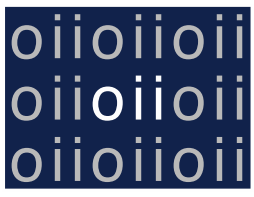
Ethical Privacy Guidelines for Mobile Connectivity Measurements

Edited by Bendert Zevenbergen, Oxford Internet Institute, University of Oxford

Contributors:
Ian Brown, Oxford Internet Institute, University of Oxford
Joss Wright , Oxford Internet Institute, University of Oxford
David Erdos, Faculty of Law, University of Cambridge

November 2013



Contact OII

Oxford Internet Institute
University of Oxford
1 St Giles
Oxford OX1 3JS
United Kingdom

Telephone: +44 (0) 1865 287210
Fax: +44 (0) 1865 287211
Email: bendert.zevenbergen@oii.ox.ac.uk
Web: http://www.oii.ox.ac.uk/research/projects/?id=107

Please cite the source of text and data excerpts as: Zevenbergen, B., Brown, I., Wright, J., Erdos, D.O. (2013) Ethical Privacy Guidelines for Mobile Connectivity Measurements. Oxford Internet Institute, University of Oxford.

Special acknowledgement for contributions to discussions that led to the creation of these guidelines:

- Ulrich Atz, Open Data Institute
- Dave Choffnes, Northeastern University
- Richard Clayton, University of Cambridge
- George Danezis, University College London
- Tim Davies, University of Southampton
- Erin E. Kenneally, University of California San Diego
- Luciano Floridi, Oxford Internet Institute
- Phillipa Gill, Citizen Lab
- Deborah Guterman, ASL19.org
- Dominic Hamon, Google Research
- Tristan Henderson, University of St Andrews
- Ethan Katz-Bassett, University of Southern California
- Douwe Korff, London Metropolitan University
- Christian Kreibich, UC Berkeley
- Masashi Crete-Nishihata, Citizen Lab
- Meredith Whittaker, Google Research
- Kieron O'Hara, University of Southampton
- Adam Senft, Citizen Lab
- Kumar Sharad, University of Cambridge
- Sam Smith, Privacy International
- Linnet Taylor, Oxford Internet Institute
- Rob Van Eijk, Dutch Data Protection Authority/Leiden University
- Stef van Grieken, Open State Foundation
- Christopher Wilson, engine room
- Mike Wittie, Montana State University
- Aristea Zafeiropoulou, University of Southampton
- Frederik Zuiderveen Borgesius, University of Amsterdam

These guidelines have been made possible by a kind donation from Google Inc.

Table of Contents

A) Introduction

The Internet is a highly complex and pervasive information environment. Everyday activities increasingly have an online component, from talking to friends and family, watching TV programs, dating, to interacting with government. To understand and make sense of the complex Internet architecture underpinning these activities, network researchers need to collect and share datasets regarding the measurements of the network, from detailed traces on an individual basis to aggregated data on a regional level. Data on individuals' Internet behaviour will frequently contain sensitive information about the data subjects' lives. On the other hand, data that only reveal an Internet users' connection to a given point on the network is not necessarily privacy invasive.

There are a number of existing large research datasets gathered from fixed line broadband Internet connections, such as those hosted by Crawdad, PREDICT, the Cooperative Association for Internet Data Analysis (CAIDA) and the Measurement Lab. Less data is available regarding mobile Internet connections, which are increasingly important as an access mechanism given the huge numbers of deployed smart phones and tablets. Measuring the mobile Internet will potentially expose information about individuals, such as location throughout the day and contact details stored on the phone, as well as the metadata (or "communications data" in the UK) of all their communications. It can be very difficult to predict how or whether records in supposedly "anonymised" datasets will be re-identified.

Sensitive data in the wrong hands – of identity thieves, malevolent (possibly authoritarian) governments, abusive spouses, aggressive marketers, etc. – can lead to serious financial, reputational, physical or other harms. In many countries, not just within the European Union, privacy is a constitutionally protected individual right seen as vital to democracy. It is therefore important that network researchers understand what privacy is, why it needs to be protected (see section D.1), and seriously consider ethical protections while collecting, processing and disseminating data from Internet measurements.

Reproducible science, secondary data use and third-party innovative re-use of research data all benefit from access to disaggregated raw data. It may be possible to find a compromise in which some level of aggregation and pre-processing to de-identify the data takes place before a dataset is released. This involves a balancing act, maintaining the maximum potential for low friction data re-use and checking of findings, whilst ensuring privacy. Above all, researchers must actively consider how to preserve the privacy of data subjects when collecting data.

If effective de-identification leads to an unacceptable level of utility loss of the data, secure data archives can help ensure data is available to trusted third parties, even if it is not made available as open data. Researchers should consider further interactive information management mechanisms to maximise the utility and manage the relationship with data subjects.

The guidelines in this document have been developed to protect the interests of both researchers and data subjects. They are based on existing examples of best practice, wide consultation with networking and privacy researchers, and a one-day workshop. Their use will contribute to public trust in networking research, which is essential for future data collection. They will also help researchers demonstrate they have taken reasonable steps to ensure data subjects' privacy.

Goal of guidelines

The aim of these guidelines is to help network researchers navigate the challenges of preserving the privacy of data subjects, publishing and disseminating datasets, while adhering to and advancing good scientific practice. The Association for Computing Machinery (ACM) highlights two relevant principles in its code of ethics:

- 1.2) Avoid harms to others, and
- 1.7) Respect the privacy of others.

These guidelines will help researchers assess the potential privacy risks and associated harms of a research project, and how these can be managed. They identify some of the common privacy problems that mobile networking researchers face, and offer ethical recommendations and considerations that need to be taken into account when designing a research project.

It is difficult to quantify privacy risks and subsequent utility trade-offs precisely, as they depend on many factors, such as the political context or the capabilities of a possible adversary. The assessment of risks and choice of appropriate de-identification technique therefore need to be based on careful deliberations, primarily between network researchers, legal experts, ethics review boards, academic journals, and conference organizers. These guidelines are designed to provide the basis for such a constructive dialogue and to guide the appropriate management of the risks involved. Further, these guidelines can be used by researchers for self-assessment or reflections on research design with colleagues.

How to use these guidelines

Section A of this document describes the scope of these guidelines. Section B describes some key ethical considerations on which the guidelines are based. Section C contains the guidelines, offers short introductions to relevant considerations for network researchers, and poses assistive questions on important topics.

The text refers to underlying explanatory section D, which explains key concepts and considerations in more detail. This section contains concrete examples and demonstrates how to think about specific privacy related issues in network research.

The assistive questions in section C should be considered during the research design phase in an iterative process, to reduce risk to a minimum, compensating newly identified higher risks in

some areas (e.g. open data disclosure) with lower risk parameters in other areas (e.g. identifiability).

Key definitions

A dataset is a collection of related sets of information that is composed of separate elements. The elements of datasets discussed in this framework are divided into three categories: (i) identifiers, (ii) key-attributes, and (iii) secondary attributes.

i) Identifiers

Identifiers are attributes that can individually distinguish the data subject more or less directly. Typical identifiers include: name, address, social security numbers, mobile phone number, IMEI number.

ii) Key-attributes

Key-attributes can be used to identity a data subject using auxiliary sources of information, by linking to databases that contain identifying information. They are indirect identifiers of a data subject, which make an individual more distinctive in a population. Typical key-attributes include: age, race, gender, date of birth and place of residence.

iii) Secondary attributes

Secondary attributes cannot individually identify a data subject directly and may require significant amounts of auxiliary data to be useful for re-identification purposes. A data subject may then be identified individually through more sophisticated methods such as *fingerprinting*, rather than mere linking of databases. Examples include the settings in an application, the battery level measured over time, or location patterns.

Auxiliary datasets

Many databases are available, publicly or in private hands, that contain identifiers, key-attributes and secondary attributes about individuals. These can be linked to de-identified databases to re-identify data subjects.

Adversary/attacker

An adversary is the entity who is interested in re-identification of a dataset, for example by attacking the de-identification technique used. In existing literature, the terms ‘adversary’ and ‘attacker’ are used synonymously.

Delimitations of guidelines

Open data

Academic publications and network measurement platforms often require researchers to make their datasets publicly available, sometimes in an open data format. Public disclosure in such formats is problematic for datasets that contain identifiers, key-attributes and secondary attributes, as these enable re-identification of data subjects by linking the records with auxiliary datasets. In general, only datasets that exclude any identifiable information (see section D.2) are fit to be published as open data. Section D.6a explains in more detail why this is difficult to

achieve. However, some identifiable information may be harmless given the context or the type of information contained in the dataset. Therefore, this should not be considered a blanket prohibition, but the researcher should strive to publish largely de-identified information where possible.

The starting point of these guidelines is open data publishing, but managed access systems are recommended for many contexts in section D.6b. Managed access enables the utility of the datasets to be calibrated for each individual dissemination, which can increase the usefulness of research data overall.

Active measurements

Network research platforms generally only allow research applications that actively measure the network. This involves a user-initiated measurement that generates data transfers, and measures how the network responds to this data.

Passive measurements – whereby a measurement tool monitors a user's behaviour and records the device's interaction with the network – are commonly not allowed on measurement platforms that mandate open data release. These guidelines will therefore not cover passive measurements.

Focus of guidelines

These guidelines focus on ethical considerations relating to a data subject's privacy. Only active measurements – initiated and consented to by data subjects – are covered. They do not concern ethical questions relating to research activities such as infiltrating botnets or observing criminal behaviour

For countries with comprehensive privacy laws, these guidelines assume that informed consent is the legal basis on which data are collected about data subjects. There are situations when research projects can also be pursued without consent from data subjects, but they fall outside the scope of these guidelines. However, most of them will still be relevant for research conducted without informed consent. Researchers should discuss with their ethical boards when and how such an approach is feasible.

Although an ethical approach is the starting point of these guidelines, privacy laws around the world already formalize and enforce some of these principles. These guidelines therefore take much inspiration from various legal frameworks, and apply it to network research.

Academic researcher focus

The focus of these guidelines is on academic researchers, who in most university systems must gain ethical approval before a research project involving human participants can commence. When the guidelines are used by private sector researchers, independent or within a company, these guidelines should be discussed with an equivalent authority, legal expert or the relevant beneficiary organization, which intends to publish the research results (such as, for example, a conference, an Internet measurement platform or the legal department of the organization).

Methodology

These guidelines are based on an extensive literature review, similar existing guidelines, discussions with computer scientists/network researchers and lawyers, and a workshop organized at the Oxford Internet Institute on 18 June 2013. A list of key literature, which offers further and in-depth reading about the issues summarized in these guidelines, can be found in section D.12.

The UK Anonymisation Network (UKAN) is a useful resource for best practices and practical advice in anonymisation of data sets that will be shared: http://www.ukanon.net/

The European Article 29 Working Party will soon publish an opinion on the use of anonymisation and de-identification techniques, which may warrant an update of these guidelines.

B) Outline of the ethical considerations

Internet measurement can impact on user privacy, especially if specific data on some aspect of user behaviour is collected. Therefore, data subjects' trust is an important foundation for the legitimacy of the research sector. Trust in network research will diminish if data subjects suffer harm as a result of the collection and dissemination of their data. It is therefore imperative – and a legal requirement in many countries - that researchers take privacy and the rights of data subjects seriously.

The utility and privacy of data are generally directly and inversely related. For many datasets, it has proven difficult – if not impossible – to increase data subjects' privacy without concurrently decreasing the overall utility of the dataset. Small privacy gains are generally achieved by far-reaching decreases in data utility. A small increase in data utility often requires much more personal information to be revealed.

Data subjects can be identified more easily when linkable information is revealed in a new dataset, because the attributes might be used for re-identification by linking the new dataset to auxiliary datasets. It is difficult to assess exactly how much auxiliary data is available in public or private sources. Some suggest it is good practice to adopt a conservative approach to auxiliary data, whereby perfect auxiliary information is assumed to exist that can be used to re-identify data subjects in new databases with relative ease. Perfect auxiliary data does not exist, but the researcher should take a cautious approach when assessing the risks of linkability.

A strong movement to open up research data currently exists, for good reasons (see section D.6a). However, the assessment of privacy risks becomes even more challenging with a free and open online dissemination of a research dataset. Once a dataset is disclosed online, the researcher has lost control over how these data will be used. Although the uses for certain datasets can be predicted to some extent with regards to the current state of technology and business or government interests, the context of uses may change significantly in future.

Therefore, privacy considerations require a conservative approach to data dissemination on the Internet. These guidelines do not use a zero-risk standard, whereby data utility would be minimal. Some reasonable risks are permissible, depending on the context. Due to the seriousness of a privacy breach and the possible sensitivity of the collected data, we advise researchers make the reasonableness assessment a cautious one.

Current privacy and data protection laws offer exceptions for datasets that have been "anonymised". However, the real possibility of re-identification of so-called "anonymised" datasets is not adequately reflected in most privacy laws. These guidelines will help researchers navigate the new challenges to privacy posed by re-identification technology, while also complying with existing laws. We use the term "de-identification" rather than "anonymisation", as it is technically more accurate.

C) Privacy-protecting Ethical Research Design

The potentially sensitive data collected by network researchers can cause harm to individuals if they are identified as the result of the disclosure of a dataset, as well as potential liability for the researcher or her institution and reputational harm for the sector. It is therefore an ethical obligation of researchers to design research carefully and to control the flow of sensitive data. Privacy-aware research does not merely control data disclosure, but manages risks during its collection and processing.

No single privacy statute or data protection law contains all the considerations set out in these guidelines. We have based them on national and international law, and existing research and ethical considerations. When assessing research design, researchers should actively consider at all steps: *What would re-identification mean for data subjects in this particular context?*

Research Design

These guidelines take the researcher through the process of designing a research project that manages privacy risks appropriately while maximizing data utility to the extent that is ethically acceptable. The assistive questions offer a series of tests, with further background information in section D. The aim is to help the researcher think about and discuss with colleagues the level of privacy risk of specific research design choices. It is difficult – if not impossible – to quantify the privacy and utility trade-offs accurately. Therefore, the questions rely on three parameters: higher risk, medium risk and lower risk.

The aim of the iterative process is to reduce privacy risk to a minimum, by taking into account the advice and criticism resulting from the discussions based on the assistive questions. The researcher should update the research design, compensating newly identified higher risks in some areas (e.g. open data disclosure) with lower risk parameters in other areas (e.g. identifiability).

Once the research design has been finalized and approved by the relevant ethical boards and/or legal experts, clear information needs to be provided to potential data subjects, which explains the research design in a transparent manner. The data subjects can then base their informed consent on this information (see section D.9a for an overview of the information that should be provided).

Privacy by design

A network research project design that protects data subjects' privacy and maximises utility requires a multi-dimensional consideration of how all the parts of the design operate together. The protection of personal information must be considered from the start and analysed at each step. Section D.3 gives an overview of some considerations about such a process of Privacy by Design. To assess how each part of the design affects the risk assessment of other parts of the research, the process must be iterative; the researcher must assess the effect of each change of the research design on the other parts.

Privacy Impact Assessment

A privacy impact assessment (PIA) is an essential exercise to assess to what extent privacy will be preserved when conducting research that might have a high privacy risk. The PIA forms an assessment of the privacy risks in a research project and helps the researcher to manage the risks. The PIA can also be used as evidence that the researcher has considered the privacy issues properly, should questions or doubts arise. These guidelines can be considered as an applied PIA, as elaborated in section D.4.

Assessing benefits

The aim, purpose, and intended methodologies of research need to be stated clearly before any further ethical judgments can be made. These will be used when assessing the proportionality of a de-identification technique and method of data dissemination:

- How will this research contribute to the state of the art in understanding network phenomena?
- Will the research results be directly relevant to and applicable in some specific government, business or academic processes?
- How will the research benefit society and specific stakeholders?
- Can the researcher formulate the research aim concretely and specify stakeholders who will directly benefit from the research?

Assessing privacy risks

Collection of data

Categorization of mobile data types

The researcher should consider which data categories are needed for the research analysis to achieve the stated research aim. An overview of the data types that will be collected informs all other risk assessments and dictates the appropriate data processing controls. Section D.10 gives an overview of the data types that can be collected from mobile phones and shows how to classify related privacy risks.

Purpose limitation and data minimisation

The amount and types of data collected must be relevant and not excessive for the research purpose. This is not only a legal obligation in many countries, but also minimises the risks of liability for researchers and simplifies the management of privacy issues. Section D.5 explains the necessary precautions in more detail:

- Is it necessary to conduct a new measurement, or do datasets containing the needed measurement already exist?
- Can the same results be achieved in a test setting, or must the data be collected in the field?

- If new measurements need to be conducted, are the identified data categories relevant and not excessive in relation to the research purposes (i.e. strictly necessary), as specified in section D.5 under 'data minimisation'?
- Where does the combination of collected data put the dataset on the identification continuum described in section D.2?
    - Does the dataset contain direct identifiers? (Level 1, Higher risk)
    - Is it possible to infer the identity of individuals through a combination of the key-attributes? (Level 2, Higher risk)
    - Is it feasible to take into consideration auxiliary data which can identify individuals when combined with the key-attributes or other collected data? (Level 3, Higher/medium risk)
    - Is the dataset void of identifiers, key and secondary attributes? (Level 4 or 5, Lower risk)
- If the raw data is disclosed, does the information reveal any sensitive data about the substance of the information the data subject interacted with?
    - Is it possible to collect less data, to reduce the sensitivity of the data?

Risk assessment

The privacy risk needs to be assessed in light of any likely *adversary* who could be motivated to use the new dataset to her advantage (see section D.7), and the broader context in which re-identified data could be used. For example, the researcher must consider what a dataset, if re-identified, tells the adversary about the data subject. Some information may be fairly benign, whereas other contexts could be sensitive when interpreted by a specific adversary.

The researcher must substantiate which types of adversaries are the most likely to want to re-identify the dataset. The overall risk level of the research project should be adjusted based on the expected capacity, motivation, skill, time and available auxiliary information the adversaries are likely to possess with regards to the re-identification of data subjects. In addition to this consideration, the researcher should give due deliberation to possible future adversaries, to the extent possible. These parameters are necessary to determine the suitable de-identification technique to be applied to the raw data. The researcher may reuse risk assessment profiles from previous, comparable research designs:

- What persons or organizations may likely be interested to re-identify the proposed dataset and for which reasons? See section D.7 for a general classification of adversaries and assess their motivation and subsequent level of risk.
- To what extent would such re-identification harm individual data subjects, or specific groups in the dataset? Could the type of information be considered a higher risk (e.g. financial and medical information, even if indirect), or rather a lower risk (only secondary identifiers)?
- Which known auxiliary information could the adversary use to re-identify data subjects, and would the available information increase the potential harm? How sensitive is the known auxiliary data that can be combined with the research dataset to re-identify data subjects? Is it reasonable to assume that more linkable auxiliary information exists?

- What capacity (in terms of time, skill, computing power, etc.) do any identified adversaries likely have to re-identify datasets?
- What activities would the dataset reveal, if re-identified?
- What are the roles, relationships and power structures of the stakeholders (data subject, likely adversary and other beneficiaries)? What is the political context? Does this change the sensitivity of the revealed activities?
- Are there any meaningful statutory privacy protections in the jurisdiction of the data subject that offer extra protection for the data subject?
- Does the benefit of the research outweigh the potential harms, given the context in which the data is collected and research results are likely to be used?

Type of dissemination

Before deciding how best to de-identify any collected data, the researcher must decide how the research data will be disseminated. Section D.6a explains that an open data format disclosure, with no obligations or restrictions attached, presents a higher risk. Therefore, the researcher will need to de-identify her dataset completely (Level 4 or 5, section D.2) before disseminating it in this manner and carry out rigorous re-identification testing (see next step on "de-identification of datasets").

Because an open data format disclosure means that datasets need to be de-identified as much as possible, thereby losing much utility, we suggest some other types of disclosure that the researcher or measurement platform may want to consider (see section D.6b). Generally, the following hierarchy of disclosure techniques can be identified:

1. Open Data - No restrictions on dissemination - Higher risk;
2. Restricted data sharing - Legally enforceable restrictions - Medium/higher risk;
3. Managed access - Lower risk;
4. Interactive methods - dissemination of statistical information about dataset - Lower risk.

- Will the research dataset be shared with specified individuals (lower risk), a wider research consortium (medium risk), or be released publicly in open data format (higher risk), possibly via a data repository or Internet measurement platform?
- Will the disclosed data be limited to fulfil certain specified tasks (e.g. developing anti-spam lists, lower risk)?
    - If the answer to the previous question is positive: Will the functionality be left to the receiving party to decide (higher risk), or will the researcher discuss the needed functionality with the recipient and tailor the data for this use (medium risk)?
- If the researcher chooses a data sharing approach, which legal restrictions will be included in the data-sharing agreement?
- How will the researcher enforce compliance by the receiving party?
- If an interactive method is chosen, does the researcher disseminate only general statistics about the data (lower risk), truncated data (medium risk) or is more detailed information including key-attributes and identifiers shared (higher risk)?

De-identification of datasets

When deciding on an appropriate de-identification technique, the resulting benefits and risks must be weighed. As such an assessment is difficult to achieve precisely, the researcher should discuss the choice of appropriate de-identification techniques with colleagues.

For example, when opting for an open data disclosure method (higher risk), a suitable de-identification method will need a 'higher' robustness level. All identifiers and key-attributes should be removed, or a method that aggregates these data to a lower risk level should be employed. The extent to which secondary-attributes can be used to identify data subjects, either via *fingerprinting* or combination with an extensive range of existing auxiliary datasets, should also be assessed and tested where possible. The chosen method must be able to obfuscate even the secondary-attributes to an extent at which the risk of successful re-identification by the potential adversary is low.

It may not be necessary to de-identify the dataset at all if the research data is disclosed by an interactive method (lower risk). Of course, this will only be feasible if the raw dataset will never be disseminated and is well secured against attacks.

Restricted data-sharing systems can have varying risk levels, as the context of disclosure will dictate the sensitivity of the data to a large extent. The level of robustness must be decided on a per-case basis. General rules can be set in accordance with colleagues, for example deciding on low robustness (high utility) when sharing datasets with a research consortium, as long as the dataset is accompanied by an enforceable non-disclosure agreement. A higher robustness level must be chosen if the researcher does not intend to control further dissemination of the dataset.

The researcher may apply multiple de-identification techniques and methods of dissemination to a single dataset (or a sample thereof), depending on a case by case basis on the assessed risk level of the recipient, amount of control exercised over the dataset and the sensitivity of the dataset.

- Is the chosen threshold of de-identification technique proportionate to the:
    - Sensitivity of the data?
    - Foreseen disclosure method?
    - Capacity of the identified adversary?
- Has the researcher consulted a re-identification expert to discuss whether the foreseen data collection categories can lead to re-identification of individuals in the dataset? (Lower/medium risk)
- Has the researcher successfully carried out experiments to test re-identifiability? (Lower risk)

To ensure a privacy-aware research design, the researcher should check whether high risks in the categories and amount of collected mobile data types, the initial risk assessment and the

type of foreseen dissemination, are counterbalanced by the robustness of the de-identification technique used.

Managing unforeseen risks

Systems and research design will never be as robust as intended. To mitigate unforeseen risks, the researcher must be prepared and manage the unknown in the best way possible. When a dataset is disclosed unexpectedly, it is part of the ethical process to alert data subjects, so they too can take precautions:

- Is the dataset stored securely?
- Does the researcher employ any encryption, for example on sensitive datasets?
- Is there a containment policy and what does it oblige the researcher to do?
- Will the researcher contact the data subjects and/or the relevant privacy regulator directly about a breach? To what extent does this depend on the seriousness of the disclosure or the sensitivity of the data?
- How will harmed data subjects or stakeholders be compensated?

Consent, Transparency and Informational Self-Determination

By this stage, the researcher has identified and assessed the benefits and risks of the research design as outlined above and made informed decisions about the collection, processing and dissemination of the foreseen dataset. For university researchers, an ethical approval of this research design is needed before the data collection can commence. Non-university researchers should discuss their research design with their funders or other relevant entities.

Clear information about the research must be communicated to the potential data subjects before any data collection can commence. Section D.9a outlines the main considerations for gaining informed consent from research participants. In practice, achieving this requirement is harder than one may think. However, user trust is essential for the benefit of sustainable network research, especially when any identifiers, key-attributes or secondary attributes are collected. Researchers must use the informed consent procedure to be transparent about the data they collect:

- Have data subjects in the dataset consented to their involvement with the research project?
- Has the researcher informed the data subject about possible and foreseen secondary uses?
- Does the data subject understand the potential risks and benefits of the chosen method of dissemination?
- Could a lay person understand to what extent the data will be de-identifiable?
- Could a lay person understand how long the data will be stored and disclosed?
- Can the data subject object to the processing?

- How will the researcher give the data subject insight into what data is collected, what secondary uses the data is used for, and share the research results with the data subject?
- If the purpose of the collected data changes, can the data subject be informed and can she retract her consent?

Once a data subject has given her informed consent, the researcher is in principle free to start the collection data in line with the notice given to the data subject.

D) Background

D.1 What is Privacy?

Privacy is a difficult term to define. Daniel Solove has developed an overview in his paper "A Taxonomy of Privacy", which builds on previous works such as William Prosser's "Privacy" and Alan F. Westin's book "Privacy and Freedom". Many types of privacy have been identified, but for the scope of these guidelines, we will only focus on informational privacy (also known as data privacy) and discuss the principle of informational self-determination.

Informational/data privacy and self-determination

Informational privacy concerns the individual's right to exercise control over the disclosure or processing of her personal information. In this sense, Westin defines privacy as "*the claim of individuals, groups, or institutions to determine for themselves when, how and to what extent information about them is communicated to others*". Personal information is generally defined as *any* information about or otherwise relating to an identified or identifiable individual. Whilst there has been a debate about whether truly innocuous data is covered by this definition, it should be stressed that the type of data used in network measurement research will, if identifiable, fall within this definition. Information will still be identifiable even if this is only possible when the data in question is matched to information stored in another (or even several other) auxiliary databases. See section D.2 on the continuum of identifiability of individuals in datasets.

Solove identifies four harmful activities with regards to privacy, all of which are directly relevant for informational privacy and the collection of mobile connectivity data:

1. Information **collection** (surveillance and interrogation);
2. Information **processing** (identification, aggregation, storing, second uses, exclusion);
3. Information **dissemination** (disclosure, breach of confidentiality, exposure, increased, accessibility, distortion, blackmail and appropriation);
4. Invasion (intrusion or interference into one's life).

Informational self-determination is a very important concept to mitigate these harms for several reasons, including (according to Westin):

1. Personal autonomy– the development of the personality and prevention of manipulation by others;
2. Emotional release – the ability to escape everyday tensions;
3. Self-evaluation – understanding events and experience from the individual's perspective;
4. Protected communication – sharing information with trusted individuals and setting interpersonal boundaries.

Privacy harms

The need to protect privacy can best be explained from the perspective of possible harms and risks. Protection from privacy harms is a legal right for data subjects in many countries. Privacy breaches can be first-order harms (e.g. identity theft, blackmail), or a second-order harm, which increases the risk of other first-order harms (e.g. disclosing data collected by surveillance).

Finally, privacy breaches can also lead to less immediately obvious harms, such as the loss of individual autonomy.

Decisions based on computer algorithms

Decisions affecting people's lives are increasingly based on inferences generated by aggregated information and automated processes, which are linked to natural persons. Modern societies have increasingly placed trust in algorithms to make these decisions. However, the linked data is often incomplete and thus only represents a facet of people's lives. Further, when aggregated incorrectly, misguided decisions can be made, which can have significant impact on people's lives.

Identity theft

It has become an easy and common criminal practice to use another person's identity to commit crimes, accessing resources or obtaining credit in another person's name. The victim of identity theft will often be left with a tainted digital identity, whereby decisions will subsequently be made about her based on information which classes her as a criminal or debtor. It can take a long time to fix such problems, and the victim may have trouble finding employment or mortgages during this time. Careless dissemination of mobile data containing personal information may give criminals more information about a specific person, to possibly make identity theft easier, or more comprehensive.

Blackmail

Blackmail is a crime in many countries, where the criminal threatens to publish certain information about the victim if certain demands are not met. Extensive mobile connectivity datasets can potentially show some incriminating information or expose certain behaviour, which may be used against a victim. The blackmailer may find further unexpected information about his victim on close inspection of a research dataset, while looking for other information.

Use of data by governments

Governments in different countries, but also over time, use data for many different purposes. It cannot be guaranteed that their motives will always be in line with the (good) intentions for which the data was originally collected. Data collected from mobile phones can be highly sensitive personal information. In her book *Privacy in Context*, Helen Nissenbaum states: "[…] *information is a more effective tool in the hands of the strong than in those of the weak.*" The revelations about the NSA and other intelligence agencies have shown, it is not only authoritarian governments that can misuse data gathering powers, but also democratically elected governments. Further, surveillance by public (e.g. government intelligence agencies) or private organizations (e.g. mobile connectivity researchers), can affect people's behaviour and sense of freedom.

Privacy in context

Privacy proponents do not simply want to restrict the flow of information, as many uses of data can be beneficial for data subjects or for the wider public good directly. However, there is a societal interest to ensure information flows *appropriately*, in order to prevent harms from being inflicted from an uncontrolled flow of information. To assess the appropriateness of the flow of

information, contextual considerations such as the capabilities of an adversary or the political environment of the data subject must be taken into account.

Information openly available online is likely to be accessible indefinitely, since it can be stored and republished by anyone. The context a researcher must take into account is therefore not limited in time, and includes a consideration of how datasets could be used to re-identify data subjects using technology, computing power and algorithms that we do not have at our disposal today.

It is not an easy task to predict future technologies and their possible interaction with existing datasets. Such a contextual consideration warrants a conservative approach to the collection, processing and dissemination of personal information with regards to open data publishing. The contextual aspect of privacy is further developed in section D.7 and is a critical part of these guidelines.

International Human Rights Based Framework for privacy

Jurisdictions worldwide vary in their approaches to the right to privacy, giving a patchwork of laws related to privacy. These guidelines will take the international human rights framework on privacy as an ethical compass, while drawing on various more specific legal concepts which have been developed in individual countries, thereby providing some globally agreed upon standards, as well as a compliance with crucial national implementations of the right to privacy or data protection.

D.2 Identifiability

The concepts of personal data (PD, EU) and personally identifiable information (PII, US) are central in determining when legal rules on privacy and data protection apply. These terms are defined below.

The European definition of personal data in the data protection Directive (95/46/EC) is:

> *"Personal data shall mean any information relating to an identified or identifiable natural person ("data subject"); 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".*

Although there is a lack of precision in US law defining PII, the National Institute of Standards and Technology of the US Department of Commerce has defined it as follows:

> *"PII is —any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."*

These concepts are binary by nature: certain data is either PD/PII and therefore protected by law, or it is not. Privacy statutes and data protection laws commonly grant exceptions for data that are anonymised, as they are considered to fall outside the scope of PD/PII. This dichotomy is criticized in academic literature, as it fails to take into account the development of re-identifying individuals through the various techniques that combine non-PD/PII data, such as key-attributes.

Each element of these legal texts carries significant weight and have been given very specific meanings by courts and government agencies over the years. This has led to wide ranging arguments about the specific meaning of “identifiable” or “linkable” (e.g. By whom? Over what time period?).

For the purpose of these guidelines, it is useful to establish an ethical standard which leans heavily on these legal definitions, combined with established authoritative academic literature. Our approach to PD/PII is based on Paul Schwartz and Daniel Solove’s PII 2.0 and Khaled El Emam’s *Identifiability Continuum*. Instead of using the terms PD or PII, these guidelines will simply refer to *personal information*.

The approaches taken by Schwartz, Solove and El Emam are fairly similar, categorising data by the extent to which they are identifiable. The table below is based on El Emam’s *Identifiability Continuum*, and expanded using the “risk of identification” continuum developed by Schwartz and Solove.

Level 1	Raw data	Least re-identification effort	Higher Risk		Identified
Level 2	Masked data		Higher/ medium risk		Identifiable
Level 3	Exposed data		Higher/ medium risk		Identifiable
Level 4	Managed Data		Lower/ medium risk	Threshold boundary	
Level 5	Aggregate/ anonymised data	Greatest re-identification effort	Lower risk		Largely unidentifiable

Level 1: Clearly identifiable data, containing direct identifiers such as name, social security number, address. Data refers directly to a certain person.

Level 2: Use of pseudonyms, but no further masking of key-attributes. Data does not refer to an individual directly, but within the context of re-identification techniques, this is effectively personal information.

Level 3: Obfuscating key-attributes. No objective measure of re-identification is applied, which leaves uncertainty with respect to the identifiability. For example, deleting the last octet of an IP-address still leaves much identifiability when combined with a key-attribute, such as a credit card number stored by an e-commerce provider. Data does not refer to an individual directly, but within the context of re-identification techniques, this could lead to a successful identification.

Level 4: Researcher can manage the risk to data subjects effectively, because an objective measure of re-identification is applied. If the researcher can substantiate his claim that the data cannot be re-identified, it should not be considered as identifiable information. However, this category does still warrant caution on behalf of the researcher and therefore should not be treated the same as Level 5 data.

Level 5: Information is clearly not identifiable, not even with sophisticated re-identification techniques. There is no risk to disseminate this data.

The table serves as an aid to researchers and ethical boards or legal experts when deciding the risk of privacy harm of the data to be collected. It gives an indication of the risk level of the data, which can inform further stages of the research design.

The classification of data must also undergo a contextual test. Further factors need to be taken into account when deciding on the risk of the data, for example available auxiliary data, duration for which the data is stored, the type of dissemination and the likelihood of future development of relevant technology (see D.7)

IP addresses

There is a strong disagreement in the international legal community as to whether an IP address or network traces can be used to identify an individual. For the purpose of these guidelines, we take a conservative and contextual approach. In almost all contexts an adversary armed with an IP address or a network trace, will, in conjunction with other relevant information, be capable of identifying some of the individuals involved.

IP addresses identify devices participating in a computer network that uses the Internet Protocol for communication. Under certain laws and policies, carriers are required to store data such as assigned IP addresses for specified periods of time, sometimes many years. IP addresses for mobile devices, however, tend to be shared among several devices more often than IP addresses for fixed-line connections. It is often stated that mobile carriers do not keep accurate logs of which IP address was assigned to whom at a given moment in time. It is therefore uncertain what auxiliary data is available via a legal order directed at the operator.

If indeed mobile telecom operators do not keep the logs of assigned IP addresses, it could be argued that they do not directly identify individuals. However, when IP addresses are assigned dynamically, the question then arises what the value of a dynamic address is for research. A mobile IP address may not provide an identifier to assign collected information to a certain data subject.

If truncation is to be used for anonymisation of IP addresses, we recommend that the truncation be applied in such a way as to ensure an appropriate level of k-anonymity (see section D.11). Note that the distribution of IP addresses in a given sample may not be uniform, and so removal of the last few digits of the address may not always achieve the expected, acceptable level of k-anonymity.

We therefore recommend researchers only collect IP addresses when they are necessary for a specific purpose (see section D.5 on data minimisation), and assign a higher risk to the disclosure of datasets containing IP addresses.

D.3 Privacy by Design

This section outlines how privacy should be addressed at each stage of research design, including consideration of technical safeguards, organizational procedures and management to improve protection of personal information. As Internet measurement technology increasingly enables data collection, processing and worldwide dissemination at a scale that current privacy laws may not provide adequate protection for, researchers have an ethical responsibility to ensure the privacy of their data subjects. Designing research with privacy in mind is important to maintain the trust data subjects have put in measurement systems. When trust is lost, data subjects may object to their data being processed.

Ann Cavoukian, the Information and Privacy Commissioner of Ontario (Canada) has suggested seven principles of privacy by design, which we apply to the context of mobile network measurements below.

1. Proactive not Reactive; Preventative not Remedial

When designing a project, it is important researchers identify and anticipate the potential privacy issues (see for example section D.7 and D.10). The project should be designed such that these problems are mitigated. It is not satisfactory to think about privacy issues only at the dissemination stage.

2. Privacy as the Default Setting

A research project should be designed such that only information that is needed for the research analysis is collected (see purpose limitation and data minimisation in section D.5). Information should be disseminated so that it does not include personal information and neither key-identifiers nor a collection of identifying secondary identifiers. Appropriate data processing techniques (section D.11) should be employed to achieve this end.

3. Privacy Embedded into Design

Privacy considerations must be part of the whole design process. At the informed consent stage (section D.9a), the choices made should be conveyed to the data subject. The more privacy is embedded in the research, the higher the trust of data subjects will ultimately be.

4. Full Functionality – Positive-Sum, not Zero-Sum
The project design should maximize privacy and utility. It is imperative to iterate the design process and tweak it where possible to achieve a positive-sum approach. These guidelines offer some ethical considerations that can be taken into account to maximize utility.

5. End-to-End Security – Full Lifecycle Protection
Privacy assessments require a holistic approach for the entire project. Data must be stored securely. Data deletion must also be completed in a secure and timely manner, especially with regards to identifiable information and sensitive data.

6. Visibility and Transparency – Keep it Open
Data collection, processing and dissemination must fulfil the promises communicated to the data subject in the informed consent procedure. When it is not possible to be transparent about methods (for example, when revealing the anonymisation technique employed would enable re-identification), it is important to have independent third parties verify the procedures, and communicate this to data subjects.

7. Respect for User Privacy – Keep it User-Centric
Data subjects' trust is of paramount importance to the success of individual research projects and the network research sector as a whole. The data subject must be and feel in control of their data. In addition to clear information for consent and demonstrable strong embedded privacy consideration, the data subject must have some way of understanding how her data and personal information is used in the research process. A simple user-centric privacy interface (such as Google's Dashboard) may be considered.

D.4 Privacy Impact Assessment (PIA)

A PIA is a structured assessment of how personal information is handled during a research project, which also addresses applicable legal, regulatory, policy and ethical requirements regarding privacy. It determines the risks and effects of collecting, processing and disseminating information and shows awareness of the context in which information may be used (see section D.7 for guidance). A crucial element is a detailed examination and evaluation of (technical) safeguards and protections for handling information to mitigate potential privacy risks, such as evaluating the effectiveness of an anonymisation technique (see section D.11). Since failures are a common feature of any system, a PIA should also address unforeseen disclosures of datasets. This will allow for early detection and inform researchers which steps to take in order to mitigate unforeseen situations.

It is appropriate to set up a PIA as part of the research design where a high privacy risk is foreseen, and before any data is collected. Following these guidelines constitutes a PIA. The UK Information Commissioner's Office has published a detailed handbook for conducting PIAs. It is recommended that researchers write up their privacy considerations and the decisions that have been made as part of the research design.

Further, The researcher should consider making the PIA available openly. Doing so will generally have no additional cost, but will convey benefits of credibility and trust, and open opportunities to share more greatly and advance the state of the art in ethical review for the research community. This improves the transparency and thus trust in the sector. PIAs could be collected centrally, for example by a data repository or Internet measurement platform, to share best practices.

Concurrent PIA and research design

These guidelines explain which considerations to include in a PIA for a network research. The outputs from a PIA must be iterated with research (re-)design until the privacy objectives are satisfied. PIAs will therefore usually be constructed through a dialogue with colleagues and/or a legal expert. Conducting a PIA is also a valuable tool to reflect on the research design as a whole.

Points to discuss with experts

The assistive questions in the main part of these guidelines should be discussed with the relevant colleagues and/or a legal expert, and should be weighed as part of the overall privacy assessment. Discussants should agree whether certain technical choices or procedures are higher risk, medium risk or lower risk, and find the appropriate balancing solutions to manage this risk in the relevant context.

D.5 Purpose limitation, compatibility and minimisation

The principles of purpose limitation, compatibility of data use and data minimisation are key issues for ethical accountability when designing a privacy aware research project. The implementation of these measures should be clearly conveyed to data subjects during the informed consent stage (section D.9a) and form an important part of the privacy by design considerations (section D.3) and privacy impact assessment (section D.4). These principles also contain the threat of mission creep, whereby more data would be collected or processed without informing the data subjects.

Purpose Limitation

When designing a research project, the researcher should specify clearly the purpose for which she will collect and process data, especially when it concerns personal information. This information must be sufficiently specific to leave no room for ambiguity or confusion, which is crucial given the complexity and opaqueness of the research field.

A clear explanation of complex technological processes ensures transparency towards data subjects and gives auditors, legal advisors, academic ethical boards and journal editors the necessary information to judge the ethical standard of the proposed project. The specified purpose may not be extended at a later date.

Compatible secondary use

Secondary uses of datasets include data kept for a further longitudinal study, data passed to other researchers doing similar work, or data published along with a paper. Using collected data

for further research is allowed by laws in many countries. However, the intention to use data further should be clearly stated to data subjects before consent is sought.

A data subject may not mind if her data is used, for example, for a longitudinal study on network neutrality when her data is collected to study the change of network speed in a certain area during the day. It is, however, very important to reassess the context of the new use of data and what the foreseeable impact on data subjects will be. The researcher should therefore discuss with colleagues and/or a legal expert whether renewed consent is needed for further research.

Data minimisation

The goal of data minimisation is to limit the amount of personal information that is collected to the least amount necessary to fulfil a specific need, such as a stated research purpose. To limit the availability of data, the researcher should also consider deleting personal information when it is no longer necessary to achieve the stated aim, as required by law in some jurisdictions.

Researchers should therefore only collect information that is relevant and not excessive in relation to the research purposes. The more information is collected, the higher the privacy preservation challenges will be. This can lead to increased legal risks or ethical obligations with regards to information management for the researcher. Data minimisation is the best strategy to mitigate linkability and privacy risks.

D.6a Open Data

Internet measurement platforms often require researchers to make available their datasets in an open data format. Several definitions of open data exist. With regards to open research data, the following aspects are relevant:

1. Making entire databases available;
2. In standardized, machine readable electronic format;
3. To any secondary user;
4. Free of charge;
5. Free of restrictions or obligations (i.e. and open license);
6. For any purpose.

Open research data has a number of values. Firstly, it supports the practice of open science, enabling other researchers to review, retest and validate the analysis a researcher has carried out. Secondly, it enables secondary research, which interrogates the data to explore questions the original research did not focus on. Thirdly, it enables a wide range of alternative re-uses, from artists and entrepreneurs who might take the data and find new non-research centred forms of value within it. There are, however, significant issues with regards to privacy.

The researcher has no effective control over future uses of her dataset once it is publicly available in an open data format. When the research dataset contains personal information, this characteristic poses a challenge to effective research design based on adequate and

proportionate privacy safeguards. Losing control of such a dataset poses a higher risk for the research design.

This problem can be (partly) mitigated by:

1. Applying full anonymisation safeguards, which section D.11 explains is very difficult;
2. Adequately informing data subjects of the intention to disseminate the dataset in an open data format and that secondary uses of their data cannot be predicted or controlled, alerting them to potential risks (see section D.9a).

Open Data Certificate

The Open Data Institute, a UK non-profit, has developed a certificate for open datasets. A certificate can be obtained through a questionnaire, which is based on a self-assessment of the proposed research. If a de-identification technique is employed, the ODI requires this technique to be audited by an independent party. Further, the certification process allows the researcher to set certain limits to the secondary use of the data. We encourage researchers to obtain a license, if it is considered to be relevant, from: https://certificates.theodi.org/

D.6b Disclosure Types

Not all datasets will be suitable to be published in an open data format - for example, when the sensitivity and granularity of the data is high. In such cases, the risk of re-identification will be too high to publish in an uncontrollable open data format. This section recommends methods that do not make a dataset available freely and without restrictions.

Restricted data sharing

The researcher only disseminates research datasets to persons or organisations on request, refusing dissemination when the level of risk is considered too high. The researcher should discuss the expected types of recipients and the corresponding risk level with colleagues and/or a legal expert. Generally, the following risk level can be assigned, although the interest of the recipient and their general information security and privacy standards need to be taken into account:

- An individual researcher from same organisation - Lower risk;
- Sharing with a research consortium - Medium risk;
- Sharing with a commercial entity or government - Higher risk.

The risk level can be lowered if the researcher attaches certain requirements or limitations to the use and further dissemination of the dataset (see section D.8 for data-sharing agreements with third parties). The researcher should inform data subjects of the restrictions on secondary dissemination (section D.9a) and also enforce the requirements and limitations when the third party does not act in accordance with what has been promised to the data subject (section D.8). Not enforcing such commitments would be even worse for public trust in network research than not making promises to the data subject at all.

Managed access

Instead of disseminating the dataset to specified third parties, the researcher can provide managed access to the dataset. Third parties can query the dataset and conduct statistical (or other) analysis. Such an approach allows the researcher to ascertain exactly who accesses the datasets, while maintaining control over its dissemination. The risk level of a managed access system can be considered to be lower.

Interactive methods

The most well-known interactive method of publishing research data is called *Differential Privacy*, developed by Cynthia Dwork and explained in her paper on the topic. It is a particularly robust method, which only gives statistical answers to queries about an underlying dataset. To protect privacy even further, a certain amount of noise is added to the disclosed statistical data. In principle, differential privacy offers a lower risk for privacy, but there are certain limitations to this approach that need to be understood. For example, the uncertainty related by the addition of noise to the data can be exhausted, which means the dissemination must then stop.

Hybrid

The researcher may consider splitting a database that contains personal information that is likely to be re-identified. For example, the identifiers, key-attributes and possibly harmful secondary attributes can be stored in a managed access system, whereas the other network data is published freely in a repository. The researcher can then attach a certificate to the dataset, giving a contact address that informs the third party how she can request access to the full dataset. Such approaches limit the risk of re-identification while maximizing utility.

D.7 Risk Assessment

A comprehensive, quantitative method to assess privacy risk does not yet exist. Therefore, the researcher must assess the balance of risks and benefits based on reasonableness, which can best be achieved through discussion of the proposed research and these guidelines with colleagues. This test should be a cautious one, taking into account future technical developments, auxiliary data and unexpected changes in a political landscape, amongst several other factors that could affect the use of sensitive information contained in mobile connectivity datasets.

This section gives some advice on how to assess the level of risk taking into account several aspects: identifying the adversary who is likely to want to de-identify the dataset and the context in which the data is collected, processed and disseminated.

Adversary

There are three main types of adversaries that need to be considered, as identified by El Emam. The names of the prototypes do not define the type of adversary:

1. “Prosecutor risk” - High risk
    - Wants to re-identify a specific data subject;

- Possesses auxiliary information which can be combined to reveal certain information about data subjects;
- Has legal powers to compel the production of stored information.

2. "Journalistic risk" - Medium to high risk
    - Searches a specific target in the dataset;
    - Possesses auxiliary information which can be combined to reveal certain information about a data subject.

3. "Marketer risk" - Lower to Medium risk (depending on how many individuals can be reidentified)
    - Adversary tries to identify as many people as possible;
    - The more people are identifiable, the higher the risk.

In addition to these criteria, the researcher must also take into account and adjust the risk level based on the expected capacity, skill, auxiliary information and time the adversaries are likely to have available with regards to the re-identification of data subjects.

For example, the US government will likely spend more than $50bn on intelligence services in 2013, which makes it a very capable adversary presenting a higher risk level. Journalists and marketers will have lower budgets, but the motivations may vary. A marketer could have more capacity than a journalist, but a journalist could be more determined to re-identify one person, therefore focusing his efforts. The level of risk depends on the sensitivity of the data.

Context

The data types identified in section D.10 and the discussion on identifiability in section D.2 give an indication of how sensitive the collected data can be, and what level of risk must be ascribed to it. The sensitivity of the collected data must always be considered in the context in which it is collected. For example, collecting the exact locations of mobile phones belonging to American adults carries a certain level of risk with it, but the possible harms are incomparable with the privacy risk of collecting and disseminating locations of rebel fighters or aid workers in war-ridden countries such as Afghanistan, Syria or Sudan.

The central question a researcher should ask herself is: *what would re-identification mean for the data subjects in the particular context?* Further considerations should include relevant technical advances and possibly changing political landscapes, both of which will be educated guesses.

A contextual consideration of privacy risk can thus be a problematic exercise, which is further complicated due to the multidimensional nature of the collected datasets. A contextual approach must incorporate the following elements, and weigh them on a per case basis:

- The activities that can be revealed if the data is disclosed, and possible repercussions for particular data subjects;
- The roles, relationship and power structures that are relevant for the data subject, the adversary and other stakeholders;

- The norms and rules with regards to privacy in the jurisdiction of the data subject (i.e. are there any further protections?);
- The broader values (goals, ends and purposes) of the research and how this benefits the data subjects in their contexts.

Extreme contexts: Do No Harm Principle

Some contexts are extreme to the extent that potential harm posed to individuals by data insecurity and personal identification can likely include arrest, torture, death and longstanding discrimination. Instead of harm mitigation, as outlined in these guidelines, a principle of do no harm is more appropriate, whereby data should be de-identified to the extent that it realistically cannot be re-identified, regardless of the data utility. If this is not feasible, the data should not be collected at all.

The do no harm principle is appropriate in situations where:

- Any of the data subjects are subjected to active discrimination or threat due to their gender, political, sexual or ethnic identities,
- Any of the research subjects live in a context that is marked by extreme social, ethnic or political tension, and in which violence sometimes occurs or could reasonably be expected to occur.
- Public release of research data and identification of individuals might feasibly result in human rights violations.

D.8 Managing unforeseen risks

The possibility that datasets will be disseminated in a way that was not foreseen when the research was designed should always be taken into account. This could be due to a third party distributing the dataset contrary to the data sharing agreement with the researcher, if one has been made. Security breaches or loss of equipment, such as a laptop, are other possible scenarios. Due to the nature of the Internet, it will be very difficult for the researcher to control or contain an unforeseen disclosure. This underlines the importance of a sound analysis of adversaries in section D.7.

Agreements with third parties

Section D.6b explains that an agreement with a third party with whom a dataset is shared, can lower the level of risk assessment of the receiving party. Where appropriate, the researcher should ask the receiving party to make a formal undertaking before sharing the dataset. This could be a data sharing agreement, such as a non-disclosure agreement, confidentiality agreement, another type of contract or a memorandum of understanding.

Such agreements should focus on the access, processing and secondary uses of the dataset, thereby limiting the circulation of possibly sensitive data. If the receiving party must disseminate the data as part of her research obligations, the standards of de-identification applied must be at least as thorough as what has been promised to the data subjects at the informed consent stage (section D.9a). Certain minimum security standards should be prescribed.

Enforcement

At the informed consent stage (section D.9a), the researcher should make certain commitments with regards to secondary uses of the data subject's personal information. The researcher should write these commitments into the data sharing agreement with third parties and should consider sharing the PIA as well (section D.4). If the third party does not act in accordance with the commitments, the researcher has an obligation to enforce the data sharing agreement. Trust in the research sector will diminish when data subjects find their data to be re-identified in some way, when they had given consent after having been promised a rigorous de-identification process.

Security

Sensitive data must be protected through good information security practice, such as physical and personnel security measures. Datasets containing personal information should be stored along with some sort of metadata that describes the dataset and its intended use, where feasible.

Systems should implement strong access controls, to ensure that only those who are authorized to do so access personal information. Access to the datasets should be logged, and logs regularly audited. Further audits should be carried out on topics such as data management, configuration control, intrusion detection and incident response.

Notification of Breaches

In some circumstances, data subjects will need to be notified when a dataset is disclosed unexpectedly. Not every unexpected breach will be serious enough to trigger a notification process. A common standard is to notify data subjects and possibly the relevant privacy regulator when the breach is likely to *"cause a significant risk of individuals suffering substantial detriment, including substantial distress"* (UK Information Commissioner's Office). Notification is only part of the containment strategy. The researcher must plan how to minimise the damage of a breach.

Accountability

If the unexpected disclosure of the dataset has caused harm to data subjects or certain stakeholders, the researcher may likely be held liable or accountable for damages. The researcher must consider the recourse available to disadvantaged data subjects.

D.9a Informed consent

Informed consent is recognized as a central and generally applicable principle and legal basis in scientific research when information is collected directly from the data subject. It is the process of obtaining a legally relevant approval from data subjects after they have been given the chance to understand and consider the use of their data for the research project. This demonstrates that participation is voluntary and that data subjects receive a comprehensible description of the research, including the risks they face and the benefits for research and society as a whole. There are alternative bases in several countries, such as implied consent or the vital interest of the data subject, but they are not covered by these guidelines.

With the current abundance of data, the importance of informed consent has often been neglected as researchers have struggled to identify and contact all data subjects and explain the research sufficiently. There is also a real tension between the ability to collect boundless data and the necessity of asking each and every data subject. However, it is important to note that obtaining the consent of the individual is a valuable legal and ethical safeguard both for the individual and for the researcher.

Informed consent gives the researcher permission to use certain data for the purposes he has specified to the data subject. The researcher still has certain duties towards the data subject. Informed consent is not a *carte blanche* for excessive, disproportionate or unfair data processing.

Consent must be freely given, specific and an informed indication of wishes of the data subject:

- **Free**: consent must be given without any pressure from the researcher;
- **Specific**: consent is for a specified purpose;
- **Informed**: The data subject must be informed how her data is processed for the specific purpose;
- **Indication of wishes**: the data subject must have indicated her wish to give consent by some action, for example ticking a checkbox. Silence, or implied consent, should not be equated with informed consent.

A thorough assessment of the information given to data subjects prior to consent should be part of the Privacy Impact Assessment, which must be developed as part of the research design if a high privacy risk is foreseen, and before any data is collected (see section D.4). The average user is not aware of the risks of complex Internet research or how the data collection and processing systems work. The notice should therefore be written in layman's terms, but at the same time not oversimplify the risks involved.

Key elements of an informed consent notice:

- What data will be collected for the research purpose (see section D.5 and D.10);
- How this data will be processed, de-identified (see section D.11);
- Whether measurements will be user-initiated or run in the background;
- Whether data will be published, and to what extent a person may be identifiable (section D.2);
- That the data will be used for research purposes and how these will benefit society and/or certain stakeholders;
- When the database is split, whether personal information will be stored securely
- Explicitly state if sensitive data will be collected, for example when IP addresses and geo-locations are collected;
- Length of time data will be stored;
- Explain that data will not be fully anonymous, but explain the measures that are taken to ensure the risk of identifiability is minimised;
- Highlight that identity may still be revealed, even after de-identification;

- Give data subjects an indication of risks they may need to consider, stating that one cannot anticipate all the secondary uses in the future, especially if the dataset is disseminated in an open data format.

Emphasis should be added to areas where the risk is identified as being higher, for example by adding bold text. When possible and feasible, the researcher should meet the data subject in person (or telecommunicate) to discuss the points above in person. This is especially important when datasets generated by the research may reveal very sensitive data or when the (political) context is particularly dangerous to a person.

D.9b Informational self-determination in practice

When a data subject gives her consent for a certain research project, it is good practice to also ask the data subject whether she consents to her data being used for further secondary research (section D.5). In the interest of transparency and trust, the data subject should be able to request information about how her data is being used in practice once she has consented to research.

Many countries have given data subjects the right to withdraw their consent to use their data if they can substantiate that there appreciable harm to them, when informed consent is the legal basis for processing data. Internet technologies have made it possible to communicate and inform data subjects about the use of their data directly and real time. Researchers should use this opportunity to allow data subjects to exercise the right to object to certain collection, processing or dissemination of their data via this channel of communication.

D.10 Categorization of mobile data types

Many types of data can be collected via a mobile phone. Some will identify the user directly (identifiers). Other data types will make a re-identification of the user likely with only a few extra pieces of auxiliary data (key-attributes), or will need to be combined with several data to re-identify the user through methods such as fingerprinting (secondary attributes).

The risk level of the collected data depends very much on the context (see section D.7), the combination of data types collected and the auxiliary information that is realistically available. This section therefore describes how the researcher should assess the risk level of the types of data she wishes to collect.

Privacy risks increase when more data types are collected, as it enables inferences to be drawn more easily. Therefore, we apply the labels described above (identifier, key-attribute and secondary-attribute) in a general manner. These labels should be considered as guidance rather than objectively correct in all contexts and must therefore be discussed in detail with colleagues and/or a legal expert.

We describe five different data types to illustrate how to think about them with regards to privacy risks and identifiability of individuals for research design. For this discussion we assume the researcher is collecting a large and high-dimensional dataset.

IMEI number
An IMEI number is the serial number of a phone, and is unique to each device. It does not directly identify a person, but as devices are generally owned and used solely by a specific individual, it should be considered a key-attribute. Auxiliary data, such as billing information from the telecom operator, can identify that device's owner or user.

Current IP address
IP addresses identify devices participating in a computer network that uses the Internet Protocol for communication. Internet service providers and telecom operators may keep logs of IP addresses that are assigned to certain fixed line broadband connections. IP addresses for mobile devices, however, tend to be shared amongst multiple devices over time.

It is frequently stated that mobile carriers do not keep accurate logs of which IP-addresses were assigned to which device at a given moment in time. This would make such data a secondary attribute. Under certain laws and policies, however, carriers are required to store data such as assigned IP addresses for specified periods of time, sometimes many years, although this can be further complicated by common techniques such as *Network Address Translation* (NAT) that allows multiple devices to share a single IP address. In the case that IP address logs do uniquely identify devices, IP addresses should be considered as key-attributes that render a data subject identifiable.

Name of carrier
Relatively few mobile carriers exist in the US, with subscriber numbers ranging into the tens or hundreds of millions per carrier. However, worldwide there are many carriers with far fewer subscribers. This is especially true in Europe and other parts of the world, where carriers exist that only have a few thousand subscribers. The name of the carrier will likely only be a secondary attribute, but the amount of auxiliary data required to identify a person will depend on the size and type of carrier. The name of an obscure carrier could easily be the necessary data point that allows a subscriber to be identified individually along with other attributes.

Battery level
The current battery level of a mobile device is widely considered to be irrelevant for the identifiability of a mobile phone user. However, the rate of decay of battery power of devices, when monitored over time, allows differences to be found. While this does require significant analysis, it is not beyond the bounds of possibility, and may be made possible due to the perceived low sensitivity of gathering and releasing such data.

Similarly, research has indicated that it is possible to identify many Internet users uniquely by analysing their browser configurations (so far only proven on desktops,). It should always be remembered that seemingly irrelevant data types can thus become critical in re-identification.

Location
A GPS location gives an accurate position of a device. Locations of mobile phones can also be collected when the GPS is switched off, by triangulating the position with regards to Wi-Fi access points or cell towers. When a mobile phone moves, it is usually in the direct possession of a person, typically its owner. It can therefore be assumed that the GPS or triangulated locations reveal the location of a specific person at a certain time and, more importantly, the series of locations through which the device and its owner have moved over time.

Identifiability is very dependent on the context of the geographical location and the local population density. GPS location may not be a key-attribute, but research has shown that human mobility traces are highly unique.

Below are some further data types that can be collected via network measurements.

- Download throughput;
- Upload throughput;
- Current DNS resolver;
- Visible networks;
- Traceroutes;
- Identify active radio antenna;
- Operation system & version;
- Current memory usage;
- Names of installed applications;
- How many applications are running;
- Cookies;
- IMSI (International mobile subscriber identity).

D.11 Categorization of de-identification control techniques

De-identification techniques are useful tools to make it more difficult for an adversary to identify individuals in a dataset. Full de-identification is very difficult to achieve, however, and "anonymised" datasets have often been re-identified. For example, a former governor of a US state was identified by combining a public "anonymised" healthcare dataset with auxiliary data.

More sophisticated methods exist, such as using *fingerprinting* techniques, where inferences about individuals can be made based on seemingly non-identifying data types. Especially in multi-dimensional datasets, such as often created with mobile Internet measurements, it has proven to be possible to uniquely identify a large part of the dataset. Human mobility traces, for example, have been found to be highly unique: only four location data points over the course of one day were necessary to identify 95% of individuals uniquely. The settings, configurations and combination of plug-ins of certain applications - such as a web browser - can also be used to distinguish individuals and identify them with adequate auxiliary data.

The premise of these guidelines is the generally accepted understanding that full de-identification is not possible without significant loss of utility. Exemptions for “anonymised” datasets in existing privacy laws are therefore not a suitable ethical standard. To guide the researcher in the choice of suitable de-identification technique, we describe some methods below and attribute a level of robustness to them (higher, medium or lower robustness). None of these techniques are perfectly resistant against determined attackers with access to sufficient auxiliary data sources, computing power and determination. This does not mean, however, that only the strongest de-identification technique is suitable for network research. The choice of technique should be guided by an assessment of the risk (section D.7).

Perturbation

One of the simplest approaches to de-identifying a datasets is to add ‘noise’ to genuine values. For numeric quantities this can simply be the addition of random figures according to an appropriate probability distribution. For categorized data this can result in attributes being re-assigned in various ways. Rick L. Wilson and Peter A. Rosen discuss the use of perturbation and its impact on the ability for knowledge discovery in their paper “Protecting Data through ‘Perturbation’ Techniques: The Impact on Knowledge Discovery in Databases.”

Truncation

In numerical data, truncation limits the number of significant digits stored, thereby making such values less accurate. Truncation can also be applied to IP addresses, postcodes or other key-identifiers.

- Lower robustness;
- Simple and useful when dealing with fields that contain sensitive data;
- Accuracy is decreased.

Randomization & permutation

This approach refers to reordering the values of a column without losing the accurate values in the dataset.

- Medium robustness;
- Useful when dealing with fields that contain sensitive data;
- Useful for maintaining statistical utility and accuracy, such as aggregate counts; averages and distribution of data;
- Individual accuracy is lost.

Quantization

Similar to truncation, quantization constrains continuous values into elements of a defined set. This could, for example, take the form of grouping values such as exact height into a number of height ranges.

- Medium robustness;
- Care must be taken that all groups contain sufficient individual entries;
- Accuracy per individual record is lost.

Pseudonymization

This method replaces directly identifying fields of a dataset with non-identifying values. A common example is to replace identifiable IP addresses with (linkable) prefix-preserving pseudonyms or hashes of data. This is typically aimed at maintaining the links between a group of records, whilst removing the ability to easily identify the record identifier itself.

- Lower robustness;
- Preserves all key and secondary-attributes, so risk of fingerprinting or linking with auxiliary data remains.

K-Anonymity

K-Anonymity is widely used in network research, which ensures that any record in a database must be identical to some number of other rows, forming a group of size *k* that is indistinguishable from each other. This approach may take the form, for example, of grouping subjects' locations into sufficiently large areas such that no set of locations is unique to any individual.

Latanya Sweeney describes this method as follows: *"A release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release."*

- Medium/higher robustness, depending on suitable threshold for "k";
- A quantifiable probability that individuals could be re-identified exists. The probability that a data subject can be identified is 1/k, where k is the size of granularity chosen;
- It may still be possible to infer sensitive information about a person, even if direct identification is impossible. Further, the attributes shared by an entire group, such as a particular disease or condition, could be sensitive;
- Knowledge of the specific k-anonymisation algorithm could be sufficient to re-identify a dataset;
- Datasets with k-anonymity applied have been re-identified. See, for example, the AOL search data case, where k-anonymity was shown to be useless for some individuals in high-dimensional datasets and the information revealed was very damaging;
- An appropriate threshold for "k" depends on the context in which data is collected and disseminated. The researcher must consider this on a case-by-case basis.

Various extensions to k-anonymity have been proposed to mitigate weaknesses against various forms of attack. A full discussion of these is not appropriate here, but for more details see the following documents:

- ❏ The paper "t-Closeness: Privacy Beyond k-Anonymity and -Diversity" [.pdf];
- ❏ The presentation "k-Anonymity and Other Cluster-Based Methods" [.ppt];
- ❏ The presentation "Data Anonymization Techniques" [.pdf].

Differential Privacy

The concept of differential privacy (developed by Cynthia Dwork) is an interactive privacy method for statistical databases (also covered in section D.6b). Differential privacy does not guarantee that a privacy breach will not occur, but it guarantees that the privacy breach will not

occur due to the data in the database. Breaches that can happen if data is in the database could have happened even if the data weren't in the database. This accommodates any and all possible auxiliary information available now or in the future. However, differential privacy has some limitations (see D.6b), so the usefulness of the concept must be well researched before it can be applied to specific network research.

D.12 Bibliography

Much has been written about privacy, also with regards to Internet measurements. These guidelines aim to provide a summary of the key considerations. Much more elaborate guidelines, papers and books are available. Some literature from which the authors have drawn inspiration is listed below:

General privacy literature (books and papers)

- Nissenbaum, H. (2010) *Privacy in Context: Technology, Policy and the Integration of Social Life*, Stanford: Stanford University Press.
- Pfitzmann, A., Hansen, M. (2010) "A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management" v0.34, available at http://dud.inf.tu-dresden.de/Anon_Terminology.shtml
- Schwartz, P. M., Solove D. J., (2011) "The PII Problem: Privacy and a New Concept of Personally Identifiable Information", 86 N.Y.U. L.Q. Rev. 1814, available at: http://scholarship.law.berkeley.edu/facpubs/1638
- Solove, D. J., (2006) "A Taxonomy of Privacy". University of Pennsylvania Law Review, Vol. 154, No. 3, p. 477, January 2006; GWU Law School Public Law Research Paper No. 129. Available at http://ssrn.com/abstract=667622
- Westin, A. F., (1967) *Privacy and Freedom*, New York: Atheneum
- Whitman, J., Q., (2004) "The Two Western Cultures of Privacy: Dignity versus Liberty", Faculty Scholarship Series. Paper 649. http://digitalcommons.law.yale.edu/fss_papers/649

Existing relevant guidelines

- Bailey, M., Dittrich, D., Kenneally, E., Maughan, D., (2011) "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research", available at http://www.cyber.st.dhs.gov/wp-content/uploads/2011/12/MenloPrinciplesCORE-20110915-r560.pdf
- Bailey, M., Dittrich, D., Kenneally, E., Maughan, D., (2012) "Applying Ethical Principles to Information and Communication Technology Research: A Companion to the Department of Homeland Security Menlo Report", available at http://www.cyber.st.dhs.gov/wp-content/uploads/2012/01/MenloPrinciplesCOMPANION-20120103-r731.pdf
- Cavoukia, A., (2009, 2011) "Privacy by Design: The 7 Foundational Principles", available at http://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/

- Chen, B., Kifer, D., LeFevre, K., Machanavajjhala, A., (2009) “Privacy-Preserving Data Publishing”, available at http://dl.acm.org/citation.cfm?id=1640479
- Coull, S.E., Keneally, E. (2012) “A Qualitative Risk Assessment Framework for Sharing Computer Network Data”, available at http://ssrn.com/abstract=2032315.
- El Emam, K., (2008) “Heuristics for de-identifying health data”, available at http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4588232
- El Emam, K., (2010) “Risk-Based De-Identification of Health Data”, available at http://www.ehealthinformation.ca/documents/parat/riskdeid.pdf
- El Emam, K., Hassan, W., (2013) “The De-identification Maturity Model”, available at http://waelhassan.com/wp-content/uploads/2013/06/DMM-Khaled-El-Emam-Wael-Hassan.pdf
- UK Information Commissioner’s Office (2012) “Anonymisation: managing data protection risk, code of practice”, available at http://www.ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation
- US National Institute of Standards and Technologies, (2010) “Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)”, available at http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf

Recommended relevant papers

- Coull, S.E., Monrose, F., Reiter, M., Bailey, M., (2009) “The challenges of effectively anonymizing network data. Conference For Homeland Security, Cybersecurity Applications & Technology, 0:230–236, 2009. Available at http://dx.doi.org/10.1109/CATCH.2009.27.
- Dwork, C., (2006) “Differential Privacy”, available at http://research.microsoft.com/apps/pubs/default.aspx?id=64346.
- Sweeney, L., (2002) “k-Anonymity: A model for protecting privacy”, available at http://epic.org/privacy/reidentification/Sweeney_Article.pdf.
- Ioannidis, J. P. A., (2013) “Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research”, The American Journal of Bioethics, 13:4, 40-42, available at http://dx.doi.org/10.1080/15265161.2013.768864.
- De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., Blondel, V. D., (2012) “Unique in the Crowd: The privacy bounds of human mobility”, available at http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html.
- boyd, d., Crawford, K., (2012): “Critical questions for big data”, Information, Communication & Society, 15:5, 662-679, available at http://dx.doi.org/10.1080/1369118X.2012.6D.678878.
- Narayanan, A., Shmatikov, V., (2008) “Robust De-anonymization of Large Sparse Datasets”, available at www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
- Ohm, P., (2009) “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization”, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.
- Ohm, P., Sicker, D., Grunwald, D., (2007) “Legal Issues Surrounding Monitoring During Network Research”, available at http://conferences.sigcomm.org/imc/2007/papers/imc152.pdf.

- Article 29 Data Protection Working Party (2013) “Opinion 06/2013 on open data and public sector information ('PSI') reuse”, available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp207_en.pdf