Square-root Zipfian Distributions in Gene Interactions

Linas Vepstas

23 January 2020
revised 30 Nov 2021

Abstract

The distribution of interactions between genes, extracted from public genome databases, is examined. Interactions between pairs of genes, and triangles of three interacting genes, is found to follow a square-root Zipfian distribution with an exponential tail. The mutual information between gene pairs is found to have a bimodal Gaussian distribution. This is an experimental result; no theoretical explanation is proposed.

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 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. If the genome network were scale-free, then one might expect Zipfian distributions of triangles and pentagons. 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.

Datasets

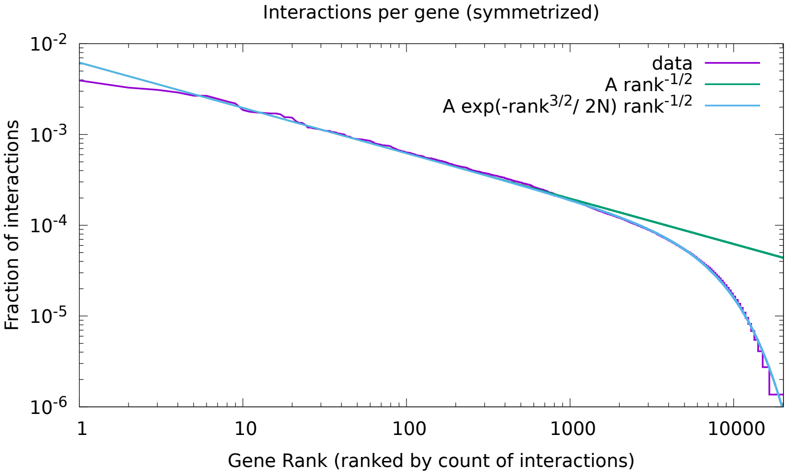
The graph networks explored here were constructed from public data-sets from MCBI, ChEBI, PubMed, UniProt, SMPDB, Entrez and BioGrid. These datasets include information about genes that regulate one-another, genes that express proteins, and proteins that appear on a common reaction pathway.

All gene-gene interactions were taken as symmetric, so that if one gene up- or down-regulates another, the interaction was represented by a symmetric (undirected) edge: namely, that two genes interact, without any specific direction information noted.

The datasets contained a total of 20123 genes participating in 365745+2939 pair-wise interactions. Of these, there were 2939 self-interacting genes; these self-interactions were removed, so as to avoid confusion during counting.

Pair-wise interactions

At the most basic level, one can ask about the distribution of gene interactions. This is straight-forward: one simply crawls over the dataset, and simply counts how many interactions a gene participates in. Ranking the genes by count produces the distribution shown below.



There are 20123 interacting genes participating in a total of $N_{tot} = 365745$ interactions. The vertical axis shows the fraction of all interactions that a given gene participates in. That is, if the k'th gene in the ranking participates in $N(k)$ interactions, then the fraction of interactions is given by

$$p(k) = \frac{N(k)}{N_{tot}}$$

so that the integral under the curve is $N = N_{tot}$.

The first unexpected surprise is that this graph is not Zipfian. That is, the distribution is not of the form $k^{-\alpha}$ with exponent $\alpha \approx 1$. The leading slope is instead given by $\alpha = 0.5$. There is also a pronounced tail that seems too large to ignore or ascribe to statistical rounding. The overlaid fitting curves are “eyeballed”, in that they look visually reasonable during plotting. The best eyeballed fit is surprisingly good, and has a fairly simple expression:

$$p(k) = A\frac{\exp\left(-k^{3/2}/2N\right)}{\sqrt{k}}$$

The normalization A can be solved for exactly in the limit of large N:

$$\int_0^\infty \frac{\exp -\beta x^{-3/2}}{\sqrt{x}} dx = \frac{2}{\beta^{1/3}} \int_0^\infty e^{-u^3} du = \frac{2}{3\beta^{1/3}} \int_0^\infty y^{-2/3} e^{-y} dy$$

The last integral is the gamma function $\Gamma(1/3) \approx 2.67893853\cdots$ and so, based on the eyeballed fit, one has

$$A = \frac{3}{2\Gamma\left(\frac{1}{3}\right)(2N)^{1/3}}$$

The author does not know of any reason why the initial fall-off should follow a square-root slope, or why there should be a simple exponential correction: that is, the theoretical underpinnings and implications for this form is unknown.

Fractal shot noise

The distribution

$$y^{-2/3} e^{-y} dy$$

is reminiscent of a Poisson distribution, but with two key differences. First, for a conventional Poisson distribution, the y is fixed, and not variable: it is the parameter. Second, for a conventional Poisson distribution, the exponent $-2/3$ is a non-negative integer, and is the support of the distribution.

Sums/integrals over Poisson distributions have been proposed as models of 1/f noise as early as 1937.[1, 2] The precise form of the noise depends on the distribution of the “parameter” y; conventional shot noise results from a uniform distribution. The distribution $y^{-2/3}$ seen here most closely resemble models of fractal shot noise.[3]

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.

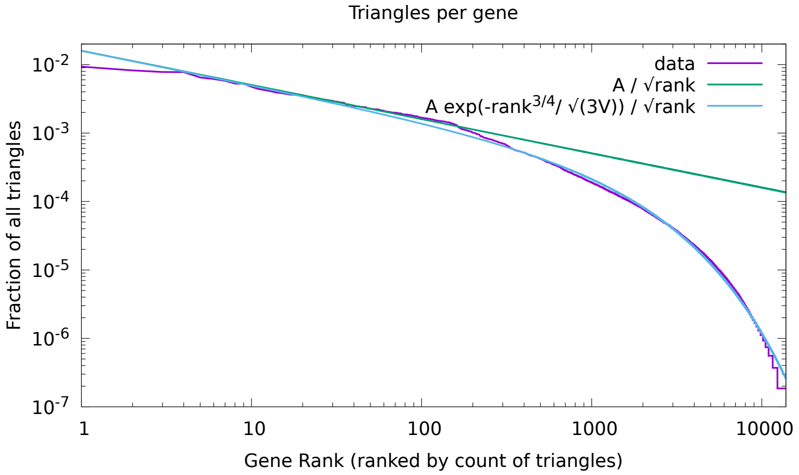
Because these explorations are performed on curated data-sets carried out by research labs, one has an open question of whether one is measuring "true biology", or whether one is measuring the social network of the scientists, and their intellectual interests. There does not appear to be any obvious way of answering this question.

Characterization

Not all of the interacting genes participate in triangular relations; only 13846 of the initial 20123 appear in such relations. In total, there were 308765 distinct pairs participating in triangles (out of 365745 pair-wise interactions in the dataset). A total of 1797281 triangles were observed.

Zipf graphs

The following figure shows the fraction of triangles in which a gene participates in.



The general slope, of a square-root-of-rank, is as before. The eyeballed fit is given by

$$p(k) = Ak^{-1/2} \exp\left(-k^{3/4}/\sqrt{3V}\right)$$

where $V = 13846$ is the number of vertexes (number of genes participating in triangles). Notice that the leading slope is one-half, as before, but the exponential fall-off is not as sharp. It still appears to have a very simple numerological value that is easily guessed. As before, the integral is again in the form of a gamma:

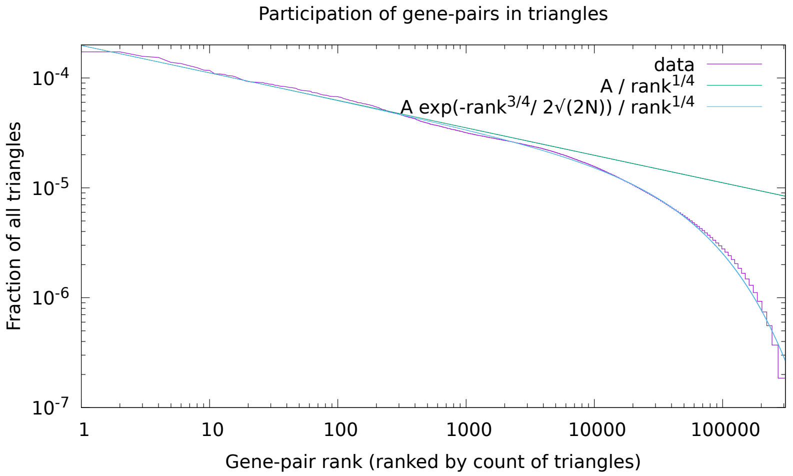
$$\int_0^\infty \frac{\exp -\beta x^{-3/4}}{x^{1/2}} dx = \frac{2}{\beta^{2/3}} \int_0^\infty \exp\left(-u^{3/2}\right) du = \frac{4}{3\beta^{2/3}} \Gamma\left(\frac{2}{3}\right)$$

so that the normalization is

$$A = \frac{3}{4\,(3V)^{1/3}\,\Gamma(2/3)}$$

with $\Gamma(2/3) \approx 1.354117939\cdots$. As before, the fit seems very good, despite effectively being numerological guess-work.

Similarly, one may ask how many triangles a given edge (a given gene-pair) participates in. This is shown below.



The rank exponent is now 1/4, instead of 1/2. The eyeballed fit is given by

$$p(k) = Ak^{-1/4} \exp\left(-k^{3/4}/2\sqrt{2N}\right)$$

where $N = 1797281$ is the total number of triangles. The integral is again in the form of a gamma, but much simpler:

$$\int_0^\infty \frac{\exp -\beta x^{-3/4}}{x^{1/4}} dx = \frac{4}{3\beta} \int_0^\infty e^{-u} du = \frac{4}{3\beta}$$

Thus, the normalization is

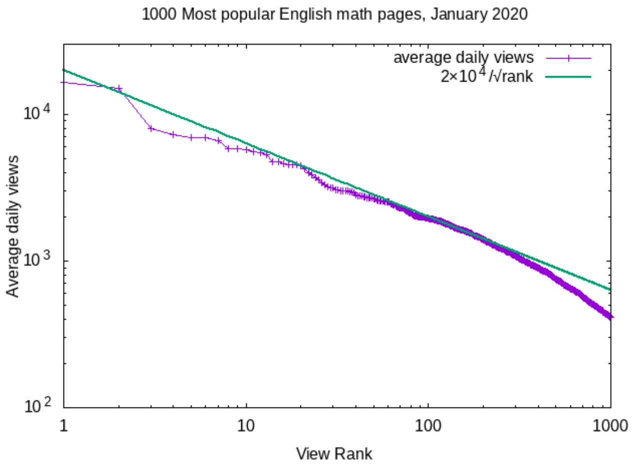
$$A = \frac{3}{8\sqrt{2N}}$$

The preciseness of the numerology is remarkable, but the underlying theory is unknown.

Similar Results

Square-root Zipfian distributions are common. Such a distribution has been observed in the grammatical network of natural language data, that is, in the network of the English language, constructed by counting nearby word-pairs in text corpora.[4] (See page 43)

Page views of the most popular pages on Wikipedia show the same square-root scaling.[5] The result is reproduced below:



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

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.[6]

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

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

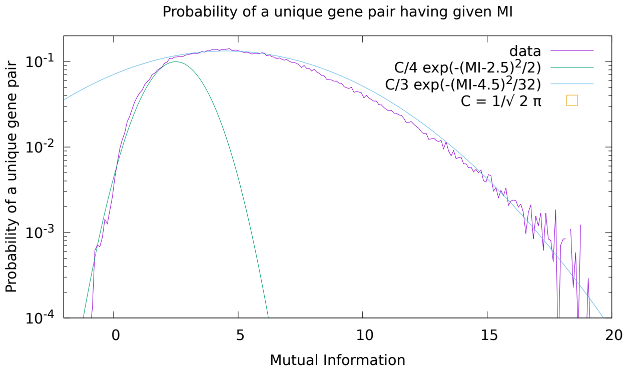
The total MI for a data-set is defined as

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

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).

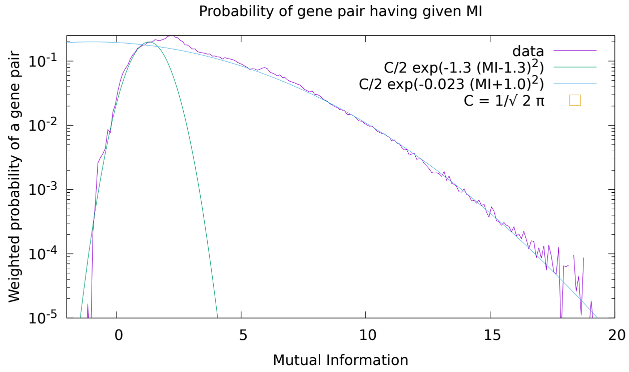
The distribution of the MI as a function of pair frequency is an interesting distribution to explore. There are two variants of this distribution that one can examine. One may consider only unique gene-pairs, or one may consider pairs weighted by their counts. So, in the first case, if two gene-pairs 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.

Conclusion

The distribution of the interaction between genes in publicly-available genetics databases has been examined. The distribution appears to be a square-root Zipfian distribution. The distribution of the mutual information between pairs of genes is described by a bimodal Gaussian distribution.

No theoretical foundation is proposed, although the author imagines that, perhaps, the matrices describing pairwise interactions might be drawn from a Gaussian Unitary Ensemble (GUE) or a Gaussian Orthogonal ensemble (GOE).

References

[1] J. Bernamont, "Fluctuations de potential aux bornes d'un conducteur metallique de faible volume parcouru par un courant", *Ann Phys (Leipzig)*, 7, 1937, pp. 71–140.

[2] Lawrence M. Ward and Priscilla E. Greenwood, "1/f noise", *Scholarpedia*, 2007, URL http://www.scholarpedia.org/article/1/f_noise.

[3] Steven Bradley Lowen and Malvin Carl Teich, *Fractal-Based Point Processes*, Wiley, 2005.

[4] Linas Vepstas, *Connector Set Distributions*, Tech. rep., 2017, URL https://github.com/opencog/learn/raw/072f0a46d143d94fc27cb5a04dd867b88c7142d0/learn-lang-diary/connector-sets-revised.pdf.

[5] Wikipedia, "Does Wikipedia traffic obey Zipf's law?", , 2020, URL https://en.wikipedia.org/wiki/Wikipedia:Does_Wikipedia_traffic_obey_Zipf%27s_law%3F.

[6] Deniz Yuret, *Discovery of Linguistic Relations Using Lexical Attraction*, PhD thesis, MIT, 1998, URL http://www2.denizyuret.com/pub/yuretphd.html.