TRANSACTIONS ON DATA PRIVACY 13 (2020) 151–170

Confidentiality challenges in releasing longitudinally linked data

Robin Mitra[1*], Stephanie Blanchard[2], Iain Dove[2], Caroline Tudor[2], Keith Spicer[2]
[1] *Department of Mathematics and Statistics, Lancaster University, UK*
[2] *Office for National Statistics, Hants. UK*
[*] *E-mail: R.Mitra@lancaster.ac.uk*

Received 6 August 2019; received in revised form 13 February 2020; accepted 8 April 2020

Abstract. Longitudinally linked household data allows researchers to analyse trends over time as well as on a cross-sectional level. Such analysis requires households to be linked across waves, but this increases the possibility of disclosure risks. We focus on an inter-wave disclosure risk specific to such data sets where intruders can make use of intimate knowledge gained about the household in one wave to learn new sensitive information about the household in future waves. We consider a specific way this risk could occur when households split in one wave, so an individual has left the household, and illustrate this risk using the Wealth and Assets survey. We also show that simply removing the links between waves may be insufficient to adequately protect confidentiality. To mitigate this risk we investigate two statistical disclosure control methods, perturbation and synthesis, that alter sensitive information on these households in the current wave. In this way no new sensitive information will be disclosed to these individuals, while utility should be largely preserved provided the SDC measures are applied appropriately.

Keywords: Data confidentiality, Disclosure risk, Matching, Perturbation, Propensity score, Synthetic data

Acknowledgement

The majority of this work was supported through a research contract between the Office for National Statistics and the University of Southampton. The authors would also like to acknowledge Professor Reiter for supplying code to facilitate implementation of the CART synthesis method.

1 Introduction

Data-holding agencies need to protect the confidentiality of survey microdata prior to releasing it to analysts while also ensuring released data is of a sufficiently high utility to analysts. One approach is to restrict analysis of the data to secure centres. However, such an approach is often cumbersome for many analysts who would need to spend considerable time and effort to access the data. Another popular approach is to release an accessible version of the data set with values altered sufficiently to protect confidentiality. This area is known as Statistical Disclosure Control (SDC) (Willenborg and de Waal, 2001; Hundepool *et al.*, 2012). Common SDC methods include recoding variables to be released at a lower level of detail, adding random noise with zero mean to certain variables' values, and perturbing information (swapping) between different records in the data.

The literature on SDC has been largely focused on cross-sectional data sets, and the understanding of risks and methods to mitigate these are well developed here. However, there has been relatively little work on understanding risks with longitudinally linked data which poses additional potential disclosure risks to the usual risks present in cross-sectional data. Even when SDC methods have been applied to longitudinal data, the focus has been on applying cross-sectional methods to measure risk, which may underestimate the risk present in the released data. This article focuses on the risk specifically associated with such a data set and considers two methods, perturbation (swapping) and synthesis, to mitigate this risk. We are thus considering a new specific type of risk: dealing with disclosure risks specific to longitudinally linked data.

Longitudinal linked household survey data comprise data that is collected on households over repeated time points; each time point is often referred to as a wave. These data sets are an important resource in social science research and related fields (Frees, 2004). They allow researchers to analyse trends over time as well as on a cross-sectional level, and thus enables more complex questions to be addressed. Developments in software packages also facilitate researchers in performing such analysis with minimal extra burden. There are numerous examples of analysis of such data in social science research (Frees, 2004; Allen and Meyer, 1990; Steinfield *et al.*, 2008) while examples can be found in other fields such as psychology (Velez *et al.*, 1989) and medicine (Von Korff *et al.*, 1992).

In order for researchers to perform such analysis, households must be linked across waves, but this will increase the possibility of disclosure risks as more information is being released about each household than if they were not linked. We focus on a risk specifically present in longitudinal data when an intruder can make use of intimate knowledge gained about the household in one wave to identify the household in that wave, and subsequently learn new sensitive information about the household in future waves. We denote such a risk as the inter-wave risk.

If the data are linked then identifying the household in future waves is straightforward, and the inter-wave risk is apparent. One option might be to remove the links between different waves altogether, but the inter-wave risk may still be present as the intruder could use the socio-demographic information present on the households, which is largely static over different waves, and apply record linkage techniques (Fellegi and Sunter, 1969) to identify the household in future waves. In addition, removing the links will not allow researchers to analyse longitudinal trends at the same level of detail as otherwise would be possible, and thus the utility of such data will be substantially reduced.

After identifying the households susceptible to this inter-wave risk, we investigate two approaches to mitigate the risk. One approach we investigate is a matching scheme that perturbs or swaps sensitive information on these households in the current wave with in-

formation from other similar households. In this way no new sensitive information will be disclosed to these individuals, while the utility of the data should be largely preserved as information is being swapped between similar households. Another approach we investigate is to impute or synthesize the sensitive information on these household from a model fit to the data. Again no sensitive information should be disclosed, and provided the synthesis model used is a plausible model then the synthetic data set should reflect the relationships present in the original data. These approaches also benefit from being a general approach that can be applied to any longitudinal household data set affected by this kind of disclosure risk. We illustrate performance of these approaches on the Wealth and Assets Survey (WAS). The WAS is a relevant important survey that contains information on many sensitive variables and is susceptible to this risk of disclosure. The survey is longitudinal in nature with six waves of microdata already published, and is thus ideal for use in our illustration.

The article is organised as follows. Section 2 gives some background to the problem area being considered and illustrates the inter-wave risk with a specific example when a household splits in one wave. Section 3 considers the specific inter-wave risk in more details, and illustrates the relevance of this risk in the WAS. Section 4 describes the proposed SDC methods to mitigate this risk. Section 5 illustrates the performance of these methods when applied to the WAS. Section 6 finishes with some concluding remarks.

2 Addressing the risk inherent in longitudinal data

Increasingly, there is a growing demand and availability of longitudinal data sources for analysis. It is thus important that the disclosure risks associated are well understood and mitigated appropriately. Clearly, there are additional risks inherent in the release of longitudinal data that are not present in the release of cross-sectional data. This is recognised in the Government Statistical Service guidelines for disclosure control of microdata which states that if data are being released with households linked over time then additional modification would be required (Office for National Statistics, 2014). As noted in the guidelines, changes to a household structure over time could potentially be cross checked with information in the public domain increasing the chance of the household being identified.

We can consider a specific example of the inter-wave risk present in a longitudinally linked household data set, when an individual leaves a household from one wave to the next. Such split households could be self-identified by the individual that has left using all the available information reported on the household from the previous waves (when it was part of the household), and using the fact the household is linked with the current wave, it could then learn new information on this household. Henceforth, we will refer to these individuals as intruders. Figure 1 illustrates this risk scenario with a particular example. Here the individual that leaves household A in wave t (now the intruder in wave t+1) can use the information it knows about household A in wave t (when it was part of that household), to identify the household, and then use the links to learn new information about the household in wave $t + 1$, for example information about a new partner joining the household in wave $t + 1$.

It is important to recognise that the inter-wave risk we investigate here is not the standard self-identification risk, of an individual identifying themselves in the survey, which some argue is not a confidentiality risk. The difference between this risk and the self-identification risk is that the individual will possess only partial information on the household it was *previously* part of in later waves. As such, it will be harder for the individual to

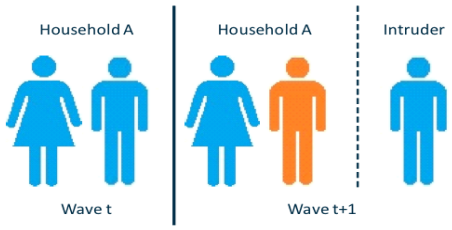

Figure 1: Ex-partner (now the intruder) can deduce new information about household A using information from wave t and the links

identify the household in later waves but there still might be a high chance of identification due to the intimate knowledge it possesses on the household. The risk is also specific to the longitudinal nature of the data set, and not something that would be of concern in a cross-sectional data set, but still very relevant in light of concerns raised by statistical agencies. To our knowledge, there has been no assessment of this type of risk when releasing longitudinally linked data.

It is also important to note that this is not the only way an inter-wave risk could arise in longitudinally linked data. Another type of inter-wave risk could be when a new household moves into the address of a household that has been taking part in the longitudinal survey, and they learn the previous occupants were in the survey. The new occupants could then use this fact, together with information it is likely to know about the previous occupants, to identify this household in the survey, and again potentially track the household in later waves. This is not an exhaustive list of all types of additional risk present when releasing longitudinally link data, and a full assessment of the specific risks associated with longitudinally linked data would form part of important future research into this area. For the remainder of this article the inter-wave risk considered will be the risk arising from households splitting as described above.

3 Inter-wave disclosure risk from split households

Suppose that our longitudinal data set at wave $t \in \{1, \ldots, T\}$ comprises two parts $X = (x_1, x_2, \ldots x_n)$ and $Y^{(t)} = (y_1^{(t)}, \ldots, y_n^{(t)})$ where $(x_i, y_i^{(t)})$ corresponds to information on household i in the data set, $i = 1, \ldots, n$. Variables in X represent those that are typically static across waves such as a measure of geography, or the number of bedrooms in the household. These variables may also include standard socio-demographic variables such as occupation or marital status but are not limited to this set. Variables in $Y^{(t)}$ represent the information on households that are likely to change from wave to wave, some examples include income or house price. For simplicity we assume that variables in X are the key variables that would be known to an intruder in any wave. The intruder may also possess knowledge on some variables in $Y^{(t)}$, in any wave, due to the intimate knowledge it possesses on the household. For example the number of children in the household, which may change across waves, could also be considered to be a key variable. To avoid complicating the notation we will refer to the set of all key variables (time constant or otherwise) by X, with the remainder of the data being referred to by $Y^{(t)}$. We also assume that the data set includes a unique household identifier $w^{(t)} = (w_1^{(t)}, \ldots, w_n^{(t)})$ that is used to link households across waves.

The risk we are addressing in this article occurs when a particular household i splits. Specifically an individual leaves household i from waves $T-1$ to T. Such an individual could be assumed to possess information on that particular household comprising $(x_i, y_i^{(1)}, \ldots, y_i^{(T-1)})$ and as such would more than likely be able to uniquely identify the household in the data, and due to the presence of the linking identifier, $w_i^{(T)}$, would be able to learn new sensitive information about the household contained in $y_i^{(T)}$. In the WAS, split households apply to approximately 10% of the households in wave 2.

3.1 Assessing longitudinal risks via record linkage and modelling intruder behaviour illustrated using the WAS

An illustration of the potential risks present in this scenario can be seen by considering the WAS household level data. This is a longitudinal survey that aims to improve understanding about the economic well-being of households. The survey collects data on levels of assets, savings and debt; saving for retirement; how wealth is distributed among households or individuals; and factors that affect financial planning. Many of these variables could be viewed as containing sensitive information and be highly disclosive. The survey also collects the usual socio-demographic information and details on the structure of the household. The survey is considered to be an important resource with currently six waves of microdata published and approval given for future waves to be collected. More details are available about this survey on the Office for National Statistics website [1] and a survey review report has also been published (Office for National Statistics, 2012). The WAS is an important resource for both academics and government departments. As the survey is relatively new, currently government analysts have produced the majority of published output, although a research article has been published that makes use of the WAS (Daniel and Bright, 2011) and we anticipate more will become available in the future. The WAS is available under End User Licence or Special Licence http://www.data-archive.ac.uk/conditions/data-access. The Special Licence version can only be accessed by 'Approved Researchers' while the End User Licence requires a user to register and sign up to some simple terms and conditions. The most detail in the variables can be found in the Special Licence version.

The End User Licence version of the WAS currently comprises six waves, although for simplicity we will only consider the first two waves here, so $T = 2$. In wave 1 there were 30587 households while in wave 2 there were 20165 households. This is to be expected due to attrition, and to keep the framework simple we assume that a household j dropping out after wave t will have "No Response" recorded for all $y_j^{(t+1)}, \ldots, y_j^{(T)}$. There were 1486 households observed to have split in wave 2. There were 499 variables recorded in wave 1 and 537 variables recorded in wave 2. The WAS is thus a complex longitudinal survey data set and typical of the type of data sets we would like to consider in general.

To keep things simple we assume two scenarios concerning X, i.e. the information the intruder might possess and use to make a disclosure. These two scenarios represent different levels of intruder knowledge that might be expected.

- In the first scenario, X comprises only variables recorded in wave 1 on output area code (a geographic variables with 53 categories), type of accommodation (a variable with three categories), and number of bedrooms in the household (a variable with

[1] http://www.ons.gov.uk/ons/guide-method/method-quality/specific/economy/wealth-and-assets-survey/wealth-and-assets-survey—user-guidance/index.html

10 categories), while $Y^{(1)}, Y^{(2)}$ comprise just one variable that measures total wealth. The variables in X were chosen as they remain largely unchanged in their values across both waves.

- In the second scenario $Y^{(1)}, Y^{(2)}$ remains the same while X additionally includes the number of children in the household recorded in wave 2 (a variable with 7 categories) and a measure of household type in wave 2 (a variable with 10 categories).

The presence of the continuous variable in $Y^{(1)}$ means that the intruder is quite likely to be able to uniquely identify its household in the wave 1 data, as these households are largely uniquely determined by these values (there are 1478 unique $Y^{(1)}$ values). We thus assume that intruders are able to uniquely identify their household in wave 1. We note that there were a few split households in wave 2 that linked back to the same household in wave 1, this is due to the nature of the survey which tries to follow up respondents rather than a fixed location. In these situations either respondent could be viewed as an intruder. For now we only focus on those households with a unique link between wave 1 and 2, i.e. where only one household was followed up after the split, as this comprised 87.7% of the split households, but note that it is relatively straightforward to handle the non-unique links in practice.

Clearly keeping the links here results in a very high inter-wave disclosure risk, with all intruders able to learn about their target household's income in wave 2. However, simply removing the links here does not necessarily help mitigate the disclosure risk. We can see that by applying a naive but intuitive record linkage/matching approach for these 1486 split households. Specifically, for a given household i, we assume the intruder has identified this household in wave 1 (due to the knowledge it possesses on the household's total wealth in wave 1). The intruder then uses the information it possesses i.e. the values in x_i for this particular household to find all the split households with exact matching information in X in the corresponding wave 2 data, suppose there are c_i such matches. If $c_i > 0$, and the correct household is one of the correct matches in the given set, the associated risk of identification is $1/c_i$. Figure 2 presents an illustration where X comprises area and number of bedrooms for a particular split household, and we can see that if the link is removed $c_i = 3$.

We apply this procedure to each split household in turn in the WAS. In scenario 1, out of the 1486 split households, 1282 households have the potential to be correctly identified here, i.e. $c_i > 0$ for 1282 households and the correct household is in the set of matches. Amongst these households, 121 households have a probability of being identified of 0.5 or greater (i.e. $c_i \leq 2$) and 56 household have an identification probability of 1 ($c_i = 1$). As expected, in scenario 2 the risks increase. As before 1282 households have the potential to be correctly identified here, but now 871 households have an identification probability of at least 0.5, with 563 households being identified with probability 1.

We note that there are numerous methods we could consider to link households across the waves. We could instead consider a discrepancy measure, where for example only those households that match on a certain number of variables in X will be considered as matches, rather than requiring exact matches. In simulations not reported here, we implemented such a procedure but found risks were not as high as those reported above in both scenarios, although in other scenarios and data sets it may well be that this discrepancy measure is more appropriate. We also note that sophisticated probabilistic record linkage techniques have been proposed in the literature (Fellegi and Sunter, 1969) and could potentially yield higher risks if such procedures were implemented. Investigating different

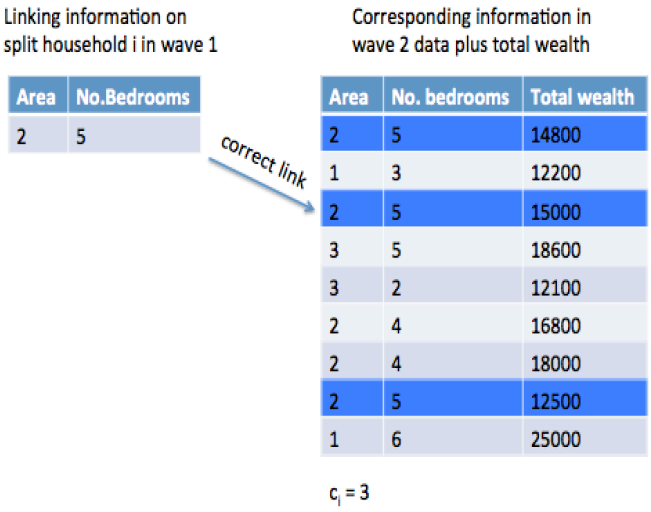

Figure 2: Illustrating how an intruder can identify a split household in wave 2. If the link is removed, a split household can be matched to three potential candidate households (one of which is the correct household)

ways to link households longitudinally here is a topic for future investigation and is not considered any further here.

As we are not seeking to prevent intruders from identifying their household in the data (in either wave), but instead seeking to limit the amount of new information the intruder can learn about the household, we might like to consider how the intruder would go about learning new information. We assume in this article that the intruder might behave in one of two ways depending on the amount of information released.

Firstly, if the links are supplied, the intruder will find the total wealth corresponding to the household it identifies in wave 2 using the household it has identified in wave 1 and the link provided to its information in wave 2. As noted earlier, without any alteration to the total wealth variable in wealth 2 this will result in a very high inter-wave disclosure risk. In Figure 2 above we can see the intruder will obtain the exact total wealth of the household in wave 2.

Secondly, if the links are not supplied, then the intruder can proceed as in the illustration above, and use the reported total wealth for the matches it has identified to determine a range of plausible values. If all the total wealth values are close to the original total wealth then we view this as a higher risk than if the total wealth values were spread out over a wide interval. We thus measure the risk by considering, for each split household, what proportion of wave 2 total wealth values, for households the intruder has identified as matches, lie within 5% of the true wave 2 total wealth value of the original household. Specifically, if we denote the set of matches by M then we calculate for split household i:

$$p_i = \frac{1}{c_i} \sum_{j \in M} I\left(\frac{|y_j^{(2)} - y_i^{(2)}|}{y_i^{(2)}} < 0.05 \right) \tag{1}$$

where $I(\cdot)$ is the indicator function taking value 1 if the condition in the bracket is met and 0 otherwise. If $c_i = 0$ then define $p_i = 0$ and otherwise $p_i \in [0, 1]$. In the example

given in Figure 2, we see that amongst the three matches, two households' total wealth lies within 5% of the original total wealth value of 15000, so the value of $p_i = 2/3$ here. The risk measure we use in this article is to consider the number of split households with $p_i \geq 0.5$, i.e. the intruder has at least a 1 in 2 chance of randomly selecting the household's total wealth to within 5%. We denote this attribute disclosure risk value by $A = \sum_i I(p_i \geq 0.5)$. In the example in Figure 2 we would flag this household as a risky household in the data set.

We note that the true household need not be in the set of matches for the intruder to obtain a value of $p_i \geq 0.5$, for example the intruder may identify a different household to the true household as a match, but with a very similar total wealth value. The choice of 5% and the threshold value of p_i is arbitrary, and we could consider investigating different choices but this is not explored further here.

In the two scenarios described above we can calculate this attribute disclosure risk. In scenario 1 this yielded a value of $A = 130$, so there were 130 households for which the intruder had at least a 1 in 2 chance of obtaining the true total wealth value of the household to within 5%. This by definition includes those households with an identification disclosure risk of 0.5 or higher. In scenario 2 the value of A was, as expected, substantially higher with $A = 897$.

The risk metric in (1) above can also be adapted to measure the risk present when the links are also released. Here there will only be one match per split household (obtained through the link) so $c_i = 1$, and we can still measure A in the same way but now p_i will be either 0 or 1. The value of A here reduces to computing the number of split households's total wealth that an intruder could obtain to within 5% of its true value using the links provided. In the example in Figure 2 we see that $p_i = 1$. Clearly here, where no alteration has taken place to the total wealth of households in wave 2, the value of A will equal 1486, but if SDC methods are applied to the total wealth in wave 2, then this value of A may be reduced. More details about the SDC methods considered are provided in the next section.

4 SDC methods

To protect the split household from being identified by the intruder would require substantial alteration of variables in $(x_i, y_i^{(1)}, \ldots, y_i^{(T-1)})$. Standard SDC methods would likely reduce the utility of the data to such an extent that there would be little benefit for analysts. There are techniques that exist in the literature that can perform this alteration and attempt to preserve the utility in the data, such as multiple imputation techniques for generating synthetic data (Reiter, 2003, 2004, 2005; Mitra and Reiter, 2006; Reiter and Mitra, 2009). However, in practice applying these techniques would be challenging due to the large number of variables involved (as noted in the WAS). Determining which variables to synthesise is not immediately obvious and the synthesis models would need to be able to deal with large multivariate distributions often with complex dependencies present between variables. Such an approach would also become very specific to a given data set and could not be generalised to any longitudinal household data set affected by this confidentiality problem, which is the main aim of this work.

The approach we take is to allow intruders to be able to identify the split household in the data, but prevent any attribute disclosure from taking place. Specifically, we alter the information present in $y_i^{(T)}$ only, and in doing so prevent the intruder from learning any new sensitive information about this household. Thus our methodology is a general all

purpose approach to handling this type of disclosure risk. The two approaches we consider to alter the information are swapping and synthesis.

4.1 Swapping

In order to preserve the utility of the released data we would like to replace the information in $y_i^{(T)}$ with information from a similar household, i.e. one that shares similar characteristics to the affected household. In order to do this we use a propensity score nearest neighbour matching scheme (Rosenbaum and Rubin, 1983, 1985).

This approach works by defining a binary indicator $z_i \in \{0, 1\}$, $i = 1, \ldots, n$ where $z_i = 1$ if household i is a split household in wave T and $z_i = 0$ otherwise. Suppose there are $n_1 = \sum_{i=1}^{n} z_i$ split households in total in the data set. We then estimate the following conditional probability for all households that are present in all waves,

$$e_i = P(z_i = 1 | x_i, y_i^{(1)}, \ldots, y_i^{(T)}) \tag{2}$$

which is denoted to be the propensity score, and is typically estimated with a logistic regression of $Z = (z_1, \ldots, z_n)$ on X and $Y = (Y^{(1)}, \ldots, Y^{(T)})$. Rosenbaum and Rubin (1983) show that if two records i and j have the same value of $e_i = e_j$, then their covariate information $(x_i, y_i^{(1)}, \ldots, y_i^{(T)})$ and $(x_j, y_j^{(1)}, \ldots, y_j^{(T)})$ respectively will come from the same distribution. Thus, by matching the household from the non-split group with the closest propensity score value to a split household i's propensity score, and using this household to swap the sensitive information $y_i^{(T)}$, we should be largely preserving the statistical properties of any released data. We implement such a scheme, selecting households from the non-split group as candidate matches without replacement for those from the split group. We note that there are alternative ways to select the matches such as matching with replacement or optimal matching strategies such as full matching (Rosenbaum, 1991). However implementing these schemes would not be straightforward, and we use the nearest neighbour matching scheme without replacement for simplicity. We note a similar approach for swapping using propensity scores was proposed by Oganian and Lesaja (2016).

One potential problem with the approach outlined above is that information from split households will be replaced with information from non-split households. This should work well in limiting the disclosure risk, but we may not be able to preserve certain characteristics of this particular sub-group. We might decide to take a less conservative approach to protecting risk where we decide to only replace a certain proportion of the split households' sensitive data with other similar split households' data. This would preserve the univariate cross-sectional properties of the split household group, e.g. the means and standard deviations of variables in the split group in any given wave would remain unchanged. Specifically we now predefine a certain proportion ρ of split households to have their sensitive information replaced, defined through a new indicator $Z_i^* \in \{0, 1\}, i = 1, \ldots n_1$ for those households i belonging to the split group, i.e. for $i : z_i = 1, i = 1, \ldots, n$. The scheme proceeds in a similar manner to that described above, but now finding matches within the restricted set of split households using the indicator Z_i^*.

If we only replace information for a certain proportion of split households, we must decide which households to select. A random selection is perhaps the simplest way to proceed but not necessarily optimal. A more effective strategy would be to select households according to some risk metric, with households more at risk picked ahead of households less at risk. The risk metric we decide to use is based on that in (1) and defined as,

$$r_i = \frac{1}{c_i} \sum_{j \in M} \frac{|y_j^{(T)} - y_i^{(T)}|}{y_i^{(T)}} \tag{3}$$

where $T = 2$ in our illustration on the WAS. This measures the absolute difference between matched households' total wealth and the true total wealth of the household in the current wave T, averaged over the set of matches. The smaller this number the riskier the record. For example if a household was uniquely and correctly matched in wave T then its r_i value would be 0. In the example given in Figure 2 we can see the value of $r_i = \frac{1}{3}(200/15000 + 0 + 2500/15000) = 0.06$. We can then select records for swapping in ascending order of r_i. In the illustration given in Section 5 we compare this targeted selection with a random selection and find that the targeted selection greatly reduces risks for a given proportion of households being protected.

4.2 Synthesis

In this approach, rather than finding a similar household with which to swap the sensitive information, we instead replace $y_i^{(T)}$ with synthetic values drawn from a model for $y^{(T)}$ fit to the data. Specifically we draw new values from a model given by the conditional distribution
$p(y^{(T)}|X, y^{(1)}, \ldots, y^{(T-1)})$. Denote the synthetic values drawn in this way by $y^{(T*)}$. As determining a plausible parametric model is typically challenging in these data sets we use a non-parametric approach based on classification and regression trees (CART) proposed by Reiter (2005). The approach works by using CART to partition the covariate space into leaves that have approximately homogenous outcome values. Then to synthesise a new value for a household, we essentially find which leaf that household belongs to, using its covariate information, and then use the outcomes in that leaf to draw a new outcome value for this household. As with the swapping approach, we can either choose to synthesise all records within the split household group, or only a proportion of split households, the metric given in (3) above is again used to target which households are most at risk of disclosure and should have their sensitive information synthesised.

Using imputation or synthesis models for protecting confidentiality is an increasingly popular approach. One potential advantage over swapping is that only the information for those records identified as being at risk will be altered, while swapping will additionally affect those records whose information will be used to swap with the sensitive records' information. We consider applying both approaches to the WAS data in the next section and evaluate the disclosure risk of the altered data.

As mentioned earlier, when evaluating the disclosure risk A defined in the previous section, we can consider the situations where the intruder uses the links to identify households in wave 2, or does not use the links (perhaps they may have been removed or the intruder does not trust the reliability of this approach in light of the SDC applied) and instead attempts to find the household using the record linkage approach described in Section 3.1. It may be the case that the approach of using the links does not result in a much larger value of A, in which case releasing the links does not substantially increase risk. Figure 3 illustrates how this can happen. We can see that after SDC has been applied (either through swapping or synthesis) the original total wealth of the split household in wave 2 has been changed from 15000 to 16000, which is more than a 5% difference. Thus, using the link to learn about the household's total wealth in wave 2 will not be considered disclosive with $p_i = 0$. However, ignoring the link and using the record linkage strategy, results in three

matches (as before), two of which have released total wealth values within 5% of 15000, so $p_i = 2/3$.

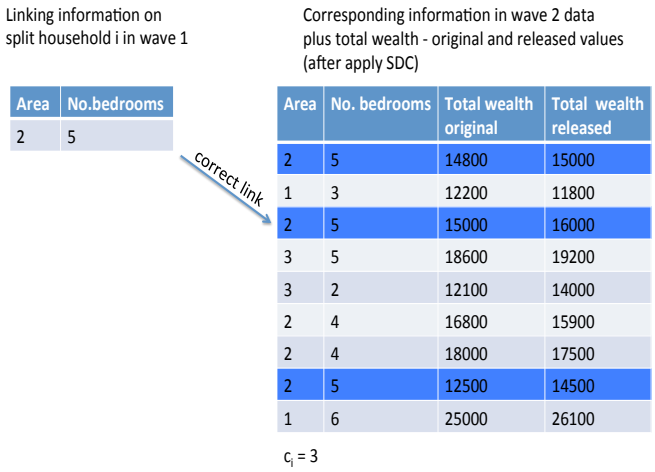

Figure 3: Illustrating the risks after applying SDC to total wealth in wave 2, the intruder could use the link provided to infer the household's total wealth in wave 2, or instead ignore the link and determine a range of total wealth values from amongst the three matches it has identified.

5 Illustration of the method on the WAS

We now apply both the proposed methods described in the previous section to the WAS data and evaluate the attribute disclosure risk defined in Section 3.1. As before, we assume both scenarios regarding the intruders' state of knowledge concerning the household.

- Scenario 1: X includes output area code, accommodation type and number of bedrooms reported in wave 1.
- Scenario 2: X includes output area code, accommodation type and number of bedrooms reported in wave 1, and also number of children and household type reported in wave 2.

As before, we assume $Y^{(t)}$ contains just one variable, total wealth, for waves $t = 1, 2$. In both scenarios, the identification disclosure risks remain the same as those reported in Section 3. However, the intruder will now be unsure as to whether the reported total wealth for the households it has identified as potential matches correspond to the originally reported values, or have been altered to protect confidentiality. We consider both swapping and synthesis as methods to protect confidentiality here.

5.1 Altering a proportion of values within the split household group

We first consider the approach of only altering a proportion of split households total wealth values. When the SDC method is swapping, Figure 4 below plots the value of the attribute disclosure risk A as ρ increases from 0.05 to 0.3 in steps of 0.05. We see that in both scenarios there is a steady decrease in the risk as ρ increases, and as before the disclosure risk is far higher in scenario 2 than for scenario 1. There is also a definite risk reduction in both scenarios when targeting records for swapping as opposed to randomly selecting records for swapping.

We can present similar illustrations when applying the synthesis method. Figure 5 gives risk summaries when only a proportion of split households total wealth values are altered through synthesis using the CART method. We see similar profiles to those in Figure 4. There tends to be a decrease in risk as ρ increases, with risks higher in scenario 2, and targeted synthesis reduces risks in both scenarios. We note that it is possible to modify the specific CART model used, e.g. by changing the minimum leaf size, but this is not explored here.

Considering the above plots, we can see that in general the risks tend to be lower when swapping as opposed to synthesising. This is reasonable given that swapping effectively alters twice the number of records compared to synthesis. We can also consider combining both approaches, for a given proportion of households to be altered we select a certain (approximate) percentage of these to be protected using swapping, and the rest are protected using synthesis. Figure 6 below presents risk summaries for both scenarios for different combinations, when records are randomly selected for alteration using the targeted approach. We see that that in general risks decrease as the percentage synthesised decreases. Thus for risk protection here it appears that swapping is more effective than synthesis. In practice however, we would want to target either swapping or synthesis to a given record based on its characteristics rather than randomly choosing between the two. For example some records might have very good candidate matches compared to others, e.g. their propensity scores are similar but wealth values are quite different, which makes them good candidates for reducing the risk when swapping. Other records might be well modelled by the synthesis method, e.g. with fitted values close to their actual values but with a high predictive variance which would help in lowering the risk if synthesised. This is an area that merits further research.

5.2 Altering all values within the split household group

Table 1 presents attribute risk summaries when altering all split household total wealth values. Results are given for both scenarios and for when values are altered using swapping or synthesis. The attribute risk when the intruder uses the links are also provided (if they are released). We note that when using the links, the value of A is independent of the scenario as the intruder is not performing a record linkage experiment to find a candidate set of matches.

When swapping, we see that risks tend to be lower when swapping all records as opposed to only swapping a proportion of records, although in Scenario 1 a similar level of risk can be achieved by only targeting a proportion of records to be swapped within the split household group. When synthesising, we see that there is a definite benefit here in reducing risk through synthesising all total wealth values for the split group. Risks in general tend to be higher when synthesising rather than swapping which is consistent with what was observed in the previous section.

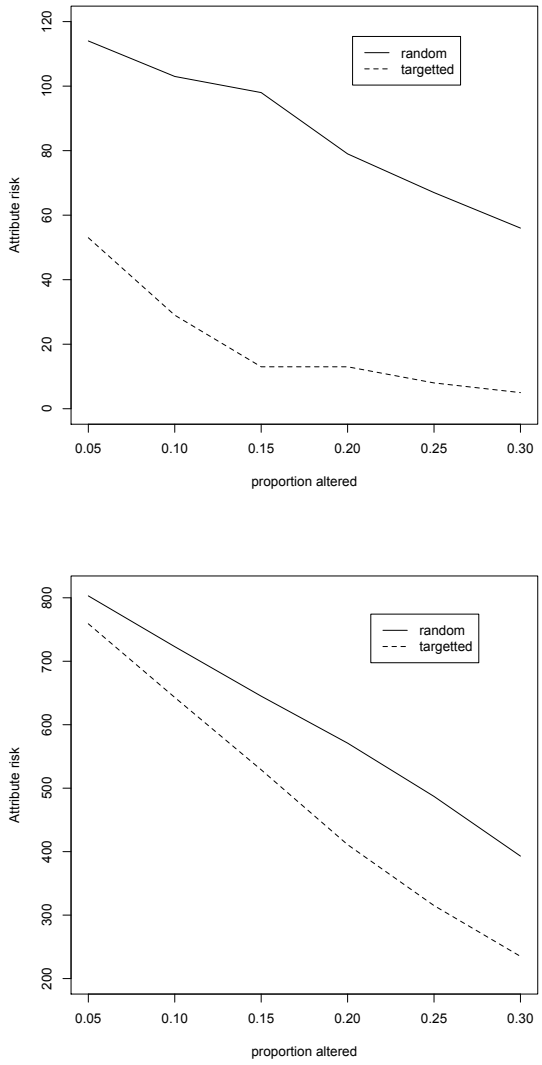

Figure 4: Attribute disclosure risks for scenario 1 (left) and scenario 2 (right) when swapping for proportions altered from 0.05 to 0.30. Note the different scales on the y axis for both plots.

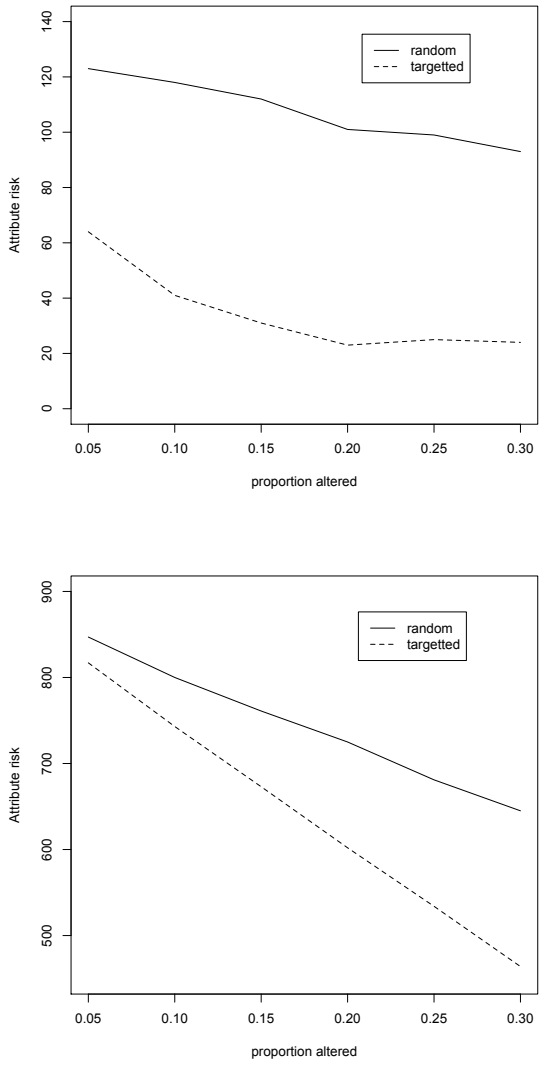

Figure 5: Attribute disclosure risks for scenario 1 (left) and scenario 2 (right) when synthesising for proportions altered from 0.05 to 0.30. Note the different scales on the y axis for both plots.

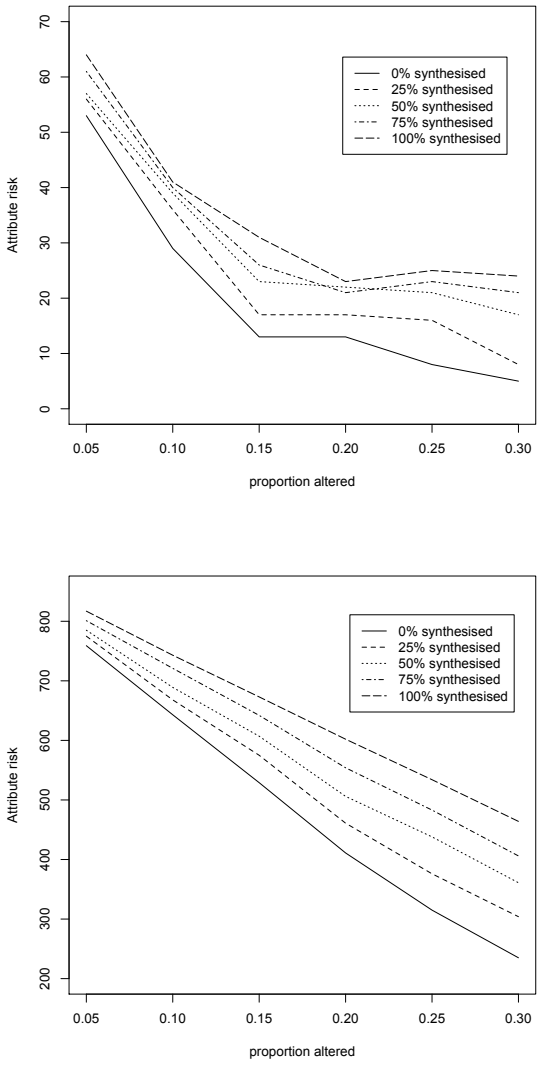

Figure 6: Attribute disclosure risks for scenario 1 (left) and scenario 2 (right), when a certain (approximate) percentage of records' sensitive information are swapped with the rest synthesised, for proportions altered from 0.05 to 0.30. Note the different scales on the y axis for both plots.

Table 1: Attribute risk summaries (value of A) when altering all split household total wealth values

SDC method	Not using links		Using links
	scenario 1	scenario 2	
Swapping	6	31	40
Synthesis	4	55	56

If the links are released here, and the intruder chooses to use them to identify households in wave 2, we see that risks increase, but not greatly. For both swapping and synthesis we see that the value of A is the highest when using the links but not substantially higher than the value of A in scenario 2. This indicates that releasing the links does not necessarily comprise confidentiality here as the attribute risk can be seen to be similar to scenarios where the links are not released. Releasing the links will also greatly improve the utility of the data, allowing longitudinal analyses to be performed. This is considered further in the next section.

5.3 Evaluating utility

We might also be concerned with what impact the SDC methods have on the utility of the data. We focus on examining the impact on a relevant longitudinal analysis that has been performed by the ONS that considers factors determining inheritance, one of which is total wealth. The analysis splits households into 10 groups (deciles) from least to most wealthy with the 5th group (50-60th percentiles) of wealth as the baseline. The analysis specifically considers a logistic regression on whether a household inherits or not on total wealth as well as age, sex, economic activity, household type and tenure from wave 1, highest level of qualification and socio-economic classification from wave 2. Total wealth and whether the household received inheritance were from wave 2. As variables are included across both waves, clearly the links will need to be available for this type of analysis to be performed.

Summaries of the regression coefficients were presented by giving their point estimates as well as associated 95% confidence intervals. We can consider how these summaries are affected when altering the total wealth using swapping or synthesis. We only consider the approach where all split household total wealth was altered, although in principle the approach of altering only a proportion of split household total wealth could also be easily considered. To simplify the illustration we only fit a logistic regression of inheritance status on total wealth, age and sex, but note that the results are similar when performing the complete full analyses.

Figure 7 below plots the confidence intervals of parameter estimates when the regression model is fit to the original and altered data, both by swapping and synthesis. We see that the main conclusions of the analysis are unchanged. For example the wealthiest households are still more likely to inherit in the protected data. In some instances, confidence intervals are wider and thus endpoints are more likely to overlap zero after applying disclosure control, particularly after swapping. On variables that have not been altered, the intercept, age, and sex, we see very little difference in the results before and after applying disclosure control.

Further evaluation of the utility could be performed by considering other longitudinal analyses that might be of interest, and would be important when considering the release of any SDC protected longitudinal data set. It is also possible to consider global utility

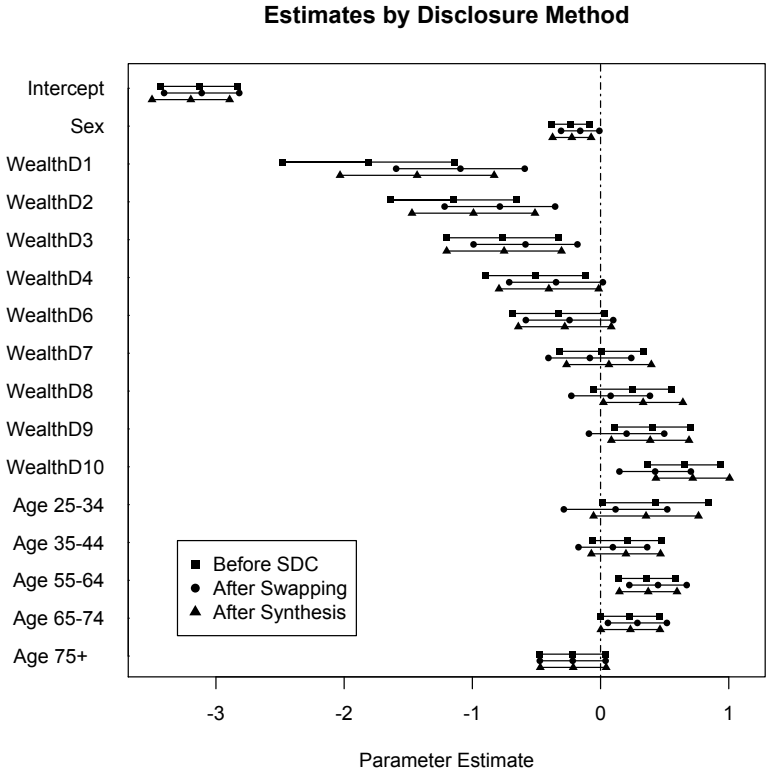

Figure 7: Point and interval estimates from the inheritance analysis based on the original data, as well as the data protected using swapping and synthesis

metrics that seeks to summarise the utility of the released data in general terms, such as those described in Woo *et al.* (2009), but this is not considered further here.

6 Concluding remarks

In this article we have identified a new type of disclosure risk present in longitudinal data; that of where an individual leaves a household in a particular wave, and is subsequently able to self-identify the household in the data and learn new sensitive information about the household. We illustrated the relevance of this disclosure risk in the WAS by applying record linkage techniques and modelling intruder behaviour under different assumptions of intruder knowledge. We also investigated SDC methods of swapping and synthesis to mitigate this disclosure risk, and we saw they were effective in doing so. We noted that the utility would be diminished as a result of applying this approach.

Future research would include a more thorough assessment of the risk/utility balance, and in particular deciding what proportion of households should be altered, as well as potentially determining which households would be better protected through swapping or synthesis. We can also consider other types of inter-wave risk that might arise in general, as well as other risks that might be inherently present due to the longitudinal nature of the data.

In our illustration we focused on a simple example involving only a few variables. In practice the data sets are likely to have a much greater number of variables and so implementing the method will become more challenging. In particular, it may not be possible to include all the variables in the propensity score matching model, and so we would need to consider what sets of variables would be most appropriate to include in the matching model. This is primarily a model selection problem and there is a lot of literature that could be explored, for example LASSO models or Bayesian model averaging strategies. Similarly, the model selection problem would apply when formulating synthesis models and is a topic for future research.

We note that there are various different synthesis methods that could be considered to generate the synthetic values here. We explored using a parametric synthesis approach, using a linear regression model to synthesis wealth values but results were not qualitatively different to the synthesis approach using CART, and in some cases did worse. We did not explore this more as it would be more complicated to determine the most plausible parametric synthesis model with such a complex data set. Various other non-parametric synthesis methods have been explored in Drechsler and Reiter (2011), and in future research it would be interesting to see whether other synthesis methods offer any advantages here, although Drechsler and Reiter (2011) concluded that CART performs the best in general.

We note we have only considered two waves in our model and would also like to ensure this method can work well when extended across more waves. A practical issue we may also encounter is that often these surveys are ongoing. So we may in the future receive new waves' data. As such, we might find records that were not deemed to be at risk previously, now are determined to be risky. We would thus need to be able to handle the dynamic nature of this problem. In addition, we need to take into account the SDC already applied to the data so that relationships across waves remain consistent. As longitudinal data becomes ever more in demand, it is vitally important we understand the risks inherently associated with such data so that confidentiality is maintained.

References

Allen, N. J. and Meyer, J. P. (1990). Organizational socialization tactics: A longitudinal analysis of links to newcomers' commitment and role orientation. *Academy of Management Journal* **33**, 4, 847–858.

Daniel, E. and Bright, G. (2011). Exploring the geographical distribution of wealth using the output area classification. *Economic and Labour Market Review* **5**, 1, 56.

Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis* **55**, 12, 3232–3243.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association* **64**, 1183–1210.

Frees, E. W. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P.-P. (2012). *Statistical disclosure control*. John Wiley & Sons.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrer and L. Franconi, eds., *Privacy in Statistical Databases*, 177–188. New York: Springer-Verlag.

Office for National Statistics (2012). Wealth and Assets Survey Review Report. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/economy/wealth-and-assets-survey/index.html.

Office for National Statistics (2014). Disclosure control guidance for microdata produced from social surveys. Available at: http://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyforsocialsurveymicrodata.

Oganian, A. and Lesaja, G. (2016). Propensity score based conditional group swapping for disclosure limitation of strata-defining variables. In *International Conference on Privacy in Statistical Databases*, 69–80. Springer.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* **1**, 99–110.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society, Series B-Methodological* **53**, 597–610.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* **39**, 1, 33–38.

Steinfield, C., Ellison, N. B., and Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology* **29**, 6, 434–445.

Velez, C. N., Johnson, J., and Cohen, P. (1989). A longitudinal analysis of selected risk factors for childhood psychopathology. *Journal of the American Academy of Child & Adolescent Psychiatry* **28**, 6, 861–864.

Von Korff, M., Ormel, J., Katon, W., and Lin, E. H. (1992). Disability and depression among high utilizers of health care: a longitudinal analysis. *Archives of general psychiatry* **49**, 2, 91–00.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* **1**, 111–124.