

# A generalized model of relational similarity

Balázs Kovács

University of Lugano, Institute of Management, Via Buffi 13., Lugano 6900, Switzerland

## ARTICLE INFO

### Keywords:

Relational similarity  
Geometrical representations  
Correlation  
Blockmodeling

## ABSTRACT

This paper introduces two principles for relational similarity, and based on these principles it proposes a novel geometric representation for similarity. The first principle generalizes earlier measures of similarity such as Pearson-correlation and structural equivalence: while correlation and structural equivalence measure similarity by the extent to which the actors have similar relationships to other actors or objects, the proposed model views two actors similar if they have similar relationships to *similar* actors or objects. The second principle emphasizes consistency among similarities: not only are actors similar if they have similar relationships to similar objects, but at the same time objects are similar if similar actors relate to them similarly. We examine the behavior of the proposed similarity model through simulations, and re-analyze two classic datasets: the Davis et al. (1941) data on club membership and the roll-call data of the U.S. Senate. We find that the generalized model of similarity is especially useful if (1) the dimensions of comparison are not independent, or (2) the data are sparse, or (3) the boundaries between clusters are not clear.

© 2010 Elsevier B.V. All rights reserved.

## 1. Introduction

The notion of similarity presents itself in most walks of life. As humans we constantly make similarity judgments: whenever we encounter a new situation, we apply the knowledge we have gained in similar situations (Hume, 2004 (1748); Shepard, 1987). Whenever we evoke a category or concept, say, “apple”, we implicitly refer to a set of objects that are similar to each other (Murphy, 2002). Similarity and categorization, through their influence on cognitive structures, shape the life of societies and organizations. People are put together into classes, races, age groups or nations. Organizations are classified based on their similarity to industries and populations. How and why things are deemed similar and are classified thus has deep-rooted consequences to how the (social) world works.

When data on the attributes of actors are not available or the weighting of the attributes is unobvious, researchers may turn to relational analysis to assess similarity. The underlying principle of relational similarity is that actors<sup>1</sup> are considered to be similar if they have similar relationships to other actors or objects. For example, 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; Doreian et al., 2004). Or, senators who tend to vote similarly are similar (Clinton et al., 2004).

This article is an endeavor to rethink the concept of relational similarity. We propose two principles we believe similarity representations should satisfy. First, we generalize the idea that “two actors are similar if they have similar relationships to other actors or objects” to “two actors are similar if they have similar relationships to *similar* actors or objects.” For example, the logic that “people who attend the same clubs are similar” can be generalized to “people who attend similar clubs are similar.” This generalized approach, as we further demonstrate in the paper, incorporates more information on similarity than first-order measures such as structural equivalence (Lorrain and White, 1971; Burt, 1976) or Pearson-correlation.

The second principle emphasizes consistency in similarity. That is, the similarity matrices have to be self-consistent and also consistent with other similarity matrices. 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.” This argument builds on the duality concept of Breiger (1974) and Breiger and Pattison (1986), and is quite similar in spirit to Correspondence Analysis (Greenacre, 1984) and to Latent Semantic Analysis (Landauer and Dumais, 1997).

These two principles provide a unified framework for the analysis of one-mode, two-mode, and multi-mode data. For one-mode data (for example social networks), the representation solves the “two persons are similar if they are linked to similar persons” problem. An example for a two-mode data is the already mentioned

<sup>1</sup> Throughout the paper we shall use the expression “actors” to denote the things that are being compared for their similarity. This is only a notational simplicity, and stands as a shortcut for whatever the unit of analysis is, be it concepts, objects, or attributes for that case. That is, in my usage “actor” do not have any specific connotation such as, for example, agency.

club-membership affiliation network, in which the “people who visit similar clubs are similar” and “clubs that are visited by similar people are similar” relations have to hold. An example for a three-mode data would be an article-scientist-academic journal dataset, for which the following consistency relationships have to hold: “scientists are similar if they publish similar articles”, “academic journals are similar if similar scientists publish in them”, “academic journals are similar if they contain similar articles” – and their symmetric relationships (see Fig. 2).<sup>2</sup>

Of course, we are not the first to generalize direct structural similarity. There exist a variety of concepts and measures that generalizes direct structural similarity, especially in the social networks literature. Just to mention a few, these include: automorphic equivalence (Winship, 1988), regular equivalence (White and Reitz, 1983), and cumulated social roles (Breiger and Pattison, 1986). In a somewhat unusual manner, we postpone the detailed discussion of the connection between the proposed representation and the models in the literature to the second half of the paper. We decided to do so because we want to emphasize the conceptual novelty of the paper, and we believe that the general insights of the principles are valid independently of the merits or disadvantages of the specific measure proposed.

The rest of the paper is structured as follows. In Section 2, we study the need for a generalized representation of similarity data. We outline the two principles for such a representation. Also, in this section we formalize the principles and describe a modified version of Pearson-correlation that meets these principles. In Section 3, to further study and understand the proposed model, we observe its behavior on simulated data. In Section 4, we reanalyze two classic relational datasets with the generalized similarity framework. In doing so, we compare the results with the findings in the similarity and clustering literature. In Section 5, we compare the proposed model to the most commonly used models of similarity in the networks literature: blockmodeling (White et al., 1976), the CONCOR algorithm (Breiger et al., 1975), concepts of more abstract equivalences (e.g., automorphic equivalence, see Borgatti and Everett, 1992), and Correspondence Analysis (Greenacre, 1984). Finally, we discuss the findings and explore directions for further research.

## 2. Two principles for similarity

We believe that a generalized model of similarity should satisfy two principles: it should take the similarity among the dimensions into account, and the similarity matrices should be consistent. Below we discuss these principles in detail and introduce a geometrical representation that satisfies them. For simplicity, we first present the principles through the setting of senators and their votes, that is, through two-mode data. Later, we demonstrate how the same principles apply to various kinds of data (for examples and illustration of the model for one-, two-, and three-mode data, see Fig. 2). In the senator-votes setting, the first principle states that “Two senators are similar if they vote similarly on similar issues.” The second principle, consistency, requires that a similar relationship holds concurrently for the similarity of issues as well, thus “Two issues are similar if similar senators vote similarly on them.”

### 2.1. Principle 1: taking the similarity among the dimensions into account

Take a dataset that consists of senators and their votes in the Senate. Roll call data is one of the most often analyzed dataset in political science (e.g., Clinton et al., 2004; Poole and Rosenthal,

<sup>2</sup> For a discussion of further applications of three-mode data, see Fararo and Doreian (1984).

	Start war A	Raise taxes	Ban Abortion
Senator 1	Yea	Nay	Yea
Senator 2	Yea	Yea	Nay

	Start war A	Start war B	Ban Abortion
Senator 1	Yea	Nay	Yea
Senator 2	Yea	Yea	Nay

Fig. 1. Two hypothetical voting scenarios to illustrate why taking the similarity of contexts into account is important.

1997). As a general approach, senators are viewed similar if they tend to vote the same. To measure similarity, correlation or the cosine distance between the vote vectors are often used. Taking a simple correlation between the voting vectors works rather well: for example, Fig. 8b shows how a Multidimensional Scaling (MDS, see Shepard, 1962) map of the 109th Senate based on correlation as a similarity measure recreates the clustering of senators into two major groups. (We analyze this case later more in detail.)

Note that Pearson-correlation assumes that the dimensions along which the senators are compared (i.e., the votes) are independent. Here, we argue that a similarity measure should take the relationships among the dimensions into account. To see why this is important, consider the two settings in Fig. 1. These settings describe two hypothetical situations in which two senators vote on three issues. Senator 1 votes “Yea”, “Nay”, and “Yea”, while Senator 2 votes “Yea”, “Yea”, and “Nay”, respectively. That is, the senators agree on one issue out of three. The correlation between the senator-vote vectors is  $-0.5$  in both examples (if one codes “Yea” as 1 and “Nay” as  $-1$ ). The votes have, however, markedly different interpretations in the two examples. In the first example, one can assume that the three issues represent three independent dimensions, thus in this case the correlation is a good measure of similarity. In the second example, however, the issues are clearly not independent. If we assume that there exists a pacifist-warmonger dimension, then the first two issues both provide information about the senators’ positions in this dimension. Senator 1 is a middle-of-the-road in questions about war and peace, while Senator 2 is a warmonger. Thus, in the second example there exist only two dimensions, war and abortion, and the vote-vectors of the senators can be rewritten as  $(0,1)$  and  $(1,-1)$ , the correlation of which two vectors is  $-1$ . Clearly, taking the relationships among the dimensions into account is important, and a good representation of similarity needs to incorporate this.

The dimensions along which actors are compared are correlated not only in the case of senators and issues, but virtually in any settings (in some settings more than in others). For example, the dimensions along which demographers group individuals, such as education, income, and gender, are correlated. Or, if the measure of similarity of people is overlap in club-membership, the same argument applies as clubs have their own similarity structure (chess clubs are more similar to other chess clubs than to karate clubs). Indeed, if actors are compared along more than one dimension, then it is hard to find two dimensions that are perfectly independent.

How should one incorporate the non-independence of dimensions into the similarity measure? Let us return to the senators’ example. At the baseline, when the two senators do not cast any votes, their similarity is zero. Principle 1 states that for all issues the two senators vote the same, the similarity between the issues should increase the similarity of the senators. This is the case, for example, if they vote on two wars. (Note that this principle includes the case in which the two issues are exactly the same: In this case, the issues are obviously similar – and their similarity is highest – so the similarity of the senators needs to increase.) Likewise, the similarity of senators should not change if they vote differently on unrelated issues, but should decrease if they vote dissimilarly on similar issues.

Let us introduce the notation. In two-mode data, the relation of one kind of actors to another kind of actors or objects is stored in a rectangular matrix. Examples include senator-issues, people-club membership, people-workplace, and word-document matrices. To stay at a high level of generality, we shall refer to this rectangular matrix as the *actor-setting* matrix (denoted by  $M$ ), with rows denoting actors and columns denoting settings. The cells of  $M$  contain the values of the “actors” along the “settings”. For example, in the senator-vote example, 1 denotes that the senator voted for the bill,  $-1$  denotes he or she voted against, and 0 means he or she abstained. From  $M$ , two similarity matrices can be derived. The first,  $O^1$  contains the pairwise similarity of actors, and  $O^2$ , which contains the pairwise similarity of settings.

$$\underline{M} = \begin{matrix} & s_1 & s_2 & \dots & s_n \\ \begin{matrix} O_1 \\ O_2 \\ \dots \\ O_m \end{matrix} & \left[ \begin{array}{cccc} & & & \\ & & & \\ & & & \\ & & & \\ & & & \end{array} \right] \end{matrix}$$

A common approach to measure similarity among the actors is to take the cosine distance or correlation between the row-vectors (Widdows, 2004). Eq. (1) shows how Pearson-correlation is calculated.

$$\text{correlation}(i, j) = \frac{(M_{i, \cdot} - \overline{M}_{i, \cdot})(M_{j, \cdot} - \overline{M}_{j, \cdot})^T}{\sqrt{(M_{i, \cdot} - \overline{M}_{i, \cdot})(M_{i, \cdot} - \overline{M}_{i, \cdot})^T} \sqrt{(M_{j, \cdot} - \overline{M}_{j, \cdot})(M_{j, \cdot} - \overline{M}_{j, \cdot})^T}}, \quad (1)$$

where  $M_{i, \cdot}$  denotes the  $i$ th row of the  $M$  matrix,  $\overline{M}_{j, \cdot}$  denotes the vector composed of the mean of the  $j$ th row, and  $T$  denotes matrix transposition.

Our starting model of similarity is the Pearson-correlation, which we modify to meet Principle 1. One main problem with Pearson-correlation is that it does not incorporate the similarities among the settings when comparing the actors. There exists, however, an easy way to incorporate this information into the correlation measure. As a basic relationship in linear algebra states, the scalar product of vectors  $x$  and  $y$  in a base space of  $A$  is  $xAy$ . Building on this relationship, we create a modified version of Pearson-correlation that incorporates the non-independence of dimensions. The main idea of the generalized measure is to use the setting-similarity matrix,  $O^2$ , as a base space for calculating the actor similarity matrix (“actors are similar if they appear in similar settings”). Formally, if  $M$  denotes the original  $m \times n$  actor-setting matrix (the input for the model),  $O^1$  denotes the  $m \times m$  actor-actor similarity matrix, and  $O^2$  denotes the  $n \times n$  setting-setting similarity matrix, then the following equation describes the similarity of actors  $i$  and  $j$ <sup>3</sup>:

$$O_{i,j}^1 = \frac{(M_{i, \cdot} - \overline{M}_{i, \cdot})O^2(M_{j, \cdot} - \overline{M}_{j, \cdot})^T}{\sqrt{(M_{i, \cdot} - \overline{M}_{i, \cdot})O^2(M_{i, \cdot} - \overline{M}_{i, \cdot})^T} \sqrt{(M_{j, \cdot} - \overline{M}_{j, \cdot})O^2(M_{j, \cdot} - \overline{M}_{j, \cdot})^T}}. \quad (2)$$

The value of this modified similarity measure is in the range of  $[-1, 1]$ , 1 denoting perfect similarity and  $-1$  denoting perfect dissimilarity. Values around 0 mean independence or neutrality between the actors. Also, note that the similarity measure is symmetric, i.e.,  $O_{i,j}^1 = O_{j,i}^1$ . In Appendix A, we show that this formula satisfies all the requirements we set out for Principle 1: (1) if two actors have similar values on similar dimensions, their similarity increases; (2) if two actors have dissimilar values on similar dimensions, their similarity decreases; (3) if two actors have similar values on dissimilar dimensions, their similarity decreases; and (4) if actors have dissimilar values along dissimilar dimensions, their similarity increases. Thus, we have provided a geometric representation for similarity that is able to incorporate the non-independence of dimensions.

Note that Principle 1 can be applied independently from Principle 2. If the similarity or correlational structure of the dimensions are known, then one can plug this similarity data into Eq. (2), and obtain the similarity of actors in the warped space. So, for example, if the similarity among the issues are given from some external source or can be calculated (for example from matching the text of the bill-proposals), then the similarity of senators can be directly calculated.

### 2.2. Principle 2: the consistency of similarity matrices

In Principle 1, we demonstrated how the pairwise similarity of actors can be obtained if the setting similarity matrix,  $O^2$  is known. But what if the setting similarity matrix is not given? To continue with the senator-vote example of Fig. 1, what if the issues are not known, and we do not know that they represent “War,” “Taxes,” or “Abortion”? With the help of Principle 2, this problem can be overcome, and the similarity of the dimensions can be inferred from the voting data (although not in this specific example, because there are only two senators in the dataset). The trick is to use Principle 1 again on the same input matrix, but instead of comparing the actors (rows), now we compare the settings (columns). In the example of senators and their votes, this would mean that “two issues are similar if similar senators vote similarly on them.” That is, for calculating the setting similarity matrix,  $O^2$ , the actor similarity matrix  $O^1$  can be used as a base space. Formally,

$$O_{i,j}^2 = \frac{(M_{\cdot, i} - \overline{M}_{\cdot, i})O^1(M_{\cdot, j} - \overline{M}_{\cdot, j})}{\sqrt{(M_{\cdot, i} - \overline{M}_{\cdot, i})O^1(M_{\cdot, i} - \overline{M}_{\cdot, i})} \sqrt{(M_{\cdot, j} - \overline{M}_{\cdot, j})O^1(M_{\cdot, j} - \overline{M}_{\cdot, j})}} \quad (3)$$

Principle 2 states that the similarity matrices have to satisfy the consistency conditions: the solution of Eq. (2) have to satisfy Eq. (3), and vice versa.

Eqs. (2) and (3) define a system of equations with two independent variables,  $O^1$  and  $O^2$ . To be more precise, Eqs. (2) and (3) define  $m^2 + n^2$  equations with  $m^2 + n^2$  variables, for each cell in the  $O^1$  and  $O^2$  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  $((m-1)^2 + (n-1)^2)/2$  equations.

Although we did not find analytical solution for the system of Eqs. (2) and (3), we can solve the equations iteratively. Start with  $O_0^2$  equal to the identity matrix (the subscript 0 denotes the 0th iteration). Plug this in to Eq. (2), which yields  $O_1^1$ , the first iteration of the actor-similarity matrix (note that this is equivalent to the similarity matrix from the Pearson-correlation). Next, use this  $O_1^1$  in Eq. (3) to get  $O_1^2$ , the first iteration for  $O^2$ . Repeat until the process converges, i.e., until  $\|O_{t+1}^1 - O_t^1\| < \epsilon$  (where  $\epsilon$  is a pre-defined convergence threshold). Although we found no

<sup>3</sup> Note the similarity of this formula to the Mahalanobis-distance (Mahalanobis, 1936), and to the formula of Breiger (1974, p. 186).



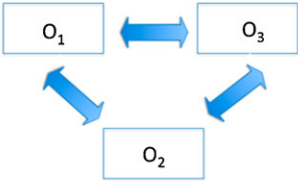
Mode of the data	Dependency between the similarity matrices	Corresponding equations	Example
One-mode		$O_{i,j}^1 = \frac{(M_i - \bar{M}_i) Q^1 (M_j - \bar{M}_j)^T}{\sqrt{(M_i - \bar{M}_i) Q^1 (M_i - \bar{M}_i)^T} \sqrt{(M_j - \bar{M}_j) Q^1 (M_j - \bar{M}_j)^T}}$	Two people are similar if they are connected to similar people.
Two-mode		$O_{i,j}^1 = \frac{(M_i - \bar{M}_i) Q^2 (M_j - \bar{M}_j)^T}{\sqrt{(M_i - \bar{M}_i) Q^2 (M_i - \bar{M}_i)^T} \sqrt{(M_j - \bar{M}_j) Q^2 (M_j - \bar{M}_j)^T}}$ $O_{i,j}^2 = \frac{(M_i - \bar{M}_i)^T Q^1 (M_j - \bar{M}_j)}{\sqrt{(M_i - \bar{M}_i)^T Q^1 (M_i - \bar{M}_i)} \sqrt{(M_j - \bar{M}_j)^T Q^1 (M_j - \bar{M}_j)}}$	Two people are similar if they are member of similar clubs. AND Two clubs are similar if they have similar members.
Three-mode		$O_{i,j}^1 = \frac{(M_i - \bar{M}_i) Q^2 (M_j - \bar{M}_j)^T}{\sqrt{(M_i - \bar{M}_i) Q^2 (M_i - \bar{M}_i)^T} \sqrt{(M_j - \bar{M}_j) Q^2 (M_j - \bar{M}_j)^T}}$ $O_{i,j}^2 = \frac{(M_i - \bar{M}_i)^T Q^1 (M_j - \bar{M}_j)}{\sqrt{(M_i - \bar{M}_i)^T Q^1 (M_i - \bar{M}_i)} \sqrt{(M_j - \bar{M}_j)^T Q^1 (M_j - \bar{M}_j)}}$ $O_{i,j}^3 = \frac{(M_i - \bar{M}_i) Q^3 (M_j - \bar{M}_j)^T}{\sqrt{(M_i - \bar{M}_i) Q^3 (M_i - \bar{M}_i)^T} \sqrt{(M_j - \bar{M}_j) Q^3 (M_j - \bar{M}_j)^T}}$ $O_{i,j}^3 = \frac{(M_i - \bar{M}_i)^T Q^1 (M_j - \bar{M}_j)}{\sqrt{(M_i - \bar{M}_i)^T Q^1 (M_i - \bar{M}_i)} \sqrt{(M_j - \bar{M}_j)^T Q^1 (M_j - \bar{M}_j)}}$ $O_{i,j}^3 = \frac{(M_i - \bar{M}_i) Q^2 (M_j - \bar{M}_j)^T}{\sqrt{(M_i - \bar{M}_i) Q^2 (M_i - \bar{M}_i)^T} \sqrt{(M_j - \bar{M}_j) Q^2 (M_j - \bar{M}_j)^T}}$ $O_{i,j}^2 = \frac{(M_i - \bar{M}_i)^T Q^3 (M_j - \bar{M}_j)}{\sqrt{(M_i - \bar{M}_i)^T Q^3 (M_i - \bar{M}_i)} \sqrt{(M_j - \bar{M}_j)^T Q^3 (M_j - \bar{M}_j)}}$	Two professors are similar if they write similar articles. AND Two articles are similar if they are written by similar professors. AND Two journals are similar if they contain similar articles. AND Two articles are similar if they appear in similar journals. AND Two journals are similar if similar professors write into them. AND Two professors are similar if they write to similar journals.

Fig. 2. Overview and illustration of the consistency conditions for one-, two-, and three-mode data.

proof for convergence, in all the empirical settings we studied, the process converged quite fast, in 10–100 iterations for  $\epsilon = 0.001$ .<sup>4</sup>

Note that this iterated solution of the equations strongly resembles the hierarchical clustering method CONCOR (Breiger et al., 1975). CONCOR takes the correlation of the original co-occurrence matrix, then takes the correlation of the resulting correlation matrix, and iterates this process until convergence. The main difference between the generalized similarity model proposed in this paper and CONCOR is that CONCOR either takes the correlation of the rows or the correlation of the columns, while the generalized similarity model provides a representation which incorporates both row and column correlations simultaneously. Although the mathematical results behind CONCOR cannot be directly applied to the generalized similarity model, it is interesting to note that CONCOR, similarly to the generalized similarity model, achieves convergence rather quickly (for a discussion on the mathematical background of CONCOR, see Chen, 2002).

The two principles, generalization and consistency, apply to relational data of any modality. For one-mode network data, there is only one similarity matrix. In this case, Principle 2 requires this similarity matrix to be self-consistent. In higher mode data, each mode corresponds to a similarity matrix, and Principle 2 requires that these similarity matrices satisfy the consistency principle. Fig. 2

provides examples and an overview of the principles for one-, two-, and three-mode data.<sup>5</sup>

### 3. Properties of the model

In this section, we turn to simulations to investigate the properties of the generalized similarity model. Through simulations, we can explore how the proposed generalized similarity model performs in recovering the underlying data generating process. (When the data is simulated, we exactly know what the underlying data generating process is, while this is not the case for empirical data.)

The structure of the simulations is as follows. First, we stochastically generate datasets based on a model of data. Next, we compare the generalized similarity model's solution with the solution of other similarity and clustering models. In this section, we focus on the comparison with correlation<sup>6</sup>; and later, in Section 5 we compare the generalized similarity model with other models in the literature. In comparing the similarity methods, we focus on two issues: how robust the methods are to local disturbances in the data; and how well they deal with the sparsity of data.

<sup>4</sup> We have investigated how the speed of convergence depends on the dimensionality and sparsity of the matrices. Preliminary results show that larger matrices converge faster than smaller matrices, while sparser matrices converge slower than dense matrices. (Caution is needed in taking these results for granted, as we have only investigated the convergence properties of the generalized similarity model in the senator-issues setting of Section 3). All in all, we can state that the generalized similarity model converges relatively fast, and does not require significant running time with the computing power of current computers.

<sup>5</sup> In this paper, we only discuss one- and two-mode data in detail, and we have not made thorough investigations regarding how the model works for three or higher mode data. For example, it is not guaranteed that the iterative solutions will converge in higher mode data. We leave this issue for further research.

<sup>6</sup> A reviewer has pointed out that because of the iterated nature of the solution the generalized model of similarity should be compared to CONCOR and not to correlation. Indeed, the way of solving the generalized similarity model strongly resembles the iterative solution of CONCOR. However, we believe that the interpretation of the solution itself relates more strongly to the interpretation of the Pearson-correlation, and not to that of CONCOR.



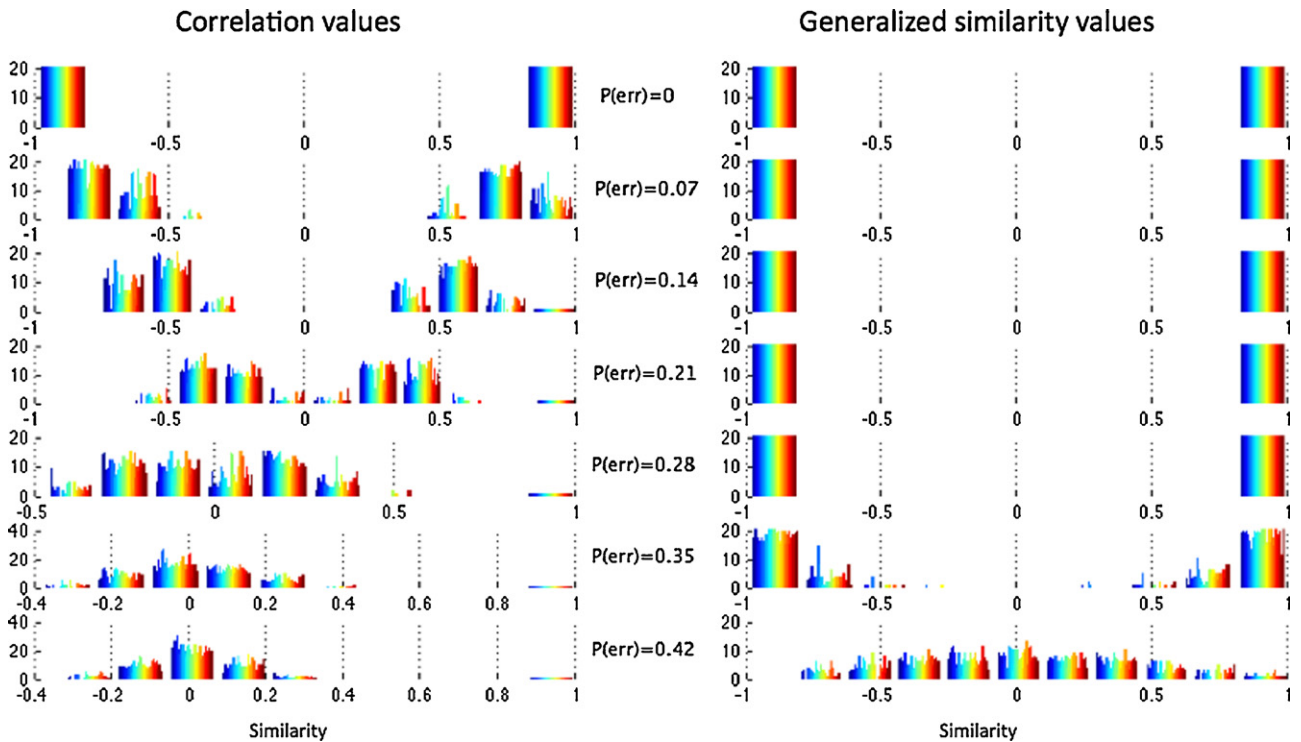


Fig. 3. Comparison of the distribution of similarity values for Pearson-correlation and generalized similarity, in the simulated Senator-vote setting, with the introduction of noise.

3.1. Robustness of classification in the Senate setting

First, we build a simulation that analyzes how the generalized model of similarity works in the U.S. Senate example. We assume that there are 100 senators,<sup>7</sup> and each senator votes on 100 issues. We assume that the voting preferences can be described with an ideal position along a single dimension, liberal-conservative. This assumption is not too far from reality, as political scientists find that most of the variance in roll-call data can be explained by the senators' position along the liberal-conservative dimension (Poole and Rosenthal, 1997). Let 0 denote the liberal, 1 the conservative end of the scale. First, we assume two parties, Red and Blue. The ideal values of the Blue senators are drawn from a normal distribution with mean 0.3 and standard deviation 0.2, and the ideal values of the red senators are drawn from a normal distribution with mean 0.7 and standard deviation 0.2. The issues are similarly located on a liberal-conservative scale, and are drawn from a 0–1 uniform distribution. We further assume that the closer an issue is to the senator's ideal position, the more likely that the senator will vote "Yea". Specifically, the probability of "Yea" is  $P(\text{"Yea"}) = |issue\ position - senator's\ ideal\ position|$ . This is an admittedly simplistic modeling of senatorial votes, but our focus is the illustration of the workings of the generalized model of similarity, and not the introduction of a senatorial choice model.

We use the above simulation framework to investigate how the Pearson-correlation<sup>8</sup> and the general similarity model performs in comparing the senators. First, we explore how robust the methods are to noise in the data. We model noise with a certain  $p$  probability that senator's vote is randomized. Fig. 3 shows the distribution of the pairwise similarity values between the senators: the generalized similarity measure is superior to the simple correlation

measure in identifying the two parties for most levels of  $p$ . The efficiency of the generalized similarity measure is surprisingly high. By taking the cross-issue information into account, it perfectly identifies the parties. This occurs even when the level of noise is high, 35%.

To investigate the workings of the generalized similarity measure more in detail, we compare how correlation and the generalized similarity measure detect the differences in the underlying ideal positions of the senators. Fig. 4 shows that both correlation and the general similarity measure move together with the simi-

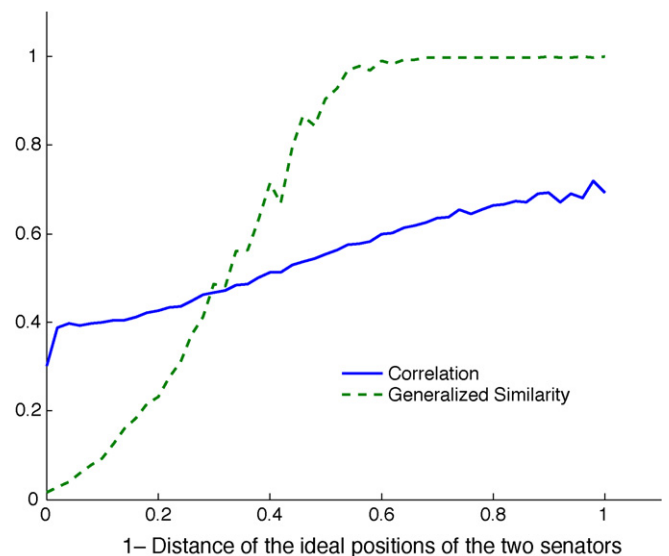


Fig. 4. Simulated two-mode senator data: the comparison of the correlation and the generalized similarity values, as a function of the real difference between the senators.

<sup>7</sup> 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.  
<sup>8</sup> Again, the "Yea" vote is coded with 1, the "Nay" with -1.

larity in the ideal positions (the graph is based on the simulations described in the previous paragraph, with no noise). The figure shows that in the lower similarity regions, the generalized similarity model is more efficient in detecting the underlying differences in the senators' positions. In the higher similarity region, however, the generalized similarity measure does not pick up the differences, but lumps the similar senators together. This graph suggests that (in a two-mode data setting) the generalized similarity measure magnifies within-group similarities and between-group dissimilarities. This magnification implies that the similarity values will gravitate toward 1 (perfect similarity), or  $-1$  (perfect dissimilarity).

We also analyzed how the generalized model of similarity works if there exist more than two groups of senators. Specifically, we assumed that the ideal positions of senators are located in a two dimensional space, and there exist four separate groups (high–high, high–low, low–high, and low–low) along these two dimensions. As previously, we simulated votes based on the senators' similarity to the issues, and analyzed the resulting vote matrix with the generalized model of similarity. We found that the model successfully recovers the underlying groups in this setting as well (results not shown here).

### 3.1.1. Data sparsity

Relational data are often sparse. Such is the case, for example, with a dataset of senatorial votes across various Senates: many senators never voted together. We expect the generalized similarity approach to be especially efficient in sparse-data settings, as it incorporates indirect similarities of actors and settings. Take, for example, three senators, one that served between 1991 and 1999, a second that served between 1995 and 2007, and a third one that served between 2001 and 2007. As the first and the third senators never voted on the same proposition as they never served in the same Senate, first-order similarity measures are not able to assess their similarity. The generalized model of similarity, however, is able to assess the similarity of senators one and three, based on their similarity to senator two.<sup>9</sup>

To investigate the behavior of the generalized similarity model with sparse data, we use the same senator voting model. We now add a probability that the senator will not vote at all (coded with 0). Fig. 5 explores how the correlation between the real similarity of the senators (measured by the difference in their ideal positions) and their similarity according to the generalized similarity measure changes, as a function of (1) the proportion of missing votes, and (2) the number of issues the senators vote on. The figure is based on simulated data (with 100 senators), and each data point represents an average of 100 simulation runs. The figure reveals important features of the generalized similarity model. As the proportion of missing data increases, the accuracy of the generalized similarity model decreases. This finding is not really surprising. Also, with the increase in the number of votes, the accuracy of the generalized similarity model increases. What is really interesting is that it is not purely the vote-matrix size or the sparsity of the matrix that matters, but the combination of the two. Namely, the matrix can be very sparse and the generalized similarity model still works, if the matrix is large. What matters is that the actors should have some overlap, and for large matrices this can be the case even if the density is low.

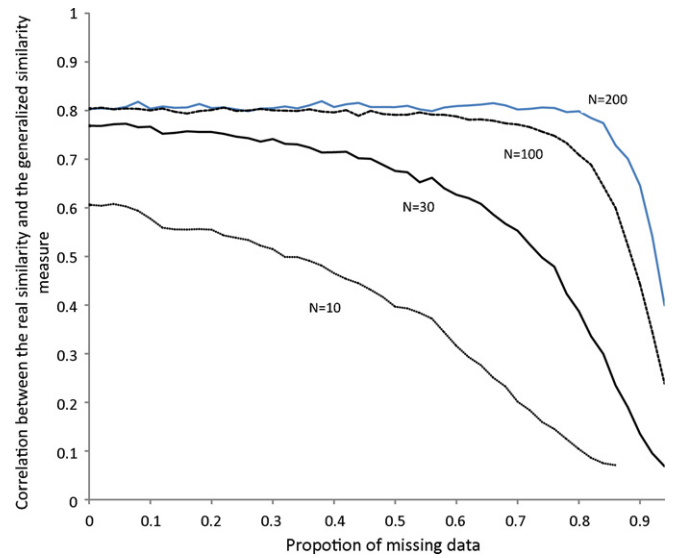


Fig. 5. The correlation between the real similarity of the senators (measured by the difference in their ideal positions) and their similarity according to the generalized similarity measure, as a function of (1) the proportion of missing votes, and (2) the number of issues the senators vote on. Based on simulated data.

### 3.2. Recovering the true social network

How does the generalized similarity framework perform on one-mode social network data? To investigate this question, we build a stochastic network formation model (Snijders et al., 2010). We specify a given distribution of attributes of the nodes, generate random networks based on these attributes, and see how the solution of the generalized similarity model compares to the correlational solution. As we shall see, the generalized similarity measure can recover the real, underlying distribution of data even in stochastic and sparse settings, even when the first-order co-appearance measures fail to do so.

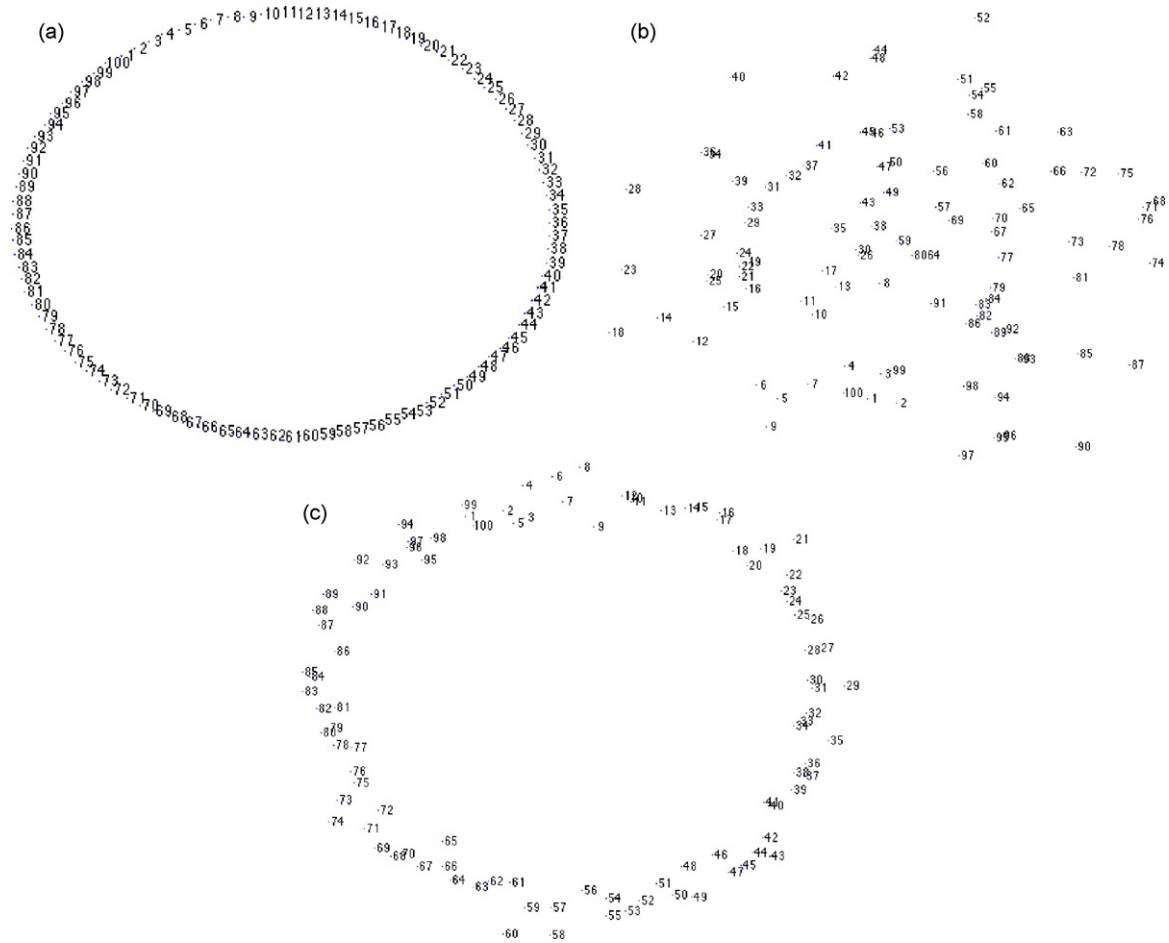
We modeled 100 individuals, indexed from 1 to 100. This number represents the individual's attribute along a dimension; we assume that the individuals are ordered along this dimension such that the ends of the distribution meet and the closer two numbers are, the more similar the individuals are. The similarity map of these individuals is thus a circle (shown in Fig. 6a). We simulate random networks in which the tie creation rule is homophily (McPherson et al., 2001; Snijders et al., 2010): the more similar the individuals are, the more likely that 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."

To calculate the generalized similarity solution for first order data, we use a slightly modified version of Eq. (2). For one mode data, the following equation encompasses both principles:

$$O_{i,j} = \frac{(M_i - \bar{M}_i)O(M_j - \bar{M}_j)^T}{\sqrt{(M_i - \bar{M}_i)O(M_i - \bar{M}_i)^T} \sqrt{(M_j - \bar{M}_j)O(M_j - \bar{M}_j)^T}}. \quad (4)$$

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 in which the individuals only have a few ties. In the first setting, we define the probability of a tie between any two individual to be  $P_{i,j} = 1/(3 \exp(10 \times |distance(i,j)/100|))$ . Thus, the probability that an individual will be connected to its closest neighbor is 30%, to its second closest neighbor is 27%, etc. In the generated networks, the individuals on average have 10 ties. In this setup, both the Pearson-

<sup>9</sup> There exist other methods to find the similarity of actors who are not directly connected. A prime example is Latent Semantic Analysis (Landauer and Dumais, 1997), which uses Singular Value Decomposition to relate terms that do not appear together (See Section 5).



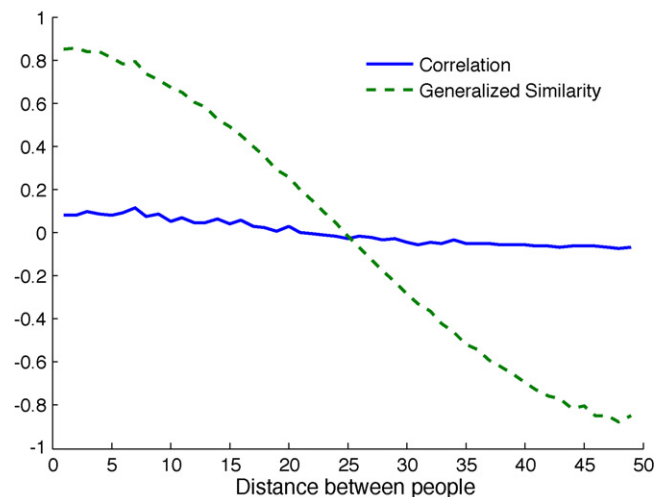
**Fig. 6.** Comparison of the two-dimensional MDS solutions based on (a) real data, (b) Pearson-correlation, and (c) the generalized similarity model for a simulated social network in which people are located along a ring and they only have a few ties.

correlation and the general similarity measure are quite efficient in recovering the original data (for brevity, we do not show the results here).

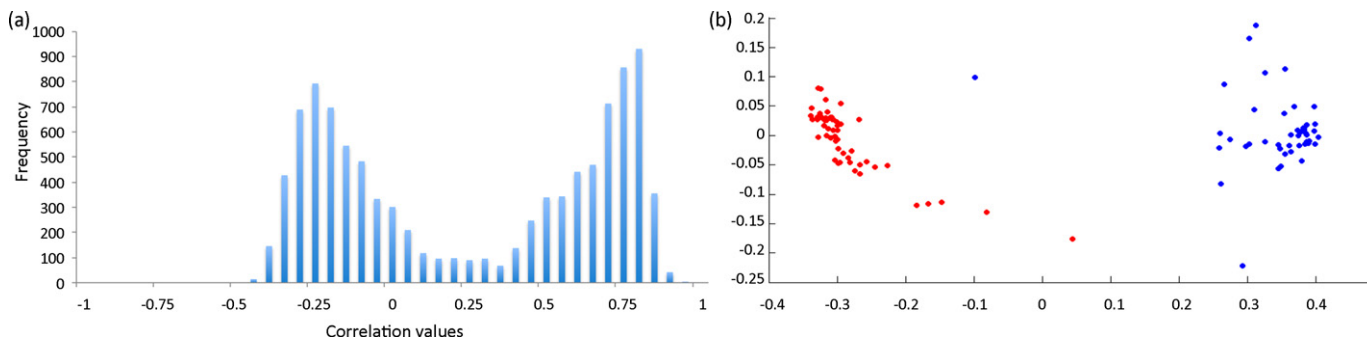
The superiority of the generalized similarity model shows more strongly if we generate networks with fewer ties. For example, if we reduce the probability of ties to  $P_{i,j} = 1/(5 \exp(10 \times |distance(i,j)/100|))$ , then the individuals have on average 7 ties. In this case, the Pearson-correlation measure cannot recover the original structure of data, but the generalized similarity measure can (see Fig. 6b and c). The reason for this is that the Pearson-correlation measure only takes the first-order relational data into account, being exclusively concerned with whether the two individuals are linked to the same individuals or not. When links are rare, most of the individuals will be relatively similar to each other because there is a high overlap in people to whom they do not link. 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 most of the dissimilar individuals (absence of ties). The generalized model of similarity, however, by taking indirect similarities into account, can recover the underlying similarity structure even if the data is sparse. As Fig. 7 shows, the generalized similarity model is much better in picking up the underlying difference among people than the correlation measure. Note that in this case the generalized similarity measure does not overestimate the similarity of the highly similar people (as it was the case in the senator simulations in the previous section), but is superior to correlation in the whole similarity range.

**4. Two empirical illustrations**

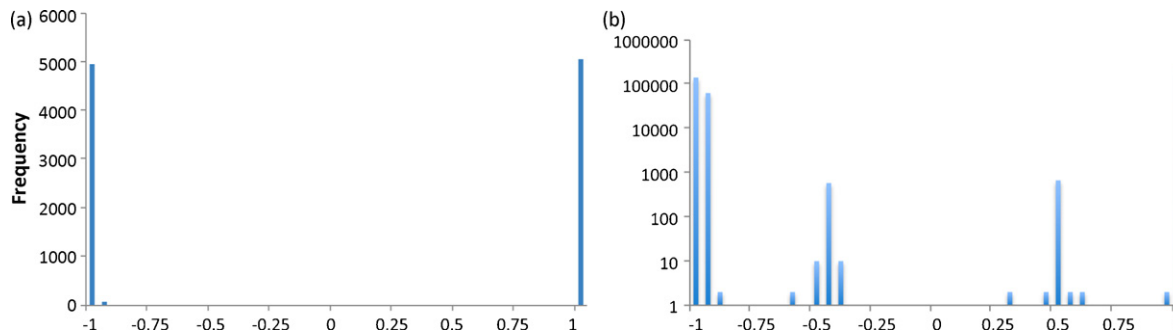
In this section, we illustrate how the generalized similarity model works on empirical data by analyzing two datasets: the roll-call data of the U.S. Senate, and the classic club-membership data



**Fig. 7.** Simulated one-mode data: the comparison of the correlation and the generalized similarity values, as a function of the real difference between the people.



**Fig. 8.** The distribution of the pairwise similarity measures of the senators (a), and the two-dimensional MDS map based on these similarity values (b). Calculated from the 109th U.S. Senate roll-call data.



**Fig. 9.** The distribution of the senator-senator similarity values (a) and the issue-issue similarity values (b), based on the result of the generalized similarity model.

of Davis et al. (1941). We chose these two datasets as they are both commonly analyzed in the literature.

#### 4.1. Similarity of senators and issues

Here 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,<sup>10</sup> and there were 644 issues on which there were no perfect consensus.<sup>11</sup> Thus, the  $M$  matrix, which contains the data, is a  $101 \times 644$  matrix. We coded the “Yea” vote with 1, the “Nay” with -1, and the “Not present” or “Abstain” with 0.

First, we analyze the similarity of the senators using Pearson-correlation. The similarity of senators  $i$  and  $j$  is set equal to the correlation of the their voting vectors,  $M_i$  and  $M_j$ . As there are 101 senators in our dataset, the senator-senator similarity matrix contains  $101 \times 101 = 10,201$  cells. This similarity matrix is symmetric, with 1s in the diagonal. Fig. 8 shows the distribution of the pairwise correlation values (Fig. 8a). The bimodal distribution in Fig. 8a reflects the bipartisan nature of the Senate. Nonetheless, it indicates some overlap between the parties. The two-dimensional MDS map (Fig. 8b) visualizes the similarity map of the senators based on correlation. As can be seen, the map identifies 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. Also, it is important to note that MDS does take the indirect relationships between the senators into account, so the MDS map is a significant enhancement above the simple pairwise correlation.

<sup>10</sup> Robert Menendez filled the seat of Jon Corzine in 2006, when the latter became the Governor of New Jersey.

<sup>11</sup> The data on the votes and senators was retrieved from the U. S. Senate’s website, [http://www.senate.gov/pagelayout/legislative/a\\_three\\_sections\\_with\\_tasers/votes.htm](http://www.senate.gov/pagelayout/legislative/a_three_sections_with_tasers/votes.htm) on April 5th, 2008.

How do the results of the generalized similarity model differ from the results based on Pearson-correlation? Fig. 9a 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 above. The generalized similarity model classifies U.S. senators into two clearly distinct and uniform subsets: Democrats and Republicans. Even if there are within party variance in given votes, the model incorporates across-vote patterns and found that there are no systematic differences between party members, only across the parties. 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 U.S. Senate example, this translates 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 not substantial, and the generalized model of similarity filters out these small differences by pooling across vote data.

The generalized similarity representation provides a similar clustering for issues. As Fig. 9b shows, the issues are bipolar in nature as well, although less 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 underlines the necessity of taking inter-issue relationships into account.

#### 4.2. Comparing senators who never voted together

As simulations show, a major advantage of the generalized measure of similarity is its efficiency in dealing with data sparsity. The 109th Senate dataset is relatively dense, in the sense that there is not much missing data. In order to demonstrate the advantages of the generalized similarity model in a setting with large amount of missing data, we expand the time-frame of the roll-call analysis. As our next step, we analyze the voting data of



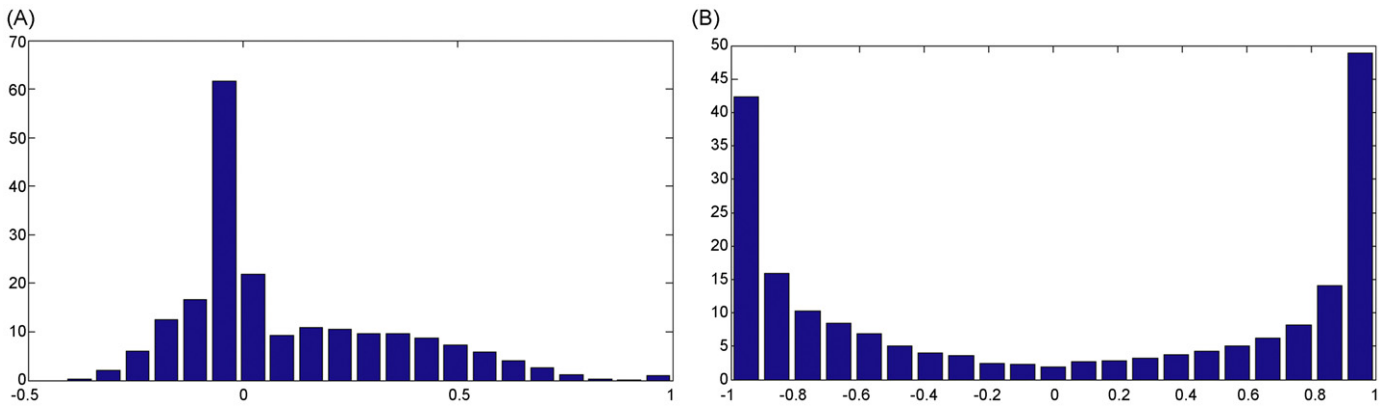


Fig. 10. Comparison of the distribution of Pearson-correlation and generalized similarity for the 202 senators serving in the 101st–110th U.S. Senate. (a) Distribution of the correlation values and (b) distribution of the generalized similarity values.

Actor	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
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	1	1	1	0	1	0	0	0	0	0	0
Eleanor	0	0	0	1	1	1	1	1	0	0	0	0	0	0
Ruth	0	0	0	1	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

Fig. 11. The original Davis et al. (1941) data on the social event participation of 18 Southern women, with the generalized blockmodel solution generated by Doreian et al. (2004).

10 consecutive Senates: the 101st–110th Senates, serving during 1989–2008. These Senates have 202 senators altogether, who voted on 6510 issues, therefore the resulting senator-issue vote matrix has  $202 \times 6,510 = 1,315,020$  cells. As no more than 100 senators can vote on any given issue, the resulting matrix is clearly sparse: 51% of the cells are missing. 28.4% of the senator-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 solutions, we coded the missing data as “Not present,” that is, with 0. Fig. 10 shows the distribution of the correlation (Fig. 10a) and the generalized similarity (Fig. 10b) values for the senator pairs. This figure clearly demonstrates 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. Using correlation, the mode of the

distribution is around zero, which indicates that correlation, not surprisingly, is not able to compare the senator-pairs who never voted together.<sup>12</sup> On the other hand, the generalized similarity measure reveals a bipartisan Senate.

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

Our second illustration uses Davis et al. (1941)'s data on the participation of 18 women in 14 social events. The orig-

<sup>12</sup> One might suspect that this finding is due to the fact that we coded the missing observations as zero. But even if we code the missing observations as missing, the main result does not change: the correlation measure is unable to compare senators who never voted together.

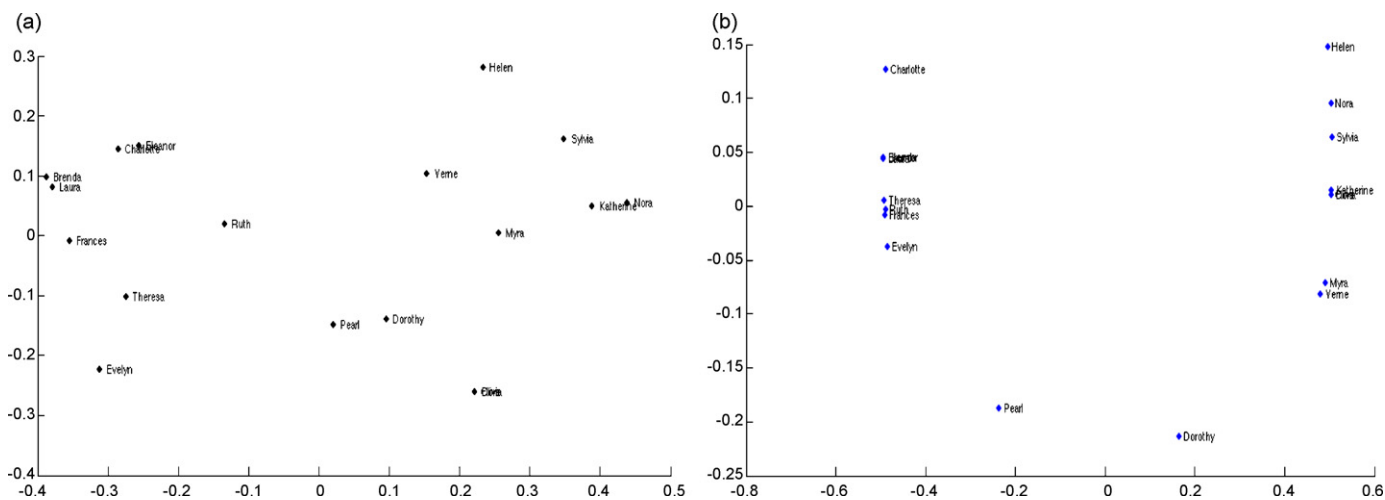


Fig. 12. Comparison of the two-dimensional MDS maps of Pearson-correlation (a) and the generalized similarity model (b) for the Davis et al. (1941) data on the club membership of 18 women.

inal data are shown in Fig. 11 (as sorted by Doreian et al., 2004).

Freeman (2003) provides an exhaustive literature review of 21 articles analyzing the Southern women data. He arrives at the conclusion that the underlying structure of the data is composed of two subgroups of women. One subgroup 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. Note that Freeman (2003) does not analyze the corresponding partition of events.

Doreian et al. (2004), in their article introducing generalized blockmodeling for two-mode network data, reanalyze the Southern women data and arrive at a slightly different conclusion. Instead of two subgroups, they find that there are actually three subgroups of women, with Pearl and Dorothy constituting the third. Simultaneously, they provide a partitioning for the social events into three main subgroups of events: (1, 2, 3, 4, 5), (6, 7, 8, 9), (10, 11, 12, 13, 14). Their partitioning is shown in Fig. 11.

Here we reanalyze the Southern women social event participation data using the generalized similarity model. Fig. 12 shows the two-dimensional Multidimensional Scaling (MDS) maps based on the Pearson-correlation similarity measure and the generalized similarity measure.<sup>13</sup> The MDS map based on correlation (Fig. 12a) is consistent with the blockmodeling results, but does not show clear clustering. The generalized similarity measure (Fig. 12b), however, clearly identifies three clusters, and perfectly recovers the generalized blockmodel solution of Doreian et al. (2004). Note that the generalized similarity model, as in other examples, have magnified the within-group similarity and the between-group dissimilarity, and therefore provides a crisper grouping of the actors.

The generalized similarity model provides a grouping for the events as well (not shown here). This grouping differs slightly 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 (1, 2, 3, 4, 5), while events 7, 8 and 9 do not fall into any group but stand separately.

<sup>13</sup> The Pearson-correlation and generalized similarity values are between  $-1$  and  $1$  ( $-1$  denoting perfect dissimilarity). However, the MDS procedure takes distances as input ( $0$  denoting the closest distance), so the similarity values had to be transformed to dissimilarity values. The rule of transformation used here was  $dissimilarity = (1 - similarity)/2$ .

## 5. Comparison with other similarity models and clustering algorithms

In this section we compare the generalized model of similarity to other major concepts in the literature on similarity and social positions. Specifically, we look at blockmodeling (White et al., 1976), CONCOR (Breiger et al., 1975), regular equivalence (White and Reitz, 1983), and Correspondence Analysis (Greenacre, 1984). We discuss these four modeling frameworks/concepts because we believe these provide the most relevant comparisons to the generalized model of similarity.

### 5.1. Blockmodeling

Blockmodeling was developed in the 1970s to partition the nodes of a network into clusters based on node positions (White et al., 1976). The rationale of this partitioning is relational similarity: nodes in a partition (block) are similar in their relations to other nodes and, therefore, to other blocks that include those nodes. There are two major approaches to blockmodeling: the first approach converts the data into a similarity matrix, and then clusters the actors into blocks based on this similarity matrix. The second approach, generalized blockmodeling, defines a so-called criterion function which evaluates how well a given blockmodeling solution describes the data, and provides as a solution a blockmodeling that minimizes this criterion function (Doreian et al., 2004).

The generalized similarity measure proposed in this paper speaks more closely to the first approach to blockmodeling by providing a new approach to measuring similarity. While most blockmodeling applications use structural equivalence to measure similarity, we provide a similarity measure that generalizes the notion of structural equivalence, and calls nodes equivalent if they have similar relationships to similar nodes. This extension, we argue, is essential for two reasons. It yields a more precise measure of similarity by incorporating indirect similarity; and it provides a better measure of similarity in sparse matrices.<sup>14</sup> Thus, the generalized similarity model can be used to come up with a similarity

<sup>14</sup> Matrices describing relations between large number of actors and settings tend to be sparse (e.g., in a matrix describing the club membership of all Americans in all clubs in the U.S. most of entries would be 0). It is worth noting that most of the applications of blockmodeling analyze small networks, in which the sparsity problem does not arise.

matrix, then the usual clustering methods can be applied to identify blocks.

## 5.2. CONCOR

One of the most commonly used algorithm to define blocks is CONCOR. CONCOR is a hierarchical clustering algorithm, introduced by Breiger et al. (1975). This algorithm has some resemblance to the generalized similarity model: it uses iterated correlations to cluster the relational matrix into blocks. The major difference between CONCOR and the generalized similarity model is that the similarity of the columns in CONCOR is not built into the similarity of the rows, and the algorithm can be used either on row similarity or on column similarity, while the generalized model can provide both row and column correlations simultaneously. The generalized similarity model shows how the row correlations and column correlations can be unified.

Unfortunately, it is not straightforward to compare the two algorithms, as they provide different outputs: the output of the generalized similarity model is a transformed similarity matrix, while the output of CONCOR is a hierarchical clustering. However, we analyzed how CONCOR performs in the social network simulation model of Section 3. Our preliminary analyses show that in sparse data settings (in this case, when people only have a few ties), CONCOR performs better than Pearson-correlation, but worse than the generalized similarity model (for brevity, results not shown here).

## 5.3. Relation to structural equivalence and more abstract equivalences

As we noted in Section 1, generalizing direct structural similarity has produced a sizable body of research, and many models were proposed to generalize structural equivalence, such as automorphic equivalence or regular equivalence (for an overview, see Borgatti and Everett, 1992). The objective of these generalized equivalences is to generalize the “actors are structurally equivalent if they are connected to the same actors” definition to, for example, “a group of actors is regular equivalent if they are connected to other regular equivalent actors.” For example, according to this definition “parents” are regularly equivalent because they have the same relationship to another regularly equivalent class, “children.”

A main difference between structural equivalence and the more general equivalences is whether they require the actors to be connected. In this sense, structural equivalence is a local concept (if two actors are structurally equivalent, it means that they are connected to the same actors, thus they cannot be more than two steps away from each other<sup>15</sup>). The generalized equivalences are, however, independent of proximity: role equivalent actors can be connected, distant, or unreachable from each other. In this sense, the generalized model of similarity lies between structural equivalence and more general equivalences: while it does not require direct connection or close proximity, it does require that the actors are reachable for each other. To see why this is the case, consider that the generalized similarity model will recognize members of two disconnected clusters as independent (thus, for example, it is unable to identify “parents” and “children”).

## 5.4. Correspondence Analysis

Correspondence Analysis (CA) is a generalized principle component analysis for two or higher mode data (Greenacre, 1984). CA transforms the data matrix into two sets of factor scores, which

**Table 1**

Comparison of (1) Pearson-correlation, (2) generalized similarity model, and (3) Correspondence Analysis. Spearman rank correlation on the senators–issues simulations, noise=.3.

	Issue distance	Correlation	Gen. similarity
Correlation	−0.6662		
Gen. similarity	−0.8217	0.7449	
Dist(CA)	0.2862	−0.1841	−0.1476

represent the similarity structure of the row and column actors. Correspondence Analysis is often used to map the similarity structures of higher mode data, especially because the resulting row and column similarity matrices are of the same structure thus can be plotted on the same graph.

Correspondence Analysis is rather close in spirit to the generalized model of similarity. To see why, consider one of its many interpretations: reciprocal averaging (Hill, 1973). In reciprocal averaging, the goal is to recover the underlying values of the rows and the columns, by calculating the row actors’ values as the weighted average of the columns, and the values of the columns as a weighted average of the rows. For example, in the senator-issues setting of the paper, this would translate to the following: the position of a senator (in the issue-space) is the average of the position of the issues it has voted for; and, concurrently, the position of an issue is the average of the position of the senators who voted for it. Correspondence Analysis solves this dual problem.

The spirit of this duality is very close to the second principle of the generalized similarity model, but the two approaches are different. On one hand, CA does not directly look for similarity, but positions the row and column actors in a space, from which the similarity can later be calculated. On the other hand, and this is the more important difference, the averaging is different: CA does not take into account Principle 1 in averaging the actors’ votes. Also, CA is only applicable for two or higher mode data, while the generalized similarity model can be applied to one mode data as well (although one could think of a reciprocal similarity model for one mode data).

How does the solution of the generalized similarity model compares to the solution of CA? Again, it is hard to compare the two models, as they provide different outputs: the generalized similarity model provides a transformed similarity matrix, while CA provides the location of the row and column actors in the issues space. However, this latter can be easily transformed to a distance matrix (by simply taking the distance among the locations), and these can be compared with the transformed similarity matrix.

Table 1 shows how correlation, generalized similarity and Correspondence Analysis compare in the senator simulation setting. To create this table, we ran the senator simulation with 100 senators and 100 issues, with 30% noise level, and ran a Spearman rank correlation between these three measures. As the table illustrates, out of these three measures the generalized similarity model seems to be the most efficient in recovering the underlying similarity of the senators. Note, however, that these results are only for illustration, and to properly compare the generalized similarity model to Correspondence Analysis, further research is required.

## 6. Discussion, applications and further work

This paper proposes two principles for similarity data. First, we emphasize the need for taking relationship among dimensions into account. As we demonstrate on a simple example of senators and issues, it is crucial that the relationship among dimensions are taken into account when comparing actors. Thus, we propose that the approach “two actors are similar if they are related to other actors or objects similarly” should be extended to “two actors are

<sup>15</sup> Of course, this statement only holds if the tie is undirected.

similar if they are related to similar actors or objects similarly.” Just to recall one of the main examples of the paper: “Senators are similar if they vote similarly on similar issues.” The second principle states that similarity matrices should be consistent with each other. Building on the duality argument of Breiger (1974), we require that not only should senators be similar if they vote similarly on similar issues, but issues should be similar if similar senators vote similarly on them.

These principles naturally imply a geometrical representation of data. In this representation, each mode of data represents a dimension. The first principle provides a way to calculate similarity of actors along these dimensions. The second principle warps the dimensions of the space such that the similarity matrices resulting from the warped space satisfy the consistency equations.

When is the generalized similarity model preferable to Pearson-correlation, or to other similarity measures that take only direct relations into account? First, the generalized similarity measure is applicable if the dimensions along which actors are being compared are not independent (if the dimensions are independent in all mode of the data, then the generalized similarity measure is equal to the Pearson-correlation). Second, we have found that the generalized similarity model is especially efficient in analyzing sparse data. The generalized similarity model is also preferable if other similarity measures do not provide a clear clustering of the data: As the generalized similarity model emphasizes within-cluster similarity and between-cluster dissimilarity, it yields a more distinct classification than the standard Pearson-correlation based similarity measures.

The approach proposed in the paper is very general, and can be applied to a wide range of social and natural phenomena. We have already mentioned a few applications in the paper. The generalized similarity measure directly applies to one-mode relational data, such as social network data. The “actors are similar if they are related similarly to similar people” approach can be used to assess the similarity of not only people, but for example organizations. For two-mode data, we analyzed in detail two settings: the senator-vote and people-club membership settings, but clearly the approach applies to a plethora of other settings, including nations belonging to alliances, or organizations employing people.

Applications pertain outside of the traditional domains of the network literature. For example, in linguistics, word co-appearance is a common measure of word-associations and word similarity (Manning and Schütze, 1999): “words that tend to co-appear in the same documents are similar”. Our approach generalizes direct word associations and states that “two words are similar if they appear in similar documents”, and, also, “documents are similar if similar words appear in them.” Although we do not pursue this argument further here, our approach seems to solve the duality between article and document similarity.

The same approach applies to similarity measures in computer science and bibliometrics, which disciplines measure similarity by co-citations. For example, to measure the similarity of webpages, link-overlap is used: two webpages are similar to the extent that they overlap in incoming citations (Dean and Henzinger, 1999). Similarly, two articles are similar if they tend to appear-together in the citation lists (Garfield, 1972). Clearly, our generalized approach to similarity could be applied in both of these settings.

The generalized similarity model might also help in handling another problem of first-order relational data. In settings in which actors serve as substitutes, it is not generally true that the more similar two actors are, the more likely they appear together. 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.” tend to appear in similar sentences, and as people who buy Beethoven con-

**Table 2**

A hypothetical voting record of two senators on three issues.

	Issue 1	Issue 2	Issue 3
Senator 1	1	-1	-1
Senator 2	1	-1	1

certos tend to make other similar purchases, 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 exist. We proposed an approach that can possibly be combined with other methods.

The third, possibly most severe limitation of the proposed model is that it builds on correlations among the dimensions, and correlation is not a useful concept if the relationship between two dimensions is not linear. Such is the case for example between age and income. The model presented in this paper is not able to capture this, and to deal with such cases an alternative model would be required.

### Acknowledgements

This paper builds on my dissertation research at Stanford University. I am grateful for the help and detailed comments of Gianluca Carnabuci, Jerker Denrell, Sasha Goodman, Balázs Geynis, Michael Hannan, Michael Macy, Florin Niculescu, Philippa Pattison, Tomasz Sadzik, Amanda Sharkey, and Károly Takács. The comments of the editor, Patrick Doreian, and of two anonymous reviewers are also appreciated. The paper benefited from the discussions at the Nagymaros Group Conference in Antwerp, the SunBelt conference in San Diego the MORSE Seminar at the University of Lugano, and the Social Networks Workshop and the Macro OB Lunch at Stanford. All remaining errors are my own. The algorithm for the generalized model of similarity can be downloaded from my website.

### Appendix A. A detailed illustration for Principle 1

This section illustrates the basic properties of the modified version of the Pearson-correlation (Principle 1). Throughout this Appendix, we use the example of two senators who vote on three issues, and we shall illustrate how the similarity of the senators change as a function of the similarity of issues.

Take two senators voting on three issues. First we go through one specific constellation of votes (see Table 2).

The baseline is that the issues are independent. That is, the issue similarity matrix is a matrix with 1s in the diagonal and 0s otherwise. In this case the similarity of senators is the Pearson-correlation value, that is, 0.5.

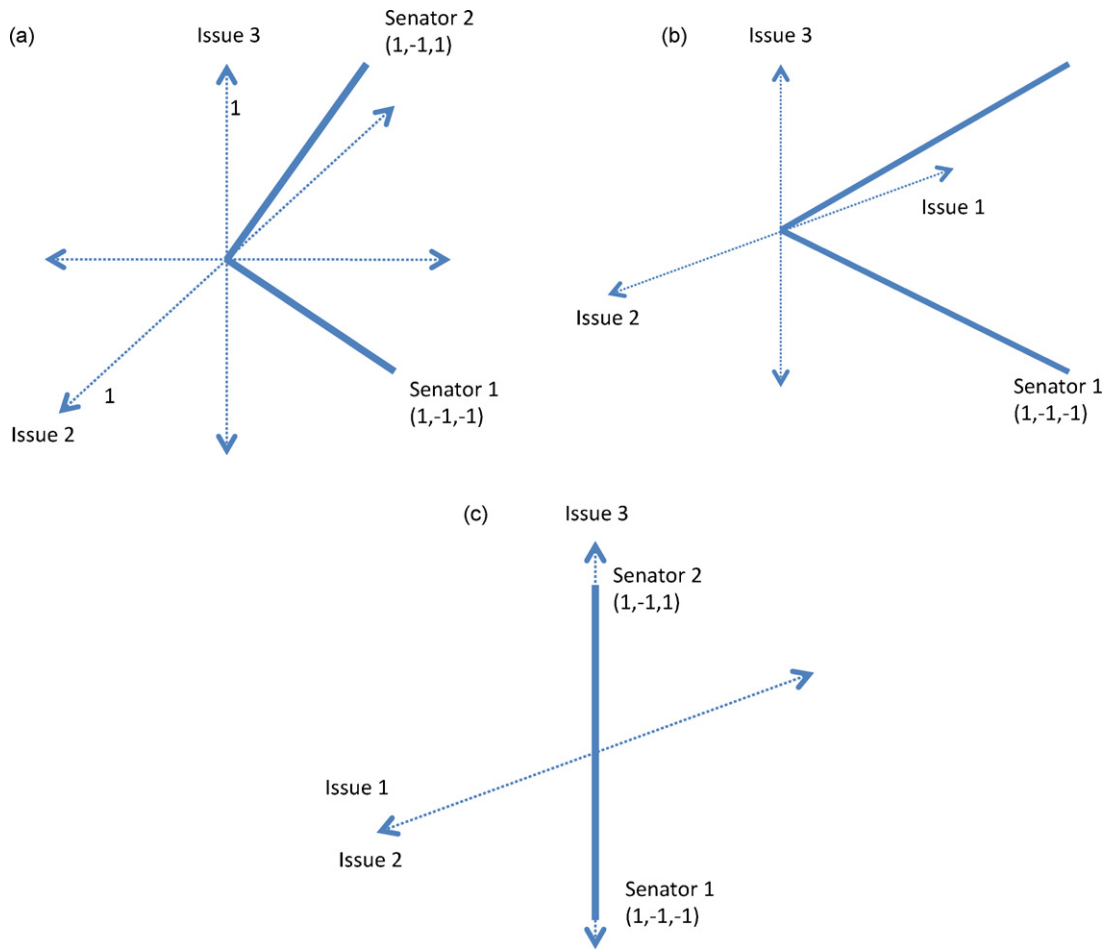
What happens if we introduce some non-independence to the issue similarity matrix? Let  $\alpha$  denote the similarity of Issue 1 and Issue 2. Thus, the similarity matrix is (Table 3):

Fig. 14 illustrates through five cases how the similarity of senators 1 and 2 changes as a function of  $\alpha$ . As the model is rather complex, we discuss each case separately, and then we sum up the main implications of the modified correlation measure.

**Table 3** $\alpha$  denoting the similarity of issues 1 and 2.

	Issue 1	Issue 2	Issue 3
Issue 1	1	$\alpha$	0
Issue 2	$\alpha$	1	0
Issue 3	0	0	1





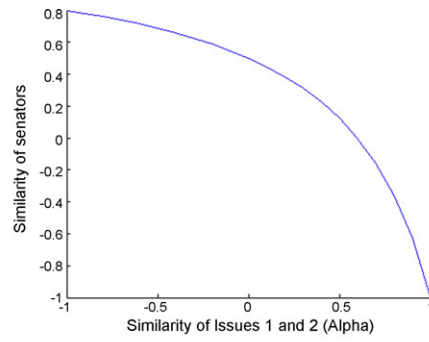
**Fig. 13.** Geometrical illustration of the votes of the two senators in the three different issue-similarity settings. The dotted lines represent the issues, the solid line represents the votes. In the first setting, the issues are independent. In the second setting, Issues 1 and 2 are opposing. In the third setting, Issues 1 and 2 are the same.

In Case 1., senator 1 votes “Yea” on Issue 1, and “Nay” on Issues 2 and 3. Senator 2 votes “Yea” on Issues 1 and 3, and “Nay” on Issue 2. As one can see, if  $\alpha$  is 0, then the similarity of senators 1 and 2 is 0.5, which corresponds to the Pearson-correlation value. Fig. 14 shows that as  $\alpha$  gets bigger, the similarity between senators 1 and 2 decreases. To explain this result, we discuss three scenarios: (1) when  $\alpha$  is  $-1$ , (2) when  $\alpha$  is 0, and when (3) when  $\alpha$  is 1. The three scenarios are illustrated in Fig. 5. When  $\alpha$  is  $-1$ , Issues 1 and 2 are opposing. For example, a “Yea” Issue 1 one means war, but a “Yea” on Issue 2 means peace. The third issue is independent of 1 and 2, it is, say, about education (for a moment assume that education is independent from war and peace). When a given senator votes opposingly on two opposing issue, she makes her position more strongly (a “Yea” on war and a “Nay” on peace). In other words, the two votes add-up. However, if Issues 1 and 2 are similar ( $\alpha = 1$ ), the opposing vote of a given senators on Issues 1 and 2 cancel each other. That is, we can not really tell what is the position of a senator who votes (a “Yea” on war and a “Nay” on another war). Putting these arguments together with the senators’ vote on Issue 3 explains the negative effect of  $\alpha$  on similarity: when Issue 1 and 2 are dissimilar, then the senators have strong opinion about the issues, and because they vote the same on Issues 1 and 2, they will be highly similar. This similarity is stronger then the dissimilarity stemming from disagreement on Issue 3. However, when Issues 1 and 2 are similar, the opposing votes on them cancel each other, thereby putting a stronger weight on Issue 3, on which the senators are dissimilar.

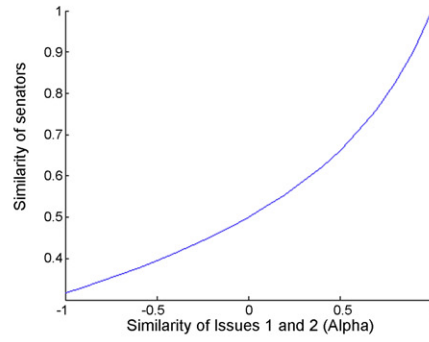
These arguments can be nicely illustrated geometrically, as shown in Fig. 13. As discussed in the “Principle 1: Taking the similarity among dimensions into account” section, the non-independence of the issues is modeled as “warping” of the base space, and the generalized similarity measure is nothing else but the standardized version of the normalized cosine distance in this warped space. Fig. 13a shows the case of  $\alpha = 0$ , and notes the votes of the two senators with a three dimensional voting vector which corresponds to their votes. Fig. 13b and c displays the same two voting vectors, but in the warped base space ( $\alpha = -1$  and  $\alpha = 1$ ).

The other for voting scenarios further illustrate the mechanics of Principle 1. In Case 2 of Fig. 14, senator 1 votes “Yea” on Issues 1 and 2, and “Nay” on Issue 3. senator 2 votes “Yea” on Issue 1, and “Nay” on Issues 2 and 3. The similarity of senators 1 and 2 increases with  $\alpha$ , but note that the similarity is always positive. When  $\alpha = -1$ , the similarity is weaker because the “Yea” and “Nay” votes of senator 1 cancel each other, and this make senator 1 dissimilar from senator 2. In Case 3, the senators vote opposingly on each issues, so they are perfectly dissimilar regardless of the content of the issues. In Case 4., senator 1 votes “Yea” on Issues 1 and 3, and “Nay” on Issue 2. Senator 2 votes “Yea” on Issues 1 and 2, and “Nay” on Issue 3. The senators similarity decreases with the similarity of Issues 1 and 2 ( $\alpha$ ), but note that they are always dissimilar. Finally, Case 5, describes a voting scenario in which senators 1 and 2 vote “Yea, Nay, Nay” and “Nay, Yea, Nay” on the issues, respectively. When Issues 1 and 2 are dissimilar, the “Yea” and “Nay” votes strengthen each other and make the senators rather dissimilar. However, when Issues 1 and

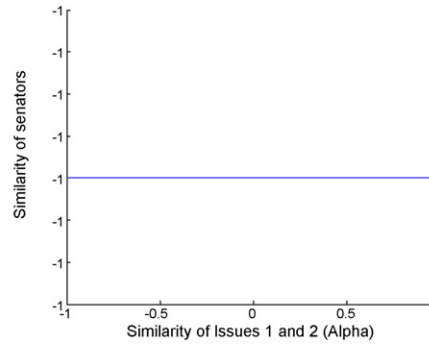
	Issue 1	Issue 2	Issue 3
Senator 1	1	-1	-1
Senator 2	1	-1	1



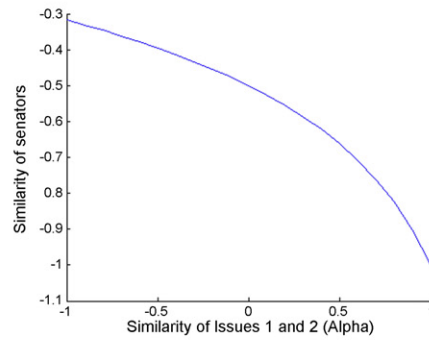
	Issue 1	Issue 2	Issue 3
Senator 1	1	1	-1
Senator 2	1	-1	-1



	Issue 1	Issue 2	Issue 3
Senator 1	1	1	-1
Senator 2	-1	-1	1



	Issue 1	Issue 2	Issue 3
Senator 1	1	-1	1
Senator 2	1	1	-1



	Issue 1	Issue 2	Issue 3
Senator 1	1	-1	-1
Senator 2	-1	1	-1

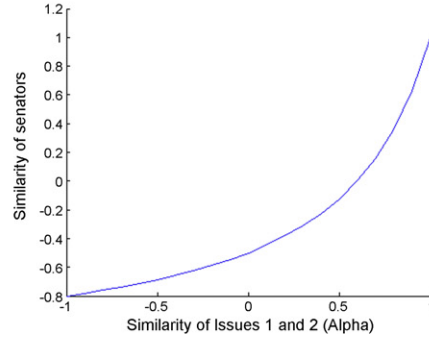


Fig. 14. Five voting scenarios for two senators on three issues.  $\alpha$  denotes the similarity of Issues 1 and 2. Issue 3 is independent from Issues 1 and 2 in all settings.

2 are similar, the “Yea” and “Nay” votes cancel each other, so the similarity of the votes on Issue 3 dominates the dissimilarity on Issues 1 and 2.

## References

- Borgatti, S.P., Everett, M.G., 1992. Notions of position in social network analysis. *Sociological Methodology* 22, 1–35.
- Breiger, R.L., 1974. The duality of persons and groups. *Social Forces* 53, 181–190.
- Breiger, R.L., Boorman, S.A., Arabie, P., 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, R.L., Pattison, P.E., 1986. Cumulated social roles: the duality of persons and their algebras. *Social Networks* 8, 215–256.
- Burt, R.S., 1976. Positions in networks. *Social Forces* 55, 93–122.
- Chen, C.-H., 2002. Generalized association plots: information visualization via iteratively generated correlation matrices. *Statistica Sinica* 12, 7–29.
- Clinton, J., Jackman, S., Rivers, D., 2004. The statistical analysis of roll call data. *American Political Science Review* 98, 355–370.
- Davis, A., Gardner, B.B., Gardner, M.R., 1941. *Deep South: A Social Anthropological Study of Caste and Class*. University of Chicago Press, Chicago.
- Dean, J., Henzinger, M.R., 1999. Finding related pages in the World Wide Web. *Computer Networks* 31, 1467–1479.
- Doreian, P., Batagelj, V., Ferligoj, A., 2004. Generalized blockmodeling of two-mode network data. *Social Networks* 26, 29–53.
- Fararo, T.J., Doreian, P., 1984. Tripartite structural analysis: generalizing the Breiger–Wilson formalism. *Social Networks* 6, 141–175.
- Freeman, L.C., 2003. Finding social groups: a meta-analysis of the Southern Women data. In: Breiger, R., Carley, P., Pattison, P. (Eds.), *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*. National Academies Press.
- Garfield, E., 1972. Citation analysis as a tool in journal evaluation. *Science* 178, 471–479.
- Greenacre, M.J., 1984. *Theory and applications of correspondence analysis*. Academic Press, London.
- Hill, M.O., 1973. Reciprocal averaging: an eigenvector method of ordination. *Journal of Ecology* 61, 237–249.
- Hume, D. 2004 (1748). *An Enquiry Concerning Human Understanding*. Dover, Mineola, NY.
- Landauer, T., Dumais, S., 1997. A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104, 211–240.
- Lorrain, F.P., White, H.C., 1971. Structural equivalence of individuals in networks. *Journal of Mathematical Sociology* 1, 49–80.
- Mahalanobis, P.C., 1936. On the generalised distance in statistics. *Proceedings of the National Institute of Science of India* 12, 49–55.
- Manning, C.D., Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.
- McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 415–444.
- Murphy, G.L., 2002. *The Big Book of Concepts*. MIT Press.
- Poole, K.T., 2007. Changing minds? Not in Congress! *Public Choice* 131, 435–451.
- Poole, K.T., Rosenthal, H., 1997. *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press, USA.
- Shepard, R.N., 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika* 27, 125–140, 219–246.
- Shepard, R.N., 1987. Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323.
- Snijders, T.A.B., van de Bunt, G.G., Steglich, C.E.G., 2010. Introduction to stochastic actor-based models for network dynamics. *Social Networks* 32, 44–60.
- White, D.R., Reitz, K.P., 1983. Graph and semigroup homomorphisms on networks of relations. *Social Networks* 5, 143–234.
- White, H.C., Boorman, S.A., Breiger, R.L., 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology* 81, 730–780.
- Widdows, D., 2004. *Geometry and Meaning*. Center for the Study of Language and Information/SRI.
- Winship, C., 1988. Thoughts about roles and relations: an old document revisited. *Social Networks* 10, 209–231.