Analyzing the Impact of Events in an Online Music Community

Juan M. Tirado  Daniel Higuero  Florin Isaila  Jesús Carretero
Carlos III University
{jtirado,dhiguero,fisaila,jcarrete}@inf.uc3m.es

Abstract

The huge popularity of on-line social networking sites has increased the likelihood that locally-relevant events propagate globally throughout the Web. Conversely, real world events captured as digital content on the Web may influence the behavior of these digital social communities. In this work we collected and analyzed event-related data from LastFM, and global volume of searches from Google Trends, with the following objectives: (1) to study the event mechanism provided by LastFM, (2) to evaluate the impact of global and local events on system utilization, and (3) to understand the event-related propagation of information over social links.

We analyze the impact of LastFM events on the user activity and interests. Our study indicates that half of LastFM events cause an increase of the interest for an artist. However, several peaks of popularity are not associated with LastFM events, while being highly correlated with global volume of Internet searches provided by Google Trends. Finally, our analysis shows that the interest for an artist appears to be disseminated over social links. We find out that there are two factors likely to make a user influential over his friends: the degree of interest and the number of social links.

Categories and Subject Descriptors
H.3.5 [Information Systems]; Web-based services; J4 [Social and behavioral sciences]; Sociology

General Terms Measurement

Keywords Social Networks, Data Analysis, Last.FM, Google trends

1. Introduction

Recent years have shown a spectacular surge in popularity of Web 2.0 based applications such as blogging, wikis, mashups, file sharing or social networking. One of the main reasons for this phenomenon is that these applications appear to reflect the behavior and interests of users in the real world and allow participants to intercommunicate in an efficient way. The ever increasing interconnectedness among the users makes Internet a "global village" in which trends, content, marketing and rumors propagate at high speed.

On-line social networking sites such as Facebook, Twitter or YouTube are currently among the most visited in the Internet. These sites integrate many Web 2.0 applications, which can be used in cooperation by communities build based on explicitly declared relationships such as friendships, follower, colleague, fan etc. Given the huge popularity of these sites, the behavior of these communities is reflected throughout the Web. Conversely, events from the real world reflected in the Web may influence the behavior of these digital communities.

One of the mechanisms that most of the social networking sites offer is the declaration, sharing and advertising of an event. The events that can be declared in advance are expected such as a concert, a political meeting, a music release, a party, etc. A category of events are unexpected (the sudden death of an artist, an earthquake etc.) and are not reflected in the system before their occurrence, unless there is some expectancy that they might happen.

Understanding the effect of events on the community and on the system can bring several benefits. The intuition is that events have the potential of triggering activity (such as message exchange, rumor diffusion, content propagation) in an on-line social network and even beyond it depending on the general interest and magnitude. Events may be used as conveyors of advertising in viral marketing strategies. On the other hand in an active social network, events may quickly influence users and cause increases in both traffic and server load. The knowledge of patterns of traffic produced by events can help service providers to forecast high workloads and adequately distribute content or provision resources in order to improve user experience.

In this paper we collected and analyzed data from LastFM, a popular music Web site and Google Trends for a period of 142 days. LastFM allows users to listen to selected artists, to build bidirectional friendship relationships, to declare and share events such as concerts or music releases, and to post comments on events. Google Trends is an indicator of the evolution of world interests in certain topics given in number of searches. Our analysis targets to answer the following questions. What impact do events declared in LastFM have on the behavior of the users and on the system? How does the local impact of events correlate with the global impact? How can events be characterized? Have events any impact on influencing behavior over social links? What are the shortcomings of the event mechanism in LastFM? How can the event mechanism in LastFM be improved?

The remainder of this paper is structured as follows. Section presents the related work. Section presents a summary of the data set and the results of our analysis. Finally, Section concludes and presents future work.
2. Related work

On-line social networking research has approached various aspects such as structural social network characteristics [1, 21], statistical properties of content popularity [1, 19], growth and evolution of social graphs [8, 16, 24], network-level activity [22, 23]. While our study shares certain analyses and data collection methodology with the cited works, it differs in its focus on characterization of events and their impact on workload and social interactions.

Understanding the behavior of social networks has derived in different studies that propose algorithms to detect communities of similar users. Community detection is achieved by different means such as analyzing the modularity [14, 25] between partitions of a graph, or the overlapping cliques [9] in a graph. While these studies are interested in discovering communities in a social network, our approach differs in that we are not interested in finding communities based only on social links, but taking also into account the profile of the users.

There are few works that examine events in social networks. Yin et al. [23] present an analysis of a particular event (Beijing 2008 Olympic games) for Video on Demand (VoD) workloads. They study user behavior patterns such as the emergence of flash crowds and the impact of advertising on video access patterns. Becker et al. [1] present a method for clustering Flickr pictures by correlating Flickr upcoming events with LastFM events. In turn, we want to extend the understanding of event impact (1) by studying the correlation between local social network events and global interest (e.g. Google Trends) and (2) by analyzing the effect of events on content utilization and propagation through social links.

Several works have investigated patterns of dynamics of user interaction and behavior in on-line social networks [4, 22, 27]. Some studies focus on the role of social links on the propagation of information and influence [8, 24]. Bakshy et al. [3] analyze contagion in Second Life on-line game by tracking the transfers of online artifacts over 130 days. They observe that 48% of all transfers occur along social links.

3. Data set and Event characterization

In this section we provide a characterization of the events based on the collected data. First, we describe the methodology and provide general information of the data set. Second, we investigate the impact of LastFM events based on the effect on the play count. Third, we evaluate the efficiency of LastFM events, by comparing explicit events from LastFM and events inferred from Google Trends for different patterns of popularity evolution. Finally, we study how events may cause users to influence each other through social links.

3.1 Data Set

We crawled LastFM data analyzed in this paper through LastFM API by using a distributed crawler deployed in cluster over 20 machines. We traversed the friendship graph in a breadth first search manner and extracted the profiles of a set of 250,000 users including the listened artists, daily for the period between January 1st to May 22nd, 2009 (142 days). The extracted social graph has an average degree of 14.9, a diameter of 8, an average path length of 4.37, and an average clustering coefficient of 0.17.

The total number of artists users listened to was 2,390,970, amounting to a total play count of 780,579,318. For a set of 68 most popular artists, representing 13% of the total observed play amounting to a total play count of 780,579,318. For a set of 68 artists, we selected four artists with relevant popularity Becker et al. [13] present a method for clustering Flickr pictures by correlating Flickr upcoming events with LastFM events. In turn, we want to extend the understanding of event impact (1) by studying the correlation between local social network events and global interest (e.g. Google Trends) and (2) by analyzing the effect of events on content utilization and propagation through social links.

Several works have investigated patterns of dynamics of user interaction and behavior in on-line social networks [4, 22, 27]. Some studies focus on the role of social links on the propagation of information and influence [8, 24]. Bakshy et al. [3] analyze contagion in Second Life on-line game by tracking the transfers of online artifacts over 130 days. They observe that 48% of all transfers occur along social links.

3.2 Event impact

In this section we evaluate the relevance of LastFM events based on the user participation in the system. The traces show that LastFM-registered events caused user activity. The average number of users announcing to attend an event of the 68 most popular artists was 143. The play counts of songs of the involved artist on event’s day was 11,919. There was an average of 7.8 shouts per event and 1.3 reviews per event. Our results show weak correlation between play counts and reviews, shouts and reviews, play counts and attendance, and shouts and attendance. However, the number of reviews and attendance are highly correlated (0.9), while play counts and shouts are moderately correlated (0.49).

Figure 1(a) shows the CDF of the ratio between the play counts on event day and the day before (pre-event impact) and Figure 1(b) shows the ratio between the day after and the day of the event (post-event impact) for the most popular 68 artists amounting to 13% of the play counts. On the event day 52% of the events appear not to have any impact on the system. The remaining 48% increase the play count 8.9% on average. In the day after, 50% of the events generate a similar average increase of 8.3%.

This lack of impact for the events is particularly interesting. Events in LastFM can be included in the system without any reliability or relevance control. Although it is possible to notify problems related with duplicated or canceled events, there is no way to determine the reliability or relevance of an event.

3.3 Impact of events on different popularity evolution patterns

Last section has shown that LastFM events are qualitatively different in terms of impact. In this section, we complement the LastFM data with information extracted from Google Trends, reflecting the global interest in an artist based on the volume of Internet searches.

For this analysis, we filtered relevant LastFM events producing an increase of 10% in the play count in two consecutive days including the day of the event. Furthermore, from the 68 most popular artists, we selected four artists with relevant popularity evolution patterns. Figure 2(a) shows the evolution of LastFM play counts, Google Trends volume search and related LastFM events.

Figure 2(a) shows a fast increase fast decrease pattern of play count evolution for “Blink 182”. The LastFM play count is strongly correlated with Google Trends volume search index (0.91). However, the surge in popularity on February 9th, one day after the group reunited for the Grammy Awards ceremony, is not represented by any LastFM event. This event can be inferred from Google Trends spike, which accurately corresponds to the surge in the LastFM play counts. This event does not produce a consolidated interest and both indicators show a fast decay. Events on May 15th and 18th, posted in LastFM, can be inferred from play counts and Google Trends, but they produce a relative small increase compared to the earlier spike.

Lily Allen’s popularity evolution plotted in Figure 2(b) shows a sluggish increase in the first days of observations when a new album was promoted in different performances followed by a four times surge coinciding with the launch of the album on February, 9th. In this case, the play count increases due to performances and an album release matched by LastFM events. Additionally, we can also observe a strong correlation (0.86) between the play count in LastFM and the search volume index of Google Trends. After the peak, the play count volume decreases slowly, with the particularity transferences occur along social links.

For selected artists, we used Google Trends to extract the search volume index for the name of the artist, daily for the period of our trace. Google Trends [11] reports the share of a search term relative to the total number of searches done on Google over time.
that the mean play count is double after the launch of the new album compared with the play count volume before.

Play count evolution also reflects how artists gain or lose popularity as they become famous. Figure 2(c) shows the slow increase of play count for "Lady GaGa" as she becomes popular. The positive trend has a local maximum after events related with her European Tour (February 2nd – 13th). The local peak is not matched by any LastFM event, while the events on the European tour are posted in the system. As in previous cases, play count is strongly correlated with Google Trends search volume (0.74).

Figure 2(d) shows the play count evolution of "Fall Out Boy". In this case, we can clearly observe a slowly decreasing trend, which shows how the artist becomes unpopular over time. The negative trend starts with a play count volume of more than 10,000 per day, and ends at roughly 6,000. The last album of the artist was presented on December, 12th, 2008. After the release of the new

Figure 1. Impact of LastFM events on the play counts.

Figure 2. Examples of four popularity evolution patterns. The continuous line indicates the LastFM play count and the dashed line indicates the Google Trends search volume index. Events from LastFM are plotted using vertical lines.

(a) Fast increase - fast decrease (Artist: Blink 182).
(b) Fast increase - slow decrease (Artist: Lily Allen).
(c) Slow increase (Artist: Lady GaGa).
(d) Slow decrease (Artist: Fall Out Boy).
album, no major events were registered in the system until May, 3rd, when a United States tour starts. The list of LastFM events shown in the figure corresponds to the different major cities toured by the band, but unlike other cases, while the system registers minor peaks in play count volume, the negative trend continues. How the artist loses popularity is also reflected in Google Trends (2.8 times less search volume), matching the trend observed in LastFM (correlation 0.69).

3.4 Influence propagation over social links

In this section, we want to understand how events may affect influence propagation over social links. For this analysis, we have selected for each of the four artists from Figure 2(a) the events associated with the highest increase in play counts. The selected events correspond either to LastFM-declared events or events inferred from correlated peaks of LastFM play count and Google Trends global volume search index (e.g. the peaks for Blink-182 from Figure 2(a) without an associated LastFM event). For the analyzed events, we select the users listening to the associated artist and the play count for three days: pre-event day, day of the event and post-event day. For the users from this set, we define a user to be active if: (1) the user listens to an artist two consecutive days D1 and D2 (2) D1 is the day of the event (post-event active) or D2 is the day of the event (pre-event active).

We classify the active users in potential influential over other users with respect to an event if the number of friends listening to the artist on day D2 is greater than on day D1, and non-influential otherwise. A potential influential user has pre-event influence if D2 is the day of the event and post-event influence if D1 is the day of the event. A user that may have been influenced by others is called infected. We use the attribute potential in order to point out that there can be other factors, which can affect the behavior of a friend listening to the same artist in the next day. One of these factors is homophily, which is defined as the tendency of people with similar tastes to behave similarly. Influence and homophily have been shown to be hard to distinguish when present [2, 24]. However, the absence of homophily between two users listening to the same artist in consecutive days, might indicate a higher probability of influence propagation over the social links.

Table 1 shows that from 2759 pre-event active users 32% may be influential over 5970 of their friends. The number of post-event active users increases to 3475, out of which 24% may be influential over 5116 friends. A number of 1516 are both pre-event and post-event active users. A number of 108 pre-event infected users become post-event active, out of which 16 become post-event influential, a result suggesting the propagation of influence over the social cascade.

<table>
<thead>
<tr>
<th></th>
<th>Pre-event</th>
<th>Post-event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active users</td>
<td>2759</td>
<td>3475</td>
</tr>
<tr>
<td>Potential Influential</td>
<td>894</td>
<td>822</td>
</tr>
<tr>
<td>Non-influential</td>
<td>1865</td>
<td>2653</td>
</tr>
<tr>
<td>Infected</td>
<td>5970</td>
<td>5116</td>
</tr>
</tbody>
</table>

Table 1. Active and potential influential users. Active users are users listening to an artist two consecutive days, including the day of the event. Potential influential users are active users, whose number of friends listening to an artist increments the second day.

First, we want to understand if there is any factor that makes an active user become or not influential over other users on day D2. Figure 3(a) shows the CDF of the play count of potential influential and non-influential users during D1 day for two cases: D1 is the pre-event day (pre-event influence), and D1 is the event day (post-event influence).
event influence). Potential influential users show a higher play count both days. The play count for non-influential users slightly increases the day of the event, but it shows a significant difference with potential influential users. On the day before the event, 50% of the potential influential users listen to less than 3 songs (with an average of 13.05 songs), while one half of the non-influential users listens to less than 2 songs (average 8.883). For the event day the play count increases in both cases. Potential influential users listen to less than 5 songs in 50% of the cases (average 18.97), while 50% of non influential users listen to less than 3 songs (average 10.43).

On average, potential influential users listen to 68% and 54% more songs than non influential users the day before the event and the event day, respectively.

Second, we study the effect that active users have on their friends on day D2. Figure 3(b) illustrates the average play count for the friends of potential influential and non-influential users. Friends of potential influential users listen on average to more songs than friends of non-influential users on event day: 14.65 songs (median: 7.889) versus 11.439 songs (median: 4.536) for friends of non-influential users. On post-event day the average play count value for friends of both potential influential and non-influential users is similar (the median differs: 9.102 for potential influential and 6 for non-influential users). This is due to the fact that, even though the aggregate play count increases, the number of friends of potential influential users listening to the same artist also increases; for the non-influential users this number remains constant.

Third, we investigate how the social degree of active users might indicate the proclivity of a user to become potential influential or non-influential. Figure 3(c) shows that potential influential users have a significant higher number of friends (median 58, mean 75.6) than non-influential users (median 24, mean 39.55). The results suggest that users with a high social degree are more likely to influence their friends than users with a lower social degree.

Finally, we analyze the pairs of friends classified by the previous analysis as potential influential and infected users, in order to estimate if the observed behavior might be a result of influence propagation or homophily [18]. We base this analysis on a metric computing the distance between user consumption profiles. The consumption profile \( P(u_i, d_k) \) is the set of tags associated to artists listened by user \( u_i \) on day \( d_k \). We calculate \( s(u_i, u_j) \), a content sharing coefficient between users \( u_i \) and \( u_j \) over a period of \( D \) days before the potential influence was observed, as the average over \( D \) of Jaccard similarity coefficient for user day profiles \( P(u_i, d_k) \) and \( P(u_j, d_k) \). If two users are listening daily to the very same set of artists the value of the sharing coefficient will be 1, while if they are listening to completely different artists it will be 0. The formula of \( s \) is given by Equation 1.

\[
s(u_i, u_j) = \frac{1}{D} \sum_{d_k \in D} \frac{|P(u_i, d_k) \cap P(u_j, d_k)|}{|P(u_i, d_k) \cup P(u_j, d_k)|} \tag{1}
\]

Figure 4 shows a CDF with the values of \( s(u_i, u_j) \), where \( u_i \) are potential influential users, \( u_j \) the potentially infected user, and \( D \) one week before an event occurs. We observe that the pairs of potential influential-infected users listening to Blink-182 appear to share less similar content than the users listening to all the other artists, whereas the pairs listening to Lady Gaga share more content than the others. This indicates that the probability that a friend was infected over a social link is larger for Blink-182 and smaller for Lady Gaga. For example, for Blink-182 around 20% of pairs do not share any tag 7 days before the event. This observation suggests that listening to the artist by a potentially infected user might be the result of a influence propagating over a social link. In the case of Lady Gaga, around 20% of the pairs have a sharing coefficient larger than 0.4. For these pairs of users it is complicated to differentiate between two cases: do we identify an influence propagated over the social link? or was the social link created because the two users share common interests (homophily)? We propose a probabilistic interpretation, which associates a higher certainty of influence propagations to the pairs sharing less content, and a higher certainty of homophily to the pairs sharing more content.

4. Conclusions and future work

The analysis performed in this work suggests that events are correlated with activity in LastFM, irrespective if they are or not posted in the local scope of the system.

First, we have observed that only around 50% of the events posted in the system are associated with an increase in the play count of the involved artist larger than 10%, while the other 50% did not appear to have any impact. On the other hand, several peaks of popularity were not associated with any LastFM event, reducing the possibility of load prediction.

Second, our study shows that there is a high correlation between the LastFM play count per artist and the volume of searches in Google Trends. This fact indicates that local LastFM activity reflects the global interest on artists. This result suggests that the lack of user-declared LastFM events may be compensated by events inferred from the global interest for an artist. This approach can be used both for advertising inside the system or predicting system load.

Third, we investigate the influence of events on the propagation of information through social links. Our analysis shows that interest for an artist appears to be disseminated over social links. We identify a set of active users, which listen to an artist both the day before the event and the day of the event. Our analysis suggests that there are two factors likely to make an active user potentially influential or non-influential. Influential users listen on average to more songs than non-influential users in the day before event: 13 versus 9. Influential users have on average more friends than non-influential users: 76 versus 40. Our results also show that 16 infected users become influential for their friends in the day after the event, suggesting the existence of a social cascade. In order to confirm this hypothesis, we measure the level of homophily as the ratio of shared tags between potential influential and potentially infected users for the period of our trace. Our study shows that the sharing coefficient remains low (under 0.4) in most cases. This suggests that with a larger certainty the potential users can be classified as influential. However, in some cases the sharing coefficient can be as high as 1. We propose a probabilistic interpretation, associating high certainty of influence propagation with low sharing coefficient values and high certainty of homophily with high sharing coefficient values.
Finally, we make some observations and suggestions related to the LastFM event mechanism. In LastFM, users can post events with a limited number of attributes (for instance an event can be only a concert or a festival), making the automatic classification of events difficult. Additionally there is no validation and reputation mechanism, which should allow users to evaluate the relevance of an event. We suggest three improvements in the LastFM event management system. First, a broader range of attributes should allow the classification of events into live performances, disc releases, TV appearances, etc. Second, we suggest a collaborative event validation mechanism including a reliability system to help users to decide by vote if an event is relevant or reliable or whether a user is to be trusted. Third, we recommend a mechanism for semi-automated addition of events, by monitoring the global interest based on services such as Google Trends. This is likely to increase both user satisfaction and system popularity.

We would like to extend this work in various directions. First, we would like to propose mechanisms for inferring events from the activity and interests reflected either in related sites or globally in the Internet. Second, we are interested in studying the relationship between events and interaction patterns in LastFM. Third, we would like to investigate how events can be leveraged for predicting activity in the system, in order to adequately provision resources during peak loads. Fourth, we wish to make a comparative study among different systems to extend our conclusions to a larger sample of users and systems.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Education under FPU program AP2007-03530 (Juan M. Tirado) and by the Spanish Ministry of Science and Innovation under the project TIN2010-16497.

References


