공학박사 학위논문

Analytic Approach of Understanding Crowd Phenomena on the Internet: Case Studies of MMORPG and BitTorrent

온라인 게임과 컨텐트 공유 네트워크 분석을 통한 온라인 군집 현상의 이해

서울대학교 대학원
컴퓨터공학부
정태중

Analytic Approach of Understanding Crowd Phenomena on the Internet: Case Studies of MMORPG and BitTorrent

지도교수 권태경

이 논문을 공학박사 학위논문으로 제출함

2014년 12월

서울대학교 대학원

컴퓨터공학부

정태중

정태중의 박사 학위논문을 인준함

위 원 장	김 종 권	(인)
부위원장	권 태 경	(인)
위 원	이 상 구	(인)
위 원	나 종 연	(인)
위 원	문 수 복	(인)

Abstract

Analytic Approach of Understanding Crowd Phenomena on the Internet: Case Studies of MMORPG and BitTorrent

Taejoong Chung
School of Computer Science & Engineering
The Graduate School
Seoul National University

Quantification of collective human behavior and understanding the group characteristics in the Internet is important in user behavior studies since people tend to gather together and form groups due to their inherent nature. On the Internet, people are also often forming a group for a specific purpose such as i) an online group in games (e.g., MMORPGs) to experience various social interactions with other players or accomplish a difficult quest with teammates or ii) a swarm in peer-to-peer network to share a content to utilize a higher download rate with an availability. To this end, we studied the two most well-known major applications in the Internet that people are actively using with different purposes; i) MMORPGs and ii) BitTorrent.

In this dissertation, we analyze the i) group activities of users in Aion, one of the largest MMORPGs, based on the records of the activities of 94,497 users and ii) crowd phenomena of BitTorrent. First, in a case study of Aion, we focus on (i) how social interactions within a group differ from the ones across groups, (ii) what makes a group rise, sustain, or fall, (iii) how group members join and leave a group, and (iv) what makes a group end. We first find that structural patterns of social interactions within a group are more likely to be close-knit and reciprocative than the ones across groups. We also observe that members in a rising group (i.e., the number of members increases) are more cohesive, and communicate with more evenly within the group than the ones in other groups. Our analysis further reveals that if a group is not cohesive, not actively communicating, or not evenly communicating among members, members of the group tend to leave.

Second, we investigate what kinds of crowd phenomena of content exist and why different patterns of crowd phenomena appears and how we can exploit content crowd phenomena considering the content category, publisher, and population of content in BitTorrent. To this end, We conduct comprehensive measurements on content locality in one of the largest BitTorrent portals: The Pirate Bay. In particular, we focus on (i) how content is consumed from spatial and temporal perspectives, (ii) what makes content be consumed with disparity in spatial and temporal domains, and (iii) how we can exploit the content locality. We find that content consumption in real swarms is 4.56 times and 1.46 times skewed in spatial (country) and temporal (time) domains, respectively. We observe that a cultural factor (e.g., language) mainly affects spatial locality of content. Not only the time-sensitivity of content but also the publishing purpose affects temporal locality of content. We reveal that spatial locality of content

rarely changes on a daily basis (microscopic level), but there is notably spatial spread of content consumption over the years (macroscopic level). Based on the observation, we conduct simulations to show that bundling and caching can exploit the content locality.

Keywords : Group Activities, Crowd Phenomena, MMORPG, BitTorrent, Peer-to-Peer

Student Number : 2009-23148

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Quantification of collective human behavior and analyzing crowd phenomena on the Internet have became an cruical factor to understand complex human behaviors. People using many applications on the Internet such as online social networks (OSNs), online games, and file sharing application via peer-to-peer network give us an opportunity to observe the socio-economic behaviors of humans with available datasets by passive ways of measurement. On the Internet, people are also often forming a group for a specific purpose such as i) an online group in games (e.g., MMORPGs) to experience various social interactions with other players or accomplish a difficult quest with teammates or ii) a swarm in peer-to-peer network to share a content to utilize a higher download rate with an availability. Understanding the dynamics of group-focused activities on the Internet with multiple perspectives is important in human behavior studies since people tend to gather together and form groups due to their inherent nature both online and offline.

To this end we studied the two most well-known major applications in the Internet that people are actively using with different purposes; i) MMORPGs and ii) BitTorrent

1.1 Crowd Phenomena in Massively Multi-player Online Role-Playing Games (MMORPGs)

It is reported that tens of millions worldwide enjoy Massively Multi-player Online Role-Playing Games (MMORPGs) as of 2013. An MMORPG typically offers its players rich virtual environments where they can engage in various real-world-like interactions including combats, trades, and conversations with others. Given the complexity, variety, and longevity of the virtual worlds, user experiences in MMORPGs are expected to be close to real life ones.

The realistic virtual environments of MMORPGs opens up new opportunities for researchers to understand complex human behaviors. That is, a game space in an MMORPG is deemed as a large scale virtual laboratory for observing the socio-economic behaviors of humans. The landscape of user behavior studies like sociology or psychology has been changed by the proliferation of online services including online social networks (OSNs) and MMORPGs [1, 2]. Since MMORPGs provide real life-like environments where users can experience various social interactions like communications, cooperations, economic activities, and so on, there has been an increasing interest in analyzing various activities in popular MMORPGs [1–4]. For example, [3, 4] investigated various kinds of social interactions (e.g., friendship, conversation, trade) among users in MMORPGs.

Although the above studies on MMORPGs reveal valuable insights into the social interactions among users, most of these studies paid little attention to the social interactions in a group (or a community), rendering the following research questions: *How groups are structured by people? How a user interacts with another in a group?*

What are the differences among groups in terms of diversity of social interactions or economic activities? How groups evolve and why? Why people leave a group? What makes a group rise and fall? We argue that understanding the dynamics of group-focused activities in MMORPGs is important in human behavior studies since people tend to gather together and form groups due to their inherent nature both online and offline.

1.2 Crowd Phenomena in BitTorrent

BitTorrent is one of the most popular peer-to-peer (P2P) applications and responsible for a substantial amount of current Internet traffic [5]. However, its network-oblivious nature has posed a few challenges from the networking perspective. First, P2P connections may incur substantial transit traffic between different networks of Internet Service Providers (ISPs) [6, 7]. Second, its peering strategy may lead to sub-optimal throughput [8]. Third, its time-varying traffic patterns can be a hurdle for traffic engineering [9, 10].

To address the above issues, both ISPs and P2P application developers have considered various alternatives [11–14]. ISPs have tried many traffic control techniques such as rate throttling or charging [15]. However, this appears to be inefficient because most P2P applications are dodging this control. On the other hand, some P2P applications have adopted techniques to improve efficiency by localizing P2P connections (i.e., preferring peers within the same ISP [11, 12]). These techniques, however, have limitations since network information such as topology, cost, and link status is at best inferred by application-level traffic observations. To overcome these

limitations, there have been efforts to promote the cooperation between ISPs and P2P applications; ISPs can provide the above information to P2P applications [13, 14].

Recently, some studies turn their attention to grouping phenomena (i.e., *locality*) among peers to fundamentally understand the above problems [16–18]. Here, the locality generally indicates how much disparity exists in content sharing patterns from the spatial and temporal perspectives. According to [17, 18], 30% more connections among peers in the same ISP compared to a random graph are observed for more than 45% of peers, which indicates that BitTorrent connections among peers are biased to local peers. They further argued that the localized nature of BitTorrent may help both ISPs to reduce inter-ISP traffic and P2P applications to improve the download speed. [16] found that substantial amount of BitTorrent traffic does not reach higher-tier ISPs, which is in line with [17, 18]. The authors of [16] also revealed that BitTorrent's temporal usage patterns are observed to vary in a diurnal fashion. Based on this observation they argued that ISPs need to devise a better price model to balance the traffic over time.

While these studies focus on how much BitTorrent traffic is localized in swarm dynamics, most of them paid little attention to investigating locality phenomena from a content perspective, which we call *content locality*. We focus on the following questions: *How are content files (spatially and temporally) consumed by human beings, and why these phenomena occur in BitTorrent? Are there any skewed patterns in the way people participating in BitTorrent swarms depending on the content properties (e.g., content types or cultural aspects)?* We argue that understanding content locality with emprically-grounded evidences is important for BitTorrent stakeholders: (i) how BitTorrent service providers deal with locality to improve system performance

and (ii) how content providers publish torrents, especially for increasing sales. For instance, ISPs may develop content caching strategies by considering locality phenomena to reduce the inter-ISP traffic.

Chapter 2

Related Work

2.1 Crowd Phenomena in MMORPGs

2.1.1 Social Interactions in MMORPGs

The landscape of user behavior studies like sociology or psychology has been fundamentally changed by the proliferation of online social networks [1, 19]. However, most of studies on online social networks have mostly focused on a single type of interactions (e.g., phones or online buddy relations), missing the wide variety of human interactions in real life [20, 21]. In contrast, MMORPGs provide rich virtual environments where users engage various real world-like interactions including combats, trades, and conversations with others, which allows researchers to explore the richer details of real-world-like complex and various social interactions [3, 4].

Consequently, there has been an increasing interest in social interactions in MMORPGs (e.g., virtual places for social purposes [22], shared experiences [23], and social bonds [2, 24]). However, most of the prior studies have used traditional methods of social science such as interviews and questionnaires that need substantial time and resources to deliver statistically meaningful assertions, which may also introduce well-known biases [24, 25]. Recently, [3] analyzed the structural equivalence among the interaction networks in an MMORPG, and [4] investigated various kinds of social interactions (e.g., friendship, conversation, trade, and etc.) among users in

an MMORPG, based on the datasets from game providers. Our work is also based on the datasets (and their analysis) of multiple social interactions of an MMORPG from game providers; however, we focus more on various social activities from a group perspective, which has been paid little attention.

2.1.2 Group Activities in MMORPGs

Understanding the group activities is important in human behavior studies due to the nature of people gathering together and forming groups, which is one of the key drives of a society [26]. Thus there have been a few studies to understand the motivation of group activities in MMORPGs, and the motivation of gaming or joining groups [23], based on the interviews or surveys [24, 27]. Also, some studies have tried to understand the structural properties [2] or stability [28, 29] of groups in MMORPGs. However they cannot see comprehensive interactions among users since their collected data is limited (i.e., querying the status of users only from the client-side interface provided in a game). [30] studied the combat groups (i.e., parties) in an MMORPG, which are formed for cooperative game play; however social aspects of groups are not investigated. To the best of our knowledge, this is the first work that comprehensively and empirically investigates the various group activities of users (i.e., social and economic aspects), using the records of users' activities provided by a game provider.

2.1.3 Group Activities in Other Online Services

Since gathering together and forming groups are the inherent structure of society, their structure and evolution have been investigated in other online services

such as [19, 31, 32]. [33] tried to study the structures of *implicit* communities and their properties by identifying the clusters within a given graph, which are characterized by implicit factors such as the density of links. Also, there have been studies to investigate how the binary relationship (e.g., friendship or co-working relationship) affects the formation of *explicit* groups in LiveJournal [32], churning of users within a group (i.e., an identical conference) in DBLP [31], or network evolution in LiveJournal, Flickr, and YouTube [19]. In the context of social science, mathematical modeling (e.g., diffusion model) has been proposed to explain group evolution and change [34], but social interactions are not considered. We investigate the group activities in MMORPGs, where users can engage various real world-like interactions including conversations, trades, and combats.

2.2 Crowd Phenomena (Locality) in BitTorrent

2.2.1 Peer Localization

The network-oblivious nature of P2P applications poses the above challenges to both ISPs and P2P application developers. To address these issues, many P2P applications have adopted techniques to improve networking efficiency by localizing application-level peering (i.e., preferably select peers within the same ISP). For example, [11] suggested a client-side solution without any help from ISP, which helps to select peers within the same ISP to achieve the better throughput and reduce the inter-ISP traffic. However, relying solely on peers has fundamental limitations because the network information such as topology, cost, and link status is at best inferred by application-level traffic observations. To overcome the limitation, [13] and [14] sug-

gest explicit cooperation between ISPs and P2P applications to localize traffic within ISP.

2.2.2 Crowd Phenomena in BitTorrent

Recently, there have been studies of BitTorrent on locality phenomena to fundamentally understand the BitTorrent's traffic characteristics [16–18]. [18] analyzed the impact of locality-aware peer selection algorithms of BitTorrent on the inter-ISP traffic and download times of end users. The authors revealed that the localized nature of BitTorrent can help both ISPs by reducing inter-ISP traffic and P2P applications by improving download speeds. The measurement results of [17] showed that 45% of peers have more than 30% peers in the same ISP compared to a random graph, which implies that the current BitTorrent mechanism preferably selects peers in the same ISP. [16] also showed that the major BitTorrent traffic does not reach high-tier ISPs (tier 1 or 2), which signifies that locality phenomena are more prevalent in the stub ISPs. The authors further revealed that BitTorrent temporal usage exhibits the peak in the evening, which implies the presence of temporal locality.

[35, 36] showed different levels of locality phenomena are observed in BitTorrent depending on the link bandwidth of peers and the popularity of content files. The authors of [35] revealed that peers with similar upload/download speeds in a swarm tend to have more connections one another, which is mainly due to the choking algorithm. While these studies on BitTorrent locality have been focusing on traffic characteristics of BitTorrent, our focus is to empirically analyze the the locality of BitTorrent in aspect of content considering multiple content properties such as content category (e.g., movie or music).

2.2.3 Locality in Other Domains

There have been many studies to understand and exploit *"locality phenomena"* in various domains. The locality phenomena also can be found in traditional libraries [37] or newspapers [38]. Interestingly, the locality phenomena also can be observed in online social networking (OSN) services in the Internet. [39] investigated the relation between popularity and geographical locality of YouTube videos and further showed that sharing videos in OSN widens the geographical reach of the videos. In Twitter, [40] showed the users' geographical proximity with their followers and further revealed that language and cultural characteristics determine the level of locality in Twitter. [41] observed that majority of communications in Facebook are occured within the same geographical region. By exploiting the spatial (i.e., geographical) locality phenomena, [41, 42] tried to design the system and improve performance of OSN with mininal infrastructural and operational cost. Inspired from prior work of locality phenomena in other domains, this is the first measurement study to empirically investigate content locality in BitTorrent.

Chapter 3

Group Activities in Online Social Game

3.1 Aion overview

In this section, we explain the key features of Aion from the perspective of social interactions, and introduce its datasets.

3.1.1 Game Features

Aion ranked as the second most played MMORPG with over 3.4 million people from more than 60 countries[1] as of early 2011 [43]. Similar to most of the other MMORPGs, a user chooses one of the virtual worlds to engage. In her world, she can do various kinds of social interactions with others including conversations, economic activities or joining a group much like in her real life.

3.1.1.1 Social Interactions

We model social interactions among users (in a world) as a graph $S = (V, E)$, where V is the set of users (or nodes) participating in the virtual world, $\{v_1, \cdots, v_n\}$, and E is the set of directional social interactions, $\{e_1, \cdots, e_m\}$, where $e_i \in \{Friendship, Whisper, Mail, GroupChat, Trade, Shop, PartyRequest\}$. A social interaction in Aion refers to one the following actions:

[1] http://www.alteredgamer.com/pc-gaming/35992-mmo-subscriber-populations/

Friendship: A user can invite another user to be her *friend*. Upon the approval of the invited user, she can easily check the status of her friends. The direction of the edge is from the inviter to the invitee. There are 103,995 Friendship records in our datasets.

Conversation: There are three types of communications; *Whisper*, *Mail*, and *Group Chat*. A Whisper can take place between any two online users, and the others cannot overhear this. If the receiver is not online, the sender can send a mail. If a user belongs to a group, she can broadcast a message to all the other online members in the same group through a Group chat. For the three types of communications, we cannot see message contents; however, we can retrieve the sizes of the messages. There are 27,479,612, 49,706,934, and 475,236 records for Whisper, Group chat, and Mail, respectively.

Economic Activities: A user can send a request to another in proximity to exchange or give items using the *Trade* interaction. A user can also open a *Shop* in a designated place (in the virtual space), and any user can go to the *Shop* to purchase or sell in-game items. In this case, an outgoing edge (out-degree) is drawn from a buyer to a seller. Our datasets contain 407,783 Trade and 57,758 Shop records.

Party Requests: A user can send a Party request to anybody to accomplish a quest together. The user who receives the Party request can accept or reject. Note that a party consists of a few people (upto 6), usually to wage a battle.

3.1.1.2 Group Membership

In Aion, a user can join a guild (or a group), which is friendship based. For this, a member in a group invites her friends or someone who would like to join. A guild member can check the status of other members (of the same guild) by accessing the

roster. A guild is a major element in the social life of online gaming communities [2], and has some similarities with a group in a real society [44]. Hence, the lifetime of a guild is relatively long, which often lasts for months, and the maximum number of members of a guild is 200. In this paper, we analyze a guild and the social interactions among the guild members. From our datasets, we identify 4,955 groups (or guilds) during the measurement period (30,690 and 19,995 users have joined and left the groups during the period, respectively).

3.1.2 Datasets

Aion uses a high-end log system that records every action of every user. We have retrieved all the user records of the *Tiamat* server, one of the 44 servers of Aion, for 91 days from December 21, 2010 to March 21, 2011. Our datasets include 94,497 (anonymized) users, 4,955 groups, 145 million social interactions. The total log size is around 918 GB. We remove the data of the bots from our datasets by using anti-bot scripts provided by NCSoft. We also exclude the groups where the number of members is below three, which leaves us 3,177 groups. We focus on social interactions among users, user affiliation with groups, and user participation such as playtime.

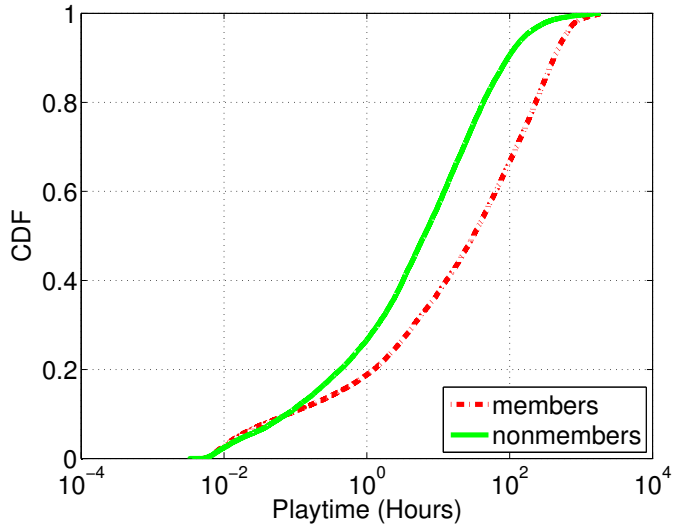
3.2 Group Affiliation

In this section, we investigate (1) how many users join groups, (2) how social interactions are affected by group membership, and (3) how users interact with each other within a group or across groups.

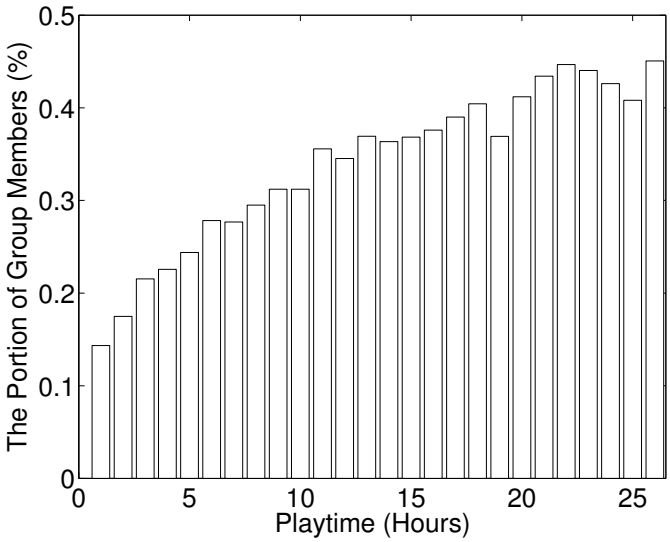
3.2.1 How prevalent are group activities in Aion?

In most of MMORPGs including Aion, a user can join no more than one group at any moment. In our datasets, 29,755 users (30%) among 94,497 users in Aion have ever joined groups. We have investigated 3,177 groups where the average number of users of a group is 8.70. The highest number of users of a group is recorded to be 148.

Figure 1(a) compares the playtimes of both (i) group members and (ii) nonmembers (i.e., users not belonging to any groups). As shown in Figure 1(a), group members play longer than nonmembers. Figure 1(b) plots the portion of group members as user's play time increase. We observe that users who play longer exhibit more group membership. Note that the portion of group members who have played longer than 10 hours (in 3 months) is higher than 50%. These results imply that group membership in MMORPGs may increase users' immersion/indulgence to play longer in a virtual world. This phenomenon is in line with the previous survey reports (e.g., [23,24,45]), which have claimed that a group membership is one of the important motivations for users to indulge in a game and to increase their playtimes by providing a social tie.



(a) Playtime distribution depending on group membership



(b) Portion of group members vs. user's playtime

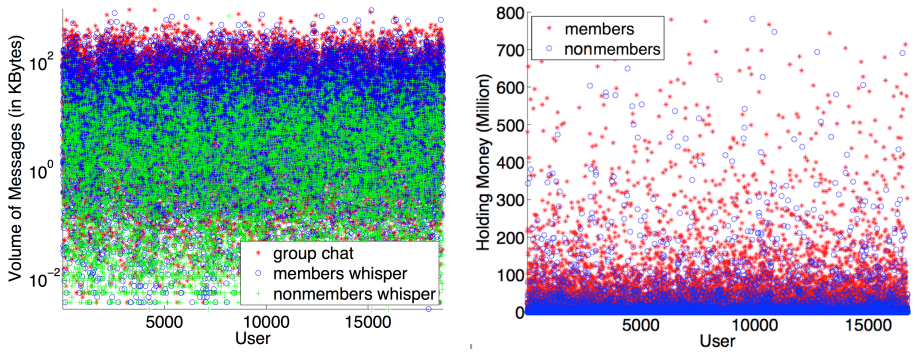
그림 **3.1.** As users spend more time in Aion, they are more likely to be group members.

3.2.2 Effect of Joining a Group

In most of MMORPGs including Aion, a user who wants to join a group should receive an invitation from a current member of the group. In this subsection, we focus on how the users' activities (e.g., communications or economic behaviors) are changed after they join groups. Figure 3.2 shows that communications and economic behaviors become more active after users join groups. We find that the number of private messages (i.e., Whispers) per user is marginally increased after users join groups. However, users having joined groups tend to send a large number of Group chats, which are broadcast only to group members. Thus, the volume of total messages (i.e., the sum of Whispers and Group chats) is increased by three orders of magnitude as compared with that of Whispers only. That is, users much more actively communicate with others by Group chats. We believe that the active communications within a group can be one of the main factors to increase the sense of group attachment, as reported in [23, 24, 46]. Interestingly, the average money of a user increases after she joins a group. This implies that the group membership may also encourage the economical activities in Aion.

3.2.3 Social Interactions Within a Group

We next investigate how users interact with one another within a group. To this end, we calculate the frequencies of social interactions occurring among users in the same group and those occurring among users across groups, respectively. Figure 3.3 shows that the social interactions of users are more active within a group. Note that most of the Friendship requests is two because a user who sends a request typically receives a response. It turns out that the average number of Whispers be-



(a) Communication (b) Economic Behavior

그림 **3.2.** Users' activities (communications and economic behaviors) are boosted after they join groups. Note that the y-axis of the left graph is log scale.

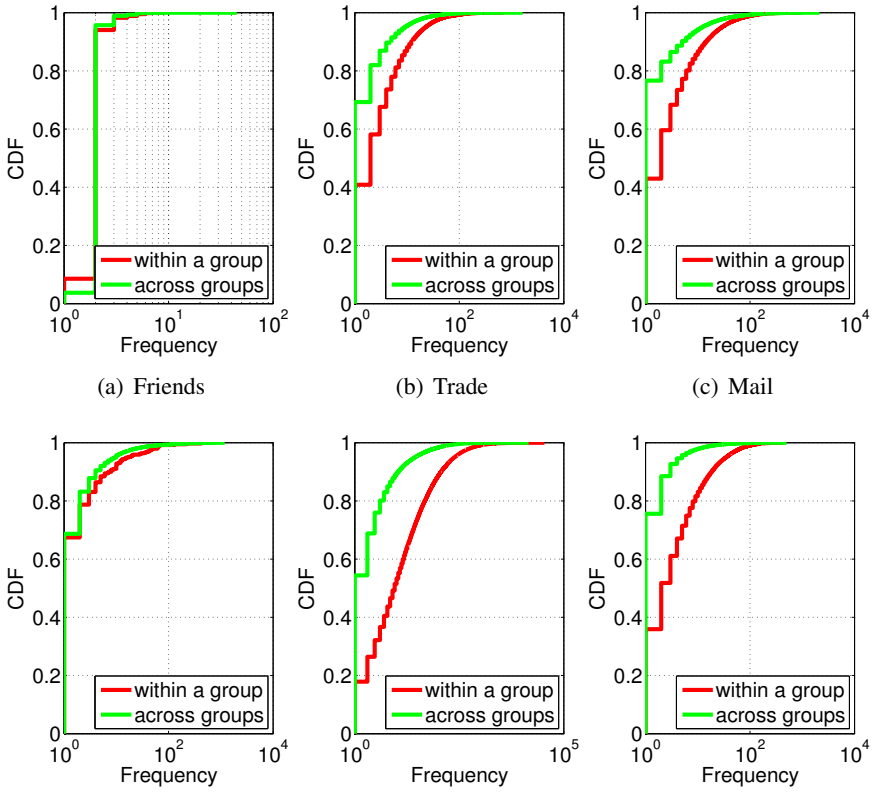
tween two members in the same group is 6.52 times larger than those between users across groups. Through this analysis, we conclude that social interactions occur more actively within a group in Aion. We next investigate the structural patterns of social interactions happening in a group by computing (i) the clustering coefficient [47], and (ii) the reciprocity[2]. Figure 3.4 shows the average clustering coefficient and reciprocity of the six interaction networks in a group, which are compared with those of the interaction networks of the entire users. We find that the average clustering coefficient and reciprocity of most of the social interaction networks within a group are higher than those of the social interaction networks of the entire users, which indicates the structural patterns of social interactions within a group are more likely to be close-knit and reciprocal. We believe that this structural pattern of social interactions in a group is one of the most important factors to give users the sense of group attachment [24].

[2]We define the reciprocity as the portion of bidirectional edges of a user to the total number of her edges.

Interestingly, the clustering coefficient is the strongest in the Party Request interactions in a group. This implies that users in the same group are more likely to play a game together in a cohesive manner. In Figure 4(b), we also find that economical behaviors (i.e., Trade and Shop) are more reciprocal within a group, which means not only communications and battle activities but also economical behaviors exhibit social characteristics (i.e., closely connected and reciprocal) more in a group. Note that the clustering coefficient of the Friend interaction network in a group is very close to that of the entire Friend interaction network. To our surprise, we find that only 4% of total Friendship requests take place among users in the same group. We conjecture that because a member of a group can easily check the status of other group members with a roster without making friends, she does not have to add other members in her friend list.

3.3 Group Dynamics

There have been a few studies to show that one of the important factors of making a user join or leave an online community (i.e., LiveJournal or DBLP) is the relationship among group members [19, 32]. To see the phenomenon of user churning from a holistic viewpoint, we try to focus on how the number of members in a group is rising, falling, or stable. To this end, we classify the groups into three types in terms of vitality as follows: (i) a rising group where the number of its joining users is greater than 120% of that of its leaving users, (ii) a stable group where the number of its joining users is between 80% and 120% of that of its leaving users, and (iii) a falling group where the number of its joining users is less than 80% of that of its

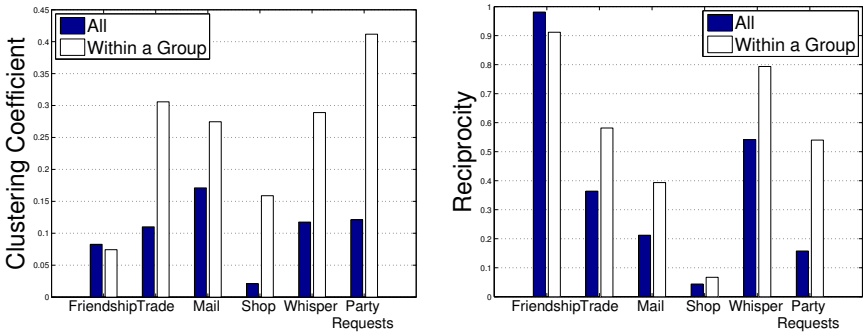


(a) Friends (b) Trade (c) Mail

(d) Shop (e) Whisper (f) Party Request

그림 **3.3.** The frequencies of social interactions among members within the same group and those among users across groups are plotted, respectively.

leaving users. It turns out that there are 1,498 rising groups, 754 stable groups, and 925 falling groups as shown in Table 3.1. We investigate what factors (i.e., group cohesion [48], diversity [23], and spatial/temporal locality) affect the group vitality, depending on group types (i.e., rising, stable, and falling).



(a) Clustering coefficient (b) Reciprocity

그림 **3.4.** Average clustering coefficient and reciprocity of the six interaction networks for group members and for entire users are plotted, respectively.

Type	# of members	# of joins	# of leaves
Rising	23.74 (23.98)	20.26 (20.14)	8.12 (10.38)
Stable	17.22 (20.95)	11.88 (14.18)	11.60 (13.36)
Falling	10.51 (11.31)	3.01 (4.89)	8.99 (9.85)

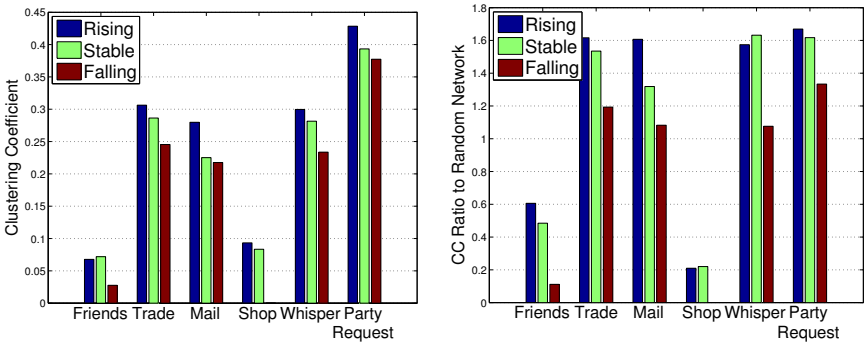
표 **3.1.** Groups are classified into three types: Rising, Stable, and Falling. Averages (and standard deviations) of the numbers of members, joins, and leaves of three group types are shown, respectively.

3.3.1 Group Cohesion

Cohesion usually refers to the tendency of people to be in unity while working towards a common goal in a group [49]. Some studies have reported that users in a cohesive group have more satisfaction than the ones in a non-cohesive group [50]. To understand the cohesiveness of groups in Aion, we investigate the group cohesion from two perspectives: (i) the structural cohesion that focuses on the structural patterns of social interactions and (ii) the interaction scope that indicates whether social interactions take place between the members of the same group or not.

3.3.1.1 Structural Cohesion

In the literature, the cohesion can be measured in various ways [51]; we adopt the clustering coefficient [52] for this purpose in this paper. We further calculate the ratio of the clustering coefficient in the given interaction network to the one in a random network. To this end, we generate 100 random networks each having the same number of nodes and edges with the given interaction network and compute their clustering coefficients, which are averaged. Figures 3.5(a) and 3.5(b) show the clustering coefficient of each social interaction network and the ratio of the clustering coefficient in each social interaction network to the one in a random network, respectively. Since there are only three Shop interactions in the falling groups, we exclude them in this analysis. First, it is worth noting that the clustering coefficients of Friends and Shop networks are lower than those of the corresponding random networks as shown in Figure 5(b). We conjecture that (i) each member in the same group can easily check its member status, without adding their members into her friend list, which makes a lower clustering coefficient in all types of group networks as mentioned in Section 4.3, and (ii) members are exchanging their items or giving them without asking money (i.e., Trade) rather than selling them. We also find that the rising groups show the highest clustering coefficient across the five social interaction networks (except Friends) than other groups, which indicates that social interactions in the rising groups tend to be more clustered. Also, the falling groups show the lowest cohesion, which means members in the falling groups usually do not interact with one another actively. This observation is in line with the prior work [23], which reported that one of the common sources of dissatisfaction in a group results from the social distance.



(a) Clustering coefficient

(b) CC ratio of an interaction network to a random one

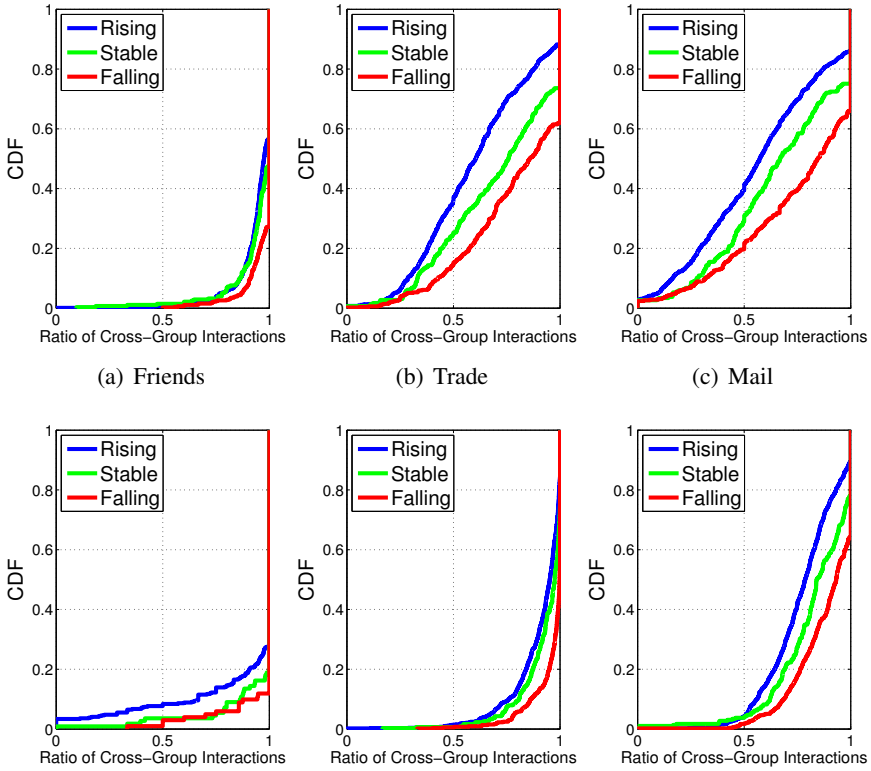
그림 **3.5.** The clustering coefficients (CCs) and the CC ratio values of the interaction networks to random networks of rising and stable groups are higher than those of falling groups across all the interactions.

3.3.1.2 Interaction Scope

From the above analysis of the structural cohesion, we conclude that the interaction patterns of the rising groups are in contrast with those of the falling groups. We further investigate the interaction scopes of the social interactions by looking at whether social interactions happen within the same group or across groups. To this end, we compute how many social interactions occur (i) between members in the same group (i.e., intra-group) and (ii) between a member and a nonmember[3] (i.e., cross-group). Figure 3.6 shows the ratio of the number of cross-group social interactions to the number of all the social interactions happening in the group. We find that rising groups have more intra-group social interactions than cross-group social interactions. Note that the rising groups have more intra-group social interactions than the stable or falling groups, which implies that the rising groups are more cohesive

[3]Note that a nonmember means a member of another group or a user without any group membership.

than the others. In case of the Trade and Mail interactions, a half of their actions take place within the same group for the 40% of groups, which means the prevalence of intra-group interactions. However, in case of the Friendship interactions, even though the rising groups have more intra-group social interactions than the others, we find that most of their Friendship requests are made towards non-members, which is in line with the results of the clustering coefficient in Section 3.2.3.



(a) Friends (b) Trade (c) Mail

(d) Shop (e) Whisper (f) Party Requests

그림 3.6. The ratio of the cross-group social interactions to all the social interactions happening in a group is plotted across the six interactions networks. Members in the falling groups tend to interact more with nonmembers. On the contrary, the rising groups have more intra-group social interactions than cross-group ones.

3.3.2 Group Diversity

In this subsection, we investigate how the social interactions, communications, and economical behaviors exhibit diversity depending on the group's vitality (i.e., rising, stable, and falling).

3.3.2.1 Social Interaction Diversity

Prior studies [23, 25, 45] suggested that one of the motivations to join a group is the social interactions taking place in the group. To examine how diverse social interactions affect the group vitality, we quantify the diversity of social interactions within a group by calculating the Shannon diversity index (or entropy) defined by:

$$H' = -\sum_{i=1}^{A} p_i \ln p_i \tag{3.1}$$

where A is the number of interaction types and p_i is the relative proportion of the i^{th} interaction type among total interactions in a group. Figure 3.7 shows the CDF of the entropy of the social interaction diversity. As shown in Figure 3.7, rising groups have the higher entropy, which signifies the balanced social interactions among the members. In other words, groups whose social interaction types have strong disparity may tend to not grow further.

3.3.2.2 Communication Diversity

We next consider the communication diversity (e.g., whether and how each member has a fair chance to talk in a group.), which may be crucial for group members to mingle. For example, if some members monopolize Group chats, this may

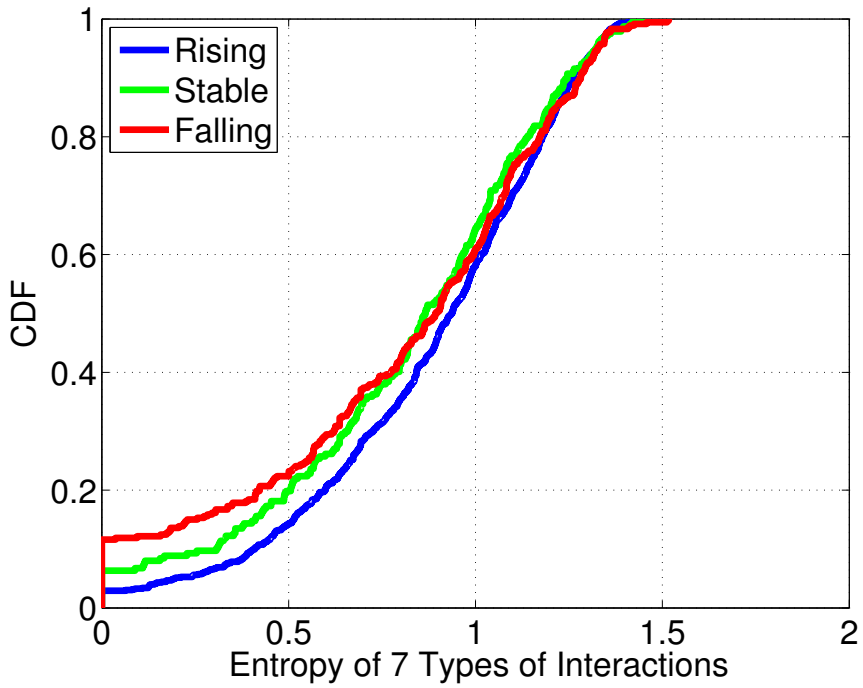
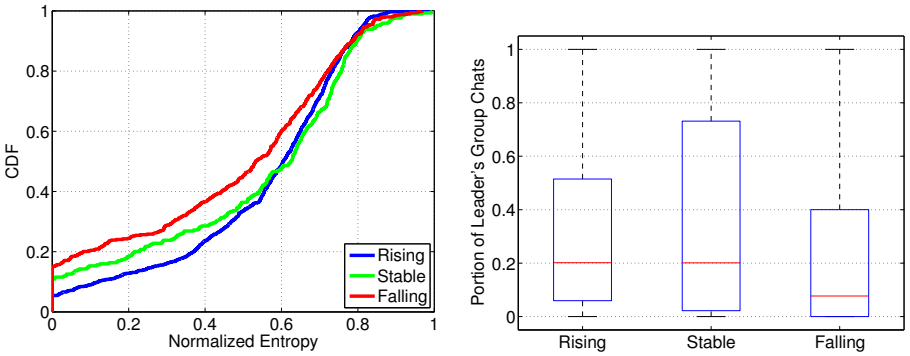


그림 **3.7.** Rising groups exhibit balanced social interactions than other groups.

make some members alienated. We conjecture that the monopolized communications can be a reason to make some members leave the group.

We estimate how users evenly communicate with each other within a group by using the Shannon's entropy. Figure 8(a) shows the normalized entropy of the Group chats in a group. It is worth noting that since every group has a different number of the members, the entropy of Group chats is divided by the maximum entropy, $\log_e M$ where M is the number of members in a group, for normalization purposes. As shown in Figure 8(a), members in rising groups communicate with other members in the same group more evenly than the other groups. [23, 53] showed that one of the motivations to join a group is to chat with various kinds of people. In this sense, it

seems that a balanced communication pattern could lead people to stay in the group and to invite more users into the group.



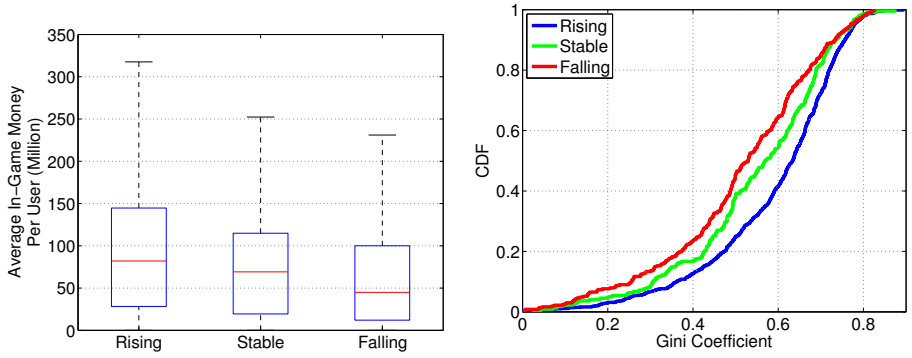
(a) Entropy of Group chats among users (b) Portion of leader's Group chats (in bytes)

그림 **3.8.** Members in rising groups communicate with other members more evenly than the ones in other groups. The portions of leaders' Group chats in stable groups are significant.

In earlier work [23,24], it is pointed out that the role of a leader in a group is critical to keep the community/group alive. For example, one of the reasons for a group to come to an end is a poor leadership of the group. We notice that many of the notifications from a leader are often conveyed through Group chats in Aion. If the portion of the leader's Group chats is high, the leadership may be deemed active in the group. Thus we calculate the portion of the leader's Group chats in each group in Figure 8(b) using boxplots (the red bar is the median). To our surprise, we find that leaders in 40% of the falling groups do not have any conversations in a group, which means the indifference of the leaders. In contrast, in the stable groups, the average value of the portion of the leader's Group chats is 37%, which implies the leaders' activeness while rising and falling groups have 32% and 26% respectively. From Figures 8(a) and 8(b), we find that the low entropy (skewed patterns of dialogues) of the stable

groups is mainly caused by the leaders.

3.3.2.3 Economical Diversity



(a) In-game Money (b) Gini Coefficient

그림 **3.9.** Rising groups tend to have more in-game money than other groups. However, economical behaviors are skewed in rising groups.

One of the key factors of the virtual world attracting millions of users is virtual money in game. By using the in-game money, users can decorate their avatars or buy virtual goods (e.g., weapons and armors) to increase their power and reputation in the virtual world. To understand the economical behaviors of users in groups, we consider not only the amount of the in-game money that individual users have, but also the economical diversity which indicates how the fortune of a group is evenly distributed in the group. To this end, we calculate (i) the average in-game money among users in a group, and (ii) *Gini coefficient* which is a well-known estimator to evaluate the disparity of a distribution in economics [54]. The Gini coefficient is always within the range of [0, 1], where 0 means a perfect uniform distribution and 1 means an extremely skewed distribution [54]. As shown in Figure 3.9, we

find that rising groups have more in-game money than other groups, which indicates that economical activities are more active in the rising groups. However, economical behaviors are significantly skewed in the rising groups. Interestingly, more than 82% of the rising groups, 80% of the stable groups, and 71% of the falling groups exhibit greater Gini coefficients than 0.4. This implies that the whole fortune of a group is skewed to a small portion of users within a group, which will be detailed in Section 3.5.

3.3.3 Group Locality

In this subsection, we investigate how group members are physically closely located in real world or play at similar timeframes depending on the group's vitality (i.e., rising, stable, and falling).

3.3.3.1 Spatial Perspective

To investigate whether (online) group members are located in close (offline) real-world locations, we estimate each user's location by using WHOIS API[4] which is provided by KISA (Korea Internet Security Agency) who manages all the IP addresses in Korea. Based on the location information of each user in a group, we calculate the *spatial locality*, which is defined as the probability that randomly-selected two members (in the same group) have the same locale [55]. Note that we deem that the members have the same locale if they belong to the same city and borough. We assume that physical proximity among (online) members implies a certain degree of

[4]http://whois.kisa.or.kr. This service provides high accuracy since it is managed by Korean government in cooperation with Internet Service Providers (ISPs) in Korea.

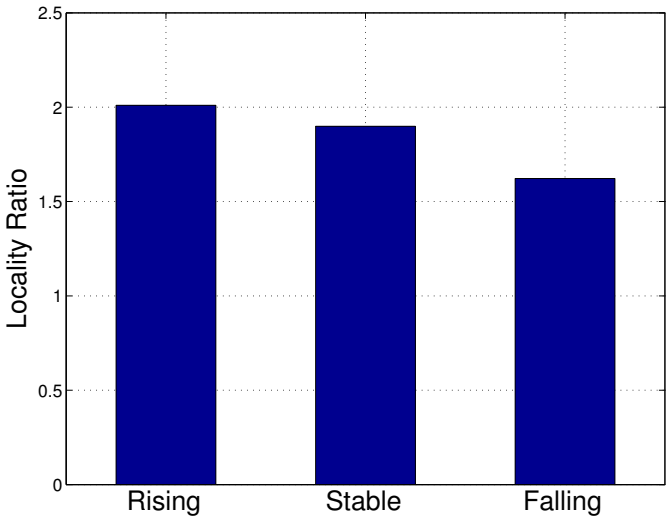


그림 **3.10.** The locality ratio of a real group to a uniform hypothetical one is plotted for each group type. Falling groups show the least spatial locality.

spatial correlation.

To verify if the spatial locality exists, we calculate the locality values of hypothetical groups (with the same numbers of members) whose members are uniformly distributed among all locales. Figure 3.10 plots the average locality of real groups in our datasets divided by that of hypothetical ones depending on the group's vitality. If the calculated value in Figure 3.10 is higher than 1, we can say there is a spatial locality in the group. We observe that the spatial locality of groups is higher than that of hypothetical ones, which indicates that members in the same group are likely to be located in similar places in real world. Interestingly, we find that the locality ratio of the rising groups is larger than that of the other groups. This implies that offline closeness (i.e., spatially correlated distribution of members) exists in growing groups whose social and economical activities are more active than other groups, which may

support the claim that offline bonding can be a factor to be members of the same group [23].

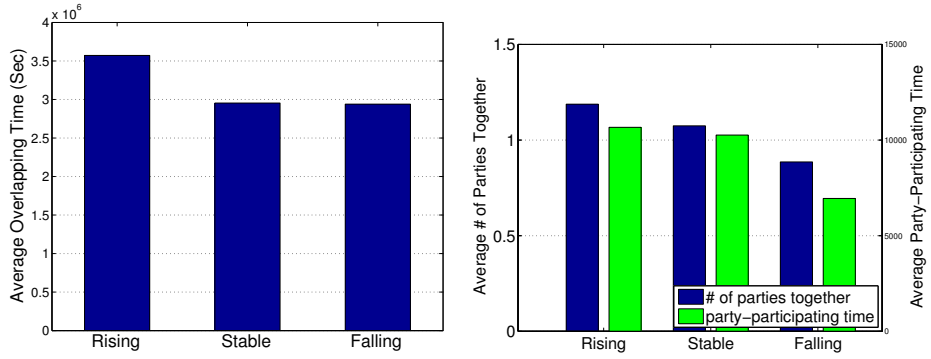
3.3.3.2 Temporal Perspective

We next investigate how group members play games at similar timeframes (i.e., temporal locality of each group). To this end, we calculate two metrics to estimate the temporal locality: (i) the overlapping time (duration) among members of the same group, and (ii) the degree of simultaneous engagement among group members. The first metric is proposed in [2], which can be used to validate that concurrently playing with other members is often mentioned (by users) as an important reason not to leave the group [23]. We estimate the average overlapping time by calculating each member's login and logout times in Figure 11(a). However, from the first metric, it is hard to say whether group members actually have played together. To check the simultaneous play among group members, we focus on a party. A party is formed for a relatively short interval to accomplish a quest (i.e., mostly battles) together by a few users. Note that any users (regardless of group membership) can constitute a party. In Figure 11(b), we plot how many parties have been formulated only by members in the same group and how long a group member has participated in parties (that each consist of the same group members) during the measurement period. Note that the overlapping and party-participating times of a user in a group are normalized by its group size, respectively.

As shown in Figure 11(a), stable and falling groups exhibit the lower overlapping time together than rising groups. Note that falling groups also show the smaller number of parties and the shorter party-participating time than other groups. This

means that even though falling group members are staying together in game, they barely play together, which is in line with the results showing the lower cohesion of groups in Figures 3.5 and 3.6. Meanwhile, stable groups exhibit comparable party-participating time to rising groups, which indicates that members in stable groups are likely to play the game together when they stay online.

In summary, members in rising groups are more spatially and temporally correlated among group members than members of other group types. This implies that not only social or economical behaviors among users in the virtual world, but also the location and time in the real world are important factors for group vitality.



(a) Average overlapping time

(b) Average number of parties together and party-participating time

그림 **3.11.** Rising groups show the highest overlapping time, number of parties together and party-participating time.

3.3.4 Survival Rate

We next investigate how many groups (in our datasets collected in 2010 and 2011) are currently alive (as of Aug. 2013) and active, which is 32 months after from our measurement period. To this end, we have collected current status (e.g., current

Type	Average # of members	# of survived groups	Survival Rate
Rising	14.07	567	37.9%
Stable	12.06	199	26.4%
Falling	8.89	318	34.4%

표 **3.2.** Average number of group members, number of survived groups, and group' survival rate are shown depending on the group's vitality.

number of members or leader's name) of groups from the Aion website[5]. We query the corresponding group information to the website by its name and creation time. If the queried group information is not available, we conclude it has ended.

Table 3.2 summarizes the average number of members and group's survival rate depending on the group's vitality. To our surprise, we observe that 34.1% (1,084 / 3,177) of groups are still alive and active. Note that rising and falling groups still have the largest and smallest number of members even after 32 months have passed, respectively. This signifies that the group vitality does not show substantial changes as time goes on. Interestingly, the survival rate of stable groups is even lower than that of falling groups, which will be detailed in next subsection. Also we will further analyze which factors (i.e., group's cohesion and diversity) affect their survival rate in Section 7.

3.3.5 Dichotomy in Stable Groups

We notice that stable groups, whose rates of joining and leaving users are similar, can be divided into two sub-types: (i) groups where most of joining users are leaving soon within the measurement period and (ii) groups where most of joining users

[5]http://search.plaync.co.kr/aion/

staying during the measurement period. The former type indicates that joining users (i.e., newcomers) of the group cannot mingle with old members, which leads the newcomers to leave soon. In contrast, the latter case means that newcomers usually stay for a while.

To quantitatively differentiate the two sub-types of stable groups, we first introduce a new metric *overlapping ratio*, which is defined as the number of users in the intersection of the joining and leaving users divided by the number of the users in the union of them. That is, if the overlapping ratio of a group is high, users who have joined the group are likely to leave soon. We compare the overlapping ratio among three group types (rising, stable, and falling) in Figure 3.12. As shown in Figure 3.12, the overlapping ratio of the half of stable groups is over the 0.5, which means more than a significant portion of newcomers leave in a short time. To investigate these groups, we quantitatively classify stable groups into two sub-types based on the overlapping ratio: (i) groups having high overlapping ratio (¿, 0.5, *stable-high*) and (ii) groups having low overlapping ratio (¡= 0.5, *stable-low*). Figure 3.14 compares the *stable-high* groups and the *stable-low* groups in terms of group cohesion, diversity, spatial/temporal locality, and survival rate. As shown in Figures 14(a) and 14(b), we observe that the *stable-low* groups exhibit the higher clustering coefficient and Group chat entropy than the *stable-high* groups. This implies that stable groups whose newcomers stay longer have stronger cohesion and more fair participation in conversations. The difference between two sub-types of stable groups is remarkable at their survival rates. Surprisingly we observe that the survival rate of the *stable-low* groups is 2.15 times higher than that of the *stable-high* groups as shown in Figure 14(c). We conjecture that frequent user churning may affect the survival rate eventually.

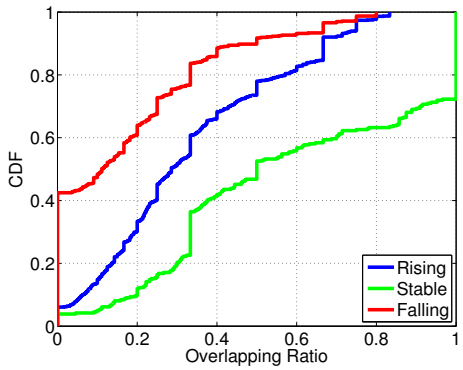


그림 **3.12.** Overlapping ratio depending on group vitality is plotted.

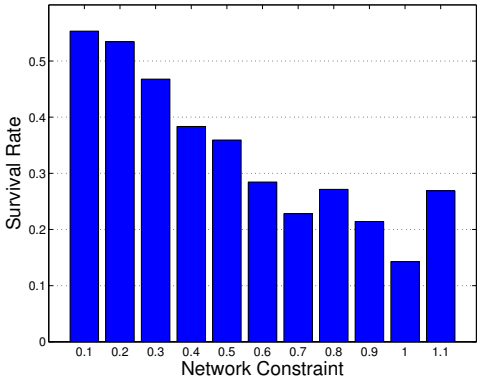
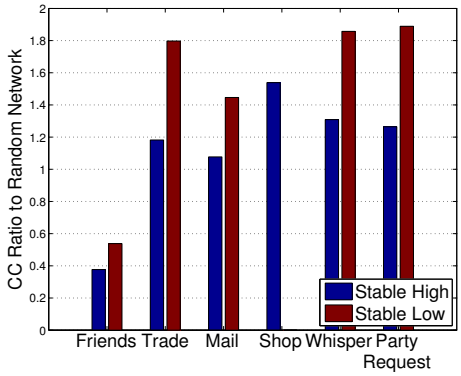


그림 **3.13.** As the network constraint of a group becomes higher, its survival rate is decreased.

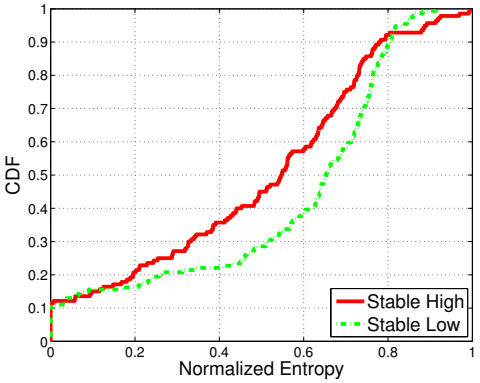
From the spatial and temporal perspectives, we similarly observe that the *stable-low* groups have higher spatial/temporal locality than the *stable-high* groups. This implies that members in the stable groups whose members are spatially/temporally localized are likely to stay longer.

3.4 Group Network

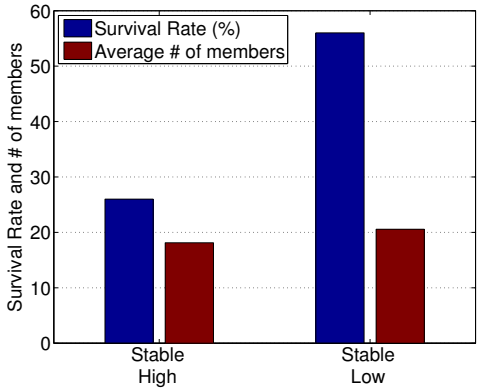
So far, we have investigated various activities which occur within a group mostly. We now turn our attention to the relations among groups or user migration across groups. To this end, we propose a *group network* whose vertex is a group and edge is a relation between two groups. We assume that there is a relation between two groups A and B if there are users who move from group A to group B. More specifically, we define a group network G as a directed weighted graph $G = (V, E, W)$, where V is the set of groups, E is the set of directional edges for migration of users, and W is the



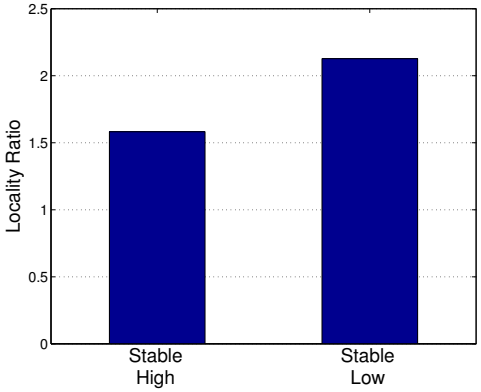
(a) CC ratio of an interaction network to a random one



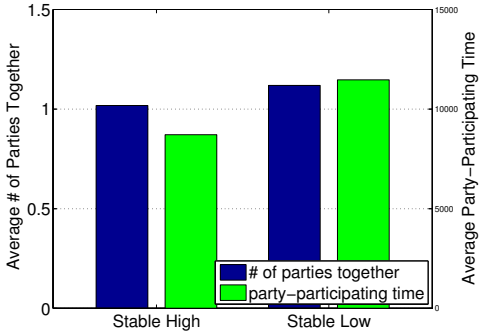
(b) Entropy of Group chats among users



(c) Survival Rate



(d) Spatial locality



(e) Temporal locality

그림 **3.14.** *Stable-low* groups show the higher clustering coefficient and communication diversity than *stable-high* groups. Member churning (and hence the survival rate) is highly related with its entropy of Group chats and spatial/temporal locality.

	# of nodes (edges)	Average degree	CC	Path length
Aion Group	4,022 (17,033)	4.24	0.40	2.26
Facebook	63,730 (817,090)	25.7	0.22	NA
Flickr	2,302,924 (23,838,276)	20.9	0.18	NA
Cyworld	11,537,961 (177,566,730)	30.9	0.16	NA

표 3.3. The main characteristics of the group network in Aion, along with online social networks (Facebook, Flickr, and Cyworld) are presented for comparison purposes.

number of users having migrated between the two groups. That is, if a user moves from group 1 to group 2, there is an outgoing edge from group 1 (or vertex 1) to group 2 (or vertex 2).

3.4.1 Properties of the Group Network

There have been many studies that investigate structural properties of online social networks (OSNs) [4,40,56,57]. Here, we focus on the set of groups (instead of the set of individual users), which are under-appreciated by the research community. Thus, we build the group network to look at the dynamics at group level, while other OSNs are for user level dynamics. That is, an edge between two users in an OSN is usually based on user's relationship (e.g., friendship, followee/follower and so on) while an edge from group A to group B in the group network is set up if a user moves from group A to group B. We believe that investigating the structural properties of a group network is important for understanding how people move across groups and what groups play more important roles in the migration of people across groups.

Table 3.3 summarizes the structural properties of the group network in Aion,

along with those of well-known OSNs (Facebook, Flickr, and Cyworld [56–58]) for comparison purposes. When we compare the group network with the other OSNs, we find that the average degree of the group network is substantially lower than the others. Interestingly, even though the average degree is low, we observe that the clustering coefficient of the group network is substantially high (0.40) compared with those of OSNs. To compare the clustering coefficient (CC) of the group network with those of random networks, we generate 100 random networks based on the Erdös & Rényi (ER) model [59] while preserving the same numbers of nodes and edges. As a result, the average and standard deviation of the CC in random networks are 0.0021 and 0.0002, respectively. We notice that the CC of the group network is significantly higher (190 times) than that of the random network while its average path length is quite small (2.26), which signifies the *small-world* property of the group network. To quantify the 'small-worldness' [60] of the group network in Aion, we calculate the small-world index σ_{SW}, which is defined as:

$$\sigma_{SW} = \frac{\gamma/\gamma_{random}}{\lambda/\lambda_{random}} \tag{3.2}$$

where γ and λ are the CC and the path length of the given group network, respectively, and γ_{random} and λ_{random} are the clustering coefficient and the path length of the random network with the same numbers of nodes and edges as the given network, respectively. If σ_{SW} is greater than one, it means that the given group network has the small-world property [60]. We find that the average and standard deviation of σ_{SW} is 343.56 and 33.96 respectively, which reveals the substantial small-worldness of the group network. It is interesting that the small-world property is found not only at the

level of users [60, 61], but also at the level of groups.

3.4.2 Structural Holes

Category	Factor	Network Constraint
User Dynamics	Interaction Diversity	0.2217
	# of joins	0.5495
	# of leaves	0.5231
Outgo	Group	0.3288
	Per member	0.1264
Fortune/Money	Group	0.0448
	Per member	0.1442

표 **3.4.** Correlation coefficient ρ between network constraint and (i) user dynamics, (ii) outgo, and (iii) fortune/money are shown, respectively. All values are statistically significant (p-value < 0.01).

In sociology, a *structural hole* [62] in a given network is defined as a bridging edge that connects two denser sub-networks, which is similar to the weak tie in Mark Granovetter's theory [33]. The structural holes are often strategically important from social and economical perspectives [63,64]. However, the presence of structural holes in a group network and their characteristics remain unexplored so far.

To find the structural holes in the group network, we first compute the network constraint of every node using Burt's formulation [62]. The network constraint of a group i is defined as:

$$C_i = \sum_j (p_{ij} + \sum_q p_{iq} p_{qj})^2, \quad q \neq i, j \tag{3.3}$$

In our case, z_{ij} is the number of users who move from group i to j. The smaller network constraint a vertex has, the more likely it is connected to a structural hole

(i.e., broker).

Table 3.4 shows the Pearson's correlation coefficient between the network constraint and group activities (i.e., economical behaviors and user dynamics). We first find that the network constraint and the diversity of the social interactions are negatively correlated ($\rho = -0.2217$), which indicates that the groups which are more of structural holes tend to have diverse social interactions. This result is in line with [65], which found that users who tweet diverse topics are likely to be the structural holes. Also, the correlation between the network constraint and the number of joins/leaves is significantly negative, which means many users migrate through the groups corresponding to the structural holes. Interestingly, the groups of structural holes seem to be successful from an economical perspective; they have more fortune than other groups. This result is consistent with [64], which showed that structural holes boost the firm performance in fund companies.

We finally investigate whether the brokerage theory holds in the longevity of groups. Thus we calculate each group's survival rate depending on their network constraints. As shown in Figure 3.13, to our surprise, the survival rate of a group decreases as its networking constraint increases. Hence, we conclude that the brokerage groups are likely to survive longer than other groups.

In summary, we observe the phenomena of the structural hole theory [62]; that is, a weak-tie (i.e., structural hole) can lead to social and economical success by providing access to diverse sources of expertise (i.e., people). By applying the metrics of the brokerage theory into the groups in Aion, we reveal that the brokerage theory holds at the level of groups.

3.5 Implications

In this section, we seek to answer the following questions: (i) what factors make people leave a group? and hence (ii) what makes a group end? To answer the above questions, we adopt a well-known machine learning technique by leveraging the RPART (Recursive PArtitioning and Regression Tree) package [66] in the statistical program *R* to train and test a classifier. Note that we set 70% of the 3,177 groups for a training set and the other 30% for a test set in the training and testing phases of a classifier (which consists of multiple features).

3.5.1 Why people leave groups?

There have been many studies to understand the churning behavior of a user from a group to another for many purposes (e.g., increasing sales in economics [67] or system performances in peer-to-peer networking [68]). In particular, social studies have found that the number of friends is crucial for users to join/leave a group [31,32], but the diverse aspects of social interactions have not been thoroughly investigated. In this subsection, we seek to understand what factors play a key role in making people leave groups. We investigate the churning behaviors from two perspectives: (i) communication patterns and (ii) economical behaviors. We believe understanding the churning behaviors of people with empirically-grounded evidences is important for stakeholders who are to encourage group activities such as social commerce events, social networking services, and MMORPGs.

Feature Categories	Features
Features related to communications within a group	Ratio of numbers of (i) Friendship requests, (ii) Mails, and (iii) Whispers within a group to the ones across groups Number and Entropy of Group chats (frequency & bytes) Clustering coefficient and Reciprocity of (i) Friendship request, (ii) Mails, and (iii) Whispers Ratio of number of leader's Group chats to that of entire members (frequency & bytes)
Features related to economical behaviors within a group	Ratio of income to outgo of a group Fortune (Sum of in-game money of group members) & Gini Coefficient Ratio of numbers of (i) Trades and (ii) Shops within a group to the ones across groups Clustering coefficient and Reciprocity of (i) Trade and (ii) Shops
Features related to the survival rate of a group	Overlapping ratio Spatial locality Average staying time with group members Average number of parties consisting of group members Average party-participating time with group members

표 **3.5.** Features in group characteristics selected for machine learning are listed.

3.5.1.1 Classifier Formulation

We classify the various social interactions in a group into two categories: (i) communication patterns and (ii) economical behaviors. Table 3.5 summarizes the features in the first two categories that we use in this analysis.

3.5.1.2 Results and Discussions

Figures 15(a) and 15(b) show the top three levels of the two classifiers made up of the features of the communication patterns and economical behaviors, respectively. Here, N is the number of groups that are classified according to the classifier tree. The communication-based classifier to characterize groups with high and low churning rates tells us that: (1) if the portion of the number of Mails sent outside the group is relatively high ($> 24\%$), (2) if the number of Group chats is not high ($<$ 8,121), (3) if the Group chats are skewed to a small number of members (i.e., low entropy), the churning rate is the highest (0.900). Note that if the portion of the Mails inside the group is high ($> 68\%$) and the number of Group chats is not low ($>$ 3,136), the churning rate exhibits the lowest.

The economical behavior-based classifier reveals that the high portion of the Trade interactions inside the group ($> 74\%$) makes the churning rate the smallest. To our surprise, however, even with the large fortune of the group ($>$ 3.78 billion), if the distribution of the money is biased (Gini coefficient > 0.8), the group's churning rate is the highest. In summary, these results indicate that the communication patterns and economical interactions in a group are critical for users to stay or leave the group, which complements [28, 29] that have not considered various social interactions.

3.5.2 Why a group ends?

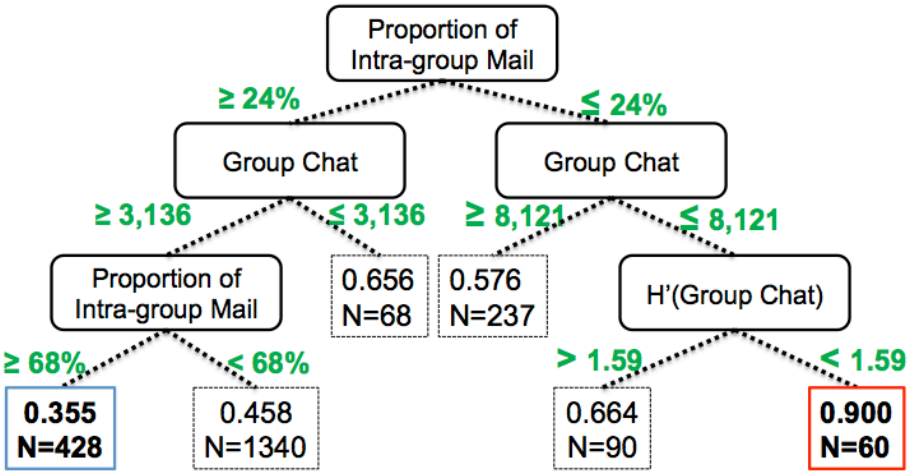
In this subsection, we seek to understand what factors play a key role in making a group end.

3.5.2.1 Classifier Formulation

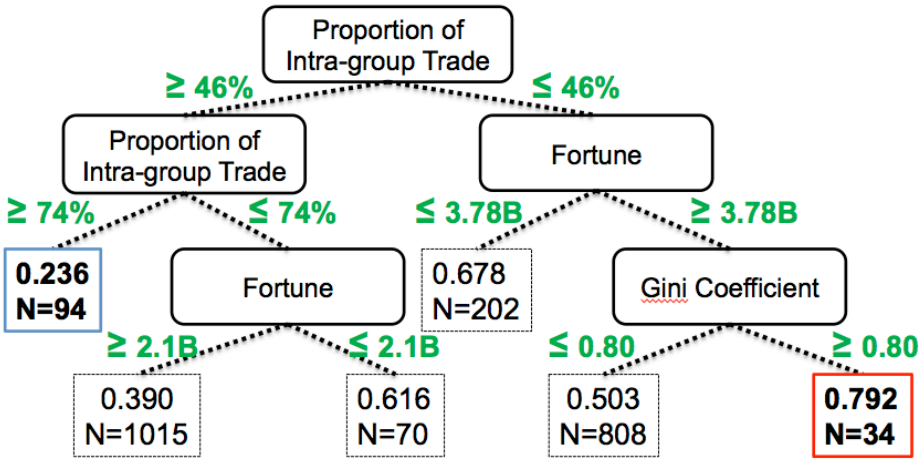
In addition to the features used in the previous subsection, we additionally consider various group properties such as group's spatial/temporal locality and overlapping ratio, which are detailed in Table 3.5. Note that the features in the last category in Table 3.5 are collectively used for constructing the decision tree in this analysis.

3.5.2.2 Results and Discussions

Figure 3.16 shows the top two levels of the classifier made up of all the features in Table 3.5. Interestingly, despite setting the maximum depth of the tree 15 in the partitioning algorithm in RPART, the final tree exhibits only two levels (the average staying time together in the game and the overlapping ratio), which means the two classification criteria are critical for the group's survival rate. We first find that if the average staying time along with members in a group is low ($<$ 150 hours), the survival rate exhibits the lowest (13%), which indicates that simultaneous playing with the group members is crucial for the group's survival. We also observe that (1) if the group's average staying time in the game is high ($>=$ 150 hours) and (2) if the overlapping ratio is low ($<$ 48.4%), the survival rate is the highest (50%). This implies that groups where members stay together in the game and newcomers mingle with old members are likely to sustain.



(a) Communication patterns



(b) Economic behaviors

그림 **3.15.** Top three levels of the decision trees as to communications patterns and economic behaviors are illustrated. The average churning rate is 0.47. *N* is the number of groups for each classification criterion. Note that root node errors of the decision trees for the communication patterns and economical behaviors are 7.5% and 6.6%, respectively.

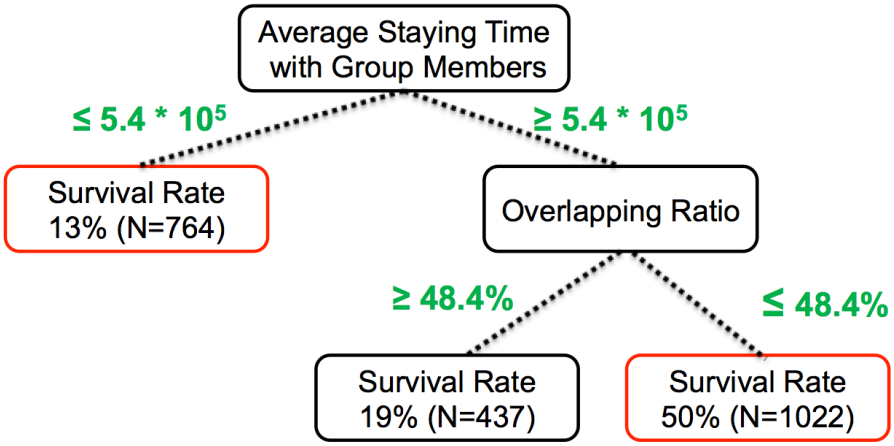


그림 **3.16.** Decision tree for a group survival rate is constructed. N is the number of groups for each classification criterion. Root note error is 21.3%.

Chapter 4

Crowd phenomena of BitTorrent in Spatial and Temporal Perspective

4.1 Methodology

We have conducted a measurement study on one of the most popular BitTorrent portals, The Pirate Bay (TPB). We developed a monitoring client to keep track of swarms by modifying *Azureus* [69]. Figure 4.1 illustrates the overall measurement framework. To monitor each swarm from its beginning, we leverage the RSS notification of a new torrent to retrieve its publisher's username and .torrent file, from which we obtain peers from its tracker and seeds on the distributed hash table (DHT). In addition to the peer list from the tracker and the DHT, a swarm monitoring client further leverages the peer exchange extension (PEX) by which we can discover new peers of the already known peers in a swarm.

4.1.1 Discovering Swarm Topology

Each monitoring client iteratively asks the list of peers from trackers and DHT every 10 minutes. Besides, to discover the topology of each swarm (i.e., how peers in a swarm are connected to each other), the monitoring client further uses the peer exchange extension (PEX) [1]. By analyzing the connectivity among peers from the

[1] Most widely used BitTorrent client software such as uTorrent and Vuze already supports PEX [70]

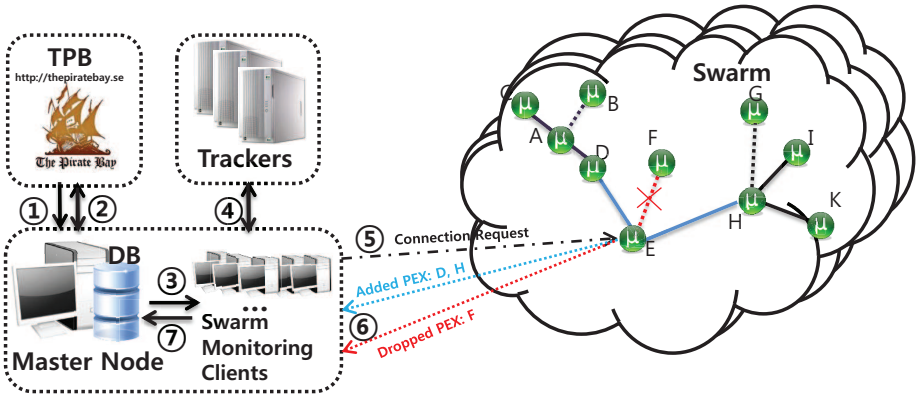


그림 **4.1.** We build a measurement framework to capture the torrent data and user behaviors of a real BitTorrent system.

PEX messages, we can retrieve each peer's (peer-connectivity-level) routing table in a swarm.

When we obtain the routing table of a particular peer by analyzing her PEX messages, an entry in the routing table does not necessarily mean that they are actually exchanging the data. In current BitTorrent systems, a peer can have many connections in the routing table; however, only a small portion of connections are used for exchanging data. To identify the connections that are actually used, we consider two types of PEX messages: PEX-Added and PEX-Dropped [71]. Suppose a swarm monitoring agent monitors peer A. Whenever peer A establishes a connection with a new peer, the monitoring client receives a PEX-Added message from peer A, and thus

	2010	2011
period	April 30 $\sim$ July 23	April 6 $\sim$ May 9
# of unique ip	13,863,126	15,884,221
# of torrents	80,173	63,793

표 **4.1.** Dataset description.

learns when a new peer is added to A's neighbor list. Similarly, a PEX-Dropped message will be sent to A's swarm monitoring agent whenever A's connection to her peer is terminated. We notice that peer A normally drops her connection to peer B because peer B has transmitted little or no data due to B's poor network status or selfish behavior. Hence, a dropped peer within a short duration is unlikely to have exchanged (much) data. Overall, we refine the connections of each peer by removing peers who are dropped shortly during each measurement period. Note that this process is carried out iteratively (e.g., every 10 seconds). Figure 4.1 also illustrates how we identify the peers of peer E. Peers D and H are included, but peer F is removed since it is dropped shortly.

4.1.2 Dataset

Our datasets are composed of two different periods as shown in Table 4.1. For the 143,966 torrents observed during the two periods, the swarm monitoring clients captured snapshots every 30 minutes. The numbers of torrents and IP addresses are described in Table 4.1.

Moreover from the publisher (ID) information on TPB, we divide publishers (in the datasets) into three types, like [72, 73]: (i) *fake* publishers who publish fake

Rank	2010 Country	Portion	2011 Country	Portion
1	United States	14%	United States	14%
2	United Kingdom	7%	India	9%
3	India	7%	United Kingdom	8%
4	Spain	6%	Canada	5%
5	Italy	5%	Korea	4%
6	Canada	5%	Italy	4%
7	France	4%	Australia	4%
8	Sweden	3%	China	3%
9	Australia	3%	Sweden	3%
10	China	3%	Brazil	3%

표 4.2. The portion of Europe is decreased (e.g, Spain: rank 4 → rank 11 and France: rank 7 → rank 15). Asian countries show an increase, e.g., Korea: rank 33 → rank 5.

content, (ii) *profit-driven* publishers who publish content for financial incentives, and (iii) *altruistic publishers* who publish only for sharing. Also, we investigate the locality depending on the different content categories given at TPB: TV, Porn, E-book, Movie, Music, Application, and Game. Lastly, we identified the user's locale using the MaxMind database [74], which maps each IP address (of a peer) to its country or autonomous system (AS) [2]. There are 168 countries and 11,191 ASes in our datasets.

4.1.3 Representativeness

We now analyze the representativeness of the above datasets. Figure 4.2 shows the distribution of peers per continent from 2010 to 2011. The portion of peers in the Europe has decreased from 46% to 38%. The peers in Asia, in contrast, have increased from 25% to 32%. For reference, we also compare the observed distribution

[2]Note that Maxmind exhibits 99.8% accuracy in country-level.

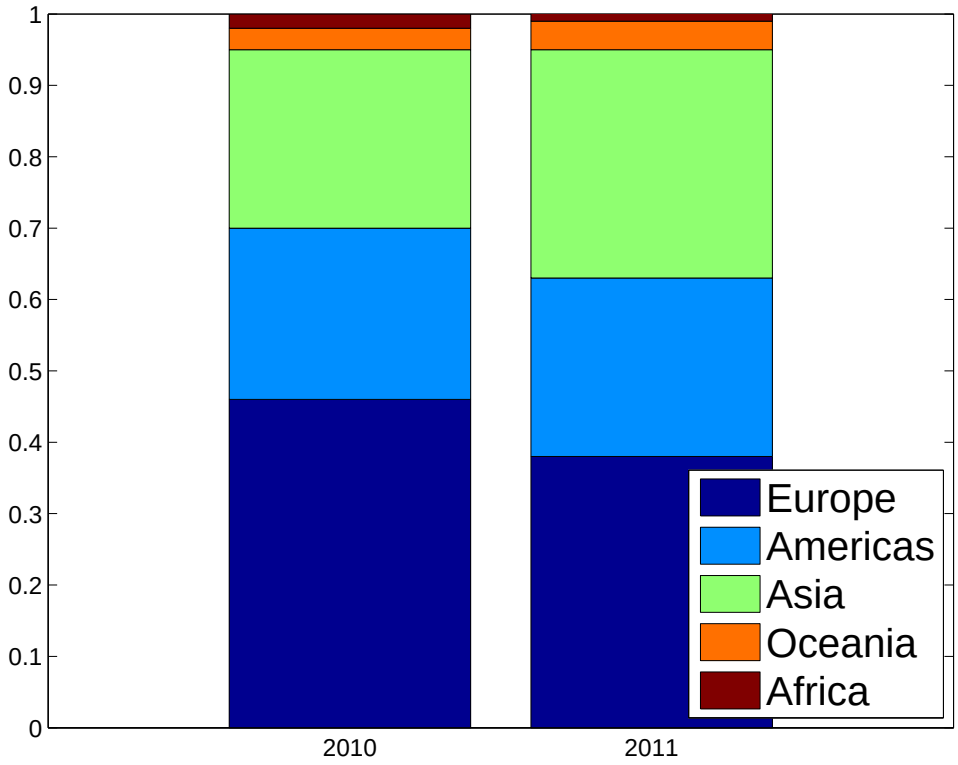


그림 **4.2.** Peer distribution in aspects of continental level is plotted.

with the previous studies [16, 75] and find a similar pattern. For example, declining BitTorrent usage in Europe is also reported in [16], which is aligned with previous reports that European users are increasingly using direct download sites instead of P2P [76]. To investigate the BitTorrent users deeper, we also mapped peers at country level. Table 4.2 shows the top-10 countries from 2010 to 2011 sorted by peer population, and we can see the decline in Europe and growth in America and Asia (e.g, the proportion of Spain, Italy, France, and Sweden are decreased. In contrast, that of India, Korea, and Australia are increased.) Overall, we find little bias of dis-

tribution in our datasets compared with prior work. Below, we analyze this dataset in aspect of content locality.

4.2 Spatial Locality

4.2.1 Locality Metrics

In this subsection, we introduce three metrics for spatial locality: (i) swarm locality, (ii) community locality, and (iii) neighbor locality. Note that these metrics have different searching scopes of BitTorrent peers for calculating the spatial locality; the swarm, community, and neighbor locality consider the whole peers in the given swarm, the peers in the same community, and the neighbor peers respectively.

We first model swarm S as a graph $S = (V, E)$, where V is the set of peers (or nodes) participating in swarm S, $\{v_1, \cdots, v_n\}$, and E is the set of bidirectional edges between peers, $\{e_1, \cdots, e_m\}$. Let $L(v)$ denote the locale (e.g., AS or country) of peer v. We define *swarm locality* as the probability that randomly-selected two nodes (in the same swarm) have the same locale[3]:

$$\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta(L(v_i), L(v_j)),$$

where $\delta(i, j)$ is the Kronecker's delta ($\delta(i, j) = 1$ if $i = j$, and $\delta(i, j) = 0$ otherwise). Here, the swarm locality considers not whether two peers have a connection (or edge), but whether they are interested in the same content.

We devise another metric: *community locality*. A community is a group of peers

[3] In this case, we consider all possible connections among peers regardless of traffic exchanged.

in a swarm, within which connections are denser, but between which connections are sparser. In that sense, we assume that it is more likely to be more traffic inside a community rather than outside of the community. We identify communities using the Louvain method [77], which is a well known algorithm that can quickly find the community and maximize the ratio of the number of edges within communities to that of edges between communities. Community locality is defined as the probability that randomly selected two peers *within the same community* have the same locale. Suppose we have c communities in a swarm, and community k consists of nodes $V^k = \{v_1^k, \cdots, v_{n_k}^k\}$. Then, community locality is

$$\frac{2}{\sum_{k=1}^{c} n_k (n_k - 1)} \sum_{k=1}^{c} \sum_{i=1}^{n_k - 1} \sum_{j=i+1}^{n_k} \delta\left(L\left(v_i^k\right), L\left(v_j^k\right)\right).$$

Since peers within the same community are likely to exchange more chunks directly or indirectly than those in different communities, community locality captures what fraction of interactions among peers (i.e., exchanging pieces) are localized.

We also consider the ratio of the actual number of peer v's neighbors with the same locale to the expected number of neighbors with the same locale assuming purely random assignment of neighbors among all the peers in its swarm [17], which we call *neighbor locality*. For instance, suppose peer v has 100 peers in the same swarm, and 40 of those are in the same locale as v. If v has currently 10 neighbors, with 5 of them in the same locale, then v's neighbor locality is 5/4. If the neighbor locality is close to unity, this means that the number of neighbors with the same locale is almost same as expected one, indicating that the peer selection mechanism exhibits marginal locality. The neighbor locality is somewhat limited in the sense that

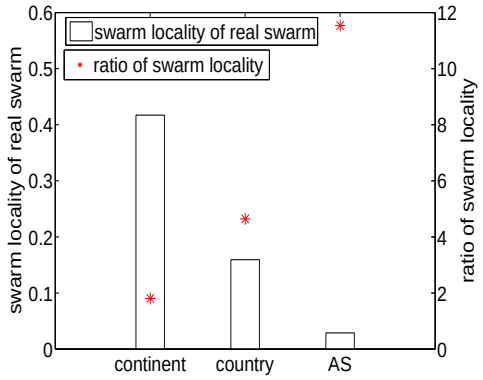


그림 **4.3.** The ratio of the swarm locality of real swarms to that of uniformly distributed hypothetical swarms is plotted.

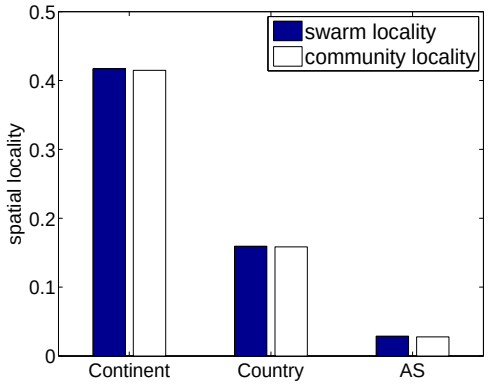


그림 **4.4.** There is no significant differences between swarm locality and community locality.

it considers only direct neighbors.

4.2.2 Swarm, Community, and Neighbor

To see if spatial locality exists, we plot the swarm locality of real swarms, compared with that of hypothetical swarms where peers are uniformly distributed among all locales. Note that the numbers of vertices and edges of the hypothetical swarms are preserved, respectively. In Figure 4.3, we observe that swarm locality of real swarms is significantly higher than that of hypothetical ones. As locale changes from continents to countries and to ASes, the ratio increases from 1.80 times to 4.56 times and to 11.49 times, respectively. This implies that users of the same torrent are spatially biased, which becomes stronger as locale size decreases.

We then examine the effect of BitTorrent's dissemination mechanism on locality. We calculate the neighbor locality for the real swarms, which equals 0.98, 0.98, and

0.99 when the locale is a continent, a country and an AS, respectively. It seems that the peer selection algorithm in BitTorrent contributes little to spatial locality.

For community locality, we first check whether and how swarms make groups (i.e., communities) by calculating the modularity from the Louvain Method [77]. The modularity is defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left[1 - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where k_i is the degree of node i, m is the summation of node degrees for all nodes, c_i is the community where node i belongs, and δ is the Kronecker delta. In general, modularity above 0.3 indicates a strong presence of community structures in a swarm [78]. We find that the average modularity is 0.75, and 70% of swarms exhibit modularity higher than 0.7. Figure 4.4 reveals that the average community locality is similar to the average swarm locality.

In conclusion, even if there is a high locality in the wild (Figure 4.3), there is little difference between a measure which reflects connections (i.e., community locality) and a measure not reflecting connections (i.e., swarm locality) and, moreover, neighbor locality is close to unity. This implies that locality seems not to be much influenced by BitTorrent's sharing mechanism. Instead, we conjecture that locality is more influenced by content itself. Hence, unlike previous studies that investigate the locality only in terms of traffic, our analysis reveals the crucial role of content for better understanding of locality.

4.2.3 Content Categories, Publishers, and Popularity

We now investigate the spatial locality in country level depending on content categories, publishers, and consumers.

Categories: We plot the CDF of swarm locality for each content category in Figure 5(a). We see that torrents in Movie and TV categories have higher swarm locality while the ones in Porn category exhibit lower swarm locality, even though the three categories are all video-centric. We believe that the disparity across the three categories is due to the style of content consumption; movies or video content typically require understanding of content through language and culture, while porn films typically do not need such background.

To further investigate the effect of languages on locality, we examine the Movie torrents in terms of the number of subtitles. Among total 1,597 torrents, 364 torrents have one or more subtitles. We observe that (i) torrents with no subtitles show 46% higher swarm locality than the others, and (ii) as the number of subtitles increases, swarm locality becomes smaller, as shown in Figure 5(b). This is because movies with more subtitles can be consumed by users in more locations (i.e., with different languages), resulting in less spatial locality. This was conjectured in [18], but there has been no empirical study. Moreover, this result is complement to two recent work in OSN [40, 41]; not only in OSN (i.e., Twitter and Facebook) but also in BitTorrent, cultural characteristics (i.e., language) determines the level of content locality.

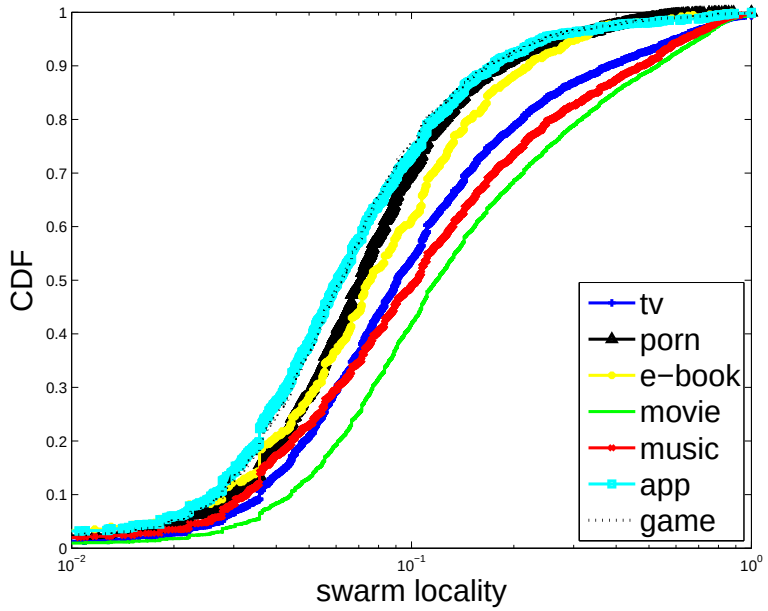
Figure 5(a) shows that torrents in App and Game categories have low swarm locality. For most App and Game torrents, multi-language-support packages are either included in the main program or downloadable from web sites. Hence, language is not an important factor. Also, application and game software often targets global

markets, and thus their torrents are usually downloaded by users without regional inclination. For example, popular software torrents (e.g., Windows 7, Photoshop, or AutoCAD) account for 45% of App torrents.

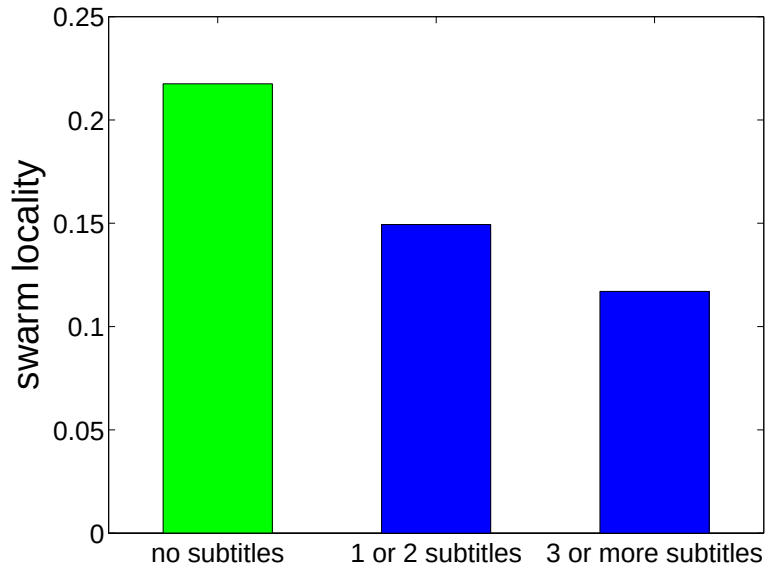
Publisher Types: We examine swarm locality depending on publisher types: altruistic, profit-driven and fake publishers in Figure 4.6. We observe that torrents of profit-driven and fake publishers have lower swarm locality than the ones of altruistic ones. To examine the difference in swarm locality depending on the publisher types, we further analyze it with content categories. We find that porn torrents constitute 39% of the torrents uploaded by profit-driven publishers, aligned with [72], while porn torrents are found in 5% and 2% of the torrents of altruistic and fake publishers, respectively. Overall, the very low swarm locality of Porn torrents explains why profit-driven publishers' torrents exhibit low swarm locality.

We examine torrent titles of fake publishers and find that most torrents of fake publishers have attractive titles like those of latest popular movies. For example, the portion of Movie torrents with titles containing '2011' (i.e., torrents of latest content) but whose publishers are not fake is only 22% (384 out of 1752). On the contrary, among all the Movie torrents of fake publishers, the ratio of torrents whose titles containing '2011' is 60% (117 out of 194). As popular titles of torrents of fake publishers are attractive to users worldwide [72, 73], their naming convention results in lower swarm locality.

Popularity: We investigate the correlation between the number of downloaders (or popularity) of a swarm and its swarm locality by calculating the Pearson's coefficient, which is -0.004. Thus, the swarm locality has no or little correlation with the number of downloaders.



(a) Category



(b) Movie

그림 **4.5.** (a) Swarm locality of each content category is shown. (b) The number of subtitle files affects spatial locality of Movie torrents.

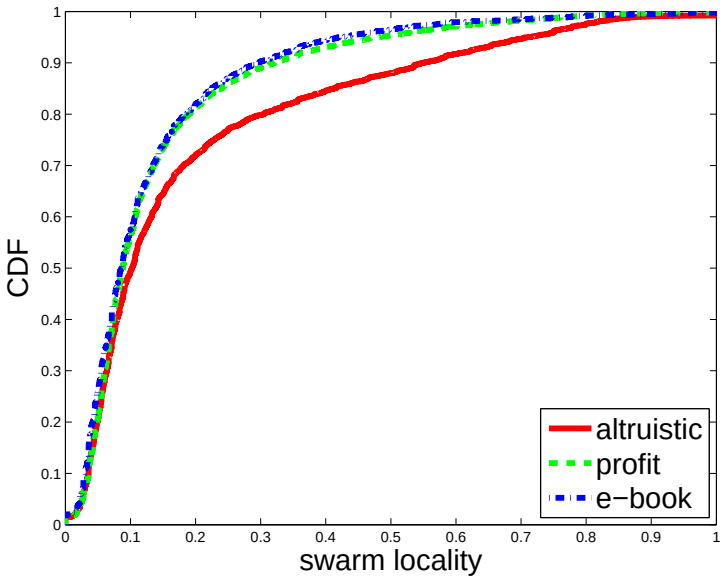


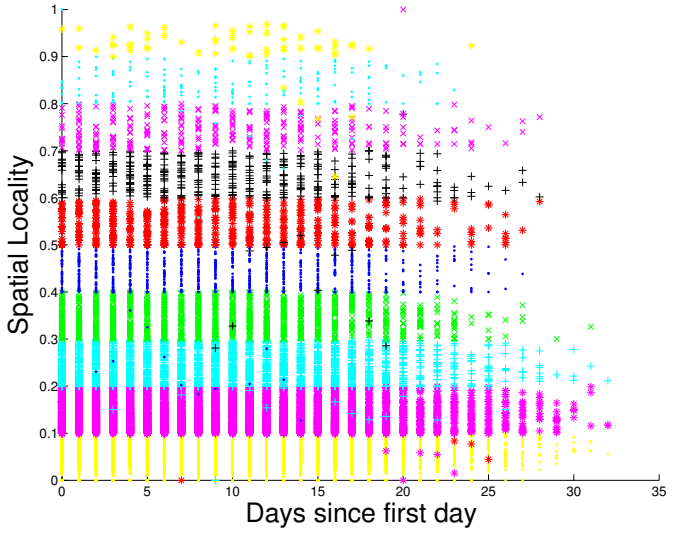
그림 **4.6.** Swarm locality of each publisher type is shown.

4.2.4 Spatial Locality Over Time

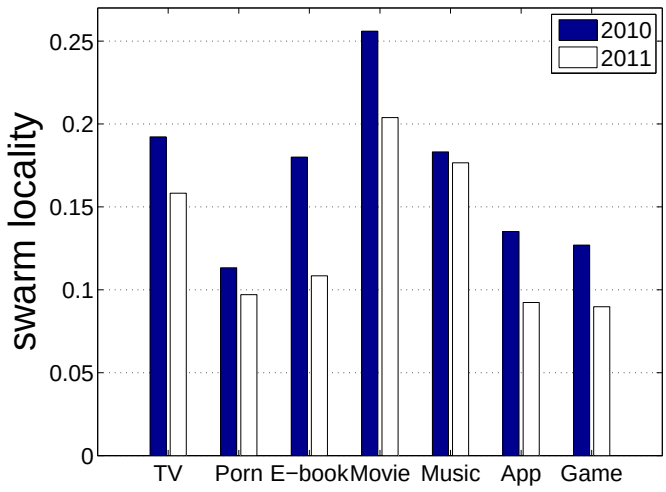
This subsection first analyzes whether and how spatial locality changes over time in Figure 7(a). To this end, we divide the locality into ten bins with a unit of 0.1 and plot localities of each torrent during its life time from the birth (i.e., the published day of the torrent) to death (i.e., the last day at which there is no more seed in the swarm of the torrent). We assign a color to each torrent according to the locality observed in the first day. For example, if the locality (in the first day) of a torrent *"A"* is 0.55, red color is assigned to *"A"*. Even if its locality changes (e.g., 0.13 in the next day), its color is still red which has been assigned in the beginning. Interestingly, spatial locality of a content rarely changes over time as shown in Figure 7(a). Among

32,489 plots in Figure 7(a), only 74 ones have moved from the original bin to another bin. This signifies that content locality has a time-invariant property from a spatial perspective. We believe this property may be helpful for content/network providers in content caching or prefetching. For example, CDN (Content Delivery Network) providers can decide an adequate server to prefetch/cache by exploiting the time-invariant property.

We next compare the spatial localities of 2010 and 2011 to investigate how the spatial locality changes over the years. Figure 7(b) shows the average spatial localities of 2010 and 2011 across diffrent content categories. Interestingly, the spatial locality decreases as years go on; this imples that content sharing patters are increasingly globalized.



(a) Spatial locality on a daily basis



(b) Spatial locality over the years

그림 **4.7.** spatial locality of content rarely changes on a daily basis (microscopic level), but there is notably spatial spread of content consumption over the years (macroscopic level) in country level.

4.3 Temporal Locality

To see how swarm dynamics behave temporally, we define a metric: *daily locality* to indicate the probability that two peers in the same swarm download the torrent in the same day.

4.3.1 Existence of Temporal Locality

Like spatial locality, we find that daily locality of swarms is higher (1.46 times) than that of the hypothetical uniform distribution, indicating that swarm dynamics in terms of population are temporally skewed [79].

We notice that the number of users downloading a torrent tend to be highest for the first day after the torrent is published. As shown in Figure 4.8, positive correlation (0.53) exists between the percentage of first-day downloaders and the daily locality. This means that for all downloads in a given period, a substantial fraction of downloads happen on the first day.

Interestingly, TV and Porn contents exhibit stronger positive correlation (0.71 and 0.72) than others. We conjecture that for TV content, the air dates of TV drama episodes are fixed each week, thus many consumers already expect its publication. For Porn torrents, we conjecture that users usually download porn contents not through searching but through navigating recent contents from the meta-torrent site (e.g., TPB), resulting in older torrents being less likely to be downloaded.

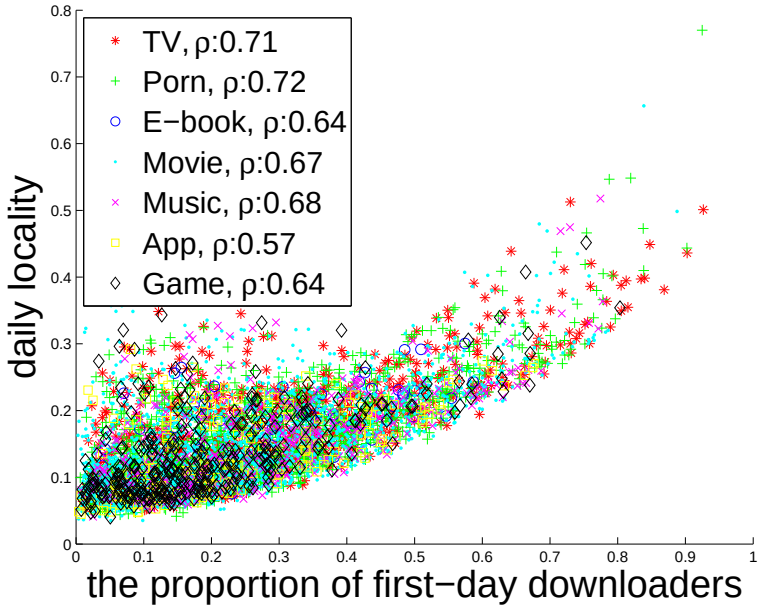


그림 **4.8.** Daily locality according to the proportion of the first-day downloaders is plotted.

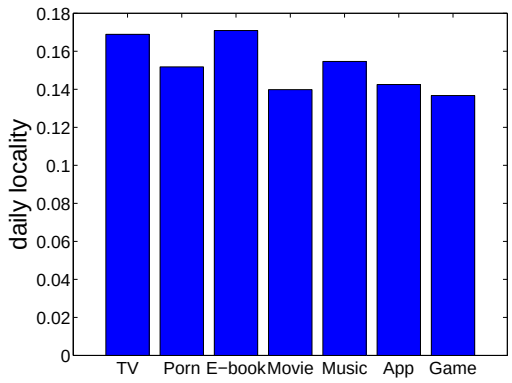


그림 **4.9.** Daily locality is plotted for seven content categories.

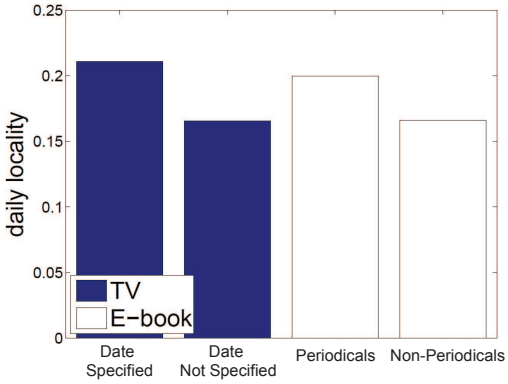


그림 **4.10.** Air dates and publication dates affect temporal locality.

4.3.2 Categories, Publishers, and Popularity

We now analyze the temporal locality of torrents depending on the content categories, publishers, and consumers.

Categories: Figure 4.9 shows the daily locality across the seven content categories. TV torrents exhibit higher temporal locality than torrents of other content categories except for E-book. To investigate why TV torrents show high daily locality, we first analyze their periodic nature by checking whether their titles have the form of 'S**E**', where 'S' and 'E' stand for season and episode, respectively. We find that 58% of titles (of TV torrents) follow this naming convention; for instance, a title of a torrent the drama "Game of Throne: Season 1 - Episode 6" can be "Game of Throne S01E06". The torrents with this naming convention are likely to be published weekly when a new episode is aired. However, even though a torrent has a title of 'S**E**', it does not guarantee that it is published recently. Thus, we further check the torrents whose titles include the air dates of the particular episodes (e.g., The.Daily.Show.2011.03.30). As shown in Figure 4.10, the TV torrents whose titles include the air dates exhibit higher temporal locality (0.211) than the others (0.165).

Interestingly, E-book torrents also show high temporal locality. Our investigation reveals that E-book torrents have high temporal locality since (i) periodicals are published at fixed intervals (weekly or monthly; thus, publication dates are easily expected), and (ii) the lifetime of an E-book torrent is shorter than torrents in other content categories. First, the torrents of E-book periodicals show higher temporal locality than those of non-periodicals as shown in Figure 4.10. Like TV, the periodic nature of periodicals leads to the higher temporal locality of E-book. Second, the average lifetime (i.e., the duration during which at least one seed is alive) of an E-book

torrent is around 8∼9 days, which is shorter than that of a torrent in other categories (around 11∼12 days). The shorter lifetimes of E-book swarms are likely to result in the high temporal locality.

Publisher Types: As shown in Figure 11(a), torrents published by fake publishers exhibit higher temporal locality than others. This is because the administrators of TPB remove fake publishers' accounts and torrents when they are reported as fake ones, which results in shorter lifetimes of their torrents/swarms. Therefore, fake torrents can only be downloaded before they are removed, which results in higher temporal locality. The average lifetime of torrents of fake publishers (9.72 days) is shorter than those of torrents of profit-driven publishers (11.47 days) and altruistic publishers (10.63 days).

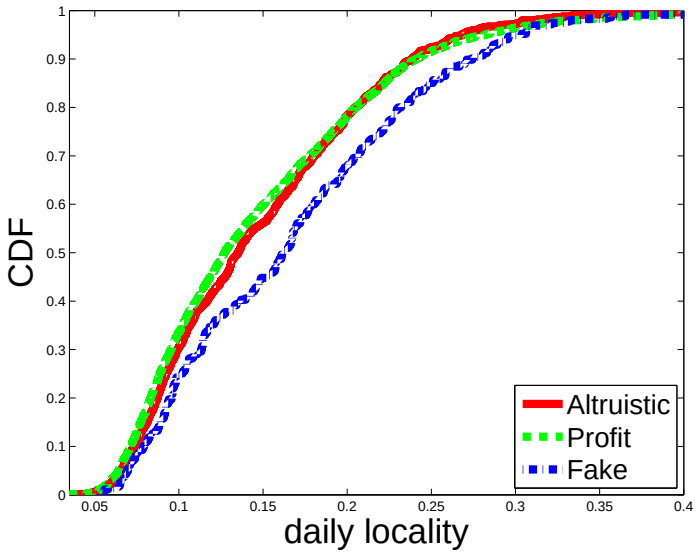
Popularity: We investigate the correlation between the total number of downloaders and daily locality by calculating the Pearson's coefficient. We found that there is a negative correlation (-0.30) between the number of downloaders and daily locality across content categories. Specifically, App (-0.43) and E-book (-0.43) show relatively strong negative correlation while TV (-0.24) shows weak negative one (TV: -0.24, Porn: -0.35, E-book: -0.43, Movie: -0.31, Music: -0.38, App: -0.43, and Game: -0.35).

To understand the disparity of correlation across the seven content categories, we analyze two contrasting content categories: TV (-0.24) and App (-0.43). As shown in Figure 11(b), TV shows a weaker negative correlation compared to the other content categories. This is because a significant portion of TV torrents have a large number of downloaders with high temporal locality. That is why the Pearson coefficient of TV torrents is low, so we take a deeper look at TV contents by accessing the air dates from

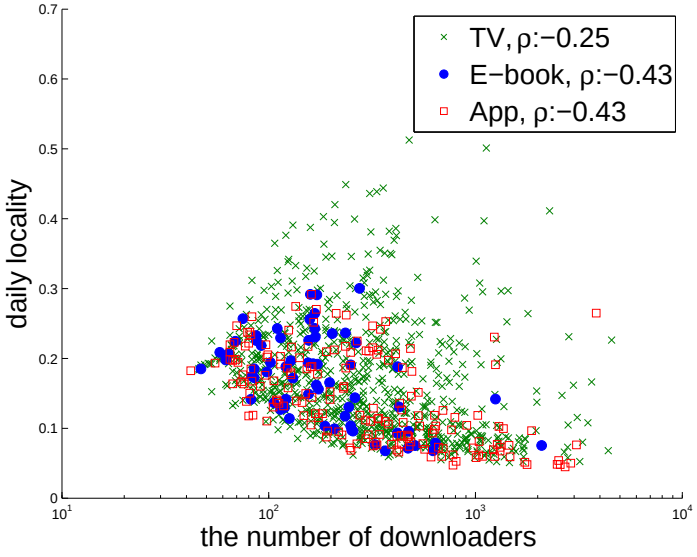
Internet Movie Database (IMDB) [80] manually and found that 94% of torrents are periodical programs (e.g., TV shows or dramas), which 61.2% of users (average 858.9 users) have downloaded within 48 hours after it having been aired. Due to the time-sensitive nature of periodic TV torrents, a majority of people download popular TV torrents early after their air dates, which results in high popularity and daily locality. Consequently, it makes weak negative correlation for TV torrents.

Interestingly, App torrents show a strong negative correlation shown in Figure 11(b). We compare two torrent groups in App: (i) torrents with relatively high daily locality (> 0.2) and a small number of downloaders (< 100) and (ii) torrents with low daily locality (< 0.1) and a large number of downloaders (> 1000). Contradicting the common belief, the torrents of the first group mostly correspond to popular software (e.g., Microsoft Windows, Winzip, or Microsoft Office), and we find that lifetimes of the popular software torrents are relatively short (6 to 10 days in our datasets). To understand the phenomena, we take a look for all the app contents and we find that this is because many publishers upload these popular torrents of the same software frequently which makes users to be more likely to download a recent torrent. For example, 9 torrents of the same software 'Windows 7' are uploaded from April 6 to 10, 2011. On the contrary, the torrents of the second group correspond to software for special customer base (e.g., CAD or graphic tools). We find that the lifetimes of the special-purpose software torrents are substantially longer (> 20 days) than those of other torrents since they are not uploaded frequently, which makes users download their torrents steadily. For example, when we look at 'AutoDesk ECSCAD' from April 6 to May 9, 2011, there is only one torrent. We conclude that not only the time-sensitive nature but also the number of customers of content (i.e.,

general-/special-purpose) affects temporal locality.



(a) Daily locality

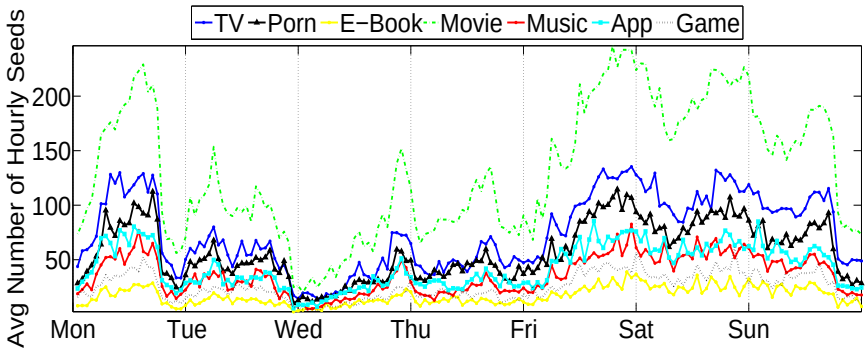


(b) Correlation

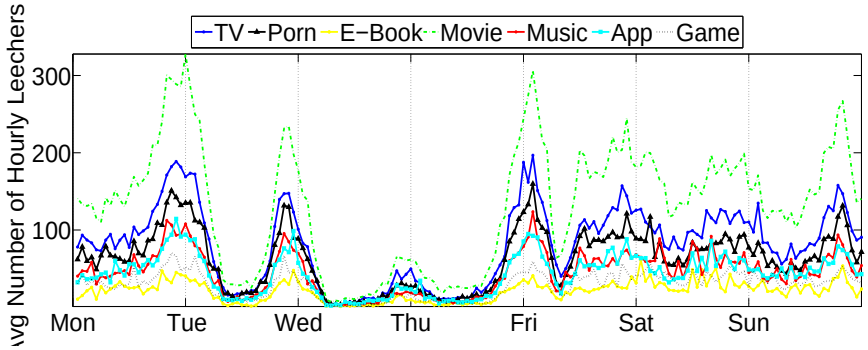
그림 **4.11.** (i) Daily locality of each publisher type is shown. (ii) App and E-book exhibit strong correlation between daily locality and popularity.

4.3.3 Temporal Usage Trends

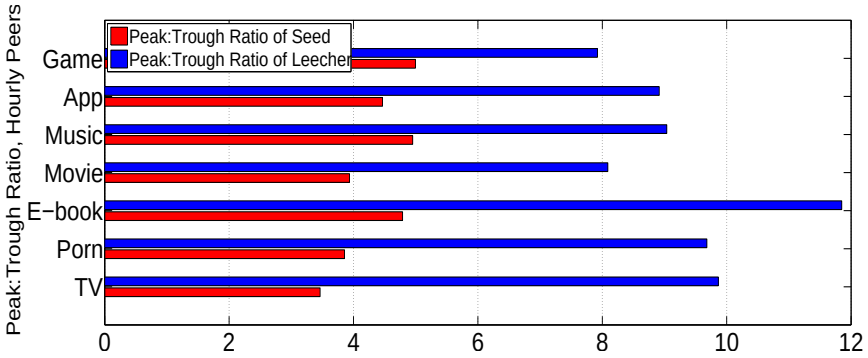
In Figure 4.12, we plot temporal changes of the numbers of seeds and leechers over the week in United States, which ranked first in terms of the number of users in Table 4.2. Moreover, to better illustrate changes of diurnal patterns, we plot the average daily peak-to-trough ratio (i.e., the number of peers at peak devided by that of trough.) in Figure 12(c). We observe a significant diurnal pattern with peak usage in the late evening, which have been already reported in prior work like [16]. In addition, we can find the following interesting patterns. First, the peak-to-trough ratio is relatively higher in weekdays than weekends. We believe this is because people mostly work in a weekday, thus they often download torrents after coming from work (say, 7 pm). Second, the peak-to-trough ratio of leechers are higher than that of seeds. We believe that this because (i) BitTorrent clients tend to keep seeding after completion of downloading by default and (ii) profit-driven publishers tend to keep seeding for their financial gains [81]. Third, we find that both of the Movie and TV torrents have low peak-to-trough of seeds, but the different patterns are observed in leechers. In other words, TV torrents show the relatively higher peak-to-trough ratio of leechers, but Movie torrents exhibit the low peak-to-trough ratio of leechers. We conjecture that the low peak-to-trough ratio of seeds in Movie and TV is due to high popularity (as shown in Figure 4.12). However, because of time-sensitive nature of TV as shown in Section 4.2.3, many users tend to download soon after live broadcast, which results in higher peak-to-trough ratio than Movie torrents.



(a) Average number of hourly seeds for United States



(b) Average number of hourly leechers for United States



(c) Average daily peak-to-trough ratio of hourly peers (seeds and leechers)

그림 **4.12.** Distributions of the number of seeds and leechers, and average daily peak-to-trough ratio of hourly peers consuming the content across the categories in United States in 2011 are plotted. Vertical grid lines in (a) and (b) correspond to midnights in its local time.

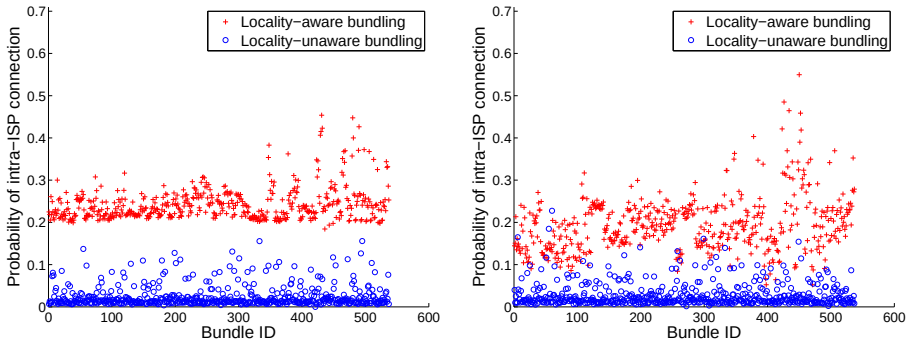
4.4 How to Exploit Locality

Let us illustrate two use cases of exploiting locality: (i) bundling for both improving content availability and decreasing inter-ISP traffic, and (ii) caching for networking efficiency.

Bundling: Bundling torrents in BitTorrent has gained attention [81] since it can mitigate the unavailability problem [82] as well as reduce download times [83]. In bundling, two or more files are disseminated via a single torrent. However, prior bundling approaches may result in substantial inter-ISP traffic because a bundled torrent brings increased file size and swarm-size (i.e., users). However bundling contents mainly consumed in the same region (i.e., ISP) can reduce inter-ISP traffic.

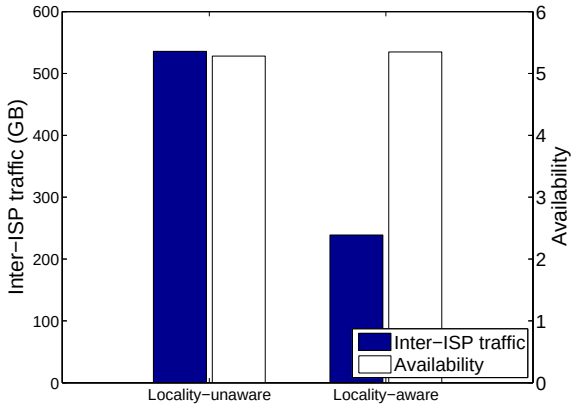
To confirm the aformentioned conjecture, we first make bundles based on data for 17 days from April 1 to April 17 with two different bundling approaches as shown in Figure 13(a): (i) bundling two contents whose locality is high and consumption regions are mainly same (locality-aware bundling) and (ii) bundling two contents randomly but considering availability (locality-unaware bundling) [82]. Then we calculate the spatial locality of (actually) bundled torrents for the following 15-days (from April 18 to May 2, 2011). Surprisingly, we observe that torrents bundled by a locality-aware strategy still exhibit high locality as shown in Figure 13(b), which signifies that we can exploit the nature that the spatial locality changes marginally over time. We next estimate the average content availability of bundled torrents and the inter-ISP traffic after April 17. Figure 13(c) shows that total inter-ISP traffic is significantly reduced (50%) in locality-aware bundling without degrading the availability. From this, we can conclude that locality need to be seriously considered in

bundling.



(a) Bundle locality before April 17

(b) Bundle locality after April 17



(c) Inter-ISP traffic and availability

그림 **4.13.** Total inter-ISP traffic is significantly reduced (50%) in locality-aware bundling without degrading the availability.

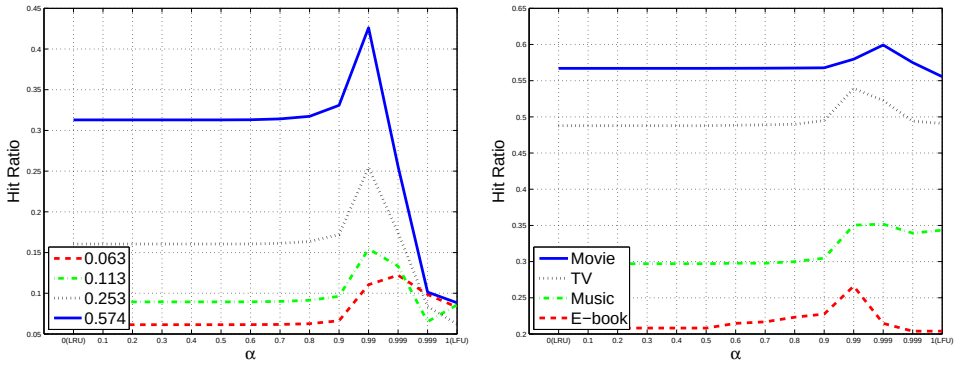
Caching: Recently, information-centric networking (ICN) [84] has gained momentum, and seeks to achieve efficient content distribution. One of the key components in ICN is in-network caching, where how to select content to be cached (and to be replaced) is critical. To investigate how the temporal locality affects in-

network caching performance, we conduct a simulation study by generating four request patterns (each consisting of 100,000 requests) with different temporal locality (i.e., 0.063, 0.113, 0.253, and 0.574) under the same Zipf-like distribution (β=0.63 [85]). We evaluate a caching strategy, *LRFU* [86], which reflects both of the *recency* and *frequency* and its reference count of each cached element is decayed by multiplying α periodically [86]. If we set $\alpha = 0$, LRFU operates like LRU because it reflects only the current reference counts, while if we set $\alpha = 1$, LRFU is reduced to LFU because the past reference counts do not decay. We vary α from 0 to 1 and examine the relation between the temporal locality and the caching performance. As shown in Figure 14(a), the cache hit ratio with high temporal locality is higher than the one with low temporal locality since the request pattern becomes more bursty as the temporal locality is higher. Therefore, we can conclude that caching performance is affected by the temporal locality. Interestingly, the point at which the hit ratio is highest is changes depending on α, which implies that lower α fits for high temporal locality torrents (e.g., TV), while higher α fits for low temporal locality torrents (e.g., Movie).

We also conduct a simulation for the torrents of four categories from real traces[4]: (i) 'TV' with high locality, (ii) 'E-book' with high locality, (iii) 'Movie' with low locality, and (iv) 'Music' with moderate locality as shown in Figure 5(a). As shown in Figure 14(b), the cache hit ratio from the real traces is in line with that of synthetic requests. The highest hit ratio in TV/E-book category is achieved at α = 0.99, while the highest hit ratio in Movie category is α = 0.999 since its locality is relatively low. Since the locality of Music torrents is moderate, the highest point of hit ratio

[4] 147,286 content requests from 'AS0920 National Internet Backbone'

lies in-between. Through this evaluation, we can notice that the caching performance depending on content type is different resulting from different localities even in the same request distribution, which implies the importance for the adaption of temporal locality to cache replacement algorithm such as GreedyDual [87].



(a) Effect of different temporal locality with synthesized requests

(b) Effect of different categories with requests from real trace

그림 4.14. Caching performance (Hit-Ratio) is affected by the temporal locality.

Chapter 5

Summary & Future Work

This dissertation studied crowed phenomena on the Internet by investigating two major applications; i) AION, which is the second most popular MMORPG. and ii) BitTorrent, which is the most popular file sharing application in the Internet.

First, we have comprehensively analyzed the group characteristics in Aion from a socio-economic point of view. Our analysis revealed that structural patterns of social interactions within a group are more likely to be close-knit and reciprocative than those across groups. We also showed that rising groups in terms of number of members exhibit more/higher (i) cohesive social interactions, (ii) balanced communication patterns, (iii) skewed economic behaviors, and (iv) spatial and temporal correlation among group members, compared to the other groups. Based on the machine learning analysis, we revealed interesting findings for group characteristics: (1) if a group is not cohesive, not actively communicating, or not evenly communicating among members, members of the group tend to leave, and (2) if a group's members stay together in the game and its newcomers mingle with old members, it exhibits the high survival rate. Our ongoing work includes investigating (i) the structural patterns of economic behaviors in virtual worlds and how they affect the group activities and social interactions, and (ii) the differences between real-world groups and virtual-world groups from a socio-economic point of view. Our domains of future work lie on analyzing the differences between offline and online behaviors of collective humans.

Secondly, we conducted comprehensive measurements on content locality of BitTorrent. From the datasets, we analyzed: (1) how content is consumed in a spatially and temporally skewed way, (2) what makes the content be consumed differently in spatial and temporal domains depending on its properties, (3) how content consuming patterns changes over the years, (4) how we can exploit the content locality. We observed that content consumption pattern is biased in both spatial and temporal domain. We also observed how cultural factors (e.g., language) affect the spatial locality and how publishing purpose or time-sensitivity also affects the temporal locality of content. We also found that content sharing patterns are increasingly globalized and the diurnal pattern of content usage in leechers is more pronounced than that of seeds. From these observations, we also demonstrated that how spatial and temporal locality can be exploited for bundling torrents and in-network caching. We plan to build the real system (e.g., trackers) to alleviate the traffics flowing between each ISPs by bundling contents dynamically and considering its properties (e.g., category, publishers, or peers) to expect the region where it will be consumed.

참고 문헌

[1] J.-K. Lou and *et al.*, "gender swapping and user behaviour in online social games," in *WWW*, 2013.

[2] N. Ducheneaut and *et al.*, "The life and death of online gaming communities: a look at guilds in world of warcraft," in *ACM SIGCHI*, 2007.

[3] M. Szell and *et al.*, "Multirelational organization of large-scale social networks in an online world," *Proceedings of the National Academy of Sciences*, 2010.

[4] S. Son and *et al.*, "Analysis of context dependence in social interaction networks of a massively multiplayer online role-playing game," *PloS One*, vol. e33918, 2012.

[5] Sandvine: Global Internet Phenomena (Fall 2012).

[6] Karagiannis and *et al.*, "Should internet service providers fear peer-assisted content distribution?" in *ACM IMC*, 2005.

[7] S. Seetharaman and *et al.*, "Characterizing and mitigating inter-domain policy violations in overlay routes," in *ICNP*, 2006.

[8] A. R. Bharambe and *et al.*, "Analyzing and improving a bittorrent network's performance mechanisms," in *IEEE INFOCOM*, 2006.

[9] R. Keralapura and *et al.*, "Can ISPs Take the Heat from Overlay Networks?" in *ACM HotNets*, 2004.

[10] L. Qiu and *et al.*, "On selfish routing in internet-like environments," in *ACM SIGCOMM*, 2003.

[11] D. R. Choffnes and *et al.*, "Taming the torrent: a practical approach to reducing cross-isp traffic in peer-to-peer systems," in *ACM SIGCOMM*, 2008.

[12] Peterson and *et al.*, "Antfarm: efficient content distribution with managed swarms," in *USENIX NSDI*, 2009.

[13] J. Seedorf and *et al.*, "Traffic localization for p2p-applications: The alto approach." in *IEEE P2P*, 2009.

[14] Xie and *et al.*, "P4p: provider portal for applications," in *ACM SIGCOMM*, 2008.

[15] M. Dischinger and *et al.*, "Detecting bittorrent blocking," in ACM IMC, 2008.

[16] J. S. Otto and *et al.*, "On blind mice and the elephant: understanding the network impact of a large distributed system," in *ACM SIGCOMM*, 2011.

[17] Kryczka and *et al.*, "Unrevealing the structure of live bittorrent swarms: Methodology and analysis." in *IEEE P2P*, 2011.

[18] R. C. Rumín and *et al.*, "Deep diving into bittorrent locality," in *IEEE INFOCOM*, 2011.

[19] E. Zheleva and *et al.*, "Co-evolution of social and affiliation networks," in *ACM SIGKDD*, 2009.

[20] Y.-Y. Ahn and *et al.*, "Analysis of topological characteristics of huge online social networking services," in *WWW*, 2007.

[21] H. Kwak and *et al.*, "Fragile online relationship: a first look at unfollow dynamics in twitter," in *ACM SIGCHI*, 2011.

[22] N. Ducheneaut and *et al.*, "Virtual third places: A case study of sociability in massively multiplayer games," *ACM CSCW*, 2007.

[23] D. Williams and *et al.*, "From tree house to barracks: The social life of guilds in world of warcraft," *Games and Culture*, vol. 1, 2006.

[24] E. M. Koivisto, "Supporting communities in massively multiplayer online role-playing games by game design," in *DIGRA Conference*, 2003.

[25] D. Williams and *et al.*, "Behind the Avatar: The Patterns, Practices, and Functions of Role Playing in MMOs," *Games and Culture*, vol. 6, pp. 171–200, 2011.

[26] J. Coleman, *Foundations of Social Theory*. Havard Press, 1990.

[27] A. F. Seay and *et al.*, "Project massive: a study of online gaming communities," in *ACM CHI*, 2004.

[28] A. Patil, J. Liu, and J. Gao, "Predicting group stability in online social networks," in *WWW*, 2013.

[29] A. Patil, J. Liu, B. Price, H. Sharara, and O. Brdiczka, "Modeling destructive group dynamics in on-line gaming communities," in *ICWSM*, 2012.

[30] Y. Huang, W. Ye, N. Bennett, and N. Contractor, "Functional or social?: exploring teams in online games," in *ACM CSCW*, 2013.

[31] S. Wu and *et al.*, "Arrival and departure dynamics in social networks," in *ACM WSDM*, 2013.

[32] L. Backstrom and *et al.*, "Group formation in large social networks: membership, growth, and evolution," in *ACM SIGKDD*, 2006.

[33] M. S. Granovetter, "The Strength of Weak Ties," *The American Journal of Sociology*, vol. 78, 1973.

[34] P. Sarkar and A. Moore, "Dynamic social network analysis using latent space models," *SIGKDD Explorations: Special Edition on Link Mining*, 2005.

[35] Legout and *et al.*, "Clustering and sharing incentives in bittorrent systems," in *ACM SIGMETRICS*, 2007.

[36] D. R. Choffnes and *et al.*, "Strange bedfellows: community identification in bittorrent," in *IPTPS*, 2010.

[37] C. L. Viles and J. French, "Content locality in distributed digital libraies," *Information Processing & Management*, vol. 35, 1999.

[38] C. L. Viles, “Content locality in time-ordered document collections,” 1999.

[39] A. Brodersen and e. Scellato, “Youtube around the world: geographic popularity of videos,” in *ACM WWW*, 2012.

[40] H. Kwak and *et al.*, “What is twitter, a social network or a news media?” in *WWW*, 2010.

[41] M. P. Wittie and e. Pejovic, “Exploiting locality of interest in online social networks,” in *ACM CoNEXT*, 2010.

[42] J. M. Pujol and e. Erramilli, “The little engine(s) that could: scaling online social networks,” in *ACM SIGCOMM*, 2010.

[43] D. Jang, “Aion: A blockbuster two years after its launch,” 2010, http://www.etnews.com/news/detail.html?id=201011100087.

[44] N. F. Johnson and *et al.*, “Human group formation in online guilds and offline gangs driven by a common team dynamic,” *Phys. Rev. E*, vol. 79, 2009.

[45] V. H.-H. Chen and *et al.*, “Enjoyment or engagement? role of social interaction in playing massively mulitplayer online role-playing games (mmorpgs),” *IFIP*, vol. 4161, 2006.

[46] D. McMillan and D. Chavis, “Sense of community: A definition and theory,” *Journal of Community psychology*, vol. 14, 1986.

[47] D. J. Watts and S. H. Strogatz, “Collective dynamics of ’small-world’ networks,” *Nature*, vol. 393, pp. 440–442, 1998.

[48] D. M. Paskevich and *et al.*, “Relationship between collective efficacy and team cohesion: Conceptual and measurement issues.” *Group Dynamics Theory Research and Practice*, vol. 3, 1999.

[49] L. R. B. Albert V. Carron, “Cohesion: Conceptual and measurement issues,” *Small Group Research*, vol. 31, 2000.

[50] M. A. Hogg, *The social psychology of group cohesiveness: from attraction to social identity*. Prentice-Hall, 1992.

[51] E. Kolaczyk, "Statistical analysis of network data: Methods and models," *Springer Series In Statistics*, p. 386, 2009.

[52] M. E. J. Newman, "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.

[53] N. Ducheneaut and R. J. Moore, "The social side of gaming: a study of interaction patterns in a massively multiplayer online game," ser. ACM CSCW, 2004.

[54] C. Dagum, "The generation and distribution of income, the Lorenz curve and the Gini ratio," *Economie Appliquée*, vol. 33, no. 2, 1980.

[55] T. Chung, J. Han, and *et al*, "Spatial and temporal locality of content in bittorrent: A measurement study," in *IFIP NETWORKING*, 2013.

[56] H. Chun and *et al.*, "Comparison of online social relations in volume vs interaction: a case study of cyworld," in *ACM IMC*, 2008.

[57] J. Ugander and *et al.*, "The anatomy of the facebook social graph," *CoRR*, vol. abs/1111.4503, 2011.

[58] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *WWW*, 2009.

[59] P. Erdös and A. Rényi, "On random graphs i," *Publicationes Mathematicae Debrecen*, vol. 6, p. 290, 1959.

[60] M. D. Humphries, K. Gurney, and T. J. Prescott, "The brainstem reticular formation is a small-world, not scale-free, network," *Proceedings of the Royal Society B: Biological Sciences*, vol. 273, no. 1585, pp. 503–511, 2006.

[61] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, no. 6684, pp. 409–10, 1998.

[62] R. S. Burt, *Structural holes: The social structure of competition.* Harvard University Press, 1992.

[63] R. S. Burt and *et al.*, "The social capital of french and american managers," *Organization Science*, vol. 11, 2000.

[64] A. Zaheer, "Benefiting from network position: Firm capabilities, structural holes, and performance." *Strategic Management Journal*, vol. 26, 2005.

[65] D. Quercia and *et al.*, "The social world of twitter: Topics, geography, and emotions." in *ICWSM*, 2012.

[66] T. M. Therneau and *et al.*, *rpart: Recursive Partitioning*, http://CRAN.R-project.org/package=rpart, 2011.

[67] H.-S. Kim and *et al.*, "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market," *Telecommunications Policy*, vol. 28, 2004.

[68] D. Stutzbach and R. Rejaie, "Understanding churn in peer-to-peer networks," in *ACM IMC*, 2006.

[69] "Open sourced bittorrent client: vuze," http://www.vuze.com.

[70] D. Wu and *et al.*, "Understanding peer exchange in bittorrent systems," in *IEEE P2P*, 2010.

[71] "Bittorrent peer exchange conventions," http://wiki.theory.org/BitTorrentPeerExchangeConventions.

[72] R. C. Rumín and *et al.*, "Is content publishing in bittorrent altruistic or profit-driven?" in *ACM CoNEXT*, 2010.

[73] S. Kim and *et al.*, "Content publishing and downloading practice in bittorrent," in *IFIP Networking*, 2012.

[74] "Maxmind," http://www.maxmind.com/.

[75] C. Zhang, P. Dhungel, and *et al.*, "Unraveling the bittorrent ecosystem," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, pp. 1164–1177, 2011.

[76] G. Maier, A. Feldmann, and P. *et al.*, "On dominant characteristics of residential broadband internet traffic," in *ACM IMC*, 2009.

[77] V. D. Blondel and *et al.*, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, 2008.

[78] H. Kwak and *et al.*, "Mining communities in networks: a solution for consistency and its evaluation," in *ACM IMC*, 2009.

[79] M. Izal and e. Urvoy Keller, "Dissecting BitTorrent: five months in a torrent's lifetime," in *PAM*, 2004.

[80] "Internet movie database," http://www.imbdb.com.

[81] J. Han and *et al.*, "Bundling practice in bittorrent: What, how, and why," in *ACM SIGMETRICS*, 2012.

[82] D. S. Menasche and *et al.*, "Content availability and bundling in swarming systems," in *ACM CoNEXT*, 2009.

[83] S. Zhang and *et al.*, "Dynamic file bundling for large-scale content distribution," in *IEEE LCN*, 2012.

[84] A. Ghodsi and *et al.*, "Information-centric networking: seeing the forest for the trees," in *ACM HotNets*, 2011.

[85] A. Abhari and *et al.*, "Workload generation for youtube," *Multimedia Tools Application*, vol. 46, no. 1, 2010.

[86] D. Lee and *et al.*, "Lrfu: A spectrum of policies that subsumes the least recently used and least frequently used policies," in *ACM SIGMETRICS*, 2001.

[87] S. Jin and *et al.*, "Greedydual* web caching algorithm," *Computer Communications*, 2001.