TINTIN: Exploiting Target Features for Signaling Network Similarity Computation and Ranking

Huey Eng Chua, Sourav S. Bhowmick
Complexity Institute, Nanyang Technological University, Singapore
hechua@assourav@ntu.edu.sg

Lisa Tucker-Kellogg
Duke-NUS Medical School, National University of Singapore, Singapore
lisa.tucker-kellogg@duke-nus.edu.sg

ABSTRACT
Network similarity ranking attempts to rank a given set of networks based on its “similarity” to a reference network. State-of-the-art approaches tend to be general in the sense that they can be applied to networks in a variety of domains. Consequently, they are not designed to exploit domain-specific knowledge to find similar networks although such knowledge may yield interesting insights that are unique to specific problems, paving the way to solutions that are more effective. We propose TINTIN which uses a novel target feature-based network similarity distance for ranking similar signaling networks. In contrast to state-of-the-art network similarity techniques, TINTIN considers both topological and dynamic features in order to compute network similarity. Our empirical study on signaling networks from BioModels with real-world curated outcomes reveals that TINTIN ranking is different from state-of-the-art approaches.

1 INTRODUCTION
The network similarity problem scores the resemblance between a pair of related networks. We can broadly classify network similarity approaches for biological networks into two classes, namely node mapping-based and network feature-based. The former is based on graph isomorphism [15, 19] as node mappings are performed using different measures and the extent of the network similarity is dependent on the mappings. In contrast, the latter class of approaches do not assume the existence of such node mappings. These approaches [4, 17] generally employ network structure-based local or global measures to determine similarity between networks. Specifically, global measures, such as those based on network features, are derived from the entire biological network and tend to be biased as biological networks are inherently noisy and incomplete [17]. In contrast, local measures, which are typically derived from regions of networks that are well-studied, are deemed more appropriate [17]. For instance, graphlet degree distribution (GDD) [17] and NetSimLe [4] both use local measures to determine network similarity. A common theme that runs through these network feature-based approaches is their generality and applicability to other types of networks such as social networks. Consequently, they are not designed to exploit domain-specific knowledge to find similar networks, although such knowledge may yield interesting insights that are unique to specific problems. In this paper, we present a novel network similarity technique called TINTIN, which is designed for signaling networks and leverages such domain-specific knowledge to identify networks that are similar in terms of target features.

Signaling networks model biological systems as networks of interacting molecules. When biological processes are dysregulated due to diseases, the activities of downstream molecule(s) (referred to as disease node(s) in the paper) are affected and typically manifest themselves as phenotypic changes related to the disease. For instance, in the MAPK–PI3K network, the hyperactivity of activated ERK, a downstream molecule, is often linked to cell proliferation [21], a cancer hallmark. This has led to increasing popularity of targeted therapeutic strategies to tackle such diseases where drugs are designed to hit molecules in a signaling network crucial for tumor growth and progression [22]. These molecules (referred to as targets) modulate the disease node(s) directly (e.g., MEK [21]) or indirectly (e.g., Raf [21]).

In this paper, we exploit the network features associated with targets in a signaling network to compute similarity between signaling networks. Intuitively, in target feature-based network similarity problem we deem a pair of signaling networks as similar if their targets have similar characteristics. Hence, our network similarity technique called TINTIN (Target-based Signaling Network Similarity Computation) is driven by similarity of features of targets that modulate specific disease-related nodes (disease nodes) associated with a pair of signaling networks. It takes as input (a) a reference network (e.g., Ras activation), its disease node (e.g., Ras) and an associated set of known targets; and (b) a set of candidate networks along with their disease nodes, known targets and features. It ranks these candidate networks based on their degree of similarity to the reference network w.r.t. the features of the known targets. TINTIN is potentially useful for several real-world applications such as network-based target prioritization [8]1 and clustering signaling networks based on the feature similarity of targets.

In summary, this paper makes the following key contributions: (a) We introduce the novel problem of target feature-based network similarity for signaling networks (Section 4). (b) We present TINTIN, a target-driven approach that is, to the best of our knowledge, the world’s first algorithm to address this problem (Section 5). (c) We conduct an empirical study on real signaling networks and drug target data to analyze the similarities and differences of TINTIN’s rankings compared to that of the state-of-the-art generic network similarity techniques (Section 6).

1 In contrast to this work, the study in [8] does not focus on details of the network similarity problem. Instead, it assumes the existence of such a network similarity technique, which is utilized for target prioritization.
2 RELATED WORK

Compared to generic target-unaware network similarity techniques such as GDD [17] and NetSIMILE [4], TINTN differs in the following ways. First, the network similarity problems are defined differently when perturbed, it modulates the activity of a specific node (referred to as disease node). A disease node is a protein that is either involved in some biological processes which may be deregulated, resulting in manifestation of a disease, or be of interest due to its potential role in the disease. Given a signaling network \( G = (V, E) \) and a disease node \( x \in V \), let the set of nodes having a path leading to \( x \) be denoted as \( V_x \subseteq V \). Then, the set of candidate target nodes in \( G \) relevant to \( x \) is denoted as \( T_x \subseteq V_x \).

Each reaction (edge) in a signaling network is further annotated with dynamic information associated with the biochemical process. In signaling networks, numerous ordinary differential equations (ODEs) containing various reaction kinetics and initial concentrations for every species are used to model the production and consumption rates of different molecular species [2]. The determination of these reaction kinetics can be technically challenging. Hence, a large proportion of these kinetics are usually estimated using parameter estimation techniques [18]. Despite this uncertainty, these under-determined ODE systems can still model real, observable biological behaviour, providing valuable means for quantitative study. Note that ODE-based models of signaling networks are expected to grow further in the future and become an important and accepted way of representing biological knowledge [2]. In this paper, we use hypergraphs containing ODEs simulation to obtain concentration-time series profiles (i.e., plots of concentration against time) of nodes.

3.2 Profile Shape Similarity Distance (PSSD)

In signaling networks, signal responses to perturbation are typically measured in terms of phosphoprotein concentrations dynamics represented as concentration-time profiles. There are certain considerations in comparing these phosphoprotein concentration-time profiles. In signaling networks, reactions occur at different and non-uniform rates [1] resulting in profiles with variable time delays. Hence, a distance measure based on one-to-one alignment is ineffective at detecting similarity in these profiles. A non-linear measure, such as dynamic time warping (DTW) distance, allows a more intuitive alignment between profiles [15] and is more suitable for biological time series data [1].

Definition 3.1. Given two discrete time series \( \varphi_u \) and \( \varphi_v \), the dynamic time warping distance between them is defined recursively as:

\[
\text{DTW}(\varphi_u, \varphi_v) = \xi(F(\varphi_u), F(\varphi_v)) + \min \left\{ \begin{array}{l}
\text{DTW}(\varphi_u, \text{Rest}(\varphi_v)) \\
\text{DTW}(\text{Rest}(\varphi_u), \varphi_v) \\
\text{DTW}(\text{Rest}(\varphi_u), \text{Rest}(\varphi_v))
\end{array} \right. 
\]

where \( F(\varphi_u) = \{\varphi_u[1]\}, \text{Rest}(\varphi_u) = \{\varphi_u[2], \varphi_u[3], \ldots, \varphi_u[n]\} \), \( \xi(\varphi_u[i], \varphi_v[j]) = (\varphi_u[i] - \varphi_v[j])^2 \) and \( \varphi_u[i] \) is the value of \( \varphi_u \) at time point \( i \) [15].

Definition 3.2. Given a concentration-time profile \( \varphi_u \) having \( n \) time points, denoted as \( \varphi_u = \{\varphi_u[0], \ldots, \varphi_u[n]\} \), let \( m \) be the median value of \( \varphi_u \). The corresponding inverted profile is denoted as \( \varphi_u' = \{\varphi_u'[0], \ldots, \varphi_u'[m]\} \) where \( \varphi_u'[i] = 2 \times m - \varphi_u[i] \).

Definition 3.3. Given a signaling network \( H = (V_H, E_H) \), let \( \varphi_u, \varphi_v \) be the Z-normalized concentration-time profiles of \( u, v \in V_H \). The profile shape similarity distance [5] of \( u \) with respect to \( v \) is defined as:

\[
\Phi_{(u, v)} = \min(\text{DTW}(\varphi_u, \varphi_v), \text{DTW}(\varphi_u', \varphi_v'))
\]
In summary, PSSD identifies the most relevant upstream regulators by assessing the similarity of the concentration-time series profiles of a target and its upstream regulators.

### 3.3 Target Characterization using TENET

In this paper, we leverage on a target characterization approach (TENET [6]) to generated the ground truth for assessing the performance of TINTIN. TENET is a network-centric, in silico target characterization system, which uses signaling networks having known targets from publicly-available signaling network repositories (e.g., BioModels) to learn for each network, a set of topological features that are predictive of targets and a characterization model that can be used to generate topological feature-based (TFB) rankings of targets. The characterization model specifies which topological features are important for discriminating the targets in a signaling network and how these features should be combined to quantify the likelihood of a node being a target. It generates different characterization models for different networks as it is unlikely for one characterization model to generalize the characteristics of known targets in all networks due to the complexity and diversity of signaling networks.

### 4 TARGET FEATURE-BASED NETWORK SIMILARITY PROBLEM

Network similarity measures the similarity between a pair of networks. In the literature, there are different strategies to measure such similarity for both directed (e.g., signaling networks) and undirected (e.g., PPI networks) networks. Hence, a set of networks can be ranked with respect to a reference network based on their similarity degrees to the latter. Formally, the similarity-based network ranking problem can be defined as follows:

**Definition 4.1.** Given a reference network $G$ and a set of candidate networks $\mathcal{L} = \{L_1, \ldots, L_N\}$, the similarity-based network ranking problem computes network similarity distance $D(G, L_i)$ between $G$ and each $L_i \in \mathcal{L}$ and ranks them in ascending (or descending) order of $D(G, L_i)$. Given the networks $L_1$ and $L_2$ where $L_1, L_2 \in \mathcal{L}$, $L_1$ is more similar to $G$ if $D(G, L_1) < D(G, L_2)$. The best matched network $L_k \in \mathcal{L}$ of $G$ is the network with the smallest network similarity distance. That is, $\forall L \in \mathcal{L}, D(G, L_k) < D(G, L_1)$.

Recall that the network similarity distance in TINTIN is measured using target features. Hence, we now formally introduce it. The network similarity distance $D(G, L)$ between a pair of signaling networks, $G$ and $L$, is based on similarity of the targets in $G$ and $L$ with respect to their network features. Table 1 lists the set of topological and dynamic features that we consider for computing $D(G, L)$ (referred to as target features). That is, given a reference network $G$ and two candidate networks $L_i$ and $L_j$, $G$ is more similar to $L_i$ if the feature distribution of the targets in $G$ is more similar to that of the targets in $L_i$ across all network features being considered.

**Definition 4.2.** Given a reference network $G$, its disease node $x_G$ and an associated set of known targets $\mathcal{T}_G$; a set of candidate networks $\mathcal{L} = \{L_1, \ldots, L_N\}$, their disease nodes $\bigcup_{i=1}^{L} x_{L[i]}$ and associated sets of known targets $\bigcup_{i=1}^{L} \mathcal{T}_L[i]$; and a set of target features $X$, the distribution similarity of $G$ and $L_i$ for a target feature $X_j$ is defined as

$$p_{G, L_i, X_j} = \mathcal{F}(G(x_G, T_G, X_j), \mathcal{G}(L_i, x_{L_i}, T_{L_i}, X_j))$$

where $\mathcal{G}(G(x_G, T_G, X_j))$ is a function that retrieves target feature $X_j$ for a set of nodes $T_G$ in a given network $G$ with disease node $x_G$, and $\mathcal{F}(A, B)$ is a statistical function that computes the similarity of two distributions $A$ and $B$. Then, the target-based similarity distance between $G$ and $L_i$ is defined as

$$D(G, L_i) = C(|p_{G, L_i, X} | 1 \leq |X|)$$

where $C(Z)$ is a function that aggregates the set of items $Z$.

In this work, we use the target features specified in Table 1. Hence, the functions $(\mathcal{G}(G(x_G, T_G, X_j)))$ used for retrieving the target features correspond directly to the formula for computing these features. For example, to compute PSSD, we use the formula given in Definition 3.3. The formula for the remaining target features can be found in [6]. In particular, we use two statistical functions $(\mathcal{F}(A, B))$, namely, Wilcoxon Rank-Sum (Wilcoxon) and Kolmogorov-Smirnov (ks) statistical measures to assess distribution similarity. Hence, distribution similarities are obtained as $p$-values. We use Stouffer’s method (Stouffer) as the aggregation function $(\mathcal{C}(Z))$ to aggregate the $p$-values in Equation 2. Note that Stouffer’s method can be applied to combine dependent $p$-values by introducing some degree of dependence (correlation) between pairs by following the approach in [11]. A larger aggregated $p$-value implies the null hypotheses (a closer target feature-based similarity) are true for every test.

### 5 THE TINTIN ALGORITHM

Algorithm 1 outlines the TINTIN algorithm. Given a reference signaling network $G$, its disease node $x_G$ and known targets $T_G$, a set of candidate networks $\mathcal{L}$, their disease nodes $x = \bigcup_{i=1}^{L} x_{L[i]}$ and known targets $T = \bigcup_{i=1}^{L} T_{L[i]}$, it identifies the best matched network $G_{best} \in \mathcal{L}$ of $G$ and a ranked list of $\mathcal{L}$ in two phases. TINTIN provides optional relaxation parameter $p_r$ to configure the criteria for filtering out dissimilar networks.

---

**Table 1: Target features. T: Topological, D: Dynamic**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_u$</td>
<td>Degree centrality of node $u$. The in, out and total degree centralities are denoted as $\theta_{in}(u)$, $\theta_{out}(u)$ and $\theta_{total}(u)$, respectively.</td>
<td>T</td>
</tr>
<tr>
<td>$\omega_u$</td>
<td>Eigenvector centrality of node $u$.</td>
<td>T</td>
</tr>
<tr>
<td>$\phi_u$</td>
<td>Closeness centrality of node $u$.</td>
<td>T</td>
</tr>
<tr>
<td>$\gamma_u$</td>
<td>Eccentricity centrality of node $u$.</td>
<td>T</td>
</tr>
<tr>
<td>$\delta_u$</td>
<td>Betweenness centrality of node $u$.</td>
<td>T</td>
</tr>
<tr>
<td>$\pi_u$</td>
<td>Bridging coefficient of node $u$.</td>
<td>T</td>
</tr>
<tr>
<td>$\xi_u$</td>
<td>Bridging centrality of node $u$.</td>
<td>T</td>
</tr>
<tr>
<td>$\kappa_u$</td>
<td>Clustering coefficient of node $u$. The undirected, in, out, cycle and middleman clustering coefficients are denoted as $\kappa_{undir}(u)$, $\kappa_{in}(u)$, $\kappa_{out}(u)$, $\kappa_{cycle}(u)$ and $\kappa_{mid}(u)$, respectively.</td>
<td>T</td>
</tr>
<tr>
<td>$\mu_u$</td>
<td>Proximity prestige of node $u$.</td>
<td>T</td>
</tr>
<tr>
<td>$\omega_i$</td>
<td>Target downstream effect of node $u$ [5].</td>
<td>T</td>
</tr>
<tr>
<td>$\Phi(u, v)$</td>
<td>Profile shape similarity distance (pssd) between nodes $u$ and $v$ [5].</td>
<td>D</td>
</tr>
</tbody>
</table>

---

1The Wilcoxon and its tests are nonparametric and are suitable for features with distribution that are unknown a priori.
Algorithm 1 Algorithm TINTIN

Require: Reference network G and its disease node xG; and known targets T; set of candidate networks L; their disease nodes x and known targets T; relaxation parameter p (optional).
Ensure: Best matched network Gbest and network ranked list r.
1: Xbest, Pbest, H, Truth ← learnBestVariant(L, x, T)
2: pt ← learnP Threshold(Truth, Pbest, pt)
3: Gbest, r ← getBestNetwork(G, xG, T, G, Xbest, pt, H)

Algorithm 2 Procedure learnBestVariant

Require: Set of training networks L, their disease nodes x and known targets T.
Ensure: Best feature type Xbest; matrix of p-values for training networks using the best feature type Pbest; matrix of all features values in all training networks H; and the ground truth Truth.
1: A ← initialize(|Xall, x_t, Xj|)
2: H ← extractFeatures(L, Xall)
3: Truth ← getGroundTruth(L, H)
4: for iteration i = 1 to |L| do
5: for iteration j = 1 to |A| do
6: M ← getRelevantFeatures(H, A)
7: Pw ← wilcox(Li, Mi, M)
8: Pk ← KS(Lj, Mk, M)
9: PVal(k, Lj) ← combineP(Pw, Pk)
10: end for
11: end for
12: end for
13: end for
14: for iteration j = 1 to |L| do
15: for iteration k = 1 to |A| do
16: Ranki,j ← Rank(PVal, i, j)
17: end for
18: Distancei ← sum(Spearman(Rank, i), Kendall(i, Truth))
19: end for
20: Xbest, Pbest, H, Truth ← getBestVariant(Distance, PVal, A)

The Learning Phase. In this phase, TINTIN learns the best variant for finding Gbest (Lines 1-2 in Algorithm 1). In particular, three variants (Table 2) of TINTIN utilizing different feature sets (only topological features, only dynamic feature and combination of topological and dynamic features) are considered. We begin by identifying the ground truth (true order of the ranking) of the candidate networks. The ground truth reflects the actual similarity of the target characteristics of the networks based on prior knowledge or empirical results. Hence, it can either be provided by experts familiar with the candidate networks and its associated targets or generated automatically by using models that characterize the known targets in these networks. We adopt the latter strategy by utilizing Tenet[6]. Specifically, in order to generate the ground truth, we exploit Tenet in the following way. Given a reference network G and two candidate networks L1 and L2, L1 is more similar to G if the characterization model of L1 produces a better characterization of known targets in G compared to L2. That is, the characterization model of L1 obtains a larger AUC (area under ROC curve) for known targets in G compared to that of L2. Note that AUC is typically used to assess classifier performance as the metric is robust for imbalanced datasets [9]. The ground truth can be interpreted as the ordering of the candidate networks based on decreasing AUC.

Next, all the features for targets in all candidate networks are extracted and the combined p-value for the Wilcoxon and the KS tests for each variant is computed. Then, for each candidate network Lj, three ranked lists (denoted as r), each corresponding to one variant (see Table 2), are obtained by ordering the remaining candidate networks in decreasing order of the combined p-value. Next, the disagreement between r and the ground truth is measured using the Spearman footrule distance and Kendall-Tau distance, denoted as Y(.) and A(.), respectively. Given two complete rankings r1 and r2 of a set of N individuals, let r1(i) be the rank of i ∈ N in the ranked list r1. Then, the distances are calculated as:

\[
Y(r_1, r_2) = \sum_{i \in N} |r_1(i) - r_2(i)|
\]

\[
A(r_1, r_2) = |\{(i, j) : r_1(i) < r_1(j) \text{ and } r_2(i) > r_2(j)\}|
\]

For each variant, the Spearman footrule and Kendall-Tau distances for all the candidate networks are aggregated to obtain the overall distance. The best variant is the one yielding the least overall distance. Furthermore, this phase determines a p-value threshold (denoted as pt) that shall be used to filter off dissimilar network (Algorithm 1, Line 2). First, it obtains a mapping between the combined p-values and “poor” AUC. Then, these combined p-values are averaged to obtain pm. The value of pt is set as the minimum of pm and pt since combined p-values less than pm are removed.

The Ranking Phase. In the next phase, the TINTIN algorithm identifies the best matched network Gbest and rank of the candidate networks with combined p-values greater than pt. First, values of predictive topological features is extracted for G. Then, for each pair of (G, Lj), the Wilcoxon and KS tests are performed for each of these features and the p-values obtained are combined. Finally, the candidate networks with combined p-values greater than or equal to pt are ranked in order of decreasing combined p-values to obtain r. The top-ranked network is Gbest.

Observe that it is possible for our ranking strategy to return no best matched network if all the combined p-values are less than pt. This indicates that none of the given candidate networks are similar. In this case, we can either explore additional candidate networks to identify Gbest or relax pt using p to obtain a suboptimal Gbest.

Theorem 5.1. The worst-case time complexity of Algorithm 1 is O(|L||L| − 1)(G(Xall) + |A|(|V L[i]| |L[j]| 3)) time in the worst case.

We deem AUC<0.5 as “poor” since it indicates performance worse than random prioritization.

Table 2: Variants of our network similarity strategy.

<table>
<thead>
<tr>
<th>Variant</th>
<th>All Features</th>
<th>T. Features</th>
<th>D. Features</th>
<th>Stouffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>TINTIN</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>TINTIN</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>TINTIN</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

Figure 1: Ground truth found using AUROC.
case, where \( G(\mathcal{X}_{d_l}) \) is the worst time complexity for extracting all features, \( |V_{L(i)}| \) is the number of nodes of the \( i \)-th network in \( L \) and \( |A| \) is the number of TINTIN variants.

The proof of the above theorem is given in [7].

6 EXPERIMENTS

TINTIN is implemented using Java. In this section, we investigate its performance. All experiments are performed on a computer system using a 64-bit operating system with 8GB RAM and a dual core processor running at 3.60GHz.

Datasets. We use a reference network (I0) and four candidate networks (I1 to I4) for our experiments as shown in Table 3. Note that although there are more than 600 curated signaling networks in Biomodels, we restrict our study to only five signaling networks. This is because we need to identify known targets of signaling networks for validating our experimental results. Unfortunately, to the best of our knowledge, there is no publicly available technique that can automatically identify known targets from signaling networks by analyzing biomedical literature. Hence, we are confined to manual target curation from a large volume of biomedical literature, a time-intensive process. Also, although larger signaling networks are desirable, to the best of our knowledge, no publicly-available large signaling networks (e.g., human cancer signaling network) contain dynamic information of all edges (dyes), preventing us to exploit dynamic features such as psd. The targets of I0 are curated from [16] (Ca++-, EGFR, EGFR, activated EGFR and Ras) and from [3] (dimerized EGFR). The curated targets of the candidate networks are given in [7].

Best TINTIN variant. First, we identify the best variant of TINTIN. Specifically, we examine how various variants (Table 2) perform on the given set of candidate networks (I1 to I4). Then, we examine if the best performing variant for the candidate networks is also effective in identifying the best matched network for the reference network. Finally, we analyze their runtime performance.

Figure 2 reports the results. Interestingly, we observe that the variant using only dynamic feature performs better (smaller distance between ground truth and TINTIN ranking) than variants that either use only topological feature set or a combination of topological and dynamic feature sets. This underscores the importance of considering dynamic feature. The best performing variant obtained from the learning phase (LEARNBDESTVARIANT) is TINTIN\(_D\).

Next, we examine the effect of applying different variants on the reference network to validate if the learning phase yields the desired best performing variant. Indeed, TINTIN\(_D\) is the top performing variant for the reference network (Figure 2, top, reference network). In fact, majority of the TINTIN variants identify I4 as the best matched network (Figure 2, middle).

In terms of the runtime performance (Figure 2, bottom), the learning process (Step 2) consumes the bulk of its execution time. Hence, learning can be performed offline to improve the runtime performance. The improvement is up to 2 orders of magnitude for certain variants (e.g., TINTIN\(_4\) for comparison of networks with fewer than 100 nodes). In addition, we observe that the runtime performance is dependent on the size of the network, the types and number of features used by the variant. In subsequent experiments, we shall use TINTIN\(_D\).

Performance of the LEARNTHRESHOLD procedure. In this set of experiments, we identify the threshold \( p_1 \) learnt from the candidate networks. Table 4 shows the combined \( p \)-values of each pair of candidate networks using TINTIN\(_D\) and the corresponding AUROC when applying the characterization model of one candidate network to another. \( p_1 \) is the average combined \( p \)-value of the cells marked with \( \$ \). That is, \( p_1 = \frac{2.2 \times 10^{-16}+2.2 \times 10^{-16}+6.045}{3} = 0.01. \) When we apply \( p_1 = 0.01 \) to the ranked list of I1 to I4 obtained using TINTIN\(_D\), we note that I1 (AUROC=0.632, combined \( p \)-value=2.2x10^{-16}) is considered as a dissimilar network and filtered off. Note that \( p_1 \) affects the number of networks being ranked, but not the actual rank of the network. Hence, it is possible to have no best matched network if the given set of candidate networks is considered dissimilar to the unseen network.

Comparison with state-of-the-art. Lastly, we compare TINTIN\(_D\) against GDD [17] and NETSIMILE [4] in terms of network ranking and runtime performance. We consider both the arithmetic and geometric versions of GDD which are denoted as GDD\(_A\) and GDD\(_D\), respectively. Since these target-unaware network similarity approaches define similarity differently from TINTIN (Section 2), it is
not possible to have a good network ranking benchmark that can be used as ground truth for comparison. Instead, we compare the similarities and differences in the rankings derived from TINTIN as compared to that from other approaches. As observed in Figure 3 (top), our approach differs in ranking of networks I₁ to I₄ when compared to traditional target-unaware network similarity approaches. The differences in ranking is more significant when we compare TINTIN against target-unaware approaches (Table 5, left) versus a comparison among traditional target-unaware approaches only (Table 5, right). In particular, I₄ was ranked best by TINTIN and worst by the target-unaware approaches (Figure 3, top).

The runtime performance is affected by the network size (Figure 3, bottom), of which the most significant impact is experienced by the GDD-based approaches. In particular, NetSIMILE performs the best and scales well even for larger networks. The order of the approaches in terms of runtime performance is NetSIMILE < TINTIN < NETSIMILE (offline learning) < TINTIN < GDD. Hence, our approach has moderate runtime performance and the ranks produced are markedly different from state-of-the-art approaches.

7 CONCLUSIONS & FUTURE WORK

In this paper, we present TINTIN, a target feature-based signaling network similarity computation and ranking technique by exploiting the topological and dynamic characteristics of the targets. It is interesting to note that the empirical study highlights a single dynamic feature (pssd) as being more important than a set of topological features in identifying the best matched network. This signals the importance of considering dynamic features in measuring network similarity. However, to the best of our knowledge, majority of the work on network feature focus on topological features instead. Hence, as part of future work, we intend to explore novel dynamic network features that could be used to study signaling networks. In addition, our empirical results demonstrate the differences in terms of network ranking of TINTIN compared to state-of-the-art target-unaware network similarity techniques.

Acknowledgments. Huey Eng Chua and Sourav S Bhowmick were partially supported by MOE AcRF Tier-1 Grant RGC 1/13.

REFERENCES