May 8th, 12:30 PM - 2:00 PM

Exploring the Application of Reconstructability Analysis to Behavior Expression Data from a Social Network

Martin Zwick
Portland State University

Teresa D. Schmidt
Portland State University

Exploring the Application of Reconstructability Analysis to Behavior Expression Data from a Social Network

Teresa D. Schmidt
Portland State University

Professor Martin Zwick
Portland State University

Abstract. Reconstructability analysis (RA) is proposed as a complementary method for evaluating social network-related phenomena. Longitudinal records of social behavior expression among members of a social network are commonly represented as a set of social network analysis (SNA) connections, but might also be usefully represented as a set of associations derived through RA methods. Reconstructability Analysis identifies individuals as being associated when their behavior patterns appear coordinated–a representation that is unavailable with standard SNA. To explore the potential usefulness of RA for analyzing social behaviors, simulated behavior patterns were evaluated with both SNA and RA, and the results were compared. Several RA data formats were tested, as RA cases can be defined by (a) individual or (b) synchronous behavior expression, or by (c) pairwise or (d) group level interactions. Associations derived with each data format were compared with the connections captured in a routine SNA adjacency matrix. Highest agreement between the two methods was found when cases were defined as instances of behavior expression at the group level. In addition, RA was shown to be able to derive triadic and higher-order associations among individuals, as well as sets of individuals whose behavior patterns were positively *or* negatively associated. Thus, RA appears to offer several capabilities to the study of social network-related phenomena that are not available with standard SNA techniques. Reconstructability analysis holds promise for advancing research on social behaviors, and can likely complement many of the efforts that are currently being made with SNA.

1 Introduction

Social Network Analysis (SNA) is often applied to the study of behaviors, with the goal of being able to account for the spread of a behavior among individuals who are socially connected (e.g., Christakis & Fowler, 2007). This line of inquiry is stunted, however, by limitations in network analysis that prevent observational research from discerning contagion, or the 'true' spread of behavior through social influence (e.g., Burt, 1987), from other forces that can mimic behavior spread in a social network. Such forces are, for example, selection effects, where similar individuals tend to establish connections with one another (Kandel, 1978), synchronous maturation, and environmental factors (see Shalizi & Thomas, 2010), which can all give rise to patterns suggestive of phenomena 'spreading' though a network. Such 'spreading' has been observed even among phenomena unlikely to be socially transmitted, such as headaches, acne, and height (Cohen-Cole & Fletcher, 2008). Thus, several scholars argue that standard network analysis of observational data cannot provide sufficient evidence that the behavior is caused by social processes (Cohen-Cole & Fletcher, 2008; Lyons, 2011; Shalizi & Thomas, 2010) even when social connections are found to be a strong statistical predictor of behavior expression.

Reconstructability analysis (RA; see Klir, 1986; Zwick, 2004) is proposed as a methodology for addressing this concern. Reconstructability analysis is a data mining methodology that can be used to detect deviations from mutual independence among variables based on patterns in behavior. When studying human behaviors that are hypothesized to be social in nature, RA can identify deviations from mutual independence among a set of individuals based on patterns in their behaviors over time. This produces a network of pairs and larger groups of individuals whose behavior patterns suggest they are associated with one another. In effect, RA offers an approach that is opposite to SNA: Instead of predicting behavior based on relationships between individuals, RA predicts relationships between individuals based on behavior. The comparison of networks derived through both approaches can provide insights. If the level of agreement is significantly higher than would be expected by chance, the comparison would suggest that a behavior *is* subject to interpersonal processes, such as contagion or selection effects. By contrast, a chance level of agreement would suggest the behavior may be driven by non-interpersonal processes, such as environmental or individual factors. By comparing this RA-network of associations to a standard SNA-network, RA can complement SNA, supporting or refuting claims that social processes drive a given behavior.

1.1 Data Requirements

There are several requirements for data to be amenable to both SNA and RA. Standard SNA requires measurements of dyadic social connections among a set of network members, usually through observation or self-report. In studying the spread of a behavior through a network, the connections measured should be plausible as pathways along which social processes, such as contagion or selection, might occur. Frequent measurements of social connections may not be necessary, especially for networks where social connections are relatively stable.

Network modeling via RA requires discrete measurements of each network member's behavior over multiple time points. Each member is treated as a variable in RA and a record of each member's behavior at one time point constitutes a single case. As a data mining methodology, RA does best with a large number of cases, and with a relatively non-sparse dataset, meaning that individuals' behavior ought to vary at least somewhat over time. For best results, the time period between behavior measurements ought to reflect a time period over which social processes, such as contagion or selection, might occur.

1.2 Relations, Connections, and Associations

Reconstructability analysis and network analysis differ in their mathematical definitions of a relation, and it is due to these differences that RA might be used to complement SNA. In SNA, the presence or absence of a *connection* is measured through observation or report of the existence (and strength) of interaction, similarity, flow, or a perceived social relation between two individuals (Borgatti, Mehra, Brass,

& Labianca, 2009). By contrast, in Reconstructability Analysis, the presence or absence of an *association* is indicated by some degree of constraint in a pattern of behaviors.

Constraint, defined as a reduction in uncertainty (Shannon & Weaver, 1949), is independent of the direction of association. Two individuals can be associated if they have a tendency to express either the same or opposite behaviors. As shown in the sample contingency tables below, the ‘positive’ association between A and B is just as strong as the ‘negative’ association between C and D. In either case, knowing the behavior of one individual reduces uncertainty about the behavior of the other. Both of their behavior patterns demonstrate constraint, and could be identified in RA. For some behaviors however, positive associations might be much more anticipated than negative associations among individuals indicated by SNA to be connected. So in comparing sets of individuals who are connected (SNA) and associated (RA), knowing the direction of association can aid in the interpretation of results.

Positive Association

A \ **B**	**0**	**1**
0	9	1
1	1	9

Negative Association

C \ **D**	**0**	**1**
0	1	9
1	9	1

A ‘1’ in the row and column headings indicates some ‘marked’ behavior. Cell (0,0) indicates the number of times neither individual expressed this behavior, cell (1,1) indicates the number of times both individuals expressed the behavior, and cells (0,1) and (1,0) indicate the number of times one person expressed the behavior.

Table 1. Sample Contingency Tables

2 Method

Several small, illustrative data sets were simulated and analyzed to explore the potential for RA to complement standard SNA approaches to the study of behavior spread. These data sets were constructed in the form of data taken from postings to online discussion forums, as those are likely candidates for future analysis. Of primary interest in this preliminary work was establishing how RA might best identify a set of associations under an ideal circumstance, when a behavior is perfectly accounted for by social connections within a stable social network. With each artificial dataset, therefore, behavior patterns were fixed so individuals' online posting behaviors were perfectly predicted by their social connections. A social connection between two individuals, such as A→B, indicated that A's posting behavior was followed by a response post from B.

Simulated data was organized into several different formats for RA. With online forum data, as these data sets were meant to imitate, a case might be defined one of four ways, depending on whether behavior expressions are considered individual or collective, and aggregated or disaggregated. First, a case might be defined as a unique instance of behavior expression by one individual. In the case of online forum data, each posting could constitute its own case. Second, a case might collapse behavior expressions over a fixed period of time, such as a day or a week, indicating a set of individuals who behaved synchronously. Third, a case might be defined as a pairwise interaction, between the person submitting a post and the person targeted when the post is a response. And finally, a case might be defined as a group-level interaction, such as a discussion 'forum' or an online conversation, where all individuals might or might not have participated. These four case formats are summarized in Table 2.

	Disaggregated	**Aggregated**
Individual	Individual Behavior	Synchronous Behavior
Collective	Pairwise Interaction	Group-Level Interaction

Table 2. Data Formats for RA by Level of Aggregation and Collective Expression

Sample datasets in each format and were simulated and evaluated through the comparison of SNA and RA networks. All datasets were submitted to *Occam*, a discrete multivariate modeling tool based on the RA methodology (Willett & Zwick, 2004). All Occam analyses were searches for loopless neutral models, beginning with the independence model and working "upward" toward the fully saturated model. This procedure allowed RA to identify a model containing a set of associations with maximal constraint in each dataset, as selected by the Bayesian Information Criterion (BIC). The application of time series analysis was also explored with each format in turn, resulting in a total of eight options that were investigated for applying RA and SNA to behavior data.

3 Results

As a general rule, RA identifies associations among those variables which exhibit significant amounts of constraint. When applied to the study of behavior spread, RA identifies pairs and sets of individuals who exhibit anomalous patterns in behavior. Depending on the way data is organized into cases, RA might be used to identify patterns in (a) individual or (b) synchronous behavior expression, or by (c) pairwise or (d) group-level interaction. The following sections describe each of these data formats in more detail, along with preliminary findings regarding their potential for complementary RA and SNA.

3.1 Cases as Individual Behaviors

In order to apply RA to data defined by individual behavior expression, an SNA network dataset was first created where six individuals were socially connected in a chain structure: A→B→C→D→E→F. Reconstructability analysis was then applied to several simulated datasets where each instance of behavior expression was defined as its own case. Behavior expression of the focal individual was set as 1 and behavior of all other individuals in that case was set as 0. Datasets in the individual behavior case format were submitted to Occam as a neutral search with 3 models retained at each level for 7 levels.

In defining cases by individual behaviors, RA identifies those individuals whose behavior patterns were most anomalous. When all individuals post with equal frequency, all pairwise associations have equal strength: AB, AC, AD, BC, BD, etc. However, when individuals post with different frequencies, those who post more often will routinely be identified in associations with each other. This stems from the sparse nature of datasets in individual behavior format, which consist mostly of 0s. Associations are identified among individuals with higher than normal frequencies of 1s, as they create patterns that are least like the otherwise homogenous dataset of 0s. And because each case can only have one instance of '1', by the person posting, this case format produces only negative associations, where individuals' posting behaviors are always conversely related to the posting of any other individual.

Consider Example Dataset 1, where individual A submits 20 responses and all others submit 10 responses. Because all of the anomalous behavior patterns stem from negative associations, the strongest associations will contain person A, who has the highest frequency of posts opposite the other members. By itself, this data format is not suitable for comparing RA associations to SNA connections. The best model will identify associations among all individuals whose behavior patterns are anomalously frequent, regardless of whether those individuals are socially connected to one another.

A	B	C	D	E	F	Case Freq
1	0	0	0	0	0	20
0	**1**	0	0	0	0	10
0	0	**1**	0	0	0	10
0	0	0	**1**	0	0	10
0	0	0	0	**1**	0	10
0	0	0	0	0	**1**	10

Example Dataset 1. Cases as Individual Behaviors
Best Model: AB:AC:AD:AE:AF

3.2 Cases as Pairwise Interactions

In order to apply RA to data defined by pairwise interactions, an SNA network dataset was first created where six individuals were socially connected in isolated pairs: A–B, C–D, E–F. Reconstructability analysis was then applied to several simulated datasets where cases were defined as instances of interaction. Behavior expression of the individuals sending and receiving a response post was set as 1 and behavior of all other individuals in that case was set as 0. This retains relational information that is present in online discussion data, and allows pairs of individuals to exhibit positive associations by being concurrently involved (and uninvolved) in online interactions.

Datasets in individual case format were submitted to Occam as a neutral search with 3 models retained at each level. When restricted to advancing only 3 levels, RA is able to derive associations that precisely match the connections specified. In Example Dataset 2, RA is able to identify anomalous patterns that stem from the positive associations among individuals who are concurrently involved in online interactions. However, when allowed to advance to higher levels, RA identifies additional associations that are negative, such as AC, and more complex, such as the three-way association, ACE.

It is important to remember that RA does not have a preference for identifying positive over negative associations, and the negative associations between, say, A and C are nearly as strong as the positive associations. Due to this feature, it is possible for the most anomalous associations to be negative, and for RA to identify pairs and sets of individuals who appear to *avoid* concurrent involvement in online interactions. This is especially likely if individuals' online participation targets various partners at different times, as the most consistent behavior patterns will be among individuals who never respond to one another. Effectively, RA's employs a wider definition for relation than is available by many standard network approaches, and can reveal individuals as being associated even when they do not appear to interact.

A	**B**	**C**	**D**	**E**	**F**	**Case Freq**
1	**1**	0	0	0	0	20
0	0	**1**	**1**	0	0	20
0	0	0	0	**1**	**1**	20

Example Dataset 2. Cases as Pairwise Interactions
Best Model: AB:CD:EF

3.3 Cases as Group-Level Interactions

In order to apply RA to data defined by group level interaction, an SNA network dataset was created where six individuals were again connected in isolated pairs: A–B, C–D, E–F. Reconstructability analysis was then applied to several simulated datasets where cases were defined as instances of group-level interaction, such as the group of individuals who were involved in a given discussion forum, or topic. Behavior expression of those participating in a forum was set as 1 and behavior of all other individuals in that case was set as 0. Pairs consistently replied to one another within a given discussion forum, and sometimes alongside other pairs, even though they never responded directly to posts from individuals who were outside their pair.

Datasets in the group level case format were submitted to Occam as a neutral search with 3 models retained at each level for 7 levels. In defining cases by group-level behaviors, RA was able to take into a count a larger context for behavioral expression than was necessary in the pairwise case format. In Example Dataset 3, the best model identifies the three isolated pairs, despite pairs' frequent participation in (and abstinence from) the same forums as other pairs. There seems to be some amount of tolerance, then, for noise in the data, so long as pairs do not consistently participate in the same forums as other pairs.

A	**B**	**C**	**D**	**E**	**F**	**Case Freq**
1	**1**	**1**	**1**	0	0	10
0	0	**1**	**1**	**1**	**1**	10
1	**1**	0	0	**1**	**1**	10
1	**1**	**1**	**1**	**1**	**1**	10
0	0	0	0	0	0	10
1	**1**	0	0	0	0	10
0	0	**1**	**1**	0	0	10
0	0	0	0	**1**	**1**	10

Example Dataset 3. Cases as Group Level Interactions
Best Model: AB:CD:EF

3.4 Cases as Synchronous Behavior

In order to apply RA to data defined as synchronous behavior, an SNA network dataset was again created with six individuals in isolated pairs: A–B, C–D, E–F.

Reconstructability Analysis was then applied to several simulated datasets where cases were defined as a series of 24-hour time segments, during which each individual might or might not participate in any online forum. This format looks identical to the group level interaction format, in that behavior expression of those who participate during a given time segment was set as 1 and behavior of all other individuals in that case was set as 0. However this format contains no information regarding direct or indirect interactions among members during a given case, so members of an isolated pair may or may not express a behavior during the same case.

Datasets in the time segment case format were submitted to Occam as a neutral search with 3 models retained at each level for 7 levels. When partners in each pair were simulated to participate during the same time segments, RA returned results identical to those from the group level interaction format. However, when pairs' participation spanned across time segments, or when multiple pairs' participation was collapsed into too few cases, the associations identified by RA were less comparable to the SNA network. In Example Dataset 4, individuals A, B, C, and D all post (and abstain) during the same time segments, so it is not possible to differentiate pair A–B from C–D. This kind of finding is likely when behavior expression is more subject to matters of time, such as the day of the week, than to social connections.

A	**B**	**C**	**D**	**E**	**F**	**Case Freq**
1	**1**	**1**	**1**	0	0	20
0	0	0	0	**1**	**1**	20
1	**1**	**1**	**1**	**1**	**1**	20
0	0	0	0	0	0	20

Example Dataset 4. Cases as Time Segments
Best Model: AB:AC:AD:EF

3.5 Time Series Analysis

A time series component was added to RA data in each format. While RA typically ignores the order of cases, focusing only on case frequencies, additional variables can be added to a behavior dataset, such that a given variable reflects an individual at one point in time. In Example Dataset 5, for a cyclical network structure, A's behavior at

Time 1, or A_1, is shown to be associated with B's behavior at Time 2, or B_2. In time series analysis, an additional set of either lagged or leading variables can be added for each time point of interest. This allows the chronological order of cases to reveal temporal relations between individuals, indicating whose behavior follows whose.

A dataset with cases defined as individual behaviors, Example Dataset 5, was submitted to Occam for a time series analysis, retaining 3 models at every level for 7 levels. This produced a set of RA associations that matched the SNA network well, as did time series analysis of a dataset with cases defined by pairwise interactions. These results indicate that both the individual and pairwise data formats can benefit by using time series to retain chronological information in the data. However, consistency is needed in the number of time steps it takes a behavior to "move" between individuals. If person B's behavior is sometimes one case behind person A, and sometimes more, it will be more difficult for RA to identify those two individuals as being associated.

A_1	B_1	C_1	D_1	E_1	F_1	A_2	B_2	C_2	D_2	E_2	F_2	Case Freq
1	0	0	0	0	0	0	1	0	0	0	0	10
0	1	0	0	0	0	0	0	1	0	0	0	10
0	0	1	0	0	0	0	0	0	1	0	0	10
0	0	0	1	0	0	0	0	0	0	1	0	10
0	0	0	0	1	0	0	0	0	0	0	1	10
0	0	0	0	0	1	1	0	0	0	0	0	10

Example Dataset 5. Time Series with Cases as Individual Behaviors
Best Model: $A_1B_1{:}A_1B_2{:}B_1C_2{:}C_1D_2{:}D_1E_2{:}E_1F_2{:}F_1A_2$

Time series was not found to offer benefits for group-level and synchronous behavior formats, mostly because of the difficulty of interpreting results. When several individuals participated in each case, associations identified tended to be complex, involving positive and negative associations among a mix of individuals who were and were not connected in the SNA network.

4 Discussion

Reconstructability analysis appears to offer several features to the study of behavior spread that are not achievable with standard SNA techniques. First, RA associations may provide a broader definition of a relation than SNA. In the current manuscript, SNA connections were based on instances of concurrent behavior among pairs of individuals, while RA associations were also able to detect patterns of concurrent abstinence as well as patterns of opposite behaviors. Second, RA may take behavioral context into account more than SNA. Associated individuals were identified not only when their behaviors appeared coordinated, but also when their coordinated behavior patterns appeared independent of the behavior of others in the network. This was apparent when RA could still identify isolated pairs in the group-level interaction format, regardless of 'noise' stemming from higher-level patterns in behavior expression across pairs. Finally, RA is capable of identifying three-way and higher-way associations among individuals, which is not commonly possible with SNA.

Analysis of simulated datasets illustrates the conditions under which RA identifies associations, and how they might best be compared with SNA data. Of the four data formats examined, defining cases at the group level appeared to produce a network of RA associations that could be compared most directly with an SNA dataset. When adding a time series component to RA, individual and pairwise cases also produced associations that matched SNA data closely, but the effectiveness of RA with time series appears to be contingent on specific network circumstances, such as a consistent amount of time over which a behavior "moves" between network members. For this reason, the group level case format is anticipated to be most generally useful for the comparative application of RA and SNA.

The most limiting aspect of this work is that RA and SNA have not yet been jointly applied to real data. The preliminary simulated datasets here can only illustrate the potential value of the proposed method under an ideal case where social networks are stable and a behavior is perfectly predicted by an individual's social connections. Work is underway to apply RA and SNA to datasets from online discussion forums, so that the utility of various formats and techniques can be tested more practically.

5 Conclusion

Reconstructability analysis appears to hold promise for advancing research on social behaviors, as it offers several features to the study of behavior spread that are not available with standard SNA. By defining relations more broadly, and by taking higher-level patterns of behavior expression into account, RA may provide a complement to the standard SNA approach to studying behavior spread. Future work is needed to apply this method to genuine datasets and to continue to explore its strengths and limitations. However, indications at this point suggest the method may be able to advance fields of research that currently employ social network analyses, and amplify the usefulness of network analysis for studying social behaviors.

References

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 13,* 892-895. doi: 10.1126/science.1165821

Brownell, K. D., Greenwoord, M. R. C., Stellar, E., & Shrager, E. E. (1986). The effects of repeated cycles of weight loss and regain in rats. *Physiology & Behavior, 38*, 459-464.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. American Journal of Sociology, 92(6), 1287-1335.

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology, 113,* 702-734.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*, 1194-97.

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine, 357,* 370-379.

Cohen-Cole, E., & Fletcher, J. M. (2008). Detecting implalusible social network effects in acne, height, and headaches: longitudinal analysis. *British Medical Journal, 337,* a2533-37.

Kandel, D. B. (1978). Homophily, selection, and socialization in adolescent friendships. American Journal of Sociology, 84(2), 427-436.

Klir, G. (1986). Reconstructability analysis: An offspring of Ashby's constraint theory. Systems Research, 3(4), pp. 267-271.

Kramer, P., Westaway, S. K., Zwick, M., & Shervais, S. (2012). Reconstructability analysis of genetic loci associated with Alzheimer disease. IEEE 6th International Conference on Soft Computting & Intelligent Systems & 13th International Symposium on Advanced Intelligent Systems, SCIS-ISIS2012, Kobe, Japan, Nov. 20-24, 2012.

Lyons, R. (2011). The spread of evidence-poor medicine via flawed social network analysis. Statistics, Politics, and Policy, 2(1), 1-26.

Marsden, P. V. (2005). Recent developments in network measurement. In P. J. Carrington, J. Scott., & S. Wasserman (Eds.), *Models and methods in social network analysis* (8-30). New York, NY: Cambridge University Press.

Milligan, G. W. (1980). Factors that affect Type I and Type II error rates in the analysis of multidimensional contingency tables. *Psychological Bulletin, 87,* 238-244.

Shalizi, C. R. & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. Sociological methods and Research, 40, 211-239.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of communication. Urbana: University of Illinois Press.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics*(5th ed.). Boston: Pearson/Allyn and Bacon.

Wasserman, S. & Foust, K. (1994). *Social network analysis: Methods and applications.* New York: Cambridge University Press.

Willett, K. & Zwick, M. (2004). A software architecture for Reconstructability analysis. *Kybernetes, 33*(5/6), 997-1008.

Zwick, M. (2004). An overview of Reconstructability Analysis. *Kybernetes, 33*(5/6), 877-905.