A Generalized Model of Similarity for Relational Data*

Balázs Kovács
Stanford University
bkovacs@stanford.edu

September 2008

Abstract

A widespread approach to assess the similarity of objects is to measure the extent to which they have similar relationships to other objects or settings. For example, the notion of structural equivalence calls two persons similar if they have similar relationships to other persons. This paper generalizes this approach and views two objects similar if they have similar relationships to *similar* objects or settings. After proposing a geometrical representation for this generalized approach, we reanalyze two classic datasets: the Davis et al. (1941) data on social event attendance of 18 women, and the roll-call data of the U.S. Senate. We show that the proposed representation surpasses previous models of relational data, and illustrate how it opens up new possibilities for sociologists and social scientists in general.

1. Introduction

Relational data is widely used to the assess the similarity of objects and to group them to clusters. As an underlying principle, objects are held to be similar if they have similar relationships to other objects or settings. Sociologists group people based on whether they have similar relationships to other people (White et al., 1976), or based on whether they attend the same clubs (Breiger, 1974; Breiger et al., 1975; Borgatti and Everett, 1997; Doreian et al., 2004). Political scientists group senators based on their votes (MacDonald

*The generalized similarity algorithm developed in this paper can be freely downloaded from the author's website for Matlab. We are grateful for the help and detailed comments of Jerker Denrell, Balázs Gyenis, Michael Hannan, and Tomasz Sadzik. The paper also benefited from the discussions at the Nagymaros Group Conference in Antwerp, and the Macro Lunch at Stanford GSB. All remaining errors are our own.

and Rabinowitz, 1987), computational linguists measure the similarity of words by their co-occurrence in documents (Manning and Schütze, 1999; Schütze, 1998), computer scientists measure similarity of webpages by the frequency that they are linked to together (Dean and Henzinger, 1999).

In all the above cases, relations are viewed as indicators or proxies of underlying features and attributes of the objects or persons. When underlying attributes are hard to assess, researchers often turn to relational data to infer their similarity. For example, we do not know the leisure-time preferences of people, but we can observe the clubs they attend, and assume that they go to similar clubs because they have similar preferences. When people or objects are grouped based on their relations, it is implicitly assumed that there is a one-to-one correspondence between the underlying features and the resulting relationships[1].

This inference from relations to underlying features, however, is often problematic. Take, as illustration, the case of measuring the similarity of people by looking at whether they attend the same clubs, and take two imaginary persons, Mary and David. The more clubs they attend together, the more similar they are; the more clubs they attend without each other, the less similar they are. What happens, however, if Mary and David both love chess and the only clubs they attend are chess-clubs, but they live in different parts of the city so they attend two different chess clubs? In this case, although Mary and David are quite similar, the club overlap measure calls them dissimilar. In general, the problem with similarity inferences based on relational data could be that the absence of relations does not necessarily means dissimilarity, and the presence of ties does not necessarily mean similarity. Thus the measures using first-order relational data can be too coarse.

Network scholars, as a response to this shortcoming of first-order relational data, developed alternative ways to infer similarity. They call for incorporating indirect relations and abstract structural patterns. Burt (1988), for example, reviews and investigates alternative approaches to incorporate indirect ties into social network analysis. Relatedly, there had been a discussion on relational algebras (Breiger and Pattison, 1986; Pattison, 1988; Faust, 1988; Borgatti and Everett, 1992). In this literature, the above mentioned authors investigate ways to generalize structural equivalence, to other, more abstract equivalences. The reason for generalizing structural equivalence is to identify roles: people have the same role if they have the same kind of relationship to other people that are of the same role (Pattison, 1988). For example, a parents and children are two roles: parents form a class of people who have same relationships to the members of the other role, children.

Although the above approaches and methods provide some kind of solution to the problem of inferring similarity from relational data, they have several other shortcomings. First, there are a number of ways to incorporate indirect relations, and a number of ways to operationalize abstract equivalences. The problem is that many of these operationalizations are rather intuitive, and there had been no consensus which approach to use. This might

[1]This approach to relational data is similar to economists' use of revealed preference.

have been the reason that this line of research has kind of disappeared since the 1980s, and that most empirical researchers still use the basic, first-order relational measures.

This paper aims to revive the discussion on inferring similarity from relational data by proposing a novel approach. We do not want, however, to propose yet another measure for generalized similarity for relational data. We identify two principles for a generalized similarity measure, and show a model that corresponds to these principles.

First, we generalize "two objects are similar if they tend to have similar relationships to other objects or settings" to "two objects are similar if they tend to have similar relationships to similar objects or settings". To stick with the same club membership example: "people who attend the same clubs are similar" can be generalized to "people who attend similar clubs are similar." This generalized approach, as we demonstrate in this paper, provides a more fine-grained description of relational similarity then using first-order measures like structural equivalence or Pearson-correlation.

As a second principle, we emphasize the need for balance in relational similarity. That is, the similarity matrices have to be self-consistent and also consistent with other similarity matrices (in case of higher mode data). To follow our earlier example, not only "people who visit similar clubs are similar", but also "clubs that are visited by similar people are similar". Although this might seem as bootstrapping something out of nothing, we show that this is a valid representation of data, quite similar in spirit to Bonacich's measure of status: people are high status if linked to by other high status people (Bonacich, 1987).

This new representation provides a novel approach to relational similarity, and brings forth a number of fresh insights into the nature of relational data. Besides the two-mode relational data, it readily handles one-mode relational data, that are, for example social networks. In this case the "two persons are similar if they are linked to similar persons" balancing problem has to be solved. Also, it naturally gives access to higher mode relational data - and this is especially important as researcher previously had no clear representation to analyze higher-mode relational. An example for a three-mode relational data would be an article-scientist-academic institution data, in which case the following six relationships have to be balanced: "scientists are similar if they publish similar articles", "academic institutions are similar if they employ similar scientists", "academic institutions are similar if they produce similar articles" - and their symmetric relationships.

The structure of the paper is as follows. In Section 2, we study the need for a generalized representation of relational data. We outline the desiderata for such a representation, and describe a modified version of Pearson-correlation that meets these desiderata. Also, we compare the proposed model to related models in the literature: the CONCOR algorithm for clustering (Breiger et al., 1975), blockmodeling (White et al., 1976), and Latent Semantic Analysis (Landauer and Dumais, 1997). In Section 3, we reanalyze two classic relational datasets with the generalized similarity framework, and compare the results with the findings in the similarity and clustering literature. In Section 4, to further study and understand the proposed model, we observe its behavior on simulated data. Finally, we discuss the findings

and sketch some ideas for further research.

2. A Framework for Generalized Relational Data

Below we present a set of desiderata that such a generalized model of relational similarity should possess, and we present a representation that fits these desiderata.

Principle 1: Generalizing across context

We expect three natural properties of such a generalized measure of similarity for a pair of objects, A and B, A appearing in setting i: (1) if B appears in settings similar to i, sim(A&B) increases, (2) if B appears in settings neutral to i, sim(A&B) does not change, (3) if B appears in settings dissimilar to i, sim(A&B) decreases. Moreover, we expect the change in A and B to be proportional to the strength of similarity/dissimilarity of the settings.

The reason for the above principles becomes apparent if one thinks back to the example of Mary and David and the club membership. The baseline is that Mary and David are not members of any club, thus their similarity is zero. Principle 1 says that for all pairs of clubs David and Mary are members of, if the clubs are similar then the similarity of Mary and David should increase. This is the case, for example, if they are members of two chess clubs. Note that this principle includes the case in which Mary and David are members of the same club: in this case the clubs are obviously similar (and their similarity is highest), so the similarity of Mary and David will increase. Likewise, we expect that their similarity will not change if Mary and David are members of unrelated clubs, but will decrease if they are members of dissimilar or opposing clubs (like outdoor club vs. couch-potato club).

The same principles can be formulated for the relationships of settings (i.e., the clubs in our example): the similarity of two settings increases if objects that are part of them are similar, and decreases if the objects are dissimilar[2].

In two-mode relational data, the relation of one kind of objects to another set of objects is stored in a rectangular matrix. Examples are people-club membership, people-workplace, senator-issues, word-document matrices. To stay at a high level of generality, we will refer to this rectangular matrix as the object-setting matrix, with rows designating objects, and columns designating settings. Let $\underline{M}$ denote the object-setting matrix.

$$\underline{M} = \begin{array}{c} \\ o_1 \\ o_2 \\ \dots \\ o_m \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & \dots & s_n \end{array} \\ \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix} \end{array}$$

[2]Generalizing across settings in human inference making (Shepard, 1987)

Scholars are usually interested in the relationships among the objects, and often want to cluster the objects to groups of similars. Although there are numerous methods to do this, the underlying approach for these methods is to measure similarity between objects by assessing to what extent they appear together, or, in the case of valued data, to what extent they have similar values in the settings. A common way for this is to take the cosine distance or correlation between the row-vectors (Widdows, 2004). Less often, scholars are interested in classifying the settings (e.g., document classification), in which case they use the same cosine distance or correlation measures, but apply these measures to the columns of the object-setting matrix.

Below we present a model of generalized similarity that satisfies the above desiderata. Our starting model of similarity is the Pearson-correlation, and we will modify the Pearson-correlation to meet the outlined principles. The main problem with correlation is that it does not incorporate the similarities among the settings when comparing the objects, and, likewise, does not incorporate the similarities among the objects when comparing the settings.

There is, however, an easy way to incorporate this information to the correlation measure. If the dimensions along which the objects or settings are compared are not independent, one can just adjust the correlation measure for the non-independence of the dimensions. As a basic relationship in linear algebra states, the scalar product of vectors x and y in a base space of A is xAy. As the cosine distance is the normed scalar product of x and y, and the Pearson-correlation is a centralized version of the cosine distance, one readily gets the formula for a correlation measure in any arbitrary A base space:

$$correlation(x,y)[in\ base\ set\ \underline{A}] = \frac{(x-\overline{x})\ \underline{A}\ (y-\overline{y})^T}{\sqrt{(x-\overline{x})\ \underline{A}\ (x-\overline{x})^T}\sqrt{(y-\overline{y})\ \underline{A}\ (y-\overline{y})^T}} \quad (1)$$

Given this formula, the main idea of the generalized measure is to use the setting similarity matrix as a base space for calculating the object similarity matrix ("objects are similar if they appear in similar settings"), and to use the object similarity matrix as a base space for calculating the setting similarity matrix. Formally, if $\underline{M}$ denotes the original $m\ X\ n$ object-setting matrix (the input for the model), $\underline{O}$ denotes the mXm object-object similarity matrix, and $\underline{S}$ denotes the $n\ X\ n$ setting-setting similarity matrix, then the following equation has to hold for all pairwise combinations of objects:

Principle 2: Balancing the similarity matrices

relationship to balance theory Heider, 1946; Cartwright and Harary, 1956

$$O_{i,j} = \frac{(M_{i,} - \overline{M_{i,}})\ \underline{S}\ (M_{j,} - \overline{M_{j,}})^T}{\sqrt{(M_{i,} - \overline{M_{i,}})\ \underline{S}\ (M_{i,} - \overline{M_{i,}})^T}\sqrt{(M_{j,} - \overline{M_{j,}})\ \underline{S}\ (M_{j,} - \overline{M_{j,}})^T}}, \quad (2)$$

where $M_{i,}$ denotes the ith row, $M_{,j}$ the jth column of the $\underline{M}$ matrix, and $\overline{M_{j,}}$ denotes

the mean of the jth row.

Similarly, for all pairwise combination of settings, the following equation has to hold (the "settings are similar if they contain similar objects" principle):

$$S_{i,j} = \frac{(M_{,i} - \overline{M_{,i}})^T \, \underline{O} \, (M_{,j} - \overline{M_{,j}})}{\sqrt{(M_{,i} - \overline{M_{,i}})^T \, \underline{O} \, (M_{,i} - \overline{M_{,i}})} \sqrt{(M_{,j} - \overline{M_{,j}})^T \, \underline{O} \, (M_{,j} - \overline{M_{,j}})}} \tag{3}$$

Equations (2) and (3) define a system of equations with two independent variables, $\underline{O}$ and $\underline{S}$ [3]. Although there seems to be no analytical solution for the above equations, one can solve the system of equations iteratively. Start with $\underline{S}$ equal to the identity matrix. Plug this in to Eq. (2), which yields $\underline{O}^1$, the first iteration of the object-similarity matrix (note that this is equivalent to the similarity matrix from the Pearson-correlation). Then use this $\underline{O}^1$ in Eq. (3) to get $\underline{S}^1$, the first iteration for $\underline{S}$. Repeat doing this until the process converges, i.e. until $\underline{O}^{t+1} - \underline{O}^t < \epsilon$ (where ϵ is a pre-defined convergence threshold). Although we found no general proof for convergence, in all empirical settings we tried, the process converges quite fast, in 5-20 iterations.

Comparison with other similarity models and clustering algorithms

In this part of Section 2 we compare the proposed generalized model of similarity to a few other major methods in the literature. Specifically, we look at CONCOR (Breiger et al., 1975), blockmodeling, and Latent Semantic Analysis (Landauer and Dumais, 1997).

CONCOR

CONCOR, a hierarchical clustering algorithm was introduced in Breiger et al. (1975). We discuss this algorithm as it has some resemblance to the generalized similarity model, at least at first glance: CONCOR uses iterated correlations to cluster the relational matrix into blocks. The resemblance to the generalized similarity model, however, ends here. First of all, the similarity of the columns is not built into the similarity of the rows. Second, the optimal number of subgroups is hard to assess with CONCOR, while it is simply an output of the model. Third, the theoretical underpinning of the CONCOR model is not clarified (on this later issue, see Schwartz, 1977). In a constructive manner, we could say that the generalized similarity model provides a theoretical motivation for taking iterated correlations, and shows how the row correlations and column correlations can be put together into a unified model.

[3] To be more precise, Equations (2) and (3) define $m^2 + n^2$ equations with $m^2 + n^2$ variables, for each cell in the $\underline{O}$ and $\underline{S}$ matrices. Excluding the equations for the diagonals (as the diagonal values are always one) and half of the off-diagonal cells (because of symmetry), there are $\frac{(m-1)^2+(n-1)^2}{2}$ equations.

Blockmodeling

Blockmodeling was developed in the 1970s to partition the nodes of networks (i.e., one-mode relational data) in to clusters based on the positions of the nodes (White et al., 1976). The rationale of this partitioning is structural equivalence: those in a partition (block) are similar in their relations to other nodes and, therefore, to other blocks that include those nodes. Blockmodeling is basically an inductive technique that involves shuffling the rows and columns to get at homogenous blocks (or kind of homogenous blocks). Doreian et al. (2004) extended the blockmodeling approach to two-mode relational data.

The generalized similarity measure proposed in this paper has the same goal as blockmodeling: finding the equivalence types, i.e. those nodes that have similar relationships to other nodes. The generalized similarity model, however, generalizes the notion of positional equivalence, and calls nodes equivalent if they have similar relationships to similar nodes. This extension, we argue, is essential for two reasons: to get a more precise measure of similarity by incorporating across settings/objects information; and to get a better measure of similarity and dissimilarity in sparse matrices. In large and sparse matrices[4], blockmodeling underestimates the similarity of many pairs of object, therefore blockmodeling would identify lots of small local blocks, while the real underlying data is much more structured.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a high-dimensional linear associative model, which is used to analyze text corpora to induce the similarity of words (Landauer and Dumais, 1997). The input of the LSA is usually text, more specifically, word-document matrices, which are essentially two-mode relational datasets. The LSA uses local co-occurence data to induce global similarities of words. In short, LSA is an application of a well-know linear algebra procedure, Singular Value Decomposition (SVD), to linguistic data. SVD breaks down the original m X n word-document matrix into three matrices: to a U matrix that contains the pairwise similarities of words, to a S matrix that contains the eigenvalues of the word-document matrix, and to a V matrix that contains the pairwise similarities of the documents.

Although not often used for analyzing relational data in sociology, in our opinion LSA (or rather, SVD) could be one of the best technique to analyze relational data, as SVD incorporates inter-setting information in calculating the similarity of objects, and also uses the inter-object similarity information to calculate the similarity of documents. There are two major points on which our generalized similarity model differs from LSA. First, the generalized similarity model provides an axiomatization for the measure and builds up the model from these axioms, while LSA is an application of SVD (without much argument why SVD is a valid way of representation). Second, LSA being a dimension reduction technique,

[4]It is worth noting that most of the applications of blockmodeling analyze small networks, in which the sparsity problem does not arise.

it contains an element, the choice of the number of dimensions, which is kind of arbitrary, and is usually chosen so as to maximize the fit of the data.

Iterating the adjacency matrix

old Breiger paper Breiger and Pattison (1986) Snijders matrix exponential Snijders (2001) MDS is already a measure that takes across settings into account

3. Applications

We illustrate the workings of the generalized model of similarity in two empirical settings: the classic Davis et al. (1941)'s dataset on the club membership of 18 women, and the voting record of the U.S. Senate. We chose to analyze these settings as they are ones of the most extensively analyzed settings social networks analysis and in political science. In both settings, we show how the similarity solution differs if we use the original Pearson-correlation or the generalized similarity measure.

Davis et al. (1941)'s data on Southern women's social event participation

Our first illustration uses Davis et al. (1941)'s data on 18 women's participation in 14 social events. The original data is shown in Figure 1 (as sorted by Doreian et al., 2004).

Actor	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}
Evelyn	1	1	1	1	1	1	0	1	1	0	0	0	0	0
Laura	1	1	1	0	1	1	1	1	0	0	0	0	0	0
Theresa	0	1	1	1	1	1	1	1	1	0	0	0	0	0
Brenda	1	0	1	1	1	1	1	1	0	0	0	0	0	0
Charlotte	0	0	1	1	1	0	1	0	0	0	0	0	0	0
Frances	0	0	1	0	1	1	0	1	0	0	0	0	0	0
Eleanor	0	0	0	0	1	1	1	1	0	0	0	0	0	0
Ruth	0	0	0	0	1	0	1	1	1	0	0	0	0	0
Verne	0	0	0	0	0	0	1	1	1	0	0	1	0	0
Myra	0	0	0	0	0	0	0	1	1	1	0	1	0	0
Katherine	0	0	0	0	0	0	0	1	1	1	0	1	1	1
Sylvia	0	0	0	0	0	0	1	1	1	1	0	1	1	1
Nora	0	0	0	0	0	1	1	0	1	1	1	1	1	1
Helen	0	0	0	0	0	0	1	1	0	1	1	1	0	0
Olivia	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Flora	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Pearl	0	0	0	0	0	1	0	1	1	0	0	0	0	0
Dorothy	0	0	0	0	0	0	0	1	1	0	0	0	0	0

Figure 1: The original Davis et al. (1941) data on the social event participation of 18 Southern women, with Doreian et al. (2004)'s blockmodel solution.

Freeman (2003) provide an exhaustive literature review of 21 articles analyzing the Southern women data, and arrive at the conclusion that the underlying structure of the data is in which there are two subgroups of women. One is composed of Evelyn, Laura, Theresa, Brenda, Charlotte, Frances, Eleanor, Pearl, Ruth; the other has Verne, Myra, Katherine, Sylvia, Nora, Helen, Dorothy, Olivia, Flora as its members. Freeman (2003) does not analyze the corresponding partition of events. Doreian et al. (2004), in their article introducing block-modeling for two-mode network data, reanalyze the Southern women data, and arrive at the conclusion that there are actually three subgroups of women, Pearl and Dorothy constituting a third subgroup. Simultaneously, they provide a partitioning for the social events: they find that there are three main subgroups of events: (1,2,3,4,5),(6,7,8,9),(10,11,12,13,14). Their partitioning is shown on Figure 1.

Here we analyze the South women social event participation data with the generalized similarity model. Figure 2 shows the two-dimensional Multidimensional Scaling maps based on both the correlation similarity measure and the generalized similarity measure[5]. The MDS map based on correlation "kind of" recovers the blockmodel results: it finds three subgroups. However, two women, Olivia and Flora are put together with Pearl and Dorothy, which is clearly a mistake. Moreover, the three clusters are not clearly distinct.

As Figure 2 shows, the generalized similarity measure perfectly recovers the blockmodel solution of Doreian et al. (2004). Also, in the generalized solution the differences between the groups is strengthened, and the groups are clearly distinct. The generalized similarity model provides a grouping for the events as well. This grouping, not shown here, slightly differs from Doreian et al. (2004)'s grouping: although the (1,2,3,4,5) and (10,11,12,13,14) clusters emerge in the generalized similarity solution as well, the picture differs for events 6,7,8, and 9. Event 6 here is clustered together with the (1,2,3,4,5), while events 6, 7, and 8 form do not fall into any group but stand separately.

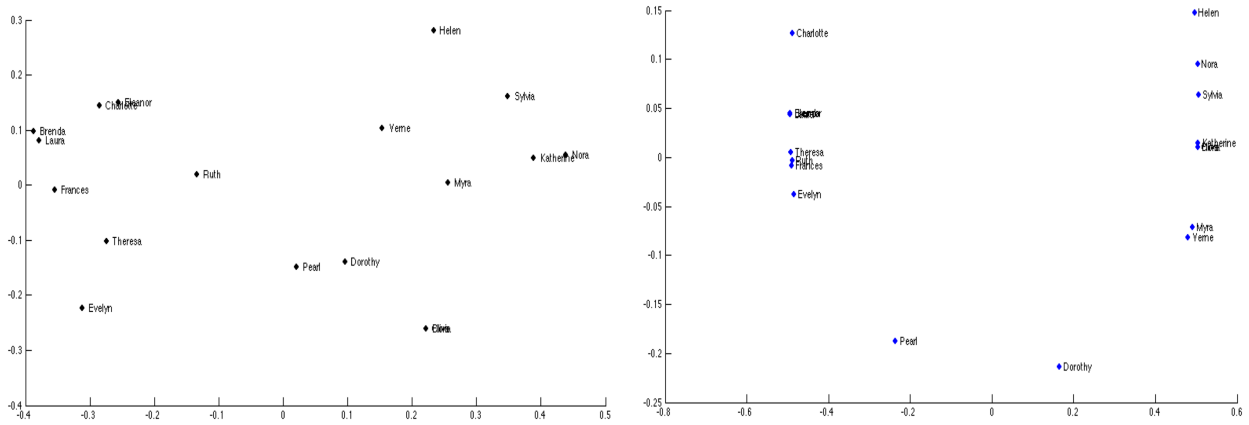
Senate roll-call data - the 109th Senate

Roll call data is one of the most analyzed kind of datasets of political scientists (e.g., Clinton et al., 2004). First we analyze the voting record of the 109th U.S. Senate (that which was in office 2005-2006). The 109th U.S. Senate had 101 members (as Robert Menendez filled the seat of Jon Corzine in 2006, when the latter became the Governor of New Jersey), and there were 644 non-unequivocal votes[6]. That is, the $\underline{M}$ matrix, which contains the data, is a $101\,X\,644$ matrix. The "Yay" vote is coded with 1, the "Nay" with -1, and the "No vote" with 0.

First, we analyze the similarity of the senators with Pearson-correlation. That is, the similarity of senators i and j is equaled to the correlation of the their voting vectors, M_i, and

[5]For the MDS procedure, the similarity values had to be transformed to dissimilarity values. The rule of transformation was $dissimilarity = (1 - similarity)/2$.

[6]The data on the votes and senators was downloaded from the U. S. Senate's website, www.senate.gov.



(a) MDS based on correlation

(b) MDS based on generalized similarity

Figure 2: Comparison of the two-dimensional MDS maps of Pearson-correlation and generalized correlation for the Davis et al. (1941) data on the club membership of 18 women.

$M_{j,}$. As there are 101 senators in our dataset, the senator-senator similarity matrix contains $101\ X\ 101\ =\ 10,201$ cells. This similarity matrix is symmetric, with 1s in the diagonals. Figure 3 shows the distribution of the pairwise similarity values, and the two-dimensional Multidimensional Scaling map based on these similarity values. The bimodal distribution reflects the bipartisan nature of the Senate, but indicates some overlap between the parties. The two-dimensional MDS map visualizes the pairwise similarities. As can be seen, the voting records identify two distinct clusters, and these clusters perfectly identify the two parties in the Senate. There is, however, a relatively large heterogeneity within the clusters, especially among the members of the Democratic Party.

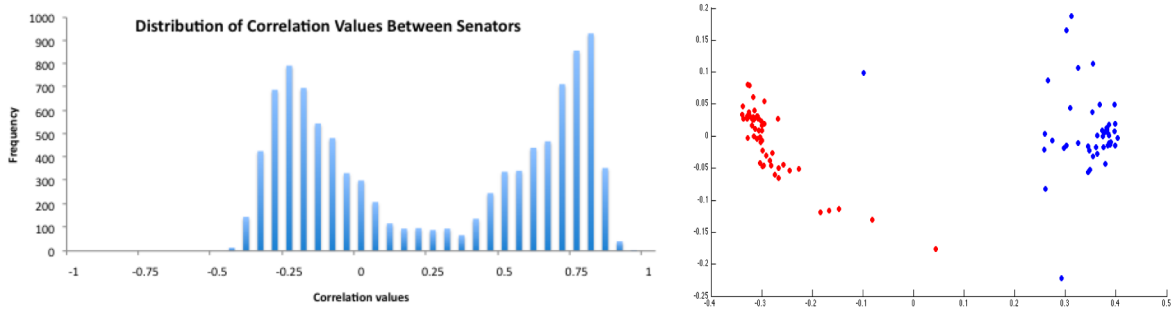


Figure 3: The distribution of the pairwise similarity measures of the senators, and the two-dimensional MDS map based on these similarity values. Calculated from the 109th U.S. Senate roll-call data.

How do the results of the generalized similarity model differ from the results based on

Pearson-correlation? Figure 4 shows the distribution of the generalized similarity values. The generalized similarity values show that the partisanship of the senate is much stronger than indicated by the Pearson-correlation. Indeed, the generalized similarity measure identifies a perfectly bipolar Senate.

The generalized similarity representation also provides a similarity clustering for issues. As Figure 4 shows, the issues are bipolar in nature as well, although less perfectly than the senators. This is consistent with earlier findings in political science showing that the issue-space in the Senate is bipolar, and constrained in the sense that position on a given issue strongly correlate with positions on other issues (Poole, 2007). The bipartisan nature of issues underline the necessity of taking inter-issue relationships into account.

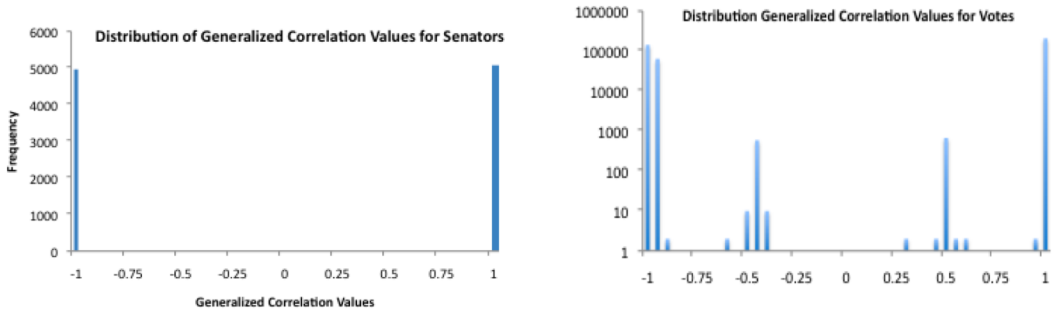


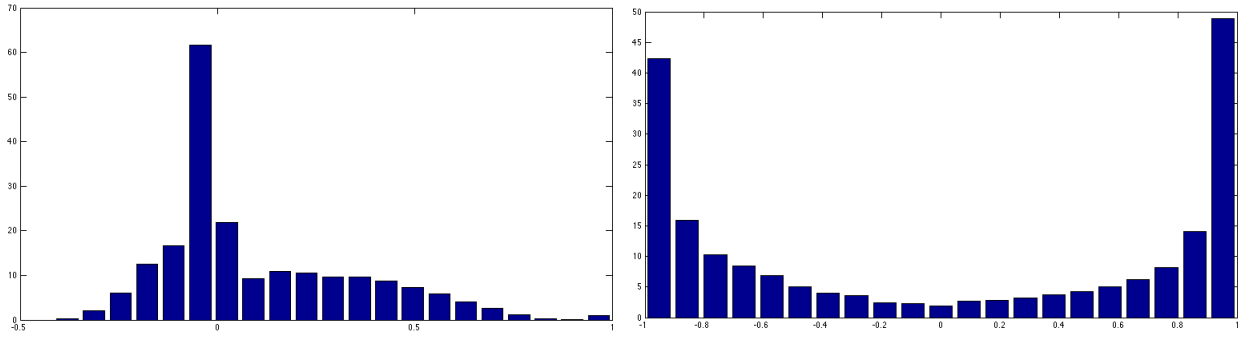
Figure 4: The distribution of the senator-senator similarity values and the issue-issue similarity values, based on the result of the generalized similarity model.

Comparing Senators who never voted together

As pointed out in the Introduction, a major advantage of such a generalized measure is its efficiency to deal with data sparsity. The previous two datasets analyzed were relatively dense, in the sense that there was not much missing data. Now we expand the time-frame of the roll-call analysis, and we analyze the voting data of ten consecutive Senates: the 101th-110th Senate, serving from 1989-2008. These Senates had/have 202 senators altogether, who voted on 6,510 issues. As on any given issue no more than 100 Senators can vote, the resulting matrix is clearly sparse (51% of the cells in the vote matrix is missing). 28.4% of the senator pairs pairs never voted together, so the measures using first-order relations cannot say anything about their similarity.

To compare the Pearson-correlation and the generalized similarity solution, we coded the missing data as "No vote", that is, with 0. Figure 5 shows the distribution of the correlation and the generalized similarity values for the senator pairs. This figure clearly shows the advantage of the generalized similarity model in settings with sparse data: while

the Pearson-correlation cannot capture the structure of the Senate, the generalized similarity can.



(a) Distribution of the correlation values (b) Distribution of the generalized similarity values

Figure 5: Comparison of the distribution of Pearson-correlation and generalized similarity for the 202 Senators serving in the 101th-110th U.S. Senate

4. Simulations

The previous empirical analyses indicate that the generalized similarity model tends to emphasize the similarity within, and the dissimilarity between the objects and between the settings, thereby leading to a crisper similarity map. Also, there is some evidence that in performs much better in inducing pairwise similarity for objects that do not appear together. In the next section we turn to simulations to further investigate these properties. The approach is as follows: we create a model of data, based on which we stochastically generate datasets, and run the different similarity algorithms to see how well they recover the underlying similarity structure. In comparing the clustering and similarity methods, we focus on two issues: how robust the methods are to local disturbances in the data, and how well the methods deal with the sparsity of the data.

Robustness of classification in the Senate setting

As we have seen in the previous, empirical section, the generalized similarity method classifies the senators into two clearly distinct and uniform subsets: Democrats and Republicans. Even if there were within party variance in given votes, the model incorporated across vote patterns and found that there are no systematic differences between party members, only across the parties. Similarly, for the Southern women data, the generalized similarity model found three distinct groups. These findings indicate that the method is robust for small, local

variations, and can pick up the real underlying data even if the local variances are relatively high. In the Senator example, this translate to saying that the individual Democrats might deviate from the other party members in their vote here and there, but overall they tend to vote with their party. In this sense, deviations are local idiosyncrasies, and not systematical differences - and the generalized model of similarity is very efficient in filtering out these idiosyncrasies by pooling across vote data[7].

To verify the above argument, we build a simulation that extends the Senator example. In the simulation, we assume that there are two parties, Red and Blue, each with 20 senators[8]. Each senator votes on 100 issues. We assume a perfect bipolar system: on issues on which the Red senators vote "Yay", the Blue-s vote "Nay", and vice versa. The "Yay" vote is coded with 1, the "Nay" with -1.

If all senators vote perfectly with their parties, then all similarity measures perfectly identify the polarization of senators and issues. What happens, however, if the senators sometimes diverge from their party consensus? If this divergence is non-systematic, we can model the divergence as error. That is, there is a certain p probability that a given senator will not vote with its party. How well the two similarity and classification measures, Pearson-correlation and our generalized similarity compare in identifying the two groups of senators and issues in such a "noisy" environment? Figure 5 shows the results. It shows that the generalized similarity measure is superior to the simple correlation measure in identifying the two parties. The efficiency of the generalized similarity measure is surprisingly high: even if the error term is 0.35, that is, even when on every third issue the senators deviate from their party's vote, the generalized similarity measure, by taking the cross issue information into account, can perfectly identify the parties. Also, the algorithm is similarly efficient in classifying the issues (results not shown here).

Next we investigate how the generalized similarity method performs if there is an underlying heterogeneity within the parties. That is, there are Red leaning Blues and Blue leaning Reds. As seen from the empirical results of Section 3, we expect that the generalized similarity measure attenuates these differences. To investigate this, we incorporate systematic differences to the previous senator simulation model. Each of the senators gets a random value between 0.7 and 1 that denotes the proportion of votes on which the senator follows the party's vote. We are interested in how well the Pearson-correlation and the generalized similarity recover this structure of the data. Simulations show that the generalized similarity measure is less good in picking up the internal differences. This is consistent with the empirical findings of Section 3.

We tried to compare the similarity algorithms under more complicated underlying models, in which there are more "real" groups, more options along the dimensions, etc., but the results

[7]This approach is rather similar to a signal detection and filtering approach.

[8]The specific parameters of the simulation model are not important: we tried a various number of other parameter settings, and the results remained very similar.

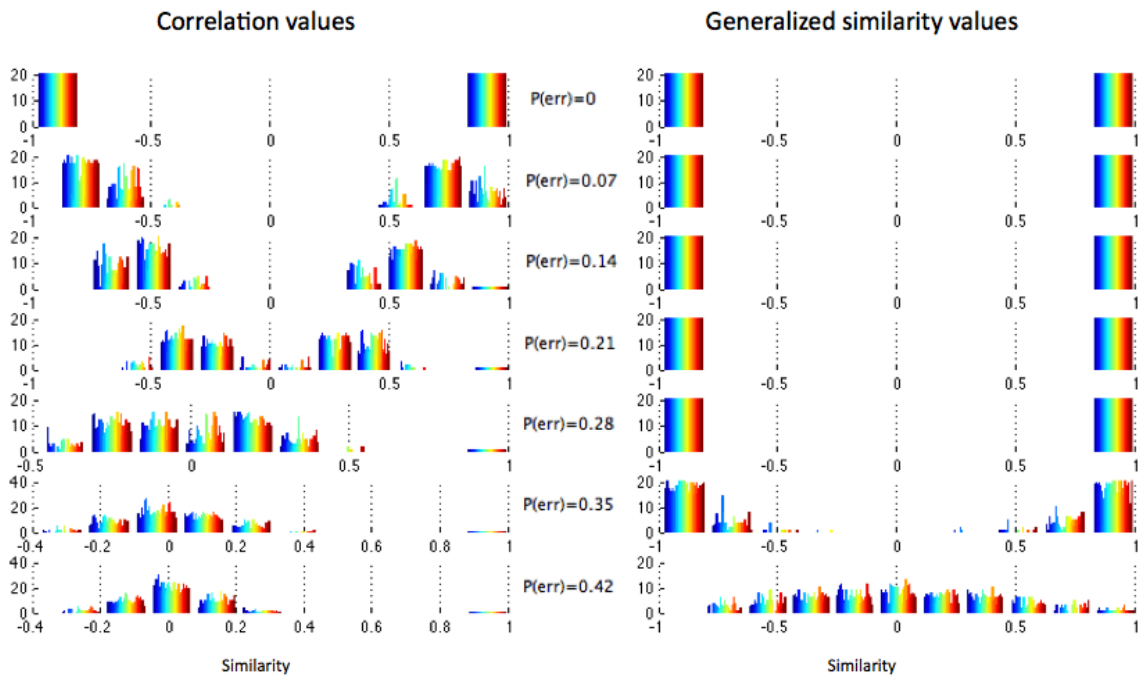


Figure 6: Comparison of the distribution of similarity values for Pearson-correlation and generalized similarity, in the simulated Senator-vote setting, with the introduction of error

remained practically the same.

As pointed out earlier, in many cases the relational data are sparse. This might be because the number of other objects an object can appear together is limited. Such is the case for example with words appearing in sentences (there are only so many words that can appear in a sentence), people and club membership, friendship ties. We expect the generalized similarity approach to be especially efficient in settings like this, as it can incorporate the across settings information, thereby guessing/imputing missing data.

To investigate the behavior of the generalized similarity model with sparse data, we use the same senator voting model, but now we add a probability that the senator will not vote at all (coded with 0). As the results on Figure 6 illustrate, the generalized similarity copes surprisingly well with missing data: even when the 70% of the votes are missing, it can identify the two underlying groups. Moreover, the findings are robust for the introduction of error: the two groups are identified when the noise is at 30%. The advantages of the generalized measure as compared to the Pearson-correlation are even more apparent when the data is sparse: correlation fails to identify the groups even if only a small random disturbance is introduced.

Recovering the true social network

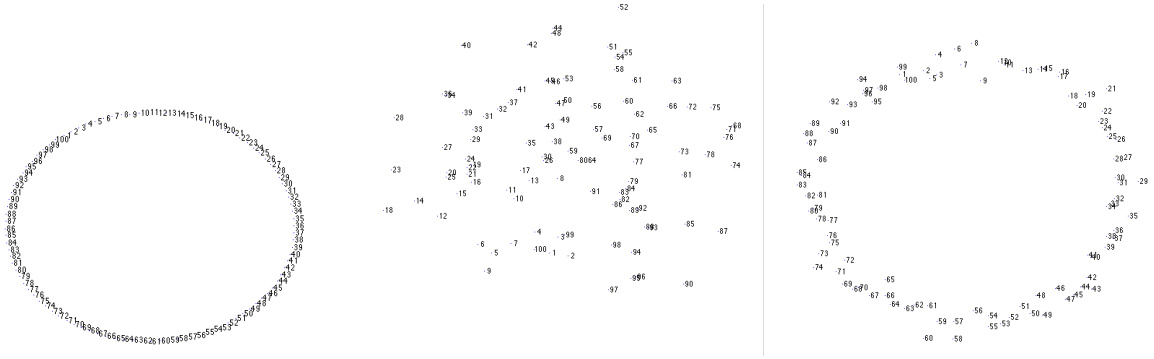
We now turn to analyze how the generalized similarity framework operates on one-mode relational data, specifically, in a social network setting. We specify a given distribution of attributes of the nodes, generate random networks based on these attributes, and see how the generalized similarity model solution compares to the correlational solution. Again, the goal here is to illustrate that the generalized similarity measure can recover the real, underlying distribution of data even in stochastic and sparse settings in which the first-order co-appearance measures fail to do so.

The specific model is as follows. We assume that there are 100 individual, numbered from 1 to 100. This number represents the individual's attribute along a dimension, and we assume that the individuals are ordered along this dimension such that the ends of the distribution meet, and the closer two numbers are, the more similar the individuals are. The similarity map of these individuals is thus a circle (shown on Figure 7(a)). We simulate random networks in which the tie creation rule is homophily: the more similar the individuals are, the more likely there will be a tie between them. This tie creation rule is consistent with the approach "two individuals are similar if they are connected to similar individuals", thus the generalized similarity model is expected to provide a better description of the underlying data than Pearson-correlation.

Here we compare the differences between the generalized model of similarity and Pearson correlation in two settings: one in which the individuals have a relatively large number of ties, and another one in which the individuals only have a few ties. In the first setting, we set the probability of a tie between any two individual to be $P_{i,j} = \frac{1}{3*exp(10*|distance(i,j)/100|)}$. In this setup, the probability that an individual will be connected to it's closest neighbor is 30%, to it's second closest neighbor is 27% etc. In the networks generated, the individuals have, on average, ties to 10 other individuals. In this setup both the Pearson-correlation and the general similarity measure are quite efficient in recovering the original data (for brevity, we do not show the results here). We can see that even here, the general similarity measure is slightly better.

The superiority of the generalized similarity model shows more strongly if we generate networks with fewer ties. For example if we reduce the probability of ties with 67%, to $P_{i,j} = \frac{1}{5*exp(10*|distance(i,j)/100|)}$, then the individuals will have on average 7 ties. In this case, the Pearson-correlation measure cannot recover the original structure of data, but the generalized similarity measure can (see Figures 7(b) and 7(c)). The reason for this is the following: the Pearson correlation measure only takes into account the first-order relational data, that is, whether the two individuals are linked to the same individuals or not. When the links are rare, most of the individuals will be relatively similar to each other, as there is a high overlap in people to whom they do not link to. There will be only small differences for individuals they do link to. In other words, the Pearson similarity measure is too crude because it lumps together a most of the dissimilar individuals (absence of ties), but cannot

differentiate between them based on how different they are.



(a) Original data (b) Correlation solution (c) Generalized similarity solution

Figure 7: Comparison of the two-dimensional MDS solutions based on **a** real data, **b** Pearson-correlation, and **c** generalized similarity solution

5. Discussion and Further Work

This paper proposed a generalized model of similarity to analyze the similarity relationships of relational data. By extending the Pearson-correlation, we get a model of data that simultaneously provides a similarity matrix for objects and settings. This model is especially efficient in analyzing sparse relational data, and it gives a more robust classification than the standard Pearson-correlation based similarity measures.

The generalized similarity model might also help in handling another problem of first-order relational data. In settings in which objects serve as substitutes, the principle that the more similar two objects are, the more likely they appear together, does not work. For example, the words "America" and "U.S." rarely appear in the same sentence (Widdows, 2004), and customers rarely buy two different recordings of the same Beethoven concerto. The proposed generalized approach solves this problem: as "America" and "U.S." tends to appear in similar sentences, their similarity will be quite high.

The presented model is, of course, not without limitations. First, if possible, an analytical solution of the model would be needed. Second, the exact model is just one of the possible frameworks for generalized similarity, other methods exists. We proposed an approach that can possibly be combined it with other methods.

References

Bonacich, Philip. 1987. "Power and centrality: A family of measures." *American Journal of Sociology* 92:1170–82.

Borgatti, Stephen P. and Martin G. Everett. 1992. "Notions of Position in Social Network Analysis." *Sociological Methodology* 22:1–35.

Borgatti, Stephen P. and Martin G. Everett. 1997. "Network analysis of 2-mode data." *Social Networks* 19:243–269.

Breiger, Ronald L. 1974. "The duality of persons and groups." *Social Forces* 53:181–190.

Breiger, Ronald L., Scott A. Boorman, and Phipps Arabie. 1975. "An Algorithm for Clustering Relational Data with Applications to Social Network Analysis and Comparison with Multidimensional Scaling." *Journal of Mathematical Psychology* 12:328–383.

Breiger, Ronald L. and Philippa E. Pattison. 1986. "Cumulates Social Roles $The Duality of Persons and Their Algebras." Social Networks 8 : 215 - -256$.

Burt, Ronald S. 1988. "Some properties of structural equivalence measures derived from sociometric choice data." *Social Networks* 10:1–28.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98:355–370.

Davis, Allison, Burleigh B. Gardner, and Mary R. Gardner. 1941. "Deep South: A Social Anthropological Study of Caste and Class."

Dean, Jeffrey and Monika R. Henzinger. 1999. "Finding related pages in the World Wide Web." *Computer Networks* 31:1467–1479.

Doreian, Patrick, Vladimir Batagelj, and Anuska Ferligoj. 2004. "Generalized blockmodeling of two-mode network data." *Social Networks* 26:29–53.

Faust, Katherine. 1988. "Comparison of Methods for Positional Analysis: Structural and General Equivalences." *Social Networks* 10:313–341.

Freeman, Linton C. 2003. "Finding Social Groups: A Meta-Analysis of the Southern Women Data." In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, edited by Ronald Breiger, Kathleen Carley, and Philippa Pattison. National Academies Press.

Landauer, Thomas and Susan Dumais. 1997. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review* 104:211–40.

MacDonald, Stuart Elaine and George Rabinowitz. 1987. "The Dynamics of Structural Realignment." *The American Political Science Review* 81:775–796.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Pattison, Philippa E. 1988. "Network Models: Some Comments on Papers in this Special Issue." *Social Networks* 10:383–411.

Poole, Keith T. 2007. "Changing minds? Not in Congress!" *Public Choice* 131:435–451.

Schütze, Hinrich. 1998. "Automatic Word Sense Discrimination." *Computational Linguistics* 24:97–123.

Schwartz, Joseph E. 1977. "An Examination of CONCOR and Related Methods for Blocking Sociometric Data." *Sociological Methodology* 8:255–282.

Shepard, Roger N. 1987. "Toward a universal law of generalization for psychological science." *Science* 237:1317–23.

Snijders, Tom A. B. 2001. "The Statistical Evaluation of Social Network Dynamics." *Sociological Methodology* pp. 361–395.

White, Harrison C., Scott A. Boorman, and Ronald L. Breiger. 1976. "Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions." *American Journal of Sociology* 81:730–780.

Widdows, Dominic. 2004. *Geometry and Meaning*. Center for the Study of Language and Information/SRI.