Journal of Network and Computer Applications 56 (2015) 88–100

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

# Lifespan and propagation of information in On-line Social Networks: A case study based on Reddit

Giannis Haralabopoulos [a,*], Ioannis Anagnostopoulos [a], Sherali Zeadally [b]

[a] Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia GR 35100, Greece
[b] College of Communication and Information, University of Kentucky, Lexington, KY 40506-0224, USA

ARTICLE INFO

*Article history:*
Received 2 April 2015
Received in revised form
22 May 2015
Accepted 21 June 2015
Available online 3 July 2015

*Keywords:*
Virality
Online Social Network (OSN)
Social network service
Information flow

ABSTRACT

Since 1950, information flows have been in the center of scientific research. Up until the Internet penetration in the late 1990s, these information flow studies were based on traditional offline social networks. From the first Online Social Network studies, various observations of "offline" information flows, such as the two-step flow of communication and the importance of weak ties, were verified in several "online" studies, also indicating that information flows from one Online Social Network (OSN) to several others. Within that flow, information is shared with and reproduced, by users of each network. Furthermore, the original content is enhanced or weakened according to its topic, as well as the dynamic nature and exposure of each Online Social Networks (OSNs). In such an informational connected environment, each OSN is considered as a layer of information flows, which interacts with other layers. We examine information flows in several social networks, as well as their diffusion and lifespan, across these networks, based on user-generated content. Our results verify the information connection in various OSNs and provide a measurement of shared information lifetime in multiple OSNs.



## 1. Introduction

Information is constantly exchanged online among friends, acquaintances, family members, colleagues and even unknown individuals. The type of information varies and includes local or world news, general and scientific facts, quotes, personal preferences, etc. This broad information exchange would not be possible without a communication environment such as the Internet. Moreover, the creation of Online Social Networks (OSNs) and their adoption in our everyday lives, have led to the development of new information-sharing schemes where users can easily disseminate information quite fast.

The information that flows in OSNs, along with its characteristics, properties and impact, have been the subject of several previous studies (as discussed in Section 2), which were mainly focused on the following:

a. Virality: the tendency of information to be circulated rapidly and widely across different Web users.
b. Diffusion and propagation: how fast information is reproduced and spread in OSNs.
c. Dynamics: properties that constitute, sustain, or modify the topology of the OSNs based on the diffusion and propagation processes (hierarchy, network partition, clustering etc.).
d. Influence: the capacity of OSN power nodes (e.g. popular persons, news media, opinion makers) to be a compelling force on behavior.

However, nowadays, OSNs are densely connected to each other, through multiple information flows.

Considering every OSN as a layer (Fig. 1) and the information as links connecting each layer, we propose the concept of multilayer information flow. In this concept, information is spread from a source layer and propagates in multiple other layers. To evaluate our proposal, we decided to focus on Reddit and its content, mainly because this OSN is comprised of original content, along with relayed information, and offers a fairly liberal data access policy, with an open Application Programming Interface (API) and the required documentation for data scraping.

Reddit is a social news and entertainment site powered by user generated content. Registered users submit content through a descriptive link that may contain; an image, meme, video, question, Ask Me Anything (AMA) session, and the community can then vote and comment on that post. In correspondence with votes, users who have created a post or commented on one, gain or lose "karma", a Reddit-oriented metric for user ranking. This

* Corresponding author.
*E-mail addresses:* ghara@dib.uth.gr (G. Haralabopoulos), janag@dib.uth.gr (I. Anagnostopoulos), szeadally@uky.edu (S. Zeadally).

http://dx.doi.org/10.1016/j.jnca.2015.06.006
1084-8045/

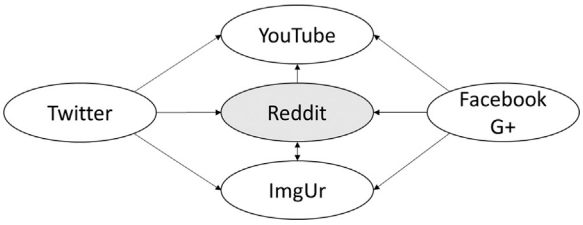


**Fig. 1.** Linkage of different social networks – this work is focused on Reddit.

metric is calculated as the sum of all the upvotes minus the sum of all the downvotes a user receives. Posts that acquire a high vote ratio (positive to negative) in a short time period after their submission are moved to the front page. It is apparent that Reddit community defines the popularity of the disseminated content and determines its "success" or "failure".

Throughout posted content on Reddit, we focused on posts that link to an external domain and their traffic could be easily measured (e.g. number of views). Thus, content linking to Wikipedia articles or news sites, is not taken into consideration in this research. One of our initial observations was that the highest rated content was mostly from the ImgUr domain. ImgUr is one of the most popular online images hosting service in Reddit community.

As illustrated in Fig. 1, content on Reddit can be (amongst others) an image or a set of images hosted in ImgUr, or a video in YouTube. In ImgUr, content is usually created at the same time as the corresponding post in Reddit. In the case of YouTube, most posts in Reddit are linked to old videos. A short time after content creation and the subsequent increase in popularity within Reddit, users from different OSNs start mentioning that content, either by citing Reddit or the domain where the content is hosted (ImgUr or YouTube in our case).

In this work, we wanted to use famous and heavily-visited OSNs. In this context, Twitter provided us with the ability to fully observe the impact of a front-paged Reddit post. In contrast, since most content in Facebook and Google Plus are private, we only discovered a fraction of the total references, derived from search through public posts.

It is an easily observed fact that information is shared and spread among these social networks. A post in any of these networks impacts the others as well. But, what is the size of this impact? How long does it last? Is it dependent over different thematic categories (e.g. politics, sports, entertainment etc.)? These are some of the questions we intend to explore in this work. Information flows and their diffusion along with information virality, are the main aspects addressed towards the answer. Through multiple social network analysis, we aim to present the multi-layered flow of information across modern OSNs.

The remainder of the paper is organized as follows. In Section 2, we present some of the most important research initiatives on information virality and information diffusion along with their results in social networks. Section 3 describes the data mining methods and the dataset we used. In Section 4, we present the results derived from our analysis. Finally, Section 5 discusses evaluation issues of our study, while Section 6 concludes this work.

## 2. Related work and contributions of this work

### 2.1. Related work

The topic of diffusion has been at the center of sociology interest for many years. Even before the emergence of OSNs, social ties and information flows have been studied in traditional real-life social networks.

The notion that information flows, from mass media to opinion leaders and later on to a wider population as final consumers, was firstly introduced during the middle of 1940s (Lazarsfeld et al., 1944). In their introduction, Lazarsfeld et al. found that, during a presidential election, mass media had a direct influence on voting intentions, but informal and personal contact was more frequently mentioned as sources of influence. Almost a decade later, Katz and Lazarsfeld (1970) revisited the subject by proposing their theory of "Two step flow of communication". Specifically, they proposed that the information from the media is first received by opinion leaders, who then pass their interpretation and actual information to other individuals. In a similar conceptualization, Granovetter and Mark (1973) suggested the analysis of social networks, as a tool for linking micro- and macrolevels of sociology theory. Small-scale interactions become large-scale patterns and feed back into small groups of users. The author concluded that weak ties are seen as the individual's opportunity towards integration into a community, while strong ties result in multiple fragmentations of the community (small disconnected clusters or islands) that eventually lead to the fragmentation of the whole community. Similar conclusions were verified in OSNs after nearly 40 years, while at the same time, interest in social media analysis skyrocketed (Bakshy et al., 2012; Rajyalakshmi et al., 2012).

Viral marketing was introduced in 1997 by Jurvetson and Draper (1997). In their work, the authors described their marketing strategy for a free email service. They found that, by sending personal messages to individuals, the marketing team boosted the number of email users. Their strategy proved to be very effective, gaining millions of users within few months. During the following decade, not only e-mail but several Internet-based services (e.g., real-time interactive services, OSNs) evolved becoming an integral part of marketing strategies worldwide.

Porter and Golan (2006) found that provocative content such as sexuality, humor, violence and nudity are crucial virality factors, in comparison with traditional TV advertising, where emotive content had always been the key. In addition to virality, information diffusion became an important research subject. Leskovec et al. (2007), modeled outbreak detection via node selection, while performing a two-fold evaluation of their model by using a water distribution network as well as a blog network. Although the aforementioned model was not verified on any OSN, yet it could be easily applied to such networks mainly because OSNs have similar diffusion properties with the ones tested in Leskovec et al. (2007). Concerning word of mouth scenarios, Allsop et al. (2007) noted that 59% of individuals frequently share online content. The authors also tried to examine the link between emotion and virality, and concluded that content sharing is mainly adopted for entertainment purposes. Berger and Milkman (2010) observed that positive content is more viral than negative. Moreover, they ended up with the conclusion that the more the content evokes emotions of activation (e.g. anger, awe, anxiety), the more viral is, in contrast to deactivating emotions (e.g. softness).

Rodriguez et al. (2010) focused on identifying the optimal network that best describes information propagation in news media and blogs. News diffusion networks were found to have a core periphery structure, where a small set of core media sites diffuses the information to the rest of the network, while blogs are mostly influenced by mass media. Iribarren and Moro (2011) analyzed word of mouth through web means, and mainly through email forwarding. Two dynamic patterns were observed by the authors, namely "Transmissibility" and "Fanout Cohesion", while direct referrals proved the best form of diffusion by creating affinity paths (e.g. the linkage created by preference and interest).

Social influence modeling was studied by Goyal et al. (2010). Using old propagation data to create such models, the authors observed that viral marketing could leverage genuine influence which, as they claimed, occurs only in real-world networks. Similarly, Ienco and Bonchi Castillo (2010) investigated the social influence in