MODELING DISAGREEMENT WHEN NETWORK DATA DISAGREE

RICHARD MCELREATH

May 23, 2017

1. Problem

Anthropologists and other interesting people sometimes ask about *social networks*. Measuring social networks can mean everything from tracking people with radio waves to interviewing people about who is in their network. For interview data in particular, it is common for there to be disagreement among individuals about the existence of a network connection, a *tie*. Common examples may include:

(1) Sexual networks: Individuals may make different reports, with either party in a potential relation being more likely to report a tie. For example, A may report a relationship with B, but B not with A. However ties must be, in reality, reciprocal. If some classes of individuals are reluctant to mention ties, or tend to overreport ties, this makes the underlying reality harder to discern.

(2) Helping networks: When both individuals in a dyad are asked about helping in both directions, both from A to B and from B to A, it is possible for the reports to disagree. While ties do not have to be reciprocal here, the underlying reality must be reconciled with conflicting reports about each direction. If some individuals are generally reluctant to mention giving or receiving aid, this may result in substantial disagreement about the reality.

Typical approaches to resolving these conflicts include (1) ignoring them, (2) declaring (outside the model) that some reports are more reliable than others, and (3) using the rule that if any individual reports a tie, the tie must be real. None of these is entirely satisfactory, as they treat ties associated with conflicting reports as equally certain as ties associated with agreement.

The question is whether a more principled approach is possible. We believe it is. We believe we can Bayes this.

Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

E-mail address: richard_mcelreath@eva.mpg.de.

2. Model

The general model asserts only a distinction between the underlying reality of a tie and any particular report about that tie. Let T_{nij} be the probability of tie state n for dyad ij. These ties could be binary 0/1 ties or continuous measures of tie strength. Then the probability an individual k reports tie state R_{nijk} between individual i to j is:

$$R_{nijk} = \sum_m T_{mij} P(n|mijk)$$

This is nothing more than a definition of total probability, summing over the probability of each true state of the tie, T_{mij}, and the conditional probability P of a reported tie given each true state. In plainer language, the probability of a reported tie is just the average probability of a reported tie, averaging over the unknown true states of the relationship.

To do anything with this, we require a specific data context and parameterization of both how ties are generated and how they are reported, conditional upon the true state of the tie. Many models are possible, but each must specify both (1) how ties actually form, conditional upon features of the individuals, as well as (2) how reports are generated, conditional upon the true state of the dyad.

Consider a very simple model of tie formation, for example. Suppose the ties of interest are binary 0/1 helping relationships. First, assume an average rate of helping across all dyads, β. Other effects will be modeling as deviations from this average rate. Let individuals i possess both general giving and receiving rates, g_i and r_i, respectively. An individual with a large value g_i tends to donate more help, and so has more directed ties $i \to j$. An individual with a large value r_i tends to receive more help, and so has more directed ties $j \to i$. In addition, allow that an individual i tends to help an individual j in particular, as measured by a dyad-specific parameter d_{ij}. The reverse, d_{ji}, may or may not be of the same magnitude, which allows dyads to be reciprocal or not.

Again, we consider the case where help from i to j, y_{ij}, is binary. Then the above model indicates:

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$
$$\text{logit}(p_{ij}) = \beta + g_i + r_j + d_{ij}$$

In principle, help could be quantitative and depend upon other factors. But this is perhaps the simplest model with differences in general helpfulness and special dyadic relationships. These parameters g, r, and d could be functions of covariates or merely varying effects.

The states y_{ij} are the real ties, but these can not be easily observed in most cases. But many different kinds of data can be used to infer them, such as self-reports and behavioral observations. All methods will entail some observation error. Consider here the case of interviews in which each individual is asked to report on both who they have helped and who has helped them. This generates symmetrical reports, in which both individual in a dyad ij report on both $i \to j$ and $j \to i$. These reports may easily disagree.

Let $q \in \{1, 2\}$ indicate whether an individual was reporting their own help or help received, respectively. Let $\bar{h}_q$ be the average rate of reported help for question q and h_{iq} be the individual-specific deviation from the average for individual i and question q. Finally, assume that true ties are reported more often, with $m > 0$ being the increased rate of reporting true ties. Note however that non-ties may also be falsely reported as existing.

With the above assumptions, a report of directed help v_{ijkq} made by individual $k \in \{i, j\}$ about $i \to j$ for question type q is:

$$\begin{aligned} v_{ijk} &\sim \text{Bernoulli}(\pi_{ijk}) \\ \text{logit}(\pi_{ijk}) &= \bar{h}_q + h_{kq} + m y_{ij} \end{aligned}$$

While the true states y_{ij} cannot be observed, they can be inferred within the context of the model, which treats them as latent variables. Posterior distributions for all y_{ij} and all of the parameters can be computed from observations v_{ijk}.

The posterior probability for $y_{ij} = 1$, conditional on the observations $v_{ij} = \{v_{iji}, v_{ijj}\}$, is given by:

$$\begin{aligned} P(y_{ij} = 1 | v_{ij}) &= \frac{P(v_{ij}, y_{ij} = 1)}{P(v_{ij})} = \frac{P(v_{iji}|y_{ij} = 1)P(v_{ijj}|y_{ij} = 1)P(y_{ij} = 1)}{P(y_{ij} = 1)P(v_{ij}|y_{ij} = 1) + P(y_{ij} = 0)P(v_{ij}|y_{ij} = 0)} \\ &= \frac{(\pi_{iji}\pi_{ijj})|_{y_{ij}=1} \cdot p_{ij}}{(\pi_{iji}\pi_{ijj})|_{y_{ij}=1} \cdot p_{ij} + (\pi_{iji}\pi_{ijj})|_{y_{ij}=0} \cdot (1 - p_{ij})} \end{aligned}$$

3. Validating the model

Model validation proceeds by first simulating data from the model. This requires plugging in values for all of the variables, except for y_{ij} and v_{ijk}. The y_{ij} values are simulated first as Bernoulli random numbers from the definition of p_{ij}. Then the observable v_{ijk} values are simulated from the definition of π_{ijk}.

We programmed the statistical model in Stan and drew samples from the posterior distribution of the model and simulated data. This allows us to validate both theoretical usefulness of the approach as well as the validity of our code. Because Stan does not allow sampling for discrete parameters, we

93 used the definition of the posterior distribution $P(y_{ij} = 1|\upsilon_{ij})$ to compute these in Stan's generated quantities block.