Rethinking web content distribution in the social media era

Alan Mislove, Northeastern University

We are witnessing the beginnings of a shift in the patterns of content creation and exchange over the web. Previously, web content—including web pages, images, audio, and video—was primarily created by a small set of entities and was delivered to a large audience of web users. However, recent trends such as the rise in popularity of online social networking; the ease of content creation using digital devices like smartphones, cameras, and camcorders; and the ubiquity of Internet access have democratized content creation. Now, individual Internet users are creating content that makes up a significant fraction of Web traffic [3, 9].

As a result, compared to content shared over the web just a few years ago, content today is generated by a large number of users located at the edge of the network, is of more uniform popularity, and exhibits a workload that is governed by the social network. Unfortunately, existing content distribution architectures—built to serve more traditional workloads—are ill-suited for these new patterns of content creation and exchange. For example, web caches have been shown to exhibit poor performance on social networking content [5, 18], due to the more uniform popularity of content, causing many online social networking sites have begun to move away from content distribution networks (CDNs) and towards highly-engineered in-house delivery solutions [10, 11, 15].

Changing workload patterns

To more closely examine these trends, we examine a data set on the photos exchanged by 63,731 users from the New Orleans Facebook regional network [14]. Because data on photo *views* is not available, we use photo comments as a proxy for views (i.e., if a user has commented on a photo, they must have viewed it). Crawling the news feed in a manner similar to previous work [16], we discovered information on a total of 1,068,787 comments placed on 816,508 different photos. While we only analyze this dataset due to lack of space, we have found similar results on other social networks.

Content is created at the edge We first explore *where* the emerging content being exchanged over the web is being created. Today, the rapid adoption of smartphones, digital cameras, digital camcorders, and professional-quality music and video production software, combined with the low cost of broadband Internet service, has greatly eased content creation by individual users. Significantly more news articles are written by bloggers than news organizations [12], more photos are shared on online social networks [8] than on professional photography websites [1], and much of the content shared on YouTube, the most popular video-sharing site, is created by end users [6, 7] empowered by the ubiquity of webcams.

Content is of more uniform popularity We now explore the *popularity distribution* of the content in emerging workloads, relative to previous workloads. To do so, we examine the popularity of photos on Facebook, comparing it to the popularity distribution to that observed in studies of traditional web workloads [4]. We note one primary distinction with respect to traditional workloads: The Facebook workload contains a significantly lower exponent of the Zipf distribution (0.44, compared to 0.64 to 0.83 [4]), implying less emphasis on popular items and resulting in a more uniform popularity distribution and a significantly longer, fatter tail (Figure 1).

Exchange is governed by the social network We turn to explore *how* users are locating content. In particular, we explore the degree to which the exchange of content is governed by the structure of the

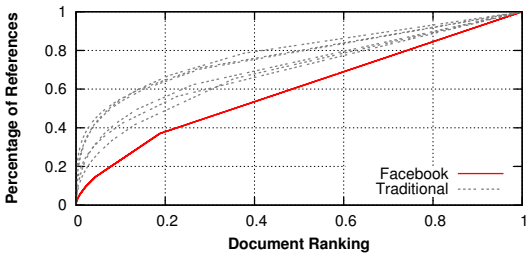

Figure 1: Cumulative distribution of Facebook views compared to five traditional web workloads [4].

social network. To do so, we calculate the fraction of comments on photos that come from the local social network of the uploader. The result of this analysis is that over 28.3% of the comments are placed by friends of the uploader, and at least 89.1% are placed by friends or friends-of-friends (compared to expected values of 0.04% and 0.30%, respectively, were the placement random). This indicates that users are significantly more interested in the content that is uploaded by their friends and friends-of-friends.

Exchange has significant geographic locality Finally, we explore the connection between content exchange and *geographic locality*. Using our Facebook data set, we find that 32.9% of the friends of New Orleans users are also in the New Orleans network; similar findings have been observed in other regional networks [16]. However, if we examine the fraction of content exchange that occurs between New Orleans network users, we observe that 51.3% of comments are placed by other users within the New Orleans regional network, even though only 32.9% of the friendship relationships lie within the network. This indicates that the significant geographic locality already present in social networks is present to an even greater degree in the content exchange that occurs over these networks [17].

Rethinking content distribution

The content that is increasingly being shared on the web today is created at the edge of the network, but is exchanged using centralized infrastructure. The usefulness of existing techniques on this workload is declining [5, 17, 18]: For example, caching the most popular 10% of the items in traditional workloads would satisfy between 55% [4] and 95% [2] of the requests; in our social network workload from the previous section, such a cache would only satisfy 27% of the requests. This also affects the ability to use CDNs, which similarly work best for popular content.

We therefore propose to work towards more decentralized content exchange over the web. While some have suggested decentralizing the provider's data center architecture [17] into many regional data centers, this requires significant changes and expense for the provider. Instead, we propose to focus on retaining the centralized provider architecture of today, while attempting to decentralize content exchange when possible.

In ongoing work, we are building WebCloud, a content distribution system designed to support the workloads present in existing online social networking websites. WebCloud works by recruiting users' web browsers to help serve content to other users and is compatible with the web browsers and web sites of today. Due to the geographic locality that often exists between friends in online social networks [13, 16], content exchange in WebCloud often stays within the user's local Internet Service Provider (ISP), thereby providing a bandwidth savings for both the site and the ISP. As a result, by deploying WebCloud, OSNs such as Facebook would enjoy most of the benefits of large centralized CDNs with lower costs and their users would benefit from faster service.

References

[1] 4,000,000,000 ≪ Flickr Blog. http://blog.flickr.net/en/2009/10/12/4000000000/.

[2] M. Arlitt and T. Jin. Workload Characterization of the 1998 World Cup Web Site. *IEEE Network*, 14(3), 2000.

[3] Alexa Top 500 Global Sites. http://www.alexa.com/topsites.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web Caching and Zipf-like Distributions: Evidence and Implications. *INFOCOM*, 1999.

[5] G. Cormode and B. Krishnamurthy. Key Differences between Web 1.0 and Web 2.0. *First Monday*, 13(6), 2008.

[6] 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. *IMC*, 2007.

[7] X. Cheng, C. Dale, and J. Liu. Statistics and Social Network of YouTube Videos. *IWQoS*, 2008.

[8] Facebook Statistics. http://www.facebook.com/press/info.php?statistics.

[9] Facebook and YouTube dominate workplace traffic and bandwidth. http://www.scmagazineuk.com/facebook-and-youtube-dominate-workplace-traffic-and-bandwidth/article/168082/.

[10] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube Traffic Characterization: A View from the Edge. *IMC*, 2007.

[11] N. Kennedy. Facebook's photo storage rewrite. http://www.niallkennedy.com/blog/2009/04/facebook-haystack.html.

[12] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cyle. *KDD*, 2009.

[13] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33), 2005.

[14] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in Online Social Networks. *WSDM*, 2010.

[15] P. Vajgel. Needle in a haystack: efficient storage of billions of photos. http://www.facebook.com/note.php?note_id=76191543919.

[16] C. Wilson, B. Boe, A. Sala, K. P.N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. *EuroSys*, 2009.

[17] M. P. Wittie, V. Pejovic, L. Deek, K. C. Almeroth, and B. Y. Zhao. Exploiting Locality of Interest in Online Social Networks. *CoNEXT*, 2010.

[18] M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications. *MMCN*, 2008.