A Generalized Model for Similarity[*]

Balázs Kovács
University of Lugano
kovacsb@usi.ch

October 2009

Preliminary draft, do not quote.

Abstract

This paper introduces two principles for similarity, and based on these principles it proposes a novel geometric similarity representation. The first principle generalizes earlier measures 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.

[*] This paper builds on my dissertation research at Stanford University. I am grateful for the help and detailed comments of Jerker Denrell, Sasha Goodman, Balázs Gyenis, Michael Hannan, Michael Macy, Florin Niculescu, Tomasz Sadzik, Amanda Sharkey, and Károly Takács. The paper also benefited from the discussions at the Nagymaros Group Conference in Antwerp, the SunBelt conference in San Diego, and the Social Networks Workshop and the Macro OB Lunch at Stanford. All remaining errors are my own. The algorithm for the Generalized Model for Similarity can be downloaded from my website.

> *In reality, all arguments from experience are founded on the similarity which we discover among natural objects, and by which we are induced to expect effects similar to those which we have found to follow from such objects. (...) From causes which appear* similar *we expect similar effects. This is the sum of all our experimental conclusions.*
>
> David Hume: *An Enquiry Concerning Human Understanding*, 2004: 21 (1748)

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.

Despite the prevalence and importance of similarity, the meaning of similarity is not agreed-upon. A natural way of thinking about similarity is the *attribute-based similarity* approach. In this approach, researchers identify the main attributes of actors[1], and assess their similarity based on their similarity along these attributes. For example, social stratification studies classify people based on their gender, age, income level, education, etc. 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

[1] 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.

similarity. The underlying principle of *relational similarity* is that actors 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). Attribute-based and relational approaches to similarity have different epistemic assumptions and markedly different data requirements. There has been much research in both approaches to similarity (for overviews, see Bailey, 1994; Borgatti and Everett, 1992; Lattin et al., 2002). These two approaches are not necessarily contradicting, and often provide corresponding similarity ratings (Kovács, 2009)[2].

This article is an endeavor to rethink the concept of similarity. We propose two principles that we believe similarity representations should satisfy. Our primary aim is thus conceptual, not methodological. The beauty of the proposed representation is, however, that the methodology naturally "pops out" from the principles.

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. We argue that 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 Bonacich's measure of status, which states that people are of high status if linked to by other high status people (Bonacich, 1987).

These two principles provide a unified framework for the analysis of one-

[2] There exists a third major approach to similarity: perceptual similarity (see e.g., Medin, 2005). We do not discuss perceptual similarity in detail here because we think that it is consistent with the relational approach.

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 three-mode data would be an article-scientist-academic journal publication dataset, for which the following six 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 Figure 2).

This paper provides another, one might say radical reading of the connection of similarity in attributes and similarity in relations. If we follow the logic to the extreme, we can view attributes as modes in relational data. For example, the statement: "persons A and B are similar because they are both educated" can be recast in the terms "persons A and B are similar because they have the same relationship to the concept *Education*." While this re-conceptualization of attributes in relational terms is not magical (later in the paper we discuss this point carefully), its implications are far reaching. If we accept this re-conceptualization, then we can apply both the generalization and the consistency principles to the attributes as well. For example, people are similar because they have similar educational levels, but at the same time educational levels are similar because similar people have them. This way the attributes become fluid. As an analogy, we might call this the generalized theory of relativity for similarity data, because it reminds of the general theory of relativity in physics: how mass influences the curvature of spacetime, which in return influences the behavior of mass (Einstein, 1916).

To make this view realistic, we need to make an adjustment to the second principle. Specifically, we need to allow for stickiness of the dimensions. This is needed because any relational data is necessarily a subsample of all relational data, and fully fitting the dimensions of the similarity space to the sample would lead to overidentification. To continue the previous example, if we have data on the education level of a few people, then we do not want the similarity of these few people to fully determine the similarity of the educational levels. The similarity of educational levels would depend on the similarity between people on the whole population, but if we only have a subsample of this population, then we need to control for this. In the model, we shall thus include a parameter that controls the stickiness of the dimension. Note that the corner solutions, when some of the dimensions are fixed, correspond to the previous models in the literature that treat dimensions as independent.

Of course, we are not the first ones to try to generalize direct structural similarity. There exist a plethora of concepts and measures that attempt this, especially in the social networks literature. Just to mention a few, these are: 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 from 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 literature: the CONCOR algorithm for clustering (Breiger et al., 1975), blockmodeling (White et al., 1976), methods for calculating more abstract equivalences (REGE and MaxSim), and Singular Value Decomposition (with an emphasis on Correspondence Analysis and Latent Semantic Analysis). 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 Figure 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.”

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). As a general approach, senators are viewed similar if they tend to vote together. To measure similarity, thus, correlation or the cosine distance between the vote vectors is often used. Taking a simple correlation between the voting vectors works rather well: for example, Figure 5b 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., 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 Table 1. These settings describe two hypothetical situations in which two senators vote on three issues. Senator 1 votes “Yeah”, “Nay”, and “Yeah”, while Senator 2 votes “Yeah”, “Yeah”, and “Nay”, respectively. That is, the senators agree on one issue out of three. The correlation between the senator-vote vectors is -.5 in both examples (if one codes “Yeah” 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.

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

(a) Example 1

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

(b) Example 2

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

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 for 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 increases.) Likewise, the similarity of senators should not change if they vote differently on unrelated issues, but should decrease if they vote similarly on opposing issues.

In two-mode data, the relation of one kind of object to another set of 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 $\underline{M}$), with rows denoting actors and columns denoting settings. The cells of $\underline{M}$ 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 $\underline{M}$, two similarity matrices can be derived. The first, $\underline{O}^1$ contains the pairwise similarity of actors, and $\underline{O}^2$, which contains the pairwise similarity of settings.

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

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

$$correlation(i,j) = \frac{(M_{i,} - \overline{M_{i,}})(M_{j,} - \overline{M_{j,}})^T}{\sqrt{(M_{i,} - \overline{M_{i,}})(M_{i,} - \overline{M_{i,}})^T}\sqrt{(M_{j,} - \overline{M_{j,}})(M_{j,} - \overline{M_{j,}})^T}}, \tag{1}$$

where $M_{i,}$ denotes the ith row of the $\underline{M}$ matrix, $\overline{M_{j,}}$ denotes the vector composed of the mean of the jth 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 can create a modified version of Pearson-correlation that handles the non-independence of dimensions. The main idea of the generalized measure is to use the setting-similarity matrix, $\underline{O}^2$, as a base space for calculating the actor similarity matrix ("actors are similar if they appear in similar settings"). Formally, if $\underline{M}$ denotes the original $m \times n$ actor-setting matrix (the input for the model), $\underline{O}^1$ denotes the $m \times m$ actor-actor similarity matrix, and $\underline{O}^2$ denotes the $n \times n$ setting-setting similarity matrix, then the following equation describes the similarity of actors i and j:

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

The value of this modified similarity measure is in the range of $[-1, 1]$, 1 denoting perfect similarity and -1 denoting perfect dissimilarity. Also, note that the similarity measure is symmetric, i.e., $O^1_{i,j} = O^1_{j,i}$. In the Appendix, we show that this formula satisfies all the requirements we set out for Principle 1: (1) if two actors have similar values on dissimilar dimensions, their similarity decreases; (2) if two actors have similar values on similar dimensions, their similarity increases; (3) if two actors have dissimilar values on similar 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 Equation (2), and obtain the similarity of actors in the warped space. So, for example, if the similarity among the issues are given or can be calculated from some external source (for example from matching the text of the bill-proposals), then the similarity of senators can be directly calculated. Similarly, from the correlation between educational level, age, income, gender, race, one can easily get a modified similarity of people, which might lead to different picture of stratification.

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, $\underline{O}^2$ is known. But what if the setting similarity matrix is not given? With the help of Principle 2, this problem can be overcome. The trick is to use Principle 1 again on the same input matrix, but instead of comparing the actors, now we compare the settings. Back to our 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, $\underline{O}^2$, the actor similarity matrix $\underline{O}^1$ can to be used as a base space. Formally,

$$O^2_{i,j} = \frac{(M_{,i} - \overline{M_{,i}})^T \, \underline{O}^1 \, (M_{,j} - \overline{M_{,j}})}{\sqrt{(M_{,i} - \overline{M_{,i}})^T \, \underline{O}^1 \, (M_{,i} - \overline{M_{,i}})}\sqrt{(M_{,j} - \overline{M_{,j}})^T \, \underline{O}^1 \, (M_{,j} - \overline{M_{,j}})}} \tag{3}$$

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

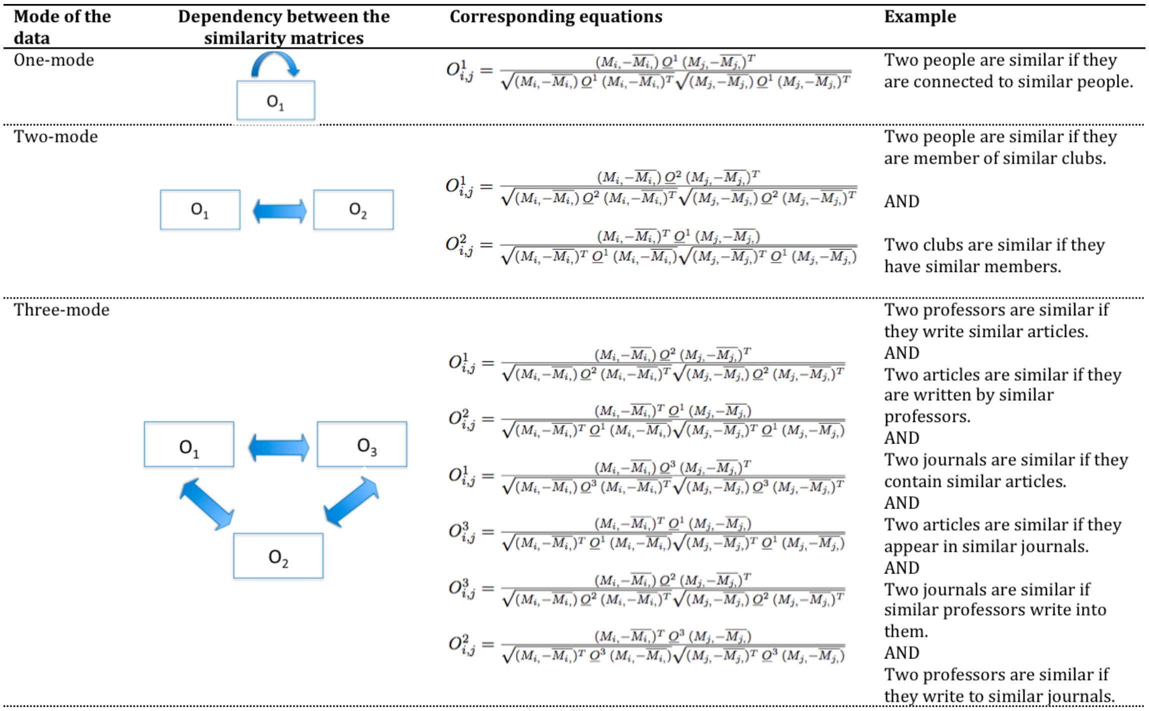
Equations (2) and (3) define a system of equations with two independent variables, $\underline{O}^1$ and $\underline{O}^2$. To be more precise, Equations (2) and (3) define $m^2 + n^2$ equations with $m^2 + n^2$ variables, for each cell in the $\underline{O}^1$ and $\underline{O}^2$ matrices. Excluding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

Although there seems to be no analytical solution for Equations (2) and (3), one can solve the system of equations iteratively. Start with $\underline{O}^2_0$ equal to the identity matrix. The subscript 0 denotes the 0^{th} iteration. Plug this in to Equation (2), which yields $\underline{O}^1_1$, the first iteration of the actor-similarity matrix (note that this is equivalent to the similarity matrix from the Pearson-correlation). Next, use this $\underline{O}^1_1$ in Eq. (3) to get $\underline{O}^2_1$, the first iteration for $\underline{O}^2$. Repeat until the process converges, i.e., until $||\underline{O}^1_{t+1} - \underline{O}^1_t|| < \epsilon$ (where ϵ is a pre-defined convergence threshold). Although we found no proof for convergence, in all the empirical settings we studied, the process converged quite fast, in 5-20 iterations for $\epsilon = 0.0001$.

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. Figure 2 provides examples and an overview of the principles for one-, two-, and three-mode data.

The "stickiness" of dimensions

Principle 2 states that at the solution of the generalized similarity model the similarity of actors and the similarity of settings should hold concurrently. This assumes that the similarity of actors within each mode of the data can be fully identified from comparing the actors along the other modes of data.



Mode of the data	Dependency between the similarity matrices	Corresponding equations	Example
One-mode	O_1	$O^1_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})\,\underline{O}^1\,(M_{j,}-\overline{M_{j,}})^T}{\sqrt{(M_{i,}-\overline{M_{i,}})\,\underline{O}^1\,(M_{i,}-\overline{M_{i,}})^T}\sqrt{(M_{j,}-\overline{M_{j,}})\,\underline{O}^1\,(M_{j,}-\overline{M_{j,}})^T}}$	Two people are similar if they are connected to similar people.
Two-mode	O_1 ⟷ O_2	$O^1_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})\,\underline{O}^2\,(M_{j,}-\overline{M_{j,}})^T}{\sqrt{(M_{i,}-\overline{M_{i,}})\,\underline{O}^2\,(M_{i,}-\overline{M_{i,}})^T}\sqrt{(M_{j,}-\overline{M_{j,}})\,\underline{O}^2\,(M_{j,}-\overline{M_{j,}})^T}}$ $O^2_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})^T\,\underline{O}^1\,(M_{j,}-\overline{M_{j,}})}{\sqrt{(M_{i,}-\overline{M_{i,}})^T\,\underline{O}^1\,(M_{i,}-\overline{M_{i,}})}\sqrt{(M_{j,}-\overline{M_{j,}})^T\,\underline{O}^1\,(M_{j,}-\overline{M_{j,}})}}$	Two people are similar if they are member of similar clubs. AND Two clubs are similar if they have similar members.
Three-mode	O_1 ⟷ O_3, O_1 ⟷ O_2, O_3 ⟷ O_2	$O^1_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})\,\underline{O}^2\,(M_{j,}-\overline{M_{j,}})^T}{\sqrt{(M_{i,}-\overline{M_{i,}})\,\underline{O}^2\,(M_{i,}-\overline{M_{i,}})^T}\sqrt{(M_{j,}-\overline{M_{j,}})\,\underline{O}^2\,(M_{j,}-\overline{M_{j,}})^T}}$ $O^2_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})^T\,\underline{O}^1\,(M_{j,}-\overline{M_{j,}})}{\sqrt{(M_{i,}-\overline{M_{i,}})^T\,\underline{O}^1\,(M_{i,}-\overline{M_{i,}})}\sqrt{(M_{j,}-\overline{M_{j,}})^T\,\underline{O}^1\,(M_{j,}-\overline{M_{j,}})}}$ $O^1_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})\,\underline{O}^3\,(M_{j,}-\overline{M_{j,}})^T}{\sqrt{(M_{i,}-\overline{M_{i,}})\,\underline{O}^3\,(M_{i,}-\overline{M_{i,}})^T}\sqrt{(M_{j,}-\overline{M_{j,}})\,\underline{O}^3\,(M_{j,}-\overline{M_{j,}})^T}}$ $O^3_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})^T\,\underline{O}^1\,(M_{j,}-\overline{M_{j,}})}{\sqrt{(M_{i,}-\overline{M_{i,}})^T\,\underline{O}^1\,(M_{i,}-\overline{M_{i,}})}\sqrt{(M_{j,}-\overline{M_{j,}})^T\,\underline{O}^1\,(M_{j,}-\overline{M_{j,}})}}$ $O^3_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})\,\underline{O}^2\,(M_{j,}-\overline{M_{j,}})^T}{\sqrt{(M_{i,}-\overline{M_{i,}})\,\underline{O}^2\,(M_{i,}-\overline{M_{i,}})^T}\sqrt{(M_{j,}-\overline{M_{j,}})\,\underline{O}^2\,(M_{j,}-\overline{M_{j,}})^T}}$ $O^2_{i,j} = \frac{(M_{i,}-\overline{M_{i,}})^T\,\underline{O}^3\,(M_{j,}-\overline{M_{j,}})}{\sqrt{(M_{i,}-\overline{M_{i,}})^T\,\underline{O}^3\,(M_{i,}-\overline{M_{i,}})}\sqrt{(M_{j,}-\overline{M_{j,}})^T\,\underline{O}^3\,(M_{j,}-\overline{M_{j,}})}}$	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.

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

This would indeed be the right view if one were to have all data. Almost any dataset, however, only contains information on a subset of the dimensions. If the similarity of dimensions in each mode would be fully fit to the observed data, then we would overfit the data. To understand why this is the case, consider the following example. Imagine we have a two-mode dataset on people and their club membership (for an example, see Figure 8). If we were to directly apply Principle 2, then we would accept that the similarity of the clubs can be perfectly measured by the overlap in membership, and the similarity of the members can be perfectly measured by the similarity in their club-memberships. But this is clearly not the case, as both the similarity of the people and the similarity of the clubs are influenced by various other dimensions, such as age, gender, or similarity in taste (people); and location, private or public, and, say, opening hours (clubs). For a perfect comparison, we would need data on all of these data types. Given that typically not all data is available, one can not be sure about the similarity matrices, and this implies that when taking the similarity matrices as a base-space one needs to be careful. For this reason, here we introduce a new concept, the "stickiness" of a dimension, which denotes the extent to which the similarity of the dimensions should be fitted to the similarities along other dimensions. For example, if we have data on the educational level of five people, we do not want the similarity of the educational levels to be determined by this dataset, so we say that the data-type "education" is sticky in this case. In other words, one can view stickiness as a proxy for the extent to which the sample is comprehensive.

To operationalize stickiness of the dimensions, we introduce a parameter, ρ, for each mode of the data. Specifically, this ρ finetunes the extent to which the similarity matrix $O^2_{i,j}$ is updated in the iterative solution of the generalized similarity model. Formally, Equation (3) changes to

$$O^2_{i,j,(t+1)} = \rho \cdot O^2_{i,j,(t)} + (1-\rho) \cdot \frac{(M_{,i}-\overline{M_{,i}})^T \, \underline{O}^1 \, (M_{,j}-\overline{M_{,j}})}{\sqrt{(M_{,i}-\overline{M_{,i}})^T \, \underline{O}^1 \, (M_{,i}-\overline{M_{,i}})}\sqrt{(M_{,j}-\overline{M_{,j}})^T \, \underline{O}^1 \, (M_{,j}-\overline{M_{,j}})}} \tag{4}$$

$O^2_{i,j,(t)}$ denotes the t^{th} iteration of the issue-similarity matrix. Note that if $\rho = 1$, then the model corresponds to the Pearson-correlation solution, and if $\rho = 0$, then the model corresponds to the extreme version of the generalized similarity model (as formulated in Principle 2).

3 Properties of the model

In this section, we turn to simulations to investigate the properties of the generalized similarity model. Simulations are needed because the properties of the model can only be evaluated if the underlying data generating process is known.

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 to the correlational solution to see how well they recover the underlying similarity structure. 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.

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 two parties, Red and Blue, each with 20 senators[3]. Each senator votes on 100 issues. We assume a perfect bipolar system: on issues on which the Red senators vote "Yeah", the Blue-s vote "Nay", and vice versa. The "Yeah" vote is coded with 1, the "Nay" with -1.

If all senators vote perfectly with their parties, then both similarity measures (correlation and the generalized model of similarity) 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 random disturbance. That is, there is a certain p probability that senators will not vote with their parties.

How the two similarity measures — Pearson-correlation and the generalized model of similarity — compare in identifying the two groups of senators and issues in such a "noisy" environment? Figure 3 shows the results: 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 can perfectly identify the parties. This occurs even when the senators deviate from their party's vote one third of the time. The algorithm is similarly efficient in classifying the issues (for brevity, the result is not shown here).

Next, we investigate how the generalized similarity method performs if there is an underlying heterogeneity within the parties, that is, if there exist Red-leaning Blues and Blue-leaning Reds. To investigate this setup, we incorporate systematic differences into 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 their party's vote. Simulations show that the generalized similarity measure is less effective in picking up the internal differences and recovering the data than is the Pearson-correlation.

We run further simulations to compare the similarity algorithms under more complicated underlying models in which there are more real groups,

[3] 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.

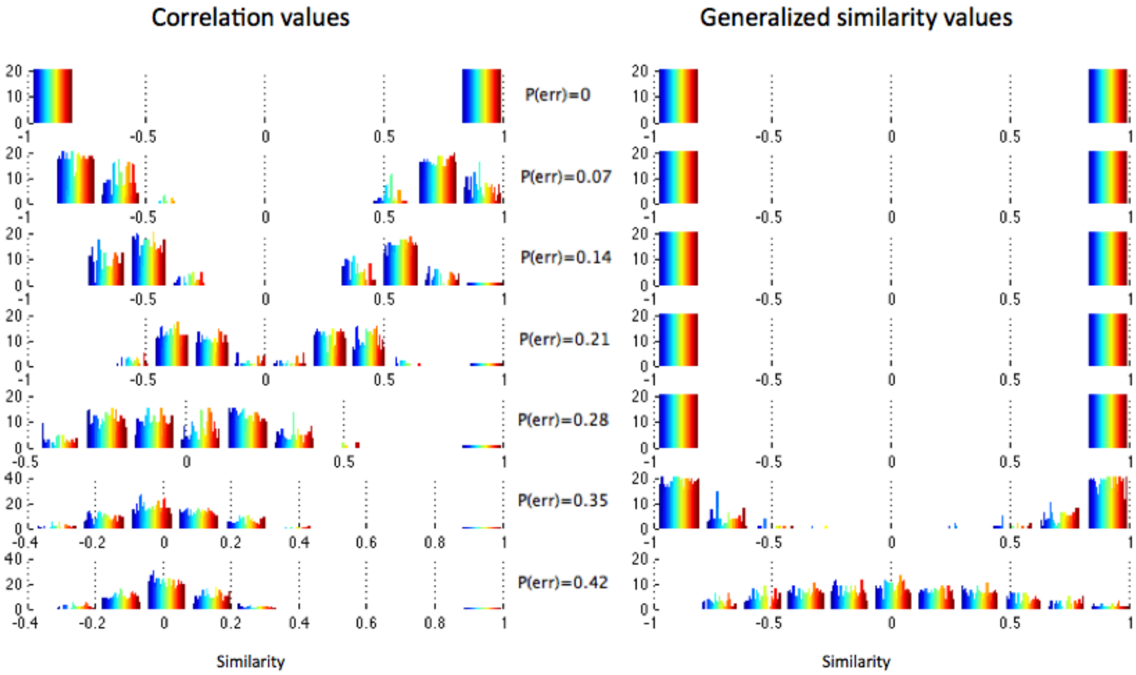


Figure 3: Comparison of the distribution of similarity values for Pearson-correlation and generalized similarity, in the simulated Senator-vote setting, with the introduction of random disturbance

more options along the dimensions, etc. The results remained practically the same.

Data sparsity. As pointed out earlier, in many cases the existing data are sparse, meaning that actors do not have any values along some of the dimensions. Sparsity makes the comparison of actors more cumbersome, as if actors do not have values along the same dimensions then they cannot be directly compared. Consider the case of the Senate again: senators are similar if they vote similarly on issues, but how could we compare two senators who never voted on the same issues? Clearly, Pearson correlation and other direct measures of similarity are not able to handle these cases. The generalized model of similarity, however, might still be able to derive the similarity of senators who never voted together by incorporating the indirect similarities. Consider three senators: senator 1 served from 1993 to 2001, senator 2 served from 1997 to 2005, and senator 3 served from 2003 to 2007. Senators 1 and 3 never voted on the same issues, so they cannot be compared directly. The generalized similarity model, however, is able to compare them as it compares both senators 1 and 3 to senator 2, and from these comparisons it can infer the similarity of senators 1 and 3. In short, we expect the generalized similarity approach to be especially efficient in sparse-data settings, as it can incorporate the across-settings information, thereby imputing missing data.

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). The generalized similarity copes surprisingly well with missing data: even when 70% of the votes are missing, it can identify the two underlying groups (results not shown on graph). Moreover, the findings are robust for the introduction of error, as the two groups are identified when the noise is at 30%. Compared to the Pearson-correlation, the advantages of the generalized measure are even more apparent when the data are sparse; correlation fails to identify the groups even if only a small random disturbance is introduced. As we have seen, the generalized similarity model has proven more efficient than Pearson-correlation in settings with high amount of missing data.

Recovering the true social network

Next we investigate how the generalized similarity framework perform on one-mode social network data. We build a stochastic network formation model (Snijders et al., 2009). 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.

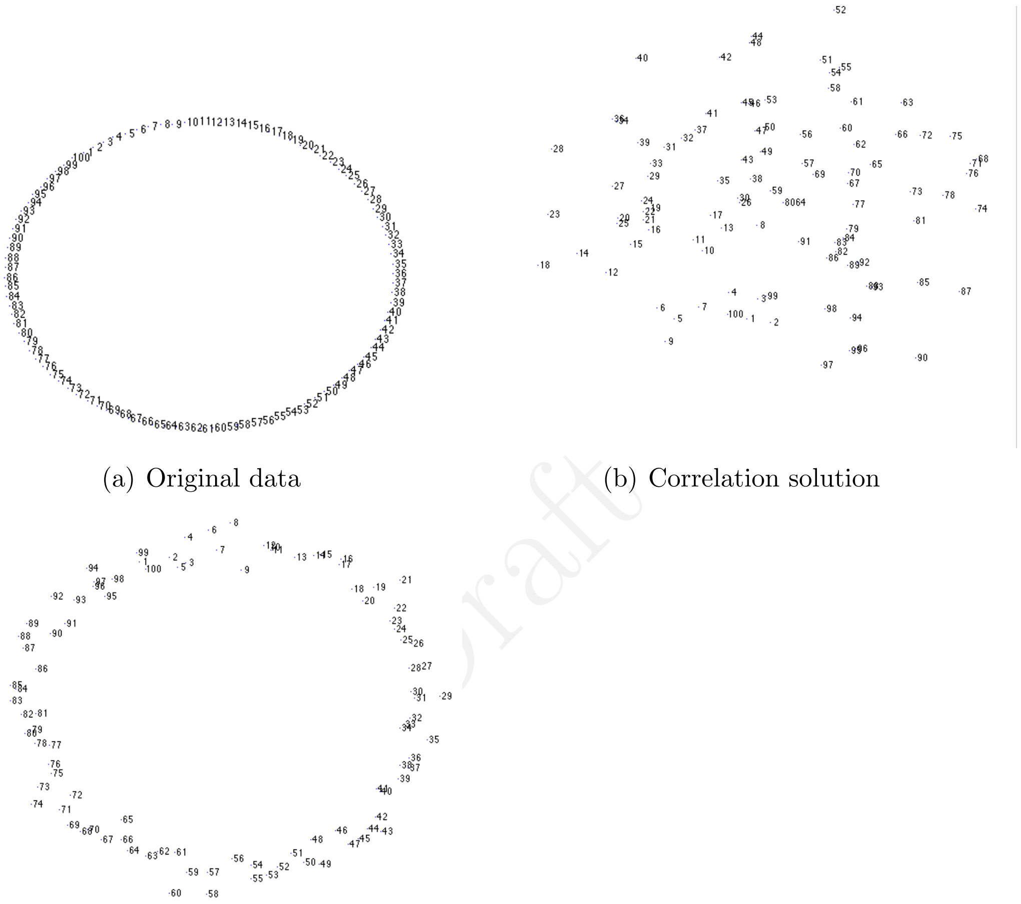
We modeled 100 individuals, indexed from 1 to 100. This number represents the individual's attribute along a dimension; we assume that the individuals are ordered along this dimension such that the ends of the distribution meet and, the closer two numbers are, the more similar the individuals. The similarity map of these individuals is thus a circle (shown in Figure 4a). We simulate random networks in which the tie creation rule is homophily (McPherson et al., 2001; Snijders et al., 2009): 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." Thus, the generalized similarity model is expected to provide a better description of the underlying data than Pearson-correlation.

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 define the probability of a tie between any two individual to be $P_{i,j} = \frac{1}{3 \cdot exp(|distance(i,j)/10|)}$. In this setup, the probability that an individual is connected to its closest neighbor is 30%, to its second closest neighbor is 27% etc. In the generated networks, 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).

The superiority of the generalized similarity model surfaces if we generate networks with fewer ties. For example, if we reduce the probability of ties to $P_{i,j} = \frac{1}{5 \cdot exp(|distance(i,j)/10|)}$, 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 Figures 4b and 4c). 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.

4 Two empirical illustrations

The previous simulations indicate that the generalized similarity model tends to emphasize the similarity within, and the dissimilarity between actors and settings. This leads to a crisper similarity map. Also, the generalized similarity model performs much better in inducing pairwise similarity for sparse data. Here we illustrate these properties of the model on empirical data. We analyze two datasets: the roll-call data of the U.S. Senate, and the



(a) Original data

(b) Correlation solution

(c) Generalized similarity solution

Figure 4: Comparison of the two-dimensional MDS solutions based on (a) real data, (b) Pearson-correlation, and (c) Generalized Similarity Model for a simulated social network in which people are located along a ring and they only have a few ties.

classic club-membership data of Davis et al. (1941).

Similarity of senators and issues

As we have seen in the previous section, the generalized similarity method 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. 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 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 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.

Roll call data is one of the most analyzed kinds of datasets of political scientists (e.g., Clinton et al., 2004). 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[4], and there were 644 issues on which there was no perfect consensus.[5]

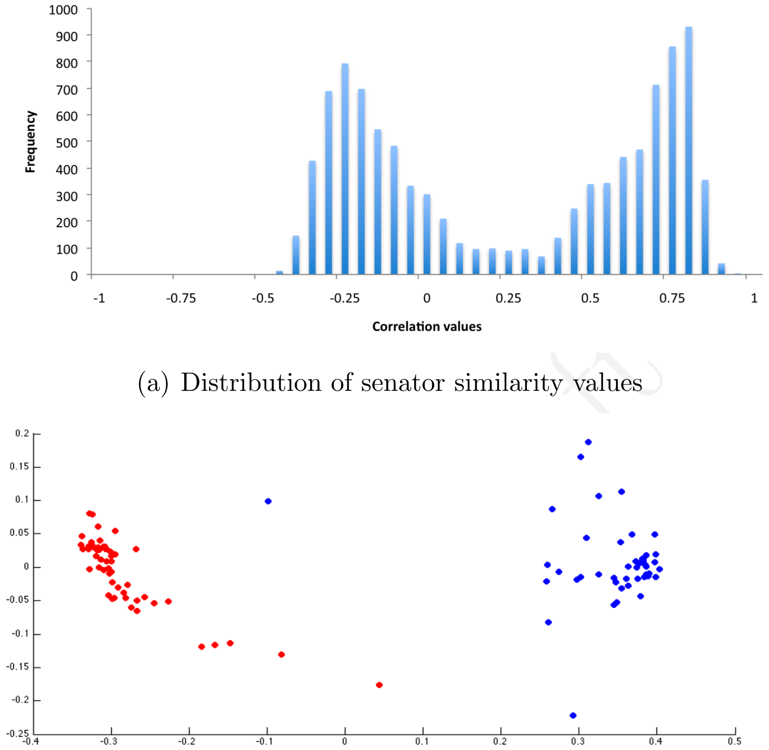
Thus, the $\underline{M}$ matrix, which contains the data, is a 101×644 matrix. We coded the "Yeah" vote with 1, the "Nay" with -1, and the "Not present" or "Abstain" with 0.

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. Figure 5 shows the distribution of the pairwise similarity values (5a), and the two-dimensional MDS map based on these similarity values (5b). The bimodal distribution

[4] Robert Menendez filled the seat of Jon Corzine in 2006, when the latter became the Governor of New Jersey.

[5] 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_teasers/votes.htm on April 5th, 2008.

in Figure 5a reflects the bipartisan nature of the Senate. Nonetheless, it indicates some overlap between the parties. The two-dimensional MDS map (5b) visualizes the pairwise similarities. 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.



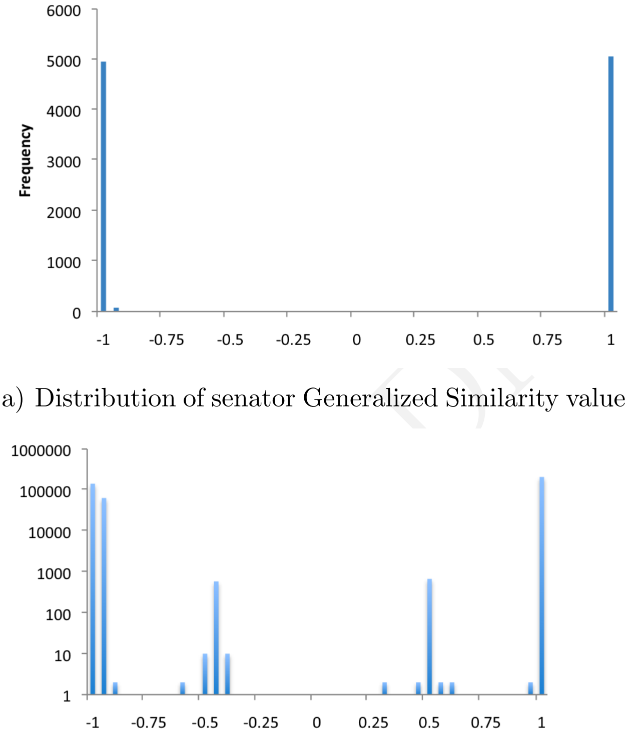
(a) Distribution of senator similarity values

(b) Senator similarity MDS map based on Correlation

Figure 5: 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 (Red dots denote Republican senators, Blue dots denote Democrats).

How do the results of the generalized similarity model differ from the results based on Pearson-correlation? Figure 6a 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. Indeed, the generalized similarity measure identifies a perfectly bipolar Senate.

The generalized similarity representation provides a similar clustering for issues. As Figure 6b shows, the issues are bipolar in nature as well, although less perfectly than for 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.



(a) Distribution of senator Generalized Similarity values

(b) Distribution of the similarity of issues

Figure 6: 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

Comparing senators who never voted together

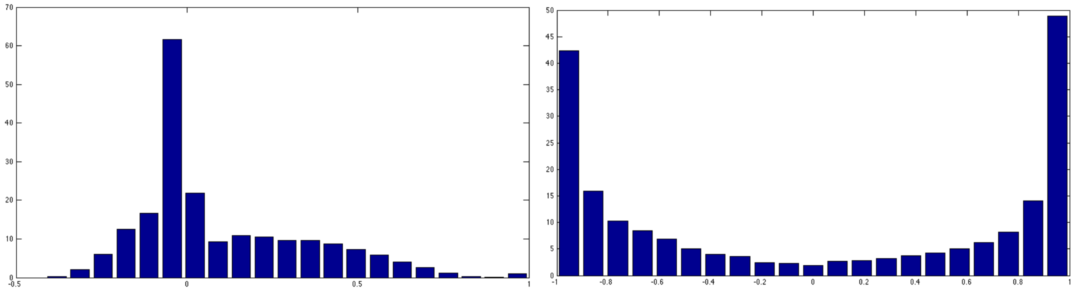
As pointed out in the Introduction, a major advantage of a generalized measure of similarity is its efficiency in dealing with data sparsity. The two datasets that we analyzed were relatively dense, in the sense that there was not much missing data. In order to model what will happen in cases where there is missing data, we expand the time-frame of the roll-call analysis. As our next step, we analyze the voting data of ten consecutive Senates: the 101th-110th Senates, serving during 1989-2008. These Senates have 202 senators altogether, who voted on 6,510 issues, therefore the resulting senator-issue vote matrix has $202 \times 6,510 = 1,315,020$ cells. No more than 100 senators can vote on any given issue, so 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. Figure 7 shows the distribution of the correlation (7a) and the generalized similarity (7b) 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. Using the correlation measure, the mode of the distribution is around zero, implying no relationship between the voting patterns of a given senator-pair. On the other hand, using the generalized similarity measure, the senate is revealed to be highly polarized. That is, most members are either highly similar or highly dissimilar, reflecting the underlying bipartisan structure.

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

Our first illustration uses Davis et al. (1941)'s data on the participation of 18 women in 14 social events. The original data are shown in Figure 1 (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,



(a) Distribution of the correlation values

(b) Distribution of the generalized similarity values

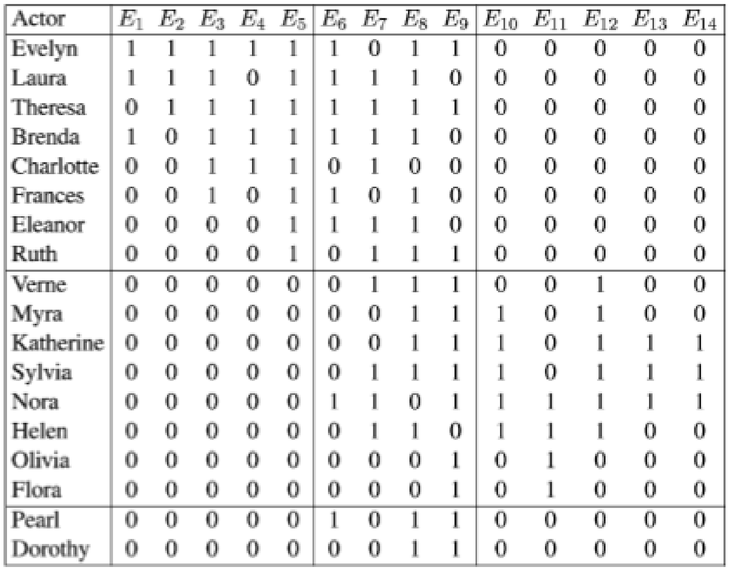
Figure 7: Comparison of the distribution of Pearson-correlation and generalized similarity for the 202 senators serving in the 101th-110th U.S. Senate

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 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 on Figure 1.

Here we analyze the Southern women social event participation data using the Generalized Similarity Model. Figure 9 shows the two-dimensional Multidimensional Scaling (MDS) maps based on both the Pearson-correlation similarity measure and the generalized similarity measure[6]. The MDS map based on correlation (Figure 9a) essentially recovers the blockmodel results, finding three subgroups. However, two women, Olivia and Flora are put together with Pearl and Dorothy, which is inconsistent with past analyses. Moreover, the three clusters are not clearly distinct.

[6] 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 $\mathit{dissimilarity} = (1 - \mathit{similarity})/2$.

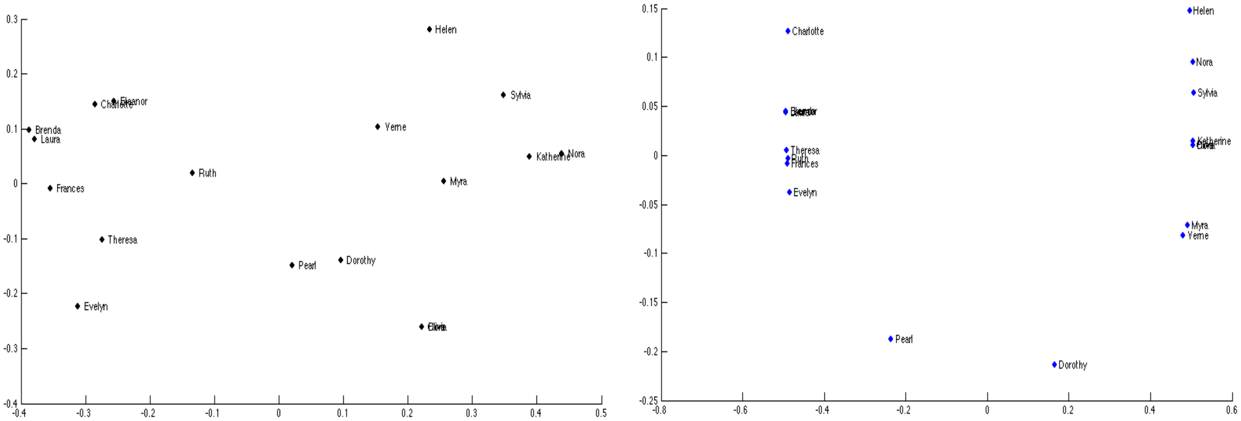


Actor	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}
Evelyn	1	1	1	1	1	1	0	1	1	0	0	0	0	0
Laura	1	1	1	0	1	1	1	1	0	0	0	0	0	0
Theresa	0	1	1	1	1	1	1	1	1	0	0	0	0	0
Brenda	1	0	1	1	1	1	1	1	0	0	0	0	0	0
Charlotte	0	0	1	1	1	0	1	0	0	0	0	0	0	0
Frances	0	0	1	0	1	1	0	1	0	0	0	0	0	0
Eleanor	0	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 8: The original Davis et al. (1941) data on the social event participation of 18 Southern women, with the blockmodel solution generated by Doreian et al. (2004).

As Figure 9b shows, the generalized similarity measure shows results similar to the analysis of Freeman (2003). Furthermore, it perfectly recovers the blockmodel solution of Doreian et al. (2004). Compared with the Pearson-correlation graph, the differences between the groups are strengthened and the groups are clearly distinct.

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 6, 7, and 8 do not fall into any group but stand separately.



(a) MDS based on correlation

(b) MDS based on Generalized Similarity Model

Figure 9: Comparison of the two-dimensional MDS maps of Pearson-correlation (a) and Generalized Similarity Model (b) for the Davis et al. (1941) data on the club membership of 18 women.

5 Comparison with other similarity models and clustering algorithms

In this section we compare the proposed generalized model of similarity to other major concepts in the literature on similarity and social positions. Specifically, we look at blockmodeling, CONCOR (Breiger et al., 1975), and Correspondence Analysis/Latent Semantic Analysis (Landauer and Dumais, 1997). Here we discuss the substantive and conceptual differences and similarities, and in Appendix 2 we illustrate how the solutions of these concepts differ on the simulated dataset of the social network presented in Section 3 of this paper.

Blockmodeling. Blockmodeling was developed in the 1970s to partition the nodes of networks (i.e., one-mode relational data) into clusters based on node positions (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 somewhat 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. It yields a more precise measure of similarity by incorporating across settings/actors information; and it provides a better measure of similarity and dissimilarity in sparse matrices. As 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), and because blockmodeling treats all 0s similarly, it underestimates the similarity of many of the actor pairs[7]. Therefore, it would identify a number of small local blocks, obscuring the real underlying data, which is much more structured.

CONCOR. CONCOR is a hierarchical clustering algorithm, introduced by Breiger et al. (1975). This algorithm has some resemblance to the generalized similarity model, at least at first glance: it uses iterated correlations to cluster the relational matrix into blocks. The resemblance to our model, however, ends there. First, the similarity of the columns in CONCOR 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 we propose. 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.

positional equivalence, automorphic equivalence Our model is somewhere halfway between the blockmodeling and role-equivalence (Borgatti and Everett, 1992) (Lorrain and White, 1971; Burt, 1976)

Singular Value Decomposition and Latent Semantic Analysis. Latent Semantic Analysis (LSA) is a high-dimensional linear associative model, which is used to induce the similarity of words in text corpora (Landauer and Dumais, 1997). The input of the LSA is usually text, more specifically a word-document matrix, which is essentially a two-mode relational dataset.

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

LSA uses local co-occurence data to induce global similarities of words. In short, it is an application of a well-know linear algebra procedure, Singular Value Decomposition (SVD), to linguistic data. SVD breaks down the original $m \times n$ word-document matrix into three matrices: a U matrix that contains the pairwise similarities of words, an S matrix that contains the eigenvalues of the word-document matrix, and a V matrix that contains the pairwise similarities of the documents.

SVD incorporates inter-setting information in calculating the similarity of actors. It also uses the inter-actor similarity information to calculate the similarity of documents. There are two major points on which our proposal 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 for why SVD is a valid way of representation). Second, LSA being a dimension reduction technique, it contains an element – the choice of the number of dimensions – which is often merely chosen to maximize the fit of the data.

6 Discussion, Applications and Further Work

In this paper we propose two principles for similarity data. First, we emphasize the need for taking the similarities among dimensions into account. As we demonstrate on a simple example of senators and issues, it is crucial that the original notion of co-locational similarity be extended. 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 similar if they are related to similar actors or objects similarly." Just to recall one of the main example of the paper: "Senators are similar if they vote similarly on similar issues." Second, we stress 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 (be it relational or attribute-based) stands for a dimension. These dimensions can be continuous, ordinal, or categorical; moreover, these dimensions need not be independent. 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.

This approach to data inherently changes how relations and attributes are perceived. By going against the common conception that "we know what the important dimensions for comparison of actors are", in our representation the structure of the dimensions (and their weighting) emerges from the structure of the data. This is, thus, a constructivist view. To prevent overfitting the dimensions to data, however, we introduced another concept: the stickiness of dimensions. With this concept, we want to capture situations in which the available data on which we do the analysis is only a subsample of all data, and thus we do not want to fully fit the similarity of dimensions to the available data.

The approach proposed in the paper is very general, and can be applied to a wide range of social and natural phenomena. First, it applies to all relational data. We have already mentioned a few applications in the paper. For one-mode relational data, such as social network data, our model directly applies. 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.

Note that applications do not have to come from the traditional domains of the network literature, but from other disciplines as well. 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 of 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 to both of these settings.

The proposed representation is especially efficient in analyzing sparse data because it heavily utilizes indirect similarities. Because of this property, the representation can be a promising framework for disciplines and settings in which sparsity is acute. These disciplines include computational linguistics, marketing, bibliometrics, or Web-analysis; but also settings from more traditional social sciences like consumption studies, social networks, demographics, or political science.

The generalized similarity representation 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 concertos tend to make other similar purchases, their similarity will be quite high.

The presented model is, of course, not without limitations, and these limitations call for further research on the framework. First, if possible, an analytical solution of the model would be needed. This we leave for scholars who are mathematically more gifted than we are. Second, as we emphasized earlier in the paper, the specific model of the generalized Pearson-correlation is just one of the possible frameworks for generalized similarity. It would be interesting to explore what other measures would satisfy the principles, and also to investigate how these principles can be combined with other approaches in the literature.

The third, possibly most severe limitation of the proposed model is that it builds on correlations among the dimensions, and correlations are not really good if the relationship between two dimensions is not linear. Such is the case for example between age and income. At this point, the model is not able capture this.

Finally, we would like to reiterate the main message of the paper.

Appendix 1: 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 voted 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 6)

Table 1: Case 1. 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

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 α denote the similarity of Issue 1 and Issue 2. Thus, the similarity matrix is:

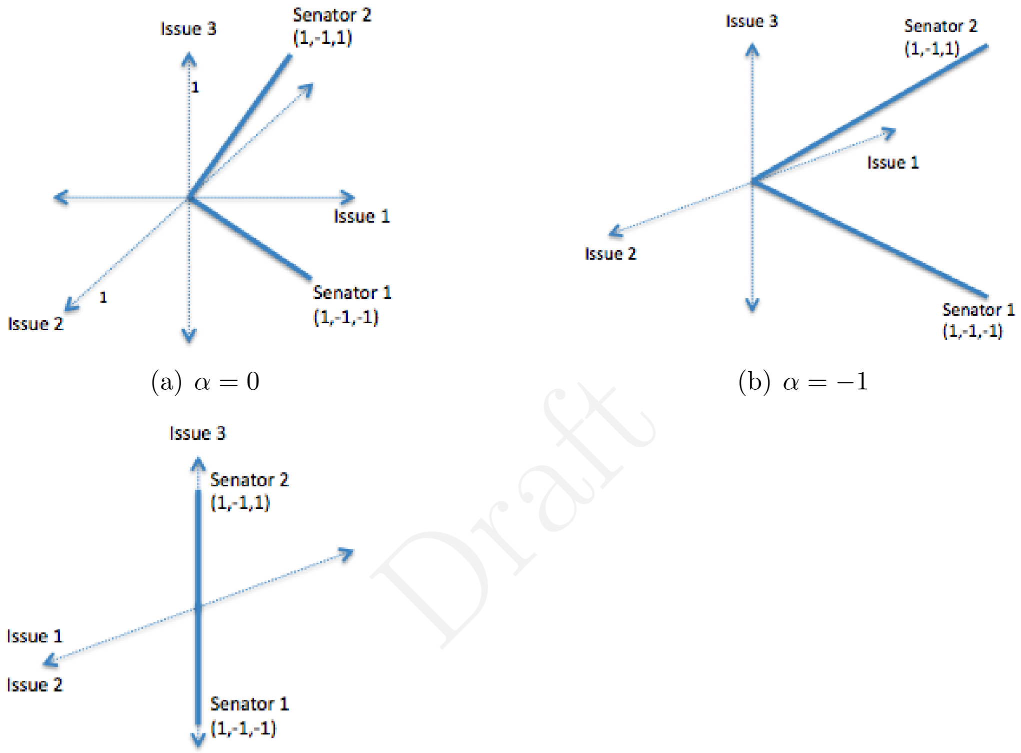
	Issue 1	Issue 2	Issue 3
Issue 1	1	α	0
Issue 2	α	1	0
Issue 3	0	0	1

Figure 11 illustrates through five cases how the similarity of senators 1 and 2 changes as a function of α. As the model is rather complex, we discuss each cases separately, and then we sum up the main implications of the modified correlation measure.

In Case 1., senator 1 votes "Yeah" on Issue 1, and "Nay" on Issues 2 and 3. senator 2 votes "Yeah" on Issues 1 and 3, and "Nay" on Issue 2. As one can see, if α is 0, then the similarity of senators 1 and 2 is 0.5, which corresponds to the Pearson-correlation value. Figure 11 shows that as α gets bigger, the similarity between senators 1 and 2 decreases. To explain this result, we discuss three scenarios: (1) when α is -1, (2) when α is 0, and when (3) when α is 1. The three scenarios are illustrated on Figure 5. When α is -1, Issues 1 and 2 are opposing. For example, a “Yeah” Issue 1 one means war, but a “Yeah” 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, s/he makes his/her position more strongly (a “Yeah” 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 “Yeah” on war and a “Nay” on another war). Putting these arguments together with the senators' vote on Issue 3 explains the negative effect of α 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 on Figure 10. As discussed in the “Principle 1: Taking the similarity among dimensions into account” section, the non-independence of the issues is modeled as a “warping” of the base space, and the generalized similarity measure is nothing else but the standardized version of cosine distance in this warped space. Figure 10a 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. Figures 10b and 10c 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 Figure 11, senator 1 votes “Yeah” on Issues 1 and 2, and “Nay” on Issue 3. senator 2 votes “Yeah” on Issue 1, and “Nay” on Issues 2 and 3. The similarity of senators 1 and 2 increases with α, but note that the similarity is always positive. When $\alpha = -1$, the similarity is weaker because the “Yeah” 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 “Yeah” on Issues 1 and 3, and “Nay” on Issue 2. senator 2 votes “Yeah” on Issues 1 and 2, and “Nay” on Issue 3.



(a) $\alpha = 0$

(b) $\alpha = -1$

(c) $\alpha = 1$

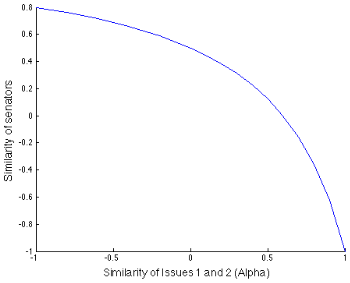
Figure 10: Geometrical illustration of the votes of the two senators in the three different issue-similarity settings. The dotted lines represent the issues, the solid line represents the votes. In the first setting, the issues are independent. In the second setting, Issues 1 and 2 are opposing. In the third setting, Issues 1 and 2 are the same.

The senators similarity decreases with the similarity of Issues 1 and 2 (α), but note that they are always dissimilar. Finally, Case 5, describes a voting scenario in which senators 1 and 2 vote “Yeah, Nay, Nay” and “Nay, Yeah, Nay” on the issues, respectively. When Issues 1 and 2 are dissimilar, the “Yeah” and “Nay” votes strengthen each other and make the senators rather dissimilar. However, when Issues 1 and 2 are similar, the “Yeah” and “Nay” votes cancel each other, so the similarity of the votes on Issue 3 dominates the dissimilarity on Issues 1 and 2.

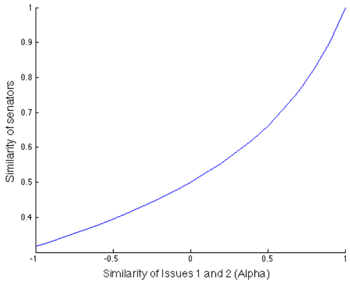
Appendix 2: Comparison of CONCOR and Correspondence Analysis with the generalized similarity model

[Here we shall compare the CONCOR and Correspondence Analysis solutions with the solution of the generalized similarity mode for the senator and the social network simulations of Section 3]

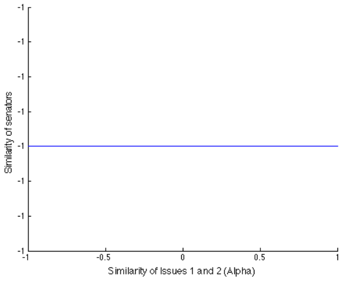
	Issue 1	Issue 2	Issue 3
Senator 1	1	-1	-1
Senator 2	1	-1	1



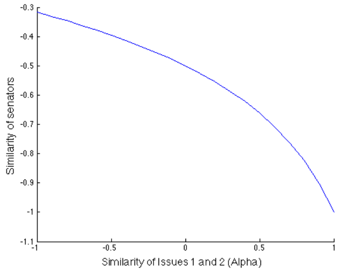
	Issue 1	Issue 2	Issue 3
Senator 1	1	1	-1
Senator 2	1	-1	-1



	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

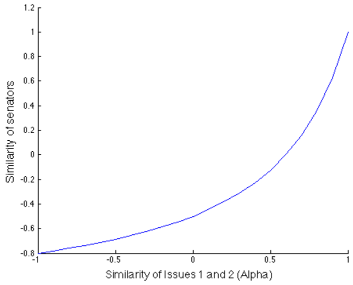


Figure 11: Five voting scenarios for two senators on three issues.

References

Bailey, Kenneth D. 1994. *Typologies and taxonomies: An introduction to classification techniques*. SAGE Publications.

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.

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. "Cumulated Social Roles: The Duality of Persons and Their Algebras." *Social Networks* 8:215–256.

Burt, Ronald S. 1976. "Positions in networks." *Social Forces* 55:93–122.

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*. Chicago: University of Chicago Press.

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.

Einstein, Albert. 1916. "Die Grundlage der allgemeinen Relativitätstheorie." *Annalen der Physik* 49:2.

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.

Gärdenfors, P. 2004. *Conceptual spaces: The geometry of thought*. The MIT Press.

Garfield, Eugene. 1972. "Citation analysis as a tool in journal evaluation." *Science* 178:471–479.

Hume, David. 2004 (1748). *An Enquiry Concerning Human Understanding*. Mineola, New York: Dover.

Kovács, Balázs. 2009. "Relational Similarity as an Indicator for Attribute-based Similarity." Unpublished manuscript, Stanford University.

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.

Lattin, James M., J. Douglas Carroll, and Paul E. Green. 2002. *Analyzing Multivariate Data*. Duxbury.

Lorrain, Francois P. and Harrison C. White. 1971. "Structural Equivalence of Individuals in Networks." *Journal of Mathematical Sociology* 1:49–80.

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

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–444.

Medin, Douglas L. 2005. "Concepts and conceptual structure." *Social Cognition: Key Readings* pp. 115–129.

Murphy, Gregory L. 2002. *The Big Book of Concepts*. MIT Press.

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

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

Shepard, Roger N. 1962. “The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function.” *Psychometrika* 27:125–140,219–246.

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

Snijders, Tom A.B., Gerhard G. van de Bunt, and Christian E.G. Steglich. 2009. “Introduction to stochastic actor-based models for network dynamics.” *Social Networks* 000:000–000.

White, Douglas R. and Karl P. Reitz. 1983. “Graph and semigroup homomorphisms on networks of relations.” *Social Networks* 5:143–234.

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.

Winship, Christopher. 1988. “Thoughts about roles and relations: An old document revisited.” *Social Networks* 10:209–231.