Bioinformatics
doi.org/10.1093/bioinformatics/btaa1036

OXFORD

Data and text mining

On Biases of Attention in Scientific Discovery

Uriel Singer [1], Kira Radinsky [1] and Eric Horvitz [2,3]

[1]Department of Computer Science, Technion—Israel Institute of Technology, Haifa 3200003, Israel
[2]Adaptive Systems and Interaction Group, Microsoft Research, Redmond WA 98052 USA
[3]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle WA 98195 USA

Abstract

Summary: How do nuances of scientists' attention influence what they discover? We pursue an understanding of the influences of patterns of attention on discovery with a case study about confirmations of protein-protein interactions over time. We find that modeling and accounting for attention can help us to recognize and interpret biases in large-scale and widely used databases of confirmed interactions and to better understand missing data and unknowns. Additionally, we present an analysis of how awareness of patterns of attention and use of debiasing techniques can foster earlier discoveries.
Availability: The data is freely available at https://github.com/urielsinger/PPI-unbias.
Contacts: urielsinger@cs.technion.ac.il, kirar@cs.technion.ac.il, horvitz@microsoft.com

1 Introduction

Indian Philosopher Jiddu Krishnamurti has said, "the finding is not in the future–it is there, where you do not look." (Krishnamurti, 2018). In the spirit of Krishnamurti's reflection, we investigate how patterns of attention in scientific discovery can influence the state of knowledge in a discipline in ways that may not be recognized. Systematic biases of attention in the formulation and confirmation of hypotheses by investigators is especially important for understanding the nature and limitations of knowledge encoded in large, general-purpose databases that see wide use as general tools. In biology and other fields, large-scale multi-use databases of findings have become the lenses on what is known. Patterns of attention, guided by the flow of interests and curiosities, and enabled by available experimental methods and laboratory set-ups, can introduce systematic biases in databases of findings. We focus on the illustrative example of the growing fund of knowledge on interactions among proteins. We have found via the analysis of graphical representations of the flow of confirmations, that patterns of exploration and confirmation have become embedded and implicit in the protein-protein interaction database.

The protein-protein interactions (PPI) database is a uniquely rich artifact for studying biases of attention and influences on discovery per its importance, size (Zhu *et al.*, 2007), and the accessibility of its temporal evolution. Decades of research have yielded large databases of protein-protein interactions. These databases have played a critical role in biomedicine, enabling the construction of biochemical cascades and larger protein interaction networks. The interaction data and resulting representations of metabolic, structural, and regulatory processes have been critical in understanding the etiologies of diseases and in identifying promising therapies, including efforts to prioritize pharmacological targets (Monod *et al.*, 1965; Krogan *et al.*, 2006; Prelich *et al.*, 1987; Collins *et al.*, 2007; LaCount *et al.*, 2005; Pu *et al.*, 2007; Komurov and White, 2007; Strong and Eisenberg, 2007; Wells and McClendon, 2007).

In Figure 1, we display a network representation of the evolution of confirmed interactions among Homo sapiens proteins over time, where nodes are proteins and arcs represent confirmed interactions. We introduce analysis over snapshots of confirmed PPI over time that reveal that discoveries about protein interactions are rooted in scientists' attention to recent findings. Such biases may be rooted in several factors, including the sequencing of attention to specific sets of biochemical pathways of interest, and pursuit of understanding of these systems via PPI testing when one or more proteins are already known to be interacting with one another.

For any two proteins p_i, p_j, we represent the probability of a confirmed interaction over a period of time as $P(p_i \leftrightarrow p_j)$. This probability of interaction can be expressed as a chain of two probabilities: (1.A) the probability that the proteins will be found to interact, given the experiment is carried out, and, (1.B) the probability of an experiment being performed to check the interaction during the period. Thus, taking both probabilities into consideration, the likelihood of an interaction being confirmed can be rewritten as follows:

$$P(p_i \leftrightarrow p_j) = \overbrace{P(p_i \leftrightarrow p_j | Check\ p_i, p_j)}^{\text{(1.A)}} \cdot \overbrace{P(Check\ p_i, p_j)}^{\text{(1.B)}} \quad (1)$$

Our knowledge of protein interactions is constrained by the keyhole of sets of decisions made over time to perform experiments. Computing $P(p_i \leftrightarrow p_j)$ requires consideration of the probability that a protein-protein interaction is examined, $P(Check\ p_i, p_j)$. This probability can be further divided into two more probabilities: (2.A) the probability that

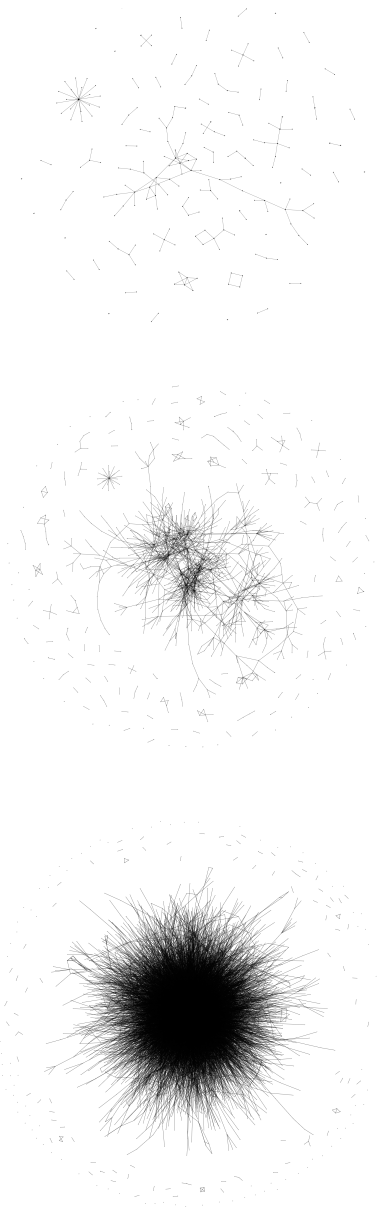

Fig. 1. Homo sapiens protein-protein interactions represented as a graph where nodes are proteins and edges are interactions. The data was drawn from (López et al., 2015) as explained in Section 4.1. From top to bottom, expansion of the protein-protein interaction graph over the years 1990, 1995, and 2005.

scientists are interested in performing a specific experiment to validate or invalidate a hypothesis about an interaction and (2.B) the probability that they have the required scientific tools, experimental resources, and affordances:

$$P\big(Check\ p_i, p_j\big) = \overbrace{P\big(Attention\ p_i, p_j\big)}^{\text{(2.A)}} \cdot \overbrace{P\big(Tools\ p_i, p_j\big)}^{\text{(2.B)}} \quad (2)$$

Decomposing the overall likelihood of a hypothesis into a chain of probabilities can provide a lens on the current state of knowledge, the influence of enabling tools, and on recurrent patterns of attention in different areas of science. We find that $P\big(Check\ p_i, p_j\big)$ captures, via patterns of confirmation of protein interactions over time in the PPI database, the influences of attention on discovery for interactions. The influences of biases of attention are typically implicit, and are not overtly considered in interpretations of confirmed, invalidated, or unknown protein interactions. We suspect that findings about systematic biases introduced by such attentional considerations may generalize to systematic patterns of confirmations and unknowns encoded in other large-scale scientific databases. We focus in this work on interactions among proteins in homo sapiens. While the human genome encodes approximately 30,000 proteins (Venter *et al.*, 2001), defining a space of nearly a half billion potential interactions, only $\sim 300,000$ interactions for $\sim 17,000$ proteins have been confirmed to date. To map the influence of attention, we consider the database of all known protein interactions at the end of each calendar year. For each year, we represent the protein-interaction database as a topological graph where nodes stand for proteins and edges for interactions that are confirmed by the end of the respective period. Within each annual PPI graph, we define the *protein distance* as the length of the shortest chain of nodes connecting pairs of proteins that interact (see Section 4.1 for more details on the database creation).

We have found that new discoveries about protein interactions in a consecutive year are highly skewed towards protein interactions with small distances in the PPI graph for the current year. Figure 2 describes the distribution of *protein distance* a year before interaction discovery, normalized by the distribution of all possible edge distances, creating a probability graph. We had first discovered this phenomenon during our unpublished precursory work centering on the use of machine learning from a database of confirmed interactions to predict future interactions. During these investigations, we were surprised and intrigued to discover that the variable with the greatest evidential power to predict the next year's discoveries was the proximity of proteins in the PPI graph. This phenomenon has been corroborated by other studies (Han *et al.*, 2005; Tanaka *et al.*, 2005; Lima-Mendez and van Helden, 2009; Fraser and Hirsh, 2004; Saeed and Deane, 2006).

We, nor expert colleagues we consulted with, could identify a biological explanation to explain the skew of discoveries based on adjacencies. We hypothesized that the observations of protein interactions are founded in the focus of attention of scientists who continue to explore certain proteins and processes that have been most recently studied. In the absence of such attentional influences, we would expect a more uniform distribution in distances among proteins that are found to interact in the next year. While the phenomenon of new interactions being linked to recent findings may not be unexpected, such biases on discovery, persisting in a widespread manner over long periods of time, can have significant influences on confirmed, disconfirmed, and unexplored hypotheses in a large database. We set out to characterize such attentional effects and their influences on the sequencing of PPI discoveries. We believe that unearthing and characterizing such biases in research on protein interactions will be valuable for interpreting distributions of knowns and unknowns at different points in time with the refinement of knowledge that comes with ongoing experimentation.

2 Characterizing and Countering Attentional Influences

We now pursue the means for recognizing, characterizing and countering the influence of a bias of attention in discoveries of protein interactions. We study the effect of proximity in the PPI graph on the inferred probability of future discoveries by employing statistical models that predict interactions between untested pairs of proteins—should they be studied. We first explore the ability of these predictive models to forecast future discoveries when they are trained on data of previously confirmed protein interactions. Next, we build and consider uses of predictive models that do not rely on graph proximity and, thus, are less influenced by biases of attention

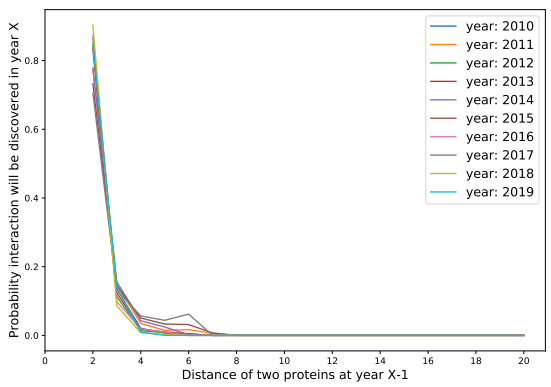

Fig. 2. Probability that a protein interaction will be confirmed in the next year as a function of the distance between two proteins in the PPI graph in a current year. Each color represents a different year.

encoded via proximity. We consider hypothetical uses of these unbiased models to guide decisions about experiments and confirmations on protein interactions.

By definition, proteins at distance $d = 1$ are confirmed as interacting. We seek to compare the growth in knowledge about previously unconfirmed interactions between proteins at $d = 2$ versus proteins that are separated by greater distances in the topological graph. Proteins confirmed at $d = 2$ are those that are nearest in proximity to those proteins participating in a currently known interaction pair. The comparison of growth in knowledge about protein interactions at $d = 2$ versus greater distances can reveal the potential for alternate sequencing of discoveries of protein interactions where attentional biases are not at play. We perform this analysis via a counterfactual study based on matching (Ho *et al.*, 2007).

We seek to understand how the protein interaction database might grow differently if the influences of proximity in the protein interaction graph were removed or minimized. To do this, we build statistical models that can predict protein interactions solely from attributes or *features* of proteins. We consider two sets of features and build a predictive model for each. The first approach is based on biological properties of proteins. We apply an embedding technique inspired by Dubchak *et al.* (1995) where we generate a vector for each protein by considering the protein's sequence of amino acids. We build the vectors by concatenating three types of descriptors, including composition (C), transition (T), and distribution (D) attributes (see Section 4.2 for details on the biological features). In a second approach to predicting protein interactions, we calculate a feature vector for each protein by using a neural graph embedding learning method (see Section 4.3 for more details). This approach is of special interest due to the graph topology that is shaped in part by attentional influences as captured in the sequence of confirmed interactions. For both approaches, the feature vector of an edge linking two proteins is calculated as the absolute differences of the protein feature vector of each of the linked proteins. Following the computing of vectors for each edge, a matching analysis is applied on edges with distance $d > 2$ ("original edges") to edges with distance $d = 2$ ("matched edges"), by selecting the specific set of features and distance metric. Using the matching method enables us to answer the question: "What if the *protein distance* was 2 instead of n?". Answering such a question can help us interpret the causal effect of the *protein distance* and normalize the skew shown in Figure 2 (see Section 4.4 for more details).

Based on the graph embedding, we train a logistic regression model on edges included in the PPI graph in the year 2017, while false edges are sampled at random. We use the trained model to infer the probability that an edge will appear as an interaction in the following period, 2018–2019. We compare four sets of prediction methods:

distance	Biased	Unbiased - graph embedding	Unbiased - amino acid	Random
2	0.664	0.664	0.664	0.506
3	0.693	**0.720**	0.482	0.503
4	0.730	**0.752**	0.464	0.516
$5 \leq$	0.830	**0.851**	0.473	0.396
all	0.797	**0.825**	0.564	0.503

Table 1. AUC results for different distances in the protein interaction graph for each method. Boldfaced results indicate a statistically significant difference.

1. *Biased*: Predictions considering the original edges in the PPI graph, without applying the matching analysis.
2. *Unbiased - graph embedding*: Prediction on matched edges via graph embedding with a Euclidean distance metric (see Section 4.5.1 for details on the Euclidean distance metric).
3. *Unbiased - amino acid*: Prediction on matched edges via biological features using a Canberra distance metric (see Section 4.5.2 for details on the Canberra distance metric).
4. *Random*: Prediction on random matched edges. Matching is applied between the original edges at $d > 2$ to edges at $d = 2$ by random sampling. The random method is necessary in order to eliminate false insights from the data.

We build four predictive models, employing each of the different methods and we test the power of each to predict future discoveries of interactions. As a performance metric, we report the area under the receiver-operator characteristic curve (AUC) of each model.

In Table 1, we show the AUC for predicting the interactions discovered in 2018–2019 for each distance and for each type of prediction method. We find that using graph-based matching as a predictive tool significantly improves the forecast of new discoveries. We note that, in the most recent years, interactions appear to be discovered via more advanced techniques and tools, resulting in a weakening of the influence of attention-centric biases in the test set as compared to earlier years. We train the predictive model on a biased training set and predict on a less biased test set. We hypothesize that checking interactions is less dependant on having the right tools or laboratory setup in later years. Assuming weak dependence on tools, and recalling Equations 2 and 1, the probability that an interaction between two proteins will be confirmed becomes dependent mainly on the attention of scientists on those specific proteins. By using graph-based matching, we strive to debias the model by normalizing the attention to potential interactions. The new techniques and tools still show bias, indicating that the graph-based matching of today's known PPI could provide guidance on studies aimed at identifying protein interactions. We attempt to weaken biases of attention by applying graph-based matching rather than use the original 'biased' edges of 2017.

While our focus has been to elucidate attentional influences on the sequence of PPIs, we foresee that such predictive models could provide exploratory directions, with promise for helping scientists to recognize and address attentional biases of proximity.

3 Defining Bias of Attention

We can further define attentional bias by using the unbiased graph embedding method from Section 2 and comparing the individual edge scores between the original and matched edges. This method starts with matching all edges at distance $d = n$ to their closest matches at distance $d = 2$. After matching, we calculate an individual treatment effect (ITE),

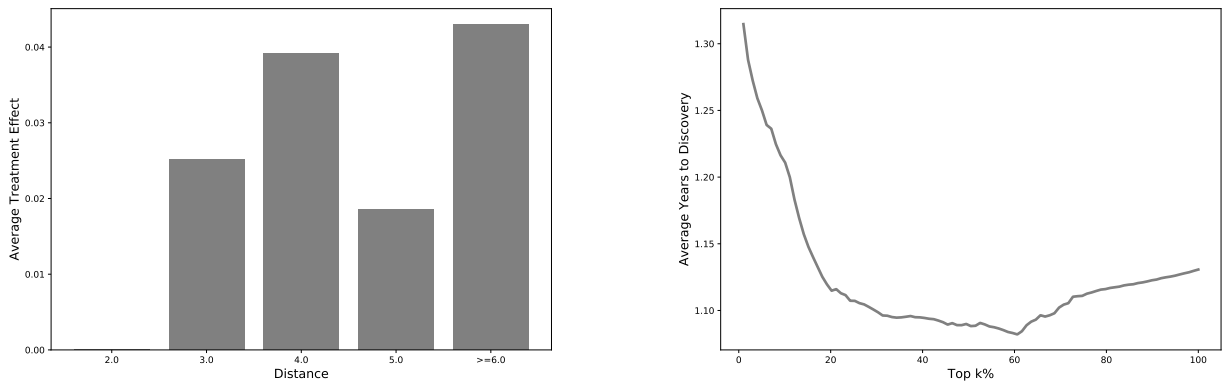

Fig. 3. Average number of years to discovery for the top k% score difference.

defined as the difference between the original edge score and the matched edge score for the test period 2018–2019: $y_i(P_{2(i)} - P_{n(i)})$. In each potential edge (i), $P_{n(i)}$ is the probability of the original edge being discovered in later years, $P_{2(i)}$ is the matched edge probability, and y_i is the true label of the original edge, where $y_i = 1$ if an edge forms an interaction between the proteins by the end of year 2019, and $y_i = -1$ if not.

The value of $y_i(P_{2(i)} - P_{n(i)})$ varies between -1 and 1, where -1 is assigned if an edge is predicted correctly between the original edge at $d > n$ but incorrectly for the matched edge at $d = 2$. Alternatively, 1 is assigned if an edge is predicted correctly between the matched edge but incorrectly for the original edge, and 0 if there is no difference between the probabilities. We then calculate the average treatment effect (ATE) for each distance:

$$ATE = \frac{1}{N}\sum_{i=1}^{N} y_i(P_{2(i)} - P_{n(i)})$$

Observing $ATE > 0$ means that the matched edges are more successful in predicting interactions than the original edges. The closer the value of ATE is to 1, the better the predictions of the matched edges over the original edges. Similarly, the closer the value of ATE is to -1, the poorer the predictions of the matched edges will be over the original edges. We propose that the ATE provides a valuable characterization of attentional bias for the discovery of protein interactions.

As displayed in Figure 3, the matched edges perform better than the original edges at $d > n$. We observe that the larger the distance, the larger the effect of the matching analysis. The only exception is distance 5, which we attribute to the small number of samples (see Section 4.6 for analysis of additional years).

As a final analysis, we seek to understand whether the use of unbiasing method techniques could have led to earlier discoveries. We continue using the same notations, $P_{n(i)}$ as the probability of the original edge being discovered in later years, and $P_{2(i)}$ as the matched edge probability using the graph embedding method. We further define $t_{n(i)}$ as the year that the original edge was discovered minus 2017, and $P_{n(i)} - P_{2(i)}$ as the score difference of the sample i. Given $k \in [0, 100]$, we average over $t_{n(i)}$ for samples with a top k% score difference. We display in Figure 3 the finding that, for smaller values of k, we would tend to discover an interaction at earlier times.

4 Methods

4.1 Data

The protein interaction information used in the study was drawn from HitPredict (López *et al.*, 2015). HitPredict contains a list of all protein interactions, where each interaction includes a list of all published articles where it is mentioned. We perform a preprocessing phase, where we filter to retrieve studies with H.sapiens proteins and then identify the interaction discovery date as the publication date of the earliest associated article.
Once established, we create a topological graph that includes all interactions with their first publication dates. By using this method, we construct a graph that evolves over time. Protein interactions with no associated articles are not included in the graph. In the experiments, the PPI graph of year y includes all interactions among proteins that are confirmed by the end of year y.

4.2 Biological Features.

The protein amino acid code used in the study was drawn from Ensembl (Zerbino *et al.*, 2018). For each protein, three different calculations were applied to its amino acid code, and then concatenated to create one biological feature vector:

1. ***Composition***: We calculate the probability of appearance for each attribute in the amino acid code, and create a list of the probabilities for each possible attribute.

$$C_i = \frac{\sum_k CODE_k = i}{L}$$

were $CODE_k$ is the kth attribute in the code, L is the code length, and C_i is the probability of appearance for attribute i.
2. ***Transition***: We calculate the probability of a consecutive appearance for each couple of attributes in the amino acid code.

$$T_{i,j} = \frac{\sum_k CODE_k = i \wedge CODE_{k+1} = j}{L}$$

where $CODE_k$ is the kth attribute in the code, L is the code length, and $T_{i,j}$ is the probability of a consecutive appearance of i and j. After calculating $T_{i,j}$, we flatten the upper triangle of the two-dimensional array.
3. ***Distribution***: The distribution of an attribute along the sequence of amino acids is described by a vector of size five. The values of the

distribution are the locations in the amino acid code for the first, 25%, 50%, 75%, and 100% appearances of the given attribute. For example, given the code ABBABBBBBAAAABBBABABBBBBAA of length 26, and attribute 'B', which appears 16 times in the code:

- The first appearance is at index 2: $2/26 = 7.7\%$
- The $16 \cdot 25\% = 4$ appearance is at index 6: $6/26 = 23.1\%$
- The $16 \cdot 50\% = 8$ appearance is at index 14: $14/26 = 53.8\%$
- The $16 \cdot 75\% = 12$ appearance is at index 20: $20/26 = 76.9\%$
- The last appearance is at index 24: $24/26 = 92.3\%$

Therefore, the distribution of the attribute 'B' will be: [7.7, 23.1, 53.8, 76.9, 92.3]. We do the same for each attribute and concatenate the results.

4.3 Node2vec

We chose to work with node2vec as our graph embedding methodology, an approach introduced by Grover and Leskovec (2016) with state-of-the-art performance on multiple benchmarks including predicting PPIs. Much research has since been done using node2vec to predict protein interactions (Grover and Leskovec, 2016; Singer *et al.*, 2019; Yue *et al.*, 2020; Zhong and Rajapakse, 2019; Ata *et al.*, 2018; Zhang *et al.*, 2019; Goyal and Ferrara, 2018; Ma *et al.*, 2018). Node2vec is based on word2vec (Mikolov *et al.*, 2013), which is a method for learning features of a representation of words. Word2vec takes as input a text corpus and outputs an embedding vector for each word. By trying to predict words' neighbors, word2vec creates an embedding that provides a representation of semantic relationships among words. Node2vec generalizes word2vec for the graph domain where, intuitively, each node is regarded as a word. The algorithm creates the equivalent of sentences by performing random walks on the graph starting at each node (i.e., each node that is sampled in the random walk is a word in the constructed sentence).

One of the key contributions of node2vec is the generalization that differentiates between space and structure. In other words, one can select whether to embed a node based on other nodes that are closer in space (i.e., same cluster) or based on nodes with a similar role in the structured graph. From a graph algorithm perspective, this can be explained as selecting whether to perform random walks with a breadth-first or depth-first search bias. Given a random walk from node u to node v, node2vec formulates this bias strategy by defining two hyperparameters, p and q, which help to adjust the transition probability $\alpha_{pq}(u, x)$ from node u to some node x, where d_{ux} stands for the distance between node u and node x:

$$\alpha_{pq}(u,x) = \begin{cases} \frac{1}{p} & if \quad d_{ux} = 0 \\ 1 & if \quad d_{ux} = 1 \\ \frac{1}{q} & if \quad d_{ux} = 2 \end{cases}$$

In this way, node2vec can bias the random walk closer or further away from the source node, creating different types of embeddings. For example, setting $p < q$ biases the random walk to nodes closer to one another. This in turn causes nodes from the same cluster to be embedded closer and nodes from different regions to be embedded further away. Setting $p > q$, biases the random walk to embed nodes of the same graph characteristics (e.g., same role in a social graph) closer together while others are embedded further away. We note that in the special case of $p = q = 1$, the algorithm operates in a similar manner to DeepWalk (Perozzi *et al.*, 2014). We believe that representing proteins in the PPI graph as node2vec vectors takes advantage of more comprehensive structural features that can boost a model's ability to predict interactions (we used the implementation published by the authors: github.com/aditya-grover/node2vec).

4.4 Causal Inference and Matched Sets

To provide intuition about the matching methods, we cast the method in a medical context. Assume we have two patients, where patient A receives treatment *a* and patient B receives treatment *b* and outcomes are recorded. We wish to understand what would have happened if patient A had received treatment b rather than treatment a). As we cannot change history, to perform the counterfactual treatment of *b*, we would have to find patient A's twin or doppelganger. An analogy between the health domain and the PPI domain can be constructed by looking at potential edges as patients, the treatments as the distances between proteins in the protein interaction graph, and the outcomes as interactions confirmed in the future. We apply the matching techniques to the PPI domain and answer the question raised at the beginning of this section. Ho *et al.* (2007) demonstrated the power of matching. They describe two groups that differ in the treatments they received, and for each individual in the first group find its closest match in the second group. By identifying these matches, we are able to estimate the ITE as the difference in the value of the outcome of the original treatment minus that of the outcome associated with the matched twin. In order to compute the ATE, we average the ITEs over all individuals. We can apply matching to the PPI domain by finding for each edge with distance $d > 2$ its match with distance $d = 2$ and calculate the distance effect.

4.5 Distance metrics

4.5.1 Euclidean

Euclidean distance is a classic and widely used metric calculated as the distance between two points in the Euclidean n-space. The Euclidean distance is a generalization of the Pythagorean theorem. If x and y are vectors, their Euclidean distance d is defined as:

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

4.5.2 Canberra

If x and y are vectors, their Canberra distance d is defined as:

$$d(x,y) = \sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Notice that each element is normalized by itself. This functionality is important when pursuing an equal value for each element in the vector regardless of its scale.

4.6 Average Treatment Effect for Additional Years

We performed ATE analyses for different years. We trained a logistic regression model on edges included in the PPI graph up to a specific year while testing on edges from the following year, up to the year 2019. In Figure 4, we observe the ATE for the years 2015 and 2016. Beyond an anomaly for a distance of four, we observe the same correlation where the larger the distance, the larger the effect of matching analysis.

5 Conclusion

We examined the sequence of confirmations of protein-protein interactions over time and found evidence of an attentional bias: Scientists tend to focus the formulation and confirmation of hypotheses in the direction of recently confirmed PPIs. The bias has shaped the progression of knowledge about protein interactions, represented as the join of findings available at different times in the widely used PPI database. The systematic biases, based in the natural pursuit of hypotheses and the ongoing evolution of experimental methods, have influenced the completeness of data and the patterns of "dark matter" of undiscovered findings. We believe that awareness of attentional

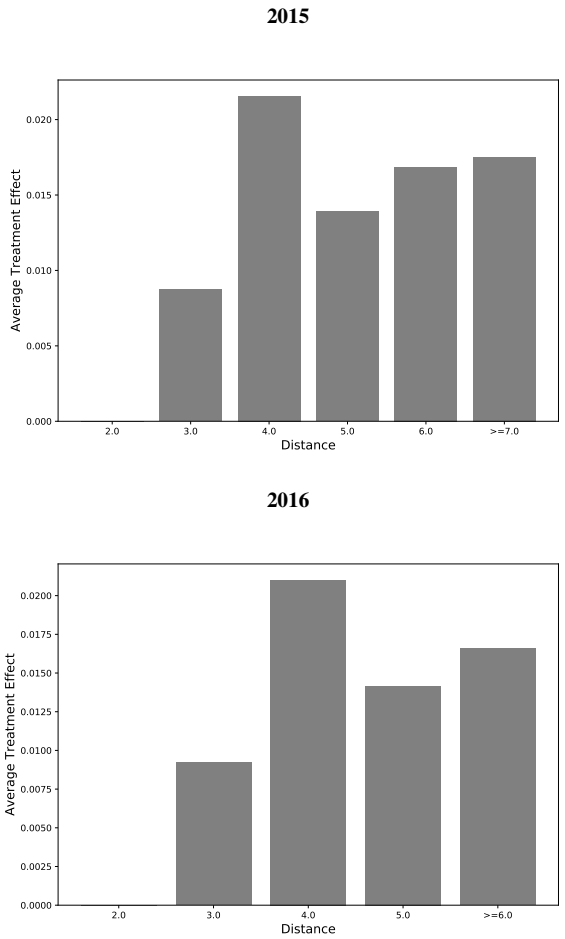

Fig. 4. Average treatment effect (ATE) as a function of the distance of matched edges. For each year, the ATE is calculated up to the maximum distance in the graph ('$\geq$' represents the distance, including the infinite distance—proteins from different connected components in the PPI graph).

biases in discovery will help us to better understand limitations in the PPI dataset and other databases. Biases of attention based on temporal and conceptual proximity highlight the potential value of adding to research portfolios the formulation and pursuit of hypotheses that make more distant leaps in conceptual spaces. We believe that biases of locality can be addressed by nurturing research practices that promote exploration of more distant conceptual relationships and of making serendipitous discoveries (Board, 2018) beyond the frontier of current knowledge. Opportunities for future work include studies of biases of attention in other biological domains such as pursuing understandings of interactions between homo sapien proteins and proteins expressed by viruses (Lasso *et al.*, 2019). The findings for the PPI database frame questions about other systematic biases of exploration, confirmation, and discovery. We suspect that similar attentional factors have influenced the content of multiple widely used databases. We hope this study will stimulate efforts to identify systematic biases of attention in other areas. Characterization of attentional biases can provide valuable insights about the structure and implications of unexplored hypotheses and missing data across the sciences.

References

Ata, S. K. *et al.* (2018). Integrating node embeddings and biological annotations for genes to predict disease-gene associations. *BMC systems biology*, **12**(9), 138.

Board, N. E. (2018). The serendipity test. *Nature*, **554**(7690), 5.

Collins, S. R. *et al.* (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **446**(7137), 806–810.

Dubchak, I. *et al.* (1995). Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences*, **92**(19), 8700–8704.

Fraser, H. B. and Hirsh, A. E. (2004). Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC evolutionary biology*, **4**(1), 13.

Goyal, P. and Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, **151**, 78–94.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, San Francisco, California, USA. ACM.

Han, J.-D. J. *et al.* (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nature biotechnology*, **23**(7), 839–844.

Ho, D. E. *et al.* (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, **15**(3), 199–236.

Komurov, K. and White, M. (2007). Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Molecular systems biology*, **3**(1).

Krishnamurti, J. (2018). The pitcher can never be filled, the bulletin of the krishnamurti trust 22, 1974. In *Meeting Life: On Finding Your Path Without Retreating from Society*.

Krogan, N. J. *et al.* (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**(7084), 637–643.

LaCount, D. J. *et al.* (2005). A protein interaction network of the malaria parasite plasmodium falciparum. *Nature*, **438**(7064), 103–107.

Lasso, G. *et al.* (2019). A structure-informed atlas of human-virus interactions. *Cell*, **178**(6), 1526–1541.

Lima-Mendez, G. and van Helden, J. (2009). The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, **5**(12), 1482–1493.

López, Y. *et al.* (2015). Hitpredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species. *Database*, **2015**.

Ma, J. *et al.* (2018). Depthlgp: Learning embeddings of out-of-sample nodes in dynamic networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA.

Mikolov, T. *et al.* (2013). Efficient estimation of word representations in vector space. *ICLR*, **abs/1301.3781**.

Monod, J. *et al.* (1965). On the nature of allosteric transitions: a plausible model. *J Mol Biol*, **12**(1), 88–118.

Perozzi, B. *et al.* (2014). Deepwalk: Online learning of social representations. In *Proc. of KDD*, pages 701–710, New York City, USA.

Prelich, G. *et al.* (1987). Functional identity of proliferating cell nuclear antigen and a dna polymerase-δ auxiliary protein. *Nature*, **326**(6112), 517–520.

Pu, S. *et al.* (2007). Identifying functional modules in the physical interactome of saccharomyces cerevisiae. *Proteomics*, **7**(6), 944–960.

Saeed, R. and Deane, C. M. (2006). Protein protein interactions, evolutionary rate, abundance and age. *BMC bioinformatics*, **7**(1), 128.

Singer, U. *et al.* (2019). Node embedding over temporal graphs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4605–4612, Macau, China. AAAI Press.

Strong, M. and Eisenberg, D. (2007). *The protein network as a tool for finding novel drug targets*, pages 191–215. Birkhäuser Basel, Basel.

Tanaka, R. *et al.* (2005). Some protein interaction data do not exhibit power law statistics. *FEBS letters*, **579**(23), 5140–5144.

Venter, J. C. *et al.* (2001). The sequence of the human genome. *science*, **291**(5507), 1304–1351.

Wells, J. A. and McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, **450**(7172), 1001–1009.

Yue, X. *et al.* (2020). Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, **36**(4), 1241–1251.

Zerbino, D. R. *et al.* (2018). Ensembl 2018. *Nucleic acids research*, **46**(D1), D754–D761.

Zhang, J. *et al.* (2019). Prone: fast and scalable network representation learning. In *Proc. 28th Int. Joint Conf. Artif. Intell., IJCAI*, pages 4278–4284, Macau, China.

Zhong, X. and Rajapakse, J. C. (2019). Predicting missing and spurious protein-protein interactions using graph embeddings on go annotation graph. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1828–1835, San Diego, California, USA. IEEE.

Zhu, X. *et al.* (2007). Getting connected: analysis and principles of biological networks. *Genes & development*, **21**(9), 1010–1024.