Semantic Hierarchies in Knowledge Analysis and Integration

Cliff Joslyn

Information Sciences Group

DIMACS Workshop on Recent Advances in Mathematics and Information Sciences for Analysis and Understanding of Massive and Diverse Sources of Data

May 2007

OUTLINE

- The challenge of semantic information for knowledge systems
- Large computational ontologies
  - Analysis
  - Induction
  - Interoperability
- Order theoretical approaches
  - Ontology anlaysis
  - Concept lattices: Formal Concept Analysis

APPLICATION CHALLENGES

Decision Support: Military, intelligence, disaster response

Intelligence Analysis: Multi-Int integration: IMINT, HUMINT, SIGINT, MASINT, etc.

Biomedicine: Biothreat response

Defense Applications: Defense transformation, situational awareness, global ISR

Bibliometrics: Digital libraries, retrieval and recommendation

Simulation: Interaction with knowledge management/decision support environments

Nonproliferation: "Ubiquitous sensing", information fusion

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

KNOWLEDGE SYSTEMS

- Challenge for database integration at the **knowledge** level:

  Connectivity: Wiring everything up, everything *accessible*

  Interoperability: Knowing *what* you have and *where* it is
- Complement *quantitative* statistical techniques with *qualitative* methods:
  - Knowledge representation, natural language processing
  - Search, retrieval, inference
  - Focus on the *meaning* (*semantics*) of information in databases: use, interpretation
- In conjunction with existing capabilities in data mining, machine learning, sensor technology, simulation, etc.
  - **Knowledge-based and data-rich sciences:** Biology, astronomy, earth science
  - **Knowledge-based technologies for national security:** Decision support, intelligence analysis
  - **Knowledge-based technologies supporting the scientific process:** Semantic web, digital libraries, publication process, communities of networked scientists

MULTI-MODAL DATA FUSION

- Qualitative difference:

  Sensors:

  - Physics sensors: nuclear, radiological, chemical
  - Electromagnetic spectrum
  - Acoustic, seismic
  - Images, video

  Information Sources:

  - Geospatial
  - Structured and semi-structured data
  - Relational databases
  - Text, documents
  - Plans, scenarios

- How to bridge?
  - Meta-data
  - Feature extraction from signals, images
  - Feature ontologies and interoperability protocols

LANL KNOWLEDGE AND INFORMATION SYSTEMS SCIENCE

http://www.c3.lanl.gov/knowledge

Semantic Hierarchies for Knowledge Systems

- Representations of *semantic* and *symbolic* information
- Approach from *mathematical systems theory*:
  - Discrete math, combinatorics, information theory
  - Metric geometry approach to order theory (lattices and posets)
- *Hybrid* methodologies combining statistical, numerical, and quantitative with symbolic, logical, and qualitative
- **Ontologies and Conceptual Semantic Systems:** Discrete mathematical approaches
- **Computational Linguistics and Lexical Semantics:** For natural language processing and text extraction
- **Database Analysis:** User-guided knowledge discovery in complex, multi-dimensional data spaces
- **Software Architectures:** Parallel and high performance algorithms

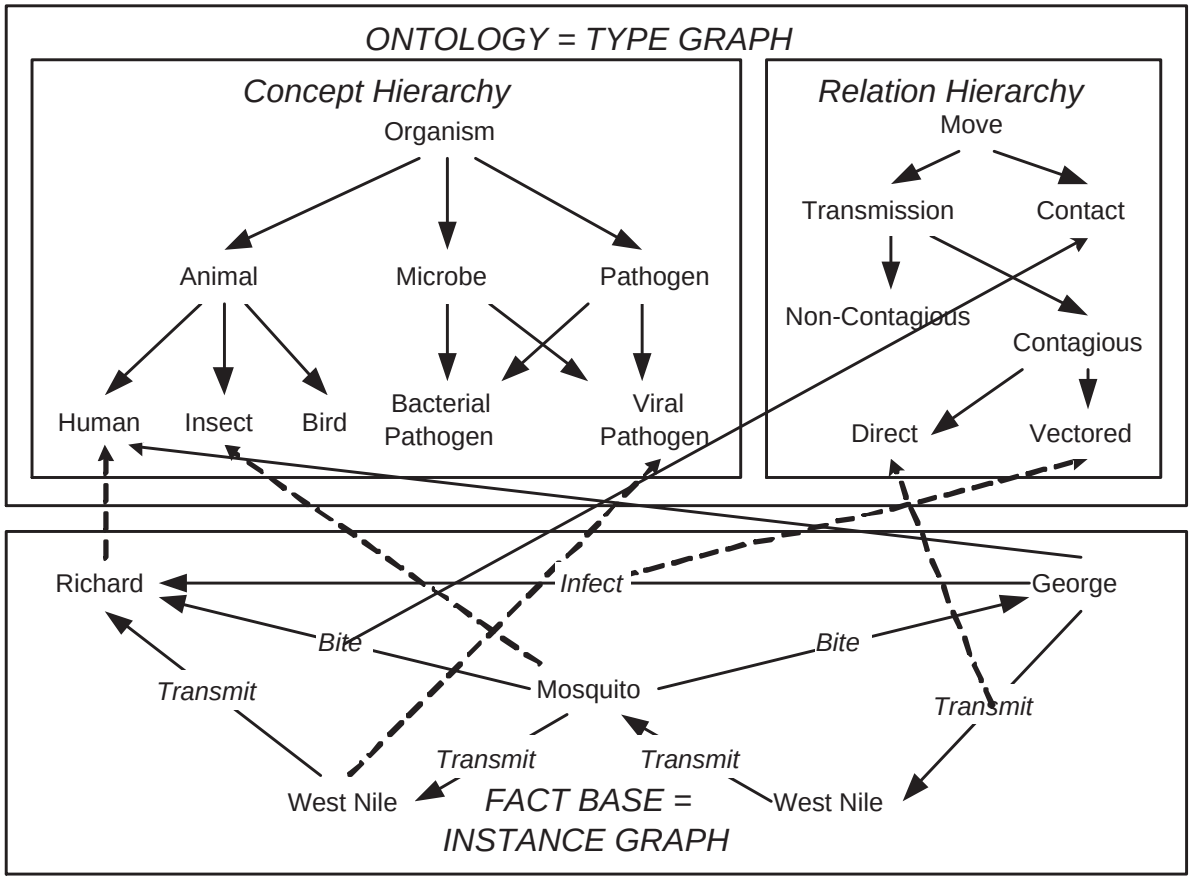
Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

PARADIGM: SEMANTIC NETWORKS

- **Lattice-labeled directed multi-graphs**
- **Increasing size and prominence for databases:** Intelligence analysis, law enforcement, computa-tional biology



- **Challenges:** Typed-link network theory; morphisms of typed graphs; ontology analysis, induction, and interoperability.

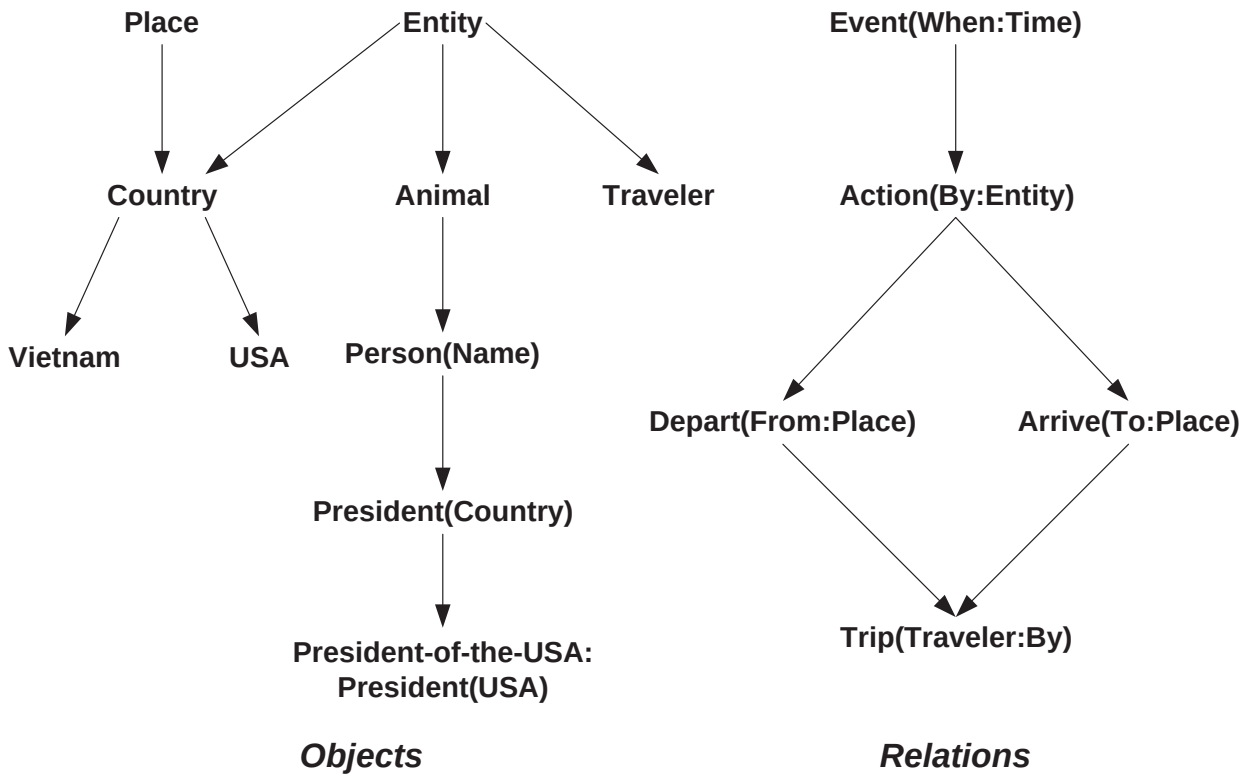
Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

REASONING WITHIN ONTOLOGIES FOR THE SEMANTIC WEB

- Proposed basis for Semantic Web
- Ontological database: interacting hierarchies of objects and relations



- Semantic relations valued on objects
- Description-logic queries

Who was the last president before Clinton to visit Vietnam?

`>>: (Name(By)) ( Trip?`x `( To:Vietman, By:President-of-the-USA ) .and. lub(When(`x`)) <` `1992)`

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

BIO-ONTOLOGIES

- Domain-specific concepts, together with *how they're related semantically*
- Crushing need driven by the genomic revolution
- *At least:*
  - **Large** terminological collections (controlled vocabularies, lexicons)
  - Organized in taxonomic, hierarchical relationships
- *Sometimes in addition:* Methods for inference over these structures
- Molecular, anatomy, clinical, epidemiological, etc.:

  Gene Ontology: Molecular function, biological process, cellular location

  Fundamental Model of Anatomy

  Unified Medical Language System: National Library of Medicine, meta-thesaurus

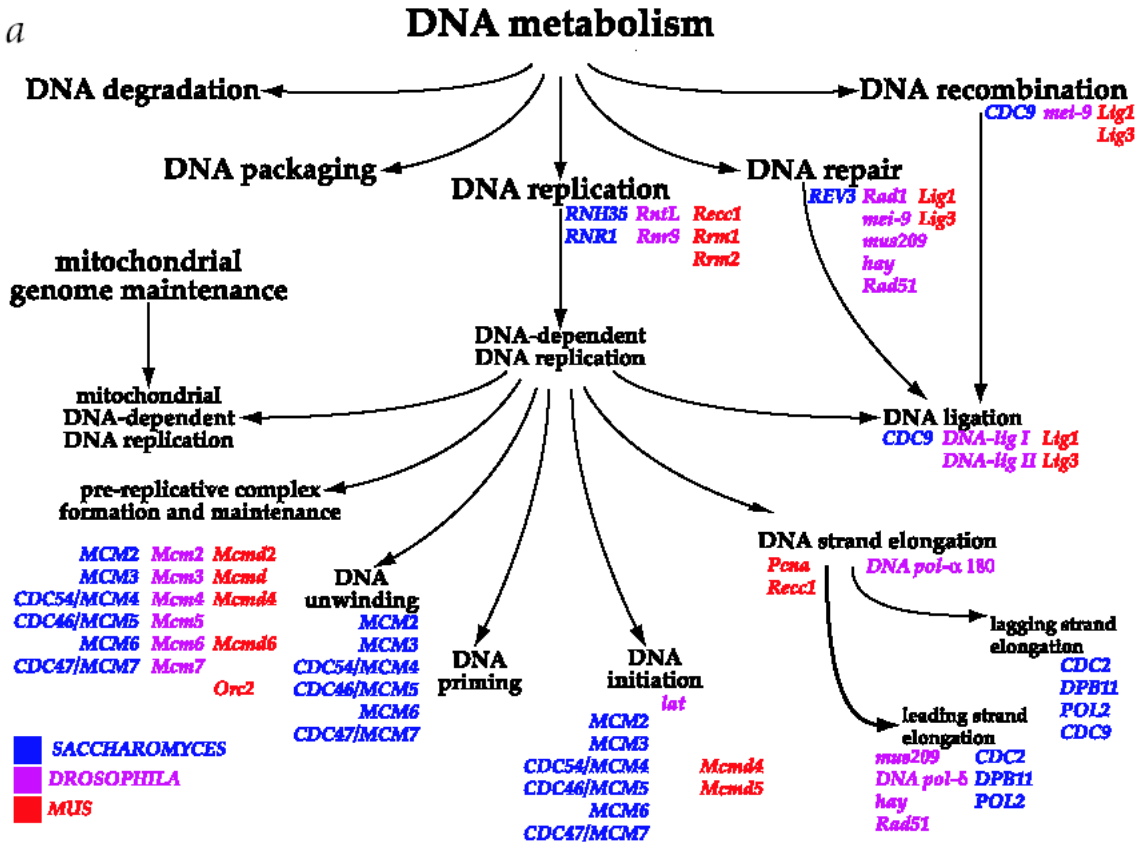
  Open Biology Ontologies

  MEdical Subject Headings (MeSH)

  Enzyme Structures Database: EC numbers

GENE ONTOLOGY (GO): DNA METABOLISM PORTION

- Taxonomic controlled vocabulary
- $\sim 20K$ nodes populated by genes, proteins
- Two orders $\leq_{isa}, \leq_{has}$
- Major community effort: assuming primary position in general bioinformatics



Gene Ontology Consortium (2000): "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, 25:25-29

- Tremendous computational resource: large, semantically rich, validated, middle ontology, first (?) in major use

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

CATEGORIZATION IN THE GENE ONTOLOLGY

http://www.c3.lanl.gov/posoc

- Develop functional hypotheses about hundreds of genes identified through expression experiments
- Given the Gene Ontology (GO) ...
- And a list of hundreds of genes of interest ...
- "Splatter" them over the GO ...
- Where do they end up?
  - Concentrated?
  - Dispersed
  - Clustered?
  - High or low?
  - Overlapping or distinct?
- POSet Ontology Categorize (POSOC)

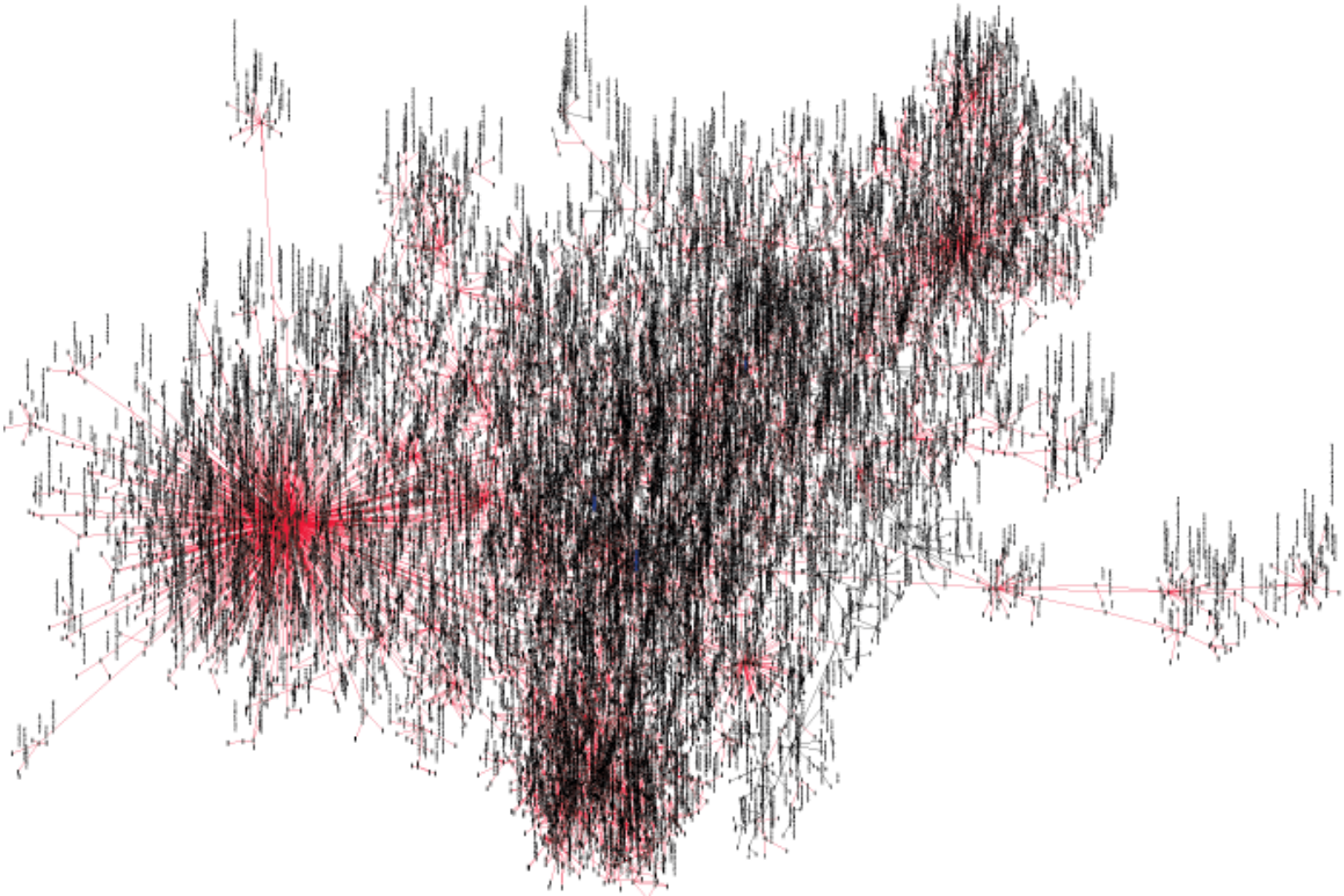
C Joslyn, S Mniszewski, A Fulmer, and G Heaton: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. **20**:s1, pp. 169-177

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

WHOLE GO CA. 2001



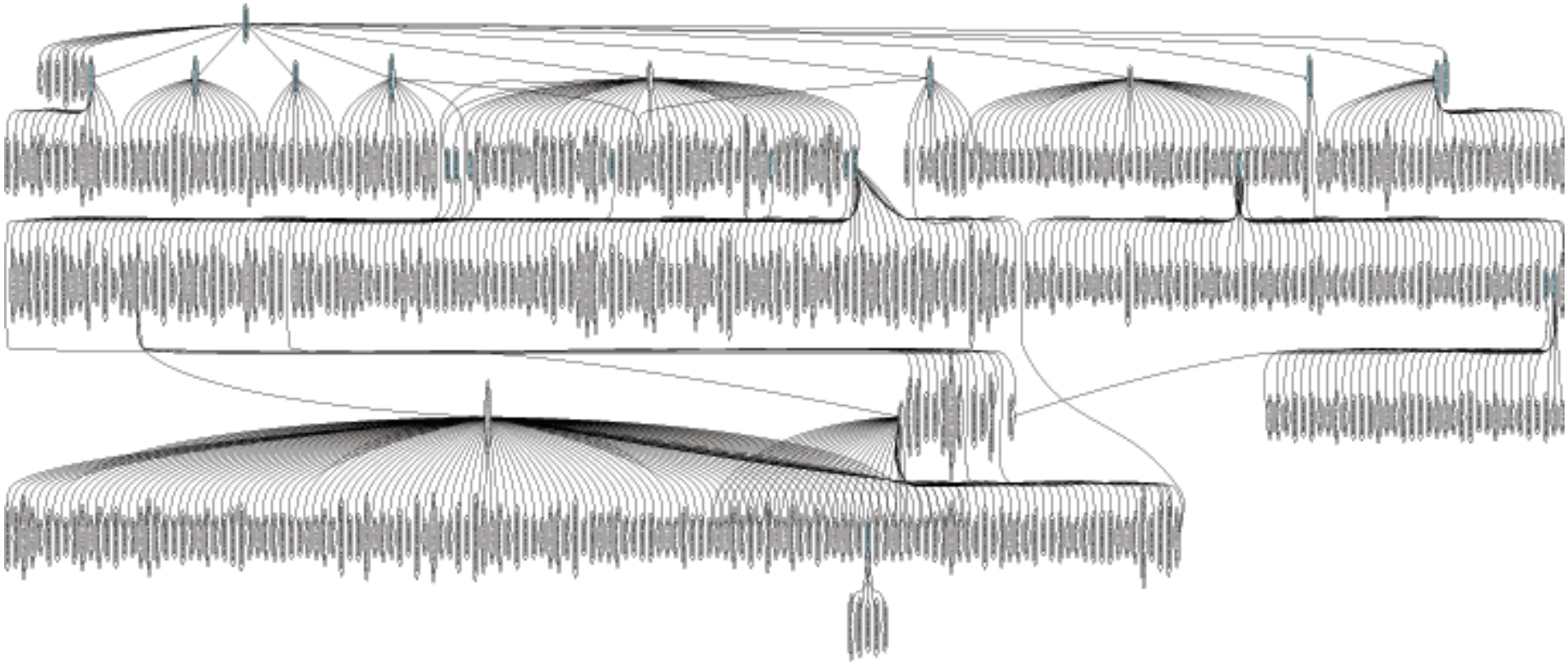
Courtesy of Robert Kueffner, NCGR, 2001

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

GO PORTION, HIERARCHICAL EYECHART

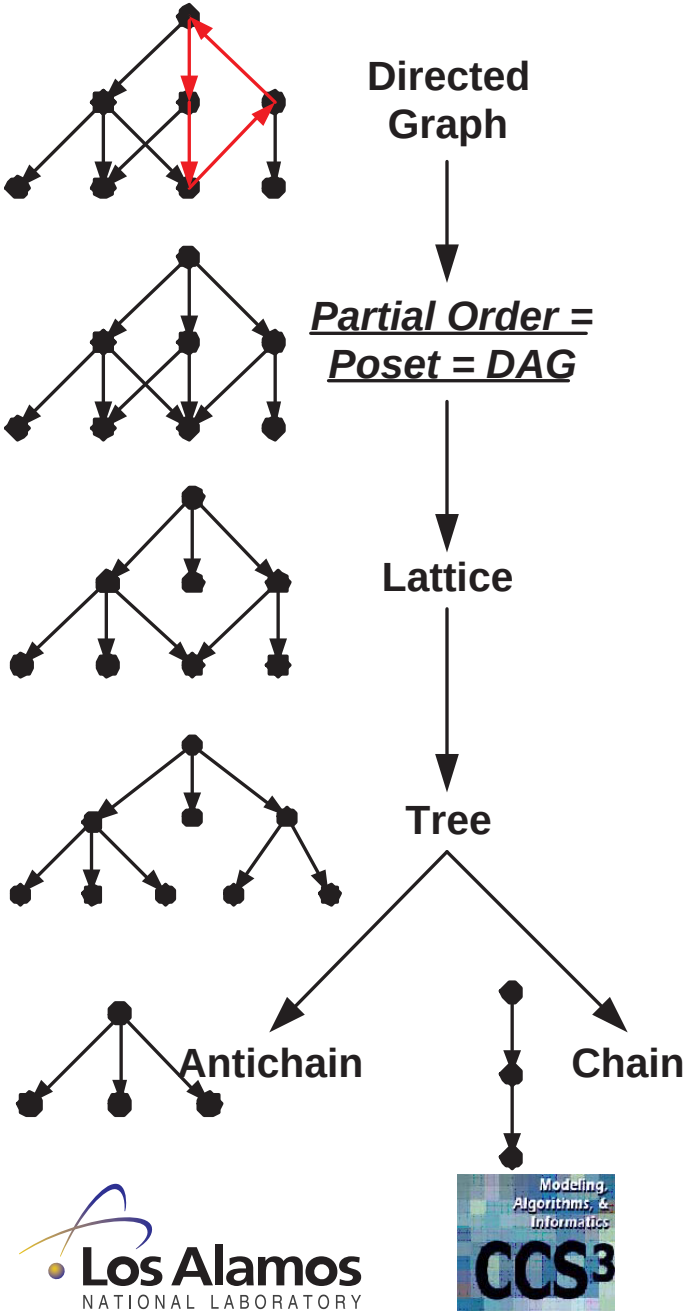


Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

HIERARCHIES AS PARTIALLY ORDERED SETS



- **Partial Order:** Set P; relation $\leq \subseteq P^2$: reflexive, anti-symmetric, transitive
- **Poset:** $\mathcal{P} = \langle P, \leq \rangle$
- Simplest mathematical structures which admit to descriptions in terms of "levels" and "hierarchies"
- More specific than graphs or networks: no cycles, equivalent to Directed Acyclic Graphs (DAGs)
- More general than trees, lattices: single nodes, pairs of nodes can have multiple parents
- Ubiquitous in knowledge systems: constructed, induced, empirical

Cliff Joslyn, joslyn@lanl.gov

BASIC POSET CONCEPTS

Poset: $\mathcal{P} = \langle P, \leq \rangle$

Comparable Nodes: $a \sim b := a \leq b$ or $b \leq a$

Up-Set: $\uparrow a = \{b \geq a\}$, **Down-Set:** $\downarrow a = \{b \leq a\}$

Chain: Collection of comparable nodes: $a_1 \leq a_2 \leq \ldots \leq a_n$

Height: Size maximal chain $\mathcal{H}(\mathcal{P})$

Noncomparable Nodes: $a \not\sim b$

Antichain: Collection of noncomparable nodes: $A \subseteq P, a \not\sim b, a, b \in A$

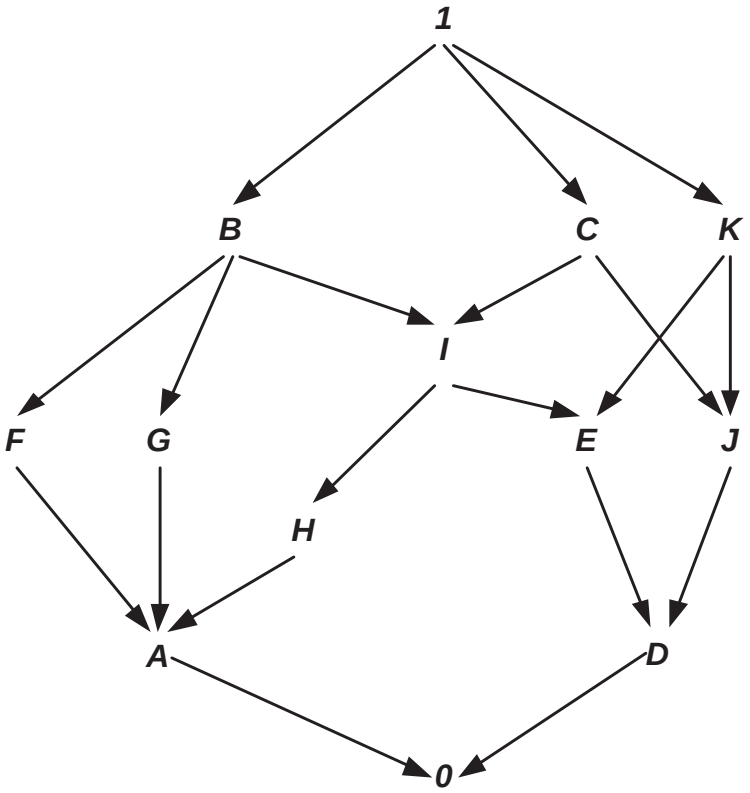
Width: Size maximal antichain $\mathcal{W}(\mathcal{P})$

Interval: $[a, b] := \{c \in P : a \leq c \leq b\}$, a bounded sub-poset of $\mathcal{P}$

Join, Meet: $a \vee b, a \wedge b \subseteq P$

Lattice: Then $a \vee b, a \wedge b \in P$

Bounded: Min $0 \in P$, Max $1 \in P$



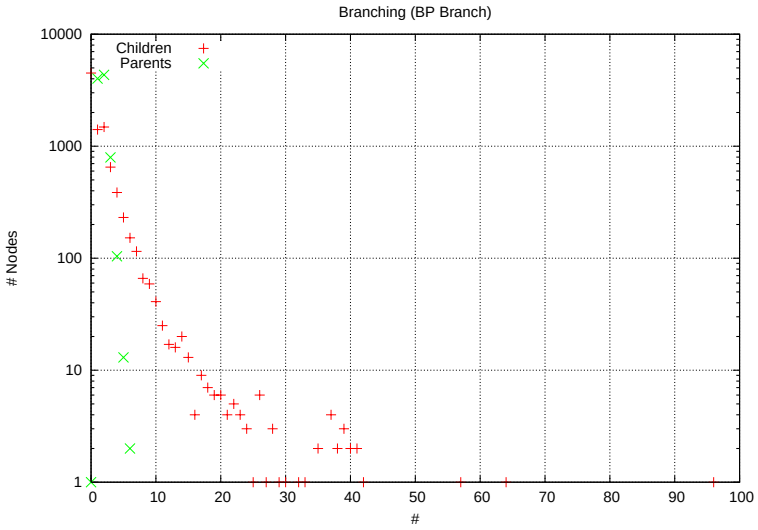
. Schröder, BS (2003): *Ordered Sets*, Birkhäuser, Boston

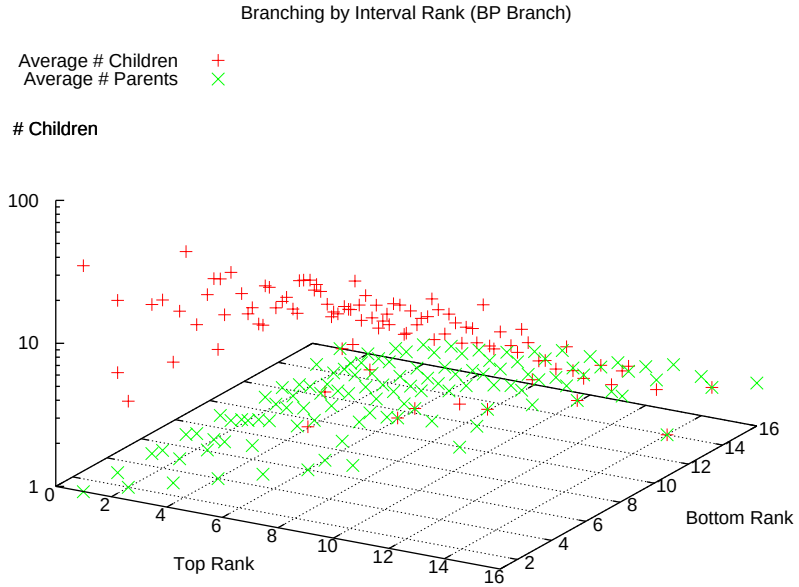
Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

SOME GO QUANTITATIVE MEASURES

	Nodes	Leaves	Interior	Edges	$\mathcal{H}$	$\mathcal{W}$
MF	7.0K	5.6K	1.3K	8.1K	13	$\geq$ 3.5K
BP	7.7K	4.1K	3.6K	11.8K	15	$\geq$ 2.9K
CC	1.3K	0.9K	0.4K	1.7K	13	$\geq$ 0.4K
GO	16.0K	10.6K	5.4K	21.5K	16	$\geq$ 5.9K





Joslyn, Cliff; Mniszewski, SM; Verspoor, KM; and JD Cohn: (2005) "Improved Order Theoretical Techniques for GO Functional Annotation", poster at *2005 Conf. on Intelligent Systems for Molecular Biology (ISMB 05)*

C Joslyn, S Mniszewski, A Fulmer, and G Heaton: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. **20**:s1, pp. 169-177

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, `joslyn@lanl.gov`

CHAIN DECOMPOSITION OF INTERVALS

Comparable Nodes: e.g. $D \leq 1 \in P$

Chain Decomposition: Set of all chains connecting them:

$$\begin{aligned}\mathcal{C}(D,1) &= \{C_j\} \\ &= \{D \prec E \prec I \prec B \prec 1, D \prec E \prec I \prec C \prec 1, \\ &\quad D \prec E \prec K \prec 1, D \prec J \prec C \prec 1, \\ &\quad D \prec J \prec K \prec 1\} \subseteq 2^P\end{aligned}$$

Chain Lengths: $h_j := |C_j| - 1$

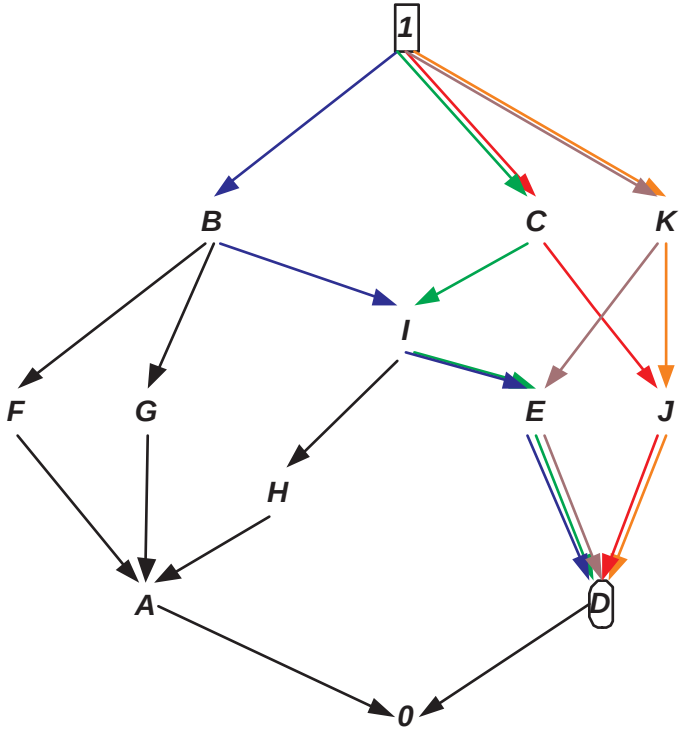
Vectors of Chain Lengths:

$$\vec{h}(a,b) := \left\langle h_j \right\rangle_{j=1}^{M} = \langle 4,4,3,3,3 \rangle$$

Extremes:

$$h_*(a,b) = \min_{h_j \in \vec{h}(a,b)} h_j = 3$$

$$h^*(a,b) = \max_{h_j \in \vec{h}(a,b)} h = 4$$

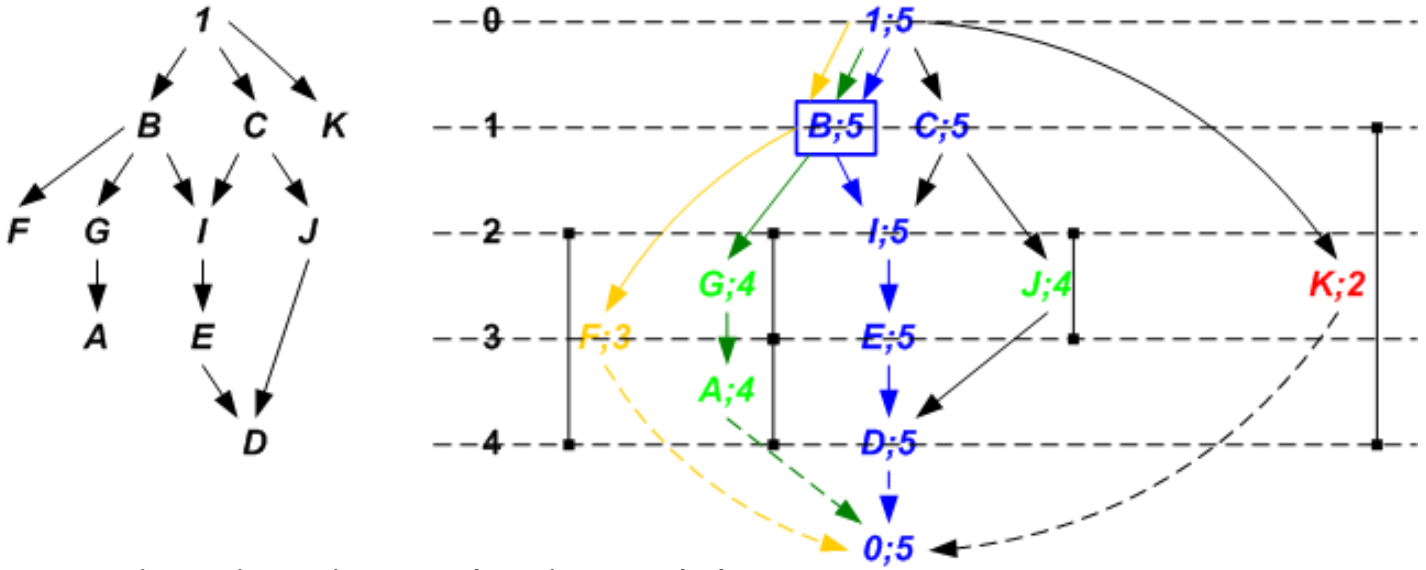


Los Alamos NATIONAL LABORATORY

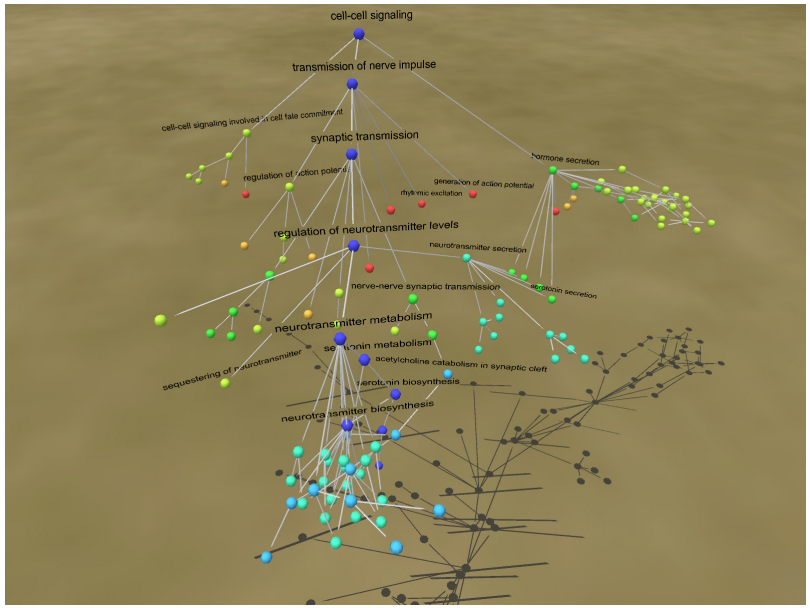
Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

INTERVAL RANK LAYOUT



- Interval valued vertical position (rank)
- Chain decomposition guides horizontal: short maximal chains to outside



CA Joslyn, SM Mniszewski, SA Smith, and PM Weber: (2006) "SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology", *Joint BioLINK and 9th Bio-Ontologies Meeting (JBB 06)*

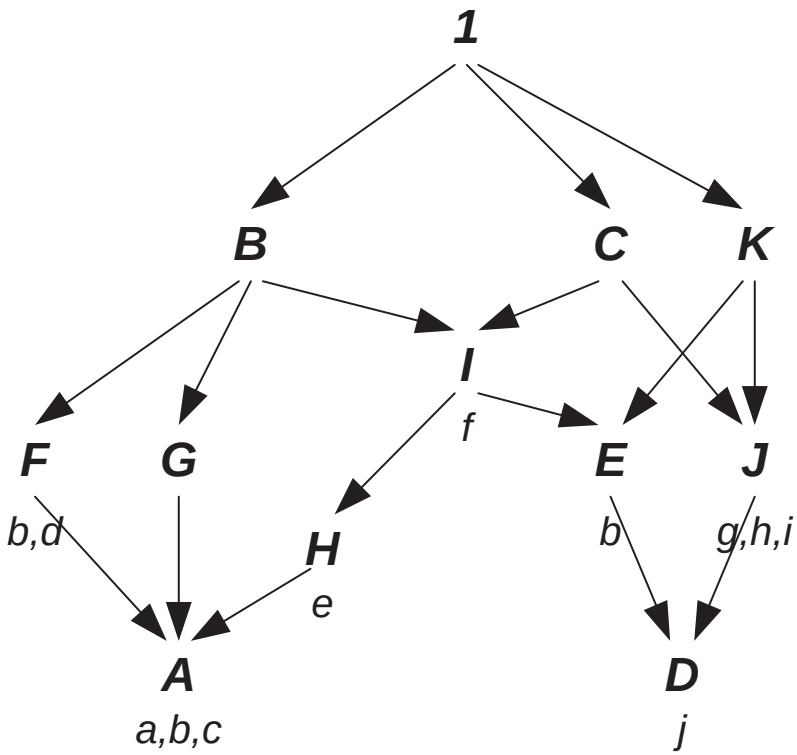
Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

CATEGORIZATION METHOD

- **POSO:** POSet Ontology
  $\mathcal{O} := \langle \mathcal{P}, X, F \rangle, \mathcal{P} = \langle P, \leq \rangle$
  Labels: finite, non-empty set X
  Labeling Function: $F: X \mapsto 2^P$
- Given labels (genes) $c, e, i \ldots$
- What node(s)
  $P = \{A, B, C, \ldots, K\}$ are best to pay attention to?



- Scores to rank-order nodes wrt/gene locations, balancing:
  - **Coverage:** Covering as many genes as possible
  - **Specificity:** But at the "lowest level" possible
- "Cluster" based on non-comparable high score nodes

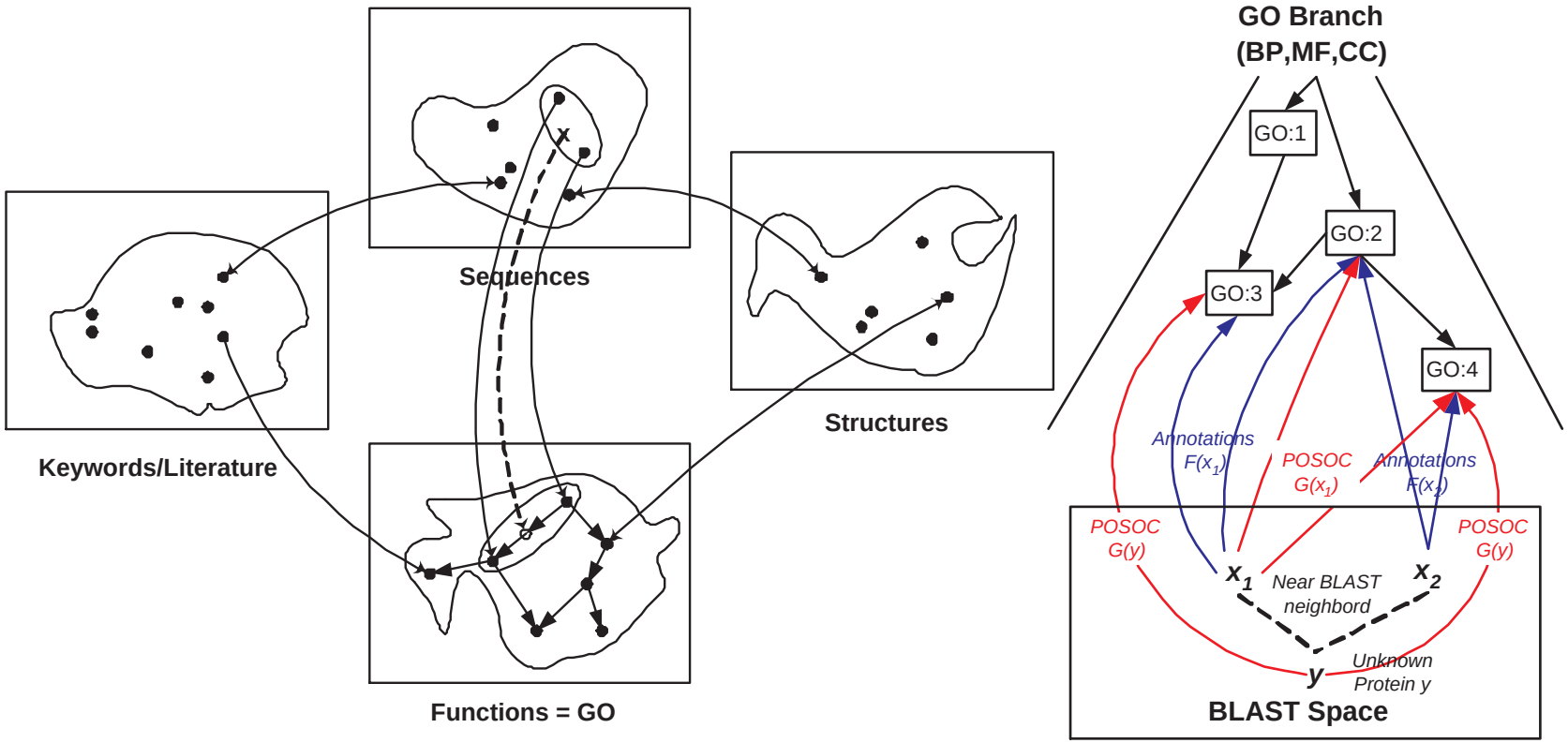
C Joslyn, S Mniszewski, A Fulmer, and G Heaton: (2004) "The Gene Ontology Categorizer", *Bioinformatics*, v. **20**:s1, pp. 169-177

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

AUTOMATED ONTOLOGICAL PROTEIN FUNCTION ANNOTATION



- Mappings among regions of biological spaces . . .
- . . . into spaces of biological functions
- POSOC annotated BLAST neighborhoods of new proteins
- *How to measure quality of inferred annotations?*

Verspoor, KM; Cohn, JD; Mniszewski, SM; and Joslyn, CA: (2006) "Categorization Approach to Automated Ontological Function Annotation", *Protein Science*, v. **15**, pp. 1544-1549

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

HIERARCHICAL EVALUATION METRICS

- Hierarchical measures:

Precision:

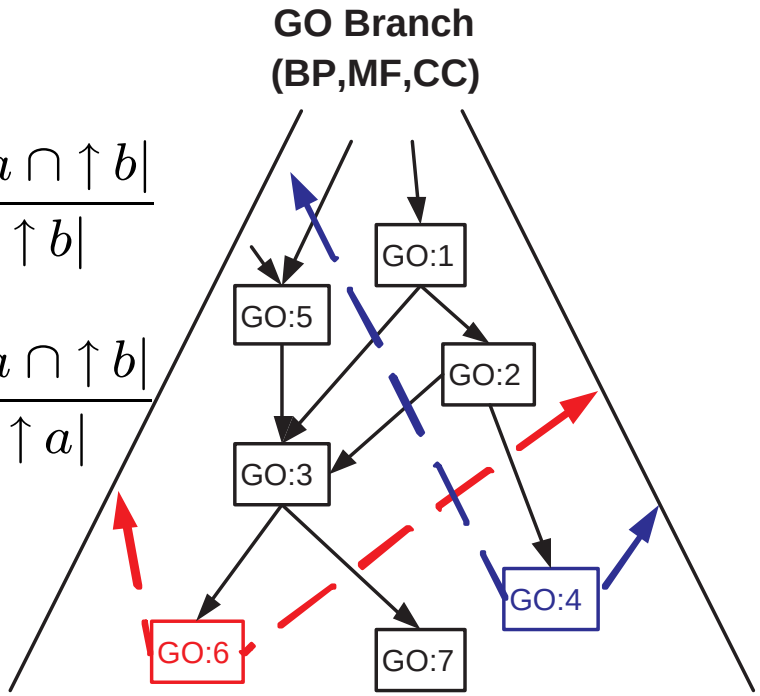
$$HP = \frac{1}{|G(x)|} \sum_{b \in G(x)} \max_{a \in F(x)} \frac{|\uparrow a \cap \uparrow b|}{|\uparrow b|}$$

Recall:

$$HR = \frac{1}{|F(x)|} \sum_{a \in F(x)} \max_{b \in G(x)} \frac{|\uparrow a \cap \uparrow b|}{|\uparrow a|}$$

F-Score:

$$HF = \frac{2(HP)(HR)}{HP + HR}$$



- Example: $F(x) = \{\text{GO:4}\}, G(x) = \{\text{GO:6}\}$
  $\uparrow a = \{\text{GO:1}, \text{GO:2}, \text{GO:4}\}, \uparrow b = \{\text{GO:1}, \text{GO:2}, \text{GO:3}, \text{GO:5}, \text{GO:6}\}$
  $HP = 2/5, HR = 3/5$

S Kiritchenko, S Matwin, and AF Famili: (2005) "Functional Annotation of Genes Using Hierarchical Text Categorization", *Proc. BioLINK SIG on Text Data Mining*

Verspoor, KM; Cohn, JD; Mniszewski, SM; and Joslyn, CA: (2006) "Categorization Approach to Automated Ontological Function Annotation", *Protein Science*, v. **15**, pp. 1544-1549

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

SEMANTIC SIMILARITIES

Poset $\mathcal{P} = \langle P, \leq \rangle$, probability distribution $p\colon P \mapsto [0,1], \sum_{a \in P} p(a) = 1$, "cumulative" $\beta(a) := \sum_{b \leq a} p(a)$

Resnik: $\delta(a,b) = \max_{c \in a \vee b} [-\log_2(\beta(c))]$

Lin:

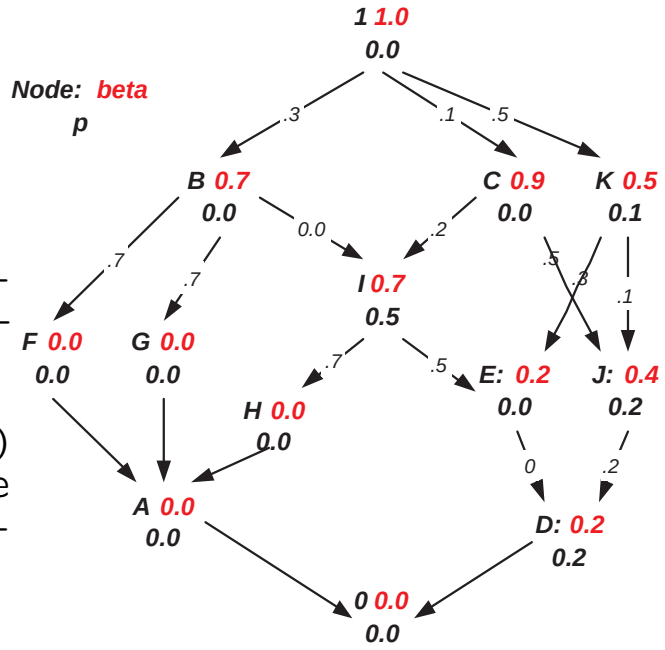
$$\delta(a,b) = \frac{2 \max_{c \in a \vee b} [\log_2(\beta(c))]}{\log_2(\beta(a)) + \log_2(\beta(b))}$$

Jiang and Conrath:

$$\delta(a,b) = 2 \max_{c \in a \vee b} [\log_2(\beta(c))] - \log_2(\beta(a)) - \log_2(\beta(b))$$

Issues:

- General mathematical grounding in poset metrics
- Not *rely* on probabilistic weighting



A Budanitsky and G Hirst: (2006) "Evaluating WordNet-based measures of semantic distance." *Computational Linguistics*, 32(1), 13–47.

Lord, PW; Stevens, Robert; Brass, A; and Goble, C: (2003) "Investigating Semantic Similarity Measures Across the Gene Ontology: the Relationship Between Sequence and Annotation", *Bioinformatics*, v. **10**, pp. 1275-1283

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

POSET METRICS

Assume $\langle P, \leq \rangle$ finite, connected, bounded

$$aub := \uparrow a \cap \uparrow b, \quad alb := \downarrow a \cap \downarrow b$$

Isotone Map: $v\colon P \mapsto \mathbb{R}, a \leq b \rightarrow v(a) \leq v(b)$

$$v^+(a,b) := \min_{w \in aub} v(w)$$

$$(aub)_v := \{w \in P : v(w) = v^+(a,b)\}$$

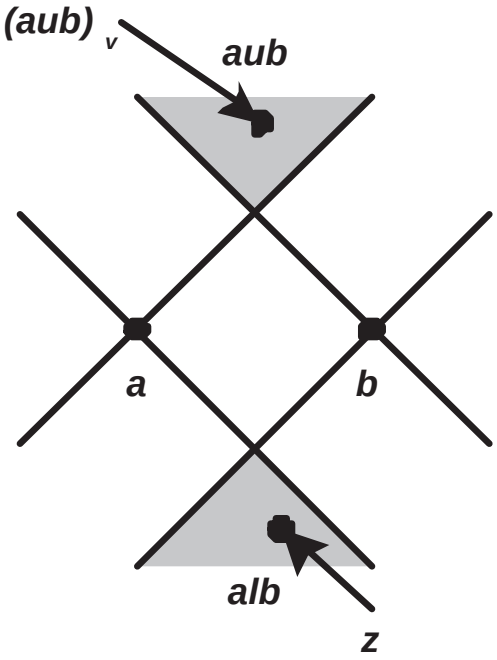
Upper Valuation: $\forall z \in alb$,

$$v(a) + v(b) \geq v^+(a,b) + v(z)$$

Distance: v is an upper valuation iff

$$d(a,b) := 2v^+(a,b) - v(a) - v(b)$$

is a distance (triangle inequality)



	Upper Valuation $z \in alb$	Lower Valuation $z \in aub$
Isotone	$v(a) + v(b) \geq v^+(a,b) + v(z)$ $d(a,b) = 2v^+(a,b) - v(a) - v(b)$	$v(a) + v(b) \leq v^-(a,b) + v(z)$ $d(a,b) = v(a) + v(b) - 2v^-(a,b)$
Antitone	$v(a) + v(b) \leq v^+(a,b) + v(z)$ $d(a,b) = v(a) + v(b) - 2v^+(a,b)$	$v(a) + v(b) \geq v^-(a,b) + v(z)$ $d(a,b) = 2v^-(a,b) - v(a) - v(b)$

Monjardet, B: (1981) "Metrics on Partially Ordered Sets - A Survey" , *Discrete Mathematics*, v. **35**, pp. 173-184

Los Alamos NATIONAL LABORATORY

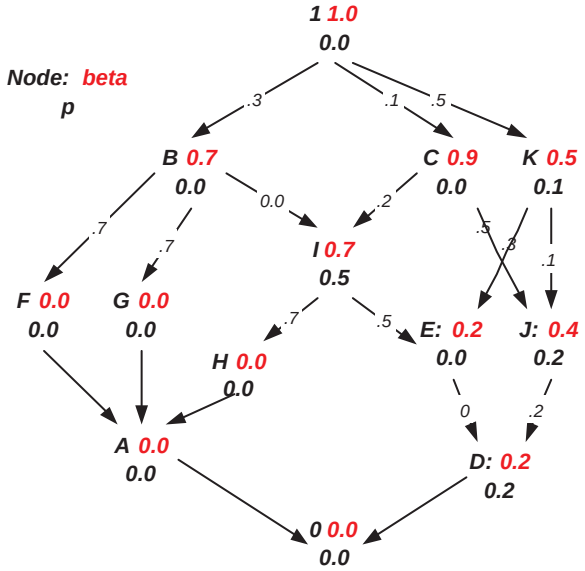
Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

SOME LATTICE METRICS

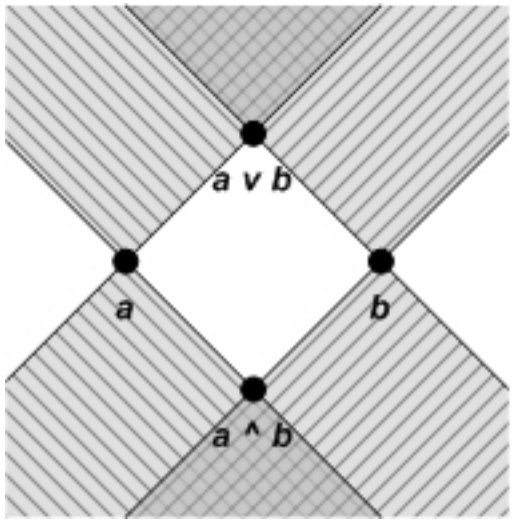
Information Theoretical: Monotone upper valuation

- Let $v(a) = \beta(a)$, "cumulative" probability
- **Proposition:** Jiang and Conrath is a metric, others are not
- $d(a,b) = 2\beta(a \vee b) - \beta(a) - \beta(b)$
- $d(I,J) = 1.53, d(E,J) = 1.64$



Purely Structural: Antitone upper valuation

- $|\uparrow a \cap \uparrow b| = |\uparrow(a \vee b)|$, $|\downarrow a \cap \downarrow b| = |\uparrow(a \wedge b)|$
- Let $v(a) = |\uparrow a|$
- $d(a,b) = |\uparrow a| + |\uparrow b| - 2|\uparrow a \cap \uparrow b|$
- $d(I,J) = 4, d(E,J) = 6$



Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

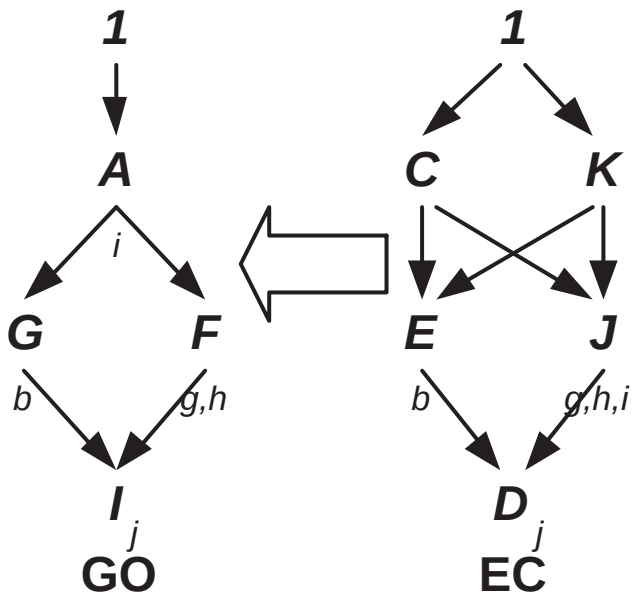
INTEROPERABILITY AND ALIGNMENT

Matching: Measure similarity between two regions of a single ontology

Comparing: Twist one ontology on a given term set into another ordering

Merging: Given two completely distinct ontologies:

- Identify structurally similar regions: intersection
- Create encompassing meta-ontologies: product or union?



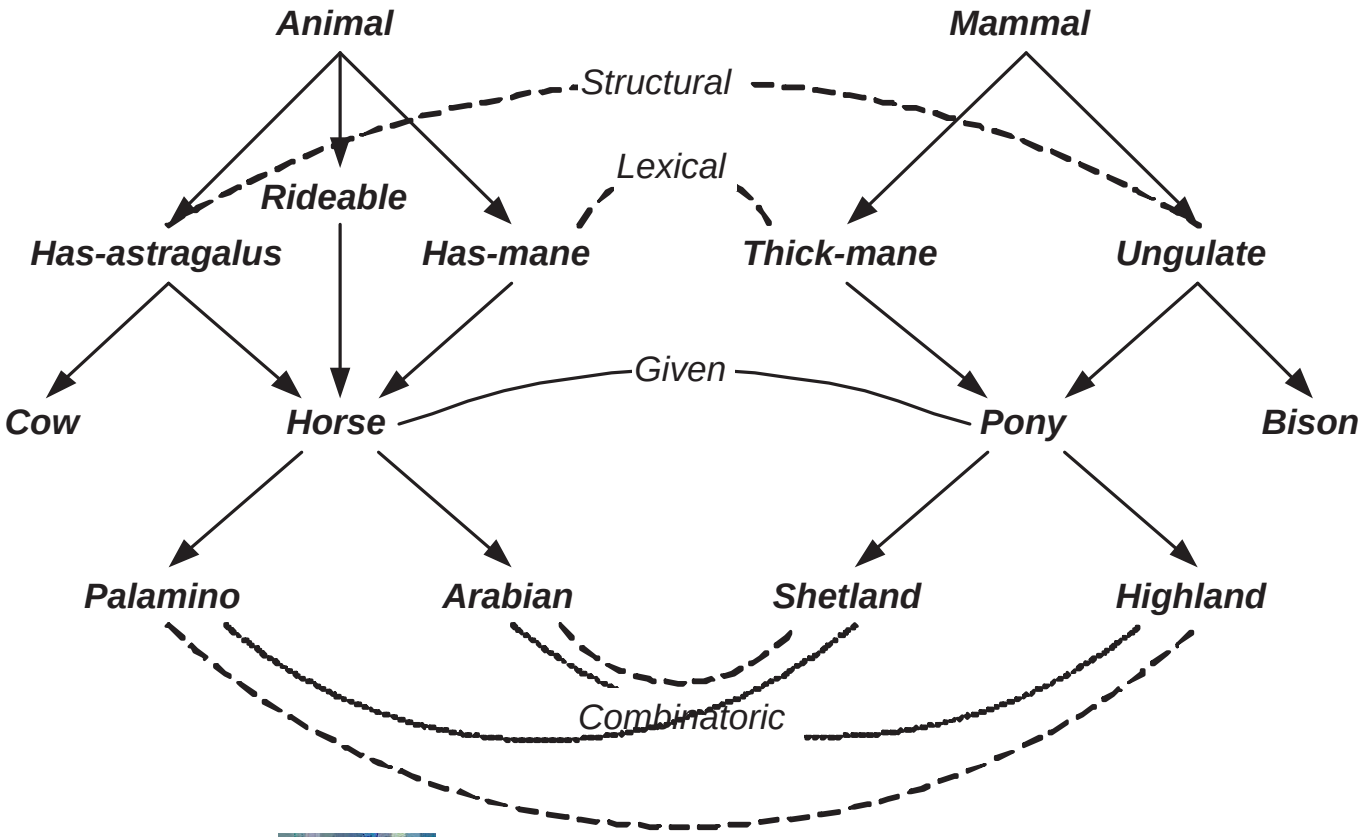
Joslyn, Cliff: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in: *Conceptual Structures at Work, Lecture Notes in Artificial Intelligence*, v. **3127**, ed. Wolff, Pfeiffer and Delugach, pp. 287-302, Springer-Verlag, Berlin

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

ALIGNMENT METHODS

Ultimate Goal: Construct order morhpisms
Neighborhoods: Around given anchors
Lexical: Matches
Structural: Nodes playing similar structural roles
Combinatoric: Sets of nodes playing similar structural roles
Poset Metrics: Measure candidate alignment, suggest new anchors



Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

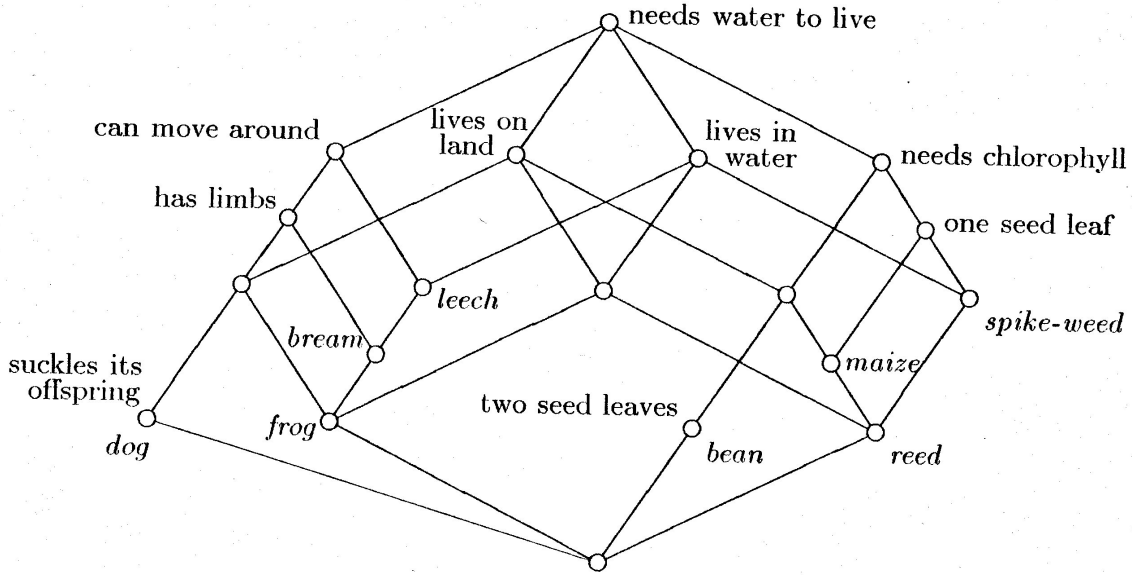
Cliff Joslyn, joslyn@lanl.gov

FORMAL CONCEPT ANALYSIS

- Semantic hierarchies from relational data
- Unbiased, graphical, visual representation
- Hypothesis and rule generation and evaluation
- For ontology induction, interoperability

		a	b	c	d	e	f	g	h	i
1	Leech	×	×					×		
2	Bream	×	×					×	×	
3	Frog	×	×	×				×	×	
4	Dog	×		×				×	×	×
5	Spike – weed	×	×		×		×			
6	Reed	×	×	×	×		×			
7	Bean	×		×	×	×				
8	Maize	×		×	×		×			

Figure 1.1 Context of an educational film "Living Beings and Water". The attributes are: a: needs water to live, b: lives in water, c: lives on land, d: needs chlorophyll to produce food, e: two seed leaves, f: one seed leaf, g: can move around, h: has limbs, i: suckles its offspring.



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

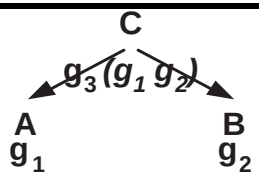
Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

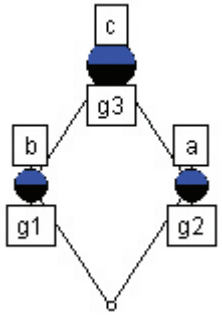
Cliff Joslyn, joslyn@lanl.gov

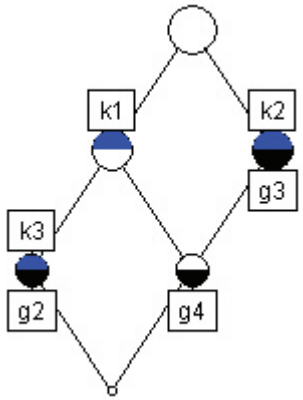
FCA ONTOLOGY MERGER, INDUCTION

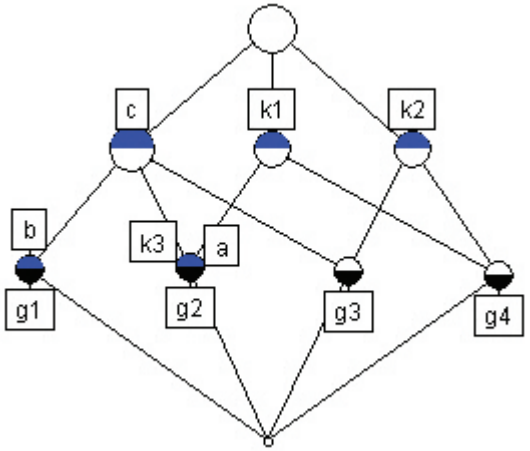
- $\{g_1, g_2, g_3\}$: annotated into an ontology O:



- $\{g_2, g_3, g_4\}$: annotated to keywords $K = \{k_1, k_2, k_3\}$
- *Induce* order on K while *incoporating* order on O
- Amenable to metric treatment of attributes, objects







	a	b	c
g_1		$\surd$	$\surd$
g_2	$\surd$		$\surd$
g_3			$\surd$

	k_1	k_2	k_3
g_2	$\surd$		$\surd$
g_3		$\surd$	
g_4	$\surd$	$\surd$	

	a	b	c	k_1	k_2	k_3
g_1		$\surd$	$\surd$			
g_2	$\surd$		$\surd$	$\surd$		$\surd$
g_3			$\surd$		$\surd$	
g_4				$\surd$	$\surd$	

Gessler, DDG, CA Joslyn, KM Verspoor: (2007) "Knowledge Integration in Open Worlds: Exploiting the Mathematics of Hierarchical Structure", in preparation for ICSC 07

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

ACKNOWLEDGEMENTS, COLLABORATIONS, AND OTHER ASSORTED NAME-DROPPING

LANL Info. Sciences:

- *Susan Mniszewski*
- *Chris Orum*
- *Karin Verspoor*
- Michael Wall

LANL Elsewhere:

- *Judith Cohn*
- Bill Bruno
- Steve Smith

U. West Indies:

- Jonathan Farley

PNNL: Joe Oliveira

U. Newcastle: Phillip Lord

NCGR: Damian Gessler

Technische Universität Dresden:

- Stephan Schmidt
- Tim Kaiser
- Bjoern Koester

New Mexico State U.:

- Alex Pogel

P&G: Andy Fulmer

Stanford Medical Informatics:

- Natasha Noy

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3