A case study in large-scale networks: Freebase.com

Benjamin P. Goldenberg

26 April 2009

1 Introduction

Over the last 20 years, large-scale networks have become more and more prominent. Much of this can be attributed to the rise of the internet which has made possible the formation of formal social networks. Before the rise of these large-scale networks, most analysis of graphs relied on being able to draw pictures of graphs. The human brain has an amazing ability to describe structure based on graphical representation. However, there is often no practical way to visualize large-scale networks with millions of nodes and edges. Many of the statistical methods for describing large networks have been developed in order to describe the overall structure of a graph when it cannot be visualized.

For the purposes of this paper, we will define a network as a collection of vertices $V = v_1, v_2, \ldots, v_n$ and directed edges $E = \{(v_i, v_j), \ldots\}$. The terms vertices and nodes, and edges and links will be used interchangeably. We define the neighborhood of a vertex $N(v_i)$ or simply N_i as the set of vertices with an incoming link from v_i. The degree of the vertex, $\deg(v)$ or k_i is the cardinality of the neighborhood of the vertex.

Freebase is an open database of the world's information. It represents data as a labelled directed graph. Topics, such as people, places, or things are represented as nodes and edges express the relationships between them. For example, there is a link labelled marriage between Barack and Michelle Obama. Similarly, there is a directed edge from Barack Obama to Honolulu, labelled "Place of Birth." Unlike many databases, Freebase aims to be comprehensive without duplicating topics. For example, Arnold Schwarzenegger may appear in a body building database as well as a political database, but in Freebase, Arnold is one entity with links to his body-building awards as well as his political positions.

In many ways, we can expect Freebase to have a similar structure to large-scale social networks because they both represent historical interactions between topics. However, there are likely to be some key differences. Social networks grow organically as friendships are formed, but most of the data in Freebase has been bulk loaded by the employees of Metaweb Technologies. Freebase is also in a very early stage of development and many topics have only a handful of outgoing links, and a considerable number have none at all. This paper examines two major statistical techniques for analyzing large-scale graphs: degree distribution and clustering coefficient.

One important simplification was made for the purposes of analysis. Compound value types (CVTs) are an important paradigm that has become common in Freebase. There are some relationships that are difficult to model as a single link between nodes. Instead of using a hypergraph which would allow edges to connect any number of vertices, Freebase introduces mediator vertices that can be connected to any number of topics. For example, the Musical Group Membership CVT has links to the band, the musician, and the role or instrument the musician plays. CVTs were ignored in this project because they introduced significant complications and cannot be queried as easily or in an efficient manner. Furthermore, the CVT model has not been studied in the context of scale-free networks and would introduce significant bias, because there would be a huge number of vertices with small degree.

2 Degree Distributions

The first major attempt to model large-scale networks was the Erdös-Rényi model, first proposed in 1959. [3] It is generally known as the "random graph model," or sometimes the "Poisson random graph model." Constructing a graph with this model is very simple. Given n vertices, place an edge between each pair of vertices with probability p. Graphs constructed using this model will have vertices whose degrees are distributed according to a binomial distribution, since the probability of an edge being present is independent of all other edges. Hence, the probability of a degree having degree k is

$$\Pr(\deg(v) = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{1}$$

As $n \to \infty$, the binomial distribution approaches the Poisson distribution and

$$\Pr(\deg(v) = k) \to \frac{z^k e^{-z}}{k!} \tag{2}$$

where $z = p(n-1)$, the expected value of the degree distribution.

However, real-world networks rarely resemble Erdös-Rényi random graphs. Real-world networks are generally right-skewed, meaning there is a long right tail of degree vertices above the mean. In fact, the degree distributions of most real-world networks follow a power-law distribution:

$$\Pr(\deg(v) = k) \propto k^{-\alpha} \tag{3}$$

for some positive α, which is generally between 2 and 3. The power law distribution usually only describes vertices with degree greater than some $x_{\min}$. For a simple power law, the normalizing constant is $(\alpha - 1)/x_{\min}$. Thus

$$\Pr(\deg(v) = k) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} \tag{4}$$

is a valid probability distribution.

2.1 Fitting power law data

People often attempt to fit data to a power law distribution by transforming the data to log-log axes and performing a linear regression. If we take the logarithm of (3), we see that the power-law obeys

$$\ln p(k) = \alpha \ln k + \text{constant} \tag{5}$$

which implies that it should follow a straight line on a log-log plot. As discussed in the appendix to Clauset, Shalizi and Newman's paper, this approach can lead to significant systematic errors. We therefore turn to the *method of maximum likelihood* which allows us to derive a maximum likelihood estimator (MLE) for the scaling parameter. For continuous data, the MLE is

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1} \tag{6}$$

where x_i for $i = 1, 2, \ldots, n$ are the values of x such that $x \geq x_{\min}$. For now, assume that we are given $x_{\min}$. However, the MLE for discrete data, such as degree distributions is much more complicated. If $x_{\min} = 1$, which is likely for our data, the estimator $\hat{\alpha}$ has been shown by Goldstein, et. al. to be given by the solution to the transcendental equation

$$\frac{\zeta'(\hat{\alpha})}{\zeta(\hat{\alpha})} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i \tag{7}$$

where ζ is the Riemann zeta function. [1] In the more general case, where $x_{\min} \geq 1$, we replace the Riemann zeta function with the generalized Hurwitz zeta function,

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i \tag{8}$$

where

$$\zeta(s, q) = \sum_{n=0}^{\infty} \frac{1}{(q+n)^s}. \tag{9}$$

Note that $\zeta(s, 1) = \zeta(s)$, so the generalized equation also applies when $x_{\min} = 1$. In practice, $\hat{\alpha}$ must be estimated numerically. We can do this either by maximizing the likelihood function, or its logarithm, which is generally simpler:

$$L(\alpha) = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i. \tag{10}$$

Instead of using the exact discrete MLE some authors choose to use a discrete approximation of the continuous MLE:

$$\hat{\alpha} \approx 1 + n \left[\sum_{i=1}^{n} \frac{x_i}{x_{\min} - \frac{1}{2}} \right] \tag{11}$$

However, Clauset et. al. have found that the approximation is only accurate for $x_{\min} \geq 6$, which is not the case for the Freebase data.

We have brushed aside the issue of finding $x_{\min}$. For Freebase degree distributions, we can expect that $x_{\min} = 1$ since topics with degree 0 are an artifact of incomplete data loading. However, it was worthwhile to verify this. Clauset, et al. propose a method of estimating $x_{\min}$based on varying $x_{\min}$to minimize the Kolmogorov-Smirnoff statistic of the computed fit. Using this method, we found that, as expected, $x_{\min} = 1 \pm 0.0$ for all of the degree distributions analyzed in this paper.

2.2 Results

I analyzed three different subgraphs of Freebase: the graph of all topic-to-topic links, the graph of person-to-topic links, and the graph of person-to-person links. In all three cases, the degrees seem to roughly follow a power-law distribution with some important deviations. All three distributions were also fitted to a Poisson distribution. The Poisson fits all had slightly higher Root Mean Square Errors than their corresponding Power Law fits. Person and topic IDs were sampled from the March 2009 Freebase data dumps, and then their degrees were queried using queries such as the one below.

```
{
    'type' : '/type/link',
    'source' : '/en/barack_obama',
    'target' : {'type' : '/people/person'},
    'return' : 'count'
}
```

The results of these fits are summarized in Figures 1 through 3 and in Table 2.2.

The person-to-person links fit a power law distribution very well, but topic-to-topic links exhibit a bias toward lower degree vertices. On the other hand, person-to-topic links exhibit a bias toward higher degree vertices. It seems that when the person-to-person and topic-to-topic links are combined in the person-to-topic subgraph, the two biases are combined which can be seen at $x = 10^1$ and $x \approx 10^{1.5}$.

Subgraph	Poisson (λ)	Poisson RMSE	Power law (α)	Power law RMSE
Person-to-person	1.325	$1.65\ \times 10^{-5}$	3.07	$8.54\ \times 10^{-6}$
Person-to-topic	1.327	$6.58\ \times 10^{-6}$	3.06	$1.63\ \times 10^{-6}$
Topic-to-topic	2.677	$5.58\ \times 10^{-6}$	2.75	$5.06\ \times 10^{-7}$

Table 1: Computed fit parameters and their corresponding RMSE for Poisson and Power law fits. Note the RMSE values are based on a normalized probability distribution that sums to 1.

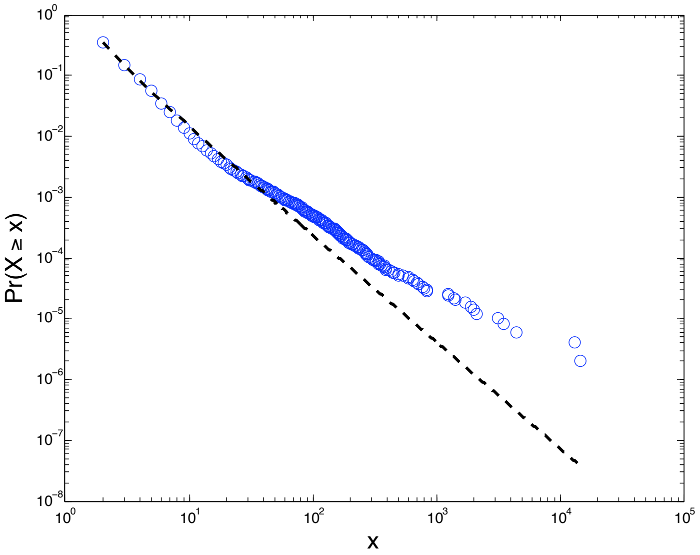

Figure 1: Complementary CDF of vertex degrees of topic-to-topic links. $\alpha = 2.75 \pm 0.13, x_{\min} = 1 \pm 0.0$

.

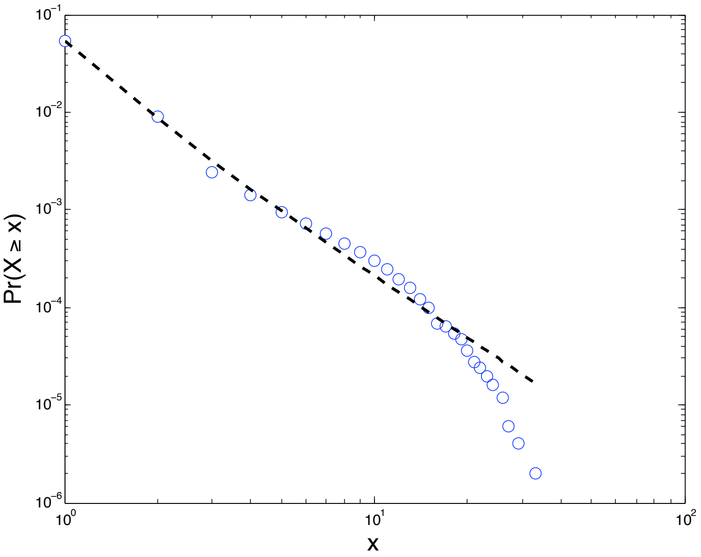

Figure 2: Complementary CDF of vertex degrees of person-to-topic links. $\alpha = 3.06 \pm 0.015, x_{\min} = 1 \pm 0.0$

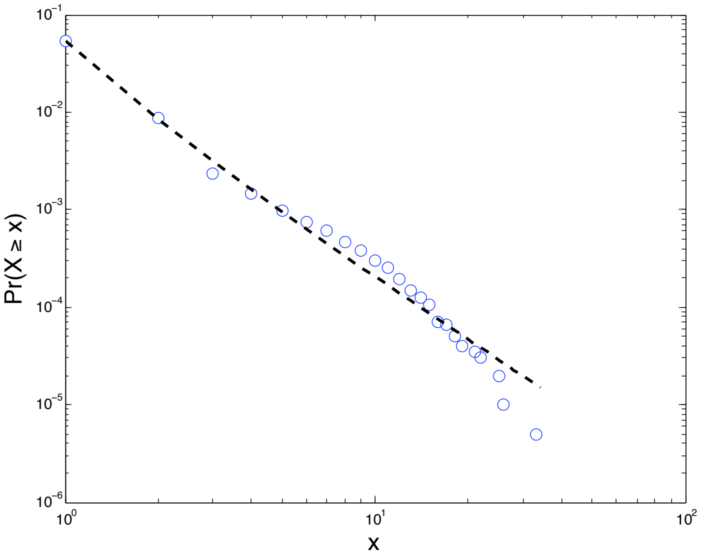

Figure 3: Complementary CDF of vertex degrees of person-to-person links. $\alpha = 3.07 \pm 0.026, x_{\min} = 1 \pm 0.0$

3 Clustering coefficient

Duncan Watts and Steven Strogatz proposed the “clustering coefficient” metric in 1998 to describe how interconnected a graph is. The metric describes the proportion of the number of links between the vertices of a neighborhood to the number of links that could exist between them. Hence, the clustering coefficient for a vertex v_i is defined as

$$C_i = \frac{|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E \tag{12}$$

where k_i is the degree of vertex v_i, N_i is the neighborhood of v_i, and E is the edge set of the graph. In Freebase, it is possible for the clustering coefficient to exceed 1, since there might exist multiple links between a pair of topics, via multiple properties, such as a person who was employed by his alma mater. Strogatz and Watts also define an overall clustering coefficient, defined as the average of the clustering coefficient for each vertex,

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} C_i. \tag{13}$$

If a graph has an average clustering coefficient that is significantly higher than a random graph of the same size, the graph could be a *small-world graph*. A small-world graph has both a high clustering coefficient and short paths between all pairs of vertices. Since the probability of two vertices being connected in a random graph is always p, regardless of whether they share a neighbor, the average clustering coefficient of a random graph has an expected value of p. [4] As of May 1, 2009, there are approximately 6.18×10^6 topics and 1.111×10^8 links between topics. There are 1.910×10^{13} pairs of topics. So, an edge occurs with probability 5.82×10^{-6}. A corresponding Erdös-Reényi graph with the same number of vertices and edges has a clustering coefficient of 4.85×10^{-7} We would expect Freebase to have a much higher clustering coefficient since users are likely to form communities around certain types of highly connected topics.

3.1 Results

The clustering coefficients of vertices in the Freebase graph do not seem to follow any obvious distribution. Because of the relatively low degrees of the graph, the clustering coefficient tends to take on discrete values like $1/4, 1/6$ and $1/2$. I computed the clustering coefficient for the person-to-topic and topic-to-topic subgraphs using samplse of 20K topics, chosen randomly from the March 2009 data dumps. [2] The person-to-topic subgraph has a mean clustering coefficient of 0.1119 and the entire graph has a mean clustering coefficient of 0.1444. These overall clustering coefficients are similar to the clustering coefficients of Roget's Thesaurus and email address books, but such comparisons do not seem very relevant. [3]

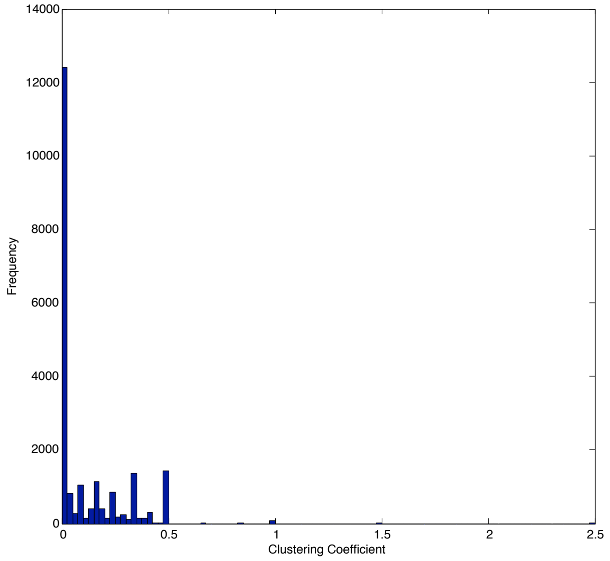

Figure 4: Histogram of the clustering coefficient of vertices in the subgraph of person-to-topic links. $\overline{C} = 0.1119$.

4 Conclusion and future work

An obvious next step is to do all computations on the whole graph, instead of taking a sample. At the current rate of 100k degree queries per hour, computing a degree distribution of all topics would take about 70 hours. Using the current method, computing the clustering coefficient would take approximately 1000 hours. However, there might be more efficient algorithms if one were to use the raw link dumps published by Metaweb. If large-scale computing resources were available, it would be very insightful to calculate

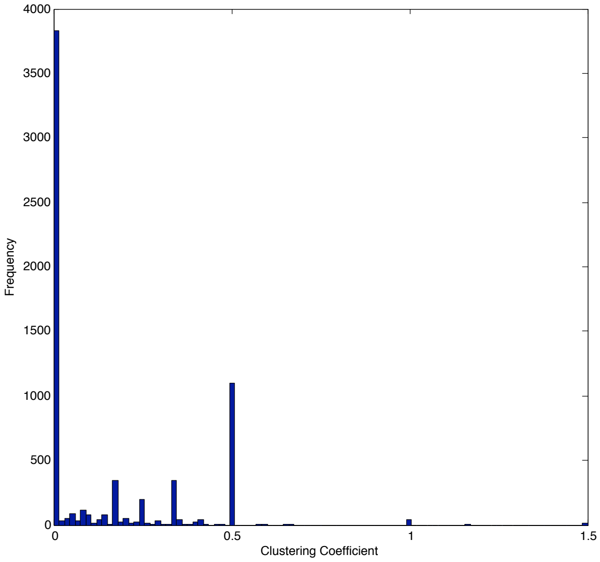

Figure 5: Histogram of the clustering coefficient of the entire topic graph.

shortest paths between pairs of vertices to determine whether the graph is a small world network.

Assortative mixing would also be an interesting phenomenon to explore. In many networks, vertices tend to be connected to other similar vertices. Often, high degree vertices are connected to other high degree vertices and low degree vertices are connected to other low degree vertices. Mixing by discrete traits has also been extensively studied. Freebase's strong type system is an ideal setting for exploring which types assortative mixing.

Finally, all of these computations could be analyzed over time. The Metaweb Query Language provides an `as_of_time` parameter that executes the query over a snapshot of the graph at a specific time. By analyzing the clustering coefficient and degree distributions over time, one could determine the extent to which Freebase has kept up with the influx of new data. Has the existing data become richer, or has there just been an increase in the number of topics?

References

[1] M. L. Goldstein, S. A. Morris, and G. G. Yen. *European Physics Journal*, (41):255, 2004.

[2] Metaweb Technologies. Freebase data dumps. `http://download.freebase.com/datadumps/`, March 23 2009.

[3] MEJ Newman, SH Strogatz, and DJ Watts. Random graphs with arbitrary degree distributions and their applications. *Arxiv preprint cond-mat/0007235*, 2001.

[4] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 293:440–442, 4 June 1998.