Modeling the Flow of Information in a Social Network

Sarita Yardi, Amy Bruckman
School of Interactive Computing
Georgia Institute of Technology
yardi,asb@cc.gatech.edu

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous

General Terms

Design

Keywords

Social networks, viral sharing

1. INTRODUCTION

Recent measurements of online social networks have observed a number of shared structural characteristics, such as power law distributions, scale-free properties, graph evolution, and information cascades (e.g. [9, 2, 1]), and, more recently, real-world networks such as blogs and viral outbreaks [8] and temporal patterns in networks (e.g. [3]). However, these studies measure networks as a function of nodes and links rather than a function of the subjective and individual relationships in the network.

There are fewer examples of real-world applications that leverage the explicitly social nature of users and content in a graph for measuring trust in distributed systems. Specifically, how do levels of trust influence who people decide to talk to and what they share? Does trustworthiness scale, such that information flow is optimized not only through friend relationships, but also by friend-of-friend associations, network affiliations (e.g. Atlanta, Georgia Tech), and shared groups (e.g. RamblinWreck Fans)? Are there ways of detecting node outliers, subgraph borders, or link decays, that cause communication breakdowns?

1.1 The Complex Nature of Friendships

Facebook networks often reflect real-world social graphs more closely than related sites such as MySpace and Orkut. In contrast to these less-structured sites, the technical and social design of Facebook encourages users to articulate personal identity markers and existing relationships by joining networks, groups, and filling out profile fields [6]. This articulation of one's real-world networks, along with the public and transparent nature of news feeds, wall postings, status updates, and friendships establishes shared trust and accountability. Thus, selecting which profile elements to fill out, who to grant profile access to, and at what granularity, enables rudimentary identity management; however, relationships in Facebook are reduced to a binary representation–a "friend"–in their social graph. The semantic meaning of "friend" is highly subjective and contextual and requires a more complex representation.

2. MINING THE SOCIAL NETWORK

Web scraping and log file methods are well-suited to studying large graphs of non-human binary representations, but may be insufficient for understanding the contextualized semantics of a social network. Such networks are culturally rich, dynamically evolving viral ecosystems, with individual and community-level norms and practices. We can model when and where information flows, but we have a poor sense of why and how.

Similarly, while marketing research is targeted towards understanding the factors that drive mass media consumption—what motivated over six million people to add Facebook's Vampires application or 20 million people to view YouTube's Free Hugs Campaign? (e.g. [7])—we know far less about the factors that influence mass media production. What factors motivate people to become content creators, and to learn new skills through their participation in these networked environments? Social networking sites that are characterized by structured peer-to-peer connectivity, transparency of activity, creative expression, and a permeable barrier to entry can be especially rich breeding grounds for viral spread.

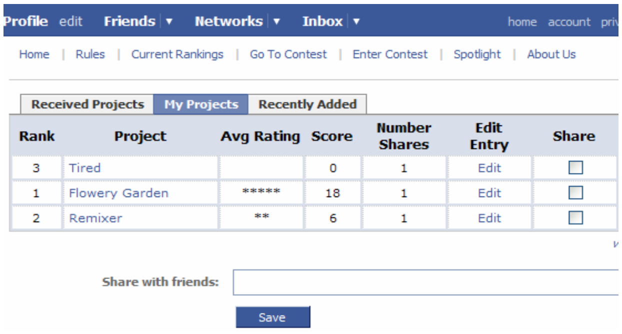
2.1 The Facebook API

We have developed a Facebook application to track viral sharing patterns and are beginning real-world deployment in Fall 2008[1]. We are conducting a series of reward-incentivized contests in which users–teens and college students–participate in short activities and games, then pass the activity on to their friends. Using the Facebook API, we are tracking time-sequenced snapshots of sharing patterns in combination with evolving friend network characteristics to capture bursty sharing patterns and stickiness of influence.

SIGCOMM'08, August 17–22, 2008, Seattle, Washington, USA.
ACM 978-1-60558-175-0/08/08.

[1] Supported by NSF-BPC Grant #0634629.

If l is the target node, G is the set of all friends in l's social graph, and G_A is the subset of friends with whom projects are shared, what are the characteristics of this subset, $G_A \subset G$, that motivate information flow? Datetime stamps of each click, project view, share, and related action of first-time project creators are tracked in order to quantitatively model breadth and depth (based on sharing activity) and link strength (based on personal node characteristics).



Profile edit Friends ▾ Networks ▾ Inbox ▾
home account priv
Home | Rules | Current Rankings | Go To Contest | Enter Contest | Spotlight | About Us
Received Projects | My Projects | Recently Added
Save

Rank	Project	Avg Rating	Score	Number Shares	Edit Entry	Share
3	Tired		0	1	Edit	□
1	Flowery Garden	*****	18	1	Edit	□
2	Remixer	**	6	1	Edit	□

Share with friends:

Figure 1: Facebook application

Breadth
How many friends F_{1-n} in subset G_A share information with l?

$$\sum_{F=1}^{n} f(F) = f(1) + f(2) + f(3) + \ldots + f(n)$$

Depth
How often does each friend F_n in G_A share with l?

$$\sum_{P=1}^{m} f(P) = f(1) + f(2) + f(3) + \ldots + f(m)$$

Link Strength
What is the strength of each link where TS_{lF} is the weighted aggregate of l and F's link strength based on shared friends, groups, networks, and wall postings?

$$\sum_{F=1}^{n} f(TS_{lF}) = f(TS_{l1}) + f(TS_{l2}) + \ldots + f(TS_{lF})$$

3. CONTRIBUTION

Facebook has over 70 million users and more than 65 billion page views per month [4]; designing authentication-based trust protocols to ensure privacy, without restricting sociality and connectedness, becomes an increasingly important goal. We hypothesize that trust will be motivated by link strength in the network as measured by: degree centrality (number of friends) clustering coefficient (shared friends), similarity (e.g. shared Interests and shared groups/networks with weighted importance based on inverse size of group/network), and node influence (viral spread).

We propose that a classification tree model can represent the subjective, non-linear nature of information flow based on these parameters. We build this learning decision tree using a top-down, greedy search, using training cases from our application. Decisions are weighted by breadth, depth, and link strength, based on binary relationships as well as derived critical thresholds for non boolean attributes (e.g. Similarity). We expect to see local trends based on demographics, cliques, interests, and hobbies, that then give way to generalized patterns such as types of users and common sharing behaviors.

3.1 Challenges

Mining a network using an API introduces a distinct class of data collection techniques, with unique methodological challenges and privacy implications [5]. Facebook's Developer Terms of Services prohibits web scraping, caching user data, and storing user data in a data repository without explicit user permission[2].

Another challenge is that social networks are human, complex, and unpredictable and the data will be noisy. Furthermore, friendship cannot be represented as a Boolean classification. For an attribute that is continuous-valued, such as age, the algorithm can dynamically create a new Boolean attribute A that is true if $A > b$ and $A < c$ and false otherwise, where the question is what is the best value for thresholds b and c that produces the greatest information gain. However, there are multiple classes of friendships (and also groups and networks) that have no linear, continuous value; thus, our challenge is to maximize information gain, represented as flow of data between nodes, while accounting for the imprecise nature of these elements.

4. REFERENCES

[1] A. Barabasi. Scale-free networks. *Scientific American*, 288:60–69, 2003.

[2] P. Domingos and M. Richardson. Mining the network value of customers. In *Proc. ACM SIGKDD '01*, 2001.

[3] D. W. Golder, Scott and B. Huberman. Rhythms of social interaction: Messaging within a massive online network. In M. A. . N. C. C. Steinfield, B. Pentland, editor, *Third International Conference on Communities and Technologies*, pages 41–46, East Lansing, MI., 2007. London: Springer.

[4] http://www.facebook.com.

[5] J. M. Kleinberg. Challenges in mining social network data: processes, privacy, and paradoxes. In *Proc. 13th ACM SIGKDD*, San Jose, CA, 2007.

[6] C. A. Lampe, N. Ellison, and C. Steinfield. A familiar face(book): profile elements as signals in an online social network. In *Proc. CHI*, pages 435–444, San Jose, Ca, 2007. ACM.

[7] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.

[8] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proc. 13th ACM SIGKDD*, San Jose, CA, 2007.

[9] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. 7th ACM SIGCOMM*, San Diego, CA, 2007.

[2] See Developer Terms of Service Sections 2.A.4, 2.A.5. We have human subjects approval to study profile information and sharing patterns and store profile data in a backend local database.