informatics
luis rocha 2005

proximity networks and semi-metric behavior

in bioinformatics, social networks, and recommendation systems

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

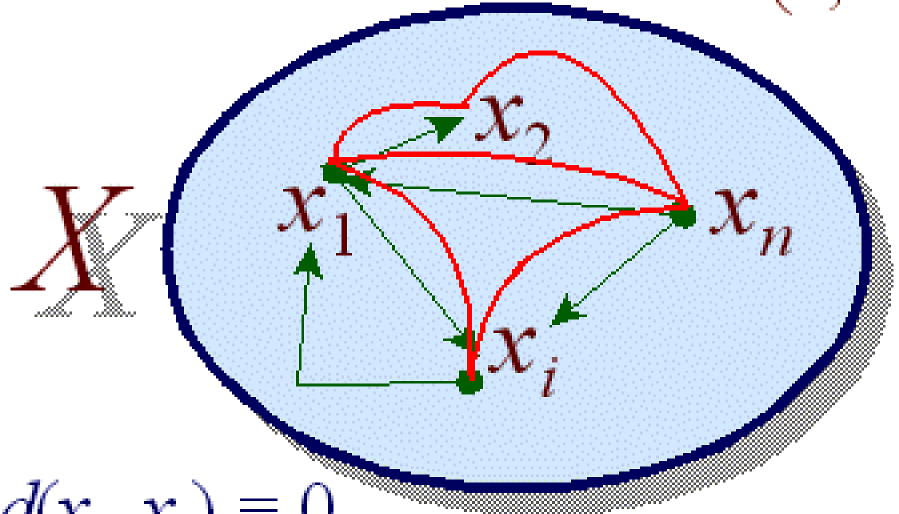
rocha@indiana.edu
http://informatics.indiana.edu/rocha

Cognitive Science Program
the biocomplexity institute
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)



$d(x_i, x_i) = 0$

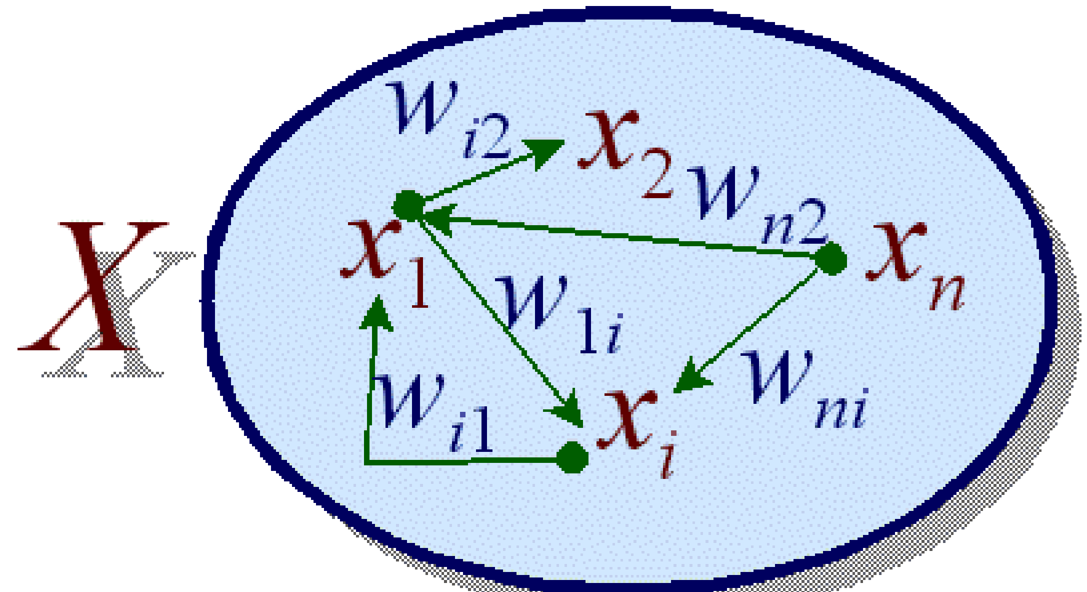
$d(x_i, x_j) = 1$, if there is an edge

$d(x_i, x_k) = d(x_i, x_j) + ... + d(x_l, x_k)$ 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

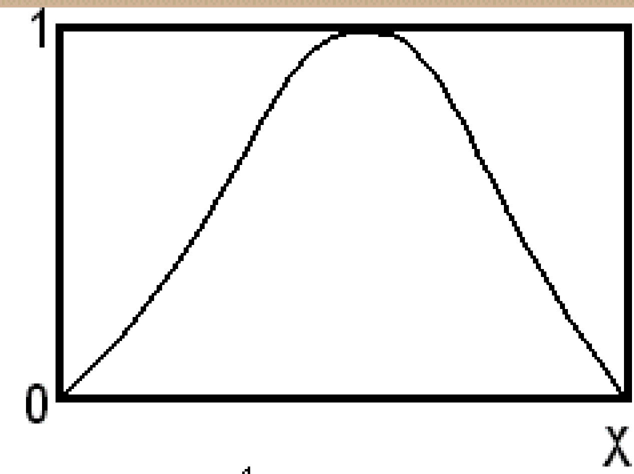
In graphs used to store “knowledge”, what does it mean?

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

INDIANA UNIVERSITY

operations

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



Fuzziness

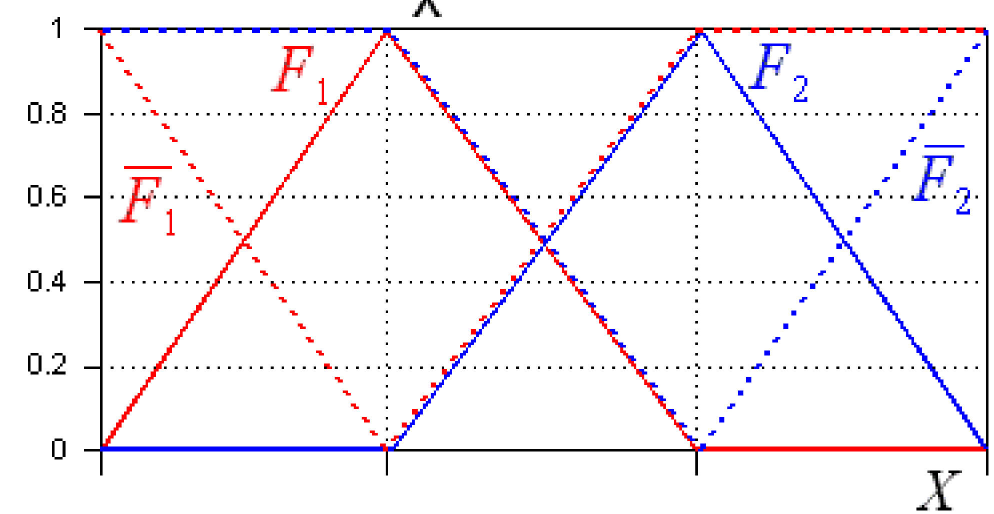
Degree of Membership/Truth

Standard Fuzzy Operations

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

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

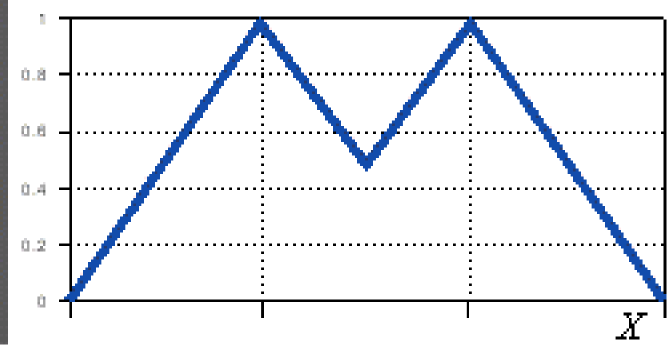
$$(A \cup B)(x) = \max[A(x), B(x)]$$



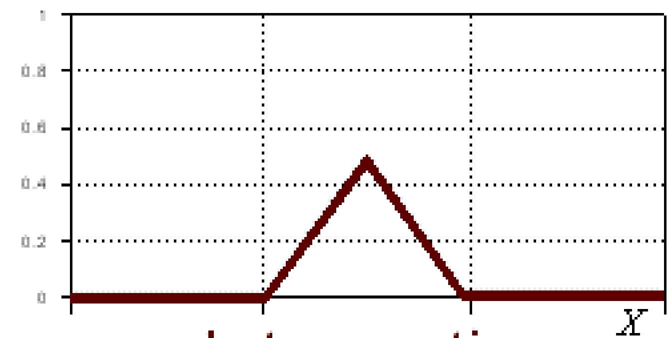
- Follows
  - Involution, commutativity, associativity, distributivity, Identity De Morgan's Laws, etc
- Does not Follow
  - Laws of contradiction and excluded middle

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

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



Union



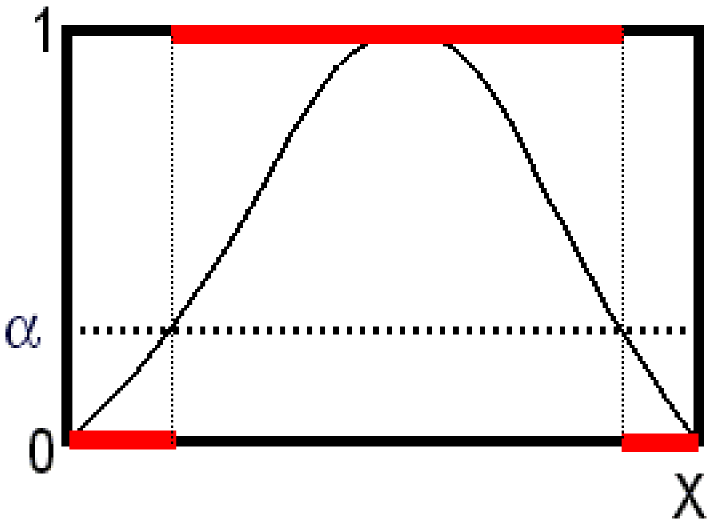
Intersection

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

α-cut

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

informatics
luis rocha 2005

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 that follow De Morgan's Laws plus
  - Complement
    - Boundary Conditions: $c(0)=1$ and $c(1)=0$
    - Monotonicity: if $a \le b$ then $c(a) \ge c(b)$
    - 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: $u(a,a) \ge a$

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

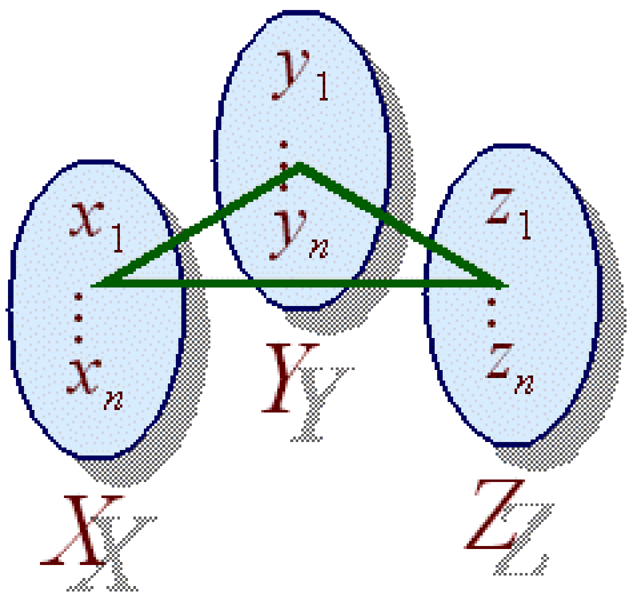
Sugeno Complement: $\lambda \in (-1,\infty)$

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

INDIANA UNIVERSITY

relations

Represent the presence or absence of association, interaction or interconnectedness between the elements of two or more sets.



$r(x_i, y_j, z_k) = 1$

- a relation R between sets $X_1, X_2, ..., X_n$ is a subset of the Cartesian product of these sets: $R(X_1, X_2, ..., X_n) \subseteq X_1 \times X_2 \times ... \times X_n$.
  - Traditional logical operations between sets can be used to modify relations

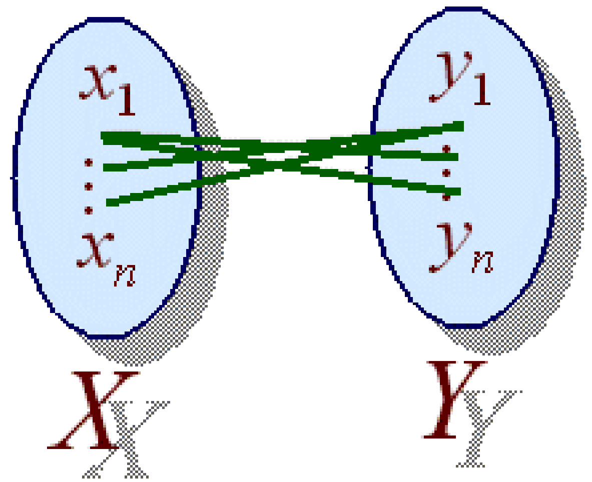
$$R(x_1, x_2, \ldots, x_n) = \begin{cases} 1 & \text{iff } (x_1, x_2, \ldots, x_n) \in R \\ 0 & \text{otherwise} \end{cases}$$

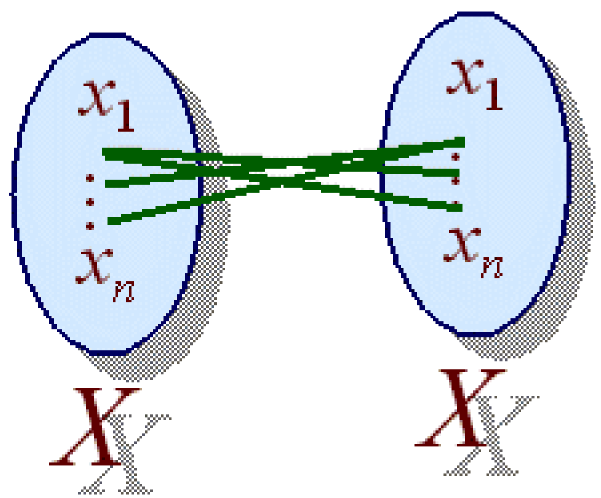
$$R(\mathbf{x}) \in [0,1], \quad \forall \mathbf{x} \in \mathbf{X}$$

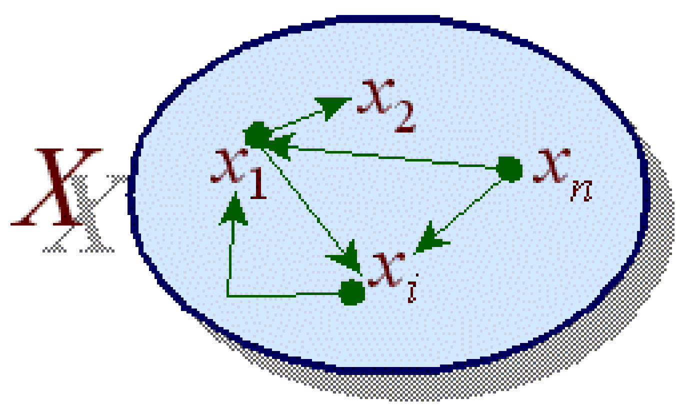
Fuzzy: Degree of Relation or association

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

binary relations







- Binary fuzzy relations are a generalization of real functions
  - Two or more elements of Y may relate to an element of X
  - Easily represented by matrices of dimension $n \times m$

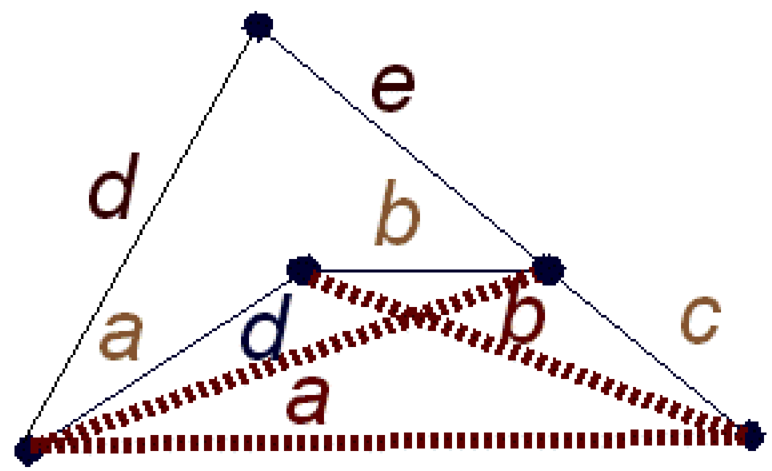
- Graphs are binary relations defined on a single set: $R(X, X)$.
  - Degrees of association between elements of the same set
  - If symmetric, R represents a non-directed graph

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

informatics
luis rocha 2005

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

composition of fuzzy graphs

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 *Min* ($\cap$) and *Max* ($\cup$) 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
    - 1. $R' = R \cup (R \circ R)$; 2. If $R' \neq R$, make $R = R'$ and go back to step 1; 3. Stop: $R_T = R'$

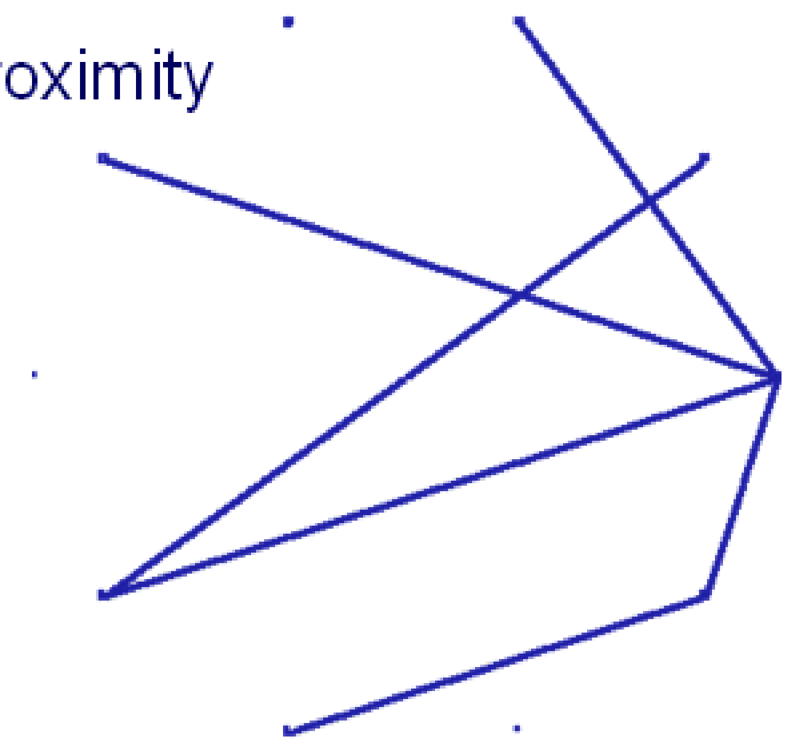
Generic Composition: $$R \circ R = \bigcup_k \bigcap (r_{ik}, r_{kj}) = r'_{ij}$$

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

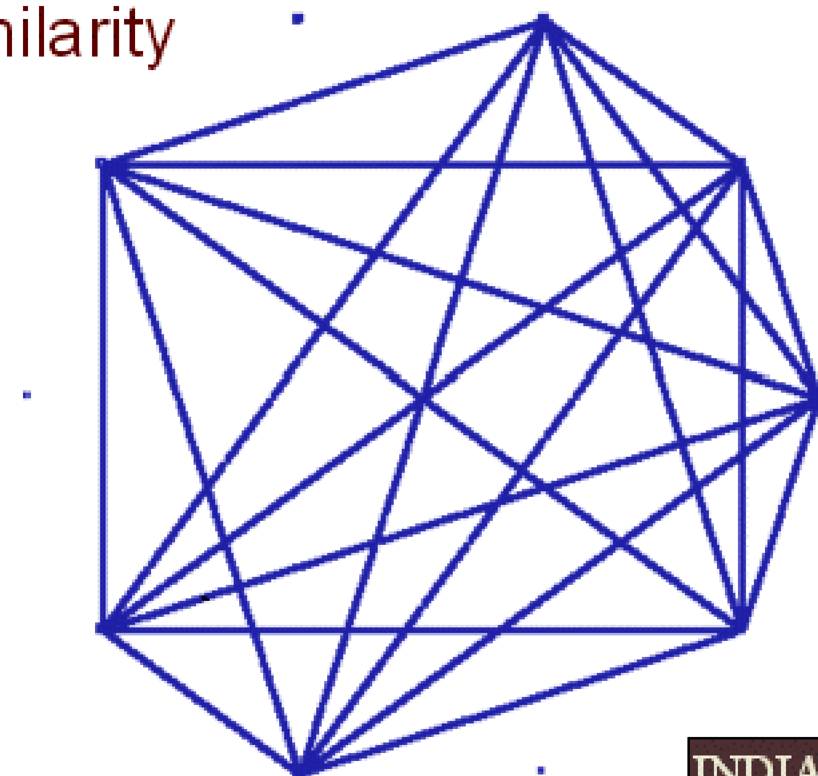
similarity and proximity relations

- **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.

Proximity



Similarity



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

informatics
luis rocha 2005

Active Recommendation Project

Luis M. Rocha, Andreas Rechtsteiner, Tiago Simas, Chien-Feng Huang, Judith Cohn, Eugene Gavrilov, Johan Bollen

http://arp.lanl.gov

Building Adaptive Webs that co-evolve with user communities

- Extraction of co-occurrence (associative) networks
  - Represent associative knowledge of information resources and users
- Identification of implicit associations in networks
  - Discovery of relevant items
  - Identify Communities of Users
- Conversation amongst information resources
  - driven by uncertainty reduction
  - Produce context-specific, proactive recommendations
- Collective Adaptation of network architecture
  - Evolving knowledge organization

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

INDIANA UNIVERSITY

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)
- Genes × MeSH Keyterms
  - Gene/keyterm Co-Occurrence

	X (Keywords)
Y (Documents)	R:X×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, 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.

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

proximity measures

produce associative (probabilistic) networks

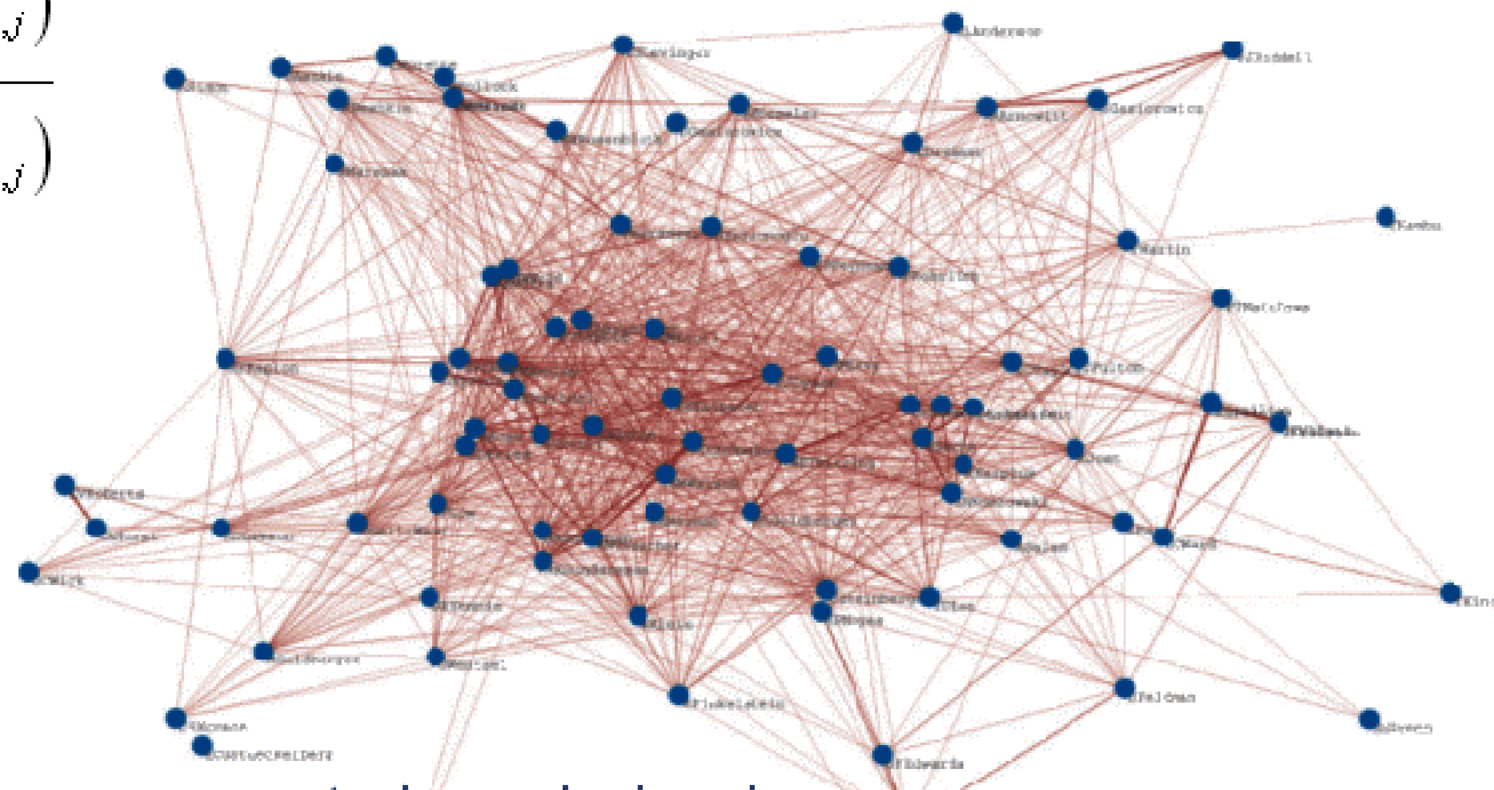
	X (Keywords)
Y (Documents)	$R: X \times Y$

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

	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)	$YXP: Y \times Y$



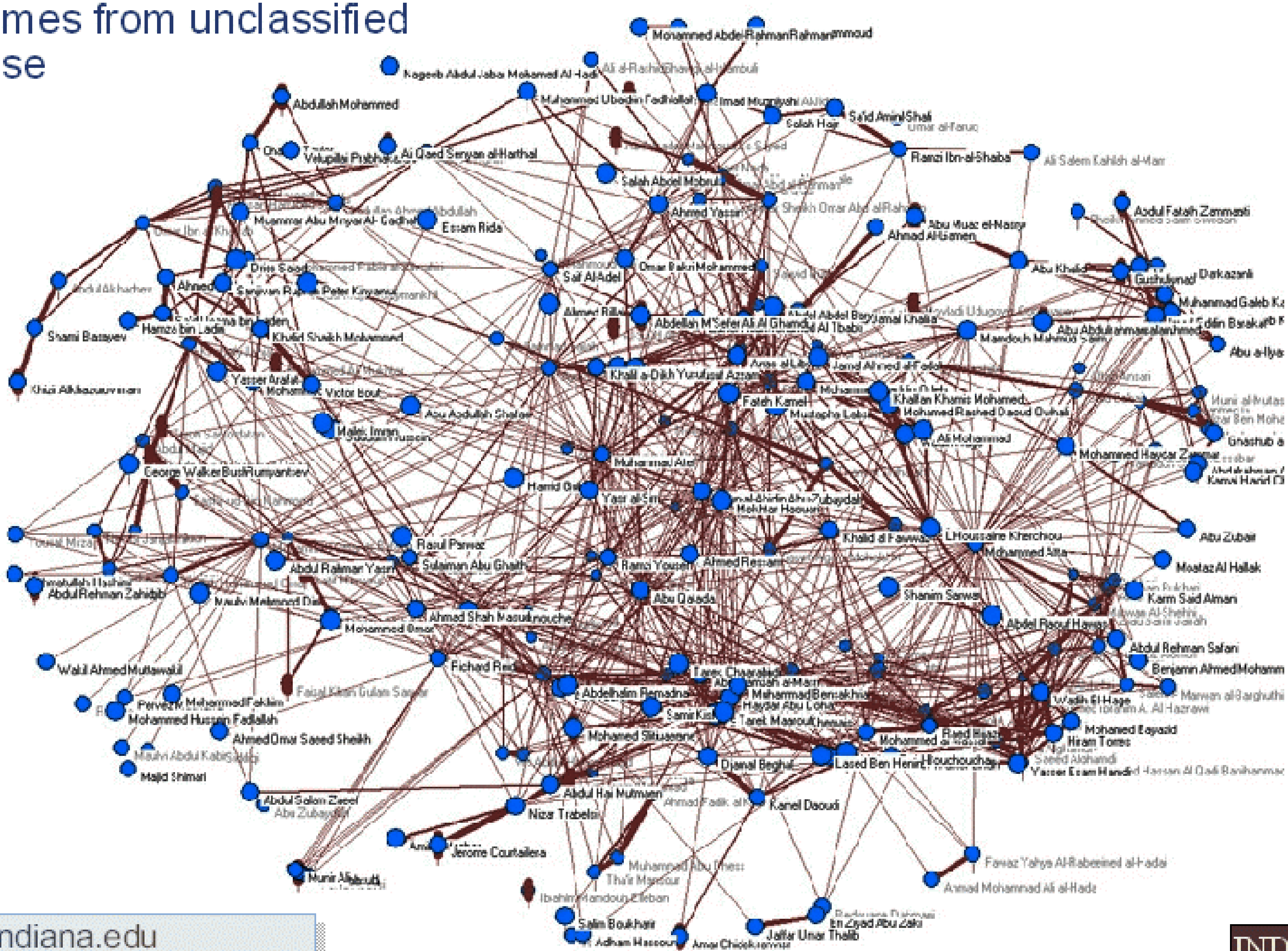
represents knowledge in associative manner

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

terrorist networks

PDP2

318 names from unclassified database



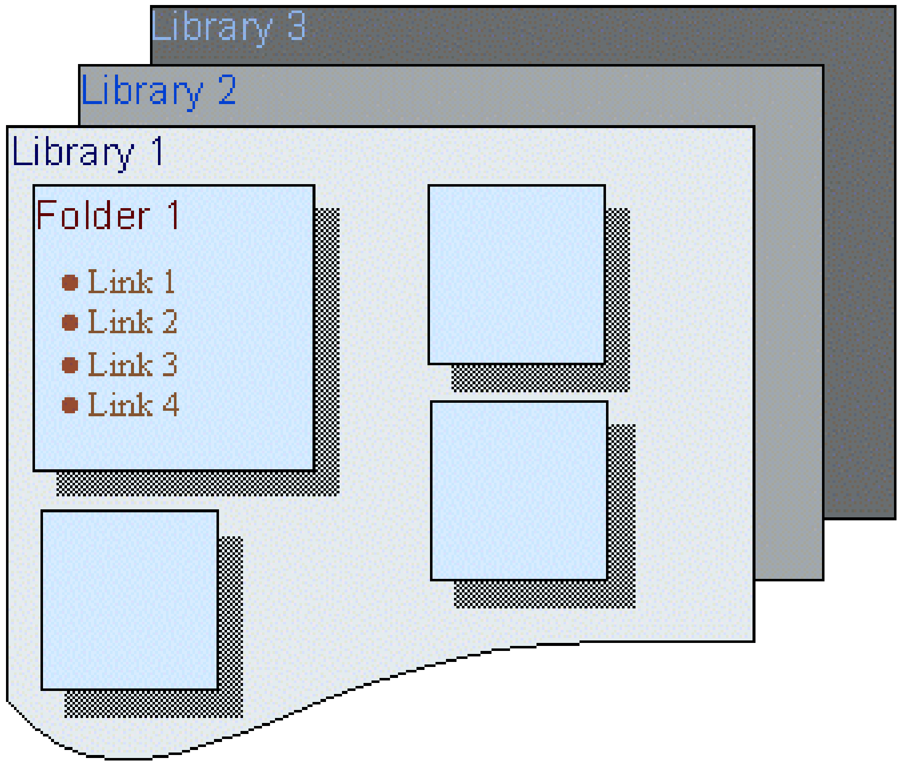
rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

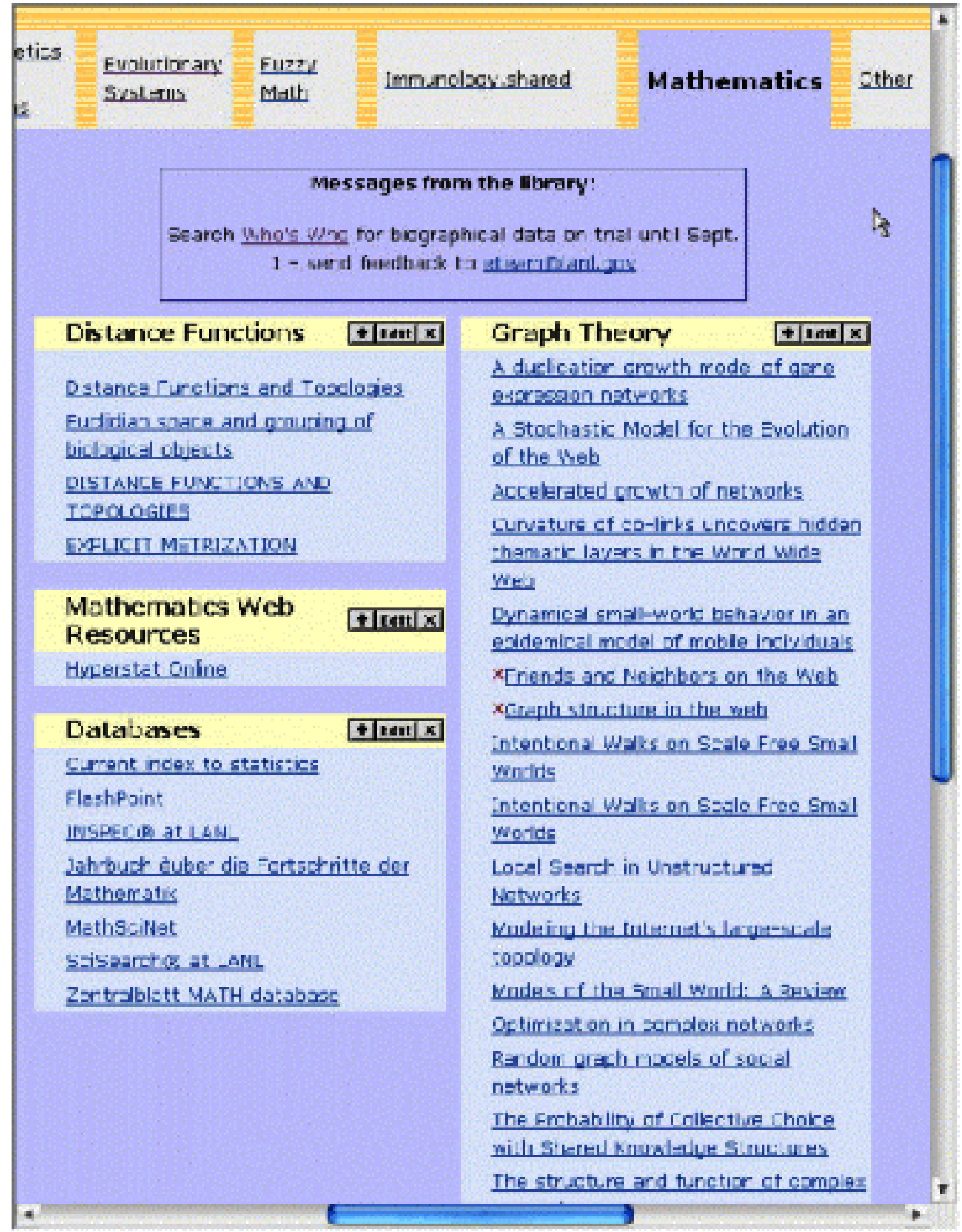
informatics
luis rocha 2005

mylibrary.lanl.gov

architecture



- 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.



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

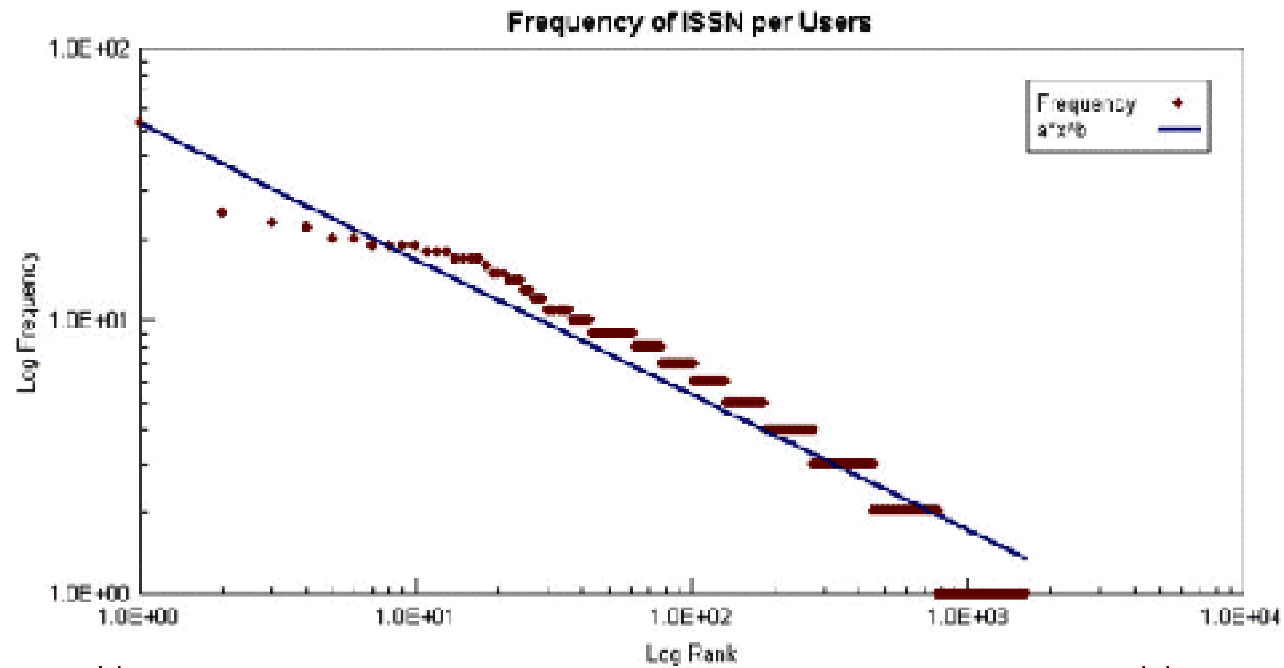
With Tiago Simas, Andreas Rechtsteiner and Chien-feng Huang

INDIANA UNIVERSITY

most frequent ISSN

occurences in 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

ISSN and personality proximity

from co-occurrence in mylibrary.lanl.gov

	ISSN
Personality	$A{:}P \times I$

326 personalities with at least one ISSN
253 users with at least one ISSN
623 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)

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

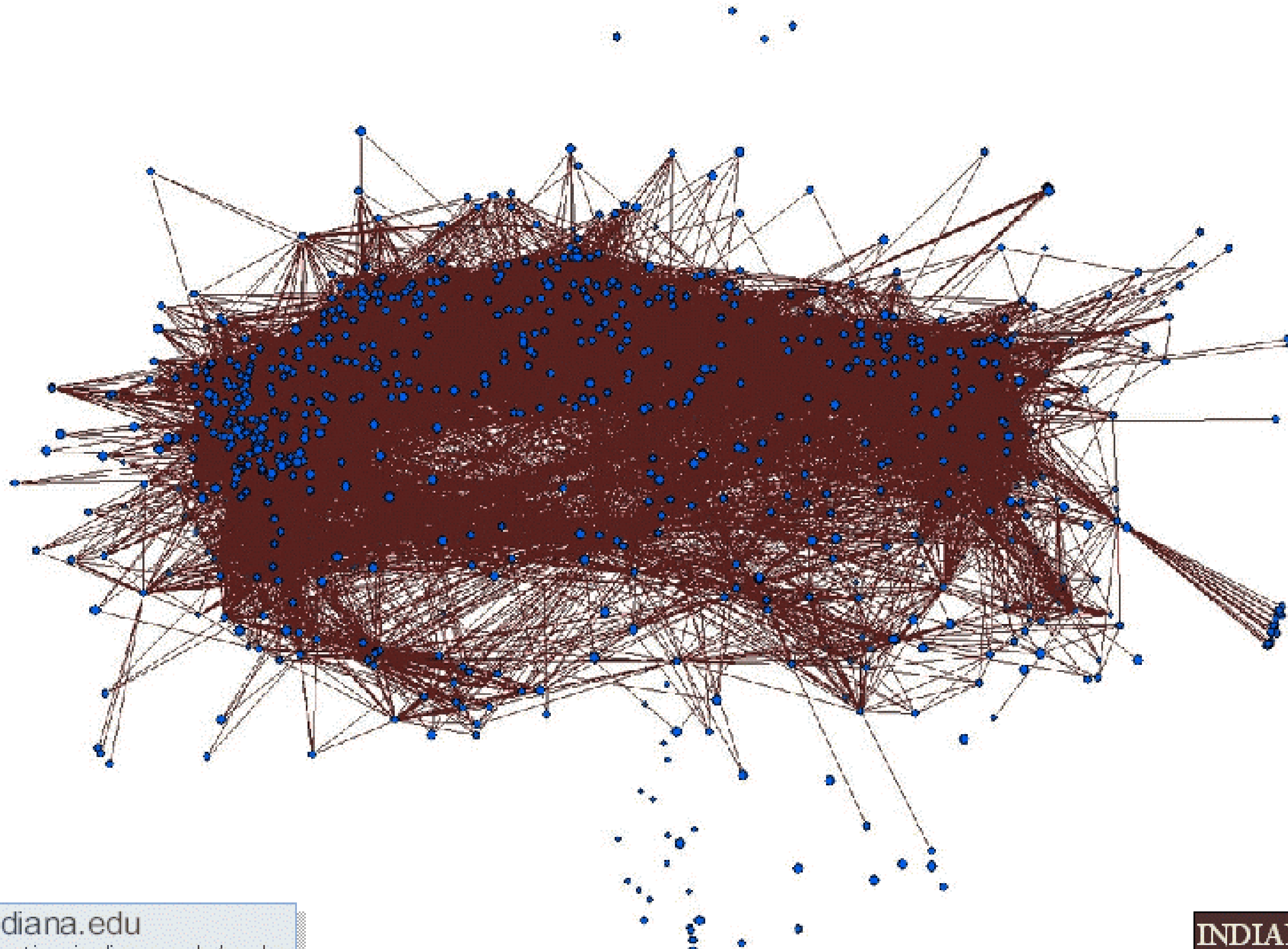
(ISSN Personality Proximity)

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

Rocha, Simas, Rechtsteiner and Gavrilov 2003.
LAUR 03-7357

journal network

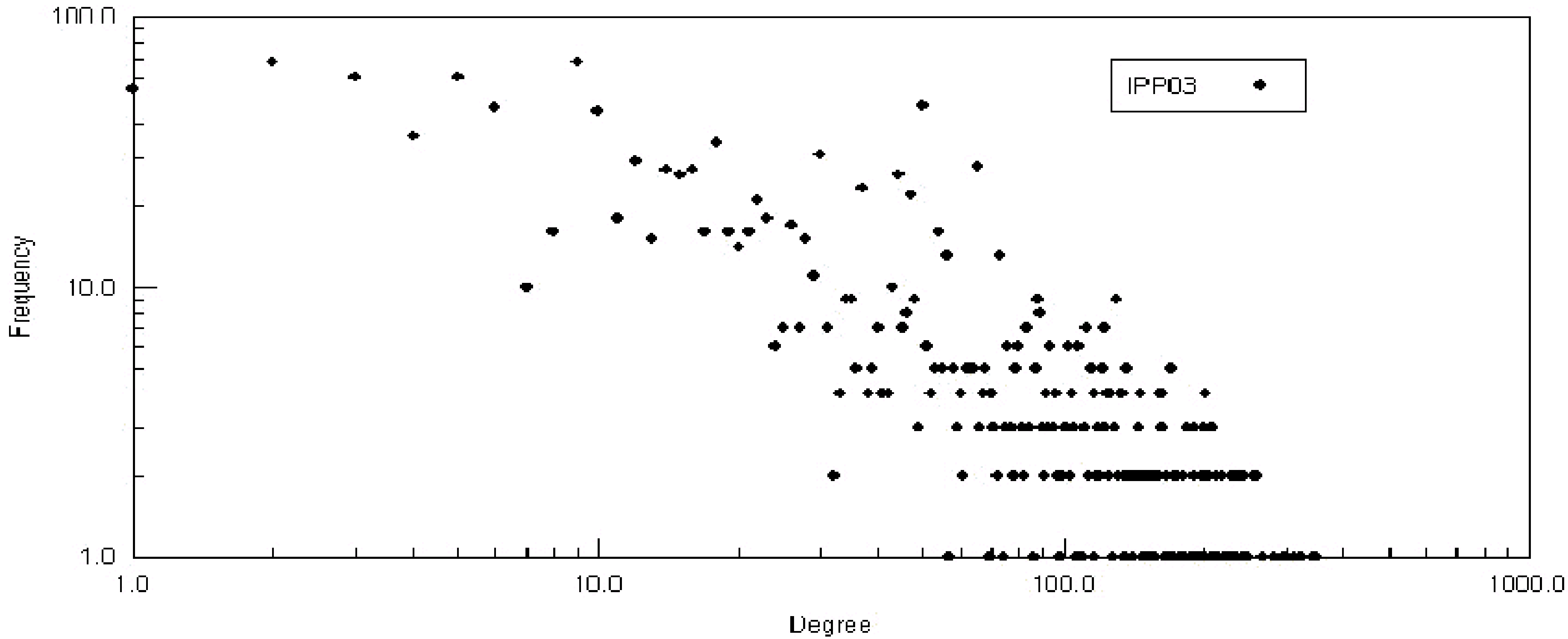
from co-occurrence in user personalities in mylibrary.lanl.gov: IPP



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

cumulative degree distribution

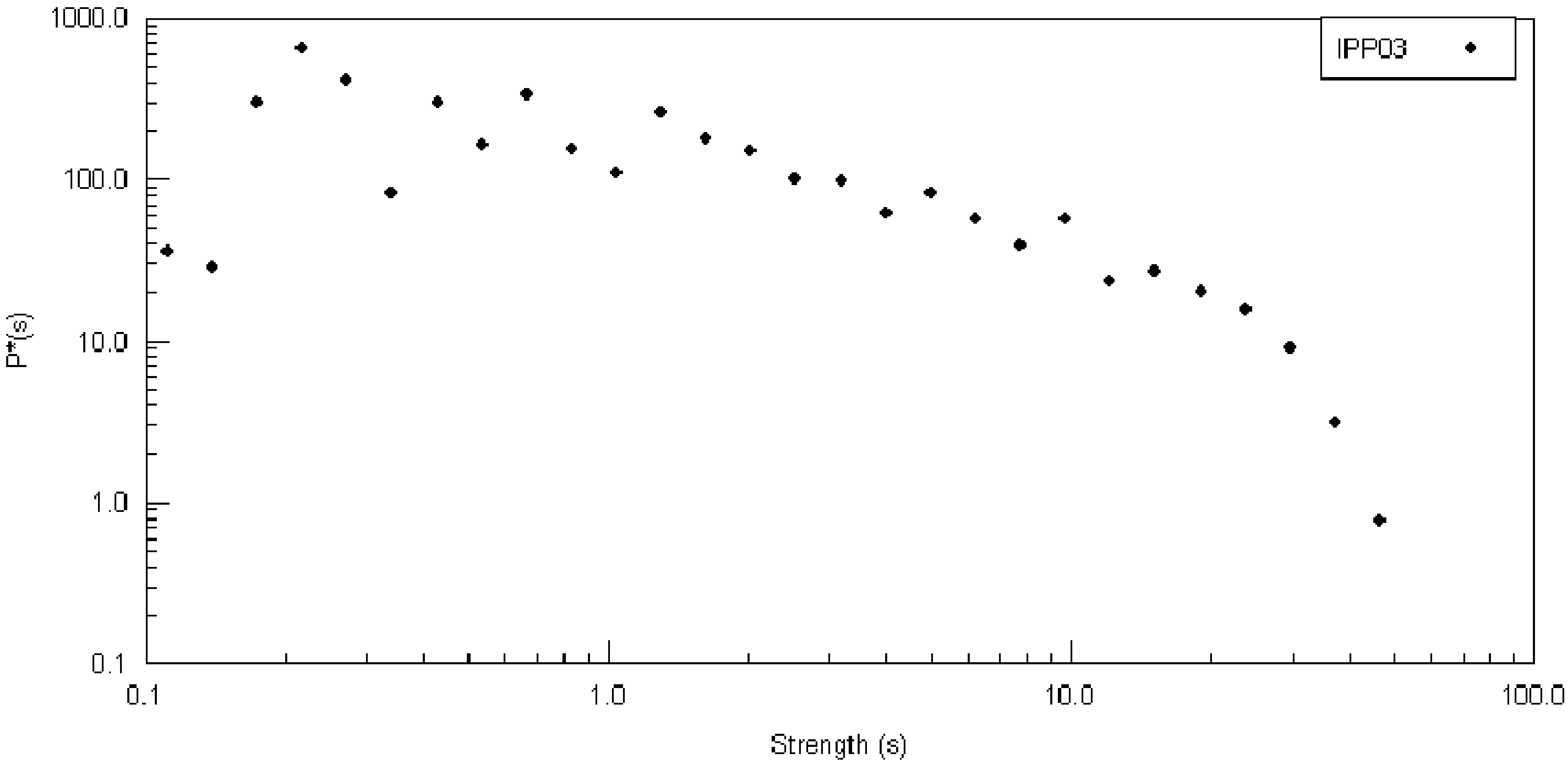
All weights



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

cumulative strength distribution (binned)

all weights

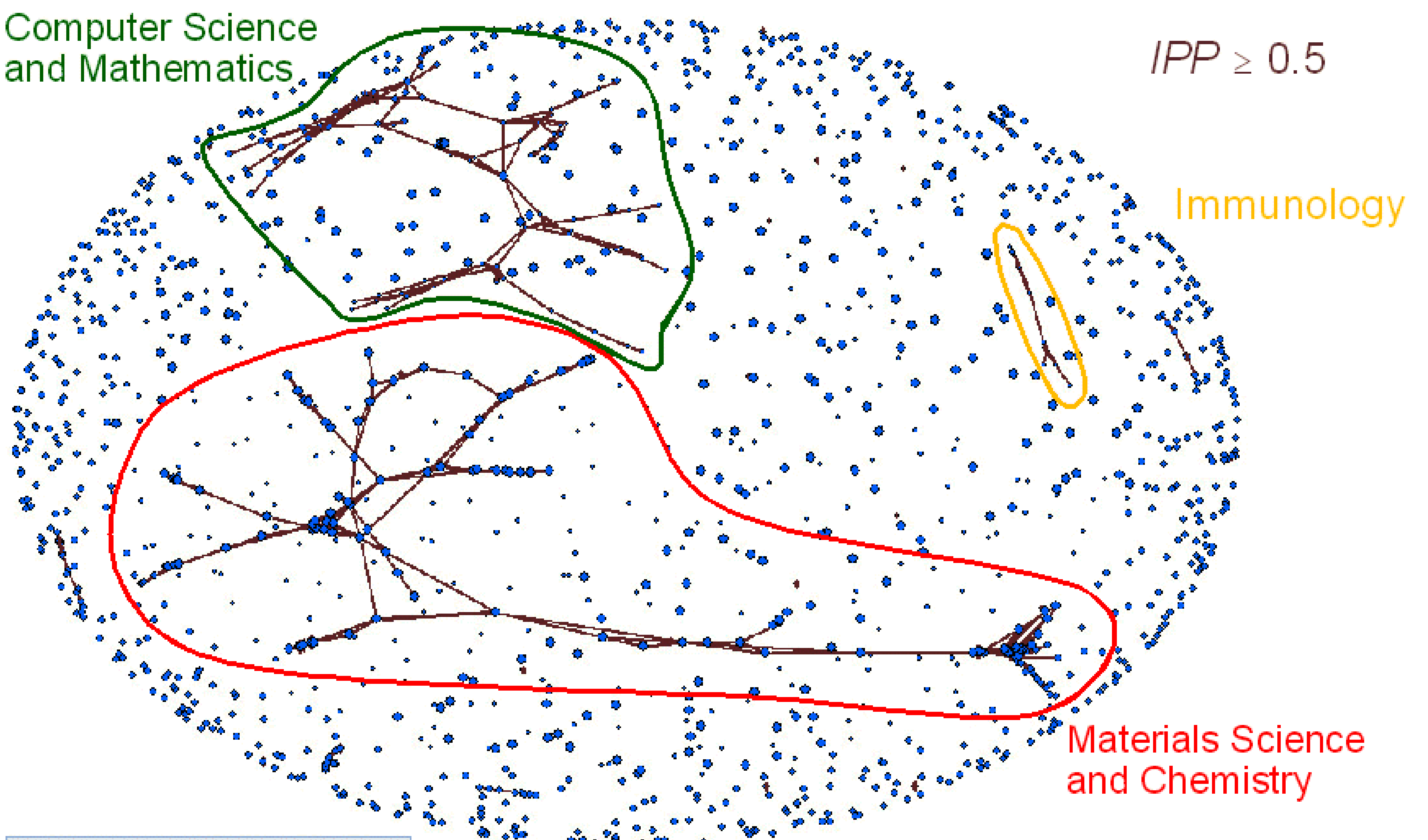


rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

journal network

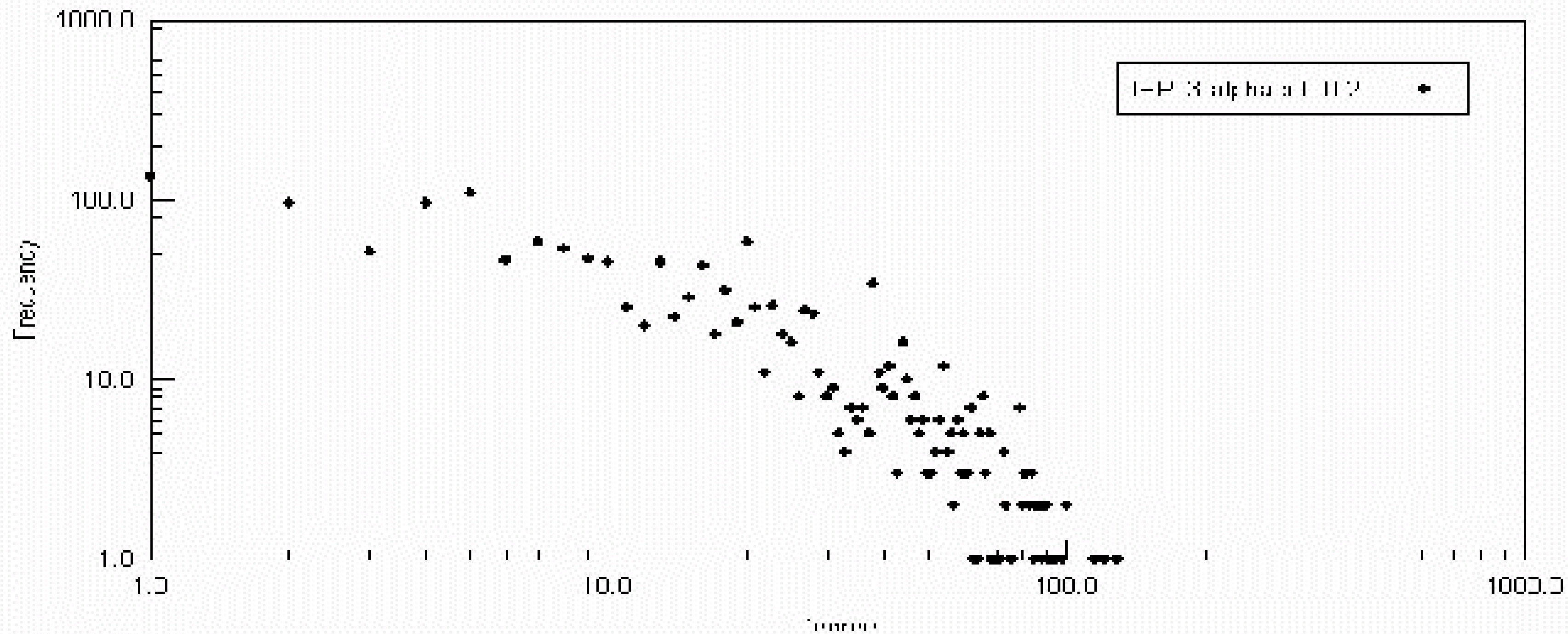
from co-occurrence in user personalities in mylibrary.lanl.gov: IPP



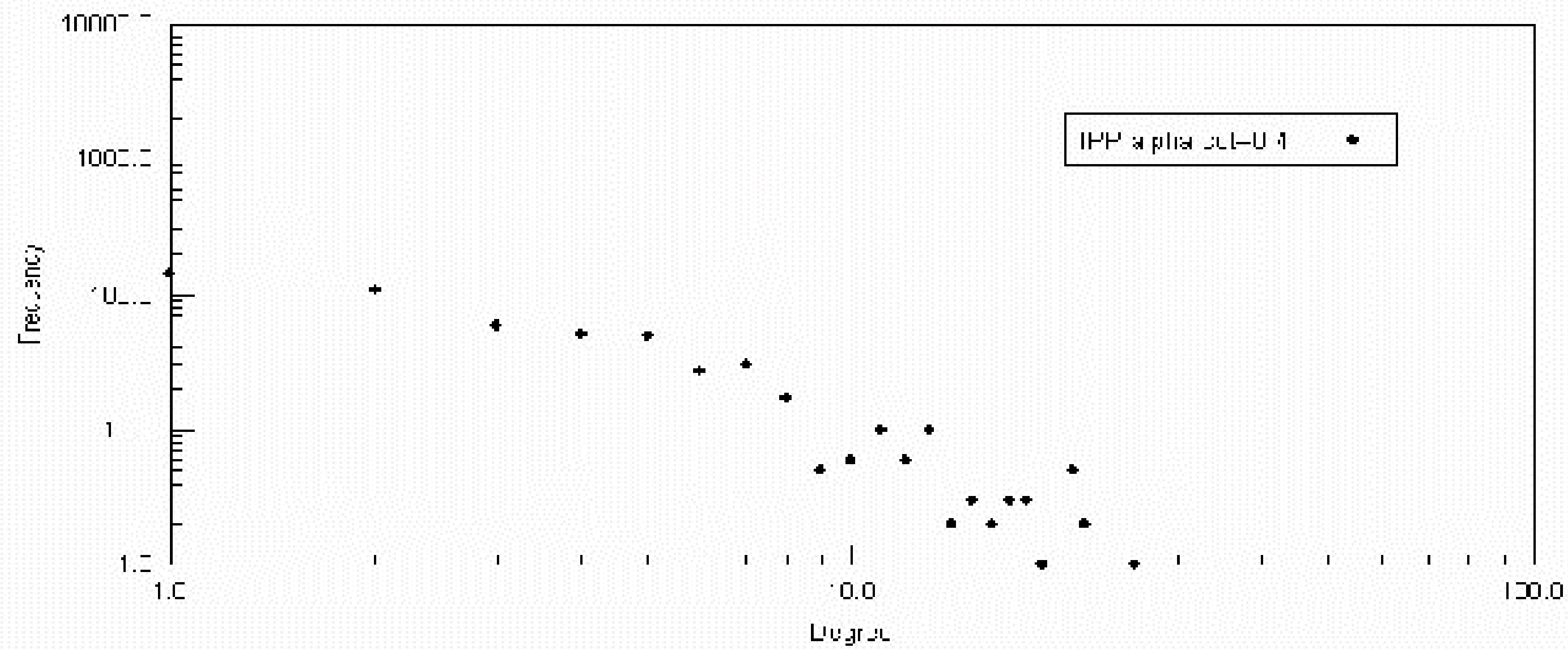
rocha@indiana.edu
http://informatics.indiana.edu/rocha

cumulative degree distribution

α-cut = 0.2



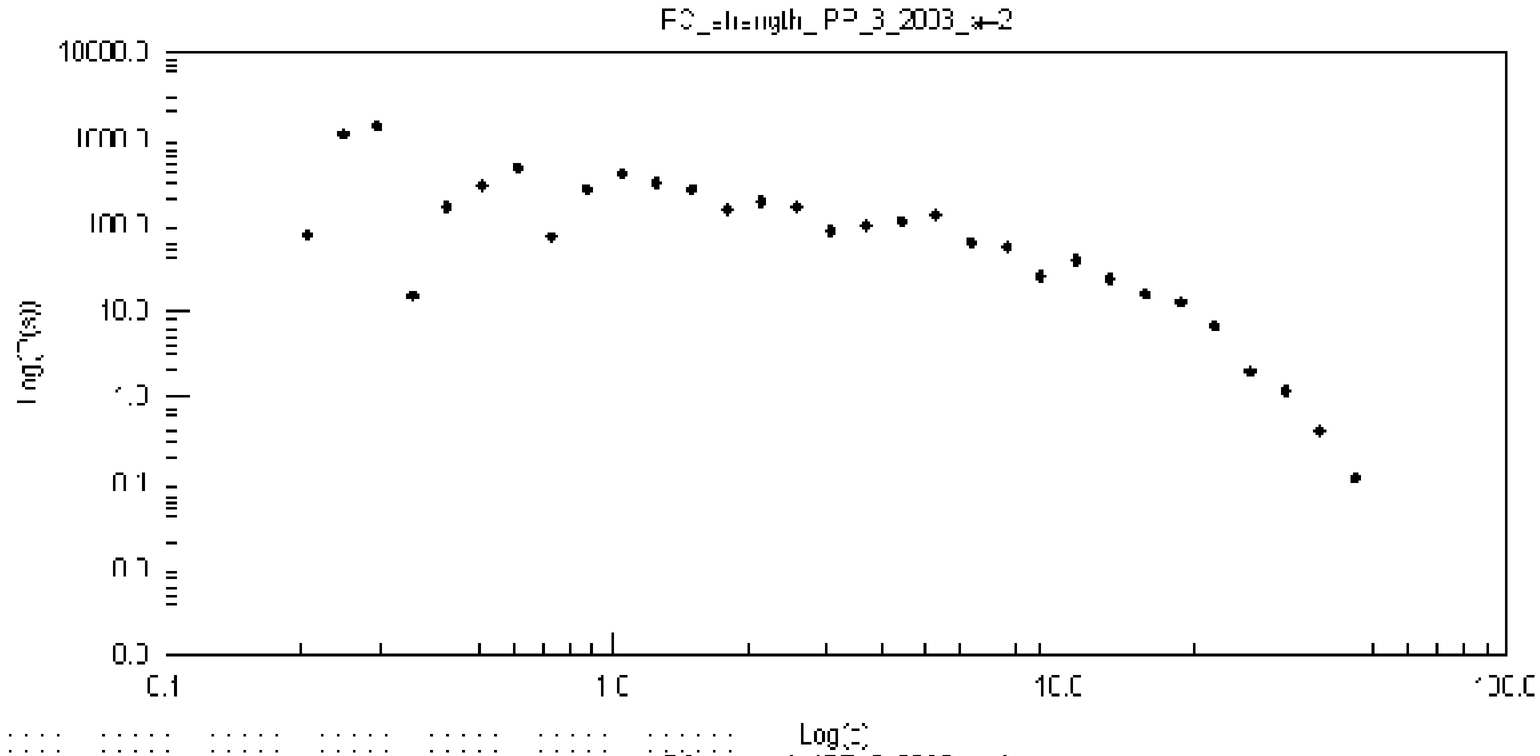
α-cut = 0.4



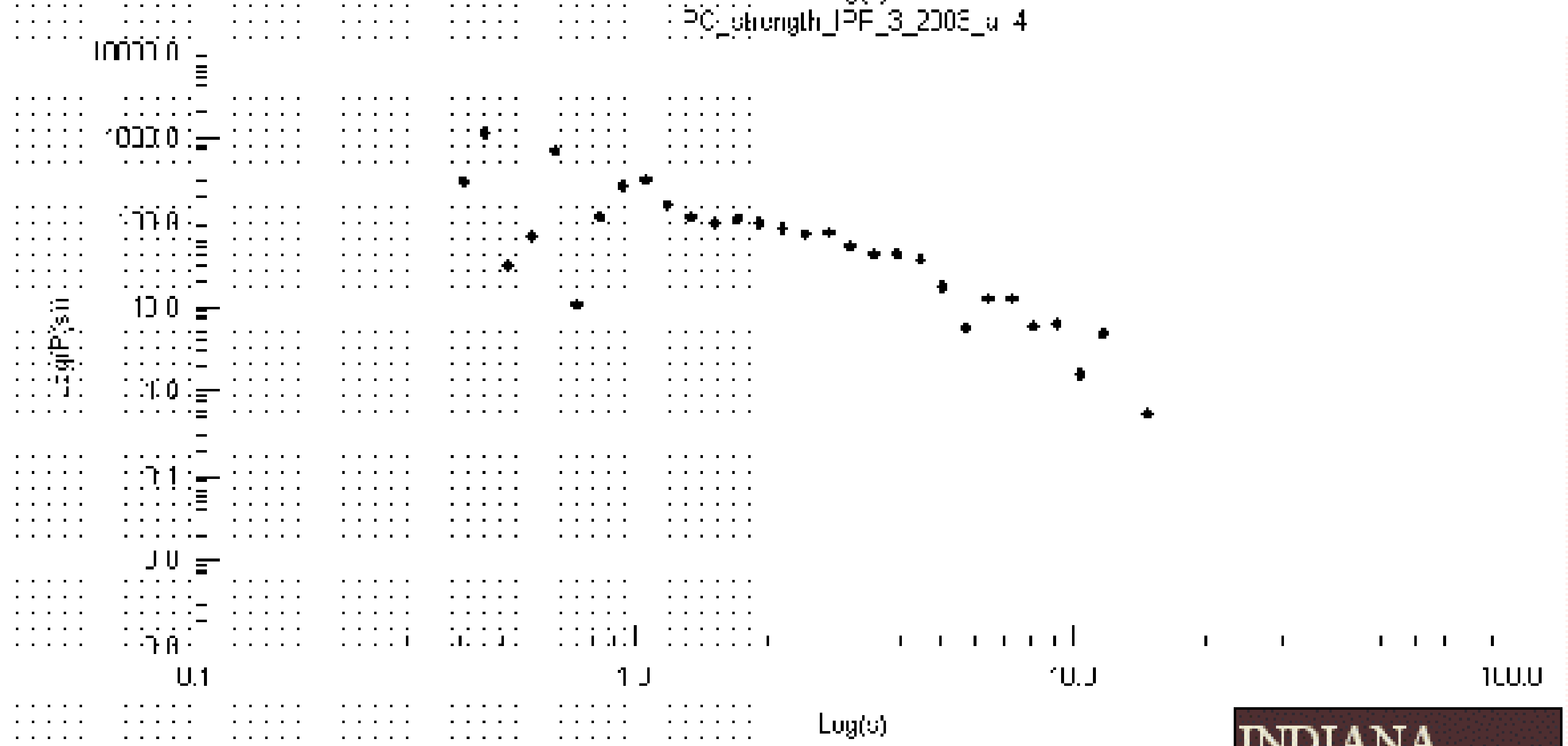
rocha@indiana.edu
http://informatics.indiana.edu/rocha

cumulative strength distribution (binned)

α-cut = 0.2



α-cut = 0.4



informatics
luis rocha 2005
rocha@indiana.edu
http://informatics.indiana.edu/rocha
INDIANA UNIVERSITY

informatics
luis rocha 2005

top 5 most frequent ISSN and their neighbors

$$ipp(i_s, i_t) = \frac{N_\cap(i_s, i_t)}{N_\cup(i_s, i_t)} \geq 3$$

$$ipp \geq 0.3$$

0031-9007--Physical review letters
- 1095-3787 --Physical review E: 0.4074
- 0556-2805--Physical Review B: 0.3621
- 0034-6861--Reviews of modern physics: 0.3333
- 0556-2791--Physical review A General physics: 0.3636

0556-2805--Physical Review B
- 0031-9007--Physical review letters: 0.3621
- 0370-1573--Physics reports: 0.3103
- 0921-4526--Physica B Condensed matter: 0.3462
- 0921-4534--Physica C Superconductivity: 0.3462
- 0034-6861--Reviews of modern physics: 0.3235
- 0556-2791--Physical review A General physics: 0.3333
- 1434-6036--European physical journal B: 0.3077

1095-3787--Physical review E
- 0031-9007--Physical review letters: 0.4074

0556-2791--Physical review A General physics
- 0031-9007--Physical review letters: 0.3636
- 0370-1573--Physics reports: 0.3077
- 0556-2805--Physical Review B: 0.3333
- 1089-490x --Physical review C Nuclear physics: 0.3913
- 0034-6861--Reviews of modern physics: 0.4643
- 1434-6036--European physical journal B: 0.3043

1089-5647--Journal of physical chemistry B
- 0002-7863--Journal of the American Chemical Society: 0.3000
- 0021-9606--Journal of Chemical Physics: 0.6250
- 1089-5639--Journal of physical chemistry A: 0.7619
- 0009-2614--Chemical physics letters: 0.6000
- 0301-0104--Chemical physics: 0.5714
- 0743-7463--Langmuir: 0.3810

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

informatics
luis rocha 2005

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

Bollen, Johan, Luis M. Rocha [2000]. "An Adaptive Systems Approach to the Implementation and Evaluation of Digital Library Recommendation Systems." In: Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000. *Lectures Notes in Computer Science*, Springer-Verlag, pp.356-359.

Rocha, Luis M. and Johan Bollen [2001]. "Biologically Motivated Distributed Designs for Adaptive Knowledge Management". In: *Design Principles for the Immune System and other Distributed Autonomous Systems*. L. Segel and I. Cohen (Eds.) Santa Fe Institute Series in the Sciences of Complexity. Oxford University Press, pp. 305-334.

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.

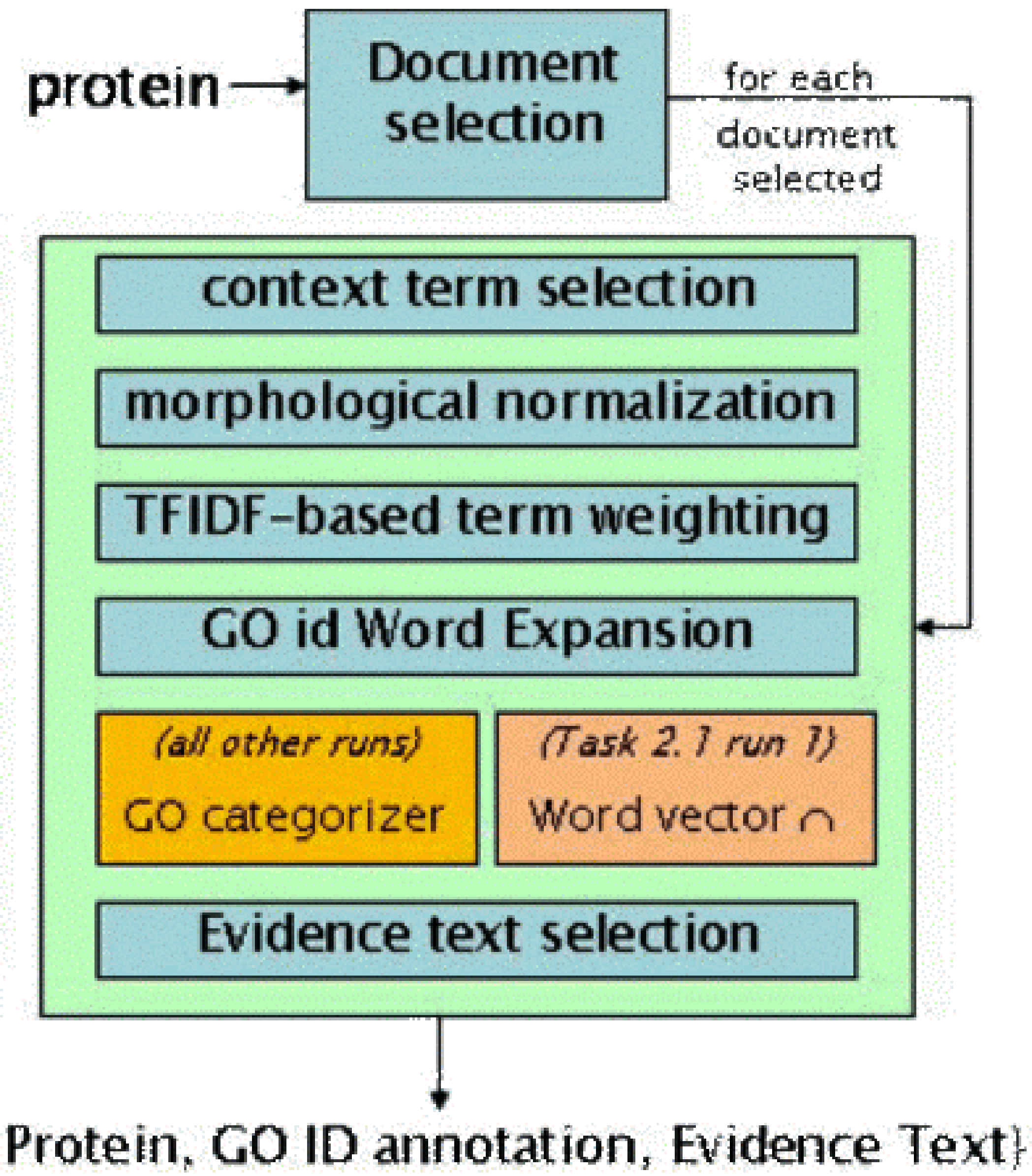
rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

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: same



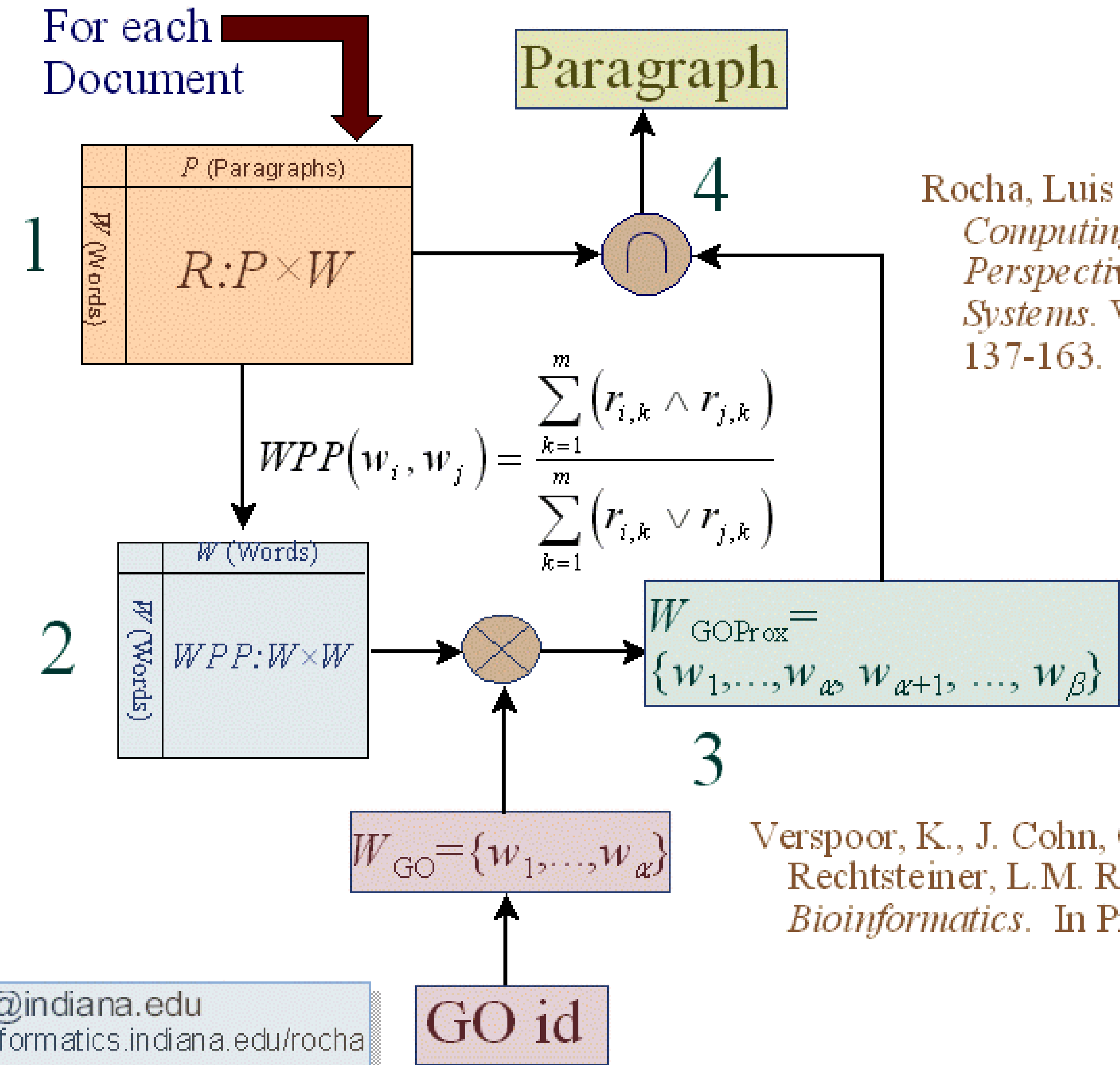
Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2004]. *BMC Bioinformatics*. In Press.

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

INDIANA UNIVERSITY

GO id word expansion

based on proximity measure



Rocha, Luis M. [2002]. In: *Soft Computing Agents: A New Perspective for Dynamic Information Systems*. V. Loia (Ed.) IOS Press, pp. 137-163.

Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2004]. *BMC Bioinformatics*. In Press.

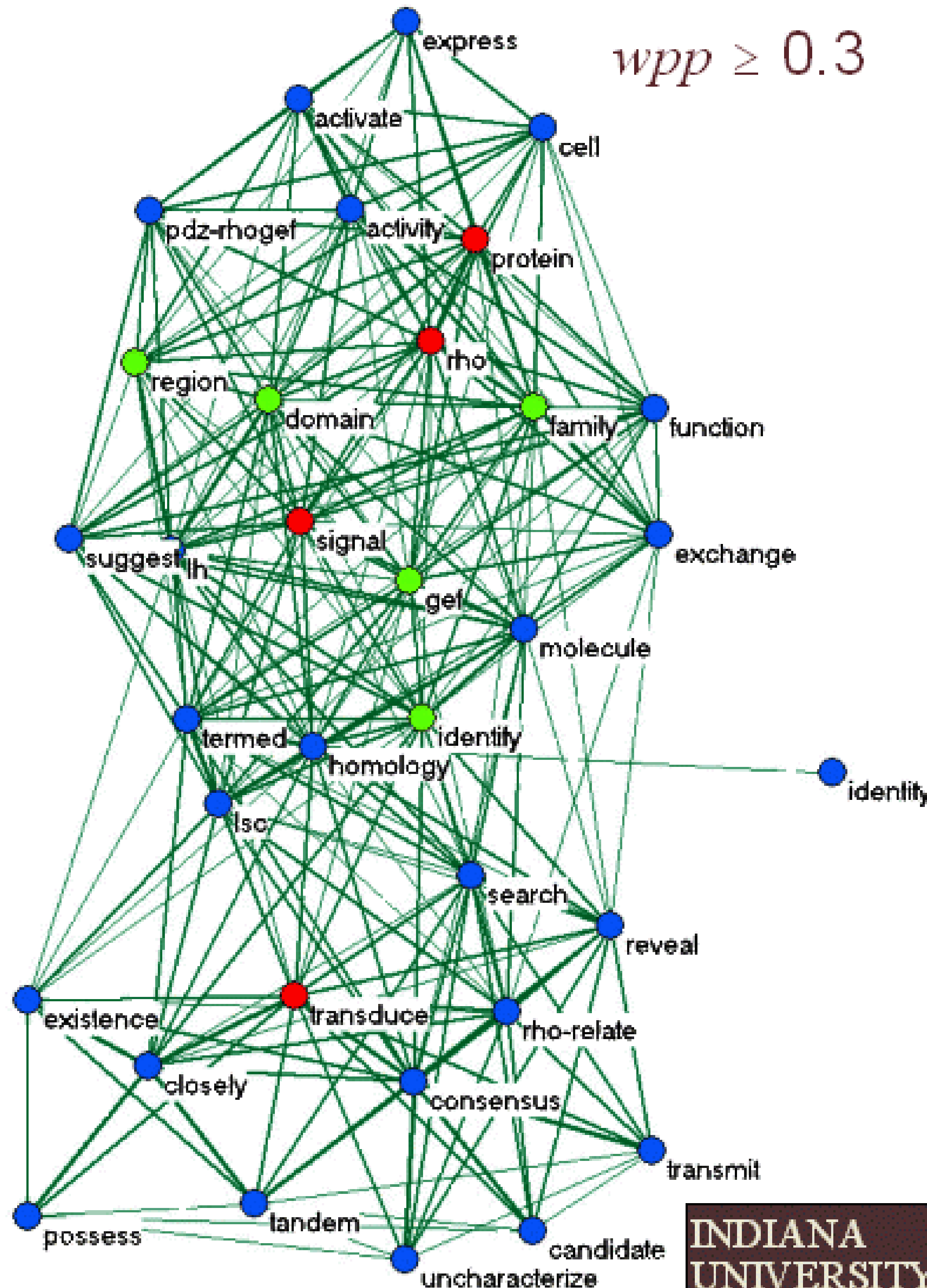
informatics
luis rocha 2005

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

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 [2004]. *BMC Bioinformatics*. In Press.

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%

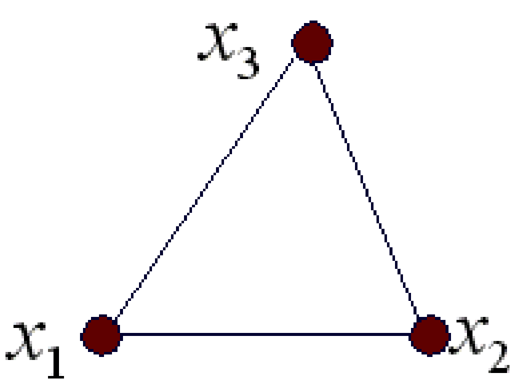
Verspoor, K., J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, T. Simas [2004]. *BMC Bioinformatics*. In Press.

informatics
luis rocha 2005
rocha@indiana.edu
http://informatics.indiana.edu/rocha
INDIANA UNIVERSITY

identification of implicit associations in networks

semi-metric behavior

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



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

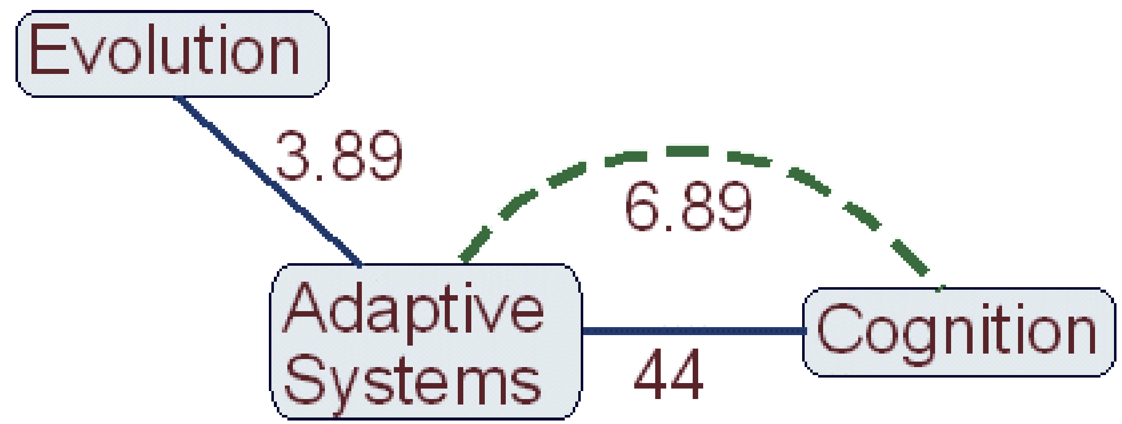
Distance from a Proximity Graph is semi-Metric
Distance from a Similarity Graph is Metric

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

Metric

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

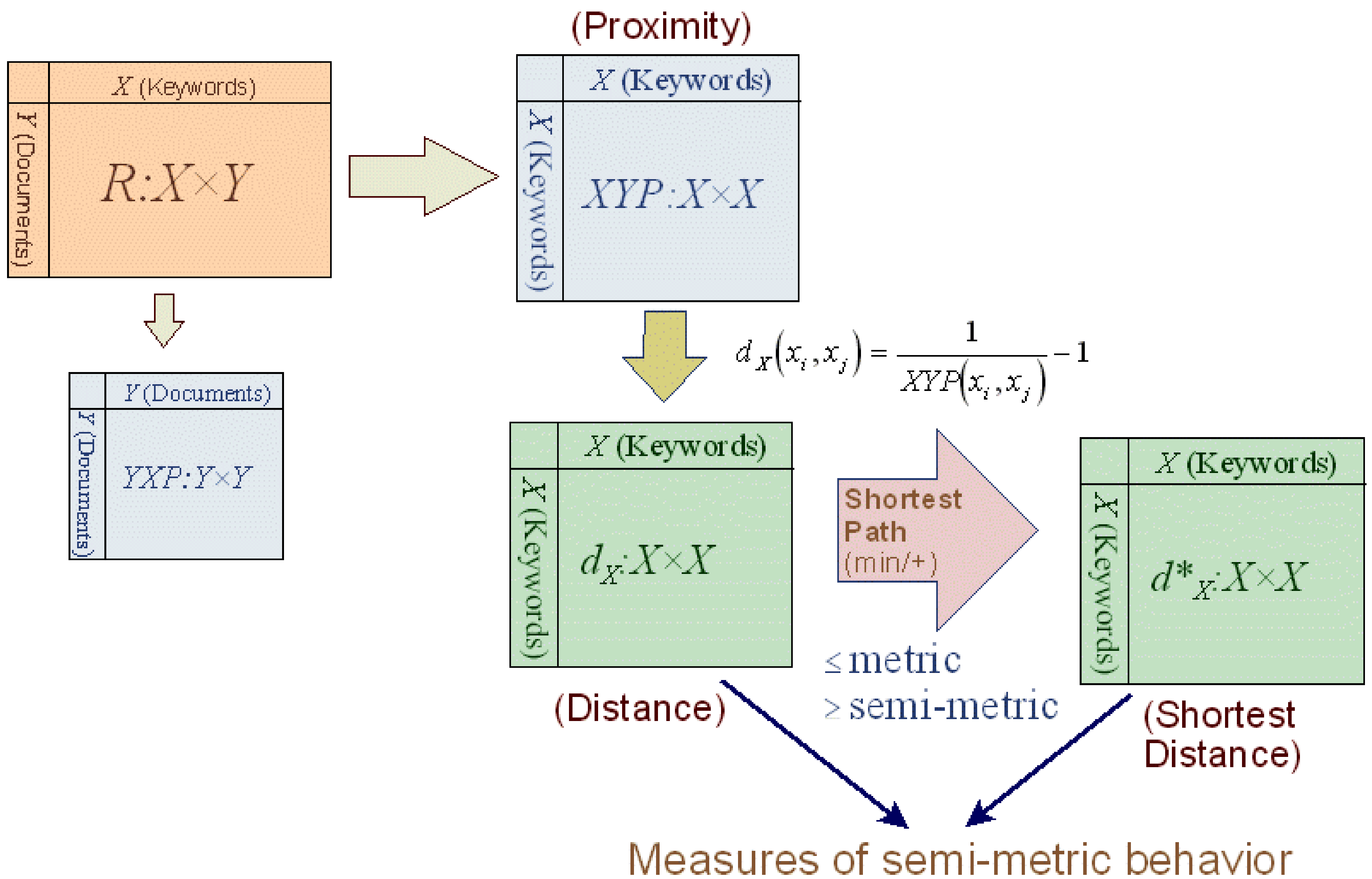
Semi-metric



Semi-metric ratio: 6.3861

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

computing semi-metric behavior



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

informatics
luis rocha 2005

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

identification of latent associations

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.
  - In recommendation system for journals now at LANL

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

informatics
luis rocha 2005

semi-metric recommendations

catching strong indirect associations in mylibrary.lanl.gov

IPP_3: parameter *rs*

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 *b*

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

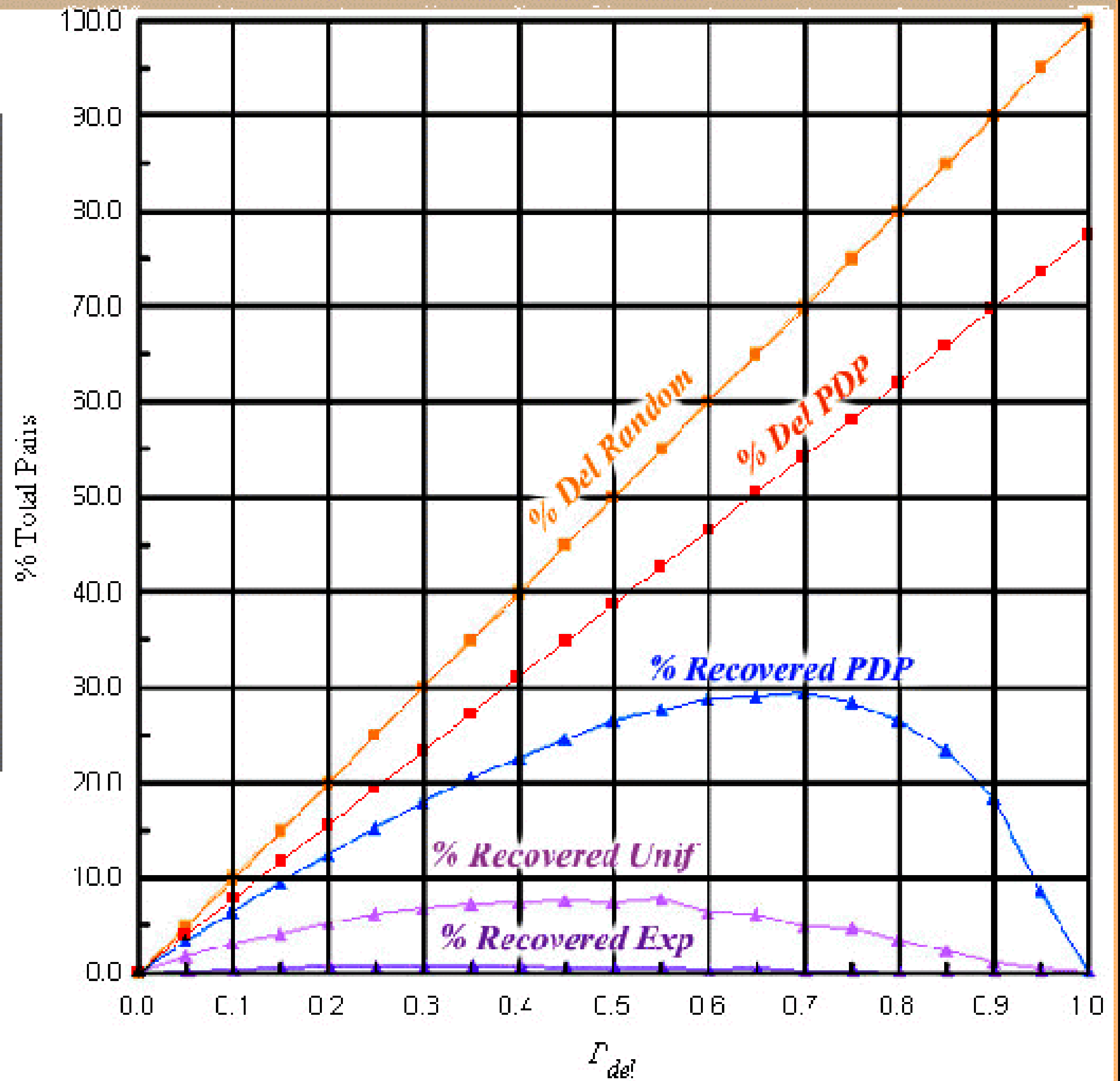
rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

detecting incomplete knowledge

random deletion experiments

- Perfect Knowledge
  - Transitive Closure of real graph
  - Metric Distance Graph
- Incomplete Knowledge
  - Each positive association is deleted with probability p_{del}
  - 100 graphs for each value of p_{del}



Full Deletion

Recovery via parameter *b*

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

Acknowledgement: John Hogden

INDIANA UNIVERSITY

shortest paths or weakest links

Edges: largest of the weakest links in all paths:

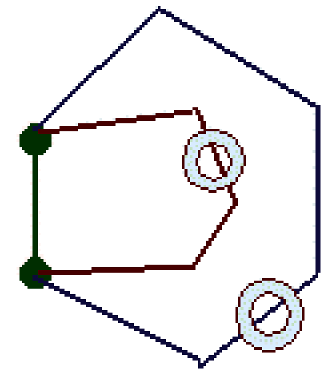
Proximity

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

Transitive Closure (max/min) →

Similarity

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



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

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

Distance

Shortest Path (min/+) →

$\leq$ metric

$>$ semi-metric

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

Edges: shortest path (sums all edges)

$\geq$

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

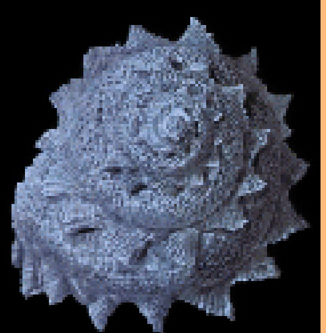
Edges: Smallest of the largest edges in each path

For any monotonic increasing distance function

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

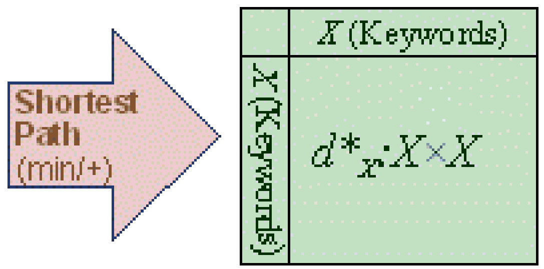
With Tiago Simas

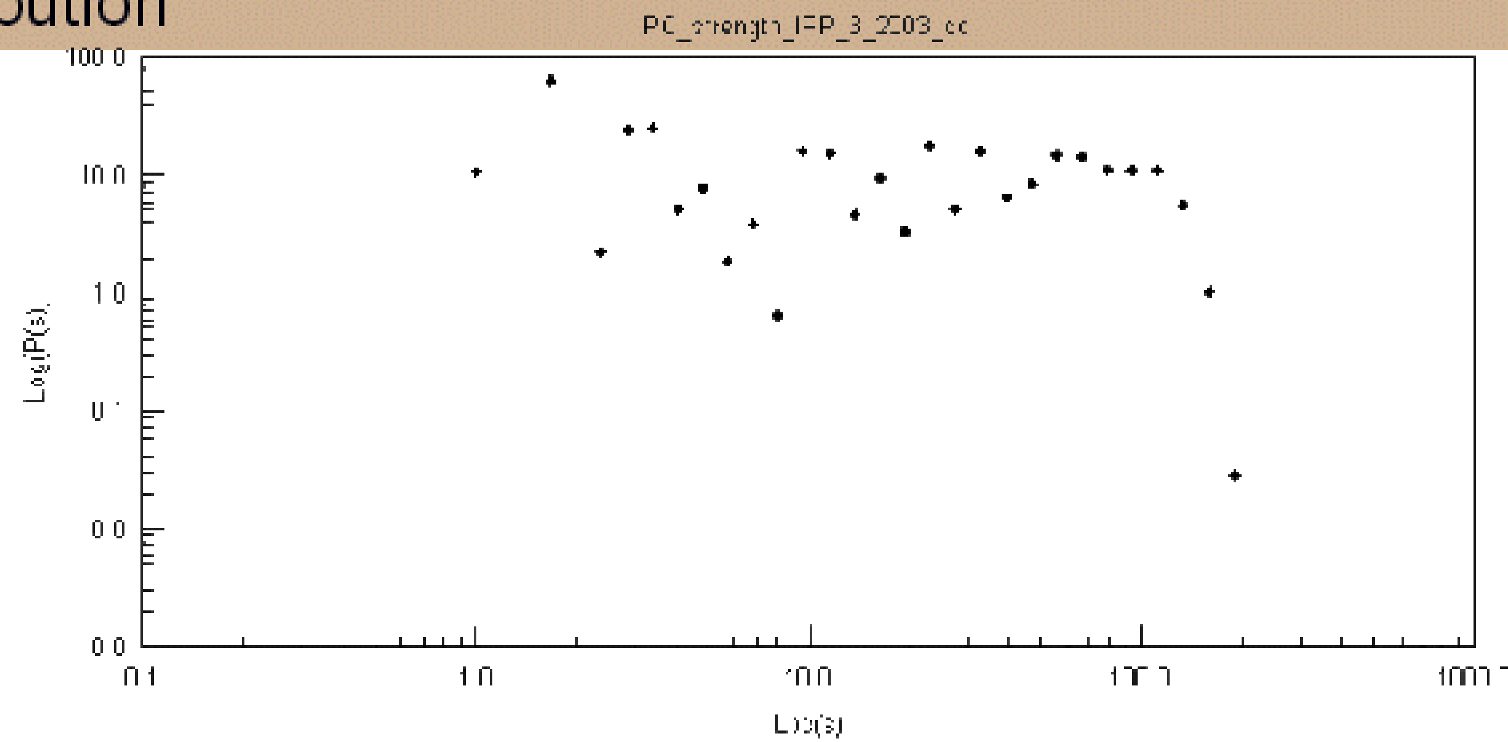
informatics
luis rocha 2005

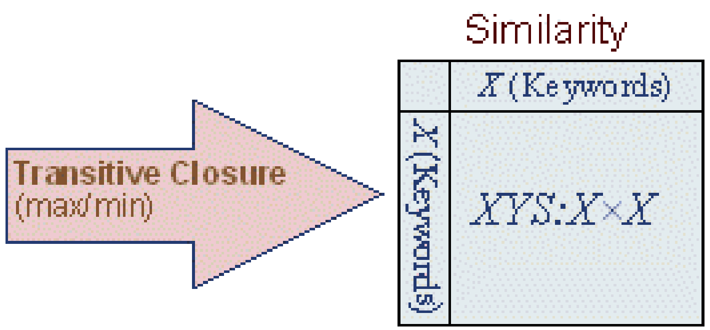


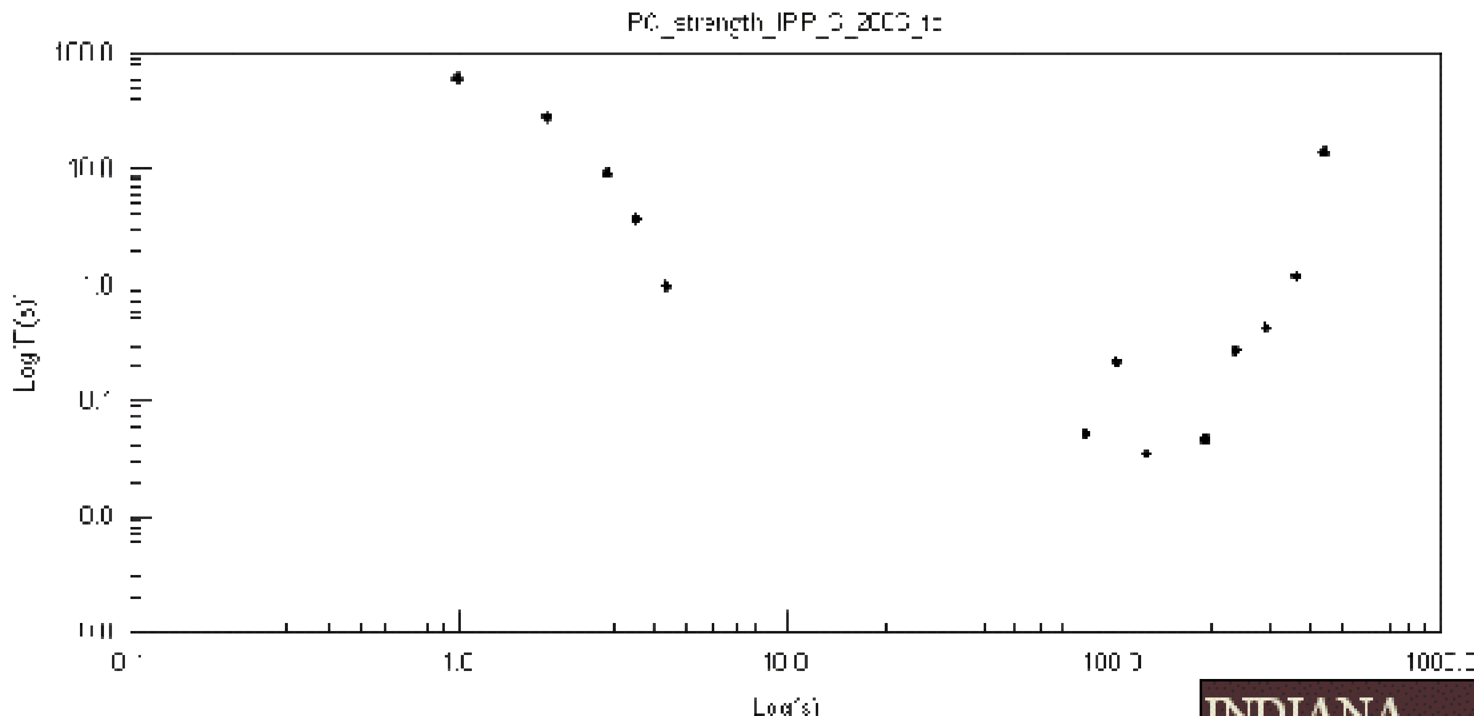
comparison of the two closures

cumulative strenght distribution









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

INDIANA UNIVERSITY

distance-equivalent fuzzy graph

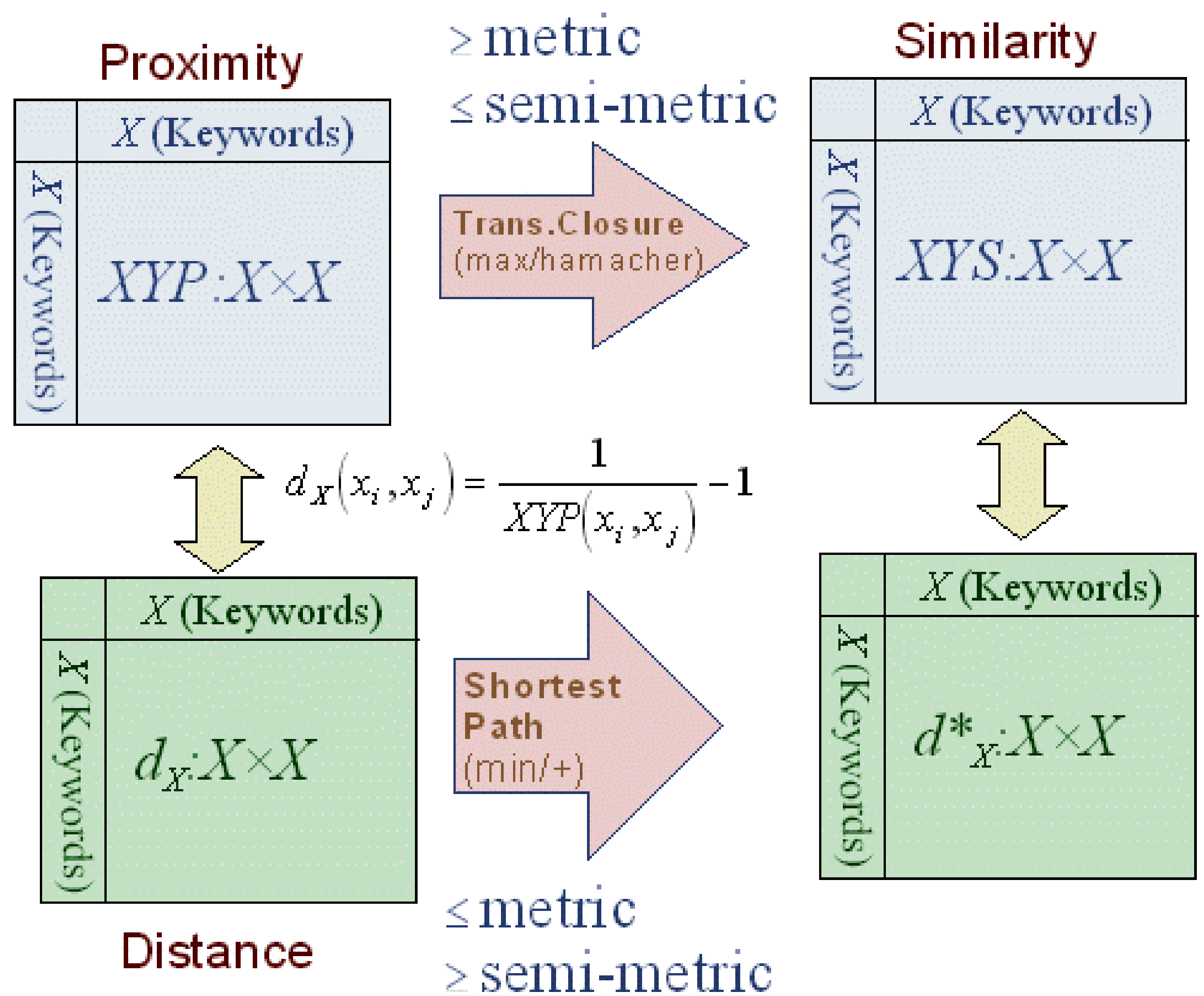
Hammacher function

If we use :

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

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

With Tiago Simas



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

distance and fuzzy graphs

Current work

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

Are not dual: no complement can satisfy de Morgan's laws!

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

- Can fuzzy graphs represent distance graphs?
- What fuzzy intersection/union comes closest to distance graphs?
- What captures the semi-metric behavior best?
  - Shortest paths on distance graphs?
  - Some transitive closure?

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

informatics
luis rocha 2005

scientific community working on feynman diagrams

as published in *Physical Review*, 1949-54

	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

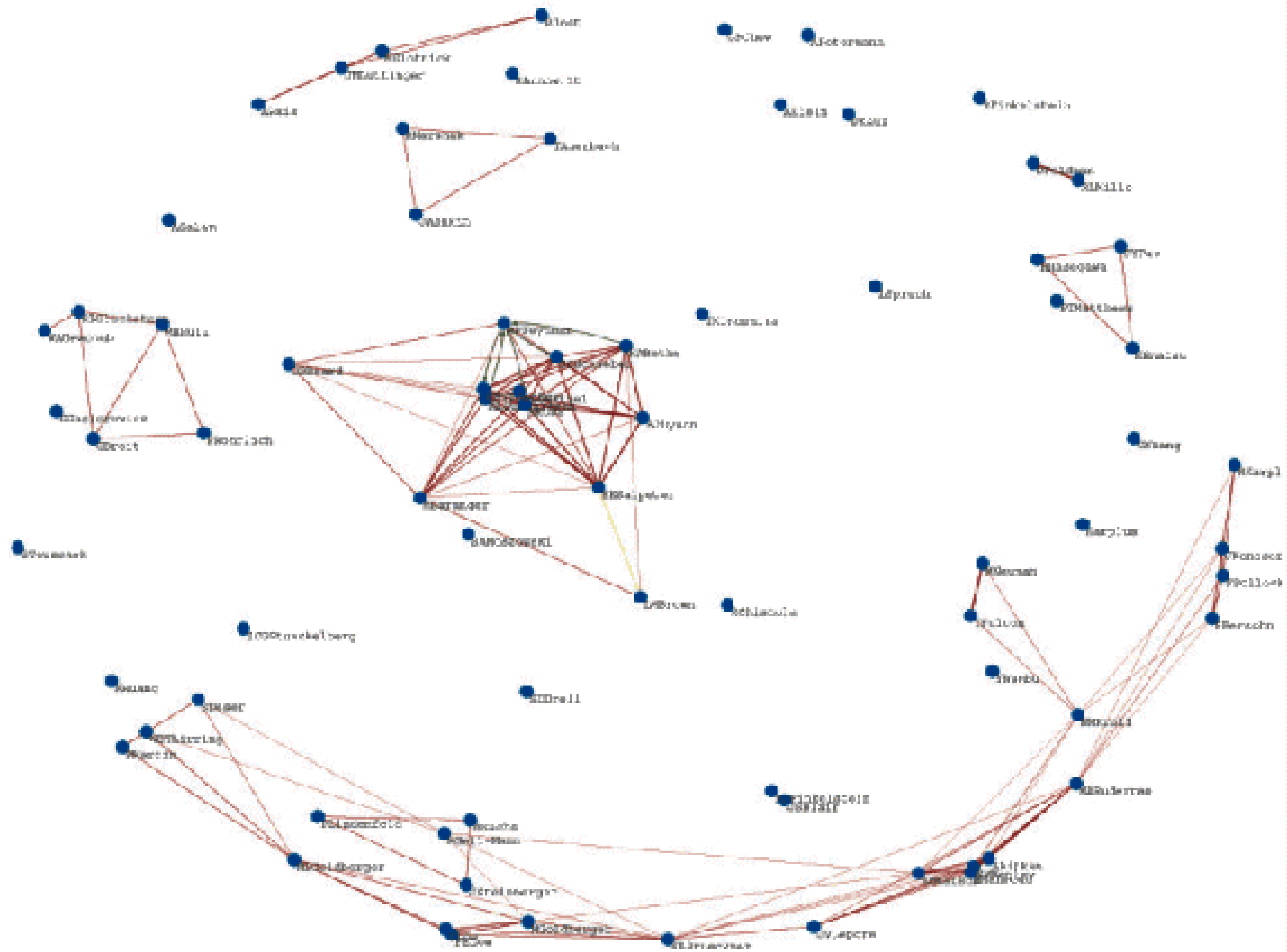
With Luis Bettencourt and David Kaiser

INDIANA UNIVERSITY

co-collaboration network

semi-metric analysis

- *CP* is almost metric
  - 139 papers, 76 authors
  - Percentage of pairs with positive semi-metric ratios (*rs* and *s* parameters): 0.667%
  - Percentage of pairs with indirect distances smaller than the average distance of direct edges to either node (*b* parameter): 0.439%
  - Very few implicit associations



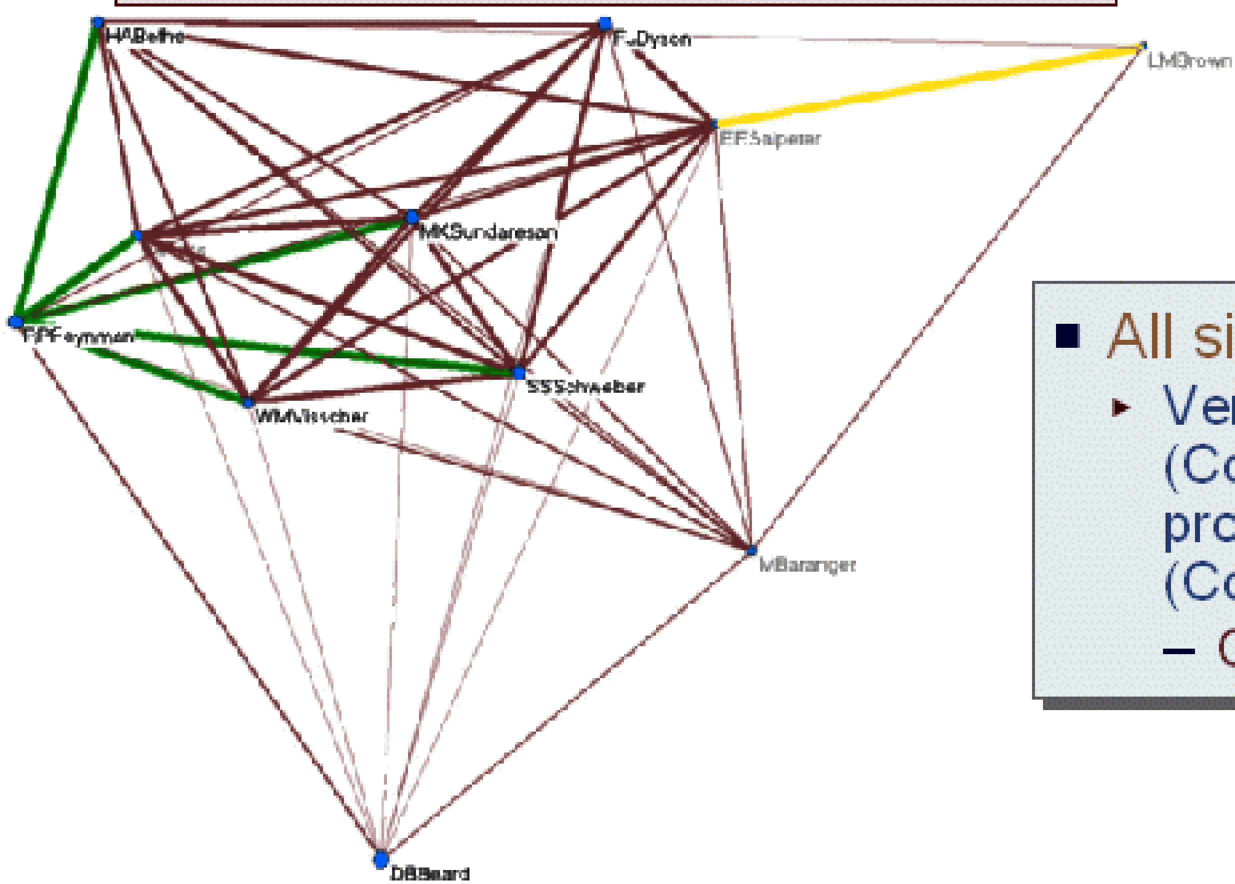
rocha@indiana.edu
http://informatics.indiana.edu/rocha

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



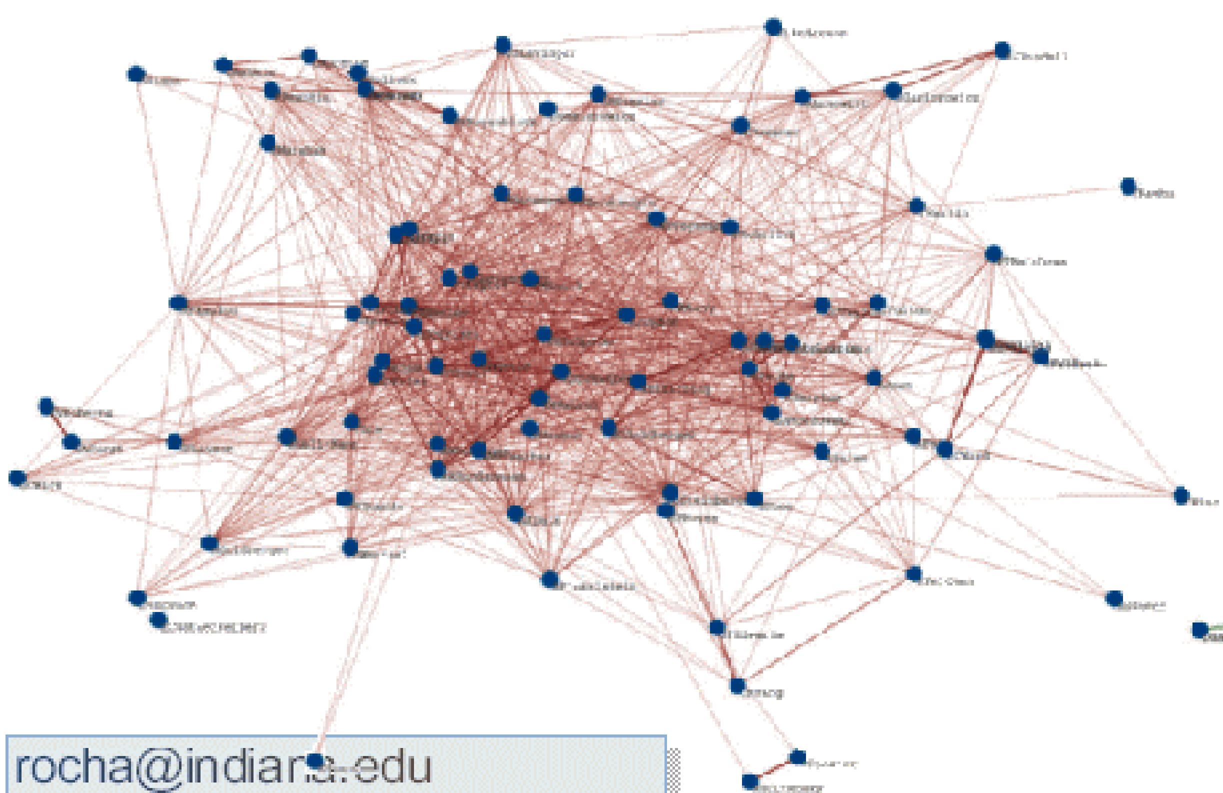
- All six authors in top pairs
  - Very strong proximities to *EESalpeter* (Cornell) and *KMWatson* (postdoc at IAS, prof at Indiana) in AP and with to *EESalpeter* (Cornell) and *FJDyson* (Cornell, IAS) in CP
    - Cornell and Institute of Advanced Studies

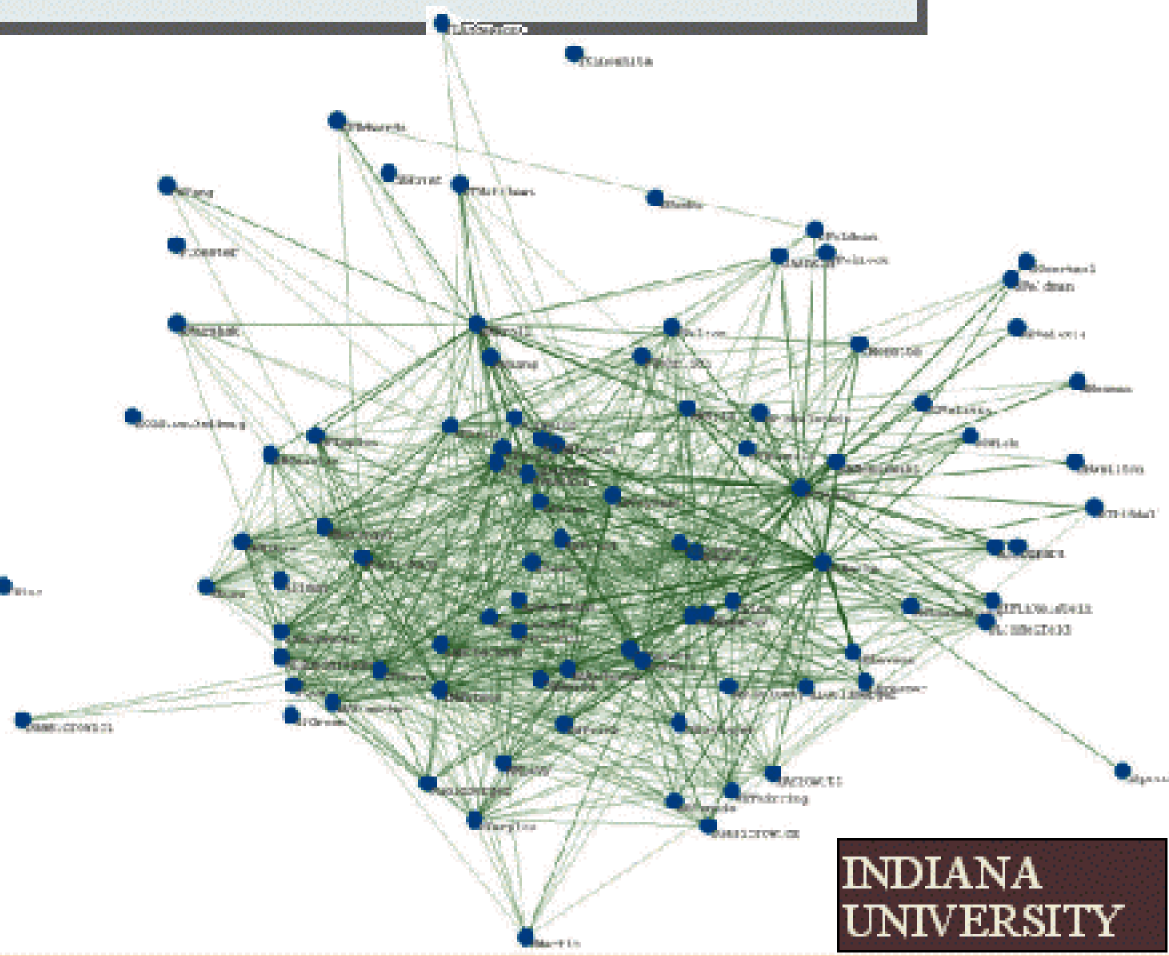
rocha@indiana.edu
http://informatics.indiana.edu/rocha

co-acknowledgment network

semi-metric analysis

- *AP* is very semi-metric
  - 139 papers, 91 authors
  - Percentage of pairs with positive semi-metric ratios (*rs* and *s* parameters): 18.3%
  - Percentage of pairs with indirect distances smaller than the average distance of direct edges to either node (*b* parameter): 30.8%
  - Many strong implicit associations

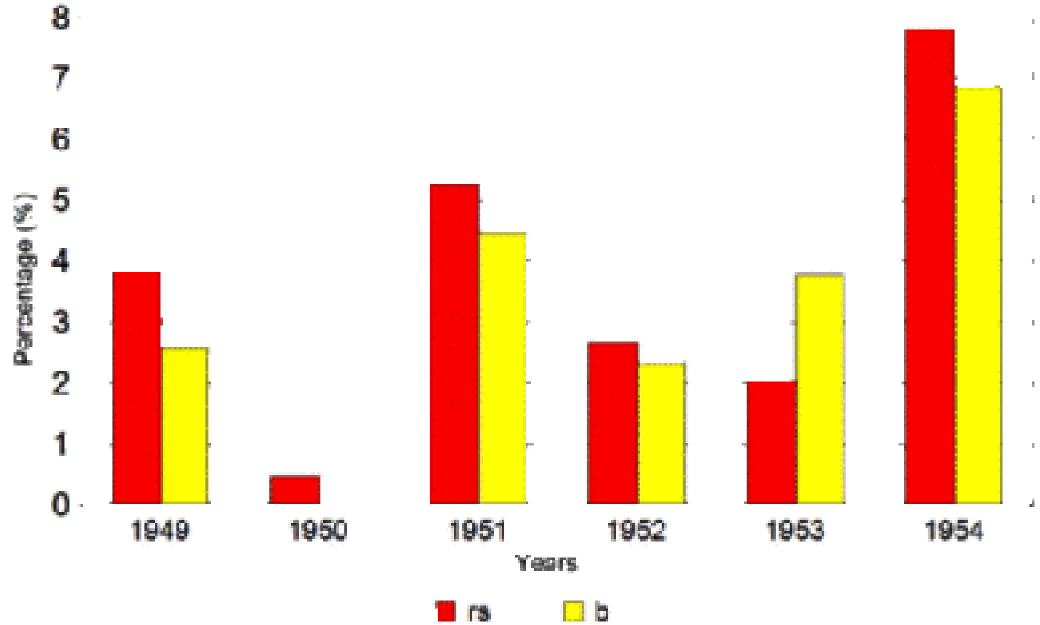


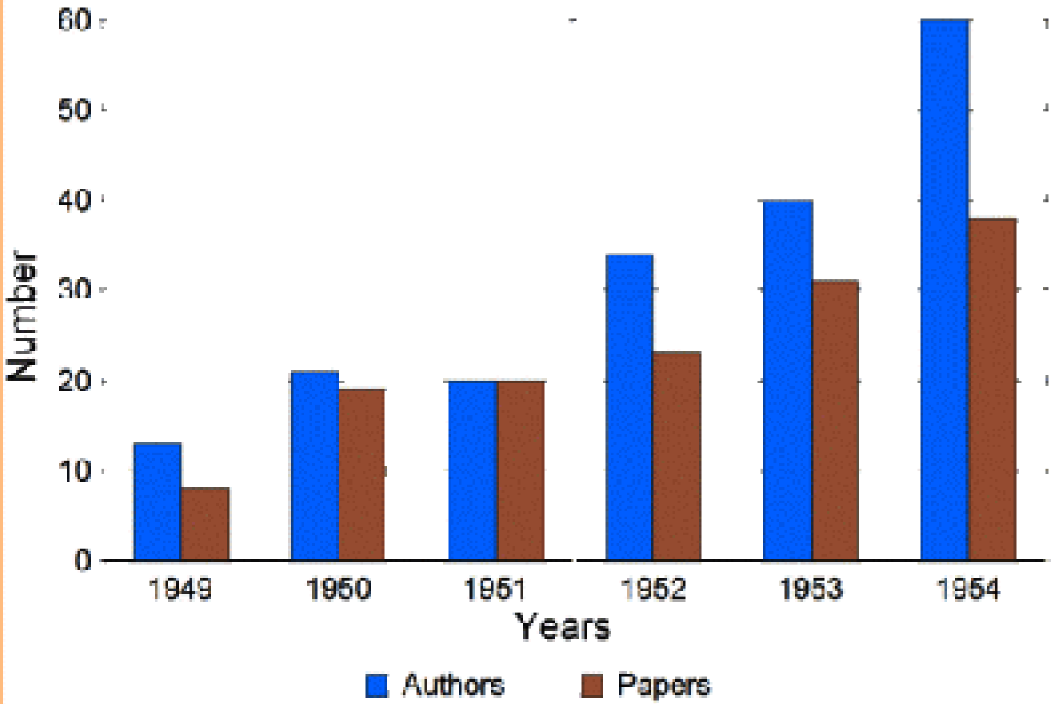


rocha@indiana.edu
http://informatics.indiana.edu/rocha

dynamics of co-acknowledgment network

semi-metric analysis





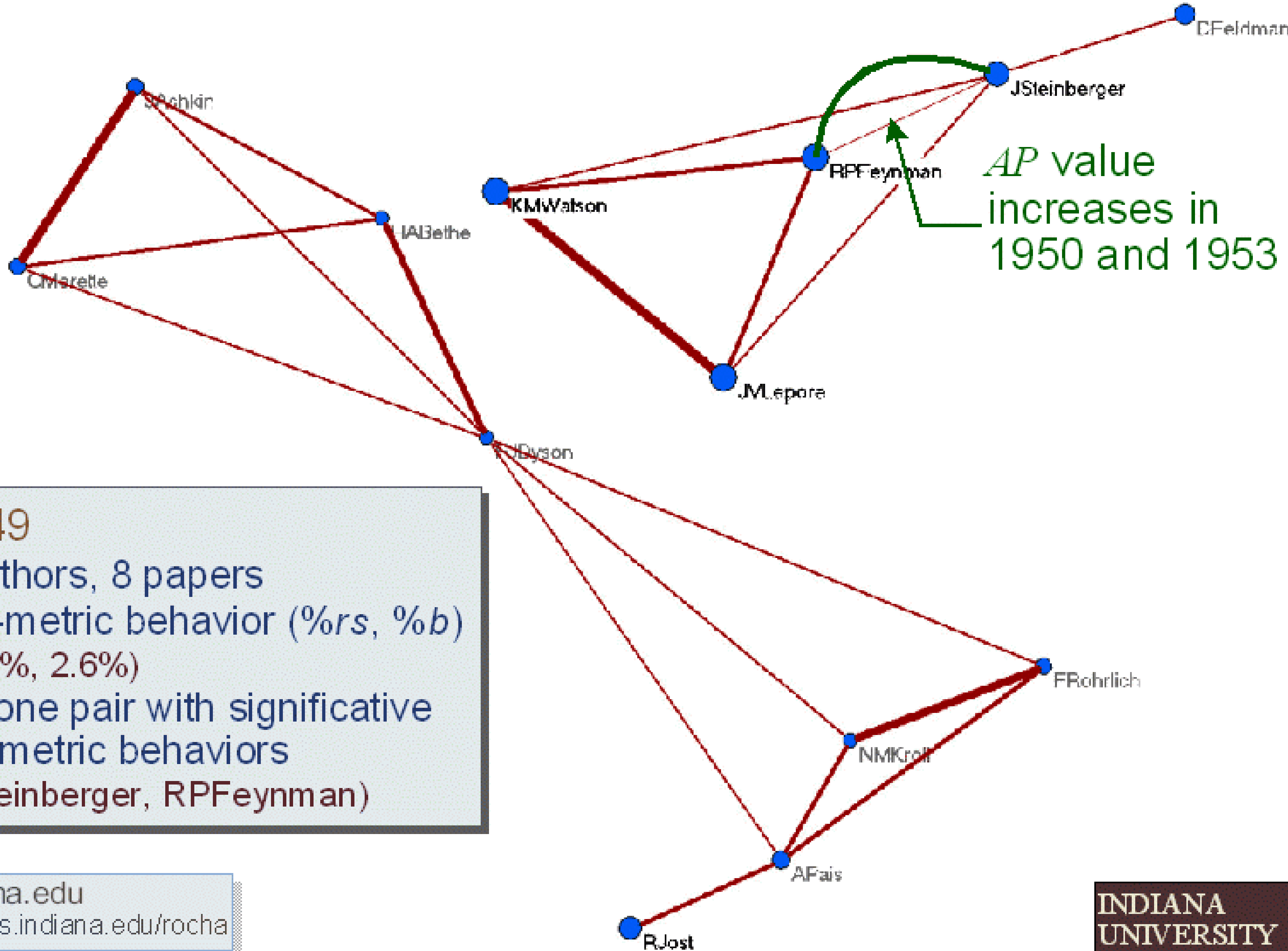
- *AP* computed for every individual year between 1949 and 1954
  - Papers, Authors
    - (8, 13); (19, 21); (20, 20); (23, 34); (31, 40); (38, 60)
  - Compared with the global *AP*, the individual years are more metric
    - 1950 is almost completely metric
  - Semi-metric behavior (%*rs*, %*b*)
    - 1949: (3.9%, 2.6%)
    - 1950: (0.5%, 0.0%)
    - 1951: (5.3%, 4.5%)
    - 1952: (2.7%, 2.3%)
    - 1953: (2.1%, 3.8%)
    - 1954: (7.8%, 6.9%)
  - Can semi-metric pairs uncover latent and future associations?

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

INDIANA UNIVERSITY

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)

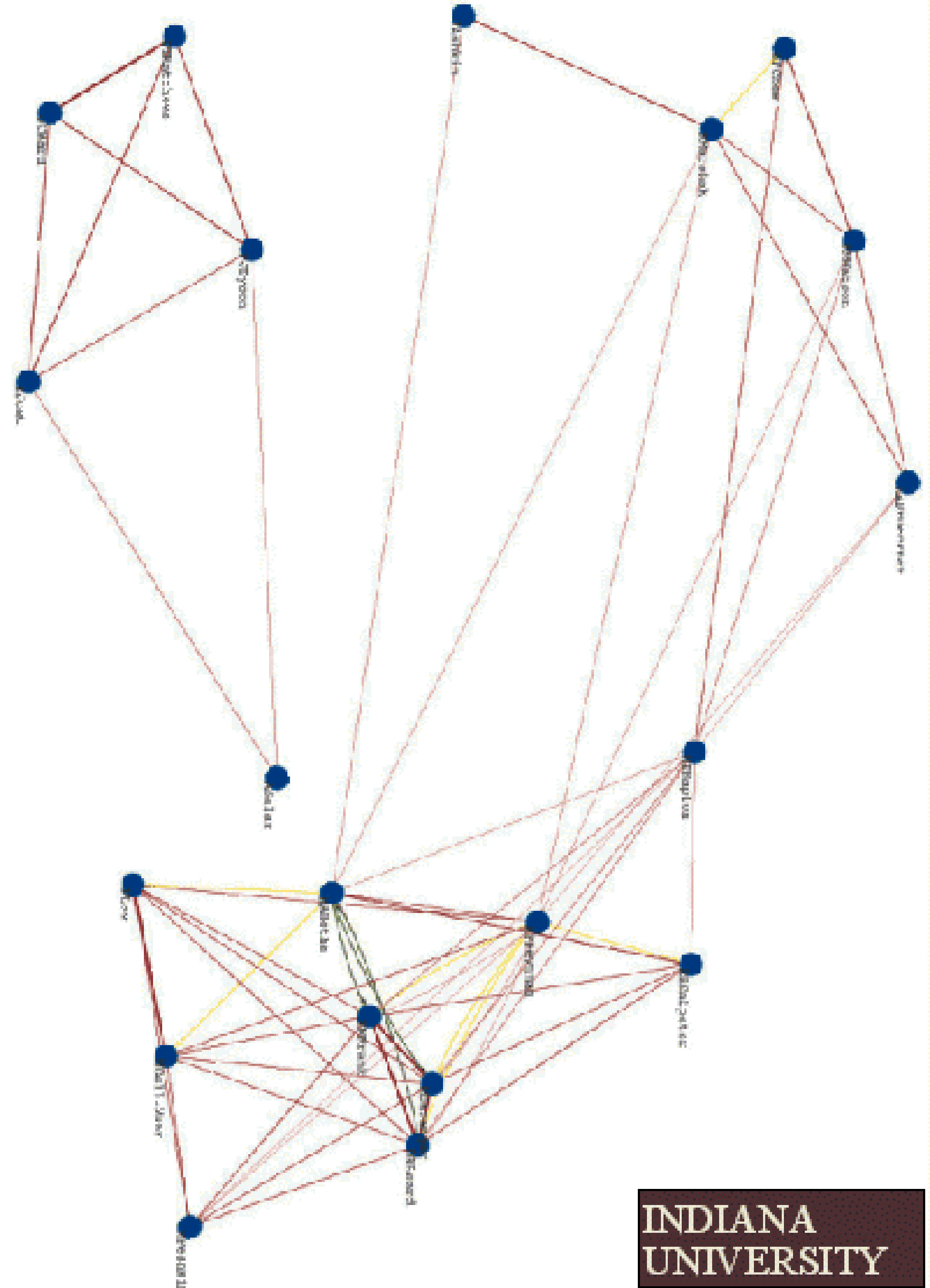
rocha@indiana.edu
http://informatics.indiana.edu/rocha

INDIANA UNIVERSITY

dynamics of co-acknowledgment network

1951

- *AP* 1951
  - Not very strong semimetric behavior
  - (HABethe, RMFrank): both rs and b
    - Advisor/ Student at Cornell
    - David Kaiser suspects Bethe learned about the diagrams via Frank
  - (DBBeard, HABethe): both rs and b
    - Wrote paper together in 1951
  - (HABethe, Mbaranger): both rs and b
    - Wrote paper together in 1953; high value of proximity in co-collaboration network (CP); value of AP increases in 1952 and 1953
  - (RPFeynman, EESalpeter): b
    - High proximity in co-collaboration network. No link in AP 1951, but AP increases in 1952 and 1953.
  - (HABethe, Flow) : b
    - No link in AP 1951, but AP increases in 1952 and 1953.
  - (HABethe, Mgell-Mann): b
    - No link in AP 1951, but AP increases in 1954.



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

latent associations in networks

second stage of adaptive webs

- Extraction of co-occurrence (associative) networks
  - Represent associative knowledge of information resources and users
- Identification of implicit associations in networks
  - Discovery of relevant items, missing information, trends, associations with higher future probability of occuring
  - Identify Communities of Users
  - Applications
    - Recommendation systems, social networks, bioinformatics
  - Complex systems methodology: network analysis and knowledge integration

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

INDIANA UNIVERSITY