Effective Estimation of Node-to-Node Correspondence Between Different Graphs

Hyundoo Jeong, *Student Member, IEEE,* and Byung-Jun Yoon, *Senior Member, IEEE*

Abstract

In this work, we propose a novel method for accurately estimating the node-to-node correspondence between two graphs. Given two graphs and their pairwise node similarity scores, our goal is to quantitatively measure the overall similarity – or the correspondence – between nodes that belong to different graphs. The proposed method is based on a Markov random walk model that performs a simultaneous random walk on two graphs. Unlike previous random walk models, the proposed random walker examines the neighboring nodes at each step and adjusts its mode of random walk, where it can switch between a simultaneous walk on both graphs and an individual walk on one of the two graphs. Based on extensive simulation results, we demonstrate that our random walk model yields better node correspondence scores that can more accurately identify nodes and edges that are conserved across graphs.

Index Terms

Graph comparison, node correspondence, random walk, pair-HMM (pair hidden Markov model).

EDICS Category: SAS-STAT

Hyundoo Jeong is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA.

Byung-Jun Yoon is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA, and the College of Science, Engineering, and Technology, Hamad Bin Khalifa University (HBKU), Doha, Qatar.

This work was supported in part by the National Science Foundation through the NSF Award CCF-1149544.

Effective Estimation of Node-to-Node Correspondence Between Different Graphs

I. Introduction

Graph-based system and data analysis techniques have become a critical tool in many fields as it can provide an intuitive way of representing interactions between variables and analyzing them [1]–[4]. In recent years, graph-based techniques have been widely applied to the analysis of social networks [5], [6], images [7], [8], and biological networks [9], [10]. Given multiple graphs, one question that is of practical importance is how the nodes in a given graph can be mapped to nodes in the other graphs based on the similarity between nodes and the topological similarity between graphs. Considering that each node may have a number of similar nodes in the other graphs and that the graphs may have significant differences in their topology, quantitatively estimating this overall similarity between nodes – or the *node correspondence* – is theoretically challenging. Furthermore, estimating these similarities can pose computational challenges, especially for large graphs, due to the combinatorial nature of the problem.

So far, several methods have been proposed for measuring the node correspondence between graphs, where random walk based methods have been popular as they are intuitive and can be efficiently implemented [10]–[15]. These methods perform a simultaneous random walk on the two graphs to be compared, where the random walk scheme is designed such that the walker more frequently visits (or stays longer at) node pairs that have higher similarity and are surrounded by a larger number of similar node pairs. The stationary probability of the resulting (semi-)Markov model gives us the long-run proportion of time that the random walker simultaneously visits (and stays at) a given node pair, which can be used as the correspondence score between the two nodes. This score provides a simple and intuitive way of measuring the overall similarity between two nodes in different graphs by integrating the node similarity and the topological similarity [10]. Recently, these random walk models have been applied to the comparative analysis of large-scale biological networks [12], [13].

For example, the scoring method adopted by IsoRank [12], a *global network alignment* algorithm that aims to find the best overall mapping between different biological networks, utilizes a Markov random walk model with restart probability. In IsoRank, the random walker moves

on the product graph of the two graphs to be compared, where each node in this product graph corresponds to a pair of nodes, one from each graph. At each time step, the random walker will randomly move to one of the neighbors in the product graph with equal probability. Additionally, the model allows the random walker to either continue the random walk or restart at new position, where the probability of the restart position is made proportional to the similarity of corresponding node pair. This restart scheme allows the random walk model to capture both topological similarity and the pairwise node similarity into the resulting node correspondence score. More recently, another global network alignment algorithm called SMETANA [13] adopted a scoring method that uses a semi-Markov random walk model [14]. In this method, the random walker spends a different amount of time at each location, which is proportional to the similarity of the corresponding node pair. Evaluation based on synthetic and real biological networks have shown that the long-run proportion of time that the semi-Markov random walker spends at a node pair can serve as an effective measure of the correspondence between the given nodes [13].

In this paper, we propose a novel random walk model that can significantly improve the accuracy of the estimation of the node correspondence between different graphs. The proposed random walker performs a random walk on the two graphs to be compared, where it can switch its mode between a simultaneous walk on both graphs and an individual walk on one of the graphs. The mode switching is determined by the presence (or absence) of similar node pairs among the current neighbors. Through extensive simulations, we show that the proposed model leads to an enhanced node-correspondence scoring method that clearly outperforms existing methods.

II. Estimating the Correspondence Between Nodes in Different Graphs

A. Problem Statement

Consider two graphs $\mathcal{G}_{\mathcal{U}} = (\mathcal{U}, \mathcal{D})$ and $\mathcal{G}_{\mathcal{V}} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{G}_{\mathcal{U}}$ consists of a set $\mathcal{U} = \{u_1, u_2, \cdots\}$ of nodes and a set $\mathcal{D} = \{d_{ij}\}$ of edges between nodes u_i and u_j and $\mathcal{G}_{\mathcal{V}}$ consists of a set $\mathcal{V} = \{v_1, v_2, \cdots\}$ of nodes and a set $\mathcal{E} = \{e_{\ell m}\}$ of edges between nodes v_ℓ and v_m. We assume that a nonnegative pairwise node similarity score $s(u_i, v_j)$ is given for every node pair (u_i, v_j). Our goal is to estimate the node correspondence score $c(u_i, v_j)$ for every node pair (u_i, v_j) that quantifies the overall similarity between these nodes by integrating the pairwise node similarity scores and the topological similarity between the two graphs in a reasonable manner. In other

words, we want the node correspondence score $c(u_i, v_j)$ to be proportional to the posterior alignment probability $P[u_i \sim v_j | \mathcal{G}_\mathcal{U}, \mathcal{G}_\mathcal{V}]$ of u_i and v_j given $\mathcal{G}_\mathcal{U}$ and $\mathcal{G}_\mathcal{V}$.

B. Motivation and Overall Approach

We propose a novel random walk model to measure the node correspondence score $c(u_i, v_j)$. Our random walk model is motivated by the pair hidden Markov model (pair-HMM), which has been widely used for the comparative analysis of biological sequences (e.g., sequence alignment) due to its simplicity and effectiveness [16], [17].

Unlike traditional HMMs, which generate a single symbol sequence, the pair-HMM generates a pair of aligned symbol sequences. A typical pair-HMM has three different states: M, I_1, and I_2. At the M state (indicates a "matched" symbol pair), the HMM emits an aligned symbol pair. On the other hand, at the I_k state (indicates an "inserted" symbol in either sequence), the HMM only emits a symbol to sequence-k alone that is aligned to a gap symbol in the other sequence. Given two (unaligned) symbol sequences, we can use the forward-backward algorithm to predict the alignment probabilities between symbols in the two sequences based on the pair-HMM.

Similarly, the proposed random walk model has three different internal states, M, I_U, and I_V, where each state corresponds to a different "mode" of random walk. At a M state, which corresponds to "matched" node pair, the random walker makes a simultaneous walk on both graphs, moving into a pair of matched nodes. This is illustrated in Fig. 1(a). On the other hand, at state I_U (or state I_V), the random walker makes an "individual" walk on graph $\mathcal{G}_\mathcal{U}$ (or $\mathcal{G}_\mathcal{V}$). Figure 1(b) illustrates the individual walk at state I_U. The random walker can switch its mode between a simultaneous walk and an individual walk, in a context dependent way by examining the neighborhood. In the presence of node pairs in the immediate neighborhood with a positive node similarity score, the random walker will make a simultaneous move on both graphs by randomly moving into one of the similar node pairs (M state). Otherwise, the random walker will make a transition to either state I_U or I_V and make a random move only on the corresponding graph.

Based on this random walk model, we estimate the stationary probabilities of this random walk, or in other words, the long-run proportion of time that the random walker will simultaneously visit a given node pair. Finally, from these stationary probabilities, we estimate the actual proportion of time that the random walker spends at a given node pair by "entering"

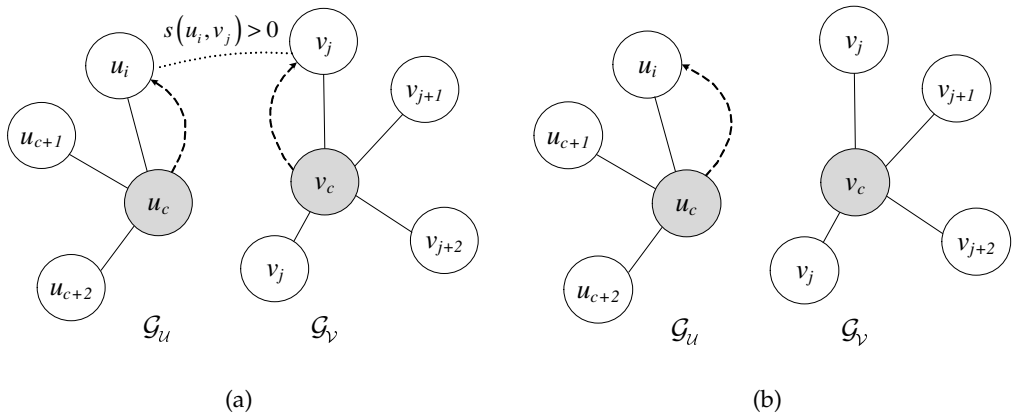

Fig. 1. Illustration of the proposed random walk model: (a) simultaneous walk on both graphs; (b) individual walk on either graph ($\mathcal{G}_\mathcal{U}$ in the example shown). The shaded nodes show the current position of the random walker on the two graphs. The dashed arrows indicate the movement of the random walker at the next time step.

the nodes *simultaneously* (i.e., at state M), which we used as the correspondence score for the node pair. It should be noted that this last step is crucial, since we are not interested in the case when the random walker happens to stay at a node pair as a result of an individual move on one of the graphs. In such cases, the simultaneous visit of the two nodes is coincidental and is not a direct result of the relevance between the given nodes.

C. Proposed Method

Let $\mathcal{G}_\mathcal{X} = (\mathcal{X}, \mathcal{E}_\mathcal{X})$ be the product graph of $\mathcal{G}_\mathcal{U}$ and $\mathcal{G}_\mathcal{V}$, where the nodes in the graph $\mathcal{G}_\mathcal{X}$ correspond to node pairs (u_i, v_j), $u_i \in \mathcal{U}$ and $v_i \in \mathcal{V}$. Two nodes in the product graph $\mathcal{G}_\mathcal{X}$ are connected if and only if the corresponding nodes are connected in both $\mathcal{G}_\mathcal{U}$ and $\mathcal{G}_\mathcal{V}$. Joint random walk on the two graphs $\mathcal{G}_\mathcal{U}$ and $\mathcal{G}_\mathcal{V}$, both simultaneous walk and individual walk, can be viewed as a random walk on this product graph $\mathcal{G}_\mathcal{X}$. We define $\mathcal{M} = \{(u_i, v_j) \,|s\,(u_i, v_j) > 0, u_i \in \mathcal{U}, v_j \in \mathcal{V}\}$ as the set of similar node pairs, where $s\,(u_i, v_j)$ is the pairwise node similarity score for the node pair (u_i, v_j). Suppose that the random walker is currently located at (u_c, v_c) for some $u_c \in \mathcal{U}$ and $v_c \in \mathcal{V}$. Let us define the set of similar node pairs in the neighborhood of (u_c, v_c) as $\mathcal{N}(u_c, v_c) = \{(u_i, v_j) \,|u_i \in \mathcal{N}(u_c), v_j \in \mathcal{N}(v_c), (u_i, v_j) \in \mathcal{M}\}$, where $\mathcal{N}(u_c)$ is the set of neighbors of node u_c in graph $\mathcal{G}_\mathcal{U}$ and $\mathcal{N}(v_c)$ is the set of neighbors of node v_c in graph $\mathcal{G}_\mathcal{V}$.

If there are similar node pairs in the current neighborhood, hence $\mathcal{N}(u_c, v_c) \neq \varnothing$, the random walker makes a simultaneous move on both graphs, from (u_c, v_c) to (u_i, v_j), according to the

following transition probabilities

$$P\big[(u_i, v_j) \,|\, (u_c, v_c)\big] = \frac{s\,(u_i, v_j)}{\sum_{(u_{i'}, v_{j'}) \in \mathcal{N}(u_c, v_c)} s\,(u_{i'}, v_{j'})}. \tag{1}$$

On the other hand, if there is no similar node pair in the neighborhood, hence $\mathcal{N}(u_c, v_c) = \varnothing$, the random walker randomly selects either $\mathcal{G}_\mathcal{U}$ or $\mathcal{G}_\mathcal{V}$ and performs an individual walk only on the selected graph. The probability that each graph will be selected is proportional to its size (i.e., number of nodes in the graph), and in the selected graph, the random walker will move into one of the neighboring nodes with equal probability. The resulting transition probabilities are given by

$$P\big[(u_i, v_c) \,|\, (u_c, v_c)\big] = \frac{|\mathcal{U}|}{|\mathcal{U}| + |\mathcal{V}|} \times \frac{1}{|\mathcal{N}\,(u_c)|} \tag{2a}$$

$$P\big[(u_c, v_j) \,|\, (u_c, v_c)\big] = \frac{|\mathcal{V}|}{|\mathcal{U}| + |\mathcal{V}|} \times \frac{1}{|\mathcal{N}\,(v_c)|} \tag{2b}$$

for $u_i \in \mathcal{N}(u_c)$ and $v_j \in \mathcal{N}(v_c)$. From (1), (2a), and (2b), we can construct the transition probability matrix $\mathbf{P}$ for the random walk on the product graph $\mathcal{G}_\mathcal{X}$. In practice, the matrix $\mathbf{P}$ will be often sparse, as the original graphs $\mathcal{G}_\mathcal{U}$ and $\mathcal{G}_\mathcal{V}$ that arise in practical applications will be typically sparse. This property makes it easy to compute the stationary probability $\pi(u_i, v_j)$ of the random walk using the power method [12], [13], [21]. Given $\pi(u_i, v_j)$, we finally compute the actual proportion of time $\hat{\pi}(u_i, v_j)$ that the random walker spends at (u_i, v_j) by entering the node pair through a simultaneous random walk (i.e., at state M) as follows

$$\hat{\pi}(u_i, v_j) = \sum\nolimits_{(u_p, u_q) \in \mathcal{N}(u_i, u_j)} \pi(u_p, u_q) \cdot P\big[(u_i, v_j) \,|\, (u_p, v_q)\big], \tag{3}$$

for all $(u_i, v_j) \in \mathcal{M}$. Finally, we define the correspondence score between two nodes u_i and v_j as $c(u_i, v_j) \equiv \hat{\pi}(u_i, v_j)$, where $u_i \in \mathcal{G}_\mathcal{U}$ and $v_j \in \mathcal{G}_\mathcal{V}$. As we will demonstrate in the following section, the proposed scoring scheme effectively quantifies the overall similarity between nodes in different graphs by seamlessly integrating the pairwise node similarity and the topological similarity between graphs.

III. Simulation Results

In order to demonstrate the effectiveness of the proposed scoring method, we performed extensive simulations based on synthetic graphs [18] as well as real biological networks [19]. To evaluate the performance, we computed the node correspondence scores using the proposed scheme, and used the scores to predict the graph alignment through greedy one-to-one mapping.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT SCORING METHODS

	Pair. Sim. Score	IsoRank	SMETANA	Proposed
CN	519	549	533.4	**704.4**
MNE	0.28	0.31	0.27	**0.15**
CE	510.2	581.5	554.3	**1,000.4**

More specifically, we started from an empty alignment and built up the graph alignment by iteratively adding one node pair at a time according to its correspondence score in a descending order. Given the final alignment, we define the equivalence class as the set of nodes that are aligned to each other. A given equivalence class is said to be correct if the aligned nodes have the same label, indicating that they belong to the same functional class. We computed three different metrics to assess the goodness of the predicted alignment: correct nodes (CN), mean normalized entropy (MNE), and conserved edges (CE). CN is the total number of aligned nodes that belong to the correct equivalence class. The coherence of the node mapping can be accessed by MNE. MNE for a given equivalence class $\mathbf{C}$ can be computed by $H(\mathbf{C}) = -\frac{1}{\log d}\sum_{i=1}^{d} p_i \log p_i$, where p_i is the relative proportion of nodes in $\mathbf{C}$ with label i and d is the total number of different labels. A mapping with higher coherence will lead to a lower entropy. CE counts the total number of conserved edges between aligned nodes in the predicted graph alignment. CE can be used to assess the performance of detecting conserved topological structures across graphs. For comparison, we repeated similar experiments by using two state-of-the-art scoring schemes used in IsoRank [12] (parameter α was set to 0.6 as in the original paper) and SMETANA [13].

A. Overall Performance using Synthetic Datasets

Using the NAPAbench package [18], we generated 10 pairs of synthetic graphs based on the crystal growth model [20], where each pair consists of a graph with 750 nodes and another graph with 1,000 nodes. On average, the smaller graphs had around 3,000 edges and the larger graphs had around 4,000 edges. For every pair of graphs, the true correspondence between the nodes in the two graphs are known, hence we can evaluate the effectiveness of the proposed scheme. Table I shows the performance of different scoring methods. The proposed method clearly outperforms all other methods. For example, the proposed scoring method finds around 30 percent more correct nodes compared to the scoring methods in IsoRank [12]

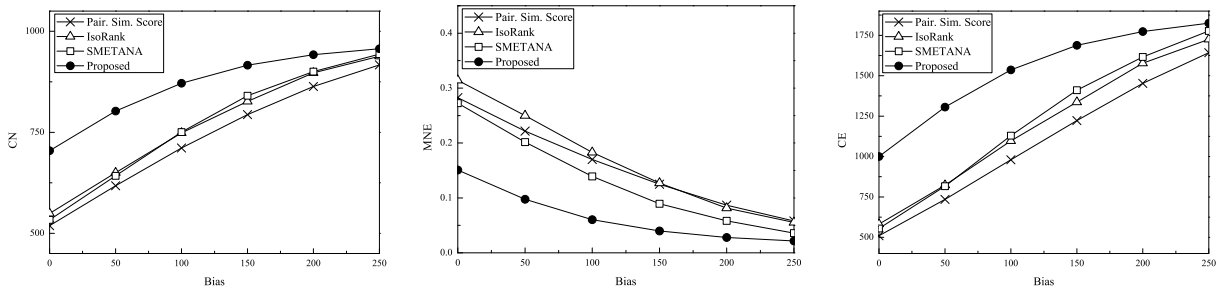

Fig. 2. Performance dependence on pairwise node similarity: correct nodes (left), mean normalized entropy (middle), and conserved edges (right).

and SMETANA [13]. Furthermore, the proposed method yields a more coherent mapping as indicated by the lower MNE. It is also important to note that our proposed method results in significantly higher CE, which implies that the resulting node correspondence scores capture the topological similarity between graphs more effectively. The mean computation time (in second) was 0.42 for SMETANA [13], 7.8 for IsoRank [12], and 16.4 for the proposed method. The increase of computational cost for the proposed method is mainly due to the larger amount of time needed to construct the transition probability matrix $\mathbf{P}$, in exchange for the substantial performance improvement shown in Table I.

B. Performance Dependence on Pairwise Node Similarity

Next, we evaluated the influence of the pairwise node similarity scores on the performance of each method. For this purpose, we introduced an additional bias term to further separate the distribution of the pairwise node similarity score between nodes with the same label and the score distribution for nodes with different labels. A higher bias makes it easier to predict the correspondence between nodes in different graphs based on the pairwise node similarity score alone (i.e., without taking topological similarity into account). Figure 2 shows that the proposed method significantly outperforms other scoring methods for a wide range of bias. As we would expect, the performance difference between the proposed method and the other methods decreases with an increasing bias, as it becomes easier to distinguish relevant nodes from irrelevant ones.

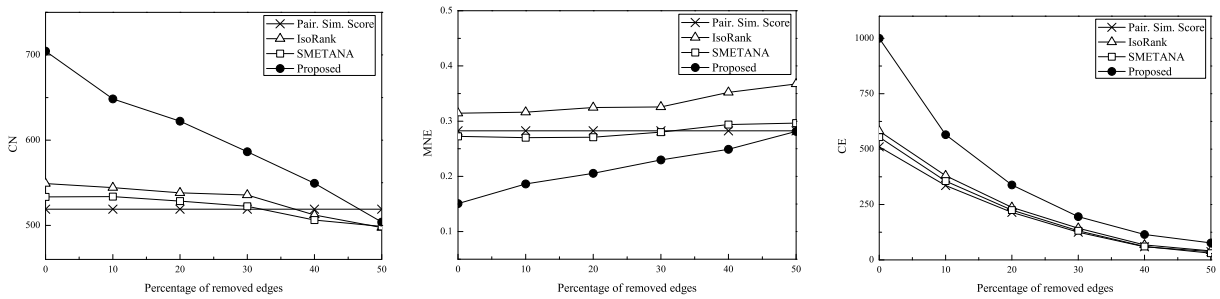

Fig. 3. Performance dependence on sparsity of graphs: correct nodes (left), mean normalized entropy (middle), and conserved edges (right).

C. Performance Dependence on Sparsity of the Graphs

We also assessed the effect of the sparsity of the graphs on the effectiveness of the estimated node correspondence scores. To this aim, we randomly removed 10%∼50% edges in both graphs. Figure 3 shows that the proposed method outperforms other methods in most cases. One interesting observation is that the performance of the proposed method degrades more quickly as we remove more edges. In fact, this is another evidence that the proposed scoring scheme incorporates topological similarity more effectively compared to other methods. As more edges are removed, the graphs will bear lower topological similarity, which should be reflected in the effectiveness of the estimated node correspondence scores.

D. Performance Evaluation using Real Datasets

For further evaluation, we performed simulations using the human protein-protein interaction (PPI) network in [19]. We randomly extracted 10 pairs of connected networks, one with 750 nodes and the other with 1,000 nodes, from the human PPI network. Each pair was obtained by randomly selecting a highly-connected seed node and independently extending the two networks from the same seed. Next, we perturbed their pairwise node similarity scores by adding random noise that follows a gamma distribution $\Gamma(k, \theta)$. We set the shape parameter k as 0.7 and change the scale parameter θ to verify effect of noise. Figure 4 shows that the proposed method is very robust to noise, while the performance of other methods is severely degraded due to their relatively strong dependence on pairwise node similarity scores. This result clearly demonstrates the potential of the proposed method in practical applications.

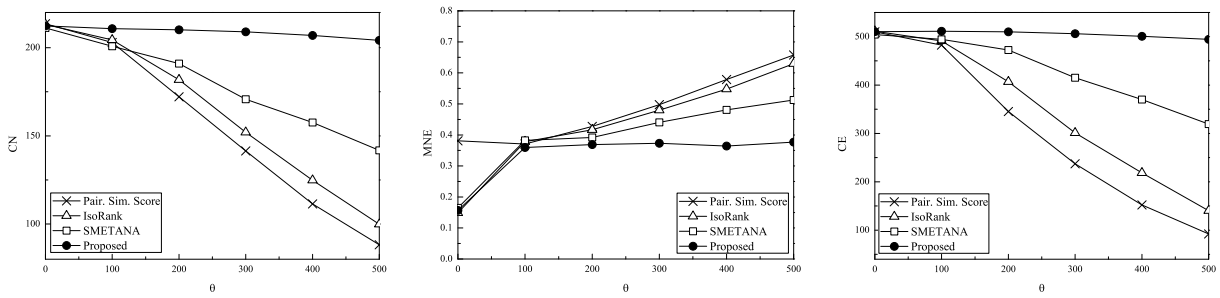

Fig. 4. Performance evaluation based on human PPI network: correct nodes (left), mean normalized entropy (middle), and conserved edges (right).

IV. Concluding Remarks

In this paper, we proposed a novel method for scoring the correspondence between nodes that belong to two different graphs. The proposed method utilizes a novel random walk model that switches between two different modes of random walk – simultaneous walk on both graphs and individual walk on either graph – in a context dependent manner. The node correspondence scores are estimated based on the stationary probabilities of the random walk. Simulation results show that the proposed scoring method significantly outperforms previous methods that rely on different random walk models in terms of accuracy and robustness. Our scoring scheme can provide an effective and computationally efficient foundation for comparative analysis of graphs, including biological networks and social networks.

References

[1] S. Umeyama, "An Eigendecomposition approach to weighted graph matching problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 5, pp. 695-703, Sept. 1988.

[2] C. Ding, T. Li, and M. I. Jordan, "Nonnegative matrix factorization for combinatorial optimization: spectral clustering, graph matching, and clique finding," in *Proc. IEEE ICDM 2008*, Pisa, Italy, pp. 183-192, Dec. 2008.

[3] M. Bayati, M. Gerritsen, D. F. Gleich, A. Saberi, and Y. Wang, "Algorithms for large, sparse network alignment problems," in *Proc. IEEE ICDM 2009*, Florida, USA, pp. 705-710, Dec. 2009.

[4] D. Koutra, H. Tong, and D. Lubensky, "BIG-ALIGN: fast bipartite graph alignment," in *Proc. IEEE ICDM 2013*, Texas, USA, pp. 389-398, Dec. 2013.

[5] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814-818, June 2005.

[6] L. Backstrom and J. LesKovec, "Supervised random walks: predicting and recommending links in social networks," in *Proc. ACM WSDM 2011*, Hong Kong, China, pp. 635-644, Feb. 2011.

[7] O. Duchenne, F. Bach, I.-S. Kweon, and J. Ponce, "A Tensor-Based Algorithm for High-Order Graph Matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2383-2395, Dec. 2011.

[8] X. Qian and B.-J. Yoon, "Shape matching based on graph alignment using hidden Markov models," *Proc. IEEE ICASSP 2010*, pp. 934–937, 2010.

[9] R. Sharan and T. Ideker, "Modeling cellular machinery through biological network comparison," *Nature Biotechnology*, vol. 24, pp. 427-433, Apr. 2006.

[10] B.-J. Yoon, X. Qian, and S. Sahraeian, "Comparative analysis of biological networks: hidden Markov model and Markov chain-based approach," *IEEE Signal Processing Magazine*, vol. 29, no. 1, pp. 22-34, 2012.

[11] S. M. E. Sahraeian and B.-J. Yoon, "RESQUE: network reduction using semi-Markov random walk scores for efficient querying of biological networks," *Bioinformatics*, vol. 28, no. 16, pp 2129-36, 2012.

[12] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, pp. 12763-12768, Sept. 2008.

[13] S. Sahraeian and B.-J. Yoon, "SMETANA: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks," *PLoS ONE*, vol. 8, no. 7, July 2013.

[14] S. Sahraeian and B.-J. Yoon, "A novel low-complexity HMM similarity measure," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 87-90, Feb. 2011.

[15] M. D. Colllins, J. Xu, L. Grady, and V. Singh, "Random walks based multi-image segmentation: quasi convexity results and GPU-based solutions," *IEEE CVPR 2012*, Rhode Island, USA, pp. 1656-1663, June, 2012.

[16] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*, Cambridge Univ. Press, 1998.

[17] B.-J. Yoon, "Hidden Markov models and their applications in biological sequence analysis," *Current Genomics*, vol. 10, no. 6, Sept. 2009.

[18] S. Sahraeian and B.-J. Yoon, "A network synthesis model for generating protein interaction network families," *PLoS ONE*, vol. 7, no. 8, Aug. 2012.

[19] D. Park, R. Singh, M Baym, and B. Berger, "IsoBase: A Database of Functionally Related Proteins Across PPI Networks," *Nucleic Acids Research*, vol. 39, pp. D295-D300, 2011.

[20] W. K. Kim and E. M. Marcotte, "Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence," *PLoS Computational Biology* vol. 4, no. 8, Nov. 2008.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Technical report, Stanford Digital Library Technologies Project, 1998.