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@2b069f9 on August 7, 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)

- **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)']

Abstract

Networks of biomedical data rarely consist of all true relationships. Instead, networks contain spurious relationships while omitting actual relationships. How a network deviates from the real set of relationships is often biased according to node degree, resulting from processes such as inspection bias and experimental methods. While degree is subject to potentially substantial biases, link prediction methods can be strongly affected by degree. In the present work, we introduce a network permutation framework to quantify the effect of node degree on network-based methods and prediction tasks. We introduce the “edge prior” to quantify the probability that two nodes are connected based only on their degree. After demonstrating that this prior feature shows excellent discrimination and calibration performance for 20 different biomedical networks (16 bipartite, 3 undirected, 1 directed), we conclude that our prior feature represents a suitable baseline for network link prediction tasks, as performance exceeding the baseline is attributable to factors other than degree alone. Additionally, we propose methods to incorporate network permutation and the edge prior into other predictive methods. Our results highlight the importance of degree for link prediction and provide a way to account for its effects when degree bias may be present. We have released a full implementation of our network permutation method and the edge prior as an open-source Python package on GitHub.

Introduction

Node degree

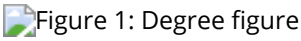
Figure 1: Degree figure

Figure 1: Degree figure

Edge prediction

Feature-degree correlation

Methods

Network permutation

XSwap algorithm

Edge prior

Edge prior approximation

Prediction tasks

Degree-grouping

Implementation and source code

Results


Figure 2: Discrimination figure

Figure 2: Discrimination figure

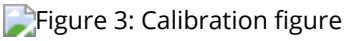
Figure 3: Calibration figure

Figure 3: Calibration figure

Discussion

Conclusion

References