

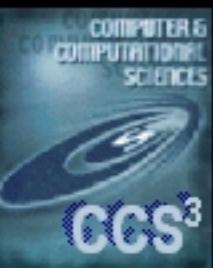


from document networks to the adaptive web

Luis Rocha
2004



the active recommendation project



luis m. rocha

CCS3 - modeling, algorithms, and informatics
los alamos national laboratory, MS B256

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>





Luis Rocha
2004



Adaptive 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
- Identification of implicit associations in networks
 - ▶ Discovery of relevant items
- Adaptive network architecture
 - ▶ Evolving organization

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



similarity and proximity relations

fuzzy graphs

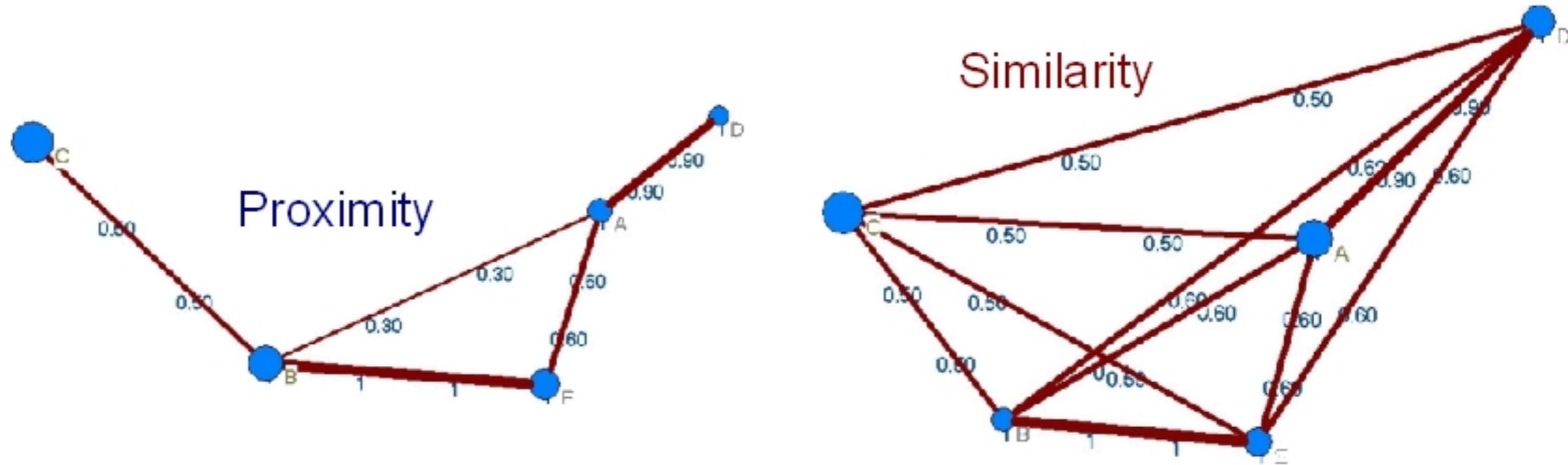
Luis Rocha
2004

■ 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@lanl.gov
<http://www.c3.lanl.gov/~rocha>



extraction of co-occurrence networks

from document relations

- Document \times Keyterms
 - ▶ Keyterm Co-Occurrence
- Document \times Document
 - ▶ Co-Citation or Hyperlink structure
- Document \times Author
 - ▶ Co-Authorship (Collaboration Network)
- Genes \times MeSH Keyterms
 - ▶ Gene/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@lanl.gov
<http://www.c3.lanl.gov/~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.





Luis Rocha
2004

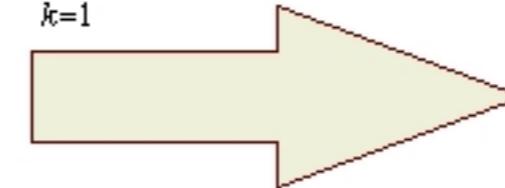


proximity measures

produce associative (probabilistic) networks

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

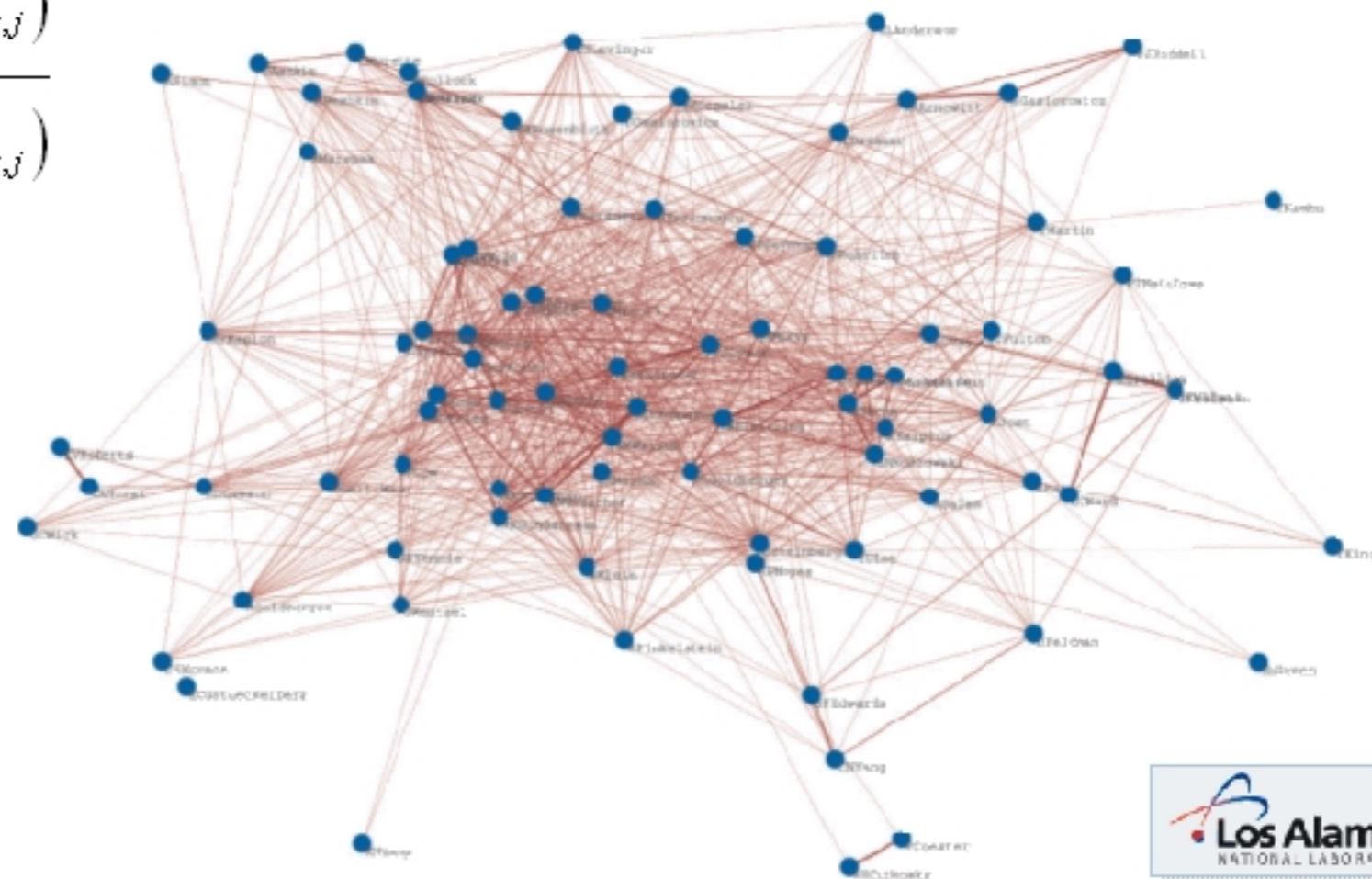
$$XYP(x_i, x_j) = \frac{\sum_{k=1}^m (r_{i,k} \wedge r_{j,k})}{\sum_{k=1}^m (r_{i,k} \vee r_{j,k})}$$



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

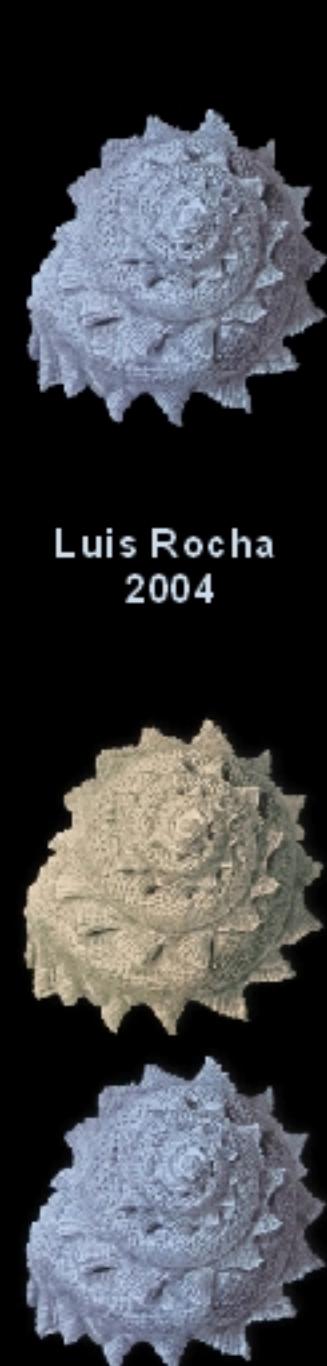
$$YXP(y_i, y_j) = \frac{\sum_{k=1}^n (r_{kj} \wedge r_{kj})}{\sum_{k=1}^n (r_{kj} \vee r_{kj})}$$

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



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>





people document proximity

from document relations

	$P(\text{people names})$
$D(\text{Documents})$	$A : P \times D$

318 names of terrorists
from unclassified database

$PDP(p_i, p_j)$ is the probability that both persons p_i and p_j co-occur in the same document $d \in D$. Conversely, $DPP(d_i, d_j)$ is the probability that both documents d_i and d_j contain the same person $p \in P$.

$$PDP(p_i, p_j) = \frac{\sum_{k=1}^m (a_{i,k} \wedge a_{j,k})}{\sum_{k=1}^m (a_{i,k} \vee a_{j,k})}; \quad DPP(d_i, d_j) = \frac{\sum_{k=1}^n (a_{k,i} \wedge a_{k,j})}{\sum_{k=1}^n (a_{k,i} \vee a_{k,j})}$$

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Rocha, 2002. LAUR 02-6557
Voss & Joslyn 2002. LAUR 02-7867



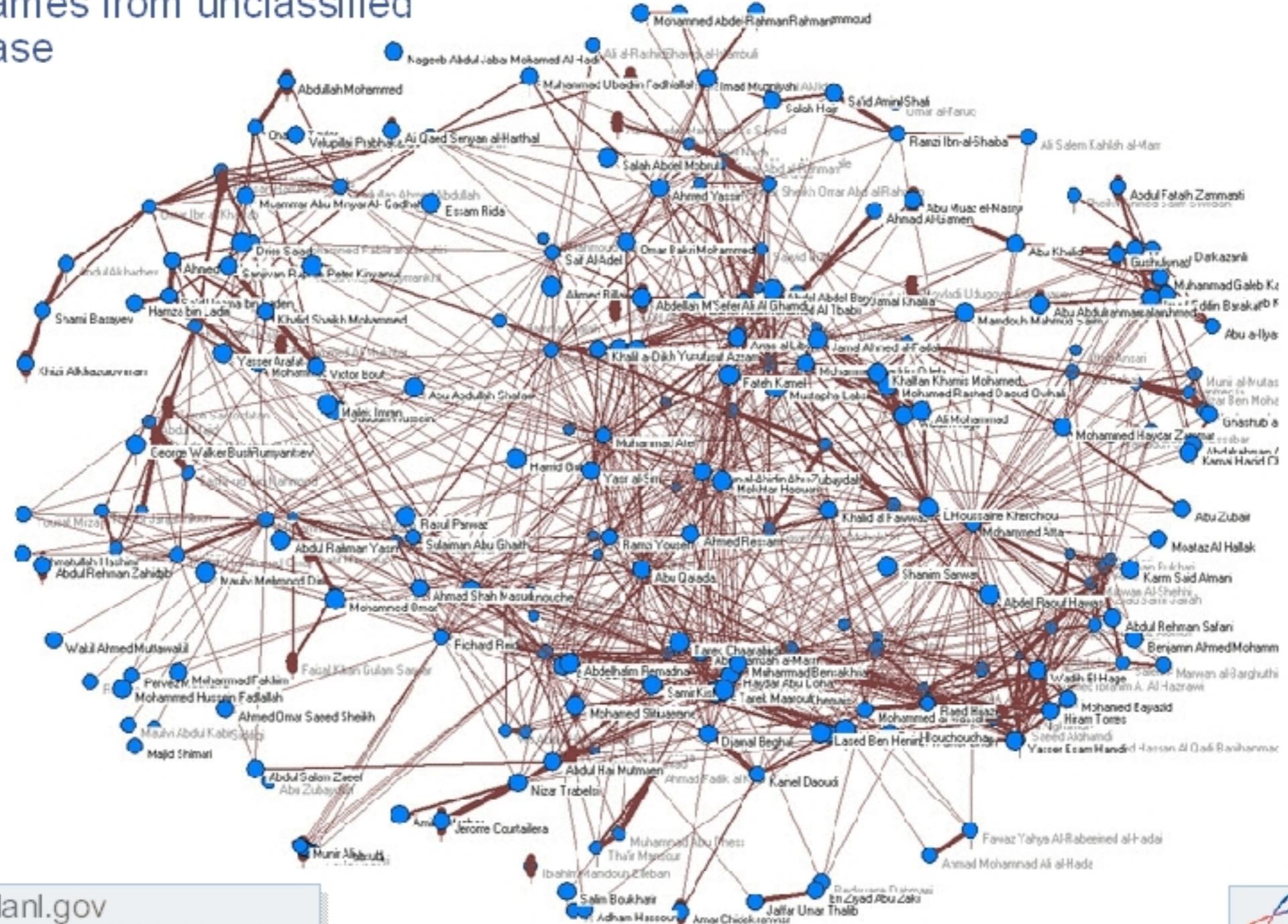


terrorist networks

PDP2

318 names from unclassified database

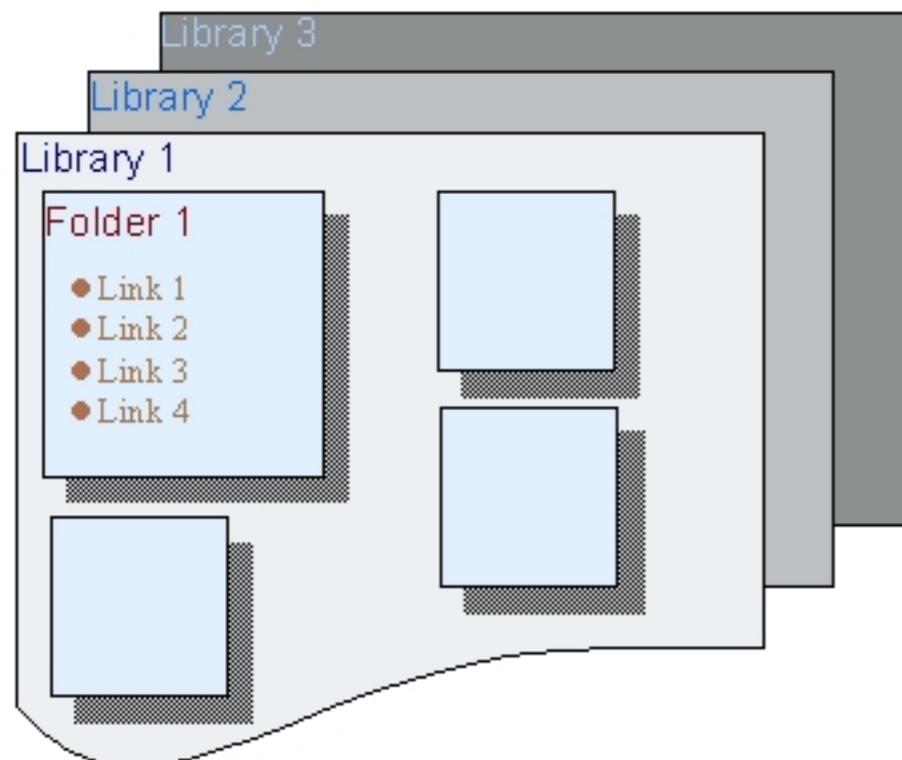
Luis Rocha
2004



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



architecture



Luis Rocha
2004

- 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@lanl.gov
<http://www.c3.lanl.gov/~rocha>

With Tiago Simas, Andreas Rechtsteiner and
Chien-feng Huang

The screenshot shows a web page with a purple header and sidebar. The header includes a navigation bar with categories like "Mathematics", "Fuzzy Math", and "Immunology-shared". Below the header is a search bar and a message box. The main content area is divided into sections:

- Distance Functions**:
 - [Distance Functions and Topologies](#)
 - [Eulerian space and grouping of biological objects](#)
 - [DISTANCE FUNCTIONS AND TOPOLOGIES](#)
 - [EXPLICIT METRIZATION](#)
- Mathematics Web Resources**:
 - [Hyperstat Online](#)
- Databases**:
 - [Current index to statistics](#)
 - [FleshPoint](#)
 - [INSPECIE at LANL](#)
 - [Jahrbuch über die Fortschritte der Mathematik](#)
 - [MathSciNet](#)
 - [SciSearcher at LANL](#)
 - [Zentralblatt MATH database](#)
- Graph Theory**:
 - [A duplication growth mode of gene expression networks](#)
 - [A Stochastic Model for the Evolution of the Web](#)
 - [Accelerated growth of networks](#)
 - [Curvature of co-links uncovers hidden thematic layers in the World Wide Web](#)
 - [Dynamical small-world behavior in an epidemiological model of mobile individuals](#)
 - [Friends and Neighbors on the Web](#)
 - [Graph structure in the web](#)
 - [Intentional Walks on Scale-Free Small Worlds](#)
 - [Intentional Walks on Scale-Free Small Worlds](#)
 - [Local Search in Unstructured Networks](#)
 - [Modeling the Internet's large-scale topology](#)
 - [Nodes of the Small World: a Review](#)
 - [Optimization in complex networks](#)
 - [Random graph models of social networks](#)
 - [The Erratability of Collective Choice with Shared Knowledge Structures](#)
 - [The structure and function of complex](#)

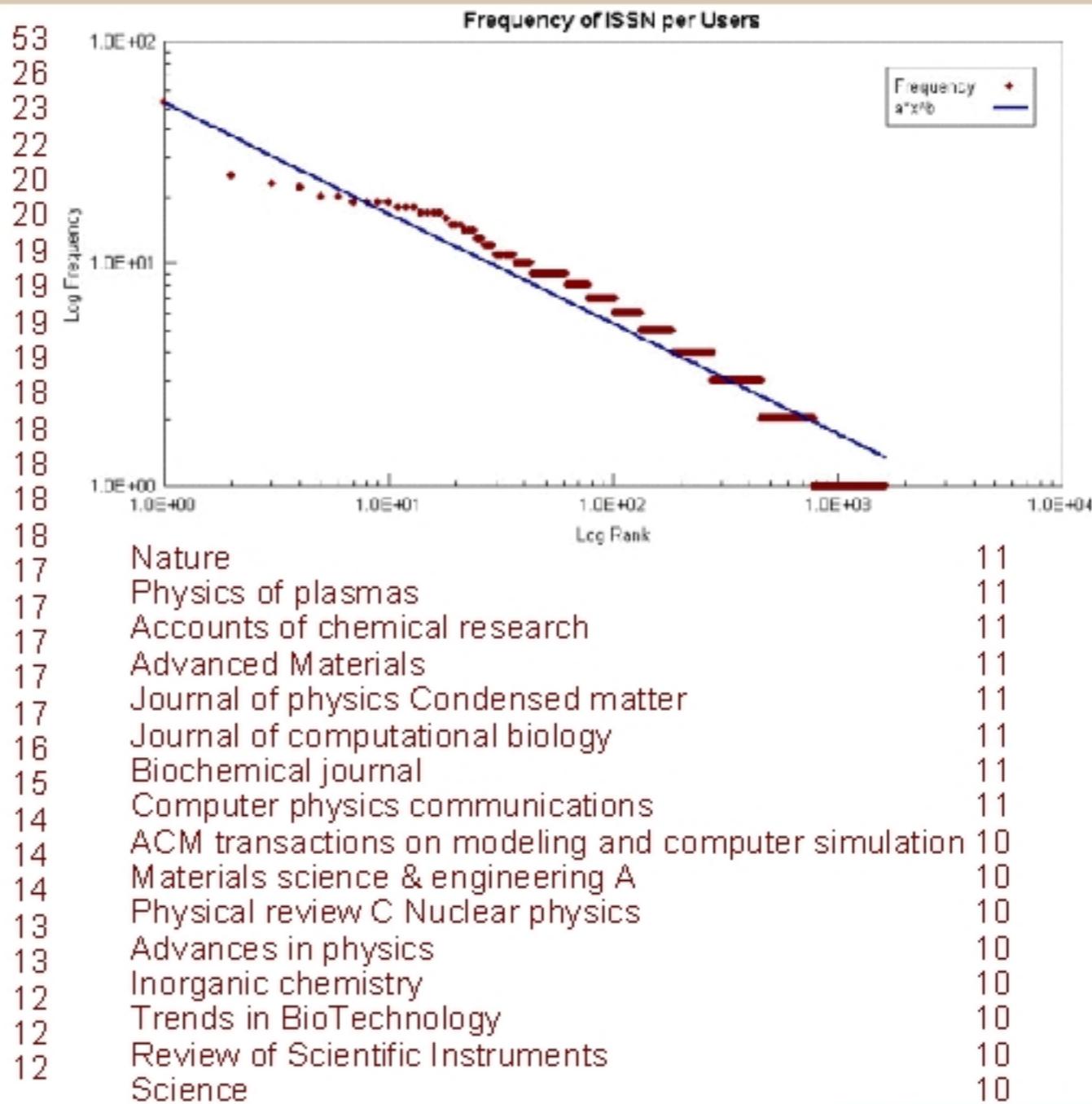
most frequent ISSN

occurrences in personalities

Luis Rocha
2004

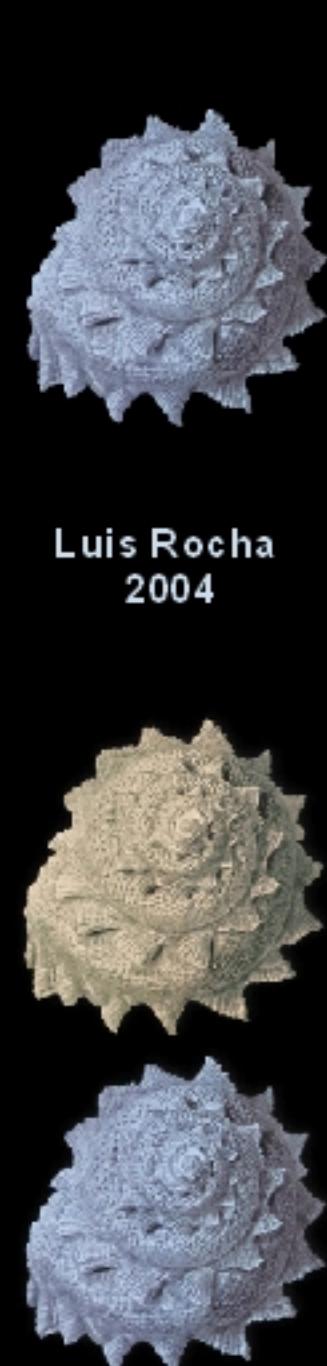


Physical review letters
Physical Review B
Physical review E
Physical review A General physics
Journal of physical chemistry B
Computers & geosciences
Scientific American
Journal of the American Chemical Society
Journal of Chemical Physics
Reviews of modern physics
Bioinformatics
IEEE trans. on geoscience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



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>





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 occurring 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)

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

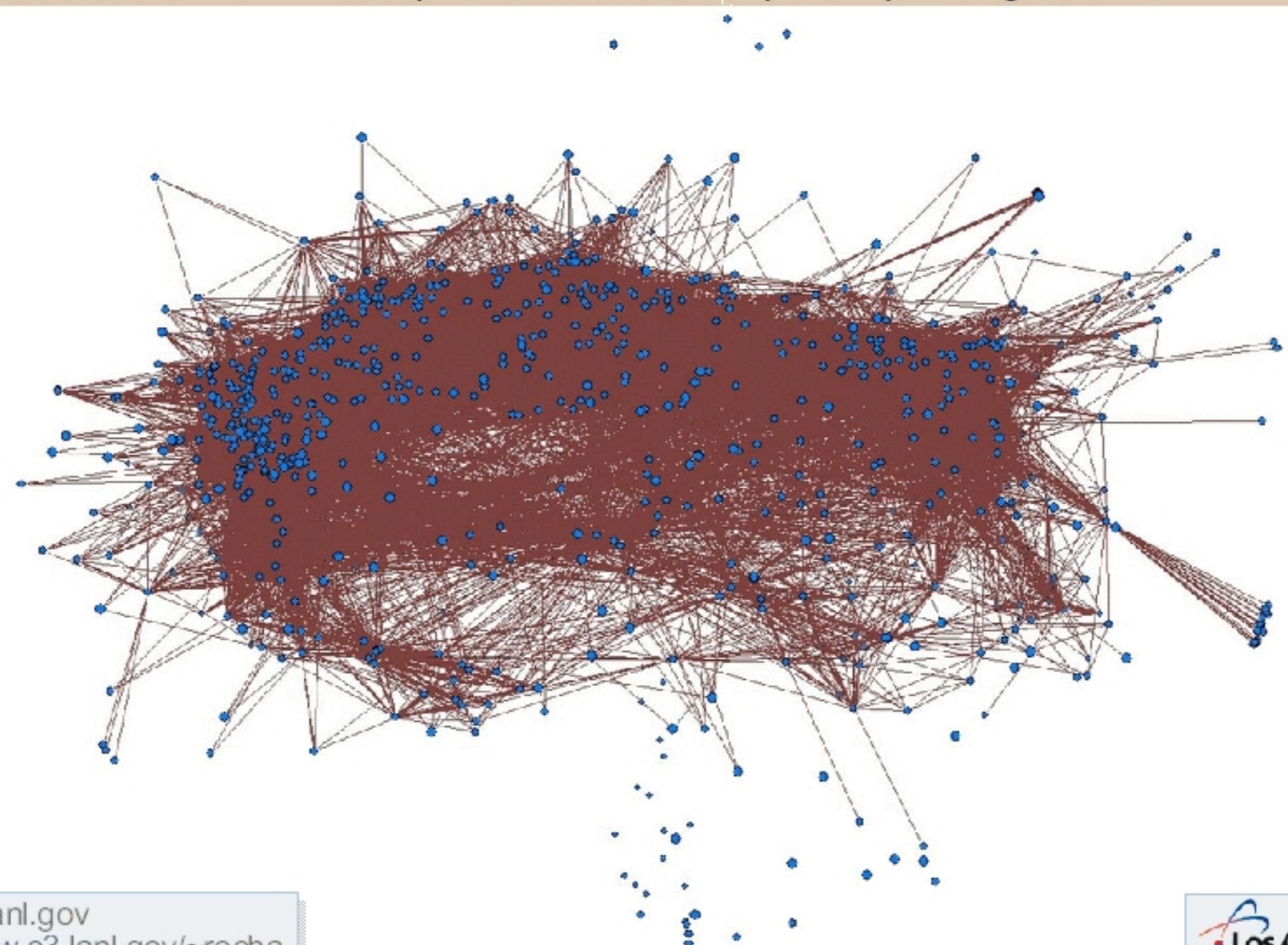
(ISSN Personality Proximity)



journal network

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

Luis Rocha
2004



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>





top 5 most frequent ISSN and their neighbors

Luis Rocha
2004



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

ipp ≥ 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



recommendations based on proximity

mylibrary.lanl.gov

Luis Rocha
2004

- 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@lanl.gov
<http://www.c3.lanl.gov/~rocha>





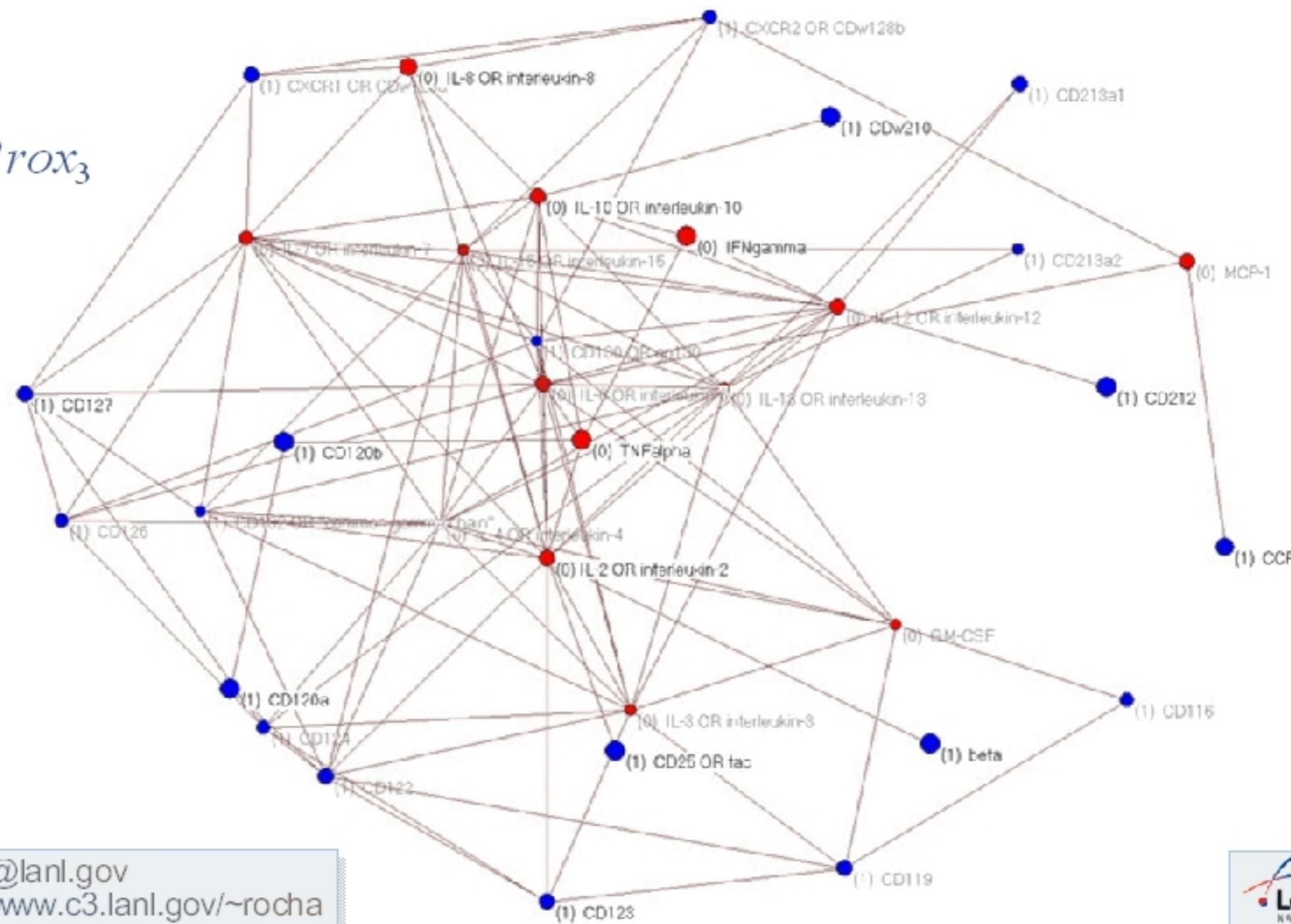
Luis Rocha
2004



cytokine-receptor molecule network

from co-occurrence proximity in AltaVista

Prox₃

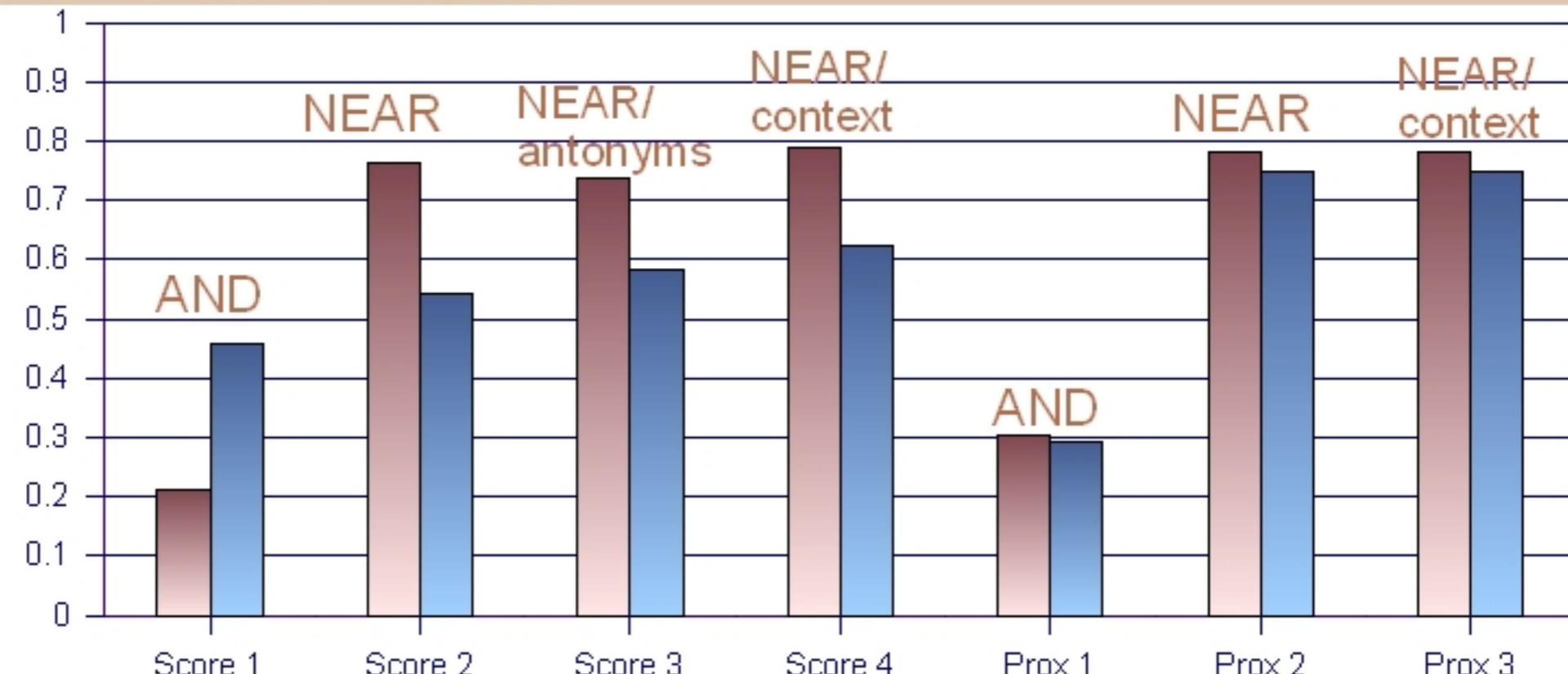


rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



precision and recall

values obtained with FaCSO



Precision: probability that an identified association is relevant

Recall: probability that an association has been identified given that it is relevant

$$precision = \frac{\|\{relevant\} \cap \{retrieved\}\|}{\|\{retrieved\}\|}$$
$$recall = \frac{\|\{retrieved\} \cap \{relevant\}\|}{\|\{relevant\}\|}$$

Rocha, Luis M. and Andreas Rechtsteiner [2003]. "Fast Cheap and Synthetic Oracle (FaCSO): Proximity Measures to capture Expert Knowledge in the Biobiume". *Pacific Symposium on Biocomputing 2003*.

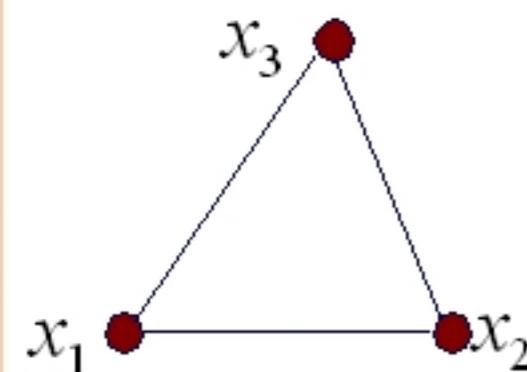
rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



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

Evolution

3.89

Adaptive Systems

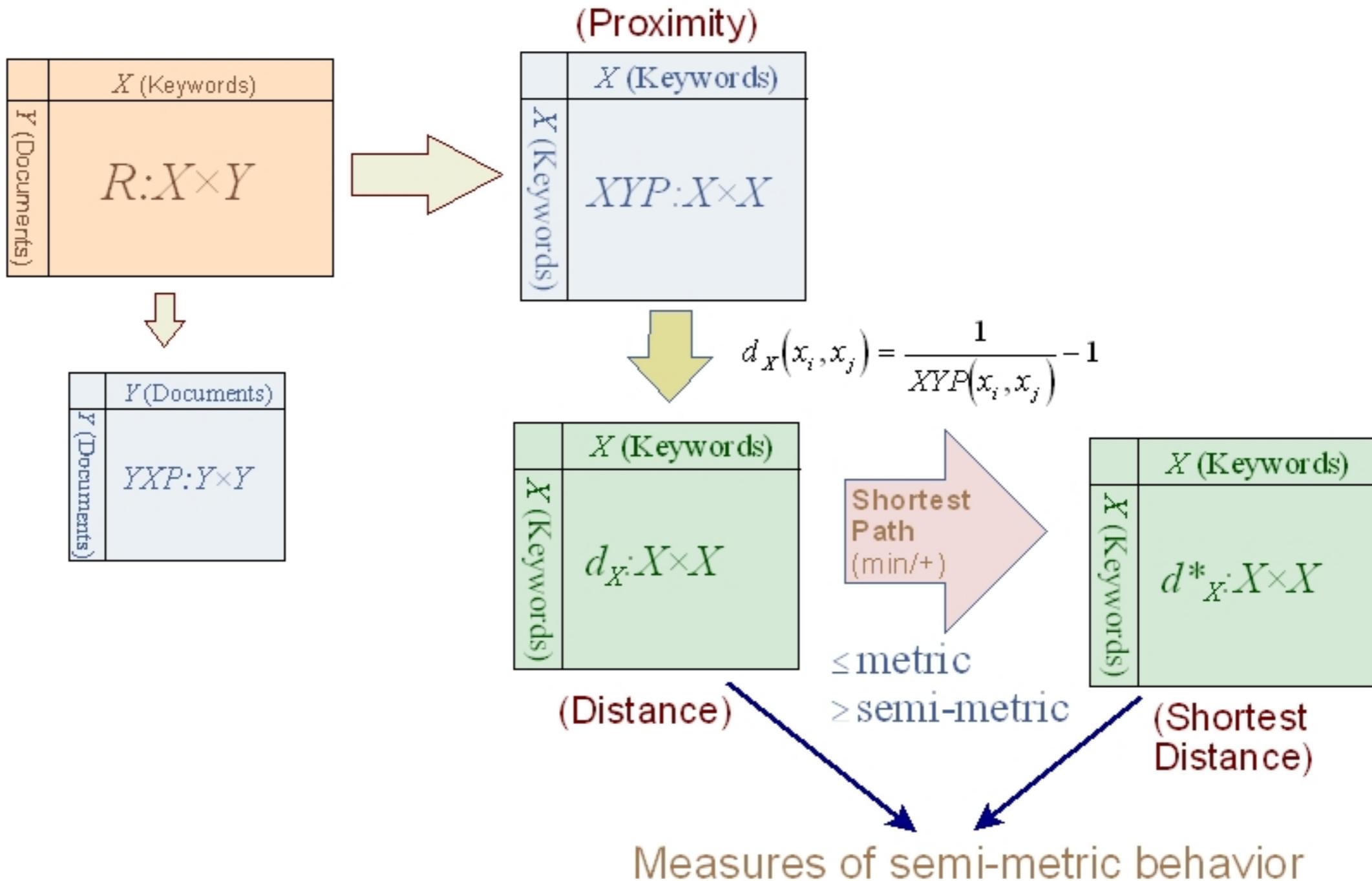
6.89

Cognition

44

Semi-metric ratio: 6.3861

computing semi-metric behavior





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@lanl.gov
<http://www.c3.lanl.gov/~rocha>



applied to many data sets

- Keywords extracted from books cited in a dissertation
 - ▶ Identified novel theme associations
- ARP Database
 - ▶ 3 million scientific articles published 1996-2000
 - ▶ Identified trends
- Web Sites, Web logs, Word Norm, Random Graphs
 - ▶ Denotes Latent Associations

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.



identification of latent associations

what do semi-metric edges imply?

- Luis Rocha
2004
- 
- 
- 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@lanl.gov
<http://www.c3.lanl.gov/~rocha>

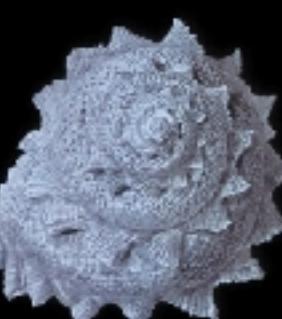




semi-metric recommendations

catching strong indirect associations in mylibrary.lanl.gov

Luis Rocha
2004



IPP_3: parameter *b*

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
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 0740-7487--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 *rs*

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



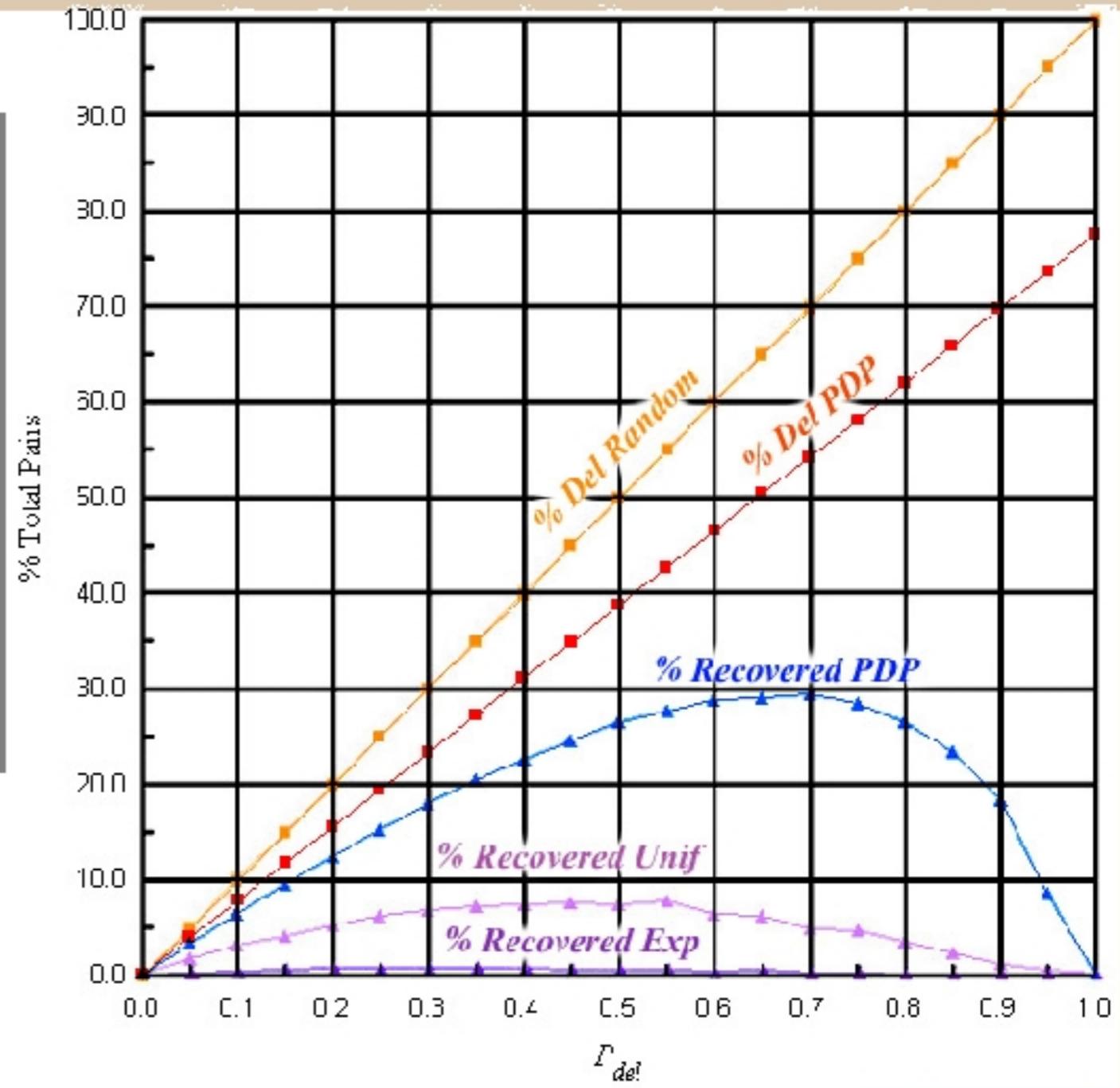
detecting incomplete knowledge

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@lanl.gov
<http://www.c3.lanl.gov/~rocha>

Acknowledgement: John Hogden



Luis Rocha
2004





scientific community working on feynman diagrams

as published in *Physical Review*, 1949-54

$P(\text{author names})$	$P(\text{author names})$
$C:P \times P$	

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

$P(\text{author names})$	$P(\text{author names})$
	$A:P \times P$

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

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

co-collaboration network

semi-metric analysis

- CP is almost metric
 - ▶ 139 papers, 76 authors
 - ▶ Percentage of pairs with positive semi-metric ratios (r_s 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@lanl.gov
<http://www.c3.lanl.gov/~rocha>

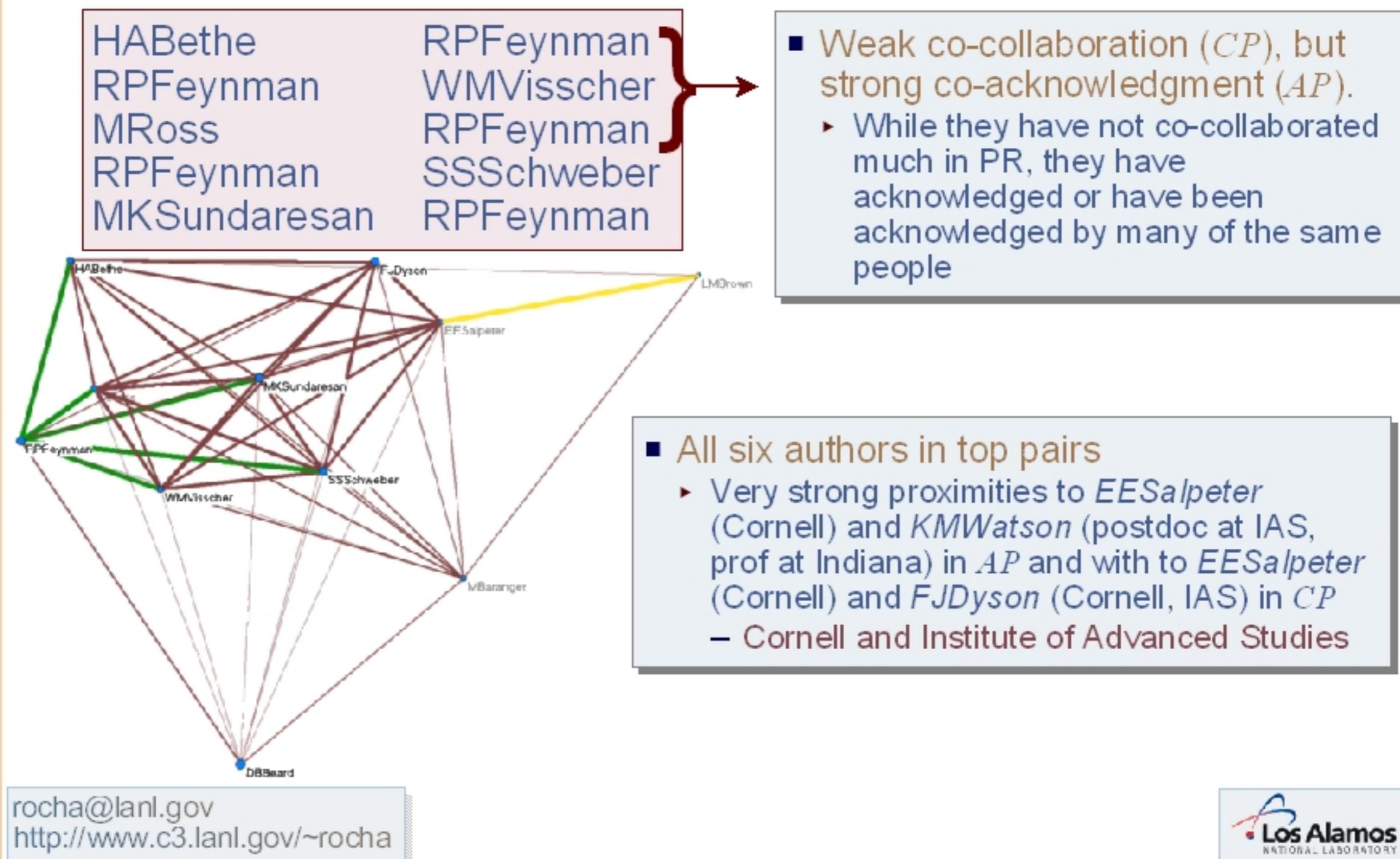


Luis Rocha
2004

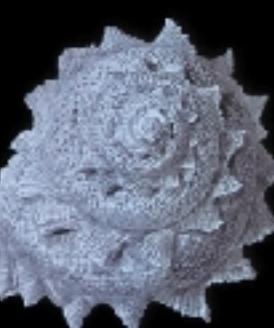


co-collaboration network

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



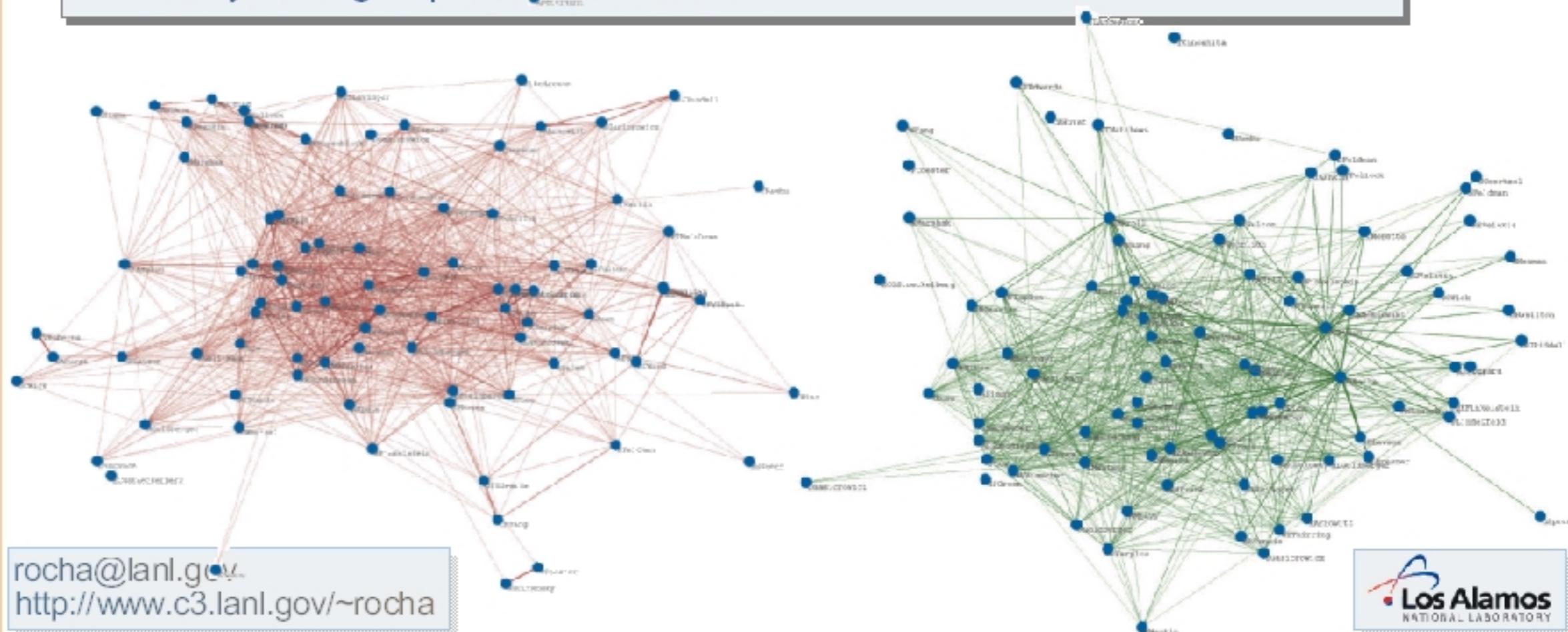
Luis Rocha
2004



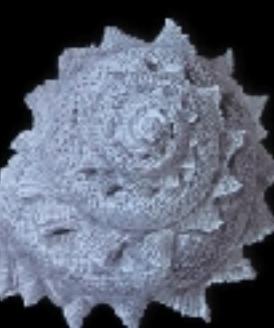
co-acknowledgment network

semi-metric analysis

- *AP* is very semi-metric
 - ▶ 139 papers, 91 authors
 - ▶ Percentage of pairs with positive semi-metric ratios (r_s 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

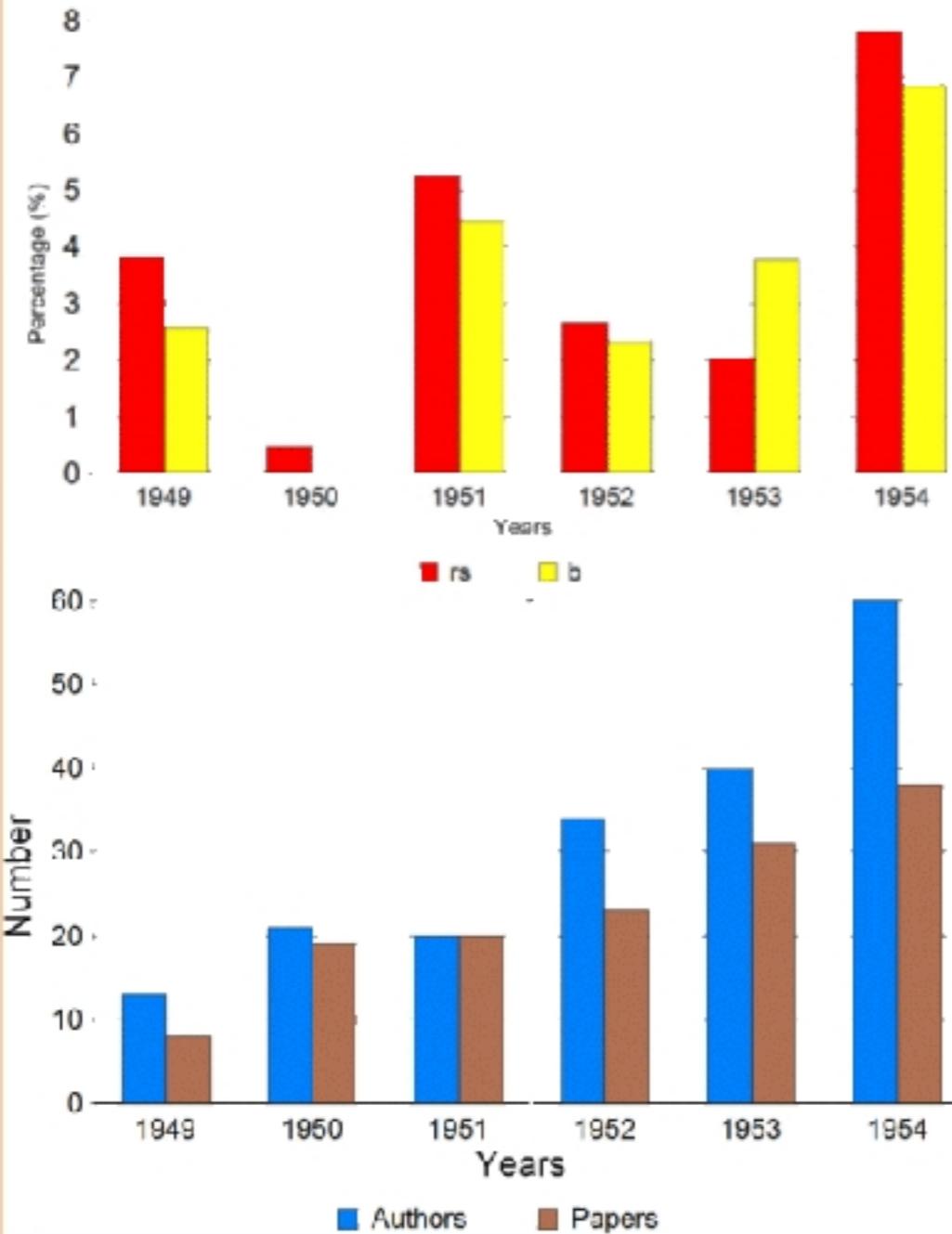


Luis Rocha
2004



dynamics of co-acknowledgment network

semi-metric analysis

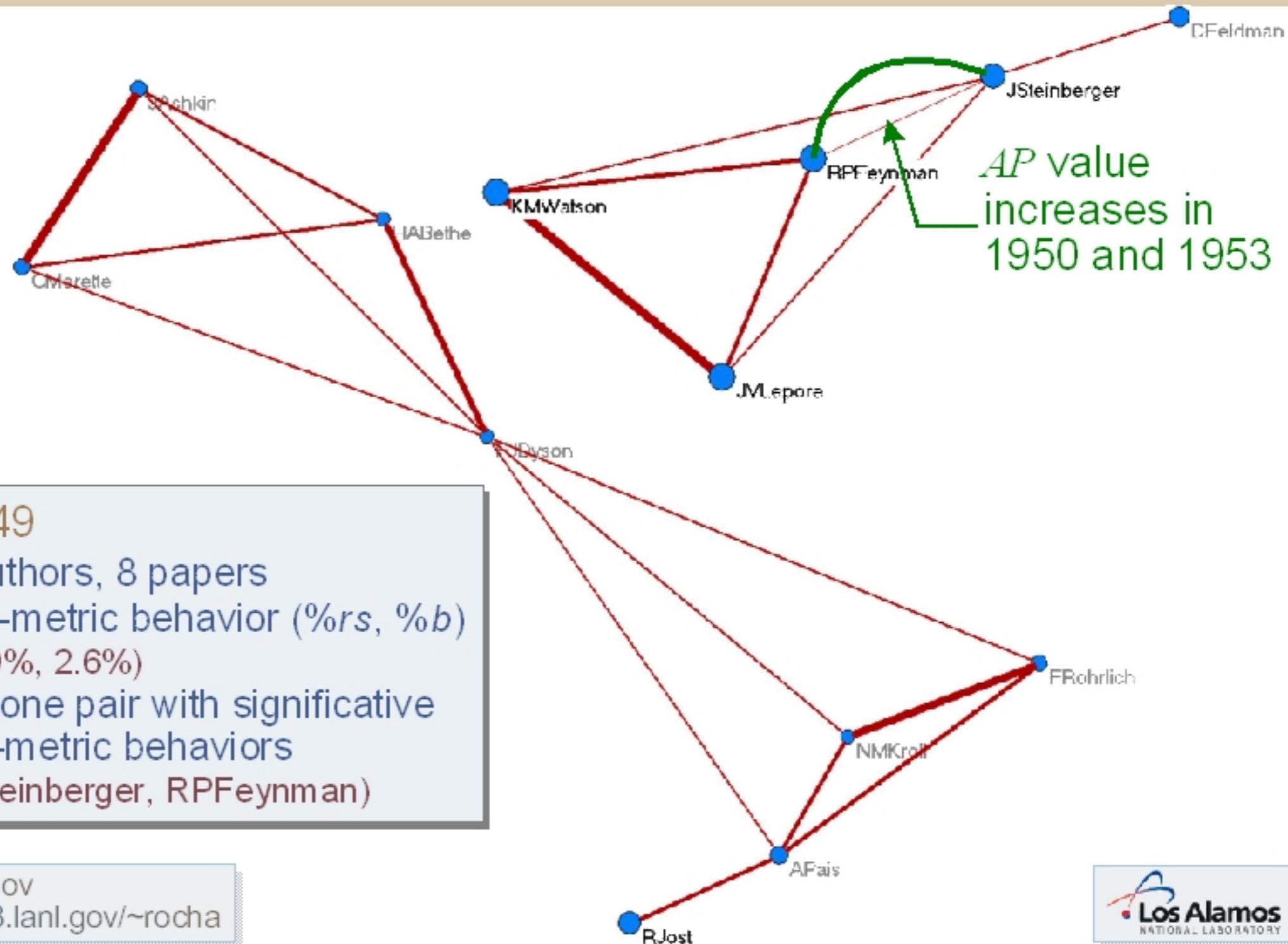


- AP computed for every individual year between 1949 and 1954
 - ▶ Papers, Authors
 - (8, 13); (19, 21); (20, 20); (23, 34); (31, 40); (38, 60)
 - ▶ Compared with the global AP , the individual years are more metric
 - 1950 is almost completely metric
 - ▶ Semi-metric behavior ($\%rs$, $\%b$)
 - 1949: (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@lanl.gov
<http://www.c3.lanl.gov/~rocha>

dynamics of co-acknowledgment network

1949

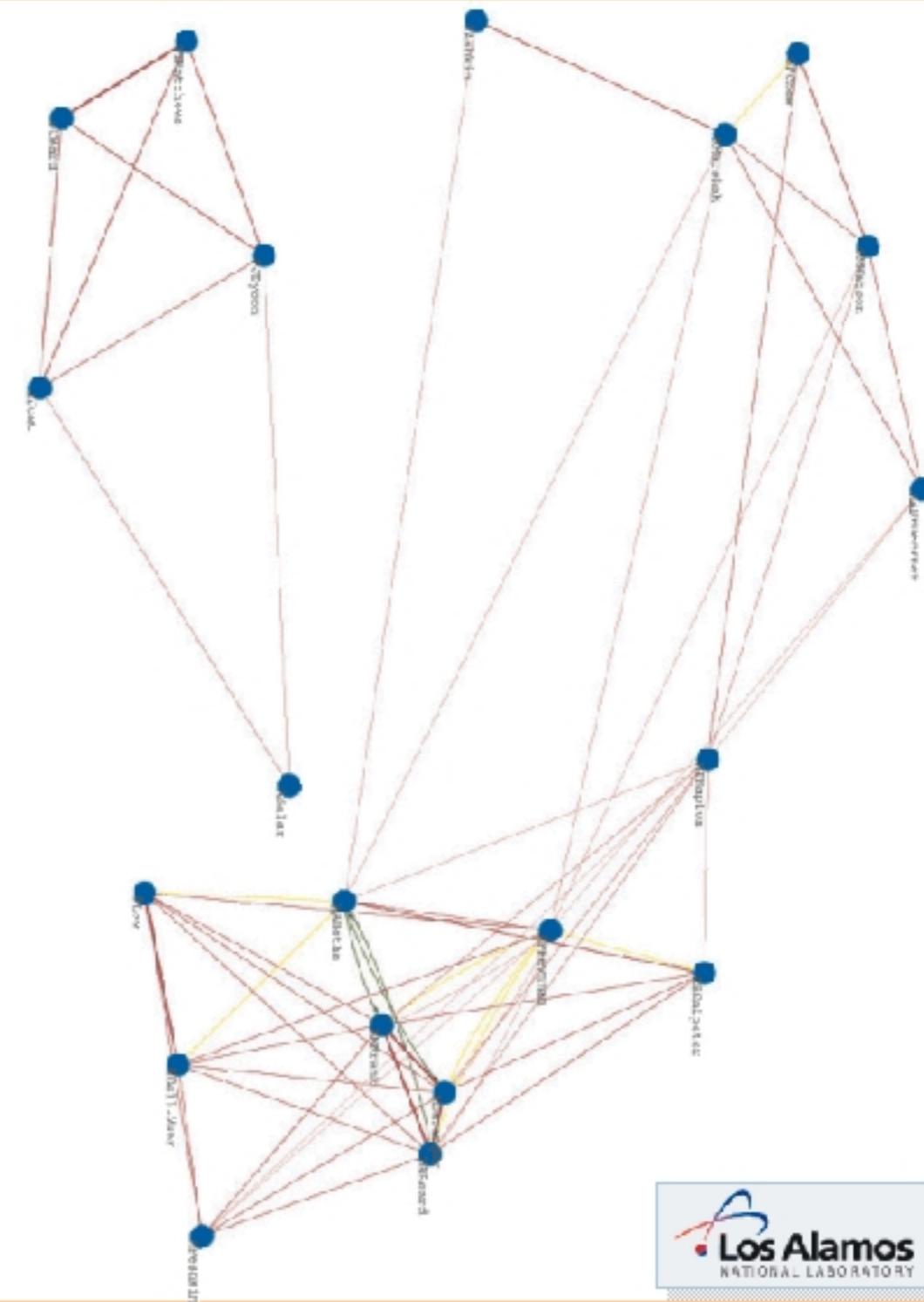


dynamics of co-acknowledgment network

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.

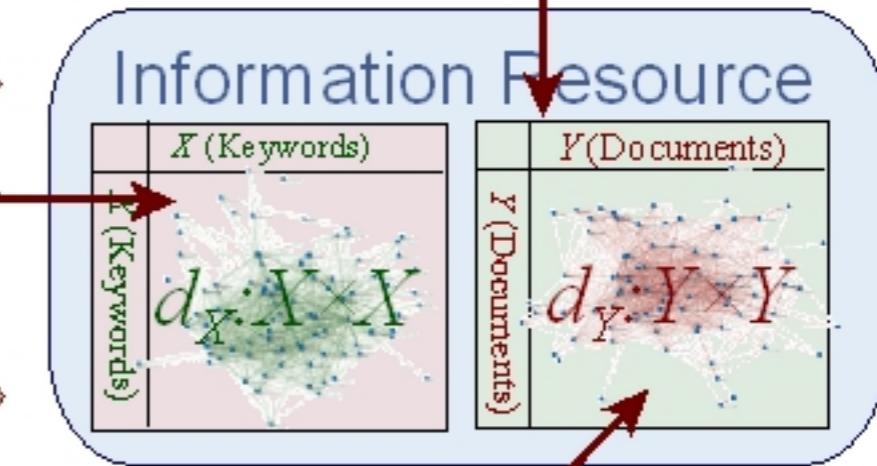
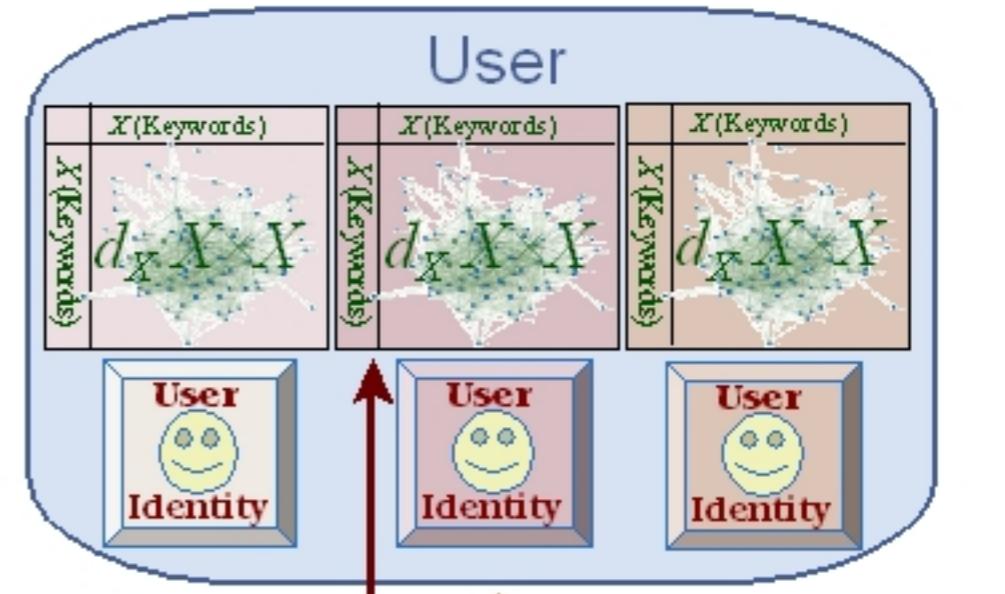
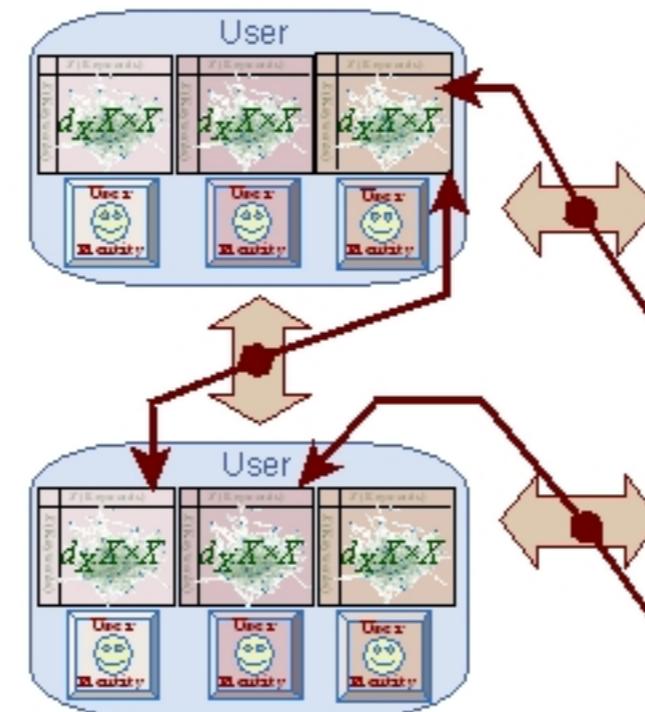
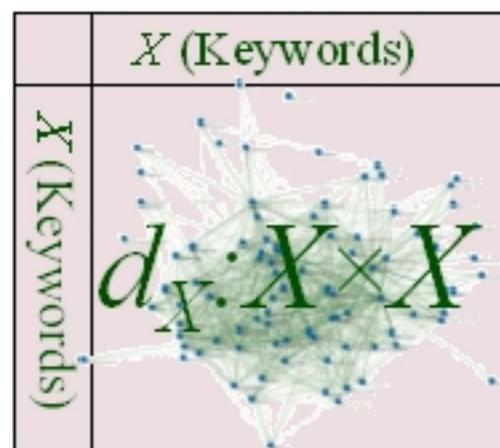


rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

active recommendation systems

adaptive webs from collective behavior

- Represent the *knowledge* of information resources and users as *distance networks*
- Define *active conversation* between users and information resources to produce appropriate *recommendations*
- *Collective Adaptation* to discover and maintain evolving knowledge



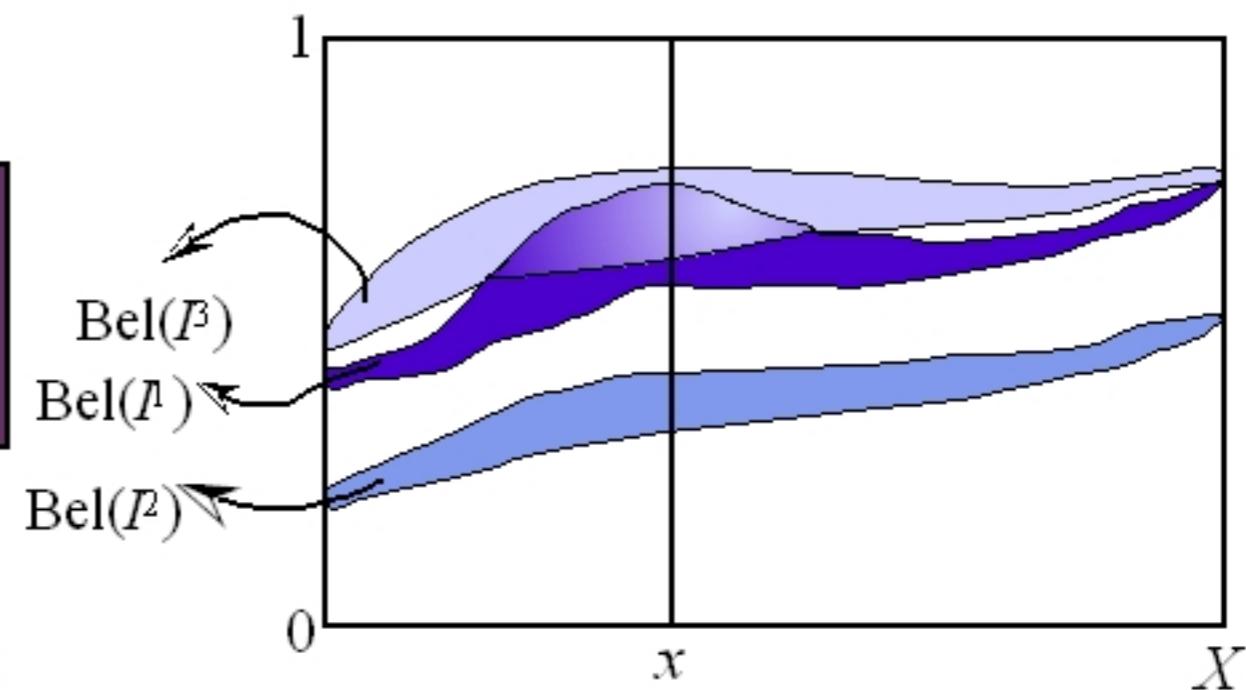
query extraction via automatic conversation

An Algorithm to enable the communication between users/agents and information resources, leading to knowledge exchange, recommendation and adaptation

Evidence Sets

$$A(x): \mathcal{B} \rightarrow [0, 1]$$

A Model of Cognitive Categories

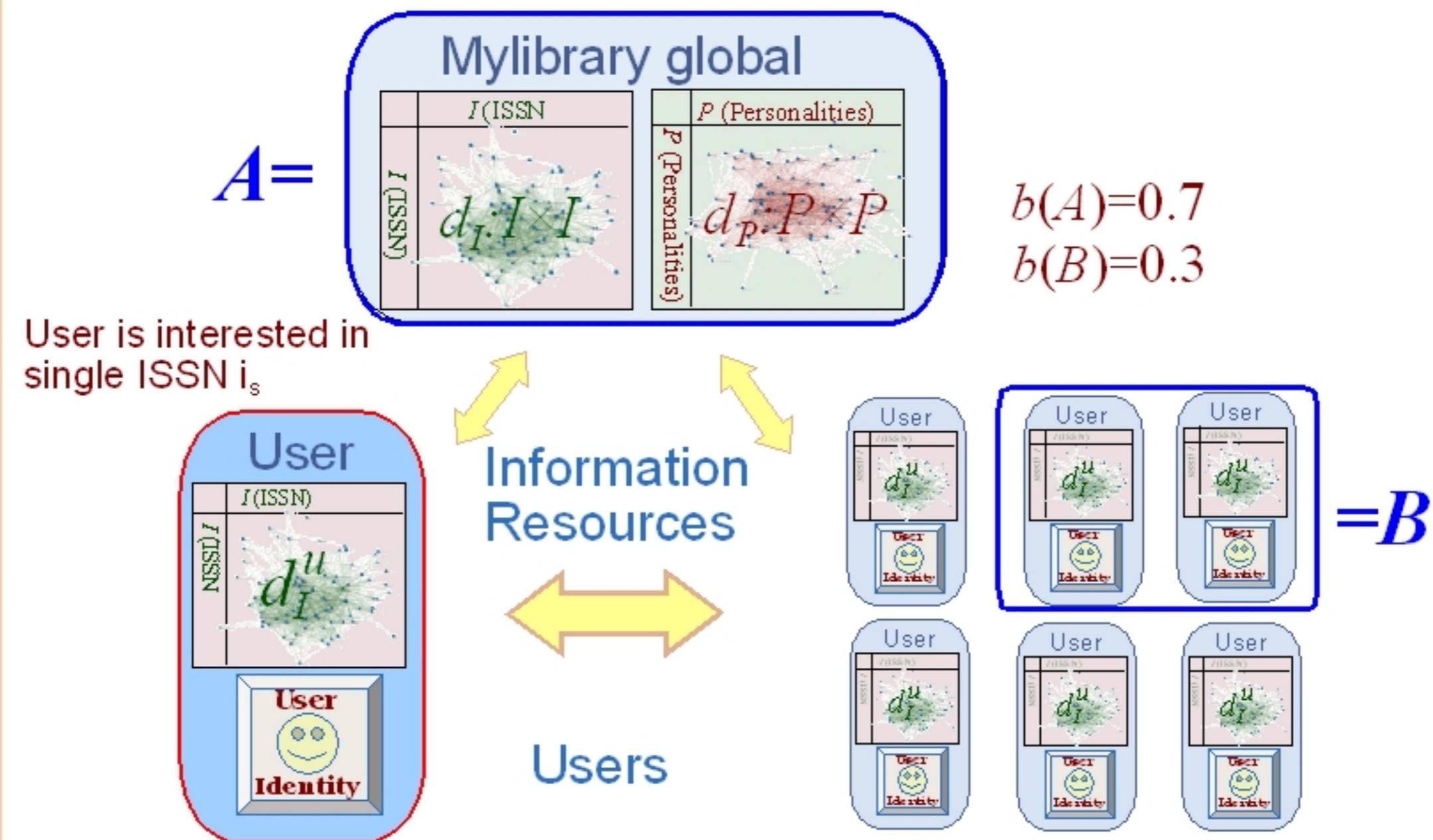


Luis Rocha
2004



TalkMine for MyLibrary

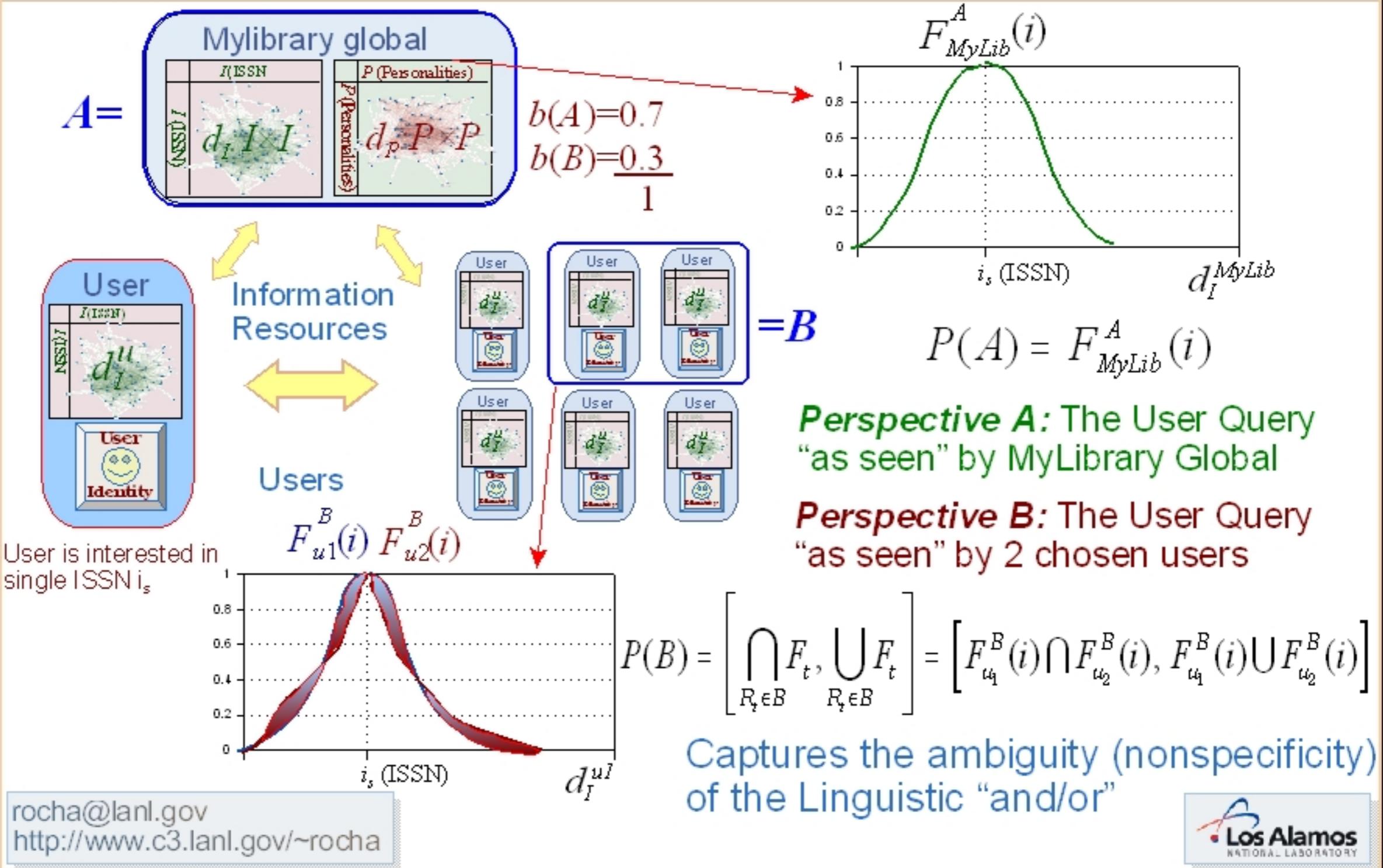
distributed conversation



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

TalkMine for MyLibrary

querying



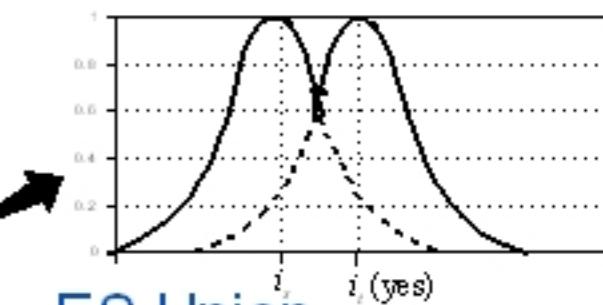
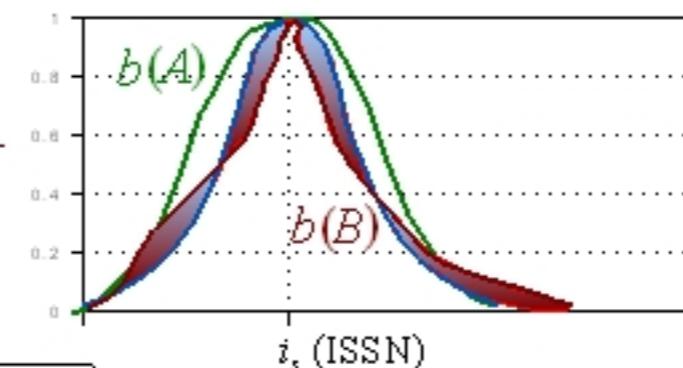
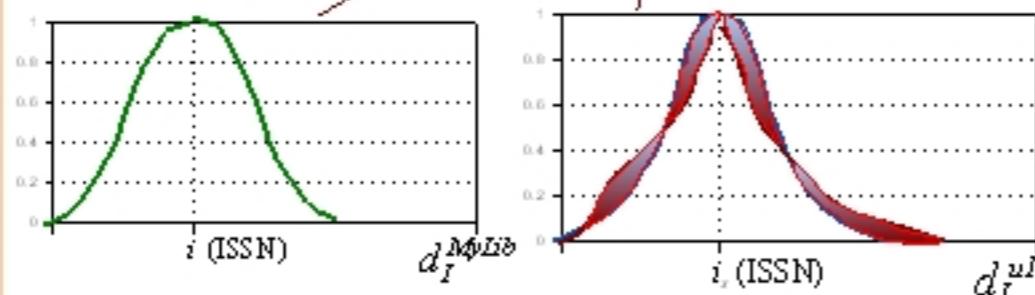
TalkMine: Categorization

interactive spreading of interest on multiple resources

Luis Rocha
2004

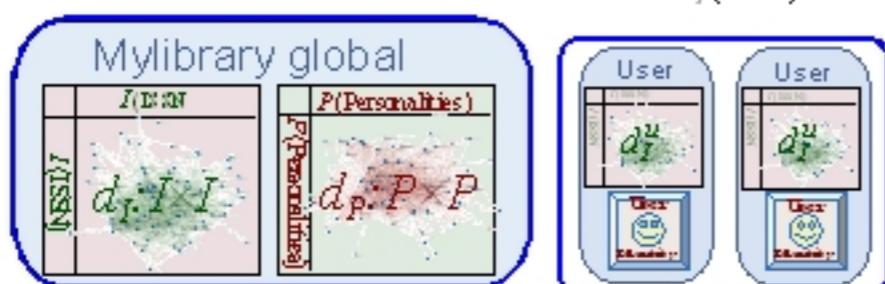
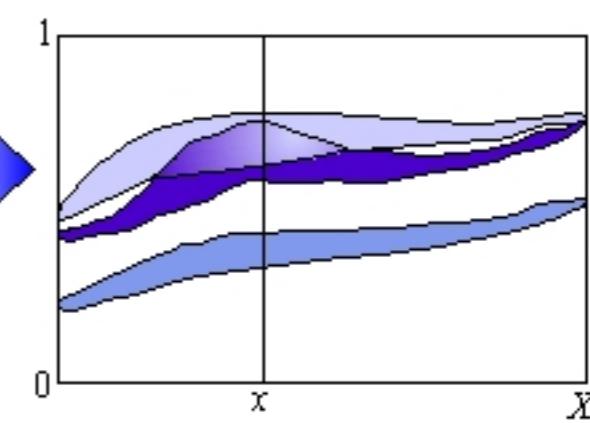
Spreading Interest

Evidence Set



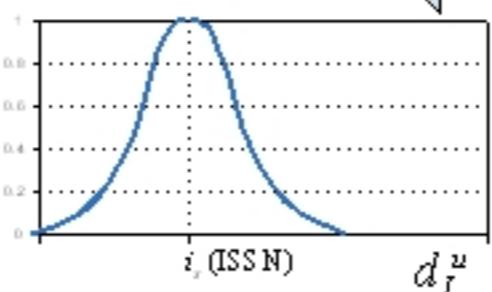
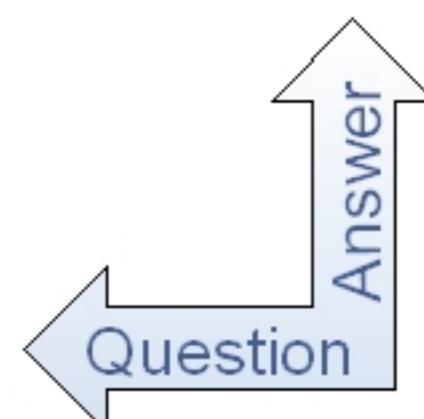
Uncertainty Reducing Process

ES Intersection



A

B



Final Category of Items of Interest

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

adaptation of associative networks



Luis Rocha
2004



Long-Term Memory

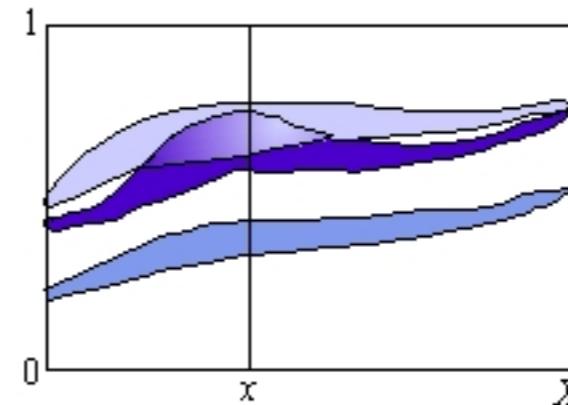
New ISSN are
recognized by
Information Resources

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

(ISSN Personality Proximity)

Short-Term Categorization

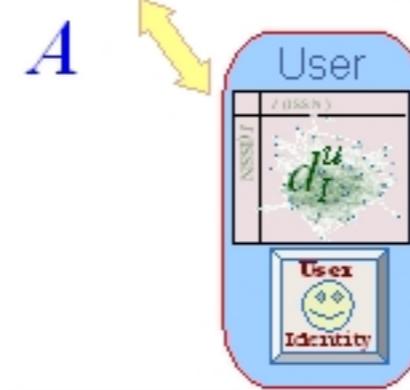
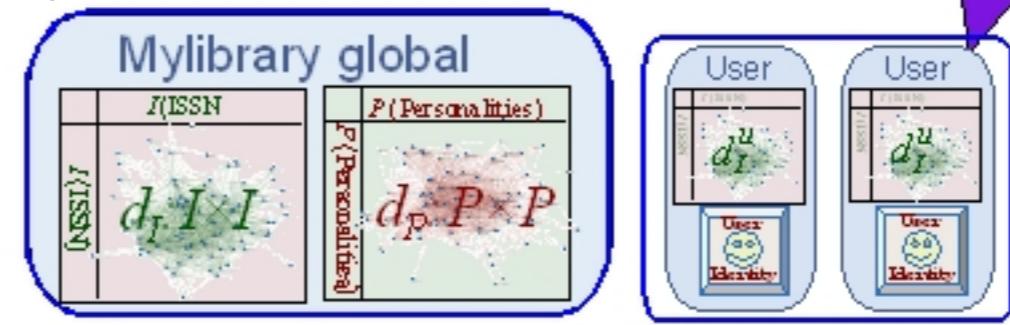
Categorization



No Categories are stored as such, but rather constructed in conversation between agents and information resources, leading to agent identification

Final Category

Adaptation



Set of
Information
Resources

$$N_{t+1}^A(i_s) = N_t^A(i_s) + w \cdot ES(i_s)$$

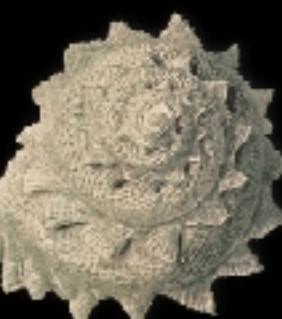
$$N_{t+1}^A(i_s \cap i_t) = N_t^A(i_s \cap i_t) + w \cdot \min[ES(i_s), ES(i_t)]$$

Publications

- Rocha, Luis M. [1991]. "Fuzzification of Conversation Theory." In: *Principia Cybernetica Conference*, Free University of Brussels, Brussels, June 1991. Ed. Francis Heylighen.
- Rocha, Luis M. [1994]. "Cognitive Categorization revisited: extending interval valued fuzzy sets as simulation tools for concept combination." In: *Proceedings of the 1994 International Conference of NAFIPS/IFIS/NASA*. IEEE Press. pp 400-404.
- Rocha, Luis M. [1995]. "Interval Based Evidence Sets." *Proceedings of the ISUMA-NAFIPS'95*. B. Ayyub (Ed.). IEEE Press. pp.624-629.
- Rocha, Luis M. [1996]. "Relative Uncertainty: Measuring Uncertainty in Discrete and Nondiscrete Domains". In: *Proceedings of the NAFIPS'96*. M. Smith et al (Eds). IEEE Press, pp. 551-555.
- Rocha, Luis M., V. Kreinovich, and R. Kearfott [1996]. "Computing Uncertainty in Interval Based Sets." In: *Applications of Interval Computation*. R.B. Kearfott and V. Kreinovich (Eds.). Kluwer Academic Publishers. pp.337-380.
- Rocha, Luis M. [1997]. "Evidence Sets: Contextual Categories". In: *Proceedings of the meeting on Control Mechanisms for Complex Systems*, New Mexico State University, Las Cruces, New Mexico, January 1997. M. Coombs (ed.). NMSU Press, pp. 339-357.
- Rocha, Luis M. [1997a]. *Evidence Sets and Contextual Genetic Algorithms: Exploring Uncertainty, Context, and Embodiment in Cognitive and Biological Systems*. PhD Dissertation. State University of New York at Binghamton.
- Rocha, Luis M. [1997]. "Relative Uncertainty and Evidence Sets: A Constructivist Framework." *International Journal of General Systems*. Vol. 26 (1-2), pp. 35-61.
- Rocha, Luis M. [1999]. "Evidence Sets: Modeling Subjective Categories." *Int. Journal of General Systems*. Vol. 27, pp. 457-494.
- Rocha, Luis M. [1999]. "TalkMine and the Adaptive Recommendation Project". In: *Proceedings of the Association for Computing Machinery (ACM) - Digital Libraries 99*. U.C. Berkely, August 1999, pp. 242-243.
- Rocha, Luis M. [2001]. "Adaptive recommendation and open-ended semiosis". *Kybernetes*. Vol. 30, No. 5/6, pp. 821-851.
- 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. [2001]. "TalkMine: a Soft Computing Approach to Adaptive Knowledge Recommendation". In: *Soft Computing Agents: New Trends for Designing Autonomous Systems*. Vincenzo Loia and Salvatore Sessa (Eds.). Physica-Verlag, Springer, pp. 89-116.
- 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. [2003]. "Automatic Conversation Driven by Uncertainty Reduction and Combination of Evidence for Recommendation Agents ". In: *Systematic Organization of Information in Fuzzy Systems*. NATO Science Series. P. Melo-Pinto, H.N. Teodorescu and T. Fukuda (Eds.) IOS Press, pp 249-265.



Luis Rocha
2004



comparison

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

■ P_X and P_Y are not symmetric

- ▶ May measure a strong degree of association between two elements, when that association is one-sided only.
- ▶ In many applications, when we think of a strong association between two elements, we expect both directions of association to be similar.
- ▶ Distances are symmetric, proximity is the semantic inverse of distance.

$$XYP(x_i, x_j) = \frac{1}{\frac{1}{P_X(x_j | x_i)} + \frac{1}{P_X(x_i | x_j)} - 1}; \quad YXP(y_i, y_j) = \frac{1}{\frac{1}{P_Y(y_j | y_i)} + \frac{1}{P_Y(y_i | y_j)} - 1}$$



pointwise mutual information

comparison

$$PMI(x_i, x_j) = \log \left[\frac{P(x_i \wedge x_j)}{P(x_i) \cdot P(x_j)} \right]$$

comparision between the observed co-occurrence probability for the two words, compared with the co-occurrence probability one would expect to see if the two words were independent

If the two words occur together exactly as frequently as one would expect by chance, $PMI = 0$; if they occur more frequently than one would expect by chance, $PMI > 0$; and conversely if they occur less frequently than one would expect by chance, $PMI < 0$.

$$\frac{P(x_i \wedge x_j)}{P(x_i) \cdot P(x_j)} = \frac{N(x_i \wedge x_j) \cdot N_{Total}}{N(x_i) \cdot N(x_j)} = \frac{N_{Total}}{N(x_i \wedge x_j)} \cdot P(x_i | x_j) \cdot P(x_j | x_i)$$

Not a pobability measure and is dependent on the size of the universal set

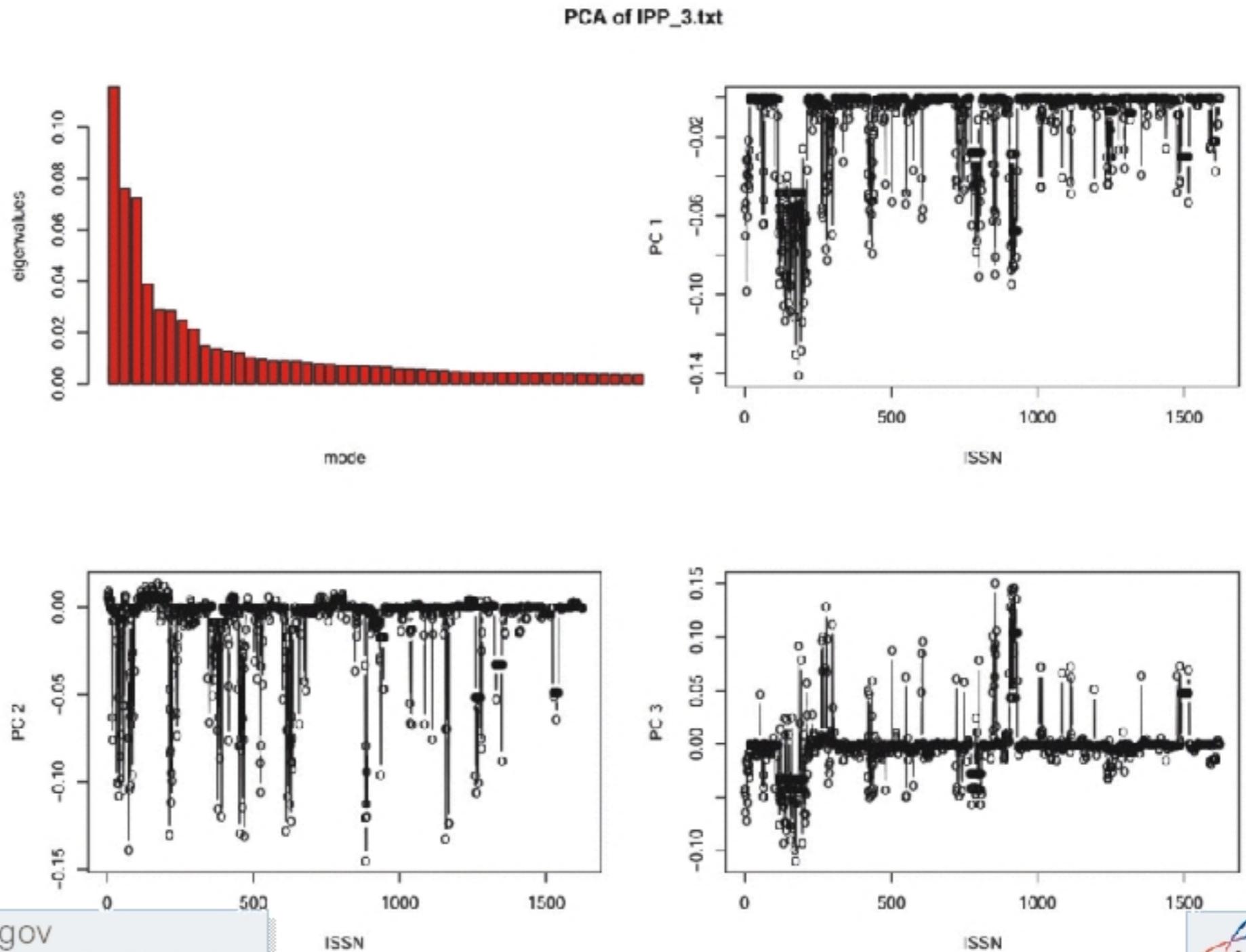
$$XYP(x_i, x_j) = \frac{1}{\frac{1}{P_X(x_j | x_i)} + \frac{1}{P_X(x_i | x_j)} - 1}; \quad YXP(y_i, y_j) = \frac{1}{\frac{1}{P_Y(y_j | y_i)} + \frac{1}{P_Y(y_i | y_j)} - 1}$$

eigen analysis of IPP

Andreas Rechtsteiner and Luis Rocha



Luis Rocha
2004

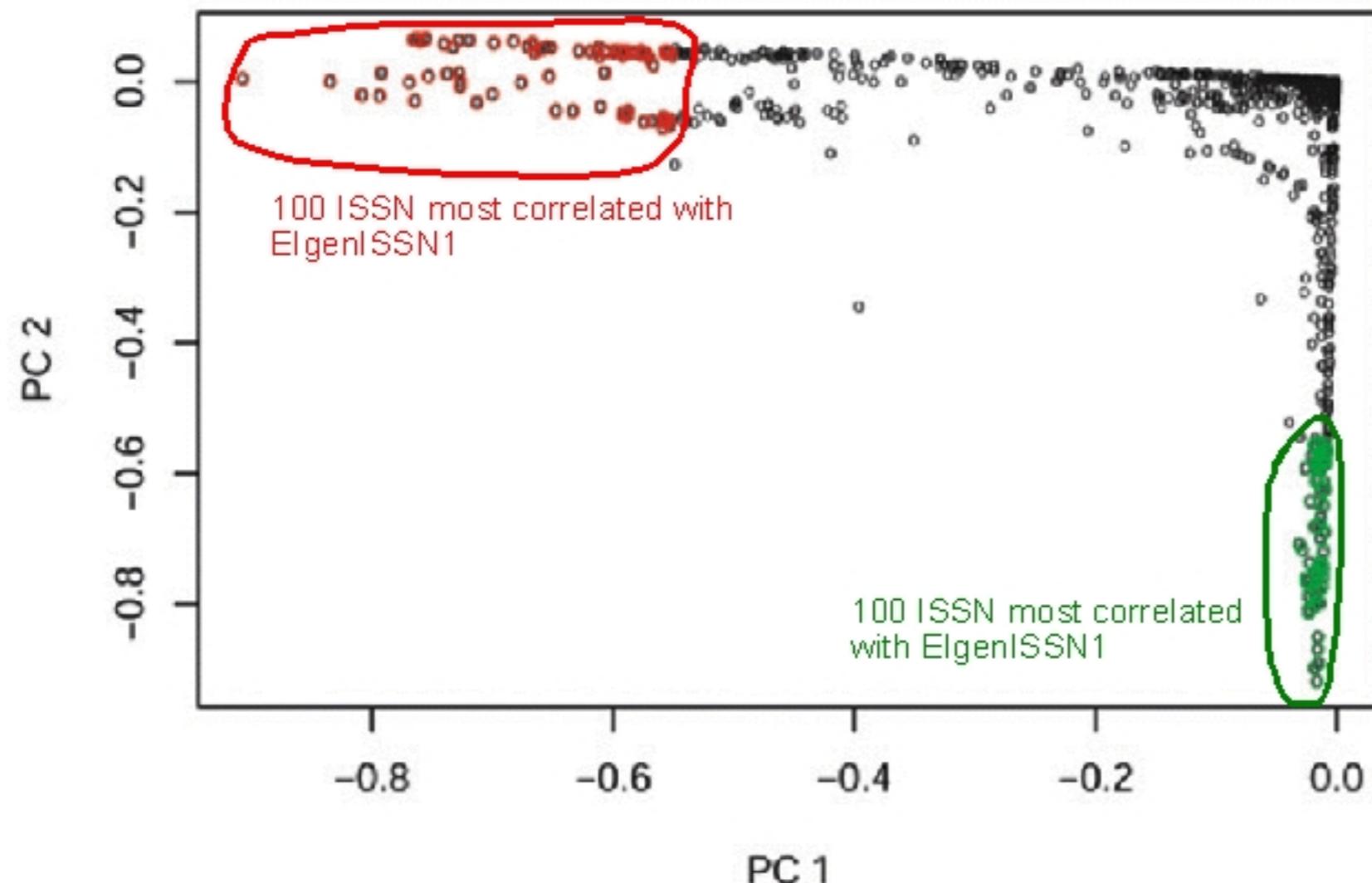


rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



PCA of IPP_3

First 2 EigenISSN



- ISSN correlated with either one of the first two EigenISSN, or with neither
 - ▶ Very few ISSN correlated with both.
 - ▶ 2 very clear main groups

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



Luis Rocha
2004



top 30 ISSN most correlated with eigenISSN 1 and 2

EigenISSN 1: Chemistry and Materials Science

Nanostructured materials 142
Journal of materials chemistry 160
International Journal of Solids and Structures 192
International journal of refractory metals & hard materials 183
Advanced functional materials 136
Applied organometallic chemistry 211
Journal of organic chemistry 154
Annual review of materials science 210
Polyhedron 156
Nano Letters 118
Tetrahedron 196
Topics in current chemistry 173
Radiation effects and defects in solids 121
European journal of inorganic chemistry 6
International journal of inorganic materials 150
Progress in solid state chemistry 202
Chemical reviews 133
Biophysical chemistry 140
Advanced Materials 198
Nuclear energy 187
Advanced engineering materials 135
Journal of Raman spectroscopy 162
Intermetallics 800
Chemistry of materials 425
Chemical Society reviews 134
Journal of biomolecular structure & dynamics 137
Bioorganic chemistry 139
Journal of biochemical and biophysical methods 206
Journal of inorganic biochemistry 170
Chempyschem 789

EigenISSN 2: Computer Science and Applied Mathematics

Algorithmica 74
Science of computer programming 470
Discrete applied mathematics 884
ACM journal of experimental algorithms 213
Mathematical programming 629
Applied mathematics and computation 80
ACM transactions on modeling and computer simulation 87
Computer languages 887
Journal of the Association for Computing Machinery 454
SIAM journal on computing 222
Mathematical and computer modelling 618
Mathematics of computation 611
ACM computing surveys 58
ACM letters on programming languages and systems 219
Computers & mathematics with applications 623
Journal of computational and applied mathematics 378
Theoretical computer science 889
IEEE computational science & engineering 76
Computational mathematics and modeling 524
SIAM journal on scientific computing 220
Applied scientific research 1261
ACM trans. on programming languages and systems 1263
Journal of graph algorithms and applications 883
Discrete mathematics 885
IEEE computer graphics and applications 1167
Computers & graphics 1168
Computer graphics 1169
ACM transactions on graphics 47
Communications on pure and applied mathematics 391
Computers & security 41



Luis Rocha
2004

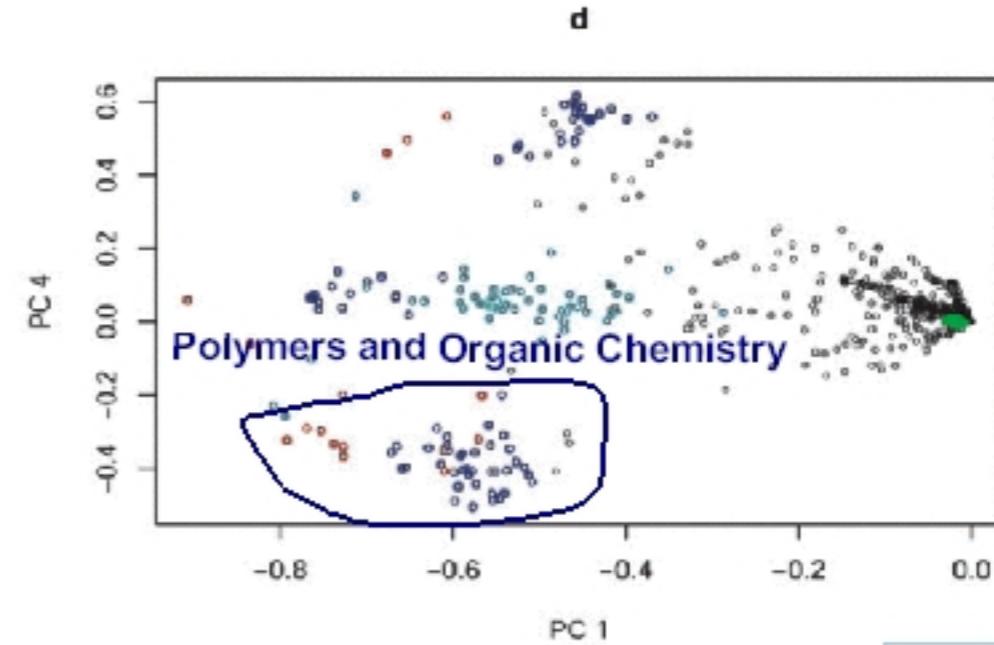
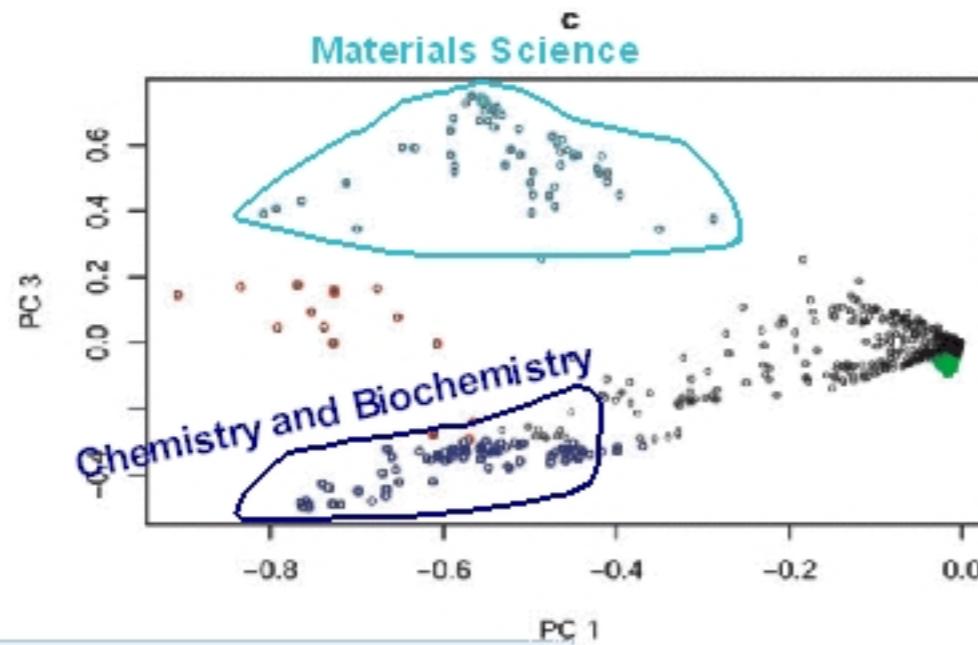
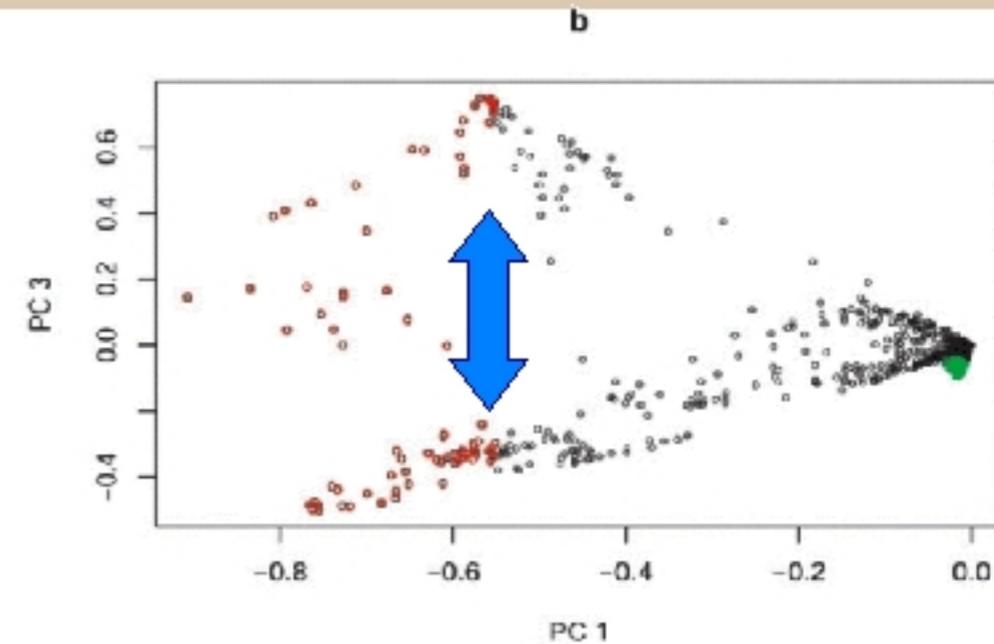
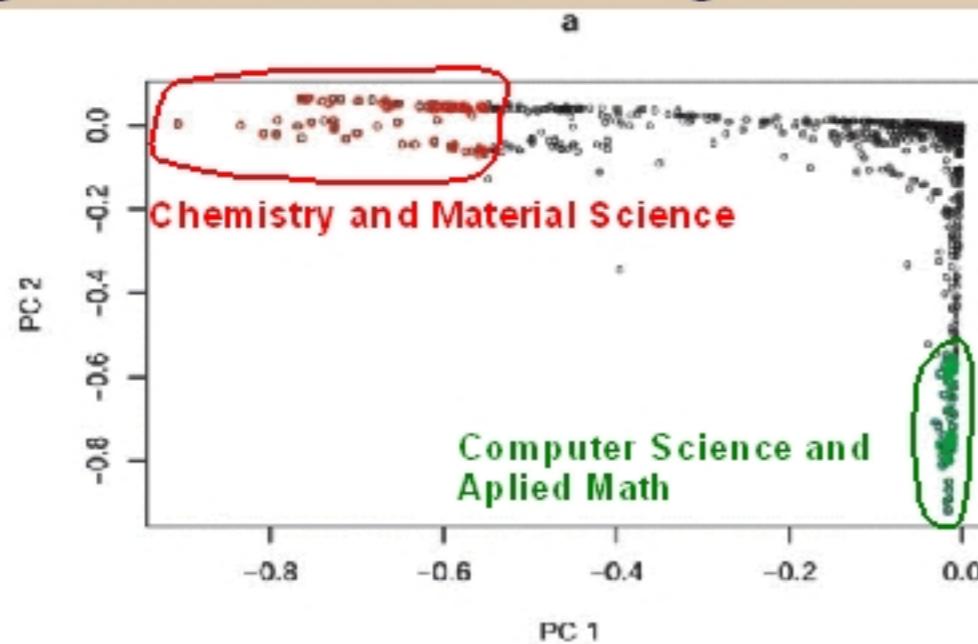


rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



PCA of IPP_3

EigenISSN1 with other EigenISSN



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

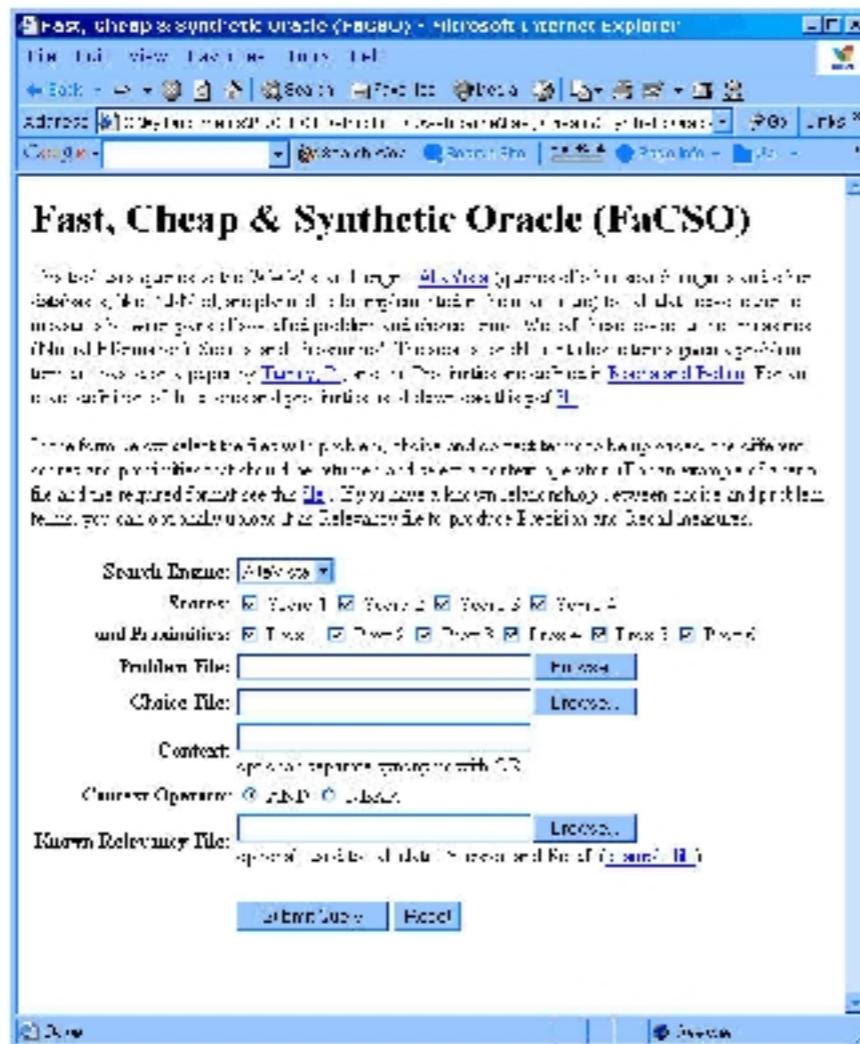


Luis Rocha
2004



co-occurrence web utility

Mining the “Weblome” and “Bibliome”



The screenshot shows a Microsoft Internet Explorer window with the title "Fast, Cheap & Synthetic Oracle (FaCSO) - Microsoft Internet Explorer". The address bar contains the URL "http://bimmer.c3.lanl.gov/~andreas/facso_beta.html". The page content is titled "Fast, Cheap & Synthetic Oracle (FaCSO)" and includes a detailed description of the tool's purpose and usage. It features several input fields for specifying search parameters: "Search Engine" (set to Altavista), "Proximity" (checkboxes for "Years 1", "Years 2", "Years 3", "Years 4", "Prox 2", "Prox 3", "Prox 4", "Prox 5", "Prox 6"), "Prohibited File" (text input), "Choice File" (text input), "Context" (text input with placeholder "specify separate query term with OR"), "Context Operator" (radio buttons for "AND" or "OR"), and "Known Relevance File" (text input). Below these fields are two buttons: "Submit Query" and "Reset".

Rocha, Luis M. and Andreas Rechtsteiner [2003]. "Fast Cheap and Synthetic Oracle (FaCSO): Proximity Measures to capture Expert Knowledge in the Bibliome". *Pacific Symposium on Biocomputing 2003*.

- Discovers Relevant Associations in Altavista and PubMed
 - ▶ Using co-occurrence proximity
 - ▶ How often two sets of words co-occur (near) in documents
- Retrieves Documents substantiating the associations

Fast, Cheap & Synthetic Oracle (FaCSO)

http://bimmer.c3.lanl.gov/~andreas/facso_beta.html

rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>

With Andreas Rechtsteiner



Luis Rocha
2004

FaCSO on Altavista

Cytokines

IL-2 OR interleukin-2
IL-3 OR interleukin-3
IL-4 OR interleukin-4
IL-6 OR interleukin-6
IL-7 OR interleukin-7
IL-8 OR interleukin-8
IL-10 OR interleukin-10
IL-12 OR interleukin-12
IL-13 OR interleukin-13
IL-15 OR interleukin-15
GM-CSF
IFNgamma
TNFalpha
MCP-1

Receptor Molecules

CD25 OR tac
CD122
CD132 OR "common gamma chain"
CD123
beta
CD124
CD126
CD130 OR gp130
CD127
CXCR1 OR Cdw128a
CXCR2 OR Cdw128b
Cdw210
CD212
CD213a1
CD213a2
CD116
CD119
CD120a
CD120b
CCR2b

Signaling molecules:
their levels affect
response

4 Scores using Altavista

Turney, P.D. (2001). "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL." *ECML-2001*, pp. 491-502.

Luis Rocha
2004

1. Conditional Probability with Altavista AND (when both terms appear in the same document)

$$\text{score}_1(\text{choice}_i) = \frac{\text{hits}(\text{problem AND choice}_i)}{\text{hits}(\text{choice}_i)}$$

2. Conditional probability with Altavista NEAR (when both terms appear within 10 words of each other in the same document)

$$\text{score}_2(\text{choice}_i) = \frac{\text{hits}(\text{problem NEAR choice}_i)}{\text{hits}(\text{choice}_i)}$$

3. Tends to reduce equal scores for synonyms and antonyms

$$\text{score}_3(\text{choice}_i) = \frac{\text{hits}(\text{problem NEAR choice}_i) \text{AND NOT}((\text{problem OR choice}_i) \text{ NEAR "not"})}{\text{hits}(\text{choice}_i \text{ AND NOT}(\text{choice}_i \text{ NEAR "not"}))}$$

4. Accounts for context words

$$\text{score}_4(\text{choice}_i) = \frac{\text{hits}((\text{problem NEAR choice}_i) \text{ AND context AND NOT}((\text{problem OR choice}_i) \text{ NEAR "not"}))}{\text{hits}(\text{choice}_i \text{ AND context AND NOT}(\text{choice}_i \text{ NEAR "not"}))}$$

$$\text{context} = \{\text{context}_1, \text{context}_2, \dots, \text{context}_m\}$$

We use NEAR



Luis Rocha
2004



$$prox_1(problem, choice_i) = \frac{\text{hits}(problem \text{ AND } choice_i)}{\text{hits}(problem \text{ OR } choice_i)}$$

$$prox_2(problem, choice_i) = \frac{\text{hits}(problem \text{ NEAR } choice_i)}{\text{hits}(problem \text{ OR } choice_i)}$$

NEAR: co-occurrence within 10 words

$$prox_3(problem, choice_i) = \frac{\text{hits}((problem \text{ NEAR } choice_i) \text{ NEAR } context)}{\text{hits}(problem \text{ OR } choice_i)}$$

context = {receptor}

possible distances

shortest paths or weakest links

Proximity

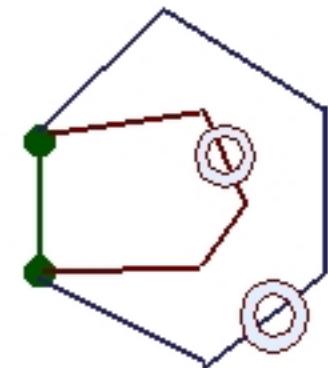
	$X(\text{Keywords})$
$X(\text{Keywords})$	$XYP : X \times X$

Transitive Closure
(max/min)

Edges: largest of the
weakest links in all
indirect paths:

Similarity

	$X(\text{Keywords})$
$X(\text{Keywords})$	$XYS : X \times X$



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

	$X(\text{Keywords})$
$X(\text{Keywords})$	$d_X : X \times X$

Shortest
Path
(min/+)

Distance

\leq metric
 \geq semi-metric

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

Edges: shortest
indirect path (sums
all edges)



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

Edges: Smallest of the
largest edges in each
indirect path



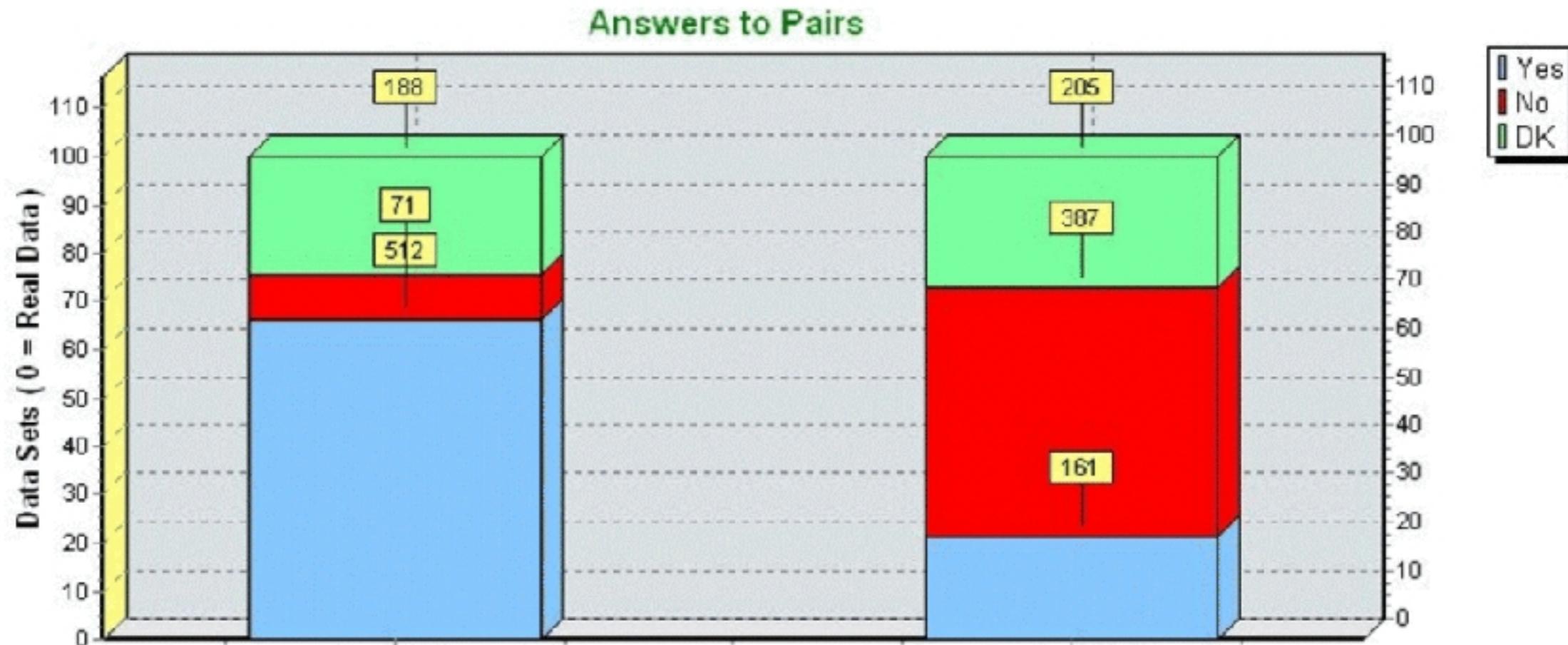
Luis Rocha
2004



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



population from STB-RL and CCS-3



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



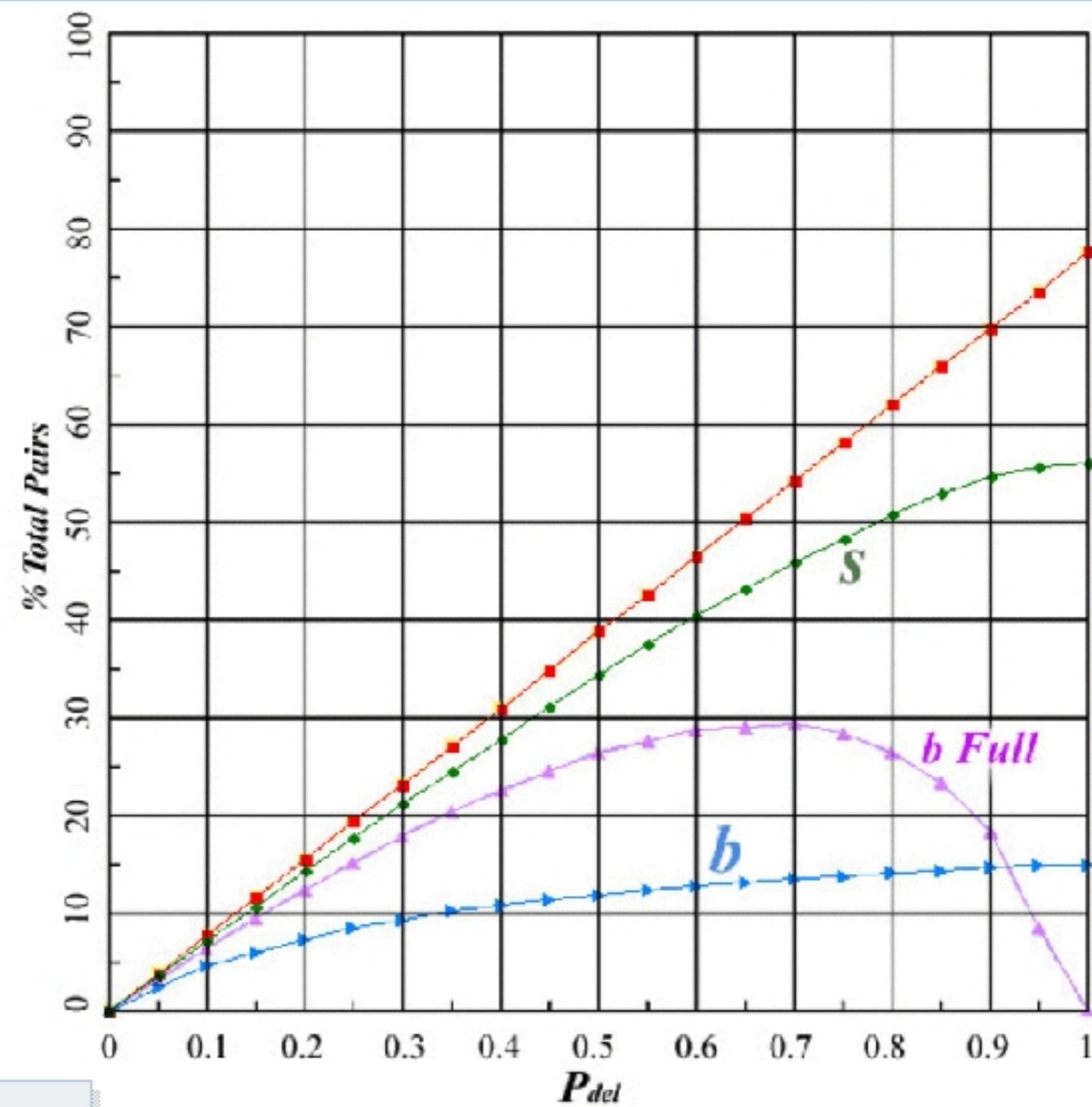
partial deletion



Luis Rocha
2004



rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



set-theoretic operations

for evidence sets

Complement $A^c(x) = \left\{ \langle (I_1(x))^c, m_1(x) \rangle, \langle (I_2(x))^c, m_2(x) \rangle, \dots, \langle (I_n(x))^c, m_n(x) \rangle \right\}$

$$A(x) = \left\{ \langle I_1(x), m_1(x) \rangle, \langle I_2(x), m_2(x) \rangle, \dots, \langle I_n(x), m_n(x) \rangle \right\}$$

$$I(x) = [I_L(x), I_U(x)] \subseteq [0,1] \quad I^c(x) = [1 - I_U(x), 1 - I_L(x)] \subseteq [0,1]$$

Intersection $m_C^x(K_k^x) = \sum_{MIN(I_i^x, J_j^x) = K_k^x} m_A^x(I_i^x) \cdot m_B^x(J_j^x)$

Union $m_C^x(K_k^x) = \sum_{MAX(I_i^x, J_j^x) = K_k^x} m_A^x(I_i^x) \cdot m_B^x(J_j^x)$

$$I = [I_L, I_U], J = [J_L, J_U]$$

$$K = MIN(I, J) = [MIN(I_L, J_L), MIN(I_U, J_U)]$$

$$K = MAX(I, J) = [MAX(I_L, J_L), MAX(I_U, J_U)]$$

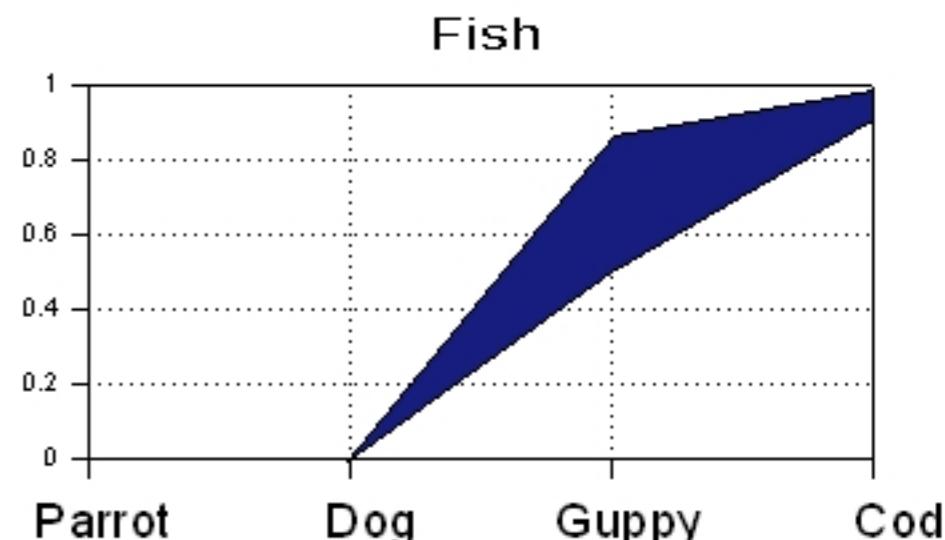
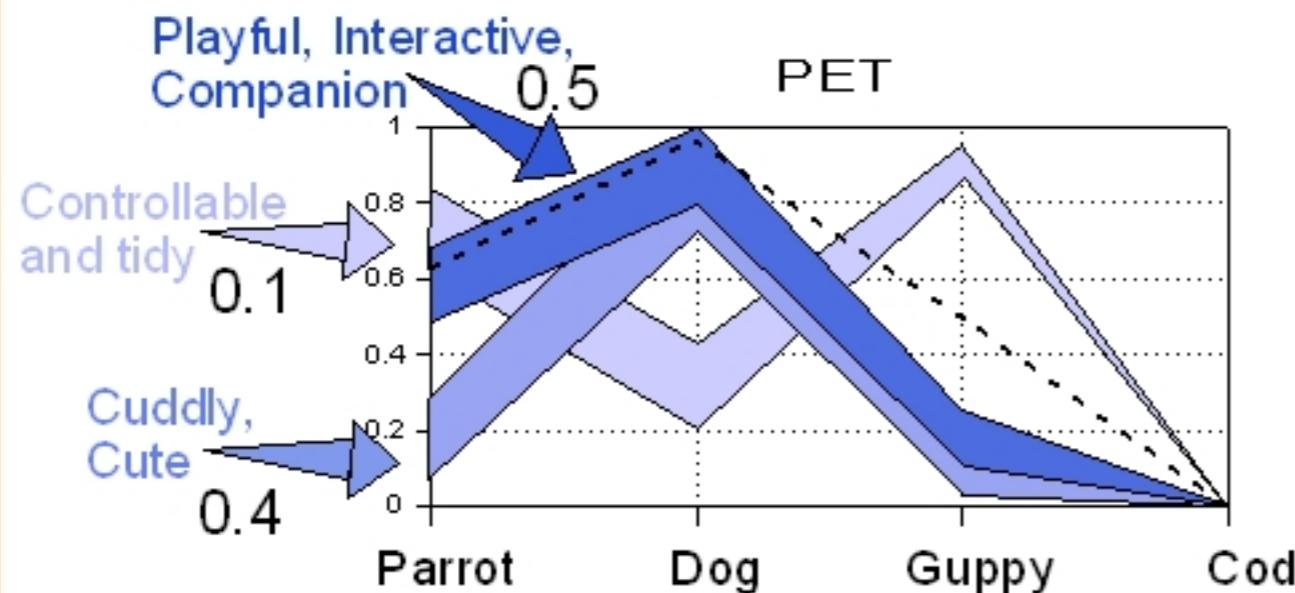
rocha@lanl.gov
<http://www.c3.lanl.gov/~rocha>



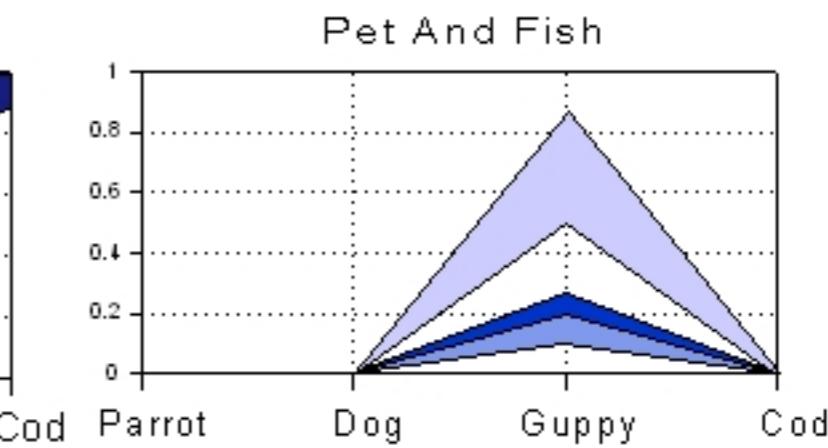
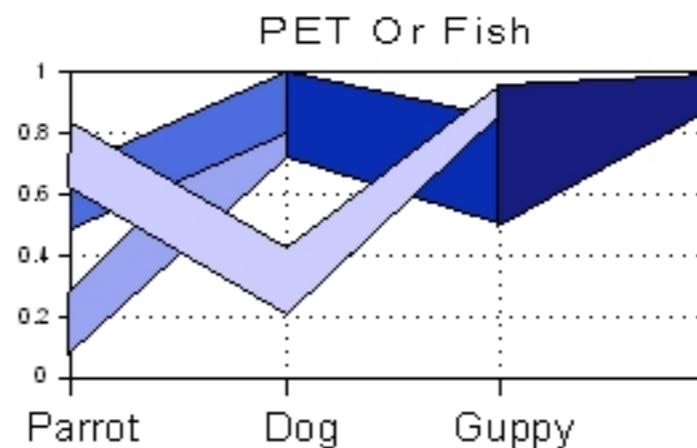
Luis Rocha
2004

evidence set combination

pet-fish example



Set-Theoretic Combinations



TalkMine: Inferring User Interest

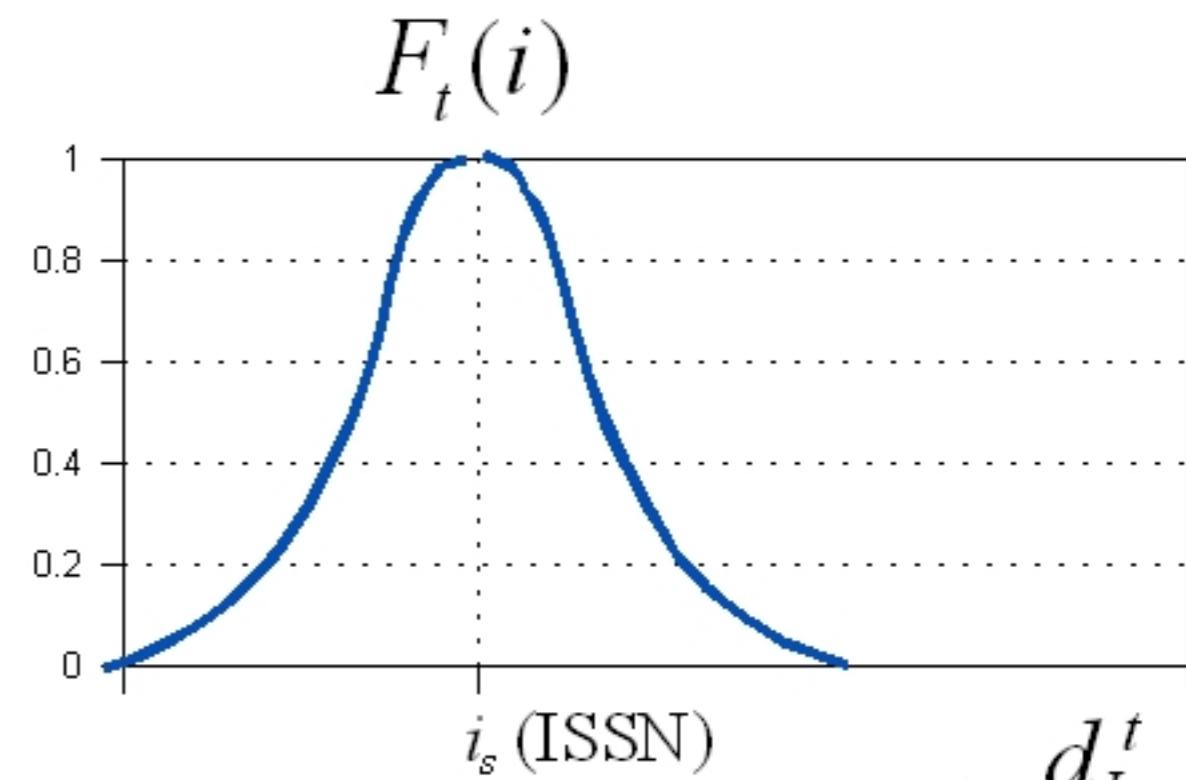
Spreading Interest on Distance Graphs

$$F_t(i) = \bigcup_{s=1}^{m_\pi} \max \left[e^{\left(-\alpha \cdot d_I^t(i, i_s)^2 \right)}, \varepsilon \right], \forall_{i \in I}, t = 1 \dots r, i_s \in \pi$$

For each information resource R_t and the set of ISSN i_s in the user's present interests, a spreading interest fuzzy subset F_t of I is created using d_I^t :

- User Present Interests
 - ▶ Set of ISSN
 - Clicks Link
 - Updates Folder
 - Updates Personality
 - Updates Library

User is interested in
single ISSN i_s



The Interests of the User

Combination of several Perspectives from distinct sets of information Resources into an Evidence Set

Given all the perspectives (IVFS) $P(A)$, and probabilistic weights $b(A)$

$$ES(i_s) = \langle (P(A), b(A)) \rangle, \forall A \in \mathcal{F}$$

