Profiling and Predicting User Activity on a Home Network

Xueli An
Huawei Technologies
Munich, Germany
xueli.an@huawei.com

Fahim Kawsar
Nokia Bell Labs
Cambridge, UK
fahim.kawsar@nokia-bell-labs.com

Utku Günay Acer
Nokia Bell Labs
Antwerp, Belgium
utku_gunay.acer@nokia-bell-labs.com

ABSTRACT

This paper reports a study on the characterization of in-home Internet activity behavior based on application usage logs. We collected online activity data from 86 Belgium households for 60 days. We analyzed the activity traces to gain insights on the temporal traffic distribution, interaction regularity, and activity correlations. This analysis is then used to develop a generic method to segment households into designated groups showing similar behavioral profiles. Our technique combines interaction frequencies and regularities across activities for segmentation, and is able to reveal interesting time-slotted profile for each segment. These profiles aim to show the strength of routine behaviors in Internet usage, based on which we present a novel algorithm to predict future Internet activities of a household. Our algorithm shows that 60% of the households online activities can be predicted accurately 70% of times.

CCS CONCEPTS

• **Networks → Network monitoring**;

ACM Reference Format:
Xueli An, Fahim Kawsar, and Utku Günay Acer. 2017. Profiling and Predicting User Activity on a Home Network. In *MobiQuitous 2017: the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, November 7–10, 2017, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3144457.3145502

1 INTRODUCTION

Hundreds of Millions of users interact with the Internet daily. The web environment provides them with avenues to support their learning, leisure entertainment, information about work and, make and maintain friends, and to find out about and engage with the world around them. Despite this massive participation, relatively simple questions regarding online activity behavior remain unanswered. For example: how frequently do different user groups engage with online games or online shopping? Which online activities are performed in tandem? What distribution is appropriate to segment Internet users into different groups and furthermore, what this segmentation tells us about the underlying users behavior? How predictable users' online activities are? Answers to such questions have implications ranging from improving web experience, to custom service design to refining marketing and advertising strategies.

Conventional approaches to understand web activity rely either on surveys, toolbar tracking or website crawling. Surveys are the most popular methodology to assess attitudes and general usage trends [22], but suffers from incomplete and sometimes inaccurate statements of the participants. In contrast, toolbar tracking approach collects web browsing data by installing dedicated plugins in the web browser and provides complete browsing histories for millions of users [16]. Although massive in scale, these data often suffer from sample bias and do not capture the entire web interactions (e.g., non browser activities). Finally, data sets obtained by crawling different social networking websites, e.g., Facebook, Twitter, etc. offer valuable insights on the properties of underlying social graphs [6, 21]. However, these data sets are tailored to the underlying websites and henceforth do not offer a comprehensive view on users online interests and behavior.

In this paper, we divert from this traditional avenue and apply a network-based approach in an attempt to characterize and model peoples online activities. We monitored Internet applications usage logs of 86 Belgium households for 2 months through network packet inspection. We analyze these activity traces to understand temporal distribution and correlation of Internet activities and to identify behavioral similarities among different households to segment and predict their future online activities. Our major contributions are threefold.

- First, we provide an in-depth view of in-home Internet activities that exposes basic usage characteristics, identifies popular Internet activities and establishes the fact that most of the online activities follow a daily pattern.
- Second, we propose a new technique to model recurrent Internet activities based on interaction frequency and temporal regularity and show that how this technique is used to segment households into a limited number of groups reflecting distinguishable behavioral profiles.
- Finally, we propose a novel activity prediction algorithm that uses historical activity data to estimate the probability of a household's future engagement with a set of Internet activities. The algorithm can successfully predict 70% of future Internet activities of 60% of households.

Our segmentation model and prediction algorithm call attention to the developers of future ubiquitous technology in a domestic environment with implications to both end user service design and residential network optimization. With a better awareness of households Internet Activity pattern, application developers can reach their intended households with promotional offers and recommendation services in a timely fashion. Network operators can design personalised dynamic pricing package tailored to household's need.

MobiQuitous 2017, November 7–10, 2017, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5368-7/17/11…$15.00
https://doi.org/10.1145/3144457.3145502

Furthermore, elastic network resources can be better managed with an informed understanding of households Internet usage, which eventually can minimize operational cost for the residential operators.

We begin by we positioning our work with respect to related research Then we describe the dataset we use for this study. After that, we give some statistical characterization of Internet application usage, temporal distribution of traffic and periodicity of interactions that we observed in our households. Next, we introduce our technique to model recurrence pattern of Internet activities along with a series of analysis. We then present the segmentation of households based on our proposed method and discuss the characteristics of different segments. This is followed by the presentation of the activity prediction algorithm and its performance evaluation. Finally we offer our concluding remarks.

2 RELATED WORK

There are three aspects of our work : i) understanding residential Internet usage, ii) profiling households based on Internet activity characteristics and iii) predicting households future Internet activity. Hence, in this section, we look at the related research from these three perspectives.

2.1 Understanding Internet Usage

One of the earliest research on the analysis of web browsing activity is the work Catledge and Pitkow [5], who used both client and server side data to study user behavior and characterized user browsing patterns as serendipitous browsing, general browsing or searching. A number of follow-up studies examined users browsing trends, cross-site visit behavior and revisitation patterns [1, 20]. The HomeNet Field Trial placed computers with logging installed into 48 households in Pittsburgh in 1995 [14]. As households went online for the first time, researchers observed participants discovery of communication tools and found teenagers were some of the heaviest users and sources of expertise. Some previous research also investigated web browsing activities in the context of demographic distribution to understand the impact of race, culture, sex, and income on users online behavior [11, 13]. In [16], Kumar and Tomkins analyzed user browsing sessions based on one week of Yahoo! toolbar data to uncover several topical and temporal patterns of user activity. Some researchers studied the behavioral dynamics of home broadband users in the context of bandwidth and speed constrains. Their analysis suggested that home broadband users alter their behavior in response to unlimited and limited internet access [19] and persuasive visualization of their internet usage can effectively shift their web activity pattern [7]. In contrast, our analysis offers the opportunity to understand home Internet usage from a networking perspective by characterizing Internet usage and by identifying popular Internet activities and co-relation across them.

2.2 User Segmentation

Literature is rich on the subject of measuring user similarities based on different feature attributes and preforming segmentation of users accordingly. A well known technique used in recommendation systems is called *collaborative filtering* whose inherent assumption is that if users X and Y rate n items similarly, or have similar behavior (e.g., buying, watching, listening), then they will rate or act on other items similarly [24]. User similarity has also been investigated in social networks to recommend potential friends and content of interest. The combination of Pearson Correlation and Nearest Neighborhood algorithm atop the network specific features is the most commonly used technique to identify homophily across people [21]. In [25], Terveen et al. proposed *social matching*, a framework that aims to determine similarity mainly using the physical locations of people. In a similar work, Li et al. [17] presented a technique for calculating user similarity based on the location history. The prime differentiating factor of our work is that we measure similarity across households using interaction frequency and temporal regularity of their engagements with different Internet activities to derive usage profiles.More recently, Beauvisage reported on a 19 month logging study of 661 French households ending in 2006, that collected data primarily from a single household computer [4]. Beauvisage classified PC users into five distinct types based on applications used using Principal Component Analysis: Web-oriented Users (42% of his user population), Instant Messaging (14%), Gaming (11%), Multimedia (14%) and Serious (18%). These types correspond to five of the top six Internet activities we observed in our sample (see Table 1). In comparison, we take a network centric approach to have a more holistic view of Internet Activity, and address behavioral profiles on a household basis.

2.3 Prediction of Next Activity

One of the first algorithms for predicting next user action is IPAM [9]. It employed a first order Markov model, i.e., it based its predictions only on the last seen action. Gorniak and Poole argued that the last action does not provide enough information to predict the next action and proposed an on-demand prediction model called ONISI [12]. It used a k-Nearest Neighbors scheme that considers previous actions as well as the state of the system. However, ONISI searches for exact matches from the historical context and therefore hardly useful for realistic scenarios. In comparison to these methods, our proposed prediction algorithm considers multiple activity patterns from the past with highest similarity and temporal proximity to current day. In addition, the a novel popularity weight ensures that more relevant activities are given highest priority. In [26], a genetic algorithm based prediction technique was proposed that finds the most probable event sequences that indicate the expected occurrence of a specific event. Although the event prediction problem defined in this work is similar to the activity prediction problem considered in our work, the major difference is that we expect a set of multiple activities that will likely to occur in the near future as output. Furthermore, their algorithm is not capable of predicting the expected occurrence time of the predicted event. There are a number of other work in the context of human mobility prediction and trajectory analysis where different probabilistic models are examined [2, 3, 15]. Our algorithm is focused on the prediction of in-home Internet activities and hence mobility is not considered. However, we expect that our algorithm can easily be extended to predict future mobile activity patterns by analyzing mobile application usage at handheld terminals.

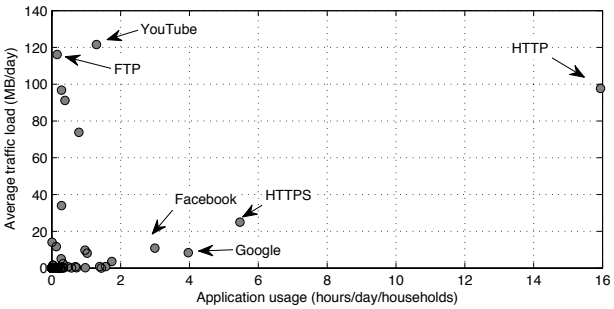

Figure 1: Daily Aggregated Traffic and Corresponding Application Usage Duration Averaged over all 86 Households for all 75 Applications

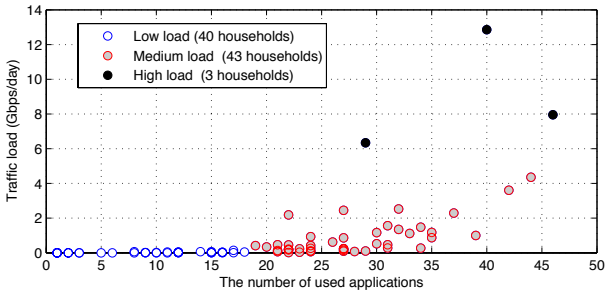

Figure 2: Aggregated Traffic and Corresponding Application Diversity for Clustered Households

3 DATASET

The dataset for this study has been collected from a Living Lab Project[1] based in the city of Kortrijk at Belgium. The project seeks to study users' experience with new fiber-based digital services especially for multimedia and health care. To do this, with the assistance of the city office of Kortrijk, Living Lab recruited 86 households that consented to having network packet inspection capabilities available on the backend service routers in exchange for free fiber optic Internet connection for two years including installation. The backend service routers monitor every single network packet and record application level information including protocol, Up/Down packet size and URL (protocol, domain name, and port number). This information is used to categorize network traffic into 75 applications and web portals (e.g., Skype, YouTube, Facebook, etc.) predefined by the router manufacturer. Due to privacy and legal concerns, our access to the data set is limited to the application (standalone or web based) or protocol name and corresponding Up/Down traffic (hourly aggregated) for each household. We collected these data on a daily basis from June 20, 2012 to August 19, 2012 that yielded 9,288,000 hourly data points for 86 households over 60 days.

Limitations of the dataset. Our dataset did not include precise session information of every application, e.g., start time, duration, etc. rather only provided hourly aggregated up/down traffic load. As a result, even if an application session spanned over a few minutes, it yielded an hourly entry in our dataset. Furthermore, we lacked information about the number of active devices connected to the Internet and the occupancy of the device owners. As such, there could be situations that some devices were active and generated traffic albeit absence or active engagement of members in the households.

3.1 Application Level Analysis

Figure 1 shows the daily aggregated traffic load and corresponding application usage duration averaged over all 86 households for all 75 applications. This plot indicates that the applications producing highest traffic loads are not necessarily used heavily. In fact most popular applications (Facebook, Google, etc.) in terms usage duration produce very low traffic. In contrast, applications with high traffic load (BitTorrent, YouTube, etc.) are not used regularly. One exception is the standard HTTP web traffic - it generates high traffic load with significantly high average usage time reaching approximately 16 hours per day. We have not filtered any traffic volume at this stage, thus this could reflect the traffic of passive activities e.g., notification from webmail, social networks etc. To understand households overall Internet activity patterns, we applied k−means algorithm [18] to segment households into three groups according to their aggregated web traffic. We then plot in Figure 2, their average daily traffic load against the number of different applications they used during the monitoring period. One interesting observation in this plot is that the heavy-weight households have high application diversity. Furthermore, the heaviest households (3 out of 86) generated majority of all traffic. These suggest that traffic load is distributed unequally across the households and hence this should not be used as an isolated property to determine similarity across them.

4 ACTIVITY LEVEL ANALYSIS

Our dataset comprises of 75 different applications (standalone, web based, and protocol specific). However, it is difficult to achieve an in-depth analysis of households online activities with such a high number of applications. In addition, many of the 75 applications offer semantically same functionalities, e.g., video watching, conferencing, social networking, etc. Henceforth, we have distributed these 75 applications into 8 distinct activity types as shown in Table 1 following to some degree the taxonomy suggested in [16]. However, this distribution is not absolute, and we acknowledge the possibilities of alternative distributions. For example, *Video Watching* and *Music Listening* can be put together as *Multimedia*, and so on. Another point of discussion is the distribution of HTTP/S traffic. In our dataset HTTP/S traffic is either partitioned into some distinct web portals, e.g., E-Bay, Facebook, etc., or as raw traffic without any portal label. We categorize the raw HTTP/S traffic as Web Browsing activity (including possible traffic from web mails), albeit several other activity groups have contributions from HTTP/S, e.g., Online Shopping, Social Networking, etc.

4.0.1 Activity Popularity Score. We discussed in the previous section that we have observed an inverse relationship between application usage frequency and corresponding traffic load, i.e.,

[1] http://www.leylab.be/english

popular applications generate low traffic whereas high traffic applications are rarely used. Taking this phenomenon into account, we argue here that interaction frequency (how many times an application is used) and temporal regularity (how often an application is used) can better characterize Internet activities of households than web traffic alone. Accordingly, we propose an *activity popularity score* - for user i and activity j, the corresponding popularity score is defined as

$$\gamma_{i,j} = \frac{1}{N_d} \sum_{x=1}^{n_{i,j}} 1 - \frac{\left| g_x - \overline{g}_{i,j} \right|}{N_d} \quad (1)$$

where, N_d is the total number of days during the monitoring period, $n_{i,j}$ is the number of days that a user is engaged with activity j, g_x is the day difference between $(x-1)^{th}$ usage day and x^{th} usage day, and $\overline{g}_{i,j}$ is the average usage gap defined as $\overline{g}_{i,j} = \frac{N_d}{n_{i,j}}$. The definition of $\gamma_{i,j}$ aims to reflect the interaction frequency and temporal regularity of an activity engagement. The popularity score is upper bounded by the number of usage days as $\gamma_{i,j} \leq \frac{n_{i,j}}{N_d}$. If g_x is very close to $\overline{g}_{i,j}$ indicating a strong temporal regularity, then $\gamma_{i,j}$ approaches its upper limit. For a practical example, consider two users *Alice* and *Bob*. They both engage with *Online Shopping* activity for n_x days. *Alice's* engagement follows a temporal pattern, e.g., she visits *E-Bay* every saturday, however *Bob* visits *E-Bay* in n_x consecutive days and then he stops visiting the site completely. In this case, *Alice's* popularity score for *Online Shopping* will be higher than that of *Bob's* using equation (1).

4.0.2 Activitiy Pattern Analysis. We discussed earlier that HTTP traffic (i.e., Web Browsing activity) was the most popular in our usage logs, and in fact had the highest interaction frequencies for all the households. At the same time, File Sharing activity and corresponding applications were very bursty in nature yet producing heaviest traffic load. Our aim is to model household Internet usage and examine recurrence patterns of activities. Due to the very nature of how these activities are performed in the home - engagement with Web Browsing all the time and with File Sharing very rarely, we have eliminated these two activities from our analysis as a feature engineering step. To justify our argument, in Figure 3, we plot the autocorrelation of the aggregated hourly traffic load for all households for all applications during the entire test period. When all applications are considered the correlogram does not show a periodic pattern. However, if we disregard the Web Browsing and File Sharing applications, and consider the rest six activities, the correlogram exhibits a periodic pattern. The high peaks are the integer multiples of 24 hours approximately, which match human daily activity routine. This observation indicates that households interact with most of the Internet applications following a temporal pattern. To gain further insights, we plot each of these six activities against their popularity score calculated using equation (1) in Figure 4. The result shows that Social Networking activity is the most popular activity collectively in our 86 households followed by Video Watching, Home Working, Web Communication, Online Shopping and Online Gaming respectively. Henceforth, we argue that activity patterns of these popular and bandwidth friendly activities with strong routine behavior can be used to characterize households Internet activities and to determine similarities across them. In a later section, we present a technique based on this premise to characterize and to find similarities across households online behavior.

Table 1: Top 6 Internet Activities and Corresponding Applications

ID	Activity	Applications and Protocols
1	Web Communication	POP3, IMAP, SMTP, MS Exchange, Domino, Skype, SIP, Betamax VoIP, Google Talk, RTP, XMPP, MSN Messenger, Asterisk, RTSP, TeamSpeak, WebEx, IRC, OoVoo
2	Social Networking	Facebook, Twitter, Google+, MySpace, Flickr, Photobucket
3	Online Gaming	Steam, World of Warcraft, XboxLIVE
4	Home Working	Teredo, TLS, GRE, Citrix ICA, SSH, Telnet, Remote Desktop, LDAP, Citrix IMA, IP Printing
5	Online Shopping	Amazon, EBay
6	Video Watching	YouTube, HTTP Video, RTMP Streaming, Shockwave Flash, SHOUTcast, Real Player, BBC iPlayer, PPTV
7	Web Browsing	HTTP, HTTPS, Google, MSN,Yahoo, Bing, Google Earth, Google Maps
8	File Sharing	BitTorrent, Gnutella, Ares, Kontiki, EMule, Tor, FTP, DepositFiles, RapidShare, Uploading.com, MediaFire, MegaVideo, MegaUpload, SendSpace, EasyShare

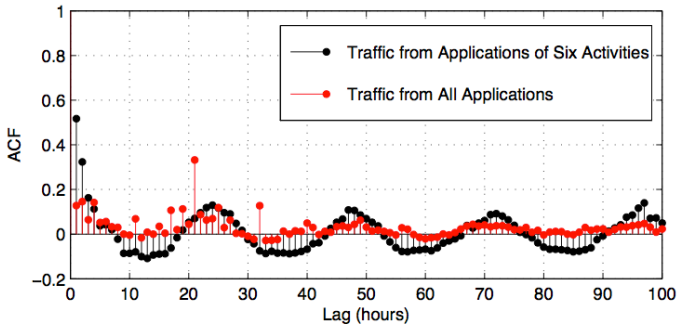

Figure 3: The Autocorrelation (Correlogram) of Hourly Traffic Time Series for All 8 Activities and 6 Activities excluding Web Browsing and File Sharing.

4.0.3 Correlation of Activities. We are often engaged with multiple Internet activities simultaneously. For instance, one might play Online Games while using Web Conferencing applications, e.g., Skype to chat with his/her gaming partners, or may prefer using Facebook while browsing through online shops. In this section, we examine such correlation across Internet activities in our households, i.e., we try to answer which Internet activities are often performed in tandem.

For a household x, we use $L_x^i(t)$ to represents the time series of the traffic load of an activity i with t expressed in hours, $i \in [1, N_a]$ and N_a is the number of activities. We use A_x^i to represent the

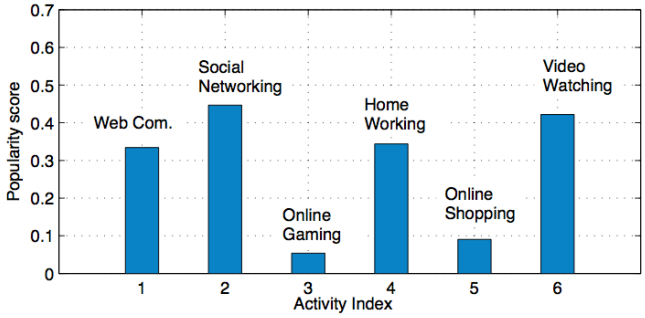

Figure 4: Popularity of Different Activities

set of hourly time slots in which $L_x^i(t) > 0$, i.e., at t^{th} time slot, a household x is engaged with i^{th} activity that produced non zero traffic. We use parameter v_x^{ij} to express the dependency coefficient between activity i and j, i.e., it represents the probability for activity i to happen when activity j is present. Hence, $v_x^{ij} = \frac{\left\|A_x^i \cap A_x^j\right\|}{\|A_x^i\|}$, where $\|A\|$ represents the cardinality of set A. The correlation between activity i and j is defined as the ratio of the number of households whose dependency coefficients are higher than a threshold ν_{th} to the total number of households who were engaged with activity i and j during the testing period and it is defined as:

$$c_{ij} = \frac{\sum_{x=1}^{N_u} I(v_x^{ij} > \nu_{th})}{\sum_{x=1}^{N_u} I(\left\|A_x^i\right\| > 0 \& \left\|A_x^j\right\| > 0)} \quad (2)$$

where, $I(r)$ is the indicator function, and $I(r) = 1$ if r is true or 0 otherwise. N_u is the total number of households. If c_{ij} is higher than a pre-defined threshold c_{th}, we consider engagement with activity i is correlated with activity j. We use threshold $\nu_{th} = 0.75$ and $c_{th} = 0.5$ to plot the correlations across the six activities as shown in Figure 5. If activity i is correlated with j, an arrow is drawn from i to j. This implies that activity i happens when activity j is present, however the reverse may not be true. As shown in this figure, most of the activities are highly correlated with Social Networking and Video Watching respectively. This observation is actually simple to interpret - as these activities are the most popular ones as shown in Figure 4. One further interesting observation of this correlation topology is the arrow from *Online Games* → *Conferencing*, which suggests that families often tend to use web conferencing tools while playing online games. Naturally, we did find any correlation between Home Working and other activities.

5 SEGMENTATION OF HOUSEHOLDS

We mentioned previously that households can be segmented based on their aggregated traffic footprint. However, such segmentation does not offer deep insights on behavioral characteristics. In the earlier sections we showed that how popularity score can elegantly address different attributes of interaction patterns, e.g., temporal regularities and frequencies of different activities, which in turn enables us to gain a better understanding of overall web activities

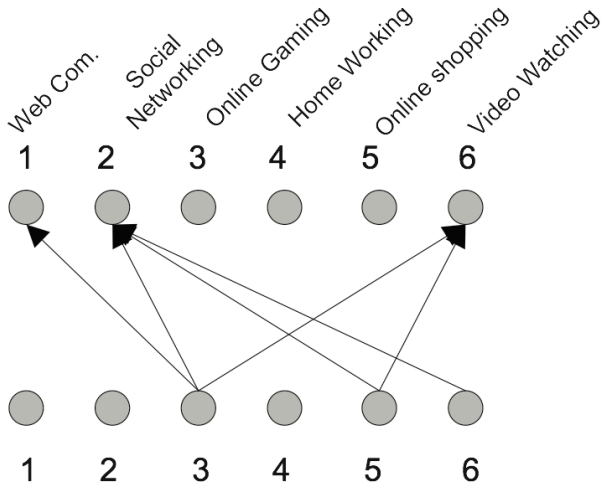

Figure 5: Correlation Topology of the six activities with $\nu_{th} = 0.75$, $c_{th} = 0.5$ showing which activities occur simultaneously

of a household. Hence in the following, we discuss a segmentation technique grounded upon activity popularity score and then highlight different behavioral profiles of these segments.

In the first step of segmentation, we build a $N_u \times N_a$ activity popularity matrix Q. Each element of Q is γ_{ij}, the popularity score of activity j for household i, $i \in [1, N_u]$ and $j \in [1, N_a]$. In the second step each of these popularity scores is multiplied by a constant $k = 10$ and passed through a floor function to obtain an integer grade which is within the range $[0, k-1]$. This grade is represented by $\gamma_{ij}^{'} = \lfloor \gamma_{ij} \times k \rfloor$. The popularity score has a range $[0, 1]$. Thus if two households have popularity scores of 0.05 and 0.95 for Online Shopping activity respectively, this transformation assigns grade 0 and 9 as their popularity scores for the respective activity. Hence, it helps us to increase the discrimination granularity of activity popularity score across households to some degree for clustering purpose. As we discussed earlier, Web Browsing and File Sharing activities are eliminated from our analysis as either they are performed almost all the time or very rarely. Hence using them as discriminating features would result in misleading segmentation. Therefore, we disregard these two activities and consider the remaining six activities as the main features for segmentation purpose and represent them as $\varpi = [1, 2, 3, 4, 5, 6]$. Hence for a household i, popularity scores (represented by grades as computed in step two) for these activity groups are denoted as $\epsilon_i = [\gamma_{i1}^{'}, \gamma_{i2}^{'}, \gamma_{i3}^{'}, \gamma_{i4}^{'}, \gamma_{i5}^{'}, \gamma_{i6}^{'}]$. In the fourth step, we do another transformation, and represent these popularity grades in a 6-bit binary feature vector κ_i. The x^{th} element of $\kappa_i(x)$ is set to 1 if the corresponding popularity grade $\gamma_{i\varpi(x)}^{'} > 0$, otherwise $\kappa_i(x) = 0$. For instance, κ_i = [1 1 1 0 01] represents the feature vector of a household that engages with activities 1, 2, 3 and 6, i.e., Web Communication, Social Networking, Video Watching and Home Working respectively. Once we have generated the feature vectors κ for all the households, we move to the final step of our segmentation technique and apply DBSCAN [10] algorithm

Step 1 : Construct an Activity Popularity Vector for each households.

Step 2 : Multiply each element of the vector with a constant K and pass through a floor function to increase discrimination granularity.

Step 3 : Select X activities based on popularity score, eliminating those activities that are non discriminant (e.g., occurs always or very rarely) **[Optional, case by case]**

Step 4 : Construct a Binary Feature Vector for the selected activities, each bit is set to 1 if popularity score γ' is higher than a threshold.

Step 5 : Run DBSCAN on these binary feature vectors to obtain segments.

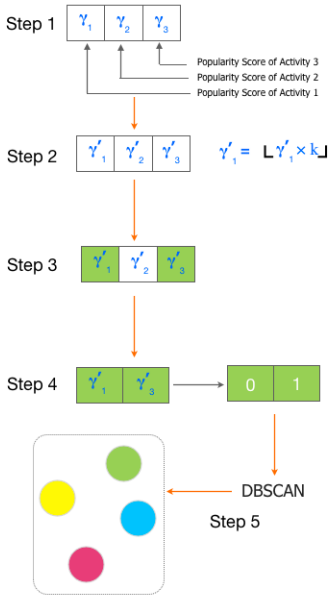

Figure 6: A Step-by-StepVisual Desctiption of the Segmentation Technique used to Segment the Households.

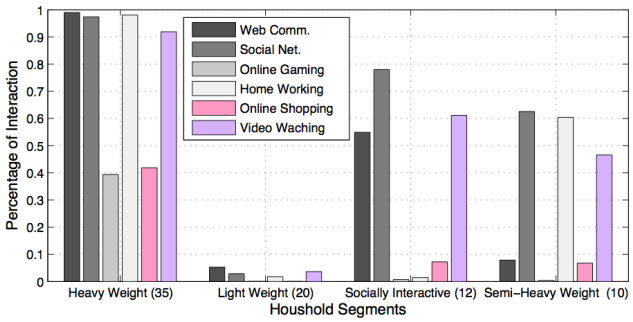

Figure 7: Activity Interaction Percentage Distribution for Six Activities in Different User Segments

to cluster the households. Figure 6 shows a visual representation of these steps.

Our segmentation technique obtained fours segments - Heavy Weight Households, Light Weight Households, Socially Interactive Households and Semi-Heavy Weight Households. Out of the N_u = 86 households, 9 households (10% of all the households) do not belong to any of these segments. Figure 7 shows the proportions of the selected six activities that each household segments were engaged during the monitoring period. In the following we further scrutinize these segments.

(1) Heavy Weight Households (HWH): This segment is represented by 35 out of 86 households (41% of all the households). As shown in the Figure 7, households of this segments are engaged with all six activities heavily comparing to other segments. For Web Communication, Social Networking, Home Working and Video Watching, the interaction proportions are over 90% meaning, there were interactions from one or more households of this segment 90% of times during our monitoring period for these activities. To further verify, we calculated the households similarity with Pearson Correlation coefficient using ϵ_i as the input for all the households within this group. The average similarity among all the 86 households is 0.2261, where as among the households of this group, the similarity is 0.83. Henceforth, we argue that they exhibit similar behavioral profiles.

(2) Light Weight Households (LWH) : 20 households represent this segment (23% of all the households) and these households had minimum number of interactions across all six activities (below 10%) as shown in Figure 7. The value of similarity coefficient for this segment is 0.97 which clearly suggest that these households share similar online usage characteristics. We concur that these households represent the population that either only engages with Web Browsing activity that is not considered in our segmentation feature or represent the extreme end of Internet population with very little online footprint.

(3) Socially Interactive Households (SIH): This segment is represented by 12 households (14% of all the households). As depicted in Figure 7, households of this segment mostly engage with Web Communication, Social Networking and Online Video Watching activities. Hence, we label them as the socially interactive group. The similarity coefficient for this segment is 0.87 indicating high similarity across representative households' online behavior.

(4) Semi-Heavy Weight Households (SHWH): Finally, this segment is represented by 10 households (12% of all the households). Looking at the Figure 7, we observe that this group is very similar to Socially Interactive group, however with one distinct activity feature, i.e., Home Working with an interaction proportion of 60% for this activity. Our segmentation

technique could determine this difference across the households representing these two segments, and could split them properly. With a similarity coefficient of 0.78, these households represent the Internet Population, who are working adults with significant exposure to social activities.

6 PREDICTION OF IN-HOME INTERNET ACTIVITY

In the earlier section, we have discussed a segmentation method based on households engagement regularity in different Internet activities. In this section we present an activity prediction algorithm grounded upon identical premise, i.e., regularity in the occurrence of an activity. The objective of this algorithm is to predict which set of activities a household will be engaged in the following hours by matching the activity pattern from the previous hours against historical activity patterns.

We mentioned earlier that individual household shows variable activity characteristics, reflecting their lifestyle, daily routines and activity preferences. It is not trivial to identify features to develop a global activity model for prediction. Henceforth, the basic working principle of the prediction algorithm is to operate on an individual household basis, i.e., prediction outcome solely depends on historical activities of the household in context. In the rest of this section, we present the prediction algorithm followed by a discussion on its performance evaluation.

6.1 Prediction Algorithm: Model and Strategy

The algorithm predicts activity patterns of future hour slots of current day by matching patterns of similar days in the past N_l days, and N_l is defined as the lookup days in our algorithm. We use a $24 \times N_l$ matrix U to represents household activities within lookup days. The element of matrix U, denoted as an activity vector u_{ij}, is a N_a-bit binary vector, and represents the engagement pattern of N_a activities. u_{ij}^k is the k^{th} bit of u_{ij}, and it is set to 1 if there is any engagement with k^{th} activity at i^{th} hour slot on j^{th} day or 0 otherwise. We denote the current day as j_c and the current hour as i_c. Hence the current activity vector is $u_{i_c j_c}$. As a day progresses, activity vector for each hour is constructed from midnight up to the current time. To predict an activity pattern of a future hour slot, these activity vectors of the current day or part of it are used against the corresponding parts of the past N_l days. We define a look up window with a size L_h hours. To predict the activity vector at h^{th} hour after current time, we compare the past L_h hours of current day with the corresponding hours of previous N_l days and select M top most similar days that provide the basis for prediction, where $M < N_l$.

Searching for similar past days in this context essentially is a case for binary similarity measures [8]. We have examined several binary similarity measures with our dataset by dividing our samples into subsets randomly. We have found that Sokal-Michener measure [23] offers the best discrimination capability for our case as it gives equal weight to presents and absence of an activity in context. However, as we have discussed in earlier sections that households engage with a subset of activities heavily, resulting in variable popularity scores for different activities. Taking this observation into account, we argue that the activities that are popular should contribute more to the measure than those that are less popular. Henceforth, to further improve discrimination capability, we have applied activity popularity score γ (defined by equation (1)) as weight to the corresponding activity bit of the vectors while comparing similarity. Accordingly, to compare the similarity between two hourly activity vectors x and y, we define a *hour similarity score* as

$$HS = \left(\sum_{k=1}^{N_a} \gamma_k\right)^{-1} \sum_{k=1}^{N_a} \gamma_k I(x_k^t y_k || \overline{x}_k^t \overline{y}_k) \quad (3)$$

where, $I(r)$ is the indicator function, and $I(r) = 1$ if r is true or 0 otherwise, $x_k^t y_k$ denotes the positive match and $\overline{x}_k^t \overline{y}_k$ denotes the negative match at k^{th} position between x and y, and γ_k is the popularity score of the activity represented by the k^{th} bit. Using equation (3), the prediction algorithm first computes the hour similarity scores between past L_h hours of current day and the corresponding hours of last N_l days and assign for each day a *day similarity score* which is simply the mean of hour similarity scores

Algorithm 1: The Activity Prediction Algorithm

Input: Lookup Days N_l, Current Day j_c, Current Hour i_c, Number of Candidate Days M, Lookup Window L_h, Prediction Hour $i_f = i_c + h$, Threshold p_{th}
Output: Activity Vector $u_{i_f j_c}$

```
Initialise the Matrix U and its element hourly activity vectors u_ij, i ∈ [0, 23] for N_l days
Initialise a Column Matrix U_c for current day j_c, and its element hourly activity vectors u_ij_c, i ∈ [0, i_c]
for j = 0 to Lookup Days (N_l − 1) do
    for i = i_c to i_c − (L_h − 1) do
        Compute the hour similarity score HS_ij between u_ij and u_ij_c using equation (3)
    end
    Compute the day similarity score DS_j using eqation (4)
end
Sort N_l Lookup Days first based on DS_j and then on Time Difference from current day and select top M Days
u_if_jc ← 0;
foreach activity bit k in N_a do
    Look at the k^th bit of each activity vector u^k_if_j of the M selected days, j ∈ [1, M] and i_f is the prediction hour.
    Compute weighted occurrence probability p combining u^k_if_j with corresponding DS_j using eqation (5)
    if p ≥ p_th then
        u^k_if_jc = 1;
    else
        u^k_if_jc = 0;
    end
end
return u_if_jc
```

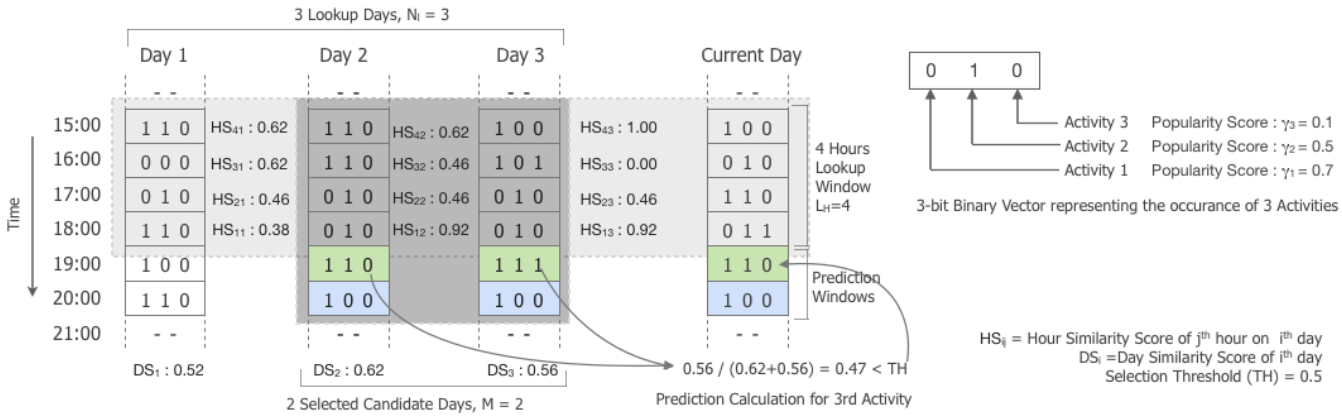

Figure 8: A simplified visual explanation of the prediction algorithm for 3 activities with $N_l = 3$ days, $M = 2$ days, $L_h = 4$ hours, and $p_{th} = 0.5$. First 4 hour similarity scores are calculated for each 3 lookup days against current day, and then 2 candidate days are selected based on highest day similarity scores and temporal proximities to current day. Finally, a prediction is made by combining day similarity scores of each selected days for each activity bit and by comparing it to the selection threshold of 0.5. For example, 3^{rd} activity bit at prediction hour is set to 0 as the combination of the day similarity scores of candidate days for this bit is less than the selection threshold.

of that day. Hence for j^{th} day, the day similarity score is defined as

$$DS_j = \frac{1}{L_h} \sum_{i=1}^{L_h} HS_{ij} \tag{4}$$

Once the day similarity scores are obtained, the algorithm moves to the selection of the M candidate days from the nearest past by picking the days that have highest day similarity scores and are closest to the current day. This selection is performed by sorting N_l past days twice, first on the day similarity score and sorting on the time difference from the current day. Based on our experiments $M = 10$ with $N_l = 30$ and $L_h = 6$ are found to be good choices for high prediction accuracy.

The final step of the prediction algorithm is to consider the activity vector of each candidate day for the target hour slot, and compute the weighted probability of occurrence of each activity. If the probability is higher than a selection threshold p_{th} then that activity bit is set to 1 or 0 otherwise. After ROC analysis, we set p_{th} to 0.44. The algorithm performs this step by taking each activity bit at a time and combining day similarity score to ensure that most similar and more recent days have highest contribution in predicting the occurrence of an activity. Hence, the k^{th} activity bit of the vector at future hour slot i_f (here $i_f = i_c + h$) for the current day is predicted as

$$u^k_{i_f j_c} = \begin{cases} 1 & \text{if } \left(\sum_{j=1}^{M} DS_j\right)^{-1} \sum_{j=1}^{M} DS_j I(u^k_{i_f j} = 1) > p_{th} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where, $I(r)$ is the indicator function, and $I(r) = 1$ if r is true or 0 otherwise. Here r indicates whether k^{th} bit of i_f^{th} hour of j^{th} day is 1 or not. When the prediction hour slot i_f is at the beginning or end of a day, i.e. before midnight or just after midnight, the lookup hour slots for similarity matching are determined from the immediate previous day, and candidate hour slots for prediction are selected from the immediate next day. This avoids complexities in making predictions that span midnight. Figure 8 provides a simplified visual explanation of the algorithm which is pseudo codified in Algorithm 1.

Table 2: Prediction Performance on Different Household Segments

Segment	Accuracy		Precision		Recall		F-Measure	
	$\overline{x}$	σ_x	$\overline{x}$	σ_x	$\overline{x}$	σ_x	$\overline{x}$	σ_x
HWH	0.87	0.11	0.80	0.13	0.75	0.03	0.78	0.07
LWH	0.93	0.08	0.94	0.02	0.76	0.10	0.84	0.10
SIH	0.83	0.10	0.78	0.08	0.79	0.03	0.79	0.03
SHWH	0.85	0.07	0.80	0.05	0.77	0.01	0.78	0.01

6.2 Performance Evaluation

Predicting an activity pattern for a future hour slot is essentially a multi-label classification problem. Henceforth, the performance of the algorithm can be evaluated by standard Information Retrieval measures for a multi-label classification setting. For each hour slot i_f, let T be the true set of activities, and S be the predicted set of activities. Accuracy is measured by the Hamming Score which symmetrically measures how close T is to S, i.e., $Accuracy(i_f) = \frac{\|T \cap S\|}{\|T \cup S\|}$. Precision ($P$), Recall ($R$) and F-Measure ($F_1$) are defined as $P(i_f) = \frac{\|T \cap S\|}{\|S\|}$, $R(i_f) = \frac{\|T \cap S\|}{\|T\|}$ and $F_1(i_f) = \frac{2P(i_f)R(i_f)}{P(i_f)+R(i_f)}$.

For evaluating the algorithm, we split 60 days of data into two parts. The data of first 45 days are used to train the algorithm, e.g., as histories of activities and the data of remaining 15 days are used to evaluate the performance of the algorithm.

Figure 9 plots the cumulative distribution of F-Measure across all 86 households for the selected six activities. As we observe,

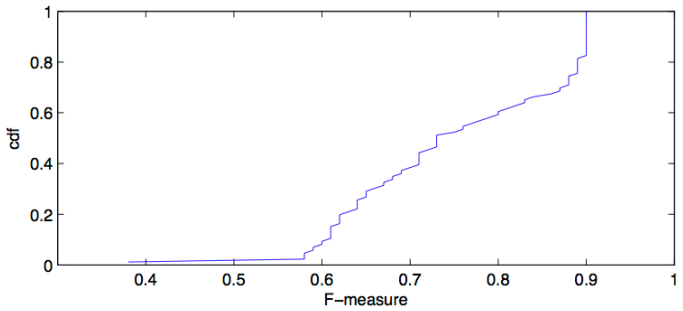

Figure 9: Cumulative Distribution of Prediction Performance over all 86 Household for the Six Selected Activites

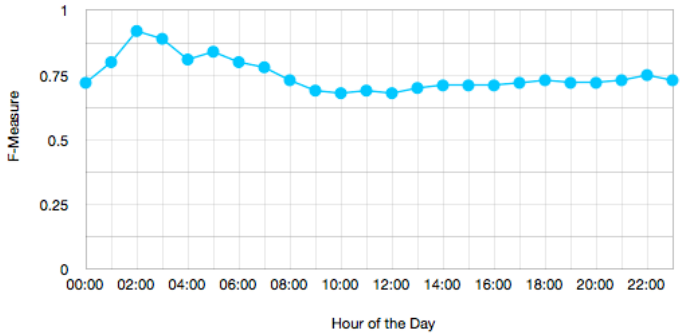

Figure 10: Prediction Performance at Different Hours of the Day

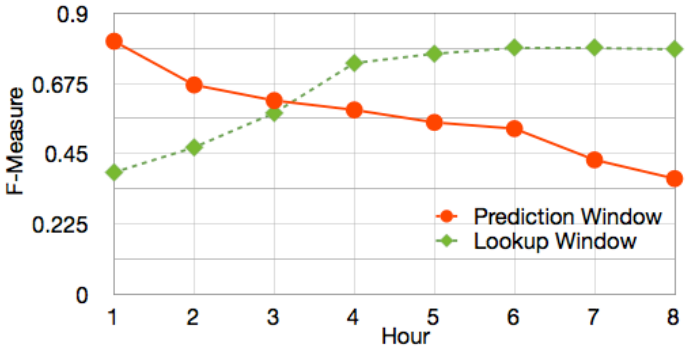

Figure 11: Impact of Lookup Hours on Prediction Performance

at least 60% households have over 0.7 F-Measure, which we consider is reasonably high. However, we have discussed in the earlier sections that a subset of activities are very popular across some households, e.g., Social Networking etc., where as some activities are occasionally performed, e.g., Online Shopping, etc.The segmentation method presented earlier addresses these issues elegantly to group households with identical online profiles based on these observations. To understand the impact of such behavioral characteristics on prediction, Table II shows the prediction performance on each household segments presented earlier. We observe that the prediction performance is consistent across all the different household segments. One interesting observation is the performance on Light Weight Household segment, as one might expect that the low activity interaction rate of these households might impact prediction performance. However, we did not notice any significant difference in the performance, suggesting the fact there were enough predictive elements in the data, ideally because of the strong routine behavior representative households exhibit.

Naturally, the prediction performance varied with time of the day as shown in Figure 10. F-Measure is highest during night times, and drops during the day when households activities are less predictable. Taking this observation into account, we have eliminated night predictions from our evaluation presented above, by starting prediction at 5 O'Clock in the morning and continuing until midnight. Figure 11 shows the prediction performance over varying lookup window in the past and in the future. As one might expect, prediction accuracy drops as we look further into the future. Past lookup window also impacts the prediction performance as increasing lookup window gradually improves prediction performance. However, looking back beyond 6 hours does not contribute to prediction performance implying the fact that the immediate past hours offer best hints of near future activities.

7 CONCLUDING REMARKS

In this paper, we present a study on in-home Internet activities of 86 households for 2 months. We show that despite individual lifestyles and behavioral dynamics, families follow a temporal recurrent pattern in their Internet activities. This observation in turn enables us to segment households with similar behavioral profiles. We present a novel technique to model this homophily across families using interaction frequency and temporal regularity of their web activities. We also present a new algorithm to predict future online activity patterns by considering historical patterns from similar days. Our algorithm shows that 60% of the households online activities can be predicted correctly 70% of times. Both the proposed techniques are generic enough and can easily be applied in other contexts (e.g., analysis of mobile web activity, social network activity, etc.) to characterize and model users behavior.

REFERENCES

[1] E. Adar, J. Teevan, and S. T. Dumais. Resonance on the Web: Web Dynamics and Revisitation Patterns. *27th international conference on Human factors in computing systems (CHI '09)*, pages 1381–1390, 2009.
[2] A.Monreale, F.Pinelli, and R.Trasarti. WhereNext: A Location Predictor on Trajectory Pattern Mining. *5th Intl. Conference on Knowledge Discovery and Data Mining (KDD-09)*, 2009.
[3] D. Ashbrook and T. Starner. Using GPS to Learn Significant Locations and Predict Movement across Multiple Users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.
[4] T. Beauvisage. Computer Usage in Daily Life. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, pages 575–584, New York, New York, USA, Apr. 2009. ACM Press.
[5] L. D. Catledge and J. E. Pitkow. Characterizing Browsing Strategies in the World Wide Web. *Computer Networks ISDN Systems*, 27(6), 1995.
[6] J. Chang, I. Rosenn, and L. Backstrom. ePluribus: Ethnicity on social networks. *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.
[7] M. Chetty, D. Haslem, A. Baird, U. Ofoha, B. Sumner, and R. E. Grinter. Why is My Internet Slow?: Making Network Speeds Visible. *2011 Annual Conference on Human Factors in Computing Systems (CHI 2011)*, pages 1889–1898, 2011.

[8] S. Choi, S. Cha, and C. C. Tappert. A Survey of Binary Similarity and Distance Measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48, 2010.
[9] B. D. Davison and H. Hirsh. Predicting Sequences of User Actions. *AAAI/ICML Workshop on Predicting the Future: AI Approaches to Time– Series Analysis*, 1998.
[10] M. Ester, H. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Second International Conference on Knowledge Discovery and Data Mining*, 1996.
[11] S. Goel, J. M. Hofman, and M. I. Sirer. Who Does What on the Web: A Large-Scale Study of Browsing Behavior. *Sixth International AAAI Conference on Weblogs and Social Media*, 2010.
[12] P. Gorniak and D. Poole. Predicting Future User Actions by Observing Unmodified Applications. *Conference of the American Association for Artificial Intelligence*, 2000.
[13] E. Hargittai. Whose Space? Differences Among Users and Non-users of Social Network Sites. *Journal of Computer- Mediated Communication*, 13(1), 2007.
[14] R. Kraut, W. Scherlis, T. Mukhopadhyay, J. Manning, and S. Kiesler. The HomeNet Field Trial of Residential Internet Services. *Communications of the ACM*, 39(12):55–63, Dec. 1996.
[15] J. Krumm and E. Horvitz. Predestination: Inferring destinations from Partial Trajectories. *Eighth International Conference on Ubiquitous Computing (UbiComp 2006)*, 2006.
[16] R. Kumar and A. Tomkins. A Characterization of Online Browsing Behavior. *19th International Conference on World Wide Web (WWW)*, pages 561–570, 2010.
[17] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining User Similarity based on Location History. *16th ACM SIGSPATIAL international conference on Advances in geographic information systems (GIS '08)*, 2008.
[18] J. MacQueen. Some Methods for Classification and Analysis of MultiVariate Observations. *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.
[19] C. Middleton and S. Chang. The Adoption of Broadband Internet In Australia and Canada. *Handbook of Research on Global Diffusion of Broadband Data Transmission*, pages 818–840, 2008.
[20] A. L. Montgomery and C. Faloutsos. Identifying Web Browsing Trends and Patterns. *Computer*, 34(7):94095, 2001.
[21] P. Singla and M. Richardson. Yes, There is a Correlation: - From Social Networks to Personal Behavior on the Web. *17th international conference on World Wide Web (WWW '08)*, 2008.
[22] A. Smith. Home Broadband Adoption 2010: Summary of Findings. *Pew Internet & American Life Project*, 2010.
[23] R. R. Sokal and C. D. Michener. A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Scientific Bulletin*, 38:1409–1438, 1958.
[24] X. Su and T. M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 2009(4), 2009.
[25] L. G. Terveen and D. W. McDonald. Social Matching: A Framework and Research Agenda. *ACM Transactions on Computer-Human Interaction*, 13(3):401–434, 2005.
[26] G. M. Weiss and H. Hirsh. Learning to Predict Rare Events in Categorical Time-Series data. *Predicting the future: AI approaches to time-series problems; Workshop in conjunction with the fifteenth national conference on artificial intelligence*, 1998.