SSC Annual Meeting, May 2022
Proceedings of the Survey Methods Section

A NEW MODEL FOR THE AUTOMATED IDENTIFICATION OF DUPLICATE RECORDS

Abel Dasylva[1] and Arthur Goussanou[2]

ABSTRACT

Duplicate records are records from the same unit in a given data source, regardless of whether they are identical. Their identification is required when the source is used to produce official statistics, such as a sampling frame or a census. To date, many Bayesian models have been described to perform this task in an automated manner. Yet, they involve computer-intensive procedures and tend to assume that the linkage variables are conditionally independent, when this is seldom the case in practice. To overcome these limitations, a new model is described for applications, where one can reasonably assume that each unit is associated with at most two records because duplication is rare, as in the private dwellings of the census of population. The duplication is modeled through the number of links from a given record as in a recent model of linkage errors, while extending the latter to account for the multiplicity of false positives from some other unit.

KEY WORDS: entity resolution, record linkage, deduplication

RÉSUMÉ

Les enregistrements en double proviennent de la même unité dans une source de données précise, qu'ils soient identiques ou non. Il est nécessaire de les identifier lorsque la source est utilisée pour produire des statistiques officielles, comme c'est le cas pour une base de sondage ou un recensement. À ce jour, plusieurs modelès bayésiens ont été décrits afin de réaliser cette tâche de manière automatisée. Cependant, ils nécessitent des procédures informatiques intensives et font généralement l'hypothèse que les variables de couplage sont conditionnellement indépendantes, alors que c'est rarement le cas dans la pratique. Pour pallier ces limites, nous décrivons un nouveau modèle pour les applications où on peut raisonnablement supposer que chaque unité est associée à au plus deux enregistrements, parce que les doublons sont rares, comme c'est le cas pour les personnes vivant dans les logements privés, dans le cadre du recensement de la population. Nous modélisons la duplication par le nombre de liens issus d'un enregistrement donné, comme dans un modèle récent des erreurs de couplage tout en généralisant ce dernier afin de tenir compte de la multiplicité des faux positifs provenant d'une autre unité.

MOTS CLÉS : résolution d'entité, appariement, déduplication

1 INTRODUCTION

In a file, two records are called duplicates if they refer to the same unit (e.g. a person, a household or a business), regardless of whether they are identical. Various problems arise when a file contains unknown duplicate records. In a sampling frame, they increase the response burden. In a population census or a register, they produce some over-count (Statistics Canada 2019). To avoid such problems, the common

[1]Abel Dasylva, R.-H. Coats building, 100 Tunney's pasture driveway Ottawa ON, K1A0T6, abel.dasylva@statcan.gc.ca
[2]Arthur Goussanou, R.-H. Coats building, 100 Tunney's pasture driveway Ottawa ON, K1A0T6, arthur.goussanou@statcan.gc.ca

solution consists in identifying the duplicate records by linking the file to itself. However linkage errors arise when linking with quasi-identifiers such as names, dates or postal codes. These errors include the false negatives (FN) and the false positives (FP), where a false negative is failing to link records from the same unit and a false positive is linking records from different units. When one is deduplicating a file, false negatives mean that some duplicate records remain, while false positives lead to under-coverage. Measuring these errors is critical to optimize the linkage decisions and report the quality of the deduplicated file. However it is a challenge. One solution uses clerical reviews, i.e. the visual inspections of a probability sample of record pairs to determine if they are from the same unit (Dasylva et al. 2016). Clerical reviews are also used to estimate the proportion of duplicate records in the Canadian census (Statistics Canada 2019). However they are best avoided due to their high cost. An alternative solution is to use a Bayesian model of entity resolution as suggested by Fortini et al. (2001), Tancredi & Liseo (2011), Steorts et al. (2016) or Sadinle (2017). However this approach has two major drawbacks. In addition to being computer intensive, all the Bayesian models to date assume that the linkage variables are conditionally independent or uncorrelated, which may be untrue in practice according to Newcombe (1988), Blakely & Salmond (2002) and Belin & Rubin (1995). To overcome these limitations, a new model is proposed for applications, where one can reasonably assume that each unit is associated with at most two records because duplication is rare, as in the private dwellings of the census of population. This model extends previous work by Dasylva & Goussanou (2020, 2021, 2022).

The remaining sections describe the notation and assumptions, theory, data experiment and conclusion, in this order.

2 NOTATION AND ASSUMPTIONS

This work considers a large finite population and a file, which is based on a Bernoulli sample from this population. Within the file, each unit is associated with one or two records. To identify the duplicate records the file is linked to itself such that the decision to link two given records involves no other records. The following paragraphs provide further details.

Finite population and file: Consider a finite population of N units and a Bernoulli sample s that is drawn from this population with the inclusion probability τ not depending on N. The file is generated by associating r_i records with typographical errors to unit i. With a positive probability, two records are generated. Otherwise, a single record is produced. The file is identified with the subset $\{(i,j)\ s.t.\ i \in s, 1 \leq j \leq r_i\}$ from $\{1,\ldots,N\} \times \{1,2\}$. For unit $i \in s$, the related records are denoted by $V_{(i,1)},\ldots,V_{(i,R_i)}$ that live on the record space $\mathcal{V}_N$, which may be discrete or continuous. Let $\mathcal{D}_N$ denote a subset of $\mathcal{V}_N$, which is of special interest (e.g. a post-stratum), and let ϕ_N denote the corresponding subset of the file, i.e. $\phi_N = \{(i,j)\ s.t.\ i \in s, 1 \leq j \leq r_i\ and\ V_{(i,j)} \in \mathcal{D}_N\}$. Also let $\delta = P\left(r_i = 2 | i \in s, V_{(i,1)} \in \mathcal{D}_N\right)$ and $\overline{\mu} = E\left[r_i | i \in s, V_{(i,1)} \in \mathcal{D}_N\right] = 1+\delta$, which are assumed to not depend on N. The file and record generation mechanism are assumed to be such that $\left[\left(I(i \in s), r_i, \left[V_{(i,j)}\right]_{1 \leq j \leq r_i}\right)\right]_{1 \leq i \leq N}$ are independent and identically distributed, and $V_{(i,1)}$ and $V_{(i,2)}$ are identically distributed given that $i \in s$ and $r_i = 2$. Also suppose that the joint distribution of $I(i \in s), r_i$ and $\left[I(V_{(i,j)} \in \mathcal{D}_N)\right]_{1 \leq j \leq r_i}$ does not depend on N.

Linkage: For distinct (i,j) and (i',j'), let $L_{(i,j)(i',j')}$ denote the indicator of a link between $V_{(i,j)}$ and $V_{(i',j')}$. The linkage is assumed to be such that $L_{(i,j)(i',j')}$ is only a function of $V_{(i,j)}$ and $V_{(i',j')}$, i.e. $L_{(i,j)(i',j')}$ is independent of all the other information (including s, $[r_i]_{i \in s}$, the other record values, and the linkage decisions at the other record pairs), given $V_{(i,j)}$ and $V_{(i',j')}$. It is further assumed that the linkage decisions are symmetric, i.e. $L_{(i,j)(i',j')} = L_{(i',j')(i,j)}$.

Linkage errors: These errors are unavoidable when linking with quasi-identifiers, including false negatives and false positives, where a false negative is failing to link records from the same unit and a false positive is linking records from different units. For completeness, define a true positive as linking two records from

the same unit and a true negative as not linking two records from the different units. These different pair types are equivalently defined by calling a pair matched if its two records are from the same unit, and calling it unmatched otherwise. Then a false negative is a matched pair that is not linked, a false positive is an unmatched pair that is linked and a true positive is a matched pair that is linked. Let TP, FN and FP respectively denote the numbers of true positives, false negatives and false positives, for the pairs that involve at least one record from ϕ_N. For these pairs, the linkage errors may be assessed by the recall equal to $TP/(TP+FN)$ and the precision equal to $TP/(TP+FP)$. According to Blakely & Salmond (2002) and Dasylva & Goussanou (2020, 2021, 2022), the recall and precision may be estimated by modeling the number of links from a given record, when linking a file to a complete census. The resulting solutions have the advantage of dispensing with assumptions about the dependence of the linkage variables, e.g. their conditional independence (Fellegi & Sunter 1969). To extend this approach to the current setting means modeling the number of links from $V_{(i,j)}$ for $(i,j) \in \phi_N$. To this end, define $n_{(i,j)|M} = \sum_{j'=1:\ j' \neq j}^{r_i} L_{(i,j)(i,j')}$, $n_{(i,j)|U} = \sum_{i' \in s-\{i\}} \sum_{j'=1}^{r_{i'}} L_{(i,j)(i',j')}$ and $n_{(i,j)} = n_{(i,j)|M} + n_{(i,j)|U}$.

Regularity conditions: To extend the model from Dasylva & Goussanou (2020, 2021, 2022) one must also extend the regularity conditions that are assumed therein. For $v \in \mathcal{V}_N$, define

$$\mu_N(v) = E\left[r_i \middle| i \in s, V_{(i,1)} = v\right], \tag{1}$$

$$p_N(v) = \mu_N(V_{(i,1)})^{-1} E\left[I(r_i = 2) r_i L_{(i,1)(i,2)} \middle| i \in s, V_{(i,1)} = v\right], \tag{2}$$

$$\lambda_{kN}(v) = \sum_{t=k}^{2} P\left(r_{i'} = t, \sum_{j'=1}^{r_{i'}} L_{(i,1)(i',j')} = k \middle| \{i,i'\} \subset s, V_{(i,1)} = v\right),\ k = 1,2, \tag{3}$$

with the convention that $p_N(v) = 0$ if $\mu_N(v) = 0$. The first proposed condition imposes an upper-bound on the expected number of false positives for any record in $\mathcal{D}_N$ based on

$$(N-1) \sup_{v \in \mathcal{D}_N} \max\left(\lambda_{1N}(v), \lambda_{2N}(v)\right) \leq \Lambda, \tag{4}$$

where Λ does not depend on N. The second condition imposes an invariant (with respect to N) conditional joint distribution of $\mu_N(V_{(i,1)})$, $p_N(V_{(i,1)})$, $(N-1)\lambda_{1N}(V_{(i,1)})$ and $(N-1)\lambda_{2N}(V_{(i,1)})$ given that $i \in s$ and $V_{(i,1)} \in \mathcal{D}_N$, with $F()$ denoting this invariant distribution, i.e.

$$\left(\mu_N(V_{(i,1)}), p_N(V_{(i,1)}), (N-1)\lambda_{1N}(V_{(i,1)}), (N-1)\lambda_{2N}(V_{(i,1)})\right) \middle| \{i \in s, V_{(i,1)} \in \mathcal{D}_N\} \sim F(). \tag{5}$$

3 THEORY

The main result of this communication is a theorem stating the convergence in distribution of the number of links from a given record, under the proposed regularity conditions. It also provides the basis for the proposed model. After stating the theorem, the following paragraphs describe the implications for the estimation of the linkage errors and the proportion of duplicate records.

Limiting distribution: Let (I,J) be drawn uniformly from ϕ_N and define

$$\widetilde{n}_M = I\left(|\phi_N| \geq 1\right) \sum_{(i,j) \in \phi_N} I\left((I,J) = (i,j)\right) n_{(i,j)|M},$$

$$\widetilde{n}_U = I\left(|\phi_N| \geq 1\right) \sum_{(i,j) \in \phi_N} I\left((I,J) = (i,j)\right) n_{(i,j)|U}$$

and $\widetilde{n} = \widetilde{n}_M + \widetilde{n}_U$. The following theorem states the convergence in distribution of $(\widetilde{n}_M, \widetilde{n}_U)$ (and thus of $\widetilde{n}$) based on the convergence of the characteristic function of $(\widetilde{n}_M, \widetilde{n}_U)$, which is defined as

$$H(\omega_1, \omega_2) = E\left[e^{\jmath(\omega_1 \widetilde{n}_M + \omega_2 \widetilde{n}_U)}\right],\ \omega_1, \omega_2 \in I\!R,$$

where $\jmath$ is the complex number such that $\jmath^2 = -1$. The limiting distributions involve two related families of discrete distributions, which are hereafter called base distributions. The first family is denoted by $\mathcal{F}_1$ and comprises the discrete univariate distributions that correspond to a random variable of the form $X+Y+2Z$, where X, Y and Z are mutually independent, $X \sim Bernoulli(p)$, $Y \sim Poisson(\nu_1)$ and $Z \sim Poisson(\nu_2)$, for $p \in [0,1]$ and $\nu_1, \nu_2 \geq 0$. For such a member distribution, the probability mass function is given by

$$\begin{aligned} q\left(t; p, \nu_1, \nu_2\right) &= e^{-\nu_1-\nu_2}\left(I(t=0)(1-p) + I(t>0)\left((1-p)\sum_{k=0}^{\lfloor t/2\rfloor}\frac{\nu_1^{t-2k}}{(t-2k)!}\frac{\nu_2^k}{k!} + p\sum_{k=0}^{\lfloor (t-1)/2\rfloor}\frac{\nu_1^{(t-1)-2k}}{((t-1)-2k)!}\frac{\nu_2^k}{k!}\right)\right), \\ & t = 0, 1, 2, \ldots. \end{aligned} \tag{6}$$

The second family is denoted by $\mathcal{F}_2$ and comprises the discrete bivariate distributions that correspond to random bivariate vectors of the form $(X, Y+2Z)$, with X, Y and Z as given above. For such vectors, the characteristic function is of the form

$$E\left[e^{\jmath(\omega_1 X + \omega_2(Y+2Z))}\right] = \left[1 + p\left(e^{\jmath\omega_1} - 1\right)\right]\exp\left(\nu_1\left(e^{\jmath\omega_2} - 1\right) + \nu_2\left(e^{2\jmath\omega_2} - 1\right)\right),\ \omega_1, \omega_2 \in I\!R, \tag{7}$$

where $\jmath$ is the complex number such that $\jmath^2 = -1$. The theorem proof is found in the appendix.

Theorem 1 *Under Eqs. 4-5*

$$\lim_{N\to\infty} H(\omega_1, \omega_2) = \int_{(\mu,p,\lambda_1,\lambda_2)\in[0,1]^2\times[0,\Lambda]^2} \left(1 + p\left(e^{\jmath\omega_1} - 1\right)\right)\exp\left(\tau\sum_{k=1}^{2}\lambda_k\left(e^{\jmath k\omega_2} - 1\right)\right)(\mu/\overline{\mu})dF(\mu, p, \lambda_1, \lambda_2). \tag{8}$$

Thus $\widetilde{n}$ converges to a mixture of distribution from $\mathcal{F}_1$ and $(\widetilde{n}_M, \widetilde{n}_U)$ converges in distribution to a mixture of distributions from $\mathcal{F}_2$.

An important special case is when $F()$ is concentrated at a single atom. It corresponds to homogeneous records, when the limit distribution of $\widetilde{n}$ is a base distribution from $\mathcal{F}_1$. Another important special case is when $F()$ is discrete with finitely many atoms. In this case, the limiting distribution is a finite mixture of base distributions from $\mathcal{F}_1$. For a finite mixture with G components, let α_g and $(p_g, \nu_{1g}, \nu_{2g})$ respectively denote the mixing proportion and the parameters for the g-th component. Also define $\overline{p} = \sum_{g=1}^{G}\alpha_g p_g$ and $\overline{\nu}_l = \sum_{g=1}^{G}\alpha_g\nu_{lg}$ for $l = 1, 2$. Given G, the $[(p_g, \nu_{1g}, \nu_{2g})]_{1\leq g\leq G}$ parameters may be estimated by maximizing the composite likelihood of the $n_{(i,j)}$'s for $(i,j) \in \phi_N$.

Implications for the linkage errors: Measuring the errors without manual interventions is required to optimize the parameters of automated linkage procedures, which may be deterministic, hierarchical or probabilistic (Fellegi & Sunter 1969). Such measures may be obtained by applying a model based on the limiting distribution since the parameters of this distribution are related to the recall and the precision. Indeed note that the above theorem implies the following limits

$$\begin{aligned} \lim_{N\to\infty} E\left[\widetilde{n}_M\right] &= \int_{(\mu,p,\lambda_1,\lambda_2)\in[0,1]^2\times[0,\Lambda]^2} p(\mu/\overline{\mu})dF(d, p, \lambda_1, \lambda_2), \\ \lim_{N\to\infty} E\left[\widetilde{n}_U\right] &= \int_{(\mu,p,\lambda_1,\lambda_2)\in[0,1]^2\times[0,\Lambda]^2} \tau\left(\lambda_1 + 2\lambda_2\right)(\mu/\overline{\mu})dF(\mu, p, \lambda_1, \lambda_2). \end{aligned}$$

Suppose that $|\phi_N|^{-1}FP = |\phi_N|^{-1}\sum_{(i,j)\in\phi_N} n_{(i,j)|U}$ and $|\phi_N|^{-1}TP = |\phi_N|^{-1}\sum_{(i,j)\in\phi_N} n_{(i,j)|M}$ converge in probability to $\lim_{N\to\infty} E\left[\widetilde{n}_U\right]$ and $\lim_{N\to\infty} E\left[\widetilde{n}_M\right]$, respectively. Since

$$\begin{aligned} TP + FN &= \sum_{i=1}^{N}\sum_{j=1}^{r_i} I(i \in s, r_i = 2, V_{(i,j)} \in \mathcal{D}_N), \\ |\phi_N| &= \sum_{i=1}^{N}\sum_{j=1}^{r_i} I(i \in s, V_{(i,j)} \in \mathcal{D}_N), \end{aligned}$$

$TP + FN$ and $|\phi_N|$ are iid sums, which satisfy the law of large numbers, so that

$$\frac{TP + FN}{|\phi_N|} \xrightarrow{p} \frac{E\left[\sum_{j=1}^{r_i} I(i \in s, r_i = 2, V_{(i,j)} \in \mathcal{D}_N)\right]}{E\left[\sum_{j=1}^{r_i} I(i \in s, V_{(i,j)} \in \mathcal{D}_N)\right]} = \frac{2\delta}{1+\delta}.$$

By continuity, it easily follows that

$$\frac{TP}{TP + FN} \xrightarrow{p} \left(\frac{2\delta}{1+\delta}\right)^{-1} \int_{(\mu,p,\lambda_1,\lambda_2)\in[0,1]^2\times[0,\Lambda]^2} p(\mu/\overline{\mu}) dF(d, p, \lambda_1, \lambda_2), \tag{9}$$

$$\frac{TP}{TP + FP} \xrightarrow{p} \frac{\int_{(\mu,p,\lambda_1,\lambda_2)\in[0,1]^2\times[0,\Lambda]^2} p(\mu/\overline{\mu}) dF(d, p, \lambda_1, \lambda_2)}{\int_{(\mu,p,\lambda_1,\lambda_2)\in[0,1]^2\times[0,\Lambda]^2} (p + \tau\lambda_1 + 2\tau\lambda_2)(\mu/\overline{\mu}) dF(\mu, p, \lambda_1, \lambda_2)}. \tag{10}$$

When the limiting distribution is a finite mixture such that $\widetilde{n} \dot{\sim} \sum_{g=1}^{G} \alpha_g q(.; p_g, \nu_{1g}, \nu_{2g})$, the recall (i.e. $TP/(TP + FN)$) converges to $(2\delta/(1+\delta))^{-1}\overline{p}$ while the precision (i.e. $TP/(TP + FP)$) converges to $\overline{p}/(\overline{p} + \overline{\nu}_1 + 2\overline{\nu}_2)$, and the proportion of records with a duplicate that is linked (i.e. $TP/|\phi_N|$) converges to $\overline{p}$. The above results show that knowing δ is not required to estimate the precision but it is needed for the recall. Conversely δ may be estimated if the recall is given. Even when δ is unknown, the estimate of $\overline{p}$ can still be used to compare the recalls for different linkage strategies.

Applications: The model may be used for two applications. In the first application, it is used to estimate the linkage accuracy when deduplicating a file, where the proportion of duplicate records (i.e., $2\delta/(1+\delta)$) is known. In this case, the measured accuracy (including the precision and recall) is useful for optimizing the linkage and reporting the quality of the deduplicated file, in terms of residual duplication and under-coverage. In the second application, the recall is known and the goal is estimating the proportion of duplicate records. For example one may assume a recall of 1.0 if applying a linkage rule that is deemed sufficiently lax. In a probabilistic linkage, such a rule may consist in linking all the pairs that meet the blocking criteria, i.e., computer-efficient criteria for selecting a manageable subset of the Cartesian product (comprising all the possible pairs) where most matched pairs are expected to be found (Statistics Canada 2017). The proposed methodology serves both applications by providing a way of estimating the precision and the proportion of records with a duplicate that is linked (i.e.,$\overline{p}$), where the latter is the product of the recall by the proportion of duplicate records. Thus the success of the methodology depends on how accurately it estimates $\overline{p}$ and the precision. This is evaluated with public data from the 2010 US Census of population in the next section.

4 DATA EXPERIMENT

Setup: The experiment involves the creation of 100 synthetic populations with $N = 1,000,000$ persons with the surname and the birth date based on public data for the 2010 US Census of population, according to Dasylva & Goussanou (2022). For each population, a complete census is created with duplicate records based on $\delta = 0.02, 0.1$ and typographical errors. The experiment follows Dasylva & Goussanou (2022) except for the addition of duplicate records in the census. When two records are generated for a unit, they are generated to be conditionally independent given the true surname and birth date. However this does not imply that these variables are conditionally independent within a record pair. The records are linked based on having the same surname and birth date. To estimate the linkage errors, the records are placed in post-strata based on the log-frequency of the surname (in base 10), and the homogeneous model is fitted within each post-stratum by maximizing the composite likelihood of the $n_{(i,j)}$'s. For simplicity, the surname frequency is given and set to the empirical census frequency, i.e., it is not re-estimated in each repetition. In practice, one would have to estimate the surname frequency from the data.

Results: They appear in Table 1, where it can be seen that the bias and the variance increase with the frequency but decrease with δ. This is expected because a link from a record with a rare surname provides more evidence that there is a duplicate than if the surname is popular. In the latter case, there is a greater chance than the record is linked to a record from a different person, who just happens to have the same surname. Even then, the bias is quite large for the post-stratum with the most frequent surnames when $\delta = 0.02$. This corresponds to a worst-case situation where the proportion to be estimated is small, the related events (whether there is actually a duplicate record) are not directly observed and the fraction of missing information according to Louis (1982) is large because the linkage variables provide little discrimination. Practical solutions exist for estimating such small proportions when the successes are directly observed. In this case, one replaces the goal of computing an accurate point estimator with the more attainable one of finding a confidence interval with good coverage, e.g., the exact interval by Clopper & Pearson (1934). Yet this approach cannot be applied here. One possible remedy is using variables with a greater discriminating power, e.g., combining the given names (one or more), surnames (one or more), birth date, province and postal code as in the over-coverage study for the Canadian Census (Statistics Canada 2019, chap. 8.2.1). When the goal is estimating the proportion of duplicate records (i.e. $2\delta/(1+\delta)$) under the assumption of a perfect recall (i.e., a recall of 1.0), another solution is to assume that δ is uniform across the post-strata and estimate $\overline{p} = 2\delta/(1+\delta)$ (e.g., their average) within the post-strata with less frequent surnames. Note that this latter solution has the major advantage of dispensing with clerical reviews that are costly.

Table 1: Results of the data experiment.

		$\delta = 0.02$		$\delta = 0.1$	
Measure	log-frequency of the surname	Bias(%)	Variance	Bias (%)	Variance
proportion of records with a duplicate that is linked	-6	0.003	3.72E-04	-0.098	1.49E-06
	-5	-0.910	3.84E-04	0.022	7.58E-07
	-4	-4.106	7.22E-03	-0.165	1.89E-06
	-3	28.196	3.26E-01	1.080	1.25E-05
precision	-6	-0.021	4.40E-05	-0.012	5.06E-07
	-5	-0.835	3.65E-04	-0.003	4.42E-06
	-4	-4.164	8.45E-03	-0.123	4.44E-05
	-3	28.122	9.24E-02	1.059	3.23E-04

5 CONCLUSION

This work has described a joint model for the duplication and the linkage errors within a file, where each unit has at most two records and the decision to link two records involves no other record. With this model, one may estimate the precision, the proportion of records with a duplicate that is linked, the recall for a known proportion of duplicate records and the proportion of duplicate records for a known recall. Like the model from Dasylva & Goussanou (2020), which it extends, it accounts for the records heterogeneity and the interactions among the linkage variables implicitly. Future work will aim at reducing the bias of the resulting estimators for small values of δ by considering linkage variables with a greater discriminating power. It will also look at extending the model to more general settings, where a unit may be associated with three or more duplicate records.

6 DISCLAIMER

The content of this paper represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that may not reflect those currently implemented by the Agency.

REFERENCES

Belin, T. & Rubin, D. (1995), 'A method for calibrating false-match rates in record linkage', *Journal of the American Statistical Association* **90**, 694–707.

Blakely, T. & Salmond, C. (2002), 'Probabilistic record linkage and a method to calculate the positive predicted value', *International Journal of Epidemiology* **31**, 1246–1252.

Clopper, C. & Pearson, E. (1934), 'The use of confidence or fiducial limits illustrated in the case of the binomial', *Biometrika* **26**, 404–413.

Dasylva, A., Abeysundera, M., Akpoue, B., Haddou, M. & Saidi, A. (2016), Measuring the quality of a probabilistic linkage through clerical reviews, *in* 'Proceedings of the 2016 International Methodology Symposium'.

Dasylva, A. & Goussanou, A. (2020), Estimating linkage errors under regularity conditions, *in* American Statistical Association, ed., 'In Proceedings of the Section on Survey Research Methods', pp. 687–692.

Dasylva, A. & Goussanou, A. (2021), 'Estimating the false negatives due to blocking in record linkage', *Survey Methodology* **47**(2), 299–311.

Dasylva, A. & Goussanou, A. (2022), 'On the consistent estimation of linkage errors without training data', *Japanese Journal of Statistics and Data Science* **5**, 181–216.
URL: *https://doi.org/10.1007/s42081-022-00153-3*

Fellegi, I. & Sunter, A. (1969), 'A theory of record linkage', *Journal of the American Statistical Association* **64**, 1183–1210.

Fortini, M., Liseo, B., Nuccitelli, A. & Scanu, M. (2001), 'On bayesian record linkage', *Research in Official Statistics* **4**, 185–198.

Louis, T. (1982), 'Finding the observed information matrix when using the em algorithm', *Journal of the Royal Statistical Society B* **44**, 226–233.

Newcombe, H. (1988), *Handbook of Record Linkage*, Oxford University Press, New York.

Sadinle, M. (2017), 'Bayesian estimation of bipartite matchings for record linkage', *Journal of the American Statistical Association* **112**, 600–612.

Statistics Canada (2019), '2016 census of population coverage technical report'. 98-303-X2016001.

Statistics Canada, ed. (2017), *Record Linkage Project Process Model*, Catalog no 12-605-X, Statistics Canada.

Steorts, R., Hall, R. & Fienberg, S. (2016), 'A bayesian approach to graphical record linkage and de-duplication', *Journal of the American Statistical Association* **111**, 1660–1672.

Tancredi, A. & Liseo, B. (2011), 'A hierarchical bayesian approach to record linkage and population size problems', *Annals of Applied Statistics* **5**, 1553–1585.

A PROOF OF THE THEOREM

We have $E[|\phi_N|] = E\left[I(i \in s, V_{(i,1)} \in \mathcal{D}_N) r_i\right] N$. It is also easy to show that there exist $\epsilon \in (0,1)$ and $c > 0$ such that

$$P\left(||\phi_N| - E[|\phi_N|]| > N^{\epsilon}\right) = O\left(e^{-N^c}\right).$$

Thus

$$\begin{aligned} H(\omega_1,\omega_2) &= P\left(|\phi_N|=0\right) + E\left[\frac{I\left(|\phi_N|\geq 1\right)}{|\phi_N|}\sum_{i=1}^{N}\sum_{j=1}^{r_i} I(i\in s,\ V_{(i,j)}\in\mathcal{D}_N)e^{\jmath\omega\left(\omega_1 n_{(i,j)|U}+\omega_2 n_{(i,j)|U}\right)}\right] \\ &= \frac{E\left[\sum_{i=1}^{N}\sum_{j=1}^{r_i} I(i\in s,\ V_{(i,j)}\in\mathcal{D}_N)e^{\jmath\omega\left(\omega_1 n_{(i,j)|U}+\omega_2 n_{(i,j)|U}\right)}\right]}{E\left[|\phi_N|\right]} + o(1) \\ &= \frac{NE\left[I(i\in s, V_{(i,1)}\in\mathcal{D}_N)r_i e^{\jmath\left(\omega_1 n_{(i,1)|U}+\omega_2 n_{(i,1)|U}\right)}\right]}{E\left[I(i\in s, V_{(i,1)}\in\mathcal{D}_N)r_i\right]N} + o(1). \end{aligned}$$

Hence

$$H(\omega_1,\omega_2) = \overline{\mu}^{-1}E\left[r_i A B^{N-1}\left|i\in s, V_{(i,1)}\in\mathcal{D}_N\right.\right] + o(1),$$

where

$$\begin{aligned} A &= E\left[e^{\jmath\omega_1 I(r_i=2)L_{(i,1)(i,2)}}\middle| i\in s, V_{(i,1)}, r_i\right], \\ B &= E\left[\exp\left(\jmath\omega_2 I(i'\in s)\sum_{j'=1}^{r_{i'}} L_{(i,1)(i',j')}\right)\middle| i\in s, V_{(i,1)}, r_i\right],\ i'\neq i \\ &= E\left[\exp\left(\jmath\omega_2 I(i'\in s)\sum_{j'=1}^{r_{i'}} L_{(i,1)(i',j')}\right)\middle| i\in s, V_{(i,1)}\right] \end{aligned}$$

Hence

$$H(\omega_1,\omega_2) = \overline{\mu}^{-1}E\left[E\left[r_i A\left|i\in s, V_{(i,1)}\right.\right] B^{N-1}\left|i\in s, V_{(i,1)}\in\mathcal{D}_N\right.\right] + o(1).$$

$$\begin{aligned} A &= 1 + E\left[I(r_i=2)L_{(i,1)(i,2)}\middle| i\in s, V_{(i,1)}, r_i\right]\left(e^{\jmath\omega_1}-1\right), \\ E\left[r_i A\left|i\in s, V_{(i,1)}\right.\right] &= E\left[r_i\middle| i\in s, V_{(i,1)}\right] + E\left[I(r_i=2)r_i L_{(i,1)(i,2)}\middle| i\in s, V_{(i,1)}, r_i\right]\left(e^{\jmath\omega_1}-1\right) \\ &= \mu_N(V_{(i,1)})\left(1+p_N(V_{(i,1)})\left(e^{\jmath\omega_1}-1\right)\right), \end{aligned}$$

As for B, we have

$$\begin{aligned} B^{N-1} &= \left(1+\frac{\tau}{N-1}\sum_{k=1}^{2}(N-1)\lambda_{kN}(V_{(i,1)})\left(e^{\jmath k\omega_2}-1\right)\right)^{N-1} \\ &= \exp\left(\tau\sum_{k=1}^{2}(N-1)\lambda_{kN}(V_{(i,1)})\left(e^{\jmath k\omega_2}-1\right)\right) + O(1/N). \end{aligned}$$

Hence

$$\begin{aligned} H(\omega_1,\omega_2) &= E\Bigg[\left(1+p_N(V_{(i,1)})\left(e^{\jmath\omega_1}-1\right)\right)\exp\left(\tau\sum_{k=1}^{2}(N-1)\lambda_{kN}(V_{(i,1)})\left(e^{\jmath k\omega_2}-1\right)\right)\times \\ &\quad \left(\overline{\mu}^{-1}\mu_N(V_{(i,1)})\right)\Bigg|\, i\in s, V_{(i,1)}\in\mathcal{D}_N\Bigg] + o(1). \end{aligned}$$

The proof is completed by rewriting the right-hand side of the above equation based on Eq. 5 and by identifying the component distributions of the mixture based on Eq. 7.