Operations_data

Unknown Author

```
In [1]:  %load_ext autoreload
         %autoreload 2
         %matplotlib inline
```

```
In [2]:  import pandas as pd
         import YT_api_generate as yt
         import seaborn
         import re
         from optparse import OptionParser
         from IPython.display import HTML
         from IPython.display import YouTubeVideo
         import matplotlib.pyplot as plt
         pd.set_option("display.max_columns", 6)
         pd.set_option("display.max_rows", 15)
         pd.set_option("display.notebook_repr_html", True)
```

1 Mirroring and Anonymous

We are trying to gain an empirical sense of the extent and shape of mirroring practices in the Anonymous movement. We are focusing on Youtube video mirroring (why? Adam). There are various problems in trying to get a sense of mirroring practice on Youtube. [TBA – a description of this]

1.1 What we'd like to produce:

1. A quantitative description of how many mirrored videos exist, when they were published and on what topics (that is, operations)
2. A timeline of mirroring events associated with Anonymous. The timeline should describe their relative importance

1.2 The list of Anonymous operations

Youtube searches using the API only return 500 results. (I guess most people don't want to look at more than a couple of pages of results). But a given operation might have 700 or 2400 mirrors. It is hard to get around this. Our main strategy has been to use *operations* to multiply our search queries. We have found around 80 Anonymous operations. We run Youtube queries on all of these. We also run another set of 80 queries with the word 'mirror' added to the operation. The idea here is that we are creating a large query out of many smaller queries tiled together.

```
In [83]:  # felipe produced a list of operations
          operations = [op.replace('\n', '') for op in open('data/operations_list.txt').readline
          operations_mirror = [op + ' mirror' for op in operations]
```

```
In [84]: operations.sort()
         operations
```

Out [84]:
```
['anonymous operation NSA',
 'anonymous operation V',
 'anonymous operation aaron swartz',
 'anonymous operation acta',
 'anonymous operation american freedom',
 'anonymous operation american spring',
 'anonymous operation anaheim',
 'anonymous operation antisec',
 'anonymous operation assange',
 'anonymous operation awake the masses',
 'anonymous operation bahrain',
 'anonymous operation bart',
 'anonymous operation big brother',
 'anonymous operation black march',
 'anonymous operation blackout',
 'anonymous operation blitzkrieg',
 'anonymous operation brotherhood takedown',
 'anonymous operation cannabis',
 'anonymous operation chanology',
 'anonymous operation chris forcand',
 'anonymous operation cispa',
 'anonymous operation darknet',
 'anonymous operation didgeridie',
 'anonymous operation egypt',
 'anonymous operation ele',
 'anonymous operation empire state',
 'anonymous operation facebook',
 'anonymous operation federal reserve',
 'anonymous operation fullerton',
 'anonymous operation game',
 'anonymous operation greece',
 'anonymous operation green rights',
 'anonymous operation guantanamo bay',
 'anonymous operation habbo',
 'anonymous operation hal turner',
 'anonymous operation hbgary',
 'anonymous operation hunthunter',
 'anonymous operation india',
 'anonymous operation indignant',
 'anonymous operation iran',
 'anonymous operation israel',
 'anonymous operation japan',
 'anonymous operation kingfish',
 'anonymous operation kiwi freedom',
 'anonymous operation kony',
 'anonymous operation last resort',
 'anonymous operation leakspin',
 'anonymous operation libya',
 'anonymous operation los zetas',
 'anonymous operation malaysia',
 'anonymous operation maryville',
 'anonymous operation mastercard',
```

```
            'anonymous operation mayhem',
            'anonymous operation megaupload',
            'anonymous operation north korea',
            'anonymous operation occupythemedia',
            'anonymous operation orlando',
            'anonymous operation palin',
            'anonymous operation payback',
            'anonymous operation paypal',
            'anonymous operation petrol',
            'anonymous operation pipa',
            'anonymous operation pirate bay',
            'anonymous operation powerbot',
            'anonymous operation quebec',
            'anonymous operation rebuild the hive',
            'anonymous operation reconnect',
            'anonymous operation rohingya',
            'anonymous operation rose awareness',
            'anonymous operation safe winter',
            'anonymous operation shooting sheriffs',
            'anonymous operation sony',
            'anonymous operation sopa',
            'anonymous operation steubenville',
            'anonymous operation support snowden',
            'anonymous operation syria',
            'anonymous operation tea party',
            'anonymous operation titstorm',
            'anonymous operation tunisia',
            'anonymous operation ukraine',
            'anonymous operation vendetta',
            'anonymous operation visa',
            'anonymous operation wall street',
            'anonymous operation westboro baptist church',
            'anonymous operation wikileaks',
            'anonymous operation world bank',
            'anonymous project mayhem']
```

```
df_ops = pd.DataFrame()
```

In [4]:

```
## the basic operations
```

In [71]:

```
for op in operations:
    df  = yt.youtube_search(op,1000, True)
    df['operation'] = op
    df_ops = df_ops.append(df)
```

```
retrieved 492 results for anonymous operation habbo
('anonymous operation habbo', (492, 15))
(492, 8)
retrieved 45 results for anonymous operation hal turner
('anonymous operation hal turner', (45, 13))
(45, 8)
retrieved 22 results for anonymous operation chris forcand
('anonymous operation chris forcand', (22, 13))
(22, 8)
retrieved 494 results for anonymous operation chanology
('anonymous operation chanology', (494, 14))
```

```
(494, 8)
retrieved 491 results for anonymous operation palin
('anonymous operation palin', (491, 15))
(491, 8)
retrieved 500 results for anonymous operation iran
('anonymous operation iran', (500, 15))
(500, 8)
retrieved 1 results for anonymous operation didgeridie
('anonymous operation didgeridie', (1, 12))
(1, 8)
retrieved 136 results for anonymous operation titstorm
('anonymous operation titstorm', (136, 15))
(136, 8)
retrieved 500 results for anonymous operation tea party
('anonymous operation tea party', (500, 15))
(500, 8)
retrieved 460 results for anonymous operation payback
('anonymous operation payback', (460, 15))
(460, 8)
retrieved 500 results for anonymous operation wikileaks
('anonymous operation wikileaks', (500, 14))
(500, 8)
retrieved 500 results for anonymous operation assange
('anonymous operation assange', (500, 15))
(500, 8)
retrieved 500 results for anonymous operation leakspin
('anonymous operation leakspin', (500, 14))
(500, 8)
retrieved 473 results for anonymous operation visa
('anonymous operation visa', (473, 15))
(473, 8)
retrieved 479 results for anonymous operation mastercard
('anonymous operation mastercard', (479, 14))
(479, 8)
retrieved 478 results for anonymous operation paypal
('anonymous operation paypal', (478, 14))
(478, 8)
retrieved 500 results for anonymous operation tunisia
('anonymous operation tunisia', (500, 15))
(500, 8)
retrieved 499 results for anonymous operation egypt
('anonymous operation egypt', (499, 14))
(499, 8)
retrieved 500 results for anonymous operation libya
('anonymous operation libya', (500, 15))
(500, 8)
retrieved 322 results for anonymous operation hbgary
('anonymous operation hbgary', (322, 14))
(322, 8)
retrieved 479 results for anonymous operation empire state
('anonymous operation empire state', (479, 14))
(479, 8)
retrieved 483 results for anonymous operation sony
```

```
('anonymous operation sony', (483, 15))
(483, 8)
retrieved 500 results for anonymous operation india
('anonymous operation india', (500, 15))
(500, 8)
retrieved 500 results for anonymous operation malaysia
('anonymous operation malaysia', (500, 15))
(500, 8)
retrieved 500 results for anonymous operation orlando
('anonymous operation orlando', (500, 14))
(500, 8)
retrieved 500 results for anonymous operation antisec
('anonymous operation antisec', (500, 15))
(500, 8)
retrieved 477 results for anonymous operation facebook
('anonymous operation facebook', (477, 15))
(477, 8)
retrieved 500 results for anonymous operation bart
('anonymous operation bart', (500, 14))
(500, 8)
retrieved 454 results for anonymous operation shooting sheriffs
('anonymous operation shooting sheriffs', (454, 15))
(454, 8)
retrieved 500 results for anonymous operation wall street
('anonymous operation wall street', (500, 15))
(500, 8)
retrieved 483 results for anonymous operation syria
('anonymous operation syria', (483, 15))
(483, 8)
retrieved 500 results for anonymous operation darknet
('anonymous operation darknet', (500, 14))
(500, 8)
retrieved 224 results for anonymous operation los zetas
('anonymous operation los zetas', (224, 14))
(224, 8)
retrieved 462 results for anonymous operation brotherhood takedown
('anonymous operation brotherhood takedown', (462, 14))
(462, 8)
retrieved 459 results for anonymous operation megaupload
('anonymous operation megaupload', (459, 15))
(459, 8)
retrieved 497 results for anonymous operation black march
('anonymous operation black march', (497, 14))
(497, 8)
retrieved 469 results for anonymous operation acta
('anonymous operation acta', (469, 13))
(469, 8)
retrieved 457 results for anonymous operation pipa
('anonymous operation pipa', (457, 14))
(457, 8)
retrieved 500 results for anonymous operation sopa
('anonymous operation sopa', (500, 14))
(500, 8)
```

```
retrieved 477 results for anonymous operation blackout
('anonymous operation blackout', (477, 15))
(477, 8)
retrieved 500 results for anonymous operation kony
('anonymous operation kony', (500, 15))
(500, 8)
retrieved 500 results for anonymous operation greece
('anonymous operation greece', (500, 14))
(500, 8)
retrieved 500 results for anonymous operation quebec
('anonymous operation quebec', (500, 14))
(500, 8)
retrieved 500 results for anonymous operation japan
('anonymous operation japan', (500, 14))
(500, 8)
retrieved 230 results for anonymous operation blitzkrieg
('anonymous operation blitzkrieg', (230, 14))
(230, 8)
retrieved 419 results for anonymous operation anaheim
('anonymous operation anaheim', (419, 14))
(419, 8)
retrieved 498 results for anonymous operation steubenville
('anonymous operation steubenville', (498, 14))
(498, 8)
retrieved 500 results for anonymous operation darknet
('anonymous operation darknet', (500, 14))
(500, 8)
retrieved 458 results for anonymous operation paperstorm
('anonymous operation paperstorm', (458, 14))
(458, 8)
retrieved 500 results for anonymous operation aaron swartz
('anonymous operation aaron swartz', (500, 14))
(500, 8)
retrieved 500 results for anonymous operation federal reserve
('anonymous operation federal reserve', (500, 14))
(500, 8)
retrieved 458 results for anonymous operation north korea
('anonymous operation north korea', (458, 13))
(458, 8)
retrieved 491 results for anonymous operation israel
('anonymous operation israel', (491, 14))
(491, 8)
retrieved 470 results for anonymous operation last resort
('anonymous operation last resort', (470, 14))
(470, 8)
retrieved 500 results for anonymous operation vendetta
('anonymous operation vendetta', (500, 14))
(500, 8)
retrieved 471 results for anonymous operation cispa
('anonymous operation cispa', (471, 14))
(471, 8)
retrieved 488 results for anonymous operation NSA
('anonymous operation NSA', (488, 15))
```

```
(488, 8)
retrieved 500 results for anonymous operation american freedom
('anonymous operation american freedom', (500, 15))
(500, 8)
retrieved 488 results for anonymous operation big brother
('anonymous operation big brother', (488, 14))
(488, 8)
retrieved 500 results for anonymous operation V
('anonymous operation V', (500, 14))
(500, 8)
retrieved 500 results for anonymous operation westboro baptist church
('anonymous operation westboro baptist church', (500, 14))
(500, 8)
retrieved 500 results for anonymous operation support snowden
('anonymous operation support snowden', (500, 14))
(500, 8)
retrieved 128 results for anonymous operation kiwi freedom
('anonymous operation kiwi freedom', (128, 14))
(128, 8)
retrieved 493 results for anonymous operation world bank
('anonymous operation world bank', (493, 14))
(493, 8)
retrieved 493 results for anonymous operation awake the masses
('anonymous operation awake the masses', (493, 14))
(493, 8)
retrieved 500 results for anonymous operation cannabis
('anonymous operation cannabis', (500, 15))
(500, 8)
retrieved 278 results for anonymous operation rohingya
('anonymous operation rohingya', (278, 14))
(278, 8)
retrieved 123 results for anonymous operation rose awareness
('anonymous operation rose awareness', (123, 14))
(123, 8)
retrieved 393 results for anonymous operation rebuild the hive
('anonymous operation rebuild the hive', (393, 14))
(393, 8)
retrieved 500 results for anonymous operation bahrain
('anonymous operation bahrain', (500, 15))
(500, 8)
retrieved 500 results for anonymous operation ukraine
('anonymous operation ukraine', (500, 14))
(500, 8)
retrieved 500 results for anonymous operation pirate bay
('anonymous operation pirate bay', (500, 14))
(500, 8)
retrieved 458 results for anonymous operation petrol
('anonymous operation petrol', (458, 14))
(458, 8)
retrieved 268 results for anonymous operation ele
('anonymous operation ele', (268, 14))
(268, 8)
retrieved 282 results for anonymous operation guantanamo bay
```

```
('anonymous operation guantanamo bay', (282, 14))
(282, 8)
retrieved 500 results for anonymous operation safe winter
('anonymous operation safe winter', (500, 15))
(500, 8)
retrieved 480 results for anonymous operation maryville
('anonymous operation maryville', (480, 14))
(480, 8)
retrieved 386 results for anonymous operation green rights
('anonymous operation green rights', (386, 15))
(386, 8)
retrieved 500 results for anonymous operation fullerton
('anonymous operation fullerton', (500, 15))
(500, 8)
retrieved 477 results for anonymous operation american spring
('anonymous operation american spring', (477, 14))
(477, 8)
```

Run all the same queries with 'mirror' as well. This might help address the unpredictability of the Youtube search results and its 'denial of search results'. So, we undertake a 'supply of service' approach here.

In [9]:
```python
## run the same operations with 'mirror'


for op in operations_mirror:
    df   = yt.youtube_search(op,1000, True)
    df['operation'] = op
    df_ops = df_ops.append(df)
```

```
retrieved 8 results for anonymous operation habbo mirror
('anonymous operation habbo mirror', (8, 13))
(8, 8)
retrieved 6 results for anonymous operation hal turner mirror
('anonymous operation hal turner mirror', (6, 13))
(6, 8)
retrieved 4 results for anonymous operation chris forcand mirror
('anonymous operation chris forcand mirror', (4, 13))
(4, 8)
retrieved 229 results for anonymous operation chanology mirror
('anonymous operation chanology mirror', (229, 14))
(229, 8)
retrieved 13 results for anonymous operation palin mirror
('anonymous operation palin mirror', (13, 13))
(13, 8)
retrieved 245 results for anonymous operation iran mirror
('anonymous operation iran mirror', (245, 13))
(245, 8)
retrieved 0 results for anonymous operation didgeridie mirror
retrieved 10 results for anonymous operation titstorm mirror
('anonymous operation titstorm mirror', (10, 14))
(10, 8)
retrieved 45 results for anonymous operation tea party mirror
('anonymous operation tea party mirror', (45, 13))
(45, 8)
retrieved 248 results for anonymous operation payback mirror
('anonymous operation payback mirror', (248, 14))
```

```
(248, 8)
retrieved 473 results for anonymous operation wikileaks mirror
('anonymous operation wikileaks mirror', (473, 14))
(473, 8)
retrieved 202 results for anonymous operation assange mirror
('anonymous operation assange mirror', (202, 14))
(202, 8)
retrieved 117 results for anonymous operation leakspin mirror
('anonymous operation leakspin mirror', (117, 14))
(117, 8)
retrieved 39 results for anonymous operation visa mirror
('anonymous operation visa mirror', (39, 14))
(39, 8)
retrieved 45 results for anonymous operation mastercard mirror
('anonymous operation mastercard mirror', (45, 14))
(45, 8)
retrieved 73 results for anonymous operation paypal mirror
('anonymous operation paypal mirror', (73, 14))
(73, 8)
retrieved 78 results for anonymous operation tunisia mirror
('anonymous operation tunisia mirror', (78, 15))
(78, 8)
retrieved 133 results for anonymous operation egypt mirror
('anonymous operation egypt mirror', (133, 15))
(133, 8)
retrieved 97 results for anonymous operation libya mirror
('anonymous operation libya mirror', (97, 13))
(97, 8)
retrieved 40 results for anonymous operation hbgary mirror
('anonymous operation hbgary mirror', (40, 13))
(40, 8)
retrieved 62 results for anonymous operation empire state mirror
('anonymous operation empire state mirror', (62, 14))
(62, 8)
retrieved 182 results for anonymous operation sony mirror
('anonymous operation sony mirror', (182, 14))
(182, 8)
retrieved 73 results for anonymous operation india mirror
('anonymous operation india mirror', (73, 14))
(73, 8)
retrieved 27 results for anonymous operation malaysia mirror
('anonymous operation malaysia mirror', (27, 14))
(27, 8)
retrieved 43 results for anonymous operation orlando mirror
('anonymous operation orlando mirror', (43, 14))
(43, 8)
retrieved 171 results for anonymous operation antisec mirror
('anonymous operation antisec mirror', (171, 14))
(171, 8)
retrieved 494 results for anonymous operation facebook mirror
('anonymous operation facebook mirror', (494, 15))
(494, 8)
retrieved 100 results for anonymous operation bart mirror
```

('anonymous operation bart mirror', (100, 14))
(100, 8)
retrieved 25 results for anonymous operation shooting sheriffs mirror
('anonymous operation shooting sheriffs mirror', (25, 13))
(25, 8)
retrieved 493 results for anonymous operation wall street mirror
('anonymous operation wall street mirror', (493, 14))
(493, 8)
retrieved 370 results for anonymous operation syria mirror
('anonymous operation syria mirror', (370, 14))
(370, 8)
retrieved 49 results for anonymous operation darknet mirror
('anonymous operation darknet mirror', (49, 13))
(49, 8)
retrieved 14 results for anonymous operation los zetas mirror
('anonymous operation los zetas mirror', (14, 13))
(14, 8)
retrieved 46 results for anonymous operation brotherhood takedown
mirror
('anonymous operation brotherhood takedown mirror', (46, 13))
(46, 8)
retrieved 228 results for anonymous operation megaupload mirror
('anonymous operation megaupload mirror', (228, 14))
(228, 8)
retrieved 123 results for anonymous operation black march mirror
('anonymous operation black march mirror', (123, 14))
(123, 8)
retrieved 494 results for anonymous operation acta mirror
('anonymous operation acta mirror', (494, 14))
(494, 8)
retrieved 239 results for anonymous operation pipa mirror
('anonymous operation pipa mirror', (239, 14))
(239, 8)
retrieved 485 results for anonymous operation sopa mirror
('anonymous operation sopa mirror', (485, 15))
(485, 8)
retrieved 469 results for anonymous operation blackout mirror
('anonymous operation blackout mirror', (469, 15))
(469, 8)
retrieved 66 results for anonymous operation kony mirror
('anonymous operation kony mirror', (66, 14))
(66, 8)
retrieved 98 results for anonymous operation greece mirror
('anonymous operation greece mirror', (98, 14))
(98, 8)
retrieved 24 results for anonymous operation quebec mirror
('anonymous operation quebec mirror', (24, 13))
(24, 8)
retrieved 106 results for anonymous operation japan mirror
('anonymous operation japan mirror', (106, 13))
(106, 8)
retrieved 28 results for anonymous operation blitzkrieg mirror
('anonymous operation blitzkrieg mirror', (28, 14))

```
(28, 8)
retrieved 18 results for anonymous operation anaheim mirror
('anonymous operation anaheim mirror', (18, 13))
(18, 8)
retrieved 29 results for anonymous operation steubenville mirror
('anonymous operation steubenville mirror', (29, 13))
(29, 8)
retrieved 49 results for anonymous operation darknet mirror
('anonymous operation darknet mirror', (49, 13))
(49, 8)
retrieved 110 results for anonymous operation paperstorm mirror
('anonymous operation paperstorm mirror', (110, 14))
(110, 8)
retrieved 62 results for anonymous operation aaron swartz mirror
('anonymous operation aaron swartz mirror', (62, 13))
(62, 8)
retrieved 162 results for anonymous operation federal reserve mirror
('anonymous operation federal reserve mirror', (162, 14))
(162, 8)
retrieved 63 results for anonymous operation north korea mirror
('anonymous operation north korea mirror', (63, 14))
(63, 8)
retrieved 215 results for anonymous operation israel mirror
('anonymous operation israel mirror', (215, 14))
(215, 8)
retrieved 107 results for anonymous operation last resort mirror
('anonymous operation last resort mirror', (107, 14))
(107, 8)
retrieved 499 results for anonymous operation vendetta mirror
('anonymous operation vendetta mirror', (499, 14))
(499, 8)
retrieved 151 results for anonymous operation cispa mirror
('anonymous operation cispa mirror', (151, 14))
(151, 8)
retrieved 320 results for anonymous operation NSA mirror
('anonymous operation NSA mirror', (320, 14))
(320, 8)
retrieved 466 results for anonymous operation american freedom mirror
('anonymous operation american freedom mirror', (466, 14))
(466, 8)
retrieved 91 results for anonymous operation big brother mirror
('anonymous operation big brother mirror', (91, 13))
(91, 8)
retrieved 500 results for anonymous operation V mirror
('anonymous operation V mirror', (500, 15))
(500, 8)
retrieved 454 results for anonymous operation westboro baptist church
mirror
('anonymous operation westboro baptist church mirror', (454, 15))
(454, 8)
retrieved 38 results for anonymous operation support snowden mirror
('anonymous operation support snowden mirror', (38, 14))
(38, 8)
```

```
retrieved 3 results for anonymous operation kiwi freedom mirror
('anonymous operation kiwi freedom mirror', (3, 13))
(3, 8)
retrieved 282 results for anonymous operation world bank mirror
('anonymous operation world bank mirror', (282, 14))
(282, 8)
retrieved 368 results for anonymous operation awake the masses mirror
('anonymous operation awake the masses mirror', (368, 14))
(368, 8)
retrieved 134 results for anonymous operation cannabis mirror
('anonymous operation cannabis mirror', (134, 14))
(134, 8)
retrieved 5 results for anonymous operation rohingya mirror
('anonymous operation rohingya mirror', (5, 13))
(5, 8)
retrieved 5 results for anonymous operation rose awareness mirror
('anonymous operation rose awareness mirror', (5, 13))
(5, 8)
retrieved 25 results for anonymous operation rebuild the hive mirror
('anonymous operation rebuild the hive mirror', (25, 14))
(25, 8)
retrieved 27 results for anonymous operation bahrain mirror
('anonymous operation bahrain mirror', (27, 13))
(27, 8)
retrieved 87 results for anonymous operation ukraine mirror
('anonymous operation ukraine mirror', (87, 13))
(87, 8)
retrieved 47 results for anonymous operation pirate bay mirror
('anonymous operation pirate bay mirror', (47, 14))
(47, 8)
retrieved 10 results for anonymous operation petrol mirror
('anonymous operation petrol mirror', (10, 13))
(10, 8)
retrieved 26 results for anonymous operation ele mirror
('anonymous operation ele mirror', (26, 14))
(26, 8)
retrieved 20 results for anonymous operation guantanamo bay mirror
('anonymous operation guantanamo bay mirror', (20, 13))
(20, 8)
retrieved 32 results for anonymous operation safe winter mirror
('anonymous operation safe winter mirror', (32, 13))
(32, 8)
retrieved 6 results for anonymous operation maryville mirror
('anonymous operation maryville mirror', (6, 13))
(6, 8)
retrieved 22 results for anonymous operation green rights mirror
('anonymous operation green rights mirror', (22, 13))
(22, 8)
retrieved 47 results for anonymous operation fullerton mirror
('anonymous operation fullerton mirror', (47, 14))
(47, 8)
retrieved 301 results for anonymous operation american spring mirror
('anonymous operation american spring mirror', (301, 14))
```

```
        (301, 8)
        retrieved 4 results for anonymous operation hunthunter mirror
        ('anonymous operation hunthunter mirror', (4, 12))
        (4, 8)
        print ('overall video details:' + str(df_ops.shape))
In [28]: overall video details:(18444, 23)
```

1.3 Save and read data for offline use

Don't want to run the queries every time, so saving them locally.

```
         # local save  -- around 25Mb
In [32]: df_ops.to_pickle('data/operations_results.pyd')
```

```
         ops_df = pd.read_pickle('data/operations_results.pyd')
 In [3]:
         print('Dataset has %d rows'% ops_df.shape[0])
 In [4]: Dataset has 57936 rows
```

1.4 Dealing with the duplicates and finding mirrors in Youtube search results

We have run searches on Youtube that yield many results. But given that we are trying to look at 'mirroring,' how do we know whether a given result is a *duplicate* produced by our overlapping search queries or an actual *mirror*? We need to get ride of duplicate search query results and then identify mirror results

```
         ops_df.columns
 In [5]: Index([u'caption', u'channelId', u'channelTitle', u'commentCount',
Out [5]: u'contentRating', u'definition', u'description', u'dimension',
         u'dislikeCount', u'duration', u'favoriteCount', u'licensedContent',
         u'likeCount', u'liveBroadcastContent', u'operation', u'publishedAt',
         u'regionRestriction', u'relevantTopicIds', u'thumbnails', u'title',
         u'topicIds', u'videoId', u'viewCount'], dtype='object')
```

The columns that separate duplicates are

1. publishedAt
2. videoId

If two results are published at the same time, and have the same videoID, we assume they are duplicates. We can drop these from the dataset. n.b. If Youtube itself aggregates mirrored copies to a single videoID as part of its content management, we have a slight problem. But presumably the publication dates of mirrors would still be different.

```
         ops_no_dups = ops_df.drop_duplicates(cols = ['publishedAt', 'videoId'])
 In [4]: print('None duplicated results: %d' % ops_no_dups.shape[0])
         None duplicated results: 26477
```

```
         ops_no_dups.head()
 In [7]:      caption                 channelId      channelTitle commentCount  \
Out [7]: 0      false  UCIWTswhrKzualx_lctljxIQ                            18
         4      false  UCqzEDEKJ93zP44nM-c2p6lA  HabboNewsFlash             0
         8      false  UC6Z4I5naM4lJiAGnO1FKjhA     Sp00ndotcom             2
         12     false  UCDuWnJRkEJ4w61_5vnLa1qg    AnonymousOHP             0
         16     false  UCn9pu-dwF0y0QfCd3dxvTsw        XCIXNews             2

             contentRating definition
         0             NaN         hd ...
```

```
4            NaN        hd ...
8            NaN        sd ...
12           NaN        sd ...
16           NaN        sd ...

[5 rows x 23 columns]
```

1.5 Identifying mirrors

This leaves a dataset that has no duplicaten results (i.e rows). Finding mirrored videos is potentially harder. Trying to think this through:

1. If durations are the same and the titles overlap, we probably have a mirror. The practical problem here is finding the overlapping titles. Slightly different wordings for titles will be hard to track. And what about people who completely drop the titles when they mirror? This requires some slightly clever matching.
2. Conversely, distinguish videos that have the same title but are different videos. They will have different durations. Obviously descriptions, number of comments, etc, are likely to be different.
3. decide which of the remain videos are Anonymous operations as opposed to media coverage, commentaries, reports, attacks, etc on Anonymous operations

Using titles to find duplicates

In [83]:
```python
def title_clean_operations(ops_df):

    # clean up the operations in terms of duplicates,
    # only include videos whose title starts with 'anomymous operation
    # returns a dataframe with title_clean column
    ops = ops_df.drop_duplicates(cols = ['publishedAt', 'videoId'])
    ops['title_clean'] = ops.title.str.lower().str.replace('\\W+|#|-', ' ').str.strip(
    ops = ops.ix[ops.title_clean.str.startswith('anonymous operation')]
    # this line gets the first 4 words
    ops['title_short'] = ops.title_clean.map(lambda x: ' '.join(re.split(' ', x)[:3]))
    return ops
```

In [5]:
```python
# this shows some of the problems
ops = ops_no_dups

# strip all the double spaces, and non-word characters
print(ops.title.str.lower().str.strip().str.replace('\\W+', ' ').value_counts()[:20])
print('\n Videos with short titles\n')
print(ops[(ops.title.apply(len) < 4)]['title'])

ops['title_clean'] = ops.title.str.lower().str.replace('\\W+', ' ').str.strip()
```
```
anonymous                          175
                                   102
we are anonymous                    37
anonymous operation last resort     36
anonymous operation anti acta       35
...
anonymous operation revenge              19
anonymous operation fullerton            19
anonymous operation black march          18
anonymous operation american freedom     18
anonymous operation blackout             17
anonymous operation defense cispa        17
Length: 20, dtype: int64
```

```
  Videos with short titles
353     4th
464       8
478     Srt
245     Lab
477       i
...
341      16
309      10
318       8
70      CAR
1044    PSA
1059    She
Name: title, Length: 20, dtype: object
```

Some of the difficulties of finding mirrors start to appear here. We have around 100 videos with no title! 175 videos just called 'anonymous' – what operation do they belong to, if any?

The good news is that there are many cases where the standard 'Anonymous Operation X' pattern holds. We can work with those as long as we keep in mind that we are losing several hundred potential mirrors in the process.

In [76]:
```python
ops_cl = title_clean_operations(ops)
ops_cl = yt.format_durations(ops_cl)
ti_du = ops_cl.groupby('title_short')['duration_time'].value_counts()
print(ti_du.sum())
```
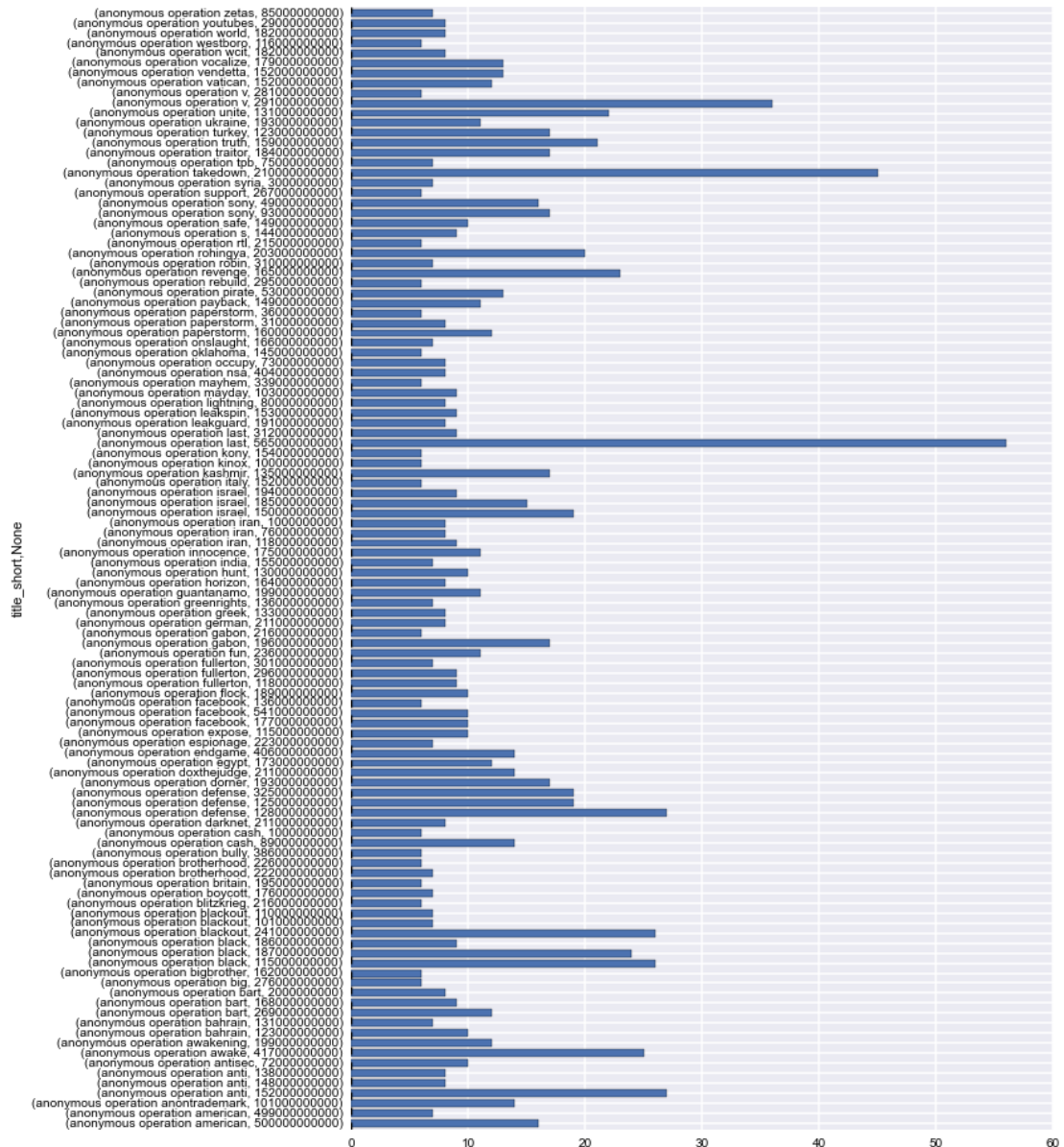```
2642
```

Working with these 2600 videos, we can get an idea of the extent and distribution of mirrors.

## 1.6  The main mirrors

Looking at the top 100 mirrored operation videos, we get this kind of distribution

In [79]:
```python
print(len(ti_du[ti_du>5]))
ti_du[ti_du>5].plot(kind='barh',figsize=(10,16))
```
```
112
<matplotlib.axes.AxesSubplot at 0x3b8acdd0>
```
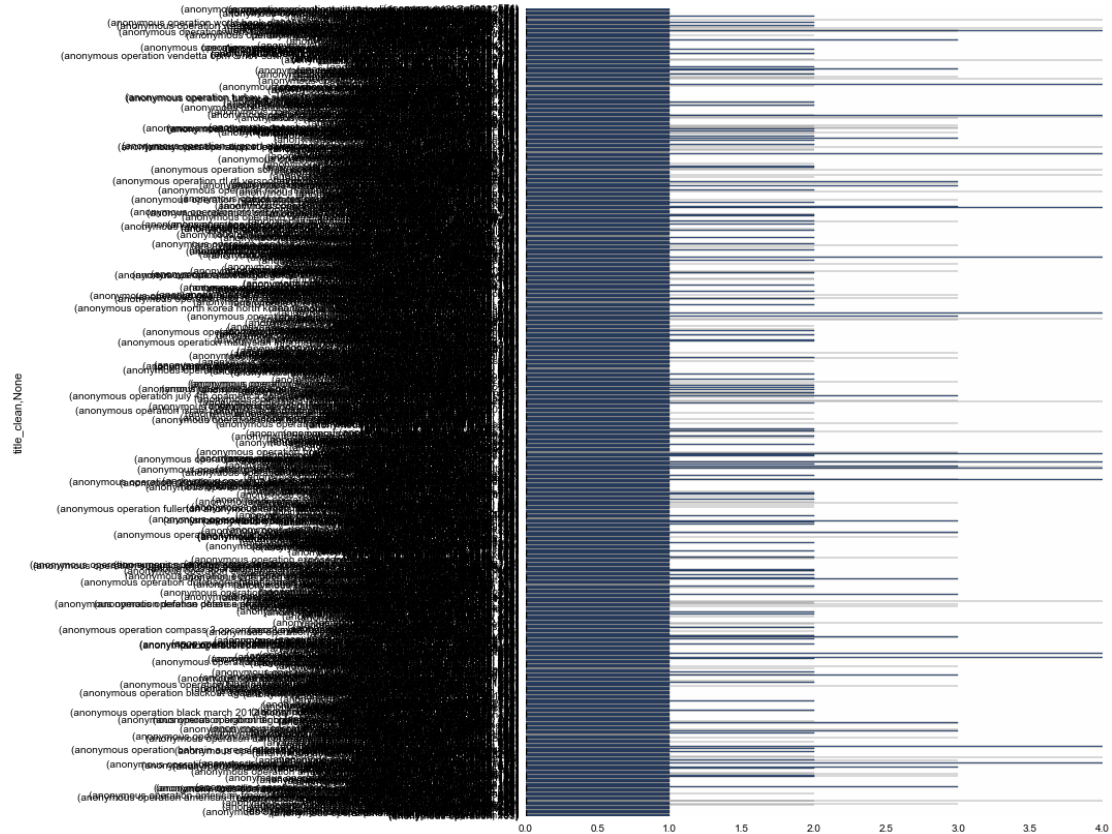
Out [79]:

Minor mirrors

Some videos are only mirrored a couple of times

```
In [35]:
print(len(ti_du[ti_du <= 4]))
ti_du[ti_du <= 4].plot(kind='barh',figsize=(10,14))
```

```
1379
<matplotlib.axes.AxesSubplot at 0x2666b6d0>
```

Out [35]:
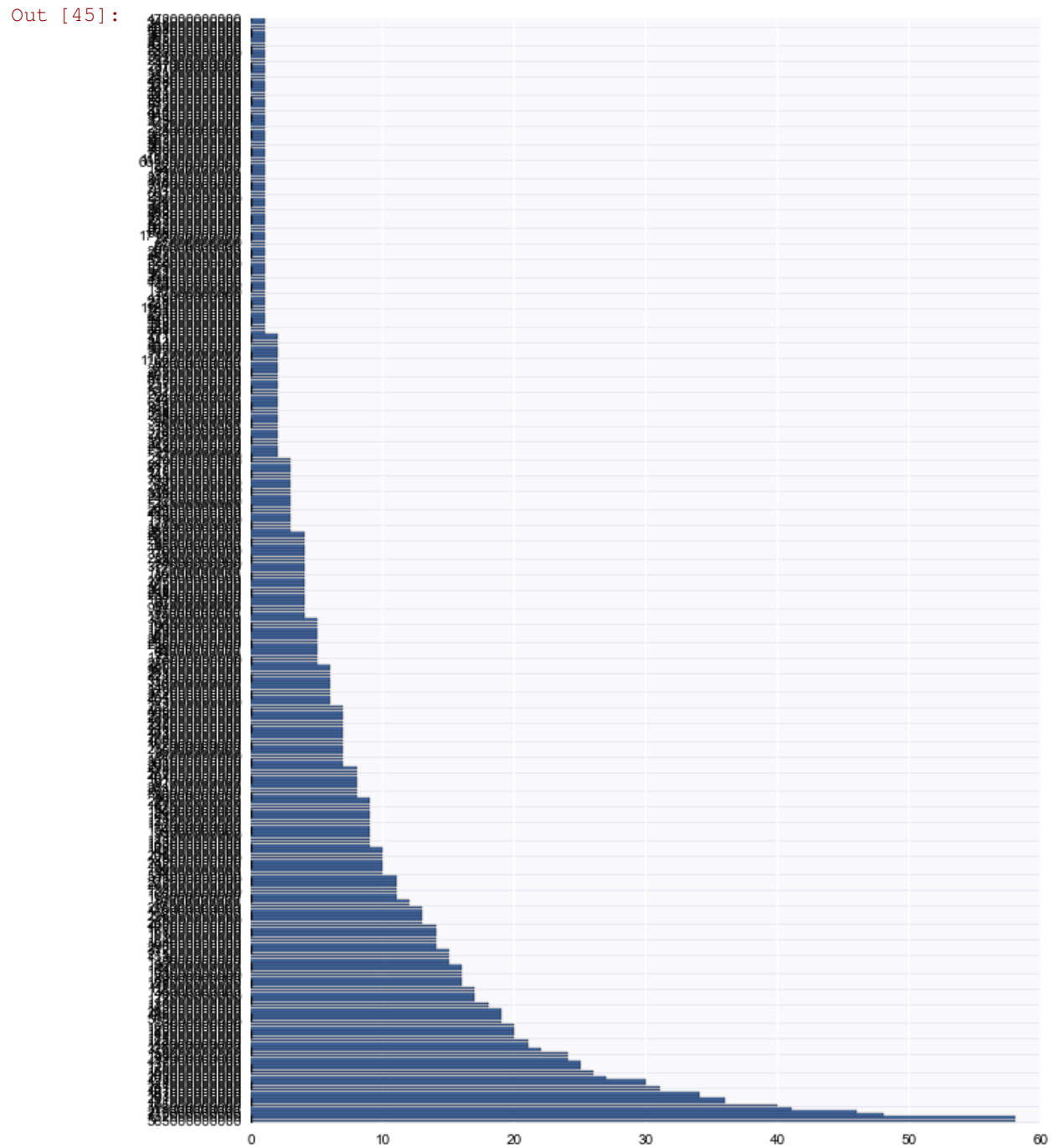
Sorting mirrored videos according to publication dates

```
         ops_cl['publishedAt'] = pd.to_datetime(ops_cl.publishedAt)
In [82]:  first_publication = ops_cl.groupby(by = ['title_short', 'duration_time']).publishedAt.
         print(first_publication)
```

```
title_short                        duration_time
anonymous operation                69000000000      2013-07-24 04:06:48
                                   103000000000     2013-08-28 04:07:22
                                   257000000000     2013-07-24 02:13:28
                                   541000000000     2011-08-31 13:35:52
anonymous operation 2012           127000000000     2012-04-26 10:32:09
                                   ...
anonymous operationgreenrights pressrelease   372000000000
2011-07-05 00:40:52
anonymous operationitaly 12                   141000000000
2012-01-11 23:05:25
anonymous operations part                     401000000000
2013-01-12 16:01:26
anonymous operationsyria chem                 271000000000
2012-12-07 22:13:43
anonymous operationtwitterstorm fr            96000000000
2012-08-11 18:37:56
anonymous operationz                          571000000000
2012-01-19 09:05:45
Name: publishedAt, Length: 992, dtype: datetime64[ns]
```

Using durations to detect mirrors

```
# might be better to cross check with other fields such as duration, publishedAt, etc
ops_cl.duration_time.value_counts().plot(kind='barh', figsize=(10,14))
<matplotlib.axes.AxesSubplot at 0x33216950>
```

```
ops_pure.duration_second.value_counts()
```

```
        565    60
        152    58
        210    48
        211    46
        241    42
        ...
        354     1
        358     1
        374     1
        410     1
        416     1
        1489    1
        Length: 375, dtype: int64
```

This suggests that durations could be some help. There are 60 videos that are 565 secs, 58 that are 152 secs, etc. If titles have some match, and durations are close, then is that a mirror?

In [ ]:
```python
def mirror_match
```

1.7 Quantifying mirroring on Anonymous

This leaves us with a dataset of 2800 videos to look at. Bear in mind this does not include the videos that do not use the term 'operation'. But it does answer the initial question we had. For around 80 Anonymous operations, there are at least 2800 mirrors.

In [67]:
```python
print('Most mirrored operations:\n')
print(ops_cl.title_clean.value_counts()[:10])
```
```
Most mirrored operations:

anonymous operation last resort            38
anonymous operation anti acta              37
anonymous operation black march            34
anonymous operation takedown november 5th 2012   33
anonymous operation israel                 29
anonymous operation facebook               23
anonymous operation unite                  22
anonymous operation revenge                21
anonymous operation defense cispa          20
anonymous operation american freedom       19
dtype: int64
```

In [32]:
```python
print('Least mirrored operations:\n')
print(ops_pure.title_clean.value_counts()[-10:])
```
```
Least mirrored operations:

anonymous operation occupy wall street downplayed by cnn      1
anonymous operation alsatia the next american revolution      1
anonymous operation cash back november 5th 2011 press release 2    1
anonymous operation mp4                                        1
anonymous operation payback wikileaks acta laws message       1
anonymous operation do not forgive bullies education of criminal
bullies    1
anonymous operation israel help gaza read description         1
habboanonymous operation regahotel se                         1
anonymous operation last defense 720p                         1
```
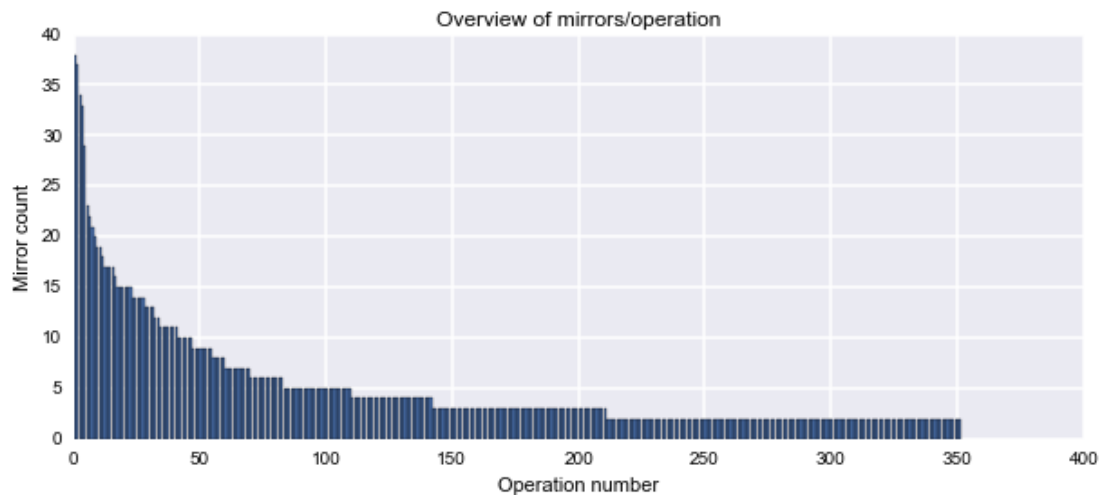
```
            anonymous operation mayhem 2012 vostfr avi                    1
            dtype: int64
```

The good thing about this list now is that it does seem to have much 'media noise' in it. This list looks like pure anonymous. Only in operations that seem to have no mirrors does noise appear.

```
In [68]:   pl = plt.figure(figsize = (10,4))
           title_counts = ops_cl.title_clean.value_counts().values
           title_counts = title_counts[title_counts>1]
           xval = range(0,len(title_counts))
           yval = title_counts
           plt.bar(xval, yval)
           plt.title('Overview of mirrors/operation')
           plt.xlabel('Operation number')
           plt.ylabel('Mirror count')

           <matplotlib.text.Text at 0x3926e110>
```

Out [68]:


```
           len(title_counts)
In [44]:   373
Out [44]:
```

This suggests that operations have anywhere between 36 and 1 mirrored videos. But there are still too many titles for operations. There are around 370 unique titles of videos that fit the 'operation anonymous X' pattern, but we know of only around 80 Anonymous operations. What is overflowing here?

```
           # begin to see the problem here -- titles are mangled in various ways
In [42]:   print(ops_pure.title_clean.unique())

           ['anonymous operation summerstorm english'
            'anonymous operation mayhem poster speed art'
            'habboanonymous operation habcity' ...,
            'anonymous operation last resort re uploaded'
            'anonymous operation nsa english' 'anonymous operationfullerton']
```

```
           # if we only look at the first three words?
In [68]:   ops_pure['title_short'] = ops_pure.title_clean.map(lambda x: ' '.join(x.split(' ')[:3]

           print(ops_pure.title_short.value_counts())
           print(sum(ops_pure.title_short.value_counts()>1))
           anonymous operation israel      77
           anonymous operation black       74
           anonymous operation last        72
           anonymous operation defense     70
           anonymous operation blackout    66
           ...
```

```
anonymous operation powerbot           1
anonymous operation kingfish           1
anonymous operation occupythemedia     1
anonymous operation game               1
operationexpose anonymous operation    1
anonymous operation indignant          1
Length: 541, dtype: int64
240
```

OK, that's a bit better. 240 operations appearing now. We might need to consider the possibility that our operations list is nowhere near complete

```
ops_pure.title_short.to_csv('ops_list_to_check.csv')
```

In [70]:

This is obviously still not clean enough because many of the titles do not begin with 'anonymous operation'. Need to apply a stricter filder

In [74]:
```
ops_pure2 = ops_pure.ix[ops_pure.title_clean.str.startswith('anonymous operation')]
ops_pure2['title_short'] = ops_pure2.title_clean.map(lambda x: ' '.join(x.split(' ')[:

print(ops_pure2.title_short.value_counts())
print(sum(ops_pure2.title_short.value_counts()>1))
```
```
anonymous operation israel      77
anonymous operation black       74
anonymous operation last        72
anonymous operation defense     70
anonymous operation blackout    66
...
anonymous operation antiwar      1
anonymous operation xbox         1
anonymous operation indonesia    1
anonymous operation reborn       1
anonymous operation kurobasu     1
anonymous operation flash        1
Length: 435, dtype: int64
220
```
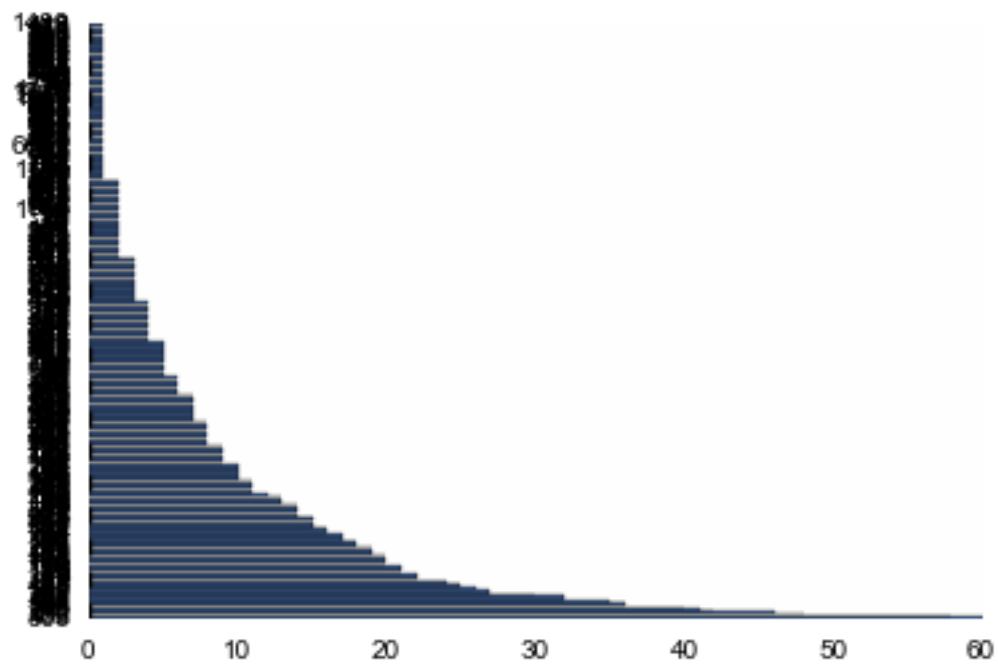
So now we are down to 220 operations.

In [82]:
```
print(ops_pure2.title_short.value_counts())
f = open('data/found_operations.csv', 'wb')
f.writelines('\n'.join(ops_pure2.title_short.unique()))
```
```
anonymous operation israel      77
anonymous operation black       74
anonymous operation last        72
anonymous operation defense     70
anonymous operation blackout    66
...
anonymous operation antiwar      1
anonymous operation xbox         1
anonymous operation indonesia    1
anonymous operation reborn       1
anonymous operation kurobasu     1
anonymous operation flash        1
Length: 435, dtype: int64
```

Ook,