Datenschutz & UbiComp

RESEARCH GROUP FOR
Distributed Systems

Am Beispiel von Location Privacy

Betreuer: Marc Langheinrich

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Thierry Bücheler

Inhalt

- Einführung: Location Based Services
- Gegenwärtige Probleme
- Privacy & Anonymity
- Lösungsvorschlag 1
- Lösungsvorschlag 2
- Was wird eigentlich gelöst?
- Fazit: Was ist zu tun?

Location Based Services

- **Location Based Services** (**LBS**) sind Mehrwertdienste, die sich auf den aktuellen Standort des Users beziehen
- Optimalerweise wird der Standort **automatisch** erkannt
- **Positionsdaten** in Verbindung mit dem **personalisierten Benutzerprofil** des Nutzers erlauben ein massgeschneidertes Angebot

LBS - freiwillig?

- Bei Active Badge/Bat kann ein User trivialerweise auf die **Herausgabe seiner Ortsinfo verzichten**, indem er sie einfach nicht nützt.

  Dies ist aber **nicht möglich** bei grösseren Netzwerken wie z.B. GSM-Interfaces oder nationalen gesichtserkennenden Kameraprojekten.

Anwendungsbeispiel

subscribe

publish

LBS: Preis

Middleware

Tankstellenzone

Inhalt

- Einführung: Location Based Services
- Gegenwärtige Probleme
- Privacy & Anonymity
- Lösungsvorschlag 1
- Lösungsvorschlag 2
- Was wird eigentlich gelöst?
- Fazit: Was ist zu tun?

Das Problem

- Zugang zu Information kontrollieren
- Aber nicht gesamten Zugang sperren
- Gegensätzliche Ziele

Attacken

- Verschiedene Arten von Attacken:
  - **Locating/Lokalisierung**: Das Subjekt kann auf verschiedene Arten lokalisiert werden, z.B. aufdecken von LBS-Informationen oder Wireless-Signale triangulieren.
  - **Identification/Identifikation**: Der Real-world-Name oder z.B. die Adresse des Subjekts wird bekannt. Netzwerkadressen werden nicht als Identifikation des Subjekts bezeichnet, sondern als Pseudonyme, welche es ermöglichen können, ein Subjekt zu identifizieren.
- Lokalisierung und Identifikation sind z.T. verknüpft.

Was will man also:

- Von Vorteilen des Ortstrackings profitieren
- Privacy gewährleisten/schützen
- Wahre Identität des Users geheim halten

Inhalt

- Einführung: Location Based Services
- Gegenwärtige Probleme
- Privacy & Anonymity
- Lösungsvorschlag 1
- Lösungsvorschlag 2
- Was wird eigentlich gelöst?
- Fazit: Was ist zu tun?

Privacy

- Die Universal Declaration of Human Rights (UNO, 1948) deklariert, dass jedermann ein Recht auf Privatsphäre (Privacy) zu Hause, in der Familie und in seiner Korrespondenz hat.

Location Privacy

- Location Privacy ist eine **spezielle Art von** sogenannter **Information Privacy**, definiert als die Fähigkeit, andere daran zu hindern, den gegenwärtigen oder vergangenen Aufenthaltsort einer Person zu erfahren [1,3].

Location Privacy

- Der Schutz der Privatsphäre wird als genügend erachtet, wenn die Verfolgung einer Person (Ortstracking) einen **vergleichbaren Aufwand benötigt wie traditionelle Überwachungs-massnahmen** (Beschatten, Funksender an Auto haften etc.). [2]

Policies

- Eine gegenwärtig schon eingesetzte Praxis ist es, dem User von LBS sogenannte **Privacy Policies** vorzulegen, welche den User über die Datenverwendung des Providers informieren und dem User als Entscheidungsbasis dienen, Informationen freizugeben.

Anonymity

- Anonymität kann einen **hohen Grad an Privacy ermöglichen**, rettet Service-Nutzer davor, sich mit den Privacy Policies (Verträge) der Serviceprovider herumzuschlagen und reduziert die Anforderungen der Serviceprovider, private Information zu schützen [1,2].
- Anonymität wird daher als geeignetes Mittel empfunden, die Privatsphäre eines Nutzers ausreichend zu schützen.

Wo Daten schützen?

- Daten sind schwieriger zu schützen, wenn sie einmal auf einem System gespeichert sind.

- Der Vorschlag von Gruteser et al. [3] erhöht die Privacy durch verteilte Anonymitätsmechanismen direkt im (Sensor-)Netz.

Inhalt

- Einführung: Location Based Services
- Gegenwärtige Probleme
- Privacy & Anonymity
- Lösungsvorschlag 1
- Lösungsvorschlag 2
- Was wird eigentlich gelöst?
- Fazit: Was ist zu tun?

Gruteser & co.

- Die Mechanismen in diesem Paper anonymisieren die Daten **direkt in einem Sensornetz**. Somit muss sich der Provider nicht um die Sicherheit kümmern.
- Ein Problem bei vielen vorgeschlagenen Privacy-Lösungen ist, dass die **Middleware von nicht vertrauenswürdigen Institutionen gesteuert** werden kann, was die Anonymität brechen kann.

Gruteser & co.

- Gewisse Applikationen benötigen **aggregierte Statistiken** über die Popularität gewisser Orte (z.B. im Supermarkt), aber nicht unbedingt präzise Informationen über den genauen Standort einer einzelnen Person zu einem bestimmten Zeitpunkt.

→ „**Zählapplikationen**“

Gruteser & co.

- Drei Voraussetzungen müssen bei diesem System gewährleistet sein:

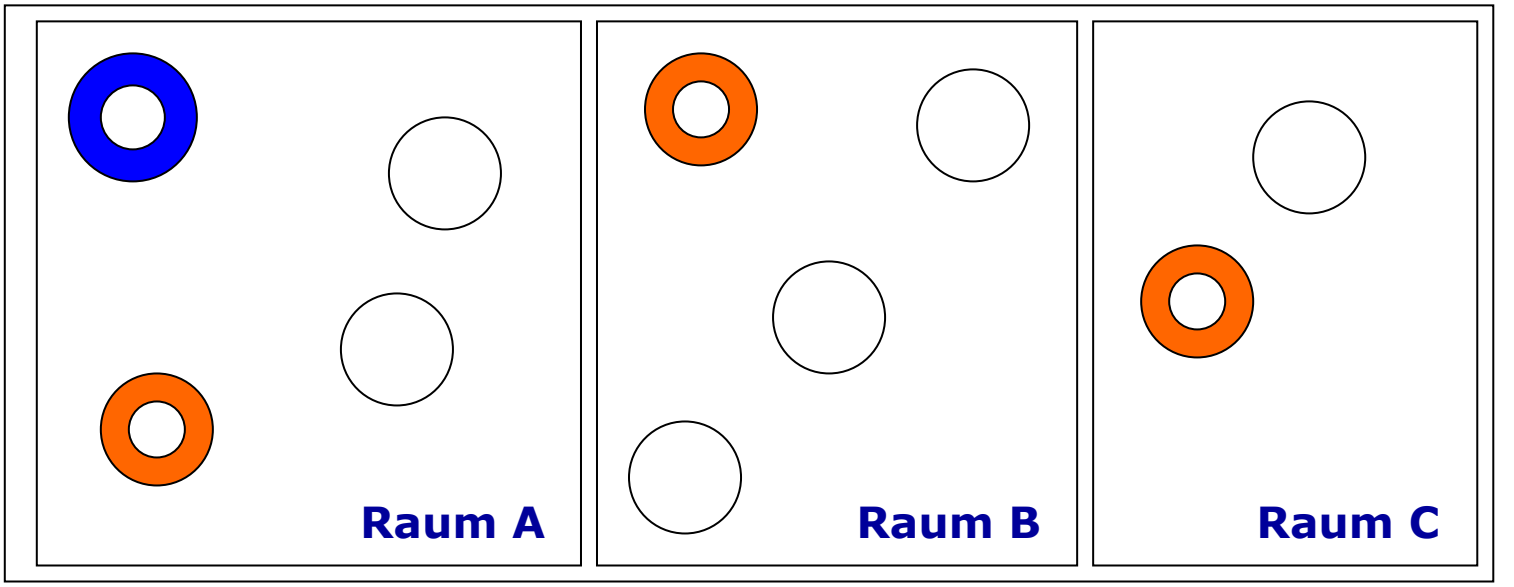
- Hierarchische Aggregation
- Data Cloaking
- Sichere Kommunikation

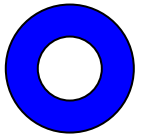
Hierarchische Aggregation

- Das Netzwerk aus Sensorknoten wird hierarchisch gegliedert
- Auf jedem Hierarchielevel benötigt man einen **CL** (**co-ordination leader**)
- Die IDs der Knoten werden mit einem „**Präfixcode**“ bestückt, d.h. Knoten haben dieselbe **Stockwerk-/Raum-ID** codiert, die IDs an sich sind aber alle eindeutig (unique)

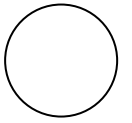
Beispiel mit CLs in Räumen

Stockwerk 1

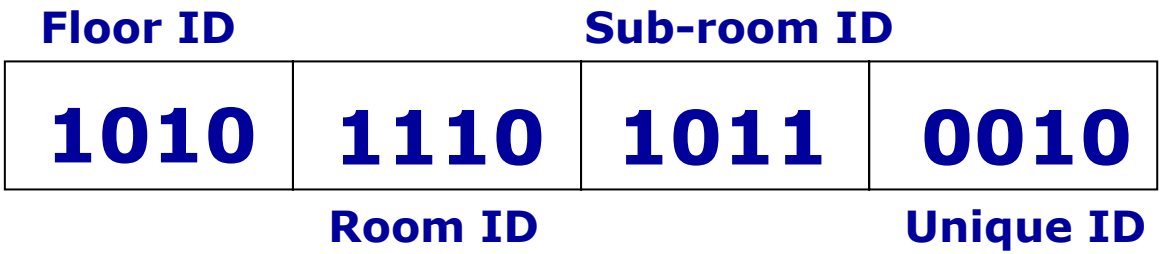


 **Stockwerk CL**

 **Raum CL**

 **Anderer Knoten**

IDs



K-Anonymität

- Dieser Begriff stammt von Samarati und Sweeney:

- Daten sind **k-anonym**, wenn jeder Ortsevent (erfasstes Subjekt) aus dem Netzwerk un-unterscheidbar ist von mindestens k-1 anderen Ortsevents (Subjekten).

Data Cloaking

- Knoten setzen zwei Techniken ein um die Anonymität zu erhöhen:
  - Biete weniger örtliche Genauigkeit
  - Verwische die Anzahl gezählter Subjekte in der Zone

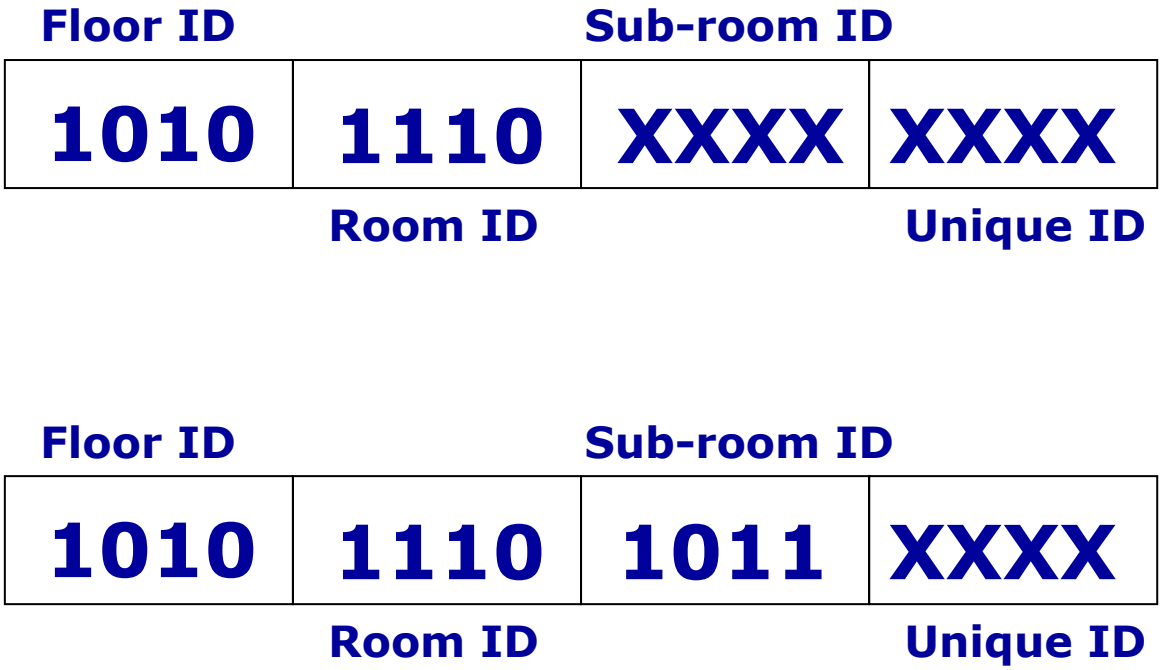
Data Cloaking

- Das Sensornetz verwischt die Daten, sodass das **k-Anonymitäts-Kriterium erfüllt** wird. Im Idealfall wird nur minimal verwischt, damit die Daten ihre Brauchbarkeit für möglichst viele Applikationen behalten.

- Daten können auf zwei Arten gecloaked werden:
  - Cloake die ID (Ort des Sensors), liefere präzise Date
  - Cloake die Daten, liefere die präzise ID

- Zwischen diesen zwei Varianten wird eine ausgewählt, indem die Anzahl Subjekte mit einem festgesetzten **Anonymitätslevel k** verglichen wird. Übersteigt die Anzahl Subjekte k, werden die Daten gecloaked und die ID wird präzis geliefert (und umgekehrt).

ID Cloaking


Floor ID
Sub-room ID
1010
1110
XXXX
XXXX
Room ID
Unique ID
Floor ID
Sub-room ID
1010
1110
1011
XXXX
Room ID
Unique ID

Sichere Kommunikation

- Knoten sollen mit **verschlüsselten** und **authentifizierten** Datenpaketen kommunizieren
- Um Traffic Analysis zu verhindern, ist jeder Knoten verpflichtet, **mindestens ein Paket** pro Datensammlungsintervall zu versenden (auch wenn kein Event eintritt)

Also:

- Lösungsvorschlag 1 ermöglicht v.a. eine **„Ortsunschärfe"** (die Verwischung der Daten agiert sozusagen als Zückerchen, welches nicht vertieft wird)
- Lösungsvorschlag 1 schützt v.a. vor **Lokalisierungsattacken**
- Es ist aber schwierig, diese Technologie auf einzelne Individuen und deren Tracking anzuwenden

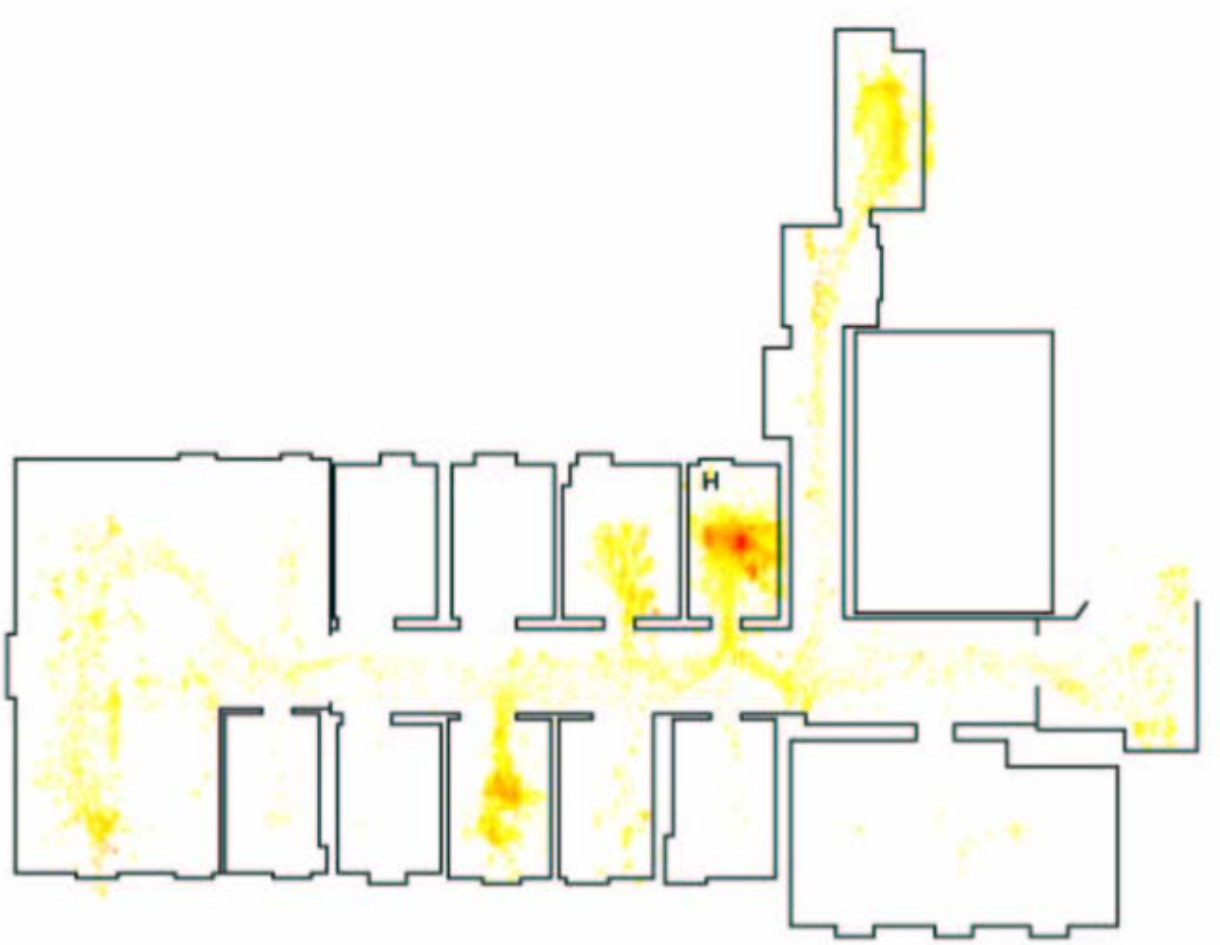
Inhalt

- Einführung: Location Based Services
- Gegenwärtige Probleme
- Privacy & Anonymity
- Lösungsvorschlag 1
- Lösungsvorschlag 2
- Was wird eigentlich gelöst?
- Fazit: Was ist zu tun?

Pseudonyme?

- Verschiedene Arten von LBS:
  - ID wichtig
  - ID unwichtig
  - Pseudonym genügt

Statische Pseudonyme



Beresford & Stajano

- Langfristige Pseudonyme sind nicht stark genug, IDs zu anonymisieren. Eine Lösung dieses Problems ist es, bei jeder Registrierung ein **ganzes Set von Pseudonymen** zu erhalten, und die Pseudonyme während des Trackings zu wechseln.

Szenario

- Normalerweise will man nicht, dass LBS-Informationen für einen bestimmten Ort einer **Applikation an einem anderen Ort** bekannt werden.
- Somit können die Event-Callbacks **an Zonen gebunden** werden.
- Was aber, wenn diese Zonen hinter unserem Rücken **miteinander Informationen austauschen**?

Beresford & Stajano

- Beresford und Stajano empfehlen ein „Privacy-protecting Framework“, welches auf **oft ändernden Pseudonymen** basiert, damit Benützer nicht einfach verfolgt oder mit beliebigen Orten in Verbindung gebracht werden können.

→ TRIVIALE LÖSUNG

Beresford & Stajano

- Wechselnde Pseudonyme sind aber auch problematisch, falls die räumliche und zeitliche Auflösung eines Systems gross genug ist
- Homes
- Footsteps

Wie wird's gelöst?

- Um dieses Problem zu beheben, führen Beresford und Stajano die „**Mix Zones**“ ein.

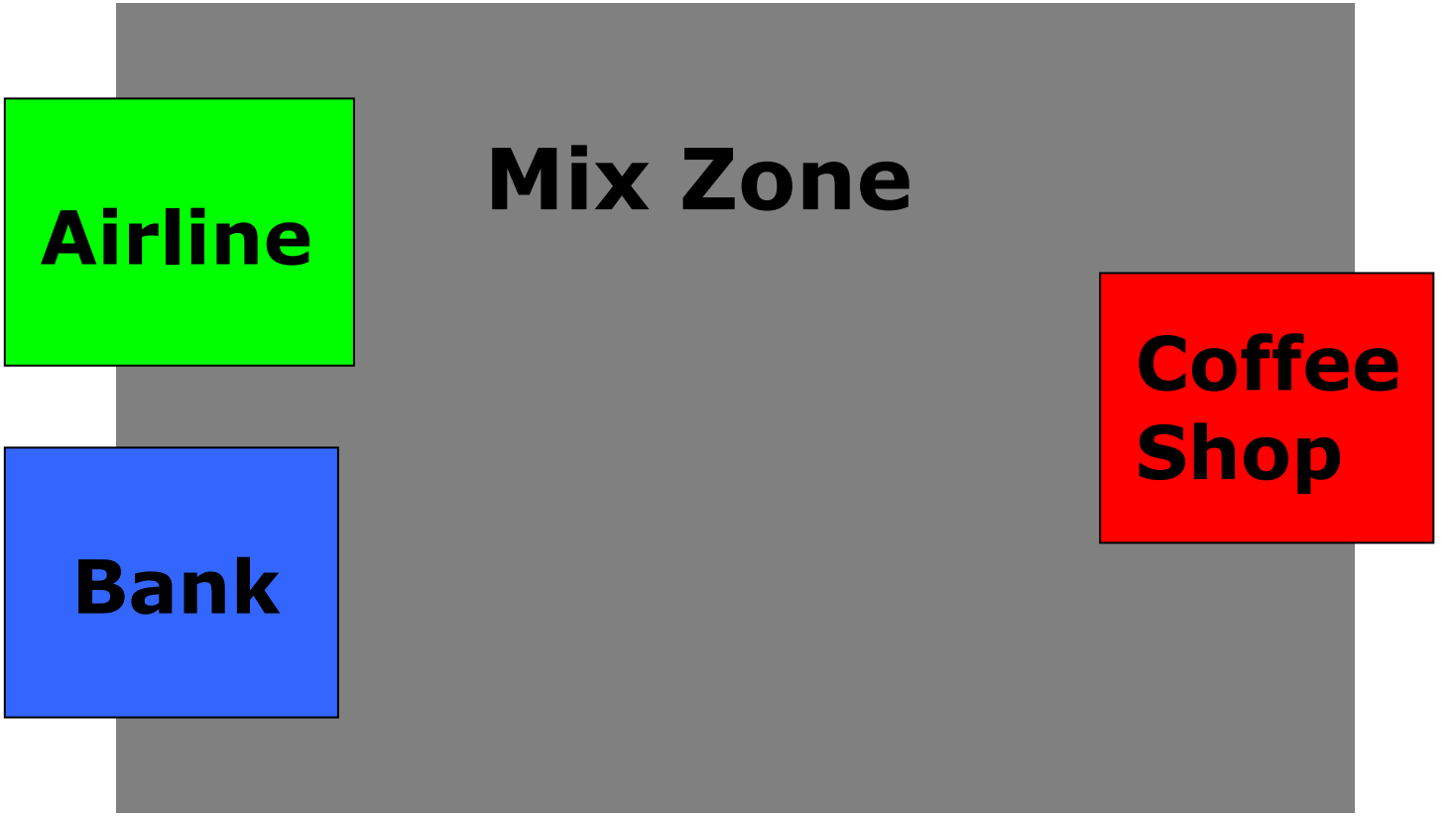
Mix Zones

- Mix zones gehen auf **Mix networks** und **Mix nodes** (von David Chaum) zurück.
- Store-And-Forward-Netzwerke
- n gleich lange verschlüsselte Pakete als Input
- Reihenfolge nach einer bestimmten Metrik (z.B. lexikografisch oder zufällig)

Mix Zone

- Beresford und Stajano definieren die Mix Zone als eine **örtliche Region maximaler Grösse**, in welcher **kein User einen Event-Callback registriert** hat.
- Im Gegensatz dazu sind „**Application Zones**“ Orte, an denen mindestens ein User Event-Callbacks registriert hat.
- Mix Zones können a priori von der Middleware definiert oder laufend berechnet werden.

Modell



Das Mixen

- Weil Applikationen **keinerlei Ortsinformationen** erhalten, während User in einer Mix Zone sind, werden Identitäten „gemixt“.
- Ein User ändert sein Pseudonym in einer Mix Zone und ein böser Beobachter kann keine Verbindung zwischen „eingehenden“ und „ausgehenden“ Usern herstellen.

K-Anonymität beim Mixen

- Die k-Anonymität wird erreicht, sobald **k Subjekte zur gleichen Zeit in einer Mix Zone** sind (oder die Mix Zone dementsprechend vergrössert wird)
- Ein User kann dann entscheiden, dass der **Anonymitätslevel k** ihm **genügend Privatsphäre** bietet.

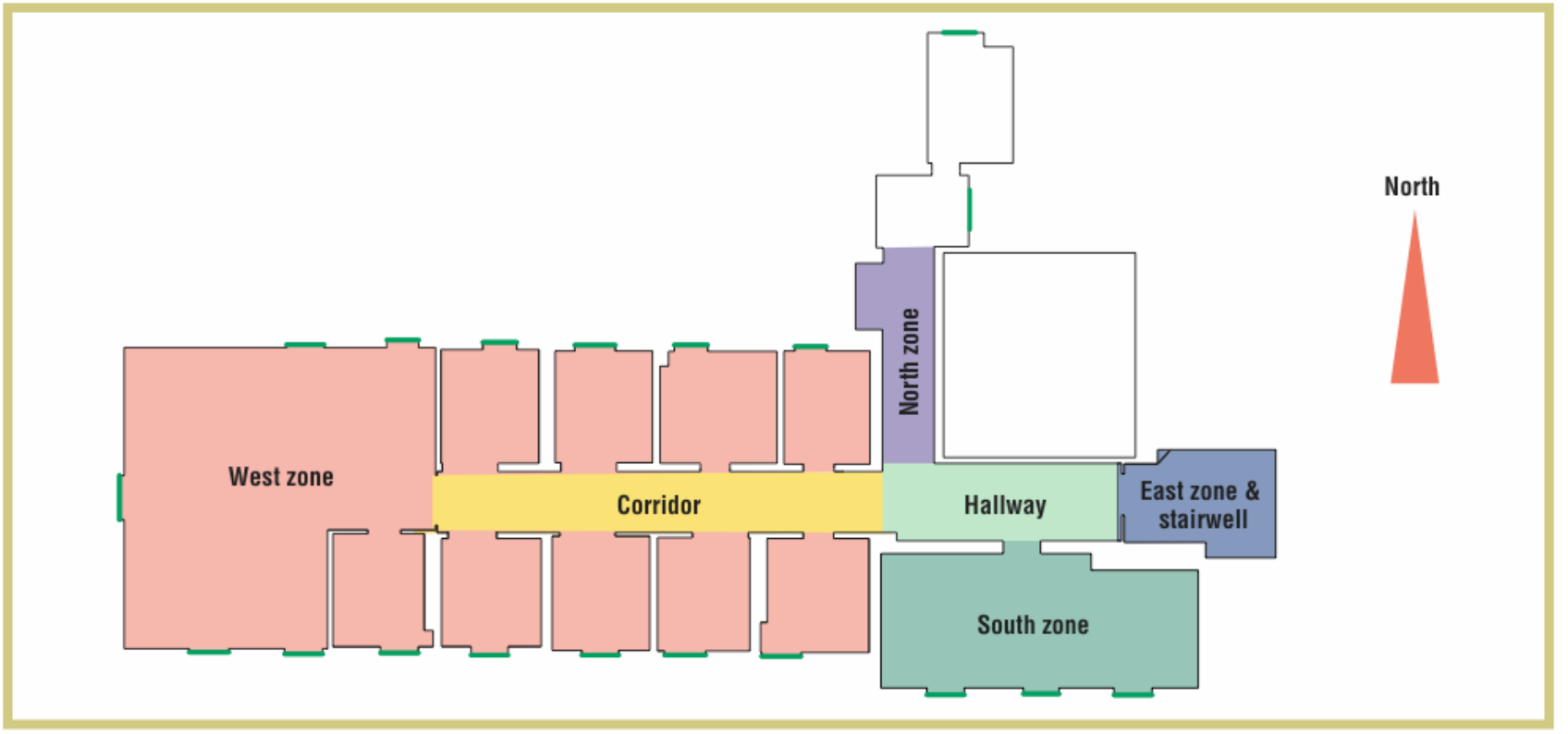
Aber: Geografie

- Es besteht eine **starke Korrelation** zwischen Eintritts- und Austrittspunkt in einer Mix Zone.
- Die k-Anonymität in einer Mix Zone ist nur ein gutes Mass, wenn **alle Subjekte** für einen bösen Beobachter **gleich interessant** sind. In diesem Fall haben wir die maximale Entropie (gemäss Shannon).

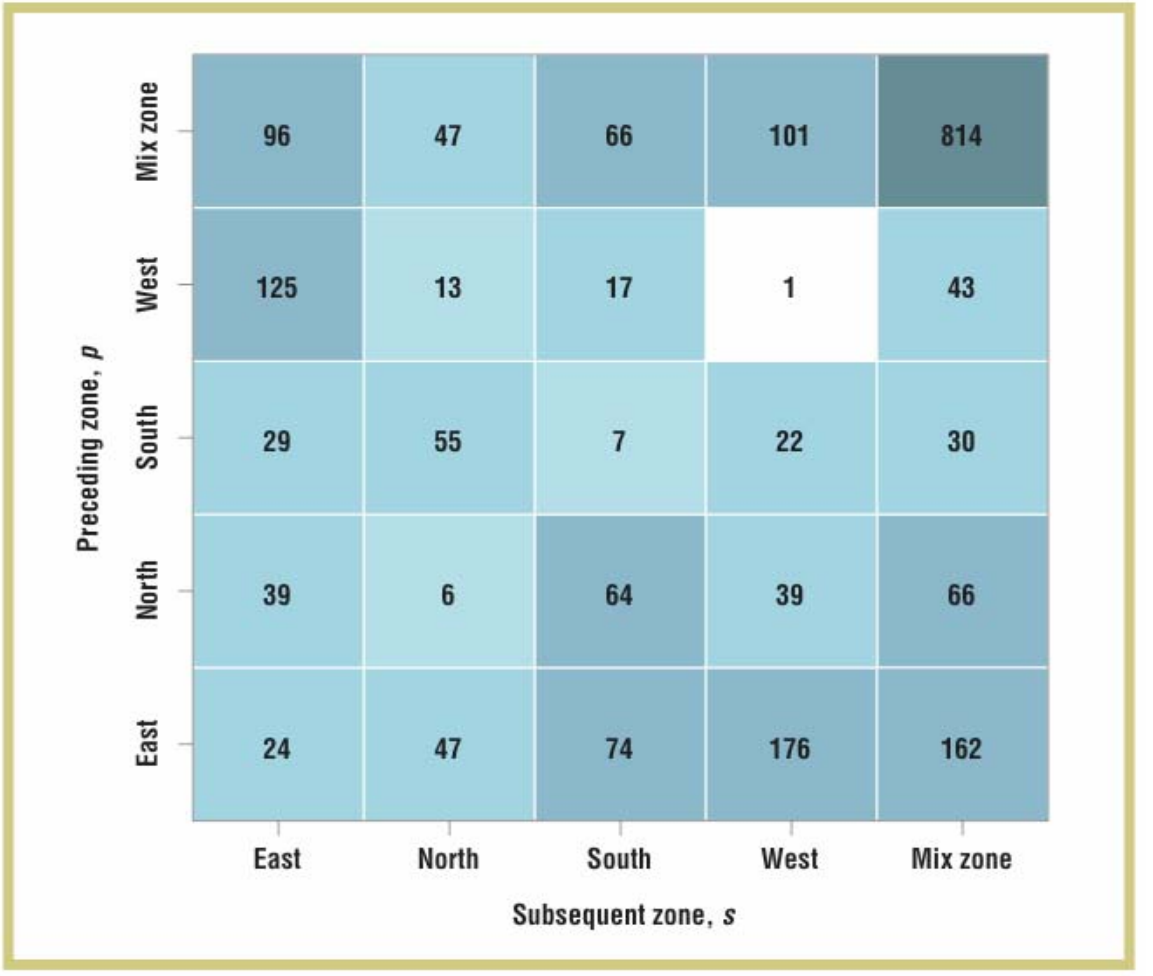
Theorie

- Beresford & Stajano haben in ihrem Feldversuch jeweils die **vorhergende Zone (p)** und die **nachfolgende Zone (s)** geloggt und eine Bewegungsmatrix erstellt.

Die AT&T Labs


North
North zone
West zone
Corridor
Hallway
East zone &
stairwell
South zone

Bewegungsmatrix



Ein bisschen Mathe

$$P(\text{vorher}=p, \text{nachher}=s) = \frac{M(p,s)}{\sum_{i,j} M(i,j)}$$

Bedingte W‘Keit

$$P(\text{nachher}=s \mid \text{vorher}=p) = \frac{M(p,s)}{\sum_j M(p,j)}$$

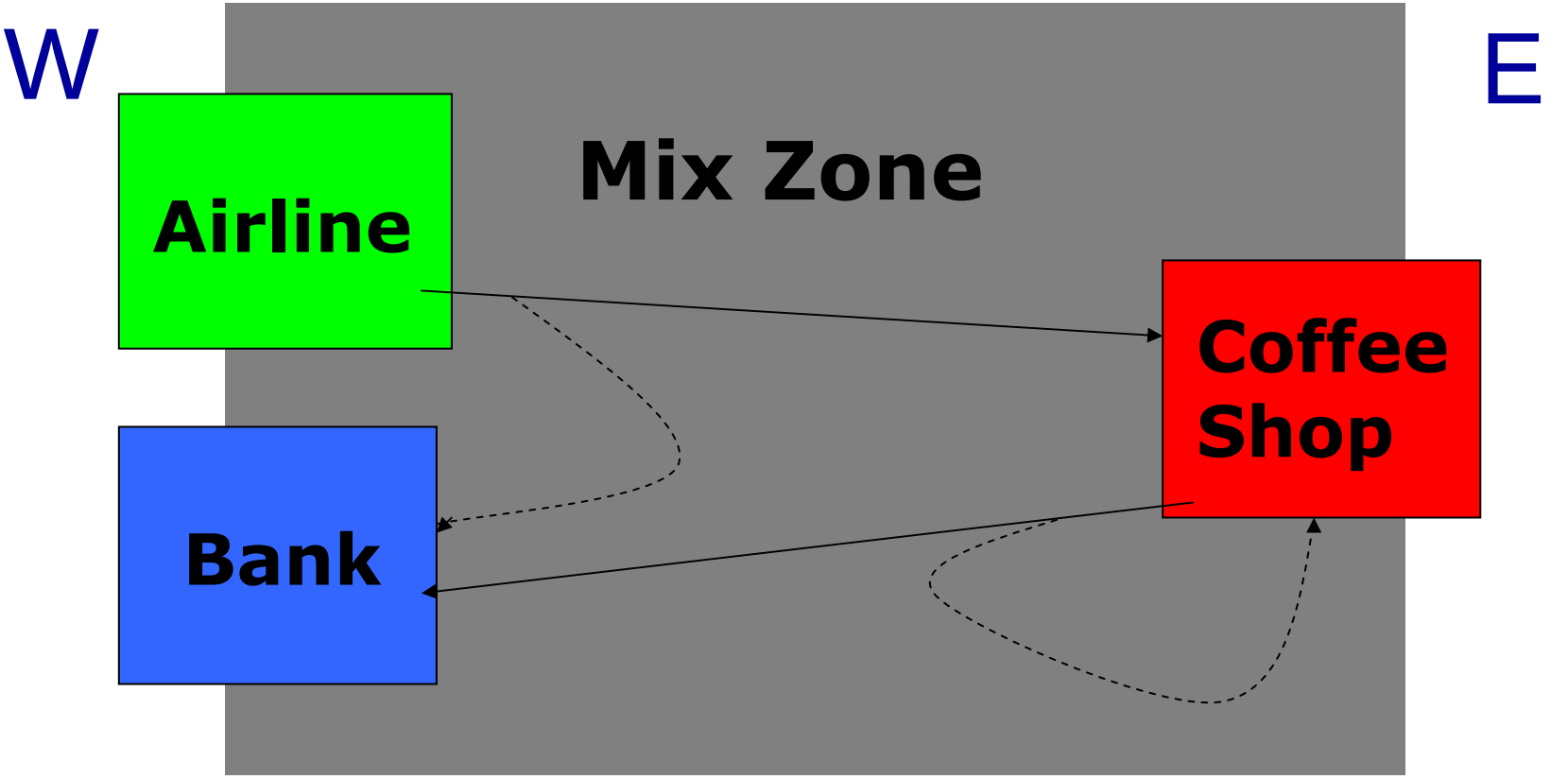
Entropie

$$H = -\sum_i p_i * \log p_i$$

Wieso Informationstheorie?

- Nehmen wir an, ein böser Beobachter verfügt über die **historischen Daten** der Bewegung (die Matrix).
- Ein Beispiel demonstriert, dass die **Anonymität geringer** ist, als die k-Anonymität suggeriert:
- Wir vergleichen die Wahrscheinlichkeit, dass zwei Subjekte von **Ost nach West** (bzw. umgekehrt) gehen, mit der Wahrscheinlichkeit, dass sie einen „**U-Turn**“ machen.

Modell



Vergleich (Herleitung)

- Wahrscheinlichkeit beobachtet:

  EEWW v EWWE v WEEW v WWEE

- Wahrscheinlichkeit U-Turn:

  EEWW v WWEE

Numerische Lösung

- Die Wahrscheinlichkeiten werden z.B. so errechnet:

p(EWWE) = p(EW) * p(WE)

Die Wahrscheinlichkeiten stammen aus der Matrix (mit den obigen Formeln):

p(beobachtet) = 0.414

p(uturn) = 0.0005

Bedingte W‘Keit

- Wir wollen wissen, wie gross die Wahrscheinlichkeit ist, dass beide Subjekte einen uturn gemacht haben:

$$p(\text{uturn}|\text{beobachtet}) = \frac{p(\text{uturn} \wedge \text{beobachtet})}{p(\text{beobachtet})} = 0.001$$

Entropie ist genauer!

- Der Informationsgehalt des Resultats ist also nicht 1 Bit, wie wenn die Wege gleichwahrscheinlich wären, sondern viel kleiner.

- Die Entropie beträgt in diesem Fall 0.012 Bits.

Also:

- Lösungsvorschlag 2 ermöglicht v.a. eine **„ID-Unschärfe"** (mit Hilfe der Pseudonymisierung)
- Lösungsvorschlag 2 schützt v.a. vor **Identifikationsattacken**
- Lösungsvorschlag 2 hat gezeigt, dass Pseudonymisieren nicht so einfach ist, wie es scheint und darum einige Anforderungen an die Sicherheit gestellt

Inhalt

- Einführung: Location Based Services
- Gegenwärtige Probleme
- Privacy & Anonymity
- Lösungsvorschlag 1
- Lösungsvorschlag 2
- Was wird eigentlich gelöst?
- Fazit: Was ist zu tun?

Was wird eigentlich gelöst?

- Gruteser und Grunwald
  - Dieser Approach stärkt Privacy-Schutz im Vergleich zu Lösungen auf dem DB-Level weil er das Sammeln von Privacy-sensitiven Daten verhindert
  - Bei „Bulk“-Applikationen (und allgemeiner Applikationen, die eine Ortsunschärfe erlauben) kann der Ort eines Einzelevents (-Subjekts) verschleiert werden.
  - Man kann eine gewisse (unscharfe) Awareness schaffen, die bei Bedarf (z.B. Emergency) geschärft werden kann (die Sensoren sind ja präzis verteilt)

Was wird eigentlich gelöst?

- Beresford und Stajano
  - Dieser Approach ermöglicht es einem Individuum, das Tracking zu verhindern
  - Es wird ein Mass für Pseudonym-Switching hergestellt und gezeigt, dass die k-Anonymität einer Mix-Zone nicht als Qualitätsmass genügt
  - Es ist möglich, pseudonymisierte Services zu nutzen, solange man ab und zu mixt.
- Wer ist besser?
  - Ortsunschärfe-Apps vs. ID-Unschärfe-Apps: verschiedene Anwendungen

Was geht (nicht)?

- Apps wie „Wo ist X?“ oder „Ich möchte ein Taxi an meine Adresse“ gehen nicht.
  - Hier darf weder der Ort noch die ID verschleiert werden.

Inhalt

- Einführung: Location Based Services
- Gegenwärtige Probleme
- Privacy & Anonymity
- Lösungsvorschlag 1
- Lösungsvorschlag 2
- Was wird eigentlich gelöst?
- Fazit: Was ist zu tun?

Fazit

- Für viele LBS ergibt die **Verwischung von Ortsinformationen wenig Sinn**. In diesem Fall ist es angebracht, an der Verwischung der Identität zu forschen, um den User anonym zu halten.
- Für verschiedene LBS-Arten müssen **verschiedene Privacy-Modelle** verwendet werden.

Fazit 2

- Ein sinnvolles Ziel ist die **Datensparsamkeit** (data minimization). Man will nur so viele Daten verbreiten wie wirklich benötigt werden.
- Lösung 2 geht davon aus, dass das häufige Peudonym-Wechseln keine Applikationen verunmöglicht. Stimmt das? Wie werden die Pseudonyme mit den wahren IDs gelinkt? Ist das nötig? Gibt‘s dann doch wieder Policies/ Contracts?

Fazit 3

- Benötigt wird ein **formales, wenn möglich probabilistisches Modell für Lokationsanonymität**, welches einen kontinuierlichen Datenstrom liefert. Dies würde eine bessere Evaluation ermöglichen.

Letztes Fazit

- Ein besseres Verständnis über die **Anforderungen an die Genauigkeit** von Lokationsdaten für **verschiedene Klassen von Applikationen** würde eine Analyse des Anonymitätslevels für LBS- und Sensornetz-Applikationen erleichtern.

Fragen?

Kommentare?

Diskussion?

Literatur

- [1] Alastair R. Beresford & Frank Stajano: *Location Privacy in Pervasive Computing*, Security & Privacy Magazine, 2003
- [2] Gruteser & Grunwald: *Anonymous Usage of LBS through Spatial and Temporal cloaking,* 2003
- [3] Gruteser, Schelle, Jain, Han, Grunwald: *Privacy-Aware Locatio Sensor Networks*, Proceedings of HotOS IX, 2003
- [4] Gruteser, Grunwald: *A Methodological Assessment of Location Privacy Risks in Wireless Hotspot Networks,* 2003
- [5] Pfitzmann & Köhntopp: *Anonymity, Unobservability, and Pseudonymity - A Proposal for Terminology*, Draft v0.14, 2003