[Advanced Content Delivery and Streaming in the Cloud], Edited by [Dr.Mukaddim Pathan, Prof. Ramesh K. Sitaraman, Dom Robinson].
ISBN 0-471-XXXXX-X   Copyright © 2013 Wiley[Imprint], Inc.

Chapter 22

On the impact of Online Social Networks in Content Delivery

Irene Kilanioti, Chryssis Georgiou, George Pallis

22.1 Introduction

As *Content Delivery Networks (CDNs)* task is the improvement of Internet service quality via replication of the content from the origin to surrogate servers scattered over the Internet, the area of CDNs faces three major issues concerning the maximization of their overall efficiency [31], [35]: (i) the best efficient *placement* of surrogate servers with maximum performance and minimum infrastructure cost, (ii) the best *content diffusion* placement, either in a global or in a local scale, i.e., which content will be copied in the surrogate servers and to which extend, since this requires memory, time and computational cost, and (iii) the *temporal diffusion*, related with the most efficient timing of the content placement.

The increasing popularity of *Online Social Networks (OSNs)* [3], [9], [15] and the growing popularity of streaming media have been noted as being the primary causes behind the recent increases in HTTP traffic observed in measurement studies [10]. The amount of Internet traffic generated every day by online multimedia streaming providers such as YouTube has reached huge numbers. Although it is difficult to estimate the proportion of traffic generated by OSNs, it is observed that there are more than 400 tweets per minute with a YouTube link [6]. These providers often rely on CDNs to distribute their content from storage servers to multiple locations over the planet. Towards this direction we can exploit information diffusion analysing the user activity extracted from OSNs. Thus, the improvement of user experience through scaling bandwidth-demanding content largely depends on the exploitation of usage patterns found in OSNs, and can be conducted either through the selective prefetching of content, also taking into account timing issues, or through the strategic placement of surrogate servers. Furthermore, the cost of scaling such content in CDNs can be expressed in different ways. For example, it might be the number of replicas needed for a specific source or it may take into account the optimal use of memory and processing time of a social-aware built system. Thus, it is crucial to support social network analysis tasks that accommodate large volumes of data requirements for the improvement of user experience (e.g., through prefetching via a CDN infrastructure).

The goal of this chapter is to present existing approaches that can be leveraged for the scaling of rich media content in CDNs using information from OSNs. Specifically, we present a taxonomy of the relative research (outlined in Figure 22.3 in the next section), taking into account phenomena related with rich media content and its outspread via OSNs, and measurement studies on OSNs that could provide valuable insight into CDN infrastructure decisions for the replication of the content, as well as systems built with the leverage of OSNs' data.

The remainder of this chapter is organized as follows: In Section 22.2 the main concepts of OSNs and social cascades are presented. In Section 22.3 the properties and approaches that characterize the social cascades and affect the CDN performance are described. The performance measurements of associating rich media content diffusion with social networks is given in Section 22.4. Section 22.5 gives the outline of existing works that exploit information extracted from OSNs for the improvement of content delivery. In Section 22.6 some future directions are given, and key areas of interest concerning the diffusion of rich media content over OSNs are explored with some commercial/practical implications for CDNs. Section 22.7 concludes our study.

22.2 Online Social Networks Background

Formally, an OSN is depicted by a directed graph $G = (V,E)$, where V is the set of the vertices of the graph representing the nodes of the network and E are the edges between them, denoting the various relationships among the nodes of the graph [15]. The semantics of these edges are different for different social networks: for Facebook, friendship is usually translated in personal acquaintance, whereas in LinkedIn means business contact. As far as the directionality of the edges of the social graph is concerned, it depends on the OSN the graph depicts. For Facebook, an edge means mutual friendship between the endpoints of a link. For Twitter, if the edge between B and A points at A, B is a follower of A, meanings that A’s posts (tweets) appear in B's main Twitter page. The *neighbours* of a node are defined as the nodes that are in a 1-hop away distance from it in the social graph. Figure 22.1 depicts an example of a Twitter social graph. Unlike other OSNs, a Twitter user may follow another user to receive his/her tweets, forming a social network of interest. Furthermore, it is not necessarily the case that two users are mutual followers. Thus, Twitter is represented by a directed graph, where nodes represent the users and a direct link is placed from a user to another user, if the first follows the tweets of the latter. Users A and G are mutual followers, while users A and B are not (A follows B but not vice-versa).

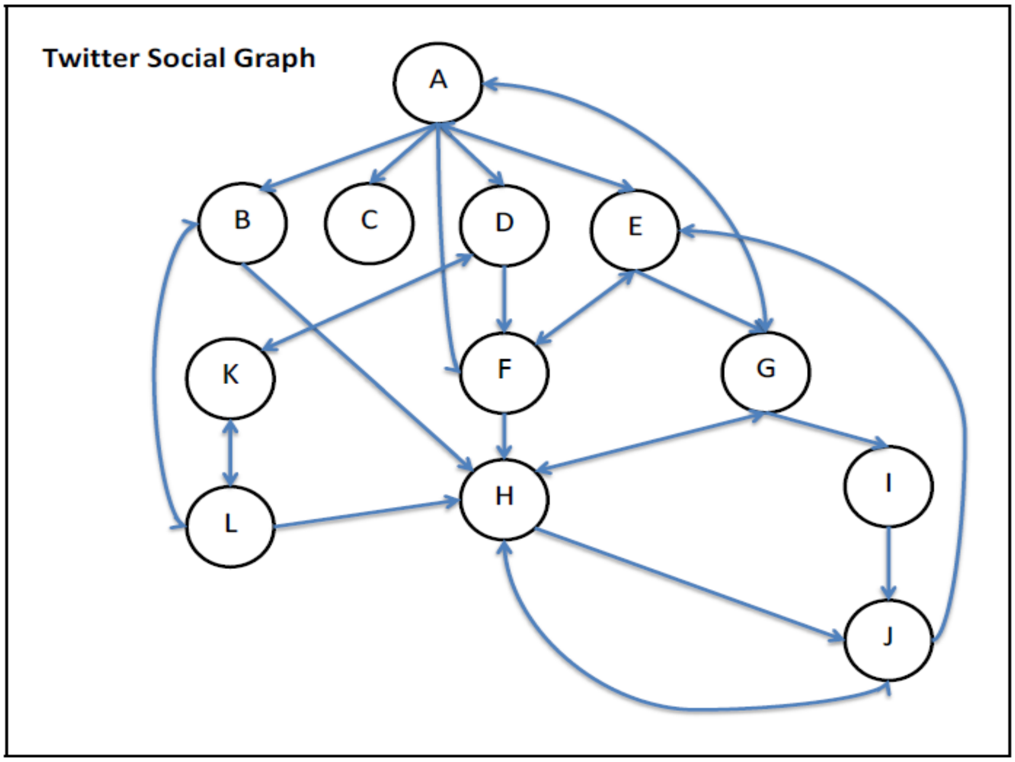

Figure 22.1 An Example of a Twitter social graph

A fusion of bandwidth and storage demanding media, which may include text, graphics, audio, video, and animation is characterized as *Rich Media*. Rich media is currently ubiquitous due to the proliferation of smartphones, video editing software and cheap broadband connections. The *diffusion* of information in a network is essentially interweaved with whether a piece of information will become eventually popular or its spread will die out quickly. A large proportion of rich media is distributed via OSNs' links (for example YouTube videos links through retweets in Twitter), that contribute significantly to Internet traffic. Facebook and Twitter users increasingly repost links they have received from others. Thus, they contribute to *social cascades* phenomena [2], a specific case of *information diffusion* that occurs in a social network, when a piece of information is extensively retransmitted after its initial publication from an originator user. Therefore, it would be beneficial to know when such cascades will happen in order to proactively replicate popular items (*prefetching*) via CDN infrastructures. *Content diffusion* placement and *temporal diffusion* could significantly benefit from such "prefetching" policies.

Social cascades can be represented as rooted directed trees where the initiator of the cascade is the root of the tree [2]. The *length* of the cascade is the height of the resulting tree. Each vertex in the cascade tree can have the information of the user, and the identity of the item replicated in the cascade. Figure 22.2 depicts an example of a cascade, initiated by user A over the Social Graph of Figure 22.1. Then, the length of the cascade is the height of the resulting tree.

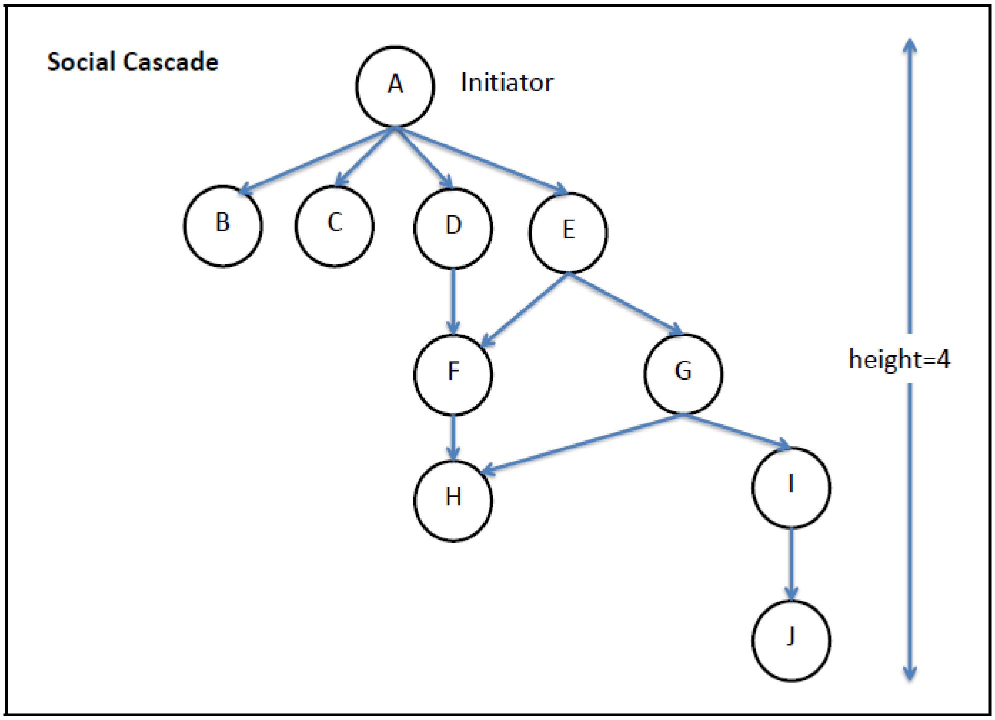

Figure 22.2. An Example of a Social Cascade.

The presented taxonomy of Content Delivery over OSNs is outlined in Figure 22.3. Our taxonomy presents the various properties and approaches in the literature for the *characterization of cascades*. The branching of topics in our presented taxonomy continues with *OSN measurement* works that focus on phenomena and measurement studies providing valuable insights into *usage analysis* and *media diffusion*. The last dimension of our taxonomy consists of content delivery *systems* built based on OSNs' data.

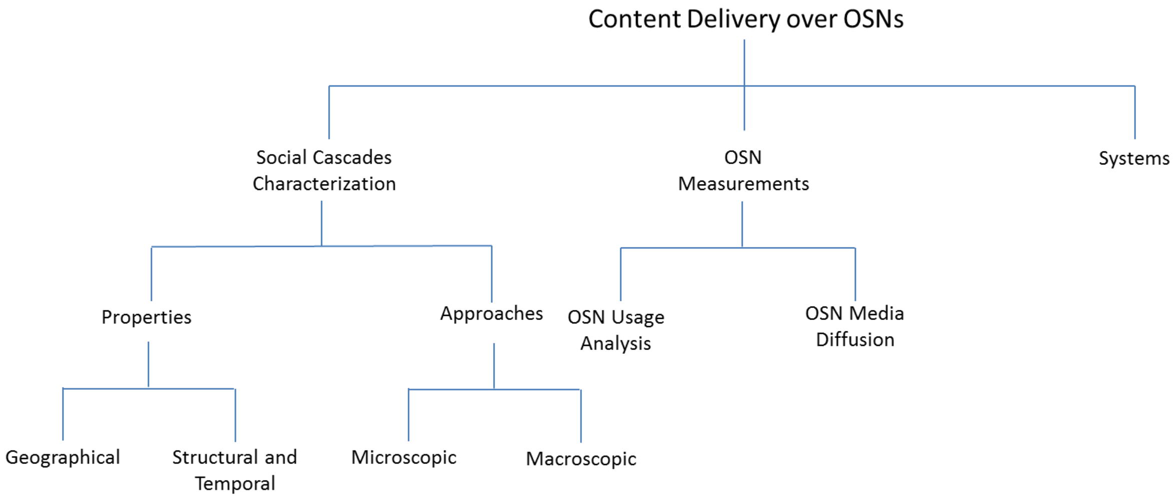

Figure 22.3. A taxonomy of Content Delivery over OSNs.

22.3 Characterization of Social Cascades

As mentioned above, it would be beneficial to characterize the social cascades in order to proactively replicate popular items via CDN infrastructures. The following sections present the key properties of cascades (geographical, structural and temporal) and the existing approaches (microscopic, macroscopic) for the characterization of the extent a cascade will receive.

22.3.1 Geographical Properties

According to a recent study [31], the geography of the requests influences the performance of CDNs. Therefore, it would be useful to understand whether an item becomes popular on a global scale or just in a local geographic area.

Local cascades affect only a relatively small number of individuals and typically terminate within one or two steps of the initiator. The size of local cascades is determined mostly by the size of an initiator's immediate circle of influence, not by the size of the network as a whole. In global cascades the opposite happens: they affect many individuals, propagate for many steps, and are ultimately constrained only by the size of the population through which they pass.

A cascade is local if it spreads in a fraction φ of the network lower than a threshold ω or else we say that the cascade is global. Let the classification function f_1 for a cascade C_z, $z \in \mathbf{N}$ be as follows:

$f_1\ (G, C_z) = \{0 \text{ if } \varphi<\omega \text{ local cascade,} 1 \text{ if } \varphi\geq\omega \text{ global cascade}\}$

Formally, a cascade should be characterized as global or local with the maximum accuracy $\alpha = \frac{\sigma}{v} * 100$, where σ is the number of the correctly classified cases and v is the number of all sampled cases, such that the cost of replicating a simple object c is minimized: min $\sum_{C_z} c = f_2(t_p,\ \varphi\ ,N_{cl,},C_z\ )$ for all cascades $C_z \in C$, where φ is the fraction of the network that the object is bound to spread, N_{cl}, $cl \in \mathbf{N}$ is the number of clients requesting a specific object and t_p the amount of traversing path p between client and the server finally serving the request.

Other geographical properties which have been presented in the literature in order to characterize social cascades are the *geodiversity* and *georange.* Specifically, the geodiversity denotes the geometric mean of the geographic distances between all the pairs of users in the cascade tree, whereas, the georange denotes the geometric mean of the geographic distances between the users and the root user of the cascade tree [31].

22.3.2 Structural and Temporal Properties

Social cascades are characterized by the structural properties of size and length. The *size* is the number of participants, including the initiator, and the *length* [2] denotes the height of the cascade tree. Social cascades are also characterized by temporal properties such as the *time delay* between two consecutive steps of the cascade [31], the *time duration,* and the *rate of the cascade* [8]. The latter, for the epidemiological model of [8] is the basic reproductive number $R_0 = \rho_0 \overline{k^2} / \overline{k}^2$, where $\rho_0 = \beta\gamma \overline{k}$, β is the transmission rate, γ is the infection duration, and k the node degree. With σ_0 the probability that a person will adopt the shared piece of information (under the assumption that duration infection is equal to the timelife of the user, much larger than duration of the cascade, and, thus, the information will be definitely shared among connections) it applies $\rho_0 = \sigma_0 \overline{k}$, and σ_0 can empirically be estimated by identifying an infected node and counting the fraction of its connected nodes subsequently becoming infected. Another temporal property related to the *susceptibility* of the network to new items is the time to the first step of the cascade from the infector's point of view and the duration of exposure to an item before infection from the infectee’s view [8].

22.3.3 Approaches

Even though a percentage of the occurred information flow in cascades is ascribed to homophily, namely the tendency of individuals to associate with similar others, as “similarity breeds connection” [26], and research has been conducted for the discrimination between the two cases (homophily or influence) [23], different approaches are presented for the characterization of the extent a cascade will receive. Some of them are related to the extent that nodes are influenced by their neighbours on a *microscopic* level, such as the “vulnerability” that Watts [13] introduces, some to factors that function as obstacles to the spread of a cascade on a *macroscopic* level, such as those that Kleinberg and Easley [15] or Ver Steeg et al. [33] introduce, and others follow an hybrid approach (Dave et al. [11]). The discrimination is based on the view of the OSN as a whole or on its study based on user-level properties.

Microscopic approaches

In [13], Watts defines a global cascade as a “sufficiently large cascade”, covering practically more than a fixed fraction of a large network. Watts introduces a simple model for cascades on random graphs.

- The network comprises of *n* nodes with threshold distribution $f(\varphi)$, and the degree distribution of the graph is p_k, namely each node is connected to *k* neighbours with probability p_k; *z* is the average node degree ($\bar{k} = z$).
- The initial state of each node is state 0 (inactive) and each node is characterized by a threshold φ. If at least a threshold fraction φ of the node’s *k* neighbours acquire state 1 (active), the node will switch from inactive to active.

- Nodes with $k \leq \lfloor 1/\varphi \rfloor$ are said to be "vulnerable" and will switch state if just one of their neighbours becomes active. Otherwise, nodes are called "stable".

Watts uses *percolation theory* [32], the theory studying how connected clusters behave in a random graph, to investigate the conditions under which a small initial set of seed nodes can cause a finite fraction of infinite nodes to switch from inactive to active. Percolation in this case is interpreted as follows: a global cascade is said to occur when the vulnerable vertices percolate. Namely the largest connected vulnerable cluster of the graph must occupy a finite fraction of the infinite network. For infinite Poisson random graphs, Watts defines a region, inside which a finite fraction of an infinite network would switch from inactive to active state if at least one arbitrarily selected node switched from inactive to active state. Simulations on finite graphs of 10000 nodes give similar results.

Watts makes the observation that the frequency of global cascades is related to the size of the vulnerable component, with the larger the component, the higher the chance for the cascade to be global. He also states that the average size of a global cascade is governed by the connectivity of the network as a whole. In sparsely connected networks, cascades are limited by the global connectivity of the network, and in dense networks cascades are limited by the stability of individual nodes.

Macroscopic approaches

In [15], Kleinberg and Easley claim that clusters are obstacles to cascades, and, moreover, that they are the *only* obstacles to cascades: "Considering a set of initial adopters of behavior A, with a threshold of *q* for nodes in the remaining network to adopt behavior A: (*i*) If the remaining network contains a cluster of density greater than 1-*q*, then the set of initial adopters will not cause a complete cascade. (A cluster of density *p* is a set of nodes, so that each node in the set has at least a *p* fraction of its network neighbors in the set.) (*ii*) Moreover, whenever a set of initial adopters does not cause a complete cascade with threshold *q*, the remaining network must contain a cluster of density greater than 1-*q*."

In [33], Ver Steeg et al. find two additional factors related with the multiple exposure of users to stories due to the highly clustered nature of Digg, that drastically limit the cascade size in Digg. The reproductive number R_0 of the epidemical model used, which intuitively expresses the average number of people infected by a single infected person, is the product of the average number of *fans* times the *transmissibility*. As far as the first factor is concerned, it is implied that only the number of new fans (those that have not already been exposed to a story) should be taken into account. For the second factor, transmissibility for actual cascades is observed to remain constant until about a number of people have voted, and then begin to decline (maybe due to decay of novelty [37] or decrease in visibility [19] as a consequence of new stories being submitted to Digg). From this point of view, cascades are limited.

In [11], Dave et al. combine microscopic and macroscopic level approaches to identify how empirical factors like users' and their neighborhood's influencing ability or a specific action's influencing capability and other user and network characteristics affect the *reach* quantity, and come to the conclusion that action dominates in the prediction of the spread of the action. Specifically, they quantify the *reach* $^{a}(u)$ of a user *u* as the number of cascades it can reach with a specific action α as:

$$reach^{\alpha}(u) = \sum_{u_i \in \vec{P}^{a}(u)} 1 + \frac{1}{2} * \sum_{u_j \in \vec{P}^{a}(u_i)} (reach^{\alpha}(u_j))$$

0, otherwise,

where a user gets the complete credit for action propagation to his immediate neighbours, or a decaying factor for non-immediate neighbours. $\vec{P}^{\alpha}(u)$ is the propagation set of user u, consisting of all his immediate neighbours u_i, such that there was an action propagation from u to u_i.

22.4 Online Social Network Measurements

OSNs can provide information, including the location of users, the items shared by users and structural and temporal properties of a social cascade. Finding ways of harnessing the potential of information constantly generated by users of OSNs is a key and promising research area for the networking community [31]. In this section, we investigate whether information extracted from social cascades can effectively be exploited to improve the performance of CDNs. Recently, thanks to the availability of large datasets, many studies have been presented. In the following subsections, we present some indicative large-scale analysis and media diffusion measurements that have been conducted in the context of OSNs and their findings have implications in CDN's performance. Section 22.4.1 focuses on OSN usage analysis, whereas Section 22.4.2 focuses on OSN media diffusion.

22.4.1 OSN Usage Analysis

A first large-scale analysis of multiple OSNs data encompassing Flickr, YouTube, LiveJournal, and Orkut, social networks for sharing photos, videos, blogs and profiles, respectively, by Mislove et al. [27] highlighted the difficulties of crawling a social network, and came to the following conclusions: Although node degrees in the studied OSNs varied by orders of magnitude, key findings are the same. The studied OSNs are power-law, small-world, scale-free, the in-degree matches out-degree distribution (due to link symmetry, an observation at odds with the web graph, that increases OSNs' network connectivity and reduces their diameter), there is a densely connected core of high degree nodes surrounded by small clusters of low-degree nodes, the average distances are lower, and clustering coefficients higher than those of the web graph (studied OSNs clustered 10.000 more times than random graphs, 5-50 times more than random power-law graphs).

Wilson et al. [36] conducted the first large-scale analysis of Facebook, by crawling and use of 'networks' (15% of total 10M users, and 24M. interactions). In [22], Kumar et al. study Flickr and Yahoo!360, finding that they follow power-law degree distributions. Low diameter and high clustering coefficient, as well as power-laws for in- and out-degree distributions were confirmed for the Twitter social graph by Java et al. in [21].

In terms of user workloads in OSNs, Benevenuto et al. [4] collected traces from a social network aggregator website in Brazil, enabling connection to multiple social networks with a single authentication, and, thus, studied Orkut, MySpace, Hi5 and Linked. Benevenuto et al. presented a clickstream model to characterise users' interactions, frequency and duration of connection, as well as frequency of users' transition to activities, such as browsing friends' profiles, sending messages etc., with their analysis showing that browsing, which cannot be identified from visible traces, is the most dominant behavior (92%). They also, reinforced the social cascade effect, since more than 80% of rich-media content like videos and photos was found through a 1-hop friend.

22.4.2 OSN Media Diffusion

Zhou et al., in [41], explore the popularity of photos in Facebook, noting that the request pattern follows a Zipf distribution, with an exponent $\alpha = 0.44$, significantly lower than that of traditional distributions (ranging from 0.64 to 0.83 [5]). They interprete this as shift of interest from popular items to items in a long tail. In the same context, Yu et al. [40] analyse PowerInfo, a Video On Demand system deployed by China Telecom and note that the top 10% of the videos account for approximately 60% of accesses, and the rest of the videos (the 90% in the tail) for 40%. Unaccessible via the official distribution channels (television networks or record companies) independent video content generated by the users, denoted as User Generated Content (UGC), becomes available to a wide number of viewers via services as YouTube or the US-based Vimeo. Cha et al. [7] investigate the long tail opportunities in the UGC services, such as YouTube video content, taking into account the fluctuation of the viewing patterns due to the volatile nature of the videos (videos may appear and disappear) and the various sources that direct to the content (recommendation services, RSS feeds, web reviews, blogosphere etc.)

In [39], Yang and Leskovec examine the temporal variations of Twitter hashtags and quotations in blogs, creating time series' of the number of mentions of an item *i* at time *t*, thus measuring the popularity given to the item *i* over time. By grouping together items so that item *i* is in the same group have a similar shape of the time series x_i with a clustering algorithm, they infer items with similar temporal pattern of popularity, and find that temporal variation of popularity of content in online social media can be accurately described by a small set of time series shapes, with most press agency news depicting a very rapid rise and a slow fading.

In a subsequent work [10], Christodoulou et al. confirm the higher impact of the social cascading effect on a more focused set of geographic regions, and, furthermore, study the social cascading effect of YouTube videos over Twitter users in terms of its impact on YouTube video popularity, dependence on users with a large number of followers, the effect of multiple sharing follows and the distribution of cascade duration. They come to the conclusion that the video retweet likelihood is increased as the number of user's follows who have already shared the same tweet increases, with the increase seeming to be exponential when the same tweet is shared by more than eight follows. This observation is consistent with [3], where it is claimed that the vast majority of YouTube videos do not spread at all, since large cascades are rare, and, finally, that links to videos can quickly spread over social networks, leading to many views in a short period of time. However, it should be noticed that Christodoulou et al. do not take factors such as the recency of the studied videos or their popularity in general into account.

From the above measurements it occurs that OSN content is different from more traditional Web content and affects significantly the navigation behavior of users.

Specifically, the above studies have shown that social cascading impacts the diffusion of information. In this context, CDNs can take advantage of the fact that social cascades have high impact on a more focused and less diverse set of geographic regions. Also, the findings regarding the temporal evolution of social cascades is a critical issue which affects the CDN performance.

22.5 Systems

In this section we present systems that could provide valuable insights concerning the exploitation of information extracted from OSNs for scaling of content diffused via OSNs. The cost for scaling of content tailored for a small number of users can be expressed in terms of required bandwidth for quick access to the content, or required storage capacity for caching of content, etc. Specifically, the long tails observed in Internet can impact system efficiency. For example, Facebook engineers developed an object storage system for Facebook's Photos application with the aim of serving the long tail of requests seen by sharing photos [1]. These optimizations are important, as requests from the long tail accounted for a significant amount of their traffic and most of these requests are served from the origin photo storage server, rather than by Facebook's CDN.

In the direction of distributing long-tailed content while lowering bandwidth costs and improving QoS, although without considering storage constraints, Traverso et al., in [35], exploit the time-differences between sites and the access patterns users follow. Instead of naively pushing User Generated Content (UGC) immediately, which may not be consumed and contribute unnecessarily to a traffic spike in the upload link, the system can follow a pull-based approach, when the first friend of a user in a Point of Presence (PoP) asks for the content. Moreover, instead of pushing content as soon as a user uploads, content can be pushed at the local time that is off-peak for the uplink, and be downloaded in a subsequent time bin, also off-peak for the downlink, and earlier than the first user in the PoP is bound to ask for it. The larger the difference between the content production bin and the bin in which the content is likely to be read, the better the performance of the system.

In [30], Sastry et al. built a prototype system, called Buzztraq, which leverages the information encoded in social network structure to predict users' navigation behaviour, which may be partly driven by social cascades. The key concept of Buzztraq is to place replicas of items already posted by a user closer to the location of friends, anticipating future requests. The intuition is that social cascades are spread rapidly through population as social epidemics. For instance, friends usually have common interests; consequently, if a user shares a video through his/her network, many of user's friends may find it interesting and share it to their network. Experimental results showed that social cascade prediction can decrease the cost of user access compared to location based placement, improving the performance of CDNs.

Zhou et al. [41] leverage the connection between content exchange and geographic locality (using a Facebook dataset, they identify significant geographic locality not only concerning the connections in the social graph, but also the exchange of content) and the observation that an important fraction of content is "created at the edge" (*is user-generated*), with a web based scheme for caching using the access patterns of friends. Content exchange is kept within the same Internet Service Provider (ISP) with a drop-in component, that can be deployed by existing web browsers and is independent of the type of content exchanged. Browsing users online are protected with *k*-anonymity, where k is the number of users connected to the same proxy and are able to view the content.

Instead of optimizing the performance of User Generated Content (UGC) services exploiting spatial and temporal locality in access patterns, Huguenin et al., in [20], show on a large (more than 650,000 videos) YouTube dataset that content locality (induced by the related videos feature) and geographic locality are in fact correlated. More specifically, they show how the geographic view distribution of a video can be inferred to a large extent from that of its related videos, proposing a UGC storage system that proactively places videos close to the expected requests. Such an approach could be extended with the leverage of information from OSNs, in the way that Figure 22.4 depicts.

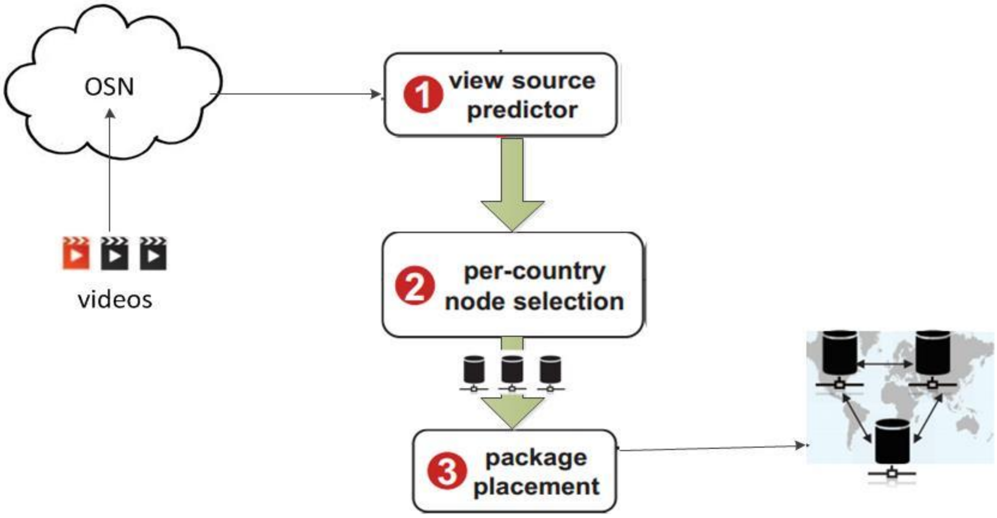

Figure 22.4. Overview of VOD service placement strategy leveraging OSNs

22.6 Future Research Directions

In order to harness the power of social networks diffusion over CDN infrastructures, the key areas of interest that need to be explored include the large-scale data sets, the OSN evolution and semantic annotation.

22.6.1 Large-scale Data Sets

The amount of information in OSNs is an obstacle, since elaborate manipulation of the data may be needed. An open problem is the efficient handling of graphs with billions of nodes and edges [44]. Facebook, for example, reported that it had one billion monthly active users as of October 2012 and 604 million monthly active users who used Facebook mobile products as of September 2012.

In order to generate aggregations and analyses that have meaning, the Facebook custom built-in data warehouse and analytics infrastructure has to apply ad-hoc queries and custom MapReduce jobs [12] in a continuous basis on over half a petabyte of new data every 24 hours, with the largest cluster containing more than 100PB of data and the process needs surpassing the 60.000 queries in Hive, the data warehouse system for Hadoop and Hadoop compatible file systems.

The desired scaling property refers to the fact that the throughput of the presented approaches should remain unchanged with the increase in the data input size, such as the large datasets that social graphs comprise and the social cascades phenomena that amplify the situation. The cost of scaling such content can be expressed in different ways. For instance, in the case of CDNs, it can be the number of replicas needed for a specific source, or it may take into account the optimal use of memory and processing time of a social-aware built system, etc.

22.6.2 OSN Evolution

Existing works examine valuable insights into the dynamic world by posing queries on an evolving sequence of social graphs (e.g., [28]) and time evolving graphs tend to be increasingly used as a paradigm also for the emerging area of OSNs [16]. However, the ability to process queries concerning the information diffusion in a scalable way remains to a great extent unstudied. With the exception of sporadic works on specialized problems, such as that of inference of dynamic networks based on information diffusion data [29], we are not aware of relative studies on the information diffusion through OSNs under the prism of graphs dynamicity.

22.6.3 Semantic Annotation

It would also be interesting to know which social cascades will evolve as global and which of them will evolve as local, possibly making some associations with their content or context features. It is challenging to discover contextual associations among the topics, which are by nature implicit in the user-generated content exchanged over OSNs and spread via social cascades. In other words, we would like to derive semantic relations. This way the identification of a popular topic can be conducted in a higher, more abstract level with the augmentation of a semantic annotation. While we can explicitly identify the topic of a single information disseminated through an OSN, it is not trivial to identify reliable and effective models for the adoption of topics as time evolves ([17], [25]) characterized with some useful emergent semantics. Such knowledge would improve caching of Web content in CDN infrastructures [31].

To sum up, OSNs create a potentially transformational change in users' behavior. This change will bring a far-reaching impact on traditional industries of content, media, and communications. In this context, the rapid proliferation of OSNs sites is expected to reshape CDN's structure and design [18]. Investigating the geographical, structural and temporal properties of social cascades, new CDN infrastructures will be built where cache replacement strategies will exploit these properties. Traditional CDN systems support content distribution with specific needs, such as efficient resource discovery, large-scale replication of popular resources that follow zipfian resource-popularity distributions, and simple access-rights. In contrast, the next generation of CDN systems are required to support a variety of social interactions conducted through an open-ended set of distributed applications, going beyond resource discovery and retrieval and involving: synchronous and asynchronous messaging; "push" and "pull" modes of information access; finer access control for reading and writing shared resources; advanced mechanisms for data placement, replication and distribution for a large variety of resource types and media formats.

22.7 Conclusions

Understanding the effects of social cascading on content over the Web is of great importance towards improving CDN performance. Given the large amount of available resources, it is often difficult for users to discover interesting content. Relying on the suggestions coming from friends seems to be a popular way to choose what to see.

Taking into account the increasing popularity of Online Social Networks and the growing popularity of streaming media, we have presented existing approaches that can be leveraged for the scaling of rich media content in CDNs using information from OSNs.

Acknowledgements

Irene Kilanioti is a PhD candidate at the University of Cyprus whose work is supported by the Greek State Scholarships Foundation, Lifelong Learning Programme (Customized Evaluation Process).

22.8 References

[1] D. Beaver, S. Kumar, H. C. Li, J. Sobel, and P. Vajgel. Finding a needle in haystack: Facebook's photo storage. In Proceedings of OSDI 10. 9th USENIX Symposium on Operating Systems Design and Implementation, Proceedings Usenix, Vol. 10, pp. 1-8, Vancouver, BC, Canada, 2010.

[2] E. Bakshy, J.M. Hofman, W.A. Mason, and D.J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international Conference on Web search and data mining*, Kowloon, Hong Kong, 2011.

[3] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international Conference on World Wide Web*, Lyon, France, 2012.

[4] F. Benevenuto , T. Rodrigues, M. Cha, M. , and V. Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet measurement*, Chicago, IL, USA, 2009.

[5] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, New York, NY, USA, 1999.

[6] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international Conference on World Wide Web*, pages 241–250, Lyon, France, 2012.

[7] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet measurement*, San Diego, CA, USA, 2007.

[8] M. Cha, A. Mislove, B. Adams, and K.P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, Seattle, WA, USA, 2008.

[9] K. Chard, ,S.Caton, O. Rana, and K. Bubendorfer. Social cloud: Cloud computing in social networks. In *Proceedings of the 3rd IEEE International Conference on Cloud Computing (CLOUD)*, Miami, FL, USA, 2010.

[10] G. Christodoulou, C. Georgiou, and G. Pallis. The role of twitter in youtube videos diffusion. In *Proceedings of the 13th International Conference on Web Information System Engineering (WISE)*, Paphos, Cyprus, 2012.

[11] K. S. Dave, R. Bhatt, and V. Varma. Modelling action cascades in social networks. In *Proceedings of the 5th AAAI International Conference on Weblogs and Social Media*, Barcelona, Spain, 2011.

[12] J. Dean, and S. Ghemawat,. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[13] D.J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

[14] M. Draief, A. Ganesh, and L. Massoulié. Thresholds for virus spread on networks. In *Proceedings of the 1st ACM international Conference on Performance evaluation methodologies and tools*, Pisa, Italy, 2006.

[15] D. Easley, and J. Kleinberg. *Networks, crowds, and markets*. Cambridge University Press, 2010.

[16] A. Fard, A. Abdolrashidi, L. Ramaswamy, and J. A. Miller. Towards efficient query processing on massive time-evolving graphs. In *Proceedings of the 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, Pittsburgh, PA, United States, 2012.

[17] A. Garcá-Silva, J.-H. Kang, K. Lerman, and O. Corcho. Characterising emergent semantics in twitter lists. In *Proceedings of the 9th international conference on The Semantic Web: research and applications (ESWC)*, Heraklion, Greece, 2012.

[18] L. Han, M. Punceva, B. Nath, S. Muthukrishnan and L. Iftode. Socialcdn: Caching techniques for distributed social networks. In *Proceedings of* IEEE 12th International Conference on Peer-to-Peer Computing (P2P), 2012 (pp. 191-202). IEEE, Tarragona, Spain, 2012.

[19] T. Hogg, and K. Lerman. Stochastic models of user-contributory web sites. In *Proceedings of 3rd international AAAI Conference on Weblogs and Social Media*, San Jose, CA, USA, 2009.

[20] K. Huguenin, A.-M. Kermarrec, K. Kloudas, F. Taäni, and others,. Content and geographical locality in user-generated content sharing systems. In *Proceedings of 22nd SIGMM International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, Toronto, Canada, 2012.

[21] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th ACM WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, San Jose, CA, USA, 2007.

[22] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. *Link Mining: Models, Algorithms, and Applications*, pages 337–357, 2010.

[23] T. La Fond, and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international Conference on World Wide Web*, Raleigh, NC, USA, 2010.

[24] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of SIAM International Conference on Data Mining (SDM) 2007*, Minneapolis, Minnesota, USA, 2007.

[25] C. X. Lin, Q. Mei, Y. Jiang, J. Han, and S. Qi. Inferring the diffusion and evolution of topics in social communities. In *Proceedings of ACM SIGKDD Workshop on Social Network Mining and Analysis (SNAKDD)*, San Diego, CA, USA, 2011.

[26] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, pages 415–444, 2001.

[27] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet measurement*, San Diego, CA, USA, 2007.

[28] C. Ren, E. Lo, B. Kao, X. Zhu, and R. Cheng. On querying historical evolving graph sequences. In *Proceedings of the 37th International Conference on Very Large Data Bases (VLDB)*, 2011.

[29] M.G. Rodriguez, J. Leskovec, and B. Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM)*, Rome, Italy, 2013.

[30] N. Sastry, E. Yoneki, and J. Crowcroft. Buzztraq: predicting geographical access patterns of social cascades using social networks. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, Nuremberg, Germany, 2009.

[31] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft. Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades. In *Proceedings of the 20th international Conference on World Wide Web*, Hyderabad, India, 2011.

[32] D. Stauffer and A. Aharony. *Introduction to percolation theory*. CRC, 1994.

[33] G. V. Steeg, R. Ghosh, and K. Lerman. What stops social epidemics? In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM)*, Barcelona, Spain, 2011.

[34] D. Towsley, A. Rao, Y.-S. Lim, C. Barakat, A. Legout, and W. Dabbous. Network characteristics of video streaming traffic. In *Proceedings of the 7th Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, New York, NY, USA, 2011.

[35] S. Traverso, K. Huguenin, I. Triestan, V. Erramilli, N. Laoutaris, and K. Papagiannaki. Tailgate: handling long-tail content with a little help from friends. In *Proceedings of the 21st international Conference on World Wide Web*, Lyon, France, 2012.

[36] C. Wilson, B. Boe, A. Sala, K.P. Puttaswamy, and B.Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer systems*, Nuremberg, Germany, 2009.

[37] F. Wu, and B.A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

[38] J. Yang, and J. Leskovec. Modeling information diffusion in implicit networks. In *Proceedings of 2010 IEEE 10th International Conference on Data Mining (ICDM)*, Sydney, Australia, 2010.

[39] J. Yang, and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international Conference on Web search and data mining*, Kowloon, Hong Kong, 2011.

[40] H. Yu, D. Zheng, B.Y. Zhao and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. In *ACM SIGOPS Operating Systems Review*, volume 40, pages 333–344, 2006.

[41] F. Zhou, L. Zhang, E. Franco, A. Mislove, R. Revis, and R. Sundaram. Webcloud: Recruiting social network users to assist in content distribution. In *Proceedings of IEEE International Symposium on Network Computing and Applications*, Cambridge, MA, USA, 2012.