# Collaborative and privacy-aware sensing for observing urban movement patterns

Nelson Gonçalves, Rui José, Carlos Baquero
Universidade do Minho and INESC TEC, PT

Esorics DPM, September 2013

# Urban Movement Patterns



Nelson Gonçalves, Rui José, Carlos Baquero
Universidade do Minho and INESC TEC, PT

Macro-level detection of aggregated urban movement can assist infrastructure management.

- In a tourism office: *"Are the individuals in this art gallery likely to have visited a given art museum first?"*
- In a shopping mall: *"Which shops are visited most likely after the movie theater? and before the theater?"*

# Individual Movement Patterns

- Individuals often carry devices than can be detected
- Local detections can be shared and allow movement tracking
- `02:27:e4:f2:cd:0a W.Foyer 11:Sep:2013:19:12:33`
- MAC pseudonyms can be correlated to individuals

# Research Questions

1. Can we design a mechanism that preserves privacy while allowing limited accuracy tracking of movement patterns?
2. Can higher accuracy collective movement result from lower accuracy individual tracking?

# Precedence Filters

Our approach, *Precedence Filters*, builds heavily on:

- Bloom Filters (for probabilistic set membership) and on,
- Vector Clocks (for distributed causality tracking).

The goal is to present a probabilist trace of past user locations, when at a given location.

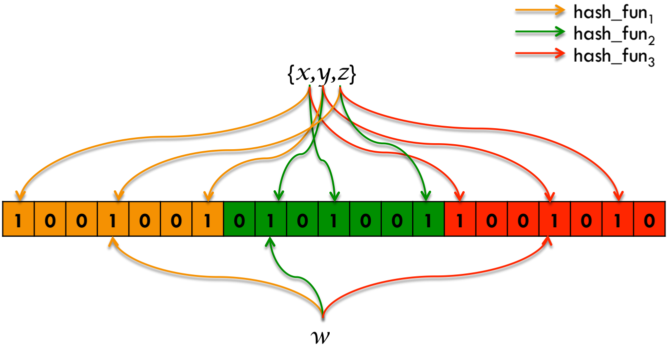
**@ Subway**

$$Bank \rightarrow Market \rightarrow Subway$$

And collectively collect common routes.

# Tools: Bloom Filters

Bloom filter for set $\{x, y, z\}$ with 3 hash functions.



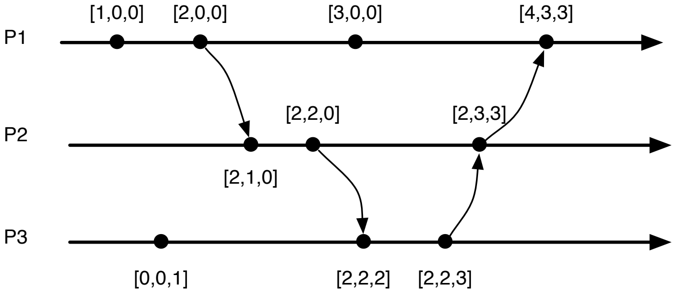

- Querying for element $w$ yields a false positive.
- Larger filters depict larger precision for the same stored set size.

# Tools: Vector Clocks



Nelson Gonçalves, Rui José, Carlos Baquero
Universidade do Minho and INESC TEC, PT

Captures *causality* (*happens before*) relations without wall clocks




$$[2,2,0] \rightarrow [2,2,3] \rightarrow [4,3,4]$$

# System Model and Design



Nelson Gonçalves, Rui José, Carlos Baquero
Universidade do Minho and INESC TEC, PT

- Network of local sensing devices (e.g. WiFi Hotspots)
- MAC/Pseudonyms cannot leave the local sensing device
- Tracking can exhibit false routes (plausible deniability)
- No network communication failures
- Network communication is faster than user movement
- A node holds a filter and caches cells from other filters

# Precedence Filters: Algorithm



Nelson Gonçalves, Rui José, Carlos Baquero
Universidade do Minho and INESC TEC, PT

- All filters have cells at 0 and they can take natural numbers
- A MAC address $a$ is sensed in scanner node $X$
- Using hashes $X$ calculates to which cells item $a$ is mapped
- Each other node sends to $X$ the value on those cells
- Node $X$ updates the caches of node's filters on those cells
- In $X$ filter, on those cells, it stores the maximum known value, plus one. This creates a fingerprint for $a$ that is after all other sightings.

From this information a node can construct its *probabilistic* view of the sequence of visits of a sensed device.

# Mobility Traces



Nelson Gonçalves, Rui José, Carlos Baquero
Universidade do Minho and INESC TEC, PT

## Trace with recurrent visits

$$Subway \rightarrow Market \rightarrow Bookshop \rightarrow Bank \rightarrow Market \rightarrow Subway$$

Precedence filters only capture the last of recurring visits

## Trace with *more recent* visits

$$Bookshop \rightarrow Bank \rightarrow Market \rightarrow Subway$$

# Metrics and Data Sets



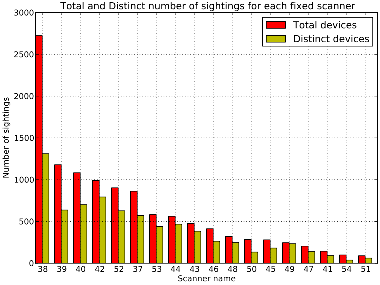
A data set of Bluetooth sightings by static nodes was used from Leguay at all, from 2006, where 18 static nodes tracked 9244 distinct users. This trace was replayed and complemented by a derived synthetic trace that expands the trace length and number of users.

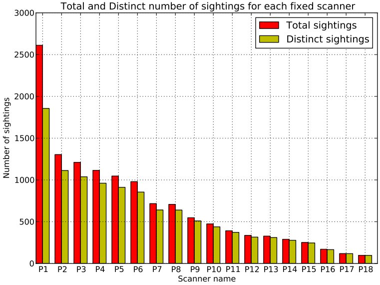
*Precedence Filters* false positives create fictitious transitions. For evaluation we observe the relative proportion of these transitions. A value of 0.5 means that 50% of transitions are false.

# Data Set: Location visits



Nelson Gonçalves, Rui José, Carlos Baquero
Universidade do Minho and INESC TEC, PT

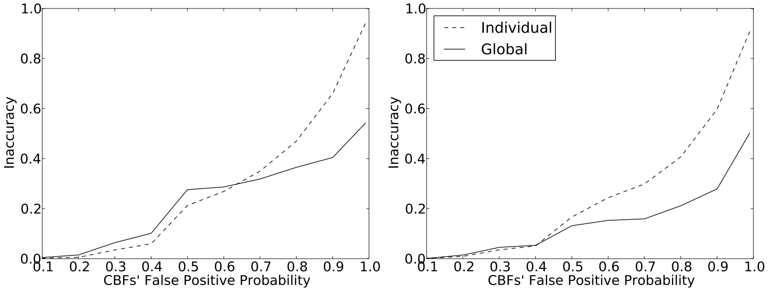



Distribution of detections in locations on real and synthetic traces

# Inaccuracy vs False Positive Probability





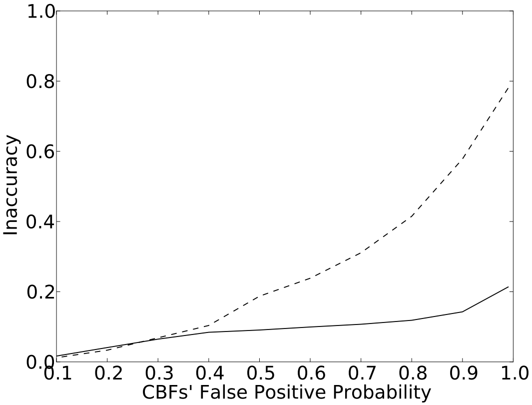

- Real and synthetic traces for the same trace length and users
- *Global* measures quality of aggregated transition prevalence

# Extended synthetic trace



Nelson Gonçalves, Rui José, Carlos Baquero Universidade do Minho and INESC TEC, PT

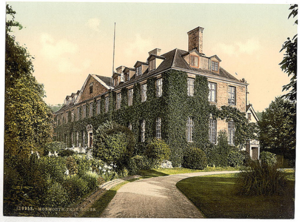
Effects of increased trace size (100) and tracked users (100000)




For longer runs higher quality aggregated data can be extracted from low quality (higher privacy) individual movement tracking.

# Take home message

- New technique, *Precedence Filters*, joins Bloom Filters and VCs
- Controlling filter size WRT number of devices, dictates accuracy
- False positives translate to fictitious visits to locations
- Proportion of fictitious visits supports plausible deniability
- 50% user inaccuracy can support aggregated 10% inaccuracy



# Photos

Attribution under Creative Commons.

- http://www.flickr.com/photos/skyjuice/
- http://www.flickr.com/photos/unknowndomain/
- http://www.flickr.com/photos/library_of_congress/