Order Theoretical Knowledge Discovery: A White Paper

Cliff Joslyn*, Joseph Oliveira[†], Chad Scherrer[‡]

INCOMPLETE DRAFT FOR DISCUSSION
NEITHER FOR DISTRIBUTION NOR ATTRIBUTION
August, 2004

Abstract

We describe a perspective on user-guided knowledge discovery techniques in large data spaces based on combinatorial algorithms rooted in order theory. In general, we propose pursuing a class of methods based on casting databases as ordered combinatorial data objects equipped both with inherent semantics and appropriate quantitative measures to support user-guided discovery tasks such as search, retrieval, discovery, anomaly detection, and linkage, with applications in intelligence analysis, homeland defense, computational biology, and law enforcement.

Motivation

Consider the following collection of tasks in the exploitation of information resources and databases, typically called Knowledge Discovery in Databases (KDD) [12, 32]:

- Helping guide users to find meaningful connections in databases containing not just thousands or millions of records (rows), but hundreds of different *kinds* of information (columns) being kept track of in each record.

- Enabling search and retrieval from databases on the basis of the semantics and meaning of the information in the request, rather than (just) explicit textual matching.

- Allowing categorization and navigation in large taxonomic catalogs available in biology, astrophysics, law enforcement, national defense, and homeland security.

- Linking multiple existing ontologies and knowledge bases according to their shared semantics.

- Identifying causal and structural connections among agents in complex interacting networks.

- Finding anomalous connections in relational and transactional data.

For many years, these tasks have concerned researchers in the knowledge-based sciences such as biology and other areas requiring knowledge exploitation such as homeland defense, intelligence analysis, and marketing. And while many have approached these tasks in many different ways, they are typically grounded in multi-dimensional *statistical* and machine learning approaches, and generally ignore how these high-dimensional, relational knowledge spaces either are natively, or admit to descriptions in terms of, large, finite, *discrete structures* with a decidedly *ordered* nature.

We assert the significance of combinatorial order theory for KDD, and in **Order Theoretical Knowledge Discovery (OTKD)** have identified a collection of both particular mathematical techniques and an overall problem-solving paradigm which we believe can be instrumental in making progress in many of these areas simultaneously. We aim to bring combinatorial and order theoretical approaches to KDD, thereby seeking conceptual and mathematical unification to fill the gap between discrete mathematics and data mining.

One can ask "why now", what is it about the current situation which suggests such a move? First, recent years have seen the emergence of new kinds of data objects (such as large taxonomic ontologies like the Gene Ontology (GO) [1] and the UMLS MetaThesaurus [3]) as useful, interactive databases, rather than laboratory experiments. And new mathematical techniques such as concept lattices [13] are available to provide order theoretical representations of relational data objects.

*Knowledge Systems and Computational Biology Team, Computer and Computational Sciences, Mail Stop B265, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, joslyn@lanl.gov, http://www.c3.lanl.gov/~joslyn, (505) 667-9096.

[†]PNNL

[‡]PNNL

Order theory [33], lattice theory [7], matroid theory [40], Sperner theory [11], and related areas comprise a robust area of discrete mathematics. But while these mathematical techniques offer significant methods for KDD, as theories they focus on mathematical objects which have far more regularity and elegance than typically available in "messy", real-world databases, and there is a corresponding lack of developments of the necessary combinatorial algorithms. Conversely, while ordered structures abound in KDD, which some exceptions [27] they are not generally approached from a strong mathematical basis.

Remarkable progress *has* been made recently in the application of discrete mathematical techniques, primarily in graph and network theory, to a wide range of challenging problems in complex systems [30]. But these have not had direct applicability in KDD, nor have they focused on *ordered* combinatorial structures: ordered structures are *not* networks!

Combinatorial Data Objects

Central to OTKD is the representation of databases and data objects as large, finite ordered, structures we call **combinatorial data objects** (CDOs). Note that CDOs can originate from quite different sources:

Native: Built explicitly by hand, and thus of moderate size ($\leq 10^6$ nodes), typically used to model conceptual spaces. Examples include:

- Taxonomic structures underlying ontologies [1] cast as collections of partially ordered sets (multi-posets) [23];
- Verb typing hierarchies in computational linguistics [8];
- Object-oriented typing hierarchies and other meta-modeling objects [3, 27] cast as Directed Acyclic Graphs (DAGs), and thus as posets; and
- Higher-order conceptual structures such as semantic networks [37] and conceptual graphs [35] cast as lattices of directed hypergraphs [16].

Derived: Automatically induced from analysis of other data objects, and can be very large. Examples include:

- Concept lattices derived from binary relational tables [13];
- Lattices of reconstruction hypotheses of relational databases cast as irredundant covers of a space of variables [25];
- Lattices of hypothetical causal "masks" over temporal or otherwise linearly ordered data streams [25];
- Graphoid structures of Bayesian belief networks [4, 36];
- The partition poset to represent the block structure of distributions of projected database vectors [2, 20, 21]; and
- Cubic and Boolean lattices to represent join spaces and projections of a given database [6].

Equipped with Semantics

These ordered structures derive their native semantics from their source databases, and perhaps additional semantics from the context of their combinatorial relations.

- Semantics can be available at the level of the nodes of the CDO, or at the level of the links (relations) between them.
- Link type semantics typically involve the semantic categories familiar to us from knowledge representation, for example, inclusion, implication, subsumption (generalization, inheritance, abstraction, or "is-a" links), composition (containment or "has-part" links), negation; or could be *ad-hoc* to a particular source database.
- Typically, semantics are related to the properties of the source data types as relational objects (for example, scalar, vector, or tensor-valued; categorical, ordinal, or real-valued), or the units of measure present.
- Even more significantly, native semantics are typically represented by or derived from textual annotations, either from free text sources, or from meta-data, keywords, categorizations, or other semantic tags.

Equipped With Quantitative Measures

Since we use CDOs to represent databases, we seek quantitative measures to help guide users to areas of interest. Examples include:

Combinatorial Measures: A novel set of questions arise when considering databases as combinatorial objects. Specifically, given large, real-world, and decidedly "messy" combinatorial objects, how can we measure their quantitative properties as a whole, and then the properties of portions of the CDO and between multiple portions of a CDO. Examples include:

- Generalized concepts of **rank** as vertical "level" within a CDO.
- Generalized concepts of **distances** between pairs of nodes or collections of nodes within a CDO.

Statistical Measures: These are the class of quantitative measures most currently developed. They are primarily related to classical statistical measures, and are thus information-theoretical entropies of one form or another, or are similar measures within a broader generalized information theory [19, 24, 26]. Examples include:

- Conditional entropies of graphical models [**?**] and reconstruction hypotheses [25].
- Relative entropies in poset partitions [20, 21].

Tasks

These statistics, preferably when used together, allow users to perform various operations with, within, and among CDOs, for example:

User-Guided Discovery: Computer-assisted navigation with and maneuvering within complex CDOs, including search and retrieval, categorization, and anomaly detection.

Construction: Abduction and induction of conceptual structures such as ontologies and taxonomies from data.

Interoperability: The ability to merge, match, compare, and link multiple CDOs in order to establish interoperability amongst their target databases.

Examples

Examples of some existing methods, techniques, and applications within our broad perspective in OTKD include:

Multi-Dimensional Link Analysis (MDLA): Network representations of relational data as binary graphical links are typically limited to a single kind of connection (semantic relation, link type). High dimensional data requires the use of multiple link types [28, 37] (see Fig. 1), and Multi-Dimensional Link Analysis (MDLA) exploits the representation of projections of high-dimensional spaces as points in a Boolean or cubic lattice; "views" of a database as projected relational subsets; and "chaining" operations as steps between views in a knowledge discovery "trajectory" in these ordered structures [17]. Joslyn and Mniszewski [20, 21] are exploring the use of the partition poset [2] to represent the extension of views by a single dimension; and informational measures in the partition poset (Fig. 2) to provide guides to users to explore possible extensions. Prototype implementations and demonstration of these techniques has been made in information analysis of terrorist networks [21, 39].

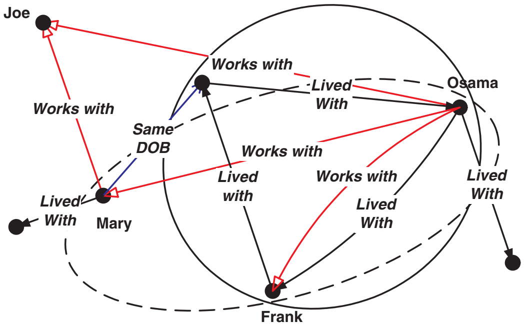


Figure 1: Labeled network representations of multi-dimensional relational data.

Anomaly Detection in Concept Lattices: Formal concept analysis [10, 13] is an increasingly exciting technology for representing relational data (e.g. as shown in Fig. 3 [21]) using ordered structures (Fig. 4 [21]). The use of FCAs for information analysis of terrorist networks has been demonstrated [21, 39], and Joslyn has proposed [18] extending rank and distance measures in concept lattices to provide a mechanism for anomaly detection in relational data.

Reconstructibility Analysis: Klir [25] and others

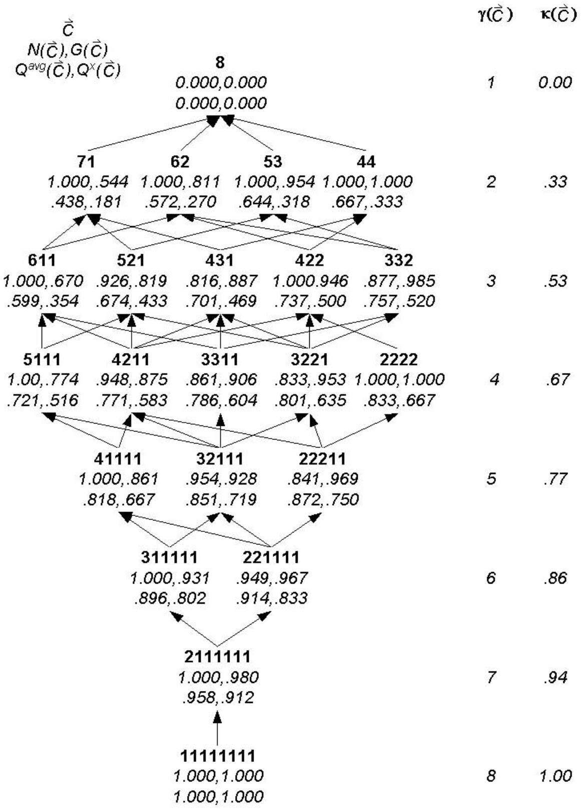
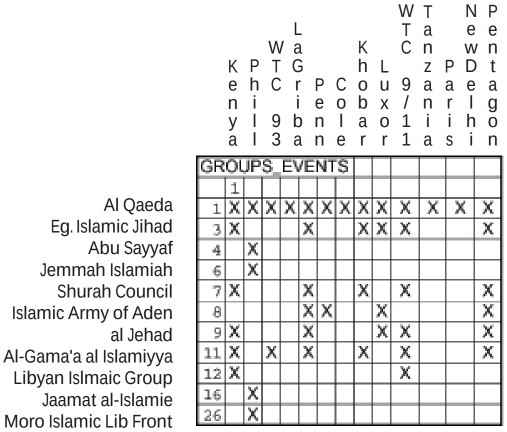


Figure 2: Measures in the partition poset ($n = 8$) for link analytical chaining [20].



		Kenya	Phill	WTC 93	La Griba	Penn	Cole	Khobar	Luxor	WTC 9/11	Tanzania	Paris	New Delhi	Pentagon
	GROUPS_EVENTS													
		1												
Al Qaeda	1	X	X	X	X	X	X	X	X	X	X	X	X	X
Eg. Islamic Jihad	3	X				X			X	X	X			X
Abu Sayyaf	4		X											
Jemmah Islamiah	6		X											
Shurah Council	7	X				X			X		X			X
Islamic Army of Aden	8					X	X			X				X
al Jehad	9	X				X				X	X			X
Al-Gama'a al Islamiyya	11	X		X		X			X		X			X
Libyan Islmaic Group	12	X									X			
Jaamat al-Islamie	16		X											
Moro Islamic Lib Front	26		X											

Figure 3: A simple data relation [21].

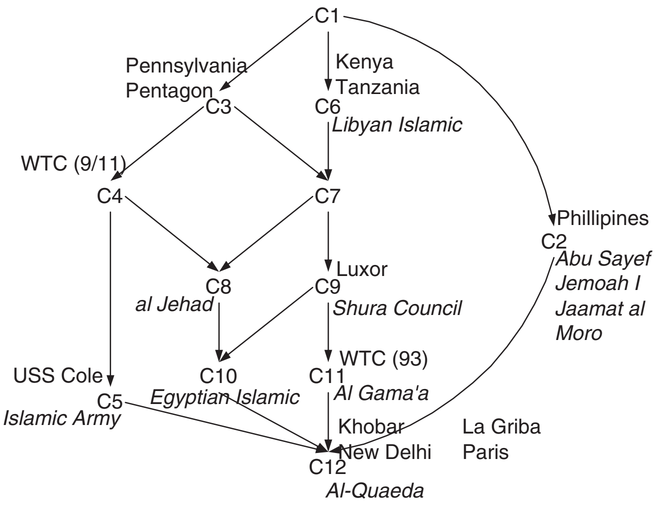


Figure 4: Concept lattice representation of Fig. 3 [21].

[5] have proposed the lattice of irredundant covers of a set (see Fig. 5) to represent the possible relations between the parts, cast as relational projections, and the whole of a system of interest. They use search optimized by conditional information (entropy) measures on these covers to identify reconstruction hypotheses of relational databases and thus induce causal and structural models. This is somewhat in the spirit of hierarchical log-linear modeling [29] and projection pursuit [31] methodologies, and holds promise as an MDLA technique.

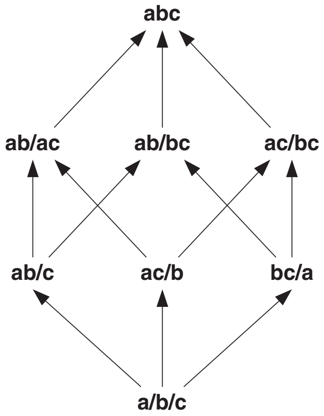


Figure 5: The lattice of irredundant covers for $n = 3$.

Mask Analysis: Mask analysis [25] is a similar technique to reconstructibility analysis where instead of structural relations, generated models predict temporal dependencies of output on input variables.

Graphical Model Identification: In the same spirit, conditional independence structures among collections of variables can be represented as lattices of graphoids [4, 36], and conditional entropy measures used to search for admissible and optimal Bayes net models [**?**].

Categorization in Poset Ontologies: Joslyn *et al.* [22, 23] are considering the mathematical properties of large ontologies, such as the GO, cast as multi-posets. They have developed methods to categorize large gene lists, annotated into the GO, using quantitative scores based on pseudo-distances between comparable nodes (see Fig. 6 [22]), and are extending these ideas to other rank-based measures of distance between general pairs on nodes in posets [18].

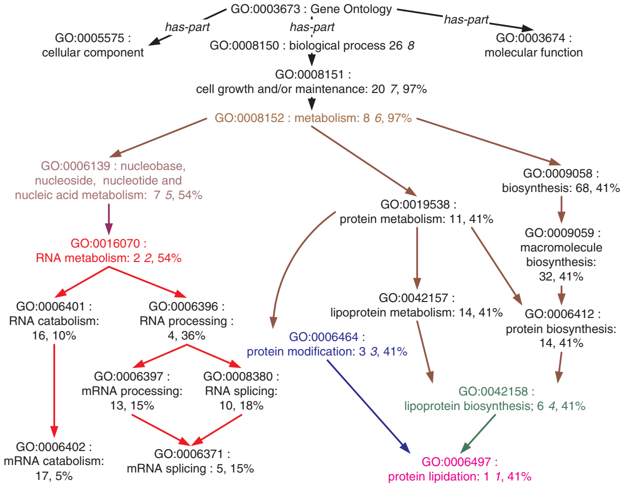

GO:0003673 : Gene Ontology
has-part
has-part
has-part
GO:0005575 : cellular component
GO:0008150 : biological process 26 8
GO:0003674 : molecular function
GO:0008151 : cell growth and/or maintenance: 20 7, 97%
GO:0008152 : metabolism: 8 6, 97%
GO:0006139 : nucleobase, nucleoside, nucleotide and nucleic acid metabolism: 7 5, 54%
GO:0009058 : biosynthesis: 68, 41%
GO:0019538 : protein metabolism: 11, 41%
GO:0009059 : macromolecule biosynthesis: 32, 41%
GO:0016070 : RNA metabolism: 2 2, 54%
GO:0042157 : lipoprotein metabolism: 14, 41%
GO:0006412 : protein biosynthesis: 14, 41%
GO:0006401 : RNA catabolism: 16, 10%
GO:0006396 : RNA processing : 4, 36%
GO:0006464 : protein modification: 3 3, 41%
GO:0006397 : mRNA processing: 13, 15%
GO:0008380 : RNA splicing: 10, 18%
GO:0042158 : lipoprotein biosynthesis; 6 4, 41%
GO:0006402 : mRNA catabolism: 17, 5%
GO:0006371 : mRNA splicing : 5, 15%
GO:0006497 : protein lipidation: 1 1, 41%

Figure 6: Scores in a poset ontology as used in the POSet Ontology Categorizer (POSOC) [23].

Ontology Comparison and Interoperability: To assist in automatic and semi-automatic linkage of multiple knowledge structures, Joslyn [18] has proposed the use of order-preserving maps between multiple poset ontologies (for example the GO and the Enzyme Commission database [9]) and the induction of semantic similarities on the basis of structural intersections and rotations between poset ontologies (Fig. 7).

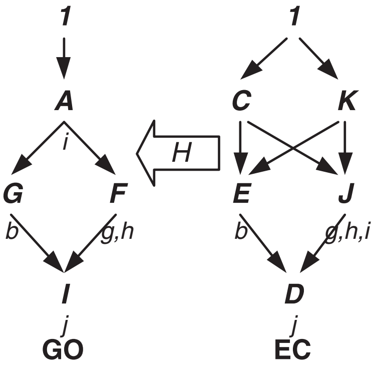


Figure 7: Construction of order-preserving maps between poset ontologies [18].

Text and Ontology Interaction: Verspoor *et al.* [38] are experimenting with the interaction between ontological and lexical information as dually ordered structures, deriving networks of lexical semantic relations as primarily shallow forests from poset ontologies (Fig. 8). The goal is to improve relation extraction from text, for example using subsumptive templates, by exploiting the rich semantic relations present in lexically equipped ontological structures.

Mathematical Treatment of Conceptual Graphs: Where the kinds of ontologies discussed above tend to be simple taxonomies (posets) equipped with additional inferential mechanisms, more complex knowledge objects like semantic networks and conceptual graphs (CGs) [35] have a richer mathematical structure as lattices of hypergraphs [16]. Some have considered the relation of CGs to other structures such as concept lattices [41], and there are tremendous opportunities to bring a stronger mathematical treatment to knowledge bases represented on such a basis.

Structures of Random Sets: Finally, nonclassical information theories are becoming critical technologies for uncertainty quantification in decision support, characterizing knowledge bases, and engineering modeling [19]. An emerging unifying paradigm in generalized information theory arises consistently with OTKD: that of statistical measures on ordered combinatorial structures, specifically generalized probability distributions on lattices [34]. Random sets, which motivate Dempster-Shafer and possibility theories, and generalize to imprecise probabilities [14], are the case of a probability distribution on the power set (see Fig. 9). When the power set is cast as a Boolean lattice, the properties of finite random sets can be characterized completely by

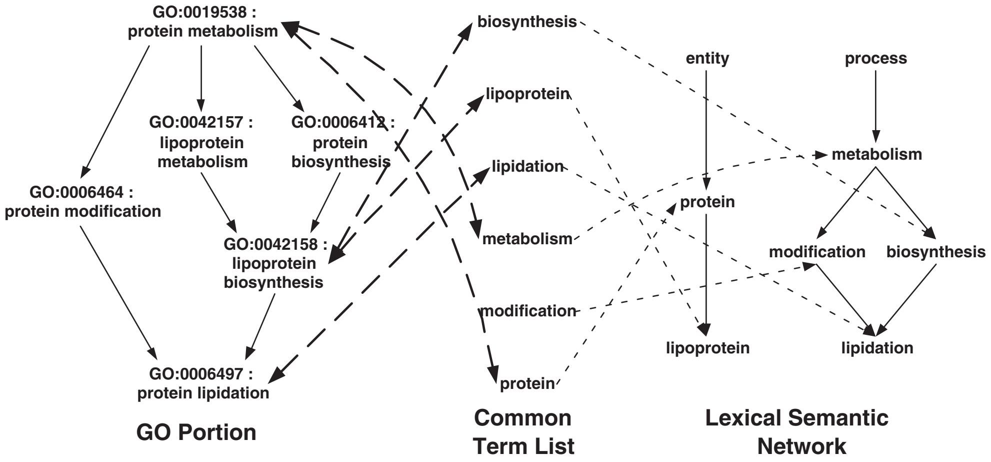


Figure 8: Ordered lexical semantic networks derived from phrasally labeled poset ontologies [38].

the topological relations among classes of sets (see Fig. 10) [15].

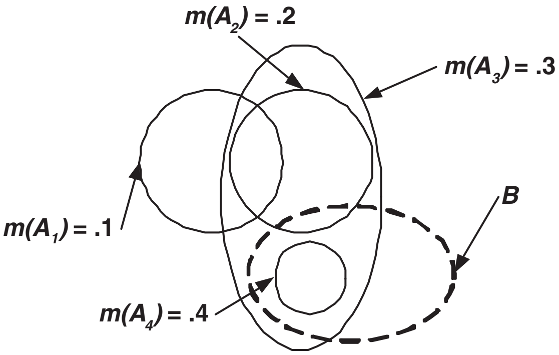


Figure 9: A random set.

Combinatorial Challenges

Not only does order theory promise to provide KDD with a set of critical tools, so KDD is challenging order theory to consider a number of new questions. In general, we can recognize a number of issues we are pursuing here:

- Measures of distance between poset nodes; generalization of poset rank to an interval-valued concept [18].
- Mapping Boolean to cubic lattice to represent relational spaces; casting relational join as an operation in a cubic lattice.

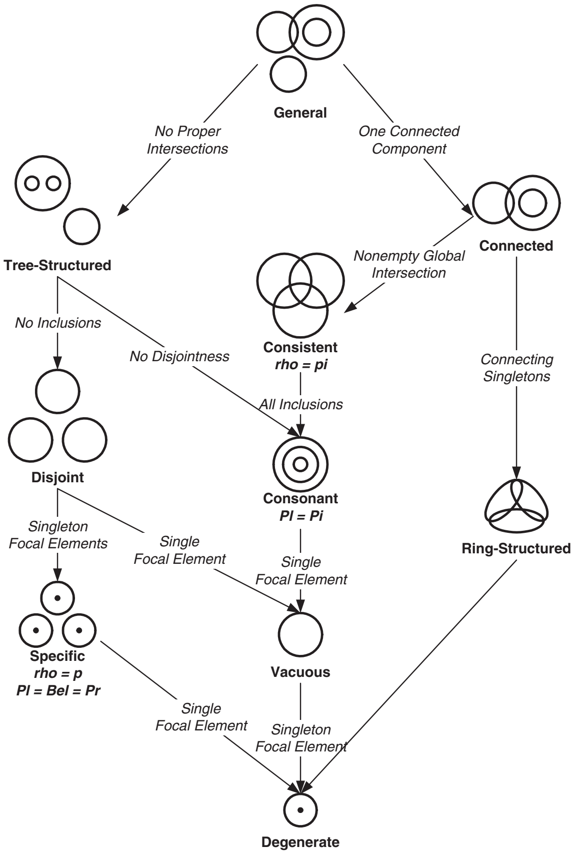


Figure 10: Topologies of classes of random sets.

- Efficient algorithms to calculate poset width.
- The role of matroid theory, ordered matroids, and rank in matroids.
- Extensions of concept lattices to higher dimensions.

References

[1] Ashburner, M; Ball, CA; and Blake, JA et al.: (2000) "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, v. **25**:1, pp. 25-29

[2] Blinovsky, VM and Harper, LH: (2002) "Size of the Largest Antichain in a Partition Poset", *Problems of Information Transmission*, v. **38**:4, pp. 347-353

[3] Bodenreider, Olivier; Mitchell, Joyce A; and McCray, Alexa T: (2002) "Evaluation of the UMLS As a Terminology and Knowledge Resource for Biomedical Informatics", in: *AMIA 2002 Annual Symposium*, pp. 61-65

[4] Borgelt, Christian and Kruse, Rudolf: (2002) *Graphical Models*, Wiley, New York

[5] Cavallo, Roger E and Klir, George: (1979) "Reconstructibility Analysis of Multi-Dimensional Relations: A Theoretical Basis for Computer-Aided Determination of Acceptable Systems Models", *Int. J. General Systems*, v. **5**, pp. 143-171

[6] Chen, William YC and Oliveira, JS: (1995) "Implication Algebras and the Metropolis-Rota Axioms for Cubic Lattices", *J of Algebra*, v. **171**, pp. 383-396

[7] Davey, BA and Priestly, HA: (1990) *Introduction to Lattices and Order*, Cambrudge UP, Cambridge UK, 2nd Edition

[8] Davis, Anthony R: (2000) *Types and Constraints for Lexical Semantics and Linking*, Cambridge UP

[9] *Enzyme Structures Database*, `http://www.biochem.ucl.ac.uk/bsm/enzymes/`

[10] Eklund, Peter, ed.: (2004) *Concept Lattices: Proc. 2nd. Int. Conf. on Formal Concept Analysis (IFCFA 04)*, Springer-Verlag, Berlin

[11] Engel, Konrad and Rota, CG, eds.: (1997) *Sperner Theory*, Cambridge UP

[12] Fayyad, Usama M; Piatetsky, G; and Smyth, P, and R. Uthurusamy, eds.: (1996) *Advances in Data Mining and Knowledge Discovery*, MIT Press, Menlo Park CA

[13] Ganter, Bernhard and Wille, Rudolf: (1999) *Formal Concept Analysis*, Springer-Verlag

[14] Hall, JW and Lawry, J: (2004) "Generation, Combination and Extension of Random Set Approximations to Coherent Lower and Upper Probabilities", *Reliability Engineering and System Safety*, v. **85**:1-3, pp. 88-102

[15] Joslyn, Cliff: (1996) "Aggregation and Completion of Random Sets with Distributional Fuzzy Measures", *Int. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, v. **4**:4, pp. 307-329

[16] Joslyn, Cliff: (2000) "Hypergraph-Based Representations for Portable Knowledge Management Environments", LAUR 00-5660, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/kenv1.pdf`

[17] Joslyn, Cliff: (2002) "Link Anlaysis of Social Meta-Networks", in: *Proc. Conf. on Computational Analysis of Social and Organizational Systems (CASOS 02)*, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/casos02_abs.pdf`

[18] Joslyn, Cliff: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in: *Conceptual Structures at Work, Lecture Notes in Artificial Intelligence*, ed. Wolff, Pfeiffer and Delugach, v. **3127**, pp. 287-302, Springer-Verlag, Berlin

[19] Joslyn, Cliff and Booker, Jane: (2004) "Generalized Information Theory for Engineering Modeling and Simulation", in: *Engineering Design Reliability Handbook*, ed. E Nikolaidis and D Ghiocel, CRC Press, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/crc.pdf`, in press

[20] Joslyn, Cliff and Mniszeiski, Susan: (2002) "DEEP: Data Exploration through Extension and Projection", LAUR 02-1330, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/deep.pdf`

[21] Joslyn, Cliff and Mniszewski, Susan: (2002) "Relational Analytical Tools: DataDelver and Formal Concept Analysis", LAUR 02-7697, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/hl1.pdf`

[22] Joslyn, Cliff and Mniszewski, Susan: (2004) "Combinatorial Approaches to Bio-Ontology Management with Large Partially Ordered Sets", in: *SIAM Workshop on Combinatorial Scientific Computing (CSC 04)*, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/csc04f.pdf`

[23] Joslyn, Cliff; Mniszewski, Susan; Fulmer, Andy; and Heaton, Gary: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. **20**:s1, pp. 169-177

[24] Klir, George: (1991) "Generalized Information Theory", *Fuzzy Sets and Systems*, v. **40**, pp. 127-142

[25] Klir, George and Elias, Doug: (2003) *Architecture of Systems Problem Solving*, Plenum, New York, 2nd edition

[26] Klir, George and Wierman, Mark J: (1999) *Uncertainty-Based Information Elements of Generalized Information Theory*, Springer-Verlag, Berlin

[27] Knoblock, Todd B and Rehof, Jakob: (2000) "Type Elaboration and Subtype Completion for Java Bytecode", in: *Proc. 27th ACM SIGPLAN-SIGACT*

Symposium on Principles of Programming Languages

[28] Lee, Yugyung and Geller, James: (2002) "Efficient Transitive Closure Reasoning in a Combined Class/Part/Containment Hierarchy", *Knowledge and Information Systems*, v. **4**, pp. 305-328

[29] Malvestuto, FM: (1996) "Testing Implication of Hierarchical Log-Linear Models for Probability Distributions", *Statistics and Computing*, v. **6**, pp. 169-176

[30] Milo, Ron; Itzkovitz, Shalev; and Kashtan, N et al.: (2004) "Superfamilies of Evolved and Designed Networks", *Science*, v. **303**, pp. 1538-1542

[31] Montanari, Angela and Lizzani, Laura: (2001) "A Projection Pursuit Approach to Variable Selection", *Computational Statistics and Data Analysis*, v. **35**, pp. 463-473

[32] Piatetsky-Sh, Gregory and Frawley, William J, eds.: (1991) *Knowledge Discovery in Databases*, AAAI Press/MIT Press, Menlo Park CA

[33] Schröder, Bernd SW: (2003) *Ordered Sets*, Birkhauser, Boston

[34] Smets, Philippe: (2002) *Matrix Calculations for Belief Functions*, `http://iridia.ulb.ac.be/ psmets/MatrixRepresentation.pdf`

[35] Sowa, John F: (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole, Pacific Grove

[36] Studeny, Milan: (1994) "Structural Semigraphoids", *Int. J. General Systems*, v. **22**:2, pp. 207-217

[37] Thomason, Richmond H and Touretzky, David S: (1991) "Inheritance Theory and Networks with Roles", in: *Principles of Semantic Networks*, ed. JF Sowa, pp. 231-266, Morgan Kauffman, San Mateo CA

[38] Verspoor, Karin; Joslyn, Cliff; and Papcun, George: (2003) "Gene Ontology as a Source of Lexical Semantic Knowledge for a Biological Natural Language Processing Application", in: *Workshop on Text Analysis and Search for Bioinformatics (SIGIR 03)*

[39] Voss, Susan and Joslyn, Cliff: (2002) "Advanced Knowledge Integration in Assessing Terrorist Threats", LAUR 02-7867, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/knowint.pdf`

[40] White, Neil and Rota, CG, eds.: (1986) *Theory of Matroids*, Cambridge UP

[41] Wille, Rudolf: (1997) "Conceptual Graphs and Formal Concept Analysis", in: *Lecture Notes in Artificial Intelligence*, v. **1257**, pp. 290-303