Cultural Consensus Theory: Estimating Consensus Graphs Under Constraints

Kalin Agrawal
William H. Batchelder
UC Irvine
kagrawal@uci.edu

Society for Mathematical Psychology
Annual Meeting
Tufts, July 2011

We acknowledge the support of
Army Research Office Grant 58261LS,
"Statistical Inference for Cultural Consensus Theory,"
PI William H. Batchelder

Agenda

Cultural Consensus Theory

Graph Aggregation

Our Constraint: Balance

CCT Tie Model

Bayesian Inference Using MCMC Sampling

Data Analysis

Conclusion

Cultural Consensus Theory (CCT)

Initially developed by Romney, Batchelder, Weller in 1980s.

E.g. Romney et al (1986, Am. Anth.)

Intuition: ‘Test Theory Without an Answer Key’.

Multiple informants’ responses to questions.

Data aggregation (‘answer key’) and informant calibration (competence, bias).

CCT-related models for different question formats, e.g.,

- True/False, Multiple-Choice E.g. Romney et al (1986, Amer. Anth.)
- Continuous Batchelder et al (2010, Adv. Soc. Comp.)
- Ranked items Romney et al (1987, Am. Beh. Scientist)
- Directed graphs Butts (2003, Soc. Net.), Batchelder et al (1997, J. Math. Soc.; 2009, Soc. Comp. & Beh. Mod.)

Why would one want to aggregate graphs where there might be some objectively true graph? This is the same question as for aggregating in any CCT-type situation.

Graph Aggregation

We ask informants to provide edge values in various types of graphs.

Social network applications:

- Friendship/advice networks (e.g. informants report on ties in their own social network).
- Covert networks (e.g. informants report on ties between others).

Imposing Constraints on Graphs

For example,

- Total order: If $a \leq b$ and $b \leq a$ then $a = b$ (antisymmetry);
  If $a \leq b$ and $b \leq c$ then $a \leq c$ (transitivity);
  $a \leq b$ or $b \leq a$ (totality).
- Equivalence (set partition): $a \sim a$ (reflexivity);
  If $a \sim b$ then $b \sim a$ (symmetry);
  If $a \sim b$ and $b \sim c$ then $a \sim c$ (transitivity).
- Structural balance (two-cell partition): A two-cell equivalence relation, but with some history in the literature of social dynamics. Cartwright, Harary (1956, Psych. Rev.)

Key Idea

We hypothesize that the consensus graph satisfies a particular constraint, but we do not presume that each informant's response satisfies the constraint due to error and/or lack of knowledge.

Our Constraint: Balance

Some notation:

M informants, indexed by i.

$\mathcal{V}$ is the set of vertices (corresponds to node items).

N vertices, indexed by $j, k \in \mathcal{V}$.

$\mathcal{E}$ is the set of undirected edges (corresponds to item-pair questions).

$\binom{N}{2}$ edges, indexed by $\{jk\} \in \mathcal{E}$, and $\{jk\} = \{kj\}$.

Balanced Graph

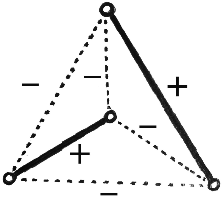
Let $G = (\mathcal{V}, \mathcal{E})$ be a *simple*, *undirected*, *complete* graph.

Let $\Sigma = (G, \sigma)$, be a *signed* graph, where $\sigma : \mathcal{E} \to \{-, +\}$.

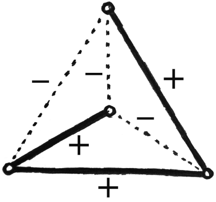
Σ is *balanced*

- iff the product of edge signs is positive along every cycle.
- iff $\mathcal{V}$ can be partitioned into complementary cells, A and A^c, such that $\forall\ j, k \in \mathcal{V}$:
  - $\sigma(\{jk\}) = +$, if $j, k \in A$,
  - $\sigma(\{jk\}) = +$, if $j, k \in A^c$,
  - $\sigma(\{jk\}) = -$, otherwise.

  That is, two-cell equivalence relations induce a balanced graph.



Balanced



Unbalanced

Tie Model I

Consensus partition $\mathbf{W}$ (with its logical complement, $\overline{\mathbf{W}}$),

$$W_k = \begin{cases} 1 & \text{if vertex } k \in A, \\ 0 & \text{if vertex } k \in A^c. \end{cases}$$

$\mathbf{Z}$ is the matrix of all (coded) edge signs on the consensus graph,

$$Z_{jk} = \begin{cases} 1 & \text{if } W_j = W_k, \\ 0 & \text{if } W_j \neq W_k. \end{cases}$$
$$= 1 - (W_j - W_k)^2$$

Tie Model II

Observed data, $\mathbf{X}$, is an *informant* × *vertex* × *vertex* array,

$$X_{i,jk} = \begin{cases} 1 & \text{if informant } i \text{ reports edge } \{jk\} \text{ is positive,} \\ 0 & \text{if informant } i \text{ reports edge } \{jk\} \text{ is negative.} \end{cases}$$

Informants' competences (probabilities of knowing the sign of an edge) is the vector $\mathbf{D}$.

Informants' guessing biases (probabilities of reporting an unknown edge is positive) is the vector $\mathbf{g}$.

Tie Model III

High Threshold Signal Detection Model,

$$\Pr(X_{i,jk} = 1 | Z_{jk} = 1, D_i, g_i) = (1 - D_i)g_i + D_i$$
$$\Pr(X_{i,jk} = 1 | Z_{jk} = 0, D_i, g_i) = (1 - D_i)g_i$$

Thus, the probability of any individual response,

$$\Pr\left(X_{i,jk} \mid Z_{jk}, D_i, g_i\right) = [(1 - D_i)g_i + D_i]^{X_{i,jk}Z_{jk}} [(1 - D_i)g_i]^{X_{i,jk}(1-Z_{jk})} \\ [(1 - D_i)g_i]^{(1-X_{i,jk})Z_{jk}} [(1 - D_i)g_i]^{(1-X_{i,jk})(1-Z_{jk})}$$

With conditionally independent responses (across edges), the likelihood is a big product,

$$L(\mathbf{X}|\mathbf{Z}, \mathbf{D}, \mathbf{g}) = \prod_{i=1}^{M} \prod_{k=2}^{N} \prod_{j=1}^{k-1} \Pr\left(X_{i,jk} | Z_{jk}, D_i, g_i\right)$$

Bayesian Inference Using MCMC Sampling

Priors are uninformative (flat):

- For partition: $W_k \sim Bernoulli(1/2)$ means two vertices just as likely to be in the same cell as different cells.
- For an informant's competence and bias: $D_i \sim Unif(0,1)$ and $g_i \sim Unif(0,1)$.

Markov Chain Monte Carlo sampler:

- Metropolis step for partition, $\mathbf{W}$, means we only sample balanced $\mathbf{Z}$.[1]
- Metropolis-Hastings step for each D_i and g_i.

[1] $\mathbf{W}$, $\overline{\mathbf{W}}$ are unidentified, but $\mathbf{Z}$ is identified.

Simulated Data

Simulated response data according to response model.

Applied sampler to estimate generating parameters.

Recovery of $\mathbf{W}$, $\mathbf{D}$, $\mathbf{g}$.

9000 iterations, 1000 burned, thinning interval of 8.

Obtaining Real Data

We wanted to have tie data with a known ground truth, but this was hard to find.

We created 'tie data' using nodal attributes of the graph.

Two surveys:

- 5 basketball players, 5 baseball players (vertices); 'Play same sport?' (edges).
- 5 Arizona cities, 5 New Mexico Cities (vertices); 'In same state?' (edges).

Survey Design Issues

Complete design involves $\binom{N}{2}$ questions.

Want to avoid logical inference from cycles, e.g. $\sigma(\{AB\}) = +$ and $\sigma(\{BC\}) = +$ implies $\sigma(\{AC\}) = +$.

Survey Design Issues

Complete design involves $\binom{N}{2}$ questions.

Want to avoid logical inference from cycles, e.g. $\sigma(\{AB\}) = +$ and $\sigma(\{BC\}) = +$ implies $\sigma(\{AC\}) = +$.

Example

Edge	Same state	Diff states
Roswell, Taos	×	

Survey Design Issues

Complete design involves $\binom{N}{2}$ questions.

Want to avoid logical inference from cycles, e.g. $\sigma(\{AB\}) = +$ and $\sigma(\{BC\}) = +$ implies $\sigma(\{AC\}) = +$.

Example

Edge	**Same state**	**Diff states**
Roswell, Taos	×	
...		
Taos, Carlsbad	×	

Survey Design Issues

Complete design involves $\binom{N}{2}$ questions.

Want to avoid logical inference from cycles, e.g. $\sigma(\{AB\}) = +$ and $\sigma(\{BC\}) = +$ implies $\sigma(\{AC\}) = +$.

Example

Edge	**Same state**	**Diff states**
Roswell, Taos	×	
...		
Taos, Carlsbad	×	
...		
Roswell, Carlsbad		

Survey Design Issues

Complete design involves $\binom{N}{2}$ questions.

Want to avoid logical inference from cycles, e.g. $\sigma(\{AB\}) = +$ and $\sigma(\{BC\}) = +$ implies $\sigma(\{AC\}) = +$.

Example

Edge		**Same state**	**Diff states**
Roswell, Taos		×	
...			
Taos, Carlsbad		×	
...			
Roswell, Carlsbad	$\Rightarrow$	×	

Survey Design Issues

Complete design involves $\binom{N}{2}$ questions.

Want to avoid logical inference from cycles, e.g. $\sigma(\{AB\}) = +$ and $\sigma(\{BC\}) = +$ implies $\sigma(\{AC\}) = +$.

Example

Edge		Same state	Diff states
Roswell, Taos		×	
...			
Taos, Carlsbad		×	
...			
Roswell, Carlsbad	$\Rightarrow$	×	

Need to sequence questions to avoid logical inferences, or make them less likely.

How to avoid logical inferences, or make them less likely?

Special order of pairwise questions for $N = 10$:

- 'Front-load' questions that complete fewer and larger cycles, 'back-load' questions that complete more and smaller cycles.
- Separates questions into three phases, based on potential for balance computation. (1-10, 11-25, 26-45)

Missing data handled in the likelihood function by setting $\Pr\left(X_{i,jk} = \text{missing} \mid Z_{jk}, D_i, g_i\right) = 1$.

- By design, if discarding later-phase questions.
- Accidental, for a skipped question.

Ball Players Survey Results

Data: 5 of 855 edges blank.
Elicited confidences:

Don't know	516
Unsure	179
Certain	148
N/A	12

Correctly recovered true partition.
Mean marginal **W**, Q1-10:
(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)
Mean marginal **W**, Q1-45:
(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)

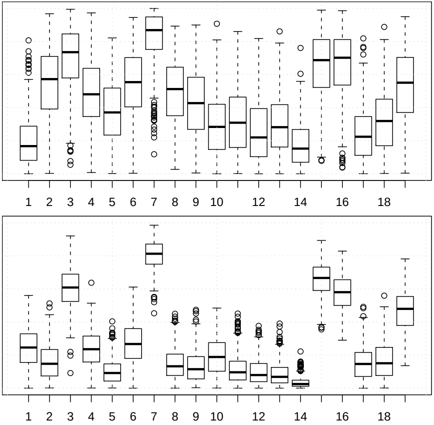


Figure: Marginal posterior **D**, Q1-10 (top), Q1-45 (bottom).

Conclusions

We think we have a good working model for *tie-based* responses.

Unfortunately, our experimental data involved *nodal-based* responses.

If one knows Hank Aaron is a baseball player in one dyad, she will know it in all dyads involving Aaron.

We need a better model for nodal-based responses.

We need good data for tie-based responses.

A nodal model we are now working with assumes that an informant either knows or doesn't know the type of each node, and knows the tie iff she knows both nodes, otherwise a guess is made.

This nodal model implies that tie responses are **not** conditionally independent, given the parameters.

This makes the MCMC sampler more complicated. Basically, data augmentation based on each informant's subset of known nodes is needed.

We are working on the sampler for the nodal model and looking for tie-based response data.

Thanks!

Appendix

Simulated Data Parameter Recovery I

Tests:

- Perfectly correct informants ⇒ correct $\mathbf{W}$, high D_i, uniform g_i; confirmed.
- Perfectly wrong informants ⇒ unchanged $\mathbf{W}$, low D_i, g_i approach (number of negative edges / number of positive edges) $= 25/45 = 0.5555$; confirmed.

Simulated Data Parameter Recovery II

- Fixed **W**, ranges of **D** and **g**.

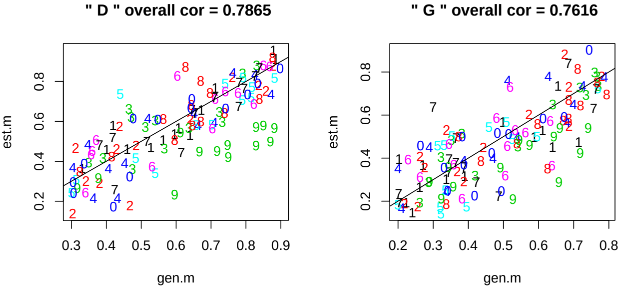


Figure: Recovery plots

Simulated Data Parameter Recovery III

- Various **W**, ranges of **D** and **g**.

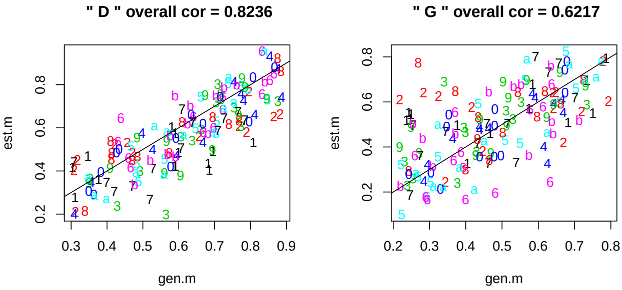


Figure: Recovery plots, range W and range D, G

Simulated Data Parameter Recovery IV

1	0000000000	0	0.1	0.015	0.035	0.16	0.105	0.12	0.04	0	0.0
2	0000000000	0	0	0	0	0	0	0	0	0	0
3	0000000001	0	0	0	0	0	0	0	0	0	1
4	0000000001	0	0	0	0	0	0	0	0	0	1
5	0000000011	0	0	0	0	0	0	0	0	1	1
6	0000000011	0	0	0	0	0	0	0	0	1	1
7	0000000111	0	0	0	0	0	0	0	1	1	1
8	0000000111	0	0	0	0	0	0	0	1	1	1
9	0000001111	0	0	0	0	0	0	1	1	1	1
10	0000001111	0	0	0	0	0	0	1	1	1	1
11	0000011111	0	0	0	0	0	1	1	1	1	1
12	0000011111	0	0	0	0	0	1	1	1	1	1

Table: Generating partitions with their per-cell marginal mean posterior. Possible evidence of the bias towards equal size cells.

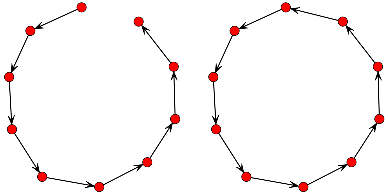
Survey design problem I

Logic could only be used (correctly) when the question is a pair that closes a cycle on a graph of pairs presented in the survey, *up to this question*.

Suppose it is harder for the informant to maintain logical consistency for questions that close larger cycles, based on *minimum closed cycle length*.

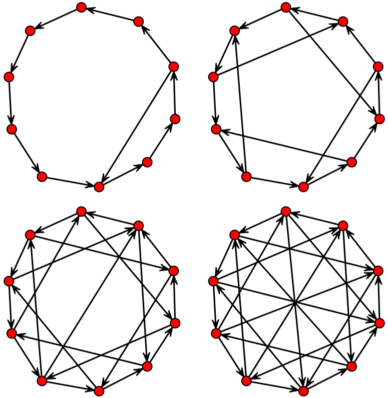
Survey design problem mitigation I

For Phase 1, present 9 + 1 directed pairs. The minimum cycle length is size 10 after 10 are presented.



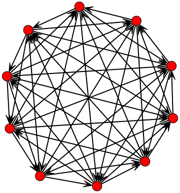
Survey design problem mitigation II

Phase 2 closes minimum cycles of length 4, “quads”. Present 15 of these quads to bring the total to 25.



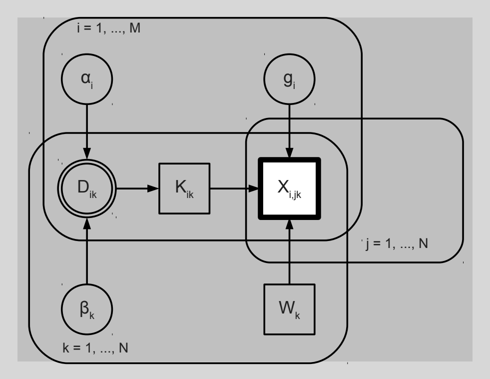
Survey design problem mitigation III

Phase 3 pairs complete the design, 20 more are added, each closing minimum cycles of length 3, “triads”.



Nodal Model Graphical Model

See the text for specific distributions. Circular nodes are continuous, square nodes are discrete. D_{ik} is double-circle is deterministic. $X_{i,jk}$ is bold, an observed datum (all other parameters are latent).



Survey Items I

Player	City	TruePartition
David Robinson	Tucson	0
Julius Erving	Flagstaff	0
Moses Malone	Kingman	0
Wilt Chamberlain	Scottsdale	0
Bill Russell	Prescott	0
Ernie Banks	Taos	1
Willie Mays	Las Cruces	1
Reggie Jackson	Los Alamos	1
Andre Dawson	Carlsbad	1
Mo Vaughn	Roswell	1

[1] William Batchelder and A. Romney.
Test theory without an answer key.
Psychometrika, 53(1):71–92, 1988.

[2] William Batchelder, Alex Strashny, and A. Romney.
Cultural consensus theory: Aggregating continuous responses in a finite interval.
In Sun-Ki Chai, John Salerno, and Patricia Mabry, editors, *Advances in Social Computing*, volume 6007 of *Lecture Notes in Computer Science*, pages 98–107. Springer Berlin / Heidelberg, 2010.
10.1007/978-3-642-12079-4_15.

[3] William H. Batchelder.
Cultural consensus theory: Aggregating expert judgments about ties in a social network.
In *Social Computing and Behavioral Modeling*, pages 1–9. Springer US, Boston, MA, 2009.

[4] William H. Batchelder, Ece Kumbasar, and John P. Boyd.

Consensus analysis of three-way social network data.
The Journal of Mathematical Sociology, 22(1):29, 1997.

[5] Carter T. Butts.
Network inference, error, and informant (in)accuracy: a bayesian approach.
Social Networks, 25(2):103–140, May 2003.

[6] Dorwin Cartwright and Frank Harary.
Structural balance: a generalization of heider's theory.
Psychological Review, 63(5):277–293, September 1956.

[7] A. Kimball Romney, William H. Batchelder, and Susan C. Weller.
Recent applications of cultural consensus theory.
American Behavioral Scientist, 31(2):163–177, 1987.

[8] A. Kimball Romney, Susan C. Weller, and William H. Batchelder.
Culture as consensus: A theory of culture and informant accuracy.

American Anthropologist, 88(2):313–338, June 1986.
ArticleType: primary_article / Full publication date: Jun., 1986
/ Copyright © 1986 American Anthropological Association.