White Paper #ICP-2016.1R
Updated 8/8/2016

Useful utilities for Social Science Researchers: Anonymization, Reconciliation, and Node Identification

Geoffrey P. Morgan[*]

Carnegie Mellon University
Pittsburgh, PA

Kelly Garbach[‡]
Point Blue Conservation Science
Petaluma, CA

Corresponding author E-mails:
[*] geoffrey.p.morgan@gmail.com
[‡] kgarbach@pointblue.org

Abstract. We introduce two open-source utilities, Anonymizer and Linker, to address common data-cleaning and data-processing needs. Both utilities work with structured table-data. Anonymizer provides support for token anonymization and reconciliation based on the Minimum Edit Distance metric. Linker allows for unique entities (people, places, things) to be identified in data based on multiple column values. Two practical scenarios are given where both utilities, in different order, have been useful: one involving coordination with a data set collected in partnership with a large, government agency and the second involving Qualtrics or Mechanical Turk data. Cross-platform distributions and source code are available to readers.

Key words: minimum edit distance, social network analysis, survey research, text-string matching

1 Introduction

Many researchers face a common challenge: how to maintain respondent confidentiality during survey research, while preserving enough detail in the data to support complex analysis, such as social and spatial network analysis. Please imagine the following scenario. As a researcher, you have gathered survey data by phone. Naturally, there are identifying features of the data, which must remain confidential, as required by Institutional Review Board (IRB) protocols and ethical research conduct. If data

were gathered through a national agency (e.g., the National Agricultural Statistics Service or similar) or a large lab with many student workers or consultants, there are likely to be transcription irregularities due to translating data from spoken language to recorded data (these errors may be more acute and/or more frequent if the survey is conducted by phone). Finally, your survey was implemented in multiple regions, so names may repeat across each survey location without referring to the same individual. This challenge involves several problems. First, to comply with agency requirements of your survey distribution partner, you must anonymize[1] your respondents, but you want to make sure they are anonymized consistently. Second, columns of your data may have transcription errors, so you would like to merge names that are very similar as probably being the same name. Finally, you want to identify the individuals embedded in your data and possibly infer links between those individuals.

To address these and related problems, we present a set of utilities in two tools: Anonymizer and Linker. The utilities embodied in these tools consume structured table-data to reconcile values, anonymize values, and identify unique entities. By structured table-data, we mean data with column headers, and each column is delimited (separated) by a consistent symbol. This type of data is often viewed as spreadsheets. To handle transcription errors, we use a text string matching algorithm as the basis of a reconciliation approach to identify non-quantitative values (e.g., "John Smith"), quantifying the minimum edit distance (Wagner & Fischer, 1974) between this value and all previously seen values (e.g., "Jonathan Goodwright", "Jon Smith"), and determining if they are so similar as to be truly the same value (e.g., "John Smith ≈ Jon Smith"; "John Smith" ≠ "Jonathan Goodwright"). To anonymize names, we replace all unique values founds with a consistent anonymous token (e.g., "John Smith = "Name_1"). This token will be consistent for all data examined during that use of the tool, but are not guaranteed to be consistent across different tool run-time instances. To identify individuals in table-data, we allow the user to use any arbitrary set of columns to generate unique node instances. Links are inferred between entities if they co-occur on the same row.

These utilities are publicly available, as both runnable programs and with full source code for examination and extension (Morgan & Garbach, 2016). We present these utilities in two distinct tools. The first, Anonymizer, is responsible for value reconciliation and anonymization, while the second, Linker, is responsible for identifying unique nodes. The two tools are only loosely coupled – there is no requirement to use them in conjunction or in any particular sequence. There are valid and useful processes that use the Anonymizer first and then the Linker, and there are other use-cases where you would want to use the Linker first and then the Anonymizer.

We created these tools because these capabilities are not otherwise available to research groups without significant programming expertise; we are not aware of any publicly available tool that provides these capabilities off-the-shelf. Anonymizer and

[1] Institutional Review Board (IRB) protocol and ethical research conduct require that respondent information remain confidential, meaning that no identifying information is released publically. Federal agencies increasingly require an additional level of anonymizing data, which can mean that researchers do not have access to any identifying information (e.g., names, and/or locations of respondents) prior to data cleaning and analysis.

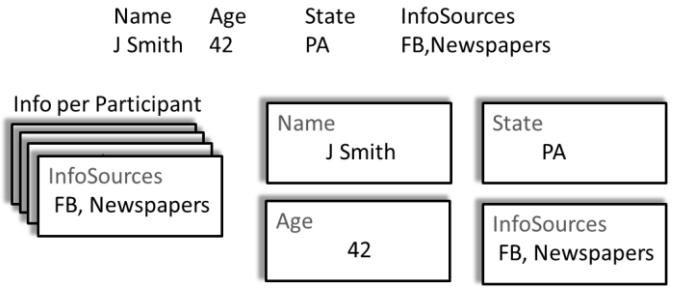
Linker are intended to work with any data of interest. Further, because both tools have configuration files – the configuration files can be stored and used repeatedly if data is collected at multiple points (e.g., longitudinal data; studies replicated in multiple regions). These configuration files and executables can also be passed to collaboration partners so that data treatment is handled consistently across multiple work-groups.

The remainder of this methods note outlines the tools and use-cases underlying their creation, and presents two practical application examples. Both Anonymizer and Linker have been used in our work, with Anonymizer successfully deployed to a third-party government partner to successfully anonymize data collected in a nation-wide survey.

2 Anonymizer

The Anonymizer is designed to work with structured table-data, and natively works with TSV (tab-separated values) that has a header row. The Anonymizer expects, but does not require, that a row represents a survey participant (this is a strong demand of the Linker, described later). When consuming the data, the Anonymizer relies on the header row to match to values described in the configuration, and thus it is important that the header row be well-formed, with unique values for each column. Anonymizer, in essence, can be thought of as that of a very eager student who creates a stack of note-cards.

Fig. 1. Anonymizer generates a stack of note-cards per row of table data



Anonymizer takes these stacks of note-cards and identifies (based on a user-supplied configuration) which columns/note-cards require reconciliation and anonymization. Configuration of the anonymizer allows for either only reconciliation to be performed, or reconciliation followed by anonymization. However, these processes are linked – if both anonymization and reconciliation are active, then values will first be reconciled and then anonymized. Reconciliation may be turned off in the configuration, so all values are matched exactly. More details can be found in the Anonymizer Quickstart Guide available, along with all code, at the open-source GitHub repository (Morgan & Garbach, 2016).

Reconciliation proceeds by working through the data, calculating the minimum edit distance (Wagner & Fischer, 1974), and then using that to determine whether two values are so similar as to be considered identical. This technique has been classically used to correct minor spelling issues (Okuda, Tanaka, & Kasai, 1976). One enhancement we offer over traditional uses of this metric is an ability to configure a variable edit threshold based on the original word length. By default, we use the configuration described in Table 1. The results of reconciliation are printed to the Graphical User Interface (GUI) window for review (Figure 2).

Table 1. The default configuration of our Minimum Edit Distance Threshold

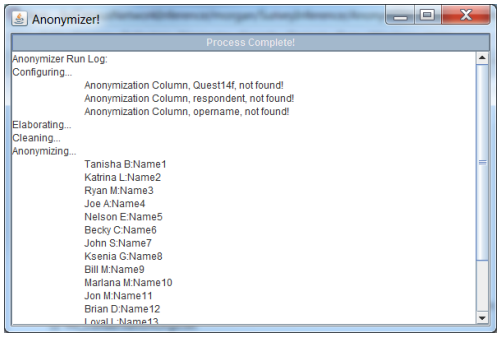
Length of String	Threshold
3	0 (exact match)
8	1
12	2
> 12	3

We configured an exact match for strings three characters or less because we wanted to ensure that frequently referenced acronyms, such as universities (e.g., "OSU", "MSU"), or government agencies (e.g., "NSF", "FDA"), were preserved.

Anonymization occurs after reconciliation. The goal of this anonymization is to mask the original values but use consistent tokens, so that relational structure can still be inferred from later examination of the data. "John Smith", if converted to the token, "Name_1", should always be replaced by "Name_1" whenever "John Smith" would have appeared in the original data. Because reconciliation has already taken place, we know that the reconciled names are now unique. We then create a new consistent token for each unique name, and apply them to set of row-data as a new "note-card". After the anonymization procedure is complete, we remove from memory the identification data before writing anything to file.

The Anonymizer was designed to be used by third parties who are not programmers, so we created a basic GUI for the tool that reports on anonymization and that the task is complete. Most standard surveys take very little time to be processed, so a GUI to provide feedback provides reassurance that the tool has done its job.

Fig. 2. The Anonymizer Feedback Window



3 Linker

Linker also operates on table-data and by default assumes that the data is separated by tabs (TSV). The Linker requires there to be a header row, and assumes that objects identified in the same row are linked. The Linker generates relational data, including node definitions and links between nodes, which are needed for social network analysis.

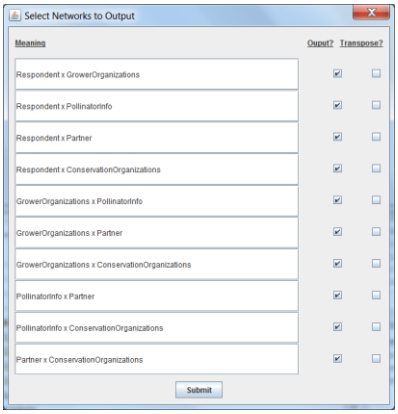
Existing tool-kits for network analysis, such as UCINET (Borgatti, Everett, & Freeman, 2002) or ORA (Carley, Pfeffer, Reminga, Storrick, & Columbus, 2013), support most of the features of Linker, but do not offer the ability to automatically concatenate columns for use in node-definition. Linker, thus, offers particular utility when considering various node requirements. For example, the data has the columns "Respondent Name", "Respondent Role", and "Respondent State". Unique nodes may initially be identified based on "Respondent Name" but then it becomes evident that multiple "John Smith"s occur in the data, and the research team may decide it's relatively unlikely that John Smiths who live in different states are actually the same John Smith. Linker makes it easy to add this term to the node definition, and then "John Smith+++PA" will be distinct from "John Smith+++FL". Further, Linker automatically adds as attributes of the node the original source columns used to create the node ID. All of these manipulations can also be done via concatenation in a spreadsheet tool or coding environment such as R (R Core Development Team, 2011), but then you increase the difficulty of consistent replication; with your data going through multiple transformative processes that must be carefully archived. Linker configurations can be saved and used on new data as required, as long as that data shares a consistent header usage.

A node has multiple attributes, including its ID, its type, identifying characteristics, and data characteristics. The ID is the unique ID of the node definition, while the type is a non-unique statement of what the node is. In the example provided in Figure 3, PartnerA and PartnerB are both node definitions with unique IDs, but both share the type "Partner". Identifying characteristics are concatenated together to identify unique nodes, while data characteristics are attached to the identified nodes as provided.

Once all node-types have been identified, the Linker enumerates the potential networks that could be drawn between the identified node-types – the user then selects which networks will be outputted. For example, the data includes entities of type "Participant", "Partner", and "InfoSource", then Linker will offer "Participant x Partner", "Participant x InfoSource" and "Partner x InfoSource", with a transpose button next to each and a provided text label to capture the relationship implied (Figure 3). The choice of what networks to output is an important one for analysis tasks. Node-Types thus help consolidate the analysis to a more manageable level, while providing needed consistency in treatment of variables.

The Linker exports both generic tab-delimited text formats (edge-lists per network and attributes per node-set) as well as the dynetml files preferred by the ORA program.

Figure 3. Network Output Dialog



4 Example Applications

Our first applied case, and the original motivation to develop both Anonymizer and Linker, was an agricultural survey of innovative crop pollination practices, completed in collaboration with the USDA's National Agricultural Statistics Service (NASS). The NASS performed phone interviews of key grower populations of interest in multiple states, but could not release the survey results (which included identifying information) without anonymization or deletion of the identifying data. Because we were interested in the network/peer effects of agricultural practice adoption, we had to provide an anonymization tool that could be used by our NASS collaborators. The tool was successfully deployed, and then Linker was used to infer network structures in pursuit of our research question.

A second applied case involves the cleaning of Qualtrics (Qualtrics, 2014) data gathered via Amazon's Mechanical Turk (Amazon, 2005). Mechanical Turk allows for much cheaper and much larger collections of survey data, however, the connection between Mechanical Turk and Qualtrics is not seamless, and further, Mechanical Turk users must be carefully managed to get quality data. We used Linker to generate participants (identified by IP-Address), and response sessions. We quickly identified 98 participant data, out of 3026, that were problematic due to discrepancies in collected demographic data (age and gender), with an additional 150 that require examination due to their number of response sessions recorded. Once the participant data has been cleaned, the Anonymizer will be used to convert the IP-Address and other identifying characteristics to generic non-identifying tokens.

5 Discussion and Conclusion

We created these tools because the utilities they contain were needed to solve our data collection and analysis challenges, namely maintaining data anonymous while retaining enough detail to support social network analysis. We believe they will be useful for many common problems in survey data collection and subsequent analysis. However, participants with unusual attributes may still be evident in the data and identifiable with sophisticated de-identification techniques, although Linker can be used to concatenate and then Anonymizer to mask these attributes. De-identification is particularly problematic with network data (Wu, Xiao, Wang, He, & Wang, 2010), but occur even in conventional data (Winkler, 2004), thus usage of both Linker and Anonymizer can minimize these risks

Although intuitive and sensible that larger Edit Distances should be allowed given more likelihood of human error, this enhancement offers a complication of order dependence in how the strings to be evaluated are processed. Imagine a scenario where strings equal to or shorter than 4 characters are exact matches (Minimum Edit Distance = 0), but strings 5 or larger are allowed to match with an Edit Distance of 1: thus the order of two tokens "Jane" and "Janet" matter. If "Janet" is processed first, then both "Janet" and "Jane" will be retained as unique names, but if "Jane" is processed first, then "Janet" will be evaluated as similar to "Jane" and only "Jane" will be retained. A future iteration of the tool might force ordering of all tokens to be processed from largest to smallest, but the current tool does not. This problem is most acute at the threshold between exact match only and minimum edit distance matching and thus is relatively rare in our default configuration.

In this short research note, we have introduced Anonymizer and Linker: open-source tools which address common data collection cleaning and processing needs. These tools are controlled via straight-forward configuration files that are easy to edit but also can be saved to ensure consistent treatment of collected data as it arrives, and provide significant capability otherwise difficult to reach without significant programming experience. We hope these open-source tools may be of interest to members of the community and that they will continue to evolve through collaboration and become more useful to the research community and our partners.

6 References

Amazon. (2005). Mechanical Turk. Retrieved from http://www.mturk.com

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). UCINET for Windows: Software for social network analysis. Harvard, MA: Analytic Technologies.

Carley, K. M., Pfeffer, J., Reminga, J., Storrick, J., & Columbus, D. (2013). *ORA User's Guide 2013*. Retrieved from Pittsburgh, PA:

Morgan, G., & Garbach, K. (2016). Network Inference Utilities. Retrieved from https://github.com/geoffrey-p-morgan/SurveyInference

Okuda, T., Tanaka, E., & Kasai, T. (1976). A method for the correction of garbled words based on the Levenshtein metric. *IEEE Transactions on Computers, 100*(2), 172-178.

Pywell, R., Warman, E., Sparks, T., Greatorex-Davies, J., Walker, K., Meek, W., . . . Firbank, L. (2004). Assessing habitat quality for butterflies on intensively managed arable farmland. *Biological Conservation, 118*(3), 313-325.

Qualtrics, L. (2014). Qualtrics [Software]: Provo: Author.

R Development Core Team. 2011. R: A language for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM), 21*(1), 168-173.

Winkler, W. E. (2004). *Re-identification methods for masked microdata.* Paper presented at the International Workshop on Privacy in Statistical Databases.

Wu, W., Xiao, Y., Wang, W., He, Z., & Wang, Z. (2010). *K-symmetry model for identity anonymization in social networks.* Paper presented at the Proceedings of the 13th international conference on extending database technology.