Graph Summarization in Annotated Data using Probabilistic Soft Logic

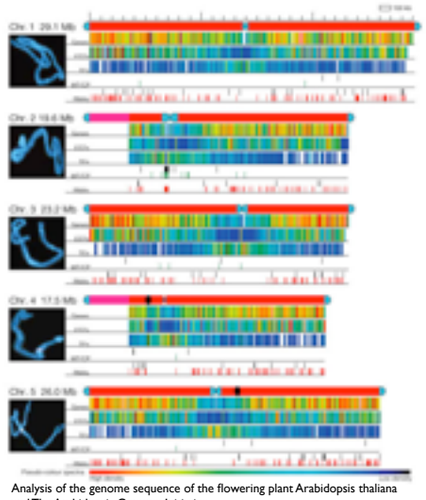
Alex Memory[1], Angelika Kimmig[1,2], Stephen H. Bach[1], Louiqa Raschid[1] and Lise Getoor[1]

[1]University of Maryland
[2]KU Leuven

Agenda

- Annotation Graphs and Graph Summarization
- Heuristics for Graph Summarization
- Probabilistic Soft Logic
- Evaluation and Results

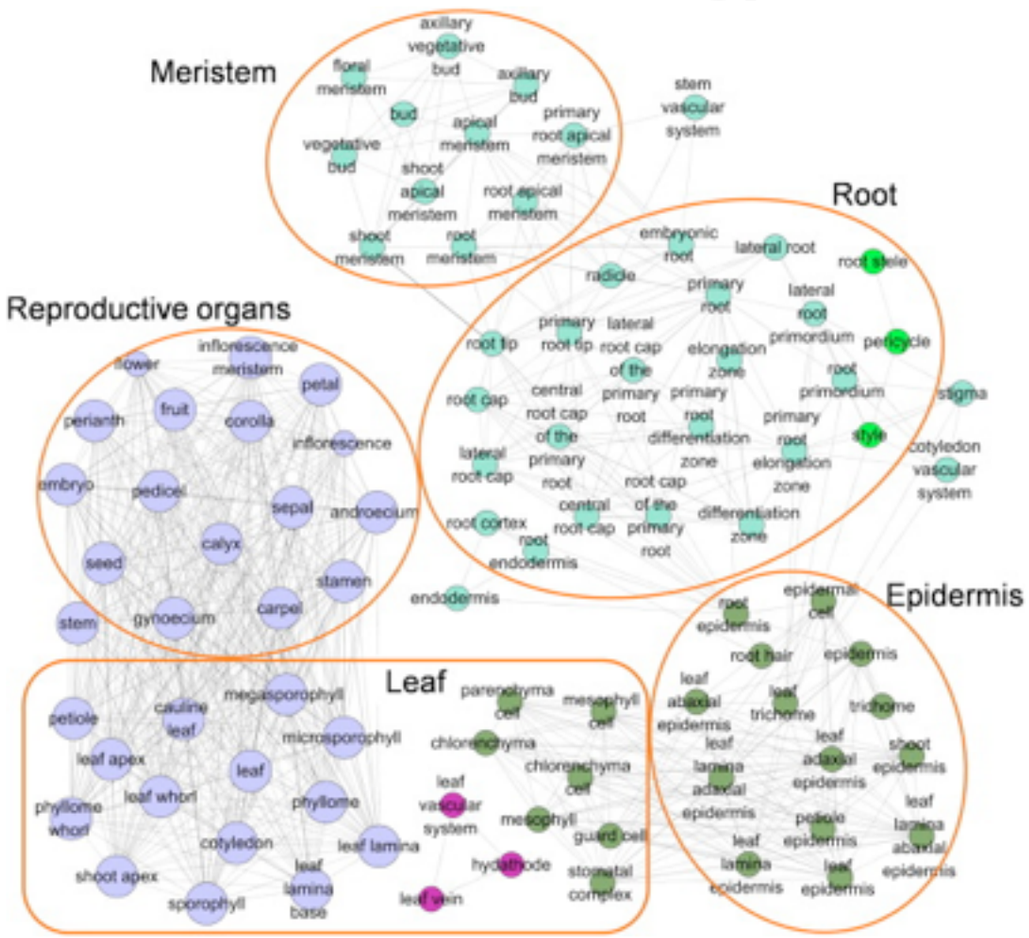
Annotated Data: the Arabidopsis thaliana Genome



Analysis of the genome sequence of the flowering plant Arabidopsis thaliana and The Arabidopsis Genome Initiative
Nature 408, 796-815(14 December 2000)
doi:10.1038/35048692

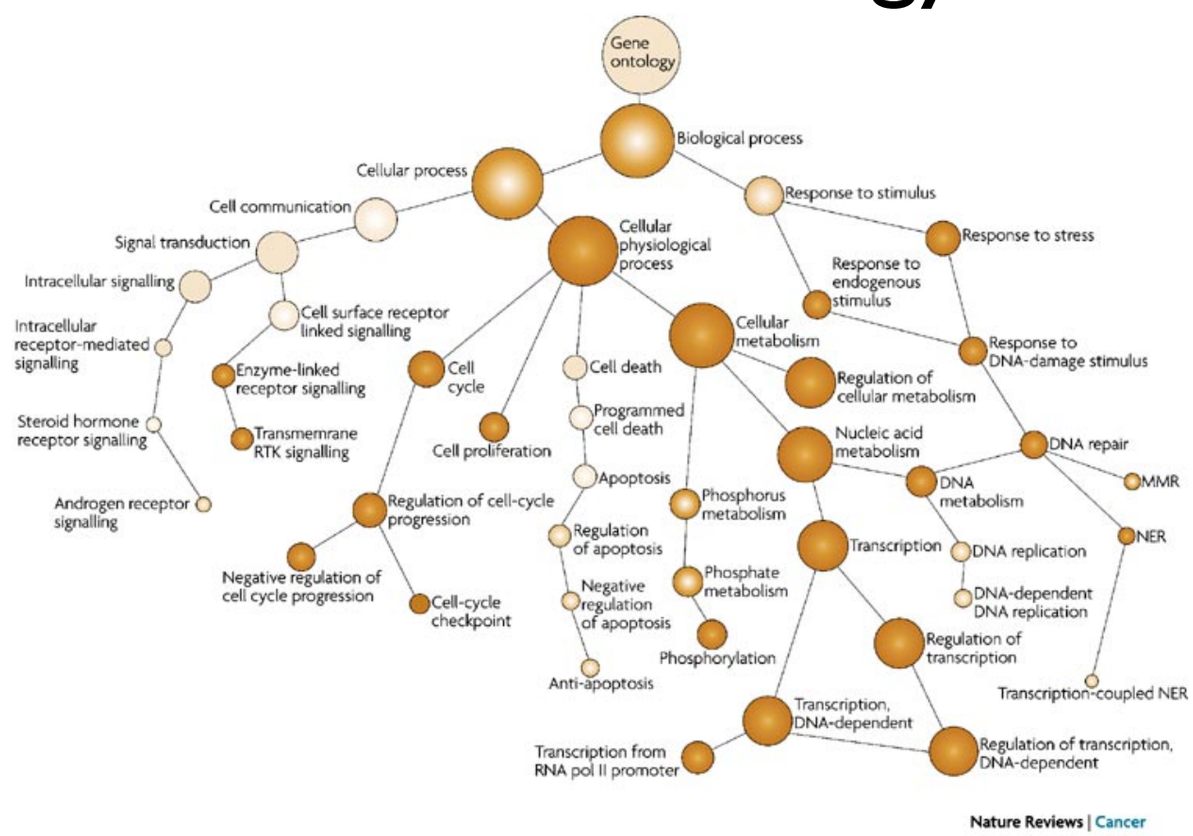
Anna Scrumillo, http://www.fotopedia.com/items/6nf9pnighbor-oaPbkJEnrX8

Plant Ontology



Xu Wanga, Yangyang Bianb, Kai Chengb, Li-Fei Gua, Mingliang Yeb, Hanfa Zoub, Jun-Xian Hea,, A large-scale protein phosphorylation analysis reveals novel phosphorylation motifs and phosphoregulatory networks in Arabidopsis, Journal of Proteomics, Oct 2012.

Gene Ontology



Hu *et al. Nature Reviews Cancer* **7**, 23–34 (January 2007) | doi:10.1038 / nrc2036

The Arabidopsis Information Resource

tair

Gene | Search

Home | Help | Contact | About Us | Login/Register

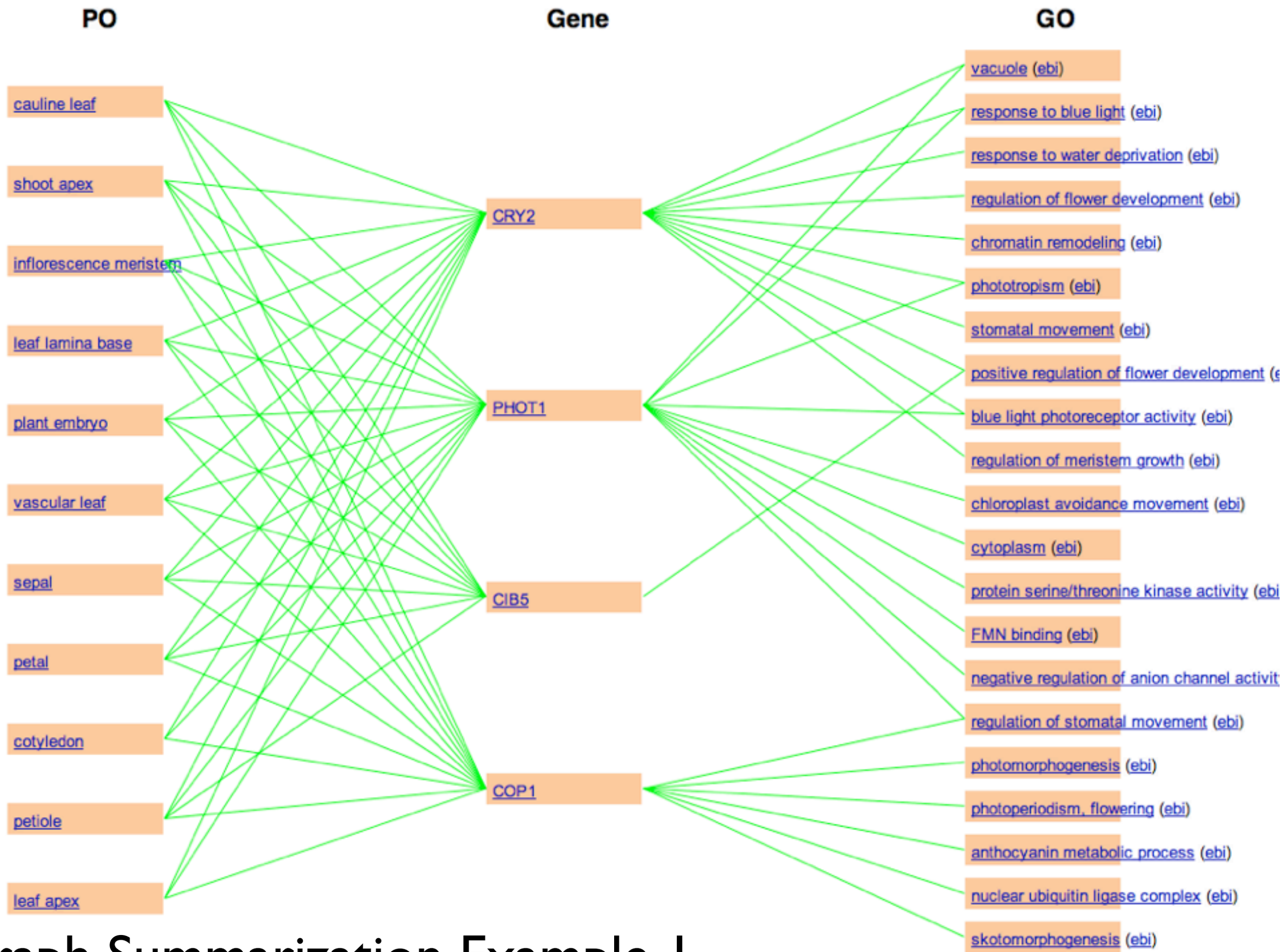
Search | Browse | Tools | Portals | Download | Submit | News | ABRC Stocks

GO Annotations

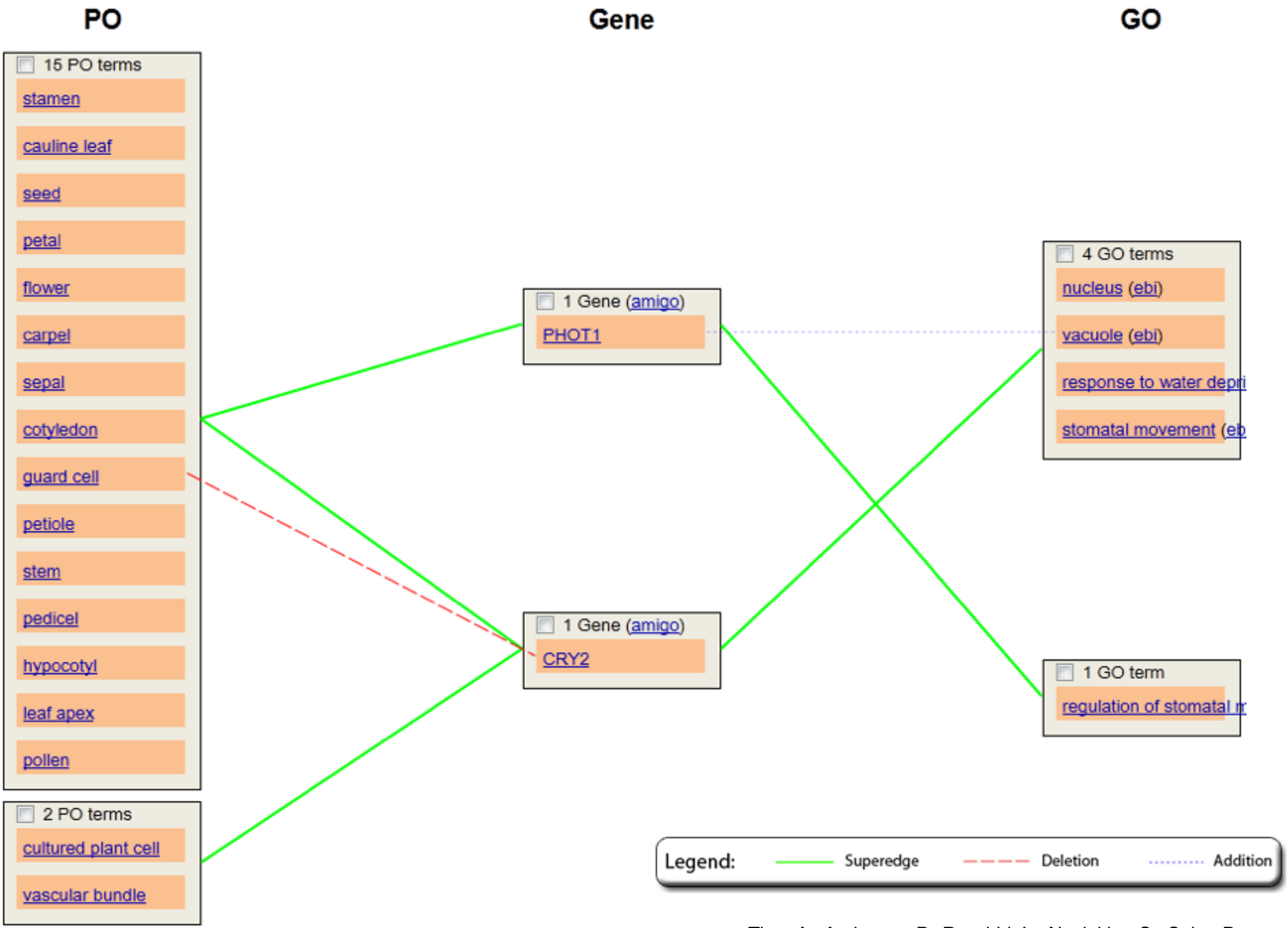
Locus	Gene Model(s)	GO term (links to Tair Keyword Browser) (GO ID)	cat	code	GO Slim	Reference	Made by: date last modified
AT1G01030	AT1G01030.1	nucleus (GO:0005634)	comp		nucleus	:501750651	NA 08/31/2012 00:00:00
	AT1G01030.1	regulation of transcription, DNA-dependent (GO:0006355)	proc		other cellular processes | other metabolic processes | transcription,DNA-dependent	:501748310	TAIR 04/18/2012 00:00:00
	AT1G01030	cellular_component_unknown (GO:0005575)	comp		unknown cellular components	:1345790	TAIR 10/19/2006 00:00:00
	AT1G01030	flower development (GO:0009908)	proc		developmental processes	:501718951	TAIR 08/29/2006 00:00:00
	AT1G01030	leaf development (GO:0048366)	proc		developmental processes	:501718951	TAIR 08/29/2006 00:00:00
	AT1G01030	sequence-specific DNA binding transcription factor activity (GO:0003700)	func		transcription factor activity	:1345963	TAIR 06/06/2003 00:00:00
	AT1G01030	regulation of transcription, DNA-dependent (GO:0006355)	proc		other cellular processes | transcription,DNA-dependent | other metabolic processes	:1345963	TAIR 03/18/2011 00:00:00
	AT1G01030.1	DNA binding (GO:0003677)	func		DNA or RNA binding	:501748310	TAIR 04/18/2012 00:00:00

v. 20040702

http://www.arabidopsis.org/cgi-bin/bulk/go/getgo.pl

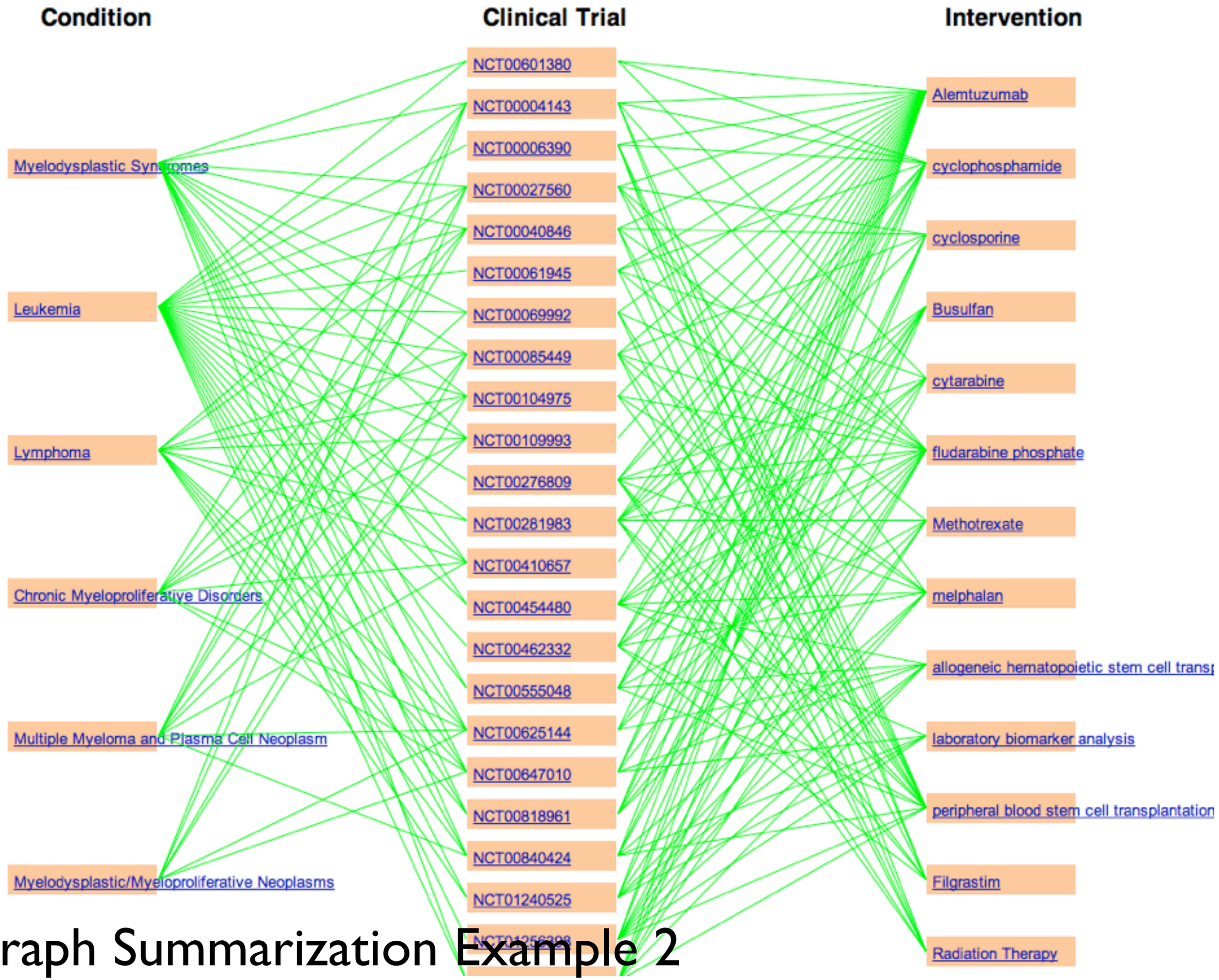


Graph Summarization Example 1

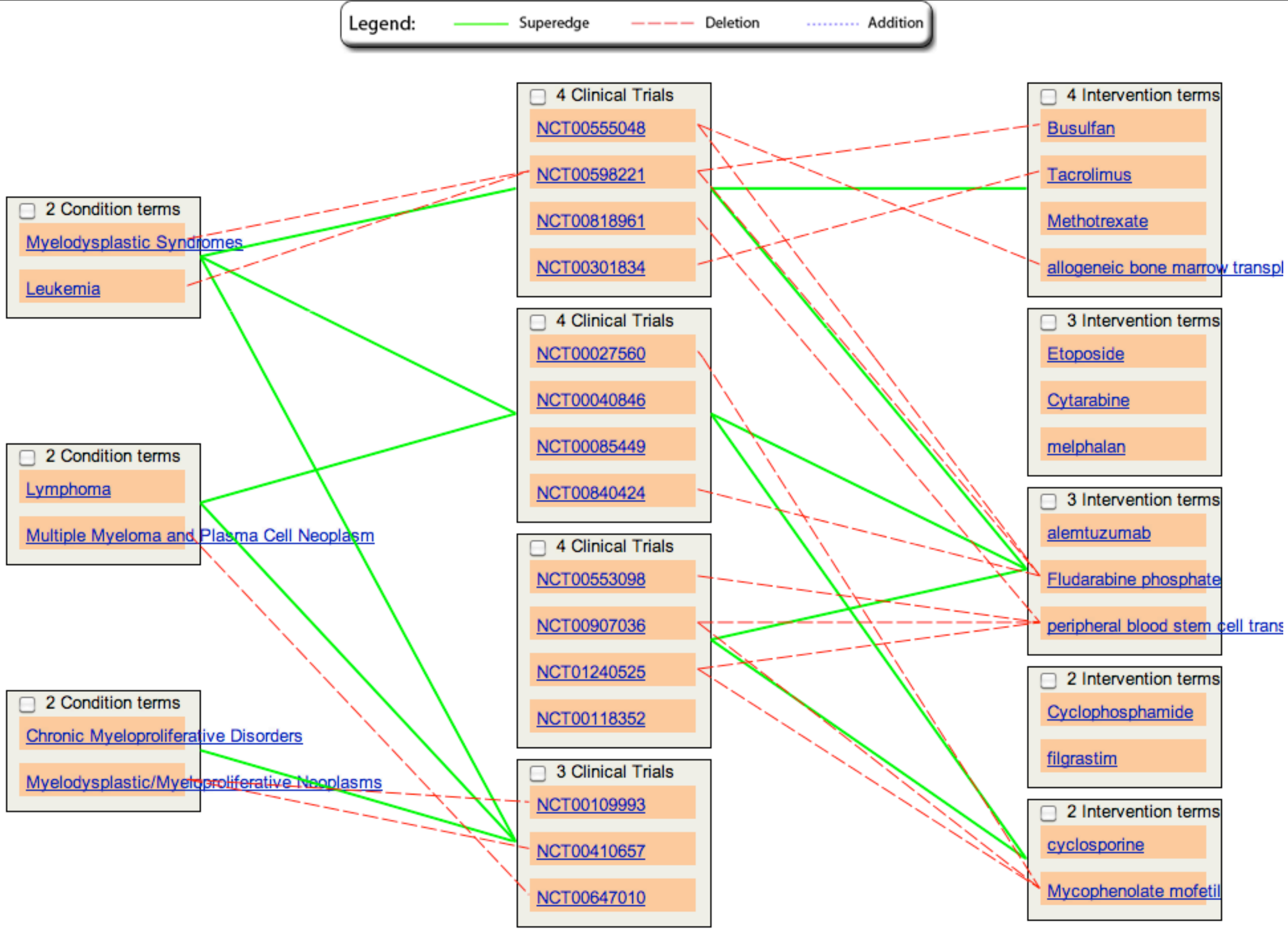


Graph Summarization Example 1

Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., Zhang, X.N.: Link prediction for annotation graphs using graph summarization. In: International Semantic Web Conference (ISWC). (2011)



Graph Summarization Example 2

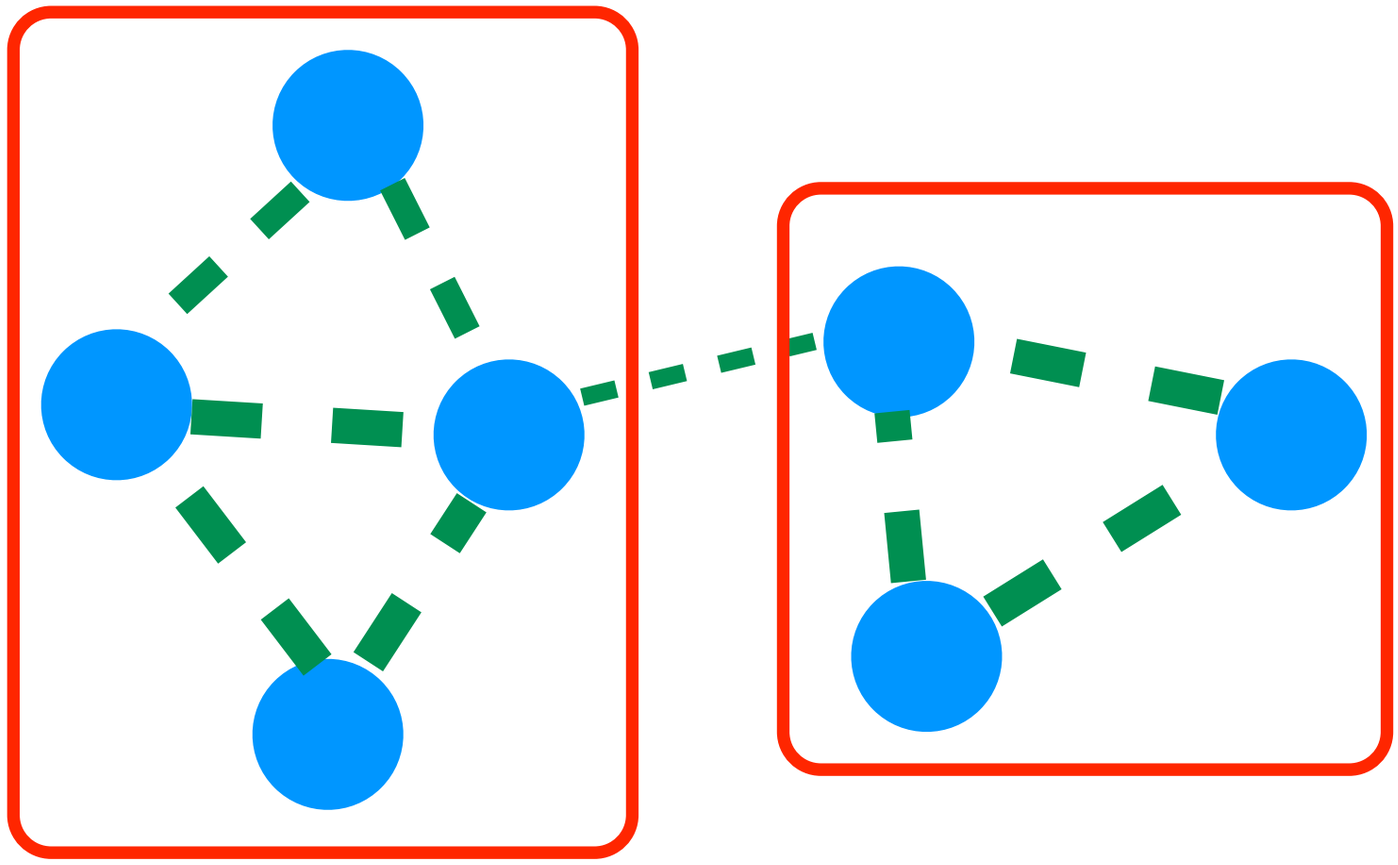


Graph Summarization Example 2

Heuristics for Graph Summarization

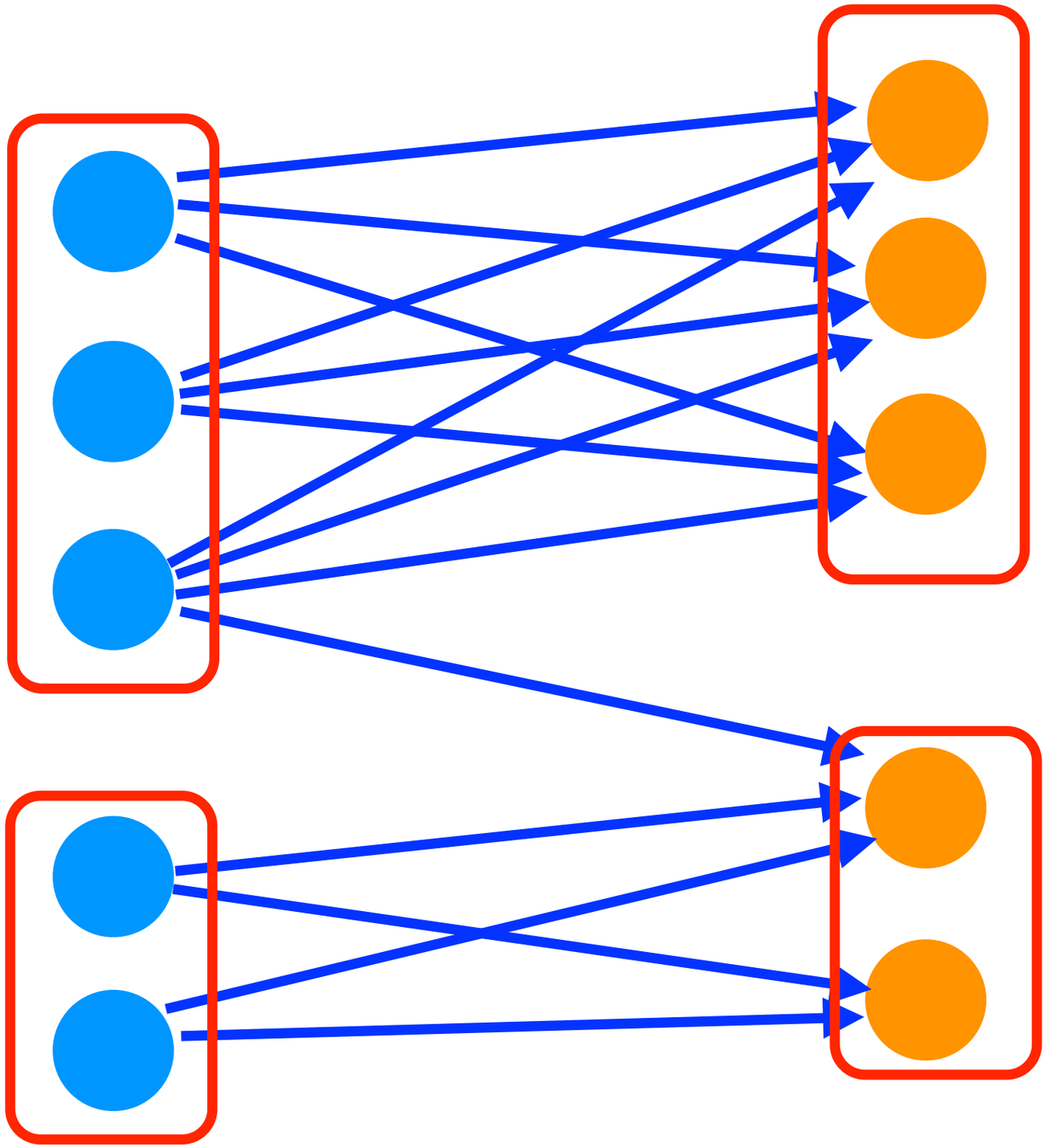
- Similarity
- Annotation Links
- Neighbor Sets

Similarity Heuristic



Nodes in same cluster should be similar

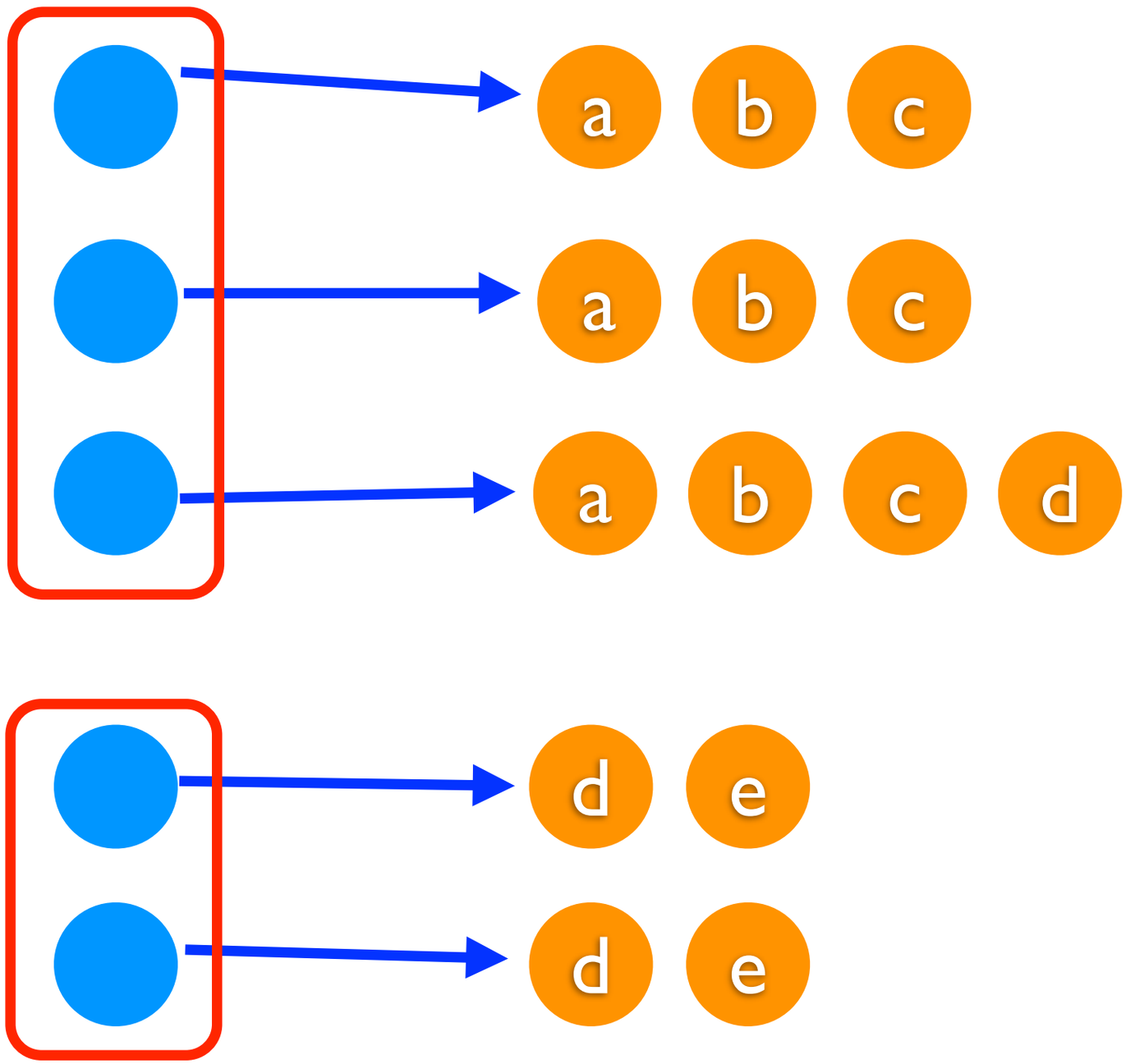
Annotation Link Heuristic



Shared neighbor → same cluster

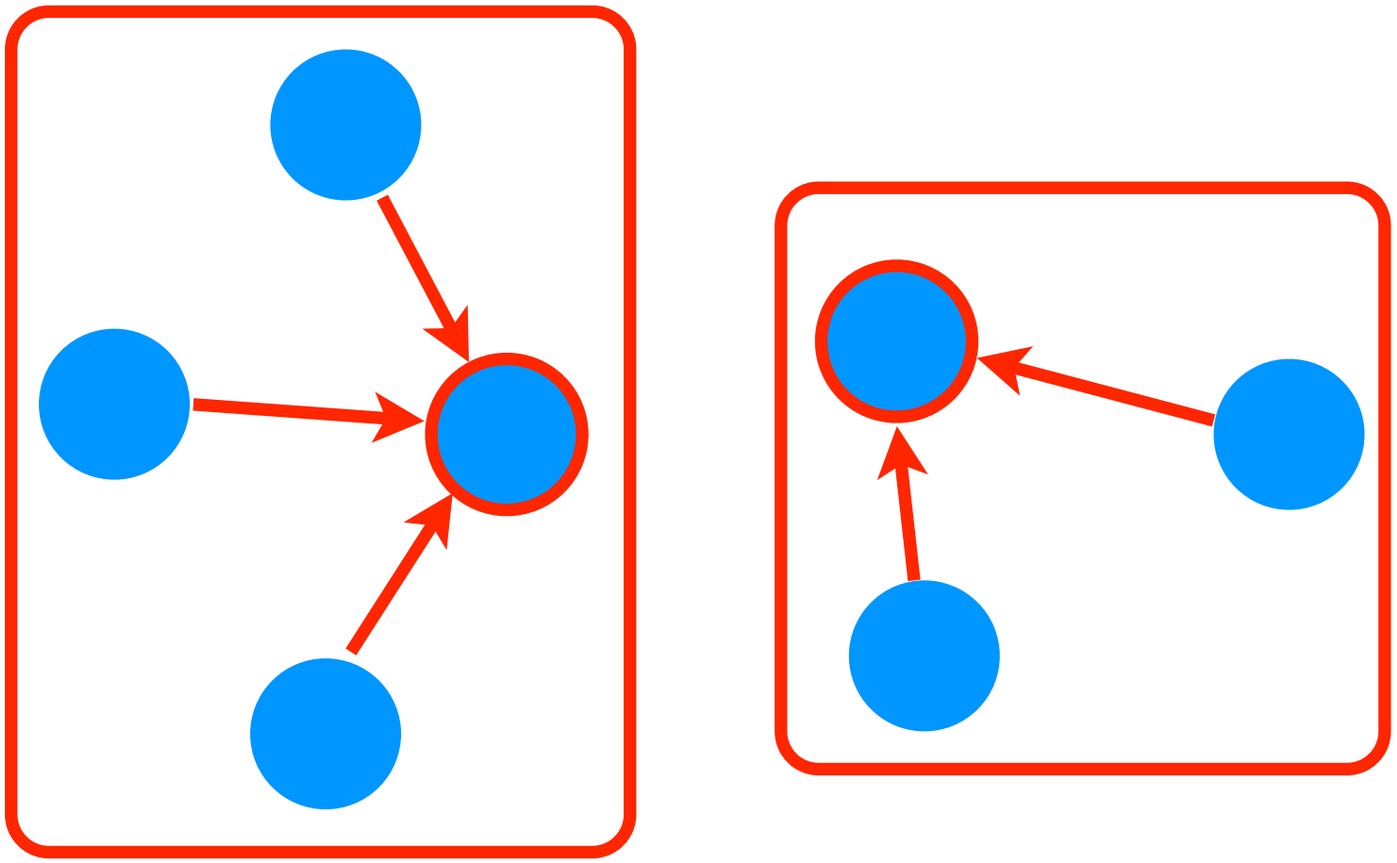
Neighbor not shared → different clusters

Neighbor Sets Heuristic



Similar set of neighbors
→ same cluster

Affinity Propagation



- one exemplar per cluster
- every node has to choose an exemplar

Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science 315 (2007) 972–976

Probabilistic Reasoning on Relational Data

- Fuzzy Logic [Lee, 1972]
- Undirected Graphical Models, FOL template language [Richardson et al, 2006]
- Probabilistic Ontologies [Costa et al, 2006]

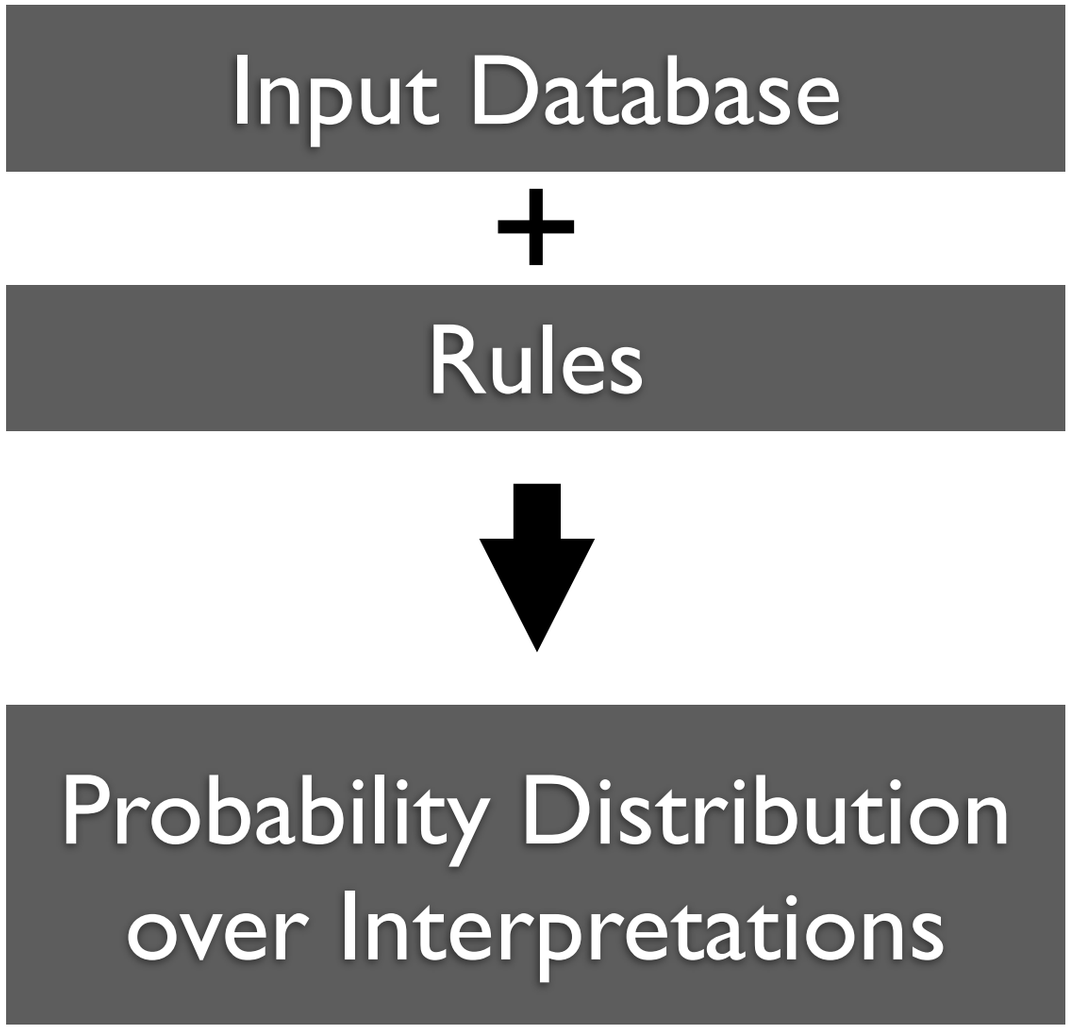
Costa, P. C. G., and K. B. Laskey. "PR-OWL: A Framework for Probabilistic Ontologies." Frontiers in Artificial Intelligence and Applications 150 (2006): 237.

Richardson, M., Domingos, P.: Markov logic networks. Machine Learning 62(1-2) (2006) 107–136

Lee, R.C.T.:Fuzzy logic and the resolution principle. J.ACM19(1)(1972)109–119

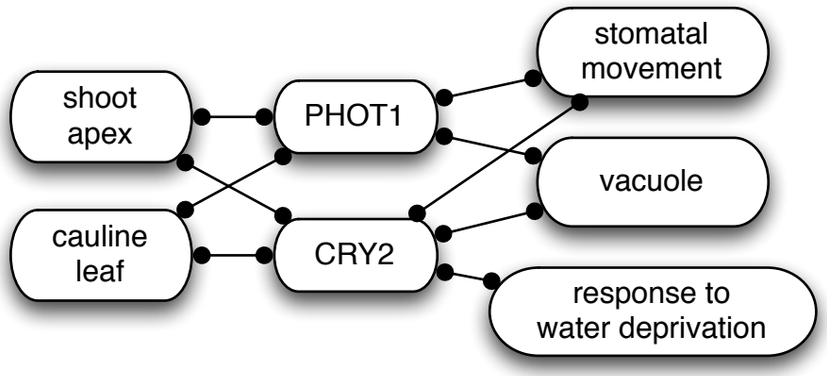
Probabilistic Soft Logic (PSL)

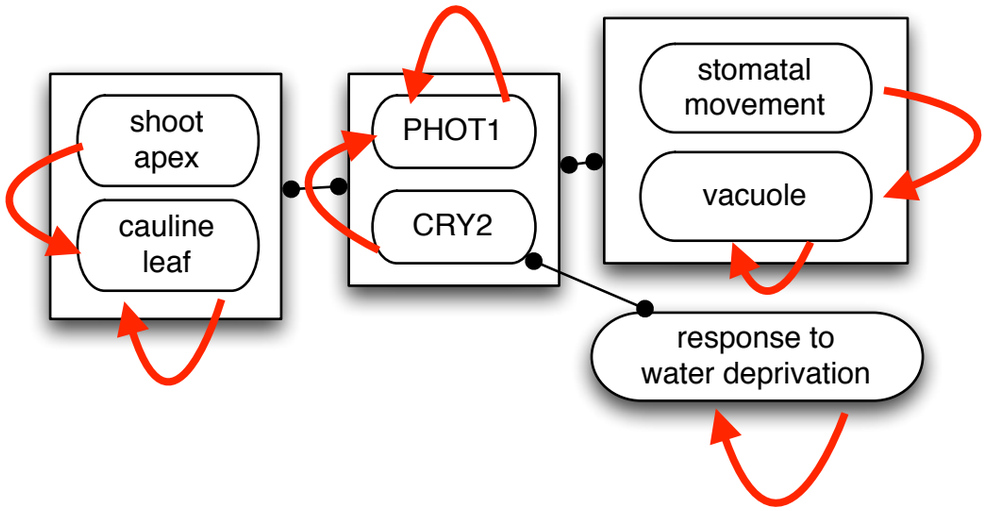
- Declarative language for probabilistic reasoning
  - First order rules
  - Soft truth values
  - Similarities
  - Set similarity
- Efficient inference
- http://psl.umiacs.umd.edu/



Broecheler, M., Mihalkova, L., Getoor, L.: Probabilistic similarity logic. In: Conference on Uncertainty in Artificial Intelligence (UAI). (2010)

Graphs and Clusters in PSL





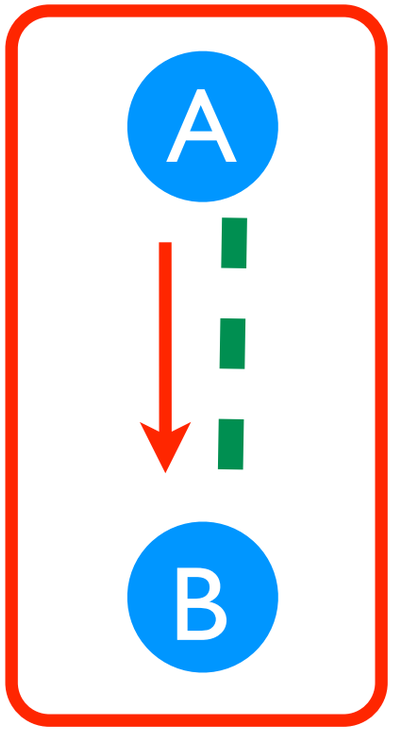

```
link(sa,p1)= 1
link(sa,c2) = 1
link(cl,p1) = 1
      ...
 link(c2,v) = 1
link(c2,rw) = 1

similar(sa,cl) = .9
similar(p1,c2) = .5
 similar(sm,v) = .1
      ...
```

```
exemplar(sa,cl)
exemplar(cl,cl)
exemplar(p1,p1)
exemplar(c2,p1)
 exemplar(sm,v)
 exemplar(v,v)
exemplar(rq,rw)
```

GS Heuristics in PSL

- Similarity Heuristic

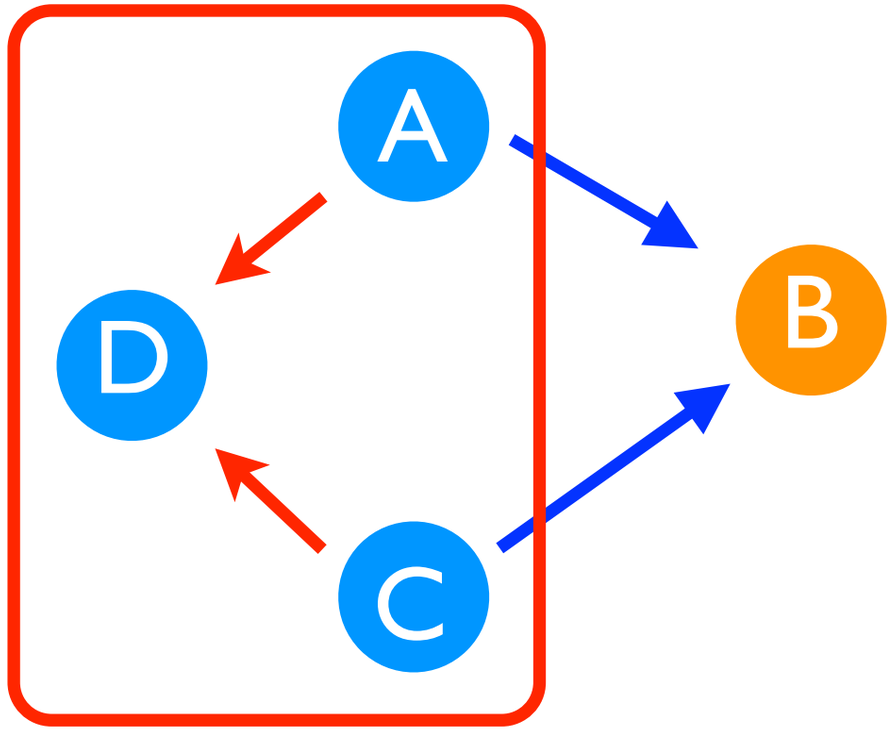
```
1 : exemplar(A,B) -> similar(A,B)
```



GS Heuristics in PSL

- Annotation Link Heuristic

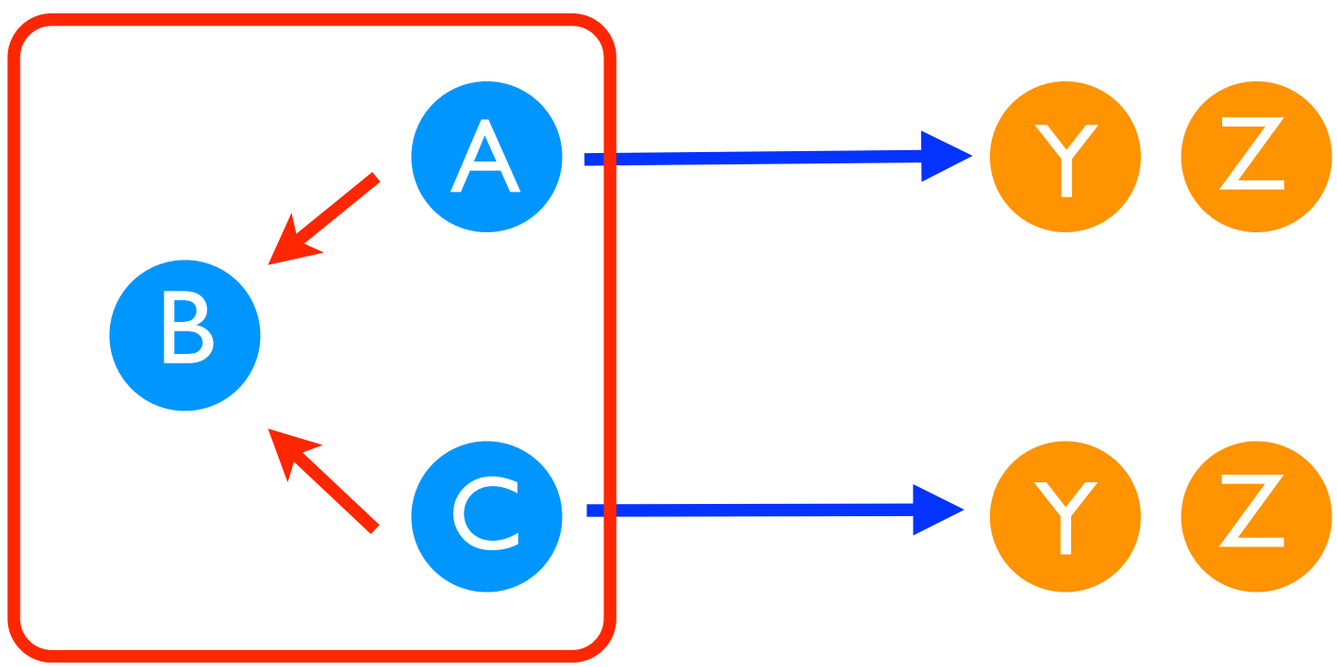
```
1 : link(A,B) & link(C,B) & exemplar(A,D) -> exemplar(C,D)
1 : link(A,B) & exemplar(A,C) & exemplar(D,C) -> link(C,B)
```



GS Heuristics in PSL

- Neighbor Sets Heuristic

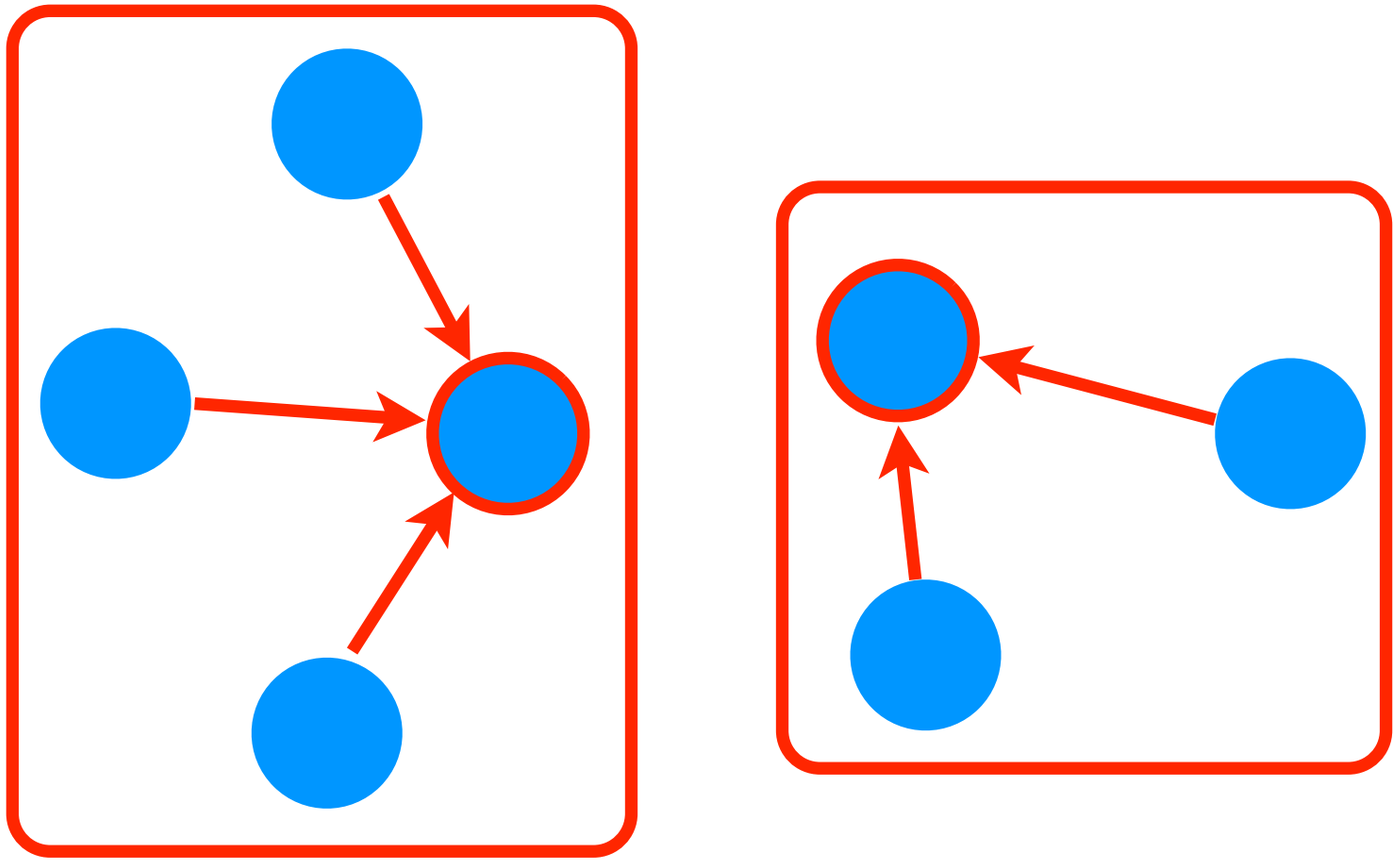
```
1 : sameNeighbors({A.link}, {C.link}) & exemplar(A,B)
                    -> exemplar(C,B)
```



GS Heuristics in PSL

- Affinity Propagation

```
1000 : exemplar(A,B) -> exemplar(B,B)
Functional: exemplar
```



GS Heuristics in PSL

- Similarity Heuristic

```
1 : exemplar(A,B) -> similar(A,B)
```

- Annotation Link Heuristic

```
1 : link(A,B) & link(C,B) & exemplar(A,D) -> exemplar(C,D)
1 : link(A,B) & exemplar(A,C) & exemplar(D,C) -> link(D,B)
```

- Neighbor Sets Heuristic

```
1 : sameNeighbors({A.link}, {C.link}) & exemplar(A,B)
                    -> exemplar(C,B)
```

- Affinity Propagation

```
1000 : exemplar(A,B) -> exemplar(B,B)

Functional: exemplar
```

Distance to Satisfaction

rule satisfied ⇔

truth value of body ≤ truth value of head

```
in1(a,b) -> out1(a,b)
```

		distance to satisfaction
1.0	0.0	1.0
0.9	0.6	0.3
0.7	0.6	0.1
0.5	0.6	0

Distance to Satisfaction

rule satisfied ⇔

truth value of body ≤ truth value of head

`out1(a,b)`	`->` `in1(a,b)`	distance to satisfaction
1.0	0.0	1.0
0.9	0.6	0.3
0.7	0.6	0.1
0.5	0.6	0

Probabilistic Model

Rule's distance to satisfaction given I

$\in\{1,2\}$

$$f(I) = \frac{1}{Z}\exp[-\sum_{r\in R}\lambda_r(d_r(I))^p]$$

Interpretation

Rule's weight

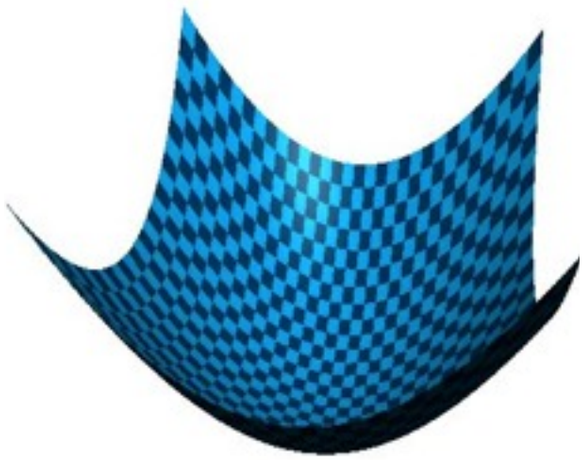
Set of rule groundings

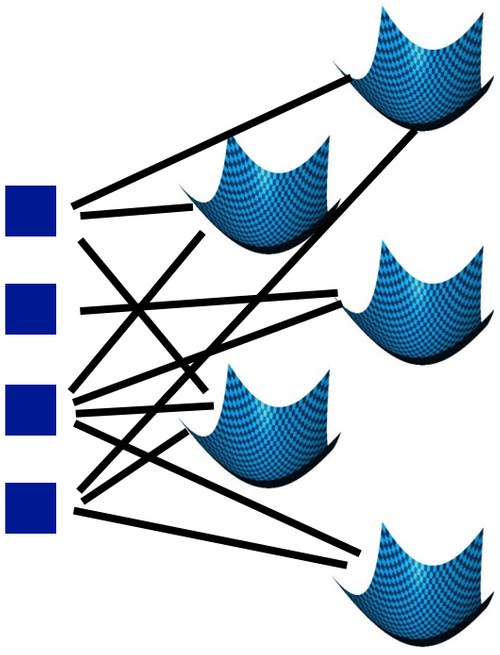
Normalization constant

$$Z = \int_I \exp[-\sum_{r\in R}\lambda_r(d_r(I))^p]$$

MPE Inference in PSL

- Convex optimization problem
- Exact inference in polynomial time by transformation to second order cone program
- New solver based on consensus optimization linear time in practice [Bach et al, NIPS 12]

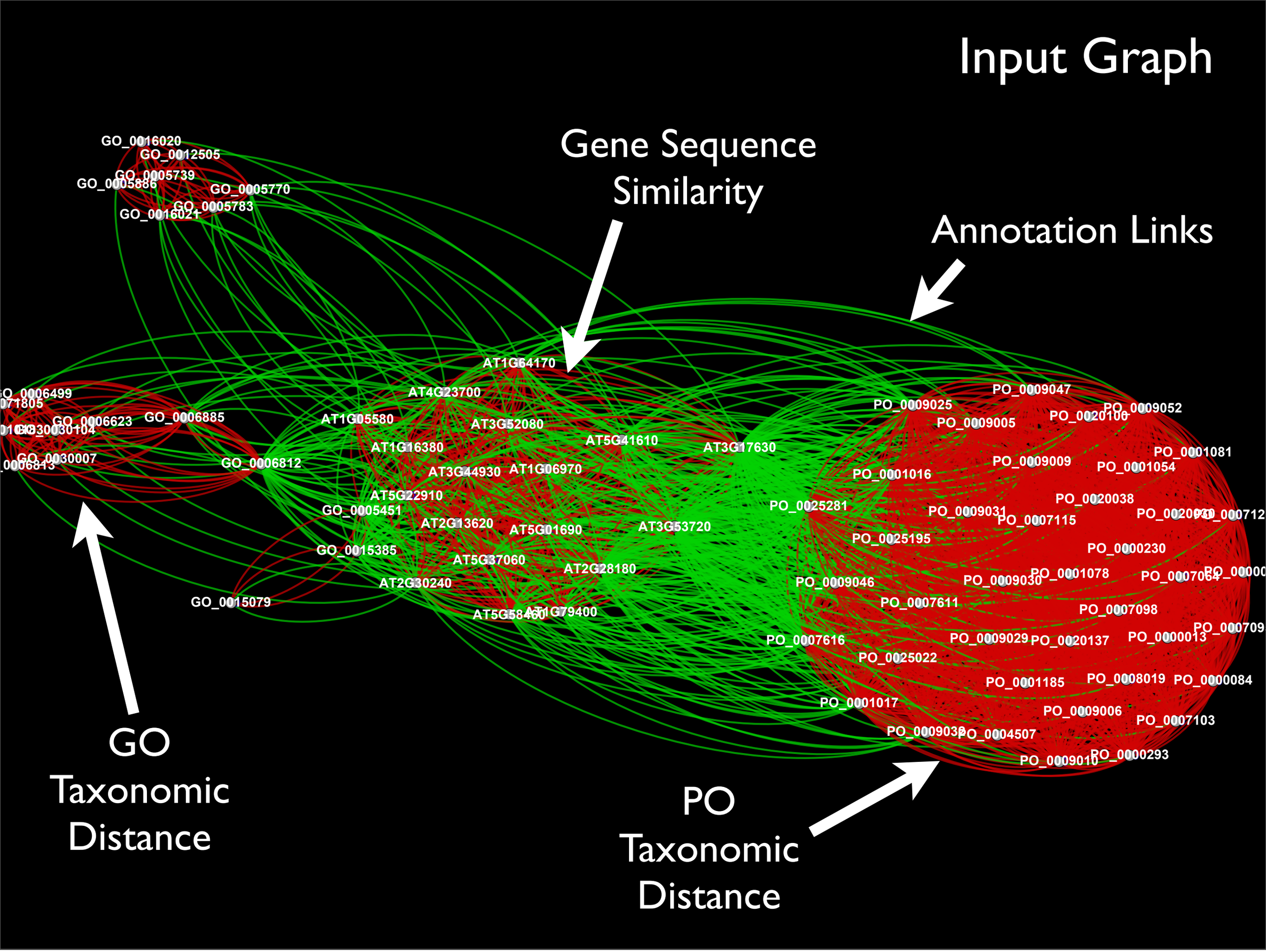




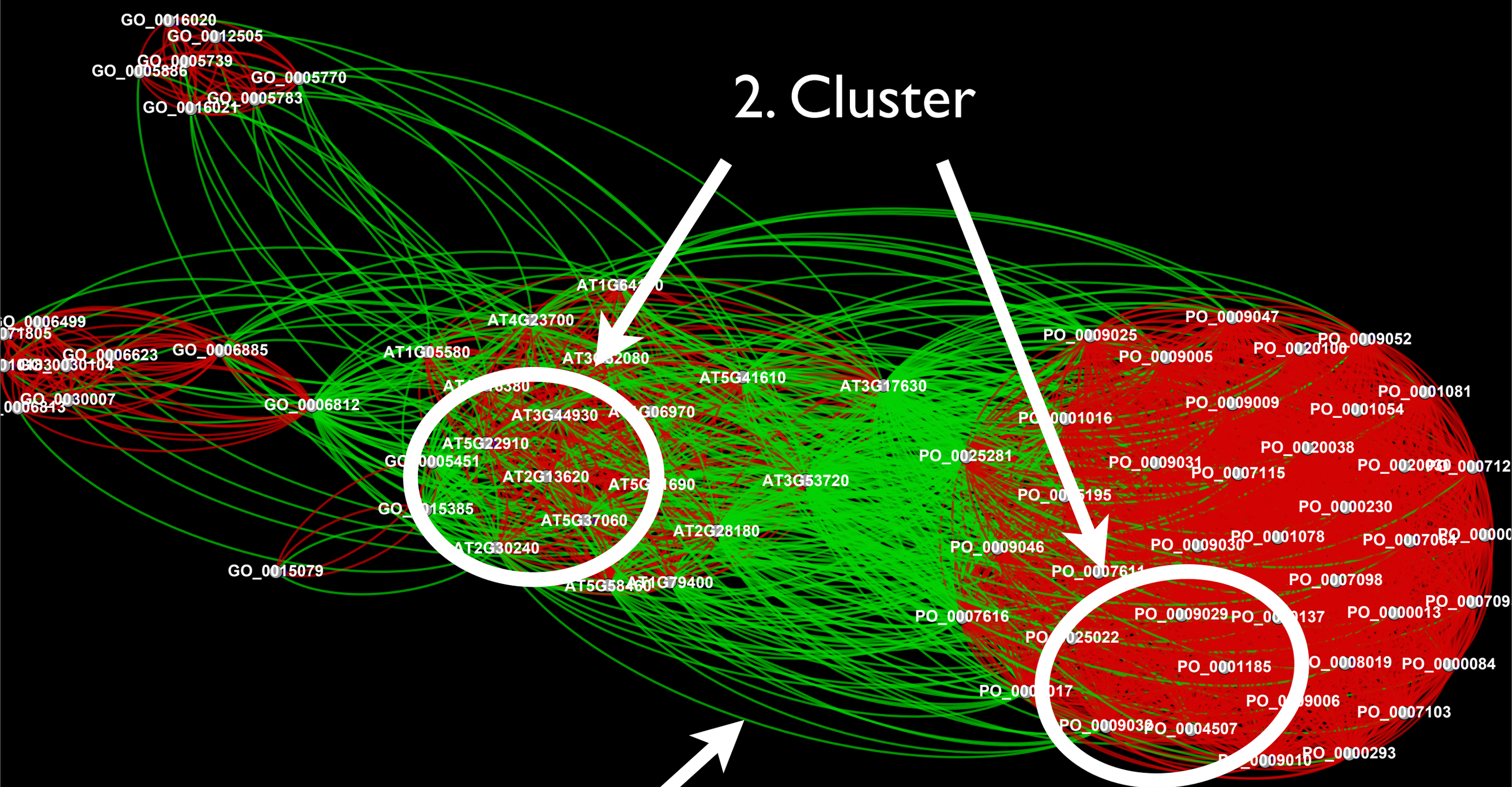
Bach, S., Broecheler, M., Getoor, L., O'Leary, D.. "Scaling MPE Inference for Constrained Continuous Markov Random Fields with Consensus Optimization" Advances in Neural Information Processing Systems (NIPS) - 2012

Evaluation

Input Graph


Gene Sequence Similarity
Annotation Links
GO Taxonomic Distance
PO Taxonomic Distance
GO_0016020
GO_0012505
GO_0005739
GO_0005886
GO_0005770
GO_0005783
GO_0016021
GO_0006499
GO_0006623
GO_0006885
GO_0030030104
GO_0030007
GO_0006812
GO_0005451
GO_0015385
GO_0015079
AT1G64170
AT4G23700
AT1G05580
AT3G52080
AT1G16380
AT5G41610
AT3G17630
AT3G44930
AT1G06970
AT5G22910
AT2G13620
AT5G01690
AT3G53720
AT5G37060
AT2G28180
AT2G30240
AT5G58460
AT1G79400
PO_0009047
PO_0009025
PO_0009052
PO_0009005
PO_0020100
PO_0001081
PO_0009009
PO_0001054
PO_0001016
PO_0020038
PO_0025281
PO_0009031
PO_0007115
PO_0025195
PO_0000230
PO_0009046
PO_0009030
PO_0001078
PO_0007611
PO_0007098
PO_0007616
PO_0009029
PO_0020137
PO_0000013
PO_0025022
PO_0001185
PO_0008019
PO_0000084
PO_0001017
PO_0009006
PO_0007103
PO_0009038
PO_0004507
PO_0009010
PO_0000293

Leave-one-out Link Prediction



2. Cluster

1. Leave one out

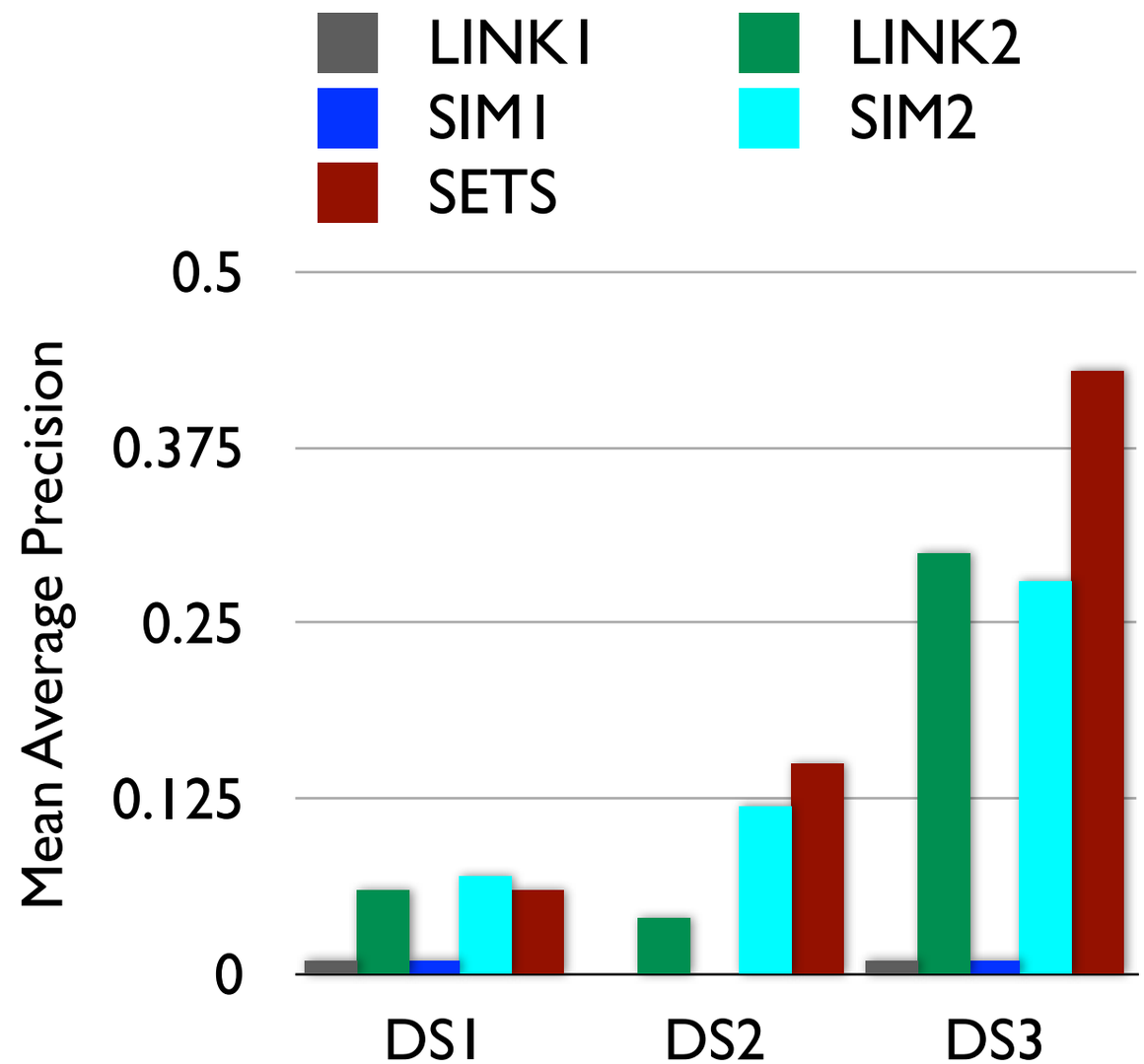
3. Predict missing link using clusters

Arabidopsis thaliana Data Sets

	DS1	DS2	DS3
Genes	10	10	18
PO Terms	53	48	40
GO Terms	44	31	19
PO-Gene	255	255	218
GO-Gene	157	157	92

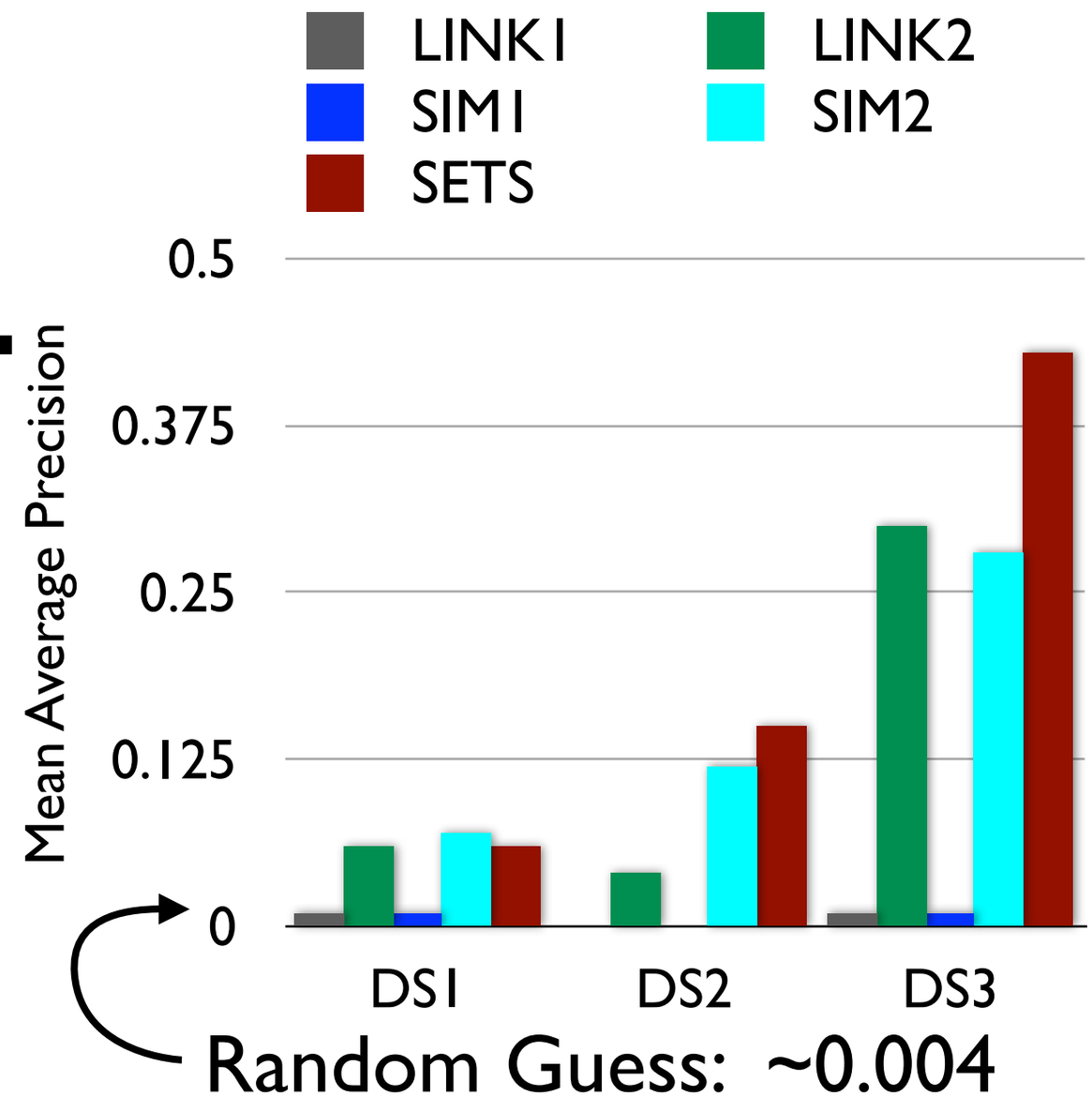
Results from Combining Heuristics

	Cluster	Link Prediction
LINK1	Similarity, Shared Neigh.	Similarity, Shared Neigh.
LINK2	Similarity, Neigh. ¬Shared	Similarity, Neigh. ¬Shared
SIM1	Similarity	Similarity, Shared Neigh.
SIM2	Similarity	Similarity, Neigh. ¬Shared
SETS	Similarity, Neigh. Sets	Similarity, Neigh. ¬Shared



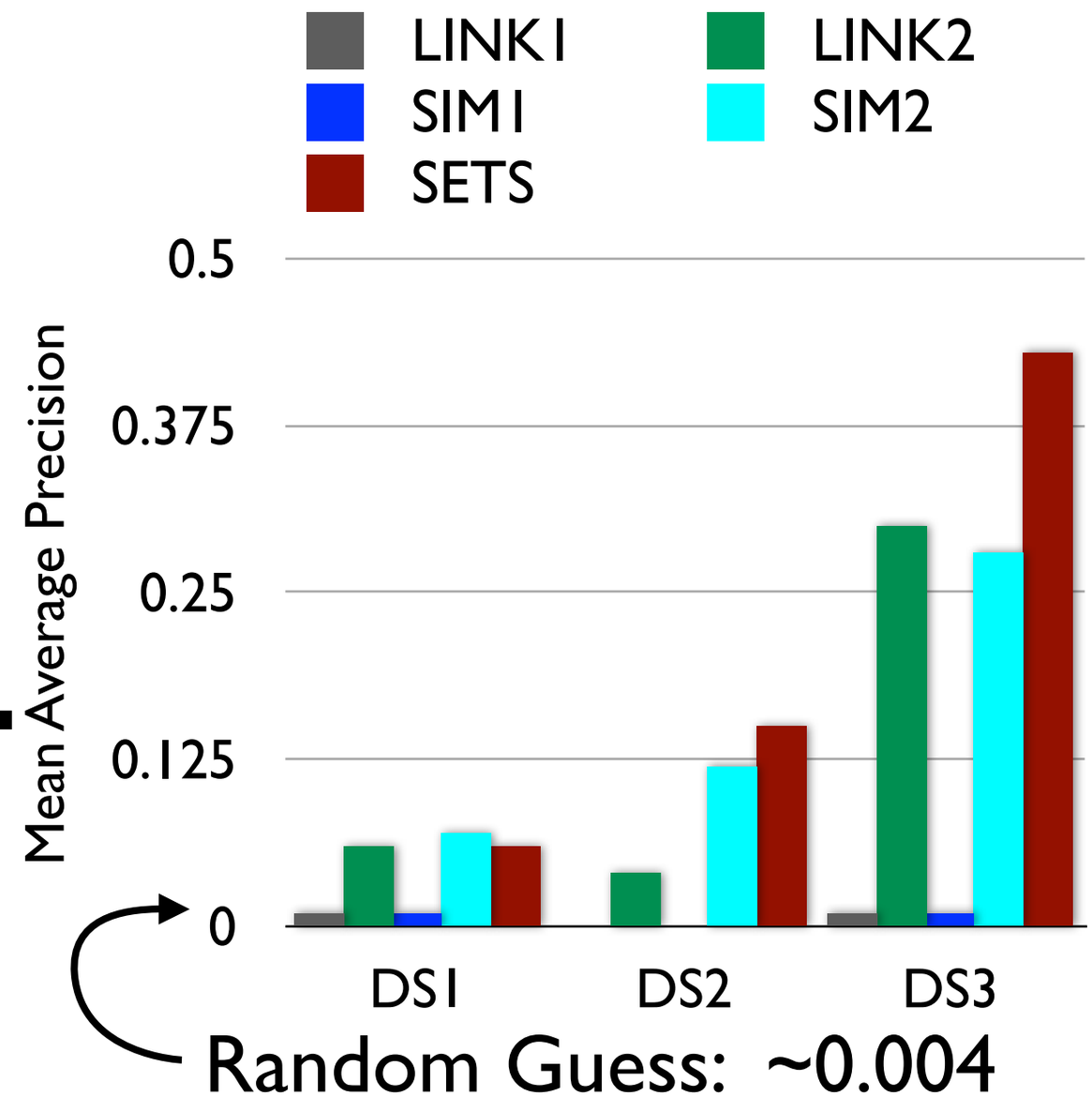
Results from Combining Heuristics

	Cluster	Link Prediction
LINK1	Similarity, Shared Neigh.	Similarity, Shared Neigh.
LINK2	Similarity, Neigh. ¬Shared	Similarity, Neigh. ¬Shared
SIM1	Similarity	Similarity, Shared Neigh.
SIM2	Similarity	Similarity, Neigh. ¬Shared
SETS	Similarity, Neigh. Sets	Similarity, Neigh. ¬Shared



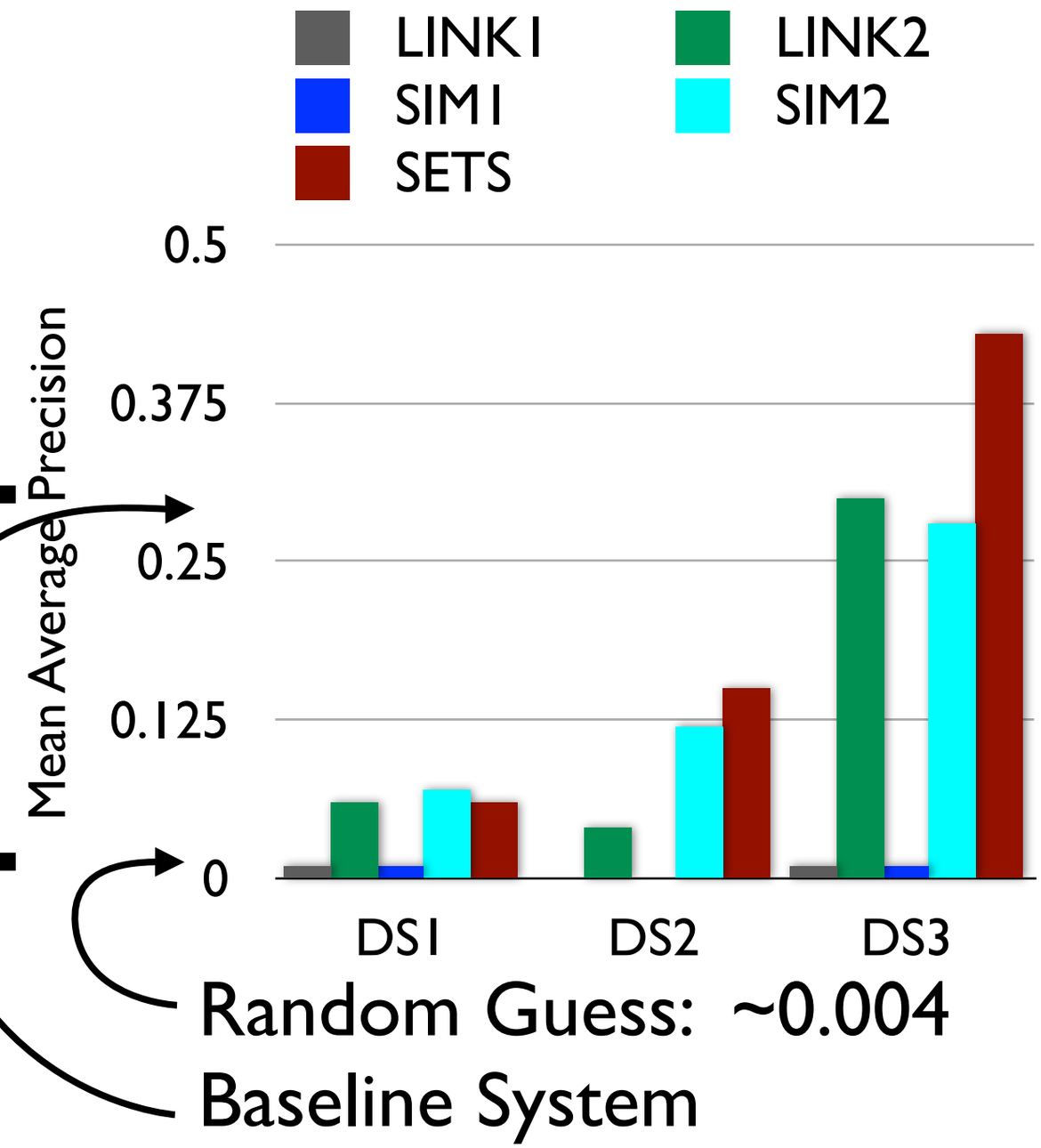
Results from Combining Heuristics

	Cluster	Link Prediction
LINK1	Similarity, Shared Neigh.	Similarity, Shared Neigh.
LINK2	Similarity, Neigh. ¬Shared	Similarity, Neigh. ¬Shared
SIM1	Similarity	Similarity, Shared Neigh.
SIM2	Similarity	Similarity, Neigh. ¬Shared
SETS	Similarity, Neigh. Sets	Similarity, Neigh. ¬Shared



Results from Combining Heuristics

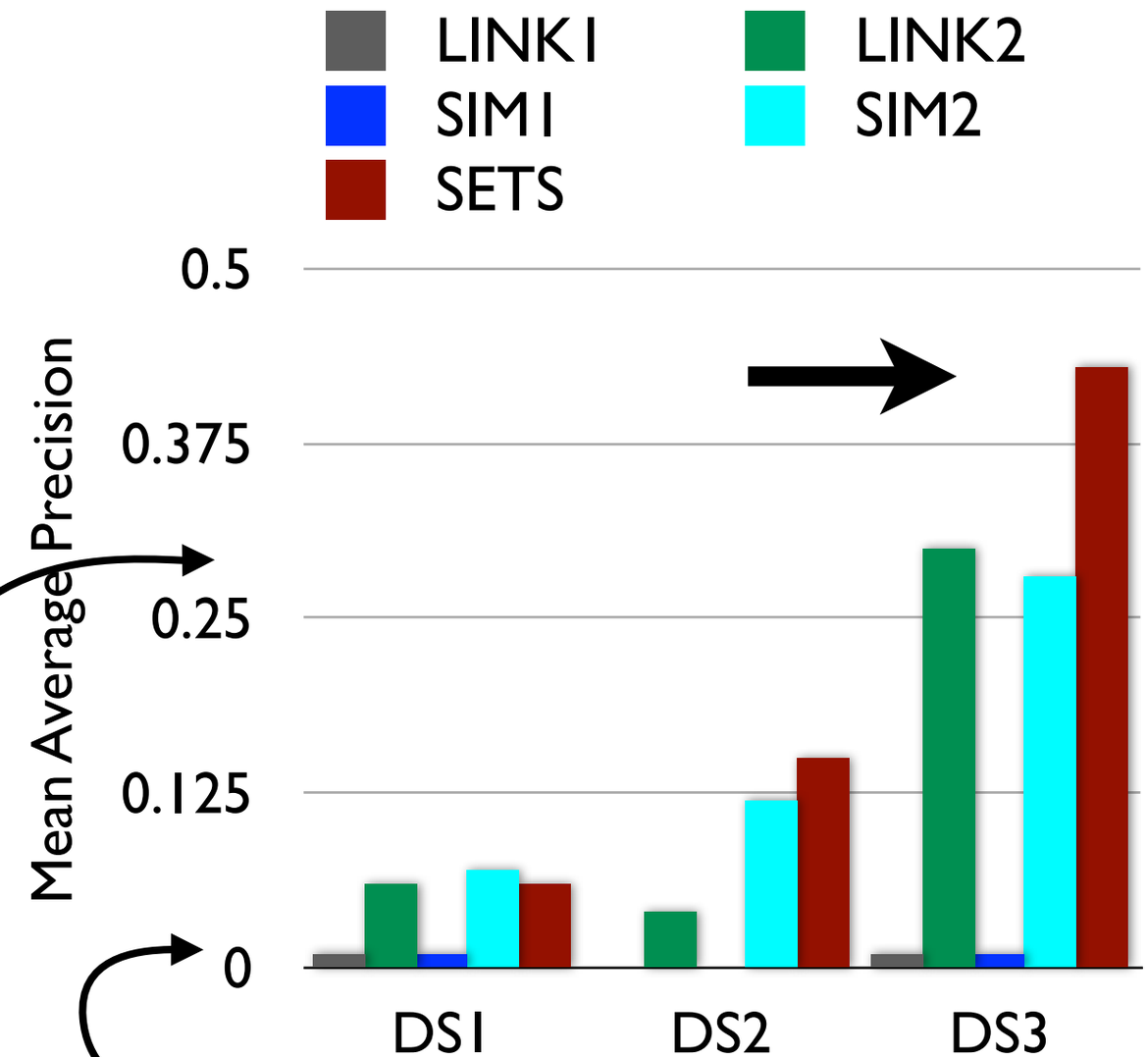
	Cluster	Link Prediction
LINK1	Similarity, Shared Neigh.	Similarity, Shared Neigh.
LINK2	Similarity, Neigh. ¬Shared	Similarity, Neigh. ¬Shared
SIM1	Similarity	Similarity, Shared Neigh.
SIM2	Similarity	Similarity, Neigh. ¬Shared
SETS	Similarity, Neigh. Sets	Similarity, Neigh. ¬Shared



Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., Zhang, X.N.: Link prediction for annotation graphs using graph summarization. In: International Semantic Web Conference (ISWC). (2011)

Results from Combining Heuristics

	Cluster	Link Prediction
LINK1	Similarity, Shared Neigh.	Similarity, Shared Neigh.
LINK2	Similarity, Neigh. ¬Shared	Similarity, Neigh. ¬Shared
SIM1	Similarity	Similarity, Shared Neigh.
SIM2	Similarity	Similarity, Neigh. ¬Shared
SETS	Similarity, Neigh. Sets	Similarity, Neigh. ¬Shared



Random Guess: ~0.004

Baseline System

Thor, A., Anderson, P., Raschid, L., Navlakha, S., Saha, B., Khuller, S., Zhang, X.N.: Link prediction for annotation graphs using graph summarization. In: International Semantic Web Conference (ISWC). (2011)

Conclusion

- Simple heuristics → competitive results
- Probabilistic Soft Logic for general reasoning over uncertainty

Thank you!

http://psl.umiacs.umd.edu

Backup

Distance to Satisfaction

$$d_r(I) = \max\{0, I(r_{body}) - I(r_{head})\}$$

using Lukasiewicz norms:

$$I(v_1 \wedge v_2) = \max\{0, I(v_1) + I(v_2) - 1\}$$
$$I(v_1 \vee v_2) = \min\{I(v_1) + I(v_2), 1\}$$
$$I(\neg l_1) = 1 - I(v_1)$$

```
1 : link(a,b) & exemplar(a,a) & exemplar(c,a) -> link(c,b)
        1            0.9            0.8              0
```

max{0,1+0.9+0.8−2}=0.7 d=0.7

```
        1            0.5            0.3              0
```

max{0,1+0.5+0.3−2}=0 d=0