Cancer drug target discovery in protein-protein interaction networks

Christopher Probert
cprobert@stanford.edu

1. Background

Protein-protein interaction networks

Proteins are linear chain biomolecules that are the basis of functional networks in all organisms. They include classes that regulate other proteins (e.g. kinases), classes that regulate gene expression (e.g. transcription factors), and classes that control response to disease (e.g. immunoglobulins). Aspects of their social networks have been extensively studied across organisms, particularly in yeast [1]–[3], and humans [4]–[7].

Relevance for treating disease and drug discovery

Protein-protein interaction (PPI) networks can be used directly to study disease and for drug discovery. The causes of disease largely manifest on a protein interaction level: most cancers are caused by increasing interaction edge weights of oncogenes and decreasing interaction edge weights of tumor suppressor genes [8]–[10]. Similarly, PPI networks are a basis of drug discovery: the majority of approved drugs target a particular protein, and especially a particular protein-protein interaction [8], [11].

Most human diseases are thought to have fewer than five causal protein-protein interactions; many have two or fewer causal interactions [11]. For example, in adult acute myeloid leukemia, the majority of patients have perturbations in the connectivity of only two proteins/nodes [12]. Given the small network effect size of many diseases, ideally, a targeted therapy for a particular disease would only affect the associated edges. However, most drugs have high-rank projections into the PPI network space from 'off-target' interactions [8]. Recent opinions in drug discovery suggest that the majority of drug effects are off target; the process of drug optimization is minimizing deleterious off-target projections while maintaining disease-associated interactions.

Cancer is an excellent disease for studying network effects of drugs: it is well studied, has many approved drugs with studied targets, and has many implicated protein interaction network features [8]. Resistance is a dominant problem in treating cancer, which occurs when network evolution allows recovery from the perturbations created by a particular drug, rendering the drug ineffective for further treatment. Treatment strategies and resistance can be seen directly as graph processes.

A common problem in modern medicine is to identify disease-associated genes. One of the most common approaches to this problem are genome-wide association studies (GWAS), which are correlation tests run on data collected from large disease / genotype cohorts. GWAS suffers from very poor sensitivity, and low reproducibility. PPI networks have a suggested role in improving GWAS: they can generate specific gene-disease hypotheses, which can significantly lower the multiple testing burden – both increasing the statistical discovery power for a particular sample size and decreasing the expected number of false positive associations. [4], [13]

Inferring protein interaction networks

Edges in PPI networks can be inferred through experimental and statistical evidence. Specific methods include experimental measurements directly on target proteins, inference using expression co-occurrence in a target species, and inference using these features extracted from evolutionarily related species [5]–[7]. These methods have different bias/variance tradeoffs, and equivalently provide different sensitivity/specificity for interaction edge discovery. Robust to the specific interaction estimator used, protein-protein interaction networks of diverse genera are approximately power law distributed [2], [3].

Edges supported by expression data hold temporal attributes: for an interaction to occur, both proteins must be expressed simultaneously [1], [2]. Using these properties, it is possible to cluster nodes with high degree based on whether they interact with partners simultaneously (a "party" node), or individually (a "date" node) [1].

Data sources

For the purposes of this study, I will use StringDB (http://string-db.org/), which is a pre-constructed PPI network. Edges are supported by different evidence, which can include direct observations of interaction, coexpression evidence, and orthology from other hypothesized interacting proteins. Edges are weighted on a scale of 0-1000, which corresponds to the types and relative strengths of evidence supporting the edge. Properties of edge weights are examined further. Additionally, I use small-molecule target data from the DrugBank (http://www.drugbank.ca/) database, and cancer gene data from the Sanger Institute's COSMIC database (http://cancer.sanger.ac.uk/cosmic).

2. Problem definition

I seek to exploit PPI graph properties for tasks relevant to disease treatment and drug discovery. The central theme of this approach is treating disease and drug treatment as graph processes, and trying to exploit projections onto the PPI network to better understand how drugs and disease interact. I decompose this topic into three related specific sub-problems:

1. How do graph properties of disease and drug associated proteins differ from general proteins at different interaction evidence thresholds?
2. Inferring drug target effect as a network propagation process to learn a pathway-specific similarity measure between drugs.
3. Develop a pathway-specific centrality measure that optimizes the edge degree / pathway count ratio, and test the method on known drug target stratification.

3. Related work

Local and global PPI network structural motifs suggest therapeutic strategies

In addition to patterns in edge attributes, PPI graphs contain local motifs that reveal network dynamics and even suggest strategies for targeting network subcomponents with drugs [3], [4]. Specific PPI network motifs of size 3 and 4 are also overrepresented in other evolved or designed networks generated through processes such as hyperlink creation, language formation, and personal social network propagation [4]. Such structures are suggestive of processes like positive and negative feedback loops [3], [4], which have important implications for therapeutic strategies (e.g. which nodes are targeted).

Using PPI network degree centrality hub nodes as candidate drug targets

The presence of central "hub" regulators is a global motif in PPI networks [4], [5], [8]. Such nodes make especially attractive drug targets, because they are often central to multiple biochemical pathways involved in processes like cell proliferation [11]. In social networks, nodes with high centrality can be called "central individuals," and are important to graph propagation processes, such as gossip [5], [14]. One processes for estimating diffusion centrality of individuals in gossip networks that is robust to partially instantiated edges and hidden nodes provides strong predictive power without direct knowledge of global network structure [5].

These papers present PPI networks as a power-law graph type that are generated by inferring interactions from experimental data [1], [2], [3]. Specific network components, such as central actors, are of particular interest because their graph theoretic properties make them effective drug targets [3]. Certain motifs are overrepresented both within local subgraphs [4], and in global PPI networks [1], [2], [3]. A recent method for detecting central nodes in gossip networks [14] is potentially useful for the same task in PPI networks because they have similar structural motifs [2], [14], and similar partial topological instantiations [1], [2], [14]. The metric I propose in section 4.2 is a direct extension of the gossip process centrality in [14].

Using pathway information jointly with PPI networks

PPI networks represent generalized protein interactions; often they do not confer specific functions to edges. Pathways represent more structured biochemical knowledge, but the total size of known pathway annotations are much smaller [8], and are biased towards interactions of studied proteins (as opposed to PPI discovery techniques such as mass spectrometry, which are thought to have approximately uniform discovery rates) [4]. Works such as [8] do use both global PPI information and pathway knowledge, but metrics are treated separately: candidate targets are found by filtering PPI hubs based on pathway memberships. Existing works on joint pathway/PPI drug discovery generally do not represent pathway knowledge as a graph process on a PPI network; they treat the two as orthogonal data sources and attempt to fit a joint discovery model.

4. Approach

4.1 – Network properties of drug and disease associated proteins

There are approximately 20,000 unique proteins in the human proteome, but fewer than 10% are implicated in disease or drug associations. Therefore, a key question is how do network properties of disease and drug associated proteins differ from the proteome-wide distribution. To address this, I analyze graph properties of protein by drug/disease association.

Threshold analysis

The full StringDB PPI network contains > 8 million edges, but the majority are only supported by weak e vidence and have low edge weights (see analysis). To compensate, I use variable thresholds to subsample the network to more significant evidence levels. W is a sorted list of edge weights from the set of edges E, and T is the vector of n unique thresholds:

$$T_i = W_{[i \cdot \frac{|E|}{n}]}$$

Then, the binary undirected graph G_i with threshold T_i applied is:

$$G_i = G(V, E : W_E > T_i)$$

Centrality measures

I will use three centrality measures to understand the network properties of disease and drug associated proteins: degree centrality (DC), eigenvector centrality (EC), and Katz centrality (KC). For graph G_i, the centrality measures for protein j are:

$$DC_{ij} = \frac{1}{|V|}\sum_{k}(G_i)_{jk}$$

$$EC_{ij} = \frac{1}{\lambda}\sum_{k}(G_i)_{jk} \cdot EC_{ik}$$

$$KC_{ij} = \sum_{c}^{\infty}\sum_{k}\alpha^c(G_i^c)_{jk}$$

Where α is a fixed attenuation factor, used as $\alpha = 0.1$ in this work.

4.2 – Modeling drug target effect as a network propagation process

There are ~1400 unique approved drugs that each have one or more known targets. A drug d_j can be viewed as a projection onto the graph G_i – it is a binary vector that encodes the influence of drugs onto the set of proteins in the human proteome. Moreover, the effect of drugs are not limited to the proteins they directly target – they are propagated to proteins that the targets interact with. Using model parameter $0 \leq \rho < 1$ as the propagative effect (which could be interpreted as the transmission probability or equivalently as a signal attenuation factor), we can describe the drug effect DE_{ij} of drug d_j on graph G_i as:

$$DE_{ij} = \sum_{c=1}^{T} d_j(G_i)^c\rho^{(c-1)}$$

We also introduce a parameter T that describes the maximum number of edges through which drug effects can propagate. This is equivalent to a gossip process, but we do not limit the number of signal origin points (each protein targeted by drug d is considered a signal origin point).

4.3 – Pathway-specific centrality measure for drug target discovery

Proteins in the network have functional annotations, such as kinase activity, metabolism, or gene regulation. Often, the desired targets of drugs are functional processes / pathways, not specific proteins. Identifying a particular candidate drug target in a pathway is non-trivial, because pathways often contain hundreds of proteins. One must balance target centrality with effect survivability: if a target is not sufficiently central, the disease can likely compensate for the drug effect (e.g. acquire resistance), but if the target is systemically central (e.g. a regulatory protein involved in many non-target pathways), drug interference can cause dangerous off-target effects. Thus, it is desirable to target proteins that are central to a particular pathway, but that don't necessarily have comparatively high global centralities.

For this purpose, I introduce the idea of pathway-specific centrality (PSC). PSC is a measure on a protein j and a pathway l that describes the centrality of the protein within the pathway:

$$PSC_{ijl} = \left(\frac{1}{N}\sum_{j} G_{ijk}\right) - \left((\alpha\sum_{l} P_{jl})\,(\beta\sum_{k} G_{ijk}\,P_{kl})\right)$$

P is a matrix that describes the memberships of each protein onto each pathway l. Additionally, α and β are parameters that determine the strength of the self and first degree pathway membership terms. Intuitively, this metric is equivalent to scaling the degree centrality (DC) by a penalty term that is the product of the number of pathways a particular protein is in, and the number of pathways its first degree connections are in. Proteins with a high number of connections spread over a small number of unique pathways score high; proteins with diffuse connections over a large number of pathways are penalized.

For this project, protein functional annotations (i.e. Gene Ontology terms) are treated as pathways. While other pathway membership metrics also exist (e.g. KEGG), this is a simplification to keep the dimensionality of the P matrix relatively small. Additionally, only the top half of GO terms are selected (the GO terms with the most annotations in the dataset).

5. Results

5.1 – Network analysis of disease and drug associated proteins

Edge score distribution

The edge scoring method used by StringDB assigns particular weights to particular categories of evidence. Some types of evidence have piecewise binary weights: an interaction with any experimental evidence is given a score boost of +600, whereas co-expression evidence is based on the number of gene expression datasets supporting co-expression. Overall the underlying distribution is roughly exponential distributed, but contains clear peaks that are due to specific evidence type scores.

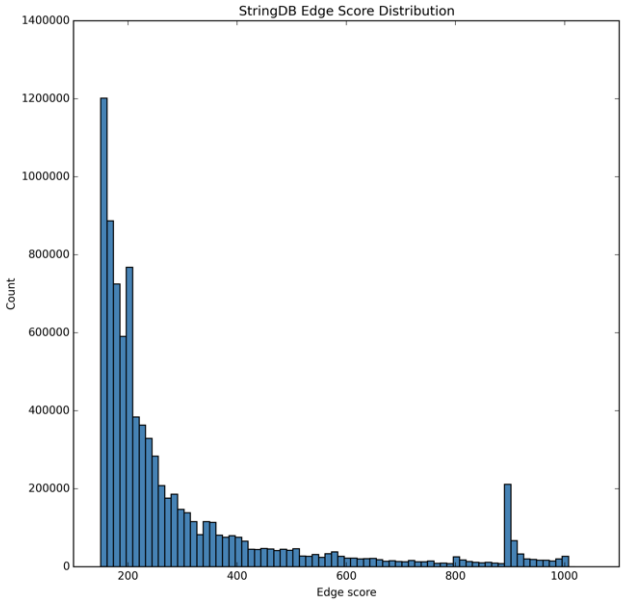

Figure 1: Edge score distribution of StringDB entries

Edge thresholding

I sought to find representative points in the edge thresholding where the graph is no longer fully connected, but still has larger connected components. The graph begins coming disconnected at threshold 500, and is fully disconnected around threshold 1000. For further analysis, I selected thresholds in the range of 550 – 950 in 100 point increments. In particular, the graph with threshold 800 has interesting properties: around 5000 connected components, and an average node degree of 20.

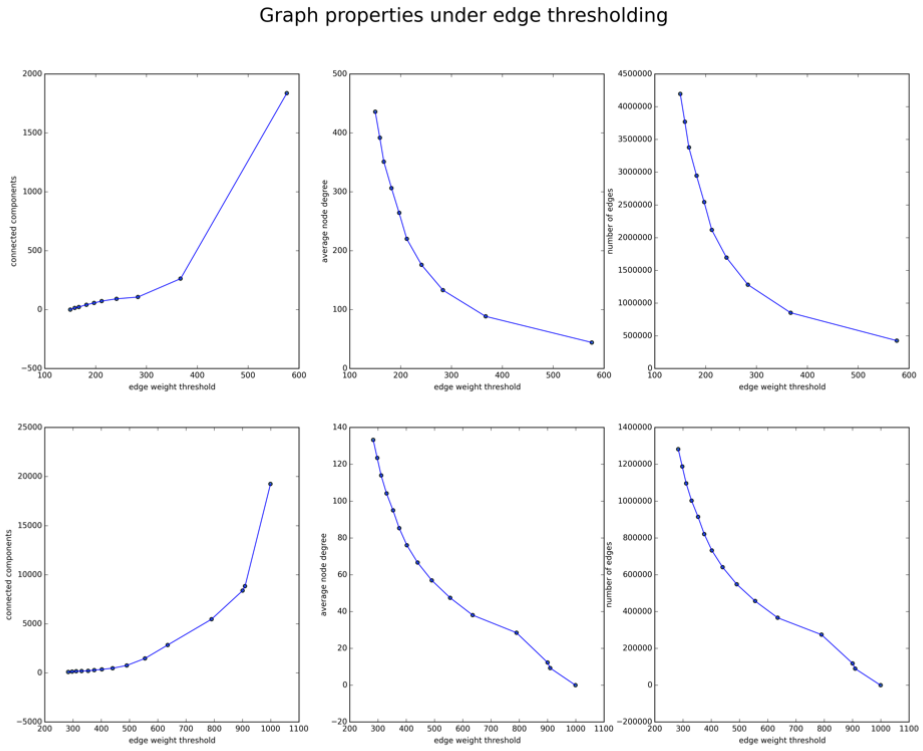

Figure 2: Network properties under edge thresholding shows graph becoming disconnected around w=400, degree decreases approximately exponentially until w=800.

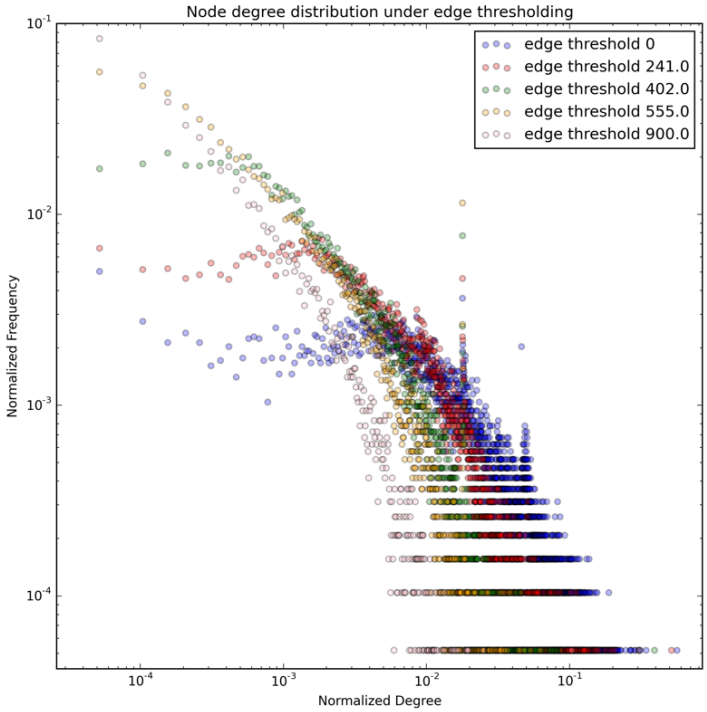

Figure 3: Node degree distributions under thresholding shows graphs with higher edge weight thresholds are closer to power-law distributed.

Network analysis of disease and drug associated proteins

I sought to understand network structural differences between proteins that are disease or drug associated and proteins that are not. I used the three centrality measures described above (degree centrality, eigenvector centrality, and Katz centrality). Cancer associated proteins were split based on whether they were implicated in cancer progression (for any type of cancer). Drug associated proteins are any protein that has been described as the primary target of any currently FDA approved drug. Sample sizes of both sets of proteins and the full set of proteins are provided in the figure captions.

Both cancer and drug associated proteins showed significantly higher degree centralities than non-cancer or drug associated proteins. I hypothesize that there are two separate bases for this fact: first, that many of these proteins are more central to the underlying PPI network than average proteins (making them valuable cancer drivers and/or drug targets), and secondly that these proteins are more intensely studied than others. As an extension of the current work, it would be interesting to build a regularized model that factors into account the total number of studies described for a protein as an approximation of how 'studied' it is – which could in turn be used to estimate the edge discovery rate for each protein, and to normalize degree bias of highly studied proteins (e.g. scaling degree centrality by the estimated edge discovery rate).

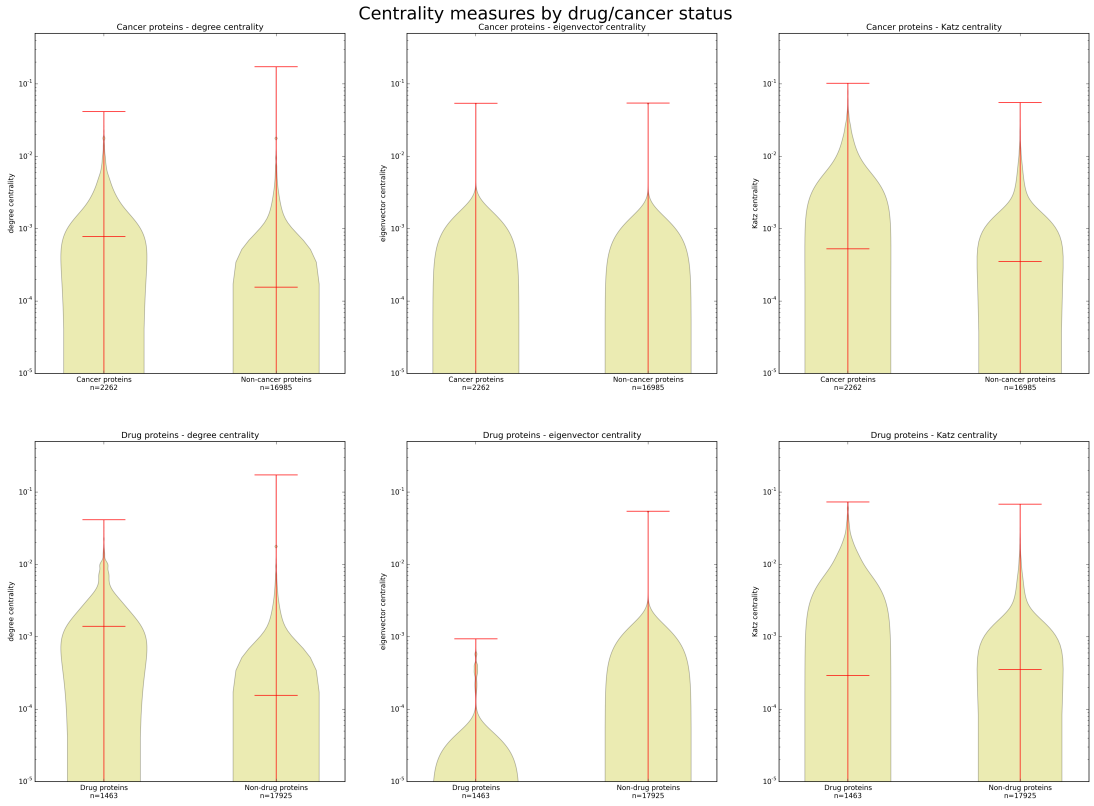

Figure 4: Centrality distributions by drug/cancer protein association shows significant differences in degree, eigenvector, and Katz centrality with drug/cancer association.

5.2 – Modeling drug effect as a propagative process

I implemented the propagation model described in section 4.2, with T=2 and ρ=0.7. This produces an embedding of the drugs onto the protein space: it represents the drugs using the proteins that each drug targets, and an attenuation cascade model of the proteins known to interact with the targets. The density of the embeddings is largely dependent on the threshold used for edge weights: networks with low edge weight threshold give very dense embeddings, high edge weight thresholds give sparse embeddings.

For this analysis, I selected an edge weight threshold of 700, which propagated non-zero signals to > 0.65 of the proteins. I used PCA and t-SNE to visualize the projections. Both plots show that the obtained embedding can be used to robustly separate the drugs, suggesting the propagative network embeddings are effective similarity measures for drugs, even when there is poor overlap between the drugs known targets (approximately half of the 100 drugs shown have unique targets).

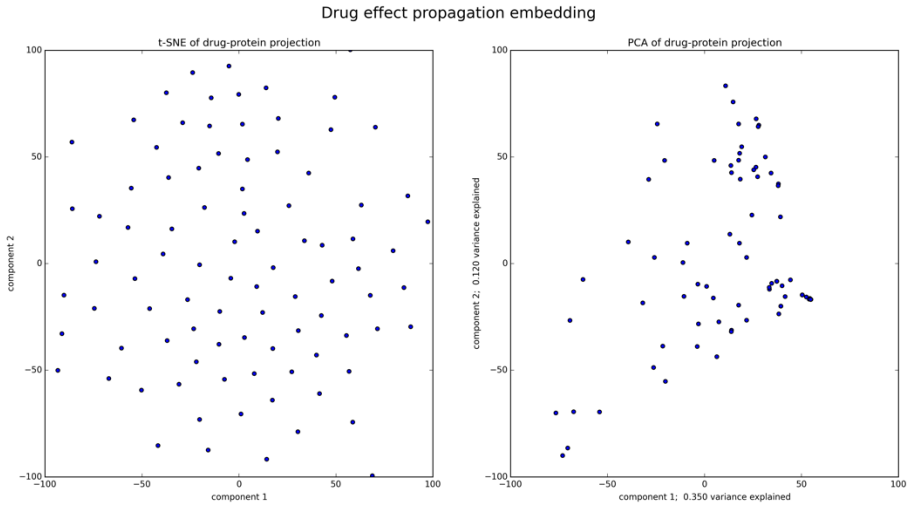

Figure 5: Dimensionality reduction on the propagated drug effect embedding shows it can be used to effectively separate drugs in both non-linear manifolds (left) and linear manifolds (right)

5.3 – Pathway-specific measure of protein centrality

I implemented the pathway-specific centrality measure described above, which is a combination of degree centrality and factors from a first and second degree protein pathway embedding matrix. After experimenting with different hyper parameters, I chose to analyze the metric with $\alpha = 0.25$ and $\beta = 0.15$, which weights first degree embeddings higher than second degree embedding terms (intuitively, the centrality measure gives more weight to the number of pathways that a particular protein is in than to the number of pathways its first degree neighbors are in.

To visualize the contribution of each of the terms (the degree centrality and the pathway embedding term) to the pathway-specific centrality measure, I plotted the measure and its basis components over a set of ~10,000 proteins:

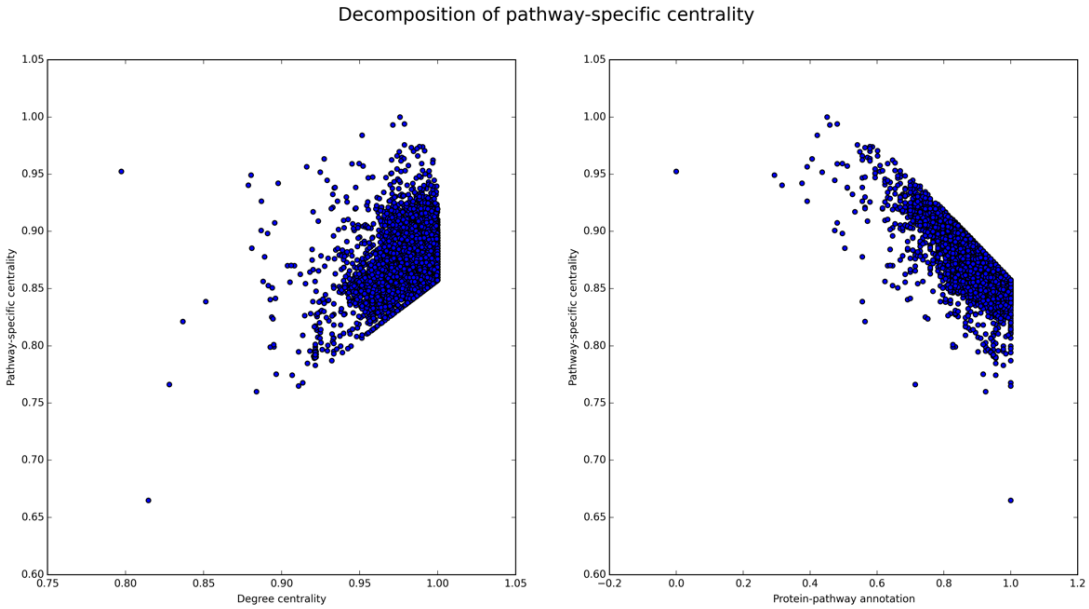

Figure 6: Decomposition of pathway-specific centrality shows positive correlation to degree centrality and negative correlation to normalized protein pathway degree.

The purpose of the pathway-specific centrality metric is to identify candidate drug targets: proteins with many connections in a limited set of pathways, but not proteins whose functions are so diverse that their regulation/inhibition is likely incompatible with cellular function. As a calibration of the metric for this purpose, I compared the PSC score with the number of FDA approved drugs targeting the protein:

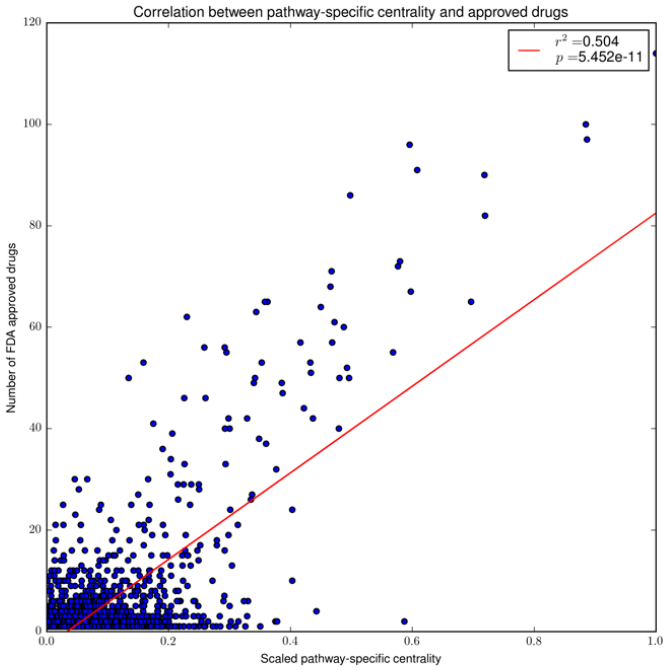

Figure 7: Significant positive correlation between PSC and approved drugs suggests PSC is able to stratify targets by properties that are relevant for drug development.

3. Conclusions

PPI networks become closer to power law distributed under edge thresholding

Previous works have stated PPI networks are thought to be approximately power-law distributed, however the raw StringDB network contains a significant bias towards intermediate degree nodes and is not power law distributed (Figure 3). Under edge thresholding, the network degree distribution becomes very close to power-law distributed. This is suggestive that thresholding removes bias present in the StringDB network: a larger relative proportion of the intermediate frequency edges are removed under thresholding, and these are also the same edges that skew the original network from a power law distribution.

There are significant differences in network properties of disease and drug associated proteins

Disease and drug associated proteins have significantly higher degree centralities (Figure 4). Interestingly, drug proteins have significantly lower eigenvector centrality than non-drug proteins but higher Katz centrality than non-drug associated proteins; and cancer-associated proteins do not display this tendency.

Drug targeting can be embedded as a propagative signal on the protein network space

The gossip / propagative model proposed in 4.2 is an effective means of stratifying different drugs over both linear and non-linear manifolds (Figure 5). This method represents the differences in drugs based on the extended interaction networks of their targets. In a further study, it would be interesting to compare classes of drugs and other measures of drug chemical similarity with this approach.

Pathway-specific drug centrality is correlated with target 'drugability'

Figure 7 shows that the PSC metric is linearly correlated with the number of FDA approved drugs for a given target. Given that different proteins have different 'drugability' (i.e. different chemical and biological characteristics that make them relatively easier/more difficult to develop drugs for), and that the number of unique approved drugs for a given target is suggestive of this difference, PSC is arguably also correlated with a hypothetical 'drugability' metric. Difference in drugability is highly important for pharmaceutical companies, who seek targets that are amendable to low-risk drug development.

[1] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant, "Gaining confidence in high-throughput protein interaction networks," *Nat. Biotechnol.*, vol. 22, no. 1, pp. 78–85, Jan. 2004.

[2] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, "Evidence for dynamically organized modularity in the yeast protein–protein interaction network," *Nature*, vol. 430, no. 6995, pp. 88–93, Jul. 2004.

[3] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein–protein interactions in yeast," *Nat. Biotechnol.*, vol. 18, no. 12, pp. 1257–1261, Dec. 2000.

[4] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore, S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J.-P. Lambert, H. S. Duewel, I. I. Stewart, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S.-L. Adams, M. F. Moran, G. B. Morin, T. Topaloglou, and D. Figeys, "Large-scale mapping of human protein–protein interactions by mass spectrometry," *Mol. Syst. Biol.*, vol. 3, p. 89, Mar. 2007.

[5] D. R. Rhodes, S. A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. M. Chinnaiyan, "Probabilistic model of the human protein-protein interaction network," *Nat. Biotechnol.*, vol. 23, no. 8, pp. 951–959, Aug. 2005.

[6] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, "Towards a proteome-scale map of the human protein–protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, Oct. 2005.

[7] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker, "A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome," *Cell*, vol. 122, no. 6, pp. 957–968, Sep. 2005.

[8] D. C. Altieri, "Survivin, cancer networks and pathway-directed drug discovery," *Nat. Rev. Cancer*, vol. 8, no. 1, pp. 61–70, Jan. 2008.

[9] P. K. Brastianos, S. L. Carter, S. Santagata, D. P. Cahill, A. Taylor-Weiner, R. T. Jones, E. M. V. Allen, M. S. Lawrence, P. M. Horowitz, K. Cibulskis, K. L. Ligon, J. Tabernero, J. Seoane, E. Martinez-Saez, W. T. Curry, I. F. Dunn, S. H. Paek, S.-H. Park, A. McKenna, A. Chevalier, M. Rosenberg, F. G. Barker, C. M. Gill, P. V. Hummelen, A. R. Thorner, B. E. Johnson, M. P. Hoang, T. K. Choueiri, S. Signoretti, C. Sougnez, M. S. Rabin, N. U. Lin, E. P. Winer, A. Stemmer-Rachamimov, M. Meyerson, L. Garraway, S. Gabriel, E. S. Lander, R. Beroukhim, T. T. Batchelor, J. Baselga, D. N. Louis, G. Getz, and W. C. Hahn, "Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets," *Cancer Discov.*, Sep. 2015.

[10] K. Chin, C. O. de Solorzano, D. Knowles, A. Jones, W. Chou, E. G. Rodriguez, W.-L. Kuo, B.-M. Ljung, K. Chew, K. Myambo, M. Miranda, S. Krig, J. Garbe, M. Stampfer, P. Yaswen, J. W. Gray, and S. J. Lockett, "In situ analyses of genome instability in breast cancer," *Nat. Genet.*, vol. 36, no. 9, pp. 984–988, Sep. 2004.

[11] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nat. Chem. Biol.*, vol. 4, no. 11, pp. 682–690, Nov. 2008.

[12] S. Fröhling, C. Scholl, R. L. Levine, M. Loriaux, T. J. Boggon, O. A. Bernard, R. Berger, H. Döhner, K. Döhner, B. L. Ebert, S. Teckie, T. R. Golub, J. Jiang, M. M. Schittenhelm, B. H. Lee, J. D. Griffin, R. M. Stone, M. C. Heinrich, M. W. Deininger, B. J. Druker, and D. G. Gilliland, "Identification of Driver and Passenger Mutations of FLT3 by High-Throughput DNA Sequence Analysis and Functional Assessment of Candidate Alleles," *Cancer Cell*, vol. 12, no. 6, pp. 501–513, Dec. 2007.

[13] T. 1000 G. P. Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, Oct. 2010.

[14] A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson, "Gossip: Identifying Central Individuals in a Social Network," *ArXiv14062293 Phys.*, Jun. 2014.