Life in the *Matrix**: Human Mobility Patterns in the Cyber Space

Tianran Hu, Jiebo Luo
Department of Computer Science
University of Rochester
{thu, jluo}@cs.rochester.edu

Wei Liu
AI Lab
Tencent Inc.
vincentwliu@tencent.com

Abstract

With the wide adoption of the multi-community setting in many popular social media platforms, the increasing user engagements across multiple online communities warrant research attention. In this paper, we introduce a novel analogy between the movements in the cyber space and the physical space. This analogy implies a new way of studying human online activities by modelling the activities across online communities in a similar fashion as the movements among locations. First, we quantitatively validate the analogy by comparing several important properties of human online activities and physical movements. Our experiments reveal striking similarities between the cyber space and the physical space. Next, inspired by the established methodology on human mobility in the physical space, we propose a framework to study human "mobility" across online platforms. We discover three interesting patterns of user engagements in online communities. Furthermore, our experiments indicate that people with different mobility patterns also exhibit divergent preferences to online communities. This work not only attempts to achieve a better understanding of human online activities, but also intends to open a promising research direction with rich implications and applications.

Introduction

Understanding human activities in online communities not only is the key to computational sociology research (Ren, Kraut, and Kiesler 2007; Zhu, Kraut, and Kittur 2014), but also can offer valuable guidance to the design of online systems (Kraut et al. 2012). Many popular platforms, such as Reddit, 4chan, and StackExchange, adopt the setting of multiple communities for user engagement. Much work has been done on human activities across communities, including user exploration and participation patterns in more than one community (Tan and Lee 2015; Zhang et al. 2017), and user loyalty under the multi-community setting (Hamilton et al. 2017). However, these studies usually focus on specific aspects of online activities, while a comprehensive and general-purpose framework for studying multi-community activities is still lacking. To overcome the problem, we introduce a novel analogy between human online activities and offline physical movements in this paper. Given the rich body of research on human mobility patterns (Barbosa-Filho et al. 2017), such an analogy allows us to borrow the existing approaches and frameworks for analyzing movements in the physical space and apply them to online scenarios with necessary adaptation. In this work, we first validate this analogy, and design experiments to reveal striking quantitative similarities between human online activities and offline movements. We then propose a frameworks for studying users of online platforms, and uncover interesting activity patterns across communities.

We draw a strong analogy between the cyber space (i.e. online multi-community platforms such as Reddit)[1] and physical space. The communities on the online platforms is then treated as the *locations* in the cyber space. The intuition of the analogy arises from two aspects:

- The activities in the cyber space resemble the movements in the physical space – people move from one community to another on the platforms, explore new communities, and regularly visit communities with which they are familiar, as do they with physical locations.
- The locations in the two spaces (i.e. communities and places) share important similarities – some locations are popular and thus gain large amounts of visitors; some are niche and attract specific groups of visitors; some are private where only authorized visitors have access (e.g. home and private subreddit).

Without causing ambiguity, in this paper we may use the *movements* across locations in the cyber space to refer to the activities across online communities.

We quantitatively validate our analogy by comparing the properties of movements in the cyber and physical spaces from three representative and progressive aspects: 1) at a coarse granularity, we first compare the overall visit distributions in the two spaces; 2) at a fine granularity, we then study the visit behavior of individuals; and 3) considering time factors, we further investigate the temporal properties of both cyber and physical movements. The results reveal striking similarities between the movements in the two spaces. For example, the number of visits to an online community is

*A metaphor taken from the hit movie *Matrix*.

[1]We use "cyber space" and "online multi-community platforms" interchangeably in this paper for the simplicity of narration. Please note that "cyber space", as sometimes used as a metaphor to the whole Internet, is a wider concept than online platforms.

found to fall into the same distribution of the number visits to a physical location. Moreover, it is shown that individuals visit locations in both spaces following random-walk behavior. More strikingly, the Zipf's law of the movements in the physical space also applies to the movements in the cyber space. In terms of temporal property, we observe an almost identical returning pattern in both spaces – people tend to return to specific communities/locations on a daily basis. Also, a temporally complementary relation between the two spaces is uncovered – human mobility level is high in the cyber space when the level is low in the physical space, and vice versa.

The analogy implies a possibility of modelling the movements in the cyber space using the existing approaches for studying physical movements. However, one of the differences between the two spaces is that distance is not universally defined in the cyber space (Hessel, Tan, and Lee 2016; Pavalanathan et al. 2017). Instead of arbitrarily defining the distance concept, we focus on entropy-based approaches that do not require an explicit measurement of distance. Entropy and its variants have been widely used for analyzing the randomness and predictability of individual mobility (Song et al. 2010b), measuring diversity of visitors to a location (Cranshaw et al. 2010), and so on. Following this line of research, we study human mobility in the cyber space.

We first study the randomness of human mobility in the cyber space. We discover that mobility randomness varies significantly across people. Some people evenly distribute their visits to many communities, and therefore exhibit high mobility randomness. In contrast, some people only visit a very limited number of communities, and exhibit very low mobility randomness. We then investigate if human mobility in the cyber space evolves over time. By comparing different stages of a user's online lifespan, our experiments reveal three interesting mobility patterns in the cyber space: 1) concentrating on a limited number of communities throughout the whole lifespan; 2) exploring many communities in the early stage, but stopping exploring and concentrating on only a few later on; and 3) focusing on a few communities in the early stage, and starting to explore more and more communities later on. Furthermore, we observe that people of the concentrated pattern are more likely to visit communities of specific topics and smaller sizes, while people of both two exploratory patterns prefer communities of general topics and larger sizes.

The main contributions of this paper are threefold:

- We introduce a novel analogy between the cyber space and the physical space. This work is intended to first achieve a better understanding of human online activities, and more importantly, initialize a promising new research direction.
- We quantitatively examine the validity of the analogy between the cyber space and the physical space. Our experiments reveal striking similarities between the two spaces.
- We investigate the individual mobility patterns in the cyber space. Our results reveal three major patterns of online activities, as well as the different preferences to online communities by the people of the three patterns.

Related Work

Online Communities

Rotman et al. suggest that users interact and share purpose in online groups, and therefore, form sociological communities (Rotman, Golbeck, and Preece 2009). Since then there has been a rich literature studying online communities from various aspects, such as the styles and evolution of communities (Tran and Ostendorf 2016; Lin et al. 2017), differences between communities (Hessel, Tan, and Lee 2016), and so on. The interaction between users and communities has also been investigated. For example, (Danescu-Niculescu-Mizil et al. 2013) studies the changes of linguistic style in online communities, and uncovers the interesting evolution of the interaction patterns between users and communities.

Given the wide adoption of the multi-community setting on online platforms, more and more attention has been paid to human activities across online communities (Chan, Hayes, and Daly 2010; Pavlick and Tetreault 2016). Zhu et al. report that user participation in multiple communities benefits the survival of communities (Zhu, Kraut, and Kittur 2014). Ren et al. study the common identities within different communities, and suggest the "bonds" between online communities (Ren, Kraut, and Kiesler 2007). Furthermore, it is reported that community characteristics can affect the user engagement pattern in a community. (Hamilton et al. 2017). For example, a community with a distinctive and dynamic identity is not only more likely to retain users, but also creates a larger "cultural" gap between senior members and newcomers (Zhang et al. 2017). Tan et al. study the involvements of users in more than one communities (Tan and Lee 2015). This work suggests that instead of gradually settling down in previously visited communities, online users keep exploring new but less popular communities.

Meanwhile, our work is also related to the studies on human online trails and web page navigation (Dimitrov et al. 2017). For example, Singer et al. focus on sequential digital trails of people, and study factors that drive the production of these online trails (Singer et al. 2015). Digital trails in the paper include web navigation, online reviews, and so on.

Human Physical Mobility Patterns

Much research work has been devoted to human mobility in the physical space on various topics such as individual and group level mobility patterns (De Montjoye et al. 2013; Simini et al. 2012), the temporal-spatial properties of human mobility (Hu et al. 2017), and the relation between individual mobility and social connections (Cho, Myers, and Leskovec 2011). Many studies in modelling human mobility suggest the relation between random-walk models and the movements in the physical space (Gonzalez, Hidalgo, and Barabasi 2008). For example, Song et al. model people in the physical space as randomly moving objects, and propose methods to predict human mobility (Song et al. 2010a). Many interesting temporal patterns of human mobility have also been discovered in previous work. For example, (Cheng et al. 2011) indicates that people return to specific locations on a daily basis, and (Noulas et al. 2011) reports the different mobility levels of people on weekdays and week-

ends. Without modelling the distance in the space, a number of entropy-based approaches are proposed to investigate the randomness of human mobility (Cranshaw et al. 2010; Smith et al. 2014). For example, Song et al. study human moving trajectories using entropy, and discover a high predictability in human mobility (Song et al. 2010b).

Data Collection and Preprocessing

We collect our data from Reddit, which was launched in 2005 and is now one of the most visited websites in the world[2]. Due to its high popularity, long time span, and almost complete data availability[3], Reddit has been used as the data source in many previous studies (Hamilton et al. 2017; Zhang, Culbertson, and Paritosh 2017). Reddit is organized into thousands topic-based communities (subreddits), and users are allowed to join any communities at will (except for private subreddits). Such a multi-community setting makes Reddit ideal for our study.

We download all the posts on Reddit from the website's inception on Dec 2005 to Dec 2016. The dataset contains 2.9 billion posts sent by 21 million users on 430 thousand subreddits. Besides text content, each post in the dataset is associated with many other types of information, such as user ID, community ID, time stamp, and so on. We first filter out 0.3 billion posts sent by deleted users (denoted by a user ID of "[deleted]" in the data). We also remove the posts by non-human accounts. To detect non-human accounts (bots), we first collect the users that post with an abnormally high frequency (50 thousand+ posts), and take them as possible examples of non-human accounts. We observe the user IDs of these accounts, and summarize a list of terms that frequently occur in these user IDs, such as "-bot", "_transcriber", and "Moderator". We take the accounts that contain at least one of these terms in their IDs as non-human accounts, and remove all the posts sent by these accounts. In total, we filter out 35 million posts sent by 28 thousand user accounts of this kind. After the data cleaning, we further process the data to extract the visit history of each user. To be specific, the visit history of a user records all the communities the user has visited in the chronological order. In other words, visit histories are users' "trajectories" in the cyber space. Our work investigates the mobility patterns across communities by focusing on user visit histories. Different slices of user visit histories may be used for facilitating the different problems studied in this paper. We will specify the data used for each problem in the corresponding section.

Analogy between the Two Spaces

We first validate the analogy between the cyber space and the physical space. To be specific, we study if the properties of human mobility discovered in the physical space still hold true in the cyber space. On the data collected from Reddit, we conduct the experiments that are originally designed for studying human physical mobility, and compare our results with the conclusions reported in the previous work. Most work on physical mobility is based on data collected in time spans of several months (Barbosa-Filho et al. 2017). To align with the previous work, we collect a data slice containing all the posts on Reddit from January to March 2016. This data slice contains 169 million posts sent by 4.6 million users on 161 thousand subreddits (after data cleaning). All the experiments in this section are conducted on this data slice.

Distributions of Visits

At a coarse granularity, we first compare the distributions of the number of visits in both spaces. (Noulas et al. 2011) reports the complementary cumulative distribution function (CCDF) of the number of visits to a physical location, as well as the CCDF of the number of visits per user. Both distributions reportedly have heavy tails. Moreover, the trend of the number of visits to a physical location follows a power-law distribution (a straight line in log-log scale). We plot the CCDF of the number of visits to an online community, as well as the CCDF of the number of visits per user in Figure 1 (a) $\sim$ (b). The plot shows that both distributions of visits in the cyber space exhibit the same trends as in the physical space. To be specific, both distributions computed from online data also have heavy tails. Moreover, the trend of the number of visits to an online community follows a power-law distribution (Figure 1 (a)).

The same trends of the distributions in the two spaces imply the similarities between communities and physical locations, as well as the similarities between the activities across communities and the movements across locations. The heavy tailed distribution of the visit amount to an online community reveals that, similar to physical locations, only a few communities receive a large number of visits, and a higher number of communities have only few visits. Meanwhile, the same distributions of visits per user suggest that, in both cyber and physical spaces, a small number of people contribute a large amount of visits, while the number of visits of most people is low.

Human Visit Behavior

At a fine granularity, we compare human visit behavior in the two spaces. Much work on human mobility in the physical space suggests the relation between random-walk models and the physical movements of individuals (Gonzalez, Hidalgo, and Barabasi 2008; Castellano, Fortunato, and Loreto 2009). Human physical mobility reportedly exhibits two important quantitative characteristics of random-walk behavior (Song et al. 2010a):

1) the number of distinct locations visited by a user, denoted by $S(t)$, follows

$$S(t) \sim t^{\mu} \tag{1}$$

where t is the time the user spent in the space.

2) the frequency f_k of the kth most visited location of a user follows Zipf's law, and formally,

$$f_k \sim k^{-\zeta} \tag{2}$$

The first characteristic describes the exploration of people in the physical space. The parameter μ is estimated to be

[2]https://www.en.wikipedia.org/wiki/Reddit

[3]Reddit data is made publicly available, and free for download at https://www.reddit.com/3bxlg7. Codes for this project is available at https://github.com/tianranhu.

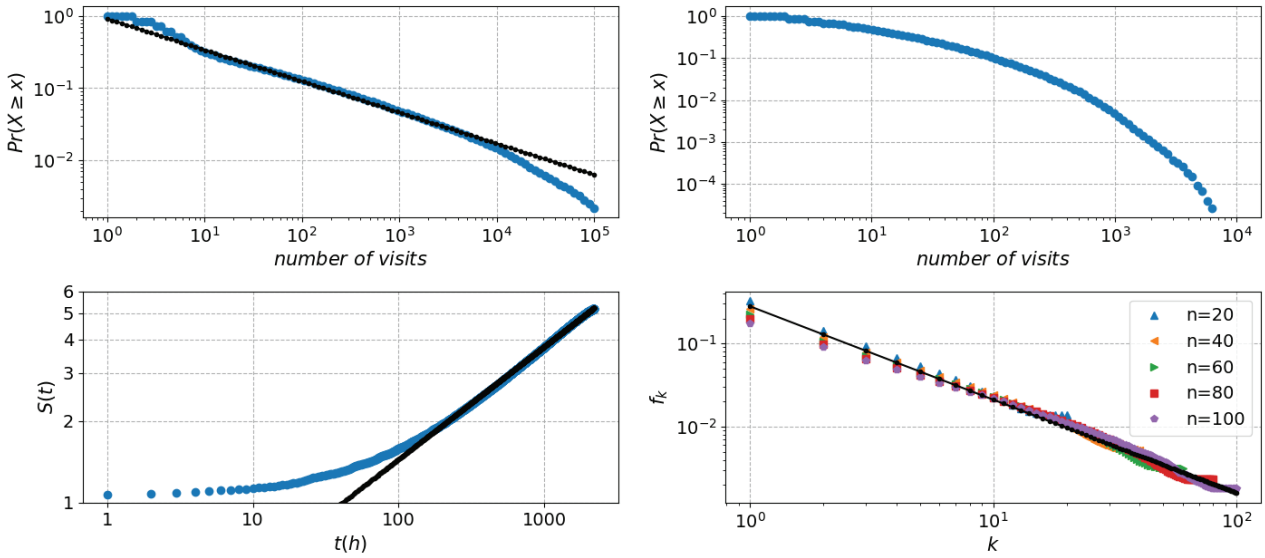

Figure 1: (a) Complementary Cumulative Distribution Function (CCDF) of the number of visits to an online community. (b) CCDF of the number of visits per user. (c) Distribution of the number of visited distinct communities $S(t)$. (d) Distributions of the frequency of the kth most visited community f_k for different S values.

smaller than 1 from physical mobility data, indicating a decreasing tendency of the users to visit new locations through time. The second characteristic indicates that the visits of users are distributed very unevenly, with most visits paid to a few most visited locations. We follow the exact steps suggested in (Song et al. 2010a) to investigate if these two characteristics still apply to the cyber space.

Distribution of Visited Locations For the first characteristic, we first extract the visit histories of all the users from the three month data slice. For a user, we split the visit history into hours, and compute the number of distinct communities visited by the end of each hour, denoted by $s(t)$. Then $S(t)$ is empirically computed as the average of $s(t)$ of all the users. We plot the relation between $S(t)$ and t in Figure 1 (c). The result reveals that in the cyber space the relation $S(t) \sim t^{\mu}$ also holds, suggesting a similar exploration pattern of people in the two spaces. Furthermore, the parameter μ is estimated to be 0.4 in our experiment, which is smaller than the estimation in the physical space (0.6 ± 0.02). This indicates that the tendency of user visiting new communities in the cyber space also decreases over time. Moreover, the decreasing tendency is faster in the cyber space than in the physical space.

Zipf's Law We then validate the Zipf's law in the cyber space. From the data slice, we select the users who visited S unique communities. Different values of S are experimented as suggested in (Gonzalez, Hidalgo, and Barabasi 2008; Song et al. 2010a). We then sort the communities a user visited according to their visit frequencies. The frequency f_k is computed as the average of the visit frequencies to the kth most visited communities of all the users. Please note that k ranges from 1 (most visited) to S (least visited). The relation between f_k and k is plotted in Figure 1 (d). The result shows that the Zipf's law $f_k \sim k^{-\zeta}$ also applies in the cyber space. More strikingly, the parameter ζ is estimated to be 1.12 from our data, which is very close to the results estimated on physical mobility data (1.2 ± 0.1). This indicates that people distribute their visits unevenly in a very similar fashion in both the cyber space and physical space.

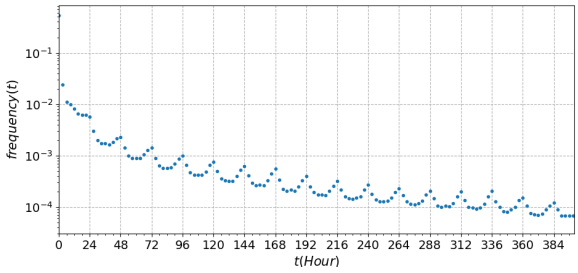

Figure 2: Distribution of the returning probability to online communities.

Our experiments reveal the two characteristics also apply to the movements in the cyber space. The results suggest that individuals' trajectories across online communities can also be described using random-walk models. This further implies the similarity between human mobility in the cyber space and the physical space.

Temporal Properties

Next, we validate our analogy for two well studied temporal properties of human physical mobility: distribution of returning probability (Gonzalez, Hidalgo, and Barabasi 2008;

Cheng et al. 2011) and hourly mobility levels (Noulas et al. 2011; Hu et al. 2016). Returning probability measures the periodic patterns of human mobility. To be specific, the returning probability is the probability that a user returns to a location that the user visited t hours before. In the physical space, the distribution of the returning probability, although having an overall decreasing tendency over time, increases sharply every 24 hours. This indicates a daily pattern of human mobility – people return to specific locations on a daily basis. Zooming into the hourly level, previous work reports a high level of mobility during the daytime and a low level during the nighttime. Furthermore, researchers have observed two peaks in the mobility level around 9am and 7pm on weekdays, matching rush hours in the morning and happy hours in the evening, respectively. As for weekends, human mobility increases rapidly in the morning, and stays at a high level from 2pm to 9pm.

Returning Probability We first compute the distribution of returning probability in the cyber space. For each community a user visited, we record the time gap between every two consecutive visits to the community (i.e. returning time). We then collect the returning time of all the users, and compute the probability of returning to a community in the tth hour. The distribution is plotted in Figure 2. Quite interestingly, we observe an almost identical distribution of returning probability as in the physical space – the probability has an overall decreasing tendency, but increases sharply every 24 hours. It indicates that people return to specific online communities also on a daily basis.

Hourly Mobility Level We then compute the mobility level in the cyber space. One of the requirements for computing the hourly mobility level is to know the clock time of each movement. However, time stamps on Reddit are recorded in UTC time while no time zone information is available. Fortunately, there are many subreddits on specific cities (e.g. /r/nyc[4]). Since the topics in a city-specific subreddit are mostly related to the life in the city, we can assume that most posts in such subreddits are sent by the city residents. Therefore, given the time zone of a city, we are able to compute the local time for each post in the city specific subreddit. Following this idea, we first collect the posts from several popular city-specific subreddits such as /r/nyc, /r/boston, /r/LosAngeles, and so on. We then convert the time stamp of each post to the local time of the corresponding city. The mobility level of one hour in the cyber space is computed as the percentage of posts in the hour. The mobility levels on both weekdays and weekends in the cyber space are shown in Figure 3. We also plot the mobility levels in the physical world as reported in (Noulas et al. 2011) for a better comparison.

Similar to the physical space, we observe that the mobility level in the cyber space is much higher during the daytime than during the nighttime. Also, on weekdays the mobility level in the cyber space varies significantly during the daytime, while on weekends the tendency of the level appears to be relatively flat. More interestingly, Figure 3 shows a

[4]On Reddit, a subreddit is denoted as "/r/" + an unique name.

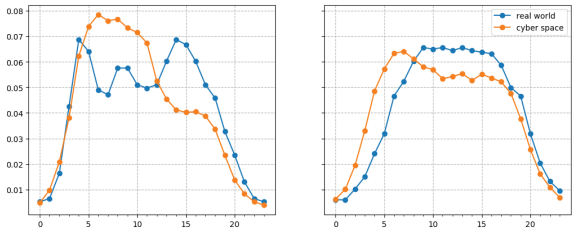

Figure 3: Mobility levels in the cyber space over weekdays (left) and weekends (right). Please note that the two subplots share the same y-axis for a better comparison between weekdays and weekends. The values of frequency of visits in the physical space are learned from (Noulas et al. 2011).

complementary relation between the two spaces – when the mobility level is high in the cyber space, the level is low in the physical space, and vice versa. For example, in the physical space, the mobility level during the work hours (from 9am to 5pm) is relatively low on weekdays. On the contrary, the highest mobility level on weekdays in the cyber space occurs exactly in these hours, indicating that people are more likely to surf the Internet during work hours. The complementary relation can also be observed from the mobility level on weekends – in the afternoon when the mobility level is high in the physical space, the mobility level in the cyber space decreases correspondingly.

The temporal properties of human online mobility reveal several interesting relations between the cyber space and the physical space. On one hand, people exhibit the same daily returning pattern, and their mobility levels follow the same day-night cycle in both spaces. This suggests that, although in two different spaces, human circadian rhythm remains unchanged. On the other hand, we discover the complementary relation between the two spaces. This makes sense intuitively – although sometimes people can access online communities while in transportation, in most cases they cannot "move" in both spaces at the same time.

Human Mobility in the Cyber Space

The similarities between the cyber space and the physical spaces imply the possibility of applying the approaches originally designed for human physical mobility to human online mobility. In this paper, we borrow the idea of entropy-based approaches (Song et al. 2010b), and study human online mobility from three aspects: 1) the randomness of human online mobility, 2) online mobility patterns, and 3) the preferences to online communities by the people of different mobility patterns. We select entropy-based approaches because such approaches do not require a distance measurement, which is not universally defined in the cyber space.

Randomness of Mobility in the Cyber Space

Entropy is widely applied to measure the randomness of human mobility in the physical space (Smith et al. 2014; Cranshaw et al. 2010). Inspired by the previous work, we use the entropy of the visit histories of a user to measure the

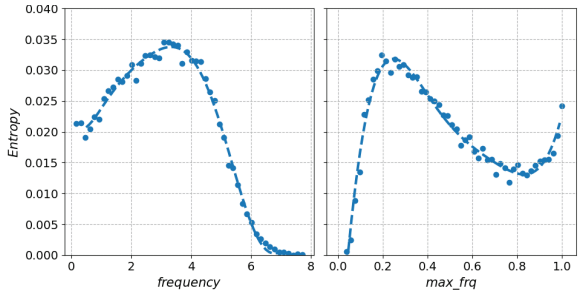

Figure 4: Distribution of the entropy of a user's visits to online communities (left). Distribution of the frequency of a user's most visited community max_frq (right). Please note that the two subplots share the same y-axis.

randomness of her mobility across online communities. Formally, the entropy of the visit history of a user u, denoted by $En(u)$, is computed as:

$$En(u) = -\sum_{n} p_i log(p_i), \tag{3}$$

where n denotes the number of unique communities the user visited, and p_i is the probability of the user visiting the ith community. In general, entropy measures how precisely the next community a user will visit can be predicted given the visit history. The lower the entropy is, the lower the mobility randomness is. Another more intuitive measurement that describes the randomness of a user's mobility is the frequency to the most visited community, denoted by $max\_frq(u)$. Formally, $max\_frq(u)$ is computed as

$$max\_frq(u) = max(p_i) \tag{4}$$

Clearly, a larger value of max_frq indicates the user is more focused on a specific community, and therefore implies a lower user mobility randomness. To investigate the mobility randomness in the cyber space, we compute both entropy and max_frq for each user. We conduct the experiments on the same data slice from January to March 2016 in this task, and apply the constraints suggested in (Song et al. 2010b) to remove inactive users (i.e. n>2 and total visits>1000). The distributions of the two measurements are plotted in Figure 4.

We observe that mobility randomness varies significantly across users. Over 25% users have an entropy value larger than 4, suggesting that a high randomness of such users' online mobility. Please note that an entropy value of 4 indicates that the next community the user will visit could be found on the average in any of $2^4 = 16$ communities. Similarly, over 30% users have a max_frq value lower than 0.3 indicating that many users do not focus on specific communities. Meanwhile, the mobility randomness of a large portion of users in the cyber space is very low. From the distribution of max_frq, we observe that about 17% users have a max_frq value larger than 0.8, and nearly 10% users with a value larger than 0.9. Please note that a $max\_frq = 0.9$ indicates that the user direct 90% of all the visits to one community. In other words, although these users are active online, they almost always devote their visits to only one community. The distribution of entropy echoes the finding by showing that over 13% users have an entropy value lower than 1.

Mobility Patterns in the Cyber Space

Given the diverse mobility randomness across people in the cyber space, a nature follow-up question is: does human online mobility change over time? Take the people with very low mobility randomness for an example. We wonder if they always focus on a specific community since they join the platform, or they explore many communities in the beginning and discover their favorites later. To study the problem, we focus on the active users whose whole lifespans on the platform are available (Danescu-Niculescu-Mizil et al. 2013; Tan and Lee 2015). In other words, these users have been active on the platform, but left the platform eventually. Therefore, we select the users who used to be active before January 2016 and never post after the time. We apply the same constraints as in the task for choosing active users (i.e. n>2 and total visits>1000). There are around 69,000 users meeting the constraints. We collect the complete visit histories of these users, and conduct our following experiments on this data slice.

Methodology Since the total number of visits varies significantly across users, we first equally divide the lifespan of a user into 20 stages[5] as suggested in (Danescu-Niculescu-Mizil et al. 2013). In other words, each stage of a user accounts for 5% of all the visits through the user's lifespan. By doing this, we are able to align the stages of all the users, and study the evolution of user mobility over different stages. We then quantify the mobility within each stage of a user. Entropy and max_frq are again used to measure mobility randomness in a stage. However, these two measurements do not quantify to what extent user explore unvisited communities over stages. For example, a user could visit two totally different sets of communities in two stages, but with the same mobility randomness. To measure the user exploration to new community in a stage, we apply another measurement $P(new\_comm)$ – the probability of a user visiting a new community that never has been visited in previous stages (Mcinerney et al. 2013; Lian et al. 2015). Formally, for the ith stage of a user u, let $V_{u,i}$ denote the set of the visits of u in this stage. The set of the visits to the communities that are visited in the ith stage but never have been visited before is denoted as $V_{u,i}^{new}$. Therefore, $P(new\_comm)$ is computed as

$$P(new\_comm) = \frac{|V_{u,i}^{new}|}{|V_{u,i}|} \tag{5}$$

For each user, we compute the entropy and max_frq for all the 20 stages in order. We also compute $P(new\_comm)$ from the second stage to the last stage. We exclude the first stage because all the visits in this stage are directed to new communities, thus the value of $P(new\_comm)$ of the

[5]We obtain similar results for other choices of number of stages.

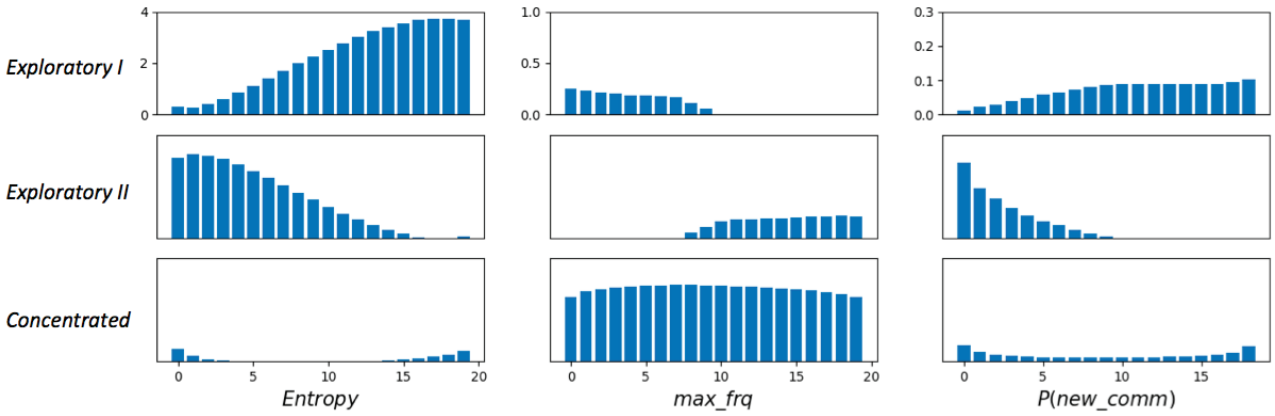

Figure 5: Three mobility patterns in the cyber spaces extracted by NMF. For a better illustration, we plot the weights of entropy, max_frq, and $P(new\_comm)$ separately.

first stage is 1 for all the users. Using the three measurements, we quantify a user's mobility over 20 stages into a 59-dimensional vector. The first 20 dimensions are the entropy values of all the stages, the next 20 dimensions are the max_frq values, and the remaining 19 dimensions are the $P(new\_comm)$ values of the 19 stages (the first stage is excluded). The vectors of all users are then stacked to create a 59×69,000 matrix. This matrix records the mobility over time for all the users. We investigate the patterns of mobility evolution by decomposing the matrix. Since Non-negative Matrix Factorization (NMF) has been successfully used for mining interpretable temporal-spatial human mobility patterns (Lee and Seung 1999), we also apply NMF to complete the decomposition.

Mobility Patterns By setting the number of component k to 3, NMF reveals three very interpretable mobility patterns in the cyber space: two exploratory patterns and one concentrated pattern[6]. We plot the three patterns in Figure 5 by showing the tendency of the three measurements over stages. We summarize the patterns as follows:

- ***Exploratory Pattern I***: The people of this pattern concentrate on a few communities in the early stages, but explore more and more new communities with high randomness in the late stages. The entropy value of these users is initially low, and increases over stages. Correspondingly, the max_frq value starts at a high level, and decreases over stages. The tendencies of both entropy and max_frq indicate that the mobility randomness is low in the beginning, and goes up as time goes on. In other words, these user only focus on a limited number of communities at first, but gradually lose their concentrations later. Meanwhile, the value of $P(new\_comm)$ is low in the beginning, and increases over stages. This tendency indicates that the users start at a low level of interest in new communities, but explore more and more unvisited communities over stages.
- ***Exploratory Pattern II***: The people of this pattern is the opposite to the first pattern. They explore many new communities in the early stages, but only focus on a few communities later. The mobility randomness is initially high, as indicated by a high entropy value and low max_frq value. This suggests that these users do not concentrate on any communities in the beginning. As time goes on, the mobility randomness monotonously decreases, as indicated by the decreasing tendency of entropy and the increasing tendency of max_frq over stages. In other words, the users discover their interested communities, and pay more and more attention to these communities. Meanwhile, the decreasing tendency of $P(new\_comm)$ also suggests that the users pay a large amount of visits to new communities in the early stages, but gradually stop exploring unvisited communities over stages.
- ***Concentrated Pattern***: The people of this pattern are very concentrated – they direct almost all their visits to a small and unchanged set of communities through the whole lifespan. In this pattern, both entropy value and $P(new\_comm)$ value are low for all the stages. This suggests that these user only visit specific communities, and rarely explore unvisited communities. Correspondingly, the max_frq is high for all the stages, also indicating the low overall mobility randomness of this pattern.

Exploratory Patterns vs. Concentrated Pattern

Given the divergent mobility patterns people exhibit in the cyber space, we further investigate the relation between mobility pattern and online community preference. For example, we wonder if the concentrated type people also like to visit communities that the exploratory people usually visit, and if the people of the two exploratory patterns share similarities in their preferences to communities. We study this problem by converting it to a classification task. In this classification task, we take the three mobility patterns of users

[6]With a k value larger than 3, the two exploratory patterns are further decomposed into smaller but less interpretable components, and the concentrated pattern is barely affected. Therefore, we set k to 3 in our experiments

	Precision	*Recall*	*F-1 Score*
Exploratory I	0.64	0.86	0.74
Exploratory II	0.32	0.10	0.15
Concentrated	**0.77**	**0.71**	**0.74**
avg / total	0.63	0.67	0.63

Table 1: Classification results among the three patterns.

as class labels, and attempt to distinguish the three classes only using the communities visited by the users as features.

Classification among Three Mobility Patterns From the results of NMF, we find the mobility pattern to which a user is assigned the highest weight among the three patterns, and take the mobility pattern as the user's class label. By doing so, we obtain around 31,000 and 13,000 users of Exploratory Pattern I and Pattern II, respectively. The around 25,000 remaining users are labelled as the Concentrated Pattern. The amounts of visits to different communities of a user are used as the classification features. To be specific, we first remove the unpopular communities that have less than 50 unique users, and obtain 7,526 unique communities. We collect the numbers of visits to the selected communities for each user. TF-IDF is then applied to weight the numbers of visits across all the users, and the weighted results are used as the classification features. We use 80% of the data for training, and the remaining 20% for testing. A logistic regression model is employed for the task. The classification results are reported in Table 1.

The results show that people of the Concentrated Pattern can be distinguished from the people of Exploratory Pattern I and Pattern II, with a high precision of 0.77 and recall of 0.71. This indicates that concentrated people have different preferences to communities from exploratory people. However, the classifier cannot distinguish people of Exploratory Pattern I and II. Because of the much larger user size of Exploratory Pattern I, a large amount of users of Exploratory Pattern II are classified as Exploratory Pattern I. This leads to the low precision (0.32) and recall (0.1) for Exploratory Pattern II. Due to the same reason, Exploratory Pattern I receives a high recall (0.86) but a low precision (0.64). The results imply that users of the two exploratory patterns share similarities in their community preferences, and therefore cannot be simply distinguished by only using the community features.

Community Preference of Different Patterns From the trained logistic model, we can tell the features (communities) assigned with the highest positive and negative coefficients for distinguishing the users of the Concentrated Pattern. These communities are reported in Table 2. Clearly, the communities with the positive coefficients are preferred by the users of the Concentrated Pattern. In contrast, the communities with the negative coefficients are preferred by the users of either Exploratory Pattern I or Pattern II. We also report the values of the coefficients and the user sizes of the communities in the table. Two interesting differences can be observed from these two groups of communities. First,

Subreddit	*User Size*	*Coefficient*
top five positive features (communities)		
/r/stopdrinking	94,187	0.013
/r/thinkpad	24,545	0.012
/r/MinecraftCirclejerk	2,508	0.011
/r/incremental_games	39,250	0.011
/r/autism	21,461	0.011
top five negative features (communities)		
/r/bestof	4,841,958	-0.083
/r/reactiongifs	1,260,843	-0.064
/r/woahdude	1,571,722	-0.051
/r/comics	836,383	-0.050
/r/technology	5,866,352	-0.041

Table 2: Five top positive and negative features (communities) for classifying the people of the Concentrated Pattern. We list the communities along with their user sizes and the values of the coefficients.

the communities preferred by the concentrated type people are much smaller than the communities preferred by the exploratory type people. None of the top five positive communities has a user size larger than 100,000, and the smallest only has a user size of 2,500. In contrast, four of the top five negative communities have a user size larger than 1 million. Furthermore, the communities preferred by the concentrated type users are usually on specific topics. For example, in the top five positive communities, two communities are on specific games (/r/MinecraftCirclejerk and /r/incremental_games), two communities are on personal issues (/r/stopdrinking and /r/autism), and one community is on a specific product (/r/thinkpad). In contrast, the top five negative communities are all on general topics such as /r/bestof and /r/reactiongifs.

Conclusion & Future Work

In this paper, we present a novel analogy between the cyber space and the physical space. We quantitatively validate the analogy from three representative and progressive aspects: visit distributions, individual visit behavior, and temporal properties. Our experiments on the three aspects all reveal striking the similarities between human mobility in the two spaces. Next, we study human online activities by treating the communities as locations in the cyber space, and activities across communities as movements across locations. By applying the framework originally designed for studying human physical mobility, we investigate the mobility patterns in the cyber space. It is observed that the randomness of online mobility varies significantly across users. Furthermore, we study the evolution of human mobility in the cyber space, and discover three interesting exploration patterns of by users of online communities. Moreover, our experiments suggest divergent preferences to different online communities across people of different patterns. Our work provides valuable insights into the human activities under the multi-community setting. More importantly, we uncover the interesting similarity between the two spaces, and suggest a

promising research direction.

In the future, we plan to build upon our work mainly from two aspects. First, we would like to quantify the "cost" of users moving among communities, i.e. the "distance" in the cyber space. Such a distance measurement would allow us to apply more well-established frameworks for modelling the physical movements to the online scenarios. Second, we would like to study online communities by borrowing the approaches developed for studying physical locations. Much previous work discusses the characteristics of physical locations. It would be interesting to see if the locations in the cyber space exhibit similar characteristics.

Acknowledgments

We thank the support of New York State through the Goergen Institute for Data Science, Tencent AI Lab, and NSF Award #1704309.

References

Barbosa-Filho, H.; Barthelemy, M.; Ghoshal, G.; James, C. R.; Lenormand, M.; Louail, T.; Menezes, R.; Ramasco, J. J.; Simini, F.; and Tomasini, M. 2017. Human mobility: models and applications. *arXiv preprint arXiv:1710.00004*.

Castellano, C.; Fortunato, S.; and Loreto, V. 2009. Statistical physics of social dynamics. *Reviews of modern physics* 81(2):591.

Chan, J.; Hayes, C.; and Daly, E. M. 2010. Decomposing discussion forums and boards using user roles. *ICWSM* 10:215–218.

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. Z. 2011. Exploring millions of footprints in location sharing services. *ICWSM* 2011:81–88.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1082–1090. ACM.

Cranshaw, J.; Toch, E.; Hong, J.; Kittur, A.; and Sadeh, N. 2010. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, 119–128. ACM.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, 307–318. ACM.

De Montjoye, Y.-A.; Hidalgo, C. A.; Verleysen, M.; and Blondel, V. D. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3:1376.

Dimitrov, D.; Singer, P.; Lemmerich, F.; and Strohmaier, M. 2017. What makes a link successful on wikipedia? In *Proceedings of the 26th International Conference on World Wide Web*, 917–926. International World Wide Web Conferences Steering Committee.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Hamilton, W. L.; Zhang, J.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Loyalty in online communities. *arXiv preprint arXiv:1703.03386*.

Hessel, J.; Tan, C.; and Lee, L. 2016. Science, askscience, and badscience: On the coexistence of highly related communities. In *ICWSM*, 171–180.

Hu, T.-r.; Luo, J.-b.; Kautz, H.; and Sadilek, A. 2016. Home location inference from sparse and noisy data: models and applications. *Frontiers of Information Technology & Electronic Engineering* 17(5):389–402.

Hu, T.; Bigelow, E.; Luo, J.; and Kautz, H. 2017. Tales of two cities: Using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. *IEEE Transactions on Big Data* 3(1):55–66.

Kraut, R. E.; Resnick, P.; Kiesler, S.; Burke, M.; Chen, Y.; Kittur, N.; Konstan, J.; Ren, Y.; and Riedl, J. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Lian, D.; Xie, X.; Zheng, V. W.; Yuan, N. J.; Zhang, F.; and Chen, E. 2015. Cepr: A collaborative exploration and periodically returning model for location prediction. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6(1):8.

Lin, Z.; Salehi, N.; Yao, B.; Chen, Y.; and Bernstein, M. S. 2017. Better when it was smaller? community content and behavior after massive growth. In *ICWSM*, 132–141.

Mcinerney, J.; Stein, S.; Rogers, A.; and Jennings, N. R. 2013. Breaking the habit: Measuring and predicting departures from routine in individual human mobility. *Pervasive and Mobile Computing* 9(6):808–822.

Noulas, A.; Scellato, S.; Mascolo, C.; and Pontil, M. 2011. An empirical study of geographic user activity patterns in foursquare. *ICWSM* 11:70–573.

Pavalanathan, U.; Fitzpatrick, J.; Kiesling, S.; and Eisenstein, J. 2017. A multidimensional lexicon for interpersonal stancetaking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 884–895.

Pavlick, E., and Tetreault, J. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics* 4:61–74.

Ren, Y.; Kraut, R.; and Kiesler, S. 2007. Applying common identity and bond theory to design of online communities. *Organization studies* 28(3):377–408.

Rotman, D.; Golbeck, J.; and Preece, J. 2009. The community is where the rapport is–on sense and structure in the youtube community. In *Proceedings of the fourth international conference on Communities and technologies*, 41–50. ACM.

Simini, F.; González, M. C.; Maritan, A.; and Barabási, A.-L. 2012. A universal model for mobility and migration patterns. *Nature* 484(7392):96–100.

Singer, P.; Helic, D.; Hotho, A.; and Strohmaier, M. 2015. Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *Proceedings of the 24th International Conference on World Wide Web*, 1003–1013. International World Wide Web Conferences Steering Committee.

Smith, G.; Wieser, R.; Goulding, J.; and Barrack, D. 2014. A refined limit on the predictability of human mobility. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, 88–94. IEEE.

Song, C.; Koren, T.; Wang, P.; and Barabási, A.-L. 2010a. Modelling the scaling properties of human mobility. *Nature Physics* 6(10):818–823.

Song, C.; Qu, Z.; Blumm, N.; and Barabási, A.-L. 2010b. Limits of predictability in human mobility. *Science* 327(5968):1018–1021.

Tan, C., and Lee, L. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*, 1056–1066. International World Wide Web Conferences Steering Committee.

Tran, T., and Ostendorf, M. 2016. Characterizing the language of online communities and its relation to community reception. *arXiv preprint arXiv:1609.04779*.

Zhang, J.; Hamilton, W. L.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Community identity and user engagement in a multi-community landscape. *arXiv preprint arXiv:1705.09665*.

Zhang, A. X.; Culbertson, B.; and Paritosh, P. 2017. Characterizing online discussion using coarse discourse sequences.

Zhu, H.; Kraut, R. E.; and Kittur, A. 2014. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 281–290. ACM.