Where the Pirates Are

Christian Bauckhage
http://mmprec.iais.fraunhofer.de

Fraunhofer IAIS
St. Agustin, Germany

B-IT, University of Bonn
Bonn, Germany

ABSTRACT

We are interested in the global characteristics and dynamics of illicit downloads of proprietary content from the Internet. Since corresponding data is difficult to obtain directly, we rely on secondary sources and analyze time series data available from search engines. In this paper, we present initial results obtained from this approach. In particular, we consider the example of blockbuster movies of 2010 and find that download related queries to the Google search engine and a special purpose Torrent search engine are strongly correlated. Moreover, aggregated data from Google Insights provides an estimate as to where these queries predominantly originate from.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communication Applications—*Internet*

General Terms

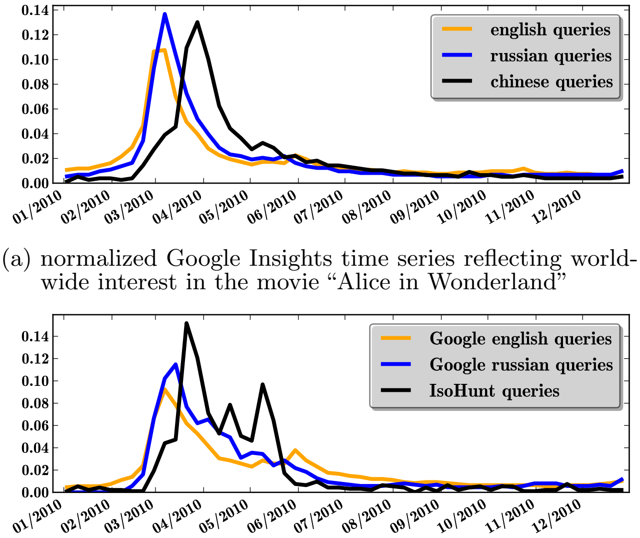
Economics, Experimentation, Measurement

Keywords

Illegal downloads, Internet piracy, Torrent searches

1. INTRODUCTION

Copyright infringements and illegal downloads of music, movies, or computer games are a common phenomenon on the Internet. Since the advent of peer-to-peer file sharing technologies in the late 1990s, software piracy has shifted to freely accessible though hardly mainstream services. Protocols such as BitTorrent distribute large files over a network of peers and users of corresponding client software can join a "swarm" of hosts for simultaneous down- and upload from each other. In order to find sources for BitTorrent downloads, users rely on special purpose search engines or indices many of which became notorious after police raids and law suits in recent years.



(a) normalized Google Insights time series reflecting worldwide interest in the movie "Alice in Wonderland"

(b) normalized Google Insights and IsoHunter time series of queries for "Alice in Wonderland" and "torrent"

Figure 1: Examples of time series that characterize the temporal dynamics of blockbuster movie related queries to the Google search engine and to IsoHunt, a BitTorrent index.

The work and results reported in this paper represent first steps towards understanding the dynamics of activities on the Internet that are related to illegal downloads. We briefly summarize an analysis of data that were retrieved from Google Insights [1] and IsoHunt [2]. We find that user queries to these search engines which indicate an interest in downloading copyrighted material are strongly correlated. Moreover, we present an estimate as to from which countries such downloads predominantly originate. We observe a correlation to countries that are known for botnet activity.

2. DATA COLLECTION

We examine the statistics of searches for the top 20 highest grossing movies of 2010. For each movie, we gathered data from Google Insights indicating global interest in its title. Also, we gathered data on searches that additionally include terms such as "download" or "torrent".

Google Insights provides statistics on queries terms users have entered into the Google search engine. It displays how

WebSci'11, June 14-17, 2011, Koblenz, Germany.

Table 1: Ranking of regions of origin of movie related, English queries to the Google search engine

	aggregate Google Insights rankings			OECD ranking
rank	queries for movie titles	queries for movie titles and 'download'	queries for movie titles and 'torrent'	# infected computers[†]
1	United States	United States	United States	India
2	Canada	India	Netherlands	Brazil
3	Australia	United Kingdom	India	Chile
4	Singapore	Portugal	United Kingdom	Poland
5	United Kingdom	Brazil	Sweden	Slovakia
6	Philippines	Indonesia	Australia	Turkey
7	Ireland	Romania	Canada	Indonesia
8	New Zealand	Philippines	Romania	Czech Republic
9	Malaysia	Australia	Portugal	Israel
10	Romania	Canada	Philippines	Russia
11	Netherlands	Germany	Italy	Hungary
12	India	Malaysia	Belgium	Portugal
13	United Arab Emirates	Netherlands	France	Greece
14	Portugal	Italy	Germany	Slovenia
15	Sweden	Egypt	Poland	Spain
16	Belgium	Spain	Croatia	Korea (Rep.)
17	Mexico	Viet Nam	Spain	Estonia
18	Denmark	Singapore	Hungary	Italy
19	Norway	Poland	Brazil	Ireland
20	Indonesia	United Arab Emirates	Greece	Mexico

[†]per subscriber at Internet Sevice Provider in 2009

Table 2: Spearman's correlation ρ between rankings in Table 1 and p-value for the null hypothesis of uncorrelatedness

	m,m+d	m,m+t	m+d,m+t	m,o	m+d,o	m+t,o
ρ	0.62	0.56	0.70	-0.15	0.13	0.06
p	0.00	0.00	0.00	0.17	0.19	0.34

m=movie title, d=download, t=torrent, o=OECD

frequently a query has been used in a given period (the year 2010 in our case). Google Insights returns normalized data which indicate relative search frequencies and do not allow for estimating interest in a topic in terms of absolute numbers. The service also provides similarly normalized rankings of the regions of origin of a query.

For the same set of movie titles, we also collected data on 12,086 torrents from IsoHunt.

IsoHunt is a BitTorrent tracker indexing more than 20 million peers and more than 10 petabytes of data. Upon entering a query, it returns related Torrents and also provides information as to how long they have been active.

The collected data were converted into a format representing average weekly activities for the period from January to December 2010. Thus, the resulting discrete time series $\vec{z} = [z_1, z_2, \ldots, z_T]$ cover a period $T = 52$ weeks. To compare activities across different sources, the data were turned into discrete probability vectors $\vec{x}$ where $x_t = z_t / \sum_i z_i$. Examples of normalized time series are shown in Fig. 1

3. BRIEF SUMMARY OF RESULTS

Overall, we observed a high degree of correlation between Google Insights time series for movie related queries in different languages (English, Russian, and Chinese). Google queries with the additional search terms "torrent" or "download" were also closely correlated across different languages, as were corresponding time series from Google Insights and IsoHunter. Figure 1 exemplifies these tendencies; for lack of space, we omit a more detailed discussion at this point.

Using Google Insights data on queries that indicate interest in movie titles and illicit downloads, we analyzed if there are regional tendencies. Table 1 displays aggregated rankings (over 20 movies) of the regions of origin of different types of English queries to the Google search engine. Searches hinting at an interest in movies as well as in illegal downloads predominantly originate from the United States. At the same time, it appears that countries from central and southern Europe, South Asia, and South America (see e.g. Germany, Italy, India, and Brazil) rank higher for searches aiming at movie downloads. For quantification, we computed Spearrman's correlation coefficient between these rankings and a recent ranking of botnet infested countries published by the OECD [3] (see 5th column of Tab. 1). While we did not find a positive correlation between purely movie related queries and botnet activity, interest in illegal downloads appears to go along with botnet activity (see 1st row of Tab. 2). Even though the null hypothesis of no correlation between the botnet ranking and the download related rankings cannot be clearly rejected, we conjecture that the rather high ranks of non-English speaking countries in columns 3 and 4 of Tab. 1 and their positive correlations with column 5 indicate that a growing number of illegal downloads happen via proxy networks and hijacked computers. We are currently collecting more data on downloads of copyrighted material to verify if similar trends can be observed for other languages and content other than movies.

4. REFERENCES

[1] http://www.google.com/insights/search/.

[2] http://isohunt.com/torrents/.

[3] M. von Eeten, J. Bauer, H. Asghari, and S. Tabatabaie. The Role of Internet Service Providers in Botnet Mitigation. STI Working Paper 2010/5, Organisation for Economic Co-operation and Development, 2010.