Algorithmic Reconstruction of Inverted Relationships Using the Alyawarra Kinship Glossary

By Amin Fozi

Advisor: David Meyer

Abstract

Woodrow W. Denham's study of a group of Alyawarra-speaking Aboriginal people lays out a set of asymmetrical kinship pairings of tribe members, and a set of rules for understanding what these kinship terms mean. Using these rules, we create an algorithm that inverts the definitions of these kinship terms, test its ability to guess Denham's full data given a subset, and use it to reconstruct data that he never took.

Acknowledgements

I would like to thank Professor David Meyer for showing me this data and research, and helping me find a way to expand on it. His guidance in devising the data-testing method for this algorithm was absolutely essential. Thanks to him I have gotten to learn a lot about a very fascinating subject, and thanks to his group meetings this thesis has been made possible.

Table of Contents

Introduction

In 1971-1972, Woodrow W. Denham did field research with a group of 264 Aboriginal Australians in the Alyawarra language group[1]. At point in this research, he surveyed 104 members of the group to describe their familial relation to 225 fellow tribespeople. The latter group included all 104 of the former. Throughout the paper, this section of Denham's data will be referred to as the "kinship survey".

This study was brought to our attention by the papers *Learning Systems of Concepts with an Infinite Relational Model*[2] and *Discovering Latent Classes in Relational Data*[3]. These papers, written by MIT's Charles Kemp, Thomas L. Griffiths and Joshua B. Tenenbaum, largely concern processes that turn relationships between objects into ways of classifying and clustering them. Both of them, as an example, use their statistical model to analyze the kinship survey.

With just the data from the kinship survey, they were able to cluster the surveyed people into groups nearly homogenous in age, gender, and culturally significant kinship section.

This paper will describe another technique for analyzing these directed two-person relationships. We look at ways in which relationships imply other relationships. Using what can be known, assumed, or guessed about the 121 people who were asked about but not surveyed in this study,

I coded and tested a method of inverting Alyawarra kinship terms.

The Alyawarra Kinship System

       Members of the Alyawarra tribe are split into four kinship sections: Kamara, Pityara, Burla, and Ngwariya[3]. These sections will be referred to by numbers from 1-4, for convenience. The way that section is defined is by parentage, as follows:

       Members of section 1 are expected to have children with members of section 2, and members of section 3 are to be with members of section 4. A child is in a different section from their parents, and two biological siblings are in the same section. In understanding relationships between Alyawarra members, section is important.
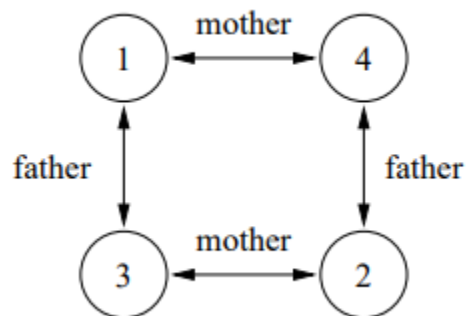
Go to [1] for diagram source

Denham had the following glossary of kinship terms[3]:

F=father, M=mother, B=brother, Z=sister, S=son, D=daughter, H=husband,
W=wife, E=elder, fs=female speaker, ms=male speaker

01	Arengiya	FF/FFZ, SS/SD (ms), BSS/BSD (fs)
02	Anyainya	MM/MMB, MMBSS/MMBSD, ZDS/ZDD (ms), DS/DD (fs)
03	Aidmeniya	MMBSS/MMBSD, ZDS/ZDD (ms), DS/DD (fs)
04	Aburliya	FM/FMB, FMBSD/FMBSS, ZSS/ZSD (ms), SS/SD (fs)
05	Adardiya	MF/MFZ, DS/DD, BDS/BDD (fs)
06	Agngiya	F
07	Aweniya	FZ, FMZD
08	Amaidya	M, SW (ms)
09	Abmarliya	MB, SWB (ms)
10	Awaadya	EB
11	Anguriya	EZ
12	Adiadya	YB/YZ
13	Angeliya	FZS/MBS
14	Algyeliya	FZD/MBD
15	Adniadya	MBS
16	Aleriya	S/D (ms), BS/BD (fs)
17	Umbaidya	S/D (fs), ZS/ZD (ms), FMBS/FMBD
18	Anowadya	W/MMBDD (ms), H/MFZDS (fs)
19	Muriya	MMBD/MMBS, WM/WMB (ms), ZDH/ZDHZ (ms)
20	Agenduriya	ZS/ZD (ms), rare form for biological (proper) sister's child
21	Amburniya	WB/ZH
22	Andungiya	HZ/BW (fs)
23	Aneriya	BWM/DHZ (fs)
24	Aiyenga	"Myself"
25	Unknown	
26	Undyaidya	WZ (ms), rare form used as reciprocal for amburniya (WB)
27	Gnaldena	YZ, rare form for biological (proper) younger sister
28	Dead	
29	No Response	

The Kinship Survey

As described earlier, Denham showed 104 of the people in the study 225 pictures of people, and asked them to use a word to describe their kinship. He used their responses to build the kinship glossary, and filled in a spreadsheet using the numbers he assigned to each word. This spreadsheet has 104 rows and 225 columns[4] – but there are another theoretical 121 rows that could be created.

Reconstructing Alyawarra Relationships

Each described Alyawarra relationship has an inverse relationship. This is taken by seeing each glossary definition as a path of relationships and genders, and reversing those relationships with the proper genders in place. E.g. the inverse of a male speaker's wife's brother – #21 WB (ms) - is a male speaker's sister's husband – #21 ZH (ms); the relationship path is spouses to siblings, the gender path is male to female to male. Note how both of these terms are described with the 21st term "amburniya". Many of the terms defined in the glossary will have both a relationship and its inverse as definitions, but there will be other options too.

Each term has a list of relationships it describes, and thus a list of relationships to which it can be opposite, which is then narrowed into the list of terms that can be used to describe at least one of those relationships. The list of relationships can be narrowed down given subjects' age ranges, kinship sections, and genders.

I created a mapping of kinship terms onto sets of kinship terms, where each relationship is given a possible inverse relationship. To test this mapping, I used it to reconstruct the survey answers of 34 participants who gave survey answers. By simply picking the first member of each inverse set, I got 1214 matches and 1096 mismatches. I further developed the algorithm until the tested accuracy was the highest it could be.

Data Testing

The logic for testing the method was as follows. Theoretically, the excel spreadsheet for the kinship data could be represented by the white rectangular diagram below, with the top rectangle being the spreadsheet itself and the bottom rectangle being data that was not collected. The vertical axis is the number for the person speaking and the horizontal axis is the number for the person being spoken about.
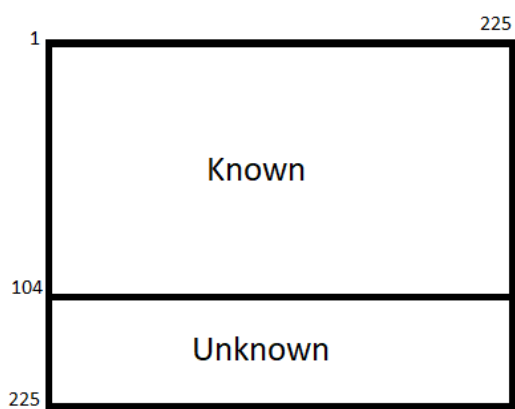
But the boxes could be split further. Of the known data, there were cells where the subject's stated relationship to the speaker (the relationship's inverse) was known, and cells where it was unknown. Similarly, some of the unknowns were invertible, while others were not.
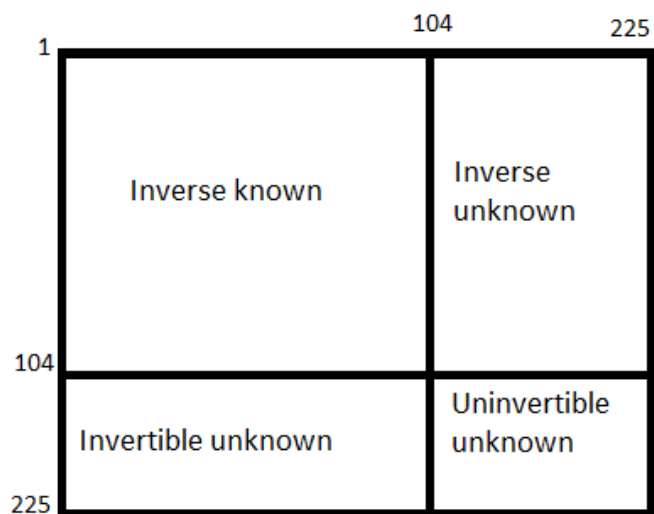
If kinship terms could be reliably inverted, the top right box could be used to predict the bottom left box. The bottom right box would remain unknown.

To indicate how accurately the algorithm could invert relationships, we must split the graph further. In the following diagram, the invertible knowns are turned into a recreation of the previous diagram. The situation is identical to the situation above if a bottom section of the invertible knowns is ignored. By "ignore", it is meant that we will try to discover and recreate this data despite already having it. That way the recreation can be tested for accuracy.

The key difference between this white 2x2 diagram and the previous one is that when predicting invertible "unknowns" using their inverses, one can check their answers.

While I optimized my answers using census and kinship data, I also took steps to make sure that the answers I formed were not biased towards the sample of answers I already had. The first step taken was to divide the square further.

This method of prediction was to maximize the accuracy of the algorithm's prediction of the bottom left white box using the top right white box. Once the accuracy of the white box was as high as it could get, the next step was to measure the accuracy of predicting the left orange box with the top orange box, using the exact same algorithm. If that accuracy were high, despite its data being no more associated with the testing than the unknown bottom red box data, then predictions of the bottom left red box data were likely to have a similar level of accuracy.

This would be a fine testing method to use, if it were the case that indices did not correlate to any demographic features. However, this was not true. Indices generally correlated with the ID numbers that Denham assigned to the tribespeople, and the first 280 ID numbers had sexes assigned to them (males 1-140, females 141-280).

A slight, but important tangent: It bears noting that the columns representing un-surveyed people in Denham's spreadsheet were spread throughout the sheet, and are drawn as on the right

side of the spreadsheet for the sake of convenient understanding. It is not the case that the columns in the rightmost red box are mostly women.

If I continued to work on the data in this manor, my results would be more useful predicting the answers of male tribe members, and not everyone equally, and they would not be provably considerate of the way sex affects answers. There were 62 male tribe members among the non-surveyed 121, and 59 female, so it would not do to only be able to predict men's responses in creating the algorithm.

So I abandoned the clustering of the indices in proximity and went ahead to tune the algorithm so that it would optimize indices based on modulus groups. I used the spreadsheet indices that were 0 (mod 3) to predict the results of the indices that were 1 (mod 3). Once I got that accuracy as high as I could, I recorded the accuracy of predicting indices that were 2 (mod 3) using the other indices. The method could be described by the following re-imagining of the diagram.

Complications and Further Elimination Using Census and Kinship Data

The reason why the inverse sets were not guaranteed to contain the desired relationship was simple: people have multiple relationships to one another, particularly in a society that allows cousin marriage under certain conditions. For example, if ones marries a cousin, then that cousin's mother could be referred to as either a mother-in-law or an aunt.

There were various other quirks that made the glossary appear flawed. These included instances of people mutually calling each other a term that had been defined as "older brother", and the word amaidya (08), a term the glossary describes as between people in each other's mother group, being the inverse to a great variety of relationships.

So while the inverse relationship sets were a promising start, they did not fully describe the range of kinship terms that could exist as inverses of each other. I updated each set to make sure it contained each of the most stated inverse relationships, and then removed members that while in theory were inverses, in practice were only a small portion of inverse responses, if any at all.

After that, there was the matter of narrowing down a set with various possible results. I used census data[5] to accomplish this. The first and most reliable tool was sex – a relationship described by a male speaker could only be inverted into a term that could be used on men, and vice-versa. A survey result describing a woman could not be inverted into a term that was exclusively for male speakers, and vice-versa.

Age was the other major contributor. Age differences between the surveyed people were used to narrow down inverted relationships.

Eventually, most terms had their inverse chosen chosen correctly a majority of the time within the optimized data subset. The exceptions were:

- 22, which was 46.667% accurate.

- 24, "Self". This relationship was never tested for inversion by the nature of the testing, but it is known to be its own inverse 100% of the time.

- 25, "unknown". Did not appear in the final tests. Defaulted to 6 and 17, two of the most popular relationships, then narrowed using sex data, as 6 was a male term.

- 26 and 27, rare forms with extremely little representation in the data.

- 28, "Dead" and 29, "No response". There was no way to predict these inverses, but they were also very rare.

Results

These are the percentages of correctly inverted reversed results. Here, "optimized percentage" refers to the data range that the code was created to optimize the given quantity. The other percentages show that the results are almost equally strong in other ranges of data.

The most indicative result is the percentage accuracy from using 1 (mod 3) and 0 (mod 3) indices to predict indices that were in 2 (mod 3). This has the same advantages as #1, but it also has more relationships to test and unlike #1, none of the data tested in it was used to optimize the second algorithm. This percentage on the second algorithm is bolded to show that it is the least biased indicator of this model's veracity.

Trained on 1-40 predicting 41-70 (potentially male-biased)

Using 1-40 to predict 41-70 (optimized percentage): 83.263246% (990/1189)

Using 1-70 to predict 71-104: 76.408451% (1736/2272)

Trained on 0 (mod 3) predicting 1 (mod3)

Using indices 0 (mod3) to predict 1 (mod3) (optimized percentage): 81.228374% (939/1156)

Using indices 0 and 1 (mod3) to predict 2 (mod3): 76.086957% (1785/2646)

Using evens to predict odds: 78.777393% (2049/2601)

Using 1-40 to predict 41-70: 80.319596% (955/1189)

Using 1-70 to predict 71-104: 78.080986% (1774/2272)

Having obtained significantly accurate results, I created a square version of the kinship data spreadsheet[7] where the unknown data cells with known inverses were predicted, and the group of cells equivalent to the bottom right red box from the diagrams was represented by the number 30.

Conclusion and Possible Further Studies

The high accuracy of the algorithm in testing suggests that the generated data is mostly accurate. Had the other 121 people in the photos been surveyed by Denham, it is likely many of them would have spoken as the new sheet describes.

It would be possible to use the principles in this paper to recreate the "lower right" box of the spreadsheet: the un-invertible unknowns. While the non-surveyed 121 do not have listed relationships with each other, the reconstructed data describes them as kin of kin. In fact, every pair between the 121 can be described as a kin's kin in the new spreadsheet.

One could create a list of which pairs of relationships have which inverse relationships (e.g. a #6's #6 would be a father's father FF, which could then be inverted into a male speaker's son's son or daughter – #1 SS/SD (ms)). From there, train the algorithm on a subset of the data until it is sufficiently accurate.

This would result in a square spreadsheet with no missing data points, and we could once again run the Infinite Relational Model on it to see if it better or worse clusters the tribespeople by age, sex, and section.

References

[1] Woodrow W. Denham. *Alyawarra Ethnographic Archive.* URL: http://www.culturalsciences.info/AlyaWeb/. Last visited 2018/06/06.

[2] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Uedo. *Learning Systems of Concepts with an Infinite Relational Model*. MIT, 2006.

[3] Charles Kemp, Thomas L. Griffiths, & Joshua B. Tenenbaum. *Discovering Latent Classes in Relational Data*. MIT, 2004.

[4] Woodrow W. Denham. *Alyawarra Ethnographic Database: Numerical Data Documentation File, Version 7*. URL: http://www.culturalsciences.info/AlyaWeb/public/dir/03dbNum/File00Doc.pdf. Last visited 2018/06/05.

[5] Woodrow W. Denham. *Kinship Data*. URL: http://www.culturalsciences.info/AlyaWeb/public/numericTOC.html as "File22", or http://www.culturalsciences.info/AlyaWeb/public/dir/03dbNum/File22aKinshipViewer.xls. Last visited: 2018/06/05.

[6] Woodrow W. Denham. *Vital Statistics and Genealogical Data*. URL: http://www.culturalsciences.info/AlyaWeb/public/dir/03dbNum/File01VSGDClean.pdf. Last visited: 2018/06/05.

[7] Available on request sent to AFozi732@gmail.com