Incorporating Record Linkage Measurement Error into Descriptive Network Metrics

Abby Smith

Northwestern University, Department of Statistics

Objectives

After conducting a literature review and identifying application areas, we aim to:

- Understand how record linkage errors arise in network analysis, and take them into account when describing a network structure
- Develop "best reporting practices" for errors in a linked dataset when performing basic network inferences

Introduction

Record linkage aims to find matches between groups of records from separate datasets in an effort to recover the true unique people across multiple datasets. The linkage process is often crucial in Census work, the creation of health indicators, and historical research. Records are susceptible to various forms of distortion: letters can be out of order or missing, punctuation can be misplaced, and abbreviations can be varied. For recent advances in record linkage, see papers by Sadinle (2017) and Steorts (2015) who specify Bayesian/latent class models for record linkage.

Datafile 1

Name	DOB	...
John M. Doe	Feb/11/1990	...
John H. Doe	Apr/24/1990	...
John G. Doe	Oct/03/1990	...
...	...	...
Juan A. Gómez	Jul/NA/1950	...
...	...	...

Datafile 2

Name	DOB	...
John Doe	NA/NA/1990	...
...	...	...
...	...	...
Juan Gómes	Jul/NA/1950	...
Juan A. Cómez	Jul/02/1950	...
...	...	...

Figure 1:An example linkage problem

There are even more sources of measurement error in a network; for example even if respondents report the correct spellings of friends names, the understanding of what qualifies as a friendship tie could vary.

Northwestern
TGS

Probabilistic Record Linkage

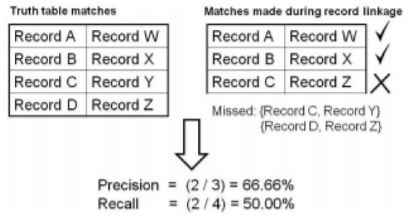
1. **Generate potential comparison pairs**. Block to reduce computation time, i.e.- require that to consider a pair of records it match on a logical string (ex. first name AND last name OR Date of Birth).
2. **Generate a comparison vector** for each potential pair. i.e.- If there are 3 fields (name, birth date, address), you could observe 100, 101, 001, etc.
3. For each of the records: **compute a m probability**: the probability that you would observe a certain "comparison pattern" given that the pair of records is a true match. Compute a u probability (non-match).
4. Assign a **weight** based on how well it matches, using the EM algorithm.
5. Calculate a **composite score** for each pair from all of the matching weights.
6. Determine whether a record pair is a match, non-match, or possible-match by **comparing each composite score** to a given threshold.

Quantifying Linkage Error

		True link status	
		Match	Non-match
Predicted link status	Match	d (true match)	b (false match)
	Non-match	c (false non-match)	a (true non-match)

$$\text{Precision} = \frac{b}{b+d} = P(\text{true match}|\text{predicted match})$$

$$\text{Recall} = \frac{c}{c+d} = P(\text{predicted match}|\text{true match})$$



In the literature, precision and recall are combined into one number, and the F_1 score is often reported:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The Big Picture

With messy social science data, erroneous linkage among individuals could have serious implications for both structural and node-centric network metrics.

Measurement Error in Networks

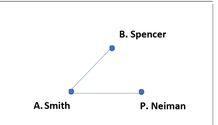
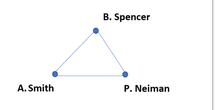
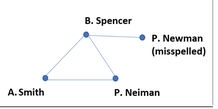
Network Types	Description	Illustration
True network	Network of true relations between entities	
Clean network	Network of relations encoded in data without measurement error	
Observed network	Network of relations encoded in data with measurement error	

Figure 2:Types of networks

Examples of Network Error

Error	Example
False negative nodes	Non-response in the Census
False positive nodes	Fake accounts in an internet sample
False negative edges	Imperfect respondent recall about friends
False "aggregation"	Coauthorship network: mistakenly treating different authors as the same author
False "disaggregation"	Coauthorship network: mistakenly treating the same author as different authors

Network Metrics

How are these affected by record linkage error? In what cases are they more/less sensitive?

- Degree distribution
- Average clustering
- Degree centrality
- Clustering coefficient
- Network constraint
- Eigenvector centrality

Questions of Interest

- How can standard reported metrics of linkage quality (such as the F_1 measure) inform network analysis? Are there are other (non-standard) metrics to consider?
- How can information about network measurement errors be meaningfully reported to the user, particular when model fitting?

Acknowledgements and References

Thank you to my advisor Bruce Spencer for his guidance and useful feedback.

For a list of references: `http://bit.ly/abby_WSDS2019`