nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05469-x OPEN

Network enhancement as a general method to denoise weighted biological networks

Bo Wang[1], Armin Pourshafeie[2], Marinka Zitnik[1], Junjie Zhu[3], Carlos D. Bustamante[4,5], Serafim Batzoglou[1,6] & Jure Leskovec[1,5]

Networks are ubiquitous in biology where they encode connectivity patterns at all scales of organization, from molecular to the biome. However, biological networks are noisy due to the limitations of measurement technology and inherent natural variation, which can hamper discovery of network patterns and dynamics. We propose Network Enhancement (NE), a method for improving the signal-to-noise ratio of undirected, weighted networks. NE uses a doubly stochastic matrix operator that induces sparsity and provides a closed-form solution that increases spectral eigengap of the input network. As a result, NE removes weak edges, enhances real connections, and leads to better downstream performance. Experiments show that NE improves gene–function prediction by denoising tissue-specific interaction networks, alleviates interpretation of noisy Hi-C contact maps from the human genome, and boosts fine-grained identification accuracy of species. Our results indicate that NE is widely applicable for denoising biological networks.

[1] Department of Computer Science, Stanford University, 353 Serra Mall, Stanford 94305 CA, USA. [2] Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford 94305 CA, USA. [3] Department of Electrical Engineering, Stanford University, 350 Serra Mall, Stanford 94305 CA, USA. [4] Department of Biomedical Data Science, Stanford University, 1265 Welch Road, Stanford 94305 CA, USA. [5] Chan Zuckerberg Biohub, 499 Illinois St, San Francisco 94158 CA, USA. [6] Present address: Illumina Inc, 499 Illinois Street, San Francisco 94158 CA, USA. These authors contributed equally: Bo Wang, Armin Pourshafeie, Marinka Zitnik. Correspondence and requests for materials should be addressed to S.B. (email: serafim@cs.stanford.edu) or to J.L. (email: jure@cs.stanford.edu)

Networks provide an elegant abstraction for expressing fine-grained connectivity and dynamics of interactions in complex biological systems[1]. In this representation, the nodes indicate the components of the system. These nodes are often connected by non-negative, (weighted-)edges, which indicate the similarity between two components. For example, in protein–protein interaction (PPI) networks, weighted edges capture the strength of physical interactions between proteins and can be leveraged to detect functional modules[2]. However, accurate experimental quantification of interaction strength is challenging[3,4]. Technical and biological noise can lead to superficially strong edges, implying spurious interactions; conversely, dubiously weak edges can hide real, biologically important connections[4–6]. Furthermore, corruption of experimentally derived networks by noise can alter the entire structure of the network by modifying the strength of edges within and amongst underlying biological pathways. These modifications adversely impact the performance of downstream analysis[7]. The challenge of noisy interaction measurements is not unique to PPI networks and plagues many different types of biological networks, such as Hi-C[8] and cell–cell interaction networks[9].

To overcome this challenge, computational approaches have been proposed for denoising networks. These methods operate by replacing the original edge weights with weights obtained based on a diffusion defined on the network[10,11]. However, these methods are often not tested on different types of networks[11], rely on heuristics without providing explanations for why these approaches work, and lack mathematical understanding of the properties of the denoised networks[10,11]. Consequently, these methods may not be effective on new applications derived from emerging experimental biotechnology.

Here, we introduce network enhancement (NE), a diffusion-based algorithm for network denoising that does not require supervision or prior knowledge. NE takes as input a noisy, undirected, weighted network, and outputs a network on the same set of nodes but with a new set of edge weights (Fig. 1). The main crux of NE is the observation that nodes connected through paths with high-weight edges are more likely to have a direct, high-weight edge between them[12,13]. Following this intuition, we define a diffusion process that uses random walks of length three or less and a form of regularized information flow to denoise the input network (Fig. 1a and Methods). Intuitively, this diffusion generates a network in which nodes with strong similarity/interactions are connected by high-weight edges while nodes with weak similarity/interactions are connected by low-weight edges (Fig. 1b). Mathematically, this means that eigenvectors associated with the input network are preserved while the eigengap is increased. In particular, NE denoises the input by down-weighting small eigenvalues more aggressively than large eigenvalues. This re-weighting is advantageous when the noise is spread in the eigen-directions corresponding to small eigenvalues[14]. Furthermore, the increased eigengap of the enhanced network is a highly appealing property as it leads to accurate detection of modules/clusters[15,16] and allows for higher-order

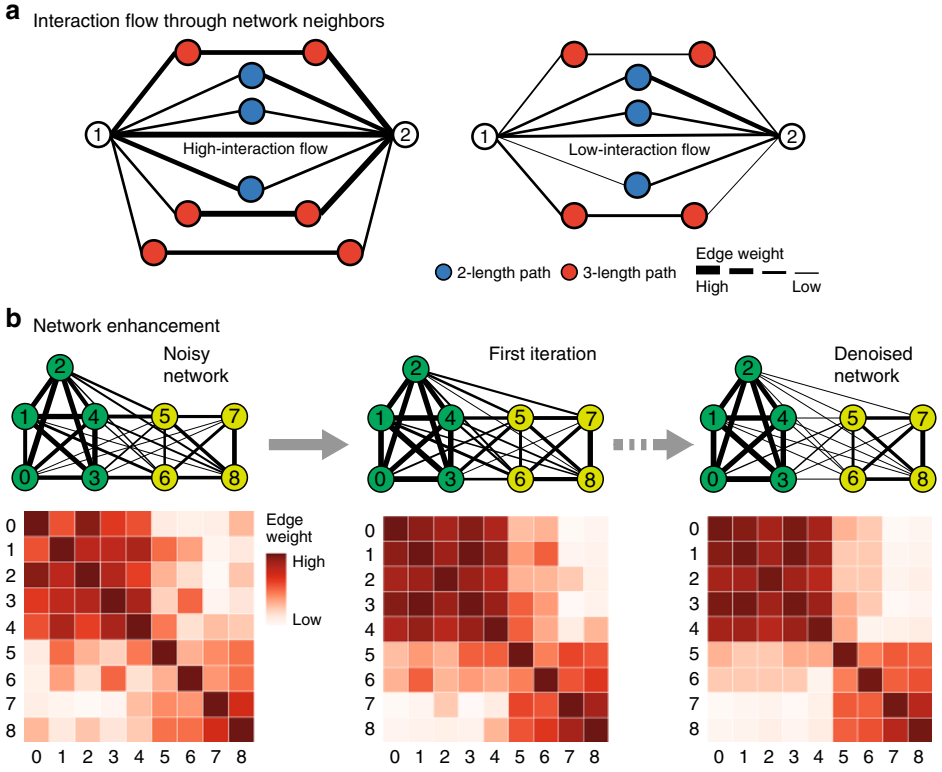

Fig. 1 Overview of Network Enhancement (NE). **a** NE employs higher-order network structures to enhance a given weighted biological network. The diffusion process in NE revises edge weights in the network based on interaction flow between any two nodes. Specifically, for any two nodes, NE updates the weight of their edge by considering all paths of length three or less connecting those nodes. **b** The iterative process of NE. NE takes as input a weighted network and the associated adjacency matrix (visualized as a heatmap). It then iteratively updates the network using the NE diffusion process, which is guaranteed to converge. The diffusion defined by NE improves the input network by strengthening edges that are either close to other strong edges in the network according to NE's diffusion distance or are supported by many weak edges. On the other hand, NE weakens edges that are not supported by many strong edges. Upon convergence, the enhanced network is a symmetric, doubly stochastic matrix (DSM) (Supplementary Note 3). This makes the enhanced network well-suited for downstream computational analysis. Furthermore, enforcement of the DSM structure leads to more sparse networks with lower noise levels

network analysis[12]. Moreover, NE has an efficient and easy to implement closed-form solution for the diffusion process, and provides mathematical guarantees for this converged solution. (Fig. 1b and Methods).

Results

Methods for network denoising. We have applied NE to three challenging yet important problems in network biology. In each experiment, we evaluate the network denoised by NE against the same network denoised by alternative methods: network deconvolution (ND)[10] and diffusion state distance (DSD)[11]. For completeness, we also compare our results to a network reconstructed from features learned by Mashup (MU)[17]. All three of these methods use a diffusion process as a fundamental step in their algorithms and have a closed-form solution at convergence. ND solves an inverse diffusion process to remove the transitive edges, and DSD uses a diffusion-based distance to transform the network. While ND and DSD are denoising algorithms, MU is a feature learning algorithm that learns low-dimensional representations for nodes based on their steady-state topological positions in the network. This representation can be used as input to any subsequent prediction model. In particular, a denoised network can be constructed by computing a similarity measure using MU's output features[17].

NE improves human tissue networks for gene–function prediction. Networks play a critical role in capturing molecular aspects of precision medicine, particularly those related to gene–function and functional implications of gene mutation[18,19]. We test the utility of our denoising algorithm in improving gene interaction networks from 22 human tissues assembled by Greene et al.[20]. These networks capture gene interactions that are specific to human tissues and cell lineages ranging from B lymphocyte to skeletal muscle and the whole brain[20,21]. We predict the cellular functions of genes specialized in different tissues based on the networks obtained from different denoising algorithms.

Given a tissue and the associated tissue-specific gene interaction network, we first denoise the network and then use a network-based algorithm on the denoised edge weights to predict gene functions in that tissue. We use standard weighted random walks with restarts to propagate gene–function associations from training nodes to the rest of the network[22]. We define a weighted random walk starting from nodes representing known genes associated with a given function. At each time step, the walk moves from the current node to a neighboring node selected with a probability that depends on the edge weights and has a small probability of returning to the initial nodes[22]. The algorithm scores each gene according to its visitation probability by the random walk. Node scores returned by the algorithm are then used to predict gene–function associations for genes in the test set. Predictions are evaluated against experimentally validated gene–function associations using a leave-one-out cross-validation strategy.

When averaged over the four denoising algorithms and the 22 human tissues, the gene–function prediction improved by 12.0% after denoising. Furthermore, we observed that all denoising algorithms improved the average prediction performance (Fig. 2a and Supplementary Note 1). These findings motivate the use of denoised networks over original (raw) biological networks for downstream predictive analytics. We further observed that gene–function prediction performed consistently better in combination with networks revised by NE than in combination with networks revised by other algorithms. On average, NE outperformed networks reconstructed by ND, DSD, and MU by 12.3%. In particular, NE resulted in an average of a 5.1% performance gain over the second best-performing denoised network (constructed by MU). Following Greene et al.[20], we further validated our NE approach by examining each enhanced tissue network in turn and evaluating how well relevant tissue-specific gene functions are connected in the network. The expectation is that function-associated genes tend to interact more frequently in tissues in which the function is active than in other non-relevant tissues[20]. As a result, relevant functions are expected to be more tightly connected in the tissue network than functions specific to other tissues. For each NE-enhanced tissue network, we ranked all functions by the edge density of function-associated tissue subnetworks and examined top-ranked functions. In the NE-enhanced blood plasma network, we found that functions with the highest edge density were blood coagulation, fibrin clot formation, and negative regulation of very-low-density lipoprotein particle remodeling, all these functions are specific to blood plasma tissue (Fig. 2b). This finding suggests that tissue subnetworks associated with relevant functions tend to be more connected in the tissue network than subnetworks of non-tissue-specific functions. The most connected functions in the NE-enhanced brain network were brain morphogenesis and forebrain regionalization, which are both specific to brain tissue (Fig. 2b). Examining edge density-based rankings of gene functions across 22 tissue networks, we found relevant functions consistently placed at or near the top of the rankings, further indicating that NE can improve the signal-to-noise ratio of tissue networks.

NE improves Hi-C networks for domain identification. The recent discovery of numerous cis-regulatory elements away from their target genes emphasizes the deep impact of 3D structure of DNA on cell regulation and reproduction[23–25]. Chromosome conformation capture (3C)-based technologies[25] provide experimental approaches for understanding the chromatin interactions within DNA. Hi-C is a 3C-based technology that allows measurement of pairwise chromatin interaction frequencies within a cell population[8,25]. The Hi-C reads are grouped into bins based on the genetic region they map to. The bin size determines the measurement resolution.

Hi-C read data can be thought of as a network where genomic regions are nodes and the normalized count of reads mapping to two regions are the weighted edges. Network community detection algorithms can be used on this Hi-C derived network to identify clusters of regions that are close in 3D genomic structure[26]. The detected megabase-scale communities correspond to regions known as topological-associating domains (TADs) and represent chromatin interaction neighborhoods[26]. TADs tend to be enriched for regulatory features[27,28] and are hypothesized to specify elementary regulatory microenvironments. Therefore, detection of these domains can be important for analysis and interpretation of Hi-C data. The limited number of Hi-C reads, hierarchical structure of TADs and other technological challenges lead to noisy Hi-C networks, and hamper accurate detection of TADs[25].

To investigate the ability of NE in improving TAD detection, we apply NE to a Hi-C dataset and analyze the performance of a standard domain identification pipeline with and without a network denoising step. For this experiment, we used 1 and 5 kb resolution Hi-C data from all autosomes of the GM12878 cell line[8]. Since true gold-standards for TAD regions are lacking, a synthetic dataset was created by stitching together non-overlapping clusters detected in the original work[8]. As a result, the clusters stitched together can be used as a good proxy for the true clusters (more details in Supplementary Note 1). Figure 3a shows a heatmap of the raw Hi-C data for a portion of chromosome 16.

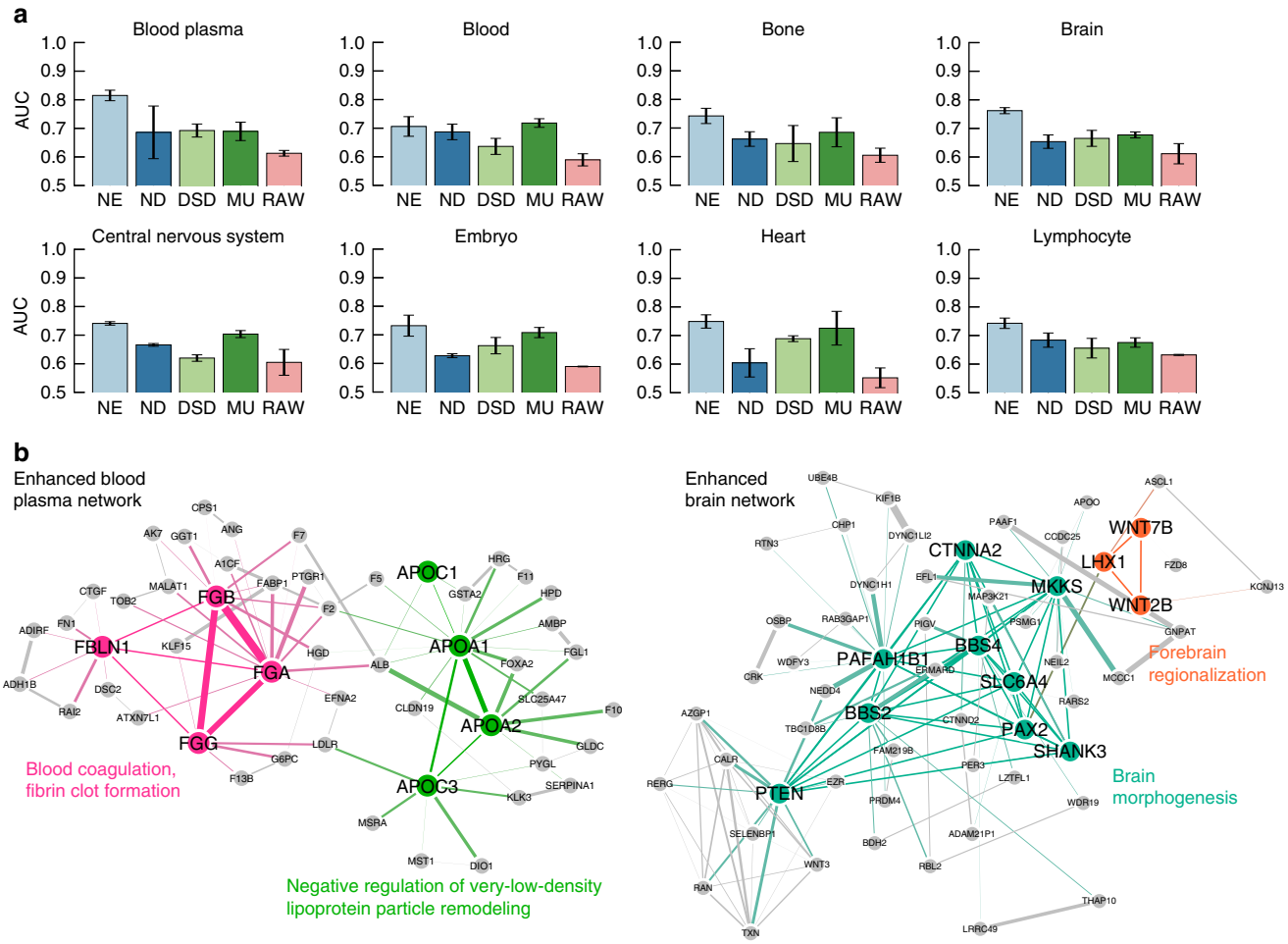

Fig. 2 Gene-function prediction using tissue-specific gene interaction networks. **a** We assessed the utility of original networks (RAW) and networks denoised using MU, ND, DSD, and NE for tissue-specific gene-function prediction. Each bar indicates the performance of a network-based approach that was applied to a raw or denoised gene interaction network in a particular tissue and then used to predict gene functions in that tissue. Prediction performance is measured using the area under receiver operating characteristic curve (AUROC), where a high AUROC value indicates the approach learned from the network to rank an actual association between a gene and a tissue-specific function higher than a random gene, tissue-specific function pair. Error bars indicate performance variation across tissue-specific gene functions. Results are shown for eight human tissues, additional fourteen tissues are considered in Supplementary Figs. 1, 2. **b** For blood plasma and brain tissues, we show gene interaction subnetworks centered on two blood plasma gene functions and two brain gene functions with the highest edge density in NE-denoised data. Edge density for each gene function (with n associated genes) was calculated as the sum of edge weights in the NE-denoised network divided by the total number of possible edges between genes associated with that function ($n \times (n-1)/2$). The most interconnected gene functions in each tissue (shown in color, names of associated genes are emphasized), are also relevant to that tissue

We applied two, off-the-shelf, community detection methods (Louvian[29] and MSCD[30]) to each Hi-C network and compared the quality of the detected TADs with or without network denoising. Visual inspection of the Hi-C contact matrix before and after the Hi-C network is denoised using NE reveals an enhancement of edges within each community and sharper boundaries between communities (Fig. 3a). This improvement is particularly clear for the 5 kb resolution data, where communities that were visually undetectable in the raw data become clear after denoising with NE. To quantify this enhancement, the communities obtained from raw networks and networks enhanced by NE or other denoising methods were compared to the true cluster assignments. We used normalized mutual information (NMI, Supplementary Note 2) as a measure of shared information between the detected communities and the true clusters. NMI ranges between 0 to 1, where a higher value indicates higher concordance and 1 indicates an exact match between the detected communities and the true clusters. The results across 22 autosomes indicate that while denoising can improve the detection of communities, not all denoising algorithms succeed in this task (Fig. 3b). For both resolutions considered, NE performs the best with an average NMI of 0.92 for 1 kb resolution and 0.94 for 5 kb resolution, MU (the second best-performing method) achieves an average NMI of 0.85 and 0.84, respectively, while ND and DSD achieve lower average NMI than the raw data which has NMI of 0.81 and 0.67, respectively. Furthermore, we note that the performance of NE and MU remains high as the resolution decreases from 1 to 5 kb, in contrast the ability of the other pipelines in detecting the correct communities diminishes. While MU maintains a good average performance at 5 kb resolution, the standard deviation of NMI values after denoising with MU increases from 0.037 in 1 kb data to 0.054 in 5 kb data due to relatively poor performances on a few chromosomes. On the other hand, the NMI values for data denoised with NE maintain a similar spread at both resolutions

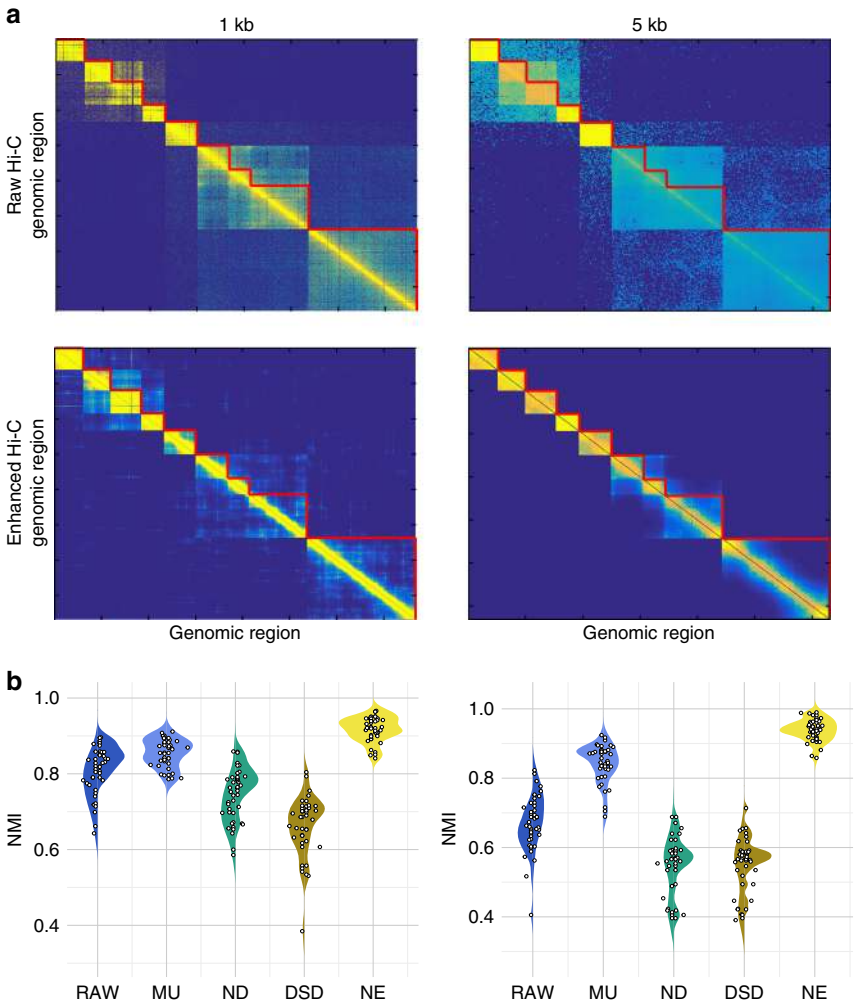

Fig. 3 Domain identification in Hi-C genomic interaction networks. **a** Heatmap of Hi-C contact matrix for a portion of chromosome 16. For 1 kb resolution data denoised with NE and clustered using Louvain community detection (Supplementary Note 1) chromosome 16 (visualized) has the median normalized mutual information (NMI) and was chosen as a fair representation of the overall performance. The top two heatmaps show the contact matrices for original (raw) data and the bottom heatmaps represent the contact matrices for data after application of NE. The images on the left correspond to data with 1 kb resolution (i.e., the bin-size is a 1 kb region) and the right images correspond to the same section at 5 kb resolution. The red lines indicate the borders for each domain as detailed in Supplementary Note 1. In each case, the network is consisted of genomic windows of length 1 kb (left) or 5 kb (right) as nodes, and normalized number of reads mapped to each region as the edge weights. The data was truncated for visualization purposes. **b** NMI for clusters detected. For each algorithm, the left side of the violin plot corresponds to Louvain community detection algorithm and the right side corresponds to MSCD algorithm. Each dot indicates the performance on a single autosome (the distance of the dots from the central vertical axis is dictated by a random jitter for visualization purposes). While for raw data and data preprocessed with DSD and ND the overall NMI decreases as resolution decreases, for NE and MU the performance remains high. MU maintains good overall performance with lower resolution, however, the spread of the NMI increases indicating that the consistency of performance has decreased compared to NE where the spread remains the same

(standard deviation 0.033 and 0.031, respectively). The better average NMI and smaller spread indicates that NE can reliably enhance the network and improve TAD detection.

NE improves fine-grained species identification. Fine-grained species identification from images concerns querying objects within the same subordinate category. Traditional image retrieval works on high-level categories (e.g., finding all butterflies instead of cats in a database given a query of a butterfly), while fine-grained image retrieval aims to distinguish categories with subtle differences (e.g., monarch butterfly versus peacock butterfly). One major obstacle in fine-grained species identification is the high similarity between subordinate categories. On one hand, two subordinate categories share similar shapes and carry subtle color difference in a small region; on the other hand, two subordinate categories of close colors can only be well separated by texture. Furthermore, viewpoint, scale variation, and occlusions among objects all contribute to the difficulties in this task[31]. Due to these challenges, similarity networks, which represent pairwise affinity between images, can be very noisy and ineffective in retrieval of a sample from the correct species for any query.

We test our method on the Leeds butterfly fine-grained species image dataset[32]. Leeds Butterfly dataset contains 832 butterflies in 10 different classes with each class containing between 55 and 100 images[32]. We use two different common encoding methods (Fisher Vector (FV) and Vector of Linearly Aggregated Descriptors (VLAD) with dense SIFT; Supplementary Note 1) to generate two different vectorizations of each image. These two encoding methods describe the content of the images differently and therefore capture different information about the images. Each method can generate a similarity network in which nodes

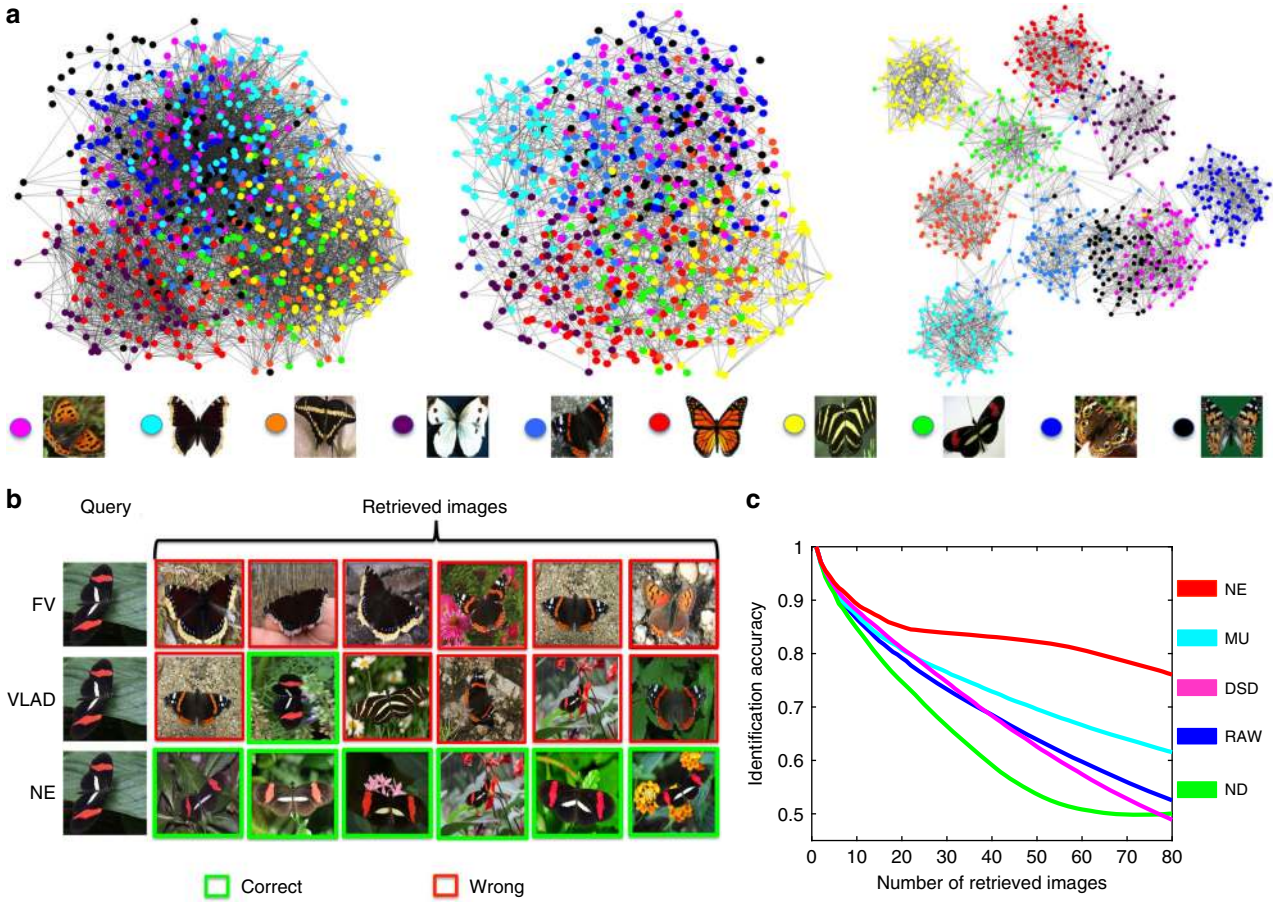

Fig. 4 Network-based butterfly species identification. (Best seen in color.) **a** Example showing that combining different image vectorizations into a similarity network, followed by a denoising of the similarity network can improve retrieval performance. Visualization of encoded butterfly images in the form of a butterfly similarity network. From left to right: Fisher Vector encoding method, VLAD encoding method (Supplementary Note 1), and the denoised similarity network by our method (NE). The legend shows an example photograph[32] of each butterfly species included in the network. **b** Retrieval by each encoding method. Given a query butterfly, raw image vectorizations fail to correctly retrieve other butterflies from the same class (i.e., same species) while the network denoised by NE correctly recovers similarities between the query and its neighbors within the same class. All photographs are taken with permission from the Leeds Butterfly image dataset[32]. **c** Species identification accuracy when varying the number of retrieved images. A detailed comparison with other methods. Each curve shows the identification accuracy (Supplementary Note 2) as a function of number of retrievals for one method

represent images and edge weights indicate similarity between pairs of images. The inner product of these two similarity networks is used as a single input network to a network denoising algorithm.

Visual inspection indicates that NE is able to greatly improve the overall similarity network for fine-grain identification (Fig. 4a). While both encodings partially separate the species, before applying NE, all the images are tangled together without a clear clustering. On the other hand, the resulting similarity network after applying NE clearly shows 10 clusters corresponding to different butterfly species (Fig. 4a). More specifically, given a query, the original input networks fail to capture the true affinities between the query butterfly and its most similar retrievals, while NE is able to correct the affinities and more reliably output the correct retrievals (Fig. 4b).

To quantify the improvements due to NE in the task of species identification, we use identification accuracy, a standard metric which quantifies the average numbers of correct retrievals given any query of interests (Supplementary Note 2). A detailed comparison between NE and other alternatives by examining identification accuracy of the final network with respect to different number of top retrievals demonstrates NE's ability in improving the original noisy networks (Fig. 4b). For example, when considering top 40 retrievals, NE can improve the raw network by 18.6% (more than 10% better than other alternatives). Further, NE generates the most significant improvement in performance (41% over the raw network and more than 25% over the second best alternative), when examining the top 80 retrieved images.

Current denoising methods suffer from high sensitivity to the hyper-parameters when constructing the input similarity networks, e.g., the variance used in Gaussian kernel (Supplementary Note 1). However, our model is more robust to the choice of hyper-parameters (Supplementary Fig. 3). This robustness is due to the strict structure enforced by the preservation of symmetry and DSM structure during the diffusion process (see Supplementary Note 3).

Discussion

We proposed NE as a general method to denoise-weighted undirected networks. NE implements a dynamic diffusion process that uses both local and global network structures to construct a denoised network from its noisy version. The core of our

approach is a symmetric, positive semi-definite, doubly stochastic matrix, which is a theoretically justified replacement for the commonly used row-normalized transition matrix[33]. We showed that NE's diffusion model preserves the eigenvectors and increases the eigengap of this matrix for large eigenvalues. This finding provides insight into the mechanism of NE's diffusion and explains its ability to improve network quality[15,16]. In addition to increasing the eigengap, NE disproportionately trims small eigenvalues. This property can be contrasted with the principal component analysis (PCA) where the eigenspectrum is truncated at a particular threshold. Through extensive experimentation, we show that NE can flexibly fit into important network analytic pipelines in biology and that its theoretical properties enable substantial improvements in the performance of downstream network analyses.

We see many opportunities to improve upon the foundational concept of NE in future work. First, in some cases, a small subset of high confidence nodes may be available. For example, genomic regions in the Hi-C contact maps can be augmented using data obtained from 3C technology or a small number of species can be identified by a domain expert and used together with network data as input to a denoising methodology. Extending NE to take advantage of the small amount of accurately labeled data might further extend our ability to denoise networks. Second, although we showed the utility of NE for denoising several types of weighted networks, there are other network types worth exploring, such as multimodal networks involving multiomic measurements of cancer patients. Finally, incorporating NE's diffusion process into other network analytic pipelines can potentially improve performance. For example, MU[17] learns vector representations for nodes based on a steady state of a traditional random walk with restart, and replacing MU's diffusion process with the rescaled steady state of NE might be a promising future direction.

Methods

Problem definition and doubly stochastic matrix property. Let $G = (E, V, W)$ be a weighted network where V denotes the set of nodes in the network (with $|V| = n$), E represents the edges of G, and W contains the weights on the edges. The goal of NE is to generate a network $G^* = (E^*, V, W^*)$ that provides a better representation of the underlying module membership than the original network G. For the analysis below, we let W represent a symmetric, non-negative matrix.

Diffusion-based models often rely on the row-normalized transition probability matrix $P = D^{-1}W$, where D is a diagonal matrix whose entries are $D_{i,i} = \sum_{j=1}^{n} W_{i,j}$. However, transition probability matrix P defined in this way is generally asymmetric and does not induce a directly usable node-node similarity metric. Additionally, most diffusion-based models lack spectral analysis of the denoised model. To construct our diffusion process and provide a theoretical analysis of our model, we propose to use a symmetric, doubly stochastic matrix (DSM). Given a matrix $M \in \mathbb{R}^{n\times n}$, M is said to be DSM if:

1. $M_{i,j} \geq 0 \quad i, j \in \{1, 2, \ldots, n\}$,
2. $\sum_i M_{i,j} = \sum_j M_{i,j} = 1$.

The second condition above is equivalent to $\mathbf{1} = (1, 1, \ldots, 1)^T$ and $\mathbf{1}^T$ being a right and left eigenvector of M with eigenvalue 1. In fact, 1 is the greatest eigenvalue for all DSM matrices (see the remark following the definition of DSM in the Supplementary Notes). Overall, the DSM property imposes a strict constraint on the scale of the node similarities and provides a scale-free matrix that is well-suited for subsequent analyses.

Network enhancement. Given a matrix of edge weights W representing the pairwise weights between all the nodes, we construct another localized network $\mathcal{T} \in \mathbb{R}^{n\times n}$ on the same set of nodes to capture local structures of the input network. Denote the set of K-nearest neighbors (KNN) of the i-th node (including the node i) as $\mathcal{N}_i$. We use these nearest neighbors to measure local affinity. Then the corresponding localized network $\mathcal{T}$ can be constructed from the original weighted network using the following two steps:

$$P_{i,j} \leftarrow \frac{W_{i,j}}{\sum_{k\in\mathcal{N}_i} W_{i,k}} \mathbb{I}_{\{j\in\mathcal{N}_i\}}, \quad \mathcal{T}_{i,j} \leftarrow \sum_{k=1}^{n} \frac{P_{i,k}P_{j,k}}{\sum_{v=1}^{n} P_{v,k}}, \tag{1}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. We can verify that $\mathcal{T}$ is a symmetric DSM by directly checking the conditions of the definition (Supplementary Note 3). $\mathcal{T}$ encodes the local structures of the original network with the intuition that local neighbors (highly similar pairs of nodes) are more reliable than remote ones, and local structures can be propagated to non-local nodes through a diffusion process on the network. Motivated by the updates introduced in Zhou et al.[34], we define our diffusion process using $\mathcal{T}$ as follows:

$$W_{t+1} = \alpha\mathcal{T} \times W_t \times \mathcal{T} + (1-\alpha)\mathcal{T} \tag{2}$$

where α is a regularization parameter and t represents the iteration step. The value of W_0 can be initialized to be the input matrix W. Equation (2) shows that diffusion process in NE is defined by random walks of length three or less and a form of regularized information flow. There are three main reasons for restricting the influence of random walks to at most third-order neighbors in the network: (1) for most nodes third-order neighborhood spans the extent of almost the entire biological network, making neighborhoods of order beyond three not very informative of individual nodes[35,36], (2) currently there is little information about the extent of influence of a node (i.e., a biological entity, such as gene) on the activity (e.g., expression level) of its neighbor that is more than three hops away[37], and (3) recent studies have empirically demonstrated that network features extracted based on three-hop neighborhoods contain the most useful information for predictive modeling[38,39].

To further explore Eq. (2) we can write the update rule for each entry:

$$(W_{t+1})_{i,j} = \alpha \sum_{k\in\mathcal{N}_i} \sum_{l\in\mathcal{N}_j} \mathcal{T}_{i,k}(W_t)_{k,l}\mathcal{T}_{l,j} + (1-\alpha)\mathcal{T}_{i,j}. \tag{3}$$

It can be seen from Eq. (3) that the updated network comes from similarity/interaction flow only through the neighbors of each data point. The parameter α adds strengths to self-similarities, i.e., a node is always most similar to itself. One key property that differentiates our method from typical diffusion methods is that in the proposed diffusion process defined in Eq. (2), for each iteration t, W_t remains a symmetric DSM. Furthermore, W_t converges to a non-trivial equilibrium network which is a symmetric DSM as well (Supplementary Note 3). Therefore, NE constructs an undirected network that preserves the symmetry and DSM property of the original network. Through extensive experimentation we show that NE improves the similarity between related nodes and the performance of downstream methods such as community detection algorithms.

The main theoretical insight into the operation of NE is that the proposed diffusion process does not change eigenvectors of the initial DSM while mapping eigenvalues via a non-linear function (Supplementary Note 3). Let eigen-pair $(\lambda_0, \boldsymbol{v}_0)$ denote the eigen-pair of the initial symmetric DSM, $\mathcal{T}_0$. Then, the diffusion process defined in Eq. (2) does not change the eigenvectors, and the final converged graph has eigen-pair $(f_\alpha(\lambda_0), \boldsymbol{v}_0)$, where $f_\alpha(x) = \frac{(1-\alpha)x}{1-\alpha x^2}$. This property shows that, the diffusion process using a symmetric, DSM is a non-linear operator on the spectrum of the eigenvalues of the original network. This results has a number of consequences. Practically, it provides us with a closed-form expression for the converged network. Theoretically, it hints at how this diffusion process effects the eigenspectrum and improves the network for subsequent analyses. (1) If the original eigenvalue is either 0 or 1, the diffusion process preserves this eigenvalue. This implies that, like other diffusion processes, NE does not connect disconnected components. (2) NE increases the gap between large eigenvalues of the original network and reduces the gap between small eigenvalues of this matrix. Larger eigengap is associated with better network community detection and higher-order network analysis[12,15,16]. (3) The diffusion process always decreases the eigenvalues, which follows from: $(1-\alpha)\lambda_0/\left(1-\alpha\lambda_0^2\right) \leq \lambda_0$, where smaller eigenvalues get reduced at a higher rate. This observation can be interpreted in relation to PCA where the eigenspectrum below a user determined threshold value is ignored. PCA has many attractive theoretical properties, especially for dimensionality reduction. In fact, MU[17], a feature learning method whose output is also a denoised version of the original network, can be fit by computing the PCA decomposition on the stationary state of the network. MU aims to learn a low-dimensional representation of nodes in the network which makes PCA a natural choice. However, a smoothed-out version of the PCA is more attractive for network denoising because denoising is typically used as a preprocessing step for downstream prediction tasks, and thus robustness to selection of a threshold value for the eigenspectrum is desirable.

These findings shed light on why the proposed algorithm (NE) enhances the robustness of the diffused network compared to the input network (Supplementary Note 3). In some contexts, we may need the output to remain a network of the same scale as the input network. This requirement can be satisfied by first recording the degree matrix of the input network and eventually mapping the denoised output of the algorithm back to the original scale by a symmetric matrix multiplication. We summarize our NE algorithm along with this optional degree-mapping step in Supplementary Note 3.

Code availability. The project website can be found at: http://snap.stanford.edu/ne. Source code of the NE method is available for download from the project website.

Data availability. All relevant data are public and available from the authors of the original publications. The project website can be found at: http://snap.stanford.edu/ne. The website contains preprocessed data used in the paper together with raw and enhanced networks.

Received: 20 October 2017 Accepted: 3 July 2018
Published online: 06 August 2018

References

1. Gao, J., Barzel, B. & Barabási, A.-L. Universal resilience patterns in complex networks. *Nature* **530**, 307–312 (2016).
2. Zhong, Q. et al. An inter-species protein–protein interaction network across vast evolutionary distance. *Mol. Syst. Biol.* **12**, 865 (2016).
3. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
4. Costanzo, M. et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).
5. Ji, J., Zhang, A., Liu, C., Quan, X. & Liu, Z. Survey: functional module detection from protein-protein interaction networks. *IEEE Trans. Knowl. Data Eng.* **26**, 261–277 (2014).
6. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
7. Menche, J. et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).
8. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
9. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133 (2015).
10. Feizi, S., Marbach, D., Médard, M. & Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* **31**, 726–733 (2013).
11. Cao, M. et al. Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS ONE* **8**, e76339 (2013).
12. Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353**, 163–166 (2016).
13. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
14. Rosipal, R. & Trejo, L. J. Kernel partial least squares regression in reproducing kernel hilbert space. *J. Mach. Learn. Res.* **2**, 97–123 (2001).
15. Spielman, D. A. Spectral graph theory and its applications. In *48th Annual IEEE Symposium on Foundations of Computer Science* 29–38 (IEEE, Providence, RI, USA, 2007).
16. Verma, D. & Meila, M. Comparison of spectral clustering methods. *Adv. Neural Inf. Process. Syst.* **15**, 38 (2003).
17. Cho, H., Berger, B. & Peng, J. Compact integration of multi-network topology for functional analysis of genes. *Cell Syst.* **3**, 540–548 (2016).
18. Radivojac, P. et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).
19. Zitnik, M. & Zupan, B. Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. *Pac. Symp. Biocomput.* **19**, 400–411 (2014).
20. Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
21. Zitnik, M. & Leskovec, J. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* **33**, 190–198 (2017).
22. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Human. Genet.* **82**, 949–958 (2008).
23. Bickmore, W. A. & van Steensel, B. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–1284 (2013).
24. De Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
25. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **17**, 743 (2016).
26. Cabreros, I., Abbe, E. & Tsirigos, A. Detecting community structures in Hi-C genomic data. In *Annual Conference on Information Science and Systems* 584–589 (IEEE, NJ, USA, 2016).
27. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
28. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
29. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.Theory Exp.* **10**, 10008 (2008).
30. Le Martelot, E. & Hankin, C. Fast multi-scale community detection based on local criteria within a multi-threaded algorithm. Preprint at https://arxiv.org/abs/1301.0955 (2013).
31. Gavves, E., Fernando, B., Snoek, C. G., Smeulders, A. W., and Tuytelaars, T. Fine-grained categorization by alignments. In *2013 IEEE International Conference on Computer Vision* 1713–1720 (IEEE Computer Society, Washington, DC, 2013).
32. Wang, J., Markert, K. & Everingham, M. Learning models for object recognition from natural language descriptions. In *Proc. British Machine Vision Conference* 1–11 (British Machine Vision Association, London, 2009).
33. Wang, B., Jiang, J., Wang, W., Zhou, Z.-H. & Tu, Z. Unsupervised metric fusion by cross diffusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* 2997–3004 (IEEE, Rhode Island, USA, 2012).
34. Zhou, D., Bousquet, O., Lal, T. N., Weston, J. & Schölkopf, B. Learning with local and global consistency. In *Advances in Neural Information Processing Systems. Proc. of the First 12 Conferences* (eds Jordan, M. I., LeCun, Y. & Solla, S. A.) 321-328 (Max Planck Institute for Biological Cybernetics, Tuebingen, Germany, 2001).
35. Pržulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics* **23**, e177–e183 (2007).
36. Davis, D., Yaveroğlu, Ö. N., Malod-Dognin, N., Stojmirovic, A. & Pržulj, N. Topology-function conservation in protein–protein interaction networks. *Bioinformatics* **31**, 1632–1639 (2015).
37. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551 (2017).
38. Goldenberg, A., Mostafavi, S., Quon, G., Boutros, P. C. & Morris, Q. D. Unsupervised detection of genes of influence in lung cancer using biological networks. *Bioinformatics* **27**, 3166–3172 (2011).
39. Mostafavi, S., Goldenberg, A., Morris, Q. & Ravasi, T. Labeling nodes using three degrees of propagation. *PLoS ONE* **7**, e51947 (2012).

Acknowledgements

M.Z. and J.L. were supported by NSF, NIH BD2K, DARPA SIMPLEX, Stanford Data Science Initiative, and Chan Zuckerberg Biohub. J.Z. was supported by the Stanford Graduate Fellowship, NSF DMS 1712800 Grant and the Stanford Discovery Innovation Fund. A.P. and C.D.B. were supported by National Institutes of Health/National Human Genome Research Institute T32 HG-000044, Chan Zuckerberg Initiative and Grant Number U01FD004979 from the FDA, which supports the UCSF-Stanford Center of Excellence in Regulatory Sciences and Innovation.

Author contributions

B.W., A.P., M.Z., and J.Z. designed and performed research, contributed new analytic tools, analyzed data, and wrote the manuscript. C.D.B., S.B., and J.L. designed and supervised the research and contributed to the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05469-x.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.