Management of Quantified Semantic Taxonomies for Biothreat Response

Cliff Joslyn

Computer and Computational Sciences
Los Alamos National Laboratory

Modeling, Algorithms, and Informatics (CCS-3)

DIMACS Tutorial and Working Group on
Order-Theoretic Aspects of Epidemiology

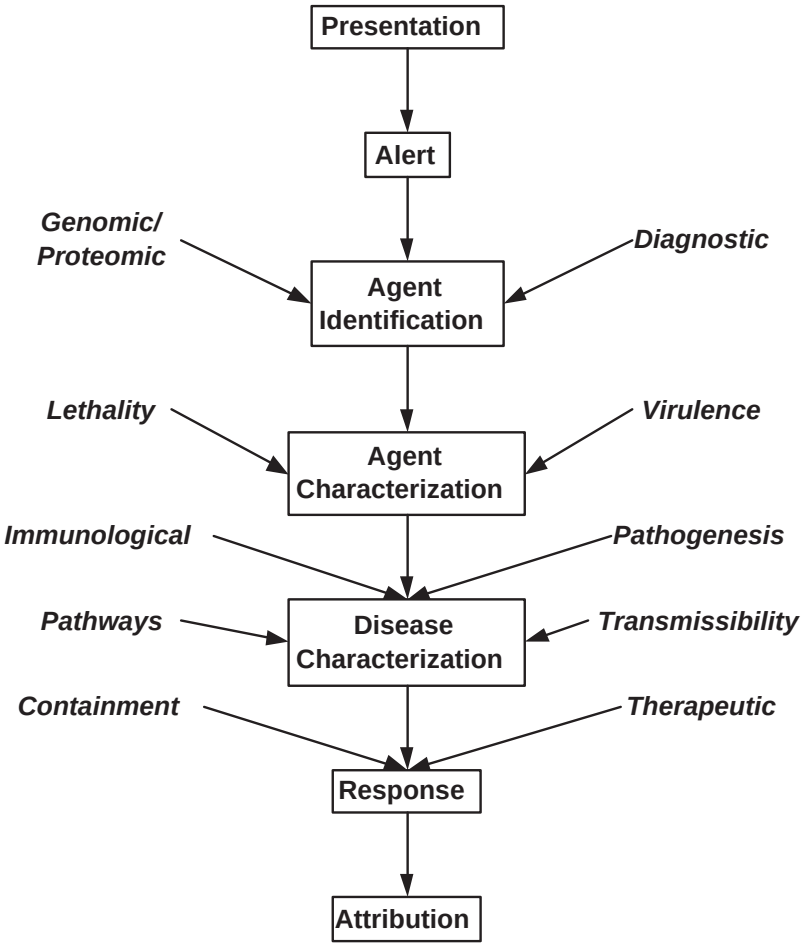
March, 2005

Los Alamos Unlimited Release
04-8407, 05-0340, 05-0640, 05-0907, 05-1621

OUTLINE

- Knowledge integration for biothtreat response
- Bio-ontologies
- Order theoretical representations and approaches: POSet Ontologies (POSOs)
- Categorization and annotation problems
- Quantified POSOs
- Interoperability problem: towards a mathematical definition

KNOWLEDGE INTEGRATION FOR BIOTHREAT RESPONSE



- Rapid response to a novel biothreat
- Past experiences: flu, resistant TB, SARS, ebola, anthrax
- Natural or engineered
- Mucho funding: NIH, NSF, DHS, DOD, DARPA, DOE
- New Los Alamos effort in computational and theoretical pathomics
- **Integration of knowledge bases within a biothreat response workflow**

KM Verspoor, CA Joslyn, JA Ambrosiano, A Bäcker, O Bodenreider, L Hirschman, P Karp, H Kelly, S Loranger, M Musen, R Sriram, C Wroe: (2005) "Knowledge Integration for Biothreat Response", *Los Alamos Technical Report 05-0907*

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

BIO-ONTOLOGIES

- Domain-specific concepts *and their semantic relations*
- *At least:* taxonomic, semantic hierarchies of typed objects and relations
- *In addition:* inference engines over these data objects
- Genomic revolution: large collections of hierarchically organized categorizations of biological objects such as genes and proteins
- IT revolution generally: anatomy, clinical, epidemeological
- **Computational biology** primary success story for ontology development
- Rapid proliferation: many more, more coming, other fields

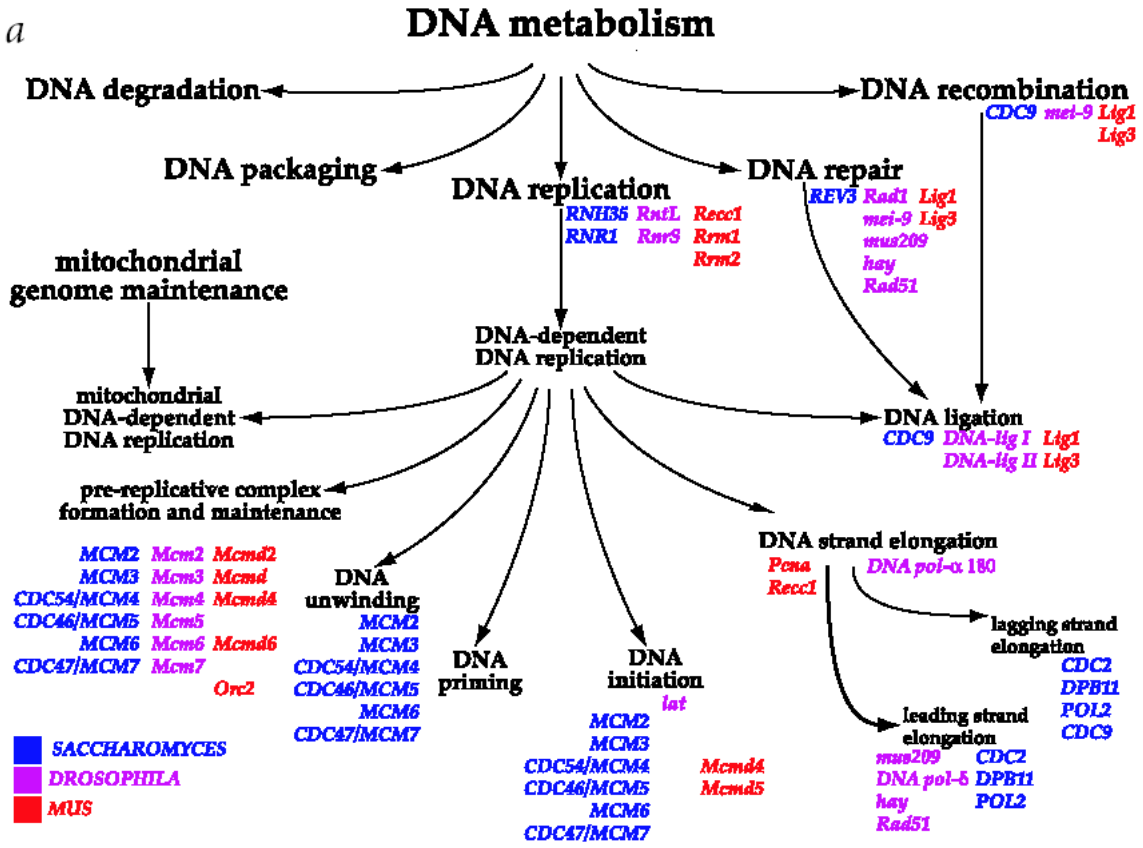
Gene Ontology	http://www.geneontology.org
Fundamental Model of Anatomy	http://sig.biostr.washington.edu/projects/fm/AboutFM.html
Unified Medical Language System	http://www.nlm.nih.gov/research/umls
Open Biology Ontologies	http://obo.sourceforge.net
MEdical Subject Headings	http://www.nlm.nih.gov/mesh/meshhome.html
Enzyme Structures Database	http://www.biochem.ucl.ac.uk/bsm/enzymes

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

GENE ONTOLOGY (GO): DNA METABOLISM PORTION

- Taxonomic controlled vocabulary
- $\sim$ $16K$ nodes P_{GO} populated by genes, proteins
- Two orders on P_{GO}: $\leq_{isa}, \leq_{has}$
- 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

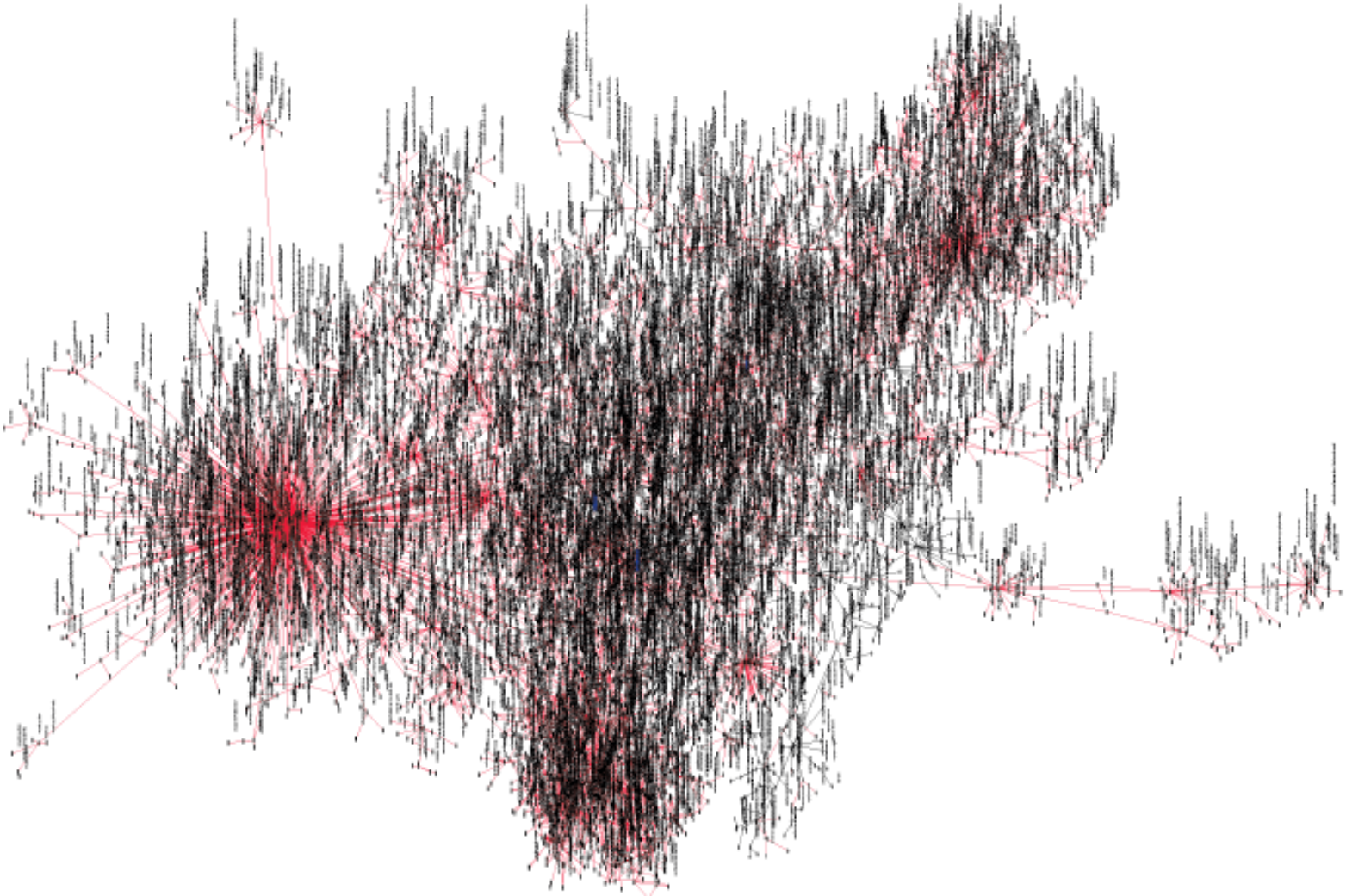
- Tremendous computational resource: large, semantically rich, validated, middle ontology, first (?) in major use

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

GO CA. 2001



Courtesy of Robert Kueffner, NCGR, 2001

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

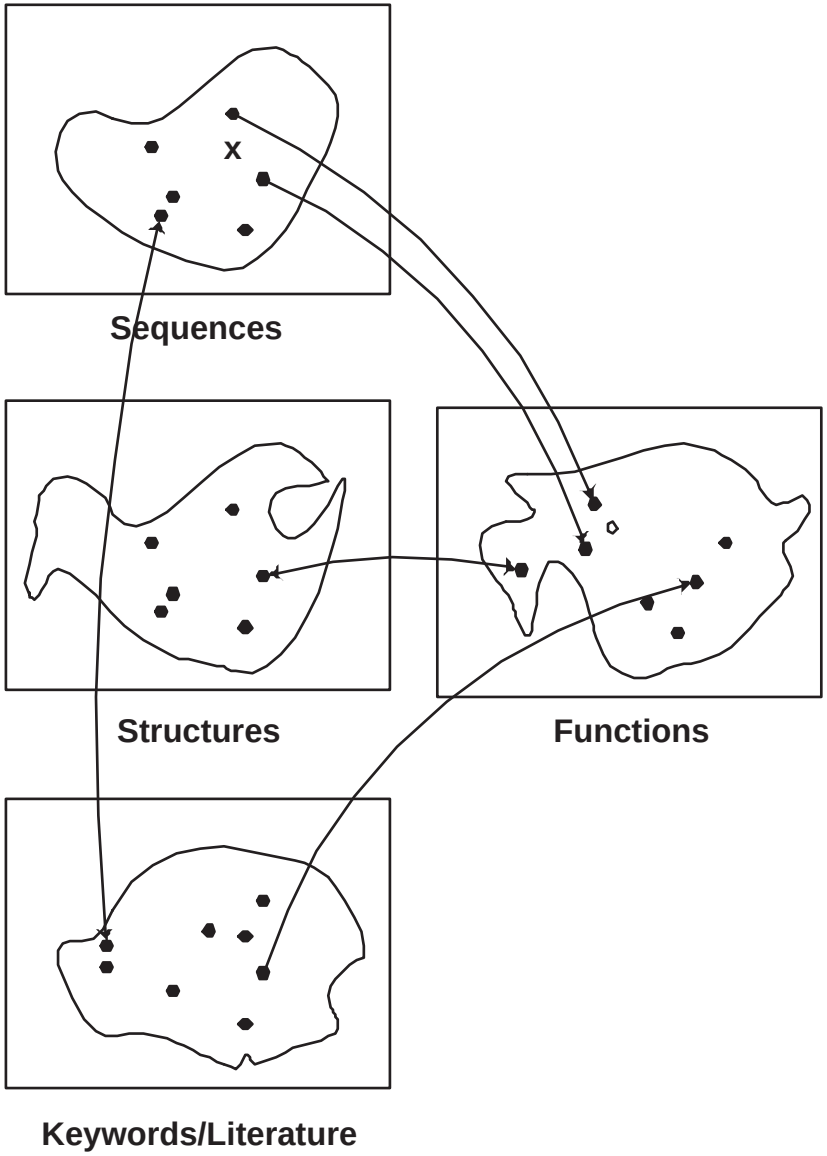
CATEGORIZATION TASK: "CLUSTER" GENES IN ONTOLOGY SPACE

- Develop functional hypotheses about 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?

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

Cliff Joslyn, joslyn@lanl.gov

ANNOTATION TASK



- Mappings among *regions* of sequence, structure, keyword spaces
- Mappings *into* regions of **biological function** space: taxonomic bio-ontologies of molecular function
- Characterize *formal* structure of bio-ontologies:
  - Order theoretical approaches
  - Combinatoric algorithms

KM Verspoor, JD Cohn, SM Mniszewski, and CA Joslyn: (2004) "Nearest Neighbor Categorization for Function Prediction", in: *Proc. 5th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP 05)*, in press

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

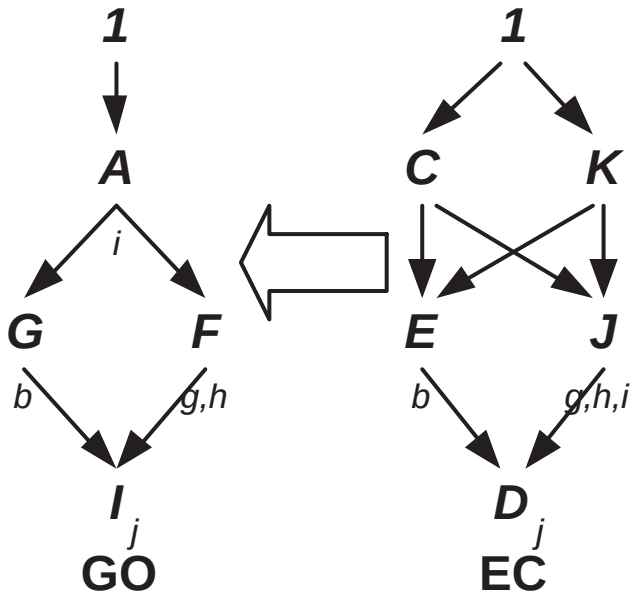
INTEROPERABILITY TASKS: MERGING AND MATCHING

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?



ORDER THEORETICAL KNOWLEDGE DISCOVERY

- Cast databases as (collections of) ordered data objects:

  Native: Constructed explicitly (e.g. ontologies)

  Induced: From other relational data (e.g. concept lattices)
- With inherent **semantics:** node, link types; metadata; text
- Equipped with measures:

  Combinatorial: Distance, rank

  Statistical: Various scores, entropy measures . . .
- **Tasks:** Induction, navigation, visualization, link analysis, search, classification, retrieval, anomaly detection, merger, linkage
- Motivated *now* by appearance of databases and methods
- **Substantial progress and value from novel applications of elementary concepts**
- Need **help**: algorithms, mathematics, applications, funding, concepts, organization?

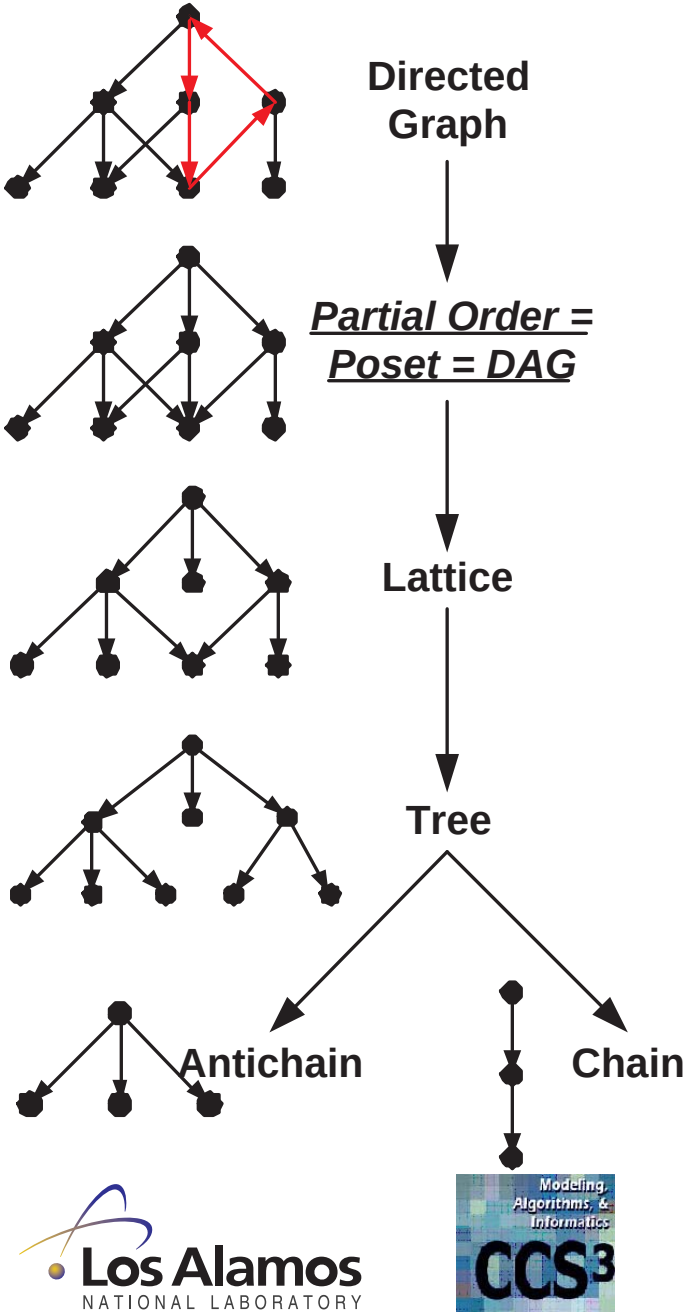
Joslyn, Cliff; Oliverira, Joseph; and Scherrer, Chad: (2004) "Order Theoretical Knowledge Discovery: A White Paper", Los Alamos Technical Report 04-5812, `ftp://ftp.c3.lanl.gov/pub/users/joslyn/white.pdf`

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, `joslyn@lanl.gov`

SEMANTIC 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

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

BASIC POSET CONCEPTS

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

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

Chains: $a \leq b \rightarrow \mathcal{C}(a,b) := \{C_1(a,b), \ldots, C_j(a,b), \ldots, C_M(a,b)\} \subseteq 2^{2^P}$, and use $C_j, 1 \leq j \leq M$.

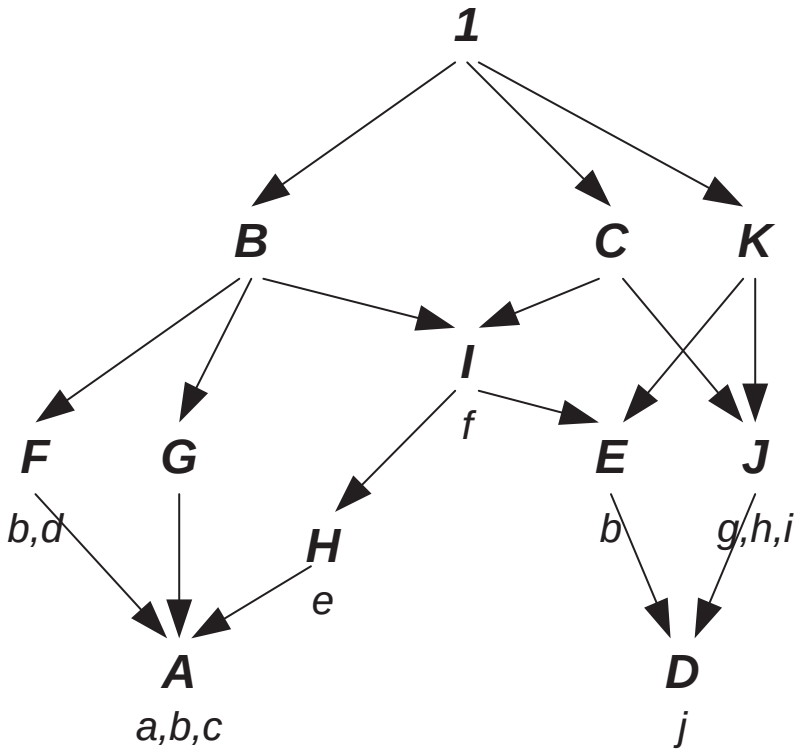
Height: Size of maximal chain: $\mathcal{H}(\mathcal{P})$

Noncomparable Nodes: $a \not\sim b$

Antichain: Collection of noncomparable nodes: $a_1 \not\sim a_2 \not\sim \ldots \not\sim a_n$

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

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



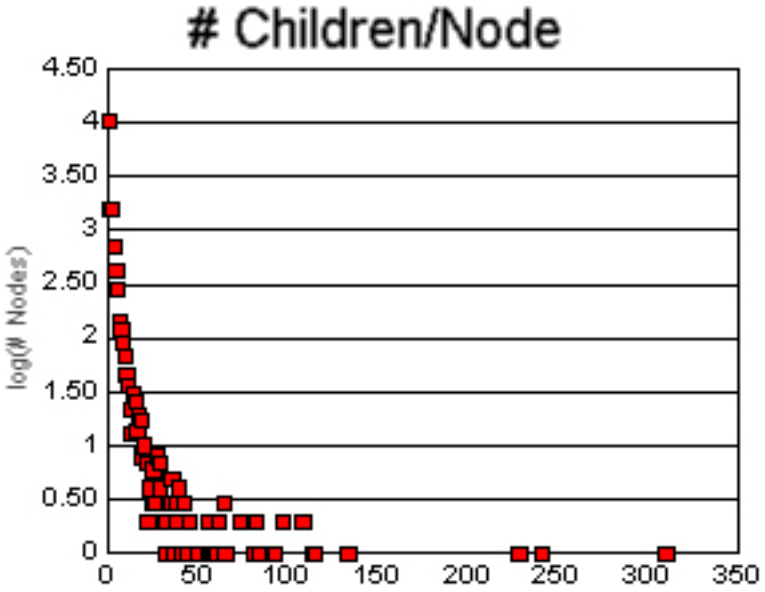
Los Alamos NATIONAL LABORATORY

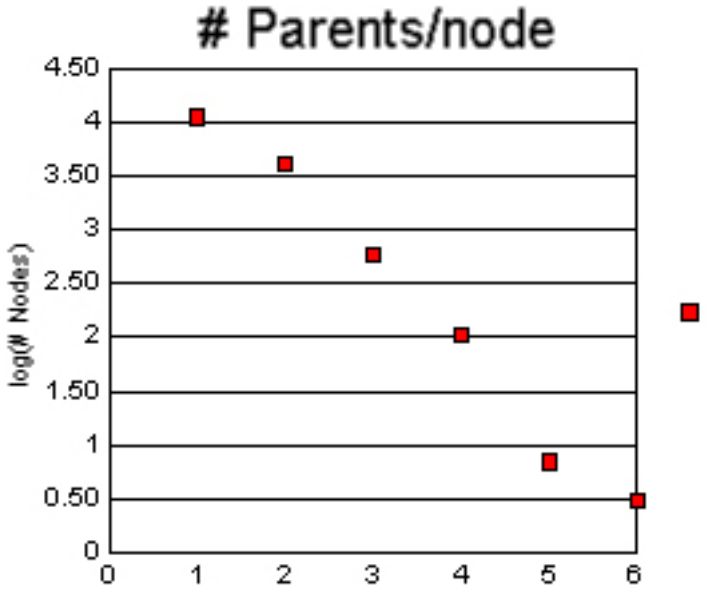
Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

SOME GO POSET STATISTICS

	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





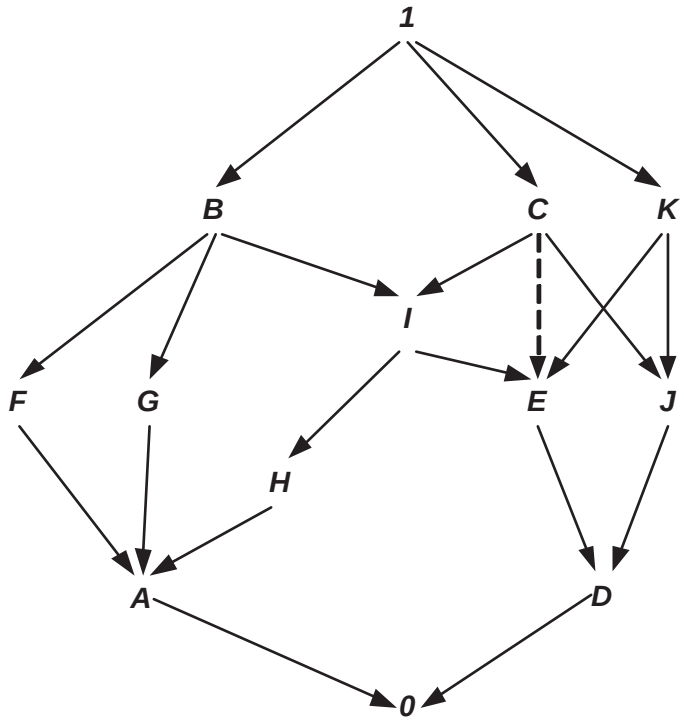
- GO for September, 2003
- Model as $\mathcal{P}_{GO} = \langle P_{GO}, \leq_{isa} \cup \leq_{has} \rangle$

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

DAGS, POSETS, AND COVERS



Graphical DAG:

$\Gamma := \{\gamma_1, \gamma_2, \ldots, \gamma_i, \ldots, \gamma_n\}$

Directed Edge:

$\gamma_i = \langle a, b \rangle \in P^2, a, b \in P.$

Also use $\gamma(a, b)$.

Relational DAG:

$\mathcal{D}(\Gamma) := \langle P, \Leftarrow \rangle$, where $\Leftarrow \subseteq P^2, \forall a, b \in P, a \Leftarrow b \leftrightarrow \langle a, b \rangle \in \Gamma$.

Cover: $\mathcal{V}(\mathcal{D}) := \langle P, <\cdot \rangle$, transitive reduction of $\Leftarrow$

Poset: $\mathcal{P}(\mathcal{D}) := \langle P, \leq \rangle$, transitive and reflexive closure of $\Leftarrow$.

Ideal, Filter: $\downarrow(a) := \{b \in P : b \leq a\}, \uparrow(a) := \{b \in P : a \leq b\}$

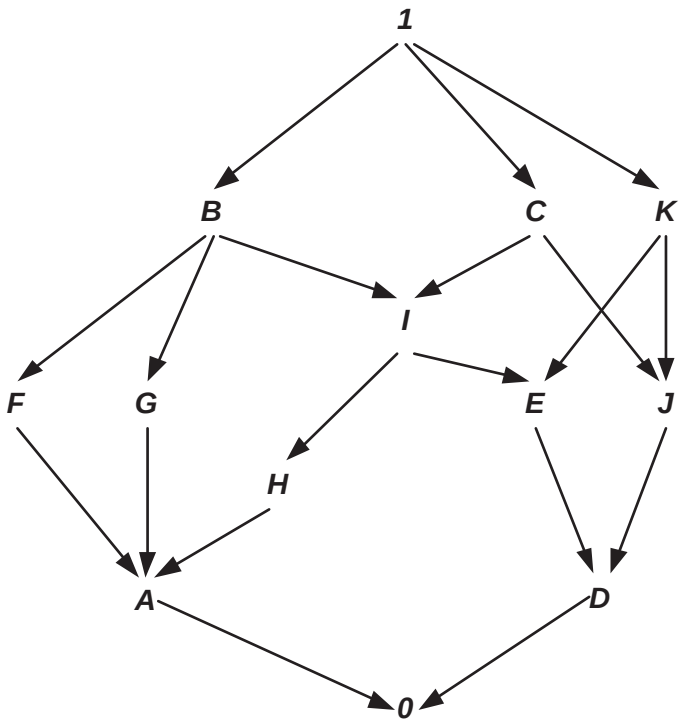
Children, Parents:

$\dot{\downarrow}(a) := \{b \in P : b <\cdot a\}, \dot{\uparrow}(a) := \{b \in P : a <\cdot b\}$

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

CHAIN DECOMPOSITION OF INTERVALS



Assume $a \leq b \in P$

Chain Decomposition:

$$[a,b] = \bigcup_{j=1}^{M} C_j$$

Dilworth: $M \geq \mathcal{W}([a,b])$

Chain Length:

$$h_j := |C_j| - 1, \bar{h}_j := h_j/(\mathcal{H}-1)$$

Vectors of Chain Lengths:

$$\vec{h}(a,b) := \left\langle h_1, h_2, \ldots, h_j, \ldots, h_M \right\rangle,$$
$$\vec{\bar{h}}(a,b) := \vec{h}/(\mathcal{H}-1)$$

Extremes:

$$h_*(a,b) = \min_{h_j \in \vec{h}(a,b)} h_j, \quad \bar{h}_*(a,b) = \min_{\bar{h}_j \in \vec{\bar{h}}(a,b)} \bar{h}_j,$$

$$h^*(a,b) = \max_{h_j \in \vec{h}(a,b)} h_j, \quad \bar{h}^*(a,b) = \max_{\bar{h}_j \in \vec{\bar{h}}(a,b)} \bar{h}_j.$$

Chains: $C_j = \{\gamma(a,c_1), \ldots, \gamma(c_{h_j-3}, c_{h_j-2}), \gamma(c_{h_j-2}, b)\}$ for some collection of nodes $\{c_1, c_2, \ldots, c_i, \ldots c_{h_j-2}\} \subseteq P, 1 \leq i \leq h_j - 2$.
$C_j = a <\cdot c_1 <\cdot \ldots <\cdot c_{h_j-3} <\cdot c_{h_j-2} <\cdot b, \gamma_i \in C_j, 1 \leq i \leq h_j$

Cliff Joslyn, joslyn@lanl.gov

PSEUDO-DISTANCES

Pseudo-Distance: Some aggregate measure of the number of "hops" between two comparable nodes: $\delta\colon P^2 \mapsto \mathbb{R}$ where $\forall a \leq b \in P, h_*(a,b) \leq \delta(a,b) \leq h^*(a,b)$

Normalized: $\bar{\delta} := \delta/(\mathcal{H}-1) \in [0,1]$

Minimum Chain Length: $\delta_m(a,b) := h_*(a,b), \bar{\delta}_m(a,b) := \bar{h}_*(a,b)$

Maximum Chain Length: $\delta_x(a,b) := h^*(a,b), \bar{\delta}_x(a,b) := \bar{h}^*(a,b)$

Average of Extreme Chain Lengths:

$$\delta_{ax}(a,b) := \frac{h_*(a,b)+h^*(a,b)}{2}, \quad \bar{\delta}_{ax}(a,b) := \frac{\bar{h}_*(a,b)+\bar{h}^*(a,b)}{2}$$

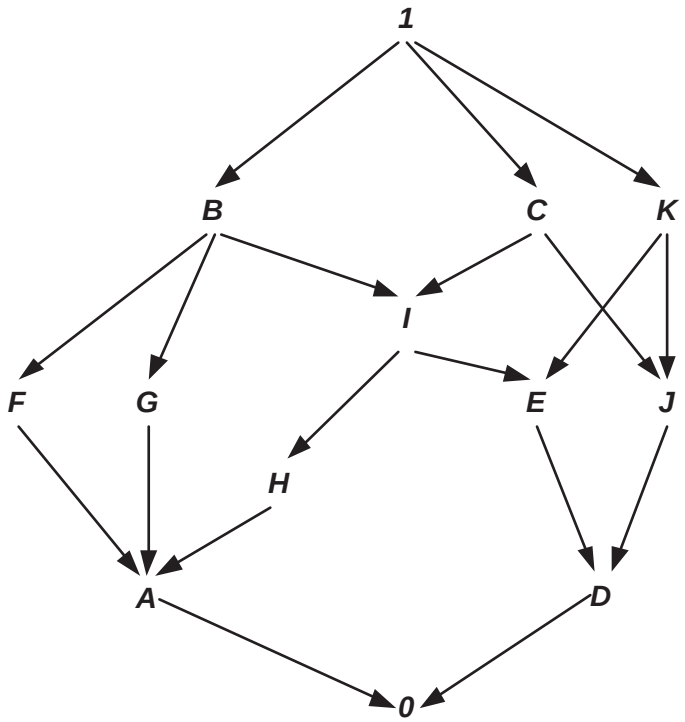
Average of All Chain Lengths:

$$\delta_{ap}(a,b) := \frac{\sum_{h_j \in \vec{h}(a,b)} h_j}{M}, \quad \bar{\delta}_{ap}(a,b) := \frac{\sum_{\bar{h}_j \in \vec{\bar{h}}(a,b)} \bar{h}_j}{M}$$

EXAMPLE

For $D \leq 1 \in P$

$$\begin{aligned}
\mathcal{H}(\mathcal{P}) &= 6, \quad \mathcal{W}([D,1]) = 2, \quad M = 5 \\
\mathcal{C}(D,1) &= \{ D <\cdot E <\cdot I <\cdot B <\cdot 1, D <\cdot E <\cdot I <\cdot C <\cdot 1, \\
&\quad D <\cdot E <\cdot K <\cdot 1, D <\cdot J <\cdot C <\cdot 1, \\
&\quad D <\cdot J <\cdot K <\cdot 1\}, \\
\vec{h}(D,1) &= \langle 4,4,3,3,3 \rangle \\
\vec{\overline{h}}(D,1) &= \langle 4/5, 4/5, 3/5, 3/5, 3/5 \rangle
\end{aligned}$$

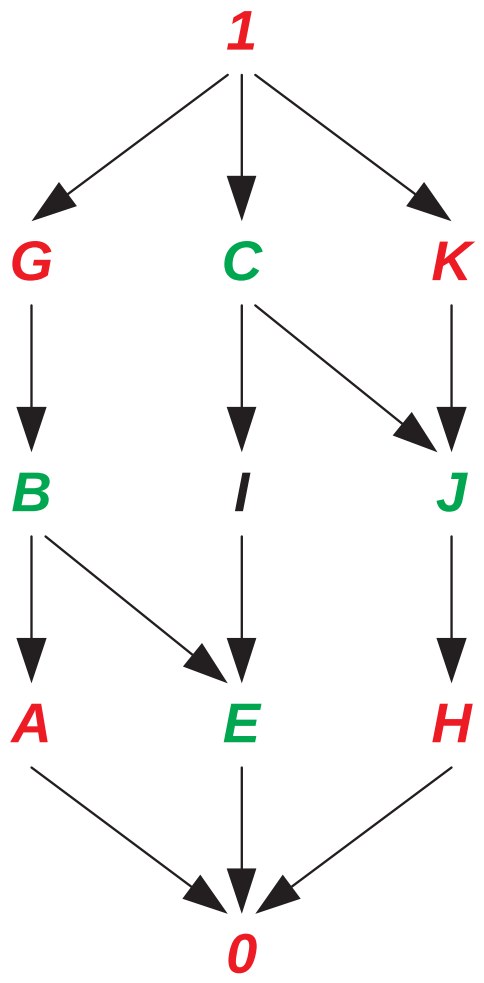


$$\begin{aligned}
\delta_m(D,1) &= 3, \quad \delta_x(D,1) = 4, \\
\delta_{ax}(D,1) &= 3.5, \quad \delta_{ap}(D,1) = 3.4, \\
\overline{\delta}_m(D,1) &= 0.60, \quad \overline{\delta}_x(D,1) = 0.80, \\
\overline{\delta}_{ax}(D,1) &= 0.70, \quad \overline{\delta}_{ap}(D,1) = 0.68.
\end{aligned}$$

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

SIMPLE "GENEOLOGICAL" SENSE OF DISTANCE



- Consider concentric regions around a node $a \in P$
- Either vertical or horizontal, towards concept of "diameter" of a poset
- Consider nodes $a, b \in P$

  Exact Match: $a = b$

  Nuclear Family: a is a parent, child, or sibling of b

  Extended Family: a is a grandparent, grandchild, uncle, nephew of b

INTERVAL RANK

- Rank defined in lower-bounded posets:

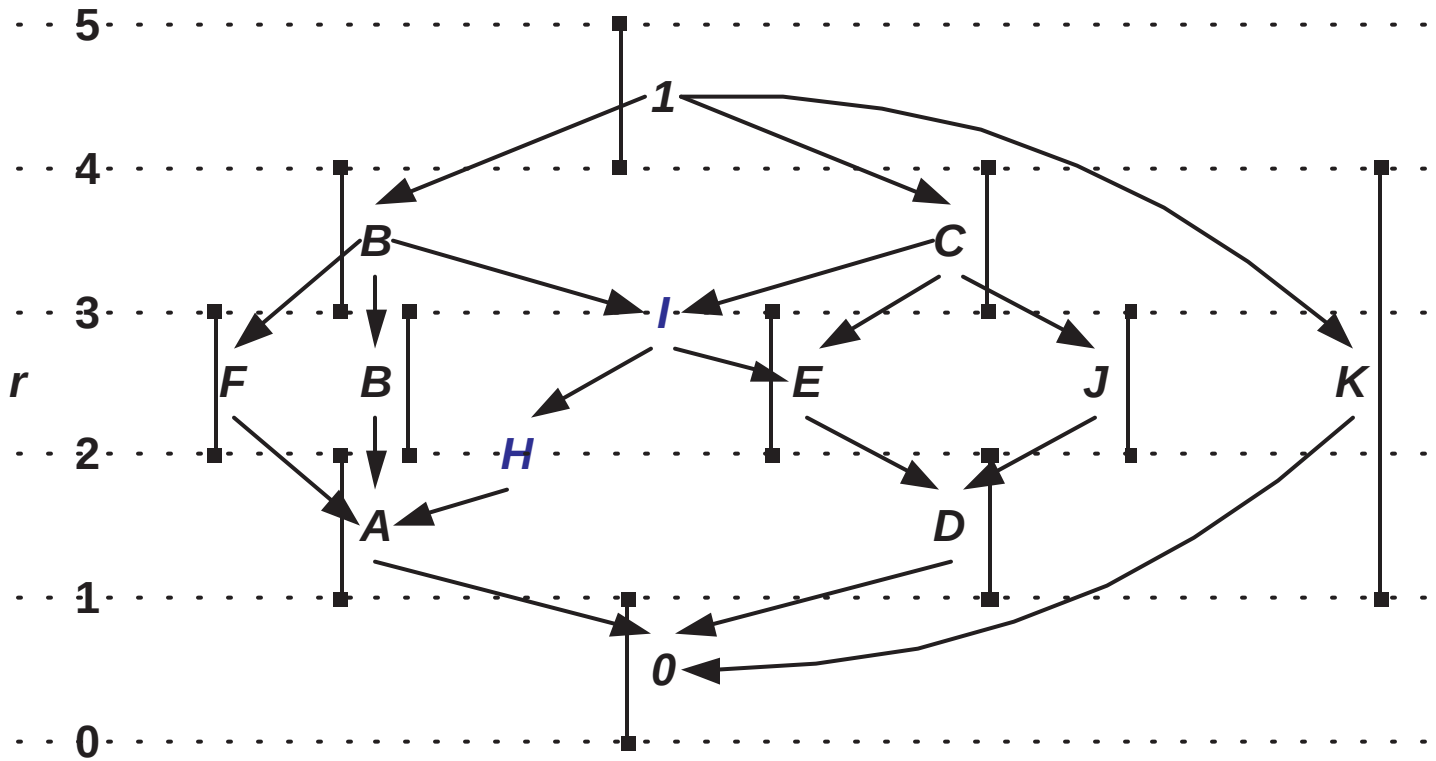
$$r_*(p) := \begin{cases} 0, & p \in \min(P) \\ n, & p \in \min\left(P - \{q : r_*(q) < n\}\right) \end{cases}$$

- **Rank Interval Function:**

$$R(p) := [r_*(p), \mathcal{H}(\mathcal{P}) - r^*(p)]$$

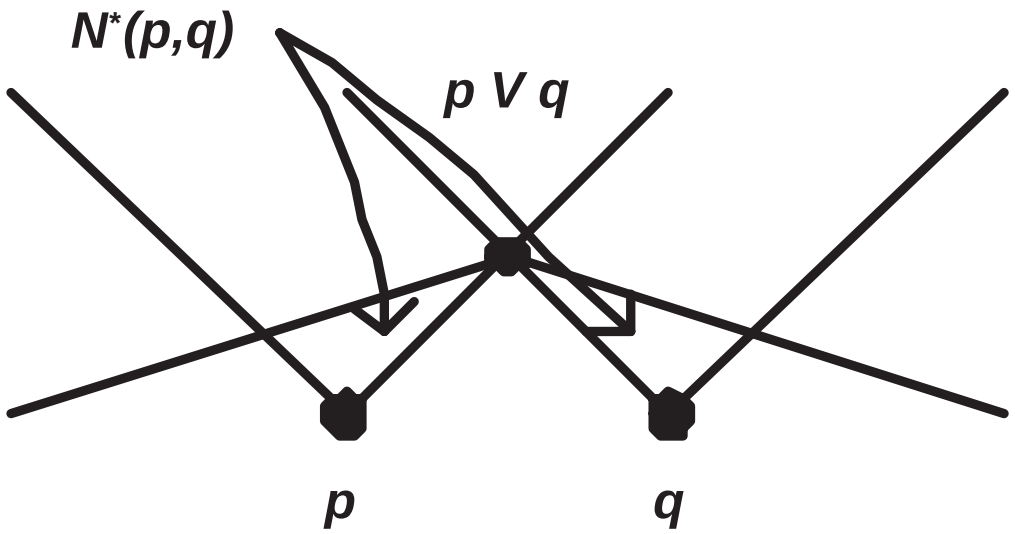
  using dual upper rank $r^*(p)$

- **Example:** $R(E) = [2,3], R(I) = [3,3] = 3, R(K) = [1,4]$



POSET NEIGHBORHOODS

Upper Neighborhood: $\bigvee Q$ exists $\rightarrow N^*(Q) := \uparrow Q \cap \downarrow (\bigvee Q)$
Lower Neighborhood: $\bigwedge Q$ exists $\rightarrow N_*(Q) := \downarrow Q \cap \uparrow (\bigwedge Q)$
Neighborhood: $\bigvee Q$ and $\bigwedge Q$ exist $\rightarrow N(Q) := \Xi(Q) \cap [\bigwedge Q, \bigvee Q]$
Pairwise: $Q = \{p, q\}$, then define for each appropriate form e.g. $N(p, q) := N(Q)$.
Theorem: $C = p_1 \leq \ldots \leq p_n \subseteq P$ is a chain $\rightarrow N(C) = [p_1, p_n]$
Corollary: $p \leq q \rightarrow N(p, q) = [p, q]$



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

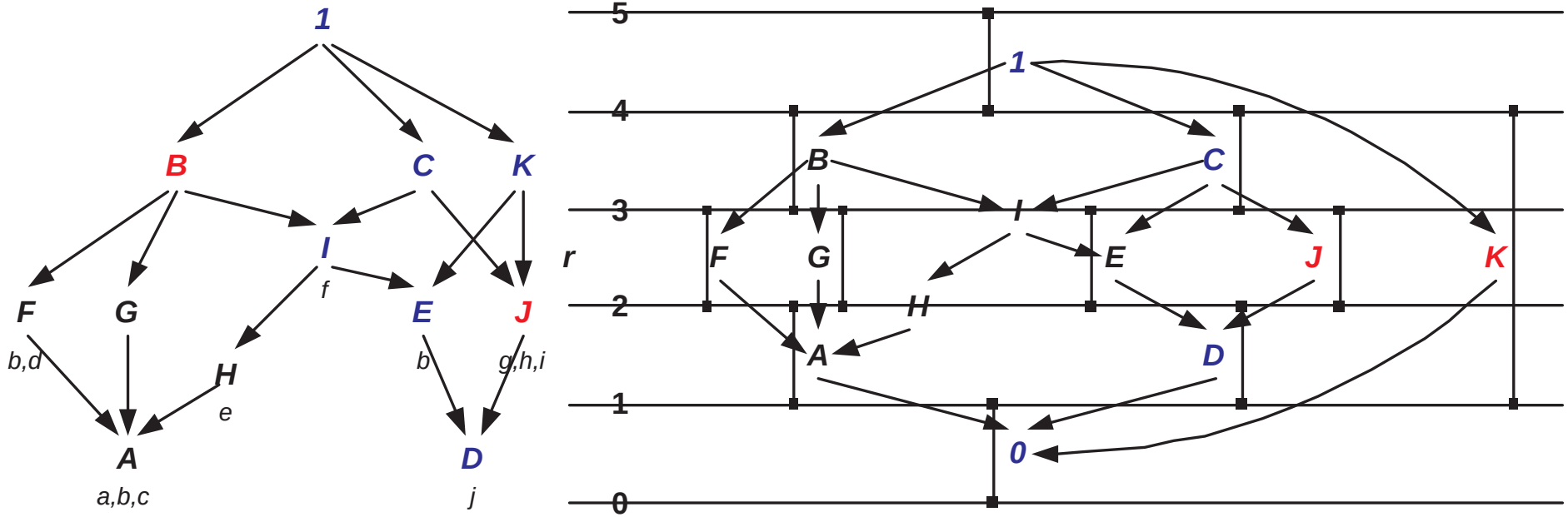
HORIZONTAL DISTANCES

Assume a bounded poset $\mathcal{P}$, subset $Q \subseteq P$, nodes $a, b \in P$

Size of Region: $\mathbf{D}(Q) := \langle \mathcal{H}(N(Q)), \mathcal{W}(N(Q)) \rangle$

Example: $\mathbf{D}(B, J) = \langle 5, 2 \rangle$ (left); $\mathbf{D}(J, K) = \langle 4, 2 \rangle$ (right)

Otherwise: Height of 2-fence between a, b; width of maximal fence between a, b



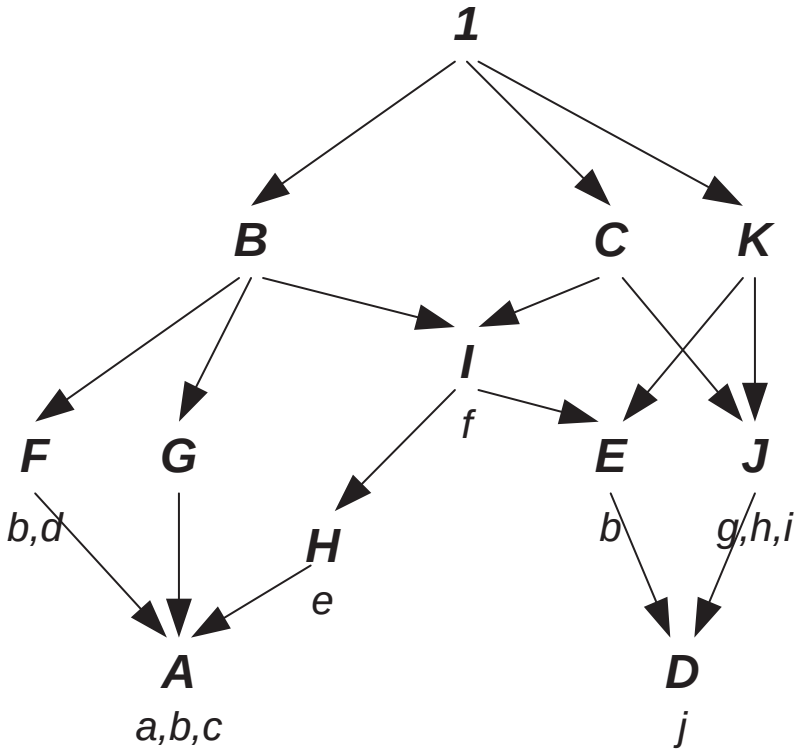
Joslyn, Cliff: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in: *Conceptual Structures at Work, Lecture Notes in Artificial Intelligence*, v. **3127**, ed. Wolff, Pfeiffer and Delugach, pp. 287-302, Springer-Verlag, Berlin

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

POSET ONTOLOGY CATEGORIZER (POSOC)

- **POSO:** POSet-based 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

Joslyn, Cliff; Mniszewski, Susan; Fulmer, Andy; and Heaton, Gary: (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

COMPONENTS OF SCORING FUNCTIONS

- Recalling $\mathcal{O} = \langle \langle P, \leq \rangle, X, F \rangle$
- Set $X = \{x\}$ of n genes (proteins)
- **Labeling function:**
  - $F: X \mapsto 2^P$
  - $F(x) =$ set of GO nodes (functions) of gene x
- $S_Y(a)$: weighted rank of node $a \in P$ based on requested genes X
- **Unnormalized Score:** $S_Y: P \mapsto \mathbb{R}^+$
- **Normalized Score:** $\widehat{S}_Y: P \mapsto [0, 1]$
- **Slider:** Balance coverage against specificity
  - $r = 2^s$
  - $s \in \{\ldots -1, 0, 1, 2, 3, \ldots\}$
  - Low $s \rightarrow$ emphasize coverage
  - High $s \rightarrow$ emphasize specificity

SCORING FUNCTIONS

Recalling $r = 2^s$

Unnormalized Distance, Unnormalized Score:

$$S_Y(a) := \sum_{x \in X} \sum_{b \in F(x): b \leq a} \frac{1}{(\delta(b,a)+1)^r}$$

Unnormalized Distance, Normalized Score:

$$\widehat{S}_Y(a) := \frac{S_Y(a)}{\sum_{x \in X} |F(x)|}$$

Normalized Distance, Unnormalized Score:

$$\bar{S}_Y(a) := \sum_{x \in X} \sum_{b \in F(x): b \leq a} \left(1 - \bar{\delta}(b,a)\right)^r$$

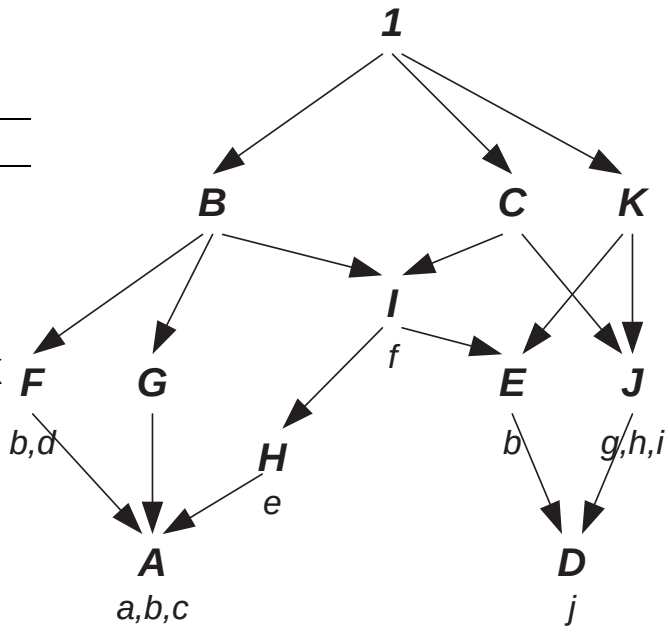
Normalized Distance, Normalized Score:

$$\widehat{\bar{S}}_Y(a) := \frac{\bar{S}_Y(a)}{\sum_{x \in X} |F(x)|}$$

POSOC EXAMPLE

- $Y = \{c, e, i\}$
- Specificity: $s = -1$ values coverage; $s = 3$ values specificity
- $\delta = \delta_m =$ min chain length between comparable nodes (many others possible)
- Normalized score $\bar{S}_Y(a)$
- Show cluster heads in bold, secondaries with $*$

	$s=-1$		$s=1$		$s=3$	
	$\bar{S}_Y(a)$	a	$\bar{S}_Y(a)$	a	$\bar{S}_Y(a)$	a
1	0.767	**C**	0.547	**H**	0.389	**H**
2	0.680	**1***	0.387	**C***	0.333	A;**J**
3	0.632	H	0.333	A;I;J		
4	0.556	I			0.062	**C***
5	0.516	B			0.062	I
6	0.333	A;J	0.240	**B***	0.056	F;G;K
7			0.227	**1***		
8	0.298	F;G;K	0.213	F;G;K		
9					0.011	B
10					0.006	**1**

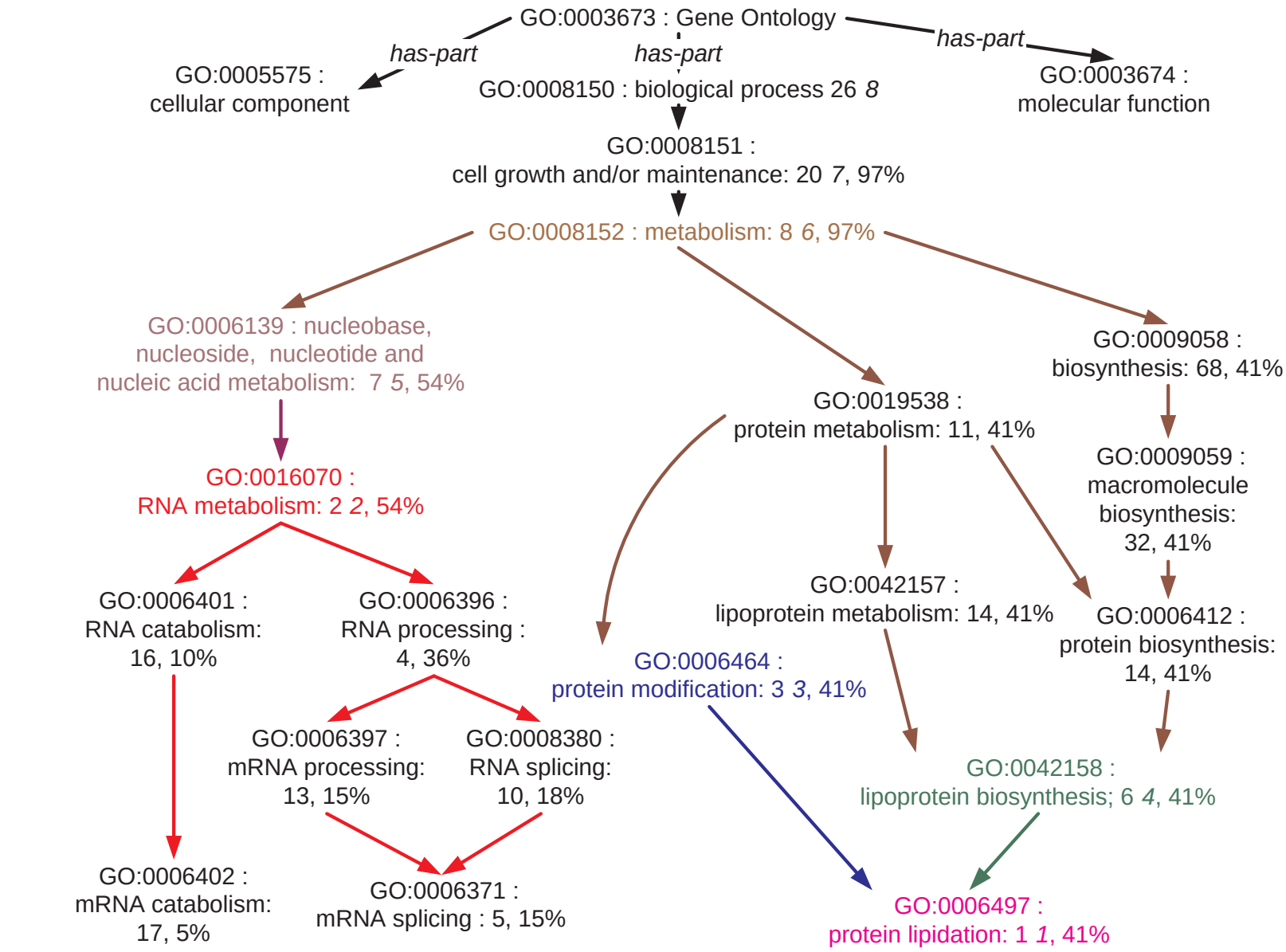


Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

QUERY $2C-$

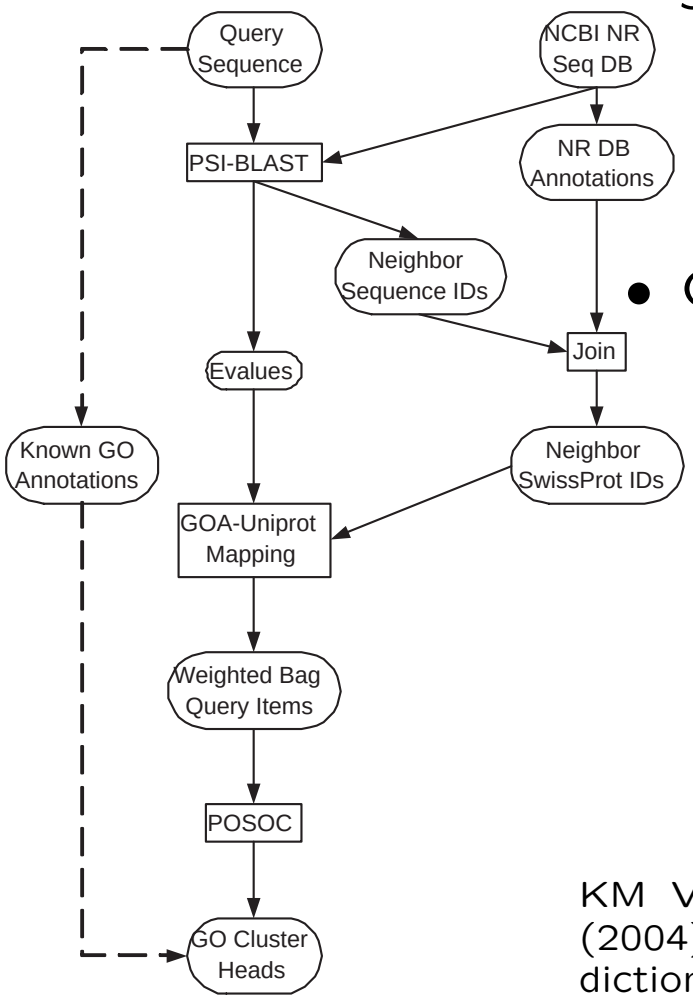


Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

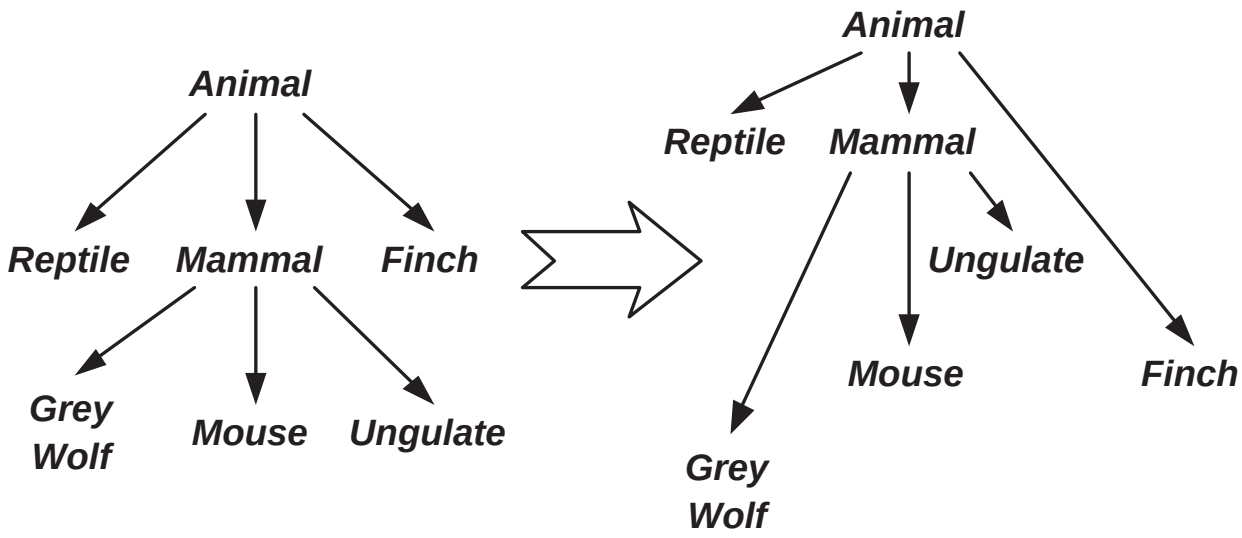
ANNOTATION AS CATEGORIZATION



- Find neighbors of target in sequence space
  - BLAST search on the target against the NCBI NR database, 5 iterations
  - Default e-value threshold of 10
- Collect GO nodes of neighbors:
  - Obtain Swiss-Prot identifiers of each PSI-BLAST match from parsed listing of the NR database headers
  - Swiss-Prot to GO mappings to find all GO nodes for proteins
  - Weight each GO node by the PSI-BLAST evalue

KM Verspoor, JD Cohn, SM Mniszewski, and CA Joslyn: (2004) "Nearest Neighbor Categorization for Function Prediction", in: *Proc. 5th CASP 05*, in press

LINK-WEIGHTED POSOC



Objections to POSOC:

- Structure of the GO doesn't reflect "reality" as much as perhaps "funding history"
- A link over here isn't the same as a link over there

Solution:

- Complement pure structural approach with statistical information source
- Shrink links where more information, stretch links where less, to reflect underlying metric

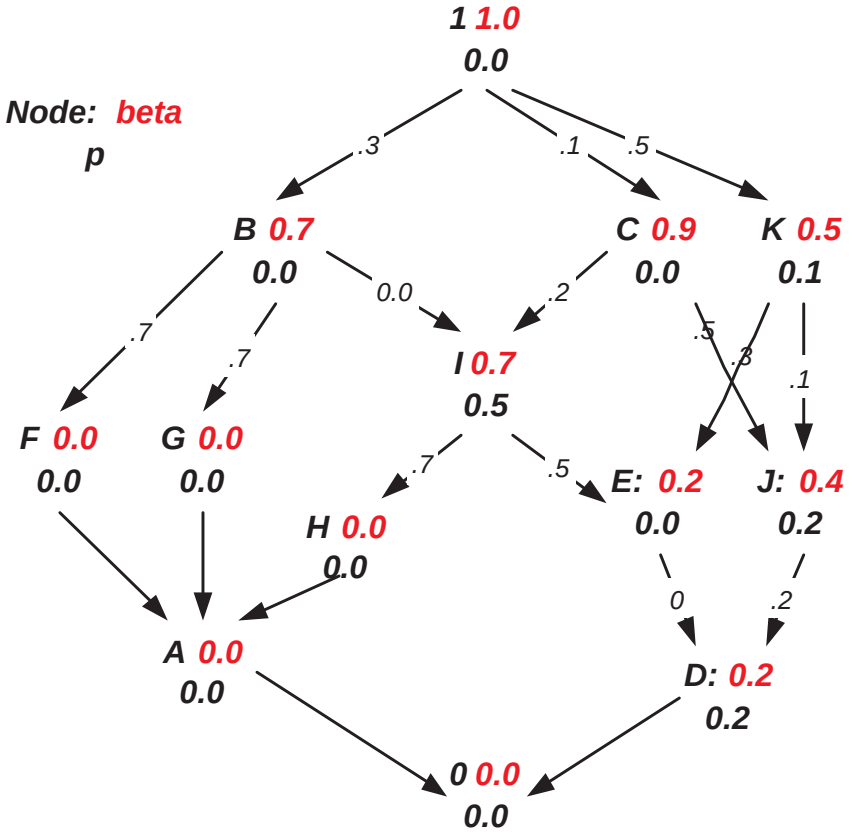
APPROACHES

1. Cast a probability distribution p onto the POSO, use information gain between comparable nodes to weight pseudo-distances.
2. Cast a discrete Markov process on the POSO's underlying poset, to derive a well-justified Markov-based pseudo-distance δ_p as the expected value of chain length between comparable nodes

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

Cliff Joslyn, joslyn@lanl.gov

WEIGHTED POSOS



Weighted Poset:

$\mathcal{O} := \langle \mathcal{D}(P), p \rangle$, where $p\colon P \mapsto [0,1]$ is a probability distribution on the nodes, so that $\sum_{a \in P} p(a) = 1$

Measure: $\beta\colon P \mapsto [0,1]$

$$\forall b \in P, \quad \beta(b) := \sum_{a \leq b} p(a) = \sum_{a \in \downarrow b} p(a).$$

Monotonicity: $a \leq b \rightarrow \beta(a) \leq \beta(b)$

Joslyn, Cliff and Bruno, William J: (2005) "Weighted Pseudo-Distances for Categorization in Semantic Hierarchies", submitted to *2005 Int. Conference on Conceptual Structures*

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

OF MATHEMATICAL INTEREST . . .

- **Base:** $\mathbf{B}(\mathcal{O}) := \{a \in P : p(a) > 0\} \subseteq P$
- When $\mathcal{P}$ is a Boolean lattice: $\forall a, b \in P$

  $$\beta(a \vee b) \geq \beta(a) + \beta(b) - \beta(a \wedge b)$$

- And in particular the power set 2^{Ω} on some underlying finite set Ω then $\beta \mapsto$ belief function Bel:

  $$\forall A, B \subseteq \Omega, \quad \mathsf{Bel}(A \cup B) \geq \mathsf{Bel}(A) + \mathsf{Bel}(B) - \mathsf{Bel}(A \cap B)$$

- If $\mathbf{B}(\mathcal{O})$ is the atoms of $\mathcal{P}$ then $\beta \mapsto$ Pr:

  $$\forall a, b \in P, \quad \beta(a \vee b) = \beta(a) + \beta(b) - \beta(a \wedge b),$$

  $$\forall A, B \subseteq \Omega, \quad \mathsf{Pr}(A \cup B) = \mathsf{Pr}(A) + \mathsf{Pr}(B) - \mathsf{Pr}(A \cap B).$$

- If $\mathbf{B}(\mathcal{O})$ is a maximal chain $C \subseteq \mathcal{P}$ with $|C| = \mathcal{H}(\mathcal{P})$, then $\beta \mapsto$ necessity function η

  $$\forall a, b \in P, \quad \beta(a \wedge b) = \min(\beta(a), \beta(b))$$

  $$\forall A, B \subseteq \Omega, \quad \eta(A \cap B) = \min(\eta(A), \eta(B)).$$

- $\mathcal{P}$ a general lattice, complemented lattice, general poset?

  $$\sum_{c \in a \vee b} \beta(c) + \sum_{c \in a \wedge b} \beta(c) \geq \beta(a) + \beta(b)?$$

Cliff Joslyn, joslyn@lanl.gov

RESNIK APPROACH

Resnik Semantic Similarity: $\forall a, b \in P$,

$$\delta_{\text{Lord}}(a, b) = \max_{c \in a \vee b} \left[-\log_2(\beta(c)) \right].$$

where $a \vee b$ is the set of least upper bounds of a and b.

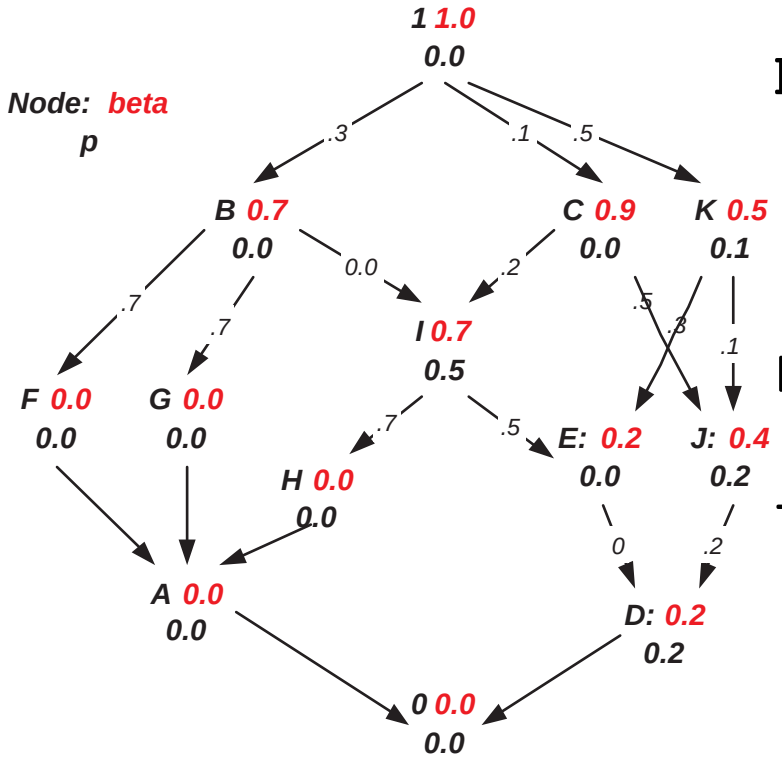
Issues:

- δ is not a distance, defined only on $a \leq b \in P$.
- β is almost never a probability measure on P
- **Theorem:** Let $b \in P$ with $\downarrow b \subseteq P$ a lattice. Then $\forall a_1, a_2, a_3, a_4 \in \dot{\downarrow} b, \delta_{\text{Lord}}(a_1, a_2) = \delta_{\text{Lord}}(a_3, a_4)$.
- **Theorem:** If $a \leq b \leq c$, then $\delta_{\text{Lord}}(a, c) = \delta_{\text{Lord}}(b, c) = \delta_{\text{Lord}}(c, c)$.

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

Resnik, Philip: (1995) "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", in: *Int. Joint Conf. on Artificial Intelligence*, pp. 448-452, Morgan Kaufmann

INFORMATION GAIN



Information Gain: For $a \leq b \in P$, let $\iota(a,b) := \beta(b) - \beta(a)$ be the amount of information gained when moving from b to a.

Edge Information Gain: For $\gamma(a,b)$, let $\iota(\gamma) := \iota(a,b)$

Theorem: $\forall a \leq b \in P, \forall C_j \in \mathcal{C}(a,b)$

$$\sum_{\gamma_i \in C_j} \iota(\gamma_i) = \iota(a,b)$$

$D \leq 1, M = 5$ chains, with

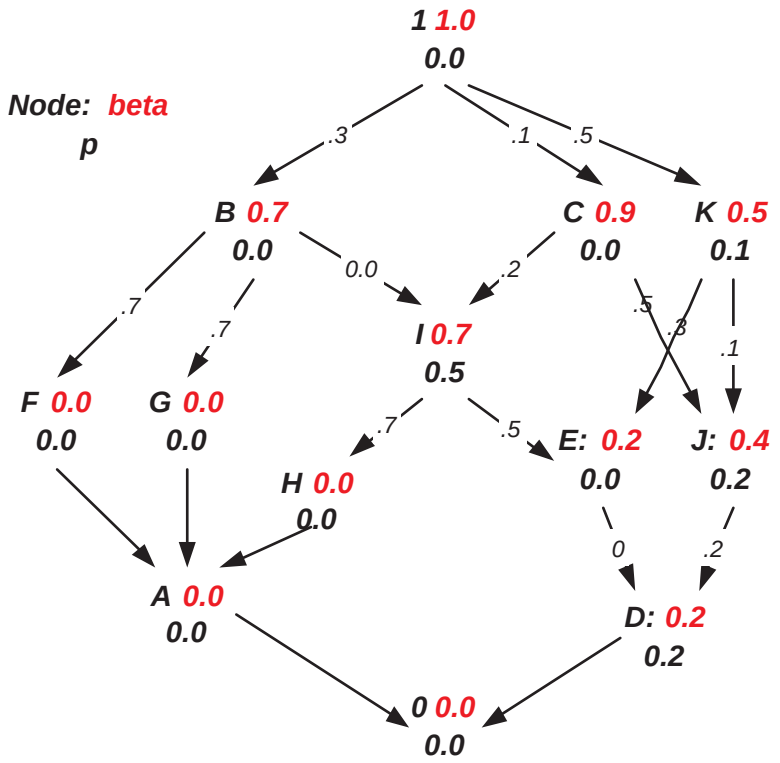
$$\begin{aligned} \iota(D,1) &= .8 = \iota(\langle D,E\rangle) + \iota(\langle E,I\rangle) + \iota(\langle I,B\rangle) + \iota(\langle B,1\rangle) \\ &= 0.0 + 0.5 + 0.0 + 0.3 \\ &= \iota(\langle D,J\rangle) + \iota(\langle J,K\rangle) + \iota(\langle K,1\rangle) = 0.2 + 0.1 + 0.5 \end{aligned}$$

Joslyn, Cliff and Bruno, William J: (2005) "Weighted Pseudo-Distances for Categorization in Semantic Hierarchies", submitted to *2005 Int. Conference on Conceptual Structures*

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

CHAIN WEIGHT MOTIVATION



- For $a \leq b \in P$, for chain $C_j \in \mathcal{C}(a,b)$, construct normalized weighted chain length $\bar{v}_j(a,b)$ as $\bar{h}_j(a,b)$ scaled up by $\iota(a,b)$
- Despite monotonicity (ι increases with chain length), any particular $\bar{h}_j$ could be small while ι is large, or *vice versa*

$f\colon [0,1]^2 \mapsto [0,1]$ with $\bar{v}_j := f\left(\bar{h}_j, \iota(a,b)\right)$

Properties for $f(h,\iota)$ with $h, \iota \in [0,1]$:

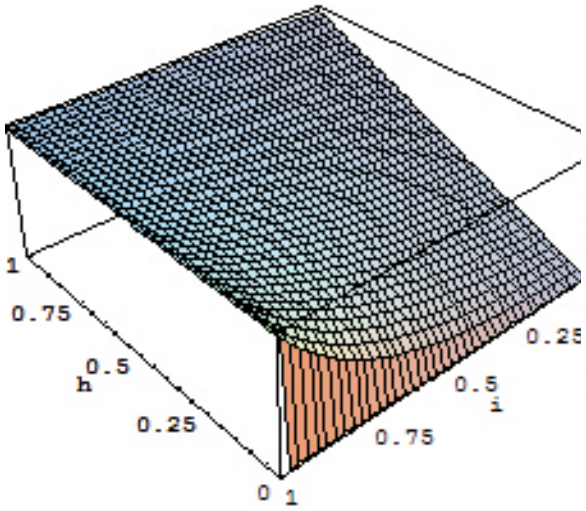
1. $a = b$, minimal distance: $f(0,0) = 0$
2. No information gain, recover h: $f(h,0) = h$
3. Chain length only lengthened: $f(h,\iota) \geq h$
4. Max chain, all mass, maximal distance: $f(1,\iota) = f(h,1) = 1$

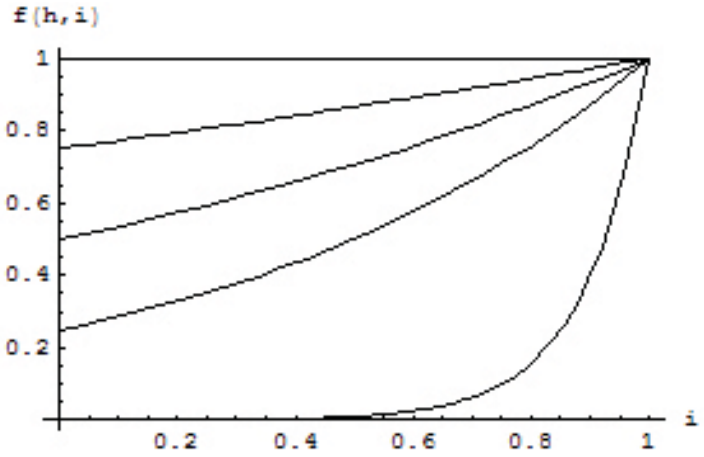
WEIGHT NORMALIZED CHAIN LENGTHS

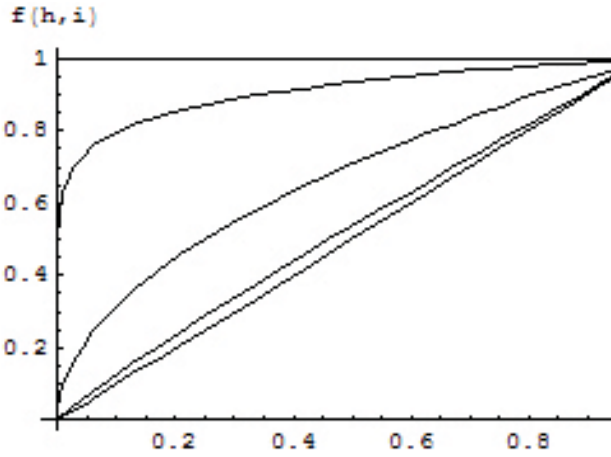
Definition: $\bar{v}_j := f\left(\bar{h}_j, \iota(a,b)\right)$

$$f(h,\iota) := h^{1-\iota}, \quad h, \iota \in [0,1]$$

Theorem: $\bar{v}_j$ satisfies the conditions above

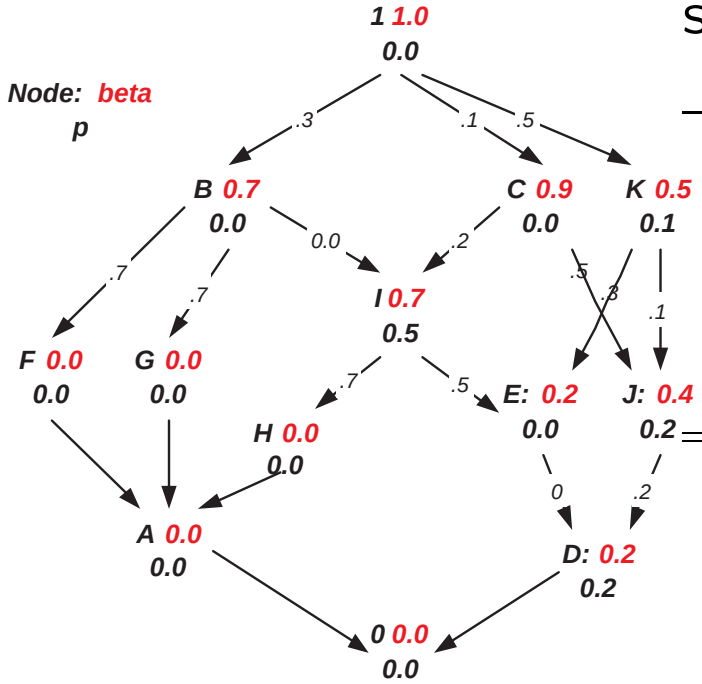






WEIGHTED NORMALIZED PSEUDO-DISTANCES

Definition: Let $\delta^w(a,b)$ be any function such that $\bar{v}_*(a,b) \leq \delta^w(a,b) \leq \bar{v}^*(a,b)$



j	h_j	$\bar{h}_j$	$\bar{v}_j$
1	3.000	0.600	0.903
2	3.000	0.600	0.903
3	3.000	0.600	0.903
4	4.000	0.800	0.956
5	4.000	0.800	0.956

	δ_*	$\bar{\delta}_*$	δ^w_*
m	3.000	0.600	0.903
x	4.000	0.800	0.956
ax	3.500	0.700	0.930
ap	3.400	0.680	0.924

Minimum: $\delta^w_m(a,b) := \bar{v}_*(a,b)$, **Maximum:** $\delta^w_x(a,b) := \bar{v}^*(a,b)$.

Average of Extremes: $$\delta^w_{ax}(a,b) := \frac{\bar{v}_*(a,b) + \bar{v}^*(a,b)}{2}.$$

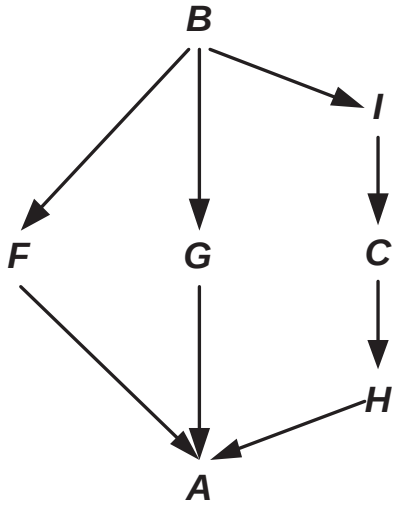
Average of All: $$\delta^w_{ap}(a,b) := \frac{\sum_{\bar{v}_j \in \vec{\bar{v}}(a,b)} \bar{v}_j}{M}.$$

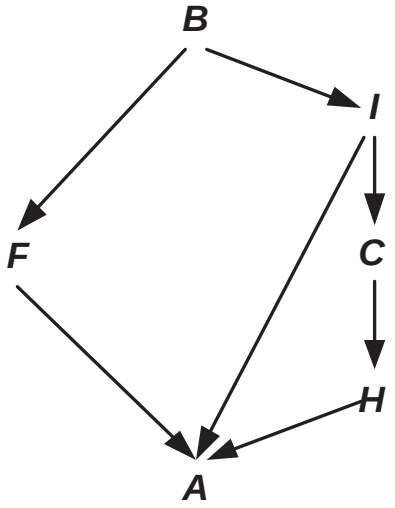
Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

LIMITATIONS OF CHAIN LENGTH VECTORS





Not all information about poset interval $[a, b]$ captured by vector of chain lengths $\vec{h}(a, b)$, nor, thus, by $\vec{v}(a, b)$

Example: $\vec{h}^1(A, B) = \vec{h}^2(A, B) = \langle 2, 2, 4 \rangle$

$$\begin{aligned} \mathcal{C}^1(A, B) &= \{A \Leftarrow F \Leftarrow B, A \Leftarrow G \Leftarrow B, A \Leftarrow H \Leftarrow C \Leftarrow I \Leftarrow B\} \\ &= \{C_1^1, C_2^1, C_3^1\} \\ \mathcal{C}^2(A, B) &= \{A \Leftarrow F \Leftarrow B, A \Leftarrow I \Leftarrow B, A \Leftarrow H \Leftarrow C \Leftarrow I \Leftarrow B\} \\ &= \{C_1^2, C_2^2, C_3^2\}, \end{aligned}$$

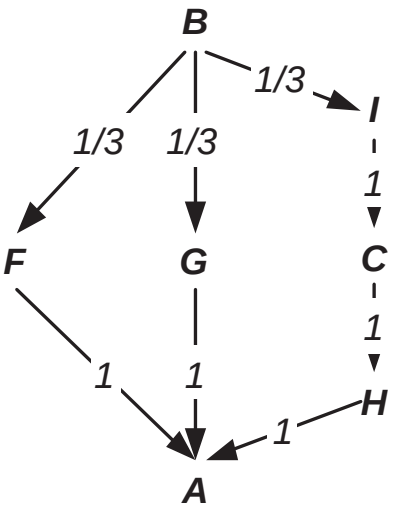
A is "closer" to B in $\mathcal{P}^2$ than in $\mathcal{P}^1$: $C_2^2 \cap C_3^2 = \{I \Leftarrow B\}$.

- $|\mathcal{P}^1| = 7 > 6 = |\mathcal{P}^2|$
- $\mathcal{W}(\mathcal{P}^1) = 3 > 2 = \mathcal{W}(\mathcal{P}^2)$

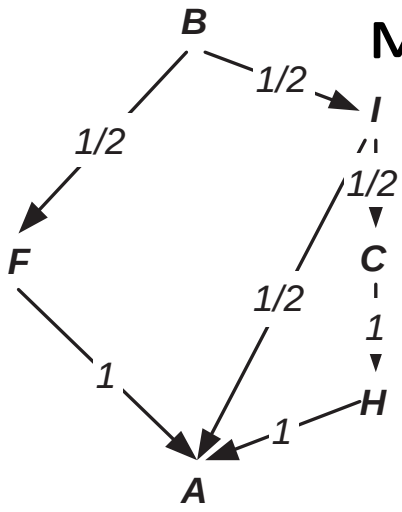
Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

MARKOV PROCESS APPROACH



<1/3,1/3,1/3>



<1/2,1/4,1/4>

Markov Process:

$p\colon \Gamma \mapsto [0,1], p(a \Leftarrow b) = p(\gamma(a,b)) = p(a|b)$ is a conditional probability of a given b:

$\sum_{a \in \downarrow\!\!\cdot(b)} p(a \Leftarrow b) = \sum_{a \in \downarrow\!\!\cdot(b)} p(a|b) = 1$

Equiprobable Distribution:

$$a <\cdot\, b \rightarrow p(a|b) = \frac{1}{|\downarrow\!\!\cdot(b)|}.$$

Chain Probability: For $a \leq b \in P, C_j \in \mathcal{C}(a,b)$:

$$p(C_j) = p(a|b) = \prod_{\gamma_i \in C_j} p(\gamma_i).$$

Vector of Chain Probabilities:

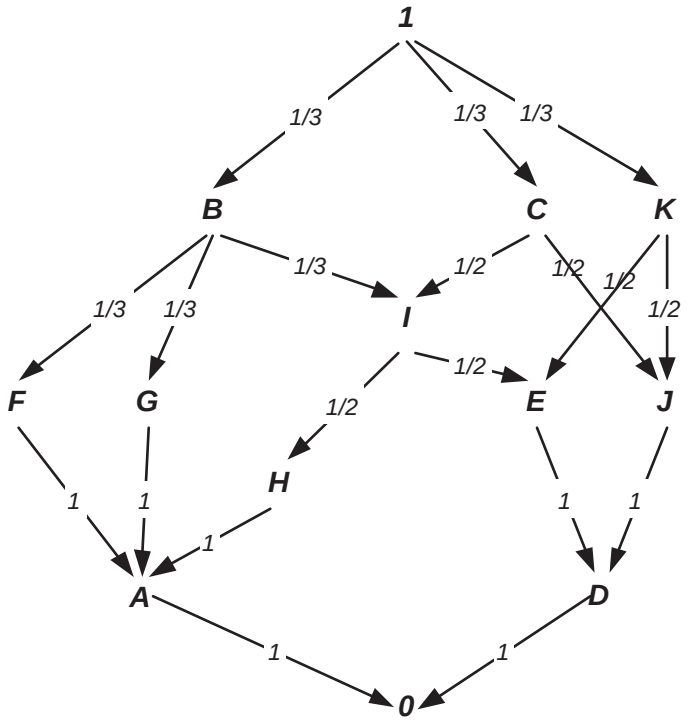
$$\vec{p}(a \leq b) := \left\langle p(C_1), \ldots, p(C_j), \ldots, p(C_M) \right\rangle = \left\langle p_1, \ldots, p_j, \ldots, p_M \right\rangle$$

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

MARKOV PROCESS APPROACH (CONT.)



- Discrete Markov processes, linear algebraic formulation;
- Bayesian nets;
- Branching processes; diffusion problems

Example: $M = 9$

$$\vec{p}(0 \leq 1) = \langle 1/18, 1/18, 1/12, 1/12, 1/9, 1/9, 1/6, 1/6, 1/6 \rangle$$

Proposition: $\forall a \leq b \in P$,

$$\sum_{C_j \in \mathcal{C}(a,b)} p(C_j) = \sum_{p_j \in \vec{p}(a \leq b)} p_j = 1.$$

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

DISTINGUISHING MEASURES

Relative Entropy:

$$\widehat{H}(\vec{p}(a \leq b)) = \frac{-\sum_{p_j \in \vec{p}(a \leq b)} p_j \log_2(p_j)}{\log_2(M)},$$

Number of Chains: $\log_2(M)$

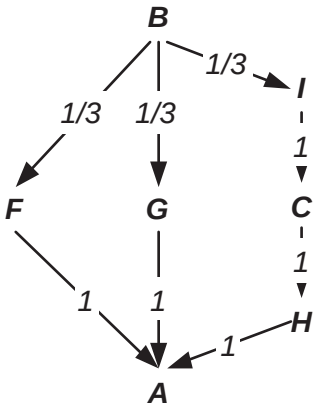
Proposition: Assume an equiprobable Markov process on Γ. Then $\forall a \leq b \in P$:

- $0 \leq \widehat{H}(\vec{p}(a \leq b)) \leq 1$.
- $\widehat{H}(\vec{p}(a \leq b)) = 1$ iff all the chains C_j are disjoint, so that $M = \mathcal{W}([a,b]) > 1$.
- Defining $\frac{0}{\log(1)} = 0$, then $\widehat{H}(\vec{p}(a \leq b)) = 0$ iff $[a,b]$ is a chain, so that $M = \mathcal{W}([a,b]) = 1$,

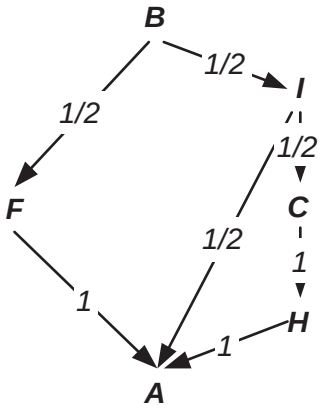
Example:

$$\widehat{H}(\vec{p}^{\,1}(A \leq B)) = \widehat{H}(\langle 1/3, 1/3, 1/3 \rangle) = 1.000$$
$$\widehat{H}(\vec{p}^{\,2}(A \leq B)) = \widehat{H}(\langle 1/2, 1/4, 1/4 \rangle) = 0.946$$
$$\widehat{H}(\vec{p}(0 \leq 1)) = \widehat{H}(\langle 2 \times 1/18, 2 \times 1/12, 2 \times 1/9, 3 \times 1/6 \rangle) = 0.965$$

MARKOV PSEUDO-DISTANCE



<1/3,1/3,1/3>



<1/2,1/4,1/4>

Definition: Expected value of the chain length from a up to b for $a \leq b \in P$:

$$\delta_p(a,b) := \sum_{C_j \in \mathcal{C}(a,b)} h_j p_j(C_j) = \vec{h} \cdot \vec{p}(a \leq b)$$

Proposition: Since $\delta_p(a,b)$ is a weighted average of the chain lengths, therefore it is a pseudo-distance, that is, $\forall a \leq b \in P, h_*(a,b) \leq \delta_p(a,b) \leq h^*(a,b)$.

Proposition: If $\widehat{H}(\vec{p}(a \leq b)) = 1$ then the Markov pseudo-distance is equivalent to the average of all chain lengths: $\delta_p = \delta_{ap}$. Thus $\delta_p(a,b) \leq \delta_{ap}(a,b)$.

Example:

$$\begin{aligned} \delta_p^1(A,B) &= \langle 2,2,4 \rangle \cdot \langle 1/3, 1/3, 1/3 \rangle = 2.67 \\ &= \delta_{ap}^1(A,B) \\ &\geq \delta_p^2(A,B) = \langle 2,2,4 \rangle \cdot \langle 1/2, 1/4, 1/4 \rangle = 2.50. \end{aligned}$$

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

ONTOLOGY MERGING AND MATCHING

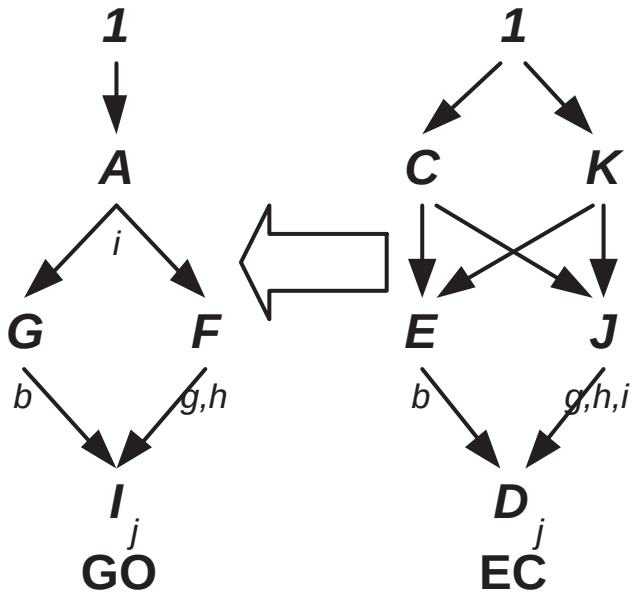
Problem: Vast, huge applicability, huge opportunity

Matching: Between two *parts* of *one* poset:

$$\mathcal{P} = \langle P, \leq \rangle, \quad P_1, P_2 \subseteq P$$

inducing

$$\mathcal{P}_1 = \left\langle P_1, \leq|_{P_1} \right\rangle,$$
$$\mathcal{P}_2 = \left\langle P_2, \leq|_{P_2} \right\rangle$$



Comparing: Two orders of the same set:

$$\mathcal{P}_1 := \langle P, \leq_1 \rangle, \mathcal{P}_2 := \langle P, \leq_2 \rangle$$

Merging: Two different posets $\mathcal{P}_1 := \langle P_1, \leq_1 \rangle, \mathcal{P}_2 := \langle P_2, \leq_2 \rangle$

- Structurally similar regions
- Similar annotations to common labels X:

$$F_1 \colon X \mapsto 2^{P_1}, \quad F_2 \colon X \mapsto 2^{P_2}$$

Cliff Joslyn, joslyn@lanl.gov

CURRENT APPROACHES

- Assume two distinct posets $\mathcal{P}_1 = \left\langle P^1, \leq_1 \right\rangle, \mathcal{P}_2 = \left\langle P^2, \leq_2 \right\rangle$
- Assume "anchoring" nodes $a^1, b^1, \ldots \in P^1, a^2, b^2, \ldots \in P^2$, which are equated between them so that $a^1 = a^2, b^1 = b^2$
- Build a common ontology around these anchors
- **Extraordinarily preliminary thoughts, looking for help**

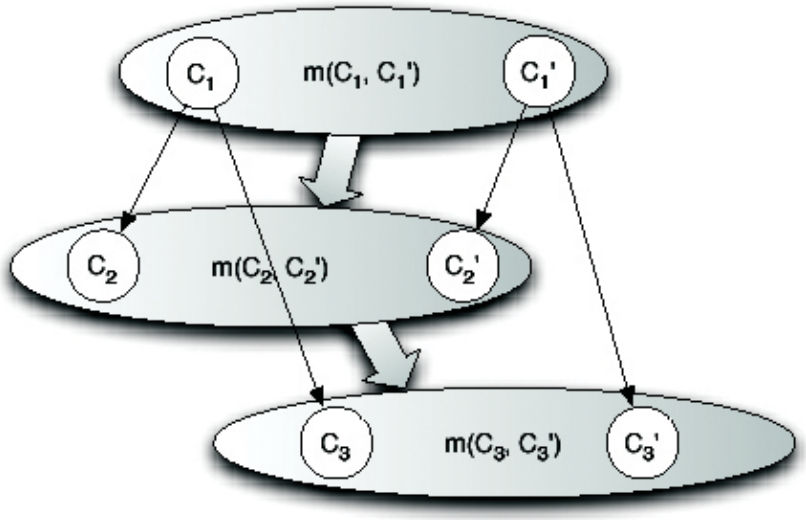


Fig. 1. Subgraphs representing some concepts in ontologies O and O' (small circles) and relations between them (thin arrows). The large gray ovals and solid arrows represent a snippet of the BN graph with nodes corresponding to matches and arrows corresponding to influences in the BN graph.

NF Noy (2004): "Semantic Integration: A Survey Of Ontology-Based Approaches", SIGMOD Record, Special Issue on Semantic Integration, 33 (4), December, 2004

M Prasenjit, NF Noy, AT Jaiswal (2004): "OMEN: A Probabilistic Ontology Mapping Tool", Workshop on Meaning Coordination and Negotiation at the 3rd Int. Conf. on the Semantic Web (ISWC-2004)

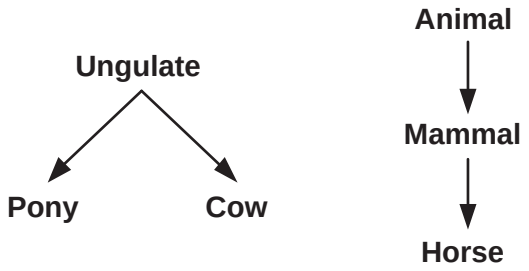
Luger, Sarah; Aitken, Stuart; and Webber, Bonnie: (2005) "Cross-Species Mapping Between Anatomical Ontologies: Terminological and Structural Support", poster at 2004 Conf. on Intelligent Systems for Molecular Biology (ISMB 04)

Los Alamos NATIONAL LABORATORY

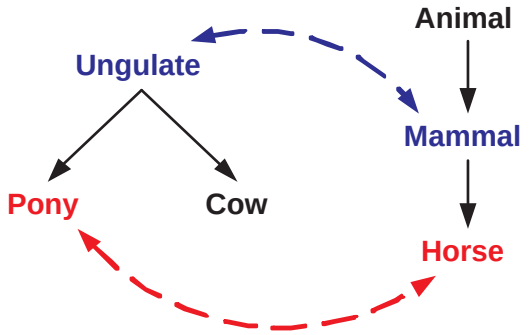
Modeling, Algorithms, & Informatics CCS3

INTEROPERABILITY FORMULATION

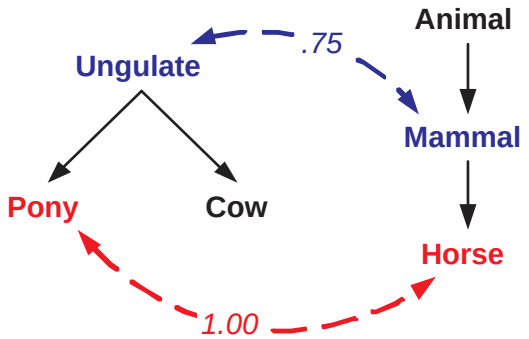
- Similar ontologies



- Anchors between posets



- Weighted anchoring nodes



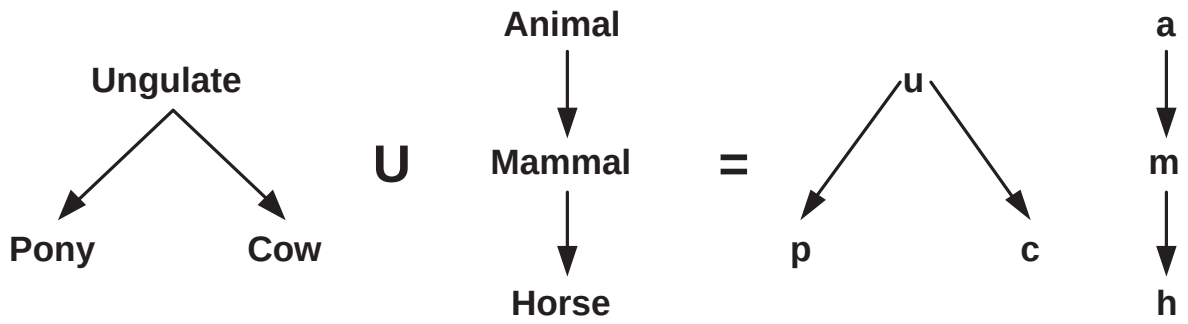
Los Alamos NATIONAL LABORATORY

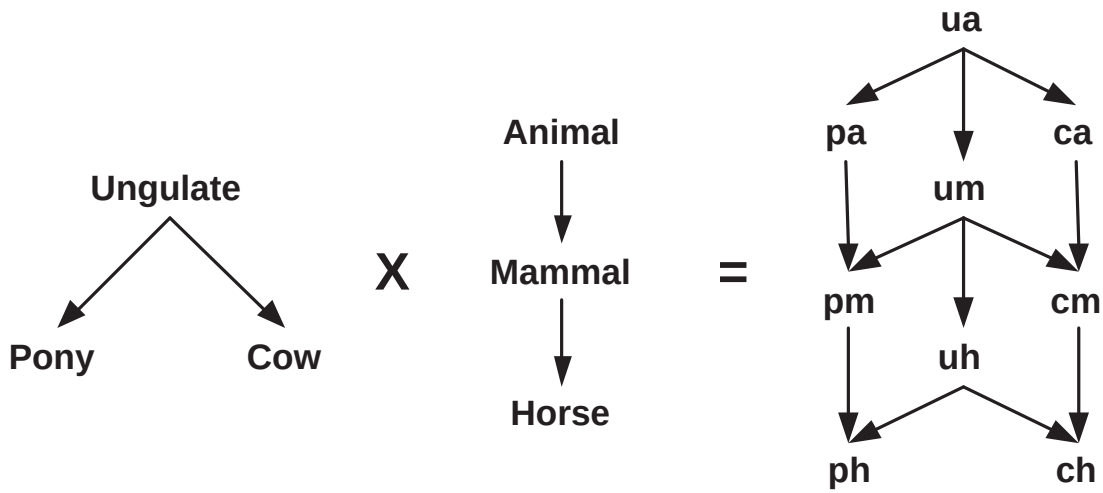
Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

FORMALIZATION

- Establishing a common universe of discourse: $P_1 \cup P_2$ or $P_1 \times P_2$?

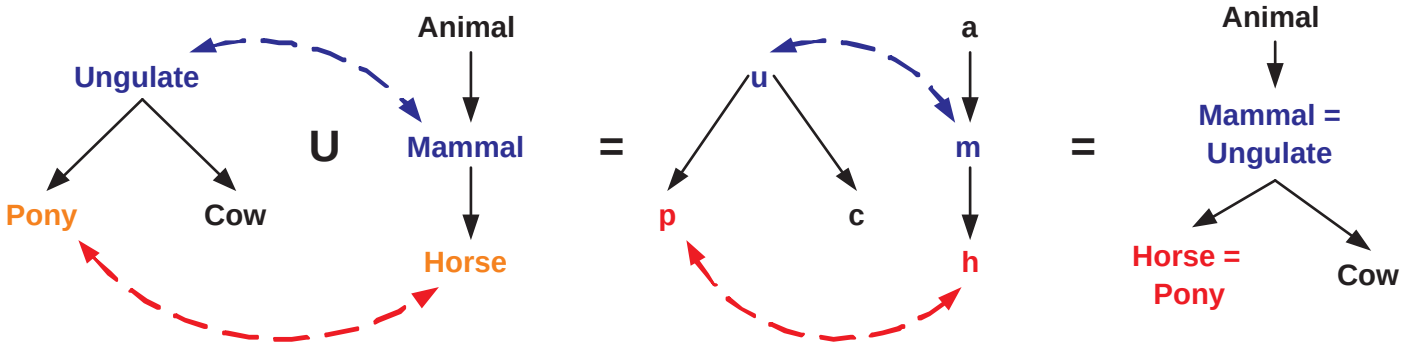




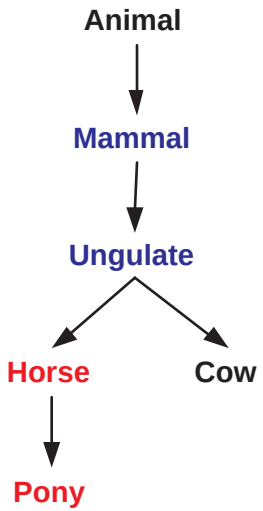
Cliff Joslyn, joslyn@lanl.gov

SUM APPROACH

- Beyond the disjoint union, identify $\mathcal{P} = \mathcal{P}_1 \oplus \mathcal{P}_2$
- Establish $P_1 \cap P_2 \neq \emptyset$
- Anchors provide intersecting nodes $p = h, u = m$



- "Real" solution:



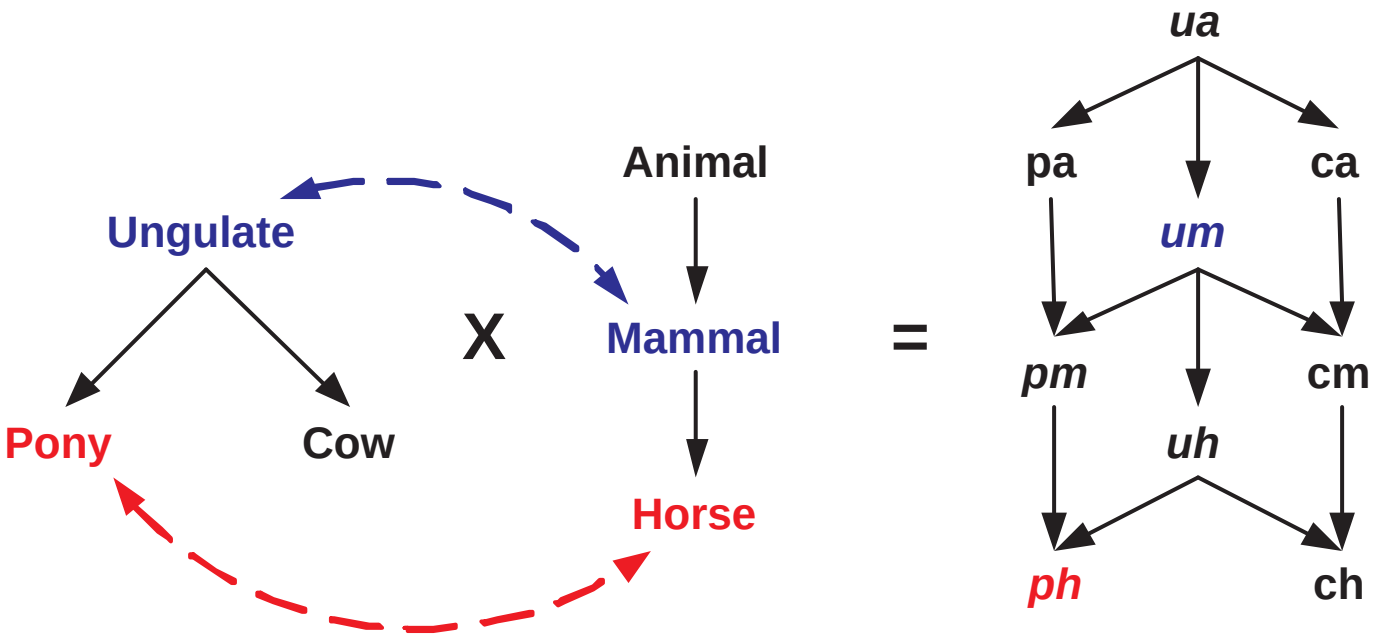
Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

PRODUCT APPROACH

- Identify sub-order of $\mathcal{P} \subseteq \mathcal{P}_1 \times \mathcal{P}_2$
- Restrict order by anochoring pairs: $\langle p, h \rangle, \langle u, m \rangle \in \mathcal{P}$



Cliff Joslyn, joslyn@lanl.gov

WEIGHTED ANCHORS

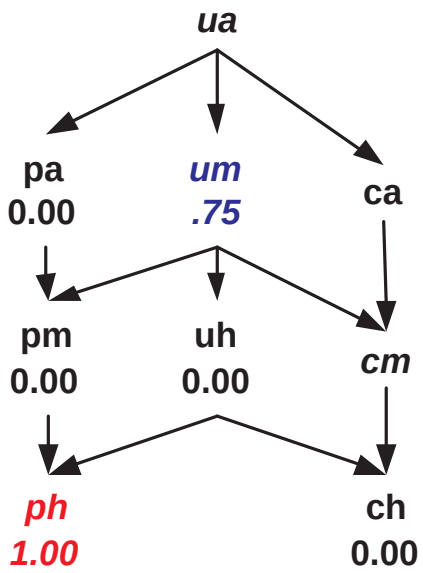
- Matrix of anchoring nodes:

	a	m	h
u		.75	
p			1.00
c			

- Normalization in a fuzzy matrix

	a	m	h
u	?	.75	0.00
p	0.00	0.00	1.00
c	?	?	0.00

- Implies working in the product:



Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3

Cliff Joslyn, joslyn@lanl.gov

ACKNOWLEDGEMENTS

LANL Computer Science:

- Susan Mniszewski
- Karin Verspoor
- Michael Wall

LANL Biosciences:

- Michael Altherr
- Judith Cohn
- Andreas Rechtsteiner
- Tom Terwilliger

LANL Theoretical Division:

- Bill Bruno

Procter & Gamble Corp.:

- Andy Fulmer
- Gary Heaton

U. Manchester CS:

- Phillip Lord
- Robert Stevens
- Alex Sanchez

Old Dominion U. CS:

- Alex Pothen

Stanford U. Med. Info.:

- Natasha Noy

This work was sponsored by the Department of Energy under contract W-7405-ENG-36 to the University of California. We would like to thank the Los Alamos National Laboratory Protein Function Inference Group for their contributions to this work.

Los Alamos NATIONAL LABORATORY

Modeling, Algorithms, & Informatics CCS3