CS 294-1 Project Proposal: Detecting Spam on Social Networking Sites

Antonio Lupher, Cliff Engle, Reynold Xin

{alupher, cengle, rxin}@cs.berkeley.edu

1. PROBLEM

Most social networks of any significant size see constant spam, scams and phishing attacks. The nature of these attacks can be quite diverse and difficult detect. Marketers can spam members with unwanted advertisements, fraudsters lure users with advance fee frauds and other confidence tricks, while others attempt to steal user information by directing users to external phishing pages. To make matters more difficult, a site with global reach sees communication among its members in a number foreign languages with varying levels of ability. This means that much benign content shares characteristics like misspellings, awkward phrases, etc. that might have made certain types of common frauds and spam more easy to distinguish on US-based (or English-language) sites.

2. APPROACH

This project will examine in detail the types of malicious and benign content that are encountered on social networks by analyzing experimental data available from InterPals, an international social network for cultural exchange and language practice. For example, the site attracts a wide variety of financial scams, ranging from Nigerian "419" scams to romance scams. Another prevalent problem is spam with links to third-party websites, directing users to various porn/webcam sites, phishing sites or various untrustworthy online marketplaces.

We will then examine various methods of detecting and preventing abuse on the site, including those measures that have already been taken (e.g. various heuristics including IP/location anomaly detection, frequency capping, duplicate account detection, etc.). However, the main focus will be on mining experimental data from the site and using features derived from this data to build and evaluate classifiers to detect unwanted behavior programmatically. The large volume of data available to us will provide a unique perspective both on the types of malicious content that exist on such sites as well as on the effectiveness of classifier/learning-based approaches to identifying these activities.

3. DATA SETS

We plan to make heavy use of our unrestricted access to the data of InterPals, which has over 1.2 million active members. This data includes a corpus of 90 million private messages and another 1.5 million messages that have been labeled as spam by users. Other data includes 40 million or so "wall" comments, 5 million photos, and 8 million photo comments.

4. TECHNIQUES

Using this data, we plan to first identify the most prevalent types of malicious activities on the site. We will investigate various machine learning techniques to automatically detect these activities, including sampling, clustering and classifiers. Sampling and clustering, in addition to the user-labeled and moderator-verified spam corpus, will help us identify training data.

Classifiers that we plan to explore include:

- Gradient boosting
- Naive Bayes
- SVMs
- Decision trees

We will then test the classifiers on new site data to evaluate their performance.

One important part of this project will be to identify features of undesirable activity that are useful in classification. We plan to examine:

- Profile and message data (keywords, n-grams, age, sex, registration date, amount of data in profile, etc.)
- User info (geographical location, IPs, browser settings)
- Complaints and user labeling of spam
- Profile photographs
- User reputation, social graph

5. FRAMEWORKS

We hope to implement the detection algorithms using Spark, an in-memory distributed computing framework that is particularly well-suited for machine learning and iterative computations.