Comp790-166: Computational Biology

Lecture 22

April 4, 2022

Good Morning Question

1. What should an ideal joint embedding of multiple modalities satisfy?
2. What is a biological application for insights that can be gleaned using the joint embedding? For example, with TCGA data.

Today

- Representing nodes with multiple relational definitions
- MASHUP + Protein Function Prediction

Intermission for Announcements

- Next homework to be assigned $\sim$ April 6. Now is a good time to work on projects!
- How are projects going?

From Last Time. Optimizing a Single Quadratic Form

Maximizing and minimizing quadratic forms

given $n \times n$ symmetric matrix A

$$\begin{aligned} \text{maximize :} \quad & x^\mathsf{T} A x \\ \text{subject to :} \quad & \|x\| = 1 \end{aligned}$$

eigenvalue decomposition of A:

$$\sum_{i=1}^{n} \lambda_i q_i q_i^\mathsf{T}$$

- $\lambda_1 \geq \cdots \geq \lambda_n$ eigenvalues of A
- $q_1, \ldots, q_n \in \mathbb{R}^n$ orthonormal eigenvectors
- solution: $x = q_1$
- optimal value: λ_1

to minimize:

- solution: $x = q_n$
- optimal value: λ_n

Figure: from `http://ee263.stanford.edu/lectures/extremal_trace.pdf`

Optimizing Multiple Quadratic Forms

Maximizing and minimizing sums of quadratic forms

$$\begin{array}{rl} \text{maximize :} & \sum_{i=1}^{k} x_i^T A x_i \\ \text{subject to :} & \|x_i\| = 1 \\ & x_i^T x_j = 0 \quad i \neq j \end{array}$$

compact representation:

$$\begin{array}{rl} \text{maximize :} & \mathbf{Tr}(X^T A X) \\ \text{subject to :} & X^T X = I \end{array}$$

- solution: $x_1 = q_1, \ldots, x_k = q_k$
- optimal value: $\lambda_1 + \cdots + \lambda_k$

to minimize:

- solution: $x_1 = q_n, \ldots, x_k = q_{n-k+1}$
- optimal value: $\lambda_n + \cdots + \lambda_{n-k+1}$

Figure: from `http://ee263.stanford.edu/lectures/extremal_trace.pdf`

Integrating Heterogeneous Information Sources

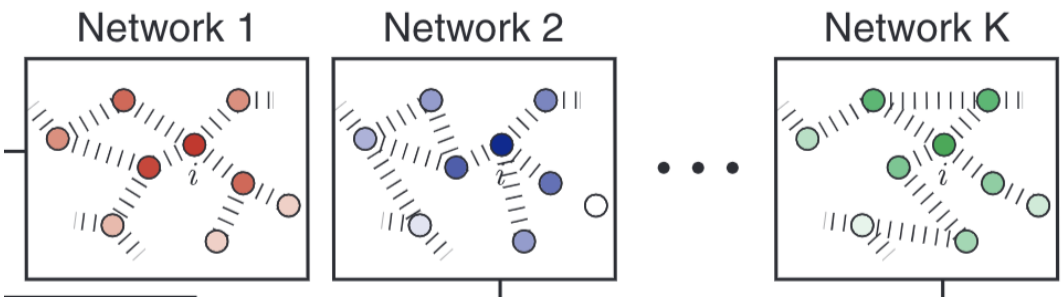

Figure: from Cho *et al.* Cell Systems. Each graph is representing a different relational definition between features.

Considering proteins, there are multiple methods for predicting whether these proteins interact .

- Physical binding
- gene expression
- co-localization
- experimentally determined
- text mined, etc.

We Seek a Unified Representations of these Nodes

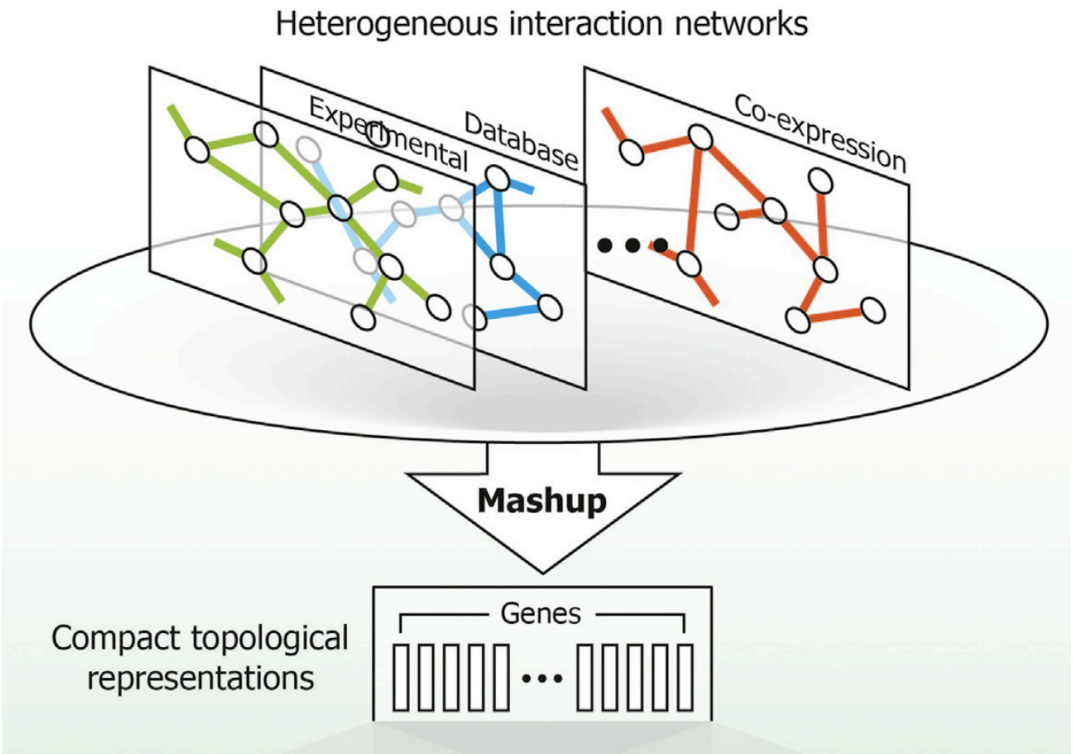

Figure: from Cho *et al.* Cell Systems. 2016.

Example from STRING

Using the STRING database, you can extract PPIs according to multiple relational definitions.

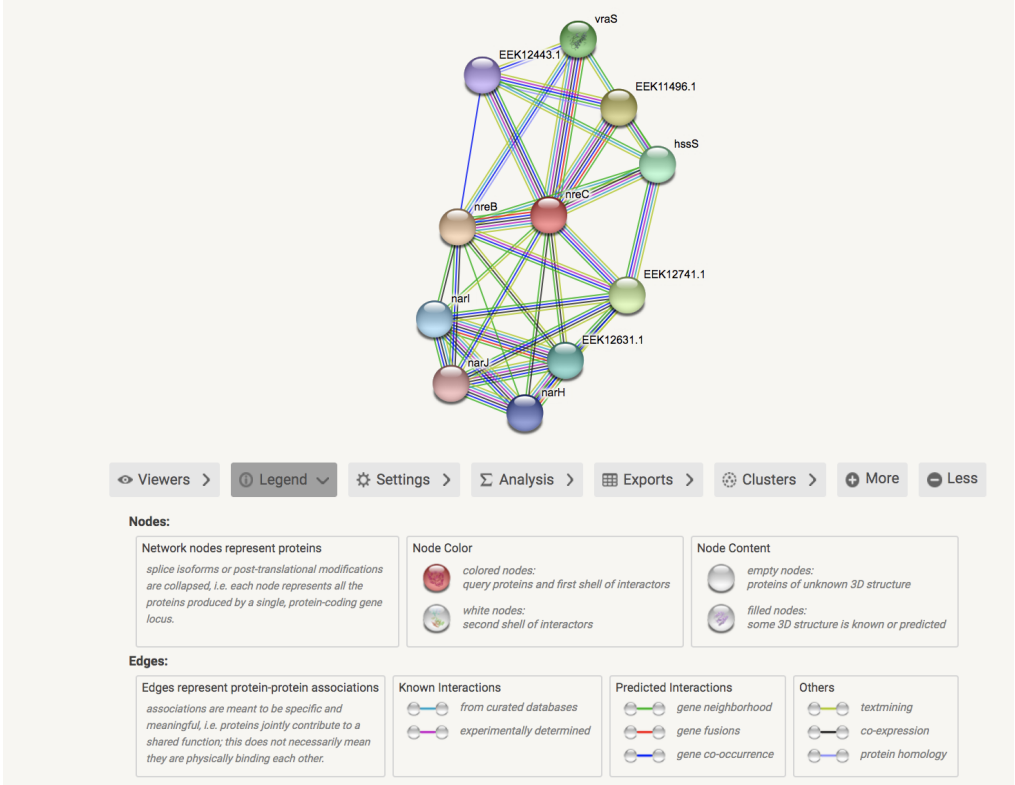

Figure: `https://string-db.org/`

Welcome Mashup

Given multiple relational definitions (e.g. multiple graphs) between a common set of nodes (features), define a consensus d-dimensional embedding vector that aligns well with each individual graph.

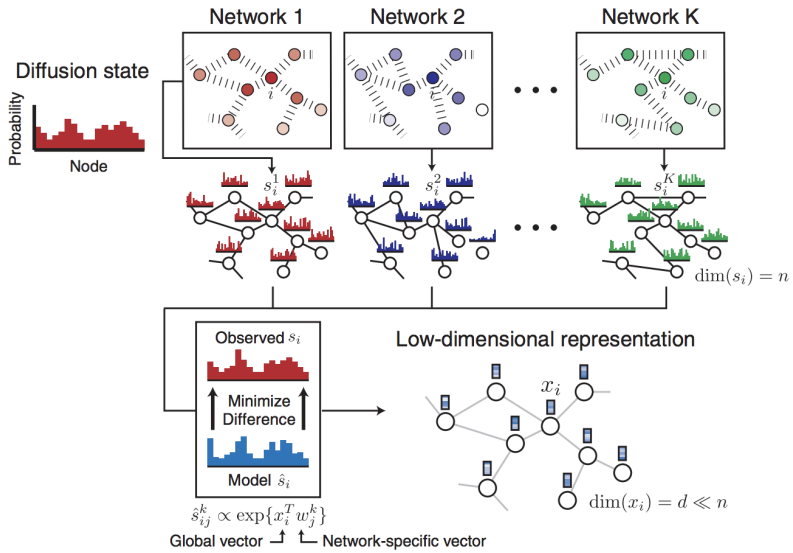

Figure: from Cho *et al.* Cell Systems. Each graph is representing a different relational definition between node (features).

Random Walk with Restart

- RWR is a way to account for both local and global ‘walk‘ information in the graph by giving your walker the chance to restart

But first, let's re-define the transition probability that a walker goes from node j to node i as,

$$B_{ij} = \frac{A_{ij}}{\sum_{i'} A_{i'j}}$$

RWR Formally Written

Given the transition matrix, B, the RWR from a node i is defined as,

$$s_i^{t+1} = (1 - p_r)Bs_i^t + p_r e_i$$

- p_r is the probability of restart
- $e_i(i) = 1$ and $e_i(j) = 0$ for $j \neq i$
- s_i^t is the vector of probabilities of each node being visited after t steps in the random walk, starting from node i

Clarifying What is Happening Here

$$s_i^{t+1} = (1 - p_r)Bs_j^t + p_r e_i$$

- The first term corresponds to following a random edge connected to the current node
- The second term corresponds to restarting from node i.
- At some point, this reaches a stationary distribution, s_i^{∞}, or fixed point
- When the diffusion states between two nodes are close, this implies they have similar positions in the graph with respect to other nodes.

Quantifying Topological Overlap Between a Node Pair

Each node is given two vector representations, $\mathbf{w}_i, \mathbf{x}_i \in \mathbb{R}^d$

- Let $\mathbf{w}_i$ refer to the context feature of a node (e.g. per relational definition)
- Let $\mathbf{x}_i$ refer to the node feature of node i (e.g. overall)

Define a new similarity measure between nodes i and j as,

$$\hat{s_{ij}} = \frac{\exp\{x_i^T w_j\}}{\sum_{j'} \exp\{x_i^T w_{j'}\}}$$

Unpacking

$$\hat{s_{ij}} = \frac{\exp\{x_i^T w_j\}}{\sum_{j'} \exp\{x_i^T w_{j'}\}}$$

- If $\mathbf{x}_i$ and $\mathbf{w}_j$ are close in direction and hence have a large inner product, then node j should be frequently visited in the random walk starting from node i.

Recap of what is happening

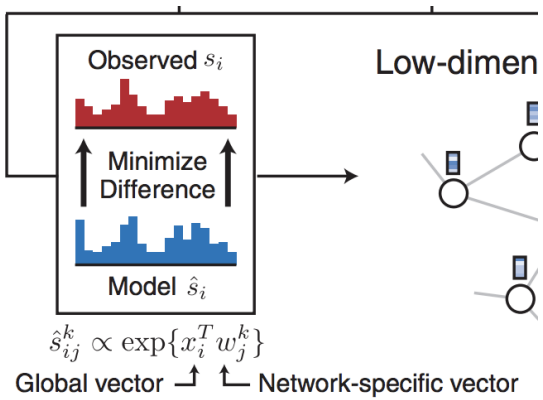

Figure: from Fig. 1. Given observed diffusion states from RWR, we should be able to find a global vector (**x**) and view-specific vector (**w**), such that a function of **x** and **w** gives a good diffusion state approximation.

Writing Out an Objective Functions for Embeddings

To find optimal d-dimensional representations for each node, formulate an optimization problem that minimizes the notation between s (RWR diffusion state) and $\hat{s}$ ('approximation') as,

$$\underset{w,x}{\text{minimize}} C(s, \widehat{s}) = \frac{1}{n} \sum_{i=1}^{n} D_{KL}\left(s_i \| \widehat{s_i}\right)$$

Written out, given our definition of $\hat{s}$ gives the following (with $H(\cdot)$ denoting entropy),

$$C(s, \widehat{s}) = \frac{1}{n} \sum_{i=1}^{n} \left[-H\left(s_i\right) - \sum_{j=1}^{n} s_{ij} \left(x_i^T w_j - \log \left(\sum_{j'=1}^{n} \exp\left\{ x_i^T w_{j'} \right\} \right) \right) \right]$$

Integrating Heterogeneous Networks

You can do these RWRs on each individual network. At the same time, you can let x be fixed across all relational definitions.
Similar to what we have seen, yet adapted for modality k, we can write,

$$\widehat{s}_{ij}^{k} := \frac{\exp\left\{x_i^T w_j^k\right\}}{\sum_{j'} \exp\left\{x_i^T w_{j'}^k\right\}}$$

Writing the Objective Function Across All Modalities

Now, the objective function can be rewritten to take into account the recently-computed $\widehat{s}_{ij}^{k}s$, and sums over all modalities as,

$$\underset{w,x}{\text{minimize}} C(s,\widehat{s}) = \frac{1}{n}\sum_{k=1}^{k}\sum_{i=1}^{n} D_{KL}\left(s_i^k \| \hat{s}_i^k\right)$$

Implementation (the slow way)

To find the optimal ws and xs for each node, you could compute gradients, which turn out to be,

$$\nabla_{w_i^k} C(s, \widehat{s}) = \frac{1}{n} \sum_{j=1}^{n} \left(\widehat{s}_{ji}^k - s_{ji}^k \right) x_j$$

and

$$\nabla_{x_i} C(s, \widehat{s}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \left(\widehat{s}_{ij}^k - s_{ij}^k \right) w_j^k$$

An SVD Formulation: Setup

- Let S^k be the $N \times N$ diffusion state matrix for network k
- Also, let s_i^k be the ith column of this matrix, S^k

This matrices can be concatenated to form an $nK \times n$ matrix, S

Remember Truncated SVD?

The authors used truncated SVD as an alternative to estimating the w_i^ks and x_is as,

$$S = U\Sigma V$$

(Remember this implies that we will have some 0s on the diagonal (e.g. zeroed out singular values) of Σ)

- $\{w_i^k\} \rightarrow \Sigma^{1/2} U^T \rightarrow (d \times d) \times (d \times (NK \times N))$
- $\{x_i\} \rightarrow \Sigma^{1/2} V \rightarrow (d \times d) \times (d \times N)$

Using the Learned $\mathbf{x}_i$s as Feature Vectors

- After Mashup each node, i has an embedding, $\mathbf{x}_i$.
- Each protein has a known function, which we can try to predict.

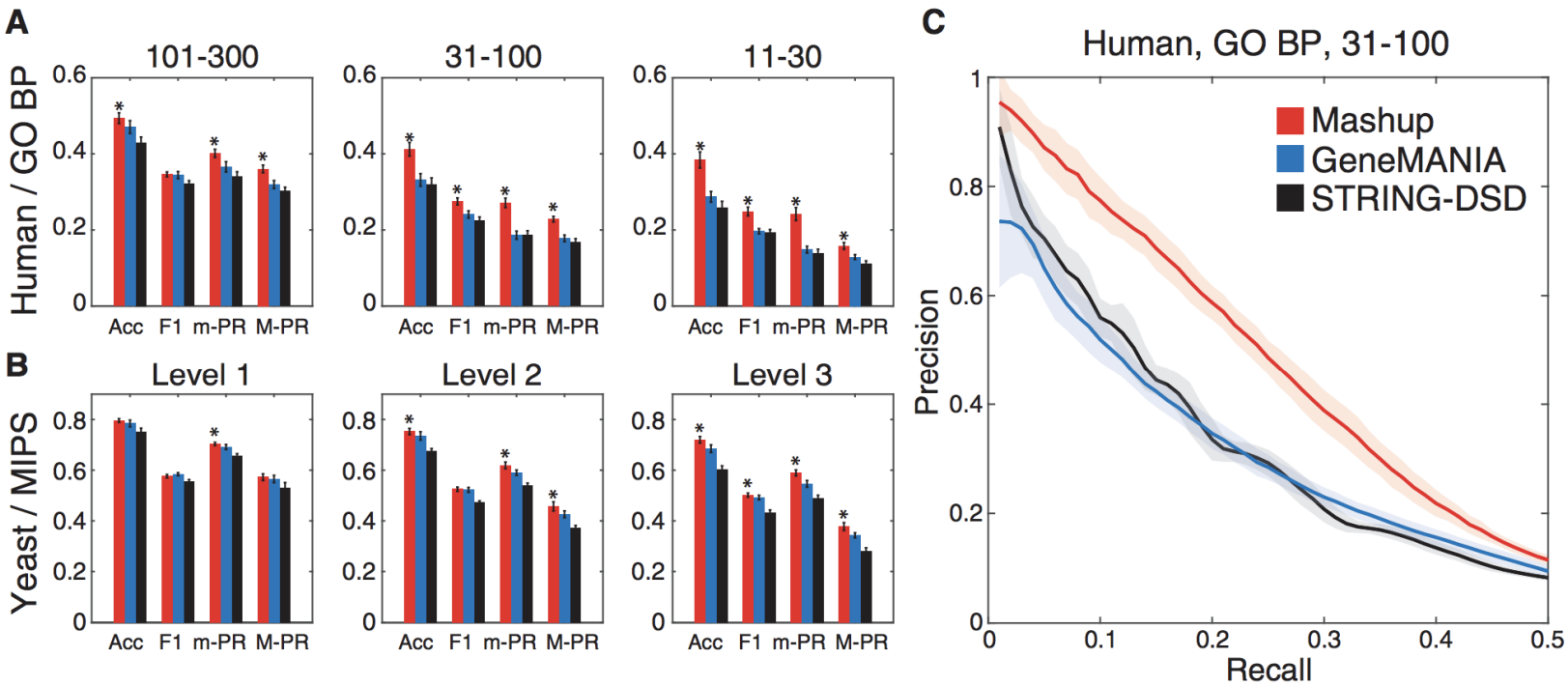

Figure: from Fig. 2. Performance is evaluated for multiple levels of annotation.

Similarly, Combining All Networks Leads to Better Protein Function Prediction

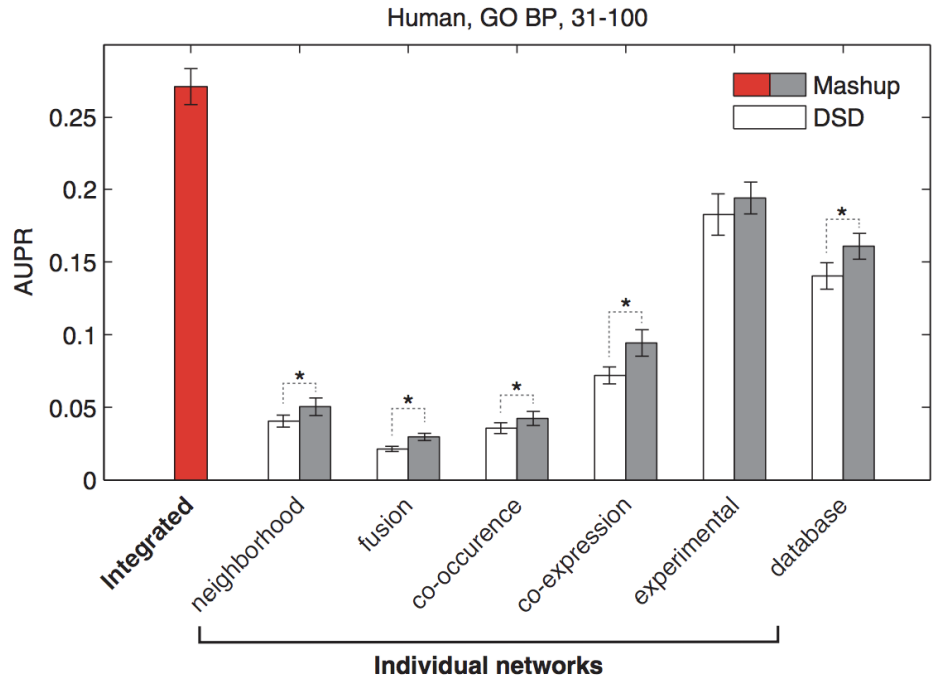

Figure: from Fig. 3. It's very reassuring to see that experimental is also a top performer!

Intuition about Parameters

The main parameters of interest is the restart probability, and the number of dimensions to keep.

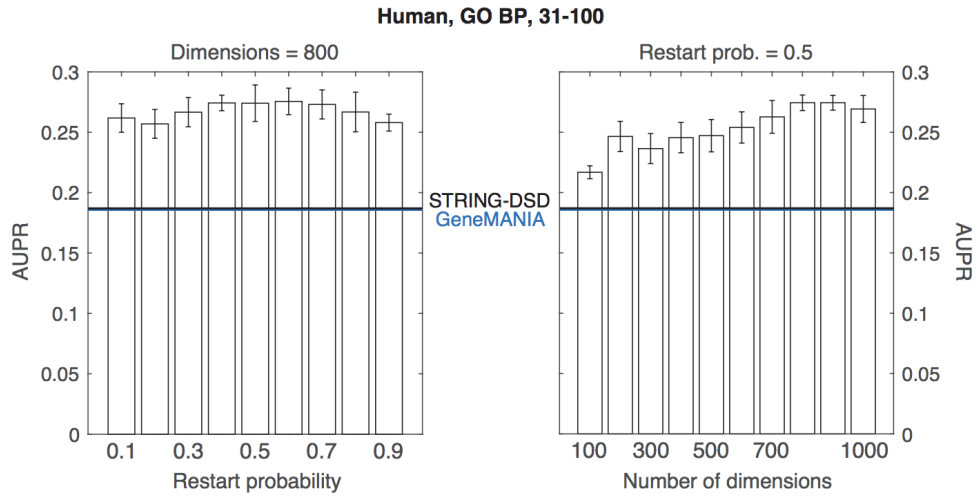

Figure: from Supp Fig. 4

*Btw, I recommend choosing your y-axis so that it is useful when making such a plot.

Mashup is More Robust to Noise in the Network

Here, noise was simulated by removing a subset of edges from the original graph.

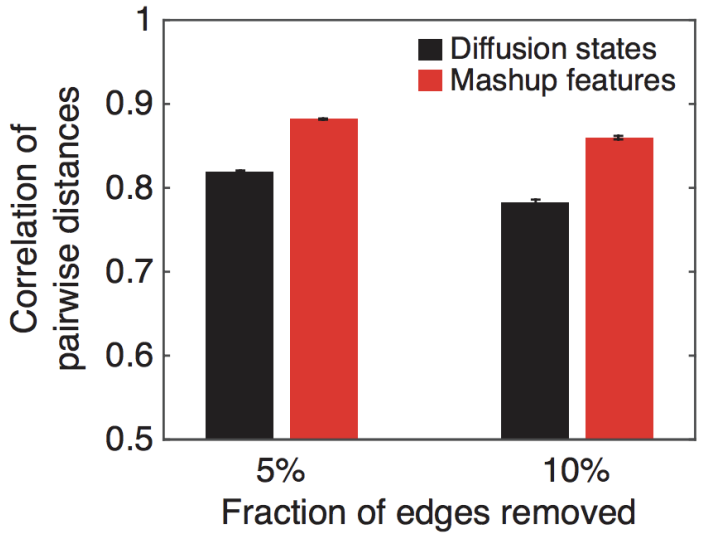

Figure: from Supp. Fig. 7. Edges were removed from the BioGrid physical interaction network. Similarities between nodes could be calculated based on diffusion state or mashup.

Since we are on the topic of multiple networks...

I'm so glad you asked. Let's talk about a related problem of graph alignment.