Person Linking for the State of Minnesota's P20W Statewide Longitudinal Data System

May 1, 2015

iBusiness Solutions, Inc.
7300 Metro Blvd, Suite 590
Minneapolis, MN 55439
www.ibusiness-solutions.com

Tim Brands
612-730-7404
tbrands@ibusiness-solutions.com

Minnesota Office of Higher Education
1450 Energy Park Drive, Suite 350
Saint Paul, MN 55108
www.ohe.state.mn.us

Meredith Fergus
651-259-3963
meredith.fergus@state.mn.us

Minnesota Information Technology Services
658 Cedar Street
Saint Paul, MN 55101

Minnesota Department of Education
1500 Highway 36 West
Roseville, MN 55113
www.education.state.mn.us

Kara Arzamendia
651-582-8599
kara.arzamendia@state.mn.us

Minnesota Department of Employment and Economic Development
1st National Bank Building
322 Minnesota Street, Suite E-200
St. Paul, MN 55101
www.mn.gov/deed

Rachel Vilsack
651-259-7403
rachel.vilsack@state.mn.us

Table of Contents

Table of Figures

Table of Tables

Overview: P20W Data Warehouse – Minnesota's Statewide Longitudinal Education Data System (SLEDS)

The P20W data warehouse will provide a valuable data repository for educational and employment analytics for many audiences to improve the educational experiences and outcomes of individuals from early childhood through college graduation and increase the likelihood of meaningful, related, and sustained employment. The P20W data warehouse serves as an umbrella structure for three separate but overlapping data projects – Early Childhood Longitudinal Data System (ECLDS, birth to grade 3 data), Statewide Longitudinal Education Data System (SLEDS, kindergarten through postsecondary and workforce), and Workforce Data Quality Initiative (WDQI, education and work).

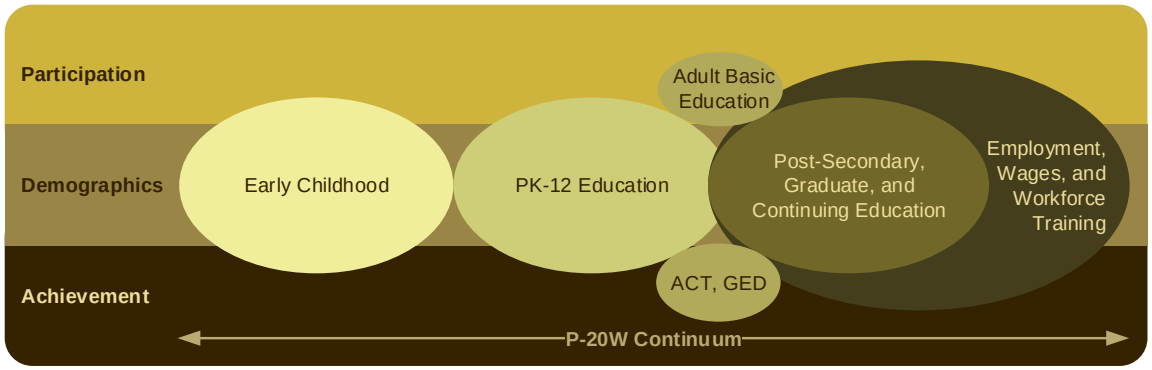
The P20W data warehouse builds on the vision created by the SLEDS Governance committee. SLEDS (and the P20W data warehouse) brings together data from education and workforce to:

- Identify the most viable pathways for individuals in achieving successful outcomes in education and work,
- Inform decisions to support and improve education and workforce policy and practice, and
- Assist in creating a more seamless education and workforce system for all Minnesotans.

The data provided to P20W can be viewed in two ways:

- Where on the P20W continuum does a data set fall, from birth through employment, and
- How does the data within the data set help paint a picture of participation and outcomes along the P20W continuum.

Figure 1: P20W Continuum



Purpose of Person Linking

The P20W SLDS data warehouse encompasses the longitudinal life of an individual from early childhood, elementary and secondary education, up to and through higher education and employment, including the various paths people take through education, to employment, and back and forth.

In order to produce a longitudinal view of a person, all records for a given individual must be matched and linked across dozens of data collection systems that capture varying data elements for personally identifying information (PII) and contain varying levels of quality and accuracy of that PII. As a result, person linking is part science and part art form that utilizes probabilistic matching algorithms to match and link people longitudinally with a relatively high degree of confidence, but always with an estimated error rate.

Evolution of Person Linking

The P20W system was originally designed for loading and linking datasets for the SLEDS data warehouse exclusively. The Statewide Longitudinal Education Data System (SLEDS) was the initial effort to bring together data primarily from K12 education, higher education, and workforce.

However, as other similar longitudinal and cross-agency data warehouse initiatives arose, it was identified that the same system could be used to integrate the datasets. The P20W loading and linking engine evolved to be able to accommodate the specific needs of differing warehouses while re-using the efforts of ETL flow and probabilistic person matching.

This person linking white paper provides the chronological evolution of the State of Minnesota's person linking engine from SLEDS only to P20W, SLEDS, and ECLDS (Early Childhood Longitudinal Data System). Terminology and references to SLEDS and P20W are retained in their historical context throughout this document.

Databases

Several databases are utilized to load dozens of data sources, match and link persons across those data sets, and ultimately feed appropriate data to multiple data warehouses and data marts for reports and analytics. The figure below depicts a high-level view of the databases and the flow of data into and through these databases.

There is much more design and architecture information related to these databases. Below are a few pieces of additional information as it pertains to the subject of person matching and linking.

P20WStage

P20WStage is the P20W staging database in which incoming datasets are initially loaded, validated, and linked before passing into the P20W operational data store (P20WODS). Each distinct data source has its own staging table and other supporting tables. P20WStage contains unlinked, or more accurately pre-linked data, and P20WStage contains PII.

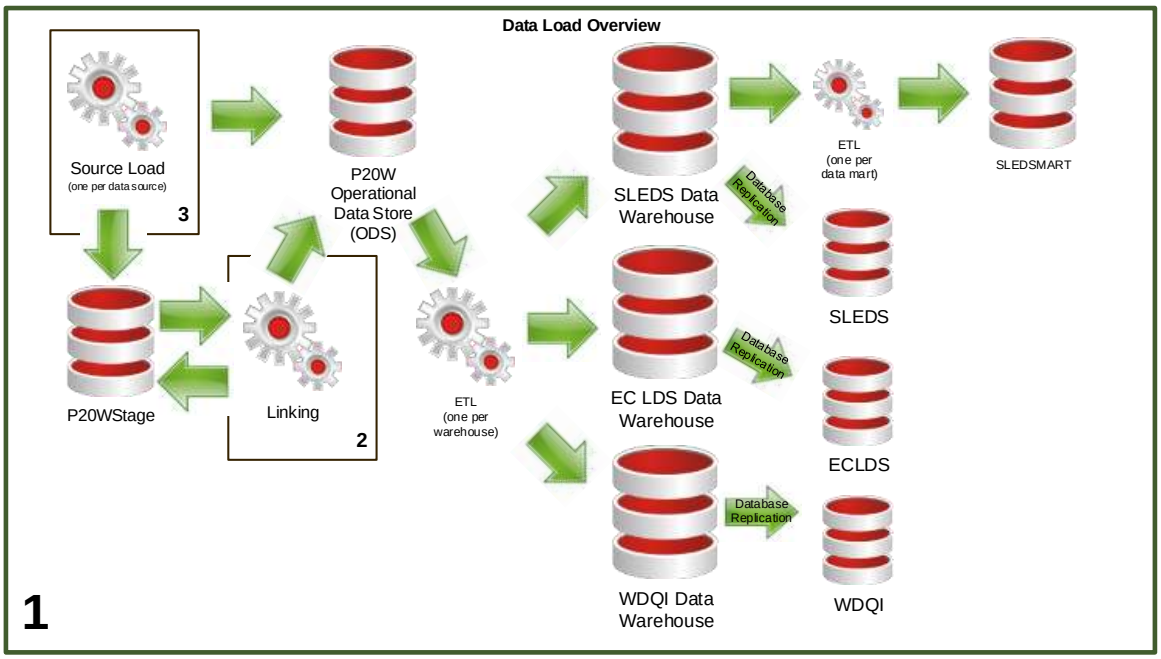
P20WODS

P20WODS contains normalized data after it has been validated and linked. Dimensions exist here as well. P20WODS retains PII.

Warehouse Databases

The ultimate destination for source data is one or more warehouses, followed by downstream data marts where necessary. Both SLEDSDW and ECLDSDW have distinct ETL processes that load the data required for each warehouse. Additionally, each warehouse has its own unique personal identification numbers that have been obfuscated from P20WODS. None of the data warehouses or data marts contains any data elements of PII.

Figure 2: Data Load Overview



Terminology

SLEDS – Statewide Longitudinal Education Data System

ECLDS – Early Childhood Longitudinal Data System

WDQI – Workforce Data Quality Initiative

PII – data elements that are considered to be personally identifying information, including name, date of birth, social security number, and others.

MARSS# – a unique statewide student identifier generated and maintained by the Minnesota Department of Education

SSN – Social Security Number

DOB – date of birth

Positive match, or match – a match, or link, made between two records with sufficient quantity and quality of PII to accurately identify that both records belong to the same person

False positive – a positive match and link made between two records that should not have been made

False negative – a positive match and link not made between two records that should have been made

SourcePerson – a unique set of PII within each data source

ReportedPerson – a unique set of PII regardless of the data source(s) from which it originated

Initial Testing for Person Linking

Rules Used

To initially test rules for linking records, sample data was selected from three sources:

- 50,000 records from K-12 enrollment
- 71,196 records from Adult Basic Education (ABE)
- 75,000 records from higher education enrollment

These records were loaded and linked, always in the same order to ensure consistency, and then evaluated. To determine the quality of matches the set of rules produced, a subset of 11,700 matches were manually reviewed to determine whether the linking was correct or not. The rules were then revised and the process started over.

We started with the same rules utilized by previous research conducted by MDE, with the exception that instead of SoundEx we used the Microsoft SSIS fuzzy lookup transform, which uses an algorithm Microsoft refers to as ETI. It has many advantages over SoundEx we hoped would show an improvement to the matching.

One of those advantages is that ETI parses a name field into tokens and matches tokens instead of the whole field allowing us to identify records where first and last names were entered in reverse order. It also scores individual tokens based on its value frequency, so if there are many Maria's in the data the weight of the "Maria" token goes down to lessen the likelihood of false positive matches.

A more significant difference from the previous research is the fact that gender was removed from the data elements considered to be PII. With data sources exhibiting few personally identifying columns, we often could not compare either SSN or MARSS# (Minnesota's state K12 student identifier) between a pair of records. As a result, if just one other field, like gender, had a typo or error, it would fall through as a false negative. Of the 11,700 total matches in the initial test set, 61 matches (134 rows) were found to be valid matches that were made as a direct result of removing gender from the PII.

The rules from the previous MDE research did not include SSN as PII so the set of rules used on the initial test set for SLEDS was also extended to include SSN.

The following set of rules resulted from this iterative process:

Table 1: Initial Research Linking Rules

MARSS #	SSN	Last Name	First Name	Middle Initial	Data of Birth
Exact Match	Ignore	ETI	ETI	Exact Match	Exact Match
Ignore	Exact Match	ETI	ETI	Exact Match	Exact Match
Exact Match	Ignore	Ignore	ETI	Ignore	Exact Match
Ignore	Exact Match	Ignore	ETI	Ignore	Exact Match
Exact Match	Ignore	ETI	Ignore	Exact Match	Exact Match
Ignore	Exact Match	ETI	Ignore	Exact Match	Exact Match
Exact Match	Ignore	Exact Match	Exact Match	Exact Match	Ignore
Ignore	Exact Match	Exact Match	Exact Match	Ignore	Ignore
Ignore	Ignore	Exact Match	Exact Match	Ignore	Exact Match

Note, a rule to identify exact name matches is not needed because ETI will return a score of 1.0 from one of the first two rules above, which is equivalent to an exact match.

Estimated Error Rates of SLEDS LDS2 Matching

False Positives

False positives occur when a positive match and link is made between two records that should not have been made. False Positive searches included queries returning a broad range of possibilities for false positives by filtering for 3 or more columns on a match that disagreed (having a comparison score of less than .5 in the JaroWinkler algorithm, discussed below). In addition to the 'fuzzy' false positive lookup, candidate results also included matches for which either the MARSS #'s did not match (and were not null) or the SSNs did not match (and were not null).

Of the 815 rows hand-checked, 683 were confirmed as valid matches, 126 unknown, and 6 definite false positives. Without verification by a data expert, an absolute statistic could not be reached. However, almost all of the unknown potential false positives lie in the set of matches containing records that have exact matches on first name, last name, and date of birth, with a disagreeing MARSS#. Thus, the false positive rate can be expected to be low. We could allow a false positive rate within this set of up to 1 in 10 and still retain an overall false positive rate of 1/10,000 for the entire set.

False Negatives

False negatives occur when a positive match and link is not made between two records that should have been made. 11 different queries tailored to search the fringe of the linking rules coverage were constructed to pull in potential false negatives. The team anticipated a much higher false negative rate due to the quality of the incoming data. Of 91 hand-checked rows, 12 were confirmed as false negatives, with 6 unknown.

Queries were also done to check for similarities in rows where it was not possible to compare either the MARSS or SSN. These queries returned over 2300 results, none of which could be verified by a non-data expert, as first name, last name, and date of birth are the only fields that can be compared (the queries allowed for error in gender and fuzziness in one of the other columns, catching potential matches that would not be caught by the 'weakest' matching rule).

A much higher false negative rate can be expected, by sheer number of results for these queries, and due to the fact that there are many instances where neither MARSS# nor SSN can be compared between two records, resulting in only matching on exact matches of all other fields (aside from gender). This leaves us with a largely unknown false negative error rate, which could be anywhere from 1 in 2,000 to 1 in 100.

Caveats

There are some other caveats that must be noted when considering these numbers and their grounds. The main reservation lies in the fact that the data being tested is not necessarily the best representation across all data sources and across all years. This could be the main cause of 183,197 SLEDS IDs coming out of the 196,196 records loaded. Additionally, the poor quality of a significant portion of the test data certainly has affected the linking quality in measurable ways, and may continue to affect future loads.

Moving Forward with Linking Rules

Linking is an ever-fluid process given the context of this project. As SLEDS continues to grow, rules will be tweaked and added to maintain an acceptable balance between false positives and false negatives. The list of the rules will be evaluated and will continue to evolve with each data set added to SLEDS. The linking rules used by SLEDS are as follows.

Table 2: Linking Rules

Rule Number	Rule Description	Order Number
1	New Record	1
1.5	FullName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	2
1.6	LastName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	3
1.7	FirstName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	4
1.1	MARSS#_Exact, FullName_Fuzzy, DateofBirth_Fuzzy	5
1.2	MARSS#_Exact, LastName_Fuzzy, DateofBirth_Fuzzy	6
1.3	MARSS#_Exact, FirstName_Fuzzy, DateofBirth_Fuzzy	7
1.8	FullName_Fuzzy, SSN_Exact	8
1.4	MARSS#_Exact, FullName_Fuzzy	9
1.11.1	LastName_Exact, FirstName_Fuzzy, DateofBirth_Exact, SSN_NULL_OR_Match_90%_or_more_to_other_SSN	10
1.12.1	LastName_Fuzzy, FirstName_Exact, DateofBirth_Exact, SSN_NULL_OR_Match_90%_or_more_to_other_SSN	11
1.11.2	LastName_Exact, FirstName_Fuzzy, DateofBirth_Exact, MARSS_NULL_OR_Match_90%_or_more_to_other_MARSS	12
1.12.2	LastName_Fuzzy, FirstName_Exact, DateofBirth_Exact, MARSS_NULL_OR_Match_90%_or_more_to_other_MARSS	13
1.9	LastName_FirstName_Exact, DateofBirth_Exact	14
1.10	FirstName_LastName_Exact, DateofBirth_Exact Against LastName_FirstName_Exact, DateofBirth_Exact	15
1.13	LastName and DateofBirth are Exact and FirstName using NameGroup	16
1.14	Exact Match on SSN and FirstName	17

Linking Gaps

Background

Each source record contains personally identifying information. As records are evaluated for linking, that PII is compared to the set of PII on records previously loaded and linked to determine:

1. If the incoming record can be linked to an existing person group, or
2. If the record starts a new person group.

The extent to which two records may be compared and the quality of the linking outcome depends upon both the amount of personally identifying information provided and the overlap between two given source records. For example, two K12 Enrollment records can be compared quite thoroughly, because they provide several personally identifying fields, all of which overlap, including a strong identifier like MARSS#.

Table 3: Linking Gap Example

Data Source	First Name	Last Name	Date of Birth	SSN	MARSS#
K12 Enrollment	Jane	Smith	1/1/1990	------	1300000000000
K12 Enrollment	Jane	Doe	1/1/1990	------	1300000000000
Workforce	Jane	Doe	------	987654321	------
Higher Ed	Jane	Doe	1/1/1990	987654321	

Linking gaps arise when there are not enough overlapping data elements between two records to even make a comparison. For example, K12 Enrollment records cannot be matched to Workforce records with any confidence, because they only share the first and last name identifiers. Consequently, the P20W warehouse will never directly link a K12 Enrollment record to a Workforce record. This will be true between any other datasets that exhibit a large gap in personally identifying information.

Bridging the Gap

However, this does not mean a K12 record will never be identified as the same person in a Workforce record. A K12 record can still link to a person's Workforce record, but only indirectly through a bridge match.

Bridge Matching

Suppose a K12 record is loaded, then a Higher Education record is loaded and linked that has enough PII to match to the K12 record. If the Workforce set is then loaded, the Workforce record can link directly to the Higher Education record and indirectly to the K12 record. The Higher Education dataset is a bridge class for the linking gap between K12 Enrollment and Workforce. Through this strategic order of linking datasets, P20W is able to at least give records an opportunity to link across an identified linking gap.

Table 4: Dataset Classifications for the K12-to-Workforce Linking Gap

K12 Type Datasets	Bridge Class Datasets	Workforce Type Datasets
K12 Enrollment K12 Assessments	ACT Higher Education Enrollment Higher Education Completion Adult Basic Education GED	Workforce Wage Data

Caveats and Implications

While there are methods to counteract linking gaps, there is still an impact on research: bridge matching can help us shrink the gap, but it still exists. Without bridge matches, you would never be able to ask SLEDS: "who was employed during high school?" With bridge matches, you can ask that question, but your answer will only contain those individuals who have a record in a bridge class record set (for example, those who have completed their GED, enrolled in college, or participated in an ABE program). Even with bridge classes, we still cannot ask the question "who went straight from high school into the workforce without participating in ABE, getting their GED, or going to college?"

Linking gaps also have an impact on data loads: when utilizing bridge matching, the order of data loads is extremely important. A bridge match can only happen if bridge records for a certain individual are loaded before any records on either side of the gap. This has implications for both historical data loads and future loads, especially when bridge records are naturally generated after records on either side of the gap, such as a Higher Education record being naturally generated after a Workforce record because a person went from K12 directly into the Workforce and then later into Higher Education.

The order in which data is loaded is the result of maximizing two main priorities when linking data:

1. That the data with the highest quality and highest completeness of personally identifying information be loaded first, and
2. Bridge class datasets must be loaded before data on either side of the linking gap.

Conclusions of Phase 1 Linking

While the SLEDS team has worked to reduce the effect of linking gaps, additional options to further reduce them exist that can be evaluated in future releases:

1. **Highly recommended:** Given the nature of bridge matching (see previous sections for details), bridge matching will be much more effective if the order of the datasets relative to each other remains in absolute. This can be accomplished only by reloading all historical data each time data is added from existing data sources in order to retain the optimal linking order.

2. Adding a mapping of primary identifiers to other missing personally identifying information. For example, known SSNs could be sent to another data source (e.g. Department of Motor Vehicles) to add additional PII data elements (e.g. Date of Birth) that a P20W data source does not capture
3. Directly adding a PII field to a dataset at collection time – for example, K12 enrollment requesting SSN's from students

Evolution to a P20W Linking System

Background

The SLEDS Linking engine was identified as a potential solution for linking data for two other longitudinal warehouse initiatives: Early Childhood Longitudinal Data System (ECLDS) and Workforce Data Quality Initiative (WDQI). Upon the approval for ECLDS to use the SLEDS linking engine, the SLEDS staging area became a P20W staging area where data from multiple data sources is linked and then sent to their respective warehouses with an obfuscated PersonID specific to the destination warehouse.

As a part of the code changes that allowed for handling data streams for multiple initiatives, a number of potential improvements, both from a performance perspective and a linking quality perspective, were identified, approved, and implemented for the newly designated P20W Linking engine. The changes in this release resulted in a large increase in linking efficiency by reducing the time for a full link by more than 90%.

Separating Loading from Linking

One of the architectural changes implemented to produce a more efficient linking engine was to separate the linking process from the loading process, meaning:

1. Data can be loaded incrementally, in any order, at any time, and
2. Data can be linked and re-linked at any time.

As a result, data are no longer linked as files are loaded so data can be pre-sorted by class (see Linking Gaps) and by quality before linking to maximize higher confidence rule usage.

The order of the data was also refactored not to be source specific. Records with the most PII are matched first, with the least matching last. This will affect linking but should not affect quality. Spot checking by rule validated this expectation.

Moving from SourcePerson to ReportedPerson

Previous releases of the SLEDS linking engine were based on a SourcePerson, a unique set of PII within each data source. This release of the P20W linking engine introduced the concept of a ReportedPerson, a unique set of PII regardless of the data source(s) from which it originated.

ReportedPersons are now derived from data as it loads, effectively eliminating the need for 'exact' match rules to run during linking. Rather, all exact matches are evaluated on data load, substantially reducing the amount of linking that needs to be performed during the linking process.

Furthermore, the remaining ReportedPersons needing to go through the linking engine use whichever rules each ReportedPerson qualifies for based upon the available PII on that record. In other words, if a ReportedPerson does not contain an SSN, all rules using SSN are ignored for linking this record. This replaced the source-specific linking rules and will allow new datasets to be added in the future with no coding and no new rules as long as the new dataset does not introduce any new data elements of PII.

Refactor Analysis

Validate Linking Quality

ReportedPerson Counts

To compare counts between the previous SLEDS linking engine and the current P20W linking engine, ReportedPerson must be counted per source in the new release. So, one unique ReportedPerson which has exact PII in 3 different sources was counted 3 times to compare it with the previous SourcePerson. This count produced 16% more ReportedPersons than SourcePerson. This was not completely unexpected because previously custom "exact" match rules were used per source. Most of these used Fullname to find a unique record where FullName did not contain middle name. The addition of middle name was expected to produce more records.

Person Counts

Included with refactor changes was an addition of 2 rules (1.8.1 and 1.8.2) to include name group index checks with SSN. This should increase the possibility of matches and therefore reduce the number of P20W Person records created.

There was a decrease of 37,883 Person records. The addition of the 2 rules accounted for 22,533 of the difference leaving 15,350 (0.22% of total Person records) that could be explained by the improvements to rules and pre-sorting records by quality of data.

Table 5: Match Rates by Rule

Description	Matches
New Record	6,798,462
FullName_Fuzzy, SSN_Exact	4,174,115
MARSS#_Exact, FullName_Fuzzy, DateofBirth_Fuzzy	2,052,156
LastName_Exact, FirstName_Fuzzy, DateofBirth_Exact, SSN_NULL_OR_Match_90%_or_more_to_other_SSN	982,980
FullName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	322,364
Exact Match on SSN and FirstName	256,070
SSN_Exact, LastName_Exact, FirstName_NameGroup, MI_null_or_exact	19,168
MARSS#_Exact, FullName_Fuzzy	13,254
LastName and DateofBirth are Exact and FirstName using NameGroup	12,185
MARSS#_Exact, FirstName_Fuzzy, DateofBirth_Fuzzy	10,720
FirstName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	9,898
LastName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	4,650
LastName_Exact, FirstName_Fuzzy, DateofBirth_Exact, MARSS_NULL_OR_Match_90%_or_more_to_other_MARSS	3,950
SSN_Exact, FirstName_NameGroup, MI_null_or_exact	3,340
MARSS#_Exact, LastName_Fuzzy, DateofBirth_Fuzzy	1,972

Description	Matches
FirstName_LastName_Exact, DateofBirth_Exact Against LastName_FirstName_Exact, DateofBirth_Exact	999
LastName_Fuzzy, FirstName_Exact, DateofBirth_Exact, MARSS_NULL_OR_Match_90%_or_more_to_other_MARSS	393
LastName_Fuzzy, FirstName_Exact, DateofBirth_Exact, SSN_NULL_OR_Match_90%_or_more_to_other_SSN	238
LastName_FirstName_Exact, DateofBirth_Exact	4

Analysis of Match Rates by Agency Identifiers

*** JW and ME abbreviations refer to the JaroWinkler and the MongeElkan string similarity algorithms respectively.**

MARSS# Research

There are approximately 49 trillion statistically possible reported person pairs in P20W. Of these, 7,695,577 of them have the same MARSS. There are 18,972 unique record **pairs** (involving 24,586 unique ReportedPersons) having the same MARSS but they were not matched by the P20W linking engine (0.2% of the 7 million MARSS match pairs). Of these 18,972 pairs:

- 135 have the same first name and the same last name
- 382 have the same first name only
  - 52 have a JW(LastName) > .8
  - 91 have a MongeElkan(FullName) > .75
  - 34 satisfy both of the above conditions
- 1,383 have the same last name only
  - 109 have a JW(FirstName) > .8
  - 299 have a ME(FullName) > .75
  - 75 satisfy both of the above conditions
- 17,072 have neither the same first name nor the same last name
  - 17 have ME(FullName) > .95
    - These look like possible multiple first/last names. Only a couple instances had a different date of birth, although that may be evidence of an incorrectly assigned MARSS# from looking up a record on first/last name and DOB.

SSN Research

There are approximately 49 trillion statistically possible reported person pairs in P20W. Of these, 19,055,930 of them have the same SSN. 18,294,777 of these pairs were matched by the P20W linking engine. There were 761,153 unique record **pairs** (involving 770,746 unique reported persons) having the same SSN but were not matched by the P20W linking engine (4.16% of the 18 million SSN pairs). Of these 761,153 pairs:

- 444 have the same first and last name

- 195,000 have a MongeElkan(FullName) > .75
- 21,718 have the same first name only
  - 2,800 have a JW(LastName) >=.8
  - 5,161 have a MongeElkan(FullName) > .75
  - 2,427 satisfy both of the above conditions
- 142,600 have the same last name only
  - 33,000 have a JW(FirstName) >= .8
  - 69,182 have a MongeElkan(FullName) > .75
  - 22,014 satisfy both of the above conditions
- 596,391 have neither the same first name nor last name
  - 16,059 have a MongeElkan(FullName) > =.95
    - These are largely cases where the first and/or last names have been transposed to one or the other field. For example, both last and first names end up in the same field for some reason, the first and last names are transposed between the pair, etc. Included also are cases where a partial last name is used in one record and a complete last name is used in the other (ex. Hernandez vs Hernandez-Perez)
  - 390,266 have a JW(FirstName) < .5
  - 321,566 have a JW(LastName) < .5
  - 204,378 have a JW(FirstName) and JW(LastName) < .5

Implications

- Many SSN matches that could possibly be made are not being made (4.16% of 18 million SSN pairs, compared to 0.2% of 7 million MARSS pairs)

Linking Issues Identified by Analysis of Match Rates

- The name group index is known to be missing several diminutive names. Another name index, taken from work done in Ancestry.com and WeRelate.org is a possible replacement.
- Rules that do fuzzy full name comparisons are currently only using MongeElkan, which is limited to a longest common substring similarity comparison. As transpositions and misspellings are frequent, combining JaroWinkler is seen as very likely to be an improvement.
- A semantic error in the 1.11.x and 1.12.x rules were allowing matches to happen in one rule that was meant to be prevented by the other. This was also the cause of almost all 1.9 linkages being enveloped in the problematic rules. A condensing of the 1.11.x and 1.12.x rules into two 1.11 and 1.12 rules will fix this issue and prevent a portion of the MARSS and SSN disagreements.
- With the changes to 1.11 and 1.12, rule 1.9 only matches records that have first name, last name, and date of birth exact that also have a disagreeing (not null) MARSS **and** disagreeing (not null) SSNs. This is going to be an almost guaranteed false positive that pairs people who just happen to have the same name and date of birth (in the most

recent link, this happened 4 times, to records with names like Carlson, Brown, and Mohammed, which are very common). This rule will be removed.

- There is no way to empirically test the linking quality other than to compare matches that disagreed with strong identifiers (SSN, MARSS, etc). A reference set is needed to fully perform a valid analysis.

Linking Adjustments: Reasoning, Research, and Impact

DOB Fuzzy Comparison

Reasoning

The current algorithm for fuzzy DOB comparison is a JaroWinkler edit distance on the full DOB string. In spot checks, several common DOB errors were identified, including DOBs with a single digit difference (typos), that put the DOB comparison below the threshold for being considered a fuzzy match. However, moving the thresholds to allow these common errors to come through as matches created too many transpositions in the DOB to also be considered a match. An alternate method of comparing DOBs was identified as necessary.

Research

To identify the most common DOB errors, identifiers such as SSN and MARSS having multiple DOBs in the system were analyzed.

1. SSN
    a. 15,966 SSN's having more than one DOB
        i. Identified differences:
            1. Day off by 1
            2. Month off by 1
            3. Year off by 1
            4. Month/day transposed
            5. Digit transposition in day
            6. Digit transposition in month
            7. Digit transposition in year
            8. Dates very different
        ii. Distribution (55,573 date pairs having same SSN):
            1. 8,014 only day different
                a. 4,938 off by 1 digit
                b. 3,076 off by both
                c. 442 cases of transposed digits
                d. 1,043 sequential days
            2. 1,799 only month different
                a. 1,524 off by 1 digit
                b. 255 off by both
                c. 186 cases of transposed digits
                d. 508 sequential months
            3. 3,513 only year different

a. 3,040 off by 1 digit
b. 387 off by 2 digits
c. 69 cases of transposed digits
d. 1,085 sequential years

4. 667 cases of transposed month/day

2. MARSS
   a. 43,051 MARSS's having more than one DOB
      i. Identified differences:
         1. Day off by 1
         2. Month off by 1
         3. Year off by 1
         4. Month/day transposed
         5. Digit transposition in day
         6. Digit transposition in month
         7. Digit transposition in year
         8. Dates very different
      ii. Distribution (118,663 DOB comparisons having same MARSS)
         1. 94,592 only day different
            a. 47,296 off by 1 digit
            b. 10,431 off by both
            c. 2,643 cases of transposed digits
            d. 12,449 sequential days
         2. 16,294 only month different
            a. 14,092 off by 1 digit
            b. 2,202 off by both
            c. 308 cases of transposed digits
            d. 7,899 sequential months
         3. 23,733 only year different
            a. 22,894 off by 1 digit
            b. 645 off by 2 digits
            c. 32 cases of transposed digits
            d. 17,449 sequential years
         4. 7,334 cases of transposed month/day

Recommended Adjustment

Currently, the date of birth fuzzy comparison runs the JaroWinkler edit distance algorithm to compare dates of birth. This could be replaced with comparisons of the date parts for these specific errors, or used in combination. However, as transpositions of single digits between date parts was not identified as a common error, running an edit distance algorithm on an entire date string is less precise than intelligent checks for human error on date parts. The recommended adjustment is to perform single digit edits, transpositions, and month/day transpositions in DateOfBirth fields.

Technical Description

Rules using JaroWinkler(dateofbirth) > (threshold) will be modified to accept one of the following criteria:

- Successful Jaro(year)
- Transposition or single-digit difference on either month or day
- Transposition of entire month and day

Impact

For MARSS, allowing even just a single digit difference in the full date of birth could restore 62% of near matches (provided other identifiers such as name are exact or similar) that may otherwise be prevented by DateOfBirth differences. For SSN, this would pick up 17% of the near matches. Allowing sequential days would pick up an additional 10% for MARSS and another 1.8% for SSN. Allowing transposed months and days would pick up 6% for MARSS and 1.2% for SSN.

Name Fuzzy Comparison

Reasoning

Particularly in the SSN-based data sources, there are many cases in which names are similar enough to qualify for a match, but do not get picked up by the linking rules due to a nuance in the comparison algorithm used (MongeElkan).

Research

See the "Analysis of Match Rates by Agency Identifiers" section.

Recommended Adjustment

While JaroWinkler should certainly be used in name comparisons, MongeElkan still provides value in finding transpositions between first and last names, as well as bunching of name parts into the same field. A combination of MongeElkan and an averaging of JaroWinkler on name tokens is recommended to pick up the potential false negatives in the research section for this analysis.

Impact

A function called FullNameCompare was created with the recommended adjustment. It returns a score of the higher score adjusted by the inverse square of the difference between the two scores. As of this release the function is used in rules 1.1, 1.4, 1.5, and 1.8. Of the records matched with these rules, 284,638 would have not matched using just MongeElkan and 171,147 would not have matched using just JaroWinkler. In addition, while difficult to measure, there should be a small number that the new rule prevented matching. This would happen when the higher score is just above the threshold and the new adjusted score pulls it below the threshold. An inverse square was used instead of an average so that if MongeElkan was 1 and JaroWinkler was 0 (or very low) it would still match this record instead of returning a 0.5 preventing the match.

Name Group Comparisons

Reasoning

In the spot checking done in the linking quality analysis, several diminutive names were identified to be missing from the current name group index. This was impacting matching enough that necessitated research into expanding the current index.

Research

The current name index group, identified by a former P20W project manager, came from a Google code project that has not been maintained for several years and appears limited to certain cultures. A more recent and vastly more extensive diminutive name lookup was found from an open source project utilizing given name to similar name lists compiled from the genealogy software projects of Ancestry.com and WeRelate.org. At the risk of being too loose in name similarities, this index could provide and continue to provide nickname lookups across many cultures for many more names than currently available to the P20W linking engine. There are 66,838 names in the Ancestry/WeRelate index (A/W), compared to the 1,694 in the current index. About 34,000 names exist in the P20W ReportedPersons that have entries in the A/W index that are missing in the current name index.

Applying the new name index to the SSN and MARSS comparisons of the record pairs having the same SSN but were not matched by P20W, 53,048 pairs find a first name match in the A/W index and 29,717 of those also share a last name. Of these, 1,234 would not be caught by the JW and ME thresholds of .8 and .75 on the full name respectively. 15,401 did not share a last name and were below the fuzzy full name thresholds.

Of the record pairs having the same MARSS but not matched by P20W, 29,385 of them find a name index match in A/W, and 106 of these also share a last name. Of these, 4 fall below JW and ME thresholds. 258 did not share a last name and fell below JW and ME thresholds.

Recommended Adjustment

The Ancestry/WeRelate index has so many 'similar names' for each given name that it may generate false positives. As an example, the following names are listed as similar names for Nick: Unis, Kai, Cole, Mikko, Klaus, and Dominik. While some could be diminutive names, a match in this index should be used with caution as not all similar names are common similar names. However, when used in conjunction with an identifier such as SSN and MARSS as well as a last name or date of birth, spot checking confirms these to be fairly strong matches. Recommended adjustment is to use this new name index in place of the current name group index, as at least an SSN, MARSS, or last name and date of birth match are required to use the current name group index. Other modifications, such as allowing last name and/or date of birth to be fuzzy on an identifier match and name index match, could be argued beyond the scope of this analysis.

Actual Adjustments

- Date of birth changes
  - Remove arbitrary JaroWinkler comparison
  - Add checks for single digit edits, transpositions within date parts, and transposition of month and day (most sequential differences are caught with 1 digit check). A date of birth will be considered partially matching (score = .8) if the comparison does not exhibit more than one of the mentioned checked edits. Exact matches are given a score of 1, and anything else is given a score of 0 (for the date of birth comparison, not the whole rule)
- Fuzzy Name comparison changes
  - Raise MoneElkan threshold to .8 on FullName comparisons
  - Add JaroWinkler algorithm to FullName comparison such that if either ME or JW > .8, the match passes
  - Replace current name group specific rules with A/W index
  - Add the A/W index to the JaroWinkler component of the FullName comparisons such that a first name match in the index will give the first name element a .8 score (this will ensure the JW threshold stays at .8 for LastName). This is added to JaroWinkler and not MongeElkan as MongeElkan tokenizes and compares last names to first names to catch transpositions, whereas JW compares first name to first name only and last name to last name only.
- Fixed 1.11.x and 1.12.x rules, condensing them into two 1.11 and 1.12 rules
  - 1.11 and 1.12 rules will prevent a match if SSN or MARSS are present on both records being compared and are different, unless both identifiers are present and one identifier agrees and the other doesn't (in which case the match will be made and the agencies responsible for providing the identifiers are welcome to discuss who is right and who is not).
- Tweaked scoring of fuzzy name matches to better reflect confidence in match quality

P20W Linking Rules

The linking rules in use by P20W as of this release are as follows.

Table 6: Linking Rules

Rule Number	Rule Description	Order Number
1	New Record	1
1.5	FullName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	2
1.6	LastName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	3
1.7	FirstName_Fuzzy, DateofBirth_Fuzzy, SSN_Exact	4
1.1	MARSS#_Exact, FullName_Fuzzy, DateofBirth_Fuzzy	5
1.2	MARSS#_Exact, LastName_Fuzzy, DateofBirth_Fuzzy	6
1.3	MARSS#_Exact, FirstName_Fuzzy, DateofBirth_Fuzzy	7
1.8	FullName_Fuzzy, SSN_Exact	8
1.8.1	SSN_Exact, LastName_Exact, FirstName_NameGroup, MI_null_or_exact	8.2
1.4	MARSS#_Exact, FullName_Fuzzy	9
1.11.1	LastName_Exact, FirstName_Fuzzy, DateofBirth_Exact, SSN_NULL_OR_Match_90%_or_more_to_other_SSN	10
1.12.1	LastName_Fuzzy, FirstName_Exact, DateofBirth_Exact, SSN_NULL_OR_Match_90%_or_more_to_other_SSN	11
1.11.2	LastName_Exact, FirstName_Fuzzy, DateofBirth_Exact, MARSS_NULL_OR_Match_90%_or_more_to_other_MARSS	12
1.12.2	LastName_Fuzzy, FirstName_Exact, DateofBirth_Exact, MARSS_NULL_OR_Match_90%_or_more_to_other_MARSS	13
1.9	LastName_FirstName_Exact, DateofBirth_Exact	14
1.10	FirstName_LastName_Exact, DateofBirth_Exact Against LastName_FirstName_Exact, DateofBirth_Exact	15
1.13	LastName and DateofBirth are Exact and FirstName using NameGroup	16
1.14	Exact Match on SSN and FirstName	17
1.8.2	SSN_Exact, FirstName_NameGroup, MI_null_or_exact	17.1

Conclusion

The evolution of P20W linking and loading generated efficiencies in how we maintain and populate the data warehouse. This is the point at which P20W (and SLEDS) began focusing on building a variety of data extracts for reporting purposes. These data extracts serve to test the linking algorithm in that the data is viewed by experienced researchers from a policy lens or "do these results make sense?".

Selective Group Matching (SGM)

One of the first data extracts from P20W and SLEDS was a data mart created for the Minnesota Office of Higher Education to produce a specific College Readiness report mandated under state law. The report was previously completed using adhoc matching of enrollment data on recent high school graduates from public post-secondary institutions to determine the percent of graduates enrolling in developmental education coursework in the first two years after high school. SLEDS built the report using K-12 and post-secondary linked records allowing for further analysis of student demographics and K-12 assessment results to provide context for the postsecondary enrollment patterns presented.

Initial data mart results were compared to student level rosters provided by the Minnesota State Colleges and Universities (MnSCU) system. As a result of the comparison, certain person linking gaps between K12 education and Post-Secondary Enrollment were identified. While these gaps represent a fairly small percentage of students, they belong to an empirically definable set of circumstances. Selective Group Matching (SGM) is part of an ongoing effort to resolve linking gaps between K12 education and Post-Secondary Enrollment.

Name Changes

A study conducted by Minnesota State Colleges and Universities (MnSCU) selected post-secondary enrollment records for students who self-reported their high school and year of graduation to the college. Those students who had a post-secondary record within two years of high school graduation were counted. To help validate the integrity of the P20W Linking Engine, the same numbers were queried from the SLEDS data warehouse. The results of the P20W Linking were used to match persons from K12 enrollment data to Post-Secondary enrollment data. The variances found in the results were profiled to see if variances could be attributed to differences in data, differences in report logic, or false positive and false negative matches in the P20W Linking. As a result, a small but significant portion of the variance was attributable to a particular class of false negatives in which students whose last names had changed in college prevented their K12 information from being identified as the same person as their Post-Secondary information. The following example details the conditions by which this gap occurs.

The P20W Linking Engine takes all distinct sets of values for personally identifying information (PII) across all data sources and stores them as records known as ReportedPerson records. ReportedPersons are then sorted by quality and sent through the Linking Engine to produce P20W Person records. This is accomplished by comparing each ReportedPerson record to all ReportedPerson records in the (initially empty) reference set. Pre-defined rules are used to determine whether any two given ReportedPersons are a match and should be classified as the same P20W Person. If no match in the reference set is found, the ReportedPerson record creates a new P20W Person and it is placed into the reference set. This continues until all ReportedPersons in the reference set have either created a new P20W Person record or has been assigned to an existing one.

Consider the following ReportedPerson records, all belonging to a Jane Doe, who married during her post-secondary career and changed her last name to Smith, but not before having workforce records associated with her maiden name. The records below are sorted by quality in descending order.

Table 7: Selective Group Matching Sample Source Records

Data Source	Reported PersonID	First Name	Last Name	Date of Birth	SSN	MARSS#
Post-Secondary Enrollment	R1	Jane	Smith	1/1/1990	987654321	------
K12 Enrollment	R2	Jane	Doe	1/1/1990	------	1300000000000
Workforce	R3	Jane	Doe	------	987654321	------

Suppose then that each ReportedPerson record is passed through P20W Linking. The first record, finding no match in the empty reference set, will create a new P20W PersonID:

P20W PersonID	Reported PersonID	First Name	Last Name	Date of Birth	SSN	MARSS#
P1	R1	Jane	Smith	1/1/1990	987654321	------

Then, the K12 Enrollment record, having only first name and date of birth in common with the first record, will fail to match and then create a second P20W PersonID:

P20W PersonID	Reported PersonID	First Name	Last Name	Date of Birth	SSN	MARSS#
P1	R1	Jane	Smith	1/1/1990	987654321	------
P2	R2	Jane	Doe	1/1/1990	------	1300000000000

The workforce record, having a match on first name and social security number, will successfully match to the first record. Consequently, a false negative between the K12 enrollment record and the post-secondary records has resulted in two P20W Persons:

P20W PersonID	Reported PersonID	First Name	Last Name	Date of Birth	SSN	MARSS#
P1	1	Jane	Smith	1/1/1990	987654321	------
P2	2	Jane	Doe	1/1/1990	------	1300000000000
P1	3	Jane	Doe	------	987654321	------

Furthermore, any subsequent records belonging to Jane will **either** match the P20W Person record created by the Post-Secondary record **or** the P20W Person record created by the K12 record. This creates a permanent break in Jane's P20W pathway that will prevent longitudinal

studies from utilizing her educational data correctly. This result will be the same regardless of the order in which the records are processed.

The P20W Linking Engine design has tried to accommodate name changes by what is known as bridge matching. Bridge matching occurs when ReportedPerson records chain together to allow two ReportedPerson records that would otherwise not be identified as the same person to match implicitly. This is accomplished by sorting the data strategically to encourage these matches to occur. In the example above, if Jane's workforce record had a date of birth, the records could have been sorted differently, and Jane's K12 record would have matched the workforce record. As this was not the case, bridge matching could not help Jane's data cross the name change gap.

Solution

In analyzing the above scenario, one can conclude that the information needed to match all of Jane's records was available. What is required is the ability to match ReportedPerson records to a group of previously matched ReportedPerson records. In other words, because Jane's workforce record was matched to Jane's Post-Secondary record, these two sets of PII could be combined to create one or more permutations of PII. These combinations of PII would produce "ReportedPerson" record(s) that, although never having actually been present on any source record, could be used to match Jane's K12 record to Jane's Post-Secondary record and Jane's workforce record on first name, last name, and date of birth.

Using the PII from the three sources records in the example above, ll possible permutations for Jane's PII are listed below:

Table 8: Possible PII Permutations

Reported PersonID	First Name	Last Name	Date of Birth	SSN	MARSS#
1	Jane	Smith	1/1/1990	987654321	------
2	Jane	Doe	1/1/1990	------	1300000000000
3	Jane	Doe	------	987654321	------
(new)	Jane	Smith	1/1/1990	987654321	1300000000000
(new)	Jane	Smith	------	------	------
(new)	Jane	Smith	1/1/1990	------	------
(new)	Jane	Smith	1/1/1990	------	1300000000000
(new)	Jane	Smith	------	987654321	1300000000000
(new)	Jane	Smith	------	987654321	------
(new)	Jane	Smith	------	------	1300000000000
(new)	Jane	Doe	1/1/1990	987654321	1300000000000
(new)	Jane	Doe	------	------	------
(new)	Jane	Doe	1/1/1990	------	------
(new)	Jane	Doe	------	------	1300000000000
(new)	Jane	Doe	------	987654321	1300000000000
(new)	Jane	Doe	1/1/1990	987654321	------

Considerations

An enhancement to the P20W Linking Engine could be implemented to allow this permutation matching to occur. However, allowing ReportedPersons to match all possible permutations of PII previously linked is not without potential for negative impact. As can be concluded from the table listing all of Jane's PII permutations, allowing all of them to be matched to other ReportedPersons would not only be redundant and inefficient, but also decrease the overall quality of the ReportedPerson reference set. This could lead to more lesser-quality matches being made.

However, this can be mitigated by selectively choosing the circumstances under which a permutation may be allowed to enter the reference set. For example, to help eliminate a well-defined gap like name changes, the selective permutation algorithm might only allow permutations on the last name where the SSN or other state agency identifier was an exact match. Another possibility would be to only allow permutations that have a very high PII completeness.

Selectively created high-quality permutations might be limited to the following:

Table 9: Selective Group PII Permutations

Reported PersonID	First Name	Last Name	Date of Birth	SSN	MARSS#
1	Jane	Smith	1/1/1990	987654321	------
2	Jane	Doe	1/1/1990	------	1300000000000
3	Jane	Doe	------	987654321	------
(new)	Jane	Smith	1/1/1990	987654321	1300000000000
(new)	Jane	Doe	1/1/1990	987654321	1300000000000

Actual Adjustments

Updated the Linking Engine with the following steps:

1. Gather unique ReportedPersons from source data and place them into the pre-link area.
2. Sort ReportedPersons by completeness of personally identifying information (PII). Note: Sorting records by "class" would no longer be needed, as this solution replaces bridge matching.
3. Perform the linking of all ReportedPersons and create P20W Person records per linking logic.
4. Permutation Phase **(implemented with the May, 2014 release)**
    a. Remove all singleton ReportedPersons from P20W Person records (those that neither matched another ReportedPerson nor had any ReportedPersons matched to them) and place them back in the pre-link area (step 1 above).

   b. From the P20W Person records that remain, produce the selective permutations from the PII in the P20W Person records and add them as P20W Person records.
5. Repeat steps 2-4 as long as a minimum threshold of matches (number or percentage) continue to occur **(implemented with the May, 2014 release)**
6. When the minimum threshold of matches no longer occurs, all remaining singleton ReportedPersons are added as new P20W Person records. **(Implemented with the May, 2014 release)**

Twins Linking

As P20W linking and loading was adjusted, the Early Childhood LDS team began to review results of linked records specific to the Birth to Pre-K population. One area of concern identified was false positive linkages found among twins.

When multiples (e.g. twins, triplets, etc.) are born, they have the same date of birth, same last name, sometimes the same middle name, MARSS numbers off by one digit, and often very similar first names. This makes it difficult for rules to differentiate the records and correctly identify them as separate people.

Research

When analyzing the Childcare Assistance Program (CCAP) data, a verified set of twins was provided with the dataset. This allowed us to know exactly how many false positives and false negatives resulted.

Most often a false positive link was made between two K12 records linked together through rules 1.11 or 1.13 (There were 23 linked by 1.11 and 7 by 1.13). Then two CCAP records joined to each K12 record placing them, incorrectly, in the same P20W Person.

Analyzing Linking Rule 1.11

It was noticed that the MARSS# between the twins was often numerically very close, often off by just 1. Normally rule 1.11 would match these records. The MARSS# being off by 1 digit would be considered the same as a typo. This is not enough by itself for 1.11 to make a match but does allow the rule to match with a fuzzy comparison on first name. If MARSS agrees then name can be fuzzy. If MARSS disagrees the name must be exact. When records being matched have exact DOB, exact last name, have a MARSS# very close numerically, there is a high probability these records belong to twins. Twins are also highly likely to have different first names (if not always). We can easily figure out how many twins have numerically close MARSS.

Count		Would match 1.13 anyway
1548	Fuzzy first name	463
211	differs by 1	20
254	differs by <10	26
274	differs by <100	31

211 out of 1548 records have a MARSS# that differs numerically by 1. 254 records differ by less than 10 and 274 records differ by less than 100. The third column also shows how many of those would be matched by 1.13 anyway showing how many twins a fix to rule 1.11 would not match only to be matched again by 1.13. Most records outside of the <10 group will likely disagree on MARSS already since more than 1 character is likely to be different. So quite a few false positive twins should be able to be fixed by concentrating on the <10 group.

Analyzing Linking Rule 1.13

There is no threshold for this rule. The rule uses name groups acquired through AncestryWeRelate. The names on the compared records are either in the name group or not. The names within the groups could be analyzed and you could decide whether to remove/add names in the name groups, realizing a possible impact on other records.

There were 182 links made through the 1.13 rule on name groups. Walking through the names we can determine how many specific name associations for each name you could remove from its group without affecting any other records. For instance, ANTHONY is used in 116 links. If you were to remove it from the name group you would likely lose many valid links. However, depending on the originating file, the name group can be customized to eliminate false positive matching twins and would not affect other links. Be forewarned that it could affect future linking so removing these may not be recommended.

Recommended Adjustment

The 1.11 rule should be changed to disagree on MARSS# when they differ numerically by <10. This will cause the rule to require the first name to be exact instead of fuzzy. As long as the twins don't have the same first name this should fix them.

Rule 1.13 should have the names removed from the name groups that were only used to match twins we know should not have matched. This will fix this handful of matches but we should not necessarily expect this to help any future datasets.

Impact

When matching the test CCAP data we identified 33 sets of twins incorrectly matched. When the rule is changed to disagree when MARSS is numerically < 10 none of the records matched on 1.11. 11 went on to be matched incorrectly by 1.13 leaving 22 corrected. This was a good result indicating a near 2/3 improvement specifically for twins (or multiple births).

When analyzing the impact it was also noticed SSN would produce a similar result, also leaving 11 false positive matches so the change was also made to the SSN version of the rules. Even though CCAP benefits from matching K12 data that already matches better on MARSS# we would suggest to also include SSN if available so that better links could also be made without the availability of MARSS data.

In Conclusion

While Minnesota's P20W SLDS has evolved to be a more robust and efficient linking and loading process, there are certain to be additional challenges ahead. Staff have already identified four additional areas that may impact the process: data integrity, data collection, ethnic group matching, and staff record matching.

Data Integrity

When looking at data integrity, one must ask "how good is good enough?" The answer to this is based largely on the researcher's expectations for the data.

Multiple times during the development of P20W, SLEDS, and ECLDS, as new stakeholders came into the project, the initial expectations were that anything short of 100% accuracy in the linking engine made P20W "completely unusable." Conducting education and managing expectations is required with all stakeholders to help them understand the nature of personally identifying information across many systems and that 100% accuracy in linking is simply not attainable when needing to utilize probabilistic linking rules.

Another challenge for P20W was the instant belief that P20W was wrong any time there were discrepancies with an existing report produced within an agency. Discrepancies between reports produced from two different systems were approached by the P20W team as an opportunity to validate linking and, if necessary, improve linking results. The analysis of the discrepancies typically required investigating individual unit records, and it was often determined that P20W was producing better linking results more consistently than the individual agency had in the past.

Data Collection

Improvements in linking results and therefore data integrity can be realized in the future by changing data collection practices. Systems feeding P20W routinely go through development cycles and these enhancement efforts should consider capturing more data elements for PII and capturing them more in alignment with the State of Minnesota's larger needs for P20W analysis. PII data elements consistently and accurately captured across all systems feeding P20W will result in more accurate linking results and even greater confidence in P20W among researchers.

A faster and more cost-effective approach for improving linking results could be to secure access to one or more data sources to provide a full set of PII back to P20W. For example, much of the workforce data includes only a SSN and a self-reported first and last name. If P20W could pass this set of PII to another system containing legal name and date of birth (e.g. Driver and Motor Vehicles), workforce data coming into P20W would contain a higher quantity and higher quality PII. This additional PII could effectively eliminate the linking gap between K12 and Workforce.

Ethnic Group Matching

Minnesota is home to several large immigrant communities whose children comprise a significant percentage of students in K-12 (e.g. Hmong, Cambodian, Vietnamese, Somali). However, the process of immigrating does impact data collected for these students:

- Many students are assigned a birth date of January 1.
- Siblings may have the same first name.
- Family name assignment varies by culture.

These differences may result in a higher rate of false positives for these groups within the P20W linking algorithm. As such, subgroup validity and testing should be undertaken to minimize the impact. Additional rules may be required taking into account any ethnic, cultural or linguistic affinities.

Similar applications have been developed for use in public health initiatives. For example, Onomap allows users to classify lists of names into groups with common cultural, ethnic and linguistic origins. The difference for P20W is that the approach desired would identify distinct individuals within an ethnic population for which rules designed for a predominantly northern European population of students fails to identify distinctly. Somali names are comprised of a personal name, father's personal name and paternal grandfather's personal name (The Financial and Banking Information Infrastructure Committee: A Guide to Names and Naming Practices, 2006). In order to identify a given individual, all three names must be used. Furthermore, women do not traditionally change their names upon marrying. However, Somali women living in Western cultures may adopt their husband's last name (that of his paternal grandfather) to adapt to local customs. The other obvious conclusion is that changes in data collected and data collection processes may also be required in order to maximize the linkages for specific ethnic populations.

Staff Record Matching

The primary purpose of the P20W linking algorithm has been to link student records across education and workforce systems. As staff and teacher data begin to be integrated, these records will naturally be assessed for linkages to postsecondary and employment data. It is uncertain if the rules designed for students will apply perfectly to those applied to staff and teachers given the different process for data collection. As such it will be important to reassess the validity of matches made for this group.