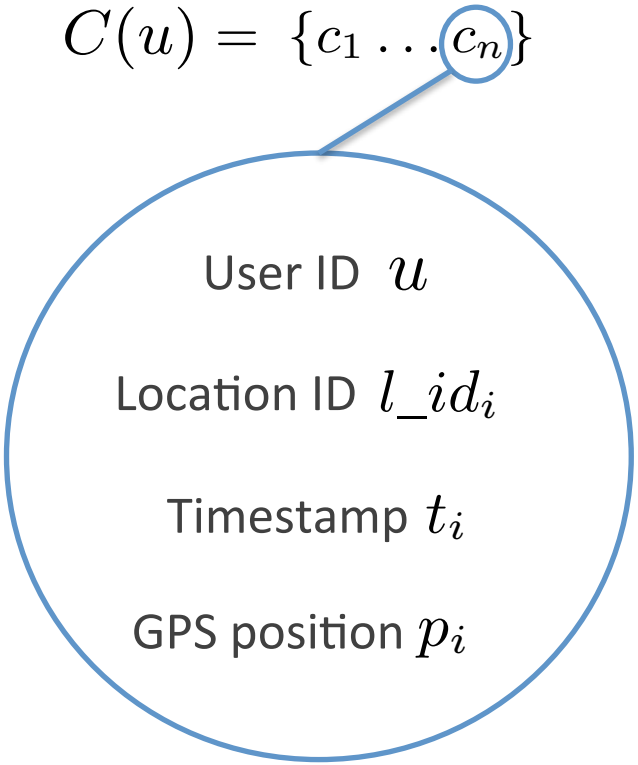
	l_1	l_2	l_3	l_4
Alice	4	4	4	4
Bob	1	1	1	4
Charlie	5	1	2	0

id	Trace	Other
u_1	l_4, l_1, l_4	s_1
u_2	l_1, l_1, l_1	s_2
u_3	l_1, l_2, l_3	s_3

How Does the Data Look Like?

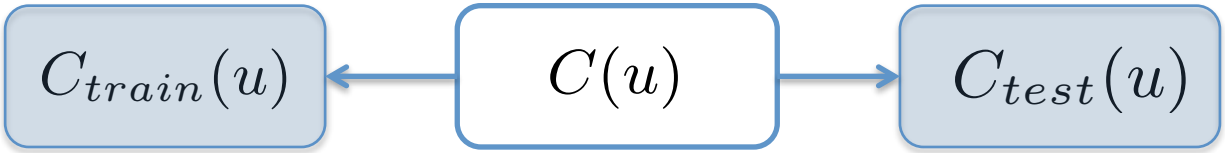
- For each user *u* we have 1 set of check-ins *C(u)*

$$C(u) = \{c_1 \ldots c_n\}$$



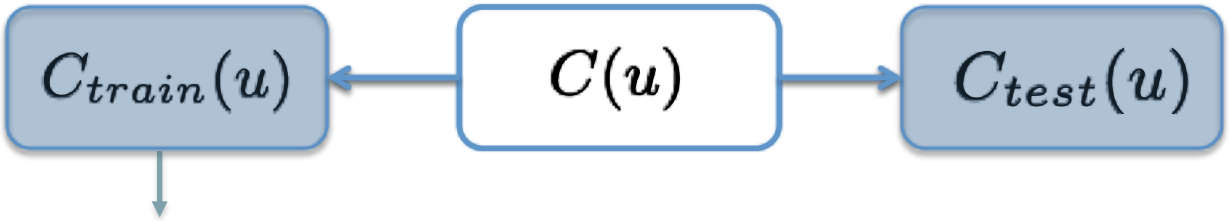
How Does the Data Look Like?

- For each user *u* we have 1 set of check-ins *C(u)*



How Does the Data Look Like?

- For each user *u* we have 1 set of check-ins *C(u)*



User ID u

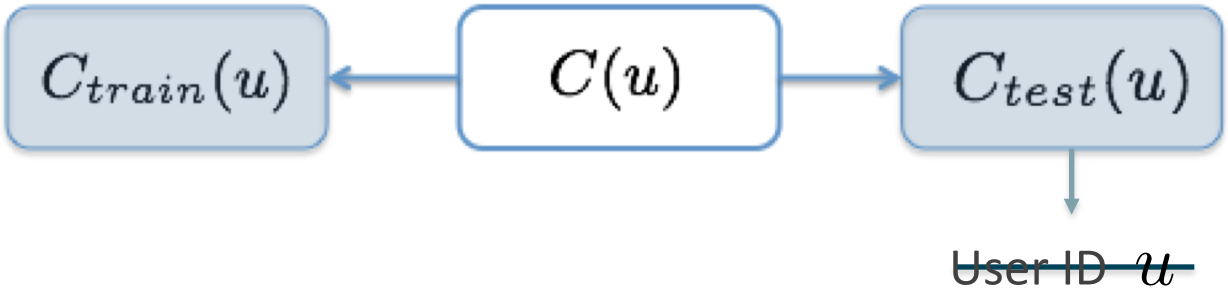
Location ID l_id_i

Timestamp t_i

GPS position p_i

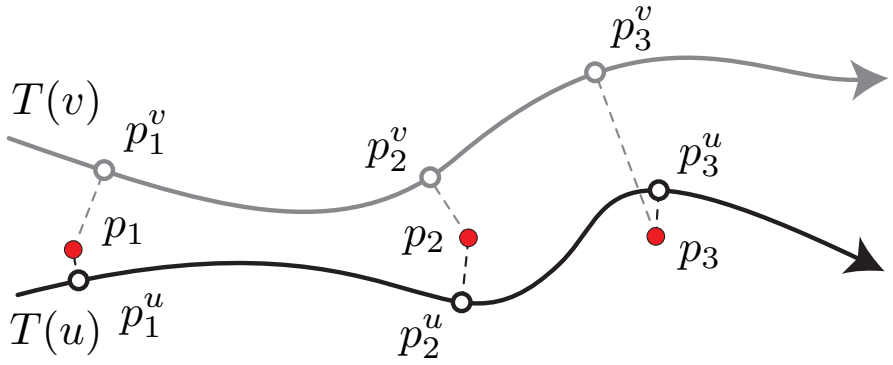
How Does the Data Look Like?

- For each user *u* we have 1 set of check-ins *C(u)*



Trajectory Based Identification

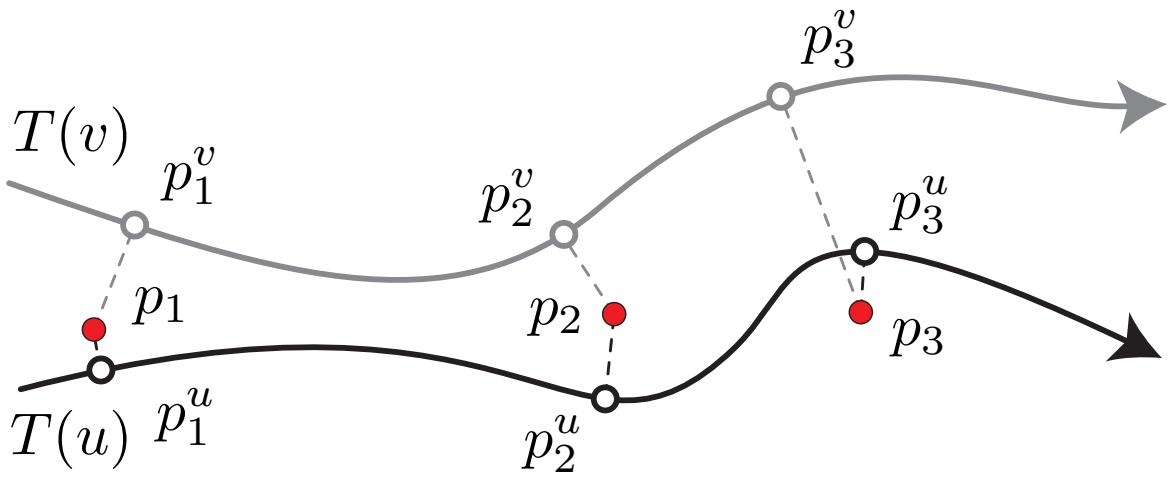
- **Assumption**: the set of unlabelled test points belongs to a single user *u*
- 1 user corresponds to 1 spatio-temporal trajectory
- **Rationale**: use the spatio-temporal information of the check-ins to assign the unlabelled points to the closest trajectory
- Let *T(v)* denote the trajectory associated to *v*



Trajectory Based Identification

- Directed Hausdorff distance from A to B: point *a* of A such that the distance from its nearest neighbour *b* in B is maximal

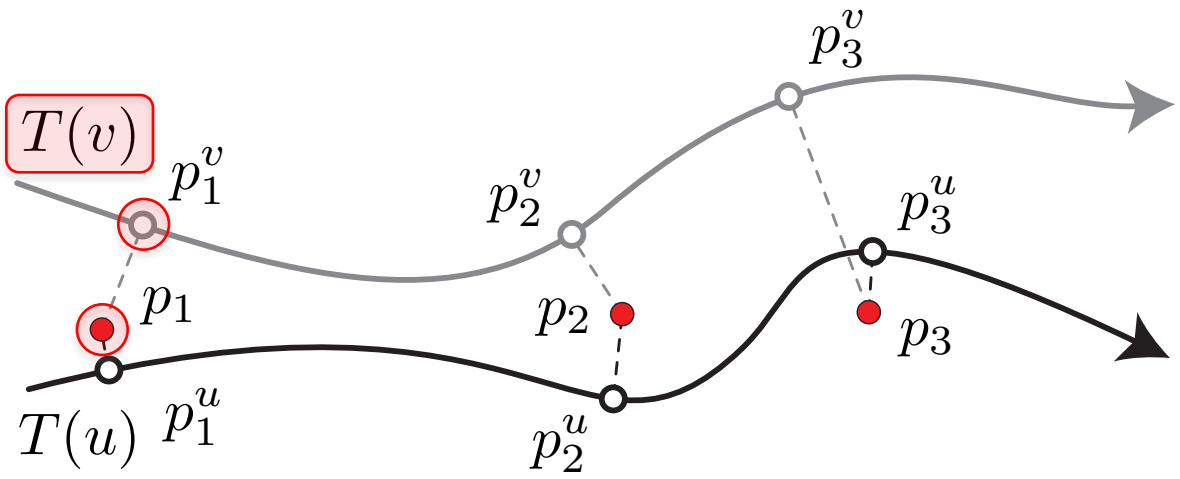
$$h_m(A,B) = \frac{1}{|A|}\sum_{a\in A}\min_{b\in B}||a-b||$$



Trajectory Based Identification

- Directed Hausdorff distance from A to B: point *a* of A such that the distance from its nearest neighbour *b* in B is maximal

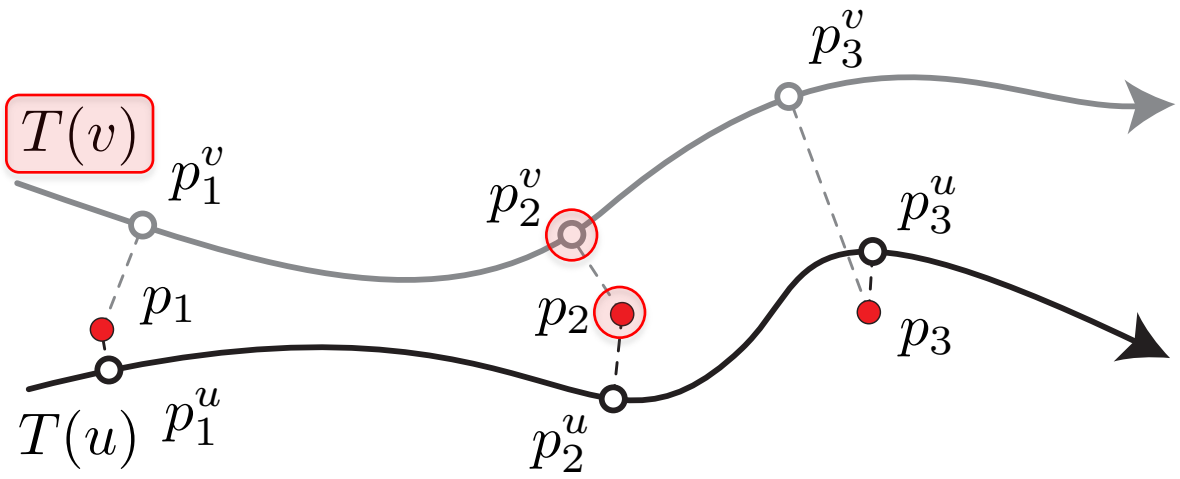
$$h_m(A,B) = \frac{1}{|A|}\sum_{a\in A}\min_{b\in B}||a-b||$$



Trajectory Based Identification

- Directed Hausdorff distance from A to B: point *a* of A such that the distance from its nearest neighbour *b* in B is maximal

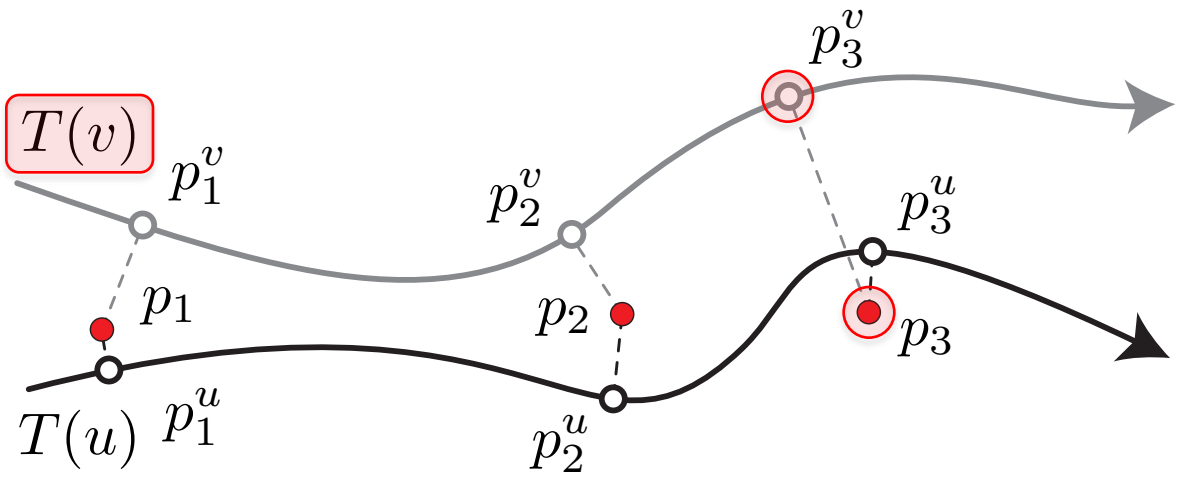
$$h_m(A,B) = \frac{1}{|A|}\sum_{a\in A}\min_{b\in B}||a-b||$$



Trajectory Based Identification

- Directed Hausdorff distance from A to B: point *a* of A such that the distance from its nearest neighbour *b* in B is maximal

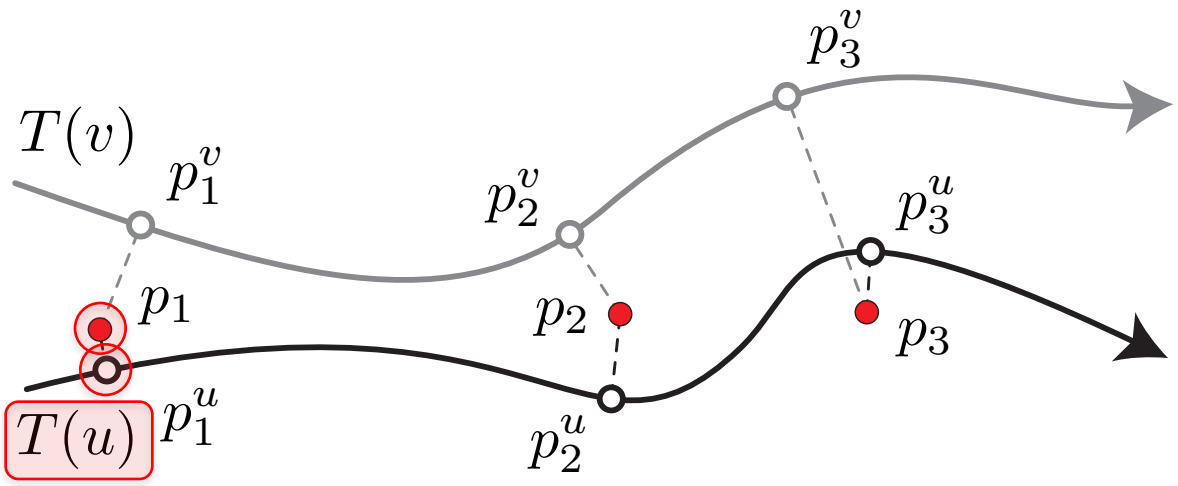
$$h_m(A,B) = \frac{1}{|A|}\sum_{a\in A}\min_{b\in B}||a-b||$$



Trajectory Based Identification

- Directed Hausdorff distance from A to B: point *a* of A such that the distance from its nearest neighbour *b* in B is maximal

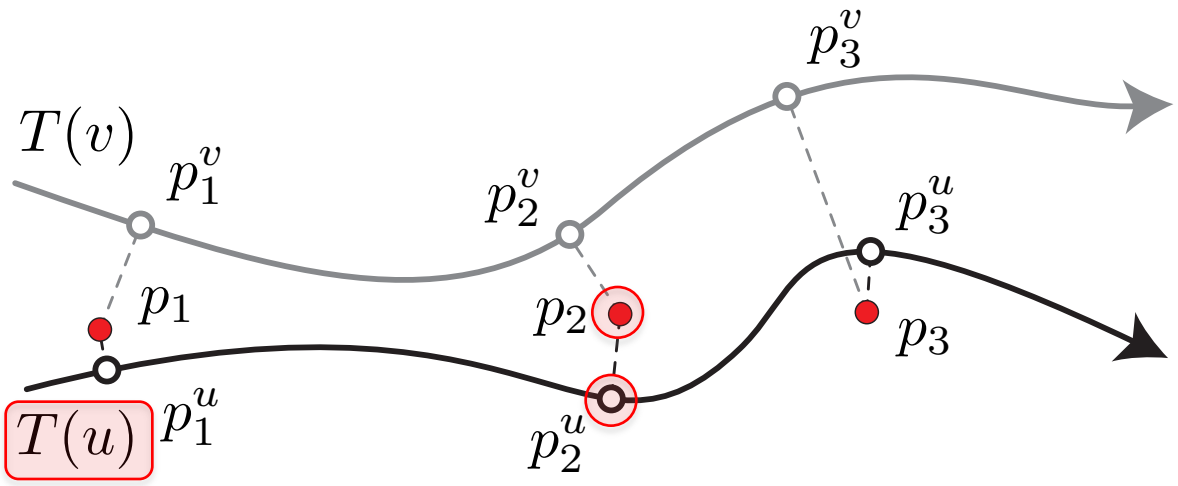
$$h_m(A, B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} ||a - b||$$



Trajectory Based Identification

- Directed Hausdorff distance from A to B: point *a* of A such that the distance from its nearest neighbour *b* in B is maximal

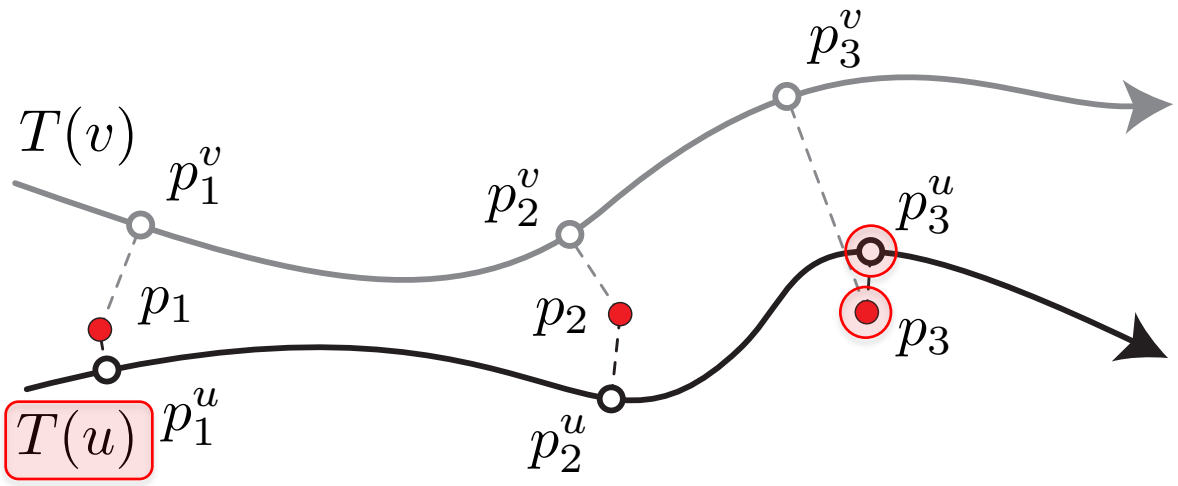
$$h_m(A, B) = \frac{1}{|A|} \sum_{a \in A} \min_{b \in B} ||a - b||$$



Trajectory Based Identification

- Directed Hausdorff distance from A to B: point *a* of A such that the distance from its nearest neighbour *b* in B is maximal

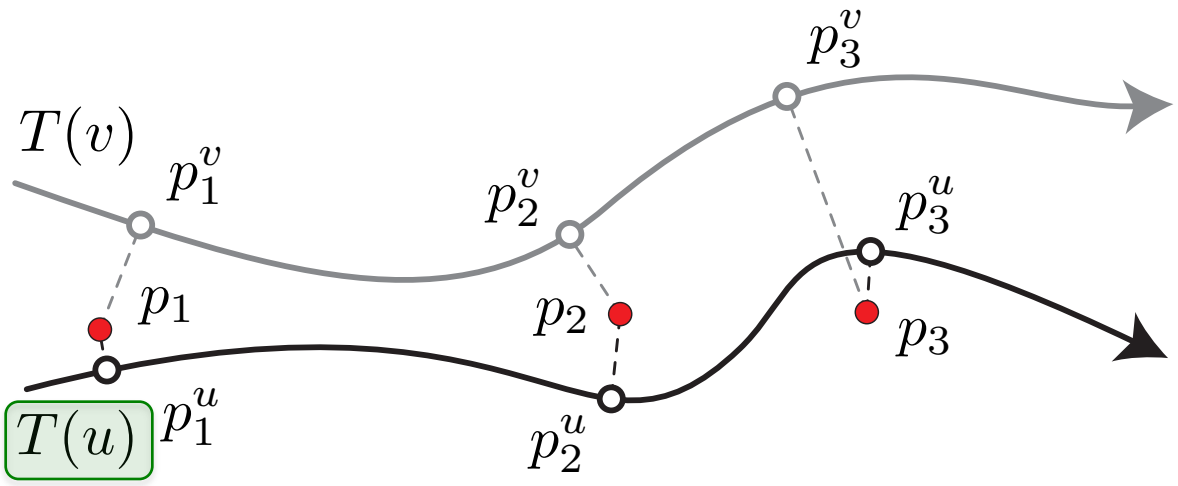
$$h_m(A,B) = \frac{1}{|A|}\sum_{a\in A}\min_{b\in B}||a-b||$$



Trajectory Based Identification

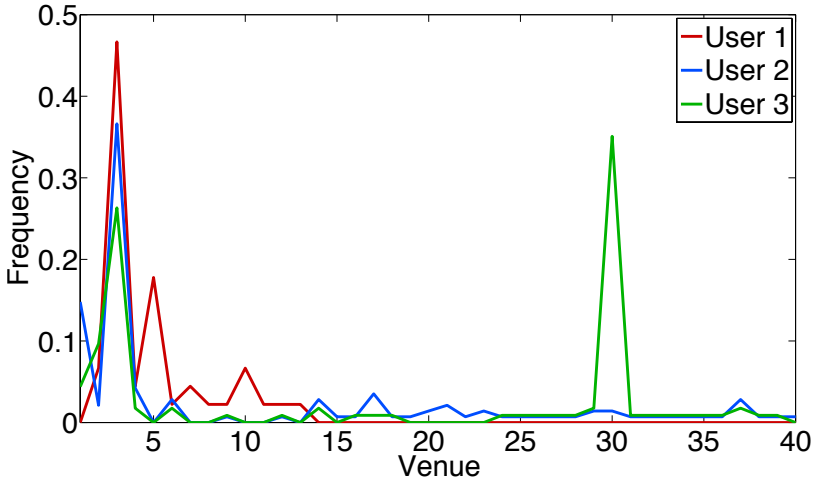
- Directed Hausdorff distance from A to B: point *a* of A such that the distance from its nearest neighbour *b* in B is maximal

$$h_m(A,B) = \frac{1}{|A|}\sum_{a\in A}\min_{b\in B}||a-b||$$



Frequency Based Identification

- **Assumption**: the set of unlabelled check-in points belongs to a single user *u*
- **Rationale**: characterise a user with the frequencies of visit to the different locations
- Simple Naïve Bayes model



Frequency Based Identification

- Alice (0.25, 0.25, 0.25, 0.25)
- Bob (0.14, 0.14, 0.14, 0.58)
- Charlie (0.62, 0.25, 0.13, 0.00)

	l_1	l_2	l_3	l_4
Alice	4	4	4	4
Bob	1	1	1	4
Charlie	5	1	2	0

id	Trace	Other
u_1	l_4, l_1, l_4	s_1
u_2	l_1, l_1, l_1	s_2
u_3	l_1, l_2, l_3	s_3

Frequency Based Identification

- Alice (0.25, 0.25, 0.25, 0.25) : 0.25 x 0.25 x 0.25 = 0.015625
- Bob (0.14, 0.14, 0.14, 0.58)
- Charlie (0.62, 0.25, 0.13, 0.00)

	l_1	l_2	l_3	l_4
Alice	4	4	4	4
Bob	1	1	1	4
Charlie	5	1	2	0

id	Trace	Other
u_1	l_4, l_1, l_4	s_1
u_2	l_1, l_1, l_1	s_2
u_3	l_1, l_2, l_3	s_3

Frequency Based Identification

- Alice (0.25, 0.25, 0.25, 0.25) : 0.25 x 0.25 x 0.25 = 0.015625
- Bob (0.14, 0.14, 0.14, 0.58) : 0.58 x 0.14 x 0.58 = 0.047096
- Charlie (0.62, 0.25, 0.13, 0.00) :

	l_1	l_2	l_3	l_4
Alice	4	4	4	4
Bob	1	1	1	4
Charlie	5	1	2	0

id	Trace	Other
u_1	l_4, l_1, l_4	s_1
u_2	l_1, l_1, l_1	s_2
u_3	l_1, l_2, l_3	s_3

Frequency Based Identification

- Alice (0.25, 0.25, 0.25, 0.25) : 0.25 x 0.25 x 0.25 = 0.015625
- Bob (0.14, 0.14, 0.14, 0.58) : 0.58 x 0.14 x 0.58 = 0.047096
- Charlie (0.62, 0.25, 0.13, 0.00) : 0.0 x 0.62 x 0.0 = 0.0

	l_1	l_2	l_3	l_4
Alice	4	4	4	4
Bob	1	1	1	4
Charlie	5	1	2	0

id	Trace	Other
u_1	l_4, l_1, l_4	s_1
u_2	l_1, l_1, l_1	s_2
u_3	l_1, l_2, l_3	s_3

Frequency Based Identification

- Alice (0.25, 0.25, 0.25, 0.25) : 0.25 x 0.25 x 0.25 = 0.015625
- Bob (0.14, 0.14, 0.14, 0.58) : 0.58 x 0.14 x 0.58 = 0.047096
- Charlie (0.62, 0.25, 0.13, 0.00) : 0.0 x 0.62 x 0.0 = 0.0

	l_1	l_2	l_3	l_4
Alice	4	4	4	4
Bob	1	1	1	4
Charlie	5	1	2	0

id	Trace	Other
u_1	l_4, l_1, l_4	s_1
u_2	l_1, l_1, l_1	s_2
u_3	l_1, l_2, l_3	s_3

Frequency Based Identification

- **Multinomial Model**: multinomial distribution associated to each users. Parameters estimation via standard MLE

$$v^* = \arg\max_{v \in U} P(v|c_1 \dots c_m)$$

$$v^* = \arg\max_{v \in U} P(v) \prod_{i=1}^{m} P(c_i|v)$$

In our setting we let P(v) = 1/number of users

Frequency Based Identification

- **Multinomial Model**: multinomial distribution associated to each users. Parameters estimation via standard MLE

$$v^* = \arg\max_{v \in U} P(v|c_1 \dots c_m)$$

$$v^* = \arg\max_{v \in U} P(v) \prod_{i=1}^{m} P(c_i|v)$$

Probability of observing the user v at location i

Frequency Based Identification

- **Multinomial Model**: multinomial distribution associated to each users. Parameters estimation via standard MLE

$$v^* = \arg\max_{v \in U} P(v|c_1 \dots c_m)$$

$$v^* = \arg\max_{v \in U} P(v) \prod_{i=1}^{m} P(c_i|v)$$

$$P(c_i|v) = \frac{N_i^v + \alpha}{\sum_{j=1}^{n} N_j^v + \alpha|L|}$$

Maximum Likelihood Estimation

Frequency based Estimation

Time dependent multinomial model

$$P_\xi(c_i|v) = \frac{N_i^v(\xi) + \alpha}{\sum_{j=1}^{n} N_j^v(\xi) + \alpha|L|}$$

Social Smoothing

$$\frac{N_i^v + \mu \sum_{w \in \mathcal{S}(v)} s(v,w) N_i^w + \alpha}{\sum_{j=1}^{n} N_j^v + \mu \sum_{w \in \mathcal{S}(v)} \sum_{j=1}^{n} s(v,w) N_j^w + \alpha|L|}$$

Frequency based Estimation

Time dependent multinomial model

$$P_\xi(c_i|v) = \frac{N_i^v(\xi) + \alpha}{\sum_{j=1}^{n} N_j^v(\xi) + \alpha|L|}$$

Social Smoothing

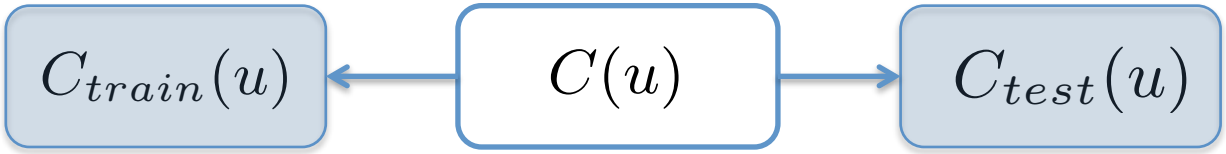
$$\frac{N_i^v + \mu \sum_{w \in \mathcal{S}(v)} s(v,w) N_i^w + \alpha}{\sum_{j=1}^{n} N_j^v + \mu \sum_{w \in \mathcal{S}(v)} \sum_{j=1}^{n} s(v,w) N_j^w + \alpha|L|}$$

Datasets

- Brightkite
  - 4,491,143 check-ins from 58,228 users over 772,764 location, from April 2008 to October 2010
- Gowalla
  - 6,442,890 check-ins from 196,591 users over 1,280,969 locations, collected from February 2009 to October 2010
- Foursquare
  - 2,073,740 check-ins from 18,107 users over 43,063 locations, from August 2010 to November 2011

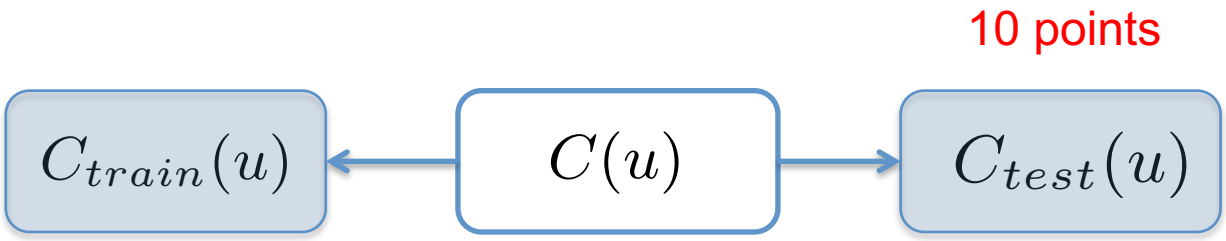
Experimental Setup

- For each user *u* we remove 10 check-ins from C(*u*) to create $C_{test}(u)$ and $C_{train}(u)$



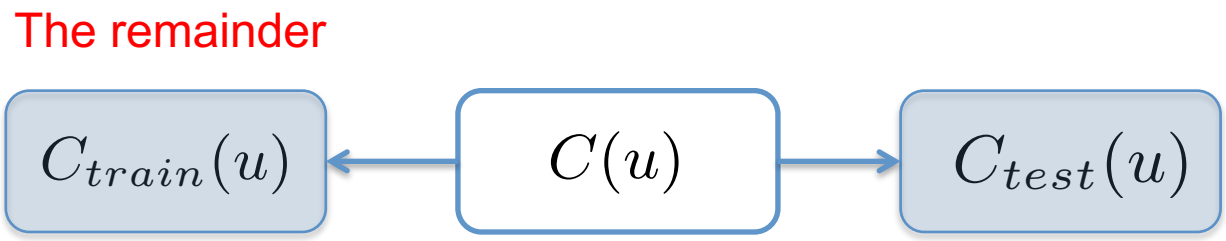
Experimental Setup

- For each user *u* we remove 10 check-ins from C(*u*) to create $C_{test}(u)$ and $C_{train}(u)$



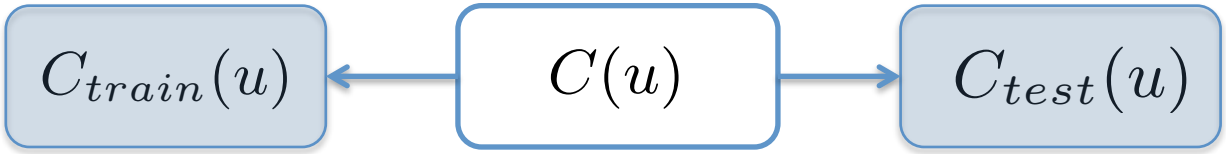
Experimental Setup

- For each user *u* we remove 10 check-ins from C(*u*) to create $C_{test}(u)$ and $C_{train}(u)$



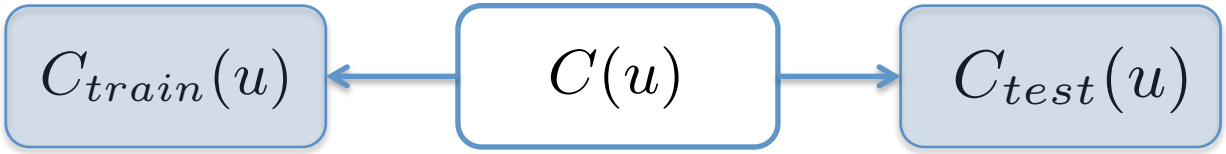
Experimental Setup

- For each user *u* we remove 10 check-ins from C(*u*) to create $C_{test}(u)$ and $C_{train}(u)$



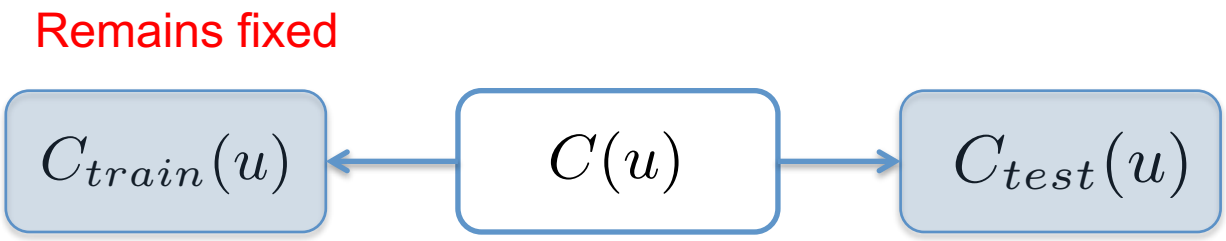
Experimental Setup

- How does the number of points observed in $C_{test}(u)$ change our ability to classify an individual?



Experimental Setup

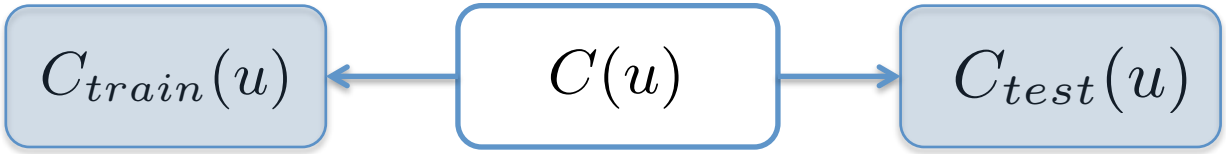
- How does the number of points observed in $C_{test}(u)$ change our ability to classify an individual?



Experimental Setup

- How does the number of points observed in $C_{test}(u)$ change our ability to classify an individual?

Varies between 1 and 10 points

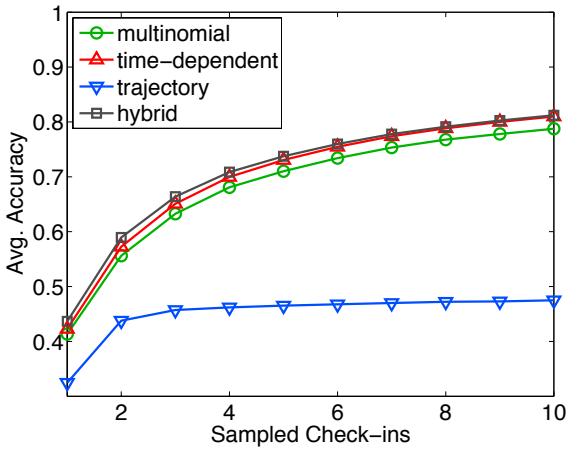


Experimental Setup

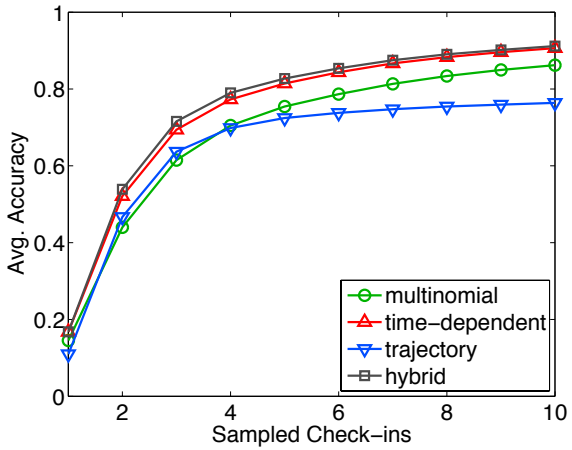
- Given $C_{test}(u)$ our task is that of finding the user u^* that originated the check-ins
- Evaluation in terms of
  - **Accuracy**: ratio of successfully identified users
- 100 repetitions: avg. accuracy +/- std. error

Experimental Results

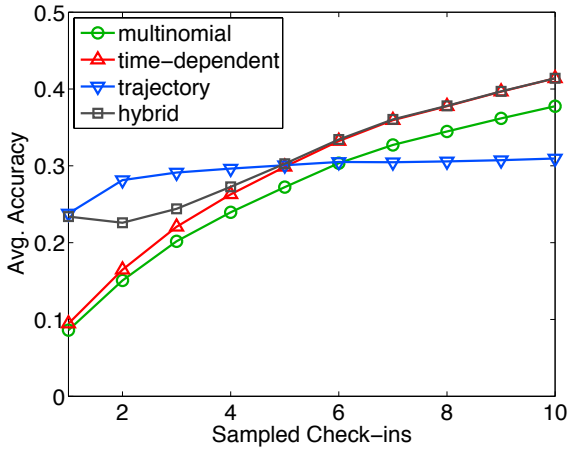
San Francisco



(a) Brightkite



(b) Gowalla

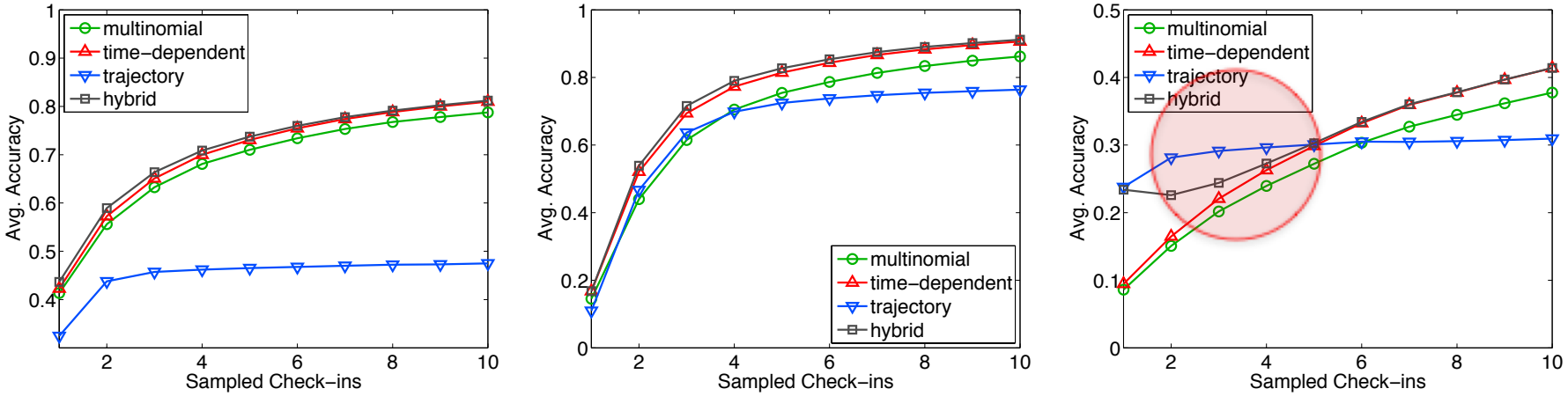


(c) Foursquare

- We measure the identification complexity (accuracy) for 4 different attack models

Experimental Results

Trajectory more efficient with few points

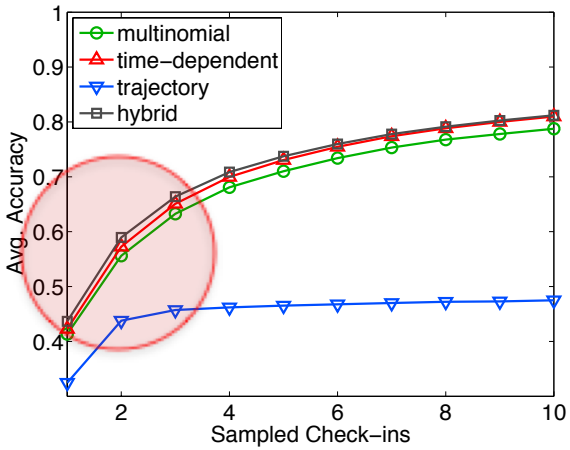


(a) Brightkite (b) Gowalla (c) Foursquare

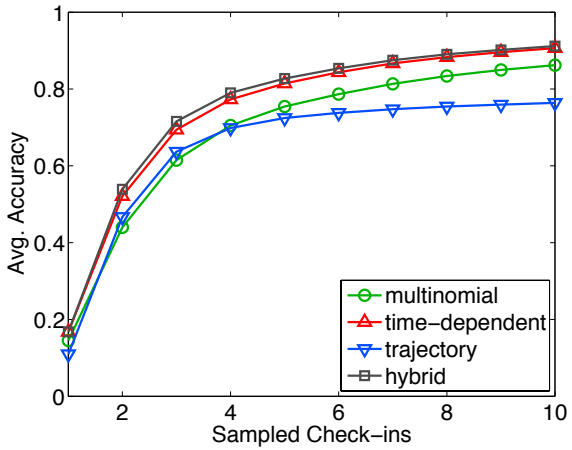
- We measure the identification complexity (accuracy) for 4 different attack models

Experimental Results

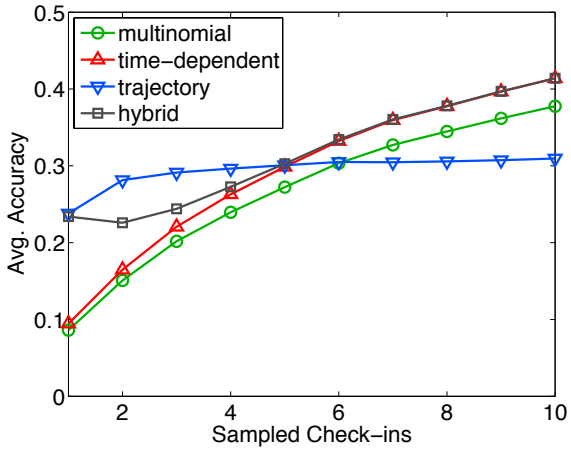
…but not always



(a) Brightkite



(b) Gowalla



(c) Foursquare

- We measure the identification complexity (accuracy) for 4 different attack models

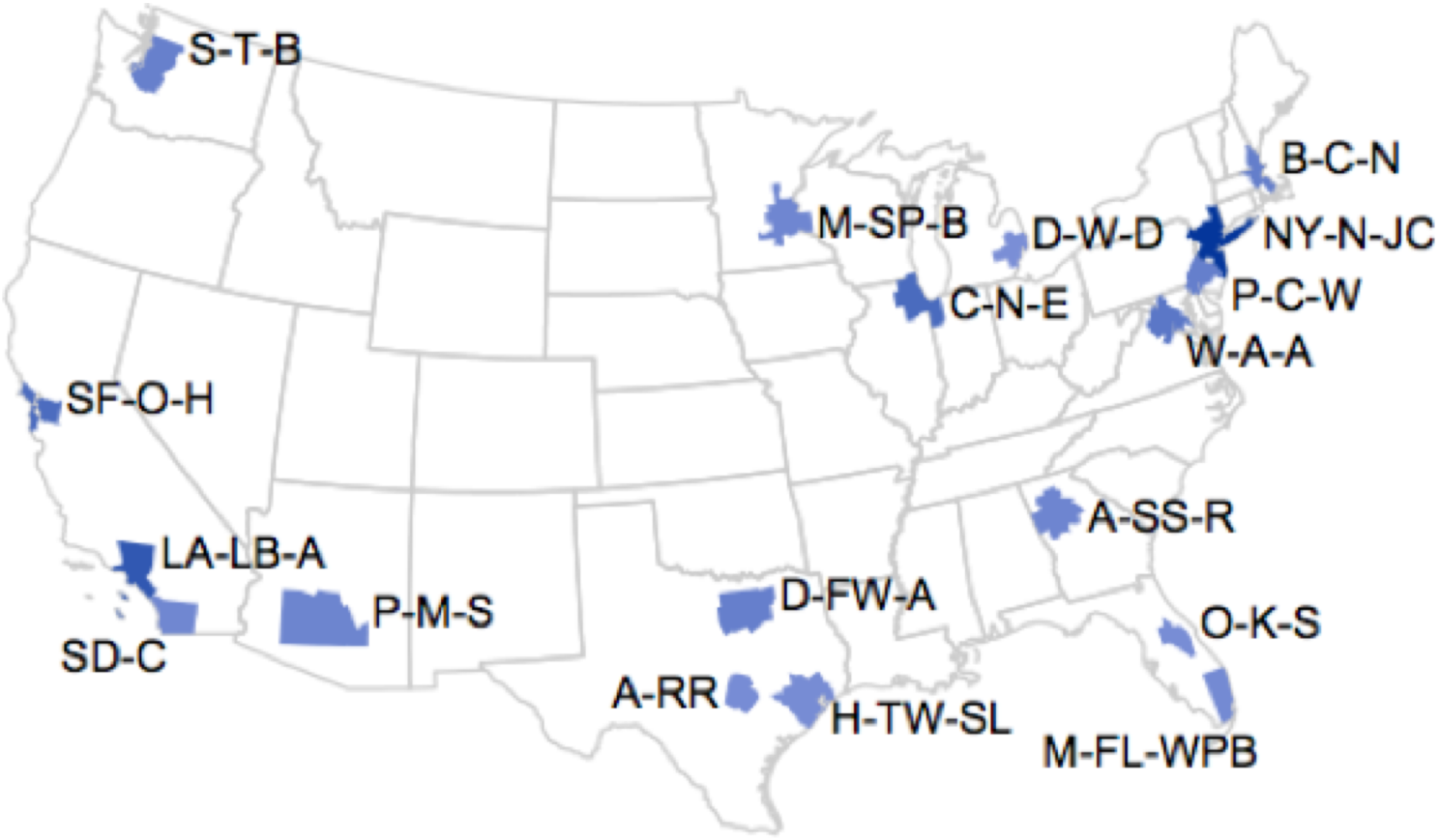
Location Semantics

- So, location data should be treated with care to protect users' privacy, but:
  - Are some locations more discriminative than others?
  - What are the types of venues that an attacker has to monitor to maximise the probability of success?
  - When should a user decide whether to make his/her check-in to a location public or not?

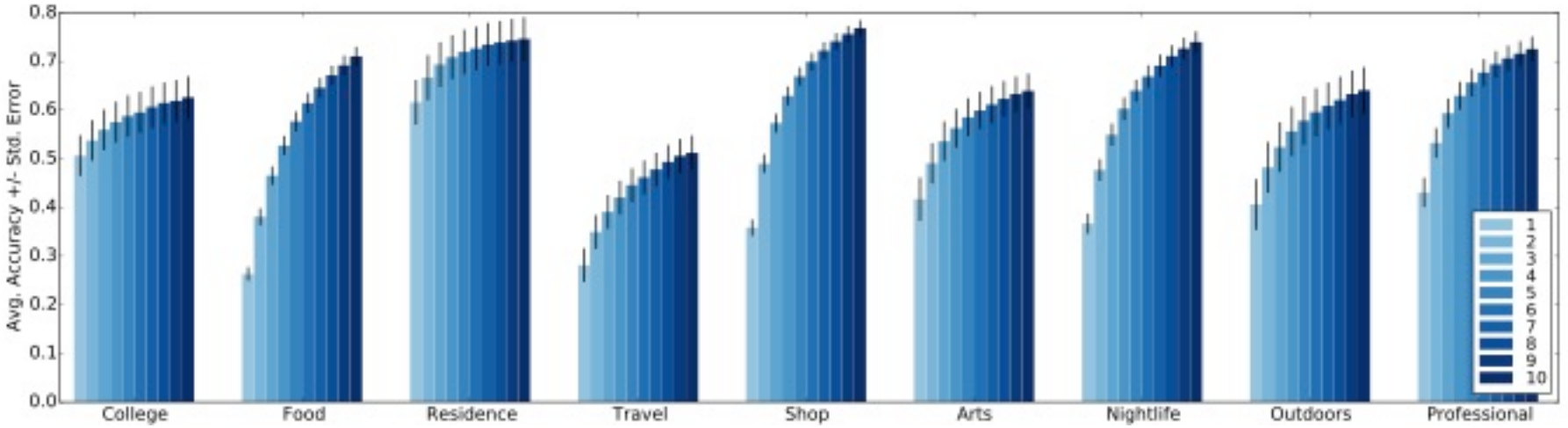
Location Semantics

- We assume that the attacker has access only to a number of check-ins in locations in specific categories
  - e.g., restaurants.
- 20,785 users and 1,391,765 check-ins over 134,989 venues in 17 *Core Based Statistical Area* (CBSA)
  - CBSA are urban regions according to the US Office of Management and Budget (OMB)

Dataset

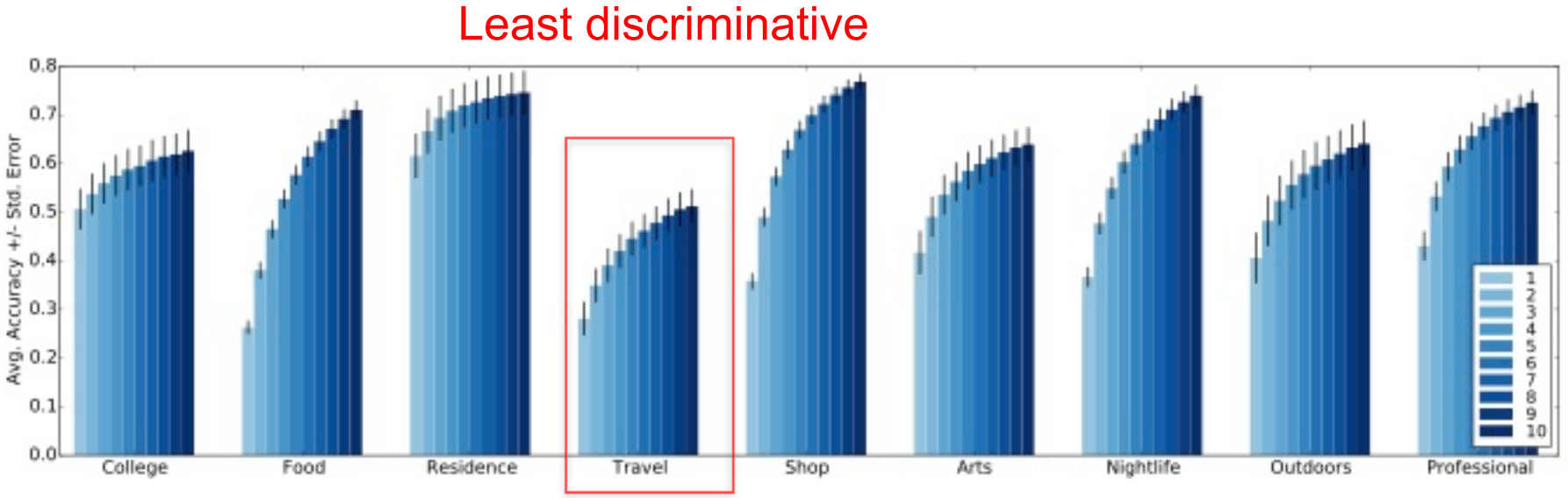

S-T-B
B-C-N
M-SP-B
D-W-D
NY-N-JC
C-N-E
P-C-W
W-A-A
SF-O-H
A-SS-R
LA-LB-A
P-M-S
D-FW-A
O-K-S
SD-C
A-RR
H-TW-SL
M-FL-WPB

Categories of Places and Identification



- Venues in the “Travel” category, which are generally characterised by a high user to venues ratio, correspond to a relatively low identification complexity

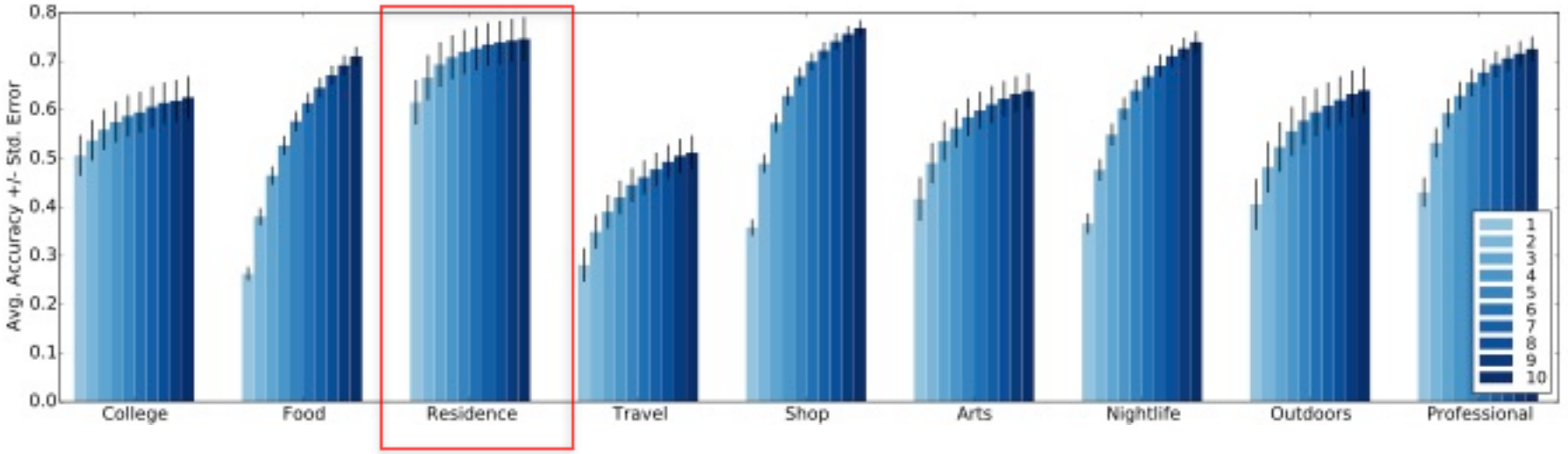
Categories of Places and Identification



- Venues in the “Travel” category, which are generally characterised by a high user to venues ratio, correspond to a relatively low identification complexity

Categories of Places and Identification

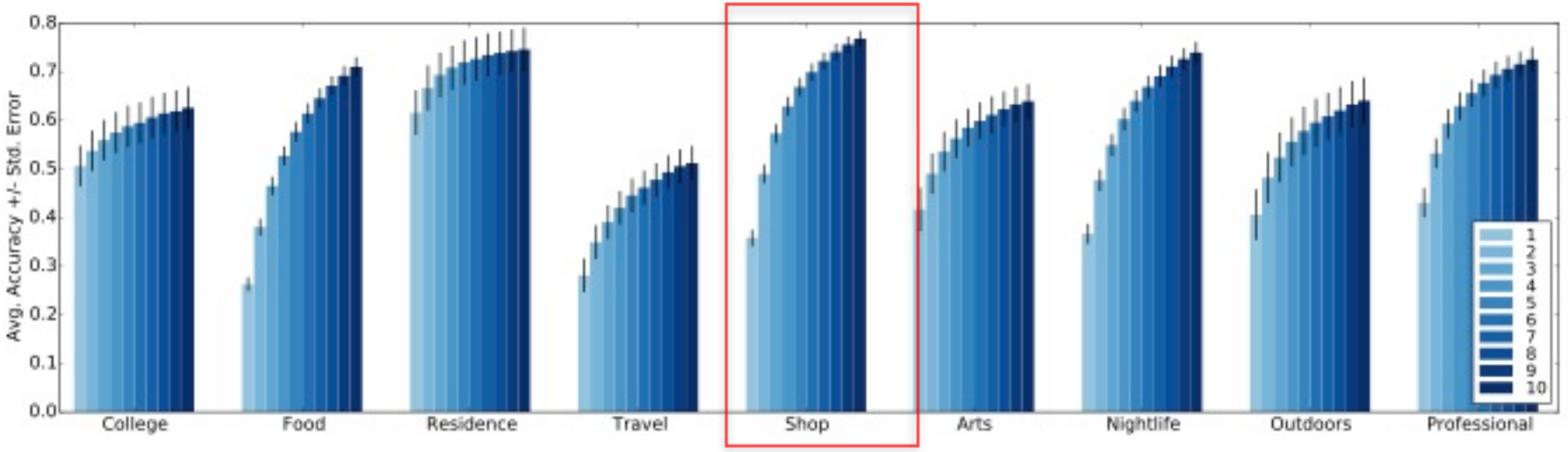
Most discriminative



- Venues in the “Travel” category, which are generally characterised by a high user to venues ratio, correspond to a relatively low identification complexity

Categories of Places and Identification

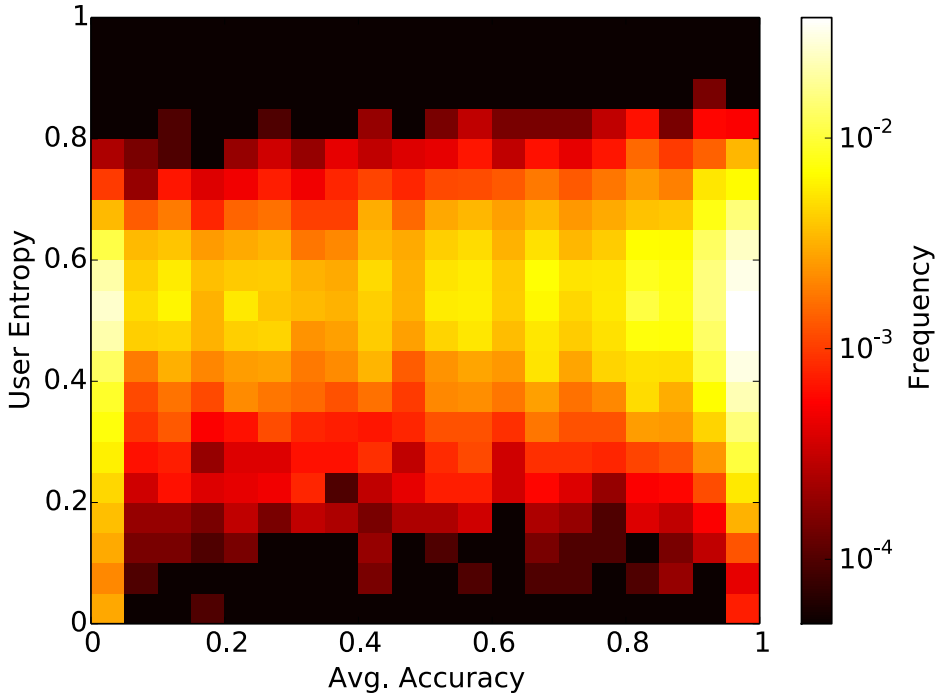
Highly discriminative if enough points are available



- Venues in the “Travel” category, which are generally characterised by a high user to venues ratio, correspond to a relatively low identification complexity

Influence of User's Entropy

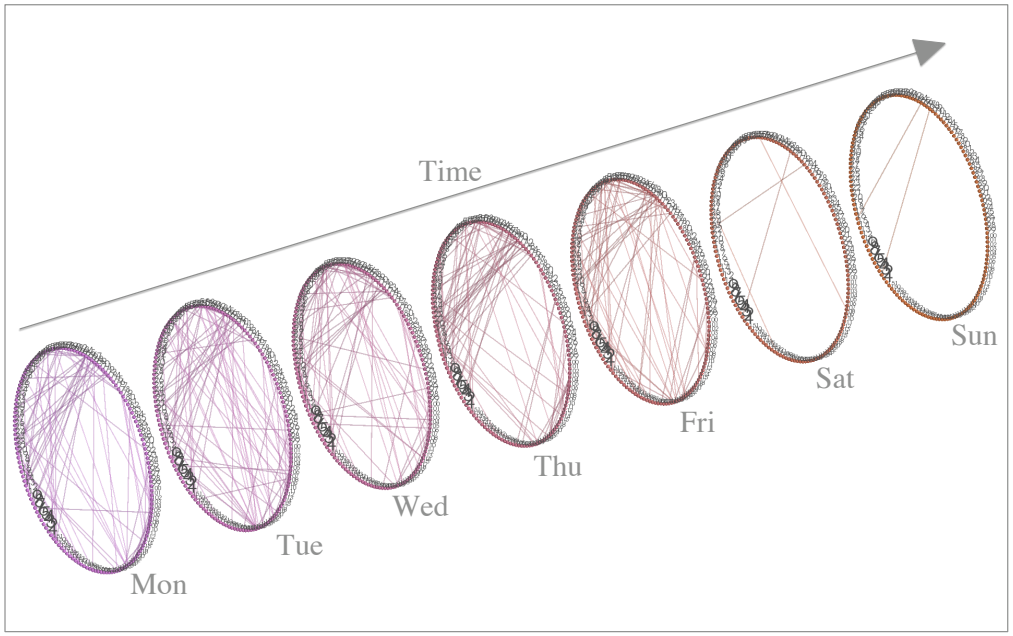
- **High** (**low**) entropy users check-in frequently in **many** (**few**) venues
- **No correlation** between a user's entropy and the complexity of identifying him/her
- **Collective behaviour** rather than individual behaviour determines the identification complexity of the individual

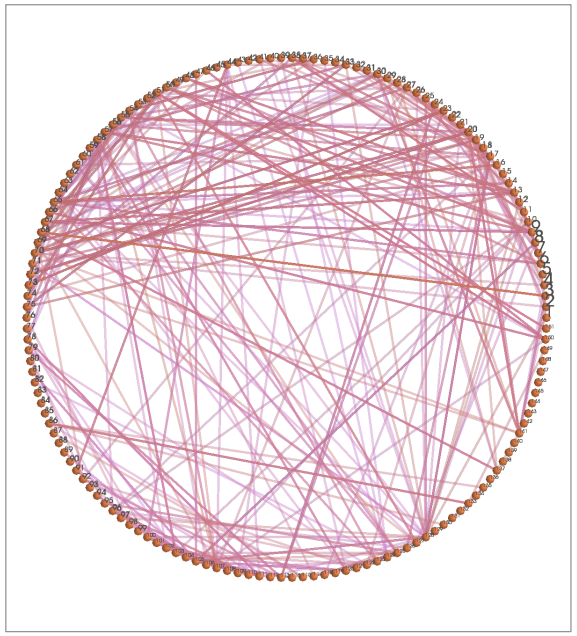


Open Questions

- To what extent the urban environment plays a part in shaping the users check-in patterns and thus their identity privacy?
- Attack model that considers sequences of check-ins
- What can we do to ensure identity privacy?
  - On the k-Anonymization of Time-varying and Multi-layer Social Graphs (AAAI ICWSM 2015)
  - Stronger privacy models? (Stochastic k-automorphism anonymity)

Anonymisation of Time-varying Graphs





Big Mobile Data Mining: Good or Evil?

- Is Big Data Mining good or evil?
- Big opportunities but also potential issues especially related to privacy
  - Differential privacy of big mobile data
  - Informed consent
- Many interesting applications:
  - Healthcare
  - Transportation
  - Development
  - …

[Mirco Musolesi. Big Mobile Data Mining: Good or Evil? In IEEE Internet Computing. January-February 2014.]

• L. Rossi and M. Musolesi, It's the Way you Check-in: Identifying Users in Location-Based Social Networks, In Proceedings of the 2nd *ACM Conference on Online Social Networks (ACM COSN'14)*. Dublin, Ireland. September 2014.

• L. Rossi, M. J. Williams, C. Stich and M. Musolesi, Privacy and the City: User Identification and Location Semantics in Location-Based Social Networks, In Proceedings of the *9th AAAI International Conference on Weblogs and Social Media (ICWSM'15*). Oxford, United Kingdom. May 2015

• L. Rossi, M. Musolesi and A. Torsello, On the k-Anonymization of Time-varying and Multi-layer Social Graphs, Proceedings of the *9th AAAI International Conference on Weblogs and Social Media (ICWSM'15*). Oxford, United Kingdom. May 2015

• L. Rossi and M. Musolesi, Spatio-temporal Techniques for User Identification by means of GPS Mobility. In EPJ Data Science. Volume 4. Issue 11. August 2015.

Questions?

Mirco Musolesi

Intelligent Social Systems Lab
Department of Geography
University College London

W: http://www.ucl.ac.uk/~ucfamus
E: m.musolesi@ucl.ac.uk
T: @mircomusolesi