Biome Distributions

Linas Vepstas

23 January 2020

Abstract

An exploration of the statistical distribution of interactions between genes, proteins and pathways, extracted from public genome, proteome and reactome datasets. This is an exploratory diary that attempts to understand the large-scale statistical properties of the network structure of bio-molecular interactions.

It was naively hypothesized that genome/proteome reaction pathways form a scale-free network, and thus would have a Zipfian distribution. Much to our surprise, this is not the case! It seems like *everything* follows a square-root Zipfian distribution! I do not know of any network theory or biology theory that would explain this, so it is a surprise.

An exploration of the mutual information of interaction pathways is also performed. It appears that these are easily fit with a bimodal Gaussian distribution.

Introduction

Publicly available genomics and proteomics databases describe a large number of interactions between genes and pathways. Taken together, these data-sets describe a large graph, and the one may reasonably wonder about the properties and general structure of that graph. For example, one might ask if the graph is scale-free, or if it has a hub-and-spoke structure, or make other graph-theoretic inquiries into it.

In this short monograph, a study is made of "triangles": three genes that mutually interact, and "pentagons": a pair of genes that express a pair of proteins that lie on a common pathway. If the genome (proteome, reactome) network were scale-free, then one might expect Zipfian distributions of triangles and pentagons.[1] This appears not to be the case. A further study is made of the mutual information ("lexical attraction") of interacting pairs. It appears to be easy to describe this as a pair of Gaussians. A theoretical grounding for these results is unknown to the author. Clarifications are solicited.

[1] These questions originally arose during the characterization of a bioinformatics data-mining benchmark, found in `https://github.com/opencog/benchmark/query-loop`, and was elaborated in the "Genome distribution!?" email discussion.

Datasets

The graph networks explored here were constructed from public data-sets from MCBI, ChEBI, PubMed, UniProt, SMPDB, Entrez and BioGrid. Two variant networks were constructed and explored. The two differ in how gene interactions were treated. In the first, genes that regulate one-another were treated as directed edges, in that gene A may regulate gene B, but not the other way around. This includes genes that may self-regulate. The second network was obtained from the first by symmetrizing all gene interactions (if A interacts with B, then B interacts with A) and removing all self-interacting genes.

Both graphs contain a total of 49050 genes. There are 540778 gene interactions in the first data-set, which are represented as directed edges. Of these, 347127 are symmetric (*i.e.* have a partner indicating the opposite direction). There were 2939 self-interacting genes. From this, we deduce a total of (347127 - 2939) / 2 = 172094 symmetric interactions, and 540778 - 347127 = 193651 non-symmetric interactions. The number of edges that have non-zero counts (i.e. appeared at least once in a triangle) is 455572.

For the second data-set, the self-interacting genes are removed, and matching symmetrized edges are created, for a total of 731490 directed edges, or half of that, 365745, when the relation is taken as symmetric.[2] The number of edges with non-zero counts (i.e. the number of edges that appear in at least one triangle) is 617530.

Pathways are from https://reactome.org and from http://smpdb.ca/

Triangles

A triangle is defined as a three-way interaction, of gene A regulating gene B, gene B regulating gene C and C regulating A. Such triangles are one of the most basic components of the topological structure of regulatory networks. It is worth understanding their structure.

Any given gene may appear in multiple triangles; clearly, some genes might participate in many, and some in only a few. So, the first statistical question becomes: what is the distribution of these genes? This is answered via a Zipf-style graph: one makes a list of all of genes, ranking them according to the number of triangles they participated in, and then graphing, on a log-log graph, the number of triangles vs. the rank. Similarly, one can pose the same question, but for edges: how many triangles does a given gene-gene interaction participate in?

By counting the number of times an edge participates in some triangle, one obtains a count $N(g_a, g_b)$ that relates gene g_a to g_b. Given the counts of such pairs, this opens the door to a stable of standard statistical questions: what are the marginal distributions, the entropies, and the mutual information of such pairs? This is explored in a subsequent section.

Because these explorations are perfomed on curated data-sets carried out by research labs, one has an open question of whether one is measuring "true biology", or

[2] Use the `(count-gene-interactions)` tool.

whether one is measuring the social network of the scientists, and their intellectual interests. There does not appear to be any obvious way of answering this question.

Characterization

The two data-sets provide sometimes-similar and sometimes-different results. The differences are noted as appropriate.

In the first data-set, there are 3452807 pointed triangles. By definition, a "pointed triangle" is one with a "distinguished point"; this helps with the tracking of rotational symmetry. By symmetry, one would expect this number to be divisible by 3; it is not because many of the triangles are degenerate: There are 145571 degenerate triangles having the only two distinct vertexes, and 2939 "bouquets" - that is, triangles where all three corners are the same vertex.[3] This leaves behind 1102412 distinct triangles having three distinct corners.

A total of 18766 genes participate in these interactions.[4] This differs from the number of genes in the data-set, as not all genes participate in triangles. There are 455572 distinct edges that appear in triangles. These edges appear a grand-total of 5050388 times in different triangles; thus, on average, any given gene interaction might appear in 11 different triangles. However, the concept of "average" is troublesome in Zipfian distributions; this and other marginals are expanded on in a subsequent section.

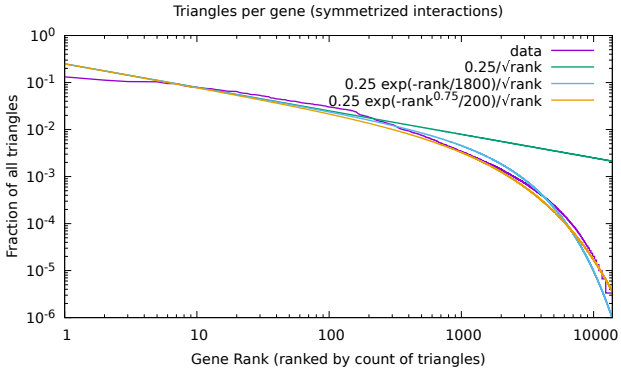
The presence of such degenerate triangles, and the non-reciprocal gene-gene interactions leads to contemplating the second data-set. In this data-set, all self-interacting genes are removed, and all gene interactions are symmetrized, so that for every edge AB there is also an edge BA: the concept of regulation is replaced by interaction. The second data-set has 10783686 pointed triangles; by symmetry, only 1/6 of these are distinct, leaving 1797281 distinct triangles. These triangles have a total of 731490 distinct edges. There is a total of 20123 genes participating in these triangles.

Zipf graphs

The following figure shows the number of triangles in which a gene participates in. This figure if for the symmetrized network; the corresponding figure for the unsymmetrized network is very similar.

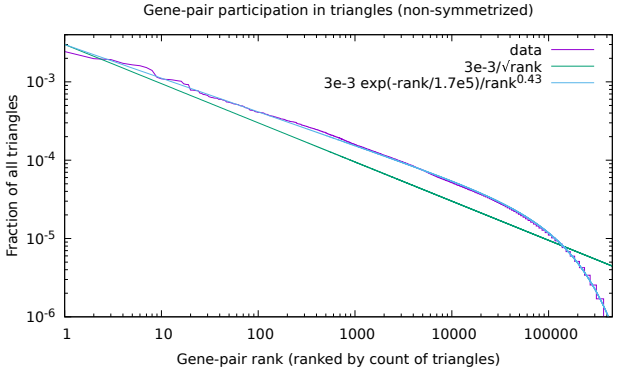
[3] Given a triangular relationship of the form "gene A regulates gene B, B regulates C and C regulates A," a degenerate triangle is obtained when B and C are the same gene. A bouquet is obtained when A, B and C are the same gene.

[4] Counted with `(length loop-participants)`.

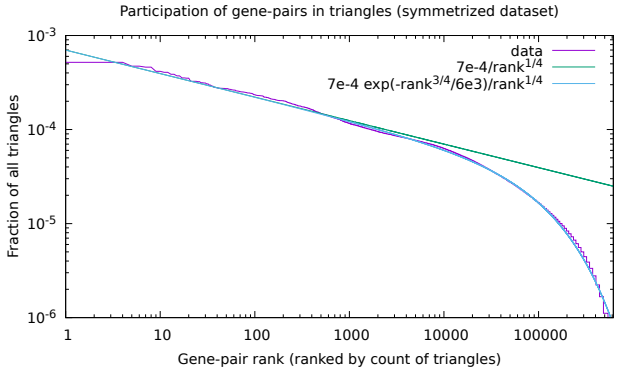


The first unexpected surprise is that this graph is not Zipfian. That is, the distribution is not of the form $1/(rank)^{\alpha}$ with $\alpha \approx 1$. Just eyeballing the general shape, as done above, one promptly arrives that $\alpha = 0.5$ for the leading part of the figure. The tail is certainly not Zipfian, either; it falls of sharply. A first guess, no better than voodoo numerology, suggests that the n-th ranked gene appears in $p(n) = 0.25e^{-\beta n}/\sqrt{n}$ of all triangles, for some constant β. Closer examination suggests $p(n) = 0.25\exp(-\beta n^{3/4})/\sqrt{n}$ provides a better fit. These forms are chosen to simply look appealing to the eye, rather than by minimizing some least-squares curve-fit. These seem like reasonable hypothesis, given the above graph. However, the situation gets a bit confused, when one looks at edges.

One may also ask how many triangles a given edge (a given gene-pair) participates in. A similar distribution results, although the suggestive $\sqrt{n}$ hypothesis does not seem to hold cleanly. As shown in the figure below, the exponent $\alpha \approx 0.43$ appears to offer a better fit. This figure is for the unsymmetrized data-set.



The symmetrization of the edges makes a dramatic change to the distribution exponent. The exponent now appears to be $\alpha \approx 0.25$, as shown in the figure below.



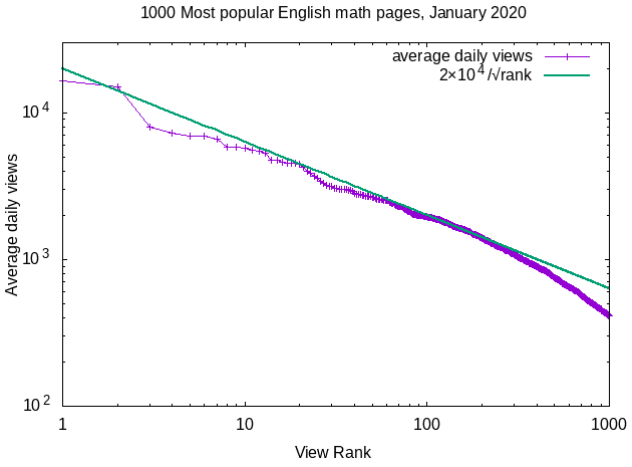
That such a technical change would have a rather profound effect on the structure of the graph is surprising.

Prior Results

Hmm. Apparently, I've seen square-root Zipfian distributions before, in language data. I'd forgotten about that. See page 43 of this PDF: https://github.com/opencog/learn/raw/072f0a46d143d94fc27cb5a04dd8 lang-diary/connector-sets-revised.pdf

Related Results

Page views of the most popular pages on Wikipedia show the same square-root scaling. See https://en.wikipedia.org/wiki/Wikipedia:Does_Wikipedia_traffic_obey_Zipf%27s_law%3F for a demonstration. Here's a graph from this:



This shows the distribution of the 1000 most commonly viewed English-language Wikipedia Mathematics articles, as of Janurary 2020. A square-root distribution appears to hold up to at least a rank of 200 or 300; after that, the behavior is unclear.

TODO

What is the degree distribution? Viz, for scale-free networks, one usually examines the degree distribution, and not the triangle distribution. What about deg(u).deg(v) for edges connecting u,v?

Mutual Information

Mutual information (MI) is an entropic concept that characterizes the degree to which a pair of objects associate with one-another. For any given pair, one may examine how often the members of that pair occur together, versus how often each member associates with other, third parties. It's "entropic", in the sense that specific sum and difference of marginal entropies, and thus fits naturally into theoretical frameworks founded on entropic concepts.

The raw genome databases do not provide any particular gene-pairing information, beyond a yes/no assertion that a pair has been observed to interact in some laboratory setting. By pattern-mining triangles, one gets far more detailed interaction information. So, again, each triangle has three corners, the genes, and three edges, the gene pairs. By counting all possible triangles, one obtains a count of how often each edge is seen -- that is, how often an edge appears in some triangle. That is, one has a count $N(g_a, g_b)$ of the number of triangles that the pair (g_a, g_b) participated in. The rest is downhill. MI is defined as usual as Yuret's lexical attraction.

Given counts $N(x,y)$, one can readily compute the lexical attraction between the elements (x,y) as

$$MI(x,y) = \log_2 \frac{N(x,y)N(*,*)}{N(x,*)N(*,y)}$$

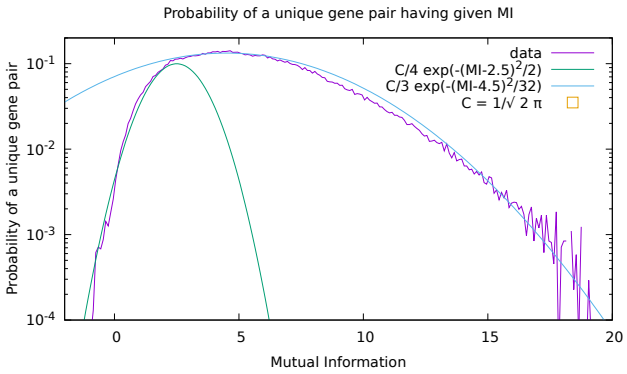
The total MI for a data-set is defined as

$$MI = \sum_{x,y} p(x,y) MI(x,y)$$

where $p(x,y) = N(x,y)/N(*,*)$ is the (frequentist) probability of observing the pair (x,y). For the symmetric data-set, it is 3.49. The total entropy for the data-set is 18.21. The sparsity is 9.36, where the sparsity is defined as $\log_2$ of the number of non-zero entries in the matrix. In this case, there are 20123 genes that participated in triangles, and 617530 non-zero entries (out of 404935129 possible; the $\log_2$ of this ratio provides the sparsity).

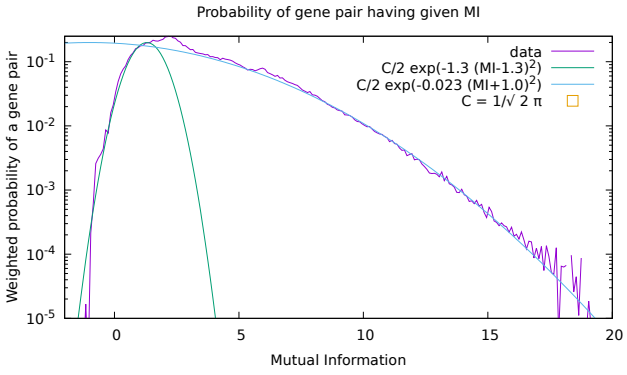
The distribution of the MI as a function of pair frequency is an interesting distribution to explore. There are two variants of this distribution that one can examine. One may consider only unique gene-pairs, or one may consider pairs weighted by their counts. So, in the first case, if two gene-pairs were found to have the same MI, then they are treated as equiprobable; when binned into a histogram, they are placed in the same bin with equal weighting. In the second case, their contribution to a bin is relative to the number of times each was observed. Equivalently, one may also say that one is counting "without and with degeneracy", or is examining "unweighted and weighted distributions".

The figure below shows the first case: the distribution of MI, counted without degeneracy (all gene pairs are treated as equi-probable).



The normalization is such that the area under the data curve will integrate to 1.0. The data is fit with two gaussian distributions. Recall that gaussians appear to be parabolas on a semi-log plot. The parameters of each fit are as marked in the legend. The fits are hand-built, or "eyeballed", chosen to look good, and are *not* the result of some automated curve-fit. The fit parameters are "numerology": rounded to integers that look pleasing. Whether or not the integers are meaningful or are incidental is unclear.

The second case, where the bin-counts are weighted by frequency, results in a graph that is similar. Again, it is easily fit with two gaussians, but this time, the centers of the gaussians are quite different, as are the widths.



As before, the fit is eyeballed, with a preference of finding numerological simple-fraction fits, so as to expose simple laws, if these exist.

Pentagons

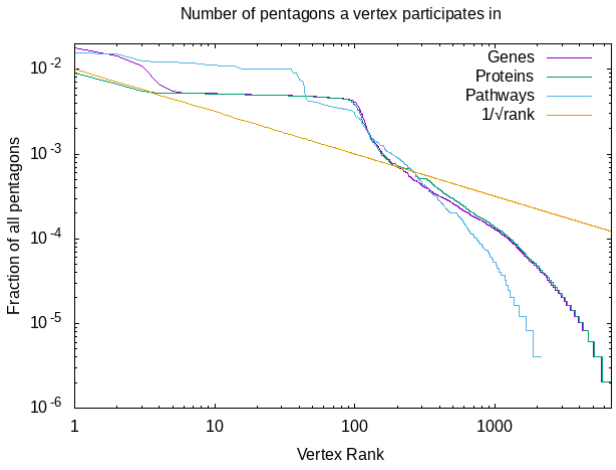
The pentagon has five vertexes: a pathway (as a single, named vertex), two proteins on that pathway, two genes, each of which express the corresponding protein, and

finally, the two genes interacting with each other. Here, there are five edges, and three edge types: (protein in pathway), (gene expresses protein) and (gene-pair interaction). Thus, characterizing this statistically requires six Zipfian graphs: three for the different vertex types, and three for the different edge types. Each edge type will also have an MI distribution, and so six MI graphs: one for each type, weighted and unweighted. So, twelve graphs in total. Not all prove to be that interesting.

There are 50566 pathways in which some protein in the data-set is a member of. These are both SMP and R-HSA-tagged pathways (from the smpdb.ca and reactome.org, respectively). Each of these is observed in at least one pentagon. There is a total of 1082860 pathway-protein pairs in the data-set, but only 38843 appear in a pentagon. There are a total of 148734 gene-protein expression pairs, of which only a paltry 6794 appear in a pentagonal relationship. As before, there is a total of 365745 pairs of interacting genes; of these only 77304 appear in pentagons. Thus, the data-set, although quite large, appears to be quite disconnected: there are many interactions that have been mapped out, but these remain sparse enough that there are relatively few of these pentagonal interactions. This is presumably an artifact of the scientific exploration done to date, rather than an innate feature of actual biology. That is, there are presumably far more interactions, but these remain unknown and uncharted in the data-set.

Distributions of Vertexes

Pattern mining the above-described pentagon resulted in the observation of 491558 distinct pentagons (taking into account mirror symmetry: there are twice as many, if the mirror image is counted separately.) There are 6694 unique genes, 6735 unique proteins and 2129 distinct pathways that appeared in pentagonal relationships. All have an unexpected distribution; they are graphed below, on the same graph.[5]



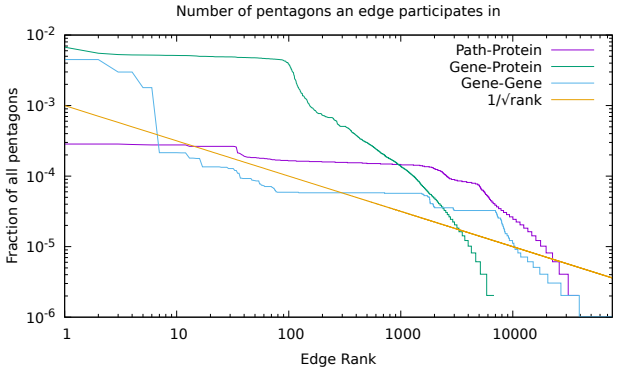
[5]Per contents of the `graphs/pentagon-paths` directory.

The normalization is with respect to the total number of pentagons that a given vertex appears in. Thus, for example, the highest-ranked gene, and the highest-ranked pathway appears in approximately 1 out of 50 pentagons; the highest-ranked protein appears in approximately 1 out of 100, which can be read directly off the y-intercept in the above graph.

I cannot come up with any plausible hypothesis for the plateaus, nor for the tails. These may be data-set artifacts: interactions that have been studied sufficiently to have been captured in the data-set. That is, a sociological artifact, rather than a statement about biology.

Distribution of Edges

There were 38843 distinct edges connecting pathways to proteins. There were 6794 distinct gene-protein expression edges. There were 77304 gene-gene interaction edges. The distribution of these edges, *viz.* the number of pentagons a given edge participates in, is shown below.



The overall lumpiness of these graphs provides no particular insight. It seems plausible that the lumpiness is an artifact of the data-set, rather than any indication that nature is structured in this way.

Gene-protein expression

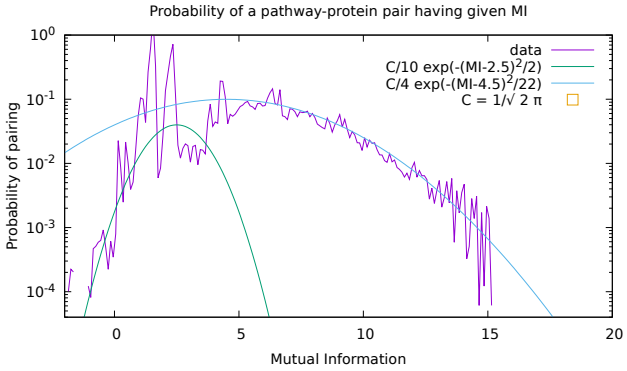
Lets look at gene-protein expression pairs that appear in pentagons. The matrix reports 6694 genes and 6735 proteins. The matrix appears to have 6794 non-zero matrix entries. On average, each gene-expression pair appeared in 144.7 pentagons. For those gene-expression pairs that appear in pentagons, very nearly all genes express just one protein, with only a few exceptions. Of the 6694 genes, there were 20 genes that expressed two proteins, and four that expressed more than two: HLA-B, HLA-A, HLA-C and HLA-DRB1. All remaining genes express just one protein. However, these same genes also participate in a (comparatively) huge fraction of all pentagons.

There is no point in computing MI for gene-protein interactions; they occur in tightly-bound and exclusive pairs; there is no free association.

Pathways

There are a large number of pathway-protein pairs available in the data-set, of which only a small fraction, 38843 pairs, appeared in pentagons. As proteins can appear in multiple pathways, it is meaningful to explore correlations and affinities between these.

Below is the pathway-protein MI graph. It is fit with two Gaussians, as suggested by the previous observations. Note that the parameters are very similar.



It's clearly much noisier, as the set of edges is much smaller. The proteome/reactome data-set although large, seems not as well-developed as the genome data-set.

TODO

weighted in-degree of vertexes?

The End

That's all.