Karlsruher Institut für Technologie

Die vernetzte Gesellschaft:
Von Datenschutz bis Datenanalyse

Erik Buchmann, Clemens Heidinger, Martin Heine, Christian Hütter, Stephan Kessler

Lehrstuhl für Systeme der Informationsverwaltung

Wer sind wir?

- Lehrstuhlinhaber: Prof. Dr.-Ing. Klemens Böhm
- Forschungsschwerpunkte
  - Verteilte Informations- und Entscheidungssysteme
  - …
- Vorlesungen
  - Datenbanksysteme
  - Verteilte Datenhaltung
  - Data Warehousing und Mining
  - Datenschutz und Privatheit in vernetzten Informationssystemen
  - …
- Website: http://dbis.ipd.kit.edu/
  E-Mail: vorname.nachname@kit.edu

SEMINARTHEMEN

Übersicht der Seminarthemen

Proseminar

1. Soziale Netzwerke (Ch)
2. Zentralitätsmaße (Ch)
3. Datenschutz in sozialen Netzwerken (Cl)
4. Datenschutz in Folksonomien (Cl)
5. Indexstrukturen für verschlüsselte Datenbanken: Anonymität vs. Performanz (Cl)
6. k-Anonymität (St)
7. e-Energy Modellregionen (St)
8. Privatheit jenseits von k-Anonymität (M)
9. Anonymisierung mengenwertiger Daten (M)
10. Verbergen von Klassifikationsregeln (E)
11. Verbergen von Assoziationsregeln (E)

Betreuer

Erik Buchmann (E), Clemens Heidinger (Cl), Martin Heine (M),
Christian Hütter (Ch), Stephan Kessler (St)

Übersicht der Seminarthemen

Seminar

1. Kleine-Welt-Phänomen (Ch)
2. Netzwerkbildung (Ch)
3. Angreifer auf verschlüsselte Datenbanken (Cl)
4. Datenschutz im Smart Grid (E)
5. Location Privacy und e-Mobility (E)
6. Privatheit in statistischen Datenbanken (St)
7. Anonymisierung von Smart-Metering Daten (M)
8. Identifikation von Haushalten mittels Smart-Metering Daten (M)
9. Datenschutz und Geodaten (M)

Betreuer

Erik Buchmann (E), Clemens Heidinger (Cl), Martin Heine (M), Christian Hütter (Ch), Stephan Kessler (St)

Soziale Netzwerke (PS)

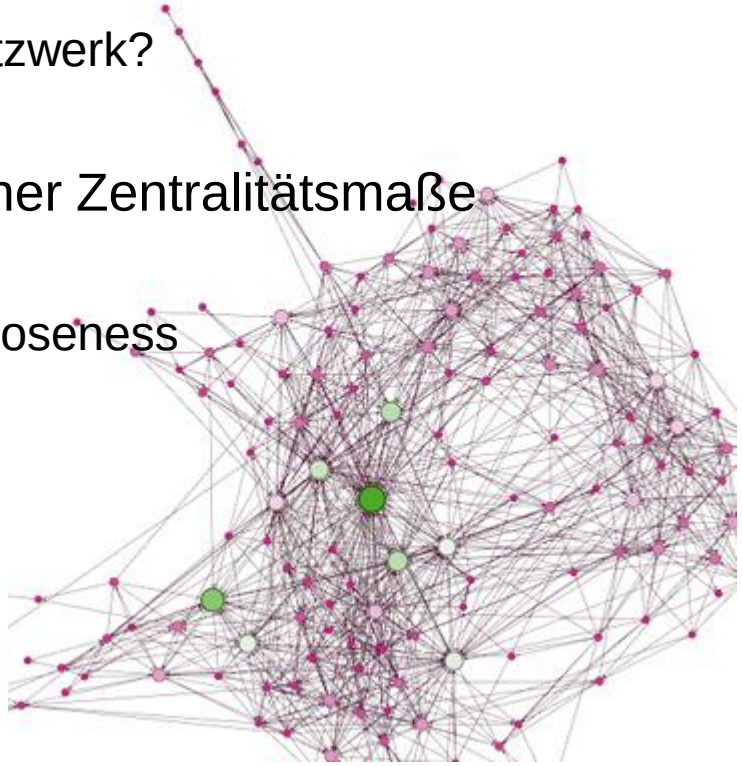
- Modell für Beziehungen zwischen sozialen Entitäten
  - Vernetzung von Freunden, Zusammenarbeit von Firmen, …
- Analyse realer Netzwerke
  - Nicht Plattformen, sondern die zu Grunde liegenden Netzwerke
- Soziale Netzwerke als Graph
  - Knoten repräsentieren Personen, Kanten die Beziehungen
  - Charakteristika: Knotengrad, Durchmesser, durchschnittliche Distanz, Clusterkoeffizient

Zentralitätsmaße (PS)

- Bestimmen die relative Wichtigkeit eines Knoten innerhalb eines Graphen

- Anwendung: Analyse von Sozialen Netzwerken
  - Wie wichtig ist eine Person in einem Netzwerk?

- Vergleich und Bewertung unterschiedlicher Zentralitätsmaße
  - Lokale Maße: Knotengrad
  - Distanzbasierte Maße: Betweenness, Closeness
  - Eigenvektorbasierte Maße: PageRank

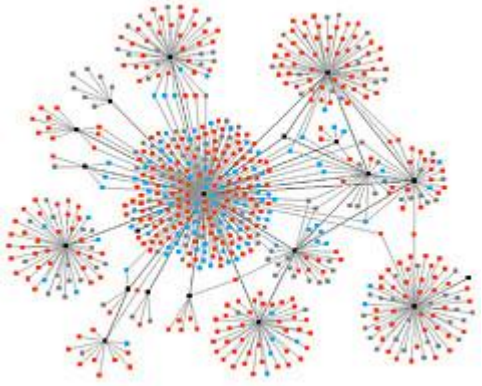


Kleine-Welt-Phänomen (S)

- Hypothese: Jede beliebige Person ist über eine kurze Kette an Bekanntschaften erreichbar

- Milgrams „Kleine-Welt-Experiment“
  - Untersuchung der durchschnittlichen Pfadlänge (Kette von Freundesfreunden)
  - Teilnehmer sollten eine Nachricht an eine fremde Person schicken, dürfen sie aber nur an Freunde weitergeben
  - ➔ “Six degrees of separation“: Jede beliebige Person ist höchstens sechs Schritte entfernt

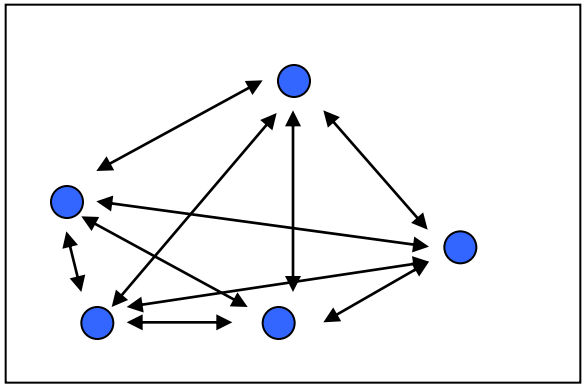
- Kleine-Welt-Phänomen im Internetzeitalter

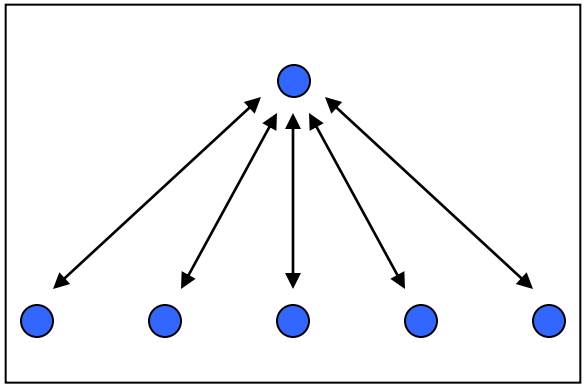


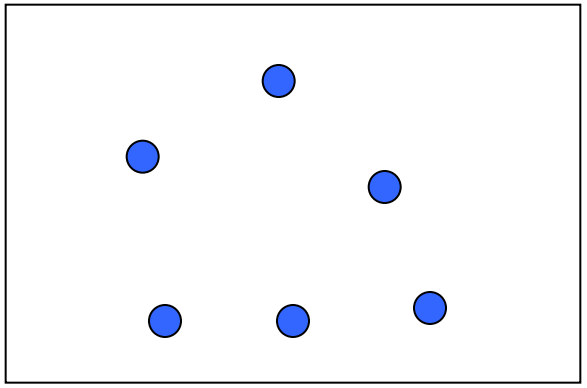
Netzwerkbildung (S)

- In sozialen Netzwerken kann man aussuchen, mit wem man interagiert
  - Wie sehen die entstehenden Netzwerke aus?
- Kleine-Welt-Netzwerke
  - Modell für Fremde, die durch gemeinsame Bekannte verbunden sind
  - Die meisten Knoten sind zwar nicht benachbart, aber durch wenige Schritte erreichbar

  ➔ Soziale Netzwerke als Kleine-Welt-Netzwerke
- Klassifikation anhand struktureller Merkmale







Datenschutz in sozialen Netzwerken (PS)

- Soziale Netzwerke wie Facebook, StudiVZ oder MySpace stark im Kommen
- Viele Daten werden preisgegeben: Bilder, Freundeskreis, Beziehungsstatus, etc.
- Über offene Schnittstellen wie OpenSocial Daten auch für Maschinen auslesbar
- Welche Daten für wen sichtbar? Risiken?
- Wie sehen Benutzer die Lage?
- Welche technischen Lösungen gibt es, Kontrolle über Daten in sozialen Netzen zu haben?

Datenschutz in Folksonomien (PS)

- Folksonomien (oder Social-Tagging-Systeme) erlauben es Objekte wie Webseiten oder Geokoordinaten mit kurzen Schlagworten (Tags) zu versehen
- Oft geben dabei Nutzer persönliche Informationen preis

- Wie muss eine rechtlich datenschutzkonforme Gestaltung einer Folksonomie aussehen?
  - Beispiel: Folksonomien können technische Grundlage für Data Mining oder Profilbildung sein. Inwieweit ist das zulässig?
- Liegen bei existierenden technischen Systemen juristischen Verstöße vor?
- Inwieweit lösen existierende Datenschutzmechanismen für Folksonomien die juristischen Probleme?

Indexstrukturen für verschlüsselte Datenbanken: Anonymität vs. Performanz (PS)

- Verschlüsselung der Daten beeinflusst die Performanz der Anfragebearbeitung negativ
- Indexstrukturen für verschlüsselte Datenbanken (sog. kryptographische Indices) als Abhilfe
- Tradeoff
  - Performance: Beschleunigung von Anfragen
  - Anonymität: Schutz gegen Angreifer
- Es gibt einige Verfahren, die jeweils einen bestimmten Tradeoff anbieten mit Performance- oder Anonymitätsgarantien

- In dieser Arbeit:
  - Untersuchung und Vergleich verschiedener Verfahren zur Erzeugung kryptographischer Indices aus der Literatur

Angreifer auf verschlüsselte Datenbanken (S)

- Verschlüsselte Datenbanken realisieren Datensicherheit und Datenschutz
  - Datenersteller verschlüsselt Daten bevor er sie zur Datenbank schickt
  - Provider der Datenbank muss nicht mehr vertrauenswürdig sein
    - Beispiel Cloud Computing: Unternehmen schicken Daten in die Cloud zur Kostenersparnis
- Anforderung: Unbefugte dürfen nicht an Klartextdaten gelangen
- Angreifer mit bestimmten Strategien und Wissen (= Angreifermodell) können aber Schutz (= Verschlüsselung) umgehen

- In dieser Arbeit:
  - Analyse verschiedener Angreifermodelle für verschiedene Anonymisierungsverfahren aus der Literatur
  - Untersuchung der Daten aus einem IPD-Experiment

k-Anonymität (PS)

- Im Bereich des Data Publishing werden Daten zur Analyse einer Partei bereitgestellt
- Im Datensatz sollen aber besonders schützenswerte Daten nicht einer Person zuordenbar sein
- Beispiel
  - Daten aus Krankenhäusern werden gesammelt dem Gesundheitsamt zur Verfügung gestellt
  - Gesundheitsamt muss den Verlauf von Krankheiten über Gebiete verfolgen können, beispielsweise zur Abwehr einer Epidemie
  - Aber: Gesundheitsamt muss und darf nicht wissen, dass eine bestimmte Person eine bestimmte Krankheit hat
- k-Anonymität ist ein Anonymitätsmaß für einen Datenbestand, das die Zuordnung „Person → <schützenswertes Attribut>“ schützt, aber noch sinnvolle Analysen auf Datenbestand erlaubt
- In dieser Arbeit soll k-Anonymität als Anonymitätsmaß und Heuristiken zur Herstellung eines k-anonymen Datenbestands untersucht werden

e-Energy Modellregionen (PS)

- Gerade mit der vermehrten Stromerzeugung aus erneuerbaren Energien, verändern sich die Anforderungen an das Stromnetz:
  - Erzeugung nicht mehr zentral (z.B. Kraft-Wärme Kopplung bei Kunden, Solarzellen)
  - Erzeugung auch “Wetterabhängig” (Wind, Sonne, etc.)
  - Steuerung und Verteilung kann somit auch nicht mehr zentral in wenigen Kraftwerken erfolgen
- Das “Smart Grid” soll helfen
  - “Internet der Energie”
  - Kommunikation von Verbrauchern und Erzeugern zur besseren (effizienteren) Energieverteilung
- In Deutschland existieren sechs Modellregionen des “e-Energy” Projekts
- Ziel dieser Arbeit:
  - Untersuchung, Zusammenstellung und Vorstellung unterschiedlicher Ansätze der Modellregionen

Anonymisierung von Smart-Metering Daten (S)

- Smart Meter sind intelligente Stromzähler, die den Stromverbrauch mit hoher Auflösung aufzeichnen.
  - Rückschlüsse auf Lebensgewohnheiten und -umstände möglich
  - Identifikation des Haushaltes anhand der Smart-Metering Daten möglich

- Verfahren zur Anonymisierung von Smart-Metering Daten verhindern die Zuordnung der Daten zum entsprechenden Haushalt.
  - Smart-Metering Daten sind Zeitreihen (spezielle Verfahren nötig)
  - Tradeoff zwischen Anonymität und Verwertbarkeit der Daten.

- Ziele der Arbeit:
  - Untersuchung und Vergleich verschiedener Verfahren zur Anonymisierung von Smart-Metering Daten in der Literatur

Privatheit in statistischen Datenbanken (S)

- Statistische Datenbanken beantworten Anfragen nach statistischen Kenngrößen, z.B.
  - Anzahl
  - Durchschnitt
  - Standard-Abweichung
- Anwendung z.B. im medizinischen Bereich: “Anzahl der Patienten über 40 mit Husten”
- Bedrohung der Privatheit einzelner “Teilnehmer” solcher Datenbanken z.B. durch
  - Geschickt formulierte Anfragen
  - Wiederholung von Anfragen
- Ansatz zur Privatheitsgarantie: “Differential Privacy”
- Ziel dieser Arbeit:
  - Vorstellung von möglichen Angriffen auf statistische Datenbanken
  - Erklärung des Konzepts der Differential Privacy inklusive der mathematischen Hintergründe

Verbergen von Klassifikationsregeln (PS)

- Objekte anhand ihrer Merkmale in vorgegebene Klassen einordnen
  - Klassifikation wichtig für viele automatische Entscheidungsprozesse, z.B. in Form von Entscheidungsbäumen
  - Klassifikator lernt anhand von Trainingsdaten mit vorgegebener Zuordnung, welche Merkmale wichtig sind

- Classification Rules können die Informationelle Selbstbestimmung gefährden
  - Regeln zur Zuordnung zu Klassen wie “unvorsichtiger Fahrer”, “ungesunder Esser”, “unsicherer Kreditnehmer”

- Seminarthema: aktuelle Verfahren zum Classification Rule Hiding
  - wie ist ein Datensatz zu verändern, damit bestimmte sensible Classification Rules nicht mehr gelernt werden können?

Verbergen von Assoziationsregeln (PS)

- Association Rule Mining ist beispielsweise für Warenkorbanalyse interessant
  - wenn Ereignis A und Ereignis B eintreten, tritt oft auch Ereignis C ein
  - wenn Objekte 1, 2 und 3 im Warenkorb liegen, interessiert sich der Kunde oft auch für die Objekte 4 und 5

- Association Rules können sensible Informationen verraten
  - wer Windeln kauft, kauft auch oft Bier

- Ziel dieser Arbeit:
  - Vergleich aktueller Ansätze zum Association Rule Hiding
    - Datensätze freigeben, aus denen sich bestimmte Regeln nicht mehr lernen lassen
    - Verfahren zum Entfernen solcher Regeln, ohne dabei neue (falsche) Regeln einzuführen oder bestehende (erlaubte) Regeln zu beeinträchtigen

Datenschutz im Smart Grid (S)

- Das Smart Grid ist die zukünftige Energieinfrastruktur
  - intelligente Stromzähler, Transparenz für den Verbraucher, flexible Tarife, erhöhte Energieeffizienz
  - derzeit in 6 Modellregionen in Deutschland getestet
- Problem: viele persönliche Details im Umlauf
  - intelligente Stromzähler messen mit sehr hoher Auflösung
  - liberalisierter Strommarkt, viele Parteien haben Zugriff auf diese Daten
- Seminarthema: wie wird heute mit diesen Datenschutzproblemen umgeganten?
  - Analyse der 6 Modellregionen
  - Review der Fachliteratur
  - kritisches Hinterfragen der vorgestellten Lösungen

Location Privacy und e-Mobility (S)

- Beim Aufladen von Elektrofahrzeugen werden Daten gesammelt, die deutlich über heutige Tankstellen hinausgehen
  - Aufbau von Bewegungsprofilen möglich
  - Lebensgewohnheiten, Reiseziele, Arbeitsverhältnisse werden offenbar

- Fragestellung dieses Seminarthemas: Sind aktuelle Verfahren im Bereich Location Privacy für Elektrofahrzeuge geeignet?
  - Welche Verfahren gibt es?
  - Welche Anforderungen ergeben sich durch die Elektromobilität
    - Abrechnung
    - liberalisierter Markt mit zahlreichen Ladestationsbetreibern
  - Welchen dieser Anforderungen werden bestehende Verfahren gerecht?

Privatheit jenseits von k-Anonymität (PS)

- k-Anonymität
  - „Ununterscheibarkeit von k-1 anderen Tupeln“
  - Äquivalenzklassen aus k ununterscheidbaren Tupeln
- Probleme von k-Anonymität
  - Geringe Diversität bei sensiblen Attributen
  - Ungleiche Verteilung von sensiblen Attributen
- Anonymitätsmaße die diese Probleme berücksichtigen:
  - l-Diversity
  - t-Closeness
- In dieser Arbeit sollen l-Diversity und t-Closeness untersucht und verglichen werden.

Anonymität mengenwertiger Daten (PS)

- Klassische Anonymisierungsverfahren arbeiten mit Tupeln fester Größe
  - Felder des Tupels und Semantik sind bekannt.
- Mengenwertige Daten
  - Größe unbekannt und verschieden (Warenkörbe, etc.)
  - Identifikation durch Kenntnis von Teilmengen
    - Kunde hat u. a. Butter und Bier gekauft
- Spezielle Verfahren nötig
  - Erweiterung von k-Anonymität für mengenwertige Daten
- Ziel der Arbeit:
  - Untersuchung und Vergleich von Anonymisierungsverfahren für mengenwertige Daten

Identifikation von Haushalten mittels Smart-Metering Daten (S)

- Smart-Metering Daten sind Zeitreihen über den Stromverbrauch eines Haushalts.
  - 1-4 Messungen pro Stunde
  - Rückschlüsse über Lebensgewohnheiten
  - Identifikation elektrischer Geräte
- Identifikation von Haushalten
  - Profil aus den gewonnenen Informationen
  - Unterscheidung verschiedener Haushalte anhand des Profils
- Ziele der Arbeit:
  - Welche Rückschlüsse können aus Smart-Metering Daten gezogen werden?
  - Welche davon sind für die Identifikation von Haushalten nüzlich?

Datenschutz und Geodaten (S)

- Umfassendes Kartenmatieral, Satellitenbilder und Geodaten sind heutzutage öffentlich verfügbar

- Geodaten geben alleine oder mit anderen Quellen verknüpft viele Informationen über eine Person preis
  - Google Maps liefert bei einer Suche nach einer Person häufig Arbeitsplatz
  - US-Zeitschrift schickte an 40 000 Abonnenten eine Ausgabe mit einem Satellitenbild von deren Haus als personalisiertes Titelblatt

- Welche Geodaten stehen zur Verfügung?
  Wo und wie werden Geodaten verarbeitet?
  Gibt es technische Systeme zum Schutz der Privatheit?

THEMENVERGABE

Themenvergabe Proseminar

Thema	Bearbeiter
Soziale Netzwerke (Ch)	Anni
Zentralitätsmaße (Ch)	Florian
Datenschutz in sozialen Netzwerken (Cl)	Kim
Datenschutz in Folksonomien (Cl)	Sebastian
Indexstrukturen für verschlüsselte Datenbanken (Cl)	Daniel
k-Anonymität (St)	May
e-Energy Modellregionen (St)	Christian G.
Privatheit jenseits von k-Anonymität (M)	Rene
Anonymisierung mengenwertiger Daten (M)	---
Verbergen von Klassifikationsregeln (E)	Moritz
Verbergen von Assoziationsregeln (E)	Matthias

Themenvergabe Seminar

Thema	Bearbeiter
Kleine-Welt-Phänomen (Ch)	Bijan
Netzwerkbildung (Ch)	---
Angreifer auf verschlüsselte Datenbanken (Cl)	Peter
Datenschutz im Smart Grid (E)	Nguyen
Location Privacy und e-Mobility (E)	---
Anonymisierung von Smart-Metering Daten (M)	Michael
Privatheit in statistischen Datenbanken (St)	Xheni
Identifikation von Haushalten mittels Smart-Metering Daten (M)	---
Datenschutz und Geodaten (M)	Thomas

ORGANISATORISCHES

Anforderungen für den Schein

- Erstellung einer
  - Gliederung und Literaturübersicht
  - Präsentation: 15 Minuten (Proseminar) bzw. 20 Minuten (Seminar)
  - Ausarbeitung: 8 Seiten (Proseminar) bzw. 12 Seiten (Seminar)

- Teilnahme an allen Vortragsterminen
  - Das Seminar ist eine Prüfungsleistung
  - Gleichberechtigt mit anderen Prüfungsleistungen
  - Unentschuldigtes Fehlen bedeutet Ausschluss aus dem Seminar

Termine

- Abgabe Gliederung und Literaturübersicht: 13.5.2011

- Blocktermine mit 4 Vorträgen pro Termin
  - werden noch bekanntgegeben
  - zwischen 06.06.2011 und 24.06.2011

- Abgabe Folien: 2 Wochen vor Vortragstermin
- Probevortrag mit Betreuer: max. 1 Woche vor Vortrag

- Erste Version der Ausarbeitung: 24.06.2011
- Abgabe Ausarbeitung: 08.07.2011

- Abgabefristen müssen eingehalten werden!

TECHNISCHES & LITERATURRECHERCHE

Technisches

- Ausarbeitung in Word / Writer oder LaTeX
  - LNCS Stylesheet
  - Vorlage wird auf Web-Seite zur Verfügung gestellt
- Präsentation mit Powerpoint, Impress oder PDF
- Abgabe der Materialien per E-Mail als PDF

Literaturrecherche

- Portal.ACM.org

- IEEE Explore

- citeseer.ist.psu.edu

- Hinweis: Zugang frei, wenn Ihr wie folgt vorgeht:
  - http://www.ubka.uni-karlsruhe.de/
  - „Digitale Biliothek“ anklicken
  - „Elektronische Zeitschriften“ auswählen

Literaturrecherche (Suchen)

Suche | Erweiterte Scholar-Suche | Scholar-Einstellungen

[BUCH] Biostatistical Analysis - Gruppe von 3 »
JH Zar - 2007 - Prentice-Hall, Inc. Upper Saddle River, NJ, USA
... The **ACM Portal** is published by the Association for Computing Machinery. Copyright © 2007 **ACM**, Inc. Terms of Usage Privacy Policy ...
Zitiert durch: 25214 - Ähnliche Artikel - Websuche - Bibliothekssuche

VIEL ERFOLG BEIM LITERATURSTUDIUM