Neil, Daniel, et al. *Interpretable Graph Convolutional Neural Networks for Inference on Noisy Knowledge Graphs.* Machine Learning for Health, 2018.

Interpretable Graph Convolutional Neural Networks for Inference on Noisy Knowledge Graphs

Daniel Neil **Joss Briody** **Alix Lacoste** **Aaron Sim** **Paidi Creed** **Amir Saffari**
BenevolentAI
Brooklyn, NY and London, UK
{daniel.neil,joss.briody,alix.lacoste,
aaron.sim,paidi.creed,amir.saffari}@benevolent.ai

Abstract

In this work, we provide a new formulation for Graph Convolutional Neural Networks (GCNNs) for link prediction on graph data that addresses common challenges for biomedical knowledge graphs (KGs). We introduce a regularized attention mechanism to GCNNs that not only improves performance on clean datasets, but also favorably accommodates noise in KGs, a pervasive issue in real-world applications. Further, we explore new visualization methods for interpretable modelling and to illustrate how the learned representation can be exploited to automate dataset denoising. The results are demonstrated on a synthetic dataset, the common benchmark dataset FB15k-237, and a large biomedical knowledge graph derived from a combination of noisy and clean data sources. Using these improvements, we visualize a learned model's representation of the disease cystic fibrosis and demonstrate how to interrogate a neural network to show the potential of PPARG as a candidate therapeutic target for rheumatoid arthritis.

1 Introduction and Motivation

In biomedicine, knowledge graphs are critical in understanding complex diseases and advancing drug discovery. The problem of identifying novel therapeutic targets for a given disease for example can be formulated as a link prediction problem, a major area of study in statistical relational learning [9]. Various tensor factorization methods have found success [19, 3, 30, 25], as have convolutional neural networks (CNNs) [24, 8, 17] and path-based methods [29, 6]. Recently, graph convolutional neural networks (GCNNs) have been applied to this problem, outperforming a number of standard models [21].

Real-world knowledge graphs tend to contain relationships from multiple sources of varying quality. For example, drug-target associations extracted from unstructured text are less reliable than manually curated ones. In this work, we make two important contributions. First, we demonstrate that introducing a learnable link weight outperforms existing tensor factorization and GCNN models in the presence of noise. This new model assigns low weights to unreliable edges, which can be viewed as learning an edge filter to remove unreliable or uninformative edges from the knowledge graph. Second, we demonstrate that this model is more interpretable because it allows measuring the impact of a particular edge on a prediction by adjusting the link weight or removing the edge entirely. Moreover, when our knowledge graph is constructed by combining information from a number of diverse sources of variable reliability, the learned link weights can be used to assess the quality or relevance of different data sources. These benefits are illustrated with applications in drug-target discovery, where the added value of interpretability is particularly great.

Machine Learning for Health (ML4H) Workshop at NeurIPS 2018.

2 Model Formulation

Let $\mathcal{E}$ denote the set of all entities and $\mathcal{R}$ the set of all relation types in a KG, represented by a directed multigraph $\mathcal{G}$. An element of the KG can be represented by the triple $(e_s, r, e_o) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$. Knowledge graph embedding models learn vector representations of entities, $\mathbf{e}_i \in \mathbb{R}^{d_e}$, as well as relations $\mathbf{r} \in \mathbb{R}^{d_r}$, and a mapping (decoder) $f : \mathbb{R}^{d_e} \times \mathbb{R}^{d_r} \times \mathbb{R}^{d_e} \rightarrow [0, 1]$ assigning a probability of existence to each triple. To learn the entity embeddings, we use a GCNN model. Our implementation re-writes the GCNN introduced in [12, 33] as a series of matrix and element-wise multiplications:

$$H^{(l+1)} = \sigma\left(B^{(l)} + \sum_{r \in \mathcal{R}} (C_r \odot A_r)(H^{(l)} W_r^{(l)})\right) \tag{1}$$

where $\sigma(\cdot)$ is an element-wise nonlinear function (e.g. a ReLU [16]), $H^{(0)} := I_N$ is the N-dimensional identity matrix, $B^{(l)} \in \mathbb{R}^{N \times k^{(l)}}$ is a $k^{(l)}$ dimensional bias for each node and $C_r \in \mathbb{R}^{N \times N}$ is a fixed scaling factor between two connected nodes. The adjacency matrix is $A_r \in \{0, 1\}^{N \times N}$, the hidden representation $H^{(l)} \in \mathbb{R}^{N \times k^{(l)}}$, and the weight matrix is denoted, $W_r^{(l)} \in \mathbb{R}^{k^{(l+1)} \times k^{(l)}} \forall r \in \mathcal{R}$. The final layer $H^{(L)}$ contains as its rows the embedding vector $\mathbf{e}_i$ for each entity, i.e $d_e = k^{(l)}$. Relation vectors are learned by look-up in a simple embedding matrix $R \in \mathbb{R}^{|\mathcal{R}|} \times \mathbb{R}^{d_r}$. To calculate probabilities, we can use any choice of the decoder $f(\cdot)$ [30, 25, 19]. This work focuses on both the Complex decoder ([25]) as well as the DistMult model:

$$f(e_s, R_r, e_o) = e_s^T R_r e_o \tag{2}$$

We propose an attention model in which each link has an independent, learnable weight designed to approximate the *usefulness* of that link. Following the success of [27], we constrain our attention weights to have a fixed total "budget":

$$C_{r,i,j} = \frac{1}{\sum_{r' \in \mathcal{R}} \sum_{j' \in \mathcal{N}_i^r} |\hat{C}_{r',i,j'}|} |\hat{C}_{r,i,j}| \tag{3}$$

where $\hat{C}_{r,i,j}$ is initialized to 1, such that $C_{r,i,j} = 1/|\mathcal{N}_i|$ for all j at the start of training. This encourages the model to select only useful links that maximally aid in predicting true facts.

3 Experiments

All hyper-parameters were optimized using the mean reciprocal rank (MRR) performance on the validation set. This work employs a single GCNN layer, with a diagonalized weight matrix W_r and no non-linearity applied. We experimented with more layers and non-linear transforms but this did not improve performance. To train the model, we minimize the cross-entropy loss on the link class $\in \{0, 1\}$, and perform negative sampling following [18, 25, 3]. We performed grid search on the number of negatives sampled for each positive, $n \in \{1, 10, 20, 50\}$, as well as the embedding dimension $d \in \{50, 100, 200, 300\}$. Our reported results all use $n = 10$ and $d = 300$. For best performance, dropout with probability 0.5 was used on both the embeddings and on the links themselves. All embeddings were initialized with an L_2-norm of 1, so that all entities initially contribute equally in magnitude to the embedded representation at the start of training.

3.1 Performance on FB15k-237, with Attention

Having established the formulation of a GCNN with scalar attention, we now examine the relationship between data volume and noise. Four different conditions were trained, with results presented in Table 1. Our setup uses a dataset of a given size ("50%"), adding in noisy edges ("Noised") in equal volume to the remaining 50%, and skipping training on these possibly noisy edges ("Skip") while the edges remain in $A_{r,i,j}$. We see that GCNNs with attention consistently outperform those without attention, with MRR of 0.283 ± 0.006 and 0.272 ± 0.001, and hits@10 of 0.482 ± 0.009 and 0.475 ± 0.004 respectively, with the "Noised" and "Skip" condition of particular relevance to standard biomedical knowledge graphs.

	Hits@10				MRR			
Algorithm	100%	50%	Skip	Noised	100%	50%	Skip	Noised
DistMult	43.2	20.2	N/A	20.6	23.9	8.69	N/A	8.93
ComplEx	44.1	24.1	N/A	24.3	25.9	10.9	N/A	11.0
GCNN	47.5	33.2	25.8	21.4	27.2	16.8	13.3	11.1
GCNN w/att	**48.2**	**34.7**	**34.0**	**35.6**	**28.3**	**18.5**	**18.8**	**19.1**
R-GCN+ ([21])	41.7	-	-	-	24.9	-	-	-

Table 1: Performance on the FB15k-237 Dataset. Our results compare favorably with those reported in previous GCNN studies.

To further explore the sensitivity of the GCNN model to noise, we replicate the experimental setup of [20] on the benchmark standard FB15k-237 dataset [23, 3]. The results presented in Fig.1, right, reinforce previous findings that our proposed attention mechanism makes GCNNs more robust to noise. With an entirely clean knowledge graph, the addition of attention yields a 7.5% improvement in performance. When approximately 20-30% of the input graph is noise, the difference is $\approx$ 25-33%.

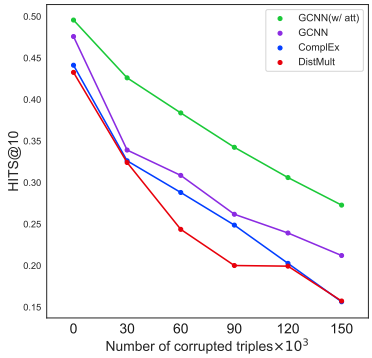

Figure 1: Test set HITS@10 for corrupted triples in FB15k-237.

3.2 Correctness and Interpretation in a Biological Knowledge Graph

As an example real-world application, we examine a subset of a proprietary knowledge graph with 708k edges, compiled from unstructured data sources including NCBI PubMed full text articles and a variety of structured sources including CTD [7], KEGG [11], OMIM [15], BioGRID [4], Omnipath [26], and ChEMBL [2]. Fig. 2, left, plots the weights of relations extracted from unstructured text between genes and diseases. It demonstrates that the weights have an inherent consistency despite uncorrelated random initial conditions (Pearson's $r = 0.9$). Further, this value can be examined as a measure of edge correctness. First, blinded manual evaluation of edges shows that low-weighted edges are three times more likely to be erroneous than high-weighted ones (Fig. 2centre). Second, edge weights are compared with their corresponding confidence scores in the Open Targets platform [13], if present, and, remarkably, GCNN weights are predictive of this score. As Fig. 2, right shows, a low-weighted edge, with score below 0.1, is 4 times more likely to be a low-scoring Open Targets edge than a high-weighted one with score above 0.9 ($p = 6 \times 10^{-28}$, two-sided KS test). This suggest that edge weight is indeed indicative of trustworthiness.

Link weights in a GCNN with attention enable the visualization of the most and least important factors underlying a representation. The left panel of Fig. 3 visualizes the connection weights for the disease cystic fibrosis (CF). The strongest drivers of CF's representation are drugs listed as a treatment for CF in curated databases, supporting their relevance to the task of predicting therapies. Two of the top six are drugs specific to CF management (Denufosol and Ivacaftor), while the other four are antibiotic drugs often used to manage infections arising in CF [28]. The six lowest-weighted links, on the other hand, consist of links arising from false extractions or weak scientific evidence [13, 5, 14]. Furthermore, connections with ABCC6 and ABCA12 extracted from text are actually false: cystic fibrosis (CF) and these genes are merely mentioned in the same sentence as part of a list but with no functional connection. The attention weight can illuminate edges to be rectified.

In addition to quality assessment, an edge's effect on the likelihood of a link can be examined by altering $\hat{C}_{r,i,j}$, because an edge can be removed after training in GCNNs. This method bears similarity to feature occlusion methods used for interpretability in high-dimensional CNNs [1, 32]. An example is shown in Fig. 3, in which all inputs to both the left and right of a relation are independently

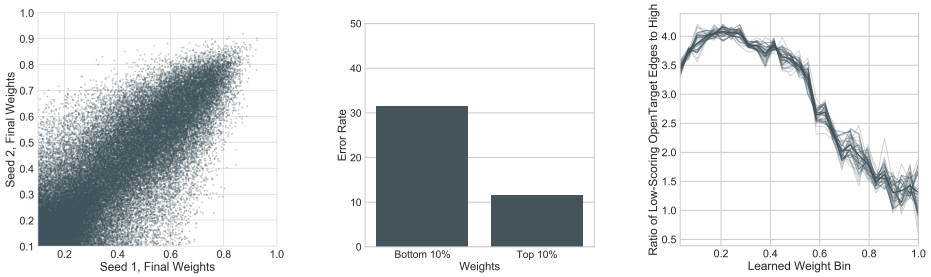

Figure 2: **Left**: Self-similarity of weight edges from two random initializations on a drug, disease, and gene dataset. Despite different initial conditions, most attention weights end training at similar magnitudes and thus lie along the diagonal. **Centre**: Rate of errors in grounding or relation extraction, when examining the top 10% and bottom 10% of weights. **Right**: Ratio of low-scoring Open Targets [13] edges to high; 5 runs with bootstrap sampling lines to show mean and variance.

removed while measuring the effect on the score of a therapeutic relation between the gene PPARG and the disease Rheumatoid Arthritis (RA) – a true edge that has been hidden during training. The top positive driver, a coexpression edge between PPARG and E2F4, uses a target in the same family as a gene implicated in RA [31]. The strongest negative driver, on the other hand, is a therapeutic link between RA and PPP3CC, a gene target associated with the rather different disease schizophrenia [15]. Future work could examine transductive methods of this analysis.

Finally, the usefulness of entire data sources can be assessed as a whole. The right panel of Fig. 3 shows the distributions of learned weights for each relation type. The edge histogram in light grey corresponds to edge types that were trained upon (and therefore typically receive a higher learned weight), while the dark grey histograms correspond to edges present only in the adjacency matrix. Patterns of weight distribution separate data sources. For example, r_1 and r_2 contain proportionally more high-weighted edges than the three below, implying that these data sources are comparatively more useful for the model. Such information enables identification of good relations and sources.

4 Conclusion

This work introduces an improvement to graph convolutional neural networks by adding an attention parameter for the network to learn how much to trust an edge during training. As a result, noisier, cheaper data can be effectively leveraged for more accurate predictions. Further, this facilitates new methods for visualization and interpretation, including ranking the influencers of a node, inferring the greatest drivers of a link prediction, and uncovering errors present in input data sources.

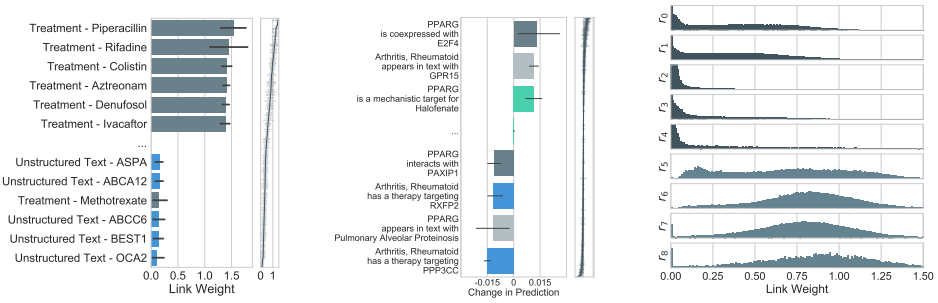

Figure 3: **Left**: Ranking of a node's influencers. The top 6 and bottom 6 known weighted-edges (+/- standard error) for cystic fibrosis are visualized as an example. **Centre**: Analyzing the drivers of link prediction, evaluating the possibility of PPARG being a drug target for Rheumatoid Arthritis. Each bar demonstrates the effect that fact has on prediction score (+/- standard error). **Right**: Distribution of edge weights across $r \in \mathcal{R}$ in the biomedical knowledge graph.

References

[1] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018.

[2] A. Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J. Bellis, Jon Chambers, Mark Davies, Felix A. Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, Michal Nowotka, George Papadatos, Rita Santos, and John P. Overington. The chembl bioactivity database: an update. *Nucleic Acids Research*, 42(D1):D1083–D1090, 2014.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

[4] Andrew Chatr-aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby-Joe Breitkreutz, Kara Dolinski, and Mike Tyers. The biogrid interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379, 2017.

[5] Alessandra Colaianni, Subhashini Chandrasekharan, and Robert Cook-Deegan. Impact of gene patents and licensing practices on access to genetic testing and carrier screening for Tay-Sachs and Canavan disease. *Genetics In Medicine*, 12(1s):S5–S14, April 2010.

[6] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *arXiv preprint arXiv:1711.05851*, 2017.

[7] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Benjamin L King, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Research*, 45(D1):D972–D978, January 2017.

[8] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. *arXiv preprint arXiv:1707.01476*, 2017.

[9] Lise Getoor and Ben Taskar. *Introduction to statistical relational learning*. MIT press, 2007.

[10] I Ispolatov, P L Krapivsky, and A Yuryev. Duplication-divergence model of protein interaction network. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 71(6 Pt 1):061911, June 2005.

[11] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, January 2017.

[12] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[13] G. Koscielny et al. Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research*, 45(D1):D985–D994, January 2017.

[14] Omar Lateef, Najia Shakoor, and Robert A Balk. Methotrexate pulmonary toxicity. *Expert Opinion on Drug Safety*, 4(4):723–730, 2005.

[15] MD) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore. Online Mendelian Inheritance in Man, OMIM, 2018.

[16] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[17] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. A novel embedding model for knowledge base completion based on convolutional neural network. *arXiv preprint arXiv:1712.02121*, 2017.

[18] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[19] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, volume 11, pages 809–816, 2011.

[20] Jay Pujara, Eriq Augustine, and Lise Getoor. Sparsity and noise: Where knowledge graph embeddings fall short. In *EMNLP*, pages 1751–1756, 2017.

[21] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*, 2017.

[22] Michael P H Stumpf, Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U. S. A.*, 105(19):6959–6964, May 2008.

[23] Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, 2015.

[24] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, 2015.

[25] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080, 2016.

[26] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature Methods*, 13(12):966–967, November 2016.

[27] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[28] "Davis S. Wishart and others". Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.

[29] Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. *arXiv preprint arXiv:1707.06690*, 2017.

[30] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[31] Rui Zhang, Lin Wang, Ji-hong Pan, and Jinxiang Han. A critical role of E2F transcription factor 2 in proinflammatory cytokines-dependent proliferation and invasiveness of fibroblast-like synoviocytes in rheumatoid Arthritis. *Scientific Reports*, 8(1):2623, February 2018.

[32] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

[33] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *arXiv preprint arXiv:1802.00543*, 2018.

Supplementary Material

S1 Synthetic Data Experiments: Duplication-Divergence Model

We also examined the specific advantages our model confers on synthetic data in a controlled setting, allowing perfect identification of true and false edges. To this end, we employ the Duplication-Divergence model, as commonly used to model protein-protein interaction networks [10].

The Duplication-Divergence (DD) Model [10] is a two-parameter model $p, q \in [0, 1]$ that mimics the growth and evolution of protein-protein interaction networks. Given a small starting seed graph $\mathcal{G}_0$ with genes as nodes and edges representing an interaction between their encoded proteins, the model simulates the process of gene duplication and mutation. To execute this model, first uniformly select a node n_{old} in $\mathcal{G}_0$ and define a new node n_{new}. Next, connect n_{new} and n_{old} with probability q and connect n_{new} with any other node linked to node n_{old}, each with probability p. This results in a network $\mathcal{G}_1$, and the process is repeated until the network grows to a specified size.

Let $\mathcal{G}_0$ be a graph of two connected nodes. We define a 1000-node graph $\mathcal{G}_{1000}$ with parameters $p = 0.75, q = 0.0$. The value of p and q are chosen to ensure that each of the sub-sampled training

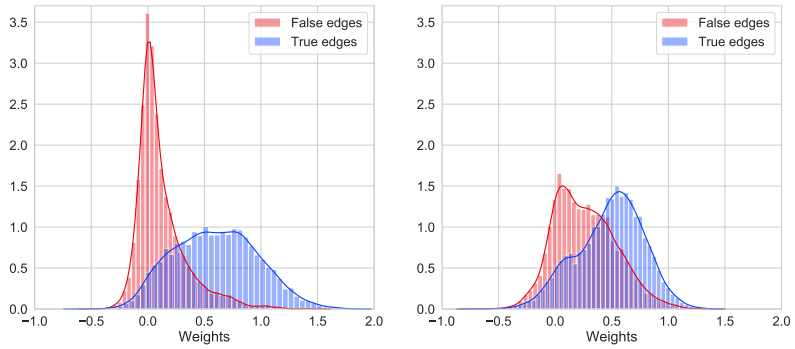

Figure S1: **Left**: Normalized histogram of link attention weights for edges in $\mathcal{G}_{\text{Add}}$ (blue) and $\mathcal{G}_{\text{Noise}}$ (red) when training only on $\mathcal{G}_{\text{Gold}}$. **Right**: Normalized histogram of link attention weights when training on edges in $\mathcal{G}_{\text{Gold}} \cup \mathcal{G}_{\text{Add}} \cup \mathcal{G}_{\text{Noise}}$.

graphs has a single connected component with edge-vertex ratio ~ 20, reflecting the estimated size of the human interactome [22].

This setup recreates the condition of an incomplete but high-precision dataset ($\mathcal{G}_{\text{Gold}}$) with which to train a model for link-prediction, together with the option of supplementing the training data with a set of possibly noisy relations ($\mathcal{G}_{\text{Add}} \cup \mathcal{G}_{\text{Noise}}$). The model here is a single-layer GCNN, with embedding size 300, and a diagonalized weight matrix W_r is chosen with no non-linearity transformation applied. Fig.S1, left two plots, demonstrate that a low attention weight can imply a faulty edge, and that the identification of false edges works best when only training on high-quality edges ($\mathcal{G}_{\text{Gold}}$) while the adjacency matrix $A_{r,i,j}$ can include edges from a noiser set ($\mathcal{G}_{\text{Add}} \cup \mathcal{G}_{\text{Noise}}$).