Essential Protein Detection by Random Walk on Weighted Protein-Protein Interaction Networks

Bin Xu and Jihong Guan

GIW 2016

Outline

- **Background**
- Contribution
- Method
- Evaluation
- Results & Discussion
- Conclusion

Background

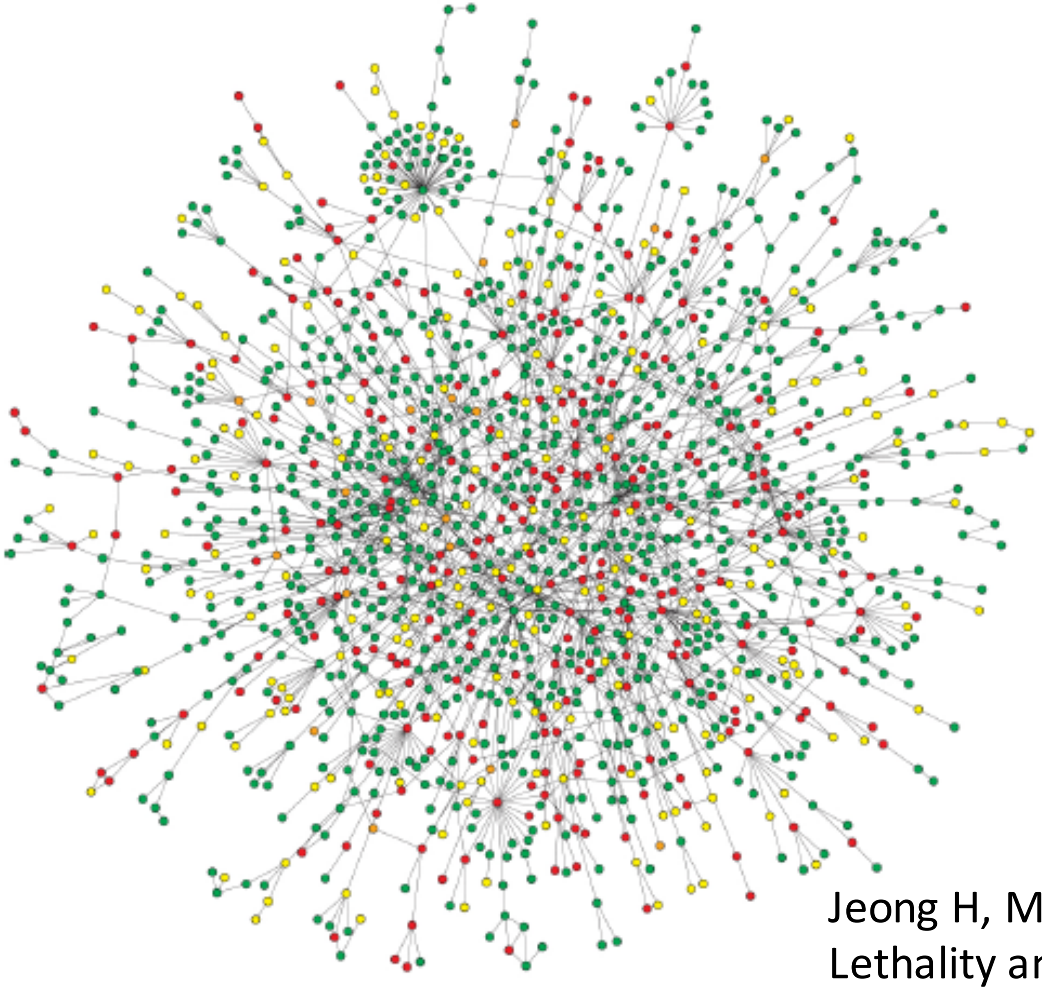
Essential proteins:

> *critical to the development and survival of cells.*

> *deletion of these proteins will result in lethality*

Identification of essential proteins is critical for the understanding of the minimal requirements of organisms

Background



Red nodes: essential

Jeong H, Mason S P, Barab´asi A L, et al.
Lethality and centrality in protein networks.
Nature, 2001, 411(6833):41–42.

Background

To detect essential proteins:

Web Lab Experiment: single gene knockouts, RNA interference, conditional knockouts, …

Disadvantages:

✗ inefficient

✗ not practical for whole-organism

Background

To predict essential genes:

- conserved
- ubiquitously present across organisms
- …

Drawback:

- limited to conserved orthologs

Background

To predict essential proteins:

Analysis topology attributes of nodes with in PIN

- Degree centrality

$$DC(p) = \sum_{q} A_{p,q}$$

- betweenness centrality
- closeness centrality
- …

Drawback:

Law confidence of protein-interaction data

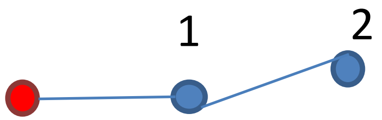
Outline

- Background
- **Contribution**
- Method
- Evaluation
- Results & Discussion
- Conclusion

Contribution

Challenge：

(1) Low confidence of PPI



(2) Neglecting
indirect protein neighbors

Our Method

★ Multiple data source &
evaluate confidence of PPI

★ Weighted PPI network

★ Discovering important nodes(essential proteins) using Personalized PageRank

Outline

Background

Contribution

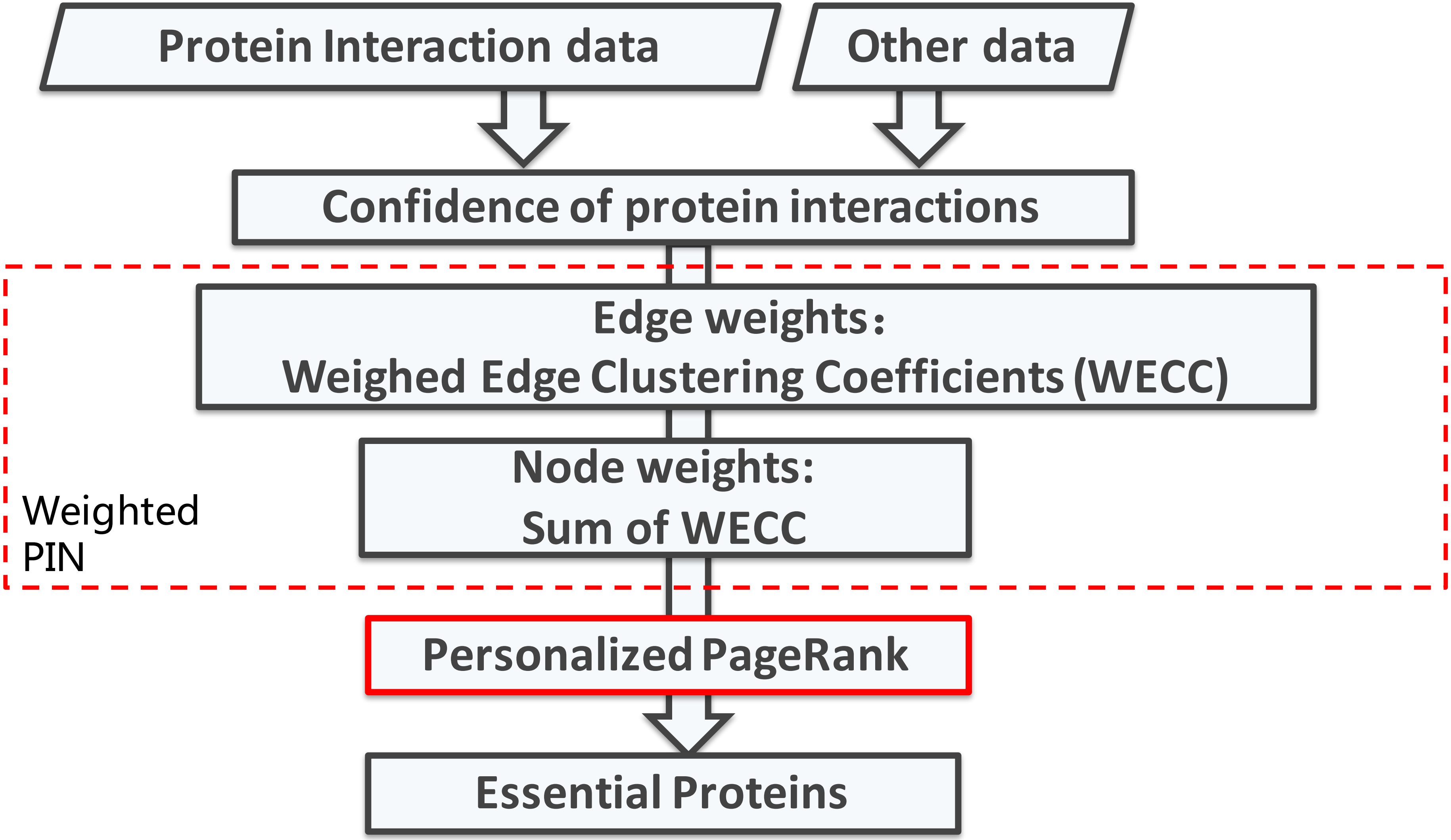
Method

Evaluation

Results & Discussion

Conclusion

Method - EssRank



Protein Interaction data

Other data

Confidence of protein interactions

Edge weights：
Weighed Edge Clustering Coefficients (WECC)

Node weights:
Sum of WECC

Weighted
PIN

Personalized PageRank

Essential Proteins

Method – Protein Interaction Confidence

PPI Interaction data

Other data source (for each interacting protein pairs):

> Number of common neighbors of proteins with in PIN

> Expression correlation

> Functional Similarity

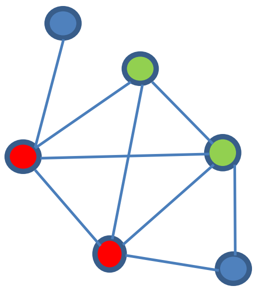
> Number of Domain-domain interactions

> Phylogenetic profile similarity

Y. Jiang, Y. Wang, W. Pang, L. Chen, H. Sun, Y. Liang, and E. Blanzieri, “Essential protein identification based on essential protein–protein interaction prediction by integrated edge weights,” Methods, vol. 83, pp. 51–62, 2015.

Method – Weighting Edges and Nodes

$$ECC(p,q)=\frac{z_{p,q}}{min(d_p-1,d_q-1)}$$



Wang J, Li M,Wang H, et al. Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2012, 9(4):1070–80.

Edge weights

$$wECC(p,q)=\frac{\sum_{r\in N_p\cap N_q}\boxed{w(p,r)}+\boxed{w(q,r)}}{min(d_p-1,d_q-1)}$$

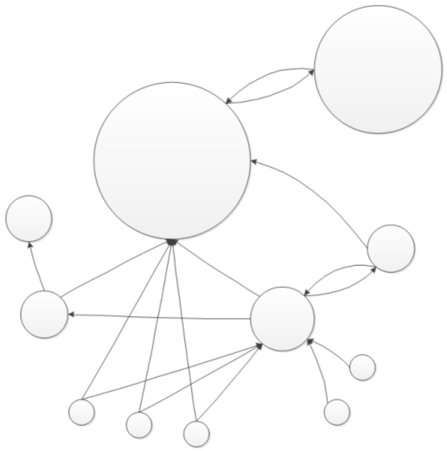
Node weights

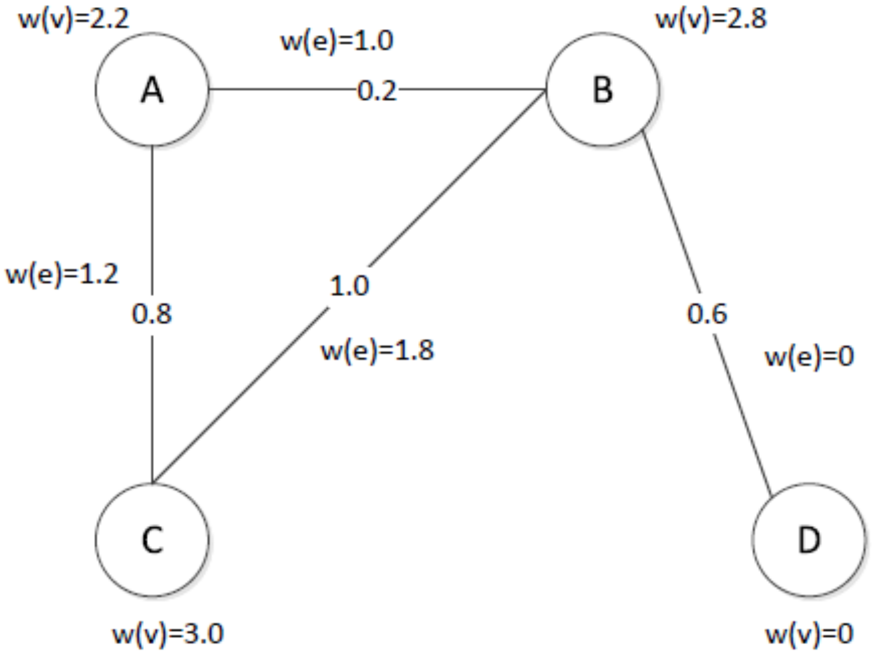
$$w(p)=\sum_{q\in N_p} wECC(p,q)$$

Method – Personalized PageRank

Personalized vector e： Node Importance

Iteration： $v' = \beta M v + (1-\beta) * e$





$$\underset{e}{\begin{pmatrix}2.2\\2.8\\3\\0\end{pmatrix}} \qquad \underset{v_0}{\begin{pmatrix}0.25\\0.25\\0.25\\0.25\end{pmatrix}} \quad \underset{v_1}{\begin{pmatrix}0.28\\0.35\\0.38\\0.00\end{pmatrix}} \quad \underset{v_2}{\begin{pmatrix}0.25\\0.41\\0.31\\0.03\end{pmatrix}} \quad \underset{v_3}{\begin{pmatrix}0.27\\0.34\\0.33\\0.06\end{pmatrix}}$$

Outline

- Background
- Contribution
- Method
- **Evaluation**
- Results & Discussion
- Conclusion

Evaluation Dataset

PPI： DIP database

Esstial proteins： DEG、MIPS、SGD、SGDP

#protein	#interactions	# essential proteins
5003	22102	1153

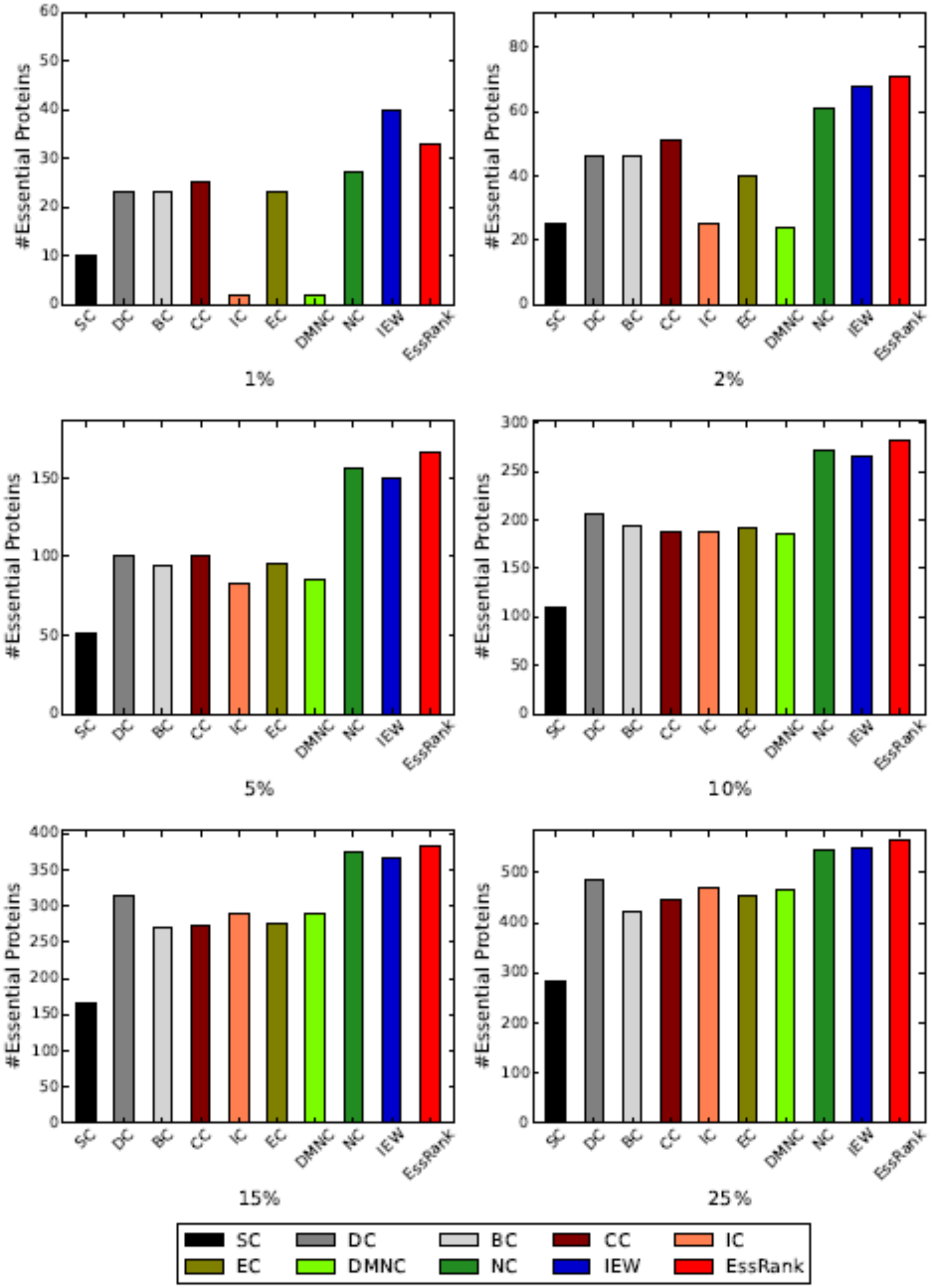
Other datasets：

- Gene expression： GSE3431
- Function, Domain： SGD database
- Phylogenetic profile: InParanoid database

Result

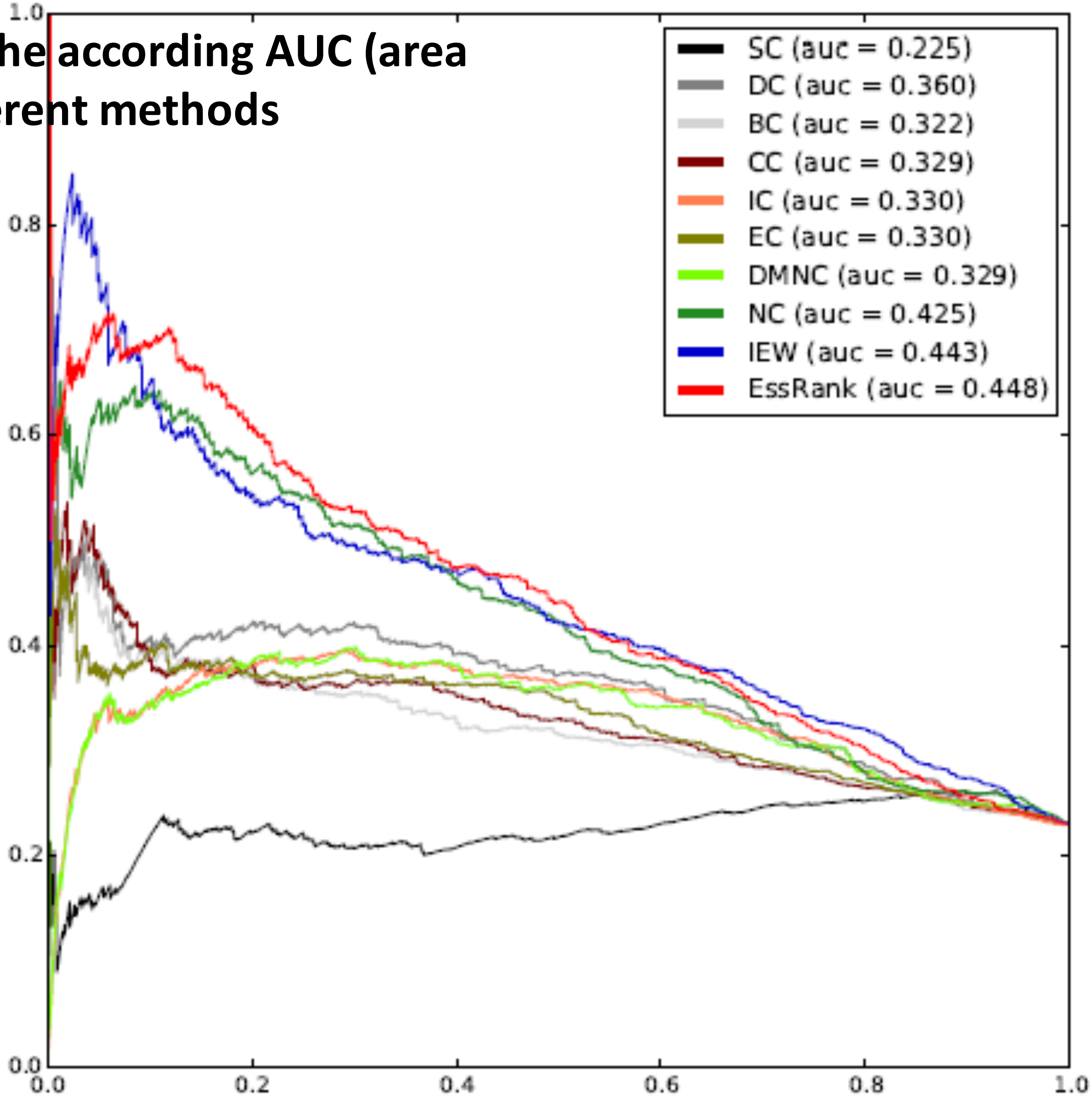
Number of Essential Proteins

in top K percent candidates



Result

Precision-recall curve and the according AUC (area under curve) scores of different methods



Conclusion

- Weighted edges and nodes in PIN.
- Involving both direct and indirect protein neighbors.
- Higher accuracy.

Thank you for listening!