The Intrinsic Geometric Structure of Protein-Protein Interaction Networks for Protein Interaction Prediction

Yi Fang, Mengtian Sun, Guoxian Dai, and Karthik Ramain

Abstract—Recent developments in high-throughput technologies for measuring protein-protein interaction (PPI) have profoundly advanced our ability to systematically infer protein function and regulation. However, inherently high false positive and false negative rates in measurement have posed great challenges in computational approaches for the prediction of PPI. A good PPI predictor should be 1) resistant to high rate of missing and spurious PPIs, and 2) robust against incompleteness of observed PPI networks. To predict PPI in a network, we developed an intrinsic geometry structure (IGS) for network, which exploits the intrinsic and hidden relationship among proteins in network through a heat diffusion process. In this process, all explicit PPIs participate simultaneously to glue local infinitesimal and noisy experimental interaction data to generate a global macroscopic descriptions about relationships among proteins. The revealed implicit relationship can be interpreted as the probability of two proteins interacting with each other. The revealed relationship is intrinsic and robust against individual, local and explicit protein interactions in the original network. We apply our approach to publicly available PPI network data for the evaluation of the performance of PPI prediction. Experimental results indicate that, under different levels of the missing and spurious PPIs, IGS is able to robustly exploit the intrinsic and hidden relationship for PPI prediction with a higher sensitivity and specificity compared to that of recently proposed methods.

Index Terms—Protein protein interaction network, complex network, computational biology

1 Introduction

This paper introduces an intrinsic geometry structure (IGS) of protein-protein interaction (PPI) network for the prediction of PPIs in network. Subsequent sections present the background of computational predication of PPI, review of the related methods, and a brief introduction to our proposed method.

1.1 Protein-Protein Interaction Prediction

Protein-protein interactions play important roles in assembling molecular machines through mediating many essential cellular activities [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. It is of important biological interest to analyze (PPI) network for deep understanding protein functions in cellular processes and biochemical events. The recent advancement in high-throughput technologies such as two-hybrid assays, tandem affinity purification, and mass spectrometry have provided tremendous amounts of PPIs in biological networks [14], [15]. The wealth of experimentally identified PPIs provide more opportunities in the exploration of protein functions and regulation in various organism. However, the labor-intensive experimental data are inherently associated with high false positives (FPs) and false negatives (FNs) which stir up many concerns in comprehensive analysis in understanding the PPI network [16], [17]. In addition, the identified PPI networks are somewhat incomplete as it is impractical to experimentally verify all individual PPIs within one cell [18]. These limitations can be complemented by the computational models for predicting PPIs from noisy experimental observations [15]. The complementary in silico approaches have been receiving more and more attentions in the assistance of PPI network analysis [1], [15], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29].

1.2 Review of Existing Prediction Methods

The computational approaches for the prediction of PPIs have been developed over the years [15]. We roughly break those approaches into two categories: the approaches based on integration of multiple data sources and the approaches solely based on topology of PPI network.

As our method falls into the later category, we mainly review the related works that address the problem of PPI prediction only using topology of network. For a review of earlier prediction approach, we refer readers to [15] for the discussion of applicability of computational methods to different types of prediction problems. The review summaries the prediction methods, including Gene neighbor and gene cluster methods, Phylogenetic profile methods, Rosetta Stone method, Sequence-based co-evolution methods, and classification methods. Although these prediction approaches can address the prediction of PPI to some extent, their performance rely on the biological

- *Y. Fang is with the Department of Electrical and Computer Engineering, New York University Abu Dhabi, Abu Dhabi, UAE. E-mail: yfang@nyu.edu.*
- *G. Dai is with the Department of Computer Science Engineering, New York University, New York, NY 11201. E-mail: guoxian.dai@nyu.edu.*
- *M. Sun and K. Ramani are with the School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907. E-mail: medavsun@gmail.com, ramani@purdue.edu.*

Manuscript received 31 Jan. 2015; revised 19 June 2015; accepted 23 June 2015. Date of publication 15 July 2015; date of current version 3 Feb. 2016. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2015.2456876

sources outside of PPI network, such as gene expression arrays, proteomics, and chromatin immunoprecipitation on chip assays.

The approaches in the later category are promising as they predict PPIs simply based on the topological connections in network and independence of any other biological sources [26], [30], [31], [32], [33], [34], [35], [36], [37]. In this category, the approaches differ from each other because they interpret the explicit topological connection of PPI network from different perspectives. The common neighbor method assumes that, in a network, two nodes are likely to interact with each other if they share many common neighbors [38]. However, as this method relies only on local topological information, the prediction will be biased by the local noisy interactions [19], [39], [40], [41]. A better way of capturing topological structure of the entire network is based on the consensus across all of individual interactions. A recent approach is proposed to generate a set of dendrograms and create a single consensus dendrogram to summarize network structure [38]. Such a consensus dendrogram captures the topological features that appear consistently across all or a large fraction of the dendrograms. It demonstrates a good performance in prediction of missing interaction [38]. However, the process of generating multiple dendrograms and creating a single consensus one is computational expensive and thus the method is inefficient in the application of large PPI network. Another approach to use topological information is propagation of local structure to a global view of network structure. In [33], [34], the authors introduced two indices called IG1 and IG2 based on the use of the local topology of a pair of proteins to rank their interaction probability. Furthermore, recently Yuan et al. [37] developed a generative network model for the prediction of protein-protein interactions in network.

One of well established propagation method, the shortest path propagation, has been recently introduced for the prediction of the PPI in networks [1]. Their approach achieves a good performance in PPI prediction with specificity of 85 percent and sensitivity of 90 percent. However, although it is able to capture the global structure of the network, it should be noticed that the shortest path propagation is known to be sensitive to short-circuit topological noise. The addition of spurious PPI would significantly affect the shortest path propagation. The random walk based diffusion propagation gains its advantage by progressively exploiting all possible linkages among proteins in the network [31], [32]. It is therefore robust to local noisy interactions. Authors in [31] introduce this propagation strategy to PPI prediction in the network and demonstrate good performance in their experiments. However, it is an open research for choosing an appropriate parameter of steps for the propagation process. This parameter determines the extent to which the global structure of a network is exploited. A greater propagation step allows to exploit more global structure however reduces the resolution to differentiate PPIs. A smaller propagation step allows to preserve a relatively higher resolution but the revealed structure is more sensitive to noise [42]. Contrast to random walk based diffusion propagation, the heat diffusion, governed by the eigenfunctions of Laplacian operator on network, can take account all of the local information at once to produce a consistent global solution.

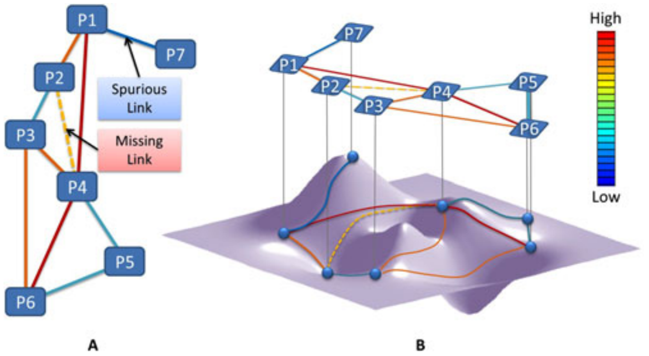


Fig. 1. Illustration of intrinsic geometric structure. (A) illustrates a toy PPI network consisted of seven proteins, which are denoted by P_1 to P_7. The solid line indicates an interaction between a pair of proteins, some of which are possibly spurious. The dashed line indicates possible missing link. (B) illustrates the intrinsic geometric structure of network on the left. The strength of revealed relationship among proteins gauged by its color. The closer the color to red end, the stronger the relationship between a pair of proteins is, the more likely they interact with each other in the original network. A possible missing link between P_2 and P_4 is identified since the revealed relationship by IGS is stronger (color in yellow). A potential spurious link between P_1 and P_7 is identified as well since the revealed relationship by IGS is weaker (color in blue).

1.3 Our Solution: Intrinsic Geometric Structure (IGS)

In this paper, heat diffusion is used to define intrinsic geometric structure of a PPI network. Because heat diffusion across network aggregates structural information about all possible paths connecting two nodes in network, it captures intrinsic relationship among nodes. Similar to random walk based diffusion propagation, the extent to which heat diffuses across network is scaled by the parameter of the dissipation time, which controls how globally the network structure is exploited. We propose a maximum likelihood based algorithm to determine the optimal dissipation time to balance the exploitation of the local and global structure of the entire PPI network. And the intrinsic geometric structure of PPI is defined as the revealed structure in heat diffusion process by the optimal dissipation time. The IGS organizes the proteins in the heat featured space according to their interactions in the network and has following three desirable properties: 1) it organizes information about intrinsic geometry of a PPI network in an efficient way, 2) it is stable under a certain number of missing and spurious interactions. 3) it faithfully interprets the implicit relationship with physics meaning supported. Fig. 1 illustrates the basic idea of IGS and how IGS detects spurious links and missing links in noisy PPI network. Fig. 1A demonstrates a toy PPI network consisting of seven proteins and nine PPIs in the network. The solid line between two proteins indicates there is a PPI between them. Fig. 1B depicts the intrinsic geometric structure for PPI network in (A). The new relationship between any pair of proteins is defined by the revealed structure (IGS). In the Fig. 1B, the strength of the relationship among proteins is represented by the color of the line between the nodes on the surface. The closer the color to red, the stronger two protein are. Spurious or missing links can be identified by investigating the strength of the revealed relationship. The stronger the relationship between a pair of nodes are, the more likely they are true PPI in the original PPI network. In the Fig. 1A, we suspect that the link between

P_1 and P_7 is a spurious link because only an isolated pair link between them. We also suspect that there should be a missing link between P_2 and P_4 because there are a couple of paths between two nodes. We can confirm the suspicion by the strength of new revealed relationships among proteins in Fig. 1B. There is a blue link between P_1 and P_7 indicating a weaker relationship, and there is a yellow link between P_2 and P_4 indicating a relatively stronger relationship. The PPI in original network could be redefined according to the strength of relationship defined by IGS.

The remainder of this paper is organized as follows. Section 2 provides the description of the method of defining the intrinsic geometric structure and the way to learn the intrinsic geometric structure. Section 3 describes experimental results on applications in PPI network. Section 4 presents additional discussion and concludes with description of the limitations of our current implementation and ways to address these limitations in the future.

2 Methods

In this section, we will first introduce fundamentals about heat kernel, and develop the proposed intrinsic geometric structure for PPI network.

2.1 Heat Kernel on Network

Heat transfer is a flow process of thermal energy from one region of matter or a physical system to another, which is mathematically governed by heat equation. Heat Kernel provides a fundamental solution to heat equation in the mathematical study of heat conduction and diffusion. The heat kernel records the evolution of temperature in a region whose boundary is held fixed at a particular temperature (typically zero), such that an initial unit of heat energy is placed at a point at some time. Intuitively, we could imagine that applying a unit amount of heat at one node i and allow the heat flow on the network across all edges, heat kernel measures the amount of the heat that passes from the node i to any other node j within a certain unit of time. Given a graph constructed by connecting pairs of data points with weighted edges, the heat kernel quantitatively codes the heat flow across a graph and is uniquely defined for any pair of data points on the graph. Suppose there is an initial heat distribution on network at time 0. The heat flow across the network is governed by the heat equation $u(x,t)$, where x denotes one node in the network and t denotes the time after the application of unit heat. The heat kernel provides the fundamental solution of the heat equation [43]. The heat kernel is closely associated with graph Laplacian by:

$$\frac{\partial H_t}{\partial t} = -LH_t, \tag{1}$$

where H_t denotes the heat kernel, L denotes the graph Laplacian and t denotes time.

2.2 Numerical Implementation of Heat Kernel

Graph Laplacian. As the graph Laplacian is important for solving the heat equation, we will introduce the graph Laplacian as follow. The PPI network under study is denoted as a graph $G = (V, E, W)$, where V is the set of notes, E is a set of edges, and W is the weight matrix,

$$W(i,j) = \begin{cases} 1 \ if \ i \neq j \\ 0 \ if \ i = j, \end{cases} \tag{2}$$

where $W_{i,j}$ is the weight of the edge connecting node i and node j and 1 denotes there is an interaction between a pair of proteins. The graph Laplacian is given as follows:

$$L = D - W, \tag{3}$$

where D is a diagonal degree matrix and its diagonal entries are given by the summation of rows of W:

$$D_{ii} = \sum_j W_{ij} \tag{4}$$

The normalized Laplacian of the graph is defined by

$$\hat{L} = D^{-1/2} L D^{-1/2}. \tag{5}$$

It is of great interest to analyze the spectral decomposition of the normalized Laplacian matrix $\hat{L}$. We can express $\hat{L}$:

$$\hat{L} = \Phi \Lambda \Phi^T, \tag{6}$$

where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_N)(\lambda_1 < \lambda_2 < \cdots < \lambda_N)$ is a diagonal matrix with ascending eigenvalues as the diagonal entries and Φ has the corresponding eigenvectors as its columns.

Heat Kernel by Eigenfunction Expansion. The heat kernel, defined in Eq. (1), can be expressed as the eigenfunction expansion by the graph Laplacian described below.

$$H_t = exp(-t\hat{L}), \tag{7}$$

where H_t is the heat kernel and $\hat{L}$ is the normalized graph Laplacian. By the Spectral Theorem, the heat kernel can be further expressed as follows:

$$H_t(i,j) = \sum_{k=1}^{|V|} e^{-\lambda_k t} \phi_k(i) \phi_k(j), \tag{8}$$

where λ_k is the k_{th} eigenvalue of the Laplacian and ϕ_k is the kth normalized eigenfunction. The eigenvalues are ordered so that

$$\lambda_1 \leq \lambda_2 \leq \cdots \lambda_{|V|},$$

and we pick an eigenvector for each eigenvalue. If the eigenvalue has geometric multiplicity one, the eigenfunction will be well-defined up to a scalar. The normalization of the eigenfunctions here refers to a choice of constants so that

$$\sum_i |\phi_k(i)|^2 = 1. \tag{9}$$

The quantity $H_t(i,j)$, defines the *heat affinity* between the pair of points i and j, is a measure of heat transfer between the two points after time t. We observe the symmetry:

$$H_t(i,j) = H_t(j,i), \tag{10}$$

reflecting the symmetry of the Laplacian, whereby the eigenvalues are guaranteed to be real and endowed with a complete basis of eigenvectors.

The heat kernel is sensitive to the structure of the network as it collects heat based information about all the possible paths between two nodes on the network. The heat flow on the network can be quantitatively approximated by the heat kernel $H_t(i,j)$, normally viewed as a function of two points i,j on the network at any given time t. The rate of diffusion over any one of the edges is determined by its weight. The value of the heat kernel $H_t(i,j)$ is the amount of heat accumulated at j after time t.

2.3 Intrinsic Geometric Structure

Heat kernel provides a transform by which the relationship among the data points is redefined according to the re-organization of all of the data simultaneously. The transform thus define a new relationship among proteins according to their topological connections. However, the heat kernel dynamically characterizes the proteins in the network from a local to global structure in the original network because it encapsulates the information about the heat flow over the time. The heat flow gradually aggregates information from local to global regions. At short time, heat kernel captures the local connectivity or topology of the network, while for long times the solution gauges the global geometry of the manifold on which the graph resides. However, there is one question remaining to answer: how to determine an appropriate time to balance how globally the structure of the entire network is exploited. The statistical interpretation of $H_t(i,j)$ arises from an exploration process: starting at node i, and exploring the entire network in all possible connections, the probability that j has been reached at the time t is $H_t(i,j)$ [42]. Based on the statistical interpretation, we proposed an approach to determining the time at which the likelihood for all of the observed PPIs is maximized. We formulate the optimization problem as follows. The likelihood function for the

$$L(t|PPI_1, PPI_2, \ldots, PPI_n) = \prod_{k=1}^{n} P(PPI_k|t), \tag{11}$$

where PPI_k denote the kth pair of the protein-protein interaction, n is the number of the total observed PPI in the network, t denotes the heat dissipation time and $P(PPI_k|t)$ denotes the probability that, starting from one node in kth pair of PPI, another node is reached by the time t, and equals to the value provided by heat kernel $H_t(PPI_k)$. In practice, it is convenient to convert to logarithm of likelihood function, defined as:

$$\ln L(t|PPI_1, PPI_2, \ldots, PPI_n) = \sum_{k=1}^{n} \ln P(PPI_k|t). \tag{12}$$

We are about to solve the following optimization problem to find the optimal time.

$$t = \arg\max_t \ln L(t|PPI_1, PPI_2, \ldots, PPI_n). \tag{13}$$

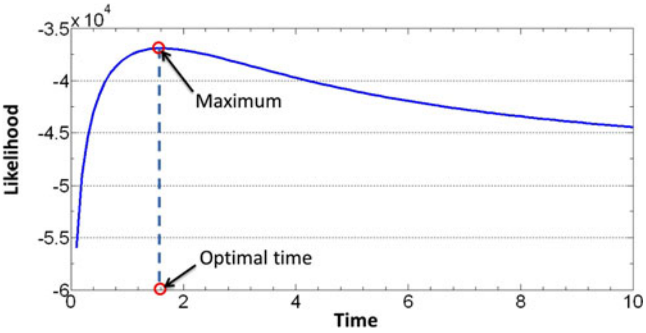


Fig. 2. Determination of optimal heat dissipation time. Figure illustrates the process of determination of the optimal time. By observation, likelihood value increases initially and monotonically decreases after obtaining its maximum at time around 1.5.

Basically, the IGS organizes the proteins in the heat featured space according to their interactions in network and has the following two desirable properties:

1) It organizes information about the intrinsic geometry of a PPI network in an efficient way,
2) It is stable under a certain number of missing and spurious interactions,
3) It faithfully interprets the implicit relationship with physics meaning supported.

The algorithm for solving the optimization problem in Eq. (13) is outlined below:

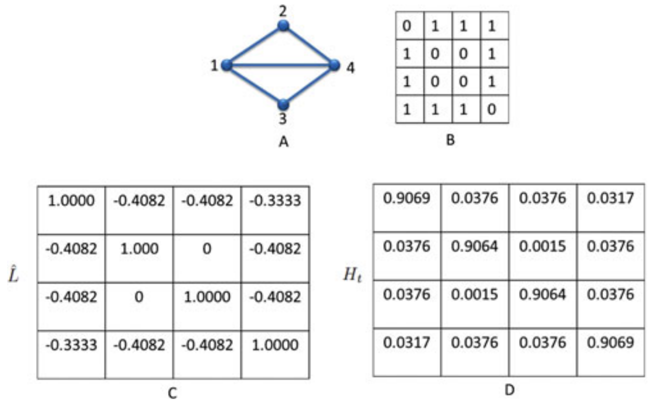
Algorithm 1. Determination of Intrinsic Geometric Structure

```
1: Given the network, $\Omega$ as the set of PPIs in the network, $i$ and $j$ denote $i$th and $j$th protein, respectively, $H_t$ denotes heat kernel and $H_t(i,j)$ denotes heat affinity between $i$th protein and $j$th protein by heat dissipation time of $t$. $IGS$ denotes the intrinsic geometric structure of the network.
2: Compute the graph Laplacian $L$ for the network and its normalized version $\hat{L}$.
3: Compute the eigenvalues and eigenvectors of $\hat{L}$,
4: Compute the heat kernel $H_t$ at each time $t$, and evaluate the likelihood defined in Eq. 13.
5: Terminate the iteration while likelihood function becomes stable or the number of the iterations reaches, we find the time $t$ at which the likelihood function reaches maximum (see Fig. 2). The located time is then chosen as $OptTime$
6: The $IGS$ is then given by $H_{OptTime}$
7: return $IGS$
```

Fig. 2 illustrates the process of determination of the optimal time. By observation, likelihood value increases initially and monotonically decreases after obtaining its maximum at time around 1.5. We then set the optimal time for the PPI network 1.5. We notice that the likelihood increases rapidly and then decreases slowly after crossing the optimal time. The likelihood remains stable while the heat distributes evenly on the network after a long dissipation time.

3 Results

To test the performance of IGS in PPI prediction in network, we carry out two experiments based on the experimental setting in [1], [38] and our new design of experiments



B

0	1	1	1
1	0	0	1
1	0	0	1
1	1	1	0

C ($\hat{L}$)

1.0000	-0.4082	-0.4082	-0.3333
-0.4082	1.000	0	-0.4082
-0.4082	0	1.0000	-0.4082
-0.3333	-0.4082	-0.4082	1.0000

D (H_t)

0.9069	0.0376	0.0376	0.0317
0.0376	0.9064	0.0015	0.0376
0.0376	0.0015	0.9064	0.0376
0.0317	0.0376	0.0376	0.9069

Fig. 3. Demonstration of Heat Kernel Calculation. Figure A illustrates the connections among four nodes. Figure B is the corresponding weight matrix (Eq. (2)) for those four nodes. Figure C is the normalized Laplacian matrix (Eq. (5)). Figure D is the heat kernel matrix at time $t = 0.1$, for example, $H_t(1,2) = 0.0376$.

against false positive and false negative PPIs in network. First, Two-class classifier test to differentiate the protein protein interaction and protein protein non-interaction (NPPI) at different noise levels. For the evaluation of the performance, we use the ROC (receiver operating characteristic curve, a graphical plot of the sensitivity versus (1 - specificity)) curve and precision-recall (PR) curve. Second, the test of the prediction of interaction from the network at different noisy level. For the evaluation of the performance, we use the area under curve (AUC). Our method is compared with two recent approaches for the prediction of PPI network [1], [38]. As our method mainly compare to the methods [1], [38], in the experiments, we denote the method proposed in [1] as MDS and the method proposed in [38] as HRG, and our method is denoted as IGS.

3.1 Data Sources

Since our method is mainly compared to the MDS method proposed in [1], we carried out our experiments on the same dataset which is used [1]. Therefore, we verify our approach on a publicly available S. cerevisiae network [44]. The S. cerevisiae network, which consists of $9,074$ interactions amongst $1,622$ proteins (we denoted it as CS2007 hereafter), is believed to be of high confidence. The high-confidence property of CS2007 makes it a good dataset for the test of a computational based method. To avoid the computational error for spectral decomposition of normalized Laplacian matrix in 6, we firstly identify the largest connected component of CS2007 which has 8,323 interactions between 1,004 proteins to form the final experimental dataset. In addition, to demonstrate the generality of our method, we also verify our approach on two other types of complex networks (non-PPI network) [38], [45], [46].

3.2 Computation of Heat Kernel Matrix

To better understand the calculation of heat kernel matrix, we provide a mini-example in this section. We created a set of four connected nodes as shown in Fig. 3A. We first calculate the weight matrix for the set of nodes based on Eq. (2). Fig. 3B shows the weight matrix with "1" indication of a connection and "0" indication of no connection. We then calculate normalized Laplacian matrix based on Eq. (5). Fig. 3C shows normalized Laplacian. The heat kernel matrix is calculated in the last step based on Eq. (7) or (8). Fig. 3D shows the heat kernel matrix at time $t = 0.1$. Each entry $H_t(i, j)$ in the matrix corresponds a heat kernel value at time t. For example, $H_t(1, 2) = 0.0376$ indicates the heat kernel value between node 1 and node 2 is 0.0376.

3.3 Two-Class Classifier Test

To validate the performance of IGS for differentiating the PPI and NPPI, we use the ROC curve and precision recall curve as the criteria. Both curves reflect how well IGS can robustly differentiates the PPI from the NPPIs based on the revealed and intrinsic relationship among proteins. To plot the ROC curve and PR curve, we should first define true position (TP), false positive, true negative (TN) and false negative. The TP measures the intersection between the new assigned PPIs set and the ground truth PPIs set, FP denotes the assigned edges which are not in the set of ground truth PPIs set, TN denotes the intersection of new assigned NPPIs and ground truth of NPPIs, and FN denotes new assigned NPPIs which are not in the set of ground truth NPPIs. The ROC and PR curve are computed based on heat affinity given by IGS as follows.

1) We vary the threshold from minimum to maximum value in the heat affinity set among all pairs of proteins.
2) For a given threshold, we compute the true positive (TP) function, true negative function, false positive function and false negative function.
3) Based on the values obtained in the previous step, we compute the sensitivity rate (TP/(TP+FN)) and specificity rate (TN/(TN+FP)), precise (TP(TP+FP)) and recall (TP/(TP+FN)). To plot the ROC curve, the horizontal axis represents (1 - specificity), and the vertical axis represents sensitivity. To plot the PR curve, the horizontal axis represents recall, and the vertical axis represents precision.

To demonstrate the robustness of IGS, we remove a fraction of true positive PPIs in the original network and plot the ROC curve at different levels of the removal of edges. The ROC curves are shown in the Fig. 4. The illustrated results are encouraging in terms of the prediction performance of PPI from the incomplete network. Without the removal of true PPI, the area under the ROC is nearly 1.00 and we can have specificity and sensitivity both over 0.95. The corresponding false positive rate (1-specificity) and false negative rate are all below 0.05. In addition, we find the IGS is very robust to the removal of the true PPIs in the network. As we can see from the Fig. 4, IGS performs pretty well even with 60 percent edges removed from the network. This result is appealing as most of the PPI network is incomplete with a large fraction of missing PPI in the observation in real scenario. Furthermore, IGS outperforms the MDS embedding method in this test [1]. They report a specificity 0.85 and sensitivity 0.90 in the experiment [1].

Following the experimental setting in [1], we further plot the precision and recall curve for accuracy analysis. The experimental result is shown in Fig. 5. Because the

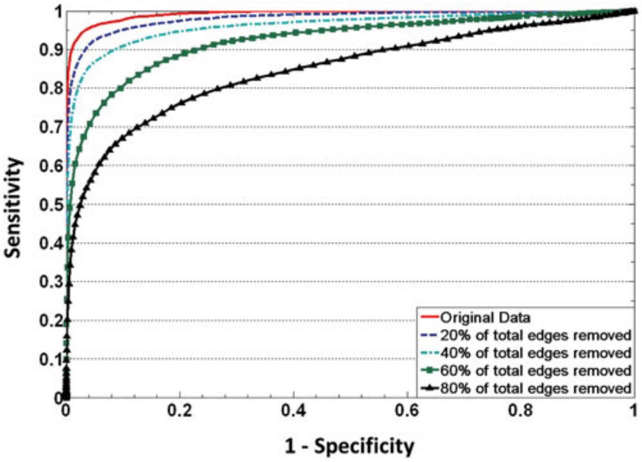


Fig. 4. Sensitivity-specificity test of IGS method. Five curves in the figure represent the ROC curves for our method at different levels of removal of edges.

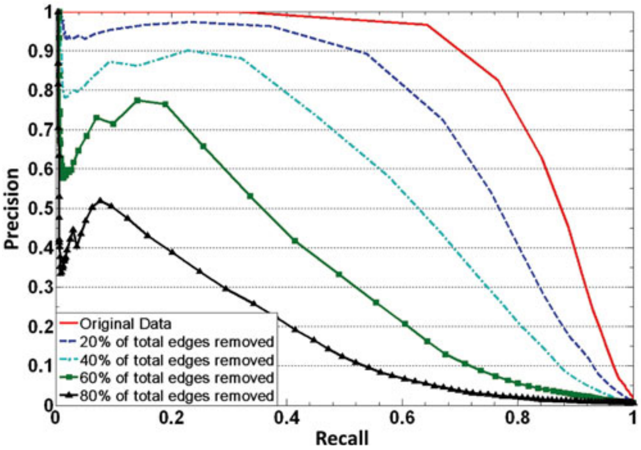


Fig. 5. Precision-recall test of IGS method. Five curves in the figure represent the PR curves for our method at different levels of removal of edges.

PPI network is really sparse, the fraction of true PPI is orderly lower than the fraction of true NPPI. A random predictor would give less than 1 correct TP in $1{,}000$ predictions, while the precision of PPI prediction of IGS can be over 0.90 at a recall about 0.60 for the original PPI network. The precision and recall analysis in [1] provides a precision of 0.15 at a recall about 0.35. With their level of precision and recall, they are able to reveal at least twice as many PPI available in BioGRID [1]. Our IGS is expected to give a much higher prediction of true PPI. In addition, we could see that the PR curve remains at a reasonable level even with 60 percent edges removed from the network. This result also indicates that the precision of prediction by IGS is not dramatically affected by the missing PPI in the network.

3.4 Prediction of Interaction in Networks

In this experiment, we demonstrate the performance of IGS in prediction of missing PPI and identification of spurious PPI in the noisy network. For an incomplete observed PPI network, we determine the IGS to fit the network and associate the heat affinity with each pair of proteins in the network. We are interested in the pairs of proteins that have high heat affinity but are not connected in the observed network, and the pairs of proteins that have low heat affinity but are connected in the observed network. The first type of pairs of proteins are most likely candidates for missing PPIs, and the second type of pairs are most likely candidates for spurious PPIs. Our method is tested with one PPI network and two other type of complex networks, and compared to two recent proposed methods in [1], [38].

For each network, we randomly remove a subset of connections for the simulation of missing PPI, and randomly insert a subset of connections for the simulation of spurious PPI. We attempt to predict the missing PPI and identify the spurious PPI. A well established criteria for quantifying the performance of prediction algorithms in machine learning area is the AUC, which can be calculated by the area under (ROC) curve. The AUC is often interpreted as the probability that a randomly chosen missing connected pair of nodes (true positive) is given a higher score by IGS than a randomly chosen unconnected pair of nodes (true negative) [38]. A random predictor will give AUC of score 0.5, and the extent to which the AUC by IGS exceeds 0.5 reflects how our prediction is better than chance.

3.4.1 Test on PPI Network

We use PPI network data (CS2007) to assess the performance of our method from two perspectives. First, we want to compare the performance in predictions of missing PPI using our method with that of MDS [1]. We evaluate the comparison by gradually increase the deletions of the true PPI and attempt to predict using the topology information remaining in the network. Second, we want to compare the performance in identification of spurious PPI using our method with that of MDS. We evaluate the comparison by gradually increase the insertions of the false PPI and attempt to identify using the topology information remaining in the network.

The first comparison result is shown in Fig. 6A. The horizontal axis represents the ratio between the number of the deleted PPIs (Fig. 6A) or the inserted PPIs (Fig. 6B), and the number of the true PPI in the network. Fig. 6A displays the the result about the first test. We can see from the figure that the AUC gradually decreases with the increase in the ratio of deletion from 0.2 to 0.8. The comparison of two curves shows that IGS consistently outperforms MDS in predicting the missing PPIs indicated by the higher values of AUC. The second comparison result is shown in Fig. 6B. Fig. 6B displays the the result about the second test. We can see from the figure that the AUC gradually decreases with the increase in the ratio of insertion from 0.2 to 0.8. The comparison of two curves shows that IGS consistently outperforms MDS in identifying the spurious PPIs indicated by the higher values of AUC.

3.4.2 Test Other Complex Network

In addition to testing PPI network, we further test IGS on other types of complex networks to demonstrate the generality of IGS. We compare IGS with one more recently proposed prediction method [38] and MDS. The method proposed in [38], name hierarchical random graphs (HRG),

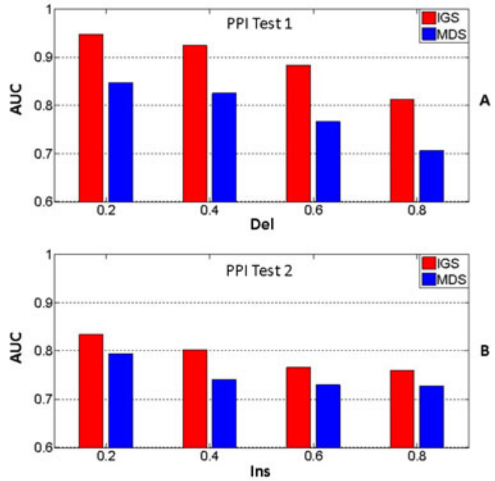


Fig. 6. Comparison of AUC test of IGS and MDS methods on CS2007. In (A), AUC values are computed for each method under certain level of deletions of true positive PPIs in the original network. “Del” on the horizontal axis displays the ratio between removed true PPIs and the number of PPIs in the original network and the vertical axis displays the corresponding AUC value. In (B), AUC values are computed for individual methods under certain level of insertions of false positive PPIs in the original network. “Ins” on the horizontal axis displays the ratio between inserted false PPIs and the number of PPIs in the original network and the vertical axis displays the corresponding AUC value.

also evaluate the probability of two nodes linking with each other. In this test, we predict the missing PPI but insert a certain levels of the spurious links to simulate the noisy links. In particular, we combine different levels of deletion and insertion and attempt to predict the missing interactions from the remaining topology information.

TABLE 1
Comparison of IGS, HRG, and MDS Methods on Terrorist Network

	Ins\Del	0.2	0.4	0.6	0.8
	0	0.8806	0.8386	0.7766	0.6473
	0.2	0.8141	0.7725	0.7061	0.6346
IGS	**0.4**	0.7814	0.7099	0.6725	0.5860
	0.6	0.7366	0.6602	0.6135	0.5832
	0.8	0.7021	0.6298	0.6001	0.5707
	1	0.6832	0.6147	0.5832	0.5638
	0	0.8139	0.8265	0.7517	0.6995
	0.2	0.6957	0.6932	0.6145	0.5764
HRG	**0.4**	0.7020	0.6204	0.6410	0.6143
	0.6	0.6646	0.6492	0.5973	0.4940
	0.8	0.5966	0.5696	0.5433	0.4749
	1	0.5614	0.6118	0.4900	0.5479
	0	0.7926	0.7529	0.7032	0.5806
	0.2	0.7655	0.7318	0.6693	0.5683
MDS	**0.4**	0.7066	0.6969	0.6445	0.5356
	0.6	0.6852	0.6672	0.6358	0.5089
	0.8	0.6622	0.6451	0.6210	0.4894
	1	0.6579	0.6374	0.6143	0.4733

Comparison of AUC value test of IGS, HRG, and MDS methods on Terrorist Network. For each method, AUC values are computed under different conditions of insertions or deletions. “Ins” stands for the ratio between the number of inserted false PPIs and the number of PPIs in the original network. “Del” denotes the ratio between the number of removed true PPIs and the number of PPIs in the original network.

TABLE 2
Comparison of IGS, HRG, and MDS Methods on Dolphin Network

	Ins \ Del	0.2	0.4	0.6	0.8
	0	0.8290	0.7703	0.7170	0.6031
	0.2	0.7541	0.7206	0.6510	0.5788
IGS	**0.4**	0.7072	0.6696	0.6396	0.5610
	0.6	0.6772	0.6214	0.5818	0.5425
	0.8	0.6325	0.6285	0.6103	0.5707
	1	0.6155	0.5858	0.5571	0.5402
	0	0.8953	0.8231	0.7478	0.6973
	0.2	0.6867	0.6365	0.6274	0.5004
HRG	**0.4**	0.6309	0.4427	0.5496	0.5815
	0.6	0.5873	0.5592	0.4921	0.5298
	0.8	0.5580	0.4878	0.5363	0.5346
	1	0.5797	0.5112	0.5143	0.5798
	0	0.7881	0.7641	0.6630	0.5503
	0.2	0.7522	0.6828	0.5900	0.5310
MDS	**0.4**	0.7210	0.6567	0.5678	0.5211
	0.6	0.6493	0.6237	0.5653	0.5177
	0.8	0.6048	0.5864	0.5536	0.5041
	1	0.5945	0.5640	0.5487	0.4952

Comparison of AUC value test of IGS, HRG, and MDS methods on Dolphin Network. For each individual method, AUC values are computed under different conditions of insertions or deletions. ‘Ins’ stands for the ratio between the number of inserted false PPIs and the number of PPIs in the original network. ‘Del’ denotes the ratio between the number of removed true PPIs and the number of PPIs in the original network.

The comparison results are shown in Tables 1 and 2 for terrorist network and Dolphin network respectively [45], [46]. In the Table 1, we can find the IGS performs best among three methods, followed by HRG. Both IGS and HRG reveal the new relationship among the nodes in the network based on the consensus metric which captures the topological features of entire or a large fraction of network. Therefore, they are more robust to both the missing and spurious PPIs. This robustness against the noise is confirmed in this test. We explain why MDS method performs worse against insertion and deletion noise. The metric revealed by MDS is based on the shortest path traveled from one protein to another protein in the network [1]. Before the deletion of the PPI between a pair of proteins, the shortest path is 1 as they are linked. However, the length of the shortest path increases if PPI no longer exists between them. Before the insertion of the PPI between a pair of proteins, the shortest path is larger than 1 as they are not directly linked. The path-length might be a very large number if two nodes are really far away. However, the shortest path would be changed to 1 if a link is introduced between them. The short-circuit noise is a very typical topological noise in computational geometry area which is usually overcome by the global geometric metric for example the graph Laplacian based representation. While IGS and HRG evaluate the probability of two node linking each other taking into account all existing connections, therefore, they are more robust to both missing PPI and spurious PPI according to the results in the table. The comparison results shown in Table 2 for Dolphin network indicated IGS and HRG have close performance in prediction, but still outperform the MDS method. However, based on our experimental test, HRG method is computationally expensive and fails to work on large size networks, like CS2007.

4 Discussion and Conclusion

The silico approaches for prediction of PPI in network have been receiving more and more attentions, however, are facing challenging because of the inherently spurious and missing PPIs presence in observed measurements. The geometric based approaches, which are only based on the topology of the PPI network, are very promising as those approaches are fully independent from other prior knowledge except for topology of the PPI network. We proposed a novel geometric description, intrinsic geometric structure for the protein-protein interaction network. IGS reflects the hidden and implicit relationship among proteins. We will discuss the properties of the intrinsic geometric structure as follows:

4.1 IGS is Learnt from Noisy Observation

As it is known in biological high-throughput data, one of the typical features is the noisy. The PPI network is usually observed with a high rate of missing and spurious interactions. An approach is likely to be overwhelmed by the high level of noises in the network if it simply relies on the individual local pairs of data or a small number of neighbors. However, two proteins are expected, though not interacted at individual pair level, to be statistically interacted given many paths bridging each other by considering all of topological connections in the network. And two proteins are expected, though connected at individual pair level, to be statistically non-interacted given few paths bridging each other by considering all of topological connections in the network. A good data analysis approach should be globally aware and robust to the perturbation of the local features. The new relationship reflected by our IGS is determined by taking into account of all local pieces of individual PPIs. The new relationship is based on the behavior of the PPI network system governed by agreement between all individual interactions. Therefore, the IGS is determined from a global structure as all of the available prior knowledge, which is very robust to noise.

4.2 IGS is Learnt from Limited Local Knowledge

PPI network are often observed in a incomplete manner as it is impractical to experimentally verify all individual PPIs within on cell or organism. f an approach is only able to capture the local neighbors for a PPI network, it is likely to fail to reveal the real relationship among proteins due to incompleteness. To complete a high fidelity PPI network, it is of great importance take into account all the pieces of local information simultaneously, in order to generate the knowledge behind the overall global structure of the data. IGS progressively collect the local information in a heat diffusion process to reach an optimal arrangement of all local piece of PPIs in a global consistent manner. Therefore, given a sufficient samples of true positive and true negatives, even they are not a complete description of network, IGS is able to reveal a high fidelity PPI network. The excellence of IGS against the incompleteness is highlighted by its good performance at a large number of insertions and deletions introduced.

4.3 Limitations and Future Directions

Although our approach demonstrates a great performance in the analysis of PPI network, we are aware of some limitations in our current implementation and experiments. Because we mainly focus on the development of computational model, the analysis of biological significance has not been emphasized in the paper. For example, the noise properties in raw PPI data can be different from the simulated random deletions and insertions used in existing experiments. The applicability and generality of IGS should be further explored in the future work by testing real challenging PPI network. For the tests on challenging PPI network, we will also compare our proposed method to a recently developed method based on generative network model [37]. We give a global optimal scale to describe the intrinsic geometric structure. However, the heat diffusion process is known to describe the structure in a multi-scale manner. Therefore, it is of interest to investigate the multi-scale properties of the PPI network and evaluate the stability and other properties at each different level. The multi-level description of PPI network will provide a grain to coarse insights and might reveal more details about the network. Furthermore, IGS is a general method and applicable to a wide range of problem domains, for example, the reconstruction of the air transportation network.

Acknowledgments

This work was supported by grants from the National Institute of Health to K.R. (GM075004) and New York University in Abu Dhabi (AD131). K.R. also acknowledges an appointment for the Donald W. Fedderson professorship in Mechanical Engineering. The authors would like to thank the Institute for Pure & Applied Mathematics for providing fellowship to support Karthik Ramani visit. They would also like to thank Amit Singer, Yoel Shkolnisky, Ronald Coifman, and Peter W. Jones for providing valuable insights on diffusion kernels. They especially thank Stephane Lafon for his discussion about the diffusion process. Y. Fang is the corresponding author.

References

[1] O. Kuchaiev, M. Rasajski, D. Higham, and N. Przulj, "Geometric de-noising of protein-protein interaction networks," *Plos Comput. Biol.*, vol. 5, 2009, http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000454

[2] T. Pawson, G. Gish, and P. Nash, "Sh2 domains, interaction modules and cellular wiring," *Trends Cell Biol.*, vol. 11, pp. 504–511, 2001.

[3] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins," *Proc. Nat. Acad. Sci. USA.*, vol. 97, pp. 1143–1147, 2000.

[4] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae," *Nature*, vol. 403, pp. 623–627, 2000.

[5] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg, "A protein interaction map of drosophila melanogaster," *Science*, vol. 302, pp. 1727–1736, 2003.

[6] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker, "A human protein-protein interaction network: A resource for annotating the proteome," *Cell*, vol. 122, pp. 957–968, 2005.

[7] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt, "Global landscape of protein complexes in the yeast saccharomyces cerevisiae," *Nature*, vol. 440, pp. 637–643, 2006.

[8] B. Wang, D.-S. Huang, and C. Jiang, "A new strategy for protein interface identification using manifold learning method," *IEEE Trans. NanoBiosci.*, vol. 13, no. 2, pp. 118–123, Jun. 2014.

[9] D. S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Current Protein Peptide Sci.*, vol. 15, no. 2, pp. 553–560, Jun. 2014.

[10] D. S. Huang and H.-J. Yu. (2013). Normalized feature vectors: A novel alignment-free sequence comparison method based on numbers of adjacent amino acids. *IEEE/ACM Trans. Comput. Biol. Bioinformat.* [Online]. *10(6)*, pp. 457–467. Available: http://bioinformatics.oxfordjournals.org/content/25/6/743.abstract

[11] K.-H. Liu and D.-S. Huang. (2008). Cancer classification using rotation forest. *Comput. Biol. Med.* [Online]. *38(5)*, pp. 601–610. Available: http://www.sciencedirect.com/science/article/pii/S0010482508000334

[12] M. Kumar, M. Bhasin, N. Natt, and G. B. Raghava. (2005). Prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM technique. *Nucleic Acids Res.* [Online]. *38(5)*, pp. W154–W159. Available: http://www.sciencedirect.com/science/article/pii/S0010482508000334

[13] J.-F. Xia, X.-M. Zhao, J. Song, and D.-S. Huang. (2010). Apis: Accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformat.* [Online]. *11(1)*, pp. 174. Available: http://www.biomedcentral.com/1471-2105/11/174

[14] B. A. Shoemaker and A. R. Panchenko, "Deciphering proteincprotein interactions. part i. experimental techniques and databases," *PLoS Comput. Biol.*, vol. 3, 2007, http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030042

[15] B. A. Shoemaker and A. R. Panchenko, "Deciphering proteincprotein interactions. part ii. computational methods to predict protein and domain interaction partners," *PLoS Comput. Biol.*, vol. 3, 2007, http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030043

[16] A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein, "Bridging structural biology and genomics: Assessing protein interaction data with known complexes," *Trends Genetics*, vol. 18, pp. 529–536, 2002.

[17] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, pp. 399–403, 2002.

[18] X. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, pp. 4394–4400, 2005.

[19] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, pp. 449–453, 2003.

[20] J. Wang, C. Li, E. Wang, and X. Wang, "Uncovering the rules for protein-protein interactions from yeast genomic data," *Proc. Nat. Acad. Sci. USA.*, vol. 106, pp. 3752–3757, 2009.

[21] D. Hwang, A. G. Rust, S. Ramsey, J. J. Smith, D. M. Leslie, A. D. Weston, P. de Atauri, J. D. Aitchison, L. Hood, A. F. Siegel, and H. Bolouri, "A data integration methodology for systems biology," *Proc. Nat. Acad. Sci. USA*, vol. 102, pp. 17296–17301, 2005.

[22] M. Koyutürk, "Algorithmic and analytical methods in network biology," *Wiley Interdisciplinary Rev.: Syst. Biol. Med.*, vol. 2, pp. 277–292, 2009.

[23] J. Li and Q. Liu. (2009). Double water exclusion: A hypothesis refining the O-ring theory for the hot spots at protein interfaces. *Bioinformatics* [Online]. *25(6)*, pp. 743–750. Available: http://bioinformatics.oxfordjournals.org/content/25/6/743.abstract

[24] J. Li and Q. Liu. (2009). Double water exclusion: A hypothesis refining the O-ring theory for the hot spots at protein interfaces. *Bioinformatics* [Online]. *25(6)*, pp. 743–750. Available: http://bioinformatics.oxfordjournals.org/content/25/6/743.abstract

[25] L. Zhu, Z.-H. You, D.-S. Huang, and B. Wang, "T-lse: A novel robust geometric approach for modeling protein-protein interaction networks," *PLoS ONE*, vol. 8, no. 4, p. e58368, 2013.

[26] Z.-H. You, Y.-K. Lei, J. Gui, D.-S. Huang, and X. Zhou. (2010). Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. *Bioinformatics* [Online]. *26(21)*, pp. 2744–2751. Available: http://bioinformatics.oxfordjournals.org/content/26/21/2744.abstract

[27] S. Martin, D. Roe, and J.-L. Faulon. (2005). Predicting Proteinprotein interactions using signature products. *Bioinformatics* [Online]. *21(2)*, pp. 218–226. Available: http://bioinformatics.oxfordjournals.org/content/21/2/218.abstract

[28] D.-S. Huang and C.-H. Zheng. (2006). Independent component Analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* [Online]. *22(15)*, pp. 1855–1862. Available: http://bioinformatics.oxfordjournals.org/content/22/15/1855.abstract

[29] B. Wang, P. Chen, D.-S. Huang, J. jing Li, T.-M. Lok, and M. R. Lyu. (2006). Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett.* [Online]. *580 (2)*, pp. 380–384. Available: http://www.sciencedirect.com/science/article/pii/S0014579305014705

[30] N. Przulj, D. Corneil, and I. Jurisica, "Modeling interactome: Scale-free or geometric?" *Bioinformatics*, vol. 20, pp. 3508–3515, 2004.

[31] H. Yu, A. Paccanaro, V. Trifonov, and M. Gerstein, "Predicting interactions in protein networks by completing defective cliques," *Bioinformatics*, vol. 22, no. 7, pp. 823–829, 2006.

[32] S. Lafon and A. Lee, "Diffusion maps and Coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1393–1403, Sep. 2006.

[33] R. Saito, H. Suzuki, and Y. Hayashizaki. (2002). Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res.* [Online]. *30(5)*, pp. 1163–1168. Available: http://nar.oxfordjournals.org/content/30/5/1163.abstract

[34] R. Saito, H. Suzuki, and Y. Hayashizaki. (2003). Construction of reliable proteinprotein interaction networks with a new interaction generality measure. *Bioinformatics* [Online]. *19(6)*, pp. 756–763. Available: http://bioinformatics.oxfordjournals.org/content/19/6/756.abstract

[35] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq. (2003). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.* [Online]. *5(1)*, p. R6. Available: http://genomebiology.com/2003/5/1/R6

[36] H. N. Chua, W.-K. Sung, and L. Wong. (2006). Exploiting indirect neighbours and topological weight to predict protein function from Proteinprotein interactions. *Bioinformatics* [Online]. *22(13)*, pp. 1623–1630. Available: http://bioinformatics.oxfordjournals.org/content/22/13/1623.abstract

[37] Y. Zhu, X.-F. Zhang, D.-Q. Dai, and M.-Y. Wu, "Identifying spurious interactions and predicting missing interactions in the protein-protein interaction networks via a generative network model," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 10, no. 1, pp. 219–225, Jan. 2013.

[38] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98–101, 2008.

[39] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, pp. 799–805, 2004.

[40] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr., and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Nat. Acad. Sci. USA*, vol. 97, pp. 262–267, 2000.

[41] F. Lu, S. Kele, S. J. Wright, and G. Wahba, "Framework for kernel regularization with application to protein clustering," *Proc. Nat. Acad. Sci. USA*, vol. 102, pp. 12332–12337, 2005.

[42] A. Vaxman, M. Ben-Chen, and C. Gotsman, "A multi-resolution approach to heat kernels on discrete surfaces," *ACM Trans. Graph.*, vol. 29, pp. 121:1–121:10, Jul. 2010.

[43] F. Chung, *Spectral Graph Theory*. Providence, RI, USA: Amer. Math. Soc., 1997.

[44] S. R. Collins, P. Kemmeren, X. C. Zhao, J. F. Greenblatt, F. Spencer, F. C. Holstege, J. S. Weissman, and N. J. Krogan, "Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae," *Mol. Cell Proteomics*, vol. 6, pp. 439–450, 2007.

[45] V. Krebs, "Mapping networks of terrorist cells," *Connections*, vol. 24, pp. 43–52, 2001.

[46] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. can geographic isolation explain this unique trait?" *Behavioral Ecol. Sociobiol.*, vol. 54, pp. 396–405, 2003.

Yi Fang is an assistant professor of electrical and computer engineering at New York University, Abu Dhabi. He is currently working on the development of state-of-the-art techniques in large-scale visual computing, deep visual learning, deep cross-domain and cross-modality model, and their applications in engineering, social science, medicine, and biology. He is a member of the IEEE.

Mengtian Sun received the bachelor's degree in mechanical engineering from the Harbin Institute of Technology in 2008, and the master's degree in electrical and computer engineering from Purdue University in 2012. He is a software development engineer at Apple Inc. His research interests include computational geometry, computer vision, and machine learning.

Guoxian Dai received the BS degree from Xi'an Jiaotong University in 2011 and the master's degree from Fudan University in 2014. He is currently working toward the PhD degree in the Computer Science and Engineering Department, NYU School of Engineering. He is a student member of the IEEE.

Karthik Ramani is a professor in the School of Mechanical Engineering, Purdue University. His research lies at the intersection of mechanical engineering and information science and technology, the areas span design and manufacturing, new kernels for shape understanding using machine learning, geometric computing, and human-computer natural user interaction, and interfaces with shapes and sketches.