Improving deduplication of identities

Matthew J. Bauman

November 30, 2018

1 Introduction

Combining datasets and performing large aggregate analyses are a powerful new way to improve service across large populations. Critically important in this task is the deduplication of identities across multiple data sets that were rarely designed to work together. Inconsistent data entry, typographical errors, and real world identity changes pose significant challenges to this process. This paper describes our solution to this problem: `pgdedupe`.

Deduplicating identities is a challenging and imperfect analysis; there is no perfect algorithm. Therefore, it's important to consider what happens when mistakes are made and if there are particular populations that are more affected by systematic biases. If, for example, we were to overzealously merge together records from a particular demographic group, that group would appear to have fewer individuals that account for more of the data. This can have a direct impact on subsequent decisions and policies that were informed by this analysis. On an individual level, this can lead to erroneously disclosing personal and confidential information to someone under the assumption that they were the same person. Alternatively, a reluctance to merge individuals means that multiple rows would be inaccurately attributed to different people. This could result in a care provider missing crucially important data that would otherwise inform and improve their level of care.

When looking across multiple datasets and even within large single datasets, there are multiple fields that can be used as evidence that two entries describe the same identity. Multiple datasets limit the analyses to only those fields that they have in common. This tends to be names, birthdate, SSN, and perhaps some demographic information such as sex or race. Sometimes, though, there may be contact information; this may include phone numbers, addresses, or actual locations of contact. Each of these fields has its own challenges and nuances, but they almost all contain varying levels of data entry problems. This most frequently manifests as typos, default values or missingness, but there are additional concerns for specific to each field.

The subsequent analyses were informed by a real-world dataset containing over a million records from a US county ("Jurisdiction A"), including sources from the criminal justice system and several behavioral health services. This real-life dataset inspired us to create a synthetic dataset with similar challenges and features, but with a known ground truth to enable rigorous validation. Two other counties (Jurisdictions "B" and "C") provided their current best-practices in data matching as examples.

1.1 Challenges in merging identities by common fields

1.1.1 Social Security Numbers

Social Security Numbers are frequently used by themselves as a key to link entries together, but they're not as robust as their nine digits would suggest. One of the biggest issues is missingness and, more importantly, biases in the distribution of missingness. Published studies have reported between 80-90% missingness, but this strongly depends upon internal policies and practices [VS02, WWL+12]. If the SSN isn't explicitly required for billing or other reasons, that number can be even lower. Worse, the distribution of missingness disproportionately affects certain populations, including immigrants, individuals with visas, and individuals who hesitate to provide private information. It's important to consider how institutional policies affect these biases. If SSNs are mandated fields, for example, then there may be a preponderance of false or shared SSNs within families or certain populations.

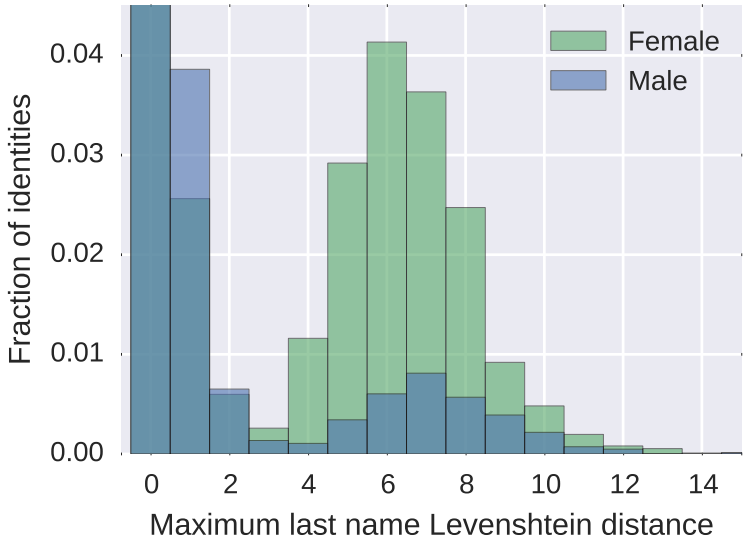

Figure 1: When looking at the real data with exact match record linkage across the first name and SSN, many more female identities have large changes to their last name as determined by the Levenshtein distance than male identities.

Until 2011, Social Security Numbers were allocated in a deterministic manner based upon the location and date of birth. This means that, depending on population mobility, certain localities may have common sets of SSN prefixes. Matching these common digits is less indicative of a matched individual as compared to less common prefixes or the final digits of the number. The distribution of SSN numbers is so deterministic that the location and date of birth is sufficient for accurately predicting the majority of the digits [AG09]. Within the Jurisdiction A's dataset, the three most common 3-digit SSN prefixes account for half (49.4%) of the population.

1.1.2 Names

Names are similarly unequally distributed across the population. The evidence that two rows might identify the same person is much weaker if the name is Mary Smith than if the name were Henrietta Raynott. But names are also much more complicated; there are lots of social and linguistic norms that must be taken into account. Nicknames are common in Western society and may not even resemble the original name or each other (e.g., Dick and Richard, or Beth and Liz). We can quantify the distance between two names with the Levenshtein distance metric, which counts the number of edits required to transform one name to the other. Within the dataset from Jurisdiction A, however, this appears to be a fairly limited issue within the set of records where the last name and SSN match exactly. In this restricted subset of matched identities, less than 6% had more than one edit to their first name.

A bigger problem, however, is in the last name. Within the United States, it's common for women to change their last name after marriage. Name changes are becoming less frequent, but still 78% of women changed their name after getting married in 2010-2013 [MW15]. The effect size here depends upon the total time that the data span, but assuming a large enough window, this would split the majority of women into two or more identities, drastically under-serving or over-counting their population, depending upon how the analyses are used. Again, we can see this effect in the real data by looking at the Levenshtein edit distance for exact record matches between SSN and first name, divided by sex. Unlike the first name data, there is a clear bump representing totally different names for both sexes. It is far more pronounced for women, however. Over 15% of the women with more than one record in this dataset have significant edits to their last name. Only 3% of men have such large edits to their last name.

There are other significant linguistic and cultural issues to consider when matching names. One prominent example is the Spanish language custom to use two surnames for both the paternal and maternal family names. Typically, the first surname is the paternal family name, whereas the

second is the maternal family name and sometimes omitted. In this context, forms and databases that simply request the "last name" are somewhat ambiguous. It's not clear if the individual should list both their surnames or just the paternal name. If someone else is entering the data, they might mistakenly record the first surname as a middle name. Occasionally, the two names get hyphenated. This leads to a greater variance in the last names of Hispanic individuals. That's precisely what's happening with the 3% of men that appear to have changed their last name in the above graph; they didn't change their last name, it just got entered inconsistently. The overwhelming majority of these men have Hispanic names.

Another strong cultural tradition is the practice of naming children with the same name as a parent, particularly for fathers and sons. In this case, all other fields could match (including points of contact like phone or address) except SSN and date of birth. Similarly, twins might have entirely identical information except their first name and the very last digit of their SSN (if they were born before 2011). Even their first names might be similar; it's common for parents to choose matching names. The three most common name pairings for twins in the US in 2010 were all very similar: Ella and Emma, Madison and Michael, and Jacob and Joshua [Adm11].

2 Approaches to deduplication

Perhaps the simplest and most robust deduplication method is simply ensuring that the data are linked to the correct individual at the time of contact, making it possible to confirm a match with the person in question. Unfortunately, this isn't always possible, and it's definitely not possible when combining multiple datasets from independent sources together. To solve this problem, many jurisdictions have implemented algorithms of varying levels of complexity.

The simplest method is to perform exact matches by SSN. This common approach, however, leaves a lot to be desired; it's not possible to account for typos, and the missingness isn't likely to be uniformly distributed, so the matched individuals probably won't be representative of the population at large. It ignores lots of other information that can be used to augment the analysis.

This is largely the approach Jurisdiction B takes, but they additionally augment that with information from the criminal justice system. When an individual is booked into jail, their fingerprints are taken and those fingerprints are used to link them to any previous records (if there are any). This allows them to accurately identify individuals without SSN as well as individuals with aliases, but its effectiveness is limited to the criminal justice system.

So some jurisdictions use all of these identifying fields along with some heuristic in weighting the match importance between them. Metrics such as the Levenshtein distance, phonetic similarity scores (SOUNDEX), or anagram checks can be used for fuzzy name matching, and SSNs and dates can be compared digit by digit.

Jurisdiction C uses such a system. They assign points to each field, scaled by how well of a match it is. Social security numbers are worth 27 points, with three points for each digit. The last name is worth 50 points and the first name worth 40 points, with partial credit for matching characters in any order. The gender is worth 90 points, and the date of birth is worth 70 points, with partial credit given for day, month and year matches. All the points are added together, and if more than 47 points are deducted then the pair of records are considered separate identities. Notice how, even with a preponderance of evidence with many exact matches, a single field typo (such as a mistake in the gender field) can cause a false split. So they've explicitly accounted for this in some special cases. If the gender field isn't a match but all other fields are exact matches or if it is missing entirely, then the full 90 points are still given. Similarly if the date of birth is missing but all other fields are exact matches, the full score is given for that field. Finally, the phonetic similarities of the first and last names are considered along with common nicknames of the first name; any pairs of records with large differences are discarded.

This system is very conservative in matching records. A large difference in the last name field loses enough points that, even if all other fields are exact matches, the two records will be considered separate. Not only does this miss many common errors with hispanic last names, it fails to link all women who have changed their last name. At the same time, though, it's lenient enough that some of the other tricky cases might earn enough points that they'd be erroneously merged. Twins born before 2011 would only miss 3-6 points from their SSNs and up to 40 points from their first names, putting them inside the 47 point cutoff so long as their first names are phonetically similar. Same-named father-son pairs would definitely pass the phonetic name check, but if they also earn more than half of the credit from the 97 points for date of birth and SSN they'd also be

falsely merged. So a linear algorithm here isn't sufficient. In some cases, large differences should be ignored (last name changes), but in other cases very small differences should be amplified (SSNs between twins), and special cases need to account for a preponderance of evidence across many other fields.

3 Our solution: probabilistic matching

Instead of designing static heuristics, we can use modern machine learning frameworks that extract multiple features from each field and combines them based upon real examples that we manually identify. This is the state-of the-art approach that has been developed over the past 10 years and is becoming a very robust approach to matching. By combining this with an exact match post-processing step, we have built a well-performing and easy-to-use tool that is extensible and accurate.

The first step is creating two sets of record pairs that a human has manually labeled as being either known to be matches or known to be distinct identities. While this process can be somewhat tedious and time consuming, it can be sped up by carefully choosing the labeled set of examples to be maximally informative. This machine learning technique, known as active labeling, prompts the user to label specific pairs of records that are ambiguous and will be more beneficial to know than a pair chosen at random. This further has the advantage that it is specifically tailored to the population in question. If, for example, there was a large population of second-language speakers with a different naming custom with additional complications with respect to merging records, those cases would be presented to a human for manual labeling. This helps prevent systematic biases in the matching algorithm when applied to the population in question. To accomplish this, we have used an open source library, `dedupe`[GE16], which is based on Mikhail Yuryevich Bilenko's dissertation[Bil06].

We have used and extended this library to create a stand-alone program, `pgdedupe`, that makes it easy to deduplicate large PostgreSQL databases. The database table and columns to use as evidence are all specified within easy-to-use configuration files, along with the choice of how each column should be interpreted (e.g., as general strings, names, or addresses). All rows with exactly identical columns are merged as a pre-processing step, and a custom filter may be specified to ensure that only records with sufficient non-missing data are considered for potential matches. The records are divided into blocks to reduce the number of required comparisons, ensuring that a sufficient percentage of the labeled examples get blocked together. Within each block, the model scores all record pairs and then they are clustered into groups. Finally, we've added the ability to merge the clusters of linked records found by the probabilistic match using an exact record linkage across one or more columns. More details on our solution, its source code, and the synthetic dataset can be found at `https://github.com/dssg/pgdedupe`.

3.1 Real data results

While there is still room for improvement, our initial results are promising. We're able to robustly match records that are missing social security numbers; the matched percentage is relatively constant between clusters that contain a record with a missing Social Security Number as compared to those that are not.

The probabilistic matching also allows for more leniency with regard to the last name and name changes. Within Jurisdiction A's dataset, there were 11,922 women with exactly matching first names and social security numbers, but a significantly different last name. Our algorithm was able to match 12,227 women with significantly different last names, 300 more than those found by exact first name and SSN match alone.

3.2 Synthetic results

It is easier to validate the results with a synthetic dataset where there is a known ground-truth. We generated a dataset that contains 20 000 true identities, each with a varying number of records (sampled from an exponential distribution with a mean of 20 and shifted by one to ensure at least one record per identity). The total dataset contained approximately 410 000 records with names, SSNs, dates of birth, genders, races, and ethnicities. The fields were artificially altered within each record such that they reflected many of the challenges listed above, including typos, missingness,

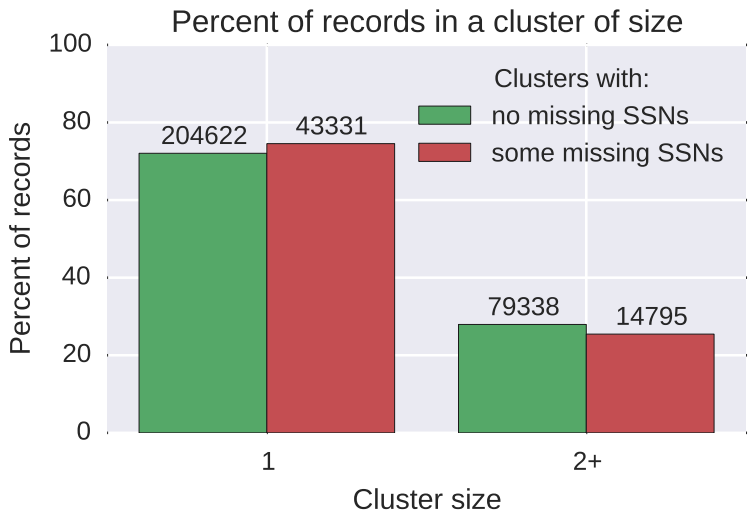

Figure 2: The existence of a record with a missing SSN did not significantly affect the ability to match to another record in the real dataset.

name changes, and the complications of two surnames. Additionally, 5% of the population was constructed such that they had a twin with the same last name, birth date, and demographic data, and their SSN only varied in the last digit.

A relatively small and conservative set of 90 labeled pairs (40 positive and 50 negative examples) were used to train the probabilistic matching. Record linkage was performed after the probabilistic matching to merge clusters where both last name and date of birth or SSN and date of birth were exact matches.

Superdeduper linked the rows with 97% accuracy, with 0.25% false positives and 2.1% false negatives. 0.75% of the rows were excluded from consideration due to having too much missing data for a conclusive match.

False negatives occur when multiple rows from the same identity fail to get clustered together. The effect this has is strongly dependent upon the sizes of the clustered records; missing just one or two records from a large cluster is much better than more evenly splitting an identity into two or more clusters. Importantly, 93% of the identities had a primary cluster that contained over 90% of their true records.

False positives, on the other hand, occur when two distinct identities get linked together into the same cluster. Only 118 true identities (less than 0.6% of the total population) were erroneously merged together. Most of these errors occurred between twins, but it's notable that twins accounted for 5% of the overall population. As such, even the vast majority of twins were correctly separated into distinct identities.

Some rows were excluded from analysis entirely due to having too much missing data for a conclusive match to be found. In this case, we excluded rows where both the SSN and date of birth were missing. By construction, no names were missing in this dataset. This ensured that there was at least a secondary datapoint beyond the name that allowed for a verification of the identity. A total of 8 true identities were entirely ignored due to having no rows that satisfied this criterion.

Even more important than the error rates per row is how this affects the analyses and resulting services on an individual interaction level. Given an arbitrary row from this dataset, 99.0% are correctly linked to a cluster where the majority of rows come from the correct identity. Of those clusters, 99.5% consist entirely of rows from the correct identity.

For population-level analyses, the conservative linking and preference toward false negatives led to a slight inflation in the number of unique identities that were found. A total of 25 993 clusters were identified, but there were only 20 000 true identities. This is a 30% inflation over the true number of unique individuals that were served. It also means that the average number of services per individual is artificially depressed.

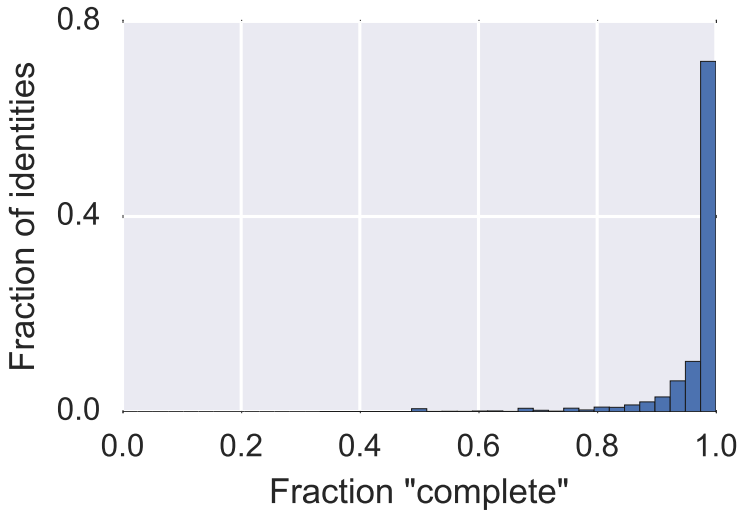

Figure 3: A large proportion of the true identities were in clusters that contained the vast majority of rows from that identity in the synthetic dataset.

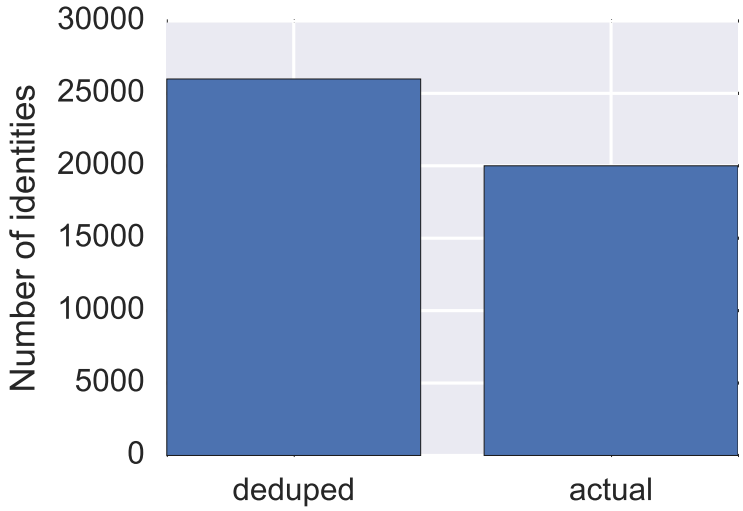

Figure 4: Approximately 30% more unique identities were found than actually occurred in the synthetic dataset.

4 Conclusion and future steps

We have developed a simple tool that enables simple specification and execution of deduplication over large PostgreSQL database tables. By combining both record linkage and probabilistic models that incorporate domain expertise in their labeled examples, our solution has demonstrated good performance and accuracy on both synthetic and real-world datasets.

We are working to improve these results. Key improvements include additional data-specific comparison metrics for dates of birth and identification numbers with known patterns (like Social Security Numbers). Further extensions include better incremental matching and stability for datasets that are periodically updated and more consistent accuracy over multiple runs.

References

[Adm11] Social Security Administration. Popular names for twins born in 2010. April 22 2011.

[AG09] Alessandro Acquisti and Ralph Gross. Predicting social security numbers from public data. *Proceedings of the National academy of sciences*, 106(27):10975–10980, 2009.

[Bil06] Mikhail Yuryevich Bilenko. *Learnable similarity functions and their application to record linkage and clustering.* PhD thesis, The University of Texas at Austin, August 2006.

[GE16] Forest Gregg and Derek Eder. Dedupe. `https://github.com/datamade/dedupe`, 2016.

[MW15] Claire Cain Miller and Derek Willis. Maiden names, on the rise again. June 27 2015.

[VS02] Denton Vaughan and Fritz Scheuren. Longitudinal attrition in sipp and spd. *US Census Bureau, SIPP Working Paper*, 242, 2002.

[WWL+12] Charmaine Smith Wright, Mark Weiner, Russ Localio, Lihai Song, Peter Chen, and David Rubin. Misreport of gestational weight gain (gwg) in birth certificate data. *Maternal and child health journal*, 16(1):197–202, 2012.