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

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).

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.

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).

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.