SOG: A SYNTHETIC OCCUPANCY GENERATOR TO SUPPORT ENTITY RESOLUTION INSTRUCTION AND RESEARCH

(Research-in-Progress)
IQ Tools

John R. Talburt
University of Arkansas at Little Rock
jrtalburt@ualr.edu

Yinle Zhou
University of Arkansas at Little Rock
yxzhou@ualr.edu

Savitha Yalanadu Shivaiah
University of Arkansas at Little Rock
sxyalandaus@ualr.edu

Abstract: This paper reports on a project to develop SOG (Synthetic Occupancy Generator), a system to create realistic, but synthetic residential occupancy (name and address) histories as input for Entity Resolution (ER) processes. ER processes are intended to link records referencing the same, or related, real-world entities. Most organizations use some type of ER process to recognize their customers or clients across different channels of contact such as name and address, telephone number, or email address. However, growing concerns over customer privacy and identity theft have made organizations reluctant to publicly release personally-identifiable customer information. The result is that it can difficult to obtain actual occupancy information to use for student exercises or to experiment with entity resolution methods and techniques. SOG was created to address this problem by providing a tool capable of automatically generating a large number of realistic, but synthetic occupancy histories. SOG control parameters allow the user to customize certain features of the simulated occupancy histories. The project reported here is the first phase of a larger project. The second phase is to develop tools that will systematically disrupt the SOG output to create ER test files that have varying degrees of data quality and file formats.

Key Words: Synthetic Data Generator, Entity Resolution, Customer Recognition, Information Quality, Occupancy History, Change-of-Address History

BACKGROUND

Entity resolution (ER) is the process of linking records that reference the same or related real-world entities. ER is a topic of growing importance for business and government. In the commercial world ER is at the heart of customer recognition systems. Often the same customer interacting with a company through different sales channels or branch offices may be treated as a different customer in each transaction. The ability to recognize the same customer using different forms of contact information, such as name and address, phone, or email, is foundational to a successful Customer Relationship Management (CRM) program. ER is also becoming increasingly important in risk management,

especially in the area of fraud detection. ER also an important tool for the intelligence community and law enforcement.

For the purposes of this paper, the term "residential occupancy" or simply "occupancy" is understood to mean a person (name) at a residence (address) for a particular period of time. Occupancy records may also have additional attributes about the person, such as date-of-birth and social security number, or about the residence such as the telephone number. An "occupancy history," also called a change-of-address (COA) history, is a set of chronologically-consecutive occupancy records for the same individual over a period of time.

All organizations, both public and private, routinely gather and store occupancy information about their customers and clients. However, growing concerns over personal privacy and identity theft have made organizations reluctant to share data containing this level of personally-identifiable information with third parties. As a result, it is increasingly difficult to obtain real data sets for use in entity resolution research and instruction. Even when data is available, it may not contain examples of a particular feature of interest, such as, occupancy records for the same person at different addresses.

When real data is not easily obtained or where specific data characteristics need to be closely controlled, synthetically generated data can often provide a solution. There are a number of generators such as DatGen [2] that simply produce tables of uniformly distributed nominal and ordinal values as a way to test sorting and data mining applications. Other generators can be very specialized. A system developed by Yu and Ganesan [12] supports irregular sampling in sensor networks. A program by Hromadka [6] provides a probabilistic simulation of rainfall runoff data, and a synthetic generator by Jiang, Gao, et al [8] produces image data to test signer-independent sign language recognition. Niagara, a system developed by Aboulnaga, Naughton, and Zhag [5], generates large volumes synthetic, complex-structured XML data to support research on XML data management.

Only a few generators are available that produce realistic occupancy data. The most notable is the Parallel Synthetic Data Generator (PSDG) developed by Hoag and Thompson [5]. PSDG was designed to generate "industrial sized" data sets of name and address information for experiments in distributing input data across a cluster of processors where the distribution is determined by certain characteristics of the address. However the primary focus of PSDG is on the geographic distribution of occupancy records. PSDG was not designed to create an occupancy history for an individual over time.

Changes in occupancy information are common occurrences as people frequently change address and in some cases use different names. According to the United States Postal Service [7] 14% of American change address annually. In addition to changes in address, other factors can create different occupancy records for the same customer such as the use of a married name or a nickname, changes is street names, and the use of non-residential addresses, such as, post office boxes, work addresses, or vacation homes.

INTRODUCTION

Figure 1 illustrates the basic elements of an occupancy history for a woman born Mary Smith. There are three occupancy records, each with a name, an address, and a period of time that the occupancy was valid. In this example, the time period is measured to the nearest month, and the periods of occupancy are consecutive. Also note the change in name between Occupancy 1 and Occupancy 2.

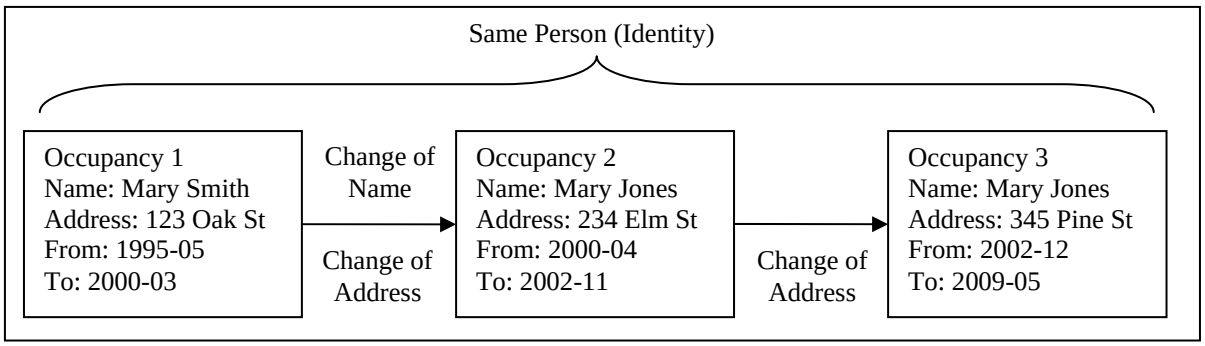

Figure 1: An Occupancy History

Internal View of Identity

The identity of an entity is a set of attribute values for that entity along with a set of "distinctness rules" that allow that entity to be distinguished from all other entities of the same class in a given context [9]. The attributes that define identity are called "identity attributes." In an occupancy history all of the occupancy records reference the same person. However the person shown in Figure 1 is known by two names at three addresses.

There are two ways to view the issue of identity shown in Figure 1. One is to start with the identity based on vital statistics, e.g. Mary Smith, a female born on December 3, 1980, in Anytown, NY, to parents Robert and Susan Smith, and to follow that identity through its various representations of name and address. This "internal view of identity" as shown in Figure 1 is the view of Mary Smith herself and might well be the view of a sibling or other close relative, someone with complete knowledge about her occupancy history. The internal view of identity represents a closed universe model in which all of the occupancy record variants are known to the internal viewer (or system) and any occupancy record that is not one of the known variants must belong to some other identity.

External View of Identity

On the other hand, an external view of identity is one in which some number of occupancy records for an identity have been linked, but it is not known if it is the complete, or even the correct history. When presented with a new occupancy record, the system must determine whether it should be linked to an existing identity, or if it represents a new identity in the system. As an example suppose that a system has only Occupancy Records 1 and 2 of the identity in Figure 1. In this case the system's knowledge of this identity is incomplete. It may be incomplete because Occupancy Record 3 is not in the system because it has not been acquired, or because it is in the system, but has not been linked to Records 1 and 2. In the latter case, the system would assume that Record 3 is part of a different identity. Even though an internal viewer would know that the Occupancy Record 3 should also be part of the Figure 1 identity, the external viewer has not made that determination.

In addition to the problem of incompleteness, the system may assemble an inaccurate view of an identity. When presented with a new occupancy record, the system may erroneous link the new record with an existing identity to which it does not belong.

In an external view, the set of occupancy records that have been linked together comprises the identity of the person. In other words, the linked set of occupancy records represents the system's entire knowledge of the identity. If additional occupancy records are acquired and correctly linked to an identity, then the knowledge about the identity increases. The external view of identity more closely resembles the experience of a business or government agency using entity resolution tools in an effort to link their records into a single view of a customer or agency client. An external view of identity represents an open

universe model. If the system is presented with a new occupancy record, it is not necessarily true that the new record must belong to a new identity. The ER process must decide whether the new record should be linked to an existing identity, or whether it represents a new identity.

The important point here is that an internal viewer is in a position to judge the quality of an external view. With complete knowledge, the internal viewer can easily determine if an external viewer has omitted some records (completeness), has linked records belonging to different identities or has failed to link records for the same identity (accuracy). Talburt, Wang, et al [11] have introduced a quality metric in the form of an index for assessing the similarity of two resolutions of identity. In cases where one resolution represents an internal view (correct) and the other is an external view, the index provides a metric for entity resolution accuracy.

The purpose of SOG is to create a realistic, but synthetic internal view of identity for a large number of identities. The utility of the SOG output is that each of the internal views can be "unlinked" into individual occupancy records. These unlinked records simulate a typical external view of the identity. If the external view of identity is then "re-linked" through an ER process, it is possible to assess the performance of that ER process by comparing the "re-linked" output to the original internal view produced by SOG. Using this methodology, it is possible to construct large test sets for entity resolution processes.

In order to make external views more realistic, they can also be disrupted in ways other than simply removing the record links. Other disruptions include splitting the records into different files (including duplicates across files), using different file formats, omitting attributes, and randomly deleting attribute values, or replacing attribute values with aliases, misspellings, and abbreviations. It is also possible to introduce other common data quality problems such as character transpositions or field truncations.

Inferred and Asserted Associations

When working with an external view of identity, there are two important methods for linking occupancy records, asserted associations and inferred associations. An inferred association is based on indirect evidence usually provided by an exact or approximate match between the set of identity attributes. The simplest form of inferred association is duplicate record matching. If all of the identity attributes in two occupancy records match, then we infer that they are for the same person under the assumption that the identity attributes are sufficient to disambiguate distinct entities in the given context. However, the degree of certainty in this determination is related to the exactness of the match between attribute values and the weights given to the attributes. Table 1 shows some of the issues related to inferred association.

Record	Name	Street Address	PO Box Address	Telephone
1	James Ray Doe	123 Oak St, Anyville	PO Box 45, Anyville	435-8783
2	Jim R Doe	123 Oak, Anyville		435-8783
3	James Doe	345 Pine, Anyville	PO Box 45, Anyville	
4	J R Doe	213 Oak St, Anyville		435-8783

Table 1: Occupancy Record Variants

Table 1 shows four records with four identity attributes, Name, Street Address, Post Office Box Address, and Telephone Number. Record 1 is the most complete with no missing values. Between Record 1 and Record 2 there is an exact of the telephone numbers. However, the two names are not an exact match, but are a close partial match based on the fact that Jim is a common nickname for James. Also the middle initial in Record 2 agrees with the first letter of the middle name in Record 1. Similarly, the Street Address values in Records 1 and 2 are almost the same except that the street suffix in Record 2 is missing. The value for the PO Box Address is missing in Record 2, therefore a match or mismatch on this field

cannot be determined. Despite the fact that Record 1 and Record 2 are not exact duplicates, most ER systems are likely to infer that they are references to the same person. The fact that some fields do not match is overcome by the weight given to the aspects that do match. For example, the exact match of the telephone number increases the confidence that these records should be linked helping overcome any doubt over the missing street suffix in Record 2.

Inferred associations rely upon the construction of a "belief function," a function that calculates a composite value of the association by evaluating all of the individual attribute-level matches and conflicts. When the value of the belief function exceeds a specified threshold, the process concludes that the records reference the same person. Belief functions were introduced by A. P. Dempster in the 1960's [3] and later extended by the work of Glenn Shafer [10]. For this reason it is often referred to as the Dempster-Shafer theory.

Inferred associations can be used to link occupancy records that have different names and addresses. For example in Table 1, Record 1 could be linked with Record 3 based on matching Name and PO Box Address values despite the mismatch on Street Address. However, it is important to note that most systems would not make a direct link between Record 2 and Record 3 because the missing PO Box and Telephone values fail to provide any evidence to overcome the clear mismatch between the Street Addresses. Nevertheless, Record 2 can be linked to Record 3 given that Record 2 and Record 3 both link to Record 1. The extension of the inference process in this way is referred to as "transitive closure", i.e. Record 2 links to Record 1 and Record 1 links to Record 3 implying that Record 2 should be linked to Record 3. A more detailed discussion of transitive closure can be found in Zhang, et al [13] and Gibbs [4].

Record 4 in Table 1 is more problematic. The key issue is whether the street number value of 213 is a possibly a typing error in the value 123. The telephone number match would provide fairly strong support for linking Record 4 to either Record 1 or Record 2 given that the name values are an exact match on last name and are at least consistent with respect to the first name and middle name initials.

In addition to inferred associations, ER systems can also utilize asserted associations. An asserted association is an explicit connection between two occupancy records. Asserted associations are available from a number of sources. Internally, customers often report changes of name and address to the companies they do business with. Companies with more mature ER systems in place will save the previous name and address as a way to link transactions still using the old information to the correct customer account.

There are many externals sources for asserted associations as well. For example, most people in the U.S. self-report their change-of-address to the postal service to assure mail forwarding. That information is then made available to large mailers for address correction. Much of this same information and more is available from other sources such as magazine and periodical publishers, utility providers, and public records. Asserted associations are high-value information because they provide explicit linkage, but they also come with a cost. Asserted associations must be maintained in some type of knowledge base. Storing and developing efficient indexing and access methods for a large knowledge base of asserted associations can be costly. For this reason, many companies prefer to buy access to this information from vendors that specialize in collecting asserted associations and offer access to this information as a product or services.

Following the previous example, suppose that a system has the three occupancy records of Figure 1, and that Occupancy 1 and Occupancy 2 have been linked together, but the system believes that Occupancy 3 is for a different identity. If is discovered that Mary Jones reported to the publishers of her favorite

magazine that effective December, 2002 the magazine mailed to Mary Jones at 234 Elm St should be mailed to 345 Pine St., then it would be straightforward to link Occupancy 2 to Occupancy 3.

SOLUTION

SOG (Synthetic Occupancy Generator) is a software system implemented in Java that was designed to provide realistic, but synthetic occupancy histories that can be used for research and instruction in entity resolution and information quality. SOG is not intended to fully simulate the behavior of individuals, but only those aspects necessary to create realistic occupancy histories.

SOG Overview

The following sections of the paper provide an overview of the operation of the SOG system including

1. Preparation of SOG seed data from a starting sample of real name and address data
    a. Pre-processing of name data
    b. Pre-processing of street address data
    c. Generation of home telephone numbers
    d. Pre-processing of PO Box data
2. Generator Logic
    a. Identity Creation
    b. Scenario Selection
        i. Single histories
        ii. Couple histories
    c. Occupancy History Generation

Preparation of Seed Data

For the experiments described in this paper, the names and addresses generated by SOG were derived from a sample of publicly available occupancy records with addresses in four U.S. states, California, Texas, Florida, and Arkansas. Several of the manipulations of the data described in the section were performed using the software tool DataFlux® dfPower® Studio 8.0 from the SAS Institute.

Pre-Processing of Name Data

Using DataFlux, the names taken from the starting sample are parsed into four components

- First Name
- Middle Name
- Last Name
- Generational Suffix (JR, SR, etc)

Using DataFlux, the first name values are classified according to gender, i.e. Male, Female, and Unknown, and together with the middle, last, and suffix values are used to create six seed tables

- Male First Name Table
- Female First Name Table
- Unknown Gender First Name Table
- Middle Name Table
- Last Name Table
- Generational Suffix Table

The values loaded in each of the name tables are not de-duplicated. This is done so that when name values are randomly selected from the tables, the probability of selecting a particular name value is proportional to its relative frequency in the original source.

The six name tables are then cleaned and standardized. For example, single letters (initials) are allowed in the Middle Name Table, but are removed from the First and Last Name Tables. The Generational Suffix Table are standardized and filtered to include only values "JR", "SR", "I", "II", "III", "IV", "V", "VI", and "VII". Other common errors are corrected including wrong-fielding of items and concatenation of name suffix with last name, e.g. SMITHJR.

Pre-Processing of Street Address Data

The address values taken from the starting sample are first separated into street addresses and post office box addresses. Using DataFlux, the street addresses are parsed into 10 components and a unique address identifier is generated.

- Address ID
- Street Number
- Pre-Directional (N, W, NW, S, E, …)
- Street Name
- Street Suffix (ST, RD, …)
- Post-Directional (N, W, NW, S, E, …)
- Secondary Identifier (APT, STE, …)
- Secondary Number (101, A, B102, …)
- City Name
- State Code
- Zip Code (5-digit)

Even thought the street addresses are parsed into the 10 components listed here, the component address values are not separated into different tables as are the name components. Each street address from the starting sample is left intact as a complete, real address. The real address structure is retained so that the occupancy data generated by SOG can be used to test commercial ER and IQ processes that expect real address data in order to operate properly.

Address cleaning operations focus on standardizing the address components according to US Postal Service guidelines, e.g. variations of Street such as "STREET" and "STR" are all standardized to "ST". The street addresses are also filtered to eliminate addresses that did not have street numbers, e.g. intersection addresses (Main and Broadway).

Generation of Home Telephone Numbers

The final step in street address processing is to add a home telephone number. The telephone number is generated using an area code value that is consistent with the city and state of the street address. Tables relating city and state to area code are widely available on the Internet, e.g. AnyWho.com [1]. After selecting the appropriate area code, the other 7 digits of the telephone number are randomly generated.

Pre-Processing of PO Box Data

PO Box addresses are parsed (fielded) into 4 components

- Box Number
- City Name
- State Code
- Zip Code (5-digit)

At the end of the street and PO Box address pre-processing phases, two tables are built

- Street Address Table
- PO Box Address Table

Identity Creation

The synthetic identity of a person in SOG comprises the following items

- Identity Number
- Name comprising the following 4 items
  - First
  - Middle
  - Last
  - Generational Suffix (populated for only for Male identities)
- Gender Code (M, F, U) consistent with the First Name value
- Social Security Number (selected from table)
- Date-of-Birth (10-digit date value in YYYYMMDD format)

The Identity Number is a sequential integer value that is unique for each identity created by SOG. The name is assembled by randomly selecting First, Middle, Last, and Suffix values from the corresponding name tables. Suffix is only populated when creating a Male (M) identity.

The Gender Code corresponds to the table from which the first name value is taken, M for Male, F for Female, and U for Unknown. The Social Security Number (SSN) is selected without replacement from the randomly generated 9-digit numbers in the SSN Table assuring that every identity has a unique SSN.

The Start-Date (start of simulation or SOS) and End-Date (end of simulation or EOS) are two SOG parameters that control the total period of time covered by the simulation. All SOG dates, including Date-of-Birth for an identity, must be within the period of the simulation. Another SOG parameter controls the age at which the first occupancy for a person in the simulation is generated. The minimum age of first occupancy constrains the generation of the Date-of-Birth to assure that each identity will have at least one occupancy record before the EOS. The generated Date-of-Birth for an identity may also be subject to other constraints described later.

Scenario Selection

The current version of SOG supports two basic occupancy history scenarios, Single and Couple Scenarios. In the Single Scenario a person has a series of consecutive occupancy records unrelated to any other identity in the simulation. Figure 2 shows a typical Single Scenario comprising a total of 6 occupancies for an identity with unknown gender. All occupancy histories begin when the person reaches the minimum age of occupancy, 18 in this example. All Single Scenarios continue until EOS. To conserve M and F names for use in Couple Scenarios, identities created for Single Scenarios are all created from the first names in the unknown gender table.

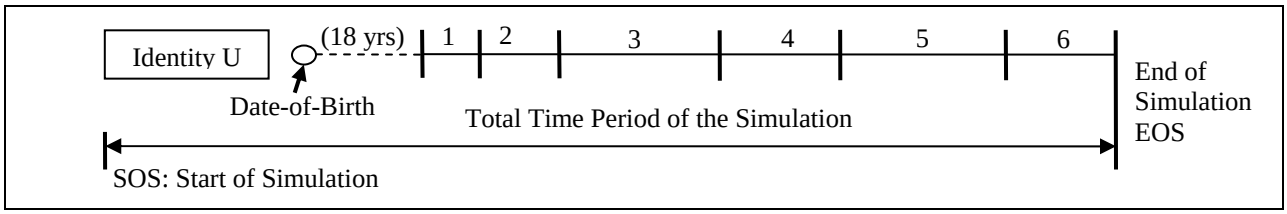

Figure 2: Single Scenario for an Identity with Unknown Gender

The Couple Scenario simulates the situation where two identities share the same address at the same time for some number of consecutive occupancies. The Couple Scenario can have several variations related to the intersection of the two occupancy histories, however in the current version of SOG all Couple Scenarios begin with a Male and Female Identity each having at least one single-occupancy record prior to sharing. Another SOG parameter controls the maximum difference in age between the identities created for a Couple Scenario.

Figure 3 illustrates the simplest form of the Couple Scenario where the two people share a series of occupancies until the end of the simulation. The Couple Scenario shown in Figure 3 comprises a total of 13 occupancies with single occupancies starting at age 18. The number of single occupancies for each identity is a random choice from 1 to the maximum number (another SOG parameter). The female identity has two single occupancies prior to marriage, and the male identity has 3. At a randomly selected date, the two identities start a series of occupancies that share the same start and end date and have the same street address. Also during the period of shared occupancy, the last name of the female identity in the occupancy record is replaced with the last name of the male identity.

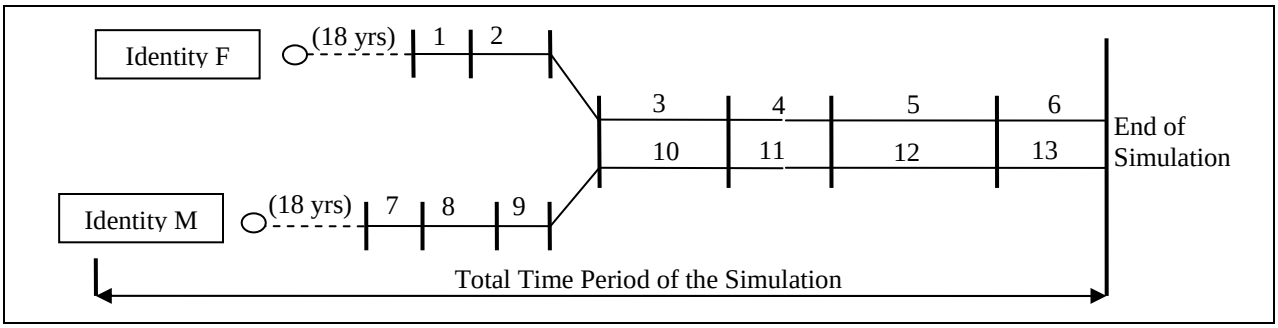

Figure 3: Shared Occupancy Continues to the End of the Simulation

Figure 4 shows a variant of the Couple Scenario in which the shared occupancy ends before the end of the simulation. The Couple Scenario in Figure 4 comprises a total of 16 occupancies. When SOG selects this variation, a second date is randomly generated that falls between the start of the shared occupancy period and the end of the simulation. The second date is the date of separation. After the separation date, each identity continues with a randomly selected number of single occupancies until the end of the simulation. In occupancy records after the separation, the last name of the female identity reverts to its original identity value, but her middle name is replaced by the last name of the male identity.

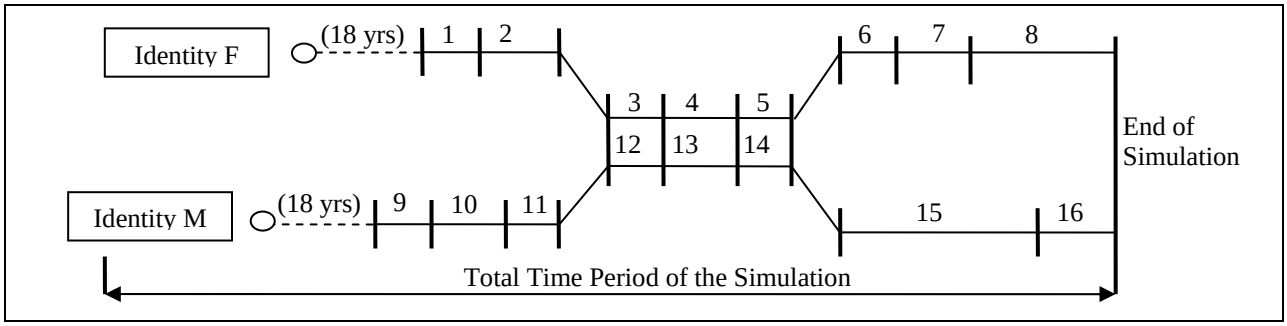

Figure 4: Shared Occupancy Followed by Separation

Figure 5 shows yet another variation of the Couple Scenario in which the shared occupancy ends by the death of one of the identities, here the male identity. In cases where the male identity dies, the occupancy records for female identity continue to use the last name of deceased male identity.

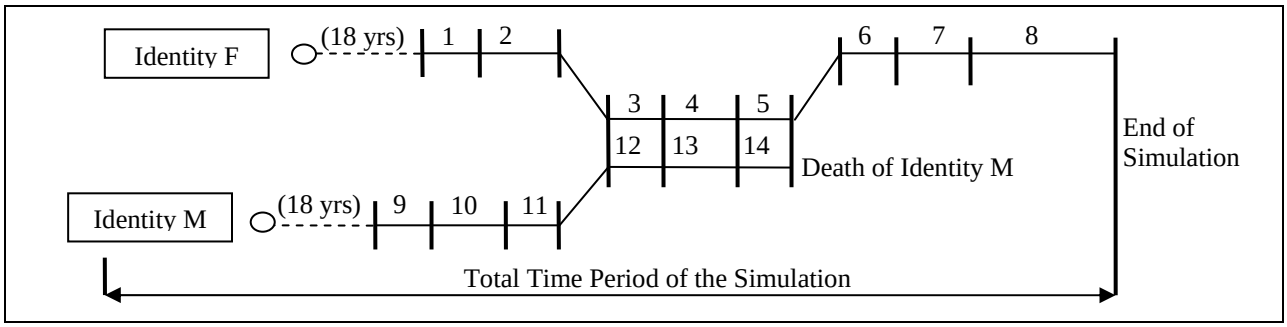

Figure 5: Shared Occupancy Ending with the Death of a Partner

Figure 6 summarizes the Scenario Selection Process that creates the Single and Couple Scenarios illustrated in Figures 2 through 5.

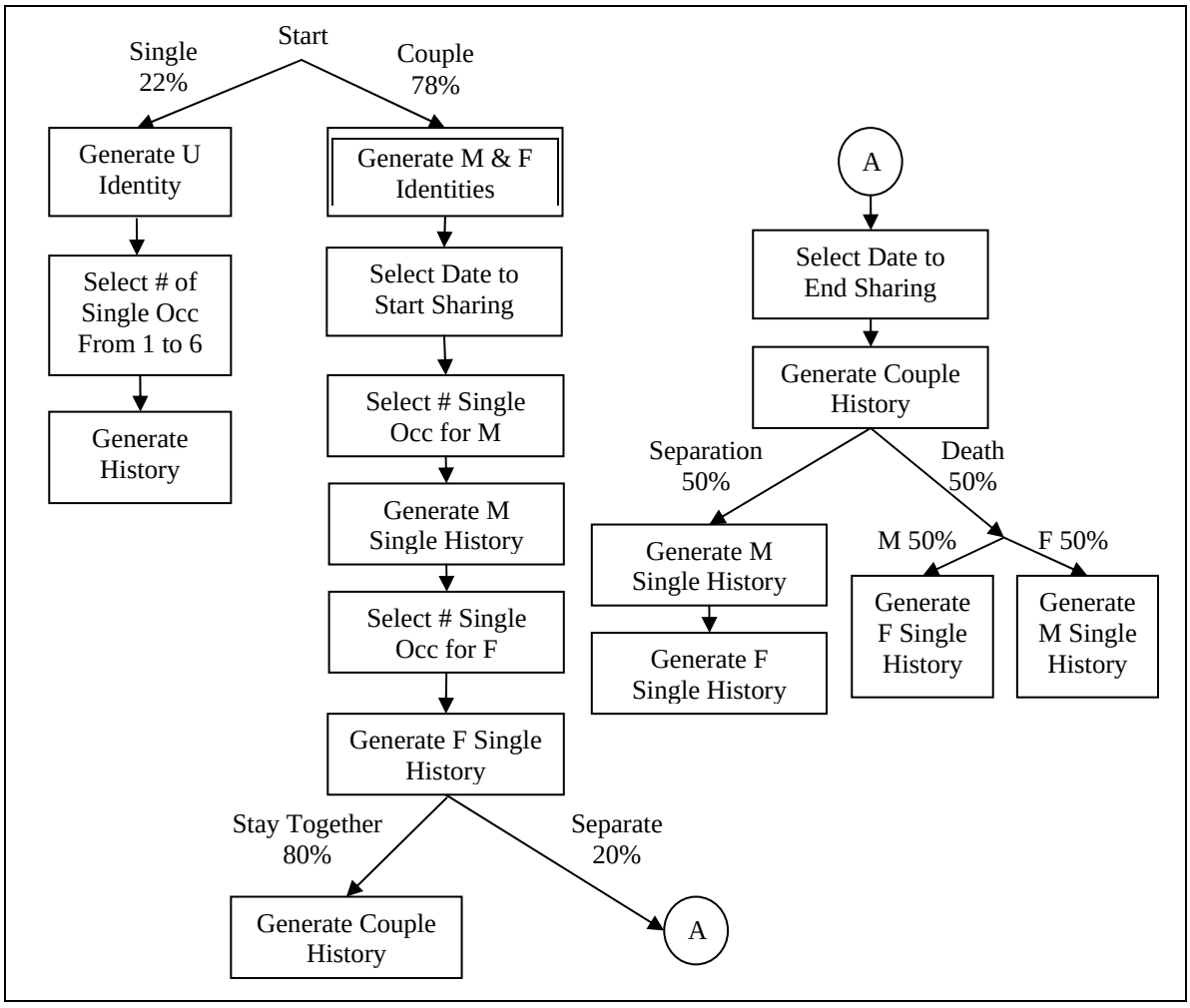

Figure 6: Scenario Selection Flow

Occupancy History Generation

For each scenario selection one or more series of occupancy records must be generated in order to build the complete occupancy history. The generation of a series of occupancy records is an iterative process driven by the number of occupancy records required and by the start and end dates of the series.

Occupancy Record Structure

The occupancy record generated by SOG has 30 items. These are

1. Occupancy ID (sequence number, unique for each occupancy record)
2. History Sequence ID (sequence number, unique for each occupancy history)
3. Identity Number (from Identity Table)
4. First Name (from Identity Record)
5. Middle Name (from Identity Record)
6. Last Name (same as Last Name of Identity except for Female as noted in Couple Scenarios)
7. Generational Suffix (from Identity Record)
8. Social Security Number (from Identity Record)
9. Date-of-Birth (from Identity Record)
10. Single/Couple Status (S/C)
11. Street Address ID
12. Street Number
13. Pre-Directional
14. Street Name
15. Street Suffix
16. Post-Directional
17. Secondary Identifier
18. Secondary Number
19. City
20. State Code
21. Zip Code
22. Telephone Number
23. Start Date of Occupancy
24. End Date of Occupancy
25. PO Box Indicator (null or “PO BOX”)
26. PO Box ID
27. PO Box Number
28. PO Box City
29. PO Box State
30. PO Box Zip Code

Occupancy Period Determination

The process for building a series of consecutive occupancy records can best be explained by example. Suppose that a series of 3 occupancy records is to be built starting at date A and ending at date B. The first step is to calculate the total number of months in the time period defined by A and B. Suppose that this period is 40 months long. Before selecting the length of the first occupancy, the total number of months must be reduced by the number of occupancies remaining to be built. This is to assure that at least one month will be left for each of the 2 remaining occupancy records. In this example, the 40 months is reduced to 38 and the length of the first occupancy is determined by randomly selecting a number from 1 to 38. Suppose this value is 18. Then the first occupancy will start at date A and continue for A+18 months. Now of the original span of 40 has been reduced to 22 months. Repeating the process, the number of remaining months (22) is reduced by 1 to allow at least one month for the last occupancy. The length of the second occupancy is then determined by randomly selecting a number from 1 to 21. Suppose this choice is 5 months. The second occupancy will start at the end date of the first occupancy and continue for 5 months. By default the third and last occupancy will start at the end of the second occupancy and continue until date B, the desired end of the occupancy segment.

Address Locality and PO Box Addresses

As the occupancy records in a series are built, another feature of SOG governs the change in locality between two consecutive occupancy periods. For example in the U.S., the change in location is not totally random. There is a bias toward changes in address that are in proximity to the previous address, i.e. there is a higher probability that a move will be within the same city rather than a move to a different city or different state. In SOG these proximity biases (probabilities) are controlled by parameters at three levels, within the same Zip Code, within the same City, and within the same State.

The assignment of PO Box address is also related to the proximity of the previous address. In the current version of SOG, the PO BOX address is an optional value for any given occupancy, and when present, it is always selected to be in the same city and state as the street address. SOG also randomly decides whether to reuse a PO Box from the previous occupancy in the next occupancy in cases where the change in street address is within the same city and state.

SOG Parameters and Probabilities

Table 2 shows a summary of the parameters that limit the maximum and minimum values in the SOG simulation along with example values from an actual run of the simulation.

Parameter	**Value**
End Date of Simulation	2009/02/01
Start Date of Simulation	1900/01/01
Age at First Occupancy	18 years
Single Scenario, Max Nbr Occupancies	6
Couple Scenario, Max Nbr Pre-Single Occupancies	3
Couple Scenario, Max Nbr Shared Occupancies	3
Couple Scenario, Max Nbr Post-Shared Occupancies	3
Couple Scenario, Max Age to Start Shared Occupancies	50 years
Couple Scenario, Max Age Difference Between Identities	10 years

Table 2: Maximum and Minimum Value Parameters

Table 3 shows a summary of the SOG parameters that govern the probabilities in selection criteria in the simulation along with examples of assigned values and actual results from a run of the simulation that generated 84,515 occupancy records comprises 11,280 occupancy histories.

Parameter	**Value**	**Actual**
Select Single Scenario	22%	22.09%
Select Couple Scenario	78%	77.91%
Share to EOS	80%	78.39%
Separate before EOS	20%	21.61%
Neither Identity Dies	50%	52.31%
One Identity Dies	50%	47.69%
Female Identity Dies	50%	52.04%
Male Identity Dies	50%	47.96%
Add PO Box to Occupancy	50%	50.26%
Keep Same PO Box	90%	90.23%
Move Within Same Zip Code	17%	17.24%
Move Within Same City	33%	32.95%
Move Within Same State	50%	49.81%

Table 3: Probability of Selections

Figure 7 shows an example occupancy history, #9037, generated by SOG as one of 11,280 histories created during the same run. The output is in the form of a pipe-delimited text file with the 30 items described in the Occupancy Record Structure section of this paper. History 9037 is a simple Couple Scenario with the female identity Connie Wasmund having 3 single occupancy records, #67783-67785, before the start of the shared occupancy period. The male identity, Robert Williams, has 2 single records, #67786-67787, before the start of the shared occupancy period. The single occupancies are followed by a series of 3 shared occupancy records beginning in April 2000 and ending with at EOS in February 2009.

```
67783|9037|16068|CONNIE|ANN|WASMUND||F|160285600|19660709|S|17632|8478||WESSE
    X|CT||||JACKSONVILLE|FL|32244|9047452377|198407|199208|||||
67784|9037|16068|CONNIE|ANN|WASMUND||F|160285600|19660709|S|27954|23712||OAK|
    AVE||||SORRENTO|FL|32776|3527170180|199208|199803|PO BOX|2720|1033|SORRENT
    O|FL|32776
67785|9037|16068|CONNIE|ANN|WASMUND||F|160285600|19660709|S|27950|5487|NW|27T
    H|PL||||OCALA|FL|34482|3525147505|199803|200004|PO BOX|1198|2152|OCALA|FL|
    34478
67786|9037|16069|ROBERT||WILLIAMS||M|160294773|19750201|S|12214|1805||ORTEGA|
    DR||||MODESTO|CA|95355|2093328747|199302|199910|PO BOX|56648|582052|MODEST
    O|CA|95358
67787|9037|16069|ROBERT||WILLIAMS||M|160294773|19750201|S|27945|293|E|69TH|WA
    Y||||LONG BEACH|CA|90805|5622620635|199910|200004|PO BOX|497|17402|LONG
    BEACH|CA|90807
67788|9037|16068|CONNIE|ANN|WILLIAMS||F|160285600|19660709|C|24257|3725||DORA
    L|CT||||PORT SAINT LUCIE|FL|34952|7722326078|200004|200406|||||
67789|9037|16069|ROBERT||WILLIAMS||M|160294773|19750201|C|24257|3725||DORAL|C
    T||||PORT SAINT LUCIE|FL|34952|7722326078|200004|200406|||||
67790|9037|16068|CONNIE|ANN|WILLIAMS||F|160285600|19660709|C|27941|13349||STE
    VES|CT||||SPRING HILL|FL|34609|3524441015|200406|200410|||||
67791|9037|16069|ROBERT||WILLIAMS||M|160294773|19750201|C|27941|13349||STEVES
    |CT||||SPRING HILL|FL|34609|3524441015|200406|200410|||||
67792|9037|16068|CONNIE|ANN|WILLIAMS||F|160285600|19660709|C|27935|1843||CLOV
    ER|CIR||||MELBOURNE|FL|32935|3216765334|200410|200902|PO BOX|8268|11|MELBO
    URNE|FL|32902
67793|9037|16069|ROBERT||WILLIAMS||M|160294773|19750201|C|27935|1843||CLOVER|
    CIR||||MELBOURNE|FL|32935|3216765334|200410|200902|PO BOX|8268|11|MELBOURN
    E|FL|32902
```

Figure 7: Example of a SOG Generated Couple Scenario

CONCLUSIONS AND FUTURE WORK

Preliminary results have demonstrated that SOG is an effective tool for generating synthetic occupancy histories. Although several refinements are planned for SOG, the current work focuses on Phase 2 of the research, the disruption of the occupancy histories.

The first experiments with disruption involve splitting the SOG output into separate files and generating change-of-address records. The Splitter program reads the occupancy histories produced by SOG and randomly assigns each occupancy record to one or more of the split files. The number of split files is a control parameter of the Splitter Program. Every SOG occupancy record is output into at least one of the split files, but the same record may be output into several of the split files creating overlap or duplication among the split files. When a SOG occupancy record is written to one of the split files, the Occupancy ID, History Sequence ID, and Identity Number are removed from the output.

Removing these identifying numbers from the SOG occupancy records creates an external view of the identities. For example in Figure 7, we have an internal view of the Connie Ann Wasmund identity. A priori we know that the first two occupancy records belong to her because they all share her unique Identity Number, #16068, and Social Security Number, 160-28-5600. However if we are presented with these same occupancy records without these identifiers, it not clear whether the Connie Ann Wasmund living at Wessex Ct in Jacksonville, FL, is the same Connie Ann Wasmund living at Oak Ave in Sorrento, FL, without other supporting evidence.

For this reason, the Splitter program also has the ability to produce Change-of-Address (COA) records to support asserted associations between occupancy records. Part of the intelligence of the Splitter (and a focus of current research) is while it removes the original SOG link from a pair of records, it must be sure that it either leaves sufficient evidence to infer a link between the records or it must produce an asserted association in the form of a COA record.

Using the same example of Connie Ann Wasmund, if the Splitter decides (randomly) to remove the Social Security Number from either of the first two occupancy records, then it should produce a COA that will allow these records to be re-linked. A COA record would explicitly state that

Connie Ann Wasmund moved from Wessex Ct, Jacksonville, FL to Oak Ave, Sorrento, FL

In this case, the Splitter has two ways to assure that the first two occupancy records for Connie Ann Wasmund can be re-linked. The first is to leave the SSN in both records to provide an inferred association. The second is to create a COA record. If either of these associations are in place, re-linking the two records is possible. However if both are removed, there is no longer an inferred or asserted association between these two records that would justify re-linking them.

The key research that must be completed for the next phase of SOG is to develop a systematic way to disrupt an occupancy history while also preserving a minimal set of inferred and asserted associations that will guarantee that an external viewer can re-construct the complete occupancy history.

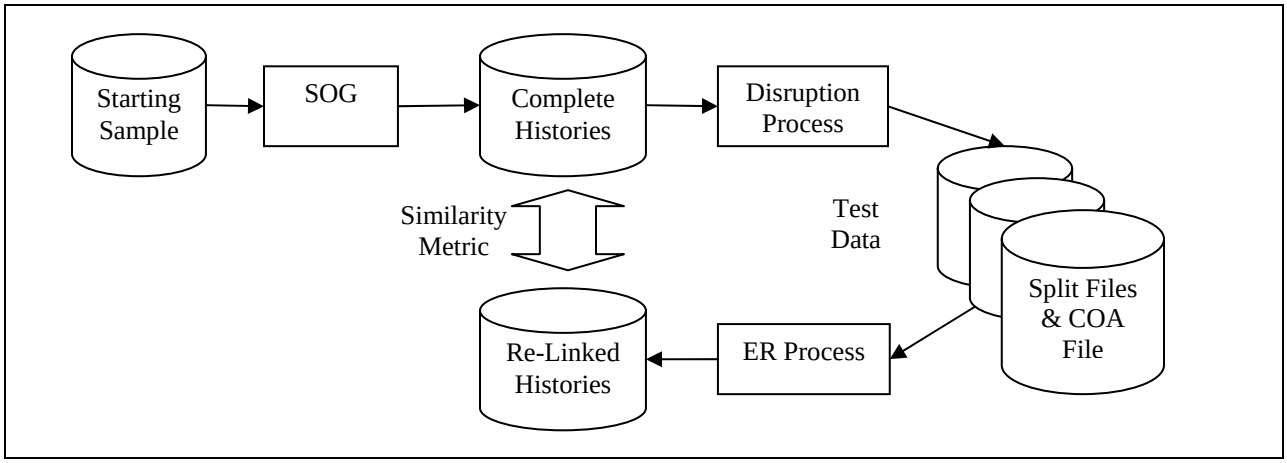

Figure 8: Cycle of Generation, Disruption, Re-Linking, and Comparison

Figure 8 shows how SOG fits into in an ER instruction or research context. SOG plays the critical role of generating a set of complete occupancy histories that serves as the benchmark for evaluating the performance an ER process. Each SOG history represents an internal view of an identity. After the histories are disrupted into external views of the same identities, it becomes the input to an ER process that re-links the disrupted records. The final step is to evaluate the effectiveness of the ER process by comparing the re-linked records to the original histories.

ACKNOWLEDGEMENT

The license for the DataFlux® dfPower® Studio software used in the project was provided by the SAS Institute as part of its sponsorship of the UALR Information Quality Graduate Program. The research described in this paper was partially supported by contributions from Market Strategies International, Livonia, Minnesota, and Acxiom Corporation, Little Rock, Arkansas.

REFERENCES

[1] AnyWho.com area code tables, http://www.anywho.com/area_codes.html.
[2] DatGen, free web application, www.datasetgenerator.com.
[3] Dempster, A.P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics* 37 355-374.
[4] Gibbs, T.H. (2006) A Declarative Approach to Record Linkage2006 Conference on Applied Research in Information Technology, pp. 82-88, posted on http://research.acxiom.com/publications.html.
[5] Hoag, J.E. & Thompson, C.W. (2007) A parallel general-purpose synthetic data generator. *SIGMOD Record* 36(1).
[6] Hromadka, T.V. (1996) A rainfall-runoff probabilistic simulation program. *Environmental Software* 11(4).
[7] Intelisent Postal Affairs Blog (2009) http://www.intelisent.com/postalaffairsblog/?p=262.
[8] Jiang, F., Gao, W., et al. (2009) Synthetic data generation technique in Signer-independent sign language recognition. *Pattern Recognition Letters* 30(5).
[9] Lim, E.P., Srivastava, J., Prabhakar, S. & Richardson, J. (1993) Entity identification in database integration. *Ninth International Conference on Data Engineering*, 294-301
[10] Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.
[11] Talburt, J., Wang, R., Hess, K., & Kuo, E. (2007). An algebraic approach to data quality metrics for entity resolution over large datasets. In L. Al-Hakim (Ed.), *Information quality management: Theory and applications* (pp. 1-22). Hershey, PA: Idea Group Publishing.
[12] Yu, Y., Ganesan, D., et al (2003) Synthetic data generation to support irregular sampling in sensor networks. *Geo Sensor Networks Oct. 2003*.
[13] Zhang, J., Bheemvaram, R., & Li, W. (2006) Transitive Closure of Data Records: Application and Computations, 2006 Conference on Applied Research in Information Technology, pp. 71-81, posted on http://research.acxiom.com/publications.html.