A Cliq of Content Curators

Angela H. Jiang
Northwestern University

Zachary S. Bischof
Northwestern University

Fabian E. Bustamante
Northwestern University

ABSTRACT

A social news site presents user-curated content, ranked by popularity. Popular curators like Reddit, or Facebook have become effective way of crowdsourcing news or sharing for personal opinions. Traditionally, these services require a centralized authority to aggregate data and determine what to display. However, the trust issues that arise from a centralized system are particularly damaging to the "Web democracy" that social news sites are meant to provide.

In this poster, we present **cliq**, a decentralized social news curator. cliq is a P2P based social news curator that provides private and unbiased reporting. All users in cliq share responsibility for tracking and providing popular content. Any user data that cliq needs to store is also managed across the network. We first inform our design of cliq through an analysis of Reddit. We design a way to provide content curation without a persistent moderator, or usernames.

Categories and Subject Descriptors

C.2.4 [**Communication Systems Organization**]: Distributed Systems

Keywords

P2P systems, social news curation

1. MOTIVATION

Unlike other media, content shown on social media sites is determined by a "Web democracy", open to anyone. However, the trust issues that arise from a centralized system are particularly damaging to this goal. Users must accept that owners of these sites have the power to present content aligned with their own agenda. For example, BBC news recently reported that moderators of the subreddit "technology" were censoring posts that included banned words such as "bitcoin" or "net neutrality"[1].

[1] `http://www.bbc.com/news/technology-27100773`

SIGCOMM'14, August 17–22, 2014, Chicago, IL, USA.
ACM 978-1-4503-2836-4/14/08.
http://dx.doi.org/10.1145/2619239.2631470.

Further, centralized management of user data leads to concerns of personal information being sold or intruded upon. In recent years, citizens have learned of government programs in multiple countries that crawl the web and archive information off of such websites. In fact, there exist multiple public services that given a username, scrape reddit and analyze the user's behavior; even showing the topics most participated in and "Fun Guessed Data" such as if the user supports Occupy Wall Street[2]. Oppressive governments may act on such data, tracking individuals. Such a threat can affect a user's willingness to share content that is political or controversial.

An effective, fair and safe Web democracy should (*i*) present content that is relevant to the user's interests and (*ii*) popular, be (*iii*) resilience to biased reporting and (*iv*) to user data aggregation. Existing social news sites have provided an effective way to accomplish the first two goals. However, their centralized nature raises concerns about content bias and users privacy. cliq leverages the anonymity and autonomy that a decentralized systems provides to fulfill these additional requirements.

2. CLIQ ARCHITECTURE

We now present the design of cliq, a fully-distributed social news site. By using a P2P model, the responsibility of tracking content and popularity is randomly distributed among peers. We remove the need for a (potentially biased) dedicated moderator. To provide anonymity, Cliq users do not have usernames. Any user information collected is distributed among random nodes and will naturally expire as the network experiences churn. A fully decentralized user content curator, however, presents its own challenges.

2.1 Managing content

We organize content through cliq similarly to what is common in most social news sites; each post is associated with one or more category tags (e.g. politics, sports, or travel). We build cliq over Kademlia [1] a popular DHT. In the DHT, nodes use content tags as keys to store and find content in the network. For each tag, a subset of nodes with similar IDs are collectively responsible for managing its content. These duties include storing, ranking and disseminating links uploaded by users with that tag, as well as monitoring for malicious nodes.

[2] `http://www.redditinvestigator.com/`

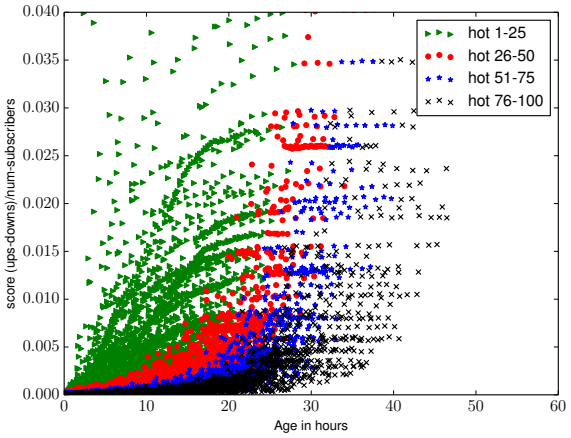

Figure 1: Popular posts on Politics subreddit. Quartiles correspond to each of the first 4 pages of the subreddit.

2.2 Load on users

As with any P2P system, we need to consider the potential load placed on cliq users. To do this, we monitor and measure activity on the social news site, reddit. We choose reddit because of its popularity (ranked 61 globally on Alexa) and very active user base.

Reddit is organized into user-created, topic-based "subreddit" communities. Subreddits serve a similar function that tags do in cliq. We look at the distribution of the sizes of the most popular subreddits. Among over 300,000 subreddits, only 212 have over 100,000 subscribers. 12 subreddits have over 1,000,000. Although the rate of activity is small compared to the number of subscribers, we find that average sized subreddits (>100,000 subscribers) can expect a popular post to gain thousands of votes within two days. We therefore expect to perform dynamic load balancing in cliq to account for a large variance in tag activity.

2.3 Determining popularity

In order to find an effective way for ranking content, we analyze popular posts of multiple subreddits. Here, popular posts refers to content on the first four pages (100 posts) of a subreddit. We find that popularity of reddit posts are strongly correlated with age of the post and it's voting score. Fig. 1 shows the age and score (normalized by the subreddit size) of popular posts of the "Politics" subreddit. Among the 100 most popular posts, we rarely see posts older than two days. This is fairly consistent among subreddits with over 100,000 subscribers.

However, ranking by score is not as straightforward, especially in cliq. Cliq tags differ from subreddits in that posts can be linked to multiple tags. However, we find that unnormalized scores of popular content vary drastically between subreddits. Therefore, a clliq node keeps track of a post's score and how many times it has shown this post to users. In effect, we rank posts by their "approval rate" rather than raw score to account for varying amounts of exposure. This feature will also be important to cliq's resilience to spam.

2.4 Resilience to spam

Without registered users or the vantage point of a centralized system, we cannot use traditional methods of detecting spammers. Instead, we take advantage of the fact that in the Kademlia P2P model, a node will be routed to a tag along the same nodes each time. Nodes along this path can collectively enforce good behavior from the users. For example, each node can store the set of IPs it has personally seen promote each link. If a user attempts to promote the same URL multiple times, the request will be dropped. Many of the randomly selected nodes along the path would need to be malicious to allow a single IP post many times. Promoters would not have to be shared among nodes but to further preserve privacy, storing promoters can be done probabilistically.

2.5 Replication

As we previously saw, nodes need to be able to dynamically offload responsibilities to other nodes. In cliq, a node can request a neighbor to replicate its most popular content. Therefore, when looking for content, a user is more likely to find popular posts. However, a malicious node could easily spam cliq with the ability to direct others to replicate arbitrary data. To account for this, if a node is asked to replicate data, it does not do so immediately. It first waits until it personally receives an upvote for the link to verify that it is indeed popular content.

Upon replicating, approval rating of the post starts back at 0. Therefore, even if enough nodes are colluding to induce replication, we would still not expect the post to get popular. Meanwhile, a post which is actually a trending will be able to regain its high ratio quickly.

3. ACKNOWLEDGEMENTS

This work is supported in part through NSF awards CNS 1211375 and CNS 1218287. The authors thank Andrew Kahn for his help implementing our first prototype of cliq.

4. REFERENCES

[1] P. Maymounkov and D. Mazieres. Kademlia: A peer-to-peer information system based on the xor metric. In *Proceeding IPTPS '01 Revised Papers from the First International Workshop on Peer-to-Peer Systems*, 2001.