Location Tracing and Potential Risks in Interaction Data Sets

Oliver Duke-Williams[1]

Received: 19 December 2016 / Accepted: 20 November 2017 /
Published online: 15 December 2017

Abstract Location-aware mobile phone handsets have become increasingly common in recent years, giving rise to a wide variety of location based services that rely on a person's mobile phone reporting its current location to a remote service provider. Previous research has demonstrated that services that geo-code status updates may permit the estimation of both the rough location of users' home locations and those of their workplaces. The paper investigates the disclosure risks of a priori knowledge of a person's home and workplace locations, or of their current and previous home locations. Detailed interaction data sets published from censuses or other sources are characterised by the sparsity of the contained data, such that unique combinations of two locations may often be observed. In the most detailed 2011 migration data 37% of migrants had a unique combination of origin and destination, whilst in the most detailed journey to work data, 58% of workers had a unique combination of home and workplace. The amount of additional attribute data that might be disclosed is limited. When more coarse geographies are used their still remain a non-trivial number of persons with unique location combinations, with considerably more attributes potentially disclosable.

Keywords UK · Census · Interaction data · Disclosure

Introduction

Amongst the outputs from recent UK censuses have been sets of interaction data (also known as 'flow data' or 'origin-destination data'). In contrast to aggregate census data which provide information about a defined area (from an entire nation to a small zone) and microdata which provide individual level observations, census interaction data

✉ Oliver Duke-Williams
o.duke-williams@ucl.ac.uk

[1] Department of Information Studies, UCL, London WC1E 6BT, UK

provide information about people moving between one location and another. The most common interaction data relating to people are migration data sets and commuting data sets; where migration data typically report moves between a present residential location and a former usual residence, and commuting data report on daily journeys between a residence and a place of work. This paper uses UK examples, although data with the same structure are available in a number of countries.

Previous research (Krumm 2007) has demonstrated that it is possible to estimate the location of a person's usual residence by examining anonymously logged data in GPS units, whilst Golle and Partridge (2009) have argued that it also possible to estimate workplace location for some people, and argued that this would pose a risk for some previously released data sets. These risk assessments rely on individual level location trace data. The use of smart phones and other portable devices which can determine – to varying degrees of accuracy – their current location (and by implication, that of an owner or user) has become widespread. Such devices allow a wide variety of location based services to be offered, some running as software on the device itself, and others running as a remote service. The term *location based services* has a number of definitions that are not necessarily consistent (Küpper 2005), and also includes many applications not related to portable devices. Data produced by location based services may permit service owners or third parties to estimate home or workplace locations of users. This paper examines the potential disclosure risks to individuals through publication of UK interaction data sets, analogous to Golle and Partridge's work on US data sets and investigates whether the level of risk is similar in the UK data as suggested for US data, and thus whether UK interaction data are potentially 'unsafe'. The extent to which there may be a risk of disclosure is affected by disclosure control procedures used in conjunction with release of the data. The paper contrasts a number of sets of interaction data released with different approaches to disclosure control in order to further explore this issue. The general risk of interaction data are considered, and possible mitigation strategies in the form of disclosure control arrangements or access restrictions.

The paper starts by reviewing general observations about the role of confidentiality and privacy in data released by national statistical agencies. The specific area of UK interaction data is considered, as these data have particular characteristics that may increase the risk of disclosure. The methods used by Golle and Partridge to analyse data from Longitudinal Employer Household Dynamics (LEHD) program are reviewed, and then applied to a number of data sets produced as outputs from UK Censuses.

Confidentiality

For any statistical agency that intends to release some data, an important consideration is the preservation of confidentiality relating to those data. The term 'confidentiality' refers to preventing disclosure of information to unauthorised parties. Fellegi (1972) characterised 'inadvertent direct disclosure' as depending on two elements: firstly, that an individual must be identifiable in the released data (identity disclosure) and secondly the released data must reveal information further to that which was used in the identification process (attribute disclosure). For statistical agencies, being seen to ensure confidentiality may be an important element of building public trust. However, Singer et al. (1993) studying the 1990 US Census, argued that trust in confidentiality had only a limited effect on response rates and that this relationship varied for black and white respondents. The effect of trust may vary depending on the nature of the survey taken: in the case of a sample study, the individual has the ability to opt out, whereas in the case of a census the individual faces legal coercion to complete a census form.

Confidentiality of public data (data gathered by public agencies with the express intention of publication of results) has two main aspects. Firstly, confidentiality must be maintained over raw data. Thus, statistical agencies must ensure that their data are stored and processed in a secure manner, without inadvertent or deliberate disclosure. Confidentiality of raw data is typically ensured by a range of legal, physical and digital data security. Media stories about problems in the protection of public data such as the loss of child benefit client records (Poynter 2008) focus on actual or potential confidentiality breaches through failures of internal data security. The second aspect of confidentiality comes into play in the preparation for and release of the data. A combination of tactics are used to ensure confidentiality in released data. Some data sets require individual or corporate users to sign license agreements; these typically contain legal undertakings not to disclose information relating to individuals. However, legal protections alone are not usually considered sufficient to ensure that confidentiality will be maintained, and thus further measures are also taken. These further measures take the form of statistical disclosure control methods which modify the data that are to be publicly released, in or order to reduce the risk of disclosure.

The first stage of ensuring confidentiality in microdata is usually a process of anonymisation. However, effective anonymity in personal data is not necessarily achieved simply by removing explicit identifiers, as a simple combination of more general personal attributes can uniquely identify many people. Sweeney (2000), using data from the 1990 US Census, found that 87% of American citizens could 'likely' be uniquely identified through a combination of 5-digit ZIP code, birth date and sex. Golle (2006), in attempting to repeat this analysis, found that 61% of Americans in 1990, and 63% in 2000, could similarly be uniquely identified. Identification of individuals in this manner in a source of concern, as these general variables - age, date of birth and area of residence - can be easily determined for many people from other means.

For data to be released in aggregate form (summed from a set of individual records), disclosure control methods exist that can be applied either prior to or after aggregation. Pre-aggregation methods involve the modification of individual records from which the results are to be aggregated. Post-aggregation methods involve the modification of the table of results, and can include various forms of rounding, random perturbation and cell suppression (Willenborg and De Waal 2012).

General attacks on a target data set (that is, an attempt by a third party to extract information from the data beyond that which was intended by the data provider) use a pre-existing attack data set. Typically, the attacker will try to match 'known' records in the attack set against records in the target data set using a set of key variables. In general, the more key variables available, the more chance there is of finding a unique match. Matching is confounded by variations in the ways in which fields are coded (for example exact age vs. classified age) and by time-dependent key variables, such as occupation, which are subject to change over time. In the case of a census, time-dependent variables are captured as of a particular known point in time. In the case of surveys collected over a fixed field period, the time can be estimated, but is probably not known (by the attacker) with any precision.

The process of using an externally sourced key to attach identities to supposedly anonymised data is known as re-identification. Whilst this is more straightforward in individual data, there is also a risk of re-identification in aggregate data if the aggregate data contain unit values (i.e. only one person has a given combination of values).

Problems with Sparsity in Interaction Data

Particular problems of disclosure arise in data that contain small values or unique observations. Interaction data are typified by the presence of small values, and might therefore represent a particularly significant risk. They conceptually take the form of a matrix, with *n* origins and *m* destinations, thus having *nm* potential flows. Each flow is typically disaggregated using a number of univariate or multivariate observations of the characteristics of the people in the flow. Depending on the reporting geography, these matrices can be exceedingly sparse. The most sparse interaction data publically available for the UK are from the 2001 census, and show migration between and within 223,060 Output Areas (OAs) in the UK, giving a total of $223{,}060^2$ potential flows, with each flow being disaggregated by age (three broad groups) and sex, giving almost 300 billion distinct cells.

Risks from Location Traces

Location tracing is increasingly feasible given the ubiquity of consumer level electronic equipment that is location aware. Levinson et al. (2011) identify three different methods used by iPhone handsets, for example, to determine current location. A growing number of services offer information to people on the basis of their current location - restaurant recommendations for example, or customised entertainment listings - and in order to use these services subscribers must allow their phone handset to report its current location. The service providers thus have the potential to gather large amounts of data that indicate the location of a given handset at different points in time. Some researchers have highlighted a potential lack of privacy through monitoring of current locations, for example Allan and Wardle (2011) highlighted the fact that Apple iPhone and iPad devices keep a log of sensed locations. Concerns raised in the popular media have focussed on the combination of location based services and social networking: services such as FourSquare directly couple these by sending details of a person's location to other members of their social network; this is particularly significant where updates are distributed using Twitter, as users can typically be 'followed' by any other Twitter user unless they take an active decision to block people. Users of services are mindful of the possible risks. The Guardian newspaper (2010) reported the results of a survey of 1645 social network users who owned devices capable of location finding; of these 55% expressed concern about privacy, with specific fears over burglary and stalking also being identified. Disclosing a person's location might also be used to embarrass them or harm their reputation by associating them with particular activities. However, it continues to be the case that many users of such services readily give up their location in exchange for the ability to use the service.

Much of the concern raised in the popular press has been based on the assumption that an individual's address might be directly revealed by location based services. It might be the case that a user of Facebook, FourSquare or other services has already openly published a residential address. However, it is not necessary for an address to be published directly; a map grid reference may also uniquely identify a particular location if given with sufficient precision and accuracy. Krumm (2007) used GPS data from in-car equipment used by volunteers to successfully determine individuals' home addresses 'to within about 60 metres at least half the time' (ibid p.123). Searching in a freely available web database, Krumm was then able to use the estimated addresses to correctly recover a person's identity in a small proportion of cases.

Golle and Partridge (2009) studied the risks of disclosure in US residence to workplace flow data, given a hypothetical location trace based attack key indicating the approximate location of both a person's home and their workplace. They found that at the census block scale, 'the majority' of the working US population could be uniquely identified given home and workplace locations. The census block is the smallest area size used in tabulations released by the US Census Bureau, At the more commonly used census tract scale (typical population of 2500–8000 persons), identification of workplace and home was uniquely identifying for 5% of US workers.

Location Trace Risks in Interaction Data Sets

The Longitudinal Employer Household Dynamics (LEHD) data set used by Golle and Partridge is a commuting data set that includes both a home and a workplace location, and thus is an example of an interaction data set. As described above, a particular characteristic of interaction data sets is that they are often sparsely populated, and contain many small numbers, and thus would seem to have an elevated risk of disclosure from location trace based attacks.

The UK Census outputs feature two main forms of interaction data – commuting data and migration data. The commuting data are generated through a census question which asks for the address of a respondent's usual place of work. The migration data are generated through a question that asks respondents whether their usual address one year prior to the census was the same as their current address. Persons who reported a different residential address one year prior to the census are thus identified as migrants. Similar questions are used to identify recent migrants in many countries' censuses, typically using either a one-year transition period or a five-year transition period. Could location tracing pose a risk to such data sets? In the case of the UK commuting data, this argument is the same as that used previously – that for many people their home and their workplace are the locations at which they spend the most time, and so these could be determined (with varying levels of precision) quite easily. For a census derived commuting data set, the available location trace would have to cover the date on which the census was held. As the trace period diverged from the census date, the confidence that an attacker could place on a presumed re-identification would diminish, due to the possibility that individuals had either changed their usual residence or their place of work. In order to attempt re-identification using the UK migration data set, the attacker would need to have an extended trace which covered both critical dates: the census date, and a date one year earlier.

UK Data Sets

The analysis in this paper is based on sets of outputs from three censuses: the 1991, 2001 and 2011 censuses. These outputs were all subject to different approaches to reducing disclosure risk. The outputs from the 2011 Census are of most interest in the context of risks posed by location tracking, but earlier censuses provide useful insight into the effect of different disclosure control approaches, and can also indicate whether there are significant changes over time in the propensity for unique combinations of origin and destination to exist.

Commuting and migration interaction data sets have been produced as part of the outputs of all three of these censuses: the Special Workplace Statistics (SWS) and the Special Migration Statistics (SMS), respectively. As part of the 2001 outputs, an additional series – the Special Travel Statistics (STS) – were created for residences in Scotland. The STS tables function as a superset of the equivalent SWS tables, as they also include information about school children and students, and their journeys to a place of study, and also include residual counts for persons not in education or employment.

The geographic scope of these varies. The 1991 SWS and SMS data were released for Great Britain, whilst the 2001 and 2011 equivalents were released for the whole of the UK, with per-country variation in coverage for some outputs. The 1991 migration data sets were reviewed in detail by Rees and Duke-Williams (1995), whilst the 2001 data sets have been reviewed by Rees et al. (2002) and Cole et al. (2002), and the 2011 data sets by Duke-Williams et al. (2018).

The 1991 and 2001 sets of outputs collected tables together in 'levels' or 'sets', based on the the reporting geography, with distinct table numbers in each group. The 2011 interaction data were not grouped into 'sets' or 'levels' in the same way as those from the preceding two censuses. Instead, outputs were published at different spatial levels, but with common table identifiers. A broader range of geographies were used for publication than in previous census rounds. Alongside the migration and commuting data, sets of tables were also released relating to students and relating to persons with a second usual residence. In total there were 223 tables published at various spatial scales and with varying levels of access control and attribute detail (ibid). This was a considerable increase over the amount of output from earlier censuses; the 2001 outputs had included a total of 16 migration tables, 14 journey to work tables and 14 'travel' tables, which in turn had been an increase in volume over the 1991 outputs.

Outputs from the three censuses were subject to different forms of statistical disclosure control. The 1991 SMS were subject to a suppression process, in which only limited counts were published for small flows. The structure of the data permitted many of the suppressed counts to be either deduced or estimated (Rees and Duke-Williams 1997). The 1991 SWS had been based on a 10% sample of data, and thus were not subject to additional modification. The 2001 data sets were subject to a process known as Small Cell Adjustment Methodology (SCAM), in which small values in aggregate tables were randomly adjusted. The processes and impact of SCAM were described by Duke-Williams and Stillwell (2007). A different approach was adopted for the outputs from the 2011 Census, with all outputs being assigned a security level: open, safeguarded or secure. Different access and usage restrictions were placed on these. Open data can be used without restriction, safeguarded data require the user to be registered, and secure data require all usage to be done by researchers who have been given Approved Researcher status, who have had a specific project approved, and for analysis to be done in a safe setting.

Methods

The analysis of the risk of disclosure in UK interaction data sets reported in this paper is based on the assumption that a location trace based attack is possible - namely that relatively high-resolution geo-location information may be recoverable by an attacker for some individuals. There are a variety of possible sources, but it is assumed here that the location traces come from smart phone applications that include a user's location amongst their metadata. There are three locations that are significant in the analysis: home location, workplace location and home location at a fixed point in the past. The latter is assumed to be less easily available – it would require longitudinal storage of location data for a particular person (or user account, or device), but risks of disclosure in migration data are included here given that such a dataset remains feasible.

The work described in this paper focuses on aggregate UK interaction data sets. There also exist a number of individual level microdata sets in the UK, both as part of Census data collection, and from other surveys. These might similarly be at risk, even though the geographic coding in them is typically coarse. Individuals who have rare combinations of workplace and residence may still be identifiable even given coarse geographies. The level of risk in microdata sets clearly varies from source to source, depending on the contents of each survey.

Data sets were analysed to examine the number and proportion of flows that formed unique observations within that set. Where there are unique observations of an origin and destination combination, then it is feasible that a location trace based attack could enable attribute disclosure. It is assumed that the location trace dataset contains location information for an identified person, but for whom other characteristics are not necessarily known, and that spatio-temporal clustering allows the location of that person's home x and workplace y to be estimated. Supposing a given table in published census data showed that there was a flow of just one person (with certain tabulated characteristics such as age and sex) between the two locations x and y, then it could be asserted that the identified person in the trace data set was the same as the unique person in the census data set, and that person therefore was now known to have certain attributes included in the census table.

This was done for outputs from the 1991, 2001 and 2011 Censuses. Whilst it is unlikely that the outputs from the 1991 and 2001 Censuses would be subject to a location-based attack, it is pertinent to assess the degree (if any) to which these unique observations have become more or less common over time.

The 2011 outputs were considered at the 'safeguarded' level. It is assumed that the usage and access restrictions placed on the 'secure' data act effectively to prevent systematic abuse. Whilst some access and usage restrictions are placed on the 'safeguarded' data, it is assumed that an attacker would not be concerned with the usage restrictions, and would find the access restrictions relatively easy to overcome.

Assessment of Aggregate Internal Migration Data

The analysis starts by considering the risks associated with publication of internal migration data. The number of unique flows and the number of potential flows were assessed in outputs from the three censuses. In the cases of the 1991 Census, this was straightforward, as a flow table exists that covers the whole of Great Britain. The number of unique flows (combinations of origin and destination) were identified at different spatial scales and compared to the total number of migrants in all flows, in order to determine the proportions of migrants that were at risk of disclosure. Equivalent analysis of the 2001 SMS is complicated by the effects of SCAM: for much of the data, unique records do not exist, as small values (cells counts of one or two) were modified. Data were examined for destinations in Scotland only, as these were not subject to modification. Again, the numbers of flows between unique combinations of origin and destination were identified, and the proportions of all migrants who were at risk was calculated. SCAM was not applied in the 2011 outputs, which thus contain small counts (i.e. cell values of 1 and 2), although usage conditions dictate that researchers should not re-publish these small values as is for data classed as 'safeguarded'.

Analysis of the 2011 SMS was done by identifying the numbers of unique records at different spatial scales on the basis of origin and destination. The 2011 and 2001 data were analysed at district, ward and OA level, whilst the 1991 data were analysed at district and ward level. 'District' level refers to an amalgam of various types of local government administrative units, including London boroughs, metropolitan districts, unitary authorities and Scottish council areas. It should be noted that due to boundary and functional change, the sets of districts, wards and OAs are not identical in any two censuses.

Assessment of Aggregate Workplace Flow Data

Analysis of the 2011 SWS data was done in a similar manner to the approach used for migration data: the number of unique flows in output data sets were identified, where 'unique' refers to a tabulated flow between a given residence and a given workplace consisting of a single person. The SWS outputs tabulate information about persons in employment or self-employed; they are referred to below using the convenient shorthand 'workers'.

The geographies used for analysis of journey to work data were not the same as those used for analysis of migration data. The most coarse level remains the same – district level – but at finer scales different reporting geographies were used. The level below 'district' is referred to in the output table code as MSOA (Middle Layer Super Output Area). MSOAs in practice are a geography used in England and Wales only; the term Intermediate Zone is used for equivalent areas in Scotland, and Super Output Area in Northern Ireland. There are two related sets of results at a more detailed level: output table WF03UKOA reported results for flows between (and within) OAs, whilst WF01UKOA (and most other detailed outputs) reported results for flows between OAs and Workplace Zones (WPZs). WPZs are a geography newly introduced for tabulating the outputs of the 2011 Census (Martin et al. 2013), with the aim of a better spatial representation of employment related statistics than can be done with extant geographies, which reflect the residential distribution of population. The 1991 SWS were based on a 10% sample of data, and were not used for this analysis, as a cell count of one does not necessarily represent a population unique.

The flows detailed in the 2001 SWS and STS data sets were more problematic, as they were affected by the SCAM disclosure control methodology. This was applied to the whole of the SWS, and to OA level STS outputs (but not to the ward or district level STS outputs). Whilst SCAM-affected outputs were ignored in the case of migration data, an estimate was made of the true proportion of unique flows for the journey to work data. This was done using two separate methods.

For 2001 data, the proportion of uniques were first considered using ward and Council Area level 2001 Scottish STS data; these data are the simplest to consider as they were not subject to SCAM. As with the migration data sets, the number of unique records for ward to ward flows were identified, and the proportions of workers in these flows was calculated.

The SWS data that were subject to SCAM were then considered. Firstly, the total number of flows with an observed (post-modification) total of three were identified, and the total proportion of all workers contained within these flows was calculated. The number of flows with a pre-modification total of one were then estimated as follows. A flow frequency table was created using the results in STS Level 2, the most detailed 2001 journey to work data available that were not subject to SCAM. These flow frequencies were then modified subject to the assumed SCAM methodology, in order to derive a new set of frequency totals (with possible totals of zero, three, four and higher). From these, an estimate was made of the ratio of pre-modification totals of one to post-modification totals of three. This ratio was then applied to the observed (post-modification) totals of three in the SWS table, in order to estimate the original number of totals of one.

A subset of the most spatially detailed journey to work data was selected in order to examine the ameliorating effect of aggregation on risk posed by the data. A set of flows from OAs in England and Wales to WPZs in England and Wales were isolated. Flows from Norther Ireland to the rest of the UK were not included as these were presented at a more aggregate level in the original data. Flows to workplaces in Scotland were not included as these used OAs as the reporting geography rather than WPZs. The data were examined with residences at OA level, and then with residences aggregated to MSOA, district and regional level (using the former Government Office Region geography as constituted at the time of the census in 2011), and with workplaces at WPZ level, and then aggregated to MSOA, district and regional level. This gave a total of sixteen combinations of residence and workplace reporting geography. In each case, a flow frequency table was constructed, and anonymity sets were constructed showing k-anonymity values (Sweeney 2002) for cumulative totals of workers. A k-anonymity value is the number of persons who share a particular combination of values. A set of size one means that the person is unique, whilst a set of size 100 means that there are 100 persons in the observed data that share the characteristics. The smaller the anonymity set, the greater the risk of re-identification.

Results

Table 1 shows the proportions of migrants that were in single person flows as shown in published data for three census periods, and at different spatial scales; the table gives a source (data set and period, table name where relevant), the numbers of origins and destinations, the total count of persons tabulated as making moves between those origins and destinations, and the proportion of persons in unique flows. A unique flow occurred where the flow between a given origin-destination pair consisted of only one person. Results are reported for variant geographies, dictated by the structure of the overall data set. Thus, the results for 1991 refer to migrants within Great Britain (the wider data set also included flows from overseas, but these are not considered here) and the results for 2011 refer to migrants within the UK. The 2001 data sets were issued with a UK scope, but flows to destinations other than those in Scotland were subject to a form of disclosure control that makes the present analysis unworkable.

The 2011 results show that for the most spatially detailed outputs – those reported at OA level – 36.7% of persons had unique combinations of origin and destination. It is these persons who may be at risk of attribute disclosure in comparison to an alternative location-trace based data set. It should be noted that this is not a proportion of *all* persons, rather it is a proportion of those persons defined as migrants given the census definition of a migrant as being someone with a change in usual residence in the year preceding the census; this figure represents about 4% of all persons in the wider population.

OAs are very small units – there were 232,296 defined in the UK for the 2011 Census, with a mean population of around 270 persons. When data are reported at a more coarse level, the proportion of unique flows falls: for flows at ward level, 11% of persons had a unique combination of origin and destination, whilst for the district level results less than 0.3% of persons were in unique flows. The extent to which any of these observations reflects an actual risk depends on the volume of published data and scope for attribute disclosure; this is considered in the discussion below.

The 2001 results are based on flows within Scotland only, but show fairly similar results, with 34.2% of migrants reported at the most detailed level being in unique flows. The 1991 outputs did not feature such a fine level of reporting; the most detailed

Table 1 Proportions of migrants in unique flows in UK census migration statistics

Source	Spatial scale	Origins x destinations	Total persons	Unique flows
1991 SMS Set 2	District	459 × 459	4,688,180	0.6%
1991 SMS Set 1	Ward	10,933 × 10,933	4,688,180	12.5%
2001 SMS Level 1[a]	District	32 × 32	473,789	<0.01%
2001 SMS Level 2[a]	Ward	1176 × 1176	473,789	8.4%
2001 SMS Level 3[a]	OA	42,604 × 42,604	473,789	34.2%
2011 Table MM01AUKLA	District	404 × 404	6,815,401	0.3%
2011 Table MM01AUKWARD	Ward	9505 × 9505	6,815,401	11.0%%
2011 Table MF01OA	OA	232,296 × 232,296	6,815,401	36.7%%

[a] For flows within Scotland only

results were given at ward level, and showed 12.5% of migrants to be tabulated in unique flows.

Table 2 shows results for similar analysis of journey to work data from the 2011 Census, and for part of the 2001 outputs. Equivalent data for 1991 are not available, as the relevant outputs were based only on a 10% sample of data. Results for 2001, as with the migration data, are affected by the disclosure control procedure used at the time, and are discussed below rather than shown in the table. The journey to work data are shown in the table with four levels of geography. The most coarse is the district level; 0.1% of persons in the reported had a journey to work with a unique combination of origin (usual residence) and destination (workplace) at this level. As with the migration data, it should be stressed that this is not 0.1% of all persons, but rather 0.1% of workers (as defined above). The number of workers is a substantially larger fraction of the total population than was the case for migrants: about 37% of the total population were employed or self-employed.

At the MSOA level, 5.3% of workers had journeys to work with a unique combination of residence and workplace. The majority of the 2011 outputs at the most spatially detailed level showed flows from OAs to WPZs, thus the geography is asymmetric. An additional table was created showing headcount data at OA to OA level, in order to facilitate change over time comparisons with 2001. The utility of this is limited due to the problems arising in the 2001 data from the SCAM adjustment methodology. WPZs were only defined for England and Wales at the time of data release, and thus OAs were used in place of WPZs outside England and Wales. In order to facilitate a comparison of the effectiveness of the WPZ geography at protecting flow data, analysis was done for England and Wales only, counting uniques in the OA to OA data and the OA to WPZ data. Both sets have high proportions of uniques: 53.2% of workers were in unique flows in the case of the OA to OA flows, and 58.1% of workers were in unique flows when tabulated using the OA to WPZ table. The number of WPZs is lower than the number of OAs, and this data is allocated between cells in the output table more evenly: using results not shown in Table 2, in the case of the OA-OA flows, 0.04% of the entire origin-destination matrix had non-zero values, whilst 0.16% of the OA-WPZ matrix had non-zero values – both are highly sparse, but the sparsity was less extreme with the OA-WPZ data. The Table also shows the proportions of persons in the 2001 STS Levels 1 and 2 who were contained in single person flows. These data were

Table 2 Proportions of workers in unique flows in UK 2001 and 2011 census workplace statistics

Source	Spatial scale	Origins x destinations	Total persons	Unique flows
2011 Table WU04AUKLA	District	404 × 404	24,240,505	0.1%
2011 Table WU01UKMSOA	MSOA	9326 × 9326	24,240,505	5.3%
2011 Table WF03UKOA[a]	OA x OA	181,408 × 181,408	21,625,060	53.2%
2011 Table WF03UKOA[a]	OA x WPZ	181,408 × 53,578	21,625,060	58.1%
2001 STS Level 1	District	32 × 32	2,238,844	<0.01%
2001 STS Level 2	Ward	1176 × 1176	2,238,844	3.1%

[a] For flows within England and Wales only

not subject to SCAM modification, and thus it is easy to determine the proportions of unique flows.

As described in the methods, similar analysis could not be carried out directly for the 2001 SWS due to SCAM adjustments. Table 3 shows the total number of persons in flows with a reported value of three, for three spatial scales. At all levels these flows will be an aggregate of: those flows with a genuine total of three, some flows with an original total of two, and (to a lesser extent) some flows with an original value of one. It is the latter component only that is of interest in this paper. The fifth column of the table lists the number of observed flows with a value of three, whilst the sixth column shows an estimate of the actual number of flows with an pre-modification total of one. These estimates shows broadly similar observations of the proportions of workers in unique flows as observed elsewhere, with differences perhaps more likely to reflect the simple estimation methodology than other factors.

Figure 1 shows the flow frequency observations and derived information for one set of flow data: the 2011 journey to work results, at OA to WPZ level, for residences and workplaces in England and Wales. The dotted line shows the number of observations in the data of flows of a given size, whilst the short-dashed line shows the number of workers accounted for by flows of that size. The solid line shows the cumulative proportion of all workers who are observed in flows up to the given size, and the long-dashed line shows the cumulative proportion of observed flows. Thus, for flows of size one (only one person observed for a fixed origin-destination pair) there were around 12.5 million flows, containing 12.5 million persons, and for flows of size two, there were around 2.07 million flows, accounting for 4.15 million people, and so on. The flows of size one accounted for 58% of workers, and 80% of all observed (non-zero) flows.

The solid line showing the cumulative proportion of all observed persons can be referred to as the anonymity set under the terminology used by Sweeney (2002). The larger the anonymity set – which is indicated on the logarithmic *x*-axis – the greater the privacy for individuals in that set, as there are more people who share their characteristics. Thus, in Fig. 1, it can be seen that 86% of workers have an anonymity set of 3 or fewer persons, and 98% have an anonymity set of 10 or fewer persons. The size of the anonymity set in this data is dependent on the level of spatial aggregation, and this can be considered in terms of both the residence (or origin) and the workplace (or destination). Considering the aggregation of these separately from each other may be

Table 3 Proportions of migrants in unique flows in UK 2001 census workplace statistics

Source	Spatial scale	Origins x destinations	Total persons	Observed flows with persons = 3	Estimated flows with persons = 1	Estimated unique flows
SWS Level 1[a]	District	394 × 394	24,205,298	27,696	24,096	0.1%
SWS Level 2[a]	Ward	9432 × 9432	24,312,420	738,210	642,243	2.6%
SWS Level 3[b]	OA	175,434 × 175,434	23,944,043	5,058,822	4,401,175	18.4%

[a] For flows in England and Wales

[b] For flows in UK excluding Scotland

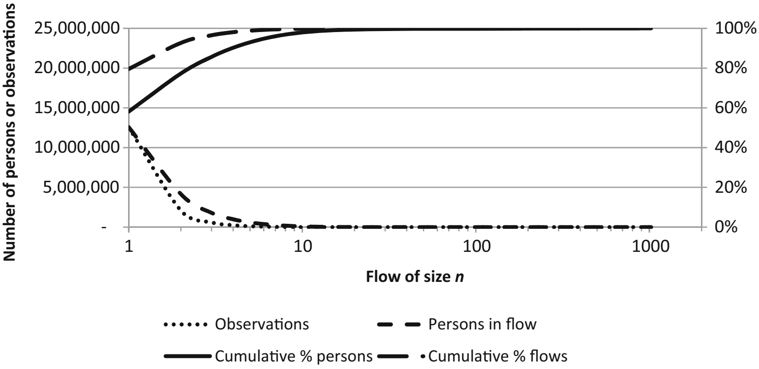

Fig. 1 Flow frequency observations, 2011 output area to workplace zone journey to work data

pertinent if location trace data were to have different levels of confidence associated with the accuracy with which home or workplace locations could be estimated. Figure 2 shows a set of re-drawn anonymity curves taken from the same data. In each case, the flow size (or anonymity set) is shown on the y-axis, whilst the proportion of all persons is shown on the x-axis.

The Figure is presented in the form of a matrix of anonymity curves, arranged by the level of spatial aggregation of residential geography (rows) by aggregation of

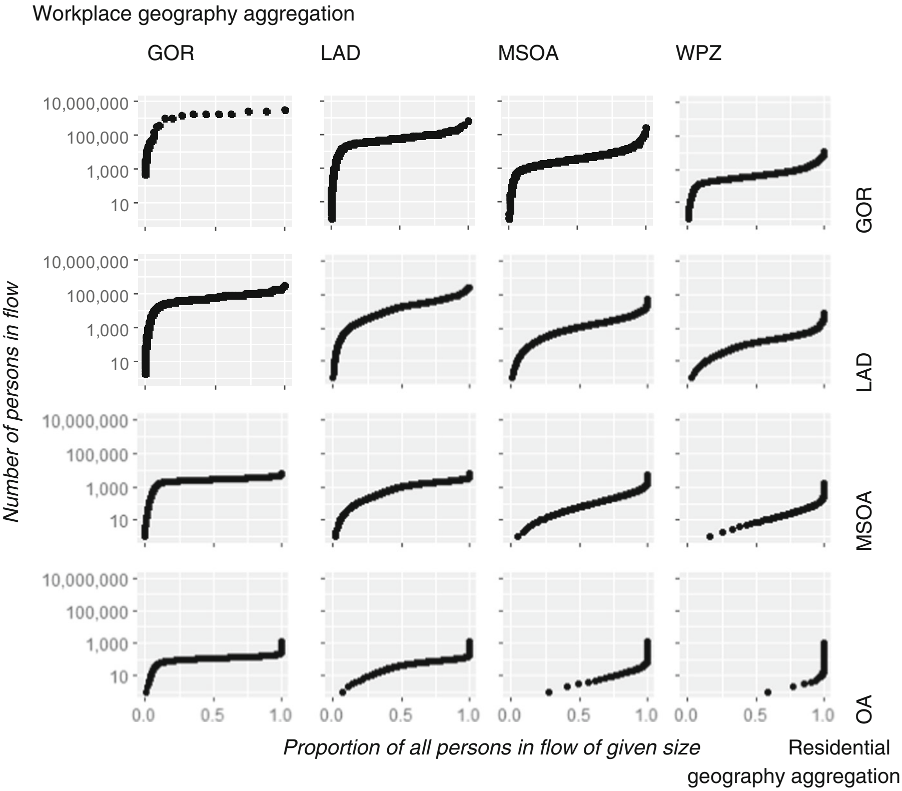

Fig. 2 Anonymity sets for 2011 journey to work data with variable aggregation of home and workplace

workplace geography (columns). The least aggregated view is seen in the lower right hand panel, whilst the most heavily aggregated is seen in the top left panel. All combinations of residence and workplace geography apart from two (district to region and region to region) featured some workers in anonymity sets of one, with proportions being shown in the summary Table 4.

Discussion of Results and Conclusions

These results show, unsurprisingly, that risk is dependent on the reporting geography, and the type of data under consideration. It can be seen from the migration data results (Table 1) that few migrants were in unique flows for data reported at district level, but larger proportions were so at ward level and especially at Output Area level. The same is true of journey-to-work data (Table 2), albeit with more people involved (a larger fraction of a bigger group of people). The number of persons at risk (through being in unique flows) in the journey-to-work data for England and Wales at the MSOA level was just over 5%, a very similar figure to that observed (5%) by Golle and Partridge (2009) for US journey-to-work data viewed at the census tract level. MSOAs in England and Wales are larger than census tracts in population terms with an average population of around 7800, whereas the census tracts used by Golle and Partridge had a mean residential population of around 1600. The set of anonymity curves (Fig. 2 and Table 4) provide further context about the relationship between area size (or population) and the number of distinct flows: unique combinations of residence and workplaces occurred at almost all scales. The level of interaction uniqueness given reporting units is one that is hard to compare between countries, given limitations on available data. However, there may be analogies with the use of Courgeau's k to measure the relationship of migration intensities to the number of reporting units (Courgeau 1973) and thus for international comparison (Bell et al. 2002).

Further characteristics of specific data sets (beyond area population) may serve to reduce the risk of identification. In the case of the 1991 SMS, there was an additional classification for migrants of 'origin unstated'; this included persons who indicated on the census form that they had not been living at the same address one year previously, but had not supplied a usable former address. At a national level, the proportion of all migrants (including those from overseas) who had an unstated origin was 13%; there

Table 4 Proportion of workers in flows with anonymity sets of size one, England and Wales, 2011 census

Residential aggregation	Workplace aggregation			
	Region	District	MSOA	Workplace zone
Region	0%[a]	<0.1%	<0.1%	0.3%
District	0%[b]	0.1%	0.9%	3%
MSOA	<0.1%	1%	5%	16%
Output Area	1%	7%	27%	58%

[a] Minimum anonymity set 460 persons

[b] Minimum anonymity set 2 persons

was considerable regional variation. At a district level, the proportion of migrants with an unstated origin ranged from 3.6% (Isles of Scilly) to 27.9% (Liverpool). If an attack data set contained a 'known' origin and destination, it may be the case a migrant with a known attack origin was in fact recorded in the census as having an unstated origin, introducing a degree of ambiguity into any claim of identification. More recent migration and commuting data have applied imputation techniques to replace missing origins and workplaces (Stillwell and Duke-Williams 2007), yet this might still be offered as a form of disclosure control, through reduced confidence in stating a particular observation to be unique.

Assuming that observed flows of one person are robust, the question then arises of whether this may pose a risk in practice. In the case of the most spatially detailed migration data, there are no additional data made available at in open or safeguarded modes that further tabulate this flow. Thus, given an assumption that one could identify a person given an estimated origin and destination, all the census data would allow an attacker to do would be to confirm that there was also an observation of a matching flow in the census. No further information would be revealed without access to the 'secure' data sets, to which access is restricted, and from which extracts cannot be retained without clearance. For the most spatially detailed journey to work data (which are assumed to be much more easily attackable with location trace data), additional attribute data are published at the OA-WPZ level in the form of a table showing mode of transport to work, in a very broad-coded version. The data publication strategy explicitly trades off scale of reporting geography and the amount of attribute data available. Table 5 summarises this for migration and journey to work data from the 2011 census, showing the range of tables made available at different spatial scales. Whilst limited data are available at the finest scales, a range of data are available at larger ward and MSOA levels. For migration data, there are two tables at ward level (a univariate age table, and an age by sex table), whilst for journey to work data, there are 12 univariate tables.

The distinction between univariate and multivariate tables is not relevant in the case of unique flows: it is clear that when only a single person moves between a residence and a workplace, all of the separate univariate classes in which that flow is tabulated relate to the same person. Thus, if a location trace dataset can be used to estimate residence and workplace to MSOA or finer level for a person, then 12 additional characteristics (to varying degrees of precision) of that person could be determined. The median size of MSOAs in England and Wales is around 3.18 km^2. Additional requirements of users are in place to protect these data, although it is assumed that an attacker wishing to breach privacy would not be concerned about terms and conditions of usage.

Here, we might be moved to ask: 'what is an acceptable level of risk?' Of those persons tabulated in the MSOA level journey to work data, 5.3% (Table 2) were in unique flows. Thus, were an attacker intending to demonstrate that for some people attribute disclosure could occur, then the risk would be tangible. Were the attacker wishing to achieve a more general acquisition of data for large numbers of people, then the opportunity would be constrained. More risk is potentially posed by the 'secure' interaction data. For the journey to work data, 58% of persons were recorded in unique flows at the most spatially detailed level. A further 3 univariate and 9 multivariate tables have been published at this level, providing scope for considerable attribute

Table 5 Cross-classifying attributes in migration and journey to work tables, 2011 census

Migration tables	Journey to work tables
District level	District level
U01: Age (single year)	U01[a]: Sex
U02: Family status	U02[a]: Age (broad group)
U03: Long term illness / disability	U03[a]: Mode of transport
U04: Economic activity	U04: Economic activity
U05: NSSeC (detailed)	U05: NSSeC
U06: Tenure	U06: Industry
U07: General health	U07: Occupation
U08: Industry	U08: Family status
	U09: Car/van availability
	U10: Hours worked
	U11: Social grade
	U12: Country of birth
Ward level	MSOA level
U01: Age (grouped)	U01: Sex
	U02: Age (broad group)
	U03: Mode of transport
	U04: Economic activity
	U05: NSSeC
	U06: Industry
	U07: Occupation
	U08: Family status
	U09: Car/van availability
	U10: Hours worked
	U11: Social grade
	U12: Country of birth
Output Area level	Output Area / Workplace zone level
n/a	U03: Mode of transport (grouped)

[a] Tables released as open data (all others released as safeguarded)

disclosure for more than half of the people included in the data. However, much stronger protections surround those data – access must take place with a secure environment, and data or notes cannot leave that environment without clearance procedures.

An obvious safety measure against location based attacks is to make the reporting geography more coarse, so that each separate flow tends toward a higher total. However, clearly this also makes the data less suitable for detailed spatial analysis. As demonstrated in Fig. 2, it is not possible to remove all risk without such coarsening of both origin and destination that no analysis could be done at a local level. The spatially detailed structure of the interaction data is intended to permit close analysis: whilst one is unlikely to want to study flows at an Output Area level – or, in the US case, at census block or tract level – the publication of detailed flows permits flexible reconstruction to any desired reporting geography.

The analysis carried out in this paper has confirmed earlier work suggesting that there is a specific risk in commuting data sets. Unlike the US example with LEHD data, the UK interaction data sets can be acquired by anyone in raw form. The paper has highlighted a more general risk for all interaction data, including migration data. For many of the data sets examined in this paper, the actual risk is minimal because of the age of the data, and there is no need to withdraw any such data sets. For the 2011 outputs, there is some evidence of risk for a limited subset of those persons in the published data. However, the generic risk remains and will apply to future interaction data outputs. One can thus return to the initial questions of whether these data now pose an unacceptable risk, and whether their release should be constrained in the future. Clearly, location based services represent a genie which is unlikely to be placed back in a bottle: location tracing seems likely to become more common rather than less common in the future; if something has to change, it may be data release strategies. One possible solution lies in the use of asymmetric data sets: these might offer a sensible hybrid of reduced risk but retained worth. The 2011 Census interaction data outputs address the issue by placing more restrictive access constraints on the data. Future study of usage levels may be instructive as to whether this approach is effective for enabling analysis of data whilst at the same time protecting it.

References

Allan, A., & Wardle, P. (2011). iPhone tracking "what your iPhone knows about you", Where 2.0 Conference, April 19–21 2011 Santa Clara CA, http://where2conf.com/where2011/public/schedule/detail/20340.

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., & Hugo, G. (2002). Cross-national comparison of internal migration: issues and measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 165*(3), 435–464.

Cole, K., Frost, M., & Thomas, F. (2002). Workplace data from the census. In P. Rees, D. Martin, & P. Williamson (Eds.), *The census data system* (pp. 269–280). Chichester: Wiley.

Courgeau, D. (1973). Migrations et découpages du territoire. *Population, 28*, 511–537.

Duke-Williams, O., & Stillwell, J. (2007). Investigating the potential effects of small cell adjustment on interaction data from the 2001 census. *Environment and Planning A, 39*(5), 1079–1100.

Duke-Williams, O., Routsis, V., & Stillwell, J. (2018). Census interaction data and the means of access. In J. Stillwell (Ed.), *The Routledge handbook of census resources, methods and applications unlocking the UK 2011 census*. New York: Routledge. Forthcoming.

Fellegi, I. (1972). On the question of statistical confidentiality. *Journal of the American Statistical Association, 67*(337), 7–18.

Golle, P. (2006). Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society* (pp. 77–80). ACM.

Golle, P., & Partridge, K. (2009). On the anonymity of home/work location pairs. In: *Pervasive computing* (pp. 390-397).

Krumm, J. (2007). Inference attacks on location tracks. In *Pervasive Computing* (pp. 127–143).

Küpper, A. (2005). Location-based services: Fundamentals and operation. Wiley.

Levinson, A., Stackpole, B., & Johnson, D. (2011). Third party application forensics on apple mobile devices. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on* (pp. 1-9). IEEE.

Martin, D., Cockings, S., & Harfoot, A. (2013). Development of a geographical framework for census workplace data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(2), 585–602.

Poynter, K. (2008). Review of information security at HM revenue and customs, final report, June. Available at: www.hm-treasury.gov.uk/media/0/1/poynter_review250608.pdf.

Rees P. & Duke-Williams O. (1995). The story of the British special migration statistics. Scottish Geographical Magazine, 111(1), 13–26.

Rees, P. H., & Duke-Williams, O. (1997). Methods for estimating missing data on migrants in the 1991 British census. *Population, Space and Place, 3*(4), 323–368.

Rees, P., Thomas, F., & Duke-Williams, O. (2002). Migration data from the census. In P. Rees, D. Martin, & P. Williamson (Eds.), *The census data system* (pp. 245–267). Chichester: Wiley.

Singer, E., Mathiowetz, N. A., & Couper, M. P. (1993). The impact of privacy and confidentiality concerns on survey participation the case of the 1990 US census. *Public Opinion Quarterly, 57*(4), 465–482.

Stillwell & Duke-Williams. (2007). Understanding the 2001 UK census migration and commuting data: the effect of small cell adjustment and problems of comparison with 1991. *Journal of the Royal Statistical Society Series A, 170*(2), 425–455.

Sweeney, L. (2000). Uniqueness of simple demographics in the U.S. Population. Laboratory for International Data Privacy, Carnegie Mellon University, Pittsburgh.

Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), 557–570.

The Guardian. (2010). *People worry about over-sharing location from mobiles, study finds.* http://www.guardian.co.uk/technology/blog/2010/jul/12/geolocation-foursquare-gowalla-privacy-concerns.

Willenborg, L., & De Waal, T. (2012). *Elements of statistical disclosure control* (Vol. 155). Springer Science & Business Media.