PRIPARE

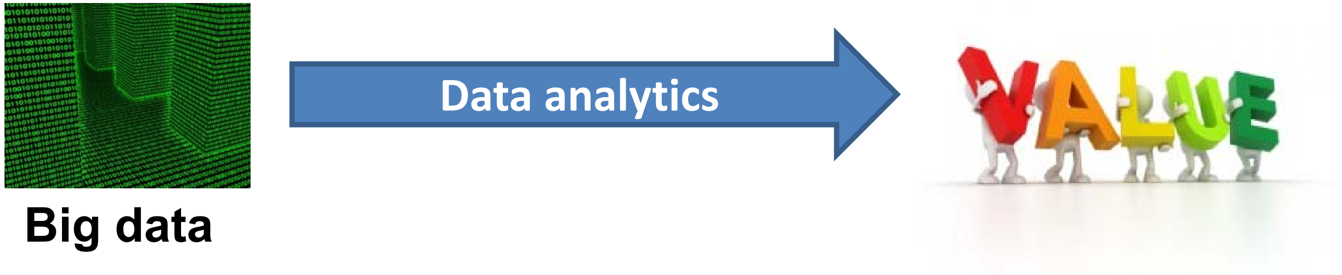
PReparing **I**ndustry to **P**rivacy-by-design by supporting its **A**pplication in **RE**search

Big Data and Privacy is it possible?

25th February 2015

Carmela Troncoso
ctroncoso@gradiant.org

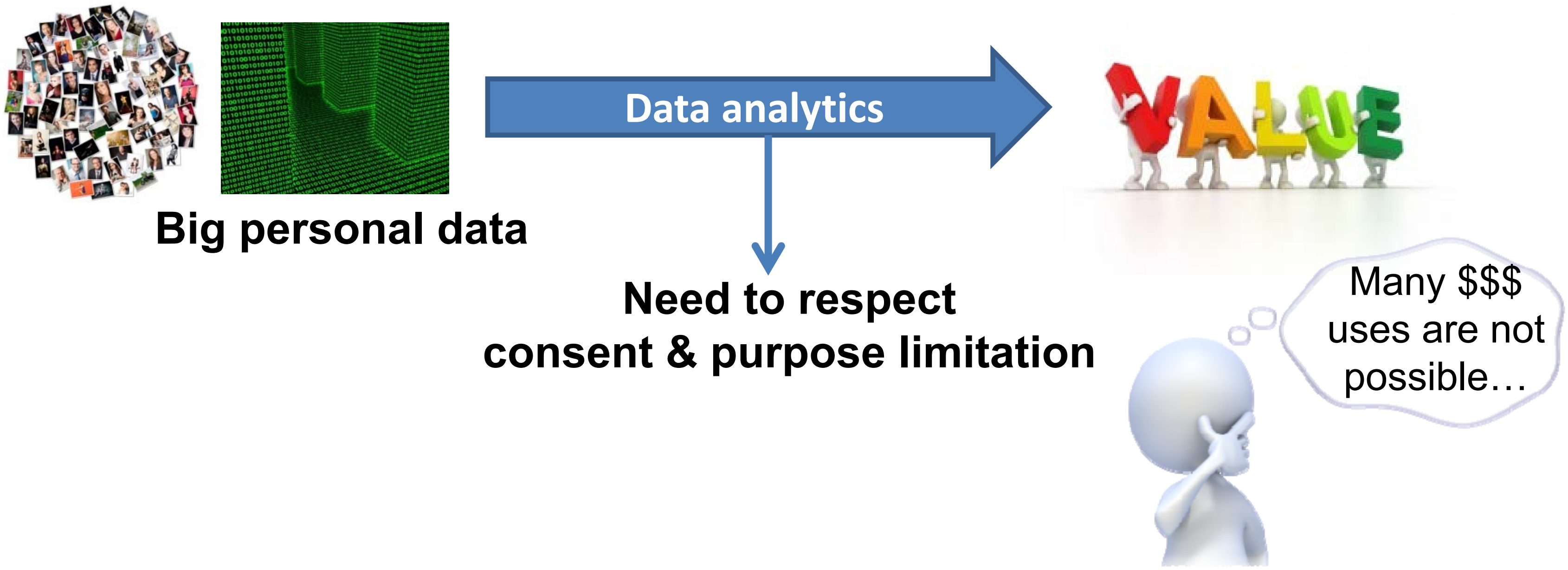
Big data and privacy



Big data

Data analytics

Big data and privacy



Data analytics

Big personal data

Need to respect
consent & purpose limitation

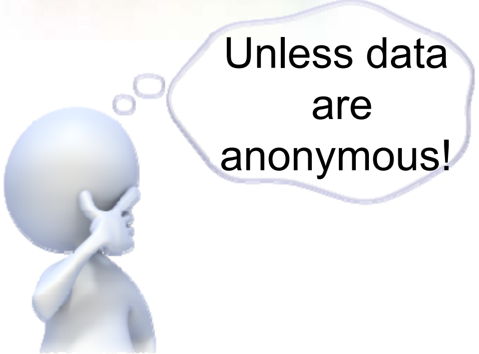
Many $$$
uses are not
possible…

Big data and privacy

Big personal data

Data analytics

Need to respect
consent & purpose limitation



Unless data
are
anonymous!

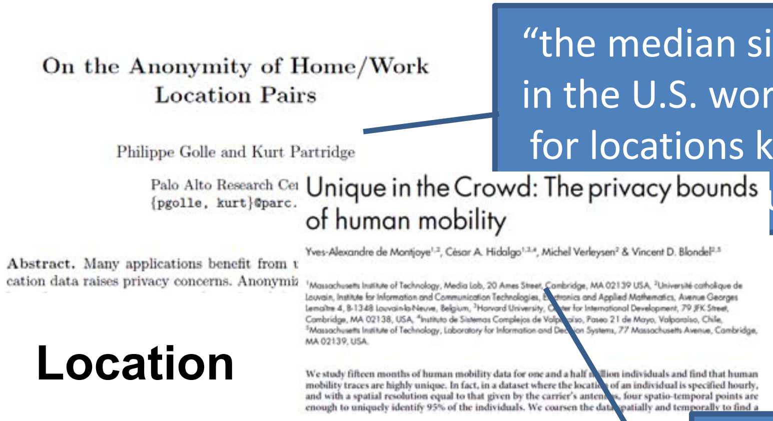
Art. 29 WP's opinion on anonymization techniques

3 criteria to decide a dataset is non-anonymous (pseudonymous):

- is it still possible to single out an individual,
- is it still possible to link two records within a dataset (or between two datasets)
- can information be inferred concerning an individual?

Is this compatible with Big Data?

Singling out - metadata tends to be unique



On the Anonymity of Home/Work Location Pairs

Philippe Golle and Kurt Partridge

Palo Alto Research Cei
{pgolle, kurt}@parc.

Abstract. Many applications benefit from u
cation data raises privacy concerns. Anonymiz

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

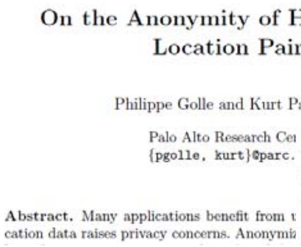
We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coursen the data spatially and temporally to find a

“the median size of the individual's anonymity set in the U.S. working population is **1, 21** and **34,980**, for locations known at the granularity of a census us track and county respectively”

Location

“if the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier’s antennas, **four spatio-temporal points** are enough to uniquely identify 95% of the individuals.” [15 montsh, 1.5M people]”

Singling out - metadata tends to be unique



On the Anonymity of Home/Work Location Pairs

Philippe Golle and Kurt Partridge

Palo Alto Research Cei
{pgolle, kurt}@parc.

Abstract. Many applications benefit from u
cation data raises privacy concerns. Anonymiz

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

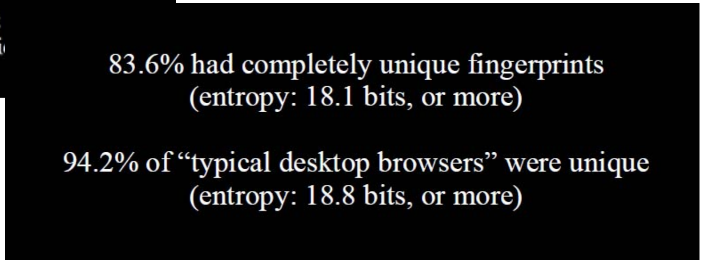
We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a

Location

How Unique is Your Browser?
a report on the Panopticlick experiment

Peter Eckersley
Senior Staff Technologist
Electronic Frontier Foundati
pde@eff.org

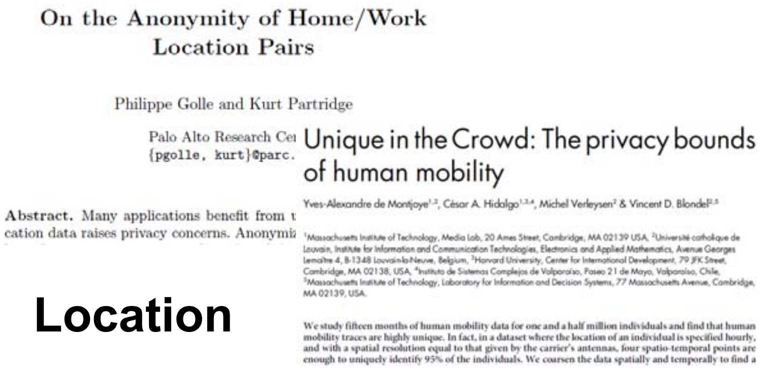
Web browser



83.6% had completely unique fingerprints
(entropy: 18.1 bits, or more)

94.2% of "typical desktop browsers" were unique
(entropy: 18.8 bits, or more)

Singling out - metadata tends to be unique



On the Anonymity of Home/Work Location Pairs

Philippe Golle and Kurt Partridge

Palo Alto Research Cer
{pgolle, kurt}@parc.

Abstract. Many applications benefit from u
cation data raises privacy concerns. Anonymiz

Unique in the Crowd: The privacy bounds of human mobility

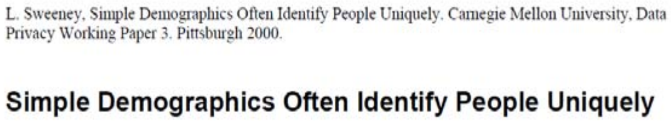
Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a

Location

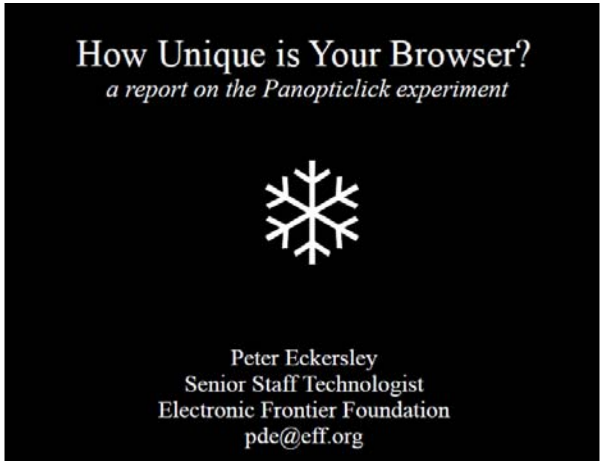
Demographics



L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

Simple Demographics Often Identify People Uniquely

"It was found that **87 % (216 million of 248 million) of the population** in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}"



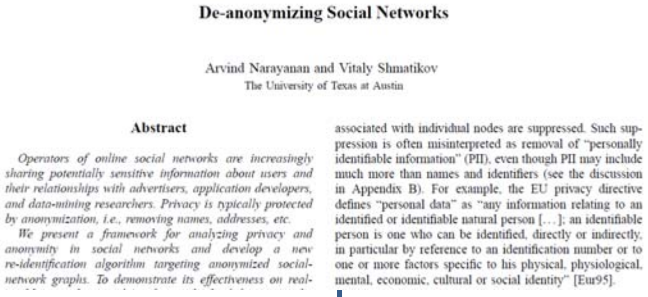
How Unique is Your Browser?

a report on the Panopticlick experiment

Peter Eckersley
Senior Staff Technologist
Electronic Frontier Foundation
pde@eff.org

Web browser

Link records relating to an individual



De-anonymizing Social Networks

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

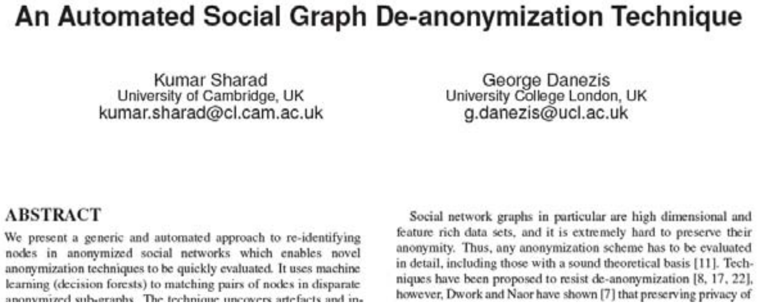
Operators of online social networks are increasingly sharing potentially sensitive information about users and their relationships with advertisers, application developers, and data-mining researchers. Privacy is typically protected by anonymization, i.e., removing names, addresses, etc.

We present a framework for analyzing privacy and anonymity in social networks and develop a new re-identification algorithm targeting anonymized social-network graphs. To demonstrate its effectiveness on real-

associated with individual nodes are suppressed. Such suppression is often misinterpreted as removal of "personally identifiable information" (PII), even though PII may include much more than names and identifiers (see the discussion in Appendix B). For example, the EU privacy directive defines "personal data" as "any information relating to an identified or identifiable natural person [...]; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity" [Eur95].

take two graphs representing social networks and map the nodes to each other based on the *graph structure alone*—no usernames, no nothing

Netflix Prize, Kaggle contest



An Automated Social Graph De-anonymization Technique

Kumar Sharad
University of Cambridge, UK
kumar.sharad@cl.cam.ac.uk

George Danezis
University College London, UK
g.danezis@ucl.ac.uk

ABSTRACT

We present a generic and automated approach to re-identifying nodes in anonymized social networks which enables novel anonymization techniques to be quickly evaluated. It uses machine learning (decision forests) to matching pairs of nodes in disparate anonymized sub-graphs. The technique uncovers artefacts and in-

Social network graphs in particular are high dimensional and feature rich data sets, and it is extremely hard to preserve their anonymity. Thus, any anonymization scheme has to be evaluated in detail, including those with a sound theoretical basis [11]. Techniques have been proposed to resist de-anonymization [8, 17, 22], however, Dwork and Naor have shown [7] that preserving privacy of

Technique to automate graph de-anonymization based on machine learning. Does not need to know the algorithm!

Inferring information about an individual

OH WAIT! What was big data about...?

Are there other avenues?

- The Big Promise: Processing in the Encrypted Domain (aka Homomorphic Encryption)
  - Advanced state of the art for particular problems
    - Privacy-preserving computation of statistics
    - Privacy-preserving billing
    - Privacy-preserving comparison
    - e.g., sharing cyberincidents data (INCIBE keynote)
      - Preserve individuals privacy and/or corporate secrecy
  - Still far away from efficient general purpose computations

Conclusions - Big data and privacy

- Is ok if no personal data involved in the analysis
  - Plenty of cases with high value!
- If there is personal data...
  - Anonymization in big data is difficult
    - Need for case-by-case evaluation of information leakage
    - Working towards an Open Source library
  - Processing in the encrypted domain
    - Not all is possible, but some things are! (come and talk to me)

PRIPARE

www.pripareproject.eu

Methodologies and research agenda

witdom

www.witdom.eu

Privacy in Cloud environments

Gradiant

www.gradiant.org

Privacy evaluation and privacy-preserving computations