НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Международный Институт Экономики и Финансов

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

по образовательной программе высшего профессионального образования, направление 080100.68 Экономика

на тему: Динамический анализ кассовых сборов фильма

Dynamic analysis of box office revenue

Студент 2 курса магистратуры
Климова Кирилла Юрьевича

Научный руководитель
Профессор Сирченко Андрей Александрович PhD,
Профессор Деан Фантаццини, PhD.

МОСКВА, 2014 год

АННОТАЦИЯ

на магистрскую диссертацию студента 2го курса магистратуры
Международного института экономики и финансов
Климова Кирилла Юрьевича
по теме «Динамический анализ кассовых сборов фильма»

Практически каждый из нас хотя бы раз в своей жизни, так или иначе, сталкивался с продуктами, относящимися к индустрии кино: многие смотрят телевизор, некоторые ходят в кинотеатры, покупают DVD диски. Всё это обеспечивает огромный охват и, соответственно, огромные возможности для данной индустрии не только в плане искусства, но также и в плане получения существенных прибылей. Таким образом, создание фильма можно рассматривать как инвестиционный проект, бесспорно, очень сложно прогнозируемый и дорогостоящий, но при грамотном менеджменте и удачном стечении обстоятельств такой проект может существенно обогатить своего инвестора. Для примера, можно рассмотреть фильм «Паранормальное явление» (2007 год), без участия звезд мирового кино, с бюджетом в 15 000 $, который только на кассовых сборах заработал более 193 миллионов долларов (то есть проект с каждого потраченного на него доллара принес в 12 000 раз больше) и тем самым существенно обогатил инвесторов, рискнувших вложить в него свои деньги. С другой стороны, известны случаи, когда фильмы с безупречными статическими начальными данными (бюджет, актерский состав, режиссеры, сюжет) «проваливались» в мировом прокате. Например, фильм «Терминал», снятый в 2004 году, входящий в список 250 лучших фильмов мира по версии известного русскоязычного портала kinopoisk.ru, с бюджетом в 60 миллионов долларов и при участии таких звезд мирового кино как Стивен Спилберг, Том Хэнкс и Кэтрин Зета-Джонс, заработал всего лишь 70 миллионов долларов (что намного ниже прогнозируемого уровня). Стоит отметить, что сборы этого фильма в первую неделю несущественно отличались от сборов более успешных проектов с участием Тома Хэнкса. Данный пример демонстрирует, что прибыльность фильма часто определяется не столько статическими начальными данными, сколько тем, что происходит в динамике (внимание к фильму, пиратство и пр.). Простой статический анализ, на наш взгляд, имеет очень ограниченные прикладные возможности в отношении описания эффектов, влияющих на сборы фильмов. Например, эффект выхода пиратского контента (который для дистрибьютеров фильмов является одним из самых важных и проблемных) просто некорректно рассматривать в статике, так как в зависимости от времени выхода копии в итоге фильм теряет существенно разные доли своей выручки. Возможно, в данном случае более правильно

говорить в терминах потерь после выхода этого контента и механизмов их минимизации, что подразумевает динамический анализ.

Учитывая всё вышесказанное, при написании данной работы мы ставили своей целью провести максимально полное динамическое исследование по оценке зависимостей подневных кассовых сборов фильмов, которые выходили в американский прокат за последние 10 лет (с 2004 по 2014 год), от таких динамически изменяющихся параметров как количество кинотеатров, выходные и праздничные дни во время проката, внимания аудитории к фильму, пиратство в зависимости от качества выпускаемой нелегальной копии и самого фильма, а также доступности этой копии для простого пользователя сети и пр. Для оценки внимания к фильму мы использовали статистику поисковых запросов Google по ключевому слову, соответствующему названию данного фильма, в период его проката в США, ограниченную по категории и региону. Что касается анализа проблемы пиратства, одной из задач нашей работы была оценка потерь кассовых сборов фильма после выхода нелегального контента в хорошем и плохом качестве для кинокартин разного уровня. Для детального анализа данной проблемы была предпринята попытка оценки так называемых «затрат на поиск» (searching costs). Одной из гипотез, которую мы хотели протестировать, было предположение, что после первого выхода пиратской копии (этот момент определялся с использованием сайта VCDQ.com) и до появления этой копии на крупных ресурсах (в нашем случае, на популярном торрент-трекере The Pirate Bay) данный контент не является легкодоступным для простого пользователя, поэтому его влияние на кассовые сборы фильма не такое разрушительное, как могло бы быть.

Стоит отметить, что для данного исследования был собран и обработан огромный объем исходных данных. Так для анализа было выбрано 1264 фильма, при этом полная база (произведение дней в прокате каждого фильма на количество фильмов) состояла из 81787 строк.

В итоге в данной работе нами были получены следующие результаты:

- Эффект зависимости выручки от количества кинотеатров крайне нелинеен и обладает свойством убывающего эффекта от масштаба (показатель степени равен 0,77 и статистически существенно отличается от единицы).
- Выходные увеличивают выручку на 86%, 125% и 79% для пятницы, субботы и воскресенья, соответственно. Праздники, к тому же, дополнительно добавляют еще 57% к выручке.
- Наибольшая связь между поисковыми запросами и сборами фильма обнаруживается с задержкой в 1 день. Более того, исследование показывает, что эффект от точечного повышения внимания к фильму рассеивается за 1-2 недели.

- Эффект пиратства зависит как от качества фильма (которое измерялось с помощью IMDb рейтинга: фильм признавался плохим, если его рейтинг меньше 6,00), так и от качества самого пиратского контента. Для условно плохого фильма при выходе экранной копии (снятой на камеру в кинотеатре) выручка падает на 33%, а при выходе копии в хорошем качестве – на 52%. Для условно хорошего фильма потери выручки составляют 26% и 33%, соответственно. Можно сказать, что чем хуже фильм и чем лучше качество пиратского контента – тем больше потери от выхода нелицензионной копии.
- «Затраты на поиск» (searching costs) существенны, так как при появлении пиратской копии на известных интернет ресурсах (в частности, на The Pirate Bay) кассовые сборы уменьшаются дополнительно на 12-20%.

По результатам динамического исследования был проведен статический анализ зависимости основных индивидуальных констант, определяющих динамику сборов фильма от начальных параметров фильма (продолжительность фильма, участие звезд, жанр и пр.).

Полученные результаты вместе с собранной базой могут быть использованы как в качестве основополагающих факторов, определяющих поведение и основные риски дистрибьютеров во время проката, так и для дальнейших исследований с применением более сложных структурных моделей. Все это может быть полезно при выработке статического плана и динамических стратегий производства и проката фильма для увеличения и стабилизации его итоговой выручки.

Table of Contents

Introduction

Almost everyone on the planet has been exposed to the movie industry in one way or another at least once in his or her life. Ever since the creation of the first film in 1906, and the first studio in 1911, the industry saw tremendous growth, quickly turning into one of the most prominent areas of business.

Creating a new film often requires a huge amount of investment, which is quite risky since it is difficult to predict the box office revenue of the movie before its production begins. Some factors, such as celebrities' participation, a well-known director or a scriptwriter, big budget or even favourable market conditions may raise potential revenues. Nevertheless, they are far from being either necessary or sufficient conditions for a movie to be a success. For example, "Paranormal activity" (2007), with the budget of 15000$ and no celebrities in the cast, earned more than $193mln, which implies the profitability ratio of over 12 000! However, there are also a lot of examples when the movie investment was not so profitable. For instance, "Sahara" (2005), despite substantial production costs of $241mln and Penelope Cruz participation, earned just a little more than $110mln. "The Terminal" (2004), ranking 147 in the list of 250 best movies at kinopoisk[1], would be another example. Although it was directed by Steven Spielberg (3 Oscars and 12 Nominations) and included such celebrities as Tom Hanks (2 Oscars and 3 Nominations) and Catherine Zeta-Jones (1 Oscar) among the cast, the film earned a relatively modest amount of $77mln in domestic total gross revenue, barely exceeding its production costs of $60mln. It is interesting to note that, while the movie earned about $13.74mln during the opening weekend, which was quite similar to some other Tom Hanks projects, such as "The Green Mile" ($13.37mln) and "Road to Perdition" ($15.47mln), the total gross of "The Terminal" turned out to be about twice as low as the average total domestic gross of all the other movies with the same actor.

These examples, in particular the latter, show that a simple consideration of the standard exogenous static parameters might be not enough for accurate and robust prediction of the box office revenue. Instead, we should consider changes in dynamic parameters throughout the whole screening period. Potentially, this could allow us to promptly identify the reason and the moment when something starts going wrong with the movie produced, and aptly respond to stabilize or even increase its box office revenue.

[1] http://kinopoisk.ru

Arguably, one of the most important dynamic factors for a movie's financial success is the marketing policy and, respectively, public attention. When discussing related literature, we demonstrate that Google search volume indices (SVI), provided by Google Trends service, could be successfully used to measure the degree of public interest. There is a set of papers which show that performance of the models with Google SVI is significantly better than the performance of the models without this indicator (e.g. Varian and Choi (2009), Kholodin et al (2010) etc.). In a different setting, it has been demonstrated that Google SVI has a substantial predictive power in the influenza epidemic spread forecasts (Ginsberg et al. (2009)), which led to the subsequent creation of a new service, Google Flu Trends, with the aim to notify people and prevent a further spread of the disease.

With widespread technological growth, however, there has emerged another important factor that could substantially affect the box office revenue - piracy. Illegal downloads have been the driving force behind many changes in the media industry during the last decade or two. Before the piracy became really widespread, musicians, for instance, could earn enough money by issuing new albums and making small concert tours in their support. Today, when every person with internet access can easily listen to any song on the web (sometimes even before the official release), the situation has changed dramatically: now musicians are forced to make big tours all over the world after the album issue to support the revenues.

The problem of piracy is especially acute in Russia due to historically relatively weak enforcement of the intellectual property laws, with only recent antipiracy law enactment. However, with the widespread internet penetration and ever increasing data exchange speeds, piracy is no longer limited to a particular country, but is a global phenomenon.

Substantial resources are spent on fighting the piracy. There are special persons whose work involves protecting a movie from being stolen on the pre-release step or during the box office period in the form of a sample copy by the cinema employees. They also monitor social networks and torrent trackers in order to send requests to block the illegal content should it become available after the movie theatre release. Such proactive monitoring may prove it more difficult for the users to find an illegal copy of a movie, essentially increasing the corresponding searching costs and thereby reducing consumption of the illegal content and mitigating the negative effect of piracy on the revenue.

Going back to the example of "The Terminal", piracy might have indeed played a significant role in its financial failure: after all, this was the year when The Pirate Bay, probably the largest torrent tracker, was founded and instantly gained a huge popularity. In fact, the first

illegal copy of “The Terminal” in good quality appeared on the site only 8 days after the actual movie release. This could explain, for instance, why the initial movie revenue was comparable to the similar Tom Hanks movies, but subsequent returns turned out to be substantially lower. Therefore, we consider piracy as one of the important factors influencing financial success of the projects in the movie industry, and incorporate it into our model.

This paper suggests a reliable dynamic reduced-form model for the movie revenue as a function of various filming and screening characteristics, including the degree of public attention, the distribution chain features, star participation, runtime, various calendar effects, etc. Most importantly, we pay particular attention to the impact of electronic piracy on the financial success of the movie. We discuss how the effect will depend on the various characteristics of the latter, i.e. film grade, the quality of the illegal copy and its availability for an unsophisticated internet user.

In order to model the effect of various film characteristics on the box office revenue, we have created a new database of the movies released in the USA during the last decade. This involved writing a special script that would extract the relevant information according to particular algorithms from the various publicly available sources and combine it in a single dataset. The database contains a lot of information regarding particular films, including (but not limited to) the presence of stars in the cast, daily revenues dynamics, distribution details, calendar effects for the state holidays/weekends, movie genre, runtime, budget, and other characteristics. Piracy presence is measured by screening the contents of The Pirate Bay tracker and VCDQ illegal content list, paying particular attention to the pirated content release date and its quality. Finally, we also trace the Google Search Volume Index that relates to a particular movie from a month prior to the film release up to the end of the screening period. This allowed us to create a unique dataset with the information covering over 1200 movies produced during the last decade (to be more precise, in average each movie had approximately 70 time points which resulted in more than 80,000 records total, used for the dynamic analysis). The full description of the dataset and its composition/algorithms could be found in the Data section.

We build a dynamic reduced-form model for the daily stream of box office revenues and make a number of interesting empirical findings.

- The effect of an additional cinema theatre on screening revenues is highly nonlinear, and there appear to be decreasing marginal returns.

- As expected, calendar effects have a substantial impact: during the public holidays the revenue on average increases by 57% and on Friday/Saturday/Sunday by 86/125/79% respectively.

- Public attention, as measured by prior values of the Google Search Volume Index, seems to have the highest effect on the future movie attendance within the span of one to 2 weeks.

- Piracy impact on the film revenue depends on the quality of the movie (as measured by the IMDB rating with a threshold of 6.0) and its electronic copy. In particular, other things being equal, the availability of a low quality illegal copy of the movie decreases daily revenues by 33% for "bad" movies compared with only 26% for "good" ones. High quality copy leads to revenue changes of 52% and 33% accordingly. We also prove that the effects of "bad" and "good" quality illegal content vary significantly inside and between these groups.

- Searching costs have a substantial impact on the illegal content consumption and its effect on box office revenue. We find that the availability of a movie on a renowned tracker (such as The Pirate Bay) has an additional negative effect on the daily revenues of 12-20% (depending on the quality of the film and its electronic copy) and this effect is statistically different between the groups.

- Overall the results suggest that the piracy is related to a substantial negative impact on the box office revenue, and that the worse is the movie, the stronger it seems to be affected by it.

The model developed in this paper has impressive explanatory power (as measured by the in-sample R-squared), and could be empirically tested to develop the recommended distribution scheme depending on the movie features. The economic impact of the piracy is assessed and shown to be decreasing with the quality of the movie and associated searching costs.

The paper is organized as follows:

- Section 1 provides a literature review
- Section 2 discusses the database collected for the research
- Section 3 introduces the main reduced-form model for the numerical dynamic analysis
- Section 4 presents the results of the dynamic analysis

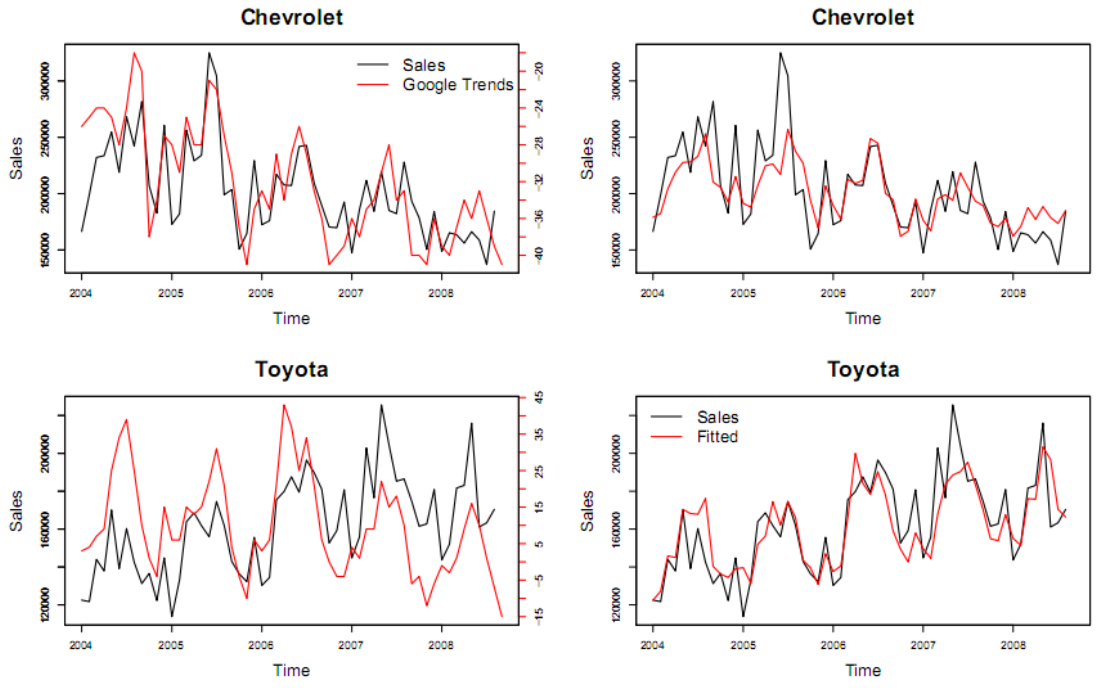
- In section 5 we discuss the static modeling results and the effects of the static movie characteristics on the box office revenue path-defining individual variables
- Section 6 provides main results and conclusions
- Section 7 contains a small survey of possible ideas for the future analysis and unresolved questions

Literature Review

Search volume data:

It is relatively recently that Google search volume index variables have started to appear in economics-related scientific works. One of the pioneering papers in this area where predictive power of the index was exposed and exploited is Varian and Choi's "Predicting the present with Google trends" (2009). In this article the authors show how Google search data might be used to predict economic activities in the variety of different areas. The authors demonstrate that sales in different areas such as automotive, home, car insurance sales etc. strongly correlate with search volume indices for the appropriate keywords in the investigated categories (see the picture below).

Furthermore, the model predictions obtained by running the simplest regressions



(a) Sales vs. Google Trends (b) Actual & Fitted Sales

Picture 1

including SVI match the real data fairly well.

Then, Schmidt and Vosen (2009) examine models based on both search volume indices and public surveys and concluded: "Google Trends is a very promising new source of data to forecast private consumption. In almost all experiments conducted by the authors the Google

indicators' in-sample and out-of-sample predictive power proved to be better than that of the conventional survey-based indicators".

Also Kolodin et al. (2010) investigate private consumption models and find statistically significant evidence that models with Google search statistics data offer an improvement over the benchmark models without SVI.

The most interesting paper concerning box-office revenue prediction, in which scientists used search volume index, was written by Goel et al (2010). In that paper the authors have shown that what consumers are searching in Google can also predict their collective future behaviour days or even weeks in advance. In the section devoted to the movie performance predictions the authors analyze a database of the opening weekend box-office revenues for 119 films released in the United States between October 2008 and September 2009 and find that the predicted result suggested by the simplest linear model concerning only search volume index matched real data well even if this prediction was based on the search index that was taken four to six weeks before the movie release. Furthermore, they come to the conclusion that search volume data improve the predictive power of these models when included as a parameter into other models. The authors also point out that in absence of other data sources, or where small improvements in the predictive performance are material, search queries might provide a useful guide to the nearest future.

Piracy and box-office revenue:

In this section I would like to start with brief introduction of a paper written by Rafael Rob and Joel Waldfogel "Piracy on the silver screen" (2007). In the article the authors analyze the data, obtained from the series of surveys administered to over the 500 undergraduates in the year 2005. Undergraduates had to choose whether they saw film or not and had to choose where they saw it from the given list of possible answers. There were 4 paid methods (the cinema, theater, television, rental and purchase) and 2 unpaid ones (watching a downloaded copy and watching a burned copy of a legally obtained copy of the film) only one of which might be chosen for a given movie. Also it was necessary to fill in an integer number: 1 or 2 etc. if they had seen the movie for the first time, for the second time and so on (choosing the way they saw the movie from the list). Moreover, the researchers asked their respondents about their family income, race and age, speed of the internet connection they have but in addition they took into consideration variables, characterizing the respondents' interest toward the movie industry: how often they go to the cinemas, how many movies are in their collection and their level of interest in watching movies. Having analyzed this database, the authors conclude that unpaid consumption constitutes a small share of the movie industry products consumption even in the

sample of technically sophisticated college students with good Internet connection. Despite the results, the authors suggest that it partially owed to the fact that file sharing systems were clumsy at that time, and the searching costs of obtaining an illegal copy may play a significant role, but if the available means for copying movies become easy to use, file sharing could evolve into a very serious threat to the film industry. Comparing movie industry with music industry, where piracy was all too common, the authors underline that even if the searching costs and any other costs of obtaining a film illegal copy fall, the two industries would still remain different because of the consumption costs: watching a movie requires a few hours of attention, which is costly. Thereby, the movie is costly to consume even if it is obtained for free. This argument is very sensible even nowadays when searching costs seem to be relatively low, but people prefer going to cinemas and pay rather than spend their time to watch a movie for free but in a worse quality. The costs of consumption in particular nowadays seem to be the key factor that mitigates the low quality piracy effect on the box office revenue because of the idea that people do not want to spend their time watching movies in a bad quality without much pleasure. Consequently, these costs force them to go to the cinema or, at least, wait for the time when high-quality illegal content will be released.

Another remarkable article, written by Ma et al (2011), which, in some sense, played the role of a starting point for our research, is titled "Effect of pre-release piracy on box-office revenue". The authors analyze effects of pre-released piracy (when the stolen copy of the movie becomes available before the official release) on the box office revenue. We should emphasize the fact that pre-released piracy differs from other types of piracy in terms of the clientele it attracts. It is a popular argument that if consumers were really interested in the content, they would buy the legitimate version, which usually has higher quality. Whereas those who are satisfied with the low quality pirated version have low willingness-to-pay for the content and would not have bought the legal version anyway. However in the case of pre-released piracy people who downloaded the illegal copy might download and watch even the copy with poor quality just because they wanted to see the film as soon as possible. On the other hand there exists a hypothesis that the pre-released piracy may also increase the box-office revenue because of the possible word-of-mouth information spread generation, which can substitute an expensive advertisement.

The data on piracy were collected from the VCDQuality.com source that played the role of the piracy release log (no direct links could be founded there). The main database contained 553 movies, 117 of which were with missing values for their budget. If the budget of the movie was unknown, the authors set its value to the average budget of all the films under the investigation and create a dummy variable for the missing budget that captures any systematic

differences between the group of the films with known budgets and unknown budgets. To avoid inadequacy of the information, the movies that were displayed in the cinema for less than six weeks were removed (58 items). Conclusions were made from the final dataset comprised of 475 movies, where 48 of them had pre-released piracy.

The structure of the dataset used and the source of the information are presented in the next table:

Variable	Description	Source
IMDBID	The unique ID assigned to the movie at IMDB.com. This is used to identify the movie.	IMDB.com
BO	The box office sales of a movie in a week.	BoxOfficeMojo.com
BUDGET	The estimated production budget of the movie. This information is not available for all movies.	IMDB.com, BoxOfficeMojo.com
VDATE	The earliest date on which a copy of the movie became available on the Internet according to vcdquality.com.	Vcdquality.com
MDATE	The official wide release date of the movie.	IMDB.com, BoxOfficeMojo.com
PIRACY	An indicator variable of pre-release piracy, with 1 representing the existence of pre-release piracy. This is inferred from VDATE and MDATE: pre-release piracy exists for a movie if VDATE is earlier than MDATE	Inferred
AUDIOQUAL	The average audio quality rating of the pirated copy according to vcdquality.com. Not all copies received a rating.	Vcdquality.com
VIDEOQUAL	The average video quality rating of the pirated copy according to vcdquality.com. Not all copies received a rating.	Vcdquality.com
DIST	The distributor of the movie.	BoxOfficeMojo.com
GENRE	The genre of the movie.	BoxOfficeMojo.com
DIRAPPEAL	A binary indicator of the presence of a star director in the movie. The indicator is set to 1 if the past average box office sales of the director is higher than $50 million. The average box office sales of the movies directed by the director of the movie over the past years were collected from BoxOfficeMojo.com. This information is not available for all directors.	BoxOfficeMojo.com Inferred
STAR	A binary indicator of the presence of stars in the cast of the movie. A movie is considered as having a star if some of the top four actor/actress of the movie have either been nominated for or won an academy award before the playing in the movie.	IMDB.com Inferred
SCREEN	The number of screens on which the movie was shown in the opening weekend.	BoxOfficeMojo.com
USERRATING	The average movie rating posted by viewers.	IMDB.com
CRITICRATING	The average critic rating of the movie.	Yahoo Movies

Table 1

The authors use a reduced-form exponential model in the form: $y_{it} = m_i e^{-n_i t + \varepsilon_{it}}$ where y_{it} is the box office revenue of movie i at time t, and m_i and n_i represent the market potential and the rate of decay of the movie, respectively, which depend on the movie static parameters

such as celebrities participations, IMDb rating etc. and the dummy variable for the pre-released illegal copy availability for movie i.

The main results of the article are the following: pre-release piracy decreases the market potential but also increases the rate of decay (presumably due to the word-of-mouth mechanisms), but as a result the piracy causes approximately a 15% reduction in the box-office sales. Nevertheless, the authors show that the pirated copy with higher audio and video quality has less severe impact on the movie box office revenues than the lower quality releases do. It may mean that if a film in a bad quality is stolen it is better for the producers to release on the internet a copy of this movie in better quality to reduce the hazard effect on the box office revenue.

The third paper, we'd like to refer in the literature revue is the article, written by Christian Peukert et al. "Piracy and Movie Revenues: Evidence from Megaupload: A Tale of the Long Tail?" In this research paper the authors aim to estimate the effect of the exogenous piracy change (Megaupload shutdown) on the box office revenue of the movie. Megaupload Ltd, a file hosting service, included in the top-15 list of the hosting services and once the 13th most visited site on the internet with more that 180,000,000 registered members and 50,000,000 visitors per day[2] was established in the year 2005 and became one of the main and the most significant illegal content sources in the world, but it was shut down by the United States Department of Justice on January the 19th, 2012. This shutdown created a quasi-experiment in the market of the illegal downloading and allowed the authors to exploit an exogenous piracy shock to perform the investigation. The authors investigate 1,344 movies in 49 countries and conclude that a positive effect after the shutdown was found only for the blockbusters (the movies shown on more than 500 screens). The rest of the database shows insignificant or even negative effect of the Megaupload closing. The scientists argued that this might be due to the social network effect because of the fact that the information about the movie could spread from consumers with low willingness to pay (who download the pirated copy) to consumers with high willingness to pay (who go to cinemas). So for this class of movies publishers may find it preferred to release piracy content by themselves instead of paying for the expensive advertisement.

Other models for box office revenue prediction:

The set of works I would like to refer to in this section might be roughly divided into two big groups. The first group of papers is devoted to prediction of the movie box office revenue using social media data such as twitter posts, IMDb comments etc. Asur and Huberman(2010)

[2] http://www.washingtonpost.com/wp-srv/business/documents/megaupload_indictment.pdf - Megaupload indictment written by the United States District court for the Eastern District of Virginia.

use the rate of chatter from almost 3 million tweets found on the popular site Twitter and construct a linear regression model for the box-office revenues of movies forecast. They show that the results outperformed the models based on the Hollywood Stock Exchange data[3] in accuracy. The authors also analyze the sentiments that were presented in the tweets and demonstrate their efficiency to improve the predictions after a movie was released. It is also demonstrated that analysis covering social media, such as twitter, may be extended to the large amount of the topics (for example, future ratings of products), but it may also be exploited for the prediction of election outcomes.

In a more recent work, Lica and Tuta (2011) analyze modern techniques of product success predictions and point out the main problems (which make the social media analysis very difficult) of the social media sources of information. The main issues presented in the article include

- Language problem
- Spam participation problem
- Difficulties in recognition posts' sentiment problem.

The second fundamental group of the articles, which might be considered as a classical models for movie's box-office revenue published in the year 1996 by Sawhney and Eliashberg and its extensions by Dellarocas et al (2007), were devoted to the attempts to construct a model which takes into account the time of making a decision to watch a movie and the time to act after this decision is made. The sum of these times was defined as the time to adopt the movie (decide to see and go to the cinema after this decision). Dellacotas proposes a new form for the hazard rate function, which plays the role of the engine in the models of this type, which take into consideration "external" force (advertising etc.) and "internal" force related to the word-of-mouth of the past viewers. This function also takes into consideration the fact that the effect of the pre-released advertisement is falling down constantly as well as the impression about the movie (which is closely linked to the word-of-mouth effect). After the main parameters estimation, the authors conclude that the prediction made with use of their model performed well enough in comparison with the out of sample movie data. Thus, the authors conclude that the model might be a good instrument for making predictions about the movie box office revenue performance.

[3] HSX.com Hollywood Stock Exchange is an artificial stock exchange where the role of stocks is playing by the movies and is considered to be a good prediction instrument. In 2007, players in the Hollywood Stock Exchange correctly predicted 32 of the 39 major-category Oscar nominees and seven out of eight top-category winners.

Database

Movie list:

Initial movie list, which were in the USA cinema box office for the last ten years, was collected from the BoxOfficeMojo.com weekend charts[4] taking into consideration all the weekends starting from the 1st of 2004 till the 10th of 2014. For the database collecting procedure a program on .net c# was written (powered by an html-parsing procedure). As a result of the data collecting, the list of the 6306 movies was initially obtained for the research (not all of them indeed had enough information for the investigations we planned to perform).

Box-office revenue data:

Box office revenue information on the daily basis for the movies from the list we described above was also taken from the BoxOfficeMojo.com web site. To be more precise, this resource provided us with the following dynamic data:

- Daily gross and gross-to-date revenues
- Daily quantity of the cinemas, average box office revenue per cinema
- Exact date
- Movie rank in the table of the top box office revenue for the given date

For instance, "The Matrix Revolution" daily box office information was presented in the following form:

Day	Date	Rank	Gross	% +/- YD / LW*		Theaters / Avg		Gross-to-Date	Day #
Wed	**Nov. 5, 2003**	1	$24,311,365	-	-	3,502	$6,942	$24,311,365	1
Thu	**Nov. 6, 2003**	1	$11,003,481	-54.7%	-	3,502	$3,142	$35,314,846	2
Fri	**Nov. 7, 2003**	1	$16,529,521	+50.2%	-	3,502	$4,720	$51,844,367	3
Sat	**Nov. 8, 2003**	1	$19,524,728	+18.1%	-	3,502	$5,575	$71,369,095	4
Sun	**Nov. 9, 2003**	1	$12,420,905	-36.4%	-	3,502	$3,547	$83,790,000	5
Mon	**Nov. 10, 2003**	1	$4,587,414	-63.1%	-	3,502	$1,310	$88,377,414	6
Tue	**Nov. 11, 2003**	2	$5,200,298	+13.4%	-	3,502	$1,485	$93,577,712	7
Wed	**Nov. 12, 2003**	1	$2,223,433	-57.2%	-90.9%	3,502	$635	$95,801,145	8
Thu	**Nov. 13, 2003**	1	$2,052,420	-7.7%	-81.3%	3,502	$586	$97,853,565	9

Picture 2

A part of static information about the distributor, production budget (unavailable for some films), movie runtime etc. was also exposed on the web page for a given movie on BoxOfficeMojo.com site.

It is necessary to mention that the daily box office revenue information was given not for all the movies from our initial dataset (only 2443 movies, which we could find in the IMDB, had at least one record and 1933 movies were mentioned in the IMDb[5] and had at least 30 records).

[4] http://boxofficemojo.com/weekend/chart/
[5] http://imdb.com, will be discussed below

All the actions on collection and processing the dataset were done in the automatic mode by the same data-analyzing program, written by the authors.

Movie static information:

The next sources of the information, which we used constructing the dataset for the analysis, was the Internet Movie Database (IMDb.com).

From this web site we got the next static data (which is constant over box office time) for the movie under consideration:

- The movie crew (director, scriptwriter, up to four main actors).

  Here we decided to identify whether an actor is a star or not using a simple list of the best 500 actors, found at the listal.com web site. We decided to apply this method for the actors star identification because there are a lot of famous ones who have almost never got any positive awards, but still his or her participation may attract people to watch the movie: for instance Steven Seagal got only Golden Raspberry award for the worst role, but movies with his participation will definitely attract attention of the audience. Speaking about the star identification among the movie directors and scriptwriters, we marked them as celebrities if they had more than one Oscar nomination as a best director or scriptwriter, respectively.

- The IMDb rating of a movie (this is the rating calculated by the special formula from the marks out of 10 points, which any user may assign on the site to the particular movie).

  It should be pointer out that all the history of the raiting changes is inaccessible for us. But we make a reliable assumption that after the box office period this mark reaches a value which deviates form its asymptotic value insignificantly.

- Date of the movie release
- Budget information (may be omitted)
- MPAA rating (age restrictive rating)
- List of genres, related to the movie (comedy, romance etc)
- Runtime
- Movie distributor in the USA (Dreamworks, Warner Bros, Fox, MGM etc.)
- Opening cinema quantity
- Maximum amount of the cinemas during the box office period
- Brief film description (using it we can introduce special dummies for the particular word participation)

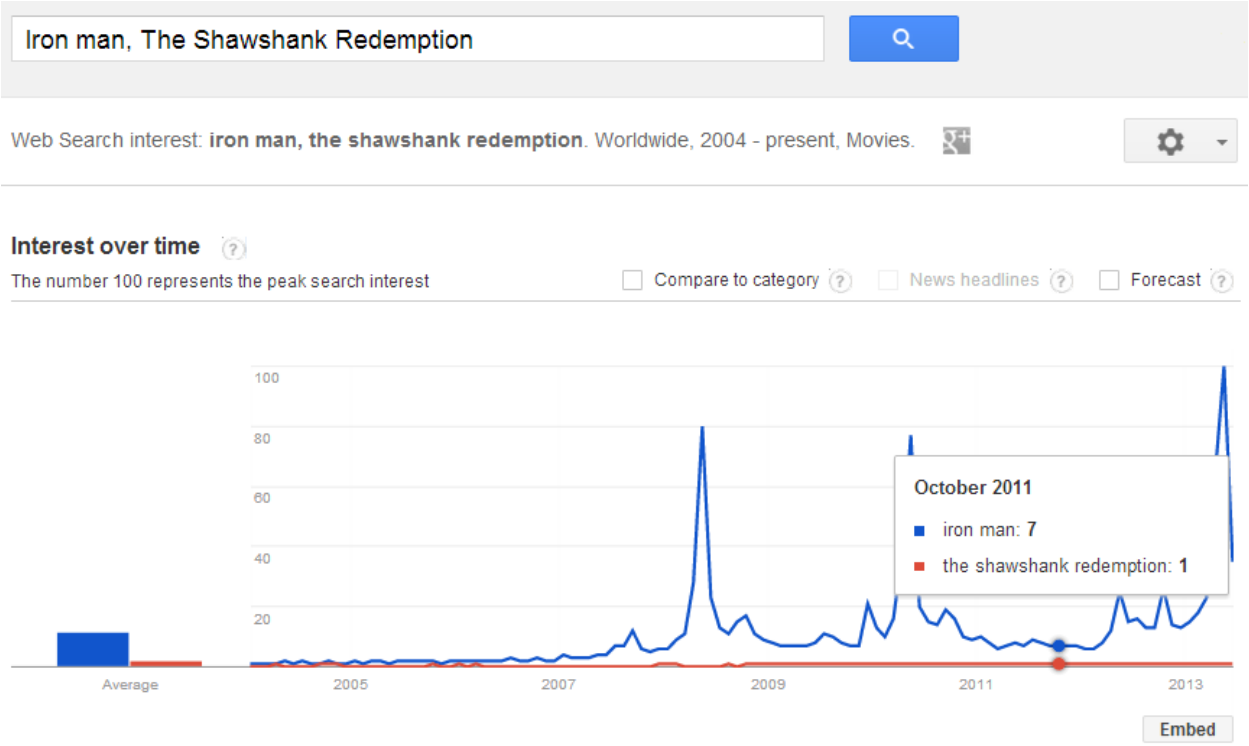
- Country-producer
- Native language of the movie
- Awards etc.

It is important to note again that all the information obtained from the IMDb was static (stay constant during the movie box office period) and not all the movies from the initial list were found in the IMDb. To be more precise, from the initial list of the 6306 movies only 4642 were found in the IMDb. Only those movies were taken for the further analysis.

Integration between the program and the IMDb web site was reached using Open Movie Database API (omdbapi.com).

Google Trends

Google Trends is a public web facility based on the Google Search engine, by the aid of which a search volume index for a given keyword could be obtained. This index is given in the form normalized to the maximum level for the given search query and overall Internet activity. The information may be obtained for a particular region of the world, time, category and content



Picture 3

(Web, News, Images etc.). News headlines, which are in some way relevant for the target search query, may be displayed on this graph. The service provides an opportunity to make the simplest prediction about the given keyword trend behavior in the nearest future. Search volume index for any keywords is available from the 1st of January of 2004 till now. The data used for plotting the graph may be downloaded as a CSV file. The main disadvantage of the Google Trends'

search volume index seems to be normalization of the data and impossibility to get an absolute value of the given search request. This might be the reason for using Yandex Wordstat index[6], which provides an absolute value for the requests, but this normalization may be also a virtue because all the data are preceded by Google algorithms (which allows to cut proxy multiple requests etc) and normalized to the overall internet activity. Moreover, it is possible to see the data in comparison with chosen search category, which provides an opportunity to see the indices for two or more independent keywords on the one graph. It allows obtaining the absolute value approximation using a benchmark keyword method.

Previously mentioned Yandex Wordstat service is a service which may supply an absolute value of the search requests for a given word based on the Yandex Search data, but these data are not suitable for our research for the following two reasons:

- Available index is limited by the last two years.
- Search data are not consistent with the USA search queries since the USA Internet users almost never use Yandex search service.

Concerning the data, as long as our investigation is devoted to the US box office revenue analysis, only Google SVI was used. Search volume index was limited on the "Movies" as a category and the US as a region. Obtained were daily data starting from the month before the movie release till the end of the box office period.

Piracy Database:

One could say that a new era of piracy began when the BitTorrent protocol was created. This protocol allowed people to carry out file exchange in peer-to-peer (P2P) mode (downloading content directly from the computers of other users, connected to the P2P network) without uploading a file to special hosting servers. It means that if one of the P2P network users reaches some content (possibly illegal) and wants to share this content using P2P network (those people are called seeds), the link to this content can be sent to other users so that they can download the data directly from the seed's computer. But at the same time users are becoming seeds of the parts they have already downloaded. Due to this protocol the information and its source started to spread through the web very fast, so that fighting with an illegal content sharing distributed by this protocol became an extremely difficult task.

Soon after the protocol was invented new torrent tracker web sites started to emerge all over the globe. These sites were just a collection of links to another computer's content, so that a huge database of content stored this way did not demand a lot of space for its storage on the host server.

[6] http://wordstata.yandex.ru

One of the first torrent trackers was The Pirate Bay. This site was founded in 2003 in Sweden and shortly after the start of its work became a very popular torrent exchange area for people from all over the world. In some sense, The Pirate Bay became one of the hearths of piracy, since the searching costs of gaining illegal content uploaded there were drastically lower compared to other sites.

Concerning our investigations, for each film from the movie database we obtained the full list of the torrents available on The Pirate Bay (given the day of release, name, size and category). Taking into consideration common trackers' rules of the torrent distribution formalization, type and quality of a torrent may be almost surely obtained from the torrent's name or its size (for instance, the DVD repacked files due to its commonly used codec has the size equals approximately 1.4 gigabytes). The full table of the world-standard piracy movie markers, their meanings and related quality, which was used for the quality identification, we obtained from the rutracker.org site[7]. The main task for our data gathering procedure in this section was to find for a given movie on The Pirate Bay web site torrents' dates of release for the illegal content in low quality format (usually in the CamRip or TeleSync format) and in high quality (frequently, this torrent files have DVDRip quality markers).

Furthermore, availability of illegal information on the Internet in general was checked with the VCDQuality.com service, which Ma et al (2011) used as a source of piracy information in their investigation. This service provides logs of almost all the illegal movie content with exact release date and quality of that content (without any direct links to the file). This information shows that a pirated copy has already been issued and now exists somewhere on the web. This information could be extremely useful for our research because for a given movie name and a given quality of the picture it shows the release date when the illegal content first emerged on the internet. It is noteworthy that the information about the illegal file quality may be obtained directly without overly complicated movie name parsing analysis as long as it is present in one of the columns of the table.

These two sources of piracy information were used simultaneously to estimate the impact of searching costs for the illegal piracy content on the box office revenue and to test the hypothesis that even if some illegal content emerged at some site on the internet and was registered by the VCDQ.com site (which, as it was mentioned above, does not provide the direct link to the content), but was not available for download on The Pirate Bay site (or any other popular piracy source) for some time, searching costs of obtaining this content for an unsophisticated internet user at that time were assumed to be substantially high. This means that

[7] http://rutracker.org/forum/viewtopic.php?p=27840514#27840514

the illegal file will not be in substantial demand (that fact was taken into consideration as a hypothesis, which will be tested in the numerical analysis part of the investigation).

Database summary:

Summing up, our daily revenue database was restricted by the 1264 most suitable for the analysis movies that were in the movie exhibition from the year 2004 till the year 2014 and the box-office revenue for them were known for twenty or more days. We restricted the initial dataset because in our research we were more interested in analysis of the movies that were initially created not just for the festivals and other competitions, but to earn the money as commercial projects.

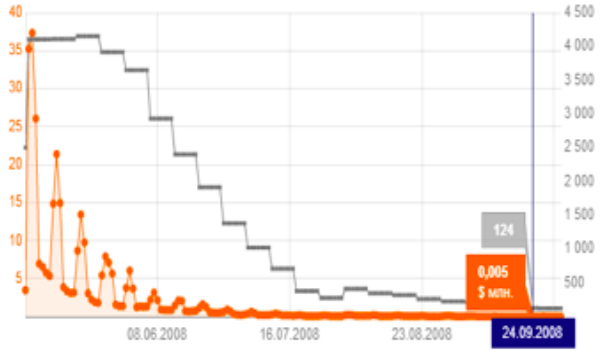
The dynamic part of the dataset we collected includes the following items:

- Daily revenue and cumulative revenue for the given date (sometimes with missed values) in terms of the US dollars
- Information about the date (whether it was weekend, holiday etc.)
- Availability of the piracy content for the given date and, if available, its quality. As it was mentioned before, the information was obtained from two sources: The Pirate Bay (with possibility to download the movie: low searching cost piracy) and VCDQ.com (without direct link for the illegal content: initial illegal copy release)
- Google search volume index for the given movie name and given date or week if interest during the movie exhibition period

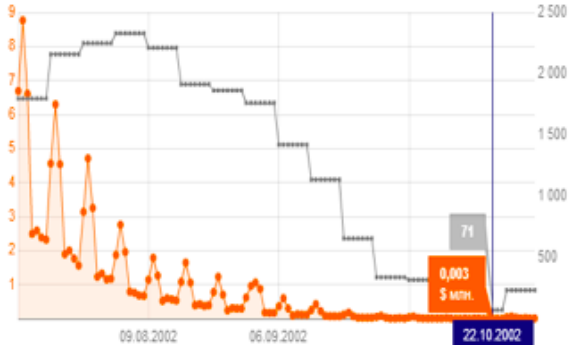
Dynamic model

As we have mentioned in the literature overview section, when we discussed the article of Ma et al (2011), for the dynamic weekly investigation the authors used an empirical exponential model in the form $y_{it} = m_i e^{-n_i t}$ (Eq. 1) where y_{it} was the box office revenue of movie i at the t weeks from the release, and m_i and n_i represented the market potential (which in some sense equivalent to the first week revenue) and the rate of decay (exponential index) of the movie.

Daily box office revenue profile for the ordinary movie
(orange line – daily revenue, gray line – cinema quantity)



Iron Man, 2008



Road to Perdition, 2002

Table 2

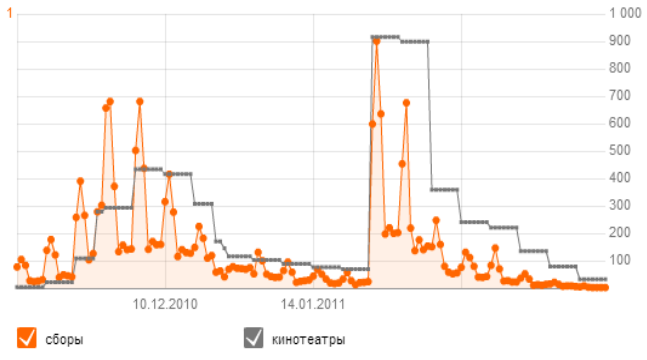
As we can see on the graphs, on which the daily box office revenue curves were ploted for the case when the cinema profile over the box office period has an ordinary path, the most common for commercial films (starting with more than 80% of the maximum cinema quantity, then growing slightly and after some time start to decrease sharply) the assumption of an exponential form of the box office revenue decay over time seems to be reasonable and trustworthy. The exponential pattern might be seen considering separately the weekday revenues and each particular weekend.

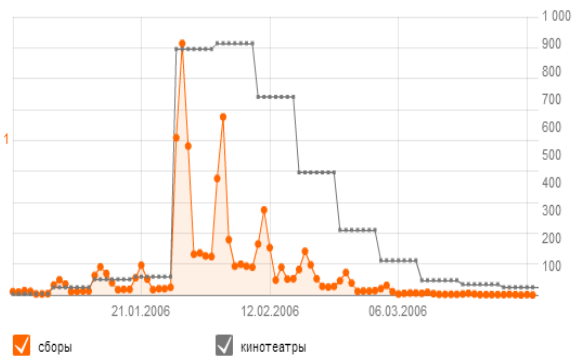
According to this fact, it was assumed that the exponential model in the form (Eq. 1), but with a special multiplicative functional, might be used to capture the daily dynamics in this case too. It is obvious that this multiplicative functional should depend on the indicators for the weekends: Friday, Saturday and Sunday (because as we can see from the graphs, if we select the group of points just for one of the given weekends, we would still observe exponential decay).

Moving further, the exponential pattern of decay might be difficult to see on the box office revenue graph of so-called “slipper” movies (the movies, cinema profile of which starts

stays some time at the low quantity level, but then the number of the cinemas increases significantly: twice or more).

Daily box-office revenue profile for the "slipper" movie





127 Hours, 2010 **The Matador, 2005**

Table 3

These examples explicitly show us that the quantity of the cinemas should be included into the exponential functional factor too. Moreover, the variables, which depend on the illegal content availability, on the people's attention to the given movie and its marketing policy (advertisement etc), should be also included into the considered functional factor as well.

Concluding, the resulting empirical model, which we used in our investigation, was taken in the following form: $y_{it} = f(c_{it}, d_{it}, p_{it}, s_{it}) m_i e^{-n_i t}$ (Eq. 2), where y_{it} was the box office revenue of the movie i, t days after the start of the box office period. Further, m_i and n_i represented the market potential and the rate of decay of the movie and depended on the static movie characteristics, which were presented in the previous section. Arguments $c_{it}, d_{it}, s_{it}, p_{it}$ represented some functions of all the available history up to time t of the next dynamic movie parameters: cinemas, exact day or week, search volume index and available pirated content (from The Pirate Bay and VCDQ) respectively for the given film i. For instance, if the search volume index at time t is high it may mean that that the people will go to the cinema at the time t+n, at which search volume index may be not so high (because, using the terminology introduced by Sawhney et al. (1996) and discussed earlier, they need time some to act). This example suggests that the lagged values of the search volume index should be also included in the SVI function.

In the analysis we performed, the function $f(c_{it}, d_{it}, p_{it}, s_{it})$ was taken in the form, such that $\ln f(c_{it}, d_{it}, p_{it}, s_{it}) =$ sum of the next summands (multiplied by their effects, which we wanted to estimate):

1) Cinema quantity part:

$\ln(cinema\_quantity_{it})$

- logarithm of the movie quantity for the given day and the given movie

2) Day of the week part:

$is\_friday\_dummy_{it}$

-dummy variable, which showed whether at the day t for the movie i was Friday

$is\_saturday\_dummy_{it}$

-dummy variable, which showed whether at the day t for the movie i was Saturday

$is\_sunday\_dummy_{it}$

-dummy variable, which shows whether at the day t for the movie i was Sunday

$is\_holiday\_dummy_{it}$

-dummy variable, which showed whether at the day t for the movie i was a public holiday in the US

$is\_holiday\_nearby\_dummy_{it}$

-dummy variable, which showed whether in the deleted neighbourhood (with radius equals 3 days) of the day t for the movie i there were any public holidays in the USA

3) Search volume index part:

$\ln(SVI\_index\_lag1day_{it})$

-logarithm of the lagged by 1 day search volume index for the movie i at time t

$\ln(SVI\_index\_lag2days_{it})$

-logarithm of the lagged by 2 days search volume index for the movie i at time t

$\ln(SVI\_index\_lag7days_{it})$

-logarithm of the lagged by 7 days search volume index for the movie i at time t

$\ln(SVI\_index\_lag14days_{it})$

-logarithm of the lagged by 14 days search volume index for the movie i at time t

Here we should mention that the SVI of the day t without lag was excluded because it was difficult to control causality: it is not obvious whether the viewer used Google search before going to the cinema (for instance, checked the ticket price, tried to find the nearest cinema etc.) or after it (for instance, because he or she was under the strong impression from the movie and tried to find some more information related to the film). Also, the multiplicative form of the model allowed us to use search volume index in the relative form (because normalizing constant was fixed for the given movie and might be included unified with the market potential (fixed effect) individual film constant).

4) Piracy part:

Before the numerical analysis start we decided to divide the collected movie database into the two big parts: bad movies (defined as movies, which IMDb rating is lower than 6.00) and good movies (IMDb rating is higher that 6.00, respectively). For each group we tried to measure the effect of piracy in the following form:

$is\_low\_quality\_piracy\_available\_lag1day_{it}$

- dummy variable, which showed that for the movie i at time t-1 there existed an illegal content of the low quality (CAMRip , TeleSync and similar) somewhere on the internet (according to the VCDQ site), but a high quality copy for the given movie at time t was not yet available (on that site)

$is\_high\_quality\_piracy\_available\_lag1day_{it}$

- dummy variable, which showed that for the movie i at time t-1 on the internet there existed a high quality piracy content

$is\_low\_quality\_piracy\_with\_low\_searching\_costs\_available\_lag1day_{it}$

- dummy variable, which showed that for the movie i at time t-1 at The Pirate Bay torrent tracker there existed a piracy content of the low quality (CAMRip or TeleSync and related), but there were no high quality copy of the film up to time t-1 on this tracker (again, here we assumed that when the piracy content emerged at The Pirate Bay site, its searching costs became low)

$is\_high\_quality\_piracy\_with\_low\_searching\_costs\_available\_lag1day_{it}$

- dummy variable, which showed that for the movie i at time t-1 there existed a torrent of the high quality, listed on The Pirate Bay torrent tracker.

All the dummies were taken lagged by one day because, in our opinion, choosing between the real time piracy values or 1 day lagged piracy values it is more reasonable to take under consideration the lagged piracy variables due to many facts. At first, we can't control explicitly what time of the day the piracy content emerged, but according to the statistical data and common sense, the content (especially torrents) usually emerges and start spreading rapidly at the end of the day because of the fact, that at the evening more people are present on the internet. Secondly, we also think that the piracy may affect the viewers' decision to go to the cinema mostly if an illegal content was released at least one day before the arranged day, because of an inertia in a decision making process and a process of booking the cinema tickets in the US.

Summing up, we decided that the most suitable reduced-from model for the dynamic analysis we wanted to perform should be taken as follows:

$$\log(spot\_daily\_revenue_{it}) = \log(market\_potential_i) - t * rate\_of\_decay_i + bX_{it} \text{ (Eq. 3),}$$

where X is the matrix of movie dynamic parameters, described above, b - is the vector of their effects, which should be estimated.

Dynamic modeling results

Before running the main regression, we tried to check the search volume index and piracy arguments, included into the model we used, on the presence of multicollinearity (first “g” and “b” here represents whether the movie is good or bad, “l” and “h” – low or high quality of the piracy content, “wsc” – without searching costs):

	bl	blwsc	bh	bhwsc	gl	glwsc	gh	ghwsc
bl	1.0000							
blwsc	0.5991	1.0000						
bh	-0.5513	-0.3063	1.0000					
bhwsc	-0.3271	-0.5574	0.6081	1.0000				
gl	-0.0002	-0.0002	-0.0005	-0.0005	1.0000			
glws	-0.0002	-0.0002	-0.0005	-0.0005	0.6330	1.0000		
gh	-0.0012	-0.0011	-0.0028	-0.0029	-0.5697	-0.3388	1.0000	
ghwsc	-0.0013	-0.0011	-0.0028	-0.0029	-0.3786	-0.5715	0.6500	1.0000
SVIl1	0.0385	0.0184	0.0172	0.0206	0.0425	0.0300	0.0318	0.0382
SVIl2	0.0463	0.0220	0.0013	0.0084	0.0480	0.0346	0.0087	0.0140
SVIl7	0.0053	0.0058	-0.0092	-0.0082	-0.0047	-0.0050	0.0026	0.0097
SVIl14	-0.0182	-0.0128	-0.0100	-0.0062	-0.0226	-0.0261	-0.0050	0.0068

Table 4

Analysing the table of correlations, we could not see any severe multicollinearity (the biggest one equals 65%, which is not so critical in our case) so that the analysis might be performed using the reduced-form model, defined above without being afraid of the strong multicollinearity.

Speaking about the regression analysis results, the main results are presented below:

```
      Source |       SS       df       MS              Number of obs =   81787
-------------+------------------------------           F(2279, 79508) =23821.02
       Model |  10557922.8  2279  4632.69977           Prob > F      =  0.0000
    Residual |  15462.6774 79508  .194479517           R-squared     =  0.9985
-------------+------------------------------           Adj R-squared =  0.9985
       Total |  10573385.5 81787  129.279537           Root MSE      =    .441

------------------------------------------------------------------------------
      lnspot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lncinemas |   .7767477   .0023521   330.24   0.000     .7721376    .7813578
     holiday |   .5701344   .0093781    60.79   0.000     .5517533    .5885155
holidaynea~y |   .1270784   .0044183    28.76   0.000     .1184186    .1357383
      friday |   .8600235   .0049027   175.42   0.000     .8504141    .8696328
    saturday |    1.25125    .005134   243.72   0.000     1.241187    1.261313
      sunday |   .7929291   .0051234   154.77   0.000     .7828873    .8029708
badfilmlqc.. |  -.3288742   .0262533   -12.53   0.000    -.3803305   -.2774179
badfilmlqc.. |  -.2018974   .0227865    -8.86   0.000    -.2465588   -.1572361
badfilmhqc.. |  -.5187364   .0331731   -15.64   0.000    -.5837556   -.4537173
badfilmhqc.. |  -.1758066   .0297999    -5.90   0.000    -.2342142   -.1173989
goodfilmlq.. |  -.2616375   .0168201   -15.56   0.000    -.2946048   -.2286703
goodfilmlq.. |  -.1215531    .015412    -7.89   0.000    -.1517604   -.0913457
goodfilmhq.. |  -.3306099   .0186609   -17.72   0.000    -.3671851   -.2940348
goodfilmhq.. |   -.184188   .0175461   -10.50   0.000    -.2185782   -.1497977
 lnsviindex1 |   .1726617   .0048836    35.36   0.000     .1630898    .1822335
 lnsviindex2 |   .0483559   .0048889     9.89   0.000     .0387736    .0579382
 lnsviindex7 |   .1140559   .0034235    33.32   0.000      .107346    .1207659
lnsviindex14 |  -.0159856   .0030317    -5.27   0.000    -.0219277   -.0100434
--more—
```

(all fixed effects and individual time trends were from the table, but they were included into the initial regression)

Table 5

From this table we can conclude the following:

1) All the individual variables are significant given any sensible probability level.
2) As we can see, revenue-cinema quantity rule is precisely nonlinear (cinema quantity effect is equal to 0.77 and significantly differs from), which implies decreasing marginal returns

```
test lncinemas = 1:

F( 1, 79337) = 9474.40 Prob > F = 0.0000
```

This means that regressions in the form ln (revenue per cinema) on all the remaining variables might be not robust and in some sense meaningless.

3) Holiday increases the daily revenue by approximately 57%. If there is a holiday in the 3 day deleted neighbourhood, it increases the revenue by more than 12%.

Film distributors may use this fact, for instance, to catch more holidays during the box office period and, respectively, increase the total revenue of the movie.

4) Weekends' levels are distinct (+86%, +125%, +79%). Moreover, Friday's and Sunday's multipliers were proved to be statistically different

```
test Friday = Sunday:

F( 1, 79337) = 85.73, Prob > F = 0.0000
```

5) The highest SVI effect was detected at the nearest (one day) lag. Moreover, the search volume index, corresponding to the one-week lag was significantly higher than zero and sufficiently large. According to the results we got, the one-week lag SVI effect appeared to be even higher that the 2 days lag effect. It may be explained by the nonzero time to act, introduced by Sawhney et al. (1996) in their paper: people after making their decisions should have some time to act, and this time usually significantly differs from zero. Our results, in some certainty, shows that the expectation of this time is in the range between one and two weeks (the 2 week SVI lag correlates negatively with the day t spot revenue, which may suggest that the effect of the intensive advertisement will go away after the 2 weeks, because the most of the advertised target audience either go to the cinema during this period or lose their attention to the movie after that time)

To conclude, we can definitely say that the viewers usually Google before going to the

cinema and that the expensive advertisement, concentrated in one single point might be not so fruitful, because the attention to the film will go away very fast.

6) Piracy hits good and bad movies in a different way. Piracy effect is more severe for the bad movies (we discuss the particular numbers later).

```
test (badfilmlqpiracy= goodfilmlqpiracy)
(badfilmlqcostlesspiracy=goodfilmlqcostlesspiracy) (badfilmhqpiracy=goodfilmhqpiracy)
(badfilmhqcostlesspiracy= goodfilmhqcostlesspiracy):

F( 4, 79337) = 17.81, Prob > F = 0.0000
```

7) Piracy content with high quality (HQ) hits revenue much more dramatically (in comparison to the low quality (LQ) piracy content) and that difference was proved to be significant:

-33%(LQ) and -52%(HQ) for bad movies

-26%(LQ) and -33%(HQ) for good movies

```
test BadFilmLQPiracy= BadFilmHQCostlyPiracy:

F (  1, 79508) =   68.98,   Prob > F =    0.0000

test GoodFilmLQPiracy = GoodFilmHQCostlyPiracy:

F(  1, 79508) =   23.36,   Prob > F =    0.0000
```

8) The searching costs matter for both bad and good movies and both low and high quality. If the searching costs are high (no torrent on The Pirate Bay site is available), it may prevent from losing additional 12-20% revenue. To be more precise the distributors lose extra:

20%(LQ) and 18%(HQ) for bad movies

12%(LQ) and 18%(HQ) for good movies

```
test (BadFilmLQCostlessPiracy=0) (BadFilmHQCostlessPiracy =0)
(GoodFilmLQCostlessPiracy =0) (GoodilmHQCostlessPiracy =0)

F(  4, 79508) =   47.55,  Prob > F =    0.0000
```

Speaking about the worst scenario when pirated content could be found without spending much time on searching even for the unsophisticated internet users, the analysis we performed suggests that the distributors lose the substantial box office revenue part because of the piracy. In

order to be more precise:

-53%(LQ) and -69%(HQ) for bad movies

-38%(LQ) and -51%(HQ) for good movies

Concluding, we want to say, that according to the numerical results we obtained, piracy is definitely not so good for the box office revenue (even for the case of the good movies with low screen number). But the better the movie of interest is the less severe will the effect of piracy (for the case of a bad movie, for instance, if the high quality piracy copy release on the popular website on the Internet, it would almost kills the movie box office revenue. In this case it may be wise enough to consider such a scenario as closing the box office at all or reduce the amount of the cinemas for the given movie or, if possible, try to fight the piracy at least on the popular sites, where an unsophisticated internet user might easily find the illegal copy without spending too much time on searching).

It will be reasonable to underline that in the dynamic analysis we performed the endogeneity of the regressors may bias the obtained results. At first, we should say that the endogeneity here affects both the box office revenue and the time of the piracy content release in the same direction because the is better the movie the higher is the attention to it, the bigger is its revenues at the first box office days and the higher is the probability that it will be stolen soon. We should also say that in the case of dynamic investigation it is very difficult to avoid endogeneity completely: in our case piracy seems to be endogenous and depends on the attention to the movie, movie quality and many other factors, but the main driving factors seem to be static or explained by the proxy variables, which were included into the model, so that the bias in some degree might be adjusted because of the individual constant effect and the individual time trend participation in the model.

It is also possible to find some way in which the piracy might became exogenous, not endogenous. For us, it seems that the only possible method to make piracy exogenous in the current context is to use quasi-experiments when the piracy change becomes determined from the outside is the case of the illegal content site shut down. The survey of the article written by Peukert et al. (2013), exploited that idea was done already in the literature review section. In this article the authors exploited the judge enforced closing of the very famous source of piracy: Megaupload. In spite of the fact that the analysis shows very interesting results (discussed above), the model considered there seems to be more or less static and limited by the very narrow case of piracy. Speaking about our investigation and its goals, we tried to assess the

effect of the every day piracy on the spot daily revenues in dynamics and to suggest the main ideas and the reference numbers for the movie distributors, which might be exploited minimizing the effect of piracy on the box office revenue.

One more possibility was also to use IV regression to instrument the piracy. But having all the data available it was unsuccessful to find any reasonable combination to instrument the piracy in dynamic in the case of the dataset we collected (all the remaining data seems to be more or less endogenous by itself and could not be used as an instruments).

Static modelling and results

In this section we have an intention to present our results considering static analysis of the key dynamic-driver individual movie constants (market potential and rate of decay), which were estimated during the dynamic research. As long as, for instance, normalization Google SVI coefficient was individual and unknown but included into the constant, even here we should run regressions using with individual intercepts. Nevertheless, we run the regression for the rate of decay estimation in the simple form with one simple global intercept.

As a result, we obtained the following values:

	log(market potential)	Rating	Rating^2	Budget	Budget^2	Runtime	Runtime^2	MPAA G-rating	MPAA PG-rating
log(market potential)	-----	2.042795	-0.144161	0.003345	0. 0000234	0. 002409	-0. 0000692	-0.900285	-1.154318
Rate of decay	.0058439	.0089099	-.001183	-7.08e-06	-2.34e-08	-.0004433	1.80e-06	.0029402	-.0032333

	MPAA R-rating	Crime	Drama	Roman-tic	Comedy	Family	Animation	Thriller	Star writer
log(market potential)	0. 0688604	-0.29806	0.103924	0. 124089	-0.4276201	0. 390177	-0.8969253	-.1657521	0.7448912
Rate of decay	.0037623	.0029767	.0003867	-.002091	-.0038624	-.0041821	-.002367	-.0013386	.0029242

	Star Director	Star Actors	Dream works dist.	Worner Bros. dist.	Fox dist.	MGM dist.	Universal dist.	Buena vista dist.	Open cinemas
log(market potential)	0. 3502806	-0. 37091	0. 6768777	-.127416	0. 0687971	-.2236115	0. 7265591	0. 945179	-.0006834
Rate of decay	-.0017074	-.001645	-.000542	-.004090	-.0053095	-.0021161	.0007409	.0128059	6.39e-07

Table 6

Interpreting the results of the regressions we obtained, we can say that the most optimal IMDb rating (the vertex of the parabola) to increase the film's market potential should be equal to 7,09 (the movie should be good enough but, in some sense, not very "sophisticated"). Furthermore, the results of our analysis shows that the best film runtime in order to decrease the rate of decay should be about 123.13 (2 hours and 3 minutes), which does not contradict the common sense (the movie should not be neither too long or too short). Moreover, it also seems to be logical, that the movie runtime affects more the rate of decay but not the market potential because when the movie had just released, it attracts mostly the audience, which wanted to see this particular movie without taking into consideration whether it is too long or too short. On the contrary, in the middle of the box office period the runtime might play significant role because people when choosing the movie to go usually take into consideration its runtime. Speaking about the movie genre, we obtained that the best ones (to optimize both: the market potential and the rate of decay simultaneously) are romantic and family. We also proved that the star participation effect is the highest when the star director participates the project (which accuses market potential to be higher and rate of decay to be lower than average). Star scriptwriters, as

we can see from the regression results, increase the initial jump, but also increase the box office revenue decay coefficient. The effect of the star actors participation seems to be counterintuitive to some extend, because we found that their participation causes decrease the initial revenue per cinema (taking into consideration that the amount of the cinemas in this case is usually higher), but the rate of decay is lower, so their participation may be useful because the movie will compensate the initial loss over the whole box office time.

Speaking about the opening cinema quantity, as we can mentioned before, the higher it is the lower will be the market potential of the movie and the higher will be its rate of decay.

Among the distributors' performance, only Fox and DreamWorks simultaneously increase the market potential and decrease the rate of decay comparing to the others.

Finally, considering MPAA rating, our investigation shows that the more restrictive this rating is, the better market potential and the rate of decay would see (in our opinion, because the main target audience of the cinemas is the persons, who are young enough, but have already reached the full age).

Conclusion

The main goal of this investigation is to analyze and estimate numerically the effects of piracy, marketing policy, holidays, weekends, cinema quantity profile change etc. in dynamics on a movie's USA box office revenue. We study some interesting cases from the movie box office history and concluded that for being as useful for the movie distributors as possible, the analysis should be performed in dynamics (for instance, the collapse of The Terimnal (2004) movie, which is now included in the list of the best movies at kinopoisk, with very strong crew, significant production budget and comparable to other successful film projects in revenue during the first box office week: the significant impact from piracy content that emerged soon after its release indicates the importance of dynamic analysis). In contrast to the previous works on related subjects discussed in the literature overview section, we aim to unite all the main ideas and merits that were point out there (for instance, Google search volume index should be included into the regressions because it increases the explanatory power of the model, searching costs of piracy matters and should be estimated, model with an exponential decay may be used for dynamic box office revenue investigations, time to act matters etc.) and do a full-strength dynamic investigation on the large database. For this work we treated and collected huge amounts of data. These are the main steps, which we performed during the data gathering process:

- Identification of the movies which were exposed on the US box office for the last 10 years
- Searching and processing static data for each particular movie in the IMDb database
- Star directors, actors and scriptwriters identification
- Dynamic daily revenue data obtaining and handling
- Public holiday identification;
- Obtaining a piracy content information for the given movie using The Pirate Bay website and the VCDQ piracy log list
- Piracy quality identification
- Normalization of the movie name and obtaining the SVI data for the result of the normalization as a keyword;
- Final data preparation for the regression analysis (clearing from a noise etc.)

As a result, we have prepared 81787 daily observations in total for 1264 movies for the numerical analysis; we have built a dynamic reduced-form model for the daily stream of box office revenues and found a number of interesting empirical results.

- The effect of an additional cinema of box office revenues is highly nonlinear and statistically different form zero, which implies decrease-to-scale function between the cinema quantity and the revenue

- Calendar effects have a substantial impact: weekends (Friday Saturday and Sunday) increase the box office revenue on average by 86%, 125% and 79% (where Friday and Saturday increases were proved to be statistically different). Public holiday increases the box office revenue by approximately 57%

- Marketing policy (as measured by prior values of the Google Search Volume Index), which shapes the level of attention to the given movie, seems to have the highest effect on the future box office revenue within the span of one to 2 weeks.

- The impact of illegal content release on the movie box office revenue depends on both the quality of the movie (measured by the IMDB rating) and the quality of the pirated content itself. In particular, ceteris paribus, the availability of the low quality illegal copy of the movie decreases causes the structural jump of the daily box office revenue by 33% for the "bad" movies compared with only 26% for the "good" ones. High quality copy has an effect of 52% and 33% respectively. We also establish that the difference between this numbers inside and between the groups is significant.

- Availability of the illegal content for an unsophisticated user (low searching costs case) has an additional negative effect on the daily revenues and depends on the quality of the movie and the pirated file quality: we suggest that if a movie had already been stolen, the low searching website monitoring may help to save (more precisely, not to lose) 20%(LQ) and 18%(HQ) for bad movies and 12%(LQ) and 18%(HQ) for good movies.

Ideas for further research

In the light of our dynamic box office revenue analysis and the best dynamic movie strategy development, we would like to discuss the main points, which may be used for the further investigations.

First, the existing analysis could be extended to account for the effect of substitution between the movies. It is obvious that for a given movie the box office revenue may be different when this movie is competing with different films-competitors, which are present on the cinema screens at the same time with the movie we tried to make as successful as possible (for instance, very often distributors decided to postpone the movie release because there is a very strong rival on the same or contiguous genre or which may pretend to the same awards that the movie under the management). For instance, a lot of prominent movie projects releases were postponed in the year 2003 because of the The Lord of the Rings: The Return of the King box office start, which eventually got 11 Oscar awards out of 11 nominations. But it is also clear that this effect may sometimes help a movie to gain more money. For instance, this might be the case when the presence of a strong competitor may push the viewers to go to the cinema, but in the cinema it might appear to be no tickets on the film they wanted, which could possibly compel them to watch a different movie (because they already reach the cinema and have an intention to go and watch at least something). Concerning the way this effect may be considered, for us it is an open-ended question which seems to be not so obvious because here we should define precisely what film we note as a rival, how strong this rival is, how severe the effect of substitution between the film considered and its competitor is (especially when the movie released with genre mixture) etc.

Social media variables may be also included into the dynamic investigation to control the word-of-mouth effect, but there are a lot of problems using it, which were pointed out by Lica and Tuta (2011) and discussed in the literature overview section.

Moving on, the main research target of the scientists working in this field should be a precise state-contingent market strategy, prescribing which static parameters of the movie should be chosen (star quantity, writer, director, MPAA rating, genre, runtime, budget etc.), at what date the movie should be released (contingent to the situation on the movie market), what should be the full cinema and advertisement profile for the given film (how many cinemas should we run each day and what should be the advertisement policy for the given movie, again, contigent to the many factors, such as previous revenues, attention to the movie etc.) and the full piracy-fighting plan with exact number and exact set of actions for each particular case (what should we monitor and how to act when the pirated content of the different quality will emerge, should we

bring an action against just the biggest web sites with low searching costs (which is not too cheap) or try to purify the internet from the illegal copy of the content we distribute (which is even more expensive), how the cinema quantity profile and the marketing policy should change in this case etc). For this part of an investigation the full value structural model should be constructed. Furthermore, the tools of the game theory analysis may be useful in solving these problems and developing the full dynamic and state-contingent strategy.

References

1. Asur, S., Huberman, B. A. (2010). Predicting the Future With Social Media, arXiv: 1003.5699
2. Dellarocas, C., Zhang, X.(M.), Awad, N. F. (2007). Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures, Journal of Interactive Marketing, Volume 21 Issue 4, Pages 2-94
3. Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant,L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232): 1012–1014.
4. Goel, S., Hofman, J., Lahaie, S., Pennock, D., and Watts, D. (2010). Predicting consumer's behavior with Web search. Proceedings of the National Academy of Sciences, 107(41): 17486–17490.
5. Kholodilin, K., M.Podstawski, and Siliverstovs, B. (2010). Do google searches help in nowcasting private consumption? Technical Report 256.
6. Lica, L., Tuta, M. (2011). Using Data From Social Media For Making Predoctopns About Product Successes And Improvement Of Existing Economic Models, International Journal of Research & Reviews in Applied Sciences; 2011, Vol. 8 Issue 3, p301
7. Ma, L., Montgomery, A. L., Param, V. S., and Smith, M. D. (2011). The Effect of Pre-Release Movie Piracy on Box-Office Revenue, eBusiness & eCommerce eJournal 03/2011; DOI:10.2139/ssrn.1782924
8. Peukert, C., Claussen, J., Kretschmer, T. (2013). Piracy and Movie Revenues: Evidence from Megaupload: A Tale of the Long Tail? Social Science Research Network, SSRN-id2313118
9. Rob, R. and Waldfogel, J. (2007). Piracy on the silver screen. Journal of industrial economics, vol.5, pp.379-395
10. Sawhney, M., Eliashberg, J. (1996). A parsimonious model for forecasting gross box-office revenues of motion pictures. Marketing Science 15(2): 113–131
11. Schmidt, T. and Vosen, S. (2009). Forecasting Private Consumption: Survey-based Indicators vs. Google Trends. Ruhr Economic Papers 0155.
12. Varian, H. R. and Choi, H. (2009). Predicting the Present with Google Trends.

Data sources

1. Box office database: http://boxofficemojo.com (Movie Web site with the most comprehensive box office database on the Internet. Founded in 1999)

2. Internet movie database: http://imdb.com (The world's most popular and authoritative source for movie, TV and celebrity content. Founded in 1990)

4. Torrents database: http://thepiratebay.sx (The biggest database of torrent files and magnet links. Founded in 2003)

5. One more piracy information source: http://vcdq.com (The biggest database containing release news of the illegal content and provide the service which allow to for the users to estimate the quality of the content. Do not provide any direct links to the content. Founded in 2001)

6. Search volume index provider: http://google.com/trends (Google service based on Google search, providing search volume indexes for particular search-term item. Founded in 2004)