Prob Set 6: out tonight (?)
Milgram small world experiment (1967)

Milgram typically chose individuals in the U.S. cities of Omaha, Nebraska, and Wichita, Kansas, to be the starting points and Boston, Massachusetts, to be the end point of a chain of correspondence.

Average path length of those that arrived between 5 and 6 “six degrees of separation”? 
Power Laws in log-log space

\[ y = cx^k \quad (k=1/2,1,2) \]

\[ \log_{10} y = k \log_{10} x + \log_{10} c \]
Power Laws in log-log space

\[ y = cx^{-k} \quad (k=\frac{1}{2}, 1, 2) \]

\[ \log_{10} y = -k \log_{10} x + \log_{10} c \]
Suppose $y = cx^k$

then $\log(y) = k \log(x) + \log(c)$,
and the plot of $\log(y)$ vs $\log(x)$ is a straight line with slope $k$
(and y-intercept $\log(c)$ when $x=1$)

Now suppose some particular $x_0$ value has associated $y_0 = cx_0^k$. Then $x_1 = 10x_0$ has $y_1 = cx_1^k = c(10x_0)^k = 10^k cx_0^k = 10^k y_0$.
So if $x$ increases by a factor of 10, then $y$ increases by a factor of $10^k$, and the value of $k$ is easily determined.

More generally, if $x_1 = r x_0$ and $y_1 = s y_0$ for two points $(x_0, y_0)$ and $(x_1, y_1)$ on the curve $y = cx^k$ then $s = y_1 / y_0 = (x_1 / x_0)^k = r^k$
and comparison of $r$ (ratio of $x_1$ and $x_0$) to $s$ (ratio of $y_1$ and $y_0$) provides the value of the exponent $k$. 
$k = 3$

$\log_{10} N = k \log_{10} t$

$k = -2$

$\log_{10} N = k \log_{10} t$
$k = 0.5$

$k = -1$
$k = \frac{1}{3}$

$k = -\frac{1}{2}$
The Anatomy of the Facebook Social Graph

Johan Ugander, Brian Karrer, Lars Backstrom, Cameron Marlow

(Submitted on 18 Nov 2011)

We study the structure of the social graph of active Facebook users, the largest social network ever analyzed. We compute numerous features of the graph including the number of users and friendships, the degree distribution, path lengths, clustering, and mixing patterns. Our results center around three main observations. First, we characterize the global structure of the graph, determining that the social network is nearly fully connected, with 99.91% of individuals belonging to a single large connected component, and we confirm the "six degrees of separation" phenomenon on a global scale. Second, by studying the average local clustering coefficient and degeneracy of graph neighborhoods, we show that while the Facebook graph as a whole is clearly sparse, the graph neighborhoods of users contain surprisingly dense structure. Third, we characterize the assortativity patterns present in the graph by studying the basic demographic and network properties of users. We observe clear degree assortativity and characterize the extent to which "your friends have more friends than you". Furthermore, we observe a strong effect of age on friendship preferences as well as a globally modular community structure driven by nationality, but we do not find any strong gender homophily. We compare our results with those from smaller social networks and find mostly, but not entirely, agreement on common structural network characteristics.

Comments: 17 pages, 9 figures, 1 table
Subjects: Social and Information Networks (cs.SI); Physics and Society (physics.soc-ph)
… we characterize the entire social network of active members of Facebook in May 2011, a network then comprised of 721 million active users. (roughly 10 percent of the world’s population)

There were 68.7 billion friendship edges at the time of our measurements, so the average Facebook user in our study had around 190 Facebook friends.

… also analyzed the subgraph of 149 million U.S. Facebook users. Using population estimates from the U.S. Census Bureau for 2011, there are roughly 260 million individuals in the U.S. over the age of 13 and therefore eligible to create a Facebook account. Within the U.S., the Facebook social network therefore includes more than half the eligible population. This subpopulation had 15.9 billion edges, so the average U.S. user was friends with around 214 other U.S. users.
Figure 1. Degree distribution $p_k$. (a) The fraction of users with degree $k$ for both the global and U.S. population of Facebook users. (b) The complementary cumulative distribution function (CCDF). The CCDF at degree $k$ measures the fraction of users who have degree $k$ or greater and in terms of the degree distribution is $\sum_{k' \geq k} p_{k'}$. For the U.S., the degree measures the number of friends also from the United States.
(pretend power law)
degree k increases by 100, fraction decreases by 1000
\[ f(k) = ck^{-3/2} = ck^{-1.5} \]
Figure 2. Diameter. The neighborhood function $N(h)$ showing the percentage of user pairs that are within $h$ hops of each other. The average distance between users on Facebook in May 2011 was 4.7, while the average distance within the U.S. at the same time was 4.3.
Feld’s observation that ‘your friends have more friends than you’ is an important psychological paradox, applying to friendship as well as sexual partners.

When people compare themselves to their friends, it is conceptually more appropriate to frame the comparison relative to the median of their friends, psychologizing the question as a matter of asking what one’s ‘class rank’ is amongst one’s peers [34].

Our finding with regard to the median is therefore perhaps more significant: we observe that 83.6% of users have less friends than the median friend count of their friends. All these individuals experience that more than half of their friends have more friends than they do. For completeness, we also note that 92.7% of users have less friends than the average friend count of their friends.
Now we have characterized the growth of the vocabulary in collections. We also want to know how many frequent vs. infrequent terms we should expect in a collection. In natural language, there are a few very frequent terms and very many very rare terms.

Zipf’s law (linguist/philologist George Zipf, 1935): The $i^{th}$ most frequent term has frequency proportional to $1/i$.

$cf_i \propto \frac{1}{i}$

$cf_i$ is collection frequency: the number of occurrences of the term $t_i$ in the collection.
In [ ]:
# Now try Shakespeare, see
# http://www.gutenberg.org/ebooks/100
# http://www.gutenberg.org/cache/epub/100/pg100.txt

In [4]:
sh_txt=urlopen('http://www.gutenberg.org/cache/epub/100/pg100.txt').read().decode('utf-8')

In [61]: sh_txt.count('Romeo')
Out[61]: 155

In [62]: sh_txt.find('1609')
# where the first sonnets start
Out[62]: 7596

In [71]: sh_words=sh_txt[7596:].lower().split()
In [65]: len(sh_words)
Out[65]: 902892
#finally plot the #occurrences of words against their rank as in ps2, #should again be a -1 power law, i.e., -1 slope in log-log coords #good fit from about rank 20 to 10,000 # vocab of 59605 distinct words
Zipf’s law: the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc. Brown Corpus:

- “the”: 7% of all word occurrences (69,971 of $\geq 1$M).
- “of”: $\sim3.5\%$ of words (36,411)
- “and”: 2.9% (28,852)

Only 135 vocabulary items account for half the Brown Corpus.
Only 135 vocabulary items account for half the Brown Corpus.

The Brown University Standard Corpus of Present-Day American English is a carefully compiled selection of current American English, totaling about a million words drawn from a wide variety of sources . . . for many years among the most-cited resources in the field.
Zipf’s law: The $i^{th}$ most frequent term has frequency proportional to $1/i$.

$\text{cf}_i \propto \frac{1}{i}$

$\text{cf}$ is collection frequency: the number of occurrences of the term in the collection.

So if the most frequent term ($\text{the}$) occurs $\text{cf}_1$ times, then the second most frequent term ($\text{of}$) has half as many occurrences $\text{cf}_2 = \frac{1}{2} \text{cf}_1 \ldots$

$\ldots$ and the third most frequent term ($\text{and}$) has a third as many occurrences $\text{cf}_3 = \frac{1}{3} \text{cf}_1$ etc.

Equivalent: $\text{cf}_i = c i^k$ and $\log \text{cf}_i = \log c + k \log i$ (for $k = -1$)

Example of a power law
Fit far from perfect, but nonetheless key insight:

Few frequent terms, many rare terms.
more from http://en.wikipedia.org/wiki/Zipf’s law

“A plot of word frequency in Wikipedia (27 Nov 2006). The plot is in log-log coordinates. \( x \) is rank of a word in the frequency table; \( y \) is the total number of the words occurrences. Most popular words are “the”, “of” and “and”, as expected. Zipf’s law corresponds to the upper linear portion of the curve, roughly following the green \((1/x)\) line.”
Another Wikipedia count (15 May 2010)

http://imonad.com/seo/wikipedia-word-frequency-list/

All articles in the English version of Wikipedia, 21GB in XML format (five hours to parse entire file, extract data from markup language, filter numbers, special characters, extract statistics):

- Total tokens (words, no numbers): $T = 1,570,455,731$
- Unique tokens (words, no numbers): $M = 5,800,280$
“Word frequency distribution follows Zipf’s law”
- rank 1–50 (86M-3M), stop words (the, of, and, in, to, a, is, ...)
- rank 51–3K (2.4M-56K), frequent words (university, January, tea, sharp, ...)
- rank 3K–200K (56K-118), words from large comprehensive dictionaries (officiates, polytonality, neologism, ...)
  above rank 50K mostly Long Tail words
- rank 200K–5.8M (117-1), terms from obscure niches, misspelled words, transliterated words from other languages, new words and non-words (euprostheneops, eurotrochilus, lokottaravada, ...)
Some selected words and associated counts

- Google 197920
- Twitter 894
- domain 111850
- domainer 22
- Wikipedia 3226237
- Wiki 176827
- Obama 22941
- Oprah 3885
- Moniker 4974
- GoDaddy 228
Project Gutenberg (per billion)

Over 36,000 items (Jun 2011), average of $> 50$ new e-books / week


- the 56271872
- of 33950064
- and 29944184
- to 25956096
- in 17420636
- I 11764797
- that 11073318
- was 10078245
- his 8799755
- he 8397205

- it 8058110
- with 7725512
- is 7557477
- for 7097981
- as 7037543
- had 6139336
- you 6048903
- not 5741803
- be 5662527
- her 5202501

... 100,000th
Power Laws more generally

E.g., consider power law distributions of the form $c r^{-k}$, describing the number of book sales versus sales-rank $r$ of a book, or the number of Wikipedia edits made by the $r^{th}$ most frequent contributor to Wikipedia.

- Amazon book sales: $c r^{-k}, k \approx 0.87$
- number of Wikipedia edits: $c r^{-k}, k \approx 1.7$

(More on power laws and the long tail here: Networks, Crowds, and Markets: Reasoning About a Highly Connected World by David Easley and Jon Kleinberg
Normalization given by the roughly 1 sale/week for the 200,000th ranked Amazon title: $40916/r^{.87}$ and by the 10 edits/month for the 1000th ranked Wikipedia editor: $1258925/r^{1.7}$
Long tail: about a quarter of Amazon book sales estimated to come from the long tail, i.e., those outside the top 100,000 bestselling titles.
If a city is 10 times as populous, does it have 10 times as many gas stations?

Empirically, $G = c P^{.77}$ (economies of scale)

similarly for miles of roadway, length of electrical cables, …, $k$ ranges from .7 to .9
Power law distributions

\[ f(x) = ax^k + o(x^k), \quad \log(f(x)) = k \log x + \log a. \]
Examples

Moby Dick

scientific papers 1981-1997

AOL users visiting sites ‘97

word frequency

citations

web hits

books sold

bestsellers 1895-1965

AT&T customers on 1 day

earthquake magnitude

California 1910-1992
Power law in networks

- For many interesting graphs, the distribution over node degree follows a power law

<table>
<thead>
<tr>
<th>Graph Type</th>
<th>Exponent $\alpha$ (in/out degree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>film actors</td>
<td>2.3</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>2.1</td>
</tr>
<tr>
<td>email networks</td>
<td>1.5/2.0</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>3.2</td>
</tr>
<tr>
<td>WWW</td>
<td>2.3/2.7</td>
</tr>
<tr>
<td>internet</td>
<td>2.5</td>
</tr>
<tr>
<td>peer-to-peer</td>
<td>2.1</td>
</tr>
<tr>
<td>metabolic network</td>
<td>2.2</td>
</tr>
<tr>
<td>protein interactions</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Consider networks with power law exponent dependent on parameter $p$ (Easley/Kleinberg 18.3)

Model has directed links (so more in the spirit of web pages than social network)

Add new page (node) $j$, give link (edge) to an earlier page, according to probabilistic rule:

(a) With probability $p$, page $j$ links to page $i$ chosen at random from all earlier pages;

(b) With probability $1 - p$, page $j$ instead links to a page $i$ chosen with probability proportional to $i$’s current number of in-links.

(a) permits discovery of pages that start with zero in-links, (b) is “preferential attachment”
Details next time, result for fraction of nodes with degree $k$ is:

$$f(k) = -\frac{dF}{dk} \sim k^{-(1+1/q)},$$

This is a power law with exponent $\alpha = 1 + 1/q = 1 + 1/(1-p)$

The limit $p \to 1$ gives back the random network, where $\alpha \to \infty$ signals loss of the power law behavior (the tail is extinguished).

In the $p \to 0$ limit, the exponent $\alpha \to 2$, and the tail of the distribution is that much more pronounced.

Smaller $p$ permits nodes with even larger in-degree, giving a longer tail.