Unveiling the Adoption and Cascading Process of OSN-based Gifting Applications

M. Rezaur Rahman
University of California-Davis
Davis, CA, USA
Email: mrrahman@ucdavis.edu

Jinyoung Han
University of California-Davis
Davis, CA, USA
Email: rghan@ucdavis.edu

Chen-Nee Chuah
University of California-Davis
Davis, CA, USA
Email: chuah@ece.ucdavis.edu

***Abstract*—This paper demystifies the adoption and cascading process of OSN-based applications that grow via user invitations. We analyze a detailed large-scale dataset of a popular Facebook gifting application, iHeart, that contains more than 2 billion entries of user activities generated by 190 million users during a span of 64 weeks. We investigate: (1) how users invite their friends to an OSN-based application, (2) what factors drive the cascading process of application adoptions, and (3) what are the good predictors of the ultimate cascade sizes. We find that sending or receiving a large number of invitations does not necessarily help to recruit new users to iHeart. We also identify a set of distinctive features that are good predictors of the growth of the application adoptions in terms of final population size. Finally, based on the insights learned from our analyses, we propose a prediction model to infer whether a cascade of application adoption will continue to grow in the future based on observing the initial adoption process. Results show our proposed model can achieve high precision (over 80%) in iHeart as well as in another OSN-based gifting application, Hugged.**

I. Introduction

The word-of-mouth information diffusion [21] over online social networks (OSNs) has been regarded as an important mechanism by which people discover and share a new idea, technology, or product. Such diffusion mechanism has gained great attentions not only from research communities but also from enterprises interested in growing their business. According to the recent New York Times article, traditional retailers like *Target* and *Walmart* have sought partners in OSN companies [12].

The growing interest in such information diffusion mechanism has attracted the research community to investigate how information (e.g., a photo, link, or product) are *reshared* in OSNs like Facebook, Twitter, Flickr, or Pinterest [4], [5], [10], [15], [21]. A news, photo, link, or product may get reshared (e.g., retweet in Twitter or repin in Pinterest) multiple times over the OSNs, hence generating a *cascade* that potentially reaches a large number of users. These studies have revealed valuable insights into the macro-level propagation patterns of messages [15], URLs [21], news [16], and photos [4], [10]. However, most of these studies paid little attention to how user recommendation of *products* leads to their actual adoption and further propagation. This is the key to understand and predict the growth, popularity, and longevity of OSN-based products or the ultimate size of the user base.

In this paper, we strive to shed lights on these issues by performing a detailed investigation of the adoption process of a gifting application on Facebook. In particular, we seek answers to the following questions: How do people invite their friends to use an OSN-based application? What factors drive a new user to adopt an invitation? How often does a recruited user recruit others to the application? Overall, how does the application adoption evolve over the OSN? What are the main drivers that determine a large, sustainable cascade of a new application? Can we predict the growth of the cascading process of an application?

It has been reported that Facebook gifting applications only rely on user-invitations (through sending gift items) to recruit new users [19]. Hence, we believe that by focusing on the gifting genre, we can minimize the effect of exogenous factors, including user recruitment through other channels such as email broadcasts and paid advertisement. As a case study, we examine the launching and spreading of *iHeart*, which was one of the top gifting applications in Facebook in terms of number of (monthly) active users as of Dec. 2009. We analyze a detailed large-scale dataset that contains more than 2 billion entries of user activities generated by 190 million users during a span of 64 weeks. A user in *iHeart* can send a gift to other existing iHeart users, or send a gift with an invitation message (referred to as *invitation*) to his/her friends who have not installed the application yet. In the latter case, the friend who has received the invitation can either accept (i.e., install the application) or ignore it. Note that a user can receive multiple invitations from different friends until he/she accepts the application. If a user accepts the invitation and gets activated, he/she can also send invitations to other friends. Finally, these inter-user communications through invitations form a sequence of adoptions generated by a contagion process where a node "recruits" its connected nodes with some probability. We refer to this sequence of adoptions as a *cascade*. Using the dataset, we analyze: (1) how users invite their friends to adopt an OSN-based application, (2) what factors drive the cascading process of application adoptions, and (3) distinctive features associated with different sizes of cascades. Based on our findings, we develop and evaluate a prediction model to infer whether a cascade of application adoption will continue to grow in the future, which gives an important implication in resource allocation of OSN-based product stakeholders, e.g., via targeted marketing.

We highlight the main contributions as follows:

- **Invitation and adoption behaviors (Section IV):** We found that sending or receiving a large number of invitations does not necessarily help to recruit new users, which may be due to the spamming phenomenon where users end up ignoring the invitations.
- **Growth patterns of application adoptions (or cascades)(Section V):** We analyzed the growth patterns of the application adoption process in iHeart from a graph-theoretical perspective. We found a set of

distinctive features (e.g., structural or evolutionary properties of cascades, and roles of their seed nodes) that are good predictors for the final population size of the cascade. To attain large cascades (e.g., beyond 100 users), effective user recruitment has to continue beyond the initial growth phase. In general, we observed that early adopters (seeds and their recruited users) play a diminishing role as cascade sizes increase.

- **Predicting large cascades (Section VI):** We explored the implications of our findings for predicting the attainable growth of different application adoption cascades. We found that the evolutionary properties of cascades (e.g., initial growth rate) are good predictors of the ultimate cascade sizes. By observing these features during the initial adoption process, our classifier can achieve over 80% precision for predicting the large cascades of application adoption. Our learning-based prediction model performs well when applied to another gifting application, *Hugged*, which suggests that our prediction model is generally applicable to gifting genre of OSN-based applications where the new users are recruited primarily through user invitations.

The rest of the paper is organized as follows. We first review related work in Section II. Section III describes our cascading model that captures the application adoption process and our datasets. Sections IV and V present our measurement-based characterization of the inviters, invitees, and application cascade process. Based on insights learned from our analyses, we build and evaluate a learning-based prediction model to forecast the growth of the cascading process in Section VI. Finally, we conclude the paper in Section VII.

II. Related Work

Information diffusion over OSNs: There have been many studies that investigate the patterns of information diffusion in OSNs, revealing valuable insights into characteristics of information spreading [3]–[5], [8], [10], [15], [21]. Rodrigues *et al.* analyzed the patterns of word-of-mouth URL exchanges in Twitter and found that users with geographically similar locations tend to share the same URL [21]. Goel *et al.* [8] investigated the cascades of URLs shared in different microblogging services such as Yahoo! or Twitter. They showed that multi-step cascading is not frequently observed; the vast majority of the diffusions occur within one hop from a seed node. Similar characteristics are also observed in photo-sharing OSNs like Flickr. Cha *et al.* [4] reported that even popular photos do not spread widely in Flickr. We focus on how *applications* are recommended, adopted, and propagated over the OSNs, which is the key to understand the growth, popularity, and longevity of OSN-based applications or products.

Some of the studies have focused on what drives information diffusion in OSNs. There have been increasing attention to effectively identify special individuals, often referred to as the "influentials" [2] who can recruit a large number of users, leading to a large-scale cascade [1]. Cha *et al.* found that the most influential users (in terms of generating retweets) are not necessarily the most followed ones [3]. Wang *et al.* showed that information diffusion depends on the social context or the users' characteristics [24]. On the other hand, Han *et al.* showed that pin propagation in Pinterest is mostly driven by pin's properties like its topic, not by user's characteristics such as the number of followers [10]. While these studies have investigated key drivers to spread URLs, news, or images, we focus on what factors drive the adoption of new OSN-based applications, which have not been thoroughly investigated.

Popularity or cascade size prediction: There have been great interests in predicting the future popularity of a given content, information, or product to increase sales. For example, content or web service providers like YouTube may want to estimate the view counts (or download counts) of their content in advance, which are directly tied to their *ad* revenue [23]. This has led to the investigations of the popularities of Digg stories [23], information on Twitter [26], and YouTube view counts [23]. Also, several studies have focused on how to predict the growth of information cascading in Twitter [14] and Facebook [5]. Cheng *et al.* showed that temporal and structural features are key predictors of the size of photo reshare cascade in Facebook [5]. In this paper, we predict the success/failure of a seed node (or early adopters) to create a large-scale cascade of application adoption using a learning-based model and insights gained from our measurement studies. Such model has great utility in resource allocation of product stakeholders, e.g., via targeted marketing.

User interactions in OSN-based applications: Facebook has become one of the most popular OSN and supported a rich application ecosystem that has become a $6+ billion industry (as of 2012) [19]. This has motivated numerous studies on the popularity of Facebook applications [7], social structures of Facebook games [13], a probabilistic growth model for user activities on Facebook gifting applications [19], roles of users to propagate Facebook applications [25], and a model of the evolution trend of Facebook gifting applications [17]. Recently, Rahman *et al.* investigated the role of the seeds in the adoption process of the Facebook gifting application, and showed that large-scale adoptions are not correlated to well-known properties of seeds (such as out-degree or active lifetime) [20]. In this paper, we perform a detailed analysis of how user invitations lead to the adoption and propagation of OSN-based applications. In addition, we apply the insights learned from our measurement study to develop a learning-based model to predict whether a cascade of application adoption will continue to grow in the future.

III. Methodology

In this section, we describe how new applications are adopted and propagated over OSN platform, using Facebook third-party gifting applications as a case study. We first define a cascading tree generated by a particular seed node. Then we describe our datasets.

A. Definition of cascading tree

To analyze how a gifting application is adopted (and propagated) in Facebook, we define a cascading tree as a directed graph, $G = (V, E)$, where V is the set of users and E is the set of invitations to adopt, based on the Krackhardt's hierarchical tree model [11]. That is, if user j adopts an application from user i's invitation, there exists an edge $E(V_i, V_j)$ from user V_i to user V_j. We refer to this event as *adoption* of an application. After a user adopts the application, he/she will in turn send

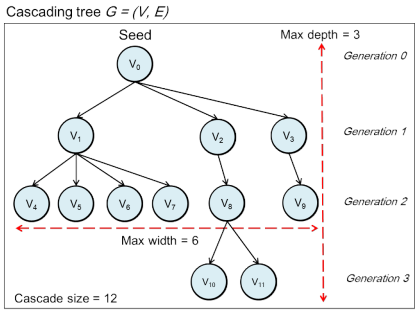


Fig. 1. Definition of cascading tree.

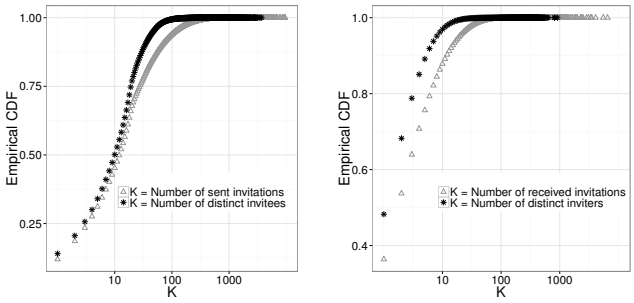
out invitations to his/her friends, thereby recruiting more users. A sequence of adoptions is denoted by a *cascade*. Note that a not-yet recruited user can receive invitations from multiple friends before installing the application. The leaf nodes in the cascading tree are the users who have received one or more invitations but are (i) silent, i.e., have never installed or used the applications, or (ii) ineffective, i.e., have sent out invitations but not successful in recruiting new users.

Fig. 1 illustrates an example of a cascading tree. The circles represent users who belong to the cascading tree G. A cascading tree is initiated by a seed user (say, v_0 in this example) who already installed the application (before receiving any invitations) and have sent invitations. The *max depth* in a cascading tree is defined as the hop count from the root (i.e., seed) to the farthest leaf node in the tree. The *max width* refers to the number of nodes in the widest generation in the tree. In Fig. 1, the size, max depth, and max width of the cascading tree G are 12, 3, and 6, respectively. Since we focus on adoption process of applications, we consider only the cascading trees that have at least one invitation-induced adoption, i.e., whose sizes are at least 2.

Since a not-yet recruited user can receive multiple invitations, his or her set of potential parents is a group of senders from which the user received invitations before his/her adoption. To identify the parent who is responsible for the adoption of a particular user, we first calculate each sender's (or inviter's) success ratio, τ, which is defined as:

$$\tau_i = \frac{A(i)}{N(i)} \tag{1}$$

where $A(i)$ is the number of activated users who received an invitation from i, and $N(i)$ is the number of total users who received an invitation from i. Since the sender whose τ is the highest is the most probable sender (or inviter) responsible for the adoption of a given user, u, we finally select that sender as the parent of u in a same cascading tree. Note that Rahman *et al.* [20] evaluated four other definitions of parents: i) the first parent who sent the invitation for the first time, ii) the last parent who sent the invitation lastly, iii) the parent who sent invitations most, and iv) the random parent. We find that our method based on the success ratio shows similar distributions of cascade sizes with the four methods in [20], but we use this method (which is based on the success ratio) since we are studying the application adoption process and are interested in identifying the most effective sender (or inviter) who succeeds in recruiting the most number of new users.



(a) Inviters' activities (b) Invitees' activities

Fig. 2. Frequency of invitations sent and received in iHeart.

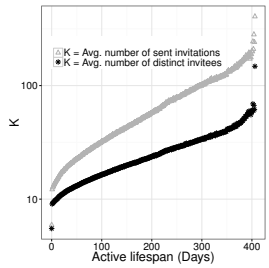
B. Datasets

We conduct a detailed measurement-based characterization of how a popular gifting Facebook application, *iHeart*, was adopted and propagated over Facebook platform. iHeart was launched in June 2009 and operated by Manakki, LLC at the time of data collection. The dataset was anonymized and donated by Manakki, LLC, which covers the major lifespan of the application (64 weeks since its initial launch until it declines in popularity) and contains more than 2 billion (2,027,488,267) entries of user activities generated by 189,989,307 users. Each user activity record includes the sender's and recipient's anonymized Facebook user IDs as well as the timestamp when the sender sends an invitation to the recipient. In this paper, we do not consider all the interactions among users (including sending and receiving a specific type of heart), but we focus on analyzing 'invitations' sent to the new users until they first become activated. To analyze the adoption process of iHeart, we first identified the *seed* nodes, i.e., users who already installed and used the applications without receiving any prior invitations/gifts from other users. We managed to uncover 2,384,686 cascading trees (i.e., cascade sizes are bigger than 1), each rooted at a unique seed. Together, these cascades contain 186,254,526 recruited users. Note that there are another 1,350,095 users who installed iHeart without prior invitations but have not recruited any users in our dataset, and hence are excluded from our study.

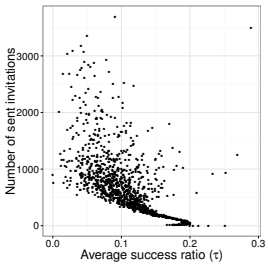
For the purpose of verifying the universality of our prediction model (in Section VI), we use another dataset from a different Facebook gifting application, *Hugged*, which was launched in February 2008 and also operated by Manakki, LLC at the time of data collection. The dataset, also anonymized and donated by Manakki, covers the major lifespan of the application (69 weeks from February 2008) and contains 114 million entries of user activities generated by 34,525,693 users. We analyzed 461,510 cascading trees generated by unique seeds and contain 33,967,052 recruited users. Note that there are another 97,131 seed nodes that fail to recruit any users and are excluded from our study.

IV. Case Study: Invitation-based Adoption and Propagation of iHeart

An iHeart user sends invitations to recruit his/her friends and some of them may accept the invitations. This results in an inviter-invitee relationship. For example, if user A sends an invitation to a friend, B (who has not yet adopted iHeart), then A becomes the inviter of B and B becomes the invitee of A. In this subsection, we explore (i) invitation patterns of inviters

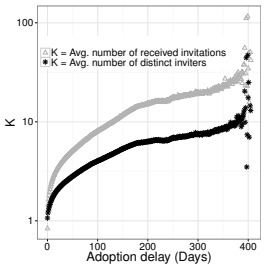


(a) Active lifespan of an inviter and her activities

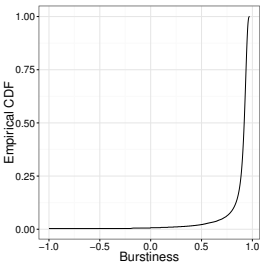


(b) Average success ratio vs. number of invitations

Fig. 3. Characteristics of inviters in iHeart.



(a) Adoption delay vs. receptions



(b) Invitees' reception burstiness

Fig. 4. Adoption behaviors of invitees in iHeart.

and (ii) adoption behaviors of invitees, both of which are key to the successful growth of iHeart user base.

A. Invitation patterns of inviters

Fig. 2(a) shows the distributions of the number of invitations sent by an inviter and the number of unique invitees targeted by the inviter, respectively. We find that more than 50% of the inviters send invitations to 10 or more distinct invitees. Moreover, 50% of the inviters send more than 13 total invitations. Note that the top 1% of inviters (in terms of number of sent invitation) send more than 84 invitations.

To investigate how long an inviter is active (i.e., using iHeart to send invitations to recruit new users), we calculate the *active lifespan* of each inviter as the time difference between the first and the last invitation sent. During the active lifespan, an inviter may send multiple invitations. We investigate the correlation between the active lifespan of the inviters and the number of invitations that are sent to their friends in Fig. 3(a). As shown in Fig. 3(a), the inviters with longer active lifespan send more invitations and invite more unique invitees. For instance, inviters with longer than 300 days of active lifespan send more than 100 invitations on average. Notice that the inviters with at least a day of active lifespan send more than 12 invitations to more than 9 distinct invitees on average.

We next explore the inviter's performance in terms of the success ratio τ introduced in Eq. 1. We observe that 3% of the inviters exhibit τ of 0, which means they may have zero impact on the growth of iHeart. However, many of these inviters with τ = 0 (who have not succeeded in recruiting new users) send a large number of invitations (the average number of invitations sent by them is 5.37, with a maximum of 892 invitations). This indicates that sending a large number of invitations does not necessarily mean that the inviter will succeed in recruiting large number of new users. In fact, the inviters sending many invitations tend to have a low average success ratio, as shown in Fig. 3(b). On the other hand, the inviters who send a small number of invitations (perhaps to more targeted recipients) show large success ratios. Note that the Pearson correlation coefficient between the number of invitations of inviters and their average success ratio is -0.612.

B. Adoption behaviors of invitees

We next investigate the adoption behaviors of invitees. Fig. 2(b) shows the distributions of the number of invitations received by an invitee and the number of unique inviters from which an invitee receives invitations, respectively. We find that 86% of the invitees receive less than 10 invitations, but 1% of invitees receive more than 46 invitations. We also investigate the time gap between the first time a user gets an invitation and the time a user gets activated, which we refer to as *adoption delay*. For example, if a user gets an invitation at $t1$ for the first time and she gets activated at $t2$, the adoption delay is $t2 - t1$. We find that the adoption delays of 29.7% of the (activated) users are less than a day; this includes the users who promptly accept the invitations after receiving them. We also observe that most (85%) of the users have less than 100 days adoption delays, but around 2.6% of the users have adoption delays longer than 200 days. Fig. 4(a) shows the correlation between the adoption delay and the average number of invitations received by an invitee. Invitees with longer adoption delays tend to receive more invitations and are targeted by more unique inviters.

We also investigate how regularly an invitee receives multiple invitations, i.e., the arrival patterns of the invitations. To this end, we use the *burstiness parameter B* [9], [22], which can be calculated as:

$$B = \frac{\sigma_X - \mu_X}{\sigma_X + \mu_X} \tag{2}$$

where X is a set of number of invitations in each time slot, and σ_X and μ_X are the standard deviation and the mean of X, respectively. That is, $x_t \in X$ where $X = \{x_1, x_2, x_3, \ldots, x_t\}$ indicates that an inviter receives x_t invitations during the given time slot t. We set the time slot as 1 hour. Note that the burstiness parameter B has values in the range between -1 and 1, where $B = -1$ means periodic activities, $B = 0$ means random activities (Poissonian), and $B = 1$ means completely bursty dynamics [22]. We plot the CDF of the burstiness of the invitations received by the invitees in Fig. 4(b). As shown in Fig. 4(b), we find that burstiness values of 90% of the invitees are mostly close to 1, which means the most of invitees receive invitations in a significantly skewed fashion, e.g., in a burst or batch, without any regularity.

Finally, we explore how many invitations (or inviters) motivate the invitees to adopt iHeart. To this end, we calculate the probability of adoption after receiving n invitations as:

$$p(n) = \frac{N_a(n)}{N_a(n) + N_{na}(n)} \tag{3}$$

where $N_a(n)$ is the number of invitees who received n invitations and got activated, and $N_{na}(n)$ is the number of the invitees who received n invitations but never got activated. Similarly, we can calculate the adoption probability $p(m)$ for m inviters by replacing n in invitations to m inviters. We plot $p(n)$ and $p(m)$ against n and m in Fig. 5, respectively, where n stands for the number of invitations, and m is the number of distinct inviters. As shown in Fig. 5, $p(n)$ and $p(m)$ decrease

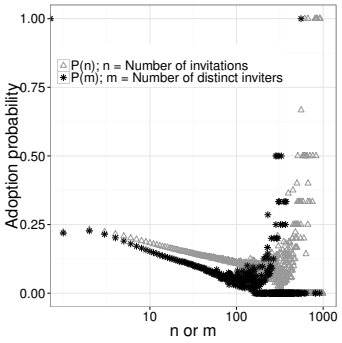


Fig. 5. Adoption probability as a function of number of invitations, n, and the number of unique inviters, m, respectively.

as n and m increase. This indicates that a large number of invitations may not necessarily help in recruiting more new users to adopt iHeart; this may be due to the spamming phenomenon and an important lesson for application recruiters. When n and m go beyond 100 invitations and inviters, $p(n)$ and $p(m)$ exhibit diverse patterns, but we observe that most of the invitees who do adopt iHeart receive only around 4 invitations from 2 distinct inviters on average.

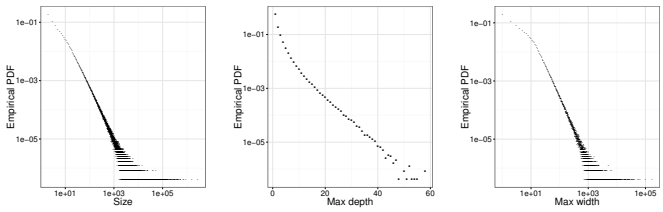
V. Characterizing Cascading Trees of iHeart

In this section, we characterize the cascading trees of iHeart from a graph-theoretical perspective, and investigate what factors drive the process of application adoptions.

A. Properties of cascading trees

We first explore the application adoption process by examining the cascading tree properties that were defined in Section III, namely the cascade size, max depth, and max width. Then we investigate the properties of cascading tree evolutions during their lifetimes as well as the roles of users based on their contributions to the growth of the cascade. Recall that the size, max depth, and max width of a cascading tree are the number of recruited users, the maximum distance from the seed to the unadopted users (leaf nodes), and the number of users in the widest generation of that tree, respectively.

1) Properties of tree structure: Fig. 6 illustrates the structural properties of cascading trees in terms of size, max depth, and max width. In Fig. 6(a), we observe that 18% of the cascading trees are of size 2, which indicates that seeds of a substantial portion of the cascading trees propagate the application to only one child. Moreover, 93% of the cascading trees are of size less than 100. Note that the maximum and average sizes of the cascading trees are 2,850,149 and 79.10, respectively. The top 10% of the cascading trees (ranked by size) have at least 70 nodes. We refer to these as the *'large'* cascading trees. We also observe that only a small fraction of the cascading trees grow deep in Fig. 6(b). The depths of 75% of the cascading trees are 1 or 2, while 7% of the cascading trees span deeper than 5 generations. The maximum and average depth of the cascading trees are 58 and 2.26, respectively. We notice that the max depth of the cascading trees in iHeart is higher than that in Twitter (11) [15] or Pinterest (35) [10], which implies that the propagation of OSN-based applications is deeper than that of other information such as tweets or pins. When we look at the cascade max width in Fig. 6(c), we find that most of the cascading trees are narrow; 80% of the cascading trees have less than 10 max



(a) Cascade size (b) Cascade max depth (c) Cascade max width

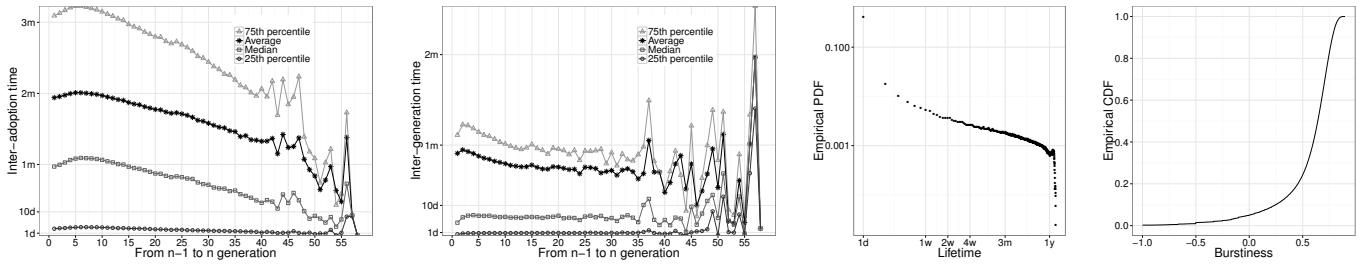
Fig. 6. Structural properties of iHeart cascading trees.

widths. The maximum and average width of the cascading trees are 171,033 and 17.64, respectively. Note that the largest max depth and max width are 58 and 171,033; they are four-orders of magnitude different, which means cascading trees in iHeart are generally much wider instead of being deeper.

2) Properties of cascade evolution: To understand the growth of application adoption process, we investigate i) inter-adoption time, ii) inter-generation time, iii) lifetime, and iv) burstiness of a cascading tree, as explained below.

Inter-adoption time is the time difference between the adoption time of a parent and each of its children in a cascading tree. Let's say, a node u and its child v in a cascading tree adopted the application at time t_u and t_v, respectively, where $t_u < t_v$. Then, the inter-adoption time i_{uv} between u and v is $t_v - t_u$. We calculate inter-adoption times for all the inviter-invitee pairs in a cascading tree. That is, a set of inter-adoption times IA of a cascading-tree is $\{i_{01}, i_{02}, \dots i_{uv}\}$. We find that 83.28% of all the inter-adoption times are longer than a day, and more than 23% of them are longer than 100 days, which implies that the propagation speed of OSN-based applications (i.e., iHeart) is slower than that of other information such as tweets in Twitter [15] and pins in Pinterest [10]. To investigate the inter-adoption times between two consecutive generations, we show the 25th percentile, median, 75th percentile, and average of inter-adoption times from $n-1$ hop to n hop in Fig. 7(a). We find that there is a large variation in the inter-adoption time at every level of the tree. Between the seeds and first generation nodes, inter-adoption time spans from minimum of 2 seconds to maximum of 227 days. This dynamic range as well as the median decreases for later generations, i.e., iHeart application propagates from one user to another at a slightly faster speed for late adopters compared to early users. This could be due to the fact that the general maturity and popularity of the application increased in the later stage. Overall, we observe that the median value decreases from 29-34 days (about a month to propagate) near the root of the cascading tree to about 13 days at generation 40.

Inter-generation time is the time difference between the first adoption time in two consecutive generations. Say, two consecutive generations g_{n-1} and g_n of a cascading tree include $n(g_{n-1})$ and $n(g_n)$ nodes with adoption times $T(g_{n-1}) = \{t_{g_{n-1},1}, t_{g_{n-1},2}, \dots, t_{g_{n-1},n(g_{n-1})}\}$ $T(g_n) = \{t_{g_n,1}, t_{g_n,2}, \dots, t_{g_n,n(g_n)}\}$, respectively. Then, the inter-generation time is defined as $IG(n-1,n) = min(T(g_n)) - min(T(g_{n-1}))$. We plot the inter-generation time in Fig. 7(b). As shown in Fig. 7(b), the median inter-generation times are below 7 days until the generation 35. After the generation 35, the inter-generation times fluctuate widely. In other words, for 50% of the time, the cascading tree takes about a week to add another level (generation). Note that the average and median of the inter-generation times



(a) Inter-adoption time (b) Inter-generation time (c) Lifetime (d) Burstiness

Fig. 7. Evolutionary properties of iHeart cascading trees.

between the seed and the first generation are 27.2 days and 4.3 days, respectively, which implies that the propagation speed of iHeart across different generations is much slower than that of tweets in Twitter [15] or pins in Pinterest [10].

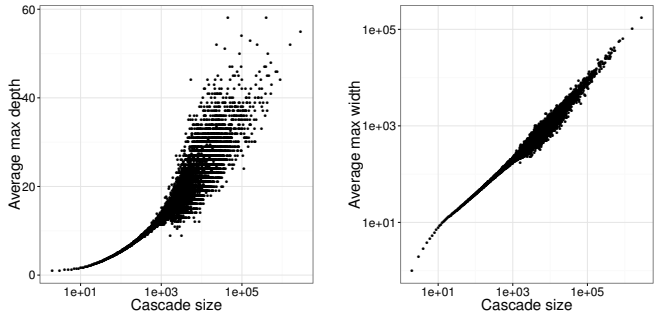
Lifetime of a cascading tree is the duration between the first time sending an invitation by a seed and the last time any node in the cascading tree sends an invitation. We show the distribution of lifetimes of all the cascading trees in Fig. 7(c). We observe that many cascading trees show long lifetimes, which signifies that they evolve for a long time during the adoption process. For instance, 3% of the cascading trees have lifetime of more than a year, and 58% of the cascading trees have lifetime of longer than a day. The average lifetime of all the cascading trees is 81.1 days.

Burstiness of tree evolution measures the regularity in the evolution process of the cascading trees. We can calculate the burstiness parameter B for the tree evolution (i.e., how regularly adoptions emerged in a cascading tree) using Eq. 2. Instead of the set of number of invitations in each time slot, we re-define X in Eq. 2 that represents the set of number of nodes in a cascading tree in each time slot. That is, $x_t \in X$ where $X = \{x_1, x_2, x_3, \dots x_t\}$ indicates that x_t nodes emerged during the given time slot t. We take time slot as 1 day. Fig. 7(d) shows the CDF of burstiness B of tree evolution for each cascading tree. As shown in Fig. 7(d), B of 5% of the cascading trees are below 0, which means they grow from regularly (closer to -1) to randomly (closer to 0) during their lifetimes. However, 80% cascading trees have B value of over 0.45, and the maximum B is 0.897, meaning that they exhibit very skewed (or bursty) growth; e.g., their growths happen predominantly during the beginning of their lifetimes.

B. What are the distinct features and driving factors for successful adoption?

We next explore what factors drive the growth of the cascading trees and whether there are distinctive features associated with different cascade sizes. To this end, we investigate i) structural factors (e.g, max depth or width of a cascading tree), ii) evolutionary factors (e.g., burstiness), and iii) roles of nodes (e.g., contribution of a seed to a cascading tree).

1) Structural properties of cascading trees: We first show the average max depth and max width corresponding to the different cascade sizes in Fig. 8. We only find a weak positive correlation between the cascade size and the average depth of the cascades (Pearson correlation coefficient is 0.365); not all big cascading trees span deep. In particular, the cascading trees whose sizes are bigger than 1000 have at least 9 generations on average. On the other hand, the average max width shows a significantly high correlation with the size; Pearson correlation

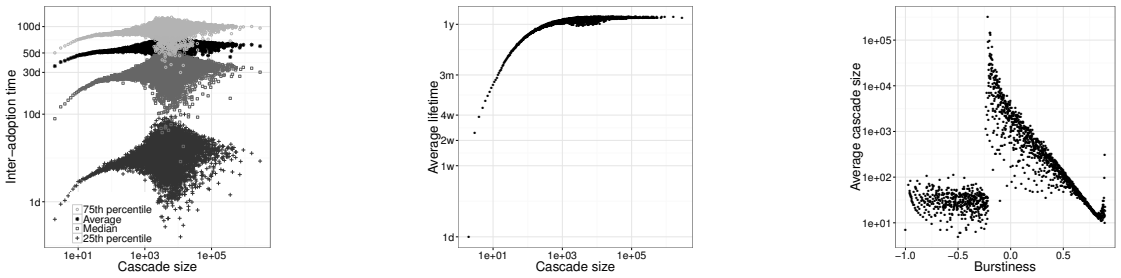


(a) Avg. max depth (b) Avg. max width

Fig. 8. Avg. max depth and width corresponding to different cascade sizes.

coefficient is 0.985. This suggests that big cascading trees are wide but small cascading trees are narrow on average.

2) Evolutionary properties of cascading trees: We investigate how the following three evolutionary properties of cascading trees vary with respect to cascade sizes in Fig. 9: i) inter-adoption time, ii) lifetime, and iii) busrtiness. Since the inter-generation time shows similar patterns with the inter-adoption time, we do not include the results for the inter-generation time due to space limitation. First, we plot the average, 25th percentile, median, and 75th percentile of the inter-adoption times of the cascading trees based on the cascade sizes in Fig. 9(a). We find that the inter-adoption time and the cascade size have a positive correlation until the cascade size is 1000. Surprisingly, the dynamic range of the average inter-adoption time changes sharply, ranging from 26 days to 77 days, after the cascade size reaches 1000. However, there is no clear correlation between the inter-adoption time (or inter-generation time) and the ultimate cascade size. Fig. 9(b) shows the correlation between the average lifetime and the cascade size. We find that smaller cascading trees have shorter lifetimes on average compared to the big ones. The average lifetime increases as the cascade size increases until the cascade size is close to 700, but it converges to around 1 year after the cascade size reaches 700. We next investigate the correlation between the average cascade size and the burstiness values of cascading trees in Fig. 9(c). Interestingly, we observe two clusters in Fig. 9(c): i) burstiness range between -1 to -0.22 and ii) burstiness range between -0.22 to 1. If a cascading tree has a uniform burstiness (closer to -1), its size tends to be less than 100. We observe that there is a negative correlation between the average cascade size and the burstiness. That is, the size of a cascading tree decreases as the burstiness increases. This implies that skewed growing patterns, such as initial spikes followed by inactivity, fails to sustain the population growth, leading to smaller cascades. We find that large cascading trees tend to have a burstiness value around zero, which implies the adoption process for large cascades tend to follow a random

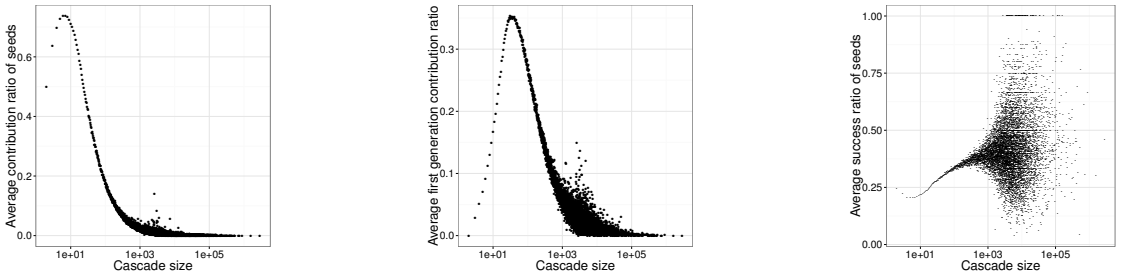


(a) Cascade size vs. inter-adoption time

(b) Cascade size vs. avg. lifetime

(c) Burstiness vs. cascade size

Fig. 9. Ultimate cascade sizes and evolutionary properties of cascading trees: (a) inter-adoption time, (b) lifetime, and (c) burstiness.



(a) Cascade size vs. avg. contribution ratio of seeds

(b) Cascade size vs. avg. first generation contribution ratio

(c) Cascade size vs. avg. success ratio of seeds

Fig. 10. Exploring the role of early adopters (seed and first-generation nodes) in the cascade growth.

(but sustainable) growth patterns over time.

3) Roles of initial adopters: Lastly, we explore whether the seeds and early adopters play a significant role in driving the growth of the cascading trees in Fig. 10. We analyze: i) the contribution ratio of the seed (i.e., founder of the cascading tree), ii) the contribution ratio of the nodes in the first generation (which may be the co-founders of the cascading tree), and iii) the success ratio (defined in Section III) of the seed. Here, the contribution ratio of a node is defined as the ratio of number of its children to the total population of the cascading tree. We observe that the average contribution ratio of the seeds decreases as the cascade size increases, as shown in Fig. 10(a). This implies that the seeds play a diminishing role in growing the user base as the cascade size increases. In fact, for large cascade sizes (70 or above), the average contribution ratio of the seeds are generally small (75th percentile is 0.012), i.e., the seeds are only responsible for recruiting a tiny fraction of the total population in the large cascades. The initial expansion of user base due to the seed is generally not a strong indicator of the ultimate size of the cascading tree. Fig. 10(b) plots the average contribution ratio of the first generation nodes, which initially increases with respect to cascade size until 31 and then it decreases as cascade size increases. On the other hand, Fig. 10(c) shows that the average success ratio of the seeds is somewhat positively correlated with the cascade size up till around 1000. Note that the Pearson correlation coefficient till the cascade size of 1000 is 0.63 while that beyond the cascade size of 1000 is 0.04, which implies that the success ratio of seeds alone may not be only the important factor for generating large cascading trees.

VI. Prediction on Cascade Growth

Our measurement-based characterization of iHeart adoption process suggests that there exist a set of distinctive features (e.g., structural or evolutionary properties of cascades and success ratio of their seed nodes) that can be combined to predict the final cascade size. In this section, we seek to answer the following questions: (1) How can we predict the growth of the cascading process of an application? (2) How long (e.g., 2 weeks or 1 month) do we need to observe the initial growth of a cascade? and (3) What features can we exploit to forecast the cascade growth with an acceptable precision? Leveraging insights gained from our measurement study, we propose a learning-based prediction model to identify *large-scale cascades of application adoption* by observing the relevant features of the cascades in the initial adoption process.

A. Prediction as a learning problem

1) Problem definition: Our goal is to identify large cascading trees whose sizes are 70 or above; this accounts for the top 10% of the cascading trees in terms of size. We cast this as a learning problem, where we observe the initial growth patterns of a cascade and forecast whether an adoption will reach a certain cascade size (i.e., 70). Note that the top 10% of the cascading trees are responsible for 86.5% of the newly recruited users in iHeart.

2) Observation window: To investigate the sensitivity of our prediction result with respect to the length of the learning phase, we vary the observation window w from one week to four weeks. For example, if the w is two, we observe the initial characteristics of a cascade until the second week. At the end of each observation window, we extract a set of features (which will be described later) associated with the cascading trees that capture the initial adoption process of iHeart. Note that we do not assume the growth of a cascade has stopped at the end of the observation window. Instead, we predict whether the observed growth will remain small (i.e., with cascade size less than 70) or grow large in the future.

3) Factors driving cascade growth: To forecast the cascade growth, we observe a set of distinctive features associated with the adoption process during a given observation window. The

Categories	Features	Descriptions
Structural	$d_w(i)$	Max depth of the cascade i at the end of the observation window w
	$w_w(i)$	Max width of the cascade i at the end of the observation window w
Evolutionary	$b_w(i)$	Daily burstiness of the cascade i during the given observation window w
	$g_w(i)$	Growth rate of the cascade i during the given observation window w
Early adopters	$c_{s,w}(i)$	Contribution ratio of the seed s of the cascade i during the given observation window w
	$c_{f,w}(i)$	Contribution ratio of the nodes f in the first generation of the cascade i during the given observation window w
	$s_{s,w}(i)$	Success ratio of the seed s of the cascade i during the given observation window w

TABLE I. FEATURE DESCRIPTIONS.

candidate features described in Table I are selected based on their demonstrated predictive power in our measurement study, i.e., how well these features can help to identify the large cascades. Note that we do not consider inter-adoption time and inter-generation time here as distinctive features for learning because of the lack of correlation between these features and cascade size. We also do not consider lifetime further since it may not be a proper measure in the initial stage; instead, we use the growth rate of the cascade i, $g_w(i)$, as a distinctive feature for evolutionary properties. $g_w(i)$ can be calculated as the cascade size of i (during the observation period) divided by w. Note that $c_{s,w}(i)$ and $c_{f,w}(i)$ are calculated as the number of children of s and f in i divided by the initial size of cascade i during w, respectively.

We first identify the specific features that contribute most towards predicting large cascades, and then we use them to create the model. For this purpose, *Chi-squared* (χ^2) statistic evaluation [18] is applied to all of the above mentioned features and a score is assigned to each one of the features, which symbolizes the relationship between the feature and the cascade size. χ^2 statistic is used to evaluate the "distance" between the distribution of the large and small cascades for an attribute. The bigger the value, the more effective is the feature in identifying the large cascades. We find very high χ^2 values for each of the aforementioned features, which are normalized by the highest value. Then we rank the features according to their normalized χ^2 value. The most important feature is g_w, which means the growth rate of a cascade is a good indicator of the ultimate cascade size. Max depth (d_w) and max width (w_w) are the next two important features for $w > 1$. The contribution ratio of the seeds is the most important feature among the three properties of early adopters. Both of the contribution ratio of the first generation users and the success ratio of the seeds have similar predictive power with respect to large cascade. Interestingly, the χ^2 values of the properties of early adopters decrease (i.e., $c_{s,w}$, $c_{f,w}$, and $s_{s,w}$) as w increases, meaning that early adopters (seeds and their recruited users) play a diminishing role as time goes on. The burstiness factor has the lowest score and, if used alone, is the least effective in predicting the large cascades.

Features	$w = 1$	$w = 2$	$w = 3$	$w = 4$
g_w	1.000	1.000	1.000	1.000
d_w	0.910	0.907	0.900	0.892
w_w	0.695	0.723	0.738	0.750
$c_{s,w}$	0.740	0.708	0.688	0.673
$c_{f,w}$	0.615	0.551	0.514	0.486
$s_{s,w}$	0.673	0.542	0.475	0.432
b_w	0.598	0.490	0.434	0.397

TABLE II. FEATURE IMPORTANCE WITH NORMALIZED χ^2.

B. Performance

We build logistic regression models based on the above features measured/estimated during the observation window from the first to fourth week of every cascade: (i) *"Structural"* considers two structural properties (i.e., d_w and w_w), (ii) *"Evolutionary"* considers two evolutionary properties (i.e., b_w and g_w), (iii) *"Early adopters"* considers three properties of initial adopters (i.e., $c_{s,w}$, $c_{f,w}$, and $s_{s,w}$), and (iv) *"All"* collectively considers all the features. We used various classifiers including SVM and decision trees, but we only report the performance of the logistic regression classifier since it performs similar or slightly better than other classifiers in most cases. We randomly set 2/3 of the 2,384,686 cascades (in iHeart) to train the model, and then we test our model with the other 1/3 of the cascades. Finally, we report the prediction precision, the true positive rate (TPR, or recall), the false positive rate (FPR), and area under the ROC curve (AUC) [6]. To calculate precision, TPR, and FPR, we use a *cutoff* probability as 0.5 to discretize continuous probabilities into binary decisions. Note that AUC resembles effectiveness of a prediction model; a perfect model has $AUC = 1$.

Features	measures	w=1	w=2	w=3	w=4
Structural	Precision	0.741	0.775	0.796	0.810
	TPR	0.344	0.438	0.492	0.535
	FPR	0.013	0.014	0.014	0.014
	AUC	0.860	0.890	0.907	0.921
Evolutionary	Precision	0.736	0.791	0.811	0.826
	TPR	0.332	0.433	0.494	0.539
	FPR	0.013	0.013	0.013	0.013
	AUC	0.901	0.926	0.938	0.947
Early adopters	Precision	0.431	0.613	0.667	0.694
	TPR	0.102	0.226	0.299	0.354
	FPR	0.015	0.016	0.017	0.017
	AUC	0.812	0.842	0.861	0.877
All	Precision	0.745	0.793	0.813	0.828
	TPR	0.415	0.495	0.548	0.587
	FPR	0.016	0.014	0.014	0.014
	AUC	0.929	0.945	0.954	0.960

TABLE III. PREDICTION PERFORMANCE ON IHEART.

Table III summarizes the performance of our prediction model. As shown in Table III, our model performs better as the observation window increases. We observe that the model based on the evolutionary properties outperforms other two models base on the structural properties and initial adopter properties, which confirms that evolutionary features are important predictors to identify the large cascades. On the other hand, the precision of the model based on the properties of early adopters is lower than other models; with four weeks of observations, the precision is only 0.694. Note that precision and TPR are already high for the observation windows $w = 1$ in our final model (i.e., "All"), meaning that one week of observation can provide a good prediction for identifying large cascades. After observing 4 weeks ($w = 4$), the precision and the recall of our final model is 82.8% and 58.7%, respectively. The AUC values of our final model for observation window $w = 4$ reaches to 0.960.

C. Universality of prediction model

We finally explore whether our learning-based prediction model can be applied to predict the large cascades observed

Features	measures	w=1	w=2	w=3	w=4
Structural	Precision	0.695	0.739	0.764	0.786
	TPR	0.252	0.356	0.417	0.460
	FPR	0.012	0.014	0.014	0.014
	AUC	0.817	0.858	0.880	0.895
Evolutionary	Precision	0.687	0.762	0.783	0.806
	TPR	0.238	0.368	0.435	0.483
	FPR	0.012	0.013	0.013	0.013
	AUC	0.874	0.913	0.930	0.941
Early adopters	Precision	0.599	0.626	0.648	0.666
	TPR	0.061	0.170	0.241	0.296
	FPR	0.005	0.011	0.014	0.016
	AUC	0.776	0.817	0.841	0.856
All	Precision	0.704	0.756	0.775	0.805
	TPR	0.316	0.425	0.482	0.527
	FPR	0.015	0.015	0.015	0.014
	AUC	0.907	0.932	0.945	0.953

TABLE IV. MODEL PERFORMANCE ON HUGGED.

in other gifting applications. To this end, we use the second dataset from *Hugged* application described in Section III-B. We followed the same methodology outlined in Section III to extract the unique seeds and construct the different cascading trees for Hugged. For each cascade, we measure the above seven features over different observation windows, w = 1, 2, 3, and 4, respectively. Likewise, we define the top 10% (ranked based on size) of the cascades whose sizes are 99 or above as large ones. Again, we train our model with randomly chosen 2/3 of the 461,510 cascades, and test with 1/3 of the cascades in Hugged. The prediction results for Hugged are very similar to iHeart, as shown in Table IV; the model based on the evolutionary properties of cascades outperforms the others. Our models achieve very high precision in identifying the large cascades also in Hugged. Note that the AUC value of our final model for observation windows $w = 4$ reaches to 0.953. Based on the prediction results for iHeart and Hugged, we believe our prediction model is generally applicable to gifting genre of OSN-based applications where new users are recruited primarily through user invitations.

VII. CONCLUSION

We investigated the adoption process of a popular Facebook application, iHeart. In particular, we analyzed: (1) how users invite their friends to an OSN-based application, (2) what factors drive the cascading process of application adoptions, and (3) what are the good predictors of the ultimate cascade sizes. We found that sending or receiving a large number of invitations does not necessarily help to recruit new users to iHeart, which may be due to the spamming phenomenon where users end up ignoring the invitations. We also discovered that the initial growth rate and shape of cascading trees (max depth and max width) are strong predictors of the final population size of the cascade. Finally, based on the insights learned from our analyses, we proposed a prediction model for identifying the large-scale cascades based on observing the initial growth process. Results showed our proposed model can achieve high precision (over 80%) with four weeks of observation period. Our model also can achieve high precision when applied to Hugged, suggesting that our prediction model is generally applicable to gifting genre of OSN-based applications where the new users are recruited primarily through user invitations.

VIII. ACKNOWLEDGMENTS

This work is supported by National Science Foundation CNS-1302691 grant.

REFERENCES

[1] Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34:441–458, 2007.

[2] J. Berry and E. Keller. The influentials: One american in ten tells the other nine how to vote, where to eat, and what to buy. Free Press, 2003.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

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

[5] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *WWW*, 2014.

[6] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[7] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang. Poking facebook: Characterization of osn applications. In *ACM WOSN*, 2008.

[8] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *ACM EC*, 2012.

[9] K. I. Goh and A. L. Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, pages 48002+, 2008.

[10] J. Han, D. Choi, B.-G. Chun, T. T. Kwon, H.-c. Kim, and Y. Choi. Collecting, organizing, and sharing pins in pinterest: Interest-driven or social-driven? In *ACM SIGMETRICS*, 2014.

[11] R. A. Hanneman and M. Riddle. Introduction to social network methods, 2005. http://www.faculty.ucr.edu/~hanneman/.

[12] E. A. Harris. Retailers seek partners in social networks, 2013. http://www.nytimes.com/2013/11/27/technology/retailers-seek-partners-in-social-networks.html?hp&_r=1&.

[13] B. Kirman, S. Lawson, and C. Linehan. Gaming on and off the social graph: The social structure of facebook games. In *International Conference on Computational Science and Engineering*, 2009.

[14] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In *ACM CIKM*, 2012.

[15] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.

[16] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *ICWSM*, 2010.

[17] H. Liu, A. Nazir, J. Joung, and C.-N. Chuah. Modeling/predicting the evolution trend of osn-based applications. In *WWW*, 2013.

[18] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *IEEE 24th International Conference on Tools with Artificial Intelligence*, 1995.

[19] A. Nazir, A. Waagen, V. S. Vijayaraghavan, C.-N. Chuah, R. M. D'Souza, and B. Krishnamurthy. Beyond friendship: modeling user activity graphs on social network-based gifting applications. In *ACM IMC*, 2012.

[20] M. R. Rahman, P.-A. Noël, C.-N. Chuah, B. Krishnamurthy, R. M. D'Souza, and S. F. Wu. Peeking into the invitation-based adoption process of osn-based applications. *Computer Communications Review*, 44(1):20–27, 2013.

[21] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *ACM IMC*, 2011.

[22] A. Sorribes, B. G. Armendariz, D. Lopez-Pigozzi, C. Murga, and G. G. de Polavieja. The origin of behavioral bursts in decision-making circuitry. *PLoS Computational Biology*, 7(6), 2011.

[23] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[24] D. Wang, Z. Wen, H. Tong, C.-Y. Lin, C. Song, and A.-L. Barabási. Information spreading in context. In *WWW*, 2011.

[25] X. Wei, J. Yang, L. A. Adamic, R. M. de Araújo, and M. Rekhi. Diffusion dynamics of games on online social networks. In *ACM WOSN*, 2010.

[26] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *IEEE ICDM*, 2010.