Datenkrake im Eigenbau

Ein Selbstversuch zur Handyüberwachung

Albert Rafetseder, Patrik Hummelbrunner,
Florian Metzger, Lukas Pühringer

Metaday 60

11. Oktober 2013

Überblick

Unsere* Motivation

- Alle sammeln Daten:
  - AAPL, GOOG, FB
  - Supermärkte, Ärztinnen und Ärzte, die Wirtschaftskammer, ...
  - ISPs und Handynetzbetreiber*
- Warum wird das gemacht?
  - Pragmatische InformatikerInnen *vs* Datensparsamkeit, -schutz („speichere, was anfällt, es wird noch nützlich sein" *vs* „so wenig und so kurz wie möglich")
  - Betriebliche Notwendigkeit*, „Netzausbau"*
  - Werbung — nicht der Dienst ist das Produkt, sondern die NutzerInnendaten
  - Strafverfolgung — `assert(ich==mein Handy)`

Unsere* Motivation (2)

- Was uns meistens bleibt, ist ein dumpfes, ungutes Gefühl – Kontrollverlust, Misstrauen, schlechte Erfahrungen
  - Was passiert mit den Daten?
  - Wer hat darauf Zugriff?
  - Wie lange werden sie gespeichert bleiben?
  - Was verraten die Daten über mich?*
- Aber wir leben ja in der Zukunft! Wir haben Techniken, um diese Frage zu erforschen!
  - In unserem Fall: Programmierbare Smartphones mit Sensoren, reichlich Speicherplatz, Phantasie, Skriptsprachen
- Also probieren wir's mit *nüchterner Forschung*.

Was wird denn überhaupt gesammelt?

- Kurze Antwort: Alles.
- Pragmatisch – Bits sind billig, Daten vielleicht brauchbar.
- Im Falle der Handynetzbetreiber müssen wir unterscheiden:
  - Passive, mobile NutzerInnen: aktuelle Cell IDs und Routing Areas für den aktuellen Betrieb
  - Telefonate und SMS: **„Call Data Records" (CDRs)**, also Uhrzeit, anrufende und gerufene Nummer, Länge, Mobilität [EP020]; Inhalt *natürlich nicht* (außer... vielleicht)
  - Datenverkehr: Alles, was offen in den Headern steht (IP-Adressen, Protokolle, Ports); vermutete Verkehrsklassifikation (Web, P2P, VoIP, Streaming); alles mögliche andere

Was machen die Operator daraus?

- Auf das Netz bezogen
  (Das machen die Operator tagein, tagaus)
  - Muster erkennen, Netzkapazität dimensionieren
    Z.B. Wo sind wann viele Anrufe abzuarbeiten?
  - Störungen erkennen – Änderungen in bekannten Mustern
- Auf den/die EinzelneN bezogen
  (Uns ist nicht bekannt, dass die Operator das ohne Rechtsgrundlage machen; wenn, dann verkaufen sie's nicht offen)
  - Name, Geburtsdatum, Wohnort sind oft bekannt
  - Muster erkennen, Störungen erkennen
  - Wann am Tag, in der Woche, im Monat werden welche Nummern von wo aus angerufen? Wann hebt NutzerIn nicht ab? Wann läuft das Gerät (nicht)? Wie sind die Anruflängen nach Nummer verteilt?
  - Mit der gebotenen Vorsicht formuliert: Rückschlüsse auf NutzerInnenverhalten durchaus möglich

Wie führt man eine solche Sammlung durch?

- Kurze Antwort im Mobilfunknetz: Dedizierte Infrastruktur.
- (Große) Teile der Kern-Infrastruktur dienen solchen und anderen Mess-, Zähl- und Abrechungsaufgaben
  (Das heißt nicht, dass die Nutzdatenbehandlung wenig aufwändig wäre)
  - Tarife je nach Uhrzeit, Quell-/Zielnetz, Freiminutenstand und sonstigen Aktionen
  - *Wirklich viel* Verkehrsaufkommen
    (2012 [FMK]: 13e6 SIMs, 23e9 Minuten, 8e9 SMS, 73e15 Bytes)
  - Relevanz für Netzbetrieb, daher Aggregation
- Kurze Antwort für das Endgerät wie in unserem Fall:
  Eine eigene App. (Details folgen in Kürze)

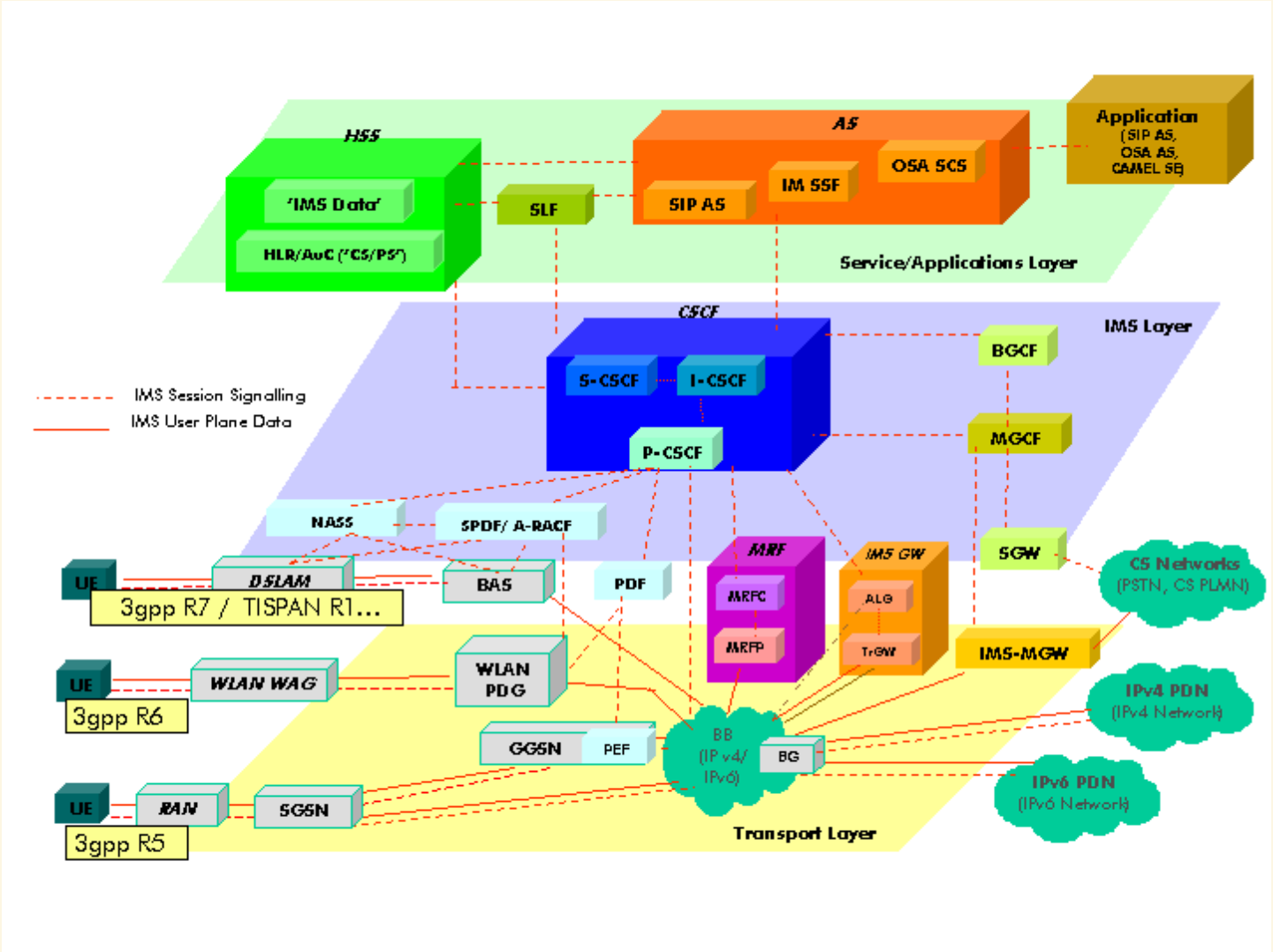
Mobilfunk-ABC

Hier nur das Notwendigste, um bei unserer Geschichte nicht den Überblick zu verlieren:

- Funkzellen!
- Überlappung für nahtlosen Betrieb
- Paging, RAs zum Weiterleiten von eingehenden Anrufen
- Leider, aber: „Das Netz" muss (über maximal einige Stunden hinweg) wissen, wo Du bist, damit Du erreichbar bleibst
- Abseits rechtlicher Erfordernisse (VDS / SiPoG?) gibt's aus Sicht des Netzbetriebs IMO *keine* Gründe, CDRs länger aufzuheben als vielleicht fünf Minuten nach dem Anruf, zur Verrechnung.

Mobilfunkarchitektur im Überblick

Einfach zum Genießen, ich werde nicht detaillierter darauf eingehen. [WIKI3G]



Unsere Sammel-App

- Für Android ab 2.1 (war halt zur Hand)
- Sammlung als Hintergrunddienst
- Auswertung erfolgt offline mit Python, AWK, Gnu R
- Gesammelt werden
  - Zeitstempel der aktuellen Messung
  - Status des Smartphones
    - Anzahl von eingehenden und ausgehenden Anrufen und SMS
    - Verpasste Anrufe
    - Datenverbindung aktiv? Up/down?
  - Informationen zur den Mobilfunkzellen
    - MCC, MNC, LAC, Cell ID, Verbindungstyp (LTE ... GPRS)
    - Signalstärke
  - Standort (abgeschätzt über Android-API)
- Außerdem: Datenschnorchler zum Auslesen von Telefonbuch, SMS-Speicher und Anrufliste

Statistiken zur Sammlung

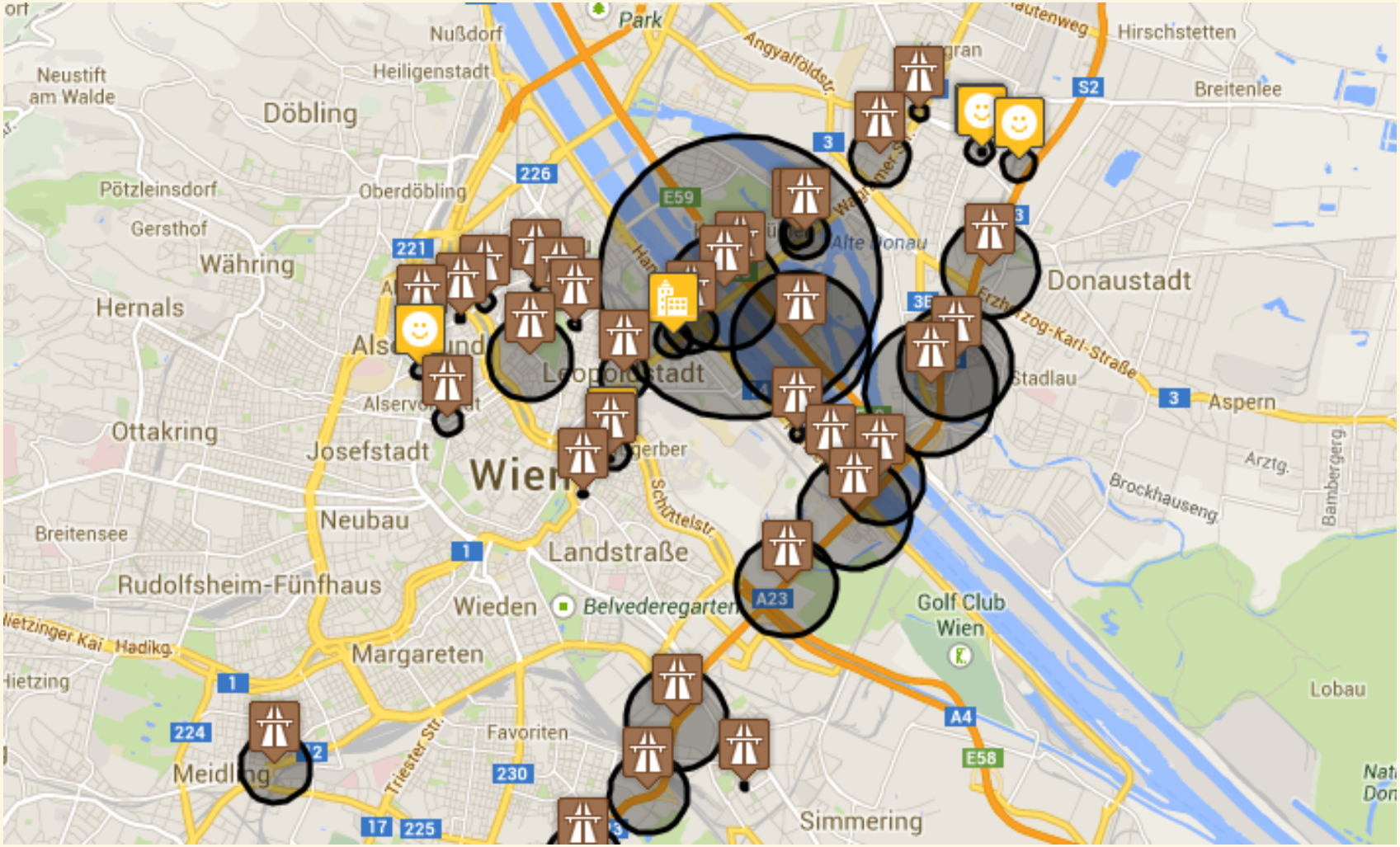
- Beobachtungszeitraum 24 Tage im Juni 2013
- 342,544 Datenpunkte
- 62 verschiedene Zellen besucht
- 48 Stunden längster Aufenthalt in einer Zelle
- 300 Stunden kumulierter Aufenthalt in der meistbesuchten Zelle
- 14/1 Anrufe bekommen/getätigt
- 118/68 Sms bekommen/gesendet

- 3 Anruflisten ($\sum$ 1.000 Einträge, ca. 300 Kontakte)
- 2 SMS-Listen ($\sum$ 1.300 Stück)

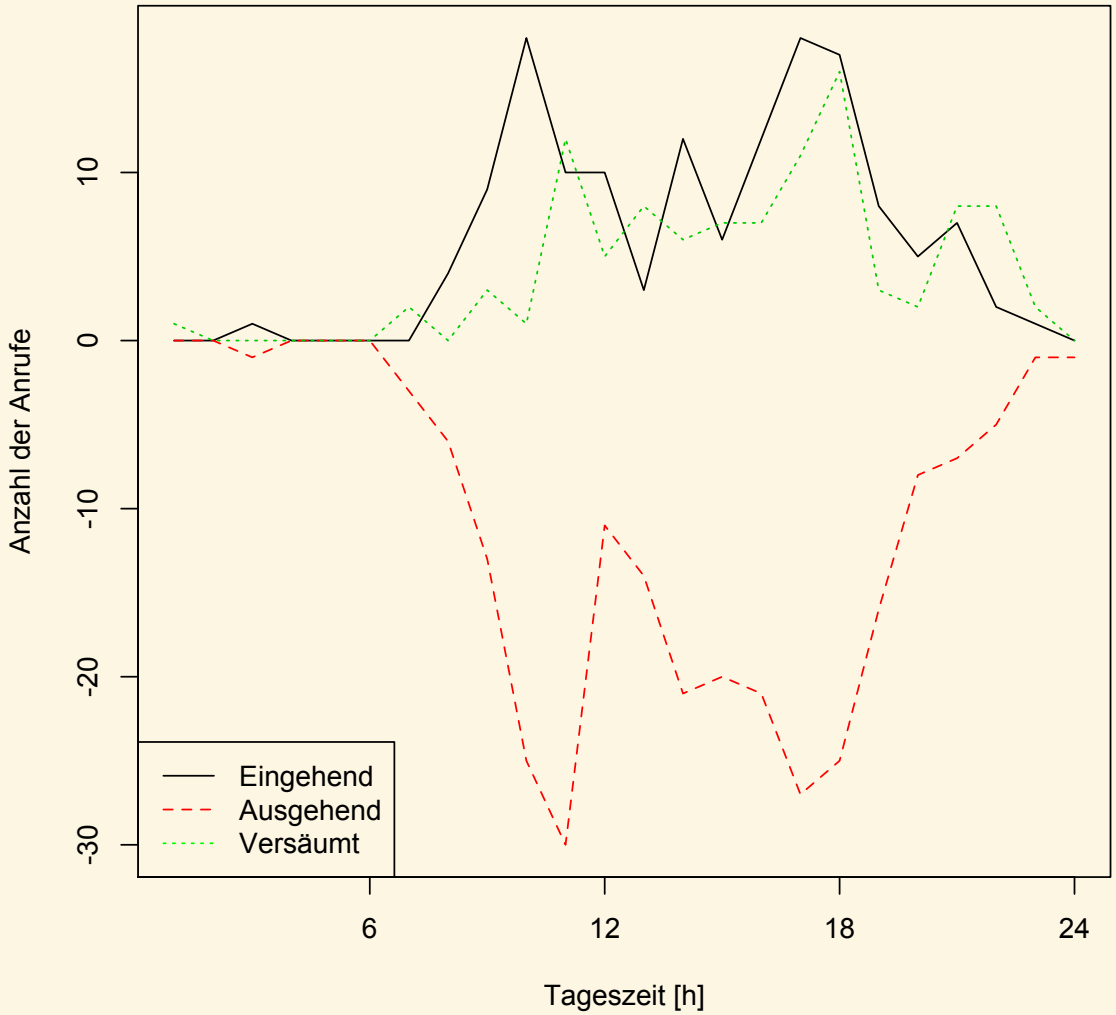
- Großes Ziel: weitgehend automatisierte Auswertung

Mobilität und Aufenthalt

Marker (Bedeutung): Autobahn (Transitzelle),
Smiley (langer Besuch), Häuschen (sehr langer Besuch)

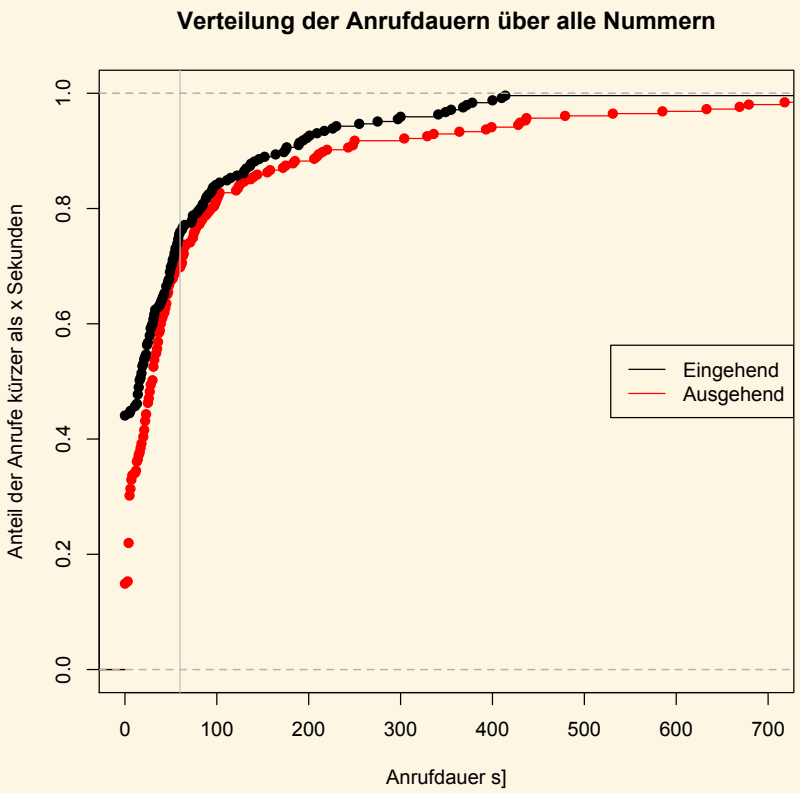


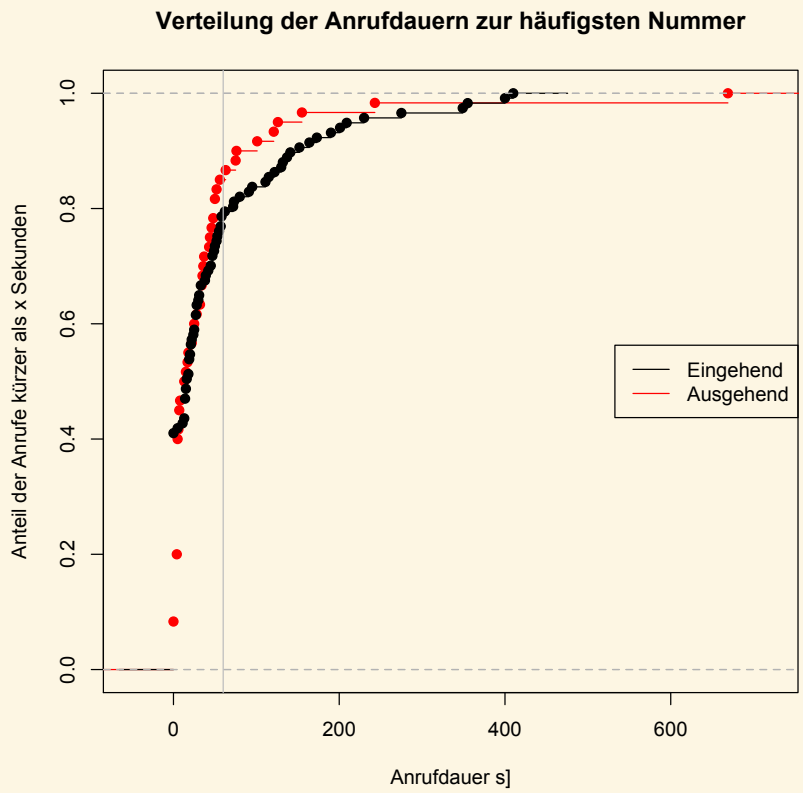
Meine Anrufhäufigkeit



Verteilung meiner Anruflängen

Links: Über alle Nummern
rechts: zur häufigst angerufenen/anrufenden Nummer





Sprache und Worthäufigkeit in SMS

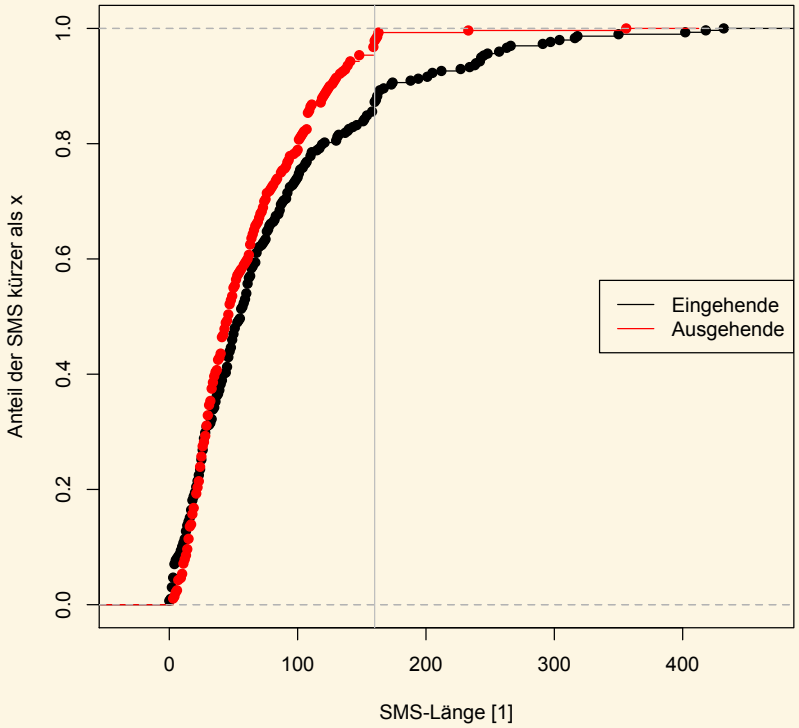
- *Buchstabenhäufigkeit*, gemessen bzw. aus Wikipedia:
  - `eniarst hudmlcgo bkwf zpvjyxq` (ich)
  - `enisrat dhulcgmo bwfk zpvjyxq` (de)
  - `etaoins hrdlcumw fgyp bvkjxqz` (en)
  - `aoiezns cwryldkt mpju bhgżfvxq` (pl)

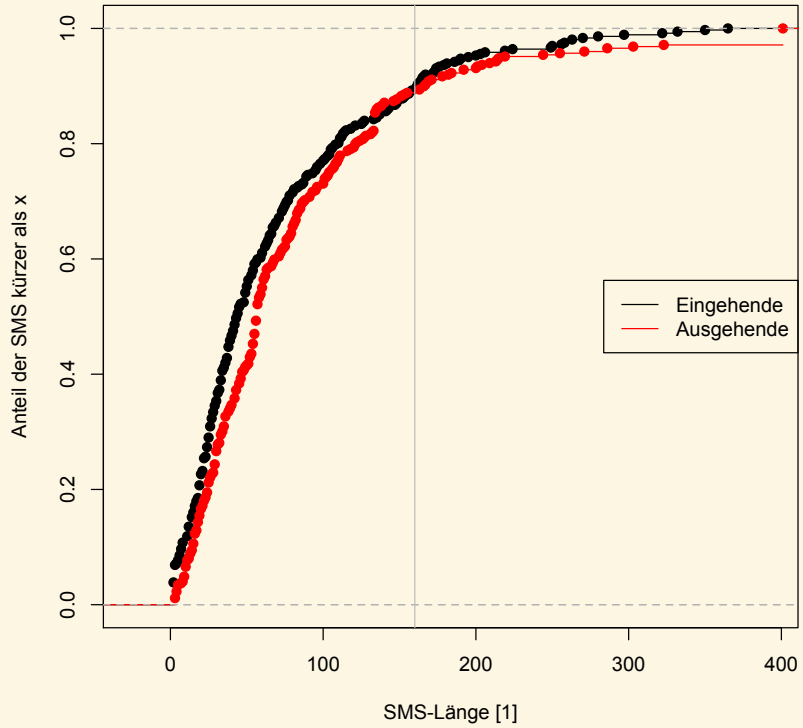
- Die 10 häufigsten Worte der beiden Listen:
  - ich – und – in – **Albert** – der – bin – im – auf – wir – ein
  - Bussi – ich – wir – und – ja – nicht – ist – auch – der – das

- Interessant sind aber eigentlich die selteneren Wörter – Seltenheit $\approx$ hoher Informationsgehalt (Entropie)
  - Metallsägeblatt – QRV – Ribisel – tenemos
  - Bussibussi – Echtzeit – Pumphebel – Tarifwechsel

Verteilung der SMS-Längen

Links: ich, rechts: einE FreiwilligeR

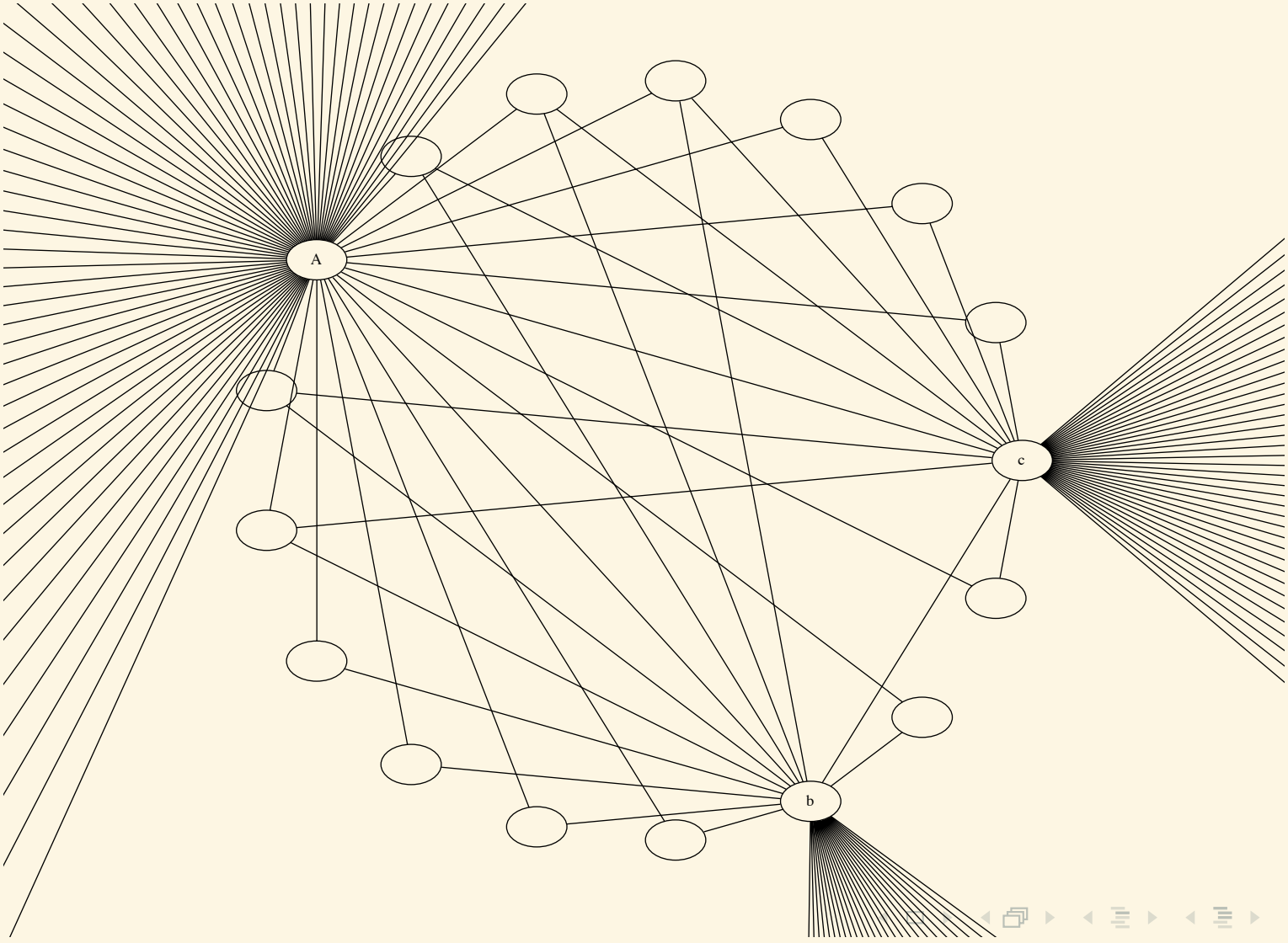




Bekanntschaften gemäß dreier Call Logs

„Fächer”: einzigartige Nummern

„Innerer Kreis”: Gemeinsame Bekannte (?)



Zum Vergleich: Facebook

Quelle [FB] im Anhang

Wohin nun weiter?

- Also ja, man kann schon einiges aus den Daten lernen.
- Was können wir tun?
- Kurze Antwort: kurzfristig lautstark unzufrieden damit leben.
- Unvermeidliches: Handy – Mast
- Ansonsten sollte vieles anders gelöst werden können
  - Möglichst wenig speichern
  - Lokal nötige Information lokal lassen
  - Randomisierung
  - „Neue" Algorithmen wie Oblivious Transfer, Secure Computational Geometry und Mental Games
- GesetzgeberInnen in aller Welt, Betreiber sowie Normungsgremien (3GPP) sind gefordert!

Ergänzende Ansätze

Was geht am Endgerät?

- Richtung Mobilfunknetz leider nicht viel, da Standardisierung
- Richtung Apps hingegen schon!
- Android 4.3 AppOps / Cyanogenmod 10.2 Privacy Guard [CM]
- TaintDroid: Genau kontrollieren und beschränken, welche Sensordaten welchen Apps zur Verfügung stehen
- Permission Spoofing Framework: Weniger invasiv — belüge alle, die unbedingt Daten wollen, und seien es Systemanwendungen
- Feinergranularer Ansatz: Detailgrad der Daten (Vollzugriff, gerundeter Wert, Salt+Hash, kein Zugriff) und Zugriffshäufigkeit pro App steuern

So! Genug der Dystopie!

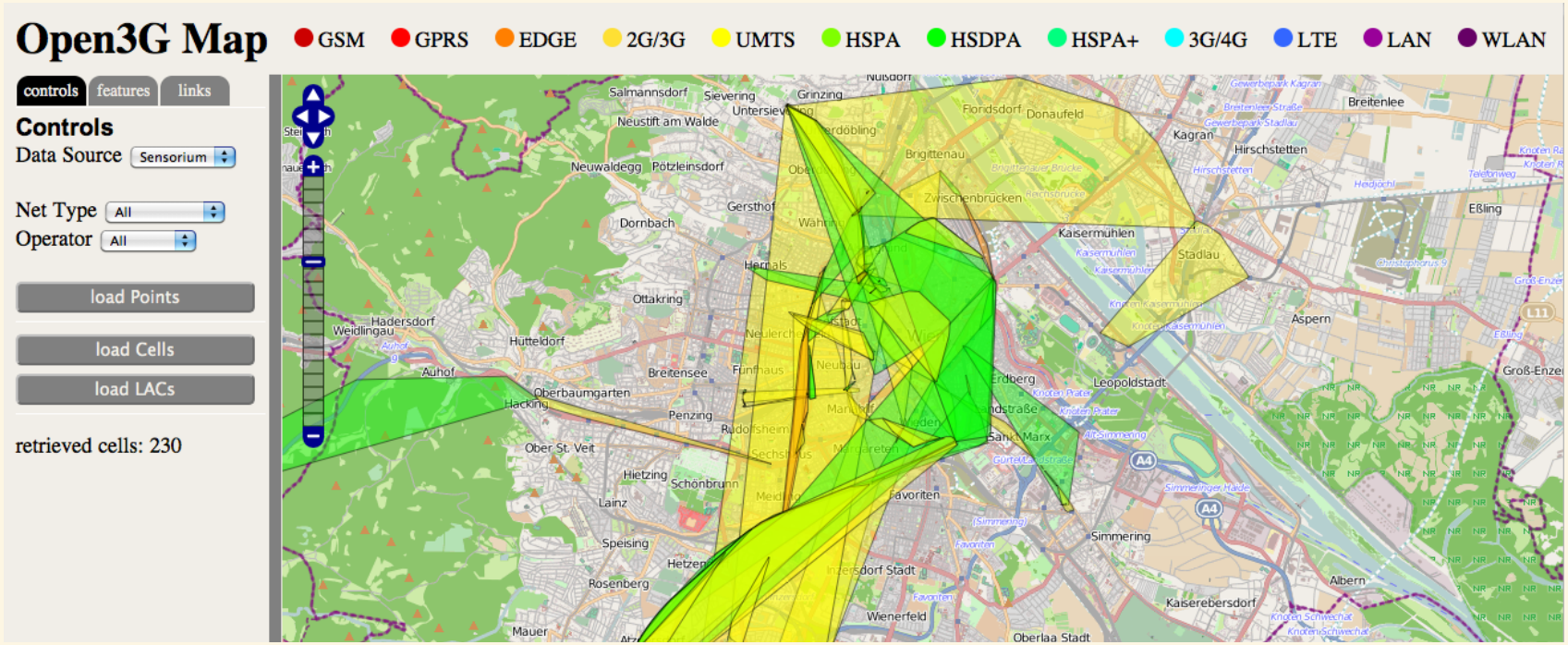
Wir wollen Daten friedlich und selbstbestimmt nutzen!

Sammeln für einen freien Zweck

- Nach manchen Schätzen kann man ewig graben, z.B. nach einem Mobilfunk-Netzplan.
- Die jeweiligen Betreiber haben ihn, exklusive Kunden vielleicht; „Konkurrenten" nicht; die großen Webfirmen sammeln selbst oder kaufen zu
- Egal ob Operator oder Webdienst — Definitiv nicht *wir*, definitiv nicht *frei*
- Das muss nicht so sein! Notwendige Daten sind wie gezeigt am Endgerät auslesbar, also speichern wir sie doch!
- Hinsichtlich Privatsphäre natürlich eine Herausforderung (siehe *Feinergranularer Ansatz* zuvor)

Und wir machen das natürlich auch!

Sensorium — Eine weitere App, die Mobilfunkdaten und hier auch GPS-Koordinaten sammelt. Serverseitig bauen wir eine Netzabdeckungs- und Qualitätskarte daraus. Open Data, bitte herunterladen, damit spielen, dazu beitragen!



https://o3gm.cs.univie.ac.at

Datenfluss

- Am Endgerät
  - Sensorium installieren
  - Sensorwerte sammeln: 3G, WLAN, GPS
  - ... und hochladen (URL konfigurierbar $\rightarrow$ für eigene Projekte verwenden!)
- Serverseitig
  - Speichern
  - Berücksichtigung anderer Datenquellen wie RTR Netztest [RTR]
  - Berechnen konvexer Hüllen aus Punkten mit derselben Cell ID
- Im Browser
  - Darstellung der Punkte, Polygone, ... mit JavaScript

Das war's schon wieder!

Vielen Dank für die Aufmerksamkeit!

Wir freuen uns auf Fragen.

Links zum Weiterspielen

- Sensorium zum Auslesen und Hochladen von Sensorwerten:
  `https://github.com/fmetzger/android-sensorium`
  `https://f-droid.org/repository/browse/?fdid=at.univie.sensorium`
  `https://play.google.com/store/apps/details?id=at.univie.sensorium`
- Datenschnorchler zum Auslesen von SMS, Anruflisten, Telefonbuch:
  `https://github.com/fmetzger/Datenschnorchler`
- Open3GMap, die offene, partizipative Netzabdeckungskarte:
  `https://o3gm.cs.univie.ac.at/`
  Backend-Sourcen:
  `https://github.com/lukpueh/Open3GMap`

Verwendete und weitere Quellen (1)

- [FMK] Jahres-Pressekonferenz des Forums Mobilkommunikation 2013 http://www.fmk.at/Medien/Pressekonferenzen/FMK-Jahrespressekonferenz-2013
- Third-Generation Partnership Project (Normungsgremium für LTE und andere Mobilfunkstandards, hier gibt es auch alle relevanten Standards gratis runterzuladen), online unter http://www.3gpp.org
- [EP020] „Schnittstelle gemäß TKG § 94 (4)", online unter http://portal.wko.at/wk/dok_detail_file.wk?angid=1&docid=1315929&conid=467524
- [WIKI3G] https://upload.wikimedia.org/wikipedia/commons/6/60/Ims_overview.png
- [CM] Cyanogenmod Privacy Guard: http://www.androidcentral.com/cyanogenmod-updating-privacy-guard-20-new-features-coming-cm102

Verwendete und weitere Quellen (2)

- [RTR] RTR Netztest – Netzmessungen, auch Open Data: `https://www.netztest.at/`
- [FB] Facebook-Visualisierung: `http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919`
- Zum Vergleich: Internet Map `http://internet-map.net/`, Visualizing Scientific Promiscuity `http://promiscuity.tentacleriot.eu/`
- Algorithmen: Secure Computational Geometry, Coin Flipping By The Telephone, Mental Poker / Mental Games