

2 Capture and curation of a data set on Facebook applications

2.1 Chapter overview

An important and extensive part of this dissertation consisted of building an original data set on Facebook developers and applications. With the start of the platform in May 2007, Facebook introduced a directory in which applications were listed, on a daily basis, with information on their functionality, their developers, their usage, and user reviews. The directory with all applications was openly accessible and was used by Facebook users to browse and discover new applications. However, Facebook did not provide an interface for accessing the information contained in the directory. Early on, developer entry rates to the platform suggested large numbers of applications which meant that data could not be collected manually but that an automated approach had to be chosen.

The opportunity of developing a novel and unique data set on digital products was realized by the author at the time of the launch of the Facebook platform in late May 2007. Since an automated approach had to be developed to collect the data from the directory, I approached Matthias Keller, a trained computer scientist, and we jointly developed a strategy to build the data set. There were several objectives. First, data collection was intended to begin as soon as possible in order to capture the dynamic early stages of the nascent market. Second, data should be collected on a daily basis, the same frequency as Facebook updated the data. Third, new applications and developers had to be discovered constantly in order to achieve a full representation of the market. Fourth, the raw data was to be saved locally in order to document the process and to be able to extract different elements at later stages.

The strategy of building the data set for economic analysis was structured in different stages which were based on approaches commonly employed in data-intensive science (Tansley 2009): the capture, the curation, and the analysis of the data. This chapter describes the first two activities of this process. The step of *data capture* involved automated programs known as “web crawlers” that automatically identified the applications in the directory and stored a web page for each application. Furthermore, an automated “parser” extracted the relevant data from the web pages and stored them in a data base. In the *data curation* step, the raw data was further processed and prepared for the subsequent empirical analysis.

The first step of the development of the web crawler was performed in parallel to the dynamically changing market. The design and implementation of the crawler was complicated by Facebook's frequent changes to the design and structure of the directory. Also, the number of applications, for which Facebook generates separate web pages dynamically, was steadily and rapidly increasing. It took several iterations of development and testing until a stable version of the program was in place: consistent data collection started on July 2, 2007. In the next step, the program that extracts the data from the collected and saved web pages (the "parser") was developed. Due to several changes by Facebook to the design and structure of the directory, this parser was improved in several iterations, the last update being implemented in September 2009. The parser was mainly developed by Jakob Keller and its scripts are in detail documented in Keller (2009). The last step of curating the data was also an iterative process which was performed by the author. In several steps, preliminary analyses of the export data identified inconsistencies in the data. Based on these findings, the parser had to be adapted and run repeatedly on the raw data collected from the crawler.²⁹ A detailed description of the steps taken in each of the three phases is given later in this chapter.

The resulting data set consists of more than 32,000 applications that were active on the Facebook platform in the period from July 2, 2007 to June 30, 2008. It captures the development of the platform and provides time-series data on the application-level for the interesting period of the first year since the launch of the platform.³⁰ In the following, I discuss the features and limitations of the data set.

The data set is a unique collection of the population of products in a digital market. Due to the approach of capturing the data, the data set is superior to many data sets that are usually compiled in retrospective. It offers a nearly complete account of all entities: applications which are active and "alive" but also the ones which are diminishing, failing, or exiting. Consequently, the data set at hand does not exhibit survivor bias that is a caveat to many studies that rely on data early in the lifetime of a firm or

²⁹ Due to the large number of web pages retrieved by the crawler of which each had to be parsed individually, one run of the parser in the last iteration in September 2009 took more than five weeks.

³⁰ The only notable drawback of the data set is that it does not cover the first six weeks of observations between the launch of Facebook Platform and the beginning of data collection.

product.³¹ Another notable feature of the data set is the variety of variables that it contains. It covers not only basic information about applications, such as their name, description and category, but also usage metrics, information about their developers, a complete download of application Wall and discussion board entries as well as various other attributes. This feature alone differentiates the data set from the ones used in other studies on Facebook applications. It appears unlikely that a comparable data set exists outside of Facebook itself since other recent studies are based on smaller data sets of Facebook application data. For example, Gjoka et al. (2008) initially attempted to collect data from the Facebook application directory, but ultimately decided to rely on an alternative data source. Onnela & Reed-Tsochas (2010), who study the effect of social influence on the adoption of applications, limit their data set to 2,720 applications and the time period from June 25 to August 14, 2007.³²

A limitation of the data set is caused by the challenging environment for the capture of the data. The design of the crawler retrieves data for applications only, when they are discovered in the application directory on that same day. Since it is not possible to make up for missed observations at a later time, it is unclear how many applications and observations are missing. However, it is unlikely and exploratory analyses do not suggest that applications are systematically missing. Also the number of applications in the data set compared with reports from Facebook and industry observers indicate that most and particularly the major applications are represented in the data set. As a consequence, one can assume that the data set as captured and later reduced is a meaningful representation of the overall population of Facebook applications.

The remainder of this chapter is structured as follows. In section 2.2, I describe the structure and content of the Facebook application directory and individual application pages in the directory. The process of capturing the data from the source as well as the resulting raw data set is described in section 0. Section 2.4 covers the additional steps of data curation which are necessary to arrive at a data set which can be used for the empirical analyses.

³¹ Data problems in form of survivor bias are often found in studies of first-mover advantage and are discussed in different review articles (Golder & Tellis 1993; VanderWerf & Mahon 1997).

³² Others studies on Facebook focus on user networks and interactions, not applications and developers (Acquisti & Gross 2006; Lampe et al. 2007; Stutzman 2010).

2.2 Data source

This section describes the Facebook directory of applications and the application directory page. Both are sources from which the data is captured and extracted.

Figure 3 provides a screenshot of the Facebook application directory (“directory”) at the time at which data was collected for this study.³³ It lists applications by different characteristics such as “newest” or “most active users”. It also allows users to browse 22 different categories of applications. The directory contains a categorized list of all applications that were approved by Facebook. Submitting an application to the directory is free of charge and requires only a few clicks on behalf of the developer. The review is done by Facebook staff and is based on the function of the application, the inclusion of an appropriate icon as well as a couple of test users. Facebook does not consider quality, (brand) name or the possibility that a similar applications is in the directory already. Also, Facebook does not take any responsibility for the developer’s actions on the directory page.

While it is not obligatory for developers to apply to list their application in the directory, there are no reasons for developers to choose to be not listed and I am not aware of any instances in which developers did. In fact, when there were some delays in the approval process during the first weeks of the platform, developers chose to release their application and list them in “underground” directories. These applications were approved later on and eventually appeared in the directory.

In addition to the directory, each Facebook application has a dedicated directory page which contains information about the application and tools for users to rate and discuss the application (“application page”). An exemplary application page is provided in Figure 4.³⁴

Each application’s directory page consists of different elements that provide the data for this data set. It contains basic information such as application name, developer name and profile/website link, a short description of the application’s functionality as well as the categories which this application is assigned to within Facebook (see (1) in

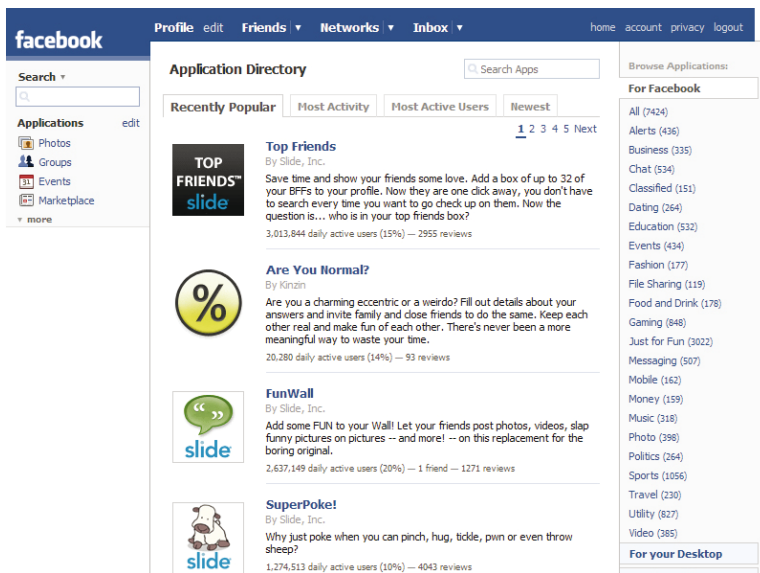
³³ The design and presentation of the directory has changed since the time period that this study examines.

³⁴ Note that the design and content of directory and the application page has changed since the time period that this study examines. The application page of Figure 4 was saved in March 2008 and represents the design and content for the period from May 2007 until June 2008.

Figure 4). This information is edited by the developer of the application. Additional information is available for the usage of the application (see (2)). Facebook displays both the number of active users (per day; i.e. 24h)³⁵ and the percentage of active users of all of the application's installations. These measures based on user engagement displaced the reporting of number of total installations 14 weeks after the platform's launch. The page also has elements in which users of the application can leave a note on the virtual blackboard ("Wall"), write a review or join the discussion forum (see (3)).

The process of how the data from the directory and the application page was captured and further processed is described in the following section.

Figure 3: Screenshot of the Facebook application directory with 22 categories



³⁵ Here active usage is defined by Facebook as a user having actively interacted with the application within the previous day, midnight to midnight. See the announcement on the Facebook developer blog for details (Facebook 2007a).



Birthday Alert

Weitere Anwendungen durchstöbern

E-Mail-Adresse:

Passwort:

☐ Angemeldet bleiben

Anmelden

Passwort vergessen?

Jeder kann mitmachen

Registrieren



Never miss another Birthday! Birthday Alert sends you emails alerting you of ALL your friends' birthdays. You can ...

- manage your daily or weekly emails and notifications,
- see photos of your friends with upcoming birthdays on your Profile,
- add friends who are not on Facebook,
- invite your Facebook friends to join!

And unlike other Birthday applications, NO invitations are required!

Thanks to deem (http://www.flickr.com/photos/dee_m/) for her fantastic photos!

Facebook is providing links to these applications as a courtesy, and makes no representations regarding the applications or any information related to them. Any questions regarding an application should be directed to the developer.

Über diese Anwendung

★★★★★ (4.3 von 5)
Auf 46 Rezensionen basierend

Benutzer:
9.515 tägliche Benutzer
2% von insgesamt

Kategorien
Meldungen, Nur zum Spaß

Diese Anwendung wurde **nicht** von Facebook entwickelt.

Über den Entwickler

 Jeff

Fans

6 von 1.780 Fans

 Disha  Khong  Lynda

 Gunjan  Christine  Hiran

Diskussionsforum

Es werden 3 von 19 Diskussions Themen angezeigt Alle anzeigen

SquidNote (aka Group Card) feedback

3 Beiträge von 3 Personen, Aktualisiert am 6. April 2008 um 08:28.

Things to Improve

20 Beiträge von 10 Personen, Aktualisiert am 26. Februar 2008 um 23:15.

New Communication Forms

1 Beitrag von 1 Person, Aktualisiert am 29. Januar 2008 um 12:43.

Rezensionen

Es werden 8 von 46 Bewertungen angezeigt Alle anzeigen

★★★★★ The implementation seems nice, the one thing that bothered me was promptly fixed!



2.3 Data capture

This section describes how the data from the above introduced data source is captured. First, section 2.3.1 provides details on the three distinct steps of the data capture process. Second, section 2.3.2 provides an overview and description of the data files that are the result of the data capture process.

2.3.1 Process

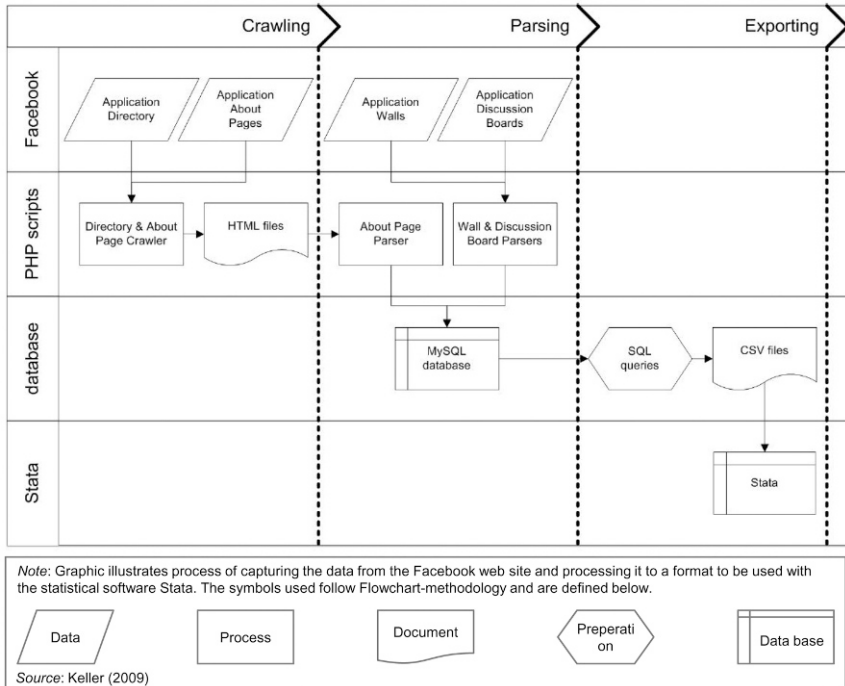
The data capture process is composed of three distinct phases: the *crawling* phase, the *parsing* phase, and the *exporting* phase. These phases need to be carried out in chronological order, since results from one phase are used as inputs for the next step. Figure 5 illustrates the three phases as well as the resources that are used during the data collection process. The process of data capture involves raw data pulled from *Facebook*, data handling and processing using *PHP scripts* (PHP: Hypertext Preprocessor), and data storage and retrieval using a relational *data base*. Finally, the export files are further processed in the *Stata* software package. This final step is described in the following section on data curation.

Crawling phase: The data source, the directory and application page, is highly dynamic and changes constantly. This poses a challenge because historic data is lost if it is not preserved in some way. The objective of the crawling phase is to discover and permanently store what would otherwise be lost data. For this purpose, a set of PHP scripts (the “crawler”) was developed shortly after the public launch of Facebook’s platform for applications in May 2007. The crawler automatically browses the application directory and attempts to identify all applications contained in it. The crawler also downloads and saves each application page in a separate text file for later analysis. These text files represent the key input for the parsing phase. The crawler started to store application pages on July 2, 2007 and continued until July 2009.

Parsing phase: Facebook provides its application directory and application pages in the form of hyper-text markup language (HTML) documents. HTML is a standardized language that contains structured text along with formatting markup and other elements for the purpose of presentation by the user’s browser. The markup allows the identification of the relevant elements of the HTML file. The data is then assigned to

specific data fields in a data base. This process is here referred to as parsing and describes the extraction of the desired content from the HTML documents.³⁶

Figure 5: Flowchart of process and resources of the data capture process



Exporting phase: The goal of the exporting phase is to prepare the content of the data base for later analysis in a statistical software package such as Stata. The data base is an intermediate representation of the gathered data as it structures the data and integrates it in tabular and relational form. However, Stata is not equipped to handle relational links between data base tables. These need to be transformed into strictly tabular data structures. To create these exports, a set of SQL (structured query language) queries is run against the data base. The results are saved into CSV (comma-separated values) files suitable for import in Stata. These export files are described in the following section.

³⁶ Three separate parsers were written for the application pages as well as for the comments and discussion entries by users. A detailed description of the parsers and the relational data base schema are available from the author upon request.

2.3.2 Export file

This section introduces the export file and the raw data that was captured by the automated crawler and parser and then saved in a data base. The main export from the data base, hereafter referred to as application export file, consists of daily observations from July 2, 2007 to June 30, 2008. It holds data on application usage metrics, some developer measures, count data on updates, Wall posts as well as an assignment to categories.³⁷ A detailed description of the data, the variables and first summary statistics are presented in the following section 2.4.1.

When compiling the application export file, the exact timing of entry had to be treated with care. Determining the market entry of most Facebook applications is straightforward. When developers decide to publicly launch their application, they submit it to the application directory. In most cases, the application is approved by Facebook within a few days and it appears in the list of “newest” applications. This event is considered the market entry of the application. The data set directly reflects the time of this event due to the daily execution of the directory crawler that downloads and stores the application pages of all applications it finds in the directory.

Later, the application page parser transforms each of these stored pages into an observation in the data base. The date of the earliest of these observations is the best estimate for the time of entry of most applications in the data set. However, the delay of six weeks between the launch of the directory and the start of data collection necessitates refinement of this general approach. Applications that are observed on the first day of data collection might have been launched exactly on that date or up to six weeks earlier, when the Facebook platform launched publicly. The estimation of the entry date of an application may be improved by analyzing contributions to the application’s Wall or discussion board. Posts that date earlier than the first observation indicate that the application was launched prior to the previously determined date. In August 2007, an automated crawler and parser collected the date of all user messages to the Wall or the discussion forums of the application page. The earliest date of these first posts is used to improve the accuracy of the estimated market entry for each application in the data set. This approach reduces the entry date of applications that launched on or before July 2, 2007, on average by 11.81 days indicating an improved

³⁷ There is a second export file that only contains data that matches applications to developers. This data is needed for the analysis of developer portfolios in chapter 3.5 and chapter 5.

estimate of the market entry for those applications. I tested the accuracy of this approach by also applying it to all applications in the data set. Here, the estimated entry date from the first Wall post is only 0.43 days before the date derived from the first appearance in the directory.

2.4 Data curation

This chapter describes the process of data transformation from the raw export data to a data set which is used for statistical analyses. First, the data is imported in the statistical software package (see 2.4.1). Second, the data is manipulated in several steps that ensure a consistent data set for subsequent empirical studies (see 2.4.2). Besides descriptions of the separate steps of the process, this section also contains first summary statistics of the data.

2.4.1 Data import

The export from the SQL data base produces a number of comma-separated text files that need to be imported and merged for subsequent statistical analysis.³⁸ During the import stage variables are named, labeled and formatted. The resulting imported data set contains data on 32,543 applications for the time period from July 2, 2007 to June 20, 2008 (i.e. 3,506,562 observations).

Table 1 lists the variables that identify each observation (unique identifiers for applications and dates) as well as 21 variables that capture different application characteristics. The data set contains information on the name and the category that each application was assigned to at a given date. The application data also provides information on the developer who is responsible for the application. Here, the distinction is made between individuals, companies, and Facebook. A developer who puts down his name and identity as a Facebook user on the application page in the directory is defined to be an individual developer. Since more than one individual can be assigned to an application, another variable counts the number of individuals for every given day. On the other hand, developers can also include the name or web site of a company or project on the application page. In this case, the developer is referred to as company developer. Finally, Facebook has its own applications (like a photo slideshow or an event management tool) which are also included in the directory.

³⁸ In this dissertation, the data is prepared with the statistical software package Stata 11 (see Stata Corp.'s website for more information: <http://www.stata.com>).

Interdependencies in the Discovery and Adoption of
Facebook Applications

An Empirical Investigation

Mayrhofer, P.

2013, XVII, 172 p. 23 illus., Softcover

ISBN: 978-3-8349-3886-2