A Unified Framework for Affine Local Graph Similarity*

T. P. Cason P.-A. Absil V. D. Blondel P. Van Dooren

Université catholique de Louvain, Department of Mathematical Engineering
Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium
http://www.inma.ucl.ac.be/∼cason , {*∼absil, ∼blondel, ∼vandooren*}

***Abstract*— In this work, we review and classify several similarity measures on undirected graphs. We show that these measures can be rewritten in terms of fixed points of a scaled affine transformation. Finally, we propose a novel definition that avoids undesirable degeneracy of the similarity matrix.**

I. Introduction

Node-to-node equivalence in graphs [11], [12], [13] is a yes or no question. It does not carry any information on how “close” two nodes are from being equivalent. The notion of node-to-node similarity remedies this by associating a real valued similarity score to pairs of nodes.

Measures of node similarity in graphs have a broad array of applications, including comparing chemical structures [2], navigating in complex networks like the World Wide Web [3], and analyzing different kinds of biological data [4]. In early work by Balaban [2], chemical compounds are considered to be graphs, whose nodes and edges are respectively atoms and inter-atomic bounds, and graph theory and similarity measure are used to identify isomers. Isomers are molecules characterized by the same graph topology and possibly the same chemical properties. This problem is actually equivalent to the graph matching problem. More recently, Blondel *et al.* [5] consider a similarity measure to extract synonyms in a monolingual dictionary. Holme and Huss [6] predict the function of proteins based on their role in a protein interaction network. Zhou *et al.* [7] use a similarity measure to predict missing links in various benchmarks. Several similarity measures use the idea of reinforcement loops, *i.e.*, the similarity score between two nodes is computed using the similarity scores between other nodes in the network [3], [5], [8], [9], [10].

In this work, we review and classify several similarity measures on undirected graphs. We further show that these measures can be rewritten in terms of fixed points of a scaled affine transformation. Moreover, we propose a novel definition of similarity that possess certain desirable properties.

II. Node-to-Node Similarity Measures

The similarity measures compare the nodes of one graph either with the nodes of an other graph, or with the nodes of the same graph. The node-to-node similarity score is conveniently stored in the so-called similarity matrix, S, whose (i, j) entry tells how the node i is similar to the node j. In essentially all cases, we are not interested in the absolute value of S_{ij} but only in the relative score of two different pairs.

We propose to classify similarity measures as follows:

- **Topological similarity** A similarity measure is termed *topological* if the similarity score S_{ij} equals 0 whenever i and j in a graph G do not belong to the same connected component.
- **Non-topological similarity** A similarity measure is termed *non-topological* if it is not *topological*.

In this paper, we assume throughout that the graphs are undirected. The case of directed graphs will be addressed in future work.

A. Topological Similarity Measures

The most simple requirement is that

$$\boxed{\begin{array}{l}\text{if } i \text{ and } j \text{ have many common neighbors,}\\ \text{then the similarity between } i \text{ and } j \text{ is large.}\end{array}} \qquad \text{(R1)}$$

Given an undirected graph $G_A = (N, E)$, let $\Gamma(i)$ denote the set of neighbors of node i, *i.e.*

$$\Gamma(i) = \{k : (i, k) \in E\} \ .$$

A natural similarity measure is to count the number of *common neighbors* (CN), *i.e.*

$$S_{ij} = |\Gamma(i) \cap \Gamma(j)| \ . \qquad (1)$$

One can notice that CN tends to give higher similarity score to the nodes that have many neighbors. As a result of this,

*This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its author(s).

two different nodes could be more similar than one node is similar to itself. This effect is commonly avoided by weighting this definition

$$S_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{w_{ij}} \ .$$

- The *Jaccard index* was defined over a hundred years ago in [14], and uses
$$w_{ij} = |\Gamma(i) \cup \Gamma(j)| \ . \qquad (2)$$
- The *Salton index* or *cosine similarity* [15] is defined with
$$w_{ij} = \sqrt{|\Gamma(i)| \cdot |\Gamma(j)|} \ . \qquad (3)$$
In [16], Leydesdorff consider the *Jaccard* and *Salton index* to analyze author co-citation networks.
- The *Sørensen index* [17] is defined with
$$w_{ij} = \frac{|\Gamma(i)| + |\Gamma(j)|}{2} \ . \qquad (4)$$
This index is mainly used to analyze ecological community data [18].
- In [8], Leicht *et al.* proposed
$$w_{ij} = |\Gamma(i)| \cdot |\Gamma(j)| \ . \qquad (5)$$
The weight w_{ij} is proportional to the number of expected common neighbors of the nodes i and j in the configuration model.

One can notice that the similarity scores of the weighted definitions are all between 0 and 1.

In all the above definitions, all nodes in $\Gamma(i) \cap \Gamma(j)$ contribute equally to S_{ij}. Other definitions sum different contributions for each node in $\Gamma(i) \cap \Gamma(j)$

$$S_{ij} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} f_{ij}(k) \ ,$$

using a particular scale function $f_{ij}(k)$.

- In [19], Adamic and Adar consider a social network in which individuals are linked to properties based on information collected on their homepages. The more individuals have properties in common, the more they are similar. Moreover, properties unique to a few individuals are weighted more than commonly occurring properties. This leads them to choose
$$f_{ij}(k) = \frac{1}{\log |\Gamma(k)|} \qquad (6)$$
- In [7], Zhou *et al.* propose
$$f_{ij}(k) = \frac{1}{|\Gamma(k)|} \ , \quad \text{or} \quad f_{ij}(k) = \frac{1}{|\Gamma(i)| \cdot |\Gamma(k)|} \qquad (7)$$
Notice that the second definition is not symmetric.

An other possible requirement is that

if there is a path between i and j,
then the similarity between i and j is large. (R2)

- The *topological overlap index* introduced by Ravasz *et al.* in [20] is a weighted measure that combines both requirements (R1) and (R2) with a path of length 1 and is defined as
$$S_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)| + A_{ij}}{\min\left(|\Gamma(i)|, |\Gamma(j)|\right)} \qquad (8)$$
where A is the adjacency matrix of G_A, *i.e.* A_{ij} is 1 if there is an edge from the node i to the node j, and 0 otherwise.
- In [7], Zhou proposes
$$S = A^2 + \alpha A^3 \qquad (9)$$
The first term gives the number of paths of length 2 between two nodes. Notice that for an undirected graph, this is equivalent to the number of common neighbors. The second term is proportional to the number of paths of length 3 between two nodes (R2).

In all the above definitions, the similarity measure between two nodes is exclusively based on direct neighbors. This implies that two nodes distant of more than 2 edges cannot be similar. In [10], Jeh *et al.* propose the *Simrank* similarity measure that extends similarity to a larger neighborhood. The basic idea is to let the similarity scores percolate through the network. More specifically, they define the similarity as the fixed point of the following iterative scheme:

```
Algorithm 1                                                        (A1)
Initialize $S^0_{ij} := \delta_{ij}$ (i.e. all nodes are self-similar)
for $t = 1, 2, \cdots, tmax$ do
    $S^t_{ij} := \frac{\alpha}{|\Gamma(i)|\,|\Gamma(j)|} \sum_{k \in \Gamma(i),\, l \in \Gamma(j)} S^{t-1}_{kl} \ , \quad \forall i \neq j.$
endfor
$S := S^{tmax}$
```

where α is a damping factor chosen sufficiently small and $tmax$ is chosen sufficiently large so that the series converges. In this algorithm, self-similarity scores S_{ii} are set to 1 and then the crossed-similarity scores S_{ij} $(i \neq j)$ are computed in the loop. Notice that for an undirected graph the *Simrank* measure is equivalent to the *P-rank* measure introduced by Zhao *et al.* in [21].

Several other definitions are defined as fixed points of an iterative scheme. In [8], Leicht *et al.* state that the node i is similar to itself and that the node i is similar to node j if the neighbors of i are similar to the node j and hence propose the following update formula in algorithm (A1)

$$S_{ij}^t := \delta_{ij} + \alpha \sum_{k \in \Gamma(i)} S_{kj}^{t-1} \ . \tag{10}$$

The fixed point of this iteration is

$$S = I + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + \cdots \ . \tag{11}$$

Recalling that $[A^l]_{ij}$ gives the number of paths of length l from node i to node j, one can notice that the terms of equation (11) say that each path from i to j contributes to S_{ij} at the rate of α^l, where l is the length of the path.

B. *Non-topological Similarity Measures*

Let us first remind that non-topological similarity measures are defined between the nodes of a graph G_A and a graph G_B (possibly different from G_A). As always in this paper, we restrict our attention to the case of undirected graphs.

Blondel *et al.* [5] introduce a simple non-topological similarity measure. They require that, given i and j,

> if the similarity between $k \in \Gamma(i)$ and $l \in \Gamma(j)$ is large, then the similarity between i and j is large.

(R3)

This leads them to define a similarity measure as a fixed point of the following iterative scheme:

```
Algorithm 2                                        (A2)
Initialize $S_{ij}^0 := 1$
for $t = 1, 2, \cdots, 2\ tmax$ do
  1: $S_{ij}^t := \sum_{k \in \Gamma(i), l \in \Gamma(j)} S_{kl}^{t-1}$
  2: The similarity matrix is normalized, i.e.
       $S^t := \frac{S^t}{\|S^t\|_F}$
endfor
$S := S^{tmax}$
```

In [9], Melnik *et al.* consider a similar iterative scheme along with the following update formula for step 1 in algorithm 2

$$S_{ij}^t := S_{ij}^{t-1} + \sum_{\substack{k \in \Gamma(i) \\ l \in \Gamma(j)}} w_{ijkl} \ S_{kl}^{t-1} \ , \tag{12}$$

where w_{ijkl}, the propagation coefficient, is chosen such that the amount of similarity flows across the *categorical product* or *Kronecker product* $G_A \times G_B$ of the graphs G_A and G_B [22], *i.e.*

$$w_{ijkl} = \frac{1}{|\Gamma(k)| \ |\Gamma(l)|} \ .$$

One can notice that these definitions tend to give a higher similarity score to the nodes that have many neighbors. This can be avoided using weighted definitions. In [6], Holme *et al.* consider a similar iterative scheme on graphs with R different types of edges. They propose the following update formula for step 1 in algorithm 2

$$S_{ij}^t := \sum_{r=1}^{R} \frac{1}{|\Gamma_r(i)| \ |\Gamma_r(j)|} \sum_{\substack{k \in \Gamma_r(i) \\ l \in \Gamma_r(j)}} S_{kl}^{t-1} \ , \tag{13}$$

where $\Gamma_r(i)$ denotes the set of neighbors of i with respect to edges of type r.

III. Generalized Affine Transformation

The methods mentioned in the previous section can all be rewritten in terms of a normalized affine transformation,

```
Algorithm 3                                        (A3)
Given: an $m \times m$ adjacency matrix $A$, an $n \times n$
adjacency matrix $B$, $C \in \mathbb{R}^{m \times n \times m \times n}$, $D \in \mathbb{R}^{m \times n}$,
and $\rho : \mathbb{R}^{m \times n} \to \mathbb{R}$.
Initialize $S^0 \in \mathbb{R}^{m \times n}$.
for $t = 1, 2, \cdots, tmax$ do
    $S_{ij}^t := \frac{\sum_{k,l} C_{ijkl} S_{kl}^{t-1} + D_{ij}}{\rho(S^{t-1})} , \quad \forall i, \ j$
endfor
$S := S^{tmax}$
```

The linear term $C_{ijkl} S_{kl}^{t-1}$ accounts for the reinforcement of the similarities at each iteration and usually propagates them from neighbors to neighbors. The constant term D_{ij} influences the fixed point of the iteration based on *a priory* knowledge on local similarities.

A. *Topological Similarity Measures*

The topological similarity, by definition, implies some closeness of similar nodes. Hence, when the similarity score S_{ij} is propagated from neighbors to neighbors (possibly far from i and j), the propagated contribution decreases and asymptotically vanishes. As a consequence, it is often not necessary to scale the iterates and $\rho(S) = 1$.

For CN (1), Jaccard (2), Salton (3), Sørensen (4), Leicht (5), Adamic (6), Zhou (7) et (9), and Ravasz (8), the parameters of the normalized affine transformation in algorithm

(A3) simply reduce to

$$C_{ijkl} = 0 \ , \quad D_{ij} = S_{ij} \ \text{, and} \quad \rho(S) = 1 \ .$$

with S_{ij} the corresponding similarity definition.

The Jeh (A1), and Leicht (10) similarity are initialized with $S^0_{ij} = \delta_{ij}$ whereas the affine transformation coefficients are given by

$$D^{\text{JEH}}_{ij} = 0 \ , \quad \text{and} \quad D^{\text{LEICHT}}_{ij} = \delta_{ij} \ ,$$

along with

$$C^{\text{JEH}}_{ijkl} = \delta_{ij}\ \delta_{ik}\delta_{jl} + (1-\delta_{ij})\ \frac{\alpha\ A_{ki}A_{lj}}{|\Gamma(i)|\ |\Gamma(j)|} \ \text{, and}$$
$$C^{\text{LEICHT}}_{ijkl} = \alpha\ A_{ik}\delta_{lj} \ ,$$

where A is the adjacency matrix of G_A.

B. Non-topological Similarity Measures

Blondel (A2), Holme (13), and Melnik (12) are initialized with $S^0_{ij} = 1$ whereas the affine transformation coefficients are given by $D_{ij} = 0$, and

$$C^{\text{BLONDEL}}_{ijkl} = A_{ik}B_{jl} \ ,$$

$$C^{\text{HOLME}}_{ijkl} = \sum_{r=1}^{R} \frac{A^r_{ik}B^r_{jl}}{|\Gamma_r(i)|\ |\Gamma_r(j)|} \ \text{, and}$$

$$C^{\text{MELNIK}}_{ijkl} = \delta_{ik}\delta_{jl} + A_{ik}B_{jl}\ w_{ijkl} \ .$$

where A and B are respectively the adjacency matrices of G_A and G_B. In all cases, the scaling function normalizes the iterates after each step

$$\rho(S) = \left\| \left[\sum_{k,l} C_{ijkl}S_{kl} + D_{ij} \right]_{i,j} \right\|_F \ .$$

IV. A Novel Similarity Measure

The similarity measures mentioned in section II-B have several counter-intuitive or arguably undesirable properties.

Blondel (A2) and Melnik (12) give a high similarity score to nodes that have many neighbors since their similarity score gets more contribution. As a consequence, all nodes from one graph tend to have a high similarity score with the highest-degree node of the other graph.

For Holme (13), if the edges are not typed (*i.e.* $R = 1$) then the similarity matrix with all entries equal and norm 1 is a fixed point of the iteration, and all nodes in one graph are equally similar to all nodes in the other graph.

Moreover, for undirected graphs, the similarity matrix introduced by Blondel *et al.* (A2) has rank 1, *i.e.*, $S = uv^T$ for some vectors u and v. This implies that $\arg\max_i S_{ij}$ (resp. $\arg\max_j S_{ij}$) is the same for all j (resp. i), *i.e.*, all nodes of one graph have the highest similarity score with one same node of the other graph.

Notice that these properties are not always undesirable. Indeed, if, for one graph, every node may be mapped by an automorphism onto any other node, then the similarity matrix has rank 1 since all its rows are equal.

We now consider an alternative definition that avoids these counter-intuitive effects. We require that, given i and j,

$$\boxed{\begin{array}{l}\text{if } k \text{ and } l \text{ are s.t. } A_{ik,ik} = B_{jl,jl}\text{, and}\\ \text{if the similarity between } k \text{ and } l \text{ is large,}\\ \text{then the similarity between } i \text{ and } j \text{ is large.}\end{array}} \qquad \text{(R4)}$$

$A_{ik,ik}$ denotes a 2×2 matrix whose elements are

$$A_{ik,ik} = \begin{bmatrix} A(i,i) & A(i,k) \\ A(k,i) & A(k,k) \end{bmatrix} \ ,$$

where A is the adjacency matrix of G_A. In other words, if the connection between i and k is identical to the connection between j and l and if the similarity between k and l is large, then the similarity between i and j is large. For simple undirected graphs, the requirement (R4) is equivalent to saying that, given i and j, if the similarity between

- $k \in \Gamma(i)$ (neighbors of i) and $l \in \Gamma(j)$ is large, or
- $k' \notin \Gamma(i)$ and $l' \notin \Gamma(j)$ is large,

then the similarity between i and j is large.

We hence define a similarity measure as the fixed point of a scaled affine iterative scheme (A3) with the following parameters $S^0_{ij} = 1$ and $D_{ij} = 0$, and

$$C_{ijkl} = 1 \quad \text{if } A_{ik,ik} = B_{jl,jl} \ , \quad 0 \text{ otherwise.} \qquad (14)$$

The similarity score between node i and node j possibly gets contributions from the similarity scores of their respective neighbors and non-neighbors, which hence avoids that all nodes from one graph tend to have a high similarity score with the highest-degree node of the other graph.

We now propose to use our method to compute the *self-similarity* matrix of G, the simple undirected asymmetric graph shown in figure 1. The self-similarity matrix compares the nodes of G with themselves. Our method gives the following self-similarity matrix for G

$$S = \frac{1}{10} \begin{bmatrix} 2.36 & 1.09 & 2.02 & 1.64 & 1.59 & 2.27 \\ 1.09 & 1.72 & 1.26 & 1.45 & 1.47 & 1.13 \\ 2.02 & 1.26 & 1.85 & 1.63 & 1.58 & 1.95 \\ 1.64 & 1.45 & 1.63 & 1.58 & 1.55 & 1.59 \\ 1.59 & 1.47 & 1.58 & 1.55 & 1.54 & 1.56 \\ 2.27 & 1.13 & 1.95 & 1.59 & 1.56 & 2.22 \end{bmatrix} ,$$

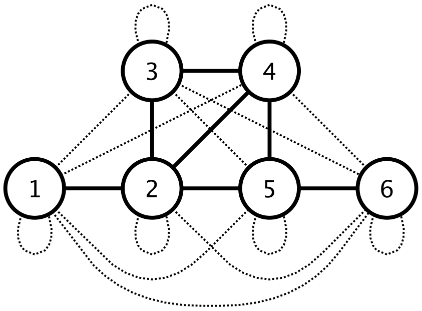

Fig. 1. The graph G (with plain edges) is a simple undirected asymmetric graph. The complementary graph of G is represented with dashed edges.

and Blondel gives

$$S^{\text{BLONDEL}} = \frac{1}{10}\begin{bmatrix} 0.42 & 1.16 & 0.8 & 1.04 & 0.92 & 0.33 \\ 1.16 & 3.18 & 2.2 & 2.87 & 2.53 & 0.92 \\ 0.8 & 2.2 & 1.52 & 1.99 & 1.75 & 0.64 \\ 1.04 & 2.87 & 1.99 & 2.59 & 2.29 & 0.83 \\ 0.92 & 2.53 & 1.75 & 2.29 & 2.02 & 0.73 \\ 0.33 & 0.92 & 0.64 & 0.83 & 0.73 & 0.27 \end{bmatrix} .$$

One can first notice that S has full rank whereas S^{BLONDEL} can be rewritten as

$$S^{\text{BLONDEL}} = uu^T \ , \quad \text{with} \quad u = \begin{bmatrix} 0.205 \\ 0.564 \\ 0.390 \\ 0.509 \\ 0.449 \\ 0.163 \end{bmatrix} .$$

As a consequence, every node is, in decreasing order, most similar to the nodes 2, 4, 5, 3, 1 and 6. Whereas, the similarity S yields that, in decreasing order, the nodes most similar

to node 1 are the nodes 1, 6, 3, 4, 5, 2, and
to node 2 are the nodes 2, 5, 4, 3, 6, 1, and
to node 3 are the nodes 1, 6, 3, 4, 5, 2, and
to node 4 are the nodes 1, 3, 6, 4, 5, 2, and
to node 5 are the nodes 1, 3, 6, 4, 5, 2, and
to node 6 are the nodes 1, 6, 3, 4, 5, 2.

V. Conclusion and Further Works

In this work, we showed that several similarity measures can be rewritten in terms of fixed points of a scaled affine transformation and proposed a novel definition of similarity that avoids undesirable degeneracy of the similarity matrix.

In future work, we will investigate the usefulness of these and other novel similarity measures as auxiliary tools in graph matching algorithms.

VI. Acknowledgments

We thank the reviewers for their kind remarks.

References

[1] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[2] A. T. Balaban. Applications of graph theory in chemistry. *Journal of Chemical Information and Computer Sciences*, (3):334–343, 1985.

[3] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[4] M. Heymans and A. K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, (1):i138–i146, 2003.

[5] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren. A measure of similarity between graph vertices: applications to synonym extraction and Web searching. *SIAM Review*, 46(4):647–666, 2004.

[6] P. Holme and M. Huss. Role-similarity based functional prediction in networked systems: application to the yeast proteome. *Journal of the Royal Society Interface*, (4):327–33, 2005.

[7] T. Zhou, L. Lu, and Y.-C. Zhang. Predicting missing links via local information. Technical Report arXiv:0901.0553, Jan 2009.

[8] E. A. Leicht, P. Holme, and M. E. J. Newman Vertex similarity in networks. *Physical Review*, (2), 2006.

[9] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In *Data Engineering, 2002. Proceedings. 18th International Conference on*, pages 117–128, 2002.

[10] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA, 2002. ACM.

[11] M. G. Everett and S. Borgatti. Role coloring a graph. *Math. Soc. Sci.*, 21:183–188, 1991.

[12] D. R. White and K. P. Reitz. Graph and semigroup homomorphisms on networks of relations. 1983.

[13] J. J. Luczkovich, S. P. Borgattiz, J. C. Johnsony, and M. G. Everett. Defining and measuring trophic role similarity in food webs using regular equivalence. 2002.

[14] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Socit Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[15] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York u.a., 1983.

[16] L. Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59:77–85, 2008.

[17] T. Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *K. Dan. Vidensk. Selsk. Biol. Skr.*, 5:1–34, 1948.

[18] E. Dahl. Some measures of uniformity in vegetation analysis. *Ecology*, 41(4):805–808, 1960.

[19] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[20] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297:1551–1555, 2002.

[21] P. Zhao, J. Han, and Y. Sun. P-rank: a comprehensive structural similarity measure over information networks. In *CIKM*, pages 553–562. ACM, 2009.

[22] M. Farzan and D. A. Waller. Kronecker products and local joins of graphs. *Canadian Journal of Mathematics*, pages 255–269, 1977.