The Students' Portal of Ilmenau: A Holistic OSN's User Behaviour Model

Thomas Paul* Stephen Stephen* Hani Salah* Thorsten Strufe†
*Technische Universität Darmstadt, Germany †Technische Universität Dresden, Germany

***Abstract*—Online Social Networks (OSNs) have become an essential part of the social life for more than one billion people. OSNs have received a considerable attention from different research communities. OSN providers, however, rarely share their data in order to protect both their business secrets as well as the privacy of their users. Data access limitations have forced researchers, in order to study *several* user behaviour aspects, to use data collected from *different* and *inconsistent* datasets. Correlating different datasets, however, is impossible or hard to validate. In this paper, we provide a *holistic* analysis of user behaviour in OSNs using the *Students' Portal of Ilmenau*[1] as a case study. Our analysis is based on a log-file level dataset, providing insights into the observed churn, usage patterns as well as social graph properties.**

I. Introduction

A core aspect of today's Internet usage is to connect people with each other. Online Social Networks (OSNs) are communication platforms that allow their users to create and maintain digital representations of themselves (profiles) to establish friendship connections amongst each other, to exchange messages and to share digital content. In 2013, more than one billion people actively used Facebook alone[2]. This new and popular communication paradigm has also attracted researchers from different fields: Psychologists and sociologists are interested in how humans integrate OSNs in their practices as well as in the impact OSNs have on their user's daily life. Computer scientists, aiming at improving OSN platforms, need to know how OSN users behave.

OSN providers, however, rarely share their data. Instead, OSN research has been done on crawler-generated data, surveys, ISP traffic analysis and data from social network aggregators. Authors of [1] correlated different profile properties with the personality of OSN users. [3] and [15] evaluated the churn behaviour (login / logout patterns) and the functionalities which are used by OSN users. [13] elaborated with whom users interact with respect to the distance in the social graph, and [10] described graph properties.

The aforementioned studies (and many others) focused on different aspects of different OSNs. The conclusions are drawn from distinct datasets, collected using different methods in different networks at different time spans. We argue that it is hard to merge all these aspects together into one *holistic* model in a valid way. The absence of a holistic user behaviour model does not only affect the quality of analysis, but also does not allow to build realistic simulation frameworks. This is the gap that we aim to fill in this paper.

[1]https://spi.tu-ilmenau.de/

[2]http://newsroom.fb.com/Timeline

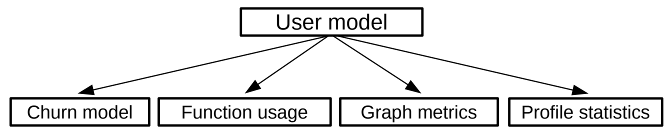

Fig. 1. Components of our user behaviour model

The students' media research community (FeM)[3] at the Ilmenau University of Technology in Germany runs and maintains a social network of more than 35,000 users, called *Students' Portal of Ilmenau* (SPI). FeM allowed us to access a privacy-preserved log-level SPI dataset. In spite of the relatively small number of users in SPI (e.g. compared with Facebook) and the very localized set of users (university students and alumni), we argue that it is worth to analyze the SPI's dataset for the following reasons:

- It is a holistic dataset, containing each single action performed by each user. This allows us to look at several aspects.
- We can observe human behaviour and compare the aspects with results of former (limited scope) studies.
- The network size is big enough for statistical evaluations and, at the same time, is small enough to afford the computational costs for the evaluations.

Our user model consists of four different components (Figure 1), each describes one user behaviour aspect: (*i*) We describe the churn behaviour by calculating session durations, inter-session time distribution as well as daily and weekly usage patterns. (*ii*) We provide statistics about the usage of several functions, (*iii*) calculate graph metrics and (*iv*) study profile statistics, showing which fields (e.g. name, nickname, address) are filled and how many pictures are included in the profile. Due to space constraints, we present in this paper the first three parts only.

The remainder of this paper is organized as follows: Section II describes the related work. Then, we introduce our methodology in Section III. Next, we describe our churn model in Section IV, the usage model in Section V and the social graph properties in Section VI. Finally, we discuss the main findings and conclude the paper in Section VII.

[3]In German, *"FeM"* stands for *"Forschungsgemeinschaft elektronische Medien e.V."*, which translates in English to: *"students' media research community"*.

II. Related Work

We classify the related work according to both: the data sources (i.e. data collection methods) and the scientific questions which can be addressed by each data source.

Crawlers are a popular tool to explore OSNs. A crawler starts at one profile and walks along friendship connections (in case they are accessible). Crawlers collect only the public parts of discovered profiles; they do not gather information about any activity without an impact on the user profile. They hence are neither suitable for elaborating which function is used (e.g. chatting or content consumption) nor provide information about churn. Studies based on crawlers' data [7], [11], [10] mainly analyze social graph properties.

Social network connectors are tools which simplify the usage of multiple OSNs. They act as a common gate to different OSNs. Data which are gathered from this type of service can only cover those actions which are performed via social network connectors. One advantage of social network connectors is that they can collect data from multiple OSNs simultaneously. Benevenuto et al. [3] elaborate churn behaviour based on social network connector data from Brazil.

HTTP header traces from ISPs represent a very powerful data source. Beside the session length, many other activities like chatting, message sharing and app usage can be monitored. That is true even for collecting data about multiple OSNs. The main drawback of HTTP header traces is that they can hardly be used to monitor the friendship graph. Instead, the analysis based on HTTP header traces [15], [6] concentrate on usage patterns of functions as well as external content retrieval.

Survey data have very different properties. Studies which are based on surveys need active involvement of users and thus are not as scalable as crawler-based studies. The main advantage of using surveys is that users can be asked about data which cannot be collected by other methods. They introduce the chance to understand why users are performing an action, and they also allow to ask about facts which are not directly related to OSNs (e.g. personal habits, preferences or diseases). Studies based on surveys [1], [9] can, for instance, gain insights into whether there is a coherence between personality and user profile, and how actions are perceived by other users in the OSN.

The ideal data source for researchers is the OSN provider itself. In rare cases, Facebook allowed researchers to perform some analysis [2] on their database. This opportunity was employed to answer: "How Facebook users allocate attention across friends?". However, [2] does not provide a holistic model which allows to simulate a system based on a single dataset nor to find different coherent aspects across different properties (e.g. profile vs. functionality usage).

III. Background, Data Description, and Ethical Considerations

The SPI platform was created in 2001 by the students' media research community (or FeM) at Ilmenau University of Technology in Germany. The users are the students and alumni of the university. SPI offers both: functions that support study as well as functions for leisure time. For instance, the students can organize lecture schedules and study groups, download presentations and exercises, view the cafeteria menu and receive weather forecasts. SPI also enables its users to maintain profiles and to connect those profiles with each other. Furthermore, users can chat, send offline messages, create photo albums, write in forums and share their opinion in diaries.

There are more than 35,000 registered users in SPI. However, we observed 12,604 active users in 650,384 sessions and 87,182 buddy links. Inactive users mainly represent alumni. The analyses which we present in this paper are conducted by analyzing the activity stream and the social graph. In total, there are 57 functions (or services) that later can be separated through various filters, such as periods (time-stamps). SPI does not require to confirm friendships, nor utilizes friendships for access control.

During the course of evaluations, we had access to an anonymized dataset not containing any communication information. We only evaluated the occurrence of user actions, and their metadata such as timestamps and their order. Nobody was able to relate actions or user profiles to individual users. In spite of the user anonymization, we do not publish any non-aggregated data.

IV. Churn Model

Churn models describe how often, how long and when users use a function. In this section, we describe the churn behavior observed in SPI. In particular, we provide information about session starting (i.e. login) frequencies with respect to daytime and weekdays, the session duration distribution and the inter-session time distribution.

A. *Frequency of Session Starts*

We observed 650,384 sessions in total. As can be seen in Figure 2, the majority (81.7%) of sessions started between 10:00 and 22:00. The lowest number of sessions (1.3%) occurred between 3:00 and 6:00. The consequence is that periods occur during night time when no single user is active. This churn model strongly impacts assumptions of P2P-based OSN approaches that have been proposed in e.g. [4], [5]. The P2P service would be unavailable in case that no nodes are connected to the network.

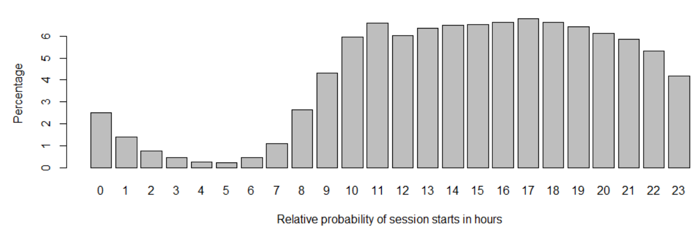

Fig. 2. Frequency of session starts during an average day (hourly binned)

Beside diurnal patterns, we also discovered weekly usage patterns. The days with the highest and lowest fractions of session starts are Tuesday (16.6%) and Saturday (10.6%), respectively (Figure 3). These results reflect the fact that SPI is a network of students (in German universities, lectures start on Monday morning and stop on Friday evening).

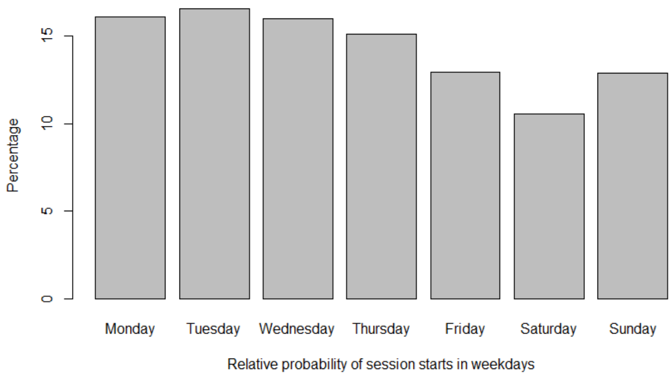

Fig. 3. Frequency of session starts during an average week

B. Session Length Distribution

As shown in Figure 4, 18.5% of the sessions consisted of a single action only, 28.2% were shorter than one minute and the durations for 24.8% of the sessions were between one and five minutes, 13.4% lasted between five and 15 minutes and 6.57% lasted between 15 and 30 minutes. We also discovered a strong peak at half an hour that can be explained by the university schedule: Since lectures last for 90 minutes and lessons are starting bi-hourly, students have a break of 30 minutes between two lessons.

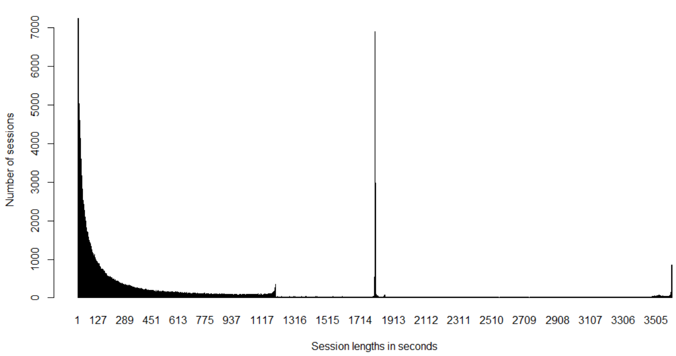

Fig. 4. Sessions length distribution during an average day

C. Inter-session Time Distribution

This is the distribution of time gaps between two consecutive session starts (i.e. logins). Figure 5 shows the results for active users only: Out of 496,458 inter-session gaps, majority (55%) lasted shorter than five minutes, 24.8% were between five and 30 minutes, and 8.9% lasted longer than six hours (several users logged in only once during the observation period). The longest observed inter-session gap lasted 5,759 hours. We observed 716 sessions and 1,069 sessions with inter-session gaps longer than 5,000 and 4,000 hours, respectively.

Furthermore, we compared the sessions of active users with those of users in general: The distribution of active users was longer. This is unexpected result: One would expect active users to have shorter inter-session gaps due to their frequent login behavior.

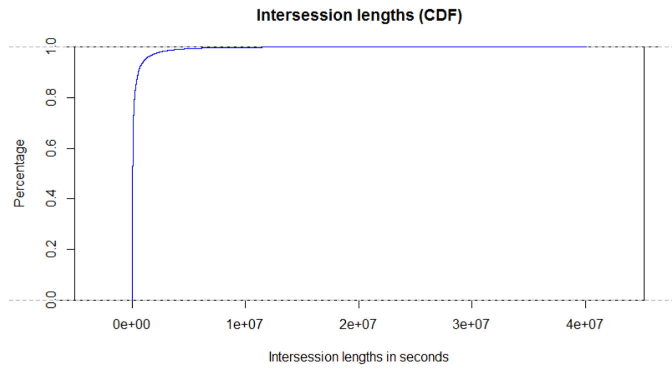

Fig. 5. Inter-session time distribution

V. Usage Model

As mentioned in Section III, SPI offers 57 different functions to its users. Knowing how often each function is used allows to: understand the reasons why users use SPI, and to predict server and network load. Understanding usage patterns gives insight into how students integrate SPI in their daily life, and is also essential to develop accurate simulation models.

Due to space constraints, we do not describe the daily and weekly patterns of all functions. Instead, we first overview the popularity of functions and than provide some patterns of the most popular functions. We close this section by describing the most popular function transitions.

A. Popularity of Functions

We distinguish among three types of functions: (*i*) functions that directly relate to studies, (*ii*) supportive functions that help users to organize their life in Ilmenau and (*iii*) functions that can be used during leisure time. The first group includes access to: the study forum (write and read) as well as to the study group's internal communications (send and receive). The second group includes: inserting, answering and access to advertisements, and access to the cafeteria menu and weather forecasts. The most popular functions in the third group are: view pictures (24.72%), access to user profiles (13.18%) and access to the universal forum (9.23%).

B. Diurnal Patterns of Function Usage

The patterns of diurnal function usage show huge differences among different functions (Table I): While patterns of some functions, e.g. show gallery (Figure 6), are very similar to the system churn (Figure 2), patterns of other functions show special diurnal patterns. For example, accessing the cafeteria menu (Figure 7), unsurprisingly, is mainly accessed around noon.

There are some more interesting daily patterns. One can be realized by comparing the access frequency of reading own guestbook (Figure 8) and the daytime when removing entries from it (Figure 9). While reading own guestbook has a very

Function	Frequency	Percentage
View pictures	1,874,978	24.7215
View profile	999,854	13.1830
Read universal forum entry	699,971	9.2291
Call start page	576,814	7.6053
Access market place	560,447	7.3895
Access study forum	526,730	6.9449
Search request	463,802	6.1152
Study group page view	259,542	3.4221
Access newsfeed	254,956	3.3616
Study group's internal forum	243,293	3.2078

TABLE I. POPULARITY OF FUNCTIONS

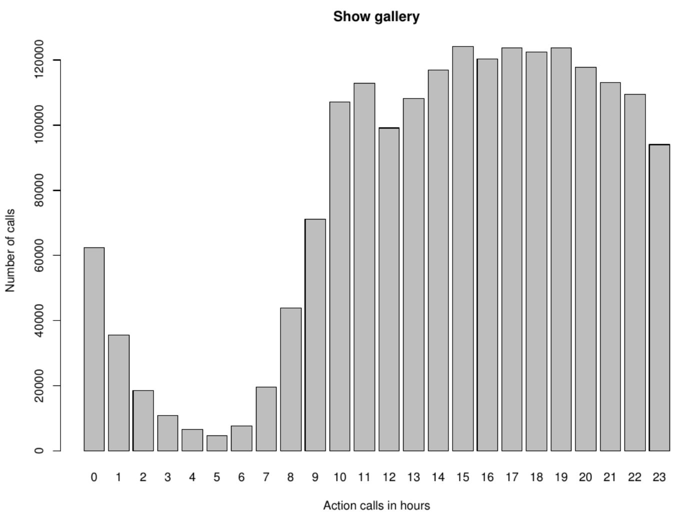

Fig. 6. Show gallery: Sessions frequencies during an average day

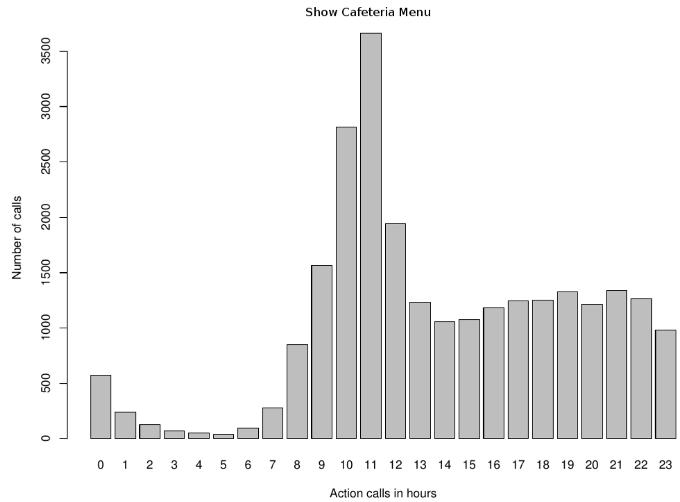

Fig. 7. Show cafeteria menu: Sessions frequencies during an average day

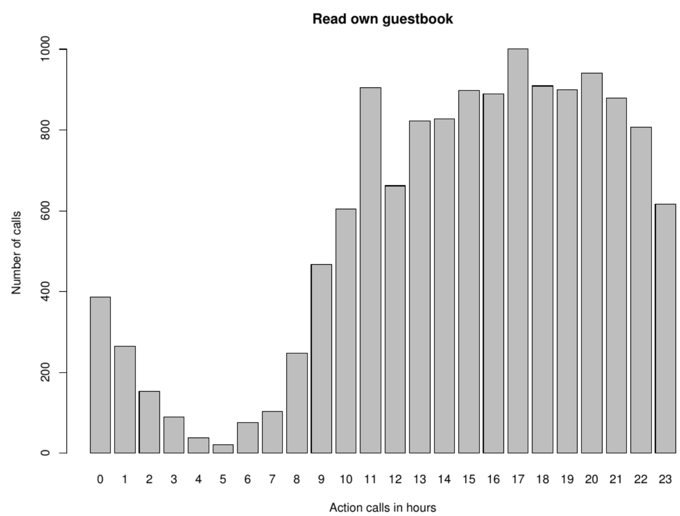

Fig. 8. Read own guestbook: Sessions frequencies during an average day

similar pattern to the system churn, removing entries is done mainly at odd hours (11:00, 13:00, 15:00 and 21:00). These observations can be correlated with the university schedule which is organized in two-hour slots (lectures start at the beginning of an odd hour; the latest starts at 19:00).

C. Transition

Figure 10 summarizes the transition from one function to another: There are several functions that lead to show profile or to show gallery. 35% and 29% of the transitions to show gallery occur after looking at the user's own gallery and after showing the profile of another user, respectively. However, the topmost predecessor is show gallery itself. The second target function after itself is show profile with 10% occurrences. For show profile, 48% of the transitions after show own group lead to show profile, 44.3%, 43.8%, 37.3%, 35.9% from read an entry in guestbook, show user details, add buddy and show homepage, respectively. Coming out from show profile beside show gallery itself with 17.8% and show homepage with 10.8%.

Out of 650,384 sessions 419,857 sessions (64.6%) started without cookies. After logging in, 16.7% called the homepage and 6% called the feeds. Aside from these functions the transitions are dispersed in another 55 functions. After showing homepage, the users with 35.9% occurrences are showing profiles of other users or just refreshing the homepage (24.2%). When the users read the feeds, they usually stay there (42.9%) or just do something else (less than 10%).

There are two ways to logout in SPI, one is to close the browser (cookies will be removed), while the other one is to logout through pressing the ”logout“ button. Among the users who use the first choice, 13.7% were looking at the homepage, while 11.1%, 10.8% and 1% were searching, showing profiles and showing feeds, respectively. For the latter choice, 24.5% were sending chain mails to their groups and 14.3% were looking at the cafeteria menu.

VI. SOCIAL GRAPHS

Social networks can be modelled by graphs in which nodes represent users and edges represent connections among users. The edges represent any kind of functional relation between two nodes. Table II provides an overview of the seven graphs (every two-sided functionality) which we elaborate. The two-hops buddy graph consists of buddies and buddies of buddies. All the graphs that we elaborate are directed graphs. While this is obvious in case of one user is viewing the profile of another

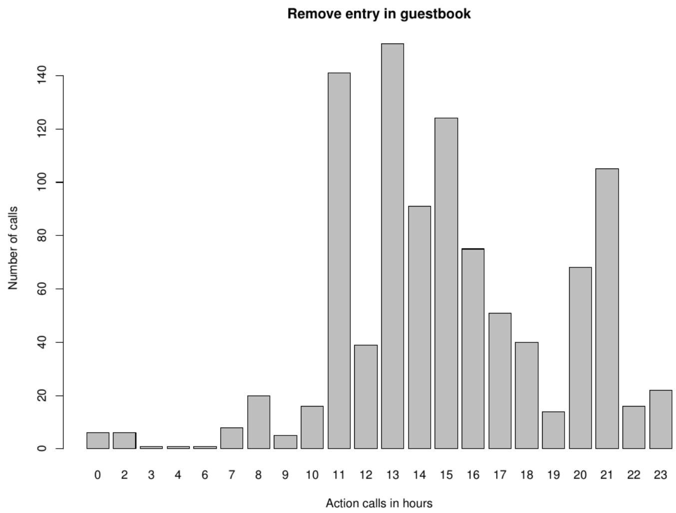

Fig. 9. Remove a guestbook entry: Sessions frequencies during an average day

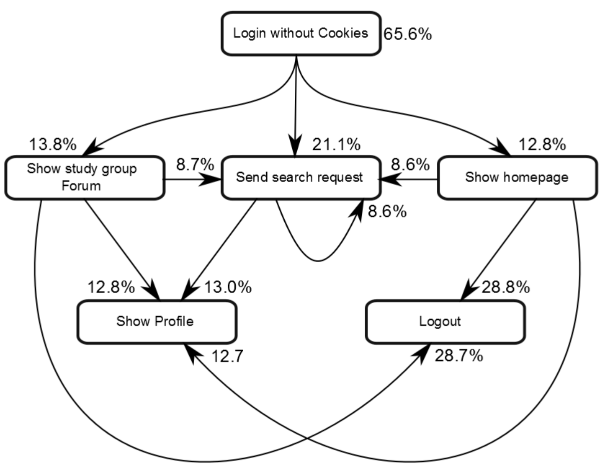

Fig. 10. Transitions of function usage

user (graph 3), it also holds for other graphs. For instance, in the buddy graph (graph 6), users can befriend with other users without approval.

Our analysis includes the size of the graphs, how tightly connected the users are, the effect on the graph connectivity in case that users leave the network and the similarities among the graphs. We performed the analyses using the GTNA framework [14].

A. Shortest Path Length

The average shortest path length is the average shortest distance among all graph nodes. The diameter of a graph is the longest shortest path between any two nodes in the graph.

Table III, Figures 11 and Figure 12 show that profile views (graph 3) create the most densely connected graph. This is to be expected since viewing a profile of another user is a predecessor for performing a couple of other functions (e.g.

Graph no.	Graph name	Users	Edges
1	Write in guestbook	1,348	1,404
2	Read guestbook entry	5,301	9,769
3	Show profile	23,524	576,985
4	Show user details	13,888	75,545
5	Show gallery	12,766	184,914
6	Buddy graph	12,746	87,182
7	Two-hops buddy	12,496	1,232,114

TABLE II. SOCIAL GRAPHS

Graph no.	Highest freq.	Avg.	Diameter
1	2	4	20
2	6	6.2	25
3	3	3.7	12
4	5	4.5	16
5	3	3.9	14
6	4	4.3	16

TABLE III. SHORTEST PATHS

viewing the gallery, viewing user details or leaving messages in the guestbook). The average shortest path length of 3.9 in accessing photo galleries of other users (graph 5) shows the importance of view pictures function in SPI. Furthermore, we can see the existence of a very heterogeneous subset of viewed pictures with many long distance links. This means that users in SPI do not concentrate on their friends' pictures. We excluded the two-hop buddy graph from this analysis since it is an artificial construction that adulterates the path lengths.

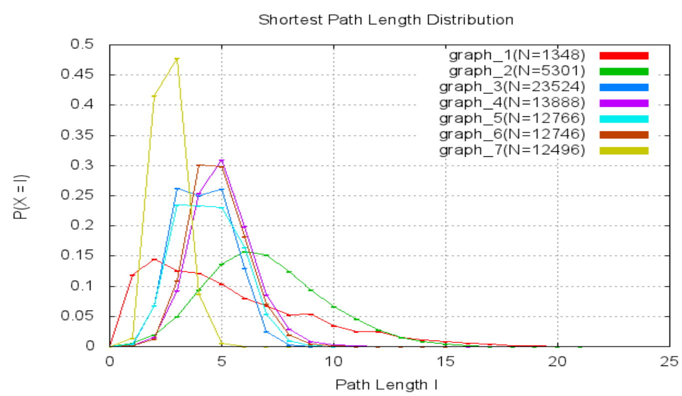

Fig. 11. Shortest path length distribution

B. Overlap of Social Graphs

We analyzed seven different social graphs (Table II) which are derived from using different functions. For example, writing guestbook entries is unpopular when compared with reading guestbook entries. However, evaluating the overlap of different social graphs helps to gain new insights. It shows the orchestration of different communication functions with respect to edges amongst users.

It is unsurprising that a user who writes guestbook entries in another user's guestbook also reads entries of the respective user's guestbook. This can be shown by combining graph 1 and graph 2. The combined graph has 9,777 edges. That means that only 8 users do not check a guestbook in which they

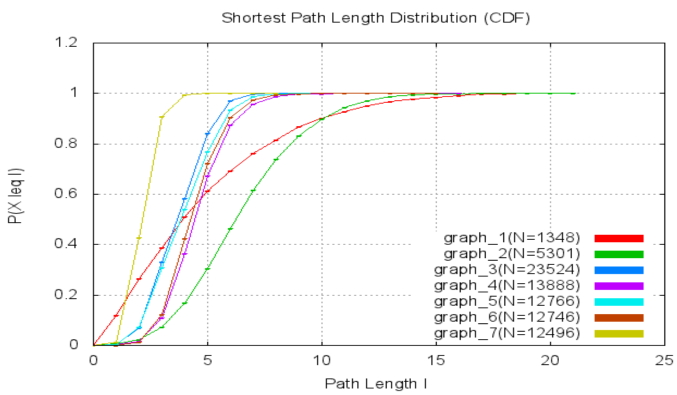

Fig. 12. Shortest path length distribution (CDF)

leave messages. Also, most users who look at photo galleries also look at the profiles of the respective users. Cumulating graph 3 and 5 leads to a graph with 582826 edges (Table II and Table IV). The combined graph has only 5841(1%) more edges than graph 3.

-	3	4	5	6	7
1	576993	76398	185824	88586	1233880
2	577010	82443	190732	96231	1241627
3	-	-	-	-	1789519
4	577026	-	247564	160086	1304457
5	582826	-	-	269239	1409995
6	654126	-	-	-	1248819

TABLE IV. MATRIX OF CUMULATED GRAPHS

Surprisingly, the buddy graph has only little overlapping edges with all other graphs. Thus, SPI is not a tool to communicate with friends.

VII. SUMMARY AND CONCLUSION

The goal of this paper is to provide the knowledge of the metrics needed to create representative simulations for online social networks. This knowledge is represented through the models described. After analyzing the complete network with 12,604 active users, user behaviours on the surface are found to be very stable periodically. By this we mean the total amount of function usage, number of sessions and session lengths per day, time gap between sessions and transition between functions.

In the following, we review and summarize the main findings: Our results show that sessions are extremely short. In particular, they are much shorter than assumed in several former evaluations of P2P-based decentralized OSNs (DOSNs, e.g. [12], [8], [8], [16]). We expect those P2P-DOSNs to perform worse with our churn assumptions (compared with former assumptions). We also provide evidence that P2P-DOSN only works in case it attracts users from many different time zones at the same time.

Usage of distinct functions shows distinct diurnal patterns. This is natural when considering the electronic cafeteria menu that is accessed mainly around noon. However, it was surprising to us that e.g. deleting messages in user's own guestbook does not correlate with reading entries in the same guestbook. Instead, it strongly correlates with the starting times of lectures. We construe this effect to be caused by the student's mood.

Friendship declarations are rare in SPI, since buddy connections are not as crucial in SPI as they are in e.g. Facebook, since they are not used for access control. Users hence add less buddy connections to their user profiles. Instead, access control to content is preferred to be done on the granularity of study groups. We thus argue that the OSN functionalities and their dependencies should be considered while interpreting friendship graphs.

REFERENCES

[1] Yoram Bachrach et al. Personality and patterns of Facebook usage. In *Annual ACM Web Science Conference*, 2012.

[2] Lars Backstrom et al. Center of attention: How facebook users allocate attention across friends. In *ICWSM*, 2011.

[3] Fabrício Benevenuto et al. Characterizing user behavior in online social networks. In *AMC IMC*, 2009.

[4] Sonja Buchegger et al. PeerSoN: P2P Social Networking - Early Experiences and Insights. In *SNS*, 2009.

[5] Leucio-Antonio Cutillo, Refik Molva, and Thorsten Strufe. Safebook: a privacy preserving online social network leveraging on real-life trust. *IEEE Communications Magazine*, 2009.

[6] Christopher W. Dunn et al. Navigation characteristics of online social networks and search engines users. In *WOSN*, 2012.

[7] Emilio Ferrara. A Large-Scale Community Structure Analysis in Facebook. *EPJ Data Science*, 1, 2012.

[8] Sonia Jahid, P Mittal, and Nikita Borisov. EASiER: encryption-based access control in social networks with efficient revocation. In *ASIACCS*, 2011.

[9] Ksenia Koroleva et al. I LIKE IT BECAUSE I (M) LIKE YOU Measuring User Attitudes Towards Information on Facebook. 2011.

[10] Marcelo Maia et al. Identifying user behavior in online social networks. In *Workshop on Social Network Systems*, 2008.

[11] Alan Mislove et al. Measurement and analysis of online social networks. In *ACM IMC*, 2007.

[12] Rammohan Narendula, Thanasis G Papaioannou, and Karl Aberer. A decentralized online social network with efficient user-driven replication. In *PASSAT*. IEEE, 2012.

[13] Khanh Nguyen and D A Tran. An analysis of activities in Facebook. In *IEEE CCNC*, 2011.

[14] Benjamin Schiller and Thorsten Strufe. Gtna 2.0 - a framework for rapid prototyping and evaluation of routing algorithms. In *Proceedings of SummerSim*, 2013.

[15] Fabian Schneider et al. Understanding Online Social Network Usage from a Network Perspective. In *AMC IMC*, 2009.

[16] Nashid Shahriar, Shihabur Rahman Chowdhury, Mahfuza Sharmin, Reaz Ahmed, Raouf Boutaba, and Bertrand Mathieu. Ensuring beta-availability in p2p social networks. In *ICDCSW*, 2013.