Provenance Tracing in Network Diffusion Algorithms

Nure Tasnina
Dept. of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA
E-mail: tasnina@vt.edu

Mark Crovella
Dept. of Computer Science, Boston University, Boston, MA 02215, USA

Simon Kasif
Dept. of Biomedical Engineering, Dept. of Computer Science, Graduate Program in Bioinformatics, Boston University, Boston, MA 02215, USA

T. M. Murali
Dept. of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA
E-mail: murali@cs.vt.edu

We propose a novel strategy for provenance tracing in random walk-based network diffusion algorithms, a problem that has been surprisingly overlooked in spite of the widespread use of diffusion algorithms in biological applications. Our path-based approach enables ranking paths by the magnitude of their contribution to each node's score, offering insight into how information propagates through a network. Building on this capability, we introduce two quantitative measures: (i) *path-based effective diffusion*, which evaluates how well a diffusion algorithm leverages the full topology of a network, and (ii) *diffusion betweenness*, which quantifies a node's importance in propagating scores. We applied our framework to SARS-CoV-2 protein interactors and human PPI networks. Provenance tracing of the Regularized Laplacian and Random Walk with Restart algorithms revealed that a substantial amount of a node's score is contributed via multi-edge paths, demonstrating that diffusion algorithms exploit the non-local structure of the network. Analysis of diffusion betweenness identified proteins playing a critical role in score propagation; proteins with high diffusion betweenness are enriched with essential human genes and interactors of other viruses, supporting the biological interpretability of the metric. Finally, in a signaling network composed of causal interactions between human proteins, the top contributing paths showed strong overlap with COVID-19-related pathways. These results suggest that our path-based framework offers valuable insight into diffusion algorithms and can serve as a powerful tool for interpreting diffusion scores in a biologically meaningful context, complementing existing module- or node-centric approaches in systems biology. The code is publicly available at `https://github.com/n-tasnina/provenance-tracing.git` under the GNU General Public License v3.0.

Keywords: protein-protein interaction, network diffusion, molecular mechanism

1. Introduction

Protein-protein interaction networks[1,2] have emerged as invaluable resources for studying the complex workings of living cells.[3,4] Within such a network, each node represents a protein, while each edge corresponds to an interaction between two proteins, e.g., physical or genetic. Diffusion algorithms are widely used to analyze such networks, enabling a variety of applications such as protein function prediction,[5,6] disease module discovery,[7] disease characterization,[8,9] and drug target prediction.[10] In recent years, they have also driven momentum in *in silico* drug repositioning[11,12]. The principle underlying these algorithms is that proteins with similar functions tend to interact with each other.[13,14] Typical network diffusion algorithms assign specific scores or labels to a set of experimentally determined nodes or seed nodes, e.g., proteins known to perform a particular function, and iteratively propagate this information through the network's edges to nearby nodes. This process continues for a fixed number of steps or until convergence. The final score of each node, which reflects its proximity to seed nodes, can be used as a surrogate for the node's association with the function of interest.

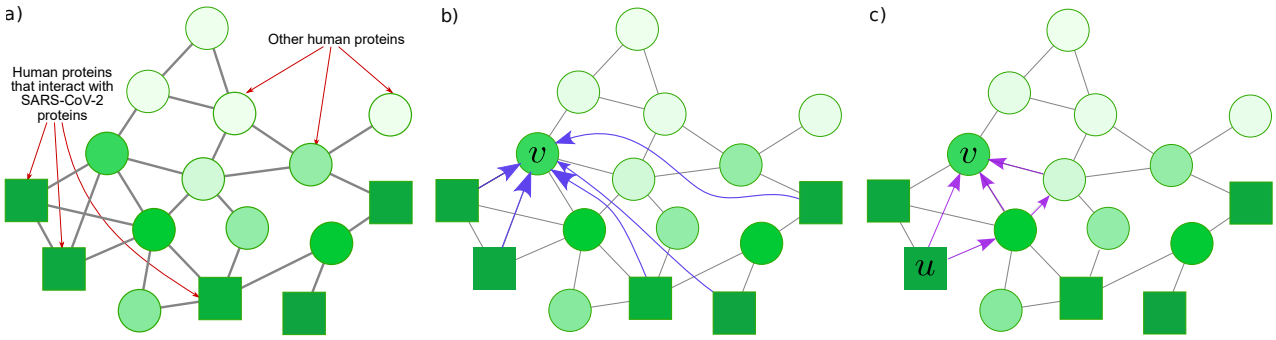

Fig. 1. **Illustration of provenance tracing**. a) A pictorial depiction of diffusion of score in a human protein-protein interaction network where nodes are the proteins and edges represent any interaction between proteins. Here, circular nodes depict all human proteins except those with experimental evidence of interaction with SARS-CoV-2 proteins. The SARS-CoV-2 protein interactors (seed nodes) are represented by square-shaped nodes. A network diffusion algorithm propagates scores from the seed nodes to all other human proteins in the network. The green hue of the nodes corresponds to the computed likelihood (score) of interaction with SARS-CoV-2, with a higher intensity indicating a greater likelihood. b) Node based algorithms of provenance tracing allow computing the contribution from each of the five seed nodes to the final score of the node v. c) Path-based algorithms of provenance tracing compute contributions coming along individual paths from seed nodes to other nodes, e.g., u to node v.

The last 10 years of research in AI has emphasized the importance of interpretability, explainability, and transparency of predictions independent of the specific prediction methodology used.[15] This is particularly important in scientific or medical application of network diffusion. In this context, "provenance" is defined as a trace from a predicted activity associated with a node (e.g., protein) to the factual (experimentally determined) evidence associated with seed nodes that were most informative in computing the predicted score.[16] Provenance tracing is the process of determining the origin of a predicted score back to the seed nodes.

While network diffusion algorithms have been deployed for more than 20 years, the question of provenance tracing is understudied.[16] Law *et al.*[17] explored a node-oriented method for provenance tracing in the important context of random walk based diffusion algorithms. For every node v, Law *et al.* computed how much each seed node contributed to the score of v. They defined "node-based effective diffusion" (NED) as the fraction of a node's score contributed by seed nodes that are not direct neighbors. Using SARS-CoV-2 interactors of humans as seed nodes,[18] they found that for each top-ranking prediction (i.e., a node with a high score), the largest contribution came from a seed node that was a direct neighbor. They also varied the algorithm's parameters to allow the random walker to travel longer paths in the network, with the expectation that this change would diffuse high scores beyond the direct neighborhood of the seed nodes. However, the outcome of these experiments only confirmed their earlier findings. This observation prompted us to delve deeper into the widely accepted belief that diffusion algorithms utilize the non-local neighborhood of the network.[14]

We developed a novel path-based strategy for provenance tracing in random walk-based algorithms by leveraging our ability to compute the magnitude of the contribution made to a node's score by each path starting at a seed node and ending at the corresponding node (Section 2.3). In this work, we focused on two well-known network diffusion algorithms: Regularized Laplacian (RL)[19] and Random Walk with Restart (RWR).[20] In these algorithms, the predicted score s_v of a node v is the linear sum of the 'contribution' made by each seed node to s_v (Section 2.2). In our path-based approach, we show for every u–v path (where u is a seed node and v is a node), how to compute its contribution to s_v (Section 2.3). We further proposed two new measures: 1) path-based effective diffusion (PED) (Section 2.3.1) and 2) diffusion betweenness (Section 2.3.3). PED measures the fraction of a node's score contributed by multi-edge paths. Diffusion betweenness measures a node's importance in score propagation by quantifying the total contribution transmitted along paths passing through that node.

We implemented RL and RWR on human protein-protein interaction (PPI)[1,2] networks with SARS-CoV-2 human protein interactors as seed nodes.[18] Analysis of PED demonstrated that a substantial amount of a node's score was contributed via paths with more than one edge. Moreover, the vast majority of the nodes in the network were involved in paths carrying non-zero contributions to the top-scoring nodes. We concluded that diffusion algorithms do exploit the non-local structure of a network. Our analysis of diffusion betweenness further revealed the biological relevance of the proteins playing a critical role in score propagation across the network, as they were enriched with essential human genes and interactors of other viruses. Finally, as a case study, we applied our path-based approach to a signaling network containing causal interactions between human proteins.[21] We observed that the edges in the paths that contributed most to the scores of top-scoring nodes had a high degree of overlap with COVID-19-related pathways. Our new path-based analysis may produce novel insights in systems biology.

2. Methods

In this section, we provided a detailed description of two network diffusion algorithms: Regularized Laplacian (RL) and Random Walk with Restart (RWR). Next, we outlined both the

node-based and path-based approaches to provenance tracing. Finally, we defined the diffusion betweenness score and presented the formula used for its computation.

2.1. *Network diffusion algorithms*

Let $G = (V, E, w)$ be a weighted network where each node $v \in V$ is a human protein, and each edge $(u, v) \in E$ represents either an interaction between proteins u and v when G is an undirected PPI network or causal effect of u on v when G is a directed signaling network, and w_{vu} specifies the weight of edge (u, v). Given G and a set $P \subset V$ of seed nodes, e.g., human proteins that interact with SARS-CoV-2 proteins, we seek to compute a score s_v for each node v indicating our confidence that it either physically interacts with or is functionally linked to the virus. In this avenue, we employ RL and RWR algorithm on the undirected PPI networks and RWR on the directed signaling network.

Regularized Laplacian (RL).

(1) Define a label vector $\vec{y}$ over the nodes in G where $y_u = 1$ if node $u \in P$ and $y_u = 0$, otherwise.
(2) Define $W \in \mathbb{R}^{n \times n}$ as the adjacency matrix of G with edge weights, i.e., W_{vu}, the entry in row v and column u of W equals w_{vu} if (u, v) is an edge in G and 0, otherwise.
(3) Define D as a diagonal matrix with $D_{uu} = d_u = \sum_v W_{vu}$. Here, D contains the degree of every node u in G.
(4) Compute $\tilde{W} = D^{-1/2} W D^{-1/2}$, which denotes the normalized network.
(5) Compute the Laplacian of G as $\tilde{L} = \tilde{D} - \tilde{W}$, where we define $\tilde{D}$ to be a diagonal matrix with $\tilde{D}_{uu} = \sum_v \tilde{W}_{vu}$.
(6) Given a parameter $\alpha > 0$ and the goal to minimize the loss function:

$$\sum_{u \in V} (s_u - y_u)^2 + \alpha \sum_{(u,v) \in E} \tilde{W}_{vu} (s_u - s_v)^2 \tag{1}$$

where s_v is the computed score of node v, we obtain the score vector from the unique, global optimization of Equation (1) as follows:

$$\vec{s} = (I + \alpha \tilde{L})^{-1} \vec{y} \tag{2}$$

Here, $\tilde{L}$ is symmetric and positive semidefinite. Hence, $(I + \alpha \tilde{L})$ is symmetric, positive definite, and invertible.

Random walk with restart (RWR).

(1) Define $\vec{y}$, W, D as defined in RL.
(2) Compute the normalized network matrix $\tilde{W} = W D^{-1}$.
(3) Given a parameter $0 < \alpha < 1$, where α is the restart probability, and the goal to minimize the loss function:

$$\alpha \sum_{u \in V} \frac{(s_u - y_u)^2}{d_u} + (1 - \alpha) \sum_{(u,v) \in E} W_{vu} \left(\frac{s_u}{d_u} - \frac{s_v}{d_v} \right)^2 \tag{3}$$

we obtain the score vector from the unique, global optimization of Equation (3) as follows:

$$\vec{s} = \frac{\alpha}{|P|}(I - (1-\alpha)\tilde{W})^{-1}\vec{y} \tag{4}$$

Here, $\tilde{W}$ is column-stochastic (every column sums to 1 and all entries are nonnegative). By the Perron-Frobenius theorem, every eigenvalue for $(I-(1-\alpha)\tilde{W})$ is strictly positive, making $(I-(1-\alpha)\tilde{W})$ invertible.

We chose a value of α in RL (and RWR) such that it balances out the two terms in the quadratic loss function in Equation (1) (and Equation (3)). We did a binary search to find out the exact value for which the two loss terms are equal. For RL, α ranged between 0.94 and 1.33 and for RWR, α was 0.49 across all networks.

2.2. *Node-based provenance tracing*

For RL, let K denote the matrix $(I+\alpha\tilde{L})^{-1}$. According to Equation (2), s_v is the sum of K_{vu} over all nodes $u \in P$ (i.e., seed nodes); $s_v = \sum_{u\in P} K_{vu}$. Hence, K_{vu} denotes the contribution of seed node u to node v's score s_v. For RWR, let K denote the matrix $\alpha/|P|(I-(1-\alpha)\tilde{W})^{-1}$ (Equation (4))

Node-based effective diffusion (NED). Law *et al.*[17] summarized node-based provenance tracing with a score, namely, *node-based effective diffusion (NED)* where NED(v) is defined as the fraction of v's score contributed from seed nodes that are not direct neighbors. Following is the corresponding formula:

$$\text{NED}(v) = 1 - \frac{\sum_{u\in P\cap N(v)} K_{vu}}{s_v}, N(v) \text{ is the set of neighbors of node } v.$$

2.3. *Path-based provenance tracing*

In this section, we show how to express the score s_v of node v as the sum of weighted 'contributions' along paths that start at some seed node and end at v. First, we describe a matrix M that captures the contribution along an edge. Then we extend the formula to compute the contribution along longer paths.

***M* for RL.** We consider diffusion by the RL algorithm as a fluid flow model where no node contains any fluid initially and then a hypothetical fluid is pumped into every seed node at a constant rate of 1.[22] The fluid diffuses from one node to another at a first-order rate of $\alpha\tilde{D}$ and fluid *leaves* the network at a first-order rate of 1. M_{ji} is the rate of incoming flow at node j from node i via edge (i,j). Applying the fluid flow model we formulated M as follows: $M = \alpha\tilde{W}(I+\alpha\tilde{D})^{-1}$

***M* for RWR.** For RWR, M_{ji} denotes the probability that a random walker moves from node i to node j in a single time step. Hence, we formulated M as follows: $M = (1-\alpha)\tilde{W}$

Having established how to compute the contribution along a single edge using M, we now extend this to longer paths involving multiple edges. Given a u and any node v in V, for a

u–v path in G, the contribution made by this path to s_v is the product of the edge weights along the path, where the edge weights are from M. For instance, via a path $u \to t \to x \to v$ containing three edges: $(u,t),(t,x),(x,v)$, u's contribution to s_v is $M_{tu}M_{xt}M_{vx}$. More generally, M^n_{vu} is the contribution from a u to a node v via all paths of length n.

2.3.1. *Path-based effective diffusion*

We define *path-based effective diffusion* (PED) of a node v as the fraction of its score, s_v contributed by paths that start at some seed node, end at v, and contain more than one edge. We formulate $PED(v)$ as follows:

$$\text{PED}(v) = 1 - \frac{c(v)\sum_{u\in P} M_{vu}}{s_v},$$

where $\sum_{u\in P} M_{vu}$ denotes the incoming contribution to s_v along paths with one edge from any seed node and $c(v)$ denotes a weight. For RL, $c(v) = (1+\alpha\tilde{D}_{vv})^{-1}$ while for RWR, $c(v) = \frac{\alpha}{|P|}$ is a constant. Note that $M_{vu} = 0$ if u is not a neighbor of v.

2.3.2. *Find top m contributing paths*

The path-based provenance tracing, more precisely the derived matrix M provided us with the ability to compute contribution coming along any path starting at a seed node to any node in the network. Though there can be infinite paths between a seed node and any other node (considering paths with cycles in them), the most interesting paths are the ones that carry substantial contributions to nodes' scores. In this avenue, from all the paths starting at any seed node carrying contribution to any of the top k (in this experiment $k = |P|$) predicted nodes, we find m most contributing paths.

Suppose, M' is the weighted adjacency matrix of a graph where $M'_{ji} = -\log_{10}(M_{ji})$. In this graph, the m-shortest paths from a seed node u to a node v are the most *contributing* m paths from u to s_v. To compute the m-shortest paths between any seed node and the top-k predicted proteins (based on their scores), we augmented the network with two dummy nodes: (1) a super-source, connected to all seed nodes, and (2) a super-sink, connected from the top-k predicted nodes. We then computed the m-shortest paths from the super-source to the super-sink using Eppstein's k-shortest paths algorithm that considers looped paths.

2.3.3. *Computing diffusion betweenness*

The diffusion betweenness score of a node aims to capture the total amount of score propagating along paths on which the corresponding node lies. For a node t, we define its *diffusion betweenness* score $b(t)$ as the sum of the contribution made to s_v along *all* π_{u-t-v} paths in which u ranges over all seed nodes, t is an internal node, and v ranges over all nodes in V.

In practice, while computing diffusion betweenness, we use a heuristic where we consider only paths with at most four edges, since a negligible fraction of the score propagates via paths with more edges.[23,24] To formulate $b(t)$, we introduce another matrix Q where Q_{vt} is the

contribution from any seed node to s_v along *all* paths (length ≤ 4) ending at v having t as an intermediate node. We formulate $b(t)$ as follows: $b(t) = \sum_{v \in V} c(v) Q_{vt}$

Computation of Q. We define X^{lm} where X^{lm}_{vt} holds the contribution to s_v along all paths of length $(l + m)$ with t being the lth intermediate node. We compute $Q_{vt} = \sum_{(l+m) \leq 4} X^{lm}_{vt}$.

Computation of X^{lm}. Let us divide any π_{u-t-v} path into two segments: (1) from the seed node to the intermediate node, i.e., u–t, and (2) from the intermediate node to the endpoint, i.e., t–v. As the contribution along a path is only the product of the edge weights (from M) along that path, the contribution of π to s_v equals the product of: (1) contribution along u–t, (2) contribution along t–v. We can extend this two-component-product concept to compute X^{lm}_{vt} as follows:

$$X^{lm}_{vt} = M^m_{vt}.(M^l \vec{y})_t,$$

Here, $(M^l \vec{y})_t$ is the sum of the contribution to s_t along *all* u–t paths of length l and M^m_{vt} is the sum of the contribution to s_v along *all* t–v paths of length m.

3. Datasets

3.1. *Experimentally determined nodes*

We considered 332 human proteins that interact with SARS-CoV-2[18] as seed nodes. We included the ACE2 receptor in this set.

3.2. *Human protein-protein interaction networks*

We considered three human protein-protein interaction (PPI) networks: STRING,[1] BioGRID-Phy,[2] and BioGRID-Y2H.[2] In STRING, an edge may represent physical binding or indirect functional interaction. While BioGRID-Y2H contains direct physical interactions only, BioGRID-Phy can additionally contain interactions denoted by the co-existence of two proteins in a stable complex. After mapping the proteins in a PPI network to their corresponding UniProt IDs and considering only the largest connected component, STRING contained $16,315$ nodes and $246,086$ edges whereas BioGRID-Phy had $17,758$ nodes and $723,772$ edges, and BioGRID-Y2H had $13,185$ nodes and $94,427$ edges.

3.3. *Causal interaction network for human*

We considered a directed signaling network from SIGNOR 3.0[21] database that contains manually curated causal interactions between biological entities such as proteins, protein complexes, and chemicals. The causal interaction denotes an up or down-regulation effect with a mechanism such as binding, phosphorylation, and transcriptional activation associated with it. After removing chemical and protein complexes and mapping the proteins to their corresponding UniProt, we retained $4,853$ nodes and $11,866$ edges.

3.4. *SARS-CoV-2 hallmark pathways and COVID-19 causal network*

We extracted causal interaction data organized in nine network modules or pathways, i.e., the hallmark pathways, representing the impact of SARS-CoV-2 proteins on cellular functions[21] of humans. To address the lack of evidence regarding SARS-CoV-2, these hallmark pathways contain manually annotated and validated causal interactions with SARS-CoV-2 as well as SARS-CoV-1, Middle East Respiratory Syndrome (MERS) proteins, and the human host from the literature.[25] We also considered a single network combining these nine hallmark pathways, i.e., COVID-19 causal network, available in SIGNOR 3.0.[21]

3.5. *Essential genes*

We extracted the essential genes for humans from the Database of Essential Genes.[26] It contained $2,452$ essential genes for humans from Liao *et al.*[27] and Georgei *et al.*,[28] where Georgei *et al.* considered human orthologs of mouse essential genes as putative essential genes for humans. This database also contained human essential proteins from exome sequencing-based experiments. We retained $4,807$ essential genes after mapping them to their UniProt IDs.

3.6. *Viral interactors*

We compiled a set of viral interactors, i.e., human proteins that physically interact with viruses from three databases of host-pathogen interactions.[29–31] After mapping these human proteins to their corresponding UniProt IDs and removing the SARS-CoV-2 interactors the final dataset contained $7,808$ proteins.

4. Results

In this section, first, we described how the path-based approach to provenance tracing confirmed that random walk based diffusion algorithms utilize the non-local network structure to propagate scores from seed nodes (Section 4.1). Next, we demonstrated that nodes with high diffusion betweenness scores are biologically significant (Section 4.2). Finally, by extending our analysis to directed networks, we showed that the paths with high contributions to top-scoring nodes (Section 2.3.2) often contain biologically relevant causal edges (Section 4.3).

For the first two experiments, using human interactors of SARS-CoV-2[18] as seed nodes, we ran both RL and RWR on three undirected human PPI networks, namely, STRING,[1] BioGRID-Physical,[2] and BioGRID-Y2H.[2] STRING consists of edges representing physical binding or indirect functional interaction. While BioGRID-Y2H contains only direct physical interactions, BioGRID-Phy also includes indirect interactions defined by the presence of two proteins in a stable complex. For our final analysis, we ran RWR on directed network SIGNOR[21] that contains directed causal interaction between proteins in human.

4.1. *Path-based provenance tracing confirms the ability of diffusion algorithms to capture non-local network structure*

We ran each diffusion algorithm to predict the score of each node in a network. We ranked the nodes that were not seed nodes based on their scores. We computed the path-based effective

diffusion (PED) and node-based effective diffusion (NED) scores for each of the top k proteins in this ranked list, setting k to be the number of seed nodes in the network.

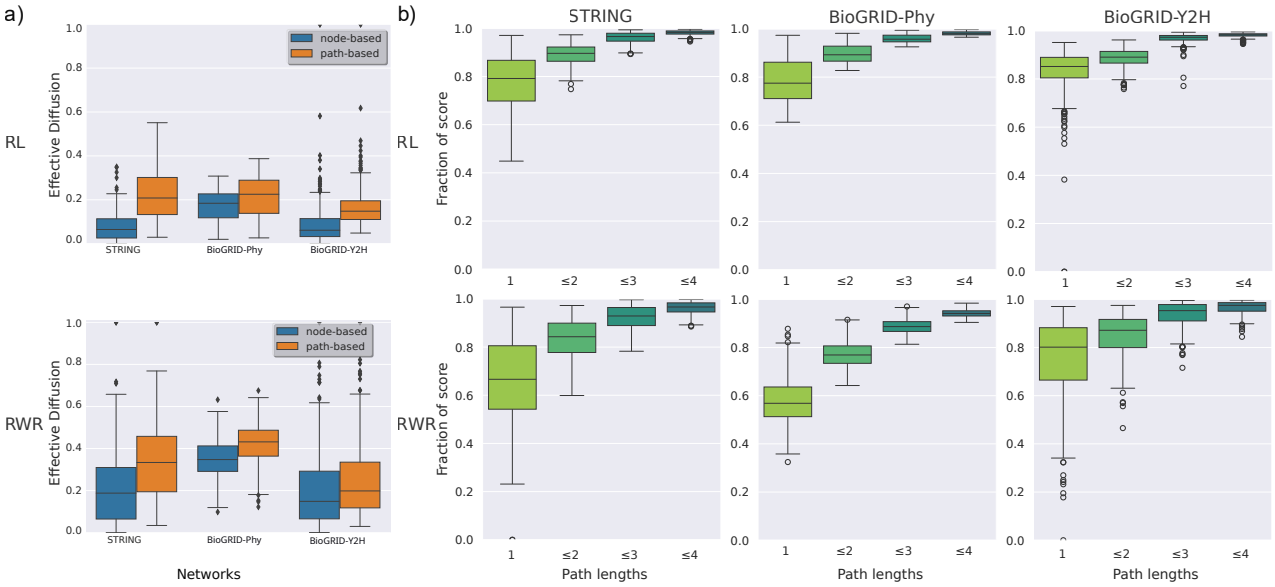

Fig. 2. **Path-based approach for provenance tracing.** a) Comparison of node-based and path-based effective diffusion. Effective diffusion score for top k predictions for three PPI networks: STRING, BioGRID-Phy, BioGRID-Y2H. Here, k is equal to the number of seed nodes in the network. b) Contribution along paths of different lengths. The x-axis shows the maximum allowed path length (e.g., ≤ 2, ≤ 3), and the y-axis shows the fraction of the score contributed by paths whose maximum length falls within that range.

We observed that for RL and RWR on all three PPI networks, a substantial amount of a node's score was contributed via paths with multiple edges (Figure 2a). For example, for RL in the STRING network, we observed that the median PED was 0.2, i.e., paths of length more than one contributed at least 20% of the score for at least half the nodes (Figure 2a). We also noted that PED was significantly higher than NED (Figure 2a, p-values $< 8.08 \times 10^{-10}$, one-sided Mann Whitney U test) across all the algorithm-network pairs. Note that PED considers any contribution from the seed nodes as long as it arises from paths of length > 1 (irrespective of whether the seed node is a direct neighbor or not), whereas NED disregards any contributions from direct neighboring seed nodes. The larger values of PED in comparison to NED scores highlight that diffusion from even a direct neighbor can and does traverse paths of length > 1, leveraging the non-local neighborhood of the network. To further investigate the algorithms' ability to leverage the network's topology, we computed the precise contribution coming along paths of length ≤ 2, ≤ 3, and ≤ 4. We observed that a substantial fraction of the diffusion scores arose from paths with more than two edges (Figure 2b).

We computed the diffusion betweenness score for each node to evaluate whether the diffusion algorithm is confined to interactions between seed nodes and top-scoring nodes or engages other nodes in score propagation. We observed that 73% to 99% of the nodes are intermediate nodes in paths carrying non-zero contributions to top-scoring k nodes with 17%–42% nodes

lying on paths carrying at least 0.1% of the score of any top-scoring node.

In summary, node-based effective diffusion collapsed contributions from all paths between two nodes into a single quantity, creating a misleading impression that network diffusion algorithms are "local" in practice. In contrast, path-based effective diffusion in combination with diffusion betweenness provided a better way to capture the diffusion algorithms' ability to exploit the non-local structure of a network.

4.2. *Proteins with high diffusion betweenness are enriched with viral interactors and essential genes*

We sought to investigate the biological relevance of the nodes that are critical for score propagation in diffusion algorithms, i.e., the nodes with high diffusion betweenness scores. Accordingly, we analyzed the overlap between these nodes and two specific categories: known human interactors of viral proteins and essential genes.

Overlap with viral interactors. Viruses from diverse families exploit shared molecular mechanisms in their interactions with host cells throughout key stages of their life cycles.[32] Hence, we investigated whether nodes with high diffusion betweenness are enriched in human proteins interacting with other viruses. To this end, we computed the overlap between viral interactors (Section 3.6) and the k ($k \in \{200, 400, 600, 800, 1{,}000, 2{,}000, 5{,}000, 10{,}000\}$) proteins with the highest diffusion betweenness scores (excluding seed nodes and top predictions). Using SARS-CoV-2 interactors as seed nodes, we observed statistically significant overlaps for both RL and RWR across all values of k in STRING and BioGRID networks (adjusted p-value < 0.01, one-sided Fisher's exact test with Benjamini-Hochberg correction for multiple hypothesis testing, Figure 3a). For instance, in the STRING network, 76% and 69% of the top 200 proteins (ranked by diffusion betweenness) were viral interactors for RL and RWR, respectively, compared to only 38% across the entire network.

To test whether these findings depended on the specific choice of seed nodes, we repeated the analysis with a set of random seed nodes of the same size as the set of SARS-CoV-2 interactors. We generated random seeds in two ways: (i) preserving the degree sequence of the original seed nodes (i.e., SARS-CoV-2 interactors),[17] and (ii) selecting nodes uniformly at random from the entire network. In both cases, the overlap between viral interactors and proteins with high diffusion betweenness remained statistically significant (adjusted p-value < 0.01). We interpreted this result to suggest that viral interactors were important in mediating diffusion even from randomly-selected seed nodes.

Overlap with essential genes. Essential genes govern the fundamental functions necessary for cell survival.[33] We hypothesized that essential genes critical for disseminating biological information in a cell might also be crucial to network diffusion algorithms in propagating scores. To test our hypothesis, we computed the overlap between the k ($k \in \{200, 400, 600, 800, 1{,}000, 2{,}000, 5{,}000, 10{,}000\}$) proteins with the highest diffusion betweenness scores (excluding seed nodes and top predicted proteins) and essential human genes (Section 3.5). using SARS-CoV-2 interactors as seed nodes, we observed statistically significant overlap (adjusted p-value < 0.01, one-sided Fisher's exact test with Benjamini-Hochberg correction for multiple

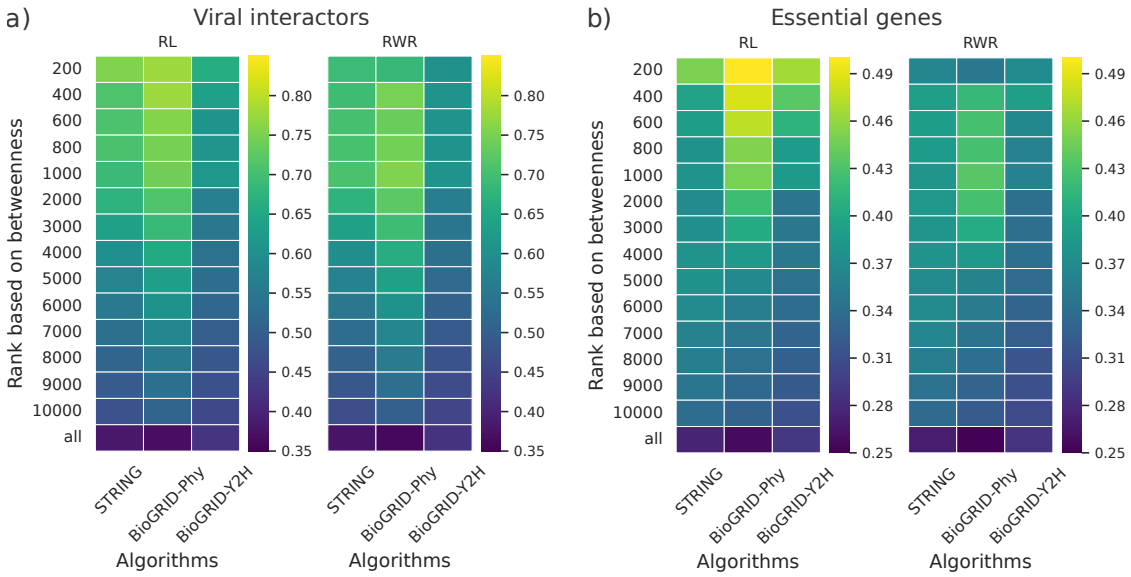

Fig. 3. **Overlap between proteins with high diffusion betweenness and the proteins of interest (i.e., viral interactors and essential genes).** Each column corresponds to a network and each row corresponds to k, protein rank based on diffusion betweenness. **a)** Each cell in the heatmap indicates the fraction of **viral interactors** in top-ranked k proteins. The row 'all' contains the fraction of viral interactors among all nodes in the corresponding network. **b)** Each cell in the heatmap indicates the fraction of **essential genes in humans** in the top-ranked k proteins. The row 'all' contains the fraction of essential genes among all nodes in the corresponding network.

hypothesis testing, Figure 3b) for each algorithm-network pair. For instance, in the STRING network, 45% and 37% of the top 200 proteins (ranked by diffusion betweenness) were essential genes for RL and RWR, respectively, compared to only 27% across the entire network. Repeating this analysis with random seed sets yielded a similar pattern of significant overlaps (adjusted p-value < 0.01), further highlighting the seed-node–agnostic importance of essential genes in network diffusion.

4.3. *Top contributing paths show significant overlap with COVID causal network and hallmark SARS-CoV pathways*

To demonstrate an application of our path-based approach in identifying biologically relevant pathways, we extended our analysis to a directed signaling network containing causal interactions between pairs of human proteins.[21] We considered all the paths in this network that started at a seed node and ended at any one of the top-k proteins (based on prediction score). We applied Eppstein's k-shortest paths algorithm to compute paths with the m highest contributions in this set (Section 2.3.2).

For $m = 1,000$, we found a significantly high overlap (Fisher's exact test p-value $= 1.76 \times 10^{-23}$) between the edges in the m most contributing paths and edges appearing in the COVID-19 casual network[25] (Section 3.4). We concluded that the RWR algorithm propagates scores from known SARS-CoV-2 interactors to other proteins in the network using causal interactions that are known to be modulated by viral infection.

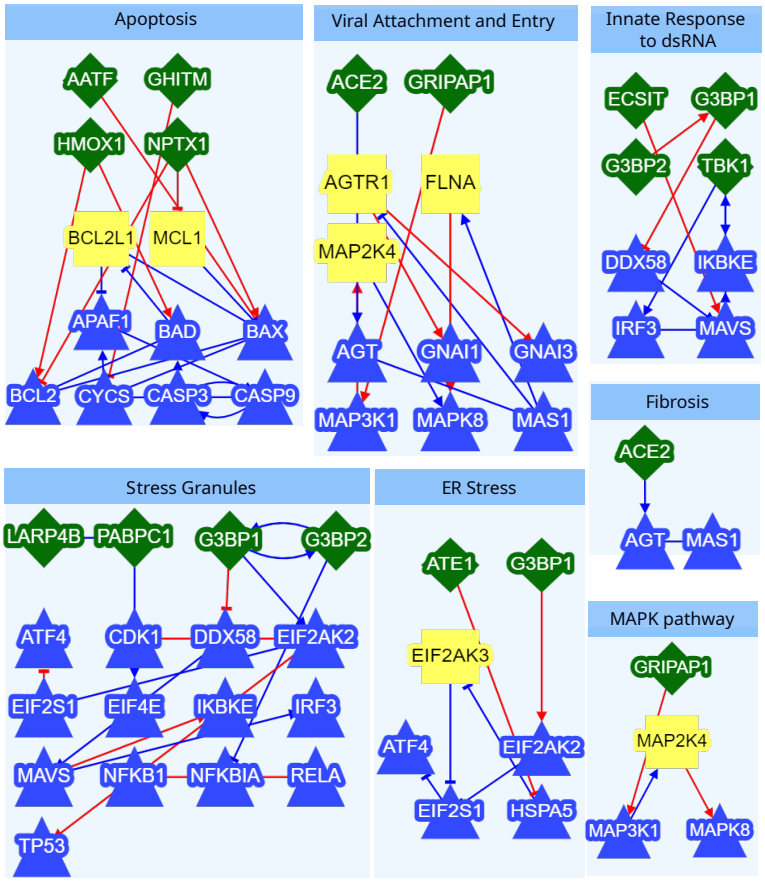

Fig. 4. **Top contributing paths assigned to hallmark SARS-CoV-2 pathways.** Each block here is for one SARS-CoV-2 causal pathway containing the top contributing paths assigned to it where we assigned a path to the pathway with the most overlap. A green diamond denotes a seed node (i.e., known SARS-CoV-2 interactors) and a blue triangle indicates protein appearing in top k predictions. Yellow rectangles denote the rest of the proteins. A blue edge indicates an edge common between a top contributing path and the corresponding SARS-CoV-2 pathway and a red edge denotes otherwise. An edge with a triangular arrow and tee denote up-regulation and down-regulation effect, respectively.

Next, we analyzed each of the top contributing paths separately. Among the $m = 1,000$ most contributing paths, we considered those with at least two edges and at most one edge missing from the COVID-19 causal network. These criteria yielded 55 paths with two to five edges. To discern the specific SARS-CoV-2 hallmark pathways[21,25] in which the top-contributing paths were involved, we assigned each path to the pathway with the most overlapping edges. We discovered that these paths appeared (either entirely or with one non-overlapping edge) in seven hallmark pathways (Section 4.3), namely, apoptosis, attachment and entry, stress granules, innate response to dsRNA, ER stress, MAPK-pathway, and fibrosis.

We further investigated the non-overlapping causal interactions. There were seven such edges across all the paths assigned to the apoptosis pathway. The source nodes in all these edges were experimentally determined SARS-CoV-2 interactors[18] with evidence of a regulatory effect on apoptosis: NPTX1,[34] AATF,[35] HMOX1,[36] and GHITM.[37] Additionally, the non-overlapping causal interaction (ATE1, HSPA5) in the ER stress pathway is crucial to the

human stress response. ATE1 facilitates the degradation of damaged proteins through arginylation,[38] while HSPA prevents apoptosis by assisting in the proper folding of nascent proteins and the refolding of misfolded protein under stress conditions.[39] Another missing signaling edge (G3BP1, DDX58) in the stress granule contains G3BP1, a known SARS-CoV-2 interactor[18] with evidence of involvement in the corresponding pathway.[40] From these observations, we concluded that non-overlapping interactions in top-contributing paths may result from incomplete annotation of COVID-19 causal pathways. These interactions could potentially be experimentally validated as being exploited by SARS-CoV-2 during viral infection.

5. Computational Setup and Runtime

We multiplied (dense) matrices using the `matmul` function in the NumPy package. We inverted matrices using SciPy's `linalg.inv` function, which computes the exact inverse up to floating-point precision. Both operations have a time complexity of $O(n^3)$. We conducted all experiments on a server equipped with an Intel 13th Gen Core i9-13900K CPU (24 cores, 32 threads, base frequency 3.0 GHz, turbo frequency up to 5.8 GHz). We observed the highest matrix multiplication runtime of 39 seconds and matrix inversion runtime of 24 seconds for BioGRID-Phy, which has the largest number of nodes among the three networks.

Acknowledgments

National Science Foundation grants 2233967 and 1759858 to TMM supported this research.

References

1. D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork *et al.*, The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets, *Nucleic Acids Research* **49**, D605 (2021).
2. R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang *et al.*, The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions, *Protein Science* **30**, 187 (2021).
3. T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding and K. Huang, MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification, *Nature Communications* **12**, p. 3445 (2021).
4. H. Göös, M. Kinnunen, K. Salokas, Z. Tan, X. Liu, L. Yadav, Q. Zhang, G.-H. Wei and M. Varjosalo, Human transcription factor protein interaction networks, *Nature Communications* **13**, p. 766 (2022).
5. R. Sharan, I. Ulitsky and R. Shamir, Network-based prediction of protein function, *Molecular Systems Biology* **3**, p. 88 (2007).
6. W. S. Noble, R. Kuang, C. Leslie and J. Weston, Identifying remote protein homologs by network propagation, *The FEBS Journal* **272**, 5119 (2005).
7. K. Mitra, A.-R. Carvunis, S. K. Ramesh and T. Ideker, Integrative approaches for finding modular structure in biological networks, *Nature Reviews Genetics* **14**, 719 (2013).
8. T. Ideker and R. Sharan, Protein networks in disease, *Genome Research* **18**, 644 (2008).
9. D.-Y. Cho, Y.-A. Kim and T. M. Przytycka, Chapter 5: Network biology approach to complex diseases, *PLoS Computational Biology* **8**, p. e1002820 (2012).

10. P. Csermely, T. Korcsmáros, H. J. Kiss, G. London and R. Nussinov, Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review, *Pharmacology & Therapeutics* **138**, 333 (2013).
11. Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen and J. Zeng, A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, *Nature Communications* **8**, p. 573 (2017).
12. D. Morselli Gysi, Í. Do Valle, M. Zitnik, A. Ameli, X. Gan, O. Varol, S. D. Ghiassian, J. Patten, R. A. Davey, J. Loscalzo *et al.*, Network medicine framework for identifying drug-repurposing opportunities for COVID-19, *Proceedings of the National Academy of Sciences* **118**, p. e2025581118 (2021).
13. J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo and A.-L. Barabási, Uncovering disease-disease relationships through the incomplete interactome, *Science* **347**, p. 1257601 (2015).
14. L. Cowen, T. Ideker, B. J. Raphael and R. Sharan, Network propagation: a universal amplifier of genetic associations, *Nature Reviews Genetics* **18**, 551 (2017).
15. M. R. Karim, T. Islam, M. Shajalal, O. Beyan, C. Lange, M. Cochez, D. Rebholz-Schuhmann and S. Decker, Explainable AI for bioinformatics: methods, tools and applications, *Briefings in Bioinformatics* **24**, p. bbad236 (2023).
16. S. Kasif and R. J. Roberts, We need to keep a reproducible trace of facts, predictions, and hypotheses from gene to function in the era of big data, *PLoS Biology* **18**, p. e3000999 (2020).
17. J. N. Law, K. Akers, N. Tasnina, C. M. D. Santina, S. Deutsch, M. Kshirsagar, J. Klein-Seetharaman, M. Crovella, P. Rajagopalan, S. Kasif *et al.*, Interpretable network propagation with application to expanding the repertoire of human proteins that interact with SARS-CoV-2, *GigaScience* **10**, p. giab082 (2021).
18. D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney *et al.*, A SARS-CoV-2 protein interaction map reveals targets for drug repurposing, *Nature* **583**, 459 (2020).
19. F. Fouss, K. Francoisse, L. Yen, A. Pirotte and M. Saerens, An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification, *Neural Networks* **31**, 53 (2012).
20. L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank citation ranking: Bringing order to the web.*, tech. rep., Stanford Infolab (1999).
21. P. Lo Surdo, M. Iannuccelli, S. Contino, L. Castagnoli, L. Licata, G. Cesareni and L. Perfetto, Signor 3.0, the signaling network open resource 3.0: 2022 update, *Nucleic Acids Research* **51**, D631 (2023).
22. Y. Qi, Y. Suhail, Y.-y. Lin, J. D. Boeke and J. S. Bader, Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions, *Genome Research* **18**, 1991 (2008).
23. S. Mostafavi, A. Goldenberg and Q. Morris, Labeling nodes using three degrees of propagation, *PloS One* **7**, p. e51947 (2012).
24. W. Ba-Alawi, O. Soufan, M. Essack, P. Kalnis and V. B. Bajic, DASPfind: new efficient method to predict drug-target interactions, *Journal of Cheminformatics* **8**, 1 (2016).
25. L. Perfetto, E. Micarelli, M. Iannuccelli, P. Lo Surdo, G. Giuliani, S. Latini, G. M. Pugliese, G. Massacci, S. Vumbaca, F. Riccio *et al.*, A resource for the network representation of cell perturbations caused by SARS-CoV-2 infection, *Genes* **12**, p. 450 (2021).
26. H. Luo, Y. Lin, T. Liu, F.-L. Lai, C.-T. Zhang, F. Gao and R. Zhang, DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools, *Nucleic Acids Research* **49**, D677 (2021).
27. B.-Y. Liao and J. Zhang, Null mutations in human and mouse orthologs frequently result in

different phenotypes, *Proceedings of the National Academy of Sciences* **105**, 6987 (2008).
28. B. Georgi, B. F. Voight and M. Bućan, From mouse to human: evolutionary genomics analysis of human orthologs of essential genes, *PLoS Genetics* **9**, p. e1003484 (2013).
29. S. Durmuş Tekir, T. Çakır, E. Ardıç, A. S. Sayılırbaş, G. Konuk, M. Konuk, H. Sarıyer, A. Uğurlu, İ. Karadeniz, A. Özgür *et al.*, PHISTO: pathogen-host interaction search tool, *Bioinformatics* **29**, 1357 (2013).
30. M. G. Ammari, C. R. Gresham, F. M. McCarthy and B. Nanduri, HPIDB 2.0: a curated database for host-pathogen interactions, *Database* **2016**, p. baw103 (2016).
31. X. Yang, X. Lian, C. Fu, S. Wuchty, S. Yang and Z. Zhang, HVIDB: a comprehensive database for human-virus protein-protein interactions, *Briefings in Bioinformatics* **22**, 832 (2021).
32. M. S. Ravindran, P. Bagchi, C. N. Cunningham and B. Tsai, Opportunistic intruders: how viruses orchestrate ER functions to infect cells, *Nature Reviews Microbiology* **14**, 407 (2016).
33. L. Funk, K.-C. Su, J. Ly, D. Feldman, A. Singh, B. Moodie, P. C. Blainey and I. M. Cheeseman, The phenotypic landscape of essential human genes, *Cell* **185**, 4634 (2022).
34. Y. Zhao, Y. Yu, W. Zhao, S. You, M. Feng, C. Xie, X. Chi, Y. Zhang and X. Wang, As a downstream target of the AKT pathway, NPTX1 inhibits proliferation and promotes apoptosis in hepatocellular carcinoma, *Bioscience Reports* **39**, p. BSR20181662 (2019).
35. D. Welcker, M. Jain, S. Khurshid, M. Jokić, M. Höhne, A. Schmitt, P. Frommolt, C. M. Niessen, J. Spiro, T. Persigehl *et al.*, AATF suppresses apoptosis, promotes proliferation and is critical for Kras-driven lung cancer, *Oncogene* **37**, 1503 (2018).
36. Y. S. Basmaeil, D. Algudiri, R. Alenzi, A. Al Subayyil, A. Alaiya and T. Khatlani, HMOX1 is partly responsible for phenotypic and functional abnormalities in mesenchymal stem cells/stromal cells from placenta of preeclampsia (PE) patients, *Stem Cell Research & Therapy* **11**, 1 (2020).
37. S. Li, L. Wang, M. Berman, Y.-Y. Kong and M. E. Dorf, Mapping a dynamic innate immunity protein interaction network regulating type I interferon production, *Immunity* **35**, 426 (2011).
38. K. Deka, A. Singh, S. Chakraborty, R. Mukhopadhyay and S. Saha, Protein arginylation regulates cellular stress response by stabilizing HSP70 and HSP40 transcripts, *Cell Death Discovery* **2**, 1 (2016).
39. G. Jego, A. Hazoumé, R. Seigneuric and C. Garrido, Targeting heat shock proteins in cancer, *Cancer Letters* **332**, 275 (2013).
40. V. Iadevaia, J. M. Burke, L. Eke, C. Moller-Levet, R. Parker and N. Locker, Novel stress granule-like structures are induced via a paracrine mechanism during viral infection, *Journal of Cell Science* **135**, p. jcs259194 (2022).