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
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)

$$d(x_i, x_i) = 0$$

$$d(x_i, x_j) = 1, \text{ if there is an edge}$$

$$d(x_i, x_k) = d(x_i, x_j) + \ldots + d(x_i, x_k) \leq 1,$$

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
Fuzzy sets

operations

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

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

\[ A \cap A \neq \emptyset \]
\[ A \cup \overline{A} \neq X \]

Fuzziness
Degree of Membership/Truth

Fuzziness

Union

Intersection
Fuzzy Sets:

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

\( \alpha \)-cut: Crisp set at threshold \( \alpha \)
De Morgan’s Laws

\[ A \cap B = \overline{A} \cup \overline{B} \]
\[ A \cup B = \overline{A} \cap \overline{B} \]

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

\[ u(a, b) = a + b - ab \]

- **Complement, Intersection and Union that follow De Morgan’s Laws plus**
  - **Complement**
    - Boundary Conditions: \( c(0)=1 \) and \( c(1)=0 \)
    - Monotonicity: if \( a \leq b \) then \( c(a) \geq c(b) \)
    - Continuity
    - Involutivity: \( c(c(a)) = a \)
  - **Intersection (T-Norm)**
    - Boundary condition: \( i(a, 1) = a \)
    - Monotonicity: if \( b \leq d \) then \( i(a,b) \leq i(a,d) \)
    - Comutativity: \( i(a,b) = i(b,a) \)
    - Associativity: \( i(a,i(b,d))=i(i(a,b),d) \)
    - Continuity
    - Strict Monotonicity: if \( a_1 < a_2 \) and \( b_1 < b_2 \) then \( i(a_1,b_1) < i(a_2,b_2) \)
    - Subidempotency: \( i(a,a) \leq a \)
  - **Union (T-Conorm)**
    - Boundary condition: \( u(a, 0) = a \)
    - Monotonicity: if \( b \leq d \) then \( u(a,b) \leq u(a,d) \)
    - Comutativity: \( u(a,b) = u(b,a) \)
    - Associativity: \( u(a,u(b,d))=u(u(a,b),d) \)
    - Continuity
    - Strict Monotonicity: if \( a_1 < a_2 \) and \( b_1 < b_2 \) then \( u(a_1,b_1) < u(a_2,b_2) \)
    - Superidempotency: \( u(a,a) > a \)

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

Sugeno Complement:
\[ \lambda \in (-1, \infty) \]
Represent the presence or absence of association, interaction or interconnectedness between the elements of two or more sets.

- 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

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

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

$$R(x) \in [0, 1], \quad \forall x \in X$$  
**Fuzzy:** Degree of Relation or association
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
fuzzy graphs

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 $(n^2-n)/2$ 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

$a < b < c$

$a < d < e$
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:

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

http://arp.lanl.gov
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 $x \in X$.

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

With some support constraint

proximity measures

produce associative (probabilistic) networks

\[
XYP(x_i, x_j) = \frac{\sum_{k=1}^{m} (r_{i,k} \land r_{j,k})}{\sum_{k=1}^{m} (r_{i,k} \lor r_{j,k})}
\]

\[
YXP(y_i, y_j) = \frac{\sum_{k=1}^{n} (r_{k,i} \land r_{k,j})}{\sum_{k=1}^{n} (r_{k,i} \lor r_{k,j})}
\]

represents knowledge in associative manner
PDP2

318 names from unclassified database
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.

With Tiago Simas, Andreas Rechtsteiner and Chien-feng Huang
**most frequent ISSN**

<table>
<thead>
<tr>
<th>Journal</th>
<th>Occurences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical review letters</td>
<td>53</td>
</tr>
<tr>
<td>Physical Review B</td>
<td>26</td>
</tr>
<tr>
<td>Physical review E</td>
<td>23</td>
</tr>
<tr>
<td>Physical review A General physics</td>
<td>22</td>
</tr>
<tr>
<td>Journal of physical chemistry B</td>
<td>20</td>
</tr>
<tr>
<td>Computers &amp; geosciences</td>
<td>20</td>
</tr>
<tr>
<td>Scientific American</td>
<td>19</td>
</tr>
<tr>
<td>Journal of the American Chemical Society</td>
<td>19</td>
</tr>
<tr>
<td>Journal of Chemical Physics</td>
<td>19</td>
</tr>
<tr>
<td>Reviews of modern physics</td>
<td>19</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>18</td>
</tr>
<tr>
<td>IEEE trans. on geoscience and remote sensing</td>
<td>18</td>
</tr>
<tr>
<td>PNAS</td>
<td>18</td>
</tr>
<tr>
<td>Journal of computational physics</td>
<td>18</td>
</tr>
<tr>
<td>Advances in water resources</td>
<td>18</td>
</tr>
<tr>
<td>Journal of applied geophysics</td>
<td>17</td>
</tr>
<tr>
<td>Applied geochemistry</td>
<td>17</td>
</tr>
<tr>
<td>APL</td>
<td>17</td>
</tr>
<tr>
<td>Journal of physical chemistry A</td>
<td>17</td>
</tr>
<tr>
<td>Phil. mag. B Physics of condensed matter</td>
<td>17</td>
</tr>
<tr>
<td>Bull. of Environmental Contamination and Toxicology</td>
<td>16</td>
</tr>
<tr>
<td>Journal of applied physics</td>
<td>15</td>
</tr>
<tr>
<td>American journal of physics</td>
<td>14</td>
</tr>
<tr>
<td>Analytical chemistry</td>
<td>14</td>
</tr>
<tr>
<td>DLib</td>
<td>14</td>
</tr>
<tr>
<td>Chemical physics</td>
<td>13</td>
</tr>
<tr>
<td>NIM</td>
<td>13</td>
</tr>
<tr>
<td>Chemical physics letters</td>
<td>12</td>
</tr>
<tr>
<td>Physics reports</td>
<td>12</td>
</tr>
<tr>
<td>Physical review A</td>
<td>12</td>
</tr>
<tr>
<td>Nature</td>
<td>11</td>
</tr>
<tr>
<td>Physics of plasmas</td>
<td>11</td>
</tr>
<tr>
<td>Accounts of chemical research</td>
<td>11</td>
</tr>
<tr>
<td>Advanced Materials</td>
<td>11</td>
</tr>
<tr>
<td>Journal of physics Condensed matter</td>
<td>11</td>
</tr>
<tr>
<td>Journal of computational biology</td>
<td>11</td>
</tr>
<tr>
<td>Biochemical journal</td>
<td>11</td>
</tr>
<tr>
<td>Computer physics communications</td>
<td>11</td>
</tr>
<tr>
<td>ACM transactions on modeling and computer simulation</td>
<td>10</td>
</tr>
<tr>
<td>Materials science &amp; engineering A</td>
<td>10</td>
</tr>
<tr>
<td>Physical review C Nuclear physics</td>
<td>10</td>
</tr>
<tr>
<td>Advances in physics</td>
<td>10</td>
</tr>
<tr>
<td>Inorganic chemistry</td>
<td>10</td>
</tr>
<tr>
<td>Trends in BioTechnology</td>
<td>10</td>
</tr>
<tr>
<td>Review of Scientific Instruments</td>
<td>10</td>
</tr>
<tr>
<td>Science</td>
<td>10</td>
</tr>
</tbody>
</table>
ISSN and personality proximity

from co-occurrence in mylibrary.lanl.gov

<table>
<thead>
<tr>
<th>Personality</th>
<th>A·P×I</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISSN</td>
<td></td>
</tr>
</tbody>
</table>

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) $p \in P$.

$$pip(p_s, p_t) = \frac{\sum_{k=1}^{m} (a_{i,k} \land a_{j,k})}{\sum_{k=1}^{m} (a_{i,k} \lor a_{j,k})} = \frac{N_{\cap}(P_s, P_t)}{N_{\cup}(P_s, P_t)}$$

(Personality ISSN Proximity)

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

(ISSN Personality Proximity)
journal network
from co-occurrence in user personalities in mylibrary.lanl.gov: IPP
cumulative degree distribution

All weights
journal network
from co-occurrence in user personalities in mylibrary.lanl.gov: IPP

Computer Science and Mathematics

IPP ≥ 0.5

Immunology

Materials Science and Chemistry
cumulative degree distribution

$\alpha$-cut = 0.2

$\alpha$-cut = 0.4
\( \alpha\text{-cut} = 0.2 \)

\( \alpha\text{-cut} = 0.4 \)
top 5 most frequent ISSN and their neighbors

\[ ipp(i_s, i_t) = \frac{N_\cap(i_s, i_t)}{N_U(i_s, i_t)} \geq 3 \]

\[ ipp \geq 0.3 \]

**0031-9007 -- Physical review letters**
- 1095-3787 -- Physical review E: 0.4074
- 0566-2805 -- Physical Review B: 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
- 0566-2791 -- 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

**0566-2791 -- Physical review A General physics**
- 0031-9007 -- Physical review letters: 0.3636
- 0370-1573 -- Physics reports: 0.3077
- 0566-2805 -- 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

<table>
<thead>
<tr>
<th>IPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendations of ISSN based on co-occurrence in Personalities</td>
</tr>
<tr>
<td>– Users who linked to this journal, also linked to...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendations of other users’ personalities: collaboration</td>
</tr>
<tr>
<td>– These personalities are similar to yours</td>
</tr>
<tr>
<td>Recommendations of specific links in close personalities</td>
</tr>
<tr>
<td>– Users who read many of the same journals where interested in these links</td>
</tr>
</tbody>
</table>


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: same


rocha@indiana.edu
http://informatics.indiana.edu/rocha
GO id word expansion

Based on proximity measure

For each Document

1. \( P: P \times W \)

2. \( WPP: W \times W \)

3. \( W_{GO} = \{w_1, \ldots, w_\alpha\} \)

4. \( W_{GO_{Prox}} = \{w_{1}, \ldots, w_\alpha, w_{\alpha+1}, \ldots, w_\beta\} \)


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)


rocha@indiana.edu
http://informatics.indiana.edu/rocha
### Task 2.1 Results

**Proximity-based run**

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


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

Adaptive Systems

Cognition

Semi-metric ratio: 6.3861

rocha@indiana.edu
http://informatics.indiana.edu/rocha
computing semi-metric behavior

\[ R \times Y \]

\[ X (\text{Keywords}) \]

\[ Y \times Y \]

\[ X (\text{Keywords}) \]

\[ Y \times X \]

\[ X (\text{Keywords}) \]

\[ Y \times Y \]

\[ X (\text{Keywords}) \]

\[ X (\text{Keywords}) \]

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

\[ d_X^* : X \times X \]

\[ X (\text{Keywords}) \]

\[ X (\text{Keywords}) \]

\[ \geq \text{metric} \]

\[ > \text{semi-metric} \]

\[ \text{Shortest Path (min/+) } \]

Measures of semi-metric behavior

rocha@indiana.edu
http://informatics.indiana.edu/rocha
Semi-metric Measures

- **Semi-metric ratio**
  - Absolute measure of indirect distance reduction
  \[ s(x_i, x_j) = \frac{d_{\text{direct}}(x_i, x_j)}{d_{\text{shortest}}(x_i, x_j)} \]

- **Relative Semi-metric ratio**
  - Distance reduction against maximum contraction
  \[ rs(x_i, x_j) = \frac{d_{\text{direct}}(x_i, x_j) - d_{\text{shortest}}(x_i, x_j)}{d_{\text{max}} - d_{\text{min}}} \]

- **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 $\infty$ distance
  \[ b(x_i, x_j) = \frac{\overline{d_{x_i}}}{d_{\text{shortest}}(x_i, x_j)} \]
Pairs with larger semi-metric behavior denote a \textit{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.


semi-metric recommendations

catching strong indirect associations in mylibrary.lanl.gov

IPP_3: parameter $r_s$

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 0828-1045--Journal of computer aided 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
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 $b$

Acknowledgement: John Hogden
shortest paths or weakest links

Comparison with transitivity

Proximity

Edges: largest of the weakest links in all paths:

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

Distance

\[ d_X : X \times X \]

\[ \leq \text{metric} \]

\[ \geq \text{semi-metric} \]

For any monotonic increasing distance function

Similarity

\[ X (\text{Keywords}) \]

\[ XYS : X \times X \]

Edges: Smallest of the largest edges in each path

\[ d**_X : X \times X \]

\[ \geq \]

\[ d^*_X : X \times X \]

Shortest Path (max/min)

Transitive Closure

\[ X (\text{Keywords}) \]

\[ X (\text{Keywords}) \]

\[ X (\text{Keywords}) \]

\[ X (\text{Keywords}) \]

With Tiago Simas
comparison of the two closures

cumulative strength distribution

Shortest Path (min/+)

Similarity

Transitive Closure (max/min)
Hammacher function

If we use:

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

\[ u(a, b) = \max\{a, b\} \]

With Tiago Simas

- Proximity
  - \( \text{metric} \)
  - \( \leq \) semi-metric

- Similarity
  - Trans.Closure
    - \( XYP : X \times X \)
  - \( XYS : X \times X \)

- Distance
  - \( \leq \) metric
  - \( > \) semi-metric

\[ d_X(x_i, x_j) = \frac{1}{XYP(x_i, x_j)} - 1 \]
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?
scientific community working on feynman diagrams

as published in *Physical Review*, 1949-54

<table>
<thead>
<tr>
<th>Collaboration Relation: $C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\triangleright$ Who wrote a paper with whom</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acknowledgment Relation: $A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\triangleright$ Who acknowledged, or informally received information from whom</td>
</tr>
</tbody>
</table>

\[
CP(p_i, p_j) = \frac{\sum_{k=1}^{m} (c_{i,k} \land c_{j,k})}{\sum_{k=1}^{m} (c_{i,k} \lor c_{j,k})}
\]

\[
AP(p_i, p_j) = \frac{\sum_{k=1}^{m} (a_{i,k} \land a_{j,k})}{\sum_{k=1}^{m} (a_{i,k} \lor a_{j,k})}
\]

76 Authors

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

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

With Luis Bettencourt and David Kaiser
**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
5 most semi-metric pairs (rs and b parameters)

- 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
- 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
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?
**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)

*AP* value increases in 1950 and 1953.
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.
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 occurring
  - 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