Privacy intrusion metrics: concepts and considerations
IPCO TAP Report
21 Oct 2021

Introduction

1. 'IPCO needs to be able to measure, weigh, rank and quantify activities under the Act; the provision of metrics to assess the level of privacy would be valuable and would aid IPCO both in correct assessment of warrant applications and in inspecting the use of the IPA powers. However, there is no simple quantification method.' [1]

2. This paper contributes to the debate by outlining key concepts and considerations that underpin quantification methods for privacy intrusion, in the context of Investigatory Powers Act 2016 (IPA).

Background

3. Law Enforcement Agencies (LEAs) have a duty to investigate the activities of subjects of interest - for National Security, prevention of serious crime, and for the economic well-being of the UK. In doing so, they seek to collect communications (content and data) relevant to these activities, as permitted under warrant by the Investigatory Powers Act 2016. All warrants under the Act must satisfy *necessity* and *proportionality* (see codes of practice [1]):

    a) The application must be necessary as meeting a requirement for National Security, prevention of serious crime or economic well-being and be within the remit of the Authority.

    b) The application must be proportionate in balancing the seriousness of the intrusion into privacy and any other rights that may be engaged, against the need for the activity in investigative, operational or capability terms. No interference with privacy should be considered proportionate if the information which is sought could reasonably be obtained by other less intrusive means.

4. Current practice in assessing proportionality is very much qualitative, relying on the professional judgment of the officer submitting a warrant application, the single point of contact (SPoC) for the Authority, the Secretary of State, and the Judicial Commissioners. In many cases this works well, when choices of approach are limited, and their impacts are well understood by the various parties involved in the approvals process. However, the rapid

[1] 1 'Metrics of Privacy Conference': Report of a discussion meeting organised by the Technical Advisory Panel of the Investigatory Powers Commissioner's Office on Metrics of Privacy, 14 November 2018.

evolution of communications systems (e.g. 5G, Internet of Things (IoT)) means this approach may not be adequate and is by no means future-proof, as it places a heavy burden on the individuals to keep abreast of developments. Given this constantly shifting technical landscape, we seek to understand how systematically calculated quantitative metrics could assist in making these judgements.

Quantification

5. A quantitative approach, employing a robust and stable quantitative model, could be of significant value, both in improving the assessment of warrant applications, and in providing a framework for assessment that is understandable, shared across the community, and which can survive disruptions in the future. This note sets out some of the key issues and considerations for a quantitative approach, applicable to both collateral privacy intrusion and protecting the rights of targeted individuals under investigation.

6. Quantifying aspects of privacy intrusion is subtle and complex because it involves differential, personal judgements, not only about the intrinsic information in a data set, but also about derivations and inferences, which may be difficult to predict or quantify. Further, it may require comparison not only of data sets, but also the methods used to acquire data sets, in which case we may not know the amount of data that will be acquired, and the scale of (direct or collateral) intrusion, until after the actual acquisition.

7. In the remainder of this paper we consider data formats, data precision and accuracy, information, cost of derivation and inference, method of acquisition, volume of data, value versus volume, type of analyst (human/machine), type of intrusion, and personal identification and anonymity.

Data Formats

8. *Structured* data has a pre-defined data model in a tabular form and is often stored in relational databases or spreadsheets; it may be sourced from online forms, GPS sensors, network logs, web server logs, etc. *Unstructured* data has no pre-defined data model and includes text, video files, audio files, mobile activity, etc.; it may be sourced from social media posts, satellite imagery, surveillance imagery, etc. *Semi-structured* data is a form of structured data that may not conform to a given tabular structure but still contains sufficient tags or other markers to separate semantic elements.

9. We assume here at least semi-structured data, where data sets contain individual *records*. We assume each record has finite arity (number and type of values) and consists of (typed) *component values*. We assume types are sets, e.g. Integers, Strings, etc.

10. *Encrypted* data is only accessible to those with a key. That is, it is data that has been encrypted by or for the end-users, for which the key is held by the end-users and/or the encryption service. If the Authority does not possess the key, and there is no likelihood that the key will subsequently become available, then we assume the encrypted data *content* has no privacy intrusion. There may however be value (and hence privacy considerations) in the number of encrypted items and their metadata, including time and volume.

Precision and accuracy of data

11. Data values may be precise or imprecise. In the latter case the value may be a range (i.e. a subset of a type such as salary or age range) or a probability. Accuracy refers to how close

the values are to the true values, e.g. they may have been corrupted during collection, either systematically or individually.

Information contained in, derived or inferred from a data set

12. Data sets (and constituent records and individual component values) *contain* information that can be inspected and used by analysts; the information and subsequent analysis will depend on the types and precision of the data. Typically, there is little information (apart from personal identification) in a single (precise or imprecise) component value, for example in a telephone number, a time interval (duration) of a call, or a car number plate. More likely, the interesting/useful information is one or more records, for example, a record consisting of an incoming telephone number, a call duration, and the name of the telephone owner, or a record consisting of a number plate and the name of the registered owner, or in a collection of records, e.g. the number of calls placed in a given time period.

13. Often, we need to consider not only the intrinsic information, but what *further information* we can *derive* or *infer* from a data set – now, or at *some point in the future*. Further information can be obtained in one of two ways.

14. Further information may be *derived* from a data set value/record/component value by applying a *mapping*, which may also involve another data set. For example, we can calculate the average of several component values in a record such as time intervals or distance measures, or derive a postcode from an address, or the registered owner from a vehicle number plate. In the first case we assume knowledge of the formula for average, in the second case we require access to the UK postcode mapping, and in the last case we require access to the DVLA service. Taking a higher order approach, we can simplify this and say there is only one mapping required (which may be higher order). Derivation is always deterministic, i.e. the result depends only on the inputs to the mapping. Success depends only on availability of the mapping, and availability may change over time (more mappings made available). However, the outputs may be probabilities, confidence intervals (the uncertainties associated with values) or confidence regions (the multi-dimensional generalisation of intervals).

15. Further information may be *inferred* from a data set value/record/component value in at least two ways. The first way is by application of an inference algorithm directly to the data of interest. For example, we can apply the (unsupervised) expectation-maximisation algorithm to cluster the data. The second way is by application of an inference algorithm to a training data set, the result of which is then applied to the data of interest. An example of this is use of a machine learning (ML) algorithm for image recognition or classification. Inference is typically non-deterministic, for example, the resulting clusters or the classifiers, may be different every application. In the machine learning case, the information that is inferred depends on properties of the training sets and the availability of "good" training sets, which are subject to the usual considerations such as bias and quality.

16. When assessing privacy intrusion, it may be important to have a likelihood of the future availability of suitable mapping, inference algorithm, or training set(s), as well as the cost of a derivation or inference. It may also be important to consider the intrusion due to the mapping itself. This is particularly important for ML algorithms which rely on historic 'real' data, which could be considered as collateral intrusion.

Costs of information derivation and inference

17. It may be pertinent to consider the cost (e.g. time, space) involved in any derivation or inference, and to set thresholds. For example, we may deem the costs so prohibitively high that we can exclude a possible derivation/inference; this may be due to a high computational cost, or time required to obtain a data set.

18. More likely we will want to consider how costs are balanced against the value of derived/inferred data and intrusiveness of the mapping/method. As an example, consider an encrypted message. The plaintext may be of high value and be correspondingly highly intrusive (it was deliberately protected). Now consider its decryption. The cost may be prohibitive if the encryption is secure, but it may be cheap if the key can be obtained by stealth or by duress, in which case the level of intrusion increases enormously.

Methods for acquisition of data

19. Data may be acquired by permission, or by warranted surveillance, access to equipment (e.g. stored on a phone), targeted equipment interference (TEI), and/or by inspection of bulk data. The last typically includes structured data sets such as electoral roll, telephone directories, travel-related data etc. Bulk *communications* data contains the "who", "where", "when", "how" and "with whom", but not the message content.

20. We may need to compare *methods* of acquisition, when we do not know volume of the data that will be obtained, or the precision of the data. This is particularly pertinent when we need to compare acquisition by different surveillance methods, and where the (likely) volumes and precision may be different. Volume depends on the environment (e.g. how many people were in range of the acquisition technique on a given day), precision may depend on both the acquisition equipment and the environment. Consideration of the method of acquisition also applies to queries on bulk data, e.g. comparison of two different queries. In all cases, the key consideration is we do not know what we will acquire until we do it.

21. For most acquisition methods, we expect collateral and targeted intrusiveness to be inversely related. For example, bulk interception (BI) has low targeted but high collateral intrusion, whereas targeted equipment interference (TEI) and directed surveillance (DS) have high targeted but low collateral intrusion. It may be useful to quantify and/or visualise the relationships.

Volume - quantitative dimensions of data sets

22. Key dimensions of a data set volume are the *cardinality* of the set (the number of elements in a set) and the *arity* of the individual records. Intuitively we would expect cardinality and arity to correlate with intrusion and there to be an interplay between the two. For example, consider comparing a small sized data set containing large and detailed records with a large sized data set containing records, each of which contain little information. We might expect to combine cardinality and arity to obtain volumes that can be compared. The obvious a *combinator* (i.e. operator) is multiplication. There are a number of interesting questions: how to assign weights to the component types, how to combine weights from the components, is multiplication the appropriate operator to combine cardinality and with arity? For example, consider an approach that sums the component weights and multiplies

that by size, comparing a) a data set of 5 records where each record consists of name, address, car registration, phone number, email address, with b) a data set of 2,000 records where each record consists solely of a phone number. Assume weights range from 1..10 and the weights of name, address, car registration, phone number, email address are 10 , 2, 1, 1, 2 resp. The weight of a) is 5*16 = 80 and b) is 2000*1= 2000, and so we might conclude that in terms of rank, a<<b. Is this what we would expect? How do we determine what we *would* expect? Is the ratio of 10:1 appropriate for weight of name and phone number?

23. Any such approach begs further questions such as: how to select the component weight values, what are the ranges and ratios between weights (e.g. is the ratio of 10:1 appropriate for weight of name and phone number?); how to scale the effect of set size, what properties do we expect of the weight and volume functions. Further, we remark that this approach assumes the information in the component data is independent, which is unlikely. If we believe that interrelationships are simple, low-degree polynomial regression may be appropriate, and if the dependencies are linear in nature, then principal component analysis (PCA), performed by singular value decomposition (SVD), could do the dimensional reduction. However, this could make interpretation by humans more difficult. Another approach would be to assign a weight to the entire record. This might more accurately reflect that overall information is more than the sum of the parts, particularly with respect to personal data.

24. Whatever the approach, we would expect the practice of regular deletion (automatic or manual) from a dataset to reduce the *intrinsic* level of intrusion, compared with a possibly growing, but never pruned set. However, there are subtleties because information may have been derived or inferred prior to deletion. For example, consider set A from which set B is derived and set C is inferred. Removing part or all of A would not affect the levels of intrusion of B and C.

Value versus volume

25. There is a trade-off to be made between value and volume. There is good evidence to suggest that both value and intrusion are not linear with respect to volume, they follow characteristic sigmoid curves. But intrusion converges more slowly (because it is measured against the collective and value against the individual). This means there is a range of volumes where we obtain little added value, but intrusion is still increasing significantly. In other words, there is a point after which the marginal value is small compared to the additional intrusion. However, this point could be a long way out, especially when combining data sets collected in different environments.

Human or computer analysis and teaming

26. To what extent does the nature of the agent performing the inspection or analysis affect a metric of intrusion, and how should we reflect humans and system working as a team? While we said earlier that intuitively volume correlates with intrusion, this may not be the case when the inspection is (possibly only partially) by a human. For example, does it matter that there are 2000 records in a data set, if a human never inspects that set? How would inspection by machine compare with our previously stated view of encrypted data?

27. The introduction of analysis by machine means that the limit of data for analysis increases significantly, but as above, there are trade-offs and the pertinent question is*: is this increase proportionate if the (value) gain is small?*

Type of intrusion

28. Common types of intrusion are collection and subsequent storage of new data, retention of (existing) data for retrieval and analysis, and disruption or denial of service.

Personal data and anonymity

29. *Personally identifiable information* (PII) is information that relates to an identified or identifiable person, for example a name (first name, surname) or number (vehicle number plate). Note the latter is an example of derived information. Data may reference a person but not be considered personal data when the information does not relate to them, for example, the data is a randomly assigned number. But if there is a means of mapping that value to other PII (de-identification), then we say that data is pseudonymised. This raises questions about k-anonymity.

30. A data set has **k-anonymity** if by suppressing or generalising the values in selected data types, then for every record, there are at least k-1 identical records. This means k-anonymity has two variables: k and l – the list of types to be suppressed/generalised. We make these explicit and say a data set is **(k,l)-anonymous** if by suppressing or generalising the data types in l, it has k-anonymity. We say there is *no anonymity* when the greatest k for which the data set is (k,[])-anonymous is 1, i.e. all records are distinct. In what way are these useful concepts for intrusiveness? For example, is a data set more or less intrusive when k is higher/l is smaller? How might we reflect these concepts in the weights assigned when determining information volume?

31. Also, while k-anonymity may be effective at making assignments irreversible to an individual, is it intrusive to make probabilistic statements (a posteriori), or to identify a group of individuals having common characteristics? Further, does pseudonymisation of data affect intrusiveness? GDPR indicates a positive answer, in that pseudonymisation makes unauthorised association harder, which concurs with our view that intrusion is concerned with both data and derivations and inferences, i.e. operations as well as the data.

References

[1] Investigatory Powers Act 2016 – codes of practice - GOV.UK (www.gov.uk)