Homework 3

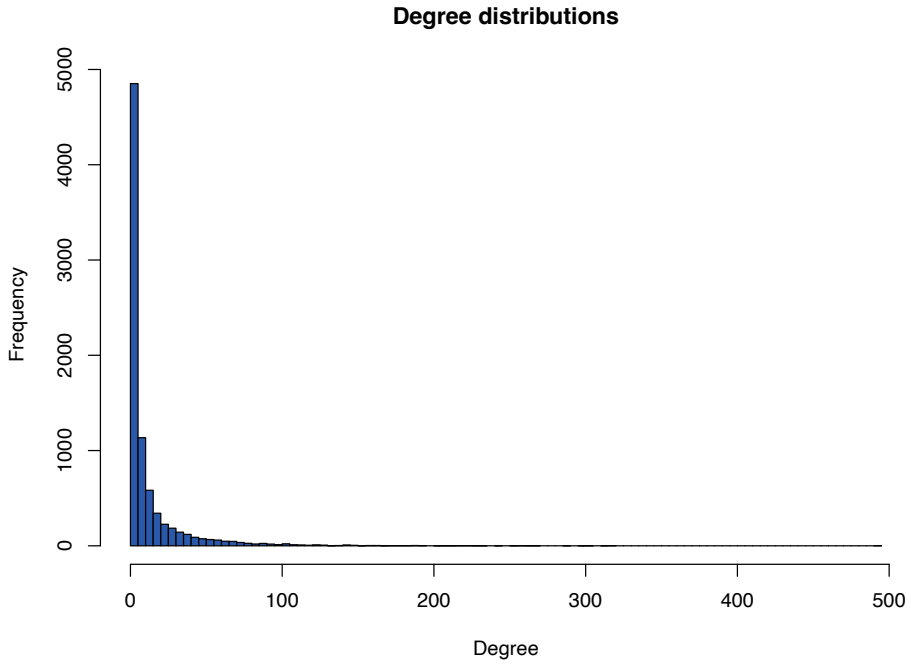
1. **Understanding the data:** Briefly describe the experimental approach used to generate the data HURI.hgnc.csv you are about to analyze. Are there any caveats associated with the approach? How many protein pairs were screened to generate this data? How many total interactions were reported?

This paper expanded the ORFeome collection to include ORFs for approximately 90% of the protein-coding genes in the human genome, and utilized three different versions of the yeast two-hybrid (Y2H) technique to screen this ORFeome collection nine times for the detection of binary protein-protein interactions (PPIs). After identifying the potential interactions using Y2H, the interactions are further validated using orthogonal assays, which can be used to confirm the findings of the Y2H screening and increase confidence in the identified interactions. This approach screened 17,500 X 17,500 proteins and generated a dataset that consists of about 53,000 PPIs.

One of the major caveats associated with the Y2H is its potential for false positives and negatives. For example, Y2H assay can only detect the interactions that occur within the nucleus, so it will miss membrane-bound proteins. Another common problem is the autoactivation of Y2H inducible reporter genes. It happens when one of the proteins activates transcription of the Y2H reporter genes independently of any PPI, and this can increase the false positives.

2. Analysis of interaction degree:

(a) Measure the degree of each protein with at least 1 interaction in the network (exclude self-interactions). Plot the degree distribution of the protein-protein interaction network (a histogram is fine).



The degree distribution of this network is highly skewed, indicating many proteins with only a small number of interactions, which is commonly observed in the scale-free network.

(b) What is the highest degree protein and how many interactions does it have? Describe what is currently known about the functional role of this protein (you can use http://www.genecards.org to learn about gene function).

The protein encoded by MEOX2 (Mesenchyme Homeo Box 2) has the highest degree, which is 495. MEOX2 encoded protein plays a role in the regulation of vertebrate limb myogenesis. It is involved in regulating the differentiation of mesodermal cells into various cell types, including smooth muscle cells, osteoblasts, and chondrocytes. Additionally, MEOX2 has been implicated in angiogenesis, the process of forming new blood vessels, and in the development of the cardiovascular system. Diseases associated with MEOX2 include Female Stress Incontinence and Alzheimer's Disease.

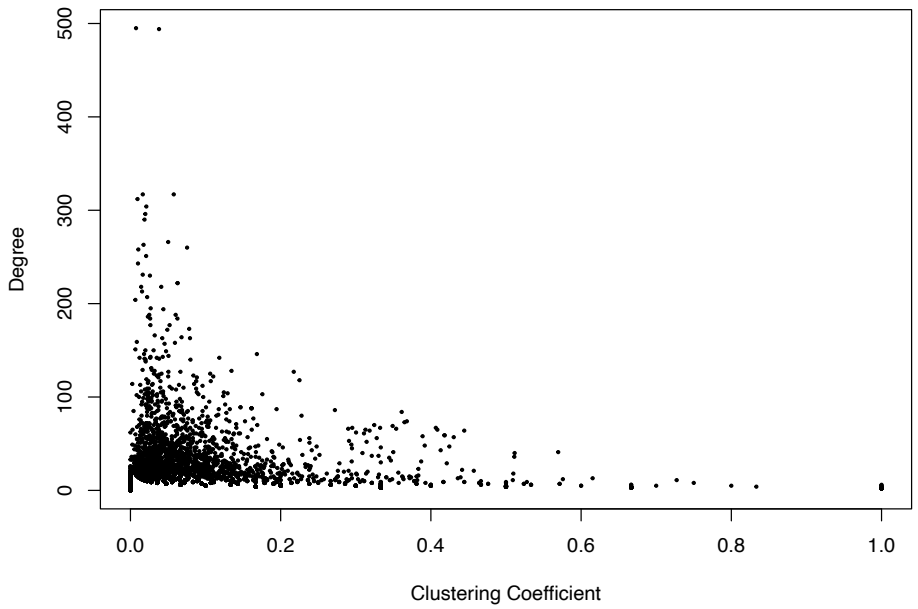
3. Analysis of clustering coefficient:

(a) What is the clustering coefficient for a node in a graph? Give an intuitive description of what topological properties it captures.

The clustering coefficient for a node in a graph measures the extent to which its neighbors are connected with each other. It provides a local structure of the network surrounding a node of interest. More specifically, the clustering coefficient quantifies how the neighbors of the protein of interest connect with each other and how close they are to forming a complete subgraph among themselves.

A high clustering coefficient (close to 1) indicates that the protein's neighbors are more likely to be connected with each other, and this protein is within this densely connected cluster or subgroup. One example is a protein serving as part of a complex. On the contrary, a low clustering coefficient (low to 0) indicates that the protein's neighbors have fewer connections. These proteins are commonly seen as part of a signaling pathway and usually serve as bridge proteins, which facilitate communication between two distinct parts of the network.

(b) Compute clustering coefficients for every protein in this network. For simplicity, exclude self-interactions in the network. To present your results, plot the clustering coefficient vs. the node degree for all proteins.



(c) Let's investigate the properties of two specific proteins, LSM6 and MAPK9. Report the number of interaction partners for both of these proteins and the clustering coefficients. Also, list the actual interacting proteins with

each of them. Comment on what's known about the function of these two proteins. Are the observed clustering coefficients and interacting partners consistent with the known functions of these proteins?

The interaction partners of LSM6 and MAPK9 are 6 and 49, respectively, and their clustering coefficients are 0.3333333 and 0.01955782, respectively. The interacting proteins with LSM6 are as following: LSM2, LSM3, PATL1, CLNS1A, LSM5, and SDCBP. The interacting proteins with MAPK9 are as following: MCRS1, NHSL2, EFHC2, KLHL8, SDCBP, L3MBTL3, CEP44, PPARG, ZC2HC1C, TEX11, DUSP10, DUSP16, CCDC36, ENKD1, LHX3, POU6F2, SMCO3, PBX4, ITGB3BP, ZNF138, DTX3, PICK1, SF3B4, GOLGA6A, ATF2, PAX5, ZNF559, DUSP4, GFAP, TCP10L, SAXO1, SHMT1, KPNA3, GSC2, LNX1, ARRB1, LZTS1, C1orf105, CBLL2, DCX, BNIP5, RASL10B, MLIP, MED12L, ARRB2, CDC16, ZNHIT1, RSPO4, and MEOX1.

The LSM6 gene encodes U6 snRNA-associated Sm-like protein LSm6. The Sm-like proteins share structural similarities with the SM proteins, which are core components of spliceosomal small nuclear ribonucleoproteins (snRNPs). The Sm-like proteins (including LSm6) are thought to form a stable heteromer present in tri-snRNP particles, which are crucial components of the spliceosome, a large machine responsible for removing introns from pre-mRNA transcripts and joining together the remaining exons to produce mature mRNA. Overall, LSm6 protein serves as a component in a large complex, and half of its neighbors (LSM2, LSM3, and LSM5) are also components of this large complex. It can also explain why the clustering coefficient is not small (0.3333).

The MAPK9 gene encodes mitogen-activated protein kinase 9. This protein is a member of the MAP kinase family. MAP kinases act as an integration point for multiple biochemical signals and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation, and development. MAPK9 protein targets specific transcription factors and thus mediates immediate-early gene expression in response to various cell stimuli. Overall, this protein (kinase) is part of the signaling pathway, which can explain its clustering coefficient is very small (0. 01956).

4. Comparison of systematically mapped and literature curated interaction networks:

(a) What is the Pearson correlation between interaction degrees in the systematically mapped network and in the literature-curated network? (Use the proteins with at least one interaction in both networks after you exclude self-interactions in the systematically mapped network). Discuss your interpretation of this.

The Pearson correlation between interaction degrees in the systematically mapped network and the literature-curated network is 0.07891022. This correlation coefficient is lower than my initial expectation, but it appears to make sense. Among the 1527 proteins shared between the systematically mapped network and the literature-curated network, 1005 of them exhibited higher interaction degrees in the former than in the latter. This observation suggests a potential underestimation of protein degrees in previous literature-based calculations.

The systematically mapped network, based on data from Luck et al. (2020), is much more comprehensive, encompassing approximately 53,000 protein-protein interactions. Based on the data, I found around 8000 proteins with at least one interaction, while the literature-curated network includes only about 4000 proteins. These differences likely contribute to the relatively low correlation between the two networks.

(b) Find an example of a protein with more than 10 interactions in the Luck et al. network, a clustering coefficient of greater than 0.2, and no interactions in the literature curated network. What was previously known about its function and what can you learn about its function from the interaction partners?

The RBM3 gene encodes RNA binding motif protein 3. RBM3 is a cold-induced protein that can help cells adapt to low temperatures by stabilizing mRNA transcripts encoding proteins involved in various cellular processes. Some research has found that RBM3 can also be induced under other stressed conditions. Additionally, RBM3 can enhance the translation of specific mRNA transcripts into proteins. The ability may be particularly important for sustaining protein synthesis in neuronal development and function, especially under conditions of stress or injury. Moreover, RBM3 exerts anti-apoptotic effects, which may contribute to its role in promoting cell survival in response to various stresses. RBM3 is also found to be a proto-oncogene that is associated with tumor progression and metastasis and has been implicated in promoting resistance to certain chemotherapy and radiation therapy for cancer.

RBM3 is found to have 14 interaction neighbors and a clustering coefficient of 0.4396. Its neighbors include RBPMS2, HNRNPK, PCBP1, RBMY1A1, RBMY1J, RBMY1F, RBMX, MYPOP, KHDRBS2, PRR3, SRSF3, CHTOP, KHDRBS3, and SNRPA. The functions of these neighboring proteins are listed below. These proteins primarily function as RNA-binding proteins and are involved in RNA processing and alternative splicing. The moderate clustering coefficient suggests a degree of interaction among these neighbors. The presence of these interacting proteins indicates that RBM3 may be part of a signaling pathway or a molecular complex involved in multiple cellular processes. This aligns with RBM3's involvement in stress response, translation regulation, and cancer biology. A deeper understanding of RBM3's function could provide insights into cellular survival mechanisms and disease pathways.

Protein	Possible functions
RBPMS2 (RNA-binding protein with multiple splicing 2)	A member of the RBPMS family, which is characterized by RNA recognition motifs (RRMs) that enable them to bind to RNA molecules.
HNRNPK (heterogenous nuclear ribonucleoprotein K)	This protein attaches (binds) to DNA or RNA to other proteins.
PCBP1 (Poly(rC)-binding protein 1)	PCBP1 is a multifunctional RNA-binding protein. Its primary function is to bind to single-stranded poly(C) tracts in RNA molecules, thereby regulating RNA metabolism and processing.
RBMY1A1/ RBMY1J/ RBMY1F	These proteins contain RNA-binding motif in the N-terminus and four SRGY (serine, arginine, glycine, tyrosine) boxes in the C-terminus. They are thought to function as a splicing regulator during spermatogenesis.
RBMX (RNA-binding motif protein, X-linked)	The RBMX protein is involved in several cellular processes, primarily related to RNA metabolism.
MYPOP	Myb proteins are a family of transcription factors involved in regulating gene expression, particularly in the control of cell proliferation, differentiation, and apoptosis. They contain a DNA-binding domain called the Myb domain, which allows them to bind to specific DNA sequences and regulate the transcription of target genes.
KHDRBS2/KHDRBS3	These two proteins are member of the signal transduction and activation of RNA (STAR) family of RNA-binding proteins. It plays a significant role in various cellular processes, particularly in RNA metabolism and signal transduction pathways.
PRR3 (Proline-rich protein 3)	Proteins containing proline-rich domains often participate in protein-protein interactions, acting as scaffolds for the assembly of larger protein complexes or as adaptors that link different proteins together. PRR3 may serve similar roles and facilitate interactions between different cellular components or mediating signal transduction pathways.
SRSF3 (Serine/arginine-rich splicing factor 3)	The main function of SRSF3 is to regulate alternative splicing by binding to specific sequences within pre-mRNA molecules and

	influencing the selection of splice sites. This activity helps in generating protein diversity, as different splice variants can have distinct functions or regulatory properties.
CHTOP (Chromatin Target of PRMT1)	The CHTOP protein plays several important roles within the cell. Its main function is associated with RNA processing and gene expression regulation. Specifically, CHTOP is involved in mRNA splicing, which is the process of removing introns from pre-mRNA to produce mature mRNA molecules that can be translated into proteins.
SNRPA (U1 small nuclear ribonucleoprotein A)	The SNRPA protein is a component of the spliceosome. Specifically, SNRPA is a key component of the U1 small nuclear ribonucleoprotein (snRNP) complex, which is involved in the recognition of the 5’ splice site during pre-mRNA splicing.