# The probability of edge existence due to node degree: a baseline for network-based predictions

*This manuscript (permalink) was automatically generated from greenelab/xswap-manuscript@2444609 on August 23, 2019.*

## Authors


- **Michael Zietz**
  0000-0003-0539-630X · zietzm · ZietzMichael
  Department of Physics & Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America · Funded by ['Roy and Diana Vagelos Scholars Program in the Molecular Life Sciences', 'the Gordon and Betty Moore Foundation (GBMF4552)']

- **Daniel S. Himmelstein**
  0000-0002-3012-7446 · dhimmel · dhimmel
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America · Funded by Pfizer Worldwide Research, Development, and Medical; the Gordon and Betty Moore Foundation (GBMF4552)

- **Kyle Kloster**
  0000-0001-5678-7197 · kkloste · kylekloster
  Department of Computer Science, North Carolina State University, Raleigh, North Carolina, United States of America · Funded by the Gordon and Betty Moore Foundation (GBMF4560)

- **Christopher Williams**
  · chrsunwil
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

- **Michael W. Nagle**
  0000-0002-4677-7582 · naglem · MikeNagle84
  Internal Medicine Research Unit, Pfizer Worldwide Research, Development, and Medical

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · greenescientist
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America · Funded by ['Pfizer Worldwide Research, Development, and Medical', 'the Gordon and Betty Moore Foundation (GBMF4552)', 'the National Institutes of Health (R01 HG010067)']

- Blair D. Sullivan
  School of Computing, University of Utah, Salt Lake City, UT, USA
  Funded by GBMF4560.


Blair 8/28-29 2019

Degree is an important metric for differentiating between nodes, and it appears in many common edge prediction features [5]. However, reliance on degree can pose problems for edge prediction. Firstly, bias in networks can distort node degree so that degree differences between two nodes may not be meaningful. Secondly, reliance on degree can lead edge prediction methods to make nonspecific or trivial predictions and fail to identify novel or insightful relationships.

word copy issue.

can you support these claims?

Most biomedical networks are imperfect representations of the true set of relationships. Real networks often mistakenly include edges that do not exist and exclude edges that do exist. How well a network represents the true relationships it attempts to represent depends on a number of factors, especially the methods used to generate the data in the network [6,7,8]. We define "degree bias" as the type of misrepresentation that occurs when the fraction of incorrectly existent/nonexistent relationships depends on node degree. Depending on the type of data being represented, degree biases can arise due to experimental methods, inspection bias, or other factors [6].

at a given node?

Inspection bias indicates that entities are not uniformly studied [9], and it is likely to cause degree bias when networks are constructed using hypothesis-driven findings extracted from the literature, as newly-discovered relationships are not randomly sampled from the set of all true relationships. Though there is a high correlation between the number of publications mentioning a gene and its degree in low-throughput interaction networks, the number of publications mentioning a gene has little correlation with its degree in a systematically-derived protein interaction network [10]. This evidence suggests that many poorly studied genes have similar numbers of interactions as those scientists have preferentially examined and that these edges are missed due to inspection bias. For networks with strong inspection bias, reliance on degree can lead to predictions that have a good metrics when assessed by cross validation but little ability to generalize.

singular/plural issue.

Another reason why a method's reliance on degree can be unfavorable is that degree imbalance can lead to prediction nonspecificity. Nonspecific predictions are made on the basis of generic characteristics rather than the specific connectivity information contained in a network. For example, Gillis et al. [11] examined the concept of prediction specificity in the context of gene function prediction and found that many predictions appear to rely primarily on multifunctionality and could be "potentially misleading with respect to causality." Real networks have a variety of degree distributions (Figure 1), and they commonly exhibit degree imbalance [1,12,2,3]. Degree imbalance leads high-degree nodes to dominate in the predictions made by degree-associated methods [13], which are effective predictors of connections in some biological networks [14]. Consequently, degree-based predictions are more likely nonspecific, meaning the same set of predictions performs well for different tasks.

ill-defined?

this isn't necessarily bad.

I feel like this point should have already been made & doesn't bear repeating in this way here.

Depending on the prediction task, edge predictions between very high degree nodes may be undesired, uninsightful, or nonspecific. Model evaluation is challenging in this context: nonspecific or trivial predictions can dominate performance evaluations and may actually be correct, even if they are not the desired outputs of the predictive model. For example, predicting that the highest degree node in a network shares edges with the remaining nodes to which it is not connected will often lead to many correct predictions, despite this prediction being generic to all other nodes in the network.

example is not "between high degree"

Degree is important in edge prediction, but it can cause undesired effects. Degree-based features should often be included in the interpretation of predictions to disentangle desired from non-desired effects and to effectively evaluate and compare predictive models. We sought to directly measure the effect of node degree on edge prediction methods. We introduce a permutation-based framework and software implementation to find edge existence probabilities due to node degree and to quantify the contribution of degree to edge prediction methods. This method allows edge predictions to be evaluated in the context of degree and its effects on the prediction task. Our results demonstrate that degree-associated methods are very effective for reconstructing a network using a subsampled holdout. However, these methods are ineffective for predicting edges between networks measuring

feels like you really want to say that prediction between high degree nodes is especially challenging.

**A**

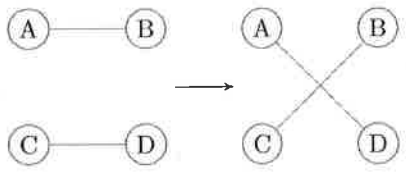
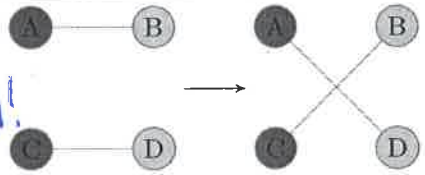
**Input:** Undirected graph G, distribution $\rho$, and number of steps $T$
**Output:** Edge-swapped graph $G_s$
**for** i = 1,...,$T$ **do**
  Select two edges $(i,j)$, $(k,l) \in E(G_s)$
  **if** $(i,l) \notin E(G_s)$ and $(k,j) \notin E(G_s)$ **then**
    $E(\widehat{G_s}) \leftarrow \left(E(G_s) \setminus \{(i,j),(k,l)\}\right) \cup \{(i,l),(k,j)\}$
    $G_s \leftarrow \widehat{G_s}$ with probability $\min(\rho(\widehat{G_s})/\rho(G_s), 1)$
  **end if**
**end for**

**B**

**Input:** Directed, undirected, or bipartite graph G, number of steps $T$, and booleans `allow_antiparallel` and `allow_loops`
**Output:** Edge-swapped graph $G_T$
**Initialize:** $G_0 \leftarrow G$
**for** i = 1,...,$T$ **do**
  Select two edges $(i,j)$, $(k,l) \in E(G_{i-1})$
  *condition_1* $\leftarrow (i,l) \in E(G_{i-1})$ **or** $(k,j) \in E(G_{i-1})$
  *condition_2* $\leftarrow$!`allow_antiparallel` **and** $((l,i) \in E(G_{i-1})$ **or** $(j,k) \in E(G_{i-1}))$
  *condition_3* $\leftarrow$ !`allow_loops` **and** $(i \neq l$ **or** $k \neq j)$
  **if** *condition_1* **or** *condition_2* **or** *condition_3* **then**
    **continue** $G_i \leftarrow G_{i-1}$
  **else** $E(G_i) \leftarrow (E(G_{i-1}) \setminus \{(i,j),(k,l)\}) \cup \{(i,l),(k,j)\}$
  **end if**
**end for**

**Figure 2: XSwap algorithm pseudocode. A.** XSwap algorithm presented by Hanhijärvi, et al. [15]. **B.** Proposed extensions to the XSwap algorithm.

**Table 1:** Applications of the modified XSwap algorithm to various network types with appropriate parameter choices. For simple networks, each node's degree is preserved. For bipartite networks, each node's number of connections to the other part is preserved, and overall node class memberships are preserved. For directed networks, each nodes' in- and out-degrees are preserved, though parameter choices depend on the network being permuted. Some directed networks can include antiparallel edges or loops while others do not.

| Network type | Degree preserved | Figure | allow_antiparallel | allow_loops |
|---|---|---|---|---|
| simple | all |  | False | False |
| bipartite | in/out |  | True | True |

## Prediction tasks

We performed three prediction tasks to assess the performance of the edge prior. We compared the permutation-based prior with two additional features: our analytical approximation of the edge prior and the product of source and target degree, scaled to the range [0, 1] to allow calibration assessment. We used 20 biomedical networks from the Hetionet heterogeneous network [4] that had at least 2000 edges for the first two tasks. In the first task, we computed the degree-based prediction features (edge prior, scaled degree product, and analytical prior approximation), and predicted the original edges in the network. We used node pairs that lacked an edge in the original network as negative examples and those with an edge as positive examples. To assess the methods' predictive performances, we computed the area under the receiver operating characteristic (AUROC) curve for all three features. In the second task, we sampled 70% of edges from each of the networks, computed features on the sampled network, then predicted held-out edges. For this task, negative examples were node pairs in which an edge did not exist in either original or sampled network, while positive samples were those node pairs without an edge in the sampled network but with an edge in the original network.

does this have a technical meaning?

how?

these should be specified in an appendix or repo. Pointer belongs here.

The third task evaluated the ability of the edge prior to generalize to new degree distributions. We used two domains where networks were available which shared nodes but had different degree distributions. Protein-protein interactions (PPI) and transcription factor-target gene (TF-TG) relationships had networks created both by literature curation of low-throughput, hypothesis-driven research and by high-throughput, systematic, hypothesis-free experimentation. For the PPI networks, we used the STRING network, which incorporates literature-mining to find relationships [18] and a combination of the high-throughput, proteome-scale interaction networks from Rual et al. [9] and Rolland et al. [10]. We used a transcription factor-target gene (TF-TG) literature-derived network from Han et al. [19] and a high-throughput network from Lachmann et al. [20]. The pairs of networks for PPI and TF-TG data sources are ideal because in one we expect inspection bias and in the other we do not.

Why PP here?

As a further basis of comparison, we added a time-resolved co-authorship network, which we partitioned by time to create two separate networks. We created the co-authorship network of bioRxiv bioinformatics preprints using the Rxivist [21,22] database, which was generated by crawling the bioRxiv server. Unlike the other two networks, co-authorship does not have degree bias, as the network faithfully represents all true co-author relationships. We include this network to offer a comparative prediction task in which the degree distributions between training (posted before 2018) and testing (posted during or after 2018) do not differ (Figure 4A). The goal of the third prediction task is to determine feature generalizability for network reconstruction between different degree distributions, especially predicting a network without degree bias using features from a degree-biased network. Further information about the networks used can be found in the supplement.

it's not immediately obvious why these dists. should be similar. They might be (& you can verify), but not necc. b/c of principal.

## Degree-grouping

Our method for degree-preserving permutation produces randomized networks that share few of their edges with the original network. The feature values for two node pairs with the same source and target degree are drawn from the same distribution in permuted networks, so nodes with equal degree can be grouped when summarizing features. We used this to augment each node pair's feature values in permuted networks, which allowed these pairs to have more permuted feature values than permuted networks. Degree grouping greatly increased the effective number of permutations for nodes with frequently observed degrees [23]. We used degree grouping throughout our analyses.

Confusing.

not entirely clear what this does

## Implementation and source code

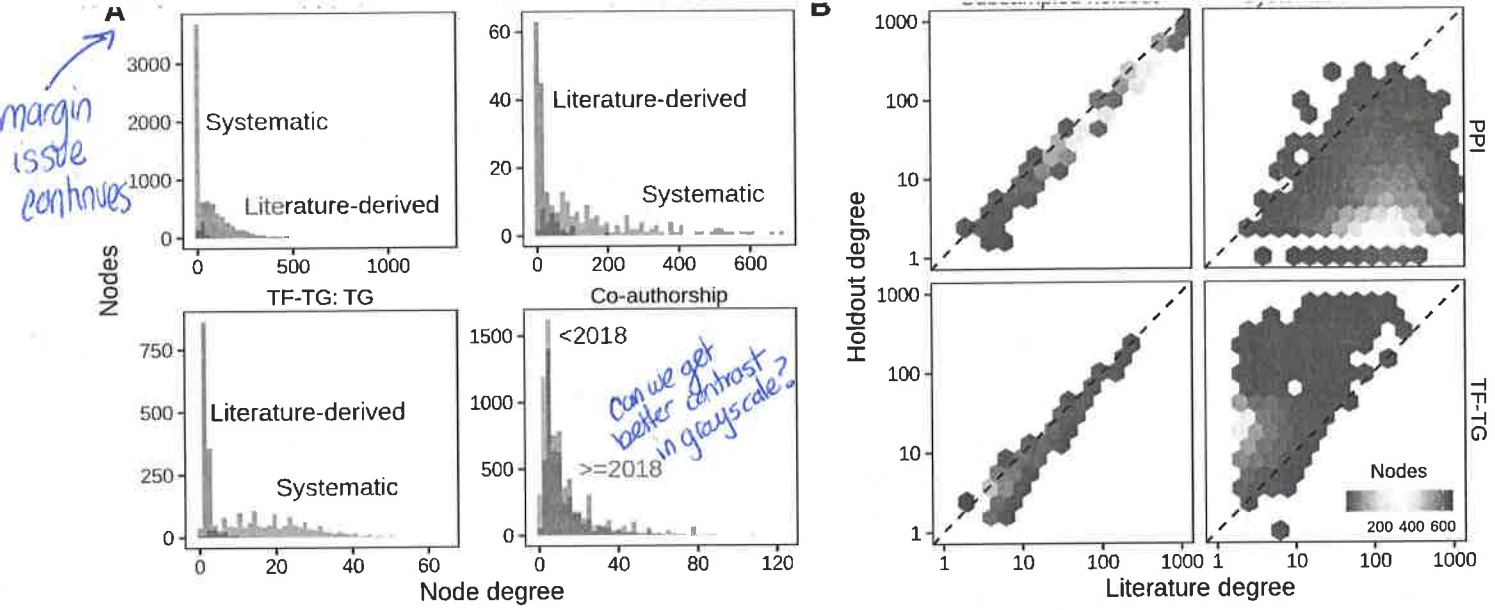
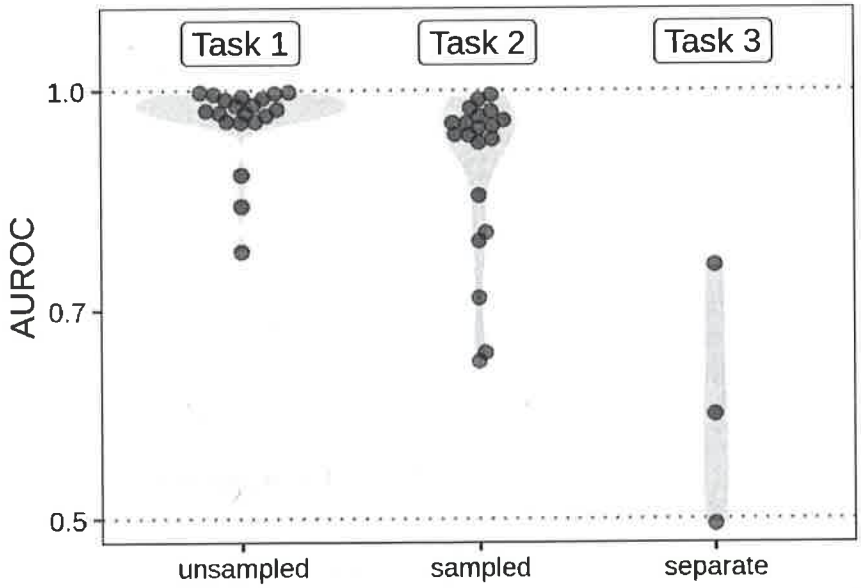



**Figure 4: A.** Degree distributions of networks with and without degree bias can be very different. Data on PPI and TF-TG were split between literature-derived and systematically-derived networks. In both cases, the networks exhibit large differences in degree distribution. Co-authorship relationship networks split by date of first co-authorship roughly share their degree distributions. **B.** Systematically-derived networks are not uniformly sampled ~~sampled~~ from literature-derived networks or vice versa. Uniform random sampling produces linearly-correlated node degree, while non-random sampling produces non-correlated degree. 70% of literature edges were sampled with uniform probability for the "Subsampled holdout" network.

## The edge prior encapsulates degree

In the first prediction task, we computed three features—the XSwap edge prior, an analytical approximation to the edge prior, and the (scaled) product of source and target node degree—on networks from Hetionet. We then evaluated the extent to which these features could reconstruct the 20 networks. The XSwap-derived edge prior reconstructed many of the networks with a high level of performance, as measured by the AUROC. Of the 20 individual networks we extracted from Hetionet, 17 had an edge prior self-reconstruction AUROC >= 0.95, with the highest reconstruction AUROC at 0.9971 (Compound–downregulates–Gene edge type). Meanwhile, the lowest self-reconstruction performance (AUROC = 0.7697) occurred in the network having the fewest node pairs (Disease–localizes–Anatomy edge type).



calibration. The edge prior has excellent calibration in unsampled and sampled networks, and each considered method is sensitive to shifts in the degree distribution.

The second prediction task mirrored the first task, but it involved reconstructing networks based on subsampled networks with only 70% of the original edges. Because edges were sampled uniformly without replacement, the subsampled networks share similar degree distributions to the original networks (see Figure 4B). Unlike in the first task, edges that were present in the sampled network were not tested and therefore are not included in the performance metrics. The results of the second prediction task further demonstrate a high level of performance for degree-sequence-based node pair features (Figure 5). The edge prior was able to reconstruct the unsampled network with an AUROC of greater than 0.9 in 14 of 20 networks. As was observed in the first task, node pair features computed in second prediction task were highly rank-correlated, meaning the AUROC values for different features were similar. While performance was slightly lower in the second task than the first, many networks were still well-reconstructed. The edge prior was the best calibrated feature for both tasks.

In the third prediction task, we computed the three edge prediction features for paired networks representing data from PPI, TF-TG, and bioRxiv bioinformatics pre-print co-authorship. The goal of the task was to compare predictive performance across different degree distributions for the same type of data. We find that the task of predicting systematically-derived edges using a network with degree bias is significantly more challenging than network reconstruction, and we find consistently lower performance compared to the other tasks (Figure 5). The edge prior was not able to predict the separate PPI network better than by random guessing (AUROC of roughly 0.5). Only slightly better was its performance in predicting the separate TF-TG network, at an AUROC of 0.59. We find superior performance in predicting the co-authorship relationships (AUROC 0.75), which was expected as the network being predicted shared roughly the same degree distribution as the network on which the edge prior was computed. The results of the third prediction task show that a difference in degree distribution between the network on which features are computed and the network to be predicted can make prediction significantly more challenging.

The edge prior can be considered a baseline edge predictor that accurately captures degree's contribution to the probability of an edge existing. The edge prior's low performance in the third task indicates that degree is less helpful for edge prediction tasks in which training and testing networks do not share their degree distributions. Many biomedical prediction tasks can be framed as edge prediction tasks between different degree distributions. In drug repurposing, for example, existing compound-disease treatment relationships are unlikely to be randomly sampled from all true treatment relationships. However, all treatment relationships between existing compounds and diseases are desirable outputs in prediction. Edge predictions can be based on both underlying biological properties and network degree distributions. However, predictions based on biological properties may be more consistent and generalizable than those based on degree. Degree's influence on edge prediction accuracy measures can reveal the relative contributions of these two factors.

## Degree can underly a large fraction of performance

We conducted a further edge prediction task as an example application of the edge prior and our permutation framework. To begin, we chose the STRING PPI network for the comparison and computed five edge prediction features (Supplemental table 2). The goal of the task was to reconstruct the network on which the features were computed. All five features were correlated with degree (Figure 7), which we quantified for a node pair using the product of source and target degrees. We expected features based on degree to show strong performance for a network reconstruction task without holdout, as found in the first prediction task.

measured degree (eg: Jaccard index), whereas features whose performances equaled the edge prior completely captured degree (eg: preferential attachment index).

Features can also capture information beyond degree, and our method can quantify this performance. For example, the superior performance on unpermuted networks relative to permuted networks indicated that RWR, resource allocation, Jaccard, and Adamic/Adar indices captured more than degree in this prediction task. These results aligned with the definitions of each feature and validated that our permutation framework accurately assessed reliance on degree.

this should be merged/condensed w/ prior paragraph

## Discussion

We focus on edge prediction in biomedical networks. Our overall goal is to predict new edges with specificity, so that predictions reflect particular connectivity rather than generic node characteristics. Our permutation framework measures the predictive performance attributable to degree to provide a baseline expectation for edge pairs. We expect that non-specificity due to degree is not a unique property of biomedical networks. For example, if node A connects to nearly all other nodes in a network, predicting that all remaining nodes share an edge with node A will likely result in many correct—though nonspecific—predictions, regardless of the type of data contained in the network. Node degree should be accounted for to make correct predictions while being able to distinguish specific from nonspecific predictions. Prediction without reliance on node degree is challenging because many effective methods for edge prediction are correlated with degree (Figure 7).

The effects of node degree are obvious when edge prediction features are functions of degree. For example, the resource allocation index is the sum of inverse degree of common neighbors between source and target nodes (in the symmetric case), while preferential attachment is the product of source and target degree [26,27]. However, because many other edge prediction methods are not explicitly degree-based, it is important to have a general method for comparing the effects of node degree on edge prediction methods.

We developed a permutation framework to quantify the edge probability due to degree. We term this probability the "edge prior", and we have identified two applications. First, a probability associated with every node pair can be treated as a classification score. Ordering these scores provides an assessment of performance based solely on degree, which can be used as a baseline for other classifiers. Second, node pair probabilities can be used to adjust edge prediction features depending on the task. If degree is a desired feature, then the edge prior can be treated like a Bayesian prior probability. Alternatively, if degree is not a desired feature, then the edge prior can be used to calibrate features and thus potentially enhance predictive specificity.

? Is this correct?

Figure 8 illustrates the utility of the edge prior and permutation framework for two purposes. First, it contextualizes feature performances relative to the baseline of nonspecific, degree-based predictions, quantified by the edge prior. Degree has varying utility for different edge prediction tasks. The edge prior's performance on a task quantifies the utility of degree toward the task. This comparison is useful because specific predictions (based on more than degree alone) are more valuable for some applications than nonspecific ones and because degree can be an expression of bias in many real-world networks.

awkward

Second, Figure 8 compares five edge prediction features computed on and unpermuted networks. This comparison identified the fraction of each feature's performance attributable to degree. Some features, such as the preferential attachment index, perfectly and exclusively measure degree. The Adamic/Adar index also incorporates degree almost completely because its performances from permuted networks are nearly at the performance of the edge prior. However, the Adamic/Adar index had much higher performance when computed on the unpermuted network, indicating that it also extracts higher-order information. This analysis, enabled by network permutation, measured the

## References


1. **Emergence of Scaling in Random Networks**
Albert-László Barabási, Réka Albert
*Science* (1999-10-15) https://doi.org/ccsmnz
DOI: 10.1126/science.286.5439.509 · PMID: 10521342

2. **Scale-free networks are rare**
Anna D. Broido, Aaron Clauset
*Nature Communications* (2019-03-04) https://doi.org/gfztz9
DOI: 10.1038/s41467-019-08746-5 · PMID: 30833554 · PMCID: PMC6399239

3. **Biology, Methodology or Chance? The Degree Distributions of Bipartite Ecological Networks**
Richard J. Williams
*PLoS ONE* (2011-03-03) https://doi.org/fmtk6x
DOI: 10.1371/journal.pone.0017645 · PMID: 21390231 · PMCID: PMC3048397

4. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini
*eLife* (2017-09-22) https://doi.org/cdfk
DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

5. **Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics**
Fei Gao, Katarzyna Musial, Colin Cooper, Sophia Tsoka
*Scientific Programming* (2015) https://doi.org/f7hvd9
DOI: 10.1155/2015/172879

6. **Bias tradeoffs in the creation and analysis of protein–protein interaction networks**
Jesse Gillis, Sara Ballouz, Paul Pavlidis
*Journal of Proteomics* (2014-04) https://doi.org/f3mn5f
DOI: 10.1016/j.jprot.2014.01.020 · PMID: 24480284 · PMCID: PMC3972268

7. **Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types**
Martin H. Schaefer, Luis Serrano, Miguel A. Andrade-Navarro
*Frontiers in Genetics* (2015-08-04) https://doi.org/gf5t46
DOI: 10.3389/fgene.2015.00260 · PMID: 26300911 · PMCID: PMC4523822

8. **Effect of sampling on topology predictions of protein-protein interaction networks**
Jing-Dong J Han, Denis Dupuy, Nicolas Bertin, Michael E Cusick, Marc Vidal
*Nature Biotechnology* (2005-07) https://doi.org/dj5cm8
DOI: 10.1038/nbt1116 · PMID: 16003372

9. **Towards a proteome-scale map of the human protein–protein interaction network**
Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F. Berriz, Francis D. Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, ... Marc Vidal
*Nature* (2005-09-28) https://doi.org/dw6q23
DOI: 10.1038/nature04209 · PMID: 16189514

Yeong Kim, Muyoung Lee, Eunbeen Kim, ... Insuk Lee
*Nucleic Acids Research* (2017-10-26) https://doi.org/gcwpcz
DOI: 10.1093/nar/gkx1013 · PMID: 29087512 · PMCID: PMC5753191

20. **ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments**
Alexander Lachmann, Huilei Xu, Jayanth Krishnan, Seth I. Berger, Amin R. Mazloom, Avi Ma'ayan
*Bioinformatics* (2010-08-13) https://doi.org/d2h98v
DOI: 10.1093/bioinformatics/btq466 · PMID: 20709693 · PMCID: PMC2944209

21. **Tracking the popularity and outcomes of all bioRxiv preprints**
Richard J. Abdill, Ran Blekhman
*Cold Spring Harbor Laboratory* (2019-01-13) https://doi.org/gftzwz
DOI: 10.1101/515643

22. **Complete Rxivist dataset of scraped bioRxiv data**
Richard J. Abdill, Ran Blekhman
*Zenodo* (2019-03-21) https://doi.org/gfz3fm
DOI: 10.5281/zenodo.2566421

23. **Degree-grouped permtuations by zietzm · Pull Request #96 · greenelab/hetmech**
GitHub
https://github.com/greenelab/hetmech/pull/96

24. **Roaring Bitmaps: Implementation of an Optimized Software Library**
Daniel Lemire, Owen Kaser, Nathan Kurz, Luca Deri, Chris O'Hara, François Saint-Jacques, Gregory Ssi-Yan-Kai
*arXiv* (2017-09-22) https://arxiv.org/abs/1709.07821v3
DOI: 10.1002/spe.2560

25. **Open collaborative writing with Manubot**
Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
*PLOS Computational Biology* (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653

26. **Predicting missing links via local information**
Tao Zhou, Linyuan Lü, Yi-Cheng Zhang
*The European Physical Journal B* (2009-10) https://doi.org/dd55vr
DOI: 10.1140/epjb/e2009-00335-8

27. **Link prediction approach to collaborative filtering**
Zan Huang, Xin Li, Hsinchun Chen
*Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05* (2005)
https://doi.org/fn39g8
DOI: 10.1145/1065385.1065415

28. **The link-prediction problem for social networks**
David Liben-Nowell, Jon Kleinberg
*Journal of the American Society for Information Science and Technology* (2007) https://doi.org/c56765
DOI: 10.1002/asi.20591

### Approximate edge prior

To approximate the edge prior, we began by making two simplifications. First, we assumed independence between node pairs. This assumption does not actually hold for the XSwap algorithm, though it is a reasonable simplification for large, sparse networks. Second, we assumed that the XSwap process is stationary. This assumption also does not actually hold, but it was made because it significantly simplifies the problem. A single node pair has two possible states, "edge" and "no edge". These states are not transient, and they are not periodic so long as more than one possible swap exists in the network. In almost all cases, then, our simplified model of the algorithm gives the state of a node pair as an ergodic process, independent of other node pairs.

Let $A_{i,j}$ represent the existence of edge $(i, j)$ For a given node pair, $(i, j)$, then, let $q_{i,j}$ represent the transition probability from the "no edge" state to the "edge" state in one successful iteration of the XSwap algorithm. Let $r_{i,j}$ represent the probability of the opposite transition ("edge" to "no edge") in one successful iteration. With "no edge" represented as $[1, 0]^T$ and "edge" represented as $[0, 1]^T$, the transition matrix, $P$, is given by the following:

$$P^T = \begin{bmatrix} 1-q & r \\ q & 1-r \end{bmatrix}$$

The stationary distribution of this system should correspond to the distribution when the number of swaps goes to infinity. It can be found by computing the eigenvectors of the system, as we know that the stationary distribution vector, $\mathbf{v}$ satisfies $P^T\mathbf{v} = \mathbf{v}$. The normalized eigenvector $\mathbf{v}$ is given by

$$\mathbf{v} = \frac{1}{r/q+1}\begin{bmatrix} r/q \\ 1 \end{bmatrix}$$

The asymptotic edge probability is therefore

$$\frac{1}{r/q+1}.$$

Since node pairs are being treated as independent, the probability of an edge being created in one successful iteration, given that the edge does not currently exist, is the ratio of the number of edge choices involving nodes $i$ and $j$ to the total number of possible swaps, $S$. Let $d(u_i)$ represent the degree of source node $i$ and $d(v_j)$ represent the degree of target node $j$.

$$q_{i,j} = \frac{d(u_i)d(v_j)}{S}$$

Similarly, the probability of an edge being eliminated in one iteration is the ratio of the number of edge choices involving $(i, j)$ and any other valid edge to the total number of possible swaps. Let $m$ be the total number of edges in the network.

$$r_{i,j} = \frac{m - d(u_i) - d(v_j) + 1}{S}$$

The approximate edge prior is, therefore,

| | | | |
|---|---|---|---|
| | GiG | 20945 | 147164 |
| | GpMF | Source: 20945, Target: 2884 | 97222 |
| | GpPW | Source: 20945, Target: 1822 | 84372 |
| | Sampled | 3992 | 255522 |
| PPI | Literature | 3992 | 364743 |
| | Systematic | 3916 | 12913 |
| | Sampled | 4587 | 30686 |
| bioRxiv | <2018 | 4615 | 43691 |
| | All time | 4615 | 44963 |
| TF-TG | Sampled | Source: 142, Target: 1396 | 2689 |
| | Literature | Source: 144, Target: 1406 | 3496 |
| | Systematic | Source: 144, Target: 1417 | 29177 |

## Edge prediction features

In the table that follows, let $k(u)$ denote the set of neighbors of node $u$. Let $\mathbf{A}$ represent the normalized Laplacian adjacency matrix, and let $y_u$ be a vector with all ones except for a one in the $u$-th position. $x$ For a directed graph, let $A(u)$ denote the set of nodes that node $u$ points to and $D(u)$ the set of nodes that point to $u$. All definitions that follow are the score between nodes $u$ and $v$.

**Table 2:** Edge prediction features.

| Feature | Definition | Citation |
|---|---|---|
| Jaccard index | $\frac{\lvert k(u)\cap k(v)\rvert}{\lvert k(u)\cup k(v)\rvert}$ | [28] |
| Preferential attachment score | $\lvert k(u)\rvert\lvert k(v)\rvert$ | [28] |
| Resource allocation index | $\sum_{w\in k(u)\cap k(v)} \frac{1}{\lvert k(w)\rvert}$ | [26] |
| Adamic/Adar index | $\sum_{w\in k(u)\cap k(v)} \frac{1}{log\lvert k(w)\rvert}$ | [29] |
| Random walk with restart score | $c\left[\left(\mathbb{I}-(1-c)\mathbf{A}\right)^{-1}\mathbf{y}_u\right]_v$ | [30,31] |
| Inference score | $\frac{\lvert A(u)\cap D(v)\rvert}{\lvert A(u)\rvert}+\frac{\lvert D(u)\cap D(v)\rvert}{\lvert D(u)\rvert}$ | [32] |