informatics
luis rocha 2007

semimetric knowledge networks

applications to information retrieval and social data mining

luis m. rocha
Indiana university
school of informatics
1900 East Tenth Street, Bloomington IN 47406
and
Instituto Gulbenkian de Ciência
Apartado 14, 2781-901 Oeiras, Portugal

rocha@indiana.edu
http://informatics.indiana.edu/rocha

FLAD
Computational Biology Collaboratorium

Cognitive Science Program

INDIANA UNIVERSITY

informatics
luis rocha 2007

agent-based model of genotype editing

CASCI Team: http://casci.informatics.indiana.edu

CASCI2RD

Alaa Abi-Haidar

jasleen kaur

Tiago Simas

Ana Maguitman

acknowledgements

Luis Bettencourt, Bharat Dravid, Mariella DiGiacomo, Fil Menczer, Pedrag Ravidojac, Andreas Retchsteiner, Elliot Smith, Karin Verspoor, Paul Wang.

rocha@indiana.edu
http://informatics.indiana.edu/rocha/editing

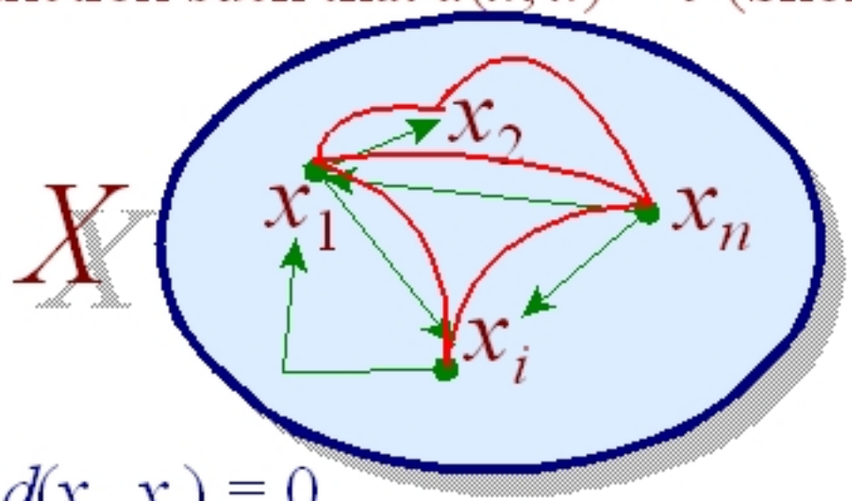
Cognitive Science Program

INDIANA UNIVERSITY

distances on graphs

measured from associative "knowledge" graphs

d is a distance function on set X if it is a nonnegative, symmetric, real-valued function such that $d(x, x) = 0$ (Shore & Sawyer 1993)



$d(x_i, x_i) = 0$

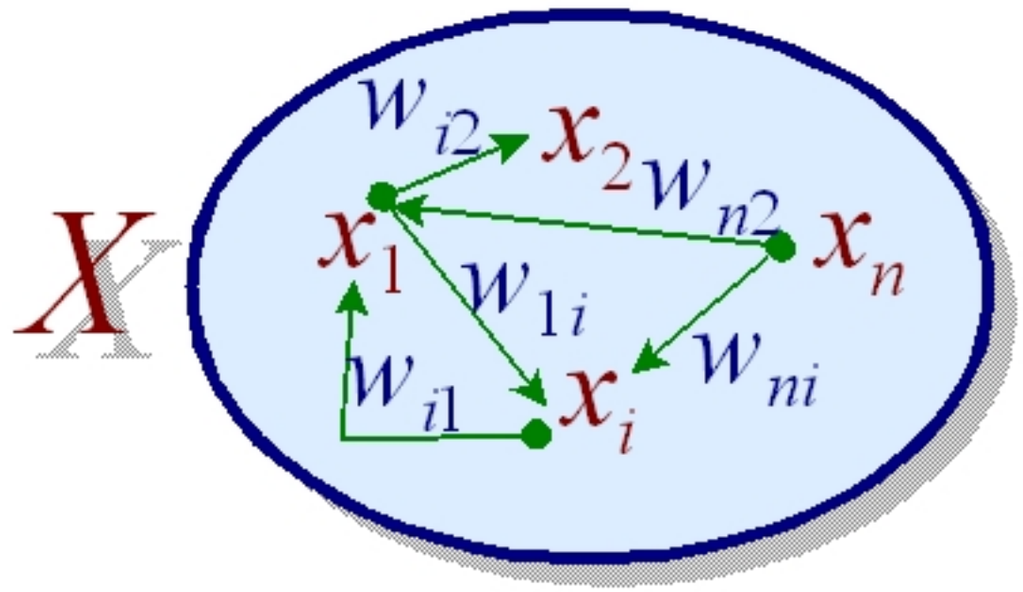
$d(x_i, x_j) = 1$, if there is an edge

$d(x_i, x_k) = d(x_i, x_j) + ... + d(x_l, x_k)\ 1$, if there is a path

Due to the symmetry requirement, distance functions yield non-directed distance graphs

$$d(x_1, x_2) \leq d(x_1, x_3) + d(x_3, x_2)$$

Metric: the smallest distance between nodes is always the most direct path



In real-valued weighted graphs, derived distance functions can be semi-metric

$$d(k_1, k_2) > d(k_1, k_3) + d(k_3, k_2)$$

Semi-metric

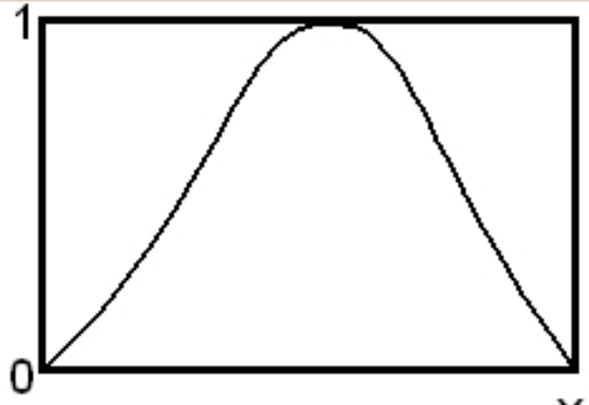
In graphs used to store "knowledge", what does it mean?

rocha@indiana.edu
http://informatics.indiana.edu/rocha

operations

informatics
luis rocha 2007

$$A(x) : X \rightarrow [0, 1]$$



Fuzziness
Degree of Membership/Truth

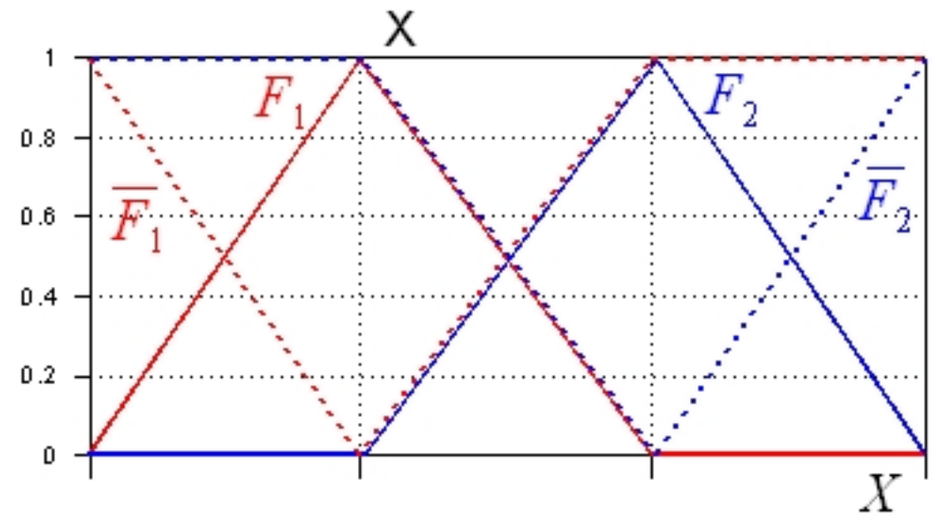
Standard Fuzzy Operations

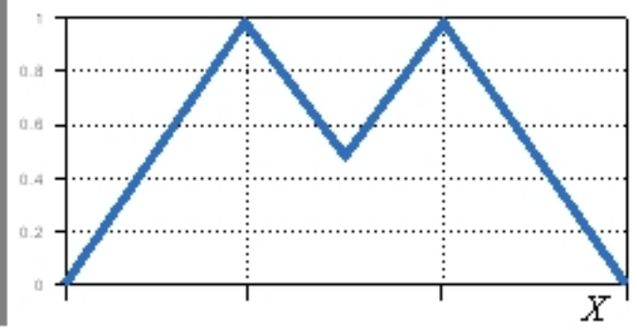
$$\overline{A}(x) = 1 - A(x)$$

$$(A \cap B)(x) = \min[A(x), B(x)]$$

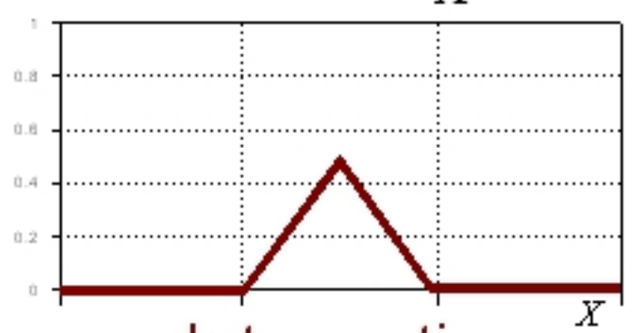
$$(A \cup B)(x) = \max[A(x), B(x)]$$

- Follows
  - Involution, commutativity, associativity, distributivity, Identity, De Morgan's Laws, etc
- Does not Follow
  - Laws of contradiction and excluded middle





Union



Intersection

$$A \cap \overline{A} \neq \emptyset$$

$$A \cup \overline{A} \neq X$$

rocha@indiana.edu
http://informatics.indiana.edu/rocha

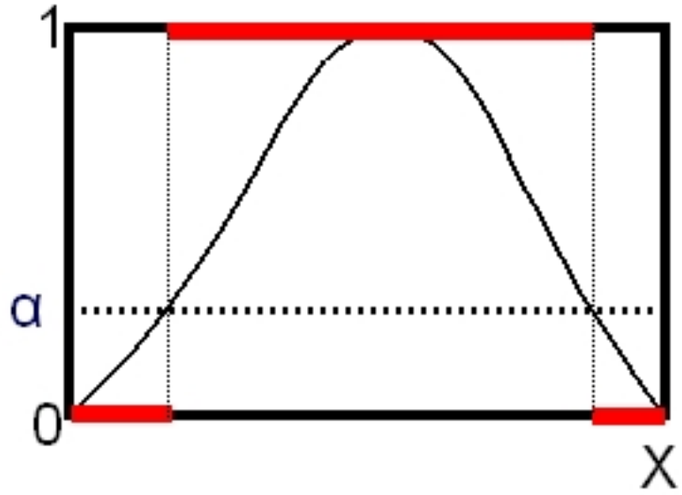
INDIANA UNIVERSITY

α-cut

informatics
luis rocha 2007

Fuzzy Sets:

$$A(x) : X \rightarrow [0, 1]$$



α-cut: Crisp set at threshold α

rocha@indiana.edu
http://informatics.indiana.edu/rocha

dual fuzzy set operations

De Morgan's Laws

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$

$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

$$i(a,b) = ab$$

$$i(a,b) = \frac{ab}{a + b - ab}$$

$$u(a,b) = a + b - ab$$

$$u(a,b) = \frac{a + b - 2ab}{1 - ab}$$

- Complement, Intersection and Union must follow De Morgan's Laws plus:
  - Complement
    - Boundary Conditions: $c(0)=1$ and $c(1)=0$
    - Monotonicity: if $a \le b$ then $c(a) > c(b)$
    - Continuity
    - **Involutive: $c(c(a)) = a$**
  - Intersection (T-Norm)
    - Boundary condition: $i(a, 1)= a$
    - Monotonicity: if $b \le d$ then $i(a,b) \le i(a,d)$
    - Comutativity: $i(a,b) = i(b,a)$
    - Associativity: $i(a,i(b,d))=i(i(a,b),d)$
    - Continuity
    - Strict Monotonicity: if $a1<a2$ and $b1<b2$ then $i(a1,b1)<i(a2,b2)$
    - Subidempotency: $i(a,a) \le a$
  - Union (T-Conorm)
    - Boundary condition: $u(a, 0)= a$
    - Monotonicity: if $b \le d$ then $u(a,b) \le u(a,d)$
    - Comutativity: $u(a,b) = u(b,a)$
    - Associativity: $u(a,u(b,d))=u(u(a,b),d)$
    - Continuity
    - Strict Monotonicity: if $a1<a2$ and $b1<b2$ then $u(a1,b1)<u(a2,b2)$
    - Superidempotency: $u(a,a) > a$

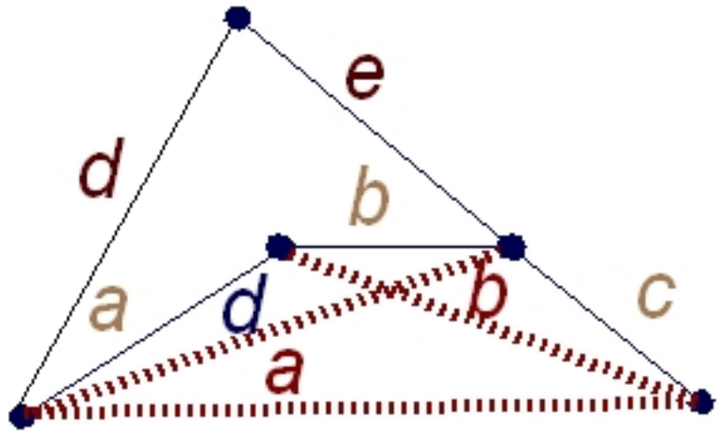
$$c_\lambda(a) = \frac{1 - a}{1 + \lambda a}$$

Sugeno Complement: $\lambda \in (-1, \infty)$

rocha@indiana.edu
http://informatics.indiana.edu/rocha

properties

- Reflexive
  - iff $R(x, x) = 1$ for all $x \in X$
    - every element of X is maximally associated with itself
- Symmetric
  - iff $R(x, y) = R(y, x)$ for all $x, y \in X$
    - Matrices require only $(n^2\text{-}n)/2$ elements to be defined
- (Max-Min) Transitive
  - iff $R(x, z) \geq \max_{y \in X} \min[R(x, y), R(y, z)]$ for all $x, z \in X$
    - For each indirect connection between x and z through some y, the weight of the connection is the smallest of each connection (x to y and y to z). Finally, the weight of the connection between x and z, is the largest of all indirect connections through all y (strongest path defined by *weakest link*)



Max-Min Transitivity
$a<b<c$
$a<d<e$

rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2007

Max-Min Composition:

$$R \circ R = \max_k \min(r_{ik}, r_{kj}) = r'_{ij}$$

where r_{ij} denotes $R(x_i, x_j)$

The max-min composition of matrices is performed in the same way as the numerical counterpart, except that *multiplication* and *summation* are substituted by the ∩ (e.g. *Min*) and U (e.g. *Max*) operations respectively.

- Transitive closure of a relation $R(X, X)$
  - The relation that is transitive, contains $R(X, X)$, and whose elements have the smallest possible membership weights that still allow the first two requirements.
    - It yields a relation where all pairs of elements which were directly or indirectly related in the original relation, are now directly related

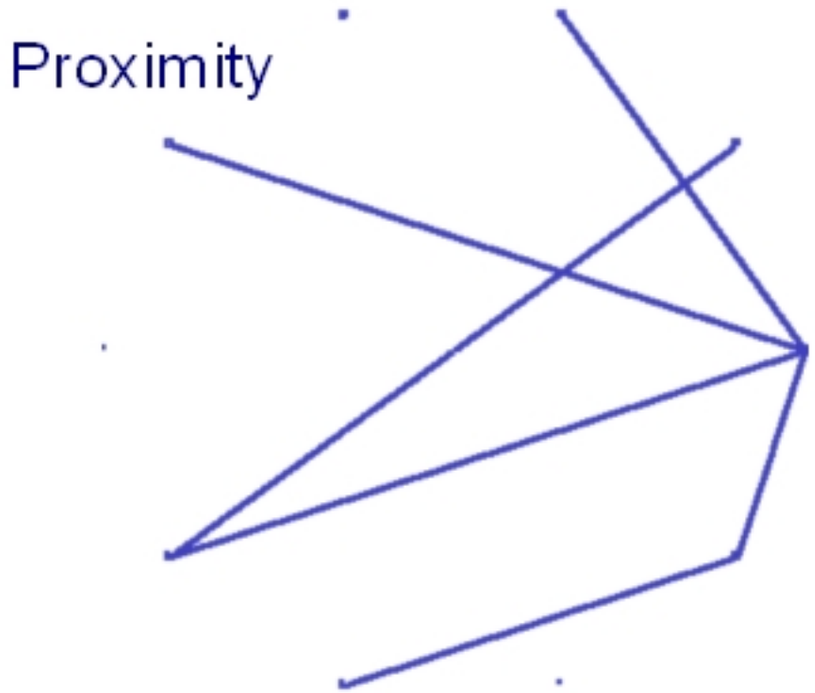
Generic Composition:

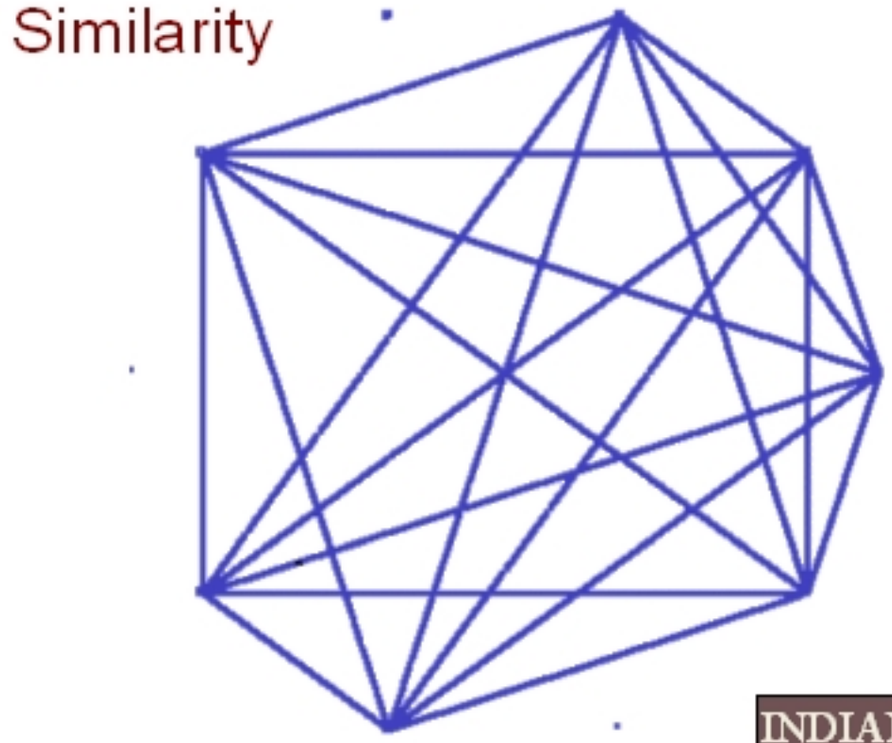
$$R \circ R = \bigcup_k \bigcap (r_{ik}, r_{kj}) = r'_{ij}$$

rocha@indiana.edu
http://informatics.indiana.edu/rocha

similarity and proximity relations

- Similarity Relation
  - A reflexive, symmetric, and transitive binary fuzzy relation
    - Also known as an equivalence relation.
- Proximity Relation
  - A reflexive and symmetric binary fuzzy relation
    - Also known as a compatibility relation
    - The transitive closure of a proximity relation is a similarity relation.





rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2007

extraction of co-occurrence networks

from document relations

- Document × Keyterms
  - Keyterm Co-Occurrence
- Document × Document
  - Co-Citation or Hyperlink structure
- Document × Author
  - Co-Authorship (Collaboration Network)
- Bio-entities × Keyterms
  - Gene/MeSH keyterm Co-Occurrence

	X (Keywords)
Y (Documents)	$R:X\times Y$

Given a binary relation R between sets X and Y we extract two proximity relations: $XYP(x_i, x_j)$ is the probability that both x_i and x_j are related in R to the same element $y \in Y$. Conversely, $YXP(y_i, y_j)$ is the probability that both y_i *and* y_j are related in R to the same element $x \in X$.

$$XYP(x_i, x_j) = \frac{\sum_{k=1}^{m} (r_{i,k} \wedge r_{j,k})}{\sum_{k=1}^{m} (r_{i,k} \vee r_{j,k})}; \quad YXP(y_i, y_j) = \frac{\sum_{k=1}^{n} (r_{k,i} \wedge r_{k,j})}{\sum_{k=1}^{n} (r_{k,i} \vee r_{k,j})}$$

With some support constraint

rocha@indiana.edu
http://informatics.indiana.edu/rocha

Rocha, 1999. *IJGS.* **27**, 457.
Rocha and Bollen, 2001. In: *Design Principles for the Immune System and other Distributed Autonomous Systems.* Segel and Cohen (Eds), 305.

INDIANA UNIVERSITY

produce associative (probabilistic) networks

informatics
luis rocha 2007

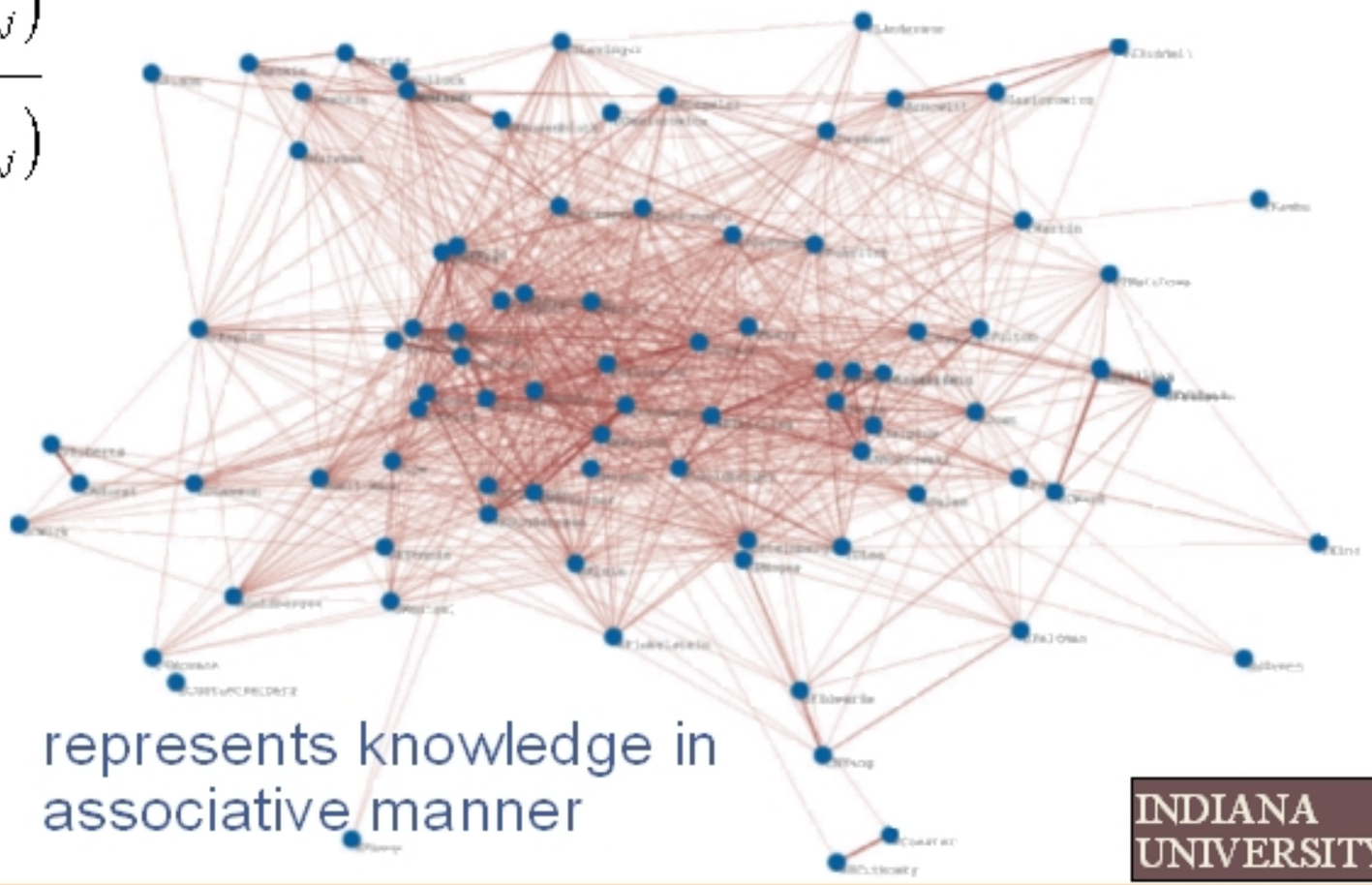
	X (Keywords)
Y (Documents)	$R:X\times Y$

$$XYP(x_i,x_j)=\frac{\sum_{k=1}^{m}\left(r_{i,k}\wedge r_{j,k}\right)}{\sum_{k=1}^{m}\left(r_{i,k}\vee r_{j,k}\right)}$$

	X (Keywords)
X (Keywords)	$XYP:X\times X$

$$YXP(y_i,y_j)=\frac{\sum_{k=1}^{n}\left(r_{k,i}\wedge r_{k,j}\right)}{\sum_{k=1}^{n}\left(r_{k,i}\vee r_{k,j}\right)}$$

	Y (Documents)
Y (Documents)	$YXP:Y\times Y$



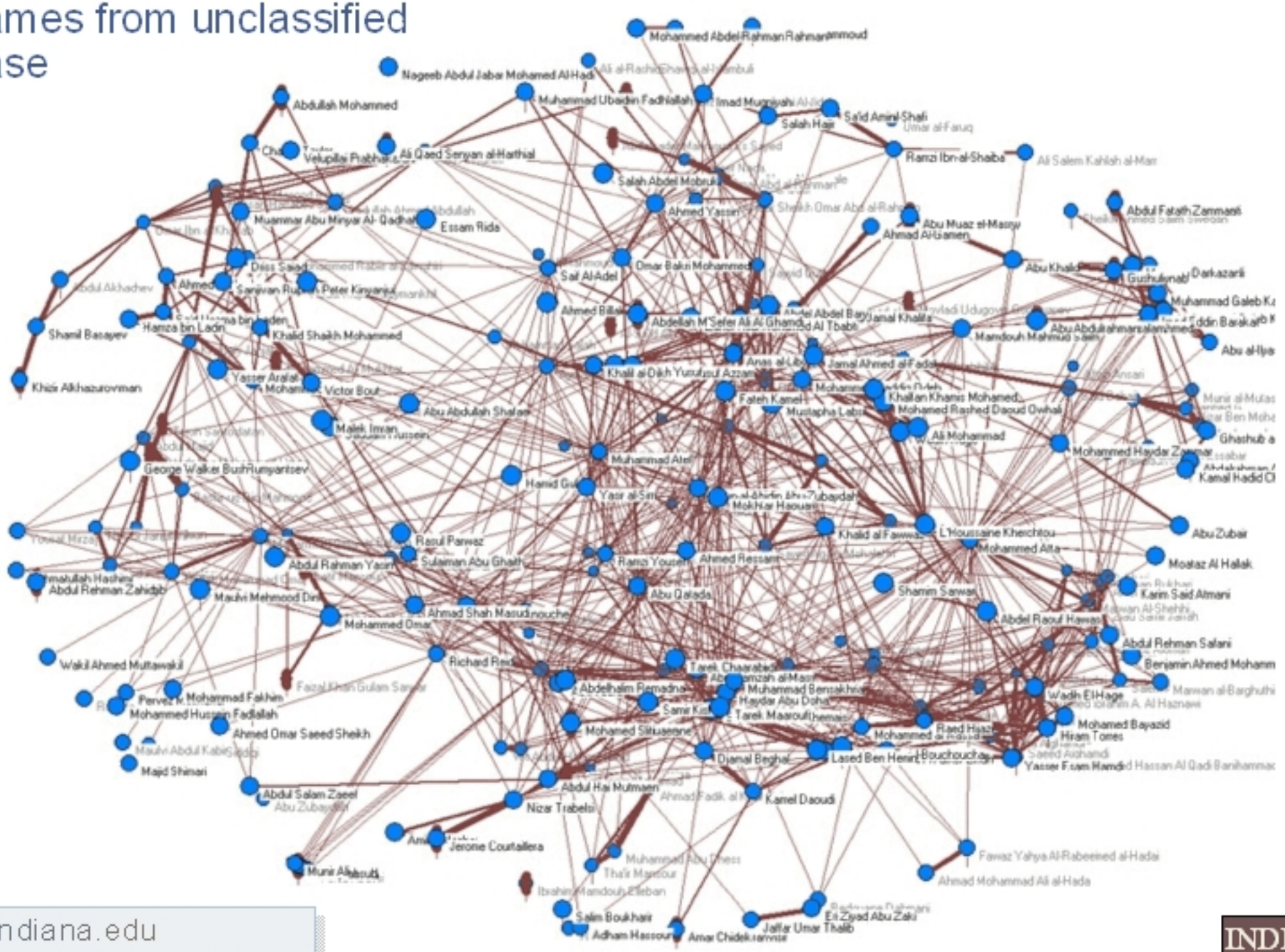
represents knowledge in associative manner

rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2007

PDP2

318 names from unclassified database



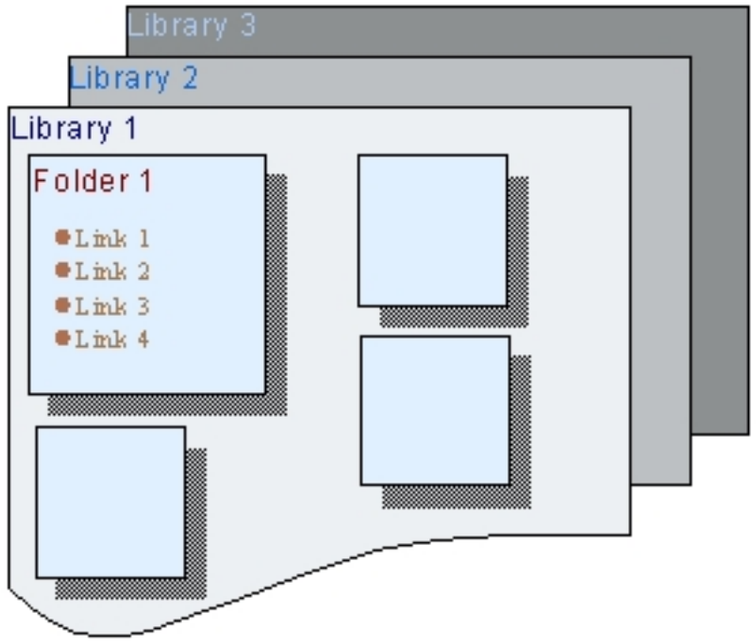
rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

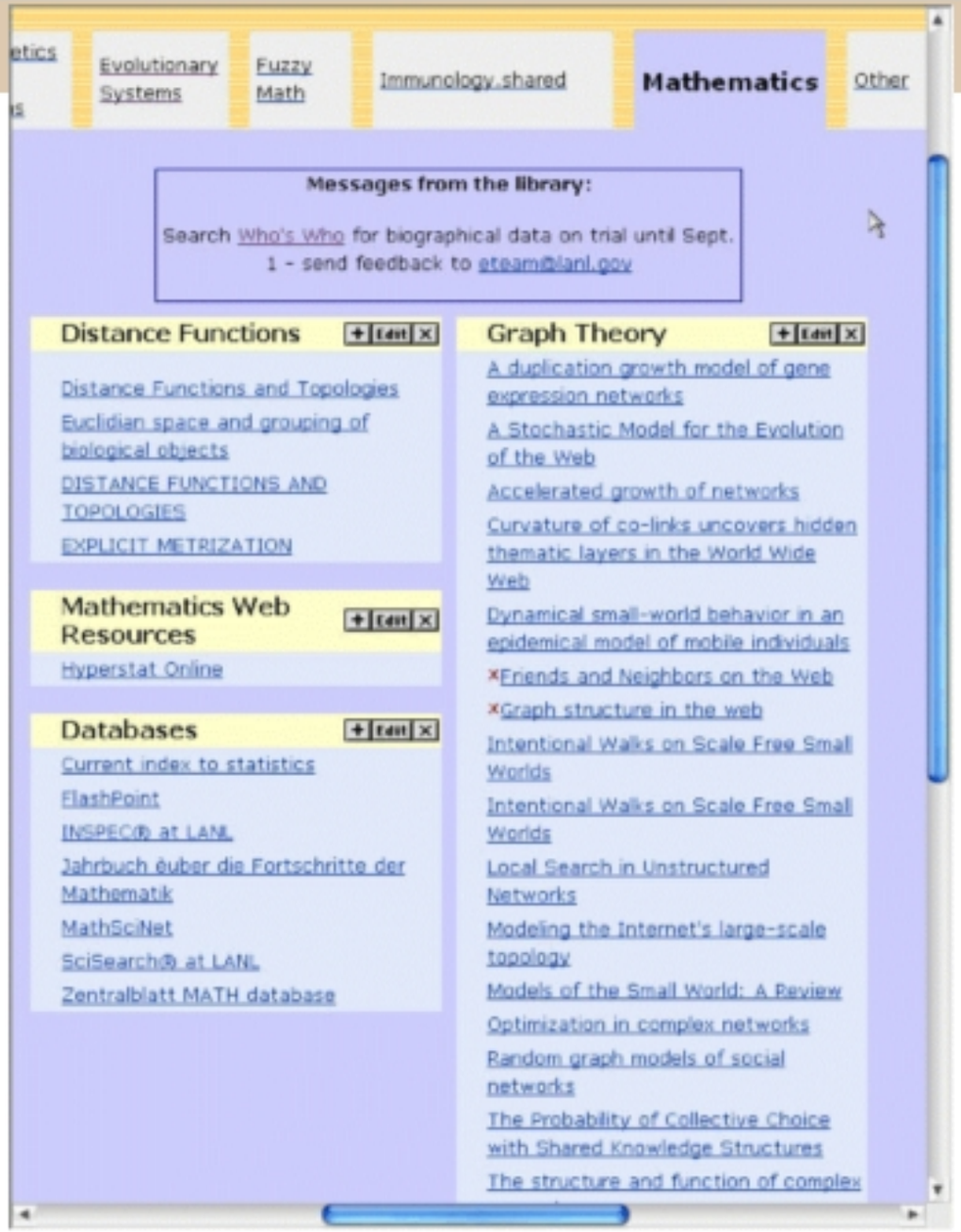
informatics
luis rocha 2007

mylibrary.lanl.gov

digital library web service



- Three nested entities:
  - Libraries ⇨ Folders ⇨ Links
    - A *library/personality* is associated with a given area of interest and consists of one or more folders.
    - A *folder* contains related types of links within a library
    - A *link* is a URL (typically scientific articles)



rocha@indiana.edu
http://informatics.indiana.edu/rocha

Rocha, L.M., T. Simas, A. Rechtsteiner, M. DiGiacomo, R. Luce [2005]. "MyLibrary@LANL: Proximity and Semi-metric Networks for a Collaborative and Recommender Web Service". In: *Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, IEEE Press, pp. 565-571.

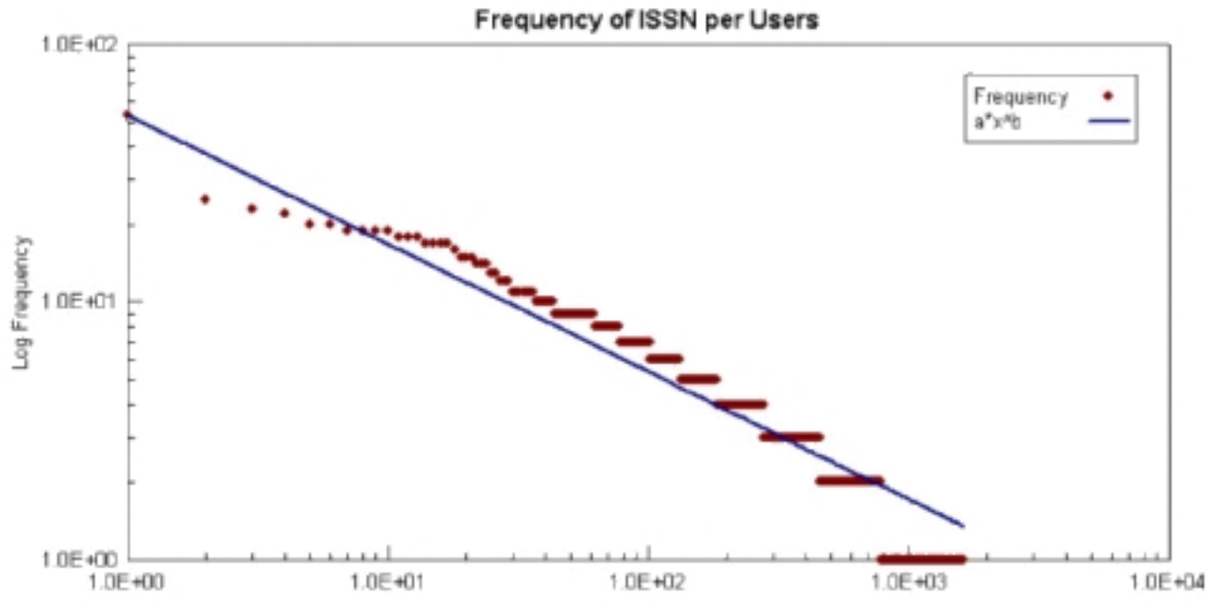
INDIANA UNIVERSITY

informatics
luis rocha 2007

most frequent ISSN

occurences in user personalities

Physical review letters	53
Physical Review B	26
Physical review E	23
Physical review A General physics	22
Journal of physical chemistry B	20
Computers & geosciences	20
Scientific American	19
Journal of the American Chemical Society	19
Journal of Chemical Physics	19
Reviews of modern physics	19
Bioinformatics	18
IEEE trans. on geoscience and remote sensing	18
PNAS	18
Journal of computational physics	18
Advances in water resources	18
Journal of applied geophysics	17
Applied geochemistry	17
APL	17
Journal of physical chemistry A	17
Phil. mag. B Physics of condensed matter	17
Bul.of Environmental Contamination and Toxicology	16
Journal of applied physics	15
American journal of physics	14
Analytical chemistry	14
DLib	14
Chemical physics	13
NIM	13
Chemical physics letters	12
Physics reports	12
Physical review A	12



Nature	11
Physics of plasmas	11
Accounts of chemical research	11
Advanced Materials	11
Journal of physics Condensed matter	11
Journal of computational biology	11
Biochemical journal	11
Computer physics communications	11
ACM transactions on modeling and computer simulation	10
Materials science & engineering A	10
Physical review C Nuclear physics	10
Advances in physics	10
Inorganic chemistry	10
Trends in BioTechnology	10
Review of Scientific Instruments	10
Science	10

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

informatics
luis rocha 2007

ISSN and personality proximity

from co-occurrence in mylibrary.lanl.gov

	ISSN
Personality	$A:P\times I$

392 personalities with at least two ISSN
253 users with at least two ISSN
1702 unique ISSN occuring at least twice

Given a binary relation A between sets of Personalities P and ISSN I we extract two proximity relations: $PIP(p_s, p_t)$ is the probability that both personalities p_s and p_t link to the same ISSN $i \in I$. Conversely, $IPP(i_s, i_t)$ is the probability that both ISSN i_s and i_t co-occur in the same personality (given that one of them occurs) $p \in P$.

$$pip(p_s, p_t) = \frac{\sum_{k=1}^{m}(a_{i,k} \wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k} \vee a_{j,k})} = \frac{N_{\cap}(p_s, p_t)}{N_{\cup}(p_s, p_t)}$$

(Personality ISSN Proximity)

$$ipp(i_s, i_t) = \frac{\sum_{k=1}^{m}(a_{i,k} \wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k} \vee a_{j,k})} = \frac{N_{\cap}(i_s, i_t)}{N_{\cup}(i_s, i_t)}$$

(ISSN Personality Proximity)

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

informatics
luis rocha 2007

journal network

from co-occurrence in user personalities in mylibrary.lanl.gov: IPP

$$ipp(i_s, i_t) = \frac{\sum_{k=1}^{m}(a_{i,k} \wedge a_{j,k})}{\sum_{k=1}^{m}(a_{i,k} \vee a_{j,k})} = \frac{N_{\cap}(i_s, i_t)}{N_{\cup}(i_s, i_t)}$$

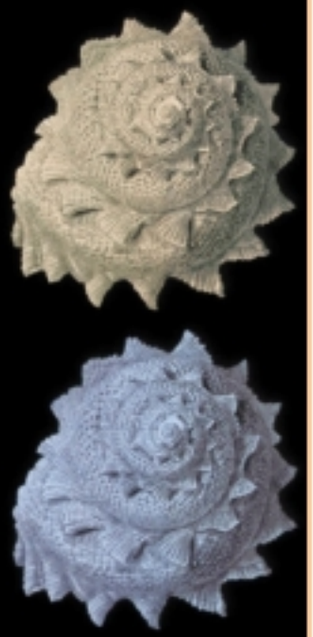
(ISSN Personality Proximity)

All Weights

rocha@indiana.edu
http://informatics.indiana.edu/rocha

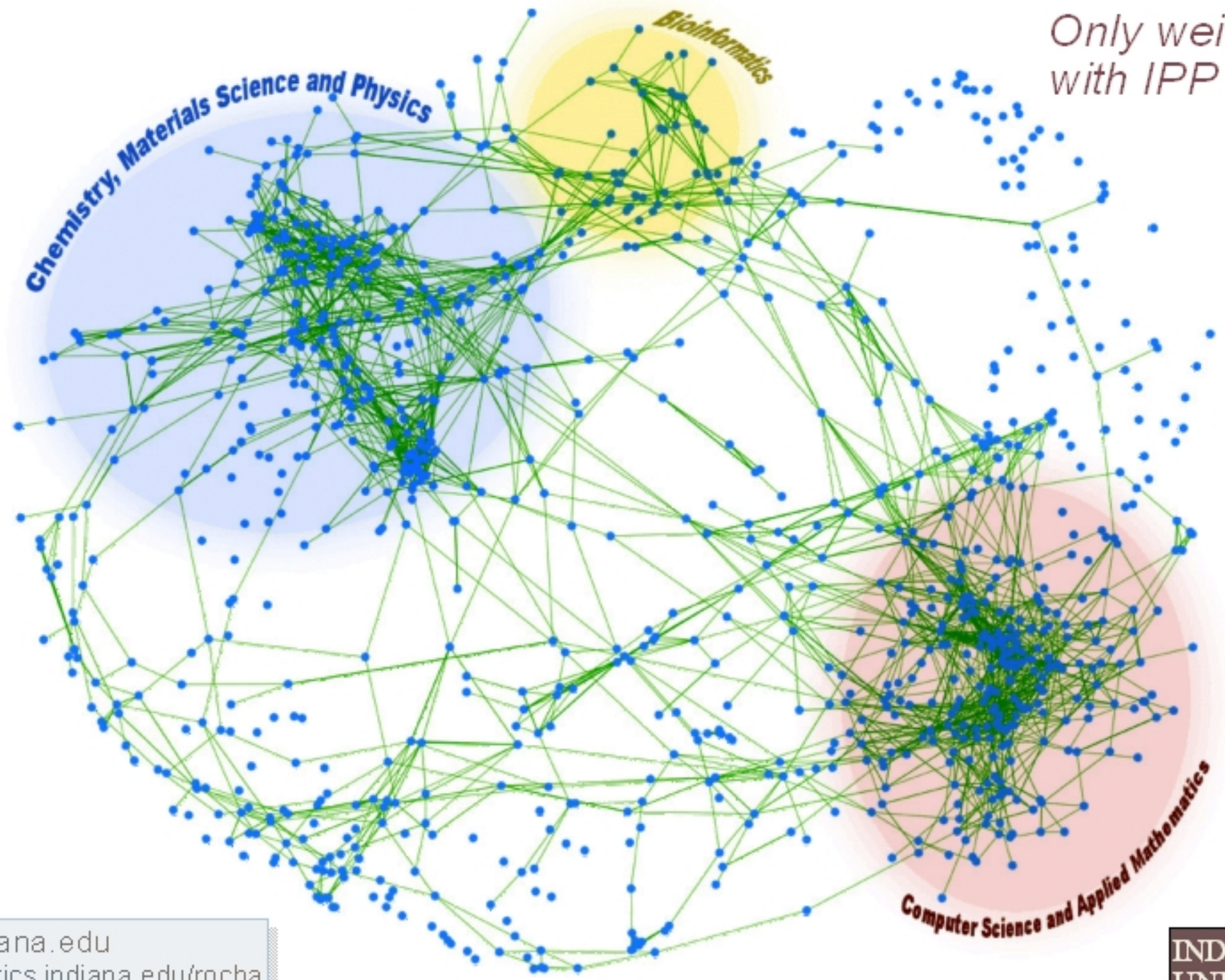
INDIANA UNIVERSITY

informatics
luis rocha 2007



IPP (journal) network

from co-occurrence in user personalities

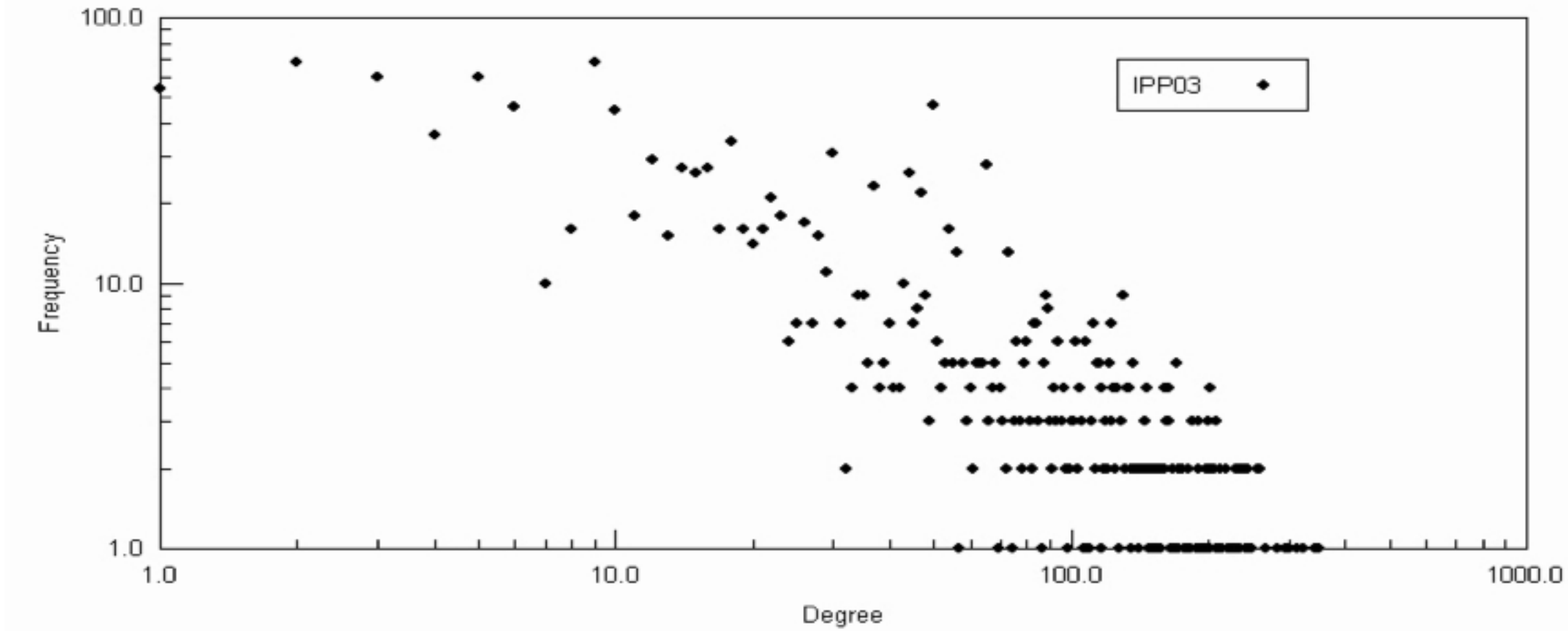


Only weights with IPP ≥ 0.3

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

IPP: all weights



informatics
luis rocha 2007

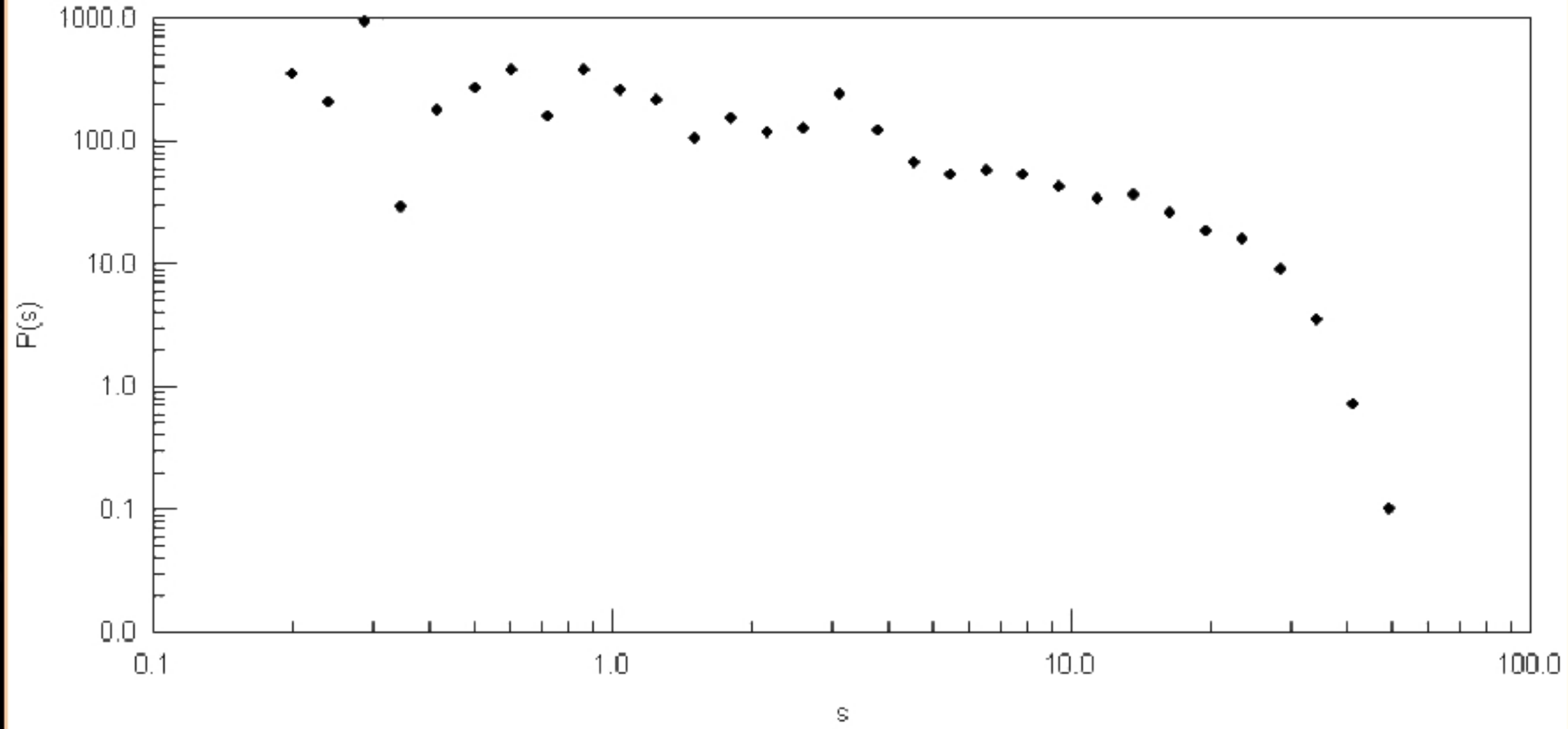
rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

informatics
luis rocha 2007

cumulative strength distribution (binned)

IPP: all weights



rocha@indiana.edu
http://informatics.indiana.edu/rocha

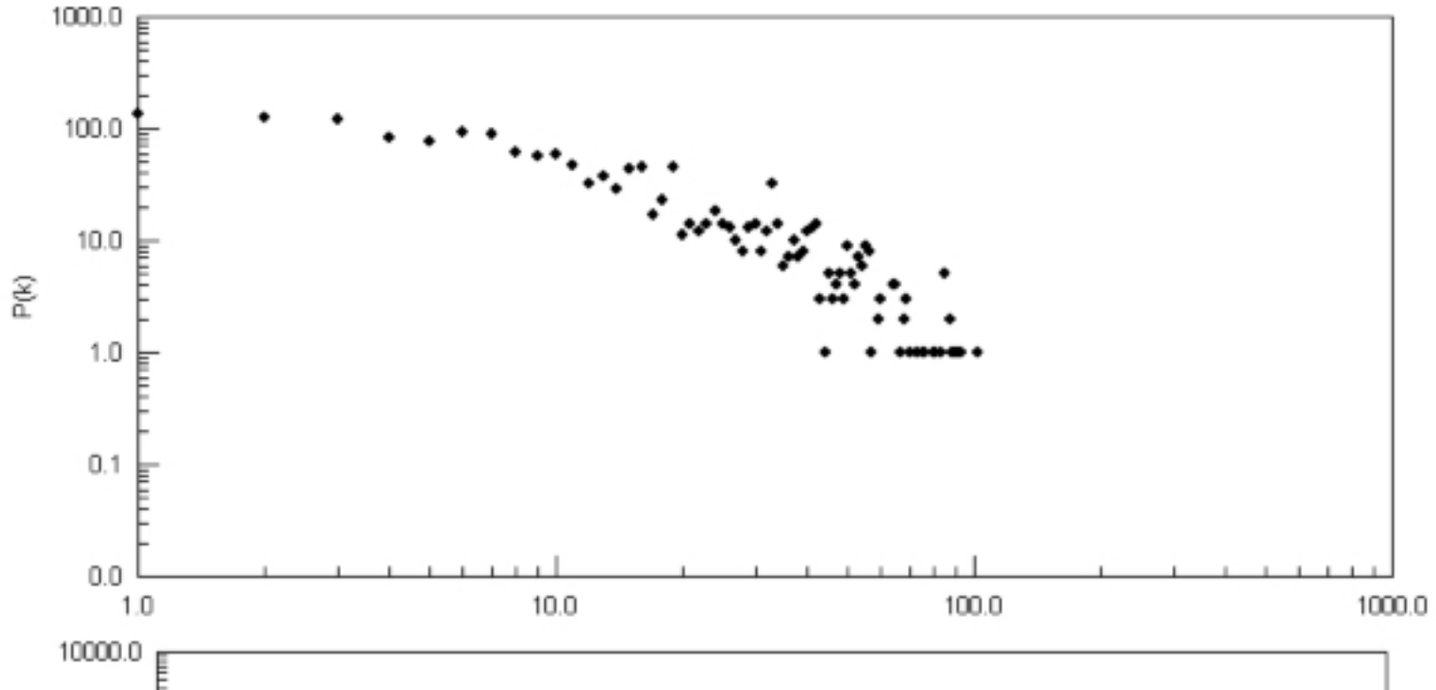
INDIANA
UNIVERSITY

informatics
luis rocha 2007

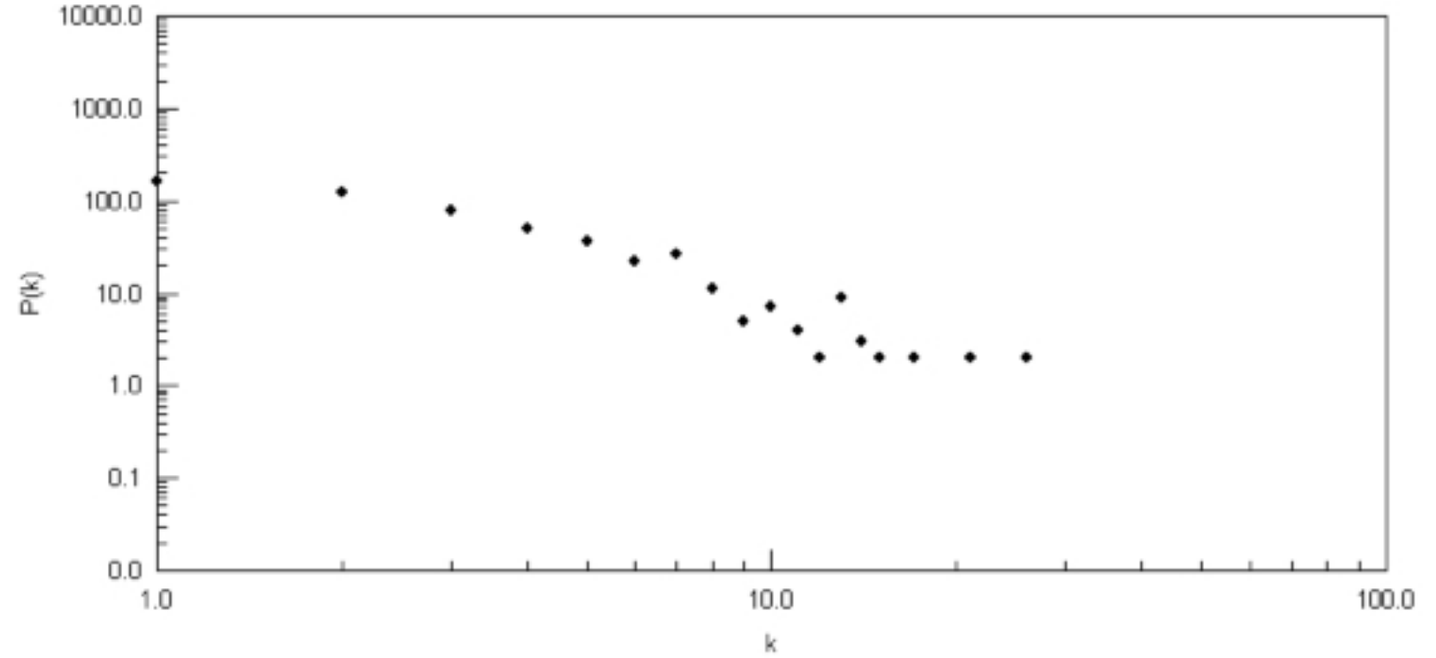
cumulative degree distribution

IPP

α-cut = 0.2



α-cut = 0.4



rocha@indiana.edu
http://informatics.indiana.edu/rocha

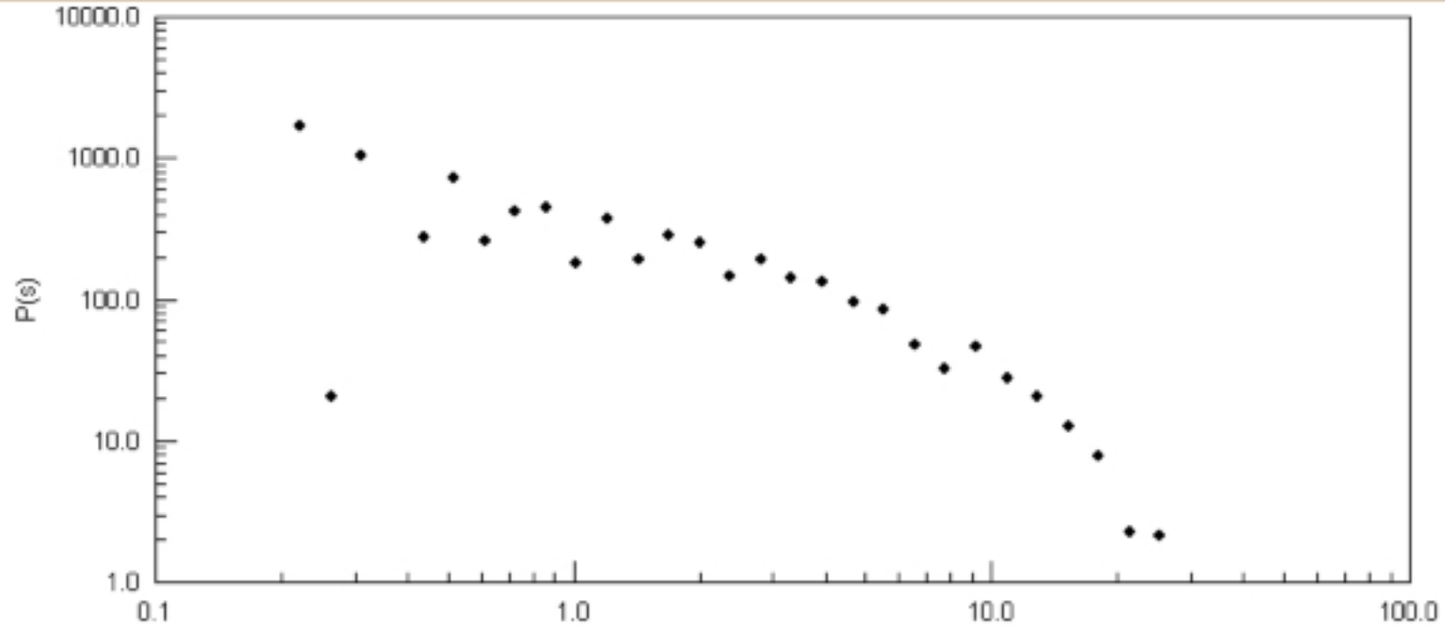
INDIANA
UNIVERSITY

informatics
luis rocha 2007

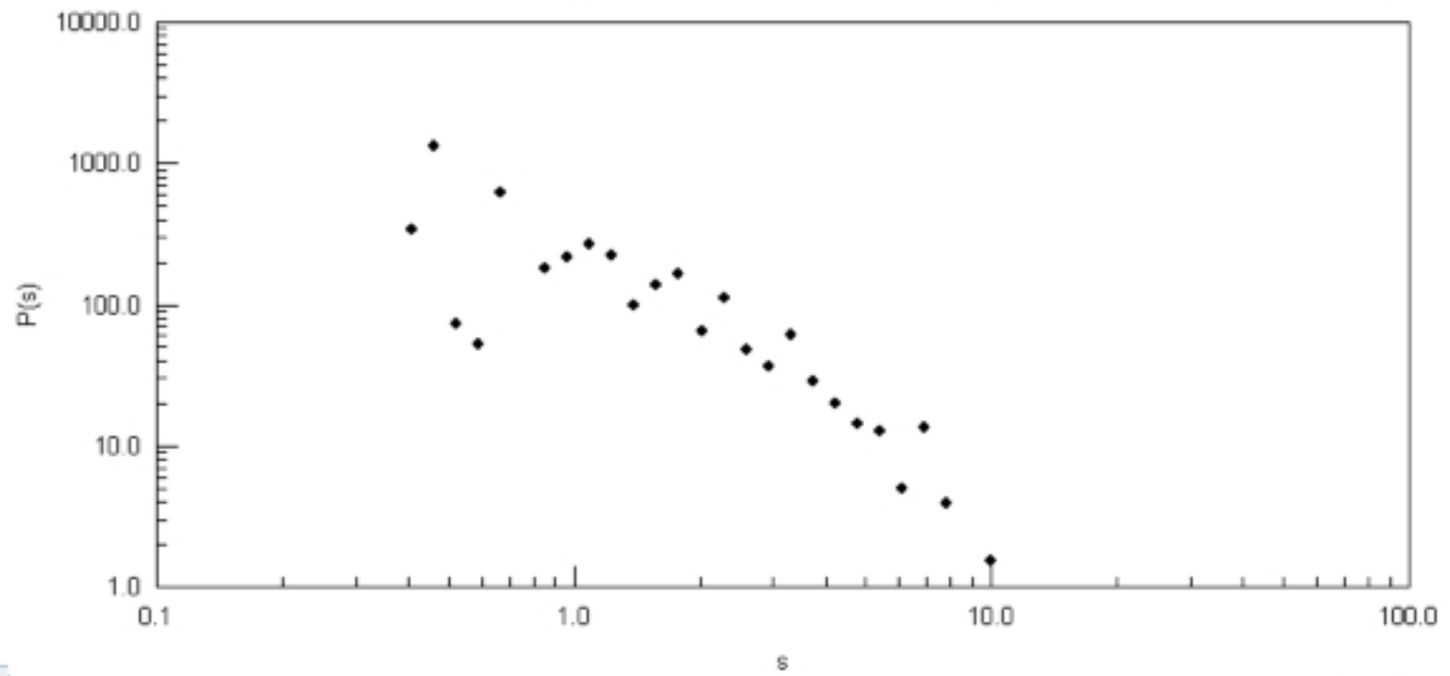
cumulative strength distribution (binned)

IPP

α-cut = 0.2



α-cut = 0.4

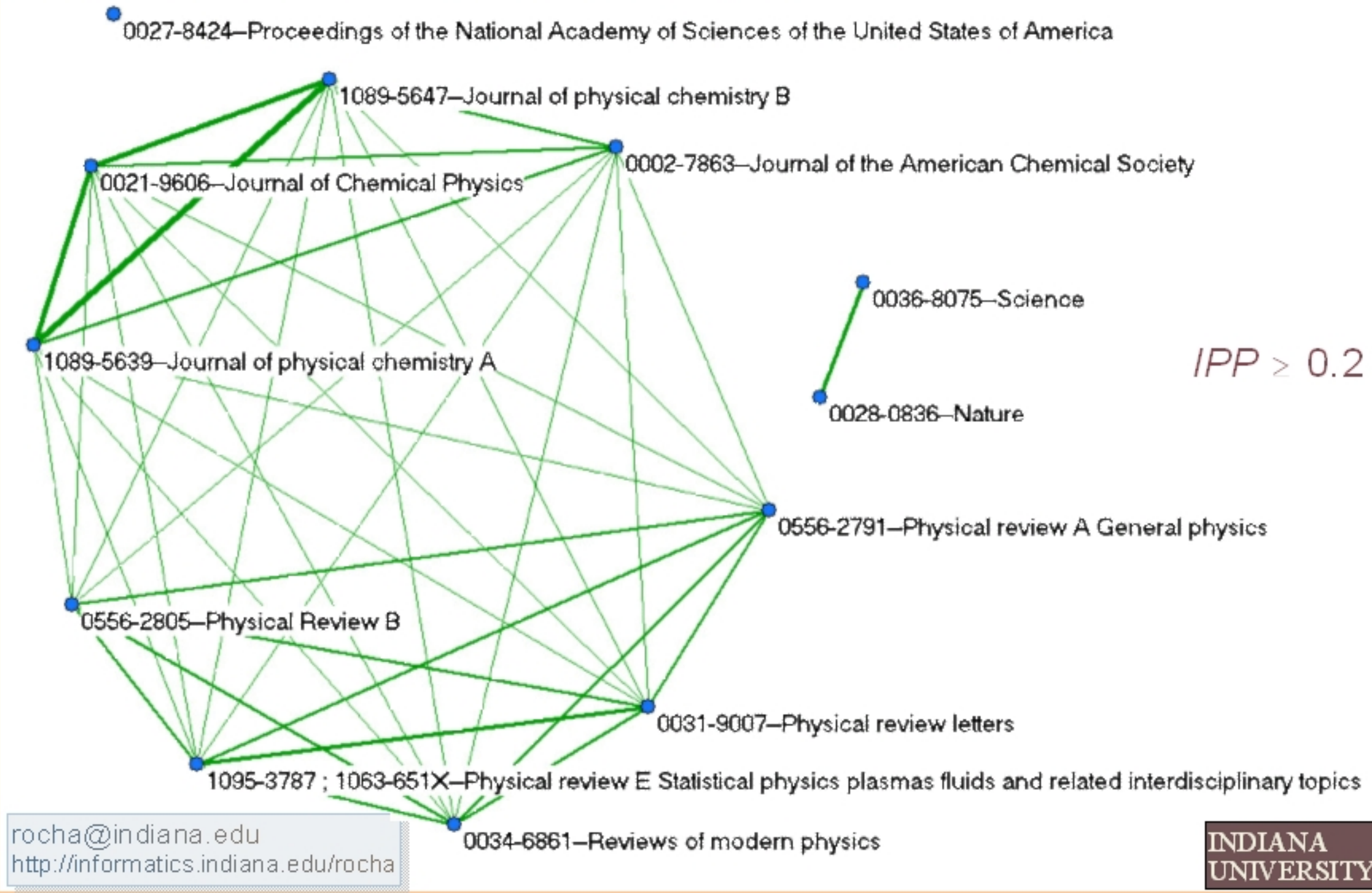


rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA
UNIVERSITY

IPP network

sub-graph with top 12 most frequent journals

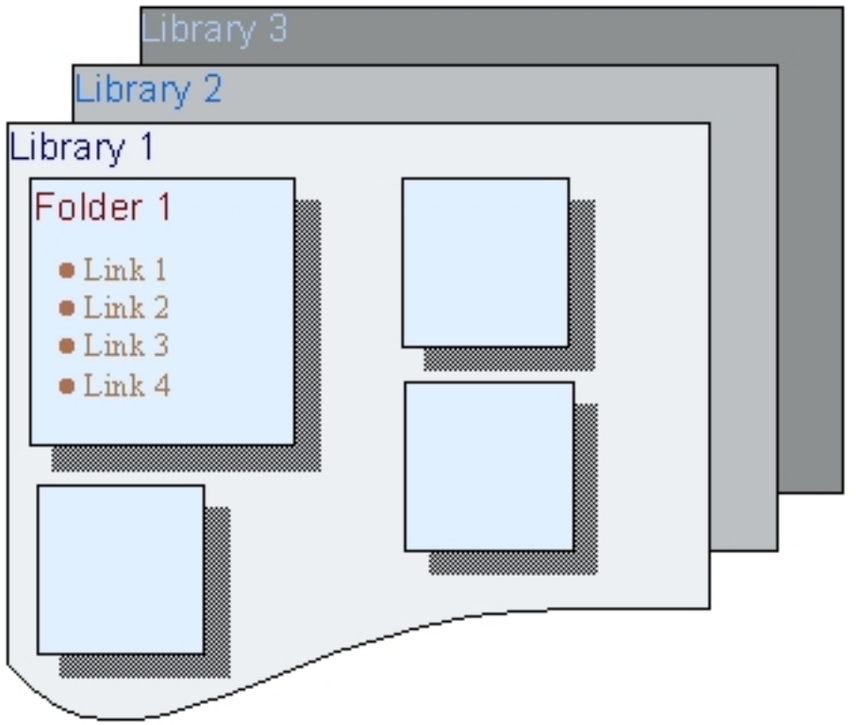


rocha@indiana.edu
http://informatics.indiana.edu/rocha

integrating proximity edges

$$I_R = \left\{ i_t : \underset{\forall i_s \in I(p_u)}{MAX} \left(IPP(i_s, i_t) \right) \geq \alpha \right\}$$

Returns all journals in the network related to at least one of the journals in personality with proximity ≥ α



$$I_R = \left\{ i_t : \underset{\forall i_s \in I(p_u)}{MIN} \left(IPP(i_s, i_t) \right) \geq \alpha \right\}$$

Returns all journals in the network related to **every** journal in personality with proximity at least ≥ α

$$I_R = \left\{ i_t : \underset{\forall i_s \in I(p_u)}{AVG} \left(IPP(i_s, i_t) \right) \geq \alpha \right\}$$

informatics
luis rocha 2007

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

informatics
luis rocha 2007

recommendation example

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

recommendations based on proximity

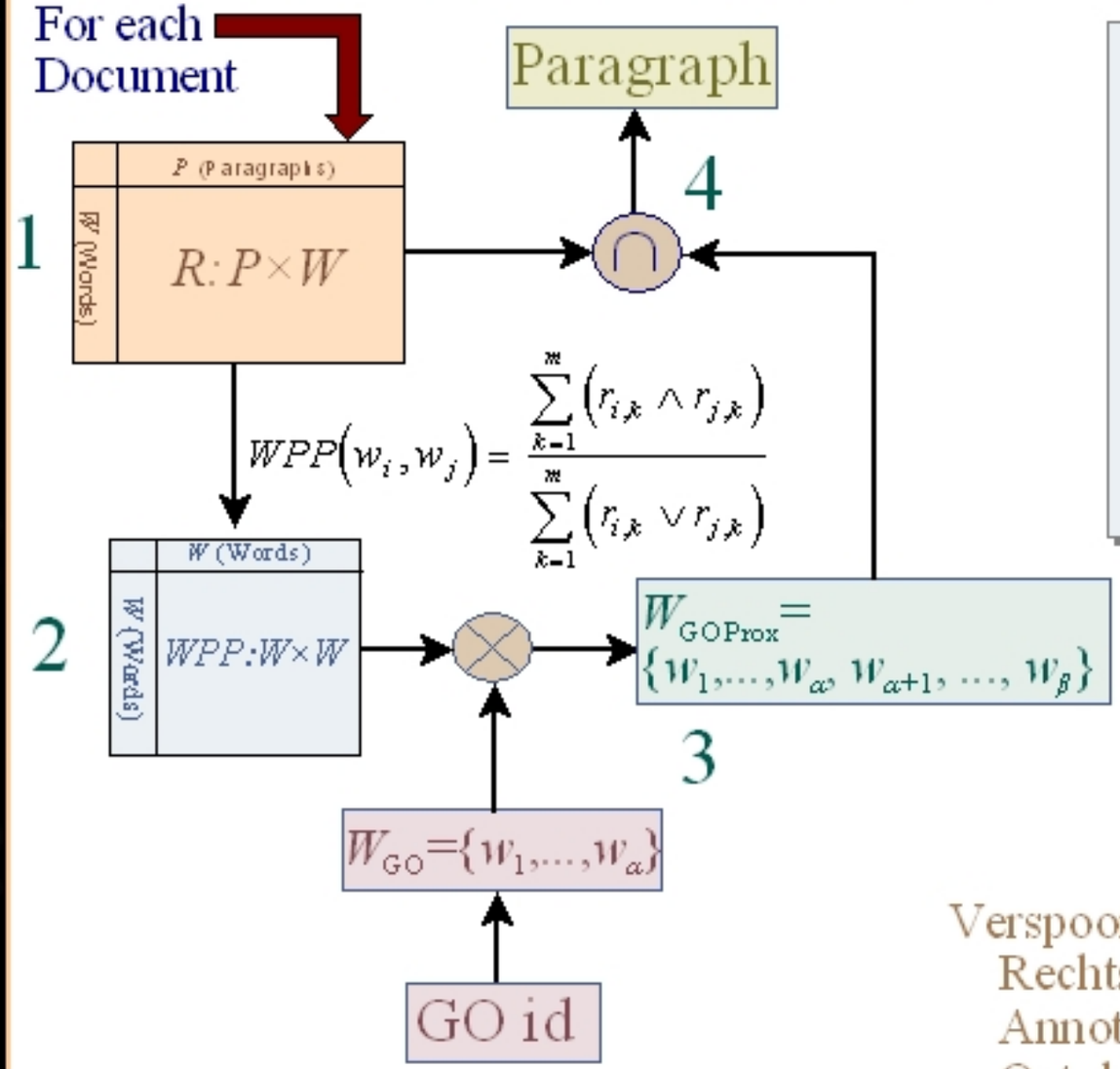
mylibrary.lanl.gov

- IPP
  - Recommendations of ISSN based on co-occurrence in Personalities
    - Users who linked to this journal, also linked to...
- PIP
  - Recommendations of other users' personalities: collaboration
    - These personalities are similar to yours
  - Recommendations of specific links in close personalities
    - Users who read many of the same journals where interested in these links

rocha@indiana.edu
http://informatics.indiana.edu/rocha

Biocreative competition (EMBO Workshop)

a critical assessment of text mining methods in molecular biology



- Task 2: Given a document, discover the portion of text most appropriate to annotate the protein's function, and produce appropriate Gene Ontology node for annotation
  - Learning set: triplets (protein, document, GO id)
  - Test set: documents

Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2005]. "Protein Annotation as Term Categorization in the Gene Ontology using Word Proximity Networks". *BMC Bioinformatics*, **6**(Suppl 1):S20. doi:10.1186/1471-2105-6-S1-S20

informatics
luis rocha 2007

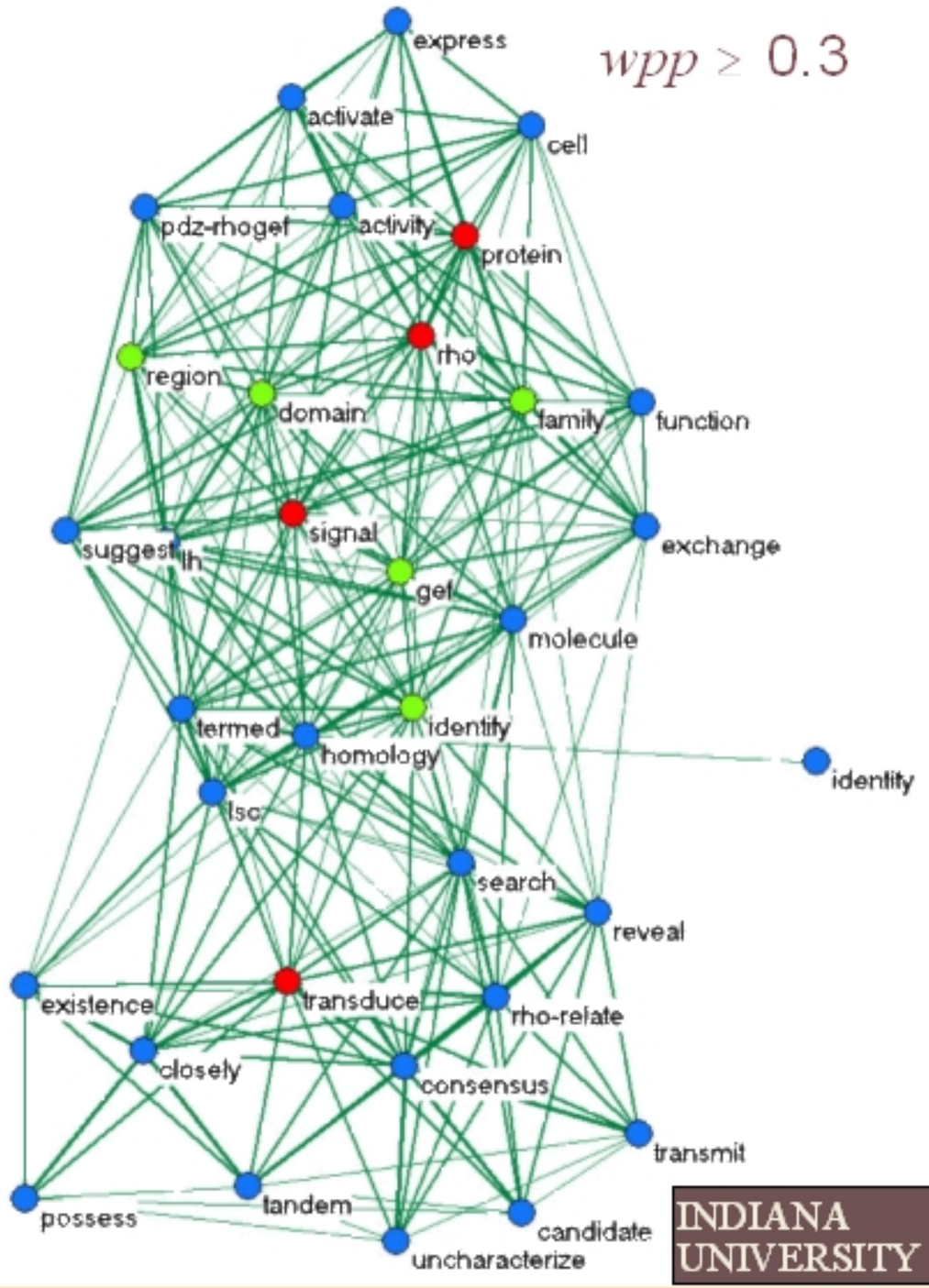
rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

word in paragraph proximity

example document

- document bc005868
  - *WPP* contains 1102 words
  - Subgraph of 34 words
    - Red nodes: words removed from the respective GO annotation (0007266): Rho, protein, signal, transducer).
    - Blue nodes: words that co-occur very frequently ($wpp > 0.5$) with at least one of the red nodes
    - Green nodes: additional words recommended with largest average proximity to all input words (red nodes)



Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2005]. "Protein Annotation as Term Categorization in the Gene Ontology using Word Proximity Networks". *BMC Bioinformatics*, 6(Suppl 1):S20. doi:10.1186/1471-2105-6-S1-S20

rocha@indiana.edu
http://informatics.indiana.edu/rocha

Task 2.1 Results

Proximity-based run

User, Run	"perfect"	"generally"	cumulative
7, 1	25.28%	14.31%	39.59%
14, 1	28.16%	6.41%	34.57%
20, 1	27.97%	5.30%	33.27%
4, 1	24.91%	6.88%	31.78%
20, 2	26.02%	5.58%	31.60%
20, 3	22.21%	5.48%	27.70%
5, 2	15.43%	8.36%	23.79%
5, 1	15.43%	7.16%	22.58%
5, 3	14.31%	7.99%	22.30%
15, 2	11.62%	6.41%	18.03%
9, 1	11.62%	1.21%	12.83%
7, 3	6.13%	3.72%	9.85%
17, 1	7.71%	1.77%	9.48%
15, 1	5.48%	2.60%	8.09%
7, 2	4.00%	3.72%	7.71%
10, 3	4.65%	0.37%	5.02%
9, 3	3.81%	0.65%	4.46%
10, 2	4.18%	0.19%	4.37%
10, 1	3.35%	0.28%	3.62%
9, 2	3.07%	0.46%	3.53%
17, 2	0.65%	0.00%	0.65%

Verspoor, K., et al [2005]. *BMC Bioinformatics*, **6**(Suppl 1):S20. doi:10.1186/1471-2105-6-S1-S20

informatics
luis rocha 2007

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

informatics
luis rocha 2007

proximity networks in Biocreative 2

Abi-Haidar et al [2008]. *Genome Biology*. In Press.

- IPS and ISS tasks: identify protein-protein interaction pairs and passages in text

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

Testing proximity network recommendation

With Movielens data

- Movielens data
  - http://www.grouplens.org/node/73
- Dual Networks for each dataset
  - Movie (item) and User Proximity
- Results (F1 Measure)
  - Top 10 Movies
    - User-based: 0.213. Item-based: 0.184: Vector/Cosine: 0.212. LSI: 0.245
  - Arbitrary Number
    - User-based: 0.305. Item-based: 0.294.Vector/Cosine: 0.306. LSI: 0.328
- Results (Fouss' variant of Sommers D Measure)
  - Top 10 Movies
    - User-based: 88.2%. Item-based: 89.53%.Vector/Cosine: 86.02%. LSI: 91.69%
  - Arbitrary Number
    - User-based: 86.21%. Item-based: 87.95%.Vector/Cosine: 85.18%. LSI: 89.01

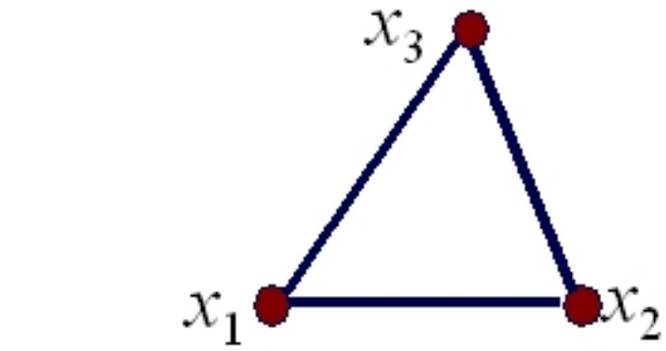
rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2007

semi-metric behavior

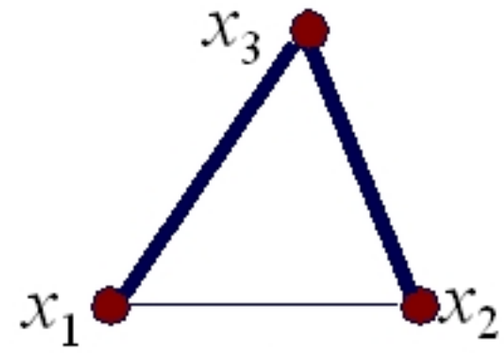
$$d_X\left(x_i, x_j\right) = \frac{1}{XYP\left(x_i, x_j\right)} - 1; \quad d_Y\left(y_i, y_j\right) = \frac{1}{YXP\left(y_i, y_j\right)} - 1$$

d is a distance function because it is a nonnegative, symmetric, real-valued function such that $d(k, k) = 0$



$d(x_1, x_2) \leq d(x_1, x_3) + d(x_3, x_2)$

Metric



$d(x_1, x_2) > d(x_1, x_3) + d(x_3, x_2)$

Semi-metric

rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2007

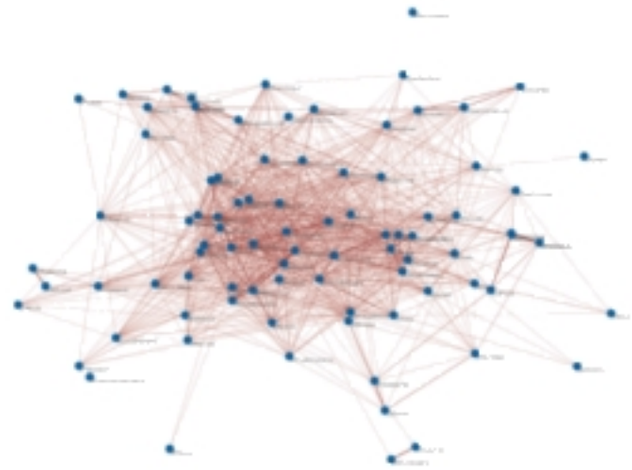
computing semi-metric behavior

	X (Keywords)
Y (Documents)	$R:X\times Y$

	Y (Documents)
Y (Documents)	$YXP:Y\times Y$

(Proximity)

	X (Keywords)
X (Keywords)	$XYP:X\times X$



$$d_X(x_i,x_j)=\frac{1}{XYP(x_i,x_j)}-1$$

	X (Keywords)
X (Keywords)	$d_X:X\times X$

Metric Closure (min/+)

	X (Keywords)
X (Keywords)	$d_X{}^{mc}:X\times X$

(Distance)

= metric
> semi-metric

(Shortest Distance Path)

Measures of semi-metric behavior

rocha@indiana.edu
http://informatics.indiana.edu/rocha

measuring semi-metric behavior

Semi-metric Measures

- Semi-metric ratio
  - Absolute measure of indirect distance reduction

$$s(x_i, x_j) = \frac{d_{direct}(x_i, x_j)}{d_{shortest}(x_i, x_j)}$$

- Relative Semi-metric ratio
  - Distance reduction against maximum contraction

$$rs(x_i, x_j) = \frac{d_{direct}(x_i, x_j) - d_{shortest}(x_i, x_j)}{d_{\max} - d_{\min}}$$

- Below Average Ratio
  - Captures semi-metric distance reductions which contract to below the average distance for a given node. Captures some of the cases of initial ∞ distance

$$b(x_i, x_j) = \frac{\overline{d_{x_i}}}{d_{shortest}(x_i, x_j)}$$

rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2007

semi-metric recommendations

catching strong indirect associations in mylibrary.lanl.gov

0020-1669--Inorganic chemistry 0031-9007--Physical review letters
0031-9007--Physical review letters 0743-7463--Langmuir
0003-2700--Analytical chemistry 0031-9007--Physical review letters
0096-3003--Applied mathematics and computation 0031-9007--Physical review letters
0031-9007--Physical review letters 0022-3115--Journal of nuclear materials
1049-3301--ACM transactions on modeling and computer simulation 0031-9007--Physical review letters
1364-548X--Chemical communications 0031-9007--Physical review letters
1064-8275--SIAM journal on scientific computing 0031-9007--Physical review letters
0965-5425--Computational mathematics and mathematical physics 0031-9007--Physical review letters
0031-9007--Physical review letters 1359-6454--Acta materialia
0003-7028--Applied spectroscopy 0031-9007--Physical review letters
0031-9007--Physical review letters 0022-2461--Journal of materials science
0031-9007--Physical review letters 1359-6462--Scripta materialia
0031-9007--Physical review letters 0022-4596--Journal of solid state chemistry
0031-9007--Physical review letters 0021-8898--Journal of applied crystallography

IPP_3: parameter *rs*

1097-6256--Nature neuroscience 1065-9471--Human Brain MApping
1097-6256--Nature neuroscience 0278-0062--IEEE transactions on medical imaging
1097-6256--Nature neuroscience 1053-8119--NeuroImage
1063-7796--Physics of particles and nuclei 0218-3013--International journal of modern physics E Nuclear physics
1053-8119--NeuroImage 1065-9471--Human Brain MApping
0031-9007--Physical review letters 0743-7463--Langmuir
0031-9007--Physical review letters 0020-1669--Inorganic chemistry
0031-9007--Physical review letters 0141-1594--Phase transitions
0031-9007--Physical review letters 0928-1045--Journal of computeraided materials design
0031-9007--Physical review letters 0042-207X--Vacuum
1097-6256--Nature neuroscience 0031-9155--Physics in medicine & biology
1097-6256--Nature neuroscience 0096-3518--IEEE transactions on acoustics speech and signal processing
1097-6256--Nature neuroscience 0740-7467--IEEE ASSP magazine
1097-6256--Nature neuroscience 1070-9908--IEEE signal processing letters
0022-5355--Journal of vacuum science and technology 0734-2101--Journal of vacuum science & technology A Vacuum surfaces and films

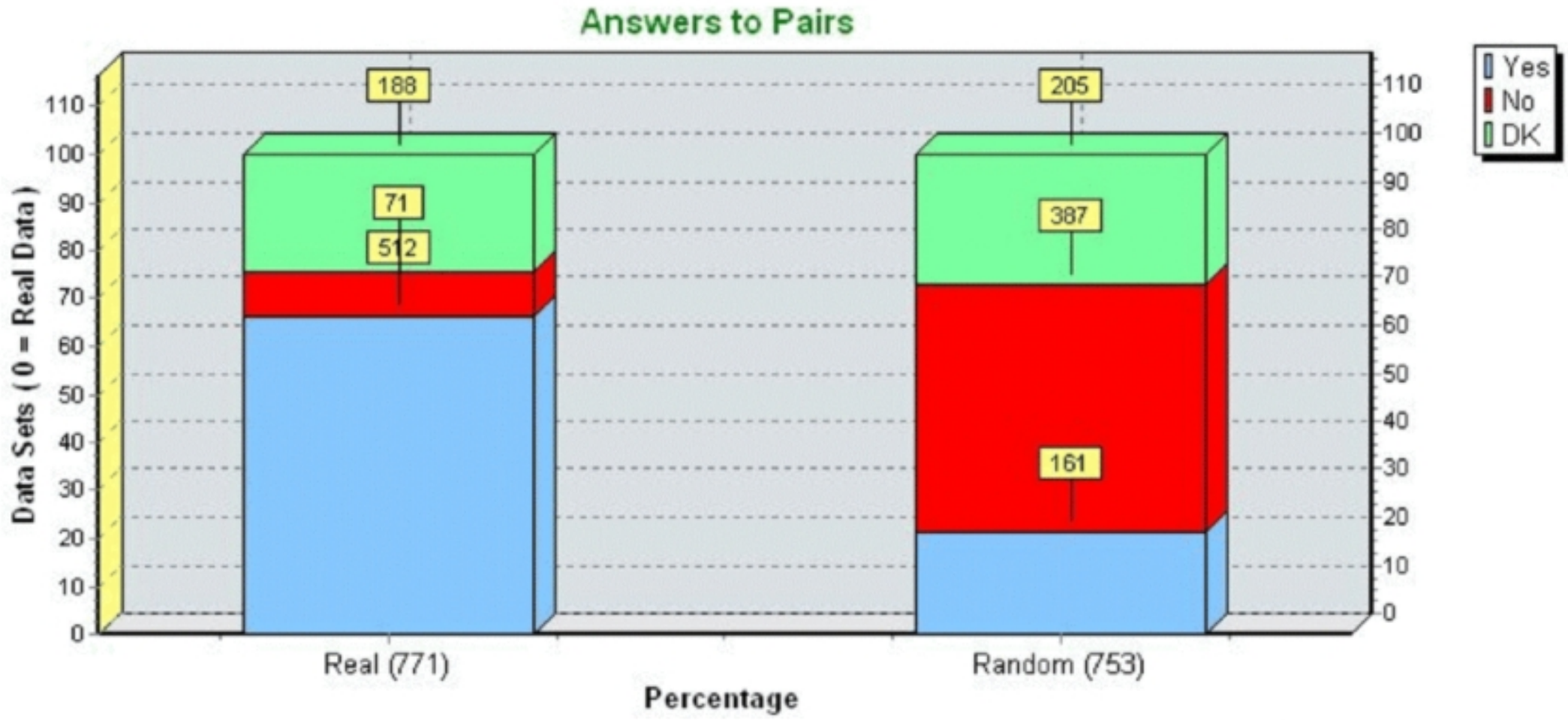
IPP_3: parameter *b*

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

informatics
luis rocha 2007

population from STB-RL and CCS-3



- 50% Pairs of ISSN with high *rs* and *b*
- 50% Random ISSN Pairs
- 1524 answers

rocha@indiana.edu
http://informatics.indiana.edu/rocha

what do semi-metric edges imply?

- Pairs with larger semi-metric behavior denote a *latent association*
  - Not grounded on direct evidence provided by the relation R, but rather implied by the overall network of associations in this relation.
  - Meaning depends on the semantics of the application
    - In graphs of keyword co-occurrence in documents: associated with novelty and can be used to identify trends.
    - In social networks it may identify pairs of people, groups, etc. for which we do not have direct evidence, in the available documents, that a real association exists, but who could easily be indirectly associated.

Rocha, Luis M. [2002]. "Semi-metric Behavior in Document Networks and its Application to Recommendation Systems". In: *Soft Computing Agents: A New Perspective for Dynamic Information Systems*. V. Loia (Ed.) International Series Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 137-163.

Rocha, Luis M. [2002]. "Combination of Evidence in Recommendation Systems Characterized by Distance Functions". In: *Proceedings of the 2002 World Congress on Computational Intelligence: FUZZ-IEEE'02*. Honolulu, Hawaii, May 2002. IEEE Press, pp. 203-208.

Rocha, L.M., T. Simas, A. Rechtsteiner, M. DiGiacomo, R. Luce [2005]. "MyLibrary@LANL: Proximity and Semi-metric Networks for a Collaborative and Recommender Web Service". In: *Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, IEEE Press, pp. 565-571.

rocha@indiana.edu
http://informatics.indiana.edu/rocha

scientific community working on feynman diagrams

as published in *Physical Review*, 1949-54

informatics
luis rocha 2007

	P(author names)
P(author names)	$C : P \times P$

- Collaboration Relation: C
  - Who wrote a paper with whom
- Acknowledgment Relation: A
  - Who acknowledged, or informally received information from whom

	P(author names)
P(author names)	$A : P \times P$

$$CP(p_i, p_j) = \frac{\sum_{k=1}^{m} (c_{i,k} \wedge c_{j,k})}{\sum_{k=1}^{m} (c_{i,k} \vee c_{j,k})}$$

76 Authors

$CP(p_i, p_j)$ is a ***co-collaboration probability***: the probability that two authors have collaborated with the same authors

$$AP(p_i, p_j) = \frac{\sum_{k=1}^{m} (a_{i,k} \wedge a_{j,k})}{\sum_{k=1}^{m} (a_{i,k} \vee a_{j,k})}$$

91 Authors

$AP(p_i, p_j)$ is a ***co-acknowledgment probability***: the probability that two authors have acknowledged or have been acknowledged by the same authors

rocha@indiana.edu
http://informatics.indiana.edu/rocha

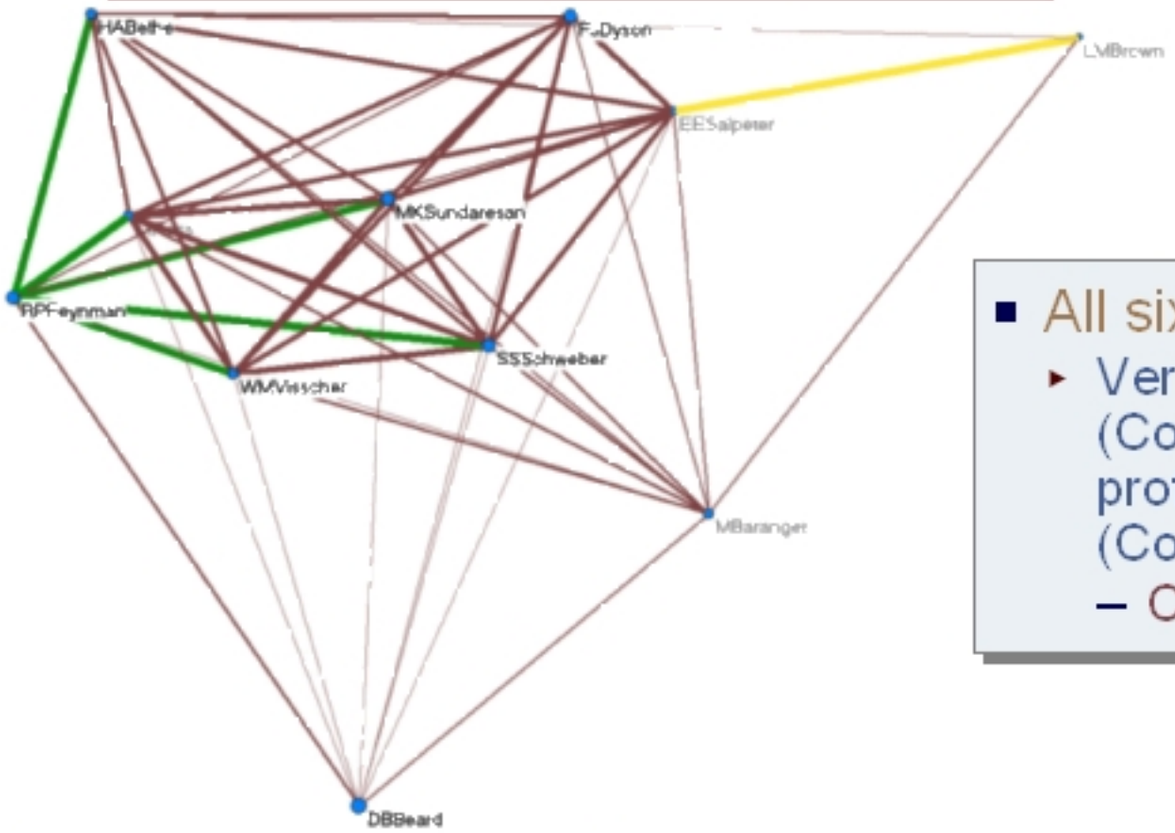
With Luis Bettencourt

INDIANA UNIVERSITY

co-collaboration network

5 most semi-metric pairs (rs and b parameters)

HABethe	RPFeynman
RPFeynman	WMVisscher
MRoss	RPFeynman
RPFeynman	SSSchweber
MKSundaresan	RPFeynman



- Weak co-collaboration (CP), but strong co-acknowledgment (AP).
  - While they have not co-collaborated much in PR, they have acknowledged or have been acknowledged by many of the same people

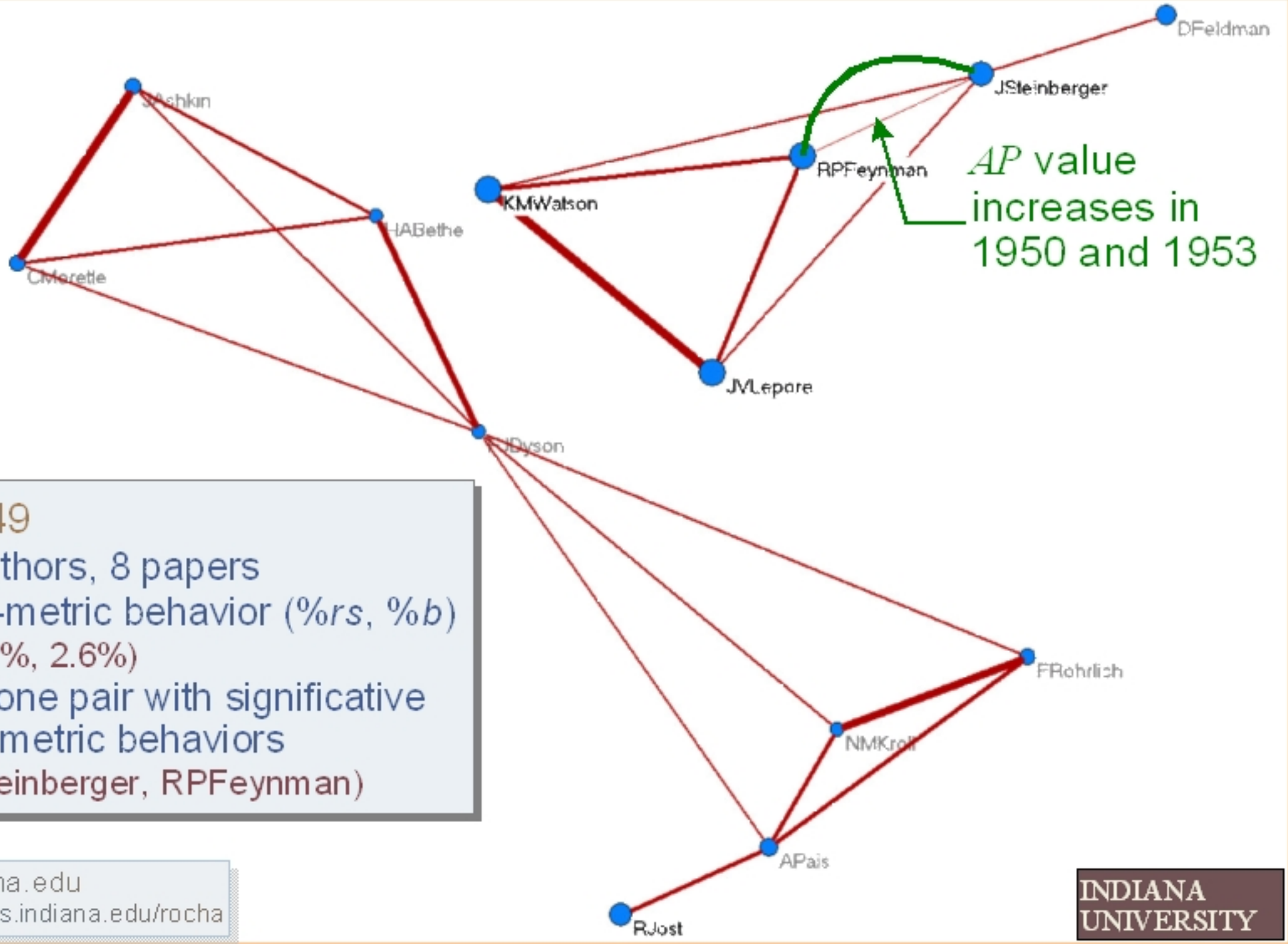
- All six authors in top pairs
  - Very strong proximities to *EESalpeter* (Cornell) and *KMWatson* (postdoc at IAS, prof at Indiana) in AP and with to *EESalpeter* (Cornell) and *FJDyson* (Cornell, IAS) in CP
    - Cornell and Institute of Advanced Studies

rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2007

dynamics of co-acknowledgment network

1949



- *AP* 1949
  - 13 authors, 8 papers
  - Semi-metric behavior ($\%rs$, $\%b$)
    - (3.9%, 2.6%)
  - Only one pair with significative semi-metric behaviors
    - (Jsteinberger, RPFeynman)

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

informatics
luis rocha 2007

Dynamics of Information Flow in Social NetworksProject

NSF HSD - Dynamics of Human Behavior

- Collected Facebook data
  - Subset of IUB volunteers
    - **Questionnaires** about themselves, their relationship partner, their use of facebook, et.
    - **Ratings**. show elements of their profile pages to raters, who can rate physical attractiveness of the photo, etc.
  - Network data
    - 2 time instances with user (333 and 260) profiles as well as friends at IU, and on those who made the most recent 10 posts on their "wall."
    - Other linked users scrambled from one time measurement to the other
  - Testing
    - Semi-metric links in co-friendship network as predictor of new friends
    - Keyword content as predictor of communities, atractivness, etc.

With Elliot Smith and Rob Goldstone

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

comparison with max-min transitivity

shortest paths or weakest links

Edges: largest of the weakest links in all paths:

Proximity

	X (Keywords)
X (Keywords)	$XYP: X \times X$

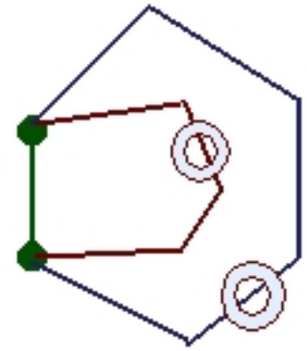
Transitive Closure (max/min)

For any monotonic increasing distance function

$$d_X(x_i, x_j) = \frac{1}{XYP(x_i, x_j)} - 1$$

Similarity

	X (Keywords)
X (Keywords)	$XYS: X \times X$



	X (Keywords)
X (Keywords)	$d_X: X \times X$

Distance

semi-metric

Shortest Path (min/+)

	X (Keywords)
X (Keywords)	$d^*_X: X \times X$

Edges: shortest path (sums all edges)

metric

$\geq$

	X (Keywords)
X (Keywords)	$d^{**}_X: X \times X$

Edges: Smallest of the largest edges in each path

ultra-metric: immune to indirect path length

rocha@indiana.edu
http://informatics.indiana.edu/rocha

With Tiago Simas

INDIANA UNIVERSITY

effect of typical transitive closure

IPP: cumulative strenght distribution (all weights)

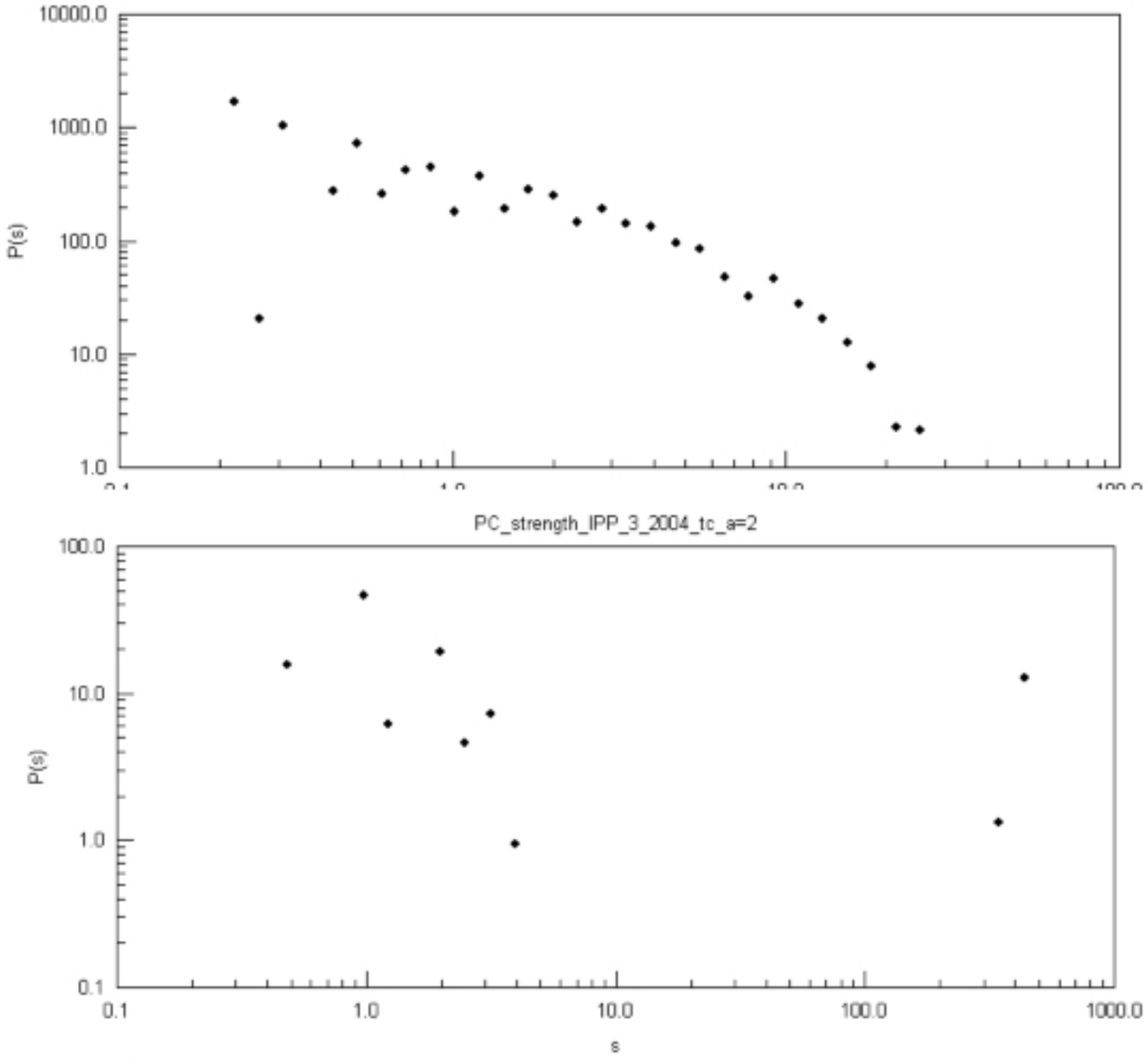
Proximity

	X (Keywords)
X (Keywords)	$XYP: X \times X$

Transitive Closure (max/min)

Similarity

	X (Keywords)
X (Keywords)	$XYS: X \times X$



α-cut = 0.2

Destroys scale free behavior

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

IPP: cumulative strenght distribution

Proximity

	X(Keywords)
X(Keywords)	$XYP: X \times X$

$$d_X(x_i, x_j) = \frac{1}{XYP(x_i, x_j)} - 1$$

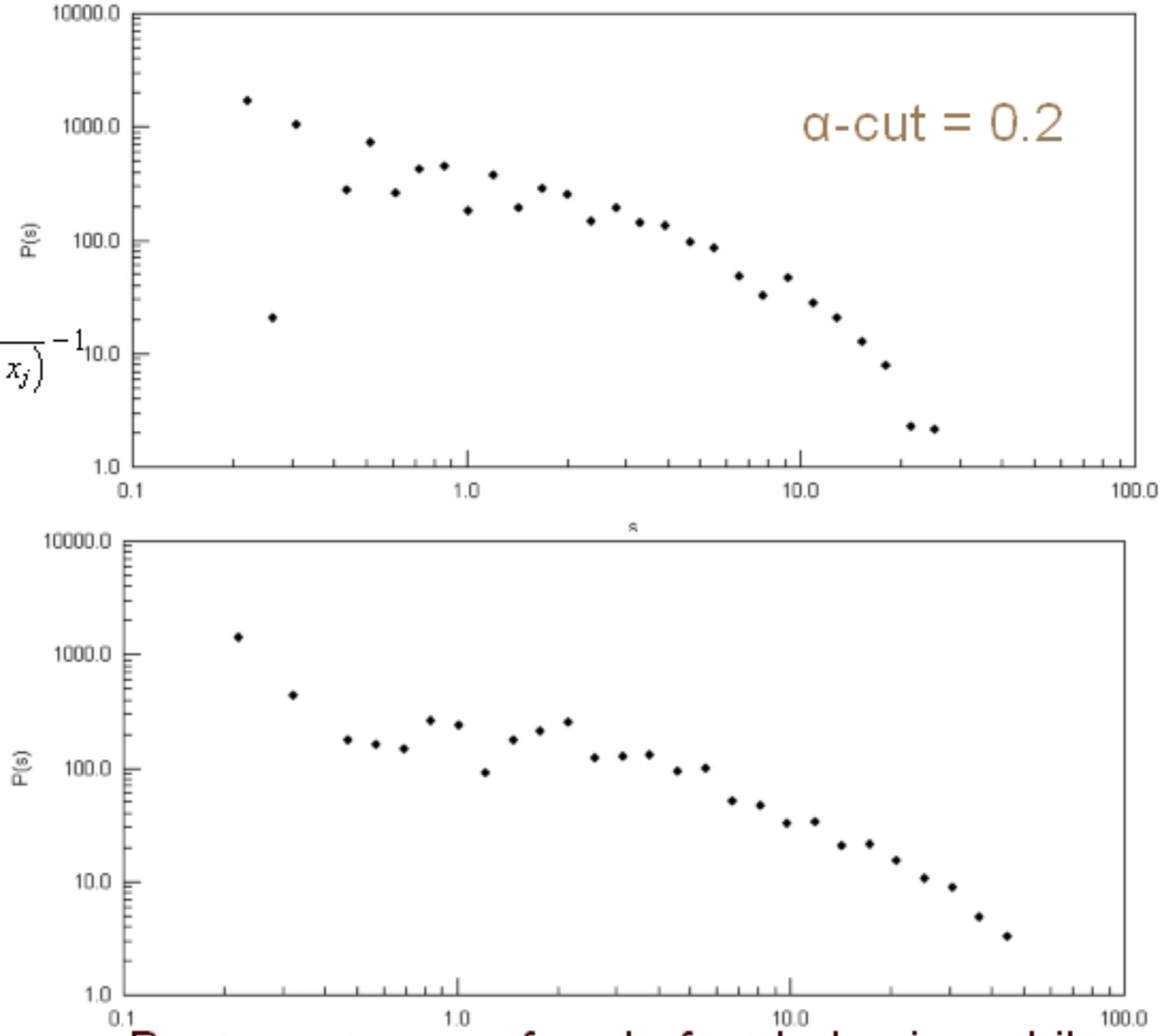
	X(Keywords)
X(Keywords)	$d_X: X \times X$

Distance

Shortest Path (min/+)

Distance Closure

	X(Keywords)
X(Keywords)	$d^*_X: X \times X$



Preserves more of scale free behavior, while still capturing indirect (transitive/latent) associations in the data

rocha@indiana.edu
http://informatics.indiana.edu/rocha

distance-equivalent fuzzy graph

Hammacher function

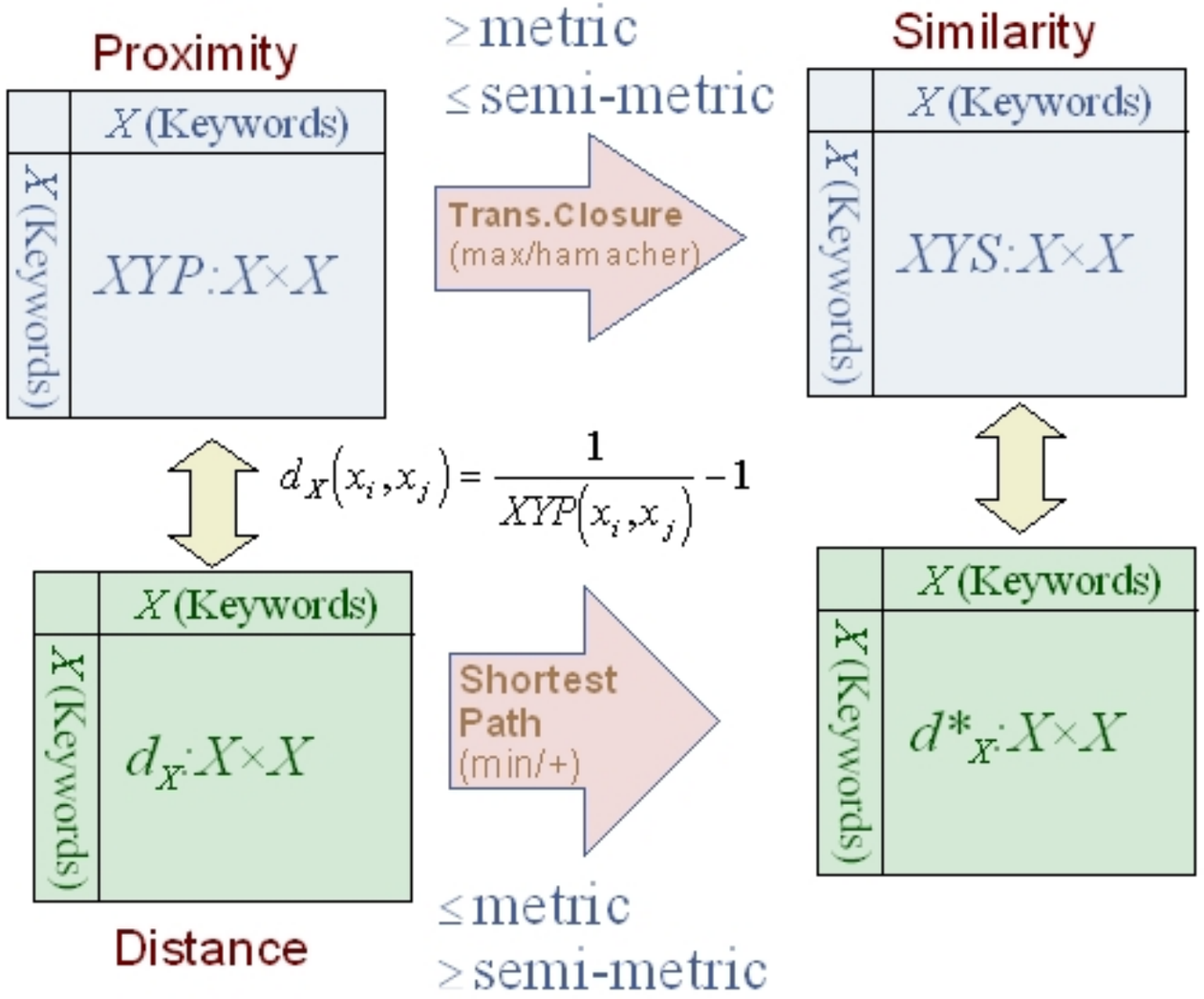
If we use :

$$i(a,b) = \frac{ab}{a+b-ab}$$

Hammacher Intersection

$$u(a,b) = \max[a,b]$$

With Tiago Simas



rocha@indiana.edu
http://informatics.indiana.edu/rocha

metric and ultra-metric closures on distance graphs

informatics
luis rocha 2007

Proximity graphs

Similarity graphs

	X (Keywords)
X (Keywords)	$XYP: X \times X$

(max,min) transitive closure

$$\max_k \min\left(xyp_{ik}, xyp_{kj}\right)$$

	X (Keywords)
X (Keywords)	$XYS^*: X \times X$

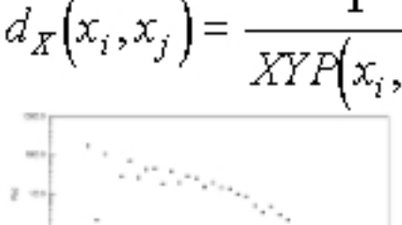
$\leq$

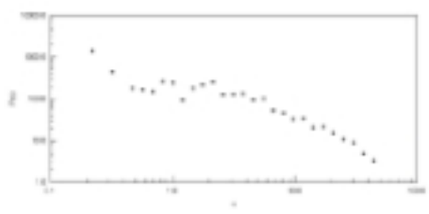
	X (Keywords)
X (Keywords)	$XYS^{**}: X \times X$

(max,$H_\wedge$) transitive closure

$$\max_k H_\wedge\left(xyp_{ik}, xyp_{kj}\right)$$

$$d_X\left(x_i, x_j\right) = \frac{1}{XYP\left(x_i, x_j\right)} - 1$$





$$d_X\left(x_i, x_j\right) = \frac{1}{XYP\left(x_i, x_j\right)} - 1$$

	X (Keywords)
X (Keywords)	$d^*_X: X \times X$

	X (Keywords)
X (Keywords)	$d^{**}_X: X \times X$

(min,+) metric closure

$$\min_k\left(d^X_{ik} + d^X_{kj}\right)$$

$\geq$

	X (Keywords)
X (Keywords)	$d_X: X \times X$

(min,max) ultra-metric closure

$$\min_k \max\left(d^X_{ik}, d^X_{kj}\right)$$

Semi-metric distance graphs

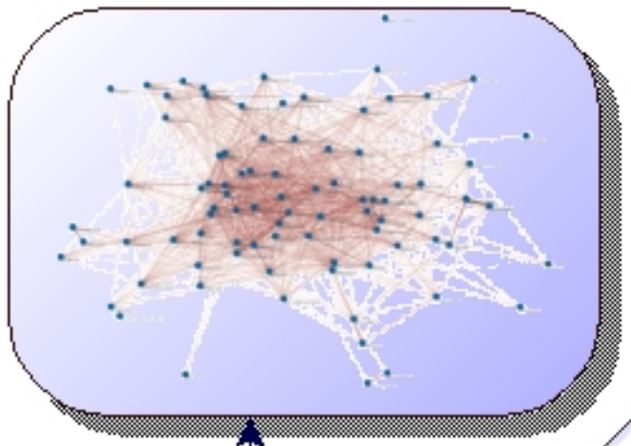
Metric distance graphs

rocha@indiana.edu
http://informatics.indiana.edu/rocha

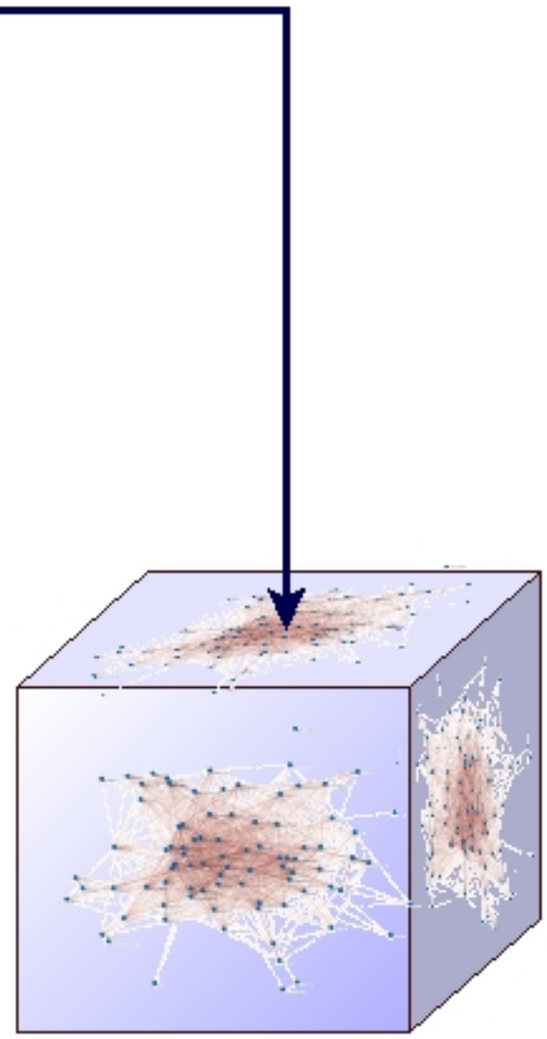
INDIANA UNIVERSITY

collecting data from users with multiple personalities

from single to multifaceted knowledge representations



Need for effective integration of various knowledge networks



rocha@indiana.edu
http://informatics.indiana.edu/rocha

What are the best ways to conbine graph knowledge representations?

informatics
luis rocha 2007

Hammacher Intersection

$$i(a,b) = \frac{ab}{a + b - ab}$$

$$u(a,b) = \max[a,b]$$

Metric closure is not dual: no complement can satisfy de Morgan's laws!

- What fuzzy intersection/union comes closest to the metric closure of distance graphs?
  - Can we have a fuzzy union/intersection that follows De Morgan's laws but make it as close as possible to a metric closure?
- What captures the semi-metric behavior best?
  - The metric closure or some other transitive closure?

rocha@indiana.edu
http://informatics.indiana.edu/rocha

Dombi family

$$u(a,b) = a \cup b = \frac{1}{1+\left[\left(\frac{1}{a}-1\right)^{-\lambda} + \left(\frac{1}{b}-1\right)^{-\lambda}\right]^{-\frac{1}{\lambda}}}$$

$$\lambda \to 0 \Rightarrow a \cup b \to 1$$

$$\lambda = 1 \Rightarrow a \cup b = \frac{a+b-2ab}{1-ab}$$

Hammacher union

$$\lambda \to \infty \Rightarrow a \cup b \to \max(a,b)$$

rocha@indiana.edu
http://informatics.indiana.edu/rocha

De Morgan Laws and Involutive Complement

$$i(a,b) = \frac{ab}{a+b-ab} \qquad u(a,b) = \frac{1}{1+\left[\left(\frac{1}{a}-1\right)^{-\lambda}+\left(\frac{1}{b}-1\right)^{-\lambda}\right]^{-\frac{1}{\lambda}}}$$

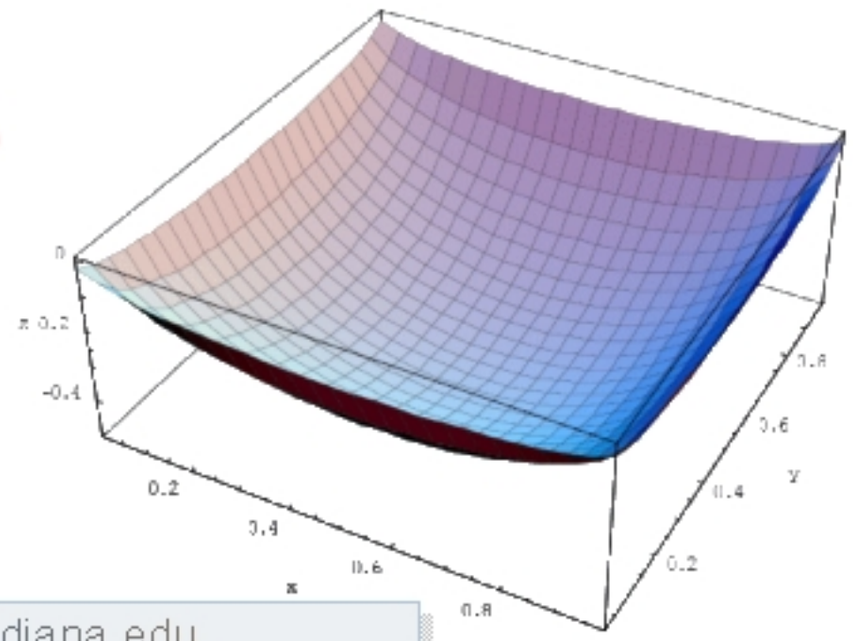
Requirement for desired axiomatics

$$-ab\left[\left(\frac{1}{a}-1\right)^{\lambda}+\left(\frac{1}{b}-1\right)^{\lambda}\right]^{\frac{1}{\lambda}}+a+b-2ab=0$$

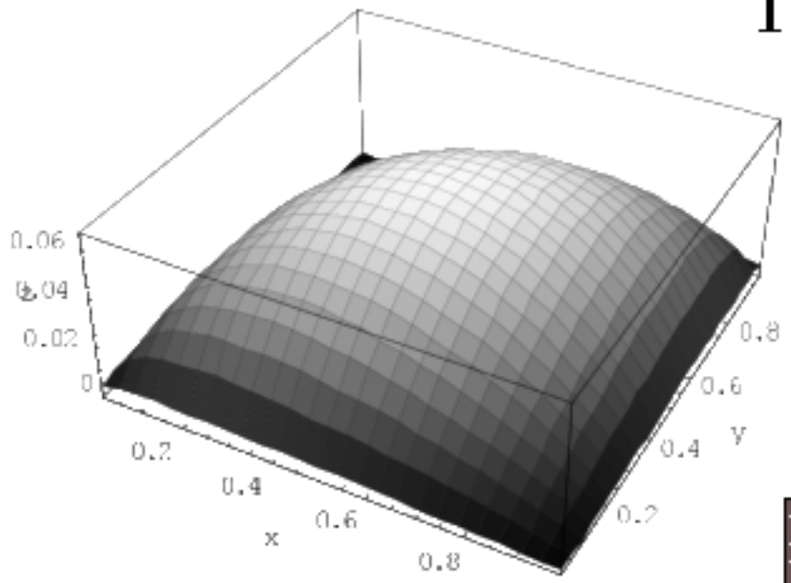
Unique solution for λ=1

$$\lambda = 1 \Rightarrow u(a,b) = \frac{a+b-2ab}{1-ab}$$

λ=0.5



λ=1.5



rocha@indiana.edu
http://informatics.indiana.edu/rocha

informatics
luis rocha 2007

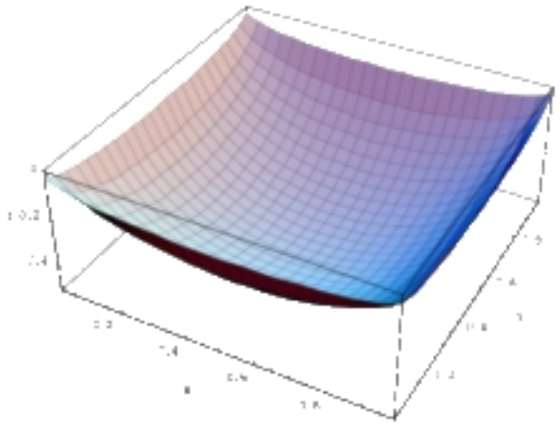
INDIANA UNIVERSITY

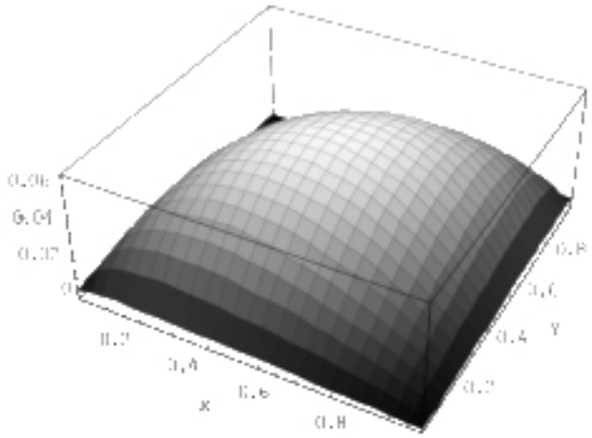
De Morgan Laws and Involutive Complement

Error for other λ

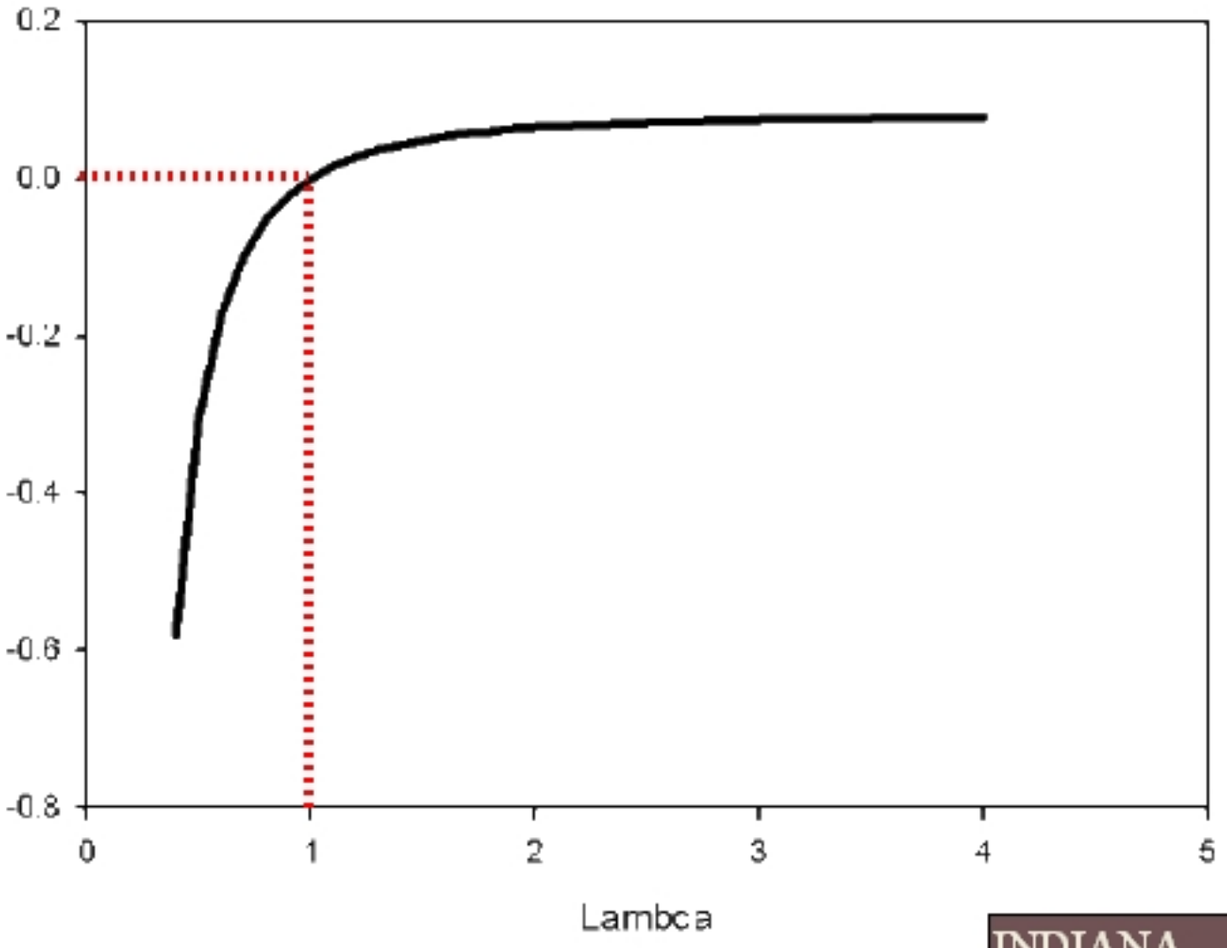
$$F(\lambda) = \int_0^1 \int_0^1 \left(-xy \left[\left(\frac{1}{x} - 1 \right)^{\lambda} + \left(\frac{1}{y} - 1 \right)^{\lambda} \right]^{\frac{1}{\lambda}} + x + y - 2xy \right) dxdy$$

λ=0.5





λ=1.5

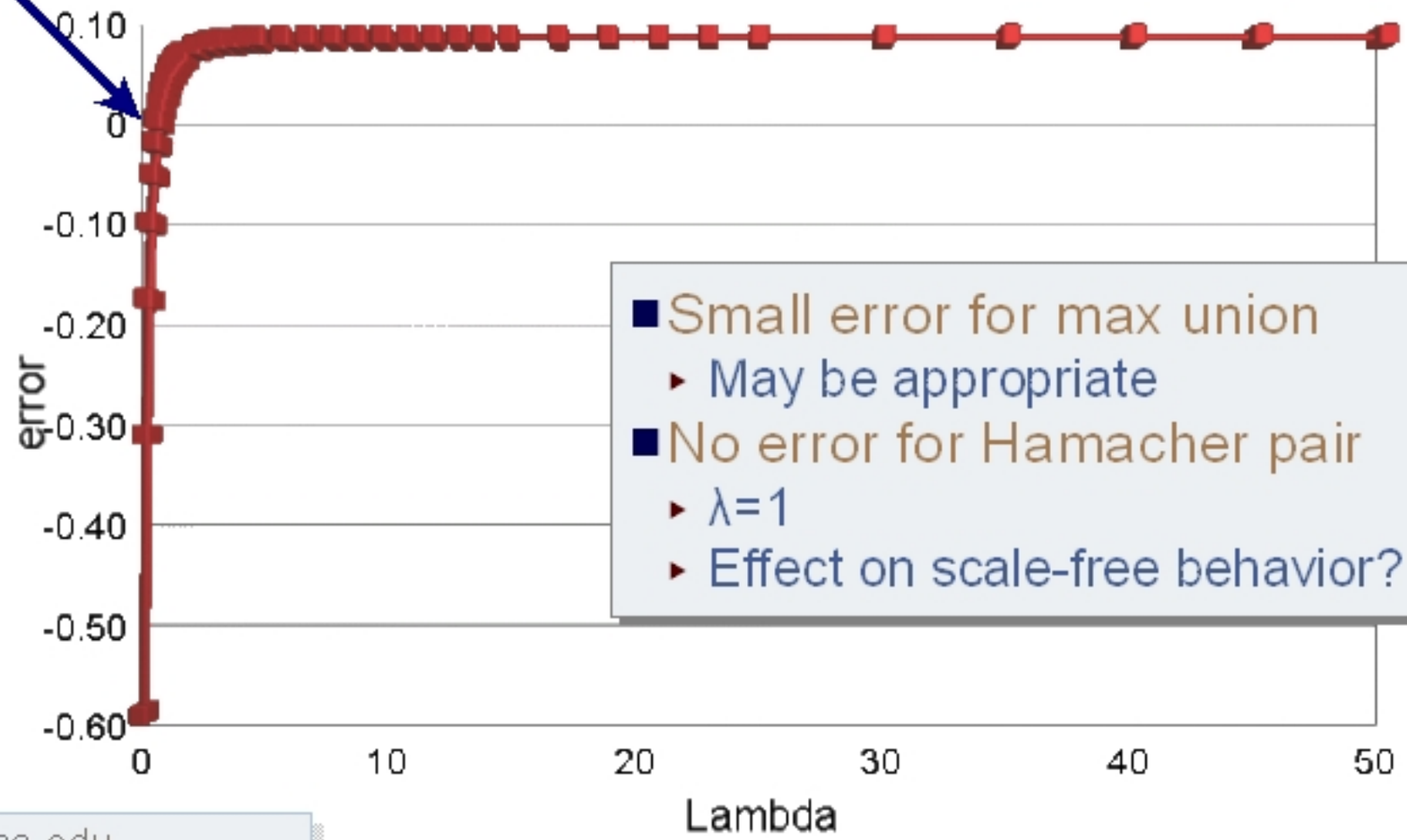


rocha@indiana.edu
http://informatics.indiana.edu/rocha

deviation from desired axiomatics

$$\lambda = 1 \Rightarrow a \cup b = \frac{a + b - 2ab}{1 - ab}$$

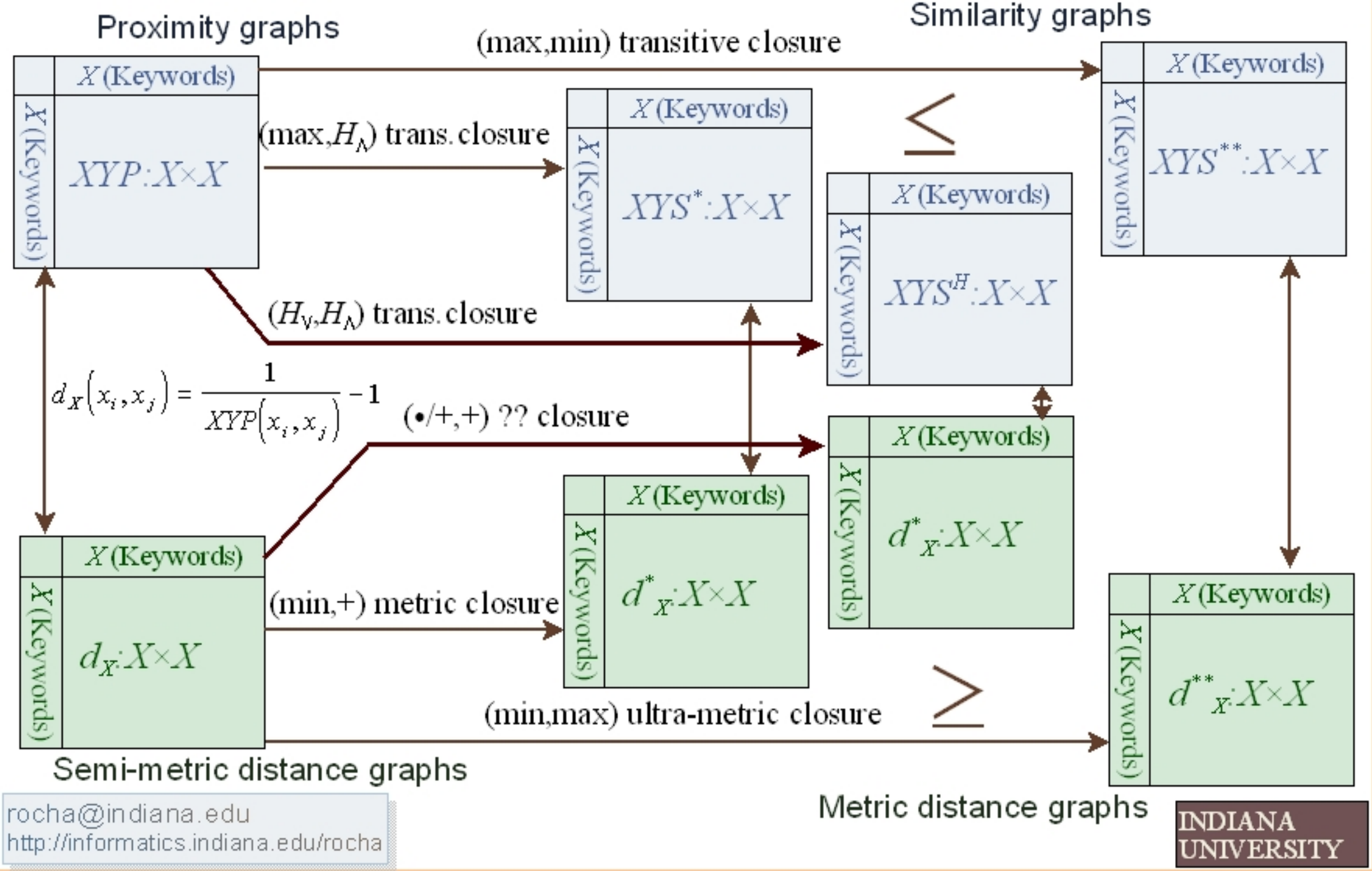
$$\lambda \to \infty \Rightarrow a \cup b \to \max(a,b)$$



- Small error for max union
  - May be appropriate
- No error for Hamacher pair
  - λ=1
  - Effect on scale-free behavior?

rocha@indiana.edu
http://informatics.indiana.edu/rocha

metric and ultra-metric closures on distance graphs



informatics
luis rocha 2007

rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY