Klippel, A., Yang, J., Wallgrün, J. O., Dylla, F., & Li, R. (2012). Assessing similarities of qualitative spatio-temporal relations. In C. Stachniss, K. Schill, & D. H. Uttal (Eds.), *Spatial Cognition 2012* (pp. 242–261). Berlin: Springer.

Assessing Similarities of Qualitative Spatio-Temporal Relations[1]

Alexander Klippel, Jinlong Yang, Jan Oliver Wallgrün, Frank Dylla*, Rui Li

Department of Geography, GeoVISTA Center, The Pennsylvania State University, PA, USA
[klippel,jinlong,wallgrun,rui.li]@psu.edu
*SFB/TR 8 Spatial Cognition, Universität Bremen, Germany
dylla@sfbtr8.uni-bremen.de

Abstract. In this article we analyze behavioral data to advance knowledge on how to assess similarities of events and spatial relations characterized by qualitative spatial calculi. We have collected a large amount of behavioral data evaluating topological relations specified in the Region Connection Calculus and Intersection Models. Several suggestions have been made in the literature on how to use associated conceptual neighborhood graphs to assess the similarities between events and static spatial relations specified within these frameworks. However, to the best of our knowledge, there are few (to none) approaches that use behavioral data to formally assess similarities. This article is contributing to this endeavor of using behavioral data as a basis for similarities (and associated weights) by (a) discussing a number of approaches that allow for transforming behavioral data into numeric values; (b) applying these approaches to nine data sets we collected in the last couple of years on conceptualizing spatio-temporal information using RCC/IM as a baseline; and (c) discussing potential weighting schemes but also revealing essential avenues for future research.

1. Introduction and Background

Every calculus with jointly exhaustive and pairwise disjoint (JEPD) relations (such as RCC and IM) has a conceptual neighborhood graph (Cohn & Renz, 2008).

To navigate through daily life, humans use their ability to conceptualize spatio-temporal information, which ultimately leads to a system of categories. Likewise, the disciplines of the spatial sciences focus on conceptualization and categorization as a means to structure spatio-temporal information. Although challenged by several researchers, similarity is one of the most important and most commonly used tools to aid in the process of conceptualization and categorization in both artificial and natural cognitive systems (Bruns & Egenhofer, 1996; Goldstone & Son, 2005; Nedas & Egenhofer, 2008; Rissland, 2006; Schwering, 2008; Tversky, 1977). In the spatial sciences and in related branches of artificial intelligence, an approach has been devel-

[1] This research is funded by the National Science Foundation (#0924534) and Deutsche Forschungsgemeinschaft (DFG) grant SFB/TR 8 Spatial Cognition.

oped that allows the specification of similarity measures for spatio-temporal data: qualitative spatio-temporal representation and reasoning (QSTR). Calculi developed in the general area of QSTR allow for meaningful processing of spatio-temporal information because they focus on categorical (discrete) changes or salient discontinuities (Galton, 2000), which are thought to be relevant to an information processing system (both human and artificial). While qualitative calculi are naturally appealing and are, on a general level, widely acknowledged in the cognitive sciences, too[2], there is comparatively little behavioral assessment of the cognitive adequacy of these calculi. This is an astonishing fact given that these calculi are often intended to improve processes at the human-machine interface and are on several occasions claimed to be cognitively adequate (Clementini, Di Felice, & van Oosterom, 1993; Knauff, Rauh, & Renz, 1997; Knauff, 1999). However, in our opinion, the systematic behavioral evaluation of QSTR is an essential missing piece that will lead to refined and improved QSTR models and in return significantly increase their value and their usability in numerous applications (e.g., information retrieval, spatial query languages, formalizing the semantics of spatial language).

This paper will discuss a framework for defining cognitively adequate similarities/weights by detailing strategies to transform results from behavioral experiments on how humans conceptualize spatio-temporal information into both qualitative (category-based) and quantitative similarity measures. These measures are tailored towards formal theories in the spatial sciences and will be applicable to theories of spatio-temporal representation and reasoning. Hence, we will contribute to the formal basis of the semantics of spatio-temporal information.

To further motivate the general questions we are addressing in this paper, consider the spatial scenes in the right part of Figure 1. The scenes show the development of an oil spill in relation to an island. We focus on relations distinguished by prominent topological calculi commonly used in spatial information theory and in the cognitive sciences to identify potentially important aspects of spatio-temporal information. The individual icons reflect distinctions made by the Region Connection Calculus (RCC, Randell, Cui, & Cohn, 1992) as well as Egenhofer's Intersection Models (IM, Egenhofer & Franzosa, 1991). An important aspect for assessing the similarity of these scenes as well as modeling spatio-temporal information is that these relations can be organized to form a so called conceptual neighborhood graph (CNG, Freksa, 1992, left part of Figure 1). Two relations, R_1 and R_2 are conceptual neighbors if it is possible for R_1 to hold over a tuple of objects at a certain point in time, and for R_2 to hold over the tuple at a later time, with no other (third) mutually exclusive relation holding in between (Cohn, 2008). A neighborhood graph has one node for each relation $R \in \boldsymbol{R}$, and an edge between two nodes if the corresponding relations are conceptual neighbors. The important aspect to keep in mind, which adds to the transformative nature of this paper, is that virtually every calculus with jointly exhaustive and pairwise disjoint (JEPD) relations (such as RCC and IM) has a conceptual neighborhood graph (Cohn & Renz, 2008), and that, hence, the methods proposed here are universally applicable amongst all such calculi.

[2] Lewin, 1936/1966; Piaget & Inhelder, 1948/56/67; Lu, Harter, & Graesser, 2009.

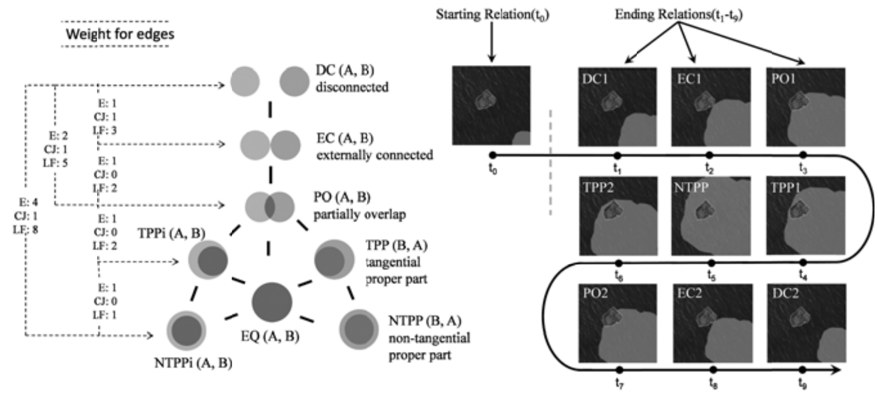

Fig. 1. The left side shows a conceptual neighborhood graph based on RCC-8 and IM. The dotted lines reflect the discussion in the text showing similarity assessments for three models exemplarily (E: equal weights; CJ: Camara and Jungert 2007; LF: Li and Fonseca 2006). On the right side the development of an oil spill is depicted in relation to an island. Each of the scenes could either be a transition or an ending relation.

One of the main characteristics of the relations displayed in Figure 1, but also relations from all other JEPD qualitative calculi, is that we can measure how similar the scenes (or their genesis) on the right side are by employing the conceptual neighborhood graph on the left side. Organizing spatial relations in this graph-like format has the advantage that graph theoretical measures can be applied to determine the similarity between these scenes which is essential for numerous information retrieval and formal semantic tasks (Bruns & Egenhofer, 1996; Papadias & Delis, 1997; Wallgrün, Wolter, & Richter, 2010). As both the spatial and the cognitive sciences focus on qualitative distinctions made, for example, by topology (Galton, 2000; Johnson, 1987; Klix, 1971), we have found—theoretically—a bridge between formal and cognitive spatial semantics. To demonstrate this aspect, we will focus on the similarity of four relations in Figure 1, DC (t1), EC (t2), PO (t3), and NTPP (t5): The simplest approach (Bruns & Egenhofer, 1996; Dylla & Wallgrün, 2007; Rada, Mili, Bicknell, & Blettner, 1989; Schwering, 2007), in a nutshell, assigns all edges in the CNG an equal weight of 1 and similarity/dissimilarity is established by counting the number of edges between two relations. Hence, the dissimilarity (weight) between DC and EC would be 1, the dissimilarity between DC and PO would be 2, and the dissimilarity between DC and NTPP would be 4.

This rather simplistic view has, of course, been challenged and several researchers have proposed (introspectively) alternative weighting schemes for CNGs. For example, Camara and Jungert (2007), in seeking to define a query language for dynamic processes, suggested a grouping of topological relations that are the basis of the CNG in Figure 1 into DC (disconnected) on the one hand and all other relations on the other. If we apply this strategy, dissimilarity between DC and all other relations would be 1, while the dissimilarity among all other relations would be basically 0. Another approach by Li and Fonseca (2006) takes into account that there may be different weights between different conceptual neighbors. They assign, for example, a weight of 3 to the edge between DC and EC, and a weight of 2 to the edge between EC and PO, a weight of 1 to the edge between TPP and NTPP. Hence the three dissimilarities from the example would be: DC-EC - 3, DC-PO - 5, and DC-NTPP - 8.

Many other approaches have been discussed. Consider the issue of the level of granularity which determines the number of basic relations that are assumed. The immediately relevant distinction here is between RCC-5 and RCC-8. While RCC-8 can be mapped onto eight relations distinguished by Egenhofer's intersection models (IM), RCC-5 cannot be directly mapped onto the coarser level of the IM (cf. Knauff et al., 1997; Renz, 2002) due to the ontologically different status of the boundaries. Hence, depending on the model we apply, the similarities will change, too (see also Clementini et al., 1993 for a different 5 relation solution, 4 plus 1 to be precise).

This simple example and associated literature demonstrate a number of important issues:

- Formal calculi using JEPD are omnipresent in research in the spatial sciences and are both theories as well as integral parts of spatial information systems. They are vital to various applications for spatial representation and reasoning, and are used frequently to establish similarities especially to aid human-computer interaction.
- There is some arbitrariness in designing and using these approaches guided by both formal constraints and requirements arising from a specific formalism (e.g., RCC versus IM) or by the introspection of a researcher. This lack of guidance as to which approach to use has been identified as a major obstacle in the usefulness of QSTR (Schultz, Amor, & Guesgen, 2011).
- There are few behavioral approaches that have evaluated QSTR[3]. However, to the best of our knowledge, except for our own work (Klippel, accepted), there is no behavioral research that addresses the possibility that similarities (as an expression of cognitive conceptualization processes) between qualitative spatial relations may change depending on the semantics of a specific domain[4]. For instance, one big question is: What happens to similarities of relations in the example in Figure 1, if we use different domains such as such as a lake and a house or a hurricane and a peninsula? Do we expect to be able to use the same similarities (weights) between relations?

We strongly believe that similarity measures should not be designed introspectively. As these measures are often intended to improve the interface of humans and computers/machines, it is essential to ground the assessment of similarities in behavioral research.

The remainder of this article is structured as follows: In Section 2, we will provide a very short overview of the behavioral data that we have collected over the last couple of years and that we will reanalyze here to discuss weights for CNGs; Section 3

[3] For an overview of research evaluating QSTR see Klippel, Li, Yang, Hardisty, & Xu, in press; Mark, 1999.

[4] Mark and Egenhofer 1995 have speculated that this might be the case.

discusses different methods that potentially allow for establishing weights; Section 4 provides a summary and lays out ideas for future research efforts.

We will be using the following abbreviations: CNG: Conceptual neighborhood graph, CN: Conceptual neighbor; TEC: Topological equivalence class; OSM: Overall similarity matrix; QSTR: Qualitative Spatio-temporal Representation and Reasoning; RCC: region connections calculus; IM: Intersection models. We will also use RCC terminology for topological relations: DC: disconnected, EC: externally connected, PO: partial overlap, TPP: tangential proper part, NTPP: non-tangential proper part.

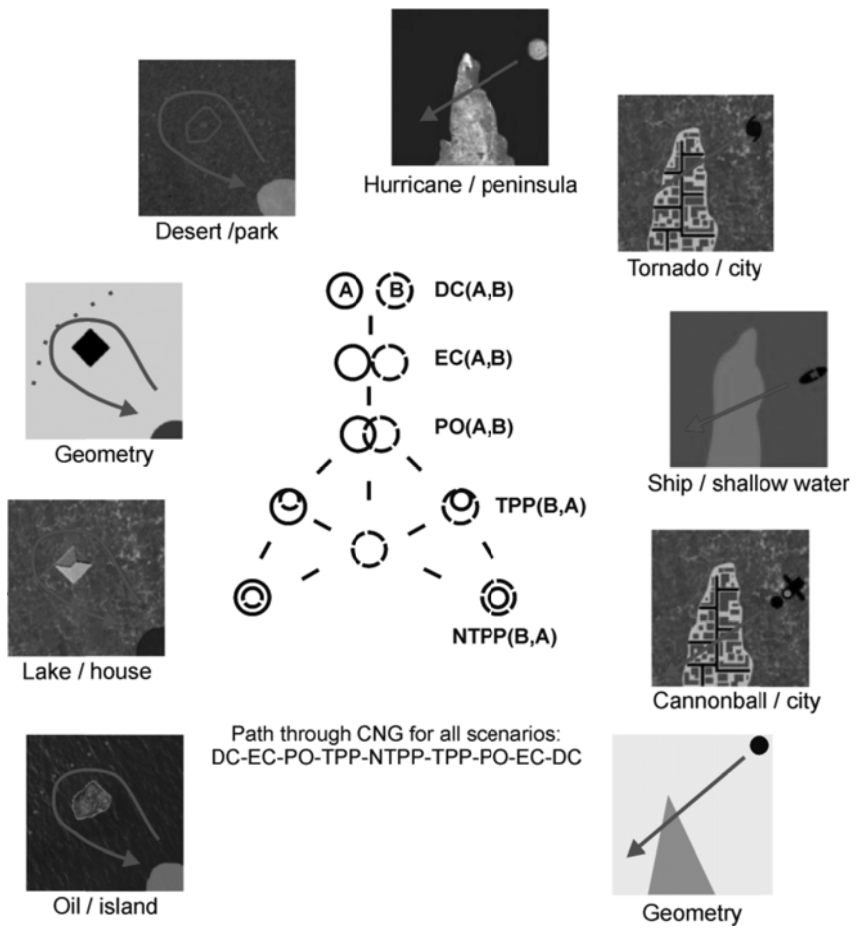

Fig. 2. Example of the nine scenarios that we reanalyze in this paper to derive weights for conceptual neighborhood graphs (translation: geometric figures, hurricane/peninsula, tornado/city, ship/shallow water, cannonball/city; scaling: geometric figures, oil spill/island, house/lake, desert/recreation park).

2. Data Collection

In this article, we reanalyze data we collected investigating cognitive conceptualizations of earth dynamics. We have designed nine experiments (for more details see Klippel, accepted and Yang, Klippel, & Li, submitted) using two different types of dynamics—movement patterns that can be considered translations and movement patterns characterized as scaling. Additionally, we used different semantic domains with different entities (translation: geometric figures, hurricane/peninsula, tornado/city, ship/shallow water, cannonball/city; scaling: geometric figures, oil spill/island, house/lake, desert/recreation park). Figure 2 gives a general idea of how the animated icons were designed for the different domains, while Figure 1 already showed static snapshots of the actual animations for the oil spill/island domain. The important aspect that makes all nine experiments comparable is that the stimuli used in each experiment are identical from a topological perspective (using either RCC-8 or IM). In each of the nine experiments, animations are designed such that nine different yet topologically equivalent movement patterns can be distinguished. The main distinguishing criterion is borrowed from cognitive theories on event conceptualization (Regier & Zheng, 2007), that is, patterns are separated based on the topological relation they can end in (see Figure 1). All movement patterns start in the DC relation and could end in one of the nine possible ending relations depicted in Figure 1. Within each topologically identical pattern we realized eight instances. This means that for each of the nine experiments (semantic domains), 72 animations were created: eight animations/instances each for nine topologically equivalent movement patterns.

In our experiments we employed a grouping paradigm, which is classically used to elicit conceptual knowledge. Participants (N = 20 in each of the nine experiments) have the task to sort the animated icons into groups with larger within-group than between-group similarities. The task can be characterized as *free classification* (Pothos, 2005) or *category construction* (Medin, Wattenmaker, & Hampson, 1987), meaning that participants created all their groups from scratch without any limitations regarding the number of groups or which icons should be placed together. All 72 animations have to be sorted into groups before the experiments were considered complete. The grouping behavior for each participant is recorded in a similarity matrix in binary form: two icons that are placed into the same group are coded as 1; two icons in different groups are coded as 0. All nine scenarios have 72 animated icons such that each matrix for each participant has 5194 cells of which 2556 are meaningful (others are redundant or encode the relation of an icon with itself). Summing over all individual similarity matrices within each domain nine overall similarity matrices (OSMs) are created. OSMs encode overall similarity assessment between icons in the following way: The highest possible similarity corresponds to N, the number of participants. For example, if all 20 participants placed a certain pair of animated icons together into a group, 20 individual ‘1’s are added up. In contrast, if a pair of two animated icons is never placed into the same group, their similarity is recorded as ‘0’.

As all nine domains are topologically identical, we have obtained a large number of similarity ratings for pairs (conceptual neighbors) of topological relations. Each topological equivalence class had eight instances in all nine domains assessed by 20 participants in each experiment. Hence, we have a total of 8 times 9 times 20 (= 1440) similarity assessments for each topological relation combination (for CNs but also for all other possible combinations). To give the reader a first impression of how these similarities are distributed across the nine different domains and across TECs, Figure 3 is visualizing the raw similarities from the OSMs in so-called heat maps. We reduced the size but the overall patterns reveal that there are potentially interesting differences across domains.

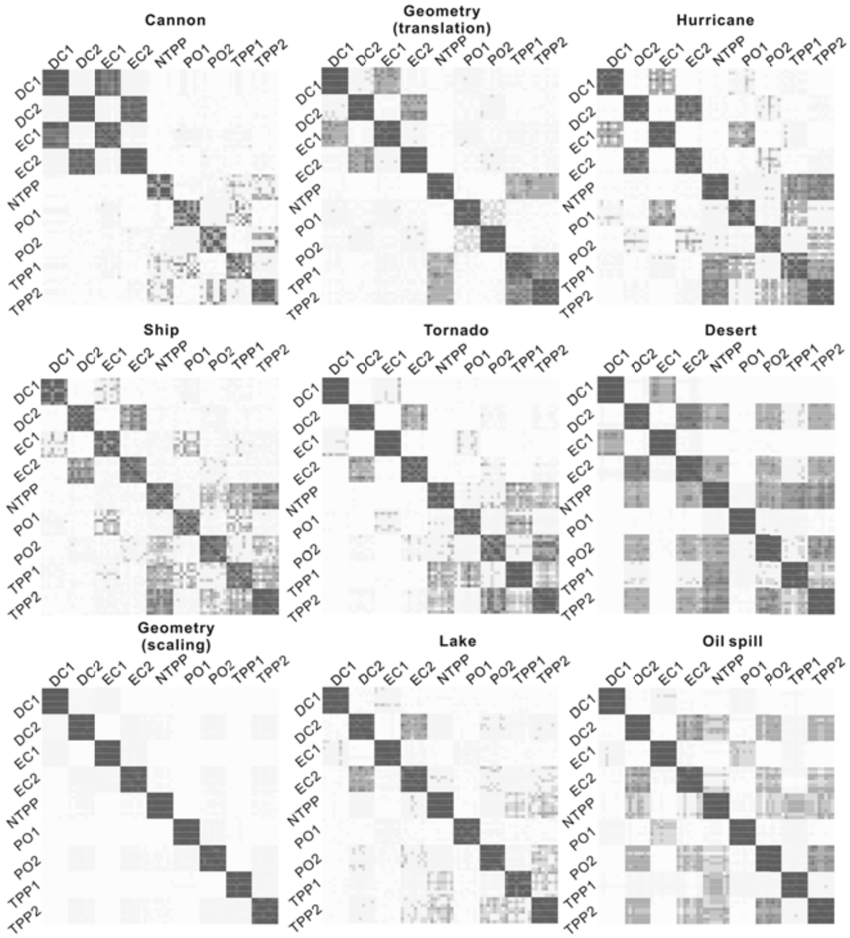

Fig. 3. Heat maps visualizing the raw similarities of all nine experiments/domains. Each heat map shows icons and TECs in the same order to allow direct comparison. Only labels for TECs are provided (not individual animations). Dark gray colors indicate high similarities, light gray to white colors indicate low or 'no' similarities (a color version of this figure can be found at min.us/m_sc2012figure3 for better readability).

3. Tailoring the Cognitive Adequacy of QSTR

In this section we are discussing how behavioral data (see Section 2) can be used to derive similarities/weights for conceptual neighborhood graphs (and potentially for pairs of topological relations that are not conceptual neighbors). We will be discussing several methods such as normalizing raw data, cluster analysis, and cluster validation techniques. In addition to using raw data, cluster analysis is chosen as it is the most common method to analyze grouping data and is thought to reveal natural groupings.

3.1 Raw Similarities

Raw similarities have been briefly introduced in Section 2. Additionally, Figure 3 provides an overview of how raw similarities are distributed within the OSMs of all nine experiments/domains. For the purpose of using raw similarities as a possibility for assessing weights of edges in conceptual neighborhood graphs (as well as, potentially, for relations that are formally not conceptual neighbors) several adjustments/standardizations have to be performed. While the behavioral data characteristics of our data allow for straightforward comparisons given that each experiment/domain had the same number of participants and the same number of icons per TEC, we will discuss normalization approaches for the purpose of creating a universal method for deriving weights on the basis of raw similarities. Raw similarities have the advantage that they can be employed not only for conceptual neighbored TECs but for all pairs of TECs.

The OSMs depicted in Figure 3 contain a large amount of redundant information. As we are concerned here with deriving weights for pairs of TECs—primarily for neighbored TECs in a CNG but the same approach can be applied to any pair of TECS—we can simply focus on the $k \times k$ sized submatrix ($k = 8$ in our case) consisting of all rows corresponding to the first TEC and all columns corresponding to the second TEC. Using $CN\_inst_{i,j}$ for the entries of this matrix, the raw similarity of the two CNs is simply computed as:

$$RSim_{CN} = \sum_{i=1}^{k} \sum_{j=1}^{k} CN\_inst_{i,j}$$

Once the raw similarities for each combination of TECs have been extracted, they have to be normalized to adjust for the specifics of the experimental setup, i.e., the number of instances in each TEC and the number of participants in each experiment. The obtained data can be normalized using row standardization taking into account all values such that individual values will be between 0 and 1. As a first step, we will look into conceptual neighbors only. Raw similarities of all CNs can be normalized to $NRSim_{CN_i}$ in the following way:

$$NRSim_{CN_i} = \frac{RSim_{CN_i}}{\sum RSim_{CN_j}}$$

3.2 Fusion Coefficients

Cluster analysis has the goal to identify natural groupings of entities (e.g., animated icons) and is frequently used in a number of disciplines (Everitt, 2001; Romesburg, 2004). There is not one specific algorithm but rather a family of cluster algorithms. In hierarchical cluster analysis, entities are stepwise assigned to groups based on similarities which are coded in similarity/proximity matrices (here: the nine OSMs). For most cluster analyses, the similarity matrix is recalculated after each clustering step, reflecting the existence of new groups. Cluster algorithms differ with respect to the way similarities are re-calculated after each grouping step (for an overview see Everitt, 2001; Romesburg, 2004). The first step in each clustering process is to combine those entities into groups that have the highest similarities. Similarities (or dissimilarities, respectively) can be considered distances and as such they are used to create so called dendrograms that reflect the clustering process (see Figure 4). Dendrograms provide an indication when two entities or groups of entities are fused (grouped) together and the distance at which they are fused is referred to as a *fusion coefficient*. All fusion coefficients are stored in a so-called *cophenetic matrix*. As similarities/dissimilarities are differently calculated by different clustering methods, there is a different cophenetic matrix for every clustering method.

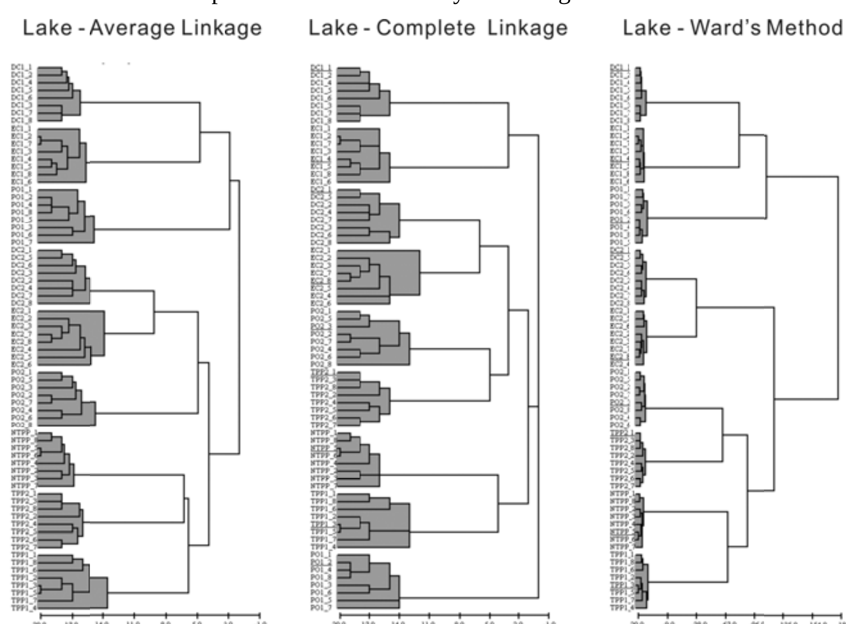

Fig. 4. Displayed are three dendrograms which reflect the clustering process for three different clustering methods (Ward's, average linkage, and complete linkage) for the same experiment/domain (lake/house). The dendrograms are visual representations of fusion coefficients indicating the value (distance) at which individual clusters are merged.

For the purpose of deriving similarities/weights for CNs on the basis of fusion coefficients, we have to briefly discuss two prerequisites. First, as different clustering methods will have different fusion coefficients (see Figure 4), we follow advice from the cluster validation literature (Ketchen & Hult, 2000; Kos & Psenicka, 2000; Milligan, 1996) and compare different methods, here: Ward's method, average linkage, and complete linkage. Second, while topology is overall a strong grouping criterion, there are situations in which TECs are indistinguishably merged (a potential indication of high similarity), individual instances of TECs might have ended up in a group with instances of a different TEC, or in the worst case, instances of certain TECs are spread across several groups.

Figure 5 shows the results for cases in which it is possible to read out one fusion coefficient as a measure of how similar two TECs are and use this as a weight for CNs. To make the data comparable, we normalized the fusion coefficients in the same way as the raw similarities (see Section 3.1).

3.3 Cluster Validation Techniques

The formal characterization based on topological equivalence classes allows for specifying a theoretical partition of the animated icons, P. This is an ideal scenario as we can employ cluster validation methods to assess whether the clustering structure C created by participants matches the theoretical partition established through topological equivalence or the CNG (Halkidi, Batistakis, & Vazirgiannis, 2002b, Halkidi, Batistakis, & Vazirgiannis, 2002a). One way of comparing C and P is to calculate indices such as Rand Statistics, Jaccard Coefficient, and the Folkes and Mallows index.

These indices build on the following information: Let $C = \{C_1, \dots, C_m\}$ be the clustering structure that results from analyzing the grouping behavior of the participants recorded either in individual similarity matrices or the OSM. Let $P = \{P_1, \dots, P_n\}$ be the partition of the stimulus (animated icons) that is based on formal requirements (such as the differences between RCC-8 and RCC-5) or some introspective assumptions made by a researcher (e.g., Li & Fonseca, 2006). To be able to compare the formally derived partitioning P with the obtained results C, the following numbers are computed by comparing the containing clusters for each pair of animated icons (x_v, x_u):

- SS-a: the number of pairs of animated icons that belong to the same cluster in both, the clustering structure C and the partition P.
- SD-b: the number of pairs that belong to the same cluster in C but to different clusters in P.
- DS-c: the number of pairs that belong to the same cluster in P but to different clusters in C.
- DD-d: the number of pairs that belong to different clusters in both C and P.

The numbers for a, b, c, and d add up to the number of pairs of animated icons M. The Jaccard coefficient J, for example, is then calculated as $J = \frac{\mathrm{a}}{\mathrm{a+b+c}}$ and provides a similarity measure for comparing C and P.

To derive actual weights for the conceptual neighborhood graph based on these indices, we adapt this general approach and compute individual indices for two conceptually neighbored TECs R1 and R2 in the following way: We focus on only those icons that belong to either R1 or R2. We then consider the individual grouping of a participant and reduce it to just these icons. The resulting clustering is used for C and compared to a clustering P in which all icons from R1 and R2 are grouped into a single cluster. This means we compare the groupings of the participants to a grouping in which TECs R1 and R2 are completely combined. The values a, b, c, and d as well as the indices are then computed as described above and averaged over all participants. Using this approach, the Rand and Jaccard indices will always be the same because there is only one cluster in P and, hence, *b and d* are always zero. It also has to be noted that in the case that a is also zero (which means that the icons from R1 and R2 form two disjoint groups in C), we consider the Folkes and Mallows index to be zero (maximally dissimilar), while it is not defined in the original definition.

3.4 Comparing Similarity/Weighting Approaches

In the following, we will compare different strategies to derive similarities/weights for conceptual neighborhood graphs applying the methods discussed above to give an overview of potential weights through the perspective of similarities. The raw similarities of the nine different experiments/domains that we re-analyzed were already shown in Figure 3. Similarities are visualized as heat maps: dark gray colors indicate high similarities; light gray to white colors indicate low or no similarities. Columns and rows are organized by TECs with eight instances (animated icons) within each TEC. The order of columns and rows is kept constant (i.e., in alphabetical order) such that the heat maps are directly comparable. From the dark gray colors along the diagonals (top left to bottom right) we can infer that for most experiments/domains, the similarities within a TEC are very high. This indicates that topology is a strong grouping criterion. There are some exceptions that we will discuss in the following (e.g., the proper part relations in the cannon scenario). We also find that other TECs form strong conceptual groups, but that these similarities are susceptible to change across different scenarios.

In addition to the visualization of the raw similarities in Figure 3, Figure 5 shows the normalized weights derived by analyzing the behavioral data using the methods discussed in Sections 3.1 to 3.3.

We ran a correlation analysis over all index combinations for all nine scenarios and, as indicated by the graphs in the figures, found higher (partially near perfect) correlations between raw similarities and validation indices and slightly lower yet high correlations between fusion coefficients and raw similarities. For the time being we are only looking at fusion coefficients for CN TECs, not for individual icons as the goal of cluster analysis is to identify natural groupings.

We can make several observations:

- Fusion coefficients are not specified for all CNs. There are two reasons for this: First, although topology overall is a strong basic grouping criterion, there are some exceptions in which instances of a TEC are split and are not members of the same group. The consequence is that fusion coefficients cannot be specified between TECs. Fortunately, these cases are rare. Second, in the case that two TECs are merged together to an extent that they become indistinguishable, specifying a fusion coefficient does not make sense. In these cases, it would be most appropriate to use dissimilarities and define dissimilarities of indistinguishably merged TECs as being minimal (e.g., '0'). Specifying an exact value for similarity is more involved and for the time being, we left the value unspecified; it should be the highest possible similarity.
- The fusion coefficients deliver more pronounced graphs, compared to, for example, the raw similarities (hence the lower correlation coefficients). This is good news and bad news. On the one side, this is the intention of clustering methods, that is, strengthening within group similarities and pronouncing between group differences. On the other hand, individual clustering methods may introduce biases. While the details on how clustering methods create groups are known (Aldenderfer & Blashfield, 1984; Romesburg, 2004), it is not necessarily transparent how this plays out in a specific calculation (e.g., why they result in differences in one experiment/domain but not in another). We found that particularly complete linkage is behaving differently than other methods: (a) weights for CNs are more often not defined (see discussion above); (b) the behavior of the graphs is sometimes contrary to graphs of other methods (e.g., EC1-PO1 in the hurricane and ship scenario).
- One important observation that we will pick up in the outlook again is that there are substantial and significant (see also Klippel, accepted) differences between the similarities across the nine different scenarios. This is exemplified by the different shapes of the graphs across the different scenarios (see Figure 5). As the scenarios are topologically identical and differences such as metric information and speed have been minimized in the experimental setup, we have to conclude that weights between CNs are not independent of the semantics of a domain.
- One additional aspect that makes the assignment of weights difficult is that contextual factors play a role. In the case of the experimental data that we reanalyzed, all paths through the CNG were identical and symmetric: DC-EC-PO-TPP-NTPP-TPP-PO-EC-DC. As several domains show, the similarity between two relations, for example DC and EC, can change in dependence on whether these CN occur at the beginning of a movement pattern (DC1-EC1) or at the end of a movement pattern (DC2-EC2). This can be nicely seen by comparing the shape of the graph of the geometry-translation (GeoT) domains in Figures 5 with all other scenarios. GeoT has a near perfect symmetric shape (in contrast to other domains) with high similarities between non-overlapping CNs as well as high similarities of proper-part CNs. In contrast, similarities of CNs involving PO are relatively low. For a modeling context, this poses a challenge as contextual factors (whether a relation occurs at the beginning or at the end of a movement pattern) have to be taken into consideration.

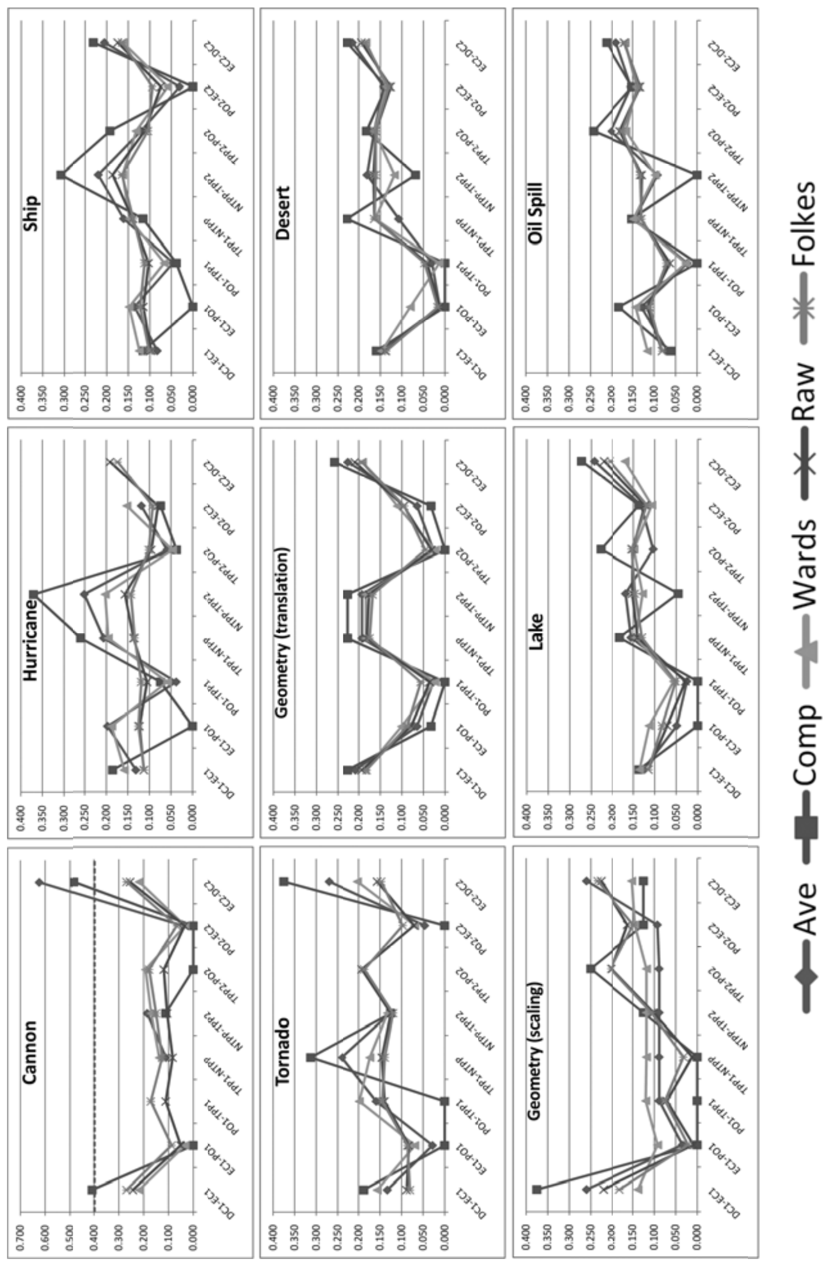

Fig. 5. Each graph visualizes the similarities / weights for CN for each of the nine scenarios using raw similarities and fusion coefficients (a color version of this figure is available at min.us/m_sc2012figure5 for better readability).

- The similarity values in Figure 5 can also be used to potentially answer the question whether an approach based on the region connection calculus is supported or whether IMs are favored. Obviously, and revealed through approaches on behaviorally evaluating QSTR approaches, TECs are forming conceptual groups. In other words, the number of JEPD relations offered by several calculi is higher than the number we would deem cognitively adequate (Clementini et al., 1993; Klippel & Li, 2009; Mark & Egenhofer, 1995). However, while coarser versions of both RCC-8 and IM exist (RCC-5 and coarse IM), these calculi do not match with respect to which relations are merged. In the case of RCC, the relations DC and EC are merged to form DR (discrete from); in the case of IM, the relations EC and PO are merged. Looking at the graphs in Figures 5, we find that some domains tend to support RCC-5, while others support coarse IM:
  – DC1-EC1 > EC1-PO1: Cannon, geometry translation, desert, geometry scaling, and lake.
  – DC1-EC1 =< EC1-PO1: Hurricane, ship, tornado, and oil.
- All scenarios, except for geometry scaling, show high similarities for CNs with proper part relations (TPP1-NTPP, NTPP-TPP2). This finding is consistent with the transition from both RCC-8 to RCC-5 and the fine to coarse transition in the IMs. This finding is also consistent with Li and Fonseca's (2006) assumption that TPP/NTPP relations are very similar to each other. However, in their model these two relations receive the highest similarity of all relations which clearly is not always supported by the data discussed here.

4 Conclusions and Outlook

4.1 Tangible Findings

All current approaches that either propose equal weights or some weighting scheme do so either introspectively or based on formal requirements. None of these approaches capture the "cognitive reality" that similarities between CNs change dependent on domain semantics. While it is still difficult to capture / derive weights directly from our data, it is clear that we need a deeper understanding of the processes at work to be able to guide weight assignments (see below for a theoretical discussion).

It is also clear that there is not simply a single formalism that will be able to capture similarity universally. We have seen a) that Clementini's (Clementini et al., 1993) proposal to use as few as five relations potentially is a step in the right direction as several TECs are very similar to each other; b) however, which TECs are considered as being more or less similar is dependent on the domain. In the analysis we showed that some scenarios follow RCC-5 while others may be better captured using the coarse version of IM.

One aspect important for using qualitative approaches to capture similarities of events is that similarities may be asymmetric (see discussion in Section 3). This aspect has been pointed out early on by Tversky (e.g., Tversky & Gati, 1978) and we do find aspects of asymmetry in nearly every domain we analyzed (e.g., whether a hurricane moves toward the coast, DC-EC-PO …, or away from the coast, PO-EC-DC). That means it matters whether two TECs are in AB or BA order. The similarity again is not something arbitrarily assigned but is an indication of an underlying (commonsense) process model that has guided participants in performing the grouping task.

4.2 A Reflection on the Methods Used

One aspect to keep in mind is that our experimental design follows a classic grouping paradigm, that is, two icons considered as being similar to each other are placed into the same group by participants. Also referred to as category construction or free classification, this method has gained widespread acceptance across a number of disciplines (Medin et al., 1987; Pothos, 2005; Roth et al., 2011), despite several limitations. The coding of similarities is binary, that is, either '0' or '1' with the implication that integers are used throughout individual similarity matrices as well as in the OSM. While the OSM reflects that individual icons may belong to more than one group or belong sort of to one but also to another group, individual matrices do not allow for such detailed distinctions. One possibility to overcome this shortcoming is to employ a different method for assessing similarities in user studies. Examples that come to mind are direct similarity assessments, that is, a participant would rate the similarity of two icons at a time on a continuous scale. The disadvantage of this method is that to achieve the same number of similarity ratings as, for example, in our experiments, 2556 comparisons would have to be made (not counting symmetric and same icon comparisons). As we (and others) have run experiments with substantially more icons, this method becomes quickly infeasible as a certain number of repetitions (per TEC) is necessary to avoid influence of individual stimuli (icons). Other methods, such as selecting the most similar icon from a group of potential target icons have a similar problem of not creating enough data points for the similarity matrix. While it is possible to simulate data, we prefer methods that provide the respective data directly (Rogowitz, Frese, Smith, Bouman, & Kalin, 1998).

One way to obtain continuous similarity rankings would be to allow participants to place icons into piles on a continuous surface (e.g., a computer screen). The advantage would be that the Euclidean properties of such a surface could be used to derive continuous similarity measures by using Pythagoras theorem. This way every pair of icons would be assigned a similarity/dissimilarity value.

The data analysis using fusion coefficients has shown that topology is often but not always the main grouping criterion. This is reflected by a few missing values in Figure 5. This problem could be prevented by using averages of fusion coefficients for all instances within TECs (in relation to all other instances of a second TEC). The raw similarity matrix would basically be replaced by the cophenetic matrix. The downside of using this method would be that the purpose of clustering methods, revealing natural groupings, would be circumvented. The bigger issue that these missing values hint at is that in cases in which TECs are split up (e.g., several overlap and proper part relations in the cannon experiment/domain), topology simply is not the main grouping criterion and as such, it is difficult to derive weights for a topological CNG from this data. Our research focused primarily on topology and as revealed by other methods we used (raw similarities and indices), the results are reasonable with respect to shedding light on similarities/weights in CNGs. However, it would be necessary to conduct similar research on other aspects of spatial knowledge such as distance and directions (Bruns & Egenhofer, 1996; Li & Fonseca, 2006). An open question is whether individual similarities in a scene (or event) can be added or whether holistic methods are necessary to assess overall similarities.

4.3 Some Theoretical Thoughts

Hirschfeld and Gelman (1994), in their introduction to their book on *Mapping the Mind*, state that "[…] much of human cognition is domain-specific." (p. 3) While domain-specificity can have multiple meanings, the one meaning important for this paper is related to semantic domains (e.g., Guarino & Giaretta, 1995) and specifically addresses geographic domains. As such, the explanation of the behavioral data might be loosely related to the concept of *theory theory* (Gopnik & Wellman, 1994), approaches to model common-sense knowledge (Davis, 1990; Hobbs & Moore, 1985), and computational approaches to semantics such as FrameNet (Fillmore & Baker, 2010). However, the main focus of this paper has been on qualitative theories of geographic event conceptualization and how they can be grounded cognitively. To this end, it is important to understand that from an ontological perspective, this paper is much closer to upper level ontologies and addresses the problem that qualitative spatial relations have been considered largely applicable in a domain-independent fashion (with the exception that different formal models are suggested based on introspections of researchers or based on formal constraints). Keil (1994) noted that "The revival of interest in domains of cognition, especially in the contexts of cross-cultural and developmental studies, is a welcome new awareness of how different sorts of concepts and belief systems might become tailored to particular kinds of lawful regularities in our physical and social worlds." (p. 234). What is needed is indeed an approach on modeling geographic event conceptualization which systematically identifies regularities in the external world and allows for providing quantitative measures that will improve the cognitive adequacy of QSTR in several information processing tasks. In the spatial sciences, process models are being developed that capture domain specific information with the goal to characterize not only entities and their relations but, additionally, underlying processes (Torrens, 2012). These approaches should be explored for the modeling of behavioral data, too.

4.4 Some Application Oriented Thoughts

Last but not least, one of the next steps in our research will be the incorporation of the developed similarity models into spatial query processing and retrieval systems. On the one hand, this would improve the usability of such systems by empowering them with the ability to provide answers based on the relational similarity to the given query and make suggestions even when no exact match can be found or in query-by-example scenarios. The provided output could then, for instance, consist of a ranked set of alternatives. On the other hand, the implemented system would allow for performing a detailed evaluation and comparison of the developed similarity models and, hence, also the different methodologies for deriving weights—using human usability studies. We are currently aiming at an implementation in the form of a generic software module that can be turned into plugins to provide similarity-based querying capabilities within existing GIS software and query interfaces to spatial information on the semantic web. The module will be instantiated with a weighted conceptual neighborhood model for an arbitrary JEPD spatial calculus together with an implementation of predicates for the different relations applicable to geometric information. It will then be able to process queries over the defined set of relations and give a similarity-ranked set of instances as a result. To deal with configurations of more than two objects, the similarity values in the weighted conceptual neighborhood graph will have to be aggregated over several relations to yield an overall similarity assessment. Investigating different approaches for this aggregation step will also be a topic of future research.

5 References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis.* Newbury Park, CA: Sage.

Bruns, H. T., & Egenhofer, M. J. (1996). Similarity of spatial scenes. In M. J. Kraak & M. Molenaar (Eds.), *Seventh international symposium on spatial data handling (SDH'96), Delft, The Netherlands* (pp. 173–184).

Camara, K., & Jungert, E. (2007). A visual query language for dynamic processes applied to a scenario driven environment. *Journal of Visual Languages and Computing, 18*, 315–338.

Clementini, E., Di Felice, P., & van Oosterom, P. (1993). A small set of formal topological relationships suitable for end-user interaction. In D. Abel & B. C. Ooi (Eds.), *Advances in Spatial Databases. Third International Symposium, SSD '93 Singapore, June 23–25, 1993 Proceedings* (pp. 277–295). Berlin: Springer.

Cohn, A. G. (2008). Conceptual neighborhood. In S. Shekhar & H. Xiong (Eds.), *Encyclopedia of GIS* (p. 123). Boston, MA: Springer.

Cohn, A. G., & Renz, J. (2008). Qualitative spatial representation and reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Foundations of artificial intelligence. Handbook of knowledge representation* (1st ed., pp. 551–596). Amsterdam: Elsevier.

Davis, E. (1990). *Representations of commonsense knowledge.* San Mateo, CA: Morgan Kaufmann Publishers.

Dylla, F., & Wallgrün, J. O. (2007). Qualitative spatial reasoning with conceptual neighborhoods for agent control. *Journal of Intelligent and Robotic Systems, 48*(1), 55–78.

Egenhofer, M. J., & Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information Systems, 5*(2), 161–174.

Everitt, B. S. (2001). *Cluster analysis* (4. ed.). London: Arnold.

Fillmore, C. J., & Baker, C. F. (2010). A frames approach to semantic analysis. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 313–339). Oxford: Oxford University Press.

Freksa, C. (1992). Using orientation information for qualitative spatial reasoning. In A. U. Frank, I. Campari, & U. Formentini (Eds.), *Theories and methods of spatio-temporal reasoning in geographic space* (pp. 162–178). Berlin: Springer.

Galton, A. (2000). *Qualitative spatial change. Spatial information systems.* Oxford: Oxford Univ. Press.

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak (Ed.), *The Cambridge handbook of thinking and reasoning* (pp. 13–36). CambridgeCambridge Univ. Press,.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* New York: Cambridge University Press.

Guarino, N., & Giaretta, P. (1995). Ontologies and knowledge bases: Knowledge building and knowledge sharing. In N. Mars (Ed.), *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing.* (pp. 25–32). Amsterdam: IOS Press.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002a). Cluster validity methods: Part I. *ACM SIGMOD Record, 31*(2), 40–45.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002b). Clustering validity checking methods: Part II. *ACM SIGMOD Record, 31*(3), 19–27.

Hirschfeld, L. A., & Gelman, S. A. (1994). Toward a topography of mind: An introduction to domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 3–35). New York: Cambridge University Press.

Hobbs, J. R., & Moore, R. C. (Eds.) (1985). *Formal theories of the common sense world.* Norwood, NJ: Ablex.

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reasoning.* Chicago: University of Chicago Press.

Keil, F. C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234–354). New York: Cambridge University Press.

Ketchen, D. J., & Hult, T. M. (2000). Validating cluster assignments. *Psychological Reports, 87*, 1057–1058.

Klippel, A. (accepted). Spatial information theory meets spatial thinking - Is topology the Rosetta Stone of spatial cognition? *Annals of the Association of American Geographers*, (67 manuscript pages).

Klippel, A., & Li, R. (2009). The endpoint hypothesis: A topological-cognitive assessment of geographic scale movement patterns. In K. Stewart Hornsby, C. Claramunt, M. Denis, & G. Ligozat (Eds.), *Spatial Information Theory, 9th International Conference, COSIT 2009, Aber Wrac'h, France, September 21-25, 2009 Proceedings* (pp. 177–194). Berlin: Springer.

Klippel, A., Li, R., Yang, J., Hardisty, F., & Xu, S. (in press). The Egenhofer-Cohn Hypothesis: Or, Topological Relativity? In M. Raubal, A. U. Frank, & D. M. Mark (Eds.), *Cognitive and Linguistic Aspects of Geographic Space - New Perspectives on Geographic Information Research* .

Klix, F. (1971). *Information und Verhalten: Kybernetische Aspekte der organismischen Informationsverarbeitung; Einführung in naturwissenschaftliche Grundlagen der allgemeinen Psychologie.* Bern: Huber.

Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. *Spatial Cognition and Computation, 1*(3), 261–290.

Knauff, M., Rauh, R., & Renz, J. (1997). A cognitive assessment of topological spatial relations: Results from an empirical investigation. In S. C. Hirtle & A. U. Frank (Eds.), *Spatial Information Theory: A Theoretical Basis for GIS* (pp. 193–206). Berlin: Springer.

Kos, A. J., & Psenicka, C. (2000). Measuring cluster similarity across methods. *Psychological Reports, 86*, 858–862.

Lewin, K. (1936/1966). *Principles of Topological Psychology*. New York: McGraw-Hill.

Li, B., & Fonseca, F. (2006). TDD: A comprehensive model for qualitative spatial similarity assessment. *Spatial Cognition and Computation, 6*(1), 31–62.

Lu, S., Harter, D., & Graesser, A. C. (2009). An empirical and computational investigation of perceiving and remembering event temporal relations. *Cognitive Science, 33*, 345–373.

Mark, D. M. (1999). Spatial representation: A cognitive view. In D. J. Maguire, M. F. Goodchild, D. W. Rhind, & P. A. Longley (Eds.), *Geographical Information Systems: Principles and Applications. Second Edition, v. 1* (pp. 81–89).

Mark, D. M., & Egenhofer, M. J. (1995). Topology of prototypical spatial relations between lines and regions in English and Spanish. In *Proceedings, Auto Carto 12, Charlotte, North Carolina, March 1995,* (pp. 245–254).

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*(2), 242–279.

Milligan, G. W. (1996). Clustering validation: results and implications for applied analyses. In P. Arabie, L. J. Hubert, & G. de Soete (Eds.), *Clustering and classification* (pp. 341–375). River Edge, NJ: Word Scientific Publ.

Nedas, K. A., & Egenhofer, M. J. (2008). Integral vs. separable attributes in spatial similarity assessments. In C. Freksa, N. S. Newcombe, P. Gärdenfors, & S. Woelfl (Eds.): *Vol. 5248. Lecture Notes in Computer Science, Spatial Cognition VI. Learning, Reasoning, and Talking about Space, International Conference Spatial Cognition 2008, Freiburg, Germany, September 15-19, 2008. Proceedings* (pp. 295–310). Springer.

Papadias, D., & Delis, V. (1997). Relation-based similarity. In *Proceedings of the 5th ACM Workshop on GIS, Las Vegas, ACM* (pp. 1–4).

Piaget, J., & Inhelder, B. (1948/56/67). *Child's Conception of Space.* New York: Norton.

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences, 28*(1), 1–49.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(1), 17–30.

Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connections. In B. Nebel, C. Rich, & W. R. Swartout (Eds.), *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning* (pp. 165–176). San Francisco: Morgan Kaufmann.

Regier, T., & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science, 31*(4), 705–719.

Renz, J. (2002). *Qualitative spatial reasoning with topological information. Lecture Notes in Artificial Intelligence: Vol. 2293*. Berlin: Springer.

Rissland, E. (2006). AI and Similarity. *IEEE Intelligent Systems, 21*(3), 39–49.

Rogowitz, B. E., Frese, T., Smith, J. R., Bouman, C. A., & Kalin, E. (1998). Perceptual image similarity experiments. In B. E. Rogowitz & N. P. Thrasyvoulos (Eds.), *Proceedings of SPIE. Human Vision and Electronic Imaging III* (pp. 576–590).

Romesburg, H. C. (2004). *Cluster analysis for researchers:* LULU Press.

Roth, R. E., Finch, B. G., Blanford, J. I., Klippel, A., Robinson, A. C., & MacEachren, A. M. (2011). The card sorting method for map symbol design. *Cartography and Geographic Information Science, 38*(2), 89–99.

Schultz, C., Amor, R., & Guesgen, H. W. (2011). Methodologies for qualitative spatial and temporal reasoning application design. In S. M. Hazarika (Ed.), *Qualitative Spatio-temporal Representation and Reasoning. Trends and Future Directions.* Hershey, Pa.: IGI Global.

Schwering, A. (2007). Semantic similarity of natural language spatial relations. In *Conference on Artificial Intelligence and Simulation of Behaviour (AISB07): Artificial and Ambient Intelligence. Symposium: Spatial Reasoning and Communication, Newcastle upon Tyne, UK, April 2nd-5th 2007.*

Schwering, A. (2008). Approaches to semantic similarity measurement for geo-spatial data: A survey. *Transactions in GIS, 12*(1), 2–29.

Torrens, P. M. (2012). *Process Models and Next-Generation Geographic Information Technology - ArcNews Summer 2009 Issue.* Retrieved February 11, 2012, from http://www.esri.com/news/arcnews/summer09articles/process-models.html.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. L. B. Rosch (Ed.), *Cognition and categorization* (pp. 79–98). Hillsdale, NJ: Lawrence Erlbaum.

Wallgrün, J. O., Wolter, D., & Richter, K.-F. (2010). Qualitative matching of spatial information. In A. E. Abbadi, D. Agarwal, M. Mokbel, & P. Zhang (Eds.): *GIS '10, Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 300-309). New York, NY, USA: ACM.

Yang, J., Klippel, A., & Li, R. (submitted). Assessing the cognitive saliency of topologically identified discontinuities in earth dynamics. *International Journal of Geographical Information Science*.