A short terminology

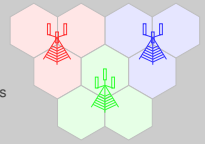
- Location Estimation - *Determining devices' physical location using properties of the data networks they are connected to.*
- Location Based Services (LBS) - *Services that provide value based on a person's or device's location. (maps, augmented reality, games, dating, etc.)*
- Location Provider - *Service that provides an estimated location using network Location Estimation*

Thesis' Four Main Parts

1. Suggesting a *privacy preserving*, *community sourced*, *open access* mobile location provider
2. Suggesting a new location estimation method tuned towards privacy
3. Creating a test system for testing location estimation methods based on field data
4. Gathering data and testing the suggested location estimation method and some of the more common methods and comparing them

Joke about that was as much as I managed to slim it down

Background - Cellular Networks



- Used to increase traffic capabilities
- Network divided into smaller cells
- Frequencies are re-used
- Cells are often sectored into three or more sectors
- Such small cells are great for location estimation
- Time Difference Multiple Access (TDMA) such as GSM and UMTS networks use timing data. Such timing data must be corrected for propagation delay, and can therefor be used for determining location.
- Neighboring cell information tracked and used for cell re-allocation

- Give example of frequency re-use. Explain what TDMA is (instead of own frequency, each unit gets a timeslot), how it relates to propagation delay and can be used for determining location.
- Cells are not really hexagons.
- Little bit about how neighboring cells work

Background - Location Estimation

- *Location Estimation* - Using features of a network to determine the spatial location of devices connected to said network.
- Any type of network information that can be translated to location can be used:
  - Signal Strength
  - Timing Data
  - ID of access point in use
  - Properties of received signal (angle, delay, etc.)
- In this thesis focus on methods using GSM/UMTS and/or WLAN networks

Background - Location Estimation

- Divided into three (often overlapping) types:
  - Network-based
  - Mobile-based
  - Mobile-assisted or hybrid
- In this thesis we focus on only Mobile-based methods
- Most common methods described in thesis. Here only the tested methods are shown.

Emphasis on more available in thesis

Background - Location Estimation Methods

- Cell Global Identity (CGI)
- Enhanced Cell Global Identity (E-CGI)
- Database Correlation Methods (DCM)
- Global Positioning System (GPS)
  - Navigation system using signals from Geo-stationary satellites
  - Often considered a mobile location estimation system
  - Here used for two things:
    - Providing true location when gathering fingerprints in the field
    - Quality control when testing location estimation methods

Background - Privacy - Cloaking

- Many different suggested methods
- All involve somehow hiding the client from the server, hence named cloaking
- Common methods:
  - Hiding one users among many
  - Hiding data among fake data
  - Onion routing
- Methods generally rely on a trusted third party cloaking service, a private network of clients, or both.

Say quickly what onion routing is. Can elaborate here if time.

[Motivation]

Two main motivational factors behind this thesis:

1. Ownership and payment
   - *Status Quo:* Corporations own your location. You have to pay to determine your own location with your privacy.
   - *Should be:* You own your own location. You should be able to determine your location freely without selling your privacy to a corporation.
2. Crowd sourced data and cloaking do not mix. Cloaking degrades crowd sourced data. By separating *location provider* from *LBS* this can be avoided, but then *location provider* must be *privacy preserving* by nature.

Also say something about the problem of licenses

Suggested Location Provider (Brief Summary)

- A system was suggested and used as a basis for creating a new location estimation method
- Started by determining threats to the system, and defined a set of goals
- Main discovery: conclusions on how to protect system and ensure privacy must be based on storage and transfer methods, which in turn must be based on location estimation method
- The open and privacy preserving nature results in a correlation between:
  - quality control
  - trust
  - incentive
  - precision

Say this will only be very brief, more in thesis

Suggested Location Provider (Brief Summary)

- In addition the following issues where addressed:
  - Data gathering:
    - Direct upload
    - Pre-generated database (estimated or gathered)
    - Clients amend query results if needed
  - Bootstrapping: *If system relies on amending queries, how to bootstrap a new area: No data exists to generate replies that can be amended*

Say that we looked into, but did not conclude, about the bootstrapping

Intersecting Areas Method

- Motivation: Combine the strengths of *DCM* with the simplicity of *CGI/E-CGI*
- Areas are stored surrounding all observations of a unique network measurement
- Areas are stored as convex hulls surrounding the extreme locations hence:
  - Small storage fingerprint
  - Few updates are needed
  - No stored data can be traced back to any individual
- Suggested improvements to areas for better precision:
  - Concave hulls
  - Limited areas, concave or convex hulls
- Location estimation: The intersection of the areas correlating to the network measurements in incoming fingerprint is calculated. The intersection, or the calculated center of the intersection is used as estimated location.

What is a unique network measurement

Intersecting Areas Method

Intersecting Areas Method

- Can fall back to *E-CGI* with no extra data or code when not enough data available
- Can fall back to *CGI* little extra data and code when not enough data available

Intersecting Areas Method - Benefits

[Intersecting Areas Method - Benefits]

- Low data transfer size and frequency (specially for updates)
- Embodies the simplicity of *CGI/E-CGI*
- Embodies the power of *CGI/E-CGI*
- Small storage, memory and processing footprint
- Extremely flexible and adaptive to different network equipment and data
- Used correctly ensures anonymity and privacy of stored data

Intersecting Areas Methods - Limitations

- Does not benefit the security and privacy of data transfer other than reducing the amount of updates needed
- *By design:* Precision cannot be gained using heuristics and statistics. Such methods require storing individuals' locations which is not compatible with privacy and open access

Make joke about nobody perfect

Test System

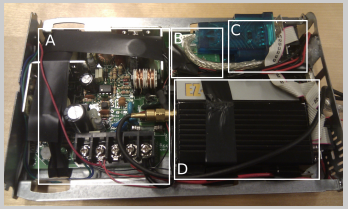
- Consists of three main parts:
  1. Data collection tools
  2. Back-end
  3. Data visualization tool

Test System - Data Collection Tools

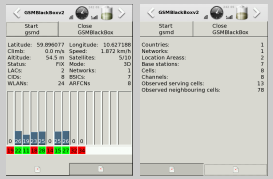
[Test System - Data Collection Tools]

- Hardware
  - Custom logging hardware
    - Created to be able to collect all information about all networks simultaneously in an area, including non-public GSM-networks
    - Less portable than mobile phone, but can be powered by any 9-24V power source for a long time
  - Android, Symbian and OpenMoko Phones
  - External or internal GPS
- Software
  - PC logger software for custom hardware logger
  - OpenMoko logger software for custom hardware logger
  - OpenMoko logger software
  - Android logger software
  - Symbian S60 logger software

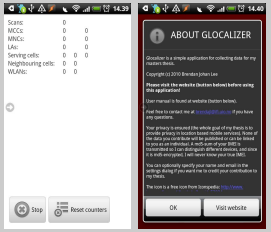
Hardware logger



OpenMoko Software



Android Software

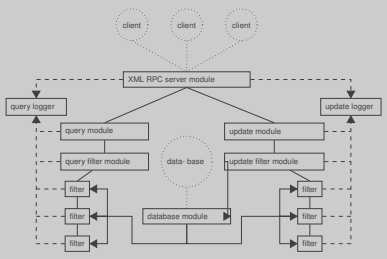


Symbian Series 60 Software

Test System - Back-end

- Created to gather data and test any location estimation method
- Completely modularize so location estimation, storage and communication methods are implemented as plug-ins
- Four main parts:
  1. Communication interface
  2. Storage/Database
  3. Query handler
  4. Update handler
- All communications and settings are logged so they can be re-played (possibly with different settings or estimation methods) at a later time

Test System - Back-end



Test System - Visualization

- Used for analyzing and visualizing gathered data and the result of location estimation methods.
- Renders maps or satellite imagery from web-services (Google maps, Bing maps, Openstreetmaps, etc.)
- Renders points, tracks and areas (polygons) on top of imagery
- Can fetch data directly from back-end database or load from files

Say that the example of Intersecting Areas method is a printout from this tool

[Tests]

- Data gathered with Android and Nokia handsets
- Algorithms tested:
  1. rxlevel *CGI* based on gathered not estimated *GSM/UMTS* data
  2. rxlevel *E-CGI* based on gathered *GSM/UMTS* not estimated data
  3. Simple, well described in literature *DCM* method, on *GSM/UMTS* serving cell and *WLAN*
  4. Simple, well described in literature *DCM* method, on *GSM/UMTS* serving cell and neighboring cells
  5. Simple, well described in literature *DCM* method, on *GSM/UMTS* serving cell, neighboring cells and *WLAN*
  6. Intersecting areas on *GSM/UMTS* serving cell and *WLAN* with and without *E-CGI* fall-back
  7. Intersecting areas on *GSM/UMTS* serving cell and neighboring cells with and without *E-CGI* fall-back
  8. Intersecting areas on *GSM/UMTS* serving cell, neighboring cells and *WLAN* with and without *E-CGI* fall-back

- For the exact handsets and amount of data, see thesis.
- For a description of the DCM method, see thesis (no time here).

[Two Individual Tests]

- First test
  - System trained on all data
  - Methods tested on all data one measurement at a time
    1. Remove training for tested point
    2. Run algorithm on measurement and log
    3. Re-add training for tested point
- Second test
  - Single dataset for Android, three for Symbian Series 60
  - Dataset randomly split in two
  - Half of set used for training, half for testing
  - Repeated on the virgin dataset 10 times
  - All algorithms tested over all datasets
  - Hence 30 Symbian and 10 Android tests for each algorithm
- Each test was done individually for Android and Symbian S60 data

Individually for each handset type since different data - S60 lacking neighboring cell info, and the possibility of comparing the performance on the different platforms/handsets

Problems and Results

- The penalty value for *DCM* is not static over different data sets, different areas and different handsets. Systems should therefore be continuously calibrated, which highly complicates using *DCM*
- The tests were comparable, only the second set of tests is presented here

No time to talk about the penalty value. Please see thesis.

Results - Training Time

Algorithm	Time on 8218	time on L7555
1	.000050	.000019
2	.000071	.000038
6	.047350	.017171
6.1	.047350	.017171
7	.027986	.024339
7.1	.027986	.024339
8	.075265	.041472
8.1	.075265	.041472

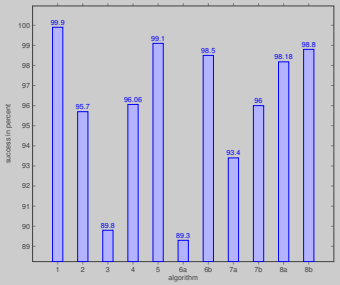
Nothing specific to note here other than that the time spent training is trivial compared to the time spent estimating locations and is also a one time event. And that training time for CGI and E-CGI as expected is much lower.

Results - Fingerprint Processing Time

Algorithm	Time on 8218	time on L7555
1	0.005383	0.008095
2	0.005621	0.008930
3	0.749295	11.088524
4	16.210149	17.802947
5	18.815485	14.625443
6	0.021477	0.023968
6.1	0.037931	0.008301
7	0.003671	0.003938
7.1	0.003632	0.004676
8	0.005185	0.005112
8.1	0.005067	0.006172

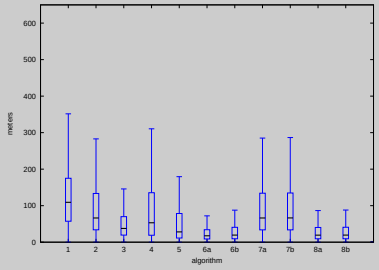
- Processing time for DCM MUCH higher than Intersecting Areas
- Processing time for Intersecting Areas somewhat higher than CGI AND E-CGI
- The weird unexplainable processor difference on algo 3

Results - Success Rate



- CGI clearly but not surprising highest success
- E-CGI fairly low, but could match CGI if fall-back to CGI was used
- Naturally DCM and Intersecting Areas relying on WLAN are much lower since WLAN not everywhere
- Intersecting Areas is somewhat outperformed by DCM
- However, when using fall-back to E-CGI Intersecting Areas outperforms or matches DCM

Results - Precision



- The Intersecting Areas method is the most precise method compared to both CGI, E-CGI (not surprising) and DCM.
- Methods relying solely on neighboring cells (not WLAN) have much lower precision, hardly providing any benefits compared to E-CGI.
- Methods relying on both neighboring cells and WLAN have a somewhat lower precision.
- This is due to neighboring cells being much larger than WLAN hot spots.
- However, relying solely on WLAN generally only works in urban areas with high WLAN-density.

Conclusions

- A privacy preserving, open access, crowd sources location estimation system is possible and will address the issues of
  - Privacy
  - Data ownership and payment
  - Location cloaking services degrading location estimation services

Conclusions

- The Intersecting Areas method is not only suited for a privacy preserving, open access, crowd sourced location estimation system, but has several other benefits:
  - Higher precision than standard DCM
  - Much lower memory, storage and processing fingerprint than standard DCM
  - The problem of the varying optimal penalty value of standard DCM is non-existent.
  - Provides a flexibility towards data, handsets, areas and future devices and technologies not found in the other tested methods.
  - Hence has a potential contribution also for other location estimation systems than the proposed
- We have discovered, and addressed, the need for a flexible location estimation test system allowing tests on any location data with any methods by anybody.

Future Work

- During the work on this thesis we have found enough possible future work for a small herd:
  - The suggested mobile location estimation system
  - The Intersecting Areas location estimation method
  - The location estimation test system
    - The same, or similar, mathematical optimization method suggested above should be implemented to allow filtering of training data
    - A module should be created to measure the density of training data needed for individual algorithms to perform and to perform optimal.
    - Several large datasets in different locations, both urban, sub-urban and rural should be gathered and released freely
    - The system should be polished and released freely

Skip everything except the first bullet point if no more time

Resources

- This slide show is located at
  http://opengsmloc.org/thesis/defence.pdf
- The thesis itself is located at
  http://opengsmloc.org/thesis/thesis-final-color-gloss.pdf
  and
  http://opengsmloc.org/thesis/thesis-final-print.pdf
- The software and code used in this thesis is located at
  http://opengsmloc.org/thesis/code.tar.gz
- The data used and generated in this thesis is located at
  http://opengsmloc.org/thesis/data.tar.gz