

In [2]:

```
bigfile = open('/Users/se/.ipython/cc2017-master/data/ows_tweets_sample.txt', 'rb')
chunkSize = 1000000
chunk = bigfile.readlines(chunkSize)
print (len(chunk))
with open("/Users/se/.ipython/cc2017-master/data/推特清洗.txt", 'wb') as f:
    for i in chunk:
        f.write(i)
```

2750

In [3]:

```
with open("/Users/se/.ipython/cc2017-master/data/ows_tweets_sample.txt", 'rb') as f:
    lines = f.readlines()
```

In [4]:

```
len(lines)
```

Out[4]:

2752

In [5]:

```
lines[0]
```

Out[5]:

```
b' "Twitter ID",Text,"Profile Image URL",Day,Hour,Minute,"Created At",Geo,"From Use
r","From User ID",Language,"To User","To User ID",Source\n'
```

In [6]:

```
varNames = str(lines[0]).replace('\n','').split(',')
varNames
```

Out[6]:

```
['b\' "Twitter ID",
\'Text\',
\' "Profile Image URL",
\'Day\',
\'Hour\',
\'Minute\',
\' "Created At",
\'Geo\',
\' "From User",
\' "From User ID",
\'Language\',
\' "To User",
\' "To User ID",
\'Source\\n\' "']
```

In [7]:

len(varNames)

Out[7]:

14

In [8]:

lines[1344]

Out[8]:

```
b'121818600490283009,"RT @chachiTHEgr8: RT @TheNewDeal: First they ignore you, then
they laugh at you, then they fight you, then you win. - Gandhi #OccupyWallStreet #O
WS #p2",http://a0.twimg.com/profile_images/326662126/Photo_233_normal.jpg,2011-10-0
6,5,26,"2011-10-06 05:26:32",N;,k_l_h_j,382233343,en,,0,"<a href="http://twi
tter.com/#!/download/iphone" rel="nofollow">Twitter for iPhone&l
t;/a&";'\n'
```

In [9]:

```
with open("/Users/se/.ipython/cc2017-master/data/推特清洗.txt", 'w') as f:
    right_line = ''
    blocks = []
    for line in lines:
        right_line += str(line).replace('\n', ' ')
        line_length = len(right_line.split(','))
        if line_length >= 14:
            blocks.append(right_line)
            right_line = ''
    for i in blocks:
        f.write(i + '\n')
```

In [10]:

len(blocks)

Out[10]:

2627

In [11]:

blocks[1344]

Out[11]:

```
'b'121818879105310720,"RT @Min_Reyes: RT @The99Percenters: New video to go viral. F
rom We Are Change\rhttp://t.co/6Ff718jk Listen to the guy begging... #ows #cdnpol
i",http://a3.twimg.com/sticky/default_profile_images/default_profile_0_normal.png,20
11-10-06,5,27,"2011-10-06 05:27:38",N;,MiyazakiMegu,260948518,en,,0,"<a href="http://www.tweetdeck.com" rel="nofollow">TweetDeck&l
t;/a&";'\n
\'
```

In [12]:

```

with open("/Users/se/.ipython/cc2017-master/data/推特清洗2.txt", 'w') as f:
    right_line = ''
    blocks = []
    for line in lines:
        right_line += str(line).replace('\n', ' ').replace('\r', ' ')
        line_length = len(right_line.split(','))
        if line_length >= 14:
            blocks.append(right_line)
            right_line = ''
    for i in blocks:
        f.write(i + '\n')

```

In [13]:

blocks[1344]

Out[13]:

```

'b\\' 121818879105310720,"RT @Min_Reyes: RT @The99Percenters: New video to go viral. From We Are Change\\rhttp://t.co/6Ff718jk Listen to the guy begging... #ows #cdnpol i",http://a3.twimg.com/sticky/default_profile_images/default_profile_0_normal.png,2011-10-06,5,27,"2011-10-06 05:27:38",N;,MiyazakiMegu,260948518,en,,0,"&lt;a href=&quot;t;http://www.tweetdeck.com&quot; rel=&quot;nofollow&quot;&gt;TweetDeck&lt;/a&gt;"\\n\\'

```

In [14]:

```

with open("/Users/se/.ipython/cc2017-master/data/推特清洗.txt", 'rb') as f:
    chunk = f.readlines()

```

In [15]:

len(chunk)

Out[15]:

2627

In [16]:

chunk[:3]

Out[16]:

```

[b'b\\' "Twitter ID",Text,"Profile Image URL",Day,Hour,Minute,"Created At",Geo,"From User", "From User ID",Language,"To User", "To User ID",Source\\n\\'\\r\\n',
 b'b\\' 121813144174727168,"RT @AnonKitsu: ALERT!!!!!!!!!!!!COPS ARE KETTLING PROTESTERS IN PARK W HELICOPTERS AND PADDYWAGONS!!!! #OCCUPYWALLSTREET #OWS #OCCUPYNY PLEASE RT !!HELP!!!!",http://a2.twimg.com/profile_images/1539375713/Twitter_normal.jpg,2011-10-06,5,4,"2011-10-06 05:04:51",N;,Anonops_Cop,401240477,en,,0,"&lt;a href=&quot;http://twitter.com/&quot;&gt;web&lt;/a&gt;"\\n\\'\\r\\n',
 b'b\\' 121813146137657344,"@jamiekilstein @allisonkilkenny Interesting interview (never aired, wonder why??) by Fox with #ows protester http://t.co/Fte55Kh7",http://a2.twimg.com/profile_images/1574715503/Kate6_normal.jpg,2011-10-06,5,4,"2011-10-06 (http://t.co/Fte55Kh7",http://a2.twimg.com/profile_images/1574715503/Kate6_normal.jpg,2011-10-06,5,4,"2011-10-06) 05:04:51",N;,KittyHybrid,34532053,en,jamiekilstein,2149053,"&lt;a href=&quot;http://twitter.com/&quot;&gt;web&lt;/a&gt;"\\n\\'\\r\\n']

```

In [17]:

```
import csv
clean_lines = (line.replace('\x00','') \
               for line in chunk[1:])
lines = csv.reader(clean_lines, delimiter=',', \
                  quotechar='"')
```

In [18]:

```
import pandas as pd

df = pd.read_csv("/Users/se/.ipython/cc2017-master/data/ows_tweets_sample.txt", \
                sep = ',', quotechar='"')
df[:3]
```

Out[18]:

	Twitter ID	Text	Profile Image URL
0	121813144174727168	RT @AnonKitsu: ALERT!!!!!!!!!!!!COPS ARE KETTLIN...	http://a2.twimg.com/profile_images/153937571:
1	121813146137657344	@jamiakilstein @allisonkilkenny Interesting in...	http://a2.twimg.com/profile_images/157471550:
2	121813150000619521	@Seductivpancake Right! Those guys have a vict...	http://a1.twimg.com/profile_images/124141283:

In [19]:

```
df.Text[1]
```

Out[19]:

```
'@jamiakilstein @allisonkilkenny Interesting interview (never aired, wonder why??) b  
y Fox with #ows protester http://t.co/Fte55Kh7' (http://t.co/Fte55Kh7')
```

In [20]:

```
df['From User']
```

Out[20]:

```
0      Anonops_Cop
1      KittyHybrid
2      nerdsherpa
3      hamudistan
4      kl_knox
5      vickycrampton
6      burgerbuilders
7      neverfox
8      davidgaliel
9      AnonOws
10     ryanjnagle
11     JanSKay
12     alyssa011968
13     fresh__air
14     fucentarmal
15     msaf
16     Seekadoo2
17     Jay_in_Omaha
18     starshollowgzt
19     TopCat_TC
20     ElizabethBrossa
21     dpfournier
22     ReTweetAgain
23     Leoslair
24     tommymiles
25     Perlins
26     SuzanneTwoTon
27     OccupySF
28     k_l_h_j
29     adpucci
...
2596    DadaIsGalletas
2597          dvnix
2598    ilmestykset
2599    Amywhere
2600    EvanPokroy
2601    shaunasmo
2602    pishka
2603    portugalthem
2604    RomperSTL
2605    hjertebraten
2606          dvnix
2607    occpal
2608    AntarianRani
2609    vaneesa
2610    OccupyKC
2611    vaneesa
2612    vaneesa
2613    singlepayer
2614    DBAYJR
2615    vaneesa
2616    guykanon
2617    DakotaNorthAnon
2618    Isajah78
2619    HuggyBear116
2620    chounesma
```

```
2020          shaunismo
2621          sindikaria
2622          trvslogan
```

```
2623          vaneeesa
2624          autolysis
2625          UsRevolt11
```

Name: From User, dtype: object

In [21]:

```
from collections import defaultdict
data_dict = defaultdict(int)
line_num = 0
lines = csv.reader((str(line).replace('\x00','') for line in chunk[1:]), delimiter=',', quotechar='\"')
for i in lines:
    line_num +=1
    data_dict[i[8]] +=1
```

In [22]:

```
list(data_dict.items())[:5]
```

Out[22]:

```
[('Anonops_Cop', 1),
 ('KittyHybrid', 1),
 ('nerdsherpa', 2),
 ('hamudistan', 1),
 ('kl_knox', 1)]
```

In [23]:

```
print (line_num)
```

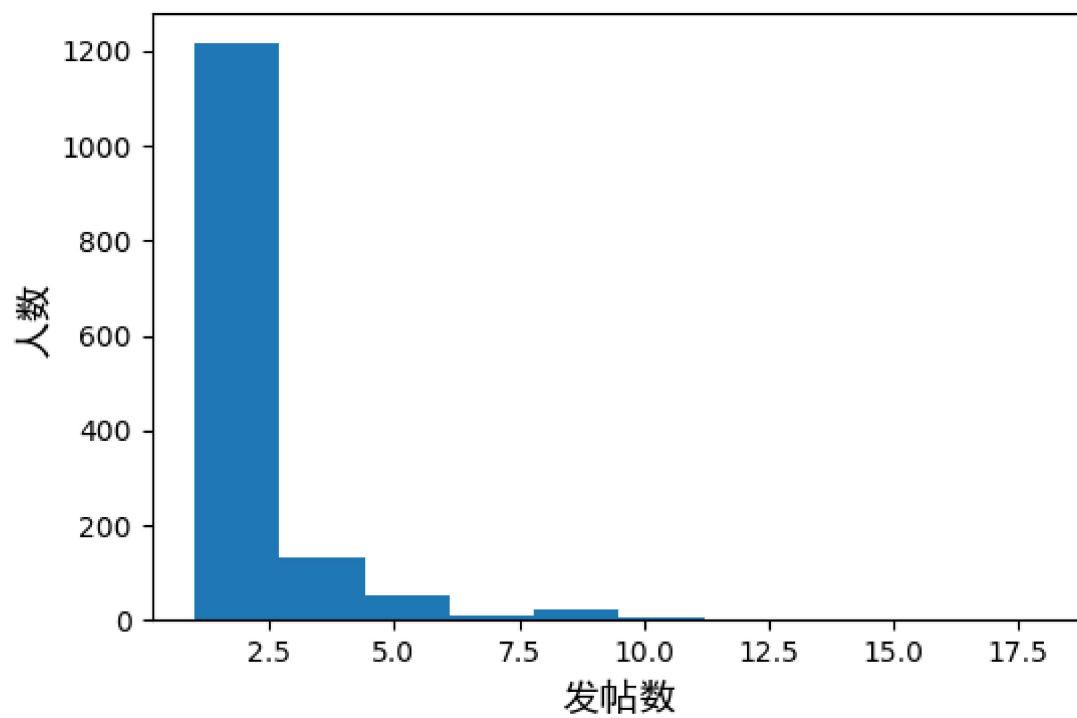
2626

In [24]:

```
%matplotlib inline
from matplotlib.font_manager import FontProperties
import matplotlib.pyplot as plt
import matplotlib
#matplotlib.rcParams['font.sans-serif'] = ['Microsoft YaHei']
matplotlib.rcParams["savefig", dpi=100)
font = FontProperties(fname=r'/Windows/WinSxS/amd64_microsoft-windows-font-truetype-simhei_31bf3856a
```

In [25]:

```
plt.hist(list(data_dict.values()))  
plt.xlabel(u'发帖数', fontproperties=font)  
plt.ylabel(u'人数', fontproperties=font)  
plt.show()
```



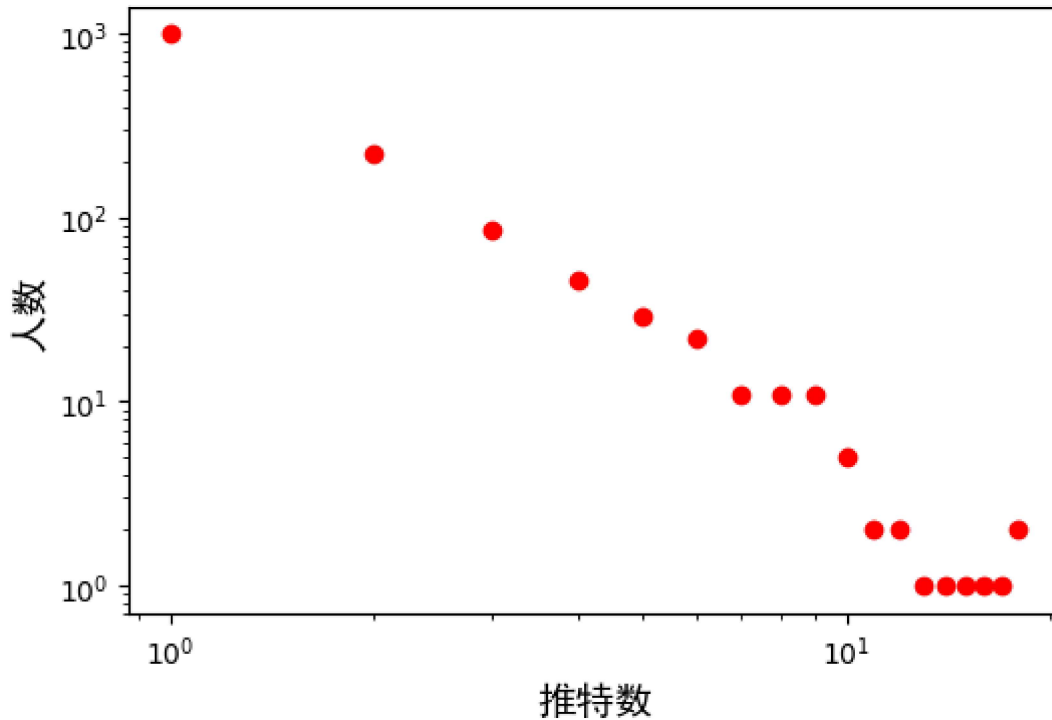
In [26]:

```

tweet_dict = defaultdict(int)
for i in data_dict.values():
    tweet_dict[i] += 1

plt.loglog(list(tweet_dict.keys()), list(tweet_dict.values()), 'ro', linewidth=2)
plt.xlabel(u'推特数', fontproperties=font)
plt.ylabel(u'人数', fontproperties=font)
plt.show()

```



In [27]:

```

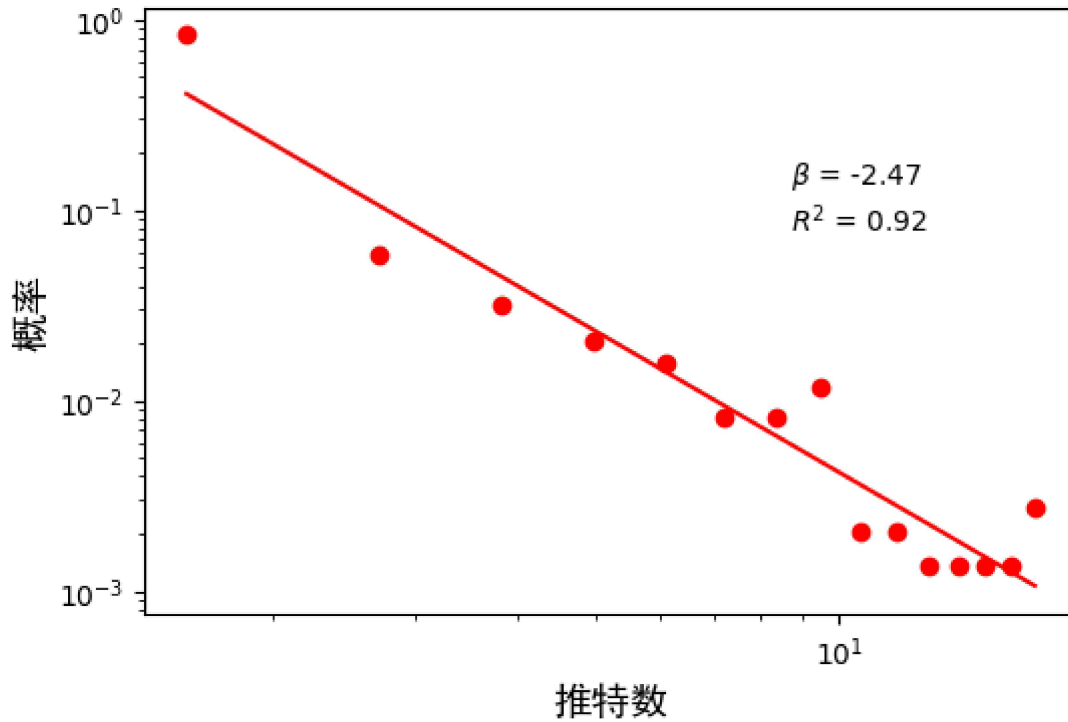
import numpy as np
import statsmodels.api as sm

def powerPlot(d_value, d_freq, color, marker):
    d_freq = [i + 1 for i in d_freq]
    d_prob = [float(i)/sum(d_freq) for i in d_freq]
    x = np.log(d_value)
    y = np.log(d_prob)
    xx = sm.add_constant(x, prepend=True)
    res = sm.OLS(y, xx).fit()
    constant, beta = res.params
    r2 = res.rsquared
    plt.plot(d_value, d_prob, linestyle='', color = color, marker = marker)
    plt.plot(d_value, np.exp(constant+x*beta), "red")
    plt.xscale('log'); plt.yscale('log')
    plt.text(max(d_value)/2, max(d_prob)/10,
             r'$\beta$ = ' + str(round(beta, 2)) + '\n' + r'$R^2$ = ' + str(round(r2, 2)))

```


In [28]:

```
histo, bin_edges = np.histogram(list(data_dict.values()), 15)
bin_center = 0.5*(bin_edges[1:] + bin_edges[:-1])
powerPlot(bin_center, histo, 'r', 'o')
plt.ylabel(u' 概率', fontproperties=font)
plt.xlabel(u' 推特数', fontproperties=font)
plt.show()
```



In [29]:

```

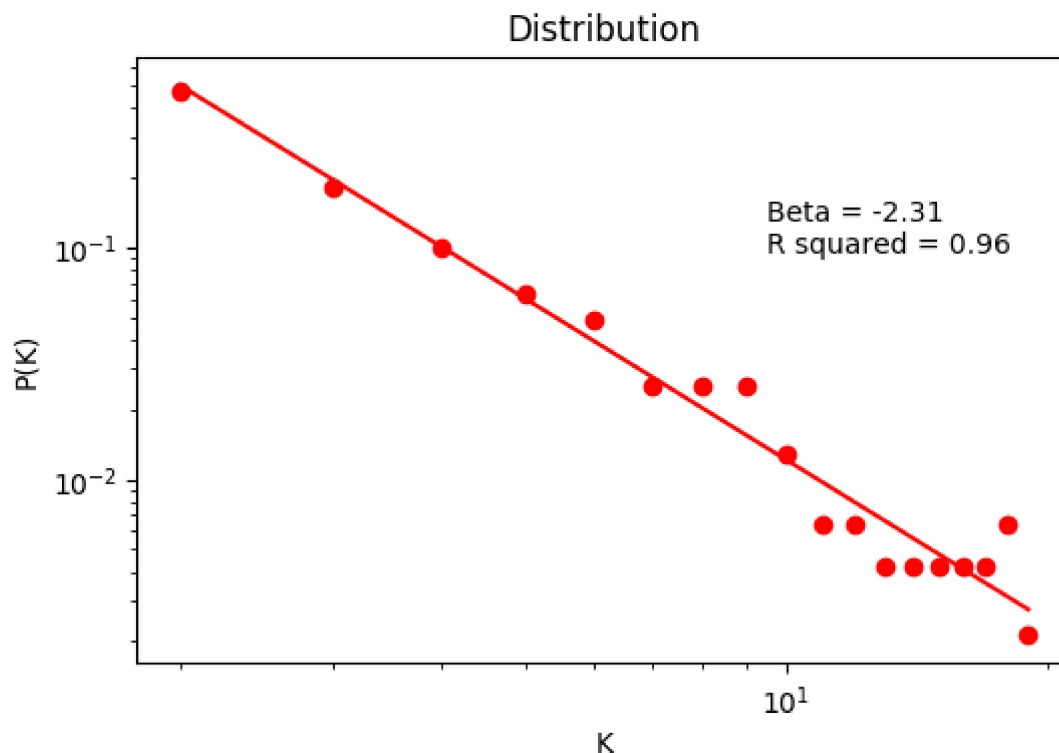
import statsmodels.api as sm
from collections import defaultdict
import numpy as np

def powerPlot(data):
    d = sorted(data, reverse = True )
    d_table = defaultdict(int)
    for k in d:
        d_table[k] += 1
    d_value = sorted(d_table)
    d_value = [i+1 for i in d_value]
    d_freq = [d_table[i]+1 for i in d_value]
    d_prob = [float(i)/sum(d_freq) for i in d_freq]
    #d_rank = ss.rankdata(d_value).astype(int)
    x = np.log(d_value)
    y = np.log(d_prob)
    xx = sm.add_constant(x, prepend=True)
    res = sm.OLS(y, xx).fit()
    constant, beta = res.params
    r2 = res.rsquared
    plt.plot(d_value, d_prob, 'ro')
    plt.plot(d_value, np.exp(constant+x*beta), "red")
    plt.xscale('log'); plt.yscale('log')
    plt.text(max(d_value)/2, max(d_prob)/5,
             'Beta = ' + str(round(beta,2)) + '\n' + 'R squared = ' + str(round(r2, 2)))
    plt.title('Distribution')
    plt.ylabel('P(K)')
    plt.xlabel('K')
    plt.show()

```

In [30]:

```
powerPlot(data_dict.values())
```



In [33]:

```
import powerlaw
```

In [35]:

```
import powerlaw
def plotPowerlaw(data, ax, col, xlab):
    fit = powerlaw.Fit(data, xmin=2)
    fit.plot_pdf(color = col, linewidth = 2)
    a, x = (fit.power_law.alpha, fit.power_law.xmin)
    fit.power_law.plot_pdf(color = col, linestyle = 'dotted', ax = ax, \
                           label = r"$\alpha = %d \:\:\:, x_{\min} = %d$" % (a, x))
    ax.set_xlabel(xlab, fontsize = 20)
    ax.set_ylabel('$Probability$', fontsize = 20)
    plt.legend(loc = 0, frameon = False)
```

In [36]:

```
from collections import defaultdict
data_dict = defaultdict(int)

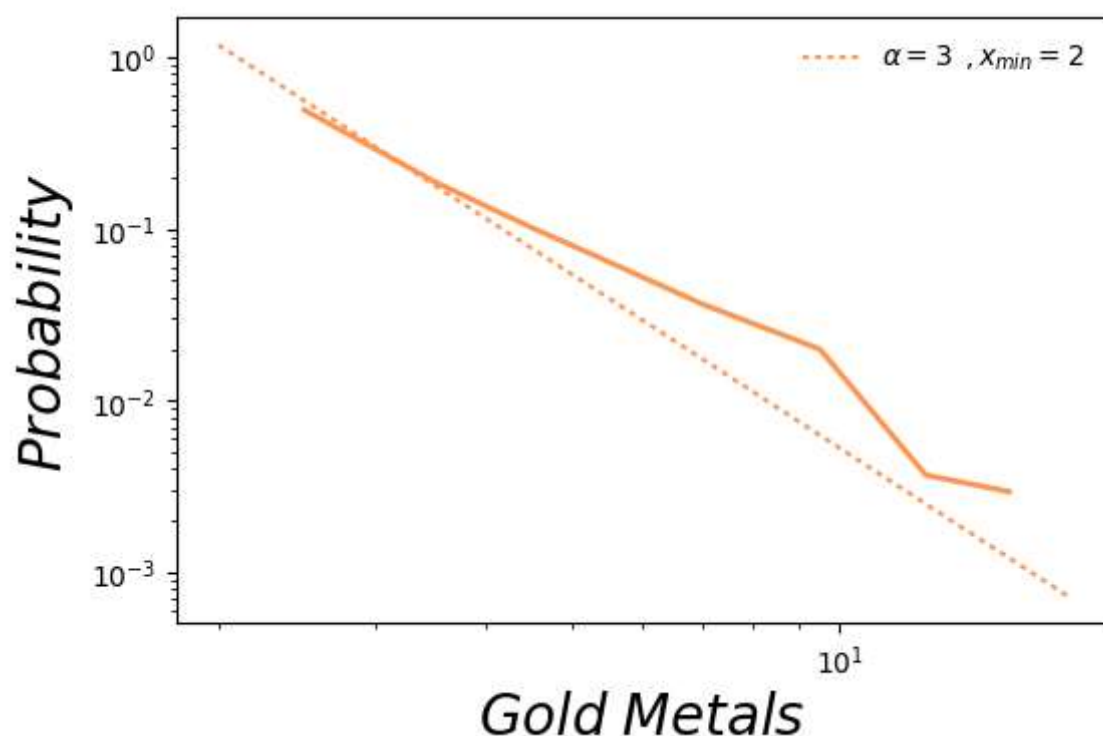
for i in df['From User']:
    data_dict[i] += 1
```

In [38]:

```
import matplotlib.cm as cm
cmap = cm.get_cmap('rainbow_r', 6)

fig = plt.figure(figsize=(6, 4), facecolor='white')
ax = fig.add_subplot(1, 1, 1)
plotPowerlaw(list(data_dict.values()), ax, cmap(1), '$Gold\;Metals$')
```

C:\Users\se\Anaconda3\lib\site-packages\powerlaw.py:697: RuntimeWarning: invalid value encountered in true_divide
(Theoretical_CDF * (1 - Theoretical_CDF))



In [119]:

```
tweet = '''RT @AnonKitsu: ALERT!!!!!!!!!!!!COPS ARE KETTLING PROTESTERS IN PARK W HELICOPTERS AND PAD  
#OCCUPYWALLSTREET #OWS #OCCUPYNY PLEASE @chengjun @mili http://computational-communicati  
http://ccc.nju.edu.cn RT !!HELP!!!!'''
```

In [6]:

```
import re

import twitter_text
```

File "C:\Users\se\Anaconda3\lib\site-packages\twitter_text\autolink.py", line 313
preceding_ellipsis = re.search(ur'\A...', display_url)

SyntaxError: invalid syntax