Order Metrics for Semantic Knowledge Systems

Cliff Joslyn and Emilie Hogan

Pacific Northwest NATIONAL LABORATORY

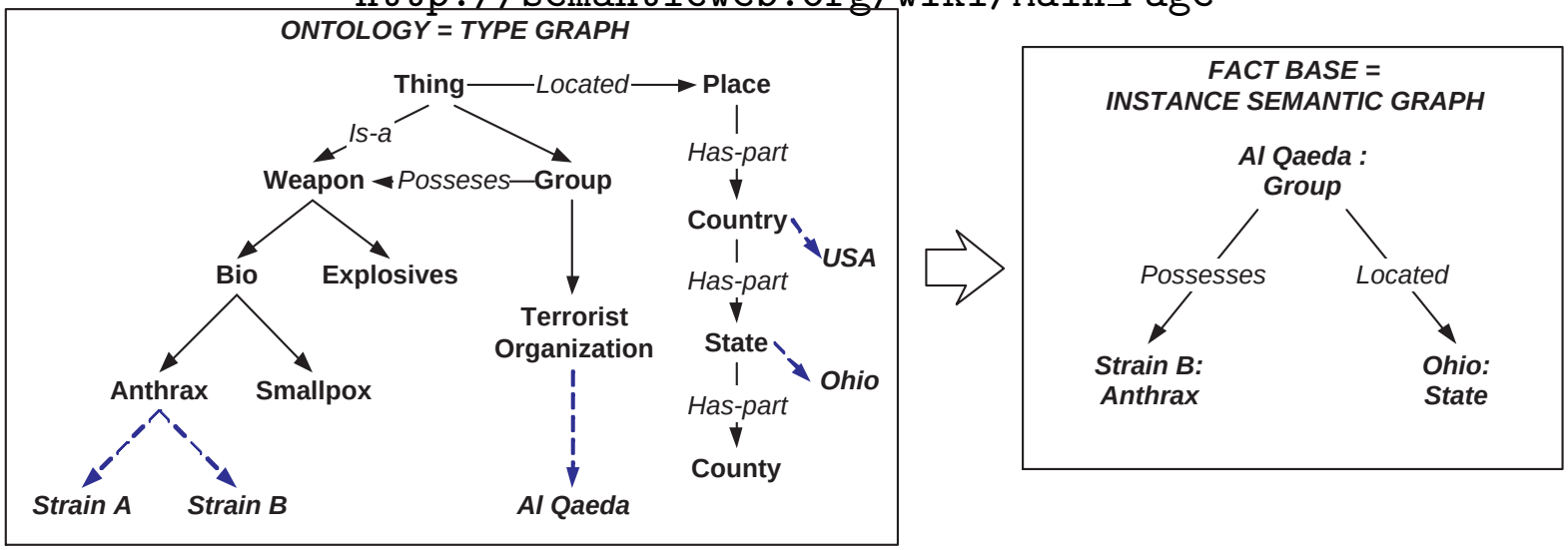
THE STATE UNIVERSITY OF NEW JERSEY RUTGERS

5th Int. Conf. on Hybrid AI Systems (HAIS 2010)

June, 2010

SEMANTIC GRAPH DATABASES

http://semanticweb.org/wiki/Main_Page

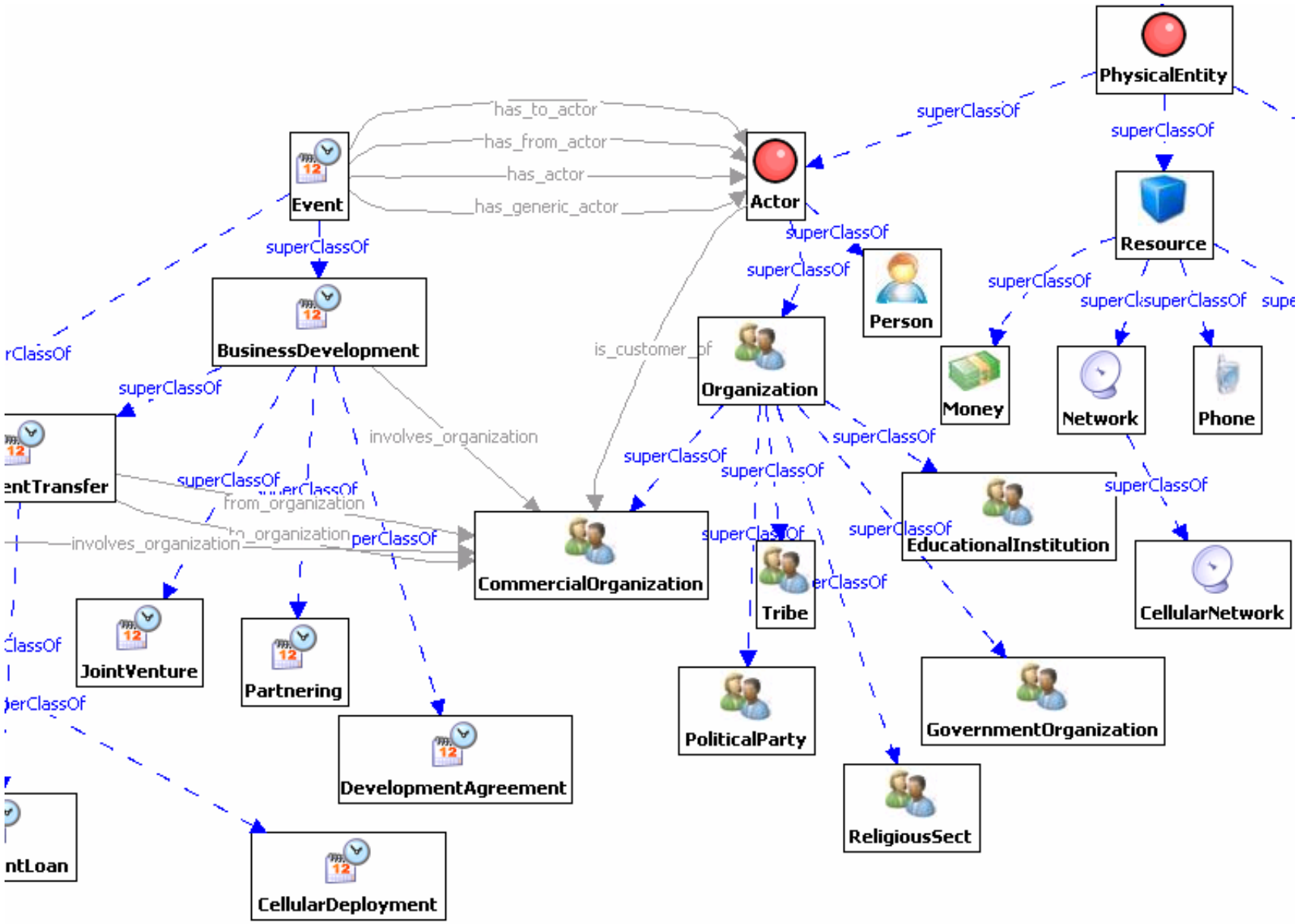


- Semantic web movement:
  - Legacy and repository of classical AI dreams
  - Propositional knowledge representation on the web
- New database paradigm for distributed data:
  - OWL, RDF, Semantic Web movement, triple store databases
  - For graph queries: seeking connections, patterns
- Increasing size and prominence: $10^9 - 10^{11}$ triples
- Linked data movement, computational science, computational biology, intelligence analysis

Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, cjoslyn@pnl.gov

ONTOLOGY FRAGMENT



Cliff Joslyn, cjoslyn@pnl.gov

Pacific Northwest
NATIONAL LABORATORY

SEMANTIC HIERARCHIES

Ubiquitous, universal: Conceptual generality and specificity

- Subsumptive, inheritance taxonomies: `is-a`
- Merenomic, compositional: `part-of`
- Inferential: `follows-from`

Claim: Cores of ontologies, computational lexicons

- Comprise the bulk of links in real-world ontological knowledgebases: $80 - 90\%$ of links
- Becoming large: $10^5 - 10^7$

Challenge: Identify/establish technology for key tasks:

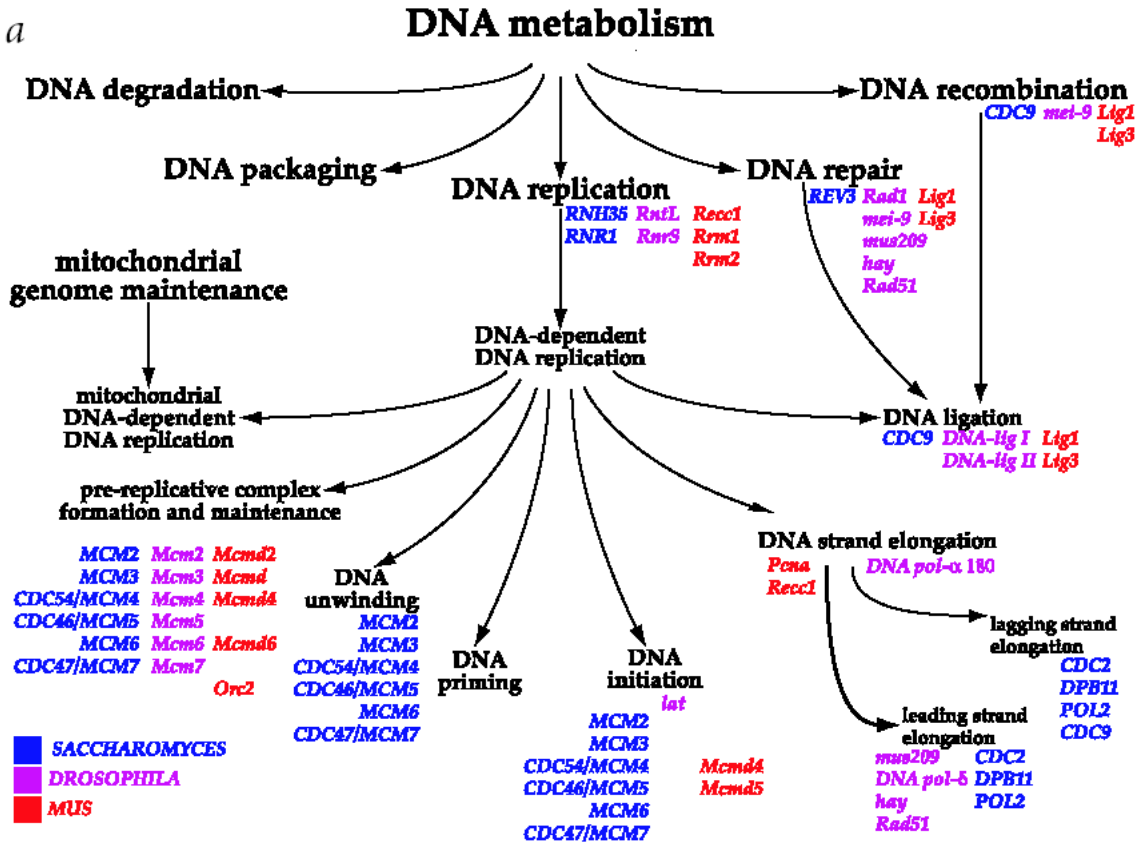
- Classification, categorization, clustering
- Induction from source data
- Navigation, anomaly detection
- Visualization
- Link analysis, search, retrieval
- Merger, linkage, interoperability

Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, `cjoslyn@pnl.gov`

GENE ONTOLOGY (GO): DNA METABOLISM PORTION

- Taxonomic controlled vocabulary
- $\sim 30K$ nodes populated by genes, proteins
- Is-a and part-of links
- **Annotation:** Assign thousands of proteins to nodes



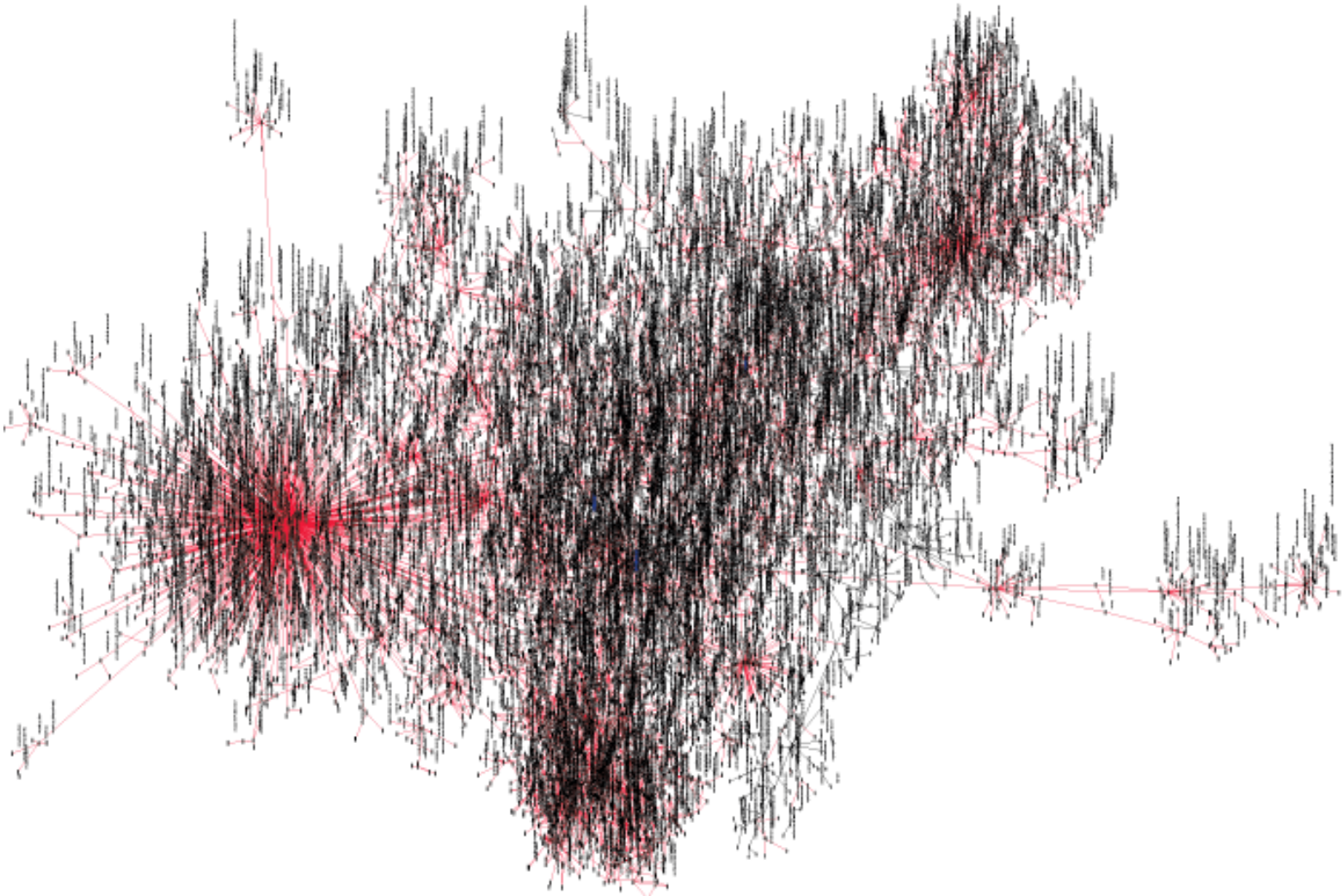
Gene Ontology Consortium (2000): "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, 25:25-29

- Tremendous community resource: large, semantically rich, validated, middle ontology, first in major use

Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, cjoslyn@pnl.gov

WHOLE GO CA. 2001

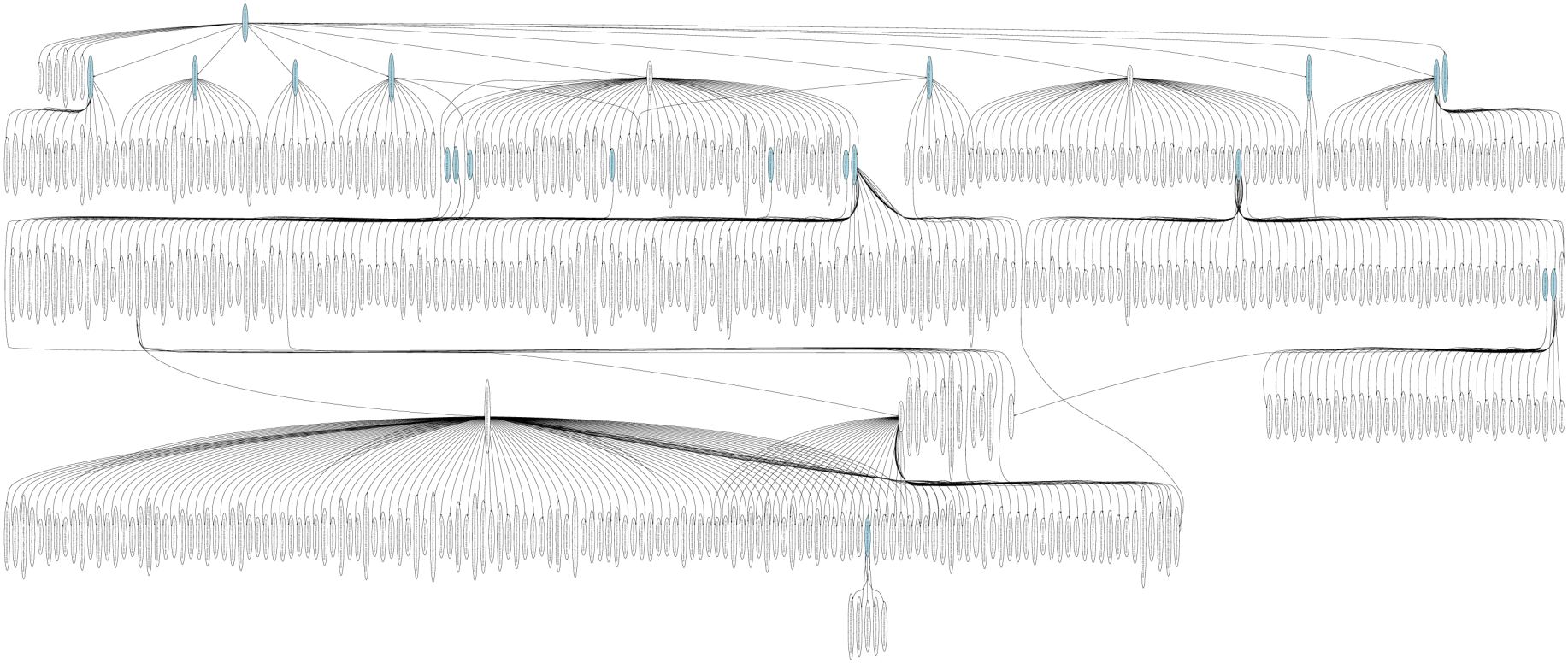


Courtesy of Robert Kueffner, NCGR, 2001

Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, cjoslyn@pnl.gov

GO PORTION



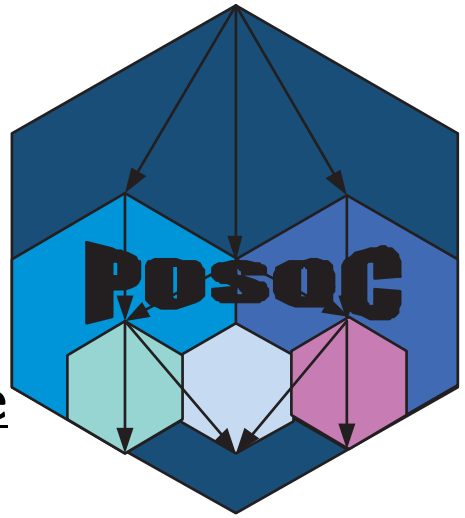
Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, cjoslyn@pnl.gov

CATEGORIZATION IN THE GENE ONTOLOLGY

http://www.ccs3.lanl.gov/posoc

- 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?
- Clustering, fundamentally about **distance** among annotated GO nodes
- 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

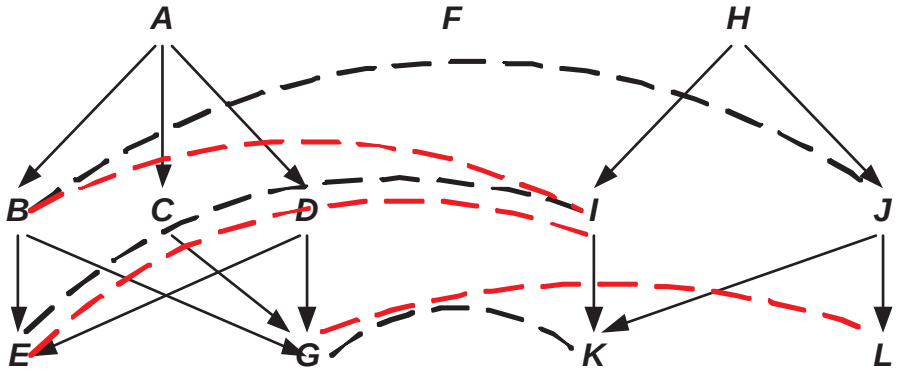
Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, cjoslyn@pnl.gov

ONTOLOGY MATCHING

http://www.ontologymatching.org

- Now a major activity:
  - Ontology Alignment Evaluation Initiative (OAEI)
  - Contest running at Int. Semantic Web Conf. (ISWC) for many years
- Identify target ontologies, e.g.:
  - UN Food and Agriculture Org. thesaurus (28K multilingual terms)
  - US National Agricultural Library Thesaurus (42K English terms)
- Build "gold standard" mapping by hand



- Measure precision and recall (exact matches) of submitted alignments

Euzenat, Jérôme and Shvaiko, P: (2007) *Ontology Matching*, Springer-Verlag, Hiedelberg

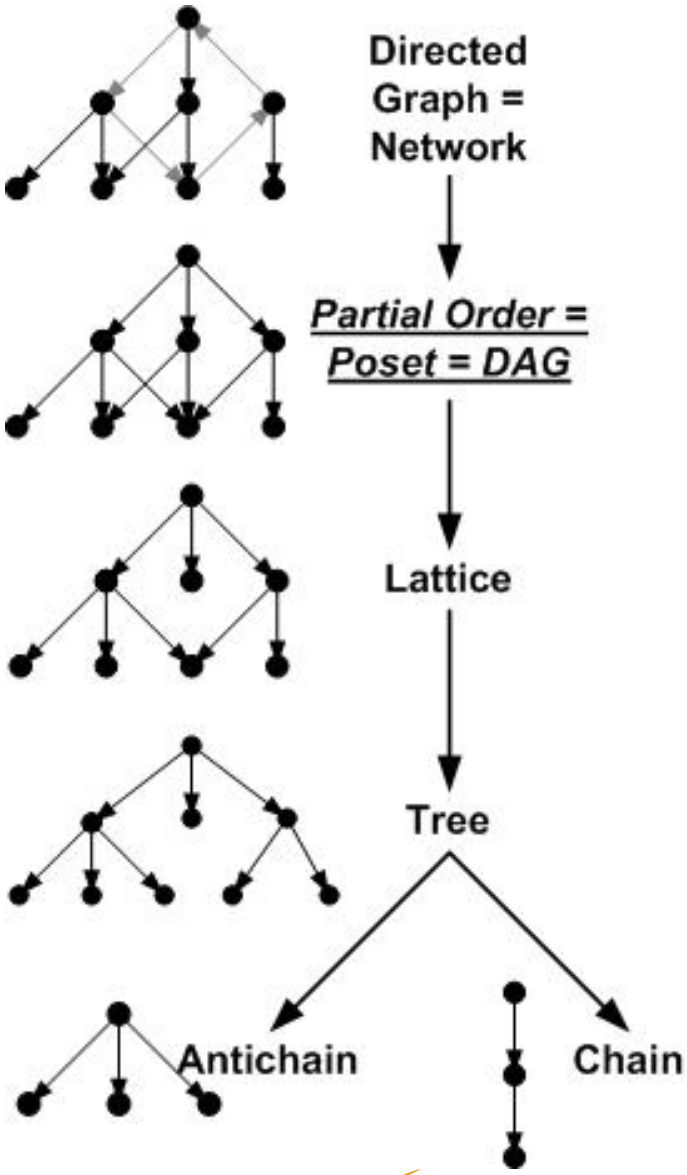
Joslyn, Cliff: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in: *Conceptual Structures at Work, LNAI*, v. **3127**, ed. Wolff, Pfeiffer and Delugach, pp. 287-302, Springer-Verlag, Berlin

Joslyn, Cliff; Paulson, Patrick; and White, Amanda: (2009) "Measuring the Structural Preservation of Semantic Hierarchy Alignments", in: *Proc. 4th Int. Wshop. on Ontology Matching (OM-2009), CEUR*, v. **551**, http://ceur-ws.org/Vol-551/om2009_Tpaper6.pdf

Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, cjoslyn@pnl.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

Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, cjoslyn@pnl.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: $A \subseteq P, a \not\sim b, a, b \in A$

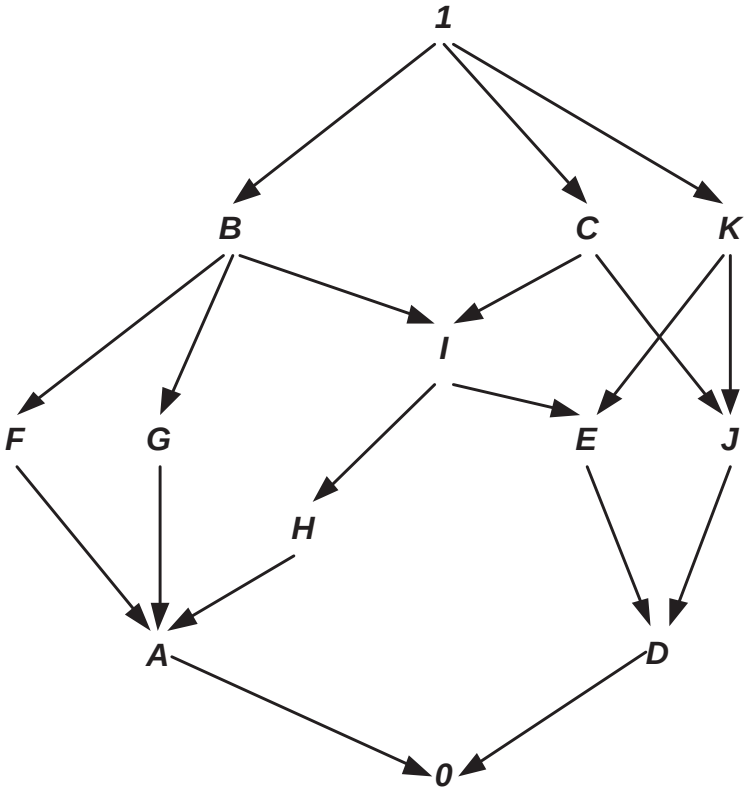
Width: Size maximal antichain $\mathcal{W}(\mathcal{P})$

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

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

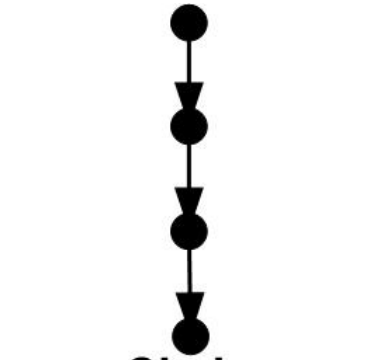
Lattice: Then $a \vee b, a \wedge b \in P$

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

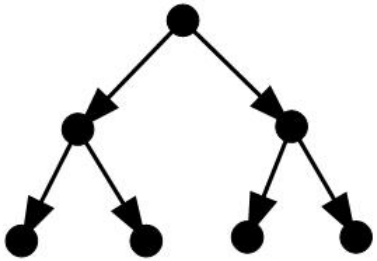


Davey, BA and Priestly, HA: (1990) *Introduction to Lattices and Order*, Cambridge UP, Cambridge UK, 2nd Edition

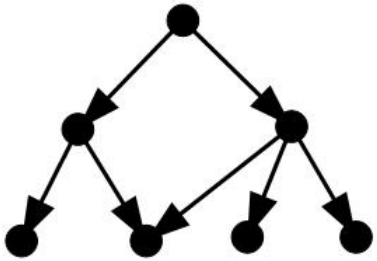
ASPECTS OF ORDERS



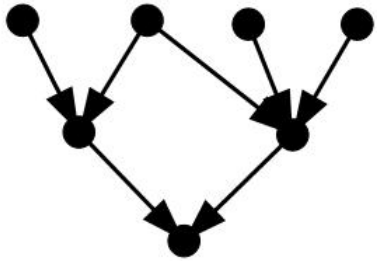
Chain
(Totally ordered)
(Unique parents
and children)



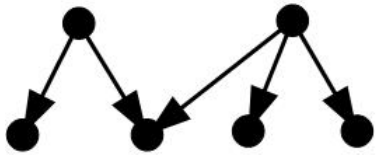
Tree
(Unique Parents)



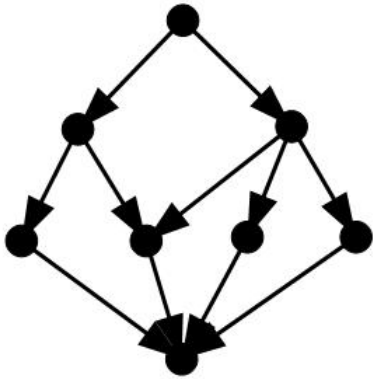
Multiple
Inheritance



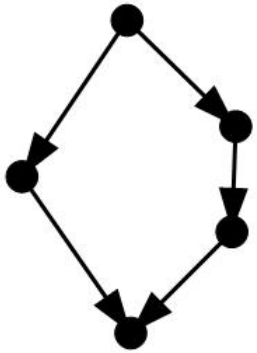
Dual
Structure



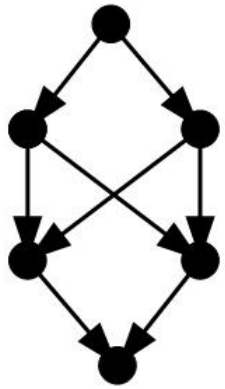
Totally
Unbounded



Totally
Bounded



Ungraded
(Unequal chain lengths)
(Ambiguous levels)



Not a lattice
(Pairs with
multiple parents)

Pacific Northwest
NATIONAL LABORATORY

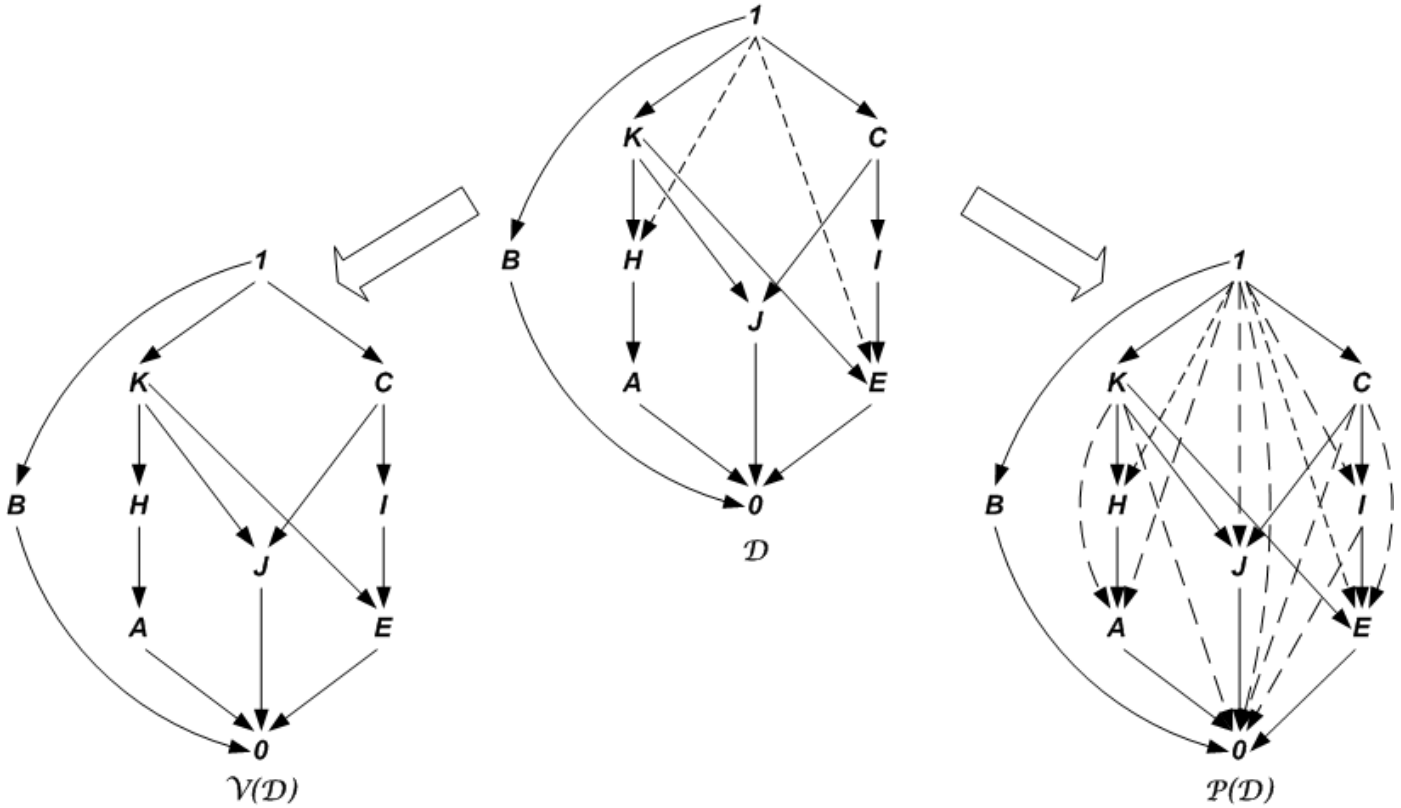
Cliff Joslyn, cjoslyn@pnl.gov

DAG STRUCTURES: POSETS AND COVERS

- Assume a Directed Acyclic Graph $\mathcal{D} = \left\langle P, E \subseteq P^2 \right\rangle$
- **Poset** $\mathcal{P}(\mathcal{D}) = \langle P, \leq \rangle$ by transitive closure
- **Cover** $\mathcal{V}(\mathcal{D}) = \langle P, \prec \rangle$ by transitive reduction: Hasse diagram
- **Degree of transitivity:** measure of "density" of $\mathcal{D}$

$$TR(\mathcal{D}) := \frac{|\mathcal{D} \setminus \mathcal{V}(\mathcal{D})|}{|\mathcal{P}(\mathcal{D}) \setminus \mathcal{V}(\mathcal{D})|} \in [0, 1]$$

- Each poset/cover determines an equivalence class of DAGs



Pacific Northwest
NATIONAL LABORATORY

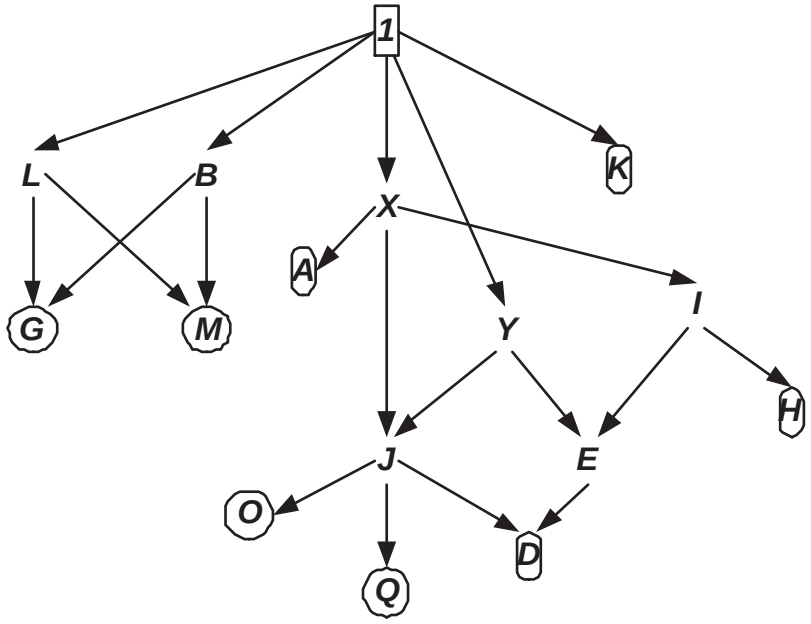
Cliff Joslyn, cjoslyn@pnl.gov

ORDER INTERVAL RANK

Motivation: Proper vertical positioning of nodes

Observation: All children of root are same distance from root, but if *also* leaves should be further down

Technique: Exploit vertical distance from *both* upper and lower bounds



Max Chain Length: $h^*(a,b)$

Order Top Rank: Max chain length a to top: $r^t(a) := h^*(a,1)$

Order Bottom Rank: Height minus max chain length bottom to a: $r^b(a) := h^*(0,1) - h^*(0,b)$

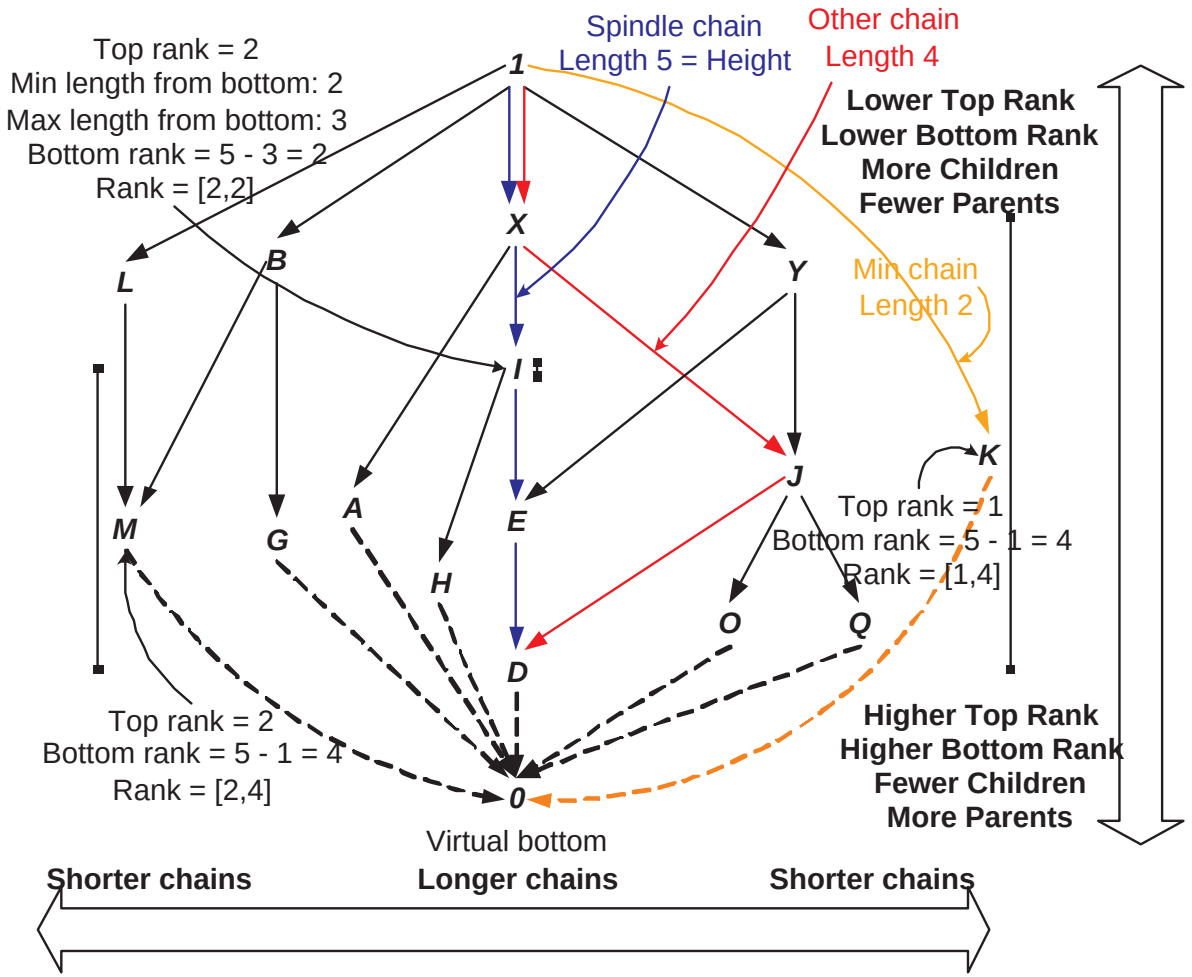
Order Interval Rank: $\bar{R}(a) := [r^t(a), r^b(a)]$

Order Rank Width: $W(a) := \|\bar{R}(a)\| = r^b(a) - r^t(a)$

Order Rank Midpoint: $\hat{W}(a) = \frac{r^b(a) - r^t(a)}{2}$

CA Joslyn, A Pogel, and S Schmidt: "Ordered Set Interval Rank for Knowledge Systems Analysis and Visualization", in preparation

CHAIN DECOMPOSITION HIERARCHICAL LAYOUT

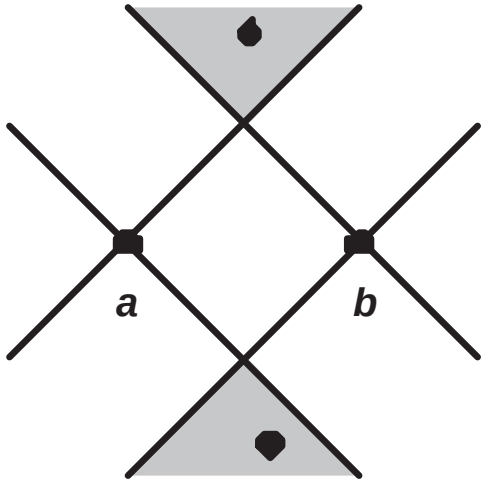


- **Spindle:** Max length max chains connecting 0,1
- All rank widths zero
- Maximal graded sub-poset (has rank function)
- Only portion which is Jordan-Dedekind
- Classical ordered structures all spindle

CA Joslyn, SM Mniszewski, SA Smith, and PM Weber: (2006) "SpindleViz: A Three Dimensional, Order Theoretical Visualization Environment for the Gene Ontology", in: *Joint BioLINK and 9th Bio-Ontologies Meeting (JBB 06)*, `http://www.bio-ontologies.org.uk/2006/download/Joslyn2EtAlSpindleviz.pdf`

ORDER METRICS

- Semimodular functions on finite, bounded posets $\langle P, \leq \rangle$
- E.g. generalized Kolmogorov probabilities, information measures in graphical models (Studeny 2005)



Generalized Join, Meet: $a \nabla b := \uparrow a \cap \uparrow b, a \Delta b := \downarrow a \cap \downarrow b$

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

$$v^{\nabla}(a,b) := \min_{c \in a \nabla b} v(c), \quad v_{\Delta}(a,b) := \max_{c \in a \Delta b} v(c)$$

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

$$v^{\nabla}(a,b) := \max_{c \in a \nabla b} v(c), \quad v_{\Delta}(a,b) := \min_{c \in a \Delta b} v(c)$$

Valuations: Super- and sub-modularity:

Lower: $v(a) + v(b) \leq v^{\nabla}(a,b) + v_{\Delta}(a,b)$

Upper: $v(a) + v(b) \geq v^{\nabla}(a,b) + v_{\Delta}(a,b)$

Metric $d(a,b)$	Lower Valuation $v(a) + v(b) \leq v^{\nabla}(a,b) + v_{\Delta}(a,b)$	Upper Valuation $v(a) + v(b) \geq v^{\nabla}(a,b) + v_{\Delta}(a,b)$
Isotone	$v(a) + v(b) - 2v_{\Delta}(a,b)$	$2v^{\nabla}(a,b) - v(a) - v(b)$
Antitone	$v(a) + v(b) - 2v^{\nabla}(a,b)$	$2v_{\Delta}(a,b) - v(a) - v(b)$

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

Orum, Chris and Joslyn, Cliff A: (2009) "Valuations and Metrics on Partially Ordered Sets", `http://arxiv.org/abs/0903.2679v1`, submitted

LATTICE METRICS

Let $\mathcal{P} = \langle P, \leq \rangle$ be a lattice

Semimodular Valuations: v is sub- or super-semimodular:

$$v(a) + v(b) \geq v(a \vee b) + v(a \wedge b)$$
$$v(a) + v(b) \leq v(a \vee b) + v(a \wedge b)$$

Cases: Also depending on whether v is monotone or antitone

$$d(a,b) = 2v(a \vee b) - v(a) - v(b)$$
$$d(a,b) = 2v(a \wedge b) - v(a) - v(b)$$
$$d(a,b) = v(a) + v(b) - 2v(a \vee b)$$
$$d(a,b) = v(a) + v(b) - 2v(a \wedge b)$$

Dempster-Shafer Evidence Theory:

- Boolean lattice of states $2^{\Omega}, A, B \subseteq \Omega$
- Bel, Pl: $2^{\Omega} \mapsto [0,1]$ super- (sub-) modular
  $$\text{Bel}(A \cup B) \leq \text{Bel}(A) + \text{Bel}(B) - \text{Bel}(A \cap B),$$
  $$\text{Pl}(A \cup B) \geq \text{Pl}(A) + \text{Pl}(B) - \text{Pl}(A \cap B)$$
- Pr: $2^{\Omega} \mapsto [0,1]$ fully modular, Kolmogorov measure
  $$\text{Pr}(A \cup B) = \text{Pr}(A) + \text{Pr}(B) - \text{Pr}(A \cap B)$$
  $$d(A,B) = \text{Pr}(A \cup B) - \text{Pr}(A \cap B)$$

SOME STANDARD VALUATIONS

- Any quantified poset: $t\colon P \mapsto \mathbb{R}$
- Lower valuations:

$$v^*(a) := \sum_{b \in \uparrow a} t(b) = \sum_{b \geq a} t(b); \quad v_*(a) := \sum_{b \in \downarrow a} t(b) = \sum_{b \leq a} t(b)$$

- **Cardinality of Filters, Ideals:** $t(a) \equiv 1$

$$v^*(a) = |\uparrow a|, v_*(a) = |\downarrow a|$$

- **Probabilities:** $t\colon P \mapsto [0,1], \sum_{a \in P} t(a) = 1$

$$\beta(a) := \sum_{b \leq a} t(a)$$

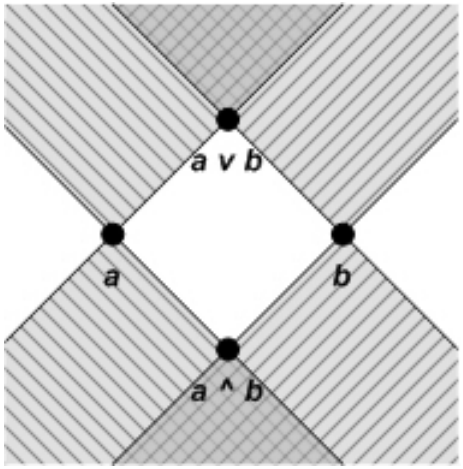
a kind of "cumulative"

Orum, Chris and Joslyn, Cliff A: (2009) "Valuations and Metrics on Partially Ordered Sets", `http://arxiv.org/abs/0903.2679v1`, submitted

SOME LATTICE METRICS

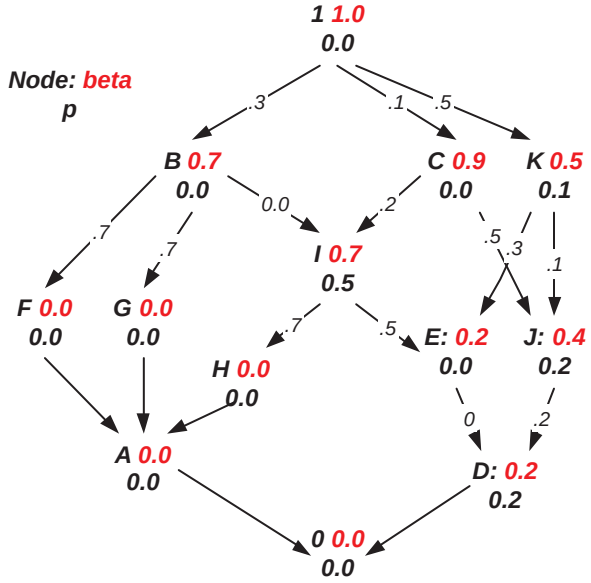
Purely Structural: Antitone upper valuation

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



Information Theoretical: Monotone upper valuation

- Let $v(a) = \beta(a) = \sum_{b \leq a} t(a)$, "cumulative" probability
- $d_p(a,b) = 2\beta(a \vee b) - \beta(a) - \beta(b)$
- $d_p(I,J) = 1.53, d_p(E,J) = 1.64$

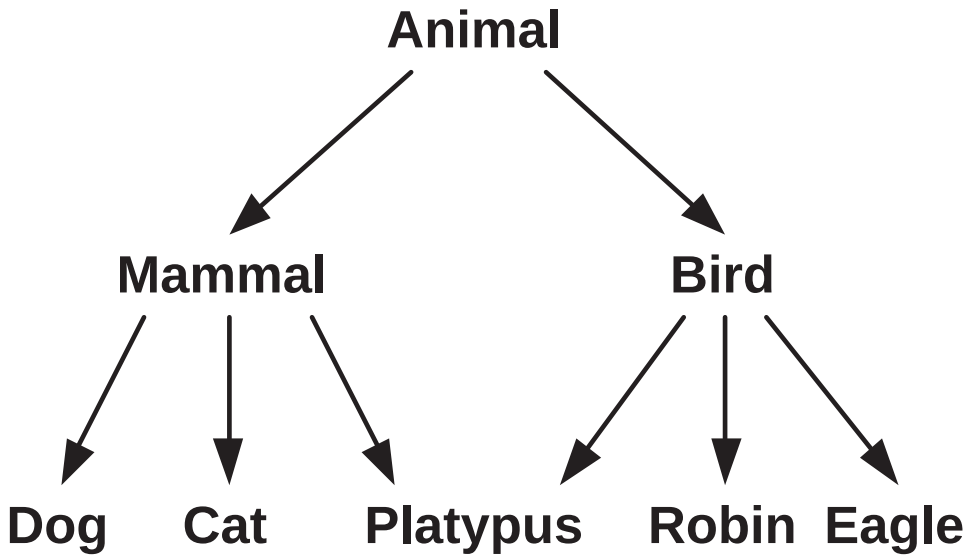


Orum, Chris and Joslyn, Cliff A: (2009) "Valuations and Metrics on Partially Ordered Sets", `http://arxiv.org/abs/0903.2679v1`, submitted

Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, `cjoslyn@pnl.gov`

COMPARING DISTANCES



	$d^*(\cdot)$	$d_*(\cdot)$	Min Path
Mammal, Bird	2	6	2
Dog,Cat	2	2	2
Dog,Platypus	3	2	2

STRUCTURAL ALIGNMENT VALIDATION

Approach: Not a *relative* measure of an alignment to a gold standard, but rather an *absolute* measure of an alignment inherently

Taxonomies: $\mathcal{P} := \langle P, \leq \rangle, \mathcal{P}' := \langle P', \leq' \rangle$; **Alignment:** $F \subseteq P \times P'$

Links: $\vec{f} = \langle a, a' \rangle, \vec{g} = \langle b, b' \rangle \in F$

Left Anchors: $a, b \in P$; **Right Anchors:** $a', b' \in P'$

Order Discrepancy (Twist): a, b should have the same structural relations in $\mathcal{P}$ as a', b' in $\mathcal{P}'$, $* \in \{\leq, \geq, \not\sim\}$

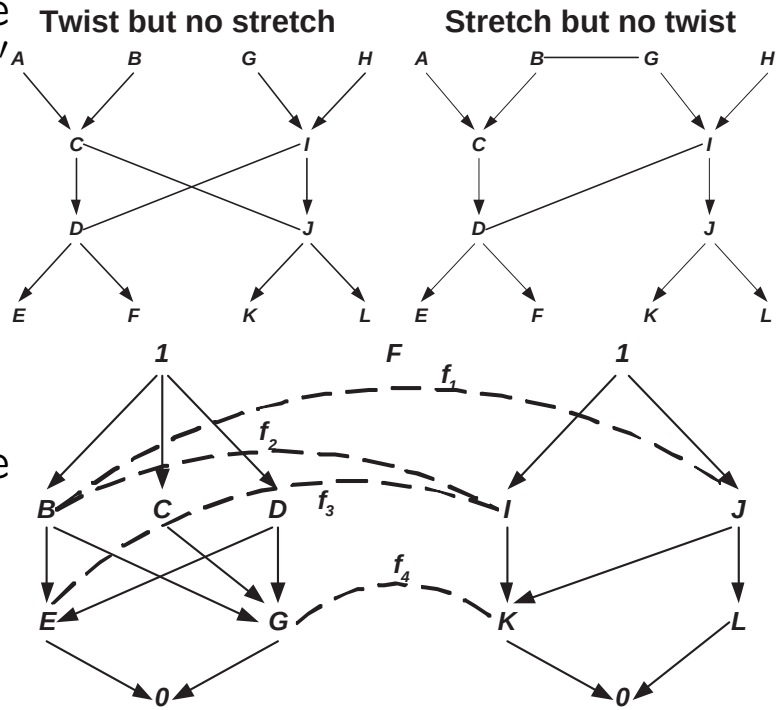
$$\Gamma_F := \frac{\sum_{\vec{f},\vec{g} \in F} \gamma(\vec{f}, \vec{g})}{\binom{|F|}{2}},$$

$$\gamma(\vec{f}, \vec{g}) := \begin{cases} 0, & a * b \text{ and } a' *' b' \\ 1, & \text{otherwise} \end{cases}$$

Distance Discrepancy (Stretch): Relative distance between $a, b \in P$ should be the same as $a', b' \in P'$:

$$D_{F,d} := \frac{\sum_{\vec{f},\vec{g} \in F} |\bar{d}(a, b) - \bar{d}'(a', b')|}{\binom{|F|}{2}}$$

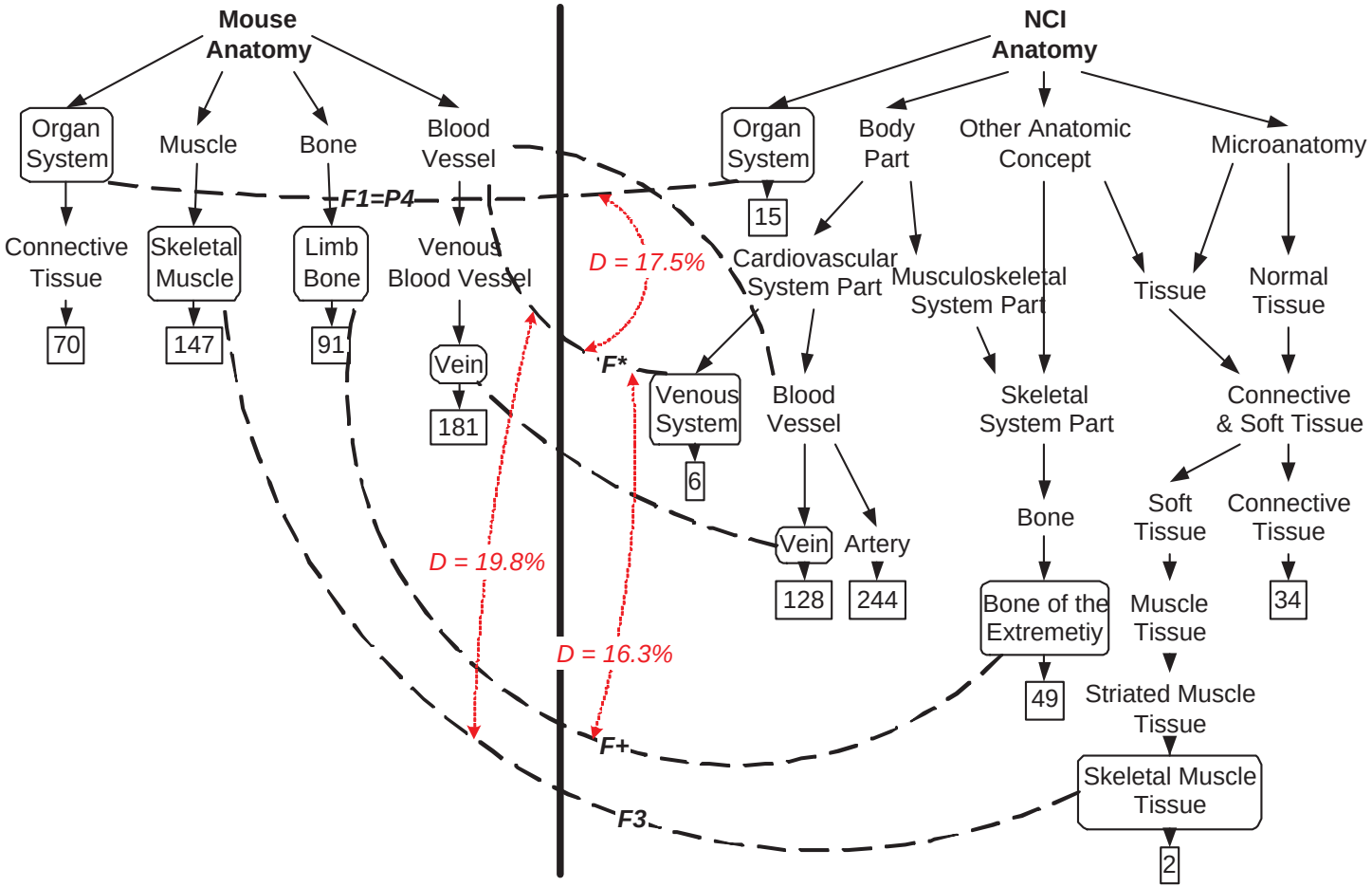
Example: $\Gamma_f = 2/3, D_{F,d^*} = 1/5$



Joslyn, Cliff; Paulson, Patrick; and White, Amanda: (2009) "Measuring the Structural Preservation of Semantic Hierarchy Alignments", in: *Proc. 4th Int. Wshop. on Ontology Matching (OM-2009), CEUR*, v. **551**, `http://ceur-ws.org/Vol-551/om2009_Tpaper6.pdf`

Cliff Joslyn, `cjoslyn@pnl.gov`

ANATOMY TRACK, OAEI 2008



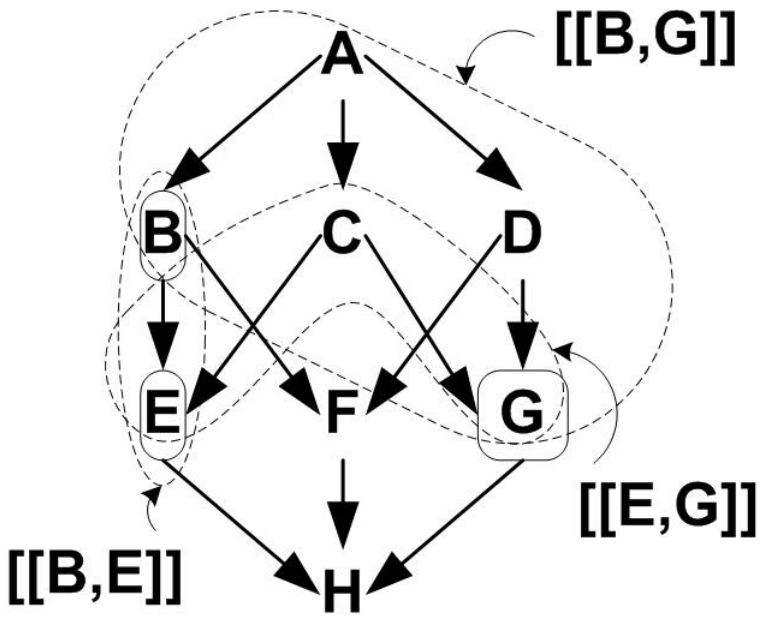
- Ability to rank-order pairs of links by discrepancy

Joslyn, Cliff; Paulson, Patrick; and White, Amanda: (2009) "Measuring the Structural Preservation of Semantic Hierarchy Alignments", in: *Proc. 4th Int. Wshop. on Ontology Matching (OM-2009), CEUR*, v. **551**, `http://ceur-ws.org/Vol-551/om2009_Tpaper6.pdf`

Pacific Northwest
NATIONAL LABORATORY

Cliff Joslyn, `cjoslyn@pnl.gov`

TOWARDS ONTOLOGY CLUSTERING



- Large ontology (121K nodes), system which returns a collection of them
- How dispersed is a collection?

Segment: $[[a,b]]_d := \{c \in P : d(a,b) = d(a,c) + d(c,b)\}$

Convex Hull: $Q \subseteq P, Q' := K(Q) = \cup_{a,b \in Q}[[a,b]]_d$, iterate to convergence

Dispersion: $D(Q) := \sum_{a,b \in C(Q)} d(a,b), \bar{D}(Q) := \frac{D(Q)}{D(P)}$

Example: $Q = \{B, E, G\}, D(Q) = 35, \bar{D}(Q) = 35/91 = 0.385$

ACKNOWLEDGEMENTS, COLLABORATORS AND OTHER NAME-DROPPING

PNNL:

- Sinan al-Saffar
- Alan Chappell
- David Haglin
- Liam McGrath
- Joe Oliveira
- Patrick Paulson
- Amanda White

LANL:

- Bill Bruno
- Judith Cohn
- Susan Mniszewski
- Steve Smith

UC Denver: Karin Verspoor

U. Arizona: Damian Gessler

U. Newcastle: Phillip Lord

Technische Universität Dresden:

- Stephan Schmidt

New Mexico State U.:

- Alex Pogel

Johannes Kepler Universität Linz:

- Jonathan Farley

Miami U Ohio: Valerie Cross

Lehigh U: Chris Orum

Pacific Northwest
NATIONAL LABORATORY