Axiomatic Ranking of Network Role Similarity *

Ruoming Jin Victor E. Lee Hui Hong
Department of Computer Science
Kent State University, Kent, OH, 44242, USA
{jin,vlee,hhong}@cs.kent.edu

ABSTRACT

A key task in analyzing social networks and other complex networks is role analysis: describing and categorizing nodes by how they interact with other nodes. Two nodes have the same role if they interact with *equivalent* sets of neighbors. The most fundamental role equivalence is automorphic equivalence. Unfortunately, the fastest algorithm known for graph automorphism is nonpolynomial. Moreover, since exact equivalence is rare, a more meaningful task is measuring the role *similarity* between any two nodes. This task is closely related to the link-based similarity problem that SimRank addresses. However, SimRank and other existing simliarity measures are not sufficient because they do not guarantee to recognize automorphically or structurally equivalent nodes. This paper makes two contributions. First, we present and justify several axiomatic properties necessary for a role similarity measure or metric: range, maximal similarity, automorphic equivalence, transitive similarity, and the triangle inequality. Second, we present RoleSim, a role similarity metric which satisfies these axioms and which can be computed with a simple iterative algorithm. We rigorously prove that RoleSim satisfies all the axiomatic properties and demonstrate its superior interpretative power on both synthetic and real datasets.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and Networks; H.2.8 [**Database Management**]: Data Mining

General Terms

Algorithms, Measurement, Experimentation

Keywords

Complex network, Social network, Vertex similarity, Role similarity, Ranking, Automorphic equivalence

1. INTRODUCTION

*This work was supported in part by a NSF CAREER grant (IIS-0953950).

KDD'11, August 21–24, 2011, San Diego, California, USA.

In social science, it is well-established that individual agents tend to play roles or assume positions within their interaction network. For instance, in a university, each individual can be classified into the position of faculty member, administrator, staff, or student. Indeed, role discovery in is a major research subject in classical social science [40]. Interestingly, recent studies have found not only do roles appear in other types of networks, including food webs [26], world trade [15], and even software systems [8], but also roles can help predict node functionality. For instance, in a protein interaction network, proteins with similar roles tend to serve similar metabolic functions. Thus, if we know the function of one protein, we can predict that all other proteins having a similar role would also have similar function [17]. In other cases, online social networks for example, there are no a priori role categories. The classifications must be learned based on the interaction patterns.

A central question in studying the roles in a network system is how to define *role similarity*. In particular, how can we rank two nodes' role similarity in terms of their interaction patterns? Despite its vital importance for network analysis and decades of work by social scientists, joined recently by computer scientists, no satisfactory metric for role similarity has yet emerged. A key issue is the encapsulation of graph automorphism into a role similarity metric: *if two nodes are automorphically equivalent, then they should share the same role and their role similarity should be maximal*. From a network topology viewpoint, automorphic nodes have equivalent surroundings. Figure 1 illustrates a graph with nodes $S1$ and $J1$ being automorphically equivalent. Automorphism can be further generalized in terms of *coloration*: assuming each node is assigned a color, then two nodes are equivalent if their neighborhoods consist of the same color spectrum [11].

The traditional social science approach for role analysis has been to define suitable mathematical equivalence relations for nodes so they can be partitioned into equivalence classes (roles). An essential property of these equivalences is that they should positively confirm automorphic equivalence, i.e., if any two nodes are automorphic, then they are role-equivalent. (The converse is not necessarily true.) Confirming automorphism is verifyig a solution, which is often algorithmically less complex than discovering a solution. Thus though there is no known polynomial-time algorithm for discovering graph automorphism [1], role equivalence algorithms [3, 5, 35] can still guarantee to satisfy the aforementioned automorphism confirmation property. These equivalence rules also directly correspond to the aforementioned coloration.

However, by relying on strict equivalence rules, these role modeling schemes can produce only binary similarity metrics: two nodes are either equivalent (similarity $= 1$) or not (similarity $= 0$).

[1]The computational complexity of graph isomorphism and automorphism are still unproven to be either P or $NP - Complete$.

In real-world networks, usually only a very small portion of the node-pairs would satisfy an equivalence criteria [27] and among those, many are simply trivially equivalent (such as singletons or children of the same parent). In addition, strict rule-based equivalence is not robust with respect to network noise, such as false-positive or false-negative interactions. Thus, it is desirable in many real world applications to rank node-pairs by their degree of similarity or provide a real-valued node similarity *metric*.

Recent research works have proposed various measures of node similarity based on similarity of interactions. In [21], a is similar to b if a's neighbor is similar to b. This definition does not fit the class-based concept of roles. SimRank [18] is based on the following principle: "two nodes are similar if they link to similar nodes". Mathematically, for any two different nodes x and y, SimRank computes their similarity recursively according to the average similarity of all the neighbor pairs (a neighbor of x paired with a neighbor of y). A single node has self-similarity value 1. This is equivalent to the probability that two simultanous random walkers, starting at x and y, will eventually meet. Most of the other node structural similarity measures [1, 12, 22, 43, 44, 45] are variants of SimRank. Though SimRank seems to capture the intuition of the above recursive structural similarity, its random walk matching does not satisfy the basic graph automorphism condition. For example, in Figure 1, though $S1$ and $J1$ are automorphically equivalent, SimRank assigns them a value of 0.226. We discuss this further in Section 3.2. To our best knowledge, there is no available real-valued structural similarity measure satisfying the automorphic equivalence requirement. Since automorphic equivalence is a pivotal characteristic of the notion of role, its lack disqualifies these existing measures from serving as authentic role similarity measures.

Thus we have an open problem: *Can we derive a real-valued role similarity measure or ranking which complies with the automorphic equivalence requirement?* In this paper, we develop the first real-valued similarity measure to solve this problem. In addition, our measure is also a metric, i.e., it satisfies the triangle inequality. The key feature of our role similarity measure is a weighted generalization of the *Jaccard coefficient* to measure the neighborhood similarity between two nodes. Unlike SimRank, which considers the average similarity among all possible pairings of neighbors, our measure counts only those pairs in the matching of the two neighbor sets which maximizes the targeted similarity function.

2. ROLE EQUIVALENCE

In social network analysis, the traditional approach for discovering role groups is to define a equivalence relation and to partition the actors into equivalence classes. Actors who fulfill the same role are equivalent. Over the years, four definitions have stood out. These four, in decreasing order of strictness, are structural equivalence, automorphic equivalence, equitable partition, and regular equivalence. Figure 1 shows how these different definitions generate different roles from the same network.

Let $G = (V, E)$ be a graph with vertex set $V = \{v_1, ..., v_n\}$ and edge set E. For any node $v \in V$, let $N(v)$ be the neighbors of v and N_v be the degree of v.

• **Structural Equivalence:** Two actors are *structurally equivalent* if they interact with the *same* set of others [24]. Mathematically, u and v are structurally equivalent if and only if $N(u) = N(v)$. For example, consider the extended family shown in Figure 1. $S1$, $J1$, and $L1$ are siblings, $S2$, $J2$, and $L2$ are spouses, and the remaining nodes are their children. Each family's children, $\{S3, S4\}$, $\{J3, J4\}$, and $\{L3, L4, L5\}$, form a nontrivial equivalence class.

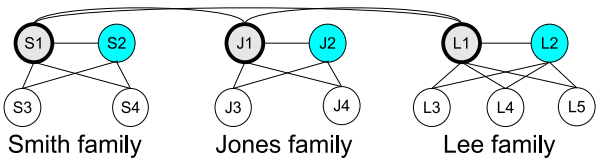


Figure 1: Example Graph for Equivalence Classes.

Equivalence	Neigh. Rule	Non-singleton Classes
Structural	exactly same	{S3,S4},{J3,J4}, {L3,L4,L5}
Automorphic, Equit. Part.	same number per class	{S1,J1},{S2,J2}, {S3,S4,J3,J4}, {L3,L4,L5}
Regular	same class	{S1,J1,L1},{S2,J2,L2}, {S3,S4,J3,J4,L3,L4,L5}

Table 1: Equivalence Classes for Figure 1

However, none of the parents can be grouped together via structural equivalence. This model is too strict to be useful for simplifying a large network and to discover meaningful roles.

• **Automorphic Equivalence:** Two actors (nodes) u and v are *automorphically equivalent* if there is an automorphism σ of $G$where $v = \sigma(u)$ [4]. An automorphism σ of a graph G is a permutation of vertex set V such that for any two nodes u and v, $(u, v) \in E$ iff $(\sigma(u), \sigma(v)) \in E$. In social terms, u and v can swap names, along with possibly some other name swaps, while preserving all the actor-actor relationships. Let $\Gamma(G)$ be the group of all automorphisms of graph G. For any two nodes u and v in G, $u \equiv v$ if $u = \sigma(v)$ for some $\sigma \in \Gamma(G)$. Note that $\equiv$ is an equivalence relation on V; if $u \equiv v$ we say that u is automorphically equivalent to v. The equivalence classes generated under $\Gamma(G)$ (or $\equiv$) are called orbits. The equivalence class for vertex $v \in V$ is called the orbit of v, and denoted as $\Delta(v) = \{\sigma(v) \in V, \sigma \in \Gamma(G)\} = \{u | u \equiv v\}$. Each orbit corresponds to a role in the automorphic equivalence. Understanding the importance of automorphic equivalence and applying it to role modeling was a major breakthrough in classical social network research. In our example Figure 1, from the topology alone, we cannot distinguish between the Smith family and the Jones family. The Lee family is distinct, because it has three children instead of two. Therefore, the equivalence classes are $\{S1, J1\}, \{S2, J2\}, \{S3, S4, J3, J4\}, \{L1\}, \{L2\}$, and $\{L3, L4, L5\}$. Interestingly, automorphically equivalent classes must have equivalent indirect relations as well, such as equivalent in-laws and cousins. However, automorphic equivalence is hard to compute and still very strict.

• **Exact Coloration (Equitable Partition):** An *exact coloration* of graph G assigns a color to each node, such that any two nodes share the same color iff they have the same number of neighbors of each color [10]. Nodes of the same color form an equivalence class. An exact coloration is also referred to as equitable partition [14] and graph divisor [7] and is often applied in the vertex classification/refinement for canonical labeling in a graph isomorphism test [32, 29]. A graph may have several exact colorations; in general we seek the fewest colors. In our example, structural equivalence and automorphic equivalance offer two different exact colorations. Exact coloration relaxes automorphism by considering only immediate neighborhood equivalence, yet it still embodies a recursive aspect to role modeling.

• **Regular Equivalence (Bisimulation):** Two actors are *regularly equivalent* if they interact with the same variety of role classes, where class is recursively defined by regular equivalence [41]. Un-

like automorphic equivalence and exact coloration, regular equivalence does not care about the cardinality of neighbor relationships, only whether they are nonzero. For example, using regular equivalence, all three families are now equivalent. There are only three equivalence classes: $sibling - parent\{S1, J1, L1\}$, $spouse - parent\{S2, J2, L2\}$, and $child$. Note that under regular equivalence, any two automorphically equivalent nodes may be partitioned into the same regular equivalence class. In computer science, the regular equivalence is often referred to as the bisimulation, which is widely used in automata and modal logic [28].

3. AXIOMATIC ROLE SIMILARITY

An equivalence relation, however, tells us nothing about non-equivalent items. The real-world need is for a measure that not only recognizes automorphic equivalence, such as Smith child/spouse/parent to Jones child/spouse/parent (Figure 1), but also tells us that a Lee child/spouse/parent has strong similarity to a Lee or Smith child/spouse/parent. Over the years, several methods have been developed for addressing various link-based similarity problems (co-citation [34], coupling [20], SimRank [18]). Recently, several researchers have tried to apply these measurements to role modeling [17, 45]. However, none of these encompass the aforementioned automorphic equivalence property and thus are inadequate for measuring role similarity. To deal with this shortcoming and to clarify the problem, we first identify a list of axiomatic properties that all role similarity measures should obey.

DEFINITION 1. *(***Axiomatic Role Similarity Properties***) Given a graph $G = (V, E)$, any $sim(a, b)$ that measures the neighbor-based role similarity between vertices a and b in V should satisfy properties P1 to P5:*

- *P1) Range: $0 \leq sim(a, b) \leq 1$, for all a and b.*
- *P2) Symmetry: $sim(a, b) = sim(b, a)$.*
- *P3) Automorphism confirmation: If $a \equiv b$, $sim(a, b) = 1$.*
- *P4) Transitive similarity: If $a \equiv b$, $c \equiv d$, then $sim(a, c) = sim(a, d) = sim(b, c) = sim(b, d)$.*
- *P5) Triangle inequality: $d(a, c) \leq d(a, b) + d(b, c)$, where distance $d(a, c)$ is defined as $1 - sim(a, c)$.*

Any node similarity measure satisfying the first four conditions (without triangle inequality) is called an **admissible role similarity measure**. *Any node similarity measure satisfying all five conditions is an* **admissible role similarity *metric***.

Property 1 describes the standard normalization where 1 means fully similar and 0 means completely dissimilar (i.e., the two neighborhoods have nothing in common). Property 2 indicates that similarity, like distance, must be symmetric. Property 3 expresses our idea that fully similar means automorphically equivalent. Property 4 claims that the similarity between two nodes is equal to the similarity between equivalent members of the first two node's respective equivalence classes. In other words, we can simply define the similarity for the orbits, i.e., $sim(\Delta(u), \Delta(v)) = sim(u, v)$. This guarantees consistency of values at an orbit-level. Property 5 assumes the measure is metric-like, i.e., satisfying the triangle inequality. This is much stronger than transitivity, enforcing an *ordering* of values.

Note that Property 5 implies Property 4. However, since most similarity measures do not necessarily satisfy the triangle inequality, we specify Property 4 separately. Further, Property 3 is an essential criterion which distinguishes the role similarity measure from other existing measures. As we discussed earlier, the automorphic equivalence can be relaxed to exact coloration or regular equivalence. In this case, we may replace Property 3 accordingly. Our work will focus on the automorphic equivalence though it can handle its generalization as well.

THEOREM 1. *(***Generalized Transitive Similarity***) For any two pairs of nodes $a, b \in V$, $c, d \in V$, if $sim(a, b) = 1$ and $sim(c, d) = 1$, then, their cross similarities are all equal, i.e., $sim(a, c) = sim(a, d) = sim(b, c) = sim(b, d)$.*

Proof: From the triangle inequality, we have $d(a, c) \leq d(a, b) + d(b, c) \leq d(b, c)$ and $d(b, c) \leq d(b, a) + d(a, c) \leq d(a, c)$ ($d(a, b) = 0$). Thus, $d(a, c) = d(b, c)$. Similarly, $d(a, d) = b(b, d)$, $d(c, a) = d(d, a)$, and $d(d, a) = d(d, b)$. Put together, we have $sim(a, c) = sim(a, d) = sim(b, c) = sim(b, d)$. □

Thus, if we partition the nodes into equivalence classes where similarity equals 1, we can simply record the similarity values between equivalent classes. Let $\Delta(x)$ and $\Delta(y)$ be the equivalence classes for node x and y, respectively. Then, we can define $sim(\Delta(x), \Delta(y)) = sim(x, y)$.

3.1 Binary-Valued Role Similarity Measures

THEOREM 2. *(***Binary Admissibility***) Given any equivalence relation that also satisfies automorphism confirmation (P3), its binary indicator function is an admissible similarity* metric.

The proof of the Theorem is omitted due to lack of space and can be found in [19]. Note that automorphic equivalence, regular equivalence, and exact coloration all satisfy P3, so they are admissible metrics. Though these similarity measures are admissible, binary-valued measures do not help us to understand the degree of similarity or dissimilarity. We would like a real-valued measure that ranks the degree of role similarity.

Before presenting our proposed real-valued role similarity metric for network roles, we first examine some similarity measures proposed in earlier works.

3.2 SimRank is Not Admissible

The SimRank [18] similarity between nodes u and v is the average similarity between u's neighbors and v's neighbors:

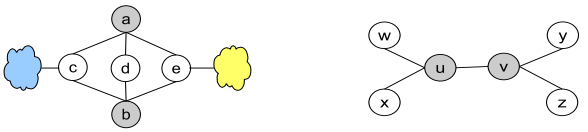
$$SR(u, v) = \frac{(1 - \beta)}{|N(u)||N(v)|} \sum_{x \in N(u)} \sum_{y \in N(v)} SR(x, y), \text{for } u \neq v,$$

$$SR(v, v) = 1,$$

where β is a decay factor, $0 < \beta < 1$, so that the influence of neighbors decreases with distance. The original SimRank measure is for directed graphs. Here, we focus on its undirected version, though our comments also hold for the directed version. SimRank values can be computed iteratively, with successively iterations approaching a unique solution, much as PageRank [31] does.

THEOREM 3. *SimRank is not an admissible role similarity measure.*

Proof: We give examples where property 3 (automorphic equivalence) does not hold. In Figure 2(a), a and b have the same neighbors. By even the strictest definition (structural equivalence), a and b have the same role. However, since SimRank's initial assumption is that there is no similarity among c, d, and e, when it computes the average similarity of a and b's neighbors, it will never discover their equivalence. Assuming the best case where c, d, and e are in fact equivalent and using the recommended $\beta = 0.15$, $SR(a, b)$ converges to only 0.667. If the neighbors are not equivalent, a to b should still be equivalent, but SimRank gives an even lower value.



(a) Structural equivalence (b) Odd Distance

Figure 2: Problematic configurations for SimRank

SimRank has an another problem (Figure 2(b)) when there is an odd distance between two nodes. Nodes u and v are automorphically equivalent, but because there are no nodes that are an equal distance from both u and v, $SimRank(u, v) = 0$!

We note that other variants of SimRank [1, 12, 22, 43, 44, 45] also do not meet the automorphic equivalence property for to similar reasons. More discussion of these variants can be found in [19].

4. ROLESIM: A REAL-VALUED ADMISSIBLE ROLE SIMILARITY

To produce an admissible real-valued role similarity measure, we face two key challenges: First, it is computationally difficult to verify the automorphic equivalence property. Though not proven to be NP-complete, the graph automorphism problem has no known polynomial algorithm [13]. Second, all the existing real-valued role similarity measures have problems dealing with even simple conditions such as structural equivalence (Subsection 3.2). To meet these challenges, we take the following approach: Given an initial simplistic but admissible role similarity measurement for each pair of nodes, refine the measurement by expressing similarity in terms of neighboring values, while maintaining the automorphic and structural equivalence properties. Next, we formally introduce RoleSim, the first admissible real-valued role similarity measure (metric) and its associated properties.

4.1 RoleSim Definition

Given a graph $G = (V, E)$, the RoleSim measure realizes the recursive node structural similarity principle "two nodes are similar if they relate to similar objects" as follows.

DEFINITION 2. *(**RoleSim metric**) Given two vertices u and v, where $N(u)$ and $N(v)$ denote their respective neighborhoods and N_u and N_v denote their respective degrees, then* $RoleSim(u, v) =$

$$(1-\beta)\max_{M(u,v)}\frac{\sum_{(x,y)\in M(u,v)} RoleSim(x,y)}{N_u + N_v - |M(u,v)|} + \beta \quad (1)$$

where $x \in N(u)$, $y \in N(v)$, and $M(u, v)$ is a **matching** *between $N(u)$ and $N(v)$, i.e., $M(u, v) = \{(x, y)|x \in N(u), y \in N(v),$ and no other$(x', y') \in M(u, v)$, s.t. , $x = x'$ or $y = y'\}$. The parameter β is a decay factor, $0 < \beta < 1$.*

The decay factor, similar to the one used in PageRank [31], both dampens the recursive effect and guarantees a minimal RoleSim score of β. We will sometimes abbreviate $RoleSim(u, v)$ as $R(u, v)$. $\mathbf{R}$ refers to the entire matrix of values. Figure 3 illustrates the matching process. The (x, y) grid is the subset of the RoleSim matrix of values corresponding to the pairings of neighbors of these two vertices. A matching selects one cell per row and column. If the number of rows differs from the number of columns, then the matching size is limited to $|M(u, v)| = min(N_u, N_v)$. A maximal matching is a matching where the total value of selected cells is maximum. In contrast, SimRank computes the average of every cell in the neighbor grid.

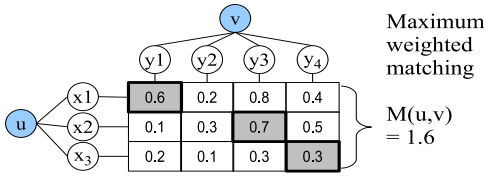


Figure 3: RoleSim(a,b) based on similarity of their neighbors

4.1.1 Relation to Jaccard Coefficient

RoleSim employs a generalization of the Jaccard coefficient, which measures the commonality between two sets A and B as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. Previous works [12] have used this index to compare node neighborhoods; several variants exist [30]. Our denominator is similar to that of the Tanimoto coefficient [37], which measures simlarity between multisets or between vectors. In our generalization, however, sets A and B are not vectors and need not share any common elements; instead, there is a weighted matching M between *similar* elements in A and B, i.e., $(a, b) \in M, a \in A, b \in B$. Let $r(a, b) \in [0, 1]$ record the similarity between a and b.

DEFINITION 3. *(**Generalized Jaccard Coefficient**) The generalized Jaccard coefficient measures the similarity between two sets A and B under matching M, defined as*

$$J(A, B|M) = \frac{\sum_{(a,b)\in M} r(a, b)}{|A| + |B| - |M|} \quad (2)$$

The original Jaccard coefficient is a special case which uses the following matching M: Let $r(x, y) = 1$ if $x = y$; otherwise 0. Then define $M = \{r(x, x)|x \in A, x \in B\}$. Thus, the generalized Jaccard coefficient $J(A, B|M)$ reduces to $J(A, B)$. Comparing Eq. (1) and (2), we see that the heart of $RoleSim(u, v)$ is equivalent to the maximum of the generalized Jaccard coefficient between $N(u)$ and $N(v)$, among all matchings $M(u, v)$. Then, $RoleSim(u, v) =$

$$(1-\beta)\max_{M(u,v)} J(N(u), N(v)|M(u, v)) + \beta \quad (3)$$

4.1.2 Relation to Weighted Matching

The definition and significance of the RoleSim for any node pair (u, v) is closely related to *maximal weighted matching*. For any nodes u and v in graph G, define a weighted bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$, with each edge $(x, y) \in N(u) \times N(v)$ having weight $RoleSim(x, y)$. Let the total weight of neighbor matching $M(u, v)$ between u and v be $w(M(u, v)) = \sum_{(x,y)\in M(u,v)} RoleSim(x, y)$. Let $\mathcal{M}$ be the maximal weighted matching for $(N(u) \cup N(v), N(u) \times N(v))$. It is clear that

$$w(\mathcal{M}) = \max_{M(u,v)} w(M(u, v)). \quad (4)$$

Using this, we can represent $RoleSim(u, v)$ in terms of maximal weighted matching $\mathcal{M}$. In Figure 3, the shaded cells represent the maximal matching: $0.7 + 0.6 + 0.3 = 1.6$.

THEOREM 4. *(**Maximal Weighted Matching**) The RoleSim between nodes u and v corresponds linearly to the maximal weighted matching $\mathcal{M}$ for the bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$, with each edge $(x, y) \in N(u) \times N(v)$ having the weight $RoleSim(x, y)$:*

$$RoleSim(u, v) = (1-\beta)\frac{w(\mathcal{M})}{\max(N_u, N_v)} + \beta \quad (5)$$

Proof: We need to show that Equations (1) and (5) are equivalent. Without loss of generality, let $N_u \geq N_v$. First, we show that *the cardinality of the maximal weighted matching* $|\mathcal{M}| = \min(N_u, N_v) = N_v$. It cannot be greater, because there are insufficient elements in N_v. It cannot be smaller, because if it were, there must exist an available edge between an uncovered node in N_u with one in N_v. Adding this edge would increase the matching (every edge has weight $\geq \beta$). If $|\mathcal{M}| = \min(N_u, N_v)$, it follows that $N_u + N_v - |M| = \max(N_u, N_v)$. Thus, the denominators in Equations (1) and (5) are constant and identical. It is then a trivial observation that the numerators are in fact the same. Therefore, the maximal value fo the entire Equation (1) is the same as the value in (5). □

Theorem 4 not only shows the key equilibrium of role similarities between pairs of nodes in a graph G, but shows that RoleSim may be computed using existing maximal matching algorithms.

4.2 RoleSim Computation

RoleSim values can be computed iteratively and are guaranteed to converge, just as in PageRank and SimRank. First we outline the procedure. In the next section, we prove that the calculated values comprise an admissible role similarity metric.

Step 1: Let the initial matrix of RoleSim scores be $\mathbf{R^0}$, estimated but admissible scores between any pair of nodes in G.

Step 2: Compute the k^{th} iteration $\mathbf{R^k}$ scores from the $(k-1)^{th}$ iteration's values, $\mathbf{R^{k-1}}$. Specifically, for any nodes u and v,

$$\mathbf{R^k}(u,v) = (1-\beta) \max_{M(u,v)} \frac{\sum_{(x,y)\in M(u,v)} \mathbf{R^{k-1}}(x,y)}{N_u + N_v - |M(u,v)|} + \beta \qquad (6)$$

Based on Theorem 4, we compute Equation (6) by finding the maximal weighted matching in the weighted bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$ with each edge $(x,y) \in N(u) \times N(v)$ having weight $\mathbf{R^{k-1}}(x,y)$).

Step 3: Repeat Step 2 until $\mathbf{R}$ values converge for each pair of nodes in G.

THEOREM 5. ***(Convergence)*** *For any admissible set of RoleSim scores $RoleSim^0$, the iterative computational procedure for RoleSim converges, i.e., for any (u,v) pair,*

$$\lim_{k\to\infty} RoleSim^k(u,v) = RoleSim(u,v) \qquad (7)$$

This can be proven by showing that the maximum absolute difference between any $\mathbf{R^k}(u,v)$ and $\mathbf{R^{k+1}}(u,v)$ is monotonically decreasing. The proof can be found in [19].

Unlike PageRank and SimRank which converge to values independent of the initialization, the convergent RoleSim score is sensitive to the initialization. Rather than being a disadvantage, this sensitivity provides the necessary relaxation to compute automorphic role similarity in polynomial time, by utilizing the initialization as prior knowledge.

4.3 Admissibility of RoleSim

Here, we present one of the key contributions of this paper: the axiomatic admissibility of RoleSim. If the initial computation is admissible, and because the iterative computation of Equation (5) maintains admissibility (i.e., is an invariant transform of the axiomatic properties), then the final measure is admissible.

THEOREM 6. ***(Invariant Transformation)*** *If the k^{th} iteration $RoleSim^k$ is an admissible role similarity metric, then so is $RoleSim^{k+1}$.*

For each axiomatic property P, we must show "If the k^{th} iteration $RoleSim^k$ satisfies Axiom P, then so does $RoleSim^{k+1}$." Properties 1 (Range) and 2 (Symmetry) are trivially invariant, so we will focus on the other three.

Automorphism Confirmation Invariance Proof: For nodes $u \equiv v$, there is a permutation σ of vertex set V, such that $\sigma(u) = v$, and any edge $(u,x) \in E$ iff $(v, \sigma(x)) \in E$. This indicates that σ provides a one-to-one equivalence between nodes in $N(u)$ and $N(v)$. Also, u and v have the same number of neighbors, i.e., $N_u = N_v$. So, it is clear that the maximal weighted matching $\mathcal{M}$ in the bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$ selects $N_u = N_v$ pairs of weight 1 each. Thus, $RoleSim^{k+1}(u,v) = (1-\beta)\frac{w(\mathcal{M})}{\max(N_u,N_v)} + \beta = 1$. □

Transitive Similarity Invariance Proof: Assume for any $a \equiv b$, $c \equiv d$, $RoleSim^k(a,c) = RoleSim^k(b,d)$. Denote the maximal weighted matching between $N(a)$ and $N(c)$ as $\mathcal{M}$. Since there is a one-to-one equivalence correspondence σ between $N(a)$ and $N(b)$ and a one-to-one equivalence correspondence σ' between $N(c)$ and $N(d)$, we can construct a matching $\mathcal{M}'$ between $N(b)$ and $N(d)$ as follows: $\mathcal{M}' = \{(\sigma(x), \sigma'(y)) | (x,y) \in \mathcal{M}\}$. Since transitive similarity holds for $RoleSim^k$, we have $RoleSim^k(x,y) = RoleSim^k(\sigma(x), \sigma'(y))$. Thus, $w(\mathcal{M}') = w(\mathcal{M})$, and

$$(1-\beta)\frac{w(\mathcal{M})}{\max(N_a,N_c)} + \beta = (1-\beta)\frac{w(\mathcal{M}')}{\max(N_b,N_d)} + \beta$$
$$RoleSim^{k+1}(a,c) = RoleSim^{k+1}(b,d). \qquad \square$$

Triangle Inequality Invariance Proof: For iteration k, for any nodes a, b, and c, $d^k(a,c) \leq d^k(a,b) + d^k(b,c)$, where $d^k(a,b) = 1 - RoleSim^k(a,b)$. We must prove that this inequality still holds for the next iteration: $d^{k+1}(a,c) \leq d^{k+1}(a,b) + d^{k+1}(b,c)$.

Observation: *if there is a matching M between $N(a)$ and $N(c)$ which satisfies $1 - ((1-\beta)\frac{w(M)}{N_c} + \beta) \leq d^{k+1}(a,b) + d^{k+1}(b,c)$, then $d^{k+1}(a,c) \leq d^{k+1}(a,b) + d^{k+1}(b,c)$.* This is because $\frac{w(M)}{N_c} \leq \frac{w(\mathcal{M})}{N_c}$, where $\mathcal{M}$ is the maximal weighted matching between $N(a)$ and $N(c)$, and thus, $1 - ((1-\beta)\frac{w(M)}{N_c} + \beta) \geq 1 - ((1-\beta)\frac{w(\mathcal{M})}{N_c} + \beta) = d^{k+1}(a,c)$.

We break down the proof into three cases:
Case 1. ($N_b \leq N_a \leq N_c$), Case 2. ($N_a \leq N_b \leq N_c$), and
Case 3. ($N_a \leq N_c \leq N_b$).

Case 1: Since N_b is smallest, $|\mathcal{M}(a,b)| = |\mathcal{M}(b,c)| = N_b$. Define matching M between $N(a)$ and $N(c)$ as $M = \{(x,z) | (x,y) \in \mathcal{M}(a,b) \wedge (y,z) \in \mathcal{M}(b,c)\}$. Then using our observation above:

$$\begin{aligned}
&d^{k+1}(a,b) + d^{k+1}(b,c) - (1 - (1-\beta)\frac{w(M)}{N_c} - \beta) \\
&= (1-\beta)[-\frac{w(\mathcal{M}(a,b))}{N_a} - \frac{w(\mathcal{M}(b,c))}{N_c} + \frac{w(M)}{N_c}] + 1 - \beta \\
&= (1-\beta)[\frac{N_b - w(\mathcal{M}(a,b))}{N_a} - \frac{N_b}{N_a} + \frac{N_b - w(\mathcal{M}(b,c))}{N_c} \\
&\quad - \frac{N_b}{N_c} - \frac{N_b - w(M)}{N_c} + \frac{N_b}{N_c}] + 1 - \beta \\
&\geq (1-\beta)[1 - \frac{N_b}{N_a} + \frac{\sum_{(x,y)\in\mathcal{M}(a,b)}(1 - R^k(x,y))}{N_c} \\
&\quad + \frac{\sum_{(y,z)\in\mathcal{M}(b,c)}(1 - R^k(y,z))}{N_c} - \frac{\sum_{(x,z)\in M}(1 - R^k(x,z))}{N_c}] \\
&\geq (1-\beta)[\frac{\sum_{(x,y,z)}(d^k(x,y) + d^k(y,z) - d^k(x,z))}{N_c}] \geq 0
\end{aligned}$$

where $(x, y) \in \mathcal{M}(a, b), (y, z) \in \mathcal{M}(b, c), (x, z) \in M$ □

Cases 2 and 3 can be proven by a similar technique; the details are in [19].

By combining the admissible initial configurations given in Sec 4.4 with Theorem 6 on invariance, we have shown that the iterative RoleSim computation generates a real-valued, admissible role similarity measure.

THEOREM 7. *(**Admissibility**) If the initial $RoleSim^0$ is an admissible role similarity measure, then at each k-th iteration, $RoleSim^k$ is also admissible. When RoleSim computation converges, the final measure $\lim_{k\to\infty} RoleSim^k$ is admissible.*

4.4 Initialization

According to Theorem 7, an initial admissible RoleSim measurement $\mathbf{R^0} = I(\cdot)$ is needed to generate the desired real-valued role similarity ranking. What initial admissible measures or prior knowledge should we use? We consider three schemes:

1. **ALL-1** : $I(u, v) = 1$ for all u, v.
2. **Degree-Binary (DB)**: If two nodes have the same degree ($N_u = N_v$), then $I(u, v) = 1$; otherwise, 0.
3. **Degree-Ratio (DR)**: $I(u, v) = (1 - \beta)\frac{min(N_u, N_v)}{max(N_u, N_v)} + \beta$.

These schemes come from the following observation: *nodes that are automorphically equivalent have the same degree*. Basically, equal degree is a necessary but not sufficient condition for automorphism. This observation is key to RoleSim: degree affects both the size of a maximal matching set and the denominator of the Jaccard Coefficient.

THEOREM 8. *(**Admissible Initialization**) ALL-1, Degree-Binary, and Degree-Ratio are all admissible role similarity measures. Moreover, Degree-Binary and ALL-1 are admissible role similarity* metrics.

Its proof is omitted due to lack of space and can be found in [19]. Note that SimRank's initialization ($SimRank^0(u, v) = 1$ iff $u = v$) is NOT admissible, because it does exactly the wrong thing: setting the initial value of any potentially equivalent nodes to 0. SimRank iterations try to build up from zero. However, due to its problems with structural equivalence and odd-length paths that we noted, SimRank will never increase the value enough to discover equivalent pairs that were neglected at the start.

In addition, we make the following interesting observation (based on the definition of RoleSim formula): *Let $\mathbf{R^1}(ALL-1)$ be the matrix of RoleSim values at the first iteration after $\mathbf{R^0} = \mathbf{1}$ (All-1 initialization); and let $\mathbf{R^0}(DR)$ be the matrix of RoleSim initialized by the Degree-Ratio (DR) scheme. Then, $\mathbf{R^1}(ALL - 1) = \mathbf{R^0}(DR)$.* Basically, the Degree-Ratio (DR) is exactly equal to the RoleSim state one iteration after ALL-1 initialization. Thus, ALL-1 and DR generate the same final results. The simple formula for DR is much faster than neighbor matching, so DR is essentially one iteration faster. On the other hand, we may consider the simple ALL-1 scheme to be sufficient, since it works as well as the more sophisticated DR. Especially, after the simple initialization, RoleSim's maximal matching process automatically discriminates between nodes of different degree and continues to learn differences among neighbors as it iterates. In our experiments, we will further empirically study these initialization schemes.

4.5 Computational Complexity

Given n nodes, we have $O(n^2)$ node-pair similarity values to update for each iteration. For each node-pair, we must perform a maximal weighted matching. For weighted bipartite graph $(N(u) \cup N(v), N(u) \times N(v))$, the fastest algorithm based on augmenting paths (Hungarian method) can compute the maximal weighted matching in $O(x(x \log x + y))$, where $x = |N(u) \cup N(v)|$ and $y = |N(u)| \times |N(v)|$.

A fast greedy algorithm offers a $\frac{1}{2}$-approximation of the globally optimal matching in $O(y \log y)$ time [2]. If an equivalence matching exists (i.e., $w(\mathcal{M}) = \max(N_u, N_v)$), the greedy method will find it. This is important, because it means that a greedy RoleSim computation still generates an admissible measure. Using greedy neighbor matching, the time complexity of RoleSim is $O(kn^2d')$, for k iterations, where d' is the average of $y \log y$ over all vertex-pair bipartite graphs in G. The space complexity is $O(n^2)$. We note that in order to handle very large graphs, in [19], we introduce an interesting *iceberg RoleSim* computation algorithm which can discover high RoleSim pairs without full materializing the entire RoleSim matrix. It is beyond the scope of this paper and we omit its further discussion.

5. EXPERIMENTAL EVALUATION

In this section we experimentally investigate the ranking ability and performance of the RoleSim algorithm for computing role similarity metric values. We analyze the effect of different initialization schemes, and compare RoleSim to several state-of-the-art node similarity algorithms Specifically, we focus on the following:

1. How do different initialization schemes perform in terms of their final RoleSim score and computational efficiency?
2. Do node-pairs with high RoleSim scores have similar network roles, and for any two nodes known to have similar network roles, do they have high RoleSim scores?

Clearly, the ideal validation study requires an explicit role model and role similarity measure, which often do not exist. In the following study, we utilize a well-known role-related random graph model and external measures of real datasets which provide strong role indication for these evaluations.

We set $\beta = 0.1$ for both RoleSim and SimRank, defining convergence to be when values change by less than 1% of their previous values. We ran several RoleSim tests with both exact matching and greedy matching. The results were nearly identical ($> 90\%$ of cells have no difference; maximum difference was small), so we focus on greedy matching from here on. We implemented the algorithms in C++ and ran all large tests on a 2.0GHz Linux machine with dual-core Opteron CPU and 4.0GB RAM.

For our tests, we use three types of graphs:

• **BL**: probabilistic block-model [39], where each block is generally considered to be corresponding to a role [42]. Here, nodes are partitioned into blocks. Each node in block i has probability p_{ij} of linking to each node in block j. Thus, the underlying block-model may serve as the ground-truth for testing role similarity.

• **SF**: Large Scale-Free random graphs [2] offer another model of large social or complex networks.

• Real-world networks, with a measureable feature similar to social role, are used for validating RoleSim performance.

5.1 Comparing Initialization

In Section 4.4 we saw that Degree-Ratio generates the same results as ALL-1 by shortcutting the first iteration. This reduces computation time by roughly 10%. Now we ask: Does Degree-Binary initialization (DB, binary indicator equaling 1 when degrees $N_u = N_v$) give similar results, quickly?

[2]http://pywebgraph.sourceforge.net/

Relative to All-1 Initialization	Degree-Binary			Degree-Ratio
	Min	Avg.	Max	
Diff. in percentile rank	0.14%	0.38%	11.17%	none
Pearson correl. coeff.	0.9994	0.9998	0.9999	1
Relative execution time	0.32	0.52	0.80	≈ 0.9
Relative # iterations	0.38	0.58	0.88	1 fewer

Table 2: Comparison of Initialization Methods

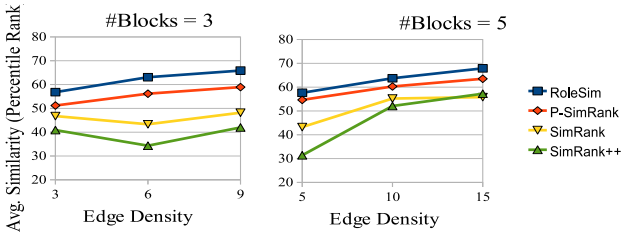


Figure 4: Avg. similarity ranking for nodes in the same block

We ran RoleSim using both ALL-1 and DB on 12 graphs, some scale-free and some block-model, having 500 to 10,000 nodes, and edge densities from 1 to 10. We then converted values to percentile ranking, where 100% means the highest value and 50% is the median value. Test results are summarized in Table 2. The high correlation coefficient means the rankings are virtually identical, so the rankings are not very sensitive to the initialization method. Moreover, DB took 20% from 68% less time to converge. Overall, DB seems to be the preferred initialization scheme in terms of computational efficiency. Thus, we adopt it for the rest of the experiments.

5.2 General Role Detection

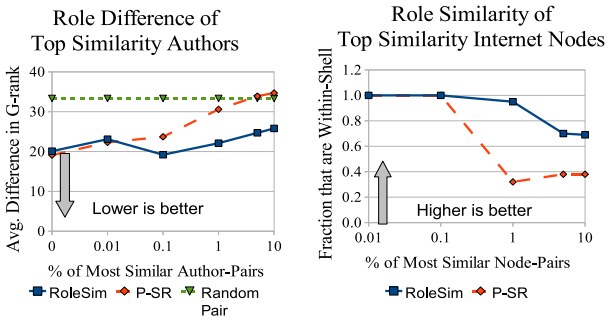
How well does RoleSim discover roles in complex graphs? Specifically, given a ground truth knowledge of roles, do nodes having similar roles have high scores? To answer this, we generated probabilistic block-model graphs, where blocks behave like "noisy" roles, due to sampling variance. We generated graphs with $N = 1000$ nodes and either 3 or 5 blocks. We varied the edge density $\frac{|E|}{|V|}$, with higher densities for graphs with more blocks. The size of each block and the p_{ij} values were randomized; we generated 3 random instances for each graph class. We compared RoleSim to the state-of-the-art SimRank, SimRank++ [1], and P-SimRank [12] measures.

For each measure and trial, we ranked its similarity scores. This normalizes the scoring among the four measures. Next, for each graph, we computed the average rank of all pairs of nodes within the same block, then averaged the three trials for each graph class.

Our results (Figure 4) show that RoleSim outperforms all other algorithms across all the tested conditions. None of the algorithms score perfectly, due to the inherent edge distribution variance of the probabilistic model. P-SimRank is better than SimRank, perhaps because it uses Jaccard Coefficient weighting, a step towards our RoleSim approach. Accuracy takes time. SimRank and SimRank++ run at the same speed. P-SimRank is about twice as slow, taking 184s to complete the most dense graph. RoleSim took 948s to complete the same graph.

5.3 Real Dataset: Co-author Network

We applied RoleSim and the best alternative measure, P-SimRank, to a real-world network having an external role measure. Our first dataset [36] is a co-author network of 2000 database researchers. Two authors are linked if they co-authored a paper from



(a) Top Coauthors (b) Top Internet nodes

Figure 5: Similarity of Nodes for Top Ranked Node-Pairs

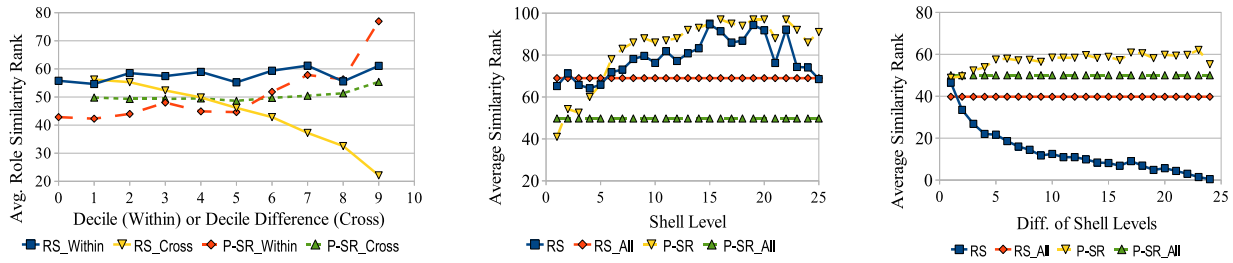
2003 to 2008. We pruned the network to the largest connected component (1543 nodes, 15483 edges). An author's role depends recursively on the number of connections to other authors, and the roles of those others. Hence, it measures collaboration. We use the G-index as a proxy measure for co-author role (H-index provides similar results and is omitted here). The G-index measures the influence of a scientific author's publications, its value being the largest integer G such that the G most cited publications have at least G^2 citations. While G-index and co-author role are not precisely the same, G-index score is influenced strongly by the underlying role. High impact authors tend to be highly connected, especially with other high impact authors. If a paper is highly cited, this boosts the score of every co-author. Thus, we expect that if two authors have similar G-index scores, their node-pair is likely to have a high role simlarity value. To normalize RoleSim, P-SimRank, and G-index values, we converted each raw value to a percentile rank.

Figure 5(a) addresses our second validation question (high rank$\rightarrow$ similar roles?). For the top ranked 0.01% of author-pairs, the average difference in G-index ranking is 20 points, for both RoleSim and P-SimRank, well below the random-pair difference of 33. A below-average difference confirms that the authors are relatively similar. However, as we expand the search towards 10%, RoleSim continues to detect authors with similar authorship performance, while P-SimRank converges to random scoring.

To validate $role \rightarrow rank$ performance, we binned the authors into 10 roles based on G-index value (bottom 10%, next 10%, etc.). For every pair of authors within the same role decile, we looked up role similarity rank and computed an average per bin. We also computed averages for pairs of authors not in the same bin (dissimilar roles). Figure 6(a) shows our results. The average within-bin RoleSim value is consistently between 55% and 60%, better than the random-pair score of 50, and independent of whether the G-index is high or low. It performs equally well for all roles. P-SimRank within-bin scores (dashed line), however, are inconsistent. Performance of P-SimRank is worse than random for low G-scores, perhaps due to low density of links in the network. For the cross-bin data, the X-axis is the difference in decile bins for the two authors in a pair. The falling line of RoleSim indicates that role similarity correctly decreases as G-index scores become less similar. For P-SimRank, however, the cross-bin scores (dashed line) hover around 50, equivalent to random scoring.

5.4 Real Dataset: Internet Network

Our second dataset is a snapshot of the Internet at the level of autonomous systems (22963 nodes and 48436 edges), as generated



(a) Similarity of Authors Binned by K-index (b) Internet Graph Intra-Shell Similarity (c) Internet Graph Cross-Shell Similarity

Figure 6: Similarity Rank for Nodes Grouped by External Role Measure

by Newman[3]. Several studies have confirmed that the Internet is hierarchically organized, with a densely connected core and stubs (singly-connected nodes) at the periphery [38, 6]. A node's position within the network (proximity to the core) and its relation to others affects its efficiency for routing and its robustness. Inspired by [6], we use K-shells to delineate roles.

The K-core of a graph is the induced subgraph where every node connects to at least K other nodes in the subgraph. If $K' > K$, then the K'-core must be an induced subgraph of the K-core. The K-shell is defined as the 'ring' of nodes that are included in a graph's $(K-1)$-core but not its K-core. Thus we can decompose a graph into a set of nested rings, becoming denser as we move inward.

Using K-shells as our roles, we perform tests and analyses similar to those of the coauthor network. In Figure 5(b) we see that both measures do well for the top 0.1%, but P-SimRank's falters significantly when the range is expanded to the top 1%.

Next, we treat K-shells the same way that we treated G-index decile bins in the previous test. See Figures 6(b) and 6(c). Unlike decile bins, the shells do not have equal sizes. K-shells 1, 2, and 3 together contain 92% of all nodes. To clarify how these three shells dominate, we also show horizontal lines representing the combined weighted average rank of all within-shell comparisons. RoleSim's within-shell values are consistently high, averaging 70%. Conversely, P-SimRank finds strong above-average similarity for the small high-K shells, but nearly random similarity for shells 1 to 3, pulling its overall performance down to 50%.

In cross-shell analysis, RoleSim is able to distinguish different shells very well: RoleSim approaches zero as shell difference approaches maximum. On the other hand, P-SimRank shows almost no correlation to shell difference. Many of its scores are above-average when they should be below-average (dissimilar). On the whole, it seems that P-SimRank is not detecting role, but something related to connectedness and density.

In all these experiments, we can see that RoleSim provides positive answer to the role similarity ranking: 1) node-pairs with similar roles have higher RoleSim ranking than node-pairs with dissimilar roles, and 2) high RoleSim ranking indicates that nodes have similar roles. P-SimRank scores, however, do not correlate with network role similarity.

6. RELATED WORK

The role similarity problem is a distinct special case of the more general structural or link similarity problems, which find applications in co-citation and bibliographic networks [25], recommender systems, [1] and Web search [16]. Link similarity means that two objects accrue similarity if they have similar links.

[3]Internet dataset, http://www-personal.umich.edu/ mejn/netdata/

Formal definitions of role, which define exactly what is being measured, arose from the social science community [24, 33, 9]. Block partitioning can be used to group nodes directly into roles [42]. However, this does not produce individual node-pair similarities, so it is not useful as a ranking method.

SimRank [18] is the best known algorithm to implement a recursive definition of object similarity: two objects are similar if they relate to similar objects. SimRank has an elegant random walk interpretation: $SimRank(a,b)$ is the probability that two independent simultaneous random walkers, beginning at a and b, will eventually meet at some node. However, the more neighbors that a and b have in common, the less likely that they will both randomly choose the same neighbor. This then explains SimRank's problem with structural equivalence. Recently, Zhao [45] has pointed out that in-neighbor and out-neighbor SimRank can be used as a univeral framework to describe co-citation (common in-neighbors), bibliographic coupling (commnon out-neighbors), or a weighted combination of the two. The number of iterations reflects the search radius for discovering similarity. As we note in Section-3.2, SimRank has an undesirable trait: its values decrease when the number of common neighbors increases. Several works have tried to address this problem. SimRank++ [1] adds an *evidence* weight which partially compensates for the neighbor matching cardinality problem. In [12], the two walkers are not fully independent, because the overall probability of a meeting b is weighted to be Jaccard coefficient $\frac{|N(u)\cap N(v)|}{|N(u)\cup N(v)|}$. Recently, MatchSim [23] has also used maximal matching of neighbors to address problems with SimRank's scoring. However, our formulations have important differences. Because they retained SimRank's initialization, their work does not guarantee automorphic equivalence in the final results. Also, their work is intuition-based, without a theory of correctness. They provide one specific formulation, while we define a theoretical framework for *any* admissible measure or metric. Because RoleSim satisfies the triangle inequality, it is a true metric.

7. CONCLUSION

We have developed RoleSim, the first real-valued role similarity measure that confirms automorphic equivalence. We have also presented a set of axioms which can test any future measure to see if it is an admissible measure or metric. Our experimental tests demonstrate RoleSim's correctness and usefulness on real world data, opening up exciting possibilities for scientific and business applications. At the same time, we see that other well-known measures, while suitable for other tasks, are not suitable for role similarity. This axiomatic approach may prove useful for developing and validating solutions to other related tasks.

8. REFERENCES

[1] Ioannis Antonellis, Hector Garcia-Molina, and Chi-Chao Chang. Simrank++: query rewriting through link analysis of the clickgraph (poster). In *WWW'08*, pages 1177–1178, 2008.

[2] D. Avis. A survey of heuristics for the weighted matching problem. *Network*, 13:475–493, 1983.

[3] Vladimir Batagelj, Patrick Doreian, and Anuška Ferligoj. An optimizational approach to regular equivalence. *Social Networks*, 14:121–135, 1992.

[4] Stephen P. Borgatti and Martin G. Everett. Notions of position in social network analysis. *Sociological Methodology*, 22:1–35, 1992.

[5] Stephen P. Borgatti and Martin G. Everett. Two algorithms for computing regular equivalence. *Social Networks*, 15:361–376, 1993.

[6] Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of internet topology using k-shell decomposition. *PNAS*, 104(27):11150–11154, July 2007.

[7] Dragos M. Cvetkovíc, Michael Doob, and Horst Sachs. *Spectra of Graphs: Theory and Applications, 3rd Revised and Enlarged Edition*. Wiley, 1998.

[8] Natalia Dragan, Michael L. Collard, and Jonathan I. Maletic. Using method stereotype distribution as a signature descriptor for software systems. In *ICSM*, pages 567–570, 2009.

[9] Martin G. Everett. Role similarity and complexity in social networks. *Social Networks*, 7:353–359, 1985.

[10] Martin G. Everett and Stephen P. Borgatti. Exact colorations of graphs and digraphs. *Social Networks*, 18:319–331, 1996.

[11] M.G. Everett and S. P. Borgatti. Regular equivalence: General theory. *J. Mathematical Sociology*, 19:29–52, 1994.

[12] Dániel Fogaras and Balázs Rácz. Scaling link-based similarity search. In *WWW'05*, pages 641–650, 2005.

[13] Scott Fortin. The graph isomorphism problem. Technical Report TR 96-20, Dept. Computer Science, Univ. of Alberta, Edmonton, Alberta, Canada, July 1996.

[14] C. Godsil and G. Royle. *Algebraic Graph Theory*. Springer-Verlag, New York, 2001.

[15] E.M. Hafner-Burton, M. Kahler, and A.H. Montgomery. Network analysis for international relations. *International Organization*, 63(03):559–592, 2009.

[16] Taher H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):784–796, 2003.

[17] Petter Holme and Mikael Huss. Role-similarity based functional prediction in networked systems: application to the yeast proteome. *J. R. Soc. Interface*, 2(4):327–33, 2005.

[18] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *KDD'02*, pages 538–543, 2002.

[19] Ruoming Jin, Victor E. Lee, and Hui Hong. Axiomatic ranking of network role similarity. Technical Report arXiv:1102.3937, http://arXiv.org, 2011.

[20] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963.

[21] E. A. Leicht, Petter Holme, and M. E. J. Newman. Vertex similarity in networks. *Phys. Rev. E*, 73:026120, 2005.

[22] Pei Li, Yuanzhe Cai, Hongyan Liu, Jun He, and Xiaoyong Du. Exploiting the block structure of link graph for efficient similarity computation. In *PAKDD'09*, pages 389–400, 2009.

[23] Zhenjiang Lin, Michael R. Lyu, and Irwin King. Matchsim: a novel neighbor-based similarity measure with maximum neighborhood matching. In *CIKM'09*, pages 1613–1616, 2009.

[24] F. P. Lorrain and H. C. White. Structural equivalence of individuals in networks. *J. Math. Sociology*, 1:49–80, 1971.

[25] Wangzhong Lu, Jeannette Janssen, Evangelos Milios, and Nathalie Japkowicz. Node similarity in networked information spaces. In *CASCON'01*, pages 11–25, 2001.

[26] J. J. Luczkovich, S.P. Borgatti, J.C. Johnson, and M.G. Everett. Defining and measuring trophic role similarity in food webs using regular coloration. *J. Theoretical Biology*, 220:303–321, 2003.

[27] Ben D. MacArthur, Rubén J. Sánchez-García, and James W. Anderson. Note: Symmetry in complex networks. *Discrete Appl. Math.*, 156(18):3525–3531, 2008.

[28] Maarten Marx and Michael Masuch. Regular equivalence and dynamic logic. *Social Networks*, 25(1):51–65, 2003.

[29] B.D. McKay. Practical graph isomorphism. *Congressus Numerantium*, 30:45–87, 1981.

[30] Guy Melançon and Arnaud Sallaberry. Edge metrics for visual graph analytics: A comparative study. In *Proc. 12th Int'l Conf. Inform. Visual.*, pages 610–615, 2008.

[31] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Univ., 1999.

[32] Ronald Read and Derek Corneil. The graph isomorphism disease. *J. Graph Theory*, 1:339–363, 1977.

[33] Lee Douglas Sailer. Structure equivalence: meaning and definition, computation and application. *Social Networks*, 1:73–80, 1978.

[34] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Amer. Soc. Information Sci.*, 24:265–269, 1973.

[35] Malcolm K. Sparrow. A linear algorithm for computing automorphic equivalence classes: the numerical signatures approach. *Social Networks*, 15(2):151–170, 1993.

[36] Jie Tang, Jing Zhang, Limin Yao, and Juanzi Li. Extraction and mining of an academic social network. In *WWW'08*, pages 1193–1194, New York, 2008. ACM.

[37] T. T. Tanimoto. An elementary mathematical theory of classification and prediction. *IBM Taxonomy Application M. and A.6*, 3, Nov. 1958.

[38] Sudhir L. Tauro, Georgos Siganos, C. Palmer, and Michalis Faloutsos. A simple conceptual model for the internet topology. In *Proc. IEEE Global Telecomm. Conf.*, pages 1667–1671, 2001.

[39] Y.J. Wang and G.Y. Wong. Stochastic blockmodels for directed graphs. *J. Amer.Stat. Assoc.*, 82(397):8–19, 1987.

[40] Stanley Wasserman and Katherine Faust. *Social network analysis: methods and applications*. Cambridge University Press, 1994.

[41] Douglas R. White and Karl P. Reitz. Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5:193–234, 1983.

[42] Harrison White, Scott Boorman, and Ronald Breiger. Social structure from multiple networks. i: Blockmodels of roles and positions. *Am. J. Sociology*, 81:730–780, 1976.

[43] Wensi Xi, Edward A. Fox, Weiguo Fan, Benyu Zhang, Zheng Chen, Jun Yan, and Dong Zhuang. Simfusion: measuring similarity using unified relationship matrix. In *SIGIR'05*, pages 130–137, 2005.

[44] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Linkclus: efficient clustering via heterogeneous semantic links. In *VLDB'06*, pages 427–438, 2006.

[45] Peixiang Zhao, Jiawei Han, and Yizhou Sun. P-rank: a comprehensive structural similarity measure over information networks. In *CIKM'09*, pages 553–562, 2009.